Project Checkpoint

Title: Multimodal ML Pipeline for Biocompatible Transparent Semiconductor Discovery

Group Members (Team 2):
Amala Vellappillil Biju, Brent Thorne, Dare Drouin, Jehad Saif, Michael Amaraut, Rachana Baskar

Abstract:

This project develops a multimodal machine learning pipeline to accelerate the discovery of biocompatible transparent semiconductors. By integrating structural, compositional, and electronic features from high-throughput Density Functional Theory(DFT) datasets, we aim to predict and rank candidate materials with tunable bandwidths, high optical transparency, and low biological toxicity. Our approach combines exploratory data analysis, graph-based representations, and multitask neural networks to model key properties such as band gap, dielectric constant, and formation energy. We incorporate biocompatibility screening and interpretability tools to ensure safe and actionable predictions. The pipeline supports applications in optoelectronics, biosensors, and sustainable materials, offering commercial and societal impact through reduced research and development (R&D) costs and improved material safety.

Project Description:

We are developing a modular, multimodal machine learning pipeline to accelerate the discovery of biocompatible transparent semiconductors. Our approach integrates tabular features, graph-based structural representations, and scalable inference tools to predict key material properties. The pipeline is designed to support multiple datasets of increasing complexity, from the curated NOMAD2018 dataset to the more expansive JARVIS-DFT and Open Quantum Materials Dataset (OQMD) repositories.

Key components of the pipeline include:

- Exploratory Data Analysis (EDA): Identify trends and correlations in formation energy, band gap, and compositional features
- Modeling Framework: Supports both structured feature models (e.g., feedforward neural networks) and structure-aware models that incorporate atomic geometry (e.g., graph neural networks). The pipeline is modular and extensible, allowing exploration of additional architectures such as ensemble methods, attention-based models, or possibly physics-informed networks.
- Flexible Inference: Command Line Interface-driven (CLI-driven) prediction with support for checkpointing, batch processing, and .CSV output
- Scoring Framework: Materials can be ranked using a composite metric that balances transparency, conductivity, and biocompatibility (planned)
- Scalable Integration: Dockerized MySQL + NVMe-backed infrastructure enables high-throughput screening of OQMD-scale datasets

Problem Statement and Scope:

Transparent semiconductors are critical for next-generation technologies in optoelectronics, biosensing, and sustainable energy. However, discovering new candidates remains computationally expensive and often neglects biocompatibility, a key requirement for medical and environmental applications.

Our goals are to:

- Model materials with tunable electronic bandwidths in the 2.5-4.0 eV range, suitable for transparent semiconductors
- Screen for thermodynamic stability using formation energy and phase competition metrics
- Incorporate biocompatibility proxies to prioritize safe, non-toxic materials
- Reduce the DFT search space by filtering large materials databases with fast, interpretable Machine Learning (ML) models
- This pipeline is designed to evolve with the data: starting with NOMAD for rapid prototyping, scaling to JARVIS for richer property prediction, and ultimately leveraging OQMD for compositional diversity and industrial relevance.

Project Status and Next Steps:

Our project has progressed through a structured, data-driven pipeline, beginning with the NOMAD2018 dataset. This dataset offered a clean, well-defined modeling task with a limited number of features and a focused chemical space: ternary oxides of Al, Ga, and In. We successfully implemented and tested two models on this dataset: a tabular-only MLP and a hybrid model combining tabular and graph-based features. Both models are fully integrated into a modular CLI, with support for training, inference, checkpointing, and reproducible output.

As we transition to the JARVIS-DFT dataset, the complexity increases. JARVIS introduces a broader range of materials, more diverse property targets (e.g., multiple band gap estimations, dielectric constants), and a richer feature space. This opens the door to more selective filtering, such as isolating transparent semiconductors, and more complex modeling strategies. While the dataset is still manageable in size, it demands more careful preprocessing and feature engineering to align with our multimodal pipeline.

The final stage involves integrating the OQMD, which is significantly larger and more heterogeneous. With over 1.3 million entries and tens of millions of atomic and structural records, OQMD introduces challenges in data access, filtering, and computational efficiency. To address this, we've deployed a Dockerized MySQL instance backed by NVMe storage, enabling high-throughput queries and scalable screening. Our next steps include extracting stable, low-energy candidates from OQMD, incorporating toxicity proxies, and unifying all datasets under a common scoring framework that balances transparency, conductivity, and biocompatibility.

Model Development Status:

We have implemented two core models for property prediction. The first, NomadMLP, is a feedforward neural network that operates on structured features such as composition and lattice parameters. The second, HybridModel, extends this by incorporating atomic structure through a graph neural network, enabling it to learn from both compositional and spatial information.

A unified command-line interface supports both training and inference for these models. It handles argument parsing, checkpoint saving and loading, device selection (CPU or GPU), and outputs predictions in CSV format. This interface has been fully tested using the NOMAD2018 dataset.

