Report

If the derived data are very similar to each other, Data doppelgängers occur and cause models to perform well no matter how they are trained. This kind of event is called the doppelgänger effect. It still remains uncharacterized, therefore, we need to show the popularity in biomedical data, the process of arising doppelgängers, the confounding effects and the way to mitigate the doppelgänger effect.

For the Abundance of data doppelgängers in biological data, in protein function prediction, the approach would be unable to correctly predict functions for proteins with less similar sequences but similar functions. And for another similar example exists in drug discovery, in some instances, because of poorly trained models and uninformative structural properties, Sorting similar molecules with similar activities into both training and validation sets confounds model validation

Regarding to the process of arising doppelgängers, One possible method is to use ordinary methods or embedding methods as well as with scatterplots. However, we found such method to be unfeasible. Then, we can use dupChecker to identifies duplicate samples although dupChecker does not detect true data doppelgängers that are independently derived samples that are similar by chance. Another measure is named PPCC, which does not detect true data doppelgängers. We can use it to identify potential functional doppelgängers.

Considering about the confounding effects, the presence of PPCC data doppelgängers in both training and validation data inflates ML performance. Moreover, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance. While regarding to the protein sequence function predictions. Also, The similarities between doppelgänger effects and leakage are evident in the training validation set although not all models are equally affected.

The final part is about mitigating the doppelgänger effect, the first method is to place all PPCC data doppelgängers together in the training set. The second way is to splitting training and test data based on individual chromosomes, as well as using different cell types to generate the training–evaluation pair. However, this method has its own difficulty to put it into practice. The third method is about using doppelgangR for the identification of doppelgängers, PPCC data doppelgängers and then removed them to mitigate their effects.