J. Aires-de-Sousa[1]

MACHINE LEARNING WITH LARGE DATASETS OF QUANTUM CHEMISTRY CALCULATIONS

[1] LAQV and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal;

The rapid access to intrinsic physicochemical properties of molecules is highly desired for large scale data mining explorations, e.g., for the discovery of new materials and drugs, toxicity risk assessment, or mass spectrum prediction in metabolomics. Data can be obtained by quantum chemistry calculations, which provide increasingly accurate estimations of several properties, but are too computationally expensive for large scale uses. Even though, high-throughput quantum chemistry calculations are being performed in projects employing enormous computational resources [1,2]. A big data scenario can be envisaged in which computational analytic techniques extract innovative knowledge from the large volumes of data produced by quantum calculations, so that they can be predicted 5–6 orders of magnitude faster in new situations. Studies have been reported in which machine learning algorithms were trained with thousands of data points to predict *ab initio*- or DFT-calculated properties (molecular, bond, or atomic properties) [3,4]. Our lab trained machine learning algorithms with >8,000 bond energies, and >35,000 atomic charges, calculated by DFT, to enable extremely fast predictions [5,6].

In this lecture, machine learning of condensed Fukui indices is presented firstly. From a chemoinformatics perspective, modeling Fukui indices presents a specific challenge – the competition of all the atoms in the molecule for a charge, so that atoms in the same substructure can exhibit very different Fukui indices in different molecules. The problem was approached either as a Random Forests regression/classification, and as a ranking of atom types with the Bradley-Terry model.

A second project is also presented involving machine learning of bond properties calculated by DFT for ca. 150,000 covalent bonds, covering a large range of molecular sizes and chemical elements. Most of the currently available QSAR/QSPR algorithms, molecular descriptors, and software have been typically designed for data sets at least 10 times smaller than this. The results obtained with various strategies to handle large data sets are presented, namely for the selection of training sets, bond descriptors, and ML algorithms.

1. Hachmann J. et al. *J. Phys. Chem. Lett.*, 2011, **2**: 2241-2251.
2. Jain A. et al. *APL Materials*, 2013, **1**: 011002; doi: 10.1063/1.4812323.
3. Hansen K. et al. *J. Chem. Theory Comput.*, 2013, **9**: 3404-3419.
4. Raj B.K., Bakken G.A. *J. Comput. Chem.*, 2013, **34**: 1661-1671.
5. Qu X., Latino D.A.R.S., Aires-de-Sousa J. *J. Cheminform.*, 2013, **5**: 34.
6. Zhang Q.-Y. et al. *Chemom. Intell. Lab. Syst.*, 2014, **134**: 158-163.

Financial support from Fundação para a Ciência e a Tecnologia (FCT), Portugal, under Project PEst-C/EQB/LA0006/2013 is acknowledged.