# Predicting the effects of mutations with deep generative models

Adam Riesselman

# Predicting the effects of mutations with deep generative models

Adam Riesselman

John Ingraham

Debbie Marks

# **Proteins** are the **workhorses** of biology

**DNA** → **Protein** → **Structure** → **Function**

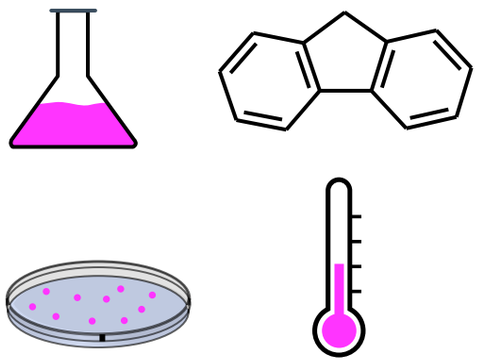
```
ATGAGTATTCAACATTTCCGTGT
CGCCCTTATTCCCTTTTTTGCGG
CATTTTGCCTTCCTGTTTTTGCT
CACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGG
GTGCACGAGTGGGTTACATCGAA
CTGGATCTCAACAGCGGTAAGAT
CCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACT
TTTAAAGTTCTGCTATGTGGCGC
GGTATTATCCCGTGTTGACGCCG
GGCAAGAGCAACTCGGTCGCCGC
ATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAG
AAAAGCATCTTACGGATGGCATG
ACAGTAAGAGAATTATGCAGTGC
TGCCATAACCATGAGTGATAACA
CTGCGGCCAACTTACTTCTGACA
ACGATCGGAGGACCGAAGGAGCT
AACCGCTTTTTTGCACAACATGG
GGGATCATGTAACTCGCCTTGAT
CGTTGGGAACCGGAGCTGAATGA
AGCCATACCAAACGACGAG...
```

**RNA**

4 different bases

```
MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW
```

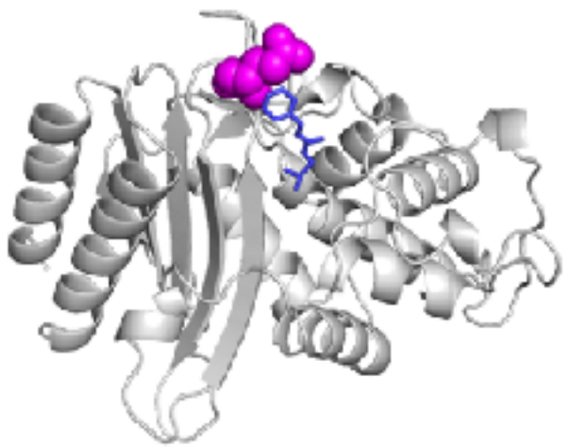
20 different amino acids



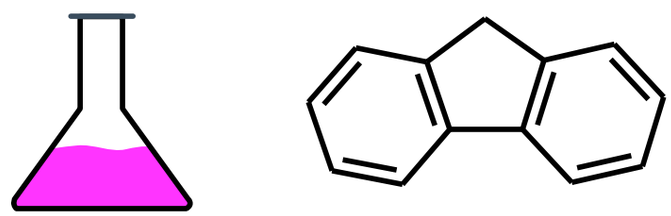
# **Mutations** impact protein **function**

MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

**E→L**

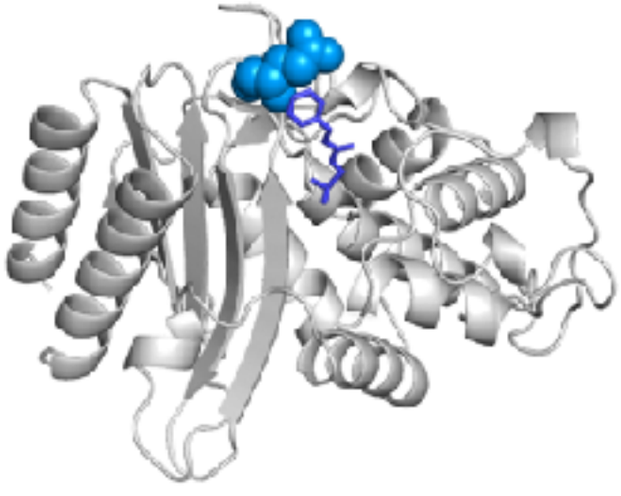


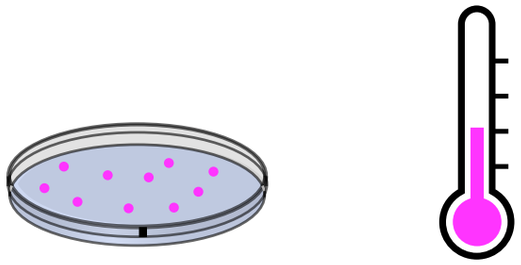
? ? ?



MSIQHFRVALIPFFAAFCLPVFA
HPETLVKVKDAEDQLGARVGYIE
LDLNSGKILESFRPEERFPMMST
FKVLLCGAVLSRVDAGQEQLGRR
IHYSQNDLVEYSPVTEKHLTDGM
TVRELCSAAITMSDNTAANLLLT
TIGGPKELTAFLHNMGDHVTRLD
RWEPELNEAIPNDERDTTMPAAM
ATTLRKLLTGELLTLASRQQLID
WMEADKVAGPLLRSALPAGWFIA
DKSGAGERGSRGIIAALGPDGKP
SRIVVIYTTGSQATMDERNRQIA
EIGASLIKHW

**E→I**





? ? ?

**Sequence** **Structure** **Function**

# **Mutation effect prediction** is **important**

| **Understanding disease** | **Biomedicine** | **Bioengineering** |
|---|---|---|

**“Does this mutation cause cancer?”**

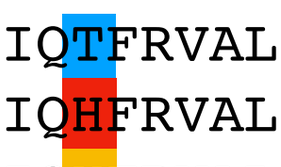
**“Is this antibody stable in a patient?”**

**“Will this protein digest wood better for biofuels?”**

# State of art methods for **measuring mutation effects**

**Input**
Pooled mutated sequences

MSIQTFRVALIPFF
MSIQHFRVALIPFF
MSIQYFRVALIPFF
MSIQRFRVALIPFF
MSIQVFRVALIPFF
MSIQCFRVALIPFF
MSIQGFRVALIPFF
MSIQIFRVALIPFF
MSIQLFRVALIPFF
MSIQVFRVALIPFF



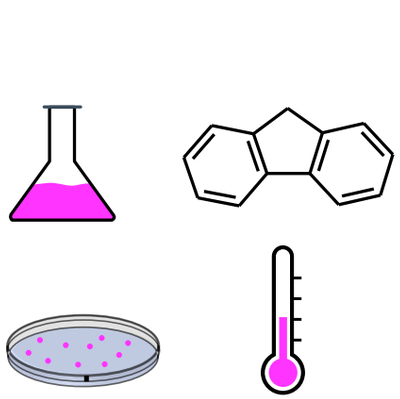
selection

MSIQHFRVALIPFF
MSIQHFRVALIPFF
MSIQYFRVALIPFF
MSIQHFRVALIPFF
MSIQYFRVALIPFF
MSIQHFRVALIPFF
MSIQHFRVALIPFF
MSIQHFRVALIPFF
MSIQYFRVALIPFF
MSIQHFRVALIPFF

**Output:**
Pooled mutated sequences

Compare ratio of sequences before and after selection

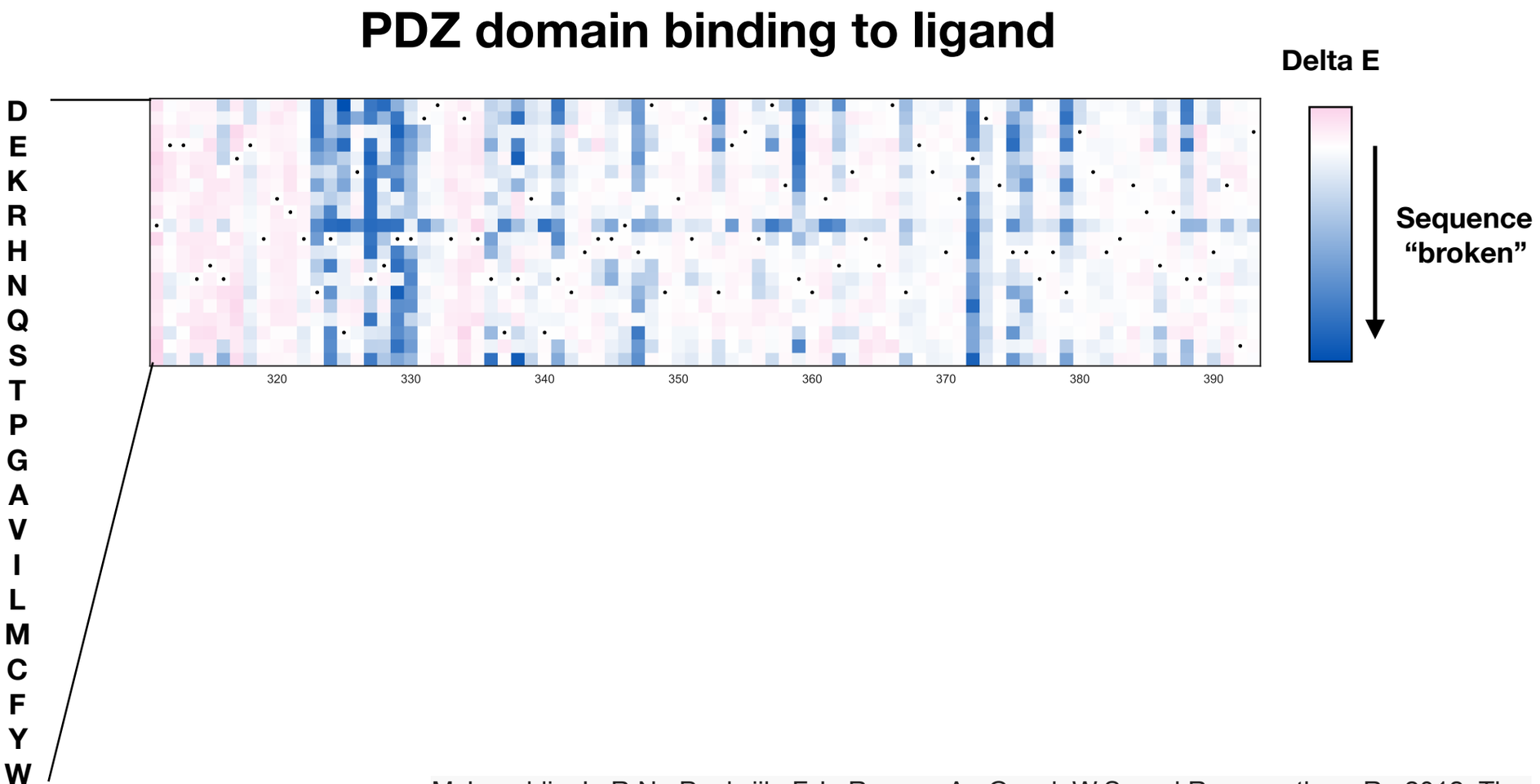
# State of art methods for **measuring mutation effects**





