Estimating Spectral Bandwidths with the Nuclear Ensemble Approach

Andrea Cocut[a] **and Marin Sapunar**[b]

andrea.cocut@chem.pmf.hr

[a]*Department of Chemistry, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia*
[b]*Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia*

The Nuclear Ensemble Approach (NEA) is a widely used method for simulating electronic absorption spectra by computing excitation energies and oscillator strengths of several electronic states across an ensemble of nuclear geometries. The absorption cross-section is then obtained by convolving each sampled point with kernel functions. In practice, Gaussian kernel functions [1] are typically employed, where the bandwidth parameter critically influences the smoothness of the resulting density estimate. A bandwidth that is too small may lead to overfitting, capturing noise rather than the true underlying distribution, while an excessively large bandwidth can cause underfitting. In the present literature [2], the rule-of-thumb bandwidth estimators are mostly used, with Silverman's rule [3] being the most common.

A key challenge in NEA simulations is the computational cost associated with large sample sizes, as each geometry requires expensive single-point electronic structure calculations. In this work, we adopt a more rigorous approach for selecting the bandwidth parameter by employing cross-validation to estimate the optimal value of the bandwidth parameter for a given sample. Our goal is to minimize the required sample size while preserving spectral accuracy, thereby reducing computational costs. To facilitate this, we developed a custom implementation based on the Fast Fourier Transform Kernel Density Estimation (FFTKDE) algorithm from the KDE.py library. We compared our method with Silverman's estimator and experimental data, assessing performance across total spectra as well as spectra decomposed into either adiabatic or diabatic (obtained using wave function overlaps [4]) state contributions. Additionally, we determined the uncertainty estimate with bootstrap resampling and conducted a thorough analysis of how sample size and the number of cross-validation splits impact the bandwidth estimation.

References:

[1] R. Crespo-Otero, M. Barbatti, *Chem. Rev.* **118** (2018) 7026-7068.

[2] (a) Š. Sršen, J. Sita, P. Slavíček, V. Ladányi, D. Heger, *J. Chem. Theory Comput.* **16** (2020) 6428–6438; (b) Š. Sršen, P. Slavíček, *J.Chem. Theory Comput.* **17** (2021) 6395-6404.

[3] B. W. Silverman, *Density Estimation for Statistics and Data Analysis; Monographs on Statistics & Applied Probability*, Chapman & Hall, 1986.

[4] M. Sapunar, T. Piteša, D. Davidović, N. Došlić, *J. Chem. Theory Comput.* **15** (2019) 3461–3469.