Investigating the Origin of Automatic Rhodopsin Modeling Outliers Using the Microbial Gloeobacter Rhodopsin as Testbed

Darío Barreiro-Lage, Vincent Ledentu, Jacopo D'ascenzi, Miquel Huix-Rotllant, Nicolas Ferré

Investigating the origin of Automatic Rhodopsin Modeling outliers using the microbial Gloeobacter rhodopsin as testbed

Darío Barreiro-Lage,[†] Vincent Ledentu,[†] Jacopo D'Ascenzi,[‡,¶] Miquel Huix-Rotllant,[*,†] and Nicolas Ferré[*,†]

†*Aix Marseille Univ, CNRS, ICR, 13013 Marseille, France*

‡*Dipartimento di Biotecnologie, Chimica e Farmacia, Università degli Studi di Siena, 53100 Siena, Italy*

¶*Dipartimento di Chimica, Biologia e Biotecnologie, Università degli studi di Perugia, 06123 Perugia, Italy*

E-mail: miquel.huix-rotllant@cnrs.fr; nicolas.ferre@univ-amu.fr

Abstract

The Automatic Rhodopsin Modeling (ARM) approach is a computational workflow devised for the automatic buildup of hybrid quantum mechanics/molecular mechanics (QM/MM) models of wild-type rhodopsins and mutants, with the purpose of establishing trends in their photophysical and photochemical properties. Despite the success of ARM for accurately describing the visible light absorption maxima of many rhodopsins, for a few cases, called outliers, it might lead to large deviations with respect to experiments. Applying ARM to Gloeobacter Rhodopsin (GR), a microbial rhodopsin with important applications in optogenetics, we analyze the origin of such outliers in the absorption energies obtained for GR wild-type and mutants at neutral pH, with a total root mean square deviation (RMSD) of 0.42 eV with respect to the experimental GR excitation energies. Having discussed the importance and the uncertainty of one particular amino-acid pK_a, namely histidine at position 87, we propose and test several modifications to the standard ARM protocol: (i) improved pK_a predictions along with the consideration of several protonation microstates, (ii) attenuation of the opsin electrostatic potential at short-range, (iii) substitution of the state-average complete active space (CAS) electronic structure method by its state-specific approach, and (iv) complete replacement of CAS with mixed-reference spin-flip time-dependent density functional theory (MRSF-TDDFT). The best RMSD result we obtain is 0.2 eV combining the protonation of H87 and using MRSF/CAMH-B3LYP.

Introduction

For at least 30 years, rhodopsins[1–3] have been used as paradigmatic transmembrane photoactive biomolecules for developing more and more advanced molecular models based on the embedding of its retinal chromophore (a polyene bound to a lysine amino-acid through a protonated Schiff base (PSB)), treated at the quantum mechanical level, with various approximate descriptions of the opsin and its surroundings (cell membrane, ions, water molecules, ...)[4] The most popular one is undoubtedly the so-called QM/MM approach[5–14] in

which the retinal electronic structure is polarized by the electrostatic potential generated by permanent (and sometimes induced) multipoles.[4,15–24] Resulting QM/MM models depend on a large number of parameters: the QM size and the corresponding affordable level of theory,[25,26] the MM forcefield and the conformational sampling,[8,27,28] the electrostatic coupling scheme between the two subsystems,[29] the treatment of the frontier bond(s),[30–33] etc. Because of this unavoidable empirical choice of parameters, it is often difficult to compare results obtained with different QM/MM setups.

Massimo Olivucci's Automatic Rhodopsin Model (ARM) approach[34–38] provides a standardized workflow, meant for ensuring repeatability, accuracy and reliability with respect to experimental trends.

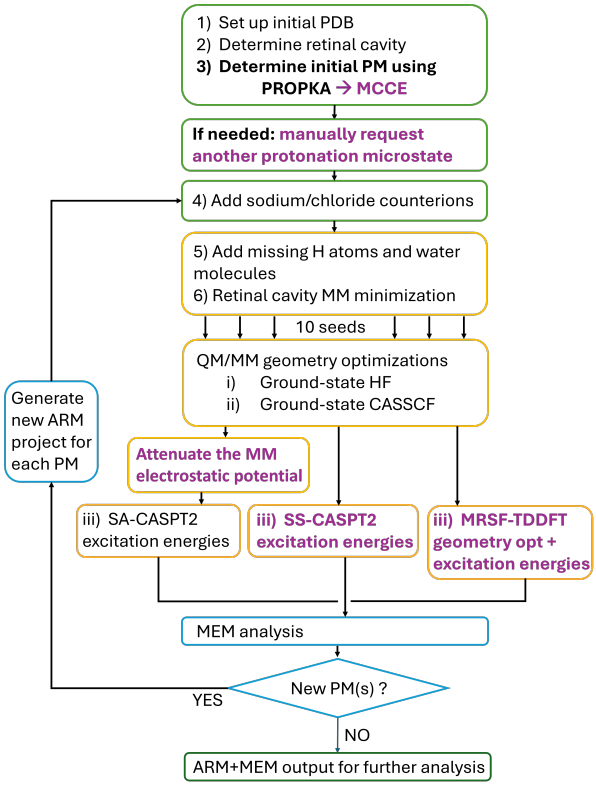


Figure 1: ARM+MEM workflow: ARM input generator (green boxes), ARM QM/MM calculator (yellow boxes), and Minimal Electrostatic Model(MEM) analysis (blue boxes). In the present work, we study the effect of the modifications written in magenta on the excitation energy of Gloeobacter Rhodopsin. PM stands for protonation microstate.

As schematized in Figure 1, green boxes, the ARM protocol starts with a structural

model of the rhodopsin of interest. After proper examination, detection of the retinal and its cavity, addition of missing atoms and extra water molecules, orientation with respect to the membrane mean planes, neutralization of the system total electric charge, structural relaxation of the cavity, a single molecule model suitable for further QM/MM calculations is produced, allowing the computation – through geometry optimizations at the Complete Active Space Self-Consistent Field (CASSCF)[39,40] wavefunctions and energies using Second-Order Perturbation Theory (CASPT2)[41] – of the rhodopsin absorption spectrum or fluorescence spectrum, the mapping of its excited state potential energy surfaces, the determination of initial conditions for subsequent non-adiabatic molecular dynamics. The ARM protocol has been packaged in the `pyARM` suite of programs[37] whose dissemination guarantees the numerical fidelity of its results.

The ARM workflow is not intended to produce numbers in quantitative agreement with experimental results. Instead, it performs well when one is interested in trends for absorption or emission maxima. For instance, a benchmark set of 44 animal and microbial rhodopsins features absorption energy values ranging from 2.71 eV to 2.15 eV, with a blue-shift with respect to experimental maximum absorption wavelengths of about 0.1 eV.[37] However, in a few special cases, the quality of the ARM trend can deteriorate because of some outliers, which are cases for which the deviation to the experimental value is much larger. For example, the ARM Krokinobacter rhodopsin 2 model features two side-chain rotamers of its retinal counterion, one of them giving closer excitation energy to experiment than the other rotamer.[37] Also, the experience with ARM accumulated over the years has pointed out a possible cure to outliers: changing the protonation state of one particular amino-acid, usually close to the retinal (but not its primary counterion). As a matter of fact, retinal protonated Schiff base interacts with two potential counterions in most bacterial rhodopsins. When one of those two is neutralized, the ARM excitation energy is often in much better agreement with the experimental one.[37] In ARM, the protonation state of a residue derives from its deprotonation probability, calculated at a given pH using its pK_a value as predicted

by the PROPKA software.[42] Since it generates pK_a values which only depend on the structure of the protein, PROPKA may face trouble when it is used on systems for which it has not been formally validated, like transmembrane rhodopsin proteins. For this reason, and also when the pK_a value is close to the working pH, a different protonation microstate (defined as a particular protonation state for each titratable residue in the protein) can be selected *manually*. Most of the time, this *ad hoc* "trick" performs well, significantly improving the quality of the ARM outcome, however at the price of requiring some expertise from its user and losing the ARM automatic character.

To avoid such an arbitrary modification of the ARM parameters, we first decide to investigate the possibility of going out of the single molecule paradigm in ARM. Indeed, at a given pH, multiple protonation states from the $2^N - 1$ possible microstates (N being the number of titrated amino-acids) can contribute to the total absorption spectrum. Accordingly, we recently extended the ARM protocol with our Minimal Electrostatic Model (MEM) analysis[43,44] (blue boxes in Figure 1). Based on how a change in an amino-acid electric charge modifies its (classical) electrostatic interaction with the retinal, the MEM analysis delivers a list of new protonation microstates and a rough estimate of the retinal excitation energy shift induced by each of them. This allows classifying and discriminating important protonation microstates, reasonably increasing the computational cost since each new protonation microstate requires the application of the full ARM workflow. This process needs to be applied iteratively. While the resulting workflow,[44] denoted ARM+MEM in the following, is expected to deliver better absorption spectra, this approach still depends strongly on the quality of the pK_a values, which we recall are provided by PROPKA in the standard ARM protocol.

Besides the treatment of multiple protonation states in ARM, we also investigate three other possible improvements to the standard ARM protocol. The first one is based on the attenuation of the MM electrostatic potential at short range. We show this is a very effective way to improve the agreement with experimental trends, albeit this does not imply

an improved description of QM/MM interactions. The second one is based on the analysis of the state-average CASSCF wavefunction in its three most stable electronic states, S_0, S_1 and S_2. In particular, we show that state-specific CASSCF wavefunction can often alleviate the problem of the identification of the bright state sometimes occurring with state-average CASSCF. Finally, we also consider a density functional-based computationally less expensive electronic structure method independent of a choice of active space, namely Mixed-Reference Spin-Flip Time-Dependent Density Functional Theory (MRSF-TDDFT), in substitution of the CASSCF/CASPT2.[45,46]

The relevance of the above-mentioned ARM modifications is tested using *Gloeobacter* rhodopsin (GR), a specific microbial rhodopsin found in the cyanobacterium *Gloeobacter violaceus*.[47] GR functions as a light-driven proton pump, transferring a proton from the cytoplasmic region to the extracellular region in a cell.[48] GR is particularly important in the field of optogenetics, due to nonoverlapping absorption and fluorescence spectral ranges.[49] GR is spectroscopically well characterized at several pH values: its absorption maximum is red-shifted from alkaline pH (8.0, $\lambda_{\max} = 545\,\text{nm}$) to acidic pH (3.0, $\lambda_{\max} = 560\,\text{nm}$). Dozens of mutants have been produced, exhibiting shifts of GR absorption maximum as large as $\pm 80\,\text{nm}$.[49]

Computational details

ARM QM/MM computations; MEM analysis

The ARM protocol has been applied using the `pyARM` package,[37] with a few changes with respect to the default parameter values, as reported below. In ARM, QM/MM calculations are performed using `OpenMolcas`[50,51] version 20.10 coupled to a patched version of `Tinker` 6.3.3.[52] The QM/MM electrostatic interaction scheme is based on the ElectroStatic Potential Fitted (ESPF) charge operators.[53,54] The Amber94 forcefield[55] is applied to the MM subsystem and the QM/MM van der Waals and bonded interactions.

The QM/MM workflow in ARM[37] includes a series of geometry optimizations: (i) HF/3-21G/MM, (ii) S_0 state-specific (SS) complete active space self-consistent field SS-CASSCF/3-21G/MM, and (iii) S_0 SS-CASSCF/6-31G*/MM levels of theory. In all cases, the selected active space in CASSCF covers the entire π-system of retinal. The final step in the standard ARM protocol involves a three-root State Average (SA3) CASSCF(12,12)/6-31G*/MM and then a CASPT2 single point calculation. $S_0 \rightarrow S_1$ and $S_0 \rightarrow S_2$ oscillator strengths are obtained using the SA-CASSCF wavefunctions and the CASPT2 energies (hereafter denoted as SA-CASPT2) computed for these 3 singlet states. In the CASPT2 calculation, an imaginary shift of $0.2\,E_\mathrm{h}$ is applied to prevent intruder states, and the IPEA shift is set to $0.0\,E_\mathrm{h}$, following literature recommendation.[56,57] Since the MEM methodology requires that the point charges of the retinal are calculated *in vacuo*, an additional single-point three-root SA-CASSCF(12,12)/6-31G* calculation is performed on the optimized geometry to obtain these charges.

