



# Hackathon

SONALDEEP HALDER AND SUBINOY ADHIKARI

TEAM NAME: CODEAI

DATE: 11/03/2023

# Method of tokenisation and Vectorization

- ▶ The tokenisation is done at atom level
- ▶ Each atom is considered as token. Eg Cl is taken as 'Cl' and not ['C', 'l'].
- ▶ Once the tokenisation is done, the vectorisation is done using count vectorizer.
- ▶ Here corresponding to each unique token that forms the vocabulary, the counts is taken as the feature value.
- ▶ The same vocabulary is used while converting the test data.

# Model Selection

- ▶ Kernel Ridge regression is selected for all the four properties.
- ▶ A grid search with cross validation is performed separately for each label (property).
- ▶ The grid was formed by the parameters: Kernel Function=('rbf', 'laplacian');  $\alpha = \text{linspace}(1e-5, 1e-2, 25)$
- ▶ The models with the best set of parameters were used for training and prediction purposes.