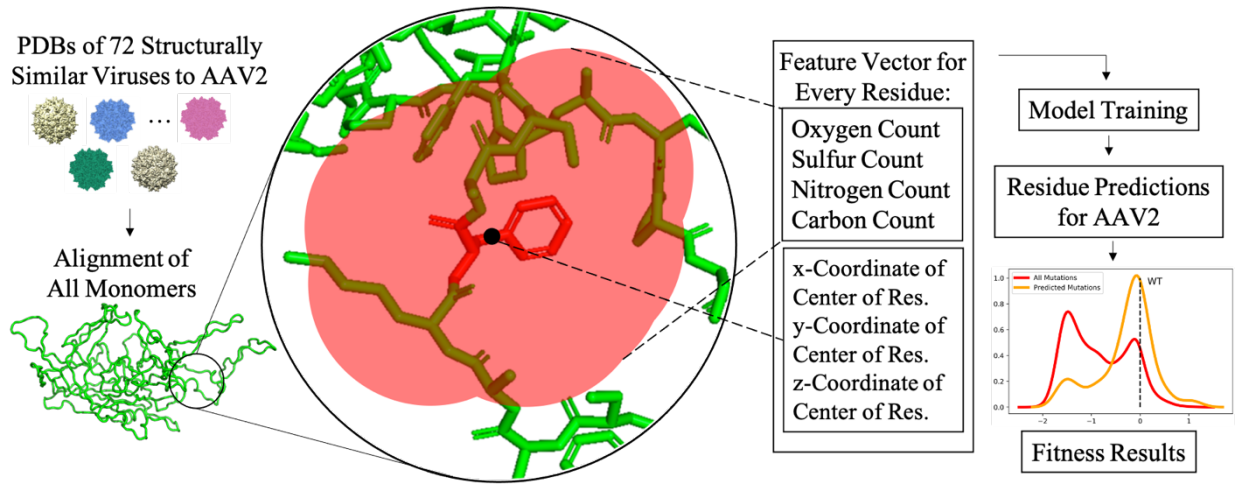
# Machine Learning Identification of Capsid Mutations to Improve AAV Production Fitness

Georgios Mikos[1], Weitong Chen[2*], Junghae Suh[1,2,3]

## ABSTRACT

The adeno-associated virus (AAV) holds great potential for gene therapy efforts by providing a viable vector. However, current efforts are constrained by a lack of AAV variants that exhibit specific tropisms or immunogenicity and a lack of sustainable industrial projection. Departing from experimental approaches to addressing these issues, we built a model to predict residue mutations to improve AAV production fitness. Our model leverages the evolutionary paradigm and microenvironment characteristics by analyzing structural AAV data without needing domain knowledge or experimental fitness data for AAV as inputs. When testing our model's predictions for AAV2 residue mutations, we found a threefold increase in the percent of mutations yielding variants with better production fitness than wild type compared to random mutations, achieving a p-value of $7.46\times10^{-12}$. Given these results, our machine learning approach of using structural data to approximate fitness data has the potential to accelerate AAV development.




**Keywords:** Adeno-Associated Virus, Machine Learning, Gene Therapy

[1] Department of Bioengineering, Rice University, 6100 Main Street, Houston, TX 77005
[2] Department of Chemical and Biological Engineering, Rice University, 6100 Main Street, Houston, TX 77005
[3] Department of Systems, Synthetic, and Physical Biology, Rice University, 6100 Main Street, Houston, TX 77005
* Corresponding Author

## INTRODUCTION

Adeno-associated virus (AAV) presents itself as one of the most promising vectors for gene therapy because of its broad tropism, persistent transgene expression, and limited immunogenicity.[1] AAV clinical trials targeting corneal blindness and hemophilia have demonstrated the promise of AAV,[2,3] with one therapy obtaining European regulatory approval.[4] Prompted by these advancements, there is an increased interest in developing new variants of AAV to increase tropism specificity and evade specific immune responses.[5]

AAV, a member of the *Parvoviridae* family and *Dependoparvovirus* genus, is an icosahedral virus composed of sixty monomers with the three capsid proteins of VP1, VP2, and VP3 present in a 1:1:10 ratio, and the capsid contains a linear single-stranded DNA genome.[6] The ease of manipulating this genome that contains a *cap* gene encoding capsid proteins has prompted significant research efforts into optimizing the virus capsid. Modification to the capsid in hopes of improving production fitness represents the first step to developing new stereotypes and may provide the key to improved industrial production, increasing the viability of AAV as a clinical pharmaceutical.

