Slides by Sergey Ovchinnikov (MIT)

Data matrix

Features

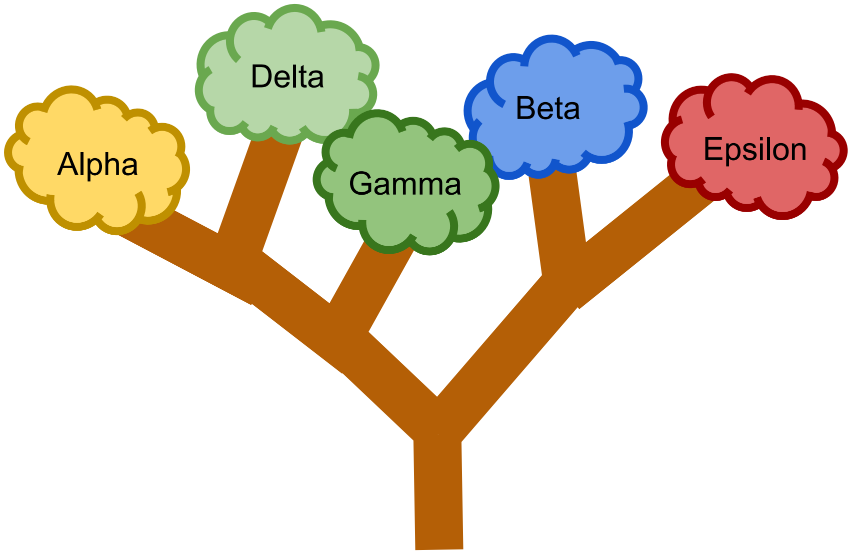
Samples

	1	2	3	4	5	6
Alpha	🟨	🟦	🟦	🟨	🟨	🟦
Beta	🟦	🟦	🟨	🟦	🟦	🟦
Gamma	🟨	🟨	🟦	🟦	🟦	🟦
Delta	🟨	🟨	🟦	🟨	🟨	🟨
Epsilon	🟦	🟦	🟨	🟨	🟨	🟦

Distance matrix

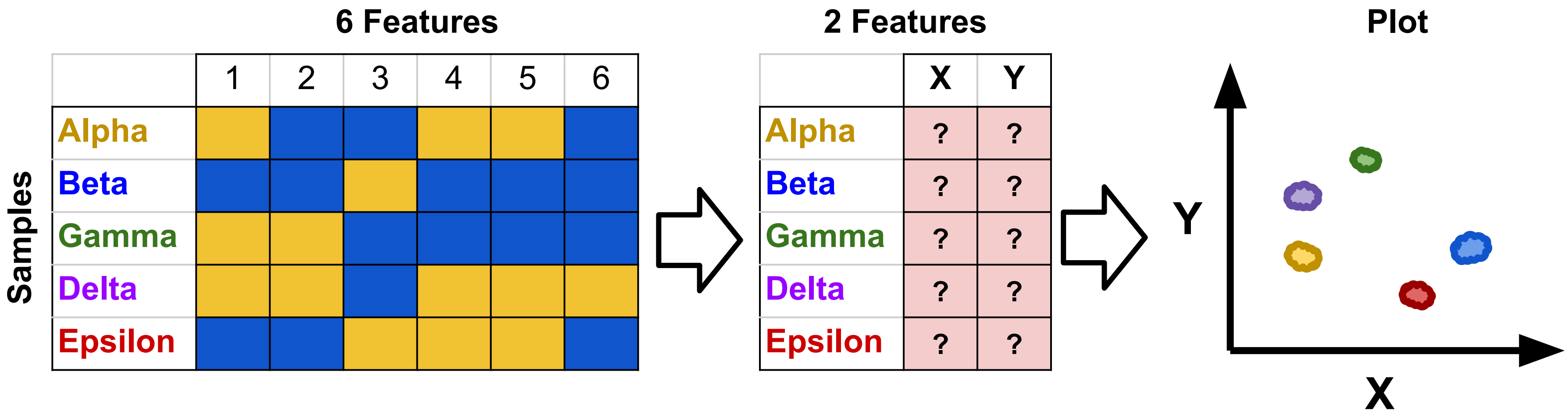
	Alpha	Beta	Gamma	Delta	Epsilon
Alpha	0	4	3	2	2
Beta	4	0	3	6	2
Gamma	4	3	0	3	5
Delta	2	6	3	0	4
Epsilon	2	2	5	4	0

