Functional doppelgängers are derived from the original samples and have similar characteristics to each other. In machine learning model training datasets, FDs are split into training and validation sets, and the accuracy of the final validation is exaggerated, i.e., the transition is inflated. In general, cross validation, such as hyperparameter tunning and feature selection, was used to validate models, but this method is affected by doppelgängers. Currently, there is also an exaggerated performance in chromatin interaction prediction systems. The same occurs for protein function and structure prediction. Data doppelgängers are present in all these biomedical data. However, this does not mean that doppelgangers effect is not present outside biomedical data. ((Kleinlogel et al., 2021))

In order to avoid the negative effect of doppelgangers, the most reliable way is to avoid mixing FDs between the training and test sets. Then it is necessary to be able to identify FDs before model training. To identify doppelgangers and spot biomedical gene expression data, doppelgangerIdentifier is a good choice. Of course, there are limitations. This identifier cannot identify PPCC data doppelgangers from more than two batches and all Functional doppelgangers from some datasets. In addition, PPCC is used as an indicator of the relevance of DD, so non-linearly correlated functional doppelgangers cannot be detected by the identifier. For the first case, the user can avoid this situation by combining all pairs of batches. For the second case, the user can experiment with other correlation coefficients for non-linear relationships, such as spearman (SROCC)

or COPULA entropy. (Wang, Fan, et al., 2022). In addition, users can use meta-data as a guide to perform careful cross-check. It is worth noting that reasonable data doppelgangers come from the same type of samples from different patients. We can use this type of information to identify potential doppelgangers and assign them all to the training set or validation set, effectively preventing the doppelgangers effect. Finally, the proportion of functional doppelgangers and data doppelgangers in the validation set will also affect the machine learning model performances.

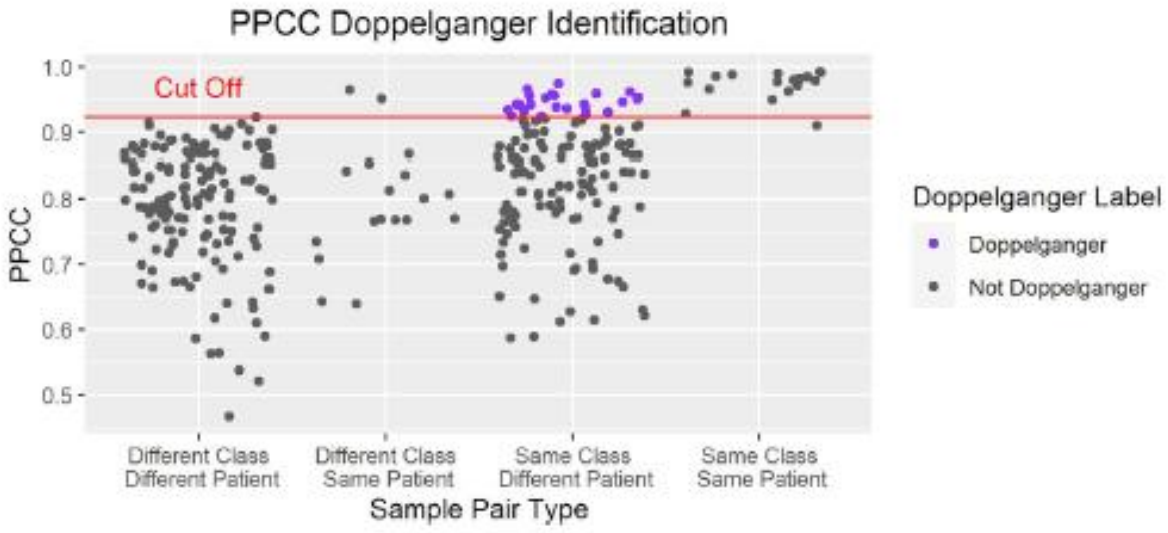

Fig.1 The red line represents where the DD identification threshold is. Samples above this threshold are marked as DD in purple, while samples below this threshold are marked as non-DD in gray.

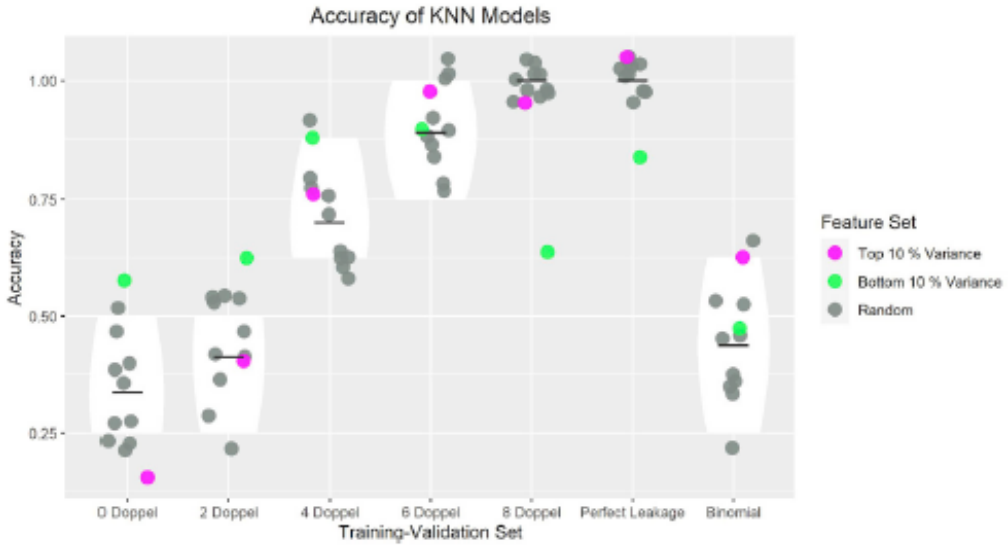

Fig. 2 The pink dots represent the model trained with the highest variance feature set, while the green dots are the opposite.
The highest variance feature set is more accurate, and vice versa.

