OS100: A Benchmark Set of 100 Digitized UV-Visible Spectra and Derived Experimental Oscillator Strengths

Astrid S. Tarleton[a], Jorge C. Garcia-Alvarez[a], Anah Wynn[a], Cade M. Awbrey[a], Tomas P. Roberts[a], Samer Gozem[a*]

[a]Department of Chemistry, Georgia State University, Atlanta, GA 30302, United States.

*To whom correspondence should be addressed: sgozem@gsu.edu.

KEYWORDS: UV-visible spectroscopy, oscillator strength, transition dipole moment, extinction coefficient, attenuation coefficient.

ABSTRACT

The scientific method involves validating computational theories and methods against experimental results. However, the comparison between theory and experiments is not always straightforward; in UV-visible spectroscopy, experiments provide a plot of wavelength-dependent molar extinction/attenuation coefficients (ε) while computations typically provide single-valued excitation energies and oscillator strengths (f) for each band. ε and f are related, but this relation is complicated by various broadening and solvation effects. We describe a protocol to fit and integrate experimental UV-visible spectra to obtain f_{exp} values for absorption bands and to estimate the uncertainty in the fitting. We apply this protocol to derive 164 f_{exp} values from 100 organic molecules ranging in size from 6-34 atoms. The corresponding computed oscillator strengths (f_{comp}) are obtained with time-dependent density functional theory and a polarizable continuum solvent model. By expressing experimental and computed absorption strengths using a common quantity, we directly compare f_{comp} and f_{exp}. While f_{comp} and f_{exp} are well correlated (linear regression R^2=0. 914), f_{comp} in most cases significantly overestimates f_{exp} (regression slope=1.31). The agreement between absolute f_{comp} and f_{exp} values is substantially improved by accounting for a solvent refractive index factor, as suggested in some derivations in the literature. The 100 digitized UV-visible spectra are included as plain text files in the supporting information to aid in benchmarking computational or machine-learning approaches that aim to simulate realistic UV-visible absorption spectra.

INTRODUCTION

In UV-visible spectroscopy, the absorbance of near-ultraviolet and/or visible light depends on the light frequency and on the electronic structure of the sample. It also depends on the concentration of the sample and the path length of the light, as expressed in the Beer-Lambert law:[1-3]

$$A(\nu) = \varepsilon(\nu)\, l\, c \qquad (1)$$

In Equation (1), $A(\nu)$ is the absorbance of the sample at frequency ν, l is the optical path length, c is the concentration of the sample, and $\varepsilon(\nu)$ is the molar attenuation coefficient at frequency ν (also often called molar extinction coefficient, or molar absorptivity). A plot of $\varepsilon(\nu)$ vs. ν (or wavelength λ or wavenumber $\tilde{\nu}$) gives the UV-visible absorption spectrum. To summarize the information contained in UV-visible spectra, spectroscopists often characterize absorption bands by reporting the wavelength of maximal absorption (λ_{max}) and the molar attenuation coefficient at maximal absorption (ε_{max}).

Excited-state quantum chemical methods have reached an evident level of maturity in predicting λ_{max}. The effort to improve these predictions continues, but hundreds of studies indicate that it is possible to predict λ_{max} to a fraction of an eV.[4] For instance, an extensive benchmark of time-dependent density functional theory (TD-DFT) vertical excitation energies indicates typical errors of ~0.15-0.25 eV for $\pi \rightarrow \pi^*$ transitions in organic dyes.[5] These errors can be reduced to chemical accuracy (<0.043 eV) by going beyond the vertical approximation and using high-level excited-state *ab initio* methods.[5-8]

In comparison, we have a limited understanding of how well quantum chemical calculations can predict absorption strengths. This is not due to a lack of interest; predicting strengths is arguably as essential as predicting energies, with important applications in the design of dyes and in the identification and quantification of analytes, for instance.[9-13] However, comparing computed and experimental absorption strengths is not straightforward. In computations, transition strengths are often represented by a single value (typically, an oscillator strength or a transition dipole moment). In experiments, $\varepsilon(\nu)$ depends on the absorption wavenumber and is not single-valued, while ε_{max} is affected by broadening and solvent effects that modulate the width, and therefore also the height, of the absorption bands. This includes intrinsic broadening effects like lifetime broadening, Doppler broadening, and vibrational and rotational effects, as well as extrinsic factors like pressure and solvent broadening (see **Fig. 1**).[14, 15] Therefore, the often-reported experimental ε_{max} is not an ideal representation of absorption strength.

