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 machine learning classifiers to identify a person based on facial features would likely face similar issues with identical twins or very similar-looking individuals.

Similarly, training handwriting recognition software may encounter similar issues as with identifying identical twins or very similar-looking individuals by facial feature recognition.

This can be solved much more easily in other datasets by filtering out possible confounders that contribute to doppelganger effects. For example, by excluding twins using metadata such as social identity number that is available. However, for biomedical data, the margin is thinner for segregating similar data sets that might contribute to doppelganger effects and to justify their exclusion from the training data. The possibility of associated meta-data contributing to unknown interactions with the phenotype cannot be excluded. Thus, biomedical data sets are more susceptible to inflated performance due to their high-dimensional and complicated nature.

# Biomedical data has high dimension and complexity

Machine learning algorithms used to train classifiers generally function on the concept of training to acquire an association between input parameters with a particular outcome and validation with independent datasets. 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, the presence of functional doppelgangers can inflate the machine learning classifier's performance when applied directly to multi-layered, multi-parametric data sets, such as those used to predict protein function from peptide or gene sequences where the similarity between data sets is high and association with unimportant features are unintentionally built.

Biological systems should be viewed from a multi-layered perspective (high-dimensional) compared to 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 the identification of vehicle license plates 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 resulting outcome result. For single-layered datasets, machine learning classifier would only need to identify the pattern of the vehicle license plates and determine the vehicle license number - a straightforward learning process that can be viewed as single layered. The same can be said for building a classifier for the 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 should be more complicated when applying to biomedical datasets where we are often missing data layers between the training inputs and the training prediction output.

# Reducing effects of functional doppelgangers

Biomedical data is also often re-analysed by the community to validate new hypotheses, made possible by the availability of databases shared online. Other than proteomics data from the NetProt software library shown to have functional doppelgangers [1], elements that contribute to doppelganger effects are also present in other biomedical dataset types, such as transcriptomic data [2].

Strategies to reduce effects of functional doppelgangers that inflate performance 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 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

Apart from the strategies suggested above to reduce the effect of functional doppelgangers on machine learning in biomedical datasets, we can try the following approaches:

1. Training a classifier based on context-specific applications with datasets with parameter(s) directly correlated with the predicted outcome. However, biomedical problems are often more complicated.
2. Using a higher resolution method (single-cell methods) to build datasets.
3. Assigning weights to certain parameters known to have direct effects on the phenotype or assigning reduced weights to similar data suspected as 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 from peptide sequence, which can subsequently be used to predict interactions with other exposed motifs.

## 1. Looking at appropriate data sets

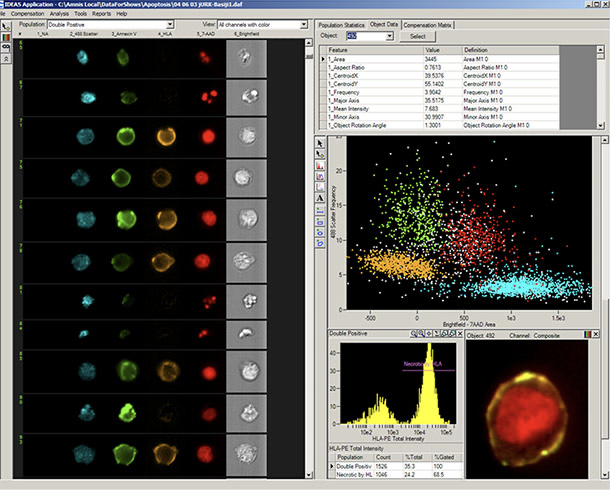
Predicting whether an individual has a good response to viral peptide stimulation (for example, EBV peptide) and a functional readout of INF-g secretion by T cells – a multi-parametric approach using surface biomarkers, TCR sequencing to determine T cell repertoire could be used to identify markers that 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 some contexts, such as 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 generate additional confounders even if the cells were from the same patient.

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 in laboratory setting, non-single-cell method datasets will introduce confounding factors if used in the context of training a machine learning classifier to identify cancer cells from patient when cancer is heterogeneous in nature.

Identification of the presence of cancer cells from histological section image with multi-color fluorescence microscopy staining for informative surface markers to complement the cell morphology information would provide more useful differentiating features that could help in reducing doppelganger effects. In contrast, identification of the presence of cancer cells from hematoxylin and eosin-stained histological slides alone treats the problem with a single-dimensional/single-layered perspective. This could lead to the machine learning algorithm picking up traits that are similar but not important to the predicted outcome – generating doppelganger effects. Emerging tools such as imaging flow cytometry [3] allow 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, which 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

Although not perfect, we can consider reducing the impact of data doppelgangers instead of eliminating them completely. Similar to how a weighted logistical regression model is constructed, we could assign weights to reduce the impact of similar data sets that might be classified as doppelgangers but needed to be included in the training data set due to the scarcity of data (such as data from clinical trials). We could also increase weights and emphasize certain parameters that are known or proven to be influential in the phenotypic outcome for training the machine learning classifier.

## 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>