Optimal Representation of the Nuclear Ensemble: Application to Electronic Spectroscopy

Štěpán Sršeň and Petr Slavíček*

Department of Physical Chemistry, University of Chemistry and Technology, Technická 5, 16628 Prague, Czech Republic

* *Corresponding author: petr.slavicek@vscht.cz*

Abstract

Nuclear densities are frequently represented by an ensemble of nuclear configurations or points in the phase space in various contexts of molecular simulations. The size of the ensemble directly affects the accuracy and computational cost of subsequent calculations of observable quantities. In the present work, we address the question of how many configurations do we need and how to select them most efficiently. We focus on the nuclear ensemble method in the context of electronic spectroscopy, where thousands of sampled configurations are usually needed for sufficiently converged spectra. The proposed representative sampling technique allows for a dramatic reduction of the sample size. By using an exploratory method, we model the density from a large sample in the space of transition properties. The representative subset of nuclear configurations is optimized by minimizing its Kullback-Leibler divergence to the full density with simulated

annealing. High-level calculations are then performed only for the selected subset of configurations. We tested the algorithm on electronic absorption spectra of three molecules: (*E*)-azobenzene, the simplest Criegee intermediate, and hydrated nitrate anion. Typically, dozens of nuclear configurations provided sufficiently accurate spectra. A strongly forbidden transition of the nitrate anion presented the most challenging case due to rare geometries with disproportionately high transition intensities. This problematic case was easily diagnosed within the present approach. We also discuss various exploratory methods and a possible extension to dynamical simulations.

1. Introduction

"The first computers were much smaller than Multivac. But the machines grew bigger and they could tell how the elections would go from fewer and fewer votes. Then, at last, they built Multivac and it can tell from just one voter." wrote Isaak Asimov in his short story Franchise in 1955.[1] The idea to represent public meaning by a single voter has not materialized yet, although from time to time, we witness an effort to reduce the size of representatives in the name of cost and efficiency. In this work, we adopted the same spirit into the field of computational spectroscopy.

In molecular simulations, we often replace continuous wavefunctions or densities in either coordinate or phase space with a set of nuclear configurations.[2–6] We use this approach when we calculate mean values of physical quantities, their distribution functions, or in molecular dynamics simulations of dynamical processes. Inevitably, we have to tackle the problem of how big the set should be. The optimal number of the representative geometries (samples) strongly depends on the particular problem and the desired accuracy. If we aim at high accuracy, we need a large number of samples, which hampers the simulation speed (especially if costly *ab initio* calculations are involved). If we use fewer samples, the accuracy might be insufficient. In the present work, we aim to reduce the number of samples while maintaining the accuracy of the simulations. We show our approach in detail for semiclassical modeling of electronic spectra within the so-called nuclear ensemble method (NEM).[7–9] However, an analogical approach can be applied to a large variety of problems involving discrete representations of continuous quantities.

The NEM is based on the so-called reflection principle approximation, representing the zero-order approximation to electronic spectroscopy.[10] The absorption cross section is obtained by projecting the ground state nuclear density onto the excited electronic state.[11] The spectral intensity for a given photon energy then reflects the probability of finding the molecule in a configuration with the corresponding transition energy. The quality of the calculations within this model strongly depends on (i) a proper description of the nuclear density, e.g. including nuclear quantum effects, (ii) the quality of the electronic structure method to calculate the transition quantities, and (iii) statistical convergence of the calculated quantities with the number of samples. The ground-state nuclear density can be sampled for example by various flavors of molecular dynamics (MD) simulations or by calculating the Wigner functions of (harmonic) wavefunctions.[6,7,12,13] The computational costs then reflect an interplay between statistical convergence of the simulation and the quality of the electronic structure method. At least thousands of sampled geometries are required to obtain a reasonably converged spectrum, which might be computationally infeasible for larger molecules or more accurate *ab initio* methods.[8]

We pose the following question: Can we find a small subset of geometries that would equally represent the nuclear density as a much larger ensemble? In other words, can we create an efficient "molecular parliament" representing the nuclear density? Such a problem does not have a unique solution. The data reduction can be achieved by minimizing the divergence between the densities estimated from a large set of samples and its subset. The optimization could be in principle performed in the original space of geometrical coordinates. However, we would soon encounter a problem with the curse of dimensionality: the space becomes too sparse for molecules containing more than 3 atoms, effectively prohibiting the optimization. Clearly, some sort of *a priori* dimensionality reduction is critical for this kind of optimization. Unfortunately, common dimensionality reduction methods, such as principal component analysis in the original space, are in general insensitive to coordinates affecting the properties of interest: the chromophore very often represents only a small part of the geometrical variability of the system.

Any successful algorithm has to learn about the important coordinates. In our approach (later referred to as representative sampling), we combine a fast exploratory quantum-chemical method with a more accurate method for production calculations; we optimize the density in the space of

excitation properties calculated with the exploratory method. Specifically, we calculate the excitation properties for a large set of geometries with the exploratory method, perform the optimization and recalculate the properties for the selected subset of geometries with the target method. The exploratory method should be, therefore, at least one or two orders of magnitude faster than the target *ab initio* method.

The idea is that most of the underlying physics is already covered by the cheaper method and we pay a disproportionate price for the last few percent of accuracy. In the present approach, we exploit the correlation between the two methods. Speaking in terms of errors, the representative sampling technique slightly increases the bias but significantly reduces the variance in comparison to spectra modeled from a random or equidistantly sampled subset of nuclear configurations. This way, it is possible to set the ratio between computational demands and accuracy almost arbitrarily, depending on the selected size of the subset and quantum-chemical methods used. In the present implementation, the subset is iteratively optimized via the simulated annealing[14,15] (SA) technique with the Kullback-Leibler (KL) divergence[16,17] as the objective function.

The algorithm is based on the correlation between the exploratory and the target *ab initio* method. Techniques based on the correlation between high-level and low-level methods are not rare in theoretical chemistry.[18–22] For example, higher-level correlation contribution corrections are added with smaller basis sets in the calculations of weak interactions.[23] Here, this type of idea is extended into the statistical treatment of large ensembles. A similar line of thought as we propose here was employed by Kossoski and Barbatti[24] to model temperature dependence of absorption spectra. While they utilized results for one ground-state density to approximate results for another density, we aim to reduce the density representation itself.

2. Computational methods

Electronic spectra within the nuclear ensemble method

Absorption cross-section for electric dipole transitions is expressed in the time-independent framework as:[25]

$$\sigma(E) = \frac{\pi e^2 E}{3\hbar\varepsilon_0 c} \sum_{a,b} P_a \left|\langle \psi_b | \hat{R} | \psi_a \rangle\right|^2 \delta(E - E_{ab}) \,, \tag{1}$$

where E is the photon energy, $\hat{R}$ is the position operator, ψ_a and ψ_b represent the wavefunctions of the initial vibronic state a and the final vibronic state b, respectively, E_{ab} is corresponding energy difference between the two states and P_a is the probability of finding the molecule in the state a (also, e is the elementary charge, $\hbar$ is the reduced Planck constant, ε_0 is the vacuum permittivity and c is the speed of light). However, evaluation of all the vibronic states is computationally intractable except for the smallest systems. Equation (1) can be approximated with the so-called reflection principle (RP):[2,11,25,26]

$$\sigma(E) = \frac{\pi E}{3\hbar\varepsilon_0 c} \sum_{b} \int \varrho_a(\vec{R}) \left|\vec{\mu}_{ab}(\vec{R})\right|^2 \delta\left(E - E_{ab}(\vec{R})\right) d\vec{R}, \tag{2}$$

where $\vec{\mu}_{ab}(\vec{R})$ is the transition dipole moment from the initial electronic state a to the final state b, $E_{ab}(\vec{R})$ is the corresponding excitation energy for a given geometry $\vec{R}$ and $\varrho_a(\vec{R})$ represents the nuclear density of the initial state a. RP is a semiclassical approach, which can be derived from both time-dependent and time-independent frameworks.[10,11,25] Perhaps the most illuminating is the path-integral formulation, where the RP emerges as a zero-order approximation in a semiclassical series.[10] RP imposes neither the Condon approximation nor the harmonic approximation and thus naturally covers the non-Condon effects (the dependence of the transition dipole moment on the nuclear coordinates), symmetrically forbidden transitions, and temperature effects.[24,27–29] On the other hand, it ignores quantum interference and it cannot describe vibrational progressions.[30,31]

Because it is usually difficult to obtain an accurate analytical density, it is more common to use the NEM, i.e., to represent the density with an ensemble of nuclear configurations.[2,6] For extended systems, for instance for liquids, the ground-state nuclear density can be sampled with classical MD. In these cases, the nuclear quantum effects are neglected. For smaller rigid systems, the nuclear ensemble is often generated via sampling of the ground state harmonic wavefunction or thermal density.[2,6] Path-integral-based methodologies represent an appealing alternative, which covers both thermal and nuclear quantum effects and also the anharmonicity of the system.[32]