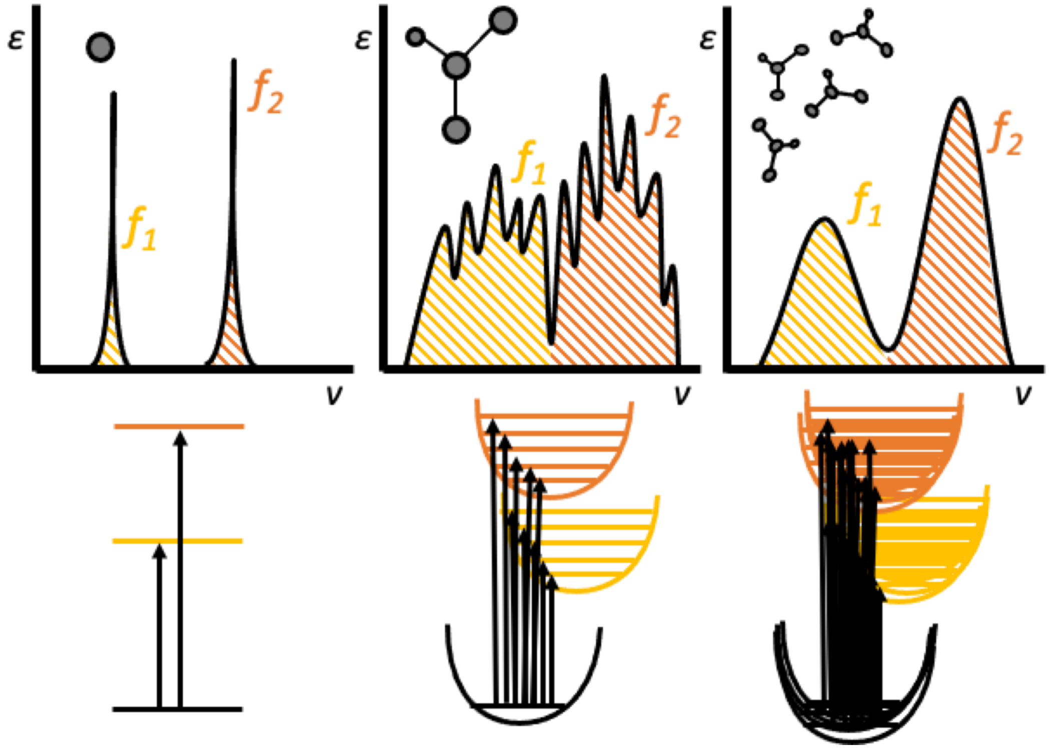


Figure 1: Line broadening in absorption spectra. Top panels show schematic UV-visible absorption spectra, while the bottom diagrams illustrate the corresponding transitions. In gas-phase atoms (left), where broadening is caused by lifetime, Doppler, and pressure effects, absorption lines typically appear sharp. Molecular spectra (center) are additionally broadened by vibrational and rotational energy levels, but the vibronic peaks may still be resolved, particularly in rigid gas-phase systems. In conformationally flexible or condensed-phase systems (right) each transition in the absorption spectrum appears as a broad band. In all cases, the total probability of the transition occurring is related to the area under the band for that transition (represented using f_1 and f_2 for each band).

A more representative way to describe the absorption strength for a transition is the total area under the corresponding band (**Fig. 1**). It is possible to derive a single-valued experimental oscillator strength (f_{exp}) from the UV-visible spectrum by integrating $\varepsilon(\tilde{\nu})$ over the range of wavenumbers, $\tilde{\nu}$, for a given band:[16]

$$f_{\text{exp}} = \frac{10^3 \ln(10) m_e c^2}{N_A \pi e^2} \int \varepsilon(\tilde{\nu}) d\tilde{\nu} = (4.32 \times 10^{-9}\ \text{M cm}^2) \int \varepsilon(\tilde{\nu}) d\tilde{\nu} \qquad (2)$$

In equation (2), m_e is the mass of the electron, c the speed of light in vacuum, N_A is Avogadro's constant, and e is the elementary charge. The numerical value in the second equality can be used when *ε($\tilde{v}$)* and $\tilde{v}$ are in units of $M^{-1}cm^{-1}$ and cm^{-1}, respectively. The limits of integration should go over the range of wavenumbers for the band of interest. This is trivial for well-defined and separated bands like those shown in Fig. 1, but for overlapping bands where multiple transitions contribute to absorbance at a common wavenumber, equation (2) cannot distinguish between different bands. This is discussed further in the Methodology section.

The oscillator strength can also be derived from quantum mechanically computed transition dipole moments:[17]

$$f_{comp} = \frac{4\pi m_e v}{3\hbar} \, |\langle \psi_I | \boldsymbol{r} | \psi_F \rangle|^2 \qquad (3)$$

In equation (3), v is the frequency for the transition, ħ is the reduced Planck constant, and $\langle \psi_I | \mathbf{r} | \psi_F \rangle$ is the transition dipole moment connecting the initial (ψ_I) and final (ψ_F) electronic state wave functions. When multiple transitions contribute to a band in a UV-visible spectrum, it is possible to sum equation (3) for several final state transitions.

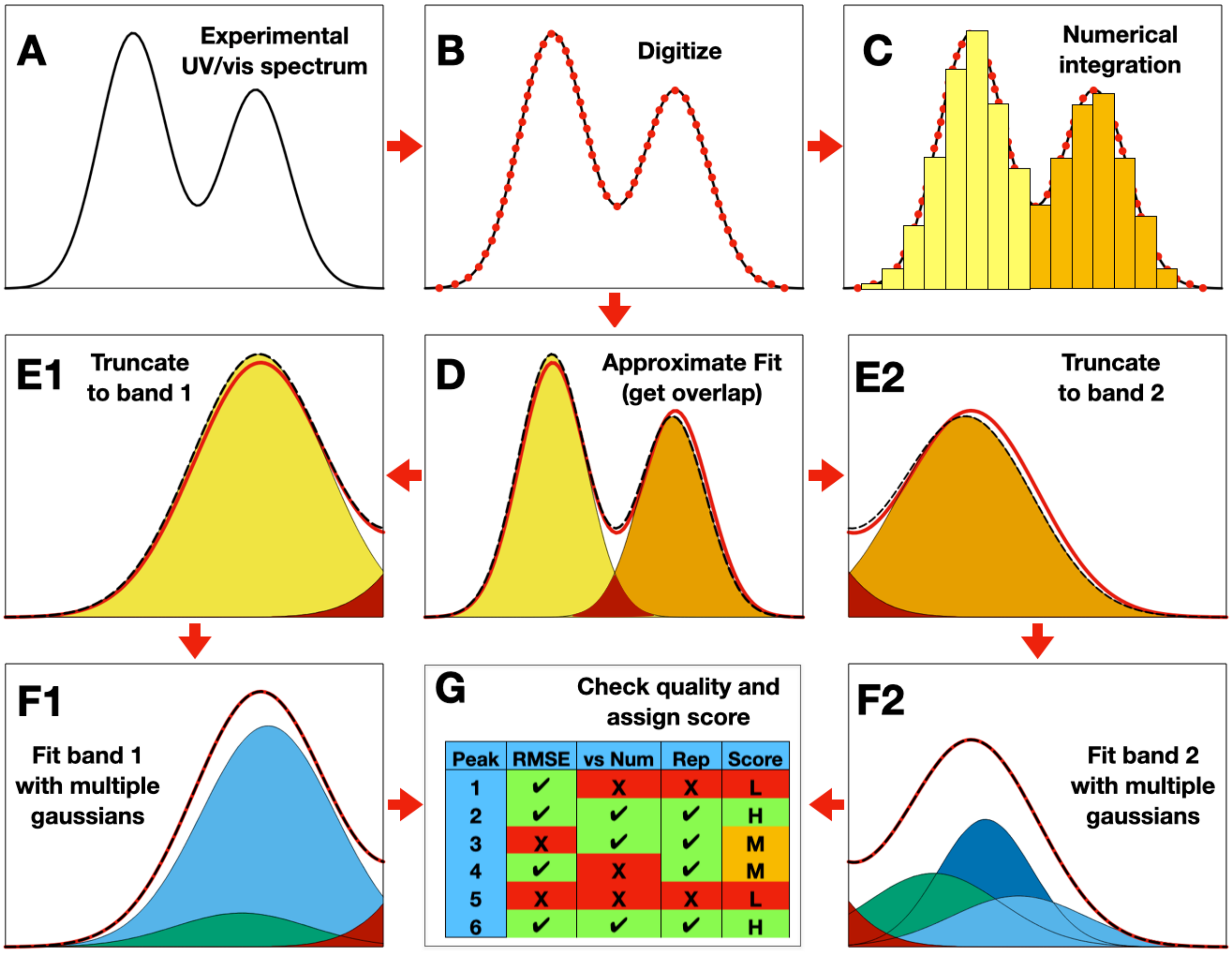
Equations (2) and (3) give experimental and computed oscillator strengths, respectively, in a common quantity that is directly comparable. However, very few experimental studies report oscillator strengths, while even fewer report the protocol used to derive f_{exp} from the UV-visible absorption spectrum. Therefore, most quantum chemical oscillator strength studies employ high-level *ab initio* methods as the reference, rather than using an experimental reference.[4] There is still limited understanding of how computed oscillator strength calculations compare to experimental ones.

