Learning from and improving upon high-throughput screens for protein fitness with Generative AI - Application to BBB-crossing AAV design

Ayan Kashyap[1], Kumari Soniya[1], Anubhooti[1], Subhalakshmi Kandavel[1], Siva Kanishka[1], Manasvi Sharma[1*] & Prasad Chodavarapu[1*]

[1]Aganitha.ai

Deep Science is enabling high-throughput experimentation (HTE) to design novel biological entities with desired properties. E.g., blood-brain-barrier (BBB) crossing adeno-associated virus (AAV) vectors, needed for systemic delivery of gene therapies to brain cells, have been identified through innovative directed evolution assays such as M-CREATE and TRACER. But, even these high-throughput experiments are only able to explore a miniscule portion of the large design space of biological entities. In this paper, we introduce autograd based maximization of protein fitness (AutoMaxProFit) to learn from and improve upon protein designs generated with high-throughput screens. Using a transformer based generative AI network and protein language models, we improve upon the design of a variant previously discovered through HTE, to yield 4x better enrichment in brain endothelial cells, as estimated by molecular dynamics (MD) simulations. This shows that Deep Tech models can learn from the observations generated by Deep Science experiments and go on to find more optimal design candidates for application in Biopharma.

Keywords: AutoMaxProFit, Protein fitness optimization, High-throughput screening, Directed evolution, Generative AI, Transformer, Protein language model (pLM), Autograd, Molecular dynamics (MD), Adeno-associated virus (AAV), Viral vector engineering, Blood-brian barrier (BBB), Gene Therapy, CNS

Introduction

Deep Science and Deep Tech are advancing Biopharma R&D

Every component of Biopharma R&D value chain - disease studies, target identification, elucidation and validation, to therapeutic design and optimization, synthesis, pharmaco-kinetic/dynamic (PK/PD) and safety analysis - is undergoing rapid transformation, with the adoption of Deep Science and Deep Tech. Deep Science is allowing researchers to precisely perturb biological systems in high-throughput multiplexed protocols, and inspect/readout molecular and cellular states and activities. These advances on the science front are generating a wide-variety of large datasets that deep tech is helping analyze - to not only produce useful and novel insights, but also to integrate, model and generate novel designs. This process is visible across all disease areas as well as current and novel therapeutic modalities.

This paper fits into the same pattern, showing how Deep Tech can build on the datasets generated by Deep Science, for developing gene therapies that address genetic disorders impacting the brain.

*Corresponding author: Prasad Chodavarapu, chap@aganitha.ai.
*Corresponding author: Manasvi Sharma, manasvi@aganitha.ai.

Gene therapies and viral vectors

Gene therapies aim to treat genetic disorders by delivering the missing functional genes into target cells. Adeno-associated virus (AAV) and lentivirus based vectors are commonly used in gene therapies as they have the natural ability to transduce cells and deliver genetic material safely.

A number of CNS disorders have genetic root causes. While local delivery of gene therapies into CNS is possible via intracranial and intrathecal administration into the brain and spinal cord respectively, systemic delivery via intravenous administration is more convenient for patients and care providers due to their non-invasive nature and are therefore, preferred over other delivery routes. However, intravenous administration requires high viral doses and are likely to provide relatively less efficient transduction of target cells, as crossing the blood-brain barrier (BBB) for reaching brain cells and blood-spinal cord barrier (BSCB) for reaching cells in the spinal cord, poses a major challenge for a successful gene therapy administration. Also, vectors used for systemic delivery should have minimal to nil penetration of cells other than the targeted cell types to minimize any off-target effects.

Success with intravenous administration of a gene therapy for spinal muscular atrophy (SMA), delivered using self-complementary (sc) AAV9 vectors [1] into mice and cats, led to human trials and eventual approval of the therapy. This encouraged research into similar vectors that can cross the BBB [2] for tackling genetic disorders of the brain such as neurodevelopmental and neurodegenerative disorders with genetic root causes.

Directed evolution (DE) based innovations in BBB-crossing AAV

A number of studies over the years have focused on discovery of AAV variants that can cross the BBB and show tropism to specific cell types in the brain. (See [3] for a list of studies up to 2022). Studies such as CREATE [4], M-CREATE [5], TRACER [6] used directed evolution (DE) experiments in select mice models to identify specific 7-mer peptide inserts into the loop VIII (VR-VIII) region of AAV9 that bestow upon the resulting AAV9 variants, distinctly higher penetration into specific brain cell types when administered intravenously.

Subsequent studies (such as [7]) identified the specific receptors (such as LY6A in C57BL/6J mice) in brain endothelial cells that the BBB-penetrant AAV9 variants are binding to. This insight enabled the easier in vitro route to screen and discover more AAV variants that penetrate BBB in a broader set of mice models [8].

Combination of directed evolution, target receptor informed screening, and additional "semi-rational" capsid engineering informed by inspection of limited data available led to discovery of variants targeting BBB of other species such as non-human primates (NHP) [9, 10] and even humans ([11] with ex-vivo brain slices, [12] using transgenic mice).

