Generative AAV capsid diversification by latent interpolation

Sam Sinai[1,2,3], Nina Jain[2,3], George M Church[2,3,*], Eric D Kelsic[1,2,3,*]

*Corresponding: church_lab_admin@hms.harvard.edu, eric.kelsic@dynotx.com

Affiliations:

1. Dyno Therapeutics, Cambridge, MA
2. Wyss Institute for Biologically Inspired Engineering, Boston, MA
3. Dept. of Genetics, Harvard Medical School, Boston, MA

Summary

Adeno-associated virus (AAV) capsids have shown clinical promise as delivery vectors for gene therapy. However, the high prevalence of pre-existing immunity against natural capsids poses a challenge for widespread treatment. The generation of diverse capsids that are potentially more capable of immune evasion is challenging because introducing multiple mutations often breaks capsid assembly. Here we target a representative, immunologically relevant 28-amino-acid segment of the AAV2 capsid and show that a low-complexity Variational Auto-encoder (VAE) can interpolate in sequence space to produce diverse and novel capsids capable of packaging their own genomes. We first train the VAE on a 564-sample Multiple-Sequence Alignment (MSA) of dependo-parvoviruses, and then further augment this dataset by adding 22,704 samples from a deep mutational exploration (DME) on the target region. In both cases the VAE generated viable variants with many mutations, which we validated experimentally. We propose that this simple approach can be used to optimize and diversify other proteins, as well as other capsid traits of interest for gene delivery.

Introduction

Protein engineering is of increasing importance in modern therapeutics. Natural proteins evolved to satisfy biological requirements for their organisms, however, some of these properties are also of use for therapeutic purposes. The Adeno-associated virus (AAV) makes a great example: this non-pathogenic virus is currently a primary candidate to be used as delivery vector in gene therapy, and is the delivery mechanism of choice in FDA-approved gene therapies (Dunbar et al., 2018; Grimm et al., 2008). However, as a naturally occurring virus, it hasn't adapted to satisfy all of the properties that are desirable for a vector. For instance, many human sera have neutralising capability against the virus (Calcedo et al., 2009; Mingozzi et al., 2015). In order to alleviate these problems, a first step would be to generate a diverse set of sequences that maintain the ability to assemble and package DNA. We assume that more distant and diverse functional sequences compared to the circulating populations of natural variants increase the probability of finding sequences that for instance, can avoid natural immunity (Bryant et al., 2021). However, finding sequences of high diversity can be challenging for reasons that we outline below.

High-throughput DNA synthesis technologies allow for direct design of thousands of sequences for experimental validation (Adachi et al., 2014; Bryant et al., 2021; Ogden et al., 2019).

Previously, high-throughput diversification was largely performed through random mutagenesis and DNA shuffling (Sarkisyan et al., 2016; Yang et al., 2019). These methods have the drawback of extremely low yield as a large majority of random mutations result in non-functional capsids in the selection phase (Bryant et al., 2021; Ogden et al., 2019). On the other hand, low throughput “rational design” approaches were used for precise design of a small number of variants that generally yield a higher proportion of functional variants (Havlik et al., 2020; Maurya et al., 2019). However, direct high-throughput synthesis scales better than what is reasonable for rational design. This provides an opportunity for modern computational approaches to step in as an intermediate where many sequences can be automatically designed in high throughput, with far better yield than random mutations.

Classical computational approaches to protein design based on biophysical first principles have enjoyed substantial success in the past decade (Huang et al., 2016). However, more recently, data-driven machine learning efforts for protein engineering have seen a rapid uptick, targeting protein structure and (to a lesser degree) function prediction (Alley et al., 2019; AlQuraishi, 2019; Hiranuma et al., 2021; Hopf et al., 2017; Marks et al., 2011; Norn et al., 2021; Rao et al., 2019; Senior et al., 2020). These efforts often make use of large protein databases to learn general rules that govern protein sequences within families and are at times followed up by fine-tuning on specific sequences. For function prediction, available mutational scanning datasets can then be used to evaluate the ability of such models to predict the effects of small perturbations(Hopf et al., 2019; Riesselman et al., 2018).

In the case of AAV, the standard approaches both in computational design as well as transfer learning through ML can come with drawbacks. The AAV capsid consists of a 60-subunit assembly of viral protein (VP) monomers, making accurate biophysical modeling challenging. Previous attempts in rational computational design of AAVs include building a phylogenetic tree and using that as a reference to reconstruct ancestral viruses(Zinn et al., 2015), though the resulting sequences did not display improved immune evasion. On the other hand, computational evolutionary (unsupervised) methods for predicting function are unknown to be effective in this setting as the number of known homologues in the dependo-parvovirus family is relatively small (<1e3).

We recently conducted a high-throughput study with a supervised ML approach to diversify AAV capsids (Bryant et al., 2021). This method shows promise for supervised design of AAV capsids It builds upon mutational scans in the sequence space near the original capsid. However, as a supervised method, it makes no use of evolutionary information. Similarly, Marques et al. use supervised methods to predict AAV capsid assembly from plasmid library data (Marques et al., 2021). In this work, we investigate the power of using simple unsupervised methods, including Variational Auto-Encoders (VAE) to extract useful information out of evolutionary data together with deep mutational explorations. In this case, a deep mutational exploration consists of randomly and additively sampled variants with up to 23 substitutions. This is a subset of the training set used in (Bryant et al., 2021) restricted to substitutions only, and with training samples that we estimate as likely to assemble and package (no negative labels). In particular, we are interested in diversifying the AAV capsid using this information. The promise of

combining evolutionary and assay-labeled data has also been observed in other contexts (Hsu et al., 2021; Wittmann et al., 2021).

