
Efficient prediction of photoabsorption cross-sections for volatile organic compounds using the nuclear ensemble approach with representative sampling

Author:

KONSTANTIN NOMEROTSKI

Supervisor:

BASILE CURCHOD

Second assessor:

LAURA RATCLIFFE



School of Chemistry

UNIVERSITY OF BRISTOL

A thesis submitted to the University of Bristol in accordance
with the requirements of the degree of MASTER OF SCIENCE
in Chemistry with Scientific Computing

APRIL 2024

Word count: 21,459

Abstract

The *in silico* calculation of photoabsorption cross-sections using the nuclear ensemble approach has gained recent popularity due to its robust and versatile applicability, producing accurate approximations for a range of volatile organic compounds. It provides a method for the description of post Condon effects allowing for the investigation of non-adiabatic phenomena, an exciting prospect for photochemists.

However, the nuclear ensemble approach is limited by the requirement of a large ensemble, often thousands of nuclear geometries in size. The need for thousands of *ab initio* calculations can become computationally prohibitive for more complex molecules, resulting in an unviable workload. Representative sampling has been proposed as a method for the reduction of the ensemble size, while minimizing the information lost, thereby producing an ensemble most "representative" of the original.

A prototype implementation of representative sampling with the nuclear ensemble approach has been created, for the automatic calculation of photoabsorption cross-sections from a given ensemble. The program implements the steps of representative sampling, automatically submitting quantum chemistry calculations, extracting the outputs, optimizing the sample and outputting the spectrum. A fast and cheap semi-empirical method, ZIndo/S, was used for a quick calculation of the excitation energies and transition dipole moments. The ensemble was optimized using 2-dimensional, weighted kernel density estimation over the transition properties $x = (\Delta E, |\mu|^2)$, weighted by spectral intensity. The Kullback-Leibler divergence (D_{KL}) was minimized between the original and subset densities using simulated annealing, producing a reduced ensemble. The prototype has been successfully tested by application to two volatile organic compounds, acrolein and toluene, with an appreciable speed-up in the calculation of *ab initio* cross-sections for the $S_0 \rightarrow S_1$ transition.

Considering acrolein, the ensemble size could be reduced by 99% from ensemble sizes previously used to just 10 geometries. This translated to an 84% decrease in overall walltime, even when increasing ensembles by an order of magnitude, there is a 48% decrease relative to the original nuclear ensemble approach. The improvement in efficiency is more pronounced for toluene, a larger and more complex molecule, due to the increased system size, and the representative algorithm's scaling with ensemble size, rather than nuclear coordinates. The divergence values were found to be higher for toluene, it required a larger ensemble of 50 geometries to account for the intensity of the transition, where more geometries are needed to account for the symmetry-forbidden transition. Where the larger ensemble is concerned, the speed-up is still appreciable, suggesting an increasing viability for larger molecule sizes using the nuclear ensemble approach.

The creation of an automated submission scheme for the nuclear ensemble approach with representative sampling provides a single workflow for the submission of more efficient calculations. The prototype presented acts as an initial step towards the integration into Atmospec, providing implementation insights. The implementation into Atmospec has been discussed and a proposed method for integration has been presented. Implementation into a web-application such as Atmospec would provide Atmospheric chemists with a more efficient alternative for the rapid estimation of photoabsorption cross-sections, increasing the applicability of the NEA to more complex molecules by reducing the computational burden of such calculations.

Dedication and acknowledgements

I would like to extend my heartfelt gratitude to my supervisor, Basile Curchod, who always provided a welcoming and lighthearted environment throughout our collaboration. His expertise and in depth understanding was invaluable, and his commitment to my success unwavering, it has been an incredibly rewarding experience working alongside him this year. I would also like to thank the second assessor, Laura Ratcliffe, for providing insightful and practical feedback, moreover, her role as the presentation chair is greatly appreciated.

A special thanks goes to Daniel Hollas for his help with algorithmic understanding and providing expert knowledge in technical aspects of this thesis. I would also like to highlight his contribution to Atmospec and the tools used herein, without which this research would not have been possible. Additionally, I am deeply grateful to Stepan Srsen, for his indispensable contribution, through his design of the representative sampling algorithm, moreover, his dedication to the accompanying documentation has been invaluable.

I am also indebted to the Centers for Computational Chemistry, and Advanced Computing Research at the University of Bristol, for providing me with the necessary resources and environment to complete my research, as well as their general support through my time at the university. Similarly, I'd like to thank my personal tutor, David Fermin, for his continued support in my academic development and his valuable advice in times of need.

I would like to express my gratitude to my family for their strong support and encouragement throughout my studies. My brother has been a great source of inspiration and motivation in the past year, which combined with his optimistic and pragmatic character has been key in my success. I would also like to thank my mother for her unfaltering belief and pride, serving as a constant drive for me.

Finally, an honorable mention goes to my cats, Barsik and Grisha, who had no idea what was going on, but supported me nonetheless

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University’s Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate’s own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.


SIGNED:  DATE: 16/04/2024

Table of Contents

	Page
1 Introduction	8
1.1 VOCs in the Atmosphere	8
1.2 Modelling atmospheric composition	8
2 Methods	9
2.1 Light induced processes	9
2.1.1 Einstein Interpretation of Radiative processes	9
2.1.2 Absorbance properties	11
2.1.3 Molecular absorption	12
2.1.4 The Born-Oppenheimer approximation	13
2.1.5 The Franck-Condon Principle	14
2.1.6 Breakdown of the Born-Oppenheimer approximation	15
2.2 Photodissociation	16
2.2.1 Photolysis rates	17
2.2.2 Photoabsorption cross-sections	17
2.2.3 Reflection principle	18
2.3 Calculating photoabsorption cross-sections	19
2.3.1 Time-dependent and independent methods	19
2.3.2 Nuclear Ensemble Approach	20
2.3.3 Statistical representation of the NEA	22
2.4 Atmospec	23
2.4.1 Current functionality	23
2.5 Developments in the nuclear ensemble approach	25
2.5.1 Importance sampling	25
2.5.2 Statistical determination of the broadening parameter	26
2.5.3 Machine learning approaches	27
2.5.4 Machine learning for quantum chemistry calculations	27
2.5.5 Probabilistic machine learning for spectral reconstruction	28
2.6 Representative sampling	28
2.6.1 Divergence	29
2.6.2 Sample optimization	30

2.6.3	Simulated annealing	30
2.6.4	Bandwidth selection	31
2.7	Data availability	32
3	Aims of this thesis	33
4	Computational Details	33
4.1	Data collection and processing	33
4.2	Representative sampling and QC recalculation	33
4.3	Spectral calculations	34
5	Results and Discussion	34
5.1	Implementation overview	34
5.2	Acrolein	35
5.3	Toluene	37
5.4	Equidistant sampling	38
5.5	Evaluating efficiency	39
5.6	Proposed integration with Atmospec workflow	41
6	Conclusion	43
7	Outlook	44
7.1	Algorithmic improvements	44
7.2	Combinatory approaches	45

List of Figures

FIGURE	Page
2.1 Jabłoński diagram showing the fundamental molecular absorption and relaxation processes, adapted from G. Insero, F. Fusi and G. Romano <i>"The safe use of lasers in biomedicine: Principles of laser-matter interaction"</i> . ²	12
2.2 Illustration of Franck-Condon principle, adapted from H. Maguire, J. Iles-Smith, and A. Nazir <i>"Environmental Nonadditivity and Franck-Condon physics in Nonequilibrium Quantum Systems"</i> ³	15
2.3 Potential energy surface diagram for photo-excitations and subsequent relaxations. adapted from M. E. Casida, B. Natarajan, and T. Deutsch <i>"Real-time dynamics and conical intersections"</i> . ⁴	16
2.4 The reflection principle, Schinke <i>"Photodissociation Dynamics"</i> . ⁵	18
2.5 Visualisation of photoabsorption cross sections of a transition from $S_0 \rightarrow S_1$ modelled by a) time-dependent, and b) time-independant methods, Prlj. et al. <i>"Calculating Photoabsorption Cross-Sections for Atmospheric Volatile Organic Compounds"</i> . ⁶	20
2.6 A visualization of the nuclear ensemble approach (NEA) for the calculation of photoabsorption cross-sections. Prlj. et al. <i>"Calculating Photoabsorption Cross-Sections for Atmospheric Volatile Organic Compounds"</i> . ⁶	21
2.7 Atmospec spectrum visualisation for 2-hydroxypropanal using EOM-CCSD with the aug-pVDZ basis set. Generated in <i>Atmospec v0.2.3</i> . ⁷	24
2.8 Photolysis rate constant calculation, rate coefficient visualization, and transition properties widgets for 2-hydroxypropanal using EOM-CCSD with the aug-cc-pVDZ basis set. Generated in <i>Atmospec v2.0.3</i> . ⁷	25
4.1 Flowchart visualising the representative sampling workflow alongside the current workflow . .	33
5.1 Flowchart pf the workflows implemented, NEA (<i>orange</i>), and NEA with Representative sampling (<i>green</i>). Ensemble sizes and data outputs are annotated.	35
5.2 Representative sampling workflow for acrolein using ZIndo/S (<i>exploratory</i>), B3LYP/6-311*G (<i>recalculation</i>) compared to single point calculation at the minimum energy geometry (<i>purple</i>), and the full sample (<i>blue</i>) calculated using B3LYP/6-311*G. Experimental values obtained from the MPI/Mainz UV-vis spectral atlas ^{8,9}	36
5.3 Representative sampling workflow for toluene using ZIndo/S (<i>exploratory</i>), B3LYP/6-311*G (<i>recalculation</i>) compared to single point calculation at the minimum energy geometry (<i>purple</i>), and the full sample (<i>blue</i>) calculated using B3LYP/6-311*G. Experimental values obtained from the MPI/Mainz UV-vis database. ^{8,10}	38
5.4 Weighted ground state densities of acrolein in the space of transition properties calculated using ZIndo/S for the full sample (<i>left</i>), subsets of 10 geometries sampled using representative and equidistant sampling, (<i>middle</i>) and (<i>right</i>) respectively	39
5.5 Weighted ground state densities of toluene in the space of transition properties calculated using ZIndo/S for the full sample (<i>left</i>), subsets of 10 geometries sampled using representative and equidistant sampling, (<i>middle</i>) and (<i>right</i>) respectively	40

5.6	Summary of representative sampling functionality alongside current Atmospec workchain . .	41
5.7	Proposed widget implementation for a submission over a range of ensemble sizes	42
7.1	General framework for the genetic algorithm meta-heuristic	44

List of Tables

TABLE	Page
2.1	Current input requirements for Atmospec workflow 23
5.1	Initial conditions for the representative sampling workflow 35
5.2	Minimum Kullback-Leibler divergence values for acrolein, obtained using 1500 cycles of simulated annealing and 28 parallel repetitions over the ground state density with weighted intensities 37
5.3	Minimum Kullback-Leibler divergence values for toluene, obtained using 1500 cycles of simulated annealing and 28 parallel repetitions over the ground state density with weighted intensities 38
5.4	A comparison of wall times of individual steps in the representative sampling workflow, varying the number of samples taken 40

1 Introduction

1.1 VOCs in the Atmosphere

Volatile organic compounds (VOCs) are compounds with a vapor pressure greater than 10 Pa at 25°C and a boiling point of up to 260° in atmospheric pressure. While they are found at relatively low concentrations in the atmosphere, their effects on both regional and global environmental stability, paired with their negative impact on human health make them critical candidates for study in atmospheric chemistry. Both biogenic and anthropogenic sources contribute to the total concentration of VOCs in the atmosphere, with the most prominent anthropogenic contributor being the exploitation of natural resources such as oil and gas, producing toxic VOCs such as benzene, toluene and isoprene.¹¹ Moreover, the VOCs emitted will rapidly undergo changes in the atmosphere due to their extreme volatility, forming secondary VOCs which are often more reactive and have extremely short lifetimes

The adverse effects on human health of volatile organic compounds are profound, with many being classified as carcinogenic or mutagenic.¹² Prolonged exposure is commonly associated with an increased risk of respiratory complications such as asthma, or more seriously, cancers of the brain, nervous system, endocrine system and skin. With the appearance of such health conditions increasing at an alarming rate,^{13,14} it is vital to monitor the air we breathe to ensure that we are controlling our exposure to harmful chemicals. Increased VOC concentration in urban areas has also been linked to prolonged, decreased regional air quality, such pollution in concentrated areas severely impacts the local health.¹⁵

In addition to affecting human health, high concentrations of VOCs can have profound effects on both local and global atmospheric conditions, with the most prominent being smog. Acting as the culmination of multiple pollutants, with tropospheric ozone as a major component, smog is produced in photochemical reactions of volatile organic compounds and nitrogen oxides in the troposphere. In addition to worsening air pollution, photochemical oxidants like ozone can damage materials such as rubber and textiles. VOCs also influence the global atmospheric temperature directly by absorbing infrared radiation from the earth's atmosphere. Although this is lessened by their short lifetimes, their photoproducts contribute to changing ozone concentrations, a notorious greenhouse gas.¹⁶

The contribution of volatile organic compounds to global warming is a double edge sword, in that the emissions of biogenic VOCs (bVOCs) are proportional to temperature, therefore increases in global temperature will be reflected in the increased emission of bVOCs.¹⁷ These factors make it important to track global emissions of VOCs, model the atmosphere's composition to predict future conditions and inform preventative actions. The emission of VOCs has been heavily regulated in many countries with extremely positive results, for example, the Clean Air Act 1970 (CAA) in the United States, and air quality standards set by the EU in Europe.¹⁸ Such legislation can have a pivotal effect on emissions when implemented and is vital to maintaining sustainable growth. emphasizing the need for informed decision-making. Additionally, there has been a significant research effort to compose reaction mechanisms for atmospheric organic compounds.¹⁹ With the advance of computational capabilities, computational atmospheric modelling has become a go-to tool for the modelling atmospheric composition.

1.2 Modelling atmospheric composition

To fully encapsulate the dynamic atmospheric environment, chemical reactions of airborne compounds must be considered. The Master Chemical Mechanism is a near-explicit model of the reaction mechanisms in the

troposphere used widely in research.²⁰ Acting as the golden standard in atmospheric modelling, its applications include investigating the impact of anthropogenic emissions and biogenic emissions on the formation of secondary organic aerosols.³ Moreover, it gives lawmakers the power to model the effects of proposed changes, thereby improving the impact of future legislation. Over the past decade, the MCM has seen substantial patches, the most recent being updates to include isoprene chemistry, significantly affecting the regeneration of OH radicals at lower NO_x concentrations as well as the degradation of nitrogen oxides.²¹ The expansion of such a large mechanism is an ongoing process, with future plans including the development of automatic mechanism generation and reduction methods.

The Master Chemical Mechanism is built on empirical data, with structure activity relationships (SARs) used where data is not available, which are still limited in their empirical derivation and are not universally applicable.²² The empirical roots of the MCM neglect ultra-fast processes such as photochemical reactions which can be difficult to measure due to their short lifetimes. Photochemical reactivity stands serve as an explanation for certain discrepancies observed in atmospheric modelling. For example, photochemical dissociation pathways of isoprene have been proposed to rationalize differences in observed conditions and values predicted by the MCM over the amazon rainforest, highlighting the limitation of only considering the ground state chemistry.^{23,24} The investigation of the photochemical properties of secondary VOCs is essential to fully understanding the composition and chemical processes of our atmosphere.

There has been a focus on the development of more rigorous models, which require less manual updates. Innovations such as automated mechanism generators and mechanism reduction methods bring us closer to a self sufficient atmospheric model, that can adapt to changing conditions.²⁵ However, such developments rely on the underlying model, which doesn't consider the excited states due to the lack of empirical data. The difficulty in obtaining experimental quantities such as rate constants of photochemical reactions gives a natural transition to computational simulations as a tool for the calculation of photoabsorption properties. Recent innovations in quantum chemistry make the accurate calculation of *ab initio* photoabsorption cross-sections possible, providing a favourable alternative to the use of structural activity relationships.⁶ Section 2 discusses light-induced chemistry, and the theoretical determination of photochemical properties.

