Question 3

Data doppelgangers refer to two sets of independently obtained data that have a high degree of similarity with each other. A data doppelganger effect arises when models are tested with data that are very similar to its training dataset, causing the model to perform exceedingly well even if they are trained poorly. More specifically, data doppelgangers are functional only if they have a confounding effect on machine learning performance, hence data that are similar based on just measurements will not be considered to produce a doppelganger effect (Wang et al., 2022a). Quantitatively, data doppelgangers are sample pairs that have high mutual correlations and pairwise Pearson's correlation coefficient (PPCC) values. They may or may not be functional doppelgangers, but it has been shown that a subset of functional doppelgangers does possess high PPCC values as well (Wang et al., 2022b).

Doppelganger effects emerge when the validation accuracy of a model increases with the number of doppelgangers in the data, producing an inflationary effect on the model performance. Binomial distributions were used as negative control to simulate validation scores of random feature sets and served as a null model for comparison purposes. Fig. 1 below shows some models that were tested to understand if the identified PPCC data doppelgangers confounded model outcomes as well. Overall, there was a positive correlation between validation accuracy and PPCC data doppelganger numbers, thereby indicating that there was a doppelganger effect in the renal cell carcinoma dataset (Wang et al., 2022a).

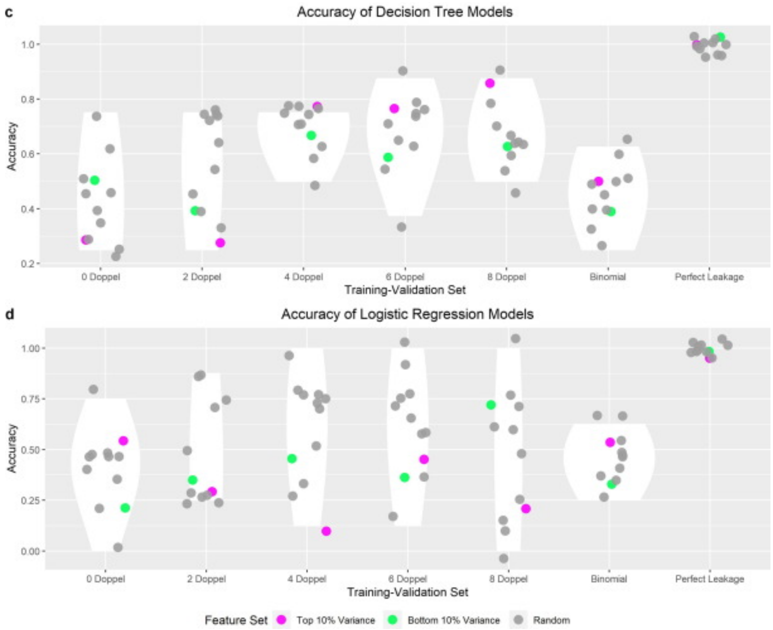

Fig. 1. Inflationary effects of PPCC data doppelgangers in renal cell carcinoma proteomics dataset.

The paper brought up two prominent examples of data doppelgangers in biological data. Firstly, during protein function prediction, it is common to see proteins with similar sequences having similar functions as well. However, there are proteins with less similar sequences that still carry out similar functions, an example being twilight-zone homologs (Wass & Sternberg, 2008). Hence, a model trained on just proteins with similar sequences may inflate its predictive performance.

Another example would be the prediction of a molecule's biological activities based on its structural properties, via the quantitative structure-activity relationship (QSAR) model (Paul et al., 2021). The doppelganger effect may occur when similar molecules with similar activities are split into both training and validation sets, as the model may not be trained on informative structural properties and yet still performs reasonably well (Cherkasov et al., 2014). Thus, it is crucial to identify and reduce such similarities as part of the model development workflow.

In addition to the above examples, doppelganger effects are also widely occurring in gene expression data. For instance, Wang et al. (2022b) investigated the confounding effect of identified PPCC doppelgangers in two RNA-sequencing datasets (derived from haematopoietic and lymphoid tissue-lung tumors (lymph_lung) and large intestine-upper aerodigestive tract tumors(large_upper)).

According to Fig. 2A, there was a noticeable positive relationship between the number of PPCC doppelganger samples and random model validation accuracy, thereby implying that majority of the PPCC data doppelgangers were functional doppelgangers that could potentially inflate model performance. A different trend was observed however in Fig. 2B for the large_upper dataset, where there was a slight decrease in accuracy as number of PPCC data doppelgangers decreased. Based on the above outcomes, there is evidence to show that the doppelganger effect does occur in gene expression data as well.

