DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

APPLICATIONS OF MACHINE LEARNING IN MATERIALS SCIENCE

MACHINE LEARNING AND MATERIALS @ JSI : RESEARCH DEPARTMENTS AT JSI

Physics

- Theoretical Physics F1
- Low and Medium Energy Physics F2
- Thin Films and Surfaces F3
- Surface Engineering F4
- Condensed Matter Physics F5
- Gaseous Electronics F6
- Complex Matter F7
- Reactor Physics F8
- Experimental Particle Physics F9

Electronics and Information Technologies

- Automation, Biocybernetics and Robotics E1
- Systems and Control E2
- Artificial Intelligence E3
- Open Computer Systems and Networks E5
- Communication Systems E6
- Computer Systems E7
- Knowledge Technologies E8
- Intelligent Systems E9

Chemistry and Biochemistry

- Biochemistry and Molecular Biology B1
- Molecular and Biomedical Sciences B2
- Biotechnology B3
- Inorganic Chemistry and Technology K1
- Physical and Organic Chemistry K3
- **Electronic Ceramics K5**
- **Nanostructured Materials K7**
- **Synthesis of Materials K8**
- **Advanced Materials K9**
- Environmental Sciences O2

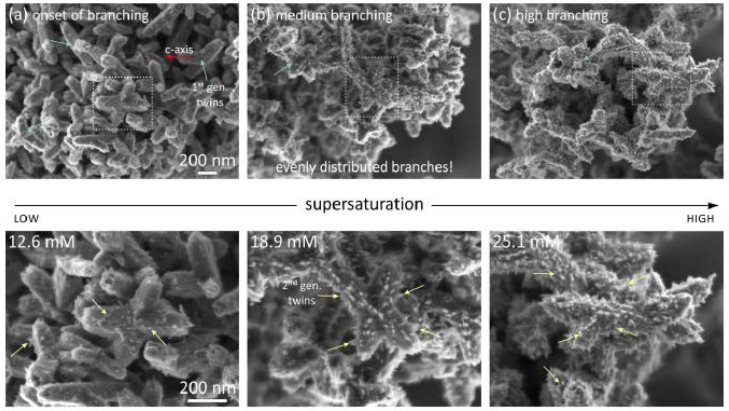
Reactor Engineering and Energetics

- Reactor Engineering R4

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Machine Learning for Materials Science
- Data-

- Material synthesis process parameters
  - Temperature, time, precursor concentration…
- Quantum-level properties
  - DFT simulation results, band-gap
- Material morphology (images)
- Functional properties of materials
  - Electrical conductivity, catalytic activity, magnetic properties, anti-corrosive properties



Machine Learning for Materials Science
- Tasks-

- Find relations between the different kinds of data
- Primarily predict functional properties of materials from the other kinds of data (synthesis parameters, morphology, quantum-level properties)
- But also find relations between other pairs/combinations of data, e.g.:
  - Parameters of synthesis and
  - Morphology of the material

  or

  - Morphology of materials and
  - Properties of material

Example Application 1: Relating the composition and mechanical properties of tungsten-based composites with ML

- Data

Independent variables

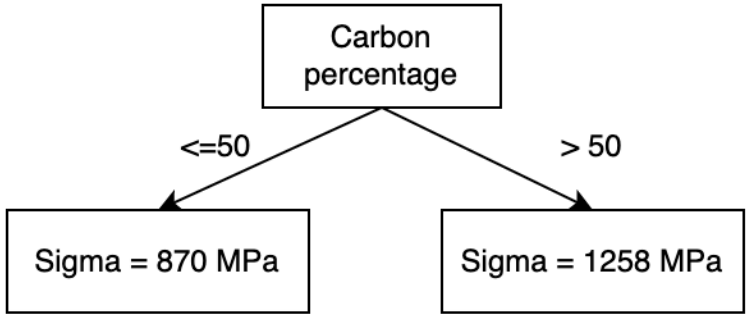
- Percentage of added carbon
- Pressure applied
- Dimensions of the sample (height, diameter, mass)
- Sintering-related measures
  - Temperature
  - Heating rate to final maximum temperature (step)
  - Heating time at maximum temperature

Targets

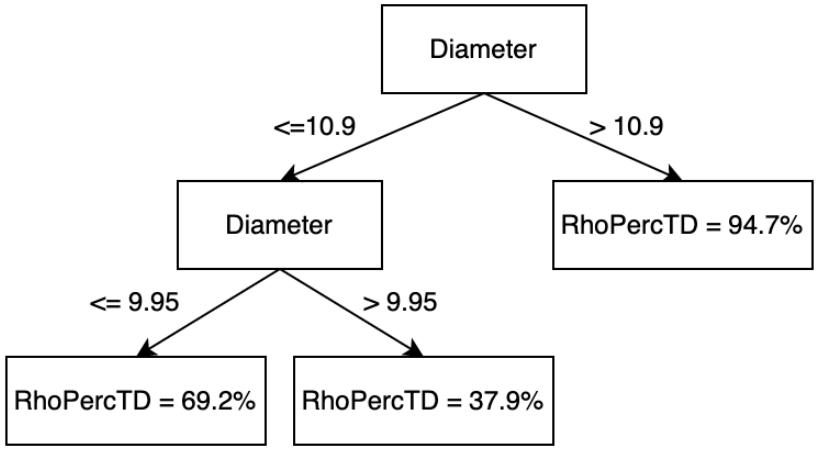
- Percentage of theoretical density of the sample (Rho%TD, RhoPercTD)
- Flexural strength (Sigma)