Additionally, we use two other electronic structure methods. At the moment of the electronic excitation computation only, we investigate the State-Specific Complete Active Space (SS-CASSCF) method,[39] producing different sets of orbitals for different electronic states, followed by CASPT2 calculations for each of them.[58] The SS-CASSCF has been shown in the past to describe better intramolecular charge-transfer states, and therefore, it could improve the description of retinal S_1.[59,60] For performing SS-CASSCF calculations, each electronic state is converged by setting a weight of 1.0 for the root of interest, and 0.0 for the other two roots. In the case that the same active space cannot be guaranteed for all electronic states, a quasi-SS-CASSCF has been performed, in which the root of interest is set to 1.0-X and the other two roots are set to X/2. The variable X is set initially to 0.01 and increased by units of 0.01 until all states can be converged with the same active space. In all cases, X<=0.1. The resulting SS-CASPT2 calculations are performed, as their state-average counterparts, using the same `OpenMolcas` and `Tinker` software versions.

We also perform calculations based on a recently developed electrostatic embedding

QM/MM method for excited states, combining the strengths of MRSF-TDDFT and electrostatic embedding QM/MM with ESPF charges. MRSF-TDDFT combines the reduced density matrices (RDMs) of the two $M_S = \pm 1$ triplet references within the linear response theory.[45,61] This approach provides additional nondynamic types of electron correlation, which are missing in conventional linear response TDDFT. The balanced dynamic and nondynamic electron correlations of MRSF-TDDFT enables accurate computation of excitation energies and conical intersections, overcoming the limitations of traditional TDDFT. Moreover, MRSF-TDDFT eliminates the major drawback of spin-flip TDDFT methods, namely the spin-contamination of states described by excitations outside the open-shell orbital space.[62] The MRSF-TDDFT/MM with ESPF coupling has been implemented in a local development version of `GAMESS-US` (R2 Patch 2)[63] interfaced with `Tinker` 8.10.1,[52] modified to incorporate the ESPF QM/MM interaction model.[46] All reported MRSF-TDDFT calculations are performed at the CAMh-B3LYP/6-31G* level of theory (additional calculations using BH&HLYP/6-31G* and rCAM-B3LYP/6-31G* are reported in the SI).[64–66]

The MEM analysis[43,44] is based on the (de)protonation-induced changes in classical electrostatic interaction energies between the retinal chromophore, considered in two electronic states I and J, and the titratable protein residues:

$$\Delta E(\mathrm{pH}) = \Delta E(\mathrm{ref}) - \sum_{r=1}^{N_p} x_r(\mathrm{pH})\delta E_r + \sum_{r=1}^{N_d} \left(1 - x_r(\mathrm{pH})\right) \delta E_r \tag{1}$$

in which $\Delta E(\mathrm{ref})$ is the excitation energy in a reference protonation microstate (i.e. the one automatically selected by ARM), N_d (respectively N_p) is the number of deprotonated (respectively protonated) residues, x_r is the deprotonation probability of the r-th residue at a given pH and δE_r is its pH-independent contribution to the excitation energy change upon (de)protonation, defined as:

$$\delta E_r = \frac{e^2}{4\pi\epsilon_0\epsilon_r} \sum_{a=1}^{N_{\mathrm{QM}}} \frac{q_a^J - q_a^I}{|\mathbf{r}_r - \mathbf{R}_a|} \tag{2}$$

with ϵ_r an effective dielectric constant, mainly depending on the location of residue r in the protein, N_{QM} the number of QM atoms in retinal, q_a^K the atomic charge of (bare) retinal a-th atom in its electronic state K, $|\mathbf{r}_r - \mathbf{R}_a|$ the distance between QM atom a and the center of charge of the r-th MM residue. In practice, the MEM step in the ARM+MEM protocol[44] is fed with CASPT2 vertical excitation energies, retinal ESPF atomic charges calculated *in vacuo* at the CASSCF level, and residue charges of the titratable amino-acids defining a particular protonation microstate. An energy threshold of 0.01 kcal/mol is used in the MEM analysis. Upon completion of the initial ARM+MEM step, the process identifies other important protonation microstates that may significantly blue-shift or red-shift the rhodopsin absorption spectrum. Inputs for these additional microstates are generated automatically, and subsequent ARM+MEM calculations are conducted until the space of relevant protonation microstates is incrementally and thoroughly sampled. ARM+MEM excitation energies correspond to the maxima of the computed absorption spectrum, as presented in Supporting Information section 1.

Gloeobacter Rhodopsin

The initial structure of our GR model is based on the X-ray structure reported in the 6NWD accessible in the RCSB Protein Data Bank.[67,68] It should be noted that it was experimentally determined in acidic pH conditions (pH=3.4). The retinal chromophore is located in a cavity featuring 3 titratable residues: aspartate D121 (associated with histidine H87[69]), aspartate D253, glutamic acid E166. D121 and D253 are bridged by a water molecule, labeled HOH401 (Figures 2(a) and 2(b)).

When `pyARM` is applied to the 6NWD PDB file, the following manual choices are made: the pH is set to 7.5, while all histidines are deprotonated, including H87. The aspartic acid E132 is protonated in accordance with its predicted pK_{a} value, and the protonation state of the main retinal counterion, aspartate D253, is kept fixed. The dark state of GR mainly contains an *all-trans* retinal isomer with a retinal composition that does not change in the

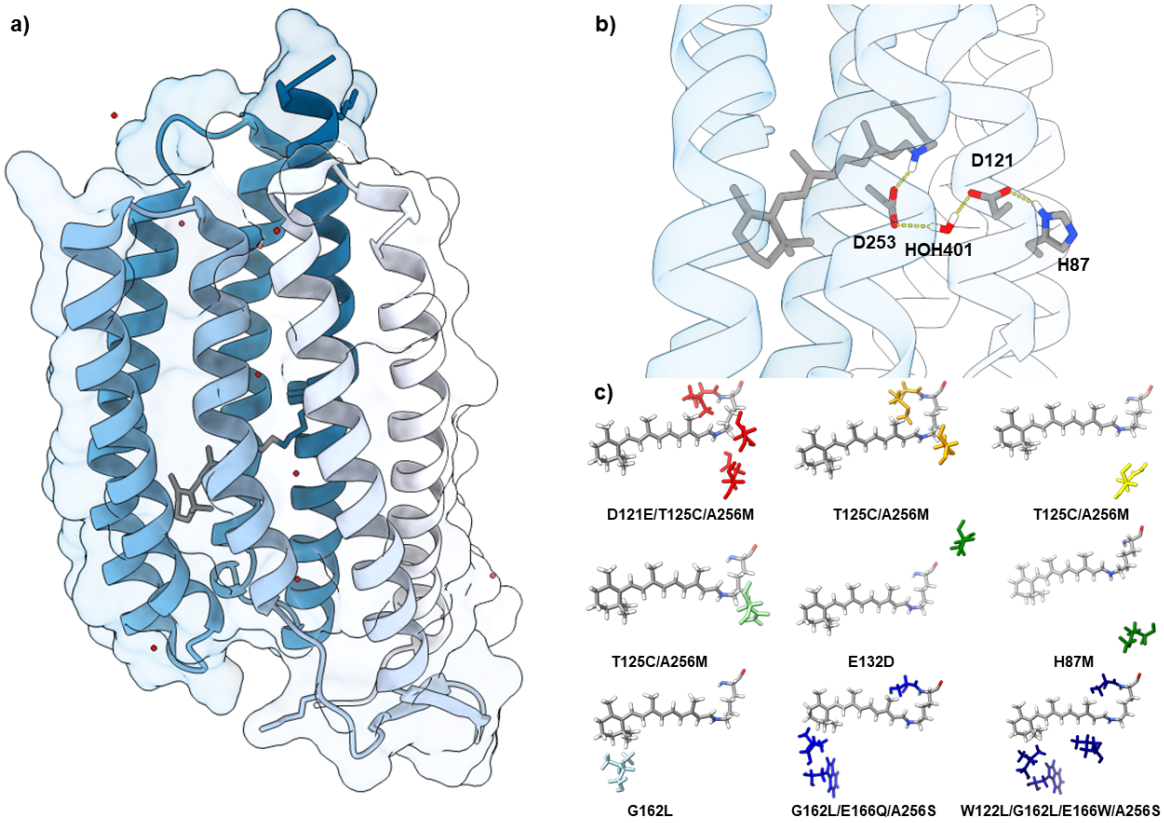


Figure 2: **a)** Experimental crystallographic structure of *Gloeobacter* rhodopsin from the RCSB Protein Data Bank entry 6NWD. The retinal chromophore and the lysine LYS257 to which retinal is bound are depicted with sticks. Small red dots represent the crystallographic water oxygen atoms. **b)** Retinal chromophore together with the main (D253) and secondary (D121) counterions as obtained from the ARM protocol at pH = 7.5. The two counterions are bridged by the water molecule HOH401. Note the presence of histidine H87 in the vicinity of D121. **c)** Graphical representations of the nine different mutants considered in this work.

light-adapted state. Accordingly, we only consider this retinal conformation in the following. Together with the GR wild type (WT), we consider nine different mutants for which experimental absorption spectra are available: D121E/T125C/A256M, T125C/A256M, D121N, T125N, E132D, H87M, G162L, G162L/E166W/A256S and W122L/G162L/E166W/A256S (Figure 2c). Their corresponding experimental excitation energies (identified to the spectrum maximum absorption wavelength) range from 2.0 to 2.7 eV.[49,70]

Results & Discussion

Multiple protonation states

We first compute the electronic absorption spectra of GR WT and several mutants using the ARM protocol: their maximum transition energies are reported in Figure 3(a). While two mutants (D121N and D121E/T125C/A256M) confirm the reported excellent accuracy (about 0.1 eV) of the ARM protocol with respect to experiments, the 8 other excitation energies are significantly worse: 0.25 to 0.70 eV blue-shifted with respect to experiment. The best agreement is thus obtained when a single counterion is present in the model: in D121N, the secondary counterion is replaced by a non-titratable arginine residue, while in D121E/T125C/A256M, a protonated glutamic acid replaces a deprotonated aspartate. Retinal has only one counterion in these two systems, while it has two counterions in WT and in the other GR mutants. In the latter ones (excluding H87M), the deprotonated D121 is not locally neutralized by H87, which is deprotonated (i.e. neutral) at pH=7.5, according to ARM. However, the `PROPKA` H87 pK_a values are quite close to this pH value (see Figure 3(b)), with the exception of the D121N and D121E/T125C/A256M, suggesting that the protonation microstate in which H87 is protonated (i.e. charged) may be also relevant. This is confirmed by the MEM analysis and by subsequent ARM calculations, as reported in Figure 3(c): protonating H87 induces a red-shift which brings the GR excitation energies closer to the experimental values (their RMSD decreases from 0.39 eV to 0.19 eV). This improvement can be easily explained: the positive charge of protonated H87 counterbalances partially the D121 negative charge and the retinal effectively has a single counterion, similar to the D121N and D121E/T125C/A256M cases.

