VAEProp: A Robust Machine Learning Algorithm for Improving AAV Capsid Performance

David Brookes, Jeffrey Chan, **Farhan Damani**, Jeff Gerold, Rishi Jajoo, Stephen Malina, Alex Mijalis, Joyce Samson, Sam Sinai, Flaviu Vadan, Cameron Webster
Dyno Therapeutics, Inc., Watertown, MA

Digital copy

Two schools of sequence design

While integrating data with models presents a huge opportunity for improving AAV capsid properties, designing optimized variants for translational applications can be costly and difficult to measure accurately.

In vitro studies can recapitulate critical aspects of the AAV delivery problem and enable more rapid development of machine learning (ML) approaches due to faster feedback cycles and lower measurement noise.

Here, we validate an advanced approach to sequence generation that is also tunable, demonstrating the power of a coupled experimental and in silico platform and illustrating our improved ability to engineer AAV capsids for characterization in non-human primates (NHPs) and eventual translation to human medicines.

Method	Advantages	Disadvantages
Generative modeling, e.g., Variational Autoencoders (VAEs)	• Faithfully recapitulates important features in the data • High packaging efficiency • Better handling of noise	• Lower transduction potential • Unlikely to generate variants different from what's observed in the data
Regression-based explorers	• High transduction potential • Novel variant generation	• Low packaging efficiency • High false positive rate • Unlikely to work well with noisy measurements • Inefficient optimization

Table 1. Trade-offs between the two dominant paradigms for ML-based sequence design.

Biologically successful ML methods emerge from rapid in-vitro feedback

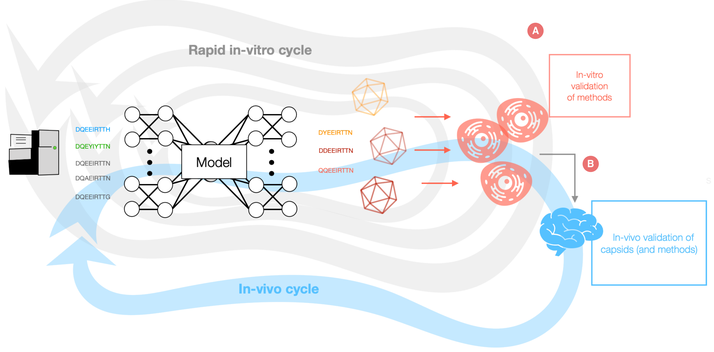

Figure 1. Fast feedback loop with in vitro design. (A) Experimental validation of machine learning using fast in vitro cycles. We use HEK293T cells in culture as a surrogate model to evaluate our machine learning methods. This surrogate enables rapid iteration of our machine learning design methods. (B) In-vivo validation. We design novel capsids for in vivo studies using our best machine learning design methods, which are first de-risked in the in vitro setting.

VAEProp: Best of both worlds

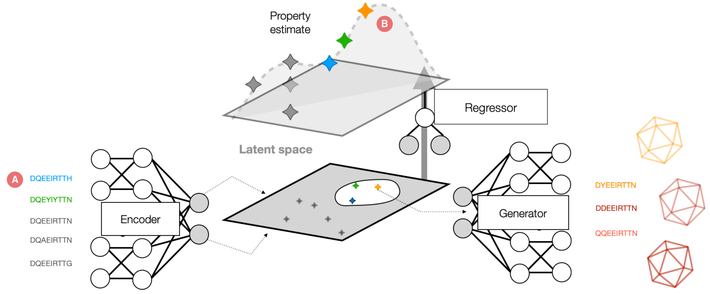

Figure 2. ML-guided design with VAEProp.
(A) The VAEProp architecture. VAEProp jointly trains a model to predict a target property (Regressor) and to reconstruct its original sequence (Encoder/Decoder).
(B) Sequence design using VAEProp. We design new sequences by searching in the low-dimensional, continuous latent space of the generative model for high scoring sequences. We stop the optimization from generating sequences that are too foreign by rejecting those below a likelihood threshold. The threshold is established by evaluating the likelihood of the training data itself under a model fit to the VAE embeddings. The new sequences are required to be at least as likely as the pth percentile of the training data, providing a tunable risk parameter.

