Doppelganger Effects on Biomedical Data Sets

# Not a unique problem to biomedical data

Doppelganger effects are not unique in machine learning classifiers for biomedical data. Training the machine learning classifier to identify a person based on facial recognition would likely face similar issues as identical twins or very similar looking individuals and training handwriting recognition software may encounter similar issues as with identifying identical twins or very similar-looking individuals. This can be solved much easily in other data sets by filtering out possible confounders that contribute to doppelganger effects (for example, by excluding twins using metadata available such as social identity number) compared to biomedical data as the margin is thinner for segregating similar data sets that might contribute to doppelganger effects and justify their exclusion from the training data. Thus, biomedical data sets are more susceptible to inflated performance due to their nature of being high-dimensional and complicated.

# Biomedical data has high dimension and complexity

Machine learning algorithms used to train classifiers function generally on the concept of training and validation with independent datasets. Linkages are built using training data sets associated with a particular outcome and validated using independent data sets.

While training methods using direct, single-layered and multi-parametric data for machine learning classifiers have provided great results in areas such as identifying obstacles for self-driving vehicles, when applied directly to a multi-layered, multi-parametric data such as datasets used to predict protein function from peptide or gene sequences where similarity between data sets is high, the presence of functional doppelgangers can inflate the machine learning classifier's performance.

Biological systems should be viewed from a multi-layered perspective (high dimensional) compared with other datasets used in machine learning such as speech to text conversion and speech recognition, adding another layer of complexity when applying machine learning algorithms directly to solve biomedical problems.

Datasets used for identification of vehicle license plate and the examples given above can be viewed as simple single layered datasets with no unknown layers hidden between the visible data sets used to train the machine learning classifier and the outcome result. The machine learning classifier would only require to identify pattern of vehicle license plates and determine the vehicle license number - a straight forward learning process that can be viewed as single layered. The same can be said for building a classifier for identification of road obstacles such as a picture of a car as a car and a picture of a bicycle as a bicycle. Training and predicting using this methodology is more complicated when applied to biomedical datasets where we are often missing data layers between the training inputs and the training prediction output.

Biomedical data is also often re-analysed by the community to validate new hypotheses. This is made possible by availability of databases shared online. Other than proteomics data from NetProt software library having functional doppelgangers[1], Doppelgangers are also present in other biomedical data sets, such as transcriptomic data [2]. However, data sets may contain unintentional duplicates and this may add unintentional confounder(s) into machine learning models when used to train them.

# Reducing effects of functional doppelgangers

Strategies to reduce effects of functional doppelgangers on machine learning has been suggested [1]:

1. Cross check meta-data for identification of functional doppelgangers (such methods based on PPCC score cut-off) and clean up by checking file MD5 signatures for preventing data leakage by removal of unintended duplicated data in data sets.

2. Training based on data stratification strategies by by grouping data based on similarities in their groups and validating them based on the same similarities instead of training the whole data set as one training set. While applicable for stratifying into specific biological function, location and features in the biological systems (ie. gene localise to chromosome, cell surface receptors and protein interaction can take reference from biological pathways and related gene), this approach assumes the existence of identifiable strata in the real-world data being queried.

3. Validating with as many data sets as possible. This would provide useful information if the trained model is useful and accurate enough for use in real world data.

# Other proposed strategies to reduce effects of functional doppelgangers

Other than above suggested strategies to reduce the effect of functional doppelgangers on machine learning in biomedical data sets, we can try the following approaches:

1. training classifier based on context specific applications with data sets with parameter(s) directly correlates with predicted outcome. However, biomedical problems are often more complicated.
2. Using a higher resolution method (single cell analysis) to build data sets
3. Assigning weights to certain parameters known to have direct effects on the phenotype or assigning reduced weights to similar data suspected as a doppelgangers
4. Adding layers to machine learning classifiers, such as between genotype (input) and phenotype (predicted outcome) or adding a layer to predict protein folding and exposed motifs, which can subsequently be used to predict interactions with other exposed motifs.

