Toward Automatic Rhodopsin Modeling as a Tool for High-Throughput Computational Photobiology

Towards Automatic Rhodopsin Modeling as a Tool for High-throughput Computational Photobiology

Federico Melaccio[†,*], María del Carmen Marín[†,¶], Alessio Valentini[†], Fabio Montisci[†], Silvia Rinaldi[†], Marco Cherubini [†], Xuchun Yang,[¶] Yoshitaka Kato[¶¶], Michael Stenrup,[‡] Yoelvis Orozco-Gonzalez,[#,¶] Nicolas Ferré,[‡] Hoi Ling Luk[¶], Hideki Kandori[¶¶] and Massimo Olivucci[†,¶,#,##*]

†Department of Biotechnology, Chimistry e Pharmacy, Università di Siena, via A. Moro 2, I-53100 Siena, Italy, ¶Department of Chemistry, Bowling Green State University, Bowling Green, OH 43403, USA, ¶¶Department of Frontier Materials, Graduate School of Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Japan, ‡Aix Marseille Univ, CNRS, ICR, Marseille, France, #Institut de Physique et Chimie des Matériaux de Strasbourg, UMR 7504 Université de Strasbourg-CNRS, F-67034 Strasbourg, France, ##USIAS Institut d'Études Avancées, Université de Strasbourg, 5 allée du Général Rouvillois, F-67083 Strasbourg, France.

Abstract

We report on a prototype protocol for the automatic and fast construction of congruous sets of QM/MM models of rhodopsin-like photoreceptors and of their mutants. In the present implementation the information required for the construction of each model is essentially a crystallographic structure or a comparative model complemented with information on the protonation state of ionizable side-chains and distributions of external counterions. Starting with such information a model formed by a fixed environment system, a flexible cavity system and a chromophore system is automatically generated. The results of the predicted vertical excitation energy for 27 different rhodopsins including vertebrate, invertebrate and microbial pigments indicate that such basic models could be employed for predicting trends in spectral changes and/or correlate the spectral changes with structural variations in large sets of proteins.

Introduction

Cheap and fast tools for gene sequencing, protein expression and analysis are routinely used for high-throughput genomics.[1,2] However, due to experimental difficulties and longer timescales (e.g. for crystallization), protein structure determination cannot presently be performed at the same fast pace: a fact that is slowing down the discovery of proteins with new features as well as their *ex novo* design.[3] These difficulties affect heavily the field of photobiology where the number of known photoreceptor structures is limited. For instance the crystal structure of bovine rhodopsin, the retina dim-light visual photoreceptor, is the only structure available for studying monochromatic and color vision in vertebrates. In principle, the gap could be filled by the construction of sufficiently accurate atomistic computer models of the set of photoreceptors of interest, which would allow *in silico* functional characterization, property screening and *ex-novo* design. However, to be useful, such models should be produced using a fast and standardized (in the sense of a well defined and replicable) protocol with known error bars.

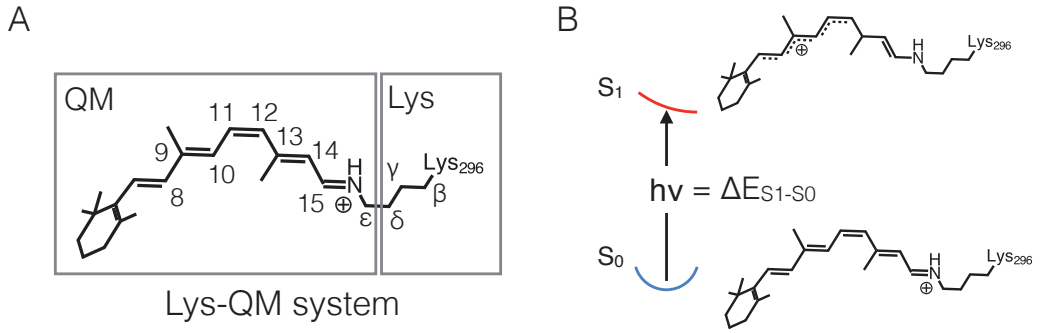
Computer models of biological photoreceptors for studying properties like spectral tuning and photoisomerization timescales are computationally complex. As originally proposed by Warshel[4], these models have to incorporate a quantum mechanical (QM) description of the chromophore (i.e. the protein spectroscopically and photochemically active prosthetic group) and account for the interaction with the apoprotein using a computationally less expensive molecular mechanics (MM) force field. In the past, our group has contributed to show that hybrid quantum-mechanics / molecular-mechanics (QM/MM) approaches combining an *ab initio* multiconfigurational QM description of the light absorbing moiety (i.e. the chromophore) with a MM description of its molecular environment can be applied to different photoresponsive proteins including bovine rhodopsin and other homologous proteins.[5-10] Similar or distinct QM/MM models of such systems have been reported by other groups and mainly used for understanding their absorption spectra.[11-14] Indeed, the rhodopsin protein family represents an impressive case of regulation of the wavelength of the absorption maximum (λ^{a}_{max}) playing a role in fundamental biological processes such as vision, chromatic adaptation, ion-gating and ion-pumping.[15,16] In these trans-membrane proteins, the chromophore λ^{a}_{max} is modulated by the environment from 420 nm (human short wave-sensitive pigment, hSWS) to 587 nm (sensory rhodopsin I).[16] As displayed in Scheme 1A, the chromophore is formed by a retinylidene isomer bound to a lysine residue via an iminium function (also known as a protonated Schiff base, PSB). The detailed knowledge of the molecular factors controlling the λ^{a}_{max} of rhodopsins is important not only for understanding the functions of these photoreceptors, but also for the rational design of novel, genetically encodable tools. In fact, there is a growing interest in

employing rhodopsins to switch "on" and "off" metabolic pathways, gene expression and ion channels.[17-20] Furthermore, rhodopsins are currently central to the field of optogenetics[21] where the gene expressing a light-gated ion-channel channelrhodopsin is used to engineer light-sensitivity into neurons.

As mentioned above QM/MM models of rhodopsins are well established, but a quick literature survey demonstrates that even models of the same protein can be hardly compared. Taking bovine rhodopsin (Rh) as a reference, being the most representative and studied system, at least 6 different QM/MM setups have been used[6,12,13,22-24] (see Supporting Information for more details) resulting in a 36 nm range for the predicted Rh visible absorption maximum wavelength (λ^{a}_{max}). The substantial differences displayed by similar approaches (e.g. Andruniow *et al*.[6] and Tomasello *et al*.[12]), must be due to the distinct manipulations used for the preparation of the initial protein structure (i.e. the structure equilibration, the external environment/solvation, the protonation states of ionizable residues, the internal waters, etc.) and the multiple variables to be selected during the construction of the final QM/MM model (e.g. the choice of the QM method or methods, the treatment of the QM/MM frontier and of QM/MM non-bonding interactions, the employed MM force field). Moreover, these manipulations are prone to operator errors and, in practice, are cumbersome to replicate. Thus, having a well defined QM/MM models which can be readily replicated in other labs and allow for consistent evaluation of the error associated to the computation of specific properties, is highly desirable, if not mandatory, for the correct application and further development of the methodology.

Here we report on the design and testing of an Automatic Rhodopsin Modeling protocol (from now on simply called ARM) addressing the above issue. Such protocol has the primary target of generating congruous QM/MM models of rhodopsins that may facilitate systematic rhodopsin studies.[25,26] More specifically, ARM automates and speeds up the construction of models based on an additive, H-link-atom scheme[27] for both wild type and mutant rhodopsins. ARM is *not* designed to produce the most accurate models possible (see for instance the models of ref. 28-32 targeting accurate spectroscopic studies), but basic, gas-phase and computationally fast models aimed to the *rationalization and prediction of trends between sequence variability and function*. Therefore, ARM models aim to satisfy the following desirable features for a high-throughput approach: (a) automation, to reduce accidental errors and avoid biased modeling; (b) speed, to deal with large sets of protein of the same family; (c) documented accuracy, to translate results into experimentally assessable hypothesis; (d) transferability, to treat rhodopsins with large differences in sequence (i.e. from different domains and kingdoms).

In the present work, a minimal chromophore cavity, a minimal QM subsystem and short molecular dynamics (MD) conformational sampling are considered characteristics of the investigated model, thus making its construction fast and avoiding unbiased construction steps (see features a-b above). On the other hand, a benchmarking set comprising vertebrate, invertebrate and microbial rhodopsins is employed to explore feautures c-d through λ^{a}_{max} computations (i.e. the corresponding vertical excitation process of Scheme 1B). More specifically we use: the all-*trans* and 13-*cis* isomers of *Anabaena* Sensory Rhodopsin (ASR, microbial, sensorial) and 5 of their mutants for which we also measure the λ^{a}_{max};[33,34] bovine Rhodopsin (Rh, vertebrate, visual pigment)[35] and 7 of its mutants for which the measured λ^{a}_{max} is reported in the literature; squid Rhodopsin (SqR, invertebrate, visual pigment);[36] human melanopsin (hMeOp, vertebrate, non-visual pigment);[9,37] the light- and dark-adapted forms of bacteriorhodopsin (bR_{LA} and bR_{DA}, an Archaea proton pump)[38,39]; blue proteorhodopsin (PR, an eubacterial proton pump)[40] and the a chimera channelrhodopsin (ChR_{C1C2}, a lab construct of microbial eukaryotic light-gated ion channels from the green alga *Chlamydomonas reinhardtii*).[31] The selected single-site mutants are S86D, S214D, L83Q, V112N, W76F for ASR and G90S, T94S, T118A, F261Y, W265F, W265Y and A292S for Rh. Moreover, to further increase the scope of our approach, a model of the Rh primary photocycle intermediate bathorhodopsin (bathoRh, experimental λ^{a}_{max} of 543 nm at low temperature[41]) has been built from its crystal structure[42]. In total the set spans 26 observed λ^{a}_{max} values comprised between 473 and 568 nm corresponding to vertical excitation energies (i.e., as illustrated in Scheme 1B, the $\Delta E_{S1\text{-}S0}$ potential energy gap computed at the S_0 equilibrium structure which is also assumed to correspond to the λ^{a}_{max} value. Of course, in doing so, we neglect the effect of the homogeneous and inhomogeneous broadening which we assume to be similar in different rhodopsin species) in the 52 to 61 kcal/mol range. The set has been designed in such a way to include evolutionary distant sensory rhodopsins from different kingdoms (Animalia and Bacteria) and phyla (Chordata, Mollusca and Cyanobacteria), two proton pumps from different kingdoms (Archaea and Bacteria) and one light-gated ion channel from the kingdom Plantae. The “evolutionary distance” between representative members of the set is documented in the Supporting Information in terms of percentage of homology and in Fig. 1A by comparing the structure of the cavities of four sensory rhodopsins.



Scheme 1 A. Structure of the retinal chromophore of bovine rhodopsin and of the Lys-QM system of the ARM model. B. Schematic illustration of the vertical excitation of the chromophore. Notice the different charge distribution in the ground (S_0) and excited (S_1) states conjugated carbon framework.[43,44]

Photobiological studies focusing on both spectroscopy and chemical reactivity require the use of accurate, correlated QM methods.[26] ARM employs the multiconfigurational Complete Active Space Self Consistent Field (CASSCF)[45] method to obtain ground state geometries. Excitation energies are then computed using multiconfigurational second-order perturbation theory (CASPT2)[46] to recover the missing dynamical electron correlation associated with the CASSCF description. Such a CASPT2//CASSCF treatment has been extensively investigated and its limitations are well understood. Here, we stress that, as previously documented, the rather limited ca. 3-4 kcal/mol error[5-8,14,25,47-50] in excitation energy reported in several studies[51-53] and comprising CASPT2//CASSCF/6-31G*/MM calculations are partly due to error cancellations associated to the limited quality of CASSCF/6-31G* equilibrium geometries.[54] Nevertheless we decided to develop ARM on the basis of such QM treatment because CASSCF allows the use of state-averaged energies for mapping thermal and photochemical reaction paths with different electronic characters and for propagating semi-classical excited state trajectories. Therefore CASSCF/6-31G*/MM equilibrium models of the selected rhodopsin benchmark set (from now on called ARM models) are suitable for consistent and comprehensive photochemical studies.

