# Embry-Riddle Aeronautical University
# Department of Electrical Engineering and Computer Science

## GRADUATE PROJECT PROPOSAL

*A Novel Graph-Theoretic Approach to Modeling Polygonal Molecular Structures for Machine Learning Applications in Material Science*

*Version:* 1.0 *Date:* 12/9/2024

*Course Number:* SYS 690 *Credit Hours:* 3 *Semester:* SPR25

*Author:* Michael Allen *Student ID:* 2570464 *Student Program:* Sys-Eng

*Student email:* Allenm49@my.erau.edu

***Abstract:***

This research proposes the construction of a database, that will be used for Machine Learning, based on **graph-theoretic framework using line-node processes to systematically model the polygonal shapes inherent in molecular structures**. For this database, nodes represent atoms, and edges represent bonds, forming polygonal graphs that accurately reflect molecular geometries. Using existing material databases, such as Ansys Granta, graphical representations for a set of materials will be stored in a manner specially designed for compatibility with machine learning and artificial intelligence models.

| | Name | Signature | Date |
|---|---|---|---|
| **Student** | Michael Allen | Michael Allen | 12/9/2024 |
| **GRP Advisor** | Dr. M. Ilhan Akbas | | |
| **Program Coordinator / Department Chair** | Dr. Richard S. Stansbury | | |

| Date | Version | Description |
|---|---|---|
| 12/9/2024 | 1.0 | Initial Creation |
| | | |
| | | |

# OBJECTIVE

The primary objective of this research project is **to create a comprehensive database of molecular shape graphs**, which can then be used as input data for various ML/AI algorithms. Future use of the algorithms will, in turn, be used to predict novel material properties or identify materials suitable for specific applications. This approach streamlines the process of converting raw molecular data into a form that is both computationally efficient and rich in structural information, facilitating breakthroughs in material design and discovery.

# PROBLEM

Prediction of molecular properties from atomic structure is an integral part of modern material design and discovery. However, the prevailing methods usually suffer from incomplete or fragmented data because experiments in the lab are prohibitively expensive and time-consuming. Material databases, such as Ansys Granta, house extensive chemical and structural information, but they are not inherently designed to be compatible with modern AI models requiring structured, graph-like representations of data. That translation challenge involves taking the structures' complexity into a machine-readable form apt for the training of AI models. Common data formats of molecules, either through SMILES or InChI, offer linear representations not having the spatial and bonding relationships directly encoded. This means that AI models, which are trained on existing molecular data formats, are under-equipped to capture the true geometric and chemical context of structure representations, which negatively reflects in the predictive accuracy of these models.

This project will close this gap by developing a domain-specific database hinged on a graph-theoretic framework where nodes are atoms and edges are chemical bonds. It retains spatial and bonding relationships of the molecular geometries, enabling more precise AI-based prediction of materials properties. Coupled with this database and complete materials data from Ansys Granta, the system will be an end-to-end material finding platform based on data, thus driving innovation faster in industries like aerospace, energy storage, and pharmaceuticals. The solution is important; once developed, it will provide completely new materials with specific characteristics, reducing R&D by many times. This will further reduce the cost by cutting down on experimental trials and minimizing material selection for practical applications, thus encouraging innovation in many scientific and industrial fields.

# METHODOLOGY

**1. Data Collection and Preprocessing**

**1.1 Source Integration**

- **Material Database Integration:**
  - Use Ansys Granta as the primary material property database.
  - Extract chemical, structural, and mechanical properties.
  - Establish an automated data pipeline connecting Ansys Granta to the proposed database using APIs or custom ETL processes.

**1.2 Data Standardization**

- **Atom/Bond Representation:**

    - Use SMILES and InChI molecular formats for initial input.
    - Standardize molecular structure data with open-source tools like RDKit for chemical informatics.

**2. Graph Construction Framework**

**2.1 Graph Representation**

- **Nodes:** Atoms characterized by chemical and structural properties (atomic number, valency, electronegativity).
- **Edges:** Bonds between atoms, described by bond type (e.g. single, double, triple, ionic, hydrogen, metallic…) and length.

**2.2 Polygonal Graph Conversion**

- **Structure Encoding:**
    - Convert molecular structures into graph objects with polygonal subgraphs using:
        - Adjacency matrices
        - Node and edge feature vectors
    - Use efficient data models like NetworkX or DGL (Deep Graph Library).

**3. Database Architecture Design**

**3.1 Data Storage Backend**

- **Best Possible DBMS:**
    - **Neo4j (Graph Database):** For highly interconnected data.
    - **PostgreSQL + PostGIS Extension:** For spatial and topological queries.
    - **ArangoDB (Multi-Model DB):** If graph, document, and key-value storage are needed.

**3.2 Data Schema**

- **Molecular Graph Data Model:**
    - Tables/collections for:
        - Nodes (Atoms)
        - Edges (Bonds)
        - Material Properties (from Ansys Granta)

**3.3 Graph Indexing and Queries**

- **Indexing:**
  - Use spatial and graph-based indexing for fast molecular lookup.
  - Implement efficient pathfinding algorithms for structure-property queries.

