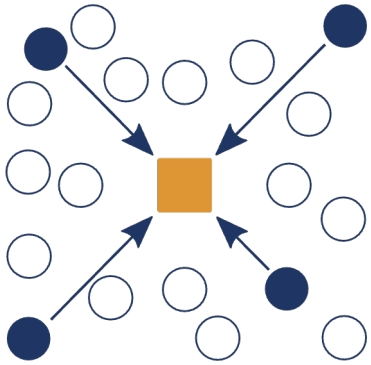
Definition and Overview

Introduction

Machine Learning



Foundations | Statistical modelling

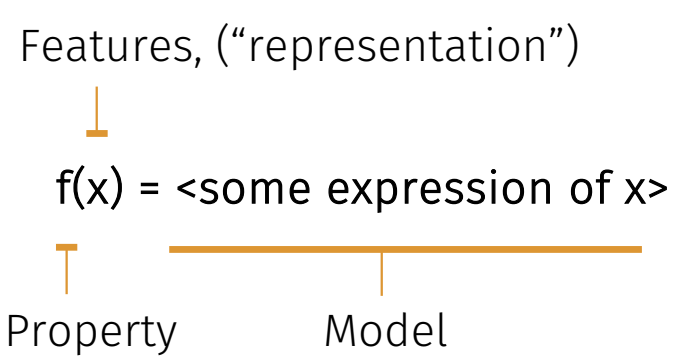
Accuracy | Systematically improvable through data and training

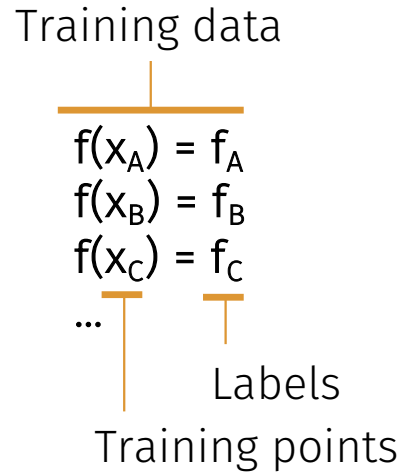
Specialty | Universal, scale-bridging, data-driven approach

Limitation | Requires training data, no black box

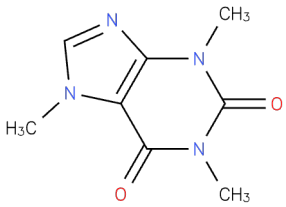
ML = Mapping compound to property using some explicit results.

Nomenclature





Quantum chemistry picture



$$\hat{\mathrm{H}}(\mathbf{R}_I, Z_I, n, \sigma) \rightarrow \Psi$$

Features

Numbers of electrons

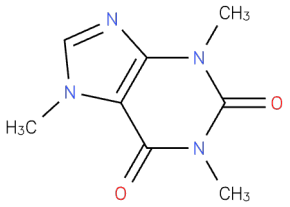
Electronic spin

Nuclear charges

Nuclear coordinates

Model

Representations



0 1 0 0 1
1 1 0 1 1
1 0 1 1 1
0 1 0 0 0

Graph
Vector, Matrix, ...
Bit field
String
...

Traditional methods

Every computational chemistry model comes from careful neglect of physical effects.

DFT accuracy

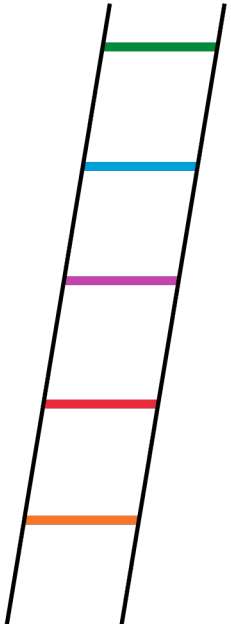
Double hybrid functionals (DH-GGA)
$\rho(\mathbf{r})$, $\nabla\rho(\mathbf{r})$, $\nabla^2\rho(\mathbf{r})$, HFX, MP2