Previous work has shown that VAEs are effective in capturing the effects of mutations when trained on evolutionary data, but most of these results are applied to small edit-distances (Riesselman et al., 2018). They have also been shown to learn relevant latent space structure that captures phylogenetic relationships and propose new proteins (Ding et al., 2019; Greener et al., 2018). In this work, we focus our efforts on generating variants of AAV2 VP3 protein that are able to successfully assemble and package (we refer to this trait as viability), a prerequisite for any downstream engineering task, using evolutionary and experimental data. We select a 28 amino acid window near the heparin binding site, for which we have collected data previously and which is known to be of immunological significance (Tseng and Agbandje-McKenna, 2014). The MSA consists of 564 sequences from AAV2-related strains. The experimental data we use consists of 36,562 variants of AAV2 (substitution only) mutants of which 22,704 we classified as viable and were designed by an in-silico random model or based on a single site mutation model (see supplement for details).

We investigate three partitions of data, and two unsupervised models. Our model consists of an independent-sites model (IS) as baseline, and a small VAE model (1e5 parameters) with two latent variables with the same architecture as in (Sinai et al., 2017). Our datasets are (i) Evolutionary data (MSA) alone (n=564) (ii) Evolutionary data supplemented 667 randomly generated viable variants (MSAr) from a deep mutational exploration (n=1,231) and finally (iii) Evolutionary data augmented by 22,704 viable (MSA+) mutants 1-21 mutations from wildtype generated at random or an additive model (n=23,268). Producing variants without access to any labeled data would have been risky as the MSA consisted of only a few hundred sequences. Having the deep mutational exploration data allowed us to also test the model's score (calculated by reconstruction probability) on a set of 7150 holdout variants (both viable and non-viable), to ensure that the VAE captures information about the variants. Both models are trained in an unsupervised manner.

To generate variants with the IS model, we constructed a position weight matrix (PWM) for the sequence and sampled each column proportional to the normalized probability of each amino acid occurring (ignoring gaps). We then sorted the variants according to their log reconstruction score and selected the top 1250 variants for each model-dataset pair (7500 in total for both models). To generate sequences from the VAE, we induced posterior distributions by generating a grid of radial latent coordinates for the VAE centered at the center of mass of the evolutionary data, and then decode the variants (i.e., emitting conditional distributions using generators seeded with the latent code). We then sampled this distribution similarly to how we sampled the IS PWMs, and we always included the maximum likelihood sequence for every conditional (local) PWM. This approach was used to investigate the ability of the VAE to interpolate between known viable regions in the latent embedding. We sorted the variants based on the model score and included the top 1250 variants per pair (See Fig 1 for a schematic).