Below we show that ARM represents a first-order answer to the need for high-throughput QM/MM model generation for rhodopsins, yielding blue-shifted $\Delta E_{S1\text{-}S0}$ values with a standard deviation of few kcal/mol and absolute deviations suggesting the presence of a systematic error. Additionally, using the same set, we document the sensitivity and limitations of ARM models with respect to: (i) the uncertainty in the assignment of the ionization state of the protein residues, (ii) the external (i.e. protein surface) counterion placement and distribution, and (iii) the stability of the side-chain conformation affecting the HBN near the chromophore.

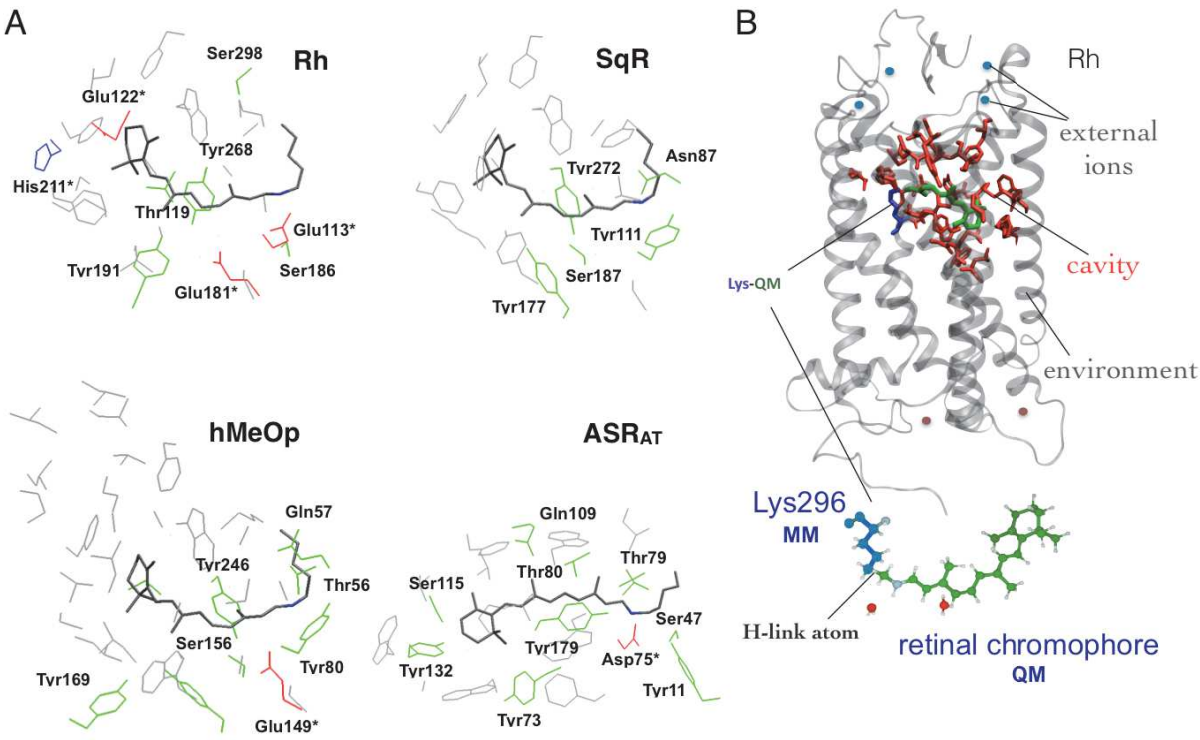


Figure 1 Rhodopsin models. A. Examples of cavity and chromophore variability in the selected benchmark set. The displayed structures are based on x-ray crystallographic data except for hMeOp that has been produced via comparative modelling (see Methods section). Ionizable side-chains are marked with a *: acidic residues are shown in red and alkaline in blue. Non-ionizable polar residues are shown in green. B. General structure of the QM/MM model constructed using ARM displaying the environment (in grey with external counterions in blue and red), cavity (residues in red) and Lys-QM systems (in blue and green).

Methods

Overview of ARM

The QM/MM model constructed by ARM is a gas-phase and globally uncharged monomer model whose structure is illustrated in Fig. 1B. The model is divided in the three systems called environment, cavity and Lys-QM. The environment has backbone and side-chain atoms fixed at the crystallographic (or comparative model) structure, and incorporates suitably placed external counterions also kept fixed at computed positions (see the details below). The cavity, which includes the Lys-QM system detailed in the left part of Scheme 1, also features fixed backbone atoms but its side-chain atoms are free to relax. The Lys-QM system comprises the atoms of the Lys side-chain in contact (through C_δ) with the QM/MM frontier and the entire QM subsystem that corresponds to an N-methylated retinal chromophore. All Lys-QM atoms are free to relax. The environment, cavity and Lys side-chain form the MM subsystem. Notice that while ARM models feature a fixed environment system (see Fig. 1B), other bare QM/MM rhodopsin models reported in the literature are based on geometrically optimized apoprotein structures.[13,14,55,56] In this respect ARM models, are

"conservative" in the sense that they ensure that the information from the x-ray crystallographic structure is retained in the final QM/MM model.

Technically, ARM is a Linux command-line tool that interfaces various publicly available or specifically designed (e.g. the ION module) computational chemistry programs (see the SI for detailed information) carrying out the different tasks required to construct QM/MM models. The only mandatory input includes a Protein Data Bank (PDB) file containing an initial (guess) rhodopsin structure with assigned ionization states, and a list of target residues required to drive the placement of external counterions. However, the original PDB may be pre-processed by completing residues with missing atoms (missing residues are not added) using the specific utility (Automatic PSF Builder) in the VMD program.[57] The sequence of steps from the input to the final results is pre-defined, and either automatic or automatable in a future version of the script (see the *Present automation limits* subsection in the Results and Discussion section).

The currently established overall ARM workflow is depicted in Fig. 2. The workflow starts by reading the PDB file mentioned above. Then a series of automatic manipulations are performed including (sequentially): external counterion placement, hydrogen atom placement, MM energy minimization, multiple MD relaxations and lower-level QM/MM calculations, which are carried out to prepare the input for multiple QM/MM geometry optimizations. The ION module (see below and the SI) adds in specific positions the external Na+ and Cl^- ions required to neutralize the model, by independently neutralizing the inner and outer (i.e. intra-cellular and extra-cellular) protein surfaces. This would better mimic the situation of the protein in a micelle rather than in a biological membrane where one has a different ion concentration in the inner and outer protein surfaces. These two surfaces are defined by a list of target residues provided by the user (see Fig. 2). The list include a sets of (solvent exposed) charged residues located above and below the chromophore, which may potentially harbor one or more counterions. Instead, the exact position of the counterions is determined by an energy minimization procedure. More specifically, for each surface, the list of target residues usually includes all residues whose charges have the same sign as the corresponding surface net charge. An exception is the retinal-bonded lysine and its counterion, which may be excluded in order to prevent the placement of ions inside the protein. Below, we refer to the chosen neutralization of the inner and outer protein surface scheme as the No Surface Charge (NSC) scheme (see also section 2.4 in the SI). Thus, as also discussed below the NSC scheme defines the *distribution* of counterions between the two surfaces while the ION module determines in which exact *position* they have to be placed. Throughout the MM energy minimization and MD calculations, the AMBER94[58] force field is used. Firstly, the hydrogen atoms are added and their positions optimized at the MM level. Consistently with the workflow in Fig. 2, the hydrogen atoms bound to water and polar atoms are added by

DOWSER, a program for hydrogen addition according to energy criteria,[59] while TINKER[60] is used elsewhere. Since water and hydrogen bond networks (HBN) affect side chain conformations and long-range electrostatics (thus ultimately modifying spectral and photochemical properties[6,61-63]), it is important to use DOWSER quantitative criteria for hydrogen and water relocation. After the MM cavity atoms have been automatically selected using CASTp[64] online server (with default settings), a single MM geometry optimization and N independent (i.e. starting with N different seeds which provide N independent sets of initial velocities. See also the SI) MD room-temperature relaxations are performed at the MM level using GROMACS[65] on the cavity and the Lys-QM systems. The chromophore non-bonding and bonding interactions are modeled according to AMBER94 rules: van der Waals parameters for retinal are taken from our custom AMBER94 parameters set[50,66]; partial charges were calculated as AMBER-like RESP charges.[58] Distinct sets of parametrized charges are used for different chromophore isomers. The MD relaxations consist of a 50 ps heating followed by 150 ps room-temperature equilibration and 800 ps production for a total of 1 ns. The MD output structures constitute the guess for N corresponding QM/MM calculations[67] performed according to the scheme described in ref. 50. Briefly, the full N-methyl retinal chromophore (53 atoms) corresponding to the QM subsystem is connected with the designated lysine side-chain of the MM subsystem (taken together these form the aforementioned Lys-QM system). The Lys residue linked to the retinal chromophore is automatically identified on the basis of the distance between the chromophore C15 atom (see Scheme 1) and the Lys residue N atoms in the input structure.

The QM/MM frontier is treated within a link atom approach,[68] whose position is restrained according to the Morokuma scheme,[69] and it is placed across the lysine Cδ-Cε bond (where Cε is a QM atom). The lysine charges are modified by setting the Cδ charge to zero, to avoiding hyperpolarization and redistributing the residual fractional charge on the most electronegative atoms of the lysine, thus ensuring a +1 integer charge of the Lys-QM system. A QM/MM geometry optimization carried out at the HF/3-21G/MM level completes the preparatory phase. In this and subsequent QM/MM computations the AMBER94[58] MM force field is used.

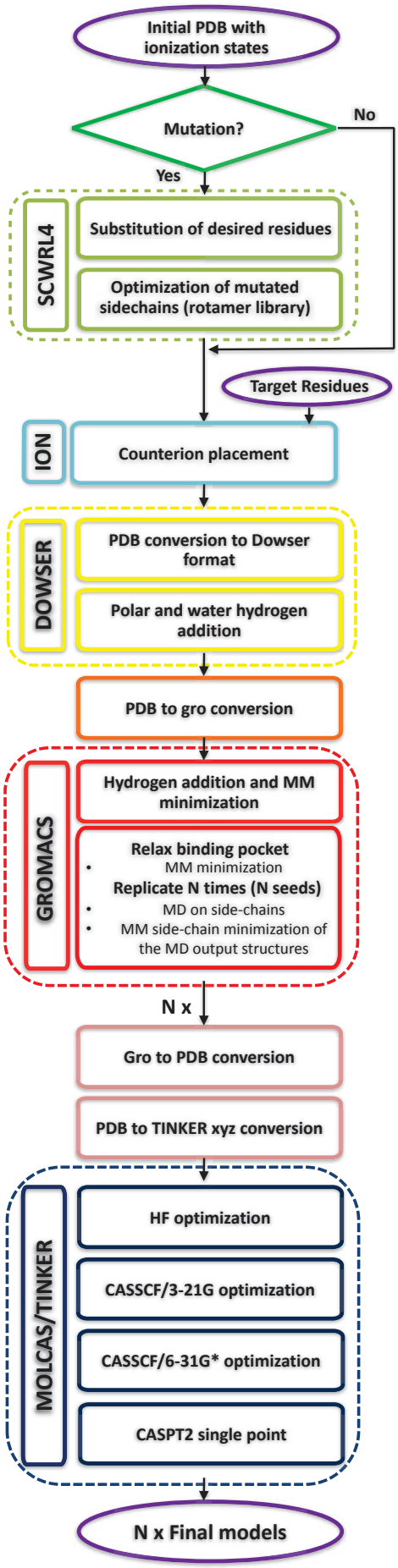


Figure 2 Automatic Rhodopsin Modeling workflow. Vertical boxes refer to programs and their associated actions are circled with dashed lines and colored in light green (SCWRL4), light blue (ION), yellow (DOWSER), red (GROMACS) and blue (MOLCAS/TINKER). Intermediate steps corresponding to file format conversions are shown in orange (PDB to GROMACS type) and pink (GROMACS to PDB, PDB to TINKER XYZ format) boxes.

