

Machine learning methods for the prediction of physicochemical and biological properties of chemical compounds

Project description

Advanced machine learning algorithms can help deal with various chemical problems, including the prediction of physicochemical and biological properties of compounds. The generation of large-scale datasets with detailed information of these compounds has allowed building up models to predict the values of features such as the retention time or collision cross-section. For instance, these values are valid for the annotation of unknown compounds detected in metabolomics studies.

In this project, different machine learning algorithms (e.g. neural networks, random forest, gradient boosting) will be used to evaluate their modelling and prediction capacities after optimising each relevant parameter for the model. Benchmark datasets will consist of several compounds structures (from which molecular descriptors will be obtained) and their related physicochemical or biological property.

Tasks to be performed by the candidate

- Application of different algorithms to model physicochemical properties of compounds found in public datasets.
- Evaluation of the performance of these algorithms considering all the dataset and selected subsets by families of compounds
- Prediction of these properties in a new set of compounds.

Tutor

Joaquim Jaumot (joaquim.jaumot@idaea.csic.es)