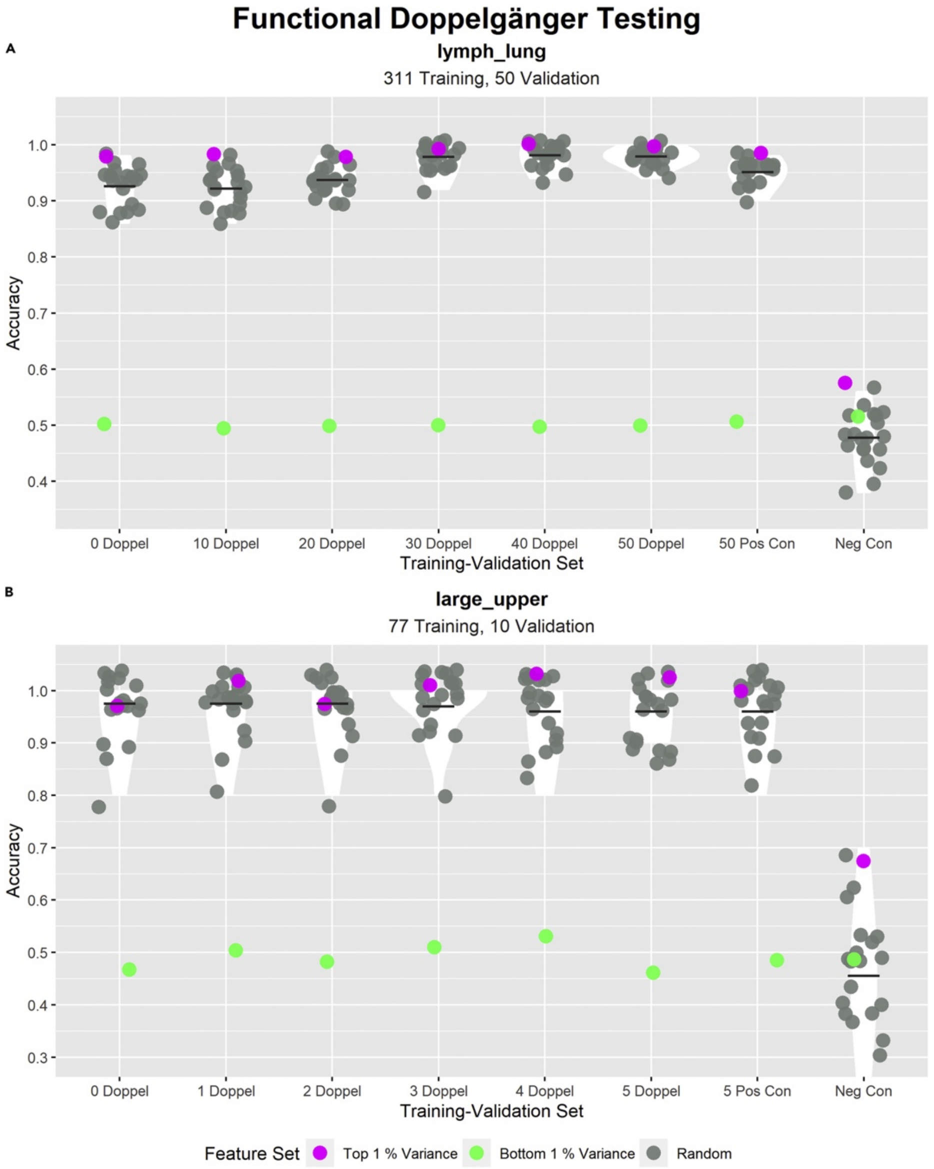

Fig. 2. Inflationary effects of PPCC data doppelgangers in lymph_lung and large_upper RNA-seq datasets

While there are many cases in the biomedical field where the doppelganger effect exists, it is not unique to this area. For instance, in natural language processing models that deal with sentiment analysis, independently derived data from two different sources may contain words that are used in common and more similar contexts (e.g. reviewing food). However, words may sometimes possess multiple meanings and how they are used within a sentence is also important, hence a model trained in just specific contexts may not perform well when interpreting words that are used in rarer situations with different connotations (Hussein, 2018). Generally, data doppelgangers can arise in any field as long as there is a probability of finding highly similar data in independent datasets.

To avoid this issue, it is vital to be able to accurately distinguish data doppelgangers within training, validation and test sets and remove them. The authors recommended several strategies for this. Firstly, stringent cross-checks should be made with the relevant metadata to ensure that training and test samples are not highly similar in nature. Data stratification was also proposed, where data is divided into strata of differing similarities and models are separately evaluated in each stratum. Lastly, independent validation checks with a large number of datasets could provide insights to how generalizable and objective the model is, even if they do not directly mitigate data doppelgangers.