2 Methods

2.1 Light induced processes

Sunlight is clearly important in global atmospheric processes, influencing chemical composition by providing alternative reaction routes for volatile organic compounds. To properly understand the changes in reactivity, we must consider the absorption process itself, as well as the subsequent relaxation pathways

2.1.1 Einstein Interpretation of Radiative processes

Introduced by Einstein in 1916, the Einstein coefficients (A and B) describe the spontaneous emission and induced absorption of light by matter.²⁶ For any given emission of photons, the energy of the photon is equal to the difference in energy between the two energy levels, due to the discrete nature of atomic energy levels given by the Bohr frequency condition. The quantum of electromagnetic radiation emitted or absorbed can be expressed using Planck's relation:

$$E_2 - E_1 = h\nu = \hbar\omega \quad (1)$$

Where E_2 and E_1 are the initial and final energies respectively, ν is the frequency corresponding to the photon energy and ω is the angular frequency (i.e. $2\pi\nu$). The absorbance of photons in molecular spectroscopy results in a vibronic excitation while the emission of light occurs from vibronic relaxation. The Einstein A coefficient is defined in terms of the rate of spontaneous radiative emission:

$$k_{21}^s = A_{21}N_2 \quad (2)$$

Where the rate of emission is inversely proportional to the population of the upper energy level. The Einstein B coefficients are defined by the rates of induced absorption and emission, independent of radiation intensity.

$$k_{21}^i = B_{21}^\omega \rho_\omega N_2 \quad (3)$$

$$k_{12}^i = B_{12}^\omega \rho_\omega N_1 \quad (4)$$

Where the rate is proportional to the product of N , the energy level population and ρ , the energy density, representing the radiation available for the transitions. Einstein showed that radiation can not only induce absorption, but also induce the emission of a photon with the corresponding photon energy. The coefficients for (induced) absorption and emission are related by the degeneracy factors of the energy levels involved. For a two-level system the relationship can be expressed as:

$$g_1 B_{12} = g_2 B_{21} \quad (5)$$

Considering both spontaneous and induced transitions, the rate of emission can be written as

$$k_{21}^{tot} = A_{21}N_2 + B_{21}\rho_\omega N_2 \quad (6)$$

Therefore, the condition for a system in thermal equilibrium, where there is no net transfer of energy, becomes

$$N_1 B_{12} \rho_\omega = N_2 (A_{21} + B_{21} \rho_\omega) \quad (7)$$

$$\rho_\omega = \frac{A_{21}}{\frac{N_1}{N_2} B_{12} - B_{21}} \quad (8)$$

Where the populations of respective energy levels are distributed according to the Boltzmann distribution. Equation 7 can be rearranged to obtain an expression for energy density ρ_ω in terms of Einstein coefficients with a substitution of the Boltzmann factor

$$\frac{N_1}{N_2} = e^{-\frac{\hbar\omega}{k_B T}} \quad (9)$$

The Planck distribution for black body radiation, similarly, follows the Boltzmann distribution.

$$\rho(\omega) = \frac{\hbar}{\pi^3 c^3} \frac{\omega^3}{e^{-\frac{\hbar\omega}{k_B T}} - 1} \quad (10)$$

Upon comparison of Equation 9 to the density of states given by the Planck distribution, Equation 10, and using the degeneracy presented in Equation 5, we yield a description of the relation between the two Einstein coefficients.

showing that the relative probability of spontaneous emission increases with the square of the transition frequency, making it a vital consideration for transitions with a high frequency such as those induced by X-rays.

$$A_{21} = \frac{\hbar\omega^3}{\pi^2 c^3} B_{21} \quad (11)$$

While quanta of light will be absorbed by matter, Einstein and Stark proposed that not all quanta absorbed will be sufficient for photochemical activation,²⁷ therefore a measure of efficiency was introduced; the quantum yield. Due to the competing nature of unproductive photochemical processes such as radiative or non-radiative transition, quantum yield is defined for a particular photochemical pathway.²⁸

$$\phi = \frac{Rate (s^{-1})}{Quanta\ absorbed (s^{-1})} \quad (12)$$

2.1.2 Absorbance properties

The absorption and emission processes identified by Einstein are transferable to a molecular picture, although complicated by the additional degrees of freedom.²⁹ It was found that the intensity of molecular absorption relates exponentially to the distance of sample passed through the path length, l ³⁰

$$I = I_0 \cdot 10^{-\epsilon Cl} \quad (13)$$

Where the intensity transmitted radiation is a product of the incident intensity and the exponential term, made up of the molar extinction coefficient, the molar concentration of the sample and the path length. Commonly considered absorption terms include a reformulation of the Beer-Lambert law gives intensity in terms of the density of particles and the molecular absorption cross-section, σ .

$$I = I_0 \cdot 10^{-\sigma\rho(\omega)l} \quad (14)$$

The intensities of electronic transitions can be described by two fundamental quantities, the first being the transition dipole moment.³¹ According to the electronic dipole approximation, the intensity of a transition is given by the square of the transition dipole moment $|\mu|^2$. The transition dipole moment for a transition between two states is calculated as a matrix element of the dipole moment operator μ

$$\mu_{ba} = \langle \Psi_b | \mu | \Psi_a \rangle \quad (15)$$

$$\mu = \sum_i q_i r_i \quad (16)$$

Where the dipole moment iterates over i particles with charge q_i and coordinates \mathbf{r}_i . Another commonly quantitative descriptor commonly reported is the oscillator strength, f , proportional to the square of the transition dipole moment for a transition.³² It is defined by comparison of rates between the molecular transition and that of a classical single electron oscillator by the relation

$$f = \frac{g_b}{g_a} \cdot \frac{2m_e \omega_{ba}}{3\hbar e^2} |\mu_{ba}|^2 \quad (17)$$

Where m_e is the mass of an electron, ω_{ba} is the angular frequency of the transition and e is the elementary charge. The transition dipole moment and oscillator strengths can be more readily related to theoretical principles, in comparison to the extinction coefficient as they are measures of integrated intensity over a band.³³

2.1.3 Molecular absorption

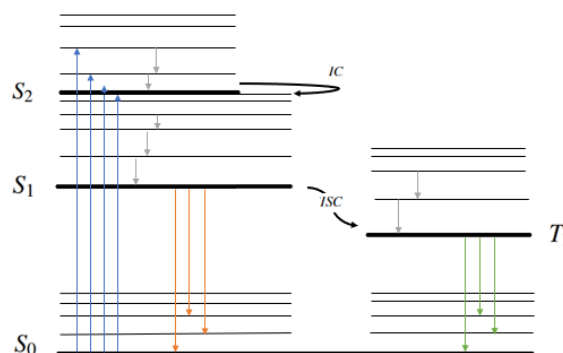


Figure. 2.1. Jabłoński diagram showing the fundamental molecular absorption and relaxation processes, adapted from G. Insero, F. Fusi and G. Romano *"The safe use of lasers in biomedicine: Principles of laser-matter interaction"*.²

The Jabłoński diagram gives us a convenient representation of the transitions available at the molecular level, describing both radiative and non-radiative processes, represented with straight and curved lines respectively.³⁴ Absorption occurs vertically into higher energy vibronic singlet states (blue), with subsequent radiationless decay by internal vibrational relaxation (grey) such that the molecule is only electronically excited. If there is a strong overlap between vibrational energy levels of two electronic states of the same spin, there is a possibility of a transition into a vibrational energy level in the lower energy state, named internal conversion (IC). Both IC and IVR are extremely fast processes and can be thought of as occurring immediately after absorption, with timescales from $10^{-11} - 10^{-14}$ s. transitions. Once in the first excited electronic state fluorescence can occur to the ground electronic state, this is not a fast process, taking place in the order of $10^{-7} - 10^{-9}$ s.³⁵ It is only at the first excited state that fluorescence competes with other non-radiative processes and does not occur from higher electronic states because of the increased likelihood of internal conversion, this is known as Kasha's rule and formally defines the emitting level of a given multiplicity to be the lowest excited level of that multiplicity.³⁶

Vibrational levels get denser as the potential energy increases, internal conversion occurs from a low vibra-

tional level of an excited state, where vibrational levels are sparsely populated to a high vibrational energy level of a lower energy excited state, where vibrational levels are a quasi-continuum of states. The rate for such a conversion is given by Fermi's rule for internal conversion, Equation 18.³⁷ Where the integral is a matrix element of the perturbation H' between the two states, and $\rho(E_b)$ is the density of target states. As such, molecules will experience rapid vibrational relaxation to the lowest excited state.

$$k_{ba} = \frac{2\pi}{\hbar} \left| \langle b | H' | a \rangle \right|^2 \rho(E_b) \quad (18) \text{eq2}$$

Triplet states are often drawn shifted horizontally from singlet states, this is done for clarity as in the Jabłoński diagram the horizontal axis is meaningless. Horizontal transitions to triplet states are formally forbidden due to the spin selection rules $\Delta S = 0$ and occur at a rate several magnitudes slower than fluorescence. Coupling between electronic and vibrational states increases the rate of transition into a triplet state, making it competitive with radiative emission from the ground state. A similar relaxation is seen with either an emission of a photon, or the dissipation of energy as heat, the emission of a photon from a triplet state is called phosphorescence and happens with a lifetime of $10^{-4} - 10^4$ s meaning that the molecule can spend a relatively long time in the triplet state.

The Jabłoński diagram acts as a simple picture of the processes occurring following the absorption of light, however there is no consideration of nuclear motion on a reaction coordinate or any relation to the nuclear degrees of freedom, as the horizontal axis does not have a scale. Introduced below is the Adiabatic approximation and its consequences in photochemistry, a key outcome being the derivation of the potential energy surface.

2.1.4 The Born-Oppenheimer approximation

Regarding electronic transitions of molecules, Born and Oppenheimer proposed the separation of electronic and nuclear (vibrational) motion, representing the total wavefunction as a product of the electronic wavefunction at fixed nuclear positions, and the nuclear vibrational wavefunction. This separation can be justified by considering the huge relative difference in mass of electrons and nuclei, with nuclei being at least $M_{\text{proton}} / M_{\text{elec}} = 2 \times 10^3$ times heavier than electrons. While nuclei and electrons possess the same charge, the velocity of nuclei in relation to electrons is extremely slow, therefore it is assumed that upon excitation of a given nuclear position, the electrons have time to distribute the energy according to their independent wavefunctions. Under the Born-Oppenheimer approximation the total Schrödinger equation can be simplified by assuming the static position of nuclei.³⁸

$$\hat{H}\Psi(r, R) = E\Psi(r, R) \quad (19)$$

$$\Psi(r, R) = \sum_i \phi_i(r; R) \chi_i(R) \quad (20)$$

$$\hat{H}_{\text{elec}} \phi_i(r; R) = E_i(R) \phi_i(r; R) \quad (21)$$

Where the total wavefunction is described as a product of ϕ and χ , and the electronic Schrödinger equation can be solved independently. The Born-Oppenheimer approximation simplifies the consideration of electronic

excited states,³¹ with the independent treatment of electronic and nuclear degrees of freedom allowing the investigation of potential energy at a given nuclear configuration. This gives rise to a vital aspect of theoretical chemistry, the potential energy surface, whereby the potential energy is evaluated as a function of nuclear configuration yielding a “surface” when plotted along specific reaction coordinates. Within the Born-Oppenheimer approximation, transfer between adiabatic surfaces of different electronic states is forbidden, therefore once a molecule is on an adiabatic surface, it can only propagate according to the given state.

2.1.5 The Franck-Condon Principle

With the separation of nuclear and electronic motion it is assumed that upon excitation of a given nuclear position the electrons have time to distribute the energy according to their independent wavefunctions. Therefore, electronic transitions occur on a much faster time scale than nuclear motion, allowing for the assumption that nuclei are fixed during the transition, giving the opportunity for the calculation of the transition probability of a fixed geometry. Upon excitation, the charge redistribution results in a change in Coulombic forces on the nuclei resulting in a vibrationally excited state.³⁹ Considering the Born-Oppenheimer approximation the transition dipole, Equation 15, can be expanded to include both electronic and nuclear components.

$$\mu_{ba} = \langle \phi_b(r;R) \chi_b(R) | \mu_E(r) + \mu_N(R) | \phi_a(r;R) \chi_a(R) \rangle \quad (22)$$

Where the transition dipole moment operator only acts on the electronic components.

$$\mu_{ba} = \langle \phi_b | \phi_a \rangle \langle \chi_b | \mu_N | \chi_a \rangle + \langle \chi_b | \chi_a \rangle \langle \phi_b | \mu_E | \phi_a \rangle \quad (23a)$$

$$= \langle \chi_b(R) | \chi_a(R) \rangle \langle \phi_b(r;R) | \mu_E(r) | \phi_a(r;R) \rangle \quad (23b)$$

The transition from Equation 22 to Equation 23a implies the Franck-Condon approximation, whereby the last integral, the electronic transition dipole moment, no longer depends on the nuclear coordinates. Equation 23a can be reduced, as the first integral over the electronic wavefunctions of the initial and final state, $\langle \phi_b | \phi_a \rangle$, is equal to zero due to the orthogonality of non-degenerate electronic states,³¹ yielding Equation 23b. The probability of a vibronic transition therefore depends on the Franck-Condon factor: $\langle \chi_b | \chi_a \rangle$, determined by the shape and overlap of the eigenfunctions of the vibrational states.

The spin-independent transition dipole moment under the Franck-Condon principle results in a set of vibrational selection rules, arising when one of the integrals is zero.⁴⁰ This leads to a *forbidden* transition, with no probability of occurrence, although only under the approximations used in Equation 23b. Notably, excitation to a higher electronic state occurs vertically from the ground state on a potential energy surface into an area called the Franck-Condon region, with the intensities of the transitions depending on the overlap of the ground state and excited state wavefunctions, Figure 2.2 shows a graphical representation of the Franck-Condon principle. The Franck-Condon principle manifests itself in a red shift of the fluorescence wavelength when there is a difference between ground state and excited state nuclear coordinates, due to the vibrational relaxation in the excited state. Moreover, vibronic progression can be observed in spectra due to the differing overlaps of vibrational wavefunctions.

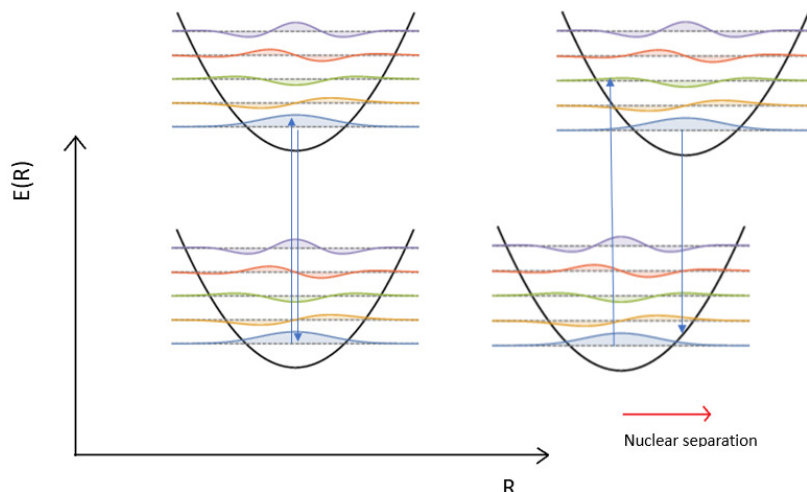


Figure. 2.2. Illustration of Franck-Condon principle, adapted from H. Maguire, J. Iles-Smith, and A. Nazir "Environmental Nonadditivity and Franck-Condon physics in Nonequilibrium Quantum Systems"³

2.1.6 Breakdown of the Born-Oppenheimer approximation

While the adiabatic approximation is a convenient representation for the treatment of electronic transitions in molecules, it fails to fully capture the behavior of molecules on potential energy surfaces, limiting them to a single surface with no probability of hopping from one state to another. The simulation of photochemical processes is an example of the breakdown of the Born-Oppenheimer approximation, with the light-matter interactions and the resulting ultrafast dynamics which gives rise to transitions and couplings between electronic states.⁴¹ Such non-adiabatic crossings can be illustrated by considering a 2×2 Hamiltonian matrix with dependence on nuclear geometry.

$$H = \begin{bmatrix} E_1(R) & V(R) \\ V(R) & E_2(R) \end{bmatrix} \quad (24)$$

Where H is the many-electron Hamiltonian, and adiabatic potential energy surfaces are found as eigenvalues. Analysis of the matrix considering degenerate adiabatic states within the Born-Oppenheimer approximation yields two conditions for degeneracy:⁴²

$$E_1(R) - E_2(R) = 0 \text{ and } V(R) = 0 \quad (25)$$

Evaluating these conditions in a three-dimensional plot yields a characteristic double cone shape around the point of intersection, giving them their name, conical intersections. An important consequence of the conditions for degenerate states is that conical intersections are not isolated but appear as seams in the configurational space.⁴³ Differentiation of Equations 25(a,b) and gives two vectors \mathbf{g} and \mathbf{h} , the branching vectors, with motion along the branching vectors lifting the degeneracy of states. Conversely, distortions orthogonal to the branching vectors upholds the degeneracy, therefore defining the seams in $(N-2)$ -dimensions, where N is the number of nuclear degrees of freedom.⁴⁴ A molecule passing sufficiently close to the seam will be 'funneled' through to a new adiabatic potential energy surface and thereby into a transition between electronic states. The vector for branching space, also known as the non-adiabatic coupling vector, can be evaluated according to Equation 26

$$h_{12} = \left\langle \Psi_2 \left| \frac{\partial \hat{H}}{\partial R} \right| \Psi_1 \right\rangle \quad (26)$$

Where the coupling arises from the application of the nuclear momentum operator to an excited vibronic state, approaching unity as the energy between states decreases.⁴⁵ The resulting intersections are topologically protected aspects of the potential energy surfaces, remaining despite small changes in the chemical environment, resilient to small perturbations of the Hamiltonian.⁴⁶ Conical intersections of different shapes will result in crossings with different non-adiabatic decay properties. A conical intersection leading to one path suggests a non-radiative relaxation to the ground state as commonly seen in biological systems such as DNA, providing stability with respect to UV radiation.⁴⁷ In contrast, conical intersections with multiple pathways suggest the formation of a photochemical product as seen in processes such as photodissociation, photosynthesis and photo-induced rearrangements. Figure 2.3 shows a summary of photophysical and photochemical processes on a potential energy surface diagram.

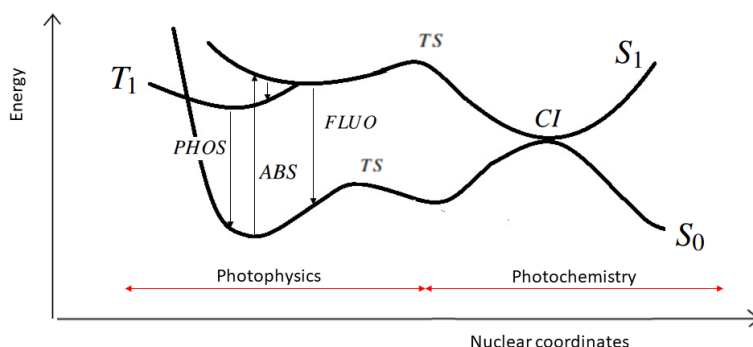


Figure. 2.3. Potential energy surface diagram for photo-excitations and subsequent relaxations. adapted from M. E. Casida, B. Natarajan, and T. Deutsch "*Real-time dynamics and conical intersections*".⁴

2.2 Photodissociation

Photodissociation is a light-initiated chemical reaction where molecules are broken down by photons, involving the interaction of one or more photons with a target molecule. Absorption of photons with sufficient energy will initiate a photodissociation pathway for a given molecule, given the inverse relationship between energy and wavelength of a photon, radiation with the energy of visible light or higher can cause photodissociation to occur. Photodissociation acts as a key contributor to secondary VOC concentrations in our atmosphere, responsible for phenomena such as photochemical smog. For example, the photodissociation pathway for nitrogen dioxide can be described as



Where a photon is absorbed, giving an excited state and a subsequent breakdown of the molecule, forming nitrogen oxide and atomic oxygen. A subsequent reaction of the oxygen triplet state with an oxygen molecule yields ozone at wavelengths below 397.8 nm. These photolysis products contribute significantly to local and global atmospheric conditions, increasing the concentration of tropospheric ozone and other harmful secondary VOCs. The lifetimes of volatile organic compounds in the atmosphere is determined by their lifetimes, ranging from seconds to months. Less reactive VOCs with higher lifetimes can be transported in the atmosphere and contribute to global concentrations of tropospheric ozone through their reactions with nitrogen oxides. On the other hand, reactive VOCs like isoprene or terpenes rapidly undergo oxidation and photodissociation, forming oxygenated products such as hydroperoxides and secondary organic aerosols. Investigating the rates of such reactions is critical to accurate and informed modelling of atmospheric molecules.