Compounding the difficulty in comparing computations and experiments is that equations (2) and (3) were derived for systems *in vacuo* and may not be exact for molecules in solution. Specifically, accounting for the refractive index of the solvent (*n*) appears to be necessary in one or in both equations. How to do that appears to be a topic of debate starting from the 1930s[18] that is still unresolved, as far as we know. For instance, reports in the literature often use equation (2) as-is,[19-22] while others have multiplied equation (2) by factors of n^{-2},[23, 24] n^{-1},[25, 26] n,[27-29] or some more complex function *n*.[30-34] In this manuscript, we initially use equation (2) as-is. We also assume that calculations using a polarizable continuum model (PCM) already account for the terms necessary to account for the solvent effect on oscillator strength (see Methodology). We then discuss the effect of the refractive index after presenting the results.

In a 2013 review on TD-DFT benchmarks, Laurent and Jacquemin cited several studies that included oscillator strength benchmarks up to that time.[4] A majority of these studies focused on a few molecules, and compared computed oscillator strengths with other theoretical results, or a mix of theory and experiments. Several benchmark studies since then have focused on comparing TD-DFT calculations with high-level *ab-initio* calculations for small molecules.[35, 36] The most extensive validation against experiments, to our knowledge, was published by Jacquemin *et al.*[37] for a series of thirty related anthraquinones.[38]

We collected and digitized one-hundred solution-phase UV/visible spectra of organic molecules ranging in size from 6-34 atoms. The benchmark set includes molecular cations and anions, and occasionally include thio-compounds and heavy atom substituents (up to bromine). The benchmark set is named OS100. We note that there is no relation to the GW100 benchmark of ionization potentials and electron affinities,[39] only that both sets of benchmarks include 100 molecules. The 100 UV-visible spectra were all obtained from a common source, the UV Atlas of organic compounds.[40] The UV Atlas is one of the resources used by the NIST online database, and reports experimental details including the solvent, concentration at which the spectra were taken, and spectral resolution for the all the measurements. 164 f_{exp} values were derived from well-defined bands appearing in these spectra. The main purpose of this work is to introduce the benchmark set and describe the fitting and scoring system. While extensive benchmarking of quantum chemical methods is left to future work, here we employ TD-DFT calculations with a popular functional, B3LYP, to provide a preliminary comparison of f_{exp} and f_{comp}. The results are well-correlated (linear regression R^2=0.914), although TD-B3LYP computations appear to systematically overestimate the oscillator strength (regression slope=1.31). This large discrepancy could be explained by a refractive index factor missing in equations (2) and/or (3).

METHODOLOGY


A
Experimental UV/vis spectrum
B
Digitize
C
Numerical integration
E1
Truncate to band 1
D
Approximate Fit (get overlap)
E2
Truncate to band 2
F1
Fit band 1 with multiple gaussians
G
Check quality and assign score
F2
Fit band 2 with multiple gaussians

Peak	RMSE	vs Num	Rep	Score
1	✔	X	X	L
2	✔	✔	✔	H
3	X	✔	✔	M
4	✔	X	✔	M
5	X	X	X	L
6	✔	✔	✔	H

Figure 2: A schematic figure of the protocol used to derive oscillator strengths from experimental UV/vis spectra. See text for details.

The protocol to digitize UV-visible spectra and obtain f_{exp} is schematically shown in **Fig. 2**. The UV-visible spectra were digitized (panel **B**) and numerically integrated using the midpoint method (panel **C**) to obtain the areas under each band. The limits of the numerical integration were set as the minima in ε at the low and high energy range of each band when a band can be distinctly identified. Bands with an oscillator strength below 0.01 were not considered. From here on, we refer to values of f_{exp} obtained by numerical integration as $f_{exp,n}$. In cases of incomplete or overlapping bands (like the example in **Fig. 2**), numerical integration does not correctly treat overlap region; it accounts for spillover from nearby bands and/or misses part of the band that is outside of the numerical integration window. This error is in addition to numerical errors associated with finite-width midpoint rule integration. However, in case of well-separated bands (like those in **Fig. 1**), $f_{exp,n}$ is expected to be a good approximation for f_{exp}. In cases of strongly overlapping bands, the bands were included in a single integration and assigned one oscillator strength value, which can be compared to the sum of their computed oscillator strengths.

Next, we performed a deconvolution of the UV-visible spectrum using a minimal number of Gaussian functions (panel **D**). Gaussian functions were used since major contributions to room-temperature broadening in the condensed phase have Gaussian profiles. While a Voigt function would also be suitable, neither a single Voigt function nor a single Gaussian function could accurately fit the absorption bands, and a more complex function was needed for quantitative fitting. Therefore, initially, the spectra were fit using only a minimal number of Gaussian curves to describe overlap regions between bands (see panels **D**, **E1,** and **E2**). Then, for each band, the area under the curve was fit using an increasing number of Gaussian curves while freezing the fits for the nearby overlapping bands, until convergence (panels **F1** and **F2**). Convergence was monitored using the normalized root-mean-square deviation (%nRMSD, defined in Sources of Error). In most cases, %nRMSD was kept below 0.5%. In difficult to converge cases, a system was considered converged when %nRMSD does not change upon adding Gaussian curves. Bands that did not converge were discarded from the benchmark set (i.e., are not included among the 164 transitions). The area under the curve was obtained by analytical integration of the Gaussian functions included in fitting that band, excluding the spillover from nearby bands represented in red in **Fig. 2**. Oscillator strengths derived in this way are labelled $f_{exp,g}$.

