Answer to Question 3

Based on my understanding of the literature, I gave my personal viewpoints on some of the issues in the title in my report. In this literature, the author raises an important question in biomedical data, the doppelganger effect. From where I stand, the doppelganger effect is that when training the classifier, we need to divide the data into the training set and the verification set, but because of the inherent correlation of the data itself, even if we randomly divide the data set and the verification set, the inevitable similarity between the two sets will result in good performance of the classifier in both data sets. However, the trained classifier may perform poorly on other datasets that do not have a high degree of similarity with the datasets we used to build the classifier, so we cannot achieve the initial purpose of building it.

This article enumerates a number of examples of data doppelgangers in the biomedical field to help readers understand their significance and illustrate the necessity of solving this problem. What impresses me is the example of inferring the function of proteins from protein sequence similarity. In previous structural biology classes, we all used this conclusion—"proteins with highly similar sequences can be thought to have the same function"—as a principle for designing proteins or interpreting the functions of proteins. So this article undoubtedly inspired me to question a lot of "guidelines."

Coming back to the question, I think the phenomenon of data doppelganger occurs not only in the biomedical field, for example, the same problem will be encountered in the process of using CNN convolutional neural networks to train image classifiers. We assume that when there are many similar pictures or even the same pictures in the same category, even if the training set and test set are randomly divided, there will be many similar pictures in the training set and test set, and the accuracy of the model will be very high in both sets. However, the accuracy of putting the model to the actual test is very low, indicating that the existence of redundant data in the example of image recognition may also make the prediction results of the model falsely high.

And as for how to avoid data doppelgangers, I think, as is mentioned in this paper, that it is not feasible to delete the same part of data directly for small data sets, but if the number of data sets is large enough and the duality data is relatively small, it may set the threshold of acceptable similarity and delete samples with high similarity. For the data set with a small number of samples and large duality data, label the same part, such as the same operators that affect the same biological process in the same organization, give it a smaller weight, increase the weight for different parts, and enlarge the influence of the difference part on the training classifier according to the actual problem. However, such a trade-off between variance and deviation often needs enough a priori knowledge to support the judge.