* Creation of affinity training datasets:

From the 1,122 PPIs contained in the’dataset\_1122.csv’, 299 of them are positive interactions that were used in the creation of the training dataset. An extra column named ‘Top\_Docking\_Score’ has been calculated using the HDOCK method was used as the endpoint of the regression models.

For these 299 interactions, 61 features were calculated (62 with the ‘Top\_Docking\_Score’)

The preprocessing procedure is implemented using InSyBio Dataset Preprocessing pipeline. In this procedure, columns that contain 100% missing values are removed. Then the rest are normalized using arithmetic sample- wise normalization and the missing values are filled using the kNN method of imputation.

The preprocessing results are the following:

- Before preprocess:

Total number of calculated features: 62

- After preprocess:

Total number of calculated features: 42

‘Affinity\_dataset\_new\_2.csv’ → The final preprocessed dataset

* ‘dataset\_MMGBSA’

The Dataset created using the MMGBSA method for Protein-Protein Binding Affinity calculation

MMGBSA\_1 - MMGBSA\_10 are the columns containing the rescored affinity scores for each protein-protein docked structure from the highest to the lowest value (highest being MMGBSA\_1).

AVERAGE\_3,AVERAGE\_5,AVERAGE\_10 contain the average score for the best 3-5-10 affinity scores for each pair.