How doppelgänger effects in biomedical data confound machine learning

Wang, Li Rong; Wong, Limsoon; Goh, Wilson Wen Bin

2022

Wang, L. R., Wong, L. & Goh, W. W. B. (2022). How doppelgänger effects in biomedical data confound machine learning. Drug Discovery Today, 27(3), 678-685. https://dx.doi.org/10.1016/j.drudis.2021.10.017

https://hdl.handle.net/10356/155991

https://doi.org/10.1016/j.drudis.2021.10.017

How Doppelgänger Effects in Biomedical Data Confound Machine Learning

Li Rong Wang [1^], Limsoon Wong [2,3], Wilson Wen Bin Goh [4*]

1. School of Computer Science and Engineering, Nanyang Technological University, Singapore
2. Department of Computer Science, National University of Singapore, Singapore
3. Department of Pathology, National University of Singapore, Singapore
4. Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
5. School of Biological Sciences, Nanyang Technological University, Singapore

^ First Author

*Corresponding Author: Wilson Wen Bin Goh, wilsongoh@ntu.edu.sg

Address for correspondence/proofs:

Wilson Wen Bin Goh, PhD

Lee Kong Chian School of Medicine, Nanyang Technological University, 59 Nanyang Drive, Singapore 636921

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Email: wilsongoh@ntu.edu.sg, Tel: +65-65162902

SUPPLEMENTARY

MATERIALS AND METHODS

Kidney tissue benchmark proteomic datasets

To explore the implications of data doppelgängers on classifier performance, a benchmark proteomics data taken from the NetProt software library [1] was used. A brief description for the dataset is as follows:

The renal cell carcinoma (RCC) study of Guo et al. [2] comprises 24 proteomics runs originating from 6 pairs of non-tumorous and tumorous clear-cell renal carcinoma (ccRCC) tissues, 4 proteomics runs originating from a pair of non-tumorous and tumorous chromophobe renal carcinoma (chRCC) tissues and 8 proteomics run originating from two pairs of non-tumorous and tumorous papillary renal carcinoma (ppRCC) tissues, in duplicates.

Software

All codes for execution and graphics are written and executed in R and are available at https://github.com/gohwils/biodatascience/blob/master/talks/iSLS9_2021/DoppelgängerCode.zip.

Identification of data doppelgängers

PPCC data doppelgängers were identified between two datasets (in RCC, we identified PPCC data doppelgängers between both batches) with the following steps:

1. Both datasets were batch corrected with sva::ComBat.
2. PPCC was calculated between all sample pairs, with each sample in a pair from different datasets.
3. Each PPCC value was labelled with “Same Patient Same Class”, “Different Patient Same Class” and “Same Patient Different Class” and “Different Patient Different Class” to represent the positive cases, the valid cases and the negative cases between same and different samples respectively.
4. The distribution of PPCC values was visualized on a scatterplot against their meta-data labels.
5. The maximum PPCC value of any sample pair in the negative case, ignoring any anomalous points observed in step 4, was calculated. This value served as the lower bound for identifying PPCC data doppelgängers.
6. Sample pairs in the valid case with PPCC values greater than the lower bound calculated in step 5 were identified as PPCC data doppelgängers.

Confounding effects of data doppelgängers

Next, the confounding effect of PPCC data doppelgängers on classifiers was tested:

1. The batch corrected datasets were min-max normalised.
2. The following feature sets were generated: (All feature sets were of equivalent sizes of 10% of the total number of features):
    1) 10 randomly generated feature sets
    2) 1 feature set containing features of the highest variance
    3) 1 feature set containing features of the lowest variance

Feature sets were randomly generated to demonstrate that the model performs well regardless of training or in this case regardless of the feature sets it receives. Features were selected with reference to variance to observe how PPCC data doppelgängers affect properly and poorly trained models.