Example Application 1: Relating the composition and mechanical properties of tungsten-based composites with ML

- Task: Single-target Prediction



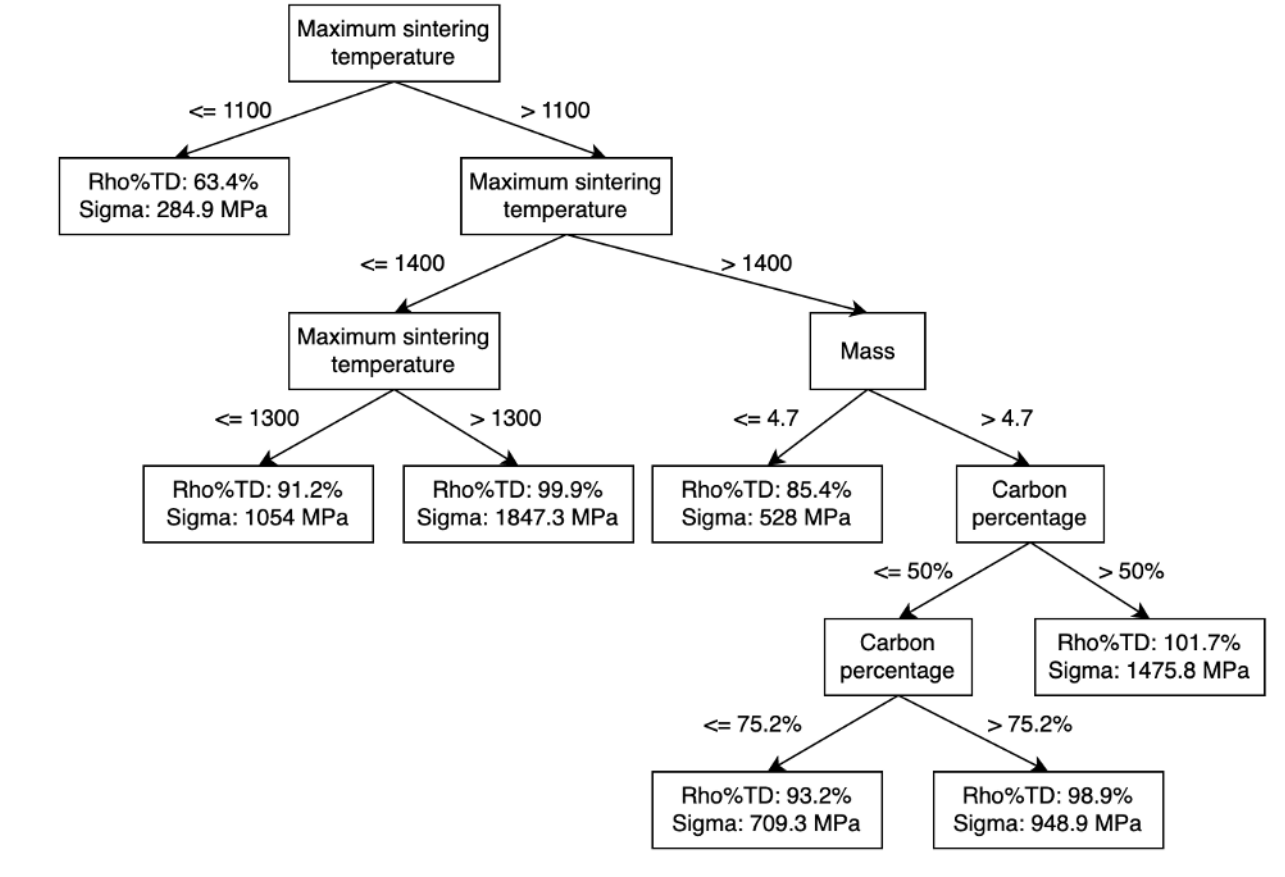
Regression tree for predicting flexural strength in MegaPascals



Regression tree for predicting percentage of theoretical density

Example Application 1: Relating the composition and mechanical properties of tungsten-based composites with ML

- Task: Multi-target Prediction



Example Application 1: Relating the composition and mechanical properties of tungsten-based composites with ML

- Task: Feature ranking

Feature ranking for flexural strength	Feature ranking for Rho%TD	Feature ranking for multi-target prediction
1. CarbonPerc	1. d	1. SinteringMax
2. SinteringMax	2. m	2. CarbonPerc
3. h	3. CarbonPerc	3. h
4. m	4. h	4. m
5. d	5. SinteringMax	5. d
6. Pressure	6. Pressure	6. Pressure
7. SinteringTime	7. SinteringTime	7. SinteringTime
8. SinteringStep	8. SinteringStep	8. SinteringStep