The obtained N independent HF/3-21G/MM models are processed via two sequential QM/MM optimizations at the single-root CASSCF(12,12)/3-21G/MM level and CASSCF(12,12)/6-31G(d)/MM level to get N final structures. These sequential optimizations have the scope to achieve a more rapid convergence of both the molecular orbital and geometry. Subsequently a CASPT2(12,12)/6-31G(d)/MM 3-single root single point calculation is performed for each model, taking a 3-roots State Average (SA)[70] CASSCF(12,12)/6-31G(d) wavefunction as a reference. Finally, N vertical excitation energy ($\Delta E_{S1\text{-}S0}$) values are computed as the difference between the first two roots. The average $\Delta E_{S1\text{-}S0}$ value is then compared with the corresponding experimentally observed value (i.e. we assume that this value corresponds to the energy carried by a photon with the observed λ^{a}_{max}). Oscillator strengths (f) are computed at the SA-CASSCF level to check if $\Delta E_{S1\text{-}S0}$ corresponds to a spectroscopically allowed transition. All QM/MM calculations are performed by the distributed MOLCAS/TINKER interface[67], as implemented in MOLCAS version 7.4 or higher (ESPF module),[71] interfaced with TINKER version 4.2 or higher.[60]

ARM can optionally perform amino acid substitutions on the starting PDB structure to generate ARM models of rhodopsin mutants. The list of desired mutations must be provided as an additional input file (see the SI for further details). Mutations are carried out by SCWRL4,[72] a

program for predicting side-chain conformations from a given protein backbone, using a backbone-dependent rotamer library. Once the mutations have been made, the ARM workflow proceeds exactly as described above.

While the results presented below are all consistently produced with *the same* standardized ARM workflow using pre-assigned *default* parameter values (see points a-f below), the following customization is possible: (a) DOWSER can retain the crystallographic water molecules in the starting guess structure (default) or neglect them. In the latter case, it generates and places water molecules in all cavities of suitable sizes, (b) the default number (N=10) of independent MD relaxations as well as the default heating (50 ps), room temperature (298 K) pre-equilibration (150 ps) and production (800 ps) stages may be changed, (c) the default chromophore cavity is obtained automatically using the CASTp[64] online server which identifies cavities and pockets in proteins by geometrical methods (we used a probe radius set at 1.4 Å). This cavity may be replaced by cavities constructed automatically specifying the maximum distance of any atom of the protein residues from any atom of the chromophore using the program VMD[57]. Alternatively, a list of residues whose side-chains will be relaxed can be provided using an additional input file. (d) The chromophore can be either allowed to relax (default) or kept frozen during the MD relaxations and (e) the final MD structures can be obtained by choosing either the last snapshot of the run, or the snapshot featuring the most similar structure to the unphysical average structure obtained from the MD production stage (nearest-to-average structure). Such similarity is evaluated as the Root Mean Square Deviation (RMSD) value of the snapshot against the average structure, so that the most similar snapshot has the lowest RMSD. The nearest-to-average structure is the default. (f) The QM/MM calculations are not customizable, except for the use of microiterations[73] to improve the convergence: during each geometry optimization step the cavity side-chains are relaxed after the optimization of the Lys-QM system by default, unless they are disabled by the user.

The above procedure is automatic once the requested input is provided by the user, who has to type values and choices and run scripts following on-screen instructions. TINKER 5.1,[60] GROMACS 4.5.4[65] and MOLCAS 7.8[67] were used throughout the testing. The final outcome is a set of N TINKER XYZ files containing the protein geometries, which can be converted to PDB by running a provided script. For each TINKER file, two files with information such as the CASPT2 absolute energies in a.u. of the computed electronic states, the associated vertical excitation energies expressed in both kcal/mol and nm and the oscillator strengths for the same transitions, are produced.

Input data for the benchmark set.

As mentioned above the construction of an ARM model requires an initial geometry and the relative

residue protonation states. The geometry is provided by PDB crystallographic structures for Rh (1U19[35]), bathoRh (2G87[42]), ASR (1XIO[34]), SqR (2Z73[36]), bR_{LA} (1C3W[38]), bR_{DA}(1X0S[39]), PR (4JQ6[40]) and C1C2 (3UG9[74]), while for hMeOp the comparative model of ref. 9, is employed. The residue protonation states are instead determined, in all cases, using the program PROPKA version 3.0[75] for all members of the benchmark set including the ASR_{AT}, ASR_{13C} and Rh mutants. The target residues required by the ION module for placing the chloride and sodium counterions on the two rhodopsin surfaces (as requested by the NSC scheme. See also the description above) are unambiguously selected according to energy criteria.

The list of residues defining the retinal chromophore cavity, the position of the water oxygens and the set up of the MD relaxation, are considered pre-assigned parameters. The positions of the crystallographic water oxygens were used as a starting guess for the DOWSER-3[59] calculation assigning the initial orientation of all water hydrogen atoms (in the initial hMeOp geometry obtained via comparative modeling we included the water oxygens present in its SqR template for consistency).

Sensitivity of ARM models to the protonation states.

As already mentioned above the assignment of the protonation states of the ionizable residues is part of the ARM input. When a different ionization state is assigned to a cavity (or near cavity) side-chain, significant changes in the $\Delta E_{S1\text{-}S0}$ values are expected. In the Result and Discussion section we investigate this effect in two cases: the E181 side-chain of Rh (Fig. 1A), whose ionization state is still debated;[76-78] and the D85 and D212 residues in bR, whose complex hydrogen bond network allows for different combinations of protonation states.[79] In the case of Rh, E181 is predicted to be neutral by PROPKA, which gives an estimated pKa>7.6, corresponding to >80% neutral molecules in a sample at pH 7. Therefore we constructed a model with an ionized E181 side-chain (labeled Rh_{181}) by using the same initial geometry as for the reference model (Rh) so that the effect of the charged E181 could be singled out. For bR we used the PROPKA prediction giving a neutral D85 and a charged D212. Nevertheless, PROPKA labels them as "correlated" residues, i.e. the choice of ionization state on one residue determines the other, mostly because they are only about 6 Å apart. Also, D85 and D212 are connected via a water molecule, so protons can, in principle, shuffle between them. However, since D212 makes hydrogen bonds with Y57 and Y185, a protonated state seems unlikely. On the other hand, D85 can accommodate a proton but the presence of many polar residues and water molecules would stabilize a charged form too. For this reason we constructed an additional model with a charged D85 side chain (bR_{85}, see the subsection *Alternative assignments of ionization states: Rh and bR* below), while keeping D212 charged as in the reference ARM model.

Sensitivity of ARM models to the neutralization scheme.

The sensitivity of the ARM models to the neutralization scheme has also been assessed because the standard NSC neutralization scheme does not necessarily reflect the physiological conditions (i.e. in general, membrane proteins are exposed to trans-membrane electrostatic fields in the range of few tens of mV originating from an asymmetrical distribution of the surface ions). As we report below (see the *Assessing the protein environment (counterion distribution and position) approximation* subsection below), we look at the $\Delta E_{S1\text{-}S0}$ variations obtained when, instead of neutral inner and outer surfaces, external Cl^- and Na^+ counterions are placed in such a way that oppositely charged surfaces are obtained, while still achieving an overall neutral model. In the latter case, an electrostatic field across the chromophore is generated. This effect is tested for Rh, Rh_{181} and the bacterial rhodopsin isomer ASR_{AT}. Below we also documented the $\Delta E_{S1\text{-}S0}$ variations between a manual (i.e. biased) external counterion placement scheme (see the SI for details) and the automatic placement obtained with the ION module for Rh, ASR, SqR, hMeOp, bR, PR, bathoRh.

Sensitivity of ARM models to the residue side-chain conformation.

Residue side chain conformations affecting the HBN in the vicinity of the chromophore could significantly affect the computed $\Delta E_{S1\text{-}S0}$ value in rhodopsins. Thus the effect of the flexibility of the side-chain conformation and HBN stability has been assessed (see the subsection *Stability of the computed side-chain conformations and hydrogen bond network* below) as a function of: (i) the number N of replicated 1 ns MD relaxations for Rh and ASR and (ii) the length (up to 30 ns) of a single MD relaxation for Rh and Rh_{181}. Such significantly longer MD relaxation is used to search for instabilities or alternate conformations that are unreachable during single 1 ns relaxation employed when building ARM models and to estimate their effects on computed $\Delta E_{S1\text{-}S0}$ values.

Preparation and λ^a_{max} measurements of ASR mutants.

In order to expand the benchmarking of mutants beyond the set of Rh mutants found in the literature, we have prepared 5 ASR mutants. The corresponding λ^a_{max} values have been determined via HPLC analysis and spectral measurements for both the ASR_{AT} and ASR_{13C} isomers providing 10 experimentally observed values to be used in the benchmark study. Full-length ASRs having six histidine residues at the C terminus were expressed in E. coli. BL21, C41(DE3) or UT5600 strains. The protein was solubilized in n-dodecyl-β-D-maltoside (DDM), and purified via Ni^{2+}-affinity column chromatography in manner previously described.[33,80,81] The λ^a_{max} of ASR_{AT} and ASR_{13C} forms were determined from the UV-vis spectroscopy and the HPLC analysis on dark- and light-adapted sample

at pH 7, according to the method described previously.[82] For these measurements, DDM solubilized sample were used. The detailed results are reported in the SI.

Present automation limits.

The automation of ARM is presently not complete. First of all the assignment of the amino acid ionization states, including the generation of models of the same rhodopsin with different possible ionizations and ionization equilibria (e.g. a possible equilibrium between Rh and Rh_{181}), are manually handled after inspection of the PROPKA output or, when available, on the basis of experimental data. Secondly, as described above, the target residues of the ION module are manually provided for the outer and inner protein surfaces even if such a list could be automatically generated on the basis of the ionization state information and the geometrical parameters of the residue. Thirdly, the list of the CASTp cavity residues is presently generated using the CASTp[64] online server whose output is pasted in a command line. While the previous limitations require only suitable coding to be removed a fourth limitation requires more serious efforts to be overcome. This is related to the selection of the CASSCF active space. In all benchmark models illustrated here the HF/3-21G optimization (see Fig. 1) yields the 12 π molecular orbitals in the HOMO-5 to LUMO+5 position. While this automatically leads to a CASSCF correct active space selection (i.e. no orbital permutation is necessary and the HF molecular orbitals can be directly used as molecular orbital guess for the successive CASSCF calculation), such favorable ordering must be considered an exception. In other words, in different rhodopsin or mutant models, the HOMO-5 to LUMO+5 positions may contain orbitals not belonging to the skeletal π-system. An automatic way of detecting this situation, based on CASSCF orbital occupations, is implemented in ARM. A wrong orbital is likely to have either 2 or 0 (larger than 1.999 or lower than 0.001) orbital occupation rather than a value in between. When this happens, it is necessary to visually inspect the orbitals and permute their order before the CASSCF calculation is restarted, thus making the workflow not automatic. However, a research line on the automatic selection of the CASSCF active space is presently carried on[83], making the possibility of a robust automation of ARM realistic.

Results and Discussion

The average (N=10) $\Delta E_{S1\text{-}S0}$ values calculated with ARM using the default parameters defined above are reported in Table 1 (wild-type rhodopsins) and Table 2 (mutants) and plotted as full circles in Fig. 3A and 3B where they are compared to the experimental values. Since the observed λ^{a}_{max} is not well defined for hMeOp, the 473 nm value (i.e. a $\Delta E_{S1\text{-}S0}$ value of 60.4 kcal/mol) corresponding to the average of the observed range (467-480 nm)[9,84,85] was taken as a reference. All input structures (i.e.

crystallographic structures and the hMeOp comparative model with assigned ionization states and external counterions) are provided as SI. As part of the SI we also provide one selected output structure for each wild-type rhodopsin and rhodopsin mutant, which corresponds to the model with the $\Delta E_{S1\text{-}S0}$ closest to the average $\Delta E_{S1\text{-}S0}$.

