Mapping AAV capsid sequences to functions through function-guided *in silico* evolution

Hanyu Zheng[1,2,3,4,5,10], Binjie Guo[1,2,3,4,5,10], Aisheng Mo[1,2,3,4,5], Hongyan Wei[1,2,3,4,5], Yile Wu[3,4], Xurong Lin[1,2,3,4,5], Haohan Jiang[1,2,3,4,5], Hengguang Li[1,2,3,4,5], Yunshuo Zhang[1,3,4], Zhuoyuan Song[1,3,4], Xuebin Ni[1,3,4], Yan Huang[6], Xiaosong Gu[7], Bin Yu[7], Ningtao Cheng[8]*, Xuhua Wang[1,2,3,4,5,9,11,*]

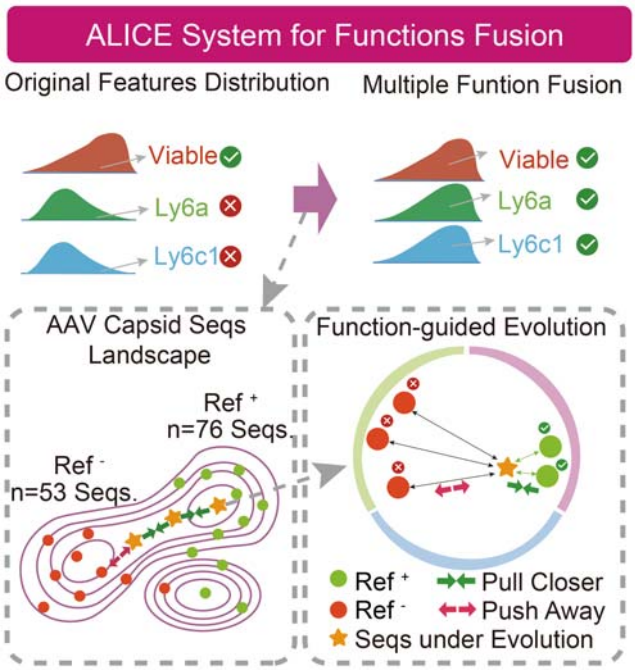
Affiliations:

1. Department of Neurobiology and Department of Rehabilitation Medicine, First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang Province 310003, P. R. China
2. Nanhu Brain-computer Interface Institute, Hangzhou 311100, China
3. Liangzhu Laboratory, MOE Frontier Science Center for Brain Science and Brain-machine Integration, State Key Laboratory of Brain-machine Intelligence, Zhejiang University, 1369 West Wenyi Road, Hangzhou 311121, China
4. NHC and CAMS Key Laboratory of Medical Neurobiology, Zhejiang University, Hangzhou 310058, China
5. Lingang Laboratory, Shanghai 200031, China
6. Department of Hepatobiliary and Pancreatic Surgery, Affiliated Hospital of Nantong University, Medical School of Nantong University, Nantong, 226001, China
7. Key Laboratory of Neuroregeneration of Jiangsu and Ministry of Education, Nantong University, Nantong, 226001 Jiangsu, P. R. China
8. School of Public Health, Zhejiang University School of Medicine, 866 Yuhangtang Road, Hangzhou, 310058, Zhejiang, People's Republic of China
9. Co-innovation Center of Neuroregeneration, Nantong University, Nantong, 226001 Jiangsu, P. R. China
10. These authors contributed equally
11. Lead contact

* Correspondence: xhw@zju.edu.cn (X.W.); ncheng@zju.edu.cn (N.C.)

Correspondence should be addressed to X.W. (email: xhw@zju.edu.cn).

Graph Abstract:



In Brief:

This study introduces ALICE, a novel machine-intelligent design method that enables direct *in silico* engineering of adeno-associated virus (AAV) capsids with multiple functions. *In vivo* experiments have validated the enhanced viability and central nervous system tropism of the variants engineered through ALICE.

Highlights:

• ALICE represents a pioneering machine-intelligent design method for direct *in silico* engineering of AAV capsids with multiple functions.

• Utilizing a limited dataset of 129 entries, ALICE successfully maps AAV capsid sequences to multiple functions.

• We elucidated the evolutionary mechanism of capsid sequences toward desired functions within the ALICE system.

• We revealed that the AAV variants engineered by ALICE exhibit increased viability and about 372-fold increased central nervous system tropism.

Keywords:

in silico engineering of the AAV capsid; interpretable artificial intelligence; sequence-to-function map; function-guided evolution; *in silico* biomolecular evolution mechanism; multifunction fusion; production fitness; CNS tropism.

Summary

Artificial intelligence (AI) has been suggested to facilitate time- and cost-effective functional engineering of adeno-associated virus (AAV) capsid sequences. Nevertheless, an AI-empowered approach to identify AAV capsid sequence-to-multifunction relationships remains elusive. To overcome this challenge, we propose a machine-intelligent design method to map an AAV capsid sequence to multiple functions, thereby enabling direct *in silico* engineering of AAV capsids. To fuse multiple functions into a single capsid sequence, a heuristic algorithm coupled with contrastive learning and reinforcement learning, named function-guided evolution (FE), was introduced to steer further evolution of the high-performing capsid sequences generated by a naïve language model toward functions. We then illustrated the evolutionary mechanism of the FE approach for function-guided generation of capsid sequences. Further optimization steers the evolution toward desired functions within a function-guided landscape. Despite the constraint of datasets of only 129 entries, we successfully constructed a model to map AAV capsid sequences to multiple functions of improved viability coupled with central nervous system (CNS) tropism. *In vivo* experiments confirmed that two of the top eight engineered variants exhibited enhanced viability and remarkable CNS tropism. This interpretable machine-intelligent design method represents a pioneering effort enabling direct *in silico* engineering of AAV capsids for effective gene delivery.

Introduction

In recent years, advances in artificial intelligence (AI) technology have significantly propelled the application of AI in functional protein design.[1-4] However, designing multifunctional large natural protein complexes, such as adeno-associated virus (AAV) capsids, remains a formidable challenge.[5] Over the past decade, conventional methodologies such as rational design[6-9] and directed evolution[10-15] have been developed to screen AAV capsids to overcome current limitations such as low viability and suboptimal tissue tropism. Nevertheless, these approaches, which are heavily dependent on wet-lab experiments, inherently suffer from inadequacy in exhaustively exploring the vast capsid sequence space of AAV variants.[16,17] Recently, computational methods have been successfully employed to engineer functional proteins under the guidance of sequence-to-function maps.[18-23] With deeper exploration into the mechanisms of action of AAV capsids,[24-27] we envisioned that a computational method might be able to map capsid sequences to their biological functions, enabling direct *in silico* engineering of AAV capsids.

Nonetheless, to map capsid sequences to functions, we anticipate that two challenges must be overcome: (i) the absence of a framework capable of mapping a single capsid sequence to multiple functions imperative for gene delivery[28] and (ii) the interpretability of the mechanism by which capsid sequences evolve into functional entities within a computational model.[16,29] Understanding their evolutionary mechanism within a computational model could guide the optimization of such a model to map sequences to desired functions. To overcome these challenges, we have developed a generative AI architecture, termed artificial intelligence custom engineering (ALICE), which is designed to map capsid sequences to multiple functions with limited data. In ALICE, a naïve generative language model is initially pretrained on a dataset containing 2.89 million peptide sequences to understand the nuances of natural peptide linguistics. Through transfer learning,[30] it was then semantically tuned on a capsid sequence dataset of 72,753 entries to learn the biology of AAV capsids. However, the naïve language model, which employs a strategy of pretraining followed by semantic-tuning, faces the challenge of mapping a single capsid sequence to multiple functions.

To address this challenge, a heuristic algorithm[31,32] coupled with contrastive learning[33] and reinforcement learning,[34] termed function-guided evolution (FE), was introduced to steer further evolution of these high-performing capsid sequences generated by the naïve language model toward functions with small datasets of $\boldsymbol{Ref}^{+}$ (n = 76 seqs) and $\boldsymbol{Ref}^{-}$ (n = 53 seqs). These $\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$ datasets provide functional information on capsid sequences, including their production fitness, ability to bind to Ly6c1 and Ly6a receptors enriched on the blood□brain barrier (BBB) of mice and performance in animals.[26,35] To optimize the FE module, we illustrated the evolutionary mechanism of capsid sequences toward functions through different algorithm approaches within the FE module. Guided by these mechanisms, we steered the FE module to map capsid sequences to multiple functions of enhanced viability coupled with CNS tropism.

After an elite ascendancy screening algorithm was used, ALICE identified 8 top potential sequences across a 20^7-dimensional sequence space. We evaluated the potency of ALICE via *in vivo* testing of these AAV variants with insertional modifications of the sequences between residues 588–589 within VP1 of AAV9 capsid individually. Two of these variants, AAV.ALICE-N2 and AAV.ALICE-N6, hitherto unreported, demonstrated enhanced viability and remarkable CNS tropism, surpassing the original AAV9 capsid. Notably, AAV.ALICE-N2 resulted in an approximate twofold increase in viability and a remarkable 372-fold increase in transduction efficiency in CNS cells compared with those of the original AAV9. More importantly, these variants effectively transduce CNS cells across various mouse strains through purposefully designed mechanisms of action by targeting both Ly6c1 and Ly6a receptors. The interpretable *in silico* AAV capsid engineering methodology described herein represents a breakthrough in the engineering of AAV capsids with desired functions and provides insights into other functional protein sequence designs.

Results

The design rationale and workflow of the ALICE platform

It is believed that the amino acid sequences within proteins share an intrinsic resemblance with the architecture of human language, with each sequence akin to the alphabet of a verbal dialect.[2,36] In our pilot experiment, we found that the peptide sequences used for the insertion modification of AAV capsids, which exhibit high production fitness and binding affinity for Ly6a and Ly6c1 receptors, display similar architectures of protein and human language (Figure S1 and Table S1). This finding motivated us to pretrain a generative language model with a large peptide sequence dataset (Figure 1A) and then semantically tune it with a limited AAV insertion peptide sequence dataset to generate functional capsid sequences (Figure 1B). We then attempted to use a ranking filtration process module to rank the generated sequences on the basis of their multifunctional properties for further evolution (Figure 1C). By combining a heuristic algorithm[31,32] with contrastive learning[33] and reinforcement learning,[34] we subsequently sought to steer the evolution of the top-performing sequences generated by the naïve language model toward the desired functions (Figure 1D). Finally, we sought to evaluate the comprehensive functions of the sequences through an elite ascendancy module (Figure 1E) and select the top candidates for wet experiment verification (Figure 1F). Thus, the workflow of the ALICE system integrates 6 essential steps: (i) pretraining, (ii) semantic tuning, (iii) ranking filtration process (RankFlitPro), (iv) function-guided evolution (FE), (v) elite ascendancy, and (vi) wet-lab validation (Figure 1).

