2019 Spring @ https://qdata.github.io/deep2Read/

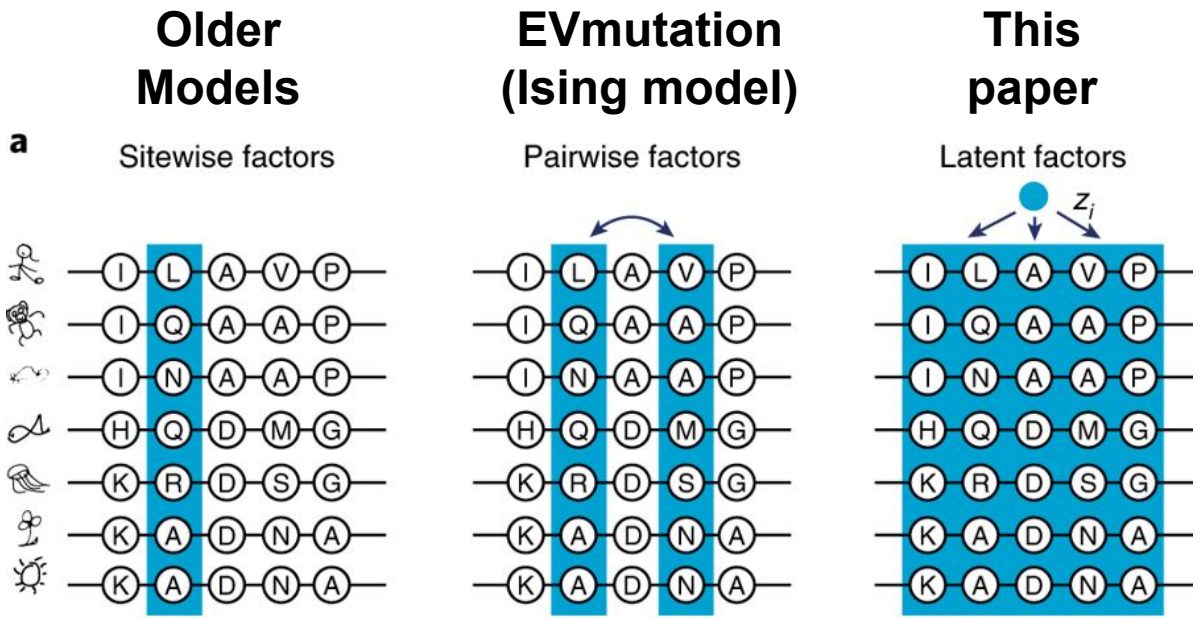
Reload this page
ARTICLES
https://doi.org/10.1038/s41592-018-0138-4
nature methods

Deep generative models of genetic variation capture the effects of mutations

Adam J. Riesselman[1,2,4], John B. Ingraham[1,3,4] and Debora S. Marks[1*]

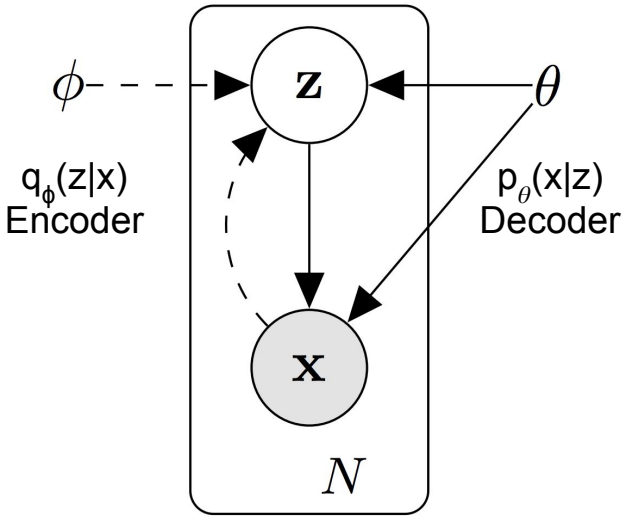
Presented by Eli Draizen
4/17/19

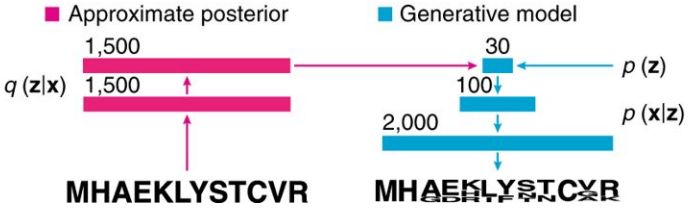
Background



- Genotype->Phenotype: How do changes in DNA present themselves in the system?
- Pairwise models cannot capture higher-order dependencies
  - Models become intractable
- **Solution:** develop nonlinear latent-variable models using Variational Autoencoders