UPGMA/NJ



Phylogenetics

Dimensionality reduction



Samples

6 Features

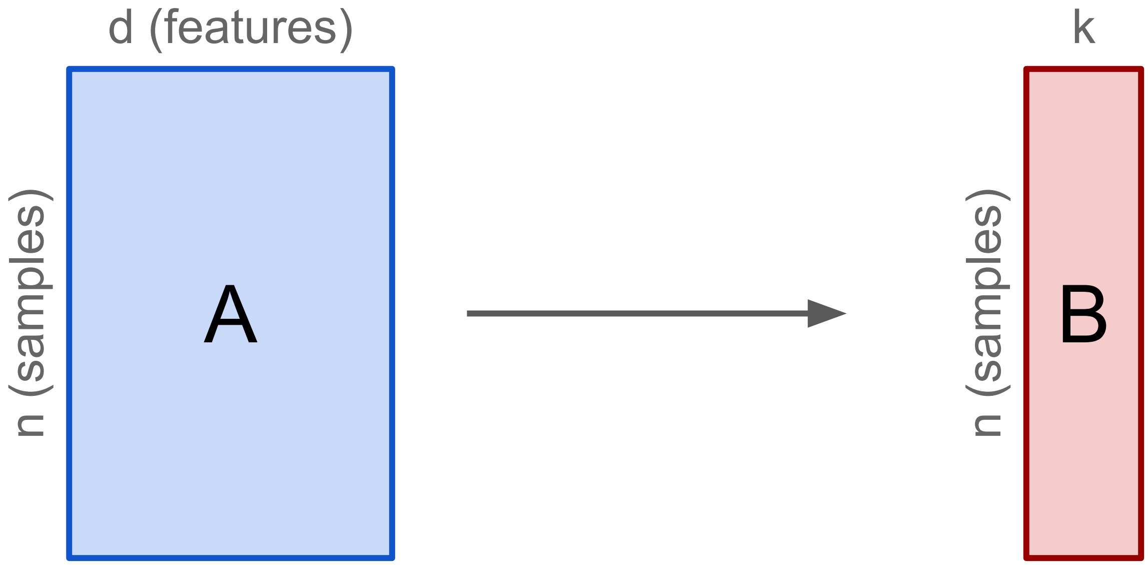
	1	2	3	4	5	6
Alpha	🟨	🟦	🟦	🟨	🟨	🟦
Beta	🟦	🟦	🟨	🟦	🟦	🟦
Gamma	🟨	🟨	🟦	🟦	🟦	🟦
Delta	🟨	🟨	🟦	🟨	🟨	🟨
Epsilon	🟦	🟦	🟨	🟨	🟨	🟦

2 Features

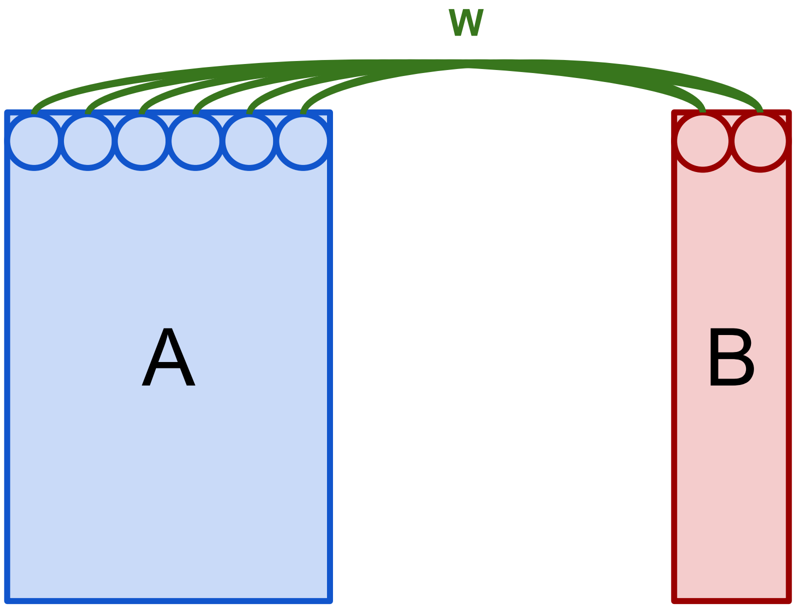
	X	Y
Alpha	?	?
Beta	?	?
Gamma	?	?
Delta	?	?
Epsilon	?	?