Recently, in situ sequencing was used to create a spatial atlas of mouse CNS and the distribution of BBB-crossing AAV9 variants across different regions of the brain was studied [13], generating more data that can be mined with Deep Tech.

Deep Tech enables further optimization

In all the discoveries described so far, while Deep Science enabled discovery of better variants, there was always a problem of limited data availability, in comparison to the size of the design space involved.

Problem of design space coverage

CREATE, M-CREATE and TRACER insert 7-mer peptides into the capsid sequence, looking for the best insert among 20^7 (=1.28 billion) possible amino acid residue sequences or, equivalently, among approximately $3^7 * 20^7$ (about 2.8 trillion) nucleotide sequences, by testing just 100 million nucleotide sequences. That represents a coverage of just 0.04% of the possible design space. Clearly, there's an opportunity for improving upon the optimal sequences found through directed evolution, if only we can find a way to do so.

Generalization to protein fitness optimization

AAV optimization fits into a broader class of protein fitness optimization problems. Optimizing enzymes for catalytic activity as well as expression, stability, solubility and other properties required for developability is a common use case for protein fitness optimization methods.

Previous attempts

Previous attempts at application of AI/ML to AAV variant generation [14–16] focused mostly on generating a good dataset that's likely to be representative of the fitness landscape over the design space. Predictive models developed on the experimental datasets were then used for screening [17–19] a large number of randomly generated sequences or sequences generated through saturation mutagenesis at select positions. Sequences with top predicted fitness scores were further validated with *in vitro* or *in vivo* studies.

Autograd based maximization of protein fitness (AutoMaxProFit)

In this paper, we describe a more direct approach, AutoMaxProFit, to finding optimal capsid sequences given the results of directed evolution experiments. Figure 1 summarizes the approach.

AutoMaxProFit leverages the availability of pre-trained protein language models (pLMs) [20] to reduce the demands on training dataset size and coverage. A deep neural network based model is trained to predict protein fitness, given the representation of a protein sequence in the pLM embedding space.

To maximize protein fitness, AutoMaxProFit carries out gradient ascent in the embedding space, using autograd, the facility to compute gradients automatically in deep neural networks [21].

Finally, embeddings computed by the optimization process are mapped back into an amino acid sequence, using the de-masking trick all transformer based generative AI models use.

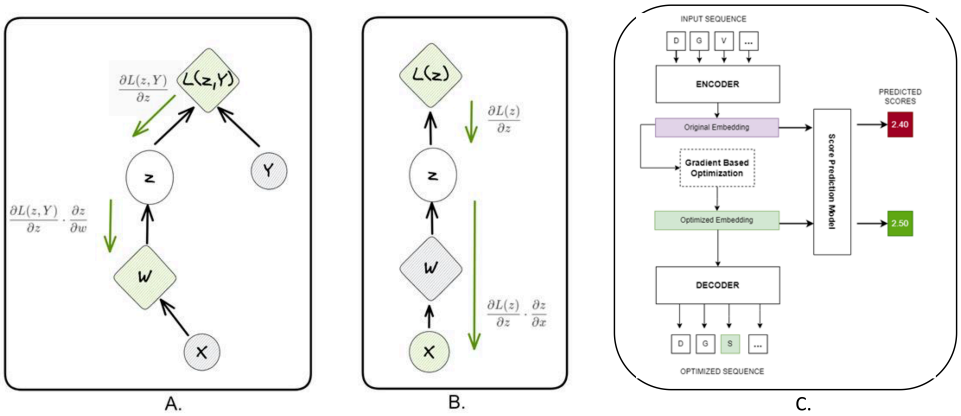

Figure 1: Schematics illustrating (A) Back-propagation of gradients during training of the model for prediction of protein fitness given sequence embedding computed by a protein language model. Loss function L compares prediction of fitness score z to ground truth Y. (B) Back-propagation of gradients during optimization. Loss function L is the predicted fitness score with its sign flipped to pose a maximization task as a minimization problem. (C) Process used to maxmimize fitness of a give protein in AutoMaxProFit.

MD for understanding tropism

We rely on Molecular dynamics (MD) simulations to validate the results from our optimization approach *in silico*.

MD simulations have proved to be useful in dissecting the molecular underpinnings of tropism, particularly by offering insights into how viruses interact with host cell receptors at an atomic level [22–25]. For instance, MD was successfully used in elucidating the role of the viral envelope protein gp120 in the binding of HIV-1 to CD4 receptor and co-receptors (CCR5 or CXCR4) on the host cells [26]. Static and dynamical analyses on MD trajectories were used to reveal how specific conformational changes, upon receptor binding, in gp120, a glycoprotein on the surface of HIV envelope, influence viral entry, shedding light on how alterations in the gp120 structure can impact tropism for different cell types [27].

Similar MD simulations have been performed on influenza virus to understand its tropism, e.g., how the hemagglutinin (HA) protein of influenza adapts to binding with sialic acid receptors on host cells. By simulating different HA variants, researchers could identify structural features critical for binding to specific sialic acid linkages, which correlate with the virus's ability to infect various host species [25]. Recently, MD studies have been extensively used in understanding SARS-CoV2 binding with ACE2 receptors and the effect of different mutations on binding [23]. These examples highlight how MD simulations can reveal the fine details of virus-receptor interactions and provide a platform for exploring how mutations or structural changes influence tropism.

