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 [4,5*]

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

Teaser (35 Words)

We describe a doppelgänger effect in biomedical data. When doppelgängers are found across training and validation sets, machine learning performance becomes inflated and is no longer reliable. Doppelgänger mitigation is necessary for objective machine learning.

Abstract (115 words)

Machine learning models have been increasingly adopted in drug development for faster identification of potential targets. Cross-validation techniques are commonly used to evaluate these models. However, the reliability of such validation methods can be affected by the presence of data doppelgängers. Data doppelgängers occur when independently-derived data are very similar to each other, causing models to perform well regardless of how they are trained (doppelgänger effect). Despite the abundance of data doppelgängers in biomedical data and their inflationary effects, they still remain uncharacterized. We show their prevalence in biomedical data, demonstrate how doppelgängers arise, and provide proof of their confounding effects. To mitigate the doppelgänger effect, we recommend identifying data doppelgängers before the training-validation split.

Keywords

Computational Biology; Data Science; Doppelgänger Effect; Machine Learning

Introduction

Machine learning (ML) models have been increasingly used in drug discovery to speed up drug development. ML increases the efficiency of drug discovery in a multitude of ways: ML models could shortlist better drug candidates (targets) faster, reducing time spent on discovery and testing. ML models can also identify existing FDA-approved drugs for the treatment of other diseases (drug repurposing), dramatically decreasing the cost of drug development.[1] Both methods have shown promise in recent years. A new anticancer drug candidate, EXS21546, was discovered by Exscientia's 'Centaur Chemist' artificial intelligence platform after 8 months and is currently undergoing clinical testing (Clinical Trial: NCT04727138).[2] Several ML-identified drugs and drug combinations for COVID-19 treatment have also advanced into clinical trials, to name a few examples: 1/ The drug combination of melatonin and toremifene, identified by network-based approaches[3] (Clinical Trial: NCT04531748) and 2/ baricitinib, identified by BenevolentAI's knowledge graph[4] (Clinical Trials: NCT04373044 and NCT04401579).

Classification models based on ML and artificial intelligence also increase efficacy in drug development. Classifiers (trained models) have been used for the prediction of new drug-disease interactions [5,6] and possible adverse drug reactions.[7] Given the expensive drug testing process, it is important that these classifiers are properly trained and tested to identify suitable drug candidates.

It is well-established in ML that when assessing the performance of a classifier, the training and test datasets should be independently derived. However, independently derived training and test sets could still yield unreliable validation results. For example, models trained and validated on data doppelgängers (where training and validation sets are highly similar due to chance or otherwise) may perform well regardless of the quality of training.[8] When a classifier falsely performs well due to the presence of data doppelgängers, we say that there is an observed doppelgänger effect. Earlier, we say that data doppelgängers are when samples appear similar across their measurements. However, this does not guarantee a doppelgänger effect. And so, we say that data doppelgängers that generate a doppelgänger effect (confounding ML outcomes) are functional doppelgängers. Despite several documented examples of data doppelgängers (see **Abundance of data doppelgängers in biological data**), it remains uncommon to check if the sample training-evaluation pairs are independent and/or dissimilar. Furthermore, data doppelgängers and their accompanying downstream analytical effects (doppelgänger effects) are poorly documented and not well understood. Here, we would like to understand better what the level of similarity

between suspected functional doppelgängers is and what the acceptable proportion of functional doppelgängers in the validation set is. Though there exist several proposed methods of identifying data doppelgängers (see **Ameliorating data doppelgängers**), most methods were not generalizable or robust enough. Hence, it is imperative to investigate the nature of data doppelgängers and propose improved methods of doppelgänger identification. Using a renal cell carcinoma benchmark dataset with appropriately designed controls,[9] we illustrate here 1/ the prevalence of functional doppelgängers given this biomedical data, 2/ the implications of data doppelgängers on ML, and 3/ ways to mitigate the doppelgänger effect.