The overall standard deviation of the average $\Delta E_{S1\text{-}S0}$ values with respect to observed values (connected by a full line in Fig. 3) and when excluding the Rh, $bR_{DA(85)}$ and $bR_{LA(85)}$ models is 3.6 kcal/mol with the absolute deviations (the difference between the calculated and observed $\Delta E_{S1\text{-}S0}$ values) ranging from -0.2 to +4.1 kcal/mol. When considering the wild-type proteins only, the standard deviation increases to 3.8 kcal/mol with the largest value found for the Rh_{181} (+2.8 kcal/mol). The trend, which parallels the observed values, and blue-shifted deviation of the computed values indicate the presence of a systematic error of few kcal/mol whose exact origin is difficult to characterize. Notice that the MS-CASPT2 level of theory yield the same trend (see Table S6 in the SI) as the CASPT2 level even if the excitation energies appear further blue-shifted. Such behaviour is attributed to an oversplitting of the electronic energies in the presence of state mixing (i.e. two large off-diagonal elements in the MS-CASPT2 hamiltonian especially when using small atomic basis) already described in the literature.[86,87]

As it will be discussed later, using the average of ten $\Delta E_{S1\text{-}S0}$ values generated from uncorrelated MD relaxations rather than using a single $\Delta E_{S1\text{-}S0}$ value, decreases the $\Delta E_{S1\text{-}S0}$ deviations. Indeed, the comparison of full (N=10) and open (N=1) circles in Fig. 3A shows a general improvement in the first set with respect to the second. One extreme case is that of PR which pass from a negative deviation (-0.6) to a positively deviation (+2.5) when considering the $\Delta E_{S1\text{-}S0}$ average value. A similar but less dramatic effect is seen for ASR_{13C}. Furthermore, in the case of bovine rhodopsin, the N=10 averaging moves the Rh_{181} model in better agreement with the general trend with respect to the Rh model. This is why in this work the Rh_{181} model is selected as the reference model of bovine rhodopsin. While we cannot presently exclude the presence of an equilibrium between Rh and Rh_{181} or that our results are a consequence of the specific ARM model setup, a further investigation of these issues is beyond the scope of the present work mainly focusing on the accuracy of computed $\Delta E_{S1\text{-}S0}$ trends and not specific $\Delta E_{S1\text{-}S0}$ absolute values.

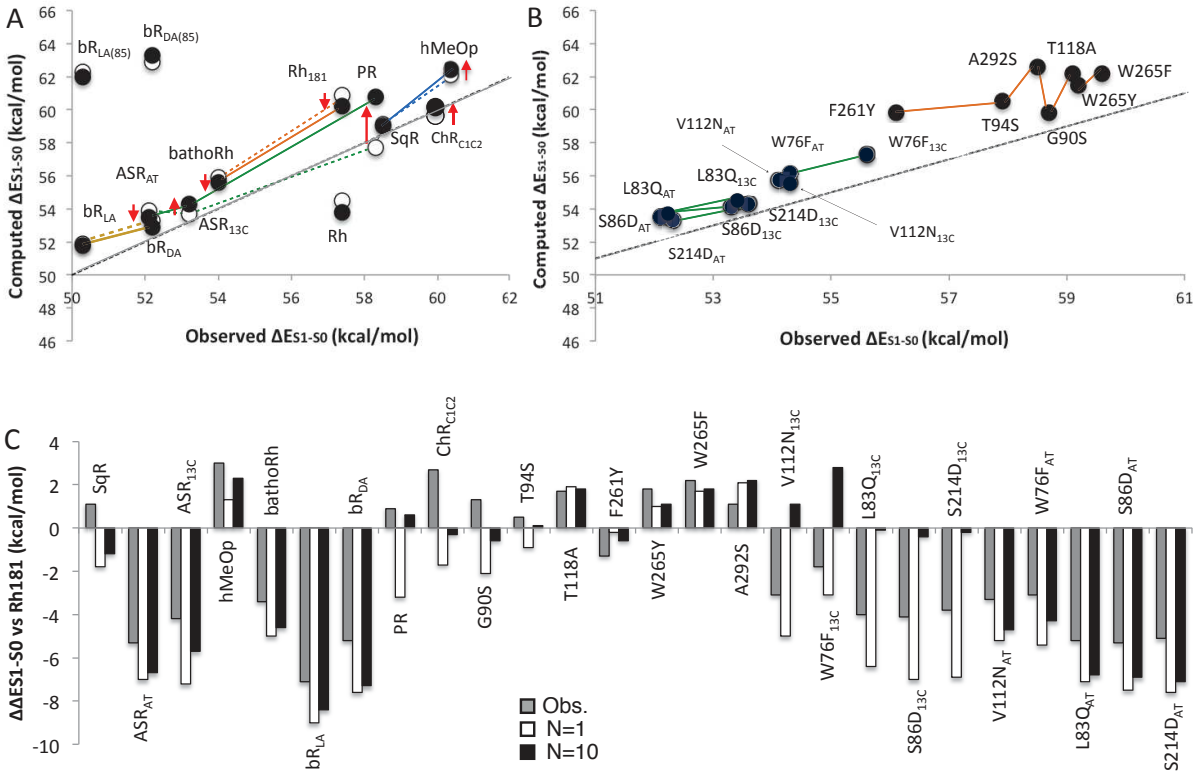


Figure 3 Benchmark results. (A) Observed vs. computed values for vertical excitation energies of wild-type rhodopsins. The diagram includes the results of sensitivity tests for input parameters, such as ionization state (for Rh_{181}, bR_{LA} and bR_{DA}), HBN variations and side-chain conformational changes. The effects of HBN variations and conformational changes are assessed by comparison between average and single $\Delta E_{S1\text{-}S0}$ values (see text). Full colored lines connect the average $\Delta E_{S1\text{-}S0}$ values obtained from N=10 ARM models generated using independent MD runs. The dashed colored lines connect the single $\Delta E_{S1\text{-}S0}$ values obtained with a single N=1 ARM model. The colors refer to vertebrate rhodopsin (orange), invertebrate and non-visual rhodopsins (blue), bacterial rhodopsins (green) and archaea rhodopsin (brownish). Channelrhodopsin ChR_{C1C2} is the only rhodopsin originating from microbial eukaryotic rhodopsins. The red arrows show the $\Delta E_{S1\text{-}S0}$ changes when passing from single (open circles) to average (full circles) values. (B) Comparison between computed and observed average $\Delta E_{S1\text{-}S0}$ values (full circles) from N=10 ARM models of a set of ASR_{AT}, ASR_{13C} (ten values on the left part of the diagram) and Rh_{181} (seven values on the right part of the diagram) mutants. (C) Relative observed and computed $\Delta E_{S1\text{-}S0}$ changes ($\Delta\Delta E_{S1\text{-}S0}$) with respect to Rh_{181}. Grey bars are observed values, white bars correspond to $\Delta\Delta E_{S1\text{-}S0}$ values computed using single $\Delta E_{S1\text{-}S0}$ values while black bars correspond to $\Delta\Delta E_{S1\text{-}S0}$ values computed using average $\Delta E_{S1\text{-}S0}$ values.

Table 1. Comparison between computed and observed vertical excitation energies $\Delta E_{S1\text{-}S0}$ in kcal/mol (maximum absorption wavelengths λ^{a}_{max} in nm) for our benchmark set of 13 wild-type ARM models computing single and multiple calculations. Computed oscillator strengths f_{Osc} are also shown. The corresponding MS-CASPT2 values of the relevant models are given in the SI.

	Obs. ΔE (λ^{a}_{max})	Calc. ΔE (λ^{a}_{max}) Single	Calc. ΔE (λ^{a}_{max}) Average (N=10)	Error Single	Error Average (N=10)	f_{Osc} Average (N=10)
Rh	57.4 (498)[81]	54.5 (524)	53.8 (531)	-2.9	-3.6	1.08
Rh_{181}	57.4 (498)[81]	60.9 (469)	60.2 (474)	3.5	2.8	0.99

bathoRh	54.0 (529)[82]	55.9 (511)	55.6 (513)	1.9	1.6	1.04
SqR	58.5 (489)[83]	59.1 (484)	59.0 (484)	0.6	0.5	0.81
hMeOp	60.4 (473)[a]	62.2 (459)	62.5 (457)	1.8	2.1	0.78
ASR_{AT}	52.1 (549)[34]	53.9 (529)	53.5 (534)	1.8	1.4	1.14
ASR_{13C}	53.2 (537)[34]	53.7 (532)	54.5 (526)	0.5	1.3	1.04
bR_{LA}	50.3 (568)[84]	51.9 (550)	51.8 (551)	1.6	1.5	1.30
$bR_{LA(85)}$	50.3 (568)[84]	62.3 (458)	62.0 (461)	12.0	11.7	0.6
bR_{DA}	52.2 (548)[84]	53.3 (535)	52.9 (540)	1.1	0.7	0.90
$bR_{DA(85)}$	52.2 (548)[84]	62.9 (454)	63.3 (451)	10.7	11.1	0.74
PR	58.3 (490)[85]	57.7 (495)	60.8 (470)	-0.6	2.5	0.88
ChR_{C1C2}	60.1 (476)[b]	59.2 (483)	59.9 (477)	-0.9	-0.2	0.95

[a] Average of available values from refs. 84, 85

[b] Available value is provided in ref. 74

Table 2. Comparison between computed and observed vertical excitation energies $\Delta E_{S1\text{-}S0}$ in kcal/mol (maximum absorption wavelengths λ^{a}_{max} in nm) for our benchmark set of ARM models for 17 ASR_{13C}, ASR_{AT} and Rh_{181} mutants, computing single and multiple calculations. Computed oscillator strengths f_{Osc} are also shown.

	Obs. ΔE (λ^{a}_{max})	Calc. ΔE (λ^{a}_{max}) Single	Calc. ΔE (λ^{a}_{max}) Average (N=10)	Error Single	Error Average (N=10)	f_{Osc} Average (N=10)
G90S	58.7 (487)[86]	58.8 (485)	59.8 (478)	0.1	1.1	0.95
T94S	57.9 (494)[87]	60.0 (475)	60.5 (472)	2.1	2.6	0.86
T118A	59.1 (484)[86]	62.8 (454)	62.2 (459)	3.7	3.1	0.78
F261Y	56.1 (510)[88]	60.7 (470)	59.8 (477)	4.6	3.7	0.91
W265F	59.6 (480)[89]	62.6 (456)	62.2 (460)	3.0	2.6	0.89
W265Y	59.2 (483)[89]	61.9 (461)	61.5 (464)	2.7	2.3	0.87
A292S	58.5 (489)[86]	63.0 (453)	62.6 (457)	4.5	4.1	0.83
$V112N_{13C}$	54.3 (526) [a]	54.5 (524)	55.6 (514)	0.2	1.3	1.07
$W76F_{13C}$	55.6 (514) [a]	57.0 (501)	57.3 (498)	1.4	1.7	0.96
$L83Q_{13C}$	53.4 (535) [a]	54.1 (528)	54.4 (524)	0.7	1.0	0.86
$S86D_{13C}$	53.3 (536) [a]	54.0 (529)	54.1 (528)	0.7	0.8	1.01
$S214D_{13C}$	53.3 (536) [a]	54.1 (528)	54.3 (526)	0.5	0.7	0.99
$V112N_{AT}$	54.1 (529) [a]	55.7 (512)	55.7 (512)	1.6	1.6	1.18
$W76F_{AT}$	54.3 (527) [a]	56.6 (504)	56.1 (509)	2.3	1.8	0.98
$L83Q_{AT}$	52.2 (548) [a]	53.2 (536)	53.6 (533)	1.0	1.4	1.17
$S86D_{AT}$	52.1 (549) [a]	54.0 (529)	53.5 (533)	1.9	1.4	1.15
$S214D_{AT}$	52.0 (550) [a]	53.6 (532)	53.3 (536)	1.3	1.0	1.16

[a] Present work. See the SI for details.

In principle, the set of mutants would be more challenging with respect to wild-type models due to the smaller changes in the observed $\Delta E_{S1\text{-}S0}$ values and the lack of crystallographic structures. Indeed, the deviations of the computed average $\Delta E_{S1\text{-}S0}$ values for ASR and Rh_{181} mutant set are 0.7-4.1 kcal/mol and therefore higher than those of the -0.2-2.8 kcal/mol of the wild-type set. In fact, upon replacement of a side chain, the local steric hindrance and polarity are perturbed upon substitution

and this induced a re-organization of the nearby residues that may not be accurately described by the SCWRL4 module and limited CASTp cavity relaxation. On the other hand, as shown in Fig. 3C, the sign of the experimentally observed $\Delta\Delta E_{S1\text{-}S0}$ are always qualitatively reproduced, with the only exception of G90S, when using $\Delta\Delta E_{S1\text{-}S0}$ computed on the basis of average $\Delta E_{S1\text{-}S0}$ values. With single $\Delta E_{S1\text{-}S0}$ values both G90S and T94S Rh_{181} mutants do not provide the correct shift direction further indicating the importance of the sampling on the HBN and side-chain conformations. However, these two "less accurate" mutant models are different from the other models since their CASTp cavity does not incorporate the mutated side-chains. Accordingly, these side-chains are kept frozen during the MD and QM/MM optimization runs and not relaxed as for the mutated side-chains of the other members of the set. However, inclusion of the 90 and 94 side chains in the chromophore cavities (which therefore does not correspond to the default CASTp cavity) does not change the trend when compared with a wild-type structure with the same expanded CASTp cavity.

