Machine learning-based library design improves packaging and diversity of adeno-associated virus (AAV) libraries

Danqing Zhu*[1], Ph.D., David H. Brookes*[2], Ph.D., Akosua Busia[3], Ana Carneiro[4], Clara Fannjiang[3] Galina Popova[5,6,7], Ph.D., David Shin[5,6,7], Edward F. Chang, MD[8], Tomasz J. Nowakowski[5,6,7,8,9] Ph.D., Jennifer Listgarten[3, 10] *, Ph.D., and David. V. Schaffer[1,4,11,12,13,14] *, Ph.D.

(* equal contributions, corresponding)

[1] California Institute for Quantitative Biosciences, University of California, Berkeley, CA, USA
[2] Biophysics Graduate Group, University of California, Berkeley, CA, USA
[3] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[4] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA
[5] Department of Anatomy, University of California, San Francisco, CA, USA
[6] Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA
[7] Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, University of California, San Francisco, CA, USA
[8] Department of Neurological Surgery, University of California, San Francisco, CA, USA
[9] Weill Institute for Neurosciences, University of California at San Francisco, San Francisco, CA, USA
[10] Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA
[11] Department of Bioengineering, University of California, Berkeley, California, CA, USA
[12] Department of Molecular and Cell Biology, University of California, Berkeley, California, USA
[13] Helen Wills Neuroscience Institute, University of California, Berkeley, CA, 94720, USA
[14] Innovative Genomics Institute (IGI), University of California, Berkeley, California, USA

Corresponding authors:
Jennifer Listgarten. Ph.D.
Professor
Electrical Engineering and Computer Science, Center for Computational Biology

University of California, Berkeley
387 Soda Hall, MC 1776
Berkeley, CA, 94720-1776
Email: jennl@berkeley.edu

David. V. Schaffer. Ph.D.
Hubbard Howe Jr. Distinguished Professor
Chemical and Biomolecular Engineering, Bioengineering, Molecular and Cell Biology, and the Helen Wills Neuroscience Institute

Director
California Institute for Quantitative Biosciences (QB3)

University of California, Berkeley
274 Stanley Hall
Berkeley, CA 94720-3220
Tel: (510) 643-5963
Fax: (510) 642-4778

E-mail: schaffer@berkeley.edu

Keywords: AAV; capsid engineering; machine-learning-based library design; diversity-fitness trade-off; maximum entropy; gene therapy.

Abstract

AAVs hold tremendous promise as delivery vectors for clinical gene therapy. Yet the ability to design libraries comprising novel and diverse AAV capsids, while retaining the ability of the library to package DNA payloads, has remained challenging. Deep sequencing technologies allow millions of sequences to be assayed in parallel, enabling large-scale probing of fitness landscapes. Such data can be used to train supervised machine learning (ML) models that predict viral properties from sequence, without mechanistic knowledge. Herein, we leverage such models to rationally trade-off library diversity with packaging capability. In particular, we show a proof-of-principle application of a general approach for ML-guided library design that allows the experimenter to rationally navigate the trade-off between sequence diversity and fitness of the library. Consequently, this approach, instantiated with an AAV capsid library designed for packaging, enables the selection of starting libraries that are more likely to yield success in downstream selections for therapeutics and beyond. We demonstrated this increased success by showing that the designed libraries are able to more easily infect primary human brain tissue. We expect that such ML-guided design of AAV libraries will have broad utility for the development of novel variants for therapeutic applications in the near future.

Adeno-associated viruses (AAVs) hold tremendous promise as delivery vectors for gene therapy. While nature has endowed AAVs with properties that have enabled initial clinical successes, these natural viruses are not optimal for human therapeutic applications. Directed evolution has emerged as a powerful strategy for selecting AAV variants with improved properties such as the ability to evade the immune system or to target specific tissues [1-3]. In this approach, a library of diverse AAV capsid sequences is subjected to multiple rounds of selection, with the aim of identifying and enriching the most effective variants [1, 4]. Despite the success of these approaches, the starting libraries have not been systematically optimized. In particular, it stands to reason that a starting library that is most diverse, while retaining the ability to package its DNA payload, will provide the highest likelihood of success for any particular downstream task. Herein, we set our sights on this problem—to rationally re-engineer the starting library for AAV to make it more optimal for any downstream directed evolution.

Primary library construction strategies include error-prone mutagenesis [1, 5], DNA shuffling [6, 7], and peptide insertion [8], and peptide insertion variants in particular have been increasingly translated to the clinic (e.g., NCT03748784, NCT04645212, NCT04483440, NCT04517149, NCT04519749). Recent studies have explored computational strategies such as leveraging comprehensive single-substitution mapping of the AAV capsid landscape to compute mutation positions and probability distributions for mutagenesis [9], reconstructing ancestral nodes from phylogenetic analysis to identify mutable positions [10], or leveraging structure to identify genomic blocks suitable for a recombination strategy [11]. While these approaches can yield millions of novel sequences, most of the variants fail to assemble into functional capsids or to package their genomes [3, 9, 12]. To mitigate this issue, one round of packaging selection is often performed prior to initiating selections for infectivity, a consequence of which is that the resulting diversity of the library gets dramatically reduced (*e.g.*, more than 50% [9, 12, 13]). If we could design the initial library to have the same packaging ability as these once-selected libraries, but with the diversity of the unselected library, the success of downstream selections would correspondingly benefit.

Deep sequencing technologies allow millions of sequences to be assayed in parallel, enabling large-scale probing of fitness landscapes [14, 15]. Such data can be used to train supervised machine learning (ML) models that predict viral properties from sequences, without mechanistic knowledge. Although some recent studies have reported applying ML models trained on experimental data to generate novel AAV variants, these studies did not systematically consider a trade-off between diversity and packaging [16, 17]. Moreover, they focused on AAV serotype 2 (AAV2), the serotype with the highest prevalence of pre-existing antibodies in a general population, limiting its translational usage for clinical therapeutics outside of the retina [18]. Among the natural variants, AAV serotype 5 (AAV5) has been suggested as a promising candidate for clinical gene transfer because of the low prevalence of pre-existing anti-neutralizing antibodies to be circumvented and successful clinical development for hemophilia B [19, 20]. Here we present a proof-of-principle application of a novel approach for ML-guided AAV5-based capsid library design that allows us to design capsid libraries by optimally balancing diversity and overall packaging fitness. We further show that a library designed in this manner can better infect brain cells as compared to the current state-of-the-art. To our knowledge, this is the first ML-guided AAV capsid library design used for selection in human tissue.

Our approach builds upon prior library design efforts to balance multiple optimization objectives. Most notably, Parker *et al*. [21] designed libraries constructed by combinatorial mutagenesis using an optimization framework that sought to balance "novelty" and "quality" scores. The quality score arises from a Potts model trained on natural sequences, while the novelty score reflects how designed library variants have sequence identities dissimilar to any natural sequences. Importantly, the novelty score encourages sequences to be different from the given set of natural sequences but does not promote library diversity. Verma *et al*. [22] extend this work to balance the novelty score with multiple fitness objectives, but still do not optimize for library diversity. In contrast to this body of work, our approach (i) allows for the use of any predictive model of fitness, (ii) explicitly addresses and controls the diversity within the library itself, and (iii) is broadly applicable to different library construction approaches.

We focused on designing an AAV5 capsid-based library wherein each viral protein (VP) monomer contains a 7-amino acid (7-mer) insertion region flanked by amino acid linkers (TGGLS) at position 587-588, within a loop at the 3-fold symmetry axis associated with receptor binding and cell-specific entry [23, 24]. The baseline insertion library that we sought to improve upon, which has been used successfully in previous studies [2], was constructed by sampling insertion sequences from 7 concatenated NNK degenerate codons. That is, a given codon is chosen at random by sampling each of the first two nucleotides with equal probability from A, C, T, G, and the last nucleotide equally from only T and G, a design intended to yield high diversity while avoiding stop codons that would ablate VP fitness. However, even beyond a stop codon, certain amino acid choices likely destabilize the structure and/or compromise protein fitness, and just one anticipated example would be placing a large hydrophobic residue on this solvent-exposed region. Thus, in these NNK libraries, a substantial fraction (>50%) of sampled sequences typically fail to assemble into viable capsids, and other do not assemble as well as natural variants, such that much of the library is effectively wasted [9, 12, 13]. The goal of the present work is to mitigate such wasted effort while maintaining diversity of the library.

Our overall workflow to enhance a library was to (1) synthesize and sequence a baseline NNK library, the "pre-packaged" library; (2) transfect the library into packaging cells (i.e., HEK 293T) to produce AAV viral vectors, harvest the successfully packaged capsids, extract viral genomes, and sequence to obtain the "post-packaging" library; (3) build a supervised regression model where the target variable reflects the packaging success of each insertion sequence found; (4) "invert" the predictive model to design libraries with optimal trade-offs between diversity and fitness; and (5) select a library design with a suitable tradeoff. We then validated both the predictive model and the designed library by experimentally measuring library packaging success and sequence diversity. Finally, we demonstrated that our ML-designed library is better able to infect primary human brain tissues as compared to the baseline NNK library.