Variational Autoencoder



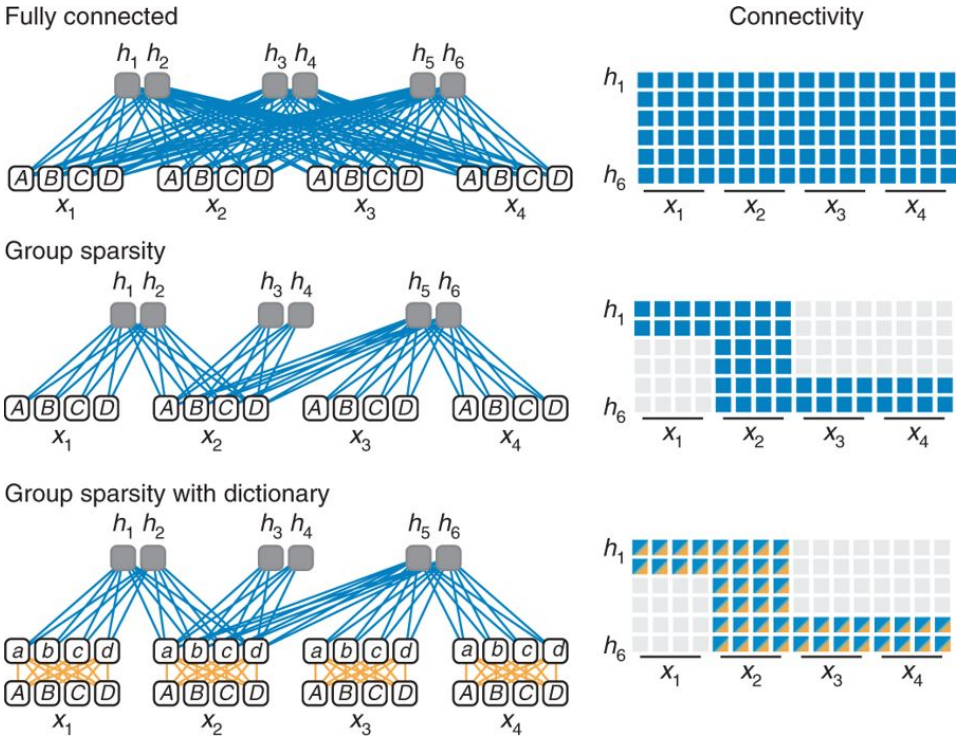


Loss: Evidence Lower Bound (ELBO), $\mathcal{L}(\boldsymbol{\phi}; \mathbf{x})$

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\phi}; \mathbf{x}) \stackrel{\Delta}{=} \mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})] - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) \| p(\mathbf{z}))$$

Specific Model

- **Group Sparsity:** Include prior to encourage small subgroups of hidden units to influence only some positions at a time
- **Dictionary:** encourage correlations b/w amino acid usage by transforming to a linear map, with shared weights
- Learns an infinite ensemble of networks since it learns distributions over weights for **p(x|z,*θ*)** with a variational approx. for global params and per-datam hidden vars.



Structured Parameterization (Dictionary)

$$\mathbf{W}^{(3,i)} = \lambda \mathbf{C}\hat{\mathbf{W}}^{(3,i)}\text{diag}(\mathbf{S}_j)$$

