QM/MM Investigation of the Spectroscopic Properties of the Fluorophore of Bacterial Luciferase

Germano Giuliani, [§] Federico Melaccio,[§] Samer Gozem,*[,#] Andrea Cappelli,[§] Massimo Olivucci.*[,§,‡]

§ Department of Biotechnology, Chemistry and Pharmacy, University of Siena, Via A. Moro 2, 53100 Siena (Italy).

Department of Chemistry, Georgia State University, Atlanta, GA 30302 (USA).

‡ Department of Chemistry, Bowling Green State University, Bowing Green, OH 43403 (USA).

Supporting Information

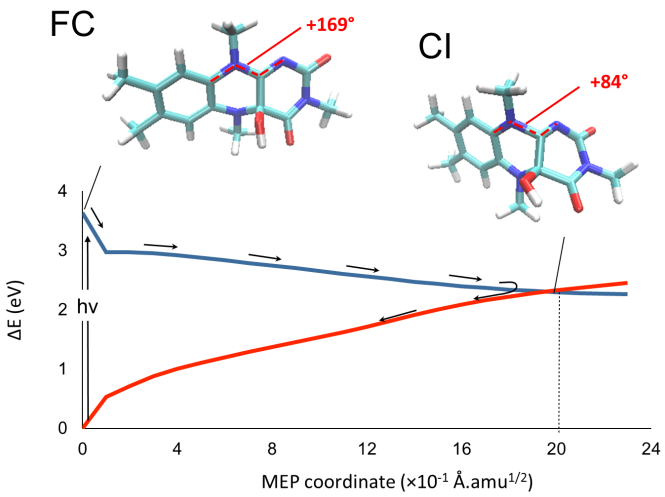

Figure S1. Computed S_0 and S_1 energy profiles of the N(3)- and N(5)-alkylated derivative of FMNHOH (**2**) in solution, connecting the Frank-Condon (FC) geometry to a Conical Intersection (CI). The arrows represent the internal conversion along the S_1 pathway where the energy decrease is associated with an out-of-plane deformation of the pyrimidine ring with respect to the original flavin plane. The values of the geometric distortion which involves the C_{9a}-N_{10}-C_{10a}-N_1 dihedral angle of the molecule, are highlighted in red. Adapted from Gozem et al., *Angew. Chem. Int. Ed.* **2014**, *53*, 9870-9875.

Preparing the protein model

The crystallographic structure of Bacterial Luciferase (PDB code: 3FGC)[1] was downloaded from the Protein Data Bank.[2] Charge parameters for the 4a-hydroxy flavin (**2** in the main manuscript) were obtained using the restrained electrostatic potential (RESP) protocol derived from HF/6-31G* electron densities computed in Gaussian 03.[3] Force constants for bonds, angles and torsions of **2** were taken from the GAFF parameter set.[4] Preliminary energy minimization of the ligand in the gas phase was performed using the Molcas 7.8 program.[5]

The hydrogen atoms, counter ions and water solvent were added using GROMACS 4.6.3[6] employing the AMBER99SB force field for the protein[7] and TIP3P force field for the crystallographic water molecules. The protonation states of ionized residue were assigned by estimating their pKa using PROPKA version 3.0.[8] and assuming a neutral reference pH. Estimated pKa's are reported in Table S1, along with the reference pKa for each side chain in its standard protonation state (pKmodel). Residues with chosen non-standard protonation states are highlighted in bold. Table S2 reports simulation parameters and the final

choice of protonation states for specific ionizable residues which required a more detailed analysis. Following the PROPKA simulation, we used the following protocol:

- All Asp residues have been kept charged, since none had a pKa higher than 5.7;

- Glu43 and Glu328 were assigned a neutral protonation state because their pKa is close to 7 due to being buried inside the protein. Both had the extra hydrogen atom added to the OE2 carboxylate atom;

- Histidines were chosen to be neutral whenever their pKa was below 6. HID/HIE tautomer selection was based on visual inspection to create a reasonable local hydrogen bond network;

- Histidines with a pKa around 6 or higher were assigned the doubly protonated HIP states. His199 is an exception and was chosen to be neutral because of its short distance with positively charged Lys202.

Fourteen Na^+ anions were placed using the *genion* module of Gromacs to obtain a total zero charge for the model. The entire system (luciferase/FMNHOH complex) was solvated with a cubic TIP3P water box such that the edge of the water box is at least 6 Å from the protein. In total, the system consists of 38985 atoms. Preliminary energy minimization was achieved with the protein (including the modelled mobile loop) and solvent with the ligand frozen during 1000-step minimization. Subsequently, a 5000-step energy minimization of part of the protein environment was performed, which includes all residues and water molecules within 8 Å of the fluorophore.

