# UV-Visible Absorption Spectra of Solvated Molecules by Quantum Chemical Machine Learning

Zekun Chen, Fernanda C. Bononi, Charles A. Sievers, Wang-Yeuk Kong, and Davide Donadio*

*Department of Chemistry, University of California Davis*

E-mail: ddonadio@ucdavis.edu

**Abstract**

Predicting UV-visible absorption spectra is essential to understand photochemical processes and design energy materials. Quantum chemical methods can deliver accurate calculations of UV-visible absorption spectra, but they are computationally expensive, especially for large systems or when one computes line shapes from thermal averages. Here, we present an approach to predict UV-visible absorption spectra of solvated aromatic molecules by quantum chemistry (QC) and machine learning (ML). We show that a ML model, trained on the high-level QC calculation of the excitation energy of a set of aromatic molecules, can accurately predict the line shape of the lowest-energy UV-visible absorption band of several related molecules with less than 0.1 eV deviation with respect to reference experimental spectra. Applying linear decomposition analysis on the excitation energies, we unveil that our ML models probe vertical excitations of these aromatic molecules primarily by learning the atomic environment of their phenyl rings, which align with the physical origin of the $\pi \rightarrow \pi^\star$ electronic transition. Our study provides an effective workflow that combines ML with quantum chemical methods to accelerate the calculations of UV-visible absorption spectra for various molecular systems.

# Introduction

Improving our understanding and ability to model light-matter interactions is essential to several branches of chemical research, including biochemistry, environmental chemistry,[1] and renewable energy harvesting conversion. The accurate prediction of UV-visible light absorption spectra is often the first step in the design molecular chromophores, light-harvesting complexes, organic photovoltaics, dyes for photoelectrochemical cells, photoresponsive materials, photocatalysts, and food dyes.[2–8]

Many-body quantum chemical approaches, such as the Bethe-Salpeter equation,[9] linear response (LR) or equation of motion (EOM) coupled-cluster (CC) theories[10] provide exci-

tation energies with accuracy that approaches the golden standard reference of full configuration interaction calculations.[11,12] However, the steep computational cost scaling of these methods (e.g. $O(N^7)$ for LR-CC3, and $O(N^6)$ for EOM-CCSD, i.e. EOM-CC with singles and doubles)[13] makes their applications to large molecular systems impractical. Moreover, to accurately predict spectral line shapes at finite temperature, it is necessary to perform statistical averages over several hundred or even thousands of configurations, thus further increasing the computational demand.[14–23]

Machine Learning (ML) is emerging as an invaluable tool to accelerate quantum chemistry (QC) calculations, providing accurate results at a fractional cost of electronic structure calculations.[24,25] Former studies applied ML to model electronic excitations in molecules inferring structure-property relations from short-range representations of the molecular geometries.[26–33] However, to the best of our knowledge, very few works have focused on accurately predicting the line shapes of UV-visible absorption spectra from ML models. Ye *et al.* used an artificial neural network with internal coordinates and Coulomb Matrices as an input layer to fit the electronic absorption spectrum of N-methylacetamide using molecular structures sampled from classical molecular dynamics (MD) trajectories.[34–36] Xue *et al.* fitted the absorption cross-sections of benzene and 9-Dicyanomethylene using a kernel ridge regression with the displacements from the equilibrium geometry as descriptors.[37,38] A similar approach, which employs linear fitting of excitation energies against molecular coordinates, was used to improve the statistical sampling of the UV-visible absorption spectra of phenol and guaiacol at the air-ice interface and in aqueous solutions.[22] These works rely on time-dependent density functional theory (TDDFT)[39] calculations of the vertical excitation energy (VEE) for several hundreds of molecular configurations to construct suitable training sets, and fit one ML model for each molecule, in order to enhance the convergence of spectral line shapes obtained via statistical averaging. The use of TDDFT limits the accuracy of the VEE calculations, and fitting an *ad hoc* ML model for each molecule limits the generalibility and the predictivity of these approaches. Westermayr *et al.*[40] stepped further to

model UV-Visible Absorption of $CH_2NH_2^+$ and $C_2H_4$ molecules using deep neural-network (DNN) based on the SchNarc[41] architecture and was able to generalize the DNN to predict UV-Visible absorption for three other small molecules. However, extending this model may not be straightforward, especially for larger molecules, due to the inherent complexity of DNN.

Here, we devise a framework that overcomes these limitations and allows one to compute spectral line shapes that are comparable to experimental measurements by coupling high-level QC calculations with a ML model that can be applied to several different molecules. To this scope, we adopt the bispectrum components (BC), an atomic environment descriptor commonly used to develop ML interatomic potentials,[42–45] as the input to a regularized regression model to predict the UV-visible absorption spectra of a set of ten aromatic molecules in aqueous solution. Even though our approach is based on linear regression, its high dimensionality and the necessity to use regularization for feature selection makes it a proper ML model. To pursue chemical accuracy we train the ML model on the EOM-CCSD excitation energies of configurations from first-principles molecular dynamics (FPMD) trajectories. We show that a single ML model predicts the shape of the lowest-energy UV-visible absorption bands of a set of similar molecules with an accuracy comparable to experimental measurements. Furthermore, by training the ML model on a subset of seven molecules we test its capability to retain its predictivity beyond the training set.

In the next section, we provide the outline of the multiscale modeling approach used to compute the reference UV-visible absorption spectra, the features of the proposed ML model, and its parameterization. In the following section we present the results obtained by fitting a unified ML model over the full dataset of ten molecules. We then tested how this ML approach would be generalized to predict the UV-visible spectra of the three molecules left out of the training set. Finally, we establish a connection between the ML descriptors and the electronic excitations, so to provide a chemical interpretation of the ML results. A concluding section summarizes the key results and highlights future perspectives.

# Methods and Models

Our development of a quantum chemically informed statistical learning method to compute spectra line shapes consists of the following steps:

1. FPMD simulations of molecules in aqueous solutions with explicit water;
2. Quantum chemical calculation of the (first) excitation energies for a few hundred frames extracted from the FPMD trajectory, in which the explicit solvent is replaced by a polarizable continuum implicit solvation model;
3. Representation of the molecular geometries in BC;
4. Fit of a linear ML model of the excitation energies as a function of the molecular configuration by regression with norm-one ($\ell_1$) regularization.

This approach is applied to a set of 10 molecules (Figure 1) consisting of a benzene ring with different combinations of the following functional groups: $-NH_2$ (amine), –OH (hydroxyl), $-OCH_3$ (methoxy).

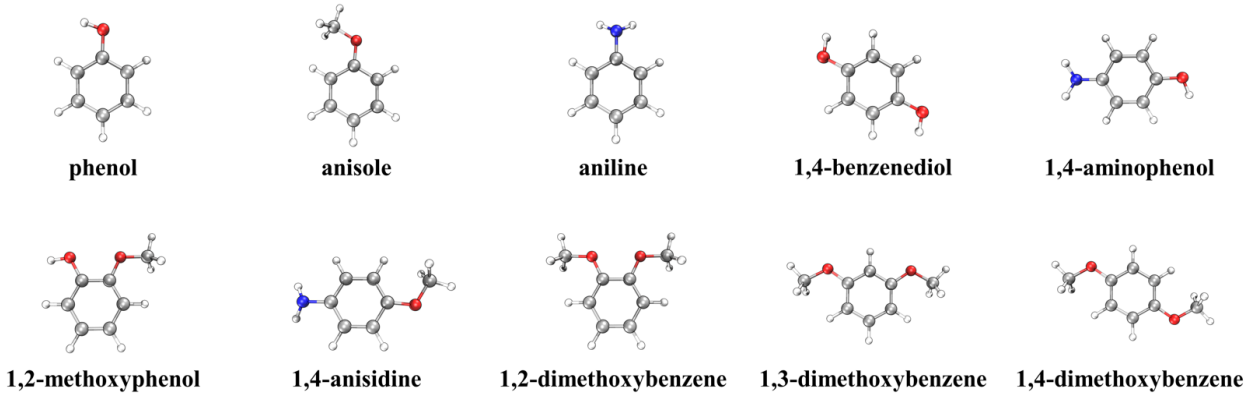



Figure (1) Schematics of the ten aromatic molecules used in this study.

