Improved description of environment and vibronic effects with electrostatically embedded ML potentials

Kirill Zinovjev[1*], *Carles Curutchet*[2,3**]

[1]Departamento de Química Física, Universidad de Valencia, 46100, Burjassot (Spain)

[2]Departament de Farmàcia i Tecnologia Farmacèutica, i Fisicoquímica, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona (UB), Barcelona, Spain

[3]Institut de Química Teòrica i Computacional (IQTCUB), Universitat de Barcelona (UB), Barcelona, Spain

Corresponding Authors

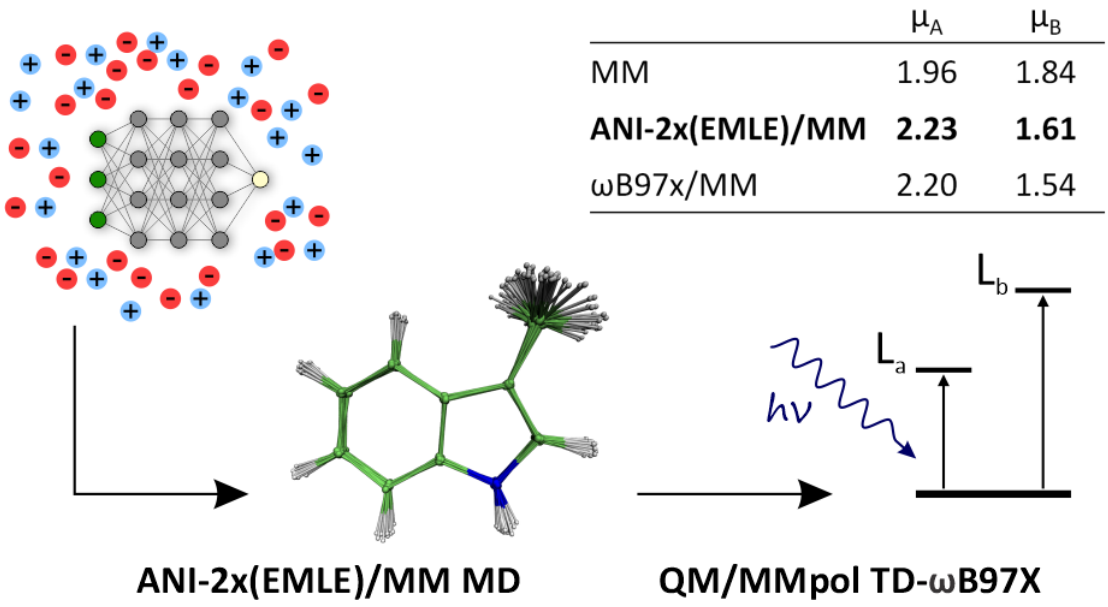
*kirill.zinovjev@uv.es, **carles.curutchet@ub.edu

Abstract

Incorporation of environment and vibronic effects in simulations of optical spectra and excited state dynamics is commonly done combining molecular dynamics with excite state calculations, which allows to estimate the spectral density describing the frequency-dependent system-bath coupling strength. The need for efficient sampling, however, usually leads to the adoption of classical force fields, despite well-known inaccuracies due to the mismatch with the excited state method. Here we present a multiscale strategy that overcomes this limitation by combining EMLE

simulations based on electrostatically embedded ML potentials with the QM/MMPol polarizable embedding model to compute the excited states and spectral density of 3-methyl-indole, the chromophoric moiety of tryptophan that mediates a variety of important biological functions, in gas phase, in water solution and in the human serum albumin protein. Our protocol provides highly accurate results that faithfully reproduce their *ab initio* QM/MM counterparts, thus paving the way for accurate investigations on the interrelation between the timescales of biological motion and the photophysics of tryptophan and other biosystems.

TOC GRAPHICS



KEYWORDS Spectral density of vibronic coupling, QM/MM, machine learning potentials, electrostatic embedding

Simulations of optical spectra and excited state dynamics observed in linear and nonlinear spectroscopies depend on the coupling between excited states and internal and environmental vibrations.[1,2] This is particularly interesting in biological systems, where heterogeneities in the environment can introduce subtle differences in the photoinduced behavior of a given chromophore. The observation of quantum beatings in the 2D electronic spectra of photosynthetic complexes, for example, has motivated extensive studies on the role of exciton-bath interactions and quantum coherence on the optimization of photosynthetic electronic energy transfer.[3–6]

Vibronic effects determine the homogeneous broadening of spectral lineshapes, whereas environment effects led to inhomogeneous broadening and solvatochromic shifts. Incorporation of these effects in theoretical simulations is commonly done combining molecular dynamics (MD) with subsequent excite state calculations on a representative number of snapshots.[7–10] Advanced multiscale excited state methods like polarizable embedding now allow highly accurate excited state calculations in condensed phases.[9,11] Sampling the configurational space, however, still represents a challenge, especially when dealing with large biosystems characterized by motions in multiple timescales.[12] In such cases, the need for efficient sampling leads to the common adoption of classical force fields, despite well-known inaccuracies in the resulting excited state properties due to the geometry mismatch between the classical potential and the excited state method. A remedy for this problem relies on collecting a number of snapshots from the trajectory, and performing multiple quantum/molecular mechanical (QM/MM) geometry optimizations of the chromophores.[1,13–15] This allows an accurate description of environment effects, and vibronic effects and the Franck-Condon spectrum can later be obtained by computing the normal modes of the ground and excited states, or only the ground state in vertical gradient methods, assuming a displaced harmonic oscillator model.[1,16]

The main limitation of vibronic models based on normal mode analysis is that environment and vibronic effects are not treated consistently. Instead of performing multiple QM/MM optimizations along classical trajectories, thus, it is more attractive to directly perform QM/MM MD simulations avoiding the use of classical potentials. Fourier transform of the classical autocorrelation function of energy gap fluctuations then allows to estimate the spectral density (SD), $J(\omega)$, a function that encapsulates the information of the frequency-dependent coupling between nuclear degrees of freedom and the excitation.[12,17–21] Whereas deriving spectral densities from classical trajectories have been shown to lead to considerable errors, recent studies have shown that highly accurate SDs can be derived from Born-Oppenheimer MD QM/MM simulations based on semiempirical or density functional theory (DFT) potentials.[15,22–25] By using a quantum mechanical (QM) method for the photoactive region the sampled conformations explore a potential energy surface (PES) more consistent with the QM excited state method of choice, resulting in more reliable predictions of excitation energies, spectral densities, and other properties like transition dipole moments. However, the cost of high-level QM/MM potentials is still too high for routine simulations, especially in multichromophoric systems like photosynthetic complexes, limiting the choice of QM Hamiltonian often to cheap semiempirical methods.

