Machine Learning for Absorption Cross Sections

Bao-Xin Xue,[a] Mario Barbatti,[b] and Pavlo O. Dral[a]

[a]State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Department of Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China

[b]Aix Marseille University, CNRS, ICR, Marseille, France

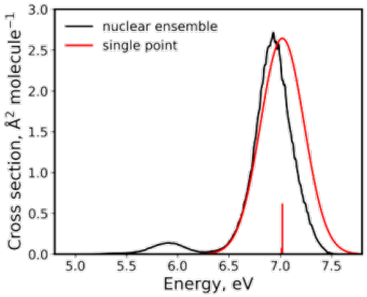
Email: bxxue@stu.xmu.edu.cn

Dral's group website: http://dr-dral.com/people

Introduction

Experimentally measured molar attenuation coefficient ε is directly related to the cross-section σ:[2]

$$\varepsilon[\mathrm{M^{-1}cm^{-1}}] = \frac{\sigma[\text{Å}\cdot\text{molecule}^{-1}]}{3.82353\cdot 10^{-5}}$$



Thus, we can simulate absorption spectra by calculating cross-section.

Commonly used approach is single point convolution (SPC), which only performs quantum chemical calculations at the ground state geometry, and then broadens oscillator strengths with the Gaussian function.

Much more accurate method is **Nuclear Ensemble Approach (NEA).** It calculates cross section by averaging over multiple normalized broadening functions at different conformations[3,4]. Compared with the single point convolution (SPC), NEA successfully makes a prediction for the absorption intensity when transitions are forbidden (have zero oscillator strength) at the ground state conformation

Machine Learning: KREG Model

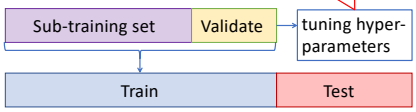
$$f(\mathbf{x}_i) = \sum_{j=1}^{N_{\mathrm{tr}}} \alpha_j k(\mathbf{x}_i,\mathbf{x}_j) \quad k(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\frac{1}{2\sigma^2}\sum_{s}^{N_x}(x_{i,s}-x_{j,s})^2\right) \quad x = \left(\cdots \quad \frac{R^{eq}}{R} \quad \cdots\right)^T$$

the Gaussian kernel function
σ is the kernel width)

RE descriptor

The regression coefficients α are found by solving the linear system of equations for the training set in matrix form:

$$\begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}_1)+\lambda & \cdots & k(\mathbf{x}_1,\mathbf{x}_{N_{\mathrm{tr}}}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_{N_{\mathrm{tr}}},\mathbf{x}_1) & \cdots & k(\mathbf{x}_{N_{\mathrm{tr}}},\mathbf{x}_{N_{\mathrm{tr}}})+\lambda \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{N_{\mathrm{tr}}} \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_{N_{\mathrm{tr}}} \end{pmatrix} \qquad (\mathbf{K}+\lambda\mathbf{I})\boldsymbol{\alpha} = \mathbf{y}$$



Model validation and hyperparameter tuning

We use the KREG model (**K**ernel ridge regression [KRR] with **RE** descriptor and the **G**aussian kernel function; RE descriptor stands for Internuclear distances **R**elative to **E**quilibrium)[5] to complete all the ML tasks. For excitation energies and oscillator strengths of each state, we train individual ML models, and then make prediction for 50000 nuclear configurations to calculate cross-section.

Absorption spectra can be calculated with *Mlatom*[5,6] and *Newton-X*[7] much faster and more accurately

B.-X. Xue, M. Barbatti, P. O. Dral, *J. Phys. Chem. A* **2020**, 124, *35*, 7199–7210

Read the blog post!

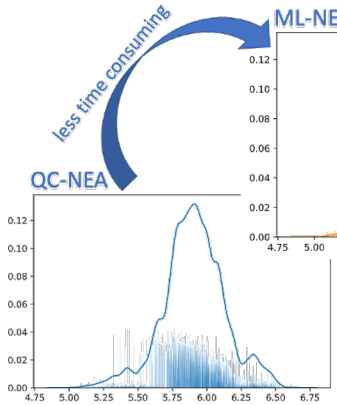
Start using MLatom! http://mlatom.com

Download this poster. password: 32CCS

ML-NEA

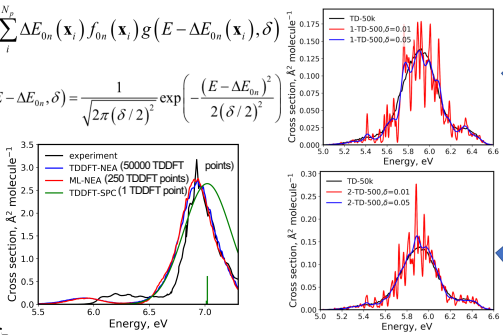
$$\sigma(E) = \frac{\pi e^2 \mathrm{h}}{2mc\varepsilon_0 E}\sum_{n}^{N_{fs}}\frac{1}{N_p}\sum_{i}^{N_p}\Delta E_{0n}(\mathbf{x}_i) f_{0n}(\mathbf{x}_i) g\left(E-\Delta E_{0n}(\mathbf{x}_i),\delta\right)$$

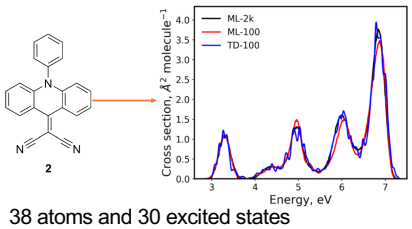
$$g(E-\Delta E_{0n},\delta) = \frac{1}{\sqrt{2\pi(\delta/2)^2}}\exp\left(-\frac{(E-\Delta E_{0n})^2}{2(\delta/2)^2}\right)$$



Nuclear ensemble is sampled in our study[1] from a Wigner distribution (normal mode sampling).

Because many conformations are quite similar, we can use ML to interpolate between them. Hence we can separate the whole ensemble into 2 parts, QC (blue sticks) and ML (orange sticks). We only use very small number of QC calculations (several hundreds) to achieve the high precision.[1]





38 atoms and 30 excited states
Only 100 training points is sufficient! [1]

In ML-NEA got rid of arbitrary parameters by fixing them to optimal values![1]
N_p = 50000 (number of points in ensemble)
δ = 0.01 eV (broadening parameter)

Optimal number of training points can be automatically determined from convergence of ML validation error! [1]

References

[1] B.-X. Xue, M. Barbatti, P. O. Dral, *J. Phys. Chem. A* **2020**, 124, *35*, 7199–7210
[2] S. Bai, R. Mansour, L. Stojanovic, J. M. Toldo, M. Barbatti, *J. Mol. Model.* **2020**, *26*, 107
[3] R. Crespo-Otero, M. Barbatti, *Theor. Chem. Acc.* **2012**, *131*, 1237
[4] https://en.wikipedia.org/wiki/Nuclear_ensemble_approach
[5] P. O. Dral, *J. Comput. Chem.* **2019**, *40*, 2339–2347
[6] P. O. Dral, B.-X. Xue, F. Ge, Y.-F. Hou, *MLatom: A Package for Atomistic Simulations with Machine Learning*, version 1.2. Xiamen University, Xiamen, China, **2013–2020**. http://MLatom.com.
[7] http://www.newtonx.org/