## First-Principles Molecular Dynamics

The calculation of the UV-visible absorption spectra follows a multiscale approach which combines FPMD and excited state calculation. This approach accurately predict line shapes

as it includes both temperature and solvation effects, within the limits of the method used to compute electronic excitations.[17,18,46,47] In this work, density functional theory (DFT) based FPMD simulations are performed using the mixed basis-set Quickstep approach, implemented in the CP2K package.[48,49] We use the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation (GGA) for the exchange and correlation functional[50] with D3 van der Waals corrections.[51] Valence Kohn-Sham orbitals are expanded in real space on a double-$\zeta$ localized valence polarizable (DZVP) basis set,[52] the electron density in reciprocal space is expanded in plane waves up to a cutoff energy of 300 Ry, and core states are treated implicitly using Geodecker-Teter-Hutter norm-conserving pseudopotentials.[53] We run MD simulations of each molecule in aqueous environment in cubic periodic cells with 128 water molecules. FPMD runs were initialized from classical MD simulations in the constant pressure canonical ensemble at ambient conditions, using the TIP4P/Ice model for water and the generalized Amber force field (GAFF) for the organic molecules.[54,55] These systems were then equilibrated for 10 ps by FPMD at constant volume at 300 K. The frames used for absorption spectra calculations were extracted from 100 ps long production runs in the constant-volume canonical ensemble, enforced using the stochastic rescaling algorithm with a coupling constant of $\tau = 10$ ps.[56] The equations of motion were integrated with a timestep of 0.5 fs.

## Absorption Spectra Calculation

For each molecular geometry sampled from the FPMD simulations, VEEs were computed by "domain-based local pair natural orbital similarity transformed equation of motion coupled cluster singles and doubles" (DLPNO-STEOM-CCSD): an efficient coupled-cluster approach with $O(N^4)$ scaling.[57] For the sake of efficiency, the explicit solvent was replaced by a conductor-like polarizable continuum model (CPCM).[58–601] All electron DLPNO-STEOM-CCSD calculations were performed with a triple-$\zeta$ valence polarized Karlsruhe basis set

[1]CPCM was used without Gaussian smearing. This method may lead to discontinuous solvation energies and engender artifacts, which however do not seem to affect our QC results and the ML training set.

(def2-TZVP)[61] using ORCA 4.2.1.[62] For the excited state calculations, the resolution of identity approximation for both Coulomb and exchange integrals was employed to speed up the self-consistent calculation. For each molecule, the lowest-energy absorption band of the UV-visible spectrum is simulated by summing Gaussian functions with a height corresponding to the oscillator strength of the transition and a width of 0.027 eV centered on the calculated VEE for 200 statistically independent configurations chosen from the FPMD trajectory. redThis approach corresponds to the "ensemble method", which provides a reasonable approximation of line widths but does not capture vibronic effects arising from nuclear dynamics.[21]

## Bispectrum Components

In principle, an ML model for absorption spectra would need to predict both VEE and oscillator strengths. Yet, as we observed that for our set of molecules the oscillator strengths are geometry independent, we decided to use a constant oscillator strength model to calculate the absorption spectra. The goal of our ML model is then to predict VEEs as a function of the molecular coordinates, so to spare the computational burden of the QC calculations. To this scope, we represent the molecular configuration in BC.[42] While originally BC was proposed as a descriptor to approximate the Born-Oppenheimer potential energy surface of single element systems,[42] it has been applied to develop ML potential of multi-component systems and predict material properties such as elastic constants, bulk modulus as well as vibrational free energies and entropies of solids.[44,63,64]Compared to other atomic environment descriptors such as smooth overlap of atomic positions (SOAP)[29,65,66] and atom-centered symmetry functions (ACSF),[29,67] BC is more keen at describing the nuances of atomic environments, as it is projected to a more complete set of basis functions with higher dimensions.[31,45] Additionally, in the development of spectral neighbor analysis potentials, BC was formulated to retain a linear relation with the target property.[43,44] This development ensures the resulting ML models to achieve robust performance using only a moderate amount of training data.

Such a trade-off between model complexity and size of the training data is the most critical factor for us to bridge BC with linear regression.

Hereafter, we summarize the key formulations of BC to supplement the following parameterization of our ML model. As reported in the original works,[43] the atomic environment is expressed as the weighted atomic density ($\rho_i(r)$):

$$\rho_i(r) = \delta(r) + \sum_{r_{ii'}<R_{cut,ii'}} f_{cut}(r_{ii'})\omega_{i'}\delta(r - r_{ii'}), \tag{1}$$

where $r_{ii'}$ is the interatomic distance between the central atom $i$ and the neighboring atom $i^{'}$, $f_{cut}$ is the cutoff function to smoothly decay the neighboring atomic density to 0 at the pair-wised cutoff radius ($R_{cut,ii'}$). $R_{cut,ii'}$ is computed by summing over cutoff radii pairs between central and neighboring atoms. The dimensionless weighting factor ($\omega_{i'}$) is used to differentiate the neighboring atoms. After expressed as a sum of $\delta$ functions, $\rho_i(r)$ is expanded in hyperspherical harmonics ($U_j(\theta, \phi, \theta_0)$):

$$\rho_i(r) = \sum_{j=0,\frac{1}{2},1,...}^{\infty} u_j \cdot U_j(\theta, \phi, \theta_0). \tag{2}$$

In equation (2), $u_j$ is the Fourier expansion coefficient given by the inner product between $\rho_i(r)$ and $U_j(\theta, \phi, \theta_0)$. In this implementation, $j$, $j_1$ and $j_2$ are truncated so that $j, j_1, j_2 \leq j_{max}$ to ensure a finite spatial resolution of the weighted atomic density. $2j_{max}$, the even integrable of $j_{max}$, represents a hyperparameter to dictate the number of BC used to fit the ML model. With $u_j$, the bispectrum components ($B_{j_1,j_2,j}$) can be defined as:

$$B_{j_1,j_2,j} = \frac{1}{2j+1} u_{j_1} \otimes_{j_1 j_2 j} u_{j_2} \cdot (u_j)^*, \tag{3}$$

where $\otimes_{j_1 j_2 j}$ represents a coupling product analogous to angular momentum coupling of spherical harmonics. The $\frac{1}{2j+1}$ prefactor ensures $B_{j,j_1,j_2}$ invariant under permutation of the atom indices. In this work, BC were computed using the FitSNAP package.[68]

## Machine Learning Model

The goal of our ML model is to predict the first VEE from the molecular geometry with quantum chemical accuracy. For this purpose we train a linear model on the DLPNO-STEOM-CCSD/def2-TZVP excited state calculations described above using BC as descriptors of the molecular configurations. As described in the previous section and in former works,[42] the BC descriptor consists of projecting the local atomic environment on hyperspherical harmonics for each atomic species. As the BC descriptor is high-dimensional, we apply the least absolute shrinkage and selection operator (LASSO),[69,70] a linear regression model with norm one ($\ell_1$) regularization. Training LASSO consists of minimizing the loss function with respect to the set of coefficients $\beta$:

$$\Delta_{loss}(\beta) = \frac{1}{2N}||\varepsilon_{ML} - \varepsilon_{QC}||_2^2 + \alpha||\beta||_1, \tag{4}$$

where $N$ is the number of molecular configurations extracted from the FPMD simulations, $\alpha$ is the regularization parameter, and $\varepsilon_{QC}$ are the first excitation energies obtained from the QC calculations. The performance of the ML model is assessed through two statistical metrics: mean absolute error (MAE) and mean signed error (MSE). The latter, defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\varepsilon_{QC,i} - \varepsilon_{ML,i}), \tag{5}$$

is useful to spot the occurrence of systematic errors in the predicted VEEs. After training the LASSO model, we compute the UV-visible absorption spectra by estimating $\varepsilon_{ML}$ for 5000 statistically independent configurations from the FPMD trajectories. The final ML spectrum consists of the envelop of Gaussian functions centered on the $\varepsilon_{ML}$ with a width of 0.014 eV. Whereas our ML model is developed in close analogy with the approach used to construct SNAP interatomic potentials,[44] the use of LASSO marks the main difference between SNAP and our excited state model. In a forthcoming section, we discuss extensively

the importance of using a norm $\ell_1$ selection operator to shrink the space of ML parameters and how it enhances the predictivity of the ML models beyond the set of molecules on which it is trained.

A further advantage of using atomic descriptors and the LASSO model is that we can compute the relative contribution of each atom or group of atoms to $\varepsilon_{ML}$, through a linear decomposition analysis, such that:

$$\varepsilon_{ML} = \varepsilon_0 + \sum_{i=1}^{N_{atoms}} \sum_{k=j,j_1,j_2}^{j_{max}} \beta_k^{\gamma_i} B_k^{\gamma_i} = \varepsilon_0 + \varepsilon_\gamma, \tag{6}$$

where $\varepsilon_0$ is the intercept of the LASSO model and $\varepsilon_\gamma$ represents prediction contributions from the atom type $\gamma$ for atom $i$. As $\varepsilon_{ML}$ is a scalar quantity and the LASSO model is linear, $\varepsilon_{ML}$ can be partitioned with respect to different functional groups:

$$\varepsilon_{ML} = \varepsilon_0 + \varepsilon_{NH_2} + \varepsilon_{OH} + \varepsilon_{OCH_3} + \varepsilon_{C_6H_n} \tag{7}$$

$$\%_{group} = \frac{(\varepsilon_{NH_2} + \varepsilon_{OH} + \varepsilon_{OCH_3} + \varepsilon_{C_6H_n})}{\varepsilon_{ML} - \varepsilon_0} \tag{8}$$

where $\varepsilon_{NH_2}$, $\varepsilon_{OH}$ and $\varepsilon_{OCH_3}$ are the prediction contributions from the amine, hydroxyl and methoxy groups. $\varepsilon_{C_6H_n}$ is the prediction contributions from carbon and hydrogen atoms within a phenyl ring ($n = 5$ for phenol, anisole, aniline and $n = 4$ for 1,2-methoxyphenol, 1,4-aminophenol, 1,4-anisdine, 1,4-benzenediol and dimethoxybenzene isomers). $\%_{group}$ is used to express the prediction contributions by percentage.

