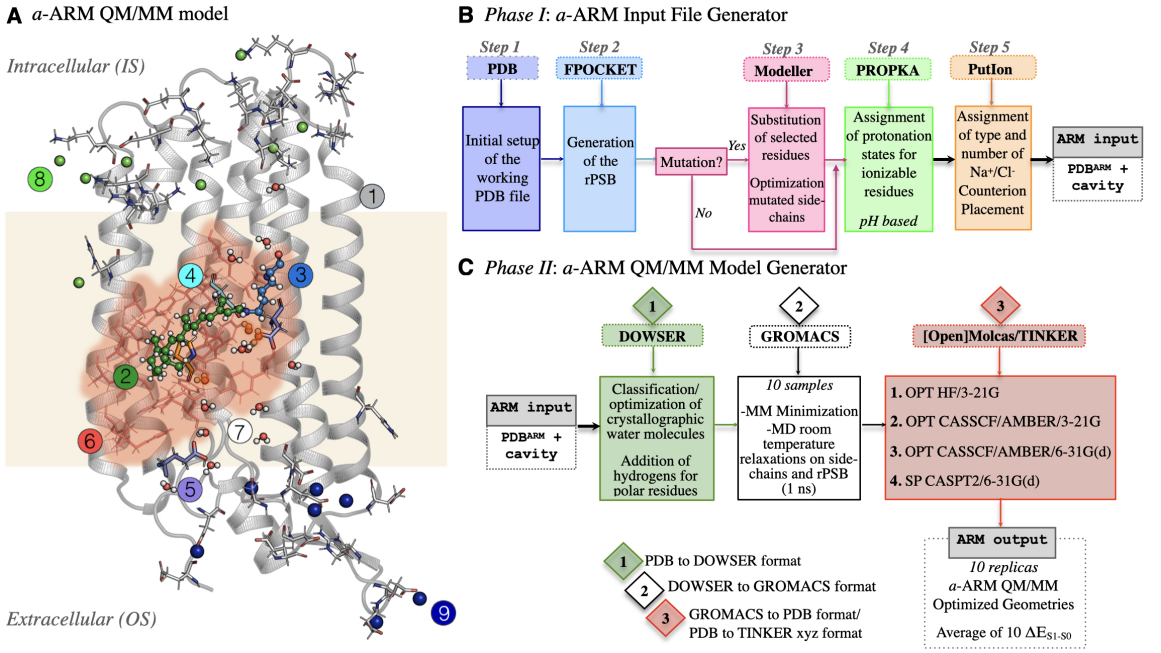

Figure 3.1: General workflow of the two phases, *i.e.*, *input file generator* and *QM/MM model generator*, of the *a*-ARM rhodopsin model building protocol. (A) (left) General scheme of a `ARM QM/MM model` for the wild-type KR2 rhodopsin. This is composed of: (1) environment subsystem (silver cartoon), (2) retinal chromophore (green tubes), (3) Lys side-chain covalently linked to the retinal chromophore (blue tubes), (4) main counter-ion MC (cyan tubes), (5) protonated residues, (6) residues of the chromophore cavity subsystem (red tubes), (7) water molecules, and external (8) Cl^- (green balls) and (9) Na^+ (blue balls) counterions. The external OS and IS charged residues are shown in frame representation. The residue P219 is presented as orange tubes. (right) The workflow of the two phases for the generation of QM/MM models of WT and mutant rhodopsins is also provided. (B) (right, top) *input file generator* phase and (C) (right, bottom) *QM/MM model generator* phase.

3.1.1.1 Methodological aspects

In the following, I will report on the most relevant features of the proposed *a*-ARM rhodopsin model building. To start with, in Figure 3.1 I illustrate the three main points concerning its framework. First, panel **A** displays the scheme of its output `ARM QM/MM model`. From the figure, and related caption, it is possible to notice that the model retains the same characteristics described in Section 2.2.1 for the *original* protocol (see Figure 2.1). Consequently, at the output level both versions are consistent. Then, panels **B** and **C** show the general workflow of *a*-ARM. As observed, the updated version comprehends two well-defined and automated phases, from now on called phase I and phase II, respectively. Whereas the latter is, substantially, the same *QM/MM model generator* phase reviewed in Section 2.2.1 (*i.e.*, in terms of methodology, not of implementation), the former is the new proposed *input file generator* phase.

The above setup allows for the automatic building of Ground-state `ARM QM/MM models` receiving, as the primary input, the structure of the rhodopsin of interest that can be provided either as the Protein Data Bank (PDB) code or as a suitable comparative (homology) model. As observed in Figure 3.1, such initial structure is processed by Phase I to obtain the `ARM input`, that is subsequently processed by Phase II to obtain the `ARM QM/MM model` (*i.e.*, gas-phase equilibrated optimized S_0 structure) and the predicted average λ^a_{max}. As

further explained in paper [**I**], the *input file generator* is implemented as an user-friendly command-line interface, where the researcher interacts with the program by typing information directly in the computer terminal[2], without the needed to manipulate text files or visualize chemical structures as was performed in the original "manual" strategy. Notice that a detailed Tutorial about the usage of the *input file generator* is provided in Appendix A.2.1.1. In papers [**I**] (see Section S1 in the SI) and [**III**] (see Section 3.1) I reported a detailed description of both *manual* (*i.e.*, *original* ARM) and *automatic* procedures employed to pursue steps 1-5 of Figure 3.1B, with particular emphasis on the improvements achieved with *a*-ARM, as well as in its higher level of automation. An overview of such improvements is presented in Figure 3.2.

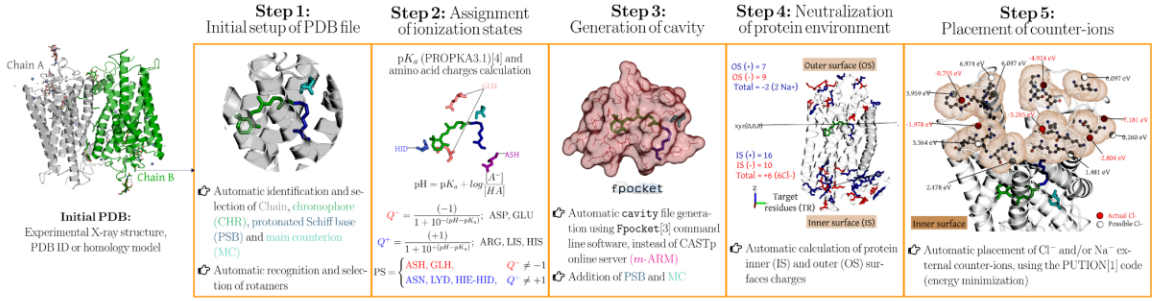

Figure 3.2: Overview of the most relevant features of the *input file generator*, introduced in the *a*-ARM version of the protocol. Methodological and automation improvements achieved with the *input file generator*, in terms of: initial setup; automatic strategy adopted for the assignment of protonation states for ionizable residues; replacement of the software (*i.e*, CASTp by fpocket) for the automatic generation of the chromophore cavity; automatic approach to define the charge of the IS and OS surfaces and automatic counterion placement based on energy minimization.

One of the most remarkable features of the new protocol is that, given the options (*i.e.*, parameters) selected in steps 1-5 of phase I, *a*-ARM allows either the automatic or semi-automatic computer-aided production of the `ARM input`. Accordingly, *a*-ARM is subdivided in *a*-ARM$_{\text{default}}$ (see Section 3.1 of paper [**I**]) and *a*-ARM$_{\text{customized}}$ (see Section 3.2 of paper [**I**]) approaches. The former refers to a fully automatic input generation, which uses default parameters as suggested by the code (*i.e.*, chain, rotamers or side-chain conformations, pH, protonation states, residues forming the chromophore cavity), whereas the latter allows the computer-aided customization of some of such parameters when the default choices are not suitable. The customized approach is used in cases where the default choices produce models that are not suitable for the reproduction of trends in absorption properties. For instance, Figure 3.3 illustrates how the customization, in terms of either selection of side-chain conformations and protonation states, is achieved for the case of KR2. As observed, in the 3X3C[78] X-ray structure the residue Asp-116, considered as the main counterion (MC) of the *r*PSB, exhibits two side-chain conformations, namely, AAsp and BAsp, labeled with occupancy numbers 0.65 and 0.35, respectively. Moreover, the residue Gln-157 that is part of the environment subsystem (*i.e.*, fixed during the QM/MM

[2]"Terminal" usually refers to a terminal program, or emulator, which provides a text-based interface for typing commands.

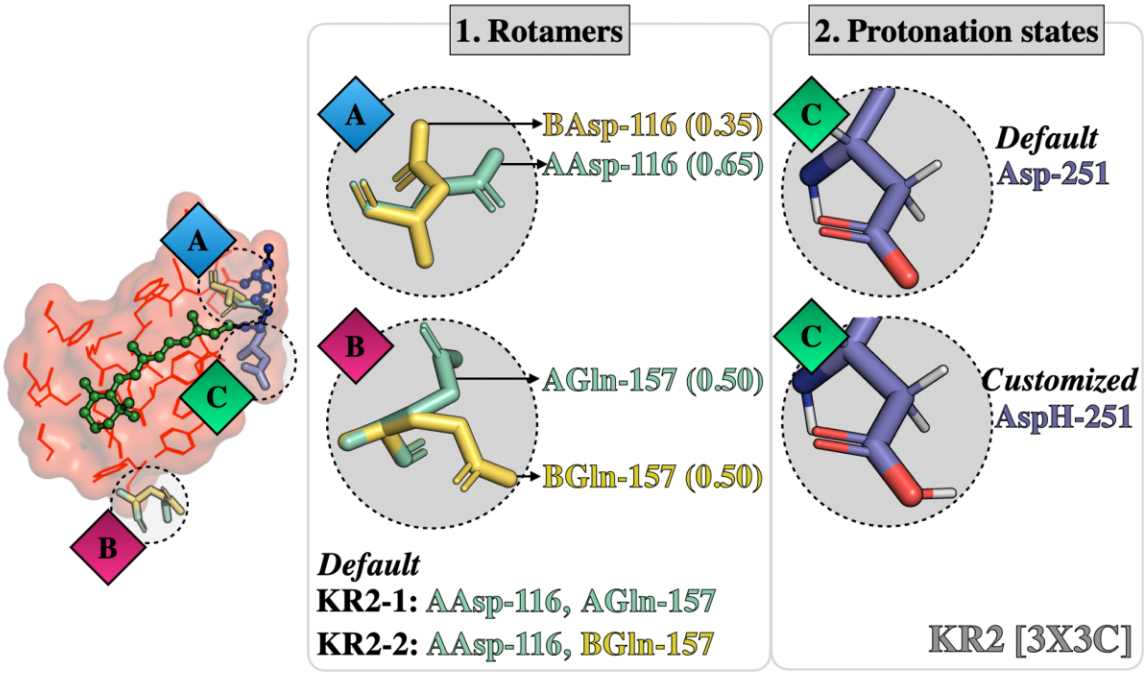

Figure 3.3: Default and customized *a*-ARM models for KR2 [PDB ID 3X3C[78]]. (left) Conformational (the occupancy factor of the rotamers Asp-116 and Gln-157 are presented in parentheses) and (rignt) ionization state variability.

calculations) presents two conformations (AGln and BGln) both with occupancy number 0.5. According to the occupancy numbers, a-ARM$_{\text{default}}$ selects the rotamer AAsp-116 and generates two models relative to Gln-157: KR2-1, which includes AAsp-116 and AGln-157, and KR2-2, which includes AAsp-116 and BGln-157. The computed ΔE^{a}_{S1-S0} for both default models, presented below in Figure 3.4, features an error of about 15.0 kcal mol^{-1} with respect to experimental data. Since the default models are unable to provide values inside the experimental trend, the a-ARM$_{\text{customized}}$ approach is necessary. As shown in the right panel of Figure 3.3, such customization is performed through a more rational assignment of the protonation states of the two aspartic acid residues forming the counterion complex of the rPSB, namely Asp-116 and Asp-251. The default model predicts that both aspartic acids are negatively charged. However, as further discussed in Section 5.1.3, the presence of these two negative charges would outbalance the single positive charge of the rPSB, generating the large blue-shifted effect mentioned above. Accordingly, in the customized model the secondary counterion (SC) Asp-251 is, instead, protonated (*i.e.*, neutral) to counterbalance the charge in the vicinity of the rPSB. As can be seen in Figure 3.4, such customization provides a model with a small error bar of about 1.5 kcal mol^{-1}.

As shown in the latter example, the customized `ARM QM/MM models` can be constructed according to well-defined operations that can be easily replicated. In fact, in paper [**I**] it had been proposed to adopt a guided-procedure focused on the selection of the ionization states and side-chain conformations only. Indeed, the novelty of the default and customized approaches is that regardless of the user or computational facility, reproducible inputs, and consequently reproducible `ARM QM/MM model`, are guaranteed when the same parameters

are employed. This represents an advancement with respect to the *original* version since it allows for the models to be reproduced in any laboratory and by any user, even when starting building an `ARM QM/MM model` from scratch (see point (c) on page 28).

3.1.1.2 Software implementation aspects

The computational implementation of both the *input file generator* and the *QM/MM model generator* phases as a Python-based modular code boosted the building of the `ARM` software package that will be introduced in Appendix A. As I will further describe in Section A.2.1, the *a*-ARM model building protocol is implemented as the *general driver* `a_arm_qmmm_protocol`, capable of executing phases I and II sequentially. Nevertheless, these phases are also implemented as the *stand-alone modules* `a_arm_input_generator` and `a_arm_qmmm_generator`, respectively, and can be executed independently. A manual user's guide for both generators is provided in Section A.2.1.1. The procedure specified there can be followed in order to reproduce all the results presented in reference [**I**] as well as reported in this thesis.

