Designing Proteins using Sparse Data

Ada Shaw
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
ayshaw@g.harvard.edu

Jung-Eun Shin
Affiliation Seismic Therapeutics; Formerly Harvard University

Nicole Thadani
Affiliation Department of Systems Biology
Harvard Medical School
Boston, MA 02115

Alan Amin
Affiliation Department of Systems Biology
Harvard Medical School
Address Boston, MA 02115

Eli Weinstein
Columbia University
New York, NY 10027

Debora Susan Marks
Affiliation Department of Systems Biology
Harvard Medical School
Boston, MA, 02115
debbie@hms.harvard.edu

Abstract

A major goal in biotechnology is to generate libraries of functional proteins that display useful phenotypes. Towards this goal, previous approaches have leveraged probabilistic models of evolutionary sequences to design proteins reflecting the constraints that govern natural evolution. Other approaches have incorporated labeled data from experiments reflecting a desired phenotype, either alone or alongside models of evolutionary sequences, to design proteins exhibiting a useful functional property. With the goal of minimizing experimental effort and accelerating design cycles, we seek to quantify the minimal amounts and types of evolutionary and experimental data required for designing novel sequences with useful properties, and to identify the best models for utilizing all available data. Using a published model dataset of AAV gene therapy vector designs developed to achieve a desired tissue tropism, we evaluate models using evolutionary and experimental data independently and in concert for their ability to predict capsid liver targeting. We find that particularly when using data on capsid formation for the related phenotype of liver tropism and when evaluating sequences farther away from the wild-type, natural sequence data becomes more important and a combination of both data-types outperforms other supervised and unsupervised benchmarks. We introduce a semi-supervised Bayesian approach trained on a combination of evolutionary sequences and capsid viability that can best predict AAV2 liver tropism for sequences greater than 3 mutations away from wild-type. This has beneficial implications for the design of diverse and functional AAV2 libraries, as well as the broader objective of protein design.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1 Introduction

Adeno-associated virus (AAV) capsids are powerful vectors for gene therapy delivery, but exhibit a broad natural tropism that can result in off-target delivery and low therapeutic efficacy for applications focused on specific tissue targets. Furthermore, the design of capsids is challenging because capsid proteins are highly multi-functional and sequence modification has a high potential for disrupting crucial elements of viral assembly or infectivity [1, 2].

Directed evolution (iterative rounds of local mutation and phenotype selection) is an effective approach for generating functional mutants locally around known viable sequences. However, exploring sequence space farther out from wildtype is challenging, as random synthesis of more distant variants yields few functional proteins [3, 4]. This loss of function is particularly drastic in AAV - of the 10,000 random capsid sequences ranging from 2-10 steps away from AAV2 generated in Bryant et al. (2021), only 2% of sequences 6 mutations away were functional and after 6 mutations, none were functional (Table S1). The AAV2 capsid fitness landscape, therefore, presents an opportunity for methods to design functional phenotypes further away from wildtype - a task that has previously been attempted by using computational approaches. Bryant et al. (2021), used single mutants with measured capsid viability to generate viable capsid proteins up to 20+ mutations away from wild-type. [3] used an unsupervised variational autoencoder trained on natural sequences and semi-supervised variational autoencoders learning from natural sequences as well as a labeled dataset of 1-30 mutations to evaluate and design sequences with viable AAV2 capsids as measured by their experimental method. This study did not explore the performance of unsupervised and semi-supervised models on classifying sequences distant from their training set.

Here, we build upon this work by establishing the utility of computational models for AAV capsid design by quantifying the types and amounts of data needed for different design tasks, including designing near and far to a known target sequence and designing sequences that exhibit specific tropism using only data on the related phenotype of capsid formation. Using available high-throughput assays to design and predict related phenotypes may reduce the cost of experimentation (as tropism assays are much more expensive than capsid formation assays, particularly when examining tropism in non-human primates) and is also pertinent to identifying proteins that exhibit multiple desirable phenotypes. We assess supervised and semi-supervised model performance at different sequence distances from the training set. We also assess how much information is added with labels by comparing with unsupervised model benchmarks.

2 Methods

2.1 Multiple sequence alignments

Multiple sequences alignments of evolutionary sequences were constructed using jackhmmer [5], an iterative profile-HMM based search tool, against the uniref100 database We optimized search depth to maximize sequence coverage and the effective number of sequences included after clustering similar sequences as previously reported [6, 7] and to optimize contact map accuracy to the 1LP3 PDB structure of AAV2 capsid [8].

