Machine Learning for Materials and Chemistry

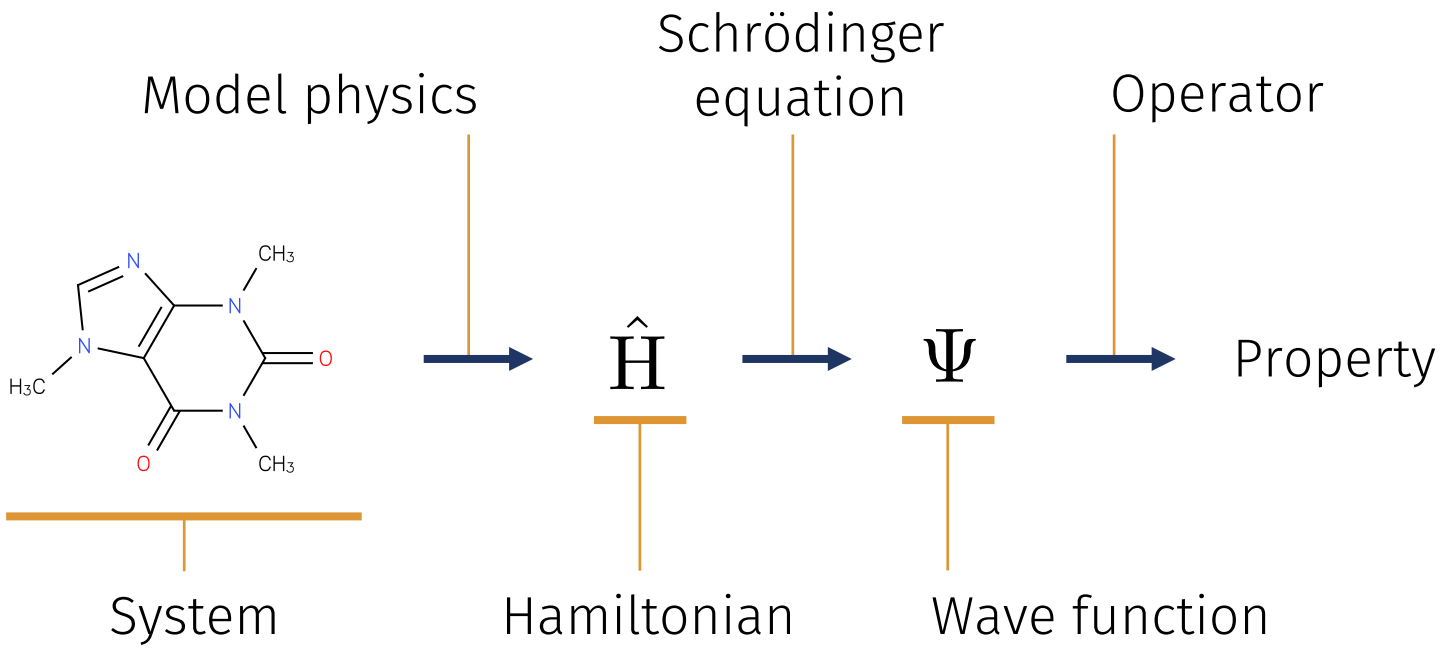
Guido Falk von Rudorff, University of Kassel

 ferchault

 @ferchault

 nablachem.org

Which problem do we try to solve?



Computational complexity

Depends on system size n

DFT	$O(n^3)$
HF	$O(n^4)$
MP2	$O(n^5)$
MP3	$O(n^6)$
MP4	$O(n^7)$
CCSD	$O(n^6)$
CCSD(T)	$O(n^7)$
CCSDT	$O(n^8)$
CCSDTQ	$O(n^{10})$
FCI	$O(n!)$

What is this course about?

Foundation of machine learning
Rules of the game

Representations
Mathematical model of molecules/materials?

Regression
Quantified prediction

Classification
Grouping, dimensionality reduction

Methods
KRR, SVM, GPR, ANN, ...

Model efficacy
Validation

Examples / challenges
Build experience and intuition

Where can I use this knowledge?