0



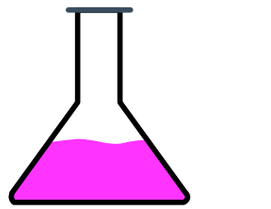


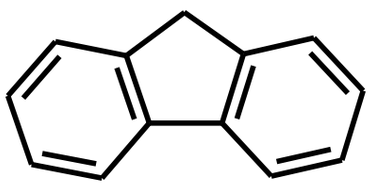

McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R., 2012. The spatial architecture of protein function and adaptation. Nature, 491(7422), pp.138-142.

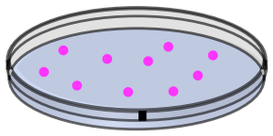
# **Understanding** the effects of **mutations** is **important**

**?**

**R**
**MHAE↑LYSTCVR**
**K**





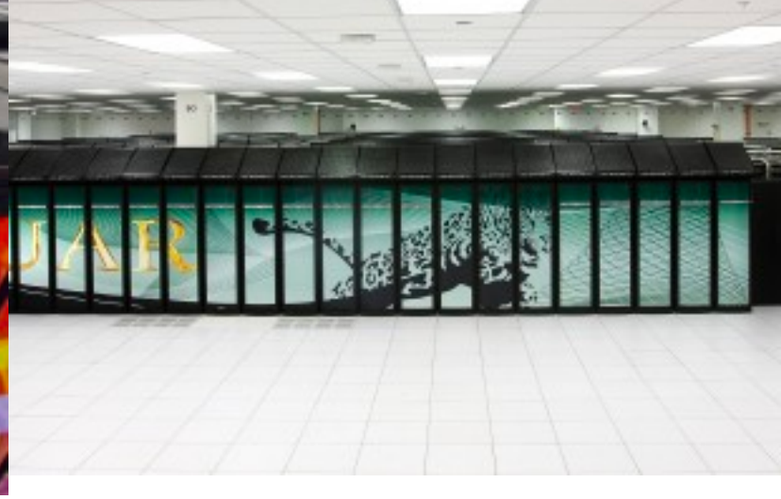


**$10^2$-$10^6$ Different experiments**

**Slow**

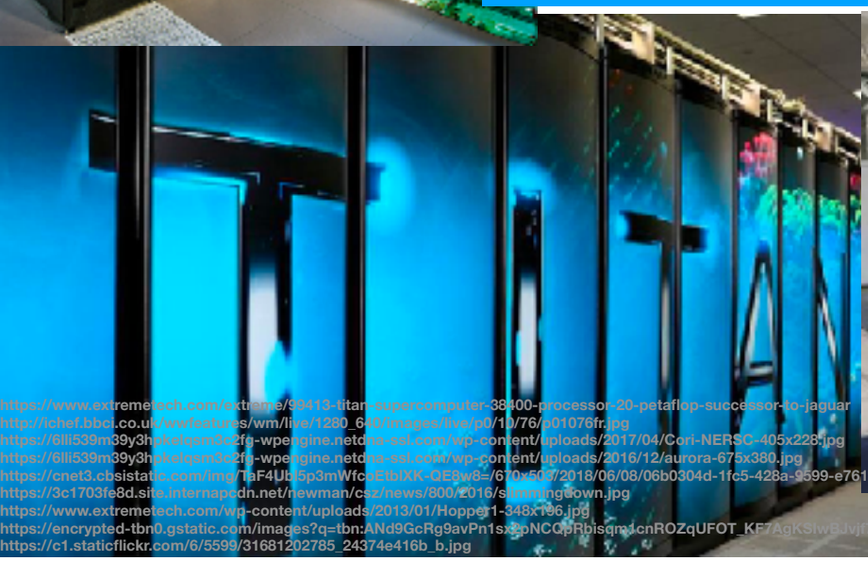
**Expensive**





# There has to be a better way…

# with computers!



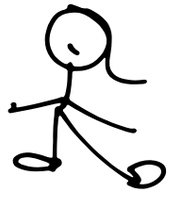




https://www.extremetech.com/extreme/99413-titan-supercomputer-38400-processor-20-petaflop-successor-to-jaguar
http://ichef.bbci.co.uk/newsbrands/wm/live/1280_640/images/live/p0/10/76/p0107 6fr.jpg
https://6lli539m39y3hpkelqsm3c2fg-wpengine.netdna-ssl.com/wp-content/uploads/2017/04/Cori-NERSC-405x228.jpg
https://6lli539m39y3hpkelqsm3c2fg-wpengine.netdna-ssl.com/wp-content/uploads/2016/12/aurora-675x380.jpg
https://cnet3.cbsistatic.com/img/TaF4Ubl5p3mWfceEtblXK-QE8w8=/670x503/2018/06/08/06b0304d-1fc5-428a-9599-e76140cc03ff/summit-supercomputer-long-shot.jpg
https://3c1703fe8d.site.internapcdn.net/newman/csz/news/800/2016/shimingtown.jpg
https://www.extremetech.com/wp-content/uploads/2013/01/Hopper1-348x196.jpg
https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRg9avPn1sxZpNCCqRbisqm1cnROZqUFOT_KF7AgKSlwBJvjf7Q
https://c1.staticflickr.com/6/5599/31681202785_24374e416b_b.jpg

# **Lots** of other **examples** of **Protein X** are **available**

ADRLYMTKIHHEFEGD

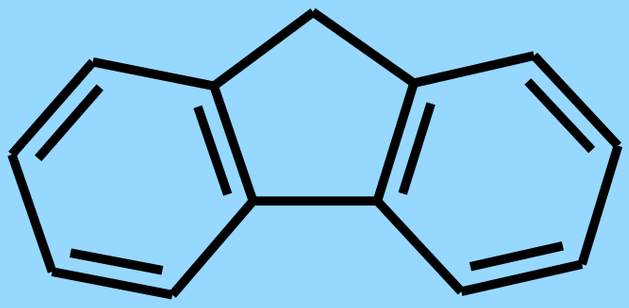


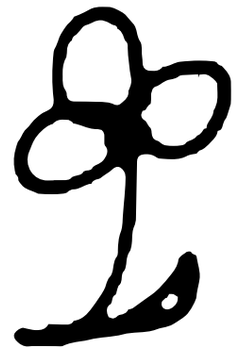
ADRLYMTKIHHQFDGD

ADRLYMTKIHHEFEGD

ADRLYLTQIRNKFKGD

Protein X





TSKMYITKIGQEFEGD

All are **functional, homologous examples** of Protein X

# **Lots** of other **examples** of **Protein X** are **available**

**Sequences** are found in public **genome databases**.

Natural **evolution** is an **experiment**, in **parallel**.

