Supplementary Information to "Modeling pH-dependent biomolecular photochemistry"

Elisa Pieri,∗,† Oliver Weingart,‡ Miquel Huix-Rotllant,† Vincent Ledentu,† Marco Garavelli,¶ and Nicolas Ferré∗,†

† Aix-Marseille Univ, CNRS, Institut de Chimie Radicalaire, Marseille, France

‡ Heinrich Heine University, Faculty of Mathematics and Natural Sciences, Institute for Theoretical and Computational Chemistry, Universitätsstr. 1, 40225 Düsseldorf, Germany.

¶ Dipartimento di Chimica Industriale Toso Montanari, Università degli Studi di Bologna, Viale del Risorgimento, 4, 40136 Bologna, Italy

E-mail: elisa.pieri@univ-amu.fr; nicolas.ferre@univ-amu.fr

Table of Contents

1. Computational details

As mentioned in the main text, the number of available microstates, here understood as combinations of individual protonation states, for a macromolecule featuring *x* titratable sites is larger or equal to 2^x. In the case of ASR, the size of this protonation state space is equal to 2^9 (aspartic acid, D) * 2^5 (glutamic acid, E) * 3^4 (histidine, H) * 2^3 (cysteine, C) * 2^{11} (tyrosine, Y) * 2^6 (lysine, K) * 2^2 (N- and C-terminal), i.e., larger than 10^{12}! We can exclude C, K, Y, R and the terminal residue titrations, since we are mainly interested in pH range 3 to 7. Nevertheless, this reduced space still contains 1327104 microstates, a computationally intractable number with the present resources. Based on our previous study regarding the pH-dependent visible light absorption spectrum of ASR[1], we have decided to consider three different pH windows (3.0-4.5, 4.5-6.0, 6.0-7.5), each of them featuring a reduced set of titrated sites. The same list is used for both retinal conformations, *all-trans* (AT) or *13-cis* (13C). The list of the titratable amino acids in ASR, and their protonation state in each window, i.e., protonated (P), deprotonated (D), titrated (T), is reported below; K, R, Y and C residues are always protonated in our simulations.

Table S1. Protonation state of the HIS, ASP and GLU residues in the protein during the CpHMD in different pH windows. "T" stands for titrated, "P" for protonated and "D" for deprotonated.

Residue	pH=3.0	pH=5.0	pH=7.0
D57	T	T	D
D75	D	D	D
D98	T	T	D
D120	T	D	D
D125	P	T	D

D166	D	D	D
D198	D	D	D
D217	P	T	T
D226	P	T	D
E4	P	T	D
E36	P	P	T
E62	T	T	D
E123	P	T	D
E160	P	T	D
H8	P	T	D
H21	P	P	T
H69	P	T	D
H219	P	T	T

The full details for the system setup from crystal structure to production can be found in the SuppInfo of our previous work[1]. We got the initial structures (both isomers) from the PDB entry 1XIO. We selected the first monomer in the PDB entry and reconstructed missing loops through homology modeling. After an initial minimization, gradual heating NVE and equilibration in the NPT ensemble, we performed 20 ns long (or 30 in the case of the 4.5-6.0 pH window to improve convergence) CpHMD in implicit solvent and pH-REMD using Amber16; the distance between replicas is 0.5 pH units. The system was modeled with the ff14SB Amber forcefield for the protein, TIP3P for water and custom retinal parameters from Hayashi et al.[2] We calculated $pK_{1/2}$ values and used them as proxy for pKa values by fitting the deprotonated fractions using a Hill equation.

One thousand snapshots, each consisting of a geometry and a distribution of charges representing the corresponding protonation microstate, were selected per isomer and per pH value (3, 5 and 7), as detailed in the main text. In summary, we built each ensemble of 1000 structures to reproduce the corresponding ASR visible absorption spectrum already obtained in our previous work[1] using a much larger number of structures (20,000) and PM7 to treat the electronic structure of retinal.

Using these 1000 structures per pH value as initial conditions, we performed excited state semi-classical MD simulations using COBRAMM 2.0.[3] Initial distributions of C13=C14 dihedral angles are given below for all cases.

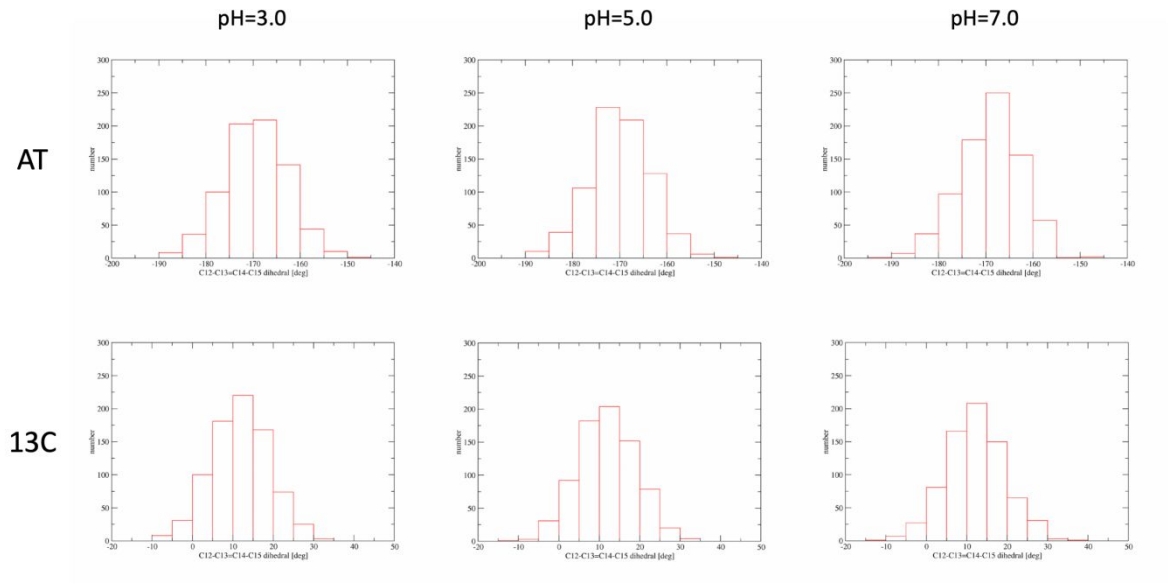


Figure S1. Histograms of the C12-C13=C14-C15 dihedral distribution among the 1000 initial conditions per set.

2.5 ps long trajectories have been propagated on hybrid quantum mechanical/molecular mechanical (QM/MM) potential energy surfaces at the semi-empirical OM3+MRCI level of theory[4-6] for retinal and Amber forcefield for the rest of the system. The validity of such a

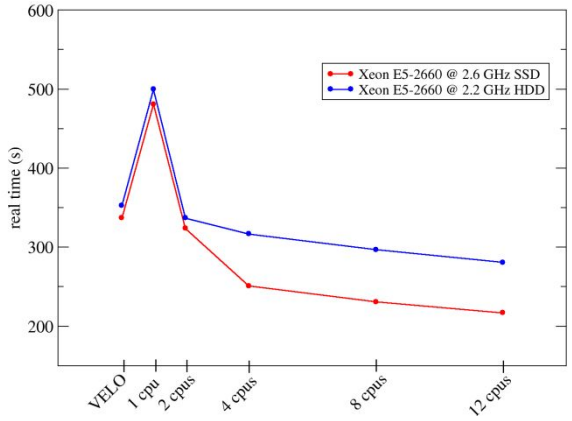
level of theory has been assessed by computing the ASR maximum absorption wavelengths at pH=3, 5 and 7.