Although path integral MD (PIMD) can be quite computationally demanding, various formulations limiting the computational costs have been suggested, for instance a quantum thermostat.[33]

The spectrum can be then modeled from the transition properties calculated for the sampled geometries:[7,18]

$$\sigma_H(E) = \frac{\pi}{3\hbar\varepsilon_0 c}\sum_b \frac{1}{nH_b\sqrt{2\pi}}\sum_{i=1}^{n} E_{ab}(\vec{R}_i)\left|\vec{\mu}_{ab}(\vec{R}_i)\right|^2 \exp\left(-\frac{1}{2}\left(\frac{E - E_{ab}(\vec{R}_i)}{H_b}\right)^2\right), \qquad (3)$$

where $E_{ab}(\vec{R}_i)$ and $\vec{\mu}_{ab}(\vec{R}_i)$ are the excitation energy and corresponding transition dipole moment between the initial electronic state a and the final state b for the i-th of n geometries. Parameter H_b is called the bandwidth and it defines the broadening of the spectrum. In other words, the spectrum is obtained as a sum of weighted Gaussian functions centered on the excitation energies of the sampled geometries with H_b being the standard deviation of these Gaussian functions. This approach corresponds to the kernel density estimation method in statistics.[8] Based on the underlying physical phenomenon, other than Gaussian functions might be used (e.g. Lorentzian functions for absorption into decaying electronic states) yet the choice has only a minor effect on the spectrum.[7]

The bandwidth H_b is usually set empirically and identically for all final states. However, it can be also derived statistically from the ensemble of nuclear configurations.[18] Here we use Silverman's rule of thumb,[34] originally derived for unimodal distributions based on normal distribution statistics, which we modified for weighted data:

$$H_b = 1.06 n_{\mathrm{eff},b}^{-1/5} s_b, \qquad (4)$$

where s_b is the corrected standard deviation obtained from the unbiased estimate of weighted sample variance[35] and $n_{\mathrm{eff},b}$ is Kish's effective sample size for state b with weights $w_{b,i}$:[36,37]

$$n_{\mathrm{eff},b} = \frac{\left(\sum_{i=1}^{n} w_{b,i}\right)^2}{\sum_{i=1}^{n} w_{b,i}^2} = \frac{\left(\sum_{i=1}^{n} E_{ab}(\vec{R}_i)\left|\vec{\mu}_{ab}(\vec{R}_i)\right|^2\right)^2}{\sum_{i=1}^{n}\left(E_{ab}(\vec{R}_i)\left|\vec{\mu}_{ab}(\vec{R}_i)\right|^2\right)^2}. \qquad (5)$$

We calculate the bandwidth for each excited state *b* separately so that the unimodality condition is fulfilled and the bias-variance ratio is optimized.

In the extreme case of representing the density in Equation (3) with only one nuclear configuration, we naturally obtain the formula for the common empirical broadening scheme, in which the density is represented only by the minimal geometry.[38] The empirical broadening scheme can serve as a fast estimate of an electronic absorption spectrum. However, it does not cover temperature and non-Condon effects and the broadening parameter becomes entirely empirical (with usual values of 0.25-0.4 eV) since it cannot be tuned statistically.[38,39] To improve the model, we employ an exploratory method to set the broadening parameter and select the most representative geometry for a given temperature. We calculate the bandwidth for the selected geometry from the standard deviation estimated for the full spectrum with the exploratory method. In the results section, we compare the common empirical broadening scheme with spectra modeled from one geometry selected by our representative sampling; for better comparison, we use the same bandwidth for the empirical broadening scheme as for our representative sampling.

Modeled spectra are accompanied by 95% confidence intervals obtained via the circular block bootstrap method to account for the sampling error.[8,40,41] In this method, we repeatedly resample with replacement nuclear configurations originally sampled by MD which we subsequently use to form a spectrum estimate. Confidence intervals are then obtained for each point of the spectrum from the distribution of these estimates.

Representative sampling

Within the representative sampling approach, we try to find a subset of structures having the same density as the full sample. More technically, we select the most representative geometries by minimizing the divergence between the density estimated from a large number of geometries and the density estimated from a subset of a desired size. As described in the introduction, we perform the optimization in the space of excitation properties calculated with a cheap exploratory method. In the simplest implementation, we could optimize the density as a function of the excitation energy $\varrho_a(E_{ab})$ only. However, it is possible to optimize the spectrum instead of the density to

emphasize spectrally significant parts of the coordinate space: a spectrum within the NEM approach is simply a weighted nuclear density. Such an approach is sufficient for the selection of a single representative geometry. However, restricting the optimization into a single dimension is in general limiting since we discard part of the information from exploratory calculations. We can instead optimize the density in the two-dimensional space of excitation energies and transition probabilities.

In the present implementation, we optimize a 2D analogue of the spectrum with an additional coordinate of transition probability: as in the 1D case, we weigh the samples by their spectral intensities to emphasize spectrally significant parts of the distribution. The distributions are then estimated from the samples by the multivariate kernel density estimation method in a similar fashion as the electronic spectra themselves:

$$\sigma_{\mathbf{H}}(\vec{x}) = \frac{\pi}{3\hbar\varepsilon_0 c}\sum_b \frac{|\mathbf{H}_b|^{-1/2}}{2\pi n}\sum_{i=1}^{n} E_{ab}(\vec{R}_i)\left|\vec{\mu}_{ab}(\vec{R}_i)\right|^2 \exp\left(-\frac{1}{2}(\vec{x}-\vec{x}_{ab,i})^{\mathrm{T}}\mathbf{H}_b^{-1}(\vec{x}-\vec{x}_{ab,i})\right), \quad (6)$$

where $\vec{x}$ is a point in the 2D space of excitation properties and $\vec{x}_{ab,i}$ is a vector of excitation properties between states *a* and *b* for the *i*-th geometry:

$$\vec{x}_{ab,i} = \left(E_{ab}(\vec{R}_i), \left|\vec{\mu}_{ab}(\vec{R}_i)\right|^2\right). \quad (7)$$

The bandwidth is now a 2×2 square matrix which takes the following form:[34]

$$\mathbf{H}_b = n_{\mathrm{eff},b}^{-1/3}\mathbf{S}_b, \quad (8)$$

where $\mathbf{S}_b$ is the weighted sample covariance matrix. Analogously to spectral evaluation, we use the sample covariance estimated for the unreduced density when selecting only one geometry.

To perform the optimization, we need to define the objective function first. The simplest option is to use the ordinary mean square error. However, as we work with densities, it is natural to use some divergence or distance metric comparing two probability density functions (PDFs) or their corresponding cumulative distribution functions. There is not a single optimal option and we can choose from several metrics based on our preferences. It is possible to use metrics based on two-

sample test statistics such as Kolmogorov-Smirnov, Kuiper, Cramér-von Mises, or Anderson-Darling statistics.[42] Alternatively, one can use information-theoretic metrics such as Kullback-Leibler (KL) or Jensen-Shanon divergence.[16,43] The third option is to employ the Wasserstein distance also known as earth mover's distance which is defined as the minimal cost of turning one PDF into the other.[44,45]

We tested several metrics and they provide comparable results; here we use the KL divergence. Compared to other divergences, it emphasizes distribution tails which are often important in spectroscopy. It is also conveniently asymmetric; the KL divergence can be interpreted as the information content lost when approximating one PDF with another.[17] Minimizing the KL divergence is also equivalent to maximum likelihood estimation if we define the problem as fitting of a gaussian mixture model.[46] The KL divergence is usually defined for normalized PDFs but we use here a generalized form suitable for unnormalized data to account for the weighting:[16,47]

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} \left(p(x) \ln \frac{p(x)}{q(x)} + q(x) - p(x) \right) dx, \tag{9}$$

where P is the target PDF and Q is an approximate PDF. We normalize both distributions by the norm of P for the sake of comparison among different systems. Note that the distribution Q might still not sum to unity.

Optimization procedure

Our goal is to find a subset of geometries that provides the minimal KL divergence. However, it is computationally intractable to test all possible combinations of geometries except for the smallest subset sizes. Instead, we can easily perform the optimization by some local or random search algorithm. We can use for example the hill-climbing technique, i.e., gradually replace geometries in the (initially random) subset while accepting only those changes that improve the solution. Alternatively, it is possible to apply more appropriate but also more complex methods of global optimization as we do here.

Here, we use the simulated annealing algorithm which is a metaheuristic devised by Kirkpatrick and co-workers[15] in 1983 to approximate global optimization. The algorithm is stochastic and the obtained minimum is not guaranteed to be the global minimum, yet it is supposed to be sufficiently close. We first generate an initial solution, i.e., an initial random subset of geometries. We then iteratively test a randomly selected neighboring solution which we define by replacing exactly one geometry in the subset. The neighboring solution is always accepted if its KL divergence is smaller. Contrary to the hill-climbing method, even worse solutions can be accepted with some probability. This probability should initially approach 100% and slowly decrease (cool down) until it reaches zero at the end of the simulation. In this way, we limit the chance of getting stuck in local minima. We use naturally inspired Boltzmann distribution to define the probability of accepting a worse solution:[15]

$$P(\Delta f) = \exp\left(-\frac{\Delta D_{\mathrm{KL}}}{T}\right), \tag{10}$$

where ΔD_{KL} is a difference in the KL divergence between the two solutions and T is a virtual temperature.

