Machine Learning for AAV Production-Fitness Modeling

Daniel H. Cox, Richard Lu, Hassan Gheisari, Netsanet Gebremedhin, Jiachen Liu, Mathieu Nonnenmacher, Weitong Chen

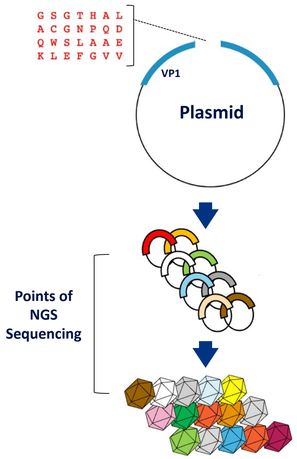
Voyager Therapeutics Inc., Lexington, MA, USA

INTRODUCTION

Adeno-associated viruses (AAV) are important vehicles for the delivery of gene therapies to target organs. Having control over their targeting is therefore of great therapeutic interest. Often, in the process of searching for AAV viruses with desirable tropisms, large libraries of mutant capsids are screened. The number of capsids screened in any one experiment, however, is typically far less than the sequence space being searched. Thus, it would be useful to restrict the search space at the outset to just those capsids that will produce viable virus. Here we present results from machine-learning models designed to aide in this process. They predict production-fitness from amino-acid sequence for AAV9 viruses carrying insertion mutations in the AAV capsid protein, VP1, in variable region 8. We demonstrate the performance of the models and show how they can be used as prescreening tools to build capsid libraries for tropism screening with high average production fitness and high diversity.

GENERATING MUTANT CAPSIDS

Figure 1. **Peptide Insertions in AAV9 VR8**



- ~100,000 random DNA sequences corresponding to 7 or 9 amino-acid peptides were cloned into a viral production plasmid.
- The resulting plasmid library of mutant capsids was then transfected into HEK293 cells where AAV9 mutant viruses were packaged.
- Next generation sequencing (NGS) was then used to quantify the amount of DNA for each capsid before and after viral production.

DATA FOR MACHINE LEARNING

Figure 2. **Next Generation Sequencing Counts Were Converted to Production Fitness**

	Sequence	Plasmid Counts	Virus Counts
0	ANWEIWY	131	72
1	CVYKWHM	91	4
2	INCYAEI	150	705
3	ECFYASI	167	27
4	QLFWIHK	79	165
...	...	...	...

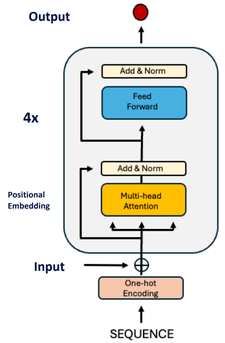
Production Fitness

	Sequence	Virus/Plasmid	Log2 virus.plasmid
0	ANWEIWY	0.303143	-1.721932
1	CVYKWHM	0.024244	-5.366229
2	INCYAEI	2.592289	1.374227
3	ECFYASI	0.089173	-3.487251
4	QLFWIHK	1.151974	0.204108
...	...	...	...

- NGS read counts for each plasmid and virus were converted to counts per million (cpm). Their ratio was taken, and the $\log_2$ of this ratio used as a measure of the quality of viral production and termed 'Production fitness'.

MACHINE LEARNING ARCHITECTURE

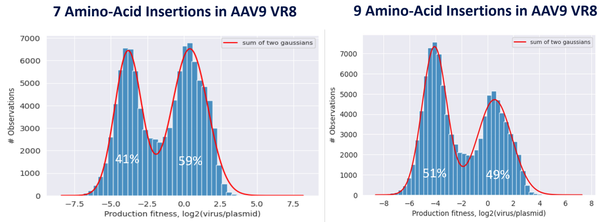
Figure 3. **A Transformer for Production Fitness Prediction**



- A Transformer encoder (left) was used for both regression and classification.
- Amino-acid seqs were one-hot encoded, passed through a positional embedding layer, and then passed through 4 transformer encoder blocks.
- Each encoder block had a multi-head dot-product self attention layer with 4 heads.
- A single node with either a linear (regression) or binary (classification) activation function was used to predict production fitness value or class.
- The training data consisted of ~ 64,000 sequences.
- The testing data consisted of ~ 20,000 sequences.
- Mean-squared error was used as the regression loss.
- Binary cross-entropy was used as the classification loss.