2.2 Experimental data collection

The capsid viability data consists of 200k AAV2 capsid sequences with mutations in a 28 amino acid region known for heparin and antibody binding. The sequences are assayed for a capsid assembly and the models are evaluated on binarized of non-viable and viable as described in [4]. The tropism data consists of 11k designed AAV2 capsid sequences with mutations in the same 28 amino acid region. We filter the capsid data for substitutions only – as are models are all alignment-based, leaving 44k sequences (Table S1, S2). To assess how well lower mutational data can generalize past their local mutational scan we split train and test sets by training on single mutants and testing on sequences 2+ mutations away from wild-type. The continuous labels describe liver biodistribution as described in [1].

2.3 Models

2.3.1 Unsupervised

The Potts model is an undirected graphical model with probability that includes terms for pairwise combinations of sites:

$$p(x|\theta) = \frac{1}{Z}\text{exp}(E(x)) \tag{1}$$

where $E(x)$ is the log-potential of a given sequence and Z normalizes over possible sequences [**?**].

$$E_{pair}(x) = \sum_i h_i(x_i) + \sum_{i<j} J_{ij}(x_i, x_j) \tag{2}$$

We implemented the Potts models as described in [9]

Two different variational autoencoders are implemented as described in [10, 3, 11]. The EVE (Evolutionary model of Variant Effects) model is a Bayesian variational autoencoder (VAE), capable of capturing complex higher-order interactions across sequence positions [7, 10]. The Sinai model is a dense-layer variational autoencoder with 2 layers in the encoder and 3 in the decoder[11, 3].

2.3.2 Supervised

Supervised Convolutional neural networks, recurrent neural networks, and logistic regression models are implemented as described in [4]. Additionally a 5-fold cross-validation l2 normalized model was implemented with one-hot sequence regression to capsid viability labels.

2.3.3 Semi-supervised

We compare a variant of the M1 model as inspired by [12, 13]. Both Kingma, et al. and Gomez-Bombarelli, et al. jointly train their semi-supervised VAE, using both the experimental data and the natural sequence data to train reconstruction. However, in our use case, it is not desirable to train the reconstruction of x on the experimental data due to a variety of factors: 1) a data imbalance of experimental data and natural sequences 2) the experimental sequence space is not what we want to focus our training of the latent space because we want to be able to extrapolate beyond the local fitness landscape; 3) the sequence space from which we learn the experimental labels can be from synthetic libraries, which are not vetted for any stability or function, so we do not want our model to try to reconstruct these sequences. So, we instead train the VAE and the supervised top model sequentially.

We implement the augmented Potts and DeepSequence as described in [14]. The augmented Potts and augmented DeepSequence model consists of a 5-fold cross validation l2-normalized linear regression trained off the one-hot-encoded sequences concatenated with the Potts Hamiltonian or DeepSequence ELBO, respectively.

We also implement the semi-supervised variational autoencoder as described in [3]. This is where the variational autoencoder is trained not only off an alignment of evolutionary sequences but also the phenotypically fit sequences of the training dataset.

We also propose a method for combining predictions of the unsupervised and supervised models trained above using Bayes rule (Eq. 3) to approximate the conditional probability of a sequence given the desired phenotype. This method is fully modular - we can use any combination of separately trained supervised (Eq. 4) and unsupervised (Eq. 5) models, requiring no additional training of a semi-supervised model.

$$p(x|y, \theta_0, \phi_0) \propto p(y|x, \theta_0) \cdot p(x, \phi_0) \tag{3}$$

$$\theta_0 = \arg\max_\theta \sum_{i=1}^{M} p(Y_i^{experiment}|X_i^{experiment}, \theta) \tag{4}$$

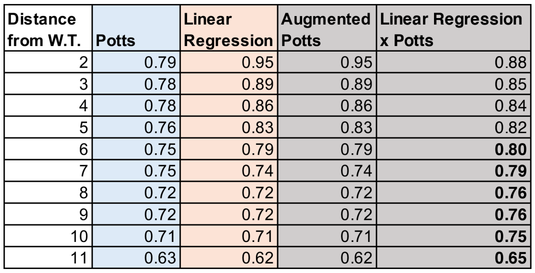
$$\phi_0 = \arg\max_\phi \sum_{i=1}^{N} p(X_i^{nature}|\phi) \tag{5}$$

3 Results

Here we evaluate a set of models trained on unlabeled evolutionary data, labeled experimental single-mutation data, or a combination (described above) for their performance at predicting which synthetic capsids assemble. Each model is assessed by how well it classifies viable and non-viable sequences a specific mutational disstance away from WT. We also evaluate these models for their performance at predicting an alternate fitness phenotype of wildtype-AAV2: liver tropism.