3. The data was partitioned to form the following training-validation sets, each training-validation set consisted of 8 samples in validation and 28 samples in training:
    1) 0 PPCC data doppelgängers in validation (4 normal samples none are PPCC data doppelgängers, 4 tumor samples none are PPCC data doppelgängers)
    2) 2 PPCC data doppelgängers in validation (4 normal samples half are PPCC data doppelgängers, 4 tumor samples none are PPCC data doppelgängers)
    3) 4 PPCC data doppelgängers in validation (4 normal samples half are PPCC data doppelgängers, 4 tumor samples half are PPCC data doppelgängers)
    4) 6 PPCC data doppelgängers in validation (4 normal samples all are PPCC data doppelgängers, 4 tumor samples half are PPCC data doppelgängers)
    5) 8 PPCC data doppelgängers in validation (6 normal samples all are PPCC data doppelgängers, 2 tumor samples all are PPCC data doppelgängers)
    6) Positive Control: Perfect leakage in validation (Duplicates of samples in Training)

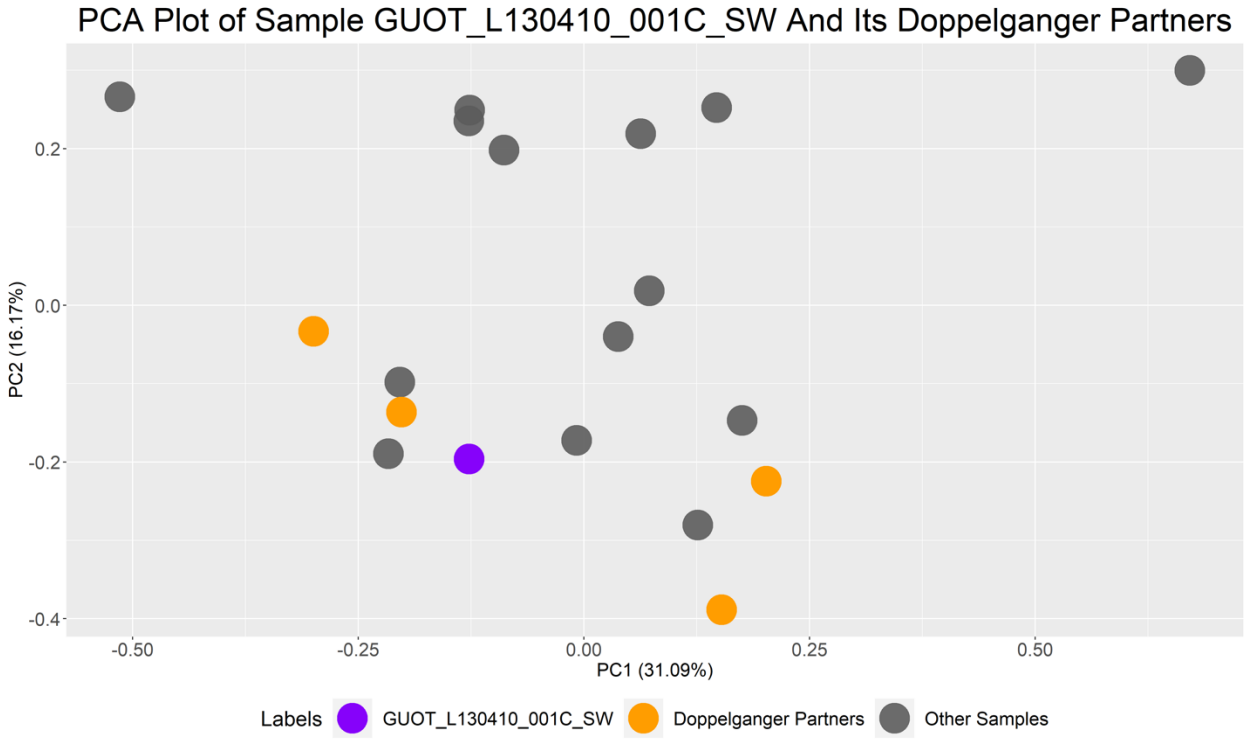
PPCC data doppelgänger samples used in the validation set were checked to be doppelgängers with at least 1 sample in the training set.

