Materials Genomics

Philipp Pelz

Corresponding author: Philipp Pelz,

Abstract

This course introduces students to materials genomics, treating the periodic table and all known crystal structures as a searchable, computable design space. Students learn how materials databases are built, how to represent matter as numbers, graphs, or fingerprints, how to interrogate and predict properties of solids, how to use ML as a surrogate for quantum mechanics, and how to design new materials algorithmically.

1 Course Information

4th Semester – 5 ECTS · 2h lecture + 2h exercises per week, together with ML for Materials Processing & Characterization

2 Course Philosophy

Materials genomics treats the periodic table and all known crystal structures as a giant searchable, computable design space.

Students learn:

- how materials databases are built,
- how to represent matter as numbers, graphs, or fingerprints,
- how to interrogate and predict properties of solids,
- how to use ML as a surrogate for quantum mechanics,
- how to design new materials algorithmically.

The angle is computational, structured, mathematically clean — the perfect foil to the messier experimental focus of other courses.

3 Week-by-Week Curriculum (14 weeks)

3.1 Unit I — Foundations of Materials Genomics (Weeks 1–3)

3.1.1 Week 1 – What is Materials Genomics?

- Genomics analogy: genes → functions vs atoms → properties.
- Brief history: AFLOW, OQMD, Materials Project, NOMAD.
- PSPP from the structure-first viewpoint.

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

3.1.2 Week 2 – Crystal structure fundamentals

- Space groups, Wyckoff positions, symmetry operations.
- How symmetry informs descriptors.

Exercise: Using pymatgen / spglib to analyze symmetries.

3.1.3 Week 3 – Materials databases & file formats

- CIF, POSCAR, PDB-like formats.
- Thermodynamic quantities in databases: formation energy, stability, convex hull.

Exercise: Parse CIF files, extract primitive cells, compute density.

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

3.2.1 Week 4 – Classical descriptors & materials fingerprints

- Magpie, matminer.
- Stoichiometric, elemental, and structural features.

Exercise: Build a small property regressor with Magpie features.

3.2.2 Week 5 – Graph-based representations

- Crystal structures as graphs: nodes, edges, periodic boundary conditions.
- CGCNN, MEGNet architecture intuition (no training from scratch yet).

Exercise: Build a simple CGCNN-like graph featurizer.

3.2.3 Week 6 – Local atomic environments

- Voronoi tessellations, coordination numbers, SOAP descriptors.
- Role in interatomic potentials and ML force fields.

Exercise: Compute SOAP vectors; perform clustering in descriptor space.

3.3 Unit III — High-Throughput Computation & Screening (Weeks 7–9)

3.3.1 Week 7 – Quantum mechanical data and DFT basics

- What DFT gives you: energies, forces, band structures, elastic constants.
- Why it's expensive; why ML surrogates matter.

Exercise: Run a toy DFT calculation (Quantum Espresso or MP workflows).

3.3.2 Week 8 – High-throughput workflows

- Automation: pymatgen, custodian, FireWorks, Atomate.
- Data generation for building surrogate models.

Exercise: Perform a small FireWorks workflow (or simulate the idea without cluster resources).

3.3.3 Week 9 – Phase stability & the convex hull

- Formation energies, metastability, hull distance.
- Mapping an entire chemical system.

Exercise: Reconstruct phase diagrams from Materials Project data.

3.4 Unit IV — Learning Properties from Atomic Structure (Weeks 10–12)

3.4.1 Week 10 – Regression on crystal data

- Predicting bandgaps, hardness, elastic moduli.
- Comparing different representation families.

Exercise: Benchmark random forest, GPR, CGCNN on a small dataset.

3.4.2 Week 11 – Machine-learned interatomic potentials

- Overview: GAP, SNAP, MTP, NequIP.
- Role in simulating defects, diffusion, mechanical behavior.

Exercise: Fit a tiny ML potential (ACE or simple SNAP-style) to toy data.

3.4.3 Week 12 – Generative models for materials

- VAEs, diffusion models for crystal generation.
- Constraints: symmetry, stability, charge neutrality.

Exercise: Sample a generative model from a pretrained online source; analyze validity.

3.5 Unit V — Mini-Project & Synthesis (Weeks 13–14)

3.5.1 Week 13 – Project workshop

Example projects:

- Predict bandgap from composition + structure representation.
- Identify new stable compounds in a chemical system.
- Build a graph-based model for elastic constants.
- Use ML to approximate formation energies for a ternary subsystem.
- Analyze SOAP fingerprints across polymorphs.

3.5.2 Week 14 – Presentations & Reflection

- Interpreting models: SHAP for materials descriptors.
- Strengths/limitations of materials genomics vs experiment-driven ML.

- How computational and experimental ML meet in modern labs.

4 Learning Outcomes

Students completing this course will be able to:

- Navigate major materials databases and extract relevant structural/property data.
- Represent crystals numerically using descriptors, fingerprints, and graphs.
- Train ML models to predict quantum-mechanical and thermodynamic properties.
- Analyze structural features via symmetry, coordination, and environments.
- Perform high-throughput screening of materials candidates.
- Understand and apply generative models for inorganic crystals.
- Critically evaluate ML results in computational materials discovery.