nature portfolio

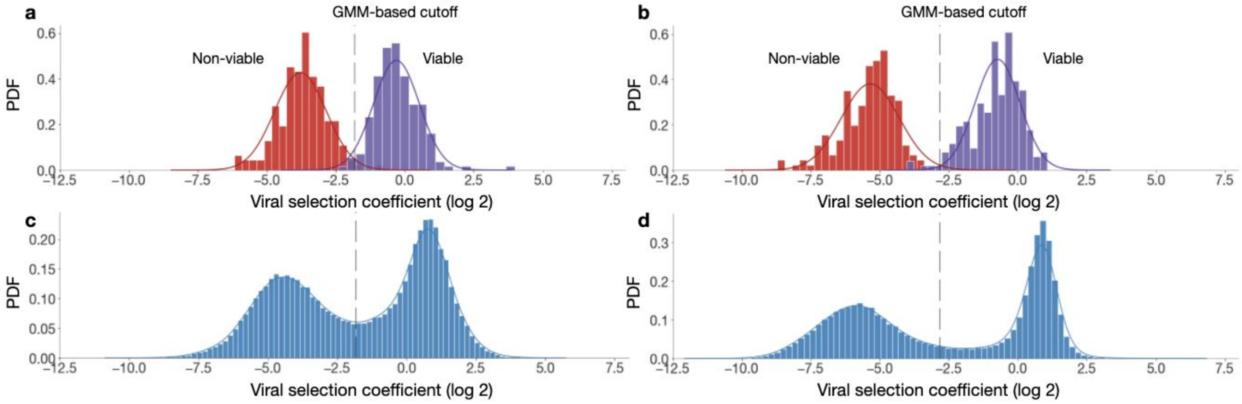
https://doi.org/10.1038/s41587-020-00793-4

Supplementary information

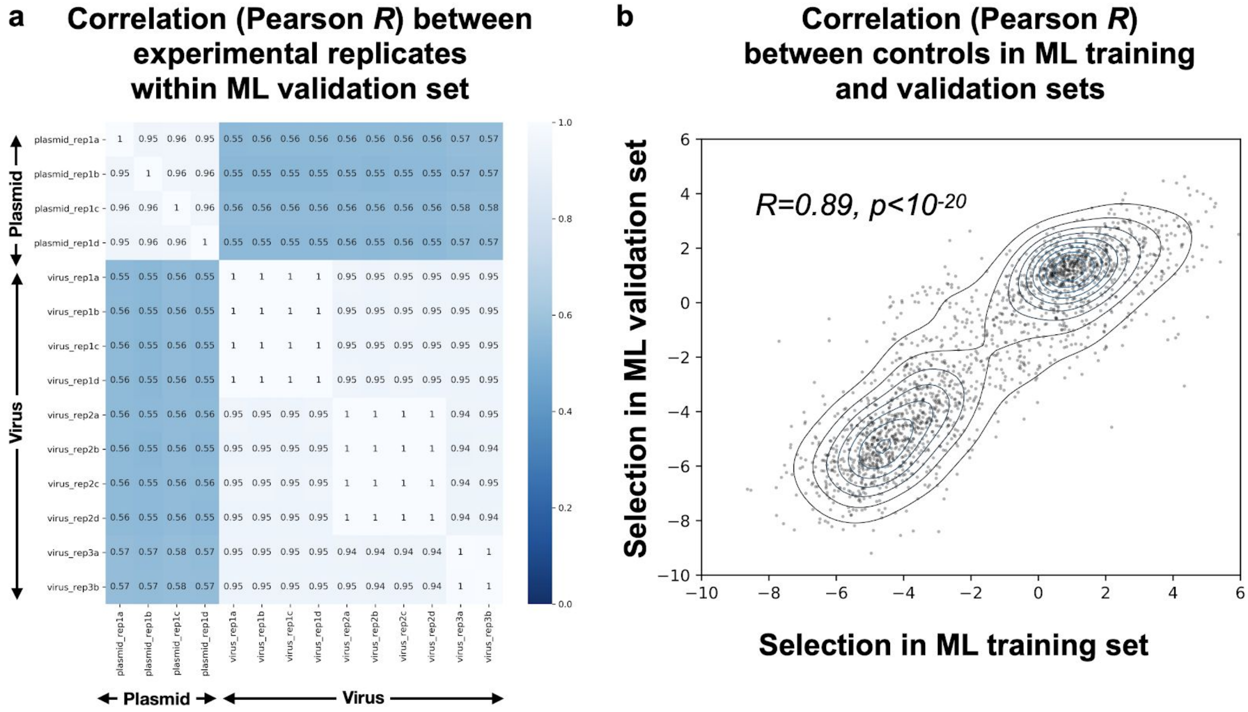
Deep diversification of an AAV capsid protein by machine learning