Conventional MD simulations (cMD) were employed in this study to check and quantify the stability and binding of various AAV9 vectors coming out of our prediction pipeline to LY6A receptors on endothelial cells. While RMSD was used to account for the stability of the AAV9-

LY6A complex, binding free energies and buried surface area were used to quantify LY6A binding relative to the known AAV9 capsids exhibiting tropism. Results are discussed in later section.

Results

Performance of predictive model

Learning from directed evolution

Results from M-CREATE are available in the form of enrichment (ER) scores, representing the relative success of a particular capsid sequence in penetration of a target tissue/cell type. We trained a model to predict the enrichment score of a capsid for endothelial cells in the brain, given its amino acid sequence.

To featurize the sequence, we relied on embeddings generated by ProtT5 [20], a protein language model that was trained on more than 2 million protein sequences, using an encoder-decoder transformer architecture [28].

We reduced the dimensionality of the capsid sequence embedding from 525 (residues in capsid sequence) * 1024 (dimensions/residue in ProtT5 embeddings) to 30 * 1024 by retaining only 30 residues in loops IV and VIII (including those in the 7-mer insert) that showed significant variation of embedding within the training set.

We used a 1-D convolutional neural network (CNN) shown in Figure 7, inspired by the best performing architecture used for secondary structure prediction in ProtT5.

Figure 2 shows (a) the distribution of ER scores for endothelial cells in M-CREATE dataset and (b) the performance of the ER score prediction model. Top-k recall values are 1, 0.6 and 0.56 for k=1, 10 and 50 respectively.

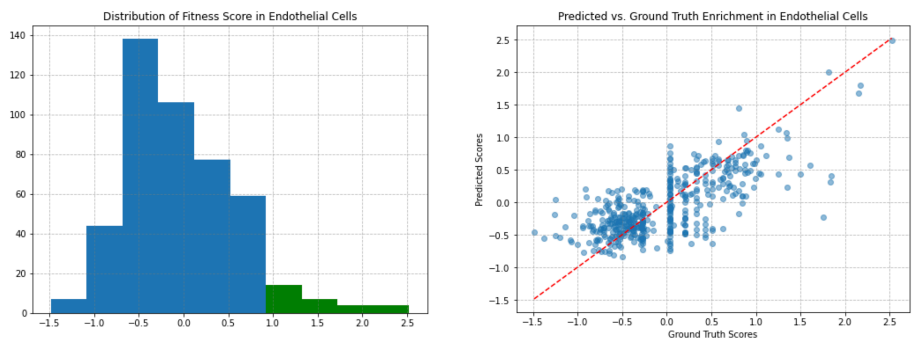

Figure 2: Distribution of ground truth enrichment scores and Ground truth vs prediction scatter plot for endothelial cells

Capsid optimization and validation

Improving upon directed evolution

To improve upon the best sequences found by M-CREATE, we conducted a numerical optimization process in the (dimensionally reduced) sequence embedding space to find the embeddings that produced the highest enrichment score prediction (using the model described in previous section) and subsequently, decoded the embeddings to produce the optimal residue sequence.

For numerical optimization, we used back-propagation with gradient descent enabled by automatic differentiation in PyTorch [21], just as in training of the predictive model, except that the gradients computed during optimization were with respect to the 30 * 1024 input embeddings, and not wrt. the parameters (weights and biases) of the 1-D CNN.

Loss function for the optimizer combined predicted enrichment score (with its sign flipped to pose a maximization task as a minimization problem) along with a penalty for straying outside the embeddings sub-space covered by the training dataset, as shown in Figure 8.

Figure 3 shows how we decoded the embeddings found by the optimization process to yield the (predicted) best enrichment score. Post decoding, we checked to make sure residues outside the 7-mer insert stayed identical to the residues in the wildtype sequence. Any embeddings that did not decode well under this constraint were discarded in favour of next best scoring embeddings that did satisfy the constraint.

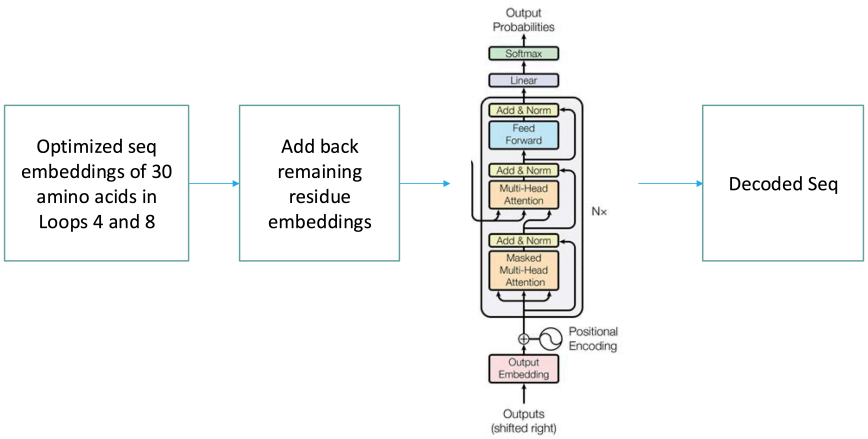

Figure 3: Decoding the sequence embeddings found by the optimization process to yield the highest enrichment score prediction

Table 1 shows the changes optimizer made to maximize the enrichment score for brain endothelial cells, starting from TLQLPFK, a top 7-mer insert found by M-CREATE to yield a good enrichment score for multiple brain cell types.