Direct

- Molecular / Materials design	Guide experiment
- Bio / Medical applications	Understand chemical process
- Method development	Widen applicability

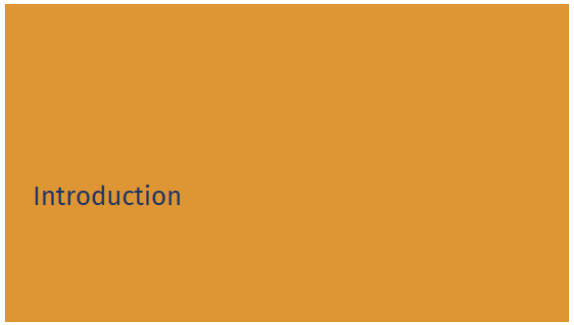
Indirect

- Data science	Extract and manage large databases
- Research in general	Strategies and methods

Lecture

Questions anytime
lecture, moodle, vonrudorff@uni-kassel.de, …
Only pre-recorded content, no live recording
Slides and notes as PDF **before** the lecture https://nablachem.org/lectures/mmc/

Related: Computational Chemistry (winter term)

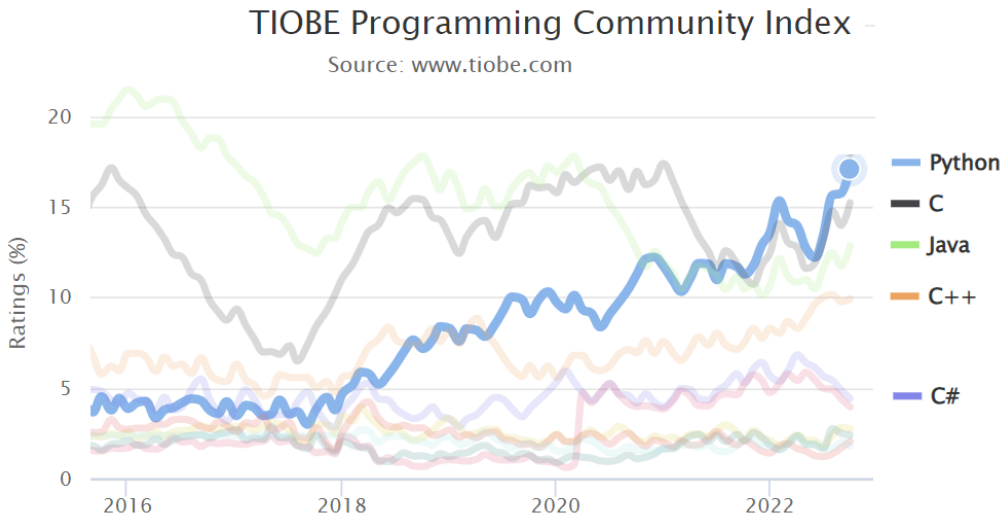


- Sections of ~30 min
- Largely self-contained
- Ask questions at latest then: new train of thought

- One slide at end of section
- Key take-aways

Exercise

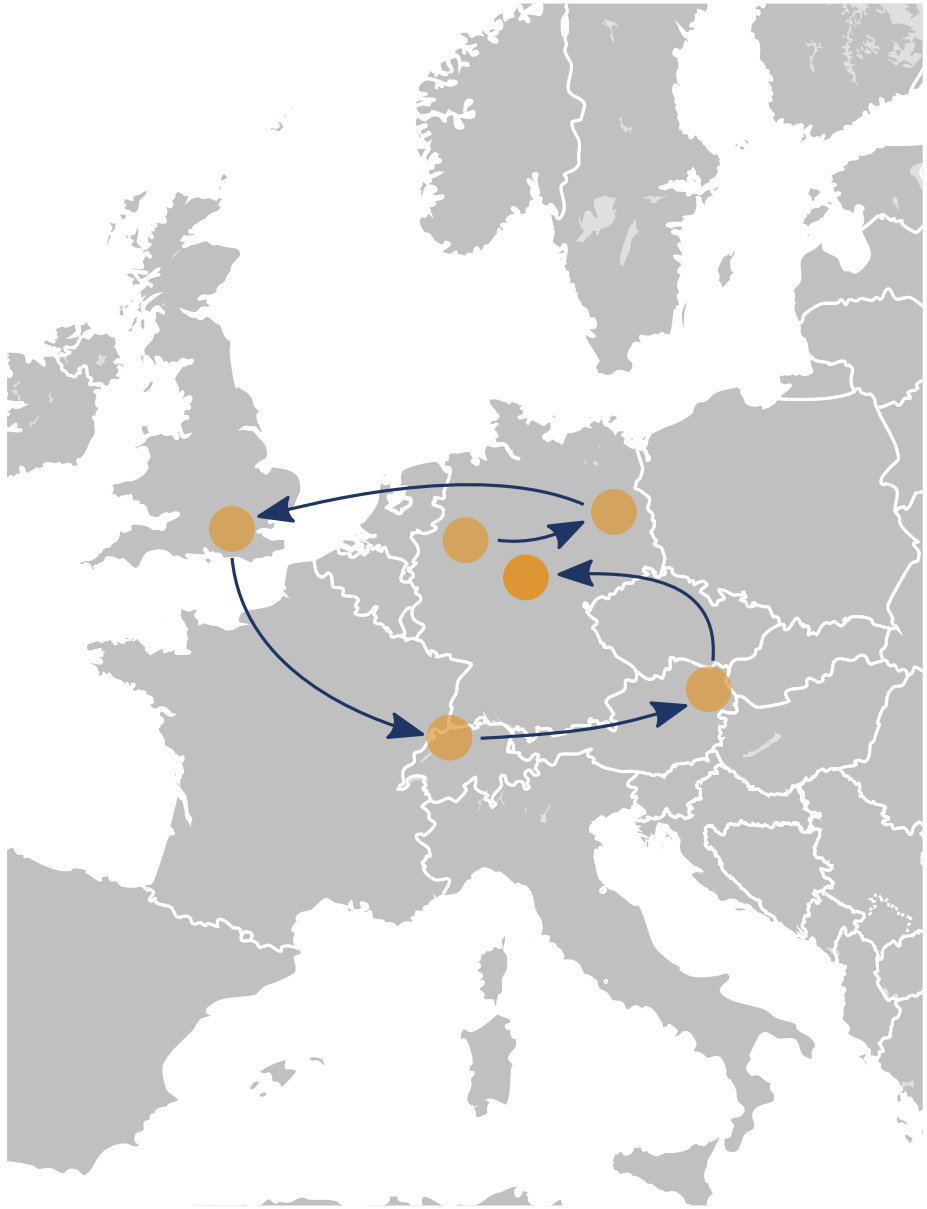
Python-based: the language of data science and glue code



