Novel machine learning protocol for the phenotypic prediction of AAV variants library production

Dylan SERILLON[1], Tiziana LA BELLA[2], Anastasia BARKOVA[1], Jean Philippe BUFFET[1], David Del Bourgo[1], Ante MIHALJEVIC[2], Giuseppe RONZITTI[2], Julien COTTINEAU[1]

[1]WhiteLab Genomics, Future4Care, 8 Rue Jean Antoine de Baïf, 75013 Paris, France
[2]INTEGRARE, Genethon, INSERM, Université Evry, Université Paris-Saclay, 91002 Evry, France

ABSTRACT

Adeno-associated virus (AAV)-based capsid libraries are becoming increasingly popular as a candidate selection tool for gene therapy vectors with a recent advance in the study of the specific VR-VIII region.

However, the amount of available data regarding variants carrying multiple mutations outside this specific region remains poorly available and functionally unexplored

Shuffling of natural adeno-associated virus (AAV) allowed to create a specific library composed by 272 capsids, mutated in 43 different positions. We decided to build a machine learning models based on two different protocols:

- *(i)* Learning from the sequence, we generated numerical descriptors that describe the physico-chemical properties of the mutated residues.

- *(ii)* Based on the sequence, the AAV2 variants were partially reconstructed as a 3D object, and a machine learning algorithm was trained on the geometric data.

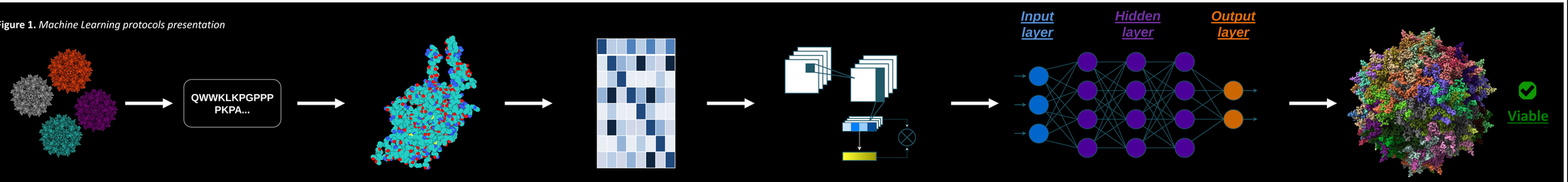
CONTACT

Dr. Dylan SERILLON
(PharmD, PhD)
dserillon@whitelabgx.com

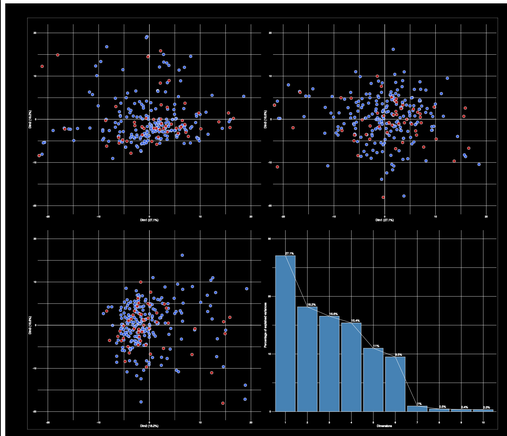
WhiteLab Genomics
https://whitelabgx.com/

MATERIAL & METHODS

Figure 1. *Machine Learning protocols presentation*



RESULTS



The space of the predictions also represents the molecular-space of the considered peptides

Figure 2. *Vizualization of the « **space** » of the predictions*

Controlling the machine learning model space is a challenging process, which allows for a little more rationality behind the classic 'black box' of artificial intelligence models. Our protocol is able not only to predict the capsid viability, but also to structure conformational changes to analyze physico-chemical variations.

STATISTICAL ANALYSIS

Table 1. *Performances of the machine learning models*

	Training-set	Test-set
	Sequential approach	
Accuracy	100 %	100 %
Sensitivity	100 %	100 %
Specificity	100 %	100 %
Prevalence	17.1 %	17.3 %
	Geometric approach	
Accuracy	98.5 %	100 %
Sensitivity	91.7 %	100 %
Specificity	100 %	100 %
Prevalence	17.1 %	17.3 %

- Prevalence : percentage of non-viable AAV variants in the datasets.
- Accuracy : global performances.
- Sensitivity : represents the capability for a model to predict non-viable variants.
- Specificity : capability to predict viable variants.

In total, more than a million models were generated using the variation of the tuning parameters, the cross-validation protocol, and different data partitions.

Both models have different ways to predict the same output and achieve powerful performances: respectively (*i*) 100% accuracy, (*ii*) and 98% accuracy with associated specificity (100%) and sensitivity (92%).

DISCUSSION

Figure 3. *AAV protein with highlighted mutated residues*

- Leading mutated residue
- Mutated residue

The models lead us to the identification of several positions which are predicted as fundamental for the stability of the viral capsid.

It is interesting to note that some of the residues that have been predicted to be very important for the model are located outside the well-known VR-VIII region

As a perspective, a list of interesting variants, which were mutated on the 43 positions identified as interesting by our AI protocols can be generated using rational approaches