2.2.1 Photolysis rates

$$\int_{\lambda_{\min}}^{\lambda_{\max}} \phi(\lambda) \sigma(\lambda) F(\lambda) d\lambda \quad (28)$$

Photolysis reactions can be described using a first-order rate constant (J), which is formulated as an integral of the product of the photoabsorption cross-section σ , the solar actinic flux F , and quantum yield ϕ over the available wavelengths of light.⁴⁸ It can be thought of as the probability of a molecule absorbing light of a given wavelength multiplied by the intensity of light at that wavelength, adjusted for the quantum efficiency. The actinic flux is independent of the molecule involved depending only on environmental factors, while the cross-section and quantum yield relate to the chosen mechanism and depends on the resolution of molecular properties. Due to the sporadic nature of photolysis reactants and products these molecular properties are often illusive empirically, allowing theoretical chemistry to take the reins and fill the knowledge gaps.

Other pathways for photochemical reactions or relaxations also need to be considered, for example collisional quenching of the excited state will reduce the lifetime of an excited VOC, thereby influencing the rate of reaction.

2.2.2 Photoabsorption cross-sections

The relation of the photoabsorption cross-section, introduced in Section 2.1.2, to the rate of photolysis makes them ideal candidates for further study, it is derived from time independent perbutation theory that the cross-section for a single transition between states a and b follows

$$\sigma_{ab}(\omega) = \frac{\pi\omega}{\hbar\epsilon_0 c} |\epsilon\mu_{ba}|^2 \delta(\omega_{ba} - \omega) \quad (29)$$

Where the total absorption cross-section can be obtained as a summation over states, moreover, for isotropic systems the expression can be simplified by averaging over all orientations and substituting $|\epsilon\mu|^2 = 1/3|\mu|^2$

$$\sigma_{ab}(\omega) = \frac{\pi\omega}{3\hbar\epsilon_0 c} \sum_b |\mu_{ba}|^2 \delta(\omega_{ba} - \omega) \quad (30)$$

Equation 30 acts as a basic formulation of cross-sections for stationary states. As discussed previously when considering molecular transitions, it is convenient to describe molecules within the Born-Oppenheimer approximation, applied to spectroscopy, this is manifested by the application of the electromagnetic interaction

only to electrons, due to the increased mass of nuclei. Following the discussion of Born-Oppenheimer in Section 2.1.4, Equation 30 can be further simplified to include only contributions of the electronic wavefunction, the common time-independent representation within the Born-Oppenheimer framework.

$$\sigma_{aj}(\omega) = \frac{\pi\omega}{3\hbar\epsilon_0 c} \sum_{b,k} |\mu_{bk,aj}|^2 \delta(\omega_{bk,aj} - \omega) \quad (31)$$

$$\text{where: } |\mu_{bk,aj}| = \langle \chi_{bk} | \mu_{ba} | \chi_{aj} \rangle \quad (32)$$

Further, upon consideration of the Franck-Condon principle, the nuclear locations remain unchanged on excitation, therefore the intensity of a transition can be taken as the overlap between the initial and final wavepackets, the Franck-Condon factor, adjusted for transition dipole magnitude from the ground state geometry.⁵

$$\sigma_{aj}(\omega) = \frac{\pi\omega}{3\hbar\epsilon_0 c} \sum_{b,k} |\mu(R_0)|^2 |\langle \chi_{bk} | \chi_{aj} \rangle|^2 \delta(\omega_{bk,aj} - \omega) \quad (33)$$

Where $\mu(R_0)$ is the zeroth constant term of a Taylor series expansion about the equilibrium geometry. The extension to a thermal ensemble of states and transitions can be achieved by a summation over transitions, weighted by the probability of the given transition, which can be found for a molecule in thermal equilibrium by considering the boltzman factors of each state compared to the total population.

$$\sigma(\omega) = \sum_j p_{aj} \sigma_{aj}(\omega) = \sum_j \left(\frac{e^{(\beta E_{aj})}}{\sum_i e^{(\beta E_i)}} \sigma_{aj}(\omega) \right) \quad (34)$$

$$\text{where: } \beta = \frac{1}{K_B T}$$

2.2.3 Reflection principle

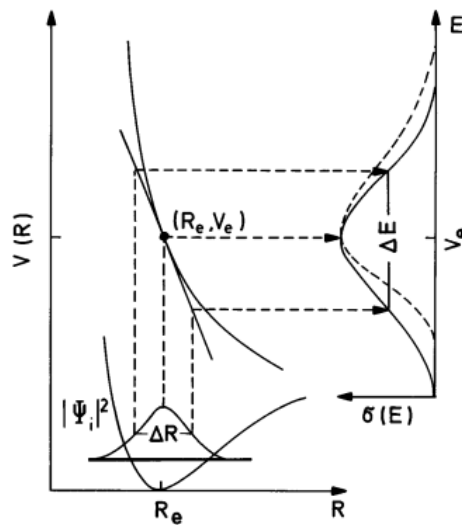


Figure. 2.4. The reflection principle, Schinke *"Photodissociation Dynamics"*.⁵

The reflection principle consists of reflecting the nuclear density onto an excited state, and subsequently from the excited state onto an energy axis of the spectrum, the one-dimensional reflection principle is shown in

Figure 2.4. Initially applied to diatomic molecules,⁴⁹ and subsequently extended to polyatomic molecules,^{50,51} it can be seen as a short-time approximation to linear electronic spectra, producing an envelope of spectral intensities. Under the reflection principle, the intensity of a spectral transition at a given energy level becomes proportional to the probability of finding the molecule in a certain configuration, thereby relating spectral intensity to the ground state nuclear distribution. By considering only the ground state the reflection principle does not impose commonly limiting approximations, allowing for the exploration of non-Condon effects and symmetry forbidden transitions. The reflection principle is particularly suitable for the modelling of dissociative states, in which the excitation is occurring to a repulsive potential, in other words a continuum of vibrational states. Due to the continuous nature of transition energies and intensities, fine structures such as vibronic progression bands are not observed.⁵ The general form of the reflection principle is given by Equation 35.

$$\sigma(E) = \frac{\pi E}{3\hbar\epsilon_0 c} \sum_b \int P_a(R) |\mu_{ba}(R)|^2 \delta(E - E_{ba}(R)) dR \quad (35)$$

Where E is the incident photon energy, R is the position operator, $P_a(R)$ is the probability density of the initial state, $E_{ba}(R)$ is the difference in energy of the two states and $\mu_{ba}(R)$ is the transition dipole moment. Under both the electronic dipole estimation and the Born-Oppenheimer principle, the probability density can be taken as the population density of the ground state ρ , and the difference in energy E_{ba} is defined as the difference between two potential energy surfaces for a given geometry.

$$P_a(R) = \rho_a(R) \quad (36)$$

$$E_{ba}(R) = V_b(R) - V_a(R) \quad (37)$$

This forms the basis of the nuclear ensemble approach, which will be discussed in Section 2.3.2

2.3 Calculating photoabsorption cross-sections

Recent advances in computational chemistry allow for the calculation of transition intensity properties, thereby paving the way for the simulation of absorption cross-sections, a key component of photolysis rate constants. The simplest calculations, considering a single structure, involve exciting the lowest energy ground state geometry to an excited state. While accurate for the chosen structure, such calculations are not representative of the molecular state, as proposed by Schrödinger in the quantum mechanical model of the atom. Quantum mechanics represents the molecular ground state as a distribution of possible structures, with the implications discussed in the previous section.

2.3.1 Time-dependent and independent methods

There are two main differences in approaches to calculating the absorption cross-section, and thereby the photolysis rate constant of a molecule. These are time independent and time dependant methods, a graphical representation is seen in Figure 7.1. In time independent simulation, the Franck Condon integrals between ground and excited state wavefunctions are considered to determine the intensity of transitions.⁵² Time independent methods fail to simulate non-Condon effects, often requiring corrections considering the dependence of the transition dipole moments on nuclear coordinates. A common strategy lies in the use of Herzberg-Teller terms to account for vibronic coupling, resulting in the Franck-Condon Herzberg-Teller theory. FCHT relies on

the harmonic approximation, therefore, it breaks down for molecules with pronounced anharmonicity or dissociative character in their excited states.⁶ Further corrections to account for anharmonicity have been proposed, however, in practice this requires the calculation of vibrational excited states which can be costly.⁵³

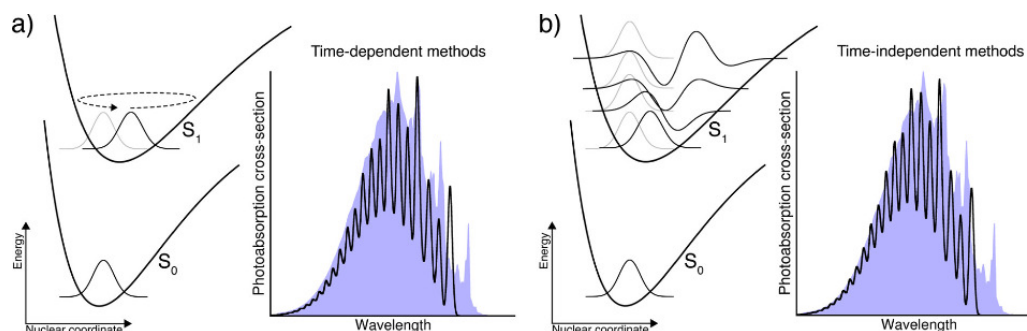


Figure. 2.5. Visualisation of photoabsorption cross sections of a transition from $S_0 \rightarrow S_1$ modelled by a) time-dependent, and b) time-independent methods, Prlj. et al. "Calculating Photoabsorption Cross-Sections for Atmospheric Volatile Organic Compounds".⁶

Similarly, exact quantum dynamics solutions using time-dependent methods require great computational resources, resulting in limitations on the size of molecules that can be considered. Further approximations such as reducing the dimensionality to the most prominent modes can be employed for simplicity, however, this can result in the production of insufficiently accurate photoabsorption cross-sections. The increased complexity of calculations for larger and flexible molecules using both time dependent and independent methods acts as a motivator for the development of new techniques for calculating the excited states and thereby absorption cross-sections.

2.3.2 Nuclear Ensemble Approach

Considering the uses of photo-absorption spectra, it is not often that the exact resolution of vibronic states is required for meaningful inferences to be made. For example, in atmospheric applications that concern only the width and shape of bands rather than their precise structure such as the calculation of photolysis rate constants. The nuclear ensemble method is rapidly gaining traction as a means of calculating photoabsorption cross-sections, and has been shown to give robust approximations over a wide range of compounds,^{54–58} notably the *ab initio* calculation of photochemical processes for transient VOCs.^{6,59–61} The method, based on the reflection principle, entails sampling a ground state wavefunction to pick a so called “nuclear ensemble”, performing single point calculations on each sample to achieve multiple vertical excitations, yielding a stick spectrum of transition intensities. These vertical transitions are broadened according to Gaussian or Lorentzian functions, like the reflection principle, to simulate the shape of absorption bands, with widths less than the bandwidth to not affect the overall shape of the spectrum.⁵⁵ Figure 2.6 shows a graphical representation of the nuclear ensemble approach, and the resulting envelope spectrum. Calculating transition dipole moments as a function of nuclear configuration accounts for non-Condon effects, moreover, the excited state wavefunction is not considered at all, greatly reducing the computational complexity. The exclusion of the excited state wavefunction results in a lack of vibronic resolution in the spectrum in comparison to more involved quantum mechanical methods, this is replaced by a spectrum showing an envelope of transitions, accounting for both bright and dark bands due to the lack of Condon integrals in its formulation.

The ground state wavefunction sampling method strongly influences the ground state properties of the resulting nuclear ensemble. Sampling can be done assuming the molecule is in thermal equilibrium (thermal) or

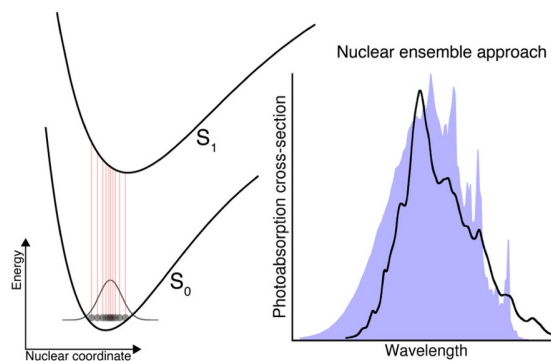


Figure. 2.6. A visualization of the nuclear ensemble approach (NEA) for the calculation of photoabsorption cross-sections. Prlj. et al. "Calculating Photoabsorption Cross-Sections for Atmospheric Volatile Organic Compounds".⁶

at the zero-point energy level (quantum), thermal sampling lacks at temperatures of 300K due to the underprediction of mean energies. On the other hand, quantum sampling of the ground state wavefunction from a Wigner distribution provides satisfactory spectra at room temperatures, outperforming thermal sampling. Wigner transformations for multidimensional molecules with anharmonic character are challenging, often resulting in the need for approximate harmonic distributions. The introduction of harmonic approximations could limit the performance of the NEA on volatile organic compounds of interest such as 2-hydroperoxy-propanal (2HPP). Regarding VOCs a combinatory approach has been composed, the quantum thermostat (QT), based on non-equilibrium dynamics fitted such that it matches a quantum harmonic oscillator. Shown to accurately sample both modes dominated by quantum effects and anharmonic modes, QT shows promise as a replacement for molecules with anharmonic characteristics.⁶ Due to the use of ab initio molecular dynamics in the sampling, QT provides improved accuracy with the sacrifice of computational cost, making it unsuitable for molecules where Wigner sampling is available.

The Wigner distribution function, under the harmonic approximation, for a non-linear system with $3N-6$ vibrational degrees of freedom where N is the number of atoms in the molecule is expressed in terms of normal mode coordinates and momenta, p and q respectively.⁶² It provides a link between the wavefunction formulation used in quantum mechanics and a joint probability distribution in the phase space, defined as:

$$\rho^w(q, p) = \frac{1}{(\pi\hbar)^{3N-6}} \prod_{i=1}^{3N-6} \exp\left(\frac{-q_i^2}{2\omega_{qi}^2}\right) \exp\left(\frac{-p_i^2}{2\omega_{pi}^2}\right) \quad (38)$$

$$\omega_{qi}^2 = \frac{\hbar}{2\mu_i\omega_i} \text{ and } \omega_{pi}^2 = \frac{\hbar\mu_i\omega_i}{2} \quad (39)$$

The terms ω_{qi} and ω_{pi} are the angular frequency parameters for the i -th normal mode, where μ is the reduced mass, ω_i is the frequency of the i -th normal mode and \hbar is the reduced planck's constant. Once obtained, the Wigner function can be used to form a ground state nuclear ensemble, by randomly choosing coordinates and momenta values, calculating the Wigner quasiprobability and comparing this to a random number generated from a uniform distribution with an interval from 0 to 1. The generated values of coordinates and momenta are accepted if they are larger than the random number, as such an ensemble is formed using a relatively low-cost method.

Following the formation of a sufficiently descriptive ensemble, the molecular photoabsorption cross-section of any sampled nuclear ensemble, derived from the reflection principle, can be calculated using Equation 40

where a Monte Carlo procedure is used to evaluate the integrals on \mathbf{R} .⁵⁵

$$\sigma(E) = \frac{\pi e^2 \hbar}{2m_e c \epsilon_0 E} \sum_L \frac{1}{N_p} \sum_n^{N_p} \Delta E_{0 \rightarrow L}(\mathbf{R}_n) f_{0 \rightarrow L}(\mathbf{R}_n) \times w_s[E - \Delta E_{0 \rightarrow L}(\mathbf{R}_n), \delta] \quad (40)$$

2.3.3 Statistical representation of the NEA

The nuclear ensemble approach lends itself to representation using statistical density estimation methods, the simplest application includes decomposing the resulting spectra, modelled as histograms with small bin widths. Subsequently, total transition dipole moments can be evaluated from the number of transition dipole moments falling in the same bin. Such an approach requires a large number of geometries sampled to produce smooth spectra and fails with small ensemble sizes. A better alternative is to use kernel density estimation applying a Gaussian kernel to each single point in the spectrum, thereby smoothing the final spectrum.^{63,64} The general form of the gaussian broadening kernel density estimation function is given in Equation 41.

$$f(x) = \frac{1}{nH} \sum_{i=1}^n K\left(\frac{x_i - x}{H}\right) \quad (41)$$

$$\text{where } K = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The nuclear ensemble approach can be formulated using kernel density estimation by applying the equation for kernel density estimation, Equation 41 to that of photoabsorption cross-sections, Equation 40. The quantum mechanical representation of the ground state wavefunction typically follows a gaussian distribution, therefore, Equation 41 can be considered without changes.

$$\sigma_H(E) = \frac{\pi}{3\hbar\epsilon_0 c} \sum_{b=1} \frac{1}{\sqrt{2\pi n H_b}} \sum_{i=1}^n E_{ba}(\mathbf{R}_i) \mu_{ba}^2(\mathbf{R}_i) e^{-\frac{\frac{1}{2}(E - E_{ba}(\mathbf{R}_i))^2}{H_b^2}} \quad (42)$$

Where the bandwidth is calculated for each separate transition to the final state b , allowing for broadening to be set for differently sharp features of the spectrum. The weight is proportional to the energy difference of the two states, and the transition dipole moment, which is a familiar relation.

With the NEA showing great potential as an accurate method for the simulation of absorption spectra, in practice, it is limited by the computational burden of launching thousands of QM calculations to produce spectra representative of experimental results. The reduction of computational intensity of the nuclear ensemble method is of paramount importance in making the algorithm suitable for application on larger systems. The starting point for cost reduction is to introduce additional approximations into the quantum mechanical calculations, however this risks the creation of spectra with poor accuracy. A more welcome alternative is to reduce the number of calculations required to create a spectrum of similar resolution, reducing the computational load required for spectral simulations. Multiple statistical approaches have been brought forward, using exploratory sampling of the ground state wavefunction to observe the regions of high density and focusing subsequent calculations in these regions.

2.4 Atmospec

Atmospec is an application for the generation and visualization of photoabsorption cross-sections using the nuclear ensemble approach, developed by Daniel Hollas within the inSilicoPhotochem group.⁷ It aims to make the *in silico* calculation of spectra more accessible for atmospheric chemists by reducing common barriers for entry, such as those unfamiliar with computational chemistry or programming. It streamlines the often complex process of structure optimization, ensemble sampling, excited state calculations and finally, spectrum calculation into one simple submissions. This simplified workflow stands to increase use of *in silico* calculations within the scientific community, allowing for more to leverage advances in computational chemistry for their research.