Furthermore, although *a*-ARM can presently build only rhodopsin models (*i.e.*, with natural retinal), it provides a template for the development and generation of an automatic QM/MM building strategy for other, more general systems such as rhodopsins incorporating artificial (i.e. unnatural) chromophores. This is straightforwardly achieved given the modular architecture of the `ARM` package and the fact that any chromophore can be treated when using the appropriate force field.

Finally, it is worthwhile to stress that the new protocol achieves all the features (a)-(e) described in section 2.2.3, overcoming the automation limits of the *original* version.

3.1.1.3 Benchmark, validation and application aspects

As shown in Figure 3.4, the validation of the *a*-ARM protocol for the prediction of trends in λ^a_{max} (*i.e.*, via the equation λ^a_{max}=$hc/\Delta E^a_{S1-S0}$; see Figure 1.3), was performed by using a benchmark set of 44 animal and microbial rhodopsin variants (*i.e.*, 25 wild type and 19 mutants) that come from different organism and are phylogenetically diverse.[62] For each rhodopsin, the ΔE^a_{S1-S0} values were obtained as the average of the energy difference between the S_0 and S_1 states in the 10 replicas generated for each model (see Section 2.2.2). The full benchmark set features values ranging from 458 nm (62.4 kcal mol^{-1}, 2.71 eV) to 575 nm (49.7 kcal mol^{-1}, 2.15 eV). Such a wide range provides information on the method accuracy, while the rhodopsin diversity provides information on the transferability and general applicability of the generated models. Figure 3.4 is divided in four different regions: *m*-set, *a*-set, $Rh-mutants$ set, and $bR-mutants$ set. The *m*-set and $Rh-mutants$ set are employed to compare the performance of *original* ARM and *a*-ARM versions, while the remaining sets focus on the performance of the *a*-ARM exclusively.

The *a*-ARM$_{\text{default}}$ approach proved to be capable of reproducing the ΔE^a_{S1-S0} values for

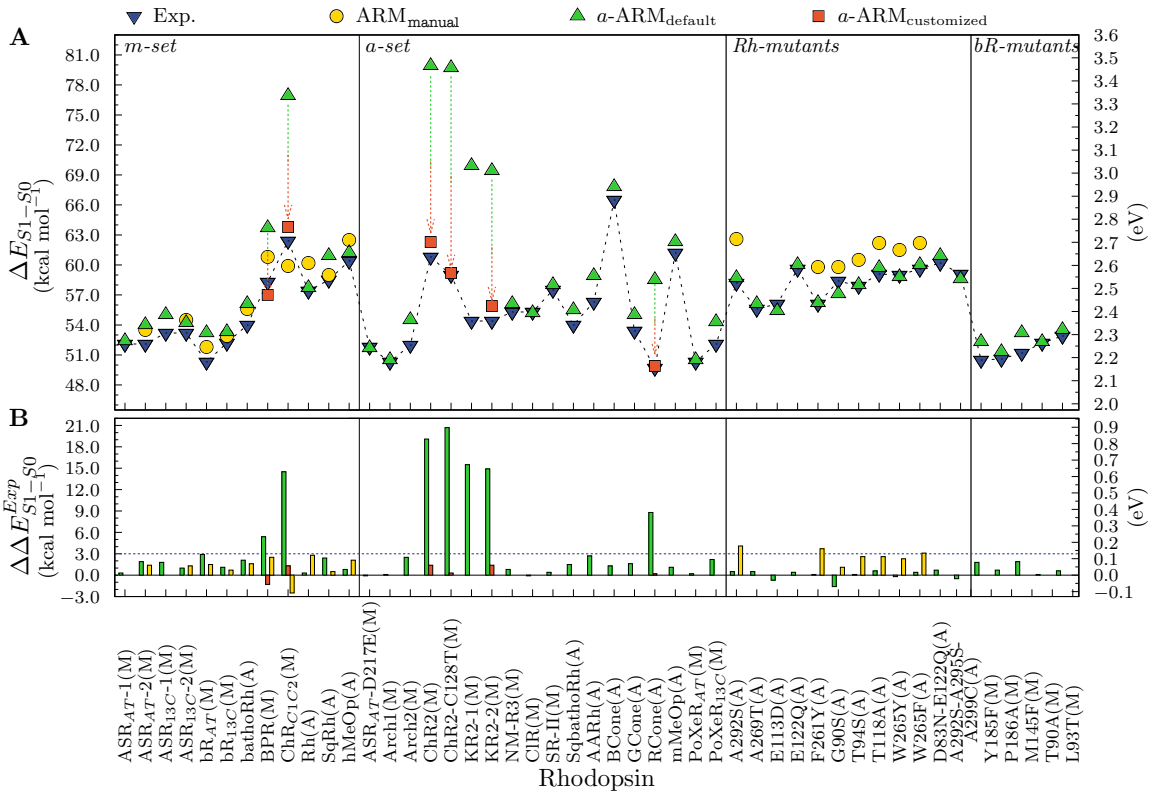

Figure 3.4: Benchmarking of the *a*-ARM version of the protocol, in terms of reproduction of experimental trends in λ^{a}_{max}. Comparison whit the *original* version that features manual input file generation. (A) Vertical excitation energies ($\Delta E_{\text{S1-S0}}$) computed with a-ARM$_{\text{default}}$ (up triangles) and a-ARM$_{\text{customized}}$ (squares),[62] along with reported ARM$_{\texttt{manual}}$ [61] (circles) and experimental data (down triangles). S_0 and S_1 energy calculations were performed at the CASPT2(12,12)//CASSCF(12,12)/AMBER level of theory using the 6-31G(d) basis set. The calculated $\Delta E_{\text{S1-S0}}$ values are the average of 10 replicas. (B) Differences between calculated and experimental $\Delta E_{\text{S1-S0}}$ ($\Delta\Delta E^{Exp}_{\text{S1-S0}}$). Values presented in kcal mol^{-1} (left vertical axis) and eV (right vertical axis). The *m*-set corresponds to WT rhodopsins forming the original benchmark set for the ARM protocol; *a*-set introduces new rhodopsins to the benchmark set of the *a*-ARM protocol; *Rh−mutants* set contains mutants of bovine Rhodopsin (Rh) belonging either to the benchmark set of the ARM or *a*-ARM protocols; and *bR−mutants* set contains mutants of bR evaluated with the Web-ARM interface [79] (see section 3.1.2). ASR: *Anabaena* sensory rhodopsin [1XIO[80]], bR$_{AT}$: Bacteriorhodopsin all-*trans* [6G7H[81]] and 13-*cis* [1X0S[82]], BPR: Blue proteorhodopsin [4JQ6[83]], Rh: Bovine rhodopsin [1U19[22]], ChR$_{C1C2}$: Chimaera channelrhodopsin [3UG9[84]], SqRh: Squid rhodopsin [2Z73[85]], hMeOp: Human melanopsin [template 2Z73[85]], ASR$_{AT}$-D217E: *Anabaena* sensory rhodopsin D217E [4TL3[86]], Arch1: Archaerhodopsin-1 [1UAZ[87]], Arch2: Archaerhodopsin-2 [3WQJ[88]], ChR2: Channelrhodopsin-2 [6EID[89]], ChR2-C128T: Channelrhodopsin-2 C128T [6EIG[89]], KR2: *Krokinobacter eikastus* rhodopsin 2 [3X3C[78]], NM-R3: Nonlabens marinus rhodopsin-3 [5B2N[90]], ClR: Nonlabens marinus rhodopsin-3 [5G28[91]], NpSRII: Sensory rhodopsin II [1JGJ[33]], SqbathoRh: Squid bathorhodopsin [3AYM[92]], AARh: Ancestral archosaur rhodopsin [template 1U19[22]], BCone: Human blue cone [template 1U19[22]], GCone: Human green cone [template 1U19[22]], RCone: Human red cone [template 1U19[22]], mMeOp: Mouse melanopsin [template 2Z73[85]], PoXeR: *Parvularcula oceani* Xenorhodopsin [template 4TL3[86]]. "Adapted with permission from Pedraza-González et al.[62]. Copyright 2019 American Chemical Society."

86% of cases (38/44), with an error lower than 4.0 kcal mol^{-1} (0.13 eV), whereas the other 14% cases were successfully obtained with the a-ARM$_{\text{customized}}$ approach[3] (*i.e.*, changing the side-chain conformation and/or protonation states pattern). Accordingly, the *a*-ARM protocol is capable of reproducing the experimental trends in vertical excitation energies for rhodopsins which structure was obtained from either X-ray crystallography or comparative

[3]A detailed description of the customization procedure employed for reproducing the experimental λ^{a}_{max} values of KR2, BPR, ChR2-C128T, ChR2 and ChR$_{\text{C1C2}}$ is provided in Section 3.2 of paper **[I]** and Section 3.3 of paper **[III]**.

modeling, as well as for rhodopsin mutants.

The final S_0 optimized equilibrium structures can be then used for further excited-state optimizations. Indeed, some of the `ARM QM/MM models` produced has been employed as input for sophisticated pH-constant dynamics,[93] the construction of fully relaxed QM/MM models embedded in a biological membrane, population analysis and for the simulation of one-/two-photon absorption spectrum.[47] Furthermore, as I will show in Chapter 5, the tool is already used for research.

3.1.1.4 Limitations and pitfalls of *a*-ARM

In spite of the encouraging outcome of the photochemical studies based on *a*-ARM (see Chapter 5), additional work is necessary to generate a tool that can be systematically applied to larger arrays of rhodopsins. Three main issues have to be tackled:

- *Assignment of the protonation states:* There are two aspects which limit the confidence in the automation of the ionizable state assignment described above. The first is that, due to the fact that the information provided by PROPKA [94] is approximated, the computed pK_a^{Calc} value may, in certain cases, be not sufficiently realistic. The second aspect regards the assignment of the correct tautomer of histidine. This amino acid has +1 charge when both the δ-nitrogen and ϵ-nitrogen of the imidazole ring are protonated (HIP), while it is neutral when either the δ-nitrogen (HID) or the ϵ-nitrogen (HIE) are deprotonated. *a*-ARM uses as default the HID tautomer for the automatic assignment [A], or allow the user to choose between the three tautomers for a not automated selection [M]. Therefore, when possible, the user should collect the available experimental data and/or inspect the chemical environment of the ionizable residues including the histidines, and propose the appropriate tautomer.[62] Alternatively, it is necessary to systematically examine all sensible choices which may not always be possible.
- *Automatic construction of comparative models*: since rhodopsin structural data are rarely available, it would be important to investigate the possibility of building, automatically, the corresponding comparative models. With such an additional tool one could achieve a protocol capable of producing QM/MM models starting directly from the constantly growing repositories of rhodopsin amino acid sequences. This target is currently pursued in our lab and a first attempt is presented in Section 5.3.1.
- *Automatic prediction of side-chain conformation for mutants*: In order to achieve a successful technology for systematically predicting mutant structures, a level of accuracy of the *a*-ARM models superior to the one currently available is needed. To deal with that, current efforts towards the improvement of the mutations routine are directed to replace SCWRL4 (*i.e.*, a backbone-dependent rotamer library) by a software based on comparative modeling. Further information about the new approach is provided in Section 3.8.

3.1.2 «Paper [II]» Web-ARM: a Web-Based Interface for the Automatic Construction of QM/MM Models of Rhodopsins

Published 3

The content of this Section is a review of the main findings reported in «Pedraza-González et al., *J. Chem. Inf. Model.* 2020, *60*, (3), pp 1481–1493». Copyright 2019 American Chemical Society.

Contribution: This is an original result of the research work carried out in this doctoral Thesis. I provided the Python-based code of the *a*-ARM rhodopsin model building protocol for its further implementation into the Web-ARM interface, carried out by Professor Luca De Vico. Moreover, I performed part of the benchmark calculations as well as the analysis of the produced results.

As reported in section 3.1.1, the computational tool driving the *a*-ARM protocol is implemented as a general driver inside the `ARM` Python-based software package (see Appendix A). Such a driver represents an easy-to-use command-line interface directed to users (*i.e.*, researchers, undergraduate students) familiar with the Linux environment. The latter is not due to a fact of usability (see user's manual in Section A.2.1.1), but rather to the technically complex initial setup of the package. More specifically, it requires the prior installation of several software (see Figure A.2) and python dependencies (see Section A.2.1). Moreover, it is recommended to install the `ARM` package in a high-performance computer cluster, which is usually required to run the underlying quantum chemical calculations (*i.e.*, MM, MD, QM/MM), rather than in a local personal machine.

The use of a user-friendly interface accessible trough the web is an appealing alternative to avoid i) to deal with complicate installations and ii) the need of a local computer facility. This would lead the access to a simplest, computationally fast and automated construction and analysis of rhodopsin QM/MM models to an interdisciplinary community that is mostly interested on the actual applications rather than in methodological development.

