Report based on

"How doppelgänger effects in biomedical data confound machine learning"

Before reading the paper "How doppelgänger effects in biomedical data confound machine learning", I felt excited to engage in the area of biomedical data, in a deeper way. And I have prepare myself for several questions regarding of doppelganger effect: What is doppelganger effect? How it is produced? Is it existed in other forms, such as RNA sequencing, and imaging? How to avoid it or how to check it before using data for machine learning? Disappointingly, maybe due to my lack of background knowledge, I failed. After reading the paper for multiple times, I was not able to gain comprehension even on the definition of doppelganger effect (in the area of data science) and thus unable to answer the question above. Therefore, I believe I should not be writing this report as expected.

After searching for relevant papers about doppelganger effect, my understanding is that doppelgänger effect is a misleading phenomena which occurs when data present occasional similarities (between training and validation sets), and would eventually results in the inflationary performance in machine learning. [i] This effect is found in various kinds of area such as biomedical data, Cloud computing, etc. For instance, confounding similarities are discovered, between similar chromosomes, RNA families, and shared ancestry.[ii] [iii] [iv]And in protein function prediction, it is intuitive to say that proteins with similar sequence probably have the same functions. Nonetheless, under the same predicting approach, for proteins with similar functions but obviously disparate sequence, the prediction would be incorrect. [v] Therefore, inflationary performance in machine learning is anticipated and requires inspection before putting into use.

[i] Wang LR, Choy XY, Goh WWB. Doppelgänger spotting in biomedical gene expression

data. *iScience*. 2022;25(8):104788. Published 2022 Jul 19. doi:10.1016/j.isci.2022.104788

[ii] Cao F, Fullwood MJ. Inflated performance measures in enhancer-promoter interaction-prediction methods. *Nat Genet*. 2019;51(8):1196-1198. doi:10.1038/s41588-019-0434-7

[iii] Szikszai M, Wise M, Datta A, Ward M, Mathews DH. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. *Bioinformatics*. 2022;38(16):3892-3899. doi:10.1093/bioinformatics/btac415

[iv] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat* Rev Mol Cell Biol. 2022;23(1):40-55. doi:10.1038/s41580-021-00407-0

v Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. Drug Discov Today. 2022;27(3):678-685. doi:10.1016/j.drudis.2021.10.017