The ability of ARM to reproduce/predict the observed $\Delta E_{S1\text{-}S0}$ values may appear surprising when considering the relatively small group of 27 values (i.e. excluding the Rh, $bR_{LA(85)}$ and $bR_{DA(85)}$ discussed below), the general uncertainty associated with the generation of comparative (i.e. hMeOp) and mutant models and, of course, the basic models generated by ARM. Furthermore, we expect the computed values to be affected by systematic error cancellations. An error of this type has been already mentioned above and occurs when the multiconfigurational CASPT2//CASSCF/6-31G* strategy is used and it is exclusively limited to the QM part of the computation.[87] However, what is most important here is the ability to reproduce the observed $\Delta E_{S1\text{-}S0}$ trends (i.e. the sign and magnitude of the $\Delta\Delta E_{S1\text{-}S0}$ values). Such ability appears to be satisfactory (see Fig. 3C) especially when comparing rhodopsin originating in similar organisms and therefore with higher homology (see orange, blue, green and brownish full lines in Fig. 3A and 3B).

Assessing the protein environment (counterion distribution and position) approximation.

While rhodopsins are trans-membrane proteins, ARM models do not incorporate the environment provided by the complex hydrated membrane bilayer and/or account for the existence of protein dimers or trimers. The effect of the protein environment is indirectly incorporated in the model at the input level by using the experimentally derived structure to build the fixed part of the protein body (see Fig. 1B), by assuming a total charge of zero and, according to our NSC reference neutralization scheme (see Method section), assuming that the solvent molecule distribution on both the inner and outer surfaces screens the surface charged residues and counterions. For instance, since Rh has a +4 total charge after the assignment of the residue protonation states (i.e. at pH=7), six Cl^- are distributed on the intracellular surface, where there are 16 positively and 10

negatively charged residues, and two Na^+ on the extracellular side, where there are 6 positively and 8 negatively charged residues (see Rh structure in Fig. 4A). This procedure results in a neutral rhodopsin model with neutral (0 charge) inner and outer surfaces.

The NSC scheme defines how the counterions shall be distributed between the two protein surfaces. However, the exact position of each counterion is decided automatically by the ION module, which employs energy criteria (see the SI for details). The sensitivity to the counterion position is assessed in Fig. 4B, where the single (N=1) $\Delta E_{S1\text{-}S0}$ values computed using ION are compared to the $\Delta E_{S1\text{-}S0}$ values computed by placing the counterions manually according to operator defined rules (i.e. as detailed in the SI, Cl^- or Na^+ are placed at qualitatively predefined distances and orientations close to the positive and negative residues respectively). The comparison shows an improvement in the trend of computed $\Delta E_{S1\text{-}S0}$ values for the set of wild-type rhodopsins of Fig. 3A when using the automatic placement, with the exception of the red-shifted Rh and PR values, whose deviations are discussed in the following sections dealing with ionization states and thermal sampling.

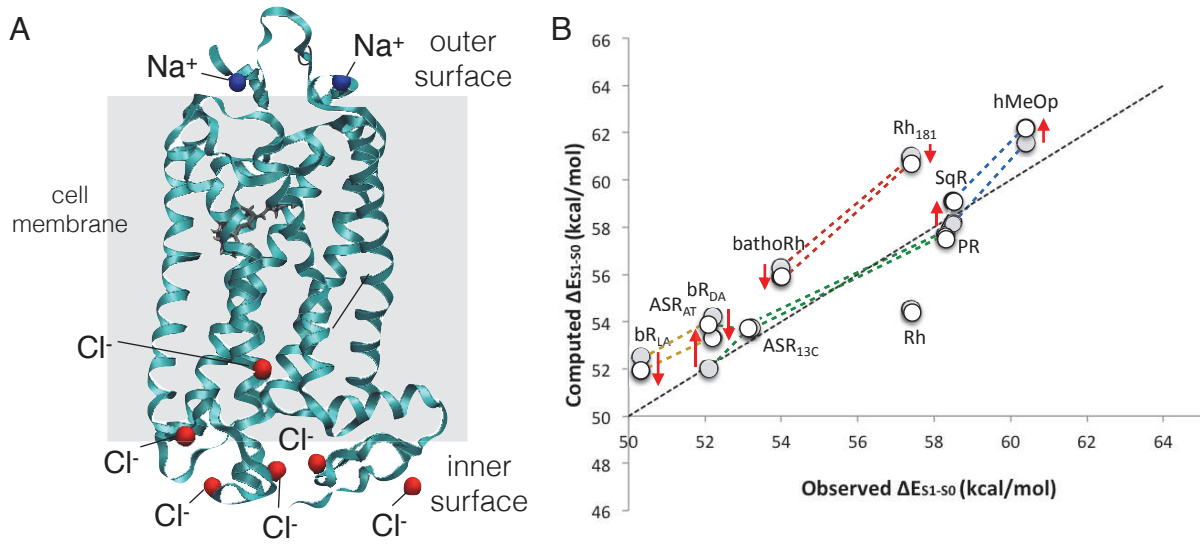


Figure 4 Features affecting the vertical excitation energy in ARM rhodopsin models. (A) ION placement of the Na^+ and Cl^- counterions in Rh. (B) Differences between the $\Delta E_{S1\text{-}S0}$ values obtained via a manual placement (see the SI) and the ION placement. The red arrows show the $\Delta E_{S1\text{-}S0}$ changes when passing from a manual (dashed circles) to an automatic (open circles) placement.

According to the NSC scheme the counterions are distributed in such a way to yield uncharged surfaces. However, in all rhodopsins the sum of the ionizable residue charges is more positive in the inner surface and more negative (less positive) in the outer surface (due to the physiological excess of extra-cellular positive ions of cell membranes compared to their intra-cellular one). This unbalanced situation creates an electrostatic field (a voltage) across the membrane, which ultimately results from differently charged inner and outer protein surfaces. For this reason, and without targeting a simulation of the physiological state, we test the sensitivity of ARM models with respect to an increase (decrease) of one unit of surface charge on the ASR_{AT}, Rh and Rh_{181} models

(see Fig. 5A for Rh). From inspection of the results displayed in Fig. 5B, it is apparent that the models have opposite sensitivities to electrostatic fields approximately aligned along the protein axis. Rh and Rh_{181} display a $\Delta E_{S1\text{-}S0}$ increase when increasing the charge on the outer (extracellular) surface and decreasing the inner (intracellular) surface. Instead, ASR_{AT} displays the opposite trend. In both cases the $\Delta E_{S1\text{-}S0}$ changes are limited to ca. 2 kcal/mol for unit charge.

The opposite sensitivities displayed by ASR_{AT} and Rh (or Rh_{181}) call for a molecular-level explanation. As previously reported,[88] the effects of the surface charges on the $\Delta E_{S1\text{-}S0}$ value can be understood using the mechanistic model illustrated in Scheme 1B. An electrostatic field stabilizing the positive charge in its S_0 location near the protonated Schiff base will increase the $\Delta E_{S1\text{-}S0}$ value, while a destabilizing field would decrease $\Delta E_{S1\text{-}S0}$. In contrast, an electrostatic field stabilizing the positive charge in its S_1 location near the β-ionone ring will decrease the $\Delta E_{S1\text{-}S0}$ value, while a destabilizing field would increase $\Delta E_{S1\text{-}S0}$. As shown in Fig. 5C left, the chromophore in Rh and Rh_{181} is placed far from the two protein sides and oriented in such a way to expose its Schiff base to the negative surface and its β-ionone region to the more positive surface. The same arrangement is found for SqR, hMeOp and the bathoRh intermediate even if bathoRh features a distorted all-*trans* chromophore. In these situations an extra negative charge placed on the outer surface together with a positive charge placed on the inner surface would stabilize the S_0 and destabilize the S_1 state, thus resulting into a $\Delta E_{S1\text{-}S0}$ value increase. Therefore the computed ca. 2 kcal/mol increase of the $\Delta E_{S1\text{-}S0}$ values is consistent with the creation of an electrostatic field by the counterion placement scheme (see open circles in Fig. 3B) with respect to the reference NSC placement.

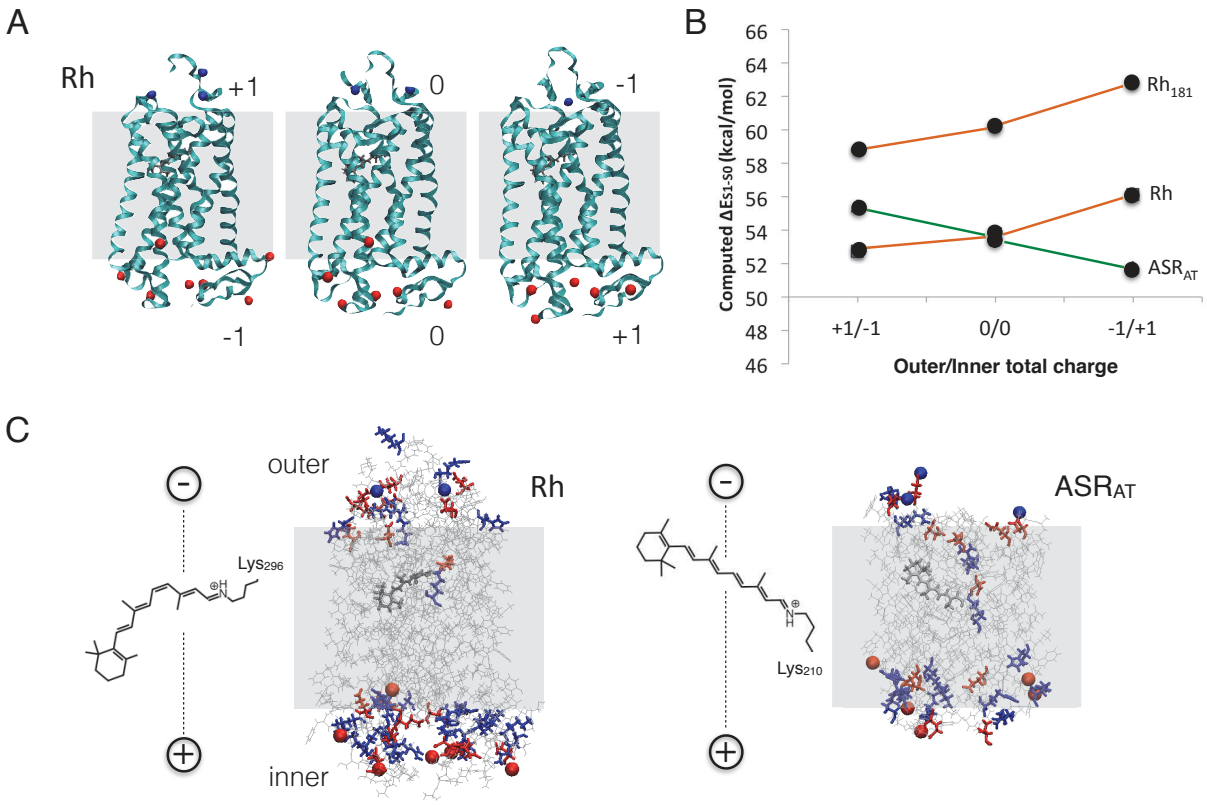


Figure 5 Features affecting the vertical excitation energy in the Rh model. (A) Different external counterion distribution. Red spheres and blue spheres correspond to Cl^- and Na^+ ions respectively. The Rh models correspond to the standard ARM model using the NSC scheme (center) and to two alternative model where the surface total charge is increased or decreased of a one unit. (B) Variations in average $\Delta E_{S1\text{-}S0}$ values as a function of excess surface charges. (C) Orientation of the electrostatic field imposed by the reference external counterion distribution in Rh (left) and ASR_{AT} (right). Positively charged side-chains are given in blue while negatively charged side-chains are given in red. The charge symbols at the top and bottom of the schematic 11-*cis* and all-*trans* chromophores indicate the direction of the protein field in the case of the reference counterion distribution and not the position of specific point charges.