Abundance of data doppelgängers in biological data

Data doppelgängers have been observed in modern bioinformatics. In one notable case, Cao and Fullwood performed a detailed evaluation of existing chromatin interaction prediction systems.[10] Their work revealed that the performance of these systems has been overstated, due to problems in assessment methodologies when these systems were reported. In particular, these systems were evaluated on test sets that shared a high degree of similarity to training sets. The presence of data doppelgängers was also observed by Goh and Wong where certain validation data were guaranteed good performance given a particular training data, even if the selected features were random.[11] Data doppelgängers are also present in established fields of bioinformatics: In protein function prediction, proteins having similar sequences are inferred to be descended from the same ancestor protein and thereby inherited the ancestor's function (i.e., the two proteins are presumed similar in function.) This naïve application of abductive reasoning is true in most cases (cases of data doppelgängers), hence giving us a false impression of highly accurate predictions. However, on greater inspection, we realise that this approach would be unable to correctly predict functions for proteins with less similar sequences but similar functions; e.g. twilight-zone homologs[12] and enzymes that are dissimilar in sequence overall but with similar active site residues.[13] A similar example exists in drug discovery: Quantitative structure–activity relationship (QSAR) models are classification and regression ML models trained to predict the biological activities of molecules from their structural properties.[14] QSAR models assume that structurally similar molecules have similar activities. In most instances, this assumption is true (cases of data doppelgängers). Assorting similar molecules with similar activities into both training and validation sets (by chance during time-split validation or random test set selection)[15], confounds model validation since poorly-trained models (trained on uninformative structural properties) may still perform well on these molecules.[16] We can only

differentiate poorly-trained models from their well-trained counterparts through testing their performance on similar molecules with different activities (SAR paradox). If the reason for the paradox is due to small variations in structure that substantially impact binding affinity,[17] a well-trained model would theoretically still perform well on these instances given that they are trained on informative structural properties and hence able to detect these small variations while a poorly-trained model would fail to identify the true biological activity. While the biomedical data science community seems to be increasingly aware of such data doppelgänger problems, it is surprising that procedures for eliminating or minimizing similarity between test and training data still do not constitute standard practice prior to classifier evaluation.

Identification of data doppelgängers

Given the potential of doppelgänger effects to confound, it is crucial to be able to identify the presence of data doppelgängers between training and validation sets before validation. One logical approach to data doppelgänger identification would be to use ordination methods (e.g. PCA) or embedding methods (e.g. t-SNE), coupled with scatterplots, to see how samples are distributed in reduced-dimensional space. However, we found this method to be unfeasible as data doppelgängers are not necessarily distinguishable in reduced-dimensional space (**Supplementary Figure 1**).

Earlier studies working on similar problems have also proposed some measures for identifying data doppelgängers. One method, dupChecker identifies duplicate samples by comparing the MD5 fingerprints of their CEL files.[18] Identical MD5 fingerprints would suggest that samples are duplicates (essentially replicates, and therefore indicative of leakage issues). dupChecker therefore does not detect true data doppelgängers which are independently-derived samples that are similar by chance. Another measure, the pairwise Pearson's correlation coefficient (PPCC) captures relations between sample pairs of different datasets.[19] An anomalously high PPCC value indicates that a pair of samples constitutes PPCC data doppelgängers (note that it is impossible to determine which one between the pair is the original). While reasonable and intuitive, the prime limitation of the original PPCC paper was that it never conclusively made a link between PPCC data doppelgängers and their ability to confound machine learning tasks (i.e., having a functional effect and therefore acting as functional doppelgängers). We also realized during reanalysis of their data that their reported doppelgängers were in fact due to leakage (between sample

replicates), and therefore do not constitute true data doppelgängers. However, the basic design of PPCC as a quantitation measure is reasonable methodologically. And so, we will use this for identifying potential functional doppelgängers (from PPCC data doppelgängers) from constructed benchmark scenarios.

To construct benchmark scenarios, we used the renal cell carcinoma (RCC) proteomics data of Guo et al.[9] taken from the NetProt software library[20] (**Supplementary Methods**). RCC was chosen for its utility in constructing clear-cut scenarios: 1/ Negative cases, where doppelgängers are non-permissible by constructing samples pairs of different class labels, 2/ Valid cases, where doppelgängers are permissible by constructing sample pairs assigned to the same class label but from different samples. These effects can then be compared against 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 1a**). We simulate these scenarios across the two batches of the RCC dataset (**Supplementary Methods**).