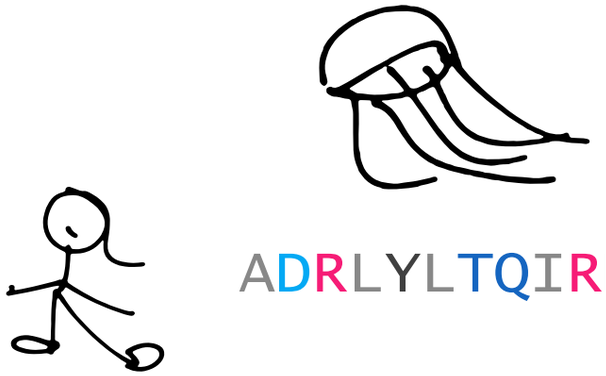
**Assumption:**

**Present in database: Tolerated**

**Not in database: Deleterious**

All are **functional, homologous examples** of Protein X

# **Natural sequences** can be grouped into **families** via **alignments**



ADRLYLTQIRNKFKGD

ADRLYMTKIHHQFDGD

ADRLYMTKIHHEFEGD

ADRLYMTKIHHEFEGD

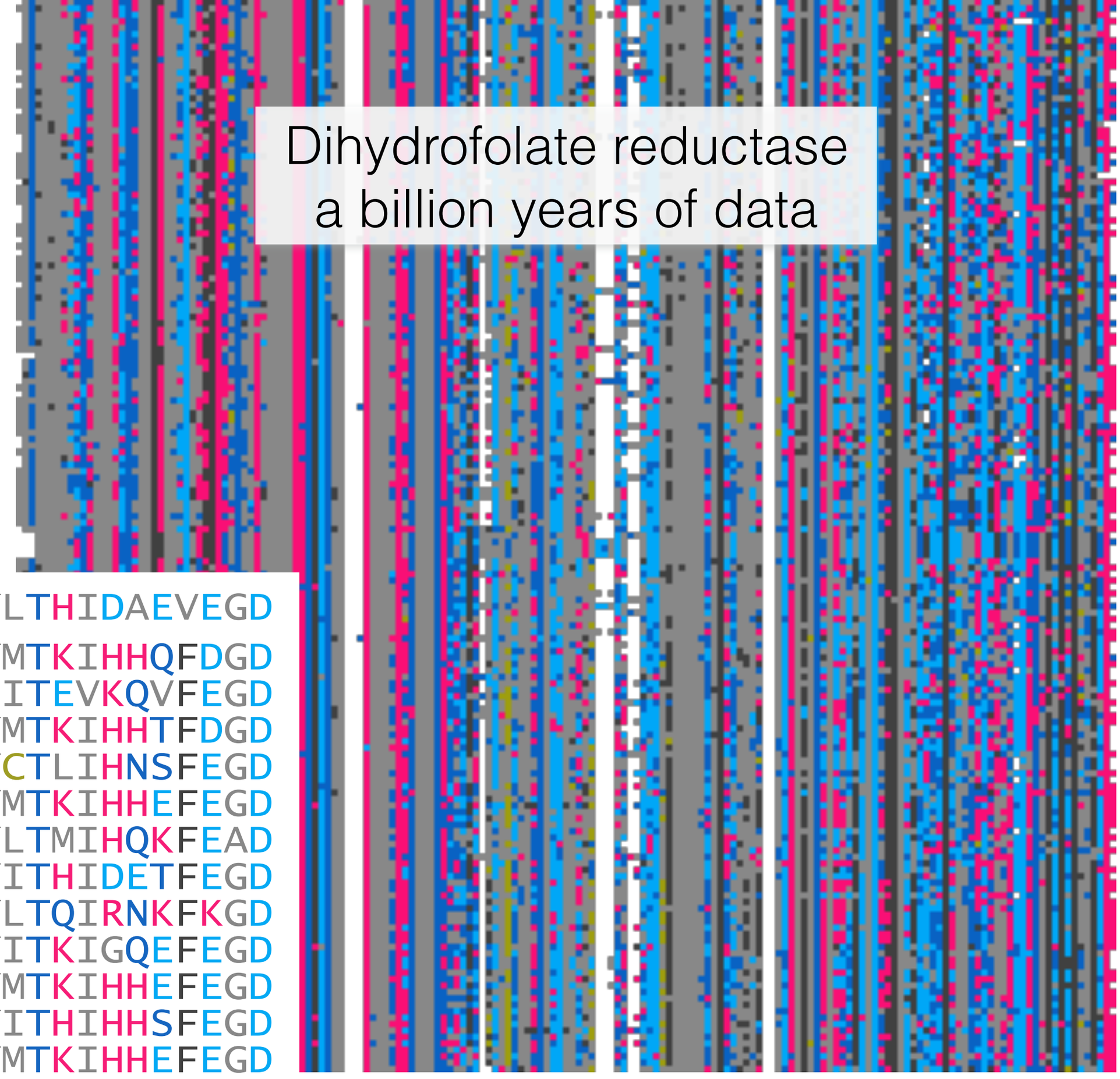
TSKMYITKIGQEFEGD

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
TSKMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD

**Examples of sequence**

**Positions with shared ancestry**

# Dihydrofolate reductase
# a billion years of data



AQKLYLTHIDAEVEGD

ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
TSKMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD

Courtesy: John Ingraham

# **Natural sequences** have been evolved to be **functional**

**x**

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
TSKMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD

**Fit generative model** →

**P(x)**

**Generative model** captures **functional constraints**

# **Mutation effect prediction** with an unsupervised model

1) Infer a **generative model** of the family

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD

$$p(\mathbf{x}|\boldsymbol{\theta})$$

2) Compute **Log Ratio** for each mutant

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

**“How much does this mutation look like what we’ve seen in nature?”**

# **Mutation effect prediction**
with an unsupervised model

1) Infer a **generative model** of the family

2) Compute **Log Ratio** for each mutant

**Uses public data (effectively free)**

**Fast**

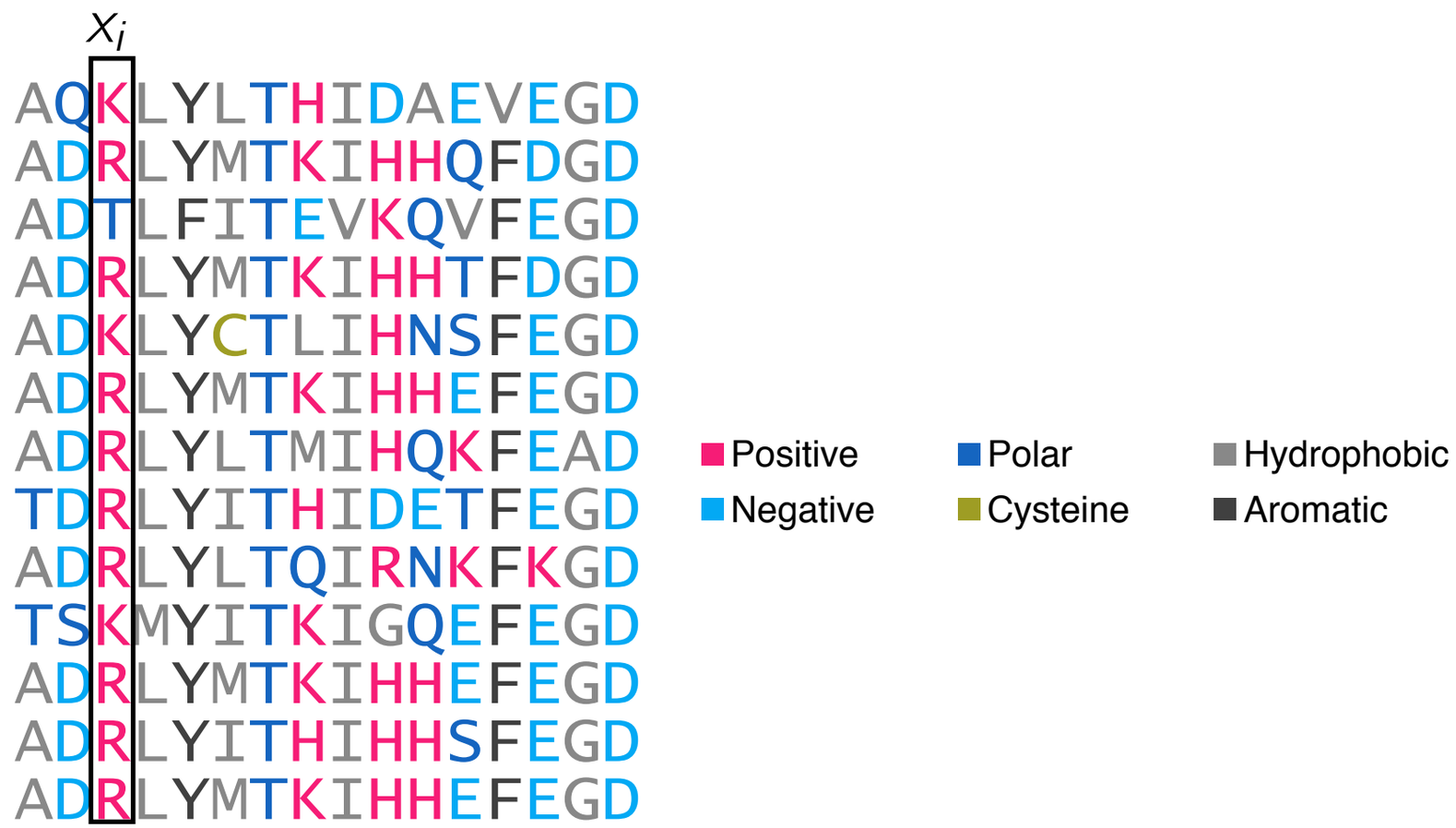
**Works on almost any protein**

**Accurate**

**“How much does this mutation look like what we’ve seen in nature?”**

# First-pass mutation predictors model **evolutionary conservation**

Amino acid i




Product of **site** factors

$$P(\mathbf{x}) = p_1(x_1)p_2(x_2)\cdots p_L(x_L)$$

R → K Neutral

R → I Deleterious

# How to capture **interactions**?


Pairwise undirected model

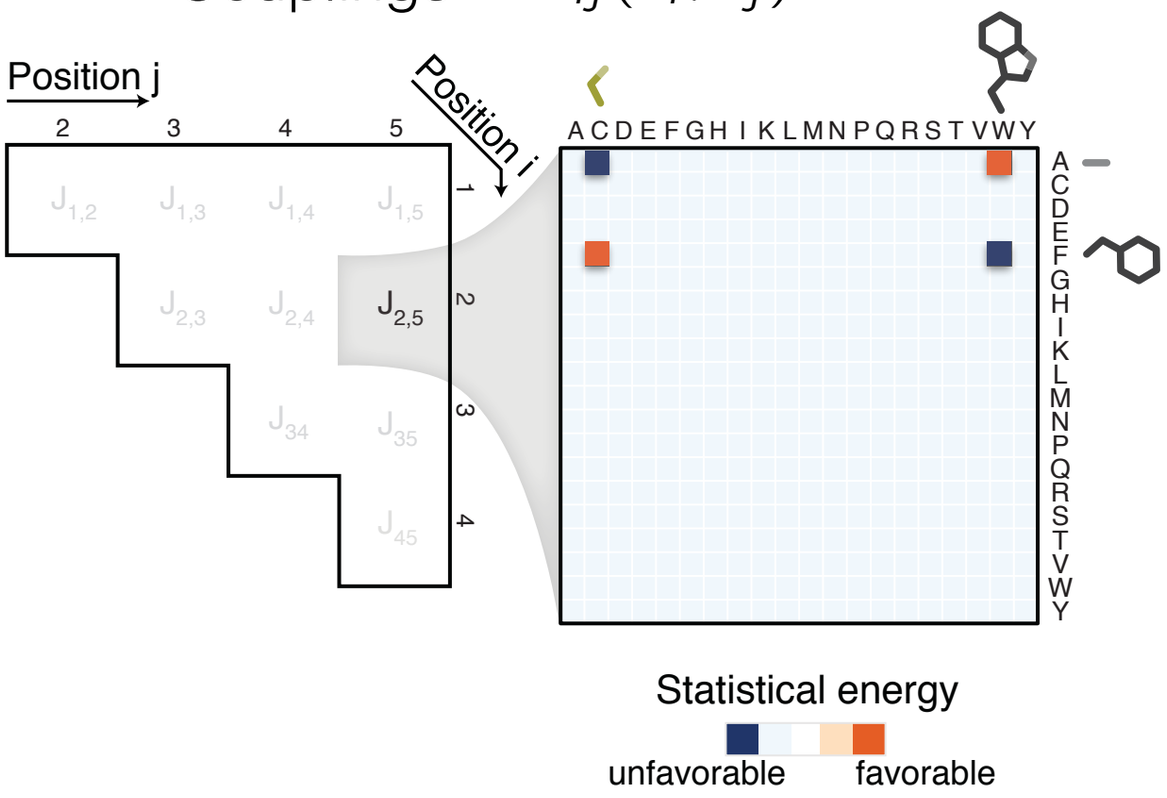
$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i<j} J_{ij}(x_i, x_j) + \sum_i h_i(x_i)\right)$$

a.k.a.
Markov Random Field
a.k.a.
Ising (Potts) model
a.k.a
Multinomial logistic regression