So far, the exploratory process for new serotypes still depends on random mutagenesis, albeit a more systematic version of it through various diversification strategies, combinatorial libraries, directed evolution methodologies, and fitness landscapes.[7] These experimental approaches, combined with new promising computational techniques using SCHEMA or other genome-based algorithms,[8] have accelerated the production of new AAVs. Machine learning has been employed to aid project workflows; by training on experimental fitness data, models can predict capsid viability[9] or improve libraries gene therapy vectors.[10] Importantly, these approaches depend on experimental data for any computational analysis.

Alternative, promising research has been conducted for entirely *in silico* development of novel AAV capsids. Zinn *et al.* previously reconstructed ancestral AAVs through *in silico* phylogenetic analysis and produced nine functioning ancestral AAVs.[11] Their approach demonstrates the predictive power of the evolutionary paradigm in optimizing a target virus.

Along a different, but relevant, line of work, Torng and Altman used an *in silico* approach to predict optimal residues for a given position for various protein structures.[12] Their result was achieved by leveraging each residue's microenvironment, the immediate area with distinct properties that surrounds the residue, to train a 3D convolution neural network that considers over

80 chemical microenvironment characteristics. Their work demonstrates the predictive power of microenvironment features.

Here, we present a machine learning model that combines the predictive power of the evolutionary paradigm and microenvironment features to suggest modifications to the AAV2 capsid in hopes of generating novel AAV capsids without compromising virus assembly fitness and accelerating the development of new serotypes. Crucially, our approach does not use viral capsid assembly experimental data to train a machine learning model. Instead, we use the 3D position of residues from structurally similar viruses to AAV2 and their microenvironment characteristics as a proxy for experimental fitness data. Since structural data for several AAV serotypes can be obtained from the Protein Data Bank[13] whereas extensive fitness data for all capsid modifications require lengthy and expensive experimental work, this ability to translate structural data to fitness data can greatly accelerate the development of new capsids and provide a new approach for AAV modification.

## RESULTS AND DISCUSSION

Our model predicted 74 viable point mutations to the AAV2 *cap* gene. To assess the fitness in virus production for each point mutation, we needed to characterize each mutation's fitness. In lieu of generating a panel of 74 different mutations in each AAV2 monomer and performing capsid assembly and characterization of each capsid, we compared our predicted mutation with an existing published set of well characterized single point mutations in AAV2 generated by Ogden et al. [14] This dataset provides the virus production fitness for every possible point mutation in the AAV2 *cap* gene. We compared the distribution in average fitness of all mutations possible to our predicted mutations. Our predicted mutations outperformed random mutations (Figure 1A,B). The density plot (Figure 1C) demonstrates how the model minimizes the number of harmful mutations shown by the peak at approximately -1.5 log10 average fitness and maximize the number of mutations similar to or better than WT, as indicated by the peak around WT. We also performed a t-test and achieved the p-value $7.46\times10^{-12}$. Crucially, we achieved over a threefold increase in the percentage of mutations that resulted in viruses with better production fitness than wild type. This result demonstrates how our model can accelerate the process of producing new serotypes, and it