We identified PPCC data doppelgängers based on the PPCC distribution of the valid scenario against the negative and positive scenarios. Surprisingly, we observed a high proportion of PPCC data doppelgängers (Half of the samples are PPCC data doppelgängers with at least one other sample, **Figure 2c**). PPCC distributions on the valid scenario exist as a wide continuum, without obvious breaks. This suggests that using outlier detection methods (as recommended in the original PPCC paper) will not be sensitive enough. It also suggests that data doppelgängers exist naturally as part of the similarity spectrum between samples (and are not spectacular anomalies). As for why this happens, we cannot say for sure if this is a problem due to PPCC itself, or due to the fact that transcriptional profile of genes is for the most part, positively correlated.[21] We checked PPCC distributions between same and different tissue pairs (**Figure 2b**). PPCC values for same tissue pairs remain high overall, suggesting high correlations between samples, even if they come from different patients. This is not surprising, seeing as how many genes share common regulators. However, PPCC distributions are assuredly are lower if we compare different tissue pairs where a class effect must also exist. In contrast, PPCCs are also extremely high when we consider replicates from the same sample or tissue. These evaluations suggests that PPCC has meaningful discrimination value.

Confounding effects of PPCC data doppelgängers

After identifying PPCC data doppelgängers in RCC, we explored their effects on validation accuracy across different randomly trained classifiers (a trained classifier is an ML model that has "learnt" from training data). This would determine if PPCC data doppelgängers act as functional doppelgängers (having an obvious inflationary effect on ML performances; **Supplementary Methods**).

We noted that the presence of PPCC data doppelgängers in both training and validation data inflates ML performance, even if the features are randomly selected (and therefore meaningless; in other words, the models should perform poorly during validation). This finding is consistently reproducible on different sets of training and validation data (**Figure 3**) and on different ML models. Moreover, the more doppelgänger pairs represented in both training and validation sets, the more inflated the ML performance. This points towards a dosage-based relationship between the number of PPCC data doppelgängers and the magnitude of the doppelgänger effect. When the validation accuracy for all properly trained models (with "Top 10% Variance" feature set) on "Doppel 4" (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 non-PPCC data doppelgängers (**Supplementary Table 2**). This result mirrors our earlier point regarding protein sequence function predictions---where there are many similar examples (many data doppelgängers), good accuracy is easily obtained without actually 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.

This result confirms that PPCC data doppelgängers (based on pairwise correlations) act as functional doppelgängers (confounds machine learning outcomes), producing inflationary effects similar to data leakage. The similarities between doppelgänger effects and leakage are evident in our experiment using KNN models where the training-validation set with 8 doppelgängers in validation showed an identical accuracy distribution to the training-validation set with perfect leakage (**Figure 3a**). However, not all models are equally affected: KNN and Naïve Bayes models have a clearer linear relationship between performance inflation and doppelgänger dosage than Decision Tree and Logistic Regression models.

By placing all doppelgängers in the training set, accuracies drop to approximately 0.5, which is the expected accuracy of a model trained on random signatures. Obviously, when all PPCC data doppelgängers are placed together in the training set, the doppelgänger effect is eliminated. This provides a possible way of avoiding the doppelgänger effect. However, constraining the PPCC data doppelgängers to either 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 may not generalize well since the model lacks knowledge. In the latter, you may end up with spectacular winner-takes-all scenarios (the doppelgängers will all either be predicted correctly or wrongly.)

Ameliorating data doppelgängers

Data doppelgängers produce undesirable inflationary effects on ML. This raises the question of how doppelgänger effects could be managed. In the previous section, we argued that enforced co-location of doppelgängers in either training or validation sets are suboptimal solutions. In Cao and Fullwood, they called for more comprehensive and rigorous assessment strategies, based on the particular context of the data being analysed.[10] This could be achieved by splitting training and test data based on individual chromosomes (instead of considering all chromosomes together), as well as using different cell types to generate the training-evaluation pair, thus establishing a good practice/standard in the field. However, this is very difficult to do practically as it predicates on the existence of prior knowledge and good quality contextual/benchmarking data.

In studies where the PPCC outlier detection package, doppelgangR, (see **Identification of data doppelgängers**) were utilised for the identification of doppelgängers, PPCC data doppelgängers could be removed to mitigate their effects.[22,23] However, this approach does not work on small datasets with a high proportion of PPCC data doppelgängers like RCC since the removal of PPCC data doppelgängers would reduce the data to an unusable size (as for what the prevalence of data doppelgängers across all instances of biomedical data is, we think this would be a consortium level effort, which may be necessary seeing as how doppelgängers effect poses a clear obstacle from good quality biomedical ML models).