Weekly assignments:

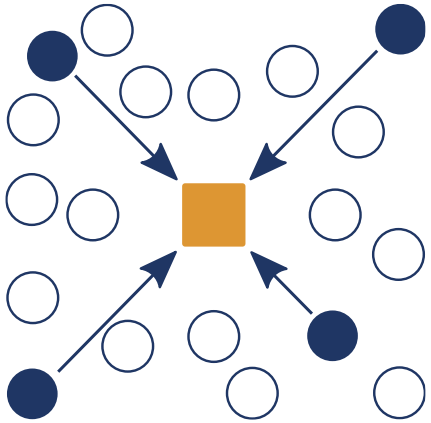
- At first: programming
- Later: machine learning
- End: modern research problems
- Typically: 2 regular tasks + one harder one if you consider research in this area

About me

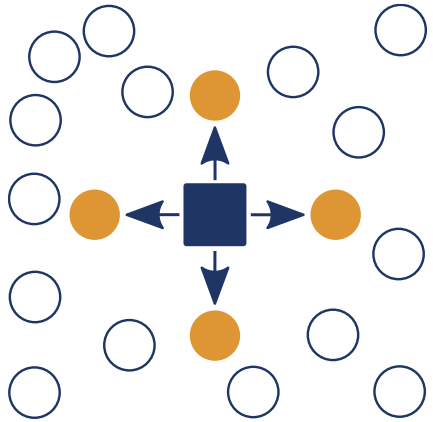


BSc/MSc Physics	Berlin	Force fields
PhD Physics	London	Quantum chemistry
PostDoc	Basel, Vienna	Machine Learning & Alchemy

Machine Learning

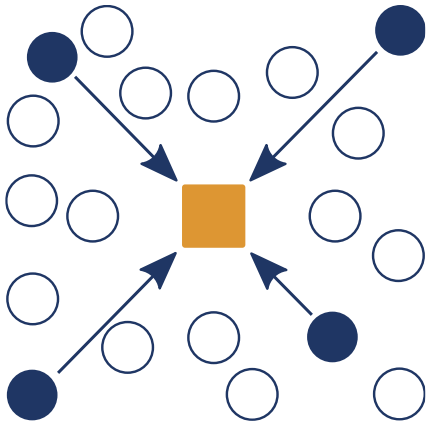


Quantum Alchemy



Definition and Overview

Machine Learning



Foundations | Statistical modelling

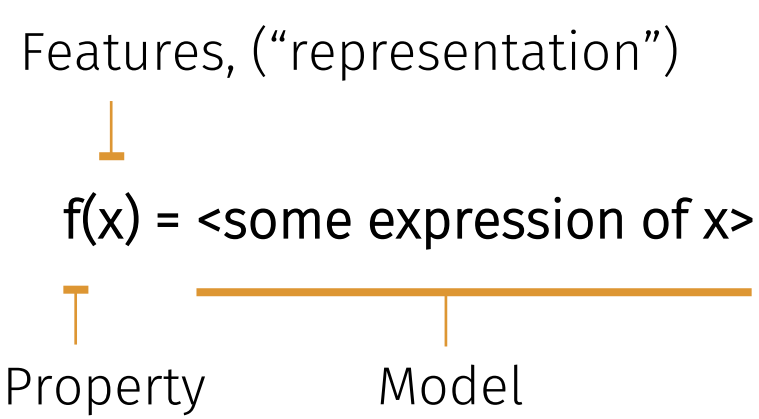
Accuracy | Systematically improvable through data and training

