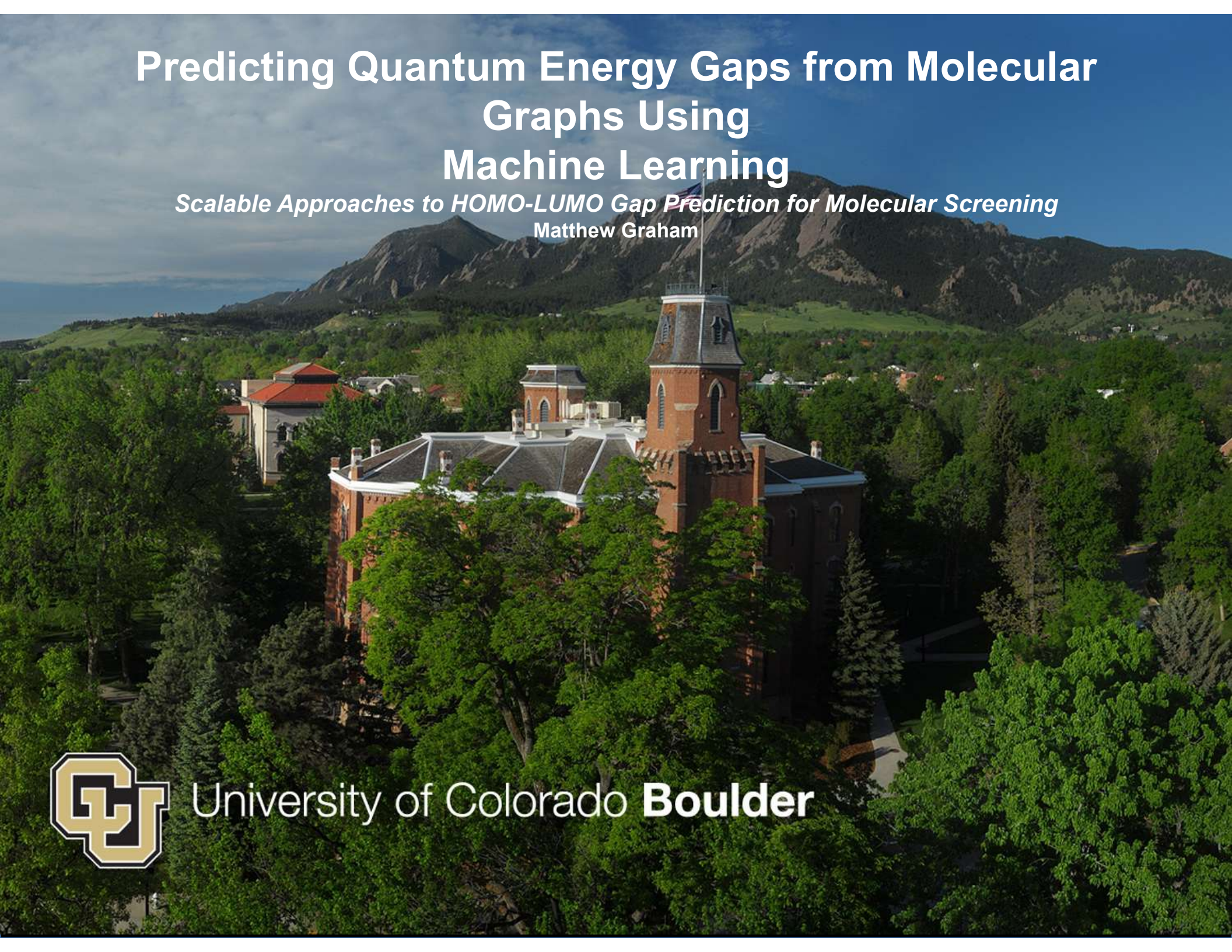
Predicting Quantum Energy Gaps from Molecular Graphs Using Machine Learning

