A Report to *How doppelgänger effects in biomedical data confound machine learning*[1]

Doppelganger effects (DEs) occur when samples exhibit chance similarities such that, when split across training and validation sets, inflates the trained machine learning (ML) model performance.[2] In this case, even though the validation sets are independently derived from training sets, the test result of ML model is not reliable. In the feature paper *How doppelgänger effects in biomedical data confound machine learning*, authors showed that DEs are prevalent in biomedical data and mentioned pairwise Pearson's correlation coefficient(PPCC) could be used to identify data doppelgangers(DDs). In detail, sample pairs with high PPCCs are referred to as PPCC DDs. However, previous papers didn't make a link between PPCC DDs and whether they are able to inflate ML model performance. So next, the authors constructed a series of experiments using renal cell carcinoma (RCC) proteomics data and found that PPCC DDs did result in inflated ML model performance. Then, the authors showed us three methods used to avoid DEs, but both of them have some limitations. First, when the DDs are placed only in training sets or validation sets, the DEs is eliminated. But when the size of each dataset is fixed, this method may cause not well-generalized models. Second, we can generate training-validation pairs from a different source. For instance, splitting training and validation datasets based on individual chromosomes, different types of cells, or different patients. Obviously, this method requires good quality contextual data. Third, we could simply remove all PPCC DDs. However, this method will cause small datasets with high proportion of PPCC DDs unusable. In addition, we attempted data trimming by removing variables contributing strongly toward data doppelganger effects. This method doesn't work well because the high correlation between data pairs cannot be simply explained by some highly correlated variables. In the end, the authors gave us some recommendations to avoid DEs in practice, which include performing careful cross-checks using meta-data as a guide, performing data stratification, and performing extremely robust independent validation checks involving as many

datasets as possible.

Actually, except proteomics data, there also are some other examples of DEs in biomedical data. For example, in Wang and Choy's study, they found DEs in gene expression data(RNA-Seq). They use the procedure which is the same as the reported paper to identify PPCC DDs: if data pairs in the same class from different patients have higher PPCCs than data pairs in different classes from different patients, then former data pairs are counted as PPCC DDs.

Moreover, I think DEs are not unique to biomedical data. They are quite common in data science research such as finance, marketing, and social science research. Typically, in research about weather or climate, DEs are also prevalent. For example, if there are two cities with similar latitude, terrain, and climate condition, they are distributed in training set and validation set separately. An ML model trained and tested on this training-validation pair is supposed to have high accuracy. However, such a model is not reliable to apply to predict climate conditions in a third city that has similar latitude and terrain with the former two cities but has totally different climate condition in fact.

In conclusion, I think no matter whether in biomedical data or other types of data, constrained by our measurement methods, we virtually are using dimension-reduced variables to represent some relatively high-dimensional objects. Cells, markets, and climate are very complex systems influenced by countless factors. Data we used to depict these systems are only their low dimensional projection. Thus, although some data points appear to be similar based on their quantitative features or variables, they actually have different underlying nature. And this phenomenon can lead to biased or inaccurate results in ML models.

According to the arguments above, it is meaningful to check and avoid DEs. Here, Wong and Fan gave a protocol using PPCC to identify functional doppelgangers.[3] As

to avoiding DEs, except the methods discussed in the reported paper, I think developing highly robust ML algorithms is also a solution. Because based on the reported paper KNN model, naive bayes model, decision tree model, and logistic regression model showed high variance in dealing with data doppelganger which implies that improving ML algorithm against doppelganger effects is possible. Besides, data augmentation is also worth attempting. We can incorporate additional data to reduce similarity statistically. Last but important, improving the measurement methods to obtain comprehensive data will be helpful to ameliorate DEs.

References

1 L.R. Wang et al., Drug Discovery Today (2021), https://doi.org/10.1016/j.drudis.2021.10.017

2 Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). iScience, 25(8), 104788. https://doi.org/10.1016/j.isci.2022.104788

3 Wang, Li & Fan, Xiuyi & Goh, Wilson. (2022). STAR Protocols. 3. 101783. 10.1016/j.xpro.2022.101783.