Chemistry data mining to discover molecular interactions

M.Sc. project 30/45/60 credits for students in Computer Science, Physics, Chemistry, Biotechnology or similar.

Prof. David van der Spoel, Department of Cell and Molecular Biology, Uppsala University
< david.vanderspoel@icm.uu.se >

Abstract
In this project you will use large databases containing structure and energies of molecules with the aim to design simple analytical functions of the atomic coordinates that reproduce these energies. Applications of this work may be in materials modeling or drug design.

Introduction
Databases with physicochemical properties of molecules, liquids and solids have existed for decades, gathering invaluable experimental results from more than hundred years of hard work all over the world. More recently, this has been complemented with quantum chemistry calculations such that there now is a large body of data available, for instance for building predictive physical models [1]. These models are used routinely in e.g. drug-design studies, recently with a focus on COVID-19 [2], however, it is well recognized that the models are inadequate and that development of improved models has stagnated [3]. My research group focuses on systematic design of force field models [1] based on both the experimental legacy and the large amounts of quantum chemistry data that has recently become available [4,5].

In this project data will be extracted from the ANI-1 database [4] that contains structures and energies for molecules that are not in their equilibrium (minimum energy state). Because of this, there is information on the effect of bond-stretching, angle bending and close contact on the energies of compounds which can be used to develop and refine the physical models (Figure 1).

Figure 1 From data to discovery using Alexandria.

Project description
First, the compounds will be processed to determine fixed atom types for all atoms and to detect the chemical bonds (based on minimum energy structures). This is a standard technique that is implemented in OpenBabel [6]. Second, the high energy structures will be analyzed using machine learning tools to “discover” relatively simple classical energy functions that reproduce the energies for all compounds. In a first ansatz, this can be done using linear regression analysis which will give a good idea of what the energy function should look like. The next step would be to use analytical functions to reproduce the
quantum chemical energies. Intriguingly, there is no known “best” analytical function to describe the energies (see references in [1]) which is why machine learning based on this large data set [4] and also others that we are developing in my group [5] may lead to novel findings for the benefit of drug design and molecular modeling in general. The proposal here is part of a Ph.D. project in my research group and both the Ph.D. student and PI will be heavily involved in this project alongside the student.

Requirements for the student
You have to have experience in Python programming, using command-line tools on Linux computers and preferably some insight into organic molecules. The dataset you will be using is many Gigabytes large which mean efficient data management will be crucial.

COVID-19 considerations.
Much of the work can be done remotely even though that is sub-optimal for learning, in particular in the start of the project. There are weekly group meetings that you are expected to participate in, either in person or via zoom.

References