The ARM+MEM approach can improve the accuracy of the model by taking into account the main protonation microstates which contribute significantly (according to the MEM analysis) to the excitation energy. However, in the case of GR, the ARM+MEM excitation energies are not better than the single ARM ones when H87 is protonated. This is expected

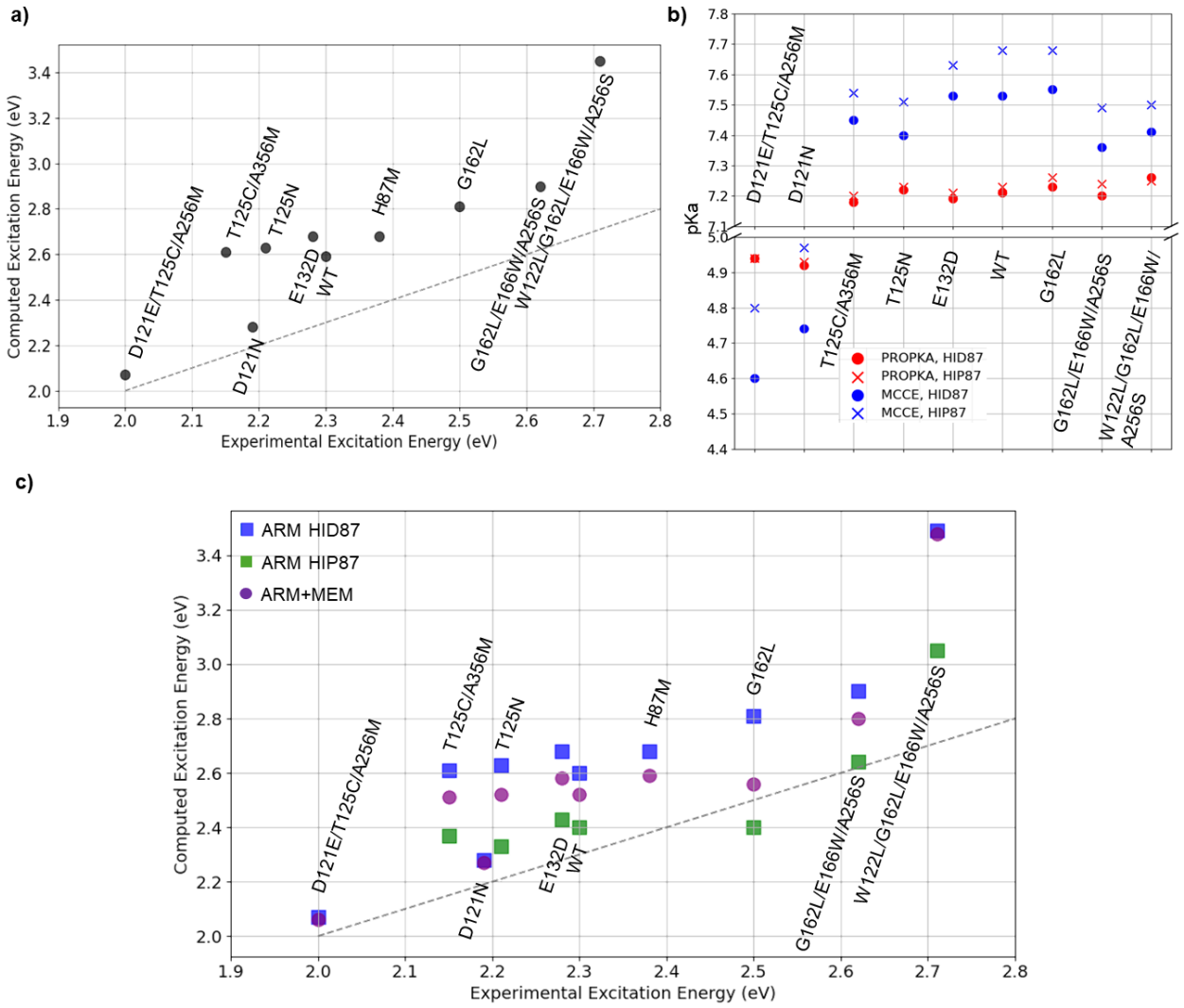


Figure 3: **a)** GR WT and mutants ARM excitation energies. Two subsets can be distinguished: the first one includes the D121E/T125C/A256M and D121N mutants which feature a single retinal counterion; the second one includes GR WT and T125C/A256M, T125N, E132D, H87M, G162L, G162L/E166W/A256S, W122L/G162L/E166W/A256S mutants in which retinal has two counterions; **b)** H87 pK_a predictions using `PROPKA` (in red) and `MCCE` (in blue) applied to two different protein structures: deprotonated H87 (circles) or protonated H87 (crosses); **c)** ARM excitation energies when H87 is deprotonated (denoted HID87, blue), is protonated (denoted HIP87, green) and ARM+MEM excitation energies (purple), compared to the experimental values. In the particular case of H87M, the ARM+MEM excitation energy results mainly from two microstates in which the D121 protonation state differs.

since, in the case of GR WT and its mutants at pH=7.5, the ARM+MEM excitation energies are mainly calculated as a weighted average between the two limiting cases: deprotonated

H87 and protonated H87. ARM with protonated H87 giving better results than ARM+MEM suggests that the H87 pK_a values may be somehow higher than the ones suggested by `PROPKA`. We have tested this hypothesis by applying an alternative pK_a prediction tool, `MCCE`, to the ARM structures for GR WT and mutants. The resulting pK_a values (Figure 3b) are found to be more structure-dependent and significantly higher than the `PROPKA` ones, some of them being even higher than 7.5, the pH value in our models, eventually indicating a predominant protonated H87 form. There is a major difference between the two tools: while `PROPKA` ignores the presence of retinal (the PSB being reduced to the usual lysine side-chain tail), `MCCE` takes into account the full retinal explicitly. This difference, as well as their different methodologies, explains why `MCCE` pK_a values are not only different from the `PROPKA` ones, but also depend more significantly on the protein structure. Previous experimental evidence supports the hypothesis of H87 protonation. In GR, the formation of a salt bridge between H87 and D121 has been shown to play a crucial role in the pH-dependent monomer-to-trimer transition.[70] Similar behaviors have been observed in other rhodopsins. In the green Proteorhodopsin, the H75-D97 cluster is suggested to be directly involved in proton transfer.[71,72] Additionally, Lanyi et al. suggested that in *Exiguobacterium sibiricum*, the H57-D85 cluster may lead to an increase in the pK_a of this histidine residue up to 9, resulting in H57 remaining protonated even under mildly basic conditions.[73] These results, if confirmed with other rhodopsins in which retinal has potentially two counterions, would suggest replacing `PROPKA` with a more advanced pK_a prediction tool.

Besides the H87 pK_a uncertainty in GR, further inspection of Figure 3(c) reveals that protonating H87 induces a shift in the retinal excitation energy ranging from virtually 0 (D121E/T125C/A256M, D121N) to more than $-0.4\,\mathrm{eV}$. This trend is surprising since the distances between H87 and the retinal PSB are always almost the same (Figure S2). In the ARM protocol, only the retinal cavity is relaxed, resulting in slight re-orientations of the residues' side-chains or water molecules inside the cavity. The case of the quadruple mutant W122L/G162L/E166W/A256S illustrates the indirect effect of H87 protonation:

while the H87–retinal distance is slightly reduced by 3% (9.8 Å → 9.5 Å), the orientation of the retinal primary counterion, D253, as well as the one of the cavity water molecule, are largely modified, eventually resulting in a large red-shift of the mutant excitation energy. Finally, the D121E/T125C/A256M mutant (in which the retinal secondary counterion is not present) shows virtually no modification of its excitation energy when H87 is protonated, despite the large reorientation of its cavity water molecule. These results evidence the interplay of several molecular factors at work when a single residue is protonated, even if it is located at distances larger than 9 Å from the rhodopsin chromophore.

The critical role of H87 can be further elucidated by considering the H87M mutant, where histidine is replaced by methionine. Initially, three different rotamers were considered for this mutant. The ARM protocol was applied to each rotamer, and the one with the lowest error relative to the experimental excitation energy was selected for further analysis. Still, the corresponding ARM excitation energy is blue-shifted by 0.3 eV, similar to the error observed for GR WT and its mutants with two counterions. The application of the MEM analysis suggests the neutralization of the secondary counterion D121 as the second most statistically significant protonation microstate (21%), with a ARM excitation energy of about 2.13 eV, which is still 0.25 eV larger than the experimental value of 2.38 eV. An important water reorientation (Figure S3), following D121 change of protonation state, rationalizes such a large excitation energy shift. Moreover, mutating H87 alters the pK_a value of D121, although not enough to change its status as the secondary counterion: the `PROPKA` pK_a value for D121 changes from 4.6 to 7.0. This is in accordance with experimental observations, where a large shift of the absorption maximum observed below pH=6 is presumed to be caused by the protonation of D121.[70] Ultimately, the combined ARM+MEM approach reduces the error to 0.2 eV, underscoring the importance of considering the two major protonation microstates in H87M at the selected pH.

Attenuating the short-range QM/MM electrostatic interactions

In the previous section, we highlighted the link between the ARM accuracy and the quality of the pK_a values, especially the H87 one, in the case of GR and its mutants. Changing the perspective and keeping the ARM protocol as it is, i.e. based on using `PROPKA` as a pK_a predictor, we can say that protonating H87 is one way, among many different ones (e.g. adding water molecules or ions, modifying the side-chain orientations of amino-acids, etc.), for modifying the external electrostatic potential the retinal is experiencing. Hereafter, we explore the possibility of directly modifying the MM electrostatic potential in QM/MM calculations, with the aim of reaching GR excitation energies as close as possible to the experimental ones.

QM/MM ESPF electrostatics are based on multi-centered charge-charge interactions. This approximation, being strictly valid at infinite distances between the QM charge density and the MM point charges, is expected to deteriorate at short distances between MM and QM particles. Accordingly, modifying the short-range QM/MM electrostatic interactions certainly changes the retinal excitation energy. To test this approach, we introduce a damping function in the usual MM electrostatic potential $\phi_a = \phi(\mathbf{R}_a; \beta)$ experienced by QM atom a:

$$\phi_a = \sum_{b \in MM} \frac{q_b}{|\mathbf{r}_b - \mathbf{R}_a|} \operatorname{erf}\left(\beta \, |\mathbf{r}_b - \mathbf{R}_a|\right) \tag{3}$$

The Gauss error function (erf) here acts as a damping function, which attenuates the short-range electrostatic interaction based on the damping parameter β. This QM/MM interaction is implemented in our local development version of the Tinker package. Note that it is only used in the final stage of the ARM protocol, specifically during the single-point CASPT2 calculations of the S_0, S_1 and S_2 energies.