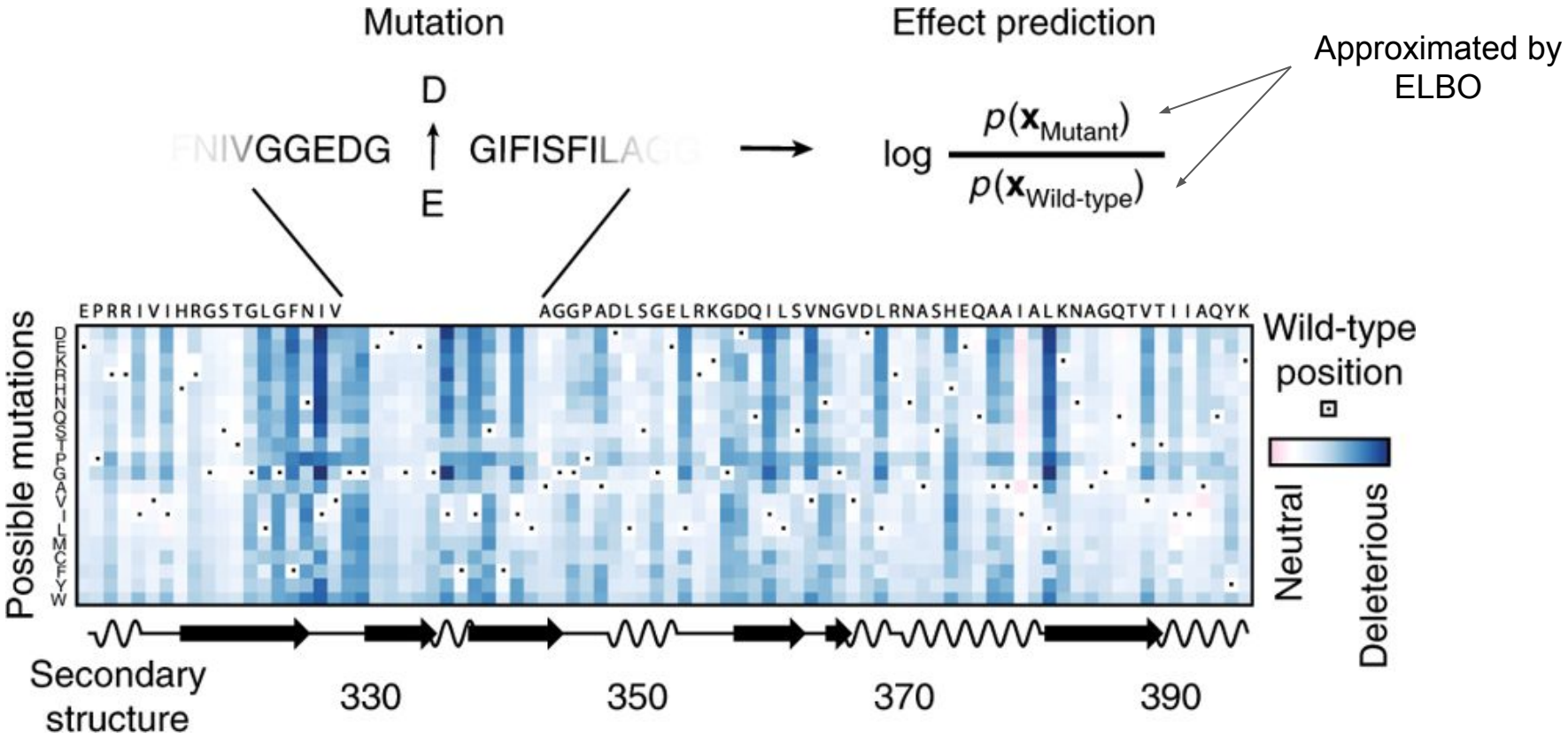
- $\mathbf{W}^{(3,i)}$: a [q x H] matrix that linearly combines H activations in final hidden $\mathbf{h}^{(2)}$ layer to q multinomial logits for different characters at position i
- **C:** matrix that captures AA correlations -- Dictionary
- **S:** matrix that gates which hidden units that can affect which positions
- λ: a scalar for the overall selective constraint across all positions

