FNSNF
Swiss National Science Foundation

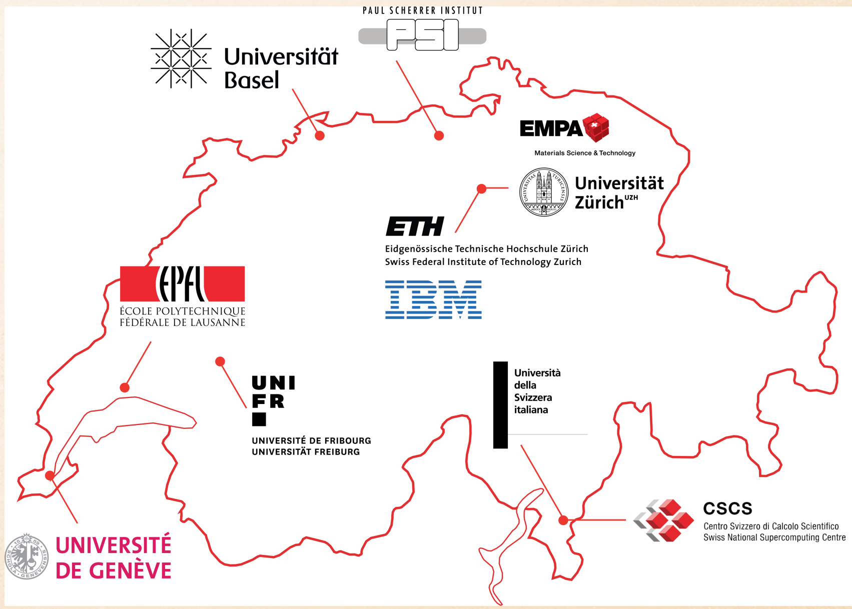
UNI BASEL

MARCO DI GENNARO

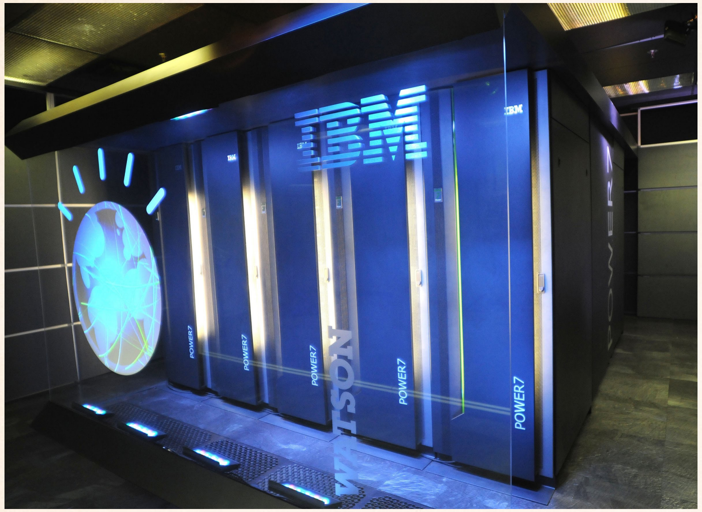
MATERIALS PROPERTIES FROM MACHINE LEARNING

MARVEL

NATIONAL CENTRE OF COMPETENCE IN RESEARCH

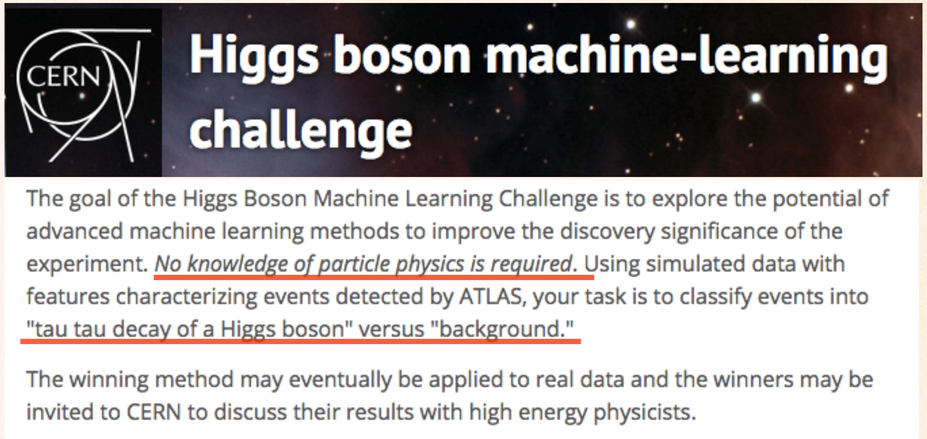


Machine Learning: extract knowledge from data



Watson

- winner of 1M$ in Jeopardy
- decision assistant in lung cancer treatments


CERN

Higgs boson machine-learning challenge

The goal of the Higgs Boson Machine Learning Challenge is to explore the potential of advanced machine learning methods to improve the discovery significance of the experiment. *No knowledge of particle physics is required.* Using simulated data with features characterizing events detected by ATLAS, your task is to classify events into "tau tau decay of a Higgs boson" versus "background."

The winning method may eventually be applied to real data and the winners may be invited to CERN to discuss their results with high energy physicists.

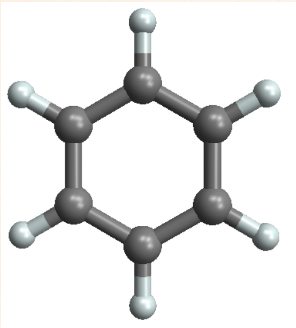
winner: *Gábor Melis* - Diósd, Hungary

- digital security,
- education,
- trade prices prediction,
- big data analysis,
- animal protection,
- …

