MONTAN UNIVERSITÄT LEOBEN

# Data Science Summer School 2022
# Machine learning in Materials Science

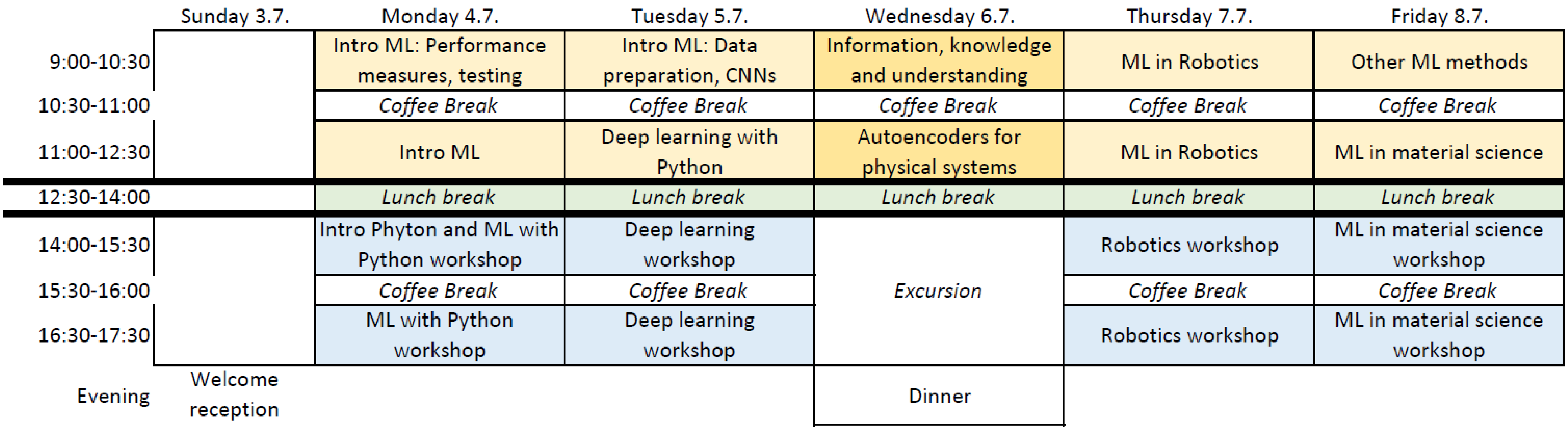
Lorenz Romaner

8.7.2022

**Computational Materials Science**
**Chair of Physical Metallurgy and Metallic Materials**
**Department of Materials Science**

WHERE RESEARCH MEETS THE FUTURE

# Schedule of the summer school



| | Sunday 3.7. | Monday 4.7. | Tuesday 5.7. | Wednesday 6.7. | Thursday 7.7. | Friday 8.7. |
|---|---|---|---|---|---|---|
| 9:00-10:30 | | Intro ML: Performance measures, testing | Intro ML: Data preparation, CNNs | Information, knowledge and understanding | ML in Robotics | Other ML methods |
| 10:30-11:00 | | *Coffee Break* | *Coffee Break* | *Coffee Break* | *Coffee Break* | *Coffee Break* |
| 11:00-12:30 | | Intro ML | Deep learning with Python | Autoencoders for physical systems | ML in Robotics | ML in material science |
| 12:30-14:00 | | *Lunch break* | *Lunch break* | *Lunch break* | *Lunch break* | *Lunch break* |
| 14:00-15:30 | | Intro Phyton and ML with Python workshop | Deep learning workshop | *Excursion* | Robotics workshop | ML in material science workshop |
| 15:30-16:00 | | *Coffee Break* | *Coffee Break* | | *Coffee Break* | *Coffee Break* |
| 16:30-17:30 | | ML with Python workshop | Deep learning workshop | | Robotics workshop | ML in material science workshop |
| Evening | Welcome reception | | | Dinner | | |

# Overview

- **9:00-10:30: Introduction and basic concepts**
  - ✧ Introduction to machine learning in materials science
    - Paradigms of materials science, data principles, materials databases
  - ✧ Additions on machine learning methods
    - Resampling methods
    - Decision trees
- **11:00-12:30: Applications of machine learning in materials science**
  - ✧ Overview
    - Structure or property prediction: Martensite start temperatures, toughness
      - Tutorial 1
    - Surrogate modeling: Machine learning density functional theory
      - Tutorial 2

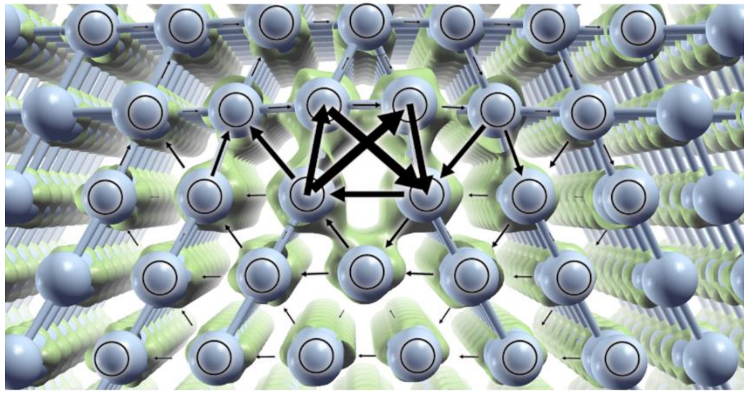
# Some information about myself

- Born and raised in Bozen, Italy.
- 1996-2003 Physics studies at TUG.
- 2003-2007 PhD at TUG.
    - 2 years at Georgia Tech, Atlanta USA.
    - PhD thesis: „Modelling of organic semiconductors and their interaction with metallic surfaces”.
- 2007-2011: PostDoc at MUL, Chair of “Atomistic Modeling and Design of Materials”.
- 2012-2020: Simulation expert at Materials Center Leoben Forschung GmbH.
    - 2018: Habilitation at the MUL, Materials Physics.
    - Title: “Atomistic investigations on the role of dislocations and interfaces for the mechanical and physical properties of metallic alloys and inorganic-organic heterostructures.”
- Seit 2021: Professorship for Computational Material Science at MUL.

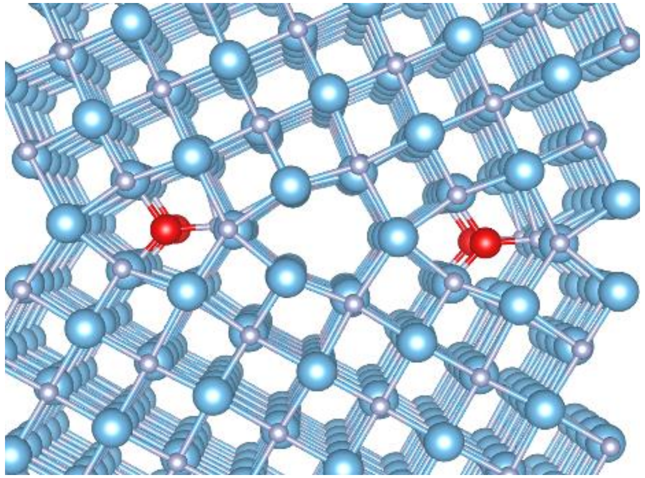
# Research fields

**Research fields:**

- Atomistic simulations
  - Density functional theory
  - Molecular dynamics
- Surface functionalization
- Crystallographic defects such as dislocations and interfaces
- Thermodynamic and kinetic materials simulations
- Design of plasticity, toughness and strength of metals
- Datadriven materials science



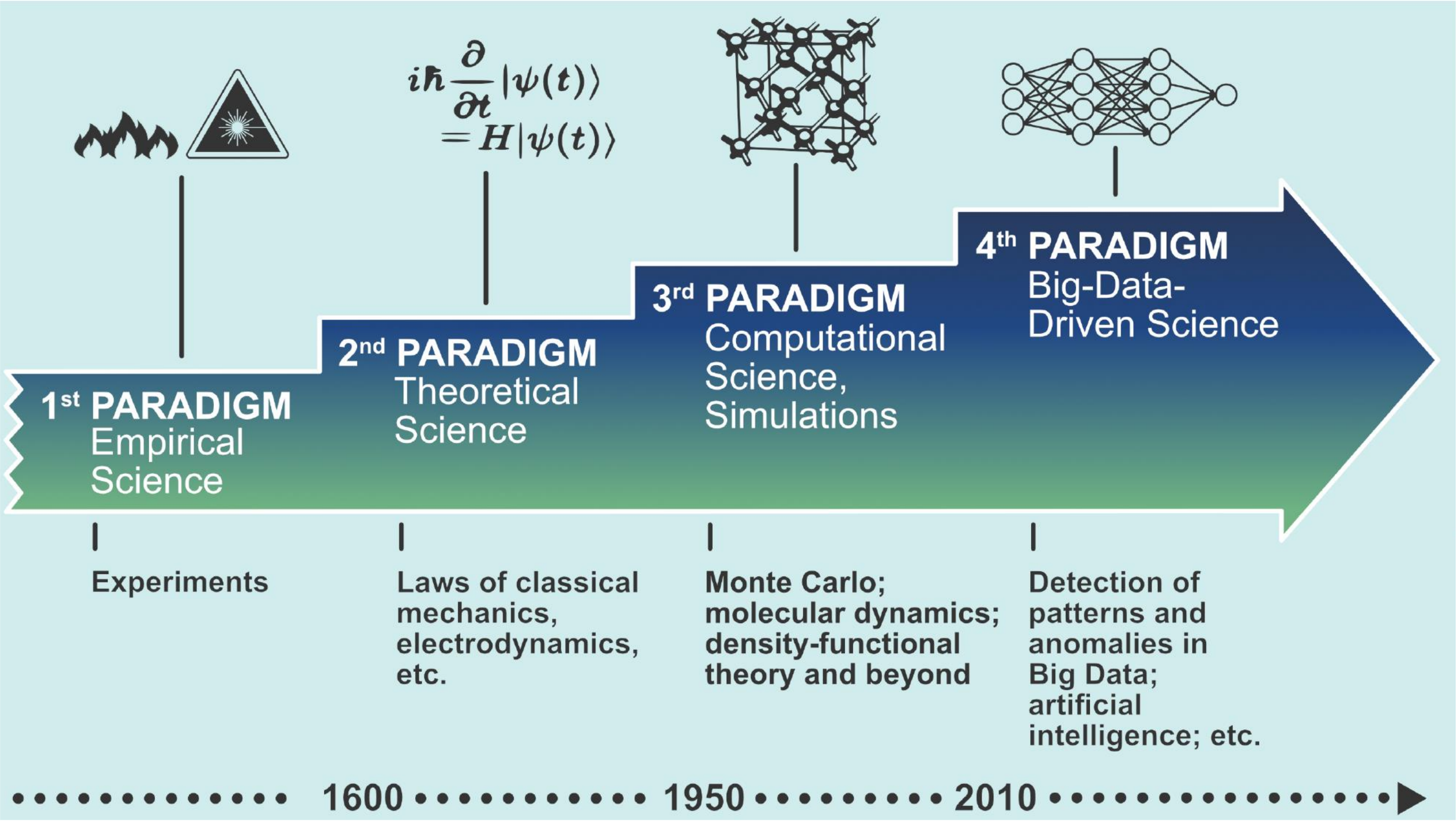
Dislocation cores



TiN grain boundary with B segregation

# Introduction to machine learning in materials science

# Paradigms of materials science



$$i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = H|\psi(t)\rangle$$

1st PARADIGM
Empirical Science

2nd PARADIGM
Theoretical Science

3rd PARADIGM
Computational Science, Simulations

4th PARADIGM
Big-Data-Driven Science

Experiments

Laws of classical mechanics, electrodynamics, etc.

Monte Carlo; molecular dynamics; density-functional theory and beyond

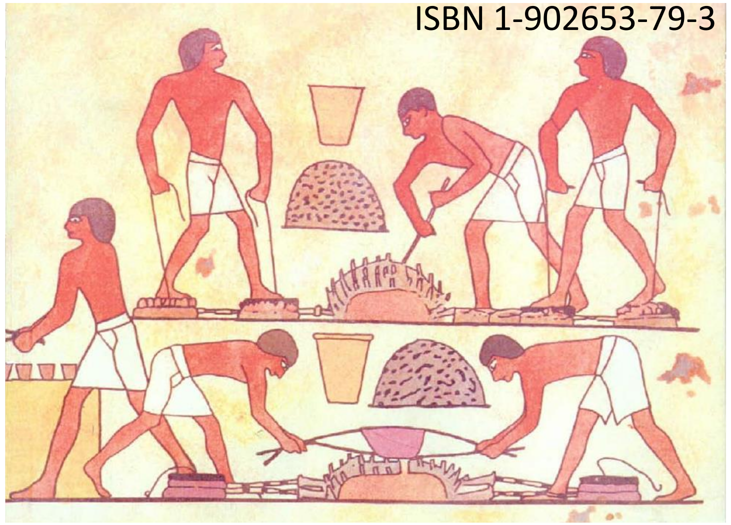
Detection of patterns and anomalies in Big Data; artificial intelligence; etc.

10.1007/978-3-319-44677-6_104
ISBN 978-0-9825442-0-4

# Paradigms of materials science

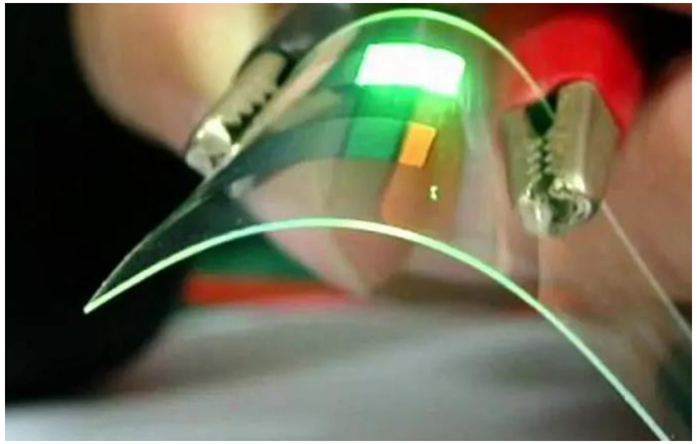
- **1st Paradigm: Empirical Science**
  - ✧Start: Copper Age (5000 BC) and Bronze Age (3000 BC)
  - ✧The basic techniques of metallurgy were developed in purely experimental research.
- **Also in recent science, examples of break-through in materials science based on trial-and-error exist**
  - ✧Conducting polymers: “A fortuitous combination of a unique […] catalyst soluble in organic solvents, […], a bizarre experimental mistake and the fact that polyacetylene is insoluble in just about everything had conspired to produce this promising metallic-looking material.” [10.1039/B210718J].
  - ✧Nobel Prize for Chemistry in 2000 for Heeger, MacDiarmid and Shirakawa.

Original drawing of metallurgy in Egypt



ISBN 1-902653-79-3

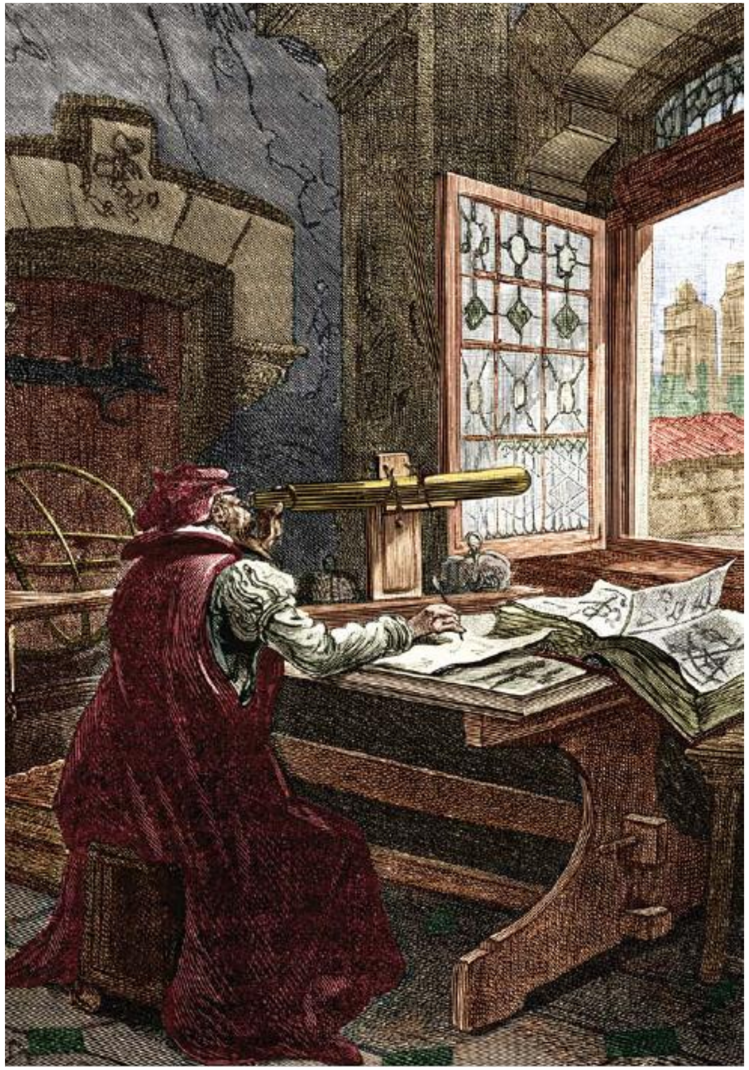
Flexible electronics



www.eenewsanalog.com/en/conducting-polymers-promise-soft-electronic-devices/

# Paradigms of materials science

- **2st Paradigm: Theoretical Science**
  - ✧Start: 1500 AD.
  - ✧Tycho Brahe (1546–1601), Johannes Kepler (1571–1630), Galileo Galilei (1564–1642), Isaac Newton (1643–1727), Gottfried Wilhelm Leibniz (1646–1716).
  - ✧Newton and Leibnitz developed the concept of the mathematical differential and derivatives.
  - ✧Analytical equations became the central instrument of theoretical physics.
  - ✧Newton's laws of motion, Maxwell equations, Schrödinger equation,…
  - ✧Nobel Prize 2013 to Francois Englert and Peter Higgs for predicting the Higgs mechanism, the process that gives elementary particles their mass.



ISBN 978-0-9825442-0-4

# The unreasonable effectiveness of mathematics in the natural sciences

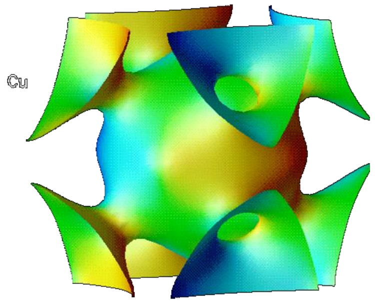
- **Article by Eugene Wigner [10.1002/cpa.3160130102]**
  - ✧“The mathematical formulation of the physicist's often crude experience leads in an uncanny number of cases to an amazingly accurate description of a large class of phenomena.”
  - ✧Example: falling rocks and observation of moon led to formulation of Newton laws which are surprisingly accurate in many circumstances.
  - ✧It is unknown why mathematics works, and impossible to tell whether a theory is unique.
  - ✧“The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve.”

10.1063/1.1522166

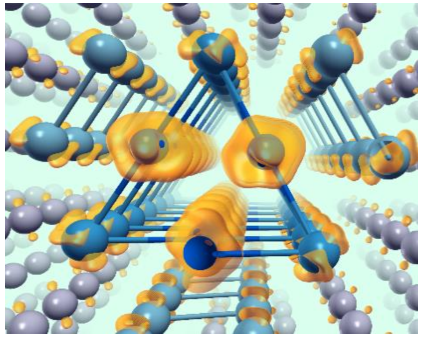
Eugene Wigner

# Paradigms of materials science

- **3rd Paradigm: Computational Science**
  - ✧ Start: With the availability of computers (about 1950).
  - ✧ Theoretical models grew too complicated to be solve analytically, and people had to start simulating.
  - ✧ Complex simulations are treated and analyzed analogously to experimental studies (computer experiment).
  - ✧ Density Functional Theory (Nobel prize 1998 for W. Kohn and J. A. Pople), Monte Carlo, Finite Elements, ….
  - ✧ Multi-scale modeling and integrated computational materials engineering



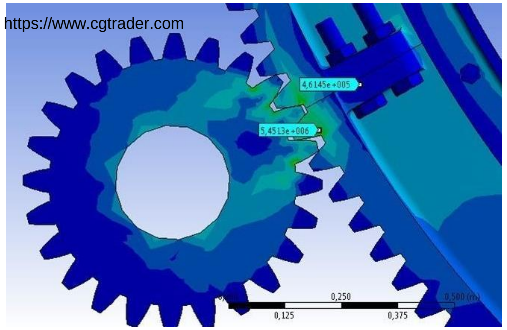
Density functional theory



Molecular dynamics
Density functional theory

paradis.stanford.edu

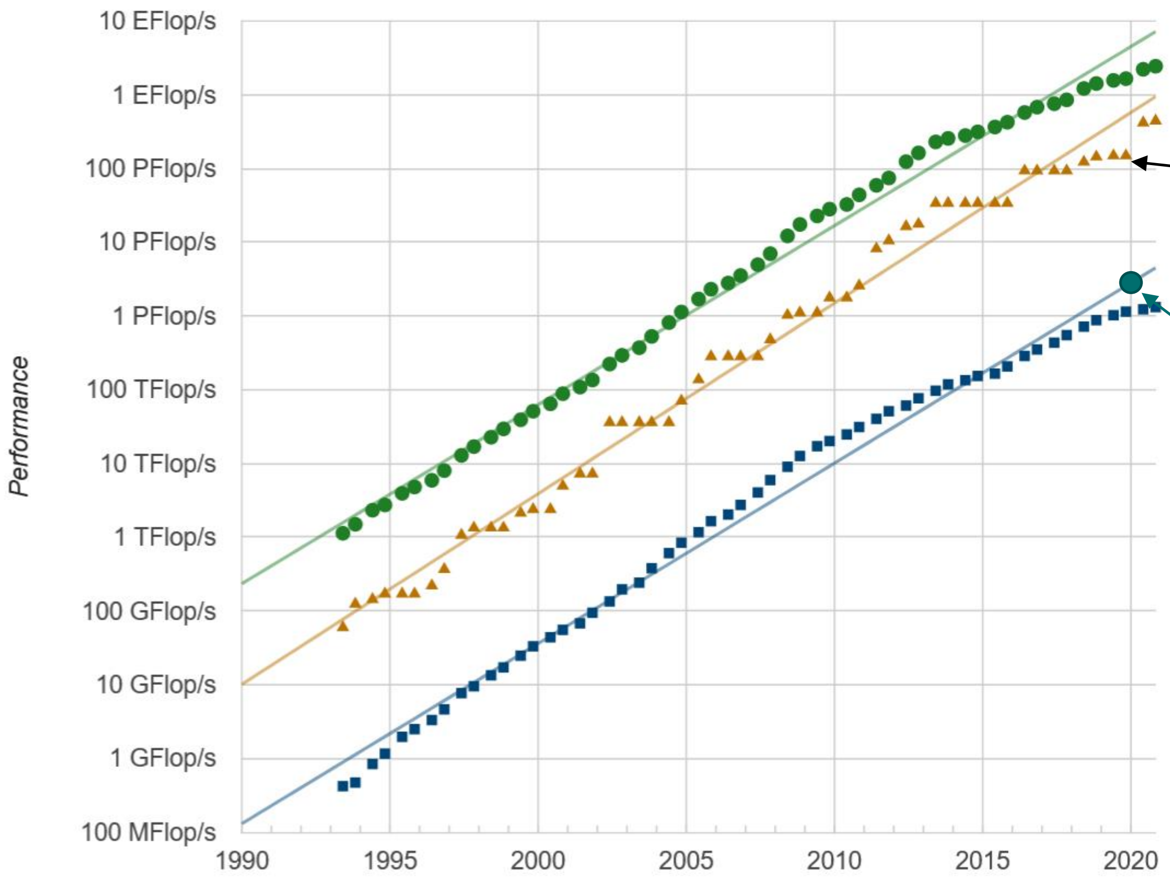
Dislocation dynamics



https://www.cgtrader.com

Finite element simulation

# Supercomputing is the tool for the 3rd paradigm



https://www.top500.org/

Fugaku, Japan

Summit, USA

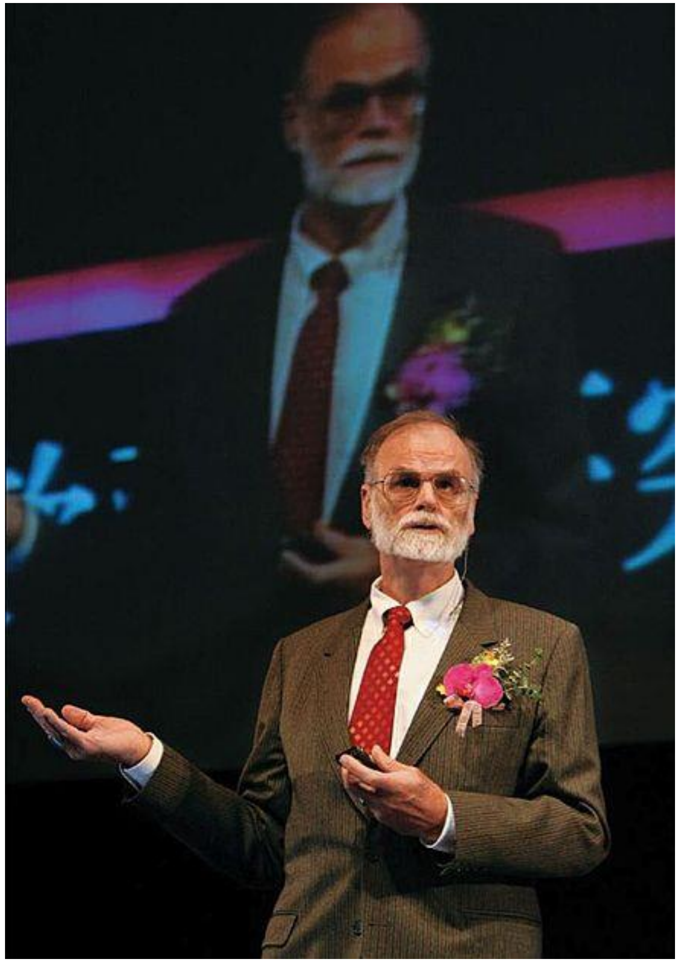
VSC, Austria

- ➢ Increasing computing power allows treating more complex phenomena.
- ➢ Improvement mostly due to parallelization → challenge for computer codes.

# Paradigms of materials science

- **4th Paradigm: Data-Driven Science**
  - ✧ Start: Probably with Jim Gray in 2007.
  - ✧ Simulations and experiments yield ever more data.
    - New techniques and technologies are needed to perform data-intensive science.
    - The techniques and technologies are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration.
  - ✧ Big data reveal correlations and dependencies that cannot be seen when studying small data sets.
  - ✧ 10.1126/science.1170411.

ISBN 978-0-9825442-0-4



Jim Gray

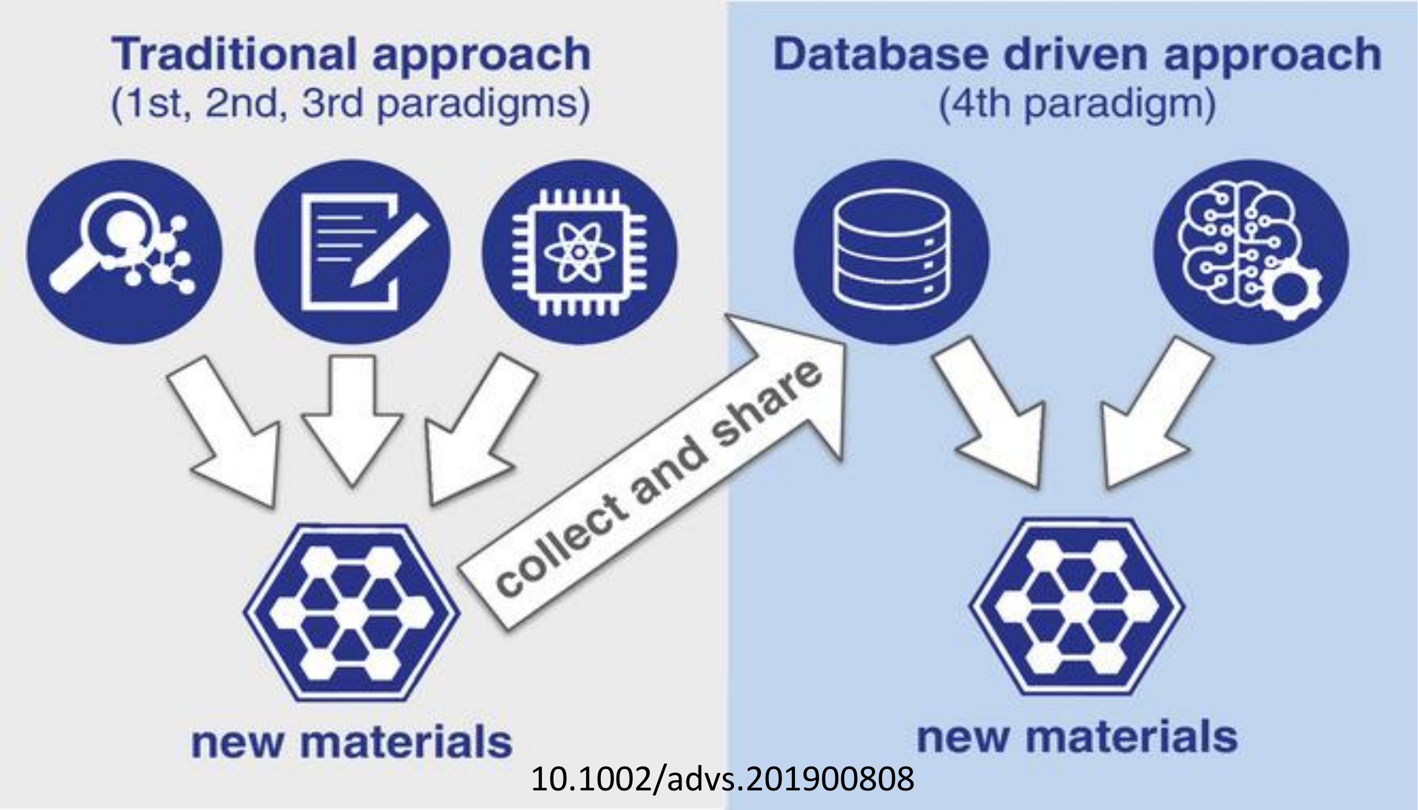
# The unreasonable effectiveness of data

- **Article by Alon Halevy, Peter Norvig, and Fernando Pere, Google** [10.1109/MIS.2009.36.]
  - “We should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.”
  - The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation.
    - Reason for these successes is not that these tasks are easier than other tasks.
    - Key is the large training set of the input-output behavior that is available in the wild (Web).

Peter Norvig

# Paradigms of materials science

- **4th Paradigm: Data-Driven Science**



10.1002/advs.201900808

Data gathered in data infrastructures. Machine learning approaches discover and propose new materials. → Recent trend (from about 2010)

# Big data in materials science

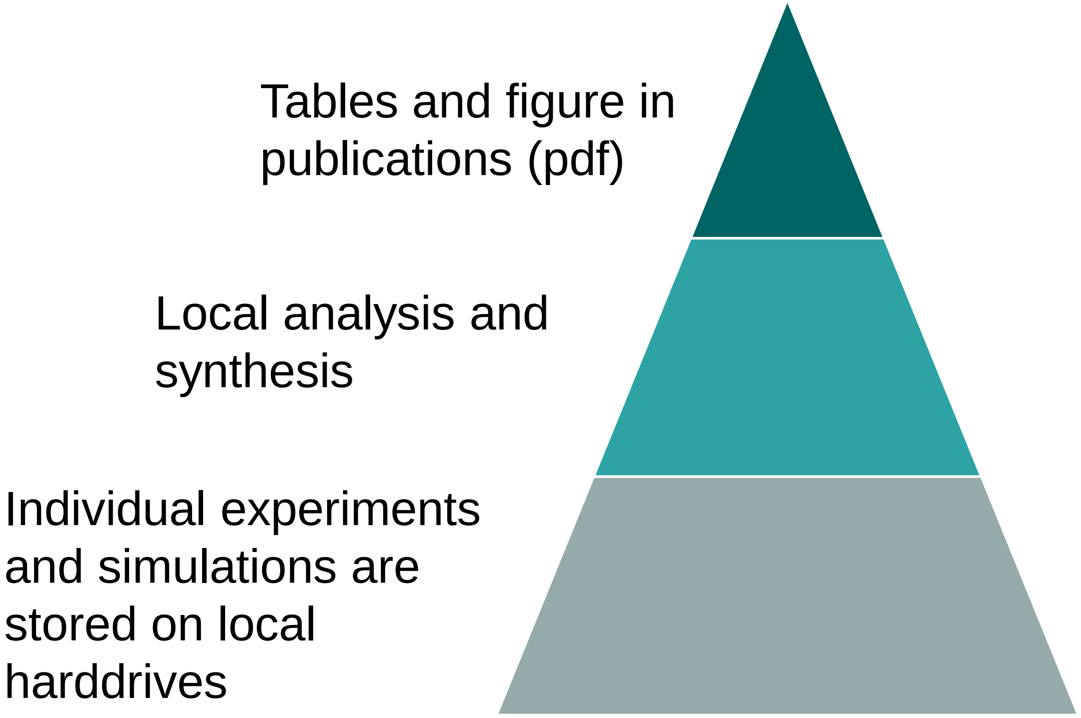
- **Big data and the four V challenge:**
  - [10.1007/978-3-319-44677-6_104, 10.1007/s43939-021-00012-0]
  - Volume: Materials science produces big data volumes
    - parallel synthesis, high-throughput screening, first-principles calculations in quantum chemistry and condensed matter physics.
  - Velocity: High pace of data generation.
  - Variety: Variety of data types and formats currently available. Problems arise when translating this language into computational models whose usage varies across research groups and even across individual researchers.
  - Veracity: potential lack of quality in data produced by imprecise simulations or collected from experiments not conforming to a sufficiently rigid protocol.

# Fair data principles

- **FAIR data principles: useful framework for thinking about sharing data in a way that will enable maximum use and reuse.**
  - **F** stands for findable
    - Requirement: Persistent identification through provision of digital object identifiers (DOIs) to datasets
    - Provide meta-data to dataset
  - **A** stands for accessible:
    - Requirements: Hardware and software tools are needed to access the data, application programming interfaces (API)
  - **I** stands for interoperable:
    - Requirement: use of metadata standards, structured file formats, translation and transformation
  - **R** stands for reusable
    - Requirement is licensing: CC-BY-4.0, DbCL v1.0, …

# Paradigm change in data sharing and publishing

- **In traditional research practice only the tip of the iceberg of data is published.**
  - ✧ Often in in pdf format which is not machine-readable and cumbersome also for humans.
- **With new databases, data can be released and published at much higher volumes and data can be used also by others.**
  - ✧ No need to repeat lengthy calculations or expensive experiments.
  - ✧ Provide grounds for data exploration and machine learning.