We report the computed excitation energies in Figure 4 for the best β values, i.e., the ones which result in excitation energies close to the experimental ones. For comparison, Figure 4 also shows the best ARM results for each system (blue squares), that is, the ARM

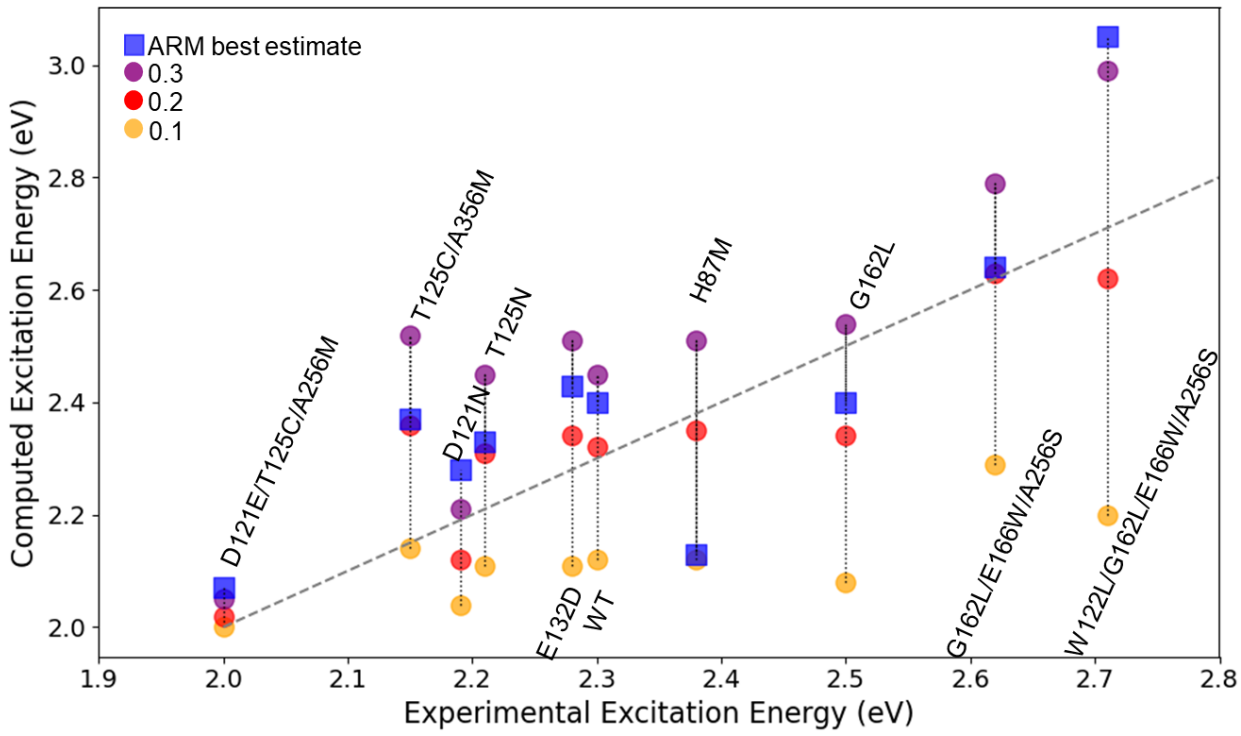


Figure 4: Comparison of ARM $S_0 \rightarrow S_1$ vertical excitation energies to experimental ones. Blue square: ARM best result, i.e. using a protonation microstate which may differ from the one suggested initially in the ARM protocol. Purple circle: scaling factor of $\beta = 0.3$ applied to the reference ARM microstate; red: $\beta = 0.2$; orange: $\beta = 0.1$.

excitation energy for the protonation microstate that gives the closest excitation energy to the experimental value. The introduction of the damping factor effectively scales down the intensity of the QM/MM electrostatic interactions. While the three β values seem to improve with respect to ARM in most cases, $\beta = 0.2$ gives the overall best results as shown by the corresponding root mean square deviations (RMSD) with respect to the experimental values: RMSD=0.113 eV ($\beta = 0.2$) to be compared to RMSD=0.234 eV ($\beta = 0.3$), and RMSD=0.260 eV ($\beta = 0.1$), and for the sake of comparison, RMSD=0.183 eV when using ARM with the protonated H87 model discussed above.

Accordingly, the attenuation of the QM/MM short-range electrostatic interactions looks like a suitable empirical improvement to the excitation energies predicted by ARM. Nevertheless, it is worth noting that using $\beta = 0.2$ induces an important decrease of the MM external potential experienced by any QM atom. For instance, ϕ_N (potential at the reti-

nal nitrogen center) is reduced by about 19% with respect to the full electrostatics of the standard ARM case. When we compare QM/MM and QM-only calculations performed on a toy model composed of retinal and two counterions (Supporting Information section 3 and Figure S4), we conclude that that the good agreement we get for the GR WT and mutants excitation energies with $\beta = 0.2$ does not primarily result from better QM/MM electrostatics. Instead, we emphasize that damping QM/MM electrostatic interactions effectively improves the cancellation of errors due to the other ARM approximations: the absence of a membrane in the GR structural model, the limited conformational sampling, etc. Accordingly, we have no real control over the damping parameter β, the value of which may change from a rhodopsin to another one.

Improving the fidelity of CASSCF/CASPT2 calculations

Having thoroughly examined and addressed the MM electrostatic potential as a possible correction to ARM outliers, we now turn our attention to the electronic structure method for describing the excited states. In rhodopsins, the retinal first excited state (S_1) is mainly of charge-transfer character, characterized by the translocation of the protonated Schiff base positive charge towards the retinal beta-ionone ring, resulting in a large increase of its dipole moment. The corresponding $S_0 \rightarrow S_1$ transition is bright and is theoretically reflected in a large transition oscillator strength. Conversely, S_2, the second excited state, has a diradical character, characterized by a dipole moment similar to the one in S_0 and a small $S_0 \rightarrow S_2$ oscillator strength. Inspection of Figure 5 top left panel shows that in 6 of the systems investigated in this study, this ordering of the retinal excited states is reversed, while in 2 of them, it is virtually impossible to decide which excited state is the absorbing one, due to state mixing. Only two systems exhibit the expected state ordering and they turn out to be the ones in which retinal has a single counterion, namely D121N and D121E/T125C/A256M. Now, if we protonate H87 in the seven other GR models (as suggested by the analysis reported above, excluding H87M), it turns out that their excited state ordering is corrected in 5 cases

out of 8 cases. This result suggests that protonating H87 is not only one way to fix the MM electrostatic potential, but it also improves the retinal electronic structure and electronic state ordering. Nevertheless, it is not improved in 3 cases: W122L/G162L/E166W/A256S in which the absorbing state is almost degenerate with the other one, G162L/E166W/A256S in which the wrong ordering is maintained, with a very large energy gap between S_1 and S_2 and, of course, in H87M since this histidine is absent.

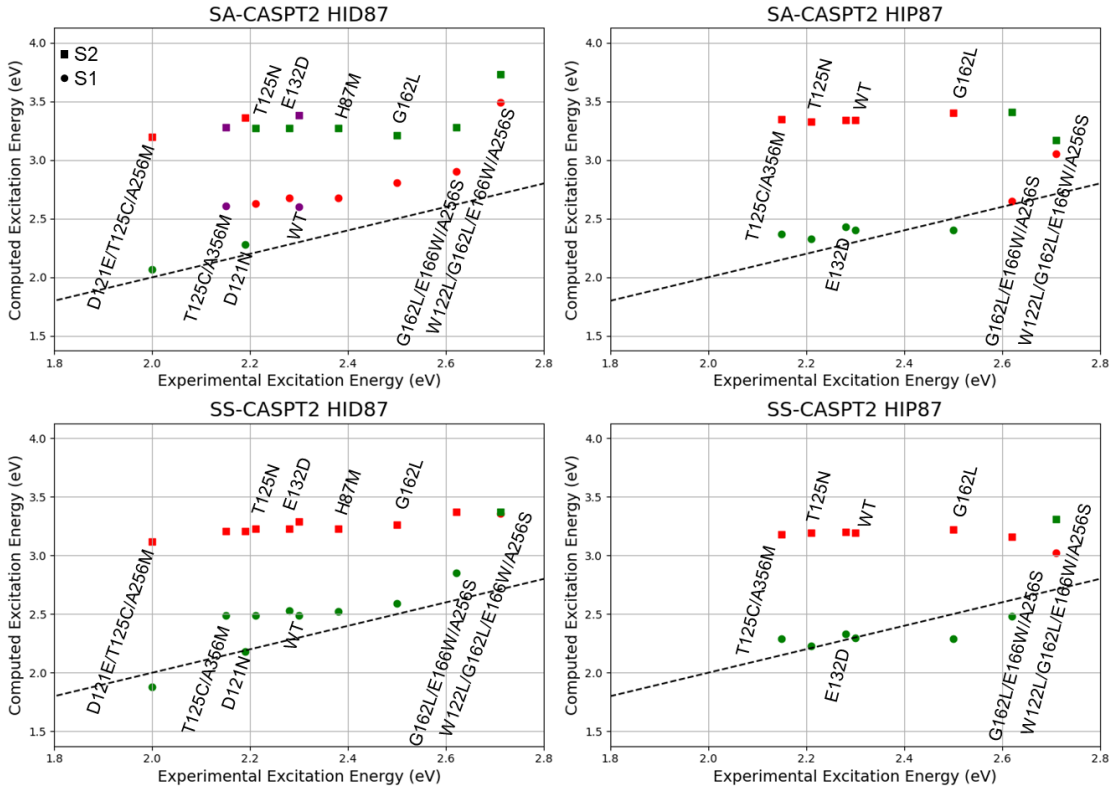


Figure 5: S_1 (circles) and S_2 (squares) for each mutant. The absorbing state with higher oscillator strength is marked in green, and the one with lower oscillator strength is marked in red. Cases where the oscillator strengths are similar are represented in purple. Top Left: Comparison of computed S_1 and S_2 states for GR and its mutants using SA-CASPT2 within the ARM protocol. Top Right: The effect of protonating H87 on the computed S_1 and S_2 states for GR and its mutants, using the ARM protocol with SA-CASPT2. Bottom Left: Computed S_1 and S_2 states using SS-CASPT2 for GR and its mutants with neutral H87, as suggested by the ARM protocol. Bottom Right: Computed S_1 and S_2 states using SS-CASPT2 for GR and its mutants with protonated H87.

The ARM protocol is based on QM/MM calculations in which the retinal electronic wavefunctions are computed at the SA-CASSCF level of theory. The latter approach features a

common set of orbitals for all the states included in the interaction state space. Accordingly, the above reported discrepancies in state ordering could be also cured if we relax this orbital constraint, i.e., using different orbitals for different states. We tested this approach at the last stage of the ARM protocol by replacing SA-CASPT2 calculations with state-specific ones, SS-CASPT2. Figure 5 bottom left panel displays the computed retinal excitation energies for GR and its mutants, keeping H87 deprotonated (neutral). With SS-CASPT2, not only is the absorbing state ordering problem resolved in all cases, but the $S_0 \rightarrow S_1$ excitation energies also show a reduced RMSD of 0.31 eV with respect to the experimental values, improving from the original 0.42 eV with SA-CASPT2 (Fig. 5 top left). The best agreement with experimental data is achieved (Figure 5, bottom right panel) when, again, H87 is protonated (but the D121N and D121E/T125C/A256M for which the H87 pK_{a} values strongly suggest to keep it deprotonated at the considered pH value). Here, the absorption energies are accurate for all mutants except W122L/G162L/E166W/A256S, and their corresponding RMSD is further reduced to 0.17 eV, i.e., the standard accuracy reported for ARM. In W122L/G162L/E166W/A256S, several hypotheses may explain the wrong excited state ordering: H87 is indeed not protonated in that particular mutant; the quality of its ARM structure is probably worse than the other ones, because it involves 4 amino-acid mutations; the inversion of S_1 and S_2 is a real property of this system; etc. Resolving this issue would require further study.