FAST, ACCURATE, SYSTEMATIC

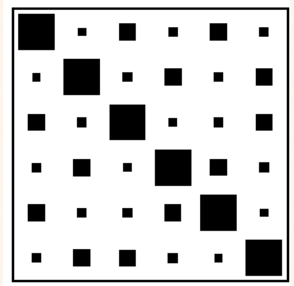
MACHINE LEARNING + QUANTUM MECHANICS = FAST MATERIALS PROPERTIES

Database



M

Representation

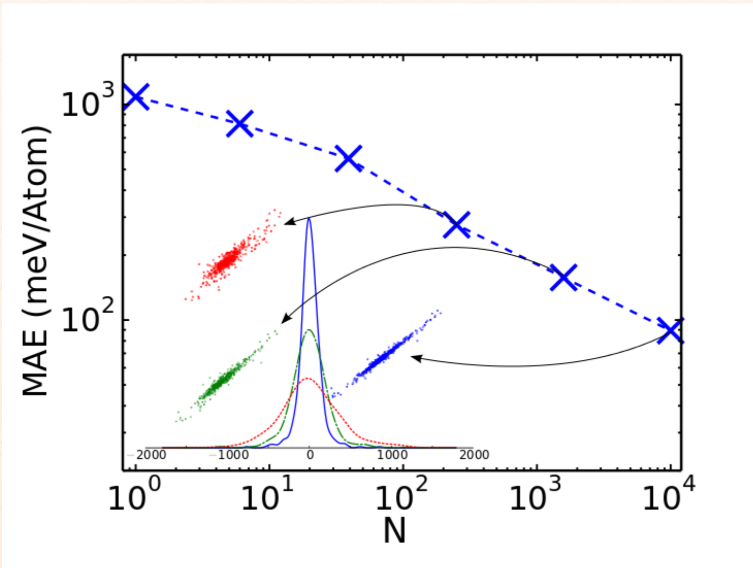


D

Algorithms
regression,
minimization

L

Prediction
Learning curve



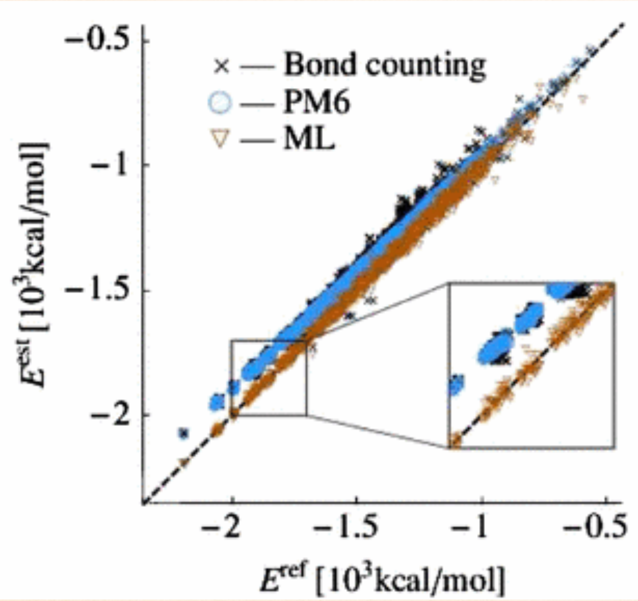
Test on
out-of-Database
population

PRL 117, 135502

DATABASE = TRAIN + TEST

from experiments or high throughput ab-initio, …

Train: ~7k molecules
Test: 1k molecules



PRL 108, 058301

ALGORITHMS

$$E^{\mathrm{est}}(\mathbf{M}) = \sum_{\mathbf{i=1}}^{\mathbf{N}} \alpha_{\mathbf{i}} \exp\left[-\frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}}\mathbf{d}(\mathbf{M},\mathbf{M_i})^{\mathbf{2}}\right]$$

M = Representation
d = Euclidean distance
α = weights: transferable and adaptable

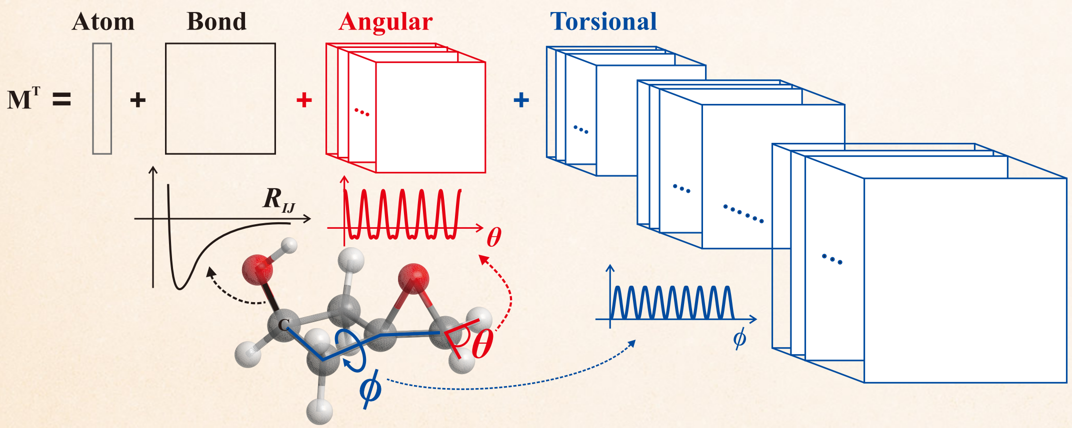
$$\min_{\alpha} \sum_i (E^{\mathrm{est}}(\mathbf{M}_i) - E_i^{\mathrm{ref}}) + \lambda \sum_i \alpha_i^2$$