Ergo, we also attempted to alleviate doppelgängers effect with methods that would not lead to a significant reduction in sample size or require a high amount of contextual data, albeit our attempts have met with failure thus far. For example, we attempted data trimming by removing variables contributing strongly towards data doppelgängers effects (see **Supplementary Figure 2**). However, we observed no change in the inflationary effects of the PPCC data doppelgängers after the removal of correlated variables. This observation hints at the extreme complexity of the doppelgänger effect, since the reason for high correlations between sample pairs cannot simply be

explained by a subset of highly-correlated variables. We are now looking towards novel feature engineering and normalization approaches. Hopefully, these may prove more successful.

Recommendations

Although removing data doppelgängers from data directly has proven elusive, we still need to guard against doppelgänger effects.

Our first recommendation is to perform careful cross-checks using meta-data as a guide. Here, we used the meta-data in RCC for constructing negative and positive cases. This allowed us to anticipate PPCC score ranges for scenarios where 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 warrants 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. In a similar vein, technical replicates arising from the same sample should also be dealt with similarly. This recommendation parallels guidelines in bioinformatics which suggest when ML models are trained on data derived from biological sequences, researchers should ensure that training and test samples are not duplicates or samples of high similarity[24].

Our second recommendation is to perform data stratification. Instead of evaluating model performance on whole test data, we may 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. 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. In RCC, the non-PPCC doppelgängers used in stratified performance assessment also happen to be papillary RCC samples. Since the proportion of kidney cancer cells of each tissue is known (papillary RCC makes up 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 which would require further improvements.

Our third recommendation is to perform extremely robust independent validation checks involving as many datasets as possible (divergent validation)[8]. 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 real-world usage) despite of the possible presence of data doppelgänger in the training set.

Future research could explore other methods of functional doppelgänger identification that do not rely heavily on meta-data. In such approaches, we may identify functional doppelgängers directly. For example, we may look for subsets of a validation set that are predicted correctly regardless of the ML method employed. 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 partners of test set samples in the training set (or conversely, the interesting question of whether dissimilar sets, or non-data doppelgängers can also act as functional doppelgängers). During model evaluation, these subsets should be avoided, as they act as functional doppelgängers, and do not give much insight on the relative performance of different models.

Conclusion

ML model performance is usually assessed through testing the model's accuracy on validation data. This approach towards model validation is only valid if validation data are independent from training data. However, this assumption is usually assumed to be true with no prior checks. This widely held assumption may not hold true in the presence of doppelgänger effects. We find that doppelgängers are fairly common in our test data, and that it has a direct inflationary effect on machine learning accuracy. This in turn, reduces the usefulness of ML for phenotype analysis and subsequent identification of potential drug leads. We also noted that the extent of this inflationary effect varies depending on two main factors: 1/ the similarity of functional doppelgängers, 2/ the proportion of functional doppelgängers in the validation set. Unfortunately, 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.

Author contributions

Wang LR implemented analyses and wrote the manuscript. LW provided critical feedback. WWBG supervised and co-wrote the manuscript.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

WWBG also acknowledges support from a Ministry of Education (MOE), Singapore Tier 1 grant (Grant No. RG35/20).

Competing interests

The authors declare no conflicting interests, financial or otherwise.

References

1. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing *The Lancet Digital Health.* 2020;
2. Savage N. Tapping Into the Drug Discovery Potential of AI. https://www.nature.com/articles/d43747-021-00045-7. Published 27 May 2021. Accessed 22 September 2021.
3. Cheng F, Rao S, Mehra R. COVID-19 treatment: Combining anti-inflammatory and antiviral therapeutics using a network-based approach *Cleve Clin J Med.* 2020;
4. Richardson P, Griffin I, Tucker C, Smith D, Oechsle O, Phelan A, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease *Lancet (London, England).* 2020; 395:e30.
5. Shi J-Y, Shang X-Q, Gao K, Zhang S-W, Yiu S-M. An integrated local classification model of predicting drug-drug interactions via Dempster-Shafer theory of evidence *Sci Rep.* 2018; 8:1-11.
6. Oh M, Ahn J, Yoon Y. A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions *PLoS One.* 2014; 9:e111668.
7. Hwang Y, Oh M, Jang G, Lee T, Park C, Ahn J, et al. Identifying the common genetic networks of ADR (adverse drug reaction) clusters and developing an ADR classification model *Mol Biosyst.* 2017; 13:1788-96.
8. Ho SY, Phua K, Wong L, Goh WWB. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability *Patterns.* 2020; 1:100129.
9. Guo T, Kouvonen P, Koh CC, Gillet LC, Wolski WE, Röst HL, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps *Nat Med.* 2015; 21:407-13.

