Materials Genomics

Philipp Pelz

Corresponding author: Philipp Pelz,

Abstract

This course introduces students to materials genomics, treating the periodic table and the space of known crystal structures as a searchable, computable design space. Students learn how materials databases are built, how simulation methods generate materials data, how atomic structure is represented numerically, how structure–property relationships are learned using machine learning, and how uncertainty-aware models enable accelerated materials discovery.

1 Course Information

5th Semester – 5 ECTS · 2h lecture

Coordinated with "Mathematical Foundations of AI & ML" (MFML) and "ML for Materials Processing & Characterization" (ML-PC)

---

2 Course Philosophy

Materials genomics views the periodic table and all known crystal structures as a **high-dimensional design space**.

In this course, students learn to:

- understand how materials data is generated by simulations and experiments,
- treat materials data as a structured, learnable representation space,
- move beyond classical descriptors toward learned representations,
- use ML models as surrogates for quantum-mechanical and continuum simulations,
- reason about uncertainty, stability, and discovery,
- understand how computational screening integrates with experiments.

The course explicitly **builds on MFML**:

- PCA and regression are assumed background,
- neural networks, representation learning, and uncertainty are used, not rederived.

---

3 Week-by-Week Curriculum (14 weeks)

3.1 Unit I — Where Materials Data Comes From (Weeks 1–4)

3.1.1 Week 1 – What is Materials Genomics? (14.04.2026)

- Genomics analogy: genes → functions vs atoms → properties.
- Structure–property–processing paradigm from a *structure-first* viewpoint.
- Materials databases as design spaces: Materials Project, OQMD, AFLOW, NOMAD.

Exercise:

Explore Materials Project; query bandgaps, formation energies, symmetries.

3.1.2 Week 2 – Simulation methods as data generators (21.04.2026)

- Why simulations dominate materials data generation.
- Simulation methods as mappings from assumptions to data.
- Overview of scales and outputs:
  - FEM: continuum fields (stress, strain).
  - MD: trajectories, forces, diffusion.
  - MC: thermodynamic sampling.
  - DFT: energies, electronic structure.
- Accuracy–cost–scale trade-offs and systematic biases.

Exercise:
For selected materials properties, identify suitable simulation methods and expected biases.

3.1.3 Week 3 – Atomistic and electronic simulations (DFT, MD, MC) (28.04.2026)

- Density Functional Theory: ground-state bias, exchange–correlation functionals, consistency vs accuracy.
- Molecular Dynamics: force fields, time averaging, limitations of timescales.
- Monte Carlo: phase-space sampling and thermodynamic averages.
- What quantities in materials databases come directly from simulations.

Exercise:
Inspect Materials Project entries; identify simulation assumptions and derived quantities.

3.1.4 Week 4 – Continuum simulations, thermodynamics, and stability (05.05.2026)

- FEM as a structure–property mapping at the continuum scale.
- Constitutive models as implicit surrogates.
- Formation energies, convex hulls, metastability.
- Why "stable" does not imply "synthesizable".

Exercise:
Analyze stability and simulated properties for a small materials system.

3.2 Unit II — Representations of Materials (Weeks 5–7)

(Aligned with early neural networks in MFML)

3.2.1 Week 5 – From classical descriptors to learned representations (12.05.2026)

- Classical descriptors: Magpie, matminer.
- Limits of hand-crafted features.
- Motivation for representation learning.

Exercise:
Build a simple property predictor using classical descriptors.

3.2.2 Week 6 – Graph-based crystal representations (19.05.2026)

- Crystals as graphs: nodes, edges, periodic boundary conditions.
- Intuition behind CGCNN and MEGNet.
- Relation to neural network concepts from MFML.

Exercise:
Construct and visualize graph representations of crystal structures.

3.2.3 Week 7 – Local atomic environments (26.05.2026)

- Local vs global representations.
- Coordination environments, Voronoi tessellations.
- SOAP descriptors as a bridge to learned representations.

Exercise:

Compute SOAP vectors and explore similarity in descriptor space.

3.3 Unit III — Learning Structure–Property Relations (Weeks 8–10)

3.3.1 Week 8 – Regression and generalization in materials data (02.06.2026)

- Predicting bandgaps, elastic moduli, formation energies.
- Bias–variance trade-off and overfitting.
- Dataset size vs model complexity.

Exercise:
Compare linear, random forest, and neural network regressors.

3.3.2 Week 9 – Neural networks for materials properties (09.06.2026)

- Neural networks as surrogates for DFT-level properties.
- Training pitfalls: data leakage, imbalance, extrapolation.
- Interpretability challenges.

Exercise:
Train a small neural network and analyze generalization behavior.

3.3.3 Week 10 – Representation learning and feature discovery (16.06.2026)

- Learned vs engineered features.
- Transferability across chemical systems.
- What networks learn about chemistry and structure.

Exercise:
Compare model performance using raw descriptors vs learned embeddings.

3.4 Unit IV — Latent Spaces, Uncertainty, and Discovery (Weeks 11–13)

3.4.1 Week 11 – Latent spaces of materials (23.06.2026)

- Autoencoders and embeddings for crystal data.
- Interpreting latent dimensions.
- Structure families and chemical intuition.

Exercise:
Train an autoencoder; visualize latent materials space.

3.4.2 Week 12 – Clustering, uncertainty, and discovery logic

- Why clustering is not discovery.
- Outliers, anomalies, and candidate identification.
- Aleatoric vs epistemic uncertainty.

Exercise:
Contrast clustering results with latent-space exploration.

3.4.3 Week 13 – Uncertainty-aware discovery and Gaussian Processes (07.07.2026)

- Gaussian Process regression as a gold standard for uncertainty.
- Exploration vs exploitation.
- Relevance to materials acceleration platforms.

Exercise:

Compare GP regression and neural network ensembles for screening tasks.

3.5 Unit V — Constraints, Trust, and Synthesis (Week 14)

3.5.1 Week 14 – Physical constraints, limits, and outlook (14.07.2026)

- Stability, charge neutrality, and symmetry constraints.
- Physics-informed learning in materials discovery.
- What ML can and cannot discover.
- Integration with experimental workflows.

Exercise:
Mini-project synthesis and presentation.

---

4 Learning Outcomes

Students completing this course will be able to:

- Explain how simulation methods generate materials data and introduce bias.
- Navigate and interrogate major materials databases.
- Represent crystal structures using descriptors, graphs, and learned embeddings.
- Train and evaluate ML models for predicting materials properties.
- Understand latent spaces and their role in materials discovery.
- Quantify and interpret uncertainty in materials predictions.
- Apply ML responsibly to accelerate materials screening.
- Critically assess the limits of data-driven materials discovery.

Source: Article Notebook