AAV-7mer library preparation and packaging selection. The theoretical size of a 7xNNK insertion library is ~10^{13}, but cloning limits an experimental library to be ~ 10^7. We propose to develop a ML model to focus on a library sequence composition on regions of sequence space that package well, so as avoid "wasting" library and experimental capacity on sequences that do not package well. First, we synthesized an NNK library with ~ 10^7 capsid-modified variants, and the resulting *pre-packaged library* was then packaged in HEK293T cells (**Methods**). After 72 hours, this *post-packaging vector library* was harvested and purified by gradient ultracentrifugation

(**Figure 1**) [25]. The NNK sequence from both *pre*- and *post*-packaged libraries were then PCR amplified and deep sequenced.

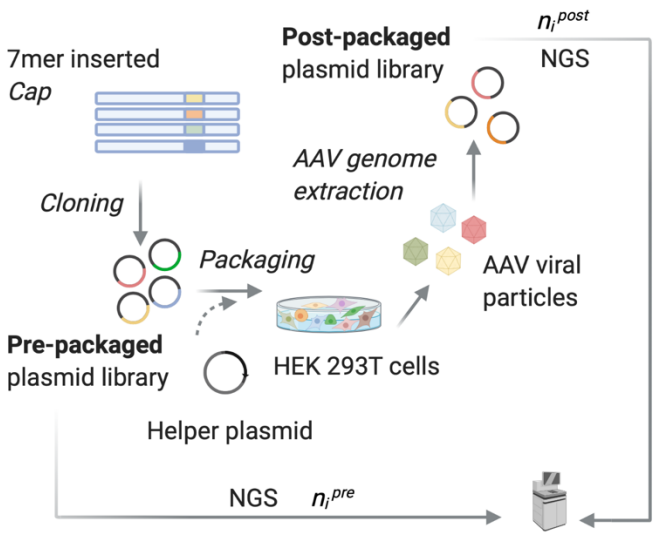


Figure 1: Schematic flowchart of generating *pre*- and *post*-packaged libraries.

We processed the raw data containing 49,619,716 *pre*- and 55,135,155 *post*-packaged sequence reads, which collectively yielded read counts for 8,552,729 unique peptide sequences (**Methods)**. We calculated an "enrichment score" for each unique sequence—the normalized log ratio of the counts of a sequence in the *post*- and *pre*-selection libraries [9, 16, 26, 27] (**Methods**). The enrichment score (**Equation 1**) represents a measure of a given sequence's ability to package. This score by itself, however, does not account for the fact that a variant that appears for example in 10 and 100 reads pre- vs. post-packaging, respectively, is better statistically sampled than one appearing 1 and 10 times. To incorporate this added level of information into our predictive modelling, we statistically derived a weight for each sequence that reflects how much impact it should have on the predictive model (**Methods**). In the running example, the 10:1 data would get a smaller weight than 100:10, as it provides less evidence of enrichment. These weights were used in all predictive models to train weighted regression models, unless otherwise noted.

Training and evaluation of predictive models. Our first goal was to find a suitable model class for our prediction task, after which we used the best performing model to perform our ML-based library design. Toward that end, we trained several ML-based regression models (**Methods**), using the log enrichment scores as the target variable and associated weights described above. The input to the model was an encoding of the 7-mer amino acid insertion sequence, described next. Seven model were evaluated: three linear models and four feed-forward neural networks (NNs). The three linear models differed in the set of input features used. One linear model used the "Independent Site" (IS) representation wherein amino acids in each sequence were one-hot encoded and a parameter assigned to each bit. Another used a "Neighbors" representation that comprised the IS features, and additionally pairwise interactions of all such features that are directly adjacent to each other in the amino acid sequence. Finally, we used a "Pairwise" representation, which comprised the IS features, and additionally all pairwise interactions of all such features, irrespective of position. All neural network models used the IS features alone, as these models

have the ability to effectively generate new features. Each NN architecture comprised exactly two densely connected hidden layers with tanh activation functions. The four NN models differed in the size of the hidden layers, with each using either 100, 200, 500, or 1000 nodes in both hidden layers.

We compared the performance of these models using the standard (unweighted) Pearson correlation between model predictions and true enrichment scores on a held-out test set. We randomly split the data into a training set containing 80% of the data points and a test set containing the remaining 20% of the points. In addition to computing the Pearson correlation on the entire test set, we computed it on subsets of the test set restricted to the K% most truly enriched sequences ("Fraction of top test sequences"); as we varied K, this traced out a performance curve, where for lower K%, the evaluation is focused more on accurately predicting the higher enrichment scores rather than the lower ones. This evaluation approach was useful because, ultimately, we planned to design a library with sequences that have high packaging enrichment, and so wanted to determine how the predictive accuracy changed in this regime (**Figure 2A**). Overall, we found that that the NN models performed better than the linear models, presumably owing to their capacity to model more complex mappings—in particular, to capture higher-order epistatic interactions in the fitness function. We selected our final model by focusing on the regime near K=0.1, finding that "NN, 100" performed best here.

To assess how the weighted part of our regression affected model performance, we re-trained with a standard (unweighted) regression loss, on the final selected model and the linear, pairwise model (**Figure 2B)**. Using the weighted loss function resulted in a clear performance benefit for K<0.25, the regime of interest. Notably, as we move toward K=1, the weighted loss function slightly degrades performance, presumably because the vast majority of the points with lower enrichment scores have few counts.

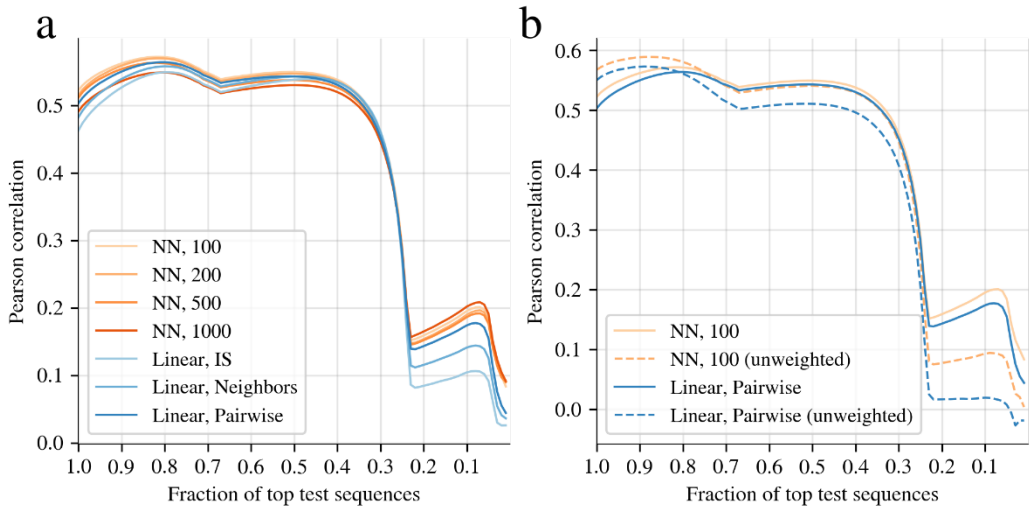


Figure 2: Comparison of models for predicting AAV5-7mer enrichment scores using Pearson correlation between the predicted and true enrichment scores. "Fraction of top test sequences" denotes sets of test sequences chosen based on their true fitness values. (a) Different neural networks (NN), where the number denotes number of nodes in the hidden layers, and a weighted loss is always used. (b) Effect of using a non-weighted loss, compared to our weighted loss, for the final selected model (NN, 100), and a baseline.

Model-guided library design. Having identified the best predictive model for our library design task, the next step was to computationally design a library of insertion sequences that packages substantially better than the NNK library yet maintains a broad diversity of insertions to enhance performance in downstream functional selection. Inherent in this challenge is a trade-off between the diversity vs. predicted packaging fitness of the library. To gain insight into this trade-off, consider that the library that optimizes the average fitness will contain only the single sequence that is the most fit, whereas the library that optimizes diversity is uniformly distributed across sequence space, irrespective of packaging fitness. The library that is most effective for downstream selections will lie between these two extremes, balancing expected packaging fitness with diversity.