Steps B-F were repeated a second time for the same molecule, by a different person, using a different resolution for the digitization. This ensured reproducibility and provided an estimate of errors introduced during each of the digitization, numerical integration, and fitting processes. Ultimately, two sets of $f_{exp,n}$ and $f_{exp,g}$ values for each molecule were produced, labeled $f_{exp,n1}$, $f_{exp,n2}$, $f_{exp,g1}$, and $f_{exp,g2}$. By convention, the higher resolution data are labeled $f_{exp,n1}$ and $f_{exp,g1}$ while the lower resolution data are labeled $f_{exp,n2}$ and $f_{exp,g2}$. The $f_{exp,n}$ and $f_{exp,g}$ data reported in this work are the averages of the two trials. Using these values, the quality of the integration was scored by checking for convergence, for consistency between $f_{exp,n}$ and $f_{exp,g}$, and for consistency between the repeats carried out independently by two different individuals (see Sources of Error). Each band fit was then assigned a score indicating confidence in the integration process (panel **G**). This process was repeated for the 100 molecules.

Quantum chemical geometry optimizations were carried out at the density functional theory (DFT) level of theory using the B3LYP functional[41, 42] and 6-31+G* basis set for all molecules. Frequency calculations were carried out at the same level of theory to ensure that exclusively positive frequencies are obtained for all molecules. TD-DFT calculations were carried out using the same functional and basis set. Preliminary calculations were performed in the gas phase by two different individuals, to check for consistency. Differences between the two trials could be reconciled by checking for differences in the structures. A final set of calculations were then performed using the integral equation formalism of PCM (IEF-PCM).[43] The TD-DFT calculations with PCM employ a non-equilibrium linear response formalism to account for the effect of solvation on the excitations.[44] The f_{comp} values reported in this work are the ones accounting for the solvent effect on excitation.

WebPlotDigitizer was used to digitize the UV-visible spectra[45] and Excel's Solver plugin was used to fit the spectra using the nonlinear generalized reduced gradient algorithm.[46] Regression and confidence interval analyses were carried out in Mathematica.[47] The two sets of computations were carried out using the random phase approximation formulation of TD-DFT[48] in Q-Chem[49] and using the default TD-DFT formulation in Gaussian 16,[50] respectively. The calculations with TD-DFT and PCM solvation were performed in Gaussian 16.[50]

SOURCES OF ERROR

When using experimental data as benchmarks, it is important to recognize sources of experimental errors. There are errors introduced by instrumentation, experimental design, sample quality (impurities, uncertainties in concentration, concentration-dependent aggregation effects, etc.), and human error. Some of these issues have been discussed at length in the literature, but it remains difficult to quantify the magnitude of the errors.[51-53] One concern raised for experimental oscillator strengths derived from photoabsorption spectroscopy and the Beer-Lambert law is that errors introduced by line-saturation effects could result in the underestimation of the experimentally measured f_{exp} relative to the "true" strength f.[36, 54] This problem is well illustrated by Chan *et al*.[53] However, it appears that such issues are particularly problematic for narrow bands (e.g., well-resolved gas-phase spectra where there are vibronic peaks with full-width at half-maxima that are on the order of hundreds of cm^{-1}). The severity of these errors is reduced for broad absorption bands that are thousands of cm^{-1} across, like most of the bands used in this work.

We do not attempt to quantify the errors in f_{exp}, although we recognize that they exist. However, the uncertainty in f_{exp} is yet another motivation for an extensive benchmark study comparing experimental and computed oscillator strengths. By determining random or systematic errors against computations, particularly for high-level quantum chemical calculations, errors in the experimental measurements could be better understood.[55]

Errors in f_{comp} exist when the transition dipole moment in equation (3) is computed using approximate rather than exact wave functions. This error can be quantified by systematically moving towards high-level quantum chemical methods,[4, 35-37] but this requires benchmarking using small molecules where these methods are affordable. Calculations of f_{comp} in this work also neglect nuclear motion, which can often be justified for rigid molecules within the Born-Oppenheimer and Condon approximations[56] but does not consider variations of oscillator strength with changes in conformation of flexible molecules. Finally, approximate solvation models may not correctly capture the effect of solvation on f_{comp}. The magnitudes of all these errors are difficult to quantify as well, which is why extensive benchmarking against experimental oscillator strengths is needed.

Here, we focus on uncertainties introduced in the fitting procedure itself, which we identify using four metrics:

- **%gn:** The percentage difference between $f_{exp,g}$ and $f_{exp,n}$. In practice, we compute two sets of %gn values then take the average:

$$\%gn = \frac{1}{2}\left(\frac{|f_{exp,g1} - f_{exp,n1}|}{f_{exp,g1}} + \frac{|f_{exp,g2} - f_{exp,n2}|}{f_{exp,g2}}\right) \qquad (4)$$

  Values of %gn ranged from 0.0% to 32.7%, with an average of 6.0%, a first quartile of 1.5%, median of 4.5%, and third quartile of 7.5%.

- **%gg:** The percentage difference between $f_{exp,g1}$ and $f_{exp,g2}$. This as an indicator of the reproducibility of the Gaussian fitting procedure performed by two different persons:

$$\%gg = \frac{2|f_{exp,g1} - f_{exp,g2}|}{(f_{exp,g1} + f_{exp,g2})} \qquad (5)$$

  Values of %gg ranged from 0.0% to 37.9%, with an average of 4.2%, a first quartile of 0.9%, median of 2.6%, and third quartile of 5.1%.

- **%nn:** The percentage difference between $f_{exp,n1}$ and $f_{exp,n2}$. Since the original spectra were plotted on a logarithmic scale, this unavoidably introduces errors in the digitization of the UV-visible spectra, particularly for intense bands. Therefore, %nn reflects the uncertainty introduced by both the digitization and numerical integration procedures. %nn is computed in the same way as %gg in equation (5) but using the numerical $f_{exp,n1}$ and $f_{exp,n2}$. Values of %nn ranged from 0.0% to 7.0%, with an average of 1.4%, a first quartile of 0.5%, median of 1.1%, and third quartile of 1.8%.

