# Prediction of *ab initio*-quality molecular spectra using machine learning

Eric Berquist

November 29, 2016

Machine learning is one of the primary tools used to process and interpret "big data" which is intractable using manual or brute-force interpretation. One area in which machine learning can be applied is regression modeling, where reference (training) data with a known relationship between some independent and dependent variables is used to predict dependent variables for unknown data. The simplest form of regression would be linear regression, where a linear relationship is assumed: $f(x) = mx + b$, where $x$ is the independent variable and $f(x)$ is the dependent variable.

I will use machine learning techniques to build regression models capable of predicting *ab initio* calculated molecular spectra, using only atomic coordinates as input. The application of machine learning to quantum chemical calculations has already been performed by Alán Aspuru-Guzik and O. Anatole von Lilienfeld. On a technical level, my focus is on extending the von Lilienfeld's work from scalar quantities ($\epsilon_{\mathrm{HOMO}}, \epsilon_{\mathrm{LUMO}}, \Delta\epsilon, \langle r^2 \rangle, \mu_{\mathrm{norm}}, \alpha_{\mathrm{iso}}, U_0, E_{\mathrm{ZPVE}}$) to quantities that must be represented using vectors or matrices, such as molecular spectra. Work by von Lilienfeld has also been done to find optimal representations of molecular data in models (the input $x$); one such representation is the Coulomb matrix, not to be confused with the Coulomb matrix in quantum chemistry.

The first test will use a subset of the existing GDB-9 database, which consists of the "subset of all neutral molecules with up to nine atoms (CONF), not counting hydrogen" (about 134K) with thermodynamic properties calculated at the B3LYP/6-31G(2df,p) level. In order to calculate the zero-point vibrational energy, $E_{\mathrm{ZPVE}}$, (harmonic) vibrational frequencies are required, so they are already present in the database. Using this publicly available data and the open-source Python library scikit-learn, I will see if current approaches to regression modeling can be extended to higher dimensional quantum chemical data.

Future work may take one of three clear directions. The first direction is use of more accurate quantum chemical models, such as modern density functionals known to perform well for thermochemistry ($\omega$B97M-V/def2-TZVP). It has not been studied if using higher-quality methods reduces the size of the training set required. The second direction is to test other regression models. [Other kinds of linear regression (not kernel ridge regression) or non-linear regression.] The third direction is to predict other kinds of molecular spectra; of particular interest are non-linear optical properties (hyperpolarizabilities) and magnetic spectroscopies (NMR and EPR).