In terms of gene expression, Researcher focused on exploring two RNA-seq datasets (large_upper and lymph_lung) in identifying PPCC DDs. Researcher observed that there was a high percentage of PPCC DDs in lymph_lung, compared with large_upper , which is a similar trend when comparing the number of PPCC DD samples. There are stronger outliers in large_upper, which leads to greater inflation of PPCC cutoffs, which in turn leads to greater decrease in DD recognition sensitivity. PPCC DD recognition approach still needs to be improved, which could be done from milly_lung, large_upper, DMD, and leukemia data sets to improve model performance. Then the reason for the poor effect on functional dichotomies may be that the PPCC threshold reduces the value of detected PPCC DDs, which may be able to change the definition of PPCC in the future. (Wang, Choy, et al., 2022)

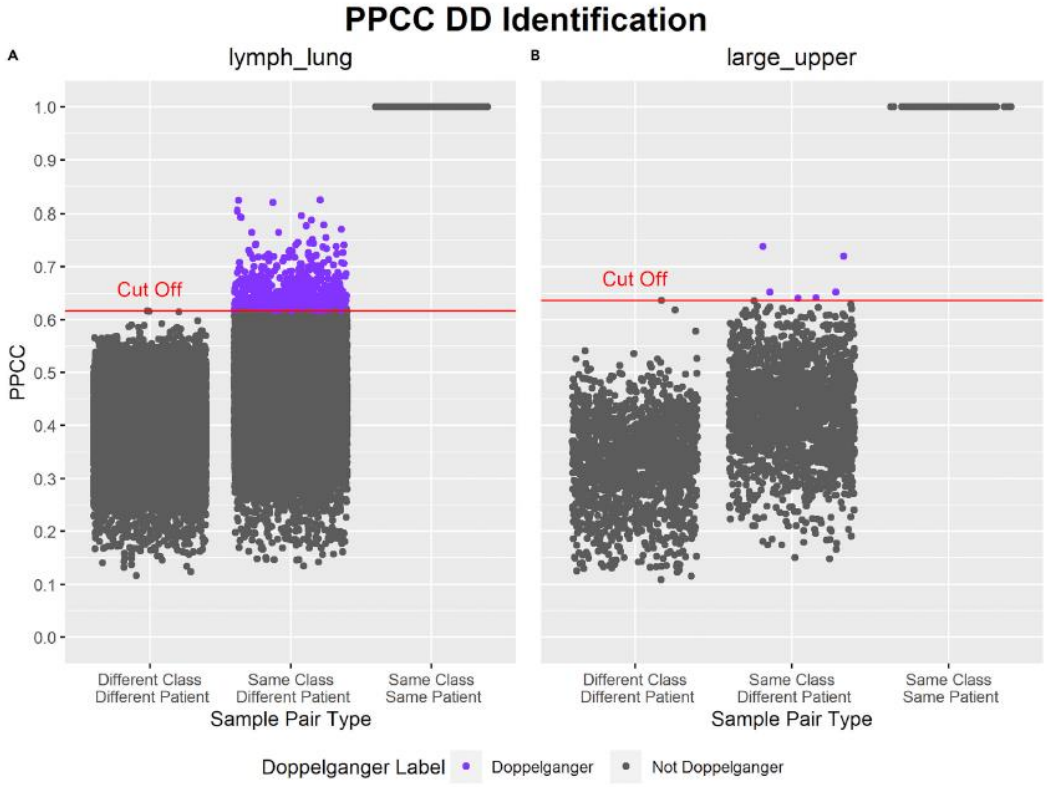

Fig.3 Results of PPCC DD Identification on lymph_lung and large_upper data sets

Reference

1. Kleinlogel, E. P., Curdy, M., Rodrigues, J., Sandi, C., & Schmid Mast, M. (2021). Doppelganger-based training: Imitating our virtual self to accelerate interpersonal skills learning. *PLoS One*, *16*(2), e0245960. https://doi.org/10.1371/journal.pone.0245960
2. Wang, L. R., Choy, X. Y., & Goh, W. W. B. (2022). Doppelganger spotting in biomedical gene expression data. *iScience*, *25*(8), 104788. https://doi.org/10.1016/j.isci.2022.104788
3. Wang, L. R., Fan, X., & Goh, W. W. B. (2022). Protocol to identify functional doppelgangers and verify biomedical gene expression data using doppelgangerIdentifier. *STAR Protoc*, *3*(4), 101783. https://doi.org/10.1016/j.xpro.2022.101783