Atmospec is built on AiiDaLab, a cloud-based environment for the development and packaging of computational workflows, offering practical user interface and workflow management solutions.⁶⁵ It is open source and distributed using a remotely hosted docker container, allowing for the easy and fast installation and initialisation for development on a local machine, encouraging collaboration. The application itself is hosted as a web server, providing a user interface using Jupyter iPyWidgets, meaning that users can access the resource without direct installation. Additionally, it provides options for the submission to a HPC cluster, lifting the often-limiting factor of local computational resources. Photoabsorption cross-sections are calculated according to Equation 40, and quantum mechanical calculation are performed using ORCA, an open source quantum chemistry package.⁶⁶ AiiDa significantly hides the complexity of a nuclear ensemble approach workflow, condensing it into a single submission.

2.4.1 Current functionality

Atmospec implements an automated workflow for the calculation of photoabsorption cross-sections, from structure selection to the calculation and visualization of cross-sections and subsequently photolysis rate constants.

Workflow initialization

The workflow begins with structure selection using either a SMILES code, a geometry file, or a structure in the AiiDa database. The selection of a structure prompts a 3D interactive display of the selected molecular structure and allows for progression to the submission step. Calculation conditions are set using a combination of drop-down menus and text inputs, with inputs validated and flagged if incorrect upon submission, thereby reducing the risk of wasting resources by submitting an incorrect input. Following the selection of calculation parameters the user is prompted to configure a submission code, which tells AiiDa where the calculation should be executed and the resources that should be used. Table 2.1 shows a summary of configurable variables for the initialization of an Atmospec workflow.

Table. 2.1. Current input requirements for Atmospec workflow

Molecular options	Ground state structure	Excited state calculations	Wigner sampling
Compound	Functional	Method	Number of samples
Geometry optimization	Basis set	Number of states	Low energy cut-off
Charge		Functional	
Multiplicity		Basis set	
PCM solvation			

Submissions are managed using the SLURM workload management system, to ensure the efficient use of

computational resources, and executed using ORCA.⁶⁷ The successful submission of a workflow results in progression to the status step, where a summary of the completed processes and overall calculation progress can be seen. This provides the user real-time details on individual steps in the workflow, allowing them to track their submission. If an error is thrown during the submission, it is here that it is displayed, and the submission directories can be observed for debugging, alternatively, if the calculation is successful, individual properties files, such as information about the molecular orbitals of an individual in the ensemble, can be viewed and downloaded for further inspection.

Spectrum visualization and analysis

Following the generation of spectrum data, an interactive graph with the photoabsorption cross-section is displayed, see Figure 2.7. The interactivity allows the user to zoom on areas of interest, as well as dynamically change the values of actinic flux and quantum yield. The broadening is determined by the user, with either gaussian or Lorentzian broadening schemes available, and can be changed using a slider in the spectrum visualization step, additionally, the user can evaluate individual conformer contributions to the spectrum using the conformer selection tool.

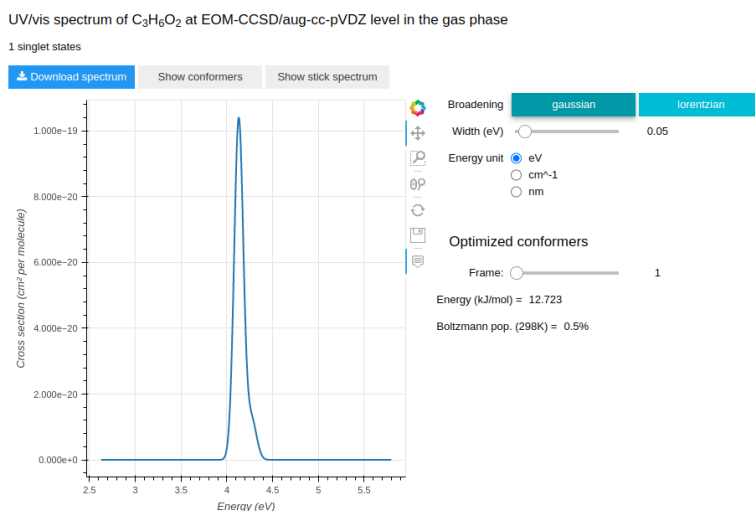


Figure. 2.7. Atmospec spectrum visualisation for 2-hydroxypropanal using EOM-CCSD with the aug-pVDZ basis set. Generated in *Atmospec* v0.2.3.⁷

The visualization of the photoabsorption cross-section is followed by a calculation of the photodissociation rate constant using Equation 28, as well as a visualization of the photolysis rate coefficients. Figure 2.8 shows an example spectrum analysis output, with the rate constant alongside a graph of rate coefficients, and individual transitions. The values of photolysis rate coefficients and the rate constant are recalculated for each change in the spectrum, providing a dynamic visualization and allowing researchers to rapidly evaluate the impact of changes to calculation conditions, such as the flux or the quantum yield, on the rate of photolysis. A plot of oscillator strengths and excitation energies is displayed in a separate tab, either as a scatterplot or a 2D density plot for further analysis of the photoabsorption cross-section contributions. The controlled nature of Atmospec workflows leaves little room for human error in the submission of calculations, making the theoretical calculation of photoabsorption cross-sections more accessible to a wider range of researchers.

With the excellent reception of the first stable release in July 2024, it is clear that these tools are a much needed addition to arsenal of experimental and theoretical chemists alike. Moreover, active development is undergoing, numerous bugfixes and enhancements provided. Unfortunately, the current approach using NEA

Spectrum analysis

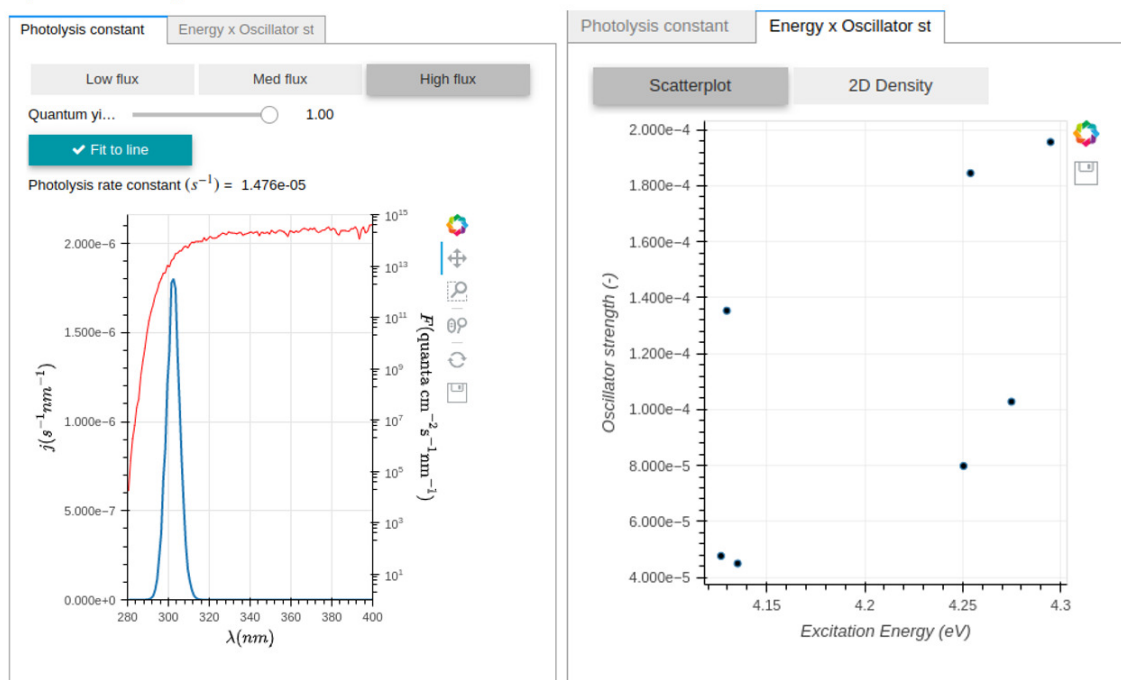


Figure. 2.8. Photolysis rate constant calculation, rate coefficient visualization, and transition properties widgets for 2-hydroxypropanal using EOM-CCSD with the aug-cc-pVDZ basis set. Generated in *Atmospec* v2.0.3.⁷

can be limited for larger, or more complex molecules due to the increasing computational intensity, resulting in calculations that can be unfeasible to run. This prompts the investigation of algorithmic improvements, leveraging recent statistical innovations to increase the efficiency of nuclear ensemble approach.

2.5 Developments in the nuclear ensemble approach

The nuclear ensemble approach has been the subject of multiple research endeavors in recent years with the uptake being accelerated by the design of automated workflows and graphical user interfaces for the submission of NEA calculations. Similarly, public programming packages for the initiation of NEA workflows, such as Newton-X,⁶⁸ allow for more widespread adoption of the algorithm, which in turn fuels scientific interest and research. Moreover, the development of graphical user interfaces such as Atmospec stands to further increase the rate of adoption by both theorists and experimentalists alike. Similarly, algorithmic developments focused on improvement of efficiency and functionality have been proposed, these advances coupled with the advent of machine learning across the sciences have led to rapid developments in the NEA algorithm. The potential for replacing costly quantum chemistry calculations with rapid and adaptable machine learning calculations has been of interest in the calculation of photoabsorption cross-sections. Additionally, the use of statistical techniques to reduce the cost of spectra calculations has shown great prospects for decreasing the resource intensity.

2.5.1 Importance sampling

Importance sampling is a technique for mapping a sampling probability distribution function to another distribution of interest, holding great potential within the context of the nuclear ensemble approach due to its probability density representation of absorption cross-sections. It was proposed as a method to reduce the cost

of recalculating spectra at different temperatures, allowing the calculation of spectra at different temperatures without the additional cost of electronic structure calculations.⁶⁹ Within the nuclear ensemble approach excitation energies and transition dipole moments are calculated for an ensemble of points distributed in nuclear phase space. This can be expressed as

$$\mu_P = \frac{1}{N} \sum_{i=1}^N f(X_i) P(X_i) \quad (43)$$

Where most of the computational cost lies in the calculation of $f(X_i)$, which is some molecular property, typically calculated using expensive quantum chemistry methods making the trivial recalculation of spectra unfeasible. By introducing a target distribution $Q(X_i)$ at a new temperature the expectation values can be rewritten as

$$\mu_Q = \frac{1}{N} \sum_{i=1}^N f(X_i) \frac{Q(X_i)}{P(X_i)} P(X_i) \quad (44)$$

Thereby expressing the new distribution as a weighting $w = \frac{Q(X_i)}{P(X_i)}$ of the original distribution, where $Q(X_i)$ can be obtained from a quantum-based distribution such as a Wigner function. This method allows for the rapid recalculation of spectra at varying temperatures with relatively little additional cost and was implemented as a module in the Newton-X package by Fabris Kossoski, with application to both photoabsorption spectra and molecular dynamics simulation.⁶⁸

2.5.2 Statistical determination of the broadening parameter

Within the nuclear ensemble approach the only undefined quantity is the gaussian broadening parameter. Silverman's rule of thumb is a general approximation of bandwidth in kernel density estimation. It is applicable to gaussian distributions and provides a process for the evaluation of bandwidth thereby completing the NEA calculation, however it is important to note that it will yield inaccurate results if the density is not sufficiently close to being normally distributed.⁷⁰ Silverman's rule of thumb for weighted distributions is given by

$$H = \left(\frac{4}{3n} \right)^{\frac{1}{5}} \sigma \approx 1.06 \sigma_s n_{eff}^{\frac{1}{5}} \quad (45)$$

Where σ is an unknown standard deviation and σ_s is the standard deviation given by the sample data, and $n_{eff,b}$ is Kish's effective sample size of state b .⁷¹ Kish's effective sample size can be calculated from data with weights $w_{b,i}$ according to Equation 46.⁷²

$$n_{eff,b} = \frac{(\sum_i^n w_{b,i})^2}{\sum_i^n w_{b,i}^2} \text{ where } w_{b,i} = E_{ba}(R_i) \mu_{ba}^2(R_i) \quad (46)$$

can be extended to a formula for multivariate bandwidth in d dimensions, derived from the multivariate normal distribution

$$H = \left(\frac{4}{(d+2)n} \right)^{\frac{1}{d+4}} \sigma_s \quad (47)$$

Silverman’s rule of thumb with Kish’s effective sample size provides a good initial guess for the bandwidth of a sample, minimizing the mean integrated squared error (MISE) between the original probability density function and the kernel density estimation. It is limited by its derivation from the normal distribution and provides good estimates only for samples closely following such a form. Incorporating the interquartile range into Equation 47 allows for slight deviations from the normal distribution.⁷³ Additionally, Silverman’s rule is only applicable to unimodal distributions, rendering it unsuitable for a large number of photoabsorption cross-sections. An alternative, yet similar, method has been proposed more recently by Feher et al, which simultaneously minimizes the MISE between the original PDF and the KDE prediction, as well as the leave-one-out cross validation error.⁷⁴ While each method has subtle differences, all of them provide comparable results

2.5.3 Machine learning approaches

Machine learning (ML) methods have the potential for rapid and accurate calculations of target variables when applied with diligence and given suitable training data. Multiple machine-learning Python libraries have been developed for applications in chemistry, such as MLatom, and RDkit,^{75,76} allowing for the straightforward implementation of molecular descriptor transformations and machine-learning models to chemical data. Considering the application of machine learning to the nuclear ensemble method, two different methodologies can be used: either supervised or unsupervised algorithms. Supervised approaches are well suited for the prediction of quantum chemical properties due to the requirement of labelled data, whereas unsupervised approaches, where data is unlabeled, are more suited to the automated calculation of bandwidth.

2.5.4 Machine learning for quantum chemistry calculations

The complexity and therefore computational intensity of *ab initio* quantum chemistry calculations is often a limiting factor in research workflows, leading to bottlenecks due to limited resource availability. Multiple machine learning methods have been suggested to increase the efficiency of such calculations, the aim being to replace those based on quantum chemistry with predictions yielded from machine learning models.⁷⁷ A well-trained machine learning model can rapidly predict quantum chemistry properties based on suitable global molecular descriptors, such as the coulomb matrix (CM) or the relative inverse nuclear distances (RE). Supervised machine learning algorithms such as neural networks and random forest models have shown great potential for the prediction of QM chemical properties with adoption in industry allowing for high throughput screening of molecules. In the context of the NEA, kernel ridge regression has been proposed to decrease the error seen in spectra calculation with relatively few sampled geometries by simulating thousands of values for the excitation values and oscillator strengths from models trained on only a handful of QM calculations.

Kernel ridge regression (KRR) is a statistical algorithm, making use of the kernel trick, a non-linear gaussian kernel function in a linear model to capture the complex relationship between descriptors and target variables. The data is mapped into higher dimensional space using the kernel functions and a linear regression model is fit in this higher dimensional space. As such a model trained on a small set of data points, where the nuclear geometries, excitation energies, and oscillator strengths have been calculated, can be used to predict those properties for new nuclear geometries, without the need for expensive calculations. Compared to more complex machine learning solutions such as neural networks, KRR is a lightweight and flexible choice, most importantly, it performs well for small datasets (<1000 geometries) and has negligible training cost. (#making it applicable to nea) The simulation of additional points in the spectrum allows for the effective removal of the

gaussian broadening parameter, which is set to a small value (e.g. $\delta = 0.001$), as the additional simulated points provide a smoothened spectrum.

2.5.5 Probabilistic machine learning for spectral reconstruction

As mentioned, one limitation of the nuclear ensemble approach lies in using an arbitrary gaussian broadening parameter, with a characteristic bandwidth set manually. The manual assignment of bandwidth is greatly subjective and should be eliminated for a truly automated workflow, with Cerdan et al. proposing gaussian mixture models as an alternative.⁷³ Gaussian mixture models are a type of probabilistic model that represents a distribution as a weighted sum of multiple gaussian distributions which can summarize complex probability distributions. They have been proposed as an alternative to the kernel density-based approach, using probabilistic machine-learning for the elimination of δ from the nuclear ensemble approach, by acting as a generator of excitation energies and transition dipole moments from a fit GMM. The simulation of additional points again allows δ to effectively be set to zero thereby negating the dependence on bandwidth, moreover it allows reliable spectra to be obtained from ensembles of hundreds, even tens of sampled geometries.

Gaussian mixture modelling for spectrum reconstruction has been implemented into Multispec,⁷⁸ and shown to provide systematically better estimates of photoabsorption cross-sections from small ensembles (< 400) for both the full spectrum and individual band shapes.⁷³

2.6 Representative sampling

The improvements discussed above focus on reducing the error in spectra calculations with a reduced sample size, however they do not consider the quality of the sample itself, often taking a random sample of the ground state wavefunction and using this as the ensemble. This leads to the question of whether there can be a subset (n) of a larger ensemble (N) which could accurately represent the density of the larger ensemble? Representative sampling, proposed by Srsen et al, aims to find subset geometries in a nuclear ensemble that minimize divergence from the original spectrum, with the goal of achieving the same density as the full sample, thereby minimizing data loss.^{79,80} The result being a smaller set of geometries which most accurately describe the initial spectrum.⁸⁰ Subsequently, high-accuracy single point calculations are performed for each geometry and a high-resolution spectrum is produced at a fraction of the cost.

The formulation of one-dimensional kernel density estimation, Section 2.3.3, is easily extended to a weighted multivariate kernel density estimation over d -dimensions with x_i samples,⁸¹ where weights w_i have been added.⁸²

$$f(x) = \frac{1}{n|H|^{\frac{1}{2}}} \sum_{i=1}^n w_i K \left(H^{-\frac{1}{2}} (x_i - x) \right) \quad (48)$$

$$\text{where } K = 2\pi^{-\frac{d}{2}} e^{-\frac{x^T x}{2}}$$

An optimization is performed over the excitation energies and transition probabilities, weighted by intensity to emphasize significant coordinates. A 2-dimensional kernel density estimation is then used to estimate the distributions, an extension of the one-dimensional process as presented in Equation 48.

$$\sigma_H(x) = \frac{\pi}{3\hbar\epsilon_0 c} \sum_b \frac{|H_b|^{-\frac{1}{2}}}{2\pi n} \sum_{i=1}^n E_{ab}(R_i) |\mu_{ab}(R_i)|^2 e\left(-\frac{(x-x_{ab,i})^T (x-x_{ab,i})}{2H_b}\right) \quad (49)$$

Where x is a point in the two-dimensional transition properties space, and $x_{ab,i}$ is a vector between states a and b for the i -th geometry.

$$x_{ab,i} = \left(E_{ab,i}(R_i), |\mu_{ab,i}(R_i)|^2\right) \quad (50)$$

The optimization is performed over multiple iterations, attempting to minimize an objective function, with the subsets represented as a density function, the cumulative distribution function naturally lends itself as a measure of divergence between two subsets. Other statistical difference measures like least squares mean are available, as well as those based on information geometry, which will be covered in the proceeding discussion.