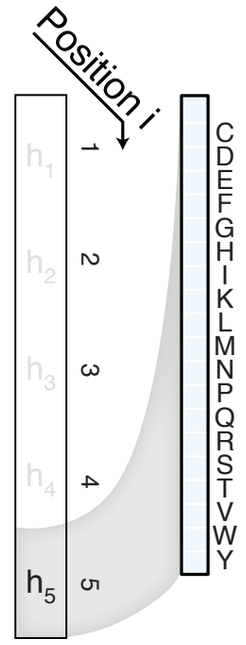
# Undirected graphical model for sequences parameterizes **pairs of letters** at **pairs of positions**

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i<j} J_{ij}(x_i, x_j) + \sum_i h_i(x_i)\right)$$

Couplings $J_{ij}(x_i, x_j)$

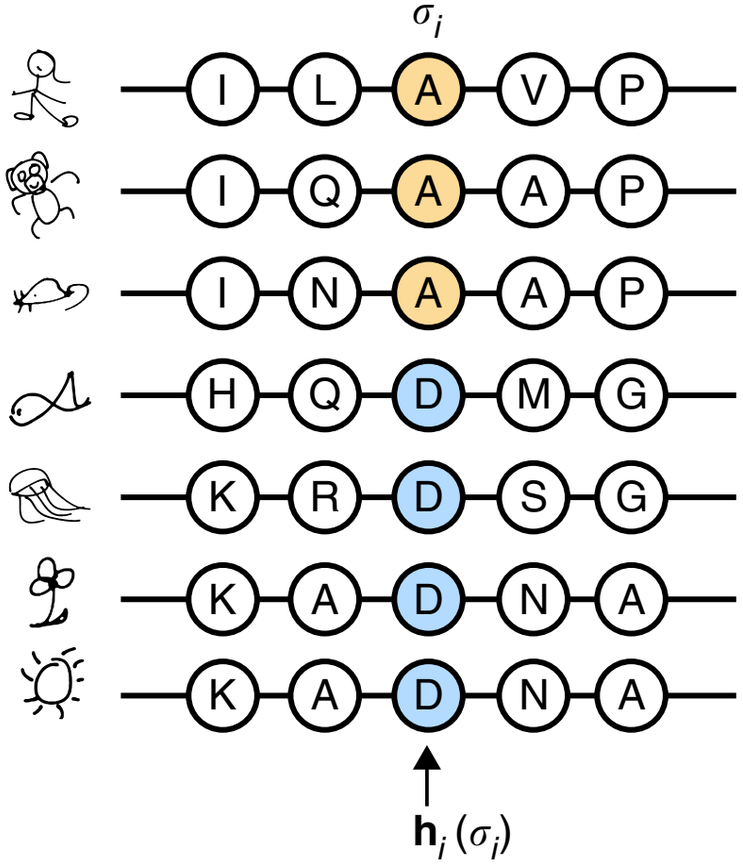



Biases $h_i(x_i)$

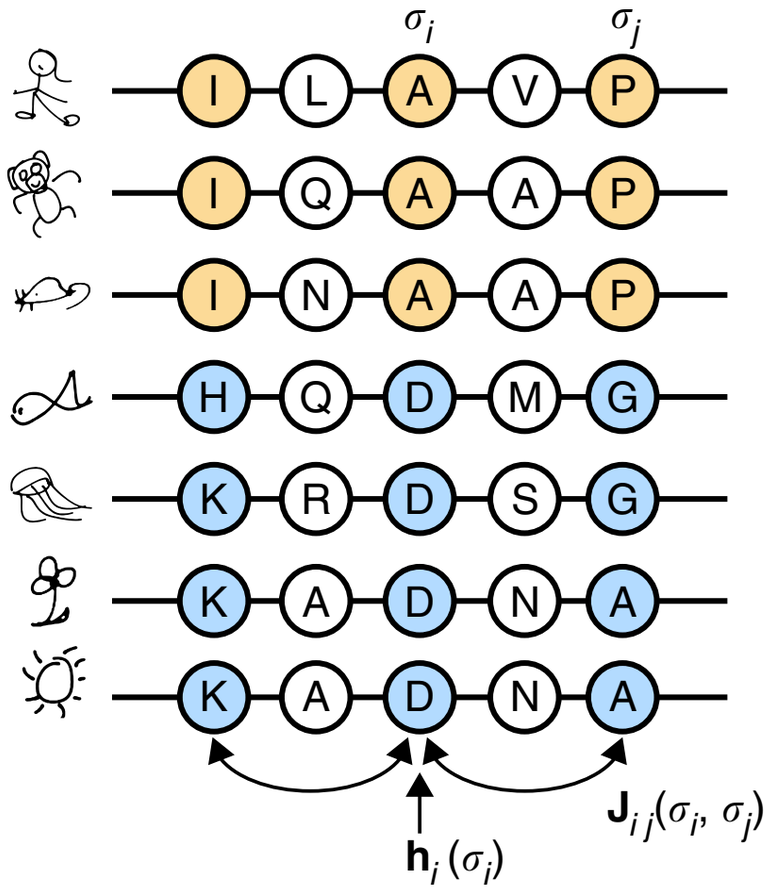


# Pairwise >> sitewise.
# Should we stop there?

Independence




Pairwise model

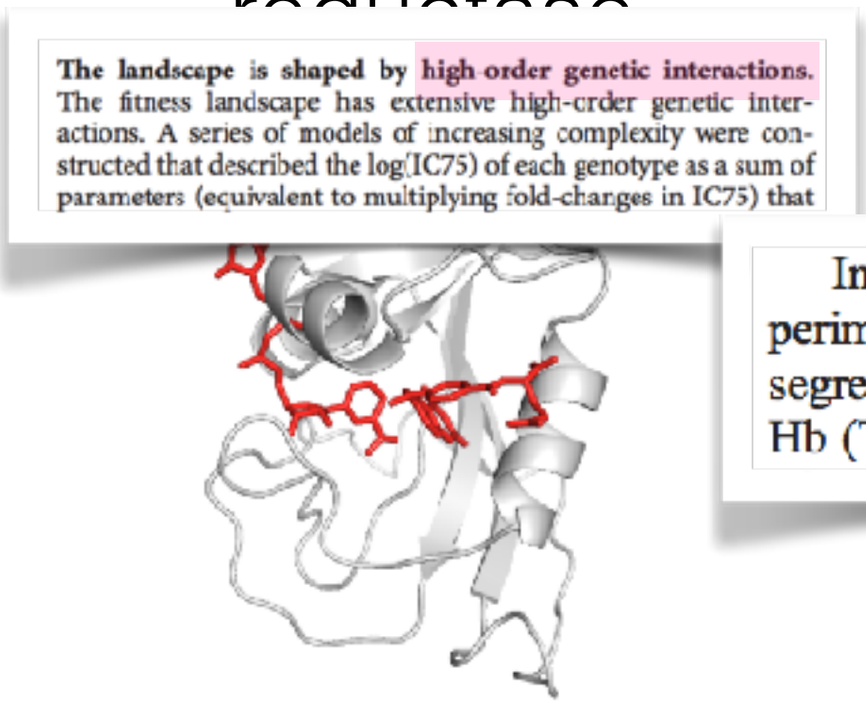



Higher order?

. . .

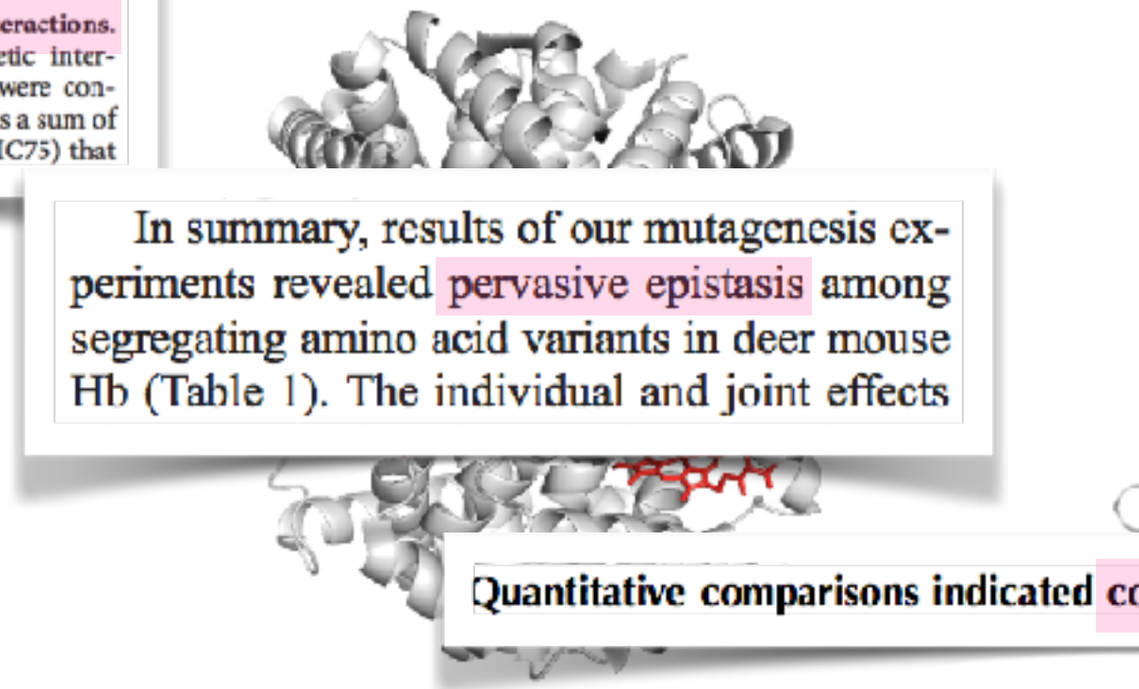
# **Pairwise** interactions insufficient for mutation effects in **proteins**