We also need to define the cooling scheme, i.e., the initial temperature, its evolution, and the terminating condition. As indicated above, the initial temperature should allow an arbitrarily wrong solution to be accepted and the probability of the acceptance should reach zero for the final temperature. However, if the temperature range is too wide, the program spends too much time in the initial or final phase not converging. We avoid it by performing a short simulation tracking the biggest and the smallest change in the KL divergence as the first step. We then plug each of these values together with a desired acceptance probability at the beginning or at the end of the simulation into Equation (10) to calculate the initial and the final temperature, respectively.[48] We set here the initial probability to 0.9 and the final probability to 0.1 and we decrease the temperature geometrically in time as proposed by Kirkpatrick.[14]

We perform the whole optimization multiple times in parallel with a smaller number of iterations as it is more efficient than one very long optimization due to the stochasticity of the algorithm.[49] Overall, the only input parameters that influence the quality of the solution and the optimization

time are the numbers of iterations and parallel jobs. Other parameters are either predefined or set automatically via a simulation. We benchmarked the SA algorithm against the exhaustive search for the selection of one or two representative geometries, where it is computationally feasible, and obtained the globally optimal solutions.

Computational details

We used the density functional theory (DFT) and the density functional based tight-binding (DFTB) for the ground-state calculations. Zerner's intermediate neglect of differential overlap for spectroscopy (ZIndo/S),[50] the third-order algebraic diagrammatic construction (ADC(3))[51], and the time-dependent DFT (TDDFT) methods were employed for excited-state calculations. ZIndo/S, DFT, and TDDFT calculations were performed in the Gaussian 09 package,[52] revision D.01. DFTB calculations were employed as implemented in the DFTB+ code,[53,54] release 18.2. ADC(3) calculations were performed in the Q-Chem 4.3 code.[55]

For dynamical simulations, we used the PIMD in combination with a quantum thermostat based on colored-noise generalized Langevin equation (GLE) to efficiently incorporate nuclear quantum effects within the so-called PI+GLE method.[33] All MD simulations were performed using our in-house program ABIN[56] interfaced with various *ab initio* codes. The parameters for the GLE thermostat were obtained from the online library.[57]

3. Results and Discussion

We used several molecular systems as a testbed for our approach. The presented spectra are always shown on the absolute scale and they are not scaled on the energy or intensity scale. The 95% confidence intervals and corresponding experimental spectra accompany the simulated data for a better comparison of the accuracy. It has been previously shown that up to tens of thousands of geometries are needed to obtain fully converged spectra.[8] However, typically only hundreds of nuclear configurations are used in production simulations.[2,58] In the present study, we model the electronic absorption spectra with 1000 geometries (the full, unreduced size).

Azobenzene: multichromophoric system

We first tested the approach on the UV/Vis absorption spectrum of the (*E*)-azobenzene molecule in methanol which represents a medium-sized system with several absorption bands and excited states including a symmetrically forbidden transition to the first excited state. Azobenzene is an archetypal molecular photoswitch that isomerizes upon exposure to the UV light of certain wavelengths.[59] Nuclear configurations were sampled by the PI+GLE MD simulation on the semiempirical DFTB potential and the transition properties were subsequently calculated at the CAM-B3LYP/6-31+g* level: a combination of methods that we previously verified for this system.[8] The solvent effects were included via an implicit solvation model.

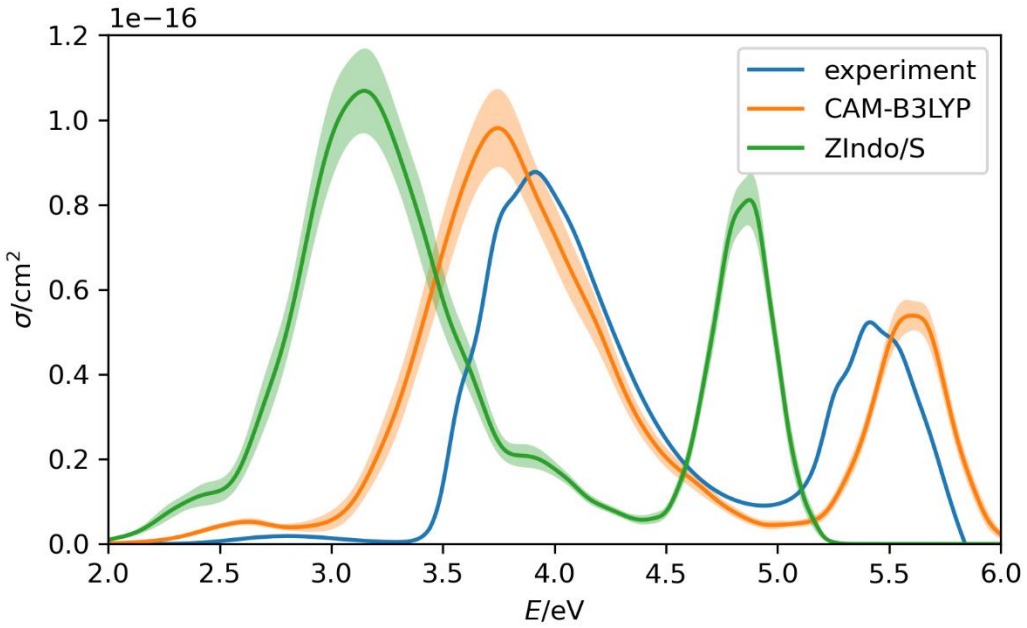

Figure 1: Comparison of simulated and exprimental[8] UV/Vis absorption spectra of the (E)-azobenzene molecule in methanol. Simulated spectra are modeled with 1000 geometries sampled with PI+GLE MD on the DFTB potential. Transitions to the first 7 excited states are considered at different levels of theory. The simulated spectra are accompanied by 95% confidence intervals accounting for the sampling error.

We use here the ZIndo/S method for exploratory calculations; semiempirical methods represent the natural choice since they are several orders of magnitude faster than any *ab initio* or TDDFT method. The comparison of spectra modeled from 1000 geometries with both the exploratory and the target method is captured in Fig. 1. Such a comparison would not be available in production runs but we present it here to provide further insight into the approach. Although the ZIndo/S

spectrum significantly deviates from both the experiment and the spectrum calculated with the target method, the correlation between the ZIndo/S and the CAM-B3LYP methods is very strong.

We optimized representative subsets of nuclear configurations for several subset sizes on the ZIndo/S level and used these geometries to recalculate spectra at the CAM-B3LYP level. For the illustration of the optimization procedure, the density estimated from 30 geometries at the end of the optimization is compared to the density modeled from all the samples in Fig. 2.

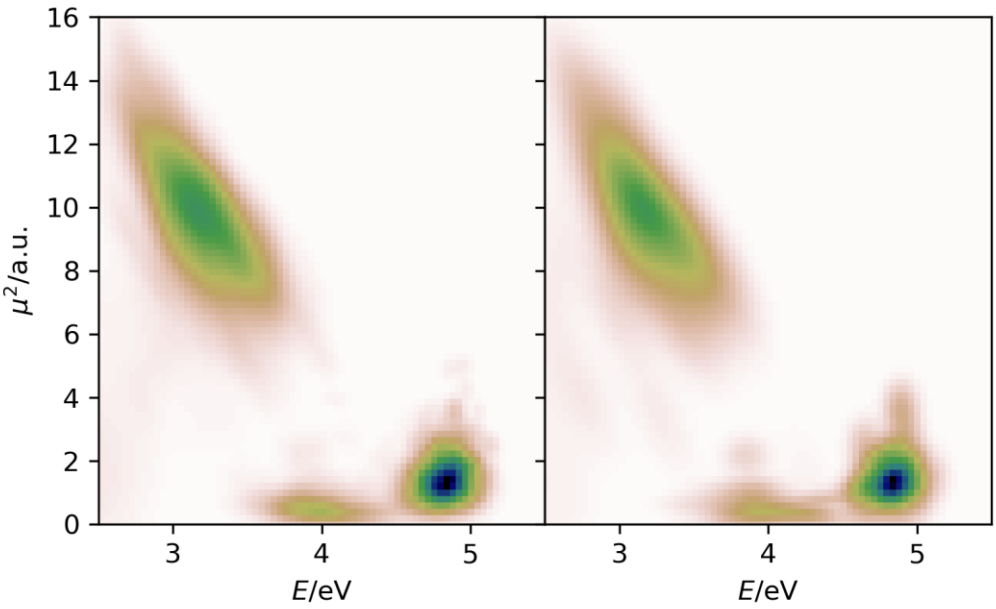

Figure 2: Weighted ground-state density for (E)-azobenzene in the space of excitation properties calculated at the exploratory ZIndo/S level for (a) all the samples and (b) optimized subset of 30 geometries.