2.6.1 Divergence

Cumulative density functions are an insufficient measure of divergence of the exploratory sample against the representative sample probability density functions, due to its symmetry the cumulative density acts only as a measure of distance between the samples. Information-theoretic approaches such as the Jensen-Shannon divergence and the Kullback-Leibler divergence, also known as relative entropy, can be used as a general measure of difference between two distributions.⁸³ Kullback-Leibler (KL) divergence is an asymmetric distance measure applicable for the comparison of exploratory and representative density functions, given two distributions of the same variable, $P(X)$ and $Q(X)$ the KL divergence is given by

$$D_{KL}(p(X)||q(X)) = \int_{-\infty}^{\infty} p(X) \log\left(\frac{p(X)}{q(X)}\right) dX \quad (51)$$

The Kullback-Leibler divergence is formulated as the expectation of the logarithmic difference between the probabilities P and Q , taken as using the probability density function P .⁸⁴ Generally, the divergence of P from Q will be different to the divergence of Q from P due to the asymmetric nature of information geometry, the exception being when $P = Q$ resulting in no divergence.

$$D_{KL}(P||Q) \neq D_{KL}(Q||P) \quad (52)$$

KL divergence can be thought of as the information lost upon accepting the proposed distribution as the original, with a zero-result given only when two identical distributions are presented, which is desirable for the comparison of exploratory and representative spectra.⁸⁵ The KL divergence is typically applied to normalized density functions, while the spectra are unnormalized, therefore an alternate form for unnormalized densities is used.⁷⁹

$$D_{KL}(p(X)||q(X)) = \int_{-\infty}^{\infty} \left(p(X) \log\left(\frac{p(X)}{q(X)}\right) - q(X) + p(X)\right) dX \quad (53)$$

In comparison to statistical error measures, such as mean squared error, KL divergence puts an emphasis on the tails of distributions, due to their disproportionate likelihood,⁸³ with such tails often of interest in UV-vis spectroscopy.

2.6.2 Sample optimization

Optimization can be performed in the space of nuclear coordinates, but the nuclear degrees of freedom for nonlinear molecules quickly become limiting regarding the convergence on a minimum solution due to $3N-6$ degrees of freedom. Transition properties can be used to as a solution space for optimization, where the density is calculated as a function of the excitation energies and transition dipole moments weighted by spectral intensity. The aim of sample optimization is to segment a subset of geometries of a desired size which is most representative of the density of the full sample, consequently, losing the least amount of information from the exploratory sample.

A simple approach would be to randomly sample the geometries, replacing those with a lower KL divergence to create an improved solution over many iterations. Random sampling with divergence comparisons is not a technique viable for thousands of density distributions, scaling exponentially with the number of samples. A slight alternation on the so-called simple hill climbing technique, would be to consider only neighbors of an initial solution, accepting the best neighboring solution until an optimal solution is found, in this way decreasing the number of comparisons made. Furthermore, it is possible to randomly select neighboring solutions with a probability of acceptance, however, despite modifications the hill climbing algorithms are limited by their tendency to get stuck in local minima, and sensitive to the initial solution.⁸⁶ The simulated annealing algorithm is presented as a meta-heuristic optimization search algorithm.

2.6.3 Simulated annealing

Drawing inspiration from metallurgy, simulated annealing mimics the controlled cooling of metals such as metals to produce products with fewer imperfections. A similar approach is seen in the algorithm, presented by Kirkpatrick et. al, where an initial guess is made, and subsequent iterations of optimization are performed at increasingly cool temperatures.⁸⁷ In other words, the probability of replacing geometries within the subsample decreases as the optimization progresses, allowing it to initially search for solutions sparsely in the transition properties space, avoiding local minima. As the temperature is cooled, the subset converges to minimize the overall KL divergence. It is important to note that simulated annealing is a stochastic sampling technique, based on probability, therefore the minimum provided by the algorithm is not guaranteed to be the global optimum, and is impacted by the simulated annealing scheme used.

The acceptance probability and the rate of cooling play a significant role in convergence of a solution on a global minimum and are closely related to the probability of ending up in a local minimum as they dictate the volatility of the initial solution. To avoid minima, the acceptance probability should evolve from 100 % to 0 throughout the progression of the optimization. A natural representation of the acceptance probability in the case of molecular configurations is a Boltzmann distribution, formulated as

$$p(\Delta f) = \exp\left(-\frac{\Delta D_{KL}}{T_{virt}}\right) \quad (54)$$

The probability of accepting a solution is dependent on the virtual temperature of the simulation and will therefore evolve based on the initial temperature and the cooling scheme within the model. Due to this relation, properly initializing the initial and final temperatures, as well as the cooling scheme aids the model in convergence on a global optimum. The initial temperature should allow for the acceptance of an incorrect solution, to move out of local minima, and the cooling rate should be such that the acceptance probability approaches zero at the final temperature. The cooling scheme can be defined in terms of a terminating condition, such

as the difference metric's convergence, but this is unpredictable in cost and runtime. To make the simulation more predictable, a short initial simulation is performed, tracking the changes in divergence, the minimum and maximum values are then used in Equation 54 to calculate the initial and final temperatures of the simulation where the initial and final probabilities are 0.9 and 0.1 respectively. Cooling is defined by the evolution of temperature over time, incrementing by a factor of α , which is calculated from the initial and final temperature states, where n is the number of optimization cycles.

$$T_{i+1} = \alpha T_i \quad (55)$$

$$\alpha = \exp\left(\frac{\log(T_f) - \log(T_i)}{n}\right) \quad (56)$$

At each decrease in temperature, the number of iterations at the new temperature should be sufficient for the system to reach equilibrium, which can be determined by the Markov chain length at a given temperature. A Markov chain in this case refers to the number of iterations at a set temperature. The chain length is set as a function of neighboring solutions and the maximum chain length and can be thought of as the maximum number of rejected solutions. Incrementation of the chain length in such a way improves convergence as initial acceptance is almost guaranteed.

Due to the stochastic nature of the algorithm, it is beneficial to run multiple smaller size optimizations and combine the results thereby increasing the amount of solution space covered and validating the global minimum result. These additional simulations can be run in parallel and have a negligible effect on the annealing runtime, in fact the cost of such parallel execution is likely to be more efficient than running a single large simulation with a slow cooling scheme.

More general approaches to probing optimal solutions using sample space searches are available, such as the genetic algorithm, taking inspiration from biological evolution processes, where neighbouring solutions compete to minimize the difference metric, with less promising "genetics" pruned at each progressive cycle. The genetic algorithm is a global search algorithm rather than a local search, making it more suited for solution spaces with multiple local minima. Moreover, in comparison to the Markov chains used in the simulated annealing algorithm, the population analysis steps can all be run in parallel, as the evolution process occurs on individuals in the populations. Therefore, the genetic algorithm can be parallelized such that each processor handles the evolution of a subset of the population providing better scalability than the simulated annealing search. The Genetic algorithm has previously been used to select descriptive, i.e. most representative features, from infrared spectroscopy data,⁸⁸ suggesting its applicability to the extraction of representative ensembles for the simulation of UV-vis spectra.

2.6.4 Bandwidth selection

Statistical methods for the determination of the broadening factor δ have been presented in section 5.2, however a more robust method is required to improve its applicability to more molecules. One such method, proposed by Srsen et al. is to perform a cross-validation over potential bandwidth values for each bin, using the bandwidth estimated by Kish's effective sample size as a starting point.⁷⁹ Further, multiple iterations of cross-validation can be performed over a dynamic potential bandwidth range, adapted after each iteration to fully explore the bandwidth solution space. The dynamic shifting of potential bandwidths over each iteration ensures that solutions outside of the initial guess range can be considered, using Silverman's rule as only a starting point. Moreover, optimizing over the Kullback-Leibler divergence puts emphasis on distribution tails which are often

of importance in spectroscopy.

The cross-validation space can be searched by checking each solution in the space, or solutions can be randomly searched, considering only a subset of solutions that cover a wide range of possibilities. While random search cross validation is much less computationally intensive, it can yield inaccurate results if too little of the space is explored. As bandwidth is the only ‘tunable’ or undefined parameter in the nuclear ensemble method, and is sensitive to slight changes, it is appropriate to explore the full solution space and converge on the best value. Such an adaptive technique improves the accuracy of automatic bandwidth determination. While recent developments with the aim to eliminate the step entirely using machine learning algorithms such as kernel ridge regression and gaussian mixture models have been proposed, in this work statistical determination of the broadening factor by cross validation is used.

2.7 Data availability

The nuclear ensemble method requires an ensemble of nuclear geometries sampled from the ground state nuclear density. This can be achieved either using *ab initio* molecular dynamics (AIMD) simulations or by calculating Wigner functions of harmonic wavefunctions. Typically, at least thousands of geometries are required to obtain a suitably converged spectrum, therefore the costly molecular dynamics simulations can be infeasible for larger molecules, or those requiring more involved and accurate *ab initio* methods. Wigner sampling, on the other hand, is a cost-effective sampling method, which naturally describes the quantum delocalization of nuclei as well as the zero-point energy (ZPE) from vibrational effects. Wigner sampling can be implemented using the Newton-X package, developed for the simulation of electronically excited molecules and molecular assemblies, which has integrated functionality for the calculation of initial conditions for excited state calculations.

With the increasing reliance on computational models for the prediction of molecular quantum mechanical properties, the requirement for high quality datasets of such target properties is growing, fueled by the need for training and validation of newly developed models, as well as benchmarking and testing improvements on previously used methods. As such, there has been a rapid development of databases containing QM properties for a wide range of organic molecules with increasing complexity. For instance, the MD17 database, containing 10 individual datasets with QM properties obtained from AIMD simulations at 500K, is commonly applied to develop and validate increasingly complex machine learning potentials.⁸⁹ Since the MD17 dataset was developed using AIMD simulations, which commonly fail to describe the ZPE, they suffer when applied to investigate quantum effects which require a broader sampling of the PES.⁹⁰

Various approaches have been taken to develop more conformationally diverse datasets, thus better representing the ground-state nuclear density. Notably, the WS-22 dataset was developed with the aim of complementing existing databases by providing an extended set of QM properties, rather than being limited to excitation energies and forces, as well as providing a broad and robust representation of potential energy surfaces with high precision.⁹¹ The dataset was created using Wigner sampling, followed by geometric interpolation applied directly to cartesian coordinates, obtaining a geodesic curve by minimizing the least squares distance between two input geometries. As a result, molecular structures further from the equilibrium geometry, closer to regions of interest in photochemistry such as transition states, are included. The WS-22 dataset contains data for several volatile organic compounds with increasing molecular complexities, including Acrolein, Toluene and 2-nitrophenol.

3 Aims of this thesis

This thesis aims to implement optimal sampling of a representative nuclear ensemble using simulated annealing into Atmospec as a fast way to explore the photoabsorption cross-sections of gas-phase volatile organic compounds. An automated nuclear ensemble approach script will be created as a proof of concept, adding representative sampling as an additional step in the workflow. The performance of the script will be evaluated by application to two volatile organic compounds, ensuring its transferability and versatility as an automated workflow. This naturally leads to the implementation of the required additional steps into the Atmospec application, saving computational resources, which can be redirected for the use of more accurate quantum chemistry methods, the investigation of larger, more complex molecules, or more frequent calculations.

4 Computational Details

Figure 4.1 shows a comparison between the nuclear ensemble approach workflow and the respective workflow for representative sampling. The representative sampling approach acts as an additional step where rather than calculating the spectrum from the initial quantum chemistry calculations an optimization of the nuclear ensemble is performed to reduce the number of geometries.

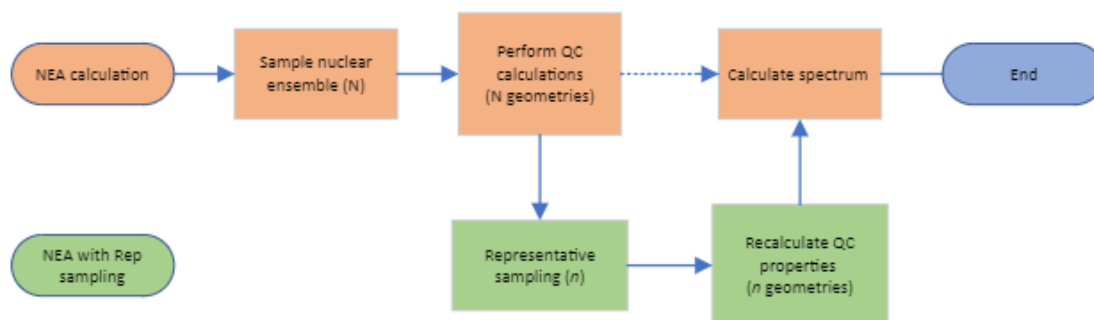


Figure. 4.1. Flowchart visualising the representative sampling workflow alongside the current workflow

4.1 Data collection and processing

Geometry data were extracted for two volatile organic compounds, Acrolein and Toluene, from the ws-22 database. The database contains 120k geometries for each molecule, obtained from equilibrium geometries optimized with density functional theory (DFT) using the hybrid density functional PBE0 and the 6-311 G* basis set without symmetry constraints. Two structures of acrolein, the *cis* and *trans* isomers, were optimized and sampled geometries are equally distributed between the two conformers, for toluene only a single ground state structure was used. 100k geometries are sampled using Wigner sampling, with the remaining 20k obtained using geometry interpolation. Nuclear ensembles of 1200 geometries for use in spectral calculations were created by taking a random subset of the dataset for both acrolein and toluene.

4.2 Representative sampling and QC recalculation

The ORCA quantum chemistry software package, version 5.0.3, was used to perform all quantum calculations for this study.⁶⁶ Calculations were executed using a single Lenovo nx360 m5 compute node on the Blue-

Crystal4 (BC4) cluster, with two 14 core 2.4 GHz Intel E5-2680 v4 CPUs and 129GiB RAM. The SLURM workload manager was employed for workflow management and submission.⁶⁷ Single point calculations for the exploratory sample were carried out using the semi-empirical ZIndo/S method for each structure, obtaining excitation energies and transition dipole moments. Following the creation of an exploratory density of transition properties, a random subset of the ensemble was taken, and optimized using representative sampling over the excitation energies and transition dipole moments, as described in section 6. Parallel instances of the algorithm were launched on individual cores of a full BC4 compute node totaling 28 repetitions, each with 1500 simulated annealing cycles. The Kullback-Leibler divergence was calculated at each cycle as a measure of difference between samples, and points were weighted by intensity to put emphasis on spectrally significant peaks. The representative sample was taken as that with the lowest divergence from the original sample, and a density plot comparing the original and final transition properties was recorded. The representative geometries were then extracted from the original ensemble, and their transition properties were recalculated using linear-response time-dependent DFT with the Tamm-Dancoff approximation.

4.3 Spectral calculations

Photoabsorption cross-sections were calculated using the nuclear ensemble approach, according to Equation 40. Kernel density estimation with a gaussian line shape was used and the phenomenological broadening factor was set using dynamic k-fold cross-validation and grid search. Whereby kernel density estimation models were fit for each bandwidth value and the log-likelihood was maximized to find the optimal broadening value. If the broadening factor chosen was one of the upper or lower bounds, additional grid searches in the direction of the respective bound were performed to find the optimal bandwidth. Error bars for the spectral intensities were estimated using bootstrapping. Resampling the original transition probabilities, recalculating the spectral intensities, and using the variability in these intensities to estimate the error.

5 Results and Discussion

5.1 Implementation overview

The representative sampling algorithm has been implemented alongside the nuclear ensemble approach, producing an automated workflow where calculations can be submitted in batches to explore the optimal sample size, an overview of a single run can be seen in Figure 5.1. The workflow was designed with Atmospec compatibility in mind, with ORCA used in quantum chemistry calculations and minimizing the additional inputs arising from the representative sampling steps. Verbose outputs are available throughout the processing if requested, providing detailed information for debugging and tracking the progress of submissions. The implementation consists of a single workhorse script, which takes a set of initial parameters and manages the workflow, creating all necessary directories for file storage and directing the outputs of separate steps. SLURM is used for the management of resources, allowing for the automated submission of ORCA calculations throughout the workflow. It is recommended that the script be ran on a high powered computing (HPC) cluster due to the computational intensity of quantum chemistry calculations, and simulated annealing in representative sampling.

Table 5.1 shows the initial conditions required for the representative sampling workflow, the notable additions being the number of cycles and repetitions of the simulated annealing algorithm, as well as the reduced ensemble settings. In order to see an improvement in calculation speed, the exploratory method should be set such that it is much faster than the final method, otherwise the reduction in ensemble size is negated by

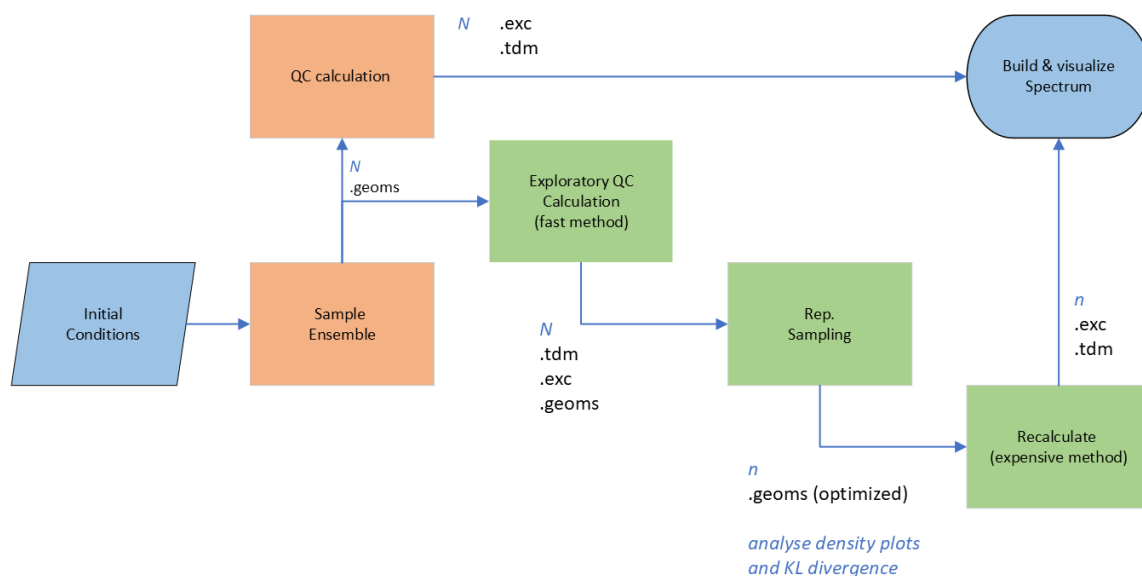


Figure. 5.1. Flowchart of the workflows implemented, NEA (orange), and NEA with Representative sampling (green). Ensemble sizes and data outputs are annotated.