Example Application 2a: Predict Foaming Glass properties from process parameters

Process parameter	Range
Water glass content [wt.%]	0 – 30
Carbon black content [wt.%]	0 – 0.84
Mn_3O_4 content [wt.%]	0 – 10
Furnace temperature [°C]	700 – 900
Foaming time [min]	0 – 60
K_3PO_4 content [wt.%]	0 – 4
Drying	Yes / No
Mixing	Yes / No

Feature space

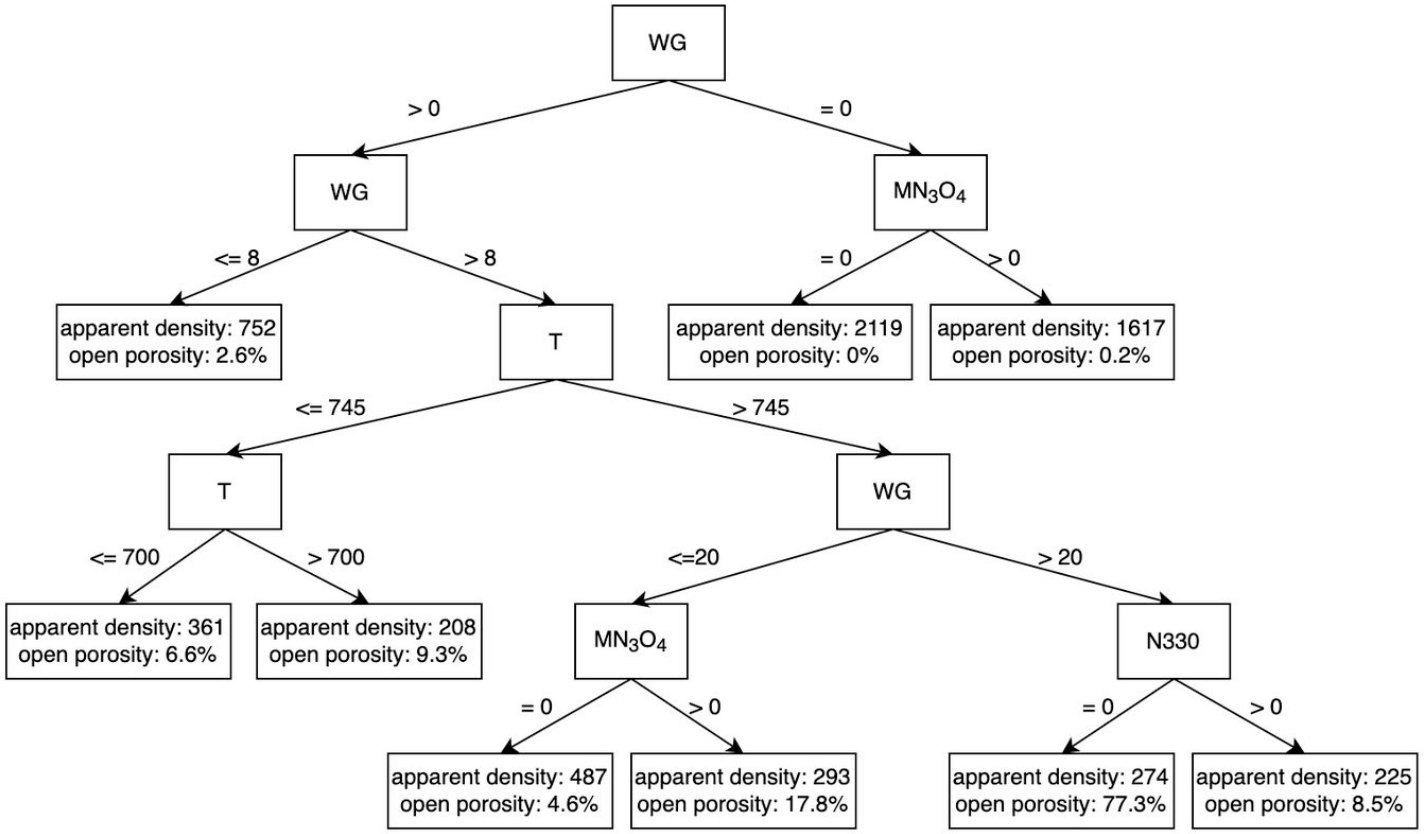
ML model

Process property
Apparent density ρ_{app}
Pycnometric density ρ_{pyc}
Overall porosity ε_{total}
Closed porosity ε_{closed}
Open porosity ε_{open}

Target space

Example Application 2a: Predict Foaming Glass properties from process parameters

- Solution: Multi-target regression tree



Example Application 2a: Predict Foaming Glass properties from process parameters

- Estimated parameter importance

Process variable	ϱ_{app}	ϱ_{pyc}	ε_{total}	ε_{closed}	ε_{open}
Water glass content	1	1	1	1	1
Carbon black content	3	2	3	3	3
Mn_3O_4 content	2	3	2	4	4
Furnace temperature	4	4	4	2	2
Foaming time	5	5	5	6	6
K_3PO_4	8	7	8	5	5
Drying	6	6	6	7	7
Mixing	7	8	7	8	8