Table S1

Residue	pKa	pKmodel
ASP 37	4.00	3.80
ASP 89	5.39	3.80
ASP 94	4.20	3.80
ASP 111	4.83	3.80
ASP 113	5.66	3.80
ASP 120	3.59	3.80
ASP 122	4.58	3.80
ASP 129	3.57	3.80
ASP 133	5.35	3.80
ASP 147	3.71	3.80
ASP 206	3.24	3.80
ASP 218	2.68	3.80
ASP 223	3.98	3.80
ASP 233	2.35	3.80
ASP 235	3.75	3.80
ASP 241	3.02	3.80
ASP 252	3.98	3.80
ASP 262	4.19	3.80
ASP 263	3.94	3.80
ASP 265	2.58	3.80
ASP 271	3.01	3.80
ASP 279	3.96	3.80
ASP 293	3.40	3.80
ASP 314	3.44	3.80
ASP 316	3.45	3.80
ASP 321	4.62	3.80
ASP 346	3.16	3.80
GLU 14	4.11	4.50
GLU 19	3.68	4.50
GLU 32	4.16	4.50
GLU 43	**8.75**	**4.50**
GLU 48	4.02	4.50
GLU 67	4.58	4.50
GLU 88	4.92	4.50
GLU 137	4.92	4.50
GLU 141	4.69	4.50
GLU 149	4.32	4.50
GLU 175	5.68	4.50
GLU 181	4.91	4.50
GLU 185	4.36	4.50
GLU 200	4.11	4.50
GLU 210	4.77	4.50
GLU 214	4.53	4.50
GLU 297	4.77	4.50
GLU 305	4.74	4.50
GLU 306	3.33	4.50
GLU 328	**6.67**	**4.50**
GLU 333	4.82	4.50
GLU 334	4.66	4.50
GLU 335	3.97	4.50
GLU 353	3.67	4.50
C- 355	3.28	3.20
HIS 44	2.23	6.50
HIS 45	3.96	6.50
HIS 61	6.17	6.50
HIS 82	5.07	6.50
HIS 150	6.00	6.50
HIS 199	6.37	6.50
HIS 215	7.01	6.50
HIS 224	3.36	6.50
HIS 234	6.48	6.50
HIS 249	6.32	6.50
CYS 34	10.17	9.00
CYS 106	13.73	9.00
CYS 130	11.27	9.00
CYS 225	19.25	9.00
CYS 243	12.08	9.00
CYS 307	12.62	9.00
CYS 324	9.29	9.00
CYS 325	12.93	9.00
TYR 10	11.77	10.00
TYR 56	15.37	10.00
TYR 110	14.16	10.00
TYR 132	12.35	10.00
TYR 143	9.92	10.00
TYR 163	10.16	10.00
TYR 171	15.03	10.00
TYR 208	13.06	10.00
TYR 217	10.52	10.00
TYR 228	16.66	10.00
TYR 251	10.22	10.00
TYR 254	10.75	10.00
TYR 270	10.14	10.00
TYR 294	10.30	10.00
TYR 296	10.82	10.00
TYR 350	10.31	10.00
LYS 2	11.05	10.50
LYS 22	10.60	10.50
LYS 29	10.31	10.50
LYS 98	10.52	10.50
LYS 112	9.25	10.50
LYS 136	9.42	10.50
LYS 152	10.59	10.50
LYS 155	10.40	10.50
LYS 201	9.21	10.50
LYS 202	11.25	10.50
LYS 221	10.65	10.50
LYS 240	10.59	10.50
LYS 259	10.56	10.50
LYS 268	9.77	10.50
LYS 274	10.59	10.50
LYS 283	10.48	10.50
LYS 341	10.56	10.50
LYS 352	10.45	10.50
LYS 354	10.33	10.50
ARG 23	11.80	12.50
ARG 85	12.53	12.50
ARG 100	11.51	12.50
ARG 102	11.55	12.50
ARG 107	10.56	12.50
ARG 115	12.26	12.50
ARG 125	11.23	12.50
ARG 186	12.53	12.50
ARG 238	14.28	12.50
ARG 244	13.49	12.50
ARG 278	12.47	12.50
ARG 290	12.44	12.50
ARG 291	12.74	12.50
N+ 1	8.85	8.00

Table S2

Residue	*pKa*	Buried (%)	Hydrogen bond	Chosen Protonation
HIS44	2.2	81%	-	HID
HIS45	4.0	70%	GLH43	HID
HIS61	6.2	2%	-	HIE
HIS82	5.1	53%	-	HIE
HIS150	6.0	29%	-	HID
HIS199	6.4	0%	-	HID
HIS215	7.0	0%	-	HIP
HIS224	3.4	99%	-	HID
HIS234	6.5	0%	-	HIP
HIS249	6.3	8%	-	HIP
HIS285	6.1	0%	-	HIP
GLU43	8.8	100%	HID45	GLH:OE2
GLU328	6.7	100%	-	GLH:OE2
ARG290	12.4	0%	GLU200	ARG:NH1

Replica-exchange molecular dynamics (REMD) simulations

REMD simulations were performed using the GROMACS 4.6.3 program.[6] The loop and all residues and water molecules having at least an atom within 6 Å of the loop residues were allowed to move, while the rest of the system was kept fixed. This is because the missing loop is a highly mobile domain and had to be better sampled to ensure a comparable accuracy with the rest of the MM model derived from crystallography.[9] All classical and REMD trajectories were carried out in the NVT ensemble with a Berendsen thermostat.[10] Preliminary standard MD simulations (Figure S2 and S3, split into two figures for clarity) were performed separately with temperatures distributed between 280 and 347 K. The simulations were conducted for 1.5 ns with the annealing-temp option for each temperature. The REMD simulation (Figure S4) was set up with replica exchange between the different temperatures every 2 picosecond, during 10 ns of dynamics at reference temperature of 298 K. The temperature of the system was controlled using Langevin dynamics with collision frequency (γ) of 1 ps.