The inference pipeline performs batch prediction with progress tracking and generates CSV files containing predicted formation energy and band gap values. It is compatible with both the MLP and hybrid models and is designed to scale to larger datasets such as JARVIS and OQMD.

Appendix:

Data Sources and Details:

NOMAD2018:

- Materials:
  - Aluminum (Al0, Gallium (Ga), Indium (In) composites
  - ~3,000 compounds of general form $(Al_xGa_yIn_z)_2O_3$
  - Composition x + y + z = 1
- Purpose:
  - Display ML models for predicting material properties from structural and composition data
- Source of Data:
  - Material properties were calculated using Density Functional Theory (DFT) methods
- Data Fields:
  - *id, spacegroup, number_of_total_atoms, percent_atom_al, percent_atom_ga, percent_atom_in, lattice_vector_1_ang, lattice_vector_2_ang*, *lattice_vector_3_ang, lattice_angle_alpha_degree, lattice_angle_beta_degree, lattice_angle_gamma_degree, formation_energy_ev_natom,* and *bandgap_energy_ev* and *latt_ang*.
- Data Splits:
  - Training set: 2,400 compounds
  - Test set: 600 compounds
  - Data Shape: (2399, 14)
- Status:
  - EDA Complete
  - Code Complete
    - Training w/checkpoint
    - Inference and Prediction
- Next steps
  - Final Analysis
  - Add models

JARVIS-DFT:

- Materials: Inorganic semiconductors
- Purpose: Predict band gaps, formation energies, and dielectric constants
- Source of Data: DFT-calculated properties from JARVIS database
- Data Fields: ~64 features including composition, structure, and electronic properties
- Filtered Subsets:
  - Semiconductors:
    - semis_vdw.csv: 8,520 rows
    - semis_mbj.csv: 2,828 rows
    - semis_hse.csv: 12 rows
  - Transparent Semiconductors:
    - transparent_semis_vdw.csv: 3,185 rows

        - transparent_semis_mbj.csv: 1,879 rows
        - transparent_semis_hse.csv: 10 rows
- Next Steps:
    - Port data featurizer and model from NOMAD model
    - Add toxicity proxies and biocompatibility heuristics for final ranking

The Open Quantum Database (OQMD):

- Materials: Inorganic compounds spanning binary to complex multi-element systems
- Purpose: Large-scale compositional and thermodynamic screening for materials discovery
- Source of Data: DFT calculations, primarily from the Inorganic Crystal Structure Database (ICSD) Data Fields:
- Data:
    - Scale:
        - entries: 1,317,701 total records
        - atoms: ~74 million atomic site records
        - structures: ~7.4 million unique crystal structures
        - formation_energies: ~2.1 million computed formation energy records
    - Key Fields (from entries table):
        - delta_e: Formation energy relative to competing phases
        - stability: Thermodynamic stability metric
        - composition_id: Chemical formula (e.g., AgCu, O_2Si)
        - ntypes: Number of unique atomic species
        - natoms: Total atoms in the unit cell
    - Insights:
        - Average formation energy (delta_e) decreases with increasing chemical complexity (ntypes)
        - Most frequent compositions include Ag-Cu alloys and common semiconductors like SiO_2 and ZnS
    - Status:
        - Database Engine:
            - MySQL 8.0 deployed in a Docker container for portability and reproducibility
        - Storage:
            - Backed by a high-throughput NVMe SSD to support fast I/O for large-scale queries (e.g., 70M+ rows in atoms, 60M+ in structures_species_set)
        - Schema Access:
            - Exposed via qmdb schema with indexed tables for entries, atoms, structures, formation_energies, and compositions
        - Query Optimization:
            - Indexed composition_id, delta_e, and ntypes for fast filtering
            - Aggregation queries (e.g., average delta_e by ntypes) complete in <1s
        - Proposed Usage:

        - Used for compositional frequency analysis, thermodynamic screening, and prototype matching
        - Enables filtering of stable, low-energy candidates for downstream ML scoring
    - Next Steps:
        - Integrate filtered OQMD candidates into the multimodal pipeline
        - Add toxicity proxies and biocompatibility heuristics for final ranking

Exploratory Data Analysis (EDA):

NOMAD2018:

Q1: How stable is the material?

- Feature: formation_energy_ev_natom
- Interpretation: Measures the energy change when forming a compound from its elements.
- Insight: Most stable compounds cluster around 0.0-0.2 eV/atom. Values near 0 indicate high thermodynamic stability.
- Visualization: Distribution of formation energy shows a skew toward low-energy, stable materials.
- Use: Filter out unstable candidates before modeling.

Q2: How transparent is the material?

- Feature: bandgap_energy_ev
- Interpretation: Band gap determines optical transparency and conductivity.
- Insight:
    - <1.5 eV => opaque (e.g., Si, metals)
    - 2.5-4.0 eV => semiconductive and potentially transparent
    - 4.0 eV => insulating, often non-conductive
- Use: Identify candidates suitable for optoelectronic applications.

Q3: Does the material exhibit semiconductive behavior?