The trajectory initial velocities are set to 0.0, hence creating one thousand ballistic trajectories for each retinal isomer and pH value. Retinal's Franck-Condon region is usually characterized by a steep S_1 potential energy surface.[7] Hence, the sampling of 1000 retinal structures performed by extracting snapshots from CpHMD trajectories is probably large enough to obtain a representative set of initial structures for non-adiabatic MD, the absence of initial velocities being compensated by the initial relaxation of the system driven by the different slopes of the Amber and OM3-MRCI/Amber potential energy surfaces. Of course, we could have used CpHMD velocities associated to each snapshot. However, to avoid large numerical instabilities, these velocities should have been transformed to adapt to the QM/MM potential energy surface (instead of the Amber one). This Amber to OM3-MRCI/Amber projection additional step would require getting access to the local topology of both the Amber and the QM/MM potential energy surfaces, i.e., it would be computationally expensive.

The resulting "wavepacket" is then deposited in the first singlet excited state, S_1, while the population transfer between electronic states (S_0, S_1 and S_2) is modeled with the Tully Surface Hopping technique[8,9] with decoherence correction.[10,11] To reduce the energy leaking towards the retinal environment, only the chromophore and its closest amino acids are free to move during the MD, while the rest of opsin and the membrane are kept fixed. Even if some trajectories are propagated up to 3.6 ps, we have considered populations in electronic states S_0, S_1 and S_2 up to 2 ps for analysis purpose. However, the trajectories were stopped 50 fs after hopping to S_0 to save computational resources.

2. Benchmark calculations

Prior to production runs, we performed a series of benchmark calculations to evaluate the performance of the routines used. The MNDO-program responsible for the QM-part is not parallelized, thus a gain in performance through parallel computation may only be expected for the MM-part. Figure S2 (left panel) shows benchmark computations for two platforms with different numbers of CPUs and with a modified force evaluation routine (VELO, see Fig. S2). Computations were performed for three gradients and three derivative couplings, defining the maximum of necessary gradient calculations when all three roots (S_0, S_1 and S_2) are included. The total computation time for one full QM/MM MD step then is ca. 480 – 500 seconds on the tested systems, where the QM step alone takes ca. 40% (190-200 s). As apparent from this figure, the speedup in total computation is only marginal beyond 2 CPUs for the system with a distributed file system: From 2 to 4 CPUs the time changes from 335 s to only 312 s. The setup with 2.6 GHz and local SSDs seems to profit from higher clock frequency and fast disk access when using up to 4 CPUs, but then the scaling drops. As a major reason for the rather slow performance, we identified a printing routine in Amber which provides the forces for the MM part (dumpfrc).



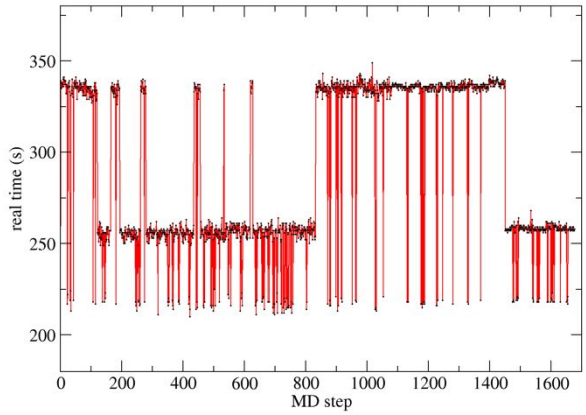


Figure S2. Left: Benchmark computations on Intel Xeon based systems with different clock frequencies and storage systems (local SSDs vs. distributed file system). Right: Full step QM/MM timings for a typical short running trajectory. Based on the state energy difference the computation switches between calculation of only one gradient and two gradients plus one coupling. Steps < 250 fs correspond to correction steps where either QM orbital mapping failed or energy conservation was violated – the time step is reduced in such steps and the computation is then repeated.

To reduce the MM timings, we propagated the MM part for one small timestep and evaluated the forces through finite differences in velocities between the two timesteps (stated as "VELO" in the graph, see this reference for more details[3]). This significantly sped up single-core computations, nearly reaching dual core performance, with the obtained forces at the same accuracy as by direct printout via dumpfrc. The final production runs were performed using the aforementioned setup with finite-difference computation of forces from velocities.

Figure S2 (right panel) shows timings for a typical trajectory on a system with a distributed file system. The total computation time for a full QM/MM step (single-core) varies between ca. 250 and 350 seconds, depending on the number of necessary gradient computations within the QM part (only one gradient or max. 2 gradients and one coupling computation in this trajectory).

For this setup, energy evaluation at the OM3/MRCI level typically takes ca. 5 s, computation of an excited state gradient ca. 30 s, evaluation of nonadiabatic coupling matrix elements ca. 48 s. A typical MM step (including at least three separate computations for high layer, charged and uncharged protein models) is produced in ca. 175 s. Another ca. 40 s are spent for file operations and computations within COBRAMM.

3. Analysis of each ensemble of trajectories

For each pH value and retinal isomer, 1000 trajectories have been produced. Some of them failed for various numerical reasons, like electronic structure calculation not

converged or total energy conservation not fulfilled during the MD simulation. Analysis of these trajectories did not yield any obvious geometric or electronic pattern for failure, we suspect that these arise due to the nature of the approximations in the OM3 and OM3/MRCI approaches. The percentage of valid trajectories in each ensemble is ~80%. The statistical relevance of our calculations is illustrated by the low uncertainties associated to the photochemical properties we are interested in.

Each ensemble of 1000 trajectories is split into numerically failed ones and valid ones. The latter trajectories are further decomposed, distinguishing the ones which don't decay to the ground state in 2 ps, the ones in which retinal successfully isomerizes, the ones in which retinal isomerization is aborted and the only one for which the retinal conformation remains undetermined (i.e., not AT or 13C) after 2 ps. We also split each ensemble of valid trajectories (denoted as a full set below) into several subsets, as indicated in the main text.

- Reactive: trajectories in which the isomerization around the C13=C14 bond is complete.
- Nonreactive: trajectories in which the isomerization around the C13=C14 bond is aborted.
- Alternative: trajectories in which the isomerization starts around bonds other than the C13=C14 one, and then gets aborted.
- Direct pathway: trajectories in which the initial population in S_1 directly transfers to the grounds state S_0.

- Indirect pathway: trajectories in which the initial population in S_1 first transfers to S_2 before turning back to S_1 and eventually transfers to the grounds state S_0.

Table S2. Statistical analysis of the ensembles relative to the AT → 13C isomerization.

AT isomer	**pH=3**	**pH=5**	**pH=7**
failed trajectories	184	168	170
valid trajectories	816 (768 hopped)	832 (764 hopped)	830 (787 hopped)
reactive	354 (43.4%)	326 (39.2%)	468 (56.4%)
unreactive	397 (48.7%)	438 (52.6%)	319 (38.4%)
undetermined	1 (0.1%)	0	0
no decay in 2000 fs	64 (7.8%)	68 (8.2%)	43 (5.2%)

Table S3. Statistical analysis of the ensembles relative to the 13C → AT isomerization.