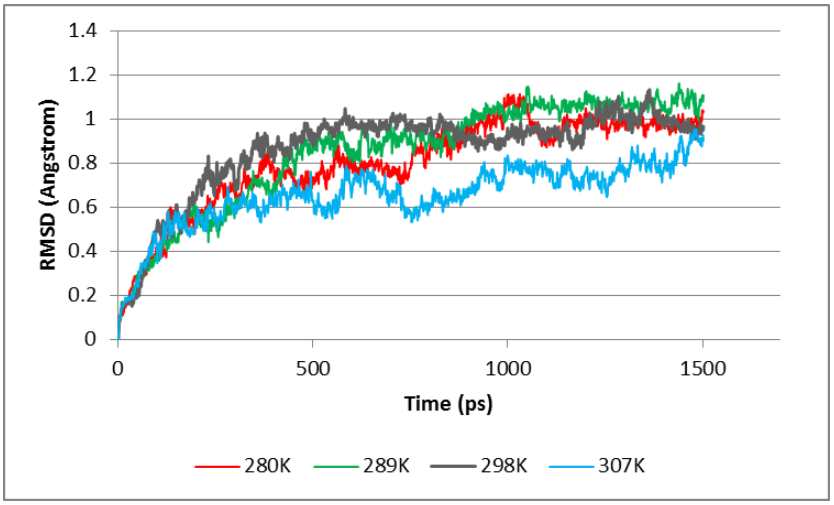

Figure S2. Molecular Dynamic trajectories obtained at different equilibration temperatures.

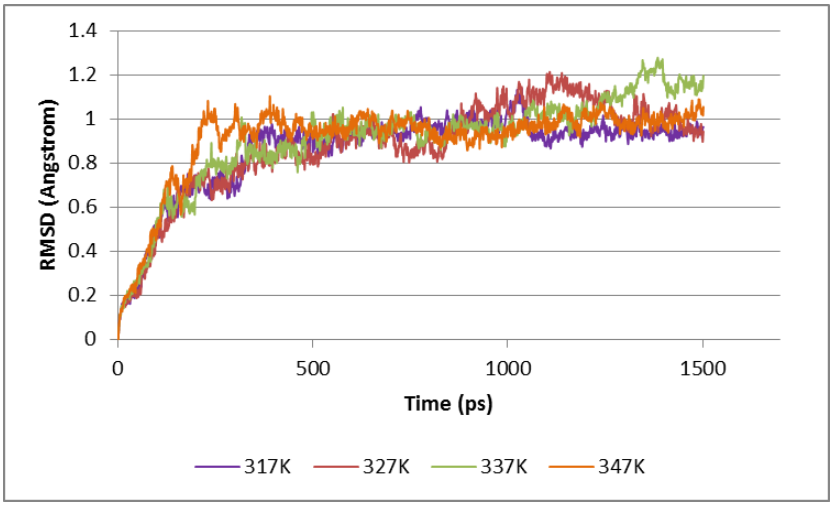

Figure S3. Molecular Dynamic trajectories obtained at different equilibration temperatures.

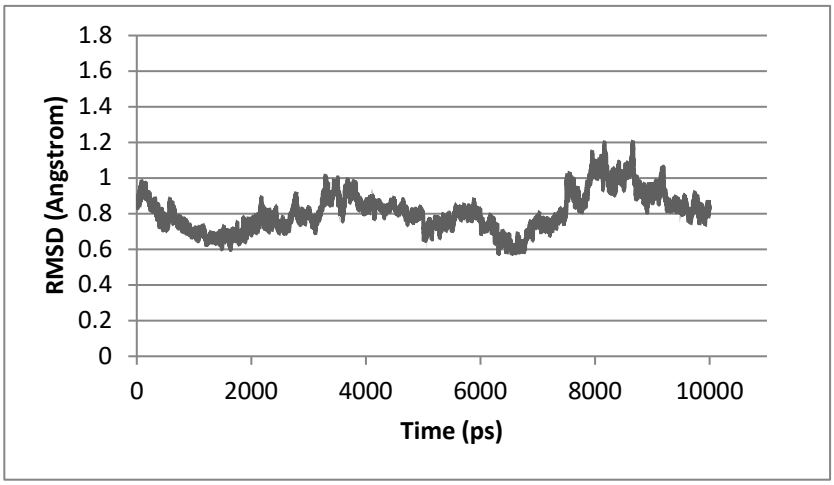

Figure S4. REMD trajectory obtained at reference temperature of 298 K. RMSD ≤ 0.6 Å.

