



Machine learning and maximum entropy for rational AAV library-design: technical commentary and analysis

Hans Zempel^{1,2}

Keywords:

Machine learning, entropy optimization, library-design, gene therapy, peptide insertion, recombinant AAV

Citation:

Zempel H. Machine learning and maximum entropy for rational AAV library-design: technical commentary and analysis. *J Transl Genet Genom.* 2026;10:14-20. <https://dx.doi.org/10.20517/jtgg.2025.100>

Received: 27 Aug 2025

First Decision: 16 Dec 2025

Revised: 30 Dec 2025

Accepted: 20 Jan 2026

Published: 26 Jan 2026

Academic Editor:

Zhu Chen

Copy Editor:

Ping Zhang

Production Editor:

Ping Zhang

Abstract

Adeno-associated virus (AAV) engineering is critical for improving and adapting these vectors for gene therapy. In this technical commentary, we evaluate the article by Zhu *et al.* on an information-theoretic, machine learning-guided approach to optimizing peptide insertion libraries for AAV engineering. To address inefficiencies in standard NNK libraries (where N represents any nucleotide and K represents G or T), Zhu *et al.* introduce a maximum entropy optimization framework that enables precise control over the trade-off between sequence diversity and functional packaging fitness. Through deep sequencing, regression modeling of log enrichment scores, and entropy-constrained library design, the authors generate synthetic AAV libraries exhibiting a fivefold increase in packaging efficiency with maintained or enhanced effective diversity. Adequate validation, including infection studies in primary human brain tissue, confirms the superior translational relevance and biological potential of their approach. This commentary outlines the methodology's mathematical and experimental rigor, emphasizes its immediate impact on gene therapy, and discusses its extensibility to broader library engineering problems. The generalizable nature of the technique opens avenues for more efficient, high-throughput synthetic and recombinant biology applications, and may be applied to modulating AAV properties, further expanding possibilities for AAV use in both basic research and translational settings.

INTRODUCTION

The article “Optimal trade-off control in machine learning - based library-design, with application to adeno-associated virus (AAV) for gene therapy” by Zhu *et al.*^[1] represents a major advance in the rational, data-driven engineering of viral vector libraries for therapeutic gene delivery. This work addresses a longstanding bottleneck in AAV-directed evolution: the inherent inefficiency and wasted diversity in standard combinatorial peptide-insertion libraries, particularly the widely used “NNK” randomization approach (using NNK codons, where N = any nucleotide and K = G or T). This method maximizes amino acid diversity while minimizing stop



¹Institute of Human Genetics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne 50931, Germany.

²Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne 50931, Germany.

Correspondence to: Dr. Hans Zempel, Institute of Human Genetics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne 50931, Germany. E-mail: hans.zempel@uk-koeln.de

codons^[2], yet a majority of variants remain nonfunctional for key requirements such as viral packaging. Through an integrated workflow combining high-throughput sequencing, supervised machine learning (ML), and an entropy-constrained optimization framework, Zhu *et al.*^[1] systematically design AAV insertion libraries that achieve markedly higher packaging fitness with minimal sacrifice of sequence diversity, outperforming established baseline libraries on every metric relevant to downstream gene therapy applications. This commentary focuses on three complementary aspects: (i) positioning the framework proposed by Zhu *et al.*^[1] within the broader landscape of AAV-directed evolution and ML-guided capsid engineering; (ii) analyzing the information-theoretic and Chemistry, Manufacturing, and Controls (CMC)-relevant implications of explicitly controlling the diversity - fitness trade-off in library design; and (iii) outlining concrete avenues for extending this approach to multi-objective optimization, serotype-agnostic campaigns, and central nervous system (CNS)-targeted gene therapy programs.

