Beyond Pairwise : Capturing the Full Complexity of Genetic Variation

Martin Auobi

Bioinformatic Course 2025 - Dr.Rezazadegan

Deep generative models of genetic variation capture mutation effects

Adam J. Riesselman*
Program in Biomedical Informatics
Harvard Medical School
ariesselman@g.harvard.edu

John B. Ingraham*
Program in Systems Biology
Harvard University
ingraham@fas.harvard.edu

Debora S. Marks
Department of Systems Biology
Harvard Medical School
debbie@hms.harvard.edu

* Equal contribution

What is the Problem ?

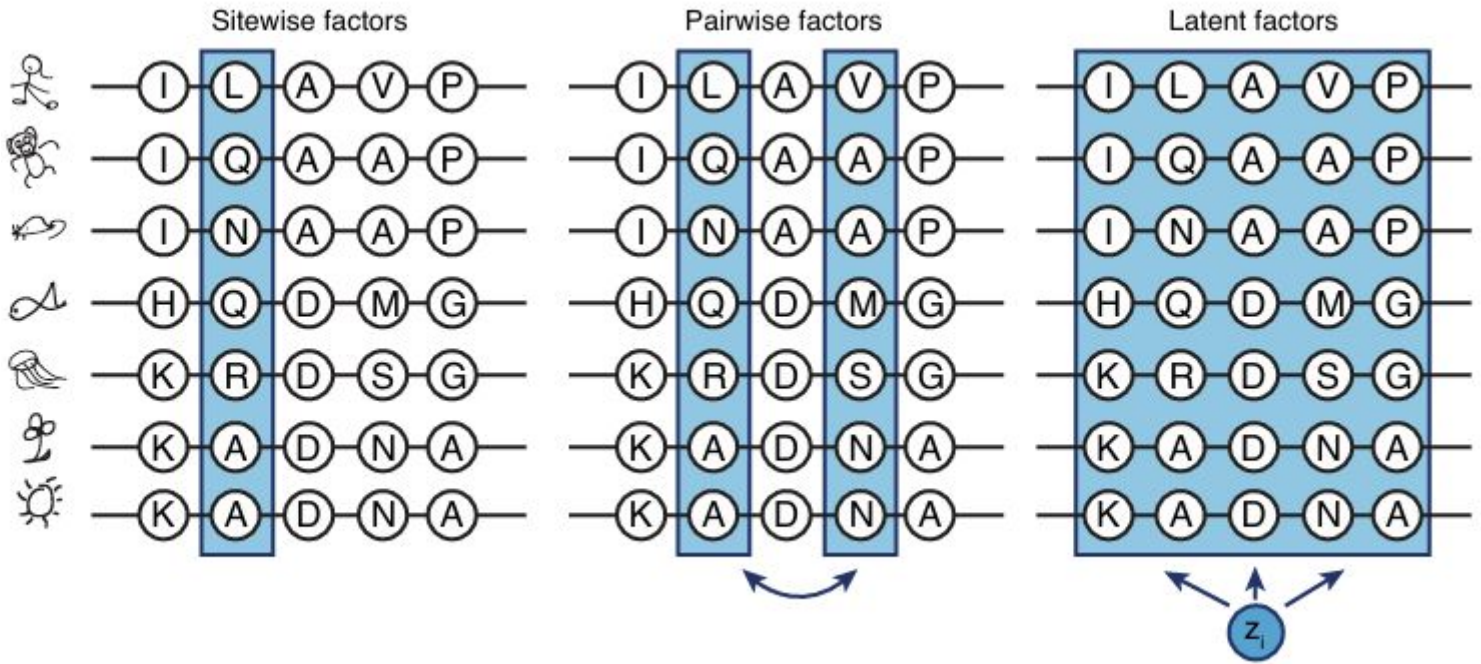
The Problem which this paper claims to solve is the difficulty of predicting how specific mutations alter the function of proteins and RNAs but the this is a just a immediate application of something that is more important .

The main Contribute of this paper is introducing a deep learning based approach to understanding the Grammar of sequences , their correlation and finding motifs and patterns not just for predicting the mutation effect but for a variations of applications .

What is the Problem ?

The functions of proteins and RNAs are determined by a myriad of interactions between their constituent residues, but most quantitative models of how molecular phenotype depends on genotype must approximate this by simple additive effects. While recent models have relaxed this constraint to also account for pairwise interactions, these approaches do not provide a tractable path towards modeling higher-order dependencies. Here, we show how latent variable models with nonlinear dependencies can be applied to capture beyond-pairwise constraints in biomolecules. We present a new probabilistic model for sequence families, DeepSequence, that can predict the effects of mutations across a variety of deep mutational scanning experiments significantly better than site independent or pairwise models that are based on the same evolutionary data. The model, learned in an unsupervised manner solely from sequence information, is grounded with biologically motivated priors, reveals latent organization of sequence families, and can be used to extrapolate to new parts of sequence space.

What is the Problem ?



What are the applications ?

The ability of predicting mutation effect on the phenotype (like the function of Protein) is a crucial problem in bioinformatic because we're aiming to answer questions like :

which genetic variants in humans underlie disease to developing modified proteins that have useful properties

What are the applications ?

The ability of predicting mutation effect on the phenotype (like the function of Protein) is a crucial problem in bioinformatic because we're aiming to answer questions like :

which genetic variants in humans underlie disease to developing modified proteins that have useful properties

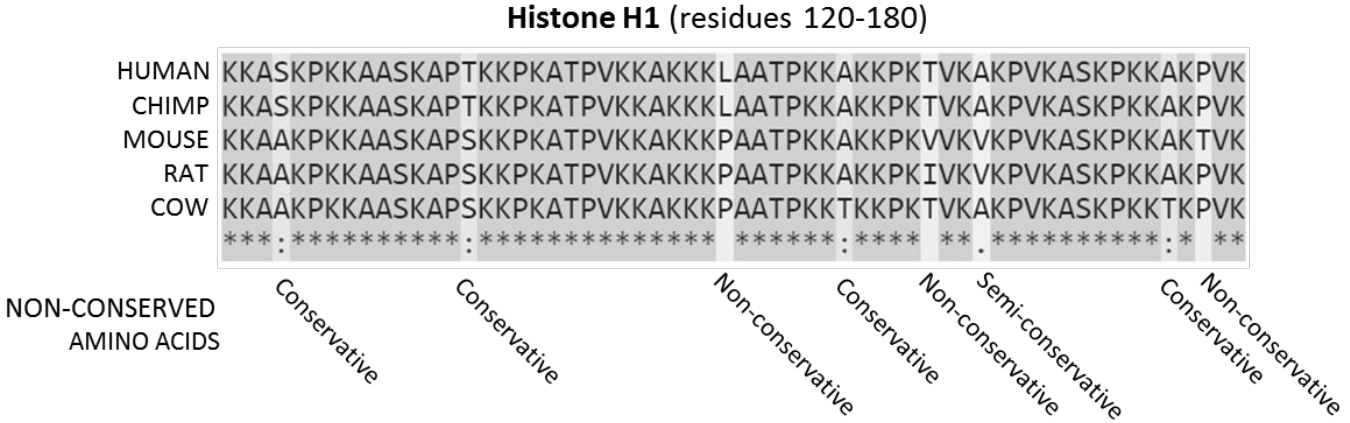
so there is need to be able to rapidly assess whether a given mutation to a protein or RNA will disrupt its function . Motivated by these diverse applications, new technologies have emerged that simultaneously assess the effects of thousands of mutations in parallel sometimes referred to as "deep mutational scans" .