Cluster analysis of REMD simulations

REMD simulations make it possible to substantially increase the conformational space sampled by molecular dynamics. However, the interpretation of results requires statistical analysis. Our aim was to find the most populated loop conformation under the assumption that it is the preferred one in physiological conditions. A well-established tool is cluster analysis, which has been routinely used for this purpose. We used the single-linkage algorithm [11] as implemented in the GROMACS 4.6.3 *g_cluster* routine, with a 0.1 nm RMSD cutoff. The analysis was performed on the REMD de-multiplexed trajectory at 298K constant-temperature (i.e. our

reference trajectory, Figure S4). 96 clusters were obtained, with only 5 clusters including more than 500 MD trajectory snapshots. Figure S5 show the number of structures for each obtained cluster.

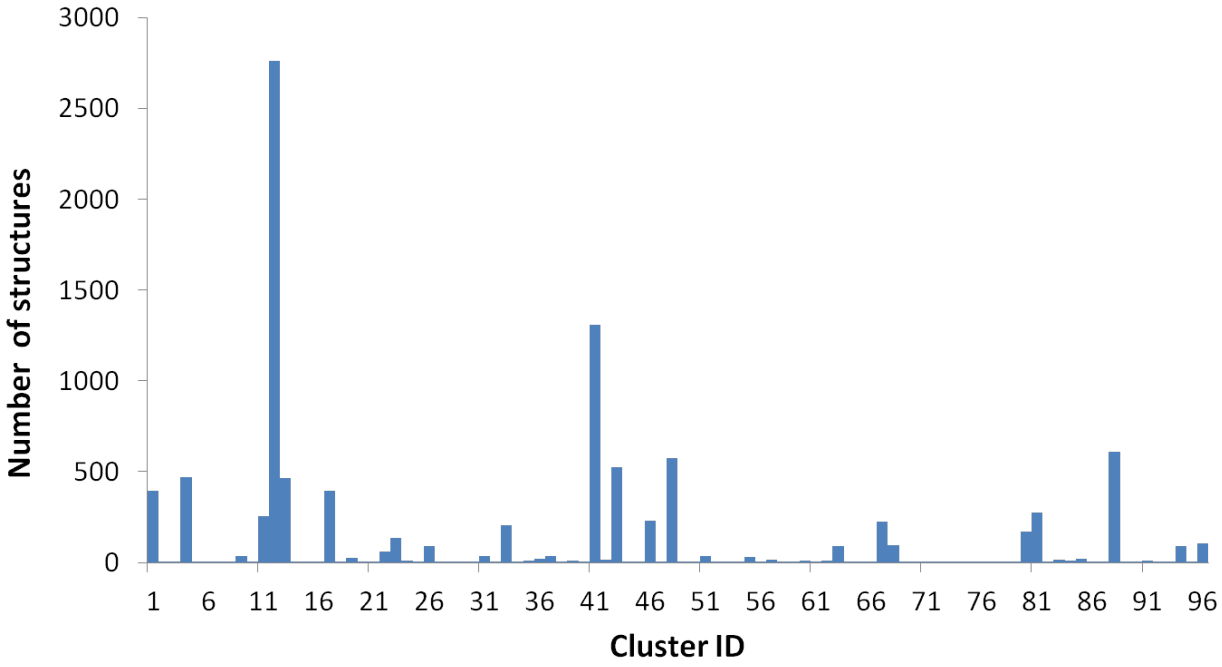

Figure S5. Number of structures for each cluster resulting from our analysis.

Cluster #12 was the most populated one with 2765 snapshots, while the second most populated is cluster #41 with less than half of cluster #12 frames. The reference structure used for subsequent QM/MM calculations has been chosen as the snapshot in cluster #12 with the lowest RMSD from the average structure produced by the *g_cluster* routine. This is the MD snapshot that would resemble the cluster "average structure" more closely, while being a realistic structure, instead of the actual average which might include chemical artifacts.

To assess the effect of the chosen RMSD cutoff on the resulting model, we have performed other cluster analyses to plot the number of obtained clusters as a function of the RMSD cutoff. The optimal value would be a compromise between a too low one which produces many clusters with a population of 1 or few-members clusters or a too large value which groups most structures in the same cluster, both of which are meaningless. As seen in Figure S6, the 0.1 nm cutoff marks the beginning of the flat section of the curve, which is the region where cluster analysis would generate non-trivial clusters. The most populated clusters

should remain the same throughout that region, regardless of the chosen cutoff. While cutoff values of 0.115 and 0.120 nm lead to a very large "best" clusters comprising 77% and 94% of the trajectory respectively, 0.105 and 0.11 nm would be suitable choices too, given the even lower number of obtained clusters. Therefore, we performed the extraction of the snapshot closest to the cluster average structure for the most populated clusters obtained with 0.105 and 0.11 nm cutoff values. Both these structures belong to the most populated cluster from our original cluster analysis (i.e. with 0.1 nm cutoff). Furthermore, our reference structure is present in the most populated clusters in both cases, which confirms that the favored loop conformation is independent of the RMSD cutoff.