In the format provided by the authors and unedited

Supplementary Information

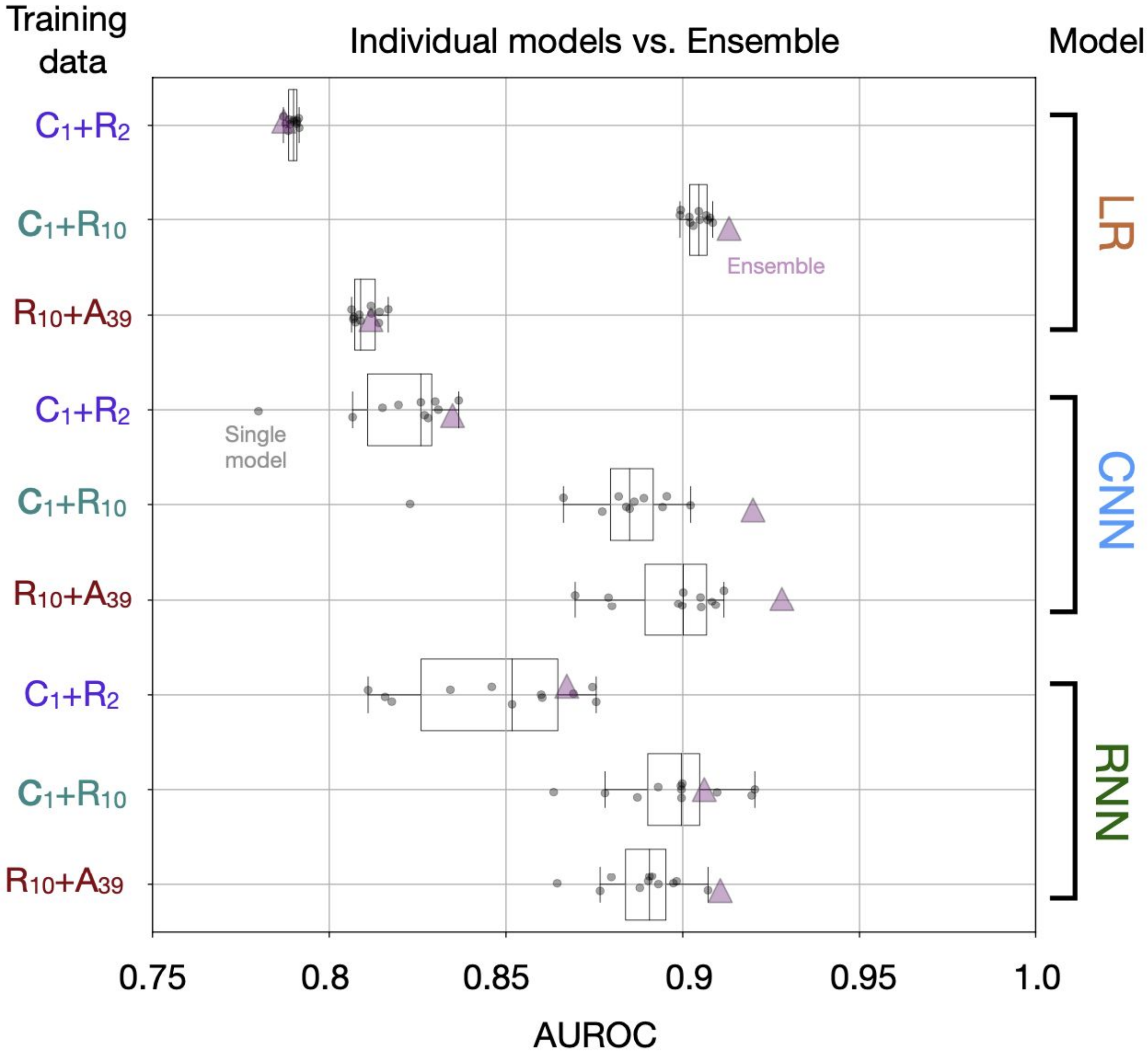


Supplementary Figure 1 | Bimodal packaging viral selection coefficient distribution. a, Viral selection coefficients for 168 sequences known to produce successfully (WT AAV2 alternate codon variants) and 162 sequences known to fail at production (capsid variants truncated via stop codon insertions) from the initial experiment. The viral selection threshold for the viable/non-viable classes was determined by fitting the 2-component GMM shown (red and purple lines), on a log2 scale. **b**, Viral selection coefficients for 200 sequences known to produce successfully (WT AAV2 alternate codon variants) and 171 sequences known to fail at production (capsid variants truncated via stop codon insertions) from the final experiment. **c, d** Distributions of all >70k variants from the initial experiment and all >240k variants from the final experiment are bimodal, motivating our use of categorical prediction models. These distributions illustrate the binary nature of the packaging assay outcomes and the intrinsic measurement variance associated with the assay.