## Parameterization of the Machine Learning Model

First, to show the baseline performance of our ML approach, we fitted a model using the full data set of ten molecules with 200 VEE per molecule. Then, to test model generalizability, we developed a 7-molecule model by leaving out of the QC calculations of 1,2-methoxyphenol, 1,4-aminophenol and 1,4-dimethoxybenzene from the training set. Be-

fore optimizing the 7- and 10-molecule models, we set the parameters of the BC descriptor, specifically the weights ($\omega_{atom}$) and the cutoff radius for each atomic species ($R_{cut,atom}$). To optimize $R_{cut,atom}$, we fixed $R_{cut,H} = 1.2$ Å, and performed a grid search on $R_{cut}$ for each heavy atom between 2.6 and 3.2 Å. We simplified the choice of $R_{cut,atom}$ parameter by setting $R_{cut,C} = R_{cut,N} = R_{cut}$ and $R_{cut,O} = 1.05R_{cut}$. Since carbon is the major building block of these aromatic molecules, we set the weighting factor of carbon ($\omega_C$) to unity. For the remaining weighting factors ($\omega_H, \omega_O$ & $\omega_N$), a constraint of $\omega_H < \omega_N \leq \omega_O$ was imposed. After determining $\omega_{atom}$ and $R_{cut,atom}$, we chose $2j_{max}$ based on the number of available training samples. At $2j_{max} = 18$, the total number of BC almost equals the number of training samples ($N_{train}$) used to develop the 7-molecule model. Thus, to retain a similar number of input features and training samples, we chose $2j_{max} = 18$. The hyperparameters $R_{cut}$ and $\omega_{atom}$ are optimized for the 7-molecule models and their value is used also for the 10-molecule model. In this process no information is used from the three molecules left out of the training set. Table 1 summarizes the hyperparameters used to compute BC.

**Table (1) Optimized hyperparameters used to compute BC**

| $\omega_H$ | $\omega_C$ | $\omega_N$ | $\omega_O$ | $R_{cut,H}$[Å] | $R_{cut,C}$[Å] | $R_{cut,N}$[Å] | $R_{cut,O}$[Å] |
|---|---|---|---|---|---|---|---|
| 0.75 | 1.0 | 0.8 | 0.9 | 1.20 | 2.80 | 2.80 | 2.94 |

The $\ell_1$ regularization in LASSO is designed to achieve feature elimination,[70] so to prevent overfitting. From Table 2, we see that for both ML models, the number of non-zero features is less than 25 % of the total number of input features ($N_{features}$). The 7- and 10-molecule models retain similar fractions of non-zero features with respect to the size of the training sets so that the model performance can be compared. A 10-fold cross validation was applied to both models to avoid bias from a specific split of training and testing data sets.

**Table (2) Hyperparameters and statistical metrics for the ML models**

| Model | $N_{features\neq 0}/N_{features}$ | $2j_{max}$ | $\alpha$ | $MAE_{avg\ \&\ std}$ [meV] |
|---|---|---|---|---|
| 10-molecule | 318/1544 | 18 | $5.8 \times 10^{-7}$ | 23.05 ± 3.92 |
| 7-molecule | 225/1544 | 18 | $1.2 \times 10^{-6}$ | 23.33 ± 4.13 |

As shown in Table 2, the testing MAE of our 10-molecule model, averaged from the 10-fold cross validation, is 23.05 ± 3.92 meV. A similar MAE, 23.33 ± 4.13 meV, is also observed from the 7-molecule model. Besides MAE, we also introduced MSE to gauge if our ML models systematically overestimate or underestimate excitation energies. As shown in Figure 2, most of the molecules have MSE of less than 5 meV, which implies that no noticeable overestimation or underestimation of excitation energies occurs, except for 1,4-anisidine, for which the predicted excitation energies have MAE and MSE of 25.91 and 11.57 meV.Figure 3 illustrates the testing performance of the 7-molecule model. It can be noticed that both 10 and 7-molecule models have similar testing MAE and MSE for molecules in the training set. It is worth highlighting, from the 10-molecule model, that the predicted excitation energies for every single molecule have MAE far below the intrinsic error of the underlying EOM-CCSD method (70 meV).[71] Even for the excitation energies interpolated from the 7-molecule model, the MAE are merely around half of the 70 meV. Hence, no significant error is introduced by both ML models.

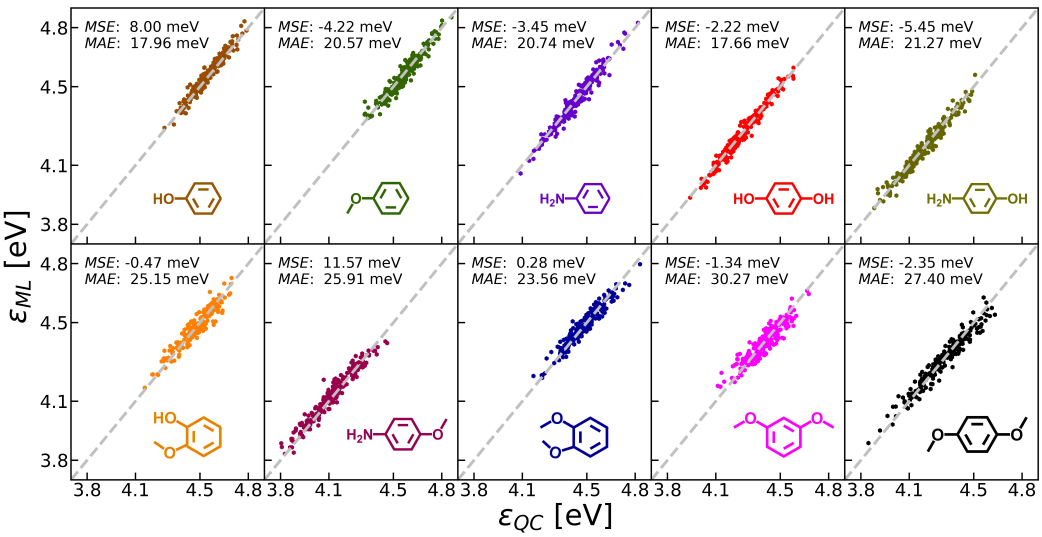



Figure (2) Testing performance for the 10-molecule model. $\varepsilon_{ML}$ is computed by averaging $\varepsilon$ predictions from the 10-fold cross-validation. $\varepsilon_{QC}$ is the quantum mechanically computed excitation energies for the first state.

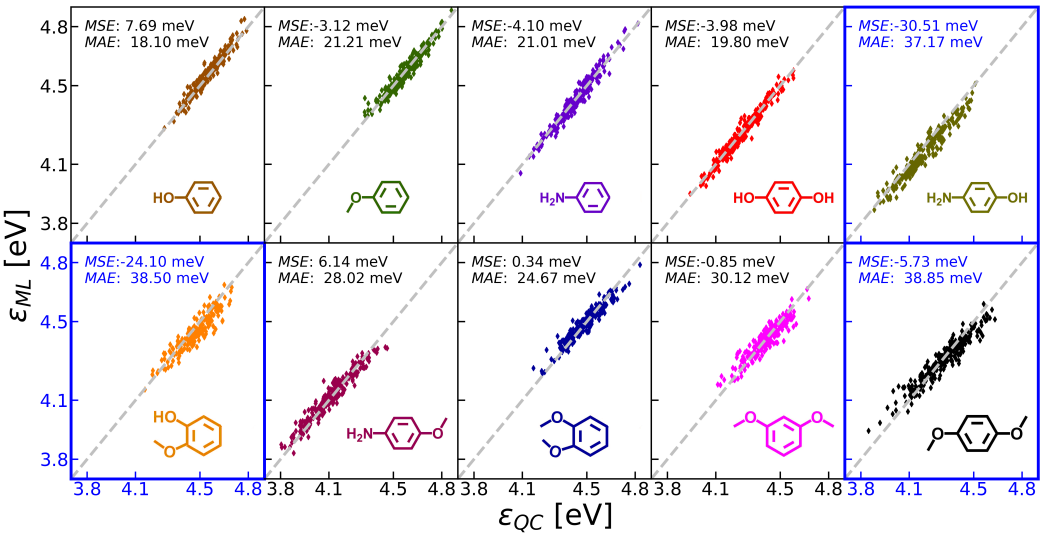



Figure (3) Testing performance for the 7-molecule model. $\varepsilon_{ML}$ is computed by averaging $\varepsilon$ predictions from the 10-fold cross-validation. $\varepsilon_{QC}$ is the quantum mechanically computed excitation energies for the first state. Molecules used in model generalization are indicated with blue panels.