Table. 5.1. Initial conditions for the representative sampling workflow

ORCA options	Ensemble options	Representative sampling	General options
Exploratory method	Number of geometries	Number of cycles	Number of cores
Recalc. method	Reduced ensemble size	Number of repeats	Name
Number of states	Initial ensemble	Weighted/Unweighted	Broadening (σ)

the exploratory calculation and representative sampling run times. The broadening is such that if none is entered, an automated broadening algorithm using grid search cross validation is used, as described in Section 2.6.4. Although some parameters are not necessarily needed to be input by the user, they are included in the configuration file for convenience.

The workflow is designed to accept a large ensemble, either obtained from databases, sampled from a Wigner distribution, or by quantum dynamics simulation. Initially, a sample of this large ensemble is taken as the exploratory ensemble, alternatively, this step can be skipped if your ensemble is already the desired size. ORCA input files are created and submitted for the exploratory sample, then simulated annealing is run over the ensemble transition dipole moments, and excitation energies. A reduced ensemble is produced, which is then used to calculate the photoabsorption cross-section molecular extinction coefficients. The representative ensemble fitness can be evaluated by considering the transition property density plots, alternatively, the Kullback-Leibler divergence can be used to evaluate the information loss of the representative ensemble. A practical application of the representative sampling workflow to two volatile organic compounds, acrolein and toluene, is presented in Sections 5.2 and 5.3 respectively.

5.2 Acrolein

Acrolein is one of the simplest multifunctional compounds present in the atmosphere, a highly reactive β -unsaturated aldehyde, which undergoes a range of oxidation processes. Present in the atmosphere from both direct emissions, and as secondary VOC from the oxidation of 1,3-butadiene, one of the most abundant pollutant olefins. It plays an important role in atmospheric chemistry, spontaneously reacting in absence of sunlight with

ozone and hydroxyl radicals to yield glyoxal and formaldehyde. Under sunlight irradiation, mixtures of acrolein and nitrogen oxide yield NO₂, O₃, and vinyl peroxyacyl nitrate (vinyl-PAN). Acrolein absorbs weakly in the actinic range due to the $n\pi^*$ transition to the first electronically excited state, with a photolysis lifetime over 6 days.

photoabsorption in actinic region

#photodissociation / photodegradation of acrolein

Applicability of nuclear ensemble approach / Wigner sampling – harmonic molecule

The comparison between experimental data, and predicted photoabsorption cross-sections, using both the nuclear ensemble approach and representative sampling, is shown in Figure 5.2. The experimental spectrum shows weak absorption in the actinic range, dominated by absorption to the low-lying $n\pi^*$ S_1 excited state. The bound nature of the $n\pi^*$ excited state is reflected with vibronic progression and a long high energy tail in the spectrum. Considering the calculated cross-sections, the single point spectrum greatly overpredicts the intensity of the absorption, while underpredicting the peak excitation energy, and the absorption range. This is because single point calculations cannot capture the dynamic behavior of molecules, such as vibrations leading to slightly different conformations. On the other hand, a calculation using the nuclear ensemble approach with an ensemble of 1200 geometries, provides a more robust description of the photoabsorption cross-section. Calculated using the B3LYP density functional and the 6-311*G basis set, it provides an accurate representation, capturing both the intensity and range of the cross-section. Vibronic bands are absent by construction of the NEA.

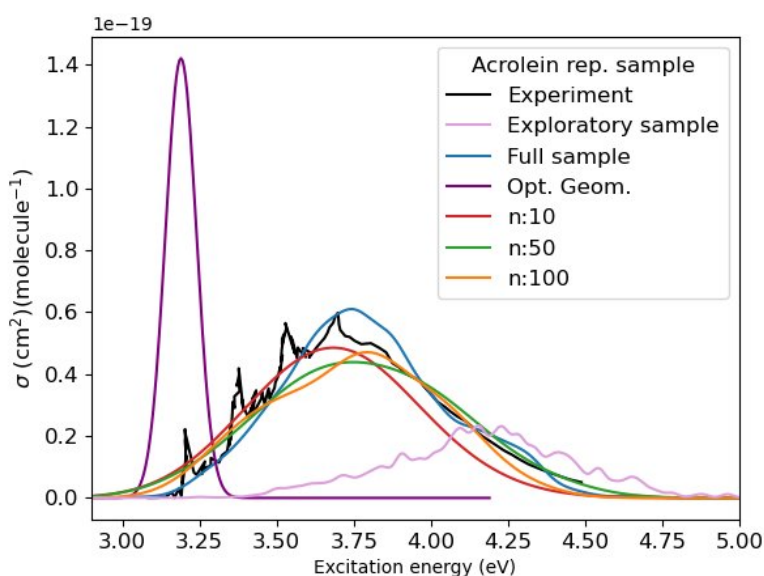


Figure. 5.2. Representative sampling workflow for acrolein using ZIndo/S (*exploratory*), B3LYP/6-311*G (*recalculation*) compared to single point calculation at the minimum energy geometry (*purple*), and the full sample (*blue*) calculated using B3LYP/6-311*G. Experimental values obtained from the MPI/Mainz UV-vis spectral atlas^{8,9}

Using the representative sampling method, the initial exploratory spectrum calculated using ZIndo/S acts as a rapid approximation of the cross-section, upon which the optimization of the ensemble is performed. It shows poor agreement with the experimental spectrum and misses both the intensity and location of the peak, however, a broad correlation between the two lines can be seen. Representative sampling was used to sample (n) geometries from the full ensemble, performing an optimization over the exploratory transition properties to

produce subset “representative” ensembles. The subsequent recalculation of transition properties for a smaller ensemble is seen to provide a massive correction of both excitation energies and transition intensities for a disproportionately low cost. While ensembles as low as 10 provide reasonable approximations of the full ensemble, it fails to capture the high energy tail, as well as being slightly red shifted. Ensembles of 100, and 50 show more robust agreement with the experimental spectrum, although still under predicting the peak intensity, they act as a good approximation of the spectrum at a greatly reduced cost. The agreement of the representative sampling spectra with that of the full ensemble show that a profound reduction in the resource intensity of the quantum chemistry calculation step is possible with an optimization of the ensemble. With an ensemble of 100 reflecting a 91.6% reduction in ensemble size, this translates directly to an increase in the overall efficiency of the spectral calculation.

A summary of Kullback-Leibler divergences for acrolein is presented in Table 5.2, with a clear decreasing trend as samples are added to the subset. An exponential decrease is seen in the lower sample sizes, emphasizing the lack of descriptive representation in the sample, with larger subset sizes being less volatile with the addition of new samples. Previous work found that KL divergence values of around 0.01-0.02 were suitable for a descriptive ensemble,⁸⁰ which agrees with the values obtained in the case of acrolein where a subset of 50 geometries captures the peak intensity and the spectrum tails on both ends.

Table. 5.2. Minimum Kullback-Leibler divergence values for acrolein, obtained using 1500 cycles of simulated annealing and 28 parallel repetitions over the ground state density with weighted intensities

<i>Sample size (n)</i>	5	10	20	30	40	50	60
<i>D_{KL}</i>	0.122	0.096	0.050	0.033	0.034	0.016	0.011

The production of a representative ensemble is an interesting prospect, not just from the perspective of efficiency, as it effectively chooses the conformations which are most important in the transition. Moreover, the construction of spectra under the nuclear ensemble approach allows for the analysis of individual conformer contributions to the overall spectrum.

5.3 Toluene

Toluene is the most abundant aromatic hydrocarbon VOC in the atmosphere, used widely in industry as a solvent in paints, adhesives and cleaning agents, as well as the production of more complex organic molecules. The absorption of atmospheric aromatic compounds is attributed to a π excitation on the benzene ring in the region of 270 nm⁹² with a small contribution from the $S_0 \rightarrow S_1$ transition.¹⁰ The photochemistry of toluene is dominated by the dissociation of bonds β to the benzene ring, the photolysis of the methyl ($-CH_3$) bond being secondary with a significantly reduced quantum yield, 86 % and 10.8 % respectively.⁹² Moreover, photochemical oxidation of toluene has been identified as an important contributor to the formation of ozone and secondary organic aerosols in urban areas.⁹³ TDDFT investigations of the $S_1 \leftarrow S_0$ transition have revealed that diabatic corrections are required for accurate photoabsorption intensities,⁹⁴ nonetheless, the same method has been used as for acrolein for the purpose of run time comparison.

Figure 5.3 shows the representative sampling results for toluene, as well as those for a regular NEA workflow. The anharmonic character of toluene is poorly described by the TDDFT approach used for acrolein, resulting in a blue shifted spectrum. Similar features are present when compared to the acrolein spectrum, again the single point calculation is a poor descriptor, overpredicting both the intensity and width of the full sample spectrum. A familiar trend is also seen in the representative sampling results, with increasing accuracy

as geometries are added to the ensemble. Interestingly, the spectrum for an ensemble of 100 geometries seems skewed in comparison to both the 50-geometry ensemble and the full ensemble. This is attributed to the inclusion of geometries responsible for the two peaks at 4.8 and 4.95 eV, which have been broadened automatically using the cross validation broadening approach. In contrast, the full spectrum has been broadened using an arbitrary broadening factor, $\delta = 0.05$, and therefore the respective low intensity bands that look like vibronic progression, are a result of an insufficiently broadened spectrum, or alternatively insufficient ensemble size to describe the intensity in the region.

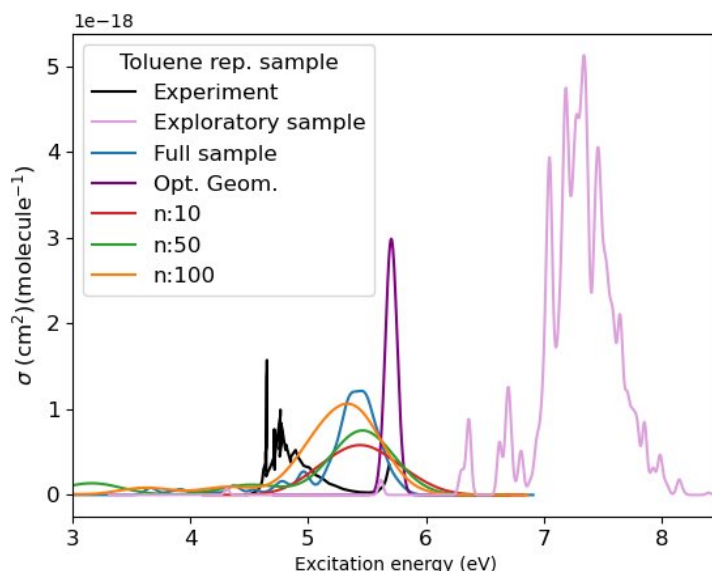


Figure 5.3. Representative sampling workflow for toluene using ZIndo/S (*exploratory*), B3LYP/6-311*G (*recalculation*) compared to single point calculation at the minimum energy geometry (*purple*), and the full sample (*blue*) calculated using B3LYP/6-311*G. Experimental values obtained from the MPI/Mainz UV-vis database.^{8,10}

Table 5.3. Minimum Kullback-Leibler divergence values for toluene, obtained using 1500 cycles of simulated annealing and 28 parallel repetitions over the ground state density with weighted intensities

<i>Sample size (n)</i>	5	10	20	30	40	50	60
D_{KL}	0.192	0.097	0.055	0.045	0.032	0.031	0.020

Like for acrolein, the divergence values were calculated at each subset size to give a quantitative evaluation of the subset representation, presented in Table 5.3. Similar trends are seen, with an exponential initial decrease, followed by smoother decreases as the subset size increases and the impact of additions diminishes. Generally, divergence values are higher than those for acrolein, however they are within acceptable bounds as the subset size reaches 60, which is still a profound decrease in sample size.

5.4 Equidistant sampling

Due to the cost of representative sampling, it is important to justify its use in comparison to more simple sampling methods, for example, the transition properties obtained from the initial nuclear ensemble calculation could be sampled equidistantly to reduce the number of geometries. Figure 5.4 shows a comparison between the ground state densities of the full exploratory sample, the representative sample, and an equidistantly sampled

ensemble of 10 geometries. For both acrolein and toluene, the representative sampling method outperforms the equidistant sampling. Considering acrolein, the equidistant sampling overpredicts the transition intensity, and doesn't capture the width of the peak, additionally, the geometries selected have relatively low transition dipole moments, suggesting they do not contribute significantly to the spectrum. Conversely, the representative sampling density provides an accurate spread of excitation energies, capturing the tails of the distribution well. Similarly, it better predicts the spectral intensity, with more spectrally significant points ($|\mu|^2 > 0$) in the reduced ensemble in comparison to the equidistant sampling method.

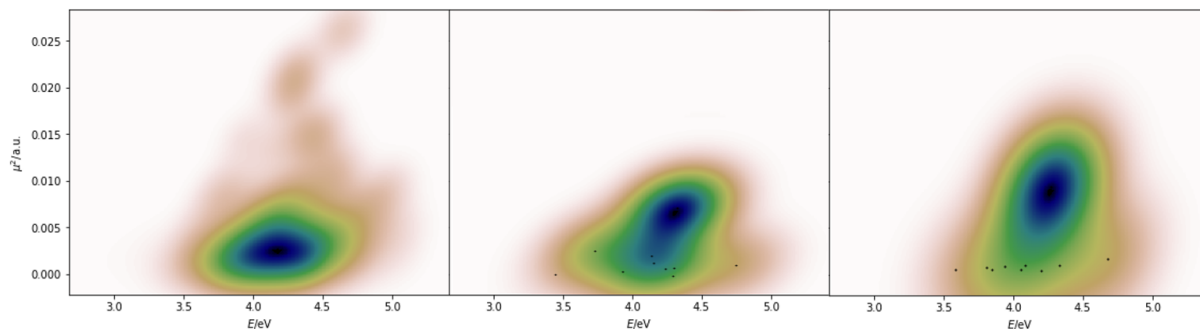


Figure. 5.4. Weighted ground state densities of acrolein in the space of transition properties calculated using ZIndo/S for the full sample (*left*), subsets of 10 geometries sampled using representative and equidistant sampling, (*middle*) and (*right*) respectively

The representative ensemble density for toluene, see Figure 5.5, also seems to be a good fit to the original, although the tails of the distribution are less pronounced geometries at both ends of the energy range have been selected. While the density is a good representation, most of the chosen geometries have dipole transition moments close to zero, skewed by a single point, suggesting that the ensemble is subpar which is reflected in a KL divergence of 0.1. The equidistant sampling method, again, fails to provide a reliable description of the original density. Although more spectrally important points have been chosen, they do not form a better representation, with a clear skew in the peak location. The small size of the ensemble means that a single point can skew the resulting spectrum, such indiscretions are avoided with an optimization of the ensemble. Overall, representative sampling outperforms the equidistant sampling method, with little additional cost providing a great improvement in representation of the original density for small ensemble sizes. Moreso, the quality of reduced ensembles can be assessed through the evaluation of density plots showing the resulting transition properties. This means that the cost of a preliminary representative sampling workflow, to investigate the number of geometries required for a representative ensemble, is dominated by a cheap and fast QM method step for small ensemble sizes.

5.5 Evaluating efficiency

The use of a fast and cheap method for the exploratory calculation of transition properties, such as the semi-empirical ZIndo/S, or alternatively TDDFT with a small basis set, allows for the use of a more complex QM method for the subsequent recalculation of the smaller ensemble at little to no additional cost in comparison to a workflow which only uses NEA. Table 1 shows the time taken for the representative sampling workflows presented in figures (# and #) compared to an NEA submission. Considering the single point calculation for both acrolein and toluene, the use of a single optimal geometry fails to describe the spectra in both cases, making it an unviable choice for the prediction of absorption cross-sections despite its speed. Within the representative sampling workflow, the additional cost of the exploratory QC calculation is minimal when compared to the

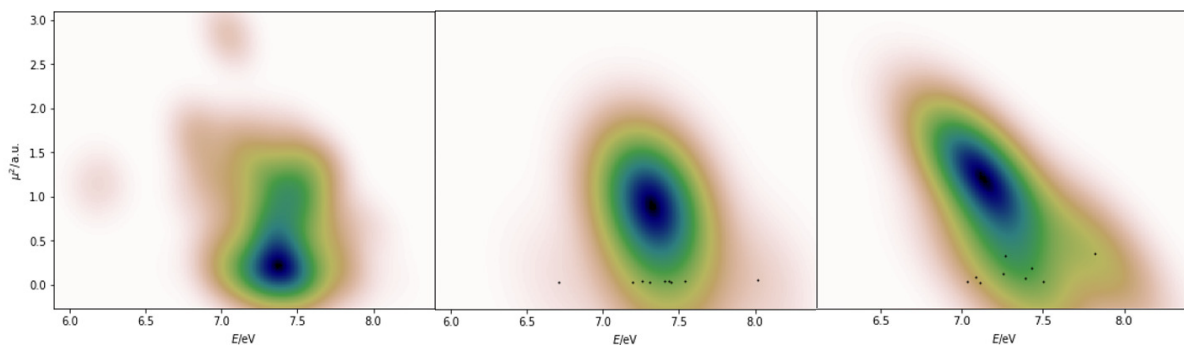


Figure. 5.5. Weighted ground state densities of toluene in the space of transition properties calculated using ZIndo/S for the full sample (*left*), subsets of 10 geometries sampled using representative and equidistant sampling, (*middle*) and (*right*) respectively

time saved in the recalculation step using the representative sampling workflow. For exploratory ensembles of the same size, considering the same number of states, the cost of representative sampling is the same because it is an optimization over transition properties and does not scale with molecular degrees of freedom. This is observed in more pronounced savings for larger molecules, where the cost of QC calculations is increased due to their increasing complexity, with molecules requiring more complex QC methods following a similar trend.