Table 1: 7-mer sequences for AAV9 loop VIII insert generated by AutoMaxProFit, starting from TLQLPFK, a top insert found by M-CREATE. Also shown are the enrichment scores predicted.

7-mer	Predicted Enrichment Score
TLQLPFK	2.81
SLQAPFK	3.04
SLQAPWK	2.87
SLQAQQQ	3.20
SLQALQA	3.39

Validation of optimized capsid fitness through MD simulations

Short 20 ns cMD simulations were performed on the trimeric AAV9 capsid protein with the 7-mer insert designed by AutoMaxProFit, in complex with the LY6A receptor.

Figure 4 shows the trimeric AAV9 variant structure. The 7-mer peptide inserts were placed in loop VIII of the protein. In the trimeric structure, loop IV, loop V and loop VIII extend outwards to interact with receptors. The trimeric AAV9 protein is bundled together by intertwining the loop V from the adjacent chain with loop IV and loop VIII of the first chain.

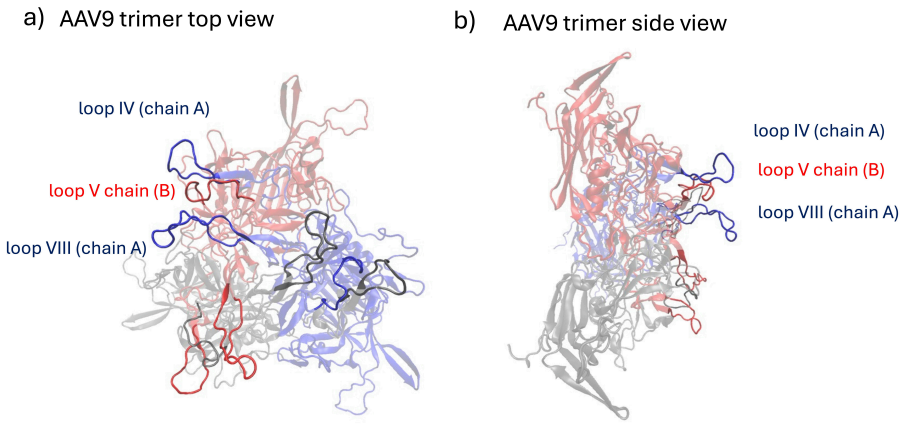

Figure 4: Positions of loops IV, V and VIII in the trimeric AAV9 as seen in a) top view and b) side view

To assess the stability of the complex, we monitored fluctuations in the root-mean-square deviation (RMSD) of the complex's structure over the length of the simulation, shown in Figure 9. We used binding free energy and buried surface area (BSA) as quantitative measures to evaluate the optimised capsid. For comparison, binding free energies and BSA values for known penetrants and non-penetrants, obtained from the M-CREATE and TRACER platforms, were used as controls.

Table 2 shows the quantitative results obtained from analysis of MD trajectories. The capsid optimized by AutoMaxProFit showed a 4x improvement in binding strength.

Table 2: Comparison of relative binding free energy and BSA of in-house optimised capsid to the known controls from M-CREATE and TRACER.

7-mer system	Binding free energy (kcal/mol)	BSA (Å^2)
M-CREATE penetrant	−224	432.2
M-CREATE non-penetrant	−22	425.2
9P16 (TRACER penetrant)	−86	627.2
9P09 (TRACER non-penetrant)	−20	383.7
SLQALQA (AutoMaxProFit designed capsid)	−984	1370.8

For successful cellular penetration through cell receptors, there must be adequate interaction between LY6A (including its N-terminal) and AAV9. The final frames from a 20 ns MD trajectory were analysed and are depicted in Figure 5. The left and middle images show proteins capable of penetrating the blood-brain barrier, while the rightmost image illustrates the optimised sequence. The lower panel displays known non-penetrants with respect to LY6A. Notably, for penetrants, the N-terminal of the LY6A receptor engages extensively with the trimer, resulting in better binding free energy and a larger buried surface area upon complex formation. Conversely, for non-penetrants, the trimeric AAV9 interacts only with loop VIII from one of the three chains in the trimer.