Plot

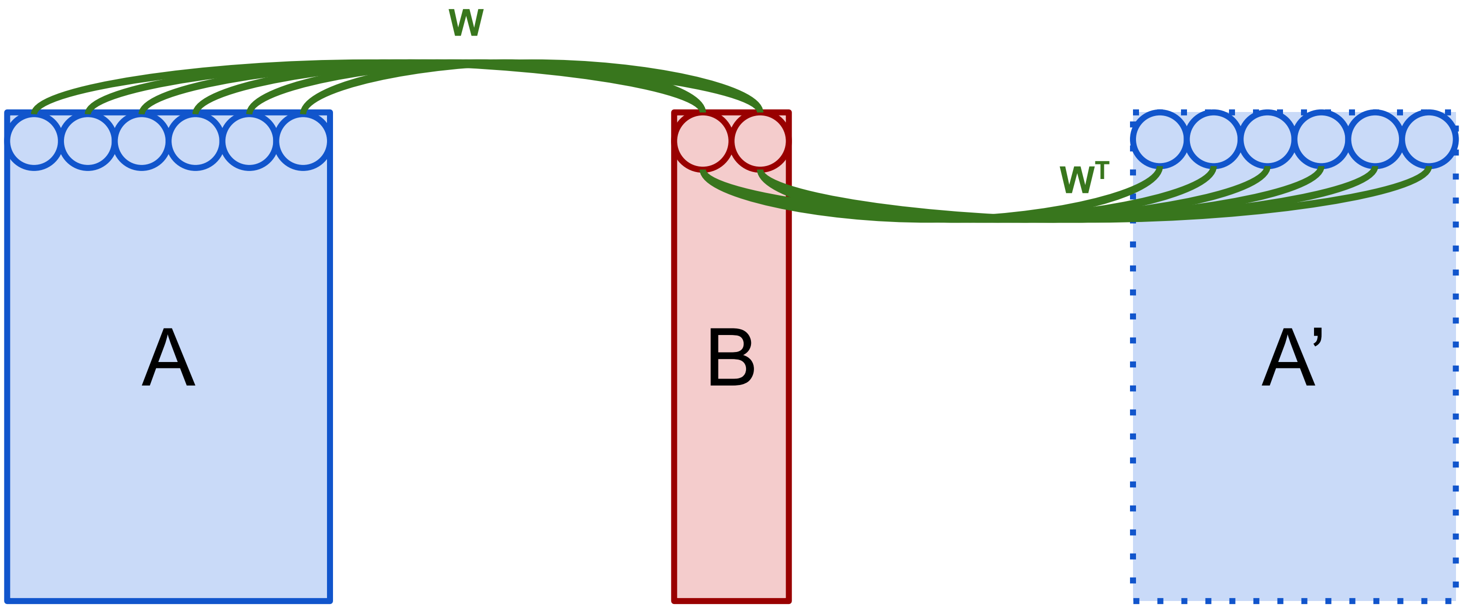
Dimensionality reduction

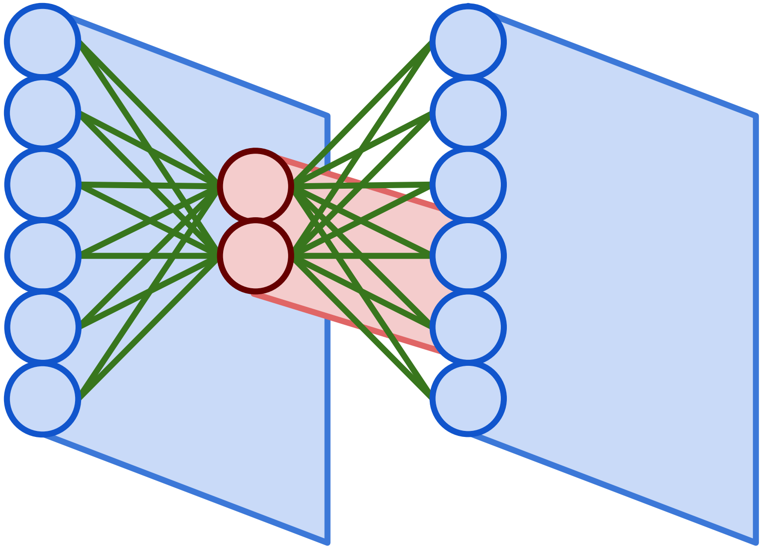


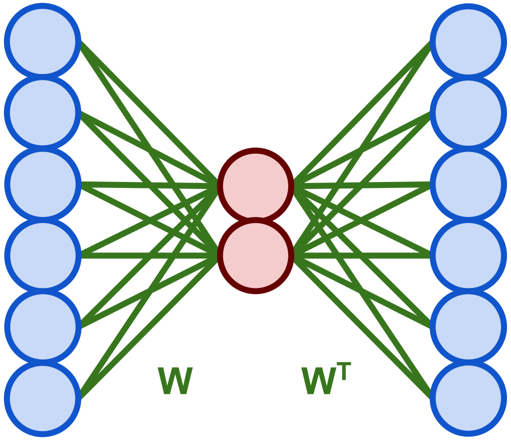
Linear transformation