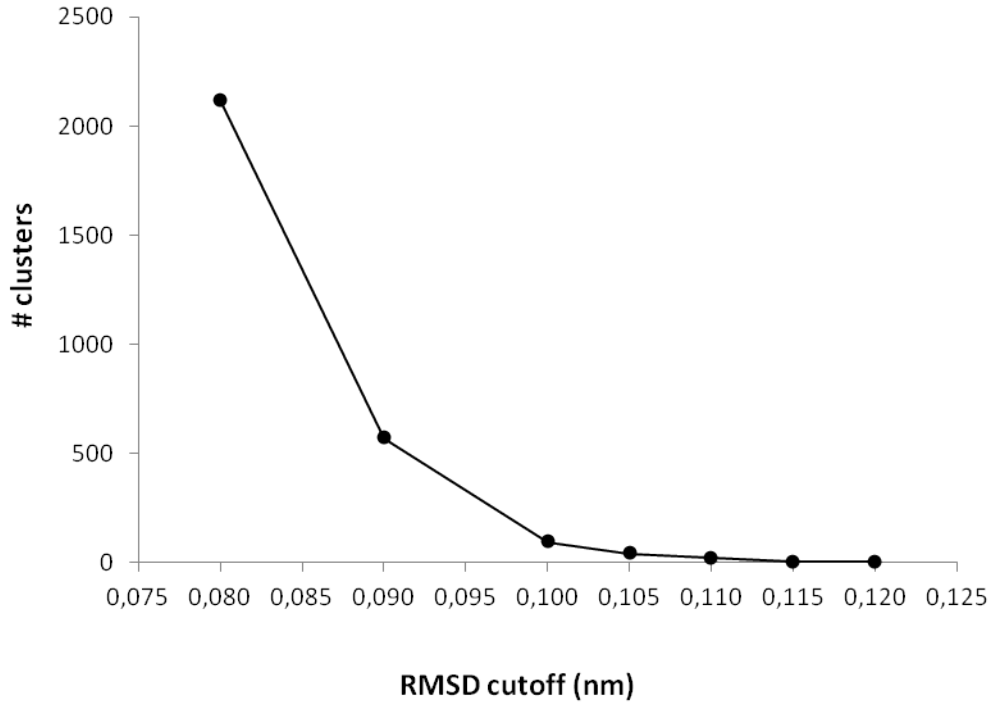

Figure S6. Number of obtained clusters as a function of RMSD cutoff values.

QM/MM Calculations

Optimizations and MEP calculations for FMNHOH (**2**), embedded in the protein environment, were performed at the 3-root state-averaged CASSCF level of theory[12] with a 6-31G* basis set. The CASSCF wave function comprised an active space of 14 electrons in 11 orbitals (see Figure S7 for a full description of the active space). These orbitals include π and π* orbitals involved in the π-conjugated system of FMNHOH. The full active space (i.e. including all π and π * orbitals) would comprise 18 electrons and 15 orbitals, but here we use a slightly reduced active space of 14 electrons and 11 orbitals such that 4 orbitals and electrons have been excluded from the active space due to the high computational cost of using the full active space. We do, however, check the effect of using the full 18,15 active space on the spectroscopic properties of FMNHOH (see manuscript). The excluded electrons and orbitals belong to lower energy π and higher energy π* orbitals of the benzene ring. Non-bonding orbitals are excluded from the active space since n,π* states are assumed to be high in energy, based on earlier TD-DFT studies,[13] and are also not relevant to the spectroscopic properties of the system (they are dark states).

Spectroscopic properties were computed using several different flavors of CASPT2 and basis sets. We test the effect of the IPEA shift (0 vs. 0.25), the basis set (6-31G*, ANO-L-VDZP, and ANO-L-VTZP), active space (18,15 vs. 14,11) and state-averaging (SS vs MS CASPT2).

Two models were generated; one with a cysteine (Cys106) included in the QM calculation, and one where it is treated at the MM level. Even with Cys106 in the QM region, however, no cysteine orbitals were included in the active space (Fig. S7).

The MEPs were computed in the form of an intrinsic reaction coordinate as implemented in Molcas 7.8, with a step constraint of 0.03 Bohr.amu$^{1/2}$ for the first eight steps, followed by 0.05 Bohr.amu$^{1/2}$ for the remainder of the MEP. Single point CASPT2 calculations[14] were then performed for stationary points and for each point along the MEP path. Our CASPT2 calculations do not include an IPEA shift[15] and employ the same basis set and active space as in the CASSCF calculations (14,11 and 6-31G*). All calculations were performed with Molcas v. 7.8.[5]