Although it is not *a priori* clear what the best trade-off should be within these two extremes, one can make use of an optimal trade-off curve, also known as a "Pareto frontier". For any library lying on this optimal frontier, it is not possible to do better on one criterion (packaging or diversity), without hurting the other. Part of our approach to library design is to provide the tools to trace out such an optimality curve. By generating, and then examining the Pareto optimal curve—which enables us to assess what levels of diversity can still allow highly fit libraries—we are able to reason about where a suitable library for downstream selection is likely to lie. To generate points that trace out the optimal curve, we built on our previously described methods for "inverting" fitness predictive models to design proteins with high fitness [28]—now additionally, coherently, enforcing a library diversity constraint (formally, the statistical entropy of the library), which we can set to different levels, λ, to trace out the Pareto frontier. We refer to these designed libraries as "maximum entropy" libraries. The frontier is approximate rather than exact because the underlying optimization problem is not convex, and thus formally intractable. Nevertheless, a frontier indeed emerges, with some striking results. Most notably, the baseline NNK library has a dramatically poor mean predicted enrichment, while a designed library D3 has substantially better predicted mean enrichment, but with very little loss in diversity. Similarly, library D2 is slightly less diverse than D3, but with substantially higher predicted enrichment. Note that each designed library specifies the marginal probabilities of individual nucleotide at the 21-bp insertion positions (**Figure 3b-d**, **Supplementary Information**).

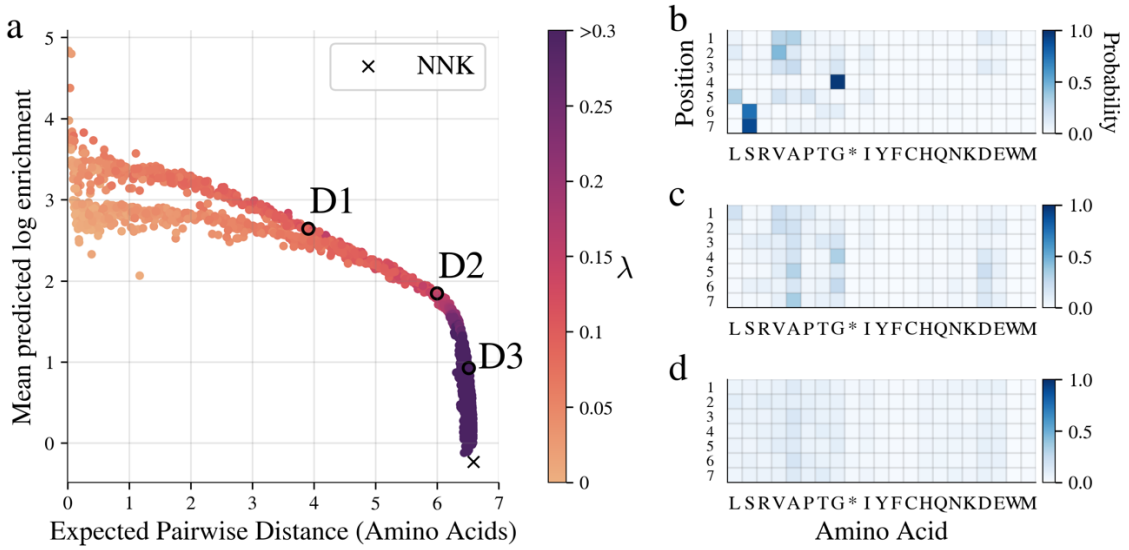
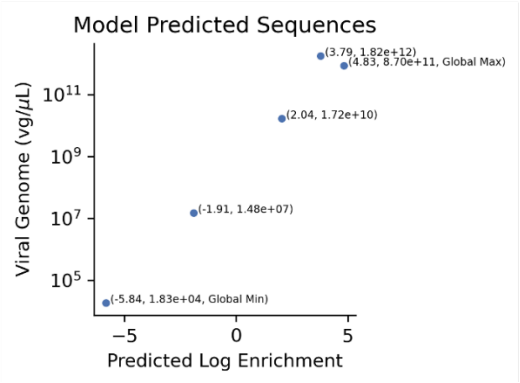


Figure 3: Results of maximum entropy library design for AAV5-based 7-mer insertion sequences. Each point in a) represents a library resulting from fitness optimization subject to a particular diversity constraint, λ (higher values yields more diverse libraries). Expected pairwise distance is a human-interpretable measure of diversity, as compared to statistical entropy, which was actually used in the library design procedure. Mean predicted log enrichment is a measure of the overall library fitness. The baseline NNK library is denoted with an "x", while several other designed libraries have been labelled D1-3. Points falling within the convex hull ("outer envelope") of the points are sub-optimal and arise from the non-convex optimization (b-d) designed library parameters (marginal probability of each nucleotide at each position) for the three libraries D1-3 highlighted in a).

Experimental validation of the predictive models. As a precursor to experimentally testing library design, we assessed the quality of the predictive models by identifying and synthesizing five individual 7-mer insertion sequences that were not present in our original dataset. These five variants were chosen to span a broad range of predicted log enrichment scores (-5.84 to 4.83—see Figure 4 for correspondence with viral titers). The strong agreement between model predictions and corresponding experimental measurement of vector production titers (1.83E+04 to 8.70E+11) (**Figure 4**) demonstrated that the predictive model was sufficiently accurate to be used for design.



Sequences	Predicted Log Enrichment	Experimental Viral Titer (vg/μL)
LSSTTAA	4.834	8.70E+11
DSRLSGT	3.793	1.82E+12
LEPDAAL	2.044	1.72E+10
IRWRATG	-1.91	1.48E+07
RWPRRVL	-5.84	1.83E+04

Figure 4: Experimental titers (viral genome/μL) versus model predicted enrichment scores. The five variants were selected with a broad range of predicted log enrichment scores.

To assess the accuracy of the designed libraries' trade-off between diversity and expected fitness, we then synthesized two designed libraries (namely D2 and D3) from our optimality curve. In particular, library D2 had predicted diversity comparable to that of the NNK library but with expected fitness in the top 25% of libraries, in contrast to NNK, which has a lower expected fitness than all designed libraries. Library D3 had slightly less diversity but an expected fitness roughly in the top 50% of designed libraries.

Following oligonucleotide synthesis following the marginal nucleotide probabilities of the D2 and D3 library designs, we constructed the corresponding AAV VP libraries. Deep sequencing showed that the library designs vs. experimental nucleotide distributions were within 5%. Next, the two designed libraries were transfected and harvested with the same methods as our NNK baseline library. Experimental titers were then measured for these two designed libraries and the NNK using digital droplet PCR (ddPCR) with Hex-ITR probes tagging the conserved regions of encapsidated viral genome of AAV. The Pearson correlation of these titers with the predicted enrichment scores revealed a strong positive correlation ($R^2 = 0.9$, **Figure 5A**). Additionally, both designed libraries (D2 and D3) outperformed NNK library in packaging without compromising their diversity (**Figure 5B**), with library D2 (predicted log enrichment score ~ 2.0) showing 5-fold higher packaging titer than that of the NNK library (predicted log enrichment score ~ -0.9). From these results we concluded that our library design approach was indeed able to trade off packaging fitness with diversity as intended. To take the validation one step further, the already-packaged NNK library was packaged again to further select for fit variants, and the resulting titer (4.38E+11) was still significantly lower than that of the initial library D2 (5.12E+11), suggesting the ML-based library design procedure was highly effective in designing for both high packaging fitness and diversity. (**Figure 5C**)

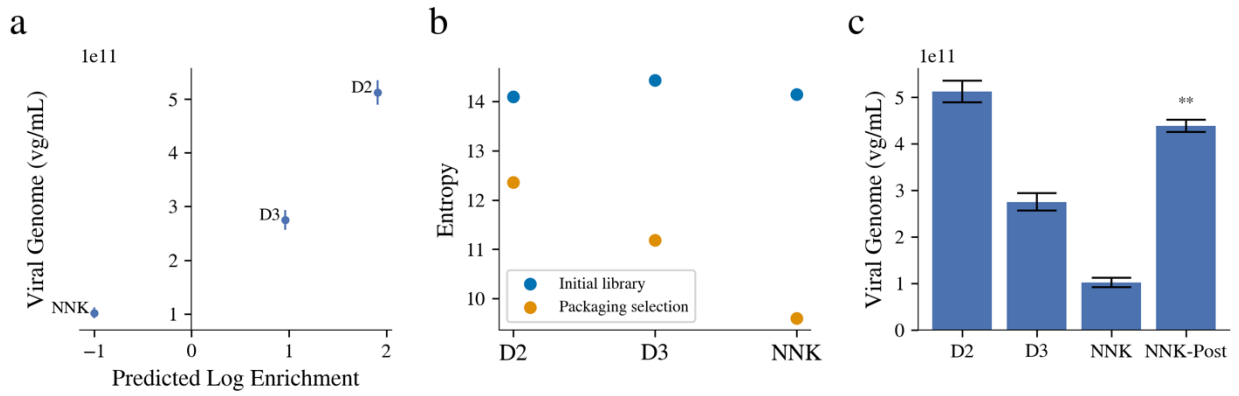


Figure 5: Comparison of ML-designed library D2 and D3 to the baseline NNK library. (a) Experimental titers (viral genome/mL) plotted against the model predicted log enrichment scores. (b) Entropy comparison represents the diversity of sequences present in each library after packaging. (c) Experimental titers across ML-designed library D2 and D3, NNK, and NNK-post library. NNK-post library represents the selected NNK library after one round of packaging. (** p <0.01; two-sided student's t-test).