The same reasoning explains the opposite effect of ASR_{AT}. As shown in Fig. 5C in ASR_{AT} the chromophore is placed closer to the negative protein surface and oriented in such a way to expose its β-ionone region to the negative charge. The consequence would be a decrease of the $\Delta E_{S1\text{-}S0}$ value (stabilization of the S_1 state with respect to the S_0 state), which explains why the value of $\Delta E_{S1\text{-}S0}$ for both ASR forms decreases upon creation of an electrostatic field by placing an extra negative counterion in the outer surface and a positive counterion in the inner surface. These qualitative explanations of the behaviour reported in Fig. 5B on the basis of the charge translocation effect (see Scheme 1) and chromophore orientation (see Fig. 5C) are supported by electrostatic potential calculations whose results are given in Section 10 of the SI.

Alternative assignments of ionization states: Rh and bR

As reported in Table 1 and Fig. 3A the effect of the ionization on the average $\Delta E_{S1\text{-}S0}$ value for rhodopsin is large. Such value increases of ca. 6 kcal/mol leading to a strong λ^a_{max} blue-shift when

going from the Rh model featuring a protonated E181 residue, to the deprotonated Rh_{181} model featuring a negatively charge E181 residue. The Rh_{181} model yields a blue-shifted $\Delta E_{S1\text{-}S0}$ value with respect to the experimental data and, in contrast to the Rh model, appears in line with the type of positive deviation observed for the other ARM models.

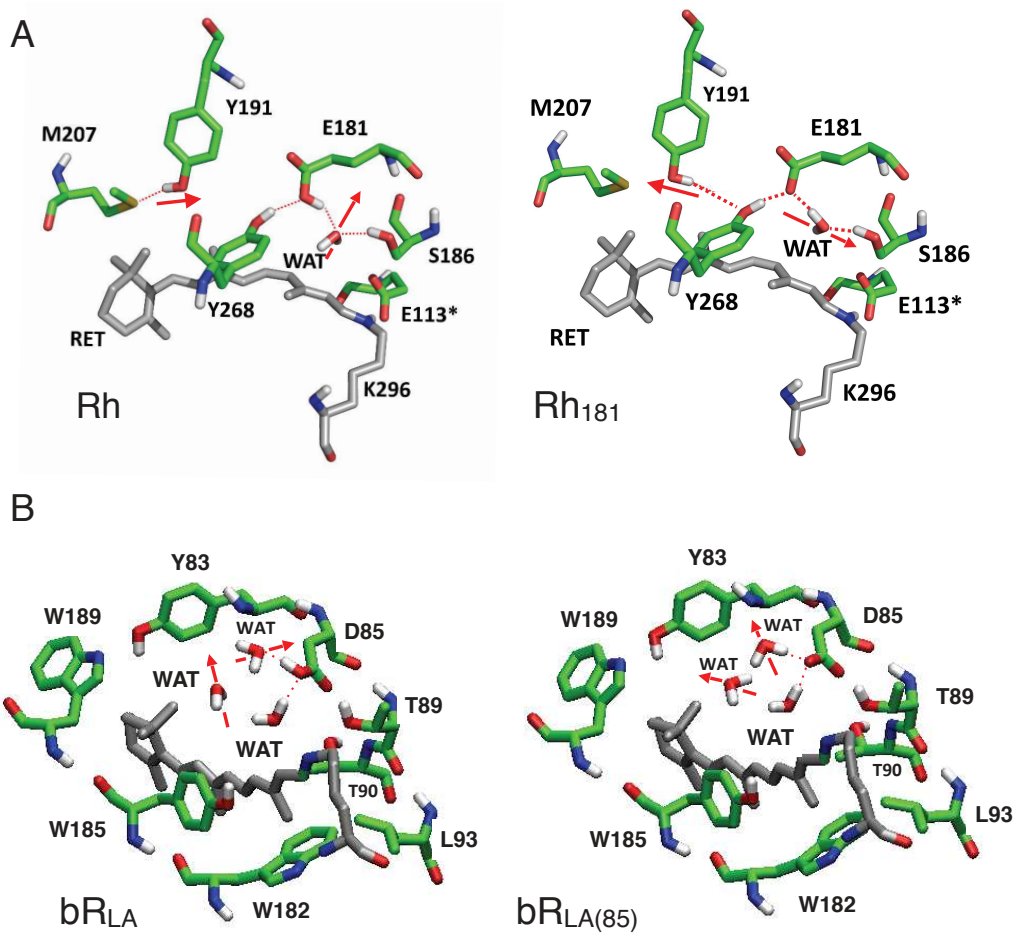


Figure 6 Features affecting the vertical excitation energy in vertebrate and archaea rhodopsin ARM models. (A) Ionization state effects of E181. The two structures illustrate the most frequent HBN and water orientation in the Rh (E181 protonated) and Rh_{181} (E181 deprotonated) models. (B) Ionization state effects of D85. The two structures illustrate the most frequent HBN and water orientation in bR_{LA} (D85 protonated) and $bR_{LA(85)}$ (D85 deprotonated) models. The red arrows indicate the changes in dipole direction.

As revealed by a comparative analysis of the chromophore cavities in Rh_{181} and Rh models, the computed $\Delta E_{S1\text{-}S0}$ change represents the response to the large variation of electrostatics and induced HBN changes in the chromophore surroundings. More specifically, as shown in Fig. 6A, Rh_{181} features a HBN where a water molecule (WAT) has, with respect to Rh (see Fig. 6A), an O-H group flipped and, rather obviously, forms a strong hydrogen bond with one of the E181 carboxylate oxygen. As a consequence the new WAT orientation modifies the local dipole moment direction leading to a further stabilization of the chromophore positive charge in its S_0 location (i.e. on the C=N bond). Another change regards the flipping of the OH group of the Y191 side-chain, which further stabilize the Y268 conformation and extends the local HBN. In conclusion, while the direct stabilization of S_0

due to the E181 negative charge may not be large in Rh (i.e. such charge is located half-way along the chromophore backbone), the Rh_{181} model indicates that the same negative charge can stabilize S_0 indirectly through a modification of the local HBN (compare left and right structures in Fig. 6A). This is confirmed by the reported HBN resulting from extensive MD simulations of bovine rhodopsin model with a ionized E181 side-chain.[28]

An even larger effect is observed for the $bR_{LA(85)}$ and $bR_{DA(85)}$ models (10 kcal/mol increase in $\Delta E_{S1\text{-}S0}$), as the negatively charged D85 residue would further stabilize the Schiff base, together with the negatively charged D212 residue. The additional charge causes a further stabilization of the S_0 state relative to the S_1, predicting a further blue-shifted value relatively to the experiment. The corresponding HBN modification is reported in Fig. 6B where we compare part of the chromophore cavity of light-adapted bacteriorhodopsin with protonated and deprotonated D85 residue (bR_{LA} and $bR_{LA(85)}$ respectively). In contrast, to the bovine rhodopsin case, these results indicate that the bR_{LA}, and not $bR_{LA(85)}$, is the model best accommodating the trend of $\Delta E_{S1\text{-}S0}$ values. A similar results is obtained for bR_{DA}.

Stability of the computed side-chain conformations and hydrogen bond network

The ten 1 ns MD relaxations incorporated, for each rhodopsin model, in the ARM workflow (see Methods section) allow the sampling of different cavity conformations and HBN variations. The sensitivity of the resulting average $\Delta E_{S1\text{-}S0}$ values as a function of the number of replicas (N=1, N=5, N=10, N=15 and N=20) has been directly assessed for the very different Rh_{181}, SqR and ASR_{13C} models. The results reported in Fig. 7 show that the average values become substantially stable, with variations less that 1 kcal/mol beyond ten replicas. N=10 is therefore assumed to be a suitable default value.

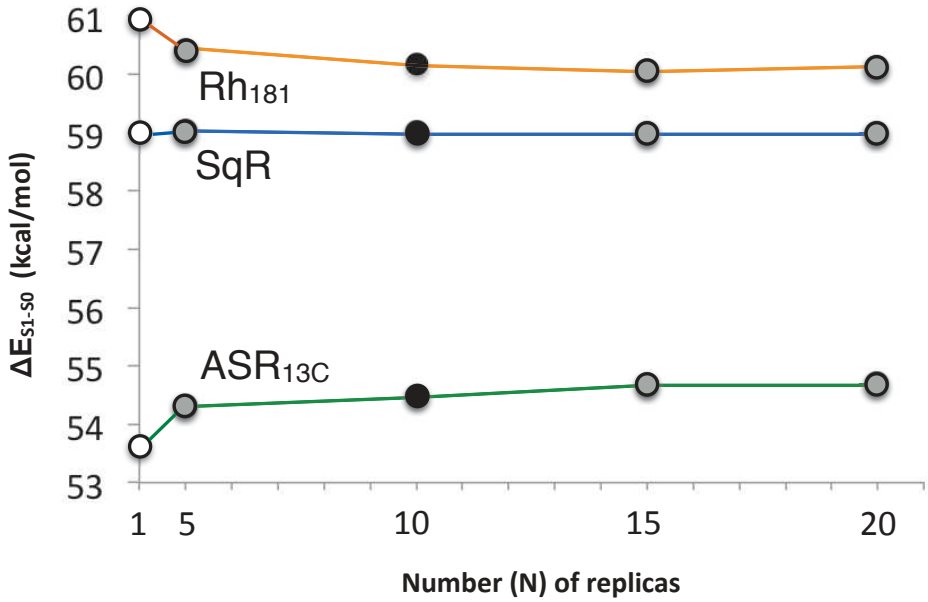


Figure 7 Features affecting the average vertical excitation energy in vertebrate, invertebrate and bacterial rhodopsin models. Effect of the sampling via N uncorrelated replicas (using different random seeds) 1 ns MD. The open circles and full circles correspond to the values of Fig. 3A and Table 1.

In Fig. 8A we instead assess the ARM model stability along a single, room-temperature 30 ns MD of the Rh model which, due to its protonated E181 residue and, therefore, weaker HBN, is more likely to undergo conformational changes with respect to Rh_{181}. It can be seen that during the first 20 ns the HBN is stable and consistent with that of the Rh model (compare Fig. 8A left structure with the Rh structure of Fig. 6A). On the other hand, at times longer than ca. 20 ns the HBN changes and displays a flipped Y191 side-chain as shown in the right structure of Fig. 8A. An ARM model (Rh_{Y191}), generated by a snapshot taken at the end of the 30 ns MD and constructed via geometry optimization produces a 1.3 kcal/mol lower $\Delta E_{S1\text{-}S0}$ value than the Rh model (see Table 3). Such difference does not change the predicted trend for wild-type rhodopsin set, but can be used to assess the model conformational flexibility. In the Rh case, the variation is attributed to a re-orientation of the dipole moment associated with the Y191 terminal O-H bond (compare the two structures in Fig. 8A). In Rh_{Y191} the O-H is hydrogen bound to the Y268 residue rather than to the M207 residue and therefore projects a positive rather than negative potential in the C=N region leading to a destabilization of the S_0 state with respect to the S_1 state. However, Rh_{Y191} has a potential energy ca. 9 kcal/mol higher than the Rh model (see the SI). Both the higher stability and the better λ^{a}_{max} prediction would favor the initial conformation, in spite of the fact that the observed HBN does not revert back during the last ten ns of the 30 ns MD. In fact, this is still a too short simulation to conclude that this is a stable HBN. As discussed above this orientation becomes stabilized when the E181 side-chain is deprotonated such as in Rh_{181}. This is because the negative E181 charge strongly stabilizes the HBN formed by Y191, Y268, E181, WAT, S186. A similar 30 ns MD

test carried out on SqR never displays an HBN change, further pointing to system-dependent results of these tests. A comprehensive investigation of the HBN dynamics would require longer and computationally demanding MD relaxations and are beyond the scope of the present work.

Table 3. Computed vertical excitation energies $\Delta E_{S1\text{-}S0}$ in kcal/mol and maximum absorption wavelengths λ^{a}_{max} (nm) for alternative ARM models of Rh and ASR. The deviation from the experimental value and the computed oscillator strengths f_{Osc} are also given.

