Infrared spectra of neutral polycyclic aromatic hydrocarbon by machine learning

G. LAURENS[1], M. RABARY[1], J. LAM[2], D. PELÁEZ[3], A.-R. ALLOUCHE[1]

[1]*Institut Lumière Matière, UMR5306 Université Lyon 1-CNRS, Université de Lyon, 69622 Villeurbanne Cedex, France*

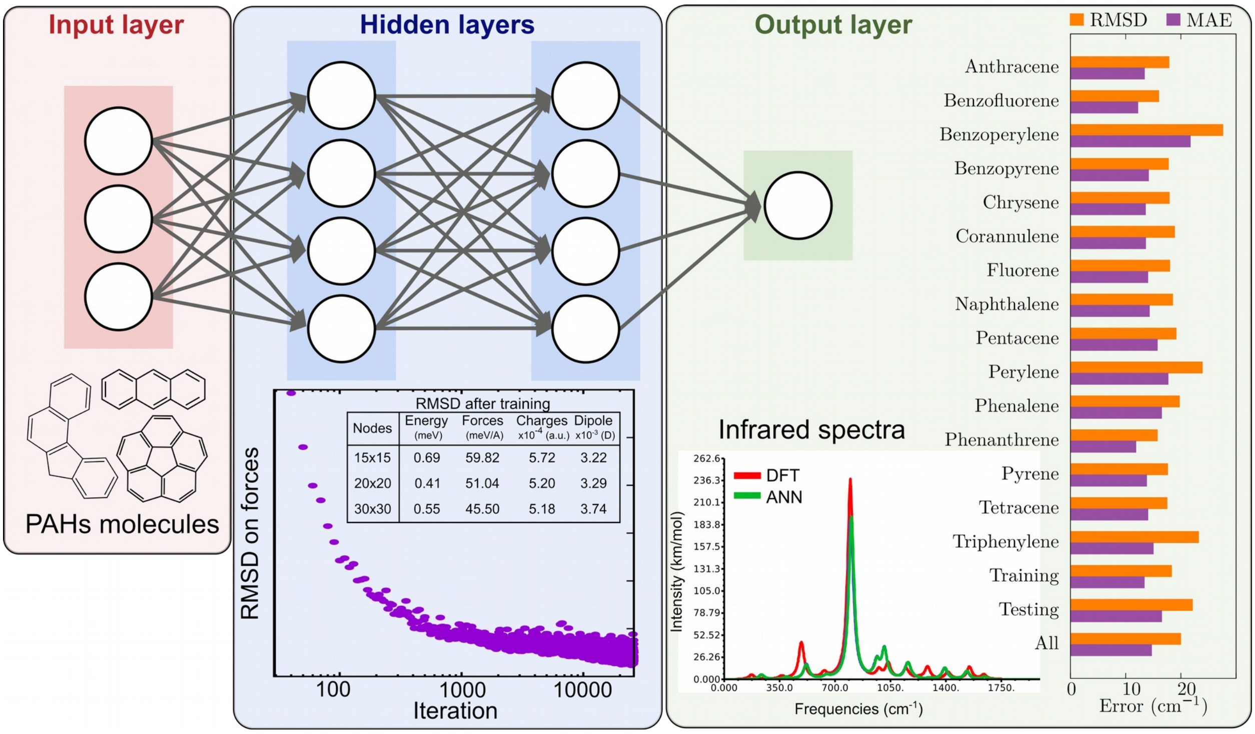
[2]*Center for Nonlinear Phenomena and Complex Systems, Code Postal 231, Université Libre de Bruxelles, Boulevard du Triomphe, 1050 Brussels, Belgium*

[3]*Institut de Sciences Moléculaires d'Orsay, UMR 8214, Université Paris-Sud - Université Paris-Saclay, 91405 Orsay, France*

Polycyclic aromatic hydrocarbons (PAHs) are organic molecules made of two or more aromatic rings. Interests on this molecule family lie on versatile fields such as interstellar chemistry [1], health issues [2], and dimerization in soot nucleation [3]. In particular, infrared spectra are measured to distinguish these molecules within interstellar medium or flames. In order to compute vibrational frequencies, numerous theoretical studies employ either quantum calculation methods, or empirical potentials based on ReaxFF interactions, but it remains difficult to combine the accuracy of the first approach with the computational cost of the second. The recent development of machine learning methods in computational science brings solutions to overcome these difficulties. In this work, machine-learning techniques were employed to obtain: (a) an artificial neural network (ANN) potential energy surface and (b) a dipole mapping based also based on a neural network architecture. Altogether, it enables us to compute infrared spectra of PAHs molecules, including anharmonic effects.

Our ANNs were trained using a database including 8 863 energies and 735 447 forces of 11 PAHs molecules, with different neuron numbers on two hidden layers. Then, we compared the computed vibrational frequencies of 17 PAHs molecules with those obtained from our DFT calculations. In overall, frequency errors, namely root-mean-square deviation (RMSD) and mean averaged error (MAE), are, respectively, around 20 and 15 cm^{-1} for all the ANN systems and for all the molecules, when compared to DFT calculations. Increasing the number of nodes per layer from 15 to 20 and 30 shows a decrease of the frequency RMSD from 23.5 to 21 and 20 cm^{-1}, respectively.

In addition to approach the accuracy of the DFT calculations, the computational cost is also largely improved. By taking our largest molecule, *i.e.* corannulene molecule, while the calculation of the fundamental frequencies using DFT lasts more than 142 days, the same calculation using our trained ANNs systems last 3h40mns.



RMSD after training

Nodes	Energy (meV)	Forces (meV/A)	Charges x10^{-4} (a.u.)	Dipole x10^{-3} (D)
15x15	0.69	59.82	5.72	3.22
20x20	0.41	51.04	5.20	3.29
30x30	0.55	45.50	5.18	3.74

[1] Sandford *et al.*, Astrophys. J. Suppl. Ser. 205(1), 8 (2013). DOI 10.1088/0067-0049/205/1/

[2] Downward *et al.*, Environ. Sci. Technol. 48(24), 14632 (2014). DOI 10.1021/es504102

[3] Mercier *et al.,* Phys. Chem. Chem. Phys. 21(16), 8282 (2019). DOI10.1039/C9CP00394