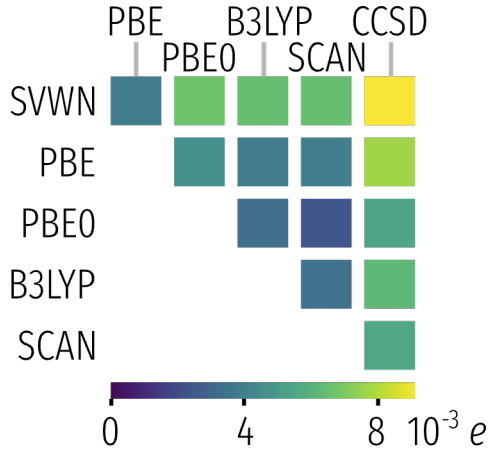
Hybrid functionals (GH-GGA)
$\rho(\mathbf{r})$, $\nabla\rho(\mathbf{r})$, $\nabla^2\rho(\mathbf{r})$, HFX

meta-GGA (mGGA)
$\rho(\mathbf{r})$, $\nabla\rho(\mathbf{r})$, $\nabla^2\rho(\mathbf{r})$

Generalized Gradient Approximation (GGA)
$\rho(\mathbf{r})$, $\nabla\rho(\mathbf{r})$

Local Density Approximation (LDA)
$\rho(\mathbf{r})$





Summary Definition and Overview

- Machine Learning is statistical modelling
- Re-use of previous information
- Traditional methods (quantum chemistry, QC) are still used as reference
- Scaling with system size of QC unfavourable
- QC does not always agree with itself
- Features = arguments of the learned function
- Labels = results of the learned function

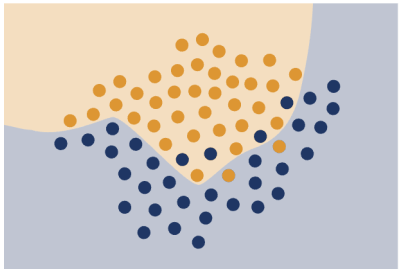
Problem Classes

Problem classes

Supervised Learning
(with labels)

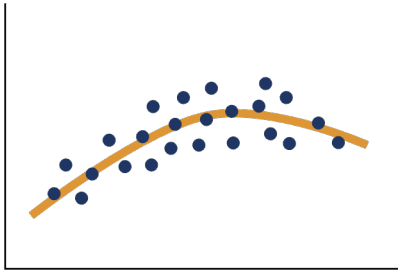
Unsupervised Learning
(without labels)

Classification



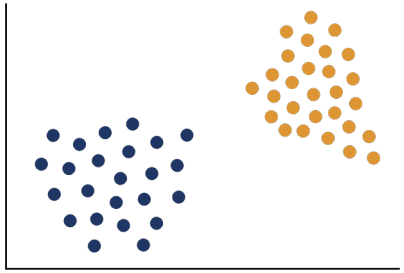
- Stability
- Reaction mechanisms
- ...

Regression



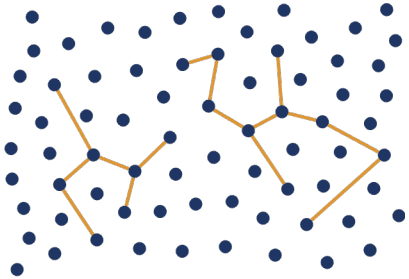
- Reaction barriers
- Geometries
- ...

Clustering



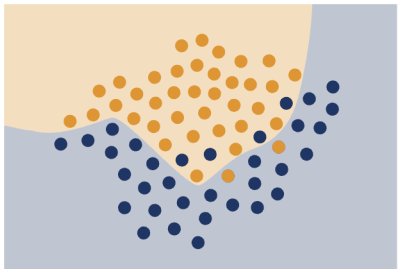
- Dimensionality reduction
- Find mechanisms
- ...

Association



- Find mechanisms
- Detect networks
- ...

Classification



Challenges

- Imbalanced frequencies
- Irrelevant features
- Overlapping classes
- Non-linear data
- High-dimensional data

Approaches

- One vs All: n classifiers
- One vs One: n*(n-1) classifiers

Common algorithms

- Decision trees / Random forest
- K-nearest neighbours
- Neural networks