How these models actually work ?

Since sequence space is exponentially large and experiments are resource-intensive, accurate computational methods are an important component for high-throughput sequence annotation and design. Many computational tools have been developed for predicting the effects of mutations, and most progress in efficacy of predictions has been driven by the ability of models to leverage the signal of evolutionary conservation among related sequences

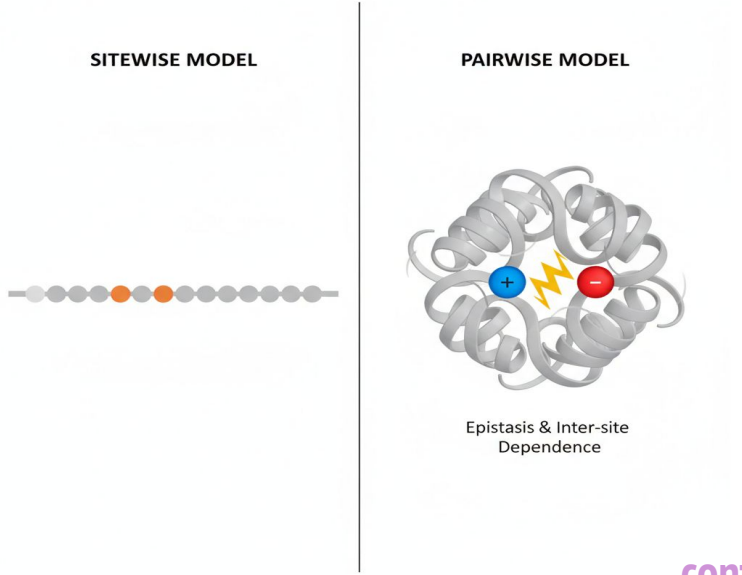
How these models actually work ?

Since sequence space is exponentially large and experiments are resource-intensive, accurate computational methods are an important component for high-throughput sequence annotation and design. Many computational tools have been developed for predicting the effects of mutations, and most progress in efficacy of predictions has been driven by the ability of models to leverage the signal of evolutionary conservation among related sequences



How these models actually work ?

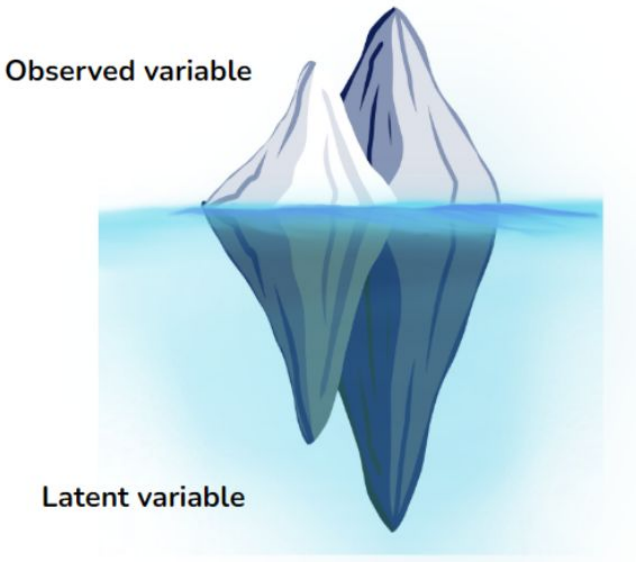
While previous approaches analyzed this signal in a residue-independent manner, recent work has demonstrated that incorporating inter-site dependencies using a pairwise model can power state of art predictions for high-throughput mutational experiments



context-dependent mutation effect

Latent Variables

Definition: Latent variables are variables that are not directly observed but are inferred from other variables that are observed (measured).



Latent Variables

We can train deep neural networks using unlabeled data to infer latent variables from our data . this latent variables act like the most important features that we can extract from our data to describe them .

After this we can use this variables to do another tasks like :

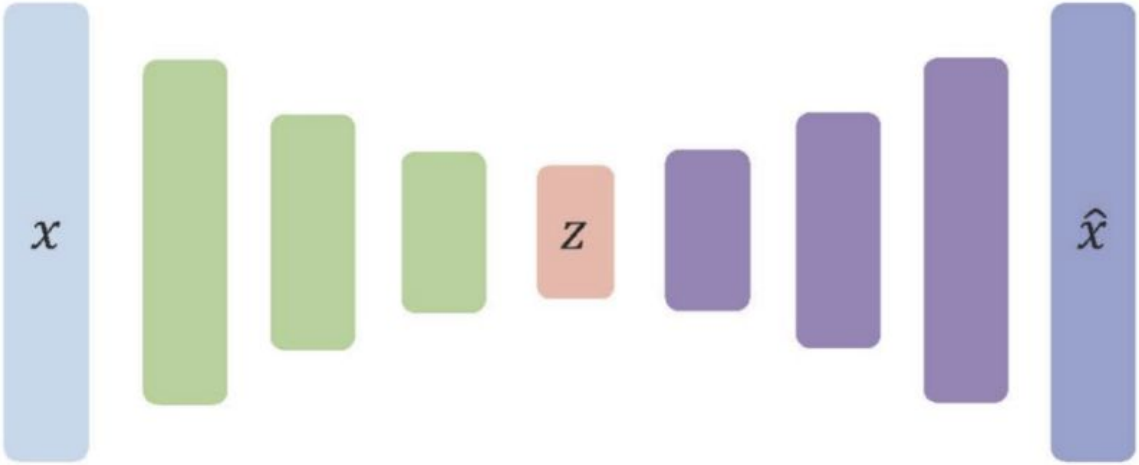
1 - Generating new Data
2 - Classification

.
.
.
.