Directed evolution, AAV library bottlenecks, and the need for optimization

AAVs are central to the field of gene therapy, but their utility is limited by suboptimal targeting, packaging, and immune evasion profiles, among other factors^[3,4]. Directed evolution of the capsid gene - using high-diversity insertion or mutagenesis libraries - has enabled iterative improvements of these properties^[5,6], with ML-guided directed evolution currently setting the benchmark (see below). Unfortunately, the efficiency of these selections is fundamentally limited by the quality of the starting library. Standard libraries, particularly NNK-based peptide insertion libraries, are highly diverse at the nucleotide and amino acid sequence level, but a substantial fraction of possible variants fail to package into viable capsids. Consequently, large-scale screening procedures are inefficient, and the probability of success in any given campaign is reduced, because a significant portion of the input cannot be utilized, wasting effort at every step of downstream experiments and analysis.

Prior approaches to library design often relied on post hoc analysis, offering limited ability to prospectively control the trade-off between diversity and effective functional fitness (i.e., model outputs that approximate packaging-related performance), or to integrate information from modern high-throughput next-generation sequencing (NGS)-based fitness assays (for review, see e.g.,^[7,8]). There is a growing recognition of the need for a quantitative, generalizable method to prospectively engineer libraries that maximize expected payoff in downstream selections - balancing fitness/performance and diversity.

METHODOLOGICAL FRAMEWORK: ML-GUIDED, ENTROPY-CONSTRAINED LIBRARY-DESIGN

The workflow reported by Zhu *et al.*^[1] comprises several innovative steps, each crucial for the realization of optimal trade-off libraries:

- **Generation and deep sequencing of an NNK 7-mer insertion baseline library:** The authors constructed a combinatorial AAV5 peptide insertion library with random 7-mer insertions, following the NNK codon scheme to maximize amino acid diversity while limiting stop codon prevalence. Deep sequencing was performed both pre- and post-packaging (i.e., before and after viral assembly selection) to establish counts of each sequence and compute enrichment, a proxy for packaging fitness.
- **Calculation of log enrichment scores:** By comparing sequence abundances pre- and post-packaging, the authors estimated the log enrichment for each variant, which quantifies the change in representation and serves as a direct surrogate for packaging capability.
- **Model training:** The authors trained both linear and neural network regression models (NN, 100) with the log enrichment score as the target and log enrichment as a proxy for packaging-related fitness. Weights for

each sequence were assigned based on sequencing depth to robustly capture statistical confidence. This represents a critical methodological improvement in high-noise, high-dimensional data regimes.

- Entropy- and fitness-constrained optimization: The core of the design strategy is a formal maximum entropy approach - a principled statistical physics framework - where the library is represented as a sequence probability distribution and optimization seeks the maximum entropy distribution under a constraint on expected predicted fitness (mean enrichment), with fitness referring to model outputs/packaging-related performance. The solution corresponds to a Boltzmann distribution where a parameter λ trades off diversity against packaging fitness. For practical synthesis constraints (e.g., degenerate oligo libraries), the optimization occurs over position-wise nucleotide probabilities rather than explicit sequence composition, making the method synthetically tractable.

From a CMC and manufacturability standpoint, it is important to recognize that the position-wise nucleotide distributions obtained *in silico* represent idealized targets, whereas actual synthesized libraries are influenced by biases in degenerate oligonucleotide synthesis, lot-to-lot variability, and context-dependent error rates in longer randomized regions. These deviations can lead to systematic under- or overrepresentation of specific codons or motifs, so that realized libraries only approximate the theoretical maximum-entropy design. Nevertheless, encoding library designs at the level of positional nucleotide frequencies, as done by Zhu *et al.*^[1], provides a direct and experimentally tractable interface for iterative calibration between design and synthesis, allowing CMC teams to quantify and progressively reduce these discrepancies across production batches.

