**Group members name and PSID:** Hao Wang 2344601, Udom Ifeanyi 2220031, Kevin Tieu 1902070

**Background:**  
Materials discovery and design play a crucial role in advancing technologies across sectors, from electronics to energy storage. Predicting material properties from their atomic structure and composition can significantly reduce the time and effort required for experimental validations. This project aims to harness the power of machine learning for predicting key properties such as the band gap, energy above the hull, and the relationship between the number of sites and volume based on crystal structures and elemental composition.

**Dataset:**

* **Source:** [Crystal System Properties for Li-ion batteries](https://www.kaggle.com/datasets/divyansh22/crystal-system-properties-for-liion-batteries/), this dataset contains data about the physical and chemical properties of the Li-ion silicate cathodes. These properties can be useful to predict the class of a Li-ion battery. These batteries can be classified on the basis of their crystal system. Three major classes of crystal systems include: monoclinic, orthorhombic and triclinic.
* **Data points:** the dataset "Crystal System Properties for Li-ion batteries” has 339 unique material records (data points).
* **Features:** Space group (Crystal Structure), Formation Energy, Energy Above Hull, Band Gap, Number of sites (Nsites), Density, Volume, Band structure, Elemental composition (e.g., Li, Mn, Si, O, Fe, Co)
* **Learning Target:** The proposed analysis is to uncover relationships between material properties, develop predictive models for these properties, classify materials based on inherent characteristics, and visually represent data for better understanding. The ultimate aim is to bridge the gap between data-driven insights and the real-world physics of materials.

**Plan of Work:**

1. **Data Preprocessing:** Parsing the chemical formulae to extract the stoichiometry of elements using the chemparse library. Merging this with the original dataset provides a richer feature set.
2. **Exploratory Data Analysis:** Using heat maps to understand correlations between features. Histograms can provide insights into the distribution of crystal structures and other properties.
3. **Regression Tasks:**
   * 2D Regression for Band Gap vs. E Above Hull and Nsites vs. Volume. Validation using parity plots.
   * 3D Regression (possibly polynomial regression) using features like Nsites, Density, and Volume.
4. **Classification Task:** Using the K-nearest neighbors (KNN) algorithm to predict the crystal space group based on the properties and composition of materials.
5. **Insights Generation:** Deep dive into the most frequently appearing crystal structures and comparing their average energy, density, etc. Further, understanding the physical origins of correlated properties.
6. **Model Evaluation:** For each machine learning task, the models will be evaluated using appropriate metrics like R2 score for regression and accuracy for classification.

**References:**

1. D. Agarwal, "Crystal System Properties for Li-ion batteries," Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/divyansh22/crystal-system-properties-for-liion-batteries/>
2. "Seaborn Heatmap: A Comprehensive Guide," [Online]. Available: <https://www.geeksforgeeks.org/seaborn-heatmap-a-comprehensive-guide/>. [Accessed: 19-10-2023].