Tables and figure in publications (pdf)

Local analysis and synthesis

Individual experiments and simulations are stored on local harddrives

# Materials databases

| Name | Website | Overview |
|---|---|---|
| AFLOW | aflowlib.org | Computational data consisting of 2 118 033 material compounds and 281 698 389 calculated properties with focus on inorganic crystal structures. Incorporates multiple computational modules for automating high-throughput first principles calculations. |
| Computational Materials Repository | cmr.fysik.dtu.dk | Computational datasets from a diverse set of applications. Data creation and analysis with the Atomic Simulation Environment (ASE). |
| Crystallography Open Database | crystallography.net | Open-access collection of crystal structures of organic, inorganic, metal–organic compounds and minerals, excluding biopolymers. |
| HTEM | htem.nrel.gov | Properties of thin films synthesized using combinatorial methods. Contains 57 597 thin film samples, across a wide range of materials (oxides, nitrides, sulfide, intermetallics). |
| Khazana | khazana.gatech.edu | Platform to store structure and property data created by atomistic simulations, and tools to design materials by learning from the data. Tools include Polymer Genome and AGNI. |
| MARVEL NCCR | nccr-marvel.ch | Materials informatics platform for data-driven high-throughput quantum simulations. Data available at materialscloud.org, powered by the AiiDA-infrastructure. |
| Materials Data Facility (MDF) | materialsdatafacility.org | Data publication network for computational and experimental datasets. Data exploration through the Forge python package. |
| Materials Project | materialsproject.org | Online platform for materials exploration containing data of 86 680 inorganic compounds, 21 954 molecules and 530 243 nanoporous materials. Develops various open-source software libraries, including pymatgen, custodian, FireWorks, and atomate. |
| MatNavi/NIMS | mits.nims.go.jp | An integrated material database system comprising ten databases, four application systems and the NIMS Structural Datasheet Online. |
| NOMAD CoE | nomad-coe.eu | Provides storage for full input and output files of all important computational materials science codes, with multiple big-data services built on top. Contains over 50 236 539 total energy calculations. |

Brown are atomistic modeling databases

Blue are crystallographic databases

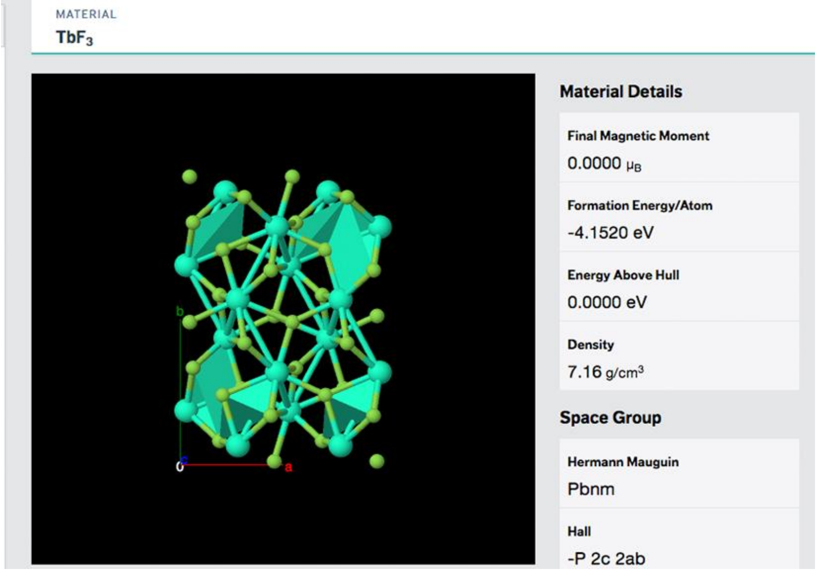
10.1002/advs.201900808

# Materials databases

10.1002/advs.201900808

| Name | Website | Overview |
|---|---|---|
| Organic Materials Database | omdb.mathub.io | Open access electronic structure database for 3-dimensional organic crystals. Contains approximately 24 000 materials. |
| Open Quantum Materials Database | oqmd.org | Database of DFT-calculated thermodynamic and structural properties with focus on inorganic crystal structures. Contains 563 247 entries with support for full download and advanced usage through the qmpy python package. |
| Open Materials Database | openmaterialsdb.se | Computational database primarily based on structures from the Crystallography Open Database. Data creation and analysis with High-Throughput Toolkit (httk). |
| SUNCAT | suncat.stanford.edu | Materials informatics center for atomic-scale design of catalysts. Online tools and computational results for 112 157 surface reactions and barriers available at catalysis-hub.org. |
| Citrine Informatics | citrine.io | A materials informatics platform combining data infrastructure and AI. Open database and analytics platform for material and chemical information available at the Citrination platform: citrination.com. |
| Exabyte.io | exabyte.io | Cloud-based modelling platform for materials informatics. |
| Granta Design | grantadesign.com | R&D organization offering data, tools and expertise for materials design. |
| Materials Design | materialsdesign.com | Software products and services for chemical, metallurgical, electronic, polymeric, and materials science research applications. |
| Materials Platform for Data Science | mpds.io | Online edition of the PAULING FILE with focus on curated experimental data for inorganic materials. |
| MaterialsZone | materials.zone | Provides a notebook-based materials informatics environment together with experimental data. |
| SpringerMaterials | materials.springer.com | Curated data covering multiple material classes, property types, and applications. A set of advanced functionalities for visualizing and analyzing data provided through SpringerMaterials Interactive. |

# Materials project



MATERIAL

$TbF_3$

Material Details

Final Magnetic Moment

0.0000 $\mu_B$

Formation Energy/Atom

-4.1520 eV

Energy Above Hull

0.0000 eV

Density

7.16 g/cm$^3$

Space Group

Hermann Mauguin

Pbnm

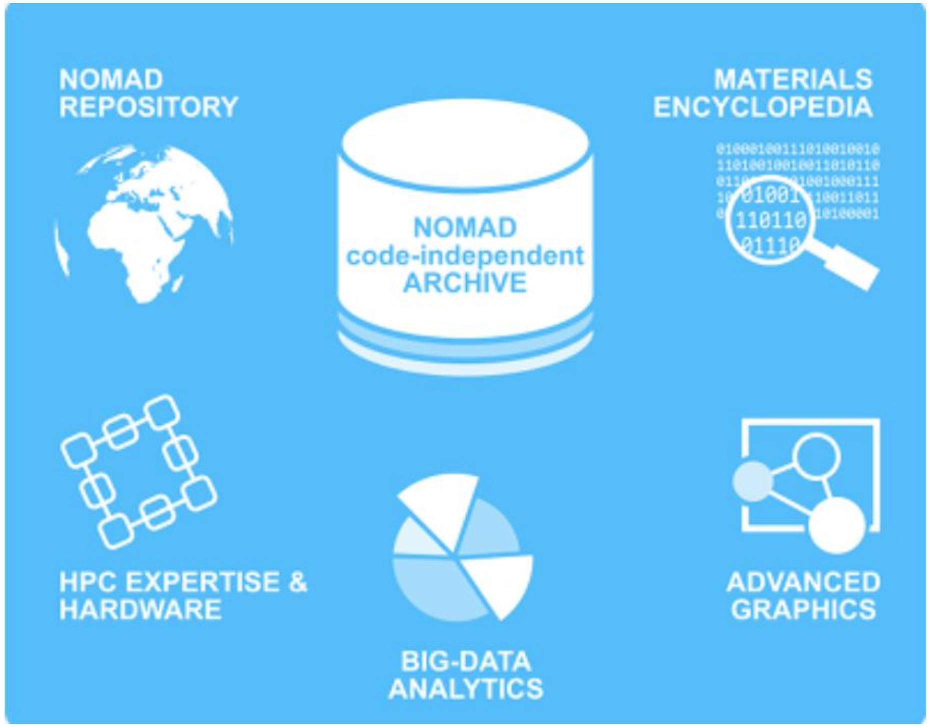
Hall

-P 2c 2ab

https://materialsproject.org

- **Open-access database offering material properties from density functional theory calculations.**
  - ✧Help predicting and selecting new materials.
  - ✧Established in 2011 with an emphasis on battery research.
  - ✧Includes property calculations for many areas of clean energy systems such as photovoltaics, thermoelectric materials, and catalysts.
  - ✧Most of the known 35,000 molecules and over 130,000 inorganic compounds are included in the database.
  - ✧Provided with an API which enables data transfer via http.

# NOMAD repository and archive

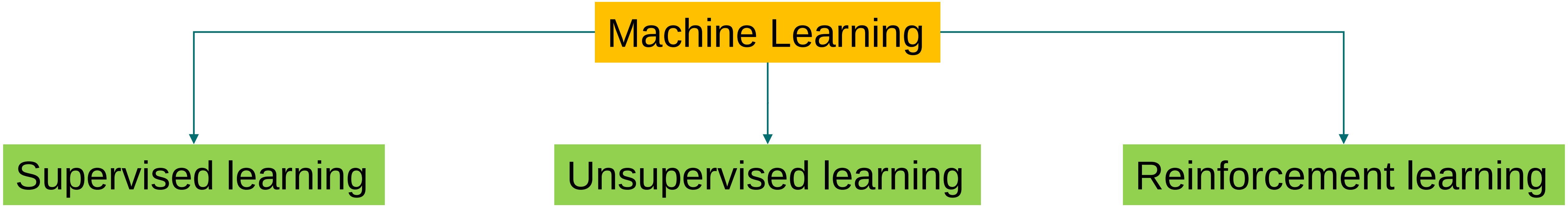


https://www.nomad-coe.eu

- **Repository and Archive**
  - ✧Open access of scientific materials data.
  - ✧Enables the confirmatory analysis of materials data, their reuse, and repurposing.
  - ✧All data is available in their raw format as produced by the underlying DFT code.
  - ✧Common, machine-processable, and well-defined data format.
  - ✧Data can be downloaded and used under the CC-BY-4.0 license.
  - ✧Data can be uploaded without any barrier: results are accepted as they are.
  - ✧You can request digital objective identifiers (DOI's) for your datasets and cite your data.

# Additions on machine learning methods

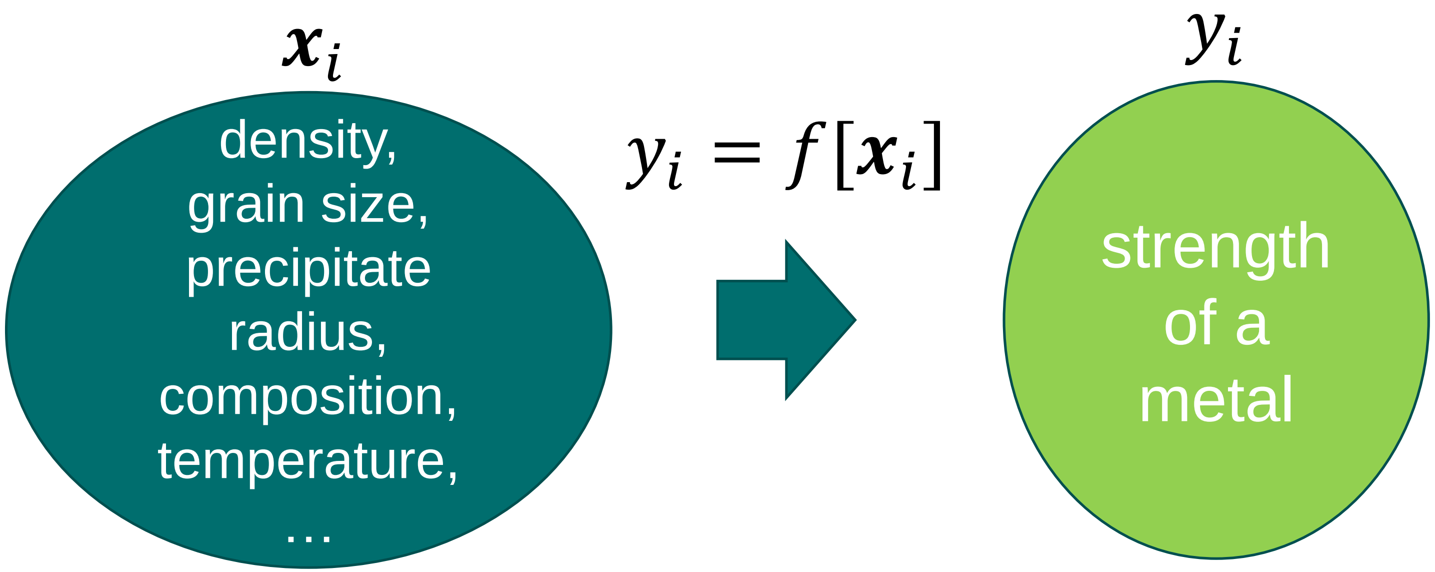
# Main categories of machine learning



- The difference is related to the application and to the information about the data.
- In general we have a set of data points, $D = \{(\boldsymbol{x}_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i$ is a vector consisting of numerical or categorical assignments.
- If we select a component of the vector as the quantity we wish to predict from the other components, the data set is labeled.
  - $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where $y_i$ is the label.

# Supervised learning

- **Supervised learning** approach, the goal is to learn a mapping $f$ from inputs $\boldsymbol{x}_i$ to outputs $y_i$. i.e. $y_i = f[\boldsymbol{x}_i]$**.**
- The $\boldsymbol{x}_i$ are called **predictors, features, or descriptors.**
- $y_i$ is the **response or target quantity**.
- Example from materials engineering:



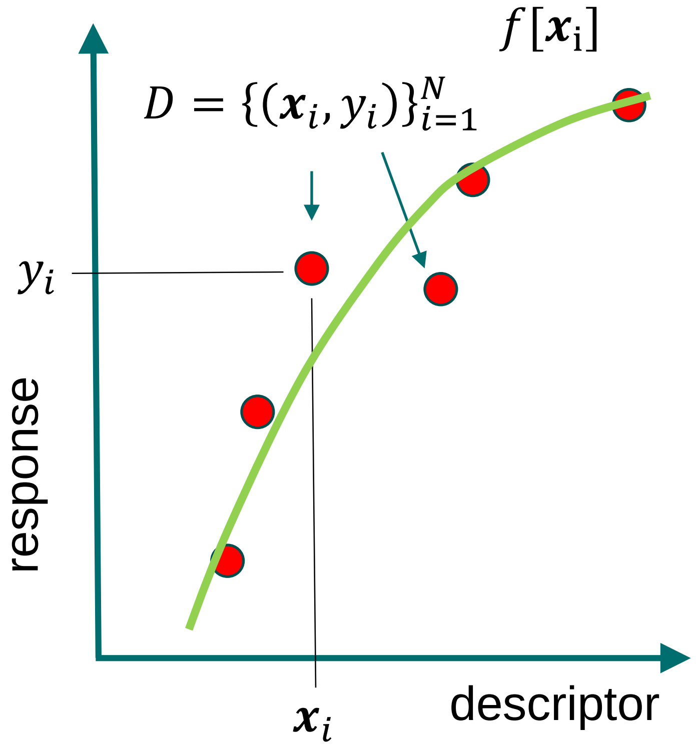
$\boldsymbol{x}_i$

density, grain size, precipitate radius, composition, temperature, …

$y_i = f[\boldsymbol{x}_i]$

$y_i$

strength of a metal



# Common machine learning methods in use in materials science

- Neural networks (deep learning)
- Support vector regression or support vector classification
- Gaussian process regression
- Decision tree based methods
  - Bagging, random forest regression or classification
  - Boosted decision tree regression or classification

# Prediction vs. inference.

- In supervised learning there are two main motivations for estimating $y_i = f[\boldsymbol{x}_i]$.
  - ✧Prediction of new response points $y_i$ for which no data exist yet → Prediction.
  - ✧Explore the relationship between $y_i$ and $\boldsymbol{x}_i$ → Inference.
- For prediction only the result and accuracy of $y_i$ is decisive while the model and its functional form are not of importance. (Black box)
- For inference, the model is the main goal. Its functional form should be transparent (parametric) to answer the following questions:
  - ✧Which predictors are associated with the response?
  - ✧What is the relationship between the response and each predictor? Is it positive or negative with $y$.
  - ✧Is the relationship linear or non-linear ?

# Examples for prediction or inference

- Example for prediction:
  - ✧Image detection: Important is the effectiveness of the image classification but not the details of the algorithm (i.e. edge detection,…)
  - ✧Email Spam filtering: The algorithm should get rid of unsuitable emails.

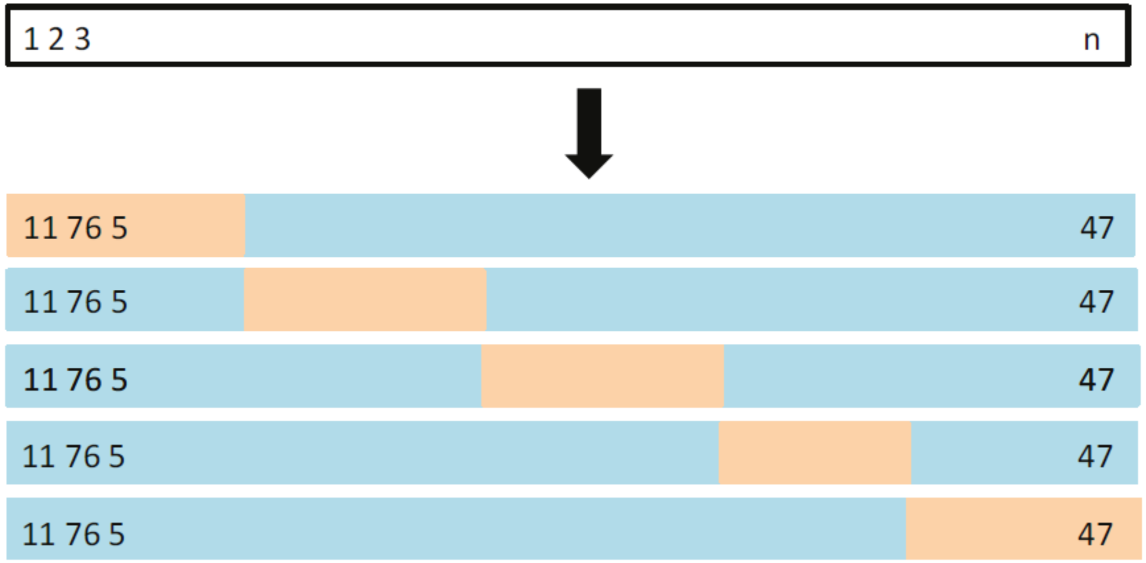
- Example for inference:
  - ✧Prediction of the martensite start temperature for steels. (Role of carbon, microstructures,…).
  - ✧Prediction of the Gibbs free energy from thermodynamic data (understanding whether solutes like to mix, impact of temperature…)

# Resampling methods

- Resampling methods involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model.
- Two common resampling methods are:
- Cross-validation (CV):
  - CV estimates the validation error of a given statistical learning method to evaluate performance, or to select the appropriate level of flexibility.
  - Measures how well the model performs for validation data.
- Bootstrap (BS):
  - BS provides a measure of accuracy of a parameter estimate or of a given selection.
  - Measures model uncertainty.

# Cross-validation

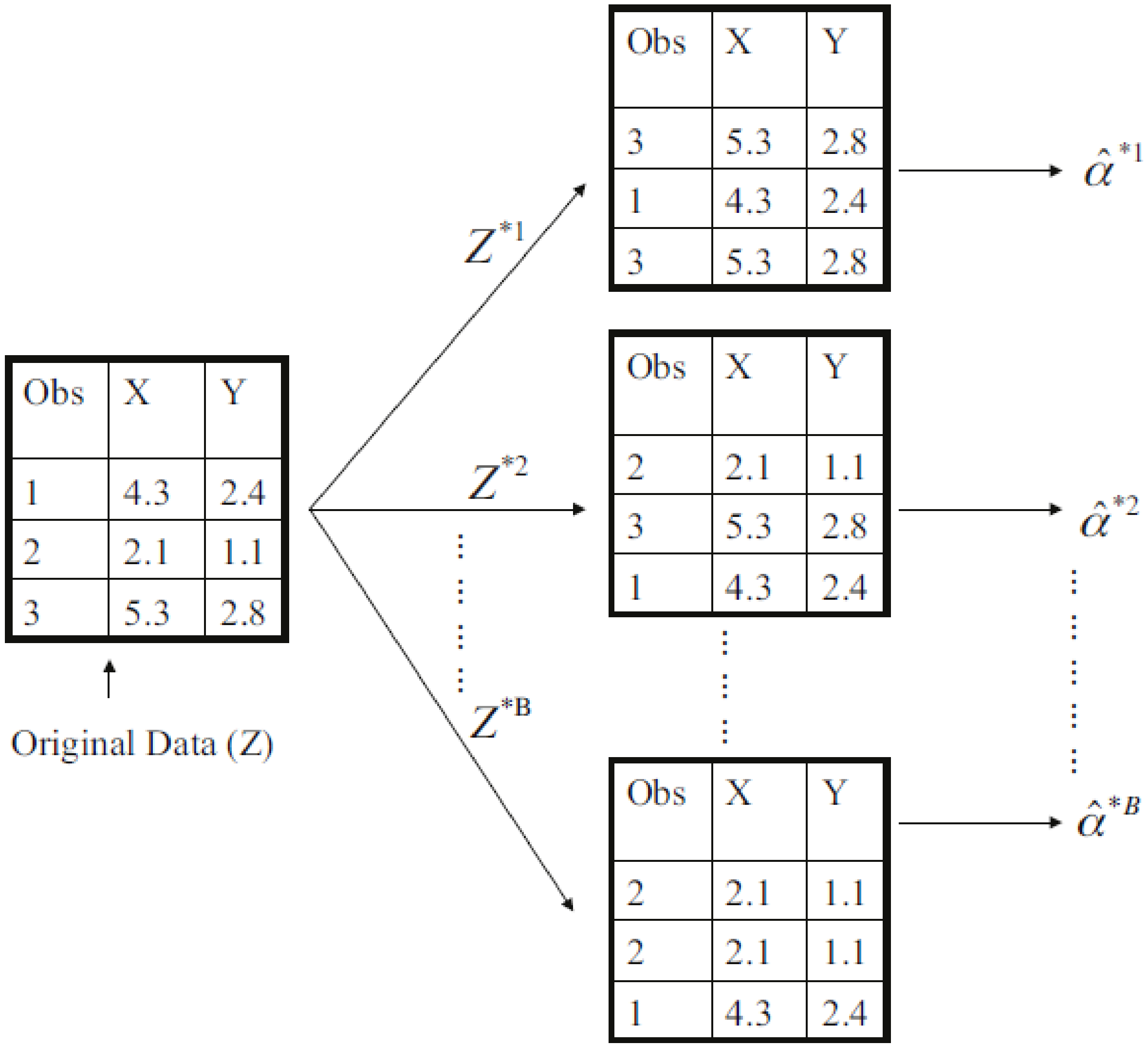
ISBN: 1461471370



- CV involves dividing the set of observations into k groups, or folds, of approximately equal size.
- The first fold is treated as a validation set, and the ML method is fit on the remaining k − 1 folds.
- The mean squared error, $MSE_1$, is then computed on the observations in the held-out fold.
- This procedure is repeated k times with a different fold as a validation set. This process results in k estimates of the validation error, $MSE_1, MSE_2, \ldots, MSE_k$
- The k-fold CV estimate of $MSE$ is computed by $CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i$.

# Bootstrapping

- Bootstrapping involves creating many BS datasets (Z*) from the original dataset (Z).
- Example of BS on a small sample.
- In contrast to CV:
  - The same data point can occur several times in the same dataset.
  - A point may never be included in the bootstrapped datasets.
- For every Z*, model training is carried out giving providing a model population



Original Data (Z)

| Obs | X | Y |
|---|---|---|
| 1 | 4.3 | 2.4 |
| 2 | 2.1 | 1.1 |
| 3 | 5.3 | 2.8 |

$Z^{*1}$

| Obs | X | Y |
|---|---|---|
| 3 | 5.3 | 2.8 |
| 1 | 4.3 | 2.4 |
| 3 | 5.3 | 2.8 |

$Z^{*2}$

| Obs | X | Y |
|---|---|---|
| 2 | 2.1 | 1.1 |
| 3 | 5.3 | 2.8 |
| 1 | 4.3 | 2.4 |

$Z^{*B}$

| Obs | X | Y |
|---|---|---|
| 2 | 2.1 | 1.1 |
| 2 | 2.1 | 1.1 |
| 1 | 4.3 | 2.4 |

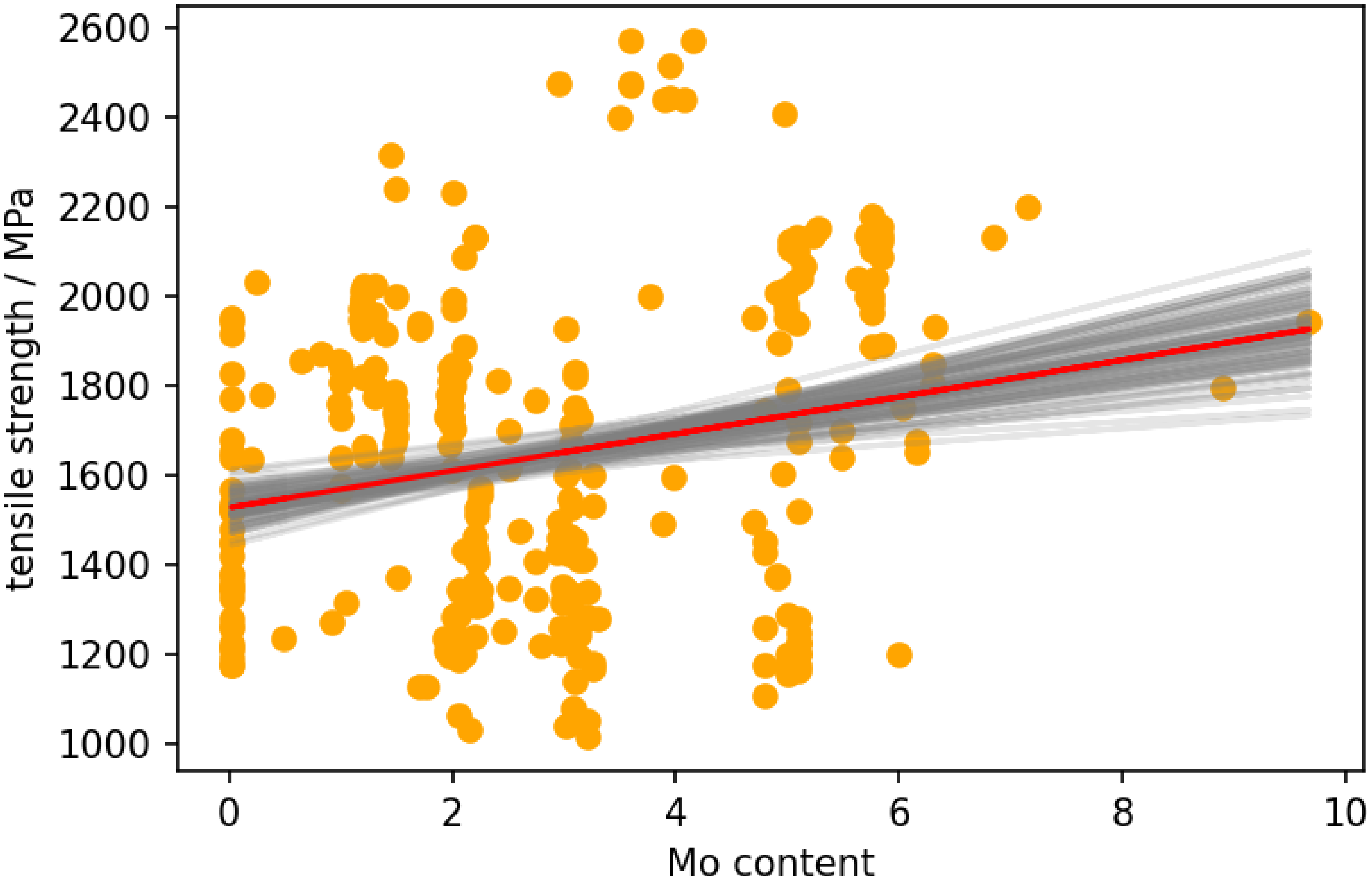
ISBN: 1461471370

# Example with linear regression

- Consider a linear regression carried out on 1000 BS data sets. $y = \beta_0 + \beta_1 x$
- We obtain 1000 linear functions with different $\beta_0$ or $\beta_1$
- For linear regression a direct simple expression exists:
  - ✧ $SE(\beta_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right],\ SE(\beta_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- However, bootstrapping can provide standard errors of model parameters for any method.

# Example with linear regression

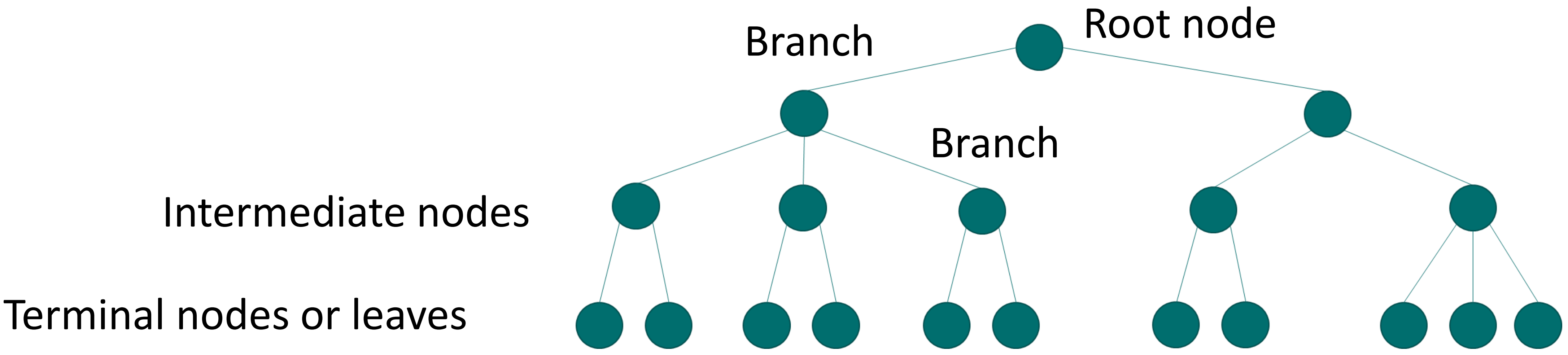
- Illustration of the variance of a linear fit.
- Gray lines illustrate linear fits to BS datasets.
- Red line is the mean of all results.



# Tree-based methods

- ➢ Tree-based methods construct decision trees to model the response from descriptors.
- ➢ Decision tree are constructed by applying a set of splitting rules used to segment the predictor space.
- ➢ Decision trees can be applied to both regression and classification problems.

Example of a decision tree



Root node

Branch

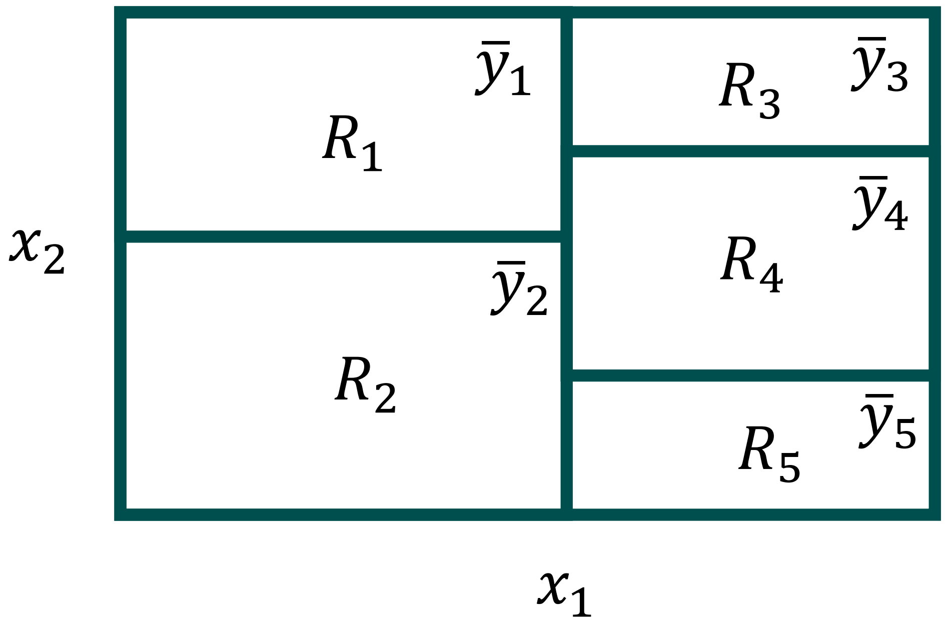
Branch

Intermediate nodes

Terminal nodes or leaves

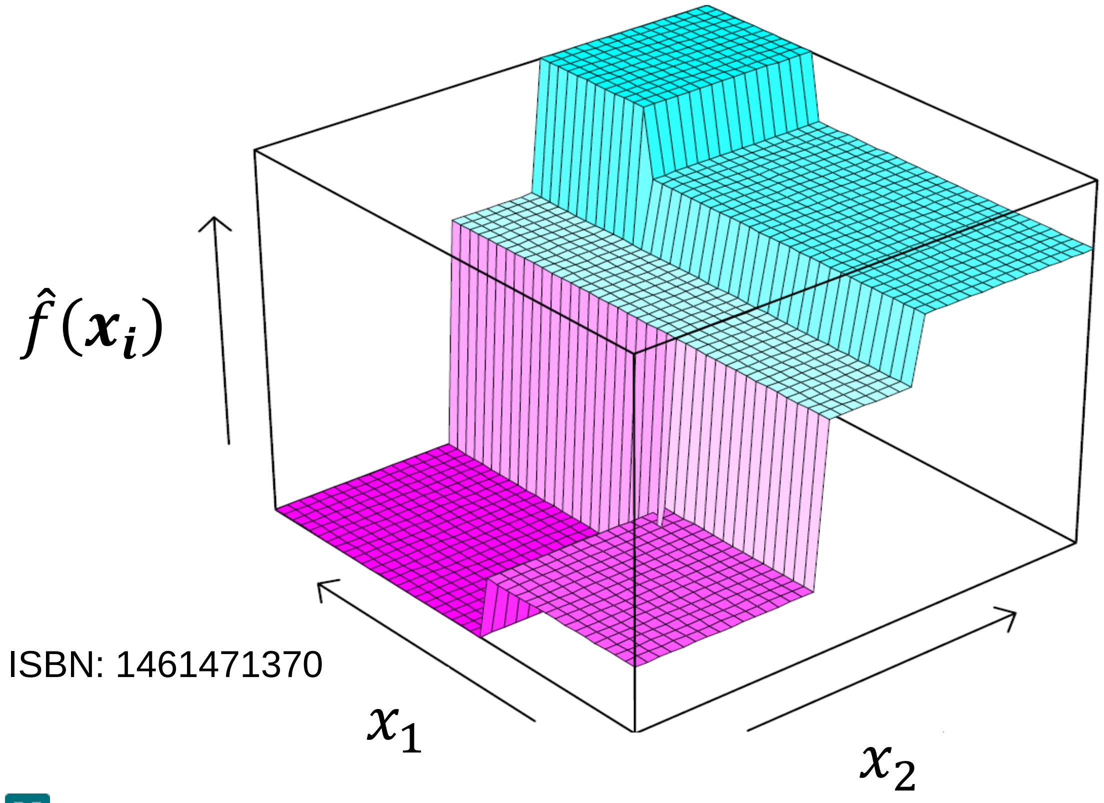
# Building the decision tree

- Steps to build a regression tree:
  - ✧Step 1: Divide the predictor space, i.e. the set of all possible $\boldsymbol{x}$, into J distinct and non-overlapping regions, $R_1, R_2, R_3, \ldots, R_J$.
  - ✧Step 2: For every observation that falls into the region $R_j$ , the prediction is equal to the mean value of the training data $\bar{y}_j$.

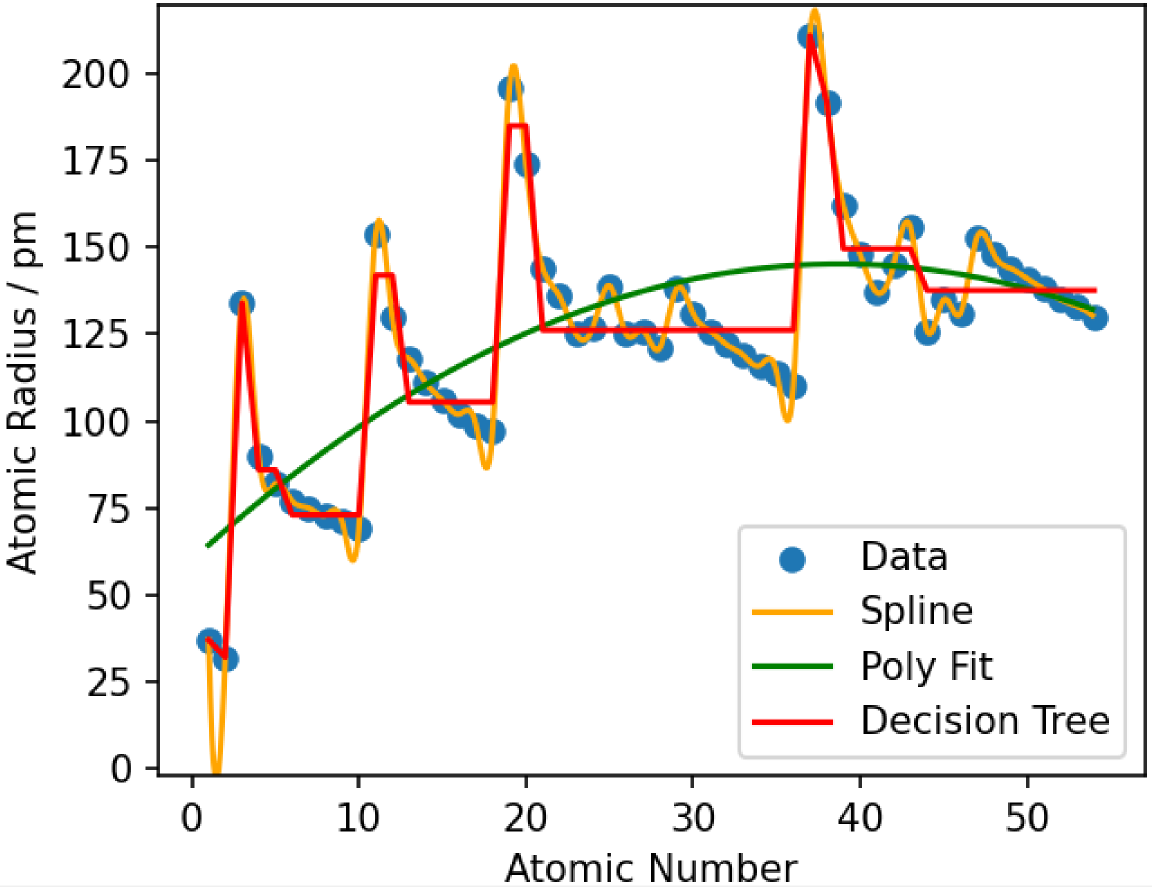


# Essentials of decision trees

- Construction of the decision tree means approximating the true $f(\boldsymbol{x_i})$ with a step function $\hat{f}(\boldsymbol{x_i})$.

