Project/Software Design Document 1.0

ERAU Graph-Theory Project

1. Project Overview

This project aims to create a graph-based materials database and accompanying AI models that:

1. Represent molecular structures as graphs (nodes as atoms, edges as bonds).
2. Predict how materials evolve over time (e.g., graphite to graphene).
3. Discover hidden correlations across diverse materials (metal behaving like rubber, etc.).
4. Provide a user-facing query interface that delivers high-quality, application-specific recommendations.

2. Goals and Objectives

1. **Database Construction**
    - Use Neo4j to store molecular and material data in a graph-based database
    - Attached metadata for bond strength and properties (Also processing methods, historical data, etc. as a stretch goal)

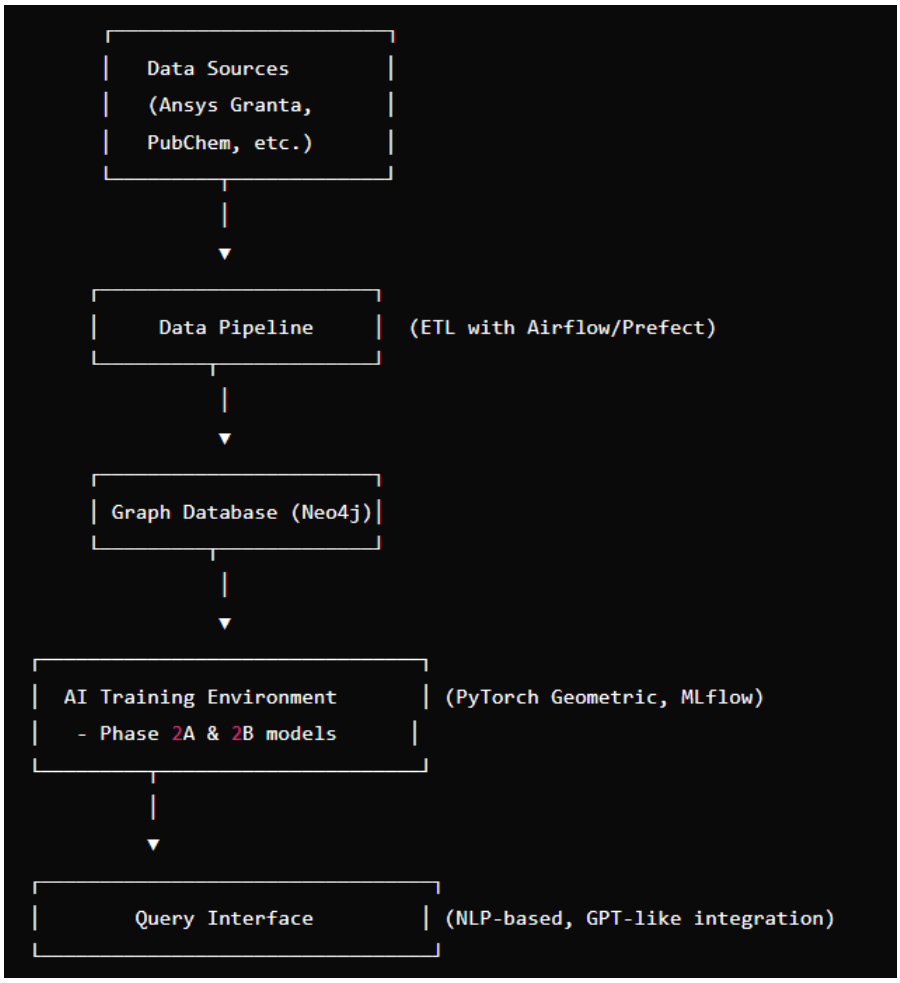
2. **AI Model Development**
    - **Evolution of Materials Model (Phase 2A):** Predict how simpler materials transform into more advanced counterparts.
    - **Hidden Connections Model (Phase 2B):** Identify unexpected relationships and correlations in materials.

3. **User Query Interface**
    - A natural language question-and-answer system that retrieves and reasons over the materials database.

3. System Requirements

- **Data Ingestion & Integration**
    - Pull data from Ansys Granta and other sources (PubChem, Materials Project) using APIs or CSV exports.
    - Regularly scheduled ETL processes for up-to-date info.
- **Graph Database**
    - Store molecular structures, bonding information, and relevant material properties.
    - Must handle node/edge indexing and advanced graph queries.
- **Machine Learning Infrastructure**
    - Must support training of Graph Neural Networks (GNNs), sequence models, and correlation-finding models.
    - Integration with experiment tracking (MLflow) and version control.
- **User Interface**
    - Intuitive UI for domain experts and non-experts to query and get recommendations.
    - Natural language processing for ease of data exploration.