# Results and Discussion

## Machine Learning Absorption Spectra

**Table (3) Summaries of $\varepsilon_{max}$ from experiment, QC and ML spectra. Absolute difference of $\varepsilon_{max}$ between experiment and QC as well as between experiment and ML are indicated in the parenthesis. Molecules used in model-generalization are indicated with blue and bold fonts.**

| Molecule | $\varepsilon_{max,experiment}$ [eV] | $\varepsilon_{max,QC}$ [eV] | $\varepsilon_{max,ML_{10-mol}}$ [eV] | $\varepsilon_{max,ML_{7-mol}}$ [eV] |
|---|---|---|---|---|
| phenol | 4.573 | 4.548 (0.025) | 4.568 (0.005) | 4.569 (0.004) |
| anisole | 4.645 | 4.560 (0.085) | 4.553 (0.092) | 4.556 (0.089) |
| aniline | 4.466 | 4.423 (0.043) | 4.424 (0.042) | 4.421 (0.045) |
| 1,4-benzenediol | 4.291 | 4.231 (0.060) | 4.265 (0.026) | 4.260 (0.031) |
| **1,4-aminophenol** | **4.239** | **4.163 (0.076)** | **4.184 (0.059)** | **4.162 (0.077)** |
| **1,2-methoxyphenol** | **4.535** | **4.481 (0.054)** | **4.479 (0.056)** | **4.451 (0.084)** |
| 1,4-anisdine | 4.263 | 4.078 (0.185) | 4.106 (0.157) | 4.106 (0.157) |
| 1,2-dimethoxybenzene | 4.556 | 4.470 (0.086) | 4.478 (0.078) | 4.479 (0.077) |
| 1,3-dimethoxybenzene | 4.559 | 4.410 (0.149) | 4.408 (0.151) | 4.409 (0.150) |
| **1,4-dimethoxybenzene** | **4.339** | **4.284 (0.055)** | **4.305 (0.034)** | **4.301 (0.038)** |

Figure 4 compares the lowest-energy UV-visible absorption band computed by the multi-scale quantum-chemical approach described in the methods section to experimental measurements for each molecule in aqueous solution.[6,72–77] The QC model is in excellent agreement with experiments for 8 out of 10 molecules, with differences in the center of the peak within less than 0.1 eV and nearly overlapping low-energy tails. Experimental bands tend to be broader on the high energy side, possibly because they encompass the tails of higher excited states. 1,4-anisidine and 1,3-dimethoxybenzene make an exception, with differences in peak positions of 0.18 and 0.15 eV. A possible source of discrepancy is that the GGA functional used in FPMD simulations leads to systematic errors in the geometry of the molecules. The differences between the gas-phase excitation energies of 1,3-dimethoxybenzene computed for geometries optimized with the PBE and PBE0[78] functionals support this hypothesis (Table S1). However, the cost of running FPMD with PBE0 would be excessive for the purpose of this study. Choosing a semilocal GGA functional strikes the desired balance between the accuracy of computing spectra and a reasonable computational cost. Additionally, the experimental conditions may be slightly different from those in the models. The absorption spectrum of 1,4-anisidine was measured at pH= 6.0, while in our FPMD simulations molecules solvated in aqueous solution at neutral pH, and this difference may cause a shift in the UV-visible absorption.[72,79]

The QC spectra were obtained by averaging only a few hundred frames for each molecule. Hence, the calculated spectra for some of these molecules, especially 1,4-dimethoxybenzene, exhibit jagged line shapes (Figure S1). As shown in previous works, ML may be employed to obtain smoother and more refined line shapes.[22,38] The same effect is achieved here, as illustrated in Figure 4. The ML spectra are obtained by averaging over 5,000 frames for each molecule, which guarantees smooth line shapes and converged statistical sampling. At this point, convergence is limited only by the capability of FPMD simulations to sample the configurational space of the molecules in aqueous solution. The main features of the ML spectra obtained with this model are almost indistinguishable from the QC references as ML

predictions have the same level of accuracy as the training data. From Figures 2 and 4, we can conclude that a unified ML model is able to accurately predict the excitation energies and thus allowing us to compute the spectra for thousands of molecular configurations at nearly no additional computational cost.

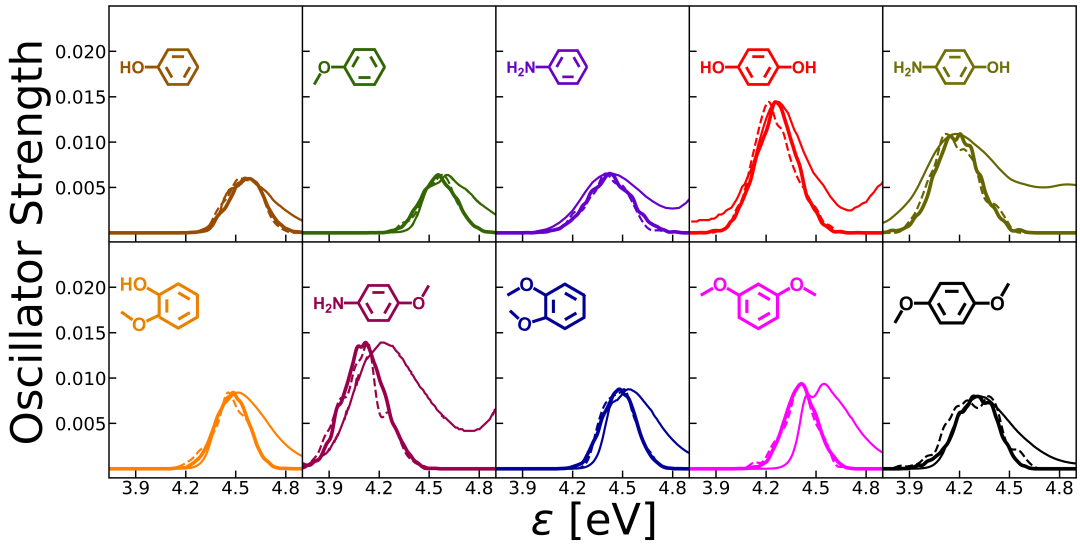



Figure (4) UV-Visible absorption spectra for all 10 aromatic molecules. Thick lines represent the ML spectra predicted using the 10-molecule model and computed with the ensemble method. Thin lines represent the experimental reference.[6,72–76] Dashed lines represent the calculated spectra using the multiscale quantum chemical method.

## Transferability of the Machine Learning Model

Although useful to save computational time to compute the lowest energy band of the absorption spectra, the application of the ML model described so far is limited to the molecules comprised in the training set. Hereafter, we explore whether our approach can be used to predict the absorption spectra of molecules that were not included in the training set, provided that these molecules are similar to those used to construct the ML model. To this scope, we employ the LASSO model fitted with the excited state calculations of 7 molecules to interpolate the excitation energies and the corresponding UV-Visible absorption spectra for 1,2-methoxyphenol, 1,4-aminophenol and 1,4-dimethoxybenzene. These three molecules

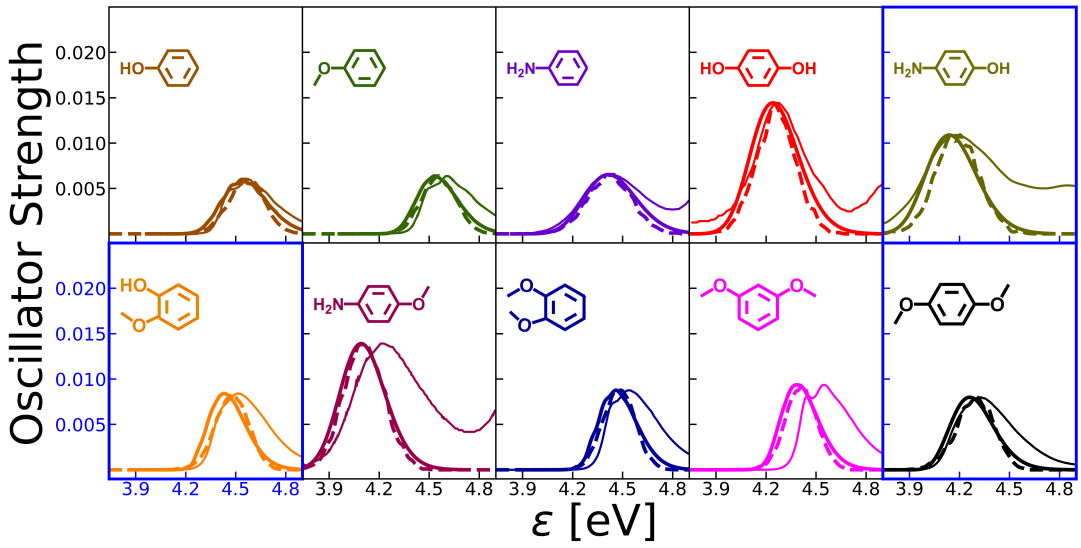



Figure (5) UV-Visible absorption spectra for all 10 aromatic molecules. Thick lines represent the ML spectra computed with the ensemble method (dashed) and with the third order cumulant scheme (solid).[21] Thin lines represent the experimental references.[6,72–76] The spectra of the molecules not included in training set are highlighted with blue graph frames and labels.