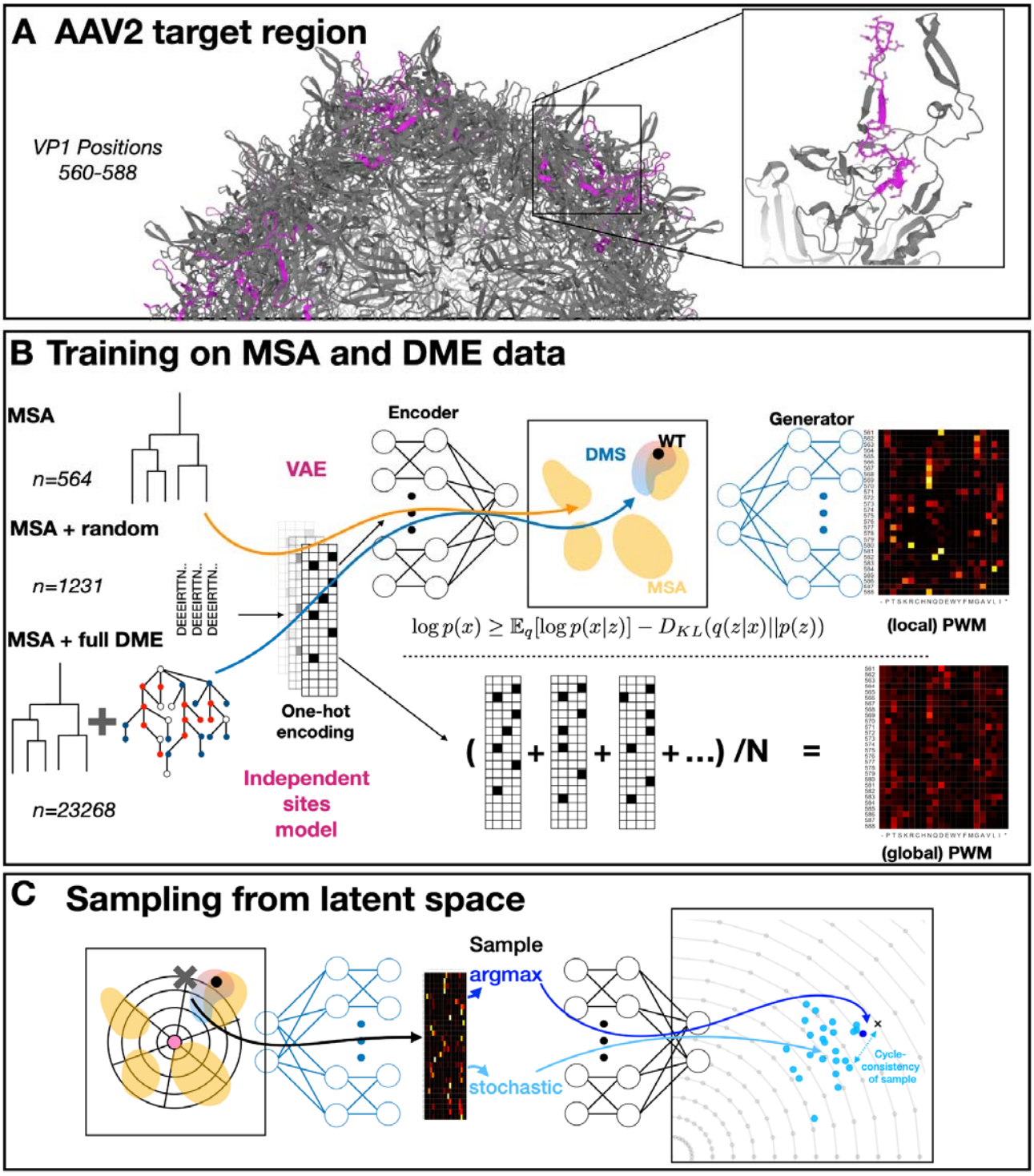


Figure 1. Overview of the generative process. (A) Targeted region (in purple) between VP1 positions 560-588 includes buried and exposed capsid segments with a variety of secondary structures and is positioned at a 3-fold axis contact point crucial for cell tropism, immune evasion and assembly. (B) Our training process on MSA+DME data. The data is one-hot encoded and passed through a VAE as well as the IS model. Both methods generate PWMs from which samples are drawn stochastically with a fixed temperature. VAEs are optimized against an objective that jointly minimizes KL-divergence between

encoding q(x|z) and a latent standard normal prior p(z) with the expected log reconstruction error of the output. This is known as the Evidence Lower BOund (ELBO), see (Kingma and Welling, 2013)(C) Sampling from the VAE latent space involves starting at latent coordinates, inducing a conditional PWM, and then sampling from it stochastically, as well as by taking the argmax per position. As a sanity check we passed each sample through the encoder and confirmed that they are generally correlated to the original point.

We compared the experimental production score of our training set variants with log reconstruction probability of the model in generating those variants. On the holdout samples, we found high correlation between IS and VAE model scores (Spearman ρ~0.95 for all models trained on the same training set), that is the models highly agree. We also found moderate correlation for the model scores and experimental holdout data (Spearman ρ (i) IS: 0.41, VAE:0.47, (ii) IS: 0.48, VAE:0.5, (iii) IS, VAE = 0.55). We deemed this sufficient to use the models for in-vitro validation. Our cloning and production assays were done exactly as described in (Bryant et al., 2021) with additional details explained in the SI. In short, we calculate a selection score $s_i = log_2(f_i{}^{virus}/f_i{}^{plasmid})$ where f_i denotes the frequency of a variant in plasmid or virus pool measured through next generation sequencing (NGS). Overall, our experiments show excellent correlation among replicates ($r^2 = 0.98$, see SI Fig 1).

Results

Our primary objective in this study was to explore the sequence space for diverse and viable variants in response to different models and datasets. We classify our viral production into three categories: Viral variants likely arising from dysfunctional genes (observed in production due to low frequency cross-packaging into viable capsids) termed *non-viable*, viral variants likely capable of generating functional assemblies termed *viable*, and viral variants that are present at very high counts, termed *viable*+ (See more detailed discussion of this classification in the SI).

Encouragingly, we found more than 100 viable variants with 25 or more substitutions in the 28-aa region, a majority of which came from the VAE designs. Of the top 10 variants produced, 9 were proposed by VAEs, and 1 by IS-MSA+. Of the top 100 scoring variants produced, 70 were proposed by the three VAE models, and the rest were proposed by IS-MSA+ (see Supp Fig 2). Overall, the IS model trained on the MSA+ dataset performed best in terms of yield with 92.6% of proposed sequences being viable (See supplementary table 1) however the yield was a result of picking the top (per-site) mutations present in the training data, and the farthest designed variant in this set was only 5 mutations away from the WT. The IS model's performance deteriorated with the removal of the large mutational scan data. For the VAEs the yield was somewhat consistent among the datasets, with the VAE-MSAr pair performing marginally better than the other methods. Even the shallow MSA dataset alone was sufficient to produce a large number of viable variants with many mutations away from the reference sequence (Fig 2E). This is in part an indication of modular interchangeability among dependo-viruses, despite significant sequence-level variation (See SI Fig 2), but as indicated above, IS models were insufficient in generating these viable variants.