4. The number of correctly classified validation samples for each feature set (12 feature sets in total) was modelled with a binomial distribution with n = 8 (number of samples in validation) and p = 0.5 (probability of randomly

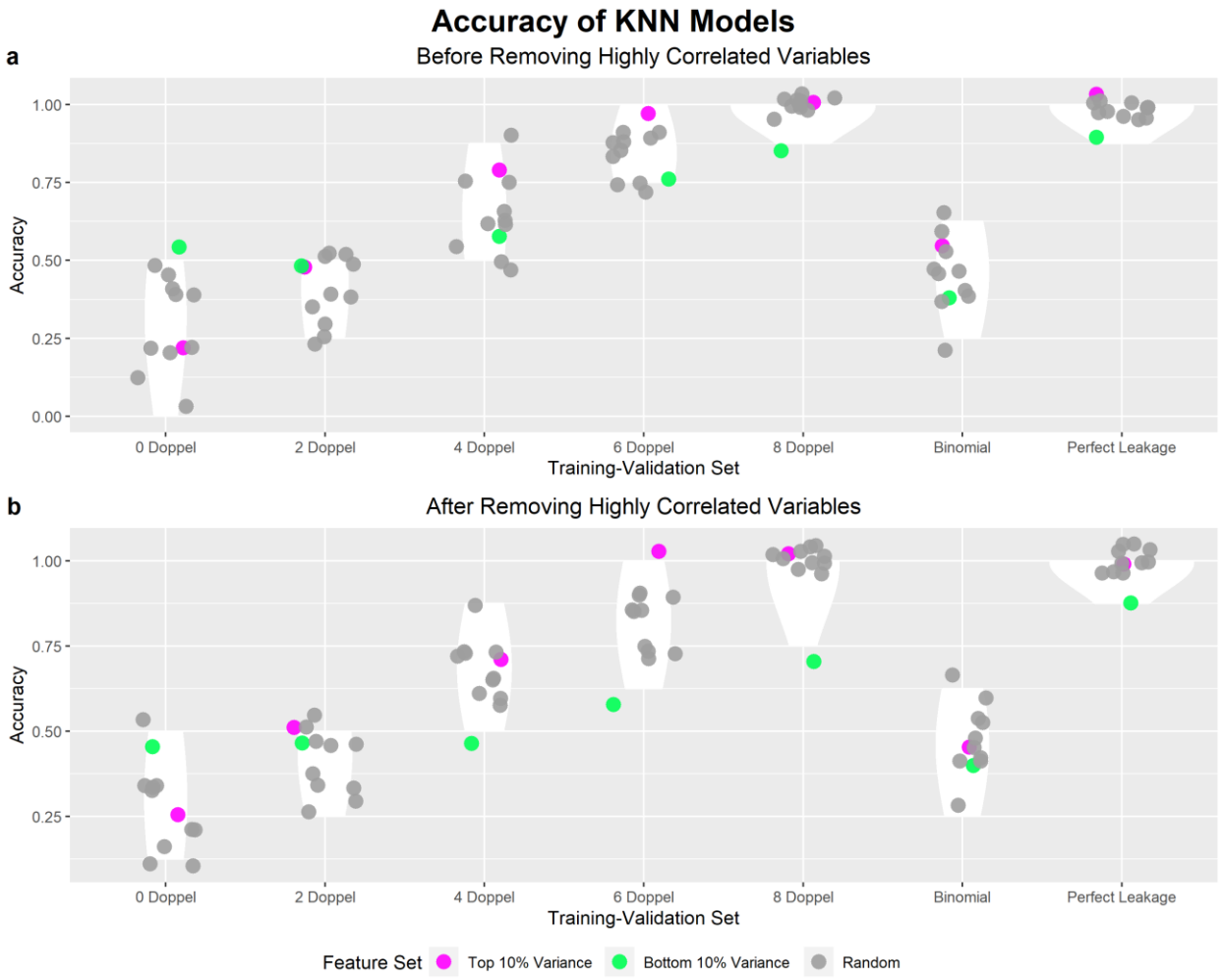
guessing the labels of a validation sample). This serves as a negative control for the experiment since a binary model trained on random signatures is expected to be equal in performance to a random coin toss.