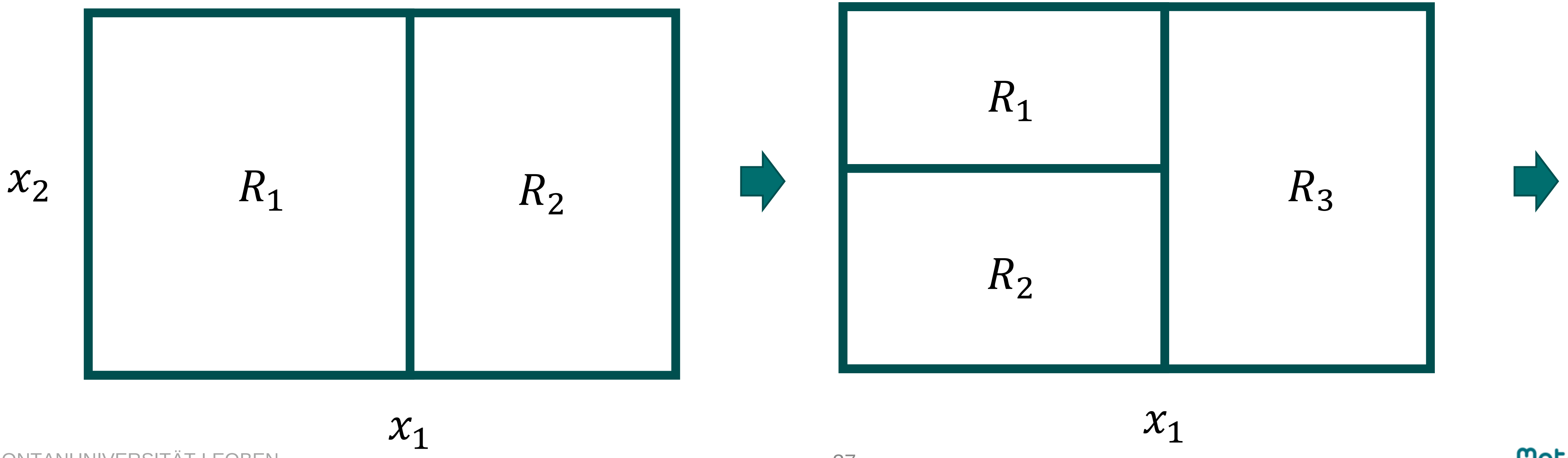


ISBN: 1461471370



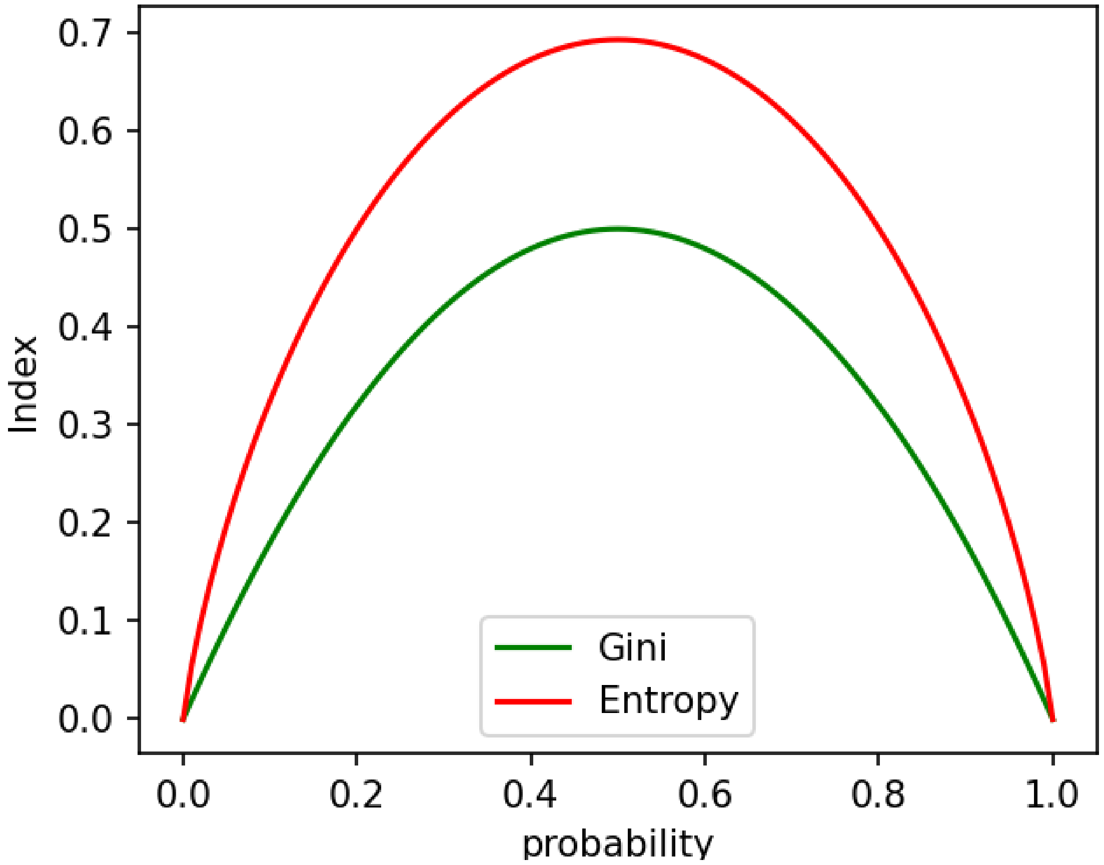
# Essentials of decision trees

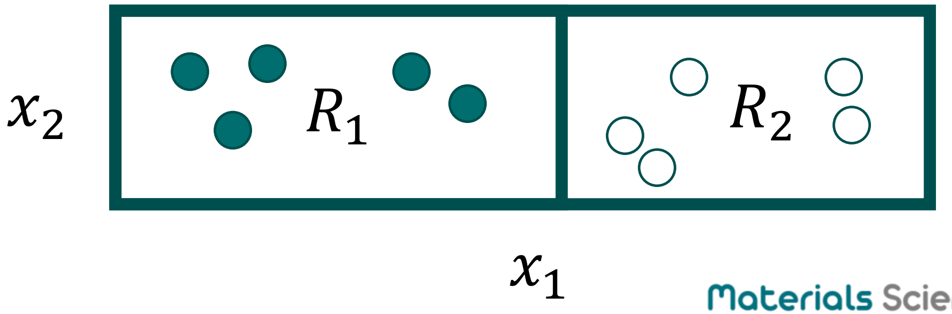
- Training the decision tree means finding the partition which minimizes
  - ✧ $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \bar{y}_j\right)^2$
  - ✧ Recursive binary splitting:
    - The regions are found by sequentially splitting into two regions.
    - Always the split with the lowest RSS is chosen.



# Classification trees

- For classification trees, not RSS but other criteria are used to determine the goodness of the split.
  - ✧ Gini index $G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$
  - ✧ Entropy $D = -\sum_{k=1}^{K} p_{mk}\log(p_{mk})$
- Here $p_{mk}$ represents the proportion of training observations in the m-th region belonging to the k-th class.
- The criteria measure node purity
  - ✧ Pure means the node consists primarily of one class → suitable split.





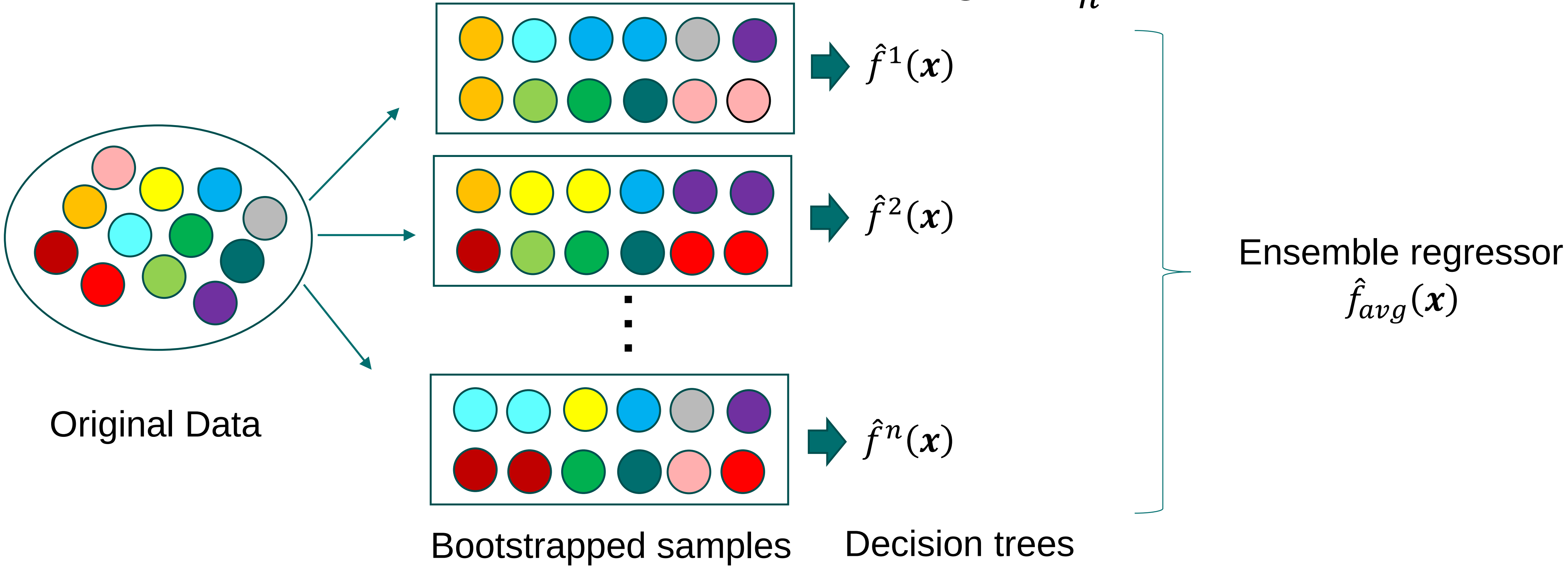
# Considerations on decision trees

- **Advantages:**
  - Trees are very easy to explain and can be displayed graphically, (especially if they are small).
  - Closely mirror human decision-making.
  - Trees can handle qualitative predictors without the need to create dummy variables.
- **Disadvantages:**
  - In the simple form they do not have the same level of predictive accuracy as some of the other regression and classification approaches.
  - Trees can be very non-robust. A small change in the data can cause a large change in the final estimated tree.
  - Approaches: Combine many decision trees, bagging, and boosting.

# Bagging

- Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method.

- Bagging combines many decision trees together $\hat{f}_{avg}(\boldsymbol{x})$=$\frac{1}{n}\sum_{b=1}^{n}\hat{f}^{n}(\boldsymbol{x})$



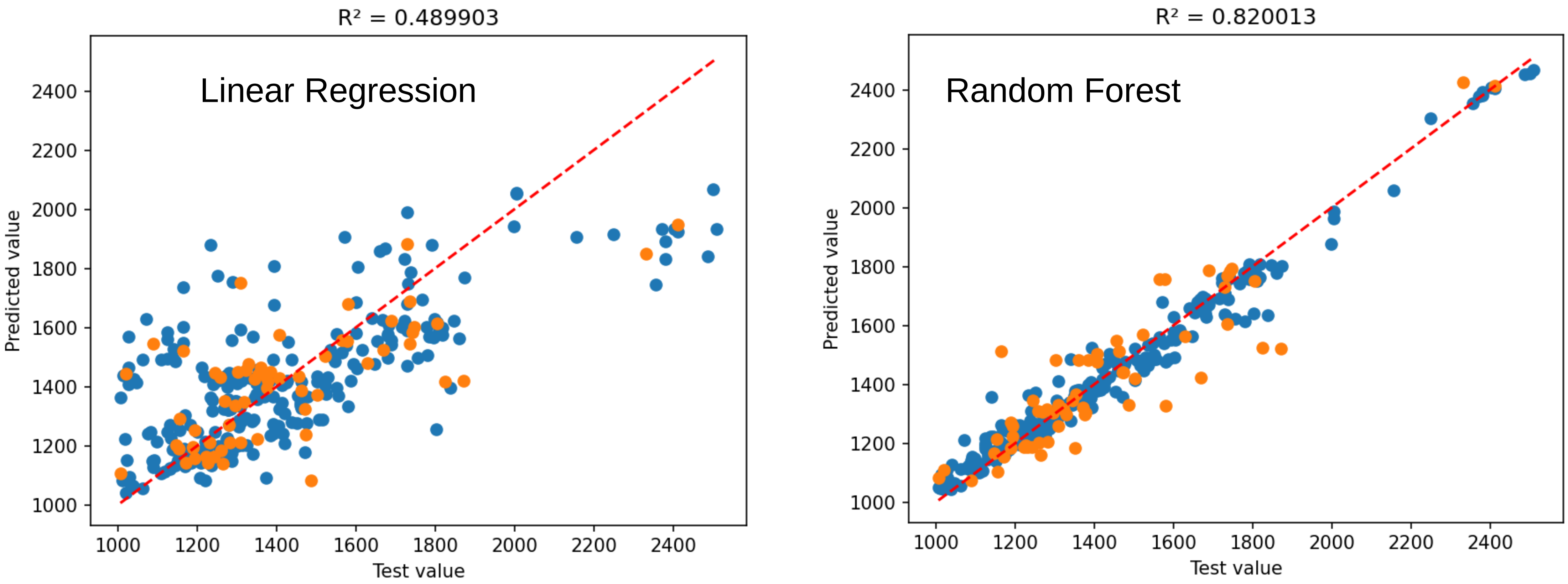
$\hat{f}^{1}(\boldsymbol{x})$
$\hat{f}^{2}(\boldsymbol{x})$
$\hat{f}^{n}(\boldsymbol{x})$
Ensemble regressor
$\hat{f}_{avg}(\boldsymbol{x})$
Original Data
Bootstrapped samples
Decision trees

- The variance of the mean of many observations, each with $\sigma^2$, is $\frac{\sigma^2}{n}$.

# Random forest (RF)

- Even when bagging, some predictors may strongly dominate over others.
  - Same predictors are favored and the trees are correlated.
- The approach of random forests is:
  - A forest of decision trees is built on bootstrapped training samples.
  - Each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors.
- The split is allowed to use only one of those m predictors.
- A fresh sample of m predictors is taken at each split,
  - typically $m = \sqrt{p}$.

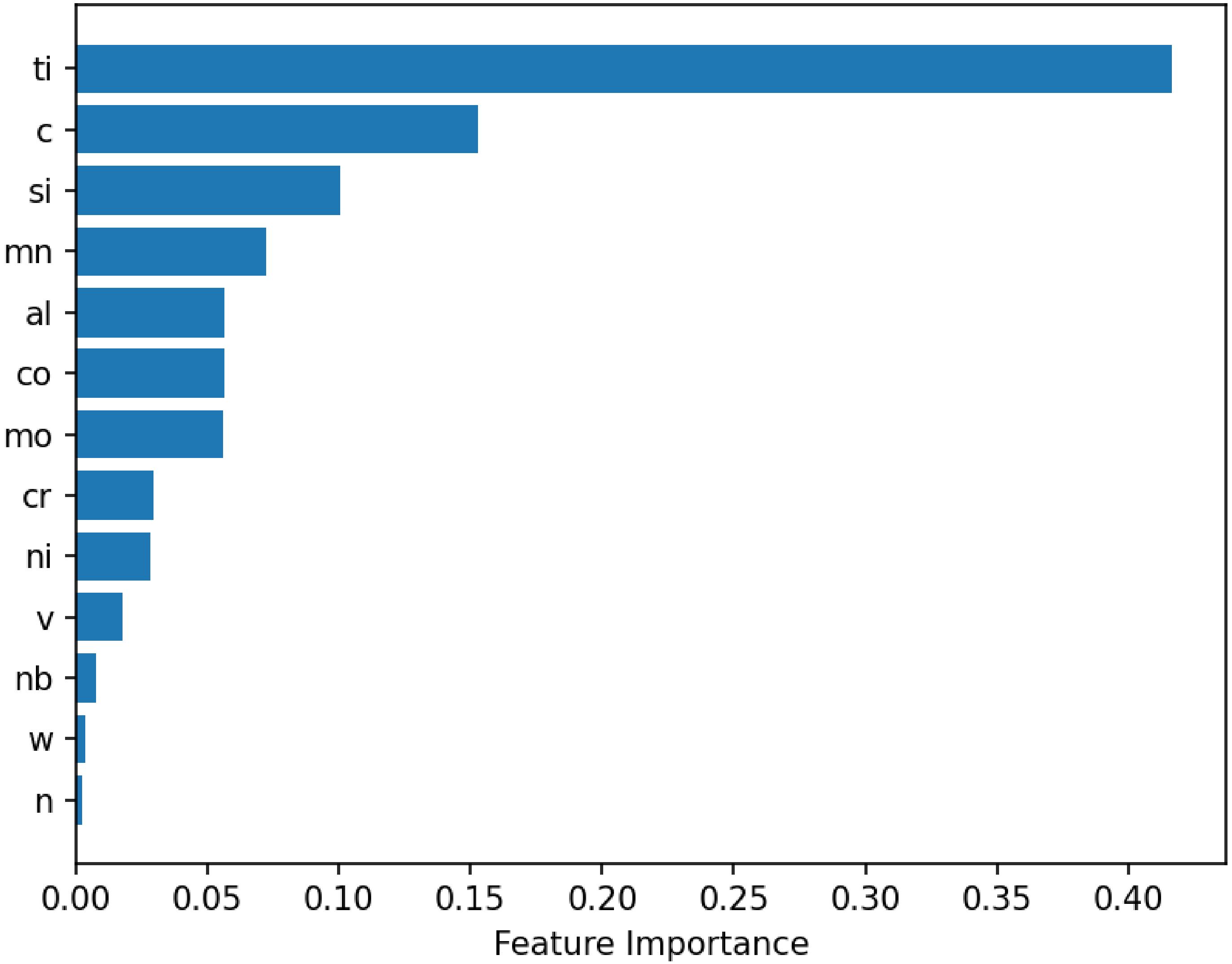
# Illustration (Afternoon tutorial 1)



- ➢ Comparison of linear regression and random forest regression for predicting steel strength from chemical composition.

# Descriptor importance

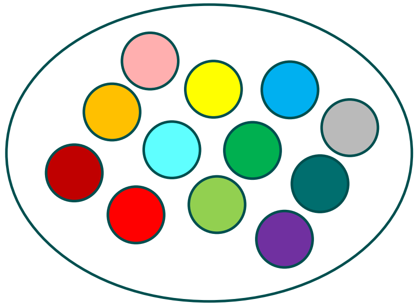
- From a single decision tree is easy to retrieve the dominant descriptors.
- Bagging makes interpretation of the decision tree difficult.
- However, the importance of a descriptor can still be inferred from the frequency a split is made with respect to that descriptor, averaged over all B trees.
- A large value indicates an important predictor.



# Boosting

- ➢ Instead of combining trees in parallel, boosting applies different trees sequentially.

Set $\hat{f}(\boldsymbol{x}) = 0$ and $r_i = y_i$ for all i in the training set.



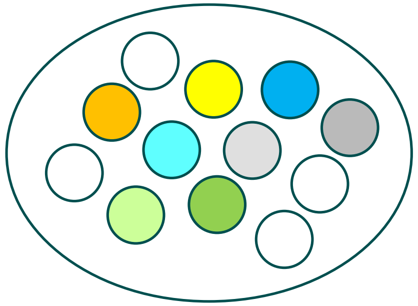
Original Data

Loop

- Train decision tree $\hat{f}^{\,i}$.
- Add a shrunken version

$$\hat{f}(\boldsymbol{x}) + \lambda \hat{f}^{\,i}(\boldsymbol{x}) \rightarrow \hat{f}(\boldsymbol{x})$$

- Update residual $r_i - \lambda \hat{f}^{\,i}(\boldsymbol{x}) \rightarrow r_i$



New Data

Ensemble regressor

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda \hat{f}^{\,i}(\boldsymbol{x})$$

# Applications of machine learning in materials science

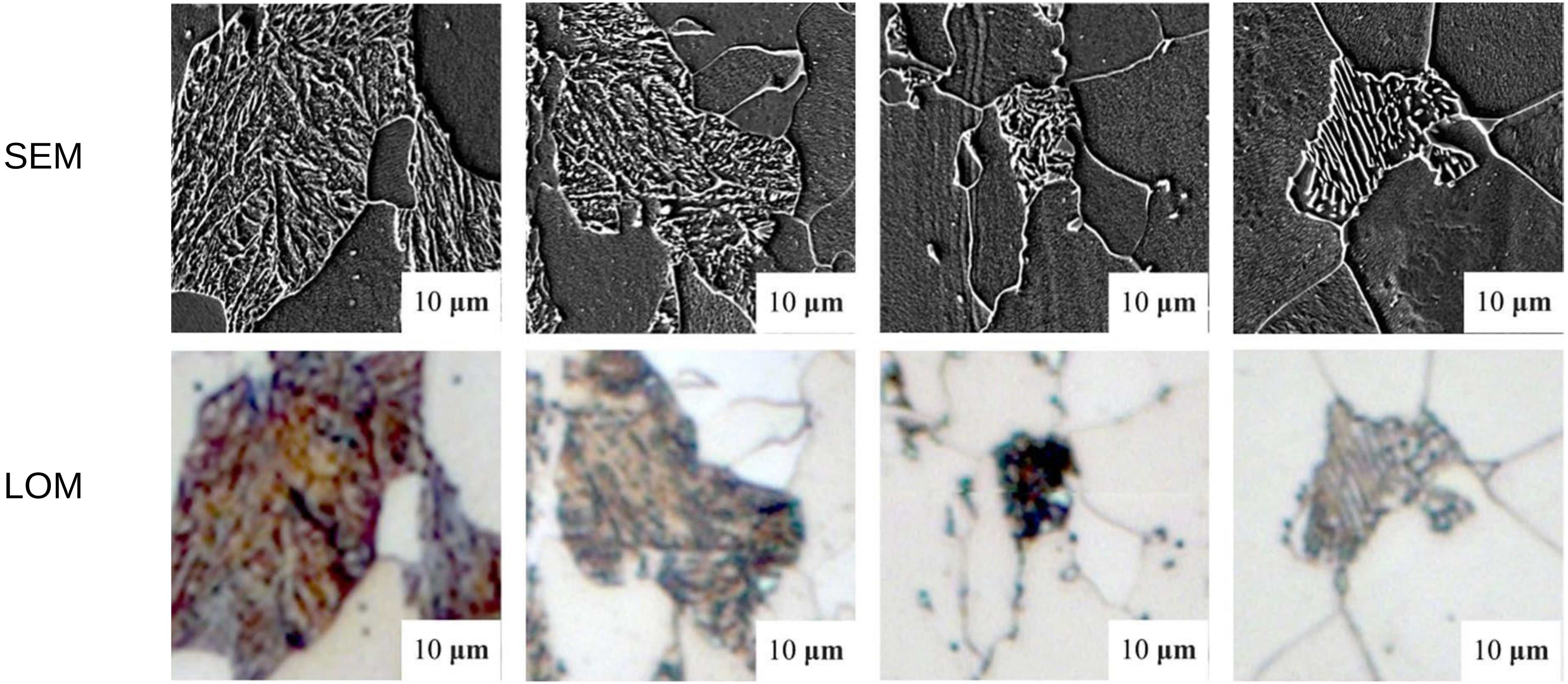
# Overview

- Automatically process or interpret analytical experiments.
  - Big data from optical or electron microscopy
  - Atom probe tomography
- Directly learn process-structure-property relationships of materials
  - Typically data is not “big”, datapoints are very expensive.
  - Predict materials properties.
  - Optimize new materials.
- Replace expensive physics based simulations (from 3rd paradigm):
  - Optimization of codes has reached a level that is difficult to improve further.
  - Use machine learning to construct surrogate models
  - Speed up calculations for optimization tasks

# Overview

- Automatically process or interpret analytical experiments.
  - ✧Big data from optical or electron microscopy
  - ✧Atom probe tomography

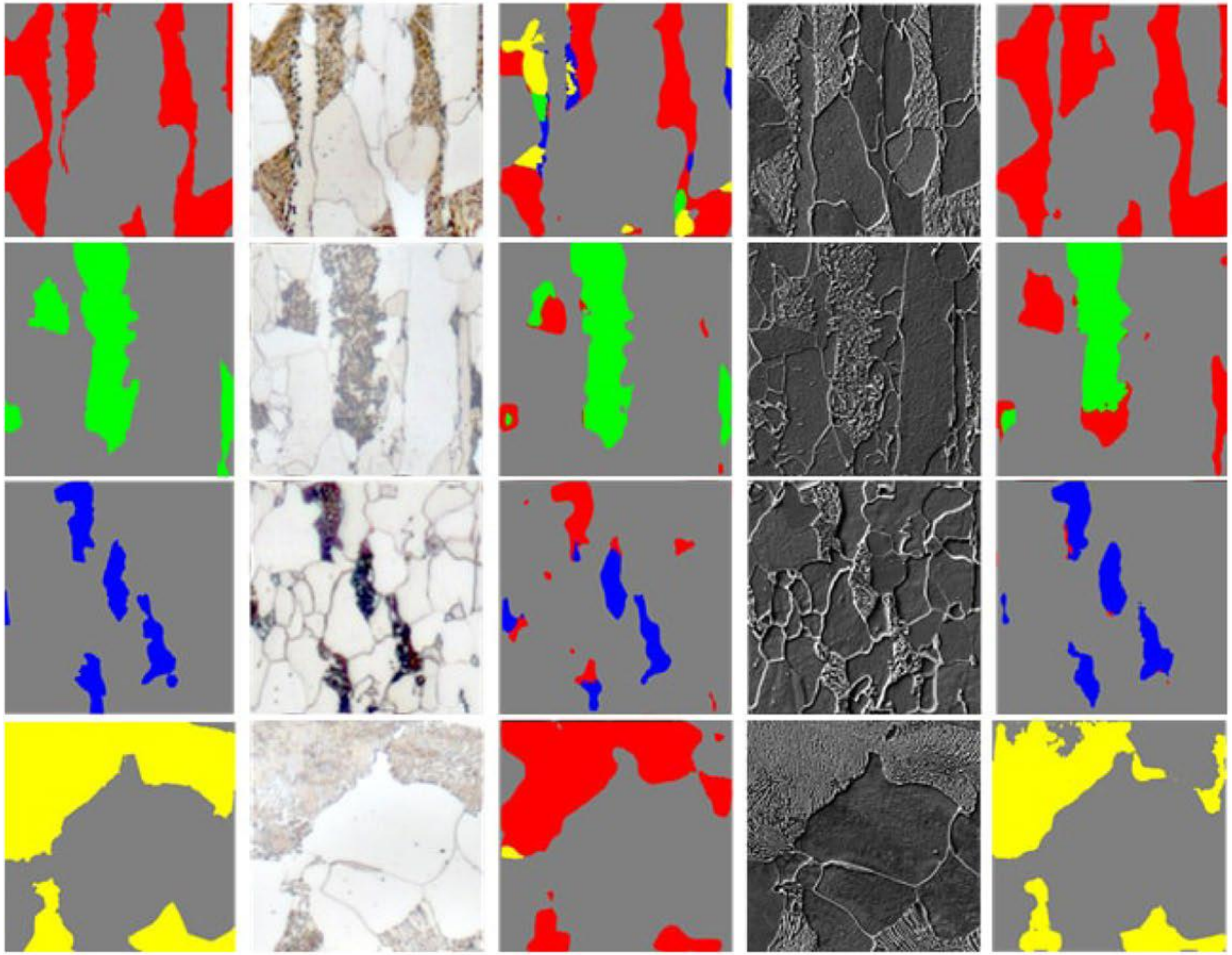
# Example: Classification of microstructure



10.1038/s41598-018-20037-5

- From left to right: martensite, tempered martensite, bainite and pearlite.
- Ferrite is the matrix phase in these images, having the role of the background.

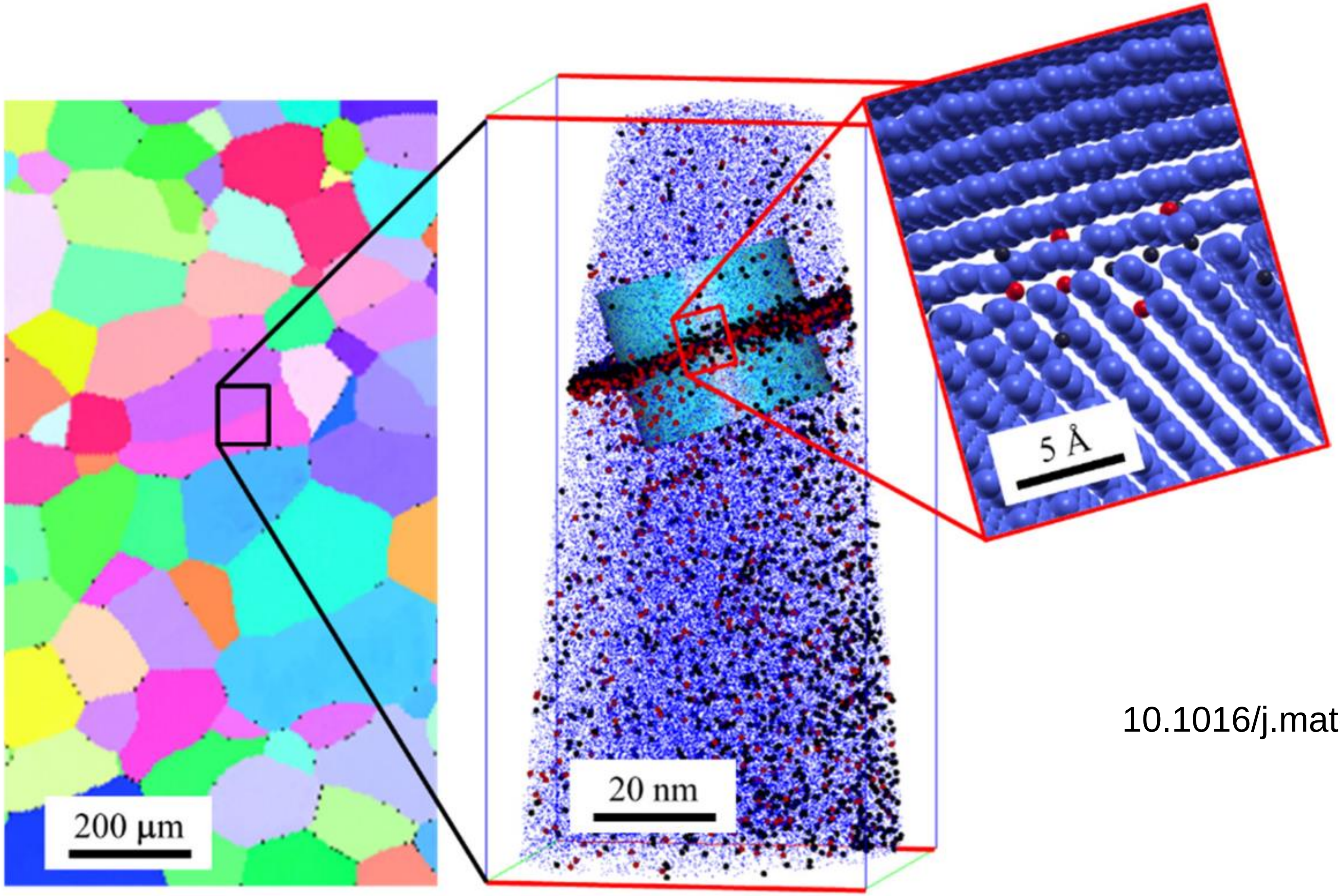
# Classification with convolutional neural networks



10.1038/s41598-018-20037-5

- Microstructure can be successfully classified
- The ground truth colors of martensite, tempered martensite, bainite and pearlite are red, green, blue, and yellow, respectively.
- Ground truth refers to data assumed to be correct.
- SEM images are more suitable for classification compared to LOM

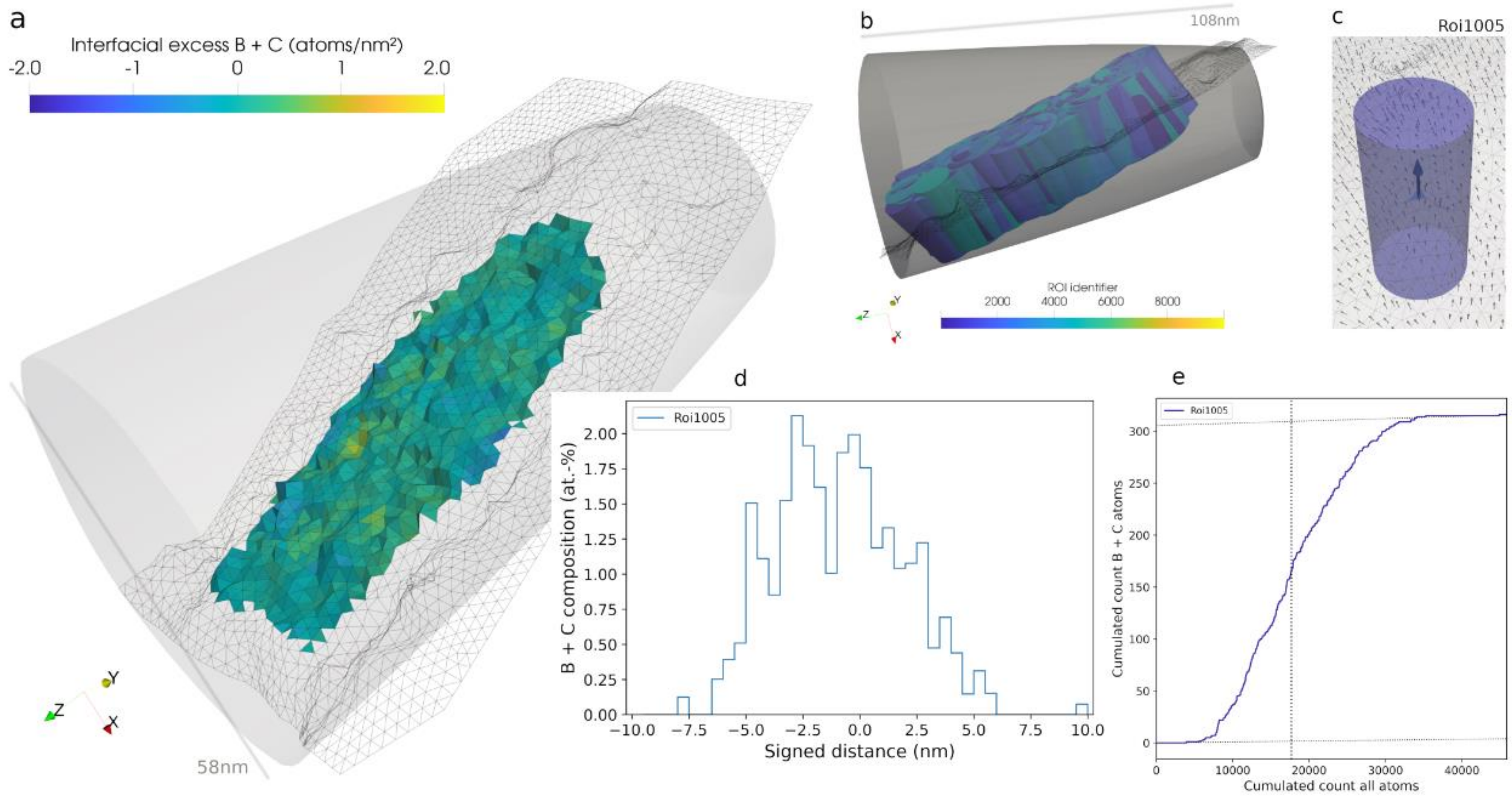
# Example: Atom probe tomography



10.1016/j.matdes.2018.01.012

# Example: Atom probe tomography

https://doi.org/10.48550/arXiv.2205.13510

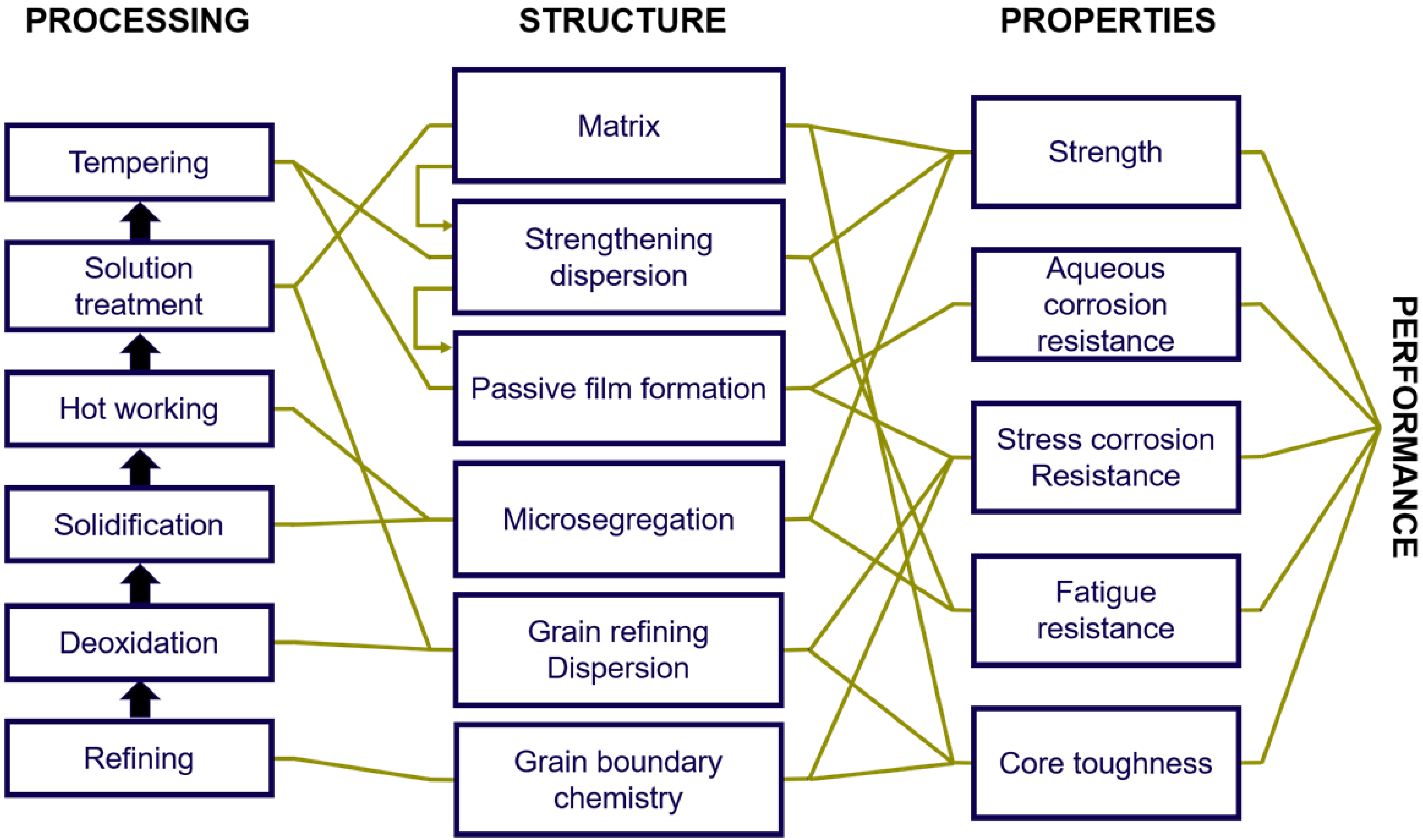


- ➢ Use of machine learning to automatically detect grain boundary surfaces and segregation excess

# Overview

- ➢ Automatically process or interpret analytical experiments.
  - ✧Big data from optical  or electron microscopy
  - ✧Atom probe tomography
- ➢ Directly learn process-structure-property relationships of materials
  - ✧Typically data is not “big”, datapoints are very expensive.
  - ✧Predict materials properties.
  - ✧Optimize new materials.