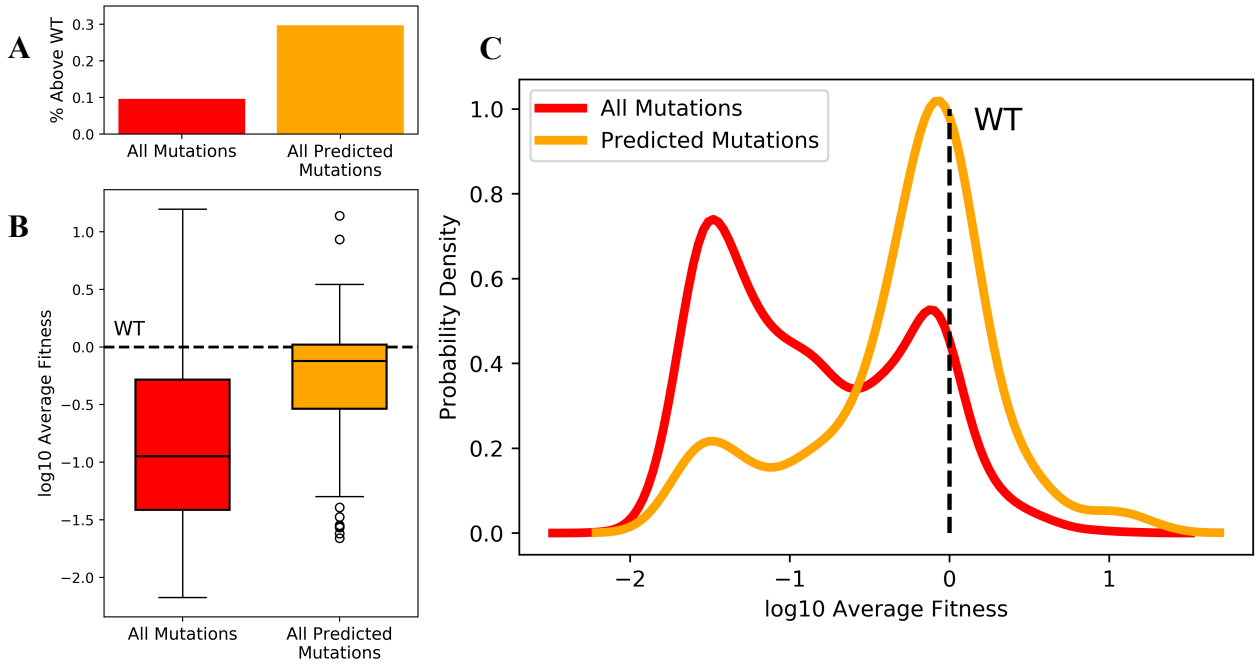



*Figure 1: Average Fitness Analysis: (A) Fraction of AAVs with fitness above WT for All Mutations and for Predicted Mutations. (B) Box-plot comparison of the average fitness of Predicted Mutations to the average fitness of All Mutations. The Predicted Mutations outperform All Mutations. (C) Density plot comparison of average fitness. The Predicted Mutations minimize the deleterious peak at ~-1.5 log10 and concentrate the peak around WT.*

proves that residue position along with microenvironment features can be accurately used instead of experimental fitness values.

We also considered whether the residue positions are on the surface or buried in the virus capsid. Of the 74 positions, 59 are on the surface, while 15 are buried. The surface positions outperformed the buried ones in terms of fitness, as expected, but the difference was not statistically significant with a p-value of 0.0688 (Figure S1).

To better understand our model, we performed a feature permutation analysis to evaluate the importance of each feature (Figure S2). The features related to the position of the residue proved most important; the only rivaling microenvironment feature is the carbon count. We hypothesize this reliance indicates our model primarily depends on identifying positions that can tolerate mutations based on the low conservation of the residue identity at a particular spatial position. Inversely, the model avoids highly conserved regions that cannot tolerate mutations and identifies any mutational “mistakes” in the AAV2 capsid that are not present in other capsids.

To prove our model targets highly variable regions, we plotted the position of the predicted residues on the AAV2 monomer (Figure 2A) and calculated the conservation for each residue in the AAV monomer based on the other serotypes of AAV (Figure 2B). We found that the average