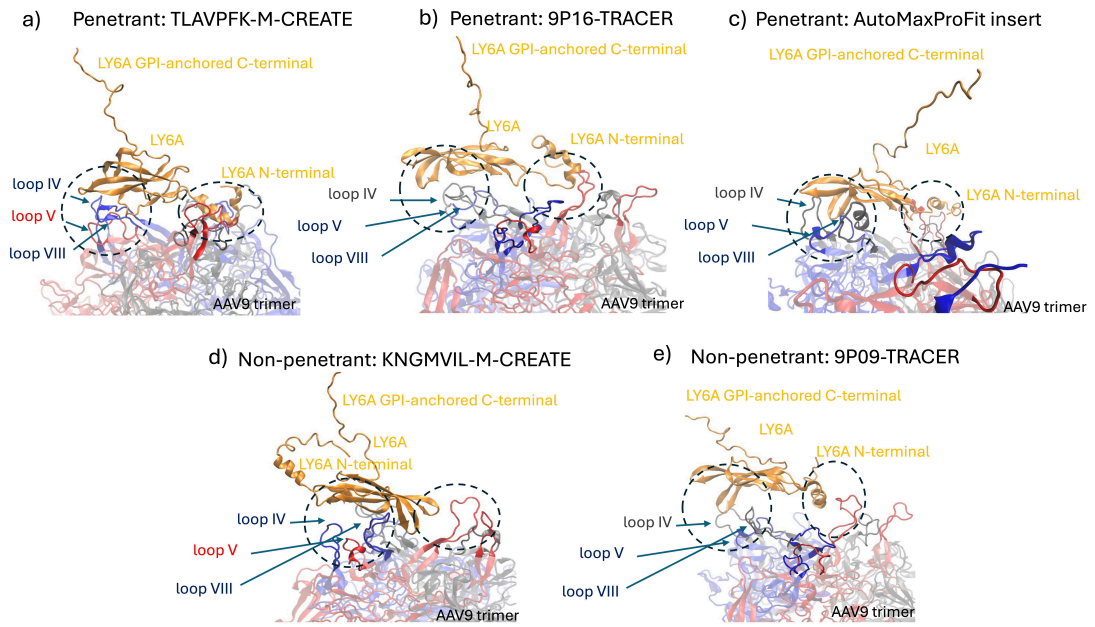

Figure 5: Final frames from a 20 ns MD trajectory. a) Optimal penetrant from M-CREATE complexed to LY6A; b) Optimal penetrant from TRACER complexed to LY6A; c) Optimal penetrant from AutoMaxProFit complexed to LY6A; d) A non-penetrant from M-CREATE complexed to LY6A; and e) A non-penetrant from TRACER complexed to LY6A. Notably, penetrants exhibit extensive engagement of the N-terminal of the LY6A receptor with the trimer, resulting in improved binding free energy and a larger buried surface area upon complex formation. In contrast, non-penetrants show interaction of the trimeric AAV9 with only loop VIII from one of the three chains in the trimer.

Discussion

In this paper, we illustrated how Deep Tech can build on the discoveries made by directed evolution studies in vivo. Similar value can be added on top of discoveries made in other studies that leveraged knowledge of BBB receptors and spatial omics studies that uncovered tissue distributions of penetrating AAV variants.

The usefulness of AutoMaxProFit approach extends much beyond design of BBB-crossing AAV variants. As noted earlier, enzyme optimization for maximizing catalytic activity while ensuring other properties are within desired ranges, can also be accomplished similarly.

In this study, we have limited validation to a rather simplistic MD simulation based approach. Multiple replicas and better sampling are needed to obtain more rigorous inferences in MD.

Given the promise of the results from this study, we hope to repeat the same on AAV variants identified via high-throughput experimentation, for penetration of BBB in NHP and humans. We also plan to conduct to a similar proof-of-concept study for enzyme optimization.

Methods

Featurization of the AAV capsid sequence

We scaled each dimension of the embeddings to [0,1] using minimum and maximum values found for each dimension across all the training dataset.