13C isomer	**pH=3**	**pH=5**	**pH=7**
failed trajectories	139	182	205
valid trajectories	861 (810 hopped)	818 (768 hopped)	795 (740 hopped)
unreactive	529 (61.4%)	481 (58.8%)	451 (56.7%)
reactive	281 (32.6%)	287 (35.1%)	289 (36.4%)
undetermined	0	0	0
no decay in 2000 fs	51 (6.0%)	50 (6.1%)	55 (6.9%)

4. BLA and dihedral torsions in the 13C sets

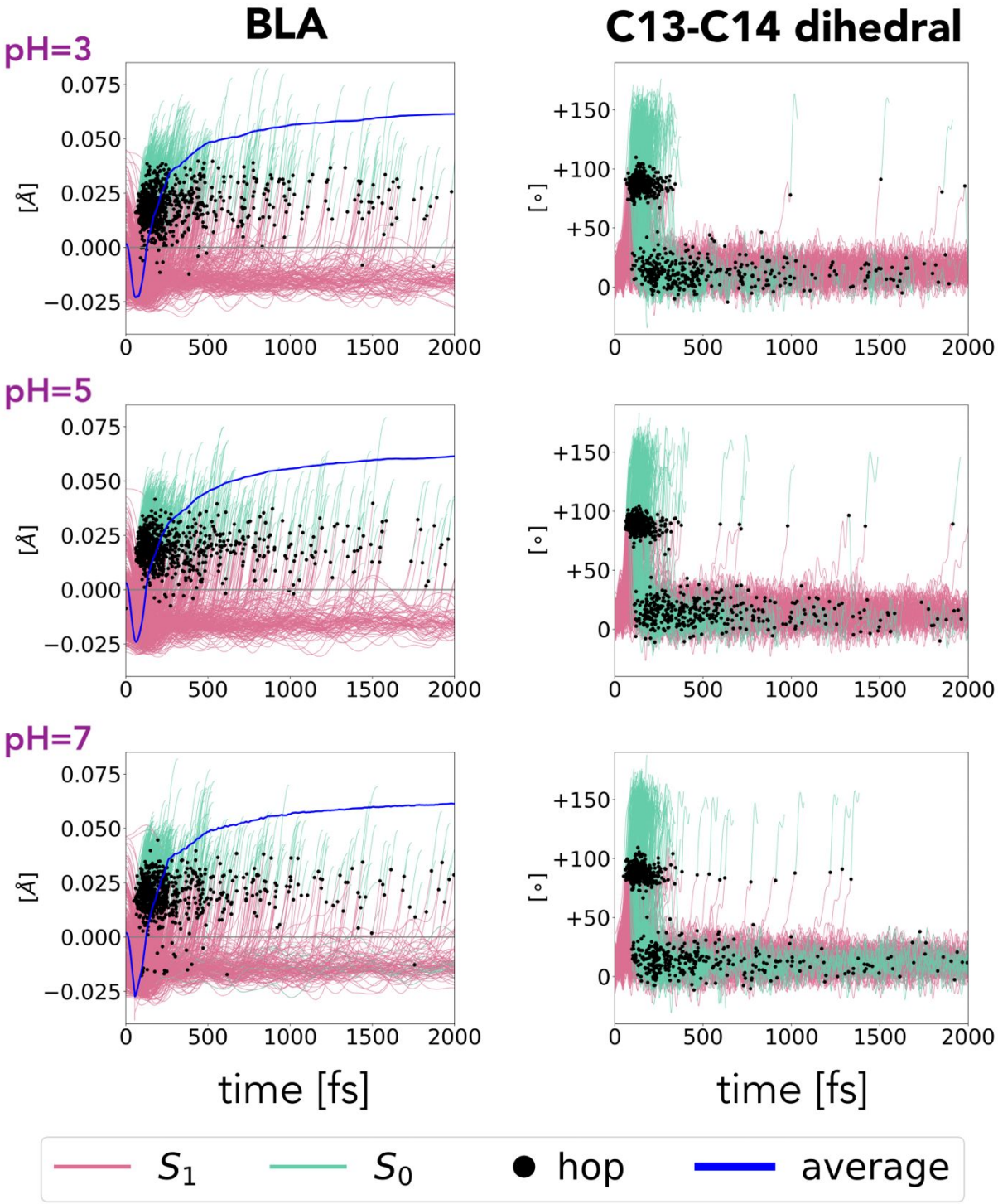


Figure S3. Time evolution of BLA (left) and torsion dihedral angle (right) during the 13C→AT retinal isomerization. S_1 (pink) and S_0 (green) parts of a trajectory are separated by a hop point (black circle). The BLA instantaneous average values are also plotted in blue. Please note that, since the trajectories are stopped shortly after reaching S_0, their last BLA value

is frozen for the reminder of the averaging to avoid noise and discontinuities. Isomerization quantum yields and hop times.

5. Isomerization quantum yields and hop times

The following tables report the isomerization quantum yield (IQY), calculated as the ratio between the number of trajectories in which the retinal isomerization is complete and the number of valid trajectories. The corresponding uncertainty is $\sqrt{IQY(1-IQY)/n}$. These tables also contain the average hop time for each set or subset (reported uncertainties are calculated as standard error of the mean), as well the corresponding number of trajectories which have hopped to the ground state.

Table S4. IQY, hop times and ensemble size for the set and subsets at pH=3 for the AT isomer.

pH3-AT	IQY (AT→13C)	Hop time (fs)	Set size
Full	0.43±0.02	379±14	816
Most pop. microstate	0.43±0.02	381±20	447
Most pop. charge state	0.42±0.02	379±19	478
Reactive	1	238±13	354
Nonreactive	0	286±26	177
Alternative	0	681±29	220
Indirect	0.46±0.05	410±38	115
Direct	0.43±0.02	373±15	637

Table S5. IQY, hop times and ensemble size for the set and subsets at pH=5 for the AT isomer.

pH5-AT	IQY (AT→13C)	Hop time (fs)	Set size
Full	0.39±0.02	480±18	832
Most pop. microstate	0.36±0.10	490±103	22
Most pop. charge state	0.41±0.02	461±31	253
Reactive	1	296±21	326
Nonreactive	0	293±30	117
Alternative	0	735±30	321
Indirect	0.46±0.04	512±48	124
Direct	0.38±0.02	474±19	640

Table S6. IQY, hop times and ensemble size for the set and subsets at pH=7 for the AT isomer.

pH7-AT	IQY (AT→13C)	Hop time (fs)	Set size
Full	0.56±0.02	364±16	830
Most pop. microstate	0.55±0.02	395±23	488
Most pop. charge state	0.56±0.02	380±21	561
Reactive	1	271±18	468
Nonreactive	0	259±27	154

Alternative	0	754±42	165
Indirect	0.63±0.04	332±32	172
Direct	0.54±0.02	373±19	615

Table S7. IQY, hop times and ensemble size for the set and subsets at pH=3 for the 13C isomer. The most populated microstate and the most populated total charge state perfectly overlap in this case.

pH3-13C	**IQY (13C→AT)**	**Hop time (fs)**	**Set size**
Full	0.33±0.02	340±13	861
Most pop. microstate	0.35±0.02	328±13	641
Most pop. charge state	0.35±0.02	328±13	641
Reactive	1	159±9	279
Nonreactive	0	181±18	157
Alternative	0	541±23	374
Indirect	0.39±0.04	329±34	140
Direct	0.31±0.02	342±14	670

Table S8. IQY, hop times and ensemble size for the set and subsets at pH=5 for the 13C isomer.

pH5-13C	**IQY (13C→AT)**	**Hop time (fs)**	**Set size**
Full	0.35±0.02	345±13	818

Most pop. microstate	0.40±0.11	329±97	20
Most pop. charge state	0.39±0.03	331±24	221
Reactive	1	155±9	287
Nonreactive	0	163±11	115
Alternative	0	550±22	366
Indirect	0.43±0.04	393±38	136
Direct	0.34±0.02	335±14	632

Table S9. IQY, hop times and ensemble size for the set and subsets at pH=7 for the 13C isomer.

pH7-13C	**IQY (13C→AT)**	**Hop time (fs)**	**Set size**
Full	0.36±0.02	348±15	795
Most pop. microstate	0.32±0.02	341±21	414
Most pop. charge state	0.33±0.02	342±19	493
Reactive	1	164±9	289
Nonreactive	0	158±5	118
Alternative	0	575±28	333
Indirect	0.50±0.04	318±37	126
Direct	0.33±0.02	354±17	614

