BEST PREPARATION AND METHOD:

Using the Rdkit we manually extracted features such as number of atoms, number of valence electrons and 9 more features. Then using Mol2vec (pre trained model of 300 dimension), we were able to extract 300 more features from the SMILES Sequence. Concatenating the features gave me the vectorized form of all molecules.

Then we trained our Neural network for an ascertained number of epoch, batch size, layers and learning rates as mentioned in the code. Regularization and Dropout was used to reduce the overfitting.

OTHER PREPS ANALYZED

1)
Variations of pretrained Mol2vec model such as model_300dim.pkl and model_100dim.pkl were tried and tested.

2)
ProtVec, another Word2vec based model, specifically for proteins and their bonds, was also tested with the given data but could not be trained well.

OTHER METHODS ANALYZED

1)
Ridge model for various solvers was tested

2)
RidgeCV model with inbuilt cross validation, for various alpha values was tested

3)
Bayesian Ridge model was tested

4)
SVR (Regression model of SVM) for learning rates was tested