Autoencoders

"Encoder" learns mapping from the data, x, to a low-dimensional latent space, z

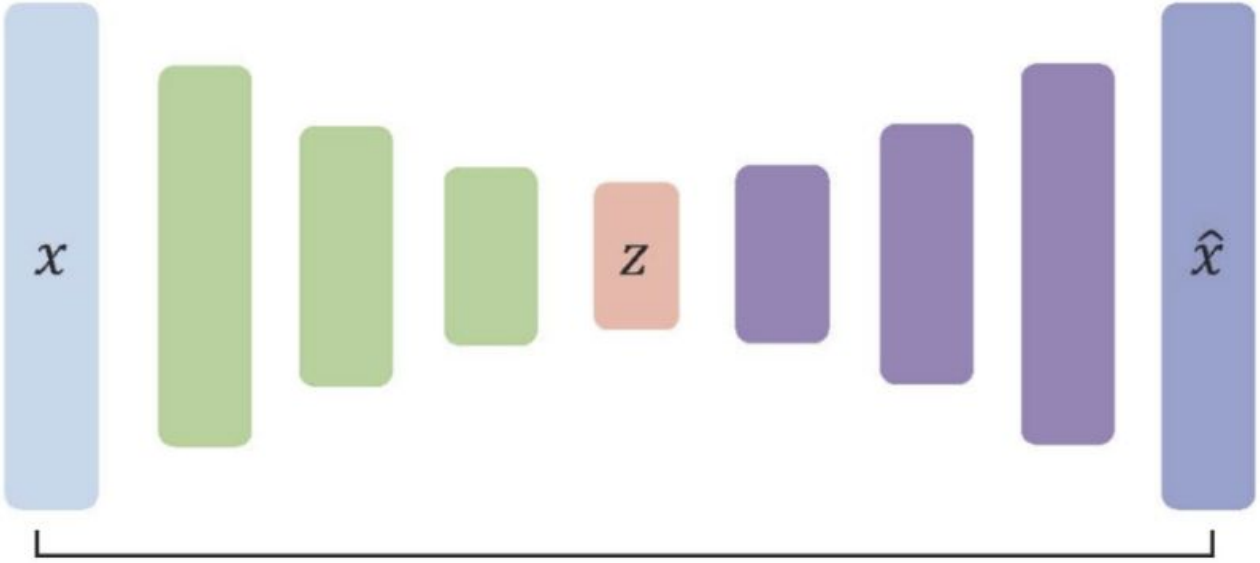
Autoencoders



"Encoder" learns mapping from the data, x, to a low-dimensional latent space, z

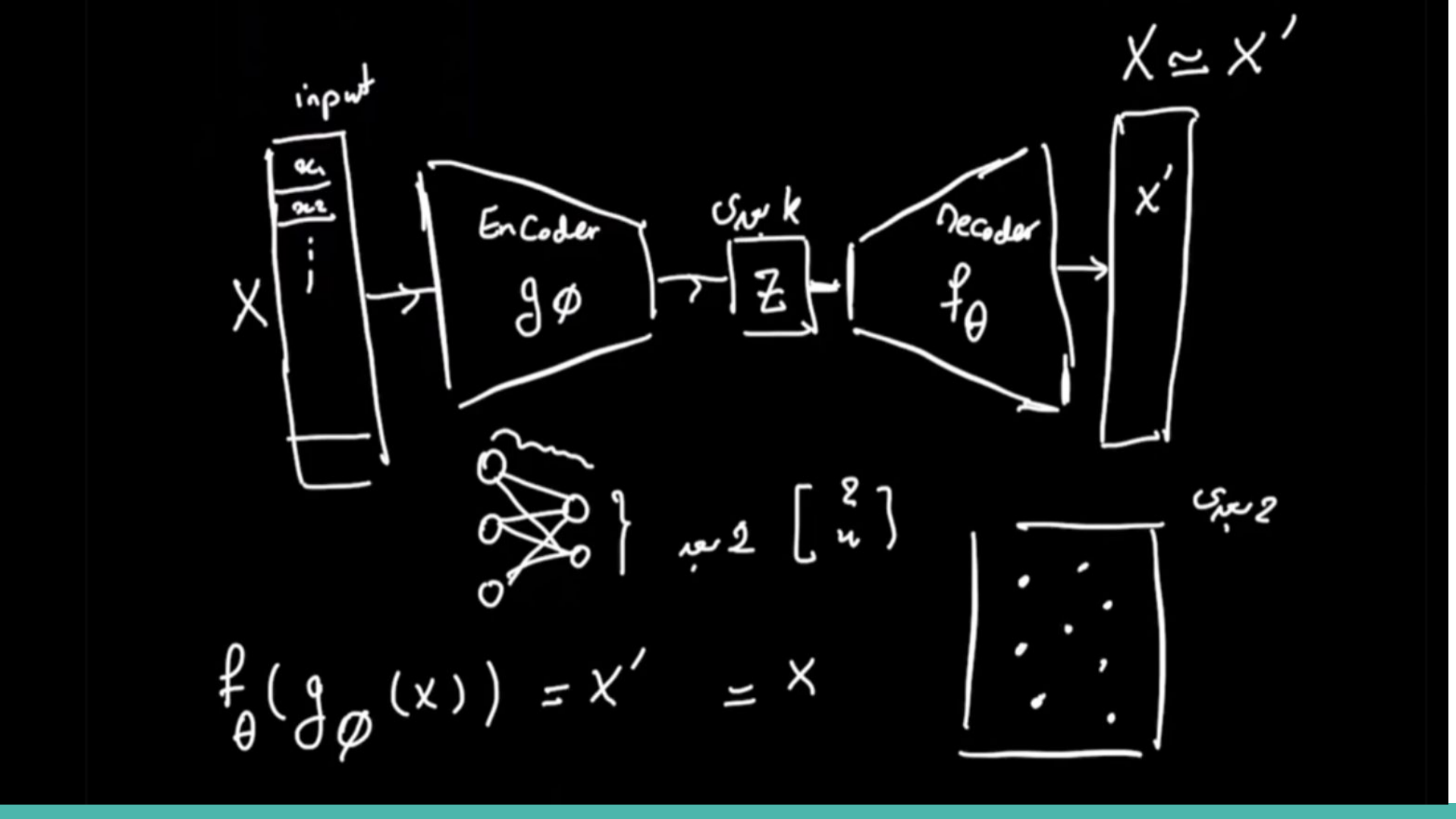
"Decoder" learns mapping back from latent z, to a reconstructed observation, x

