Solutions to Q3

1 Brief summary of literature:

The **data doppelganger** usually appears when training and validation sets are highly similar for some reasons and **doppelganger effect** represents an inflation performance because of data doppelganger. Thus, the study mainly focuses on **prevalence of functional doppelgangers**, **identification of data doppelgangers (mainly PPCC)**, **confounding effects of PPCC data doppelgangers in machine learning (ML)** and **possible ways to mitigate the doppelganger effect**.

In the 1st part prevalence of functional doppelgangers, several examples have been presented to show the common phenomenon of data doppelganger in various areas, such as in **protein function prediction** and **drug discovery**:

(1) Protein function prediction is based on similar sequences but not functions, which could result in incorrect predictions, such as twilight-zone homologs and enzymes that are dissimilar in sequence overall but with similar active site residues (different functions).
(2) In drug discovery, take QSAR for an example. It is used to predict the biological activities of molecules from their structural properties but small variations in structure may result in different biological activities.

In identification of data doppelgangers, the study demonstrate different methods to identify the data doppelganger, such as ordination method/ embedding method in reduced dimensional space and dupChecker. However, these methods are sometimes unfeasible or have bad effects. As a result, pairwise Pearson's correlation coefficient (PPCC) is used in identification.

When it comes to confounding effects of PPCC data doppelgangers in machine learning (ML), the study shows that PPCC data doppelgangers could inflate ML performance (should perform poorly) in both training and validation data and the more doppelganger pairs, the more inflated the model is. At the same time, it is noticeable that not all models are equally affected. **KNN and naïve bayes models** have a clearer linear relationship between performance inflation and doppelganger dosage than **decision tree and logistic regression models** (in Figure 3).

Last but not least, it is important to find ways to mitigate the doppelganger effect and the study proposes some possible ways:

(1) Place all PPCC data doppelgangers together in the training set or validation set but it is a suboptimal solution.
(2) Split training and test data based on individual chromosomes or use different cell types for 2 data sets to ensure the independence of training data set and validation data set.
(3) Simply remove all the PPCC data doppelgangers but it is not applicable to small data sets with a high proportion of PPCC data doppelganger.

2 Thinking and speculation of the doppelganger effect and possible solutions:

Generally, the doppelganger effect exists in all machine learning models when training data set is similar to validation data set. Thus, it is reasonable that this kind of phenomenon exists in other types of data as well. Some possible examples are as below:

(1) **MRI brain images** used for diagnosis of stroke lesions have multi-modals and the locations of stroke lesions, which are usually close to the brain blood vessels, could be very similar in different patients.
(2) **Sequencing** is very important in genomic analysis and the basic technique is **Sequence Alignment**. It is quite possible that two sequences have similar **regions of exons** and if these two sequences are in training data set and validation data set respectively, it could result in doppelganger effect.

As far as I am concerned, the most effective way to cut down the doppelganger effect is to try to ensure the independence between training data set and validation data set. For example, we can use different samples from patients with different diagnostic dates, use samples from patients of different ages or some other ways.