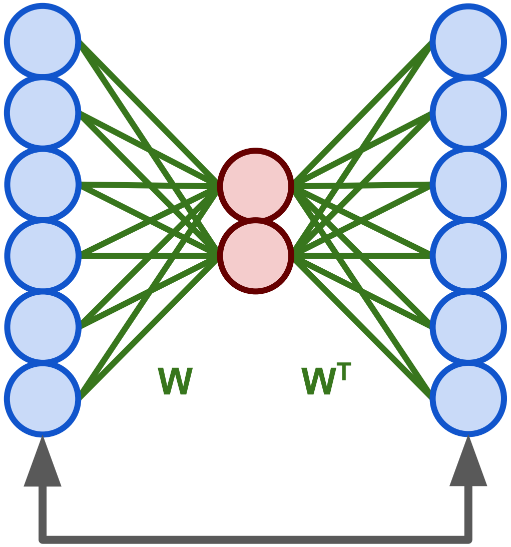


Linear transformation





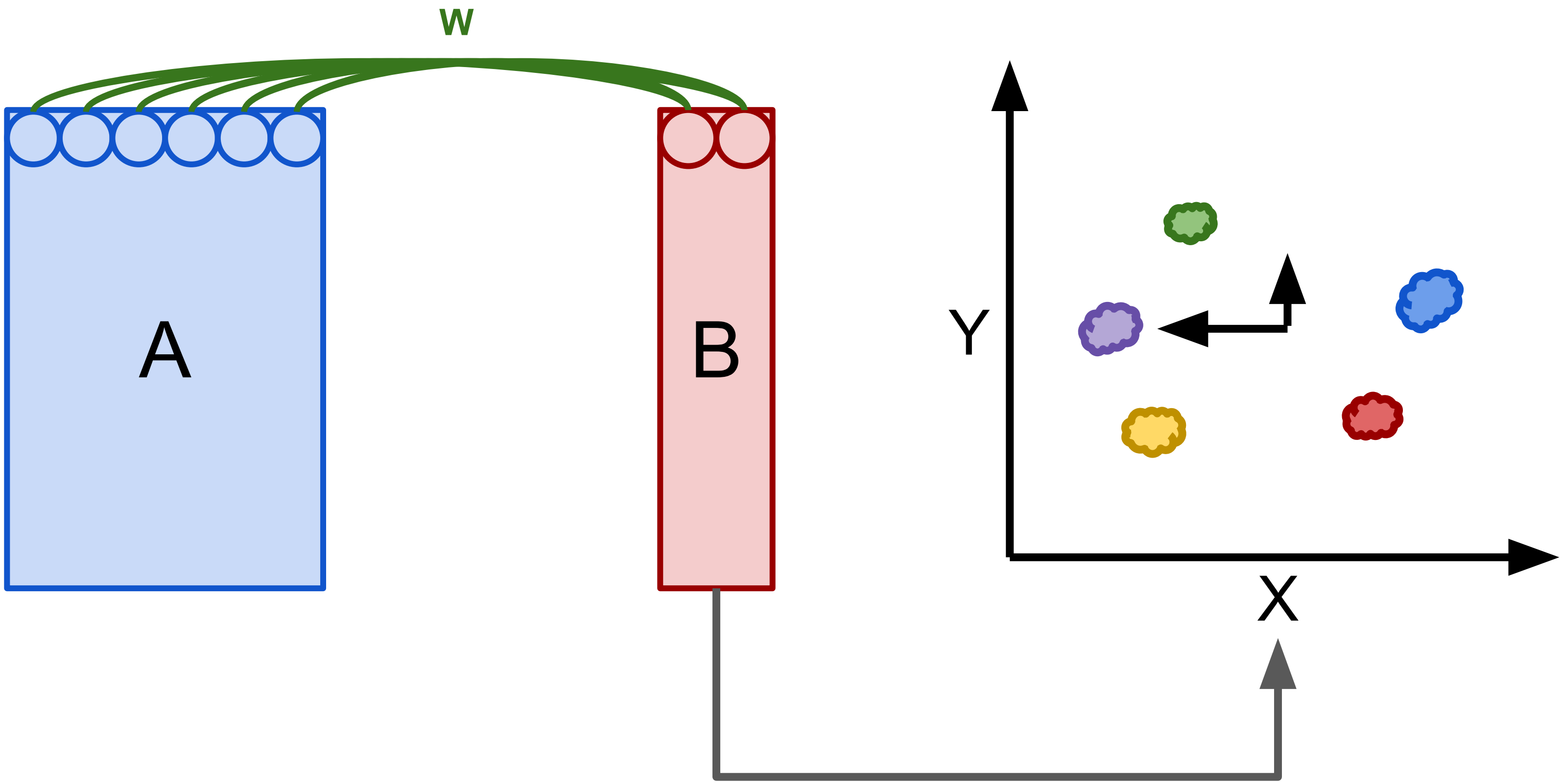

W
W^T



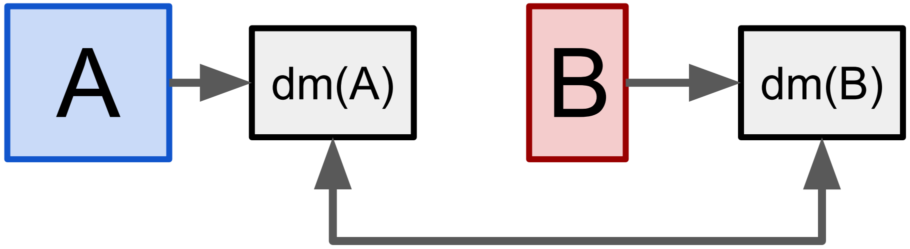
Minimize difference
(find **W**, so that **A** ≈ **A**')

