Case study: Rare variant burden testing and phenotype prediction

For this case study, suppose you are a scientist investigating which genes regulate Low Density Lipoprotein (LDL) levels in blood using whole exome sequencing data from a cohort of 2000 individuals. For each individual, you have performed whole exome sequencing, measured the blood LDL level and recorded some additional covariates such as genetic sex, BMI and age. Also, you have assessed the pathogenicity of all genetic variants using the AlphaMissense[1] tool.

All this data is stored in 3 files:

- **Genotypes.csv:** Contains individual-genotype data from the whole exome sequencing. Rows represent individuals, columns represent variant IDs. For simplicity, we only provide the variants from 50 protein coding genes.
- **Phenotypes.csv:** Includes measured LDL levels, age, genetic sex, and BMI.
- **Annotations.csv:** Provides AlphaMissense scores and gene IDs for each variant

a) Your goal is to find out which genes have a strong influence on measured LDL levels using a rare variant burden test. For this, you will have to follow the following steps:
    i) Compute the gene burden for each gene as the weighted sum of the variants present in the gene for each individual.
    ii) Use the burdens across all individuals to test whether it is significantly associated with the measured LDL levels. You can use a simple linear model for this.
b) Use your computed gene burdens to build a model that predicts the LDL level for each individual. You can use any model architecture. For example, start with a simple linear model first and assess later if a fully connected neural network improves predictive performance. Make sure you objectively measure the model's performance.

[1] Cheng, Jun, et al. "Accurate proteome-wide missense variant effect prediction with AlphaMissense." *Science* 381.6664 (2023): eadg7492.