Example Application 2b: Predict Foaming Glass properties from process parameters - SSL

- Task: Predict foamed glass properties (5 of them) from the values of the parameters of the glass foaming process
- The data describe 165 samples of foamed glasses
- All parameters of the glass foaming process were known
- But for the properties, values were missing for closed and open porosities for 41 samples
- Partially labeled data

Example Application 2b: Predict Foaming Glass properties from process parameters - SSL

- Task: Predict foamed glass properties (5 of them) from the values of the parameters of the glass foaming process

Process parameter	Range
Water glass content [wt.%]	0 – 30
Carbon black content [wt.%]	0 – 0.84
Mn_3O_4 content [wt.%]	0 – 10
Furnace temperature [°C]	700 – 900
Foaming time [min]	0 – 60
K_3PO_4 content [wt.%]	0 – 4
Drying	Yes / No
Mixing	Yes / No

Process property	Value
Apparent density ρ_{app}	Low / High
Pycnometric density ρ_{pyc}	Low / High
Overall porosity ε_{total}	Low / High
Closed porosity ε_{closed}	Low / High
Open porosity ε_{open}	Low / High

Example Application 2b: Predict Foaming Glass properties from process parameters - SSL

- Results: Area under the precision/recall curve
- Average across the five targets

Average AUC	Fully-supervised learning	Semi-supervised learning (PC)	Semi-supervised learning (self training)
Single MLC tree	0.470	0.557	/
Ensemble (RF of 100 MLC trees)	0.625	0.633	0.666

Example Application 2b: Predict Foaming Glass properties from process parameters - SSL

- Feature rankings obtained from the supervised and semi-supervised (PC) ensembles

Fully-supervised rank	Semi-supervised (PC) rank	Process variable
1	5	Carbon black content
2	1	Water glass content
3	6	Mn_3O_4 content
4	8	Furnace temperature
5	3	K_3PO_4 content
6	4	Mixing
7	7	Foaming time
8	2	Drying

Example Application 2b: Predict Foaming Glass properties from process parameters - SSL

Example supervised PCT for MLC

```
FURNACE_TEMPERATURE > 745.0
+--yes: [0,0,0,1,1] [54.0,52.0,47.0,34.0,34.0]: 54
+--no:  FURNACE_TEMPERATURE > 700.0
      +--yes: [0,1,1,1,1] [15.0,18.0,20.0,22.0,22.0]: 22
      +--no:  FOAMING_TIME > 30.0
            +--yes: Mn3O4 > 0.0
            |     +--yes: [1,1,1,1,1] [2.0,2.0,2.0,2.0,2.0]: 2
            |     +--no:  [0,0,0,1,1] [4.0,4.0,4.0,4.0,4.0]: 4
            +--no:  [0,0,0,1,1] [29.0,28.0,27.0,29.0,29.0]: 29
```

Example SSL PCT for MLC

```
Mn3O4 > 0.0
+--yes: MIXING = ADDITIONAL
|     +--yes: WG > 12.0
|     |     +--yes: [0,0,0,1,1] [7.0,5.0,5.0,8.0,8.0]: 8
|     |     +--no:  [0,0,0,0,0] [13.0,13.0,11.0,7.0,7.0]: 13
|     +--no:  K3PO4 > 2.0
|           +--yes: WG > 12.0
|           |     +--yes: [1,1,1,1,1] [3.0,6.0,6.0,6.0,6.0]: 6
|           |     +--no:  [0,0,0,1,1] [5.0,4.0,4.0,5.0,5.0]: 5
```