ML-designed AAV library for primary brain tissue infection. As a final assessment of how the designed libraries' optimal tradeoff properties benefit a downstream selection task, we investigated the ability of each library to infect primary adult brain tissue. We applied ~ 50 µL (equal viral particles, corresponding to an approximate MOI of 10,000) of library D2 or the NNK library virus onto a ~300 µm human adult brain slice and harvested the tissues after 72 hours of infection. Fragments containing the 7-mer sequences were amplified by PCR and subjected to Illumina sequencing. We evaluated the success of each library on this task by computing the diversity of sequences found after infectivity selection, with the premise that more diversity suggests that the starting library had more chances for success. We found that designed library, D2, had a 1.29-fold higher post-selection entropy than the NNK library (**Figure 6**). This increase in entropy corresponds to approximately 33,000 more effective sequences in the designed library after selection, where the effective number of sequences refers to the number of equally abundant sequences required to obtain the same entropy as the one observed. This result suggests that our designed library D2 is likely an effective, general starting library for downstream selections, and in particular substantially better than the widely used NNK library.

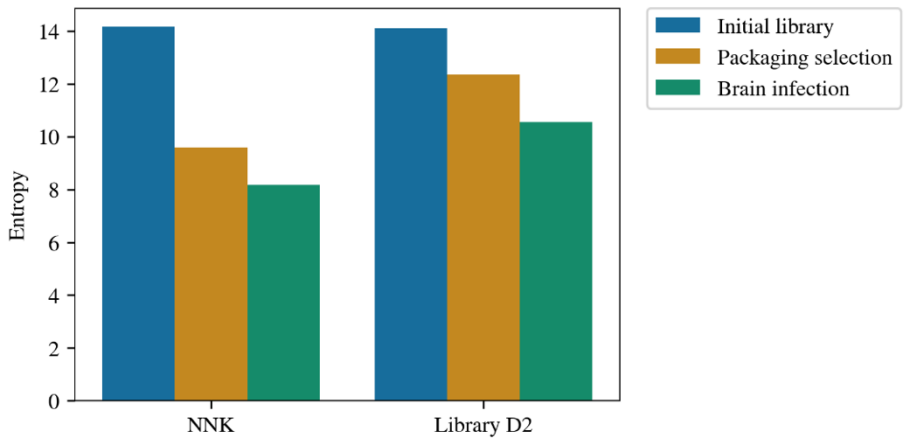


Figure 6: Entropy (diversity) comparison between NNK and ML-designed library D2 after packaging and infection. Library D2 presents comparable level of initial diversity to that of the NNK library but outperforms the NNK library after both packaging and primary brain infection.

Altogether, we have shown (i) that we can build accurate predictive models for AAV library packaging fitness; (ii) that we can leverage these libraries to design libraries that optimally trade off diversity with packaging fitness; (iii) and that these designed libraries are likely to more useful for downstream selection than standard libraries used today. Our approach can also be used to optimize libraries for any downstream selection desired, including those with fitness for therapeutic applications such as gene replacement in the nervous system or evasion of pre-existing antibodies. Moreover, we can in principle use our approach for libraries with any kind of design parameters, such as specifying specific sequences one-by-one, and so forth. We expect that these ML-designed AAV libraries will have broad utility for the development and selection of novel variants targeting different cells and tissues for therapeutic applications in the near future.

Acknowledgments
D.Z. was supported by Siebel Fellowships. D.Z., T.J.N. and J.L were supported by the Chan Zuckerberg Biohub. A.B., C.F., A.C. were supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814, Grant No. DGE 2146752, and Grant No. DGE 2146752; any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. G.P. was supported by NIH NRSA F32 1F32MH118785. This work was supported, in part, by a grant from The Shurl and Kay Curci Foundation (T.J.N) and gifts from the Schmidt Futures and the William K. Bowes Jr Foundation (T.J.N.). Schematic illustrations were created with BioRender.com.

Declaration of Interests
D.Z., D.H.B., J.L., and D.V.S. are inventors on patent related to improving packaging and diversity of AAV libraries with machine learning. David V. Schaffer is a co-founder of 4D Molecular Therapeutics. Jennifer Listgarten is on the Scientific Advisory Board for Foresite Labs and Patch Biosciences. Other authors declare no competing interests.

Online Materials and Methods:

Construction of the NNK-based 7mer Insertion Library and Vector Packaging

$(NNK)_7$ oligo was first synthesized (Elim) and introduced to the 5' end of the right fragment by a primer overhang. Left and right fragments were each PCR amplified by primers Seq_F/Seq_R and 7mer_F/7mer_R, respectively (**Supplementary Table**). PCR products of the two fragments were then purified individually and proceeded to the overlap extension PCR using Vent DNA polymerase (Thermofisher) with equimolar amounts of the left and right fragments for a total of 250ng DNA templates. The resulted library was then digested with *HindIII* and *NotI* and ligated into the replication competent AAV packaging plasmid pSub2. The resulting ligation reaction was electroporated into *Escherichia coli* for plasmid production and purification. Replication competent AAV was then packaged as been described previously [11, 23]. In short, AAV library vectors were produced by triple transient transfection of HEK293T cells with the addition of the pRepHelper, purified via iodixanol density centrifugation, and buffer exchanged into PBS by Amicon filtration.

AAV Viral Genome Extraction and Titer

Packaged AAV vectors were first combined with equal volume of 10X DNase buffer (New England Biolabs, B0303S) and 0.5 µL 10 U/µL DNase I (New England Biolabs, M0303L) incubate for 30 min at 37 ºC. Then equal volume of 2x Proteinase K Buffer (xx) was added with sample to break open capsid. After heat inactivating for 20 min at 95 ºC, the sample was further diluted at 1:1000 and 1: 10,000 and use as templates for titer. DNase-resistant viral genomic titers were measured using digital-droplet PCR (ddPCR) (BioRad) using with Hex-ITR probes (CACTCCCTCTCTGCGCGCTCG) tagging the conserved regions of encapsidated viral genome of AAV. After primary tissue infection, capsid sequences were recovered by PCR from harvested cells using primers HindIII_F and NotI_R (**Supplementary Table**). A ~75-85 base pair region containing the 7mer insertion was PCR amplified from harvested DNA. Primers included the Illumina adapter sequences containing unique barcodes to allow for multiplexing of amplicons from multiple libraries. PCR amplicons were purified and sequenced with a single read run-on Illumina NovaSeq 6000.

Data filtering and processing

The raw sequencing data consisted of 49,619,716 and 55,135,155 sequencing reads corresponding to the pre- and post-selection libraries, respectively. Each read contained (i) a 5 bp unique molecular identifier, (ii) a fixed 21 bp primer sequence, (iii) a 6 bp sequence representing the pre-insertion linker (two fixed amino acids that connect the insertion sequence to the capsid sequence at position 587), (iv) a variable 21 bp sequence containing the nucleotide insertion sequence, and (v) a 9 bp representing the post-insertion linker (three fixed amino acids that connect the insertion sequence to the capsid sequence at position 588). We filtered the reads, removing those that either contained more than 2 mismatches in the primer sequences or contained ambiguous nucleotides. After this filtering, the pre- and post- libraries contained 46,049,235 and 45,306,265 reads, respectively. The insertion sequences were then extracted from each read and translated to amino acid sequences.

Enrichment score and variance

We calculated the log enrichment scores (**Equation 1**) for each insertion sequence using the (filtered) sequencing data to quantify each sequence's effect on packaging. Note that only 218,942 of the 8,552,729 unique sequences appear in both the pre- and post-selection libraries. A pseudo-count of 1 was added to each count so that the enrichment score could still be calculated when the sequence appeared in only one of the libraries. In all cases, the natural log was used.

$$y_i = \log\frac{n_i^{post}}{n_i^{pre}} - \log\frac{N^{post}}{N^{pre}} \quad \text{(Eqn.1)}$$

We estimated a variance associated with each log enrichment score using Equation 2, which follows by noting that each of the raw counts associated with an enrichment score is a random variable. Specifically, the count associated with a sequence can be modeled as a Binomial random variable [26]. The log enrichment score (**Equation 1**) is then the log ratio of two Binomial random variables; it can be shown with the Delta Method [29] that, in the limit of infinite samples, the log ratio of two Binomial random variables converges in distribution to a Normal random variable with mean and variance approximated by Equations 1 and 2, respectively [26, 27].

$$\sigma_i^2 = \frac{1}{n_i^{post}}\left(1 - \frac{n_i^{post}}{N^{post}}\right) + \frac{1}{n_i^{pre}}\left(1 - \frac{n_i^{pre}}{N^{pre}}\right) \quad \text{(Eqn. 2)}$$

<u>Model training and evaluation</u>

Our data processing yields a data set of the form $\{(x_i, y_i, \sigma_i^2)\}_{i=1}^{M}$ where the x_i are unique insertion sequences, y_i are log enrichment scores associated with the insertion sequences, σ_i^2 are the estimated variances of the log enrichment scores, and $M = 8{,}555{,}729$ is the number of unique insertion sequences in the data. We randomly split this data set into a training set containing 80% of the data and a test set containing the remaining 20% of the data.