The final values of the KL divergences (see Table 1) measure how well we approximated the density. The KL divergence for identical distributions would vanish. Values approaching 0.01-0.02 seem to be sufficient (as shown below). The KL divergence for one geometry is smaller than for 3 nuclear configurations; the reason is that we do not estimate the sample covariance in the case of a single geometry but we take it from the full density. The optimization is then simpler but we use larger subsets to encode the covariance for a better description of the density. We can compare the minimal structure and the most representative structure selected by the representative

sampling technique (see Fig. 3); while the minimal geometry is planar, the selected representative geometry is already twisted, allowing for symmetry-forbidden transitions.

Table 1: Optimized KL divergences for different subset sizes for the (E)-azobenzene molecule in methanol.

N	1	3	5	10	30	50
D_{KL}	0.146	0.208	0.111	0.066	0.026	0.016

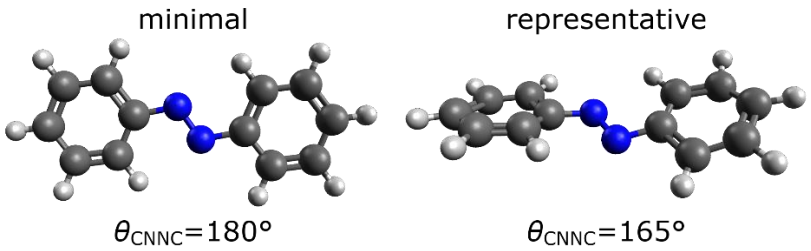

Figure 3: The minimal nuclear configuration (left) and the most representative nuclear configuration selected by the proposed scheme (right). Values of the C-N=N-C dihedral angle are presented below respective structures.

The resulting absorption spectra modeled with the target CAM-B3LYP method are compared for different subset sizes in Fig. 4. Spectra modeled only from dozens of geometries selected by the representative sampling approach are in very good agreement with the full spectrum. Small deviations for the third absorption band can be partially attributed to a different order of excited states between the exploratory and the target method for some geometries. These minor deviations could be further reduced by including more excited states while optimizing the spectrum only up to a fixed energy/wavelength. The spectrum is sufficiently converged already with 30 geometries as it coincides with the full spectrum within the confidence intervals. For comparison, we also plot in Fig. 4 the spectrum modeled from 30 geometries selected equidistantly from the MD. Our approach clearly surpasses this naive reduction. Reduced spectra modeled with less than ten geometries deviate significantly from the full spectrum. However, the spectrum from one representative geometry surpasses the empirical broadening scheme: not only is the overall agreement with the full spectrum better but it also contains the first band which is symmetry-forbidden in the minimal geometry and thus completely missing in the empirical broadening scheme.

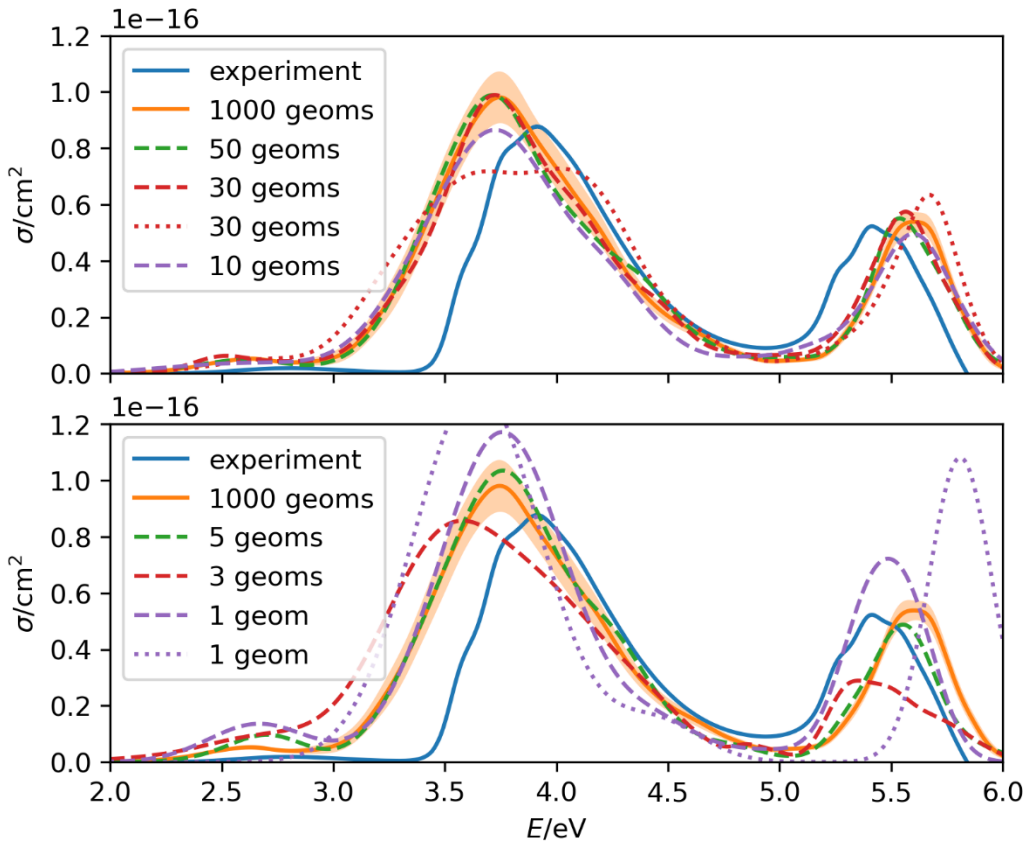

Figure 4: Comparison of the (E)-azobenzene UV/Vis absorption spectrum modeled with 1000 geometries with an experiment[8] (solid lines), absorption spectra modeled from subsets of geometries selected via the representative sampling scheme (dashed lines), equidistantly sampled geometries, and empirical broadening scheme (dotted lines). Transitions to the first 7 excited states were calculated at the CAM-B3LYP/6-31+g level with the reduction performed at the ZIndo/S level. The full spectrum is accompanied by 95% confidence intervals accounting for the sampling error.*

Criegee intermediate: representative sampling with a reduced basis set

As a second test case, we selected the simplest Criegee intermediate CH_2OO. It is a very small system with only one bright state, but it requires high-level *ab initio* treatment because of its problematic electronic structure.[18] Criegee intermediates play a key role in tropospheric chemistry; they serve as a non-photolytic source of OH radicals and contribute to the removal of volatile organic compounds from the atmosphere.[60] Based on our previous work,[18] we sampled the nuclear configurations by PI+GLE MD on the PBEPBE/aug-cc-pVDZ potential energy surface and we calculated the transition properties at the ADC(3)/aug-cc-pVDZ level.

In contrast to the azobenzene case, we used here a different strategy for exploratory calculations. Instead of changing the electronic structure method, we only reduced the basis set. Such an approach is especially advantageous for high-level electronic structure methods as they scale unfavorably with the basis set size. We used here a small 6-31g basis set without polarization or diffuse functions; it consists of only 31 basis functions for the studied molecule compared to 87 basis functions in the target aug-cc-pVDZ basis set. The ADC(3) method scales formally as $O(n^6)$ with the basis set size,[61] and the exploratory calculations were almost by two orders of magnitude faster than the production calculations. It is also possible to further extend the approach by using several exploratory basis sets and reducing the number of geometries gradually: a methodology similar to the virtual screening of molecules where a smaller number of candidates is selected in each iteration by a more accurate method.[62]