10. Cao F, Fullwood MJ. Inflated performance measures in enhancer–promoter interaction-prediction methods *Nat Genet.* 2019; 51:1196-8.
11. Goh WWB, Wong L. Turning straw into gold: building robustness into gene signature inference *Drug Discov Today.* 2019; 24:31-6.
12. Wass MN, Sternberg MJ. ConFunc—functional annotation in the twilight zone *Bioinformatics.* 2008; 24:798-806.
13. Friedberg I. Automated protein function prediction—the genomic challenge *Brief Bioinform.* 2006; 7:225-42.
14. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development *Drug Discov Today.* 2020;
15. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders *Chem Soc Rev.* 2020; 49:3525-64.
16. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? *J Med Chem.* 2014; 57:4977-5010.
17. Chen Q, Wu L, Liu W, Xing L, Fan X. Enhanced QSAR model performance by integrating structural and gene expression information *Molecules.* 2013; 18:10789-801.
18. Sheng Q, Shyr Y, Chen X. DupChecker: a bioconductor package for checking high-throughput genomic data redundancy in meta-analysis *BMC Bioinformatics.* 2014; 15:323.
19. Waldron L, Riester M, Ramos M, Parmigiani G, Birrer M. The Doppelgänger effect: hidden duplicates in databases of transcriptome profiles *JNCI: Journal of the National Cancer Institute.* 2016; 108:
20. *NetProt: Complex-based feature selection* [computer program]. Journal of Proteome Research 2017; 16(8):3102--3112.
21. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome *PLoS Comput Biol.* 2011; 7:e1002240.
22. Lakiotaki K, Vorniotakis N, Tsagris M, Georgakopoulos G, Tsamardinos I. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology *Database.* 2018; 2018:bay011.
23. Ma S, Ogino S, Parsana P, Nishihara R, Qian Z, Shen J, et al. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis *Genome Biol.* 2018; 19:1-14.
24. Bioinformatics. Instructions to Authors (Machine learning). https://academic.oup.com/bioinformatics/pages/instructions_for_authors#General%20Policies. Published 2021. Accessed 22 September, 2021.
25. Muglia VF, Prando A. Renal cell carcinoma: histological classification and correlation with imaging findings *Radiologia brasileira.* 2015; 48:166-74.

Figure Legends

Figure 1. Diagram illustrating the PPCC data doppelgänger identification method (a) Naming convention for different types of sample pairs 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.

Figure 2. (a) Distribution of PPPC across different sample pairs. X-axis: Types of sample pairs grouped by the similarities of their patient and class. Y-axis: Pairwise Pearson's Correlation Coefficient (PPCC, Pearson's correlation coefficient between two samples). The 26 PPCC data doppelgängers are labeled in purple. **(b) Distribution of PPCC values of different sample pairs by the sample pair's histological types.** X-axis: Types of sample pairs grouped by histological type pairs. Clear cell RCC is represented as cc, chromophobe RCC is represented as ch and papillary RCC is represented as p. Y-Axis: Pairwise Pearson's Correlation Coefficient (PPCC). **(c) 26 PPCC data doppelgängers visualised as a graph.** Each node represents different sample, the first number stands for the patient number, the following alphabet represents the class ("N" stands for normal, "T" stands for tumor), the number after the hyphen represents the batch of the sample e.g. "5N_2" stands for the second replicate of a normal sample from the 5th patient. The presence of an edge between each node/sample means the two samples are PPCC data doppelgängers. There are 18 nodes in this graph (meaning 18 samples out of 36 total samples are doppelgängers with at least one other sample).

Figure 3. The prediction performance of different machine learning models on pairs of training-validation sets with varying numbers of PPCC data doppelgängers in the validation set. Machine learning models assessed includes: **KNN (a), Naïve Bayes (b), Decision Tree (c) and Logistic Regression (d)** models. 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 proteins of the highest variance i.e., top 10% amongst the total number of proteins in the dataset, "Bottom

10% Variance" refers to the feature set comprising proteins of the lowest variance at 10% of the total number of proteins in the dataset, "Random" refers to the feature set comprising randomly select proteins at 10% of the total number of proteins in the dataset.