consist of a phenyl ring with side groups of the same kind (methoxy, hydroxyl and amine) as the molecules included in the training set, but combined to form different isomers or different molecules altogether. Table 3 and Figure 5 show that the peak positions and the absorption line shapes for the seven molecules in the training set remain the same as those obtained with the 10-molecule ML model. Figure 3 indicates an increase of both MAE and MSE for the three molecules excluded from the training set. MAE, in particular, raises beyond 20 meV for 1,4-aminophenol and 1,2-methoxyphenol. A larger error for these molecules is expected, as the side-group arrangements (e.g., ortho positioning of hydroxyl and methoxyl group, para positioning of amine and hydroxyl group and of two methoxyl groups) are not explicitly known by the ML model and their effects on the excitation energies are inferred from the other seven molecules. Nevertheless, a close comparison between Figures 4 and 5 suggests that the effect of these errors on the predicted UV-visible absorption spectra is small. In fact, the model-generalized spectra remain very similar to those obtained with the 10-molecules

ML model, are in excellent agreement with the reference QC spectra. The difference between $\varepsilon_{max,experiment}$ and $\varepsilon_{max,ML}$ is 0.084 eV, 0.077 eV and 0.038 eV for 1,2-methoxyphenol, 1,4-aminophenol and 1,4-dimethoxybenzene. Considering that the 7-molecule model is developed without using any QC calculations of the three excluded molecules, less than 0.1 eV differences are remarkable.

These results suggest that the 7-molecule model can properly interpolate the electronic excitations of the molecules not included in the training set without further tuning the hyperparameters of the BC descriptors and provide UV-visible absorption spectra in good agreement with experiments. However, while the low-energy tail of the absorption band is reproduced very well, our approach fails to reproduce the skewed experimental line shapes. To improve the line shape predictions we have applied the dynamics-based third-order cumulant scheme, using the fluctuations of the VEE estimated by the ML model along the FPMD trajectory (Fig. 5 solid line).[21,80,81] Whereas we get a minor systematic improvement in the prediction of the line shapes with this approach, the theory still underestimates the intensity of the high-energy part of the absorption band. This may be due to the fact that nuclei are treated as classical particles in FPMD, thus neglecting nuclear quantum effects.

This result shows that local geometric descriptors are sufficient to predict very accurately excitation energies of a group of molecules with similar features. This is a promising starting point for future work on more complex molecules. It is somewhat surprising that small changes in the molecular configurations are sufficient to fully capture solvation effects on excitation energies and lineshapes. While this is auspicious for future works, it is not guaranteed that this will be the case for different types of molecules in different solvation environments. To extend the transferability of this approach to other classes of molecules, given the non-local nature of electronic excitations in large molecules, it may turn out necessary to supplement our approach by including richer physically-based descriptors, e.g. electronic orbitals, and to compute excitation energies as the eigenvalues of ML effective Hamiltonians.[82–84]

## Effect of the $\ell_1$ Regularization

Hereafter, we analyze the features that make the ML model developed in this work accurate and predictive. We first examine the importance of $\ell_1$ regularization. To this aim, we built another 7-molecule model using ordinary least square (OLS)[85] as the underlying ML model. BC for this model were computed using the hyperparameters as summarized in Table S2. The $2j_{max}$ was chosen to be 8 so that both 7-molecule models have similar $N_{features\neq 0}$. Table S2 shows that both 7-molecule models have very similar overall MAE. The 7-molecule + OLS model even achieves lower standard deviation (std) in overall MAE than the 7-molecule model. Surprisingly, two 7-molecule models show striking differences in MAE with respect to each molecule. From Figure 3 and S2, one can see that the 7-molecule model has lower MAE for 5 molecules in the training set. The 7-molecule model also records 8.93, 1.37 and 13.63 meV lower in MAE than the 7-molecule + OLS model when interpolating the excitation energies for 1,2-methoxyphenol, 1,4-aminophenol and 1,4-dimethoxybenzene. The std of MAE from the 7-molecule model (0.724 meV) is only about 12.5% of the std from the 7-molecule + OLS model (5.76 meV) for the three model-generalized molecules. The 7-molecule model is fitted against 225 BC selected from 1544 BC generated with $2j_{max}$ up to 18. As shown in Figure S3, $2j_{max} = 18$ is a sufficiently high order to generate BC with optimal MAE. Meanwhile, feature elimination schemes, such as $\ell_1$ regularization, help capture the essential features of the atomic environments and ensure little to no performance loss as the final model is fitted on a carefully-chosen subset of the initial feature space.[86] Therefore, these 225 BC are composed of comprehensive descriptions of the atomic environments at resolution as high as $2j_{max} = 18$. However, as no feature elimination is imposed in the 7-molecule + OLS model, its feature space can only be confined at lower order of $2j_{max}$ to prevent overfitting. In particular, each of 224 BC generated with $2j_{max} = 8$ is used to fit the 7-molecule + OLS model. Thus, despite almost identical $N_{features\neq 0}$, the difference in the detailed description of the atomic environment eventually leads to a drastic difference in the capability for the two 7-molecule models. Whereas a more standard $\ell_2$ regularization may

be used, the latter is prone to overfitting when the ratio between the number of samples and features approaches unity. It is therefore very advantageous to impose the $\ell_1$ regularization as it not only prevents overfitting, but it also enables the ML model to have a sufficiently large initial feature space, which is critical for the model to be generalized to systems not included in the training set.

## Chemical Interpretation of Machine Learning Results

To trace how the ML models explore the structural similarity among these aromatic molecules, we performed the linear decomposition analysis as formulated above in the methods section. Figures 6 and S4 show that, for $\varepsilon_{ML}$ from both ML models, contributions from the aromatic ring are predominant. This trend is common to all molecules. This result suggests that the observed generalizability across the whole family of aromatic molecules lays primarily on geometric variations of the aromatic ring during the FPMD simulations, and that the primary effects of solvation are local and can be tracked down to configurational changes in the solvated molecule.[87] To interpret this result from a quantum mechanical standpoint, we computed the Natural Transition Orbitals (NTOs)[88–90] of all the aromatic molecules using their equilibrium structures in aqueous solution. As shown in Figures S5, the dominating NTOs for all these aromatic molecules exhibit the characteristics of a $\pi \rightarrow \pi^*$ transition. Such excited state characters are mostly contributed by the aromatic ring, along with a secondary participation of lone pairs on functional groups. Whereas relative contributions of the hydroxyl and amine groups from the NTOs are noticeably larger than those illustrated in the linear decomposition analysis, NTOs confirm the predominant relevance of the carbon atoms in the aromatic ring. Quantitative differences in the relative contributions of the functional groups indicate that the geometrical interpretation of the excitations from ML is qualitative.

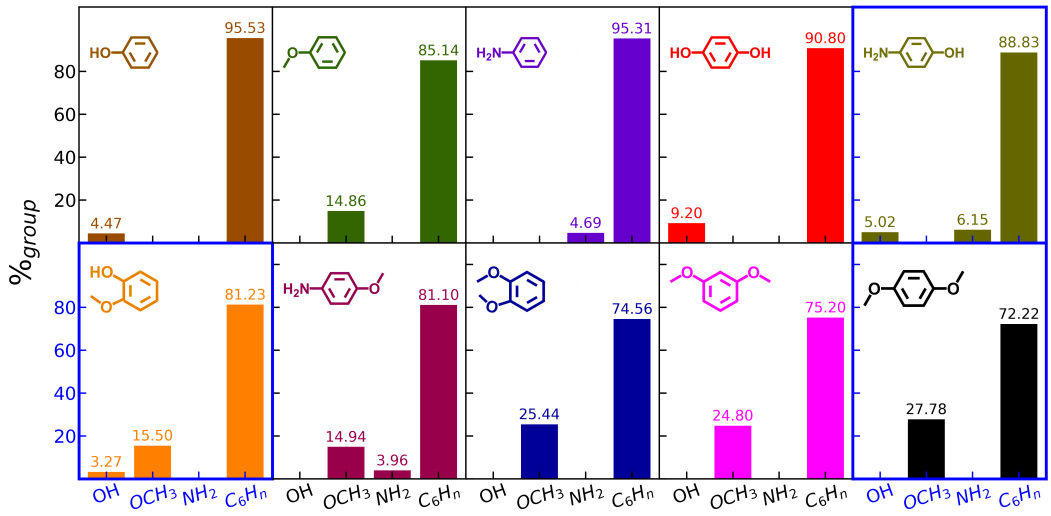



Figure (6) Linear decomposition analysis from the 7-molecule model. $\%_{group}$ are computed by averaging the excitation energy predictions of 5000 frames.