We assume that the distribution of an enrichment score given the associated insertion sequence is

$$y_i | x_i, \sigma_i^2 \sim N(f_\theta(x_i), \sigma_i^2)$$

where f_θ is a function with parameters θ that parameterizes the mean of the distribution, and represents a predictive model for enrichment scores. We determined suitable settings of the parameters θ with Maximum Likelihood Estimation (MLE). The log-likelihood of the parameters of this model given the training set of $M' \leq M$ data points is given by

$$\ell\left(\theta; \{x_i, y_i, \sigma_i^2\}_{i=1}^{M'}\right) = \frac{M'}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{M'}\left[\log \sigma_i^2 + \frac{1}{\sigma_i^2}\left(y_i - f_\theta(x_i)\right)^2\right]$$

Performing MLE by optimizing this likelihood with respect to the model parameters, θ, results in the weighted least-squares loss function in **Equation 3**.

$$L(\theta) = \sum_{i=1}^{M}\frac{1}{\sigma_i^2}\left(y_i - f_\theta(x_i)\right)^2, \quad \text{(Eqn. 3)}$$

For the linear forms of f_θ, the loss (**Equation 3**) is a convex function which can be solved exactly for the minimizing ML parameters. In order to stabilize training, we used a small amount of l_2 regularization for the Neighbors and Pairwise representations (with regularization coefficients 0.001 and 0.0025, respectively, chosen by cross-validation). For the neural network forms of f_θ, the objective (**Equation 3**) is non-convex and we use stochastic optimization techniques to solve for suitable parameters. We implemented these models in TensorFlow [30] and used the built-in implementation of the Adam algorithm [31] to approximately solve **Equation 3**.

To assess the prediction quality of each model, we calculated the Pearson correlation between the model predictions and observed enrichment scores for different subsets of the sequences in the test set. Our

ultimate aim is to use these models to design a library of sequences that package well (*i.e.*, would be highly enriched in the post-selection library). We, therefore, assess how well the models perform for highly enriched sequences by progressively culling the test set to only include sequences with the largest observed enrichment scores (**Figure 2**).

Maximum Entropy Library Design

We developed a general framework for sequence library design that can be used with any predictive model of fitness, is broadly applicable to different library construction mechanisms (e.g., error prone PCR, site-specific marginal probability specification, individual synthesized sequences), and is simple to implement and extend. This framework balances expected predicted packaging fitness with entropy, a measure of diversity for probability distributions which has been used extensively in ecology to describe the diversity of populations [29]. Our approach is based on a maximum entropy formalism: we represent libraries as probability distributions and aim to find "maximum entropy distributions" that maximize entropy while also satisfying a constraint on the expected fitness, which is predicted by a user-specific model such as a neural network.

Let χ be the space of all sequences that may be included in a library (*e.g.*, all amino acid sequences of length 7). We consider a library to be an abstract quantity represented by a probability distribution with support on χ. Let $\wp$ represent all such libraries and $p \in \wp$ one particular library. The entropy of this library is given by [32]:

$$H[p] = -\sum_{x \in \chi} p(x) \log p(x)$$

Now, let $f(x)$ be a predictive model of fitness (*e.g.*, from a trained neural network). Our goal is to find a diverse library, p, where the expected predicted fitness in the library, $\mathbb{E}_{p(x)}[f(x)]$, is as high as possible. Formally, we want to find the library with the largest entropy such that the expected predicted fitness is above some cutoff. This objective is written

$$\begin{aligned} &\max_{p \in \wp} H[p] \\ &\text{s.t. } \mathbb{E}_{p(x)}[f(x)] \geq a \end{aligned}$$

where a is the cutoff on the expected predicted fitness. It is straightforward to show that the solution to this optimization problem is given by [33]:

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp\left(\frac{f(x)}{\lambda}\right) \quad \text{(Eqn. 4)}$$

where $\lambda > 0$ is a Lagrange multiplier that is a monotonic function of the cutoff a and $Z(\lambda) = \sum_{x \in \chi} \exp(f(x)/\lambda)$ is a normalizing constant. **Equation 4** gives the probability mass of what is known as the maximum entropy distribution. The parameter λ controls the balance between diversity and expected fitness in the library (higher λ corresponds to more diversity). Each library, p_λ, represents a point on a Pareto optimal frontier of libraries, which balances diversity and expected predicted fitness; these distributions cannot be perturbed in such a manner as to both increase the entropy and the expected fitness. Theoretically, the entire Pareto frontier could be traced out by calculating the expected predicted fitness and entropy of p_λ for every possible setting of λ. In practice, we pick a discrete set of λ that traces out a practically useful curve. Note that an equivalent view of this maximum entropy library design framework is to add an entropy-regularization term, $\lambda H[p]$, to the DbAS algorithm objective, $\mathbb{E}_{p(x)}[f(x)]$, [28] yielding an overall objective, $\mathbb{E}_{p(x)}[f(x)] + \lambda H[p]$. The CbAS algorithm of [28] additionally employs a "soft trust region" that is used to modulate the design process to avoid pathological input (sequence) areas of the predictive model. However, we did not employ such a trust region for the AAV library design herein

because the amount of data relative to the size of the design space was deemed sufficient in itself to mitigate such risks. For similar reasons, we did not employ the autofocusing, domain-adaptation techniques presented for design in [34].

As written so far, this framework can be used to select a particular library distribution, $p_\lambda(x)$, with value λ, from the Pareto optimal curve; then, if designing libraries comprised of individually specified sequences, to sample individual sequences from this distribution (see **Supplementary Information**), thereby designing a realizable, synthesizable library. However, for many cases of practical interest, it will not be cost-effective to synthesize individual sequences; rather, we will set the parameters of a library mechanism, such as the probabilities of the codons at each position, to generate a library of oligonucleotides in a stochastic manner. Next, we describe how to handle such cases, what we refer to as *constrained* library designs (constrained because we cannot specify each individual sequence).

Maximum Entropy Design for Constrained Libraries

For the capsid insertion library designs of AAV focused on herein, we are designing libraries for which the "control knobs" (those experimental design parameters that we can change to create different libraries) are less precise than being able to specify individual sequences. In particular, we controlled the marginal probabilities of each nucleotide at each position. The probability mass of a distribution representing such a site-specific marginal probability library of sequences of length L and alphabet size K (*i.e.*, K = 4 for nucleotide libraries) is given by:

$$q_\phi(x) = \prod_{j=1}^{L} \sum_{k=1}^{K} q_{\phi_j}(x^j = k)\delta_k(x^j)$$

where $\phi \in \mathbb{R}^{L\times K}$ is a matrix of distribution parameters, ϕ_j is the jth row of ϕ, $\delta_k(x^j) = 1$ if $x^j = k$ and zero otherwise,

$$q_{\phi_j}(x^j = k) = \frac{e^{\phi_{jk}}}{\sum_{l=1}^{K} e^{\phi_{jl}}} \quad \text{(Eqn. 5).}$$

For an arbitrary predictive model (such as a neural network to predict log enrichment scores from sequence), the maximum entropy distribution of **Equation 4** will generally not have the form of **Equation 5**, the latter being the most general form, and thus unconstrained. To apply the maximum entropy formulation to the design of libraries with constraints, what we refer to as *constrained library* design, we take a variational approach and find the constrained library, q_θ, that is the best approximation to the maximum entropy library, p_λ (for a single, fixed value of λ, chosen from the estimated Pareto optimal frontier), in terms of KL divergence:

$$\phi_\lambda = \operatorname{argmin}_\phi D_{KL}[q_\phi || p_\lambda] = \operatorname{argmax}_\phi \mathbb{E}_{q_\phi(x)}[f(x)] + \lambda H[q_\phi] \quad \text{(Eqn. 6).}$$