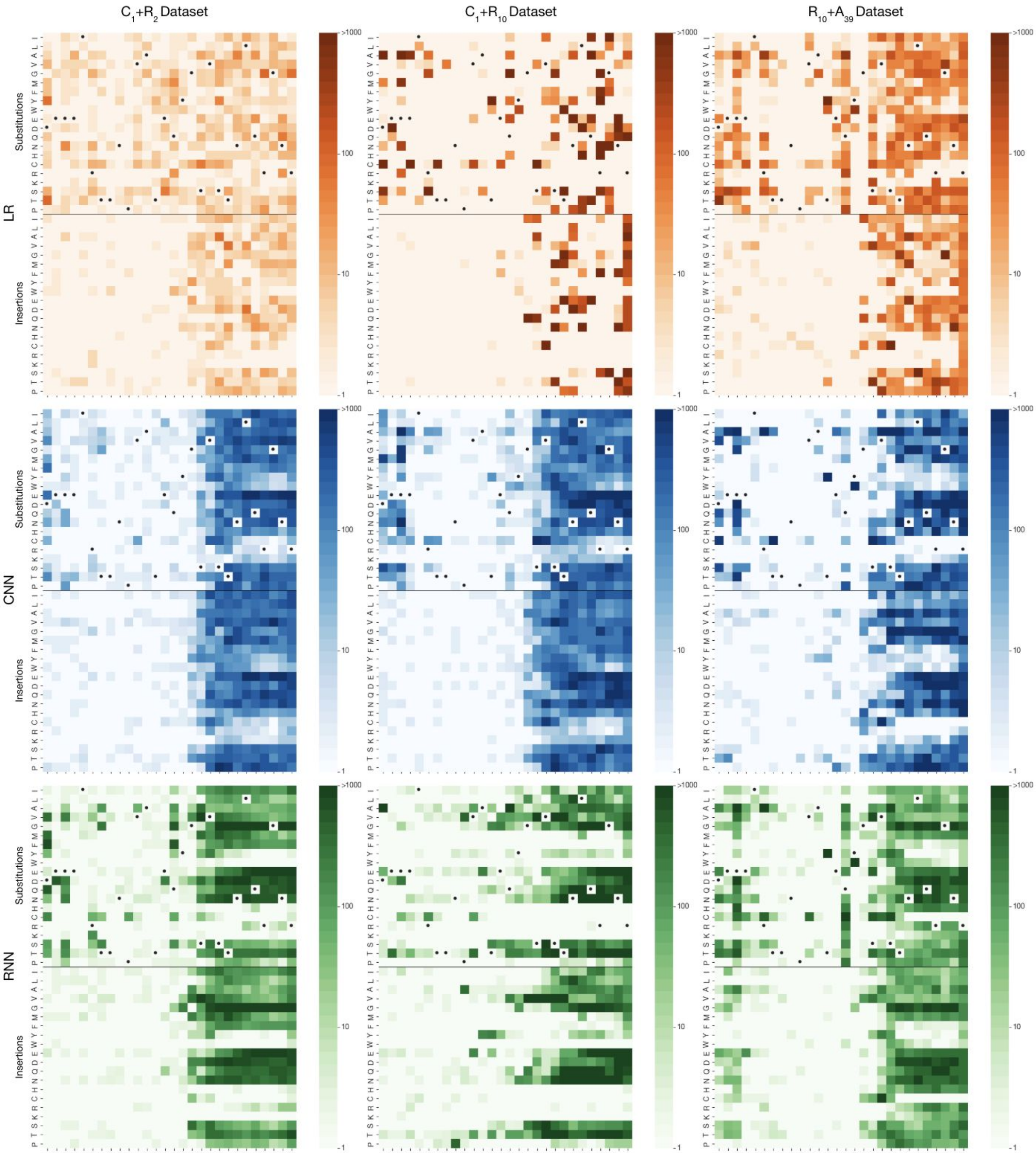


	plasmid_rep1a	plasmid_rep1b	plasmid_rep1c	plasmid_rep1d	virus_rep1a	virus_rep1b	virus_rep1c	virus_rep1d	virus_rep2a	virus_rep2b	virus_rep2c	virus_rep2d	virus_rep3a	virus_rep3b
plasmid_rep1a	1	0.95	0.96	0.95	0.55	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.57	0.57
plasmid_rep1b	0.95	1	0.96	0.96	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.57	0.57
plasmid_rep1c	0.96	0.96	1	0.96	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.58	0.58
plasmid_rep1d	0.95	0.96	0.96	1	0.55	0.55	0.55	0.55	0.56	0.55	0.56	0.55	0.57	0.57
virus_rep1a	0.55	0.55	0.56	0.55	1	1	1	1	0.95	0.95	0.95	0.95	0.95	0.95
virus_rep1b	0.56	0.55	0.56	0.55	1	1	1	1	0.95	0.95	0.95	0.95	0.95	0.95
virus_rep1c	0.56	0.55	0.56	0.55	1	1	1	1	0.95	0.95	0.95	0.95	0.95	0.95
virus_rep1d	0.56	0.55	0.56	0.55	1	1	1	1	0.95	0.95	0.95	0.95	0.95	0.95
virus_rep2a	0.56	0.55	0.56	0.56	0.95	0.95	0.95	0.95	1	1	1	1	0.94	0.95
virus_rep2b	0.56	0.55	0.56	0.55	0.95	0.95	0.95	0.95	1	1	1	1	0.94	0.94
virus_rep2c	0.56	0.55	0.56	0.56	0.95	0.95	0.95	0.95	1	1	1	1	0.94	0.95
virus_rep2d	0.56	0.55	0.56	0.55	0.95	0.95	0.95	0.95	1	1	1	1	0.94	0.94
virus_rep3a	0.57	0.57	0.58	0.57	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	1	1
virus_rep3b	0.57	0.57	0.58	0.57	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.94	1	1