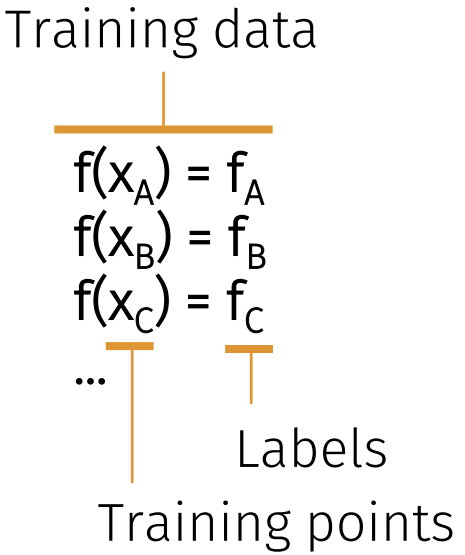
Specialty | Universal, scale-bridging, data-driven approach

Limitation | Requires training data, no black box

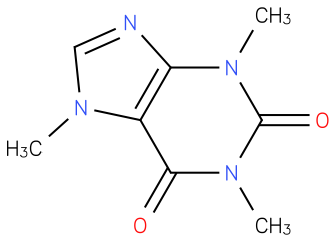
ML = Mapping compound to property using some explicit results.

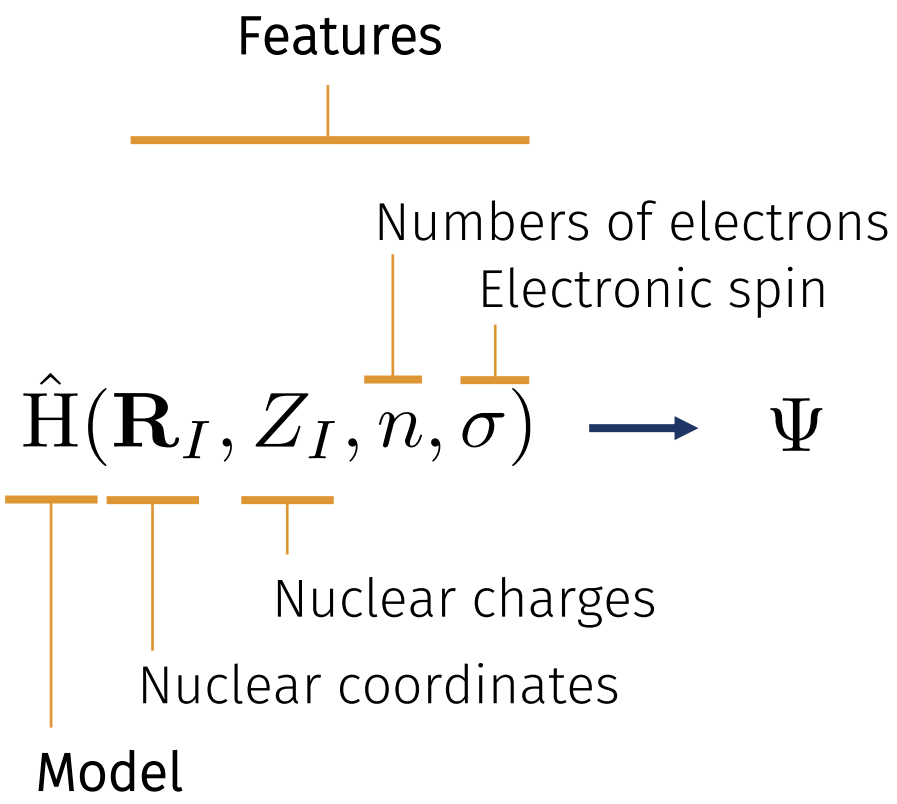
Nomenclature



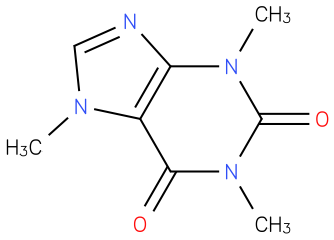


Quantum chemistry picture





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

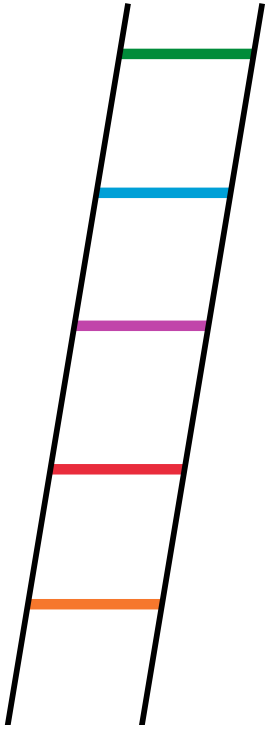
Double hybrid functionals (DH-GGA)
$\rho(\mathbf{r})$, $\nabla\rho(\mathbf{r})$, $\nabla^2\rho(\mathbf{r})$, HFX, MP2