A possible solution to the high cost of QM/MM simulations is to employ a machine learning (ML) potential. The promise of ML potentials (MLPs) is to provide energies and geometries with quality comparable to the reference QM method used for model training while at significantly lower computational cost.[26–28] Nevertheless, precise MLPs are still computationally much more expensive than molecular mechanics (MM) forcefields. This is particularly the case for biomolecular simulations, where the size of the system and large variety of atomic environments makes the description of the entire system with an ML impractical. Moreover, ML potentials are

not designed to handle accurately long-range interactions, which are key in biosystems. A convenient strategy thus relies on developing hybrid ML/MM approaches that similar to QM/MM allow to employ ML potentials only for the region of the system where a precise description is necessary, while treating the rest with a cheap MM forcefield.[29] Among these approaches is the recently introduced electrostatic machine learning embedding (EMLE) scheme.[30,31] EMLE has three key advantages. First, it explicitly treats the polarization of the ML part using a model based on atomic polarizabilities.[32,33] Second, it treats the impact of the MM environment as a correction term to the *in vacuo* energy of the ML part. This allows to employ ML potentials that were trained to predict gas phase energies (as most existing ML potentials are) without considering the existence of the MM environment. Third, it is designed to be used as a drop-in replacement for a QM engine in standard electrostatic embedding. Therefore, it can be directly employed in existing QM/MM codes without any modifications. As a result, EMLE allows to run ML/MM simulations with existing gas phase ML potentials and standard QM/MM MD packages, without any code changes.

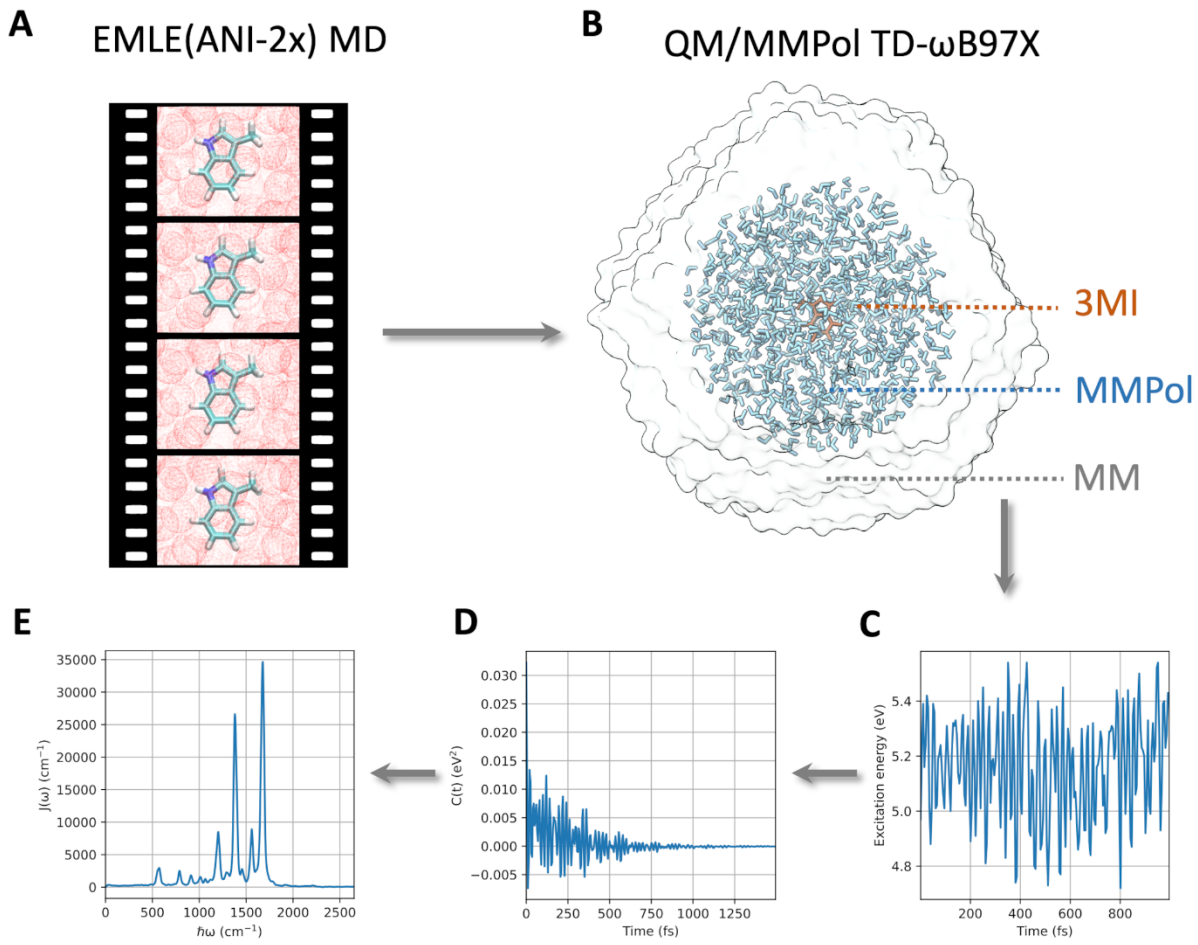


Figure 1. Protocol for the simulation of the spectral density of vibronic coupling of 3-methyl-indole based on EMLE MD simulations and QM/MMPol excited state calculations. **A** ANI-2x(EMLE) MD simulations are performed to sample the vibrations of the chromophore embedded in the MM environment. **B** The trajectory is postprocessed performing TD-ωB97X QM/MMPol excited state calculations every 5 fs. **C** Fluctuations of excitation energies computed along MD trajectories. **D** Autocorrelation function of excitation energy fluctuations along MD trajectories. **E** Spectral densities of vibronic coupling derived from the Fourier transform of the autocorrelation function of excitation energy fluctuations.

Here, we combine EMLE simulations with the QM/MMPol polarizable embedding model[11,34] to compute the excited states and the spectral density of a key biological chromophore, 3-methyl-indole (3MI), in aqueous solution and in the protein human serum albumin (HSA), following the protocol outlined in Fig. 1. 3MI represents the chromophoric moiety of tryptophans, which mediate a variety of important biological functions. After UV light absorption, tryptophan shows a strong fluorescence that is highly sensitive to the surrounding environment, which explains its wide use as a probe to inspect and monitor protein structures and biological processes.[35,36] Accurately understanding how tryptophan photochemistry is tuned by different biological microenvironments is thus of interest for a variety of techniques. We thus generated structures of 3MI solvated in water and embedded in the HSA protein at the ANI-2x/MM level of theory for subsequent calculation of excitation energies and transition moments with QM/MMPol.[34] ANI-2x is a generic gas phase ML potential trained on ωB97X/6-31G(d) energies of small molecules (including H, C, N, O, S, F, Cl).[37] To separately estimate the importance of a good *in vacuo* potential and of a correct treatment of embedding in different environments, three separate sets of simulations were performed. First, we compared the performance of ANI-2x to classical MM and *ab initio* potentials by propagating the MD trajectories of 3MI *in vacuo* using the Amber ff14SB[38] force field and the ωB97X functional with the 6-31G(d) basis set, respectively.[39] In MM simulations, thus, 3MI was described adopting the bonded and Lennard-Jones parameters corresponding to the side chain of Trp in ff14SB, whereas RESP charges where derived at the HF/6-31G(d) level of theory on a B3LYP/cc-pVTZ optimized geometry. In this case, the simulation system contained only 3MI, so no solvent was present, and no embedding scheme was employed. Five replicas of 20ps were obtained with each Hamiltonian, taking snapshots every 5 fs. The second and third sets of simulations followed the same protocol, but the system contained solvated 3MI or solvated HSA

bound to (S)-flurbiprofen. In all cases, MM atoms within 12 Å from 3MI were considered in QM or ML calculations. For the simulations of HSA, we started from an equilibrated structure taken from previous work.[40] Water was described by the TIP3P model,[41] whereas for the protein and flurbiprofen we adopted Amber ff12SB parameters. In HSA simulations, the 3MI side chain of Trp214 was chosen as the chromophore described using QM, MM or ML potentials, and QM/MM and ML/MM boundaries were treated using the link atom scheme.