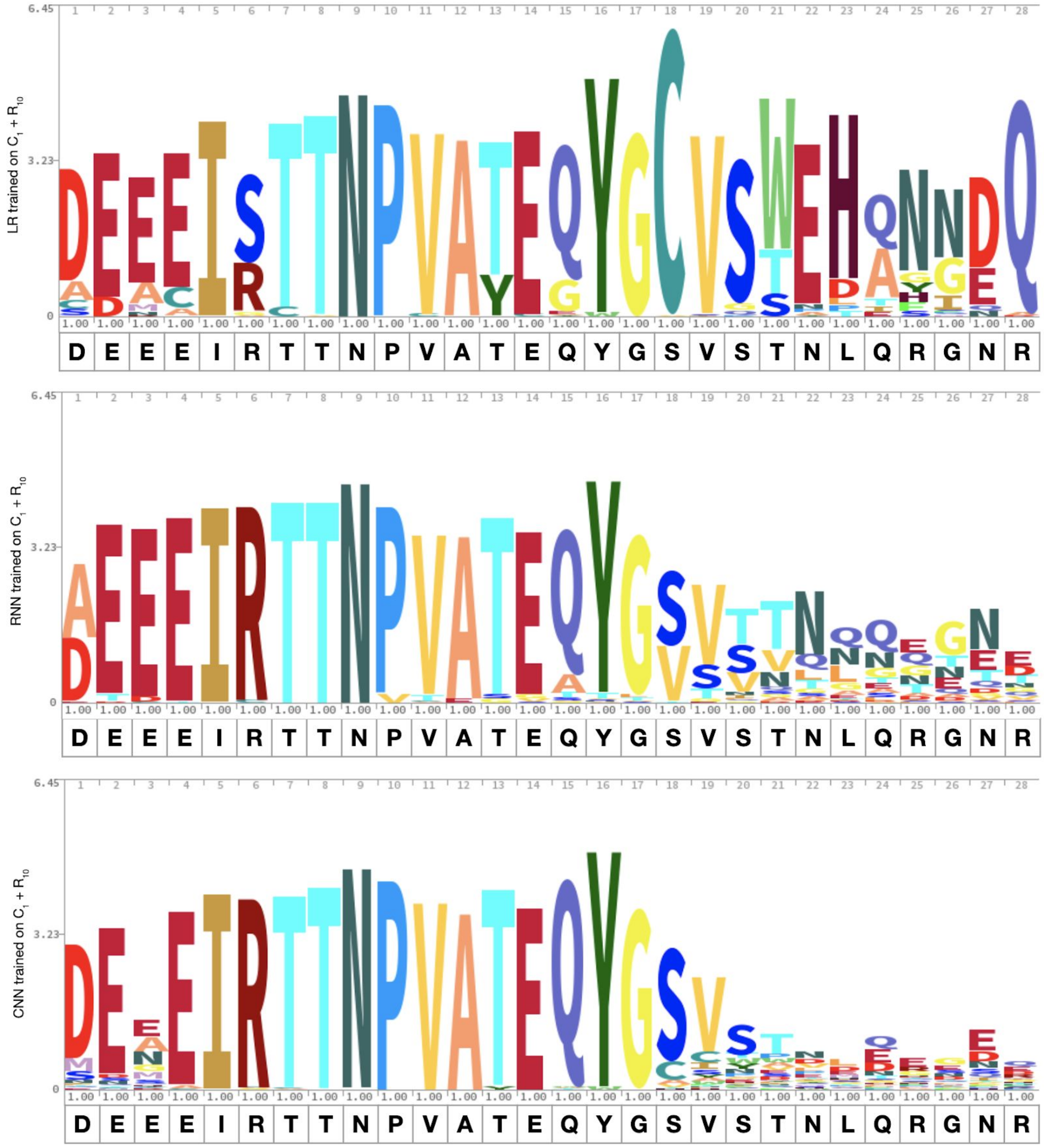
Supplementary Figure 2 | Reproducibility within and across experiments, a, Pearson correlation between plasmid replicates and virus replicates in ML validation set (243,481 DNA-level variants). For each of the four transfection replicates for virus production (numbered), we have at least two PCR replicates (denoted by letters). **b**, We remeasured fitness for n = 2000 sequence variants with a range of selection scores from our ML training data as a control on the validation chip as designed by the classifiers, to calibrate our comparison with the additive model and ensure reproducibility of results. The p-value is calculated using a two-sided *t*-test with n-2 degrees of freedom.