PCA (Principal component analysis)
W = top-k eigenvectors of the covariance matrix of **A**

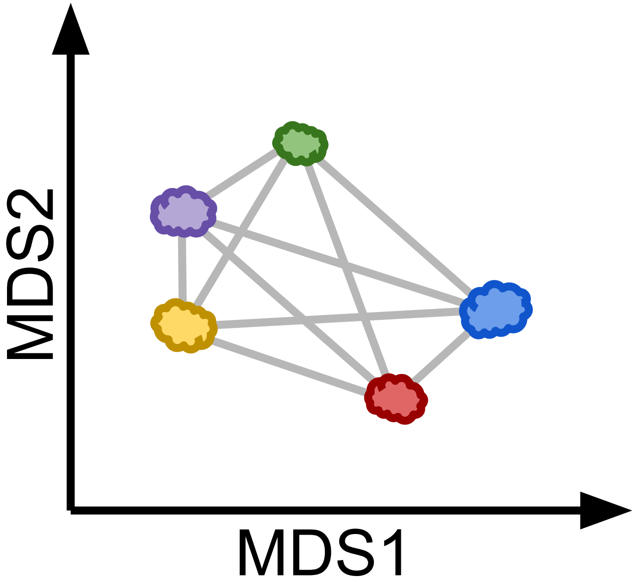
Transform (rotate) and plot the data!



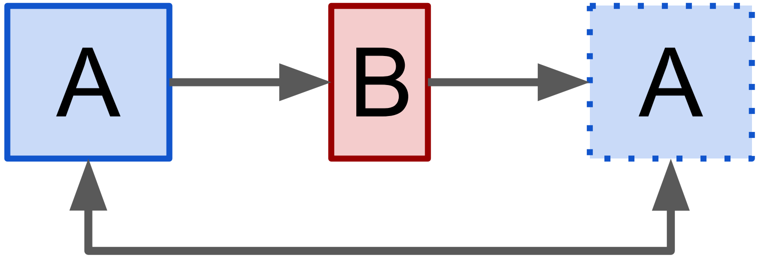
MDS - multidimensional scaling



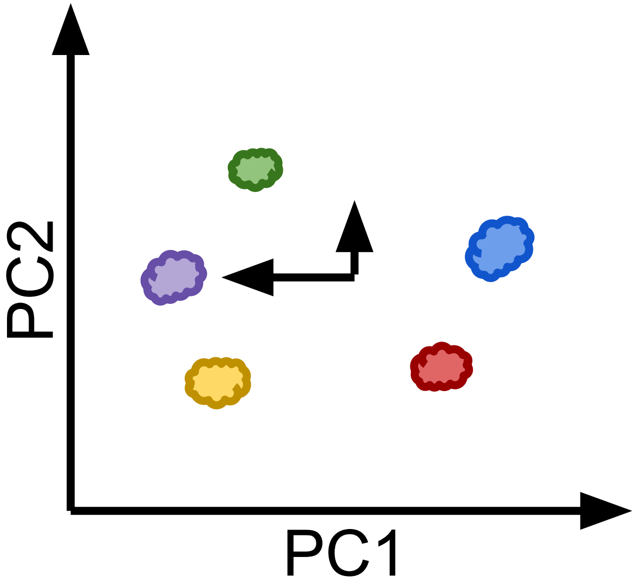
MDS finds a lower-dimensional manifold that preserves the distances.

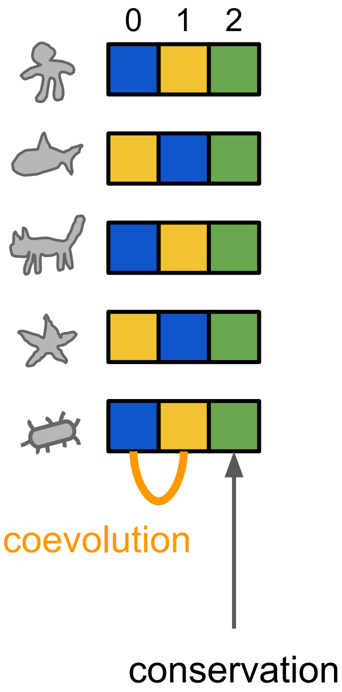


PCA - principal component analysis



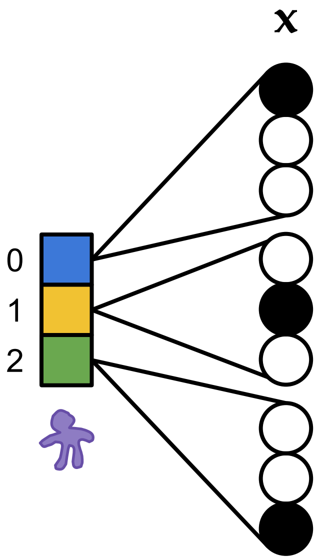
PCA “rotates” the data to find axes that maximize the variance.