Key results and performance metrics

The ML-guided library, particularly exemplar library D2, exhibits a fivefold higher packaging-fitness compared to the NNK baseline, as measured both in bulk (overall viral genome titer) and at the individual variant level (distribution of log enrichment scores). In this framework, log enrichment is used as a proxy for packaging competence at the library stage, reflecting how well a given variant survives the packaging bottleneck. This metric captures only one specific dimension of AAV critical quality attributes (CQAs): it relates to capsid assembly and genome encapsidation, but does not report on full/empty capsid ratios, VP1/VP2/VP3 stoichiometry, aggregation, higher-order structural integrity, infectivity per genome, or immunogenicity. Enrichment-derived “fitness” should therefore be understood as an upstream, packaging-focused indicator that must be complemented by orthogonal analytical assays in later development stages. Critically, this gain in packaging fitness was achieved with negligible or even improved diversity, as measured by entropy and the “effective number” of library variants (i.e., accounting for both unique variants and evenness of their distribution). Whereas one round of packaging selection on the NNK library results in substantial losses of diversity, the ML-designed library preserves a much larger effective variant space. The ML-designed library outperformed the NNK library by yielding approximately tenfold more viable, infectious variants after selection for brain tissue infection, including variants with specificity for glial cell populations, highlighting potential for cell-type-targeted therapy. Using the same optimization framework, the authors show via Pareto frontier analysis^[9] that, while libraries synthesized with individually specified sequences could theoretically outperform positionally degenerate libraries, the current method achieves near-optimal balance within practical constraints. From the perspective of multi-round directed evolution, the improved effective diversity after a single packaging step is particularly consequential: by front-loading the library with variants more likely to survive early packaging bottlenecks, ML-designed libraries can attenuate the rapid loss of diversity and premature convergence that often plague conventional NNK campaigns. While Zhu *et al.*^[1] primarily analyze one major packaging step, their framework naturally lends itself to iterative cycles in which updated fitness models are re-trained on later-round data, enabling adaptive re-design or re-weighting of sequence space to preserve exploration while progressively enriching

Table 1. Comparison of NNK insertion-library and ML-/entropy-based library-design

Aspect	Baseline NNK peptide insertion-library	ML-/entropy-designed AAV library
Library construction principle	Uniform use of NNK codons to maximize nominal amino acid diversity with minimal stop codons	Library encoded as an optimized probability distribution over sequences (approximated by position-wise nucleotide frequencies) subject to an entropy (diversity) and predicted fitness constraint
Use of functional data	No explicit incorporation of prior functional information; library is agnostic to packaging or infectivity	Fitness model trained on deep sequencing-derived log enrichment (pre/post-packaging) guides the design, so higher-probability sequences are predicted to package better
Diversity efficiency	High nominal diversity but large fractions of non-packaging or poorly packaging variants; substantial “wasted” sequence space	Effective diversity is enriched for variants that are both diverse and likely to package, yielding higher “usable” diversity per screening round
Packaging-related performance	One round of packaging imposes a strong bottleneck and causes marked loss of effective diversity in the surviving pool	Designed libraries show several-fold higher packaging efficiency and retain a broader effective sequence repertoire after packaging
Treatment of diversity-fitness trade-off	Implicit and uncontrolled; diversity and packaging-fitness are not explicitly balanced or tunable	Trade-off is explicit and tunable via a Lagrange multiplier (λ), generating a Pareto front of designs that systematically span different diversity-fitness combinations.
CMC-relevant manufacturability	Simple to specify (degenerate codons), but prone to synthesis biases and high fractions of unproductive variants, which can reduce effective titers and complicate analytical characterization	Designs are expressed as positional nucleotide distributions mapping naturally onto degenerate oligos, allowing iterative calibration between in silico distributions and realized lot-to-lot composition, still enriching for functional variants
Applicability to multi-round directed evolution (DE)	Early bottlenecks and rapid convergence can limit exploration of sequence space over successive rounds of selection	By enriching for packaging-competent variants up front, ML-designed libraries can mitigate early bottlenecks and are conceptually better suited for multi-round, adaptive DE workflows that update the model across rounds
Extensibility to other objectives and serotypes	Not designed to account for serotype-specific constraints or multi-attribute objectives (e.g., immune evasion, tropism, CQAs)	Framework is serotype-agnostic in principle and can be extended to multi-objective optimization that includes packaging, tropism, and CQA-relevant properties in the same information-theoretic design

NNK: Where N represents any nucleotide and K represents G or T; AAV: adeno-associated virus; ML: machine learning; CMC: chemistry, manufacturing, and controls; CQAs: critical quality attributes.

functional variants. It should be emphasized that these enrichment-derived packaging-fitness estimates capture only one dimension of AAV CQAs: while they strongly relate to capsid assembly and genome encapsidation, they do not directly report on full/empty capsid ratios, VP1/VP2/VP3 stoichiometry, aggregation, or higher-order structural integrity, all of which are central CMC-relevant CQAs for clinical translation. In this sense, ML-designed libraries optimize a key upstream constraint on capsid viability, but must be integrated with orthogonal analytical assays in later development stages to comprehensively qualify vector quality.