3.1 Predicting capsid viability

For mutations up to 6 mutations away from wild type a simple ridge regression trained on single mutation expression data best distinguishes viable and non-viable capsids, with marginal or no gain from using evolutionary information (Table 1). This is perhaps not surprising as the multiples in the test set were designed based on the single mutation results. When training on both double and single mutations, a simple ridge regression outperforms all unsupervised and supervised models regardless of distance from wildtype.



Distance from W.T.	Potts	Linear Regression	Augmented Potts	Linear Regression x Potts
2	0.79	0.95	0.95	0.88
3	0.78	0.89	0.89	0.85
4	0.78	0.86	0.86	0.84
5	0.76	0.83	0.83	0.82
6	0.75	0.79	0.79	**0.80**
7	0.75	0.74	0.74	**0.79**
8	0.72	0.72	0.72	**0.76**
9	0.72	0.72	0.72	**0.76**
10	0.71	0.71	0.71	**0.75**
11	0.63	0.62	0.62	**0.65**

Table 1: Here are the top performing supervised (red), semi-supervised (gray), and unsupervised (blue) capsid viability classifications denoted via AUC. The AUC performances are stratified by distance from wildtype. Below 5 mutations, the linear regression outperforms all models in AUC when classifying binary labels of capsid viability. However after 5 mutations, evolutionary information allows unsupervised and semi-supervised models to outperform linear regression. For more model performances see Table S2.

3.2 Predicting liver tropism

We next explored the question of whether a model that is trained on and successfully predicts the phenotype of capsid formation can predict the alternate, yet related phenotype of liver tropism. The supervised ridge regression trained on single and double mutant capsid viability is the best model for predicting capsid viability of more distant sequences (Table S3). However, when evaluating models trained on capsid viability for their ability to predict liver tropism, the supervised model trained on capsid viability under-performs an unsupervised Potts model trained on natural sequences for designs more than 6 mutations away from the wildtype – no matter how much training data from the capsid viability experiment is incorporated (Table S4).

We next evaluated a bayesian model that combines supervised learning with an unsupervised model. We find that this approach outperforms unsupervised models in mutational regimes where the supervised training dominates (mutational distances closer to wildtype) although it underperforms compared to supervised models for very close mutational distances. This approach also outperforms both supervised and unsupervised models in regimes where the unsupervised training dominates (further away from experimental wildtype) (Table 2).

4 Discussion

Intuitively, experimental data closely conforms to the features of the experiment and is very useful for predicting the function of designed sequences for the same experiment. Natural sequence data, on the other hand, captures the broad constraints acting on evolution of these sequences, which are generally

Distance from W.T.	Linear Regression	Potts	Linear Regression x Potts
1	**0.78**	0.32	0.60
2	**0.62**	0.40	0.55
3	**0.53**	0.39	0.50
4	0.40	0.34	**0.41**
5	0.35	0.29	**0.35**
6	0.28	0.30	**0.31**
7	0.28	0.30	**0.32**
8	0.07	0.12	**0.13**

Table 2: Spearman correlations of model predictions with observed liver tropism for designs at various distances from wildtype. Above 3 mutations, the evolutionary data allows semi-supervised models to outperform the linear regression trained on capsid viability. For sequences per distance see table S5

reflected in a range of phenotypes related to the protein's natural function. Especially in the case of deep mutational scans where the experimental data is quite local in its sample of sequence space, similarly, the predictions of a model trained on such a space are limited to a smaller neighborhood around the starting sequence, whereas a model trained on a broader range of natural evolution is better suited for predicting further away and generalizing to fitness phenotypes beyond the ones directly measured in the experiment.

5 Future Directions

Future directions to this project involve:

- Evaluating unsupervised, supervised, and semi-supervised models for their performance at generating functional AAV sequences with the desired phenotype by characterizing designed sequences from each model in-silico and in experiments.