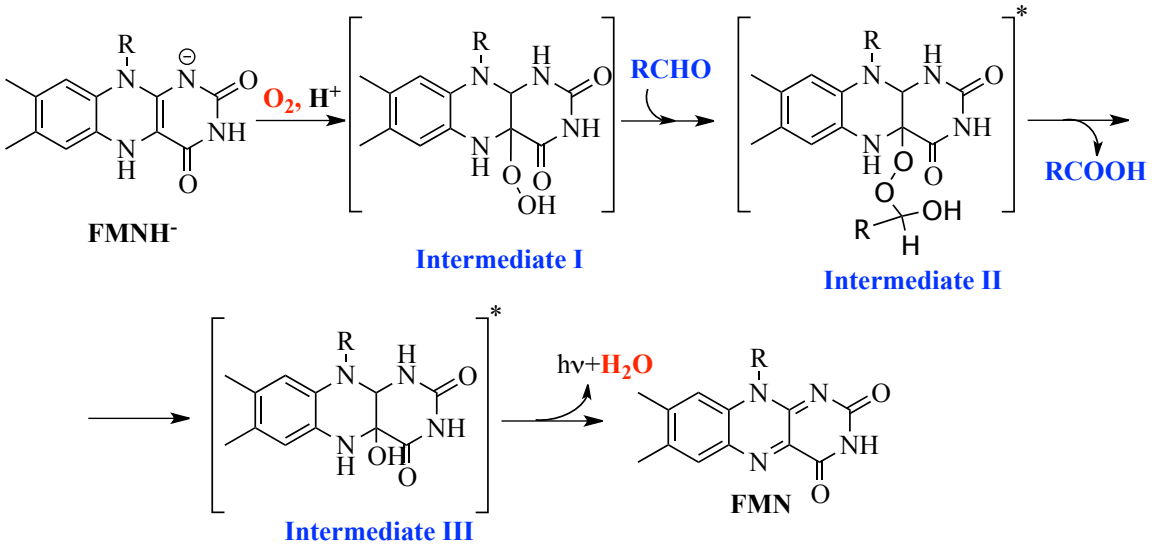

Figure S7: Proposed mechanism for bioluminescence in bacterial luciferase. Adapted from ref. [16] The lowest-energy absorption band of 4a-hydroxy flavins are taken as the signature for the participation of a 4a-hydroxy flavin intermediate (e.g. Intermediate III) to the catalytic cycle of various flavoprotein monooxygenases (such as styrene monooxygenase[17] and p-hydroxybenzoate hydroxylase[18])

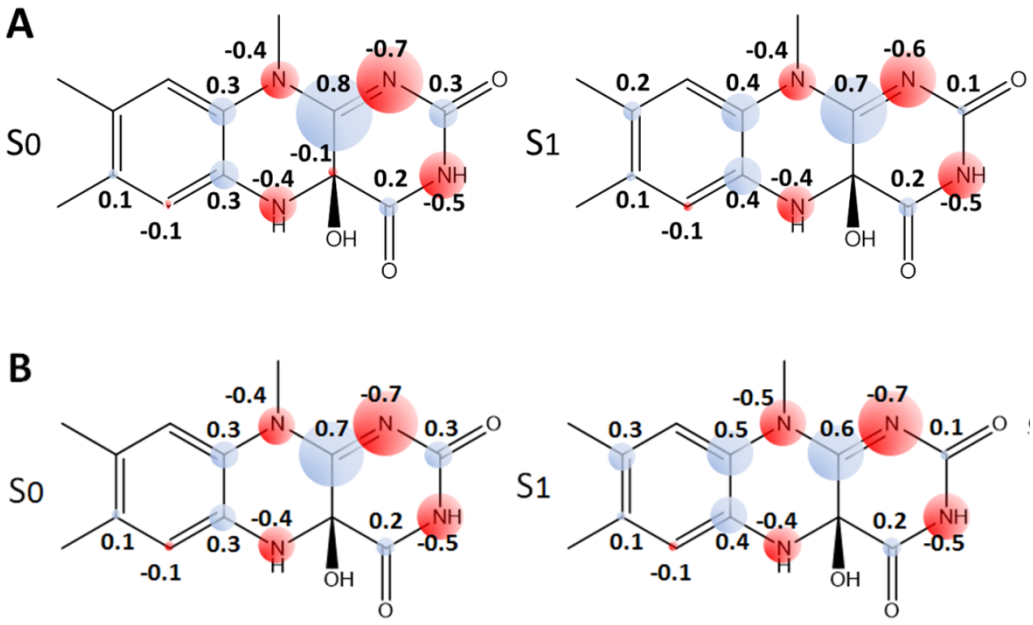

Figure S8: CASPT2 Mulliken charge distributions for FMNHOH at select geometries. Charges are shown for the ground state (S_0, left), and the first singlet excited state (S_1, right) for: A. The Frank-Condon geometry. B. The energy minimum (EM) geometry. The charges of all hydrogen atoms and other substituents were summed onto the atoms involved in the backbone ring structure of FMNHOH.