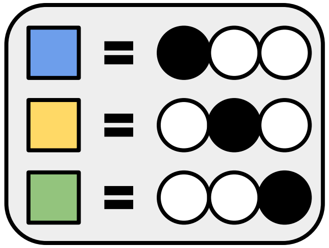


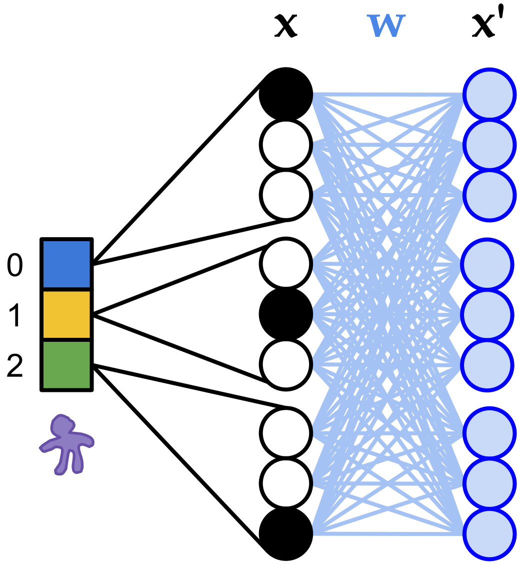
Let's pretend there are:

- 3 positions
- 3 characters: 🟦 🟨 🟩



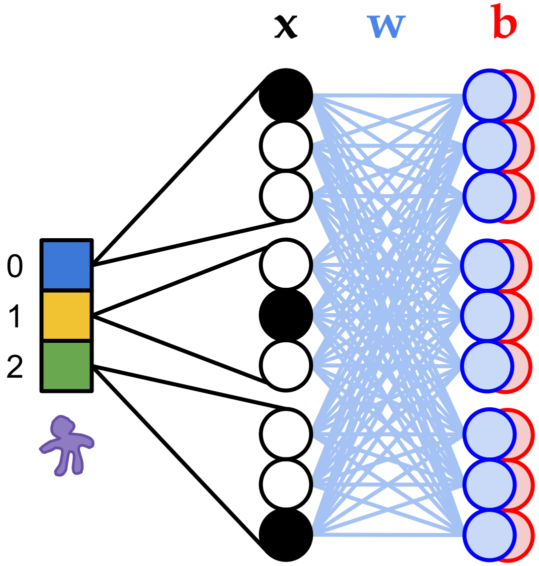
$\mathbf{x}$ = one_hot(**MSA**)