Accordingly, paper **[II]** reports the Web-Version of the *a*-ARM protocol[62]. This is the Web-ARM interface,[79] a user-friendly interface written using Python 3, available to all the scientific community at the following address: `webarm.org`. Therefore, the potential user only needs to use an up-to-date browser in any operating system (*i.e.*, Linux, macOS, Windows, Android, iOS) and, thus, without the need to install software and scripts on his/her computer can access and made use to the Web-ARM interface. Actually, to access and use the interface, the user can employ his/her own smartphone or tablet. In order to generate a `ARM QM/MM model`, the Web-ARM features four phases explained in Section 2.1 of paper **[II]**, and illustrated in Figure 3.5. As described for the command-line version, the procedure starts with the initial structure of the rhodopsin variant and finishes with

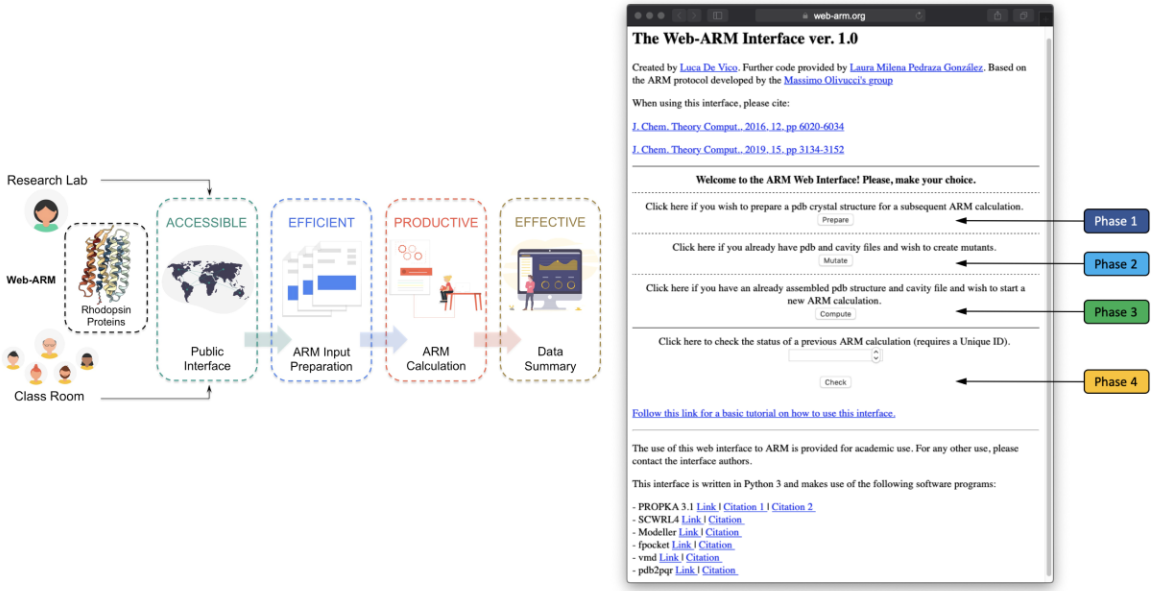

Figure 3.5: General overview of the Web-ARM interface. (left) Main features oand (right) home page of the Web-ARM interface. "Adapted with permisions from Pedraza-González et al.[79]. Copyright 2019American Chemical Society"

the generation of the S_0 equilibrium geometry along with the calculations on absorption properties. During such procedure, the interface gives to the user enough flexibility to generate either a-ARM$_{\text{default}}$ or a-ARM$_{\text{customized}}$ inputs, the former automatically and the latter by modifying some of the default choices. This is made on top of the implementation of the *input file generator* inside the framework of the web interface. Then, the `ARM input` is employed to generate a QM/MM model, by running the corresponding 10 repetitions, by using the implementation of the *QM/MM model generator*. The Web-ARM internal driver takes care of performing all the necessary steps, as well as submitting the calculations to the dedicated computational facilities.

One feature of the interface is that, once a QM/MM model is generated, the user is provided with a summary of all the relevant data (*i.e.*, energetics, oscillator strengths), along with a downloadable file (in compressed format) containing the major output files. Further information, and a complete walk-through, are provided in a Tutorial that can be accessed/downloaded from the Web-ARM main web page.

Finally, I would like to remark the fact that Web-ARM is intended as both a research, as well as a teaching tool. In fact, in the paper I and my collaborators show that the interface can be employed to systematically screen rhodopsin variants, and thus obtain a qualitative check prior to, *e.g.*, an experimental study. We also explain that the interface can be successfully used in teaching and learning activities, *e.g.*, to introduce students to the idea of QM/MM models and corresponding computed data. Therefore, we envision Web-ARM as a tool employed in teaching and training, as well as by non experienced users, as previously noted, mostly for bulk production. However, we believe that also an experienced computational chemist can take advantage of the web interface, to produce rhodopsin

QM/MM models in a standardized manner being aware of the documented accuracy and rate of success. Of course, one, possibly very useful, application of such model is to provide high quality guesses for more sophisticated calculations, such as the above mentioned pH-constant dynamics or, for instance, the generation of a fully relaxed QM/MM model embedded in explicit membrane and solvent. Such, likely few, models could prove useful as a starting substrate to which apply further, high level refinements methods.

In conclusion, by using Web-ARM both junior researchers and trainees will thus be able to perform meaningful QM/MM calculations focusing on the underling research targets, methodological concepts, and data analysis, while remaining confident that the calculations are internally consistent.

3.1.2.1 Limitations and pitfalls of Web-ARM

- *Limited computational resources:* Before using the Web-ARM interface, the user is asked to provide an email address to be registered into our database. Registered users are allowed to build as many concurrent `ARM QM/MM model`s as wished (default 10). However, given our current limited computational resources, guest users are allowed to build only one `ARM QM/MM model` model at a time on the developer's dedicated resources.
- *Technical issues:* The Tutorial of the Web-ARM interface also reports on possible errors or issues in the execution of the interface, and how to solve them.
- *Current implementation:* Presently, the capability of the Web-ARM interface is limited to the construction of ground-state models. Future work will be devoted to implement all of the features of the `ARM` software package, illustrated in Figure 1.4, inside the interface.

3.1.3 A standard protocol for the analysis of color tuning

Color tuning is the modulation of the absorption wavelength of the chromophore by the opsin surrounding amino acid residues. In the last few decades, the origin of such color tuning mechanism in members of the Rhodopsin family (*i.e.*, animal, microbial and heliorhodopsins) has been the focus of many both experimental and theoretical research efforts.[27–29, 33–40, 42] Varying the λ^a_{max} value of the light captured by a specific rhodopsin variant, obtained as a consequence of the "color tuning effect", can generate a pool of biological functions that allow several applications (see Section 1.1.2).[10]

While the color tuning mechanism is still not fully understood, it is apparent that it must be determined by interactions (*i.e.*, steric and electrostatic) between the chromophore and the surrounding amino acid residues. In turn, each amino acid features a side-chain that may have a different nature, *e.g.*, charged, dipolar, aromatic and capable of hydrogen-bonding and steric contact effects. Accordingly, gaining insights into the precise "rules" for controlling such interaction seems to be of basic importance for the understanding of

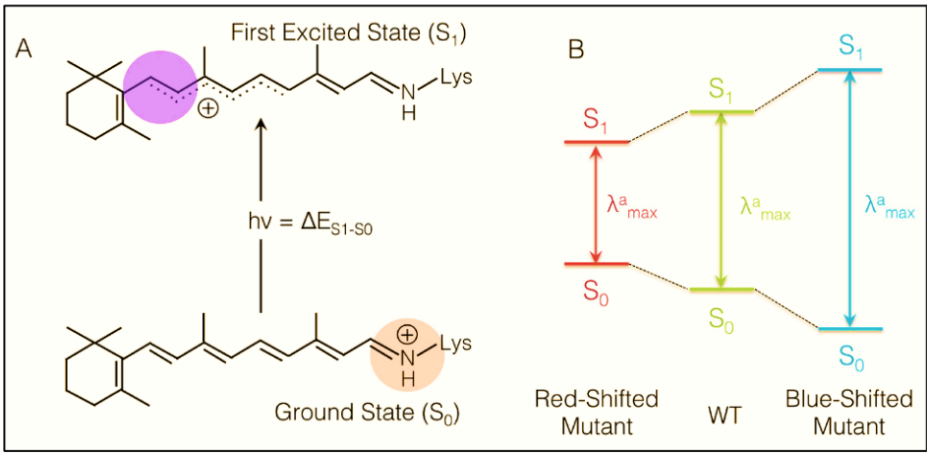

Figure 3.6: Schematic representation of color tuning mechanism in rhodopsins. Color tuning effects on the Maximum absorption wavelength, modulated by positively or negatively charges introduced near the β-ionone ring or the -CH=NH- moiety of the *r*PSB. A decrease in the energy gap between the S_0 and S_1, *i.e.*, Vertical Excitation energy (ΔE^{a}_{S1-S0}), induces a red-shifted effect, whereas an increase produces a blue-shifted effect.

different facets of rhodopsins biology, including their evolution, ecology, biophysics and the laboratory engineering required for optogenetic applications (see Section 1.1.3.1).

The general principle of color tuning states that the introduction of a polar residue in the vicinity of the β-ionone ring or the Schiff base moiety of the chromophore causes spectral red and blue shift, respectively. In order to improve the understanding of color tuning, theoretical studies rely on the analysis of electronic structure and, in turn, positive charge distribution, of the ground (S_0) and first excited (S_1) states of the *r*PSB (see Figure 3.6). Briefly, while in S_0 the positive charge is substantially localized in the -CH=NH- moiety of the *r*PSB, in the S_1 charge-transfer state the positive charge is delocalized towards the β-ionone ring. Therefore, negatively charged atoms located in the vicinity of the β-ionone ring would stabilize S_1 with respect to S_0 leading to a red-shift of the λ^{a}_{max} value. In contrast, if negative atoms are located in the vicinity of the Schiff base moiety they would stabilize S_0 with respect to S_1 leading to a blue-shift. Evidently, positively charged atoms will have an opposite effect.

As a second objective of this Thesis (see Section 1.3), I attempt at designing an automatic protocol to perform color tuning analysis over the `ARM QM/MM models`, in terms of the elucidation of steric and electrostatic effects that modulate the energy of either the S_0 or the S_1 states and, consequently, the absorption wavelength, in series of rhodopsin mutants. In Figure 3.7 I provide a visual representation of three fundamental quantities (*i.e.*, $\Delta\Delta E^{TOT}_{S1-S0}$, $\Delta\Delta E^{STR}_{S1-S0}$, $\Delta\Delta E^{ELE}_{S1-S0}$) that are considered in the excitation energy analysis and whose values are a function of the structural changes (both at the chromophore and protein cavity levels) of each mutant with respect to the WT.

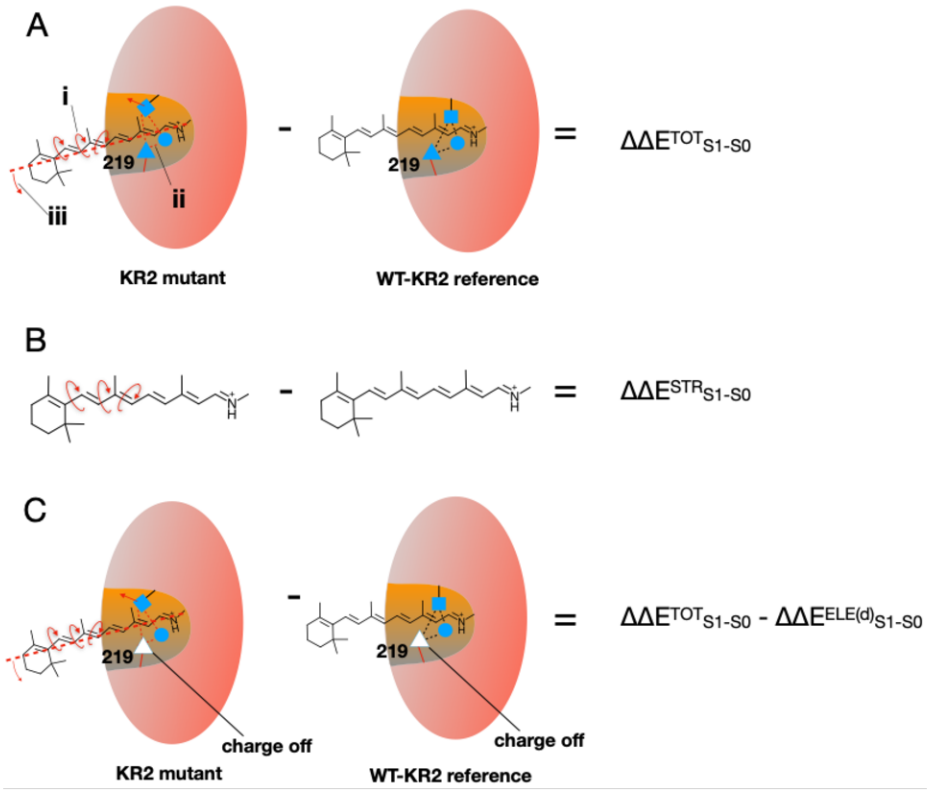

Figure 3.7: Scheme of the excitation energy analysis for the elucidation of its electrostatic and steric contributions, that modulate color tuning mechanisms in rhodopsins. Excitation energy analysis. A. Difference between the *a*-ARM models excitation energies of a microbial rhodopsin mutant with respect to the wild-type due to (i) structural deformation (see curly arrows) and (ii) structural deformation of the cavity residues (see dashed lines indicating the hydrogen bond network) and (iii) *r*PSB reorientation. B. Difference between the excitation energies of the mutant and isolated chromophores taken with their protein equilibrium structures. C. Difference between the *a*-ARM models excitation energies of a P219X mutant with respect to the wild-type due to Difference between the excitation energies of the isolated chromophores taken with their protein equilibrium structure.