Scalable Approaches to HOMO-LUMO Gap Prediction for Molecular Screening

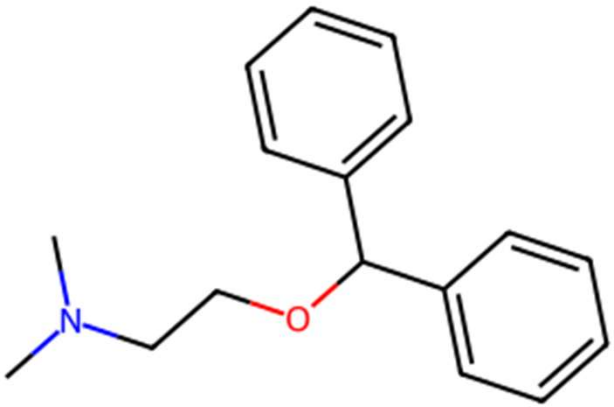
Matthew Graham

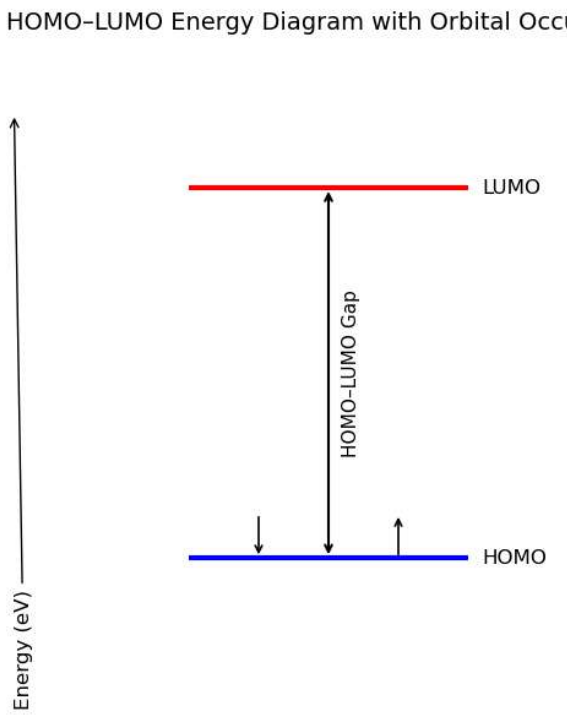


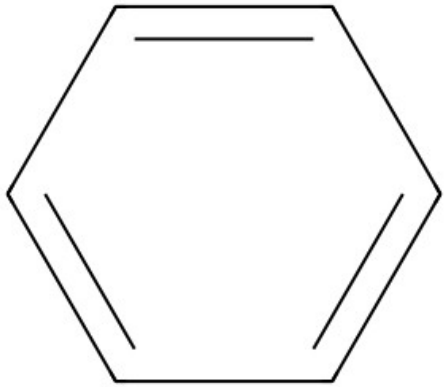
University of Colorado **Boulder**

Problem Statement & Importance

- Accurate prediction of quantum chemical properties is crucial for materials science, drug discovery, and electronics.
- The size of the HOMO-LUMO gap provides insights into molecular reactivity and stability.
- While accurate, Density Functional Theory (DFT) is computationally expensive for large-scale screening.