All simulations were performed using the sander MD code from the AmberTools23 package.[42] Reference QM/MM calculations were performed using ORCA package.[43] In case of simulations involving the ANI-2x potential, sander was coupled to the emle-engine[31] code. emle-engine implements EMLE embedding and provides an interface to the torchani[44] package implementing the ANI-2x potential.[37] In all cases, the snapshots were used to calculate excitation energies (ΔE) and transition dipole moments (μ^T) at the ωB97X/6-31G(d) level of theory using the time-dependent density functional theory (TD-DFT) implementation of QM/MMPol in the development version of Gaussian.[45] For comparison, we also calculated the excited states using the B3LYP hybrid functional. The MMPol environment was described using the Amber pol12 AL polarizable force field, with water charges taken from previous work[46] and polarization-consistent ESP charges for flurbiprofen derived at the B3LYP/6-31G(d) level of theory using the Polchat tool.[47] We used MMPol and MM cutoff radius values equal to 15 and 30 Å, respectively. 3MI has two nearly degenerate π–π^* excited states that dictate low-energy absorption and fluorescence, termed L_a and L_b according to the Platt nomenclature.[48] These states were carefully identified in each snapshot from the two lowest-lying excited states based on the orientation of the transition dipoles. We used as references the position vector from atom NE1 to CE3 to assign L_a, and the vector CG to CZ2 to assign L_b, where the atom names correspond to the Trp definition in Amber force fields. In our

simulations, the orientations of the L_a and L_b states deviated an average of ~10-20 and ~30-45 degrees from the references, respectively, indicating that state mixing was small. In Fig. S5 from the Supporting Information we show examples of the L_a/L_b energies in trajectories obtained in the different environments.

Spectral densities of vibronic coupling[18,49] where then computed from the Fourier transform of the autocorrelation function of excitation energy fluctuations along MD trajectories:

$$J(\omega) = \frac{\beta\omega}{\pi}\int_0^\infty C^{cl}(t)cos(\omega t)dt \quad (1)$$

where $\beta = 1/(k_B T)$ and $C^{cl}(t)$ is the classical autocorrelation function:

$$C_i^{cl}(t_j) = \frac{1}{N-j}\sum_{k=1}^{N-j} \Delta E_i(t_j + t_k)\Delta E_i(t_k) \quad (2)$$

The autocorrelation function for the L_a and L_b states of 3MI was found to decay within 1 ps, so it was averaged over multiple windows of 4 ps following the protocol reported previously.[22]

For the sake of comparison, we also computed the SD of 3MI in vacuum using the vertical gradient (VG) method from a normal mode analysis of the ground state and TD-DFT calculations performed at the ωB97X/6-31G(d) level of theory. The resulting SDs were broadened using a Lorentzian lineshape with HWHM values of 15 and 35 cm^{-1}.[2,46]

In case of the solvated system, different ML embedding options were tested:

1. EMLE: full EMLE embedding scheme, including models of molecular electronic density and polarization of the ML part.

2. EMLE-NoPol (no polarization): EMLE scheme but without the polarization component.

3. EMLE-Mech (mechanical): as “no polarization” but the atomic electronic densities are collapsed to the nucleus.

4. MM: atomic partial charges taken from the MM forcefield and stay fixed during the simulation.

The results for gas phase, water and HSA protein simulations are shown in Table 1, where we report excitation energies and transition dipole moments computed for the L_a and L_b states of 3MI. In Figure 2 we then report the corresponding errors with respect to benchmark data derived from ωB97X Born-Oppenheimer MD simulations. The agreement of transition energies and dipoles derived from ANI-2x ML-based simulations in gas phase is striking, with errors < 0.01 eV and < 0.01 D for energies and dipoles, respectively. In comparison, the simulations based on the Amber ff14SB force field introduce significant absolute errors ~0.06-0.14 eV and ~0.15-0.22 D. The gas phase results clearly indicate that ANI-2x is superior to the employed MM forcefield and provides values close to the ones obtained at the reference DFT level of theory.

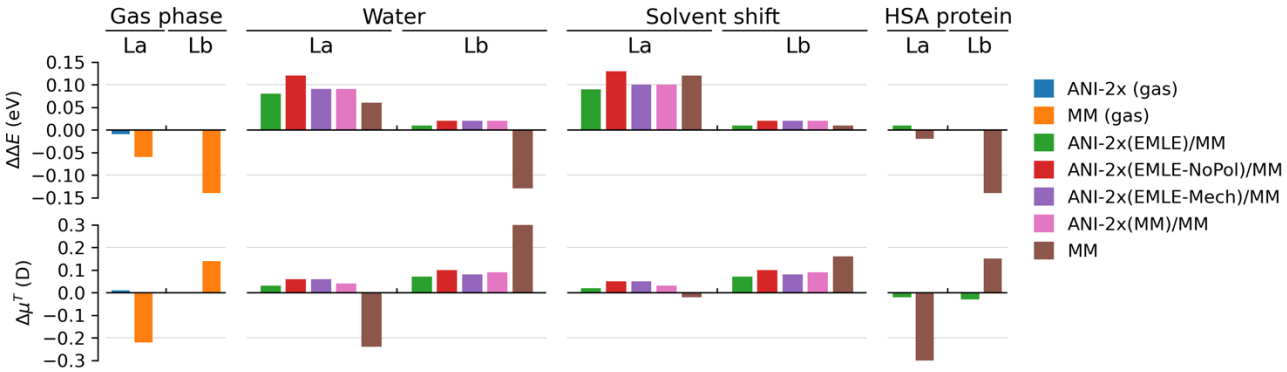


Figure 2. Deviation of mean QM/MMPol TD-ωB97X excitation energies and transition dipole moments of 3-methyl-indole derived from ML and MM-based MD trajectories with respect to benchmark ωB97X Born-Oppenheimer MD simulations. Top (left to right): excitation energies in

gas phase, excitation energies in aqueous solution, gas to water solvent shifts in excitation energies, and excitation energies in HSA. Bottom (left to right): transition dipoles in gas phase, transition dipoles in aqueous solution, gas to water solvent shifts in transition dipoles, and transition dipoles in HSA.

Table 1. Mean values of QM/MMPol TD-ωB97X excitation energies (eV) and transition dipole moments (D) of 3-methyl-indole derived from QM, MM and ML-based MD trajectories.