Semi-supervised and Active learning

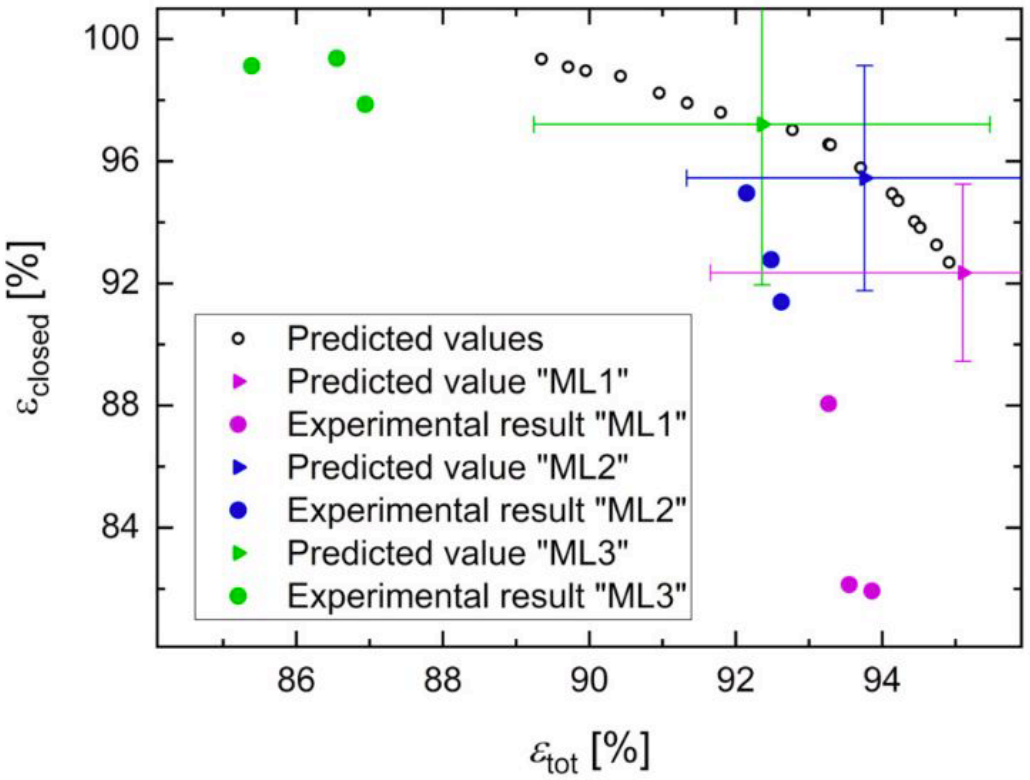
- Similarity: Both use labeled as well as unlabeled data
- Difference:
  - Semi-supervised learning has to use the unlabeled data as is
  - Active learning gets to pick some of the unlabeled examples and ask for their labels
- Returning to our foamed glass example
  - Unlabeled data refer to foamed glass samples (that may or may not have been synthesized, say they have been), but (some of) their properties have not been measured
  - Active learning can ask for property measurements for some foamed glass samples (which may require their synthesis)

Example Application 2c: Combining machine learning and optimization

- First, use machine learning to learn a model M predicting material properties (P), from synthesis parameters (S)
- P = M(S)
- The task of material design is to find synthesis parameters' values s that optimize the values of the properties P
- For this task, we can use an optimization algorithm over the space of synthesis parameter values
  - If we are optimizing the value of one parameter, we can use single-criterion optimization
  - If we are looking at multiple properties, we need to use multi-criteria/objective optimization (MOO)

Example Application 2c: Combining machine learning and optimization

- Learn neural network predicting two properties (density and porosity)
- Use the NN to search for synthesis parameters that lead to non-dominated combinations of properties



Application 3: Predicting Hydrogen adsorption energies on Platinum Nanoparticles and surfaces

The problem

- Platinum is a highly active catalyst for the hydrogen evolution reaction (HER) — a key step in green hydrogen production from water electrolysis
- The hydrogen adsorption energy is the key descriptor of catalytic activity
- Computing it with DFT is accurate but computationally expensive — infeasible for large nanoparticles with many adsorption sites

The ML approach

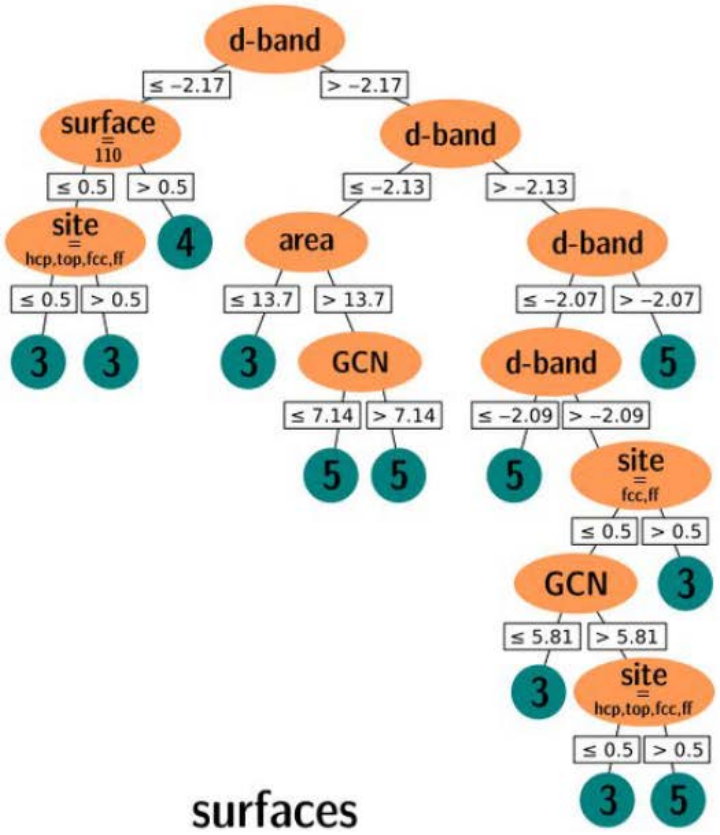
- Two datasets: platinum surfaces (46 data points) and nanoparticles (85 points)
- Features: geometric descriptors (coordination number, generalized coordination number) and electronic descriptors (d-band center)
- Methods: linear regression, random forests, regression trees (M5')