5. Several machine learning models were trained and validated on the training-validation sets identified above with all feature sets identified previously. The validation accuracy of each model was recorded. The machine learning models trained includes K-Nearest Neighbours (KNN) (class package), Naïve Bayes (e10171 package), Logistic Regression (Stats package) and Decision Tree (rpart package) models.

SUPPLEMENTARY FIGURES

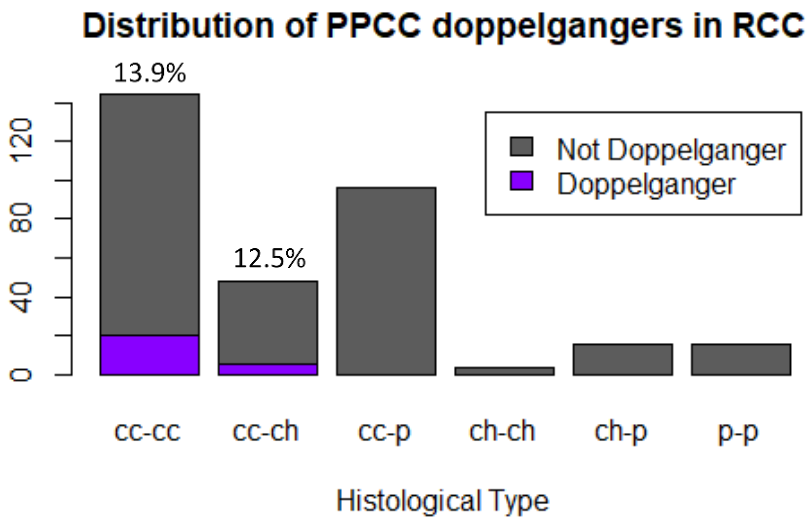


Supplementary Figure 1. Plot of PC1 against PC2 for GUOT_L130410_001C_SW and its PPCC data doppelgänger partners. GUOT_L130410_001C_SW is represented by the purple dot. The PPCC data doppelgänger partners of GUOT_L130410_001C_SW are represented by orange dots. The samples in dark grey represent other samples of different batch that are not PPCC data doppelgängers with GUOT_L130410_001C_SW. No visible distinction between PPCC data doppelgängers and non-PPCC data doppelgängers can be observed from the PCA plots above.



Supplementary Figure 2. Accuracies of KNN models before (a) and after (b) removal of highly correlated variables. X-axis: Type of validation set: "i Doppel" refers to a validation set with i number of PPCC data doppelgängers in the training set (where i = 0, 2, 4, 6, 8), "Binomial" refers to the accuracies generated by 12 (number of feature sets) binomial distributions with n=8 (since 8 samples in validation) and p=0.5 (probability of guessing the correct label for each validation sample) (negative control), "Perfect Leakage" refers to a validation set with 8 duplicates with the training set (positive control). Y-axis: Accuracy of machine learning models on a validation set of 8 samples with the lowest accuracy being 0 and the highest accuracy being 1. Legend: "Top 10% Variance" refers to the feature set comprising of proteins of the highest variance at 10% of the total number of proteins in the dataset, "Bottom 10% Variance" refers to the feature set comprising of proteins of the lowest variance at 10% of the total number of proteins in the dataset, "Random" refers to the feature set comprising of randomly select proteins at 10% of the total number of proteins in the dataset. The inflationary effects of PPCC data doppelgängers still remain after removal of these variables (seen from the presence of dosage-dependency and overall inflation of performance when PPCC

data doppelgängers were present in the validation set). Hence, removal of highly correlated variables is not a viable method of PPCC data doppelgänger amelioration.



Supplementary Figure 3. Distribution of PPCC data doppelgängers across different histological type pairs. Doppelgängers are not necessarily found amongst samples of the same histological type. 13.9% of clear cell RCC – clear cell RCC sample pairs (Both samples having clear cell RCC) and 12.5% of clear cell RCC – chromophobe RCC sample pairs (One sample having clear cell RCC and the other having the chromophobe RCC) are PPCC data doppelgängers.

SUPPLEMENTARY TABLES

Supplementary Table 1. Table of Metadata for RCC dataset.