Table. 5.4. A comparison of wall times of individual steps in the representative sampling workflow, varying the number of samples taken

Time taken (s)	Full sample	Single point	Exploratory sample	Rep. sample			Recalculation			Relative speed		
N	1200	1	1200	10	50	100	10	50	100	10	50	100
Acrolein	897	21	108	11	80	264	25	44	87	0.16	0.26	0.52
Toluene	1932	88	159	12	88	265	81	138	265	0.13	0.20	0.36

The overall representative sampling workflow, even using a subset of 100 geometries, provides a generous speed-up when compared to the nuclear ensemble approach using the full ensemble. With the combination of ZIndo/S and TDDFT using B3LYP/6-311*G the calculation workflow for acrolein has been reduced by 6 times, similarly, for toluene a decrease by almost an order of magnitude is seen. Considering the use *ab initio* electronic structure methods for the exploratory calculation and subsequent recalculation, large basis sets containing polarization or diffuse functions are required for an accurate calculation of transition properties. The size of such basis sets greatly impacts the structure method's cost, with common methods scaling poorly, such as ADC(3) scaling as $O(n^6)$ with the basis set size. This unfavorable scaling makes representative sampling particularly attractive for high-level structure methods, where the relative speed-up is increased due to the computational intensity of such methods when a large basis set is used.

More complex molecules have been shown to require more geometries to create a descriptive, representative ensemble, with molecules such as the nitrate anion needing 30 geometries,⁸⁰ and toluene needing 60 geometries. In such cases, representative sampling still provides appreciable speed-ups, moreover, the number of geometries required for a representative ensemble can be evaluated by launching multiple representative sampling optimizations on the exploratory sample and observing the resulting densities. In this way, an optimal number of geometries can be found by either evaluating the density fit, or the KL divergences of samples, at little extra cost after calculating the exploratory sample properties. Additionally, the multiple representative sampling runs are independent of each other, meaning they can be run in parallel, and stopped once a suitable sample size has been found.

5.6 Proposed integration with Atmospec workflow

This work proposes the integration of representative sampling as an option in the Atmospec calculation workflow, a more efficient tool for the approximation of absorption cross-sections. Such an implementation would allow for theoretical investigation of more complex molecules by reducing the overall resource intensity of NEA workflows, speeding up calculations and freeing up resources. Representative sampling has the potential to greatly increase Atmospec’s use case for the rapid calculation of photoabsorption cross-sections for volatile organic compounds, as it provides accurate spectra at a reduced cost. The representative sampling workflow would be offered as an optional method, with results visualized in a separate analysis widget; as seen for the photolysis rate constant calculation and the energy vs oscillator strength modules. Figure 5.6 shows the changes that would have to be made for the implementation of representative sampling into the current Atmospec workchain.

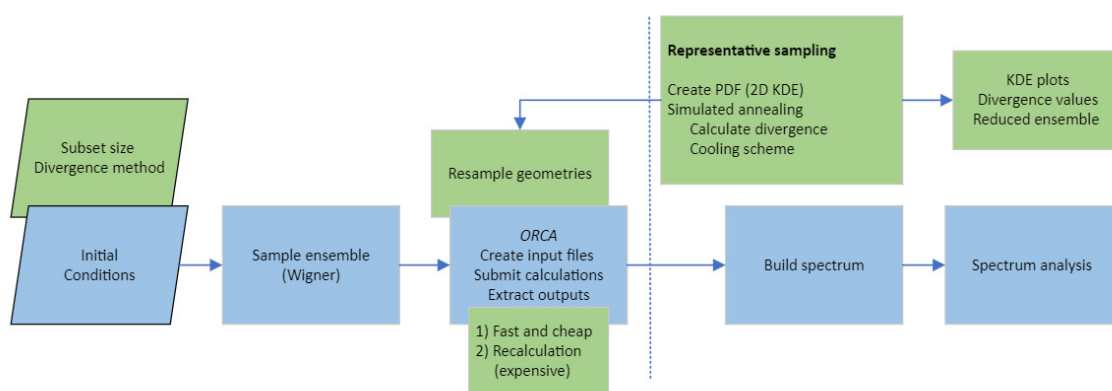


Figure. 5.6. Summary of representative sampling functionality alongside current Atmospec workchain

The representative sampling algorithm integration would begin with the calculation submission step, where the rep. sampling parameters would be changed if needed, with recommended default values being automatically filled. The representative sampling step should be optional; therefore, a toggle could be implemented to ensure that representative sampling options are not selected when it is not required. The representative sampling algorithm has multiple hyperparameters, such as the number of annealing cycles, the number of repetitions, the reduced sample size, as well as options for the optimization using spectral weighting, and the calculation of errors. Such options should be limited, to reduce the complexity for an unfamiliar user. For example, spectral weighting should always be enabled as it provides better estimations of the original density, as such, it is not something the user needs to be aware of. The simplest implementation would include the launch of a single representative sampling workflow; however, the problem arises that the optimal subset size is unknown. As a workaround, the user could be asked whether they want to optimize the subset size, which is possible by launching multiple representative sampling repetitions for a range of sample sizes.

Following a successful submission, the calculation status section should reflect the representative sampling workflow, showing the additional steps such as the exploratory sampling and calculation, the representative sampling and finally, the spectrum recalculation.

A workflow submitted using representative sampling would output the final recalculated spectrum, accompanied by the subset and full sample densities to allow for evaluation of the representative sampling workflow. A summary of the representative sampling workflow should be available for the user, including detailed informa-

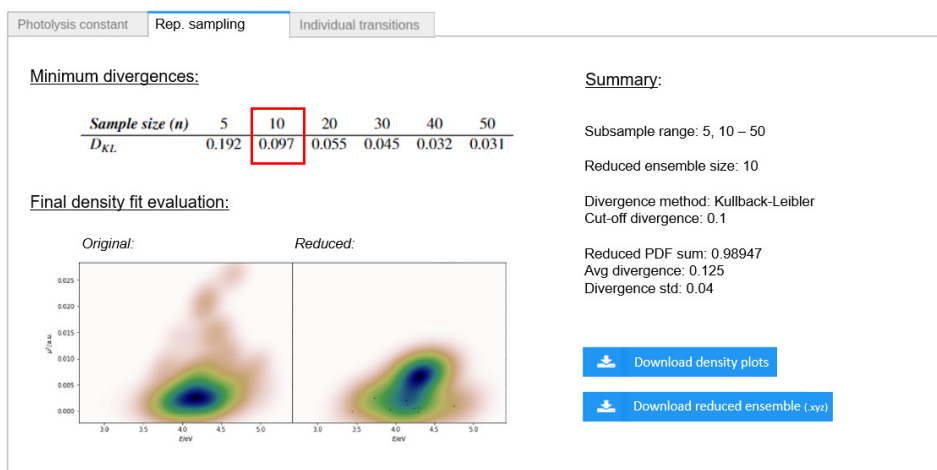


Figure. 5.7. Proposed widget implementation for a submission over a range of ensemble sizes

tion about the divergence values and subsets that have been explored, encouraging investigation of suboptimal representations. A copy of the representative subset geometries and values for the transition properties (μ , ΔE) should be available to allow for the reproduction of spectra. Figure 5.7 shows a graphical depiction of a proposed Atmospec representative sampling analysis widget, containing a pair of density plots, for comparison between the representative sample and the full ensemble, as well as information about the representative ensemble workflows that have been performed. The representative sample tab could also contain a 3D image of the most representative geometry, obtained using simulated annealing as a subset of size 1, as well as a downloadable .xyz file containing the output representative ensemble, and the divergence values for each ensemble that has been calculated.

6 Conclusion

Representative sampling with the nuclear ensemble approach has been implemented into an automated workflow, combining the multiple steps into a single submission. This workflow was applied to two volatile organic compounds of varying complexity, acrolein and toluene, demonstrating the method’s applicability in the efficient calculation of photoabsorption cross sections. The representative sampling approach yielded a reasonable estimation of the spectrum for acrolein with an ensemble as small as 10 geometries. While toluene showed higher D_{KL} values, the sample could be reduced to 50 geometries, resulting in an 80% decrease in runtime and using just 0.05% of the original ensemble. These results are extremely positive, combined with recent published work, where subsets up to 30 geometries were found to give a representative description of complex VOCs like azobenzene, and a Creigee intermediate, they highlight representative sampling as an invaluable tool for increasing the efficiency of *in silico* calculations photoabsorption cross-sections.

By optimizing over transition dipole moments and excitation energies of each configuration, the simulated annealing algorithm scales with the size of the initial ensemble and the size of the reduced ensemble, rather than molecular size, making savings more pronounced for larger molecules. When coupled with the unfavourable scaling of quantum mechanical methods with system size, representative sampling provides a favourable alternative for larger molecules. An evaluation of the divergence is possible at the cost of a cheap semi-empirical or QM method, the latter with a reduced basis set, making the exploration of optimal subset size viable for smaller subset sizes. Moreover, the geometries forming the representative sample can provide insight into geometrical properties that are important in an electronic transition. The use of representative sampling with the nuclear ensemble approach has great potential to increase its applicability to larger and more complex molecules and it would be an essential tool for atmospheric chemists Atmospec.

The preliminary implementation of representative sampling into an automated workflow provides an initial framework for integration as an optional workflow in Atmospec. A method of implementation has been discussed, outlining major changes that need to be made to the Atmospec workchain for the integration of representative sampling. The majority of the functionality required, such as plotting ground state densities, launching ORCA calculations, and the calculation of spectra, are already present,

7 Outlook

7.1 Algorithmic improvements

While simulated annealing provides satisfactory results for finding a global optimum subset, it is a stochastic algorithm, meaning that it does not necessarily provide the actual global minimum. To mitigate this, multiple simulated annealing (SA) runs can be initiated, and the best overall solution can be taken. Although the additional SA iterations can be run in parallel, they use additional resources and parallelization adds a runtime overhead. Additionally, the Markov Chain Monte Carlo (MCMC) approach used within the simulated annealing algorithm requires many sequential model evaluations which can become computationally prohibitive. Considering increasing the efficiency of the representative sampling approach, alternative global optimization algorithms are available which could improve the optimization cost and solution quality. For example, the genetic algorithm is an approach that takes inspiration from natural selection observed in nature and considers populations of solutions rather than the single solution considered in simulated annealing, thereby providing for a broader search of the solution space.⁹⁵

Applied to the sampling of an optimal subset of nuclear geometries, a population of randomly selected subsets would be created, then like in simulated annealing, the divergences between the subsets and the full samples are calculated giving a measure of each individual's fitness. After fitness is determined, the predecessors of the following generation are chosen in pairs with a probability weighted by their fitness, meaning more fit individuals will be more likely to influence the following generation. A crossover operation is performed between each pair, exchanging segments of “genetic material” (subset samples) from the parents to create two new offspring. Random changes are introduced in a mutation step, where small changes to the solutions are made to help explore new solutions and avoid local minima. Finally, a new generation is formed from a mixture of a few fittest individuals from the previous generation and those formed with crossover and mutation. The optimization terminates upon reaching a maximum number of iterations, or when a terminating criterion, such as a minimum divergence, has been met. A flowchart of a typical genetic algorithm can be seen in Figure ??.



Figure. 7.1. General framework for the genetic algorithm meta-heuristic

The genetic algorithm is favorable over the simulated annealing approach due to the consideration of a population, rather than attempting to optimize a single solution in the search space multiple solutions are proposed and the population is optimized. Moreover, the genetic algorithm is inherently parallelizable and lends itself to distributed computing paradigms, encouraging scalability.⁹⁶ The implementation of the genetic algorithm would involve optimization by tuning the ratios of the steps involved, for example the rate of genetic crossovers or the number of mutations introduced in each generation

The high tunability of the genetic algorithm approach gives it the flexibility to find a balance between exploring the solution space and converging on an optimal solution. The convergence on a solution can be measured using a fitness threshold, although this can inevitably lead to long runtimes in cases where even the optimal fitness is unsatisfactory. A better convergence measure is to look at the stability of the solution

over time, if the fitness is not improving it is likely to have reached a plateau or converged on the optimal solution. The rate of convergence is determined by the optimization of GA parameters, with convergence onto a sub-optimal solution occurring when not enough diversity is introduced into the population. To conclude, the genetic algorithm could be used as a search algorithm to find a global optimum, “representative” sample of the total population, improving on simulated annealing by the consideration of a population of samples rather than optimizing a random initial subset.

7.2 Combinatory approaches

The development of machine learning approaches for the nuclear ensemble approach, including the prediction of quantum chemistry properties and the reconstruction of spectra, is exciting when considering the possibility for compound methods to increase the efficiency of calculations. Such approaches could complement each other to combat limitations seen in the standalone methods, for example, representative sampling produces a representative subset of nuclear geometries, but the small subset impacts the error in the spectrum calculation. In this case, machine learning approaches, such as kernel ridge regression or GMM, can be used to simulate additional values for excitation energies and transition properties, effectively eliminating the broadening parameter and reducing the increased error in spectra calculation due to the small sample size. Gaussian mixture models stand as the favourable solution, as they show improved performance in comparison to KREG and statistical estimators of bandwidth such as those based on Silverman’s rule of thumb.⁷³ Moreover, they eliminate the need for bandwidth altogether, bringing us closer to the fully automated calculation of photoabsorption cross-sections.

References

- [1]
- [2] Giacomo Insero, Franco Fusi, and Giovanni Romano. The safe use of lasers in biomedicine: Principles of laser-matter interaction. *Journal of Public Health Research*, 12(3):22799036231187077, 2023.
- [3] Ra Saunders, Nicola Carslaw, Stephen Pascoe, Michael Pilling, Michael Jenkin, and Richard Derwent. Development of the master chemical mechanism (mcmv2.0) web site and recent applications of its use in tropospheric chemistry models. 12 1999.
- [4] Mark E. Casida, Bhaarithi Natarajan, and Thierry Deutsch. *Non-Born–Oppenheimer Dynamics and Conical Intersections*, page 279–299. Springer Berlin Heidelberg, 2012.
- [5] Reinhard Schinke. *Photodissociation Dynamics: Spectroscopy and fragmentation of small polyatomic molecules*. Cambridge Univ. Press, 1995.
- [6] Antonio Prlj, Emanuele Marsili, Lewis Hutton, Daniel Hollas, Darya Shchepanovska, David R. Glowacki, Petr Slaviček, and Basile F. Curchod. Calculating photoabsorption cross-sections for atmospheric volatile organic compounds. *ACS Earth and Space Chemistry*, 6(1):207–217, Dec 2021.
- [7] InSilicoPhotochemistry group and Daniel Hollas. Atmospec: Ab initio uv/vis spectroscopy for everyone, May 2023.
- [8] H. Keller-Rudek, G. K. Moortgat, R. Sander, and R. Sörensen. The mpi-mainz uv/vis spectral atlas of gaseous molecules of atmospheric interest. *Earth System Science Data*, 5(2):365–373, Dec 2013.
- [9] I. Magneron, R. Thévenet, A. Mellouki, G. Le Bras, G. K. Moortgat, and K. Wirtz. A study of the photolysis and oh-initiated oxidation of acrolein and trans-crotonaldehyde. *The Journal of Physical Chemistry A*, 106(11):2526–2537, Feb 2002.
- [10] C. Serralheiro, D. Duflot, F. Ferreira da Silva, S. V. Hoffmann, N. C. Jones, N. J. Mason, B. Mendes, and P. Limão-Vieira. Toluene valence and rydberg excitations as studied by ab initio calculations and vacuum ultraviolet (vuv) synchrotron radiation. *The Journal of Physical Chemistry A*, 119(34):9059–9069, Aug 2015.
- [11] Roger Atkinson and Janet Arey. Atmospheric degradation of volatile organic compounds. *Chemical Reviews*, 103(12):4605–4638, Oct 2003.
- [12] Wen-Tien Tsai. Toxic volatile organic compounds (vocs) in the atmospheric environment: Regulatory aspects and monitoring in japan and korea. *Environments*, 3(4):23, Sep 2016.
- [13] K Rumchev. Association of domestic exposure to volatile organic compounds with asthma in young children. *Thorax*, 59(9):746–751, Sep 2004.
- [14] Michael L. Boeglin, Denise Wessels, and Diane Henshel. An investigation of the relationship between air emissions of volatile organic compounds and the incidence of cancer in indiana counties. *Environmental Research*, 100(2):242–254, Feb 2006.
- [15] Xihe Zhou, Xiang Zhou, Chengming Wang, and Handong Zhou. Environmental and human health impacts of volatile organic compounds: A perspective review. *Chemosphere*, 313:137489, 2023.

- [16] Elena David and Violeta-Carolina Niculescu. Volatile organic compounds (vocs) as environmental pollutants: Occurrence and mitigation using nanomaterials. *International Journal of Environmental Research and Public Health*, 18(24), 2021.
- [17] John V. Constable, Alex B. Guenther, David S. Schimel, and Russell K. Monson. Modelling changes in voc emission in response to climate change in the continental united states. *Global Change Biology*, 5(7):791–806, Oct 1999.
- [18] Karolina Kuklinska, Lidia Wolska, and Jacek Namiesnik. Air quality policy in the u.s. and the eu – a review. *Atmospheric Pollution Research*, 6(1):129–137, 2015.
- [19] R ATKINSON. Gas-phase tropospheric chemistry of organic compounds: A review. *Atmospheric Environment*, 41:200–240, 2007.
- [20] Michael E. Jenkin, Sandra M. Saunders, and Michael J. Pilling. The tropospheric degradation of volatile organic compounds: A protocol for mechanism development. *Atmospheric Environment*, 31(1):81–104, Jan 1997.
- [21] M. E. Jenkin, J. C. Young, and A. R. Rickard. The mcm v3.3.1 degradation scheme for isoprene. *Atmospheric Chemistry and Physics*, 15(20):11433–11459, Oct 2015.
- [22] L. Vereecken, B. Aumont, I. Barnes, J.W. Bozzelli, M.J. Goldman, W.H. Green, S. Madronich, M.R. McGillen, A. Mellouki, J.J. Orlando, and et al. Perspective on mechanism development and structure-activity relationships for gas-phase atmospheric chemistry. *International Journal of Chemical Kinetics*, 50(6):435–469, Apr 2018.
- [23] J. Peeters, T. L. Nguyen, and L. Vereecken. Hox radical regeneration in the oxidation of isoprene. *Physical Chemistry Chemical Physics*, 11(28):5935, 2009.
- [24] J. Lelieveld, T. M. Butler, J. N. Crowley, T. J. Dillon, H. Fischer, L. Ganzeveld, H. Harder, M. G. Lawrence, M. Martinez, D. Taraborrelli, and et al. Atmospheric oxidation capacity sustained by a tropical forest. *Nature*, 452(7188):737–740, Apr 2008.
- [25] Andrew R Rickard. Development, applications and strategic future for detailed chemical mechanisms, 2020.
- [26] Albert Einstein. The collected papers of albert einstein; volume 6 the berlin years: Writings 1914 - 1917. *European Journal of Physics*, 18(1), Jan 1997.
- [27] A. J. Allmand. Part i.—einstein’s law of photochemical equivalence. *Trans. Faraday Soc.*, 21(February):438–452, 1926.
- [28] Brian Wardle. *Principles and applications of Photochemistry*. Wiley, 2009.
- [29] Lionel Goodmans. Theory and applications of ultraviolet spectroscopy. *Journal of the American Chemical Society*, 85(24):4056–4057, 1963.
- [30] D. F. Swinehart. The beer-lambert law. *Journal of Chemical Education*, 39(7):333, Jul 1962.
- [31] J. A. Barltrop and J. D. Coyle. *Excited states in Organic Chemistry*. Wiley, 1975.
- [32] Robert C. Hilborn. Einstein coefficients, cross sections, f values, dipole moments, and all that. *American Journal of Physics*, 50(11):982–986, Nov 1982.