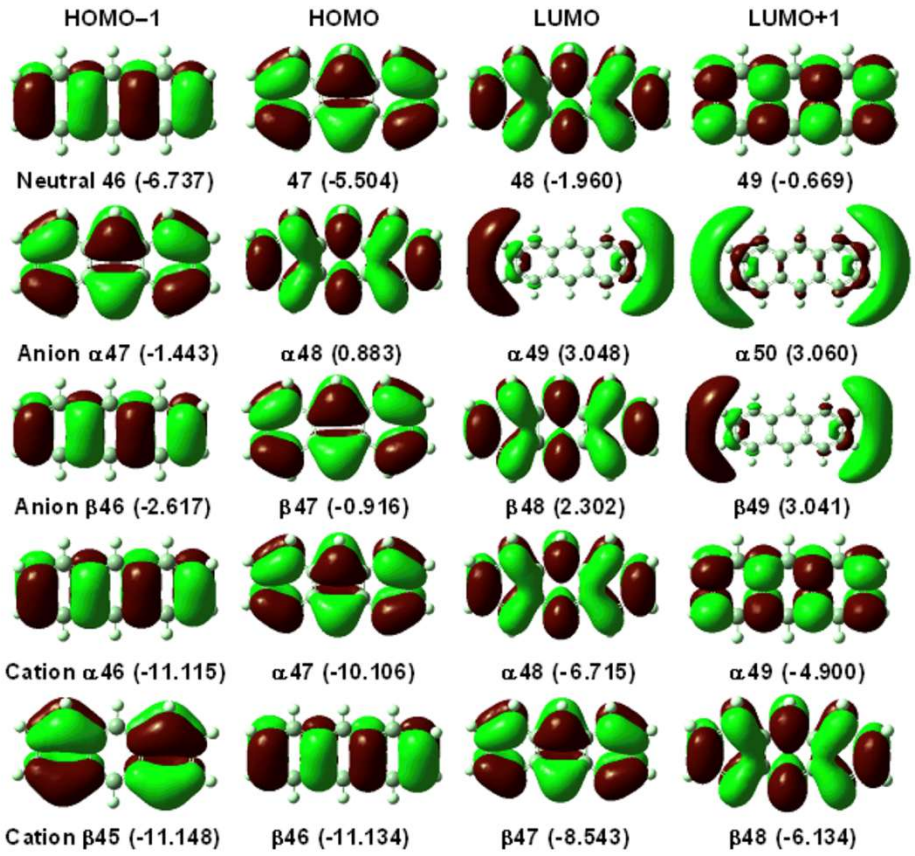






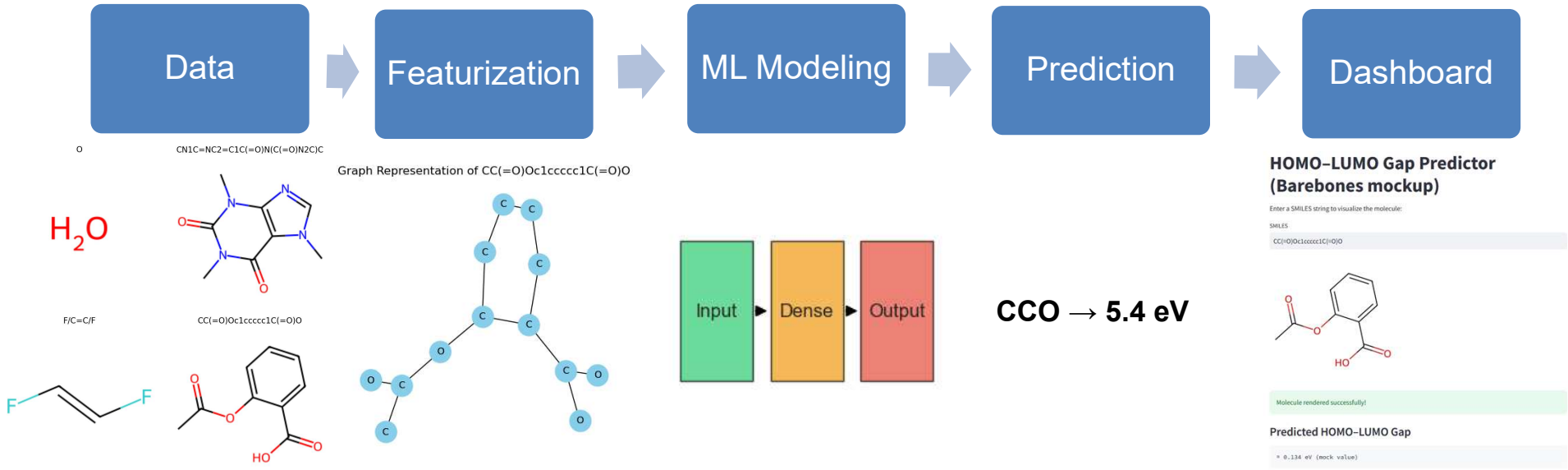
Related Work & Gaps

- **Related Work:**
  - Accurate QM methods (DFT) are computationally expensive for large scale screening.
  - Machine Learning (GNN, kernel method) shows promise but often requires large models and high-performance hardware.
- **Gap:** Need for fast, lightweight, and potentially interpretable models for broader accessibility and local use.