Generation of datasets and establishment of virtual evaluation methods

To train ALICE, we generated four main datasets: (i) 2.89 million peptide sequences from UniProt[37,38] ($\boldsymbol{D_{Pretrain}}$) for pretraining (details in STAR Methods) and (ii) a publicly accessible dataset of 7-mer amino acid sequences, which were inserted between residues 588–589 in VP1 for AAV9 capsid modification: $\boldsymbol{Lib_{seq}}$.[25] These data include capsid sequence insertion-modified variant production fitness, binding affinity to the target proteins Ly6a (widely expressed in C57BL/6J and FVB/NJ mice) and Ly6c1 (expressed in Ly6a-deficient BALB/cJ mice).[25] The capsid sequences with high production fitness and target binding ability are selected for semantic tuning in the dataset of $\boldsymbol{D_{seq}}$ (n = 72,753, details in STAR Methods). (iii) To promote capsid sequence evolution toward desired functions, we prioritized sequences in the FE reference dataset by focusing on two key factors: (1) high viability and strong binding affinity to Ly6a and Ly6c1 and (2) capsid sequences that perform the desired functions in animals reported in previous studies.[12,35,39,40] Sequences meeting these criteria were designated $\boldsymbol{Ref^{+}}$ (n = 76, Table S1), whereas those exhibiting weak binding to either Ly6a or Ly6c1 and low viability were classified as $\boldsymbol{Ref^{-}}$ (n = 53, Table S1, detailed in the STAR Methods). (iv) To evaluate the production fitness, Ly6a binding, and Ly6c1 binding capabilities of the designed capsid sequences, three datasets ($\boldsymbol{D_{Production\ Fitness}}$, $\boldsymbol{D_{Ly6a}}$ and $\boldsymbol{D_{Ly6c1}}$) were constructed from $\boldsymbol{Lib_{seq}}$ (details in the STAR Methods).

To virtually evaluate the functionalities of the sequences designed by each module within ALICE and thereby optimize them, we established predictive methodologies to comprehensively examine the functionalities of the designed sequences. This includes evaluating the fitness of the sequence insertion-modified variants, as well as their ability to bind to the targets of Ly6a or Ly6c1. To identify the best prediction model, we selected a total of seven widely used candidate machine learning models, including support vector regression (SVR),[41] eXtreme gradient boosting (XGBoost),[42] K-nearest neighbors (KNN),[43] random forest (RF),[44,45] gradient boosting (GB),[46,47] Bayesian Ridge,[48] and AdaBoost.[45] To evaluate the performance of each model, six evaluation indicators (Pearson, Spearman, RMSE, R^2, MedAE, and C-index) were selected. We trained and tested these models on the $\boldsymbol{D_{Production\ Fitness}}$ (Figure S2A), $\boldsymbol{D_{Ly6a}}$ (Figure S3A) and $\boldsymbol{D_{Ly6c1}}$ (Figure S4A) datasets to predict production fitness, Ly6a binding, and Ly6c1 binding capabilities, respectively (details in the STAR Methods). We found that the predictive performance of the GB model was the best among these candidate models (Figure S2-4), with Pearson correlations of 0.96, 0.81, and 0.81, respectively (Table S4-6). Therefore, we use the GB model in the following evaluation tasks.

The pretraining followed by a semantic-tuning strategy confronts a multifeature fusion conflict issue

To train a language model to learn and generate functional capsid sequences, we selected the notable models of BERT[49] and RoBERTa[50] as candidate models for the pretraining module (Figure 1A, details in the STAR Methods) and employed a SeqGAN architecture[51,52] in the semantic tuning module (Figure 1B). To investigate the necessity of the pretraining module, we train a SeqGAN model directly on $\boldsymbol{D_{seq}}$ without pretraining (SeqGAN-only). To optimize the pretraining module, we embedded the pretraining models of BERT and RoBERTa that were trained on $\boldsymbol{D_{Pretrain}}$ into the SeqGAN framework (Figure S5A).

Through uniform manifold approximation and projection (UMAP)[53] analysis of the properties of the generated sequences, we found that the features of the sequences generated by the RoBERTa-SeqGAN are more segregated than those generated by the SeqGAN-only or BERT-SeqGAN (Figure S5B-D). This suggests that RoBERTa-SeqGAN can generate sequences

with features different from those in the original $\boldsymbol{D_{seq}}$. However, the amino acid bias of the sequences generated by RoBERTa-SeqGAN was closer to those in the original training dataset $\boldsymbol{D_{seq}}$ than those generated by SeqGAN-only or BERT-SeqGAN (Pearson = 0.835 vs 0.643 or 0.158, Figure S6A-F). This suggests that RoBERTa-SeqGAN has learned the semantic and syntactic information inherent in natural analogues. By evaluating the functions of the generated sequences with the virtual evaluation model of GB, we found that, compared with those of the training sequences from $\boldsymbol{D_{seq}}$ (Figure S5E), the functionalities, including the production fitness and Ly6a and Ly6c1 target binding, of the sequences generated by SeqGAN-only did not significantly improve (Figure S5E-H, J), highlighting the critical role of pretraining. In contrast, while the production fitness of sequences generated by BERT-SeqGAN marginally increased, this resulted in a considerable cost to the binding affinity for both Ly6a and Ly6c1 targets (Figure S5E-H, K). Interestingly, the sequences generated by RoBERTa-SeqGAN exhibited improved binding capacity to both Ly6a and Ly6c1 targets but suffered from profoundly diminished production fitness (Figure S5E-H, L). These findings align with the observation of an antagonistic relationship between the target binding capacity and production fitness in natural capsid sequences (Figure S5I), pointing to a complex issue of multifeature fusion conflict in mapping a single capsid sequence to multiple functions with a naïve language model.

Promotion of the evolution of capsid sequences via a function-directed evolution strategy

Inspired by the mechanism of biological evolution, in which organisms undergo evolutionary breakthroughs under environmental pressures[54] and learn from limited samples to survive and adapt, we attempt to develop a new algorithmic paradigm simulating natural evolutionary progress to solve the multifeature fusion conflict issue in mapping single sequences to multiple functions. To do so, we first employed a ranking filtration process module to select the top 1,000 capsid sequences (for details, see the STAR Methods) generated by the RoBERTa-SeqGAN model, prioritizing those with high binding affinity and production fitness for further evolution (Figure 1C). Then, we introduced a function-guided evolution (FE) module by coupling a heuristic algorithm[31,32] with contrastive learning[33] and reinforcement learning[34] to improve the evolution of these sequences to desired functions. To control the sequences' evolution toward desired functions within the FE module, we used the positive ($\boldsymbol{Ref}^{+}$, n = 76) and negative ($\boldsymbol{Ref}^{-}$, n = 53) sample datasets to guide their evolution direction. The sequences in $\boldsymbol{Ref}^{+}$ or $\boldsymbol{Ref}^{-}$ are segregated by their features of production fitness, multiple target binding capacities and reported capsid sequences verified by *in vivo* studies (Figure 2A, details in the Supplementary Method). To quantify the evolutionary progress and guide further optimization, we employed the multisequence alignment (MSA) method to calculate the similarity of capsid sequences under evolution to the reference sequences ($\boldsymbol{Ref}^{+}$ or $\boldsymbol{Ref}^{-}$). The function-guided evolution function (FE function) is defined as follows:

$$max: FE_function(Seq, Ref^{+}, Ref^{-}) \propto MSA(Seq, Ref^{+}) - MSA(Seq, Ref^{-})$$

where MSA indicates the multisequence alignment algorithm implemented by the Smith–Waterman algorithm.[55] This function enables adaptive search and assessment of landscape distances between the candidate capsid sequences and reference sequences (Figure 2A). Furthermore, we employed a functional evolution score (FE score, details in STAR Methods) calculated from the FE function to quantify and analyze the performance of sequences during evolution. These FE scores reflect the traits learned from both positive ($\boldsymbol{Ref}^{+}$) and negative ($\boldsymbol{Ref}^{-}$) reference sequences.

To identify the most suitable models for the FE module, we initially employed four well-established heuristic algorithms: simulated annealing-based (SA),[32] genetic algorithm-based (GA),[31] estimation of distribution algorithm (EDA),[56] and a hybrid method combining the estimation of the distribution algorithm and genetic algorithm (EDG).[56] Among these algorithms, only the EDA and EDG algorithms demonstrated rapid convergence (Figure 2B), achieving high FE scores. Moreover, the capsid sequences generated by EDA and EDG exhibited more concentrated distributions than those generated by $\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$ (Figure 2C-F), which is in agreement with the high FE scores of the sequences identified by EDA or EDG (Figure 2B). Thus, we focused further optimization of the FE module on the EDA and EDG models.

Refining the FE module by exploring the evolutionary mechanisms of capsid sequences within the model

To elucidate the mechanisms underlying the evolution of capsid sequences within the FE module, we analyzed the evolutionary trajectory of the capsid sequences with increasing number of epochs (Figure 3A-D). By plotting the distances between the capsid sequences under evolution and the positive references ($\boldsymbol{Ref}^{+}$) against those with negative references ($\boldsymbol{Ref}^{-}$), we visualized the evolutionary trajectory across multiple well-established metrics (details in the STAR Method section "Metrics"): sequence evolutionary divergence (Figure 3A, Kullback□Leibler (KL) divergence),[57] functional similarity (Figure 3B, Blocks Substitution Matrix (BLOSUM) 62 similarity),[58] and both local (Figure 3C, N-gram similarity)[59] and global (Figure 3D, longest common subsequence (LCS)[60] semantic similarity. We designated the evolutionary changes with reference to $\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$ as $\Delta\mathrm{PE}$ (positive evolutionary change) and $\Delta\mathrm{NE}$ (negative evolutionary change), respectively.