- **%nRMSD:** The normalized root-mean-square deviation between the digitized UV-visible spectrum and the Gaussian deconvoluted spectrum. The %nRMSD is computed using:

$$\%nRMSD = \frac{\sqrt{\sum_{i=1}^{n} \frac{(\varepsilon_{i,fit} - \epsilon_i)^2}{n}}}{\epsilon_{max} - \epsilon_{min}} \qquad (6)$$

In equation (6), for each point i in the digitized spectrum, ε_i is the experimental attenuation coefficient at that point, $\varepsilon_{i,fit}$ is the corresponding value at that point obtained from the sum of the Gaussian curves used in the fitting of the band, and n is the total number of points used in the digitized spectrum. The numerator in equation (6) is the RMSD, but since RMSD is not informative on its own (it is typically larger for high intensity bands, smaller for low intensity bands), we divide the RMSD by the range of attenuation coefficients spanned by the UV-visible spectra ($\varepsilon_{max} - \varepsilon_{min}$). The final %nRMSD used is an average from the two trials. Values of %nRMSD ranged from 0.2% to 6.0%, with an average of 0.7%, a first quartile of 0.4%, median of 0.5%, and third quartile of 0.8%.

The above percentages are associated with uncertainties in the fitting process and are not errors. For instance, It is expected that $f_{exp,g}$ and $f_{exp,n}$ could be different, particularly for incomplete or overlapping bands, yielding large %gn. A large %gn is therefore an indicator of uncertainty (i.e., an incomplete band causes uncertainty in how to fit it). %gg is also an important indicator of uncertainty; if the same band is fit differently by two different individuals, it indicates a lower confidence in the fitting. Thus, we have developed a scoring system to indicate confidence in the analytical integration. The scores are based on the four metrics listed above, with the relative weights based on the ratio of the first quartiles:

%gn : %gg : %nn : %nRMSD = 1.5 : 0.9 : 0.5 : 0.4 ≈ 9 : 6 : 3 : 2

Each transition is scored out of 20 points (9+6+3+2). Integer scores are assigned by deducting points for each 1% introduced in each of %gn, %gg, %nn, and %nRMSD, up to the maximum score for that category. E.g., a score of 20 is obtained if all four metrics have a value <1%. If, for instance, %gn is 4.1% %gg is 3.8%, %nn is 4.2%, and %nRMSD is 1.02%, then the total score is

9 (9-4 points for %gn, 6-3 points for %gg, 0 points for %nn, and 2-1 point for %nRMSD). Molecules are categorized as “very high” confidence if the score is a perfect 20, “high” confidence for a score in the upper quartile (17-19), “medium” confidence for a score in the second quartile (12-16), “low” confidence for a in the third quartile (8-11), and “very low” confidence for a score in the lowest quartile (0-7).

RESULTS

Out of 164 transitions, 14 had a “very high” confidence score, 29 scored as “high”, 42 scored as “medium”, 42 scored as “low”, and 37 as “very low.” A list of the 100 molecule names and solvent information (**Table S1**), f_{comp}, $f_{exp,g1}$, $f_{exp,g2}$, $f_{exp,n1}$, $f_{exp,n2}$, and scores for the 164 transitions (**Table S2**), and ν_{min}, ν_{max}, ε_{max}, and ν at ε_{max} (**Table S3**) are included in the Supporting Information (SI) document. Plain text files with the digitized UV-visible spectra, spreadsheets used for the numerical integration and gaussian deconvolution for $f_{exp,n1}$ and $f_{exp,g1}$, ChemDraw chemical structures, and B3LYP/6-31+G*/PCM geometry-optimized structures for the 100 organic molecules are also included in the SI.

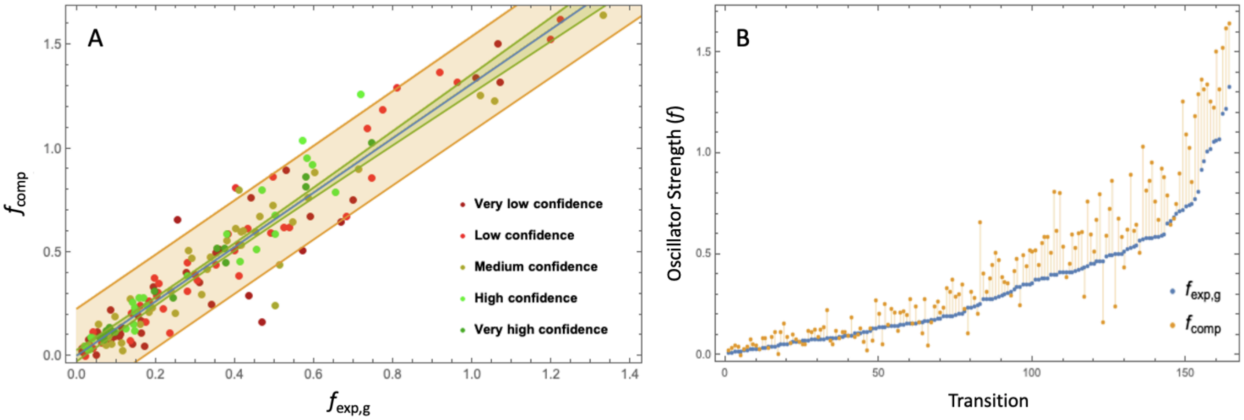


Figure 3: **A.** f_{comp} vs. $f_{exp,g}$ for 164 transitions. The points are colored by confidence level, as indicated in the legend. The blue line is a linear regression (equation: y=1.3099x+0.0028, R^2=0.9140; see comment[57]). The orange lines and area indicate 95% confidence bands for the data, while the green lines and area indicate the 95% mean confidence prediction bands. **B.** f_{comp} and $f_{exp,g}$ for each transition ordered by increasing strength of $f_{exp,g}$. The lines connecting the points represent the absolute differences.

Fig. 3 shows the correlation between $f_{exp,g}$ and f_{comp}. The linear regression analysis (R^2=0.914) indicates a strong correlation between experimental and computed oscillator strengths. However, the slope of the plot (1.31) indicates a clear systematic error; The computations substantially overestimate the oscillator strength. The mean absolute error (MAE) between f_{comp} and $f_{exp,g}$ is 0.13.

Most of the green and yellow dots in **Fig. 3A** appear close to the linear regression line, indicating that low and very low confidence points increase the spread of the data. If “very low confidence” points are excluded from the plot, the R^2 value increases to 0.930 and the slope becomes 1.33 (see SI **Fig. S1**), while the MAE stays 0.13. If plots used ε_{max} instead of $f_{exp,g}$ as the reference, the R^2 value for f_{comp} vs. ε_{max} would be 0.78 (see SI **Fig. S2**). Therefore, ε_{max} is correlated with f_{comp} as expected, but this correlation is weaker than f_{comp} with the area under the band.