## Machine Learning Higher Excited States

To explore the possibility of extending our machine learning method, we constructed another model, using the same parameters as the 10-molecule model, to predict the second excitation energies. From Figure 7, one can see that the MAE and MSE of the second excited state are noticeably higher than the MAE and MSE of the first excited state but still far below the 70 meV[71] limit of the deployed DLPNO-STEOM-CCSD method. With both the first and second excited-state energies predicted concurrently, the resultant ML absorption enrapture long-wavelength tails convoluted from both excited states. Since not all the experimental spectra are available up to second excited states, we compared QC and ML UV-visible absorption bands for the first two excited states in the space of oscillator strength (Figure 8). Both Figures 7 and 8 prove that our ML methods can be promising in predicting higher excited states.

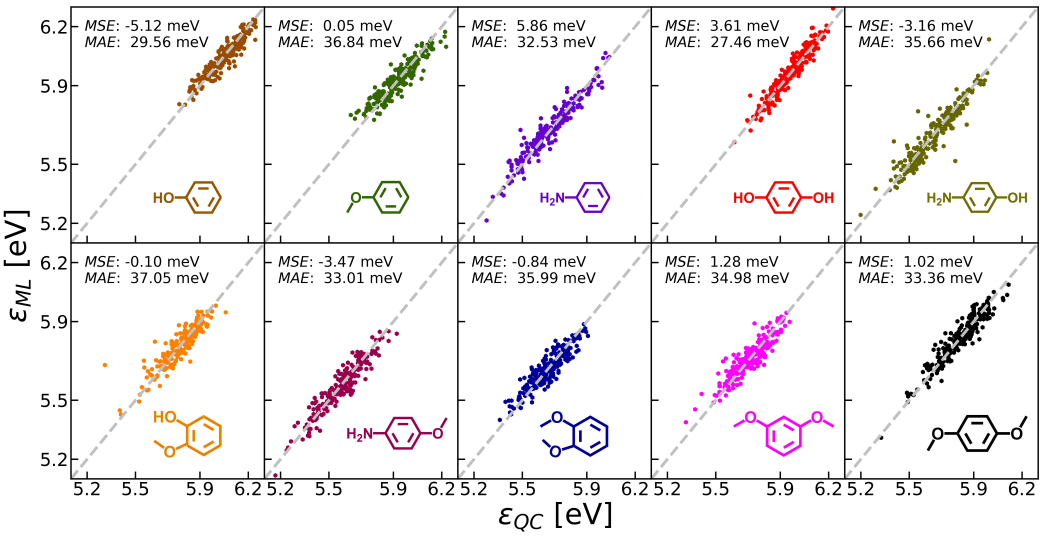



Figure (7) Testing performance for the 10-molecule model. $\varepsilon_{ML}$ is computed by averaging $\varepsilon$ predictions from the 10-fold cross-validation. $\varepsilon_{QC}$ is the quantum mechanically computed excitation energies for the second state.

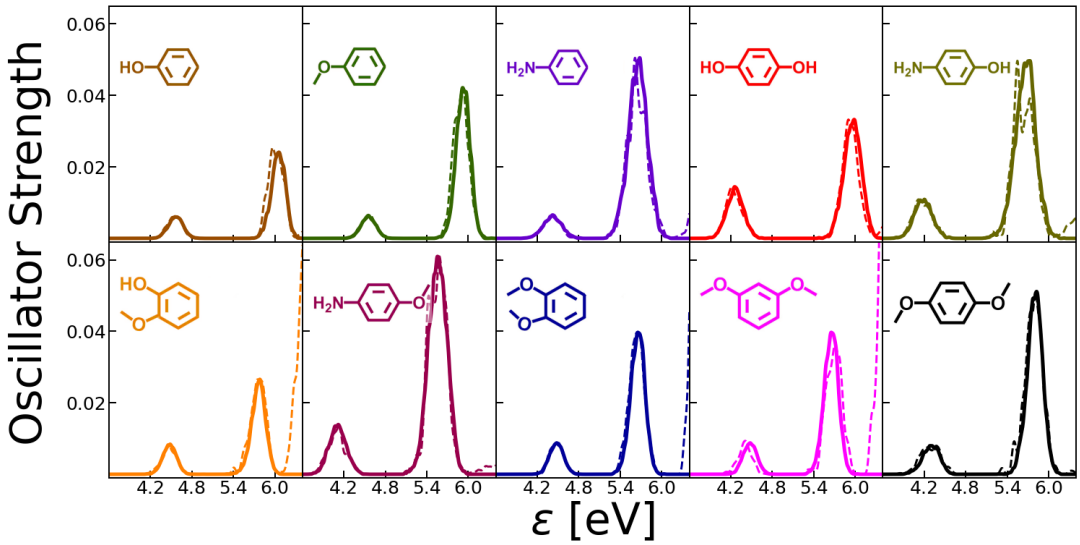



Figure (8) UV-Visible absorption spectra for all 10 aromatic molecules, including the first two excited states. Thick lines correspond to the ML spectra predicted using the 10-molecule model. Dotted lines represent the calculated spectra using the multiscale quantum chemical method.

# Conclusions

From the ML model performance and the corresponding UV-visible absorption spectra, we have shown that our ML approach can be applied to predict the electronic transitions for a class of solvated aromatic molecules. As a baseline, our 10-molecule model predicts the excitation energies of solvated aromatic molecules with MAE well below the intrinsic error of the underlying QC method. Our 7-molecule model proves that the atomic environment represented by BC can be generalized to interpolate excitation energies for molecules that are structurally similar to molecules in the training set. By applying our ML models over an ensemble of configurations sampled from FPMD simulations, converged line shapes of the lowest-energy absorption band can be readily attained. The linear decomposition analysis on the predicted excitation energies suggests the aromatic ring to be the key motif to modulate the electronic excitation, which can be explained by the $\pi \rightarrow \pi^*$ excited state character of these aromatic molecules.

This work outlines an efficient strategy to model light absorption spectra for solvated aromatic molecules by combining QC calculations and ML. We have shown that, thanks to its modular nature, our workflow can be extended to predict excitation energies of higher-energy states. Since QC-ML allows us to compute efficiently VEE for thousands of frames along a trajectory, we obtained more accurate spectral line shapes combining it with the cumulant expansion method.[81] Further improved results may be achieved by taking into account nuclear quantum effects in the MD trajectories, e.g., by using path-integral MD and/or a quantum thermostat.[91] It would also be possible to attain higher computational efficiency by using accurate *abinitio* potentials, e.g. neural network potentials or extensions of the MBpol model,[92–94] instead of FPMD simulations to sample the configurational space of solvated molecules. Besides the computational advantage, some of these models are more accurate than plain DFT with semilocal GGA functionals, as used in this work for FPMD. In the realm of perspective applications, as the ML model identifies a direct dependence of the excitation energies on the molecular configurations, it would be straightforward to apply this

approach to different solvation environments, as shown in the calculation of bathochromic shifts of molecules adsorbed in snow-packs.[6,77] Finally, given the accuracy and the generalizability of our ML approach, we envisage its extension to broader classes of organic molecules, with potential applications in energy materials, such as organic photovoltaic and dyes for photoelectrochemical systems.[95,96]

## Acknowledgments

We are grateful to Ted Hullar, Cort Anastasio and Michele Ceriotti for fruitful discussions on UV-Visible adsorption spectra of aromatic molecules in solution and ML models in computational chemistry. We thank Tim Zuehlsdorff for assisting us with the use of the cumulant scheme code.[2] We acknowledge support by the National Science Foundation under Grant No. 1806210. Computational resources were provided by the Extreme Science and Engineering Discovery Environment (XSEDE)[97] (project CHE190009), which is supported by the National Science Foundation, grant number ACI-1548562.

## Supporting Information Available

Supporting Information contains excitation energy calculations for the optimized frames of 10 aromatic molecules in gas phase, UV-Vis absorption spectra from multiscale quantum chemical calculations, comparison summaries of two 7-molecule models, parity plot of 7-molecule + OLS model, linear decomposition analysis for the 10-molecule model as well as Natural Transition Orbitals for all 10 molecules. The data, code and input files that supplement our study are publicly available at https://github.com/ZKC19940412/mluvspec.

[2]https://github.com/tjz21/Spectroscopy_python_code

# References


(1) Dominé, F.; Shepson, P. B. Air-Snow Interactions and Atmospheric Chemistry. *Science* **2002**, *297*, 1506–1510.