Data Sources

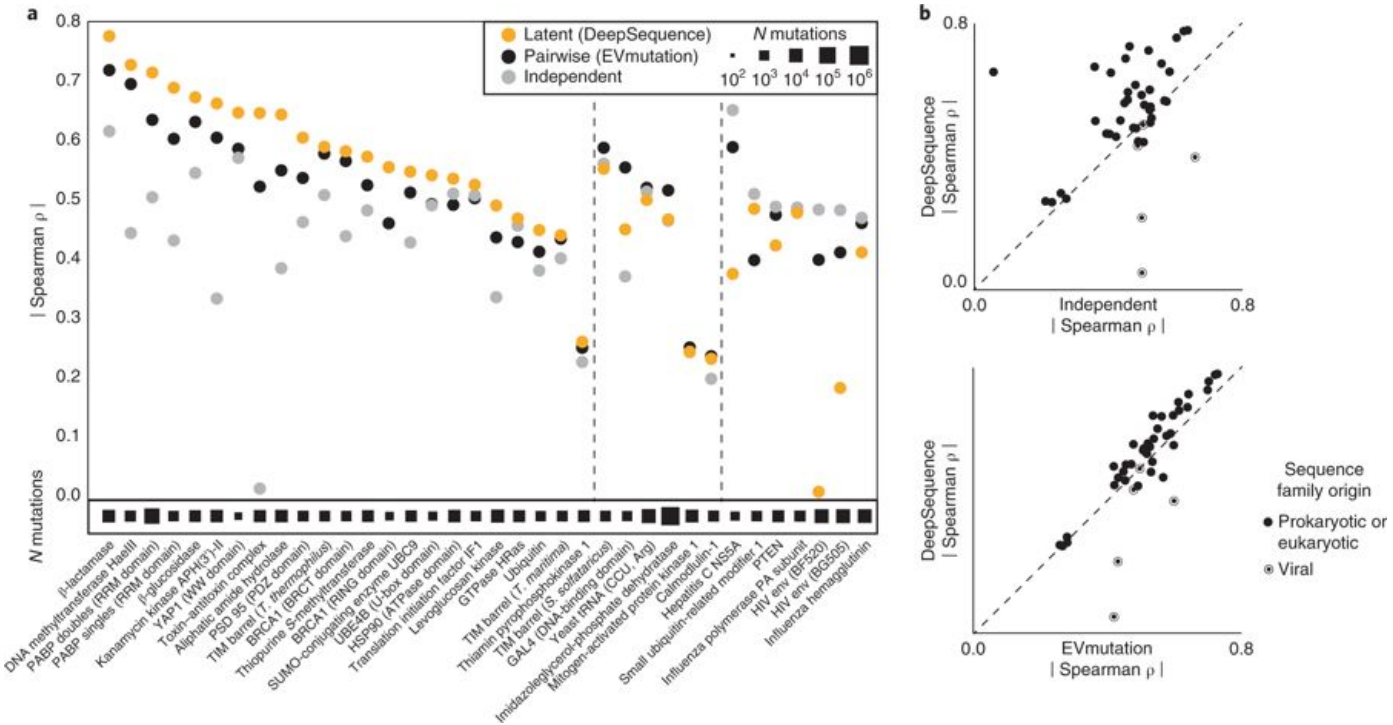
For 35 protein or RNA domains:

1) Multiple Sequence Alignments (MSA) generated from HMM searches
    a) filtered to account for redundancy and biases due to human and evolutionary sampling
2) Deep Mutational Scanning (DMS) Results

Results: PDZ Domain (from MSA)



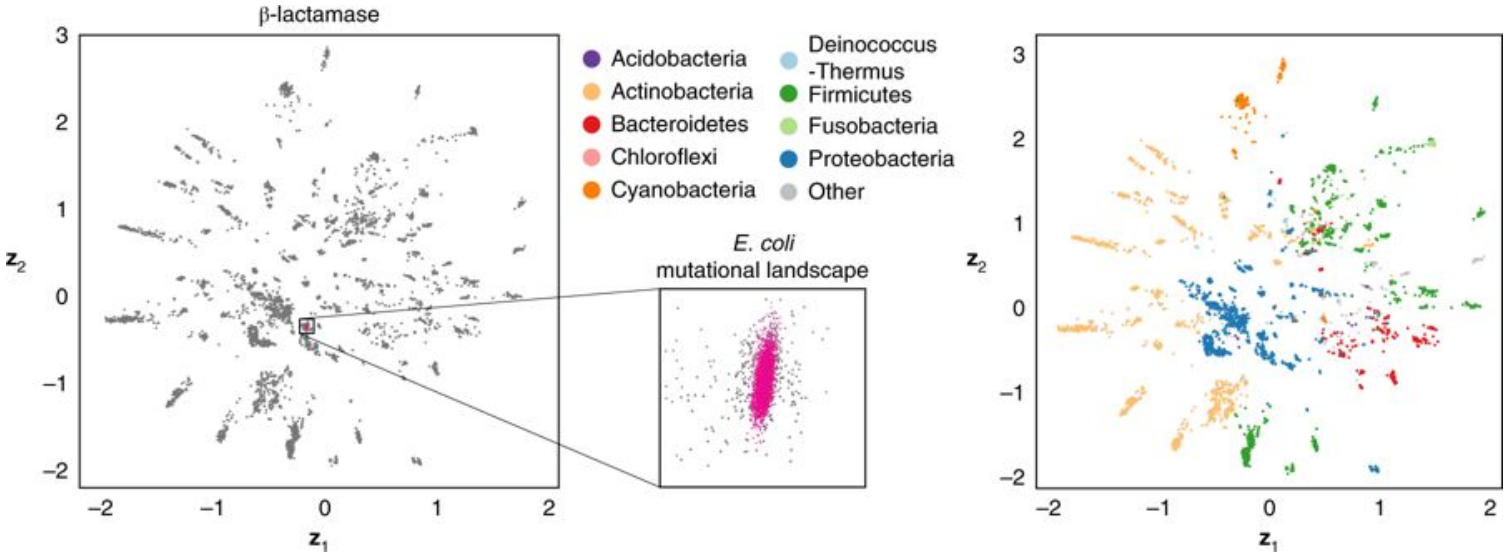
Results: All domains (MSA vs DMS)



