The doppelganger effect of biomedical data

1. What is doppelganger effects? How did it appear?

The doppelganger effects effect refers to the fact that the machine learning model performs well on the verification set regardless of how it is trained. This exaggerates the performance of model and may complicate the process of model selection based solely on validation accuracy. Data similarity between the training set and the verification set leads to doppelganger effects[1].

2. Is it unique to biomedical data?

I personally don't think doppelganger effects is unique to biomedical data. First of all, it is undeniable that doppelganger effects is more likely to occur in the biomedical field, which may be caused by the small amount of data in the biomedical field and the high similarity between the data. In other areas, there may be large differences between data and sufficient data to choose from, and the repetition rate will be greatly reduced. However, in the biomedical field, such as MRI images of a specific area, the similarity between images is too high, and it is difficult to obtain data. Sometimes, data sets can only be expanded by means of data augmentation. In this case, the training set and the verification set will be highly similar, thus doppelganger effects appears. But it's also a problem in areas where there's also a lack of data, where there's little variation between the data, so I don't think it's unique to biomedical field.

3.Some examples

On the basis of functional genomic data, TargetFinder is a machine learning technique that forecasts enhancer-promoter interactions. Cao and Fullwood[2] detected doppelganger effects in targetFinder datasets.

They found most of the positive enhancer-promoter pairings in the TargetFinder datasets had significantly overlapping window areas with at least another positive enhancer-promoter pair in the same dataset (Fig.1a). For instance, with a 99%

reciprocal-overlapping-fraction criterion, around 53-76% of positive samples had windows that overlapped with the windows of other positive samples, but just 0.16% of positive samples had windows overlap with the windows of negative samples at this cutoff (Fig. 1b). As the overlapping percentage grew, the mean absolute differences between the features in window sections of overlapping samples became less than the difference between two randomly selected positive samples (Fig. 1c). When the samples are positive, the strong similarity across window characteristics is likely to skew the cross-validation findings[2].

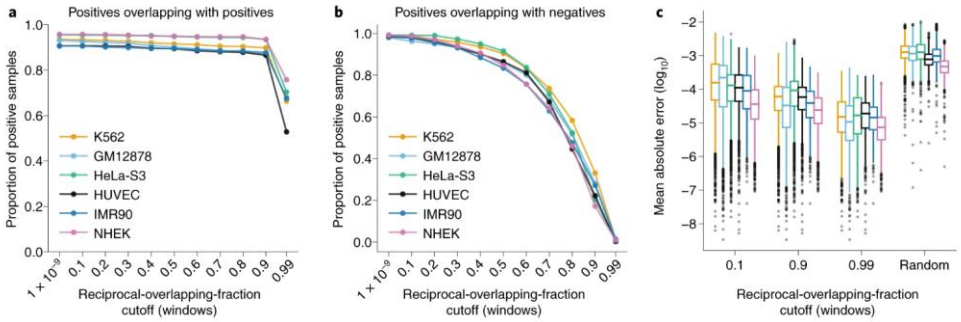

Fig 1 **:** Analyses of TargetFinder[2]

In various cancer data sets, investigators frequently share or re-use specimens in later studies, but this will lead to doppelgängers. Levi Waldron et al[3] analysed cell lines and datasets of ovarian, breast, bladder, and colorectal cancers. They manually identified and counted the occurrence of doppelgängers. Some of their findings are shown in Table 1

Table 1[3]

Dataset identifier by type of cancer	total number of samples	Total number of doppelgängers	Source
GSE19915, GSE32894(Bladder)	490	84	University Hospital of Lund, Sweden
TRANSBIG, UNT, UPP(Breast)	586	78	Uppsala County, Sweden
GSE4526, GSE14095(Colorectal)	225	37	Teikyo University School of Medicine, Japan

GSE14333, GSE17538(Colorectal)	754	569	H. Lee Moffitt Cancer Center, USA

4. How to avoid doppelgängers

There's an article on chromosome research that showed that they used chromosome segmentation strategies to avoid doppelgängers. Their idea is that all samples of the same chromosome are either all in the training set or in the validation set[4].

For protein sequences, one solution to avoid doppelgängers is to use a sensitive hidden Markov model profile comparison tool, like HH-suite, to search the test data. This tool can detect sequences that are only loosely connected to the training data[5].

Pairwise pearson's correlation coefficient (PPCC) is a great way to solve this problem. It is used to measure the degree of correlation between two variables. Data pairs of different sample sets with abnormally high PPCC values are identified and thus avoid doppelgängers[6].

References

[1] L. R. Wang, X. Y. Choy, and W. W. B. Goh, "Doppelgänger spotting in biomedical gene expression data," *iScience,* vol. 25, no. 8, p. 104788, 2022/08/19/ 2022, doi: https://doi.org/10.1016/j.isci.2022.104788.

[2] F. Cao and M. J. Fullwood, "Inflated performance measures in enhancer–promoter interaction-prediction methods," *Nature Genetics,* vol. 51, no. 8, pp. 1196-1198, 2019/08/01 2019, doi: 10.1038/s41588-019-0434-7.

[3] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. Birrer, "The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles," *JNCI: Journal of the National Cancer Institute,* vol. 108, no. 11, 2016, doi: 10.1093/jnci/djw146.

[4] F. Cao and M. J. Fullwood, "Inflated performance measures in enhancer-promoter interaction-prediction methods," (in eng), *Nat Genet,* vol. 51, no. 8, pp. 1196-1198, Aug 2019, doi: 10.1038/s41588-019-0434-7.

[5] M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, and J. Söding, "HH-suite3 for fast remote homology detection and deep protein annotation," (in eng), *BMC Bioinformatics,* vol. 20, no. 1, p. 473, Sep 14 2019, doi: 10.1186/s12859-019-3019-7.

[6] L. R. Wang, L. Wong, and W. W. B. Goh, "How doppelgänger effects in biomedical data confound machine learning," (in eng), *Drug Discov Today,* vol. 27, no. 3, pp. 678-685, Mar 2022, doi: 10.1016/j.drudis.2021.10.017.