	MD	$\Delta E\ L_a$	$\mu^T\ L_a$	$\Delta E\ L_b$	$\mu^T\ L_b$
gas phase	ωB97X	5.29	2.11	5.14	1.51
	MM	5.23	1.89	5.00	1.65
	ANI-2x	5.28	2.12	5.14	1.51
solvated	ωB97X/MM	5.10	2.20	5.13	1.54
	MM	5.16	1.96	5.00	1.84
	ANI-2x(EMLE)	5.18	2.23	5.14	1.61
	ANI-2x(EMLE-NoPol)	5.22	2.26	5.15	1.64
	ANI-2x(EMLE-Mech)	5.19	2.26	5.15	1.62
	ANI-2x(MM)	5.19	2.24	5.15	1.63
HSA protein	ωB97X/MM	5.17	2.30	5.12	1.76
	MM	5.15	1.95	4.98	1.91
	ANI-2x(EMLE)	5.18	2.28	5.12	1.73

The same trend is maintained in water and protein simulations – ANI-2x with EMLE embedding improves the prediction of both excitation energies and transition dipoles. The only exception is the excitation energy of L_a state in water, where MM benefits from error cancelation and provides a value that is slightly closer to the DFT reference compared to EMLE(ANI-2x). Indeed, this trend

is not observed for the dipole, in which MM shows a considerable error -0.24 eV for L_a, and in the HSA protein environment both excitation energies and dipoles for the L_a state are nicely reproduced by ANI-2x(EMLE). Overall, the ML/MM simulations show an excellent behavior. In all cases, they reproduce transition dipole moments with errors < 0.1 D, compared to values ~0.15-0.35 eV observed in MM data. In terms of excitation energies, in water the energy of the L_b state is also nicely reproduced with an error ~0.02 eV, whereas for L_a the errors are somewhat larger in the range 0.08-0.12 eV, as previously discussed. In the HSA protein, however, both L_a and L_b excitation energies are nicely reproduced, with errors below 0.01 eV.

For all observables in water solution, inclusion of polarization in the full EMLE embedding scheme provides a minor, but non-negligible improvement of the predictions. In water, however, polarization is not expected to have a significant impact on the generated geometries, but this could change in complex biological environments, like in the HSA protein. Tryptophan indeed establishes frequent cation-π interactions with Lys or Arg residues, an interaction known to depend heavily on the polarization of the indole ring.[50–53] Thus, inclusion of polarization could be important to describe the high sensitivity of Trp spectroscopic properties to the surrounding environment. Polarization is indeed important in a variety of key biological processes, and extensive efforts are devoted for example to the development of polarizable force fields in molecular simulations.[54] In our case, we address the properties of Trp214 in HSA, which establishes a cation-π interaction with Lys199 and interacts with nearby Arg218, as shown in Fig. 3. Our full EMLE embedding results in HSA lead to an excellent agreement with the data derived from benchmark QM/MM MD simulations, both in terms of excitation energies and transition dipole moments, suggesting EMLE can handle the polarization exerted by those cationic residues on Trp adequately.

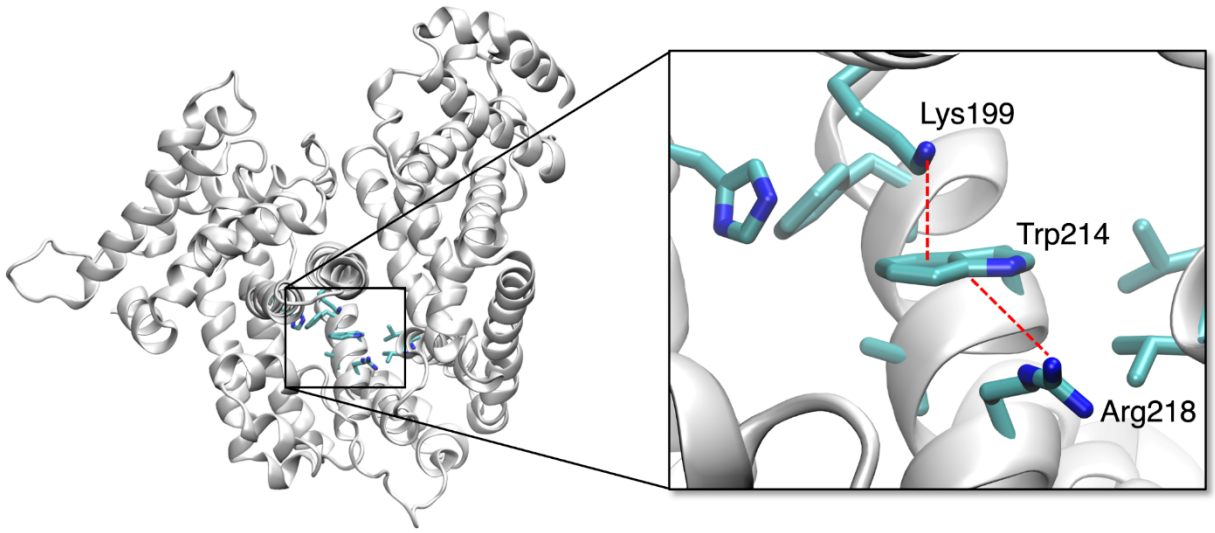


Figure 3. Arrangement of Trp214 in the HSA protein in the last frame of a QM/MM MD replica. Trp214 establishes a cation-π interaction with Lys199 and interacts with nearby Arg218.

Compared to experimental data, our QM/MMPol TD-ωB97X calculations correctly predict that L_b is the lowest excited state in gas phase, although they lead to overestimated energies and an energy gap between L_a and L_b that is too small. This behavior for indole has been previously observed for long-range corrected functionals,[35,55] with hybrid and meta-GGA functionals leading to an incorrect order of the states, this latter point also confirmed by our TD-B3LYP data (see Supporting Information). Beyond this, TD-B3LYP results are qualitatively very similar to those based on TD-ωB97X, supporting the fact that ANI-2x with EMLE embedding dramatically improves the prediction of excitation energies and transition dipoles.

On the other hand, the computed solvatochromic shifts between gas phase and water solution indicate that L_a is strongly stabilized by the polar environment compared to L_b, a well-known behavior leading to the level inversion mechanism often invoked to explain Trp properties in biosystems.[35] In this case, solvent shifts for L_a obtained with ML/MM underestimate those based on QM/MM. However, in the HSA protein we observe a similar strong stabilization of the L_a state compared to L_b, which is nicely reproduced by ML/MM results.