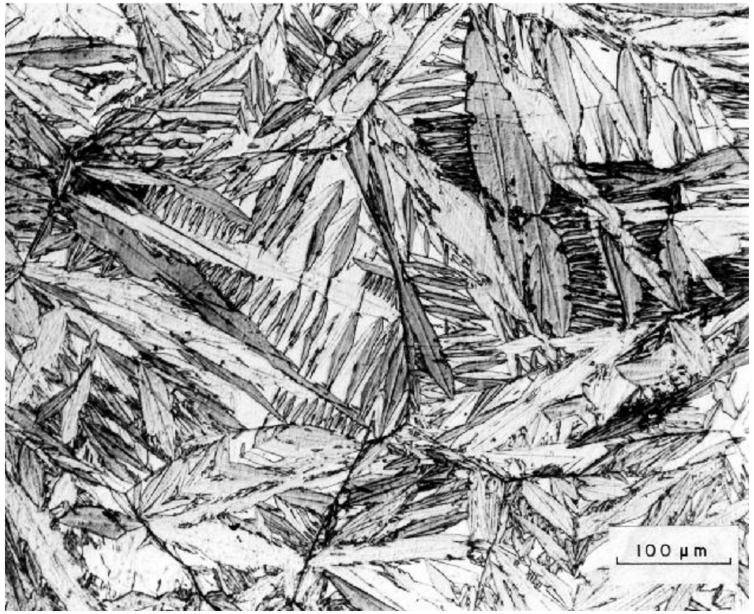
# Process-structure-property relationships



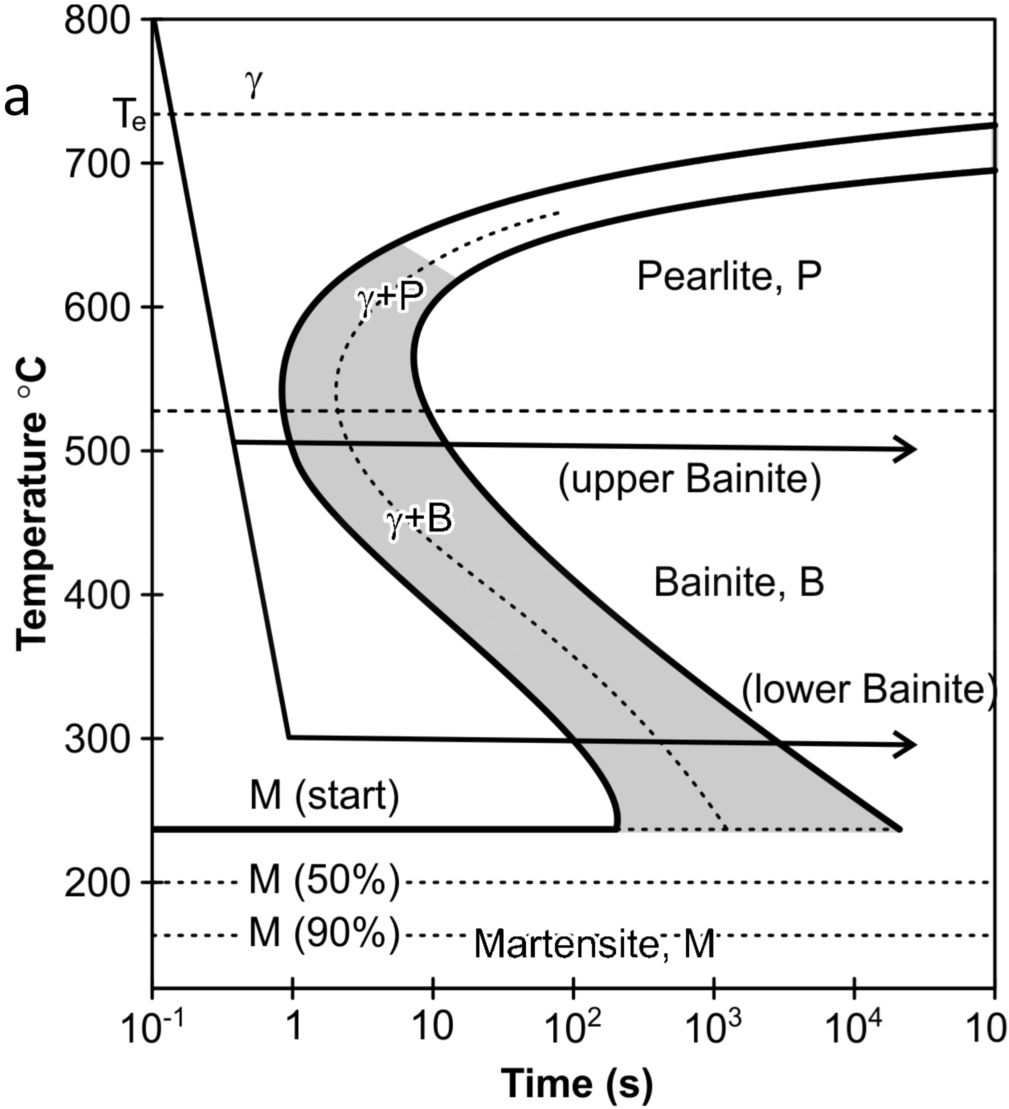
https://questekeurope.com/materials-by-design/icme-and-aim/

# Predicting the martensite start temperature

- Martensite is a metastable and very hard form of a steel microstructure.
- Depending on chemical composition and quench rate, martensite forms instead of other steel microstructures such as bainite or pearlite.



10.1590/1516-1439.000215



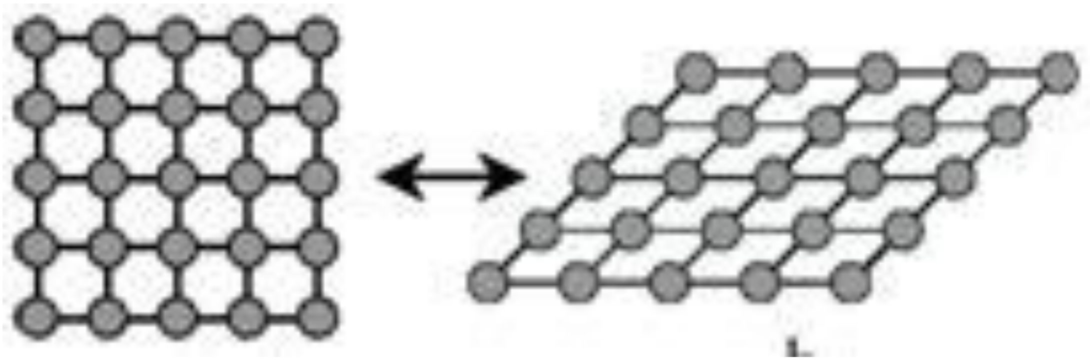
https://testbook.com/objective-questions/mcq-on-time-temperature-transformation-ttt-curve--5eea6a0939140f30f369d8f6

# Video of the martensitic transformation

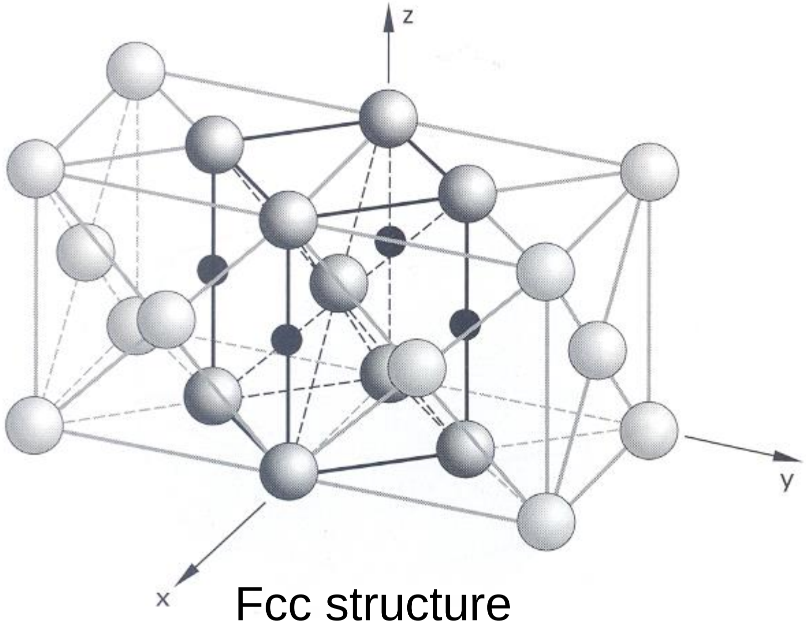
https://www.youtube.com/watch?v=OQ5lVjYssko

# The martensitic transformation

- Martensitic transformations are diffusionless shear transformations.
  - Cooperative motion of a large number of atoms, each being displaced by only a small distance (a fraction of an interatomic spacing) relative to its neighbours. → bct structure created.
  - Transformation is not thermally activated and proceeds at half the speed of sound.

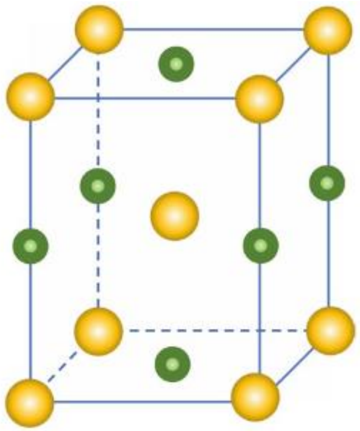


martensitic transformation from of a square lattice



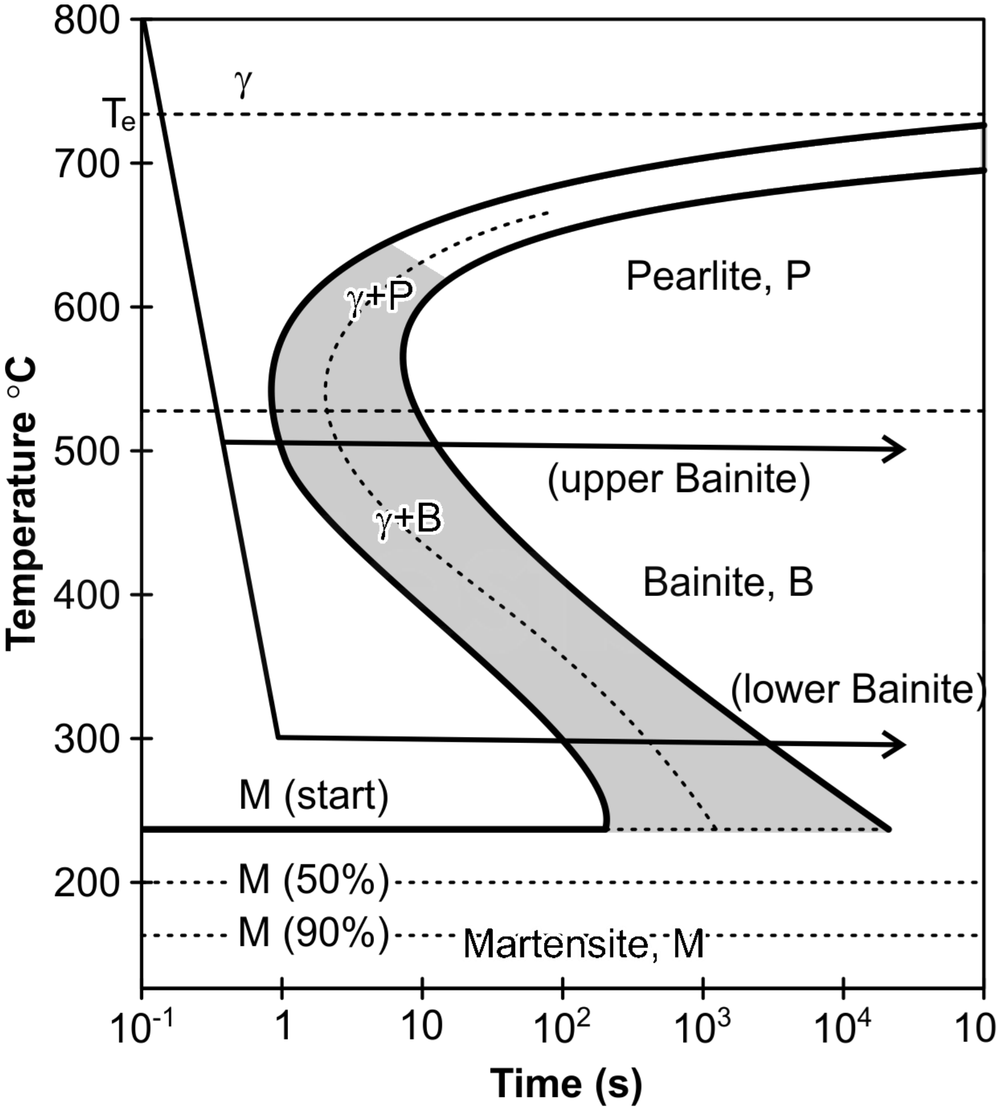
Fcc structure

Bct structure



# The martensite start temperature

- Trigger:
  - Free energy change for transformation without a composition change ($\Delta G^{\gamma\alpha'}$) reaches a critical driving value ($\Delta G_c$ ),
  - The exact magnitude of is determined by stored energies and kinetic phenomena.
- The temperature at which martensite is formed depends
  - On chemistry.
  - To some extent also on prior austenite grain size.



https://testbook.com/objective-questions/mcq-on-time-temperature-transformation-ttt-curve--5eea6a0939140f30f369d8f6

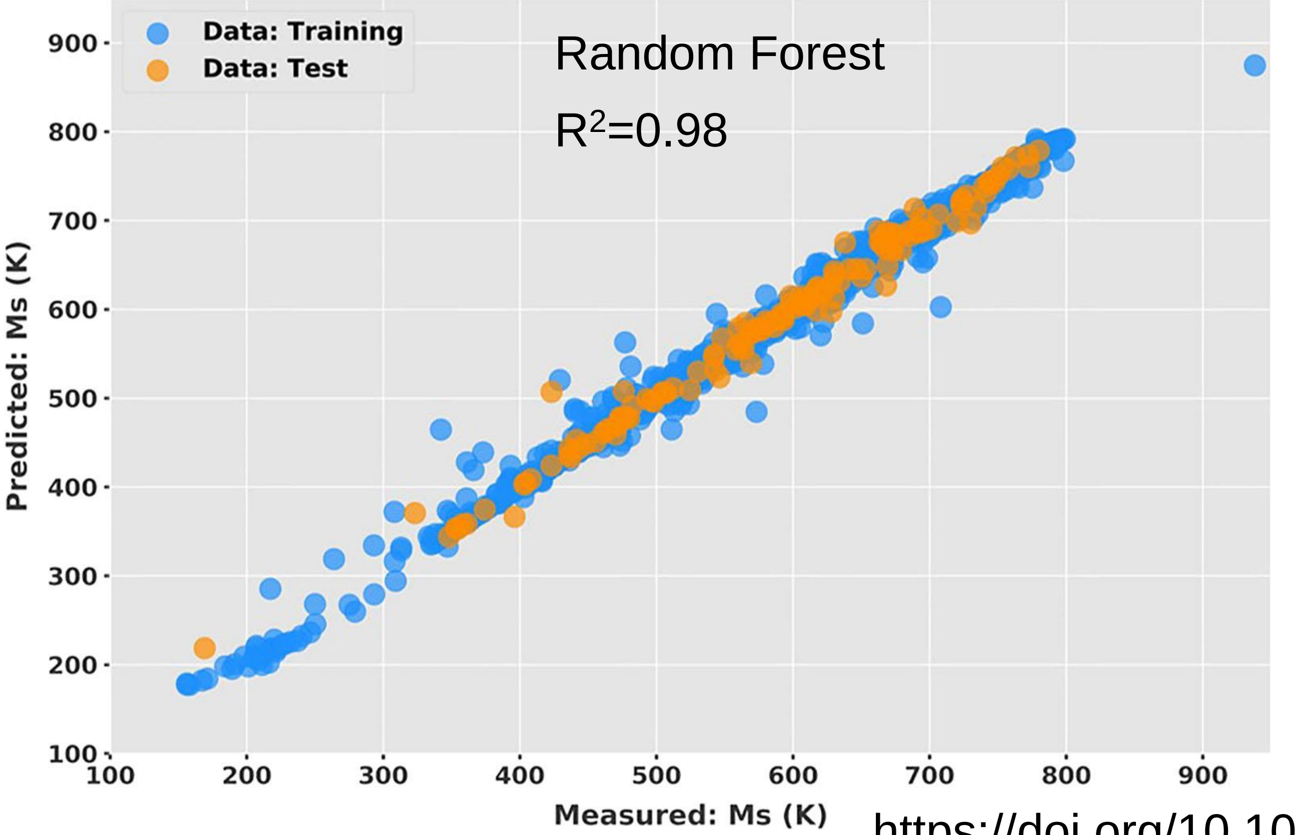
# Machine learning the martensite start temperature

- Two recent works available:
  - Paper 1: Rahaman, M., Mu, W., Odqvist, J. *et al.* “Machine Learning to Predict the Martensite Start Temperature in Steels”. *Metall Mater Trans A* 50, 2081–2091 (2019). 10.1007/s11661-019-05170-8.
  - Paper 2: Qi Lu, Shilong Liu, Wei Li, Xuejun Jin, „Combination of Thermodynamic Knowledge and Multilayer Feedforward Neural Networks for accurate Prediction of MS Temperature in Steels“, Materials & Design, 192, 108696, (2020). 10.1016/j.matdes.2020.108696.

# ML approach

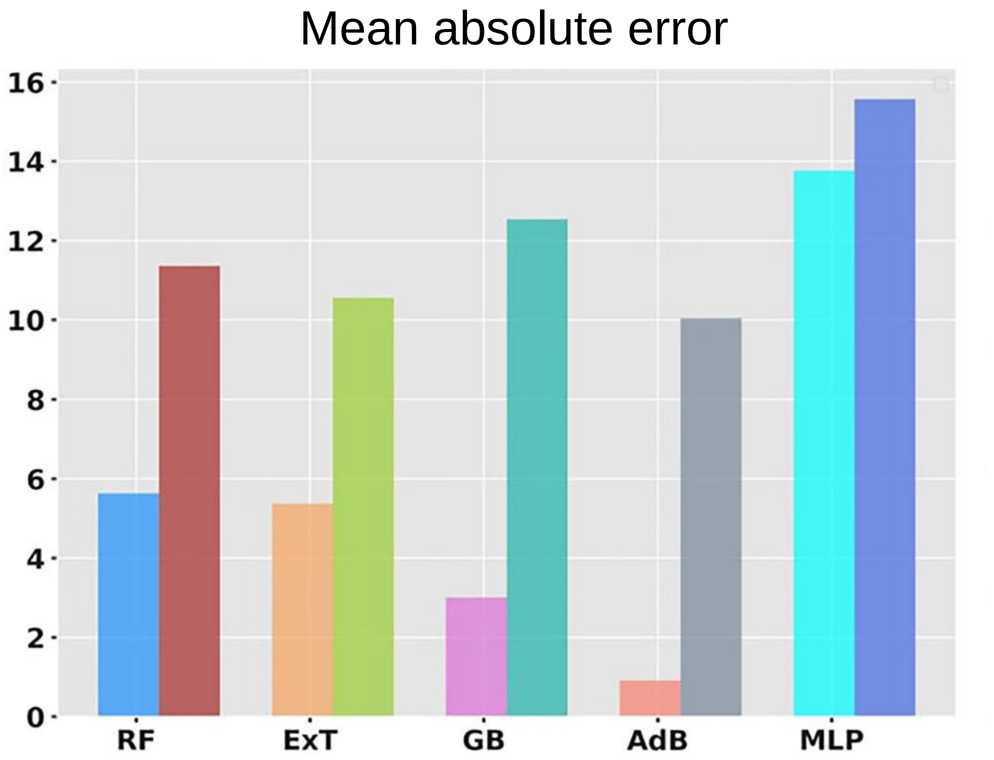
- Descriptors:
  - ✧ Chemical composition, i.e. content of C, Mn, Si, Cr, Ni, Mo, V, Co, Al, W, Cu, Nb, Ti, N, S, P, and B.
- Machine learning algorithms:
  - ✧ Bagging: Random forests (RFs), extremely randomized trees (ExT).
  - ✧ Boosting: Gradient boosting (GB), Adaboost (AdB).
  - ✧ Multilayer Perceptron (MLP).
- Data:
  - ✧ 2277 entries of $M_s$ vs chemical composition for binary, ternary, and multicomponent steel alloys.
  - ✧ Based on https://www.phase-trans.msm.cam.ac.uk/map/data/neural/neur-ms-b.html and extended.

# Results paper 1



Random Forest

$R^2$=0.98

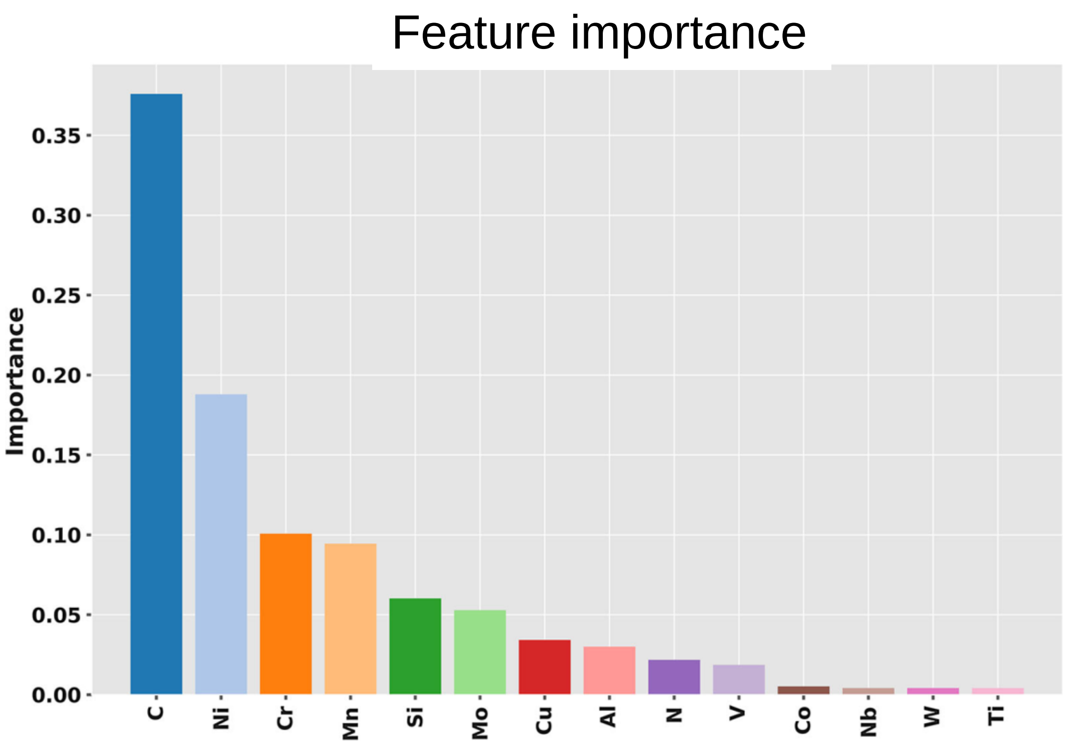


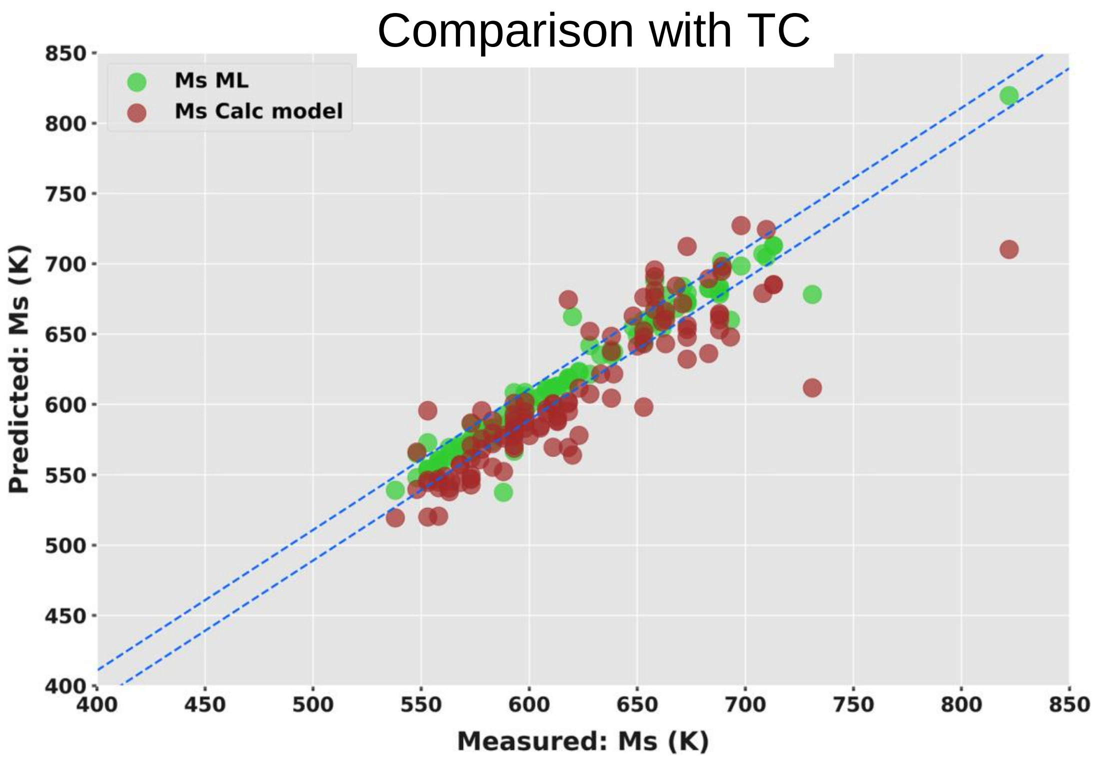
Mean absolute error

https://doi.org/10.1007/s11661-019-05170-8

- Good prediction is possible with all methods exhibiting $R^2$> 0.95.
- Best method is based on boosting, AdaBoost.

# Results paper 1


Feature importance


Comparison with TC

https://doi.org/10.1007/s11661-019-05170-8

- Carbon has the strongest influence on $M_S$, reducing it. Similar effect is with Ni, a austenite stabilizer. The only two solutes increasing $M_S$ are Al and Co.
- The model performs slightly better compared to the semi-empirical model implemented in ThermoCalc (10.1007/s11661-012-1171-z).

# Approach paper 2

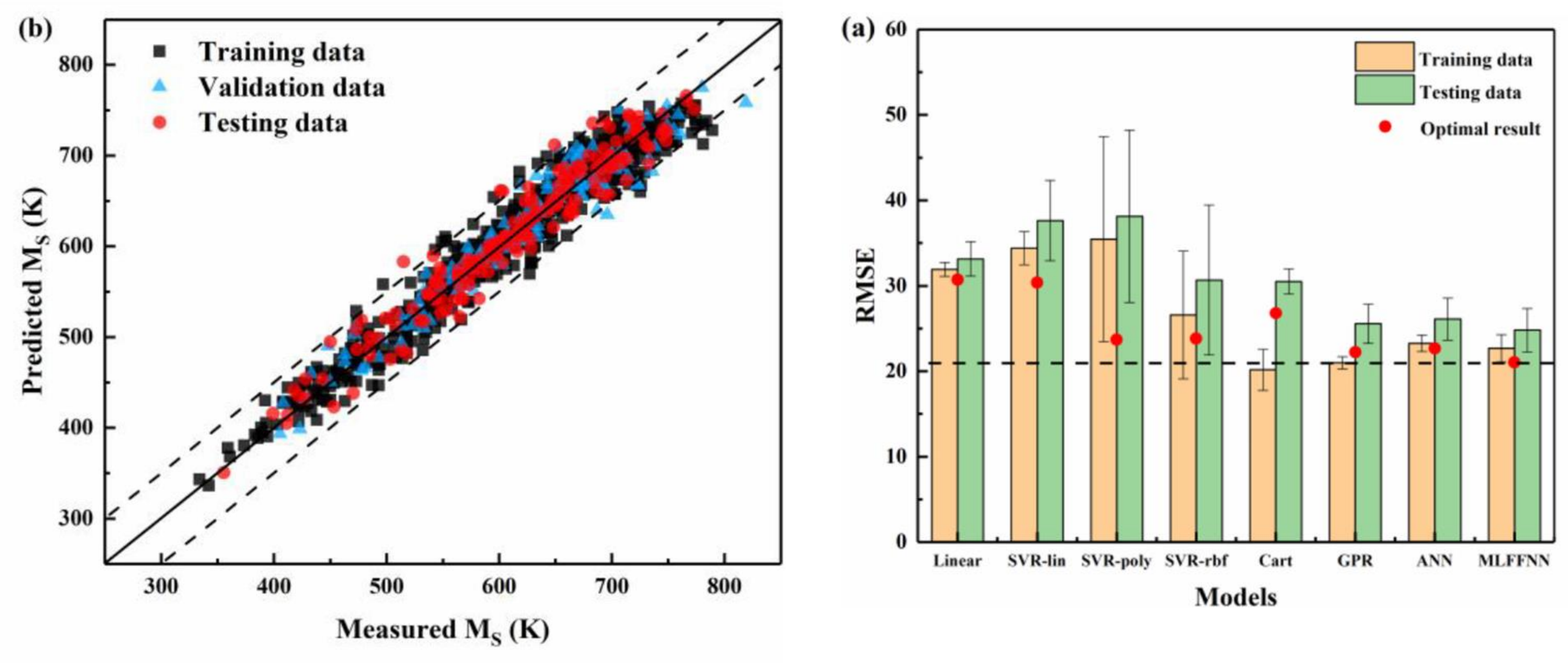
- Descriptors:
  - As paper 1, concentrations: C, Mn, Si, Cr, Ni, Mo, V, Co, Al, W, Cu, Nb, Ti, N, S, P, and B.
  - Austenitization temperature ($T_\gamma$) as a rough indicator of prior austenite grain size which has been shown to contribute in 10.1016/j.actamat.2016.12.029
  - Critical driving force ($\Delta G_c$) which provides $M_s$ with an available CALPHAD model

$$\Delta G_c = K_1 + \sqrt{\sum\left(K_\mu^i X_i^{0.5}\right)} + \sqrt{\sum\left(K_\mu^j X_j^{0.5}\right)} + \sqrt{\sum\left(K_\mu^k X_k^{0.5}\right)} + K_\mu^{Co} X_{Co}^{0.5}$$

Gosh et al. https://doi.org/10.1016/0956-7151(94)90468-5

- Machine learning algorithms
  - Linear regression, Gaussion Process Regression, Support Vector Regression
  - Multilayer Feed Forward Neural Network
- Data
  - About 2000 data points from literature. Test experiments.

# Results paper 2



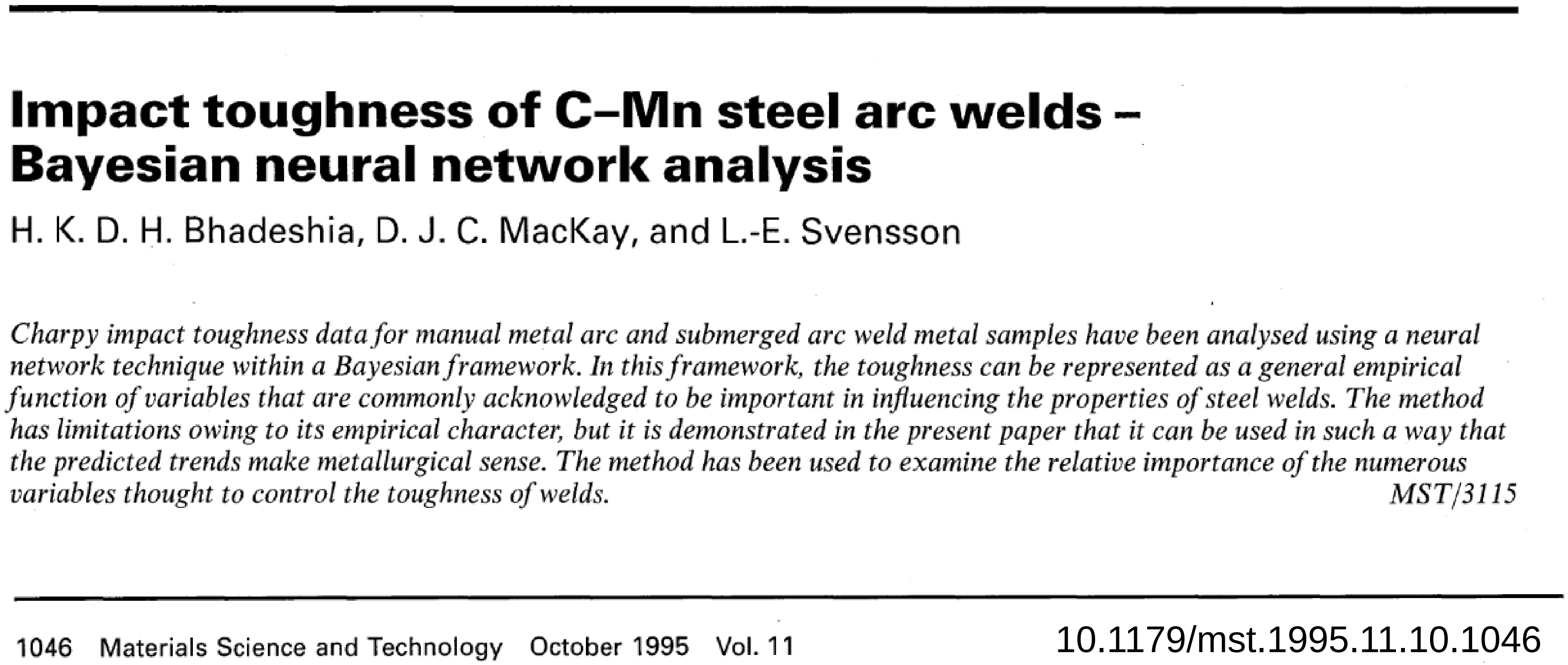
➢ Comparison of ML models:  In contrast to paper 1, neural network is the best approach.

# Afternoon tutorial on steel properties

- Predict the martensite start temperature from steel compositions
  - https://www.sciencedirect.com/science/article/pii/S0264127520302306

# Towards property prediction

- Historically one of the first works on predicting steel properties with ML.