DFT-based electronic structure QM/MM method in ARM

In the previous section, we have shown that, in some circumstances, the identification of the retinal absorbing state can be difficult, especially when S_1 and S_2 are mixed, as reflected by the similar $S_0 \rightarrow S_1$ and $S_0 \rightarrow S_2$ transition oscillator strengths. Since ARM is expected to be an automatic workflow and to accelerate the discovery of mutants with specific wavelength absorption/emission in a high-throughput screening of rhodopsins, we now investigate the possibility of replacing SA-CASPT2 (or SS-CASPT2) with a different level of theory, still able

to qualitatively describe the retinal excited states and the excitation energy shifts among GR WT and mutants. Among the plethora of methods, we decide to select one which is significantly faster than CASSCF/CASPT2, namely MRSF-TDDFT, which is a highly accurate theoretical chemistry approach for describing multiconfigurational S_0, S_1 and S_2 electronic states including dynamic correlation. MRSF-TDDFT has the further advantage of eliminating the spin contamination problem inherent in SF-TDDFT methods. As any DFT-based method, it offers the flexibility of selecting different density functionals, for which it can have a strong impact on the quality of the resulting excitation energies and state characters. With this in mind, we tested several functionals by performing various QM/MM geometry optimization calculations starting with the structures obtained from the standard ARM protocol. The selection of functionals in spin-flip methodologies is restricted to the ones with a large amount of exact exchange, since they can couple the resulting spin-flip configurations. For this test of MRSF-TDDFT/MM in ARM, we thus select three functionals: CAMh-B3LYP, a version of CAM-B3LYP including a portion of long-range HF exchange and tuned for describing accurately excitation energies within linear-response TDDFT,[64] BH&HLYP,[65] and rCAM-B3LYP, which is similar to CAM-B3LYP but aims at reducing the many-electron self-interaction error in the exchange-correlation functional.[66]

Figure 6 presents a comparative study of the ARM SA-CASPT2 excitation energies, alongside those obtained with the best-performing functional, CAMh-B3LYP (results for BH&HLYP and rCAM-B3LYP are provided as Supporting Information, see Table S6 and Figure S5). In all cases, following the ARM protocol, the values shown in Fig. 6 represent the average of ten different seeds. CAMh-B3LYP qualitatively reproduces the SA-CASPT2 trend and quantitatively agrees with experimental excitation energies for the most blue-shifted mutants, namely, G162L, G162L/E166W/A256S, and W122L/G162L/E166W/A256S. In these cases, CAMh-B3LYP shows better accuracy with experimental results for both protonated and neutral H87 compared to SA-CASPT2. Additionally, the excitation energy red-shift caused by the H87 protonation is significantly smaller with MRSF-TDDFT, with an aver-

age difference being 0.25 eV for SA-CASPT2 and 0.08 eV for CAMh-B3LYP. However, it is worth noting how, except for the last three mutants, the tendency has been slightly worse, apparently converging around a value of 2.50 eV. Nevertheless, MRSF-TDDFT seems to describe the blue/red shift trends of GR mutants with respect to its wild type in a similar fashion as with the SA-CASPT2 protocol, making it a valid method for an accelerated screening of larger mutant spaces in combination with the ARM protocol. These results show the promising value of MRSF-TDDFT for describing retinal photochemistry, and specific exchange-correlation functional forms designed for such molecular systems can be tuned, as some of us showed recently.[74]

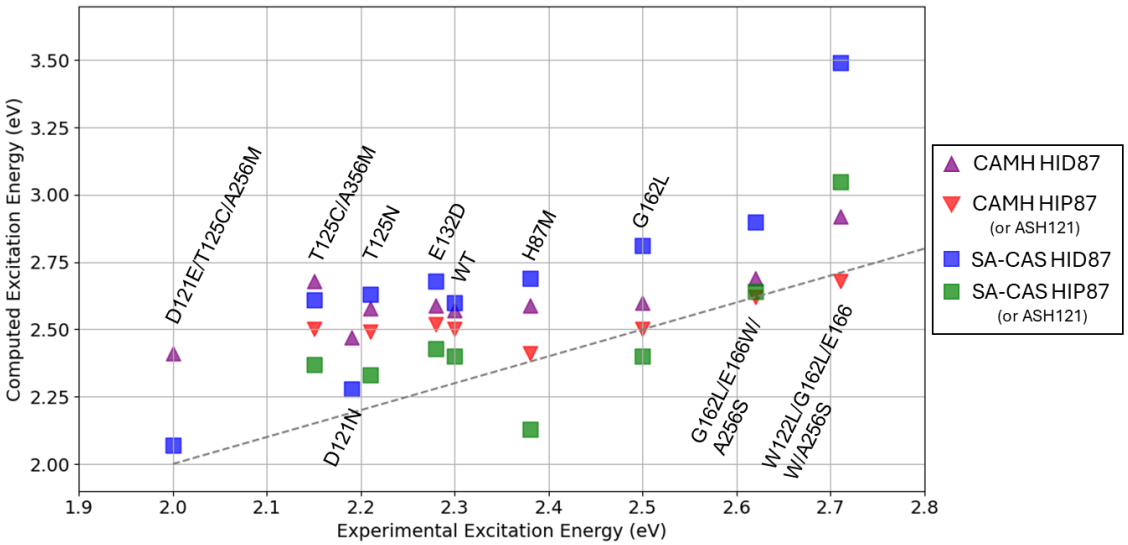


Figure 6: Comparison of AR excitation energies using SA-CASPT2 as implemented in the ARM protocol and MRSF-TDDFT using the CAMh exchange-correlation functional, for both systems with neutral H87 (HID87) or protonated H87 (HIP87). For the case of H87M, red triangles and green squares relate to the protonation state of D121. RMSD values for each method are SA-CASPT2 HID87: 0.39 eV, SA-CASPT2 HIP87: 0.19 eV, CAMh HID87: 0.39 eV and CAMh HIP87: 0.19 eV.

Conclusion

The optimization of a multi-step computational protocol is usually paved with pitfalls whose identification and resolution only come after the consideration of cases which were overlooked

when the protocol was designed. In the particular case of Massimo Olivucci's ARM protocol, meant to provide an automatic tool for studying photophysical and photochemical properties of light-activated rhodopsin proteins, the presence of two counterions close to their retinal chromophore was found to induce too large light absorption blue shifts with respect to experimental values. Besides the usual protonation of one of these two counterions, we have explored different alternative solutions which suggest further modifications of the ARM protocol: the replacement of the `PROPKA` software with a more advanced pK_a prediction tool, the systematic use of the ARM+MEM extended protocol, the scaling of the MM external potential in QM/MM calculations, the replacement of the SA-CASPT2 energy calculation with SS-CASPT2 or with the faster MRSF-TDDFT method.

Each of these suggested modifications of the ARM protocol may break the current consistency and accuracy of the ARM approach. For this reason, we will continue to explore other possible improvements, like the addition of a simple model for the membrane in which the rhodopsin is inserted, or improved identification and location of crystallographic water molecules, or enhanced sampling of the rhodopsin conformations.

CRediT

Darío Barreiro-Lage: Software, Validation, Investigation, Analysis, Writing - Original Draft, Visualization. Vincent Ledentu and Jacopo D'Ascenzi: Investigation, Visualization. Miquel Huix-Rotllant: Methodology, Writing - Original Draft, Writing - Review & Editing. Nicolas Ferré: Conceptualization, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition.

Acknowledgement

The authors thank Agence Nationale de la Recherche for grants ULTRArchea (ANR-21-CE11-0029-03) and MAPPLE (ANR-22-CE29-0014-01). Centre de Calcul Intensif d'Aix-

Marseille is acknowledged for granting access to its high-performance computing resources. Nicolas Ferré expresses his deep gratitude to Massimo Olivucci, professor at Università degli Studi di Siena, Italy and Bowling Green State University, USA, Doctor Honoris Causa of Aix Marseille University, France, for the inspiration and long-standing collaboration.

Supporting Information Available

Details of ARM+MEM excitation energy computations; Markov-chain and absorption spectrum Python scripts; optimized retinal cavity structures and superposition; toy model for the comparison of QM and QM/MM results; MRSF-TDDFT results for all exchange-correlation functionals.

References

(1) Mathies, R. A.; Lugtenburg, J. In Handb. Biol. Phys.; Stavenga, D. G., de Grip, W. J., Pugh Jr., E. N., Eds.; Elsevier Science B.V., 2000; Chapter The primary photoreaction of rhodopsin, pp 55–90.

(2) Okada, T.; Ernst, O. P.; Palczewski, K.; Hofmann, K. P. Activation of rhodopsin: new insights from structural and biochemical studies. Trends Biochem. Sci. **2001**, 26, 318–324.

(3) 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.

(4) Warshel, A. Bicycle-pedal model for the first step in the vision process. Nature **1976**, 260, 676–683.

(5) Warshel, A.; Levitt, M. Theoretical studies of enzyme reactions: dielectric, electrostatic

and steric stabilisation of the carbonium ion in the reaction of lysozyme. J. Mol. Biol. **1976**, 103, 227–249.

(6) Lin, H.; Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? Theor. Chem. Acc. **2006**, 117, 185–199.

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

(8) Meier, K.; Thiel, W.; van Gunsteren, W. F. On the effect of a variation of the force field, spatial boundary condition and size of the QM region in QM/MM MD simulations. J. Comp. Chem. **2012**, 33, 363–378.

(9) Monari, A.; Rivail, J.-L.; Assfeld, X. Theoretical Modeling of Large Molecular Systems. Advances in the Local Self Consistent Field Method for Mixed Quantum Mechanics/Molecular Mechanics Calculations. Acc. Chem. Res. **2013**, 46, 596–603.

(10) Mroginski, M. A. Encyclopedia of Biophysics; Springer Berlin Heidelberg, 2013; pp 2149–2154.

(11) Pan, X.; Rosta, E.; Shao, Y. Representation of the QM Subsystem for Long-Range Electrostatic Interaction in Non-Periodic Ab Initio QM/MM Calculations. Molecules **2018**, 23, 2500.

(12) Olsen, J. M. H.; Bolnykh, V.; Meloni, S.; Ippoliti, E.; Bircher, M. P.; Carloni, P.; Rothlisberger, U. MiMiC: A Novel Framework for Multiscale Modeling in Computational Chemistry. J. Chem. Theory Comput. **2019**, 15, 3810–3823.

(13) Loco, D.; Lagardère, L.; Adjoua, O.; Piquemal, J.-P. Atomistic Polarizable Embeddings: Energy, Dynamics, Spectroscopy, and Photoreactivity. Acc. Chem. Res. **2021**, 54, 2812–2822.

(14) Nottoli, M.; Bondanza, M.; Mazzeo, P.; Cupellini, L.; Curutchet, C.; Loco, D.; Lagardère, L.; Piquemal, J.-P.; Mennucci, B.; Lipparini, F. QM/AMOEBA description of properties and dynamics of embedded molecules. WIREs Comput. Mol. Sci. **2023**, 13, e1674.

(15) Hayashi, S.; Ohmine, I. Proton transfer in bacteriorhodopsin: structure, excitation, IR spectra, and potential energy surface analyses by an ab initio QM/MM method. J. Phys. Chem. B **2000**, 104, 10678–10691.

(16) 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.

(17) Gascón, J. A.; Sproviero, E. M.; Batista, V. S. Computational Studies of the Primary Phototransduction Event in Visual Rhodopsin. Acc. Chem. Res. **2006**, 39, 184–193.

(18) Sekharan, S.; Altun, A.; Morokuma, K. QM/MM Study of Dehydro and Dihydro β-Ionone Retinal Analogues in Squid and Bovine Rhodopsins: Implications for Vision in Salamander Rhodopsin. J. Am. Chem. Soc. **2010**, 132, 15856–15859.

(19) Gozem, S.; Schapiro, I.; Ferré, N.; Olivucci, M. The Molecular Mechanism of Thermal Noise in Rod Photoreceptors. Science **2012**, 337, 1225–1228.

(20) Campomanes, P.; Neri, M.; Horta, B. A. C.; Röhrig, U. F.; Vanni, S.; Tavernelli, I.; Rothlisberger, U. Origin of the Spectral Shifts among the Early Intermediates of the Rhodopsin Photocycle. J. Am. Chem. Soc. **2014**, 136, 3842–3851.

(21) Caprasecca, S.; Jurinovich, S.; Viani, L.; Curutchet, C.; Mennucci, B. Geometry Optimization in Polarizable QM/MM Models: The Induced Dipole Formulation. J. Chem. Theory Comput. **2014**, 10, 1588–1598.

(22) Gozem, S.; Luk, H. L.; Schapiro, I.; Olivucci, M. Theory and Simulation of the Ultrafast Double-Bond Isomerization of Biological Chromophores. Chem. Rev. **2017**, 117, 13502–13565.

(23) Dokukina, I.; Nenov, A.; Garavelli, M.; Marian, C. M.; Weingart, O. QM/MM Photodynamics of Retinal in the Channelrhodopsin Chimera C1C2 with OM3/MRCI. ChemPhotoChem **2019**, 3, 107–116.