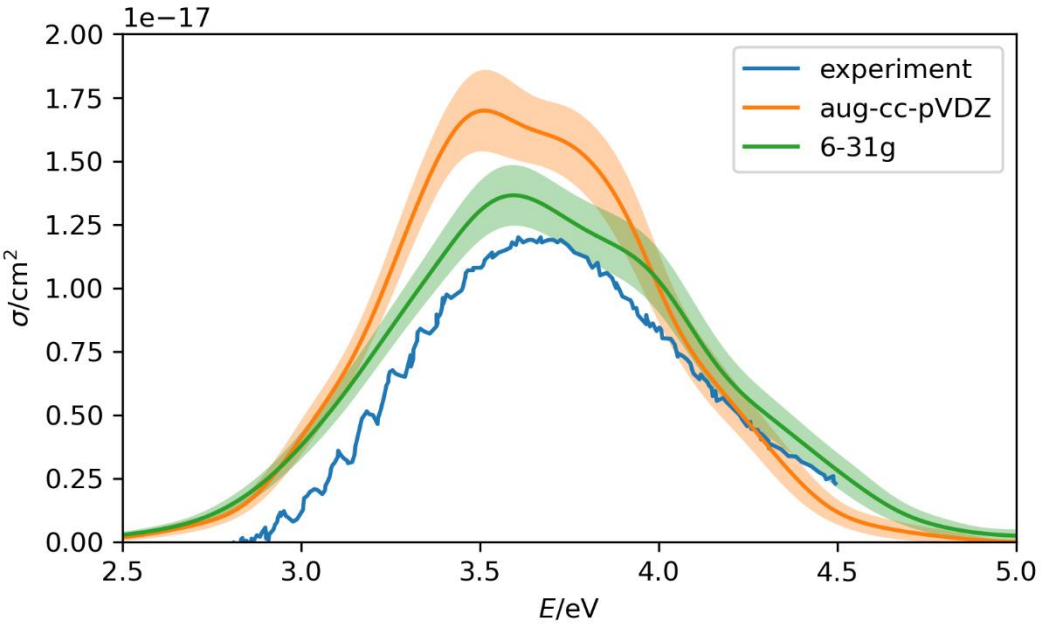

Figure 5: Comparison of simulated and experimental[63] UV/Vis absorption spectra for the simplest Criegee intermediate CH_2OO. Simulated spectra are modeled with 1000 geometries sampled with PI+GLE MD on the PBEPBE/aug-cc-pVDZ potential. Only the transition to the second excited state is considered. The simulated spectra are accompanied by 95% confidence intervals accounting for the sampling error.

The comparison of spectra modeled with both the exploratory and the target basis set is shown in Fig. 5. Both basis sets provide similar spectra yet with statistically significant differences. The

spectrum modeled with the exploratory basis set is broader and shifted to higher energies. The values of KL divergence for optimized subsets (see Table 2) are in general very low indicating an easily reproducible density. The KL divergence is approaching 0.02 already for 10 nuclear configurations.

Table 2: Optimized KL divergences for different subset sizes for the CH_2OO molecule.

n	1	3	5	10	30	50
D_{KL}	0.085	0.052	0.042	0.023	0.009	0.006

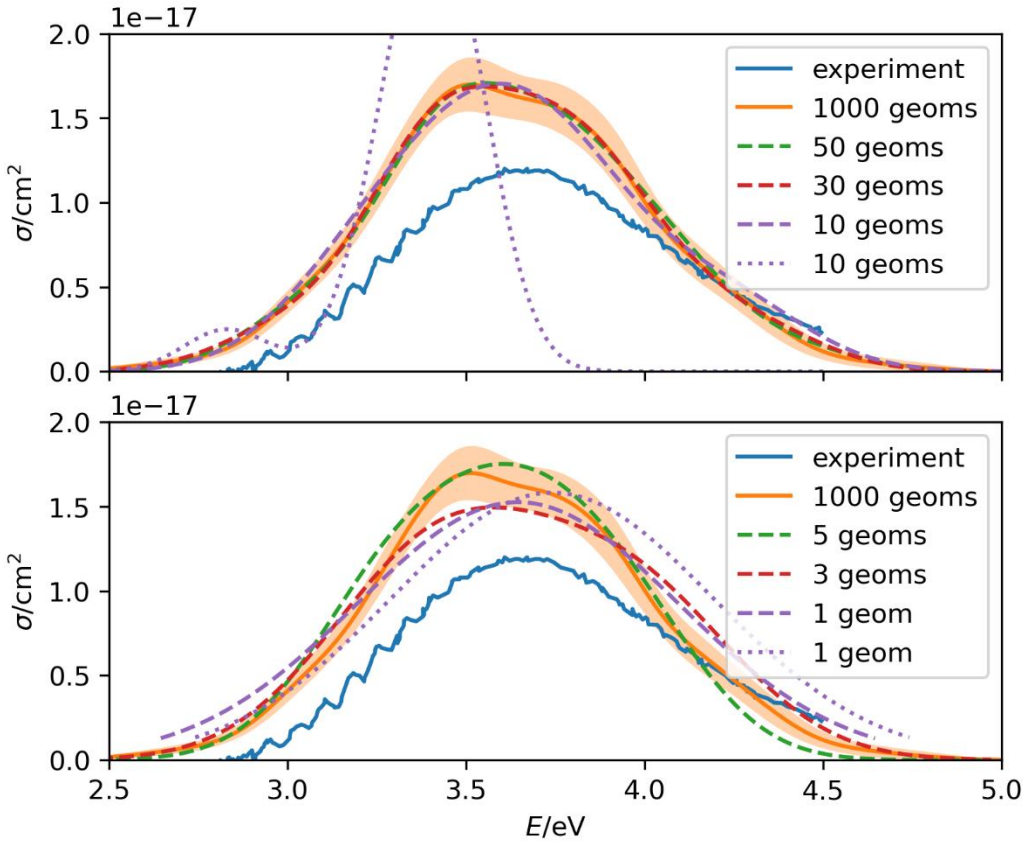

Figure 6: Comparison of the CH_2OO UV/Vis absorption spectrum modeled with 1000 geometries with an experiment[63] (solid lines), spectra modeled from subsets of geometries selected via the representative sampling scheme (dashed lines), equidistantly sampled geometries, and empirical broadening scheme (dotted lines). Only the transition to the second excited state is considered. Transition properties were calculated at the ADC(3)/aug-cc-pVDZ level with the reduction performed on the ADC(3)/6-31g level. The full spectrum is accompanied by 95% confidence intervals accounting for the sampling error.

Simulated results for reduced ensemble sizes are compared in Fig. 6. In agreement with the KL divergence values, the spectrum modeled from as little as 10 geometries is in very good agreement with the full spectrum. By contrast, the spectrum modeled from 10 equidistantly sampled geometries completely fails to reproduce the target spectrum. Smaller reduced subsets deviate a little but the accuracy is still acceptable. Our selection of one representative geometry again outperforms the empirical broadening scheme as it provides a better maximum position.

Nitrate anion: the case of a symmetry forbidden excitation

Strongly forbidden absorption into the first excited state of the nitrate anion serves as a complex third test case. The transition dipole moment for the excitation into the first excited state has a zero value in the optimal structure; the first and second derivatives of the transition dipole moments have zero values as well. A coupling between at least two vibrational modes is needed to observe the transition for an isolated nitrate anion.[27] Yet this weakly absorbing dark state is responsible for a significant generation of OH radicals in aqueous systems,[64] explaining e.g. extraordinary oxidation capacity of the polar troposphere.[65] The absorption is enhanced by several orders of magnitude by solvation which breaks the symmetry of the nitrate electronic wavefunction. We employed a hybrid solvation model: we included explicitly ten water molecules and used the polarizable continuum model to account for bulk solvent effects. The explicit solvation is essential for a proper description of the absorption.[27]

We sampled the nuclear configurations by PI+GLE MD on the DFTB potential energy surface with the DFT-D3[66,67] dispersion correction and we calculated the transition properties at the previously proposed[27] CAM-B3LYP/aug-cc-pVDZ level. We used a smaller basis set as the exploratory method again, specifically the 3-21g basis set. Spectra modeled from 1000 geometries with both basis sets are compared in Fig. 7. We were able to recover the integral intensity from the experiment even though the transition is strongly forbidden. However, theoretical spectra are shifted to lower energies. The spectrum modeled with the 3-21g basis set is significantly broader, more intense and shifted to lower energies in comparison with the target aug-cc-pVDZ basis set.

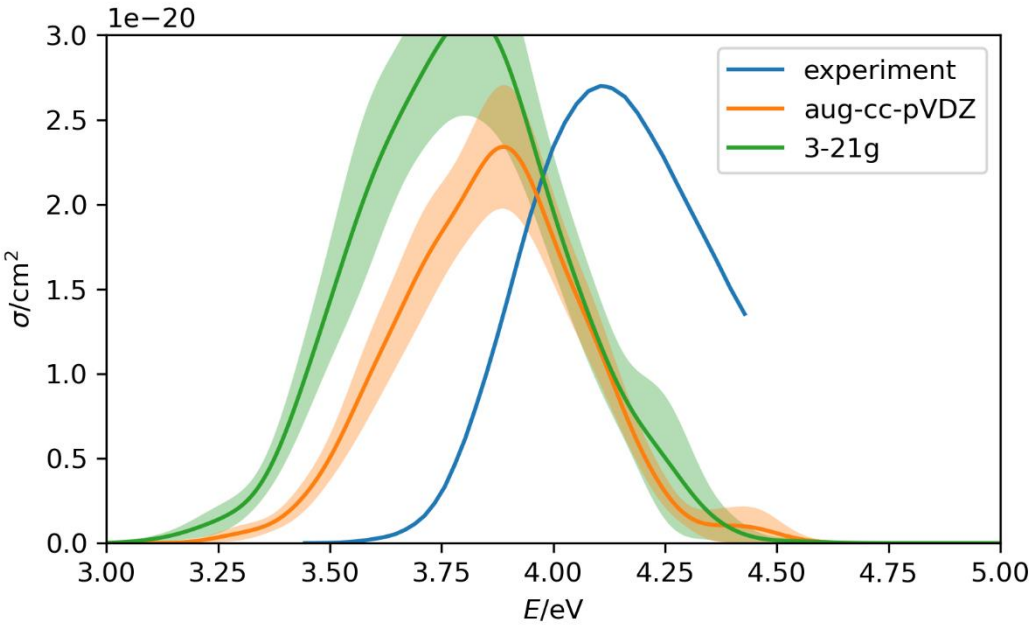

Figure 7: Comparison of simulated and experimental[68] UV/Vis absorption spectra for the first absorption band of hydrated nitrate anion. Simulated spectra are modeled with 1000 geometries sampled with PI+GLE MD on the DFTB+D3 potential. Only the transition to the first excited state is considered. The simulated spectra are accompanied by 95% confidence intervals accounting for the sampling error.