6. Fitting S_1 decay time evolution

For each pH value, each isomer and each set (or subset) of trajectories, we have found a successful kinetic model that accurately fits the S_0 and S_1 population evolution with a small root mean square deviation between the model $S_1 \rightarrow S_0$ decay curve and the one coming out of the MD simulations. As already explained in the main text, the possible S_2 population during the early stages of the decay process led us to model the S_1 decay using a time window in which only S_0 and S_1 are populated. Accordingly, t_{start} is defined as the lower bound of this window and is always set as the last time step for which the ground state population is zero. The chosen model distinguishes two types of S_1 populations, one characterized by a fast kinetic rate (dubbed $P_{S1(fast)}$ for the population and k_{fast} for the rate) and a slower one ($P_{S1(slow)}$ and k_{slow}). We consider that both decay to the S_0 state with its own rate, but no interconversion between fast and slow S_1 populations is allowed. When we considered interconversion between both at fixed initial populations, we obtained an overall error superior to the model described hereafter. The kinetic equations are thus given by:

$$\begin{aligned}
\frac{dP_{S1(fast)}^{model}(t)}{dt} &= -k_{fast}P_{S1(fast)}^{model}(t) \\
\frac{dP_{S1(slow)}^{model}(t)}{dt} &= -k_{slow}P_{S1(slow)}^{model}(t) \\
\frac{dP_{S0}^{model}(t)}{dt} &= k_{fast}P_{S1(fast)}^{model}(t) + k_{slow}P_{S1(slow)}^{model}(t)
\end{aligned}$$

This system of differential equations is solved numerically using the python module *scipy.integrate.odeint*, from which we get the model populations at any discrete time *t*, namely, $P_{S0}^{model}(t)$, $P_{S1(fast)}^{model}(t)$ and $P_{S1(slow)}^{model}(t)$. The model populations depend on an initial

condition (populations at the t_{start}) and an initial guess of the kinetic rates. Instead of fitting these parameters at once, we have decided to take advantage of particular trajectory subsets: (i) the alternative subset of aborted isomerizations and (ii) the subset of successful and failed isomerizations around the C13=C14 bond. The former subset is associated to the slow decay while the latter subset corresponds to the fast component of the decay. It turns out that these two subsets feature a mono-exponential decay behavior that allow to obtain initial estimates for k_{fast} and k_{slow}. In a second step, keeping these k_{fast} and k_{slow} fixed, initial populations $P^{model}_{S1(fast)}$ and $P^{model}_{S1(slow)}$ for the full population decay are obtained using the bi-exponential model. These parameters are then optimized by minimizing the root mean square difference (RMSD) population of S_0 and S_1 between the model and the non-adiabatic dynamics, namely, $\frac{\partial RMDS}{\partial k} = 0$ and $\frac{\partial RMSD}{\partial P(t=0)} = 0$ with an error function definition given by

$$RMSD(\{k,P(t=0)\}) = \sqrt{\frac{1}{t_1 - t_0}\sum_{t=t_0}^{t_1}\left[(\Delta P_{S0}(t))^2 + (\Delta P_{S1}(t))^2\right]}$$

where $\Delta P_{S0}(t) = P^{model}_{S0}(t) - P^{NAMD}_{S0}(t)$ and $\Delta P_{S1}(t) = P^{model}_{S1(fast)}(t) + P^{model}_{S1(slow)} - P^{NAMD}_{S1}(t)$. The minimization is done in a double self-consistency, namely, first the kinetic rates are minimized at fixed initial populations (using python module *scipy.optimize.fmin*) and then the 0th time population are optimized at fixed kinetic rates (using python module *scipy.optimize.minimize*). In the latter minimization, a constraint is imposed during optimization, namely, all populations must be positive or 0 and sum up to 1. Finally, a last step is performed using the python *scipy.curve_fit* module using as model the following equation:

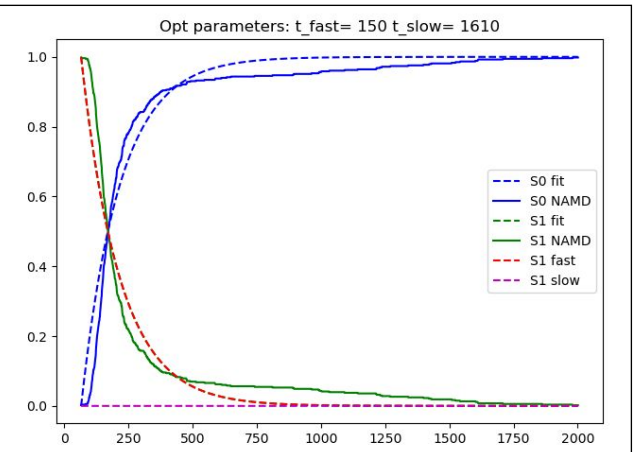
$$P_{S_1}(t) = a_1 e^{-t/\tau_{fast}} + a_2 e^{-t/\tau_{slow}}$$

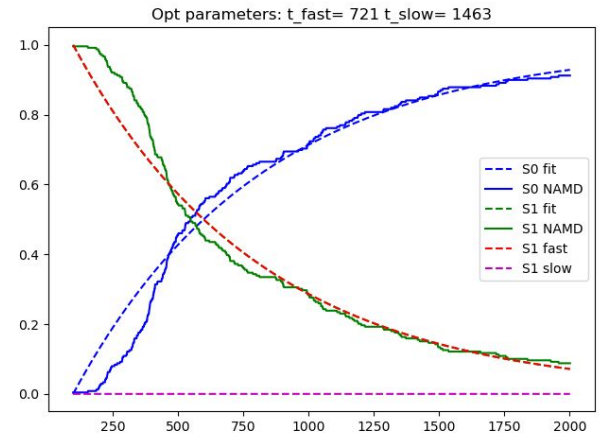
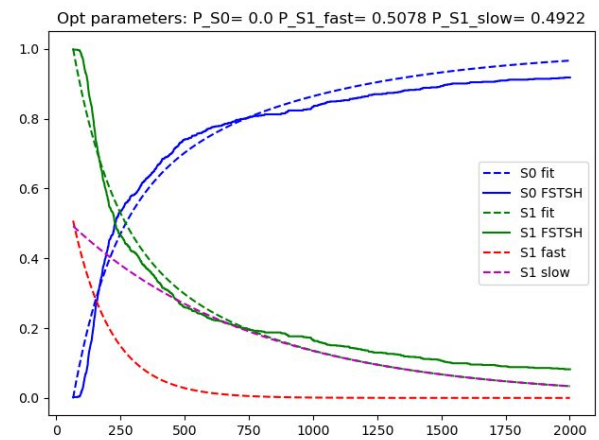
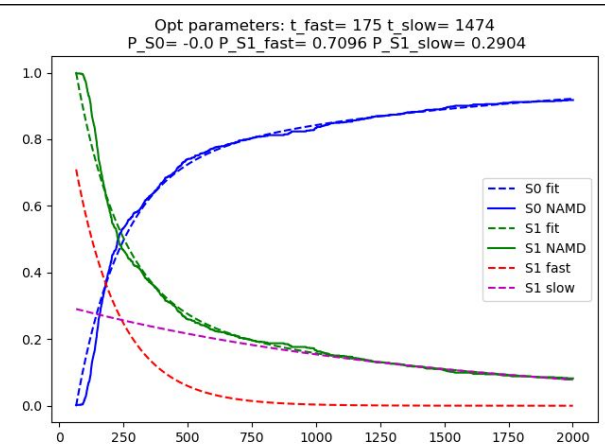
Initial parameters for the time constants τ_{fast} and τ_{slow} are simply the inverse of the decay rates obtained in the previous step. The ratio between a_1 and a_2 is set equal to the one between $P^{model}_{S1(fast)}(t = t_{start})$ and $P^{model}_{S1(slow)}(t = t_{start})$. Errors ($\Delta\tau_{fast}$, $\Delta\tau_{slow}$, Δa_1 and Δa_2) are calculated as the square root of diagonal elements in the covariance matrix. Since $P^{model}_{S1(fast)}(t = t_{start}) + P^{model}_{S1(slow)}(t = t_{start}) = 1$, their errors derive from the a_1 and a_2 ones:

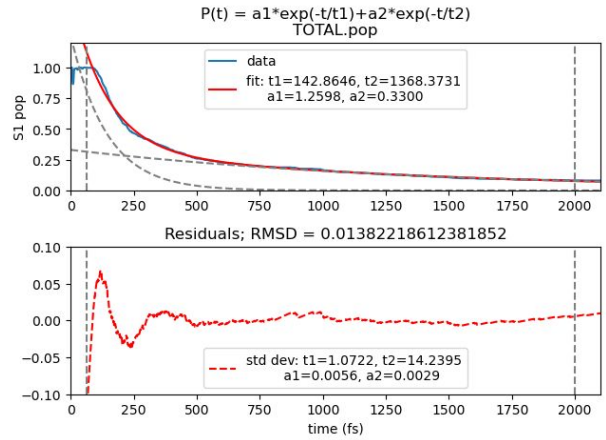
$$\Delta P^{model}_{S1(fast)}(t = t_{start}) = \left(\frac{\Delta a_1}{a_1} + \frac{\Delta a_2}{a_2}\right)\left(1 + \frac{a_1/a_2}{1 + a_1/a_2}\right) P^{model}_{S1(fast)}(t = t_{start})$$

$$\Delta P^{model}_{S1(slow)}(t = t_{start}) = \left(\frac{\Delta a_1}{a_1} + \frac{\Delta a_2}{a_2}\right)\left(\frac{a_1/a_2}{1 + a_1/a_2}\right) P^{model}_{S1(slow)}(t = t_{start})$$

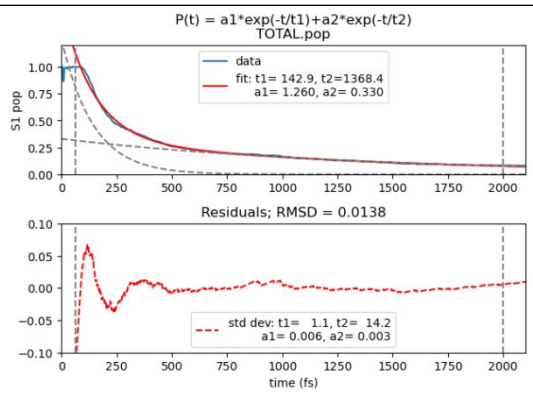
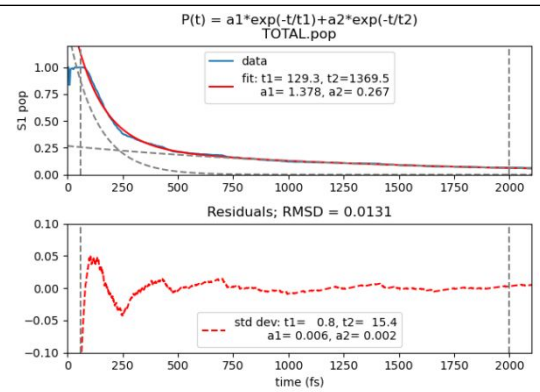
This workflow is illustrated below in the case of the pH3-AT model (plots have been generated with python matplotlib).

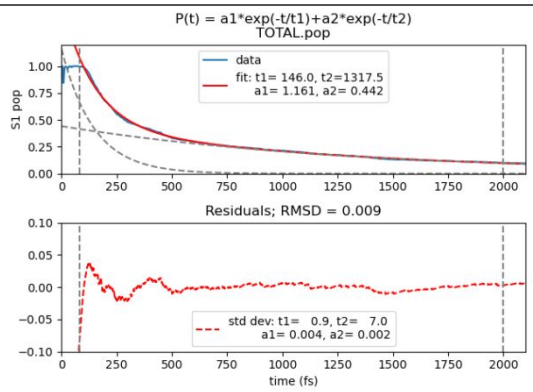
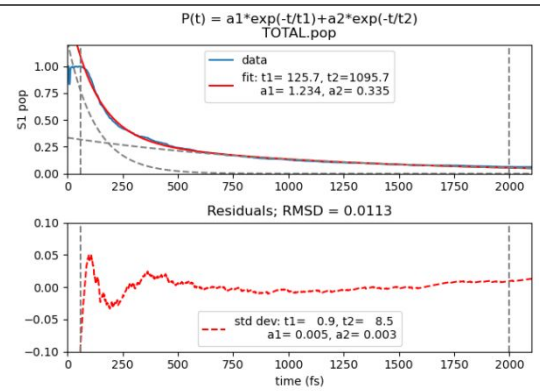
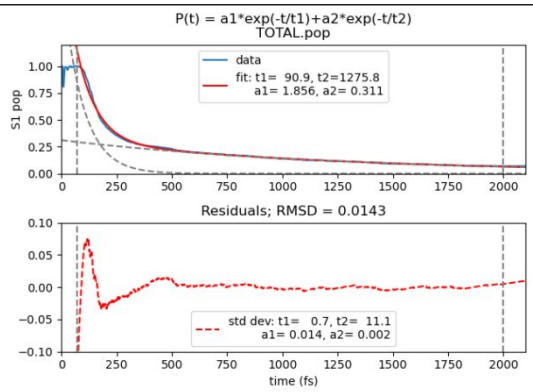
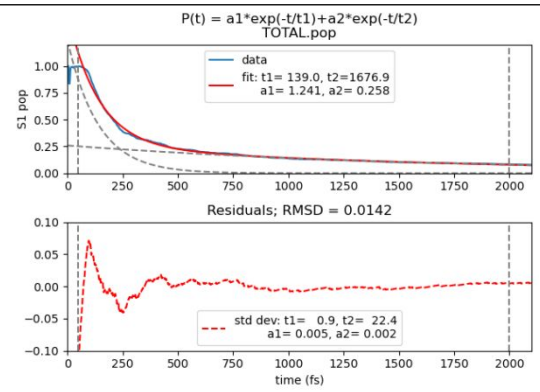
Step 1: guess of the rate k_{fast} The second exponential (in pink) does not contribute to the model -> initial value for τ_{fast} 150 fs	

Step 1': guess of the rate k_{slow} The second exponential (in pink) does not contribute to the model -> initial value for τ_{slow} 721 fs	
Step 2: guess of the rates $P^{model}_{S1(fast)}(t = t_{start})$ and $P^{model}_{S1(slow)}(t = t_{start})$ The two exponential model with fixed decay rates gives $P^{model}_{S1(fast)}(t = t_{start}) = 0.51$ and $P^{model}_{S1(slow)}(t = t_{start}) = 0.49$	
Step 3: parameter iterative optimization Improved parameters are now: τ_{fast} = 175 fs, τ_{slow} = 1474 fs, $P^{model}_{S1(fast)}(t = t_{start}) = 0.71$ and $P^{model}_{S1(slow)}(t = t_{start}) = 0.29$	

Step 4: final curve fitting for S_1 decay Final parameters are: $\tau_{fast} = 143 \pm 1$ fs, $\tau_{slow} = 1368 \pm 14$ fs, $P^{model}_{S1(fast)}(t = t_{start}) = 0.79 \pm 0.02$ and $P^{model}_{S1(slow)}(t = t_{start}) = 0.21 \pm 0.00$ for a global RMSD=0.0138 (Grey curves show individual contributions to the model, grey vertical lines define the time window used for the fitting).	

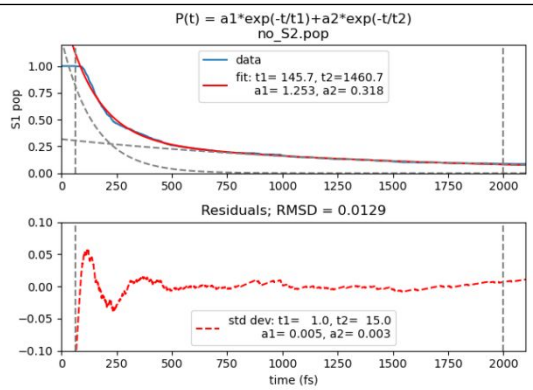
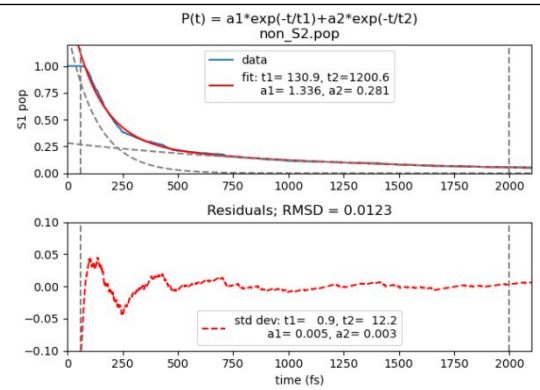
Below are reported the pictures of the final fitting step. NAMD-based S_1 populations are plot in blue line, while the red ones are coming out of the fitted models. Optimal parameters are also indicated, as reported in the main text.