	Calc. ΔE (λ^{a}_{max}) Average (N=10)	Error Average (N=10)	f_{Osc}
Rh_{Y191}	52.5 (544)	-4.9	0.90
$ASR_{AT/Q109}$	54.2 (527)	2.1	1.08
$ASR_{13C/Q109}$	55.0 (519)	1.8	1.13

An investigation of the side chain stability and conformational effects has also been carried out for the two ASR models. Inspection of the conformations of side chains involved in the HBN of both ASR_{AT} and ASR_{13C} reveals that these are similar in the two models, so the HBN effect on the computed $\Delta E_{S1\text{-}S0}$ value is thought to be approximately the same. However, as shown in Fig. 8B, such conformations are different from those initially assigned to the crystallographic structure templates (left structure in Fig. 8B). In fact a conformational change occurs already during the initial 150 ps MD equilibration, which modifies the orientation of the Q109 side-chain (yielding the structure of Fig. 8B right). To estimate the effect of such change we have constructed two models ($ASR_{AT/Q109}$ and $ASR_{13C/Q109}$) starting from average MD structures where the Q109 side-chain orientation was modified to resemble the crystal structure conformation. As reported in Table 3 a slightly larger $\Delta E_{S1\text{-}S0}$ value is obtained for these models (e.g. for $ASR_{AT/Q109}$ has an excitation energy 0.7 kcal/mol higher than ASR_{AT}), consistently with the change in orientation of the dipole associated to the amino acid side chain (see Fig. 8B), which stabilizes S_0 with respect to S_1 and increases the energy gap. However the $ASR_{AT/Q109}$ S_0 energy is ca. 10 kcal/mol higher than the default ARM model. So, effectively the ARM protocol seems to have "improved" on the ASR crystal structure by generating a more stable structure. Nevertheless, these results show again the importance of side chain orientation and the potentially large effect they can have on the computed vertical excitation energy.

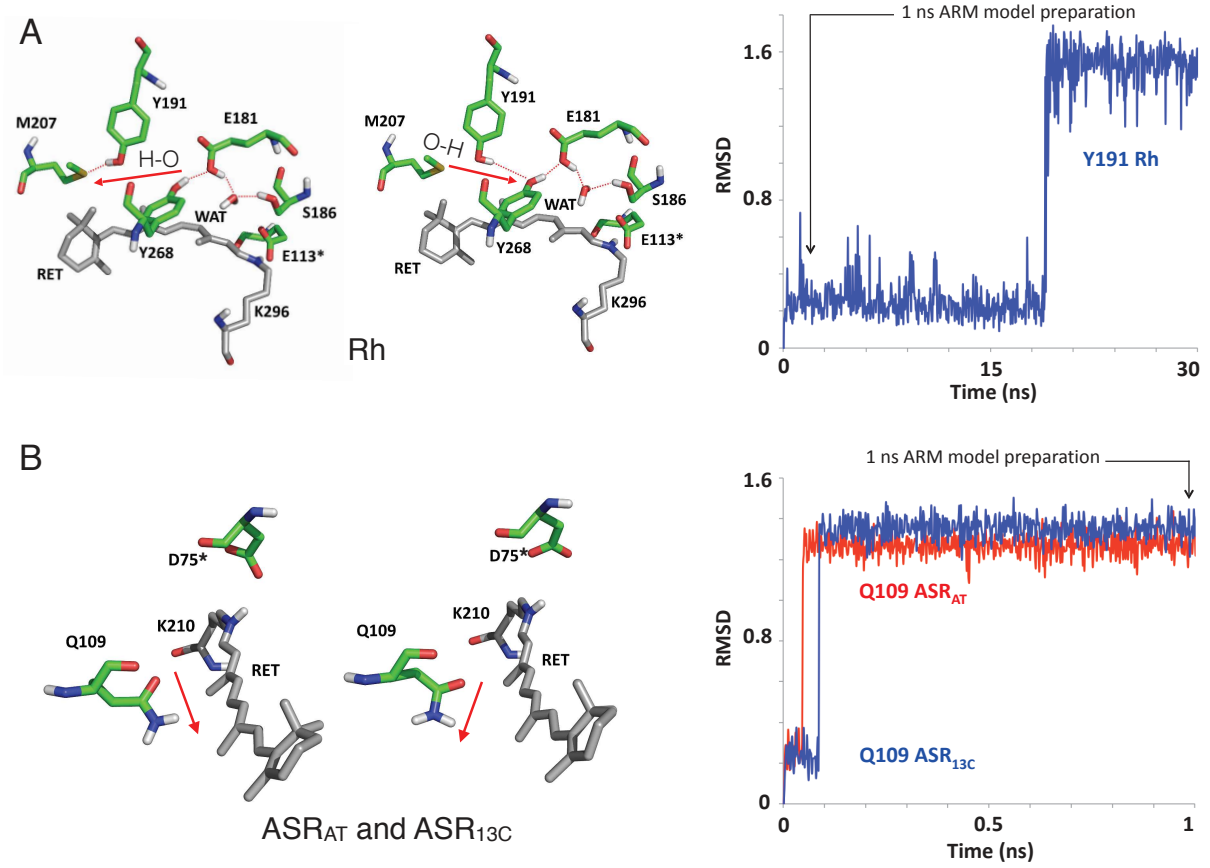


Figure 8 Changes in the side-chain conformation and related HBN. (A) Change induced in the reference ARM Rh model when a 30 ns room-temperature MD run is employed for the rhodopsin cavity sampling. The red arrow illustrates the change in orientation of the dipole moment of the Y191 residue characterizing the conformational change occurring after 20 ns MD as shown in the RMSD vs time plot. The ARM model generated starting from the 30 ns snapshot is called Rh_{Y191} and has a different HBN with respect to Rh. (B) Change in the HBN occurring during the protocol 1 ns MD run of ASR. The left structure corresponds to the original assigned X-ray crystal structure and it is obtained by geometry optimization of the initial snapshot. The right structure is the one of the final ARM model. RMSD vs time plot for the MD runs showing the corresponding conformational changes of ASR_{AT} and ASR_{13C} Q109 side-chain.

Conclusions and Outlook

Accurate QM/MM models of rhodopsins require a realistic treatment of the protein environment. This includes a membrane patch hosting the protein; inner and outer membrane water molecule layers with chloride and sodium ions (i.e. the external counterions) to counterbalance the solvent exposed ionized amino acids and obtain a globally neutral model; and the right number of protein monomers according to experimental data. A set of models of this type, targeting accurate spectroscopic studies,[28,30-32] were indeed reported including an hydrated bilayer preparation, relaxed and equilibrated with several nanoseconds of molecular dynamics (MD), followed by multiple QM/MM treatments for further sampling the protein interior and chromophore before computing the λ^{a}_{max} values. Even more resource-intensive methodologies were applied to artificial soluble retinal proteins.[30] Such computationally demanding methodologies are, presently, not adequate for

a fast *in silico* screening because they are time-consuming and their complex model building makes the workflow control difficult to automate. In order to overcome such difficulties, we have developed and benchmarked ARM as a prototype protocol for high throughput computational screening of photoreceptors and a reference for further model improvement. Accordingly, ARM generates approximated rhodopsin models useful for quickly screening experimental or computer generated sequences with the target of tracking trends in spectral or reactivity properties. Such analysis would provide candidates worth of more demanding and accurate modeling or direct experimental assessment. Thus, possible ARM applications would be the screening of the sequences associated to an evolutionary tree for tracking a specific ancestor or generated via *in silico* mutagenesis looking for sequences where a certain property is optimized.

We have shown that the information required for the construction of each ARM model is essentially a crystallographic structure or a comparative model provided in a PDB file format. However, in the current version of ARM such primary input needs to be completed by assigning the state of all ionizable side-chains and the list of the target amino acid required for counterion placement. In contrast, features such as the HBN structure (including the internal water molecule position), the Na^+ and Cl^- counterion position and the side-chain conformation of the amino acid forming the chromophore cavity are generated automatically. Thus, presently, ARM models would be more correctly defined as semi-automatically generated models.

The ARM models (see the SI for the corresponding PDB files) generated for 6 different sensory rhodopsins, 3 proton-pumping rhodopsins and one ion channel show that it is possible to reproduce the observed general trends in vertical excitation energies with blue-shifted values not significantly displaced from the observed data. This is encouraging when considering that the benchmarked rhodopsins come from evolutionarily distant organisms with different physiological functions. Remarkably, a similar level of accuracy is preserved when predicting the more challenging trend of 17 bacterial and vertebrate rhodopsin mutants. While we have shown that different factors affect the computed $\Delta E_{S1\text{-}S0}$ value, the changes due to different external counterion distributions and side-chain conformations are usually small (e.g. ca. 3-2 kcal mol^{-1}). Larger changes have been shown to be associated with different ionization state assignments of the internal residues that would also affect the stability of the HBN. For cases when these situations occur during the preliminary MD run, our results suggest that distinct models of the same rhodopsin with different ionization state assignments should be included in the set in order to examine the possible variations in predicted trends.

Future work will address the present uncertainties, by introducing a more accurate placement and optimization of all hydrogen atoms, a better treatment of the external environment, an improved treatment of the chromophore cavity. Furthermore, in the case of hMeOp, we used a comparative model as starting structure (there is no X-ray crystal structure for this protein) derived according to a procedure discussed in ref. [9] and taking as a template SqR (see the Supporting Information). The sensitivity of the predicted $\Delta E_{S1\text{-}S0}$ values to the specific procedure used to generate the starting hMeOp model and its impact on the error relative to SqR remain to be investigated.

In conclusion, ARM constitutes the first step towards a fast and cheap computational tool filling the gap between slow experimental structure determination and fast protein expression and characterization achieved by genetic engineering techniques. It also represents an attempt to provide a basis for the standardization and reproducibility of rhodopsin QM/MM models. If such research line will be proven to be fruitful, our ultimate goal would be the extension of the same methodologies to other photoresponsive protein families, so that ARM could eventually become a general tool in computational photobiology.

ASSOCIATED CONTENT

Supporting text and coordinates of the input wild-type structures (11 models) and optimized (28 models) benchmark wild-type and mutant models in PDB format. This material is available free of charge via the Internet at http://pubs.acs.org.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under grant no. CHE-1152070 and CHE-1551416 and the Human Frontier Science Program Organization under grant RGP0049/2012CHE09-56776. M.O. is grateful to the Center for Photochemical Sciences and School of Arts and Sciences of the Bowling Green State University. The authors are indebted to NSF-XSEDE and Ohio Supercomputer Center for granted computer time. M.O. and Y.O. are grateful to the Foundation of the University of Strasbourg for an USIAS fellowship supporting a visiting professorship at the IPCMS. M.S. and N.F. thank the French Agence Nationale de la Recherche for funded project FEMTO-ASR, n° ANR-14-CE35-0015-02.

References

(1) Kircher, M.; Kelso, J. *BioEssays* **2010***, 32,* 524-536.

(2) Doyle, S.; Koehn, J.; Hunt, I. in *High Throughput Protein Expression and Purification;* Humana Press: Totowa, NJ, 2009; 498, pp 1-18.

(3) Berry, S. M.; Lu, Y. Protein Structure Design and Engineering in *eLS;* John Wiley & Sons, Ltd: published on-line, 2001.

(4) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976***, 103,* 227-249.

(5) Altoè, P.; Cembran, A.; Olivucci, M.; Garavelli, M. *Proc. Natl. Acad. Sci. U. S. A.* **2010***, 107,* 20172-20177.

(6) Andruniow, T.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U. S. A.* **2004***, 101,* 17908-17913.

(7) Coto, P. B.; Strambi, A.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U. S. A.* **2006***, 103,* 17154-17159.

(8) Strambi, A.; Durbeij, B.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U. S. A.* **2010***, 107,* 21322-21326.

(9) Rinaldi, S.; Melaccio, F.; Gozem, S.; Fanelli, F.; Olivucci, M. *Proc. Natl. Acad. Sci. U. S. A.* **2014***, 111,* 1714-1719.

(10) Huntress, M. M.; Gozem, S.; Malley, K. R.; Jailaubekov, A. E.; Vasileiou, C.; Vengris, M.; Geiger, J. H.; Borhan, B.; Schapiro, I.; Larsen, D. S.; Olivucci, M. *J. Phys. Chem. B.* **2013***, 117,* 10053-10070.

(11) Schreiber, M.; Sugihara, M.; Okada, T.; Buss, V. *Angew. Chem. Int. Ed.* **2006***, 45,* 4274-4277.

(12) Tomasello, G.; Olaso-Gonzàlez, G.; Altoè, P.; Stenta, M.; Serrano-Andrés, L.; Merchàn, M.; Orlandi, G.; Bottoni, A.; Garavelli, M. *J. Am. Chem. Soc.* **2009***, 131,* 5172-5186.

(13) Altun, A.; Yokoyama, S.; Morokuma, K. *J. Phys. Chem. B.* **2008***, 112,* 16883-16890.