Table 3: Optimized KL divergences for different subset sizes for the first absorption band of the nitrate anion.

N	1	3	5	10	30	50
D_{KL}	0.503	0.710	0.637	0.671	0.454	0.376

The final KL divergence values (see Table 3) obtained from the optimization are very large for all sample sizes in comparison with previous cases, which signalizes that the reduction does not work properly in this case. We plot the sampled points recalculated at the exploratory level together with the weighted density in Fig. 8, both for all the samples and for 30 representative geometries. We can see that a large fraction of the intensity is caused by only a few isolated geometries with non-proportionally strong transition probabilities; most of the geometries have transition probabilities close to zero. Such density cannot be easily approximated by a small number of samples. The character of the excitation process is reflected also in the relatively large error bars of the simulated spectra.

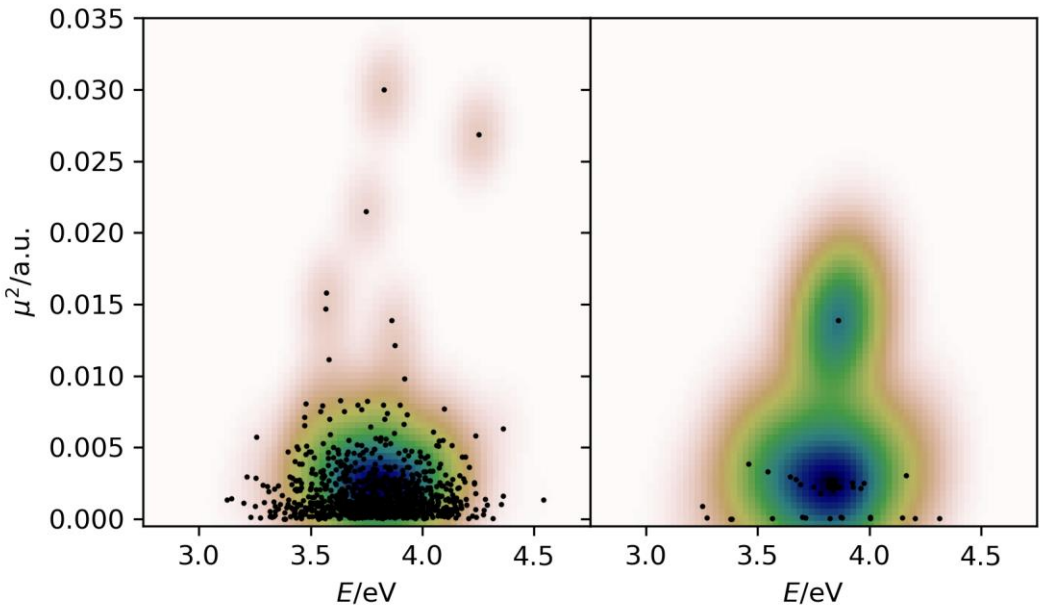

Figure 8: Weighted ground-state density for hydrated nitrate anion in the space of excitation properties calculated at the exploratory CAM-B3LYP/3-21g level for (a) all the samples and (b) optimized subset of 30 geometries.

The comparison of spectra obtained with the representative sampling procedure (using the 3-21g basis set as the exploratory approach) is shown in Fig. 9. As already indicated by the KL divergences and densities plots, the agreement is suboptimal yet still acceptable considering the complexity of this case, computational savings, and larger uncertainty of the full spectrum. Spectra modeled with 30 and 50 geometries can be found mostly within confidence intervals. The representative sampling approach also outperforms the equidistant sampling as we show on spectra modeled from 50 geometries. Spectrum modeled from just one selected geometry is slightly more intense and a little bit shifted to lower energies than the full spectrum. In this case, the empirical broadening scheme cannot be applied in a simple fashion; the peak is not present for the isolated anion and there is a huge number of local minima on the potential energy surface when including water molecules explicitly. We found the limit of the present approach. However, we were able to identify the problem already with the exploratory method and the results are still acceptable.

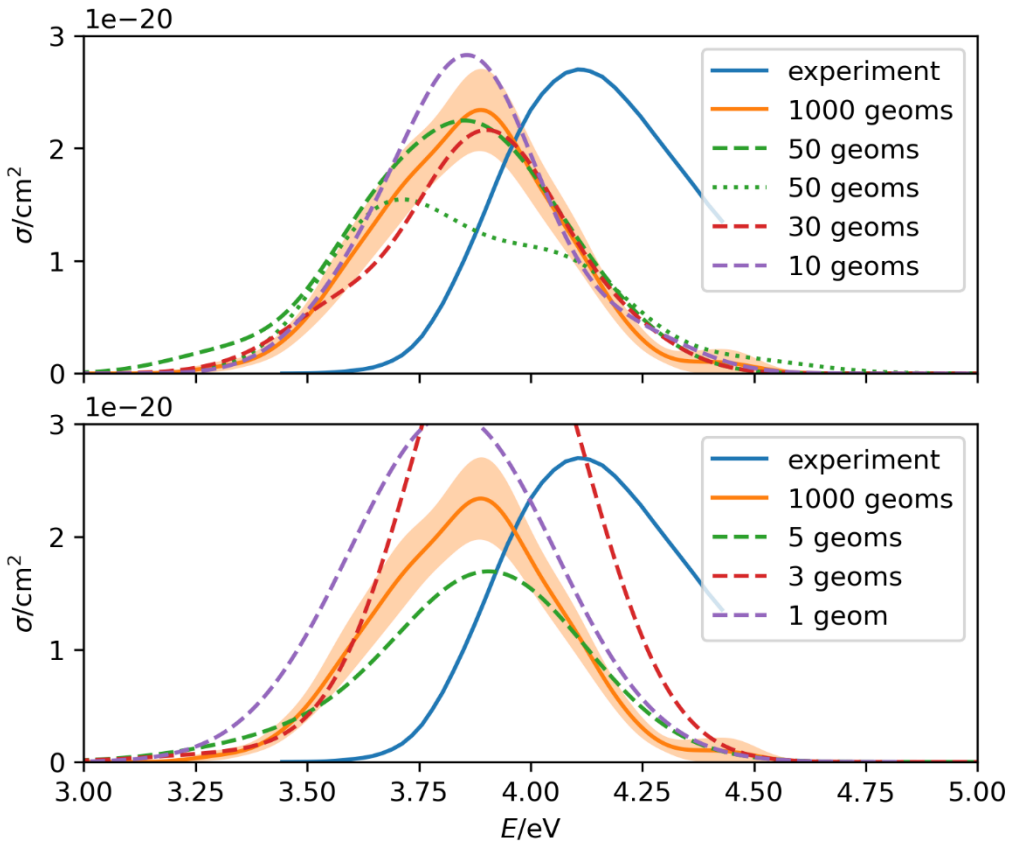

Figure 9: Comparison of the UV/Vis absorption spectrum modeled with 1000 geometries with an experiment[68] (solid lines), spectra modeled from subsets of geometries selected via the proposed scheme (dashed lines), equidistantly sampled geometries, and empirical broadening scheme (dotted lines) for the hydrated nitrate anion. Only the transition to the first excited state is considered. Transition properties were calculated at the CAM-B3LYP/aug-cc-pVDZ level with the reduction performed on the CAM-B3LYP/3-21g level. The full spectrum is accompanied by 95% confidence intervals accounting for the sampling error.

4. Conclusions

We proposed and tested a scheme to reduce the number of geometries representing nuclear density in the NEM for simulations of the electronic spectra, such as UV absorption or photoelectron spectra. We tested the proposed sampling scheme for several cases, including larger multi-chromophoric systems (azobenzene), molecules with problematic electronic structure (Criegee intermediate), symmetry-forbidden transition, and solvated molecules (nitrate anion). The error introduced by the reduction scheme is usually much smaller than the discrepancy between theoretical and experimental spectra caused by the electronic structure itself. Simulated reduced

spectra also usually coincide with full spectra within confidence intervals when we sample the system with as little as dozens of points. However, the quality of this approach is given to large extent by the selected exploratory method.