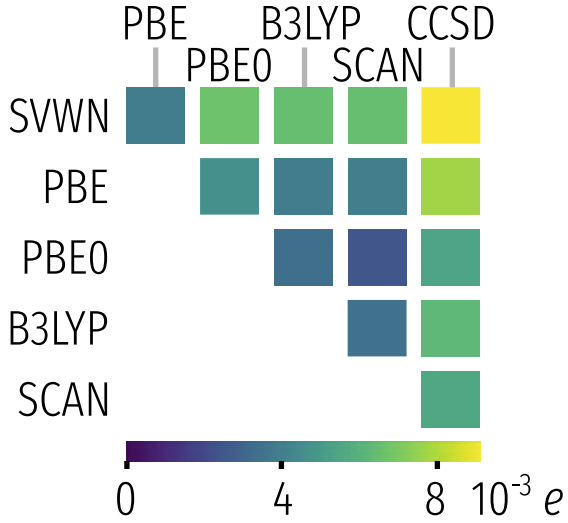
Hybrid functionals (GH-GGA)
$\rho(\mathbf{r})$, $\nabla\rho(\mathbf{r})$, $\nabla^2\rho(\mathbf{r})$, HFX

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

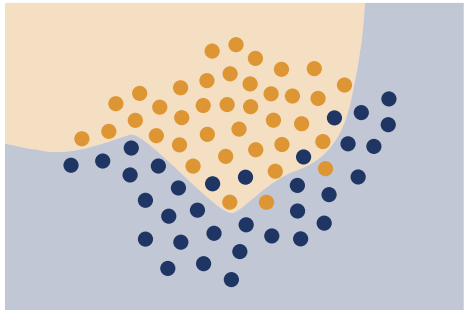
Problem Classes

Problem classes

Supervised Learning
(with labels)

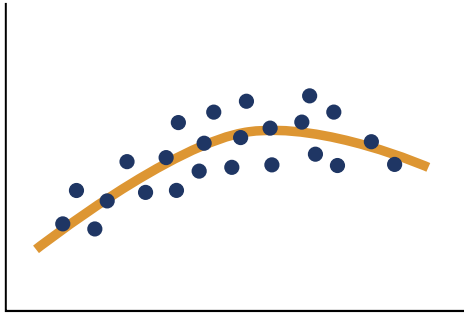
Unsupervised Learning
(without labels)

Classification



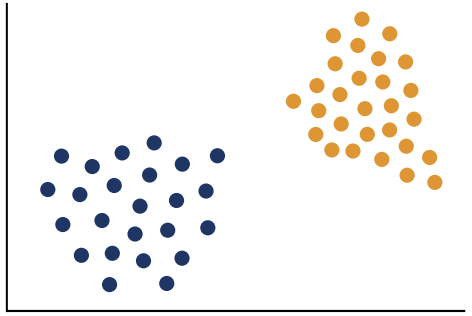
- Stability
- Reaction mechanisms
- …

Regression



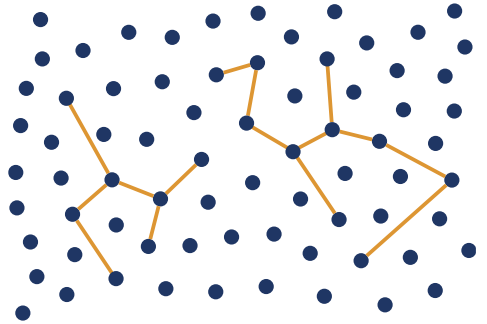
- Reaction barriers
- Geometries
- …

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