How doppelgänger effects in biomedical data confound machine learning

1.Background

1.1 Concept definition

Data doppelgängers: Independently derived data are very similar to each other.

Doppelgänger effect: Data doppelgängers cause models falsely perform well regardless of how they are trained, which yield unreliable validation results.

1.2 Present situation

It remains uncommon to check whether the sample training–evaluation pairs are independent and/or dissimilar.

Data doppelgängers and their accompanying downstream analytical effects (doppelgänger effects) are poorly documented and not well understood.

Although several proposed methods of identifying data doppelgängers exist, most methods are not generalizable or robust enough.

Procedures for eliminating or minimizing similarity between test and training data still do not constitute standard practice before classifier evaluation.

2.Abundance of data doppelgängers in biological data

Performance of existing chromatin interaction prediction systems has been overstated because of problems in assessment methodologies.

Certain validation data were guaranteed a good performance given a particular training data, even if the selected features were random.

Proteins with similar sequences are inferred to be similar in function, but this approach would be unable to correctly predict functions for proteins with less similar sequences but similar functions.

QSAR models assume that structurally similar molecules have similar activities. Sorting similar molecules with similar activities into both training and validation sets confounds model validation because poorly trained models (trained on uninformative structural properties) might still perform well on these molecules.

3.Identification of data doppelgängers

3.1 Problems

Ordination methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE) is unfeasible because data doppelgängers are not necessarily distinguishable in reduced-dimensional space.

DupChecker, identifies duplicate samples by comparing the MD5 fingerprints of their CELfiles. Identical MD5 fingerprints would suggest that samples are duplicates (essentially replicates and indicative of leakage issues). DupChecker does not detect true data doppelgängers 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 doppelgängers (impossible to determine which one between the pair is the original). Although reasonable and intuitive, it never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks. Reported doppelgängers were in fact the result of leakage (between sample replicates), do not constitute true data doppelgängers.

3.2 Our work

We use PPCC as a quantitation measure for identifying potential functional doppelgängers from constructed benchmark scenarios.

To construct benchmark scenarios, we used the renal cell carcinoma (RCC) proteomics data.

Negative cases: doppelgängers are nonpermissible by constructing samples pairs of different class labels.

Valid cases: doppelgängers are permissible by constructing sample pairs assigned to the same class label but from different samples.

Positive cases: pairs constructed by taking technical replicates arising from the same sample; these constitute obvious leakage issues and, therefore, are not considered doppelgängers.

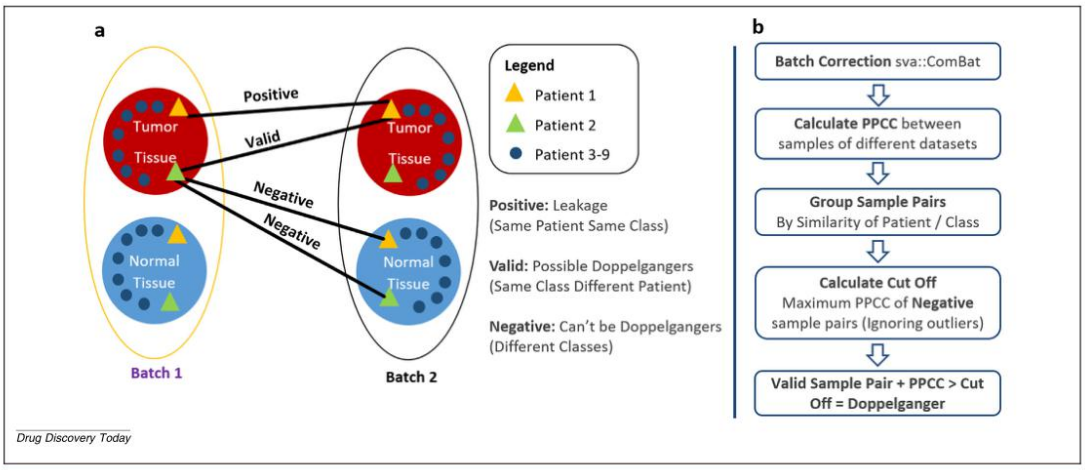
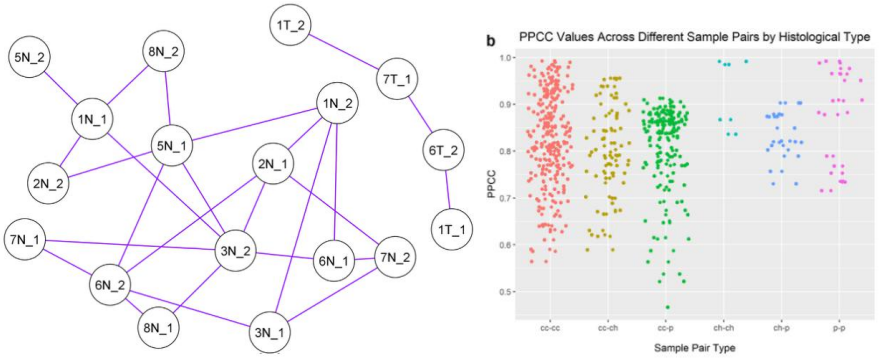


FIGURE 1

Diagram illustrating the pairwise Pearson's correlation coefficient (PPCC) data doppelgänger 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 doppelgänger identification. PPCC data doppelgängers are defined as valid sample pairs with PPCC values greater than all negative sample pairs.