3.1.3.1 Steric effects.

The steric contribution, from now on called $\Delta\Delta E^{STR}_{S1-S0}$, allows to "quantify" the effect of the geometrical rearrangement of the retinal, after mutation, on the vertical excitation energy and, therefore, on the λ^{a}_{max} value. The analysis to compute $\Delta\Delta E^{STR}_{S1-S0}$ consists in extracting, for WT and for each variant, the *r*PSB structure from the protein environment and computing its vertical excitation energy without relaxing. Then, the computed ΔE^{RET}_{S1-S0} of each variant ($\Delta E^{RET,MUT}_{S1-S0}$) is compared with the corresponding WT ($\Delta E^{RET,WT}_{S1-S0}$) for the isolated chromophore, as follows:

$$\Delta\Delta E^{STR}_{S1-S0} = \Delta E^{RET,MUT}_{S1-S0} - \Delta E^{RET,WT}_{S1-S0} \tag{3.1.1}$$

A graphical representation of the above described procedure is given in Figure 3.7B. For the interpretation of the results, a small value of $\Delta\Delta E^{STR}_{S1-S0}$ (with respect to the total $\Delta\Delta E^{TOT}_{S1-S0}$) indicates that the geometrical distortion of the retinal due to the point mutation has only a limited effect.

3.1.3.2 Electrostatic effects.

The electrostatic contribution, hereinafter referred to as $\Delta\Delta E^{ELE}_{S1-S0}$, allows to investigate the electrostatic effects, induced by the mutation, due to the interaction of the *r*PSB with the modified protein environment. The total electrostatic effect can be decomposed in two parts: (i) a *direct* component ($\Delta\Delta E^{ELE(d)}_{S1-S0}$) due to the variation in number, magnitude, and position of the point charges of the mutated residue caused by the side-chain replacement and (ii) a more *indirect* component produced from the reorganization of the local environment and Hydrogen-bond networks (HBN) induced by the same replacement and due to the fact that conserved residues and water molecules change in position or orientation (see Figure 3.7C). The former $\Delta\Delta E^{ELE(i)}_{S1-S0}$ component can be evaluated in two steps by first computing the differences between the vertical excitation energy of the mutant and WT obtained after setting the mutated residue charges to zero, $\Delta E^{MUT,OFF}_{S1-S0}$ and $\Delta E^{WT,OFF}_{S1-S0}$ respectively, and then by subtracting from such difference the steric effect $\Delta\Delta E^{STR}_{S1-S0}$ described above, as follows:

$$\Delta\Delta E^{ELE(i)}_{S1-S0} = (\Delta E^{MUT,OFF}_{S1-S0} - \Delta E^{WT,OFF}_{S1-S0}) - \Delta\Delta E^{STR}_{S1-S0}. \tag{3.1.2}$$

The direct component is then simply defined as,

$$\Delta\Delta E^{ELE(d)}_{S1-S0} = (\Delta\Delta E^{TOT}_{S1-S0} - \Delta\Delta E^{STR}_{S1-S0}) - \Delta\Delta E^{ELE(i)}_{S1-S0}. \tag{3.1.3}$$

Finally, the total electrostatic contribution is computed as

$$\Delta\Delta E^{ELE}_{S1-S0} = \Delta\Delta E^{ELE(d)}_{S1-S0} + \Delta\Delta E^{ELE(i)}_{S1-S0} = \Delta\Delta E^{TOT}_{S1-S0} - \Delta\Delta E^{STR}_{S1-S0}. \tag{3.1.4}$$

Both the steric and electrostatic analysis described above are implemented into the `ARM`. On the other hand, the main application of the above described protocol, presented in this Thesis, is devoted to investigate how all possible P219X mutants, where X stands for all alternative 19 natural amino acids, modulate the λ^a_{max} value of the reference WT light-driven sodium pump Krokinobacter rhodopsin 2 from *Krokinobacter eikastus* (KR2) (*i.e.*, 525 nm), and understand the molecular-level mechanism driving color tuning. The results are discussed in Section 5.1.3 and paper **[V]**.

3.1.3.3 Limitations and pitfalls of the protocol for color-tuning analysis

- ⊘ *Elucidation of indirect electrostatic contribution*: Although this methodology allows to "quantify" the energy contribution of *indirect* electrostatic effects to the total ΔE^a_{S1-S0}, due to residue replacement, it does not provide a tool to discern between the different individual components described above (*i.e.*, reorganization of the local environment, Hydrogen-bond networks (HBN) rearrangement)
- ⊘ *Results dependence on the selected seed*: The analysis on color tuning is performed

employing the *a*-ARM S_0 QM/MM structure whose λ^a_{max} value is closest to the average of 10 replicas. Therefore, the output result is limited to the description of a single structure that can slightly differ from the other 9 seeds (mainly for models with high standard deviation of λ^a_{max}). Future work will be devoted to evaluate the impact of using 1 value instead of the average of 10 repetitions.

3.1.4 A strategy for the prediction of side-chain conformations in mutants

It is well-known that the success of *in silico* modeling of point mutations in proteins, relies on the selection of a robust methodology for the prediction of the side-chain conformation of the replaced amino acid.[95–107] As specified above, both *original* ARM[61] and updated *a*-ARM[62] versions of the protocol, employ the software SCWRL4[99, 107, 108] to predict the side-chain of the mutated residues. Such a prediction is based on rotamer libraries[107] from public databases of experimentally-resolved protein structures. As shown in Sections 3.1.1 and 3.1.2 (see also Section 4 of paper **[III]**), this approach has demonstrated to be effective for the production of single, double and triple point mutants in different rhodopsins that are phylogenetically diverse. More specifically, previous studies, carried out in the LCPP laboratory, were focused on modeling mutants for the type II Rh,[61, 62] and type I ASR,[49, 61] bR[79] and KR2[109] rhodopsins.

On the other hand, to exploit the new features of the updated *a*-ARM rhodopsin model building protocol and, specifically, its current level of automation, I struggle to move the research target from few uncomplicated mutations to large arrays of more complex mutations. For instance, as it will be further explained in Section 5.1.3 and paper **[V]**, I have recently attempted to use *a*-ARM to systematically mutate a specific residue of the microbial Krokinobacter rhodopsin 2 from *Krokinobacter eikastus* (KR2), with each of the remaining 19 essential amino acids. More specifically, I performed single mutations of the residue P219, that is located near the β-ionone ring of the *r*PSBAT, by replacing the side-chain of the proline (P) (*i.e.*, P219X, with X= A, C, D, E, F, G, H, I, K, L, M, N, Q, R, S, T, V, W, Y). Following the workflow of *a*-ARM (see Figure 3.1B), and motivated by the encouraging results presented in Ref. 109 for the case of P219A and P219G, I first used SCWRL4 for the side-chain prediction. In doing so, I found that the replacement of P219 for larger side-chains performed by SCWRL4 may generate mutated side-chain structures, sterically clashing with either the *r*PSBAT or neighboring amino acids. In such cases, the produced mutant cannot be considered as a suitable input geometry for the Molecular Dynamics (MD) step in the Phase II of the protocol (see Section 2.2.2 and Figure 3.1C). As will be described below, this issue also happened when producing mutants for other rhodopsins such as ASR.

In order to overcome the above drawbacks and, thus, achieve a successful technology in terms of predicting mutant spectral properties, it is needed a level of accuracy of the corresponding *a*-ARM models which is superior to the one currently available. Considering the different tools available for side-chain predictions (see for instance Ref. 95), and evaluating

their advantages and pitfalls in terms of i) performance and ii) accessibility as command-line tool, I have modified the routine for mutations in *a*-ARM by substituting the SCWRL4[108] software with the program for comparative modelling Modeller.[110]

This alternative approach allows the production of mutants with side-chains suitable for the prediction of absorption wavelengths in either an automatic or a computer-aided semi-automatic fashion. The general workflow of the proposed subroutine, that constitutes the Step 3 of the Input file generator phase of *a*-ARM (see Figure 3.1B), is illustrated in Figure 3.8. As observed, at the input level of the protocol, each point mutation is generated with a customized version of the `mutate_model.py` routine implemented by Modeller,[110] were the conformation of the modeled side-chain is optimized by conjugate gradient and refined using a short MD (MD^{mod}). To start the procedure the user should provide a file with extension `".seqmut"`, that contains the list of residues to be mutated. Then, the structure (non-hydrogen atoms protein representation) of the WT is used as an input to execute the mutant generator subroutine with the customized setup shown in Figure 3.8. Briefly, the `mutate_model.py` script has been designed to model point mutations via side-chain replacement in a fixed environment, assuming that single mutations do not generally determine deep conformational changes of the protein backbone. Accordingly, and also consistently with the structurally "conservative" approach of the *a*-ARM protocol where our models are designed to retain information from the X-ray crystallographic or comparative structures, our methodology replaces only the side-chains of the mutated residues keeping the backbone atoms at fixed positions. To this aim, the optimization of the mutated side-chains is obtained using a combined approach which alternates conjugate gradient minimizations and short MD^{mod} simulations with simulated annealing. This intends to minimize a scoring function including homology-derived restraints, force field energy terms (CHARMM22), and a statistical potential for non-bonded interactions. Notice that the MD^{mod} used in this step differs from the MD employed in the QM/MM model generator phase that is described in Section 2.2.2.

In the above procedure, the script `mutate_model.py` uses a default initial condition or "seed" (variable rand_seed= -49837) for the MD^{mod} simulation. Therefore, since Modeller is deterministic, if such seed value is not modified the MD^{mod} run will always produce the same side-chain conformation when a certain template is used as input. In order to sample more extensively the conformational space of a mutated residue and evaluate its effect on the ΔE^{a}_{S1-S0}, our customized setup produces multiple rotamers of the same mutated side-chain by providing the script with different initial seeds (*i.e.*, initial velocities) for the MD^{mod} run. Thus, our customized approach uses 30 different seeds to potentially generate 30 representative side-chain conformations of a single mutant. Considering that probabilistically different initial conditions (seeds) may yield the same side-chain conformation, a strategy to discard the duplicates is required. To this aim, the subroutine compares recursively every rotamer with one another and establishes a Root-mean square-deviation (RMSD) threshold

of 0.025 Å, below which we decide that two conformers are identical and one of the two needs to be discarded. Although not particularly efficient from a computational standpoint, given the low number of conformations to evaluate, this procedure allows for the quick selection of a set of non-redundant rotamers for a single mutant. Subsequently, the remaining structures are evaluated by using the scoring function Discrete optimized protein energy (DOPE)[4] and molpdf, and ordered from lowest to highest DOPE, implemented by Modeller and ranked from lowest to highest DOPE. The DOPE score is a statistical potential developed by Shen et al.[111] which can be used for external assessment of model accuracy (i.e., it is not involved in the building routine).

Then, in order to explore the performance of different rotamers of the mutated side-chain, three QM/MM models featuring the three highest scored mutated side-chain rotamers selected by Modeller, are produced and their ΔE^{a}_{S1-S0} is evaluated. Subsequently, the model that better reproduces the observed ΔE^{a}_{S1-S0} is selected. The three highest scored mutated side-chain rotamers (as selected by Modeller) are used to produce three QM/MM models. The corresponding computed ΔE^{a}_{S1-S0} is used to evaluate the performance of different rotamers of the mutated side-chain. Subsequently, the model that better reproduces the observed ΔE^{a}_{S1-S0} is selected. To perform such selection, the difference between the computed and observed ΔE^{a}_{S1-S0} of the WT, hereafter referred to as $\Delta^{Exp}_{calc}\Delta E^{a}_{S1-S0}$, is used as a baseline. The equivalent quantity calculated for each rotamer ($\Delta^{Exp}_{rotX}\Delta E^{a}_{S1-S0}$, with X=1,2,3) is then contrasted with the $\Delta^{Exp}_{WT}\Delta E^{a}_{S1-S0}$ via the equation:

$$rotX = (\Delta^{Exp}_{rotX}\Delta E^{a}_{S1-S0} - \Delta^{Exp}_{WT}\Delta E^{a}_{S1-S0}). \tag{3.1.5}$$

The rotamer that features the lowest $rotX$ value (preferring blue-shifted values) is chosen as the representative `ARM QM/MM model` (see Figure 3.9B). Although this approach relies on experimental information and does not represent a predictive tool, it automates the side-chain conformation selection during the construction of mutant QM/MM models.

3.1.4.1 Benchmarking of side-chain predictor

In order to validate the, above described, subroutine for the automatic prediction of side-chains conformations of mutants, I have used a set of 30 mutants with available experimental data on λ^{a}_{max}. More specifically, the `ARM QM/MM models` were generated[5] for i) the WT-ASR$_{AT}$ and 15 of its mutants as well as for the WT-ASR$_{13C}$ and 15 of its mutants. Details on the procedure for the `ARM QM/MM model` generation is provided in Table 3.1, as follows:

Figure 3.10 reports the computed ΔE^{a}_{S1-S0} for the three rotamers for each of the 30 mutants, along with experimental data taken from the database reported in Ref. 29. As observed, in all the cases a $\Delta^{Exp}_{calc}\Delta E^{a}_{S1-S0}$ lower than 4.0 kcal mol^{-1} is obtained and a semi-

[4]DOPE is an atomic distance-dependent statistical potential calculated from a sample of native protein structures. It is grounded entirely in probability theory. See Ref. 111

[5]The QM/MM calculations were performed by the student Michał Marszałek, during the training period of his internship our my co-supervision.