Dihydrofolate reductase



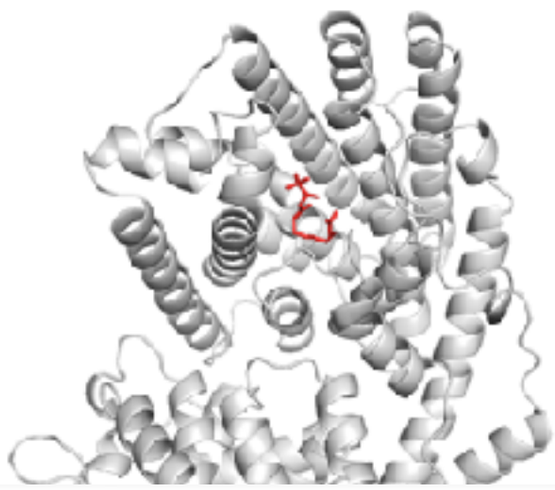

Drug resistance

Hemoglobin




Physiology

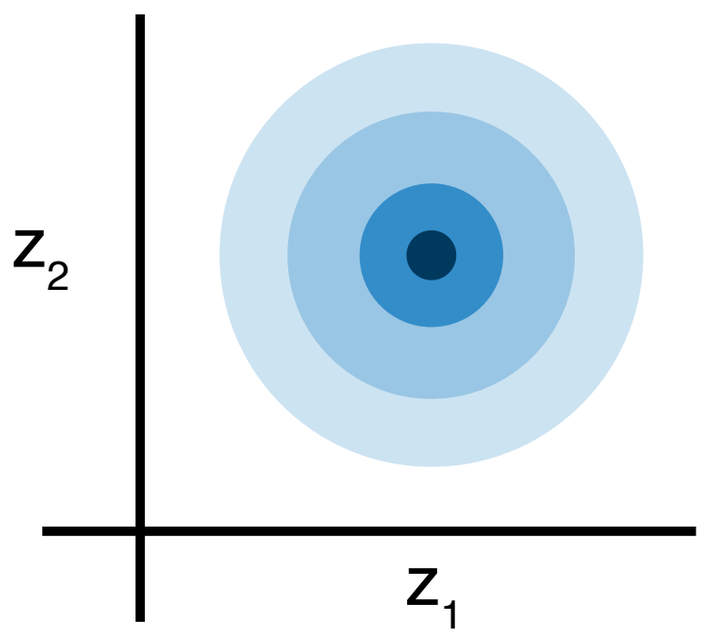
Terpene Synthase




Synthetic biology

# Neural networks make powerful latent variable models

Latent variable
$p(\mathbf{z})$



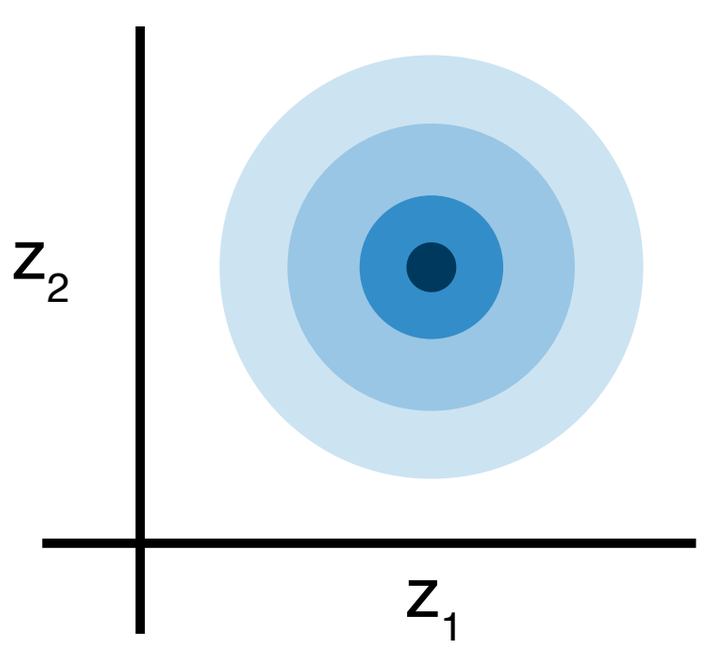

Nonlinear mapping
$p(\mathbf{x}|\mathbf{z})$

Sequence family
**X**

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
TSKMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD

# p(**x**) for latent variable models is generally **intractable**

Latent variable
$p(\mathbf{z})$




Nonlinear mapping
$p(\mathbf{x}|\mathbf{z})$

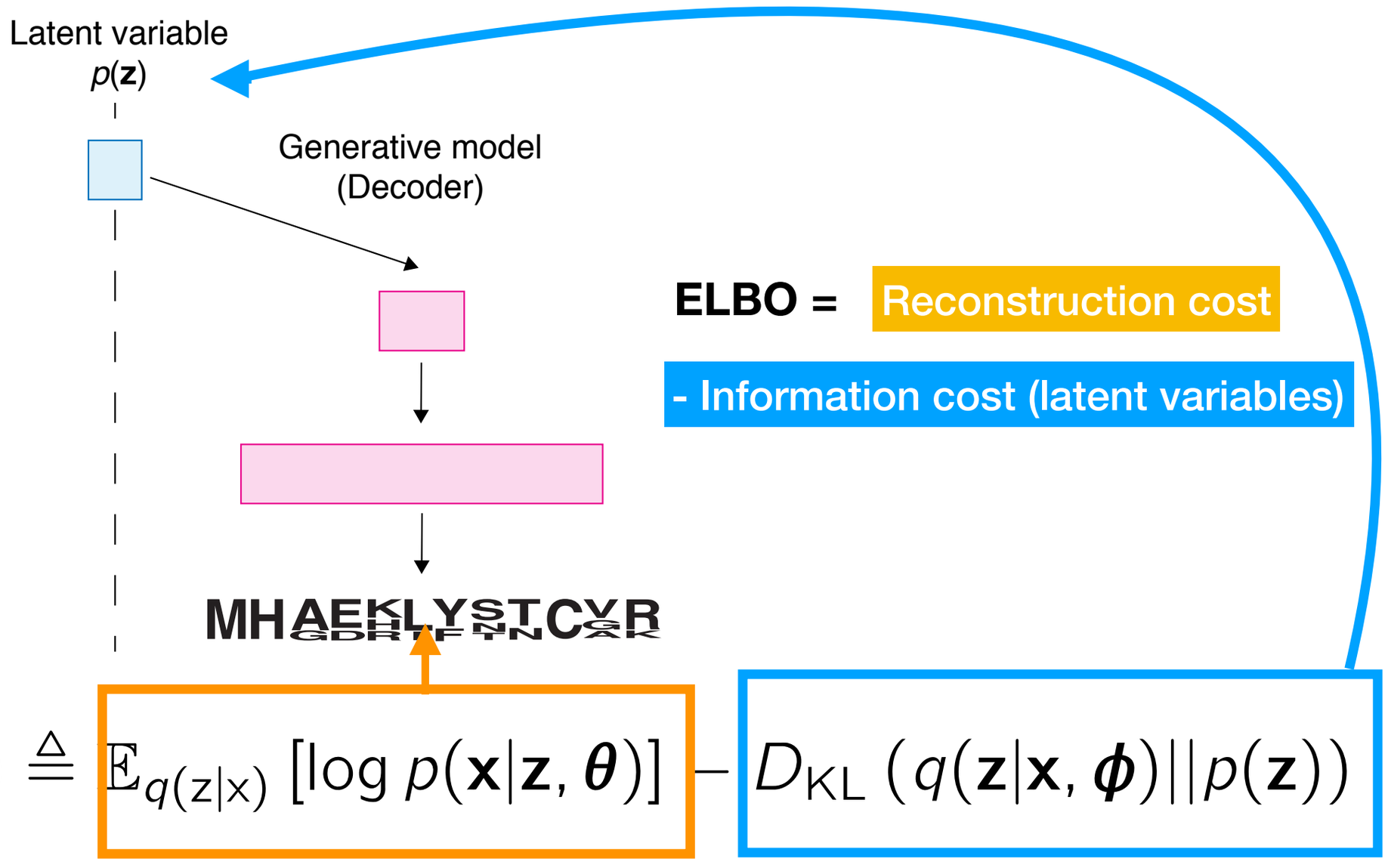
Sequence family
**X**

AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
TSKMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD

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

have to account for all possible **z** for each **x**

# **Variational autoencoders** provide a tractable lower bound on p(**x**)

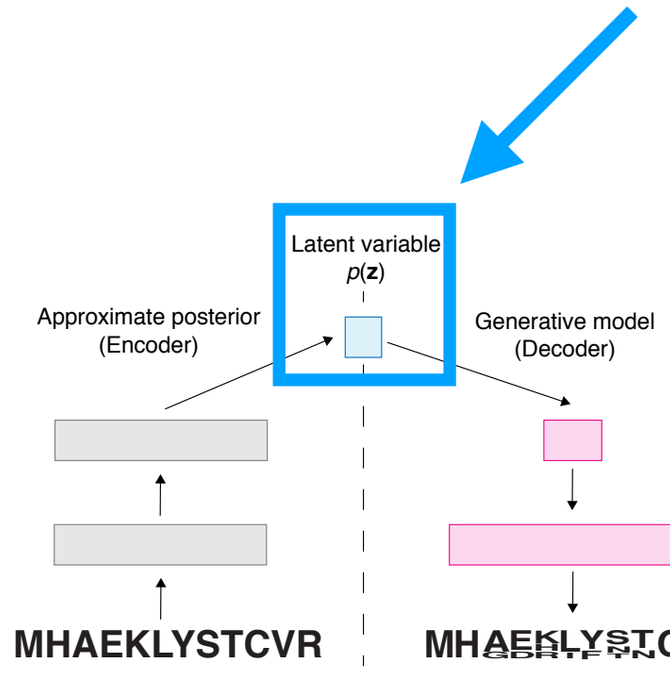




Kingma, D.P. and Welling, M., (2013). **Auto-encoding variational bayes**.