Ideally, FE should drive sequence evolution toward $\boldsymbol{Ref}^{+}$ sequences with enhanced functional properties while diverging from $\boldsymbol{Ref}^{-}$ (Figure 3A, B, green arrows). However, the EDA-based approach drives sequence evolution in a divergent direction that is far from the ideal direction during the first 5 epochs (Figure 3A, B). It then exhibited a substantial shift of approximately 180 degrees after 5 epochs, resulting in sequences with diminished functional correlations with positive samples (Figure 3A, B). Similar trends were also observed from a semantic perspective (Figure 3C-D). On the basis of these observations, we defined it as a form of partial evolution with an EDA-based approach (Figure 3A, B). In contrast, the EDG-based approach steered sequence evolution toward $\boldsymbol{Ref}^{+}$ and away from $\boldsymbol{Ref}^{-}$, particularly during the initial 5 epochs (Figure 3A, B). This initial phase was characterized by a nearly equivalent $\Delta\mathrm{PE}$:$\Delta\mathrm{NE}$ ratio with respect to KL divergence (Figure 3A). Following epoch 5, it not only propelled evolution toward $\boldsymbol{Ref}^{+}$ but also simultaneously fostered a shift toward $\boldsymbol{Ref}^{-}$, resulting in final $\Delta\mathrm{PE}$:$\Delta\mathrm{NE}$ ratios of approximately 1.5:1 (KL divergence, Figure 3A) and 3.7:1 (BLOSUM62 similarity, Figure 3B). Despite this, the evolutionary efficiency demonstrated a more significant increase in positive samples, with a notable increase in the functional correlation to $\boldsymbol{Ref}^{+}$ from -6.578 to 1.439, whereas the correlation with $\boldsymbol{Ref}^{-}$ exhibited only a modest shift from -8.002 to -5.853 (Figure 3B). This indicated a learning mechanism within the FE module, in which capsid sequences adeptly incorporate more beneficial functional features from $\boldsymbol{Ref}^{+}$ while making minor concessions in terms of resistance to the nonfunctional features of $\boldsymbol{Ref}^{-}$. Intriguingly, the EDG-guided evolution revealed an unexpected trend: sequences under evolution showed increasing similarity to both positive and negative samples from local and global semantic perspectives (Figure 3C-D, Figure S7). We reason that this phenomenon stems from the inherent nature of capsid sequences. The sequences from $\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$, despite their functional

divergence, likely adhere to the common structural and semantic principles[61,62] of capsid proteins because both of them are naturally occurring sequences. The EDG-based FE module appears to capture these shared characteristics, essentially learning a general 'grammar' of capsid sequences. Within this framework, the functional differences in sequences between $\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$ manifest as distinct 'expressions' of this common grammar, which is consistent with previous findings.[63] This understanding enables this model to generate diverse sequences[64] that conform to naïve capsid protein rules while exploring the functional spectrum between $\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$. These findings are further supported by comparable observations when single positive and negative samples are utilized for trajectory tracking (Figure S8). On the basis of these findings, we use the EDG algorithm for the FE module in the following study.

Resolving the conflict in multifunction fusion through evolution across a function-guided landscape

To determine the biological implications of the capsid sequences generated through the selected EDG-based approach, we sought to analyze them from both functional and semantic perspectives. We found, from a functional perspective, that despite similar evolutionary distances to positive and negative samples, these sequences exhibited greater functional similarity to $\boldsymbol{Ref}^{+}$ (Figure 3E, top panel). Interestingly, we found that the capsid sequences under evolution exhibit a heightened sensitivity to local semantic information (N-gram similarity, Figure 3E, bottom panel), as opposed to global information (LCS ratio, Figure 3E, bottom panel), in agreement with previous observations that the functionality of proteins is dictated by specific regions or motifs within the sequence (local semantic information) rather than the overall sequence semantics.[65,66] This sensitivity to local motifs is crucial, as the AAV capsid protein often contains specific motif substructures.

To visualize the evolutionary process of capsid sequences within the model, we depicted the dynamic landscape of capsid sequences as they evolved from epoch 0 (near $\boldsymbol{Ref}^{-}$) to epoch 25 (close to $\boldsymbol{Ref}^{+}$) (Figure 3F). We found that throughout the evolutionary process, both the FE score and production fitness of the sequences improved significantly (Figure S9A). We also found that FE ensures that evolved capsids adhere to specified functional goals without introducing unintended features (Figure S10A-D). As the diversity of the sequences increased gradually, they evolved into novel sequences featuring the 'GYSS' motif, which extends beyond the patterns observed in reference sequences (Figure S9B). Thus, the FE adeptly recognizes local semantic structures in close proximity to positive samples, thus gradually exploring the functional motifs related to positive samples while diverging from negative ones (Figure S9B). This approach successfully integrated production fitness and multiple-target binding ability simultaneously into the capsid sequences (Figure 3G). In conclusion, through evolution across a function-guided landscape, FE elegantly resolved the conflict inherent in multifunction fusion (Figure S5I-L), thereby enabling the mapping of a single sequence to multiple functions.

An ablation study interprets the contribution of each module to the evolutionary landscape

To elucidate the contribution of each module within the ALICE architecture, we conducted a comprehensive ablation study encompassing nine recombinant architectures for mapping capsid sequences to functions. The performance of the sequences identified by these architectures was quantitatively assessed via production fitness evaluation and FE scoring methods. The results revealed that the RoBERTa+SeqGAN+RankFiltPro+FE (RSRF) combination outperformed all other

architectures in terms of the FE score while maintaining competitive production fitness (Figure 4A). Further analysis revealed that sequences identified by different architecture combinations presented distinct clustering patterns (Figure 4B). Notably, we found that sequences generated by the architectures lacking the FE module gravitated toward the distribution of $\boldsymbol{Ref}^-$ and away from that of $\boldsymbol{Ref}^+$ (Figure 4B). This observation aligns with previous findings that, without the guidance of the FE module, a naïve generative language model employing the pretraining and fine-tuning strategy fails to effectively address the issue of multiple function fusion conflicts (Figure S5I-L). In contrast, combinations incorporating the FE module consistently displayed a strong tendency toward $\boldsymbol{Ref}^+$, with the RSRF combination exhibiting the closest proximity to $\boldsymbol{Ref}^+$ (Figure 4B). This underscores the crucial role of the FE module in steering the evolutionary process of sequences toward desired functions (Figure 3G). FE also helps identify sequences that exhibit low-level, stable single-site information entropy (Figure S11A), particularly at positions 4 to 7. The RSRF combination demonstrated optimal stability while maintaining a balance between diversity and specificity in amino acid distribution (Figure S11B). Notably, sequence logos highlighted the recurrent presence of the GYSS motif in the terminal four sites (Figure S9B), a motif implicated in prior pulldown experiments as potentially facilitating interaction with the Ly6c1 target.[25]

To investigate the specific influence of the FE module on the ablation combinations, we conducted further analysis of the evolutionary trajectories of the identified sequences related to the $\boldsymbol{Ref}^+$ and negative $\boldsymbol{Ref}^-$ samples. We defined four types of evolutionary trajectories (Figure 4C) on the basis of the comprehensive consideration of evolutionary degree (KL divergence), functional characteristics (BLOSUM62 similarity), and semantic information (N-gram, LCS). The sequences generated via the NNK method served as controls. We found that the absence of the FE module in combination resulted in a Type III pattern characterized by a progressive deviation from both positive and negative samples, as indicated by metrics such as KL divergence, BLOSUM62 similarity, and semantic information (Figure S12A-D), indicating diminished mapping performance (Figure 4A). Interestingly, from a semantic perspective, combinations including the FE module displayed Type I behavior: approaching both $\boldsymbol{Ref}^+$ and $\boldsymbol{Ref}^-$, albeit with a greater magnitude toward $\boldsymbol{Ref}^+$ (Figure S12C, D; Figure S11D, E). This suggests that the FE module adeptly captures both global structural similarities and local functional differences. The FE module pays attention to the overall semantic structure that shares commonalities with both positive and negative samples because of their shared 'capsid grammar'. Concurrently, it preserves a heightened sensitivity to local semantic details, particularly functional motifs that distinguish $\boldsymbol{Ref}^+$ from $\boldsymbol{Ref}^-$. This allows the FE module to find sequences that maintain some general semantic features common to both $\boldsymbol{Ref}^+$ and $\boldsymbol{Ref}^-$ (Figure S11D, E) while simultaneously incorporating specific local motifs crucial for desired functions (as evidenced by the frequent appearance of the GYSS motif). This balance enables the FE module to find sequences that are functionally similar to $\boldsymbol{Ref}^+$ (as shown by KL divergence and BLOSUM62 similarity) while still exploring novel semantic spaces (Figure S13). On the other hand, from a functional perspective, combinations incorporating the FE but lacking the pretraining module exhibited Evo-Type I behavior: evolving toward both $\boldsymbol{Ref}^+$ and $\boldsymbol{Ref}^-$ (Figure 4E, yellow arrow). In contrast, combinations including both FE and pretraining modules demonstrated a remarkable shift toward Type II behavior: evolving toward $\boldsymbol{Ref}^+$ while moving away from $\boldsymbol{Ref}^-$ (Figure 4D, blue arrow; Figure S12A). This observation underscores the positive impact of the pretraining strategy in guiding multifunctional fusion evolution toward the desired functions. The architectures, including the FE model, also demonstrate the ability to map sequences with motif segments that satisfy the fusion

of multiple functions while maintaining distinct distributions from the training data (Figure S13A, B), effectively balancing the novelty of the sequences and enhancing the desired functions. In conclusion, we found that the FE and pretraining modules work synergistically in steering the evolution of capsid sequences toward the desired functions.

In vivo experiments verified the potency of AAV variants engineered by ALICE

To validate the potency of the AAV variants designed by the ALICE platform, we employed the elite ascendancy module to screen for sequences with strong versatility for wet-lab validation (Figure 1E, details in the STAR Methods). The top 500 sequences were initially selected on the basis of FE scores, and then the top 100 were chosen on the basis of capsid sequence production fitness. The eight most promising candidates were identified on the basis of both EA scores (see STAR Methods for details) and binding affinity to Ly6c1 and Ly6a targets (Figure S14A, B). Interestingly, we found that the functional characteristics of these eight sequences were distributed closer to those of $\boldsymbol{Ref}^{+}$ and far from those of $\boldsymbol{Ref}^{-}$, indicating the potential for CNS tropism (Figure S14B). The hitherto unreported variants carrying these sequences inserted at the position between VP1 residues 588-589 of the AAV9 capsid[67,68] were packaged and designated AAV.ALICE-N1 through N8 for wet-lab experimental validation (Figure 5A). To visualize the transduced cells, these single-stranded AAV variants were produced carrying a reporter cassette expressing histone H2B-tagged red fluorescent protein (RFP) (H2B-RFP) under the control of the ubiquitous EF1α promoter (AAV-EF1α-H2B-RFP). H2B tagging allows nuclear localization of the expressed RFP. We found that, via the same production protocol, AAV.ALICE-N2, AAV.ALICE-N6, and AAV.ALICE-N8 exhibited increased viral titers, with an approximately 2-fold increase over the original AAV9 (Figure 5B, Table S10, ~5 × 10^{13} viral genomes per ml). In contrast, AAV.ALICE-N3, AAV.ALICE-N4, and AAV.ALICE-N6 displayed normal viral titers (Table S10, ~1 × 10^{13} viral genomes per ml). The AAV.ALICE-N1, AAV.ALICE-N5, which did not yield adequate viral titers, was excluded from subsequent *in vivo* experiments (Table S10; ~9 × 10^{10} viral genomes per ml).

To test the CNS tropism of these variants, we intravenously injected AAV9 or AAV.ALICE-N2, AAV.ALICE-N3, AAV.ALICE-N4, AAV.ALICE-N6, AAV.ALICE-N7, and AAV.ALICE-N8 (5 × 10^{11} vg per animal) was inoculated into BALB/cJ adult mice, which were left for three weeks, after which the brain tissues were examined for AAV transduction by detecting the RFP signal. Only a few RFP-positive cells were observed in the brain sections of the mice treated with AAV9 or AAV.ALICE-N3, AAV.ALICE-N4, AAV.ALICE-N7, or AAV.ALICE-N8 cells (Figure 5C, D; Figure S15A-D). However, a remarkable number of RFP-positive cells were detected in the brains of the mice treated with AAV.ALICE-N2 and AAV. Compared with that of AAV9, the transduction efficiency of ALICE-N6 (Figure 5C, D) was increased by 372-fold and 64-fold, respectively (Figure 5C). To test the multiple-targeting capacity of these variants, we additionally tested AAV.ALICE-N2 and AAV.ALICE-N6 in C57BL/6J and FVB/NJ mouse strains. Among these mouse strains, AAV.ALICE-N2 and AAV.ALICE-N6 also significantly enhanced brain transduction compared with AAV9 (Figure 5C, D).