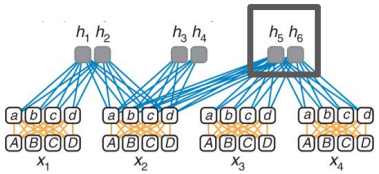
- VAE better than Pairwise and Independent
- Does not do well for viral domains
- VAE also does better when all seqs <60% id are removed

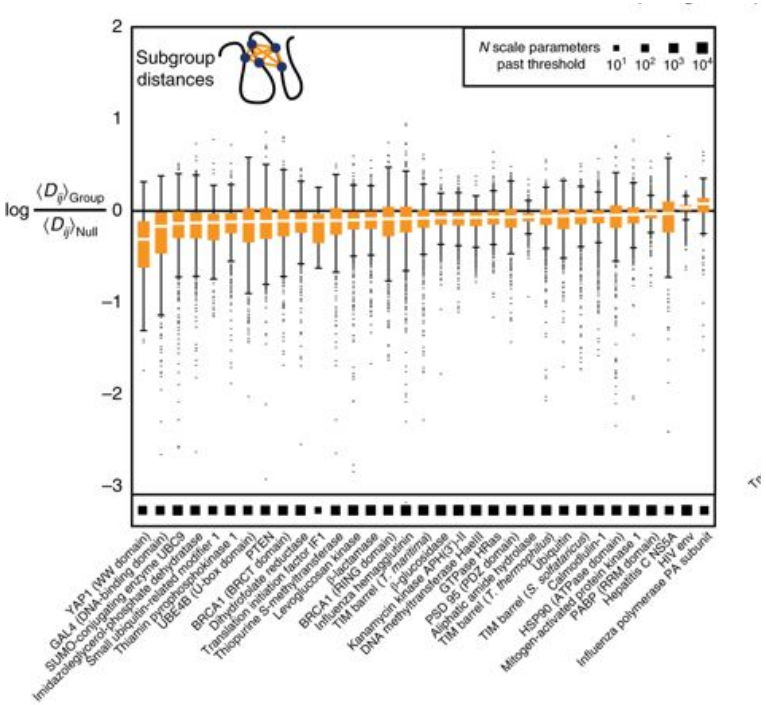
Visualizing latent space

- Fit an identical model with a 2 dimensional z, rather than 30
- Deep Mutational Scans are actually quite shallow (clustered together)