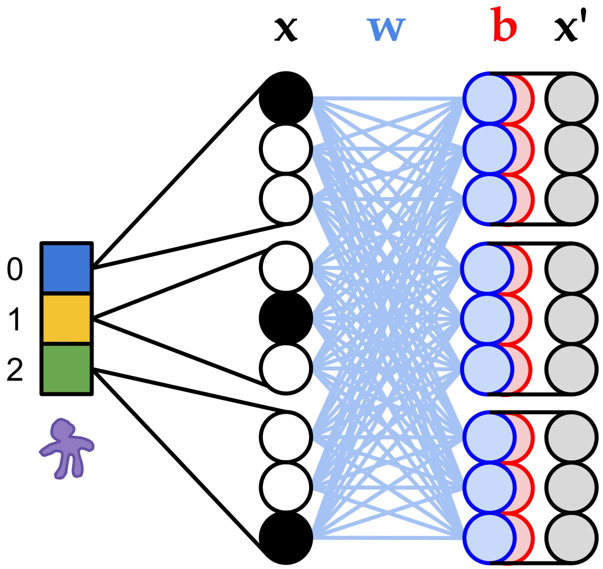
[w]eights = coevolution

$x' = x@w$



[w]eights = coevolution
[b]ias = conservation

x' = x@w + b



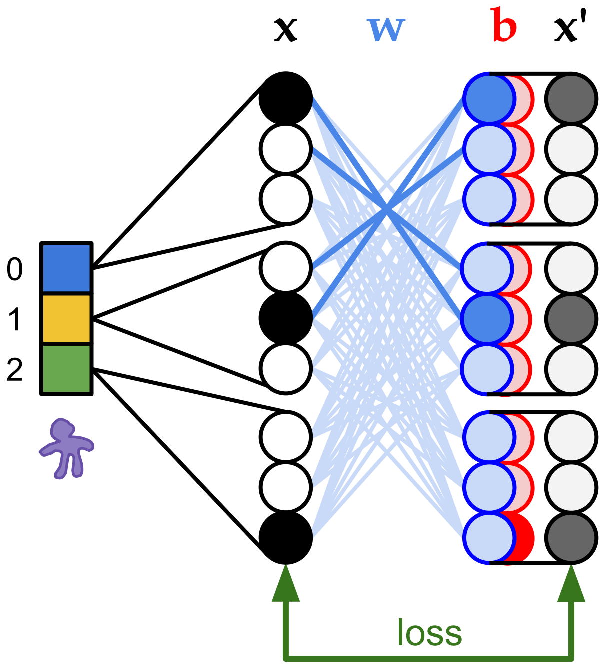
softmax(x) = exp(x)/sum(exp(x))

To make sure probabilities at each position sum to 1.0

x' = softmax(**x**@**w** + **b**)

Remove self-connections (Pseudo-likelihood)

$P(x_1,x_2,x_3) \approx P(x_1|x_2,x_3) * P(x_2|x_1,x_3) * P(x_3|x_1,x_2)$



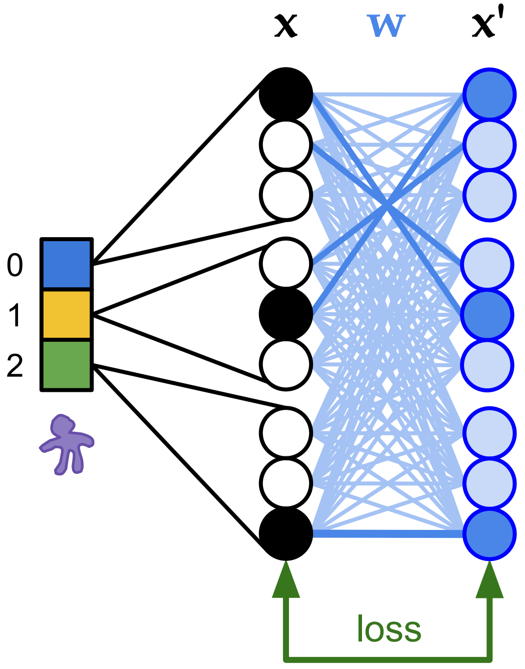
Minimize difference using categorical-crossentropy

$\mathbf{x}' = \mathrm{softmax}(\mathbf{x}@\mathbf{w} + \mathbf{b})$

$\mathrm{loss} = -\mathbf{x}*\log(\mathbf{x}')$

Balakrishnan, S., Kamisetty, H., Carbonell, J.G. and Langmead, C.J., 2009. Structure Learning for Generative Models of Protein Fold Families.

Direct Coupling Analysis
(analytical solution: inverse covariance)



[w]eights = coevolution

$$\mathbf{x}' = \mathbf{x} @ \mathbf{w}$$

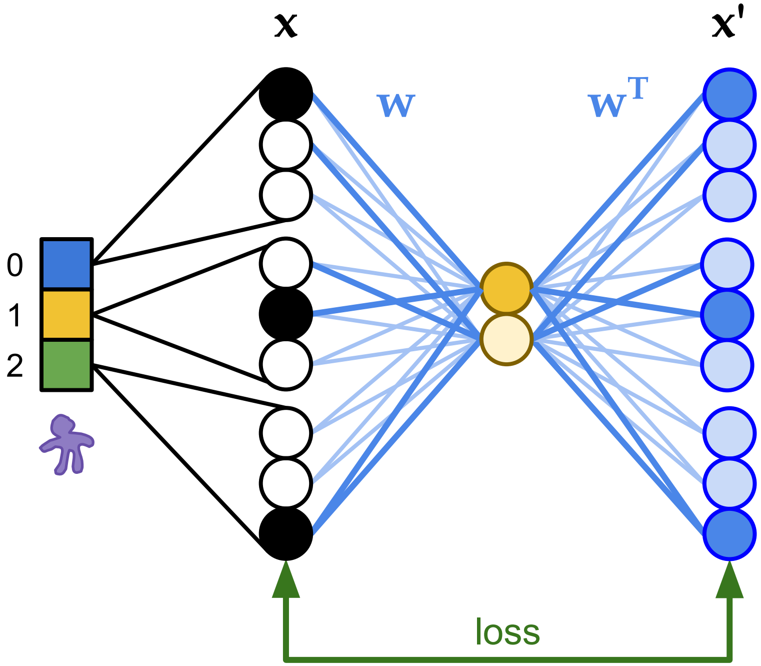
$$\text{loss} = (\mathbf{x}' - \mathbf{x})^2 / N - 2\mathrm{Tr}(\mathbf{w})$$
$$\mathbf{w} = \mathrm{cov}(\mathbf{x})^{-1} + I$$