Our objective (**Equation 6**) is a non-convex function of the library parameters. The Stochastic Gradient Descent (SGD) algorithm has been shown to consistently find optimal or near-optimal solutions to a variety of non-convex problems, particularly in machine learning [35]. We use a variant of SGD based on the score function estimator [36] to solve **Equation 6**. We randomly initialize a parameter matrix, $\phi^{(0)}$, with independent Normal samples, and then update the parameters according to

$$\phi^{(t)} = \phi^{(t-1)} + \alpha \nabla_\phi F(\phi^{(t-1)}) \quad \text{(Eqn. 7)}$$

for $t = 1, \ldots, T$, where we define $F(\phi) \coloneqq \mathbb{E}_{q_\phi(x)}[f(x)] + \lambda H[q_\phi]$ to be the objective function in **Equation 6**. The number of iterations, T, was set such that we observed convergence of the objective function values in most runs of the optimization. After T iterations, we assumed that we had reached a near-optimal solution (*i.e.*, $\phi^{(T)}$ can be used as an approximation of ϕ_λ). The components of the gradient in **Equation 7** are given by

$$\begin{aligned}\frac{\partial}{\partial \phi_{jk}} F(\phi) &= \mathbb{E}_{q_{\phi(x)}}\left[w(x)\frac{\partial}{\partial \phi_{jk}} \log q_{\phi_j}(x^j)\right] \\ &= \mathbb{E}_{q_{\phi(x)}}\left[w(x)\left(\delta_k(x^j) - q_{\phi_j}(k)\right)\right], \quad \text{(Eqn. 8)}\end{aligned}$$

where we define the weights $w(x) \coloneqq f(x) - \lambda\left(1 + \log q_\phi(x)\right)$ (**Supplementary Information**). The expectation in **Equation 8** cannot be solved exactly, so we use a Monte Carlo approximation:

$$\frac{\partial}{\partial \phi_{jk}} F(\phi) \approx \frac{1}{M}\sum_{i=1}^{M} w(x_i)\left(\delta_k(x_i^j) - q_{\phi_j}(x_i^j)\right), \quad x_i \sim q_\phi(x),$$

where M is the number of samples used for the MC approximation. We applied this maximum entropy framework to design site-specific marginal probability libraries of the 21 nucleotides corresponding to the 7 amino acid insertion using the (NN, 100) predictive model of fitness. **Figure 3** shows the near-optimal Pareto frontier resulting from 2,238 such library optimizations with $\alpha = 0.01, T = 2000,$ and $M = 1000$ and a range of settings of λ.

Comparing libraries designed with different diversity penalties, λ

To assess the extent to which our solutions to **Equation 6** trade-off diversity and predicted fitness, for the site-specific amino acid, constrained libraries, we compared two quantities corresponding to each library: the mean predicted enrichment (*i.e.*, fitness) of amino acid sequences samples from the library and the expected hamming distance between any two sequences sampled from the library, which we call the Expected Pairwise Distance (EPD). The EPD is an easily calculable measure of diversity whose numerical values carry more intuition than entropy. The EPD of a constrained library can be calculated exactly as

$$EPD(\phi) = L - \sum_{j=1}^{7}\sum_{k=1}^{21}\left(\tilde{q}_{\tilde{\phi}_j}(k)\right)^2$$

where $\tilde{q}_{\tilde{\phi}}$ are the amino acid probabilities at 7 positions corresponding to q_ϕ, the nucleotide probabilities for 21 positions, which can be calculated by summing over the probabilities of codons (**Supplementary Information**). Qualitatively, EPD is correlated with diversity; the probabilities of each amino acid position for three designed libraries (**Figure 3B-D**) show that, as the EPD increases, the probability mass is spread out over more sequences.

Consent statement UCSF

De-identified tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations. Sample use was approved by the Institutional Review Board at UCSF and experiments conform to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.

Primary Human Adult Brain Slices Culture and Library Infection

Adult surgical specimens from epilepsy cases were obtained from the UCSF medical center in collaboration with neurosurgeons with previous patient consent. Surgically excised specimens were immediately placed in a sterile container filled with N-methyl-D-glucamine (NMDG) substituted artificial cerebrospinal fluid (aCSF) of the following composition (in mM): 92 NMDG, 2.5 KCl, 1.25 NaH_2PO_4, 30 $NaHCO_3$, 20 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 25 glucose, 2 thiourea, 5 Na-ascorbate, 3 Na-pyruvate, 0.5 $CaCl_2 \cdot 4H_2O$ and 10 $MgSO_4 \cdot 7H_2O$. The pH of the NMDG aCSF was titrated pH to 7.3–7.4 with 1M Tris-Base at pH8, and the osmolality was 300–305 mOsmoles/Kg. The solution was pre-chilled to 2–4 °C and thoroughly bubbled with carbogen (95% O_2/5% CO_2) gas prior to collection. The tissue was transported from the operating room to the laboratory for processing within 40–60 min. Blood vessels and meninges were removed from the cortical tissue, and then the tissue block was secured for cutting using superglue and sectioned perpendicular to the cortical plate to 300 µm using a Leica VT1200S vibrating blade microtome in aCSF. The slices were then transferred into a container of sterile-filtered NMDG aCSF that was pre-warmed to 32–34 °C and continuously bubbled with carbogen gas. After 12 min recovery incubation, slices were transferred to slice culture inserts (Millicell, PICM03050) on six-well culture plates (Corning) and cultured in adult brain slice culture medium containing 840 mg MEM Eagle medium with Hanks salts and 2mM L-glutamine (Sigma, M4642), 18 mg ascorbic acid (Sigma, A7506), 3 mL HEPES (1M stock) (Sigma, H3537), 1.68 mL $NaHCO_3$ (892.75 mM solution, Gibco, 25080-094), 1.126 mL D-glucose, (1.11M solution, Gibco, A24940-01), 0.5 mL penicillin/streptomycin, 0.25 mL GlutaMax (at 400x, Gibco, 35050-061), 100 µL 2M stock $MgSO_4.7H_2O$ (Sigma, M1880), 50 µL 2M stock $CaCl_2.2H_2O$ (Sigma, C7902), 50 µL insulin from bovine pancreas, (10 mg/mL, Sigma, I0516), 20 mL horse serum-heat inactivated, 95 mL MilliQ H_2O (as previously described [37]). The following day after plating, adult human brain slices were infected with the viral library at an estimated of 10,000 MOI (N=3 per group) based on the number of cells estimated per slice. Slices were cultured at the liquid–air interface created by the cell-culture insert in a 37 ºC incubator at 5% CO_2 for 72 hours post infection.

Slice Culture Dissociation, Cell Purification and Hirt Extraction

Seventy-two hours after infection with the viral library, cultured brain tissue slices were first rinsed with DPBS (Gibco, 14190250) twice and detached from the filters. Then mechanically minced to 1mm^2 pieces and enzymatically digested with papain digestion kit (Worthington, LK003163) with the addition of DNase for 1 hr at 37ºC. After the enzymatic digestion, tissue was mechanically triturated using fire-polished glass pipettes (Fisher Scientific, cat#13-678-6A), filtered through a 40 µm cell strainer (Corning 352340), pelleted at 300xg for 5 minutes and washed twice with DBPS. Following mechanical digestion, the slices were first treated with lysis buffer (10% SDS, 1M Tris-HCL, pH 7.4-8.0, and 0.5M EDTA, pH 8.0) with the addition of RNase A (Thermo Scientific, EN0531) for 60 min at 37 ºC and proteinase K for 3 hours at 55 ºC. The enzymatically digested tissue homogenate was then proceeded to the Hirt column protocol as previously published [38].

References:

1. Maheshri, N., et al., *Directed evolution of adeno-associated virus yields enhanced gene delivery vectors*. Nat Biotechnol, 2006. **24**(2): p. 198-204.
2. Dalkara, D., et al., *In vivo-directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous*. Sci Transl Med, 2013. **5**(189): p. 189ra76.
3. Tse, L.V., et al., *Structure-guided evolution of antigenically distinct adeno-associated virus variants for immune evasion*. Proc Natl Acad Sci U S A, 2017. **114**(24): p. E4812-e4821.
4. Bartel, M.A., J.R. Weinstein, and D.V. Schaffer, *Directed evolution of novel adeno-associated viruses for therapeutic gene delivery*. Gene Therapy, 2012. **19**(6): p. 694-700.
5. Jang, J.H., et al., *An evolved adeno-associated viral variant enhances gene delivery and gene targeting in neural stem cells*. Mol Ther, 2011. **19**(4): p. 667-75.
6. Choudhury, S.R., et al., *In Vivo Selection Yields AAV-B1 Capsid for Central Nervous System and Muscle Gene Therapy*. Mol Ther, 2016. **24**(7): p. 1247-57.
7. Koerber, J.T., J.H. Jang, and D.V. Schaffer, *DNA shuffling of adeno-associated virus yields functionally diverse viral progeny*. Mol Ther, 2008. **16**(10): p. 1703-9.
8. Deverman, B.E., et al., *Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain*. Nat Biotechnol, 2016. **34**(2): p. 204-9.
9. Ogden, P.J., et al., *Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design*. Science, 2019. **366**(6469): p. 1139-1143.
10. Santiago-Ortiz, J., et al., *AAV ancestral reconstruction library enables selection of broadly infectious viral variants*. Gene Ther, 2015. **22**(12): p. 934-46.
11. Ojala, D.S., et al., *In Vivo Selection of a Computationally Designed SCHEMA AAV Library Yields a Novel Variant for Infection of Adult Neural Stem Cells in the SVZ*. Mol Ther, 2018. **26**(1): p. 304-319.
12. Adachi, K., et al., *Drawing a high-resolution functional map of adeno-associated virus capsid by massively parallel sequencing*. Nature Communications, 2014. **5**(1): p. 3075.
13. Byrne, L.C., et al., *In vivo-directed evolution of adeno-associated virus in the primate retina*. JCI insight, 2020. **5**(10): p. e135112.
14. Araya, C.L., et al., *A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function*. Proceedings of the National Academy of Sciences, 2012. **109**(42): p. 16858-16863.
15. Poelwijk, F.J., M. Socolich, and R. Ranganathan, *Learning the pattern of epistasis linking genotype and phenotype in a protein*. Nature Communications, 2019. **10**(1): p. 4213.
16. Marques, A.D., et al., *Applying machine learning to predict viral assembly for adeno-associated virus capsid libraries*. Molecular Therapy - Methods & Clinical Development, 2021. **20**: p. 276-286.
17. Bryant, D.H., et al., *Deep diversification of an AAV capsid protein by machine learning*. Nature Biotechnology, 2021. **39**(6): p. 691-696.
18. Calcedo, R., et al., *Worldwide Epidemiology of Neutralizing Antibodies to Adeno-Associated Viruses*. The Journal of Infectious Diseases, 2009. **199**(3): p. 381-390.
19. Boutin, S., et al., *Prevalence of serum IgG and neutralizing factors against adeno-associated virus (AAV) types 1, 2, 5, 6, 8, and 9 in the healthy population: implications for gene therapy using AAV vectors*. Hum Gene Ther, 2010. **21**(6): p. 704-12.

20. Von Drygalski, A., et al., *Etranacogene dezaparvovec (AMT-061 phase 2b): normal/near normal FIX activity and bleed cessation in hemophilia B*. Blood Adv, 2019. **3**(21): p. 3241-3247.
21. Parker, A.S., K.E. Griswold, and C. Bailey-Kellogg, *Optimization of combinatorial mutagenesis*. J Comput Biol, 2011. **18**(11): p. 1743-56.
22. Verma, D., G. Grigoryan, and C. Bailey-Kellogg, *Pareto Optimization of Combinatorial Mutagenesis Libraries*. IEEE/ACM Trans Comput Biol Bioinform, 2019. **16**(4): p. 1143-1153.
23. Müller, O.J., et al., *Random peptide libraries displayed on adeno-associated virus to select for targeted gene therapy vectors*. Nat Biotechnol, 2003. **21**(9): p. 1040-6.
24. Perabo, L., et al., *In vitro selection of viral vectors with modified tropism: the adeno-associated virus display*. Mol Ther, 2003. **8**(1): p. 151-7.
25. Zolotukhin, S., et al., *Recombinant adeno-associated virus purification using novel methods improves infectious titer and yield*. Gene Therapy, 1999. **6**(6): p. 973-985.
26. Matuszewski, S., et al., *A Statistical Guide to the Design of Deep Mutational Scanning Experiments*. Genetics, 2016. **204**(1): p. 77-87.
27. Katz, D., et al., *Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies*. Biometrics, 1978. **34**(3): p. 469-474.
28. Brookes, D., H. Park, and J. Listgarten, *Conditioning by adaptive sampling for robust design*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 773--782.
29. Keener, R.W., *Theoretical Statistics: Topics for a Core Course*. 2010: Springer New York.
30. Abadi, M., et al., *TensorFlow: a system for large-scale machine learning*, in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*. 2016, USENIX Association: Savannah, GA, USA. p. 265–283.
31. Kingma, D.P. and J. Ba, *Adam: A Method for Stochastic Optimization*. 2015.
32. MacKay, D.J.C., *Information Theory, Inference & Learning Algorithms*. 2002: Cambridge University Press.
33. Jaynes, E.T., *Information Theory and Statistical Mechanics*. Physical Review, 1957. **106**(4): p. 620-630.
34. Clara Fannjiang, J.L., *Autofocused oracles for model-based design* Advances in Neural Information Processing Systems 2020. **33**
35. Murphy, K.P., *Machine learning : a probabilistic perspective*. 2012.
36. Kleijnen, J. and R.Y. Rubinstein, *Optimization and sensitivity analysis of computer simulation models by the score function method*. 1995, Tilburg University, Center for Economic Research.
37. Ting, J.T., et al., *A robust ex vivo experimental platform for molecular-genetic dissection of adult human neocortical cell types and circuits*. Sci Rep, 2018. **8**(1): p. 8407.
38. Arad, U., *Modified Hirt procedure for rapid purification of extrachromosomal DNA from mammalian cells*. Biotechniques, 1998. **24**(5): p. 760-2.

Supplementary data

Table S1. Primer sequences for PCR reactions.

Table S2. Marginal probabilities of library D2 and D3 nucleotides at 21-bp position chart.

Figure S1: Comparison of maximum entropy unconstrained and constrained libraries.

Table S1. Primer sequences for PCR reactions.

Primer	Sequence (5'-3')
Seq_F	GGTGGAGCATGAATTCTACGTC
Seq_R	GCTCTGGTTGTTGGTGGCC
7mer_F	GGCCACCAACAACCAGAGCACCGGTNNKNNKNNKNNKNN KNNKNNKGGCTTAAGTTCCACCACTGCCC
7mer_R	GCTCTGGTTGTTGGTGGCC
Vg_F	GCGGAAGCTTCGATCAACTACG
Vg_R	CGCAGAGACCAAAGTTCAACTGA
HindIII_F	TTCCACGTCTTTATATGGTGCCCAGTC
NotI_R	CGCAGAGACCAAAGTTCAACTGA

Table S2. Marginal probabilities of library D2 and D3 nucleotides at 21-bp position chart.

Library D2	A	T	C	G	**Library D3**	A	T	C	G
1	0.12	0.04	0.39	0.45		0.21	0.09	0.43	0.27
2	0.18	0.47	0.3	0.05		0.22	0.25	0.37	0.16
3	0.21	0.19	0.28	0.32		0.25	0.28	0.27	0.2
4	0.14	0.02	0.19	0.65		0.27	0.12	0.13	0.48
5	0.23	0.33	0.29	0.15		0.2	0.35	0.27	0.18
6	0.28	0.24	0.25	0.23		0.22	0.23	0.22	0.33
7	0.35	0	0.14	0.51		0.22	0.08	0.16	0.54
8	0.13	0.17	0.36	0.34		0.32	0.19	0.34	0.15
9	0.21	0.31	0.31	0.17		0.25	0.17	0.27	0.31
10	0.13	0	0.06	0.81		0.24	0.07	0.43	0.26
11	0.26	0.12	0.22	0.4		0.34	0.35	0.2	0.11
12	0.16	0.29	0.36	0.19		0.22	0.26	0.28	0.24
13	0.09	0	0.08	0.83		0.28	0.06	0.17	0.49
14	0.36	0.12	0.37	0.15		0.45	0.14	0.29	0.12
15	0.13	0.49	0.24	0.14		0.2	0.27	0.3	0.23
16	0.22	0	0.13	0.65		0.32	0.08	0.19	0.41
17	0.29	0.08	0.24	0.14		0.23	0.2	0.35	0.22
18	0.1	0.42	0.34	0.14		0.23	0.27	0.36	0.14
19	0.16	0.01	0.09	0.74		0.2	0.04	0.32	0.44
20	0.28	0.11	0.47	0.14		0.39	0.17	0.3	0.14
21	0.17	0.35	0.3	0.18		0.26	0.25	0.24	0.25
Total reads assessed by deep sequencing: 193228 (library D2) and 212388 (library D3)									

Supplementary Methods

Maximum entropy design of unconstrained libraries

In the main text, we consider constrained library designs, where one specifies experimental control knobs, such as the marginal probabilities of observing each amino acid at each position. Contrasting the constrained libraries, are unconstrained ones, where one constructs a list of oligonucleotide sequences that comprise the library. Unconstrained libraries provide more control over the contents of the library than constrained libraries but are substantially more expensive per oligonucleotide (each of which must be synthesized). Therefore, in considering constrained *vs* unconstrained libraries, one is trading off control for library size. Note that technically, a fully unconstrained library is the probability distribution itself, $p_\lambda(x)$, and that in drawing samples from such a distribution, the resulting library becomes an approximation to the unconstrained library in the sense of having only finitely many samples.