Rezende, D.J., Mohamed, S. and Wierstra, D., (2014). **Stochastic backpropagation and approximate inference in deep generative models**.

# **Latent variables** are generated for **each sequence** in alignment




Update 10

**β-lactamase sequence family**

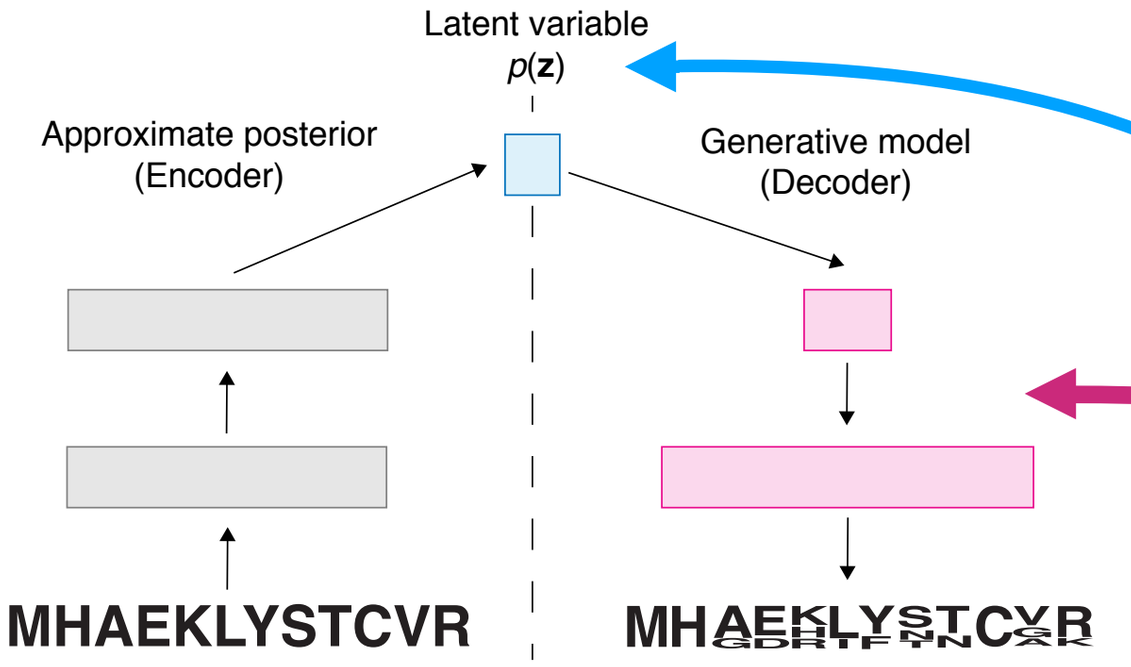
- Acidobacteria
- Actinobacteria
- Bacteriodetes
- Chloroflexi
- Cyanobacteria
- Deinococcus-Thermus
- Firmicutes
- Fusobacteria
- Proteobacteria
- Other

# **Variational inference** on **decoder** weights **prevents overfitting**

**ELBO =** Reconstruction cost

- Information cost (latent variables)

- Information cost (model parameters)

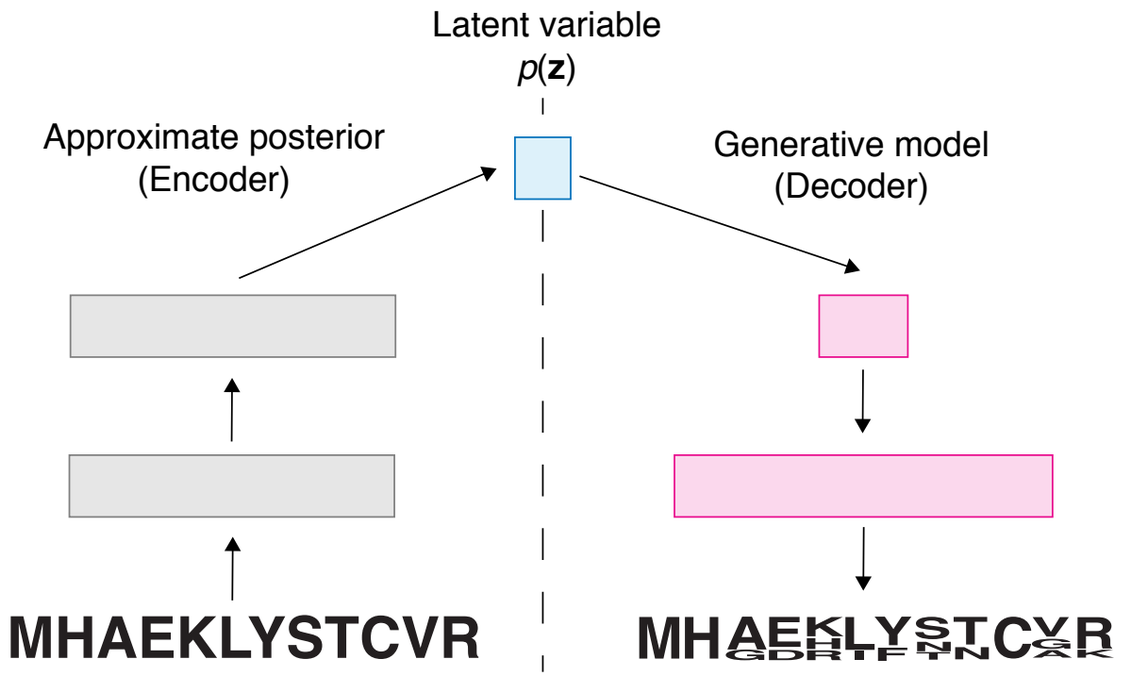



$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \mathbf{X}) = N \, \mathbb{E}_{\mathbf{x} \in \mathbf{X}} \left[ \mathbb{E}_{q(\boldsymbol{\theta}) q(\mathbf{z}|\mathbf{x})} \left[ \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) \right] - D_{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})) \right] - \sum_{\boldsymbol{\theta}^{(i)}} D_{KL}(q(\boldsymbol{\theta}^{(i)}) || p(\boldsymbol{\theta}^{(i)}))$$

Kingma, D.P. and Welling, M., (2013). **Auto-encoding variational bayes**.

# **Mutation prediction** with a variational autoencoder

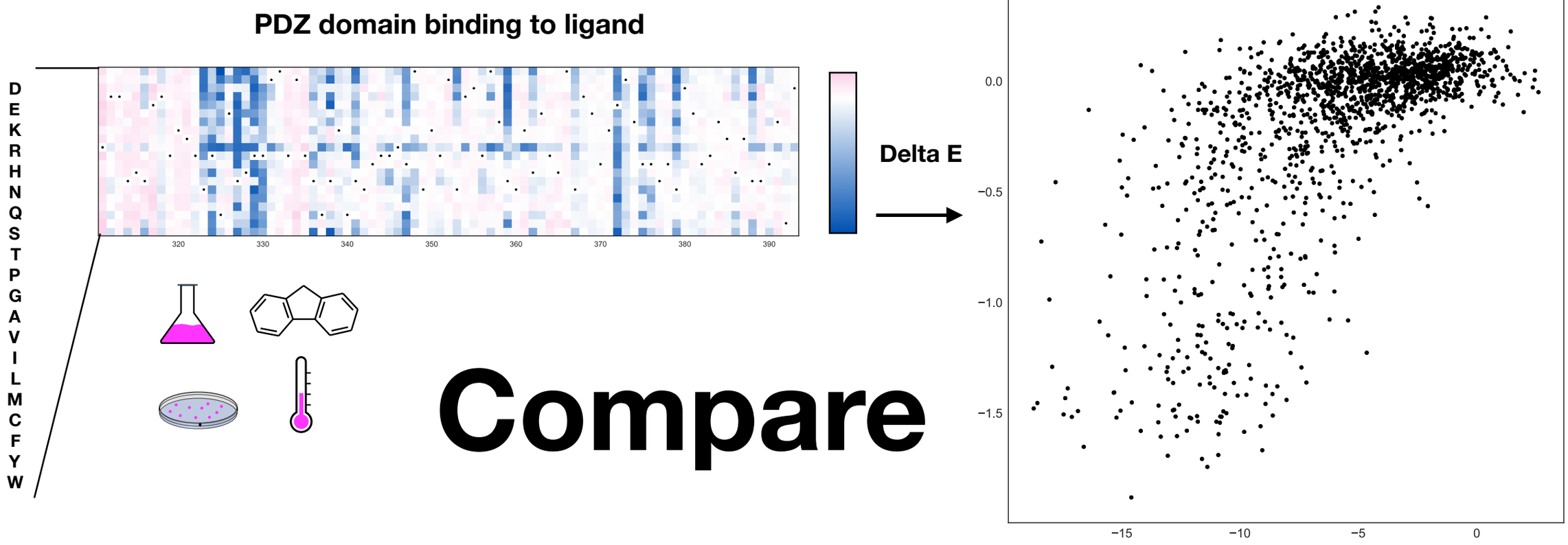
1) Infer a **generative model** of the family