4. High-Level Architecture



5. Implementation Plan

Phase 1: Database and Training Data Creation

1. **Data Ingestion and Cleaning**
    - **Tools:** Python (Pandas), RDKit for molecular descriptor calculation.
    - **Process:**
        - Ingest from Ansys Granta (properties, chemical composition).
        - Use scripts to parse, clean, and standardize data (SMILES, InChI, etc.).
2. **Graph Construction**
    - **Tools:** RDKit to extract bonding info, Neo4j for storage.

- **Process:**
  - Convert each molecule to a graph representation.
  - Store node (atom) attributes (atomic number, electronegativity) and edge (bond) attributes (single/double/triple).
  - Associate each graph with material property metadata.

3. **ETL Pipeline Setup**

   - **Tools:** Apache Airflow or Prefect.
   - **Process:**
     - Automate regular data imports/updates.
     - Validate data integrity and track changes over time.

Phase 2: Model Development

1. **Core GNN for Property Prediction**

   - **Tools:** PyTorch Geometric / TensorFlow GNN.
   - **Process:**
     - Train baseline GCN, GAT, or GraphSAGE models on the newly created dataset.
     - Evaluate performance with standard metrics (RMSE, MAE, etc.).

2. **Model Versioning and Experiment Tracking**

   - **Tools:** MLflow.
   - **Process:**
     - Track hyperparameters, datasets, and outcomes.
     - Store best-performing models for production inference.

Phase 2A: Evolution of Materials Model

1. **Data Preparation for Evolution**

   - **Tools:** Matminer or custom scripts.
   - **Process:**
     - Gather time-series or historical data of how certain materials changed (e.g., processing steps from graphite to graphene).
     - Represent “state transitions” in the graph database.

2. **Model Development**

   - **Tools:** Sequence-to-sequence Transformers (PyTorch).
   - **Process:**

        - Model evolves a base material (input) to a target advanced material (output).
        - Incorporate domain knowledge (pressure, temperature processes) as additional features.

3. **Complex Material Evolution**

    - **Tools:** Temporal Graph Networks or extended Transformer architecture.
    - **Process:**
        - Handle multi-step transitions for composites (e.g., from “low-tier carbon composites” to “high-performance CFRP”).
        - Validate predictions against known advancements in materials science.

Phase 2B: Hidden Connections Materials Model

1. **Correlation & Pattern Mining**

    - **Tools:** Autoencoders, Contrastive Learning, or Graph Attention Networks.
    - **Process:**
        - Identify latent spaces where unexpected similarities (e.g., metal vs. rubber) are revealed.
        - Use unsupervised methods to cluster or link materials based on functional behavior.

2. **Cross-Validation**

    - **Tools:** Clustering metrics (e.g., Silhouette Score).
    - **Process:**
        - Validate whether discovered connections align with known phenomena or highlight new research directions.

Phase 3: Query Interface / Recommendation Engine

1. **NLP-Based Query System**

    - **Tools:** GPT-style LLMs (OpenAI API or local large language model).
    - **Process:**
        - Fine-tune on domain-specific text and knowledge base.
        - Convert user queries to Cypher queries on Neo4j.
        - Return processed, user-friendly results with context.

2. **Recommendation Engine**

    - **Tools:** Knowledge Graph-based recommendation or collaborative filtering.

- **Process:**
  - Suggest materials or potential evolutions to meet user criteria (strength, cost, process constraints).
  - Incorporate feedback loops to refine recommendations.

6. Timeline & Milestones

1. **Month 1–4: Phase 1 Completion**
   - Database schema finalized, ETL pipeline operational, initial data loaded.
2. **Month 1–5: Phase 2 (Core GNN Model)**
   - Baseline property-prediction model trained and validated.
3. **Month 1–4: Phase 2A (Material Evolution)**
   - Prototype evolution model for simpler materials completed and tested.
4. **Month 5–8: Phase 2A (Complex Material Evolution)**
   - Evolution model extended to handle multi-component materials.
5. **Month 1–4: Phase 2B (Hidden Connections)**
   - Hidden correlation model trained, results validated.
6. **Month 5–6: Phase 3 (User Interface)**
   - NLP-based query system and recommendation engine deployed.

7. Next Steps

- Expand the dataset with more specialized materials (e.g., biopolymers).
- Integrate real-time updates from lab experiments for model retraining.
- Explore advanced neural architectures (graph transformers, hypergraph networks) for improved performance.