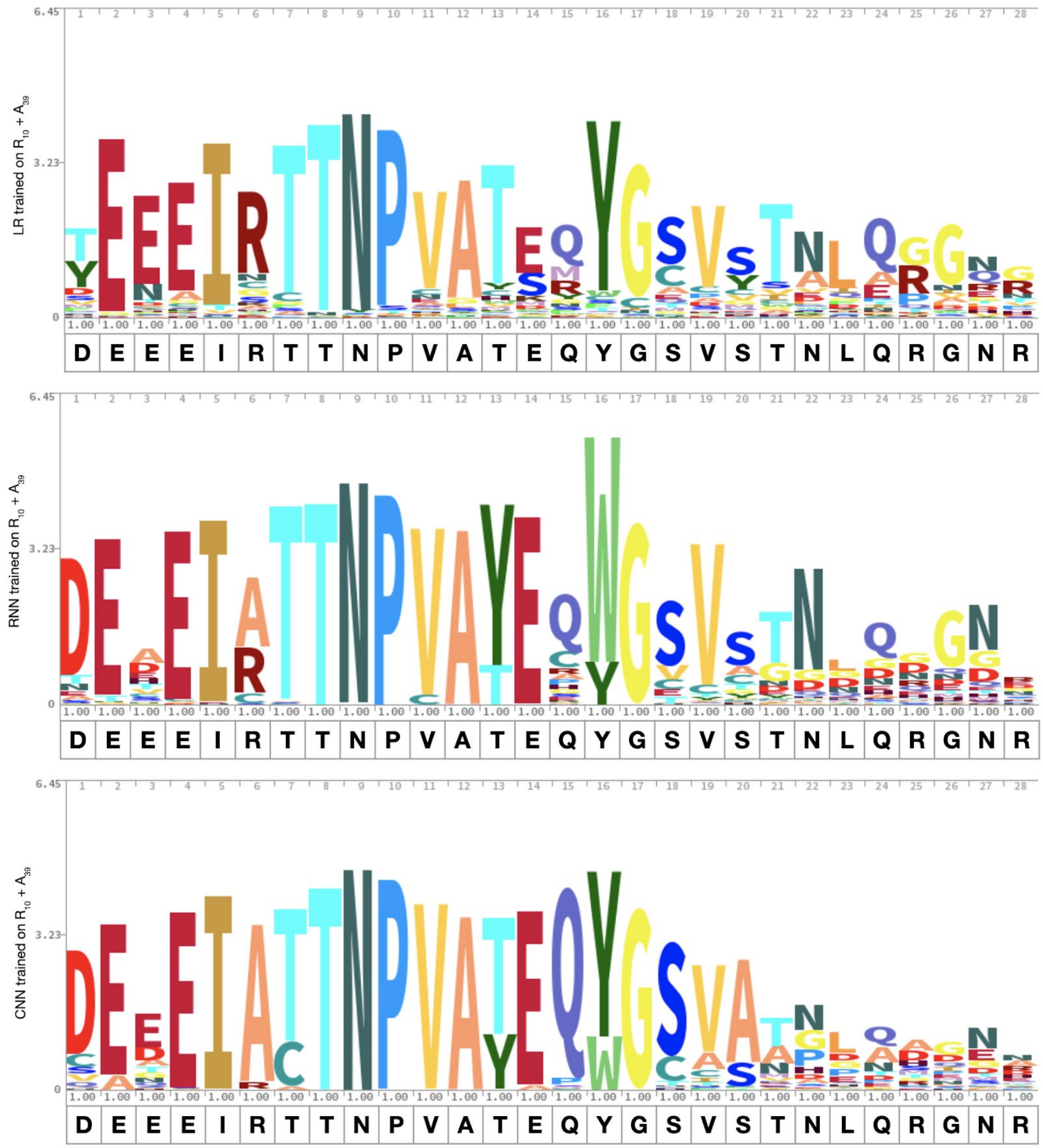
Supplementary Figure 3 | Comparison of individual and ensemble model performance. Evaluation of the performance of both the single (black dots) and ensemble (pink triangles) models built for each architecture/training set combination using the area under the receiver operating characteristic (AUROC) for all model generated sequences. For the ensemble, we average the scores of each eleven individual models before computing the AUROC. Overall we find that the ensembles consistently outperform the median performance of individual models, in some cases outperforming the best individual model as well. Note that logistic regression replicate models tend to display highly similar performance regardless of initialization, while the effects of random initializations can be quite significant for the neural networks. As a result, the performance gain due to ensembling is particularly notable for the neural network models.



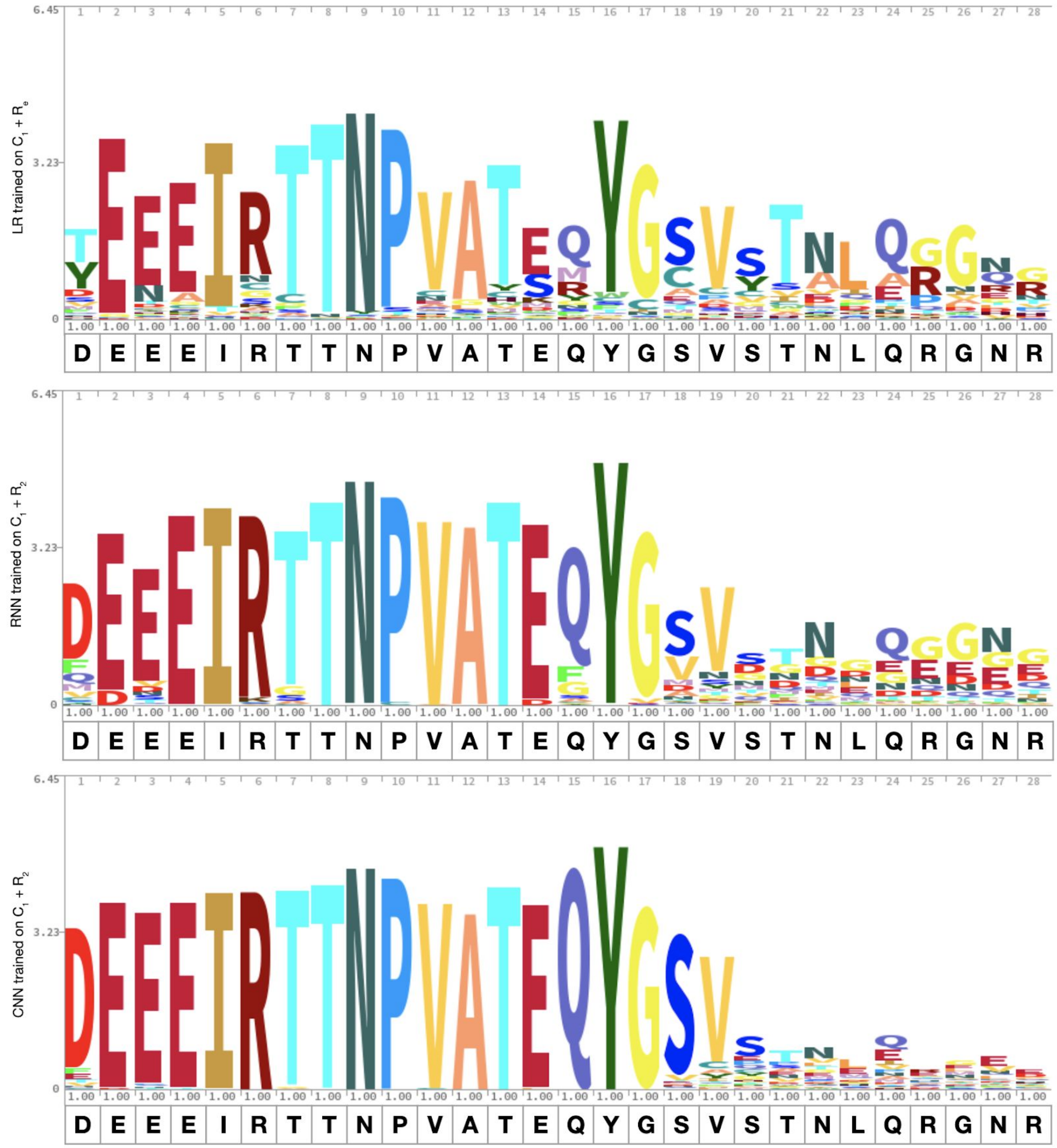
Supplementary Figure 4 | (a) Mutation preference distribution for all ML models. Heatmaps showing counts of substitutions (top) and insertions (bottom) within viable mutant capsids with ≥12 mutations as designed by each model architecture (LR, CNN, RNN), trained on each dataset.