- [33] Robert S Mulliken. Intensities of electronic transitions in molecular spectra i. introduction. *The Journal of Chemical Physics*, 7(1):14–20, 1939.
- [34] A. Jablonski. Efficiency of anti-stokes fluorescence in dyes. *Nature*, 131(3319):839–840, Jun 1933.
- [35] Ralph Sherman Becker. Theory and interpretation of fluorescence and phosphorescence. (*No Title*), 1969.
- [36] Michael Kasha. Characterization of electronic transitions in complex molecules. *Discussions of the Faraday Society*, 9:14, 1950.
- [37] J M Zhang and Y Liu. Fermi’s golden rule: Its derivation and breakdown by an ideal model. *European Journal of Physics*, 37(6):065406, Oct 2016.
- [38] P. W. Atkins and Ronald Friedman. *Molecular quantum mechanics / Peter Atkins, Ronald Friedman*. Oxford University Press, 2005.
- [39] Edward U. Condon. Nuclear motions associated with electron transitions in diatomic molecules. *Physical Review*, 32(6):858–872, Dec 1928.
- [40] Heinz Mustroph. Potential-energy surfaces, the born–oppenheimer approximations, and the franck–condon principle: Back to the roots. *ChemPhysChem*, 17(17):2616–2629, Jun 2016.
- [41] Federica Agostini and Basile F. Curchod. Chemistry without the born–oppenheimer approximation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2223), Mar 2022.
- [42] João Pedro Malhado, Michael J. Bearpark, and James T. Hynes. Non-adiabatic dynamics close to conical intersections and the surface hopping perspective. *Frontiers in Chemistry*, 2, Nov 2014.
- [43] Benjamin G. Levine, Michael P. Esch, B. Scott Fales, Dylan T. Hardwick, Wei-Tao Peng, and Yinan Shu. Conical intersections at the nanoscale: Molecular ideas for materials. *Annual Review of Physical Chemistry*, 70(1):21–43, Jun 2019.
- [44] P. W. Atkins, De Julio Paula, and James Keeler. *Atkins’ physical chemistry*. Oxford University Press, 2023.
- [45] Leticia González, Daniel Escudero, and Luis Serrano-Andrés. Progress and challenges in the calculation of electronic excited states. *ChemPhysChem*, 13(1):28–51, Sep 2011.
- [46] Marco Garavelli. Computational organic photochemistry: Strategy, achievements and perspectives. *Theoretical Chemistry Accounts*, 116(1–3):87–105, Jan 2006.
- [47] G.A. Worth and L.S. Cederbaum. Mediation of ultrafast electron transfer in biological systems by conical intersections. *Chemical Physics Letters*, 338(4–6):219–223, Apr 2001.
- [48] Martin King. Ecg environmental briefs.
- [49] E. A. Gislason. Series expansions for franck–condon factors. i. linear potential and the reflection approximation. *The Journal of Chemical Physics*, 58(9):3702–3707, May 1973.
- [50] Eric J Heller. Quantum corrections to classical photodissociation models. *The Journal of Chemical Physics*, 68(5):2066–2075, 1978.

- [51] Soo Y. Lee, Robert C. Brown, and Eric J. Heller. Multidimensional reflection approximation: Application to the photodissociation of polyatomics. *The Journal of Physical Chemistry*, 87(12):2045–2053, Jun 1983.
- [52] Fabrizio Santoro, Roberto Improta, Alessandro Lami, Julien Bloino, and Vincenzo Barone. Effective method to compute franck-condon integrals for optical spectra of large molecules in solution. *The Journal of chemical physics*, 126(8), 2007.
- [53] Fabrizio Santoro and Denis Jacquemin. Going beyond the vertical approximation with time-dependent density functional theory. *WIREs Computational Molecular Science*, 6(5):460–486, 2016.
- [54] Ana Borrego-Sánchez, Madjid Zemmouche, Javier Carmona-García, Antonio Francés-Monerris, Pep Mulet, Isabelle Navizet, and Daniel Roca-Sanjuán. Multiconfigurational quantum chemistry determinations of absorption cross sections () in the gas phase and molar extinction coefficients () in aqueous solution and air–water interface. *Journal of Chemical Theory and Computation*, 17(6):3571–3582, May 2021.
- [55] Rachel Crespo-Otero and Mario Barbatti. Spectrum simulation and decomposition with nuclear ensemble: Formal derivation and application to benzene, furan and 2-phenylfuran. *Theoretical Chemistry Accounts*, 131(6), Jun 2012.
- [56] Mario Barbatti, Adelia J. Aquino, and Hans Lischka. The uv absorption of nucleobases: Semi-classical ab initio spectra simulations. *Physical Chemistry Chemical Physics*, 12(19):4959–4967, Mar 2010.
- [57] Weixuan Zeng, Shaolong Gong, Cheng Zhong, and Chuluo Yang. Prediction of oscillator strength and transition dipole moments with the nuclear ensemble approach for thermally activated delayed fluorescence emitters. *The Journal of Physical Chemistry C*, 123(15):10081–10086, Mar 2019.
- [58] Fabio Della Sala, Roger Rousseau, Andreas Görling, and Dominik Marx. Quantum and thermal fluctuation effects on the photoabsorption spectra of clusters. *Physical Review Letters*, 92(18), May 2004.
- [59] Sršeň, D. Hollas, and P. Slavíček. Uv absorption of criegee intermediates: Quantitative cross sections from high-level ab initio theory. *Physical Chemistry Chemical Physics*, 20(9):6421–6430, 2018.
- [60] Stepan Srsen, Jaroslav Sita, Petr Slavicek, Vít Ladányi, and Dominik Heger. Limits of the nuclear ensemble method for electronic spectra simulations: Temperature dependence of the (e)-azobenzene spectrum. *Journal of Chemical Theory and Computation*, 16(10):6428–6438, 2020.
- [61] Emanuele Marsili, Antonio Prlj, and Basile F. Curchod. A theoretical perspective on the actinic photochemistry of 2-hydroperoxypropanal. *The Journal of Physical Chemistry A*, 126(32):5420–5433, Jul 2022.
- [62] William B. Case. Wigner functions and Weyl transforms for pedestrians. *American Journal of Physics*, 76(10):937–946, 10 2008.
- [63] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.
- [64] Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. *Selected Works of Murray Rosenblatt*, pages 95–100, 2011.

- [65] Aliaksandr V. Yakutovich, Kristjan Eimre, Ole Schütt, Leopold Talirz, Carl S. Adorf, Casper W. Andersen, Edward Dittler, Dou Du, Daniele Passerone, Berend Smit, Nicola Marzari, Giovanni Pizzi, and Carlo A. Pignedoli. Aiidalab – an ecosystem for developing, executing, and sharing scientific workflows. *Computational Materials Science*, 188:110165, 2021.
- [66] Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger. The ORCA quantum chemistry program package. *The Journal of Chemical Physics*, 152(22):224108, 06 2020.
- [67] Andy B. Yoo, Morris A. Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn, editors, *Job Scheduling Strategies for Parallel Processing*, pages 44–60, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [68] Mario Barbatti, Mattia Bondanza, Rachel Crespo-Otero, Baptiste Demoulin, Pavlo O. Dral, Giovanni Granucci, Fábris Kossoski, Hans Lischka, Benedetta Mennucci, Saikat Mukherjee, and et al. Newton-x platform: New software developments for surface hopping and nuclear ensembles. *Journal of Chemical Theory and Computation*, 18(11):6851–6865, Oct 2022.
- [69] Fábris Kossoski and Mario Barbatti. Nuclear ensemble approach with importance sampling. *Journal of Chemical Theory and Computation*, 14(6):3173–3183, 2018. PMID: 29694040.
- [70] B.W. Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, Feb 2018.
- [71] Bin Wang and Xiaofeng Wang. Bandwidth selection for weighted kernel density estimation. *arXiv preprint arXiv:0709.1616*, 2007.
- [72] Leslie Kish. *Survey sampling*. Wiley, 1995.
- [73] Luis Cerdán and Daniel Roca-Sanjuán. Reconstruction of nuclear ensemble approach electronic spectra using probabilistic machine learning. *Journal of Chemical Theory and Computation*, 18(5):3052–3064, 2022.
- [74] Péter P Fehér, Ádám Madarász, and András Stirling. Multiscale modeling of electronic spectra including nuclear quantum effects. *Journal of chemical theory and computation*, 17(10):6340–6352, 2021.
- [75] Pavlo O Dral, Fuchun Ge, Bao Xin Xue, Yi-Fan Hou, Max Pinheiro Jr, Jianxing Huang, and Mario Barbatti. Mlatom 2: An integrative platform for atomistic machine learning. *New Horizons in Computational Chemistry Software*, pages 13–53, 2022.
- [76] RDKit: Open-source cheminformatics. <http://www.rdkit.org>. [Online; accessed 11-April-2024].
- [77] Bao-Xin Xue, Mario Barbatti, and Pavlo O. Dral. Machine learning for absorption cross sections. *The Journal of Physical Chemistry A*, 124(35):7199–7210, Aug 2020.
- [78] Sebastian Sitkiewicz, Javier Carmona-Garcia, Luis Cerdán, and Daniel Roca-Sanjuán. Qcexval/multi-spec: Tool for predicting electronic spectra with nuclear ensemble approach., 2023.
- [79] Štěpán Sršeň. *Data science approach to electronic spectroscopy*. PhD thesis, 2021.
- [80] Stepan Srsen and Petr Slavicek. Optimal representation of the nuclear ensemble: Application to electronic spectroscopy. *Journal of Chemical Theory and Computation*, 17(10):6395–6404, 2021.

- [81] Matt P Wand and M Chris Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the american Statistical Association*, 88(422):520–528, 1993.
- [82] Mark Anthony Wolters and Willard John Braun. A practical implementation of weighted kernel density estimation for handling shape constraints. *Stat*, 7(1):e202, 2018.
- [83] James V Stone. *Kullback-Leibler divergence*, page 148–155. Sebtel press, 2015.
- [84] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar 1951.
- [85] Sumin Wang and Fasheng Sun. Deterministic sampling based on kullback–leibler divergence and its applications. *Statistical Papers*, Apr 2023.
- [86] Stuart J. Russell. *Artificial Intelligence: A modern approach*. Pearson, 2016.
- [87] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [88] Fan Fan, Zhou Changwei, Zhang Xiaojun, Wu Di, Tao Zhi, and Xu Yishen. Feature wavelength selection in near-infrared spectroscopy based on genetic algorithm. In *2021 International Conference on Sensing, Measurement Data Analytics in the era of Artificial Intelligence (ICSMD)*, pages 1–5, 2021.
- [89] Joel M. Bowman, Chen Qu, Riccardo Conte, Apurba Nandi, Paul L. Houston, and Qi Yu. The md17 datasets from the perspective of datasets for gas-phase “small” molecule potentials. *The Journal of Chemical Physics*, 156(24), Jun 2022.
- [90] Mario Barbatti and Kakali Sen. Effects of different initial condition samplings on photodynamics and spectrum of pyrrole. *International Journal of Quantum Chemistry*, 116(10):762–771, Nov 2015.
- [91] Max Pinheiro Jr, Shuang Zhang, Pavlo O. Dral, and Mario Barbatti. Ws22 database, wigner sampling and geometry interpolation for configurationally diverse molecular datasets. *Scientific Data*, 10(1), Feb 2023.
- [92] Robert R. Hentz and Milton Burton. Studies in photochemistry and radiation chemistry of toluene, mesitylene and ethylbenzene^{1,2,3}. *Journal of the American Chemical Society*, 73(2):532–536, Feb 1951.
- [93] Yuemeng Ji, Jun Zhao, Hajime Terazono, Kentaro Misawa, Nicholas P. Levitt, Yixin Li, Yun Lin, Jianfei Peng, Yuan Wang, Lian Duan, Bowen Pan, Fang Zhang, Xidan Feng, Taicheng An, Wilmarie Marrero-Ortiz, Jeremiah Secrest, Annie L. Zhang, Kazuhiko Shibuya, Mario J. Molina, and Renyi Zhang. Re-assessing the atmospheric oxidation mechanism of toluene. *Proceedings of the National Academy of Sciences*, 114(31):8169–8174, 2017.
- [94] David Robinson, Saleh S Alarfaji, and Jonathan D Hirst. Benzene, toluene, and monosubstituted derivatives: diabatic nature of the oscillator strengths of $s_1 \leftarrow s_0$ transitions. *The Journal of Physical Chemistry A*, 125(24):5237–5245, 2021.
- [95] T Ghose. Optimization technique and an introduction to genetic algorithms and simulated annealing. In *Proceedings of international workshop on soft computing and systems*, pages 1–19, 2002.
- [96] Erick Cantú-Paz and David E. Goldberg. Efficient parallel genetic algorithms: Theory and practice. *Computer Methods in Applied Mechanics and Engineering*, 186(2–4):221–238, Jun 2000.

- [97] Ahmad Hassanat, Khalid Almohammadi, Esra'a Alkafaween, Eman Abunawas, Awni Hammouri, and V. B. Prasath. Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach. *Information*, 10(12):390, Dec 2019.
- [98] Paul Crutzen. A review of upper atmospheric photochemistry. *Canadian Journal of Chemistry*, 52(8):1569–1581, Apr 1974.
- [99] Marye Anne Fox. Excited states in photochemistry of organic molecules edited by martin klessinger (university of munster) and josef michl (university of colorado). vch: New york. 1995. isbn 1-56081-588-4. *Journal of the American Chemical Society*, 118(7):1815–1816, Jan 1996.
- [100] Christian George, Barbara D'Anna, Hartmut Herrmann, Christian Weller, Veronica Vaida, D. J. Donaldson, Thorsten Bartels-Rausch, and Markus Ammann. Emerging areas in atmospheric photochemistry. *Topics in Current Chemistry*, page 1–53, 2012.
- [101] David Goldberg. Genetic algorithms in search, optimization, and machine learning. *Choice Reviews Online*, 27(02), Oct 1989.
- [102] Zhen Liu, Vinh Son Nguyen, Jeremy Harvey, Jean-François Müller, and Jozef Peeters. The photolysis of -hydroperoxycarbonyls. *Physical Chemistry Chemical Physics*, 20(10):6970–6979, 2018.
- [103] Fernando G Lobo, David E. Goldberg, and Martin Pelikan. Time complexity of genetic algorithms on exponentially scaled problems. *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, Jul 2007.
- [104] Russell K. Monson, Manuel T. Lerdau, Thomas D. Sharkey, David S. Schimel, and Ray Fall. Biological aspects of constructing volatile organic compound emission inventories. *Atmospheric Environment*, 29(21):2989–3002, Nov 1995.
- [105] Colin R Reeves. Genetic algorithms. *Handbook of metaheuristics*, pages 109–139, 2010.
- [106] Krassi Rumchev, Helen Brown, and Jeffery Spickett. Volatile organic compounds: Do they present a risk to our health? *Reviews on Environmental Health*, 22(1), Jan 2007.
- [107] William R. Stockwell, Charlene V. Lawson, Emily Saunders, and Wendy S. Goliff. A review of tropospheric atmospheric chemistry and gas-phase chemical mechanisms for air quality modeling. *Atmosphere*, 3(1):1–32, Dec 2011.
- [108] D. Taraborrelli, M. G. Lawrence, J. N. Crowley, T. J. Dillon, S. Gromov, C. B. Groß, L. Vereecken, and J. Lelieveld. Hydroxyl radical buffered by isoprene oxidation over tropical forests. *Nature Geoscience*, 5(3):190–193, Feb 2012.
- [109] L. K. Xue, T. Wang, H. Guo, D. R. Blake, J. Tang, X. C. Zhang, S. M. Saunders, and W. X. Wang. Sources and photochemistry of volatile organic compounds in the remote atmosphere of western china: Results from the mt. waliguan observatory. *Atmospheric Chemistry and Physics*, 13(17):8551–8567, Sep 2013.
- [110] David R. Yarkony. Nonadiabatic quantum chemistry—past, present, and future. *Chemical Reviews*, 112(1):481–498, Nov 2011.
- [111] Guy Brasseur and Daniel J. Jacob. *Modeling of atmospheric chemistry*. Cambridge University Press, 2017.

- [112] Theodore W Manikas and James T Cain. 1996.
- [113] D. T. Pham and Dervis Karaboga. *Intelligent optimisation techniques: Genetic Algorithms, Tabu Search, simulated annealing and Neural Networks*. Springer-Verlag, 2000.
- [114] J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. Cambridge University Press, 2021.
- [115] George William Series. *Laser spectroscopy and other topics: Selected papers of g.w. series, Raman Professor, 1982-83*. Indian academy of sciences, 1985.
- [116] Azzam Charaf-Eddin, Thomas Cauchy, François-Xavier Felpin, and Denis Jacquemin. Vibronic spectra of organic electronic chromophores. *RSC Adv.*, 4(98):55466–55472, 2014.
- [117] A. Einstein. 7. zur quantentheorie der strahlung. *Quantentheorie*, page 209–228, Dec 1969.
- [118] Henry Maguire, Jake Iles-Smith, and Ahsan Nazir. Environmental nonadditivity and franck-condon physics in nonequilibrium quantum systems. *Physical Review Letters*, 123, 08 2019.
- [119] A Einstein. Strahlungs-emission und-absorption nach der quantentheorie. *The Collected Papers of Albert Einstein*, 6:Doc–34, 1996.