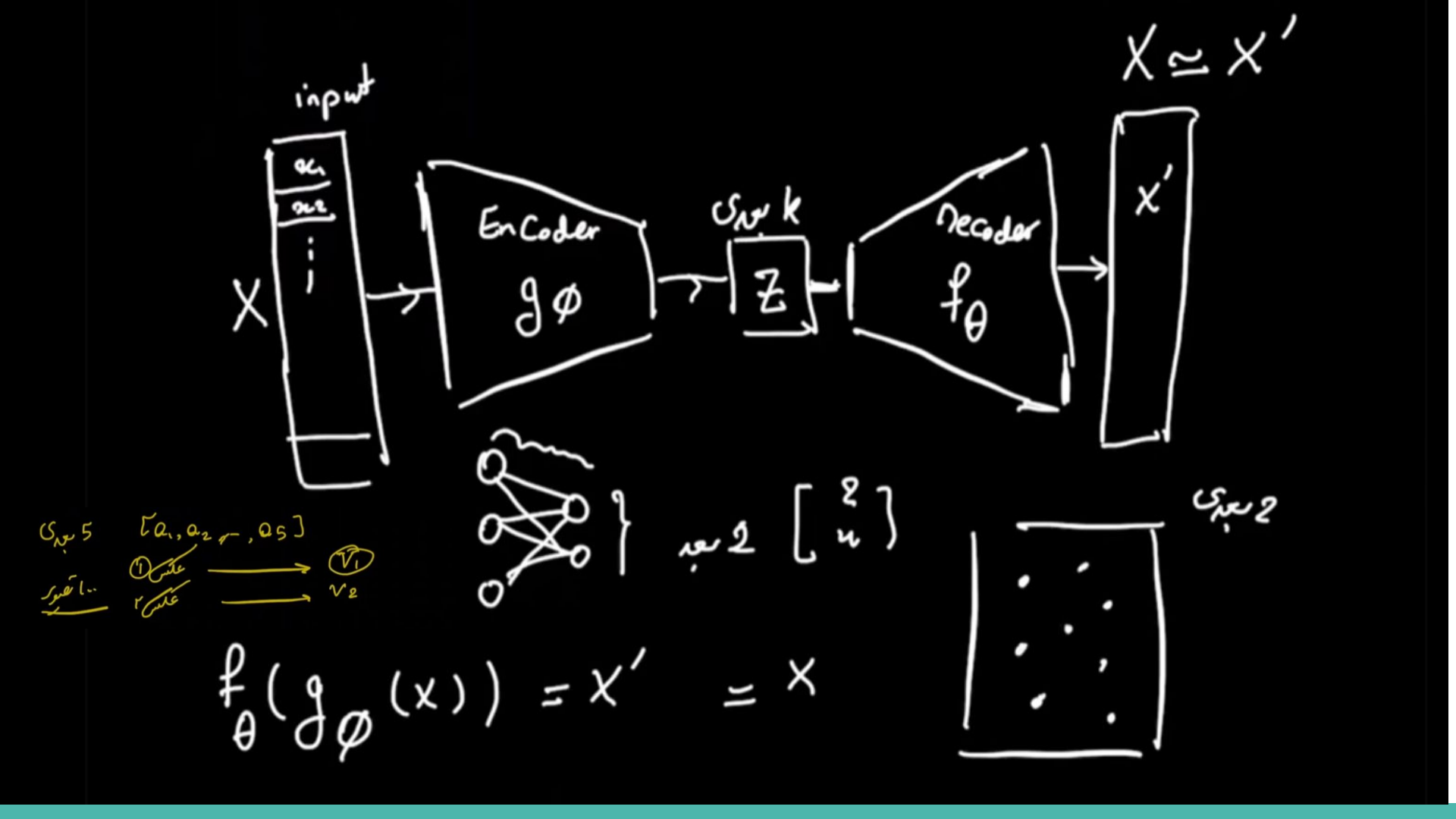
Autoencoders



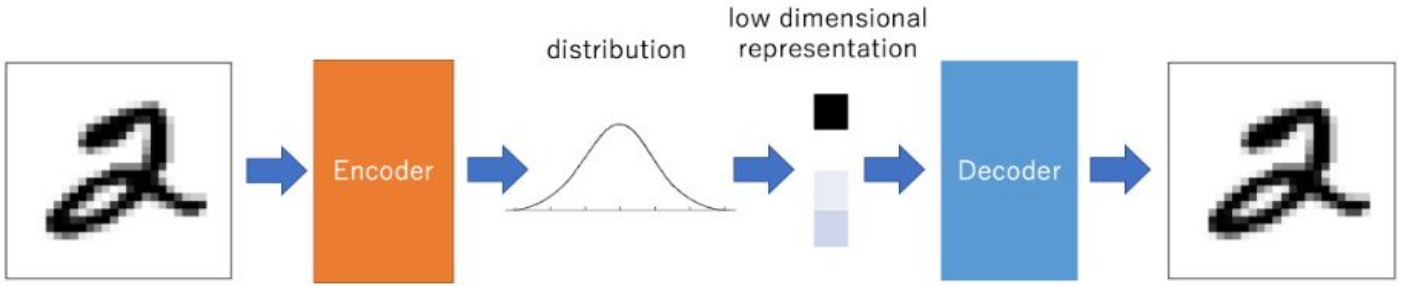
$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

Loss function doesn't use any labels!!

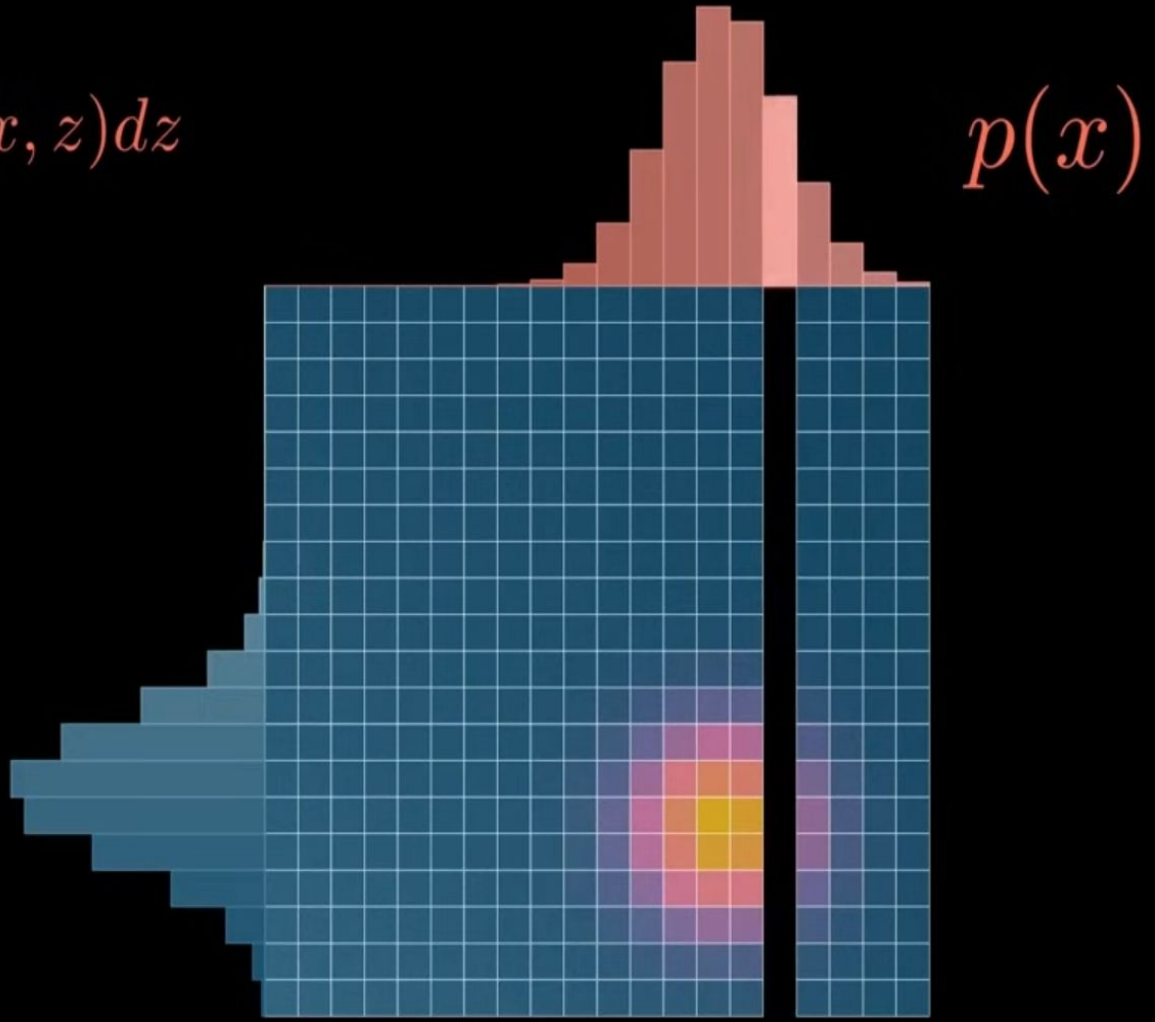

$X \simeq X'$
input
x_1
x_2
x_n
X
Encoder
g_ϕ
Latent
z
Decoder
f_θ
X'
Dense
Dense
$f_\theta(g_\phi(x)) = X' \simeq X$