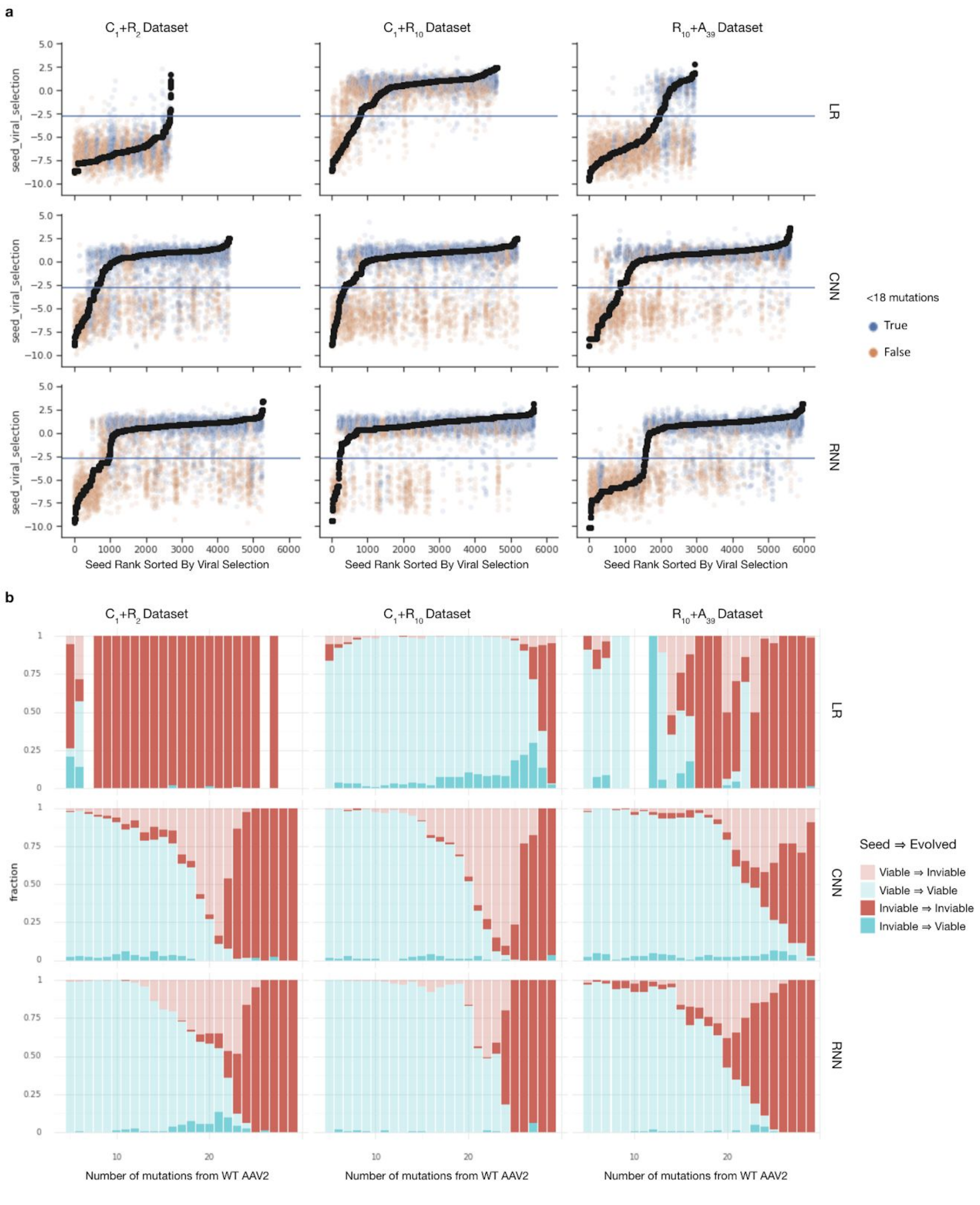
Supplementary Figure 4 | (b) Logos showing viable model-designed sequences for ML models trained on the C_1 + R_{10} dataset. Sequence logos showing amino acid usage within viable mutant capsids with ≥12 mutations from the AAV2 wildtype sequence as designed by each model architecture (LR, CNN, RNN). The wildtype AAV2 sequence is shown in black below each logo.