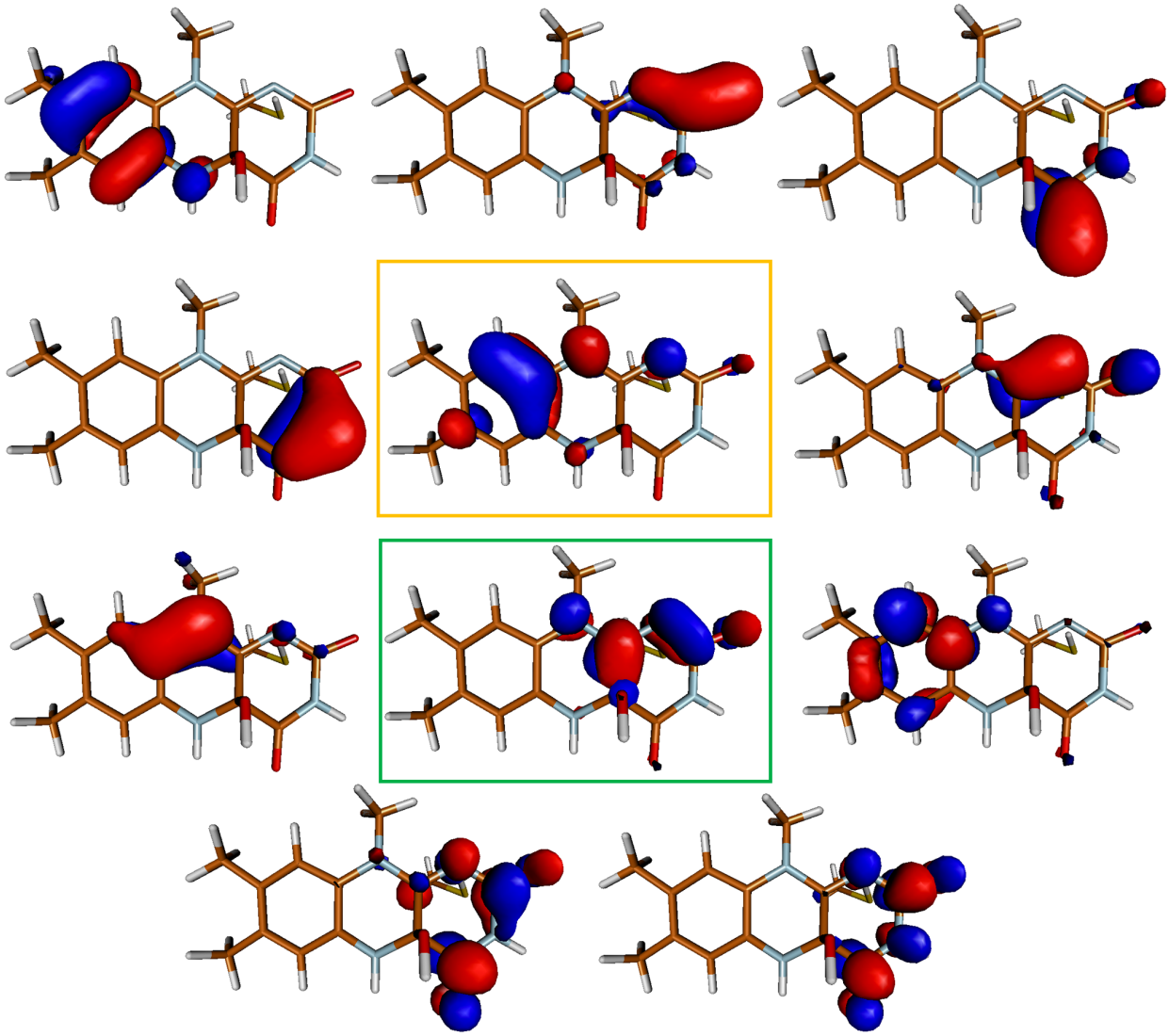

Figure S9: The active space orbitals used in CASSCF and CASPT2 calculations with the 6-31G* basis set. Shown here are the active space orbitals for the calculation with Cys106 in the QM region. The same orbitals were used in the calculation with Cys106 treated with MM. The 18,15 calculations include the full set of valence π and π*, i.e., including two π and two π* orbitals on benzene missing above. The HOMO (orange frame) and LUMO (green frame) orbitals are highlighted.

Coordinates and absolute CASSCF, 6-31G* QM/MM energies of stationary points:

Ground state optimized geometry (GS)
35

N 35.04146674 40.02835991 34.23404388
C 34.60039296 40.01657206 32.95813985
O 33.46131978 40.28182990 32.65701521
N 35.46414498 39.62424754 31.92657785
C 36.79878856 39.45014245 32.06361589
O 37.53081687 39.07902326 31.17914975
C 37.37328036 39.86720531 33.40507003
N 38.36821507 38.93918561 33.77664625
C 38.95660609 39.16781343 35.03067914
C 40.24517582 38.79519406 35.31448025
C 40.79174061 38.95151612 36.59100268
C 42.18674002 38.43931080 36.87381508
C 40.01834845 39.56679910 37.57449053
C 40.49512627 39.72392312 39.00080337
C 38.73023639 39.99438971 37.26685293
C 38.16516108 39.76684306 36.02038938
N 36.79494521 40.00287066 35.73061068
C 36.32290985 39.94093894 34.49712057
C 35.88138266 40.31454383 36.83917113
H 35.07878832 39.53939197 31.01048163
H 40.82541961 38.33445373 34.53144316
H 42.18998986 37.71473951 37.68144759
H 42.60649950 37.96152858 35.99649902
H 42.86588611 39.23690951 37.15810035
H 40.00852040 40.56538588 39.47937486
H 40.24395185 38.83563534 39.57717561
H 41.56860293 39.86733216 39.05761895
H 38.16637650 40.46878198 38.04582401
H 36.21224039 41.23552467 37.29580134
H 34.88946233 40.44334532 36.45107810
H 39.02761807 38.70910701 33.05738640
O 37.76237759 41.20455852 33.26477724
H 38.46944017 41.27580060 32.63396544
C 35.87895549 39.22097402 37.87485282
H 35.87960111 39.51186359 37.59936149