(14) Altun, A.; Yokoyama, S.; Morokuma, K. *J. Phys. Chem. B.* **2008***, 112,* 6814-6827.

(15) Ernst, O. P.; Lodowski, D. T.; Elstner, M.; Hegemann, P.; Brown, L. S.; Kandori, H. *Chem. Rev.* **2014***, 114,* 126-163.

(16) Spudich, J. L.; Yang, C. S.; Jung, K. H.; Spudich, E. N. *Annu. Rev. Cell. Dev. Bi.* **2000***, 16,* 365-392.

(17) Gorostiza, P.; Isacoff, E. Y. *Science* **2008***, 322,* 395-399.

(18) Boyden, E. S.; Zhang, F.; Bamberg, E.; Nagel, G.; Deisseroth, K. *Nat. Neurosci.* **2005***, 8,* 1263-1268.

(19) Han, X.; Boyden, E. S. *PLoS One* **2007***, 2,* e299.

(20) Schoenenberger, P.; Gerosa, D.; Oertner, T. G. *PLoS One* **2009***, 4,* e8185.

(21) Deisseroth, K. *Nat. Methods.* **2011***, 8,* 26-29.

(22) Sekharan, S.; Sugihara, M.; Buss, V. *Angew. Chem. Int. Ed.* **2007***, 46,* 269-271.

(23) Bravaya, K.; Bochenkova, A.; Granovsky, A.; Nemukhin, A. *J. Am. Chem. Soc.* **2007***, 129,* 13035-13042.

(24) Fujimoto, K.; Hayashi, S.; Hasegawa, J.; Nakatsuji, H. *J. Chem. Theory Comput.* **2007***, 3,* 605-618.

(25) Schapiro, I.; Ryazantsev, M. N.; Ding, W. J.; Huntress, M. M.; Melaccio, F.; Andruniow, T.; Olivucci, M. *Aust. J. Chem.* **2010***, 63,* 413-429.

(26) El-Khoury, P. Z.; Schapiro, I.; Huntress, M. M.; Melaccio, F.; Gozem, S.; Frutos, L. M.; Olivucci, M. in *CRC Handbook of Organic Photochemistry and Photobiology;* Griesbeck, A., Oelgemöller, M., and Ghetti, F., Eds.; CRC Press: Boca Raton, FL, 2012; 1029-1056.

(27) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986***, 7,* 718-730.

(28) Valsson, O.; Campomanes, P.; Tavernelli, I.; Rothlisberger, U.; Filippi, C. *J. Chem. Theory Comput.* **2013***, 9,* 2441-2454.

(29) Campomanes, P.; Neri, M.; Horta, B. A. C.; Röhrig, U. F.; Vanni, S.; Tavernelli, I.; Rothlisberger, U. *J. Am. Chem. Soc.* **2014***, 136,* 3842-3851.

(30) Cheng, C.; Kamiya, M.; Uchida, Y.; Hayashi, S. *J. Am. Chem. Soc.* **2015***, 137,* 13362-13370.

(31) Kato, H. E.; Kamiya, M.; Sugo, S.; Ito, J.; Taniguchi, R.; Orito, A.; Hirata, K.; Inutsuka, A.; Yamanaka, A.; Maturana, A. D.; Ishitani, R.; Sudo, Y.; Hayashi, S.; Nureki, O. *Nat. Commun.* **2015***, 6,* 7177.

(32) Guo, Y.; Beyle, F. E.; Bold, B. M.; Watanabe, H. C.; Koslowski, A.; Thiel, W.; Hegemann, P.; Marazzi, M.; Elstner, M. *Chem. Sci.* **2016***, 7,* 3879-3891.

(33) Jung, K. H.; Trivedi, V. D.; Spudich, J. L. *Mol. Microbiol.* **2003***, 47,* 1513-1522.

(34) Vogeley, L.; Sineshchekov, O. A.; Trivedi, V. D.; Sasaki, J.; Spudich, J. L.; Luecke, H. *Science* **2004***, 306,* 1390-1393.

(35) Okada, T.; Sugihara, M.; Bondar, A. -N.; Elstner, M.; Entel, P.; Buss, V. *J. Mol. Biol.* **2004***, 342,* 571-583.

(36) Murakami, M.; Kouyama, T. *Nature* **2008***, 453,* 363-367.

(37) Provencio, I.; Jiang, G.; De Grip, W. J.; Hayes, W. P.; Rollag, M. D. *Proc. Natl. Acad. Sci. U. S. A.* **1998***, 95,* 340-345.

(38) Luecke, H.; Schobert, B.; Richter, H. -T.; Cartailler, J. -P.; Lanyi, J. K. *J. Mol. Biol.* **1999***, 291,* 899-911.

(39) Nishikawa, T.; Murakami, M.; Kouyama, T. *J. Mol. Biol.* **2005***, 352,* 319-328.

(40) Ran, Tingting; Ozorowski, Gabriel; Gao, Yanyan; Sineshchekov, Oleg A.; Wang, Weiwu; Spudich, John L.; Luecke, Hartmut *Acta Crystallogr. D* **2013***, 69,* 1965-1980.

(41) Yoshizawa, T.; Wald, G. *Nature* **1963***, 197,* 1279-1286.

(42) Nakamichi, H.; Okada, T. *Angew. Chem. Int. Ed.* **2006***, 45,* 4270-4273.

(43) Mathies, R.; Stryer, L. *Proc. Natl. Acad. Sci. U. S. A.* **1976***, 73,* 2169-2173.

(44) Bonacic-Koutecky, V.; Köhler, K.; Michl, J. *Chem. Phys. Lett.* **1984***, 104,* 440-443.

(45) Roos, B. O. in *Ab. initio. Methods. in. Quantum. Chemistry.;* Lawley, K. P., Ed.Wiley: New York, 1987; pp 399-446.

(46) Andersson, K.; Malmqvist, P. -; Roos, B. O.; Sadlej, A. J.; Wolinski, K. J. *J. Phys. Chem.* **1990***, 94,* 5483-5488.

(47) Coto, P. B.; Martí, S.; Oliva, M.; Olivucci, M.; Merchán, M.; Andrés, J. *J. Phys. Chem. B* **2008***, 112,* 7153-7156.

(48) Navizet, I.; Liu, Y. -J.; Ferré, N.; Xiao, H. -Y.; Fang, W. -H.; Lindh, R. *J. Am. Chem. Soc.* **2009***, 132,* 706-712.

(49) Altoè, P.; Climent, T.; De Fusco, G. C.; Stenta, M.; Bottoni, A.; Serrano-Andrès, L.; Merchàn, M.; Orlandi, G.; Garavelli, M. *J. Phys. Chem. B* **2009***, 113,* 15067-15073.

(50) Ferré, N.; Olivucci, M. *J. Am. Chem. Soc.* **2003***, 125,* 6868-6869.

(51) Schapiro, I.; Ryazantsev, M. N.; Frutos, L. M.; Ferré, N.; Lindh, R.; Olivucci, M. *J. Am. Chem. Soc.* **2011***, 133,* 3354-3364.

(52) Polli, D.; Altoe, P.; Weingart, O.; Spillane, K. M.; Manzoni, C.; Brida, D.; Tomasello, G.; Orlandi, G.; Kukura, P.; Mathies, R. A.; Garavelli, M.; Cerullo, G. *Nature* **2010***, 467,* 440-443.

(53) Frutos, L. M.; Andruniow, T.; Santoro, F.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U. S. A.* **2007***, 104,* 7764-7769.

(54) Gozem, S.; Melaccio, F.; Lindh, R.; Krylov, A. I.; Granovsky, A. A.; Angeli, C.; Olivucci, M. *J. Chem. Theory Comput.* **2013***, 9,* 4495-4506.

(55) Sekharan, S.; Altun, A.; Morokuma, K. *Chem-Eur. J.* **2010***, 16,* 1744-1749.

(56) Sekharan, S.; Wei, J. N.; Batista, V. S. *J. Am. Chem. Soc.* **2012***, 134,* 19536-19539.

(57) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996***, 14,* 33-38.

(58) 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. *J. Am. Chem. Soc.* **1995***, 117,* 5179-5197.

(59) Zhang, L.; Hermans, J. *Proteins* **1996***, 24,* 433-438.

(60) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987***, 8,* 1016-1024.

(61) Strambi, A.; Coto, P. B.; Frutos, L. M.; Ferré, N.; Olivucci, M. *J. Am. Chem. Soc.* **2007***, 130,* 3382-3388.

(62) Jardón-Valadez, E.; Bondar, A. -N.; Tobias, D. J. *Biophys. J.* **2009***, 96,* 2572-2576.

(63) Jardón-Valadez, E.; Bondar, A. -N.; Tobias, D. J. *Biophys. J.* **2010***, 99,* 2200-2207.

(64) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. *Nucleic. Acids. Res.* **2006***, 34,* W116-W118.

(65) Pronk, S.; PÃ¡ll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013***, 29,* 845-854.

(66) Ferré, N.; Cembran, A.; Garavelli, M.; Olivucci, M. *Theor. Chem. Acc.* **2004***, 112,* 335-341.

(67) Aquilante, F.; De Vico, L.; Ferré, N.; Ghigo, G.; Malmqvist, P. A.; Neogrady, P.; Pedersen, T. B.; Pitonak, M.; Reiher, M.; Roos, B. O.; Serrano-Andres, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comput. Chem.* **2010***, 31,* 224-247.

(68) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986***, 7,* 718-730.

(69) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996***, 105,* 1959-1967.

(70) Stålring, J.; Bernhardsson, A.; Lindh, R. *Mol. Phys.* **2001***, 99,* 103-114.

(71) Ferré, N.; Angyan, J. G. *Chem. Phys. Lett.* **2002***, 356,* 331-339.

(72) Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. *Proteins* **2009***, 77,* 778-795.

(73) Melaccio, F.; Olivucci, M.; Lindh, R.; Ferré, N. *Int. J. Quant. Chem.* **2011***, 111,* 3339-3346.

(74) Kato, H. E.; Zhang, F.; Yizhar, O.; Ramakrishnan, C.; Nishizawa, T.; Hirata, K.; Ito, J.; Aita, Y.; Tsukazaki, T.; Hayashi, S. *Nature* **2012***, 482,* 369-374.

(75) Olsson, M. H. M.; Sondergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011***, 7,* 525-537.

(76) Främcke, J. S.; Wanko, M.; Phatak, P.; Mroginski, M. A.; Elstner, M. *J. Phys. Chem. B.* **2010***, 114,* 11338-11352.

(77) Sandberg, M. N.; Amora, T. L.; Ramos, L. S.; Chen, M. -H.; Knox, B. E.; Birge, R. R. *J. Am. Chem. Soc.* **2011***, 133,* 2808-2811.

(78) Hall, K. F.; Vreven, T.; Frisch, M. J.; Bearpark, M. J. *J. Mol. Biol.* **2008***, 383,* 106-121.

(79) Lanyi, J. K. *BBA-Bioenergetics* **2006***, 1757,* 1012-1018.

(80) Kandori, H.; Shimono, K.; Sudo, Y.; Iwamoto, M.; Shichida, Y.; Kamo, N. *Biochemistry* **2001***, 40,* 9238-9246.

(81) Furutani, Y.; Kawanabe, A.; Jung, K. H.; Kandori, H. *Biochemistry* **2005***, 44,* 12287-12296.

(82) Kawanabe, A.; Furutani, Y.; Jung, K. H.; Kandori, H. *Biochemistry* **2006***, 45,* 4362-4370.

(83) Stein, C. J.; Reiher, M. *J. Chem. Theory Comput.* **2016***, 12,* 1760-1771.

(84) Matsuyama, T.; Yamashita, T.; Imamoto, Y.; Shichida, Y. *Biochemistry* **2012***, 51,* 5454-5462.

(85) Bailes, H. J.; Lucas, R. J. *P. Roy. Soc. B-Biol. Sci.* **2013***, 280,* 20122987.

(86) Granovsky, A. A. *J. Chem. Phys.* **2011***, 134,* 214113-214127.

(87) Gozem, S.; Huntress, M.; Schapiro, I.; Lindh, R.; Granovsky, A. A.; Angeli, C.; Olivucci, M. *J. Chem. Theory Comput.* **2012***, 8,* 4069-4080.

(88) Melaccio, F.; Ferré, N.; Olivucci, M. *Phys. Chem. Chem. Phys.* **2012***, 14,* 12485-12495.

TOC