To further examine the cell types transduced in the CNS, AAV.ALICE-N2-EF1α-H2B-RFP (5 × 10^{11} vg) and AAV.ALICE-N6-EF1α-H2B-RFP (5 × 10^{11} vg) was injected intravenously into adult BALB/J mice, and the tissues were processed for immunostaining 3 weeks after injection. Compared with those of AAV9-labeled cells, higher percentages of RFP-labeled cells were again observed across brain regions in AAV-labeled cells. ALICE-N2 and AAV.ALICE-N6-treated animals (Figure

16A-C). Among all the AAVs. Approximately 30% of ALICE-N2-transduced cells were neurons colabeled with the neuronal marker NeuN (Figure S17A, B); approximately 40% were oligodendrocytes colabeled with the oligodendrocyte marker Olig2 (Figure S17A, B); and approximately 3% were astrocytes colabeled with the astrocyte marker glial fibrillary acidic protein (GFAP) (Figure S18A, Figure S17B). For each cell type, significantly greater percentages of cells were transduced with AAV.ALICE-N2 compared with AAV9 (Figure S17B). For the brain regions sampled, in adult BALB/J mice, AAV.ALICE-N2 transduced 19–51% of total neurons, 26–54% of total oligodendrocytes and 1–7% of total astrocytes.

Compared with AAV9, improved gene delivery of AAV.ALICE-N2 and AAV.ALICE-N6 was also evident in spinal cord cells. Indeed, AAV.ALICE-N2-EF1α-H2B-RFP labeled 8–12% of the spinal cord cells compared with 0–1% for AAV9 (Figure S18B, C) and labeled 9–10% of the ChAT-labeled motor neurons compared with 0–1% for AAV9 (Figure S18B, D), suggesting that AAV.ALICE-N2 efficiently targets spinal cord cells and spinal motor neurons in mice. For the examination of peripheral tissues, systemic administration of AAV.ALICE-N2 and AAV.ALICE-N6 resulted in transduction in the liver, skeletal muscle, heart and spleen of the mice (Figure S19 A, C). Lower transduction levels were observed in the kidney, lung and dorsal root ganglion (DRG) (Figure S19 A-C).

The variants engineered by ALICE perform functions through purposefully designed mechanisms of action

To further investigate the mechanism of action of AAV.ALICE-N2 and AAV. We tested the transduction ability of ALICE-N6 in transducing cells in the CNS through its ability to bind to the purposefully designed targets Ly6c1 and Ly6a via in vitro cell experiments. To do so, we used lentiviruses to stably express the Ly6a, Ly6c1, Car4, and Slco1c1 proteins on the surface of HEK293T cells (Car4 and Slco1c1 were used as controls; details are provided in the STAR Methods). Following an established protocol,[69] we first tested the infectivity of different doses of AAV9 and PHP.eB in HEK293T and Ly6a-transfected HEK293T cells. We selected 1×10^7 v.g. per well and a low multiplicity of infection (MOI) because at this dose, neither AAV9 nor PHP.eB had the ability to infect HEK293T cells. Moreover, the expression of membrane-localized Ly6a in HEK293T cells selectively increased the infectivity of PHP.eB, whereas AAV9 did not, significantly increasing the extent of infection and brightness of the positive cells (Figure S20 A, B). Next, we tested the infectivity of the AAVs. ALICE-N2 and AAV.ALICE-N6 for four receptor proteins (Ly6a, Ly6c1, Car4, and Slco1c1) in triplicate at a dose of $1x10^7$ v.g. per well (Figure S21 A□C). PHP.eB and Ly6a were used as positive controls, and HEK293T cells without protein expression served as blank controls.

We discovered that AAV.ALICE-N2 and AAV.ALICE-N6 displayed notable strong transduction in Ly6a- and Ly6c1-overexpressing cells (Figure S21B, C). In contrast, AAV9 demonstrated no significant transduction in cells overexpressing the target proteins. PHP.eB, employed as a positive control, successfully transduced Ly6a-overexpressing cells but failed to transduce Ly6c1-overexpressing cells (Figure S21B, C), which is consistent with a previous report.[70] Furthermore, the lack of significant infection with AAV.ALICE-N2 and AAV.ALICE-N6 to cells overexpressing Car4 or Slco1c1 target proteins suggests that the sequence evolutionary path of ALICE is not only robust but also well adapted to the Ly6a and Ly6c1 target landscape, implying a controlled evolutionary process (Figure S21B, C).

437

Discussion

ALICE represents the pioneering computational approach for accomplishing direct i*n silico* engineering of AAV capsids with multiple functions, independent of the *in vivo* selection method used. Although AI technology has been effectively utilized in constructing capsid libraries with enhanced fitness or diversity,[71,72] these approaches still depend upon an unpredictable *in vivo* selection process. Despite the use of small datasets ($\boldsymbol{Ref}^{+}$ and $\boldsymbol{Ref}^{-}$) of 129 entries (Table S2), ALICE successfully mapped AAV capsid sequences to desired functions. Even after exclusion of the 12 reported sequences from the $\boldsymbol{Ref}^{+}$ dataset (Table S2), ALICE persistently identified AAV.ALICE-N2 is one of the top candidates for CNS gene delivery (data not shown), suggesting the potential ability of ALICE to engineer AAV capsids, even in the absence of known functional sequences that engage a particular AAV receptor in animal models. However, ALICE is not intended to replace traditional AAV vector design methods. These traditional methods have been refined over nearly two decades, rigorously tested in many scenarios, and offer a wealth of functions yet to be fully exploited. Instead, leveraging data from these experiments, AI-driven AAV capsid design promises to accelerate the progress of AAV gene therapy development in a more time- and cost-effective manner.

ALICE overcomes the challenge of multifunctional fusion conflicts in AAV capsid sequence engineering by adopting a function-guided evolutionary approach. The problem of engineering functions into capsid sequences is NP-hard, given the absence of a known polynomial-time method to map capsid sequences to functions.[73] Although the naïve generative language model, which employs pretraining followed by a semantic-tuning strategy (Figure 1B), successfully generates capsid sequences with enhanced targeting binding affinity, they exhibit remarkably diminished production fitness (Figure S5L). This constraint is presumably rooted in the limitations of limited data availability. The naïve model tended to focus on surface-level data distribution, resulting in global semantic distinctions from natural capsids (Figure S5D, Figure S12C). In doing so, it failed to capture the local semantic information (motifs) crucial for capsid functions (Figure S12D). Perhaps through direct fitting, the naïve model improved only certain aspects, ultimately leading to an imbalance in overall capsid performance. To address this issue, we invented a function-guided evolution strategy to successfully fuse multiple traits, such as target binding and high production fitness, into capsid sequences (Figure 3G). This strategy results in a refined learning trajectory that effectively balances the acquisition of beneficial features with the avoidance of detrimental features, ultimately mapping a capsid sequence to multiple functions imperative for efficient gene delivery. This represents a significant breakthrough in fusing multiple functions into a single sequence, is primed to accelerate the practical implementation of gene delivery through a computational strategy, and provides insights for the design of other functional protein complexes.

Exploring the evolutionary mechanism of capsid sequences within ALICE enables the refinement of this computational strategy toward mapping a capsid sequence to multiple functions. The attributes of interpretability and explainability are crucial for a computational method to advance gene delivery techniques. As evidenced by this study, insights into ALICE's operational mechanism directed the optimization of this approach, enabling a deeper understanding of AAV capsid biology. This capability not only facilitates the mapping of capsid sequences to multiple desired functions but also deepens our mechanistic insights into the complex sequence□function relationships of AAV capsids. By illustrating the multilevel relationship between the capsid sequence and function, ALICE enables a method for the knowledge-driven design of AAVs. Moreover, by illustrating the evolution

mechanism of capsid sequences within ALICE, we demonstrated the possibility of engineering viral capsids in a function-guided manner, offering insights into the interpretability, controllability, and safety of viral capsid protein engineering, a vital aspect of biosafety concerns.[74,75]

Limitations of the study

The variants, AAV.ALICE-N2 and AAV.ALICE-N6 engineered by ALICE, which operates by engaging the Ly6c1 and Ly6a receptors, is anticipated to exhibit limited CNS tropism in other species that do not express these receptors. This limitation is somewhat restricted by the unavailability of data of different AAV targets in other species in the datasets of $\boldsymbol{Lib_{seq}}$, $\boldsymbol{Ref^{+}}$ and $\boldsymbol{Ref^{-}}$. Currently, ALICE does not incorporate structural information as a component of its coevolutionary strategy. The incorporation of structural data via methods such as AlphaFold3[76] might further augment the performance of ALICE. In this study, the design scope of ALICE is confined to insertion peptides of 7-mer amino acids between VP1 residues 588-589 of the AAV9 capsid, limiting its exploration to a 20^7-sequence space. Expanding the training dataset to include other AAV targets, varied sequence lengths, alternative insertion/replacement sites in AAV9 or other serotypes could expand its search space, thereby enhancing ALICE's ability to engineer AAV vectors across a wider landscape and diverse application scenarios. Despite these limitations, ALICE has already demonstrated its ability to design multifunctional AAV vectors with available data, offering valuable insights into gene delivery vector design strategies.

SUPPLEMENTAL INFORMATION

Further information on the research design is available in the Supplementary Information linked to this paper.

ACKNOWLEDGMENTS

The authors acknowledge the excellent technical staff at the Imaging Facility, Core Facility of Zhejiang University School of Medicine and Center of Cryo-Electron Microscopy of Zhejiang University for their assistance with confocal laser scanning microscopy and scanning electron microscopy. This study was supported by the National Natural Science Foundation of China (82371374 to X.W.), the Scientific and Technological Innovation 2030 Program of China - major projects (2021ZD0200408 to X.W.), the grants from Nanhu Brain-computer Interface Institute (No. 010904009 to X.W.), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (No. 2023-PT310-01 to X.W.), the Fundamental Research Funds for the Central Universities (2023ZFJH01-01, 2024ZFJH01-01, K20210195 to X.W.).

AUTHOR CONTRIBUTIONS

H.Z. and B.G. contributed equally to this work. X.W., H.Z. and B.G. conceptualized and designed the study. H.Z., B.G., Y.W., A.M., H.W., X.L., H.L., and H.J. conducted the experiments and collected the data. X.W., H.Z., B.G., Y.W., A.M., H.W., Y.H., X.G. and B.Y. analyzed and interpreted the data. H.Z., B.G., and X.W.N.C. drafted the paper. All the authors critically revised the manuscript and approved the final version for submission.