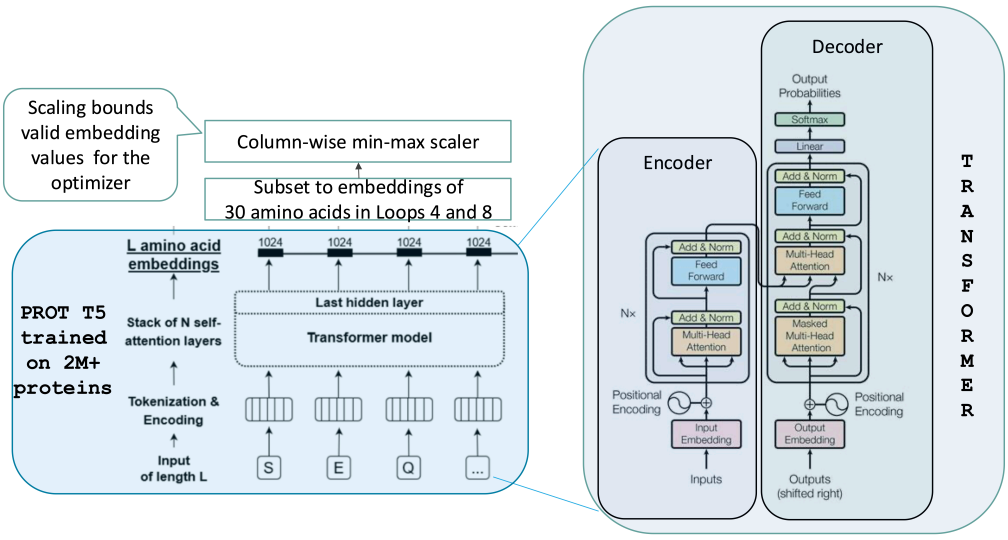

Figure 6: Featurization of AAV capsid sequence using a protein language model

Prediction model for enrichment score

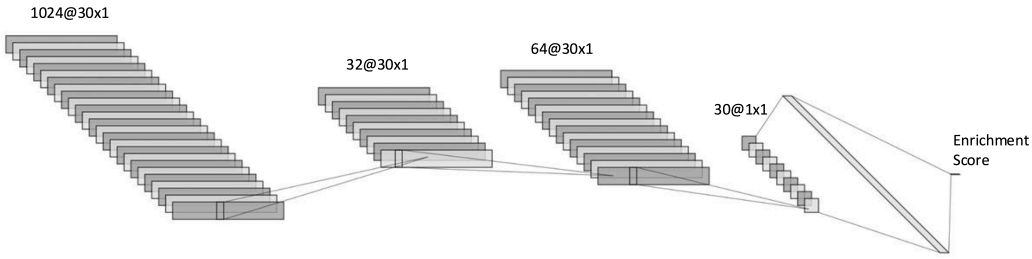

Figure 7: 1-D convolutional neural network (CNN) architecture used for training models to predict enrichment score of a given capsid sequence, for a target tissue/cell type.

Optimization using prediction model

Maximize

- Enrichment Score

By

- Gradient ascent in the sequence embedding space

Penalizing

- Deviation from valid embedding values ([0,1] as we used a min-max scaler)

$$y = f(x)$$

$$pos = max(0, x - 1)$$

$$neg = max(0, -x)$$

$$L(x) = -y + \lambda \|(pos + neg)\|$$

$$\underset{x}{argmin}\ L(x)$$

Figure 8: Loss function used in optimization

Molecular dynamics (MD) simulations

Structure preparation: The structure for PHP.eB variant of the AAV9 vector was extracted from PDB ID 7WQO [29]. To study tropism, the essential structure is the trimeric AAV9 in complex with the receptor LY6A. Experiments using AF2-multimer [30] (via Colabfold [31]) showed that while it can reliably predict a monomeric AAV9-LY6A complex, it struggles with predicting a complex involving three AAV9 and one LY6A chains. To address this, we constructed a trimeric AAV9 model using VIPERdb [32], then aligned this trimeric

AAV9 with the monomeric AAV9-LY6A complex. We saved the coordinates for the trimeric AAV9-LY6A complex, which was then energy minimised, equilibrated in water, and subjected to a ~20 ns production run.

Five systems were prepared, including four that served as the controls for tropism analysis via molecular dynamics (both penetrants and non-penetrants of the blood-brain barrier) and one that was AutoMaxProFit optimised penetrant for endothelial cells.

MD Simulation Details: The molecular dynamics (MD) simulations of the trimeric AAV9-LY6A complex in water was carried out in GROMACS 2022.3 [33, 34].

For protein, amber.ff99SB [35] force field was used. The complex was solvated in a dodecahedron box containing SPC/E water model under neutral conditions. The prepared systems contain more than 300,000 atoms in each case.

To ensure accurate dynamics, the system was first energy minimised using the steepest descents algorithm, followed by equilibration in two phases: NVT (constant volume and tempearature) and NPT (constant pressure and temperature).

The C-terminal of LY6A was kept fixed to mimic the restaurants due to glycosylphosphatidylinositol-anchored (GPI) lipid membrane. Since the LY6A receptor is not embedded directly inside the lipid membrane, no lipid bilayer was considered during system preparation. During NVT equilibration, temperature was controlled using the velocity rescaling thermostat [36] at 300 K. In the subsequent NPT equilibration, pressure was maintained at 1 bar using the Parrinello-Rehman barostat [37].

Production simulations were carried out in the NPT ensemble with a time step of 2 fs.

The Particle Mesh Ewald (PME) algorithm [38] was used to account for the long-range electrostatic interactions. Hydrogen atoms were constrained using the LINCS algorithm [39]. MMGBSA and BSA calculations were done on the generated trajectories to obtain the trimeric AAV9-LY6A binding affinities and the resulting surface area due to complexation. gmx_MMPBSA tool [40] was used for binding free energy calculations. Dr_SASA [41] was used on the final frame from MD trajectory to evaluate buried surface area between the trimeric AAV9 and the LY6A receptor. Root mean square deviations were calculated on the MD trajectory using GROMACS.