We observed a high proportion of PPCC data doppelgängers. PPCC distributions on the valid scenario exist as a wide continuum, without obvious breaks. This suggests that using outlier detection methods will not be sensitive enough.



PPCC distributions are lower if compare different tissue pairs. By contrast, PPCCs are extremely high when we consider replicates from the same sample or tissue. These evaluations suggest that PPCC has meaningful discrimination value.

4.Confounding effects of PPCC data doppelgängers

4.1 The dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect

We explored their effects on validation accuracy across different randomly trained classifiers.

The presence of PPCC data doppelgängers in both training and validation data inflates ML performance, even if the features are randomly selected. This finding is consistently reproducible.

The more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance.

Where there are many similar examples, good accuracy is easily obtained without in fact assuring generalizability to less-similar examples.

4.2 A possible way of avoiding the doppelgänger effect

When all PPCC data doppelgängers are placed together in the training set, the doppelgänger effect is eliminated. However, constraining the PPCC data doppelgängers to either the training or validation set are suboptimal solutions. In the former, when the size of training set is fixed, it leads to models that might not generalize well because the model lacks knowledge. End up with spectacular winner-takes-all scenarios (the doppelgängers will all either be predicted correctly or wrongly.)

5.Ameliorating data doppelgängers

Strategies based on the particular context of the data could be achieved by splitting training and test data based on individual chromosomes, as well as using different cell types to generate the training–evaluation pair. This is difficult to do practically because it predicates on the existence of prior knowledge and good quality contextual/benchmarking data.

PPCC data doppelgängers could be removed to mitigate their effects. But the removal of PPCC data doppelgängers would reduce the data to an unusable size.

We attempted data trimming by removing variables contributing strongly toward data doppelgängers effects. However, we observed no change in the inflationary effects. This observation hints at the extreme complexity of the doppelgänger effect, given that the reason for high correlations between sample pairs cannot simply be explained by a subset of highly correlated variables.

6.Recommendations

6.1 Performing careful cross-checks using meta-data as a guide

This allowed us to anticipate PPCC score ranges for scenarios in which doppelgängers cannot exist and where leakage exists. The plausible data doppelgängers that warrant concern are samples arising from same class but different patients. With this information from the meta-data, we are

able to identify potential doppelgängers and assort them all into either training or validation sets, effectively preventing doppelgänger effects, and allowing a relatively more objective evaluation of ML performance.

6.2 Performing data stratification

Instead of evaluating model performance on whole test data, we can stratify data into strata of different similarities. Assuming each stratum coincides with a known proportion of real-world population, we are still able to appreciate the real-world performance of the classifier by considering the real-world prevalence of a stratum when interpreting the performance at that stratum.More importantly, strata with poor model performance pinpoint gaps in the classifier.

6.3 Performing extremely robust independent validation checks involving as many datasets as possible

Although not a direct hedge against data doppelgängers, divergent validation techniques can inform on the objectivity of the classifier. It also informs on the generalizability of the model despite the possible presence of data doppelgängers in the training set.

6.4 Future plan

Methods of functional doppelgänger identification do not rely heavily on meta-data. In such approaches, we could identify functional doppelgängers directly. For example, we might look for subsets of a validation set that are predicted correctly regardless of the ML method used. These subsets are potential functional doppelgängers of the training set. Further pairing this approach with PPCC subsequently may allow us to discern the doppelgänger partners of test set samples in the training set. During model evaluation, these subsets should be avoided because they act as functional doppelgängers, and give little insight into the relative performance of different models.

7.My personal view

Firstly,I do not think that doppelganger effects are unique to biomedical data.For example, face recognition. If we only input face photos to build our training and validation set, let we imagine a scene: on this premise, can the model distinguish the face photo/face in the mirror from the real person? The model can easily capture the mirrored facial features and compare them with the information in the database, but it is obvious that the model recognizes the illusion as a living person, which brings hidden dangers to the face recognition application.

I think two of the reasons for doppelganger effect are the limited extent and depth of information we obtain. If we set up the training and validation set through a large number of tiger pictures, even if the model only learns the features of yellow and black, the model can easily identify the tigers in the sheep, but if the cheetah or white tiger is mixed, the result can be very bad. The above problems can be solved by widening and mining the information.Widening: Increasing the diversity of the data set (adding cheetahs photos) to deal with multiple situations and adding negative samples or attention mechanism (adding white tigers photos) to force the model to distinguish. Mining: Use deeper or more unique models (YOLO, U-net, etc.) for training, so that enough features of roots can be extracted from limited data.