DECLARATION OF INTERESTS

Zhejiang University has filed patent applications related to this work, with X.W. H. Z, B.G. and H.J. listed as inventors. X.W. is a scientific cofounder of WeQure AI Ltd. All the other authors declare that they have no conflicting interests.

References

1. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Jr., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. Nat Biotechnol *41*, 1099-1106. 10.1038/s41587-022-01618-2.
2. Ruffolo, J.A., and Madani, A. (2024). Designing proteins with language models. Nature Biotechnology *42*, 200-202. 10.1038/s41587-024-02123-4.
3. Kotsiliti, E. (2022). De novo protein design with a language model. Nat Biotechnol *40*, 1433. 10.1038/s41587-022-01518-5.
4. Notin, P., Rollins, N., Gal, Y., Sander, C., and Marks, D. (2024). Machine learning for functional protein design. Nat Biotechnol *42*, 216-228. 10.1038/s41587-024-02127-0.
5. Fu, X., Suo, H., Zhang, J., and Chen, D. (2024). Machine-learning-guided Directed Evolution for AAV Capsid Engineering. Curr Pharm Des *30*, 811-824. 10.2174/0113816128286593240226060318.
6. Chan, C., Harris, K.K., Zolotukhin, S., and Keeler, G.D. (2023). Rational Design of AAV-rh74, AAV3B, and AAV8 with Limited Liver Targeting. Viruses *15*. 10.3390/v15112168.
7. Pan, D., Büning, H., and Ling, C. (2019). Rational Design of Gene Therapy Vectors. Molecular Therapy Methods & Clinical Development *12*, 246-247. 10.1016/j.omtm.2019.01.009.
8. Yao, Y., Wang, J., Liu, Y., Qu, Y., Wang, K., Zhang, Y., Chang, Y., Yang, Z., Wan, J., Liu, J., et al. (2022). Variants of the adeno-associated virus serotype 9 with enhanced penetration of the blood–brain barrier in rodents and primates. Nature biomedical engineering *6*, 1257-1271. 10.1038/s41551-022-00938-7.
9. Lee, E.J., Guenther, C.M., and Suh, J. (2018). Adeno-Associated Virus (AAV) Vectors: Rational Design Strategies for Capsid Engineering. Curr Opin Biomed Eng *7*, 58-63. 10.1016/j.cobme.2018.09.004.
10. Tabebordbar, M., Lagerborg, K.A., Stanton, A., King, E.M., Ye, S., Tellez, L., Krunnfusz, A., Tavakoli, S., Widrick, J.J., Messemer, K.A., et al. (2021). Directed evolution of a family of AAV capsid variants enabling potent muscle-directed gene delivery across species. Cell *184*, 4919-4938.e4922. 10.1016/j.cell.2021.08.028.
11. Deverman, B.E., Pravdo, P.L., Simpson, B.P., Kumar, S.R., Chan, K.Y., Banerjee, A., Wu, W.L., Yang, B., Huber, N., Pasca, S.P., and Gradinaru, V. (2016). Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. Nat Biotechnol *34*, 204-209. 10.1038/nbt.3440.
12. Ravindra Kumar, S., Miles, T.F., Chen, X., Brown, D., Dobreva, T., Huang, Q., Ding, X., Luo, Y., Einarsson, P.H., Greenbaum, A., et al. (2020). Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. Nat Methods *17*, 541-550. 10.1038/s41592-020-0799-7.
13. Lisowski, L., Dane, A.P., Chu, K., Zhang, Y., Cunningham, S.C., Wilson, E.M., Nygaard, S., Grompe, M., Alexander, I.E., and Kay, M.A. (2014). Selection and evaluation of clinically relevant AAV variants in a xenograft liver model. Nature *506*, 382-386. 10.1038/nature12875.
14. Zengel, J., Wang, Y.X., Seo, J.W., Ning, K., Hamilton, J.N., Wu, B., Raie, M., Holbrook, C., Su, S., Clements, D.R., et al. (2023). Hardwiring tissue-specific AAV transduction in mice through engineered receptor expression. Nat Methods *20*, 1070-1081. 10.1038/s41592-023-01896-x.

15. Lin, R., Zhou, Y., Yan, T., Wang, R., Li, H., Wu, Z., Zhang, X., Zhou, X., Zhao, F., Zhang, L., et al. (2022). Directed evolution of adeno-associated virus for efficient gene delivery to microglia. Nat Methods *19*, 976-985. 10.1038/s41592-022-01547-7.
16. Li, C., and Samulski, R.J. (2020). Engineering adeno-associated virus vectors for gene therapy. Nature Reviews Genetics *21*, 255-272. 10.1038/s41576-019-0205-4.
17. Wang, D., Tai, P.W.L., and Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. Nat Rev Drug Discov *18*, 358-378. 10.1038/s41573-019-0012-9.
18. Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat Methods *16*, 687-694. 10.1038/s41592-019-0496-6.
19. Norn, C., Wicky, B.I.M., Juergens, D., Liu, S., Kim, D., Tischer, D., Koepnick, B., Anishchenko, I., Foldit, P., Baker, D., and Ovchinnikov, S. (2021). Protein sequence design by conformational landscape optimization. Proc Natl Acad Sci U S A *118*. 10.1073/pnas.2017228118.
20. Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. (2020). An evolution-based model for designing chorismate mutase enzymes. Science *369*, 440-445. 10.1126/science.aba3304.
21. Yang, A., Jude, K.M., Lai, B., Minot, M., Kocyla, A.M., Glassman, C.R., Nishimiya, D., Kim, Y.S., Reddy, S.T., Khan, A.A., and Garcia, K.C. (2023). Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. Science *381*, eadh1720. 10.1126/science.adh1720.
22. Goudy, O.J., Nallathambi, A., Kinjo, T., Randolph, N.Z., and Kuhlman, B. (2023). In silico evolution of autoinhibitory domains for a PD-L1 antagonist using deep learning models. Proc Natl Acad Sci U S A *120*, e2307371120. 10.1073/pnas.2307371120.
23. Raven, S.A., Payne, B., Bruce, M., Filipovska, A., and Rackham, O. (2022). In silico evolution of nucleic acid-binding proteins from a nonfunctional scaffold. Nat Chem Biol *18*, 403-411. 10.1038/s41589-022-00967-y.
24. Huang, Q., Chan, K.Y., Wu, J., Botticello-Romero, N.R., Zheng, Q., Lou, S., Keyes, C., Svanbergsson, A., Johnston, J., Mills, A., et al. (2024). An AAV capsid reprogrammed to bind human transferrin receptor mediates brain-wide gene delivery. Science, eadm8386. 10.1126/science.adm8386.
25. Huang, Q., Chen, A.T., Chan, K.Y., Sorensen, H., Barry, A.J., Azari, B., Zheng, Q., Beddow, T., Zhao, B., Tobey, I.G., et al. (2023). Targeting AAV vectors to the central nervous system by engineering capsid-receptor interactions that enable crossing of the blood-brain barrier. PLoS Biol *21*, e3002112. 10.1371/journal.pbio.3002112.
26. Shay, T.F., Sullivan, E.E., Ding, X., Chen, X., Ravindra Kumar, S., Goertsen, D., Brown, D., Crosby, A., Vielmetter, J., Borsos, M., et al. (2023). Primate-conserved carbonic anhydrase IV and murine-restricted LY6C1 enable blood-brain barrier crossing by engineered viral vectors. Sci Adv *9*, eadg6618. 10.1126/sciadv.adg6618.
27. Pillay, S., Meyer, N.L., Puschnik, A.S., Davulcu, O., Diep, J., Ishikawa, Y., Jae, L.T., Wosen, J.E., Nagamine, C.M., Chapman, M.S., and Carette, J.E. (2016). An essential receptor for adeno-associated virus infection. Nature *530*, 108-112. 10.1038/nature16465.
28. Fatma-Elzahraa Eid1, Albert T. Chen1, Ken Y. Chan1, Qin Huang1, Qingxia Zheng1, Isabelle G. Tobey1,, Simon Pacouret1, P.P.B., Casey Keyes1, Megan Powell1, Jencilin Johnston1, Binhui Zhao1, Kasper, and Lage1, 4,5, Alice F. Tarantal6, Yujia A. Chan1, Benjamin E.

Deverman1* (2023). Systematic multi-trait AAV capsid engineering for efficient gene delivery. bioRxiv.

29. Perabo, L., Endell, J., King, S., Lux, K., Goldnau, D., Hallek, M., and Büning, H. (2006). Combinatorial engineering of a gene therapy vector: directed evolution of adeno-associated virus. The Journal of Gene Medicine *8*, 155-162. https://doi.org/10.1002/jgm.849.

30. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., and Ellinor, P.T. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616-624. 10.1038/s41586-023-06139-9.

31. Coello, C.A. (2000). An updated survey of GA-based multiobjective optimization techniques. ACM Comput. Surv. *32*, 109–143. 10.1145/358923.358929.

32. Wicky, B.I.M., Milles, L.F., Courbet, A., Ragotte, R.J., Dauparas, J., Kinfu, E., Tipps, S., Kibler, R.D., Baek, M., DiMaio, F., et al. (2022). Hallucinating symmetric protein assemblies. Science *378*, 56-61. 10.1126/science.add1964.

33. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. Advances in neural information processing systems *33*, 18661-18673.

34. Lutz, I.D., Wang, S., Norn, C., Courbet, A., Borst, A.J., Zhao, Y.T., Dosey, A., Cao, L., Xu, J., Leaf, E.M., et al. (2023). Top-down design of protein architectures with reinforcement learning. Science *380*, 266-273. 10.1126/science.adf6591.

35. Huang, Q., Chan, K.Y., Tobey, I.G., Chan, Y.A., Poterba, T., Boutros, C.L., Balazs, A.B., Daneman, R., Bloom, J.M., Seed, C., and Deverman, B.E. (2019). Delivering genes across the blood-brain barrier: LY6A, a novel cellular receptor for AAV-PHP.B capsids. PLoS One *14*, e0225206. 10.1371/journal.pone.0225206.

36. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology *41*, 1099-1106. 10.1038/s41587-022-01618-2.

37. The UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research *49*, D480-D489. 10.1093/nar/gkaa1100.

38. The UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research *47*, D506-D515. 10.1093/nar/gky1049.

39. Yao, Y., Wang, J., Liu, Y., Qu, Y., Wang, K., Zhang, Y., Chang, Y., Yang, Z., Wan, J., Liu, J., et al. (2022). Variants of the adeno-associated virus serotype 9 with enhanced penetration of the blood-brain barrier in rodents and primates. Nature biomedical engineering *6*, 1257-1271. 10.1038/s41551-022-00938-7.

40. Chan, K.Y., Jang, M.J., Yoo, B.B., Greenbaum, A., Ravi, N., Wu, W.L., Sánchez-Guardado, L., Lois, C., Mazmanian, S.K., Deverman, B.E., and Gradinaru, V. (2017). Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. Nature neuroscience *20*, 1172-1179. 10.1038/nn.4593.