(2) Pastore, M.; Mosconi, E.; De Angelis, F.; Grätzel, M. A computational investigation of organic dyes for dye-sensitized solar cells: benchmark, strategies, and open issues. *J. Phys. Chem. C* **2010**, *114*, 7205–7212.

(3) Yu, Z.; Li, F.; Sun, L. Recent advances in dye-sensitized photoelectrochemical cells for solar hydrogen production based on molecular components. *Energy Environ. Sci.* **2015**, *8*, 760–775, Publisher: The Royal Society of Chemistry.

(4) Romero, E.; Novoderezhkin, V. I.; van Grondelle, R. Quantum design of photosynthesis for bio-inspired solar-energy conversion. *Nature* **2017**, *543*, 355–365.

(5) Bertrand, O.; Gohy, J.-F. Photo-responsive polymers: synthesis and applications. *Polym. Chem.* **2017**, *8*, 52–73.

(6) Hullar, T.; Bononi, F. C.; Chen, Z.; Magadia, D.; Palmer, O.; Tran, T.; Rocca, D.; Andreussi, O.; Donadio, D.; Anastasio, C. Photodecay of guaiacol is faster in ice, and even more rapid on ice, than in aqueous solution. *Environ. Sci.: Process. Impacts* **2020**, *22*, 1666–1677.

(7) Denish, P. R.; Fenger, J.-A.; Powers, R.; Sigurdson, G. T.; Grisanti, L.; Guggenheim, K. G.; Laporte, S.; Li, J.; Kondo, T.; Magistrato, A. et al. Discovery of a natural cyan blue: A unique food-sourced anthocyanin could replace synthetic brilliant blue. *Sci. Adv.* **2021**, *7*, eabe7871.

(8) Karuthedath, S.; Gorenflot, J.; Firdaus, Y.; Chaturvedi, N.; De Castro, C. S. P.; Harrison, G. T.; Khan, J. I.; Markina, A.; Balawi, A. H.; Peña, T. A. D. et al. Intrinsic

efficiency limits in low-bandgap non-fullerene acceptor organic solar cells. *Nat. Mater.* **2021**, *20*, 378–384.

(9) Blase, X.; Duchemin, I.; Jacquemin, D. The Bethe–Salpeter equation in chemistry: relations with TD-DFT, applications and challenges. *Chem. Soc. Rev* **2018**, *47*, 1022–1043.

(10) Dreuw, A.; Head-Gordon, M. Single-reference ab initio methods for the calculation of excited states of large molecules. *Chem. Rev.* **2005**, *105*, 4009–4037.

(11) Christiansen, O.; Koch, H.; Jørgensen, P.; Olsen, J. Excitation energies of H2O, N2 and C2 in full configuration interaction and coupled cluster theory. *Chem. Phys. Lett.* **1996**, *256*, 185–194.

(12) Veril, M.; Scemama, A.; Caffarel, M.; Lipparini, F.; Boggio-Pasqua, M.; Jacquemin, D.; Loos, P.-F. QUESTDB: A database of highly accurate excitation energies for the electronic structure community. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, e1517.

(13) Loos, P.-F.; Scemama, A.; Jacquemin, D. The quest for highly accurate excitation energies: A computational perspective. *J. Phys. Chem. Lett* **2020**, *11*, 2374–2383.

(14) Barone, V.; Bloino, J.; Monti, S.; Pedone, A.; Prampolini, G. Theoretical multilevel approach for studying the photophysical properties of organic dyes in solution. *Phys. Chem. Chem. Phys.* **2010**, *12*, 10550–10561.

(15) Malcıoğlu, O. B.; Calzolari, A.; Gebauer, R.; Varsano, D.; Baroni, S. Dielectric and Thermal Effects on the Optical Properties of Natural Dyes: A Case Study on Solvated Cyanin. *J. Am. Chem. Soc.* **2011**, *133*, 15425–15433.

(16) De Mitri, N.; Monti, S.; Prampolini, G.; Barone, V. Absorption and emission spectra of a flexible dye in solution: A computational time-dependent approach. *J. Chem. Theory Comput.* **2013**, *9*, 4507–4516.

(17) Ge, X.; Timrov, I.; Binnie, S.; Biancardi, A.; Calzolari, A.; Baroni, S. Accurate and Inexpensive Prediction of the Color Optical Properties of Anthocyanins in Solution. *J. Phys. Chem. A* **2015**, *119*, 3816–3822.

(18) Timrov, I.; Micciarelli, M.; Rosa, M.; Calzolari, A.; Baroni, S. Multimodel approach to the optical properties of molecular dyes in solution. *J. Chem. Theory Comput.* **2016**, *12*, 4423–4429.

(19) Zuehlsdorff, T. J.; Isborn, C. M. Combining the ensemble and Franck-Condon approaches for calculating spectral shapes of molecules in solution. *J. Chem. Phys.* **2018**, *148*, 024110.

(20) Zuehlsdorff, T. J.; Napoli, J. A.; Milanese, J. M.; Markland, T. E.; Isborn, C. M. Unraveling electronic absorption spectra using nuclear quantum effects: Photoactive yellow protein and green fluorescent protein chromophores in water. *J. Chem. Phys.* **2018**, *149*, 024107.

(21) Zuehlsdorff, T. J.; Montoya-Castillo, A.; Napoli, J. A.; Markland, T. E.; Isborn, C. M. Optical spectra in the condensed phase: Capturing anharmonic and vibronic features using dynamic and static approaches. *J. Chem. Phys.* **2019**, *151*, 074111.

(22) Bononi, F. C.; Chen, Z.; Rocca, D.; Andreussi, O.; Hullar, T.; Anastasio, C.; Donadio, D. Bathochromic Shift in the UV–Visible Absorption Spectra of Phenols at Ice Surfaces: Insights from First-Principles Calculations. *The Journal of Physical Chemistry A* **2020**, *124*, 9288–9298.

(23) Fehér, P. P.; Madarász, Á.; Stirling, A. Multiscale Modeling of Electronic Spectra Including Nuclear Quantum Effects. *Journal of chemical theory and computation* **2021**, *17*, 6340–6352.

(24) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*.

(25) Westermayr, J.; Gastegger, M.; Schütt, K. T.; Maurer, R. J. Perspective on integrating machine learning into computational chemistry and materials science. *J. Chem. Phys* **2021**, *154*, 230903.

(26) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.

(27) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; Von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **2015**, *143*, 084111.

(28) Pronobis, W.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning. *Eur Phys J B* **2018**, *91*, 178.

(29) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(30) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater.* **2017**, *3*, 1–13.

(31) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of atomic structure representations. *Phys. Rev. Lett.* **2020**, *125*, 166001.

(32) Townsend, J.; Micucci, C. P.; Hymel, J. H.; Maroulas, V.; Vogiatzis, K. D. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat. Commun* **2020**, *11*, 1–9.

(33) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-

Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **2021**, *121*, 9759–9815, PMID: 34310133.

(34) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(35) Abiodun, O. I.; Jantan, A.; Omolara, A. E.; Dada, K. V.; Mohamed, N. A.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* **2018**, *4*, e00938.

(36) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A neural network protocol for electronic excitations of N-methylacetamide. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 11612–11617.

(37) Murphy, K. P. *Machine learning: a probabilistic perspective*; MIT press, 2012.

(38) Xue, B.-X.; Barbatti, M.; Dral, P. O. Machine learning for absorption cross sections. *J. Phys. Chem. A* . **2020**, *124*, 7199–7210.

(39) Runge, E.; Gross, E. K. Density-functional theory for time-dependent systems. *Phys. Rev. Lett.* **1984**, *52*, 997.

(40) Westermayr, J.; Marquetand, P. Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space. *The Journal of Chemical Physics* **2020**, *153*, 154112.

(41) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc machine learning approach for excited-state dynamics. *The journal of physical chemistry letters* **2020**, *11*, 3828–3834.

(42) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(43) Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316–330.

(44) Cusentino, M. A.; Wood, M. A.; Thompson, A. P. Explicit Multi-element Extension of the Spectral Neighbor Analysis Potential for Chemically Complex Systems. *J.Phys. Chem. A* **2020**,

(45) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **2020**, *124*, 731–745.

(46) Timrov, I.; Andreussi, O.; Biancardi, A.; Marzari, N.; Baroni, S. Self-consistent continuum solvation for optical absorption of complex molecular systems in solution. *J. Chem. Phys.* **2015**, *142*, 034111.

(47) Zuehlsdorff, T. J.; Isborn, C. M. Modeling absorption spectra of molecules in solution. *Int J Quantum Chem* **2019**, *119*, e25719.

(48) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comput. Phys. Commun.* **2005**, *167*, 103–128.

(49) Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. cp2k: atomistic simulations of condensed matter systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 15–25.

(50) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(51) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(52) Feller, D. The role of databases in support of computational chemistry calculations. *Journal of computational chemistry* **1996**, *17*, 1571–1586.