- Derived from: Band gap range and formation energy
- Interpretation: Materials with band gaps between ~1.5-4.0 eV and low formation energy are likely semiconductors.
- Use: Classify materials for downstream conductivity modeling.

JARVIS-DFT:

Q1: What is the predicted band gap across different DFT methods?

- Features: optb88vd, mbj, hse band gap estimates
- Insight:
    - Different methods yield varying band gap values;
    - MBJ and HSE tend to predict higher gaps.
- Use: Cross-validate transparency predictions and filter for consistent semiconductors.

Q2: How does dielectric constant vary across materials?

- Feature: epsx, epsz, or averaged dielectric constants
- Interpretation: High dielectric constants may indicate strong polarization and suitability for sensors.

- Use: Rank materials for biosensing and optoelectronic applications.

Q3: Are there correlations between formation energy and band gap?
- Insight:
  - Preliminary analysis suggests weak correlation;
  - stable materials span a wide band gap range.
- Use: Supports independent modeling of stability and transparency.

Q4: Can we isolate biocompatible candidates?
- Approach: Filter out toxic elements (e.g., heavy metals) and prioritize known safe compositions.
- Use: Build a biocompatibility proxy for final scoring.

Open Quantum Materials Database (OQMD):
Q1: What compositions dominate the dataset?
- Feature: composition_id
- Insight: Ag-Cu alloys and common semiconductors (e.g., SiO_2, ZnS) are highly represented.
- Use: Prioritize underexplored compositions for novelty.

Q2: How does formation energy vary with chemical complexity?
- Feature: delta_e, grouped by ntypes
- Insight:
  - Simple binaries (1-2 types) tend to have higher formation energies
  - Complex systems (>=3 types) show lower average delta_e, suggesting more stable configurations
- Use: Target multi-element systems for stability and diversity.

Q3: What is the distribution of atomic site counts?
- Feature: natoms, atoms table
- Insight: Wide range of unit cell sizes; large structures may require graph-based modeling
- Use: Guide model selection (MLP vs GNN) based on structural complexity.

Q4: How do prototypes and spacegroups relate to stability?
- Feature: prototype_id, spacegroup
- Insight: Certain prototypes consistently yield low-energy configurations
- Use: Incorporate prototype priors into model features or filtering logic.

Model Usage Examples:

Example Training with Checkpointing:

```
brent@ml-team2:~/src/ml-final-project-team2/code$ python -m
nomad_hybrid.cli   --csv /shared/data/nomad2018/train.csv   --xyz_dir
/shared/data/nomad2018/train   --model mlp   --epochs 1   --save_path
/shared/data/checkpoints/mlp.pt
Loading data from: /shared/data/nomad2018/train.csv
Geometry directory: /shared/data/nomad2018/train
Model type: mlp
Training for 1 epochs on CUDA
Epoch 01 | Train Loss: 1.3811 | Val Loss: 0.6407 (New best model)
Saving best model (val loss 0.6407) to:
/shared/data/checkpoints/mlp.pt

brent@ml-team2:~/src/ml-final-project-team2/code$ python -m
nomad_hybrid.cli   --csv /shared/data/nomad2018/train.csv   --xyz_dir
/shared/data/nomad2018/train   --model mlp   --epochs 1
0   --save_path /shared/data/checkpoints/mlp.pt --load_path
/shared/data/checkpoints/mlp.pt
Loading data from: /shared/data/nomad2018/train.csv
Geometry directory: /shared/data/nomad2018/train
Model type: mlp
Training for 10 epochs on CUDA
Loading model weights from /shared/data/checkpoints/mlp.pt
Epoch 01 | Train Loss: 0.1014 | Val Loss: 0.0706 (New best model)
Epoch 02 | Train Loss: 0.0754 | Val Loss: 0.0616 (New best model)
Epoch 03 | Train Loss: 0.0648 | Val Loss: 0.0522 (New best model)
...
Saving best model (val loss 0.0267) to:
/shared/data/checkpoints/mlp.pt
```

Example Inference:

```
brent@ml-team2:~/src/ml-final-project-team2/code$ python -m
nomad_hybrid.cli \
  --predict \
  --test_csv /shared/data/nomad2018/test.csv \
  --xyz_dir /shared/data/nomad2018/test \
  --model mlp \
  --load_path /shared/data/checkpoints/mlp.pt \
  --output predictions_mlp.csv
Loading model weights from: /shared/data/checkpoints/mlp.pt
🔍 Predicting: 100%|████████████████| 19/19 [00:02<00:00,  7.08it/s]
Predictions saved to: predictions_mlp.csv
brent@ml-team2:~/src/ml-final-project-team2/code$ head
predictions.csv
id,formation_energy_ev_natom,bandgap_energy_ev
1,0.18480927,1.5783442
2,0.060803212,3.7262857
3,0.14032331,3.3954437
4,0.027698196,3.0036652
5,0.12271825,1.7080216
6,0.0347682,4.2817163
7,0.056691565,3.2069077
8,0.1164797,2.092504
9,0.056368865,2.6976058
```