(24) Punwong, C.; Hannongbua, S.; Martínez, T. J. Electrostatic Influence on Photoisomerization in Bacteriorhodopsin and Halorhodopsin. J. Phys. Chem. B **2019**, 123, 4850–4857.

(25) Valsson, O.; Filippi, C. Photoisomerization of Model Retinal Chromophores: Insight from Quantum Monte Carlo and Multiconfigurational Perturbation Theory. J. Chem. Theory Comput. **2010**, 6, 1275–1292.

(26) Guareschi, R.; Valsson, O.; Curutchet, C.; Mennucci, B.; Filippi, C. Electrostatic versus Resonance Interactions in Photoreceptor Proteins: The Case of Rhodopsin. J. Phys. Chem. Lett. **2016**, 7, 4547–4553.

(27) Valsson, O.; Campomanes, P.; Tavernelli, I.; Rothlisberger, U.; Filippi, C. Rhodopsin Absorption from First Principles: Bypassing Common Pitfalls. J. Chem. Theory Comput. **2013**, 9, 2441–2454.

(28) Hermosilla, L.; Prampolini, G.; Calle, P.; de la Vega, J. M. G.; Brancato, G.; Barone, V. Extension of the AMBER Force Field for Nitroxide Radicals and Combined QM/MM/PCM Approach to the Accurate Determination of EPR Parameters of DMPO-H in Solution. J. Chem. Theory Comput. **2013**, 9, 3626–3636.

(29) Nottoli, M.; Lipparini, F. General formulation of polarizable embedding models and of their coupling. J. Chem. Phys. **2020**, 153, 224108.

(30) Kairys, V.; Jensen, J. H. QM/MM boudaries across covalent bonds: a frozen localized molecular orbital-based approach for the effective fragment potential method. J. Phys. Chem. A **2000**, 104, 6656–6665.

(31) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. Frontier bonds in QM/MM methods: a comparison of different approaches. J. Chem. Phys. **2000**, 104, 1720–1735.

(32) Ferré, N.; Olivucci, M. The peptide bond: pitfalls and drawbacks of the link atom scheme. J. Molec. Struct. (Theochem) **2003**, 632, 71–82.

(33) Loos, P.-F.; Assfeld, X. Self-Consistent Strictly Localized Orbitals. J. Chem. Theory Comput. **2007**, 3, 1047–1053.

(34) 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.

(35) Pedraza-González, L.; Vico, L. D.; del Carmen Marín, M.; 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.

(36) Pedraza-González, L.; Marín, M. D. C.; Jorge, A. N.; Ruck, T. D.; Yang, X.; Valentini, A.; Olivucci, M.; Vico, L. D. Web-ARM: A Web-Based Interface for the Automatic Construction of QM/MM Models of Rhodopsins. J. Chem. Inf. Model. **2020**, 60, 1481–1493.

(37) Pedraza-González, L.; Barneschi, L.; Padula, D.; De Vico, L.; Olivucci, M. Evolution of the Automatic Rhodopsin Modeling (ARM) Protocol. Top. Curr. Chem. **2022**, 380, 21.

(38) 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.

(39) Roos, B. O.; Taylor, P. R.; Siegbahn, P. E. M. A Complete Active Space SCF Method (CASSCF) Using a Density-matrix Formulated Super-CI Approach. Chem. Phys. **1980**, 48, 157–173.

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

(41) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A.; Wolinski, K. Second-Order Perturbation Theory with a CASSCF Reference Function. J. Phys. Chem. **1990**, 94, 5483.

(42) 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.

(43) 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.

(44) Cárdenas, G.; Ledentu, V.; Huix-Rotllant, M.; Olivucci, M.; Ferré, N. Automatic Rhodopsin Modeling with Multiple Protonation Microstates. J. Phys. Chem. A **2023**, 127, 9365–9380.

(45) Lee, S.; Filatov, M.; Lee, S.; Choi, C. H. Eliminating spin-contamination of spin-flip time dependent density functional theory within linear response formalism by the use of zeroth-order mixed-reference (MR) reduced density matrix. J. Chem. Phys. **2018**, 149, 104101.

(46) Huix-Rotllant, M.; Schwinn, K.; Pomogaev, V.; Farmani, M.; Ferré, N.; Lee, S.; Choi, C. H. Photochemistry of Thymine in Solution and DNA Revealed by an Electrostatic Embedding QM/MM Combined with Mixed-Reference Spin-Flip TDDFT. J. Chem. Theory Comput. **2023**, 19, 147–156.

(47) Choi, A. R.; Shi, L.; Brown, L. S.; Jung, K.-H. Cyanobacterial Light-Driven Proton Pump, Gloeobacter Rhodopsin: Complementarity between Rhodopsin-Based Energy Production and Photosynthesis. PLOS ONE **2014**, 9, e110643.

(48) Jana, S.; Eliash, T.; Jung, K.-H.; Sheves, M. Retinal Binding to Apo-Gloeobacter Rhodopsin: The Role of pH and Retinal-Carotenoid Interaction. J. Phys. Chem. B **2017**, 121, 10759–10769.

(49) Engqvist, M. K. M.; 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.

(50) Galván, I. F.; 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.

(51) Aquilante, F.; Autschbach, J.; Baiardi, A.; Battaglia, S.; Borin, V. A.; Chibotaru, L. F.; Conti, I.; Vico, L. D.; Delcey, M.; Galván, I. F. et al. Modern quantum chemistry with [Open]Molcas. J. Chem. Phys. **2020**, 152, 214117.

(52) 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.

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

(54) Melaccio, F.; Olivucci, M.; Lindh, R.; Ferré, N. Unique QM/MM potential energy surface exploration using microiterations. Int. J. Quantum Chem. **2011**, 111, 3339–3346.

(55) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem. **2000**, 21, 1049–1074.

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

(57) Giuliani, G.; Melaccio, F.; Gozem, S.; Cappelli, A.; Olivucci, M. QM/MM Investigation of the Spectroscopic Properties of the Fluorophore of Bacterial Luciferase. J. Chem. Theory Comput. **2021**, 17, 605–613.

(58) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M.; Molina, V. Applications of the level shift corrected perturbation theory in electronic spectroscopy. J. Molec. Struct. (Theochem) **1996**, 388, 257–276.

(59) Segado, M.; Gómez, I.; Reguero, M. Intramolecular charge transfer in aminobenzonitriles and tetrafluoro counterparts: fluorescence explained by competition between low-lying excited states and radiationless deactivation. Part I: A mechanistic overview of the parent system ABN. Phys. Chem. Chem. Phys. **2016**, 18, 6861–6874.

(60) Tran, L. N.; Neuscamman, E. Improving Excited-State Potential Energy Surfaces via Optimal Orbital Shapes. J. Phys. Chem. A **2020**, 124, 8273–8279.

(61) Lee, S.; Kim, E. E.; Nakata, H.; Lee, S.; Choi, C. H. Efficient implementations of analytic energy gradient for mixed-reference spin-flip time-dependent density functional theory (MRSF-TDDFT). J. Chem. Phys. **2019**, 150, 184111.

(62) Park, W.; Komarov, K.; Lee, S.; Choi, C. H. Mixed-Reference Spin-Flip Time-Dependent Density Functional Theory: Multireference Advantages with the Practicality of Linear Response Theory. J. Phys. Chem. Lett. **2023**, 14, 8896–8908.

(63) Barca, G. M. J.; Bertoni, C.; Carrington, L.; Datta, D.; De Silva, N.; Deustua, J. E.; Fedorov, D. G.; Gour, J. R.; Gunina, A. O.; Guidez, E. et al. Recent developments in the general atomic and molecular electronic structure system. J. Chem. Phys. **2020**, 152, 154102.

(64) Shao, Y.; Mei, Y.; Sundholm, D.; Kaila, V. Benchmarking the performance of time-dependent density functional theory methods on biochromophores. J. Chem. Theory Comput. **2020**, 16, 587–600.

(65) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. J. Chem. Phys. **1993**, 98, 1372–1377.

(66) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Development of exchange-correlation functionals with minimal many-electron self-interaction error. J. Chem. Phys. **2007**, 126, 191109.

(67) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide protein data bank. Nat. Struct. Mol. Biol. **2003**, 10, 980.

(68) Morizumi, T.; Ou, W.-L.; Van Eps, N.; Inoue, K.; Kandori, H.; Brown, L. S.; Ernst, O. P. X-ray crystallographic structure and oligomerization of Gloeobacter rhodopsin. Sci. Rep. **2019**, 9, 11283.

(69) Vogt, A.; Wietek, J.; Hegemann, P. Gloeobacter Rhodopsin, Limitation of Proton Pumping at High Electrochemical Load. Biophys. J. **2013**, 105, 2055–2063.

(70) Tsukamoto, T.; Kikukawa, T.; Kurata, T.; Jung, K.-H.; Kamo, N.; Demura, M. Salt

bridge in the conserved His–Asp cluster in Gloeobacter rhodopsin contributes to trimer formation. FEBS Letters **2013**, 587, 322–327.

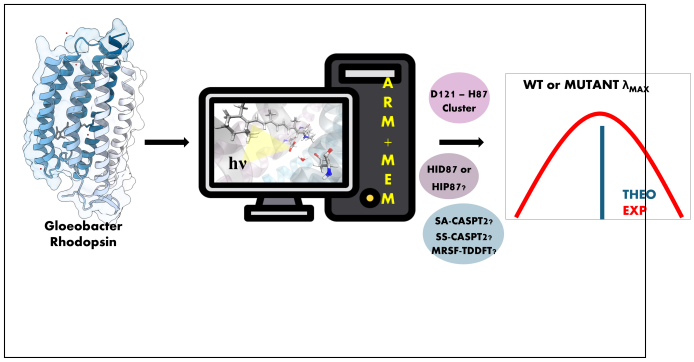
(71) Bergo, V. B.; Kralj, J. M.; Spudich, J. L.; Rothschild, K. J. His75 in proteorhodopsin, a novel component in light-driven proton translocation by primary pumps. Biophys. J. **2009**, 96, 526a.

(72) Hempelmann, F.; Hölper, S.; Verhoefen, M.-K.; Woerner, A. C.; Köhler, T.; Fiedler, S.-A.; Pfleger, N.; Wachtveitl, J.; Glaubitz, C. His75-Asp97 Cluster in Green Proteorhodopsin. J. Am. Chem. Soc. **2011**, 133, 4645–4654, PMID: 21366243.

(73) Balashov, S.; Petrovskaya, L.; Lukashev, E.; Imasheva, E.; Dioumaev, A.; Wang, J.; Sychev, S.; Dolgikh, D.; Rubin, A.; Kirpichnikov, M. et al. Aspartate–Histidine interaction in the retinal Schiff base counterion of the light-driven proton pump of Exiguobacterium sibiricum. Biochemistry **2012**, 51, 5748–5762.

(74) Komarov, K.; Park, W.; Lee, S.; Huix-Rotllant, M.; Choi, C. H. Doubly Tuned Exchange–Correlation Functionals for Mixed-Reference Spin-Flip Time-Dependent Density Functional Theory. J. Chem. Theory Comput. **2023**, 19, 7671–7684.

TOC Graphic


Gloeobacter
Rhodopsin
hν
A
R
M
+
M
E
M
D121 – H87
Cluster
HID87 or
HIP87?
SA-CASPT2?
SS-CASPT2?
MRSF-TDDFT?
WT or MUTANT λMAX
THEO
EXP

Supporting Information: Investigating the Origin of Automatic Rhodopsin Modeling Outliers Using the Microbial Gloeobacter Rhodopsin as Testbed

Darío Barreiro-Lage,[†] Vincent Ledentu,[†] Jacopo D'Ascenzi,[‡,¶] Miquel Huix-Rotllant,[*,†] and Nicolas Ferré[*,†]

†*Aix Marseille Univ, CNRS, ICR, 13013 Marseille, France*

‡*Dipartimento di Biotecnologie, Chimica e Farmacia, Università degli Studi di Siena, 53100 Siena, Italy*

¶*Dipartimento di Chimica, Biologia e Biotecnologie, Università degli studi di Perugia, 06123 Perugia, Italy*

E-mail: miquel.huix-rotllant@cnrs.fr; nicolas.ferre@univ-amu.fr

Calculating ARM+MEM excitation energies

In this section, we explain how to obtain a single excitation energy value at the end of the ARM+MEM protocol. As a reminder, ARM calculates the QM/MM retinal excitation energy of 10 molecular models of the same rhodopsin in a given protonation microstate. Practically, they correspond to 10 different "SEEDS", i.e. random numbers which are used to generate initial velocities during the molecular dynamics step in the ARM workflow.