Mathematical underpinnings, experimental validation and reproducibility

The distributional optimization relies on maximum entropy principles. For constrained library synthesis, this distribution is approximated by optimizing over product distributions (independent positional probabilities), solved using stochastic gradient descent with score-function gradients and Monte Carlo approximation. This technical development is crucial for real-world library synthesis. The optimal balance between diversity (sequence space coverage, measured by entropy) and mean fitness is visualized as a Pareto frontier. Each point on this curve represents a feasible library design; traditional NNK libraries lie far from the frontier^[10], dominated in both fitness and effective diversity by ML-guided libraries (see Table 1).

Experimental quantification of titer, diversity, and downstream infectivity is robust, with careful controls for cross-packaging and sequencing depth. Validation extends to human adult brain tissue, addressing translational relevance. All code, data, and library specifications are made available, with repository access provided for reproduction and extension. Nonetheless, output is restricted to packaging-related performance and effective sequence diversity at the library level; comprehensive product quality and clinical suitability

cannot be inferred from enrichment scores alone and require dedicated CMC characterization.

IMPACT, EXTENSIONS, LIMITATIONS, OPEN QUESTIONS AND FUTURE OUTLOOK

By interpreting the study by Zhu *et al.*^[1] through the combined lenses of information theory, CMC constraints, and translational AAV engineering, this commentary highlights how entropy-constrained design can rescue wasted diversity in classical NNK libraries and inform the structuring of future multi-round and multi-objective directed evolution campaigns. In particular, for CNS indications where tissue access, safety, and manufacturability are tightly interlinked, such principled design frameworks help bridge the gap between exploratory library screening and regulatory-grade product development. The generality of the approach extends beyond AAV and peptide-insertion libraries to mutagenesis, recombination, or antibody library design - any scenario in which a trade-off exists between function and diversity and library construction is subject to practical constraints. Beyond these experimentally supported findings, several potential implications for *in vivo* CNS gene therapy and blood-brain barrier (BBB) biology are necessarily more speculative. Nonetheless, by maximizing the probability of success downstream, these methods enable faster discovery and translation for gene therapies, especially for hard-to-access indications such as CNS disorders. This is particularly relevant for late-age-of-onset conditions, such as genetic tauopathies, which require careful consideration^[11,12]. Open questions remain regarding how best to extend the approach to optimize multiple properties simultaneously (e.g., packaging, cell-type specificity, immune evasion) or to incorporate higher-order interaction models in library design. For CNS disorders, both differential cellular tropism and BBB permeability must be considered, similar to the organ tropism on which current AAV therapies are built. Future models must first maintain or improve packaging, immune evasion, and BBB permeability/organ tropism^[13,14], before refining cellular or neuronal subtype specificity as previously demonstrated^[15]. The scalability of the framework to larger sequence spaces, as well as the impact of experimental noise and model uncertainty, will be important for future extensions. As synthesis costs continue to decrease^[16], unconstrained sequence-level library design becomes increasingly attractive, and the same framework will be directly applicable. Although Zhu *et al.*^[1] demonstrate their framework using an AAV5-based 7-mer insertion library, the underlying methodology is conceptually serotype-agnostic: any capsid scaffold for which robust high-throughput fitness readouts (e.g., packaging, infection, or transgene expression) can be obtained is amenable to similar entropy-constrained optimization. Serotype-specific differences in insertion tolerance, capsid assembly stability, and structural constraints mean that learned fitness models and optimal trade-off curves will differ for AAV2, AAV9, or engineered backbones; nonetheless, the same workflow - generate a baseline library, measure enrichment, train a predictive model, then re-optimize the library distribution - should scale across capsids, provided the experimental dynamic range and assay noise are well characterized. Regions of sequence space that are poorly sampled, yet potentially biologically interesting, may be systematically underexplored, although they might provide qualitatively distinct functional solutions (see below).