Although we did not experimentally realize any constrained libraries in this work, here we demonstrate that it is possible to apply our maximum entropy formulation to the design of unconstrained libraries. It is conceptually straightforward to build a list of sequences that approximates the maximum entropy library of **Equation 4** by sampling from this distribution with, for instance, Markov Chain Monte Carlo (MCMC) algorithms. In particular, letting, $f(x)$ be the same predictive model used to design the constrained libraries of Figure 3, we used the Metropolis-Hastings algorithm with **Equation 4** as the stationary distribution from which to sample sequences. We allowed a short burn-in period of T_{burn} iterations of the algorithm, after which we assumed the algorithm was equilibrated and used the T subsequent iterations as the specified sequences. This set of samples represents a particle-based approximation to Equation 4 and thus will approximately respect the Pareto optimal property of the maximum entropy library.

The results of applying this scheme with $T_{burn} = 1000$ and $T = 10{,}000$, for 404 settings of λ are shown in Figure S1, below. We can see that unconstrained library construction allows one to build a library with higher expected predicted fitness at the same level of diversity of constrained libraries. As oligonucleotide synthesis becomes cheaper, unconstrained library synthesis will became correspondingly cheaper. Therefore, our results suggest that at some point, it is likely that unconstrained libraries may become the libraries of choice.

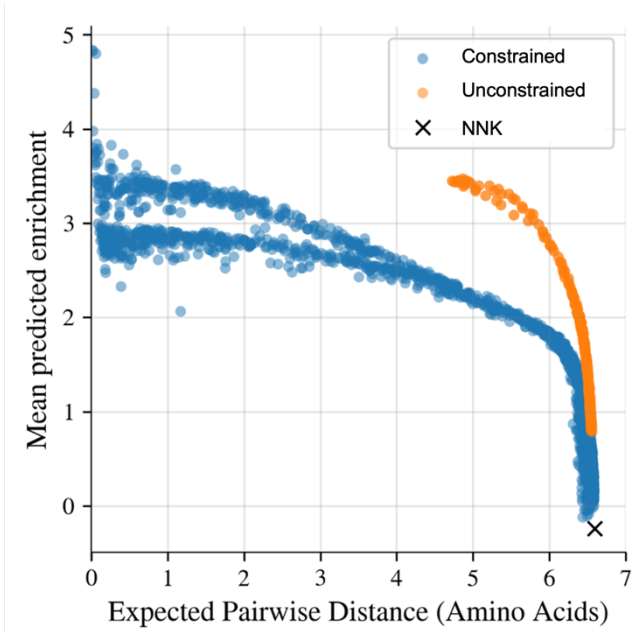


Figure S1: Comparison of maximum entropy unconstrained and constrained libraries. Blue points are identical to the points in Figure 3a, with the colors removed. Orange points represent unconstrained libraries constructed using an MCMC algorithm to sample from the distribution of Equation 4 at a range of settings of, with the same as used to construct the constrained libraries.

Gradients for maximum entropy constrained library

To solve the non-convex objective (**Equation 6**) for the library parameters, ϕ, we use the Stochastic Gradient Descent (SGD) algorithm, which requires computing the gradient

$$\nabla_\phi \left[\mathbb{E}_{q_\phi(x)}[f(x)] + \lambda H[q_\phi] \right].$$

The gradient of the entropy is given by

$$\begin{aligned}
\nabla_\phi H[q_\phi] &= -\nabla_\phi \mathbb{E}_{q_\phi(x)}[\log q_\phi(x)] \\
&= -\sum_{x\in\chi} \nabla_\phi [q_\phi(x) \log q_\phi(x)] \\
&= -\sum_{x\in\chi} \left(\log q_\phi(x)\, \nabla_\phi q_\phi(x) + q_\phi(x) \nabla_\phi \log q_\phi(x)\right) \\
&= -\sum_{x\in\chi} \left(\log q_\phi(x)\, q_\phi(x) \nabla_\phi \log q_\phi(x) + q_\phi(x) \nabla_\phi \log q_\phi(x)\right) \\
&= -\sum_{x\in\chi} q_\phi(x) \nabla_\phi \log q_\phi(x) \left(1 + \log q_\phi(x)\right) \\
&= -\mathbb{E}_{q_\phi(x)}\left[\left(1 + \log q_\phi(x)\right) \nabla_\phi \log q_\phi(x)\right]
\end{aligned}$$

where in the third line we used the equality $\nabla_\phi q_\phi(x) = q_\phi(x)\nabla_\phi \log q_\phi(x)$. For $\nabla_\phi \mathbb{E}_{q_\phi(x)}[f(x)]$, we use the equality $\nabla_\phi \mathbb{E}_{q_\phi(x)}[f(x)] = \mathbb{E}_{q_\phi(x)}\big[f(x)\nabla_\phi \log q_\phi(x)\big]$ which is well-known from its use in the score function estimator [36] (sometimes also called the 'log derivative trick'). We then have

$$\begin{aligned}\nabla_\phi\Big[\mathbb{E}_{q_\phi(x)}[f(x)] + \lambda H\big[q_\phi\big]\Big] &= \nabla_\phi \mathbb{E}_{q_\phi(x)}[f(x)] + \lambda \nabla_\phi H\big[q_\phi\big] \\ &= \mathbb{E}_{q_\phi(x)}\big[f(x)\nabla_\phi \log q_\phi(x)\big] - \lambda \mathbb{E}_{q_\phi(x)}\big[\big(1+\log q_\phi(x)\big)\nabla_\phi \log q_\phi(x)\big] \\ &= \mathbb{E}_{q_\phi(x)}\big[w(x)\nabla_\phi \log q_\phi(x)\big] \qquad \text{(Eqn. S1)}\end{aligned}$$

where $w(x) \coloneqq f(x) - \lambda\big(1+\log q_\phi(x)\big)$. The individual components of $\nabla_\phi \log q_\phi(x)$ are given by

$$\begin{aligned}\frac{\partial}{\partial \phi_{jk}} \log q_\phi(x) &= \frac{\partial}{\partial \phi_{jk}} \log q_{\phi_j}\big(x^j\big) \\ &= \frac{\partial}{\partial \phi_{jk}} \log \frac{e^{\phi_{j,x^j}}}{\sum_{l=1}^K e^{\phi_{jl}}} \\ &= \frac{\partial}{\partial \phi_{jk}} \phi_{j,x^j} - \frac{\partial}{\partial \phi_{jk}} \log \sum_{l=1}^K e^{\phi_{jl}} \\ &= \delta_k\big(x^j\big) - \frac{1}{\sum_{l=1}^K e^{\phi_{jl}}} \frac{\partial}{\partial \phi_{jk}} \sum_{l=1}^K e^{\phi_{jl}} \\ &= \delta_k\big(x^j\big) - \frac{e^{\phi_{jk}}}{\sum_{l=1}^K e^{\phi_{jl}}} \\ &= \delta_k\big(x^j\big) - q_{\phi_j}(k) \qquad \text{(Eqn. S2)}\end{aligned}$$

Using **Equation S2** within **Equation S1** gives **Equation 8**.

<u>Expected pairwise distance</u>

Here we derive the Expected Pairwise Distance (EPD) between pairs of sequences sampled from a constrained library design. Consider a constrained library, $q_\phi(x)$, for sequences of length L and alphabet size K. The Hamming distance between two sequences, x_1 and x_2, is $d(x_1, x_2) = L - \sum_{i=1}^L \delta\big(x_1^j, x_2^j\big)$, where $\delta\big(x_1^j, x_2^j\big)$ is equal to one if $x_1^j = x_2^j$ and zero otherwise.

The expected distance between two sequences samples from $q_\phi(x)$ is then:

$$\begin{aligned}EPD(\phi) &= \mathbb{E}_{x_1\sim q_\phi(x), x_2\sim q_\phi(x)}[d(x_1,x_2)] \\ &= L - \mathbb{E}_{x_1\sim q_\phi(x), x_2\sim q_\phi(x)}\left[\sum_{i=1}^L \delta\big(x_1^i, x_2^i\big)\right] \\ &= L - \sum_{i=1}^L \mathbb{E}_{x_1^i\sim q_{\phi_i}(x^i), x_2^i\sim q_{\phi_i}(x^i)}\big[\delta\big(x_1^i, x_2^i\big)\big] \\ &= L - \sum_{i=1}^L \sum_{j=1}^K \sum_{l=1}^K q_{\phi_i}(j) q_{\phi_i}(l)\delta(j,l) \\ &= L - \sum_{i=1}^L \sum_{j=1}^K q_{\phi_i}(j)^2 .\end{aligned}$$