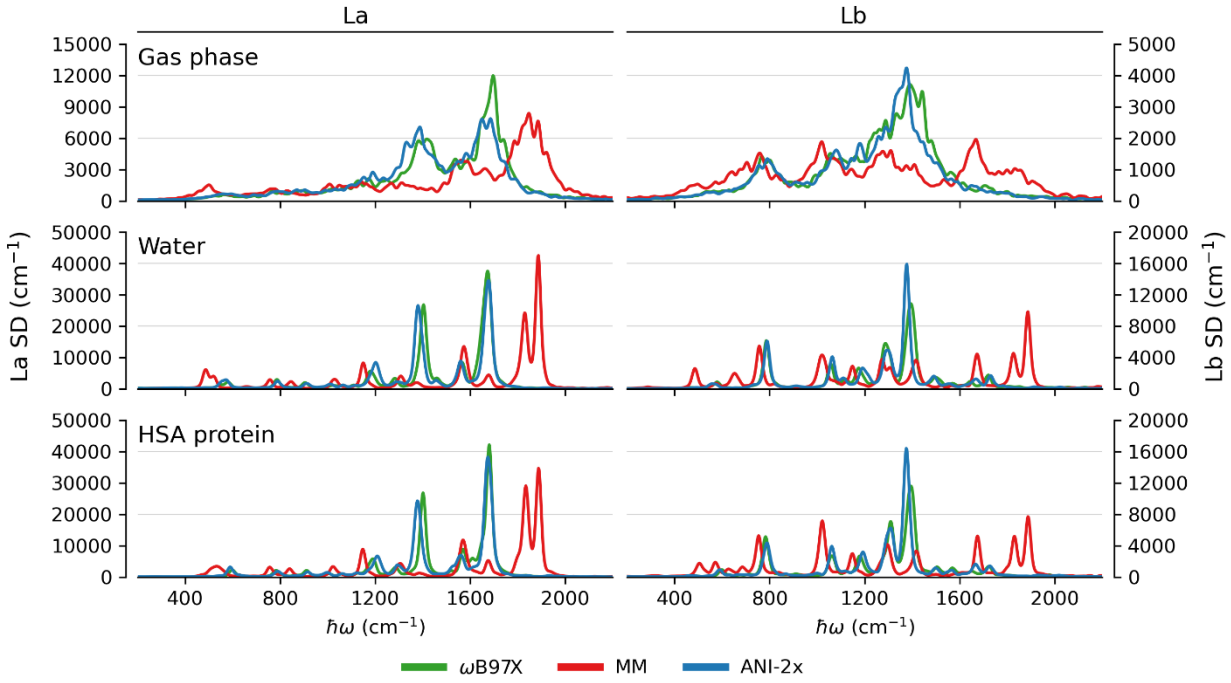


Figure 4. Comparison of spectral densities derived for the L_a and L_b states of 3-methylindole. ANI-2x stands for pure ANI-2x for gas phase and ANI-2x(EMLE) for the systems in water and in the HSA protein environment.

Excitation energies typically depend on the choice of QM method, and *ad hoc* corrections can be applied to eventual systematic errors. Much more challenging is however to describe the coupling between excitations and vibrations that are encapsulated in the SDs. In this task, the improvement

introduced by ML compared to MM is especially evident, both in gas phase, in water solution and in HSA, as shown in Figure 4. In vacuum, the SD of L_a is characterized by two large peaks at ~1400 and ~1700 cm^{-1} whereas for L_b the SDs still display the peak at ~1400 cm^{-1} but lack the one at ~1700 cm^{-1}. These features are beautifully quantitatively reproduced for both states at all frequency ranges, with only the intensity of the peak at ~1700 cm^{-1} in L_a being slightly underestimated. This result is in stark contrast with MM data, which miss the position of the main peaks, and instead leads to high-frequency peaks at ~1850 cm^{-1} and ~1670 cm^{-1} in the L_a and L_b SDs, respectively.

In water solution or in the HSA environment, indole vibrations are modulated by the surrounding solvent and protein motions, an interplay that cannot be tackled by methods based on normal modes. For 3MI, this leads to some visible changes in the structure of the SD and sharper peaks compared to gas phase, as shown in Fig. 4. In this case, again the performance of ML-based simulations based on ANI-2x with EMLE embedding is striking, reproducing almost quantitatively the SDs at all frequency ranges, whereas MM SDs again miss the main peaks and lead to overestimated high-frequency features. The inclusion of polarization in the full EMLE embedding scheme in this case provides minor variations in the intensity of the peaks (see Fig. S2 in the Supporting Information).

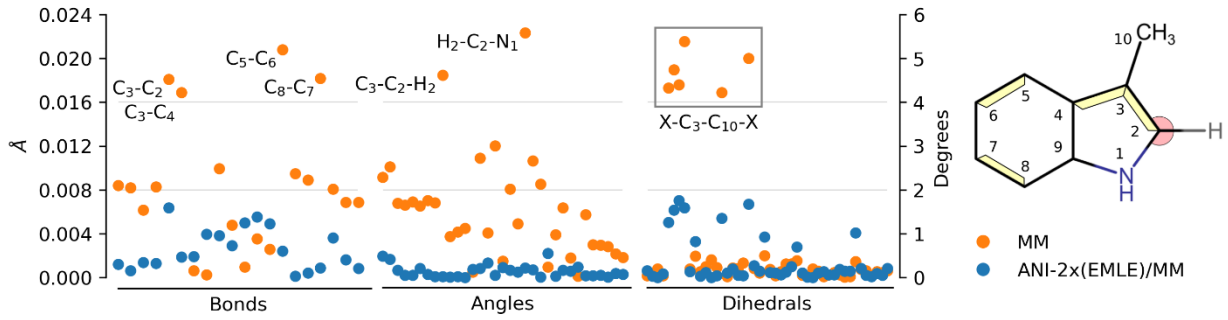


Figure 5. Absolute deviations of average bond, angle and dihedral angle values obtained from ANI-2X(EMLE)/MM simulations of 3MI in water compared to reference ωB97X/MM values. Bonds and angles with the largest deviations at the MM level are highlighted on the 3MI structure on the right: yellow for the bonds and pink for the angles. All the dihedrals with largest deviations correspond to the rotation of the methyl group (C_3-C_{10} bond).

To understand the basis behind the observed differences between ANI-2x(EMLE) and MM predictions, we compared the geometries of 3MI sampled along MM and ANI-2x(EMLE)/MM MD trajectories in water solution to the reference geometries from ωB97X/MM MD. Figure 5 shows the absolute deviations of the mean values of all bonds, angles and dihedrals present in the MM topology of 3MI. Clearly, ANI-2x(EMLE)/MM results in a conformational ensemble much closer to QM/MM than the one provided by MM. Notably, the bonds and angles with the largest deviations at the MM level play an important role in the vibrational normal mode at $\hbar\omega \approx 1700$ cm^{-1}, characterized by in-plane bendings and C–C and C–N stretchings that lead to ring breathing. This explains, thus, why this vibration, which displays the most intense peak in the observed ωB97X and ANI-2x SDs (Figure 4), is missed by MM.

On the other hand, the six dihedral angles with largest errors all correspond to the rotation of the methyl group. As shown on Fig. S11, MM fails to describe the rotational barrier present at the ωB97X level and results in an almost flat distribution of the dihedral angles, whereas ANI-2x perfectly reproduces the reference distribution.

For completeness, we have also checked the impact of the potential on the solvent structure around 3MI by comparing radial distribution functions of the water oxygens around selected atoms of 3MI. We have not observed any notable differences (Fig. S12). Therefore, we conclude that the

improvement provided by EMLE compared to MM is fully explained by the impact on the conformational ensemble of 3MI, not the surrounding solvent.

To shed light into the sharpening of the SD features in condensed phase, in Fig. S9 and S10 we compare the SDs derived from MD simulations in gas phase and in water to those computed from normal modes using the VG method (in vacuum). The SDs obtained using both approaches in gas phase are in good agreement, once a broadening with a HWHM of 35 cm^{-1} is applied to the Lorentzian lineshape used to build the VG SD. In contrast, the water SDs derived from MD trajectories are in better agreement with the VG SDs derived adopting a HWHM of 15 cm^{-1}. The broader features in the vacuum SDs then explains why the SDs are less smooth than those obtained in solution, as they probably require larger samplings to be well converged.

