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 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

4th/5th Semester – 5 ECTS · 2h lecture + 2h exercises per week

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:

- treat materials data as a structured, learnable representation space,
- move beyond classical descriptors toward learned representations,
- use ML models as surrogates for quantum-mechanical calculations,
- 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 — Materials Data as a Design Space (Weeks 1–3)

3.1.1 Week 1 – What is Materials Genomics?

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

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

3.1.2 Week 2 – Crystal structures, symmetry, and low-dimensional structure

- Crystal structures as data objects.
- Space groups, Wyckoff positions, symmetry constraints.
- PCA as an *exploratory tool* for structural/property data (refresher).

Exercise:
Use pymatgen/spglib to analyze symmetry; visualize PCA of structural features.

3.1.3 Week 3 – Materials databases & thermodynamic quantities

- File formats: CIF, POSCAR, database schemas.
- Formation energies, convex hulls, metastability.

- What databases do *not* contain (bias, incompleteness).

Exercise:
Parse CIF files; compute basic structural properties; analyze stability.

3.2 Unit II — Representations of Materials (Weeks 4–6)

(Aligned with early neural networks in MFML)

3.2.1 Week 4 – From classical descriptors to learned representations

- Classical descriptors: Magpie, matminer (composition-based).
- Limits of hand-crafted features.
- Why representation learning matters.

Exercise:
Build a simple property predictor using classical descriptors.

3.2.2 Week 5 – Graph-based crystal representations

- Crystals as graphs: nodes, edges, periodicity.
- Intuition behind CGCNN, MEGNet (no architecture deep dive).
- Relation to MFML neural network concepts.

Exercise:
Construct a graph representation of crystals; visualize connectivity.

3.2.3 Week 6 – Local atomic environments

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

Exercise:
Compute SOAP vectors; cluster structures in environment space.

3.3 Unit III — Learning Structure–Property Relations (Weeks 7–9)

3.3.1 Week 7 – Regression and generalization in materials data

- Predicting bandgaps, elastic moduli, formation energies.
- Bias–variance and overfitting in materials datasets.
- Dataset size vs model complexity.

Exercise:
Compare linear, random forest, and NN regressors on a materials dataset.

3.3.2 Week 8 – Neural networks for materials properties

- Neural networks as universal surrogates for DFT-level properties.
- Training pitfalls: data leakage, imbalance, extrapolation.
- Physical interpretability concerns.

Exercise:
Train a small NN for property prediction; analyze overfitting.

3.3.3 Week 9 – Representation learning and feature discovery

- Learned vs engineered features.
- What networks “learn” about chemistry and structure.

- Transferability across chemical systems.

Exercise:
Compare performance using raw descriptors vs learned embeddings.

3.4 Unit IV — Latent Spaces, Uncertainty, and Discovery (Weeks 10–12)

3.4.1 Week 10 – Latent spaces of materials

- Autoencoders and embeddings for crystal data.
- Interpreting latent dimensions.
- Relation to chemical intuition and structure families.

Exercise:
Train an autoencoder; visualize latent materials space.

3.4.2 Week 11 – Clustering vs discovery in materials spaces

- Why clustering discovery.
- Structure in latent space.
- Identifying families, outliers, and anomalies.

Exercise:
Compare k-means clustering with latent-space organization.

3.4.3 Week 12 – Uncertainty-aware discovery & Gaussian Processes

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

Exercise:
GP regression vs NN ensembles; visualize uncertainty-driven screening.

3.5 Unit V — Constraints, Trust, and Synthesis (Weeks 13–14)

3.5.1 Week 13 – Physical constraints and informed learning

- Stability, charge neutrality, symmetry constraints.
- Physics-informed ML in materials discovery.
- Failure modes of unconstrained models.

Exercise:
Train a constrained model using penalty-based approaches.

3.5.2 Week 14 – Integration, limits, and outlook

- Explainability of materials ML models.
- What ML can and cannot discover.
- How computational genomics meets experiment-driven workflows.

Exercise:
Mini-project synthesis and presentation.

---

4 Learning Outcomes

Students completing this course will be able to:

- 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 to accelerate materials screening responsibly.
- Critically assess the limits of data-driven materials discovery.