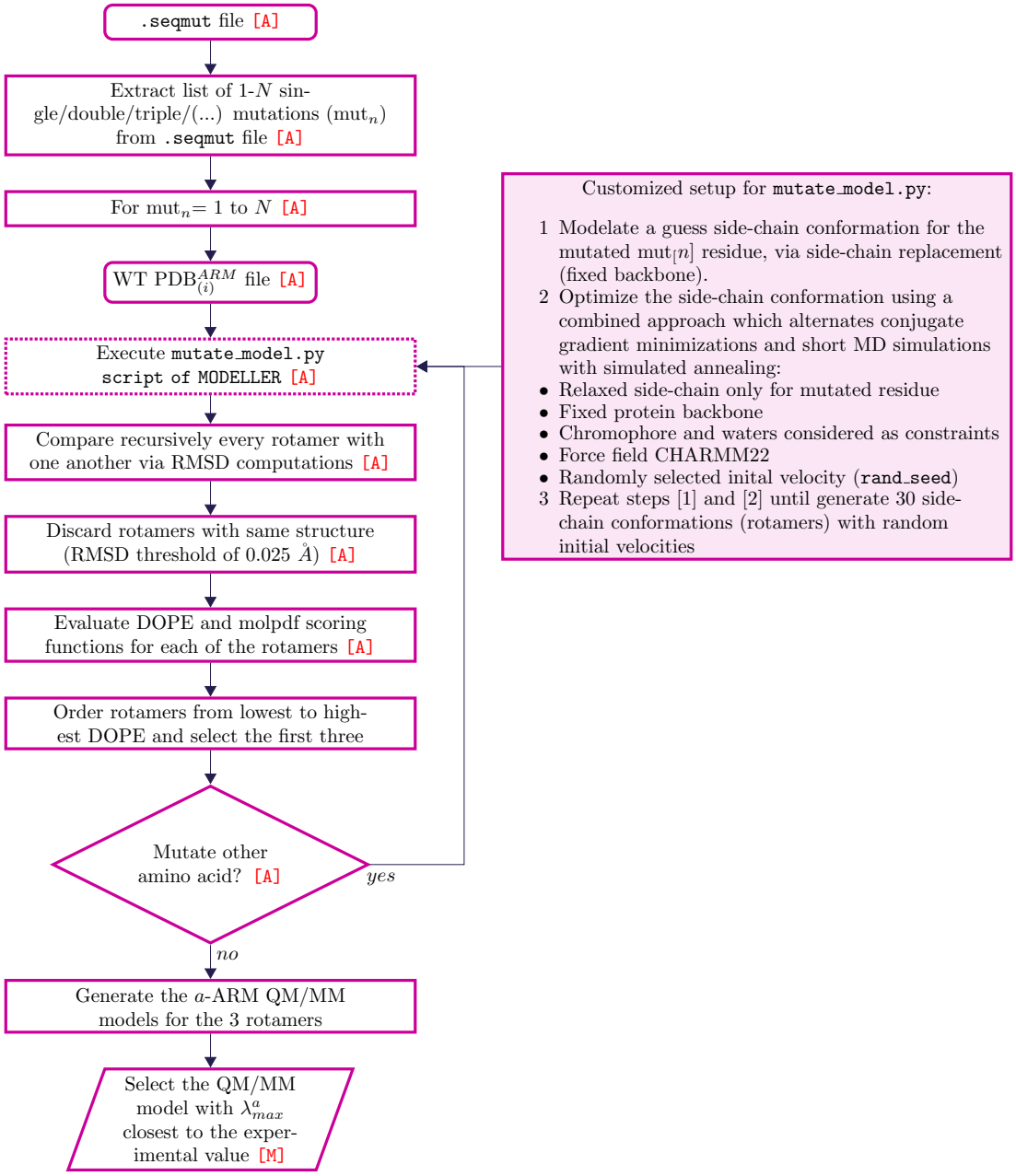

Figure 3.8: Proposed approach for the modelling and choice of side-chain conformations, included as the mutant generation routine in the *input file generator*. General workflow of the modified routine for the mutant's generator of *a*-ARM, in which the SCWRL4 rotamer library is replaced by Modeller, a software for comparative modelling.

automatic selection is achieved. However, 8 out of the 30 systems present a red-shifted effect with respect to experimental data instead of the expected blue-shifted effect documented for the *a*-ARM protocol (see Refs. 61 and 62). Such systems represent punctual cases where mutant prediction does not work satisfactory even when three different rotamers are considered. Actually, cases as S247A and S290A, that is cases where there is not a flexible side-chain, are an example that the current strategy is not enough evolved to model

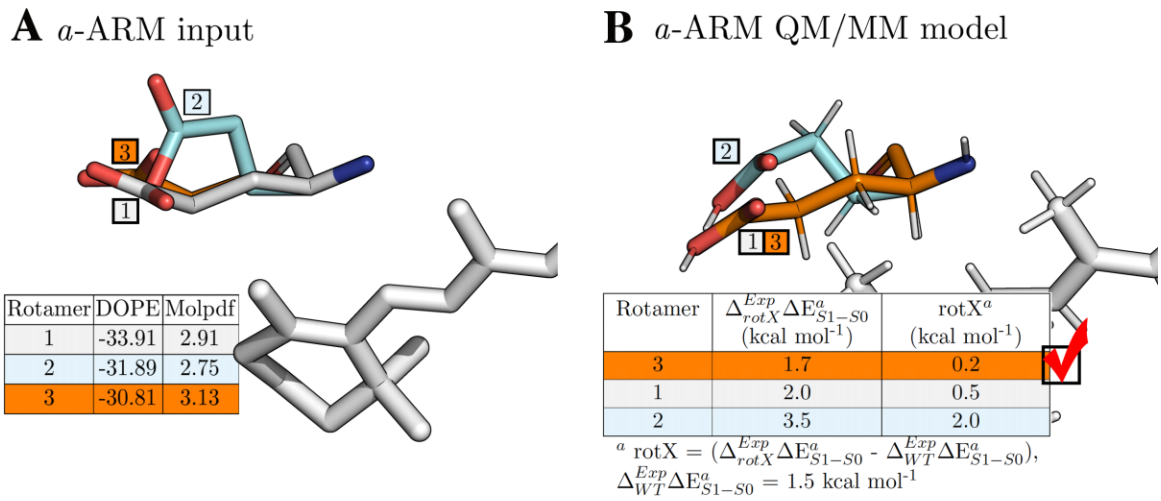

Figure 3.9: Schematic representation of the procedure employed for the selection of the side-chain conformation in mutants generation. The side-chain of the E219 residue of the KR2 rhodopsin, is modeled by using the procedure specified in Figure 3.8. (A) First, the DOPE and molpdf scoring functions for all possible rotamers are evaluated and the three best values are ranked. (B) Then, the *a*-ARM QM/MM model for each rotamer is generated and the rotamer model featuring the lower difference in ΔE^{a}_{S1-S0} with respect to experimental data (rotamer 3) is selected.

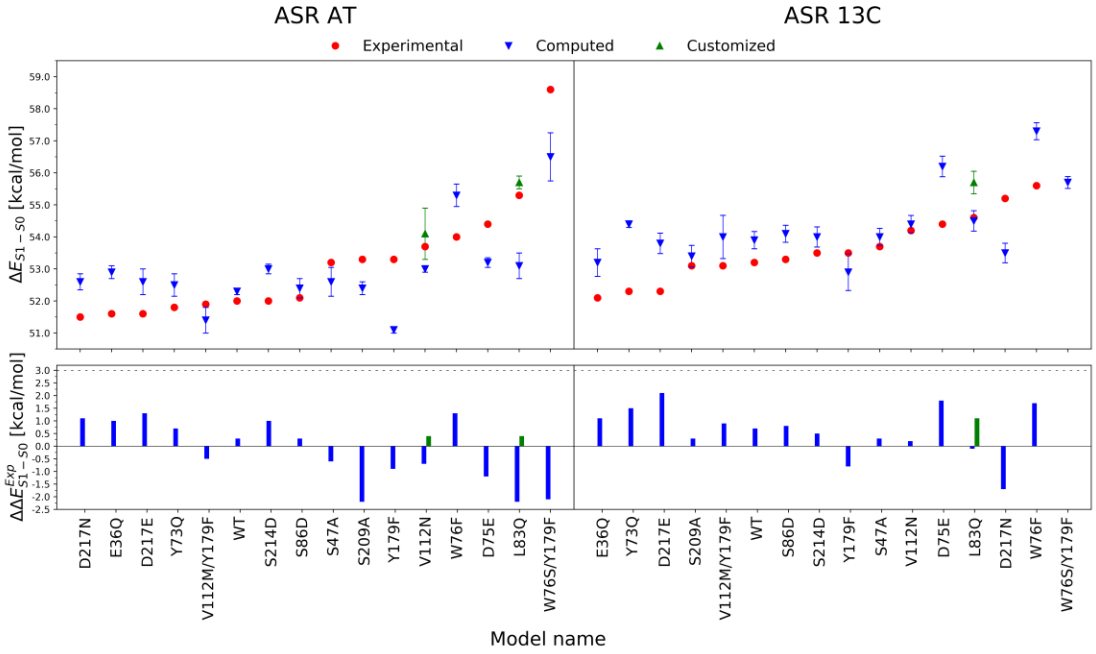

Figure 3.10: Benchmarking of the mutants generator routine based on Modeller. The Vertical Excitation energy (ΔE^{a}_{S1-S0}) values computed for the three best ranked rotamers for a set of 30 mutants of ASR (15 all-*trans* and 15 13-*cis*) are presented. The procedure to select the rotamers is illustrated in Figure 3.9.

Table 3.1: Overview of structural features and both experimental and computational data for the `ARM QM/MM model` models of *Anabaena* sensory rhodopsin (ASR).

***Anabaena* sensory rhodopsin (ASR)**			
General Information			
PDB ID:	1XIO[80]	Chromophore:	Retinal (RET)
RET Configuration:	all-*trans*	Lysine Linker:	A-Lysine 210[a] (K210)
RET Configuration:	13-*cis*	Lysine Linker:	B-Lysine 210[a] (K210)
Proton acceptor:	Aspartic 75 (D75)	Proton Donor:	—
pKa Analysis at crystallographic pH 5.6			
Protonated residues:	ASH 198 ASH 217 GLH 36 HID 8		
Counterion Distribution:	Inner surface: $7Cl^-$	Outer surface: $1Na^+$	
Cavity Residues:	11[b], 43[b], 47[b], 73, 75, 76, 79, 80, 83, 86[b], 109, 112, 113, 116, 119, 131, 132, 134[b], 135, 136, 137[b], 139, 176, 179, 180, 183, 198, 202, 203[b], 206[b], 209, 210, 211[b], 214[b]		

[a] The X-ray structure contains information on the two configurations of the retinal, as well as on the two conformations for the linker lysine-210.
[b] Residues added to the customized cavity.

the possible steric/electronic rearrangements of the chromophore cavity when a mutation is introduced.

3.1.4.2 Limitations and pitfalls of side-chain predictor

- *Insufficient description of possible cavity rearrangements after mutation*: The procedure for modeling the side-chain conformation with Modeller comprises a short MD, where the modeled side-chain is allowed to relax, whereas the rest of the cavity remains fixed. Therefore, possible local steric/electronic rearrangements of the residues of the chromophore cavity surrounding the mutated residue are not correctly described. Although during the QM/MM calculation phase the geometry of this side-chain along with the side-chain of the residues in the chromophore cavity are refined via a more sophisticated Molecular Dynamics (MD) (see Section 2.2.2), in some cases this step is not sufficient to achieve a proper description of the impact of the new side-chain on the protein environment.
- *Mutations only allowed in the chromophore cavity:* Currently, *a*-ARM only allows mutations of residues that belong to the chromophore cavity sub-system, as well as, backbone relaxation is not allowed. The latter is to ensure that, during the calculations phase, the geometry of the new modeled side-chain as well as the side-chain of its neighbors (belonging to the chromophore cavity) can be re-adjusted during the MD phase, while assuming that the general structure of the protein is conserved.
- *Lack of a predictive tool*: The fact that the mutants generator relies on the use of experimental data to select the correct rotamer, limits the usability of the protocol that can not be considered as a predictor tool.

As a perspective of this work, I suggest to introduce a mutant generator routine using proper comparative (homology) modeling instead of just modifying locally the mutated side-chain conformation. I believe that this would solve some of the issues described above.

Chapter 4

Automated QM/MM Model Screening of Rhodopsin Variants Displaying Enhanced Fluorescence

The computational modeling of excited-states behavior of microbial rhodopsins that exhibit fluorescence properties, can be instrumental for gaining insights, at the molecular level, into the details of the associated fluorescence enhancement mechanisms. This would allow to elucidate and understand the factors that determine rhodopsin fluorescence and, in turn, learn how to modulate specific photophysical properties, with the ultimate goal of achieving the «in silico» design of highly fluorescent candidates. In this regard, the methodological and technical frameworks of the a-ARM protocol represent a suitable architecture for the development of computational tools, aimed at the automated and standard production/analysis of excited-state QM/MM models. In this Chapter we will see that, beyond their utility for predicting absorption properties, the S_0 QM/MM models generated with the a-ARM protocol, serve as a template for the further production of excited-states QM/MM models. More specifically, I will introduce the methodological development, computational implementation and benchmarking of a protocol, composed of three different phases, for the automated QM/MM model screening of rhodopsin variants displaying enhanced fluorescence.