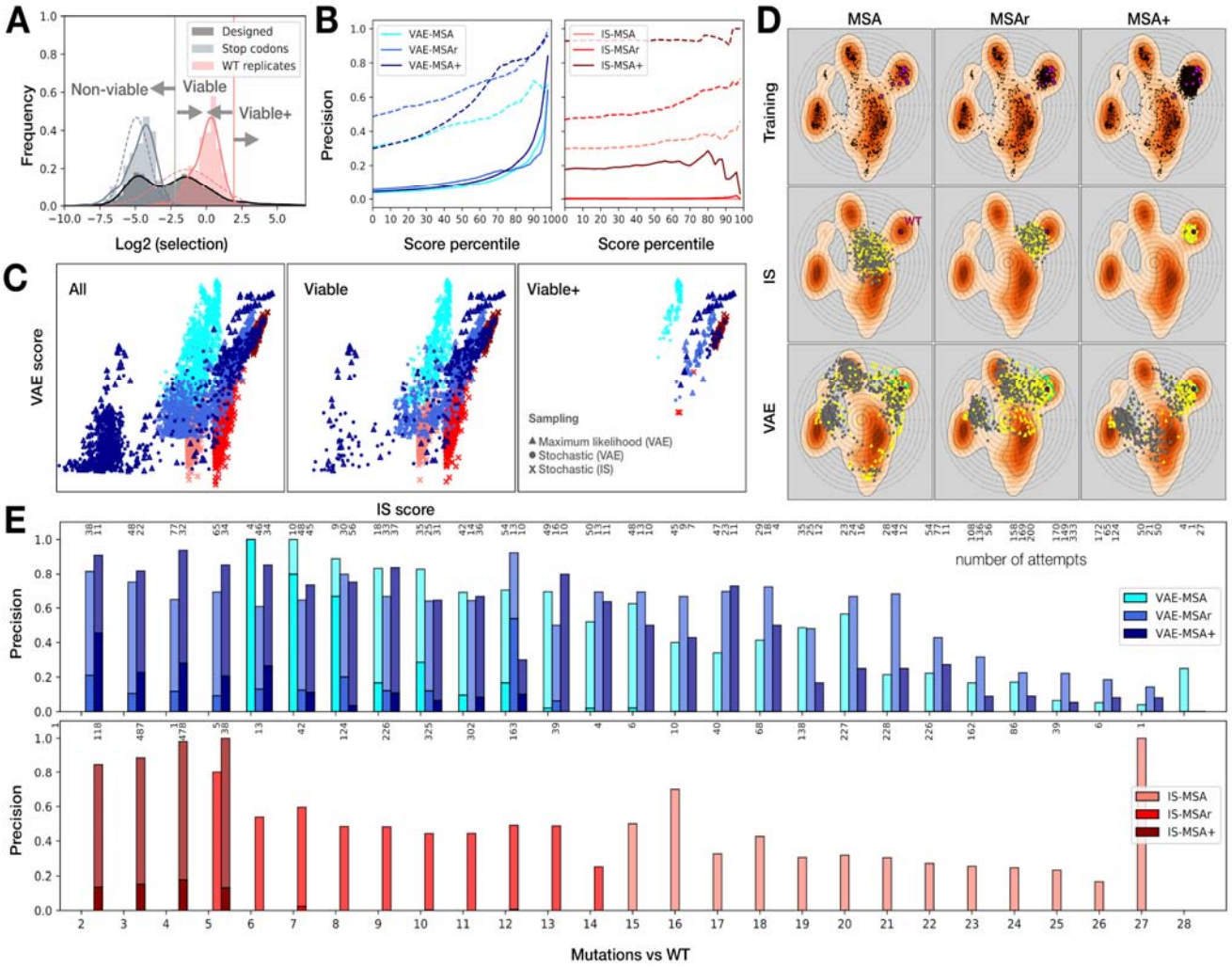


Figure 2. Successful generation of variants with VAE and IS models. (A) Classification of variants into non-viable, viable and viable+ is done by conservative decision bounds computed based on the distribution of the two Gaussian distributions (approximated by dashed lines), representing viable and non-viable distributions (see supplement for details). These are largely in agreement with WT and stop-codon replicates from the Bryant et al. study shown for comparison. Variants with selection score above 3 standard deviation of WT average and 97 percentiles of the Gaussian fit for successful producers are classified as viable+. (B) The relationship between model-denoted score and probability of success. The x-axis denotes the percentile score for the variant based on the model's evaluation. Dashed lines denote viable fraction, solid lines denote viable+. Precision denotes fraction successful by the given criteria (viable or viable+). (C) Comparison between the IS and VAE scores for variants. The first panel includes all variants, the second panel viable variants, and the third panel viable+. Triangles denote argmax sampling, circles denote stochastic sampling from VAE posterior, x denote stochastic sampling from IS prior. Colored as in part B. (D) Distribution of sampled variants projected onto the VAE-MSA latent dimensions. Orange contours are density plots for evolutionary data alone (points in the first panel). WT denoted by a purple dot. Magenta points show projected location of AAV1-12 onto the same space (top row). Grey(non-viable), yellow(viable), Green (viable+) points in the bottom two rows show the projection of experimental samples onto the latent coordinates. Circles are stochastic samples, whereas triangles are maximum likelihood (argmax). (E) Distribution of samples with viable (lighter bars) and viable+ (darker

bars) variants at each distance sampled by each algorithm. The number of attempted samples is denoted at the top of each column. Precision denotes the fraction of those attempts that resulted in success.

A key observation however is that both the IS and VAE models performed significantly better in terms of yield when “fine-tuned” through augmentation with the mutational scan data. As shown in Fig 2B, adding more data increases the quality of the IS model substantially. For the VAE however, it appears that the intermediate MSAr dataset is comparable to the full MSA+ for model quality despite having 10-fold fewer samples in its training data. Comparing how the two models score variants we find that every viable+ variant scores highly with both models (although a large majority are sampled by a VAE), however, some viable variants proposed by the VAE score poorly with the IS model (when trained on the same dataset, Fig 2C). Curiously, high-scoring VAE variants were far more likely to be in the viable+ category (Fig 2B, C).

Overall, our generative models manage to produce a significant number of viable capsids with more than 50% of positions substituted, with multiple viable variants that had substitutions at almost every position. This is not a trivial task: Most random changes in the amino-acid space within this region result in a non-functional capsid (Bryant et al., 2021).

Discussion

We mentioned in previous segments that the predictive accuracy of the IS and VAE models is very similar. But the sampling schemes introduced here exhibit drastic differences in both the quality and the diversity of samples between VAEs and the IS model. Yet, both methods perform decently as generative methods, with far better yield than random sampling as seen in (Bryant et al., 2021; Ogden et al., 2019). The performance of the IS model is consistent with previous studies where independent effect terms explain a large fraction of trait variability (Otwinowski et al., 2018). Furthermore, our VAE model can be thought of as a mixture model, with each set of initial coordinates emitting a single sequence profile, within which additive effects explain the majority of variation (Dauparas et al., 2019; Marshall et al.).