Overall, our results show the dramatic improvement of ML/MM over pure MM simulations. Indeed, important discrepancies between MM and QM-derived SDs have already been observed in systems like pigment-protein complexes.[22] For example, a shifted equilibrium position of the PES with respect to the QM one can lead to exaggerated fluctuations and overestimated peaks,[20] and an SD derived from classical MD will feature peaks at the MM frequencies rather than at frequencies corresponding to the QM method used for excited states, leading to errors in the position of the SD peaks and a redistribution of their intensities.[2] This is reflected in the fact that SDs based on MM are typically more sensitive to the choice of force field than to the QM method used for excited states.[56] Overall, our results indicate that ML/MM simulations explore a PES highly consistent with the QM/MM one. Of course, this agreement is helped by the fact that we run dynamics with ωB97X/6-31G(d), the same level of theory used to train ANI-2x on energies and forces of small molecules and that 3MI is a relatively rigid system, facilitating training of an MLP. Nevertheless, our results here show that ANI-2x combined with EMLE electrostatic

embedding is also able to reproduce with high fidelity the PES of a small molecule explored in aqueous solution and in a protein environment.

Performance-wise, ML potentials provide significant reductions in computational cost even compared to the most efficient QM codes. For consistency with the reference QM/MM calculations, the ML/MM MD simulations presented in this work were performed with the sander interface of the emle-engine, for which the computational bottleneck is the socket-based communication between sander and the emle-server code. Nevertheless, even in this case the performance of ML/MM MD is ≈ 200 ps/day, 300-fold higher than that of sander+ORCA running on a single core (≈0.5 ps/day). However, such comparison is not fair, since emle-engine employs a graphical processing unit (GPU) that has much more computing power. For a fair comparison and to get an estimate of "best-case scenario" for both QM/MM and ML/MM, we have performed additional QM/MM calculation using QUICK, a GPU-accelerated QM package, and an ML/MM calculation using the OpenMM interface with emle-engine available in the Sire framework.[57] Again, resulting ML/MM simulations are ≈ 300-fold faster (6 ns/day) than their QM/MM counterpart (19 ps/day). A summary of the performance obtained with different codes is provided in Table S2.

Overall, we conclude that the combination of ML potentials with the EMLE electrostatic embedding scheme is optimally suited for studying the interrelation between the photophysics of biological chromophores and the multiple timescales of biomolecular motion. We envision relevant future applications in photobiology, for example the investigation of the role of vibronic coupling in coherent energy transfer or the role of the protein in tuning pigment energies in photosynthetic light harvesting complexes. Alternatively, in cases where long timescales need to be sampled, it will be possible to use accurate simulations based on ML/MM to train ML models

directly designed to derive properties from structure, for example the recent method proposed by Cignoni and co-workers to predict excitonic Hamiltonians of light harvesting complexes.[58]

ASSOCIATED CONTENT

Data Availability Statement

The data that support the findings of this work are publicly available at https://github.com/kzinovjev/emle-qmmmpol

Supporting Information

The Supporting Information is available free of charge and includes: figures of autocorrelation functions, spectral densities and excitation energy trajectories from TD-ωB97X and TD-B3LYP calculations, figure and table with TD-B3LYP mean excitation energies, transition dipoles, and corresponding errors, figures of dihedral angles and radial distribution functions, and table with performances of QM/MM and ML/MM with different implementations.

AUTHOR INFORMATION

Corresponding authors

Kirill Zinovjev — Departamento de Química Física, Universidad de Valencia, 46100, Burjassot (Spain); Email: kirill.zinovjev@uv.es

Carles Curutchet — Departament de Farmàcia i Tecnologia Farmacèutica, i Fisicoquímica, Facultat de Farmàcia i Ciències de l'Alimentació and Institut de Química Teòrica i Computacional (IQTCUB), Universitat de Barcelona (UB), Barcelona, Spain; Email: carles.curutchet@ub.edu

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENTS

KZ acknowledges financial support from grant PID2021-123332OB-C22 funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe", a Maria Zambrano fellowship by Ministerio de Universidades (Spain) and computational resources provided by Barcelona Supercomputing Center (Activity Id QH-2023-3-0004). CC acknowledges financial support from the MICIU/AEI (AEI/10.13039/501100011033; grants PID 2020-115812GB-I00, PID2023-151584NB-I00 and CEX 2021-001202-M) and the Catalan Agència de Gestió d'Ajuts Universitaris i de Recerca from Generalitat de Catalunya (GENCAT; 2021SGR00671).

REFERENCES

(1) Zuehlsdorff, T. J.; Shedge, S. V; Lu, S.; Hong, H.; Aguirre, V. P.; Shi, L.; Isborn, C. M. Vibronic and Environmental Effects in Simulations of Optical Spectroscopy. *Annu Rev Phys Chem* **2021**, *72* (1), 165–188. https://doi.org/10.1146/annurev-physchem-090419-051350.

(2) Lee, M. K.; Huo, P.; Coker, D. F. Semiclassical Path Integral Dynamics: Photosynthetic Energy Transfer with Realistic Environment Interactions. *Annu Rev Phys Chem* **2016**, *67* (1), 639–668. https://doi.org/10.1146/annurev-physchem-040215-112252.

(3) Cao, J.; Cogdell, R. J.; Coker, D. F.; Duan, H.-G.; Hauer, J.; Kleinekathöfer, U.; Jansen, T. L. C.; Mančal, T.; Miller, R. J. D.; Ogilvie, J. P.; Prokhorenko, V. I.; Renger, T.; Tan, H.-S.; Tempelaar, R.; Thorwart, M.; Thyrhaug, E.; Westenhoff, S.; Zigmantas, D. Quantum Biology Revisited. *Sci Adv* **2020**, *6* (14), eaaz4888. https://doi.org/10.1126/sciadv.aaz4888.

(4) Wang, L.; Allodi, M. A.; Engel, G. S. Quantum Coherences Reveal Excited-State Dynamics in Biophysical Systems. *Nat Rev Chem* **2019**, *3* (8), 477–490. https://doi.org/10.1038/s41570-019-0109-z.

(5) Jumper, C. C.; Rafiq, S.; Wang, S.; Scholes, G. D. From Coherent to Vibronic Light Harvesting in Photosynthesis. *Curr Opin Chem Biol* **2018**, *47*, 39–46. https://doi.org/10.1016/j.cbpa.2018.07.023.

(6) Scholes, G. D.; Fleming, G. R.; Chen, L. X.; Aspuru-Guzik, A.; Buchleitner, A.; Coker, D. F.; Engel, G. S.; van Grondelle, R.; Ishizaki, A.; Jonas, D. M.; Lundeen, J. S.; McCusker, J. K.; Mukamel, S.; Ogilvie, J. P.; Olaya-Castro, A.; Ratner, M. A.; Spano, F. C.; Whaley, K. B.; Zhu, X. Using Coherence to Enhance Function in Chemical and Biophysical Systems. *Nature* **2017**, *543* (7647), 647–656. https://doi.org/10.1038/nature21425.

(7) Curutchet, C.; Mennucci, B. Quantum Chemical Studies of Light Harvesting. *Chem Rev* **2017**, *117* (2), 294–343. https://doi.org/10.1021/acs.chemrev.5b00700.

(8) Nogueira, J. J.; González, L. Computational Photophysics in the Presence of an Environment. *Annu Rev Phys Chem* **2018**, *69* (1), 473–497. https://doi.org/10.1146/annurev-physchem-050317-021013.