Key findings

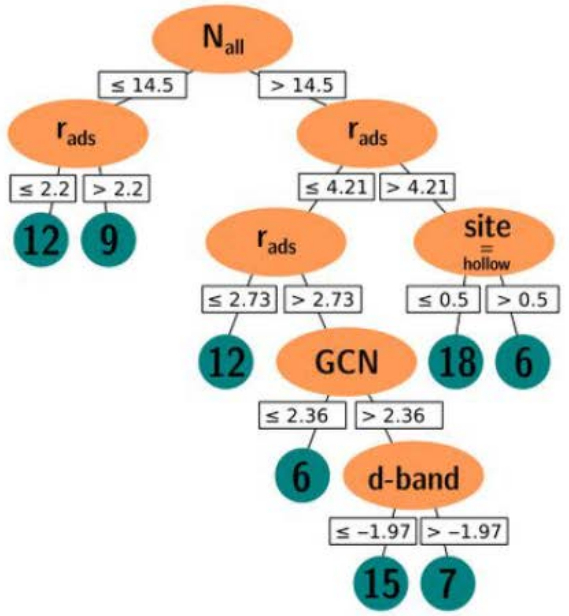
- Surfaces predicted better than nanoparticles (RMSE ±0.13 eV vs ±0.22 eV)
- Most important feature for surfaces: d-band center (electronic structure)
- Most important feature for nanoparticles: coordination number (local geometry)
- Adsorption is a **local phenomenon** — local environment descriptors dominate in both cases

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Example Application 3:
Predicting Hydrogen adsorption energies on Platinum Nanoparticles and surfaces with ML



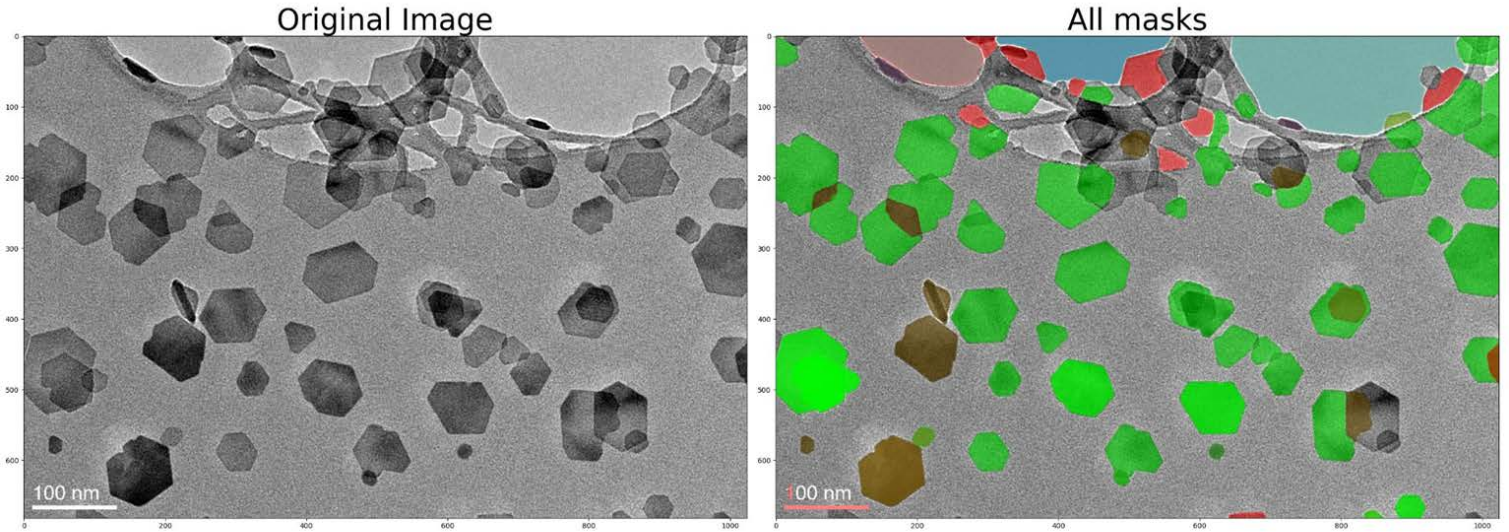
surfaces



nanoparticles

Example Application 4: Using Visual Foundational Models in Materials Science

- Data: TEM images of Barium hexaferrite nanoplatelets
  - Images from Transmission Electron Microscopy (TEM) experimental analyses of Barium hexaferrite nanoplatelets (BHF NPLs)
  - Particle size (diameter) distribution predicts magnetic properties
  - The task is to segment/identify the particles in the images



