Report

1 Article Recap

This section includes the main points and key parts of the article [1].

Background

Machine Learning models offered tremendous convenience in data analysis of biomedical data. Moreover, to evaluation the reliability of model prediction, cross validation techniques is commonly used. However, this popular validating measure does not always guarantee an excellent model.

There is a phenomenon called **data doppelgängers** which refers to the situation where samples are similar across training and validation sets. Data doppelgängers can ensure a model to perform well on validation set, given such similar input, while in practise, the actual performance may not be what it appeared in validation stage. This can be defined as **doppelgängers effect**. Encountering data doppelgängers does not guarantee the occurrence of doppelgängers effect, thus data doppelgängers that causes doppelgängers effect is specifically defined as **functional doppelgängers**.

Examples

The occurence of data doppelgängers has already been identified in varied fields of bioinformatics.

- **Modern Bioinformatics**

  In a detailed evaluation of existing chromatin interaction prediction systems [2], it is found that the model has a exaggerated performance due to the data doppelgängers, which ensures the high performance of model.

- **Protein Function Prediction**

  Proteins are assumed to have the same function when they have similar sequence and it seems to be valid using abductive reasoning with data doppelgängers. However, this would not be the case for protein with similar function but differentiated sequences.

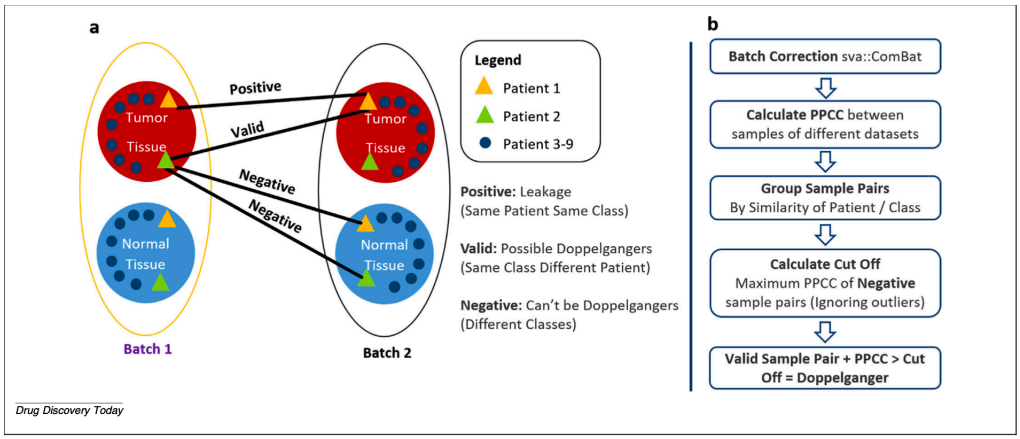
- **Drug Discovery**

  To predict biological activities of molecules, quantitative structure–activity relationship (QSAR) models are proposed. It assumes similar molecules with similar structural properties would have similar activities. Thus there is not differences between training and validation dataset (data doppelgängers). Real poor performed model can only be identified by testing on similar molecules with different activities.

Identification of data doppelgängers and doppelgängerseffect

Given that doppelgängers effect would probably confound the model performance, it is crucial to identify data doppelgängers. Early methods proposed like ordination methods and embedding methods ordupCheccker does

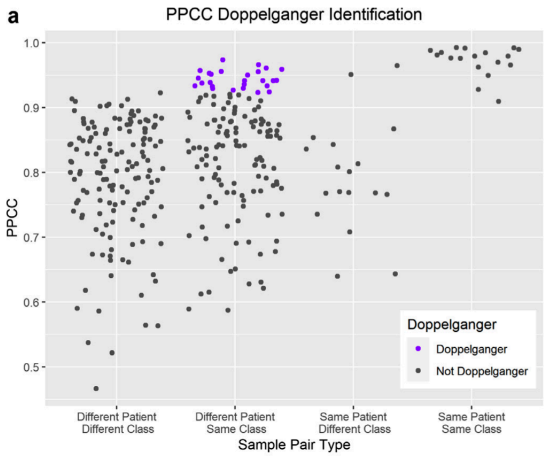
not effective enough, here the writer utilized a quantitative method, which can identify data doppelgängers through calculation. **Pairwise Pearson's Correlation Coefficient (PPCC)** [3] is a reasonable quantitation method. Abnormally high PPCC value can identify data doppelgängers, but it does not indicate that the ML model would definitely be confounding. In the article, the researchers build benchmark scenarios with renal cell carcinoma (RCC) pro-



teomics data from [4]. High rate of PPCC data doppelgängers was identified, suggesting the lack of reliability and sensitivity of original method. Moreover, the PPCC value is naturally high for replicates or samples from the same tissue but different patients. Emphasizing the usefulness of PPCC value.

Effect

After identifying data doppelgängers, whether they have an obvious inflationary effect (functioning doppelgängers) is investigated on varied ML model. As displayed in the following figjure, ML models showed higher performance on PPCC data doppelgängers with no exception.



Through the result, it seems that models can perform accurately on similar samples while showing bad performance on less-similar data. This confirms the assumption before that PPCC data doppelgängers leads to functional doppelgängers. Moreover, the affected level appears to vary among different types of models.

Identification and Amelioration To identify doppelgängers effect, it is resonable to find data doppelgängers, for which PPCC is definitely a useful method. Additionally, except for finding potential data doppelgängers that might lead to doppelgängers effect, it is also possible to test whether doppelgängers effect occurs on a model is also feasible.

For instance, we can used independent validation sets to check multiple times. Any sign of low accuracy would infer a gap in models.

To ameliorate doppelgängers effect, it is best to remove data doppelgängers, and valid methods are used, however, the method is proven elusive for not working on small data sets composing small proportion of non-doppelgängers data. Thus suboptimal solutions are proposed————eliminating factors that might lead to doppelgängers effects. To achieve this goal, the following measures are proposed.

- Cross-checks using meta-data as a guide, especially samples arising from same class but different patients and technical replicates arising from the same sample.
- Data stratification. Testing model performance on each stratum separately and gaps would be identified in model when it performs badly on specific strata.
- Checking robust independent validation that involving as many data sets as possible. It can indicate if the classifier is objective.

2 Supplementary Information

This section would include the response to the questions and requirement.

Doppelgängers effects do not belong explicitly to biomedical data, it appears as an emerging problem in data analysis models in varied areas. For instance, in medical imaging, the evaluation of models relies on training and testing the models with independent sets of data, which is very easy to encounter leakage under incorrect implementation, leading to overoptimistic results [5]. Though not identified in the paper, it is surely a doppelgängers effect by definition. Moreover, the existence of doppelgängers effect is also found in RNA-Sequencing, using the PPCC method, identifying PPCC data doppelgängers [6].

References

[1] Wang L R, Wong L, Goh W W B. How doppelgänger effects in biomedical data confound machine learning[J]. Drug Discovery Today, 2021.

[2] Cao F, Fullwood M J. Inflated performance measures in enhancer–promoter interaction-prediction methods[J]. Nature genetics, 2019, 51 (8): 1196–1198.

[3] Waldron L, Riester M, Ramos M, et al. The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles[J]. JNCI: Journal of the National Cancer Institute, 2016, 108 (11).

[4] Guo T, Kouvonen P, Koh C C, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps[J]. Nature medicine, 2015, 21 (4): 407–413.

[5] Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future[J]. NPJ digital medicine, 2022, 5 (1): 1–8.

[6] Wang L R, Choy X Y, Goh W W B. Doppelgänger spotting in biomedical gene expression data[J]. Iscience, 2022, 25 (8): 104788.