Additional Information

There are a total of **4294 molecules** (data points) and **1433 features** in the molecules\_rdkit.csv. We first process the dataset for further analysis and use it to answer each question.

The following data-processing steps are performed (in the same order):

1. Data Normalization: we first **apply z-score normalization** to all numeric features in molecules\_rdkit.csv and output variables (performance, damage) in molecules\_results.csv.

Feature Selection:

1. Using missing value %: here, we measure the missing value ratio for each feature/column in the dataset. From the analysis, there are some columns where the percentage of missing values is less (between 0-10%, 10-20%, etc.), and some columns have missing value percentages as high as 70-80%.

**We pick a threshold of 20% and drop columns where the percentage of missing values is greater than 20%.** The missing values in the remaining columns are then replaced by the mean values of those columns. This method results in a total of 1325 features in the dataset. This is still a very high number. We use another feature selection method to reduce the number of features further.

1. Discarding constant columns: we **discard columns having same value** throughout. There are 122 such columns. After this step, we end up with total 1203 features.
2. Using correlation: In a loop, **pairwise correlation** is performed between the given feature and the remaining unexplored features. Using 0.7 as a threshold magnitude, we then select the given feature as a replacement for features that show correlation >= +-0.7 with the given feature. The process is repeated for all the features.

Toward the end of this process, we get 321 features. We use these selected features in further analysis.

## Question 1:

A clustering step is performed to recommend the first 20 molecules. We use a **k-means clustering algorithm with k=20** (number of clusters) and randomly choose any one molecule from each cluster for the experiment.

## Question 2:

Since the number of features after feature selection (321) is significantly larger than the dataset size (20), there is a high chance that it would lead to overfitting. Therefore, we would ideally perform another feature selection method based on the correlation of features with output variables (performance and damage) and select features with a high correlation with output. With a few samples (20), the dataset may not provide sufficient information to reliably estimate the relationships between features and output variables. We could use all the molecules in molecules\_results.csv to estimate the correlation of features with output variables, but assuming we don't have access to all performance and damage values for all the remaining molecules (4274) based on the nature of the problem, we skip this step.

Thus, for this question, we use all 321 features for selected 20 molecules (from question 1) for training the model. We use a **random forest algorithm** as a starting model as they generally perform well on higher dimensional data and are less prone to overfitting due to the ensemble nature of the model. Ideally, we would perform model selection- train multiple models and pick one that performs better than other models. Additional hyperparameter tuning/regularization may be required depending on the models used.

Because of the smaller dataset size, we measure the model's performance using the **leave-one-out cross-validation technique** and report average values of a mean squared error on the train and validation/test set. These are the resulting scores:

Mean Train MSE: 0.6599872963871211

Mean Validation MSE: 5.169555650237333

## Based on these scores, we conclude that **the first 20 iterations are not enough**. This is because **the model is overfitting**, as the mean squared error on the validation set is much higher than the train set. It is also quite evident why it would overfit since the model is trained on only 20 molecules, and the number of features (321) is very high compared to the train set size. To solve this problem, as a first step, the train set size must be increased, and the number of features in training must be decreased with the help of domain knowledge and additional feature selection or dimensionality reduction techniques.

## Question 3:

To select the next 20 molecules to try, we first evaluate the random forest model trained in question 2 on all the remaining molecules. We use the following objective function to pick molecules:

**Objective function = (normalized) performance – (normalized) damage**

The goal is to target molecules that maximize the "performance" and minimize the "damage." **Molecules with high performance and low damage would have high value for objective function**. Thus, we calculate objective function values for all the remaining molecules and sort the molecules based on the objective function values in descending order. We pick the top 20 molecules with higher values for objective function than the rest.

**Note**: We have used normalized outputs in training. This is generally not required, but in this question, we use normalized values for "performance" and "damage" to ensure we give equal weightage to both "performance" and "damage" while calculating objective function as actual "performance" and "damage" values range over different scales.