Excited state minimum geometry (EM)
35
-946.210866308881
N 35.00102716 39.99676099 34.22691256
C 34.56344548 39.98129980 32.98723757
O 33.42217184 40.26805441 32.61313145
N 35.46432180 39.56799584 31.95006172
C 36.78404601 39.40150064 32.08849009
O 37.54980130 39.00498970 31.22895376
C 37.31566218 39.87859761 33.42552272
N 38.51692484 39.16273430 33.74946906
C 38.96417769 39.18980758 35.00495771
C 40.27198827 38.73970825 35.29989879
C 40.80826809 38.81500547 36.55768862
C 42.19144972 38.27475503 36.83353825
C 40.03310930 39.41768386 37.57247500
C 40.49148237 39.53778469 38.99234834
C 38.70907497 39.88284166 37.28006506
C 38.10968435 39.69085080 36.05076865
N 36.77660528 39.95514275 35.79227507
C 36.27869762 39.68497968 34.49831954
C 35.87720934 40.28092212 36.88264175
H 35.07056964 39.41057944 31.04793483
H 40.84126057 38.33431797 34.48086214
H 42.16622372 37.50280656 37.59512887
H 42.62598707 37.84995635 35.93751214
H 42.86288274 39.05190177 37.18071558
H 40.11605087 40.45944680 39.42599874
H 40.07372095 38.70942683 39.57268559
H 41.57112786 39.50209358 39.07276042
H 38.17906489 40.35570800 38.08352114
H 36.18641881 41.20509564 37.35835526
H 34.89099638 40.41896449 36.47727757
H 39.13994555 38.87465019 33.01975805
O 37.54368744 41.25451394 33.33926660
H 38.12808394 41.44712716 32.61548464
C 35.83510313 39.19260189 37.92980650
H 35.84630338 39.48209507 37.65126068

References

1. Z.T. Campbell, A. Weichsel, W.R. Montfort, T.O. Baldwin, *Biochemistry* **2009**, *48*, 6085-6094.
2. http://www.rcsb.org/.
3. P. Cieplak, W.D. Cornell, C. Bayly, P.A. Kollman, *J. Comput. Chem.* **1995**, *16*, 1357-1377.
4. J.M. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case *J. Comput. Chem.* **2004**, *25*, 1157-1174.
5. F. Aquilante, L. De Vico, N. Ferré, G. Ghigo, P. A. Malmqvist, P. Neogrády, T. B. Pedersen, M. Pitonák, M. Reiher, B. O. Roos, L. Serrano-Andrés, M. Urban, V. Veryazov, and R. Lindh, *J. Comput. Chem.* **2010**, *31*, 224-47.
6. a) H.J.C. Berendsen, D. Van der Spoel, R. Van Drunen, *Comp. Phys. Comm.* **1995**, *91*, 43-56. b) E. Lindhal, B. Hess, D. Van der Spoel, *J. Mol. Model.* **2001**, *7*, 306-317.
7. V. Hornak, R. Abel, A. Okur, B. Strockbine, C. Simmerling, *Proteins*, **2006**, 65, 712-725.
8. PROPKA 3 Mats H.M. Olsson, Chresten R. Sondergard, Michal Rostkowski, and Jan H. Jensen, *J. Chem. Theory Comput.* **2011**, 7, 525-537.
9. A. Onufriev, D. Bashford, D.A. Case, *Proteins* **2004**, *55*, 383-394. J. Chen, C.L. Brooks, J. Khandogin, *Curr. Opin. Struct. Biol.* **2008**, *18*, 140-148.
10. G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.* **2007**, *126*, 014101.
11. A. E. Torda, W. F. van Gusteren, *J. Comput. Chem.***1994**, *15*, 1331-1340.
12. B. O. Roos, in *Advances in Chemical Physics: Ab Initio Methods in Quantum Chemistry Part 2, Volume 69* (Ed.: K.P. Lawley), Wiley Online Library, **1987**, p 399-445.
13. D. Zhou, E. Mirzakulova, R. Khatmullin, I. Schapiro, M. Olivucci, K. D. Glusac, *J. Phys. Chem. B* **2011**, *115*, 7136-43.
14. K. Andersson, P. A. Malmqvist, B. O. Roos, A. J. Sadlej, K. Wolinski, *J. Phys. Chem.* **1990**, *94*, 5483-5488.
15. G. Ghigo, B. O. Roos, P. K. Malmqvist, *Chem. Phys. Lett.* **2004**, *396*, 142-149.
16. B. Lei, Q. Ding, S. C. Tu, *Biochemistry* **2004**, *43*, 15975-82.
17. a) A. Kantz, G. T. Gassner, *Biochemistry* **2011**, *50*, 523-32. b) D. Tischler, M. Schlömann, W. J. van Berkel, G. T. Gassner, *FEBS letters* **2013**, *587*, 3848-3852.
18. B. Entsch, D. P. Ballou, V. Massey, *J. Biol. Chem.* **1976**, *251*, 2550-63.