We note again that equations (2) and (3) are valid *in vacuo*. To understand the effect of the refractive index *n* on the agreement between computed and experimental oscillator strengths, we plot f_{comp} against $f_{exp,g}/n$ (SI **Fig. S3**) and $f_{exp,g} \times n$ (SI **Fig. S4**). Refractive indices for the solvents are listed in **Table S1**.

For f_{comp} against $f_{exp,g}/n$, the linear regression equation is y=1.7697x+0.0059, with R^2=0.910. Removing “very low confidence” data gives y=1.7956x+0.0104, with R^2=0.926. The MAE increases to 0.20 (0.21 after removing “very low confidence” data). Therefore, dividing the experimental oscillator strength from equation (2) by the refractive index *n* worsens the agreement with computations significantly, and slightly reduces the R^2.

For f_{comp} against $f_{exp,g} \times n$, the linear regression equation is y=0.9674x+0.0004, with R^2=0.917. Removing “very low confidence” data gives y=0.9796x+0.0047, with R^2=0.932. The MAE decreases to 0.08 (0.07 after removing “very low confidence” data). Therefore, multiplying the experimental oscillator strength from equation (2) by the refractive index *n* increases the absolute agreement between f_{comp} and f_{exp} for most points and slightly improves the R^2.

CONCLUSIONS

Using a benchmark set of 164 transitions from 100 molecules, we have compared quantum chemically computed oscillator strengths with oscillator strengths we derived from experimental UV-visible absorption spectra.. Using equation (2) to obtain experimental oscillator strengths, the discrepancy between experimental and computed oscillator strengths is large (slope 1.31). One reason for this discrepancy could be the refractive index. Reports in the literature disagree on how equation (2) must be modified for absorption of molecules in solvent.[18-34] Our benchmark cannot conclusively resolve this debate, although the discrepancy between experiments and computations can be largely reconciled by a factor of n (either multiplying $\mathrm{f}_{\mathrm{exp,g}}$ by n or dividing $\mathrm{f}_{\mathrm{comp}}$ as computed in the TD-DFT/PCM non-equilibrium linear response formalism[44, 50] by n). This conclusion would be better supported through continued benchmarking, such as by expanding this benchmark set to include additional “high” and “very high” confidence data and performing additional computations to determine the sensitivity of the f_{comp} and $f_{exp,g}$ plot on the quantum chemical level of theory used.

An important next step would be to quantify the accuracy of continuum or explicit solvent models in predicting solvent effects oscillator strengths.[58] Both implicit and explicit models appear to successfully predict trends in oscillator strengths for the same molecule in different environments,[59, 60] but it is not yet clear if such predictions are quantitative.

Finally, while one way to move experiments and theory closer together is by deriving easily computable quantities from experimental data, as done in this work, another approach is to continue to develop computations that simulate realistic experimental spectra through combined quantum and (semi-)classical methods or machine learning approaches.[61-69] To aid in benchmarking for such efforts, we have provided the digitized UV-visible data and optimized molecular geometries in the SI.

ACKNOWLEDGEMENTS

The idea for this project began after a discussion with Prof. Gregory Poon. Emrah Trumic, Andy Nguyen, and Audi Jeter aided in the selection and preliminary digitizing/fitting of UV-visible spectra for inclusion in earlier versions of the OS100 set. S.G. acknowledges NSF XSEDE for computational resources from Research Allocation CHE180027. We also acknowledge the use of Georgia State University's research computing resources[70, 71] that are supported by Georgia State's Research Solutions. S.G. acknowledges the donors of the American Chemical Society Petroleum Research Fund for support through PRF# 60837-DNI6. This material is in part based upon work supported by the National Science Foundation (NSF) under Grant CHE-2047667.

Declaration of competing financial interests

The authors declare no competing financial interests.

