citations

Data doppelgängers occur when independently derived data are very similar to each other, causing models to perform well regardless of how they are trained. The report will show the prevalence of data doppelgängers in biomedical data, provide proof of their confounding effects, give some ways to identificate data doppelgängers, and provide some recommendation to guard against doppelgänger effects.

1 Introduction of data doppelgängers

1.1 Definition

It is well established in ML that independently derived training and test sets could still yield unreliable validation results. When a classifier falsely performs well because of the presence of data doppelgängers, we say that there is an observed **doppelgänger effect**.

Data doppelgängers that generate a doppelgänger effect (confounding ML outcomes) are termed **functional doppelgängers**.

1.2 Data doppelgängers in biological data

Data doppelgängers have been observed in modern bioinformatics.

- Cao and Fullwood performed a detailed evaluation of existing chromatin interaction prediction systems. Their work revealed that the performance of these systems has been overstated because of problems in assessment methodologies when these systems were reported.
- Goh and Wong shows the presence of data doppelgängers, whereby certain validation data were guaranteed a good performance given a particular training data, even if the selected features were random.
- In protein function prediction, proteins with similar sequences are inferred to be descended from the same ancestor protein and thereby inherit the function of that ancestor.
- Data doppelgänger exists in drug discovery: QSAR models assume that structurally similar molecules have similar activities. Sorting similar molecules with similar activities into both training and validation sets (by chance during time-split validation or random test set selection) confounds model validation because poorly trained models (trained on uninformative structural properties) might still perform well on these molecules.

1.3 Data doppelgängers in other fileds

Data doppelgängers are not unique to bioinformatics data. It is common for ML training and test sets to be similar. In many cases, training and test sets are split from uniformly measured datasets.

For example, when using machine learning for image recognition training of an object, if there are only similar images of the object in the training set and test set, or there are many similar examples in the training set and test set, it will lead to misjudgment for model training . A poorly trained model may achieve high accuracy.

2 Identification and effects of data doppelgängers

2.1 Identification of data doppelgängers

Identification Method	Limitation
Use ordination methods (e.g., principal component analysis) or embedding methods (e.g., t-SNE), coupled with scatterplots, to see how samples are distributed in reduced-dimensional space.	The method is unfeasible because data doppelgängers are not necessarily distinguishable in reduced-dimensional space
Compare the MD5 finger- prints of their CEL files.	Identical MD5 fingerprints would suggest that samples are duplicates. 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.	It never conclusively made a link between PPCC data doppelgängers and their ability to confound ML tasks. Their reported doppelgängers were in fact the result of leakage (between sample replicates) and do not constitute true data doppelgängers.

Although all these methods have their limitation, the basic design of PPCC as a quantitation measure is most reasonable methodologically. Thus, we always choose this method for identifying potential functional doppelgängers.

2.2 Confounding effects of PPCC data doppelgängers

The researchers provide 3 experiments to show the effects of PPCC data doppelgängers. In addition to describing the phenomenon and visualizing the data, the researchers give their thoughts on the experiments and show how the experiments will support their views.

- RCC experiment

  After identifying PPCC data doppelgängers in RCC(Renal Cell Carcinoma (RCC) proteomics data of Guo et al.9 taken from the NetProt software library), the Researchers explored their effects on validation accuracy across different randomly trained classifiers. They noted that 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 on different sets of training and validation data and on different ML models. Moreover, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance.

- 'Doppel 40' (training-validation set) experiment
  - Phenomenon

    When the validation accuracy for all properly trained models (with 'Top 10% Variance' feature set) on 'Doppel 40 (training-validation set) were stratified into PPCC data doppelgängers and non-PPCC data doppelgängers strata, all ML models showed higher performance on PPCC data doppelgängers than on non-PPCC data doppelgängers .
  - Conclusion

    Where there are many similar examples (many data doppelgängers), good accuracy is easily obtained without in fact assuring generalizability to less-similar examples. However, where there are few similar examples (few data doppelgängers), gaps in the model are revealed and, thus, the model tends to underperform.

- K-nearest neighbor (kNN) models experiment
    - PPCC data doppelgängers act as functional doppelgängers, producing inflationary effects similar to data leakage

      The similarities between doppelgänger effects and leakage are evident in the experiment using k-nearest neighbor (kNN) models in which the training-validation set with eight doppelgängers in validation showed an identical accuracy distribution to the training-validation set with perfect leakage.
    - Doppelgängers affect different models unequally

      Not all models are equally affected: kNN and naïve bayes models have a clearer linear relationship between performance inflation and doppelgänger dosage compared with decision tree and logistic regression models.

3 Ameliorating data doppelgängers

3.1 Flawed attempts

The undesirable inflationary effects on ML produced by doppelgängers raises the question of how doppelgänger effectscould be managed. Researchers provide 3 prossible way of avoiding the doppelgänger effect.

Ameliorating Method	Limitation
Putting all PPCC data doppelgängers are placed together in the training set can make the doppelgänger effect eliminated.	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 (thus, each data doppelganger that gets included causes a less similar sample to be excluded from the training set), it leads to models that might not generalize well because the model lacks knowledge. In the latter, the doppelgängers will all either be predicted correctly or wrongly.
In studies in which the PPCC outlier detection package, doppelgangR was used for the identification of doppelgängers, PPCC data doppelgängers could be removed to mitigate their effects.	This approach does not work on small data sets with a high proportion of PPCC data doppelgängers, such as RCC, because the removal of PPCC data doppelgängers would reduce the data to an unusable size
Removing variables contributing strongly toward data doppelgängers effects.	There is no change in the inflationary effects of the PPCC data doppelgängers after the removal of correlated variables.

3.2 Recommendation against doppelgängers

Although removing data doppelgängers from data directly has proven elusive(see 3.1 Flawed attempts), we still need to guard against doppelgänger effects.

- Perform careful cross-checks using meta-data as a guide

  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.

  - Example

    We used the meta-data in RCC for constructing negative and positive cases. This allowed us to anticipate PPCC score ranges for scenarios in which doppelgängers cannot exist (different class; negative cases) and where leakage exists (same-patient and same-class based on replicates; positive cases). The plausible data doppelgängers that warrant concern are samples arising from same class but different patients.

- Perform 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.

  - Example

    PPCC data doppelgängers and non-PPCC data doppelgängers, and evaluate model performance on each stratum separately

  - Limitation

    In RCC, the non-PPCC doppelgängers used in stratified performance assessment also happen to be papillary RCC samples. Given that the proportion of kidney cancer cells of each tissue is known (papillary RCC comprises 10% 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.

- Perform divergent validation

Perform extremely robust independent validation checks involving as many data sets as possible. Divergent validation techniques can inform on the objectivity of the classifier. It also informs on the generalizability of the model

- Explore new method to identify functional doppelgängers directly

  Explore other methods of functional doppelgänger identification that do not rely heavily on meta- data.

  - Example

    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 (sample pairs between training and validation sets that inflate model accuracies regardless of how we train the model). Further pairing this approach with PPCC subsequently may allow us to discern the doppelgänger part-ners of test set samples in the training set. During model evaluation, these subsets should be avoided becausethey act as functional doppelgängers, and give little insight into the relative performance of different models.

4 Conclusion

We find that doppelgängers are fairly common in 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 doppelgängers and the proportion of functional doppelgängers in the validation set. Doppelgänger effects are not easy to resolve analytically. To avoid performance inflation, it is important to check for potential doppelgängers in data before assortment in training and validation data.