While IS models and VAEs share similarities, VAEs seem much better suited to use for generative interpolation. We are encouraged by the observation that even a shallow MSA is sufficient to generate interesting and high-performing variants with the VAE, and furthermore, adding only a few hundred randomly sampled variants (achievable through simple error-prone PCR instead of direct synthesis) drastically increases the performance of these models. This suggests an experimental and computationally accessible framework can be used to generate thousands of viable variants, both for AAVs and other proteins of interest.

This work stands to benefit from further investigations in multiple directions. First, in training these models, we ignored the full sequence (i.e., background) in which these AA substitutions were being made. We chose a segment that could be directly synthesized and was of sufficient challenge and clinical relevance. However, at least the VAEs have a clear possibility of learning

from the context and improving their predictions as a result. Furthermore, for both IS and VAE models, we use a fixed temperature for sampling variants. It would be interesting to tune the temperature as a method of generating further diversity. The IS model can also be augmented with clustering to generate local sequence profiles, increasing diversity in the samples. In addition, while we kept the latent dimensions of the VAE limited to 2 to simplify interpolation, it is a tunable hyperparameter, and we expect that higher capacity models could be used to perform more sophisticated sampling with better-fit models. Our study is a proof-of-concept for augmenting MSAs with experimental measurements with generative models, and was conducted in a fully unsupervised manner. However, it is also possible to conduct training in a semi-supervised fashion, where available labels are used to help the VAE learn relevant latent representation. Finally, generative models like those in this study can be paired with supervised methods to optimize sequences toward particular traits (Sinai and Kelsic, 2020; Sinai et al., 2020). Brookes et al. develop such a method which uses a VAE together with supervised oracles to optimize GFPs *in-silico* (Brookes and Listgarten, 2018; Brookes et al., 2019). Our study indicates an increased probability of success for these methods to be applicable in AAV design. We show that even without the need for deep mutational data, VAEs have potential as generative approaches for designing proteins that are hard to model structurally (like capsids). This should open avenues for low-cost and fast computational design with much higher yield than random mutagenesis, in a manner accessible to many wet lab scientists.

Acknowledgements

We thank Javen Tan with assistance in software development. We also thank Debora Marks for assistance with the alignments. We thank David Ding and Martin Nowak for related discussions. We thank Jakub Otwinowski, Elina Locane, Farhan Damani, Michael Stiffler, Jeffrey Gerold from Dyno Therapeutics for helpful comments on the manuscript. We thank members of the Church lab for their support, in particular Surge Biswas, Gleb Kuznetsov, and Pierce Ogden. This work was supported by the Wyss Institute.

Code and Data Availability

Code and data for this paper can be accessed here:
https://github.com/churchlab/Generative_AAV_design

Conflict of Interest

EK, NJ, SS, GMC performed research while at Harvard University and EK, SS also performed research while at Dyno Therapeutics. EK, SS, and GMC hold equity at Dyno Therapeutics. A full list of GMC's tech transfer, advisory roles, and funding sources can be found on the lab's website: http://arep.med.harvard.edu/gmc/tech.html. Harvard University has filed a patent application for inventions related to this work.

References

Adachi, K., Enoki, T., Kawano, Y., Veraz, M., and Nakai, H. (2014). Drawing a high-resolution functional map of adeno-associated virus capsid by massively parallel sequencing. Nat. Commun. *5*, 3075.

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods *16*, 1315–1322.

AlQuraishi, M. (2019). End-to-End Differentiable Learning of Protein Structure. Cell Syst *8*, 292–301.e3.

Brookes, D.H., and Listgarten, J. (2018). Design by adaptive sampling.

Brookes, D., Park, H., and Listgarten, J. (2019). Conditioning by adaptive sampling for robust design. In Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri, and R. Salakhutdinov, eds. (PMLR), pp. 773–782.

Bryant, D.H., Bashir, A., Sinai, S., Jain, N.K., Ogden, P.J., Riley, P.F., Church, G.M., Colwell, L.J., and Kelsic, E.D. (2021). Deep diversification of an AAV capsid protein by machine learning. Nat. Biotechnol.

Calcedo, R., Vandenberghe, L.H., Gao, G., Lin, J., and Wilson, J.M. (2009). Worldwide epidemiology of neutralizing antibodies to adeno-associated viruses. J. Infect. Dis. *199*, 381–390.

Dauparas, J., Wang, H., Swartz, A., Koo, P., Nitzan, M., and Ovchinnikov, S. (2019). Unified framework for modeling multivariate distributions in biological sequences.

Ding, X., Zou, Z., and Brooks, C.L., Iii (2019). Deciphering protein evolution and fitness landscapes with latent space models. Nat. Commun. *10*, 5644.

Dunbar, C.E., High, K.A., Joung, J.K., Kohn, D.B., Ozawa, K., and Sadelain, M. (2018). Gene therapy comes of age. Science *359*.

Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. (2015). HMMER web server: 2015 update. Nucleic Acids Res. *43*, W30–W38.

Greener, J.G., Moffat, L., and Jones, D.T. (2018). Design of metalloproteins and novel protein folds using variational autoencoders. Sci. Rep. *8*, 16189.

Grimm, D., Lee, J.S., Wang, L., Desai, T., Akache, B., Storm, T.A., and Kay, M.A. (2008). In vitro and in vivo gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. J. Virol. *82*, 5887–5911.

Havlik, L.P., Simon, K.E., Smith, J.K., Klinc, K.A., Tse, L.V., Oh, D.K., Fanous, M.M., Meganck, R.M., Mietzsch, M., Kleinschmidt, J., et al. (2020). Coevolution of Adeno-associated Virus Capsid Antigenicity and Tropism through a Structure-Guided Approach. J. Virol. *94*.

Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., and Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. Nat.

Commun. *12*, 1340.

Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. *35*, 128–135.

Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy, A., Brock, K., Riesselman, A.J., Palmedo, P., et al. (2019). The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics *35*, 1582–1584.

Hsu, C., Nisonoff, H., Fannjiang, C., and Listgarten, J. (2021). Combining evolutionary and assay-labelled data for protein fitness prediction.

Huang, P.-S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320–327.

Kingma, D.P., and Welling, M. (2013). Auto-Encoding Variational Bayes.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766.

Marques, A.D., Kummer, M., Kondratov, O., Banerjee, A., Moskalenko, O., and Zolotukhin, S. (2021). Applying machine learning to predict viral assembly for adeno-associated virus capsid libraries. Mol Ther Methods Clin Dev *20*, 276–286.

Marshall, D., Wang, H., Stiffler, M., Dauparas, J., Koo, P., and Ovchinnikov, S. The structure-fitness landscape of pairwise relations in generative sequence models.

Maurya, S., Mary, B., and Jayandharan, G.R. (2019). Rational Engineering and Preclinical Evaluation of Neddylation and SUMOylation Site Modified Adeno-Associated Virus Vectors in Murine Models of Hemophilia B and Leber Congenital Amaurosis. Hum. Gene Ther. *30*, 1461–1476.

Mingozzi, F., Büning, H., Basner-Tschakarjan, E., and Galy, A. (2015). Immune responses to AAV vectors, from bench to bedside (Frontiers Media SA).

Norn, C., Wicky, B.I.M., Juergens, D., Liu, S., Kim, D., Tischer, D., Koepnick, B., Anishchenko, I., Foldit Players, Baker, D., et al. (2021). Protein sequence design by conformational landscape optimization. Proc. Natl. Acad. Sci. U. S. A. *118*.

Ogden, P.J., Kelsic, E.D., Sinai, S., and Church, G.M. (2019). Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. Science *366*, 1139–1143.

Otwinowski, J., McCandlish, D.M., and Plotkin, J.B. (2018). Inferring the shape of global epistasis. Proceedings of the National Academy of Sciences *115*, E7550–E7558.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y.S. (2019). Evaluating Protein Transfer Learning with TAPE. Adv. Neural Inf. Process. Syst. *32*, 9689–9701.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic

variation capture the effects of mutations. Nat. Methods *15*, 816–822.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. Nature *533*, 397–401.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710.

Sinai, S., and Kelsic, E. (2020). A primer on model-guided exploration of fitness landscapes for biological sequence design. arXiv Preprint arXiv:2010.10614.

Sinai, S., Kelsic, E., Church, G.M., and Nowak, M.A. (2017). Variational auto-encoding of protein sequences. arXiv Preprint arXiv.

Sinai, S., Wang, R., Whatley, A., Slocum, S., Locane, E., and Kelsic, E. (2020). AdaLead: A simple and robust adaptive greedy search algorithm for sequence design. arXiv Preprint arXiv.

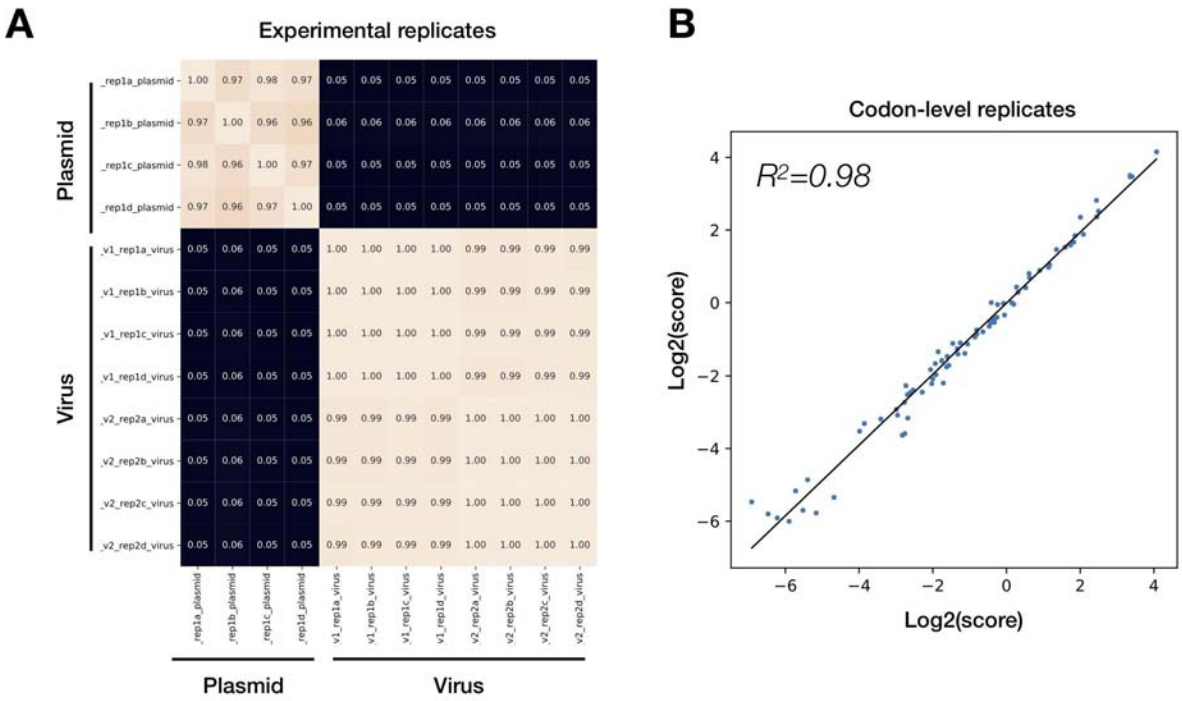
Tseng, Y.-S., and Agbandje-McKenna, M. (2014). Mapping the AAV Capsid Host Antibody Response toward the Development of Second Generation Gene Delivery Vectors. Front. Immunol. *5*, 9.