$X \simeq X'$
input
x_1
x_2
x_n
X
Encoder
g_ϕ
Bottleneck
z
Decoder
f_θ
x'
بعد 2
بعد 5
$[a_1, a_2, \ldots, a_5]$
عکس ۱
عکس ۲
v_1
v_2
بعد 2
$f_\theta(g_\phi(x)) = x' \simeq x$


encode
decode
hidden


distribution
low dimensional representation
Encoder
Decoder

$$p(x) = \int p(x, z) dz$$

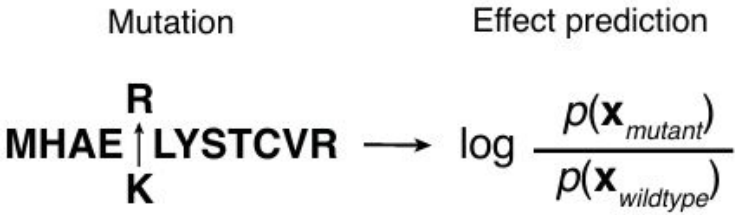


How to Use a VAE for predicting mutation effect

One strategy for reasoning about the consequences of mutations to genes is to develop models of the selective constraints that have been relevant throughout evolution. Since the genes that we observe across species today are the results of long-term evolutionary processes that select for functional molecules, a *generative model* of the outputs of evolution must implicitly learn some of these functional constraints. If we approximate the evolutionary process as a "sequence generator" with probability $p(\mathbf{x}|\boldsymbol{\theta})$ that has been fit to reproduce the statistics of evolutionary data, we can use the probabilities that the model assigns to any given sequence as a proxy for the relative plausibility that the molecule satisfies functional constraints.

How to Use a VAE for predicting mutation effect

How to Use a VAE for predicting mutation effect



How to Use a VAE for predicting mutation effect

The paper innovation is introducing another class of probabilistic model which is **nonlinear latent variable**

We introduce a nonlinear latent variable model $p(\mathbf{x}|\theta)$ to implicitly capture higher order interactions between positions in a sequence in a protein family. For every observed sequence $\mathbf{x}$, we posit unobserved latent variables $\mathbf{z}$ together with a generative process p(z)p(x|z) that specifies a joint distribution over hidden variables and observed variables.

We introduce a nonlinear latent variable model $p(\mathbf{x}|\boldsymbol{\theta})$ to implicitly capture higher order interactions between positions in a sequence in a protein family. For every observed sequence **x,** we posit unobserved latent variables **z** together with a *generative process* p(**z**)p(**x**|**z**) that specifies a joint distribution over hidden variables and observed variables. Inference this model is challenging, as the marginal probability of the observed data, p(**x**), requires integrating over all possible hidden **z** with

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z})d\mathbf{z}.$$

While directly computing this probability is intractable in the general case, we can use variational inference[50] to instead form a lower bound on the (log) probability. This bound, known as the Evidence Lower Bound (ELBO), takes the form

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) \geq \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})] - D_{KL}\big(q(\mathbf{z}|\boldsymbol{x},\boldsymbol{\phi})||p(\mathbf{z})\big),$$

where q(**z**|**x**) is an *approximate posterior* for hidden variables given the observed variables p(**z**|**x**). Modeling both the conditional distribution p(**x**|**z**) of the generative model and the approximate posterior q(**z**|**x**) with neural networks results in a flexible model-inference combination, known as a Variational Autoencoder [47, 48] (Figure 1b).

Neural network-parameterized latent variable models can in principle model complex correlations in data, but without additional architectural and statistical considerations may be hard to interpret and unlikely to generalize. We encourage generalization in three ways: First, we encourage *sparse interactions* by placing a group sparsity prior over the last layer of the neural network for $p(\mathbf{x}|\mathbf{z})$ that encourages each hidden unit in the network to only influence a few positions at a time. This is motivated by the observation that higher order interactions in proteins, while importantly higher than second order, are nevertheless low-valence compared to the number of residues in the protein. Second, we encourage *correlation between amino acid usage*, by convolving the final layer with a width-1 convolution operation. Thirdly, we estimate all global parameters with variational Bayes by estimating approximate posterior distributions over each model parameter. The result is that rather than learning a single neural network for $p(\mathbf{z}|\mathbf{x})$, we learn an infinite *ensemble* of networks. This joint variational approximation is then optimized by stochastic gradient ascent on the ELBO to give a fully trained model (Methods).