Supplementary Figure 4 | (c) Logos showing viable model-designed sequences for ML models trained on the R_{10} + A_{39} dataset. Sequence logos showing amino acid usage within viable mutant capsids with ≥12 mutations from the AAV2 wildtype sequence as designed by each model architecture (LR, CNN, RNN). The wildtype AAV2 sequence is shown in black below each logo.



Supplementary Figure 4 | (d) Logos showing viable model-designed sequences for ML models trained on the C_1 + R_2 dataset. Sequence logos showing amino acid usage within viable mutant capsids with ≥12 mutations from the AAV2 wildtype sequence as designed by each model architecture (LR, CNN, RNN). The wildtype AAV2 sequence is shown in black below each logo.



Supplementary Figure 5 | Relationship between model-designed sequences and their model-selected starting seeds. a, The set of model-designed sequences with experimentally tested seeds are shown within each facet. Model-designed sequences for a particular seed are rendered at the same x-axis position and colored by whether they were <18 (blue) or >=18 (orange) mutations from wildtype. The seeds are sorted by their viral selection value (y-axis). The horizontal blue line corresponds to the viability cutoff. Most models show a strong preference for viable model-designed sequences from viable seeds. **b,** The relative fraction of viable (blue) and non-viable (red) model-designed sequences that came from viable seeds (dark alpha) and non-viable seeds (light alpha). Most models start from viable seeds and identify viable children close to WT. Far from WT, models become less reliable and more likely to start from non-viable seeds.that came from viable seeds (dark alpha) and non-viable seeds (light alpha). Most models start from viable seeds and identify viable children close to WT. Far from WT, models become less reliable and more likely to start from non-viable seeds.

Supplementary Table 1 | ML-generated AAV2 capsid statistics by mutation count. Cumulative viable capsid generation statistics across all machine learning models (LR, CNN and RNN), including both model-designed and model-selected sequences across all training datasets (C_1+R_2, C_1+R_{10}, and R_{10}+A_{39}), for a range of mutations-from-WT thresholds. The bolded row corresponds to the mutation distance at which the models first exceed the additive model in % viable capsids.

Min Mutations Threshold	# Generated Capsids	# Viable Capsids	% Viable Capsids	%Viable Capsides (Additive Model)
2	201,426	110,689	55.00%	62.50%
3	201,426	110,689	55.00%	59.50%
4	**201,424**	**110,687**	**55.00%**	**53.70%**
5	201,368	110,633	54.90%	46.10%
6	193,413	103,403	53.50%	36.40%
7	184,424	95,422	51.70%	21.30%
8	175,443	87,571	49.90%	17.30%
9	166,361	79,628	47.90%	13.70%
10	157,294	72,180	45.90%	10.70%
11	148,167	64,678	43.70%	8.30%
12	138,815	57,348	41.30%	6.30%
13	129,433	50,330	38.90%	4.70%
14	119,469	43,236	36.20%	3.50%
15	109,474	36,173	33.00%	2.40%
16	99,137	29,326	29.60%	1.60%
17	88,694	22,901	25.80%	1.00%
18	78,951	17,588	22.30%	0.60%
19	69,612	13,233	19.00%	0.40%
20	60,049	9,710	16.20%	0.30%
21	51,164	7,048	13.80%	0.10%
22	42,202	4,952	11.70%	0.00%
23	33,500	3,301	9.90%	0.00%
24	24,879	1,983	8.00%	0.00%
25	16,977	1,038	6.10%	0.00%
26	11,089	484	4.40%	0.00%
27	7,350	196	2.70%	0.00%
28	4,094	52	1.30%	0.00%
29	1,489	10	0.70%	0.00%

Supplementary Table 2 | ML-designed AAV2 capsid statistics by mutation count. Cumulative viable capsid generation statistics across all machine learning models (LR, CNN and RNN), for only model-designed sequences (i.e., excludes model-selected) across all training datasets (C_1+R_2, C_1+R_{10}, and R_{10}+A_{39}). The bolded row corresponds to the mutation distance at which the models first exceed the additive model in % viable capsids.

Min Mutations Threshold	# Generated Capsids	# Viable Capsids	% Viable Capsids	%Viable Capsides (Additive Model)
2	183,466	106,665	58.10%	62.50%
3	183,466	106,665	58.10%	59.50%
4	**183,464**	**106,663**	**58.10%**	**53.70%**
5	183,411	106,612	58.10%	46.10%
6	176,351	100,150	56.80%	36.40%
7	168,231	92,923	55.20%	21.30%
8	160,096	85,766	53.60%	17.30%
9	151,805	78,411	51.70%	13.70%
10	143,464	71,416	49.80%	10.70%
11	135,099	64,267	47.60%	8.30%
12	126,589	57,157	45.20%	6.30%
13	118,046	50,243	42.60%	4.70%
14	108,965	43,188	39.60%	3.50%
15	99,868	36,138	36.20%	2.40%
16	90,448	29,299	32.40%	1.60%
17	80,932	22,879	28.30%	1.00%
18	72,082	17,571	24.40%	0.60%
19	63,657	13,217	20.80%	0.40%
20	55,026	9,698	17.60%	0.30%
21	47,032	7,039	15.00%	0.10%
22	38,986	4,946	12.70%	0.00%
23	31,190	3,297	10.60%	0.00%
24	23,441	1,980	8.40%	0.00%
25	16,395	1,037	6.30%	0.00%
26	11,089	484	4.40%	0.00%
27	7,350	196	2.70%	0.00%
28	4,094	52	1.30%	0.00%
29	1,489	10	0.70%	0.00%