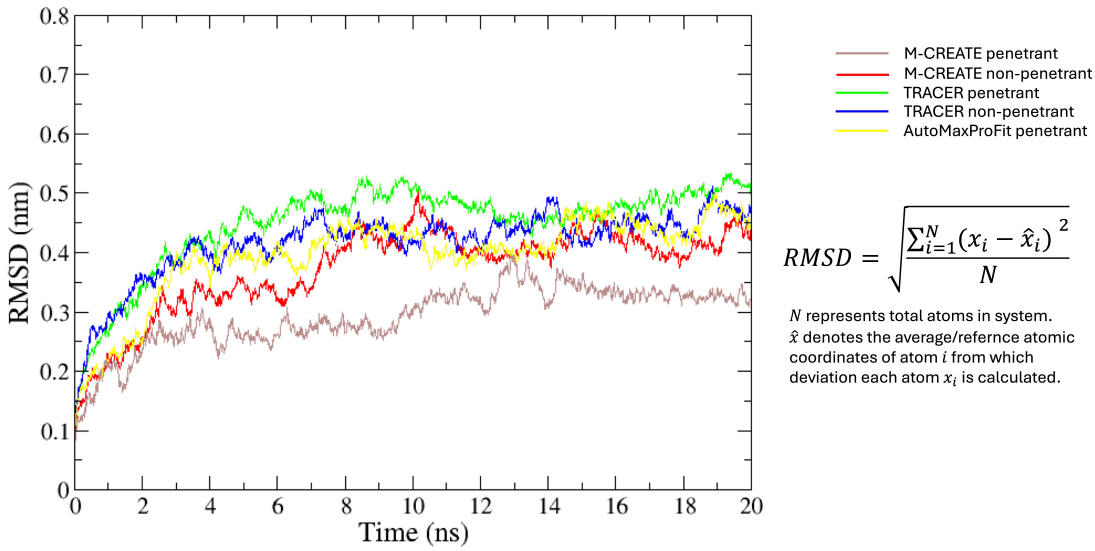

Figure 9: RMSD fluctuations for trimeric AAV9-LY6A complexes with time

AUTHOR DECLARATIONS

Conflict of Interest: All the authors are the employees of Aganitha, the company that holds interests in any commercial gains arising out of this work.

References

1. Duque S, Joussemet B, Riviere C, et al Intravenous Administration of Self-complementary AAV9 Enables Transgene Delivery to Adult Motor Neurons. Molecular Therapy 17:1187–1196. https://doi.org/10.1038/mt.2009.71

2. Chan Y A, Deverman B E Crossing the Blood-Brain Barrier with AAVs: What's After SMA?. In: Lange E C M de, Hammarlund-Udenaes M, Thorne R G (eds) Drug Delivery to the Brain: Physiological Concepts, Methodologies and Approaches. Springer International Publishing, Cham, pp 629–654

3. Challis R C, Ravindra Kumar S, Chen X, et al (2022) Adeno-Associated Virus Toolkit to Target Diverse Brain Cells. Annual Review of Neuroscience 45:447–469. https://doi.org/https://doi.org/10.1146/annurev-neuro-111020-100834

4. Deverman B E, Pravdo P L, Simpson B P, et al Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. Nature Biotechnology 34:204–209. https://doi.org/10.1038/nbt.3440

5. Ravindra Kumar S, Miles T F, Chen X, et al Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. Nature Methods 17:541–550. https://doi.org/10.1038/s41592-020-0799-7

6. Nonnenmacher M, Wang W, Child M A, et al Rapid evolution of blood-brain-barrier-penetrating AAV capsids by RNA-driven biopanning. Molecular Therapy Methods & Clinical Development 20:366–378. https://doi.org/10.1016/j.omtm.2020.12.006

7. Hordeaux J, Yuan Y, Clark P M, et al The GPI-Linked Protein LY6A Drives AAV-PHP.B Transport across the Blood-Brain Barrier. Molecular Therapy 27:912–921. https://doi.org/https://doi.org/10.1016/j.ymthe.2019.02.013

8. Huang Q, Chen A T, Chan K Y, et al (2022) Targeting AAV vectors to the CNS via de novo engineered capsid-receptor interactions. bioRxiv. https://doi.org/10.1101/2022.10.31.514553

9. Chen X, Ravindra Kumar S, Adams C D, et al Engineered AAVs for non-invasive gene delivery to rodent and non-human primate nervous systems. Neuron 110:2242–2257. https://doi.org/10.1016/j.neuron.2022.05.003

10. Chuapoco M R, Flytzanis N C, Goeden N, et al Adeno-associated viral vectors for functional intravenous gene transfer throughout the non-human primate brain. Nature Nanotechnology 18:1241–1251. https://doi.org/10.1038/s41565-023-01419-x

11. Chen X, Wolfe D A, Bindu D S, et al Functional gene delivery to and across brain vasculature of systemic AAVs with endothelial-specific tropism in rodents and broad tropism in primates. Nature Communications 14:3345–3346. https://doi.org/10.1038/s41467-023-38582-7

12. Huang Q, Chan K Y, Wu J, et al An AAV capsid reprogrammed to bind human transferrin receptor mediates brain-wide gene delivery. Science 384:1220–1227. https://doi.org/10.1126/science.adm8386

13. Shi H, He Y, Zhou Y, et al Spatial atlas of the mouse central nervous system at molecular resolution. Nature 622:552–561. https://doi.org/10.1038/s41586-023-06569-5

14. Bryant D H, Bashir A, Sinai S, et al Deep diversification of an AAV capsid protein by machine learning. Nature Biotechnology 39:691–696. https://doi.org/10.1038/s41587-020-00793-4