References

[1] Pierce J Ogden, Eric D Kelsic, Sam Sinai, and George M Church. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science*, November 2019.

[2] Cynthia E Dunbar, Katherine A High, J Keith Joung, Donald B Kohn, Keiya Ozawa, and Michel Sadelain. Gene therapy comes of age. *Science*, 359(6372), January 2018.

[3] Sam Sinai, Nina Jain, George M Church, and Eric D Kelsic. Generative AAV capsid diversification by latent interpolation. April 2021.

[4] Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.*, 39(6):691–696, June 2021.

[5] Sean R Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195, October 2011.

[6] Thomas A Hopf, Lucy J Colwell, Robert Sheridan, Burkhard Rost, Chris Sander, and Debora S Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, June 2012.

[7] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, October 2018.

[8] wwPDB: 1LP3. `https://www.wwpdb.org/pdb?id=pdb_00001lp3`. Accessed: 2022-10-4.

[9] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D structure computed from evolutionary sequence variation, 2011.

[10] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data, 2021.

[11] Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. December 2017.

[12] Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised learning with deep generative models. June 2014.

[13] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a Data-Driven continuous representation of molecules, 2018.

[14] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Combining evolutionary and assay-labelled data for protein fitness prediction. March 2021.

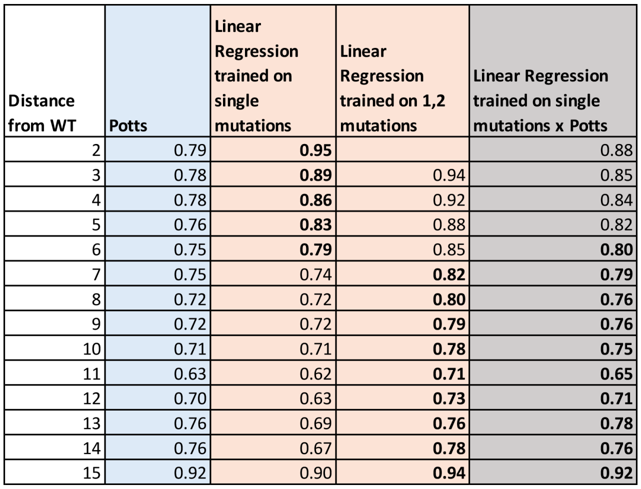
A Appendix

	Randomly synthesized			Designed		
Distance from W.T.	**No. sequences**	**No. viable**	**% Viable**	**No. sequences**	**No. viable**	**% Viable**
1				518	285	55%
2	7154	2206	31%	4045	3615	89%
3	806	131	16%	6306	5423	86%
4	727	61	8%	6126	4795	78%
5	669	31	5%	5783	4040	70%
6	626	12	2%	4919	2855	58%
7	102	0	0%	783	436	56%
8	103	0	0%	667	283	42%
9	101	0	0%	571	157	27%
10	96	0	0%	500	107	21%
11				455	59	13%
12				424	28	7%
13				384	14	4%
14				340	8	2%

Table S1: Random sequences lack viable sequences above 6 mutations from wild-type where designed sequences have viable sequences up to 15 mutations away from wild-type

