The understanding of doppelganger effects

More and more machine learning is being used in biopharmaceuticals and disease research. By integrating and analyzing the data in the database, an effective model is established to find the target. However, because there may be the same or similar data in many databases, repeated use will affect the simulation effect of the model, doppelganger effects is problematic as it could exaggerate the performance of the ML model on real-world data and potentially complicate model selection processes that are solely based on validation accuracy.

Therefore, identifying doppelganger data in advance and avoiding repeated use of doppelganger data has become a topic worth studying. According to the literature read by the author, doppelganger effects have not been found in fields other than biomedicine. Due to protect patient privacy,publicly available human genomic data is therefore normally summarized at a level that cannot be identified uniquely. Other fields don't have that problem.

Without the use of any software, the effect of doppelganger can be reduced by following ways:

Firstly, using pairwise Pearson's correlation coeffificient (PPCC) as the benchmark to identify whether data doppelgangers(DD) are functional doppelganger(FD) and affected the ML producing inflationary effects

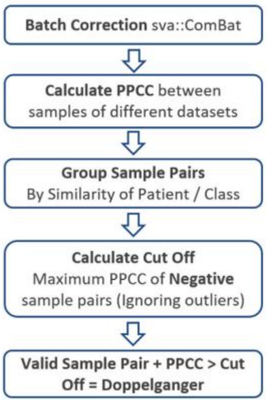

Batch Correction sva::ComBat
Calculate PPCC between samples of different datasets
Group Sample Pairs By Similarity of Patient / Class
Calculate Cut Off Maximum PPCC of Negative sample pairs (Ignoring outliers)
Valid Sample Pair + PPCC > Cut Off = Doppelganger

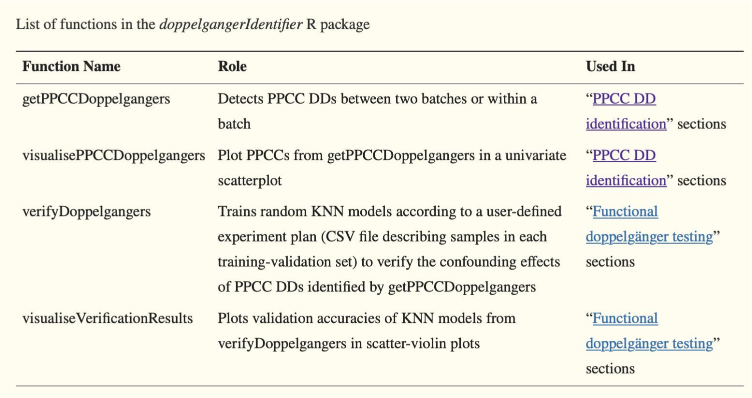
Fig 1.Process of PPCC data doppelgänger identification

According to the literature, removing data doppelgangers from data directly has proven elusive, and these are some recommandations to improve the accuracy of the ML model:

1. To 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.
2. To perform data stratifification， Instead of evaluating model performance on whole test data, we can stratify data into strata of different similarities (e.g., PPCC data doppelgängers and non-PPCC data doppelgängers, and evaluate model performance on each stratum separately).
3. To perform extremely robust independent validation checks involving as many data sets as possible (divergent validation).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 (in terms of realworld usage) despite the possible presence of data

doppelgängers in the training set.

Using software:*doppelgangerIdentifier R package*, dentifing FD and verifing impact
Here, there has doppelgangerIdentifier (DI), an R package with 4 main functions for identifying FDs
and verifying their inflflationary effects on ML mode accuracy.



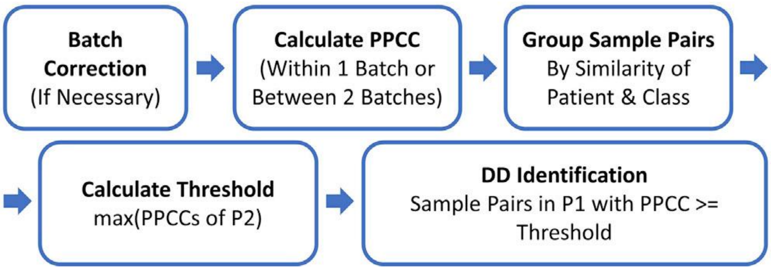
List of functions in the *doppelgangerIdentifier* R package

Function Name	Role	Used In
getPPCCDoppelgangers	Detects PPCC DDs between two batches or within a batch	"PPCC DD identification" sections
visualisePPCCDoppelgangers	Plot PPCCs from getPPCCDoppelgangers in a univariate scatterplot	"PPCC DD identification" sections
verifyDoppelgangers	Trains random KNN models according to a user-defined experiment plan (CSV file describing samples in each training-validation set) to verify the confounding effects of PPCC DDs identified by getPPCCDoppelgangers	"Functional doppelgänger testing" sections
visualiseVerificationResults	Plots validation accuracies of KNN models from verifyDoppelgangers in scatter-violin plots	"Functional doppelgänger testing" sections

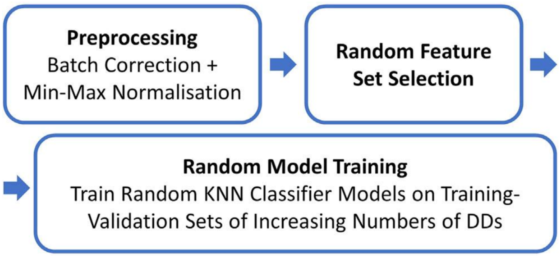
Fig2. The function of each Rpackage

The the basic idea of program implements:

1. Preparing the gene expression data and meta data
2. Data doppelgnger (DD) identifification with PPCC


Batch Correction (If Necessary)
Calculate PPCC (Within 1 Batch or Between 2 Batches)
Group Sample Pairs By Similarity of Patient & Class
Calculate Threshold max(PPCCs of P2)
DD Identification Sample Pairs in P1 with PPCC >= Threshold

3. Functional doppelganger testing and functional doppelganger (FD) verifification


Preprocessing Batch Correction + Min-Max Normalisation
Random Feature Set Selection
Random Model Training Train Random KNN Classifier Models on Training-Validation Sets of Increasing Numbers of DDs

After the program is finished, the software will generate Scatter plot to visualize the recognition results of functional doppelgange.Removing the recognized FD can increase the accuracy of ML.

Reference:

[1]Belorkar A., Wong L. GFS: fuzzy preprocessing for effective gene expression analysis. BMC Bioinformatics. 2016;17:169–184.

[2]Goh W.W.B., Wang W., Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. 2017;35:498–507.

[3]Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. Drug Discov Today. 2022 Mar;27(3):678-685.

[4]Wang LR, Choy XY, Goh WWB. Doppelgänger spotting in biomedical gene expression data. iScience. 2022 Jul 19;25(8):104788.

[5]Wang LR, Fan X, Goh WWB. Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier. STAR Protoc. 2022 Oct 26;3(4):101783.

[6]Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M. The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles. J Natl Cancer Inst. 2016 Jul 5;108(11):djw146.

[7]Piccolo SR, Mecham A, Golightly NP, Johnson JL, Miller DB. The ability to classify patients based on gene-expression data varies by algorithm and performance metric. PLoS Comput Biol. 2022 Mar 11;18(3):e1009926.

[8]Ma S, Ogino S, Parsana P, Nishihara R, Qian Z, Shen J, Mima K, Masugi Y, Cao Y, Nowak JA, Shima K, Hoshida Y, Giovannucci EL, Gala MK, Chan AT, Fuchs CS, Parmigiani G, Huttenhower C, Waldron L. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. Genome Biol. 2018 Sep 25;19(1):142.