Development of a Machine Learning Model for Predicting Photophysical Parameters of Organic Molecules

J M Araujo[1]

[1]*Photonics Group, São Carlos Institute of Physics, São Carlos, Brazil*
Contact Email: joao_araujo@usp.br

Prediction models based on machine learning have been applied in a wide diversity of fields and tasks. Through a loss function to be minimized by the model, it can iteratively map a characteristic that one wants to predict or understand based on other known characteristics. In recent years, machine learning-based modeling has also been used in the field of molecular physics and have proven to be extremely accurate in predicting molecular characteristics. Currently, to determine the dipole moment of an organic molecule one can either perform experimental measurements or use computationally expensive quantum chemistry calculation for theoretical determination. Machine learning models trained for this purpose can determine this value with relative precision within seconds, being an excellent option as they require low computational capacity. The challenge, in this case, is to efficiently translate a molecule into a data structure that serves as input for such algorithms. This project is an introduction to the topic of predicting molecular characteristics using machine learning, in which the main goal was to replicate a machine learning model to predict certain photophysical parameters. Therefore, an initial engagement with the QM9 database [1] was necessary. The QM9 dataset contains over 134-kilo molecules with various physicochemical properties calculated through quantum chemistry calculations. For a better understanding of the data, key theoretical concepts were studied, such as HOMO-LUMO energy gap and the dipole moment of a molecule. Additionally, the SMILES [2] notation was introduced to extract the necessary information and, more importantly, to transform the QM9 data into inputs for the regression algorithm. The model developed was able to generate various results, which were subsequently compared to the values found in the literature. Based on these results, the accuracy of the model was determined and found to be suitable for the initial project proposal. Furthermore, the model shows potential for improvement as the project moves forward.

References

[1] N R Ramakrishnan, P O Dral, M Rupp and O A von Lilienfeld, Sci. Data **1**, 140022 (2014), DOI: 10.1038/sdata.2014.22

[2] D Weininger, J. Chem. Inf. Comput. Sci. **28**, 31 (1988)