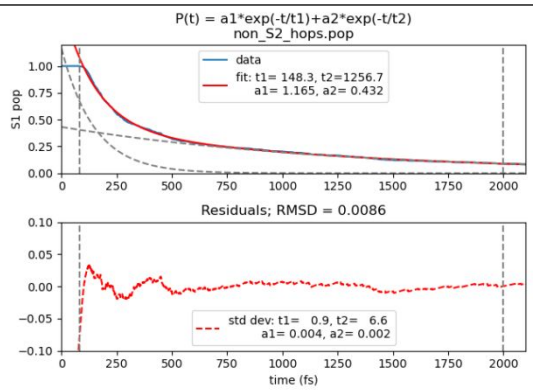
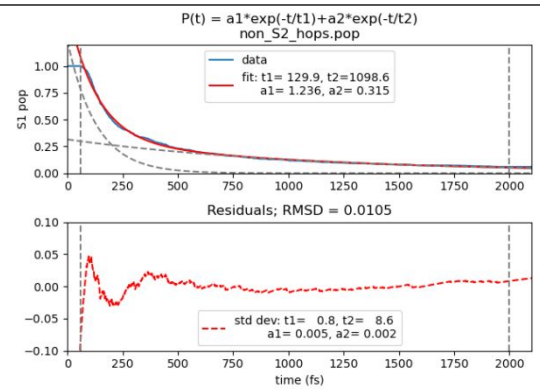
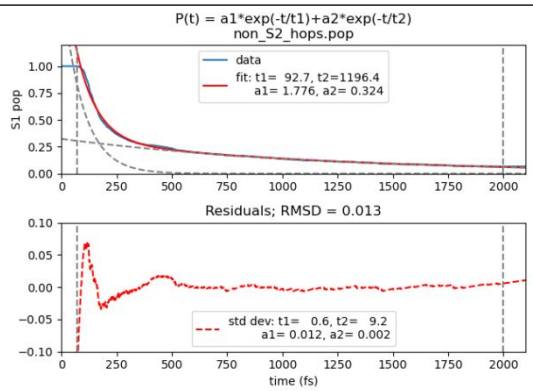
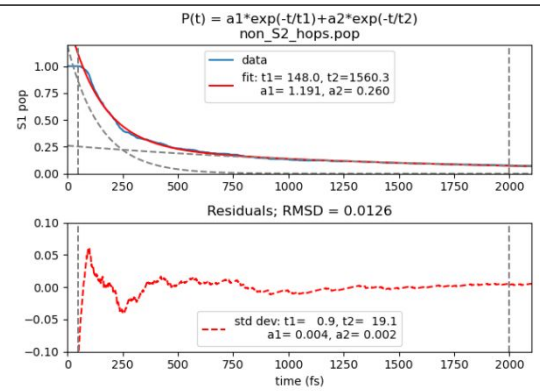
	AT retinal	13C retinal
pH=3	 t_{start} = 64 fs $\tau_{fast} = 143 \pm 1$ fs	 t_{start} = 60 fs $\tau_{fast} = 129 \pm 1$ fs $\tau_{slow} = 1369 \pm 15$ fs

	$\tau_{slow} = 1368 \pm 14$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.79 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.21 \pm 0.00$ RMSD = 0.0138	$P^{model}_{S1(fast)}(t = t_{start}) = 0.84 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.16 \pm 0.00$ RMSD = 0.0131
pH=5	 t_{start} = 81 fs $\tau_{fast} = 146 \pm 0$ fs $\tau_{slow} = 1318 \pm 6$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.72 \pm 0.01$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.28 \pm 0.00$ RMSD = 0.0090	 t_{start} = 61 fs $\tau_{fast} = 126 \pm 0$ fs $\tau_{slow} = 1096 \pm 8$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.79 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.21 \pm 0.00$ RMSD = 0.0113
pH=7	 t_{start} = 71 fs	 t_{start} = 49 fs

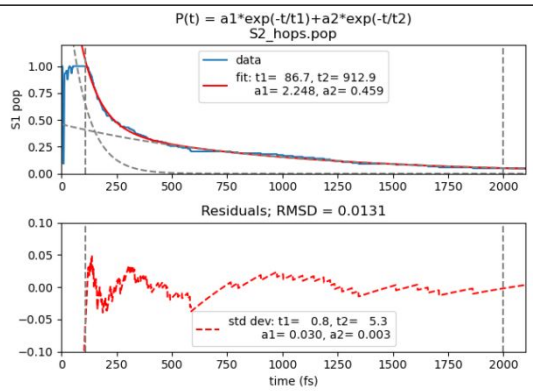
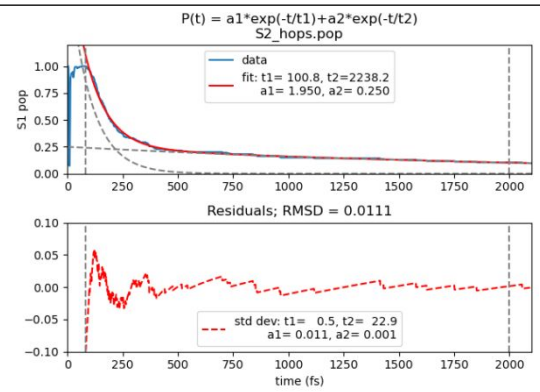
	$\tau_{fast} = 91 \pm 0$ fs $\tau_{slow} = 1276 \pm 11$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.86 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.14 \pm 0.00$ RMSD = 0.0143	$\tau_{fast} = 139 \pm 0$ fs $\tau_{slow} = 1677 \pm 22$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.83 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.17 \pm 0.00$ RMSD = 0.0142