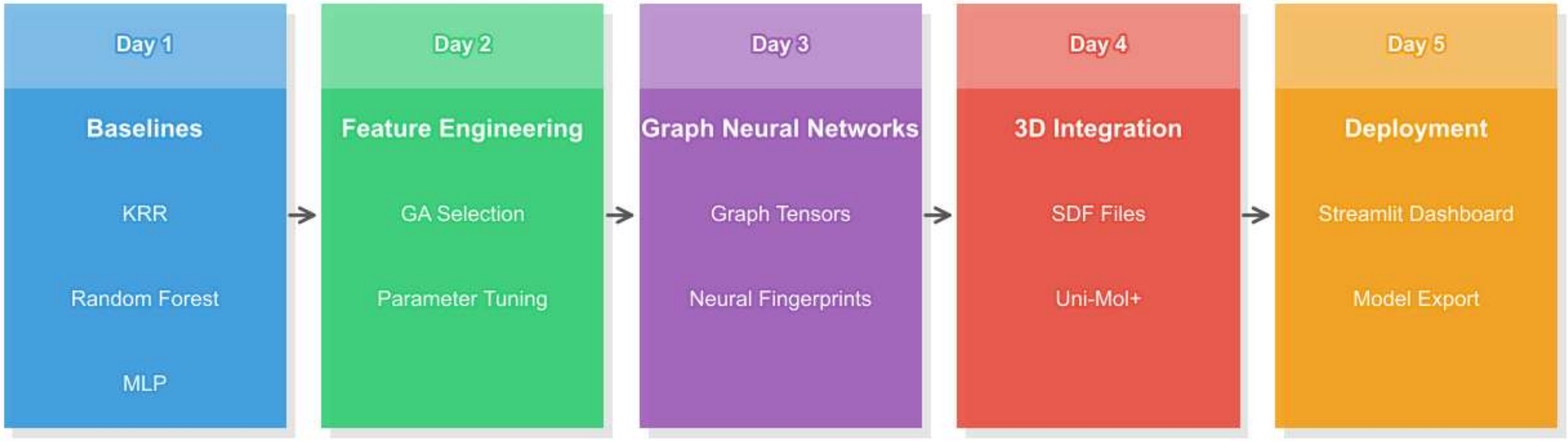


Proposed Approach & Tools

- **Dataset:** OGB PCQM4Mv2 (>3.7M DFT-calculated HOMO-LUMO gaps)
- **Tools:** Python (RDKit, ChemML, Scikit-learn, TensorFlow, PyTorch, Streamlit)
- **Molecular Representations:**
  - 1D: SMILES
  - 2D: Graph-based (ChemML & OGB), Coulomb Matrix, Fingerprints
  - 3D: Coordinates & Distances
  - Global Features (RDKit)
- **Models:** Baseline (KRR, RF, MLP), GNNs (2D & 3D Hybrid)

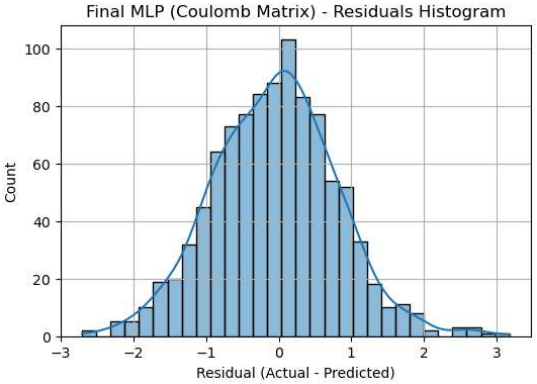


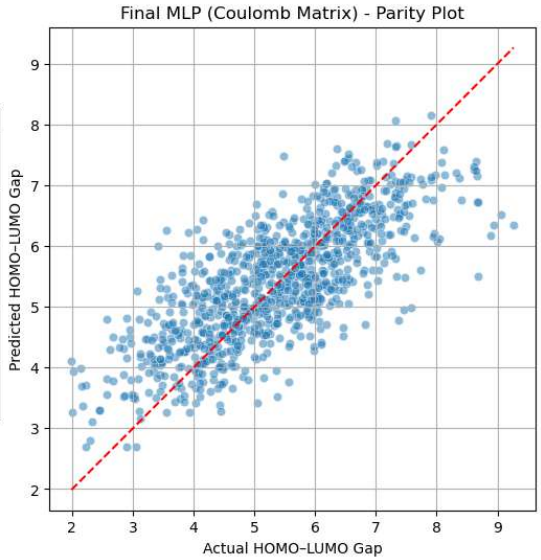
Task Phases


Day 1
Baselines
KRR
Random Forest
MLP
Day 2
Feature Engineering
GA Selection
Parameter Tuning
Day 3
Graph Neural Networks
Graph Tensors
Neural Fingerprints
Day 4
3D Integration
SDF Files
Uni-Mol+
Day 5
Deployment
Streamlit Dashboard
Model Export