Initial personal remarks

In the previous chapter I reported the methodological and computational improvements achieved, during the development of this doctoral Thesis, on the production of ground-state `ARM QM/MM models`. In this Chapter I will, instead, introduce the methodological and technical aspects[a] of a protocol designed for the production of excited-states QM/MM models suitable for the prediction of trends in light emission and light-induced dynamical properties. We will see how the proposed protocol allows for the screening of rhodopsin variants with enhanced fluorescence, with respect to a reference, possibly useful in optogenetics-oriented studies. The result of applying the protocol to two different systems exhibits a level of success that encourages further research in this direction. To the best of our knowledge, this is the first reported effort aimed at providing a computational tool for the automatic search of fluorescent proteins.

[a] Considering that, unlike the other chapters, the research work reported in this Chapter has not yet been published, a detailed explanation that includes methodological, technical and computational details, as well as scientific results, is provided.

Table 4.1: Experimental fluorescent properties for the Archaerhodopsin-3-based variants of the *application set*.[a]

Rhodopsin variant	Ref.	Maximum emission wavelength[b] (nm)	(kcal mol^{-1})	(eV)	ϕ^f
application set					
WT$^{\text{Arch3}}_{\text{AT}}$	[120]	687	41.6	1.80	1-9x10^{-4}
Arch5$^{\text{Arc3}}_{\text{AT}}$	[120]	731	39.1	1.70	8.7x10^{-3}
Arch7$^{\text{Arch3}}_{\text{AT}}$	[120]	727	39.3	1.70	1.2x10^{-2}
Archon2$^{\text{Arch3}}_{\text{AT}}$	[118]	735	38.9	1.69	1.0x10^{-2}
QuasAr1$^{\text{Arch3}}_{\text{AT}}$	[116]	715	40.0	1.73	8.0x10^{-3}
QuasAr2$^{\text{Arch3}}_{\text{AT}}$	[116]	715	40.0	1.73	4.0x10^{-3}
D95E/T99C$^{\text{Arch3}}_{\text{AT}}$	[120]	731	39.1	1.70	3.3x10^{-3}
D95E/T99C/P60L$^{\text{Arch3}}_{\text{AT}}$	[120]	731	39.1	1.70	4.0x10^{-3}
D95E/T99C/P196S$^{\text{Arch3}}_{\text{AT}}$	[120]	731	39.1	1.70	5.7x10^{-3}
D95E/T99C/V59A$^{\text{Arch3}}_{\text{AT}}$	[120]	728	39.2	1.70	6.2x10^{-3}

[a] The definition of *application set* will be given in Table 4.2

[b] experimental Maximum emission wavelength (λ^f_{max}), expressed in nm and eV and as first vertical emission energy (ΔE^f_{S1-S0}), in kcal mol^{-1}, along with the Fluorescence quantum yield (ϕ^f), unit-less.

As introduced in Section 1.1.3.1, specific microbial rhodopsins, exhibiting ion-transporting functions (*e.g.*, light-gated channels or light-driven pumps), have been found to be instrumental to the development of tools in optogenetics, an innovative technology for controlling biological activities with light.[10, 12, 14, 50, 112, 113] The related role of microbial rhodopsins can be assigned to as light-driven actuators (*i.e.*, action potential triggers), light-driven silencers (*i.e.*, action potential quenchers), and fluorescent reporters (i.e., action potential probes) of neuronal activity.[12] Regarding the latest point, particular light-driven ion-pumping microbial rhodopsins are employed to construct rhodopsin-based Genetically Encodable Voltage Indicator (GEVI) for optical imaging of the membrane voltage. In such application, a change in membrane voltage causes a variation in fluorescence intensity of the system, that is directly used as the voltage indicator.[14] The main advantages of using rhodopsin-based voltage sensors are: (i) they can be expressed in targeted neurons using genetic engineering, and (ii) it is possible to directly visualize the absolute membrane voltage with high temporal (500 μs to 40 ms) resolution even below the threshold potential (*i.e.*, the minimum potential that must be reached to initiate an action potential in neurons).[15]

From the above information, it is evident that rhodopsins used as GEVIs must exhibit intrinsic fluorescence behavior. Currently, the prototype fluorescent reporter is Archaerhodopsin-3, a WT archaeal rhodopsin from *Halorubrum Sodomense*, with light-driven outward proton pump activity.[114, 115] However, since the fluorescence of Arch3 is extremely dim featuring a Fluorescence quantum yield (ϕ^f) of ca. 1.1x10^{-4}, the design of better fluorescent reporters, with enhanced fluorescent properties, is required. In this regard, extensive efforts have been directed toward experimental engineering of novel Arch3-based variants (*i.e.*, via directed evolutionary approaches and random mutagenesis). For instance, a set of Arch3-based red-shifted variants with enhanced ϕ^f have been synthesized and tested as fluorescence probes. Among them, the most promising have been reported as: QuasArs,[116] Archons,[117, 118] Archers,[119, 120] Arch5 and Arch7[120] (see Table 4.1).

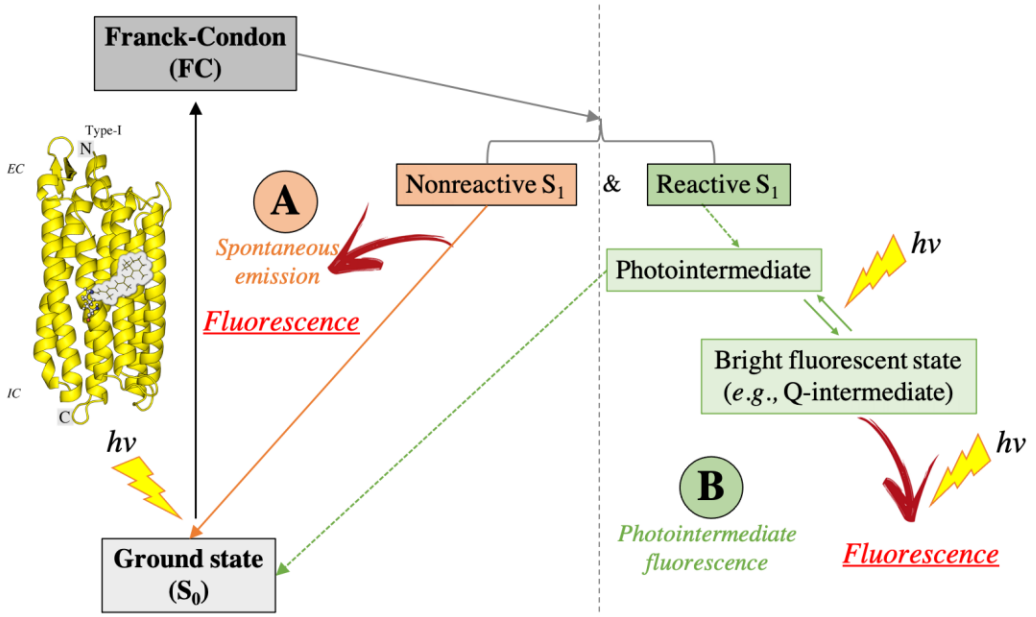

Figure 4.1: Photoreaction scheme of two potential fluorescence mechanisms for microbial rhodopsins. (A) *spontaneous emission*, and (B) *photointermediate fluorescence*. Spontaneous emission occurs from nonreactive and reactive S_1 state, while photointermediate fluorescence occurs from a bright fluorescent state (*e.g.*, Q-intermediate) that is produced by photon absorption of the precursor photointermediate (*e.g.*, N-intermediate).

These variants feature a brighter fluorescence than Arch3, enabling applications in imaging of thin brain slices, but also in living mammals and invertebrates.[121–123] Nevertheless, as shown in Table 4.1, their ϕ^f value are still in the range 10^{-3}-10^{-2} and therefore not yet as bright as desirable.

The use of of computational photochemistry and photobiology tools can be an alternative strategy to deal with the more rational design of rhodopsin variants, aimed at the actual modulation of desired fluorescence properties. Indeed, we have seen in section 1.1.3 that the computational simulation of photo-excited states of rhodopsins allows for the modeling of those light-emission properties that determine the fluorescence (see Figure 1.3). We also have seen in section 3.1.3 that it is possible to modulate photophysical properties (*i.e.*, color) by changing either the electrostatic or steric interactions of the *r*PSB and the protein environment, via the induction of point mutations. In the context of this doctoral Thesis, I expect that the thoughtful study (*i.e.*, via QM/MM computational modeling) of a set of fluorescent rhodopsins such as Arch3 and its derived variants, can be useful for gaining insights, at the molecular level, into their fluorescence enhancement mechanisms.

In this regard, two possible mechanisms for rhodopsin fluorescence have been proposed, based on experimental evidence. A schematic representation of both mechanisms is provided in Figure 4.1. As observed, the first scenario (orange path), from now on called mechanism A, hypothesize that the Franck-Condon (FC) state (see Section 1.1.3) generates the nonreactive S_1 state, which relaxes back to the original ground-state S_0 through *spontaneous emission*. In the second scenario (green path), from now on called mechanism B, the FC state generates, instead, the reactive S_1 state that initiates the photocycle and the emission

is produced from one photointermediate.

It is controversial which of these mechanisms explains the origin of the fluorescence used for GEVIs. For instance, while experimental evidence suggests that Arch3 fluorescence is dim because it does not come directly from the DA state but from a later photointermediate and three photons must be absorbed to generate one emitted photon (*i.e.*, mechanism B),[15, 124] it has been suggested that the enhanced fluorescence in neurons of its above mentioned red-shifted mutants is generated from the DA via one-photon process (*i.e.*, mechanism A).[116, 117] Additionally, one combined experimental and computational study recently carried out, in part, in our laboratory [49] suggests that the nonreactive S_1 state is correlated with the fluorescence of the possible GEVI *Anabaena* sensory rhodopsin (ASR), according with mechanism A (see Section 4.1). On the other hand, Kojima et al. have recently reported on an experimental comparative study of the fluorescence properties of 15 microbial WT rhodopsin-expressing neurons (*i.e.*, Arch3, ASR, GLR, GR, HwBR, IaNaR, KR2, MNaR, MR, NpHR, NpSRII, RmXeR, RxR, SyHR, and TR), comprising a spectroscopic analysis to elucidate the mechanism of fluorescence.[15] They proposed five rhodopsins, namely, GR, HwBR, IaNaR, MR, and NpHR, as new candidates for GEVIs, claiming that, similar to what reported for Arch3, the origin of their fluorescence observed in neurons is from the photointermediate fluorescence (*i.e.*, mechanism B). The latest information seems crucial for the development of both experimental and computational strategies aimed at the further improvement of fluorescent properties for microbial rhodopsin variants.

In this Chapter, I report on the blueprinting, implementation and benchmarking of a three phases computational protocol that largely automates the search of microbial rhodopsins with light-emission properties (*i.e.*, fluorescence). Our methodological framework (see Section 4.1) is specific for studying rhodopsin variants whose fluorescence is originated from mechanism A. The proposed protocol, called *a*-ARM rhodopsin fluorescence screening, is implemented into the `ARM` package as the `a_arm_fluorescence_searcher` driver (see Appendix A). It represents a "one-click" command-line architecture capable of generating a list containing the selected, potentially enhanced-fluorescent candidates, along with their excited-states S_1 QM/MM models. The three-phases protocol requires as input only a list of target rhodopsin variants, along with their ground-state equilibrium S_0 QM/MM models, and no further user decision/intervention.

From a computational chemistry perspective, any method or protocol has the purpose of reaching predictive capabilities. This is also true for a protocol aimed to automatically screen large arrays of rhodopsin, with the final scope of ranking them in terms of dim or enhanced fluorescence. This notwithstanding, such standard and automatic protocol will suffer from shortcomings, due to the adopted, necessarily approximated, computational methodology. In our case, the *a*-ARM rhodopsin model building [61, 62, 79], which provides the input QM/MM S_0 models to the `a_arm_fluorescence_searcher`, defines these limits (see Chapter 3). Thus, according with the philosophy of *a*-ARM, the research effort reported

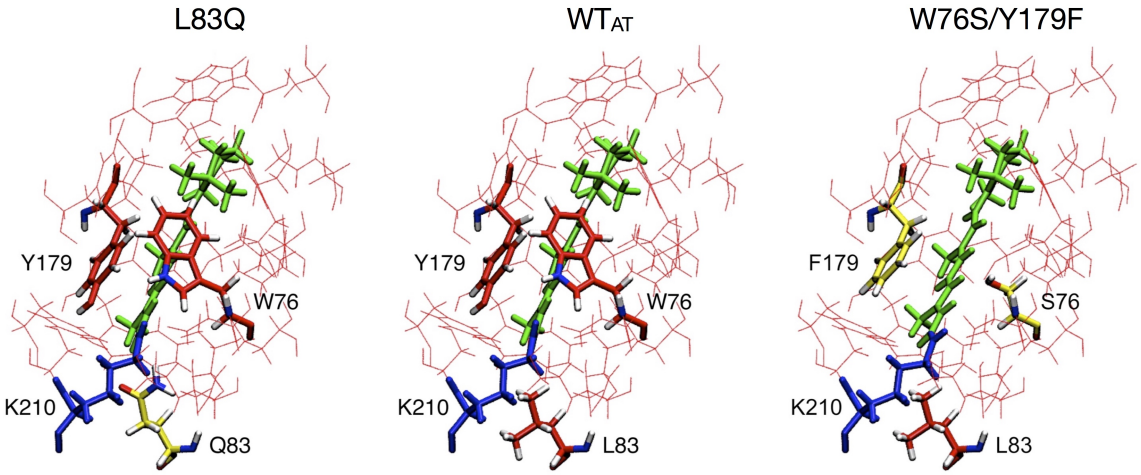

Figure 4.2: Cavity residues of the S_0 QM/MM models of ASR$_{AT}$ and two of its mutants, studied in Ref. 49 with the *original* ARM protocol. The amino acids of the chromophore cavity are represented by a thin framework representation. The variable cavity residues 76, 83 and 179 are shown in tube representation. The *r*PSB chromophore and lysine side-chain (K210) covalently linked to the chromophore are also shown in tube representation. Reprinted with permission from Marín et al.[49]. Copyright 2019 American Chemical Society.

here is limited to predicting and reproducing experimental *trends* in excited-states photochemical/photophysical properties, instead of attempting to reproduce the corresponding absolute values.