Table S1: Seeds HID87 Data (Part 1)

D121E/T125C/A256M	T125C/A256M	D121N	T125N	E132D
13308	17726	15765	19982	19641
22222	20991	20837	21345	22222
22322	30339	28640	23412	25926
23322	42561	35204	43212	30997
32333	50023	36040	49321	48844
41438	60991	40614	59767	52064
50541	64824	57686	67029	59797
56405	74244	69920	83611	63290
57078	77161	90324	89123	83661
63191	80246	99931	90432	99789

Table S2: Seeds HID87 Data (Part 2)

WT	H87M	G162L	G162L/E166W/A256S	W122L/G162L/E166W/A256S
29521	14171	11646	14318	13478
29963	33934	19465	22446	14234
32034	37060	31525	24332	15023
34723	47274	32035	45675	28755
35199	47615	33333	58175	38832
36489	50573	47688	66632	41196
38129	55549	65675	79213	54001
61398	84237	66385	86676	77691
70908	85821	95376	88448	82494
97784	92925	99624	96632	93980

Table S3: Seeds HIP87 Data (Part 1)

D121E/T125C/A256M	T125C/A256M	D121N	T125N	E132D
-	14175	-	11232	16259
-	18212	-	21982	18204
-	23312	-	30087	24026
-	25308	-	35683	41085
-	42054	-	36184	46012
-	53047	-	38262	46938
-	66586	-	46964	60577
-	73953	-	55821	60577
-	84035	-	77213	65743
-	86922	-	79212	74293

Table S4: Seeds HIP87 Data (Part 2)

WT	G162L	G162L/E166W/A256S	W122L/G162L/E166W/A256S
10223	19047	50844	11182
10831	29596	63550	16542
27821	31821	64014	19882
30428	36180	66443	22420
38173	37913	74230	24232
52134	53946	79307	27682
57405	58105	84207	45625
69065	86534	87452	51805
70065	91993	91141	53637
80857	99331	96901	56019

Table S5: Seeds H87M ASH121 Data

H87M
16026
28329
40463
49503
64789
66283
79940
81517
82036
97985

Accordingly, a given ARM model is characterized by an excitation energy and oscillator strength averaged over the 10 seeds. Then, a MEM analysis is performed to evaluate which other protonation microstates are populated at a given pH. Following the ARM+MEM philosophy, the `pyARM` program automatically performs subsequent ARM+MEM calculations on these new microstates until the protonation microstate space has been sufficiently explored. Eventually, the application of the ARM+MEM protocol results in a series of excitation energies, transition oscillator strengths and probabilities for each protonation microstate. We now present the simple method we use for transforming these results into an absorption spectrum.

Markov chains allow to determine event probabilities between different resting states. In the current work, these states are rhodopsin protonation microstates while events are protonation or deprotonation of a single rhodopsin residue. By constructing a transition matrix, which encapsulates the probabilities of moving from one microstate to another, we can easily calculate the steady-state distribution of these microstates, i.e. the probability of the system to be in each considered microstate. This is particularly useful because it allows to account for both direct transitions between microstates and the "self" transitions where the system remains in the same microstate. Thus, Markov chains provide a comprehensive way to determine each protonation microstate probability. Eventually, we can calculate the intensity of each rhodopsin electronic transition as the product of the Markov chains-based microstate probability with the corresponding averaged oscillator strength.

Using these single transition intensities and excitation energies, we can build the rhodopsin absorption spectrum by convoluting each transition with a Gaussian function per electronic transition. It is important to select a Gaussian width that correctly encompasses all different transitions; in our case, a width of 0.2 was used. Finally, when the spectrum is produced, the transition energy at its maximum intensity is considered as the ARM+MEM excitation energy.

Markov Chains program

Below is reported the `python` code used for applying the Markov Chains algorithm to protonation microstate probabilities.

```
import numpy as np
import sys

# MARKOV CHAINS FOR THE CALCULATION OF THE GLOBAL PROBABILITIES
# OF SEVERAL DIFFERENT INTERCONNECTED MICROSTATES.

# The state space for a case where only 2 protonation microstates are relevant
# for the excitation energy
states = ["A", "B"]

# Probabilities matrix (transition matrix)
# Please change each label for your MEM probabilities
ProbM = np.array([[AA, AB],
                  [BA, BB]])

# Sum of each row
row_sums = np.sum(ProbM, axis=1)

# Define a tolerance for equality (adjust as needed)
tolerance = 1e-6

# Check if the sum of each row is close enough to 1
is_row_sum_one = np.allclose(row_sums, 1, rtol=tolerance)
```

```
if not np.all(is_row_sum_one):
    print("---------------------------------------")
    print("Not all rows of the Probability matrix sum to 1. Normalizing...")
    print("---------------------------------------")
    # Normalize each row to sum to 1
    ProbM = ProbM / row_sums[:, np.newaxis]
    # Recalculate row sums to verify normalization
    row_sums = np.sum(ProbM, axis=1)
    is_row_sum_one = np.allclose(row_sums, 1, rtol=tolerance)
    if np.all(is_row_sum_one):
        print("Normalization successful. Probability matrix well defined.")
    else:
        print("Normalization failed. Please check the input matrix.")
        sys.exit(1)
else:
    print("---------------------------------------")
    print("The Probability matrix is well defined")
    print("---------------------------------------")

# Obtain eigenvalues and eigenvectors
# We need to find a vector a such that a*ProbM = a
eigenvalues, eigenvectors = np.linalg.eig(ProbM.T)

index = np.where(np.isclose(eigenvalues, 1))[0][0]

stationary_distribution = eigenvectors[:, index].real
stationary_distribution /= stationary_distribution.sum()
```

```
print("--------------------------------------")
print("Transition Probability Matrix:")
print(ProbM)
print("--------------------------------------")
print("\nStationary Distribution a: A, B")
print(stationary_distribution)
print("--------------------------------------")
```

Absorption spectrum plotting program

Below is given a simple `python` program to plot a ARM+MEM-based absorption spectrum. This program needs two input files, 'ener.inp' and 'inten.inp'. In the first one, the average ARM energies are listed. In the second one, the corresponding intensities are listed.

1. **ener.inp:**

```
#ener.inp
X
#Being X the number of states
a
b
...
#Being the list of the energies in eV
```

2. **inten.inp:**

```
 #inten.inp
X
#Being X the number of states
```

```
    a
    b
    ...
    #Being the list of the intensities
```

3. Python **program source code:**

```
import numpy as np
import matplotlib.pyplot as plt

def normalize_data(data):
    data_min = min(data)
    data_max = max(data)
    normalized_data = [(x - data_min) / (data_max - data_min) for x in data]
    return normalized_data

# Read input data from files
with open('ener.inp', 'r') as xps_file:
    npeaks = int(xps_file.readline())
    peak = [float(xps_file.readline()) for _ in range(npeaks)]

with open('inten.inp', 'r') as inten_file:
    ndata = int(inten_file.readline())
    intens = [float(inten_file.readline()) for _ in range(ndata)]

# Creating the gaussian
nener = 28000 #This should be modified for each case
deltaener = 0.00005
ener0 = 2.0 #This should be modified for each case
```

```
enerfin = ener0 + nener * deltaener
width = 0.2 #Be sure the width is big enough to englobe all peaks

# Create arrays to store energy and spectrum data
energies = np.linspace(ener0, enerfin, nener)
spectrum = np.zeros(nener)

# Calculate the spectrum
for j in range(nener):
    ener = ener0 + deltaener * j
    for i in range(npeaks):
        gauss = intens[i] * np.exp(-(ener - peak[i]) ** 2 / (2 * width ** 2))
        spectrum[j] += gauss

# Normalize spectrum using the defined function
spectrum_normalized = normalize_data(spectrum)

# Plot the spectrum
plt.plot(energies, spectrum_normalized)
plt.xlabel('Energy (eV)')
plt.ylabel('Normalized Intensity (a.u.)')
plt.title('Normalized Spectrum')

# Find maximum intensity and its corresponding energy
max_intensity = max(spectrum_normalized)
max_energy_index = np.argmax(spectrum_normalized)
c_energy = energies[max_energy_index]
```

```
# Plot a bar at the maximum intensity
plt.axvline(x=energies[max_energy_index], color='red', linestyle='--')
plt.text(energies[max_energy_index], 0, f'Energy: {c_energy:.2f}', ha='center')
#plt.show()

print("Corresponding Energy:", corresponding_energy)
plt.ylim(0, 1.1)
```

Absorption spectra

The spectra obtained with the above-mentioned tools are reported below for each Gloeobacter rhodopsin model (wild type and nine mutants) considered in the study.

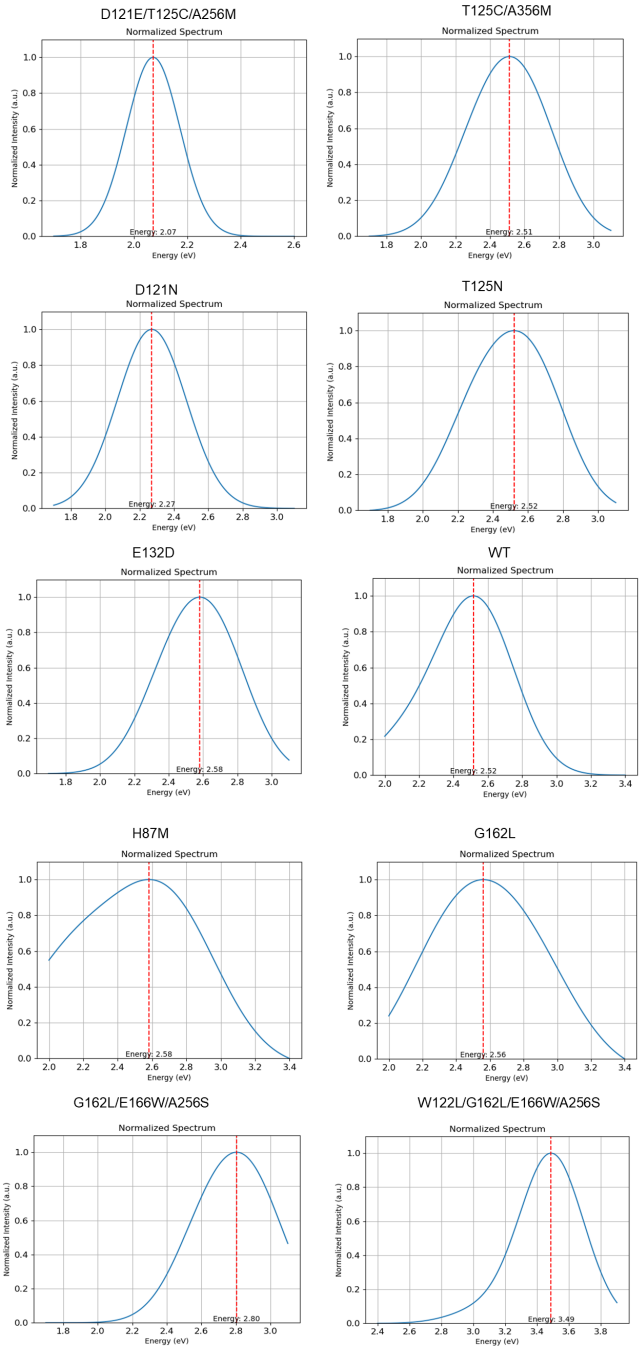


Figure S1: ARM+MEM absorption spectra of 10 Gloeobacter rhodopsin systems. Spectra has been obtained using the programs and protocol described in this section.