(53) Goedecker, S.; Teter, M.; Hutter, J. Separable Dual-Space Gaussian Pseudopotentials. *Phys. Rev. B* **1996**, *54*, 1703–1710.

(54) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(55) Abascal, J.; Sanz, E.; García Fernández, R.; Vega, C. A potential model for the study of ices and amorphous water: TIP4P/Ice. *J. Chem. Phys.* **2005**, *122*, 234511.

(56) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(57) Berraud-Pache, R.; Neese, F.; Bistoni, G.; Izsák, R. Unveiling the photophysical properties of Boron-dipyrromethene dyes using a new accurate excited state coupled cluster method. *J. Chem. Theory Comput.* **2019**, *16*, 564–575.

(58) Pascual-Ahuir, J. L.; Silla, E. GEPOL: An improved description of molecular surfaces. I. Building the spherical surface set. *J. Comput. Chem.* **1990**, *11*, 1047–1060.

(59) Silla, E.; Tunon, I.; Pascual-Ahuir, J. L. GEPOL: An improved description of molecular surfaces II. Computing the molecular area and volume. *J. Comput. Chem.* **1991**, *12*, 1077–1088.

(60) Pascual-ahuir, J.-L.; Silla, E.; Tunon, I. GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface. *J. Comput. Chem.* **1994**, *15*, 1127–1138.

(61) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(62) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152*, 224108.

(63) Legrain, F.; Carrete, J.; van Roekeghem, A.; Curtarolo, S.; Mingo, N. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* **2017**, *29*, 6220–6227.

(64) Li, X.-G.; Hu, C.; Chen, C.; Deng, Z.; Luo, J.; Ong, S. P. Quantum-accurate spectral neighbor analysis potential models for Ni-Mo binary alloys and fcc metals. *Phys. Rev. B* **2018**, *98*, 094104.

(65) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*, e1701816.

(66) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc Natl Acad Sci USA* **2019**, *116*, 3401–3406.

(67) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(68) FitSNAP. `http://github.com/FitSNAP/FitSNAP`.

(69) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso: a Retrospective. *J. Royal Stat. Soc. B* **2011**, *73*, 273–282.

(70) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* . **2011**, *12*, 2825–2830.

(71) Dutta, A. K.; Nooijen, M.; Neese, F.; Izsák, R. Exploring the accuracy of a low scaling similarity transformed equation of motion method for vertical excitation energies. *J. Chem. Theory Comput* **2018**, *14*, 72–91.

(72) Shiobara, S.; Tajima, S.; Tobita, S. Substituent effects on ultrafast excited-state proton transfer of protonated aniline derivatives in aqueous solution. *Chem. Phys. Lett* **2003**, *380*, 673–680.

(73) Stalin, T.; Devi, R. A.; Rajendiran, N. Spectral characteristics of ortho, meta and para dihydroxy benzenes in different solvents, pH and $\beta$-cyclodextrin. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2005**, *61*, 2495–2504.

(74) Malongwe, J. K.; Nachtigallová, D.; Corrochano, P.; Klán, P. Spectroscopic Properties of Anisole at the Air–Ice Interface: A Combined Experimental–Computational Approach. *Langmuir* **2016**, *32*, 5755–5764.

(75) Corrochano, P.; Nachtigallová, D.; Klán, P. Photooxidation of aniline derivatives can be activated by freezing their aqueous solutions. *Environ. Sci. Technol* **2017**, *51*, 13763–13770.

(76) Chen, C.; Zhao, D.; Wang, B.; Ni, P.; Jiang, Y.; Zhang, C.; Yang, F.; Lu, Y.; Sun, J. Alkaline phosphatase-triggered in situ formation of silicon-containing nanoparticles for a fluorometric and colorimetric dual-channel immunoassay. *Anal. Chem.* **2020**, *92*, 4639–4646.

(77) Hullar, T.; Tran, T.; Chen, Z.; Bononi, F.; Palmer, O.; Donadio, D.; Anastasio, C. Enhanced photodegradation of dimethoxybenzene isomers in/on ice compared to in aqueous solution. *Atmospheric Chemistry and Physics Discussions* **2021**, 1–28.

(78) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(79) Zheng, D.; Yuan, X.-A.; Ma, H.; Li, X.; Wang, X.; Liu, Z.; Ma, J. Unexpected solvent effects on the UV/Vis absorption spectra of o-cresol in toluene and benzene: in contrast with non-aromatic solvents. *R. Soc. Open Sci.* **2018**, *5*, 171928.

(80) Mukamel, S. Fluorescence and absorption of large anharmonic molecules - spectroscopy without eigenstates. *J. Phys. Chem.* **1985**, *89*, 1077–1087.

(81) Chen, M. S.; Zuehlsdorff, T. J.; Morawietz, T.; Isborn, C. M.; Markland, T. E. Exploiting machine learning to efficiently predict multidimensional optical spectra in complex environments. *The Journal of Physical Chemistry Letters* **2020**, *11*, 7559–7568.

(82) Zhang, L.; Chen, M.; Wu, X.; Wang, H.; Weinan, E.; Car, R. Deep neural network for the dielectric response of insulators. *Phys. Rev. B* **2020**, *102*, 041121.

(83) Westermayr, J.; Maurer, R. J. Physically inspired deep learning of molecular excitations and photoemission spectra. *Chem. Sci.* **2021**, *12*, 10755–10764.

(84) Nigam, J.; Willatt, M. J.; Ceriotti, M. Equivariant representations for molecular Hamiltonians and N-center atomic-scale properties. *The Journal of Chemical Physics* **2022**, *156*, 014115.

(85) Puntanen, S. Methods of multivariate analysis, by alvin c. rencher, william f. christensen. *Int Stat Rev* **2013**, *81*, 328–329.

(86) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

(87) P., S. K.; Genova, A.; Pavanello, M. Cooperation and Environment Characterize the Low-Lying Optical Spectrum of Liquid Water. *J. Phys. Chem. Lett.* **2017**, *8*, 5077–5083.

(88) Martin, R. L. Natural transition orbitals. *J. Chem. Phys.* **2003**, *118*, 4775–4777.

(89) Badaeva, E.; Tretiak, S. Two photon absorption of extended substituted phenylenevinylene oligomers: A TDDFT study. *Chem. Phys. Lett* **2008**, *450*, 322–328.

(90) Lu, T.; Chen, F. Multiwfn: a multifunctional wavefunction analyzer. *J. Comput. Chem.* **2012**, *33*, 580–592.

(91) Ceriotti, M.; Parrinello, M.; Markland, T. E.; Manolopoulos, D. E. Efficient stochastic thermostatting of path integral molecular dynamics. *The Journal of Chemical Physics* **2010**, *133*, 124104.

(92) Schran, C.; Thiemann, F. L.; Rowe, P.; Müller, E. A.; Marsalek, O.; Michaelides, A. Machine learning potentials for complex aqueous systems made simple. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2110077118.

(93) Galib, M.; Limmer, D. T. Reactive uptake of N $_2$ O $_5$ by atmospheric aerosol is dominated by interfacial processes. *Science* **2021**, *371*, 921–925.

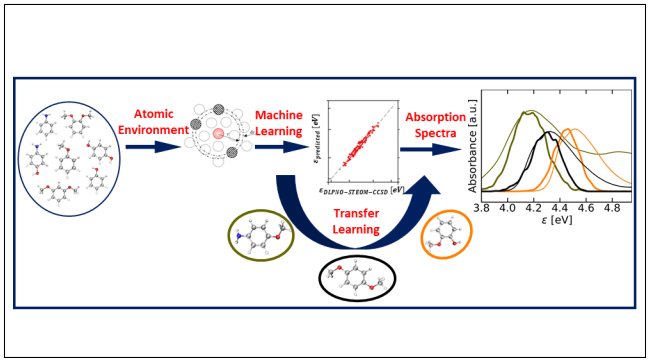
(94) Cruzeiro, V. W. D.; Lambros, E.; Riera, M.; Roy, R.; Paesani, F.; Goetz, A. W. Highly Accurate Many-Body Potentials for Simulations of $N_2$ $O_5$ in Water: Benchmarks, Development, and Validation. *J. Chem. Theory Comput.* **2021**, *17*, 3931–3945.

(95) Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S. et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *Npj Comput. Mater.* **2018**, *4*, 67.

(96) Gupta, A.; Chakraborty, S.; Ghosh, D.; Ramakrishnan, R. Data-driven modeling of S0→ S1 excitation energy in the BODIPY chemical space: High-throughput computation, quantum machine learning, and inverse design. *The Journal of Chemical Physics* **2021**, *155*, 244102.

(97) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. et al. XSEDE: Accelerating Scientific Discovery. *Comput Sci. Eng.* **2014**, *16*, 62–74.

# TOC Graphic


Atomic Environment
Machine Learning
Absorption Spectra
Transfer Learning
Absorbance [a.u.]
3.8 4.0 4.2 4.4 4.6 4.8
ε [eV]