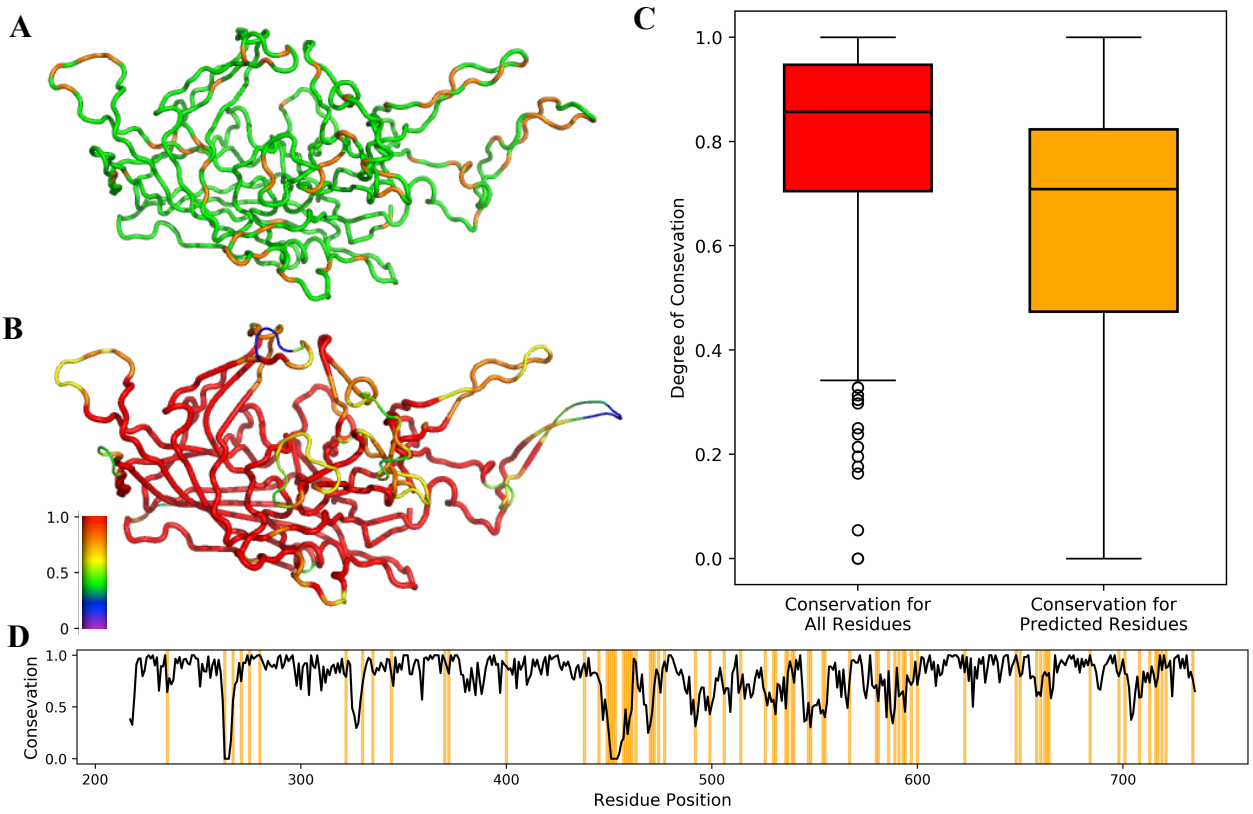



*Figure 2: Conservation Analysis for Predicted Residue Positions: (A) Position of predicted mutations in orange on AAV2 monomer. (B) Average conservation of residues on AAV2 monomer compared to other AAVs. Red (1) indicates complete conservation whereas purple (0) indicates no conservation by position. (C) Box-plot comparison of average conservation of Predicted Residues versus All Residues. (D) Plot of conservation of all residue positions with positions where predicted mutations are highlighted.*

conservation based on sequence alignment at the residue positions where mutations were predicted was significantly lower than the average conservation at all residue positions (Figure 2C). This difference in conservation reaffirms our model's predictive power since our model is targeting residue positions with low conservation. This suggests both that our model is focusing on residue positions that can tolerate mutations and on residue positions where other serotypes may have a beneficial residue that is not present in the AAV2 capsid.

The use of structurally similar viruses leverages an evolutionary paradigm to differentiate between positions that can tolerate mutations and positions that cannot tolerate mutations. Additionally, because we train on highly functioning viruses, our work allows for the implicit consideration of crucial macroscopic constraints that cannot be concretely defined. Further, in comparison to previous work that focuses on sequence similarity, our work's use of structural data

allows for the comparison of different secondary and tertiary structures, providing a more complete analysis of viral properties.

We performed a goodness of fit chi-square test to compare the frequency of each predicted residue to their respective frequencies in the training data. The test resulted in a p-value of 0.027, but only serine was a key contributor to the chi-statistic (Figure S3). We did not expect a significant result since the model is chiefly reliant on the position of each residue and not the microenvironment. This might indicate our model established a consistent link between the carbon count, which is the only important microenvironment feature, and the suitability of serine.

We also investigated the number of PDBs necessary to achieve the average fitness obtained by our model. We found that subsets as small as ten from our list of 72 PDBs could achieve an average fitness greater than our model (Figure 3A). However, certain other subsets of ten performed abysmally compared to the model with all the PDBs. To analyze this difference, we created a phylogenetic tree of all the PDBs. On average, the subsets of ten where the PDBs are closer to AAV2 on the phylogenetic tree perform better, indicating overfitting (Figure 3B). However, despite this limitation, since only a small set of PDBs is required to achieve a useful model, the intelligent use of a small number of PDBs could allow for the creation of AAV serotypes with specific tropisms by only training the model on PDBs with the desired tropism.