Maintains high packaging efficiency

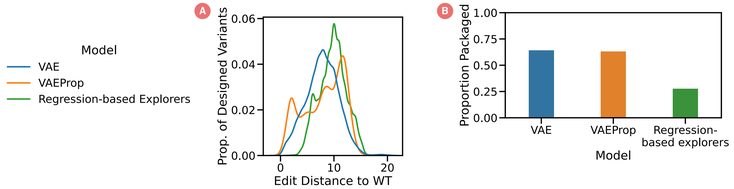

Figure 3. Packaging data. (A) VAEProp designs sequences at various edit distances while the VAE is biased towards the edit distance distribution of the training data and Regression-based explorers are biased towards high edit distance sequences. (B) VAEProp designed sequences that package at a similar rate to sequences designed by generative modeling and at a higher rate than sequences designed by Regression-based explorers.

Designs higher transducing variants compared to baselines

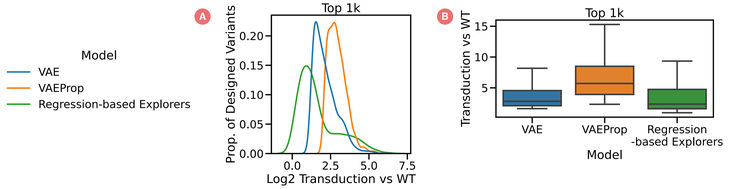

Figure 4. In vitro transduction data. We apply VAEProp to optimize the transduction of HEK293T cells in culture by designing variants of wild-type AAV9 in an area comprising of the VR-IV and VR-V loops of the VP3 region. Sequences designed by VAEProp have a higher transduction rate than sequences designed by a VAE or a Regression-based explorer. Each method received a budget of 8k variants.

VAEProp enables robust and controlled exploration

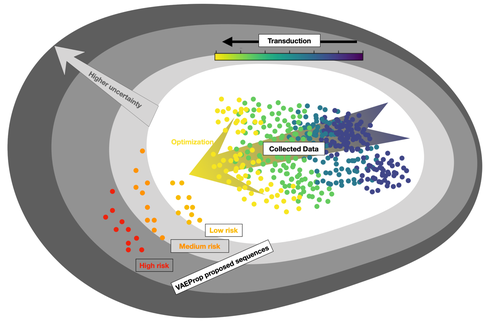

Figure 5. Sequence design using VAEProp. VAEProp allows the user to take risks between designing variants that are similar to those in the data or venturing deeper into the sequence space for potentially higher gains. We consider three flavors of our risk parameter to test the performance of the approach in conservative and high-risk explorations.

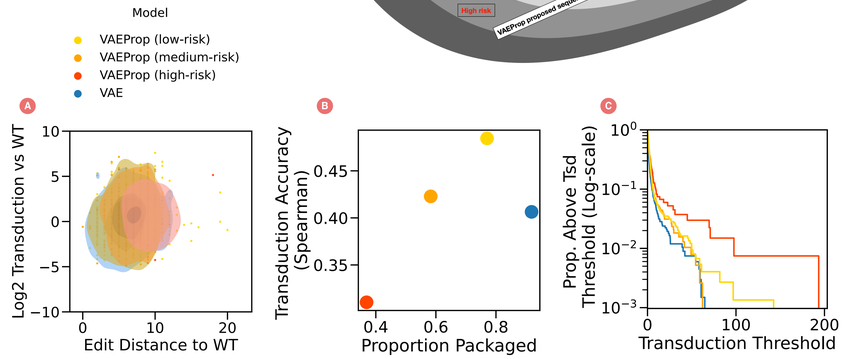

Figure 6. Experimental validation of risk-reward tradeoff. As a follow-up in vitro transduction experiment, we evaluated how this risk parameter translates to different packaging and transduction rates. Each method received a budget of 2k variants. (A) High risk setting translates to designed sequences with higher edit distance to wildtype. (B) Transduction rates for high risk sequences are harder to predict (y-axis) and result in fewer packaged variants (x-axis). (C) High-risk sequences have higher transduction upside on the tail of the distribution despite lower predictive accuracy.

Impact

- VAEProp is a machine learning method for protein design that combines the best aspects of generative modeling and regression-based explorers and experimentally validated with AAVs using a fast experimental transduction feedback loop. Here we show the results for validation in cell culture, and experiments are now ongoing in NHPs
- This method enables controlled exploration by exposing a risk parameter that the user can tune depending on the scenario, including adapting it for experimental scenarios with differential time and resource costs
- VAEprop is one of the methods that enables Dyno to deliver high-performing capsids optimized across multiple properties to our partners towards improving the safety and efficacy of gene therapy products

References

• Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.
• Bryant, Drew H., et al. "Deep diversification of an AAV capsid protein by machine learning." Nature Biotechnology 39.6 (2021): 691-696.
• Ogden, Pierce J., et al. "Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design." Science 366.6469 (2019): 1139-1143.