Histological Type	Patient ID	Tissue	Sample ID	
			Replicate 1	Replicate 2
Clear Cell RCC	1	Normal	1	19
		Tumor	2	20
	2	Normal	3	21
		Tumor	4	22
	3	Normal	5	23
		Tumor	6	24
	6	Normal	11	29
		Tumor	12	30
	7	Normal	13	31
		Tumor	14	32
	8	Normal	15	33
		Tumor	16	34
Chromophobe RCC	5	Normal	9	27
		Tumor	10	28
Papillary RCC	4	Normal	7	25
		Tumor	8	26
	9	Normal	17	35

		Tumor	18	36

Supplementary Table 2. Table of Stratified Model Accuracies for ML Models Trained with "Top 10% Variance" Feature Set on the "Doppel 4" Training-Validation Set. This table shows how stratified accuracies (evaluating (non-) doppelgänger accuracy as the accuracy on the subset of validation data which are (not) doppelgängers)) could be used to evaluate the performance of a properly trained model (models trained with a non-random feature set) on PPCC data doppelgängers and non-PPCC data doppelgängers.

Models	Doppelgänger Accuracy	Non-Doppelgänger Accuracy
K-Nearest Neighbours	1	0.5
Naïve Bayes	1	0.5
Decision Tree	1	0.5
Logistic Regression	0.25	0

SUPPLEMENTARY DATA

Supplementary Data 1 The PPCC data doppelgängers (each sample is represented by their SWATH file names) and their PPCC values

Sample 1	Sample 2	PPCC
GUOT_L130410_001C_SW	GUOT_L130410_021_SW	0.941594524
GUOT_L130410_001C_SW	GUOT_L130410_023_SW	0.941730821
GUOT_L130410_001C_SW	GUOT_L130410_027_SW	0.95524861
GUOT_L130410_001C_SW	GUOT_L130410_033_SW	0.973659654
GUOT_L130410_002_SW	GUOT_L130410_030_SW	0.941557897
GUOT_L130410_003_SW	GUOT_L130410_019_SW	0.932562118
GUOT_L130410_003_SW	GUOT_L130410_023_SW	0.935647469
GUOT_L130410_003_SW	GUOT_L130410_029_SW	0.929630963
GUOT_L130410_003_SW	GUOT_L130410_031_SW	0.926934301
GUOT_L130410_005_SW	GUOT_L130410_019_SW	0.951640511
GUOT_L130410_005_SW	GUOT_L130410_029_SW	0.958783422
GUOT_L130410_005_SW	GUOT_L130410_031_SW	0.957194794
GUOT_L130410_009_SW	GUOT_L130410_019_SW	0.952804981
GUOT_L130410_009_SW	GUOT_L130410_021_SW	0.937719585
GUOT_L130410_009_SW	GUOT_L130410_023_SW	0.938754092
GUOT_L130410_009_SW	GUOT_L130410_029_SW	0.923238048
GUOT_L130410_009_SW	GUOT_L130410_033_SW	0.95548111
GUOT_L130410_011_SW	GUOT_L130410_019_SW	0.924191762
GUOT_L130410_011_SW	GUOT_L130410_023_SW	0.96586284
GUOT_L130410_011_SW	GUOT_L130410_031_SW	0.960803492

GUOT_L130410_013_SW	GUOT_L130410_023_SW	0.933678456
GUOT_L130410_013_SW	GUOT_L130410_029_SW	0.941877653
GUOT_L130410_014_SW	GUOT_L130410_020_SW	0.929264768
GUOT_L130410_014_SW	GUOT_L130410_030_SW	0.945079706
GUOT_L130410_015_SW	GUOT_L130410_023_SW	0.949797665
GUOT_L130410_015_SW	GUOT_L130410_029_SW	0.933455956

REFERENCES

1. Goh WWB, Wong L. NetProt: Complex-based feature selection. Journal of Proteome Research 2017; 16(8):3102--3112.

2. Guo T, Kouvonen P, Koh CC et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps, Nature medicine 2015;21:407-413.