Our study considered the potential of spatial position and microenvironment factors in predicting the optimal residue. Future work could expand this study to AAVs with specific tropisms or consider how our model could harmoniously use the singular predicted mutation to inform a decision about predicting an additional mutation.

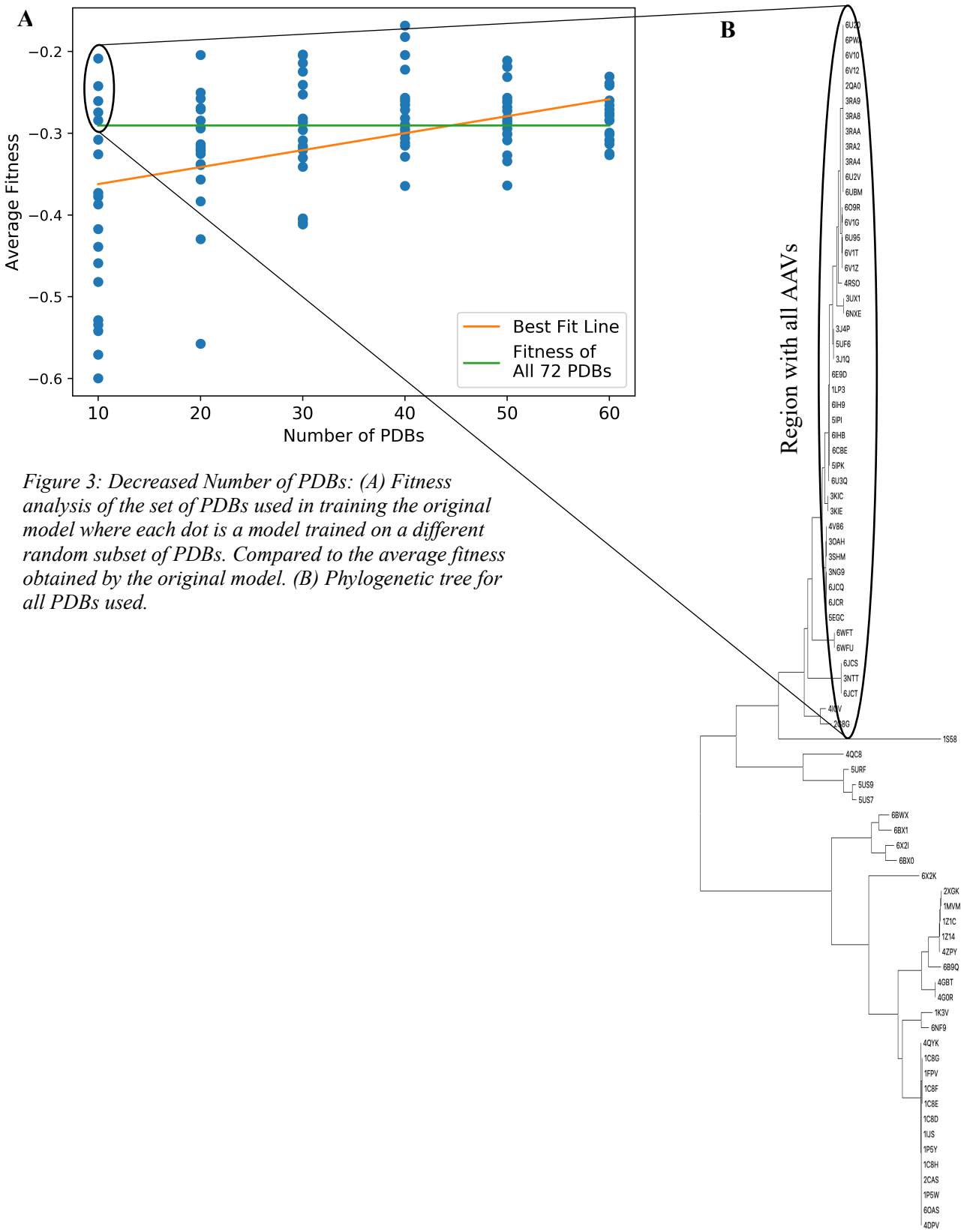



*Figure 3: Decreased Number of PDBs: (A) Fitness analysis of the set of PDBs used in training the original model where each dot is a model trained on a different random subset of PDBs. Compared to the average fitness obtained by the original model. (B) Phylogenetic tree for all PDBs used.*

# METHODS

## Preparation of Data

We chose the AAV2 PDB 1LP3 as our target virus.[15] Using the structure similarity search in RCSB, we identified 80 structurally similar PDBs after setting a threshold at 30% structure similarity. From those 80, we removed 1LP3 along with seven other AAV2 PDBs we identified through AAV2's UniProt reference sequence of P03135.[16] Hence, we excluded all occurrences of AAV2 in the training set. The remaining 72 PDBs became our training data (Table S1).