2) Approximate **Log Ratio** with difference in ELBO

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

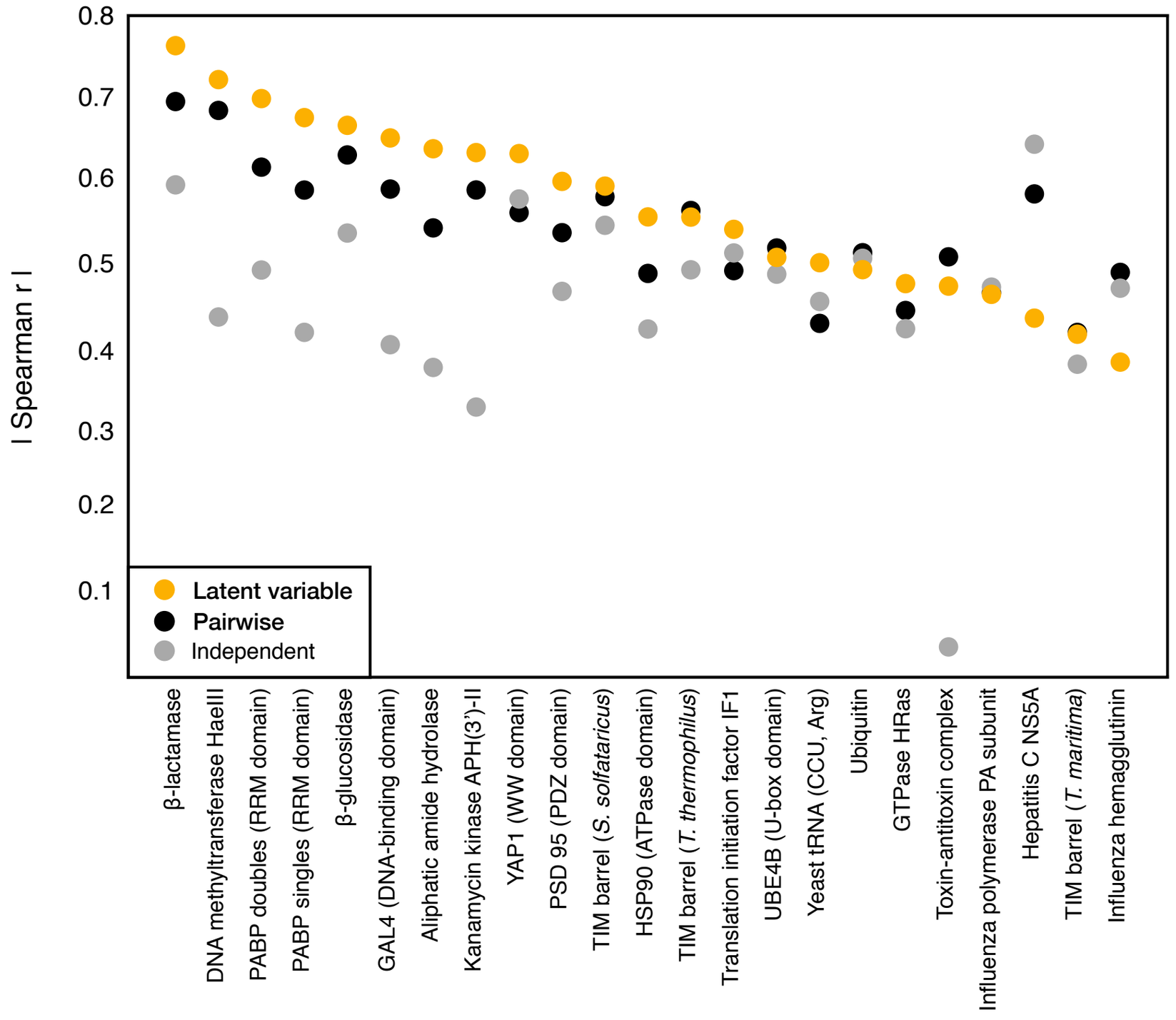
# To evaluate performance, we collected ~30 **saturation mutagenesis experiments**



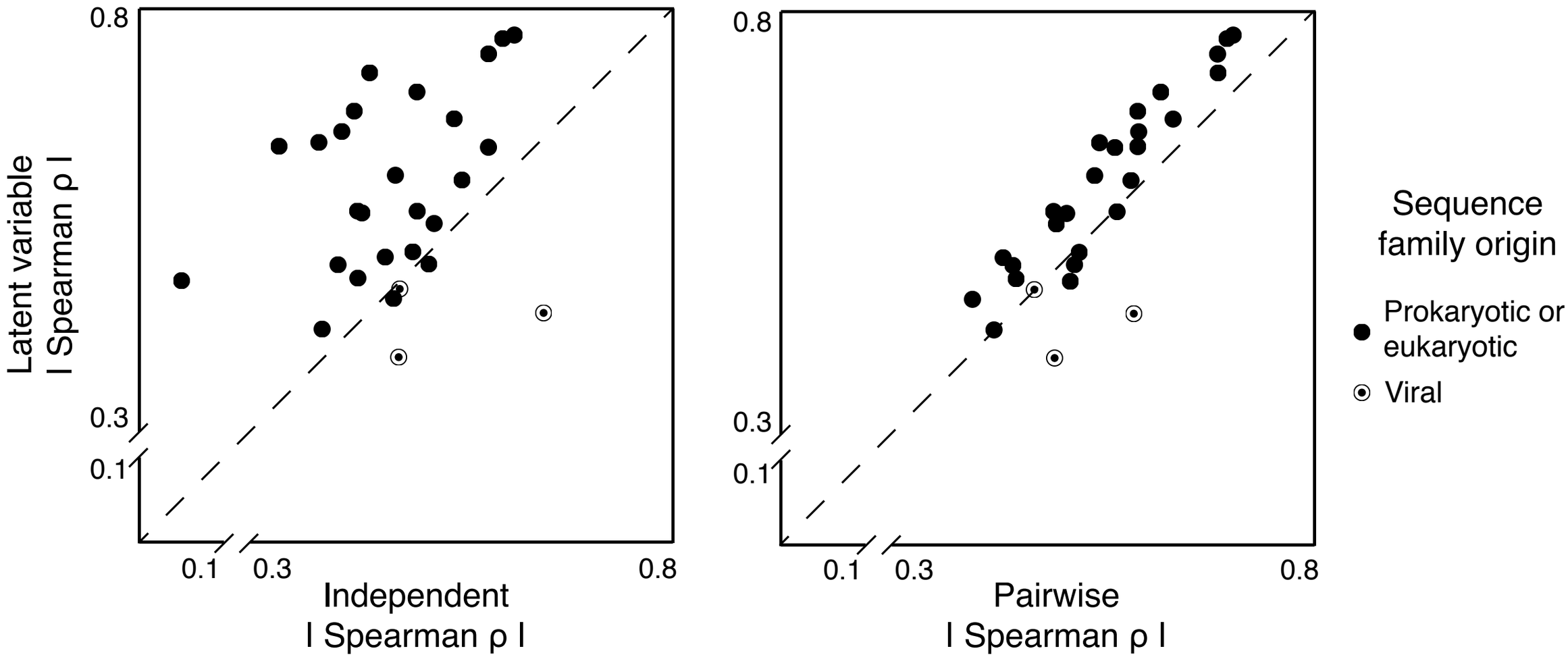

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S. and Ranganathan, R., 2012. The spatial architecture of protein function and adaptation. Nature, 491(7422), pp.138-142.

# **Latent variable** model is **more predictive** than pairwise model



# **Deeper** sequence alignments lead to **more predictive** models



Can we **interpret** our model by building **biology** into the **components**

# Encoding **biological knowledge** in a structured matrix with **parameter sharing**

**z** = latent variable
**h** = hidden vector
**x** = sequence
**W,b** = weights

Categorical VAE decoder

$$\mathbf{h} = \texttt{MLP}(\mathbf{z})$$

$$p(x_i|\mathbf{z}) = \texttt{Softmax}(\mathbf{W}^{(i)}\mathbf{h} + \mathbf{b}^{(i)})$$

Parameterized by:

$$p(x_i|\mathbf{z}) = \texttt{Softmax}\left(\mathbf{D}\tilde{\mathbf{W}}^{(i)}\left(\texttt{Sigmoid}\left(\mathbf{s}^{(i)}\right) \odot \mathbf{h}\right) + \mathbf{b}^{(i)}\right)$$