## Impact toughness of C–Mn steel arc welds – Bayesian neural network analysis

H. K. D. H. Bhadeshia, D. J. C. MacKay, and L.-E. Svensson

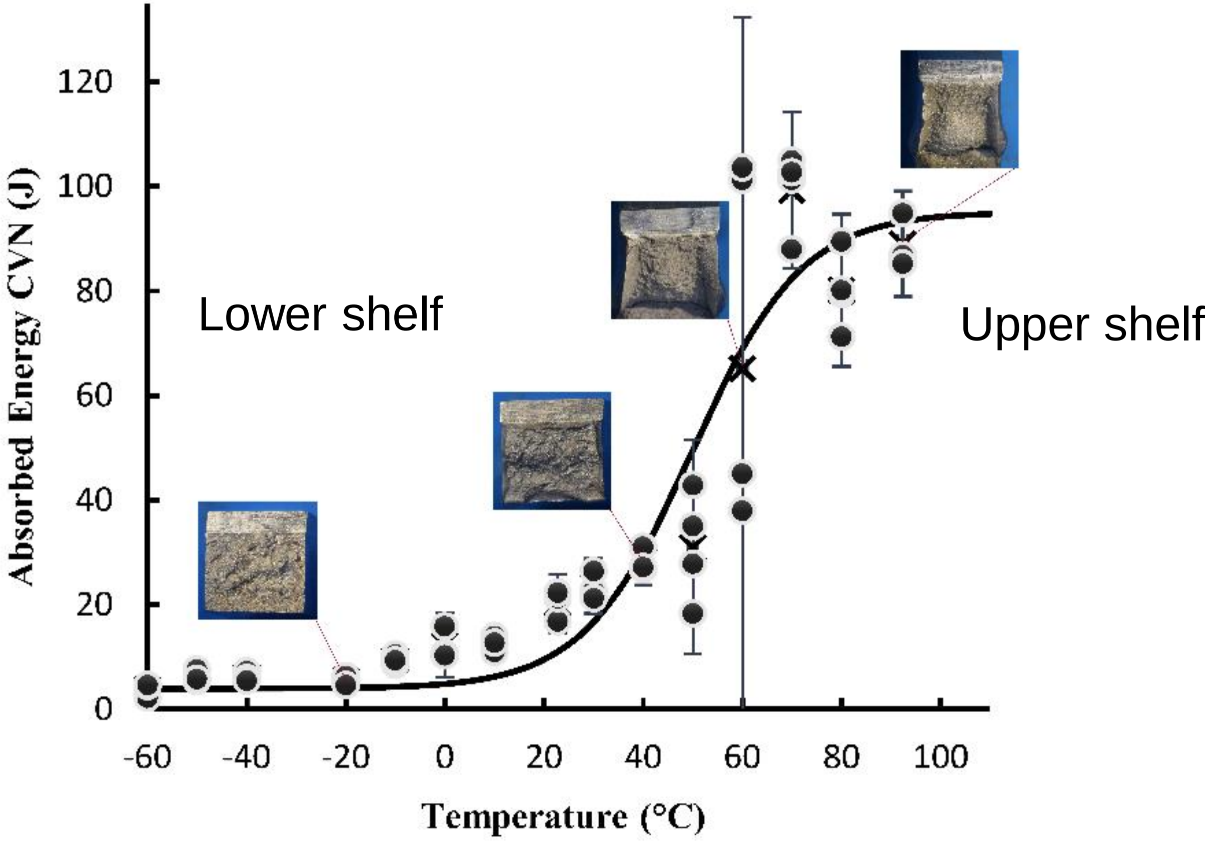
*Charpy impact toughness data for manual metal arc and submerged arc weld metal samples have been analysed using a neural network technique within a Bayesian framework. In this framework, the toughness can be represented as a general empirical function of variables that are commonly acknowledged to be important in influencing the properties of steel welds. The method has limitations owing to its empirical character, but it is demonstrated in the present paper that it can be used in such a way that the predicted trends make metallurgical sense. The method has been used to examine the relative importance of the numerous variables thought to control the toughness of welds.* *MST/3115*

1046 Materials Science and Technology October 1995 Vol. 11

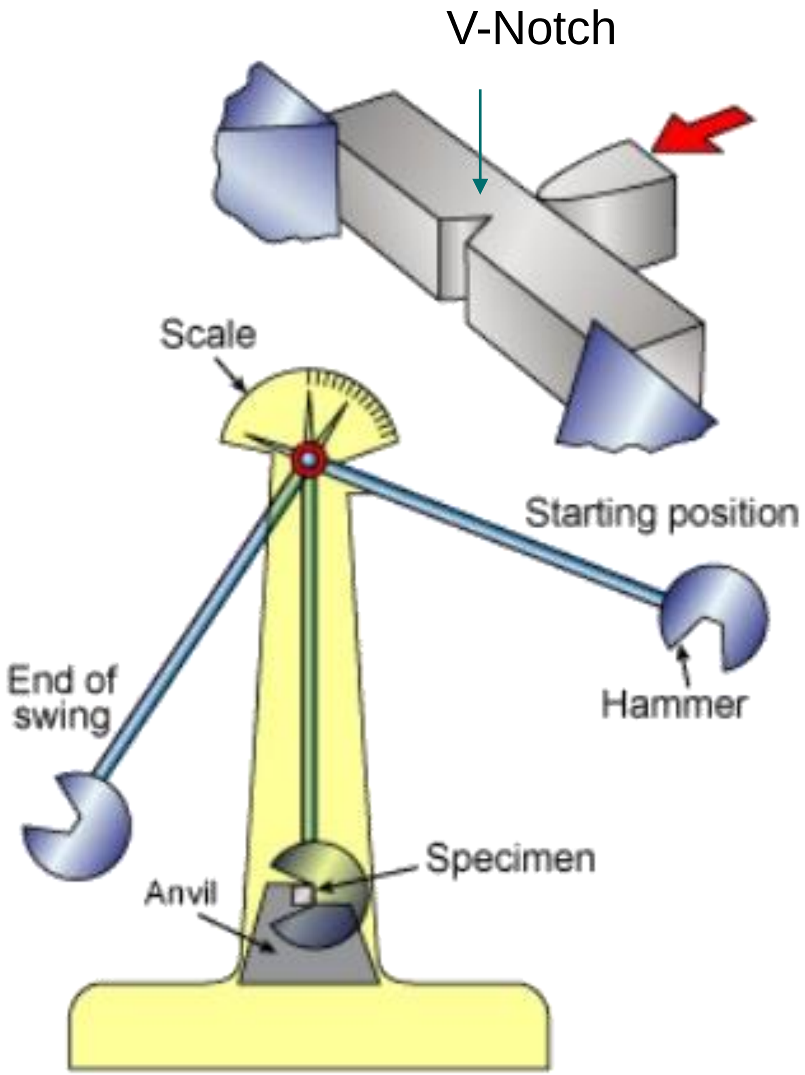
10.1179/mst.1995.11.10.1046

# Target quantity

- Charpy impact toughness of steel arc welds.



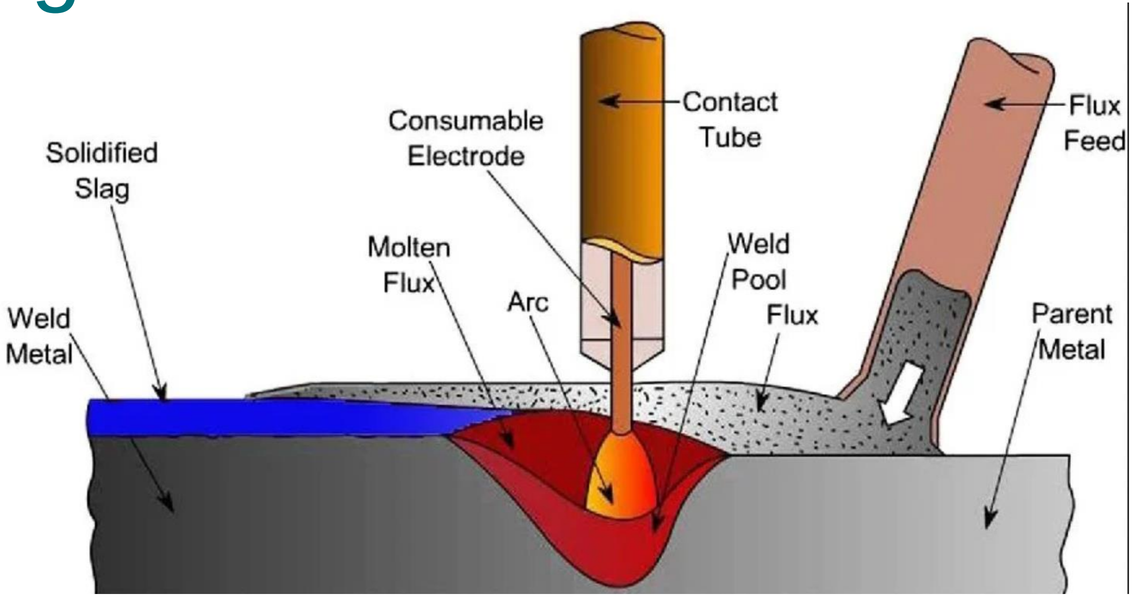
https://www.redalyc.org/journal/849/84961237003/html/



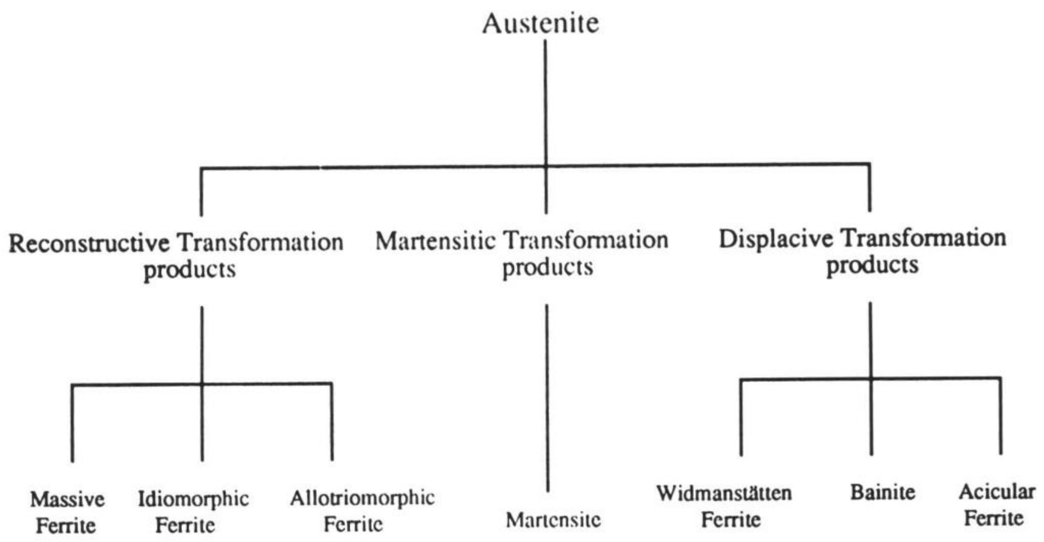
https://www.totalmateria.com/page.aspx?ID=CheckArticle&site=kts&NM=534

# Welding

- Two types of welding employed:
  - ✧Submerged arc welding.
  - ✧Manual welding.
- Several runs applied, metal is re-heated several times → secondary microstructure formed.
- Several different ferritic microstructures are involved:
  - ✧Allotriomorphic ferrite
  - ✧Acicular ferrite
  - ✧Widmannstätten ferrite

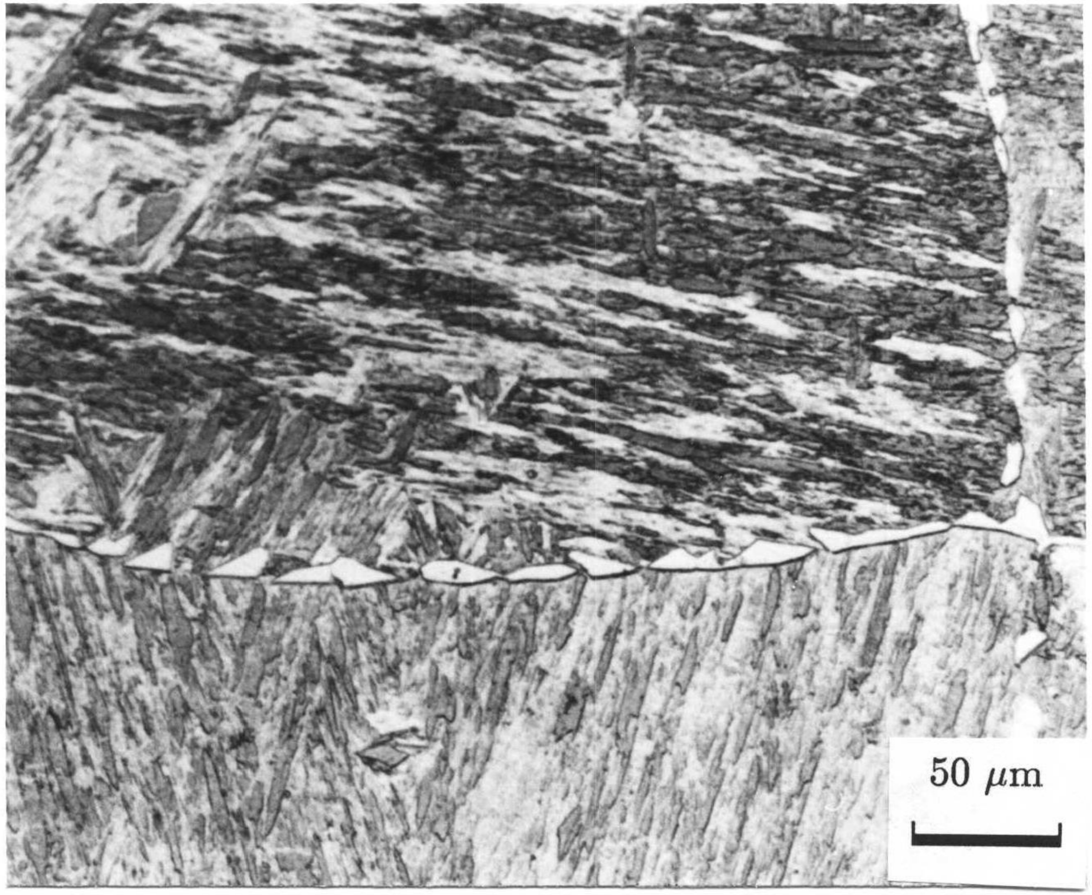


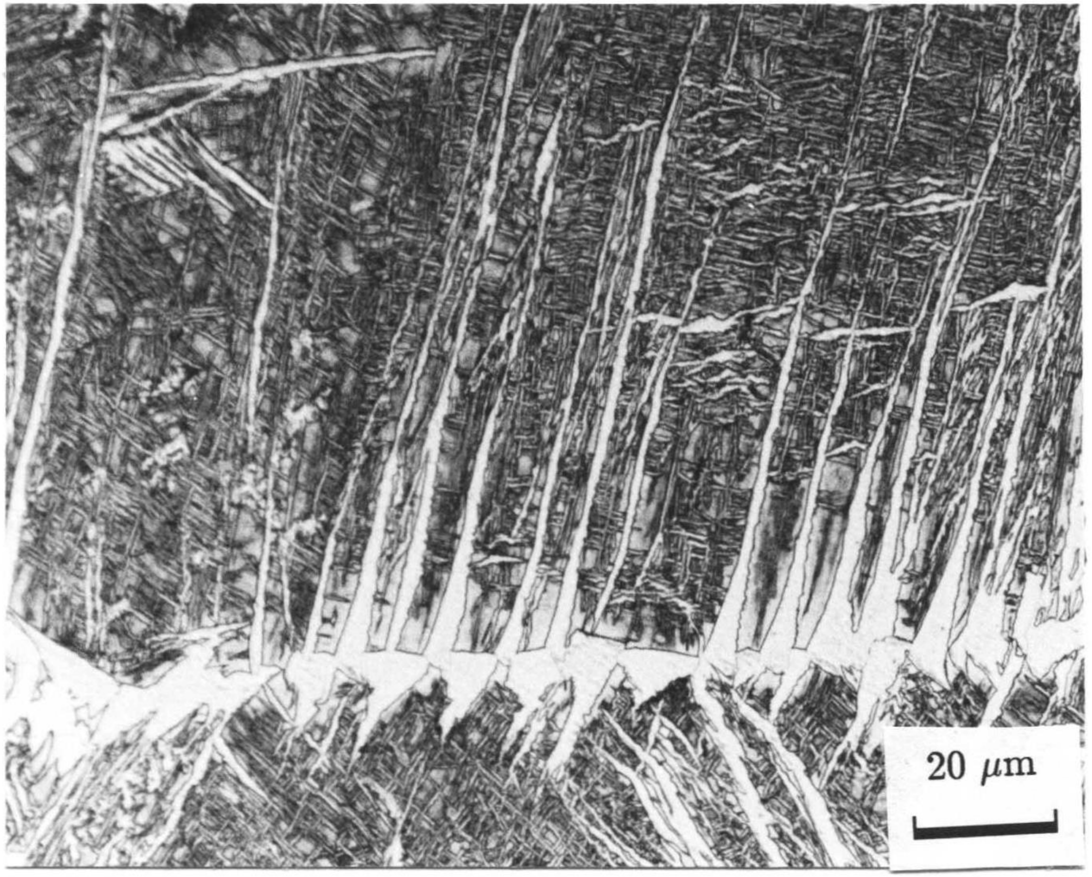
https://www.westermans.com/submerged-arc-welding-process.aspx



PhD thesis Ashraf Ali „Widmanstaetten Ferrite and Bainite in Ultra High Strength-steels“

# Microstructure elements





- Allotriomorphic ferrite, reconstructive.
- Widmannstätten ferrite, displacive.

# ML descriptors

- In total 14 descriptors used.
- High yield strength, low toughness
- Main alloying elements steering phase transformation and strength.
- Embrittling elements and inclusion formers.
- Microstructure description.
- Ductile-to-brittle transition.

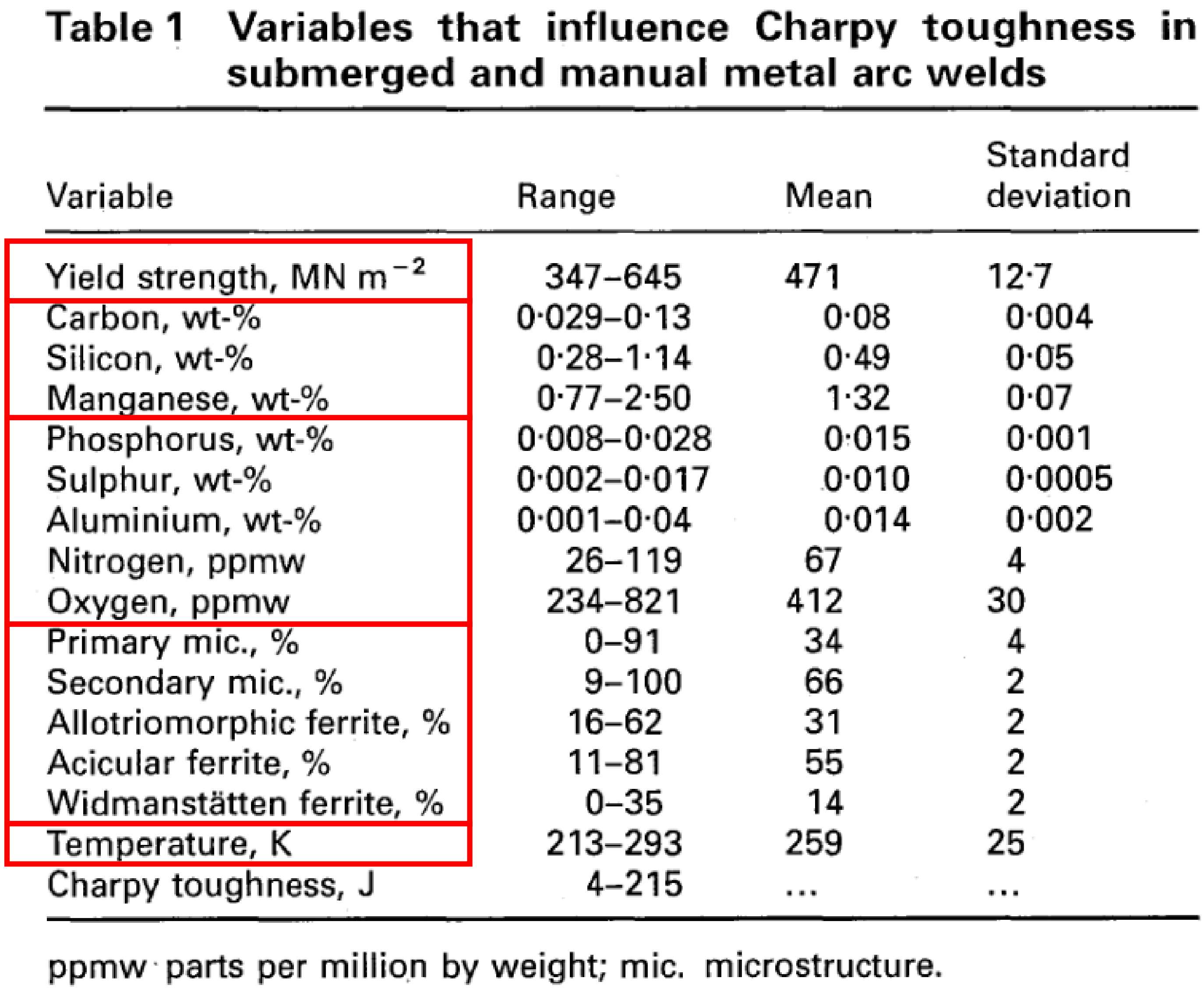


**Table 1 Variables that influence Charpy toughness in submerged and manual metal arc welds**

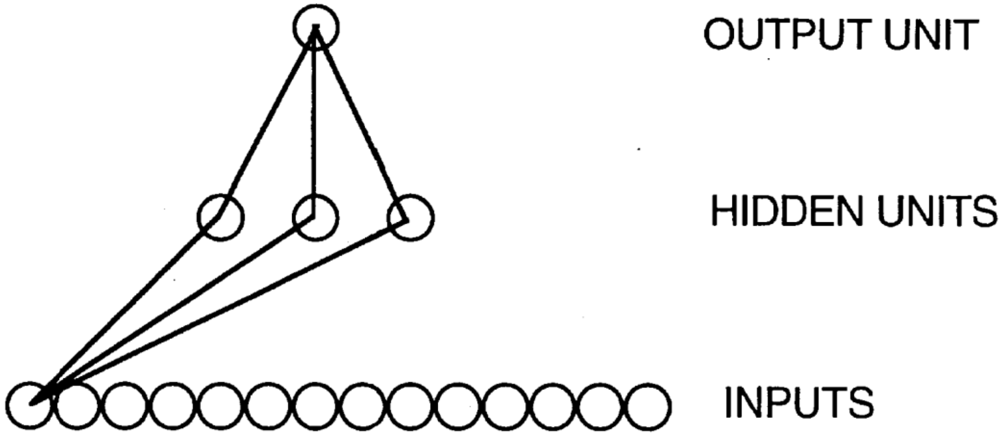
| Variable | Range | Mean | Standard deviation |
|---|---|---|---|
| Yield strength, MN m$^{-2}$ | 347–645 | 471 | 12·7 |
| Carbon, wt-% | 0·029–0·13 | 0·08 | 0·004 |
| Silicon, wt-% | 0·28–1·14 | 0·49 | 0·05 |
| Manganese, wt-% | 0·77–2·50 | 1·32 | 0·07 |
| Phosphorus, wt-% | 0·008–0·028 | 0·015 | 0·001 |
| Sulphur, wt-% | 0·002–0·017 | 0·010 | 0·0005 |
| Aluminium, wt-% | 0·001–0·04 | 0·014 | 0·002 |
| Nitrogen, ppmw | 26–119 | 67 | 4 |
| Oxygen, ppmw | 234–821 | 412 | 30 |
| Primary mic., % | 0–91 | 34 | 4 |
| Secondary mic., % | 9–100 | 66 | 2 |
| Allotriomorphic ferrite, % | 16–62 | 31 | 2 |
| Acicular ferrite, % | 11–81 | 55 | 2 |
| Widmanstätten ferrite, % | 0–35 | 14 | 2 |
| Temperature, K | 213–293 | 259 | 25 |
| Charpy toughness, J | 4–215 | ... | ... |

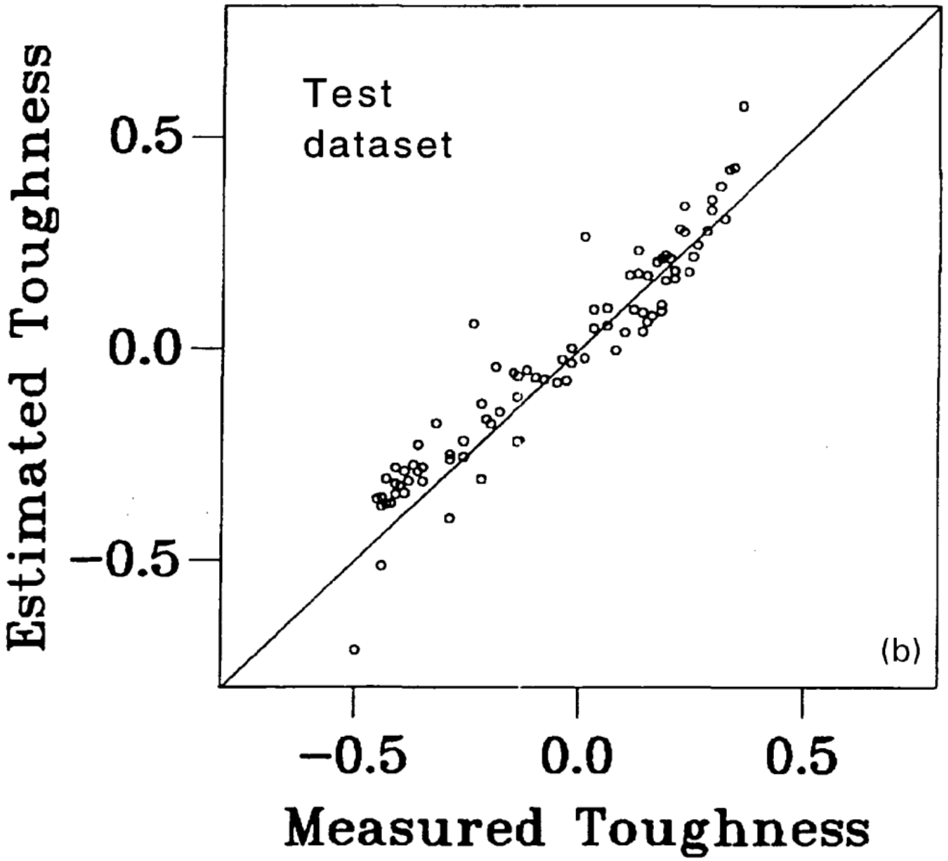
ppmw parts per million by weight; mic. microstructure.

10.1179/mst.1995.11.10.1046

# Results

- Bayesian Neural Networks:
  - ✧Feedforward Neural Network with 1 hidden layer and 4 units.
  - ✧Parameter identification with Bayesian approach.



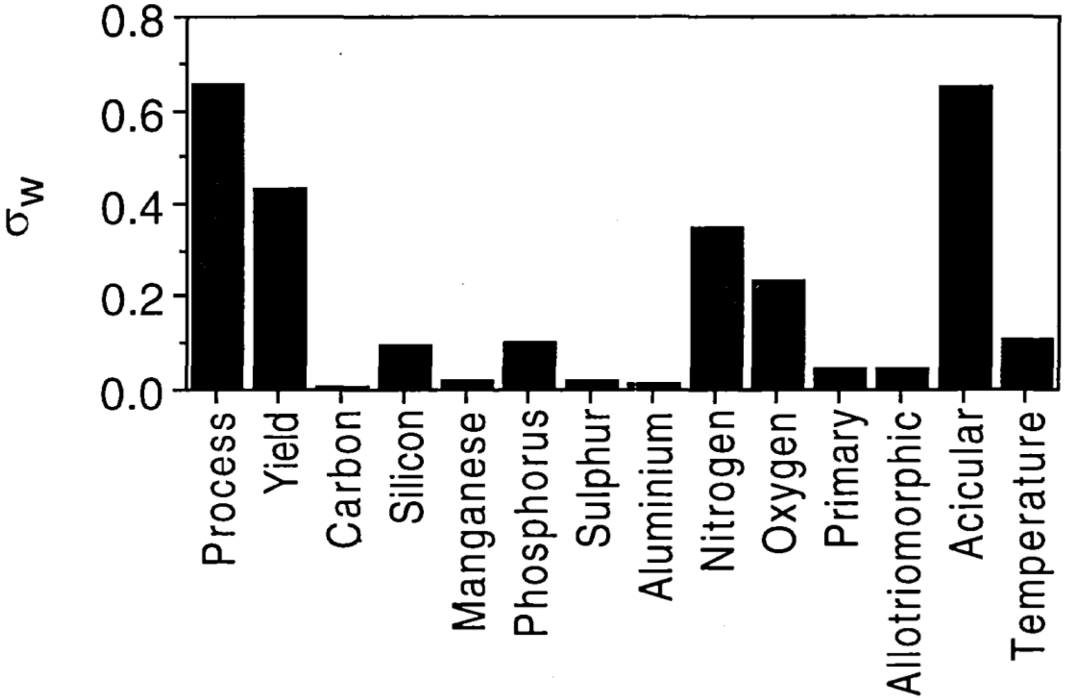


10.1179/mst.1995.11.10.1046

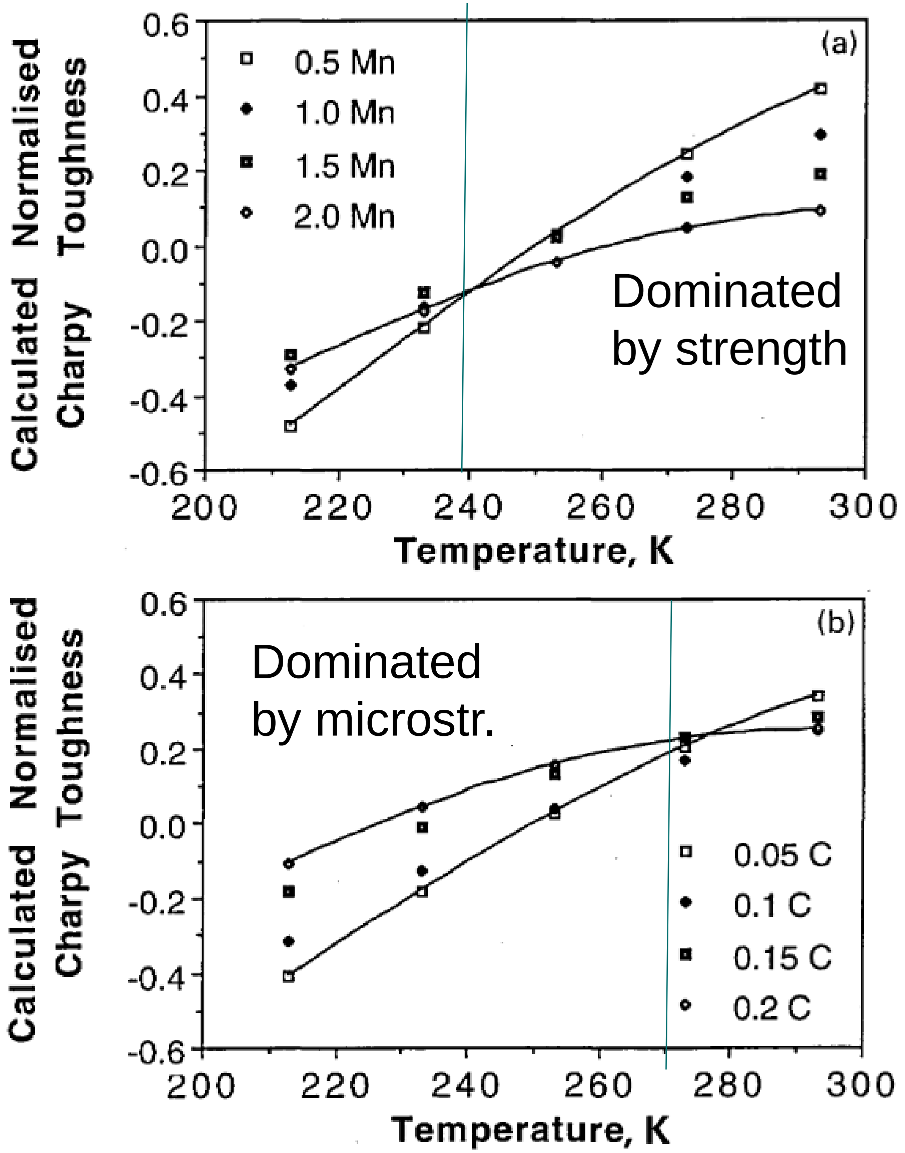
# Results

- Analysis of descriptor importance:
  - ✧Process (manual vs submerged) is the most important quantity.
  - ✧Yield strength is important.
  - ✧Acicular microstructure.

10.1179/mst.1995.11.10.1046



- Dependence on chemistry:
  - ✧Mn and C increase acicular fraction and strength.



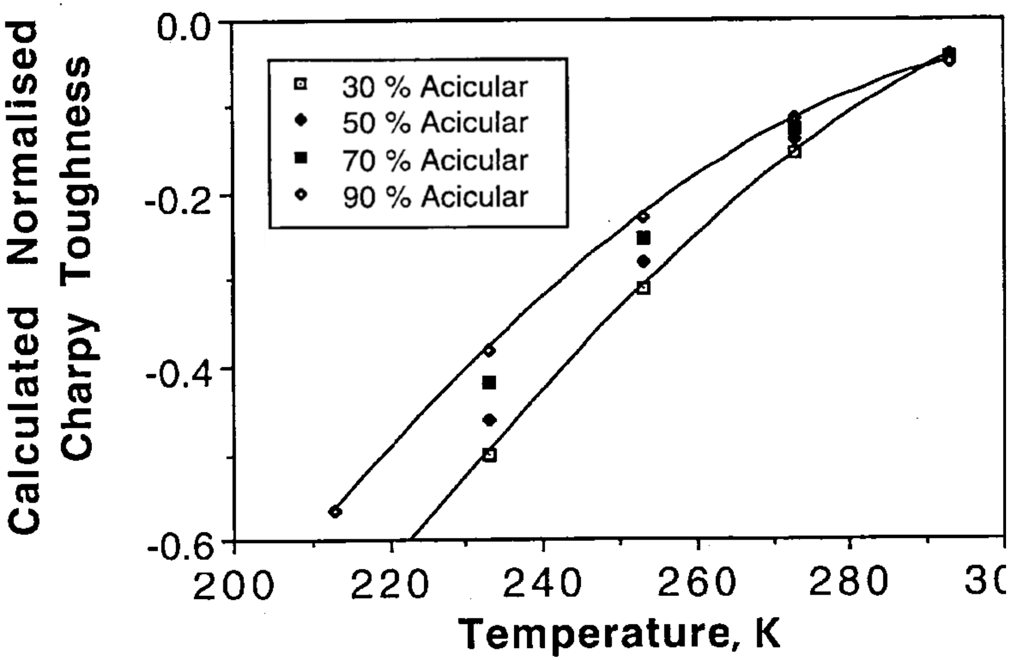
Dominated by strength

Dominated by microstr.

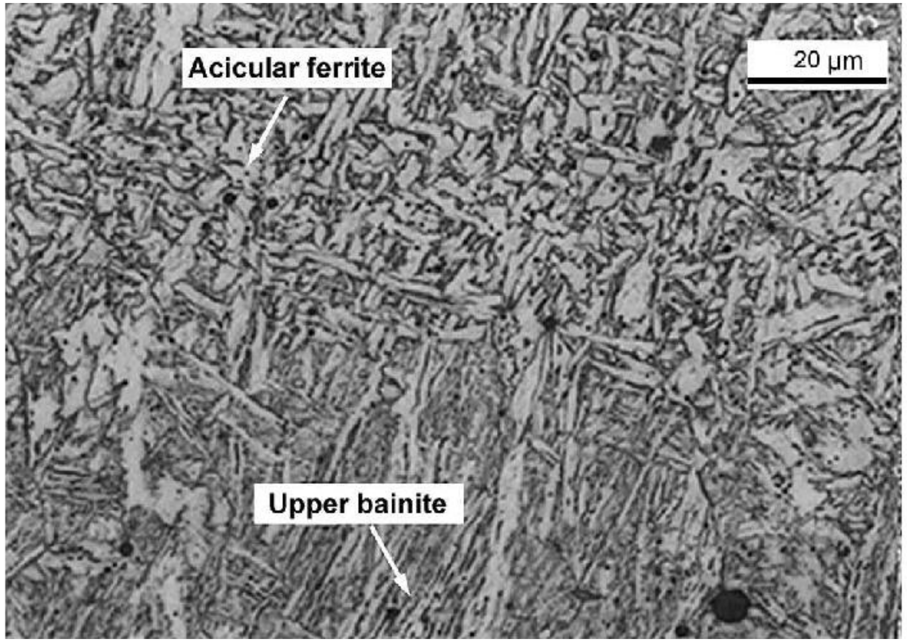
# Acicular microstructure

- Acicular ferrite is a better microstructure than Widmansstätten ferrite:
  - ✧less organised arrangement of ferrite plates.
  - ✧greater capacity to deflect crack.