We aligned all the monomers of each PDB to 1LP3 through superposition, using PyMol's to normalize the coordinate system by minimizing the root mean square deviation of the atomic position. The alignment allows for the consideration of the residue position among all the viruses.

To generate our feature vector, for each residue in each monomer, we identified the center of the residue and the microenvironment. We defined the center of the residue as the average of the coordinates of each atom in the residue, and we defined the microenvironment as everything within 5 Armstrong from any atom in the selected residue. Thus, we identified the x, y, and z coordinates of the residue and the count of carbon, nitrogen, oxygen, and sulfur atoms in the microenvironment. These data points are the input to the model.

## Training and Evaluating the Model

We performed a leave-one-out training on our data set, with the one left out being AAV2, so that the model can train on all the 72 obtained AAVs while preventing overfitting on AAV2. Using the sci-kit learn package for python, we trained a Random Forest model with hyperparameters optimized through a random search with cross-validation.[17] The parameters were bootstrap false, criterion gini, max features 3, min samples split 2, n-estimators 449. Random forest is an ensemble learning model that constructs independent decision trees.[18] This model was chosen because it can be used for classification tasks and can achieve a high level of accuracy even on limited training data and reduce overfitting. Hence, given the position and microenvironment features as inputs, our model outputs a predicted residue. When tested on 1LP3, the model predicted 86% of the amino acids correctly while it produced an alternative amino acid prediction for 14% of the cases (74 residues).

For the surface versus buried analysis, we defined a surface residue as a residue having surface exposure of greater than 6.25 Armstrong$^2$.[19]

We used the ELI5 package to perform the permutation analysis. A permutation analysis involves scrambling the data for each feature, running the model, and then comparing the decrease in accuracy for each feature.

Figure 2 was computed by analyzing the conservation found after the structural alignment of the monomers. In the alignment, we used PDBs for AAV1, AAV2, AAV4, AAV5, AAV6, AAV8, and AAV9 which are 6JCR, 1LP3, 2G8G, 3NTT, 3OAH, 2QA0, and 3UX1, respectively. We then compared the conservation values for AAV2 for the residue position predicted by our model with the conservation values for all the residue positions.[20]

For the analysis considering smaller sets of PDBs, we randomly selected 20 subsets of 10, 20, 30, 40, 50, and 60 PDBs, resulting in 120 tests. For each of these sets, we ran the entire methodology and compared the average production fitness, plotted in Figure 3. To assess the relation between different PDBs in a set, the phylogenetic tree was created in MEGAX following the instructions outlined by Hall.[21,22]

**Acknowledgments**:
This work is supported by a National Science Foundation Fellowship to WC (2018253392), and National Institutes of Health to JS (R01HL138126 and R01CA207497).