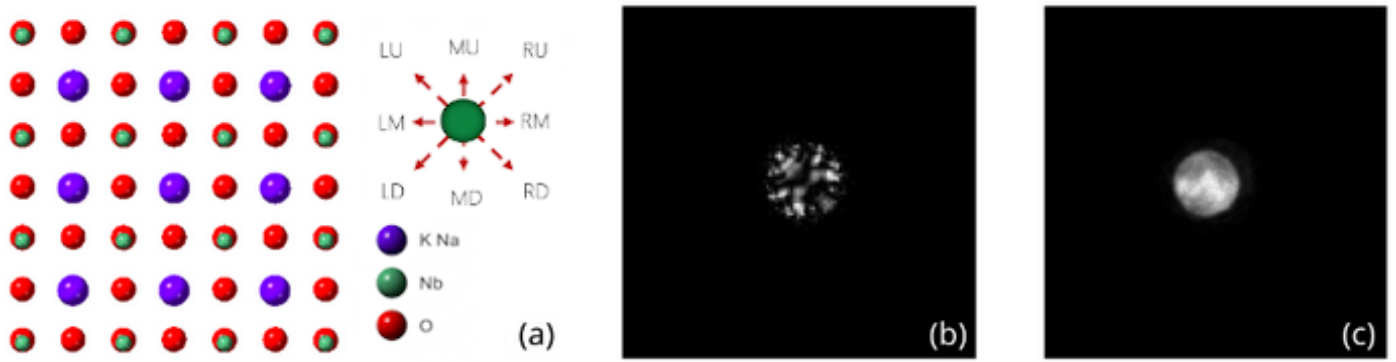
Example Application 4: Using Visual Foundational Models in Materials Science

- Segment Anything Model (SAM2): visual foundation model specialized in segmentation, used to identify the particles in zero-shot mode
  1. SAM2 creates candidate masks
  2. Mask feature extraction encodes shape/size and color/texture
  3. Custom agglomerative clustering groups masks into semantic sets
  4. Outliers are discarded to extract the targets

- Post-segmentation clustering significantly improves SAM2's zero-shot pipeline performance on complex EM images.
- The approach eliminates the requirement for manual fine-tuning annotations in data-scarce settings.

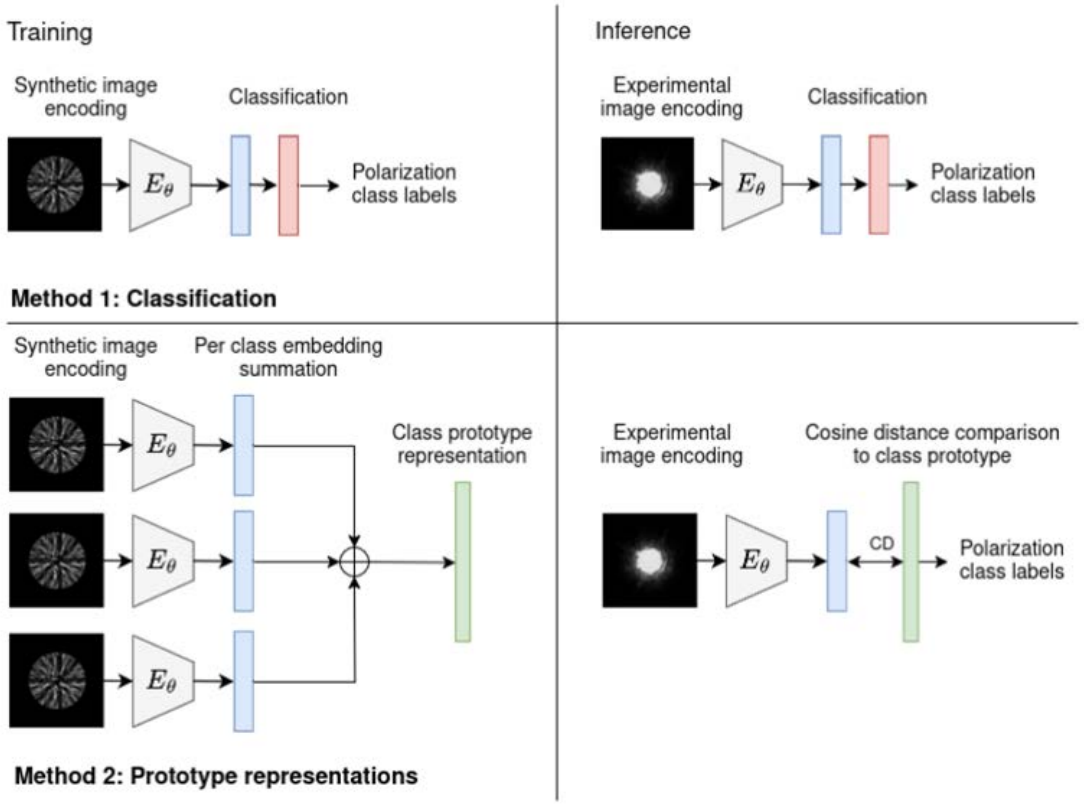
Example Application 5:
ML approaches for polarization mapping in ferroelectrics using 4D-STEM

- 4D-STEM synthetic-to-synthetic and synthetic-to-experimental evaluation
- Comparison of supervised strategies for classifying spontaneous polarization direction in KNN perovskite crystals using synthetic training data and experimental 4D-STEM scans
- Data:
  - 8 polarization directions in synthetic training set
  - 4 learning paradigms compared



Example Application 5:
ML approaches for polarization mapping in ferroelectrics using 4D-STEM

- Approaches:
  - Classification
  - Prototype Representation
  - Regression
  - PCA + K-NN
- Models:
  - ResNet
  - VGG
  - Custom CNN



Example Application 5:
ML approaches for polarization mapping in ferroelectrics using 4D-STEM