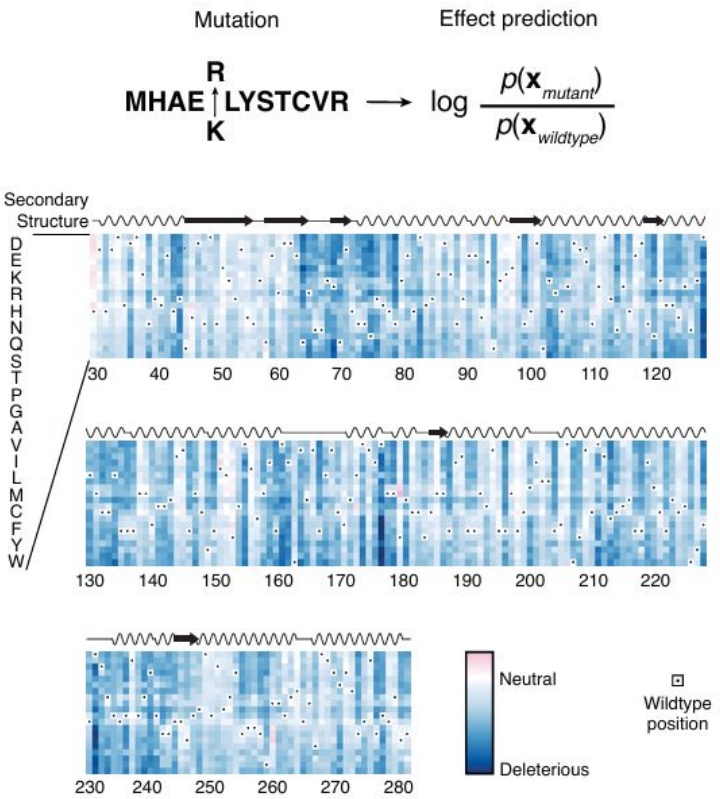
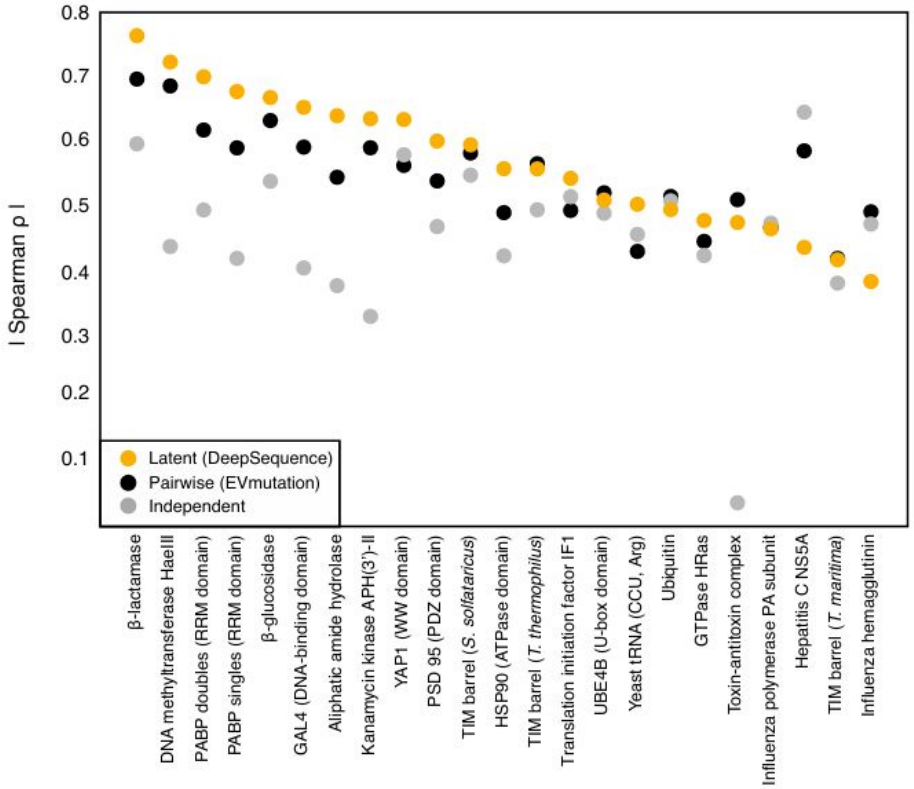
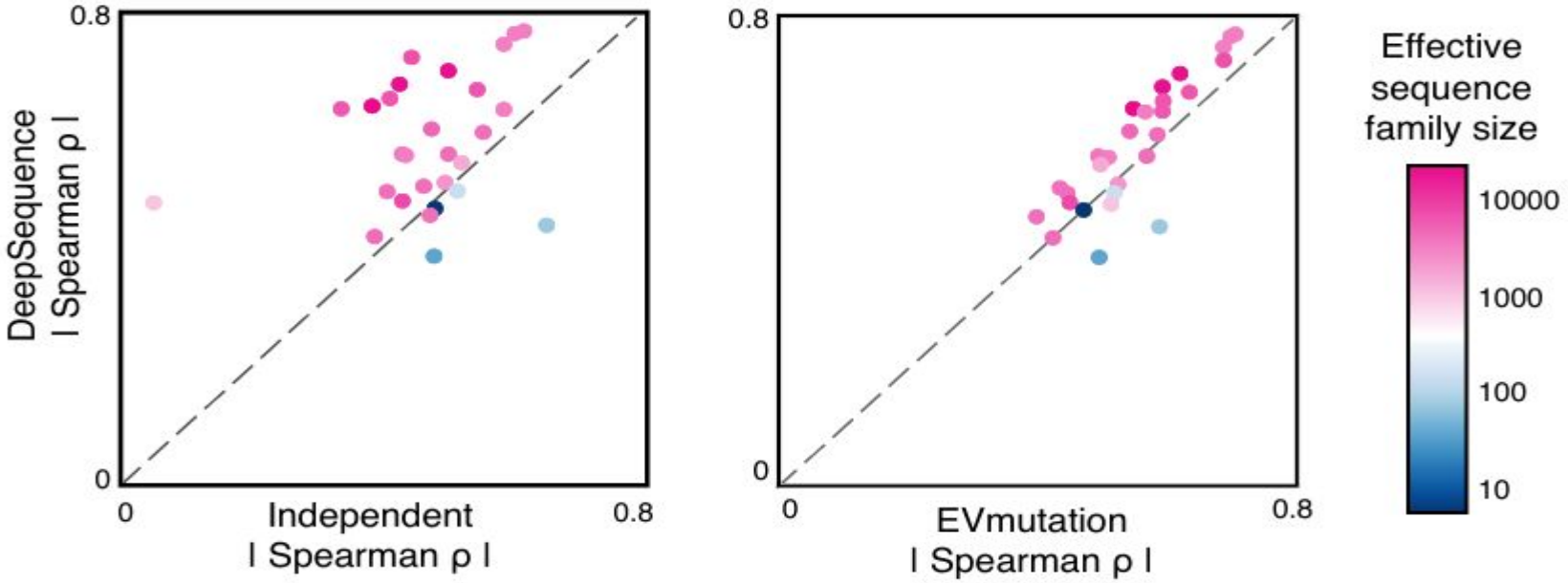


Figure 2. Mutation effects can be quantified by likelihood ratios. *After fitting a probabilistic model to a family of homologous sequences, we heuristically quantify the effect of mutation as the log ratio of mutant likelihood to wild type likelihood (as approximated by the ELBO; Methods). Below: mutation effect scores for all possible point mutations to β-lactamase.*

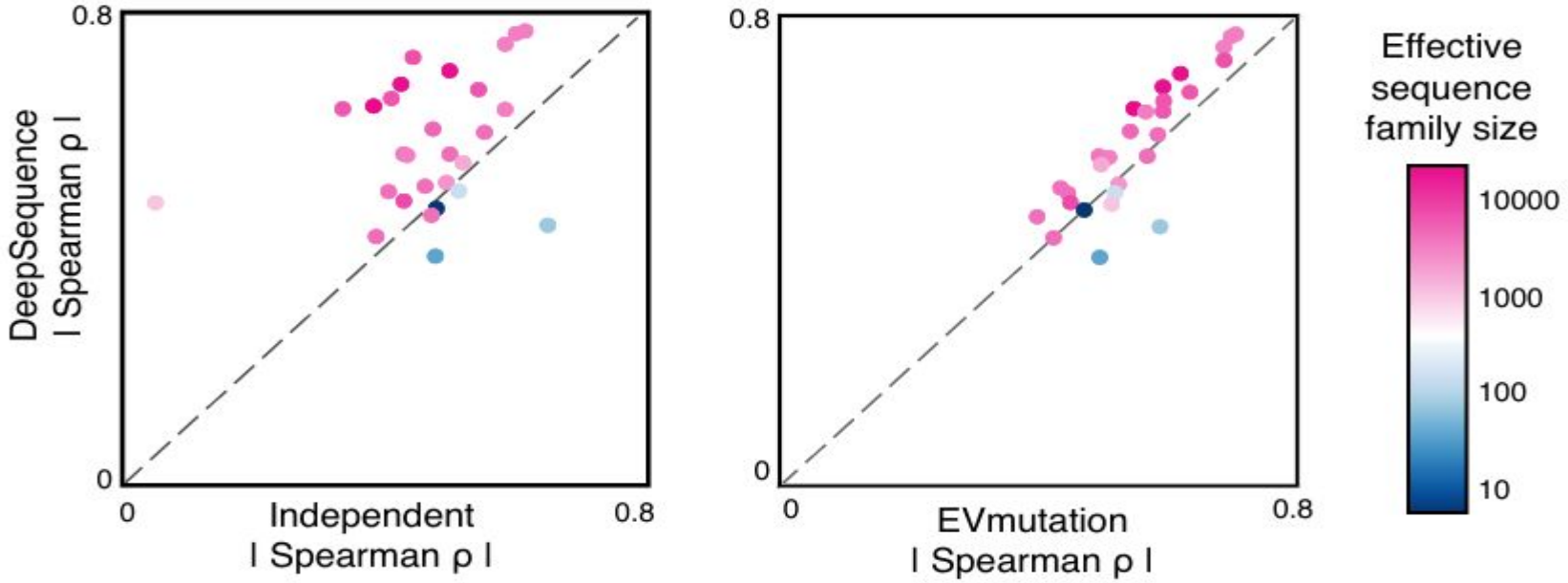
How to evaluate the Model ?



