Thesis Project description

Thesis topic

Web-based service for training data generation in trajectory-based molecular machine learning

Objective of the thesis

Virtual design of new materials and drugs is one of the currently heavily investigated fields in which the use of computational methods promises to strongly support science discovery. One potential part of this design process is the prediction of quantum chemical properties (e.g. ground or excited state energies) of molecules that are part of the new materials or drugs. Often, we are interested to do such a prediction for trajectories of molecules, hence sequences of identical molecules whose atoms are however slightly shifted due to some external process. Given a subset of a trajectory of molecules and calculated properties for each molecule in this subset trajectory, the objective is then to find a machine learning model that can predict these properties for any remaining molecule on that trajectory or other similar trajectories.

One part of the overall process to carry out this type of trajectory-based molecular machine learning starts from a given trajectory of molecules, which is then split into individual molecule descriptions, which are passed to a quantum chemistry calculation software, e.g. Orca [1]. This software then computes for each molecule respective properties (e.g. energies). The sequence of these properties become that labels of a training data set.

The objective of this project is two-fold. First, a web-based front-end is developed, where a user can upload a (compressed) file with a trajectory of molecules. Within the front-end, the user can visualize the molecules along the trajectory. In addition the user can select, which property is calculated. With a simple “Compute” button, the actual calculation process is issued, whose process is monitored in the web interface. Later, the user can download the calculated properties as a single file.

Second, a text-file configurable back-end is developed, which implements the functionalities of the front-end. The main task of the back-end is specifically to take care of the scheduling, distribution and collection of compute tasks.

Sources

[1] Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger, "The ORCA quantum chemistry program package", J. Chem. Phys. 152, 224108 (2020)

Deliverables

D1. Design of front- and back-end, including requirements catalog and GUI draft.

D2. Prototype implementation of front- and back-end.

D3. Verification of system functionality for exemplary data.

D4. All source codes for D1 – D3 are provided including a reasonable amount of comments in the source code.

D5. Thesis based on the BSc/MSc thesis template.