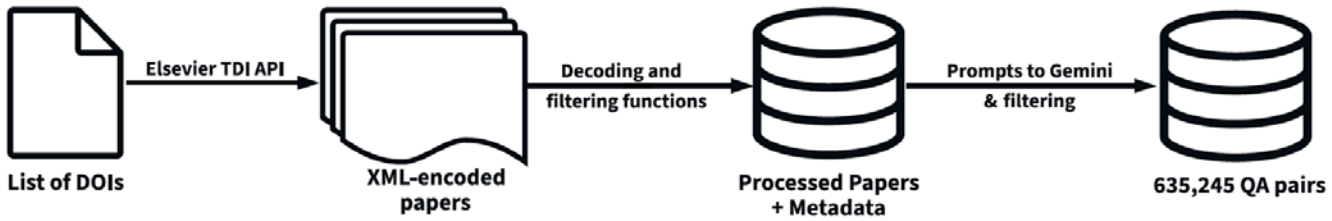
- All models generalize well to other synthetic datasets of the same sample thickness
- Extensive preprocessing helps with bridging the domain gap between synthetic and experimental datasets
- PCA-based approach showed the most robust performance, predicting correct polarization majority classes for all experimental datasets in two out of three preprocessing configurations and also shows potential in anomaly detection
- The low mean accuracy and high variance of the best two approaches (PCA and CNN (Proto)) on the experimental datasets indicate that these methods are not yet reliable for fine-grained, pixel-level classification

Example Application 6: Materials Science Literature Q and A dataset (MSLitQA)

- Open-weight LLMs are attractive for scientific assistants
  - Support reproducibility, auditability, local deployment, and open scientific infrastructure
  - Materials science remains under-served by existing instruction datasets
- Materials science literature is uniquely challenging for LMs
  - Combination of chemistry, physics, crystallography, specialized notation, and complex interconnected scientific knowledge
- Existing domain resources improve materials-text understanding and extraction
  - Unclear whether large-scale synthetic instruction data from scientific papers can effectively transfer materials-science reasoning into lightweight generative LLMs

Example Application 6: Materials Science Literature Q and A dataset (MSLitQA)

- Dataset creation



Dataset	Size	Description
PaperDS	70,229 papers	Cleaned full-text MS corpus
MSLitQA	635,245 QA pairs	Synthetic dataset derived from PaperDS
Open-ended QA	477,920	Explanatory domain questions
Numerical Short QA	58,714	Numerical questions w\numbers isolated
Numerical Long QA	11,570	Remaining numerical questions
Multiple-choice QA	87,041	MCQ with shuffled answer labels

Example Application 6: Materials Science Literature Q and A dataset (MSLitQA)

- Retrieval accuracy on *MaScQA* benchmark

Setting	Overall	MCQ	MATCH	NUM-MCQS	NUM
Gemma 3 4B IT, zero-shot	26.8%	44.9%	35.7%	11.9%	5.7%
Gemma 3 4B IT + RAND5 retrieval	30.6%	51.6%	**41.4%**	19.4%	4.4%
Gemma 3 4B IT + TOP5 retrieval	**35.4%**	**58.2%**	38.6%	**26.9%**	**8.3%**

- Retrieval augmentation improves *MaScQA* performance over zero-shot prompting, even with random examples
- Semantic retrieval (TOP5) achieves the best results
  - Accuracy increased from 26.8% to 35.4%
  - Largest gains on multiple choice questions
- Numerical free-form questions remain challenging
  - Numerical reasoning requires more than contextual retrieval alone

Example Application 6: Materials Science Literature Q and A dataset (MSLitQA)

- Fine-tuning accuracy on *MaScQA* benchmark
  - MSLitQA has useful knowledge but does not reliably transfer to the model with standard fine-tuning

Model / Setting	Base Model	Overall	MCQS	MATCH	NUM-MCQS	NUM
Gemma 3 4B IT baseline	4B IT	26.8%	44.9%	35.7%	11.9%	5.7%
Gemma 3 4B IT + TOP5	4B IT	35.4%	58.2%	38.6%	26.9%	8.3%
Gemma 3 4B IT + CoT	4B IT	32.0%	51.6%	38.6%	23.9%	7.9%
PaperDS SFT	4B IT	24.2%	43.5%	25.7%	13.4%	2.6%
QuestionDS SFT	4B IT	25.5%	42.8%	22.9%	29.9%	3.5%
QuestionDS DPO	4B IT	13.5%	24.6%	15.7%	7.5%	0.9%
QuestionDS ORPO	4B IT	26.2%	39.3%	35.7%	28.4%	6.1%
Response-only SFT	4B IT	22.6%	38.6%	22.9%	16.4%	4.4%
Format SFT	4B IT	30.0%	47.7%	40.0%	28.4%	5.3%
PT + CFT	4B PT	31.4%	52.6%	40.0%	17.9%	6.1%
Gemma 3 12B IT baseline	12B IT	54.3%	65.6%	78.6%	26.9%	40.8%
Gemma 3 12B IT full FT	12B IT	45.8%	66.7%	65.7%	16.4%	22.4%