10.1179/mst.1995.11.10.1046

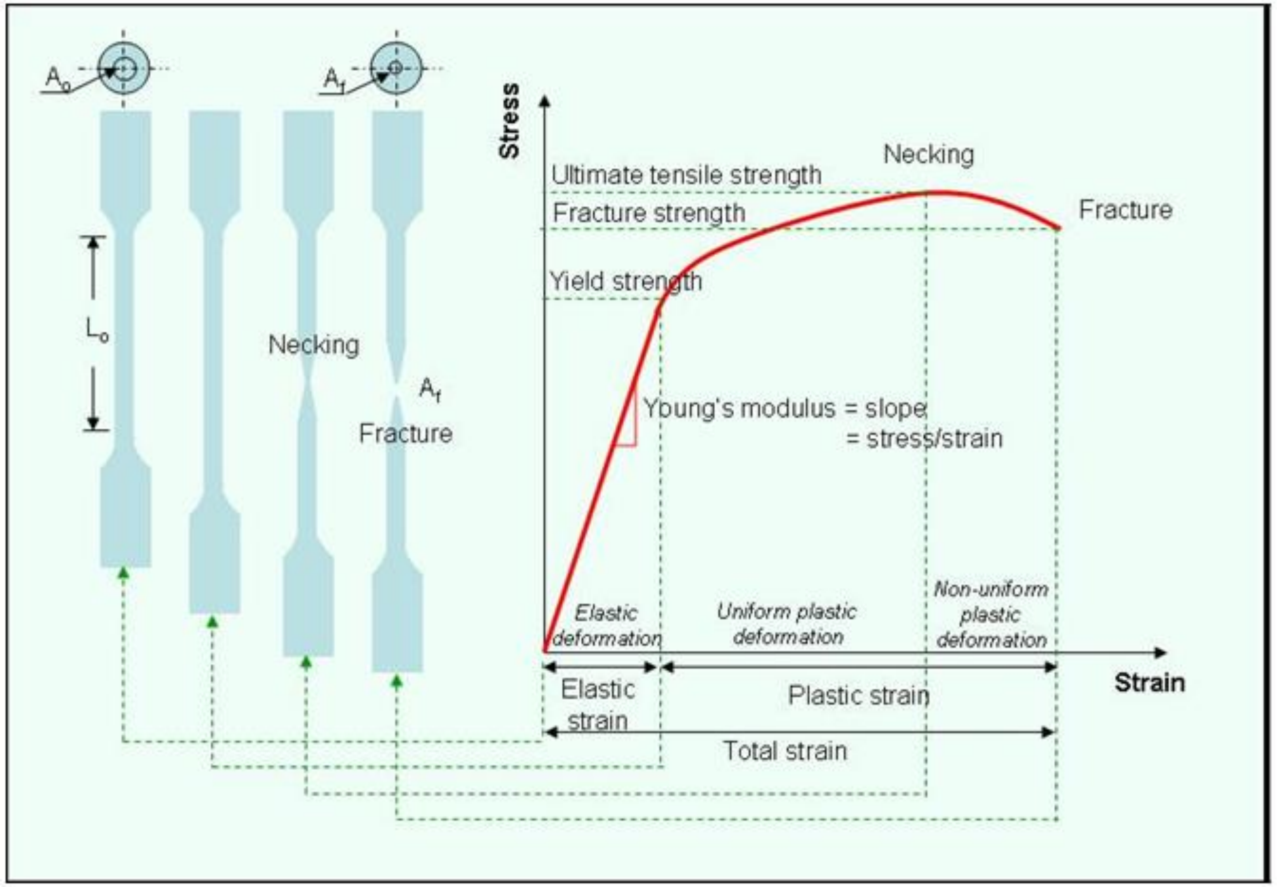


10.1007/s13632-019-00536-1



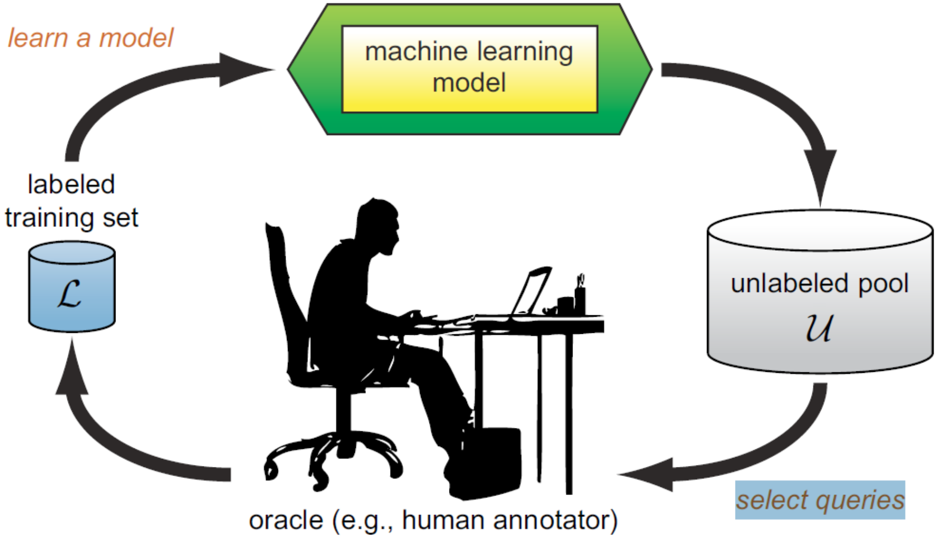
# Afternoon tutorial on steel properties

- ➢ Predict steel strength from composition
  - ✧ Database from https://citrination.com/datasets/153092/

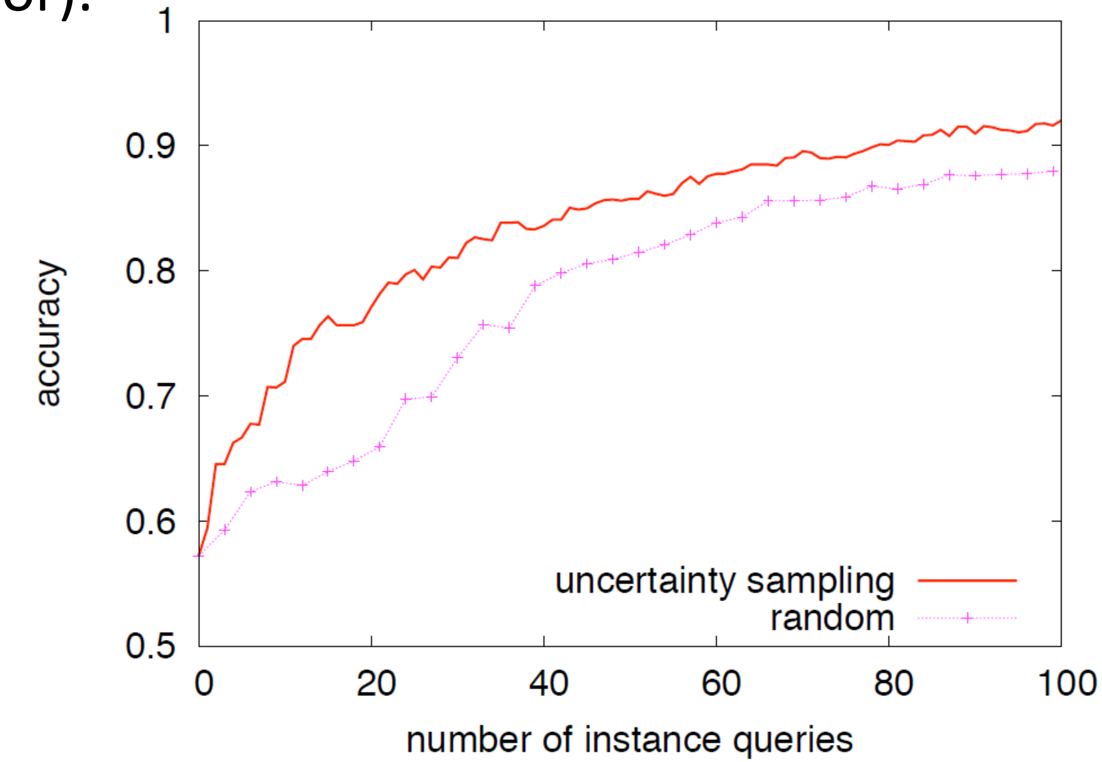


# Relation to active learning

- Machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns.
- Active learner poses queries, usually in the form of unlabeled data instances to be labeled by an oracle (e.g., a human annotator).



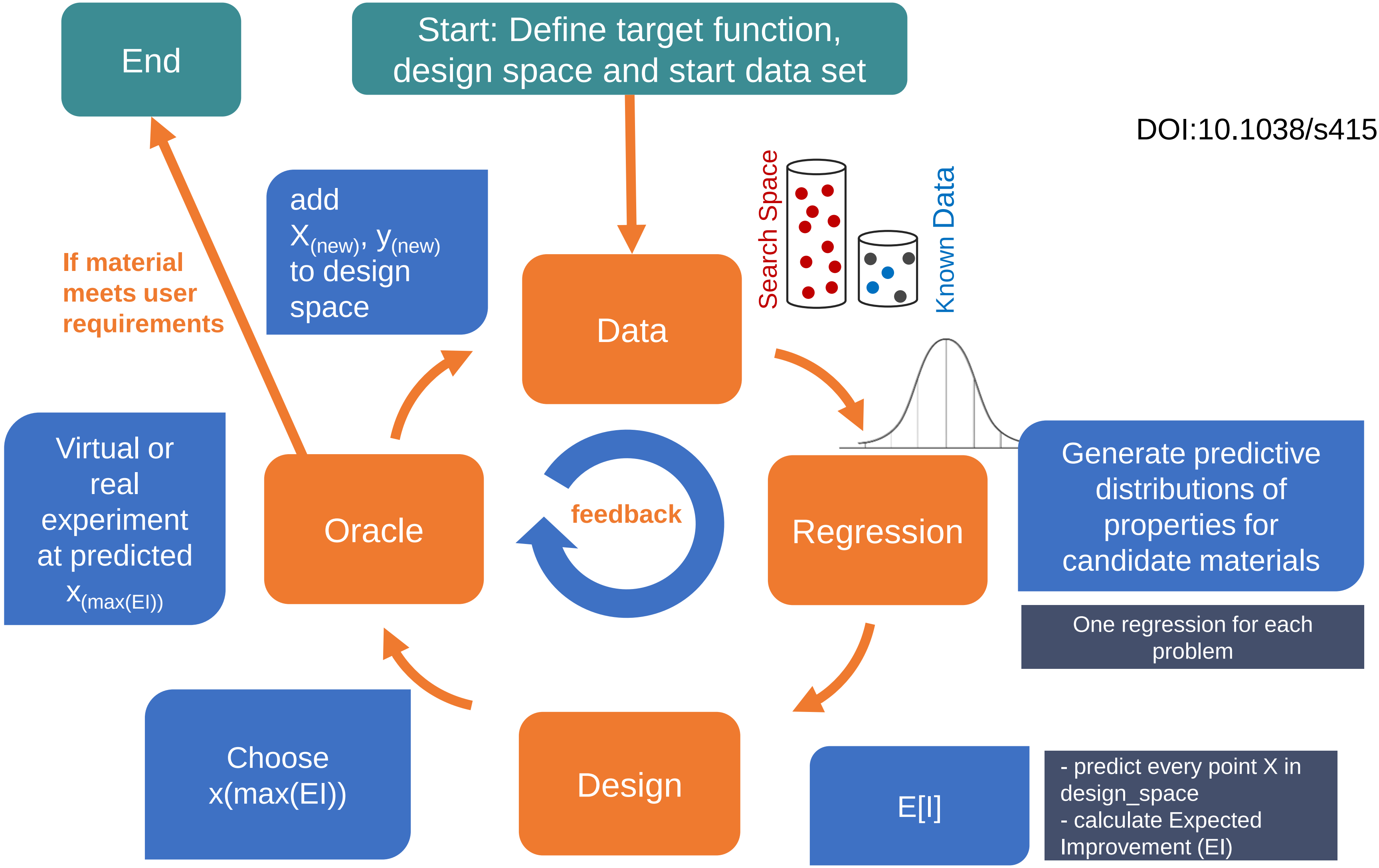
Labeling is expensive!
→ Experiments in materials science



*Settles, Burr (2010). "Active Learning Literature Survey" (PDF). Computer Sciences Technical Report 1648. University of Wisconsin–Madison.*

# Active Learning Loop (ALL) for materials acceleration

DOI:10.1038/s41598-018-21936-3



Start: Define target function, design space and start data set

Data

Regression

Generate predictive distributions of properties for candidate materials

One regression for each problem

Design

E[I]

- predict every point X in design_space
- calculate Expected Improvement (EI)

Choose x(max(EI))

Oracle

Virtual or real experiment at predicted $x_{(max(EI))}$

add $X_{(new)}$, $y_{(new)}$ to design space

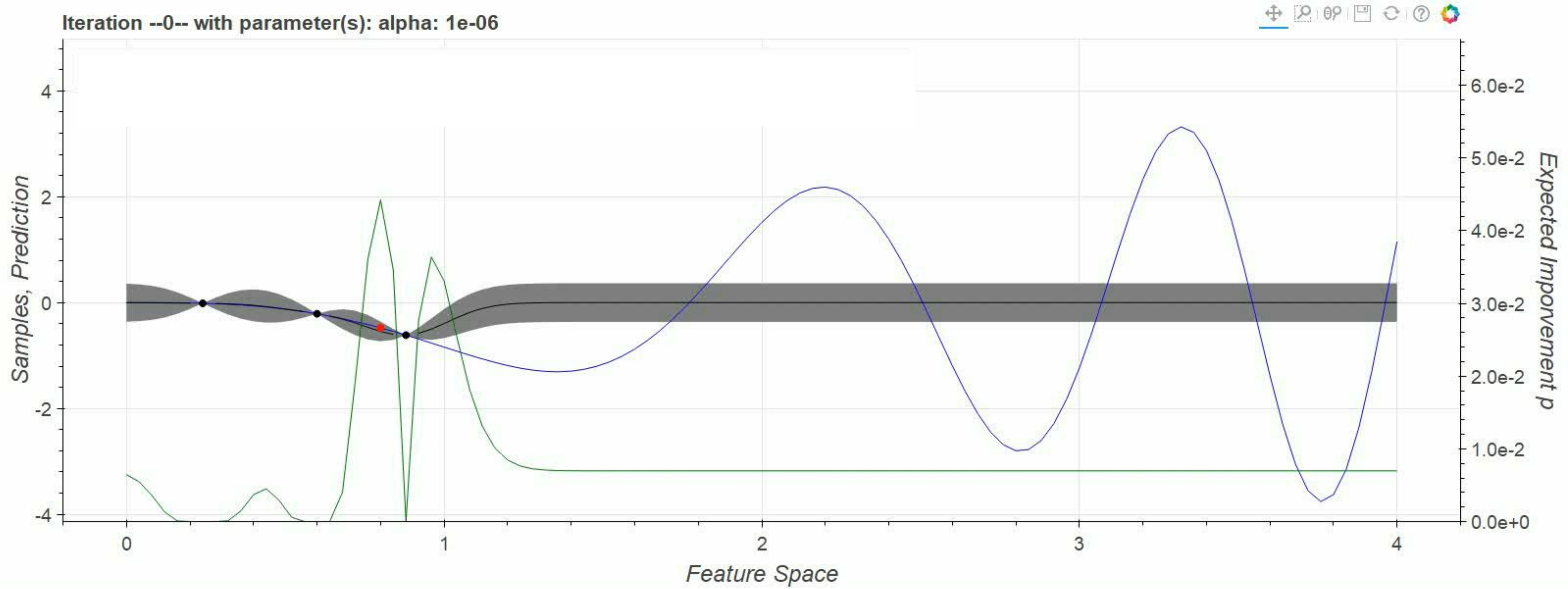
feedback

**If material meets user requirements**

End

# Bayesian optimization

- Bayesian optimization is a sequential design strategy for global optimization.
  - ✧ Goal $y_{min}$ is set upfront and the corresponding descriptor vector $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$ determined.

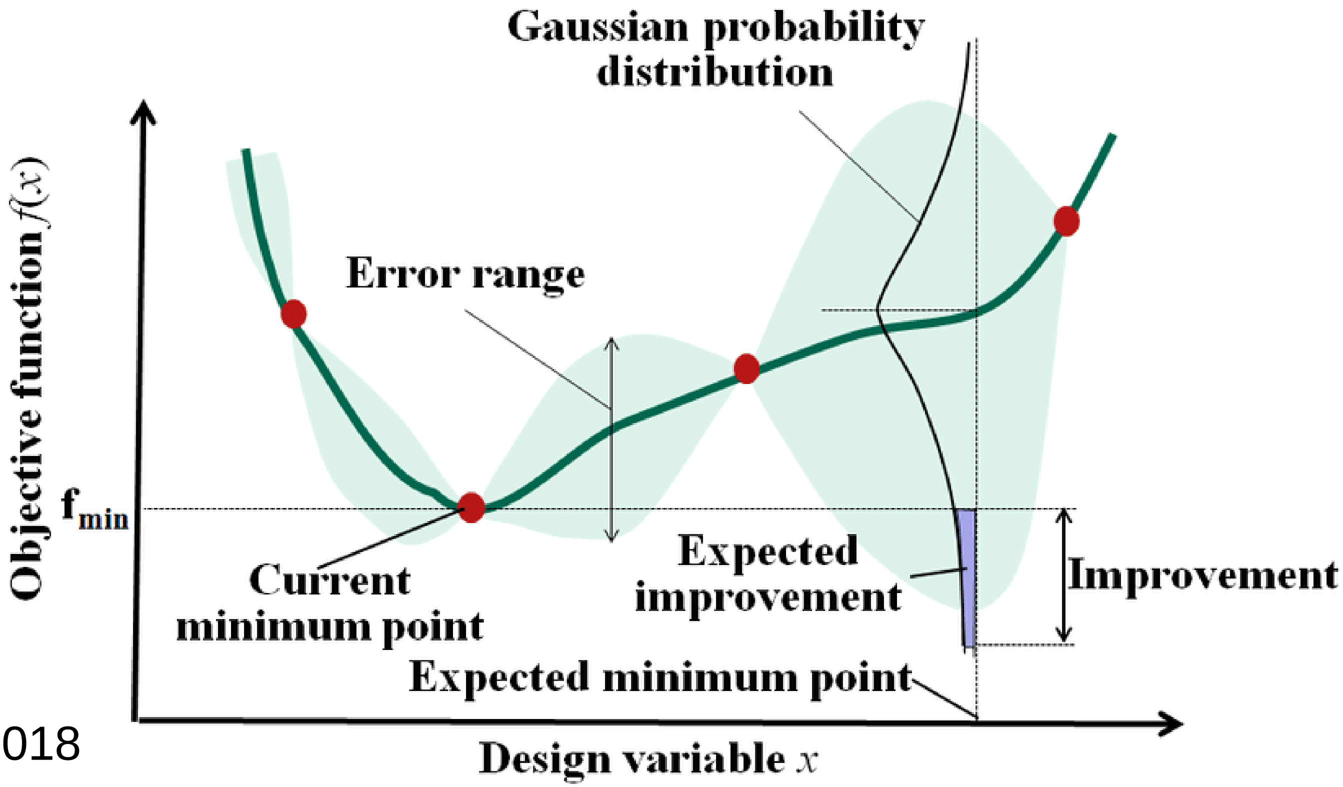


# Acquisition functions

Design

- ➢ Acquisition function proposes sampling points in the design space.
- ➢ Trade off between exploitation and exploration.
  - ✧Exploitation: Sampling where the surrogate model predicts a high objective
  - ✧Exploration means sampling at locations where the prediction uncertainty is high.

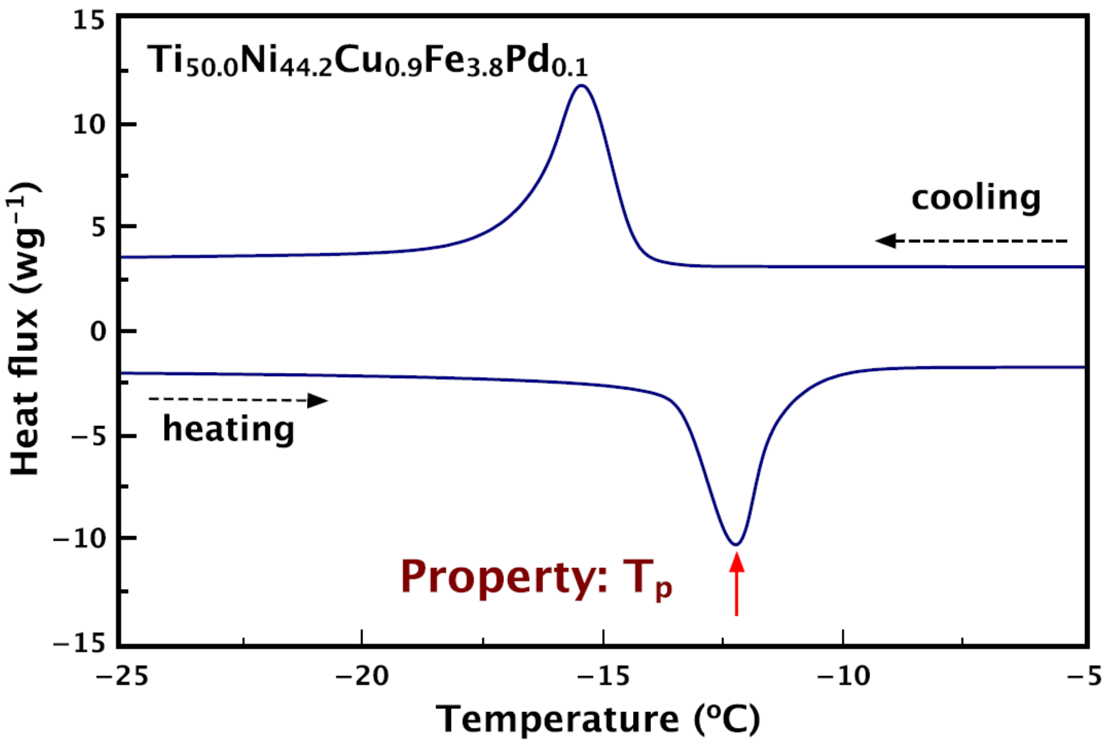
- ➢ **Popular acquisition functions are**
  - ✧Expected improvement (EI)
  - ✧Maximum probability of improvement (MPI),
  - ✧Upper confidence bound (UCB)



10.1299/transjsme.15-00018

# Examples from literature (Lookman et al.)

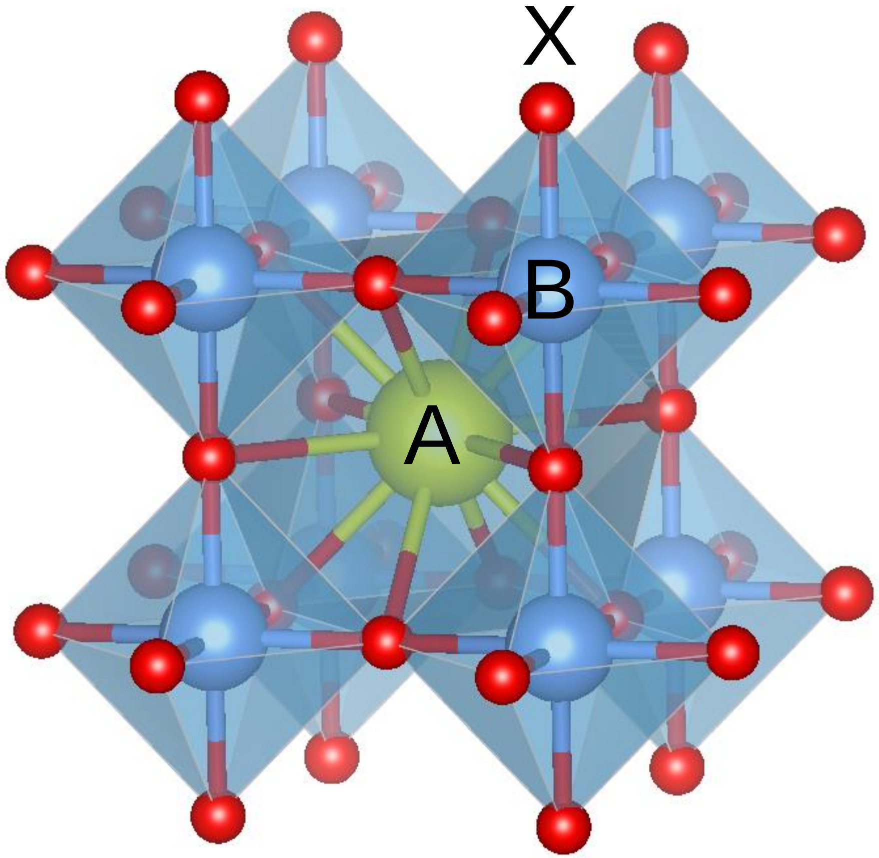
http://dx.doi.org/10.1016/j.actamat.2016.12.009



- Goal: maximize transformation temperature of shape memory alloy.
- Design space: $Ti_{50}Ni_{50\text{-}x\text{-}y\text{-}z}Cu_xFe_yPd_z$
- $Ti_{50}Ni_{25}Pd_{25}$ identified as the alloy with the highest transition temperature of 189 °C in the design space. Two iterations carried out, 50-60 alloys produced.

# The perovskite crystal structure

- Perovskites are materials described by the formula $ABX_3$
  - X is an anion, e.g. oxygen.
  - A and B are cations, A being larger than B.
  - A has a coordination number of 12, B of 6.
- Discovered by Gustav Rose and named after Lev Perovski (10.1002/hlca.2020000).
  - The first analyzed perovskite was $CaTiO_3$.
- Perovskites are one of the most abundant structural families, exhibiting an enormous number of compounds.
- Many oxides with the perovskite structure have physical and chemical properties that make them useful in electronic devices.



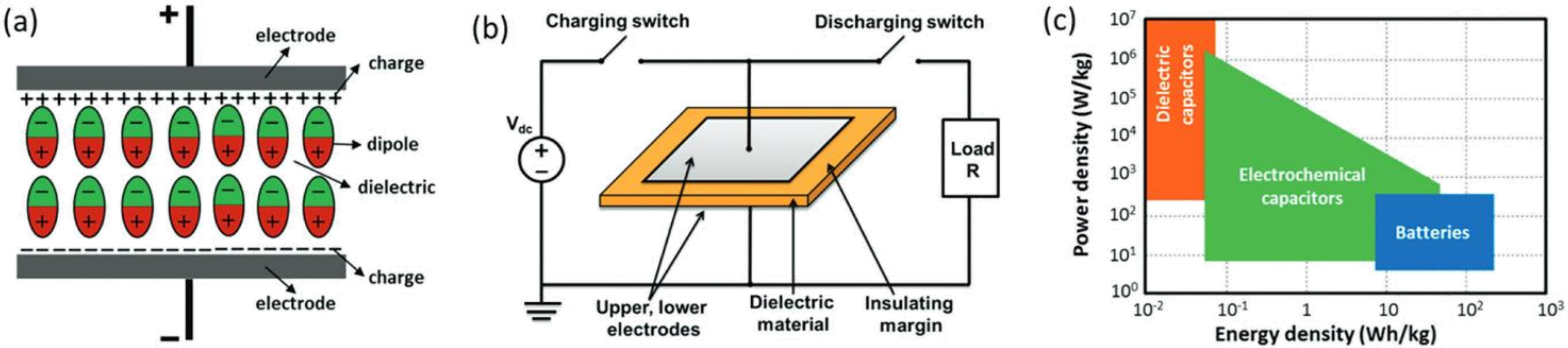
# Use of perovskites in energy storage

- Perovskites are used for energy storage in dielectric capacitors.
- Energy density stored in the capacitor $U$ relates to:

✧ $$U = \frac{\int_0^Q VdQ}{dA} = \int_0^Q EdD = \int_0^Q EdP$$

$V$ ...Voltage
$Q$ ...Charge on capacitor plates
$A$ ...Area of capacitor plates
$d$ ...Distance between capacitor plates
$D$ ...Electric displacemnet

- In high permittivity materials, $D = P + \epsilon_0 E \approx P$ with $P$ the polarization.
- In general, $E$ is a function of $P$, which depends on the intrinsic properties of the perovskite.



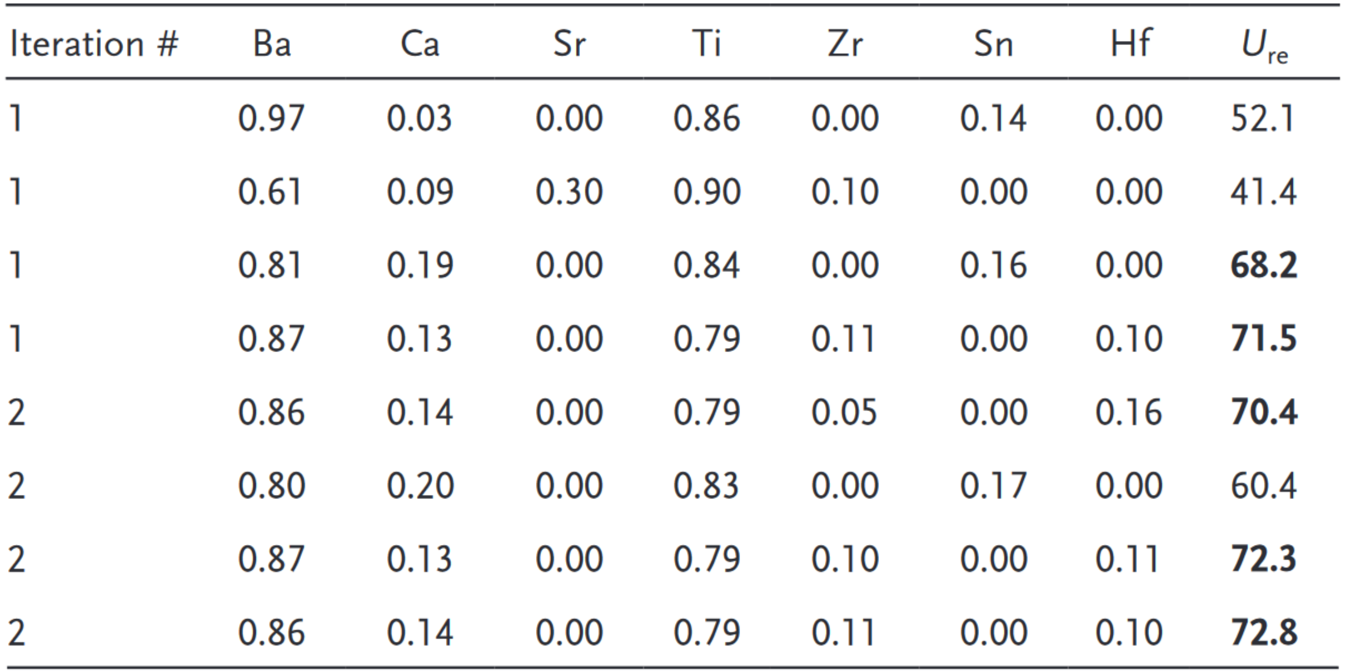
10.1002/adfm.201803665

# The machine learning task

- Objective
  - ✧Regression: Predict energy density from the chemical composition.
  - ✧Classification: Assign electric behavior (ferroelectric, relaxor, crossover) from the chemical composition.
- Algorithm
  - ✧Support Vector Regression with RBF kernel.
  - ✧Random Forest, Gradient boosting, Kernel Ridge Regression.
  - ✧Uncertainty estimation with bootstrapping.
- Data
  - ✧Training data are 182 perovskites.
  - ✧Design space $\left(Ba_{1-x-y}Ca_xSr_y\right)(Ti_{1-u-v-w}Zr_uSn_vHf_w)O_3$
  - ✧9 Million possible perovskites possible.

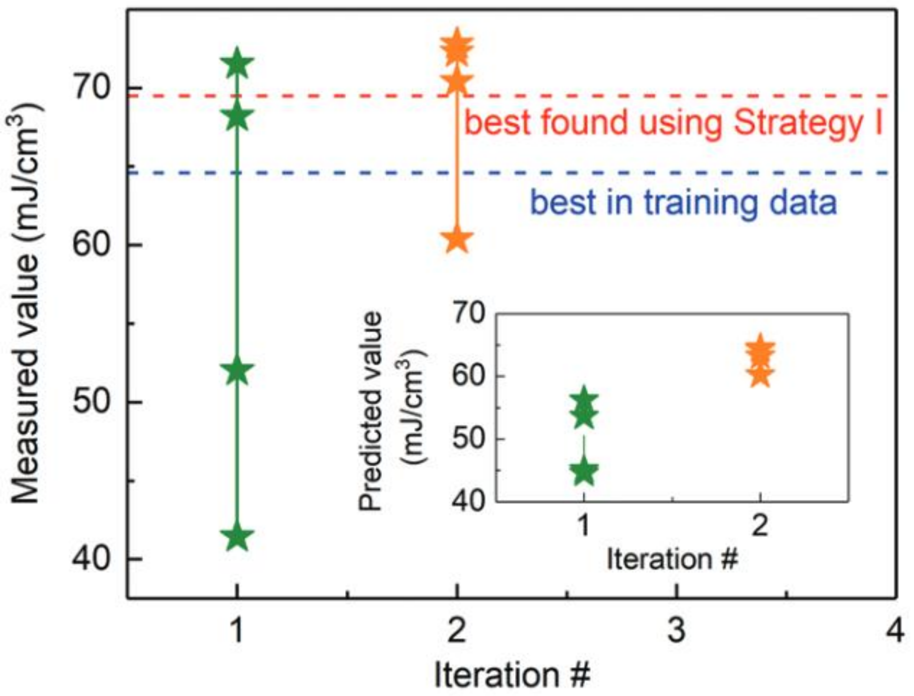
# Active learning of the energy density

- Strategy II improves over approach I.
- 8 new data points identify a new “champion”.
  - ✧ $(Ba_{0.86}Ca_{14})(Ti_{0.79}Zr_{0.11}Hf_{0.10})O_3: U_{re}$ of ≈73 mJ cm$^{-3}$,
  - ✧ 14% improvement compared to the best in the training data.



| Iteration # | Ba | Ca | Sr | Ti | Zr | Sn | Hf | $U_{re}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.97 | 0.03 | 0.00 | 0.86 | 0.00 | 0.14 | 0.00 | 52.1 |
| 1 | 0.61 | 0.09 | 0.30 | 0.90 | 0.10 | 0.00 | 0.00 | 41.4 |
| 1 | 0.81 | 0.19 | 0.00 | 0.84 | 0.00 | 0.16 | 0.00 | **68.2** |
| 1 | 0.87 | 0.13 | 0.00 | 0.79 | 0.11 | 0.00 | 0.10 | **71.5** |
| 2 | 0.86 | 0.14 | 0.00 | 0.79 | 0.05 | 0.00 | 0.16 | **70.4** |
| 2 | 0.80 | 0.20 | 0.00 | 0.83 | 0.00 | 0.17 | 0.00 | 60.4 |
| 2 | 0.87 | 0.13 | 0.00 | 0.79 | 0.10 | 0.00 | 0.11 | **72.3** |
| 2 | 0.86 | 0.14 | 0.00 | 0.79 | 0.11 | 0.00 | 0.10 | **72.8** |

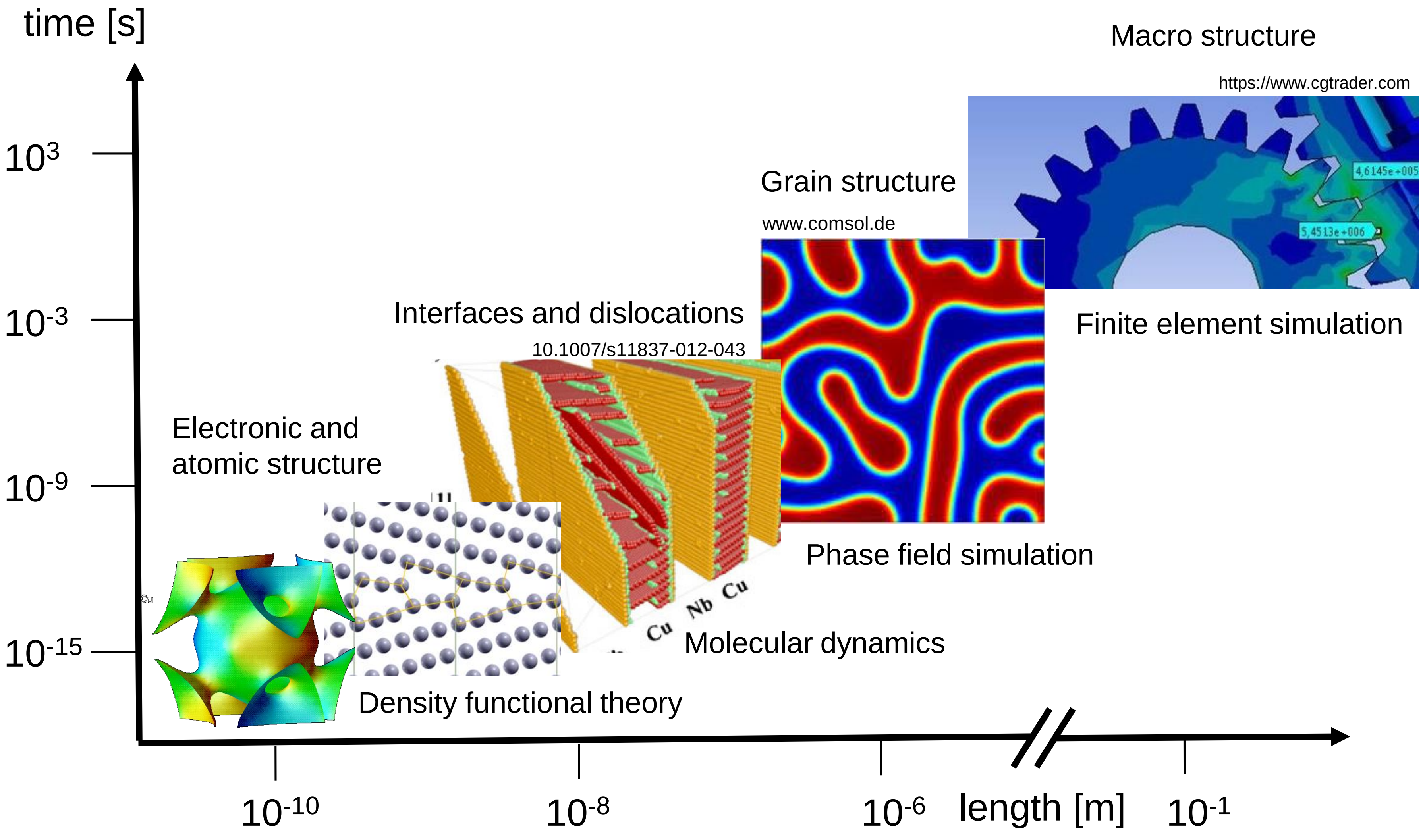
10.1002/advs.201901395



# Overview

- ➢ Automatically process or interpret analytical experiments.
  - ✧Big data from optical  or electron microscopy
  - ✧Atom probe tomography
- ➢ Directly learn process-structure-property relationships of materials
  - ✧Typically data is not “big”, datapoints are very expensive.
  - ✧Predict materials properties.
  - ✧Optimize new materials.
- ➢ Replace expensive physics based simulations (from 3rd paradigm):
  - ✧Optimization of codes has reached a level that is difficult to improve further.
  - ✧Use machine learning to construct surrogate models
  - ✧Speed up calculations for optimization tasks