15. Damani F, Brookes D H, Sternlieb T, et al (2023) Beyond the training set: an intuitive method for detecting distribution shift in model-based optimization. https://arxiv.org/abs/2311.05363

16. Eid F-E, Chen A T, Chan K Y, et al Systematic multi-trait AAV capsid engineering for efficient gene delivery. Nature Communications 15:6602–6603. https://doi.org/10.1038/s41467-024-50555-y

17. Huang Q, Chen A T, Chan K Y, et al Targeting AAV vectors to the central nervous system by engineering capsid–receptor interactions that enable crossing of the blood–brain barrier. PLOS Biology 21:1–30. https://doi.org/10.1371/journal.pbio.3002112

18. Ding X, Chen X, Sullivan E E, et al Fast, accurate ranking of engineered proteins by target-binding propensity using structure modeling. Molecular Therapy 32:1687–1700. https://doi.org/10.1016/j.ymthe.2024.04.003

19. Tran T V T, Hy T S (2024) Protein Design by Directed Evolution Guided by Large Language Models. bioRxiv. https://doi.org/10.1101/2023.11.28.568945

20. Elnaggar A, Heinzinger M, Dallago C, et al (2022) ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 44:7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

21. Paszke A, Gross S, Chintala S, et al (2017) Automatic differentiation in PyTorch. In: NIPS-W

22. Yin R, Zhou X, Zheng J, Kwoh C K (2018) Computational identification of physicochemical signatures for host tropism of influenza A virus. Journal of bioinformatics and computational biology 16:1840023–1840024. https://doi.org/10.1142/S0219720018400231

23. Saha P, Majumder R, Srivastava A K, et al (2020) Mutations in spike protein of sars-cov-2 modulate receptor binding, membrane fusion and immunogenicity: An insight into viral tropism and pathogenesis of covid-19. ChemRxiv. https://doi.org/10.26434/chemrxiv.12320567.v1

24. Ode H, Nakashima M, Kitamura S, et al (2012) Molecular dynamics simulation in virus research. Frontiers in microbiology 3:258–259. https://doi.org/10.3389/fmicb.2012.00258

25. Yokoyama M, Fujisaki S, Shirakura M, et al (2017) Molecular dynamics simulation of the influenza A (H3N2) hemagglutinin trimer reveals the structural basis for adaptive evolution of the recent epidemic clade 3C. 2a. Frontiers in Microbiology 8:584–585. https://doi.org/10.3389/fmicb.2017.00584

26. Li Y, Deng L, Yang L-Q, et al (2019) Effects of CD4 binding on conformational dynamics, molecular motions, and thermodynamics of HIV-1 gp120. International Journal of Molecular Sciences 20:260–261. https://doi.org/10.3390/ijms20020260

27. Sang P, Yang L-Q, Ji X-L, et al (2014) Insight derived from molecular dynamics simulations into molecular motions, thermodynamics and kinetics of HIV-1 gp120. PLoS One 9:e104714. https://doi.org/10.1371/journal.pone.0104714

28. Vaswani A, Shazeer N, Parmar N, et al (2023) Attention Is All You Need. https://arxiv.org/abs/1706.03762

29. Xu G, Zhang R, Li H, et al Structural basis for the neurotropic AAV9 and the engineered AAVPHP.eB recognition with cellular receptors. Molecular Therapy Methods & Clinical Development 26:52–60. https://doi.org/10.1016/j.omtm.2022.05.009

30. Evans R, O'Neill M, Pritzel A, et al (2022) Protein complex prediction with AlphaFold-Multimer. bioRxiv. https://doi.org/10.1101/2021.10.04.463034

31. Mirdita M, Schütze K, Moriwaki Y, et al ColabFold: making protein folding accessible to all. Nature Methods 19:679–682. https://doi.org/10.1038/s41592-022-01488-1

32. Montiel-Garcia D, Santoyo-Rivera N, Ho P, et al (2021) VIPERdb v3. 0: a structure-based data analytics platform for viral capsids. Nucleic Acids Research 49:D809–D816

33. Berendsen H J, Spoel D van der, Drunen R van (1995) GROMACS: A message-passing parallel molecular dynamics implementation. Computer physics communications 91:43–56

34. Bauer P, Hess B, Lindahl E (2022) GROMACS 2022.3 Source code

35. Lindorff-Larsen K, Piana S, Palmo K, et al (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins: Structure, Function, and Bioinformatics 78:1950–1958

36. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. The Journal of chemical physics 126:14101–14102. https://doi.org/10.1063/1.2408420

37. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied physics 52:7182–7190

38. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. The Journal of chemical physics 98:10089–10092

39. Hess B, Bekker H, Berendsen H J, Fraaije J G (1997) LINCS: a linear constraint solver for molecular simulations. Journal of computational chemistry 18:1463–1472

40. Valdés-Tresanco M S, Valdés-Tresanco M E, Valiente P A, Moreno E (2021) gmx_MMPBSA: a new tool to perform end-state free energy calculations with GROMACS. Journal of chemical theory and computation 17:6281–6291

41. Ribeiro J, R\ios-Vera C, Melo F, Schüller A (2019) Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions. Bioinformatics 35:3499–3501