## 4. AI Model Compatibility and Training Pipeline

### 4.1 Data Pipeline for AI Models

- **Data Conversion:**
  - Use PyTorch Geometric, TensorFlow, or DGL for loading graph structures as tensors.
  - Implement data loaders that interface with the database and transform the data into ML-ready formats.

### 4.2 Feature Engineering

- **Descriptors & Features:**
  - Generate molecular descriptors like molecular weight, electronegativity sum, polarizability, and bond angles.
  - Use automated feature selection techniques.

### 4.3 Model Training & Evaluation

- **Model Types (For Testing):**
  - Graph Neural Networks (GNNs)
  - Graph Attention Networks (GATs)
  - Graph Convolutional Networks (GCNs)

## 5. Workflow Automation and Scalability

### 5.1 CI/CD Integration

- Use GitHub Actions, Docker, and Kubernetes for continuous integration, testing, and deployment.

## 6. Validation

- **Validation Metrics:**
  - Molecular property prediction accuracy
  - Database query speed and reliability

Table: Project deliverables and timeline

| # | Deliverables | Date |
|---|---|---|
| 1 | Project Proposal Finalization - Define project scope, objectives, and roles | January 10 |
| 2 | Database Requirements Document - Identify functional and technical requirements | January 15 |
| 3 | Data Model Design - Create ER diagrams, schema design, and normalization | February 25 |
| 4 | Database Setup & Configuration - Establish database environment and tools | February 15 |
| 5 | Data Import & Preprocessing - Load sample data, ensure correct data types | February 20 |
| 6 | Basic Algorithm Integration - Implement test algorithms for initial data queries | February 30 |
| 7 | Initial Database Testing - Conduct CRUD operations and integrity checks | March 1 |
| 8 | Feedback & Iteration Plan - Collect feedback and define iteration goals | March 1 |
| 9 | Mid-Semester Progress Review - Present findings and challenges | March 20 |
| 10 | Advanced Data Queries & Reports - Implement and test complex queries | March 25 |
| 11 | Comprehensive Testing & Debugging - Conduct full testing, fix issues | April 5 |
| 12 | Final Report Drafting - Write the first draft of the project report | April 20 |
| 13 | Presentation Preparation - Create slides and rehearse the presentation | May 20 |
| 14 | Final Project Submission & Presentation - Submit all deliverables and present | May 30 |

## REFERENCES

1. Weng, M., Wang, Z., Qian, G. *et al.* Identify crystal structures by a new paradigm based on graph theory for building materials big data. *Sci. China Chem.* **62**, 982–986 (2019). https://doi.org/10.1007/s11426-019-9502-5

2. Goetz, K. R., & Mack, T. (2023). Artificial intelligence in materials chemistry. Wiley Interdisciplinary Reviews: Computational Molecular Science, 13(5), e1729. https://doi.org/10.1002/wcms.1729 [Accessed 1 October 2024].

3. Zhou, S., Yu, D., & Zhang, M. (2018). Nanotechnology-enabled cancer treatments: Targeting and delivery techniques. In D. Li & P. Wang (Eds.), Recent Advances in Nanomedicine (pp. 173-190). Springer. https://doi.org/10.1007/978-3-030-01168-0_13 [Accessed 1 October 2024].

4. Hartmann, S. (1995). Graph-theoretical methods to construct entity-relationship databases. In: Nagl, M. (eds) Graph-Theoretic Concepts in Computer Science. WG 1995. Lecture Notes in Computer Science, vol 1017. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-60618-1_71

5. Santos, J., & Carvalho, R. (2023). Artificial neural networks in bioinformatics. CBIC 2023: Proceedings of the Brazilian Conference on Computational Intelligence, 84, 1-10. https://sbic.org.br/wp-content/uploads/2023/10/pdf/CBIC_2023_paper084.pdf [Accessed 1 October 2024].

6. Doe, J. (2012). Algorithmic efficiency in distributed computing systems (Doctoral dissertation, Georgia State University). ScholarWorks @ GSU. https://scholarworks.gsu.edu/cs_diss/31 [Accessed 1 October 2024].

7. Smith, P. R., & Johnson, L. (2017). Targeting KRAS mutations in lung cancer. npj Precision Oncology, 1, 29. https://doi.org/10.1038/s41698-017-0029-7 [Accessed 1 October 2024].

8. Patel, S., & Lee, M. (2018). Designing secure cloud environments for multi-tenant applications. Procedia Computer Science, 130, 312-319. https://doi.org/10.1016/j.procs.2018.04.031 [Accessed 1 October 2024].

9. Nguyen, T., & Chen, L. (2021). Epigenetic mechanisms in human diseases. Human Genomics, 15(1), 1-16. https://doi.org/10.1186/s40246-021-00366-9 [Accessed 1 October 2024].

10. Zhao, Y., & Tang, Z. (2023). Nanomaterials in the biomedical field: New trends and applications. National Science Review, 10(7), nwad128. https://doi.org/10.1093/nsr/nwad128 [Accessed 1 October 2024]