The following steps were carried out to solve this task. First, different kinds of feature matrices were created to test them on different classifiers. The idea for the initial feature matrix was to create 4 one-hot-encoded vectors for each of the amino acid letters. For each of the 20 letters featured in the training set one vector slot was created. If, let’s say, A is the first letter of the 4 letter sequence, then the first vector has a 1 at the 0th index followed by straight zeros. However, in the end it turned out that another feature matrix variant yielded better scores (about 2 % for each trained and tested model). This matrix consists of the same four vectors as described above and makes use of special letter combinations. Specifically, the number of pairs, triplets, and quadruplets in the 4 letter sequence is counted and stored as well. E.g. the sequence AABC would be stored as [1, 0, 0, 0]. The 0th index represents the number of pairs, the 1st index the number of triplets, the 2nd index the number of quadruplets, and lastly, 2 pairs at index 3. To find the most promising ML model we carried out a grid search, that is, several models with different parameter combinations were run through a k-fold cross-validation to find the optimum parameter combination for each model. This ultimately leads to finding the best model of the best models. The classifiers that managed to pass the public score baseline were RandomForestClassifier (~ 83 %), HistGradBoostingClassifier (~ 88 %), and SVC (~ 88-90 %).