**Conflict of Interest**:
JS is an employee of Biogen as of August 2019.
WC is an employee of Biogen as of March 2020.

# REFERENCES

[1] Yang Lu, "Recombinant Adeno-Associated Virus As Delivery Vector for Gene Therapy—A Review," *Stem Cells and Development* 13, no. 1 (February 2004): 133–45, https://doi.org/10.1089/154732804773099335.
[2] Melisa Vance et al., "AAV Gene Therapy for MPS1-Associated Corneal Blindness," *Scientific Reports* 6, no. 1 (April 2016): 22131, https://doi.org/10.1038/srep22131.
[3] Amit C. Nathwani et al., "Adenovirus-Associated Virus Vector–Mediated Gene Transfer in Hemophilia B," *New England Journal of Medicine* 365, no. 25 (December 22, 2011): 2357–65, https://doi.org/10.1056/NEJMoa1108046.
[4] Laura M. Bryant et al., "Lessons Learned from the Clinical Development and Market Authorization of Glybera," *Human Gene Therapy Clinical Development* 24, no. 2 (June 2013): 55–64, https://doi.org/10.1089/humc.2013.087.
[5] William B. Guggino and Liudmila Cebotaru, "Adeno-Associated Virus (AAV) Gene Therapy for Cystic Fibrosis: Current Barriers and Recent Developments," *Expert Opinion on Biological Therapy* 17, no. 10 (October 3, 2017): 1265–73, https://doi.org/10.1080/14712598.2017.1347630.
[6] S. Daya and K. I. Berns, "Gene Therapy Using Adeno-Associated Virus Vectors," *Clinical Microbiology Reviews* 21, no. 4 (October 1, 2008): 583–93, https://doi.org/10.1128/CMR.00008-08.
[7] Dirk Grimm and Sergei Zolotukhin, "E Pluribus Unum: 50 Years of Research, Millions of Viruses, and One Goal—Tailored Acceleration of AAV Evolution," *Molecular Therapy* 23, no. 12 (December 2015): 1819–31, https://doi.org/10.1038/mt.2015.173.
[8] David S. Ojala et al., "In Vivo Selection of a Computationally Designed SCHEMA AAV Library Yields a Novel Variant for Infection of Adult Neural Stem Cells in the SVZ," *Molecular Therapy* 26, no. 1 (January 2018): 304–19, https://doi.org/10.1016/j.ymthe.2017.09.006.
[9] Drew H. Bryant et al., "Deep Diversification of an AAV Capsid Protein by Machine Learning," *Nature Biotechnology*, February 11, 2021, https://doi.org/10.1038/s41587-020-00793-4.
[10] Andrew D. Marques et al., "Applying Machine Learning to Predict Viral Assembly for Adeno-Associated Virus Capsid Libraries," *Molecular Therapy - Methods & Clinical Development* 20 (March 2021): 276–86, https://doi.org/10.1016/j.omtm.2020.11.017.
[11] Eric Zinn et al., "In Silico Reconstruction of the Viral Evolutionary Lineage Yields a Potent Gene Therapy Vector," *Cell Reports* 12, no. 6 (August 2015): 1056–68, https://doi.org/10.1016/j.celrep.2015.07.019.
[12] Wen Torng and Russ B. Altman, "3D Deep Convolutional Neural Networks for Amino Acid Environment Similarity Analysis," *BMC Bioinformatics* 18, no. 1 (December 2017), https://doi.org/10.1186/s12859-017-1702-0.
[13] Helen M. Berman et al., "The Protein Data Bank," *Nucleic Acids Research* 28, no. 1 (January 1, 2000): 235–42, https://doi.org/10.1093/nar/28.1.235.
[14] Pierce J. Ogden et al., "Comprehensive AAV Capsid Fitness Landscape Reveals a Viral Gene and Enables Machine-Guided Design," *Science* 366, no. 6469 (November 29, 2019): 1139–43, https://doi.org/10.1126/science.aaw2900.
[15] Q. Xie et al., "The Atomic Structure of Adeno-Associated Virus (AAV-2), a Vector for Human Gene Therapy," *Proceedings of the National Academy of Sciences* 99, no. 16 (August 6, 2002): 10405–10, https://doi.org/10.1073/pnas.162250899.
[16] A Srivastava, E W Lusby, and K I Berns, "Nucleotide Sequence and Organization of the Adeno-Associated Virus 2 Genome.," *Journal of Virology* 45, no. 2 (1983): 555–64, https://doi.org/10.1128/JVI.45.2.555-564.1983.
[17] F. Pedregosa et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–30.
[18] Leo Breiman, "Random Forests," *Machine Learning* 45, no. 1 (2001): 5–32, https://doi.org/10.1023/A:1010933404324.
[19] Jason Vertress, "FindSurfaceResidues," PyMol Wiki, March 29, 2019, https://pymolwiki.org/index.php/FindSurfaceResidues.
[20] Jason Vertress, "Conservation by Color," PyMol Wiki, January 15, 2012, https://pymolwiki.org/index.php/Color_by_co….
[21] Glen Stecher, Koichiro Tamura, and Sudhir Kumar, "Molecular Evolutionary Genetics Analysis (MEGA) for MacOS," ed. Claudia Russo, *Molecular Biology and Evolution* 37, no. 4 (April 1, 2020): 1237–39, https://doi.org/10.1093/molbev/msz312.
[22] B. G. Hall, "Building Phylogenetic Trees from Molecular Data with MEGA," *Molecular Biology and Evolution* 30, no. 5 (May 1, 2013): 1229–35, https://doi.org/10.1093/molbev/mst012.