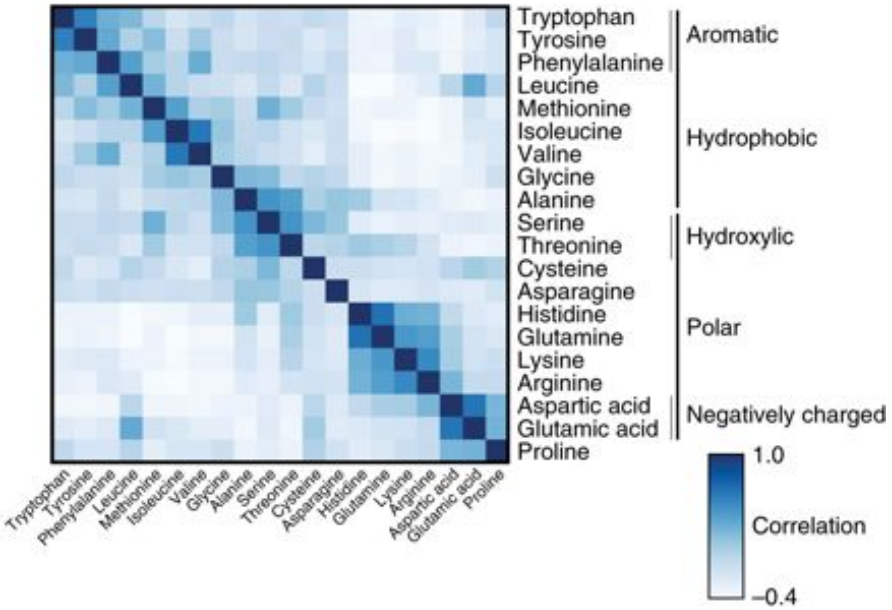


Group Sparsity Captures Residues Close in 3D



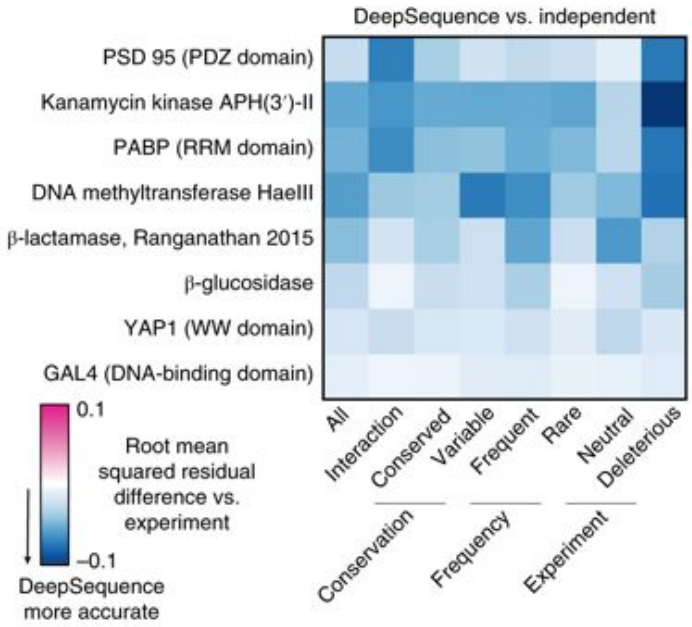


Final Weights Reflect Known Substitution Matrices



- Convolution (width -1) taken across all models

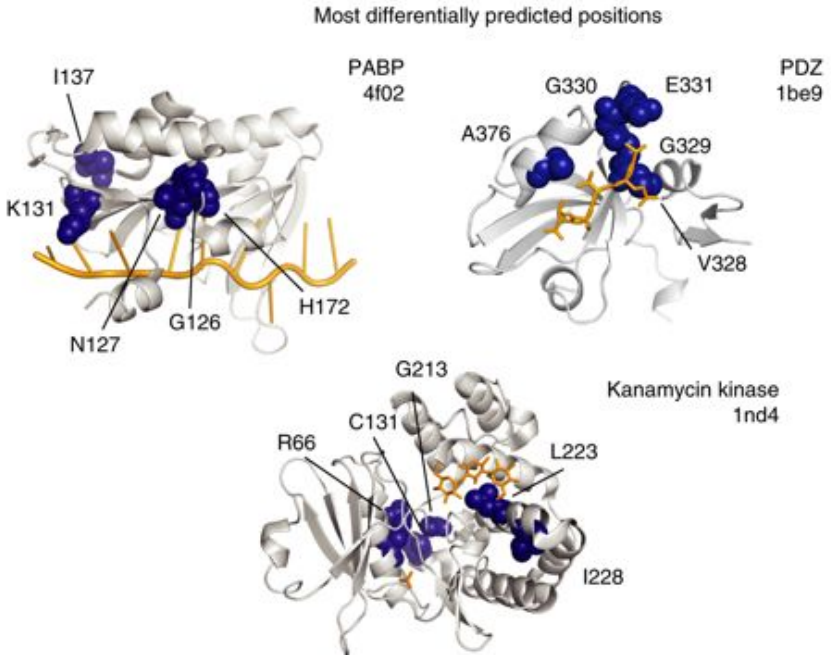
Residual Analysis: VAE better than independent



- Spearn p calculated by transforming paired data to rank quantiles, then calculating Pearson correlation b/w ranks
- Fit a least squares linear fit from normalized ranks of predictions to normalized ranks of the data
- Positive residuals from LS: over prediction of rank of experimental effect, over prediction of deleteriousness

Mutations effect functionally important residues

- Top 5 positions with greatest reduction in rank error from independent to VAE
- In PDZ, G330 is used for specificity switching



Conclusion

- The VAE predicts mutation effects better than site- and pair-wise models
- Evolutionary Info can make better predictions than Deep Mutational Scanning
- Group Sparsity highlights functionary related residues and can be used to predict 3D structure.

Does this outperform Graph NN, BiLSTM or 1D-CNN models?

How many of these sites map to PPI and PLI binding sites? Hot spots?