Nonetheless, the current framework remains fundamentally constrained by the information content of the training data. Fitness models are learned from a baseline library and specific experimental assays, and thus primarily interpolate within, or modestly extrapolate from, sufficiently sampled sequence space. Variants that are rare, structurally unusual, or located in poorly covered regions of sequence space may contribute negligibly to the training signal and are therefore unlikely to be assigned high probability in the optimized library. Accordingly, the maximum-entropy constraint increases functional diversity within the modeled sequence space but does not guarantee systematic exploration of genuinely uncharted regions that may harbor qualitatively novel or disruptive capsid variants. This means the framework proposed by Zhu *et al.*^[1] is powerful for refining and densifying functionally productive regions of sequence space but should be complemented by exploratory or unguided library components when the primary goal is to uncover rare, out-of-distribution variants with fundamentally new properties.

Despite its impressive performance, the framework inherits all the usual caveats of supervised ML trained on high-throughput sequencing data: fitness predictions are limited by the experimental dynamic range, noise structure, and coverage of the training set. Poorly sampled regions of sequence space, while potentially biologically interesting, may be systematically underexplored, and reliance on a single assay (packaging plus a specific infection readout) risks encoding assay-specific biases rather than globally relevant AAV properties. Principally, a key limitation of the maximum-entropy, ML-guided approach is that it operates within the informational boundaries of the training data and assay. Fitness models trained on a baseline library predominantly interpolate within regions sufficiently sampled and rewarded by the chosen readout. Variants that are rare, structurally unconventional, or poorly covered contribute little to the training signal and are therefore unlikely to receive substantial probability mass in the optimized library, limiting qualitative novelty. For early stages of directed evolution, this suggests a pragmatic mixed strategy: rational, maximum-entropy libraries can front-load campaigns with packaging-competent variants, while deliberately exploratory or partially unguided library components are maintained in parallel to probe rare or unconventional sequence motifs. Only after such exploratory phases have mapped additional productive regions does it become optimal to rely more heavily on purely exploitative, ML-refined library designs.

Conceptually, the current implementation is predominantly single-objective, focusing on packaging and one infection context. For translational programs, capsid design must ultimately reconcile multiple partly antagonistic objectives (immune evasion, manufacturability, tissue specificity, safety), which will require either genuinely multi-objective optimization or hierarchical, stage-specific design. Extending the maximum-entropy framework to such multi-attribute landscapes, while controlling overfitting and maintaining interpretability for regulatory review, remains an important open challenge. Moreover, the robustness of the learned fitness landscape across different production platforms, analytical pipelines, and tissue models remains to be systematically evaluated. For example, shifting from *ex vivo* adult human brain tissue to *in vivo* small-animal or non-human primate models could expose shifts in the Pareto frontier that necessitate re-training and re-optimization, raising practical questions about how often and at which stages design loops should be closed in real development programs. Additional open questions remain regarding AAV-based gene therapies in general, including clinical general^[17] and liver^[18] toxicity, cargo capacity^[3,19], and targeting difficult-to-access organs such as the brain/CNS^[20]. The study by Zhu *et al.*^[1] demonstrates that recombinant AAV-based carrier systems represent a competitive and efficient platform for gene therapy.

DECLARATIONS

Authors' contributions

The author contributed solely to the article.

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the Alzheimer Forschung Initiative e.V. (Grant No. 22039).

