The Report of the Essay

This paper focuses on the fact that doppelganger data is common in biomedical data and that when doppelganger data is present, machine learning (ML) models are affected, resulting in models that perform well no matter how they are trained, which are doppelganger effects are mentioned in the paper. The paper first gives an explanation of what doppelganger data is and defines what kind of doppelganger data are functional doppelgangers. Next, the paper explains that there is an abundance of doppelganger data in biomedical data and proposes that protein function prediction can have doppelganger effects if trained on uninformative structures, allowing poorly trained models to train well on certain molecules. In order to solve the problem of doppelganger data, the paper mentioned that doppelganger data should be identified first and then trained or validated, first of all, the method of dimensionality reduction is denied because the data loses some of the information of the previous data, making it difficult to distinguish the doppelganger data from the previous data by the dimensionality reduction. Later in the paper, two early approaches to identifying doppelganger data were mentioned, the first one is dupChecker which has leakage issues, where the leakage issues I understand that they are equivalent to a put-back sampling, so they can cause duplication. The other is to calculate the pairwise Pearson correlation coefficient between the sample pairs, if the calculated PPCC is quite high, it means that the correlation between the two data pairs is very high, and it is considered that the two sample pairs form PPCC doppelganger data, however, the paper does not mention whether this method of determining doppelganger data by calculating Pearson correlation coefficients between sample pairs has the ability to confound ML models. Although the data in the paper proposing PPCC suffers from data duplication, the design of PPCC as a quantitative indicator (in statistical terms the absolute magnitude of the Pearson correlation coefficient represents the degree to which two data can be explained by each other, the higher the absolute value of PCC, the higher the degree of explanation) is methodologically sound, so The authors used the RCC proteomic data of NetProt software library which was specifically designed to validate the effect of PPCC doppelganger data on ML. Ultimately, it was verified that PPCC doppelganger data produce inflationary effects similar to data leakage as do functional doppelgangers, it just produces different effects in different ML models. The article goes on to discuss ways to improve the doppelganger data. Improving the data by removing PPCC doppelganger datasets using the PPCC method discussed previously seems feasible, but would make the amount of data too small to use. The authors then give several suggestions, the first being to use information from the metadata to construct positive and negative scenarios to predict the range of PPCC scores and the presence of leakage. The second suggestion is to stratify the data and evaluate the performance of the ML model on each stratum separately, while the third is to start with the validation set, using a strong independent validation set and including as many datasets as possible.

I think doppelganger effects are not unique for biomedical data. In living organisms, the value of individual biological differences can be large or small, but the biological information fed back by the organism exists in a confidence interval with a high level of confidence, and within this confidence interval the data is likely to be real life, and these biological data are also very similar to each other. This makes ML more sensitive to data occurring within this confidence interval, with higher accuracy in discrimination, classification, prediction and validation, while there is a high probability of error when processing data outside the confidence interval. And this can happen in computer vision, for example, if we train our ML model to recognise bottles on the street, and probably most of the bottles on the street in life are plastic bottles and only a few are glass bottles, so there will be a large similarity between our training and validation sets, which can also lead to doppelganger effects that inflate the ML training effect. So all in all I think doppelganger effects are not unique for biomedical data.

I think that in the future development of machine learning in health and medicine, reducing doppelganger effects can be achieved by reducing the impact of doppelganger data on the accuracy of the model in the training set, for example by setting a penalty function on the doppelganger data during training, the more doppelganger data are trained the less they contribute to the accuracy or by not training them at all, then we perform extremely robust independent validation checks involving as many data sets as possible like the third recommendation to adjust our penalty function so that the accuracy of the model trained on the training set is similar to that of the validation set, which may lead to a more unaffected ML model.