4.1 Methodological framework

The workflow defining the *a*-ARM rhodopsin fluorescence screening protocol, implemented through the `a_arm_fluorescence_searcher` driver, takes its inspiration[1] from the different analyses performed by Marín et al. in Ref. 49, who investigated, both experimentally and theoretically (in part in our laboratory), the mechanism explaining both the enhanced or diminished fluorescence displayed by specific blue-shifted mutants of Anabaena Sensory Rhodopsin (ASR), a light sensor from the fresh water eubacterium *Anabaena*. The importance of this microbial rhodopsin relies in the fact that it exhibits a dim fluorescence similar to Arch3, but has only a weak (inverse) proton pumping activity. Moreover, it exists in two forms; all-*trans* ASR (ASR$_{AT}$) and 13-*cis* ASR (ASR$_{13C}$), which can be inter- converted with light of different wavelengths.[80] Such bistability (*i.e.*, photochromism) is an attractive feature for optogenetics, since it provides the basis for engineering photoswitchable fluorescent probes.

The contribution of Marín et al. was focused on supporting and explaining a set of experimental measurements of fluorescence ϕ^f and Excited state lifetime (ESL) of two variants with respect to wild-type ASR$_{AT}$ (WT$_{\mathrm{AT}}^{\mathrm{ASR}}$) (see Figure 4.2). It was observed that, whereas the double mutant W76S/Y179F$_{\mathrm{AT}}^{\mathrm{ASR}}$ displays a nearly 10-fold increase in red-light emission (*i.e.* enhanced fluorescence with respect to WT$_{\mathrm{AT}}^{\mathrm{ASR}}$), the single mutant L83Q$_{\mathrm{AT}}^{\mathrm{ASR}}$ is not emis-

[1] During the design of our protocol, we have to deal with some methodological and technical pitfalls of the strategy employed by Marín et al. (see Section 5 of paper **[III]**)

sive (*i.e.* dimmer fluorescence than WT_{AT}^{ASR}).[49] This opposite fluorescence behavior makes these two mutants suitable to characterize the photochemical features of both emissive and not emissive rhodopsins.[49] Remarkably, the work showed that the excited-state barrier for chromophore bond rotation (E_{S1}^{f}) (see Figure 1.3), controls the competition between fluorescence and photoisomerization, since the computed E_{S1}^{f} magnitude appears to correlates with the magnitude of the experimental ϕ^{f}. More importantly, the authors proposed a mechanism that justifies the ESL and fluorescence ϕ^{f} enhancement of blue-shifted variants based on an electronic state mixing between the first S_1 and second S_2 excited state of the molecule.

Considering that their computational strategy has been demonstrated to be suitable for reproducing the observed trends in photophysics properties such as absorption (λ_{max}^{a}), emission (λ_{max}^{f}), and (indirectly) ESL (*e.g.*, through the characterization of energy barriers to ground- and excited-state photoisomerization), one natural step forward is to translate all of this into a protocol for modeling the mechanism which modulates the rhodopsin fluorescence, not only as an interpretative but also as a predictive tool.

With the aim of establishing an appropriate pipeline for the automatic screening of fluorescent variants, we have designed and implemented three phases, inspired by the analyses carried out by Marín et al. In the following, I will explain each of these phases, using as explanatory aids the set of blue-shifted ASR mutants presented in Figure 4.2, along with the main findings reported in Ref. 49.

4.1.1 Phase I): Location of the first excited state (S_1) minimum

In the first part of the protocol (Phase I), we are interested in answering the question whether a target microbial rhodopsin exhibits a fluorescent excited state planar minimum (PLA) structure, located close to the FC point on the S_1 PES (see Figure 1.3) or not. Such information provides a first, relatively fast, filter to determine if a rhodopsin can be considered as a potential fluorescent candidate or, otherwise, it should be immediately discarded. The simplicity of this analysis, in terms of computational time and resources, represents an advantage when large arrays of rhodopsins are studied in parallel.

The general workflow of Phase I, in terms of either the employed QM/MM calculations and the criteria used to classify the rhodopsin variants as potential candidates, is presented in Figure 4.3, while a more specific/technical one (*i.e.*, including required input/output files, calculations, computational details) is provided in Figure A.4. As observed in these figures, the only required input is a list of rhodopsin variants, along with their representative *a*-ARM ground-state S_0 equilibrium geometry[2], that corresponds to the replica with λ_{max}^{a} closest to the average value of N=10 replicas. [61, 62, 79] The output is a list of potentially fluorescent candidates, along with their PLA geometry and λ_{max}^{f}. Once all the required input files (see Section A.2.1.3) for each rhodopsin are in the same folder, the user has to

[2]Optimized at the single state CASSCF(12,12)/AMBER/6-31G(d) level.

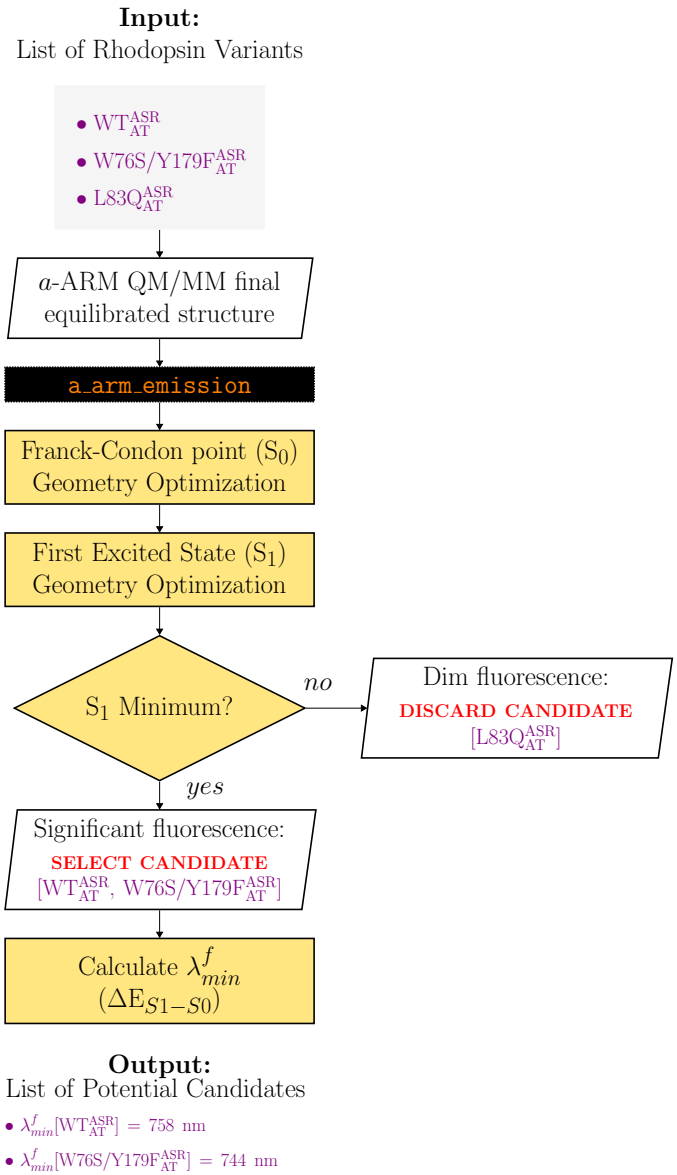

Figure 4.3: Phase I of the *a*-ARM rhodopsin fluorescence screening. Location of the first excited state minima. The procedure starts with the ground state (S_0) *a*-ARM QM/MM model of the target rhodopsin and, if a minimum structure for the first excited state (S_1) is located (PLA), ends with its generated PLA QM/MM model along with the calculated maximum absorption (λ^{a}_{max}) and emission (λ^{f}_{max}) wavelengths.

launch the module by typing the command line `a_arm_emission` (see Figure A.5) to start the procedure, which then operates as a fully automated tool. A detailed explanation on the `a_arm_emission` module structure and usage is provided in Section A.2.2.2.

In the following, I will explain, qualitatively, how the `a_arm_emission` module works, using as an example the set of ASR variants studied by Marín et al.[49] In this regard, the input list of rhodopsin variants is composed of three rhodopsins: WT^{ASR}_{AT}, $W76S/Y179F^{ASR}_{AT}$ and $L83Q^{ASR}_{AT}$ (see Figure 4.3). For each rhodopsin, the FC point located with *a*-ARM is re-evaluated by re optimizing the input S_0 optimized geometry, but employing a state-averaged[3] (*i.e.*, same weights) *n*=2-root SA-CASSCF(12,12)/AMBER/6-31G(d) level, al-

[3]In State-Averaged (SA) CASSCF calculations, one single set of molecular orbitals is used to compute

though following the gradient of the S_0 state, as described in section A.2.2.1. Commonly, this implies a negligible change in geometry with respect to the *a*-ARM structure computed with single state CASSCF (see Section 2.2.2). Then, the updated FC point is used as the starting structure to search for the first excited state PLA minimum along the S_1 Potential Energy Surface (PES), by performing a geometry optimization following the gradient of the S_1 state. These geometry optimizations are performed by using the microiterations technique (see section 2.2.2.2 and Ref. 125). Subsequently, the convergence of this calculation is evaluated by using the following criteria: if the geometry optimization calculation reached convergence within 100 steps, the rhodopsin exhibits a PLA structure and is considered as a potential fluorescent candidate and it continues to Phase II (section 4.1.2); otherwise, the rhodopsin is discarded. In our example, as illustrated in Figure 4.3, $L83Q^{ASR}_{AT}$ (*i.e.*, no located PLA) is discarded whereas WT^{ASR}_{AT} and $W76S/Y179F^{ASR}_{AT}$ (*i.e.*, located PLA) are used to build the list of potentially fluorescent candidates. Finally, according with the philosophy of *a*-ARM, the energy of the PLA structure is corrected at the 3-roots CASPT2/6-31G(d) level and the λ^{f}_{max} is calculated (see Figure 1.3) for each of the potentially fluorescent candidates.

4.1.2 Phase II): Computation of Quantum-Classical Franck-Condon (FC) trajectories

In the second part of the protocol (Phase II), I attempt to investigate, qualitatively, the S_1 PES driving the structural and electronic evolution of the *r*PSB chromophore for a target rhodopsin, as a function of time (see Figure 1.3). In particular, we are interested in identifying whether or not the S_1 excited state PES features a barrierless (or nearly barrierless) decay from the FC region to the Conical intersection (CI), by using S_1 trajectory calculations.

It is well known that properties of the PES (*e.g.*, reaction coordinate, slope and electronic character) may be controlled by the interactions of the *r*PSB with the protein environment (*i.e.*, amino acid residues forming the chromophore cavity).[126] Theoretically, in the kind of highly fluorescent rhodopsins we are searching for, such interactions may slow down or even stop the photoisomerization of the retinal. Commonly, an energy barrier is generated between the FC point and the CI, allowing the characterization of an elusive fluorescent excited state intermediate in the context of nonadiabatic dynamics.

Otherwise, in the case of dim fluorescent proteins such interactions instead accelerate the photoisomerization of the retinal, allowing an ultrafast chemical reaction in which the S_1 PES must feature a barrierless reaction path connecting the FC point to the CI. Thus, in principle, the decay time of S_1 into S_0 estimated via trajectory computations provides

a number of states of a given spatial and spin symmetry. The obtained density matrix is the average for all included states, although each state will have its own set of optimized configuration interaction (CIn) coefficients. The use of a SA-CASSCF procedure has a number of advantages, *e.g.*, all states in a SA-CASSCF calculation are orthogonal to each other.

