Question 3

I don’t think doppelganger effects are unique to biomedical data. The doppelganger effects should be applied to all datasets that have characteristics that are very similar to each other when the data are obtained independently. Models trained and validated on such dataset doppelgangers (where the training and validation sets are highly similar) may result in a model that performs well regardless of the quality of training. More often, this doppelganger effect is evident in biomedical data, where there is a large amount of doppelganger data and its inflationary effect.

This effect is seen in many areas of bioinformatics, for example, in protein function prediction, drug discovery, molecular structure characterisation, etc. Given the possibility of confounding doppelganger effects, we need to be able to identify data doppelgangers between the training and validation sets prior to validation. The use of ranking methods and scatter plots can be used to obtain information on how the samples are distributed in the reduced dimensional space.

In order to avoid this doppelganger effects in the practice and development of machine learning models for health and medical science, there are several approaches:

Firstly, the doppelganger effect can be eliminated by placing all the doppelgangers in the training set with an accuracy down to 0.5, when all the doppelgangers of the PPCC data are placed in the training set. However, it has certain drawbacks: when the size of the training set is fixed, which means that each contained data doppelganger results in a less similar sample being excluded from the training set, it may lead to a model that may not generalise well due to the model's lack of knowledge;

Secondly, this could establish good practice/standards in the field by implementing individual chromosomes based on splitting the training and test data (rather than considering all chromosomes together) and by using different cell types to generate training-evaluation pairs. However, it is still flawed: this is difficult to do in practice as it presupposes the existence of a priori knowledge and high quality contextual/benchmark data;

Further, it is possible to attempted data trimming by removing variables contributing strongly towards data doppelgängers effects;

Fourth, perform careful cross-checks using meta-data as a guide.With this information from meta-data, we are able to identify potential doppelgängers and classify them all into training or validation sets, effectively preventing doppelgänger effects and allowing for a relatively more objective evaluation of ML performance;

Fifth, data stratification is performed. Rather than evaluating model performance on the entire test data, the data can be divided into layers of varying similarity;

Sixth, perform very robust independent validation checks involving as many datasets as possible (divergent validation). Although not a direct hedge against data doppelgangers, different validation techniques can inform the objectivity of the classifier. It also illustrates the generality of the model (in terms of real-world use), despite the possible existence of data doppelgangers in the training set.

In Li Rong Wang, Xiuyi Fan, Wilson Wen Bin Goh (2022), it shows the pervasiveness of the DE across other data modalities such as high-throughput gene expression (genomics). In their article they demonstrate that DEs are also present in widely used gene expression profiling techniques. They explore DEs in well-studied microarray gene expression data from the Belorka and Wong study (Belorkar and Wong, 2016) and in widely available RNA-Seq gene expression data from the Cancer Cell Line Encyclopedia (CCLE) project (Broad, 2018; Ghandi et al., 2019). they use doppelgangerIdentifier for analysis, presenting that results are widely observed in multiple diseases and in high-throughput assay platforms capturing gene expression, and that doppelganger may be confounded by batch effects.

References:

1. Li Rong Wang, Xin Yun Choy, Wilson Wen Bin Goh.(2022).Doppelgänger spotting in biomedical gene expression data,iScience,Volume 25, Issue 8,2589-0042, https://doi.org/10.1016/j.isci.2022.104788.

2. Wang, L. R., Wong, L., & Goh, W. W. B. (2022). How doppelgänger effects in biomedical data confound machine learning. Drug discovery today, 27(3), 678–685. https://doi.org/10.1016/j.drudis.