REFERENCES

1. Bouguer, P., *Essai d'optique sur la gradation de la lumière*. Claude Jombert: 1729.
2. Lambert, J., Photometria sive de mensura et gradibus luminis colorum et umbrae augsburg. *Detleffsen for the widow of Eberhard Klett* **1760**.
3. Beer, A., Bestimmung der absorption des rothen lichts in farbigen flussigkeiten. *Ann. Physik* **1852,** *162*, 78-88.
4. Laurent, A. D.; Jacquemin, D., TD-DFT benchmarks: a review. *Int. J. Quantum Chem* **2013,** *113* (17), 2019-2039.
5. Jacquemin, D.; Wathelet, V.; Perpete, E. A.; Adamo, C., Extensive TD-DFT Benchmark: Singlet-Excited States of Organic Molecules. *J. Chem. Theory Comput* **2009,** *5* (9), 2420-35.
6. Loos, P. F.; Galland, N.; Jacquemin, D., Theoretical 0-0 Energies with Chemical Accuracy. *J Phys Chem Lett* **2018,** *9* (16), 4646-4651.
7. Santoro, F.; Jacquemin, D., Going beyond the vertical approximation with time-dependent density functional theory. *WIREs Comp. Mol. Sci* **2016,** *6* (5), 460-486.
8. Gozem, S.; Krylov, A. I., The ezSpectra suite: An easy-to-use toolkit for spectroscopy modeling. *WIREs Comp. Mol. Sci* **2021**, e1546.
9. Liu, X.; Xu, Z.; Cole, J. M., Molecular design of UV–vis absorption and emission properties in organic fluorophores: toward larger bathochromic shifts, enhanced molar extinction coefficients, and greater stokes shifts. *J. Phys. Chem. C* **2013,** *117* (32), 16584-16595.
10. Lee, Y.; Jo, A.; Park, S. B., Rational Improvement of Molar Absorptivity Guided by Oscillator Strength: A Case Study with Furoindolizine-Based Core Skeleton. *Angew. Chem. Int. Ed. Engl* **2015,** *54* (52), 15689-93.
11. Wang, J.; Cong, S.; Wen, S.; Yan, L.; Su, Z., A rational design for dye sensitizer: density functional theory study on the electronic absorption spectra of organoimido-substituted hexamolybdates. *J. Phys. Chem. C* **2013,** *117* (5), 2245-2251.
12. Beard, E. J.; Sivaraman, G.; Vazquez-Mayagoitia, A.; Vishwanath, V.; Cole, J. M., Comparative dataset of experimental and computational attributes of UV/vis absorption spectra. *Sci. Data* **2019,** *6* (1), 307.
13. Chen, X. K.; Tsuchiya, Y.; Ishikawa, Y.; Zhong, C.; Adachi, C.; Bredas, J. L., A New Design Strategy for Efficient Thermally Activated Delayed Fluorescence Organic Emitters: From Twisted to Planar Structures. *Adv. Mater* **2017,** *29* (46).
14. Hollas, J. M., *Modern spectroscopy*. John Wiley & Sons: 2004.
15. Myers, A. B., Molecular electronic spectral broadening in liquids and glasses. *Annu. Rev. Phys. Chem* **1998,** *49*, 267-95.
16. Turro, N. J.; Ramamurthy, V.; Ramamurthy, V.; Scaiano, J. C., *Principles of molecular photochemistry: an introduction*. University science books: 2009.
17. Hilborn, R. C., Einstein coefficients, cross sections, f values, dipole moments, and all that. *Am. J. Phys.* **1982,** *50* (11), 982-986.
18. Chako, N. Q., Absorption of light in organic compounds. *J. Chem. Phys* **1934,** *2* (10), 644-653.
19. Myers, A. B.; Birge, R. R., The effect of solvent environment on molecular electronic oscillator strengths. *J. Chem. Phys* **1980,** *73* (10), 5314-5321.
20. Birge, R. R., Reply to Abe's comments on ''The effect of solvent environment on molecular electronic oscillator strengths''[AB Myers and RR Birge, J. Chem. Phys. 7 3, 5314 (1980)]. *J. Chem. Phys* **1982,** *77* (2), 1075-1076.
21. Mohanty, J.; Nau, W. M., Refractive index effects on the oscillator strength and radiative decay rate of 2,3-diazabicyclo[2.2.2]oct-2-ene. *Photochem. Photobiol. Sci* **2004,** *3* (11-12), 1026-31.
22. Morris, J. V.; Mahaney, M. A.; Huber, J. R., Fluorescence quantum yield determinations. 9, 10-Diphenylanthracene as a reference standard in different solvents. *J. Phys. Chem* **1976,** *80* (9), 969-974.
23. Abe, T., Theory of solvent effects on oscillator strengths for molecular electronic transitions. *Bull. Chem. Soc. Jpn* **1970,** *43* (3), 625-628.
24. Abe, T., Comments on ''The effect of solvent environment on molecular electronic oscillator strengths''. *J. Chem. Phys* **1982,** *77* (2), 1074-1074.
25. Iweibo, I.; Oderinde, R.; Faniran, J., Electronic absorption spectra and structures of aniline and its 4-chloro, pentafluoro and pentachloro derivatives. *Spectrochim. Acta A* **1982,** *38* (1), 1-7.
26. Jinnouchi, Y.; Kohno, M.; Kuboyama, A., Solvent effects on the intensity of the nπ* absorption spectra of 1-phenylethylenetrithiocarbonate and p-benzoquinones. *Bull. Chem. Soc. Jpn* **1984,** *57* (4), 1147-1148.
27. Sklar, A. L., Electronic absorption spectra of benzene and its derivatives. *Rev. Mod. Phys* **1942,** *14* (2-3), 232.
28. Mulliken, R. S.; Rieke, C. A., Molecular electronic spectra, dispersion and polarization: The theoretical interpretation and computation of oscillator strengths and intensities. *Rep. Prog. Phys* **1941,** *8* (1), 231.

29. Hare, P. M.; Price, E. A.; Stanisky, C. M.; Janik, I.; Bartels, D. M., Solvated electron extinction coefficient and oscillator strength in high temperature water. *J. Phys. Chem. A* **2010,** *114* (4), 1766-1775.
30. Shibuya, T.-i., The refractive-index correction to the radiative rate constant. *Chem. Phys. Lett* **1983,** *103* (1), 46-48.
31. Hirayama, S.; Phillips, D., Correction for refractive index in the comparison of radiative lifetimes in vapour and solution phases. *J. Photochem* **1980,** *12* (2), 139-145.
32. Abe, T.; Iweibo, I., The experimental and theoretical expressions for the molecular electronic oscillator strength in solution. *J. Chem. Phys* **1985,** *83* (4), 1546-1550.
33. Shibuya, T. i., A dielectric model for the solvent effect on the intensity of light absorption. *J. Chem. Phys* **1983,** *78* (8), 5175-5182.
34. Dexter, D., Absorption of light by atoms in solids. *Phys. Rev* **1956,** *101* (1), 48.
35. Sarkar, R.; Boggio-Pasqua, M.; Loos, P. F.; Jacquemin, D., Benchmarking TD-DFT and Wave Function Methods for Oscillator Strengths and Excited-State Dipole Moments. *J. Chem. Theory Comput* **2021,** *17* (2), 1117-1132.
36. Chrayteh, A.; Blondel, A.; Loos, P. F.; Jacquemin, D., Mountaineering Strategy to Excited States: Highly Accurate Oscillator Strengths and Dipole Moments of Small Molecules. *J. Chem. Theory Comput* **2021,** *17* (1), 416-438.
37. Jacquemin, D.; Duchemin, I.; Blondel, A.; Blase, X., Assessment of the Accuracy of the Bethe-Salpeter (BSE/GW) Oscillator Strengths. *J. Chem. Theory Comput* **2016,** *12* (8), 3969-81.
38. Labhart, H., Zur quantitativen beschreibung des einflusses von substituenten auf das absorptionsspektrum ebener molekeln. Anwendung auf anthrachinon. *Helv. Chim. Acta* **1957,** *40* (5), 1410-1420.
39. van Setten, M. J.; Caruso, F.; Sharifzadeh, S.; Ren, X.; Scheffler, M.; Liu, F.; Lischner, J.; Lin, L.; Deslippe, J. R.; Louie, S. G., GW 100: Benchmarking G 0 W 0 for molecular systems. *J. Chem. Theory Comput* **2015,** *11* (12), 5665-5687.
40. UV Atlas of Organic Compounds. In *UV Atlas of Organic Compounds / UV Atlas organischer Verbindungen*, Springer US: Boston, MA, 1967; pp 5-605.
41. Becke, A. D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988,** *38* (6), 3098.
42. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988,** *37* (2), 785.
43. Miertuš, S.; Scrocco, E.; Tomasi, J., Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chem. Phys* **1981,** *55* (1), 117-129.
44. Cossi, M.; Barone, V., Time-dependent density functional theory for molecules in liquid solutions. *The Journal of chemical physics* **2001,** *115* (10), 4708-4717.
45. Marin, F.; Rohatgi, A.; Charlot, S., WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: application to ultraviolet spectropolarimetry. *arXiv preprint arXiv:1708.02025* **2017**.
46. Harris, D. C., Nonlinear least-squares curve fitting with Microsoft Excel Solver. *J. Chem. Educ* **1998,** *75* (1), 119.
47. Wolfram Research, Inc., *Mathematica*, Version 12.3.1; Champaign, Illinois, 2021.
48. Bouman, T. D.; Hansen, A. E.; Voigt, B.; Rettrup, S., Large-scale RPA calculations of chiroptical properties of organic molecules: Program RPAC. *Int. J. Quantum Chem* **1983,** *23* (2), 595-611.
49. Epifanovsky, E.; Gilbert, A. T.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L., Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *J. Chem. Phys* **2021,** *155* (8), 084801.
50. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
51. Reule, A., Errors in spectrophotometry and calibration procedures to avoid them. *Journal of research of the National Bureau of Standards. Section A, Physics and chemistry* **1976,** *80* (4), 609.