How to evaluate the Model ?



How to evaluate the Model ?



The latent variable model tends to be more predictive of mutational effects than pairwise and site-independent models when fit to deeper, more evolutionarily diverse sequence alignments as measured by the effective family size .

How to evaluate the Model ?

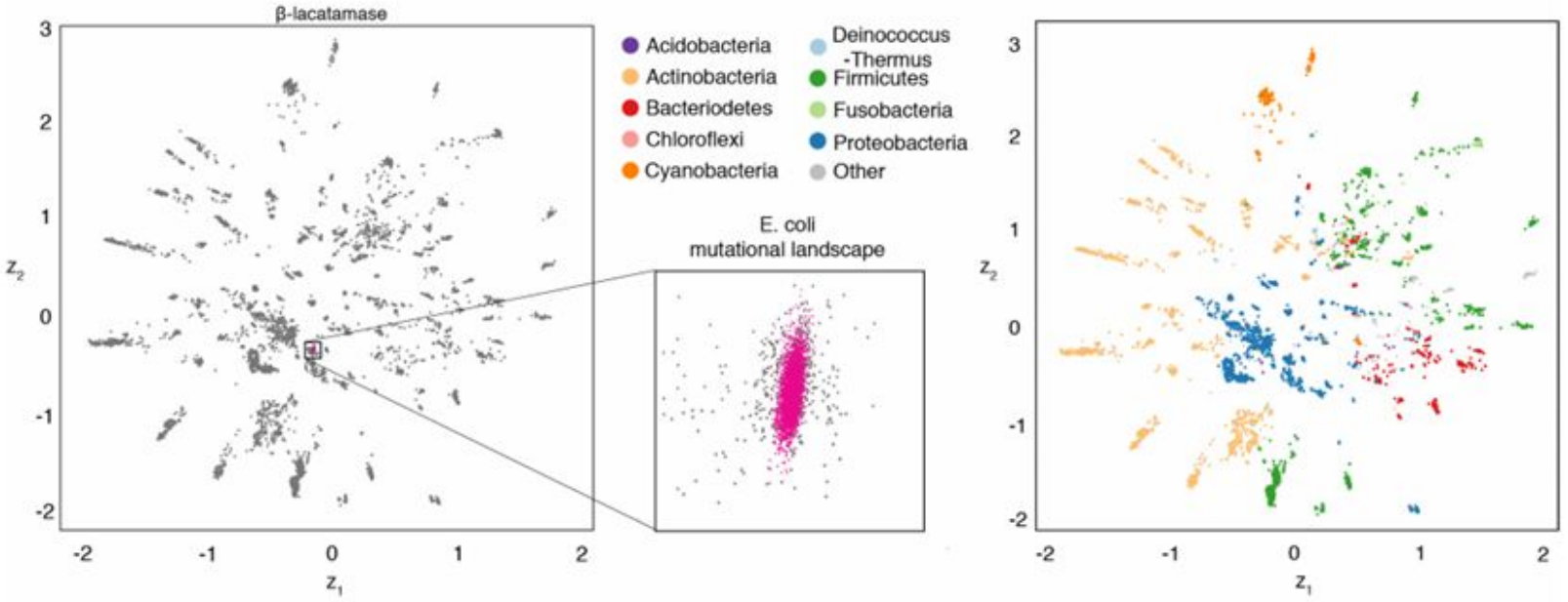
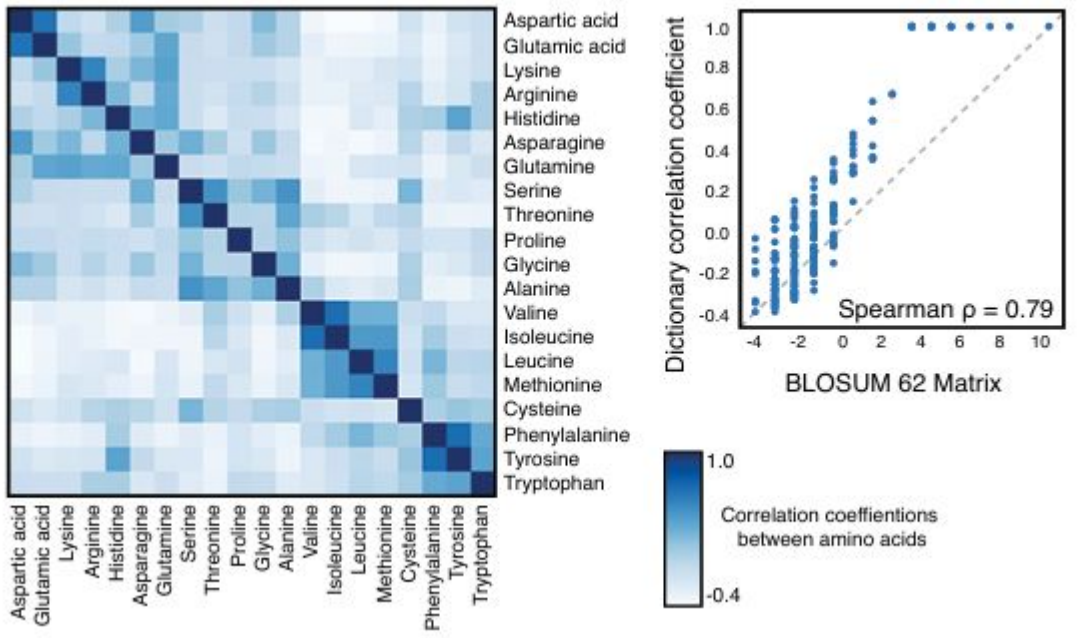


Figure 4. Latent variables capture organization of sequence space. *In a two-dimensional latent space for the β-lactamase family, closeness in latent space reflects phylogenetic groupings. When examining the variation within a single deep mutational scanning experiment, it occupies only a very small portion of the sequence space of the entire evolutionary family.*

It's not just about bayesian models !