Conflicts of interest

The author declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Zhu D, Brookes DH, Busia A, et al. Optimal trade-off control in machine learning-based library design, with application to adeno-associated virus (AAV) for gene therapy. *Sci Adv.* 2024;10:eadj3786. DOI
2. Müller OJ, Kaul F, Weitzman MD, et al. Random peptide libraries displayed on adeno-associated virus to select for targeted gene therapy vectors. *Nat Biotechnol.* 2003;21:1040-6. DOI
3. Wang JH, Gessler DJ, Zhan W, Gallagher TL, Gao G. Adeno-associated virus as a delivery vector for gene therapy of human diseases. *Signal Transduct Target Ther.* 2024;9:78. DOI PubMed PMC
4. Wang D, Tai PWL, Gao G. Adeno-associated virus vector as a platform for gene therapy delivery. *Nat Rev Drug Discov.* 2019;18:358-78. DOI PubMed PMC
5. Li W, Samulski JR. 743. Directed evolution of adeno-associated virus (AAV) by DNA shuffling yields enhanced gene delivery vectors. *Mol Ther.* 2006;13:S287. DOI
6. Maheshri N, Koerber JT, Kaspar BK, Schaffer DV. Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nat Biotechnol.* 2006;24:198-204. DOI PubMed
7. Fu X, Suo H, Zhang J, Chen D. Machine-learning-guided directed evolution for AAV capsid engineering. *Curr Pharm Des.* 2024;30:811-24. DOI
8. Guo J, Lin LF, Oraskovich SV, Rivera de Jesús JA, Listgarten J, Schaffer DV. Computationally guided AAV engineering for enhanced gene delivery. *Trends Biochem Sci.* 2024;49:457-69. DOI PubMed PMC
9. Yang C, Ye W, Li Q. Review of the performance optimization of parallel manipulators. *Mech Mach Theory.* 2022;170:104725. DOI
10. Kawakami T, Murakami H. Genetically encoded libraries of nonstandard peptides. *J Nucleic Acids.* 2012;2012:713510. DOI PubMed PMC
11. Zempel H. Genetic and sporadic forms of tauopathies-TAU as a disease driver for the majority of patients but the minority of tauopathies. *Cytoskeleton.* 2024;81:66-70. DOI PubMed
12. Langerscheidt F, Wied T, Al Kabbani MA, van Eimeren T, Wunderlich G, Zempel H. Genetic forms of tauopathies: inherited causes and implications of Alzheimer's disease-like TAU pathology in primary and secondary tauopathies. *J Neurol.* 2024;271:2992-3018. DOI PubMed PMC
13. Walkey CJ, Snow KJ, Bulcha J, et al. A comprehensive atlas of AAV tropism in the mouse. *Mol Ther.* 2025;33:1282-99. DOI PubMed PMC
14. Keng CT, Guo K, Liu YC, et al. Multiplex viral tropism assay in complex cell populations with single-cell resolution. *Gene Ther.* 2022;29:555-65. DOI PubMed PMC
15. Ravindra Kumar S, Miles TF, Chen X, et al. Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. *Nat Methods.* 2020;17:541-50. DOI PubMed PMC
16. Hoose A, Vellacott R, Storch M, Freemont PS, Ryadnov MG. DNA synthesis technologies to close the gene writing gap. *Nat Rev Chem.* 2023;7:144-61. DOI
17. Zhao Q, Peng H, Ma Y, Yuan H, Jiang H. In vivo applications and toxicities of AAV-based gene therapies in rare diseases. *Orphanet J Rare Dis.* 2025;20:368. DOI PubMed PMC
18. Piccolo P, Brunetti-Pierri N. Current and emerging issues in adeno-associated virus vector-mediated liver-directed gene therapy. *Hum Gene Ther.* 2025;36:77-87. DOI PubMed
19. Zwi-Dantsis L, Mohamed S, Massaro G, Moeendarbary E. Adeno-associated virus vectors: principles, practices, and prospects in gene therapy. *Viruses.* 2025;17:239. DOI PubMed PMC
20. Kantor B, O'Donovan B, Chiba-Falek O. Trends and challenges of AAV-delivered gene editing therapeutics for CNS disorders: implications for neurodegenerative disease. *Mol Ther Nucleic Acids.* 2025;36:102635. DOI PubMed PMC

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.