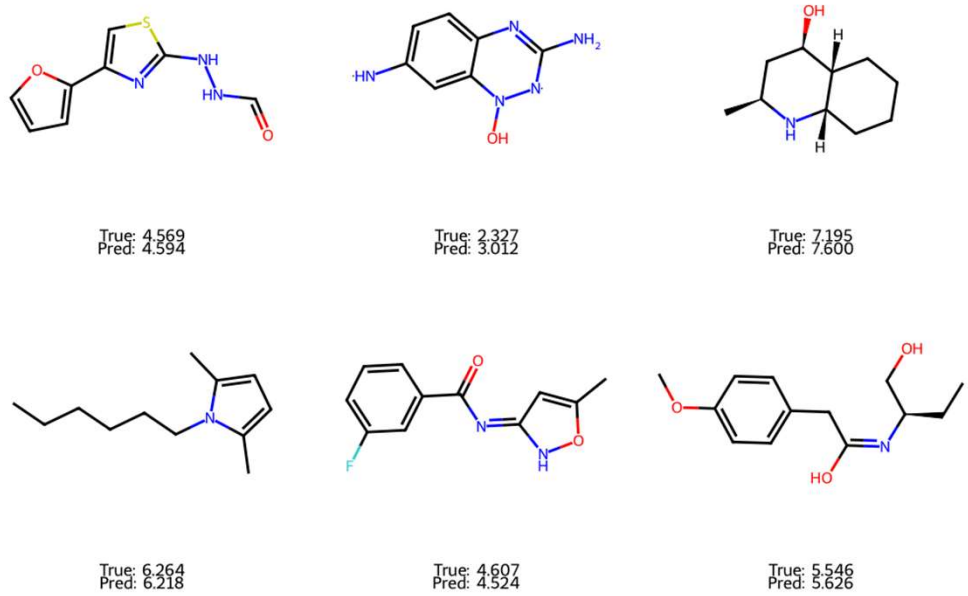
Evaluation Plan

- **Key Metric:** MAE (Target ~0.15 eV)
- **Other Metrics:** RMSE, R^2
- **Visualizations:** Parity and Residual plots, Predicted vs. True Labels





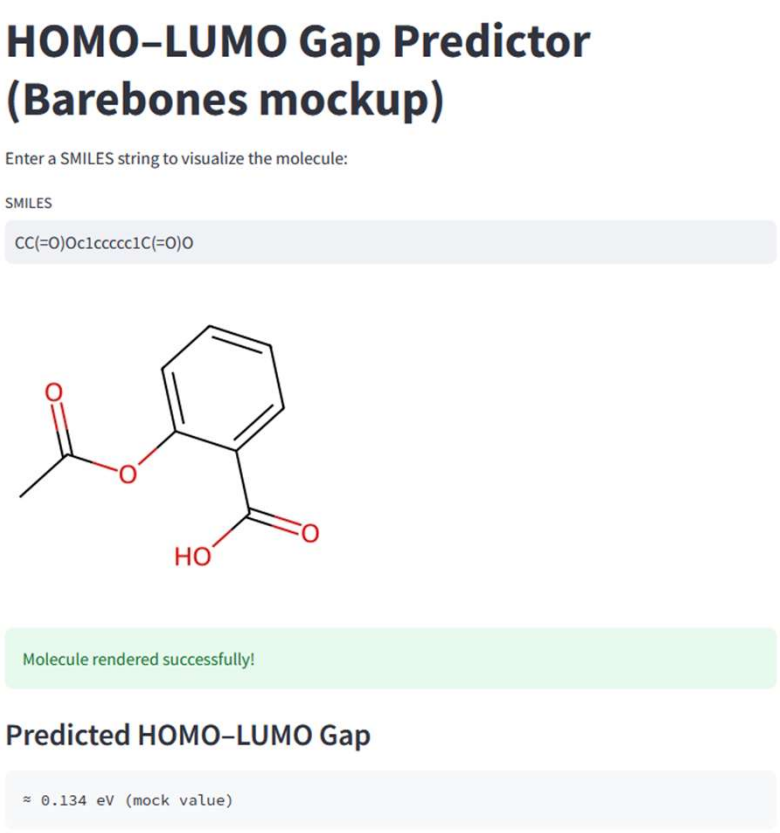
Molecule Label Plots (Predicted vs. True)