Supplementary Table 3 | NN-designed AAV2 capsid statistics by mutation count. Cumulative viable capsid generation statistics across all neural network models (CNN and RNN) for only model-designed sequences across all training datasets (C_1+R_2, C_1+R_{10}, and R_{10}+A_{39}) for a range of mutations-from-WT thresholds (i.e., excludes model-selected sequences). The bolded row corresponds to the mutation distance at which the models first exceed the additive model in % viable capsids.

Min Mutations Threshold	# Generated Capsids	# Viable Capsids	% Viable Capsids	%Viable Capsides (Additive Model)
2	**123,331**	**79,837**	**64.70%**	**62.50%**
3	123,331	79,837	64.70%	59.50%
4	123,329	79,835	64.70%	53.70%
5	123,280	79,788	64.70%	46.10%
6	117,855	74,431	63.20%	36.40%
7	112,376	69,020	61.40%	21.30%
8	106,907	63,624	59.50%	17.30%
9	101,326	58,145	57.40%	13.70%
10	95,698	52,658	55.00%	10.70%
11	90,035	47,192	52.40%	8.30%
12	84,291	41,688	49.50%	6.30%
13	78,449	36,219	46.20%	4.70%
14	72,332	30,635	42.40%	3.50%
15	65,960	24,953	37.80%	2.40%
16	59,277	19,247	32.50%	1.60%
17	52,702	13,997	26.60%	1.00%
18	46,774	9,856	21.10%	0.60%
19	41,028	6,559	16.00%	0.40%
20	35,300	4,092	11.60%	0.30%
21	30,053	2,482	8.30%	0.10%
22	24,771	1,385	5.60%	0.00%
23	19,712	670	3.40%	0.00%
24	15,145	338	2.20%	0.00%
25	10,854	165	1.50%	0.00%
26	7,306	72	1.00%	0.00%
27	4,887	47	1.00%	0.00%
28	2,666	15	0.60%	0.00%
29	942	4	0.40%	0.00%

Supplementary Table 4 | Model-selected AAV2 capsid statistics per ML model.

Model	# Generated Capsids	# Viable Capsids	% Viable Capsids
LR{C_1+R_2}	2,071	114	5.5%
LR{C_1+R_{10}}	1,989	486	24.4%
LR{R_{10}+A_{39}}	2,030	340	16.7%
CNN{C_1+R_2}	2,022	381	18.8%
CNN{C_1+R_{10}}	1,924	476	24.7%
CNN{R_{10}+A_{39}}	1,898	529	27.9%
RNN{C_1+R_2}	2,045	575	28.1%
RNN{C_1+R_{10}}	1,916	412	21.5%
RNN{R_{10}+A_{39}}	2,065	711	34.4%

Supplementary Table 5 | Model-designed AAV2 capsid statistics per ML model.

Model	# Generated Capsids	# Viable Capsids	% Viable Capsids
LR{C_1+R_2}	19,999	1,483	7.4%
LR{C_1+R_{10}}	20,456	19,211	93.9%
LR{R_{10}+A_{39}}	19,680	6,134	31.2%
CNN{C_1+R_2}	20,454	11,229	54.9%
CNN{C_1+R_{10}}	20,395	13,086	64.2%
CNN{R_{10}+A_{39}}	20,759	14,968	72.1%
RNN{C_1+R_2}	20,154	13,056	64.8%
RNN{C_1+R_{10}}	20,838	15,525	74.5%
RNN{R_{10}+A_{39}}	20,731	11,973	57.8%

Supplementary Table 6 | Additive model (A_{39}) capsid statistics. Cumulative across edit distance thresholds.

Min Mutations Threshold	# Generated Capsids	# Viable Capsids	% Viable Capsids
2	56,372	35,217	62.5%
3	50,572	30,068	59.5%
4	41,232	22,129	53.7%
5	31,561	14,551	46.1%
6	22,407	8,159	36.4%
7	13,892	2,953	21.3%
8	12,603	2,181	17.3%
9	11,387	1,561	13.7%
10	10,245	1,101	10.7%
11	9,171	757	8.3%
12	8,160	511	6.3%
13	7,195	340	4.7%
14	6,312	224	3.5%
15	5,495	134	2.4%
16	4,757	74	1.6%
17	4,102	42	1.0%
18	3,522	22	0.6%
19	2,994	13	0.4%
20	2,541	8	0.3%
21	2,148	2	0.1%
22	1,790	0	0.0%
...	...	...	...
37	30	0	0.0%
38	16	0	0.0%
39	3	0	0.0%

Supplementary Table 7 | Randomly generated (R_{10}) capsid statistics. Cumulative across edit distance thresholds.

Min Mutations Threshold	# Generated Capsids	# Viable Capsids	% Viable Capsids
2	9,885	964	9.80%
3	8,129	461	5.70%
4	6,378	213	3.30%
5	4,631	93	2.00%
6	2,883	32	1.10%
7	1,154	3	0.30%
8	866	2	0.20%
9	576	1	0.20%
10	284	1	0.40%