We found that the combination of using biologically motivated priors and Bayesian approaches for inference on the weights was important to learning models that generalize. To test the importance of these various aspects of the model and inference, we performed an ablation study across a subset of the proteins. We found that using (i) Bayesian variational approximations on the weights, (ii) sparse priors on the last layer, (iii) a final width 1 convolution for amino acid correlations, and (iv) a global temperature parameter all improved the ability of the model to predict the effects of mutations across this subset of the experiments.

It's not just about bayesian models !



It's not just about bayesian models !

	Bayesian θ						MAP θ									
Sparsity [S]	✓	✓	✓													
Convolution [C]	✓	✓		✓												
Temperature [T]	✓		✓	✓	✓											
L2 Regularization							✓	✓	✓	✓	✓	✓				
Dropout							✓	✓			✓					
[S+C+T]							✓		✓							
Final ReLU								✓		✓			✓		Pair	Site
β-lactamase	0.73	0.73	0.73	0.73	0.73	**0.74**	0.53	0.61	0.04	0.40	0.56	0.37	0.34	0.42	0.70	0.60
PSD 95 (PDZ domain)	0.58	**0.60**	0.58	0.57	0.57	0.55	0.55	0.48	0.32	0.47	0.50	0.41	0.37	0.47	0.54	0.47
GAL4 (DNA-binding domain)	0.61	0.46	0.50	**0.62**	0.60	0.58	0.60	0.53	0.26	0.47	0.52	0.43	0.42	0.47	0.59	0.41
HSP90 (ATPase domain)	**0.54**	**0.54**	**0.54**	0.51	0.52	0.52	0.48	0.45	0.03	0.34	0.44	0.26	0.22	0.33	0.49	0.43
Kanamycin kinase APH(3')-II	**0.62**	**0.62**	**0.62**	0.60	0.59	0.60	0.53	0.49	0.09	0.38	0.49	0.40	0.39	0.38	0.59	0.33
DNA methyltransferase HaeIII	**0.70**	**0.70**	0.69	**0.70**	0.68	0.68	0.64	0.64	0.12	0.54	0.64	0.50	0.49	0.54	0.69	0.44
PABP singles (RRM domain)	**0.67**	**0.67**	0.66	0.65	0.63	0.65	0.64	0.62	0.44	0.59	0.63	0.58	0.58	0.59	0.59	0.42
Ubiquitin	**0.50**	0.46	0.46	0.44	0.48	0.43	0.37	0.39	0.09	0.38	0.37	0.29	0.31	0.38	0.43	0.46
YAP1 (WW domain)	**0.64**	**0.64**	**0.64**	0.63	0.63	**0.64**	0.63	0.58	0.28	0.50	0.61	0.49	0.44	0.50	0.57	0.58

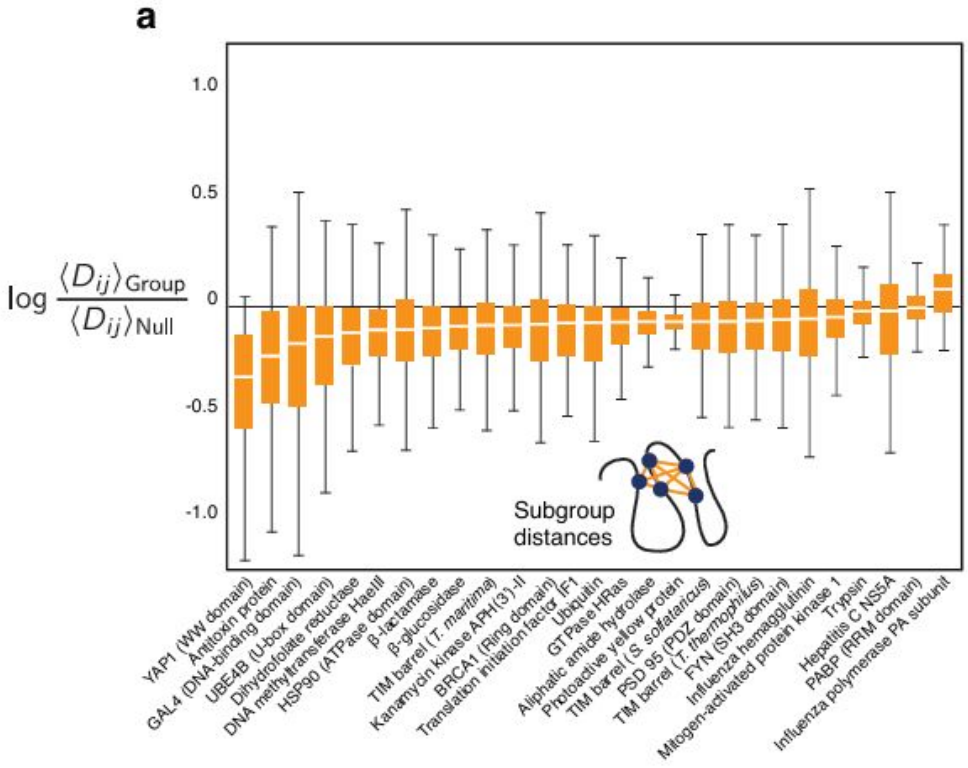
Table 1. Biologically motivated priors and Bayesian learning improve model performance.
Ablation studies of critical components of DeepSequence, showing the average Spearman ρ of predictions from five randomly-initialized models. We include combinations of components of the structured matrix decomposition and use either Bayesian approximation or Maximum a posteriori (MAP) estimation of decoder weights. These can be compared to predictions made from EVmutation (Pair) and the site-independent model (site). Inclusion is indicated with (✓), and top performing model configurations for each dataset are bolded.

It’s not just about bayesian models !

For the pairwise model of sequence families, it is well established that strongly coupled positions in the model are also close in protein 3D structure . Assessing an analogous pattern in a latent variable model is difficult.

Since these dependencies are mediated by the neural network for p(x|z) and the observed variables x are only directly affected via connections from the last hidden layer, we can focus our attention on those neural network weights. The group sparsity prior over this set of weights learns 500 soft sub-groups of positions, which can be seen as subsets of the entire sequence that are jointly influenced by the same hidden activations.

It's not just about bayesian models !



We tested if these subgroups tend to be closer in 3D structure than might be expected by chance. For each of these subgroups, we computed the average pairwise distance between positions in the group

Conclusion

1- DeepSequence outperforms both pairwise interaction models and traditional supervised predictors in predicting mutation effects. This is because it can model higher-order dependencies between residues that cannot be captured by pairwise couplings alone

2- Although DeepSequence is a deep model, it learns biologically meaningful and interpretable representations. The latent variables capture phylogenetic relationships between sequences, while sparse hidden units identify groups of residues that are structurally or functionally related. This shows that the model is not just predictive but also reveals underlying biological structure. Carefully designed priors make this interpretability possible.

Future Directions

Evolution provides a massive natural dataset, and models like DeepSequence help extract the statistical constraints underlying molecular function. Future work can combine these models with supervised data and richer biological priors to improve task-specific predictions. This approach offers a powerful framework for both understanding and engineering biological systems.