52. Mayerhöfer, T. G.; Pahlow, S.; Popp, J., The Bouguer-Beer-Lambert law: Shining light on the obscure. *ChemPhysChem* **2020,** *21* (18), 2029.
53. Chan, W.; Cooper, G.; Brion, C., Absolute optical oscillator strengths for the electronic excitation of atoms at high resolution: Experimental methods and measurements for helium. *Phys. Rev. A* **1991,** *44* (1), 186.
54. Chan, W.; Cooper, G.; Brion, C., The electronic spectrum of water in the discrete and continuum regions. Absolute optical oscillator strengths for photoabsorption (6–200 eV). *Chem. Phys* **1993,** *178* (1-3), 387-400.
55. Mata, R. A.; Suhm, M. A., Benchmarking quantum chemical methods: Are we heading in the right direction? *Angew. Chem. Int. Ed. Engl* **2017,** *56* (37), 11011-11018.
56. Condon, E., A theory of intensity distribution in band systems. *Phys. Rev* **1926,** *28* (6), 1182.
57. Linear regression could also be performed while forcing the y-intercept equal to zero, which is well justified in this case (the y-intercept in the linear regression is indeed small). Linear regression without a constant would yield a slope of 1.315 and R^2=0.962. The improved R^2 is not an indication of a better fit; setting the y-intercept to 0 results in the use of a different null model for computing R^2, and R^2 computed with and without a y-intercept cannot be compared directly. We have opted to keep the y-intercept unconstrained for the linear regressions in this work because it employs the more widely recognized definition of R^2, even if this yields a lower apparent R^2.
58. Herbert, J. M., Dielectric continuum methods for quantum chemistry. *WIREs Comp. Mol. Sci* **2021,** *11* (4), e1519.
59. Dratch, B. D.; Orozco-Gonzalez, Y.; Gadda, G.; Gozem, S., Ionic Atmosphere Effect on the Absorption Spectrum of a Flavoprotein: A Reminder to Consider Solution Ions. *J Phys Chem Lett* **2021,** *12* (34), 8384-8396.
60. Kabir, M. P.; Orozco-Gonzalez, Y.; Gozem, S., Electronic spectra of flavin in different redox and protonation states: a computational perspective on the effect of the electrostatic environment. *Phys. Chem. Chem. Phys* **2019,** *21* (30), 16526-16537.
61. Crespo-Otero, R.; Barbatti, M., Spectrum simulation and decomposition with nuclear ensemble: formal derivation and application to benzene, furan and 2-phenylfuran. In *Marco Antonio Chaer Nascimento*, Springer: 2014; pp 89-102.
62. Petrenko, T.; Neese, F., Analysis and prediction of absorption band shapes, fluorescence band shapes, resonance Raman intensities, and excitation profiles using the time-dependent theory of electronic spectroscopy. *J. Chem. Phys* **2007,** *127* (16), 164319.
63. Borrego-Sánchez, A.; Zemmouche, M.; Carmona-García, J.; Francés-Monerris, A.; Mulet, P.; Navizet, I.; Roca-Sanjuán, D., Multiconfigurational Quantum Chemistry Determinations of Absorption Cross Sections (σ) in the Gas Phase and Molar Extinction Coefficients (ε) in Aqueous Solution and Air–Water Interface. *J. Chem. Theory Comput* **2021**.
64. Cerezo, J.; Avila Ferrer, F. J.; Prampolini, G.; Santoro, F., Modeling Solvent Broadening on the Vibronic Spectra of a Series of Coumarin Dyes. From Implicit to Explicit Solvent Models. *J. Chem. Theory Comput* **2015,** *11* (12), 5810-25.
65. Ončák, M.; Šištík, L.; Slavíček, P., Can theory quantitatively model stratospheric photolysis? Ab initio estimate of absolute absorption cross sections of ClOOCl. *J. Chem. Phys* **2010,** *133* (17), 174303.
66. Zutterman, F.; Liégeois, V.; Champagne, B., Simulation of the uv/visible absorption spectra of fluorescent protein chromophore models. *ChemPhotoChem* **2017,** *1* (6), 281-296.
67. Xue, B.-X.; Barbatti, M.; Dral, P. O., Machine learning for absorption cross sections. *J. Phys. Chem. A* **2020,** *124* (35), 7199-7210.
68. Dral, P. O.; Barbatti, M., Molecular excited states through a machine learning lens. *Nature Reviews Chemistry* **2021,** *5* (6), 388-405.
69. Zuehlsdorff, T. J.; Isborn, C. M., Modeling absorption spectra of molecules in solution. *Int. J. Quantum Chem* **2019,** *119* (1), e25719.
70. Sarajlic, S.; Edirisinghe, N.; Lukinov, Y.; Walters, M.; Davis, B.; Faroux, G. In *Orion: discovery environment for HPC research and bridging XSEDE resources*, Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale, 2016; pp 1-5.
71. Sarajlic, S.; Edirisinghe, N.; Wu, Y.; Jiang, Y.; Faroux, G., Training-based workforce development in Advanced Computing for Research and Education (ACoRE). In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, 2017; pp 1-4.