41. Noble, W.S. (2006). What is a support vector machine? Nature Biotechnology *24*, 1565-1567. 10.1038/nbt1206-1565.

42. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery.

43. Yao, Z., and Ruzzo, W.L. (2006). A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics *7*, S11. 10.1186/1471-2105-7-S1-S11.
44. Stephan, J., Stegle, O., and Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. Nature Communications *6*, 7432. 10.1038/ncomms8432.
45. Asnicar, F., Thomas, A.M., Passerini, A., Waldron, L., and Segata, N. (2024). Machine learning for microbiologists. Nature Reviews Microbiology *22*, 191-205. 10.1038/s41579-023-00984-1.
46. Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. Computational Biology and Chemistry *85*, 107200. https://doi.org/10.1016/j.compbiolchem.2020.107200.
47. Huang, J., Xu, Y., Xue, Y., Huang, Y., Li, X., Chen, X., Xu, Y., Zhang, D., Zhang, P., Zhao, J., and Ji, J. (2023). Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. Nature Biomedical Engineering *7*, 797-810. 10.1038/s41551-022-00991-2.
48. Saqib, M. (2021). Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. Applied Intelligence *51*, 2703-2713. 10.1007/s10489-020-01942-7.
49. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.
50. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv *abs/1907.11692*.
51. Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. Proceedings of the AAAI Conference on Artificial Intelligence *31*. 10.1609/aaai.v31i1.10804.
52. Han, X., Zhang, L., Zhou, K., and Wang, X. (2019). ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework. Computers & Chemical Engineering *131*, 106533. https://doi.org/10.1016/j.compchemeng.2019.106533.
53. Hie, B., Zhong, E.D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. Science *371*, 284-288. 10.1126/science.abd7331.
54. Johansson, J. (2008). EVOLUTIONARY RESPONSES TO ENVIRONMENTAL CHANGES: HOW DOES COMPETITION AFFECT ADAPTATION? Evolution *62*, 421-435. 10.1111/j.1558-5646.2007.00301.x.
55. Pearson, W.R. (1991). Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics *11*, 635-650. https://doi.org/10.1016/0888-7543(91)90071-L.
56. Pang, S., Li, W., He, H., Shan, Z., and Wang, X. (2019). An EDA-GA Hybrid Algorithm for Multi-Objective Task Scheduling in Cloud Computing. IEEE Access *7*, 146379-146389. 10.1109/ACCESS.2019.2946216.

57. Mohapatra, A., Mishra, P., and Padhy, S. (2009). Modeling Biological Signals using Information-Entropy with Kullback-Leibler-Divergence.
58. Trivedi, R., and Nagarajaram, H.A. (2020). Substitution scoring matrices for proteins - An overview. Protein Science *29*, 2150-2163. https://doi.org/10.1002/pro.3954.
59. Ganapathiraju, M., Weisser, D., Rosenfeld, R., Carbonell, J.G., Reddy, R., and Klein-Seetharaman, J. (2002). Comparative ngram analysis of whole-genome sequences. Biology, Computer Science.
60. Bonizzoni, P., Vedova, G.D., and Mauri, G. (2001). Experimenting an approximation algorithm for the LCS. Discrete Applied Mathematics *110*, 13-24. https://doi.org/10.1016/S0166-218X(00)00300-0.
61. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nature Methods *16*, 1315-1322. 10.1038/s41592-019-0598-1.
62. Marks, D.S., Hopf, T.A., and Sander, C. (2012). Protein structure prediction from sequence variation. Nature Biotechnology *30*, 1072-1080. 10.1038/nbt.2419.
63. Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. Science *355*, 294-298. 10.1126/science.aah4043.
64. Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nature Methods *16*, 687-694. 10.1038/s41592-019-0496-6.
65. Bork, P., and Koonin, E.V. (1996). Protein sequence motifs. Current Opinion in Structural Biology *6*, 366-376. https://doi.org/10.1016/S0959-440X(96)80057-1.
66. Sobhy, H. (2016). A Review of Functional Motifs Utilized by Viruses. Proteomes *4*. 10.3390/proteomes4010003.
67. Deverman, B.E., Pravdo, P.L., Simpson, B.P., Kumar, S.R., Chan, K.Y., Banerjee, A., Wu, W.L., Yang, B., Huber, N., Pasca, S.P., and Gradinaru, V. (2016). Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. Nature Biotechnology *34*, 204-+. 10.1038/nbt.3440.
68. Ravindra Kumar, S., Miles, T.F., Chen, X., Brown, D., Dobreva, T., Huang, Q., Ding, X., Luo, Y., Einarsson, P.H., Greenbaum, A., et al. (2020). Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. Nature Methods *17*, 541-550. 10.1038/s41592-020-0799-7.
69. LaGory, Edward L., Wu, C., Taniguchi, Cullen M., Ding, C.-Kuang C., Chi, J.-T., von Eyben, R., Scott, David A., Richardson, Adam D., and Giaccia, Amato J. (2015). Suppression of PGC-1α Is Critical for Reprogramming Oxidative Metabolism in Renal Cell Carcinoma. Cell Reports *12*, 116-127. https://doi.org/10.1016/j.celrep.2015.06.006.
70. Shay, T.F., Sullivan, E.E., Ding, X., Chen, X., Ravindra Kumar, S., Goertsen, D., Brown, D., Crosby, A., Vielmetter, J., Borsos, M., et al. Primate-conserved carbonic anhydrase IV and murine-restricted LY6C1 enable blood-brain barrier crossing by engineered viral vectors. Science Advances *9*, eadg6618. 10.1126/sciadv.adg6618.
71. 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. Nature Biotechnology *39*, 691-696. 10.1038/s41587-020-00793-4.

72. Zhu, D., Brookes, D.H., Busia, A., Carneiro, A., Fannjiang, C., Popova, G., Shin, D., Donohue, K.C., Lin, L.F., Miller, Z.M., et al. (2024). Optimal trade-off control in machine learning-based library design, with application to adeno-associated virus (AAV) for gene therapy. Sci Adv *10*, eadj3786. 10.1126/sciadv.adj3786.
73. Bazille, H. (2014). Protein design: a NP-hard problem in bioinformatics. Biology, Computer Science, 42.
74. Chhabra, A., Bashirians, G., Petropoulos, C.J., Wrin, T., Paliwal, Y., Henstock, P.V., Somanathan, S., da Fonseca Pereira, C., Winburn, I., and Rasko, J.E.J. (2024). Global seroprevalence of neutralizing antibodies against adeno-associated virus serotypes used for human gene therapies. Molecular Therapy Methods & Clinical Development *32*. 10.1016/j.omtm.2024.101273.
75. Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence *4*, 189-191. 10.1038/s42256-022-00465-9.
76. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493-500. 10.1038/s41586-024-07487-w.

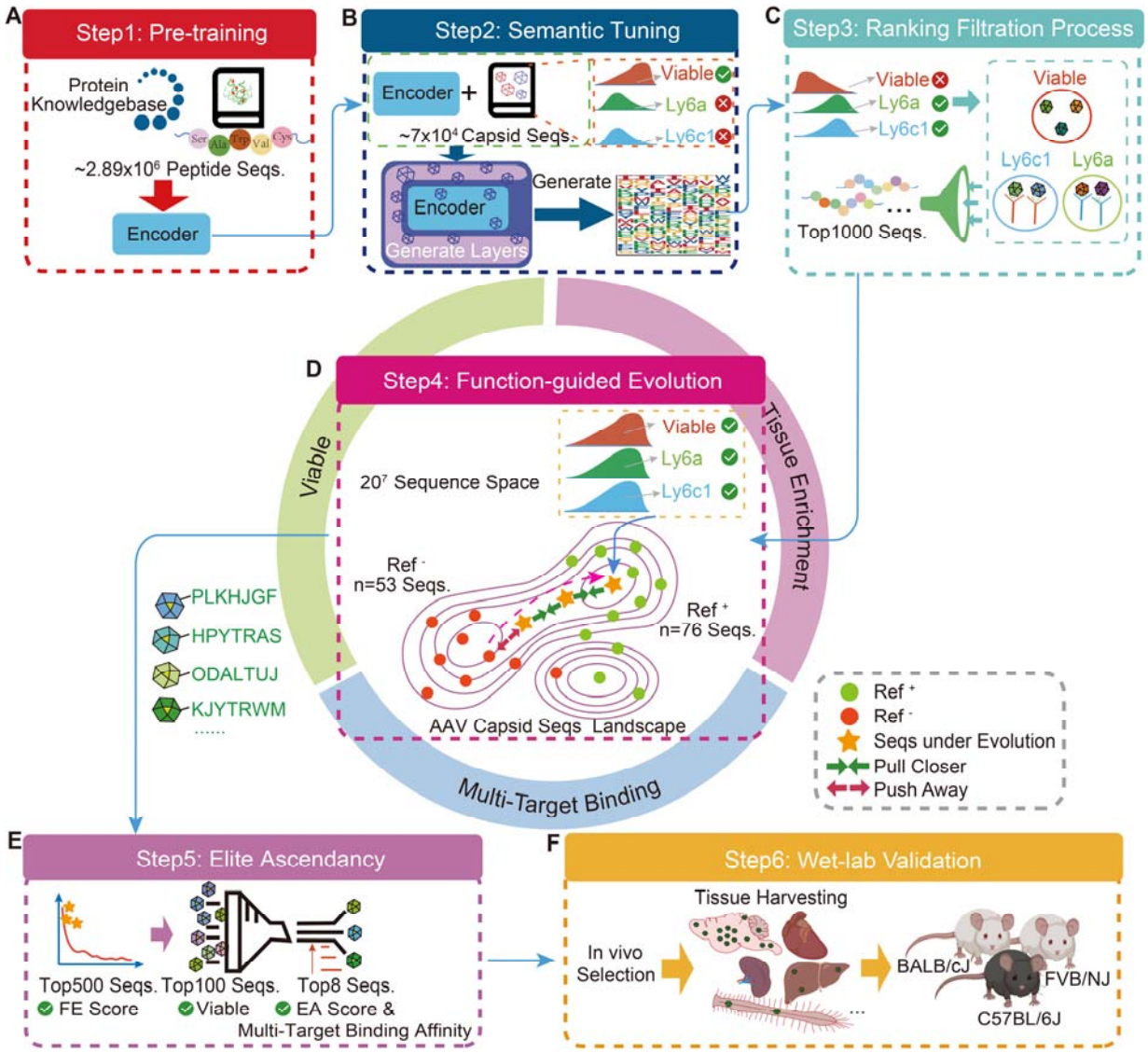


Figure 1. Workflow of the ALICE platform in mapping capsid sequences to functions. The workflow of the ALICE platform integrates 6 steps, including pretraining, semantic tuning, the ranking filtration process, function-guided evolution, elite ascendancy, and wet-lab validation. **(A)** A language model learns the foundational knowledge of peptide sequences in the pretraining step. **(B)** The language model continues learning the knowledge of AAV capsid sequences with high viability but low Ly6a or Ly6c1 binding affinity in the semantic tuning step. Then, a generative model designs AAV capsid sequences with enhanced semantics and various functions. **(C)** The Ranking Filtration Process (RankFiltPro) module selects the top 1000 capsid sequences, prioritizing those with high production fitness coupled with a strong binding affinity for both Ly6a and Ly6c1 targets.