Several possible exploratory methods were tested and discussed. One can use a completely different electronic structure method with semiempirical methods being especially efficient. We successfully tested the ZIndo/S method. We also tested another scheme for the exploratory method: the target electronic structure method but with a reduced basis set. This approach proved to be especially efficient for high-level *ab initio* methods due to their unfavorable scaling with the number of basis functions. However, many other options could be possibly used. For example, the simplified Tamm-Dancoff approach (sTDA)[69] or simplified TDDFT (sTD-DFT)[70] semiempirical methods can utilize already calculated ground-state wavefunctions when using DFT potential for the ground-state MD.

One might ask whether there is some deeper physical wisdom hidden in the procedure of reducing the ensemble size. The algorithm is stochastic and the geometries produced are thus not unique. However, the inspection of a small number of geometries provides information about important geometrical features contributing to the electronic transitions. When we look for a single most representative sample, as in Asimov's short story, we can examine the character of the excitation by looking at the difference between the minimal and the most representative geometry.

Another important piece of information from the simulation process is the minimum size we can reach. Dozens of geometries seem to be sufficient to reproduce the full spectrum within confidence intervals. However, it was demonstrated on the example of the nitrate anion that for symmetry forbidden transitions we need a large number of samples. This is unfortunate from the perspective of computational efficiency yet it does not invalidate the algorithm. We diagnosed the problem based on the KL divergences before performing the full set of calculations at the higher level. Furthermore, a difficult reduction is indicative of the character of electronic transitions – with rare but important events playing a major role.

The present technique can be easily extended to other types of processes. For example, we usually represent an initial wavefunction or density matrix by a swarm of points in the phase space, giving

rise to classical trajectories in the mixed quantum-classical simulations.[3–5] Such simulations are computationally demanding and a justified reduction of the number of trajectories would be rather helpful. Setting a suitable objective function would allow for example combining different electronic structure methods (e.g. CASSCF/CASPT2 methods for excited-state calculations). Alternatively, the ensemble used for the simulation of the electronic spectrum of the system can be used as a reasonable starting point for subsequent dynamical simulations as well.

Acknowledgments

The support of Czech Science Foundation project No. 20-15825S is gratefully acknowledged. Š. S. is a student of the International Max Planck Research School for "Many-Particle Systems in Structured Environments". This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140).

References

(1) Asimov, I. Franchise. *If: Worlds of Science Fiction*. August 1955, pp 2–15.

(2) Crespo-Otero, R.; Barbatti, M. Spectrum Simulation and Decomposition with Nuclear Ensemble: Formal Derivation and Application to Benzene, Furan and 2-Phenylfuran. *Theor. Chem. Acc.* **2012**, *131*, 1237.

(3) Suchan, J.; Janoš, J.; Slavíček, P. Pragmatic Approach to Photodynamics: Mixed Landau–Zener Surface Hopping with Intersystem Crossing. *J. Chem. Theory Comput.* **2020**, *16*, 5809–5820.

(4) Curchod, B. F. E.; Martínez, T. J. Ab Initio Nonadiabatic Quantum Molecular Dynamics. *Chem. Rev.* **2018**, *118*, 3305–3336.

(5) Barbatti, M.; Sen, K. Effects of Different Initial Condition Samplings on Photodynamics and Spectrum of Pyrrole. *Int. J. Quantum Chem.* **2016**, *116*, 762–771.

(6) Ončák, M.; Šištík, L.; Slavíček, P. Can Theory Quantitatively Model Stratospheric Photolysis? Ab Initio Estimate of Absolute Absorption Cross Sections of ClOOCl. *J. Chem. Phys.* **2010**, *133*, 174303.

(7) Barbatti, M.; Aquino, A. J. A.; Lischka, H. The UV Absorption of Nucleobases: Semi-Classical Ab Initio Spectra Simulations. *Phys. Chem. Chem. Phys.* **2010**, *12*, 4959.

(8) Sršeň, Š.; Sita, J.; Slavíček, P.; Ladányi, V.; Heger, D. Limits of the Nuclear Ensemble Method for Electronic Spectra Simulations: Temperature Dependence of the (E)-Azobenzene Spectrum. *J. Chem. Theory Comput.* **2020**, *16*, 6428–6438.

(9) Bergsma, J. P.; Berens, P. H.; Wilson, K. R.; Fredkin, D. R.; Heller, E. J. Electronic Spectra from Molecular Dynamics: A Simple Approach. *J. Phys. Chem.* **1984**, *88*, 612–619.

(10) Vaníček, J.; Cohen, D. Path Integral Approach to the Quantum Fidelity Amplitude. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150164.

(11) Lee, S. Y.; Brown, R. C.; Heller, E. J. Multidimensional Reflection Approximation: Application to the Photodissociation of Polyatomics. *J. Phys. Chem.* **1983**, *87*, 2045–2053.

(12) Lukeš, V.; Šolc, R.; Barbatti, M.; Lischka, H.; Kauffmann, H. F. Torsional Potentials and Full-Dimensional Simulation of Electronic Absorption Spectra of Para-Phenylenevinylene Oligomers Using Semiempirical Hamiltonians. *J. Theor. Comput. Chem.* **2010**, *9*, 249–263.

(13) Wigner, E. On the Quantum Correction For Thermodynamic Equilibrium. *Phys. Rev.* **1932**, *40*, 749.

(14) Kirkpatrick, S. Optimization by Simulated Annealing: Quantitative Studies. *J. Stat. Phys.* **1984**, *34*, 975–986.

(15) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(16) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

(17) Burnham, K. P.; Anderson, D. R. *Model Selection and Multimodel Inference*; Springer, 2002.

(18) Sršeň, Š.; Hollas, D.; Slavíček, P. UV Absorption of Criegee Intermediates: Quantitative Cross Sections from High-Level Ab Initio Theory. *Phys. Chem. Chem. Phys.* **2018**, *20*, 6421–6430.

(19) DeYonker, N. J.; Cundari, T. R.; Wilson, A. K. The Correlation Consistent Composite Approach (CcCA): An Alternative to the Gaussian-n Methods. *J. Chem. Phys.* **2006**, *124*, 114104.

(20) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. Gaussian-1 Theory: A General Procedure for Prediction of Molecular Energies. *J. Chem. Phys.* **1998**, *90*, 5622.

(21) Zaspel, P.; Huang, B.; Harbrecht, H.; Von Lilienfeld, O. A. Boosting Quantum Machine Learning Models with a Multilevel Combination Technique: Pople Diagrams Revisited. *J. Chem. Theory Comput.* **2019**, *15*, 1546–1559.

(22) Karton, A. A Computational Chemist's Guide to Accurate Thermochemistry for Organic Molecules. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6*, 292–310.

(23) Alessandrini, S.; Barone, V.; Puzzarini, C. Extension of the "Cheap" Composite Approach to Noncovalent Interactions: The Jun-ChS Scheme. *J. Chem. Theory Comput.* **2019**, *16*, 988–1006.

(24) Kossoski, F.; Barbatti, M. Nuclear Ensemble Approach with Importance Sampling. *J. Chem. Theory Comput.* **2018**, *14*, 3173–3183.

(25) Schinke, R. *Photodissociation Dynamics*; Cambridge Monographs on Atomic, Molecular, and Chemical Physics; Cambridge University Press, 1993; Vol. I.

(26) Della Sala, F.; Rousseau, R.; Görling, A.; Marx, D. Quantum and Thermal Fluctuation Effects on the Photoabsorption Spectra of Clusters. *Phys. Rev. Lett.* **2004**, *92*, 183401.

(27) Svoboda, O.; Kubelová, L.; Slavíček, P. Enabling Forbidden Processes: Quantum and Solvation Enhancement of Nitrate Anion UV Absorption. *J. Phys. Chem. A* **2013**, *117*, 12868–12877.

(28) 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.

(29) Kojić, M.; Lyskov, I.; Milovanović, B.; Marian, C. M.; Etinski, M. The UVA Response of Enolic Dibenzoylmethane: Beyond the Static Approach. *Photochem. Photobiol. Sci.* **2019**, *18*, 1324–1332.

(30) Lee, S. Semiclassical Theory of Radiation Interacting with a Molecule. *J. Chem. Phys.* **1982**, *76*, 3064–3074.

(31) Mukamel, S. *Principles of Nonlinear Optical Spectroscopy*; Oxford series in optical and imaging

sciences; Oxford University Press, 1999.

(32) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J.; Klein, M. L. Efficient Molecular Dynamics and Hybrid Monte Carlo Algorithms for Path Integrals. *J. Chem. Phys.* **1993**, *99*, 2796–2808.

(33) Ceriotti, M.; Manolopoulos, D. E.; Parrinello, M. Accelerating the Convergence of Path Integral Dynamics with a Generalized Langevin Equation. *J. Chem. Phys.* **2011**, *134*, 84104.

(34) Silverman, B. W. *Density Estimation for Statistics and Data Analysis*; Monographs on Statistics & Applied Probability; Chapman & Hall, 1986.