Wittmann, B.J., Johnston, K.E., Wu, Z., and Arnold, F.H. (2021). Advances in machine learning for directed evolution. Curr. Opin. Struct. Biol. *69*, 11–18.

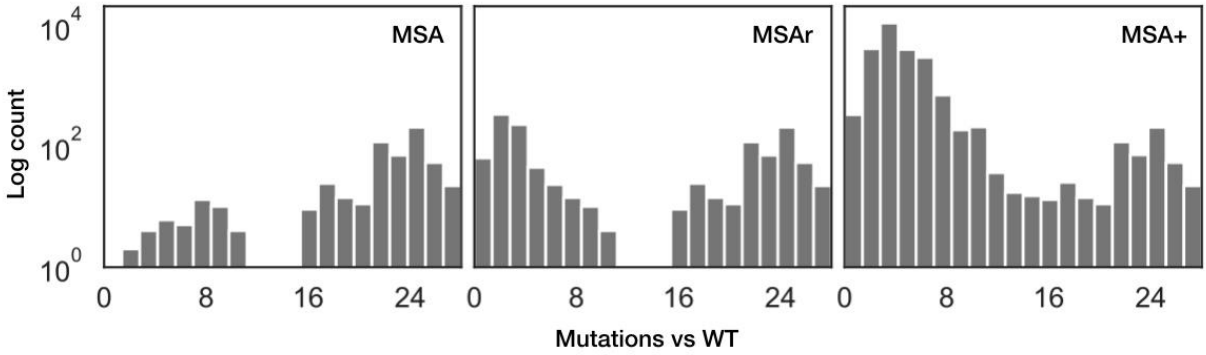
Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat. Methods *16*, 687–694.

Zinn, E., Pacouret, S., Khaychuk, V., Turunen, H.T., Carvalho, L.S., Andres-Mateos, E., Shah, S., Shelke, R., Maurer, A.C., Plovie, E., et al. (2015). In Silico Reconstruction of the Viral Evolutionary Lineage Yields a Potent Gene Therapy Vector. Cell Rep. *12*, 1056–1068.