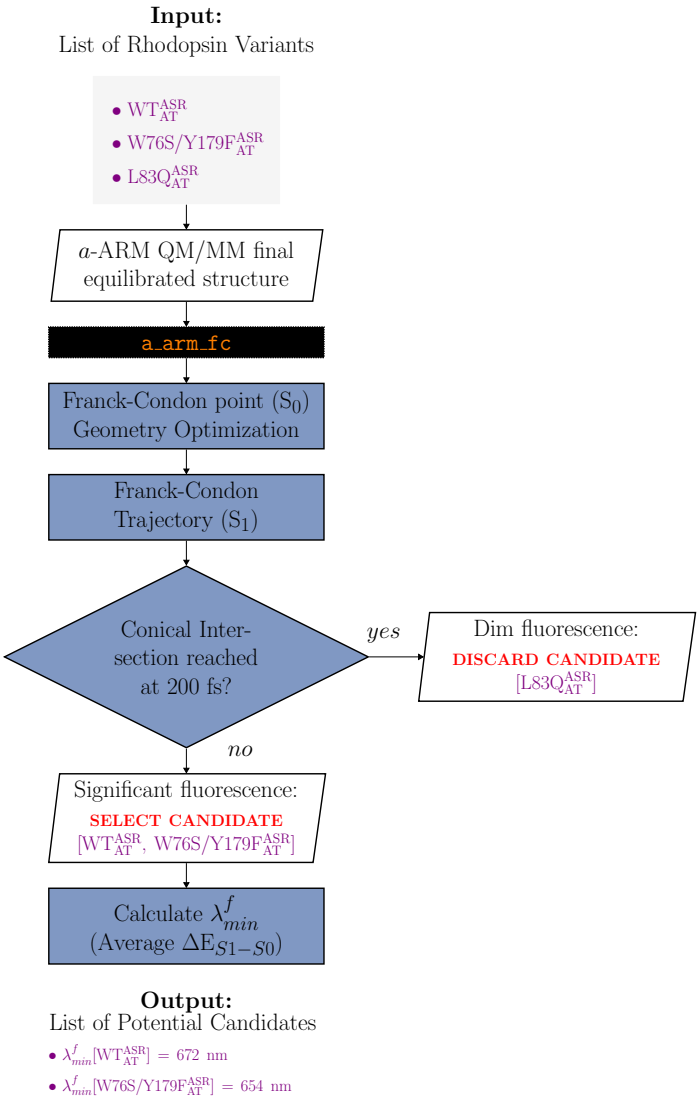

Figure 4.4: Phase II of the *a*-ARM rhodopsin fluorescence screening. Computation of semi-classical Franck-Condon (FC) trajectories. The procedure starts with the ground state (S_0) *a*-ARM QM/MM model of the analyzed rhodopsin and ends with the representation (graphics and tables) of the FC trajectory energy profile re-evaluated at the *n*-roots SA*n*-CASSCF(12,12)/AMBER//CASPT2(12,12)/6-31G(d) level, along with the calculated maximum absorption (λ^{a}_{max}) and emission (λ^{f}_{max}) wavelengths.

useful information about the fluorescent character of the rhodopsin. Based on previous studies,[19, 47, 49, 126] we define a threshold of 200 fs for the decay time to categorize rhodopsins as having dim fluorescence.

The general workflow of Phase II is presented in Figure 4.4, while a more specific/technical one (*i.e.*, including required input/output files, calculations, computational details) is presented in Figure A.6. Similar to what described for Phase I (Section 4.1.1), the only required input is a list of rhodopsin variants, along with their representative ground-state S_0 *a*-ARM QM/MM structure. The output is a list of potentially fluorescent candidates, along with the corrected λ^{f}_{max} (*i.e.*, considering the kinetics energy), this time calculated as the average ΔE_{S1-S0} along the FC trajectory.[49] The module driving Phase II starts by typing the command line `a_arm_fc` (see Figure A.7) along with the required command-line

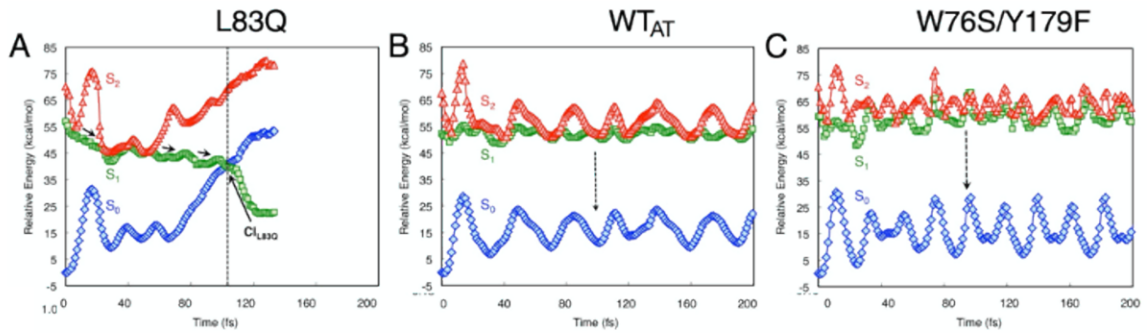

Figure 4.5: Franck-Condon (FC) trajectory computation on S_1 for $L83Q^{ASR}_{AT}$, WT^{ASR}_{AT} and $W76S/Y179F^{ASR}_{AT}$, respectively, computed in Ref. 49 with the *original* ARM.. (A)-(C) QM/MM FC trajectories computed at two-root state-averaged-CASSCF/AMBER level of theory and corrected at the CASPT2 level. S_0 (diamons), S_1 (squares) and S_2 (triangles) CASPT2//CASSCF/AMBER energy profiles along the FC trajectories. Adapted with permission from Marín et al.[49]. Copyright 2019 American Chemical Society.

arguments. Subsequently, the module operates as a fully automated tool providing the final graphical representation of the FC trajectory. A detailed explanation on the `a_arm_fc` module structure and usage is provided in Section A.2.2.3.

In order to illustrate how the `a_arm_fc` works, once again we use as an example the set of ASR rhodopsins of Marín et al.[49] described above. As observed in Figure 4.4, the first part of the procedure (*i.e.*, characterization of the structure corresponding to the FC point for each rhodopsin) is quite similar to that of Phase I; however, in this case the geometry optimization is performed without using the microiterations technique (see Ref. 125 for further details).Then, the FC point is employed as starting structure to compute the FC trajectory. When the FC trajectory calculation reaches a threshold time of 200 fs, the following criteria is used: if the CI has not been reached, the rhodopsin is considered as a potential fluorescent candidate and the FC calculation continues until it completes 500 fs and then pass to the Phase III (section 4.1.3); otherwise, the rhodopsin is discarded. In our example (see Figure 4.4 and Figure 4.5), $L83Q^{ASR}_{AT}$ reaches the photochemically relevant CI and decays to S_0 in ca. 100 fs (see Figure 4.5A) and is discarded, whereas WT^{ASR}_{AT} (see Figure 4.5B) and $W76S/Y179F^{ASR}_{AT}$ (see Figure 4.5C) are not reactive and are used to build the list of potentially fluorescent candidates. Please note that, in the context of the general use of the `a_arm_fluorescence_searcher` procedure, $L83Q^{ASR}_{AT}$ would have already been discarded during Phase I.

4.1.3 Phase III): Calculation of the excited state reaction path along the photoisomerization coordinate

In the last part of the protocol (Phase III), I simulate the excited-stated photo-isomerization path of the *r*PSB (in protein environment and in vacuum), along the torsional motion of the reactive double-bond, in order to gain further information into the ESL and, consequently, into the fluorescent character of the target rhodopsin. It has been reported that the ESL in rhodopsins is mainly determined by the S_1 reactivity of the retinal chromophore,[49] being

the photo-isomerization motion on S_1 responsible for the subpicosecond ESL and the low ϕ^f. Therefore, the presence of an energy barrier E^f_{S1} along the S_1 CASSCF isomerization path is a manifestation of the fluorescent character of the rhodopsin (See Figure 1.3). Moreover, both the ESL and ϕ^f are quantities proportional to the E^f_{S1} magnitude. It means that, when studying a large array of rhodopsins, the magnitude of the E^f_{S1} can help to categorize the items in terms of dim- or enhanced-fluorescence. Consistently, in Phase III of the protocol I focus the discussion on the characterization of the presence/absence of such an energy barrier on the S_1 PES along the torsional motion, as the last filter to determine if a rhodopsin can be considered as a potential fluorescent candidate or, otherwise, if it should be immediately discarded.

I stress that, although for the purposes of this work we are only interested on the characterization of E^f_{S1}, the analysis of the computed photo-isomerization paths provide relevant information on both the electronic states and structural changes driving the photo-isomerization of the retinal. Such information can be used to further investigate and rationalize the mechanism of fluorescence enhancement (see for instance Refs. 126 and 49).

The general workflow of Phase III is presented in Figure 4.6, while a more specific one (*i.e.*, including required input/output files, calculations, computational details) is presented in Figure A.8. In this case, the only required input is a list of rhodopsin variants, along with their S_1 (*i.e.*, PLA structure) *a*-ARM QM/MM model, previously generated in Phase II (notice that unlike Phases I and II, here the ground-state S_0 structure is not required). The main output is a list of potentially fluorescent variants, along with the calculated E^f_{S1}. In addition, the output files include both a graphical representation and raw-data of the computed photo-isomerization path. This includes information not only on the energy profile, but also on complementary properties, such as: Mulliken charges calculated for the reactive fragment, oscillator strength, bond length alternation (BLA) and hydrogen-out-of-plane (HOOP).

The module driving Phase III starts by typing the command line `a_arm_relaxed_scan` (see Figure A.9) along with the required command-line arguments. Subsequently, the module operates as a fully automated tool providing the previously described output. A detailed explanation on the `a_arm_relaxed_scan` module structure and usage is provided in Section A.2.2.4.

With the aim of illustrating how the `a_arm_relaxed_scan` module operates, once again we use the same set of ASR variants. As observed in Figure 4.6, the procedure starts with the automatic identification of the retinal configuration, at the PLA structure, to define the torsional angle to be used as isomerization coordinate during the computation of the photo-isomerization path, from now on called relaxed scan (RS) (see Section A.2.2.4). Simultaneously, the structural parameters of the PLA point are calculated to define the first point of the RS.

Starting from this structure, a series of subsequent constrained optimizations are carried

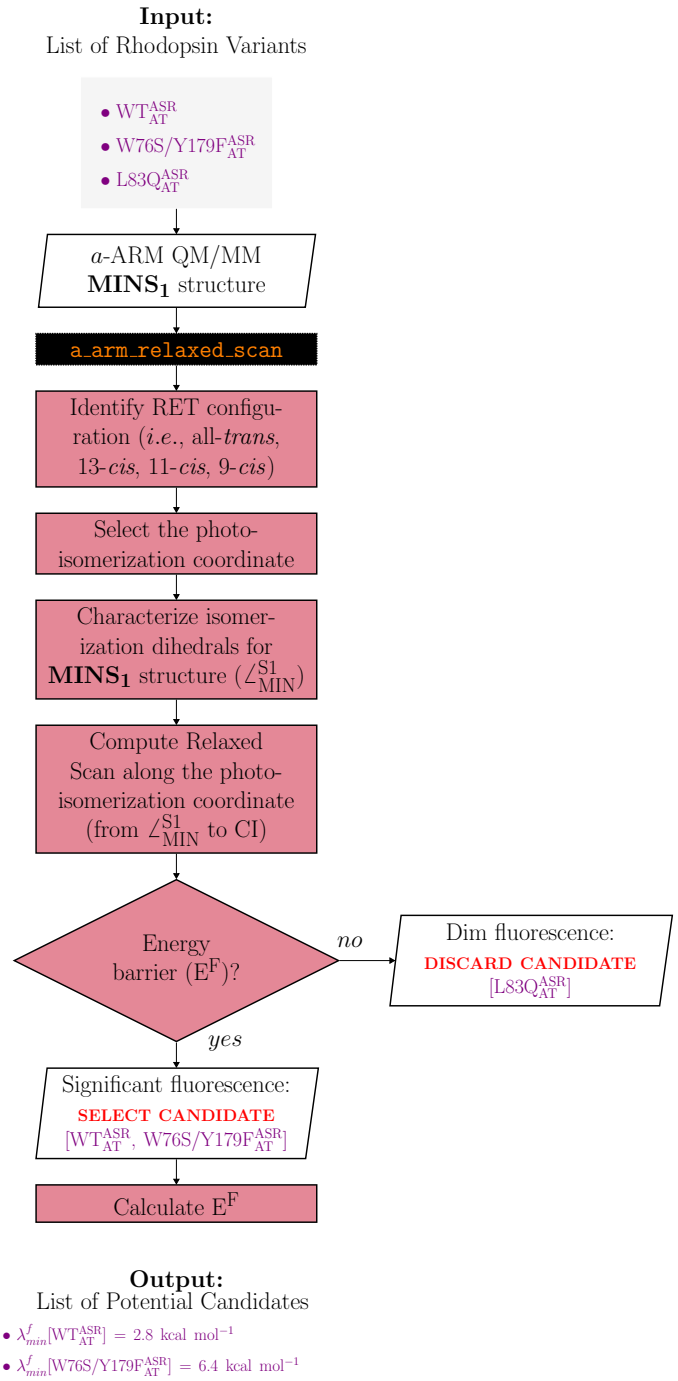

Figure 4.6: Phase III of the *a*-ARM rhodopsin fluorescence screening. Calculation of the excited state reaction path along the photoisomerization coordinate. The procedure starts with the first excited state minimum (S_{1min}) *a*-ARM QM/MM model of the target rhodopsin and ends with its relaxed scan along the photo-isomerization coordinate.

out, to explore the photo-isomerization characteristic of the given *r*PSB, employing a step size of 5 degrees for the change in torsional angle. The torsional angle is changed until it reaches a value of |90| degrees (*i.e.*, 90 for clockwise and -90 for counterclockwise rotation motion), around which the S_0/S_1 CI is, commonly, located. Further details are given in section A.2.2.4. Afterwards, the following criteria is used: if there is an energy barrier, the rhodopsin is considered as a potentially fluorescent candidate and the E^f_{S1} is calculated;