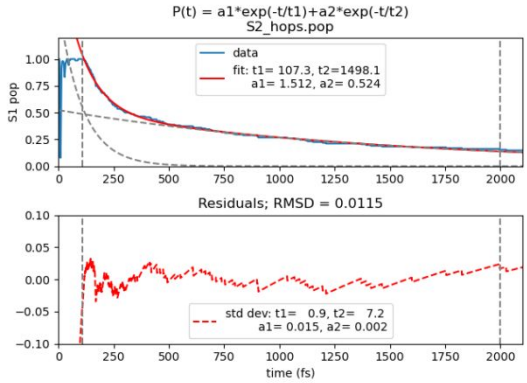
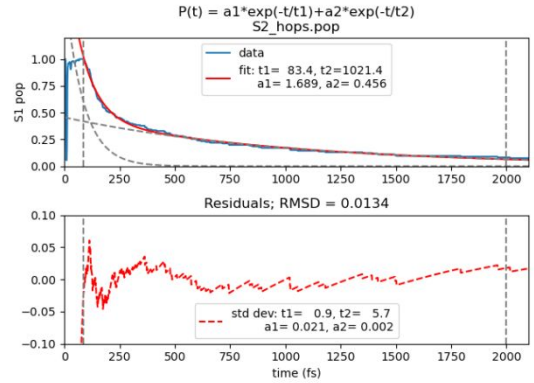
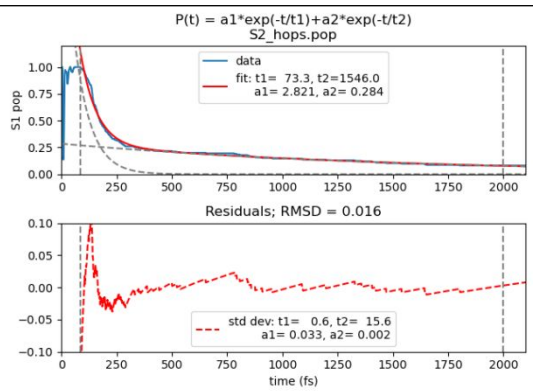
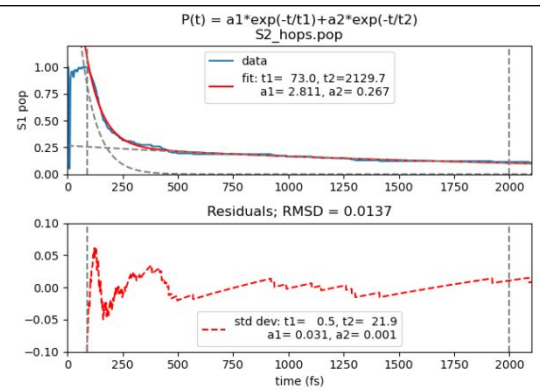
Since the importance of the second excited state S_2 cannot be understated, we also performed the same fitting-based analysis of the direct (all trajectories never hop to S_2) and indirect subsets (all trajectories hop to S_2). As evidenced in the main text, Table 2, the latter subset is 3.6 to 5.5 times smaller than the former one, hence the lower number of data points. Nevertheless, we achieved a model of similar quality as the ones for the full set or direct subset. Note that we assumed the same mechanistic scheme for both subsets than for the full ensemble of trajectories and used the corresponding optimized parameters as input guess.

Direct subset	AT retinal	13C retinal
pH=3	 t_{start} = 64 fs	 t_{start} = 60 fs

	$\tau_{fast} = 146 \pm 1$ fs $\tau_{slow} = 1461 \pm 14$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.80 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.20 \pm 0.00$ RMSD = 0.0129	$\tau_{fast} = 130 \pm 1$ fs $\tau_{slow} = 1201 \pm 12$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.83 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.17 \pm 0.00$ RMSD = 0.0123
pH=5	 t_{start} = 81 fs $\tau_{fast} = 148 \pm 0$ fs $\tau_{slow} = 1257 \pm 6$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.73 \pm 0.01$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.27 \pm 0.00$ RMSD = 0.0086	 t_{start} = 61 fs $\tau_{fast} = 130 \pm 0$ fs $\tau_{slow} = 1099 \pm 8$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.80 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.20 \pm 0.00$ RMSD = 0.0105
pH=7		

	t_{start} = 71 fs $\tau_{fast} = 93 \pm 0$ fs $\tau_{slow} = 1196 \pm 9$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.85 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.15 \pm 0.00$ RMSD = 0.0130	t_{start} = 49 fs $\tau_{fast} = 148 \pm 0$ fs $\tau_{slow} = 1560 \pm 19$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.82 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.18 \pm 0.00$ RMSD = 0.0126

Indirect subset	AT retinal	13C retinal
pH=3	 t_{start} = 108 fs $\tau_{fast} = 87 \pm 0$ fs $\tau_{slow} = 913 \pm 5$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.83 \pm 0.03$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.17 \pm 0.00$ RMSD = 0.0131	 t_{start} = 83 fs $\tau_{fast} = 100 \pm 1$ fs $\tau_{slow} = 2238 \pm 22$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.89 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.11 \pm 0.00$ RMSD = 0.0111

pH=5	 t_{start} = 107 fs $\tau_{fast} = 107 \pm 0$ fs $\tau_{slow} = 1498 \pm 7$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.74 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.26 \pm 0.00$ RMSD = 0.0115	 t_{start} = 86 fs $\tau_{fast} = 83 \pm 0$ fs $\tau_{slow} = 1021 \pm 5$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.79 \pm 0.02$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.21 \pm 0.00$ RMSD = 0.0134
pH=7	 t_{start} = 86 fs $\tau_{fast} = 73 \pm 0$ fs $\tau_{slow} = 1546 \pm 15$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.91 \pm 0.03$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.09 \pm 0.00$	 t_{start} = 90 fs $\tau_{fast} = 73 \pm 0$ fs $\tau_{slow} = 2130 \pm 21$ fs $P^{model}_{S1(fast)}(t = t_{start}) = 0.91 \pm 0.03$ $P^{model}_{S1(slow)}(t = t_{start}) = 0.09 \pm 0.00$

	RMSD = 0.0160	RMSD = 0.0137

7. Residue-based analysis

In this section, we report the average hop times, the isomerization quantum yields as well as the decay times when each ensemble of trajectories (pH value, retinal isomer) is split into 2 subsets, each of them corresponding to a given protonation state of a single amino acid. In the following tables, D means the residue is deprotonated while P means it is protonated. In the case of histidine, we don't distinguish the protonation sites on nitrogen δ and ε. Decay time constants are fitted using as guess the best parameters obtained for the full set, however with a different lower bound of the time window, always choosing the time at which the ground state population starts rising. Beware that some subsets can contain only a small number of trajectories (see the * in the second column below). In that case, the reported properties are not reliable. As expected, when the number of trajectories in each subset is much lower than the one in the other subset, the properties calculated for the latter subset are close to the ones calculated for the full ensemble. Results obtained for D57, D98, D120, D217 are discussed in the main text.

Table S10. Dataset analysis for the AT isomer at pH=3.

	Nr. traj	**Av. hop time (fs)**	**IQY**	**t_{start} (fs)**	**P_{fast}**	**P_{slow}**	**τ_{fast}(fs)**	**τ_{slow}(fs)**
Full	816	379 +/- 14	0.43 +/- 0.02	64	0.71	0.29	175	1474
D57 D	174	409 +/- 30	0.42 +/- 0.04	95	0.54	0.46	107	922
D57 P	642	371 +/- 16	0.44 +/- 0.02	64	0.74	0.26	174	1669

E62 D	19*	354 +/- 114	0.47 +/- 0.12	108	0.63	0.28	66	2558
E62 P	797	380 +/- 14	0.43 +/- 0.02	64	0.70	0.30	173	1397
D98 D	154	375 +/- 30	0.46 +/- 0.04	82	0.68	0.32	152	1173
D98 P	662	380 +/- 16	0.43 +/- 0.02	64	0.71	0.29	171	1515
D120 D	738	385 +/- 15	0.43 +/- 0.02	64	0.69	0.31	169	1420
D120 P	78	323 +/- 35	0.44 +/- 0.06	95	0.79	0.16	185	2517

Table S21. Dataset analysis for the AT isomer at pH=5.

	Nr. traj	**Av. hop time (fs)**	**IQY**	**t_{start} (fs)**	**P_{fast}**	**P_{slow}**	**τ_{fast}(fs)**	**τ_{slow}(fs)**
Full	832	480 +/- 18	0.39 +/- 0.02	81	0.60	0.40	172	1371
D57 D	601	441 +/- 20	0.42 +/- 0.02	81	0.63	0.37	154	1189
D57 P	231	588 +/- 39	0.31 +/- 0.03	96	0.51	0.49	223	1676
E62 D	432	444 +/- 23	0.41 +/- 0.02	81	0.63	0.37	165	1332
E62 P	400	518 +/- 28	0.38 +/- 0.02	81	0.56	0.44	178	1378
D98 D	740	499 +/- 20	0.38 +/- 0.02	81	0.58	0.42	176	1358
D98 P	92	327 +/- 35	0.51 +/- 0.05	81	0.76	0.24	139	1329
D217 D	99	450 +/- 55	0.42 +/- 0.05	96	0.64	0.36	118	1826
D217 P	733	484 +/- 19	0.39 +/- 0.02	81	0.59	0.41	176	1304
H219 D	134	483 +/- 45	0.43 +/- 0.04	96	0.54	0.46	129	1031
H219 P	698	479 +/- 20	0.38 +/- 0.02	81	0.60	0.40	173	1417

Table S32. Dataset analysis for the AT isomer at pH=7.