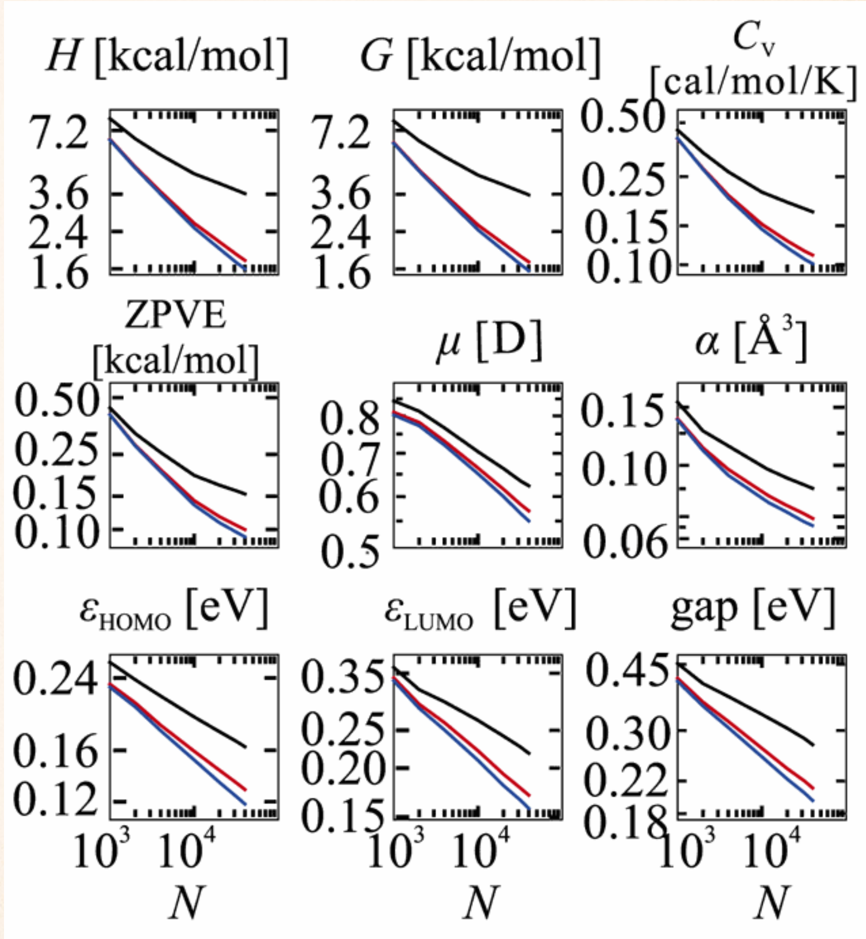
λ = regularization parameter

REPRESENTATION

more physics (chemistry), quicker learning

- complete, non-degenerate,
- compact, unique,
- efficient,
- simple





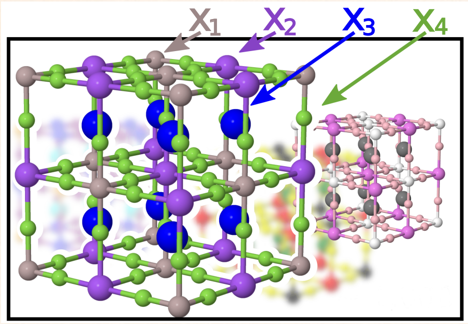
- 134k QM9 molecules
- BAML representation