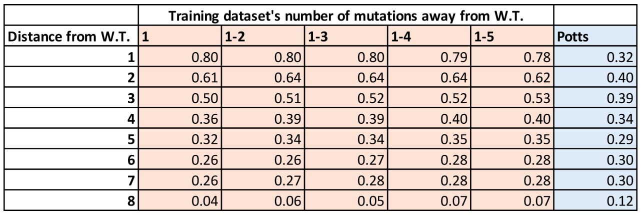
Distance from W.T.	EVE	Potts	VAE (Sinai 2021)	CNN	Logistic Regression	RNN	Ridge Regression	M1	VAE+ (Sinai 2021)	UniRep	Linear Regression x Potts	Augmented Potts	Num Sequences	Sequences with positive labels
2	0.61	0.79	0.50	0.77	0.80	0.79	0.95	0.95	0.51	0.55	0.88	0.95	11199	5821
3	0.61	0.78	0.50	0.65	0.76	0.77	0.89	0.89	0.51	0.54	0.85	0.89	7112	5554
4	0.63	0.78	0.49	0.60	0.73	0.75	0.86	0.86	0.51	0.55	0.84	0.86	6853	4856
5	0.64	0.76	0.49	0.57	0.73	0.74	0.83	0.82	0.51	0.55	0.82	0.83	6452	4071
6	0.62	0.75	0.49	0.56	0.70	0.71	0.79	0.79	0.51	0.55	0.80	0.79	5545	2867
7	0.61	0.75	0.50	0.49	0.70	0.69	0.74	0.74	0.51	0.56	0.79	0.74	885	436
8	0.63	0.72	0.49	0.45	0.71	0.66	0.72	0.71	0.55	0.56	0.76	0.72	770	283
9	0.62	0.72	0.50	0.49	0.71	0.64	0.72	0.71	0.51	0.56	0.76	0.72	672	157
10	0.59	0.71	0.52	0.56	0.70	0.63	0.71	0.71	0.52	0.55	0.75	0.71	596	107
11	0.62	0.63	0.54	0.52	0.66	0.55	0.62	0.61	0.53	0.52	0.65	0.62	455	59
12	0.62	0.70	0.52	0.48	0.61	0.57	0.63	0.63	0.52	0.54	0.71	0.64	424	28
13	0.64	0.76	0.50	0.46	0.70	0.55	0.69	0.69	0.48	0.55	0.78	0.70	384	14
14	0.64	0.76	0.60	0.48	0.76	0.48	0.67	0.66	0.43	0.51	0.76	0.67	340	8
15	0.98	0.92	0.77	0.47	0.67	0.39	0.90	0.88	0.29	0.59	0.92	0.91	297	1

Table S2: All model results of capsid viability with number of sequences per distance from wild-type



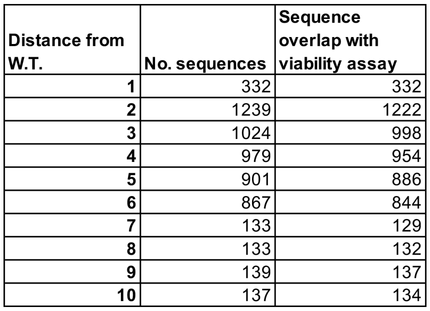
Distance from WT	Potts	Linear Regression trained on single mutations	Linear Regression trained on 1,2 mutations	Linear Regression trained on single mutations x Potts
2	0.79	**0.95**		0.88
3	0.78	**0.89**	0.94	0.85
4	0.78	**0.86**	0.92	0.84
5	0.76	**0.83**	0.88	0.82
6	0.75	**0.79**	0.85	**0.80**
7	0.75	0.74	**0.82**	**0.79**
8	0.72	0.72	**0.80**	**0.76**
9	0.72	0.72	**0.79**	**0.76**
10	0.71	0.71	**0.78**	**0.75**
11	0.63	0.62	**0.71**	**0.65**
12	0.70	0.63	**0.73**	**0.71**
13	0.76	0.69	**0.76**	**0.78**
14	0.76	0.67	**0.78**	**0.76**
15	0.92	0.90	**0.94**	**0.92**

Table S3: Model results of capsid viability with number of sequences per distance from wild-type. When training on 1-2 mutations, the ridge regression can outperform unsupervised Potts model at all distances from wild-type



	Training dataset's number of mutations away from W.T.					
Distance from W.T.	1	1-2	1-3	1-4	1-5	Potts
1	0.80	0.80	0.80	0.79	0.78	0.32
2	0.61	0.64	0.64	0.64	0.62	0.40
3	0.50	0.51	0.52	0.52	0.53	0.39
4	0.36	0.39	0.39	0.40	0.40	0.34
5	0.32	0.34	0.34	0.35	0.35	0.29
6	0.26	0.26	0.27	0.28	0.28	0.30
7	0.26	0.27	0.28	0.28	0.28	0.30
8	0.04	0.06	0.05	0.07	0.07	0.12

Table S4: Ridge results of capsid viability with number of sequences per distance from wild-type. Increasing the amount of training data does not improve spearman correlations of liver tropism above 6 mutations away from wild-type



Distance from W.T.	No. sequences	Sequence overlap with viability assay
1	332	332
2	1239	1222
3	1024	998
4	979	954
5	901	886
6	867	844
7	133	129
8	133	132
9	139	137
10	137	134

Table S5: Almost all sequences from liver tropism assay overlap with sequences from capsid viability assay.