	Nr. traj	Av. hop time (fs)	IQY	t_{start} (fs)	P_{fast}	P_{slow}	τ_{fast}(fs)	τ_{slow}(fs)
Full	830	364 +/- 17	0.56 +/- 0.02	71	0.72	0.28	110	1341
H21 D	122	325 +/- 43	0.62 +/- 0.04	78	0.74	0.26	83	1523
H21 P	708	371 +/- 18	0.56 +/- 0.02	71	0.72	0.28	114	1321
E36 D	86	335 +/- 42	0.57 +/- 0.05	71	0.66	0.34	92	747
E36 P	744	368 +/- 18	0.56 +/- 0.02	78	0.72	0.28	101	1379
D217 D	773	337 +/- 18	0.56 +/- 0.02	71	0.71	0.29	107	1281
D217 P	57*	276 +/- 39	0.58 +/- 0.07	101	0.79	0.21	91	1877
H219 D	664	378 +/- 20	0.55 +/- 0.02	71	0.72	0.28	114	1432
H219 P	166	307 +/- 29	0.61 +/- 0.04	78	0.73	0.27	89	991

Table S43. Dataset analysis for the 13C isomer at pH=3.

	Nr. traj	Av. hop time (fs)	IQY	t_{start} (fs)	P_{fast}	P_{slow}	τ_{fast}(fs)	τ_{slow}(fs)
Full	861	340 +/- 14	0.33 +/- 0.02	60	0.76	0.24	153	1441
D57 D	99	371 +/- 42	0.27 +/- 0.05	84	0.69	0.31	120	1365
D57 P	762	336 +/- 14	0.33 +/- 0.02	60	0.77	0.23	154	1496
E62 D	8*	244 +/- 83	0.75 +/- 0.15	87	0.61	0.29	65	423

E62 P	853	341 +/- 14	0.32 +/- 0.02	60	0.77	0.23	156	1526
D98 D	2*	-	-	-	-	-	-	-
D98 P	859	338 +/- 13	0.33 +/- 0.02	60	0.76	0.24	153	1420
D120 D	149	398 +/- 34	0.23 +/- 0.04	89	0.74	0.26	174	1752
D120 P	712	328 +/- 15	0.35 +/- 0.02	60	0.76	0.24	139	1342

Table S54. Dataset analysis for the 13C isomer at pH=5.

	Nr. traj	**Av. hop time (fs)**	**IQY**	**t_{start} (fs)**	**P_{fast}**	**P_{slow}**	**τ_{fast}(fs)**	**τ_{slow}(fs)**
Full	818	345 +/- 14	0.35 +/- 0.02	61	0.70	0.30	142	1122
D57 D	551	361 +/- 17	0.33 +/- 0.02	61	0.68	0.32	151	1071
D57 P	267	310 +/- 22	0.41 +/- 0.03	64	0.72	0.28	119	1231
E62 D	553	329 +/- 16	0.35 +/- 0.02	61	0.74	0.26	148	1209
E62 P	265	377 +/- 27	0.35 +/- 0.03	61	0.62	0.38	130	1026
D98 D	108	432 +/- 46	0.18 +/- 0.04	98	0.68	0.31	191	2331
D98 P	710	332 +/- 14	0.37 +/- 0.02	61	0.69	0.31	130	981
D217 D	188	384 +/- 32	0.30 +/- 0.03	61	0.66	0.34	151	1430
D217 P	630	333 +/- 15	0.37 +/- 0.02	61	0.70	0.30	138	1019
H219 D	107	365 +/- 38	0.25 +/- 0.04	61	0.65	0.35	142	844
H219 P	711	341 +/- 15	0.37 +/- 0.02	61	0.70	0.30	142	1186

Table S65. Dataset analysis for the 13C isomer at pH=7.

	Nr. traj	Av. hop time (fs)	IQY	t_{start} (fs)	P_{fast}	P_{slow}	τ_{fast}(fs)	τ_{slow}(fs)
Full	795	348 +/- 16	0.36 +/- 0.02	49	0.77	0.23	167	1817
H21 D	122	401 +/- 51	0.47 +/- 0.05	49	0.68	0.32	125	1977
H21 P	673	339 +/- 16	0.35 +/- 0.02	59	0.77	0.23	156	1644
E36 D	104	328 +/- 36	0.47 +/- 0.05	68	0.68	0.32	108	1027
E36 P	691	351 +/- 17	0.35 +/- 0.02	49	0.78	0.22	170	2038
D217 D	676	345 +/- 17	0.36 +/- 0.02	49	0.76	0.24	160	1804
D217 P	119	366 +/- 41	0.40 +/- 0.05	59	0.80	0.20	185	1942
H219 D	649	356 +/- 18	0.37 +/- 0.02	49	0.75	0.25	154	1920
H219 P	146	314 +/- 24	0.36 +/- 0.02	59	0.84	0.16	200	1204

Bibliography

1 Pieri, E. *et al.* CpHMD-Then-QM/MM Identification of the Amino Acids Responsible for the Anabaena Sensory Rhodopsin pH-Dependent Electronic Absorption Spectrum. *Journal of Chemical Theory and Computation* **15**, 4535-4546, doi:10.1021/acs.jctc.9b00221 (2019).

2 Hayashi, S., Tajkhorshid, E. & Schulten, K. Structural Changes during the Formation of Early Intermediates in the Bacteriorhodopsin Photocycle. *Biophysical Journal* **83**, 1281-1297, doi:https://doi.org/10.1016/S0006-3495(02)73900-3 (2002).

3 Weingart, O. *et al.* COBRAMM 2.0 — A software interface for tailoring molecular electronic structure calculations and running nanoscale (QM/MM) simulations. *Journal of Molecular Modeling* **24**, 271, doi:10.1007/s00894-018-3769-6 (2018).

4 Weber, W. & Thiel, W. Orthogonalization corrections for semiempirical methods. *Theoretical Chemistry Accounts* **103**, 495-506, doi:10.1007/s002149900083 (2000).
5 Fabiano, E., Keal, T. W. & Thiel, W. Implementation of surface hopping molecular dynamics using semiempirical methods. *Chemical Physics* **349**, 334-347, doi:https://doi.org/10.1016/j.chemphys.2008.01.044 (2008).
6 Dral, P. O. *et al.* Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Theory, Implementation, and Parameters. *Journal of Chemical Theory and Computation* **12**, 1082-1096, doi:10.1021/acs.jctc.5b01046 (2016).
7 Sen, S., Kar, R. K., Borin, V. A. & Schapiro, I. Insight into the isomerization mechanism of retinal proteins from hybrid quantum mechanics/molecular mechanics simulations. *WIREs Computational Molecular Science* **12**, e1562 (2022).
8 Tully, J. C. Molecular dynamics with electronic transitions. *The Journal of Chemical Physics* **93**, 1061-1071, doi:10.1063/1.459170 (1990).
9 Tully, J. C. & Preston, R. K. Trajectory Surface Hopping Approach to Nonadiabatic Molecular Collisions: The Reaction of H+ with D2. *The Journal of Chemical Physics* **55**, 562-572, doi:10.1063/1.1675788 (1971).
10 Zhu, C., Jasper, A. W. & Truhlar, D. G. Non-Born-Oppenheimer Liouville-von Neumann Dynamics. Evolution of a Subsystem Controlled by Linear and Population-Driven Decay of Mixing with Decoherent and Coherent Switching. *Journal of Chemical Theory and Computation* **1**, 527-540, doi:10.1021/ct050021p (2005).
11 Granucci, G. & Persico, M. Critical appraisal of the fewest switches algorithm for surface hopping. *The Journal of Chemical Physics* **126**, 134114, doi:10.1063/1.2715585 (2007).