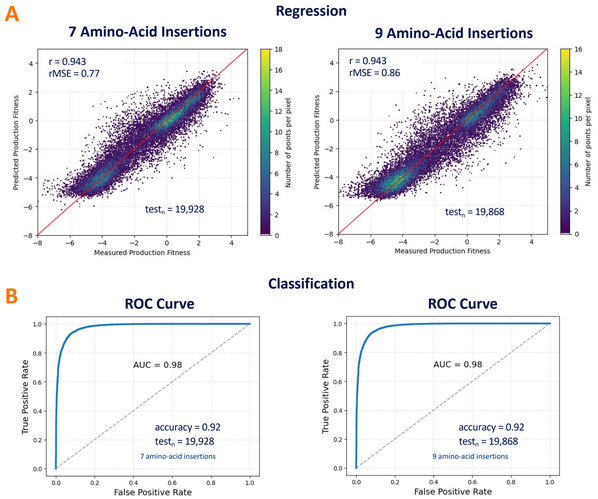
PRODUCTION FITNESS HISTOGRAMS

Figure 4. **Histograms of Production Fitness for 7 and 9 Amino-Acid Insertions**



MACHINE LEARNING

Figure 5. **Predicting Production Fitness**

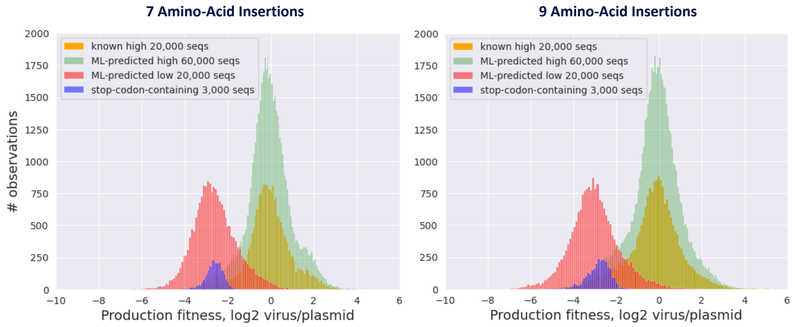


- (**A**) *Regression:* A transformer with a single linear output node was used to predict production fitness from sequence, (left) 7 amino-acid insertions, (right) 9 amino-acid insertions.
- (**B**) *Classification:* A transformer with a single binary output node was used to classify sequences into high or low production fit. High | Low threshold = -2.0. (left) 7 amino-acid insertions, (right) 9 amino-acid insertions,

AN EXPERIMENT TO TEST THE MODELS

- Capsid plasmid libraries were built that consisted of the following:
  - 20,000 known high-production fit sequences.
  - 60,000 ML-regression-model-predicted high-production fit sequences.
  - 20,000 ML-regression-model-predicted low-production fit sequences.
  - 3,000 sequences containing stop codons that should not produce well.
- The libraries were packaged into virus.
- Production fitness values were measured for each capsid via NGS sequencing before and after production.
- Distributions of these values were compared to model predictions.

Figure 6. **Results of Model Testing**



For both 7 and 9-amino-acid insertions:

- Sequences known to be high production fit were in fact so.
- Sequences predicted by ML to be high production fit were in fact so.
- Sequences predicted by ML to be low production fit were in fact so.
- Sequences carrying stop codons were low production fit as expected.

CONCLUSIONS

- We have developed machine-learning models as prescreening tools for the construction of AAV mutant capsid libraries that have a high proportion of capsids that produce virus well.
- After screening several architectures, we found that transformer networks are effective at this task.
- Our regression models predict production fitness values —for both 7 and 9 amino-acid insertions in AAV9 VR8— with a correlation coefficient between measured and predicted values of 0.943 and rMSEs of 0.77 and 0.86.
- Our classification models can distinguish between low and high production fit capsids with accuracies of 0.92 and AUC ROC values of 0.98.
- To test these models further, we generated new mutant-capsid libraries and compared the production fitness values of these capsids with those predicted by the models. The predictions were found to be accurate.
- We are now using these models to prescreen capsids for inclusion in larger libraries to be tested with our TRACER system for brain tropism.