Doppelganger effects in Biomedical data confound machine learning

Introduction:

ML plays important role drug development for faster identification of potential targets. ML techniques improve the decision-making in pharmaceutical data across various applications like QSAR analysis, hit discoveries, de novo drug architectures to retrieve accurate outcomes. Data doppelgangers occur when independently derived data are very similar to each other, causing models to perform well regardless of how they are trained.This report concentrates on doppelganger's prevalence in biomedical data, demonstrate how doppelgangers arise, and provide proof of their confounding effects.

ML models in drug discovery:

ML models helps in shortlisting better drug candidates faster, reducing time spent on discovery and testing. A new anti- cancer drug candidate, EXS21546, was discovered. Several ML-identified drugs and drug combinations for coronavirus 2019(COVID-19) treatment have discovered. Classifiers are trained models that are responsible for the prediction of new drug disease interaction. When assessing the performance of a classifier, the training and test datasets should be independently derived. The independently derived raining and test sets could yield unreliable validation results. It is imperative to investigate the nature of data doppelgangers and propose improved methods for doppelganger identification.

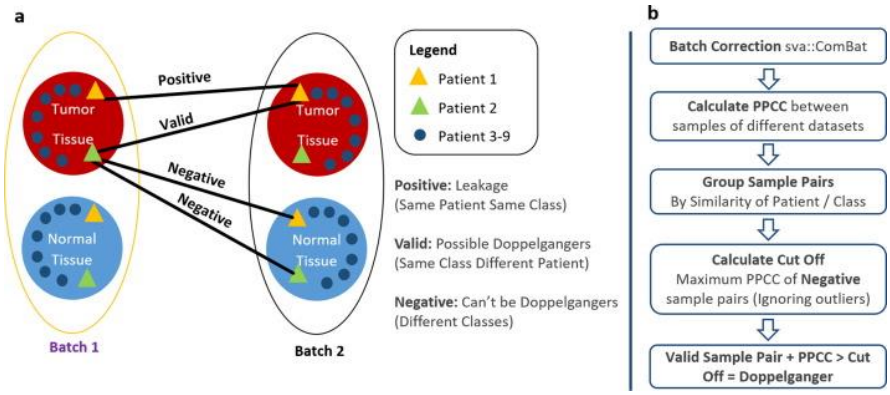
Abundance of data doppelgangers:

Doppelgangers have been observed in modern bioinformatics. Cao and Fullwood performed a detailed evolution of existing chromatin prediction systems. It has been overstated because of problems in assessment methodologies. Doppelgangers present in protein functions prediction, protein with similar sequences are descended from the same ancestor protein. QSAR models are classification and regression models. They assume structurally similar molecules have

similar activities. Well trained models are able to detect small variations. Poor trained models would fail to identify true biological activity.

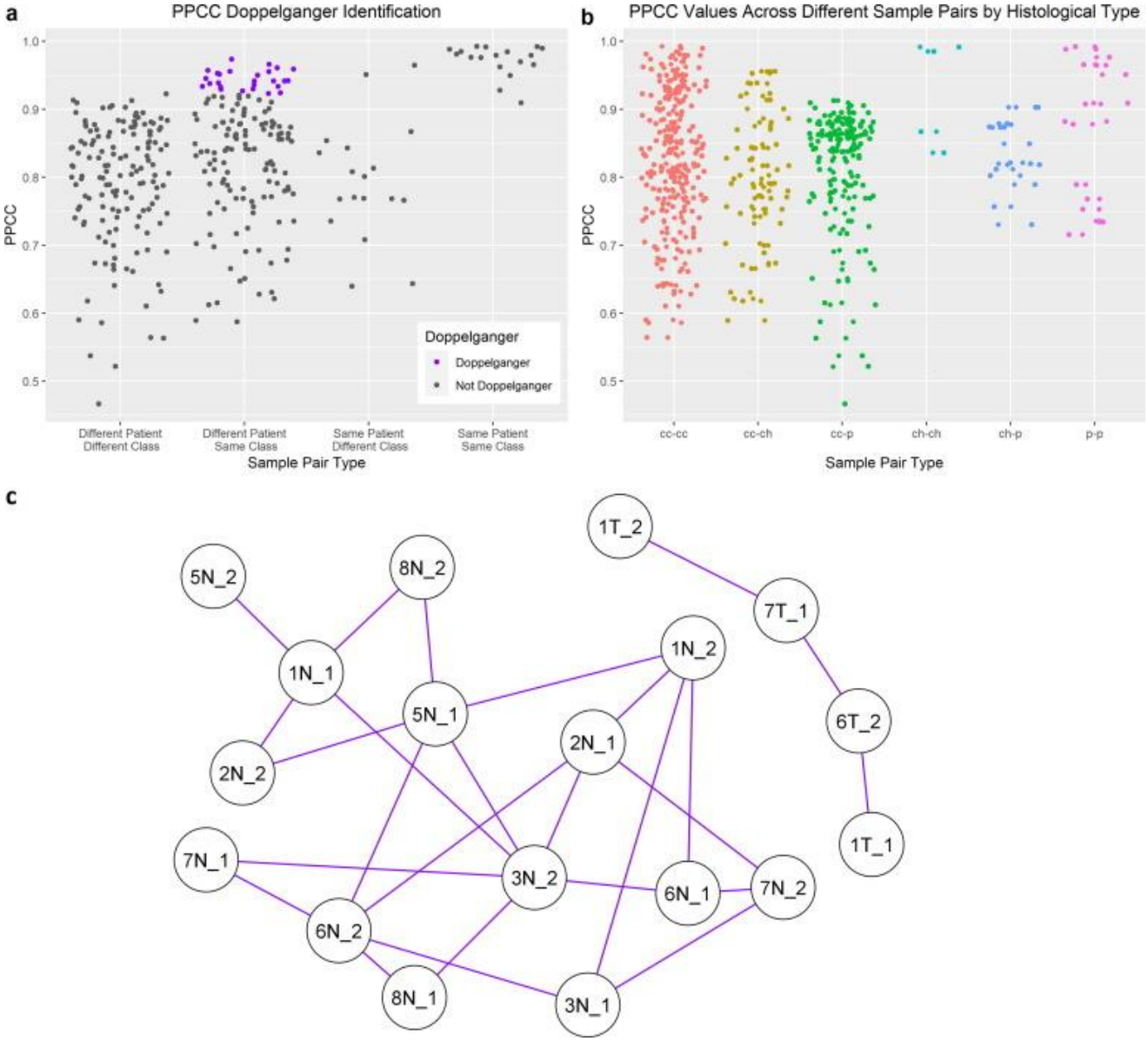
Identification of data doppelgangers:

It is difficult to identify the presence of doppelgangers between training and validation sets before validation. The identification of doppelganger would be to use ordination methods or embedding methods. This method is unfeasible because doppelgangers are not distinguishable in reduced dimensional space. Dupchecker identifies duplicate samples by comparing the MD5 fingerprints of CEL files. It doesn't detect true data doppelgangers that are independently derived samples that are similar by chance. The pairwise Pearson's correlation coefficient (PPCC), captures relations between sample pairs of different data sets. An anomalously high PPCC value indicates that a pair of samples constitutes PPCC data doppelgangers. During reanalysis of their data that their reported doppelgangers were in fact the result of leakage. The basic design of PPCC as a quantitation measure is reasonable methodologically. Thus, we use this for identifying potential functional doppelgangers.



This Diagram illustrating the pairwise Pearson's correlation coefficient(PPCC) data doppelganger identification method.(a)Naming convention for different types of sample pair based on the similarities of their patient and class. (b)Process of PPCC data doppelganger identification. PPCC data doppelgangers are defined as valid sample pairs with PPCC values greater than all negative sample pairs.

To construct benchmark scenarios, we used the renal cell carcinoma (RCC) proteomics data of Guo et al. RCC was chosen for its utility in constructing clear cut scenarios: (i) negative cases, in which doppelgangers are nonpermissible by constructing samples pairs of different class labels; (ii) valid cases, in which doppelgangers are permissible by constructing sample pairs assigned to the same class label but from different samples. These effects can then be compared against positive cases. we observed a high proportion of PPCC data doppelgangers.



(a) Distribution of pairwise Pearson's correlation coefficients (PPCCs) across different sample pairs. The X-axis indicates the types

of sample pair grouped by the similarities of their patient and class. The Y- axis indicates the PPCC (i.e. Pearson's correlation coefficient between two samples).The 26 PPCC data doppelgangers are labelled in purple. (b)Distribution of PPCC values of different sample pairs by their histological types. X-axis indicates the types of sample pairs grouped by histological type pairs. Clear cell renal cell carcinoma (RCC) is indicated by cc, chromophobe RCC by ch, and papillary RCC by p. Y-axis indicates the PPCC.(c) 26 PPCC data doppelgangers visualised as a graph. Each node represents a different sample, the first number indicates the patient number, the following letter represents the class(N, normal; T, tumour). The presence of an edge between each node/sample means that the two samples are PPCC data doppelgangers. There are 18 nodes in this graph (i.e.,18 samples out of a total of 36 samples are doppelgangers with atleast one other sample).

Recommendation:

To guard against doppelganger effects. Our first recommendation is to perform careful cross-checks using meta-data as a guide. The plausible data doppelgangers that warrant concern are samples arising from same class but different patients. We are able to identify potential doppelgängers and assort them all into either training or validation sets, effectively preventing doppelganger effects, and allowing a relatively more objective evaluation of ML performance.

Our second recommendation is to perform data stratification. Instead of evaluating model performance on whole test data, we can stratify data into strata of different similarities. Given that the proportion of kidney cancer cells of each tissue is known(papillary RCC comprises10% of kidney cancer cells),25 the poor performance of the classifier on papillary RCC would indicate that this 10% of kidney cancer cell samples is an area of weakness for our classifier.

Our third recommendation is to perform extremely robust independent validation checks involving as many data sets as

possible. In future approach we could identify functional doppelgangers directly.

Conclusion:

We find that doppelgangers are fairly common in our test data, and that it has a direct inflationary effect on ML accuracy. This reduces the usefulness of ML for phenotype analysis and subsequent identification of potential drug leads. The extent of this inflationary effect varies depending on two main factors: the similarity of functional doppelgangers and the proportion of functional doppelgangers in the validation set. Unfortunately, doppelganger effects are not easy to resolve analytically. To mitigate the doppelganger effect, we recommend identifying data doppelgangers before the training validation split.