Answer to Q1

Ziyuan Wang

Answer to Q3

Brief Summary of the paper (Wang, Wong, and Goh 2021)

What is doppelgänger effects

Training and validation sets are highly similar for accidental or other reasons(Wang, Wong, and Goh 2021).In many cases, the model generalization ability of training is poor due to the influence of the split body effect, and the performance is poor when the data is not learned. The current methods for identifying and improving the doppleganger effect are not universal enough and need to be improved. This paper discusses the popularity of functional doppleganger of biomedical data, the influence of data doppleganger on ML, and the methods to reduce the doppleganger effect.

Abundance of data doppelgänger in biomedical data

Some data doppelgänger pointed out in the article

- protein function prediction(Wass and Sternberg 2008; Friedberg 2006). *The functions proteins with less similar sequences but similar functions cannot be predecited*
- quantitative structure–activity relationship (QSAR) models(Paul et al. 2021). *The QSAR model assumes that molecules with similar structures have similar activities and will encounter problems when facing molecules with different structures.*

Other data doppelgänger in bioinformatics

- Single-cell RNA-seq/ATAC *Many single-cell machine learning analyses rely on data sets obtained in similar sequencing environments, similar donor sampling, and other conditions, which can lead to some misleading analyses.Batch effects in different data sets can also interfere with predictions.(Luecken et al. 2021)*
- Sequence Analysis *Sequence analysis is based on the assumption that DNA sequences with similar sequences have similar functions, which is the same as data doppelgänger in the paper.*

PPCC applied in datasets

The pairwise Pearson's correlation coefficient (PPCC) can capture the relation between different samples if the value is high we can deduce data doppelgängers.

When using PPCC to judge data doppelgängers, we can see from the figure that, in general, more doppelgängers will interfere with the machine learning discriminator. As the number of duplicates in the Validation dataset increases, the accuracy of machine learning will mistakenly increase.

Ameliorate data doppelgängers

Current methods:

- Analyze the specific context of the data to develop a more comprehensive and rigorous assessment strategy(Cao and Fullwood 2019).
- Delete PPCC data clone. However, it needs to lose too many samples(Ma et al. 2018; Lakiotaki et al. 2018).

Challenges : **How to solve data doppelgängers without significantly reducing data.**

Whether doppelgänger effects are unique to biomedical data

No

One example is Face recognition. (Rathgeb et al. 2022)The recognition accuracy of doppelgänger effects was improved significantly after the problem was solved.

Avoid doppelgänger effects in biomedical data

Using generative model to extract embeddings(Luecken et al. 2021; Lotfollahi, Wolf, and Theis 2019; Seninge et al. 2021).For scRNA-seq, data doppelgängers can be significantly reduced by using self-supervised models, and we can explore potential relationships between different gene expressions through hidden layers. When our model is used to predict the samples of new donations, even if the effect is caused by batch effect or the difference in physical condition between people from different regions due to different reasons, we can still capture a lot of effective information and make further prediction.

For chemical structure data, we can adopt embedding, including graph neural network, graph convolution neural network and other methods, which have been proved in some studies. In these neural networks, the model learns the relationship of edges (chemical bonds) between different types of nodes (representing chemical elements or groups). In this case, the model is very generalizing and can predict chemical structures that are not present in the data set (Ding et al. 2021; Stokes et al. 2020).