(D) The function-guided evolution (FE) module constructs a conformational AAV capsid sequence evolution landscape in which capsid sequences evolve and map to multiple functions of viability, multiple targeting, and tissue enrichment (crossing the BBB) from a large potential sequence space (20^7).

(E) In the elite ascendancy (EA) module, the top 500 sequences were first selected, prioritizing the FE score. The top 100 sequences were then selected, prioritizing viability. Finally, the top eight candidate capsids were selected, prioritizing the EA score and the target binding affinity of both Ly6a and Ly6c1.

(F) In the wet-lab validation step, the potency of the AAV variants engineered by ALICE was evaluated through *in vivo* experiments using three distinct mouse strains: BLAB/cJ, C57B/6J and FVB/NJ.

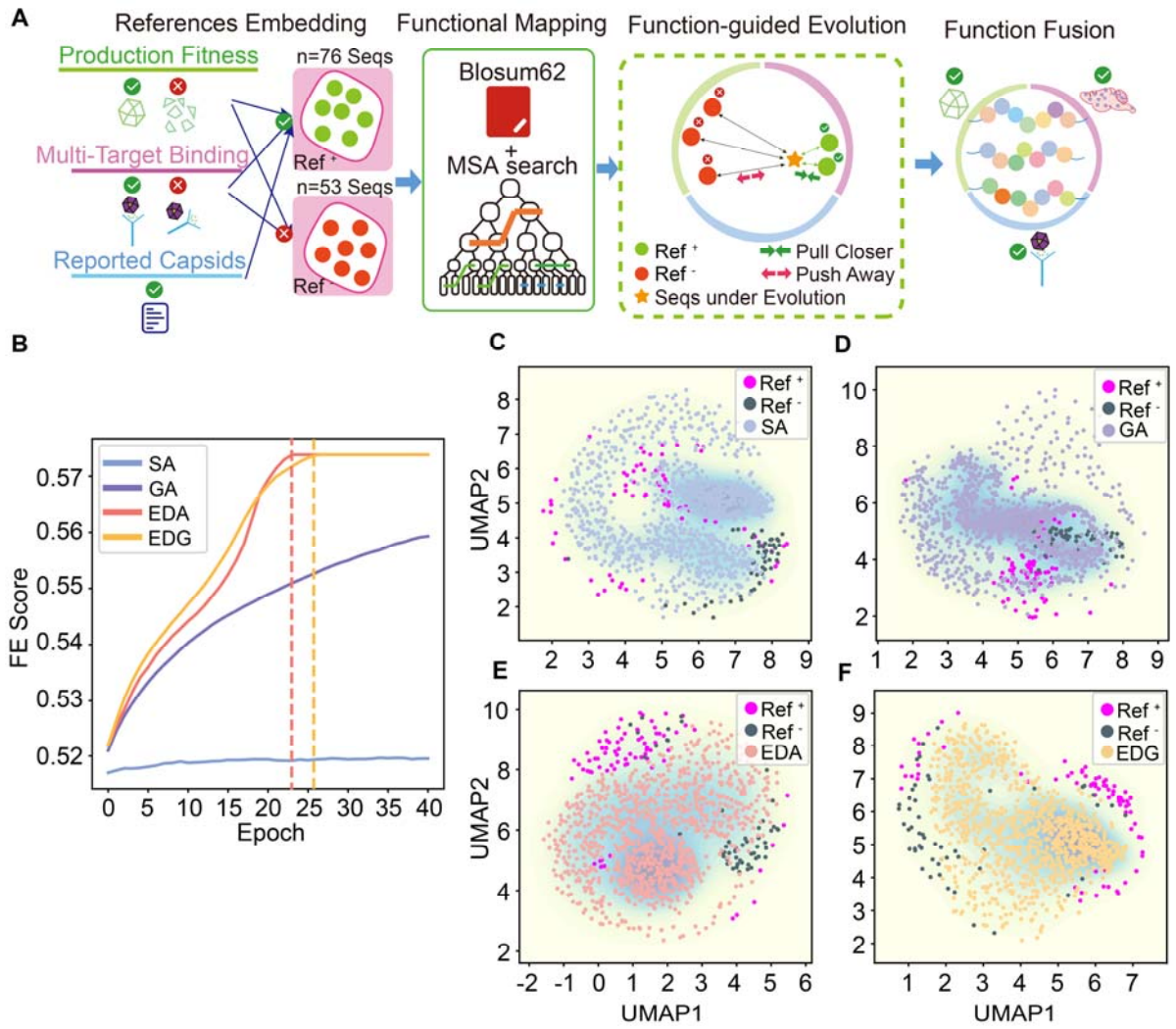
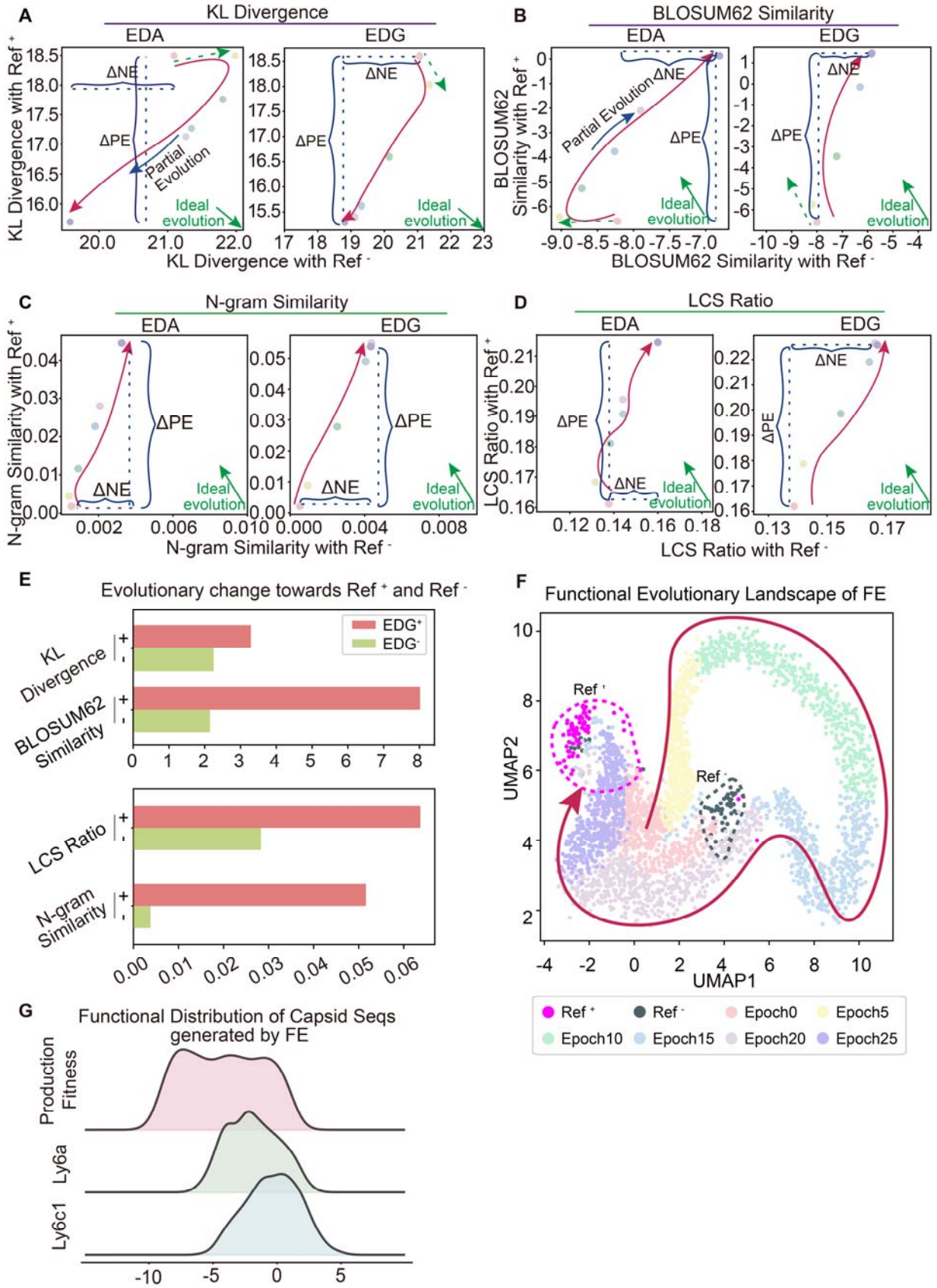


Figure 2. Operating mechanism and functionality of the function-guided evolution (FE) module.

(A) Positive and negative samples (Ref^+ and Ref^-) for the FE module were chosen on the basis of viability, multitargeting, and reported capsid (76 Ref^+ sequences, 53 Ref^- sequences). The sequences filtered from RankFiltPro evolve toward the features of positive samples and away from those of negative samples in a model employing a heuristic algorithm coupled with contrastive learning (Multiple Sequence Alignment, MSA) and reinforcement learning. This enables the fusing of multiple functions into a single sequence.

(B) The plot illustrates the FE scores of sequences engineered by the SA, GA, EDA, or EDG model at different epochs on their progress toward the Ref^+/Ref^- sample.

(C-F) The plots show the functional distributions of sequences that have been engineered via SA (C), GA (D), EDA (E), and EDG (F) compared with those of the sequences from the Ref^+ and Ref^- datasets within a multidimensional space.


A
KL Divergence
EDA
EDG
KL Divergence with Ref +
KL Divergence with Ref -
ΔNE
ΔPE
Partial Evolution
Ideal evolution
B
BLOSUM62 Similarity
BLOSUM62 Similarity with Ref +
BLOSUM62 Similarity with Ref -
C
N-gram Similarity
N-gram Similarity with Ref +
N-gram Similarity with Ref -
D
LCS Ratio
LCS Ratio with Ref +
LCS Ratio with Ref -
E
Evolutionary change towards Ref + and Ref -
EDG+
EDG-
KL Divergence
BLOSUM62 Similarity
LCS Ratio
N-gram Similarity
F
Functional Evolutionary Landscape of FE
Ref +
Ref -
UMAP2
UMAP1
Epoch0
Epoch5
Epoch10
Epoch15
Epoch20
Epoch25
G
Functional Distribution of Capsid Seqs generated by FE
Production Fitness
Ly6a
Ly6c1

800

Figure 3. The FE module steers the evolutionary trajectory of sequences toward positive samples with desired functions.