(35) Narsky, I.; Porter, F. C. *Statistical Analysis Techniques in Particle Physics*; Wiley-VCH, 2013.

(36) Kish, L. *Survey Sampling*; John Wiley & Sons, 1965.

(37) Kish, L. Weighting for Unequal P_I. *J. Off. Stat.* **1992**, *8*, 183–200.

(38) Grimme, S. Calculation of the Electronic Spectra of Large Molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Larter, R., Cundari, T. R., Eds.; John Wiley & Sons, 2004; pp 153–218.

(39) Reyes-Gutiérrez, P. E.; Jirásek, M.; Severa, L.; Novotná, P.; Koval, D.; Sázelová, P.; Vávra, J.; Meyer, A.; Císařová, I.; Šaman, D.; Pohl, R.; Štěpánek, P.; Slavíček, P.; Coe, B. J.; Hájek, M.; Kašička, V.; Urbanová, M.; Teplý, F. Functional Helquats: Helical Cationic Dyes with Marked, Switchable Chiroptical Properties in the Visible Region. *Chem. Commun.* **2015**, *51*, 1583–1586.

(40) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26.

(41) Lahiri, S. N. *Resampling Methods for Dependent Data*; Springer Series in Statistics; Springer, 2003.

(42) Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Stat. Assoc.* **1974**, *69*, 730–737.

(43) Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.

(44) Dobrushin, R. L. Prescribing a System of Random Variables by Conditional Distributions. *Theory Probab. Its Appl.* **1970**, *15*, 458–486.

(45) Rubner, Y.; Tomasi, C.; Guibas, L. J. Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121.

(46) Broniatowski, M. Minimum Divergence Estimators, Maximum Likelihood and the Generalized Bootstrap. *Entropy* **2021**, *23*, 185.

(47) Zhu, H.; Rohwer, R. Bayesian Invariant Measurements of Generalization. *Neural Process. Lett.* **1995**, *2*, 28–31.

(48) Sanvicente-Sánchez, H.; Frausto-Solís, J. A Method to Establish the Cooling Scheme in Simulated

Annealing like Algorithms. *Lect. Notes Comput. Sci.* **2004**, *3045*, 755–763.

(49) Lee, D. Y.; Wexler, A. S. Simulated Annealing Implementation with Shorter Markov Chain Length to Reduce Computational Burden and Its Application to the Analysis of Pulmonary Airway Architecture. *Comput. Biol. Med.* **2011**, *41*, 707–715.

(50) Ridley, J.; Zerner, M. An Intermediate Neglect of Differential Overlap Technique for Spectroscopy: Pyrrole and the Azines. *Theor. Chim. Acta* **1973**, *32*, 111–134.

(51) Trofimov, A. B.; Stelter, G.; Schirmer, J. A Consistent Third-Order Propagator Method for Electronic Excitation. *J. Chem. Phys.* **1999**, *111*, 9982–9999.

(52) 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.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; 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.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision D.01. Gaussian, Inc. 2013.

(53) Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(54) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(55) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; Ghosh, D.; Goldey, M.; Horn, P. R.; Jacobson, L. D.; Kaliman, I.; Khaliullin, R. Z.; Kuś, T.; Landau, A.; Liu, J.; Proynov, E. I.; Rhee, Y. M.; Richard, R. M.; Rohrdanz, M. A.; Steele, R. P.; Sundstrom, E. J.; Woodcock, H. L.; Zimmerman, P. M.; Zuev, D.; Albrecht, B.; Alguire, E.; Austin, B.; Beran, G. J. O.; Bernard, Y. A.; Berquist, E.; Brandhorst, K.; Bravaya, K. B.; Brown, S. T.; Casanova, D.; Chang, C.-M.; Chen, Y.; Chien, S. H.; Closser, K. D.; Crittenden, D. L.; Diedenhofen, M.; DiStasio, R. A.; Do, H.; Dutoi, A. D.; Edgar, R. G.; Fatehi, S.; Fusti-Molnar, L.; Ghysels, A.; Golubeva-Zadorozhnaya, A.; Gomes, J.; Hanson-Heine, M. W. D.; Harbach, P. H. P.; Hauser, A. W.; Hohenstein, E. G.; Holden, Z. C.; Jagau, T.-C.; Ji, H.; Kaduk, B.; Khistyaev, K.; Kim, J.; Kim, J.; King, R. A.; Klunzinger, P.; Kosenkov, D.; Kowalczyk, T.; Krauter, C. M.; Lao, K. U.; Laurent, A. D.; Lawler, K. V.; Levchenko, S. V.; Lin, C. Y.; Liu, F.; Livshits, E.; Lochan, R. C.; Luenser, A.; Manohar, P.; Manzer, S. F.; Mao, S.-P.; Mardirossian, N.; Marenich, A. V.; Maurer, S. A.; Mayhall, N. J.; Neuscamman, E.; Oana, C. M.; Olivares-Amaya, R.; O'Neill, D. P.; Parkhill, J. A.; Perrine, T. M.; Peverati, R.; Prociuk, A.; Rehn, D. R.; Rosta, E.; Russ, N. J.; Sharada, S. M.; Sharma, S.; Small, D. W.; Sodt, A.; Stein, T.; Stück, D.; Su, Y.-C.; Thom, A. J. W.; Tsuchimochi, T.; Vanovschi, V.; Vogt, L.; Vydrov, O.; Wang, T.; Watson, M. A.; Wenzel, J.; White, A.; Williams, C. F.; Yang, J.; Yeganeh, S.; Yost, S. R.; You, Z.-Q.; Zhang, I. Y.; Zhang, X.; Zhao, Y.; Brooks, B. R.; Chan, G. K. L.; Chipman, D. M.; Cramer, C. J.; Goddard, W. A.; Gordon, M. S.;

Hehre, W. J.; Klamt, A.; Schaefer, H. F.; Schmidt, M. W.; Sherrill, C. D.; Truhlar, D. G.; Warshel, A.; Xu, X.; Aspuru-Guzik, A.; Baer, R.; Bell, A. T.; Besley, N. A.; Chai, J.-D.; Dreuw, A.; Dunietz, B. D.; Furlani, T. R.; Gwaltney, S. R.; Hsu, C.-P.; Jung, Y.; Kong, J.; Lambrecht, D. S.; Liang, W.; Ochsenfeld, C.; Rassolov, V. A.; Slipchenko, L. V.; Subotnik, J. E.; Van Voorhis, T.; Herbert, J. M.; Krylov, A. I.; Gill, P. M. W.; Head-Gordon, M. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(56) Hollas, D.; Suchan, J.; Oncak, M.; Slavicek, P. PHOTOX/ABIN: Pre-Release of Version 1.1. **2018**. https://doi.org/10.5281/zenodo.1228463.

(57) GLE4MD Project: Automatic input generation http://gle4md.org/index.html?page=matrix (accessed 2018 -01 -20).

(58) Svoboda, O.; Ončák, M.; Slavíček, P. Simulations of Light Induced Processes in Water Based on *Ab Initio* Path Integrals Molecular Dynamics. I. Photoabsorption. *J. Chem. Phys.* **2011**, *135*, 154301.

(59) Hartley, G. S. The Cis-Form of Azobenzene. *Nature* **1937**, *140*, 281.

(60) Osborn, D. L.; Taatjes, C. A. The Physical Chemistry of Criegee Intermediates in the Gas Phase. *Int. Rev. Phys. Chem.* **2015**, *34*, 309–360.

(61) Dreuw, A.; Wormit, M. The Algebraic Diagrammatic Construction Scheme for the Polarization Propagator for the Calculation of Excited States. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2015**, *5*, 82–95.

(62) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45*, 195–216.

(63) Ting, W. L.; Chen, Y. H.; Chao, W.; Smith, M. C.; Lin, J. J. M. The UV Absorption Spectrum of the Simplest Criegee Intermediate CH2OO. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10438–10443.

(64) Daniels, M.; Meyers, R. V.; Belardo, E. V. Photochemistry of the Aqueous Nitrate System. I. Excitation in the 300-Mμ Band. *J. Phys. Chem.* **1968**, *72*, 389–399.

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

(66) 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.

(67) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(68) Chu, L.; Anastasio, C. Quantum Yields of Hydroxyl Radical and Nitrogen Dioxide from the Photolysis of Nitrate on Ice. *J. Phys. Chem. A* **2003**, *107*, 9594–9602.

(69) Grimme, S. A Simplified Tamm-Dancoff Density Functional Approach for the Electronic Excitation

Spectra of Very Large Molecules. *J. Chem. Phys.* **2013**, *138*, 244104.

(70) Bannwarth, C.; Grimme, S. A Simplified Time-Dependent Density Functional Theory Approach for Electronic Ultraviolet and Circular Dichroism Spectra of Very Large Molecules. *Comput. Theor. Chem.* **2014**, *1040–1041*, 45–53.