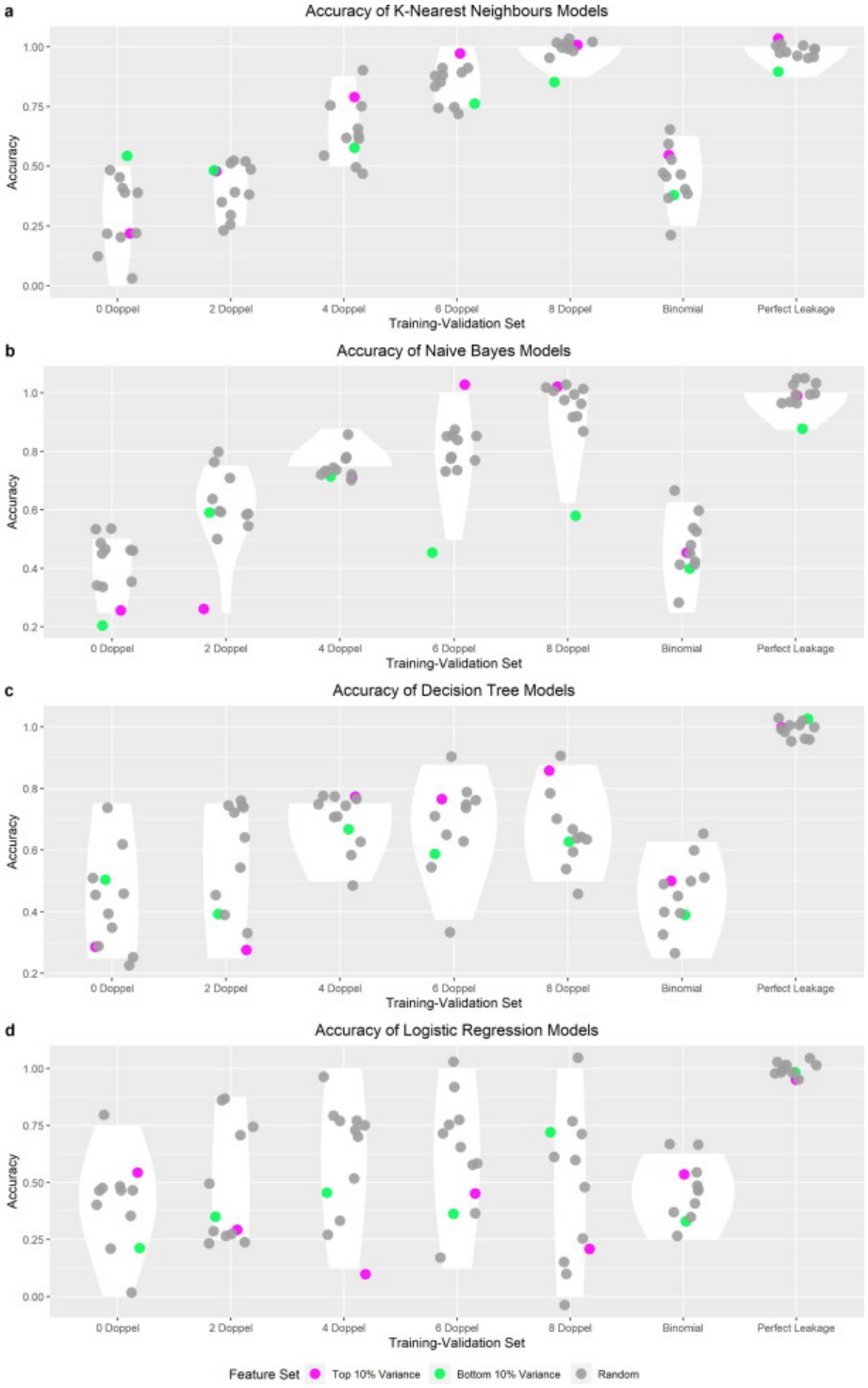

Figure 1: **PPCC-Fig**

Reference

Cao, Fan, and Melissa J. Fullwood. 2019. "Inflated performance measures in enhancer–promoter interaction-prediction methods." *Nature Genetics* 51 (8): 1196–98. https://doi.org/10.1038/s41588-019-0434-7.

Ding, Hongxu, Ioannis Anastopoulos, Andrew D. Bailey, Joshua Stuart, and Benedict Paten. 2021. "Towards inferring nanopore sequencing ionic currents from nucleotide chemical structures." *Nature Communications* 12 (1): 1–9. https://doi.org/10.1038/s41467-021-26929-x.

Friedberg, Iddo. 2006. "Automated protein function prediction - The genomic challenge." *Briefings in Bioinformatics* 7 (3): 225–42. https://doi.org/10.1093/bib/bbl004.

Lakiotaki, Kleanthi, Nikolaos Vorniotakis, Michail Tsagris, Georgios Georgakopoulos, and Ioannis Tsamardinos. 2018. "BioDataome: A collection of uniformly preprocessed and automatically annotated datasets for data-driven biology." *Database* 2018 (2018): bay011. https://doi.org/10.1093/database/bay011.

Lotfollahi, Mohammad, F. Alexander Wolf, and Fabian J. Theis. 2019. "scGen predicts single-cell perturbation responses." *Nature Methods* 16 (8): 715–21. https://doi.org/10.1038/s41592-019-0494-8.

Luecken, Malte D, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, et al. 2021. "Benchmarking atlas-level data integration in single-cell genomics." *Nature Methods* 19 (January). https://doi.org/10.1038/s41592-021-01336-8.

Ma, Siyuan, Shuji Ogino, Princy Parsana, Reiko Nishihara, Zhirong Qian, Jeanne Shen, Kosuke Mima, et al. 2018. "Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis." *Genome Biology* 19 (1): 1–14. https://doi.org/10.1186/s13059-018-1511-4.

Paul, Debleena, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh K. Tekade. 2021. "Artificial intelligence in drug discovery and development." *Drug Discovery Today* 26 (1): 80–93. https://doi.org/10.1016/j.drudis.2020.10.010.

Rathgeb, Christian, Daniel Fischer, Pawel Drozdowski, and Christoph Busch. 2022. "Reliable Detection of Doppelgängers based on Deep Face Representations." http://arxiv.org/abs/2201.08831.

Seninge, Lucas, Ioannis Anastopoulos, Hongxu Ding, and Joshua Stuart. 2021. "VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics." *Nature Communications* 12 (1): 5684. https://doi.org/10.1038/s41467-021-26017-0.

Stokes, Jonathan M, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, et al. 2020. "A Deep Learning Approach to Antibiotic Discovery." *Cell* 180 (4): 688–702.e13. https://doi.org/https://doi.org/10.1016/j.cell.2020.01.021.

Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. 2021. "How doppelgänger effects in biomedical data confound machine learning." *Drug Discovery Today*. https://doi.org/10.1016/j.drudis.2021.10.017.

Wass, Mark N., and Michael J. E. Sternberg. 2008. "ConFunc - Functional annotation in the twilight zone." *Bioinformatics* 24 (6): 798–806. https://doi.org/10.1093/bioinformatics/btn037.