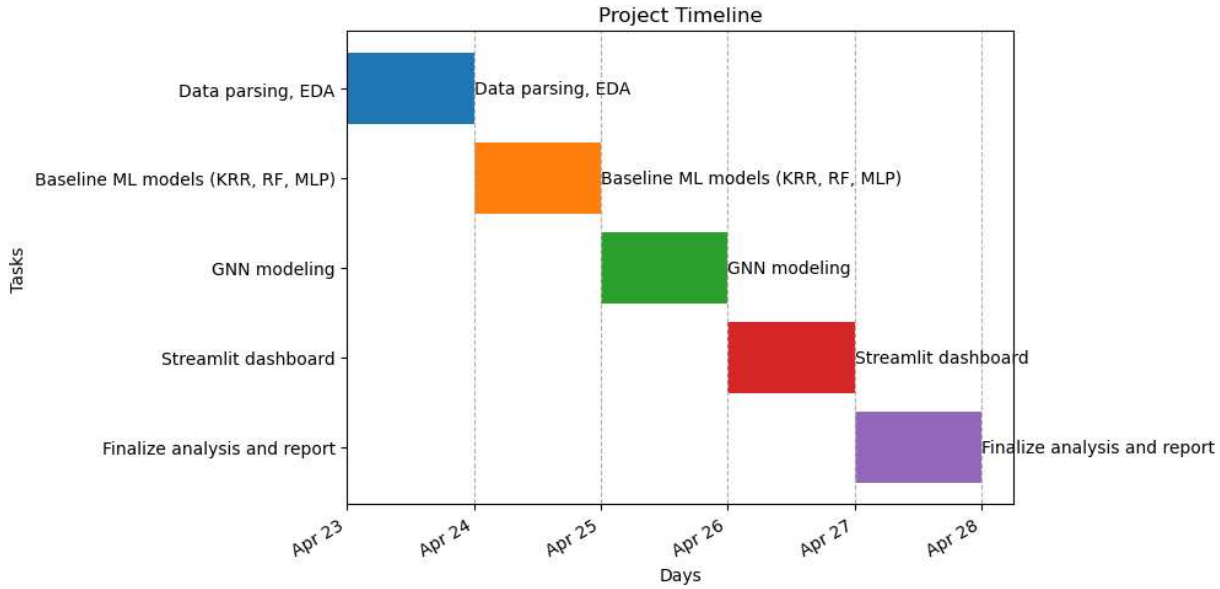

```
=== Final Validation Set Evaluation ===
        MAE      RMSE  r_squared
0  0.158608  0.233842   0.965001
```

Interactive Streamlit Dashboard

- Web app with SMILES input and model prediction
- **Output:** HOMO-LUMO gap and molecular visualization
- **Tools:** Streamlit, SQLite/CSV query



Current Status & Timeline



Current Status:

- Baseline models trained and tuned
- Final 2D Hybrid GNN trained and tuned-achieved .15 eV MAE goal
- Dashboard barebones created, but needs to have model hosted on HuggingFace and provide real time inference
- 3D model being explored-facing memory issues, even with reduced dataset size. Exploring LMDB for efficient handling of 3D graph data.
- On track to meet minimum viable deliverables

Challenges, Backup Plans, Broader Impact

Challenges:

- Challenge: Dataset size → use subsets, explore efficient ways to store 3D graph data
- Challenge: GNN complexity → fallback to MLP
- Challenge: Dashboard latency → limit predictions

Broader Impact:

- Scalable virtual screening
- Reduces cost and time in materials & drug discovery
- Educational and research tool for chemical property prediction