# Multiscale modeling methods



Macro structure

https://www.cgtrader.com

Finite element simulation

Grain structure

www.comsol.de

Phase field simulation

Interfaces and dislocations

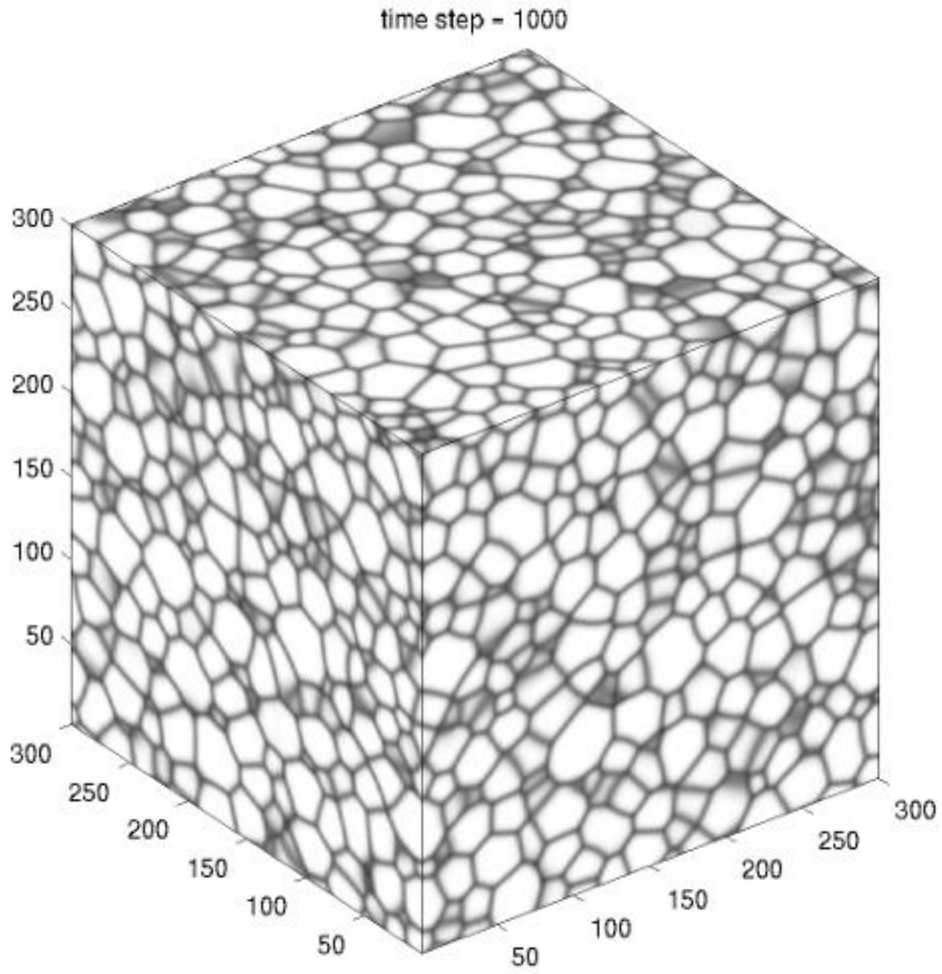
10.1007/s11837-012-043

Molecular dynamics

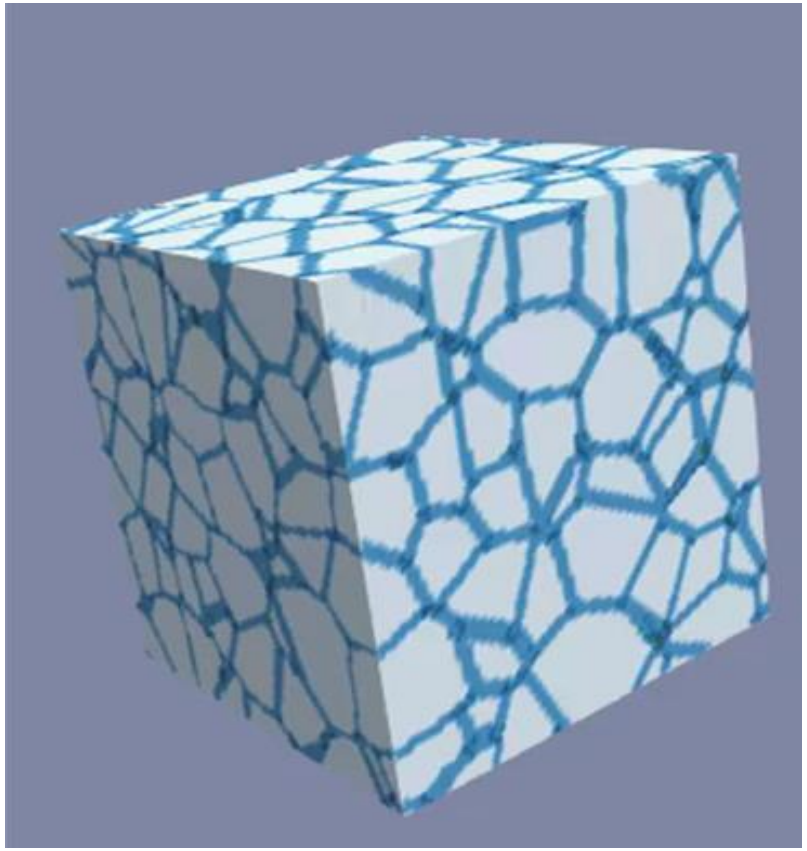
Electronic and atomic structure

Density functional theory

# Grain growth simulations

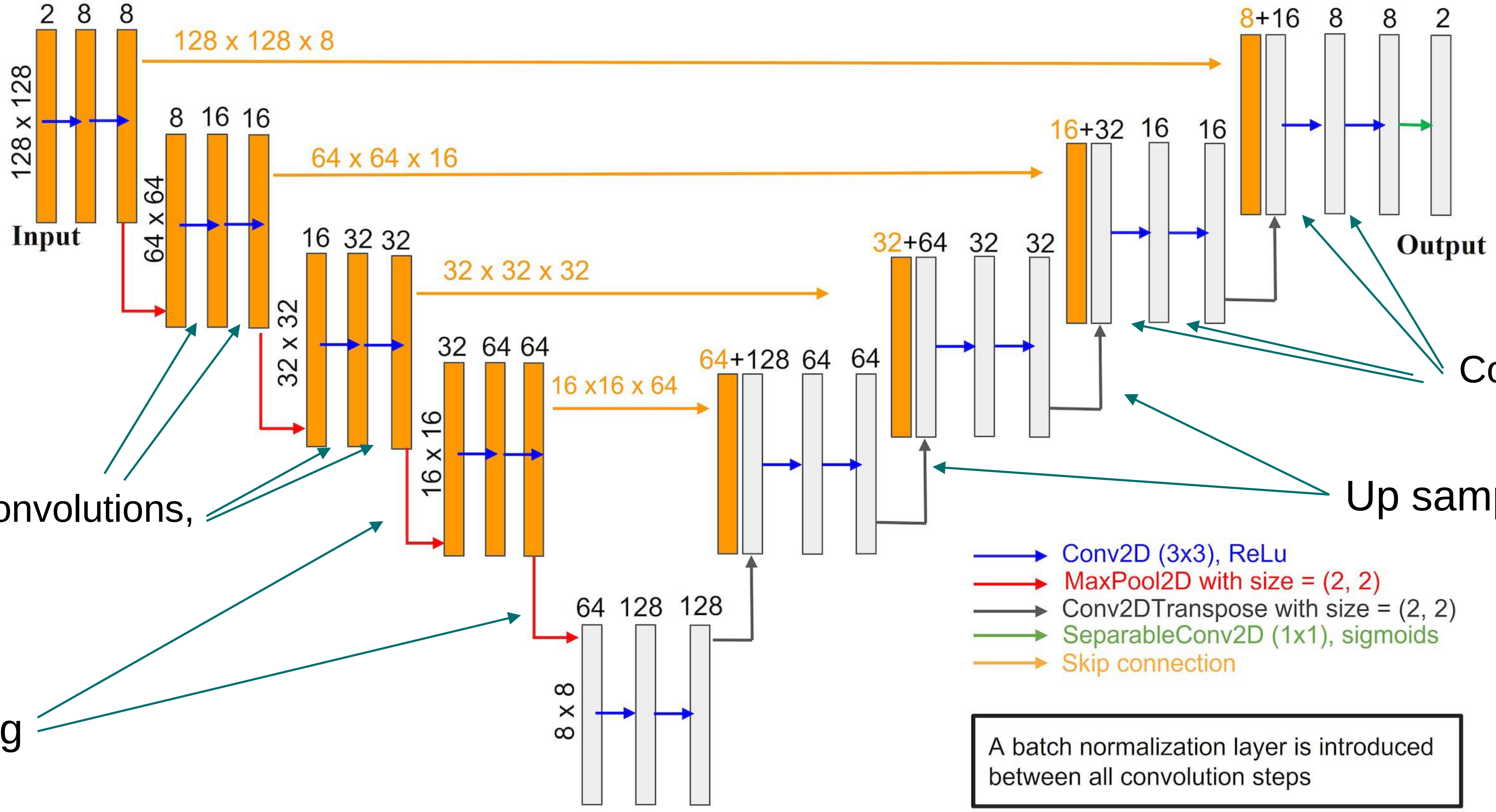


https://en.wikipedia.org/wiki/Grain_growth



# U-Net convolutional neural network

Image pixels are 128x128 for two phases



Convolutions,

Pooling

Convolutions

Up sampling

Conv2D (3x3), ReLu

MaxPool2D with size = (2, 2)

Conv2DTranspose with size = (2, 2)

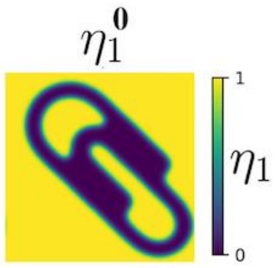
SeparableConv2D (1x1), sigmoids

Skip connection

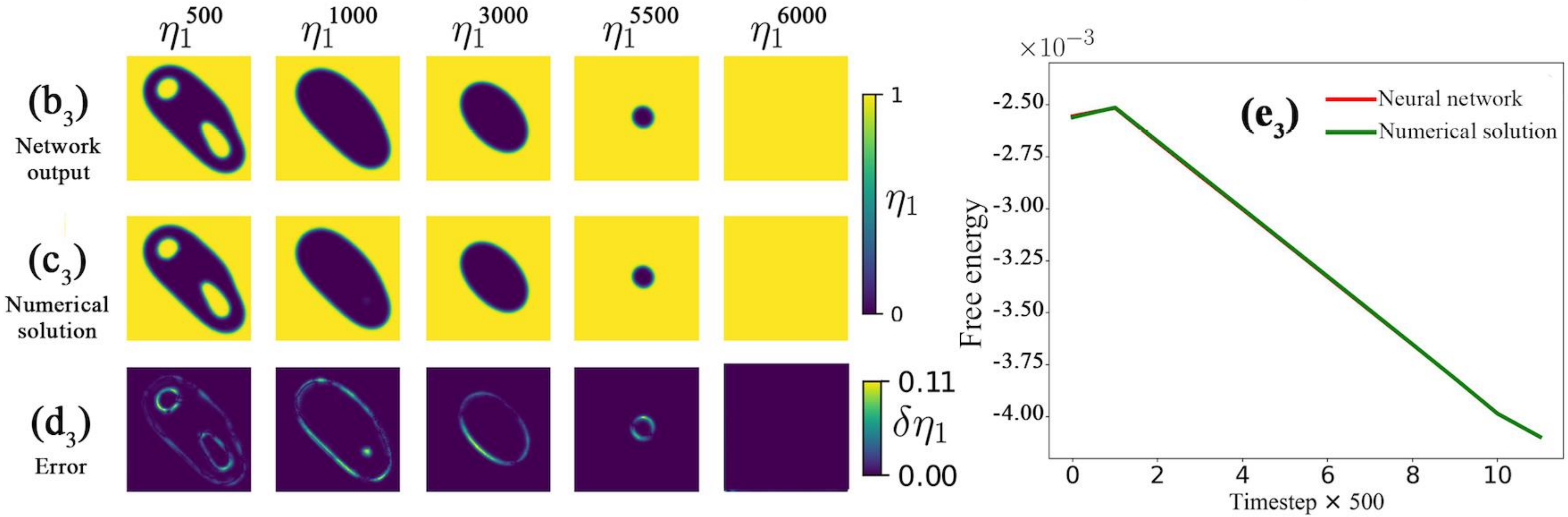
A batch normalization layer is introduced between all convolution steps

arXiv:2205.02121

# Prediction of time series

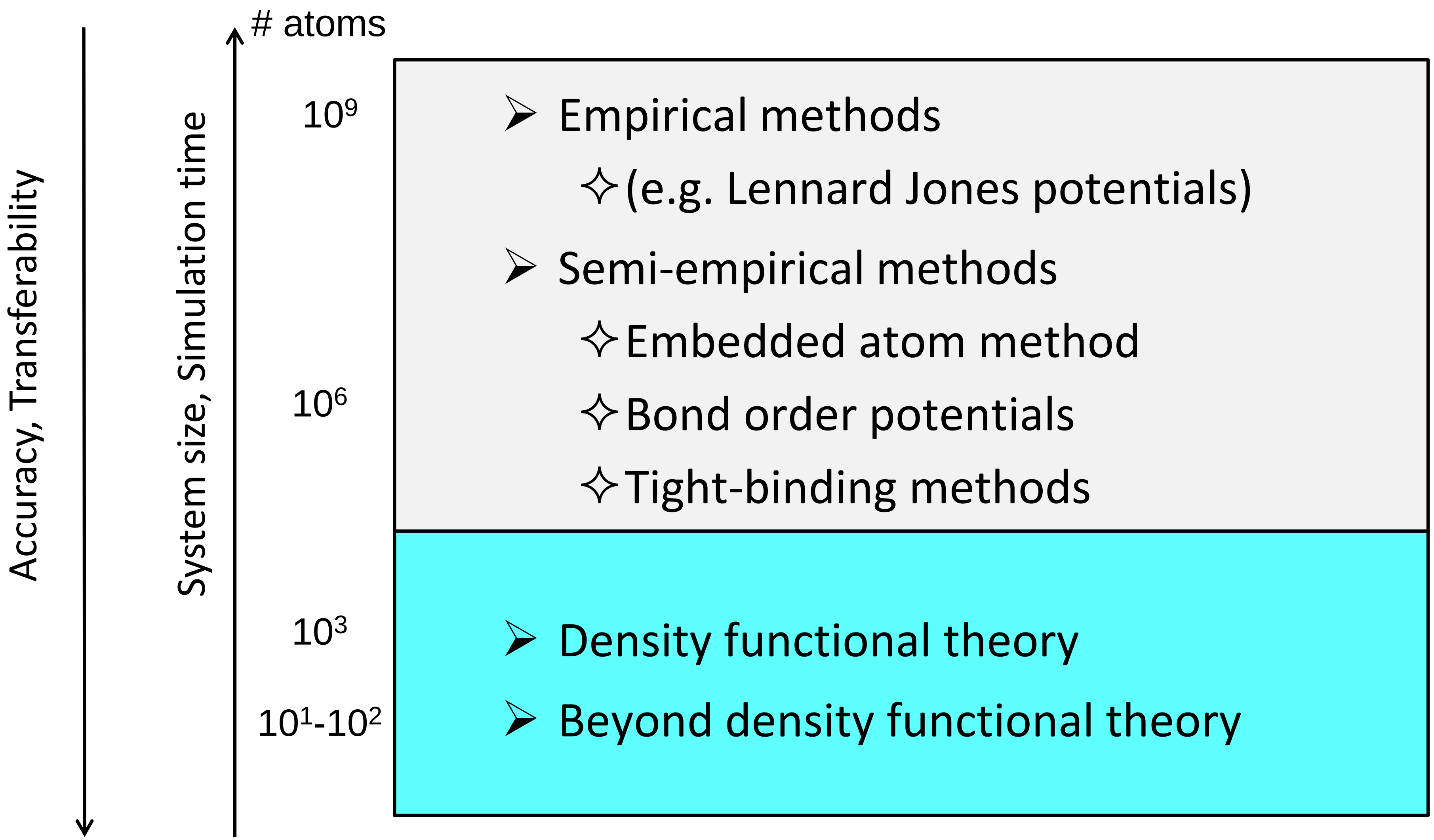


- ➢ Repeatedly feeding output  to input gives the time series.
- ➢ CNN is 90 times faster than the FFT based solver.



arXiv:2205.02121

# Hierarchy of physics-based atomistic methods



Accuracy, Transferability

System size, Simulation time

# atoms

$10^9$

$10^6$

$10^3$

$10^1$-$10^2$

- Empirical methods
  - ✧(e.g. Lennard Jones potentials)
- Semi-empirical methods
  - ✧Embedded atom method
  - ✧Bond order potentials
  - ✧Tight-binding methods

- Density functional theory
- Beyond density functional theory

Empirical methods:

Parameters are material-specific.

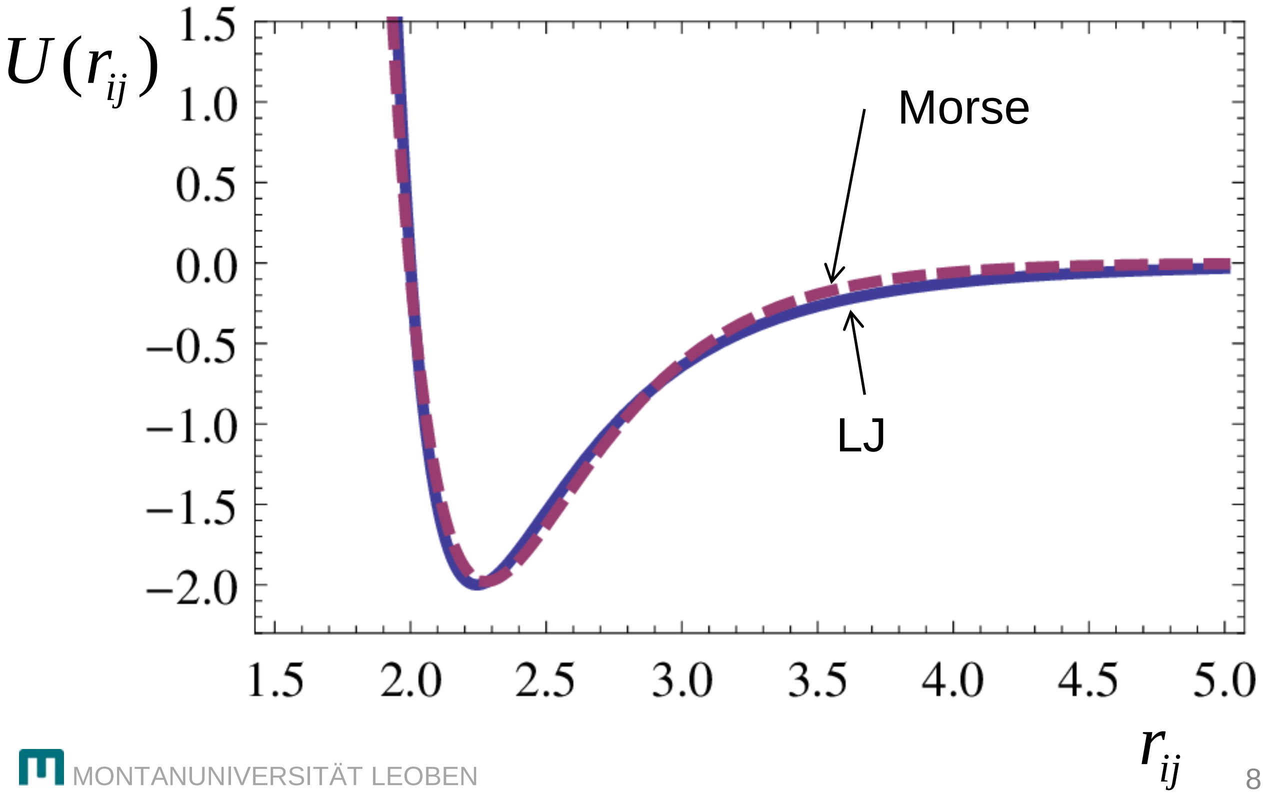
Ab-initio methods:

Parameters are universal.

# Empirical methods

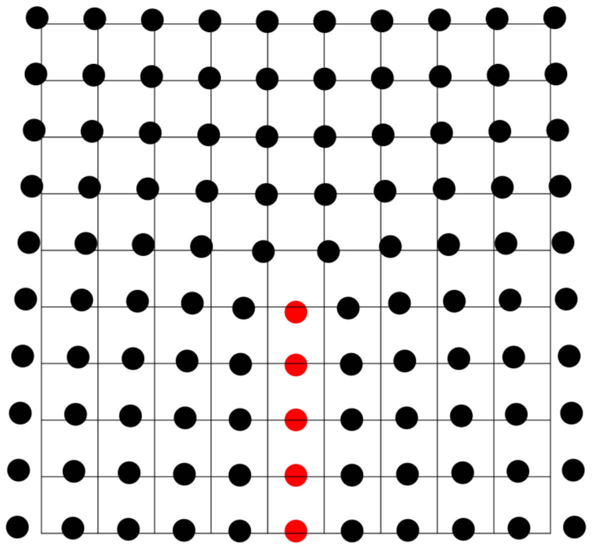
- Lennard-Jones potential $$U(r_{ij}) = 4a\left[\left(\frac{b}{r_{ij}}\right)^{12} - \left(\frac{b}{r_{ij}}\right)^{6}\right]$$
- Morse potential: $$U(r_{ij}) = a\left(e^{-2b(r_{ij}-r_0)} - 2e^{-b(r_{ij}-r_0)}\right)$$

Methods are usefull for purely van der Waals bonded systems.



Sum over all pairs

$$U_{tot} = \sum_{i,j>i} U(r_{ij})$$



# Density functional theory (DFT)

- „Father“ of DFT:
  - Walter Kohn
- Born in Vienna 1923.
- Cornerstones laid in 1964-65.
- Nobel Prize 1998.
- Theory still under active developement.
- It provides a parameter-free, quantum mechanical description of interatomic bonding.
- Allows treating up to 1000 atoms.
- Typically requires supercomputers when treating defects

Walter Kohn
1923-2016

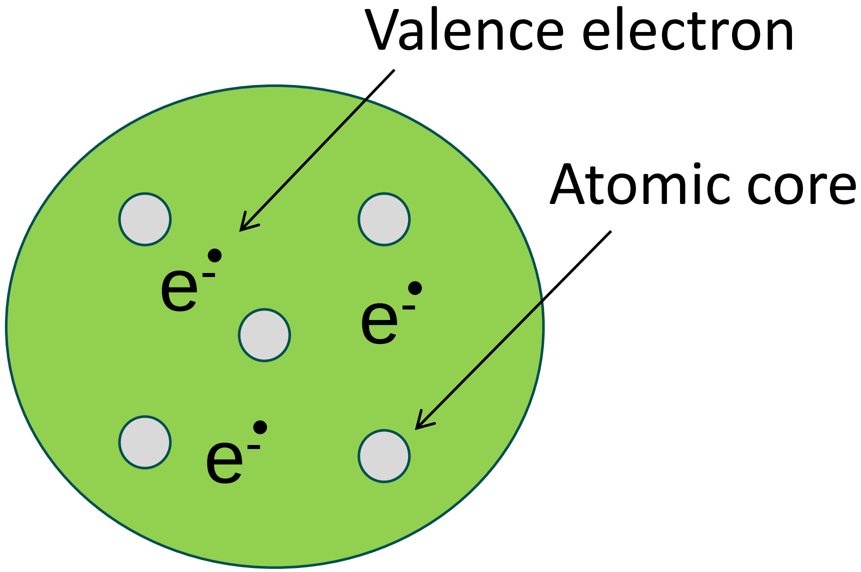
# DFT

- Heart of DFT: Kohn-Sham equation.

$$\left(-\frac{1}{2}\nabla^2 + V_{eff}[n]\right)\psi_i = \varepsilon_i \psi_i \qquad V_{eff}[n] = V_{ext} + V_H[n] + V_{xc}[n]$$

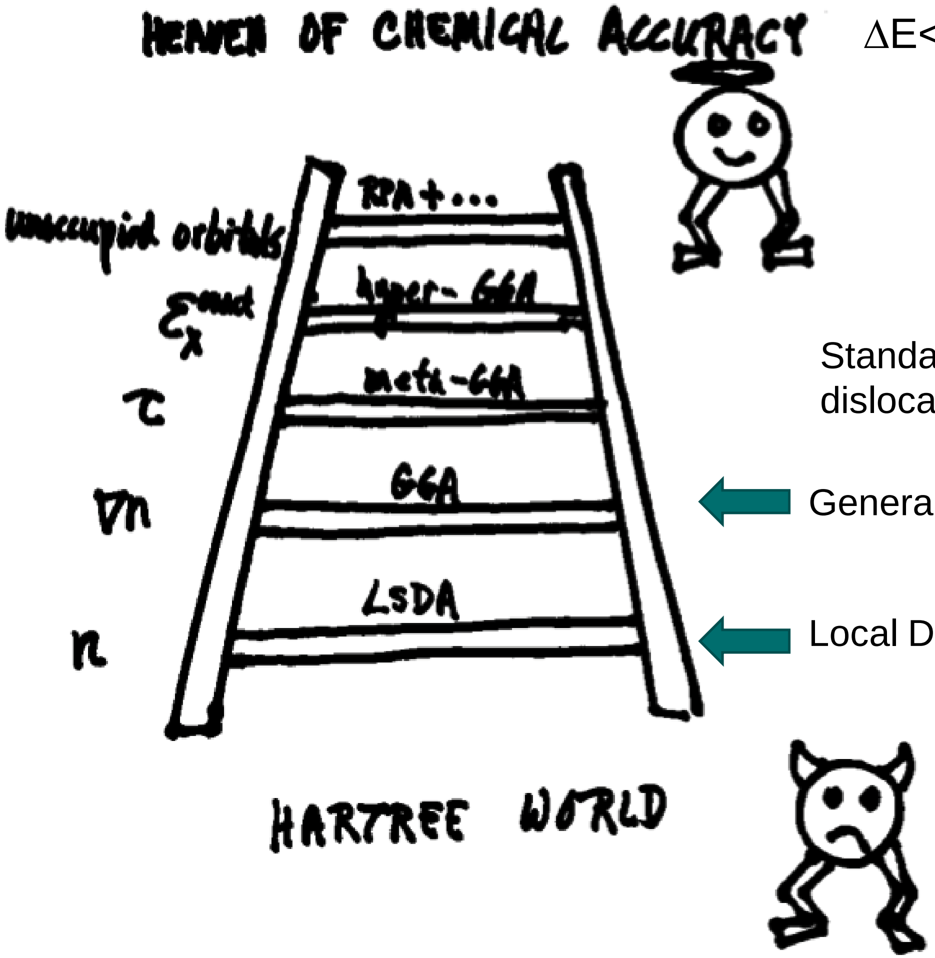
- Electrons move independently in an effective potential.
- All approximations of DFT are in the exchange-correlation (XC) potential.
- Self-consistent solution procedure.

due to dependence of $V_{eff}[n]$ on the charge density.



# Hierarchy of XC functionals

- Jacob ladder by J. Perdew



ΔE< 0.043 eV

Standard XC functionals for dislocation simulations:

Generalized Gradient Approximation

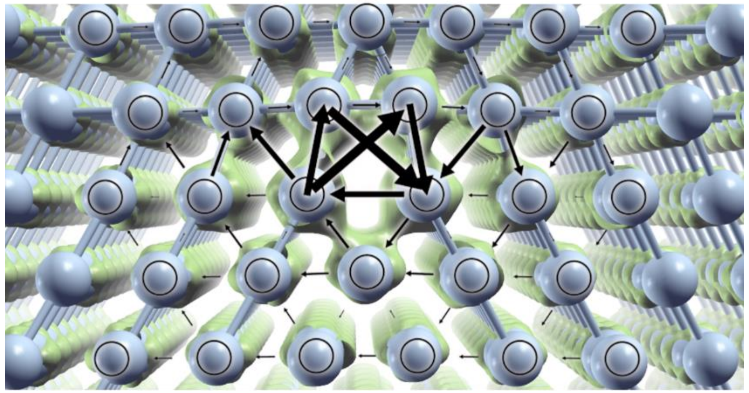
Local Density Approximation

J. Perdew

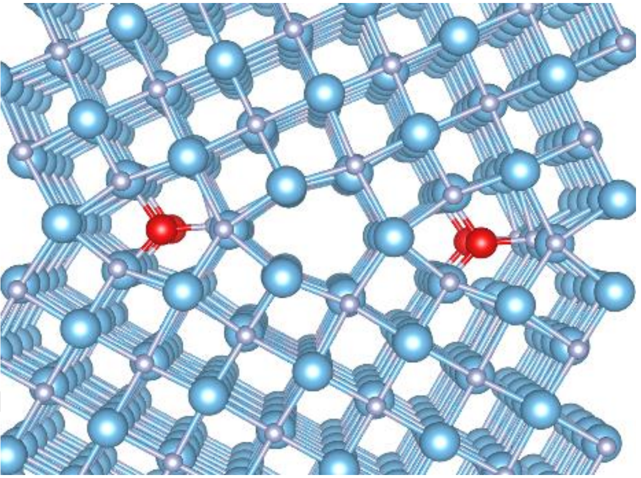
Perdew et al. J. Chem. Theory Comput. 5, 902 (2009).

# What DFT provides

- Electron density for any collection of atoms
- Energy and forces
- Ground state geometry of the the atoms
- Applications in condensed matter
  - Estimates for band gaps, optical properties
  - Stability of crystallographic phases → Thermodynamics
  - Crystallographic defects
    - Vacancies → Diffusion
    - Dislocations → mechanical properties
    - Interfaces → Grain growth, embrittlement phenomena
- Application in chemistry
  - Chemical reactions, optical properties, …



Dislocation cores



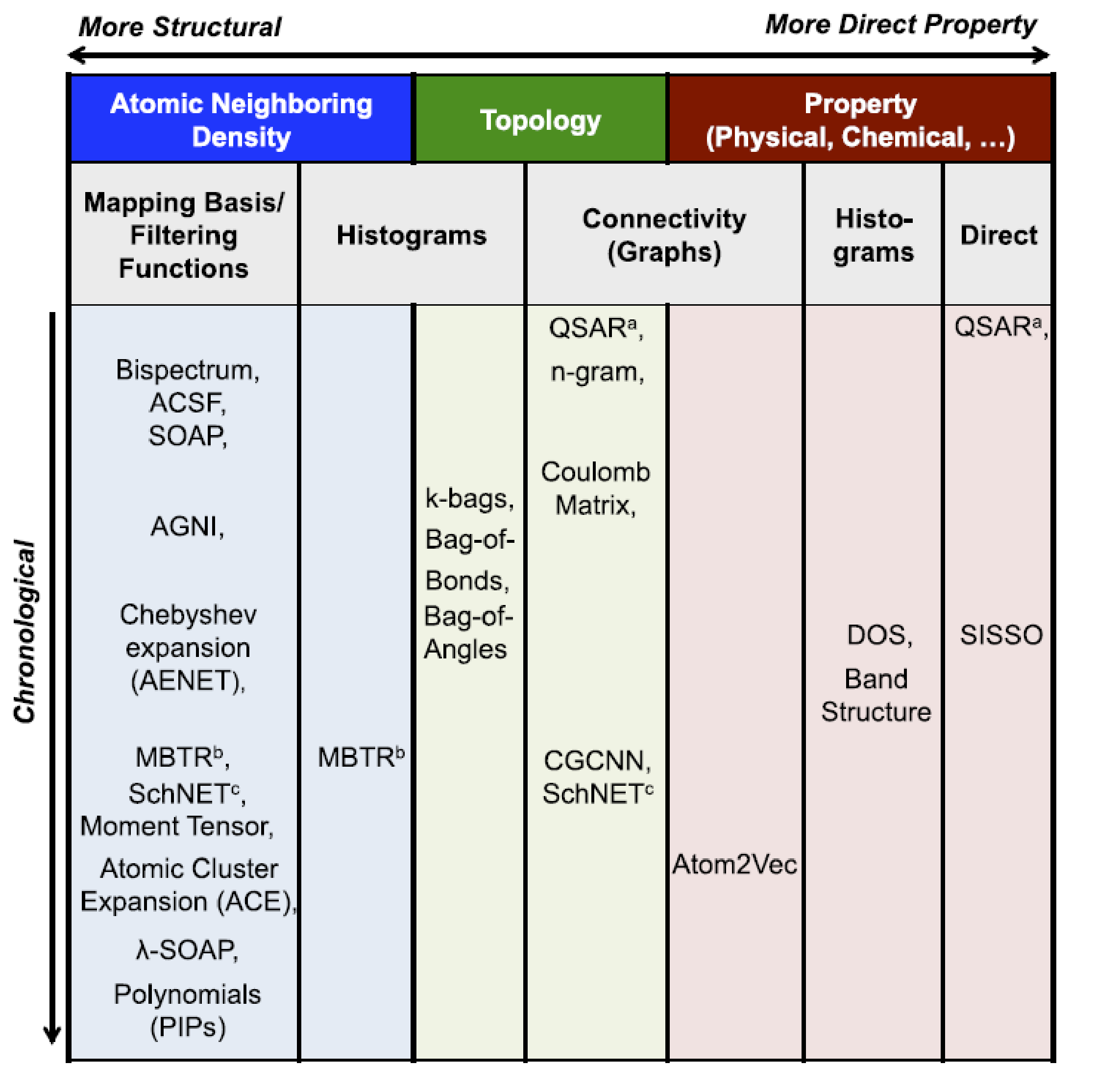
# Pros and cons of DFT

- Pros:
  - Precise and highly transferable method.
  - Can cope with different crystal structures and bonding situations at defects.
  - Particularly convenient for treatment of changes in chemistry -> Alloying.
- Cons:
  - Expensive, reduced system sizes.
  - Different xc-potentials can give different results.
  - Calculations are technically demanding: Carefull convergence with respect to k-points, cutoff energies necessary. Good choice of pseudo-potentials necessary.

VSC, Austria

# Machine learning DFT

- The task consists in predicting properties $y$ of atomistic entities, i.e. molecules or condensed matter from atomic coordinates $\boldsymbol{r}_k$, atomic species $s_k$, and physical properties $\theta_i$.
- $y = f(\{\boldsymbol{r}_k, s_k, \theta_i\})$,
- Many methods have been developed for this purpose.
  - ✧Overview
  - ✧Activities can be distinguished
    - Prediction of properties
    - Replacement of DFT energies and forces → Development of machine-learned interatomic potentials.