One other useful way of avoiding the doppelganger effect, in the case of imaging datasets, is to perform image augmentation. Image augmentation is a process generates new images through manipulating the properties of the original image. It is typically used to improve generalizability of the model, and involves a series of random geometric transformations, flipping, rotation, translation, noise injection etc, such as those shown in Fig. 3. (Shorten & Khoshgoftaar, 2019) Implementing data augmentation on the training set could help reduce similarity in imaging data from the validation set, thereby decreasing chances of having a doppelganger effect. This should still be followed up by a functional doppelganger test before model validation to ensure that the model learnt the right information.

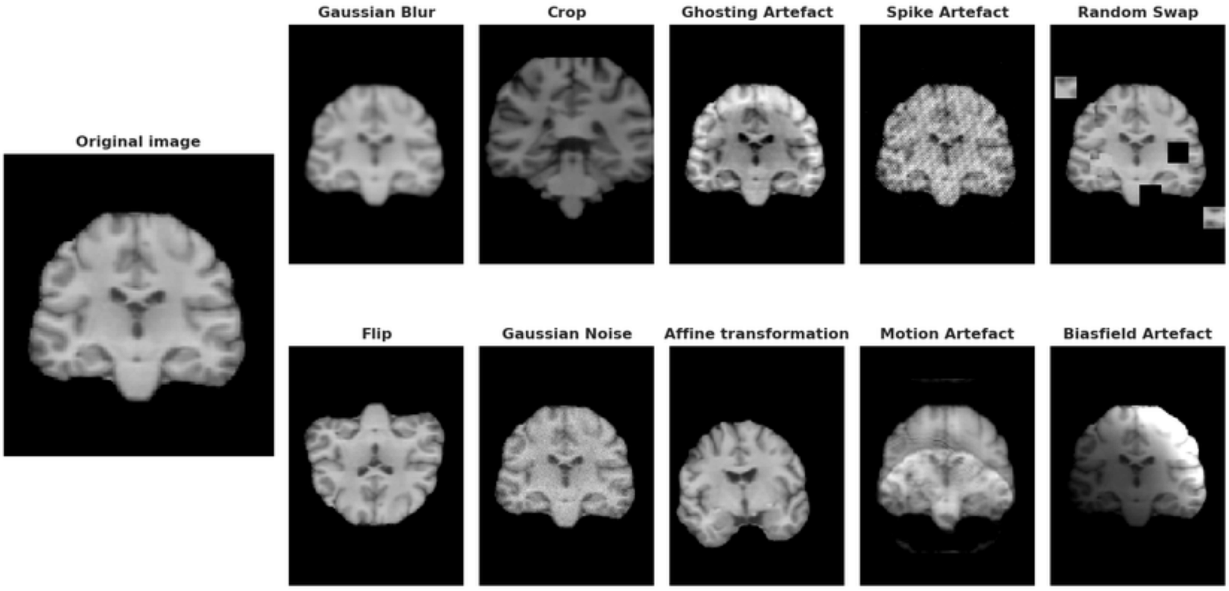

Fig. 3. Examples of data augmentation techniques used in MRI images (Dufumier et al., 2021).

With the advent of big data and large consortiums of biobanks, it will become increasingly crucial to check through these vast volumes of data. This is to ensure that the data used to train machine learning models are as unbiased as possible, given their critical role in driving model performance. Measures should be taken to incorporate doppelganger checks within the data processing workflow, in order to minimize inflated model performance and produce generalizable and high performing models.

References

Cherkasov, A. et al. (2014). QSAR modelling: where have you been? Where are you going to? *Journal of Medicinal Chemistry 57(2014), 4977-5010.* dx.doi.org/10.1021/jm4004285

Dufumier, B., Gori, P., Battalglia, I., Victor, J., Grigis, A. & Duchesnay, E. (2021). Benchmarking CNN on 3D Anatomical Brain MRI: Architectures, Data Augmentation and Deep Ensemble Learning. *Neuro Image.*

Hussein, D. M. E. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University – Engineering Sciences, 30(4), 330-338.* https://doi.org/10.1016/j.jksues.2016.04.002

Paul, D. et al. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today 26(1), 80-93.* https://doi.org/10.1016/j.drudis.2020.10.010

Shorten, C. & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data 6(60).* https://doi.org/10.1186/s40537-019-0197-0

Wang, L. R., Choy, X. Y. & Goh, W. W. B. (2022). Doppelganger spotting in biomedical gene expression data. *iScience, 25(8).* https://doi.org/10.1016/j.isci.2022.104788

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

Wass, M. N. & Sternberg, J. E. (2008). ConFunc – functional annotation in the twilight zone. *Bioinformatics 24(6), 798-806.* https://doi.org/10.1093/bioinformatics/btn037