**Dictionary** shared across all positions

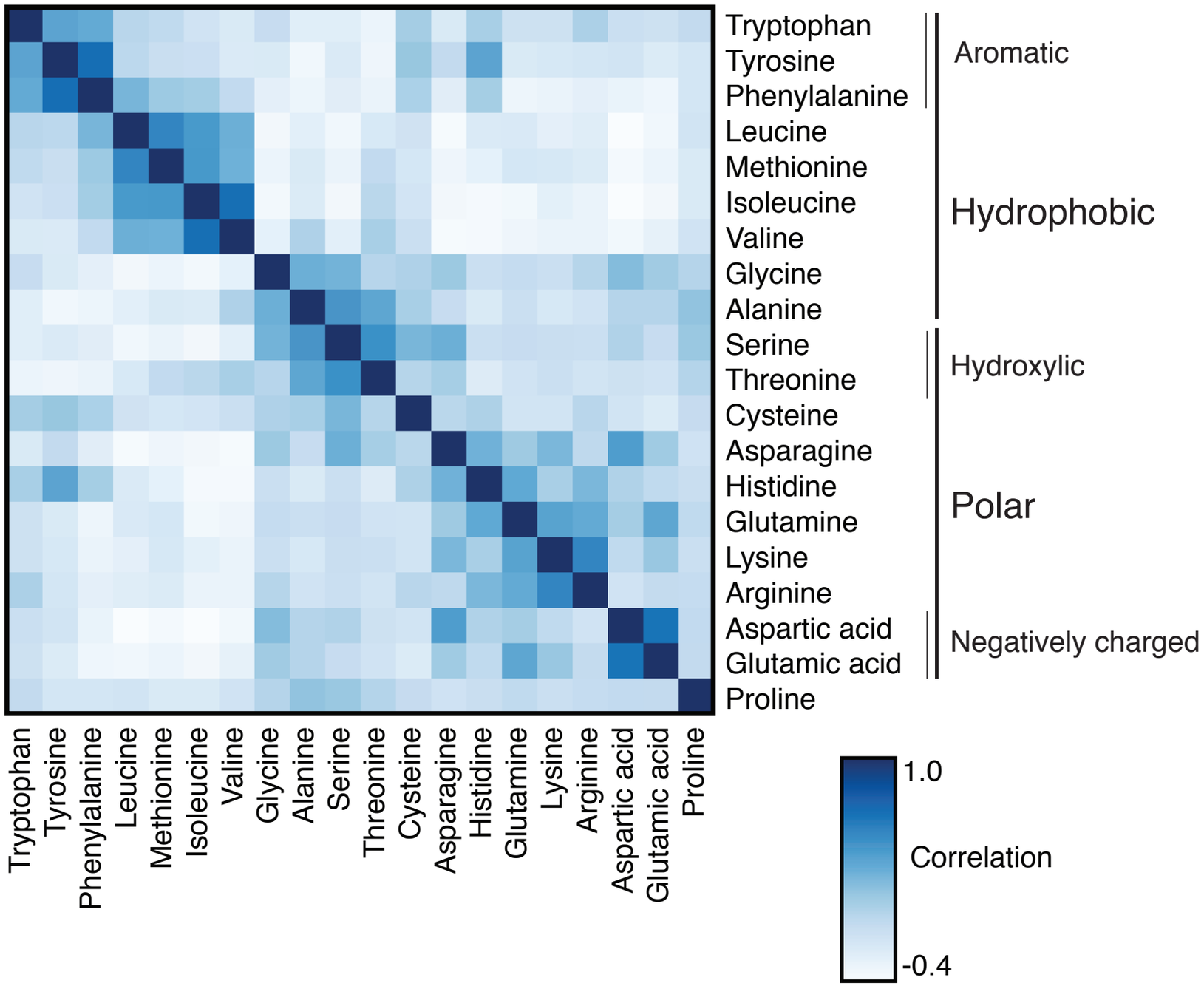
**Scale** shared across a position

# **Biological constraints** were included in model parameterization

$$p(x_i|\mathbf{z}) = \texttt{Softmax}\left(\mathbf{D}\tilde{\mathbf{W}}^{(i)}\left(\texttt{Sigmoid}\left(\mathbf{s}^{(i)}\right)\odot\mathbf{h}\right)+\mathbf{b}^{(i)}\right)$$

**Dictionary** shared across all positions

# The dictionary encodes **amino acid preferences**



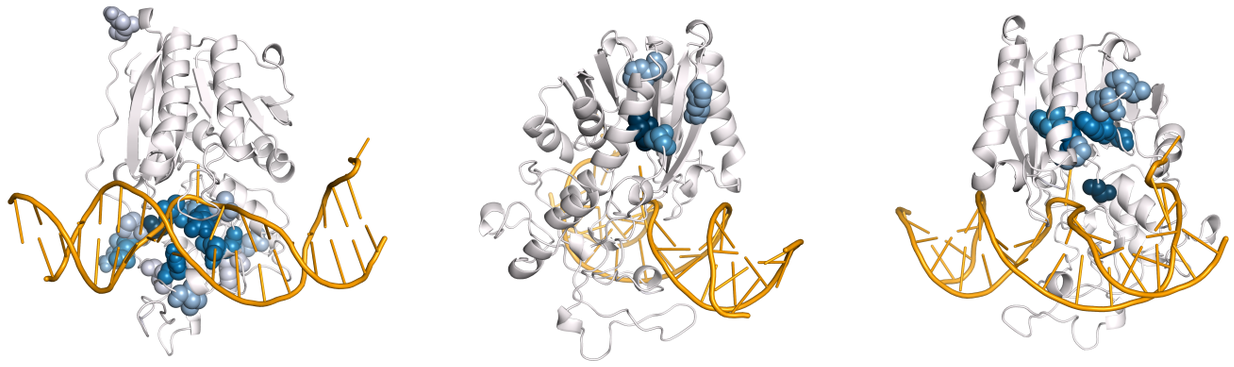
# **Biological constraints** were included in model parameterization

$$p(x_i|\mathbf{z}) = \texttt{Softmax}\left(\mathbf{D}\tilde{\mathbf{W}}^{(i)}\left(\texttt{Sigmoid}\left(\mathbf{s}^{(i)}\right)\odot\mathbf{h}\right)+\mathbf{b}^{(i)}\right)$$

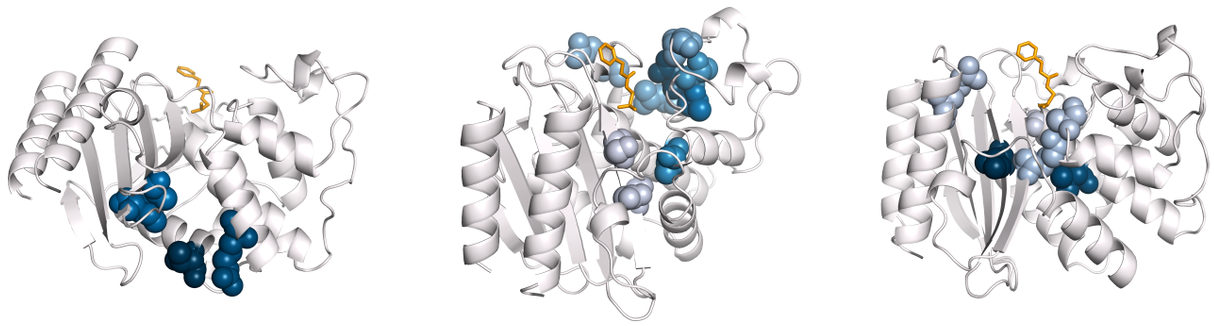
**Scale** shared across a position

# Sparse scale factors are **localized in 3D**

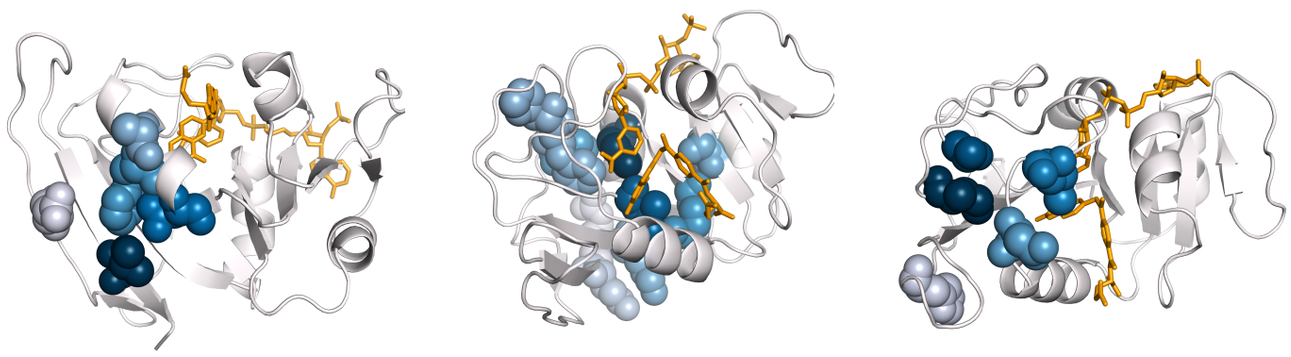
DNA methyltransferase HaeIII



β-lactamase

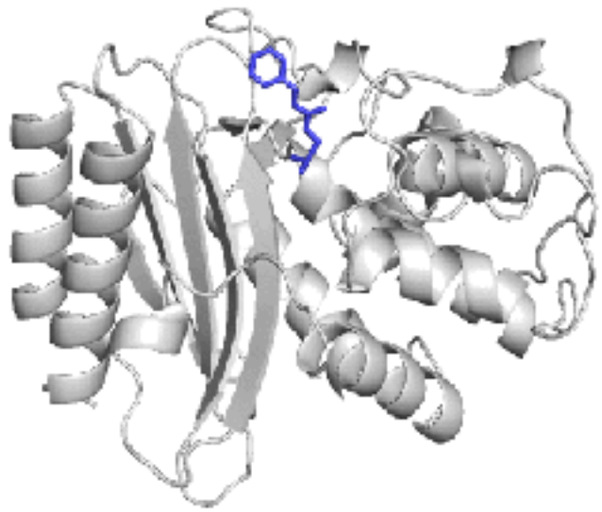


Dihidrofolate reductase



# Recap

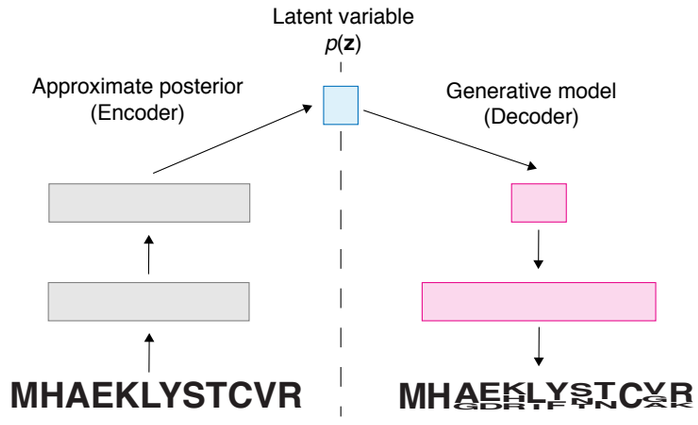
**Predicting the effects of mutations is important**



**Building good generative models of sequence families is useful**

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

**Latent variable models predict the effect of mutations better than state-of-art**



# Thank you!

DOE CSGF

Debora Marks
John Ingraham

Chris Sander

**DeepSequence github:**

https://github.com/debbiemarkslab/DeepSequence

**EVcouplings python package:**

https://github.com/debbiemarkslab/EVcouplings

## Marks Lab +

Thomas Hopf
Anna Green
Charlotta Scharfe
Benni Schubert
Eli Weinstein
Kelly Brock
Rohan Maddamsetti
David Ding
June Shin
Hailey Cambra
Agnes Toth-Petroczy
Perry Palmedo
Frank Poelwijk
Nick Gauthier

Jennie Epp

# Thank you!