(9) Mroginski, M.; Adam, S.; Amoyal, G. S.; Barnoy, A.; Bondar, A.; Borin, V. A.; Church, J. R.; Domratcheva, T.; Ensing, B.; Fanelli, F.; Ferré, N.; Filiba, O.; Pedraza-González, L.; González, R.; González-Espinoza, C. E.; Kar, R. K.; Kemmler, L.; Kim, S. S.; Kongsted, J.; Krylov, A. I.; Lahav, Y.; Lazaratos, M.; NasserEddin, Q.; Navizet, I.; Nemukhin, A.; Olivucci, M.; Olsen, J. M. H.; Pérez de Alba Ortíz, A.; Pieri, E.; Rao, A. G.; Rhee, Y. M.; Ricardi, N.; Sen, S.; Solov'yov, I. A.; De Vico, L.; Wesolowski, T. A.; Wiebeler, C.; Yang, X.; Schapiro, I. Frontiers in Multiscale Modeling of Photoreceptor Proteins. *Photochem Photobiol* **2021**, *97* (2), 243–269. https://doi.org/10.1111/php.13372.

(10) Giovannini, T.; Egidi, F.; Cappelli, C. Molecular Spectroscopy of Aqueous Solutions: A Theoretical Perspective. *Chem Soc Rev* **2020**, *49* (16), 5664–5677. https://doi.org/10.1039/C9CS00464E.

(11) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable Embedding QM/MM: The Future Gold Standard for Complex (Bio)Systems? *Physical Chemistry Chemical Physics* **2020**, *22* (26), 14433–14448. https://doi.org/10.1039/D0CP02119A.

(12) Gilmore, J.; McKenzie, R. H. Quantum Dynamics of Electronic Excitations in Biomolecular Chromophores: Role of the Protein Environment and Solvent. *J. Phys. Chem. A* **2008**, *112* (11), 2162–2176. https://doi.org/10.1021/jp710243t.

(13) Lee, M. K.; Bravaya, K. B.; Coker, D. F. First-Principles Models for Biological Light-Harvesting: Phycobiliprotein Complexes from Cryptophyte Algae. *J Am Chem Soc* **2017**, *139* (23), 7803–7814. https://doi.org/10.1021/jacs.7b01780.

(14) Ozaydin, B.; Curutchet, C. Unraveling the Role of Thermal Fluctuations on the Exciton Structure of the Cryptophyte PC612 and PC645 Photosynthetic Antenna Complexes. *Front Mol Biosci* **2023**, *10*. https://doi.org/10.3389/fmolb.2023.1268278.

(15) Kjellgren, E. R.; Haugaard Olsen, J. M.; Kongsted, J. Importance of Accurate Structures for Quantum Chemistry Embedding Methods: Which Strategy Is Better? *J Chem Theory Comput* **2018**, *14* (8), 4309–4319. https://doi.org/10.1021/acs.jctc.8b00202.

(16) Loco, D.; Cupellini, L. Modeling the Absorption Lineshape of Embedded Systems from Molecular Dynamics: A Tutorial Review. *Int J Quantum Chem* **2019**, *119* (1). https://doi.org/10.1002/qua.25726.

(17) Cupellini, L.; Corbella, M.; Mennucci, B.; Curutchet, C. Electronic Energy Transfer in Biomacromolecules. *Wiley Interdiscip Rev Comput Mol Sci* **2019**, *9* (2), e1392. https://doi.org/10.1002/wcms.1392.

(18) Valleau, S. S.; Eisfeld, A.; Aspuru-Guzik, A. A. On the Alternatives for Bath Correlators and Spectral Densities from Mixed Quantum-Classical Simulations. *J Chem Phys* **2012**, *137* (22), 224103. https://doi.org/10.1063/1.4769079.

(19) Damjanović, A.; Kosztin, I.; Kleinekathöfer, U.; Schulten, K. Excitons in a Photosynthetic Light-Harvesting System: A Combined Molecular Dynamics, Quantum Chemistry, and Polaron Model Study. *Phys Rev E* **2002**, *65* (3), 031919. https://doi.org/10.1103/PhysRevE.65.031919.

(20) Zwier, M. C.; Shorb, J. M.; Krueger, B. P. Hybrid Molecular Dynamics-Quantum Mechanics Simulations of Solute Spectral Properties in the Condensed Phase: Evaluation of Simulation Parameters. *J Comput Chem* **2007**, *28* (9), 1572–1581. https://doi.org/10.1002/jcc.20662.

(21) Mercer, I. P.; Gould, I. R.; Klug, D. R. A Quantum Mechanical/Molecular Mechanical Approach to Relaxation Dynamics: Calculation of the Optical Properties of Solvated Bacteriochlorophyll-a. *J. Phys. Chem. B* **1999**, *103* (36), 7720–7727. https://doi.org/10.1021/jp990284d.

(22) Rosnik, A. M.; Curutchet, C. Theoretical Characterization of the Spectral Density of the Water-Soluble Chlorophyll-Binding Protein from Combined Quantum Mechanics/Molecular Mechanics Molecular Dynamics Simulations. *J Chem Theory Comput* **2015**, *11* (12), 5826–5837. https://doi.org/10.1021/acs.jctc.5b00891.

(23) Blau, S. M.; Bennett, D. I. G.; Kreisbeck, C.; Scholes, G. D.; Aspuru-Guzik, A. Local Protein Solvation Drives Direct Down-Conversion in Phycobiliprotein PC645 via Incoherent Vibronic Transport. *Proceedings of the National Academy of Sciences* **2018**, *115* (15), E3342–E3350. https://doi.org/10.1073/pnas.1800370115.

(24) Maity, S.; Bold, B. M.; Prajapati, J. D.; Sokolov, M.; Kubař, T.; Elstner, M.; Kleinekathöfer, U. DFTB/MM Molecular Dynamics Simulations of the FMO Light-Harvesting Complex. *J Phys Chem Lett* **2020**, *11* (20), 8660–8667. https://doi.org/10.1021/acs.jpclett.0c02526.

(25) Bondanza, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. The Multiple Roles of the Protein in the Photoactivation of Orange Carotenoid Protein. *Chem* **2020**, *6* (1), 187–203. https://doi.org/10.1016/j.chempr.2019.10.014.

(26) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem Rev* **2021**, *121* (16), 10142–10186. https://doi.org/10.1021/acs.chemrev.0c01111.

(27) Martin-Barrios, R.; Navas-Conyedo, E.; Zhang, X.; Chen, Y.; Gulín-González, J. An Overview about Neural Networks Potentials in Molecular Dynamics Simulation. *Int J Quantum Chem* **2024**, *124* (11), e27389. https://doi.org/10.1002/qua.27389.

(28) Yang, Y.; Zhang, S.; Ranasinghe, K. D.; Isayev, O.; Roitberg, A. E. Machine Learning of Reactive Potentials. *Annu Rev Phys Chem* **2024**, *75* (1), 371–395. https://doi.org/10.1146/annurev-physchem-062123-024417.

(29) Jaffrelot Inizan, T.; Plé, T.; Adjoua, O.; Ren, P.; Gökcan, H.; Isayev, O.; Lagardère, L.; Piquemal, J. P. Scalable Hybrid Deep Neural Networks/Polarizable Potentials Biomolecular Simulations Including Long-Range Effects. *Chem Sci* **2023**, *14* (20), 5438–5452. https://doi.org/10.1039/d2sc04815a.