(A-D) The plots illustrate the evolutionary trajectories measured by KL divergence (A), BLOSUM62 similarity (B), N-gram similarity (C), and the LCS ratio (D) reveal the evolving functionality and semantic characteristics. The green arrow indicates the ideal evolution direction, whereas the green dashed arrow shows the direction of evolution within the initial 5 epochs.

(E) Evolutionary changes in the sequences toward $\boldsymbol{Ref}^{+}$ or $\boldsymbol{Ref}^{-}$ engineered via the EDG model.

(F) Conformational evolution landscape of capsid sequences engineered with EDG-based FE at different epochs, showing progression toward $\boldsymbol{Ref}^{+}$ and divergence from $\boldsymbol{Ref}^{-}$. The colors of the dots represent the outcomes after different epochs, as indicated by the color bar provided at the bottom.

(G) The plot shows the functional distribution of FE-designed capsid sequences, including production fitness and Ly6a or Ly6c1 binding capacity**.**

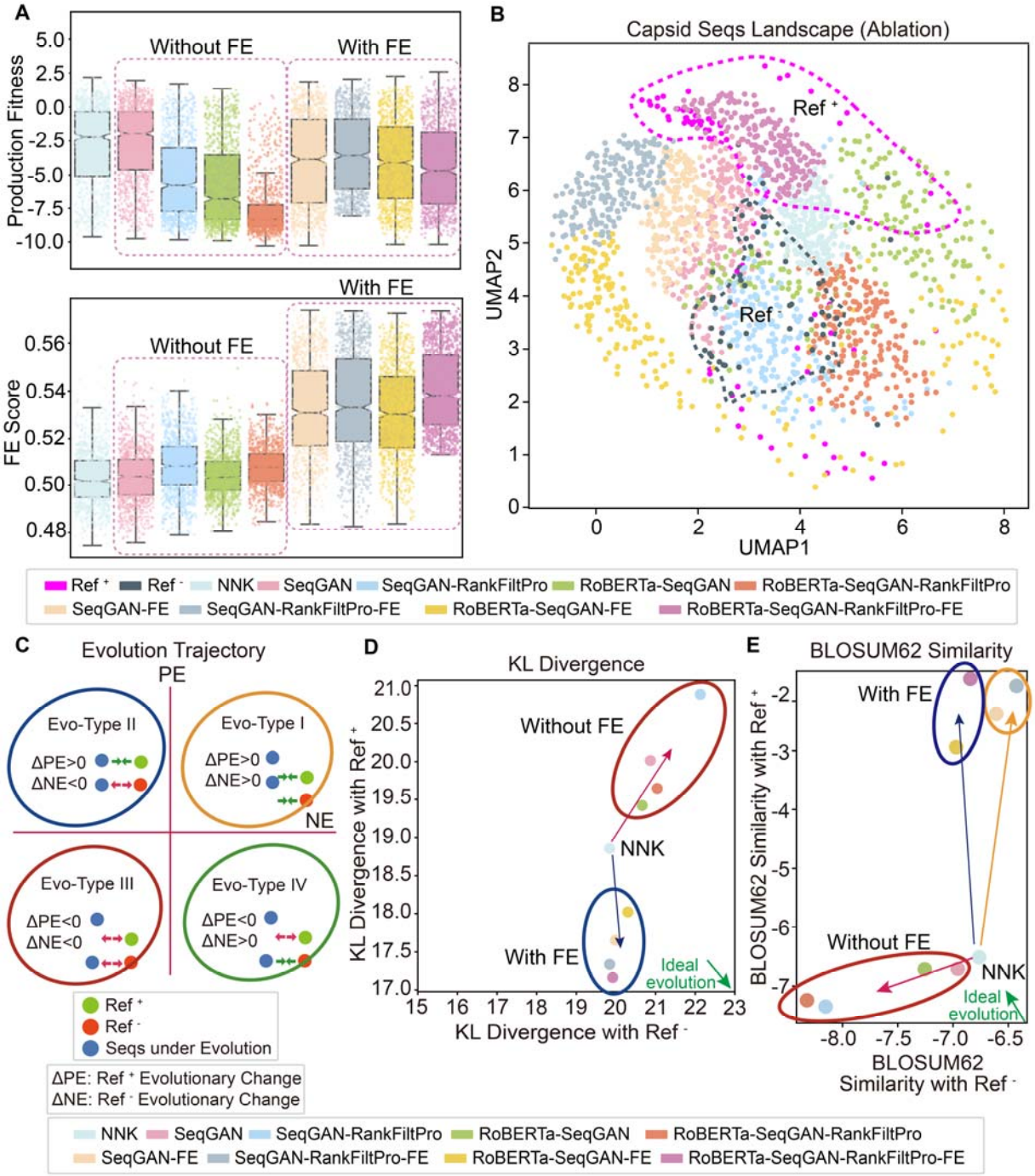


Figure 4. The ablation of each module significantly affects the evolutionary trajectory of sequences.

(A) Performance of production fitness (top) and the FE score (bottom) across architectures with different ablations. The colors of the dots and histograms represent the outcomes from architectures with distinct ablations, as indicated by the color bar provided at the bottom.

(B) The diagram illustrates the functional distribution of the capsid sequence landscape engineered by architectures with different ablations. The colors of the dots represent the outcomes from architectures with distinct ablations, as indicated by the color bar provided at the bottom.

(C) Definitions of the four evolutionary trajectory types.

(D-E) The narrative depicts the evolutionary changes in sequences engineered by architectures with different ablations, as measured by KL divergence (D) and BLOSUM62 similarity (E), in comparison to the sequences in the control NNK dataset. The colors of the dots represent the outcomes from architectures with distinct ablations, as indicated by the color bar provided at the bottom. The green arrow signifies the ideal evolution direction. The arrows colored accordingly show the direction of evolution of sequences engineered by architectures with different ablations.

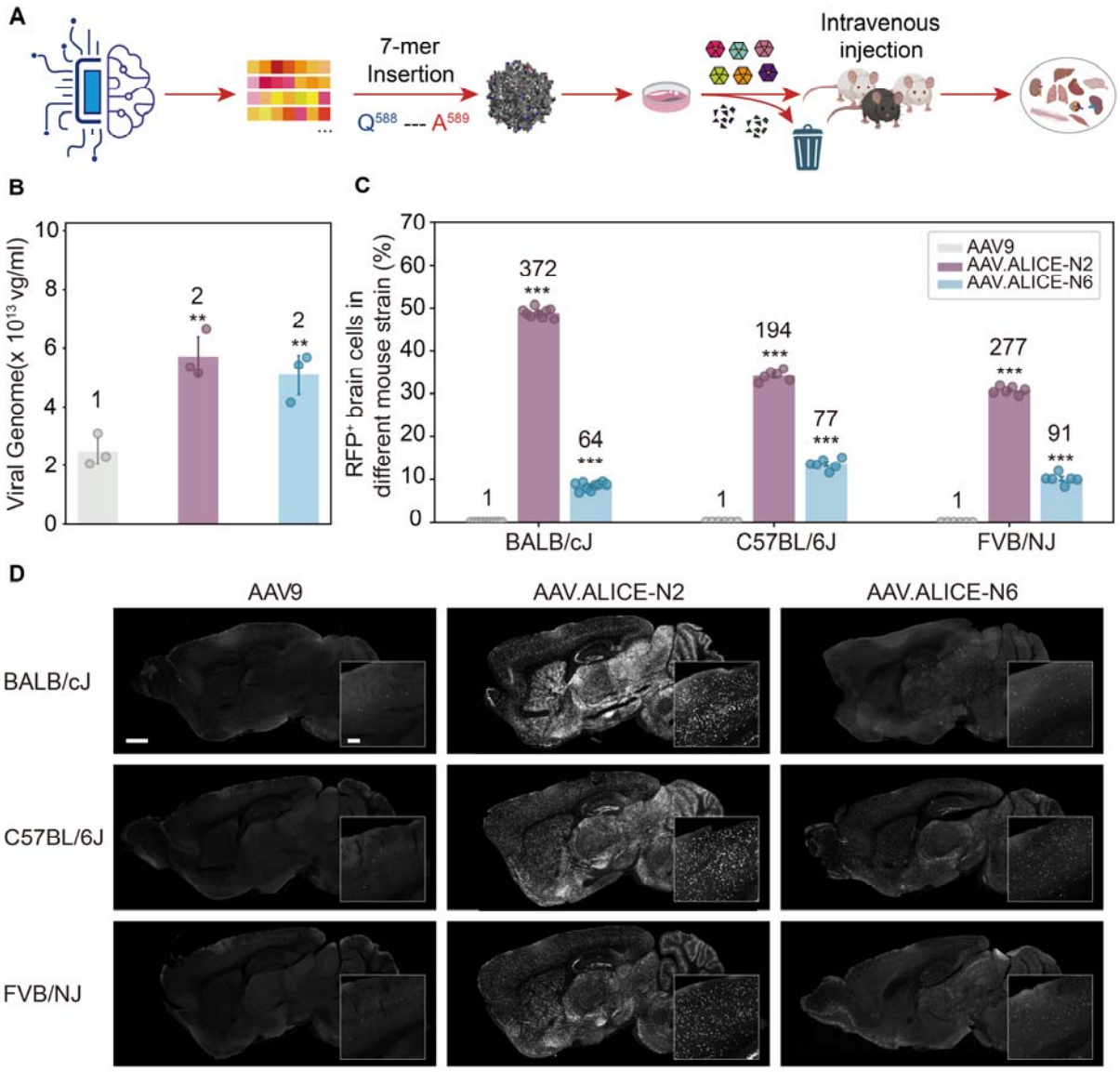


Figure 5. An *in vivo* study verified the ability of AAV variants engineered by ALICE to transduce CNS cells from different mouse strains.

(A) Workflow diagram of the wet laboratory validation process of the ALICE platform.

(B) Comparison of the titers of AAV9 and AAV.ALICE-N2 and AAV.ALICE-N6. The data are presented as the means ± s.e.m., n = 3 biologically independent replicates. The color of the bar indicates the results of different AAVs, as indicated by the color bar provided in (c).

(C) Quantification of the percentage of total RFP^+ brain cells and the fold change relative to AAV9 in BALB/cJ, C57BL/6J, and FVB/NJ mice after intravenous injections of AAV variants (5×10^{11} vg per animal). The data are presented as the means ± s.e.m; n = 3–4 mice per group. Statistical significance was determined via one-way ANOVA with Tukey's post hoc test (*P < 0.05, **P < 0.01, ***P < 0.001 versus AAV9). The color of the bar indicates the outcome of treatment with different AAVs, as indicated by the provided color bar.

(D) Representative images of brain regions in BALB/cJ, C57BL/6J, and FVB/NJ mice showing RFP-labeled cells (white dots) 21 days after intravenous administration of AAVs-EF1α-H2B-RFP (AAV9 on the left, AAV.ALICE-N2 in the middle, AAV.ALICE-N6 on the right; 5×10^{11} vg per animal). Scale bars: 1,000 μm for low-magnification images and 200 μm for enlarged images.