Cavities

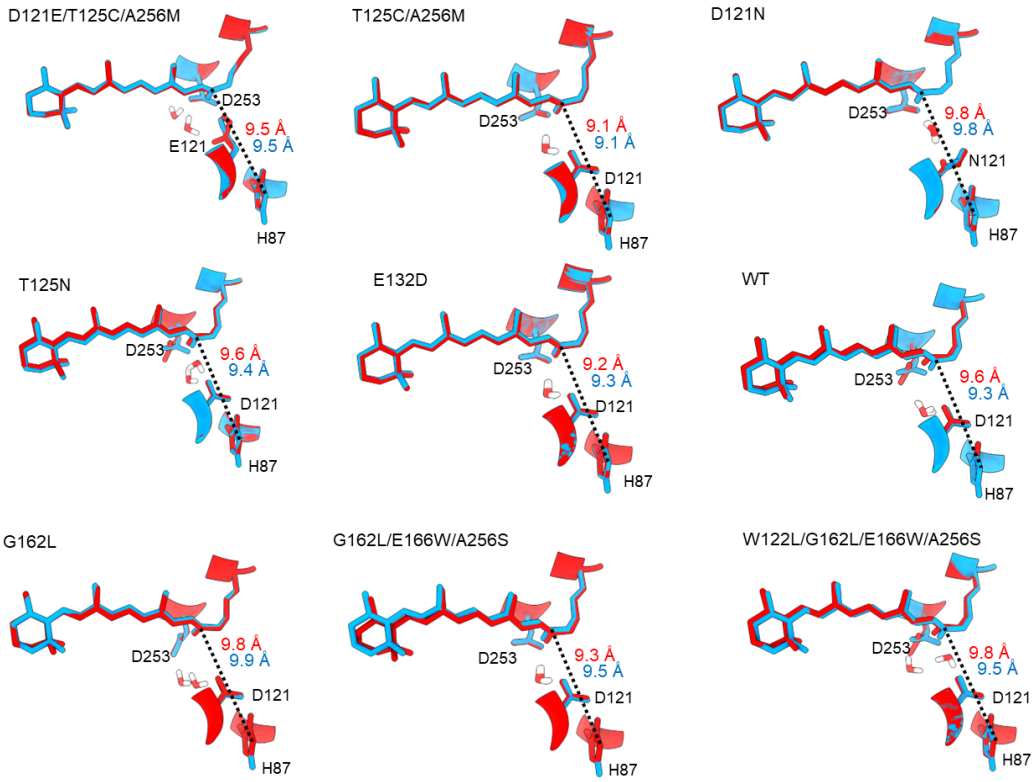


Figure S2: Superposition of the final ARM structures (wild type Gloeobacter Rhodopsin and 8 mutants), highlighting the retinal chromophore, the residues D253, D121, and H87 and the cavity water molecule HOH401. Structures with deprotonated H87 are shown in red, while those with protonated H87 are shown in blue. Distances between the retinal N center and the H87 Cϵ center are also reported.

Figure S3: Superposition of the final ARM structures for H87M, highlighting the retinal chromphore, the residues D253, D121, and M87 and the cavity water molecule HOH401. Structures with deprotonated D121 are shown in red, while those with protonated D121 are shown in green. Distances between the retinal N center and the M87 S center are also reported.

Toy model

In the main text, we find that while the damping factor $\beta = 0.2$ results in the best agreement with the experimental excitation energies, it also significantly reduces the MM electrostatic potential experienced by the QM retinal. Accordingly, we test the validity of such an empirical QM/MM approach against QM-only calculations, considering a toy model composed of retinal and two formiate anions, mimicking the two counterions case met in WT GR. In Figure S4, we report the vertical excitation energy of this toy model using SA-CASPT2/MM as it is used in ARM, attenuated SA-CASPT2/MM with different damping factors β and SA-CASPT2 for the whole complex system (while ensuring the consistency of the SA-CASSCF active space, always comprising 12 electrons in 12 π orbitals mainly located on the retinal moiety).

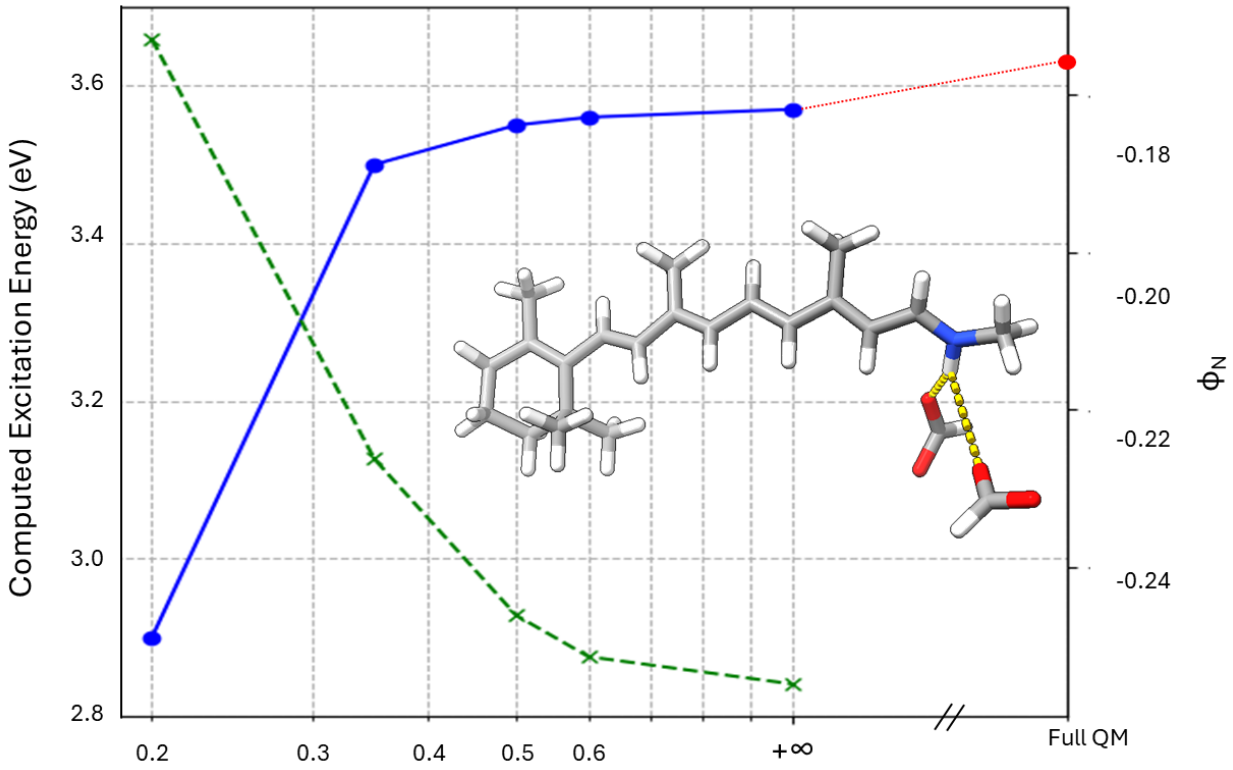


Figure S4: Retinal excitation energy of retinal (left y-axis) in the presence of two counterions (distances in yellow are 1.74 and 4.04 Å, similar to the GR WT case). Blue: QM/MM values with different damping factors; red: full QM value. Green (right y-axis): MM electrostatic potential at the retinal N center.

On one hand, we can immediately realize that the CASPT2/MM excitation energy ($\beta = +\infty$) is very close to the QM-only CASPT2 one. This result is not surprising since the CASPT2/SA-CASSCF level of theory using the 6-31G* basis set has been used for several decades to study retinal proteins, thanks to an excellent error cancellation. On the other hand, when the QM/MM electrostatic interactions are attenuated using $\beta = 0.2$, the electrostatic potential experienced by the retinal N center is reduced by about 30% while the retinal excitation energy is severely reduced by about 17%.

MRSF-TDDFT

Below are reported the ARM excitation energies using MRSF-TDDFT instead of SA-CASPT2.

Table S6: MRSF-TDDFT excitation energies (in eV)

GR system	exp	CAMh HID87	CAMh HIP87	BHHLYP HID87	BHHLYP HIP87	RCAM HID87
D121E/T125C/A256M	2.00	2.41	2.38	2.56	2.53	3.12
T125C/A256M	2.15	2.58	2.51	2.82	2.70	3.39
D121N	2.19	2.48	2.41	2.66	2.56	3.23
T125N	2.21	2.58	2.48	2.81	2.68	3.40
E132D	2.28	2.57	2.53	2.80	2.70	3.39
WT	2.30	2.57	2.51	2.78	2.70	3.36
G162L	2.50	2.62	2.49	2.82	2.68	3.41
G162L/E166W/A256S	2.62	2.73	2.57	2.99	2.79	3.73
W122L/G162L/E166W/A256S	2.71	2.78	2.68	3.06	2.93	3.78

In the case of the H87M mutant, the experimental excitation energy is 2.38 eV. Its MRSF-TDDFT excitation energy using CAMh for the most probable protonation microstate is 2.59 eV, while it drecreases to 2.41 eV when protonated.

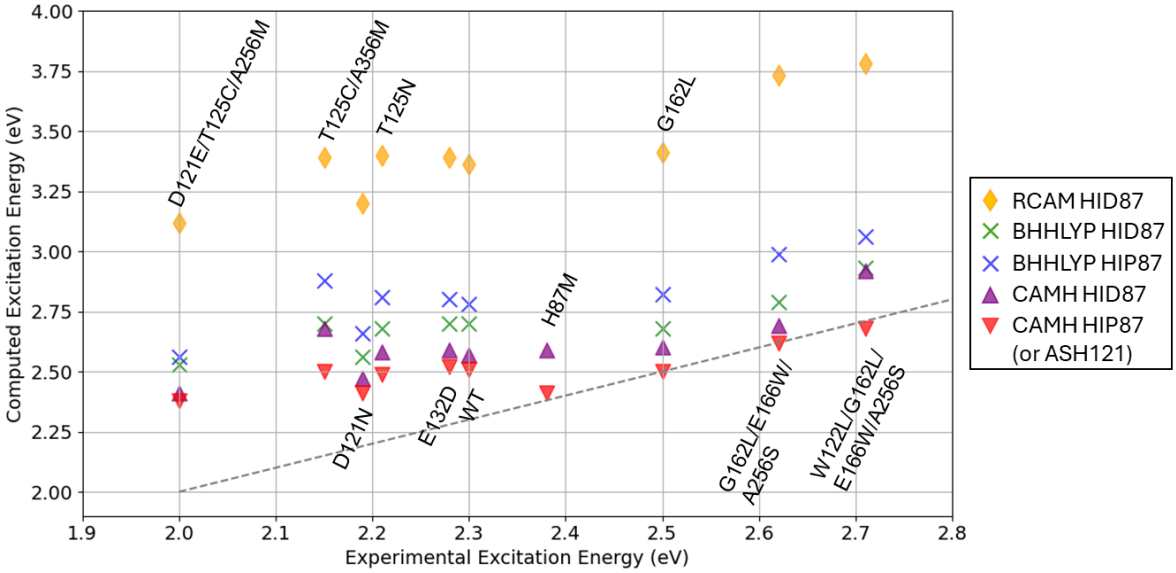


Figure S5: Comparison of GR excitation energies using MRSF-TDDFT and either CAMh, or MRSF/BHHLYP or MRSF/rCAM exchange-correlation functionals. CAMh is proven to be the best-performing functional. In the case of H87M, the red triangle corresponds to protonated D121.