(30) Zinovjev, K. Electrostatic Embedding of Machine Learning Potentials. *J Chem Theory Comput* **2023**, *19* (6), 1888–1897. https://doi.org/10.1021/acs.jctc.2c00914.

(31) Zinovjev, K.; Hedges, L.; Montagud Andreu, R.; Woods, C.; Tuñón, I.; van der Kamp, M. W. Emle-Engine: A Flexible Electrostatic Machine Learning Embedding Package for Multiscale Molecular Dynamics Simulations. *J Chem Theory Comput* **2024**, *20* (11), 4514–4522. https://doi.org/10.1021/acs.jctc.4c00248.

(32) Thole, B. T. Molecular Polarizabilities Calculated with a Modified Dipole Interaction. *Chem. Phys.* **1981**, *59* (3), 341–350. https://doi.org/10.1016/0301-0104(81)85176-2.

(33) Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non-Covalent Interactions across Organic and Biological Subsets of Chemical Space: Physics-Based Potentials Parametrized from Machine Learning. *J Chem Phys* **2018**, *148* (24). https://doi.org/10.1063/1.5009502.

(34) Curutchet, C.; Munoz-Losa, A.; Monti, S.; Kongsted, J.; Scholes, G. D.; Mennucci, B. Electronic Energy Transfer in Condensed Phase Studied by a Polarizable QM/MM Model. *J. Chem. Theory Comput.* **2009**, *5* (7), 1838–1848. https://doi.org/10.1021/ct9001366.

(35) Brisker-Klaiman, D.; Dreuw, A. Explaining Level Inversion of the La and Lb States of Indole and Indole Derivatives in Polar Solvents. *ChemPhysChem* **2015**, *16*, 1695–1702. https://doi.org/10.1002/cphc.201500073.

(36) Callis, P. R. Binding Phenomena and Fluorescence Quenching. II: Photophysics of Aromatic Residues and Dependence of Fluorescence Spectra on Protein Conformation. *J Mol Struct* **2014**, *1077*, 22–29. https://doi.org/10.1016/j.molstruc.2014.04.051.

(37) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J Chem Theory Comput* **2020**, *16* (7), 4192–4202. https://doi.org/10.1021/acs.jctc.0c00121.

(38) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from

Ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(39) Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J Chem Phys* **2008**, *128* (8), 1–15. https://doi.org/10.1063/1.2834918.

(40) Pinheiro, S.; Curutchet, C. Can Förster Theory Describe Stereoselective Energy Transfer Dynamics in a Protein–Ligand Complex? *J Phys Chem B* **2017**, *121* (10), 2265–2278. https://doi.org/10.1021/acs.jpcb.7b00217.

(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(42) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E.; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M. AmberTools. *J Chem Inf Model* **2023**, *63* (20), 6183–6191. https://doi.org/10.1021/acs.jcim.3c01153.

(43) Neese, F. Software Update: The ORCA Program System—Version 5.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. John Wiley and Sons Inc September 1, 2022, p e1606. https://doi.org/10.1002/wcms.1606.

(44) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J Chem Inf Model* **2020**, *60* (7), 3408–3415. https://doi.org/10.1021/acs.jcim.0c00451.

(45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Janesko, B. G.; Lipparini, F.; Zheng, G.; Sonnenberg, J. L.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, T.; Keith, T.; Kobayashi, R.; Normand, J. Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian Development Version, Revision H.36. Gaussian, Inc.: Wallingford CT 2010.

(46) Corbella, M.; Cupellini, L.; Lipparini, F.; Scholes, G. D.; Curutchet, C. Spectral Variability in Phycocyanin Cryptophyte Antenna Complexes Is Controlled by Changes in the A-Polypeptide Chains. *ChemPhotoChem* **2019**, *3* (9), 945–956. https://doi.org/10.1002/cptc.201900045.

(47) Caprasecca, S.; Curutchet, C.; Mennucci, B. PolChat: A polarisation-consistent charge fitting tool, ver. 4.1.2, Molecolab Tools; University of Pisa. molecolab.dcci.unipi.it/tools. 2014–2018.

(48) Platt, J. R. Classification of Spectra of Cata-Condensed Hydrocarbons. *J Chem Phys* **1949**, *17* (5), 484–495. https://doi.org/10.1063/1.1747293.

(49) *Principles of Nonlinear Optical Spectroscopy*; Mukamel, S., Ed.; Oxford University Press: New York, NY, 1995.

(50) Cubero, E.; Luque, F. J.; Orozco, M. Is Polarization Important in Cation-π Interactions? *Proceedings of the National Academy of Sciences* **1998**, *95* (11), 5976–5980. https://doi.org/10.1073/pnas.95.11.5976.

(51) Sherrill, C. D. Energy Component Analysis of π Interactions. *Acc Chem Res* **2013**, *46* (4), 1020–1028. https://doi.org/10.1021/ar3001124.

(52) Pinheiro, S.; Soteras, I.; Gelpí, J. L.; Dehez, F.; Chipot, C.; Luque, F. J.; Curutchet, C. Structural and Energetic Study of Cation–π–Cation Interactions in Proteins. *Phys. Chem. Chem. Phys.* **2017**, *19* (15), 9849–9861. https://doi.org/10.1039/C6CP08448F.

(53) Liu, H.; Fu, H.; Shao, X.; Cai, W.; Chipot, C. Accurate Description of Cation–π Interactions in Proteins with a Nonpolarizable Force Field at No Additional Cost. *J Chem Theory Comput* **2020**, *16* (10), 6397–6407. https://doi.org/10.1021/acs.jctc.0c00637.

(54) Inakollu, V. S.; Geerke, D. P.; Rowley, C. N.; Yu, H. Polarisable Force Fields: What Do They Add in Biomolecular Simulations? *Curr Opin Struct Biol* **2020**, *61*, 182–190. https://doi.org/10.1016/j.sbi.2019.12.012.

(55) Arulmozhiraja, S.; Coote, M. L. 1L a and 1L b States of Indole and Azaindole: Is Density Functional Theory Inadequate? *J Chem Theory Comput* **2012**, *8* (2), 575–584. https://doi.org/10.1021/ct200768b.

(56) Chandrasekaran, S.; Aghtar, M.; Valleau, S.; Aspuru-Guzik, A.; Kleinekathoefer, U. Influence of Force Fields and Quantum Chemistry Approach on Spectral Densities of BChl a in Solution and in FMO Proteins. *J Phys Chem B* **2015**, *119*, 9995–10004. https://doi.org/10.1021/acs.jpcb.5b03654.

(57) Woods, C. J.; Hedges, L. O.; Mulholland, A. J.; Malaisree, M.; Tosco, P.; Loeffler, H. H.; Suruzhon, M.; Burman, M.; Bariami, S.; Bosisio, S.; Calabro, G.; Clark, F.; Mey, A. S. J. S.; Michel, J. Sire: An Interoperability Engine for Prototyping Algorithms and Exchanging Information between Molecular Simulation Programs. *J Chem Phys* **2024**, *160* (20), 202503. https://doi.org/10.1063/5.0200458.

(58) Cignoni, E.; Cupellini, L.; Mennucci, B. Machine Learning Exciton Hamiltonians in Light-Harvesting Complexes. *J Chem Theory Comput* **2023**. https://doi.org/10.1021/acs.jctc.2c01044.