## 1. Looking at appropriate data sets

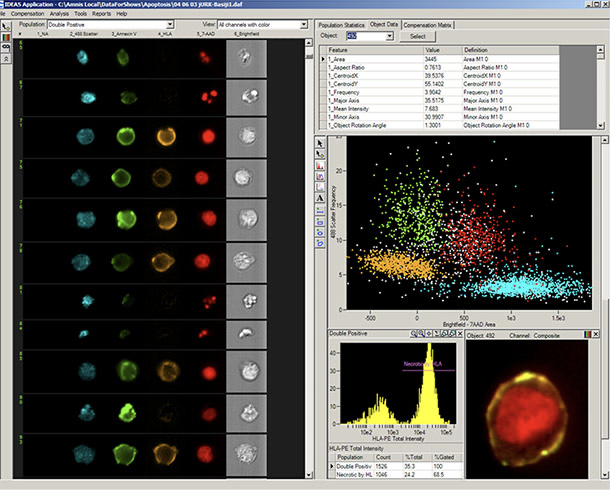
Predicting if an individual has good response to viral peptide stimulation (for example EBV peptide) and a functional read out of INF-g secretion by T cells of the donor – a multi-parametric approach in using surface biomarkers, TCR sequencing to determine T cell repertoire could be used to identify markers that could be useful for screening instead of genomic sequencing data.

## 2. Use a higher resolution data sets such as those from single cell methods

Considering on some context such as a cancer which is known to be heterogeneous in expression despite sometimes sharing the same genome (due to environmental factors such as hypoxia), the Data used to train machine learning classifiers that were not from the same source (i.e., single cell) could generated additional confounders. While not an issue when using data from pooled methods such as cDNA based microarray on cell lines that were subjected to the same conditions, non-single cell method data sets will introduce confounding factors if used in context of training classifier to identify cancer cells when cancer is heterogeneous in nature.

Identification for presence of cancer cells from histological section image multi-colour fluoresce microscopy with staining for informative surface markers to compliment the cell morphology information that could be extracted the histology slide. Identification of presence of cancer cells from haematoxylin and eosin stained histological slides alone is treating the problem with a single dimension / single layered perspective. This could lead to the machine learning algorithm picking up traits that is similar but not important to the predicted outcome.

Emerging tools such as imaging flow cytometry [3] allows for single cell identification of surface or internal cell stained biomarkers with the cell’s fluorescence image and associated forward scatter (associated with cell size) and side scatter (associated with cell complexity) to be captured at a single cell level may help elucidate the middle unknown layers from functional protein expression and pathways to morphology.



*Above* Example: Introducing a new dimension into an already high-dimension, high resolution (single cell level) method (flow cytometry) helps to link together layers [3].

## 3. Assigning weights

Much like how a weighted logistical regression model is constructed, the possibility of assigning weights to reduce impact of similar data sets that might be close to be classified as a doppelgangers but needed to be included in the training data set due to the scarcity of the data (ie. data from clinical trial). The same can be done to increase weights and emphasise on certain parameters known or proven to be influential to the phenotypic outcome that the machine learning classifier is trained to predict.

## 4. Adding additional layers to machine learning classifiers

The biological systems resulting phenotype is a complex environment with close interplay between genotype and their environment and biomedical datasets typically lack information on intermediates between genotype and phenotype.

When training machine learning classifier to identify protein function based on peptide sequence, there lack of information on how the protein is folded, which peptide is exposed for interaction with substrate or other proteins after protein folding. This is made even more complicated when the protein is multimeric (has multiple subunits) or perhaps require help to fold from a chaperone protein. The outcome of a machine learning classifier trained on the above datasets would simply predict protein function based on peptide sequence similarity. Missing layers such as this is a problem unique in biomedical data sets and not a problem that could be solved by implementing mathematical equation to substitute the missing layers.

*Above*: Missing layer not captured in data sets used to train machine learning classifiers would often directly affect outcome in biological systems

**References**

[1] Wang, Li Rong, Limsoon Wong, and Wilson Wen Bin Goh. "How doppelgänger effects in biomedical data confound machine learning." *Drug Discovery Today* 27.3 (2022): 678-685.

[2] Waldron, Levi, et al. "The doppelgänger effect: hidden duplicates in databases of transcriptome profiles." *JNCI: Journal of the National Cancer Institute* 108.11 (2016).

Bottom of Form

[3] Product information - <https://www.accela.eu/luminex/amnis-imagestream-x-mkii>