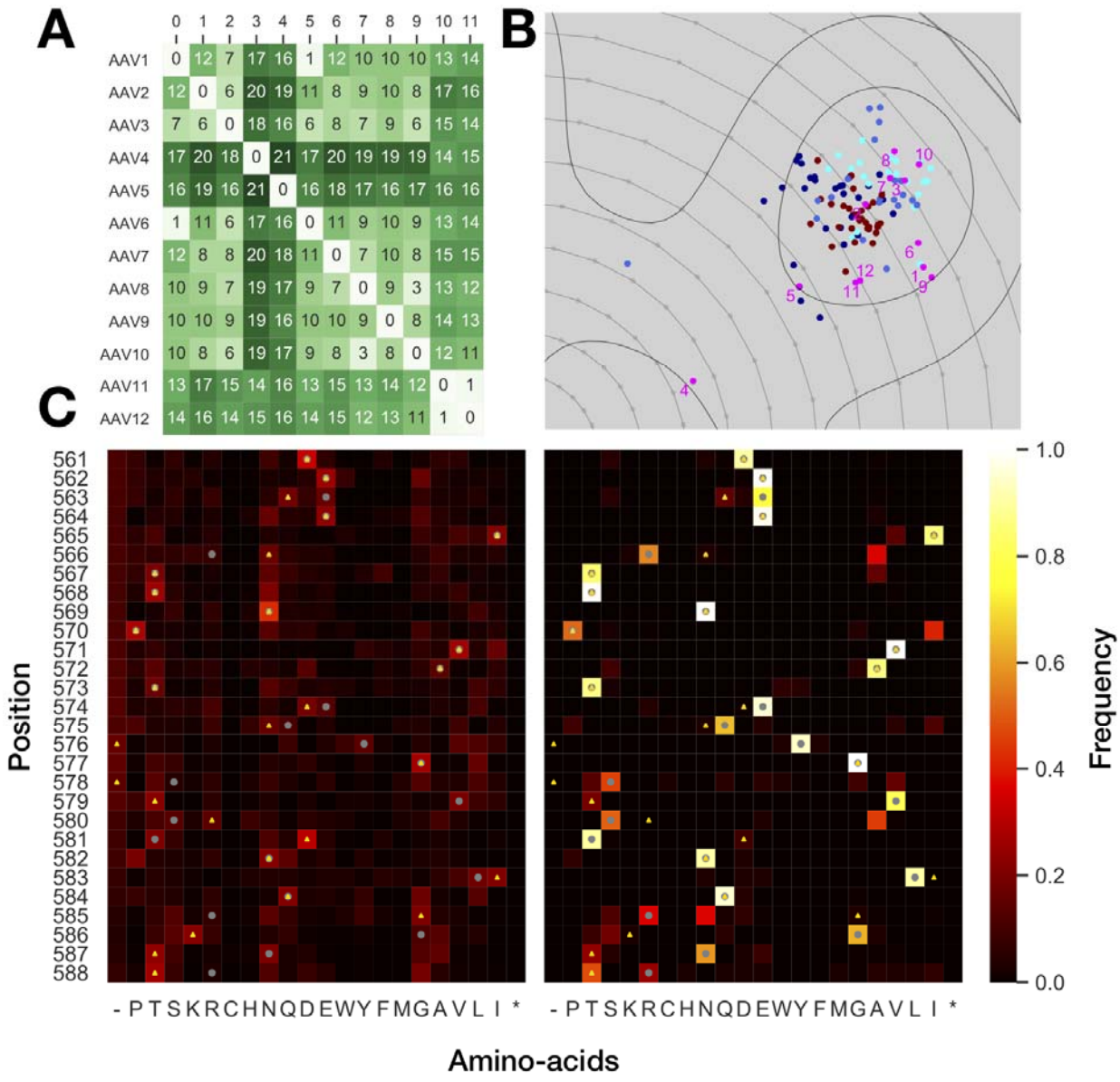
Supplementary material



Supplementary Figure 1. Experimental reproducibility. (A) We show correlation between counts among 4 separate plasmid and 4 separate virus production replicates (each with two PCR replicates). As expected, plasmid and viral counts per variant show no correlation, whereas there is excellent correlation within virus and plasmid replicates. (B) We included n=196 codon-level controls (different codons but same amino-acid sequence) corresponding to n=78 unique amino-acid sequences in our synthesis, confirming that there is excellent correlation between codon-level replicates.



Supplementary Figure 2. Distribution of substitution counts in training data. A majority of MSA variants are more than 16 mutations away from AAV2. On the other hand, a majority of the MSA+ falls within 5 substitutions of the reference variant.



Supplementary Figure 3. (A) Levenshtein edit-distance of the 560-588 segment among common AAV serotypes. (B) Projection of the top 100 produced variants onto the VAE latent space, colored as in Figure 2 for each method, along with the AAV serotypes for reference. (C) Left PWM for the IS model, with wildtype positions marked with a grey dot, and the maximum row-wise value marked by a yellow triangle. Right: PWM for top 100 variants. Star (*) represents the stop codon, and (–) indicate gaps in the alignment.

Supplementary Table 1.
Number and proportion of variants designed by each method resulting in viable and viable+ samples.

	MSA	**MSAr**	**MSA+**
IS	Viable+: 0 (0%) Viable: 369 (29.8%) Total: 1237	Viable+: 3 (0.2%) Viable: 583 (46.9%) Total: 1244	Viable+: 116 (10%) Viable: 1040 (92.6%) Total: 1122
VAE	Viable+: 34 (2.7%) Viable: 385 (31.0%) Total: 1242	Viable+: 44 (3.5%) Viable: 587 (47.3%) Total: 1242	Viable+: 42 (3.4%) Viable: 370 (29.8%) Total: 1243

Supplementary information

Library Production

Libraries were produced precisely as and in parallel with the experiments described before in (Bryant et al., 2021) (see methods). We also used the same original additive training set as our starting point to construct our training library, described therein. We successfully cloned 7446 out of the 7500 (99.28% yield) of our designed sequences as plasmids but only analysed the 7409 for which we sampled at least 100 plasmids in the cloned pool.

Viral production classification

We have observed previously that production assays tend to follow a bimodal mixture of Gaussian distributions ((Bryant et al., 2021), see supplementary Fig 1) , with scores in the first mode closely resembling those of the negative controls.

We numerically fit the distribution of scores within our library with a Gaussian mixture-model (GMM), which yields ${\mu_{gmm}}^{pos} = -1.35, {\mu_{gmm}}^{neg} = -4.95$ and ${\sigma_{gmm}}^{pos} = 2.04, {\sigma_{gmm}}^{neg} = 0.85.$

Furthermore, we compare our fit results against wildtype (WT) produced (which is expected to score higher than the typical variant as many mutations are deleterious) and stop-codon controls produced in a separate experiment. There we had μ_{wt}=-0.80, μ_{stop} =-5.47 and $\sigma_{wt} = 0.91, \sigma_{stop} = 1.03$. Using these two sets of approximations for scores we bin the data into three categories: Non-viable variants that do not package DNA efficiently (similar to stop-codon in performance), viable variants (between non-viable and highest WT replicate score), and viable+ which we categorize as extremely good capsid producers.

We calculate the boundary between viable and non-viable as follows. We approximate the distribution of WT and stop-codon controls by Gaussian distributions and calculate $b_{viable} = (\mu_{wt} - 2\sigma_{wt} + \mu_{stop} + 2\sigma_{stop})/2$. For the better than WT boundary, we calculate $b_{viable+} = (\mu_{wt} + 3\sigma_{wt})$. These boundaries yield the following values for the cumulative distribution function of our GMM fits: $CDF\ (b_{viable}; GMM_{neg})$ =0.999, $CDF\ (b_{viable}; GMM_{pos}) = 0.335, CDF\ (b_{viable+}; GMM_{pos}) = 0.977.$ Given this evidence, we believe these are strongly conservative but trustworthy decision boundaries minimizing false positives.

Sampling from PWMs

We sample variants from PWMs in both IS and VAE model by normalizing frequencies $f_i = \frac{count_i}{\sum count_j}$ in each position and sampling an amino acid at each position (possibly the WT variant) with probability $p_i = \frac{f_i}{\sum f_j}$.

For Maximum likelihood samples, we take the most likely amino acid for each position. We don't use a temperature parameter (that is, we assume a fixed temperature).

Alignments

We performed a Jackhmmer search (2015 update (Finn et al., 2015)) against UniprotKB for the VP positions 200-735 yielded a clear clustering of high-scoring hits based and were pruned based on qualitative inspection of the significance scores. We then extracted the 28 amino acid regions of interest from this data, resulting in 564 sequences.