IJQC 115, 1094

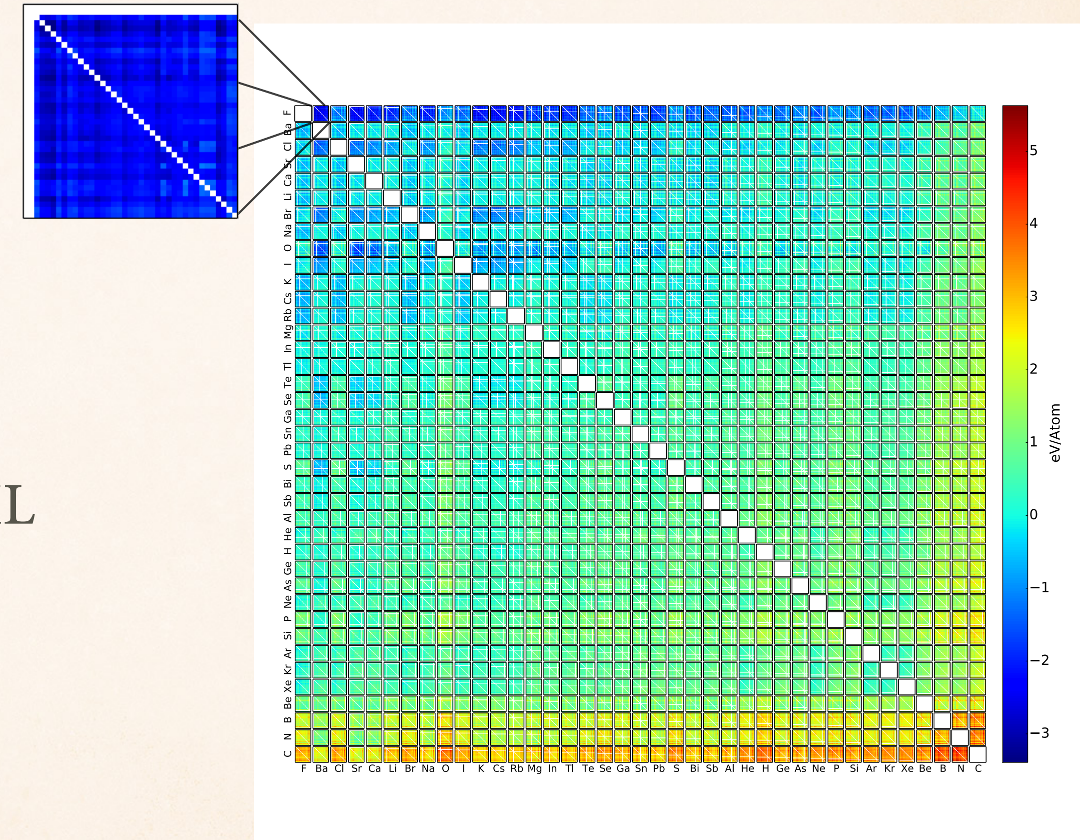
JCP 145, 161102

NEW KNOWLEDGE FROM ML

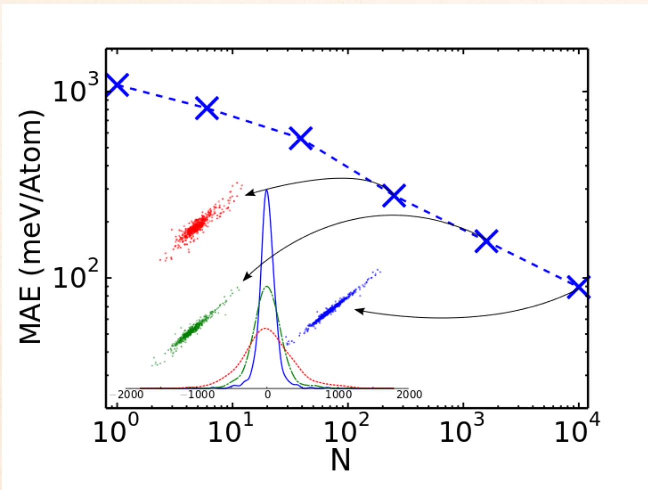
Elpasolite: K_2NaAlF_6



- 90 new stable stoichiometries
- New oxidation states for Al



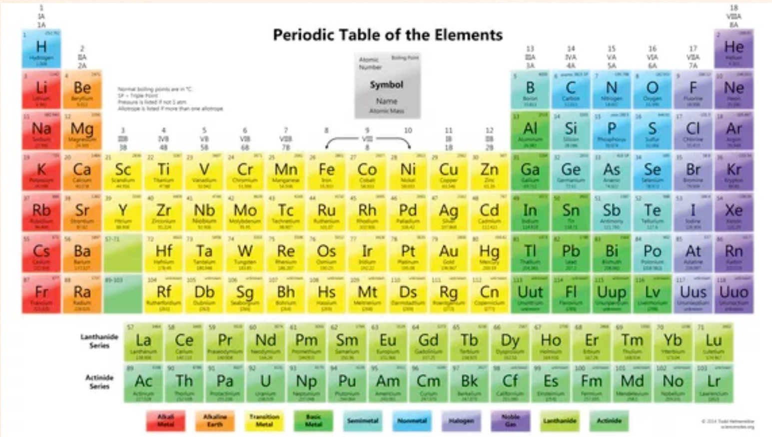
- training set: 10^4 DFT
- accuracy of 0.1eV/atom
- test: 2M (sp elements)
- 20M cpu-hrs DFT vs. 1 cpu-hr ML

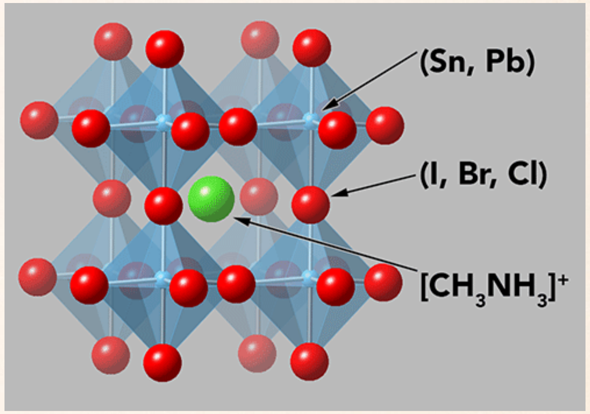


PRL 117, 135502

SCAN OF HUGE COMPOUND SPACE

- ACS: ~70M known substances (alloys, minerals, mixtures, polymers and salts) +10k daily
- 10^{60} (small) organic molecules





SCI. 351.6269, 151

ML ADVANTAGES

- Same accuracy as ab-inito models
- several orders of magnitude faster
- transferable and adaptable
- removes human bias

ACKNOWLEDGEMENTS

- Prof. A. von Lilienfeld
- Felix Faber
- Dr. Bing Huang
- Prof. M. Rupp

THANK YOU FOR YOUR ATTENTION

COMPUTATIONAL RESOURCES

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

AiiDA

FUNDING

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

UNI BASEL