# Doppelgänger Effects in Biomedical Data

Xinyi LIU

suplxy@126.com

**Abstract** Data doppelgängers occur when validation sets share a high degree of similarity to training sets, causing models to perform well regardless of how they are trained (i.e., the doppelgänger effect). The report illustrates that the effects are not unique to biomedical data, but also exist in other different fields. In addition, methods for the avoidance of doppelgänger effects are also discussed. It provides the reader with an understanding of the topic and perspective on further potential developments.

## 1 Instruction

In the past few years, the rapid development of advanced computing and imaging systems in biomedical engineering has given rise to new aspects of study. Additionally, the increasing scale of biomedical data necessitates precise data mining methods that are based on machine learning [1]. However, using machine learning algorithms to analyze biomedical data for classification and prediction is challenging because of the existence of data doppelgängers, which occur when validation sets share a high degree of similarity to training sets. Despite several documented examples of data doppelgängers, it remains uncommon to check whether the sample training–evaluation pairs are independent. Therefore, it is imperative to investigate the nature of data doppelgängers and propose methods for the avoidance of doppelgänger effects.

## 2 Prevalence of Doppelgänger effects

Doppelgängers in biomedical data include evaluation of existing chromatin interaction prediction systems, protein function prediction, drug discovery and so on [2]. However, generally, doppelgänger effects exist in all machine learning models when the training data set is similar to the validation data set. Therefore, it is reasonable that this kind of phenomenon is not unique to biomedical data, and there are data doppelgängers in different kinds of fields, such as environmental science, business, etc.

## 3 Avoidance of Doppelgänger Effects

To avoid doppelgänger effects, identifying data doppelgängers before the training-validation split is indispensable. Ordination methods, embedding methods, dupChecker, and the pairwise Pearson’s correlation coefficient are all not feasible due to certain problems, but the basic design of PPCC as a quantitation measure is reasonable methodologically.

There are a few ways available to help lessen the impact of doppelgänger effects. The first one is to conduct thorough cross-checks using meta-data as a guide. We can put all PPCC data doppelgängers together into either training or validation sets, effectively preventing doppelgänger effects. Secondly, we can assort data into strata of different similarities, rather than judging model performance on whole test data. In addition, we can also perform extremely robust independent validation checks involving the maximum number of data sets [2]. As far as I am concerned, the most effective way to cut down the doppelgänger effect is to try to ensure the independence between training data sets and validation data sets.

## 4 Conclusion

This report illustrates that doppelgänger effects are prevalent in biomedical data, and they are also likely to happen in other fields. A few available methods are put forward in order to avoid or mitigate the impact of doppelgänger effects.

## References:

[1] Park, C., Took, C. C., & Seong, J. K.. (2018). Machine learning in biomedical engineering. Biomedical Engineering Letters.

[2] Wang LR, Wong L, Goh WWB. (2021). How doppelgänger effects in biomedical data confound machine learning, Drug Discovery Today.