More Structural ↔ More Direct Property

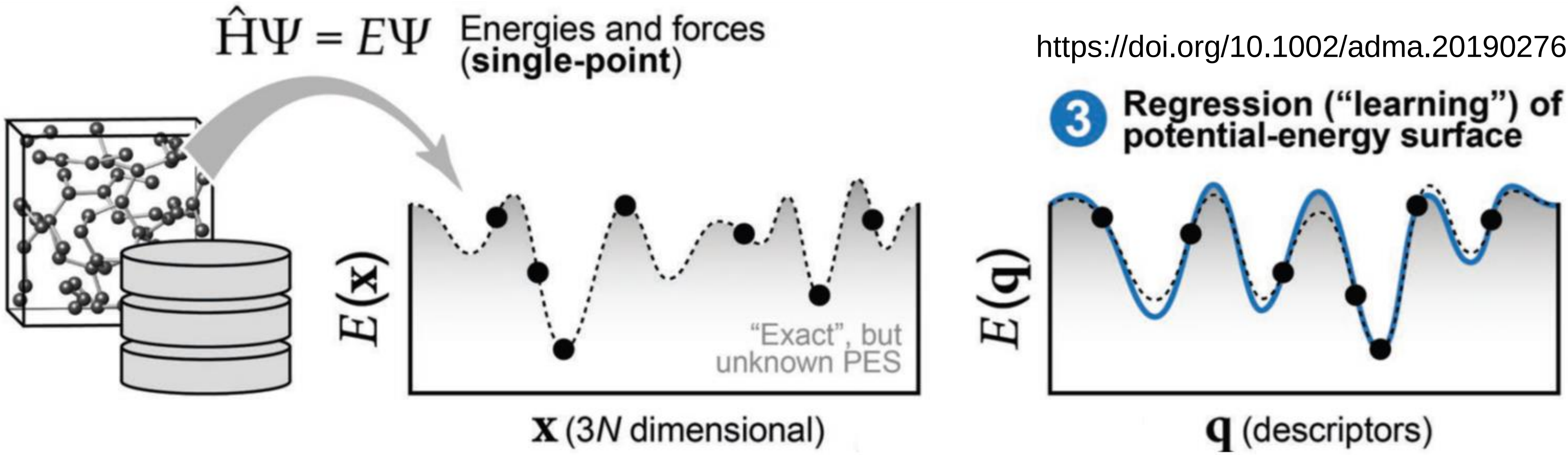
Chronological

| Atomic Neighboring Density | | Topology | | Property (Physical, Chemical, …) | | |
|---|---|---|---|---|---|---|
| Mapping Basis/ Filtering Functions | Histograms | | Connectivity (Graphs) | | Histo-grams | Direct |
| Bispectrum,<br>ACSF,<br>SOAP,<br>AGNI,<br>Chebyshev expansion (AENET),<br>MBTR[b],<br>SchNET[c],<br>Moment Tensor,<br>Atomic Cluster Expansion (ACE),<br>λ-SOAP,<br>Polynomials (PIPs) | MBTR[b] | k-bags,<br>Bag-of-Bonds,<br>Bag-of-Angles | QSAR[a],<br>n-gram,<br>Coulomb Matrix,<br>CGCNN,<br>SchNET[c] | Atom2Vec | DOS,<br>Band Structure | QSAR[a],<br>SISSO |

https://doi.org/10.1063/5.0016005

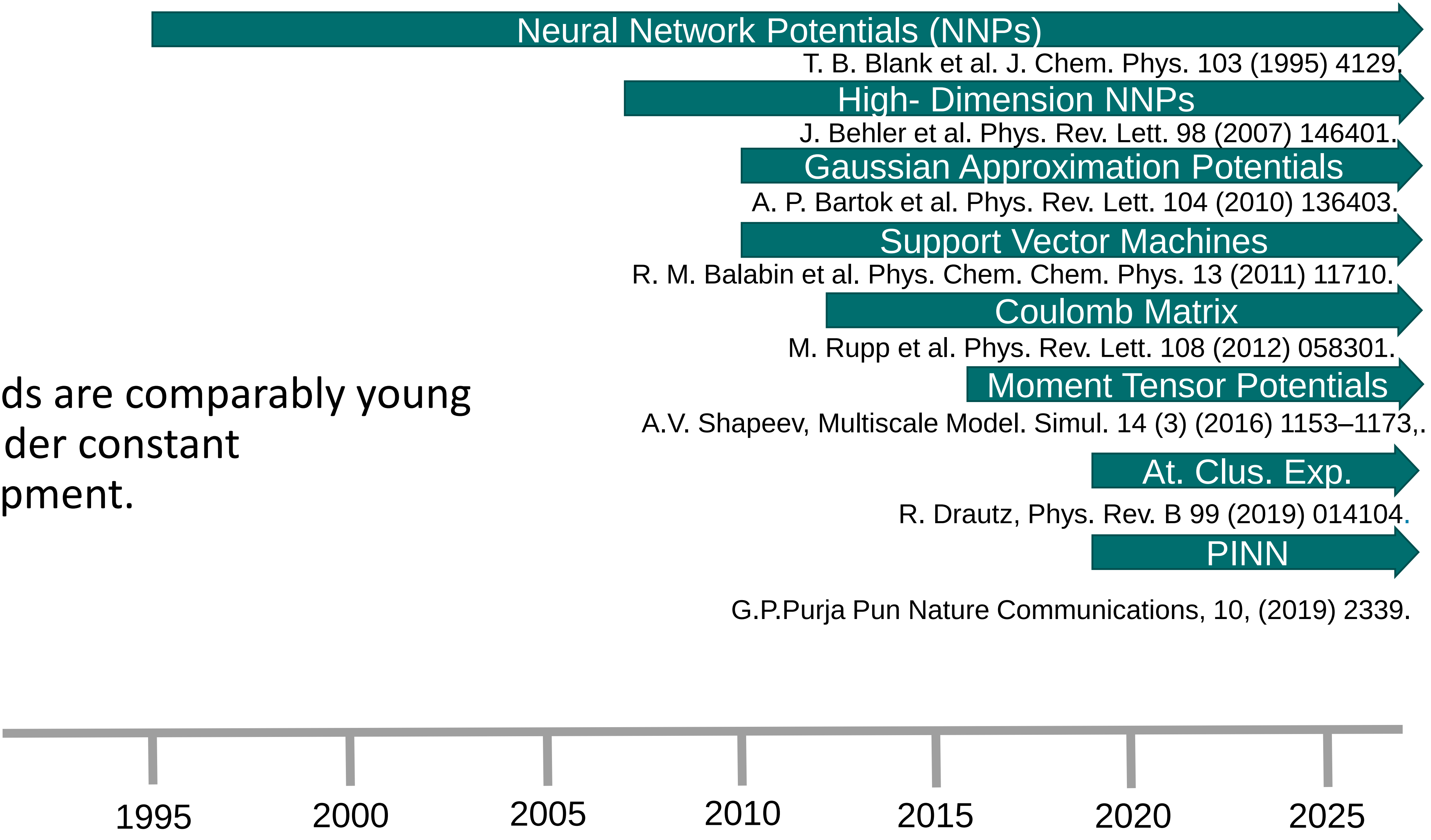
# Machine learning interatomic potentials

- General overview of how ML-based interatomic potentials are constructed:
  - ✧Assembling a database of representative structural models,
  - ✧Computing energies and forces with DFT,
  - ✧Expressing the atomic structure in “machine-readable” form using descriptors, and
  - ✧Regressing (“learning”) the potential-energy surface.


$\hat{H}\Psi = E\Psi$
Energies and forces (single-point)
$E(\mathbf{x})$
“Exact”, but unknown PES
$\mathbf{x}$ (3$N$ dimensional)
3
Regression (“learning”) of potential-energy surface
$E(\mathbf{q})$
$\mathbf{q}$ (descriptors)

https://doi.org/10.1002/adma.201902765.

# Historical evolution of ML interatomic potentials



Neural Network Potentials (NNPs)

T. B. Blank et al. J. Chem. Phys. 103 (1995) 4129.

High- Dimension NNPs

J. Behler et al. Phys. Rev. Lett. 98 (2007) 146401.

Gaussian Approximation Potentials

A. P. Bartok et al. Phys. Rev. Lett. 104 (2010) 136403.

Support Vector Machines

R. M. Balabin et al. Phys. Chem. Chem. Phys. 13 (2011) 11710.

Coulomb Matrix

M. Rupp et al. Phys. Rev. Lett. 108 (2012) 058301.

Moment Tensor Potentials

A.V. Shapeev, Multiscale Model. Simul. 14 (3) (2016) 1153–1173,.

At. Clus. Exp.

R. Drautz, Phys. Rev. B 99 (2019) 014104.

PINN

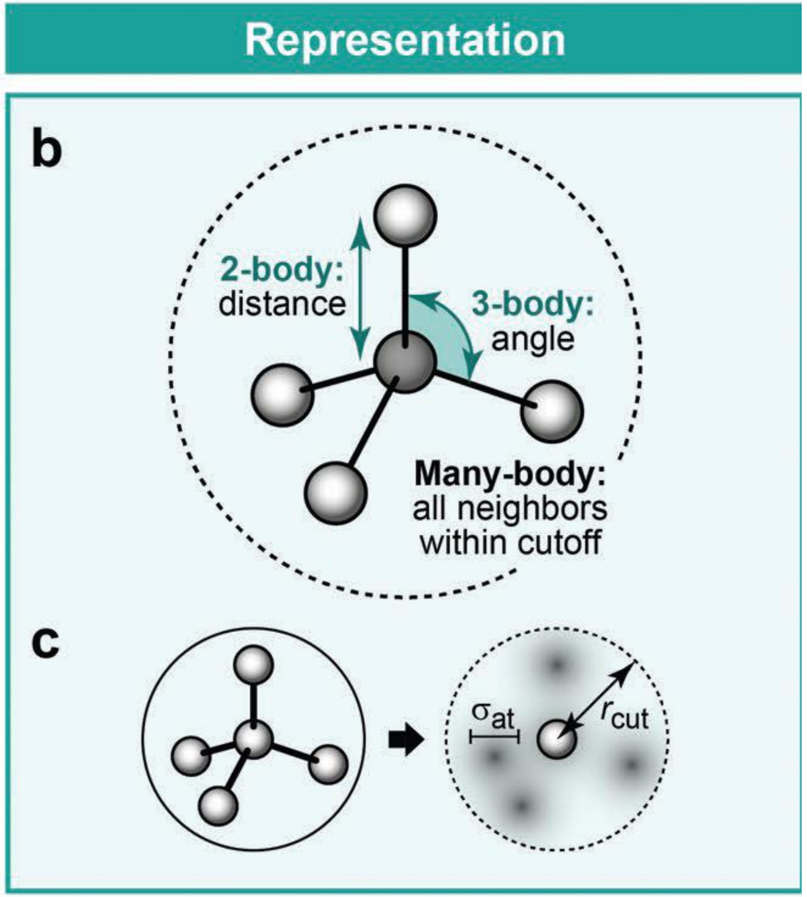
G.P.Purja Pun Nature Communications, 10, (2019) 2339.

- ➢ Methods are comparably young and under constant development.

# Descriptors

- Assumptions
  - ✧Atomic interactions are short-range.
  - ✧Energy assigned to atom *i* only depends on its local environment.
  - ✧The total energy is predicted based on the information about all local environments in the system.
- Requirements:
  - ✧Resolution must be high, different environments must be represented by sufficiently different descriptors.
  - ✧Computational cost must be low.
- Descriptors range from two- or three-body terms all the way to complex “many-body” formalisms.

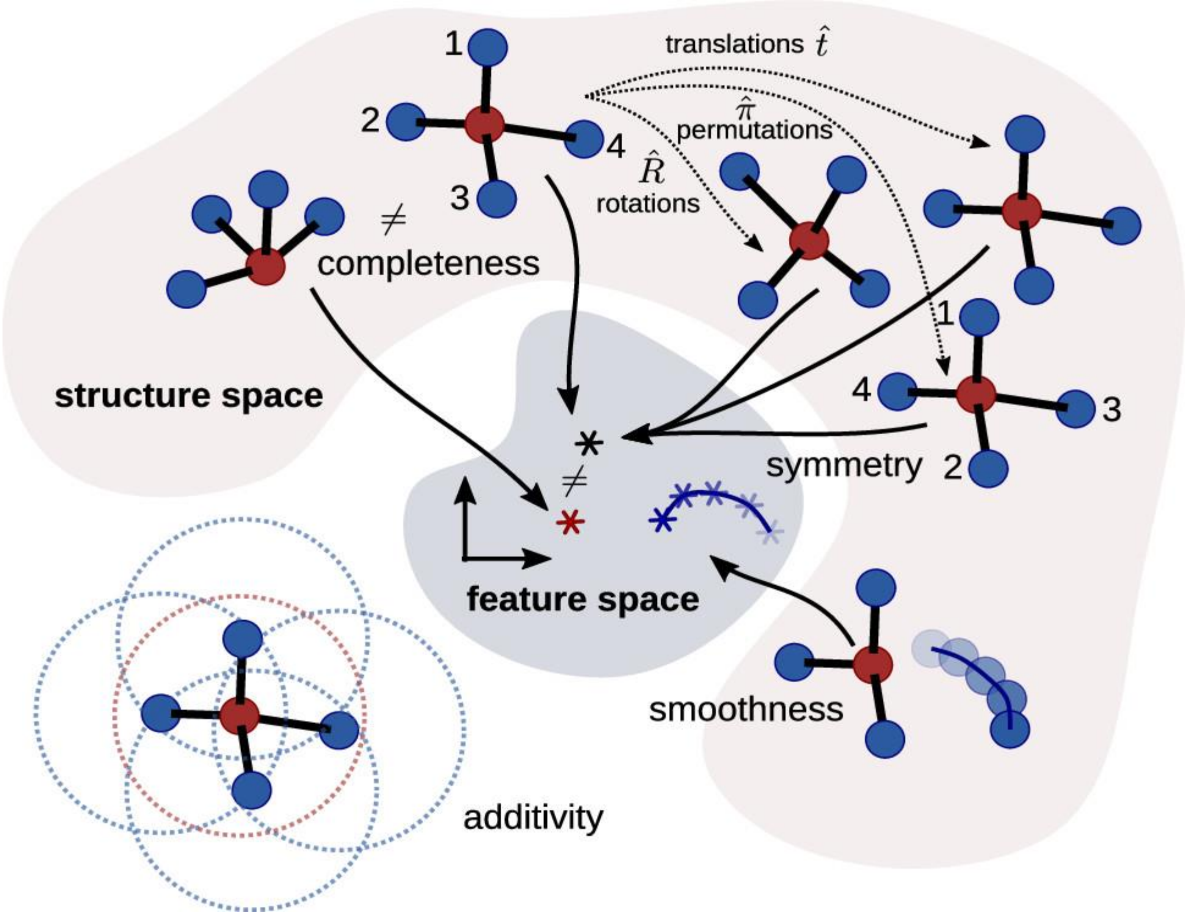
https://doi.org/10.1002/adma.201902765.



Different types of descriptors for atomic environments, as commonly used in empirical as well as ML interatomic potentials.

# Descriptors

- Descriptors must fulfill various symmetry requirements,
  - ✧Permutational invariance regarding exchange of two atoms of the same kind,
  - ✧Translational invariance,
  - ✧Rotational invariance.
- For a descriptor to be usable in practice, it is normally confined to a local environment of the atom, up to a given cutoff radius (typically, 5 or 6 Å).



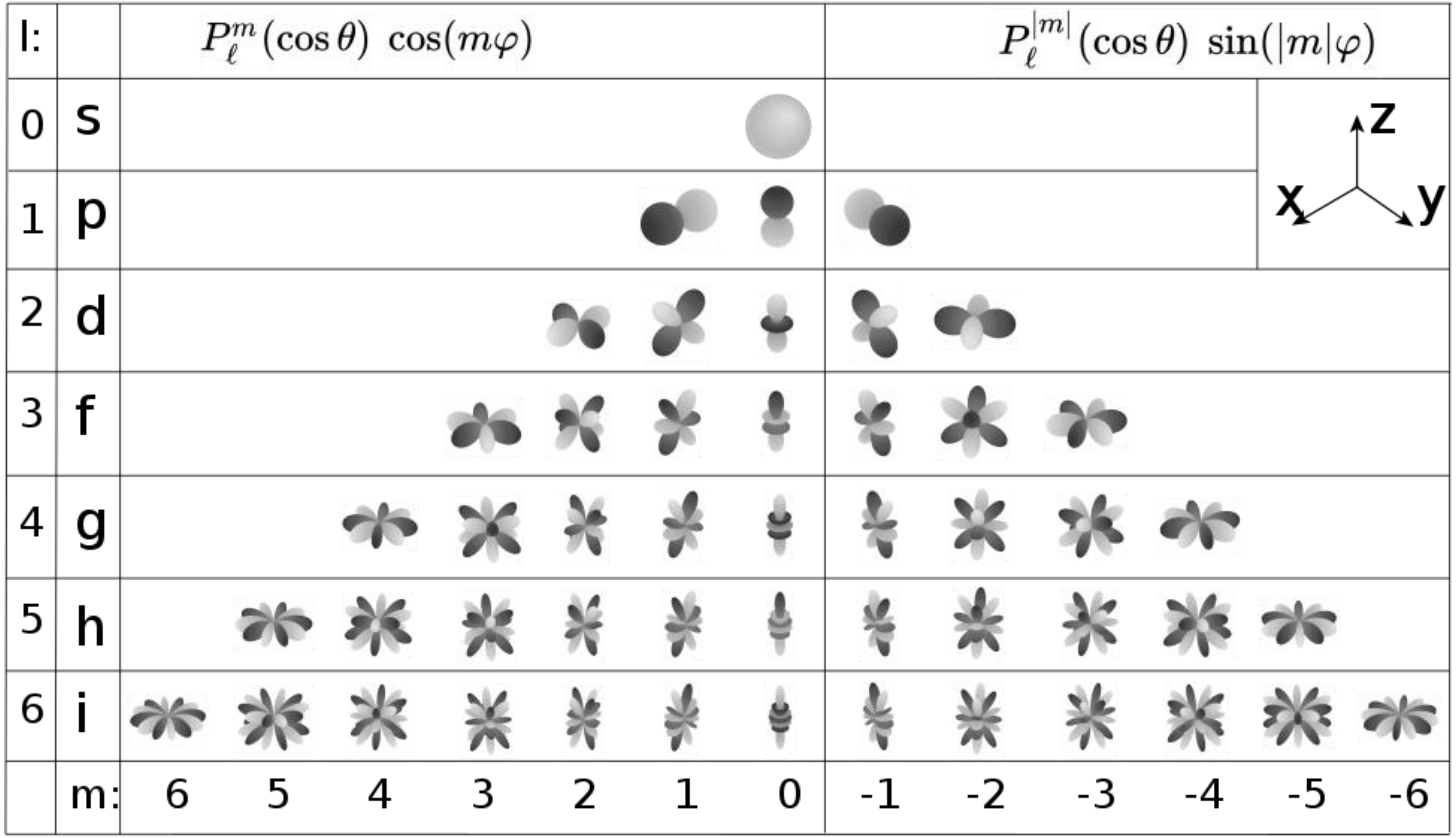
https://doi.org/10.1021/acs.chemrev.1c00021

# Common descriptors

- Atom-centered symmetry functions (ACSF)
  - ✧Used with NN.
- Smooth Overlap of Atomic Positions (SOAP)
  - ✧Used with GPR, called Gaussian approximation potentials (GAP) .
- Atomic cluster expansion (ACE)
  - ✧Used with LR.
- Moment tensor potentials
  - ✧Used with LR.

- More details can be found e.g in  https://doi.org/10.1063/5.0016005

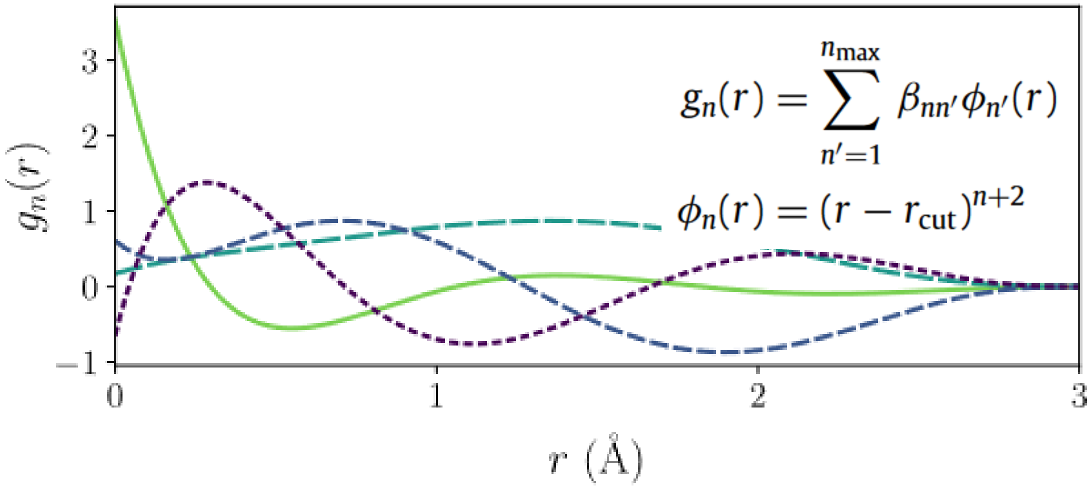
# Spherical harmonics $Y_m(\theta, \phi)$

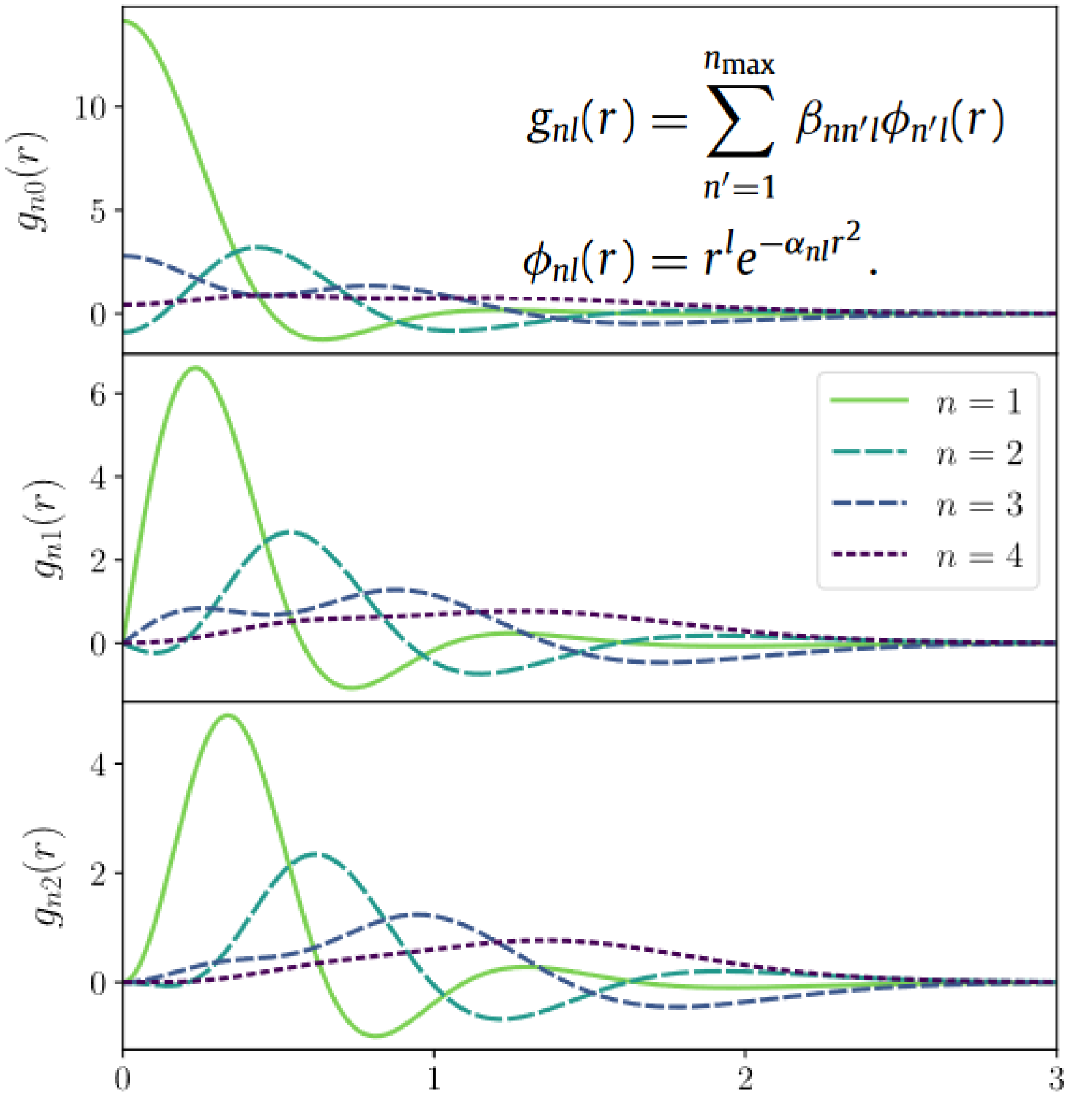

l:
$P_\ell^m(\cos\theta)\ \cos(m\varphi)$
$P_\ell^{|m|}(\cos\theta)\ \sin(|m|\varphi)$
0 s
1 p
2 d
3 f
4 g
5 h
6 i
m: 6 5 4 3 2 1 0 -1 -2 -3 -4 -5 -6
z
x
y

https://en.wikipedia.org/wiki/Spherical_harmonics

# Radial basis functions

- ➢ Common choices for orthonormal radial basis functions include
  - ✧Gaussian functions
  - ✧Polynomial functions





https://doi.org/10.1016/j.cpc.2019.106949

# Smooth Overlap of Atomic Positions (SOAP)

- The atomic density encodes the position of neighbouring atoms in a function

✧ $\rho_i(\boldsymbol{r}) = \sum_j^{neigh.} e^{\left(-\frac{\left|\boldsymbol{r}-\boldsymbol{r}_{\boldsymbol{ij}}\right|^{\boldsymbol{2}}}{2\,\sigma^2}\right)}$

- The SOAP kernel is calculated numerically by first expanding the atomic density in a basis set

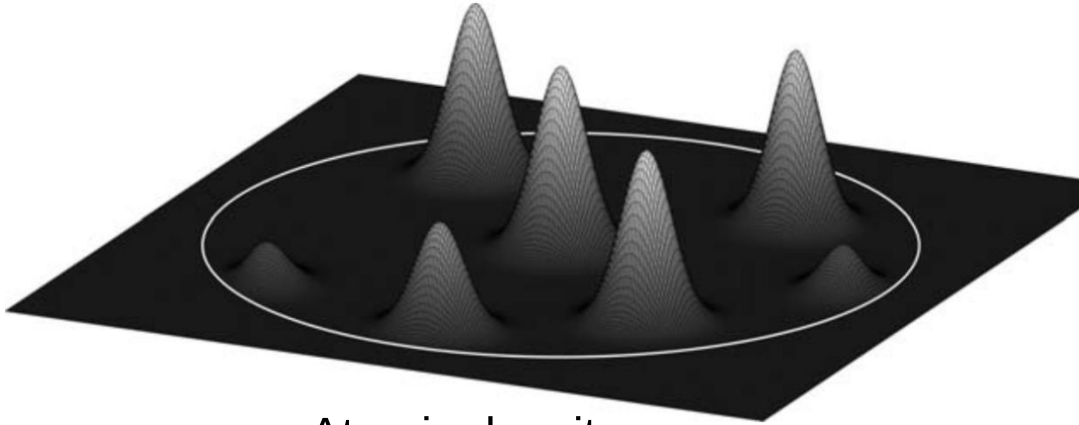
✧ $\rho_i(\boldsymbol{r}) = \sum_{nlm} c^i_{nlm} g_n(r) Y_m(\theta, \phi)$

- From these coefficients the SOAP descriptors are found by calculating the power spectrum:

✧ $p^{ij}_{nn'l} = \pi\sqrt{\frac{8}{2l+1}} \sum_m \left(c^i_{nlm}\right)^* c^j_{n'lm}$

- Descriptors are used with GPR to predict energies and forces

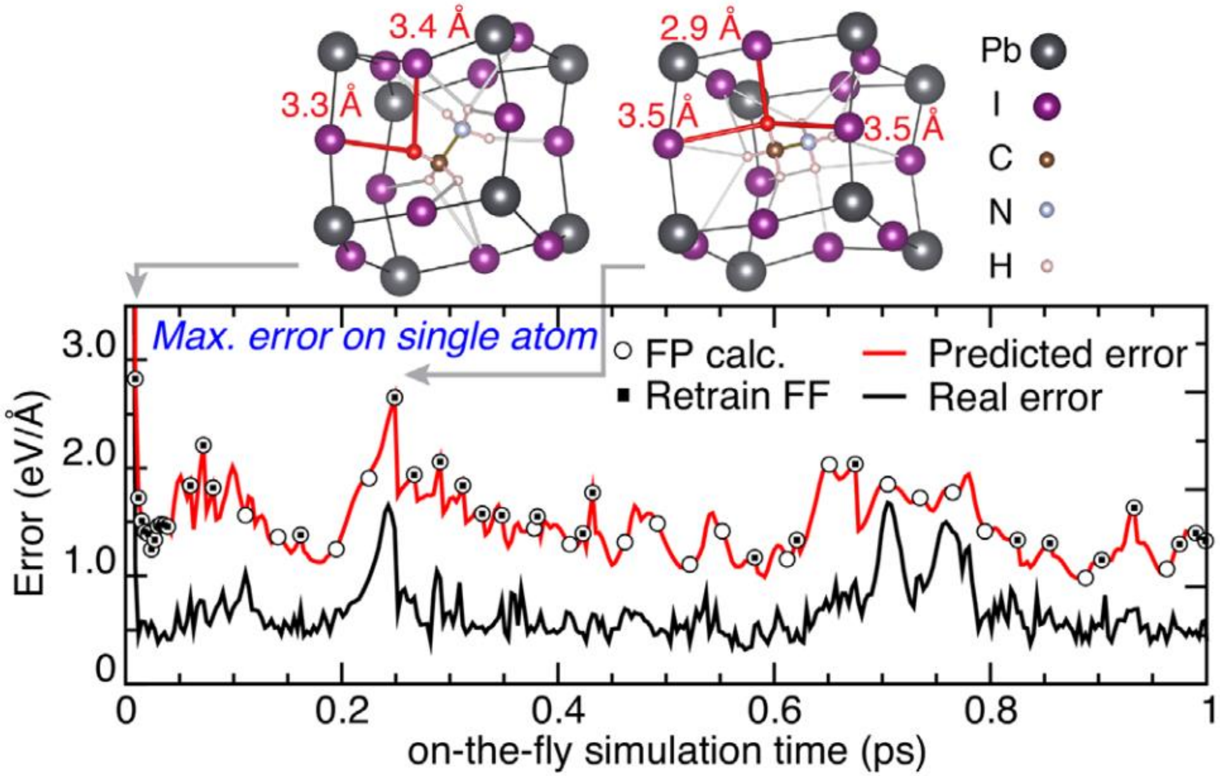
https://doi.org/10.1002/qua.24927



Atomic density

# Application to phase transition

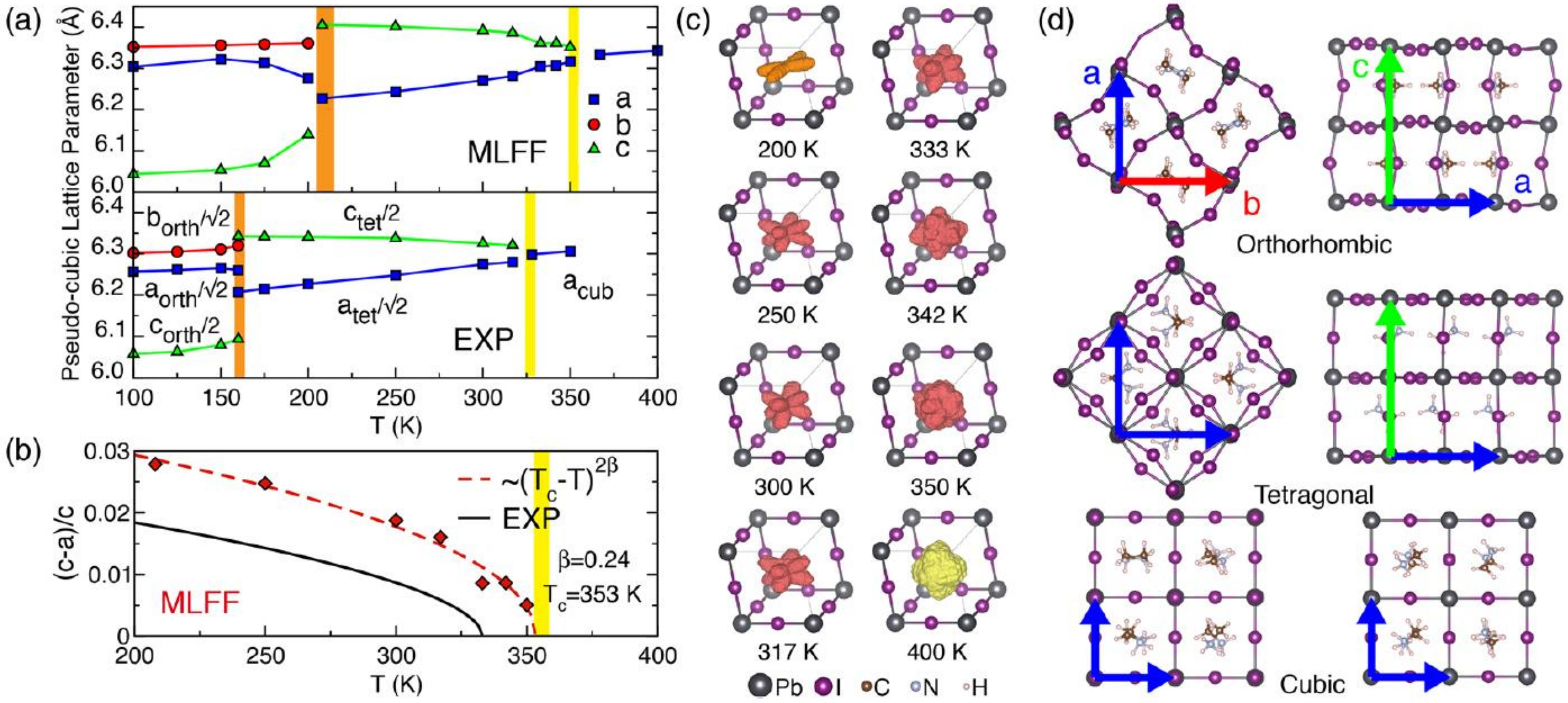
- Application to the entropy driven phase transitions of hybrid perovskites.
- Isothermal-isobaric simulations give direct insight into the underlying mechanisms.



10.1103/PhysRevLett.122.225701

# Application to phase transition

- Application to the entropy driven phase transitions of hybrid perovskites.
- Isothermal-isobaric simulations give direct insight into the underlying mechanisms.



# Atomic Cluster Expansion (ACE)

- Similar to SOAP, the ACE starts with a density projection

$$C_{nlm} = \sum_j g_n(r_j) Y_l^m(\hat{\boldsymbol{r}}_j),$$

- Isometry invariant features are then obtained by forming tensor products and integration of over the O(3) symmetry group

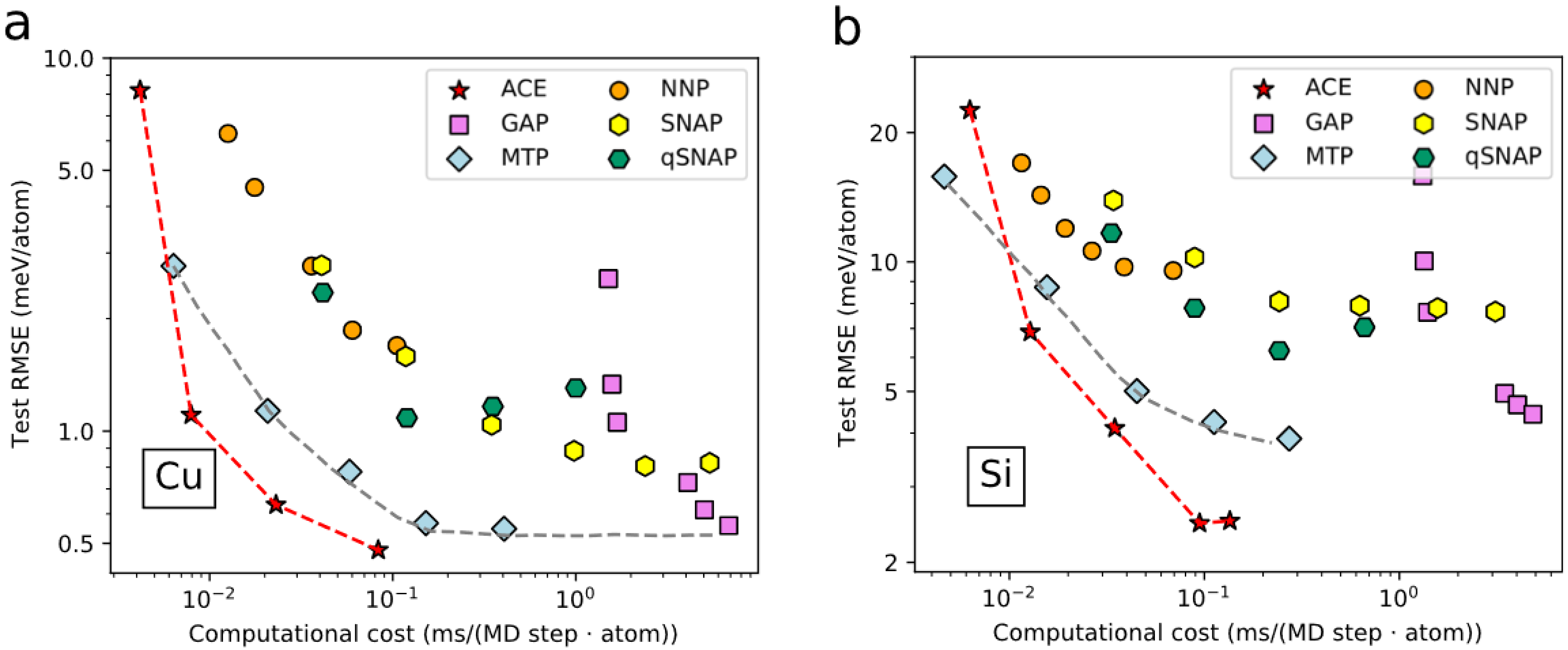
$$B_{\boldsymbol{nlm}} := \int_{\mathrm{O}(3)} \prod_{\alpha=1}^{N} C_{n_\alpha l_\alpha m_\alpha}.$$

- Furthermore, a linearly independent set is finally chosen.

- It can be shown that the energy scales linearly with the number of neighbors independent of the order of the expansion. https://doi.org/10.1103/PhysRevB.99.014104

- This is absolutely critical for a fast evaluation of higher-order terms in close-packed materials with hundreds of atoms within the cutoff sphere.

# Performance



https://doi.org/10.1038/s41524-021-00559-9

# Afternoon Tutorial 2

- ➢ Construct a GAP potential for W.
  - ✧ Build SOAP descriptors.
  - ✧ Training data from molecular dynamics runs of different atomic structures.
  - ✧ Predict atomic energies in grain boundaries.