Dauparas, J., Wang, H., Swartz, A., Koo, P., Nitzan, M. and **Ovchinnikov, S.**, 2019. Unified framework for modeling multivariate distributions in biological sequences. *arXiv*

Morcos, F.,… Weigt, M., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS*

Autoencoder

(analytical solution: principle component analysis)



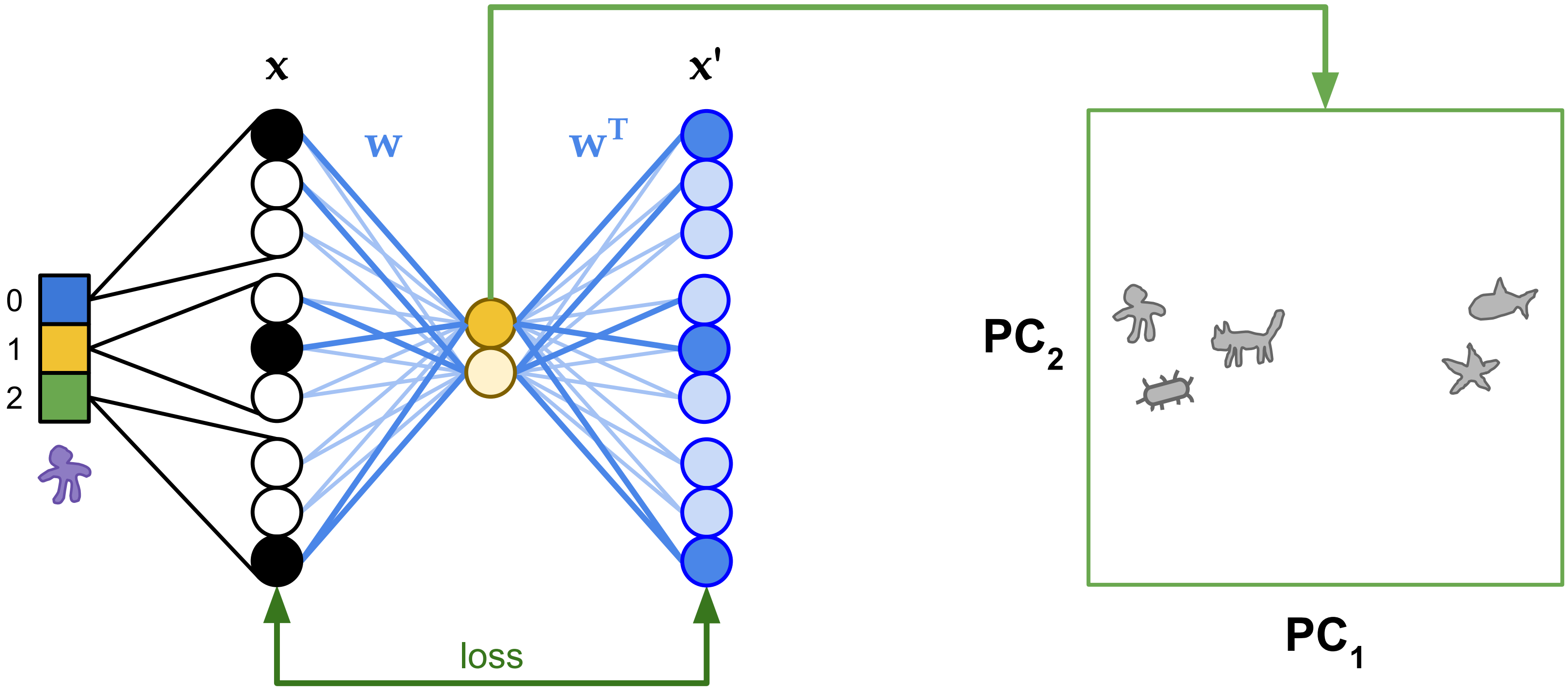
$\mathbf{x}' = \mathbf{x} @ \mathbf{w} @ \mathbf{w}_T$

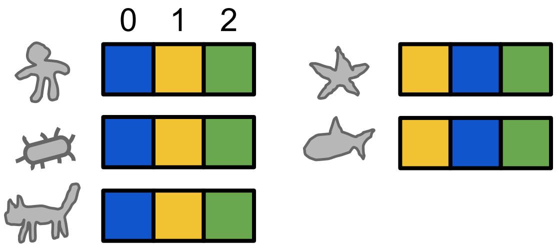
$\text{loss} = (\mathbf{x}' - \mathbf{x})^2$

$\mathbf{w} = \text{PCA}(\mathbf{x}).\text{components\_}$

Autoencoder

(analytical solution: principle component analysis)



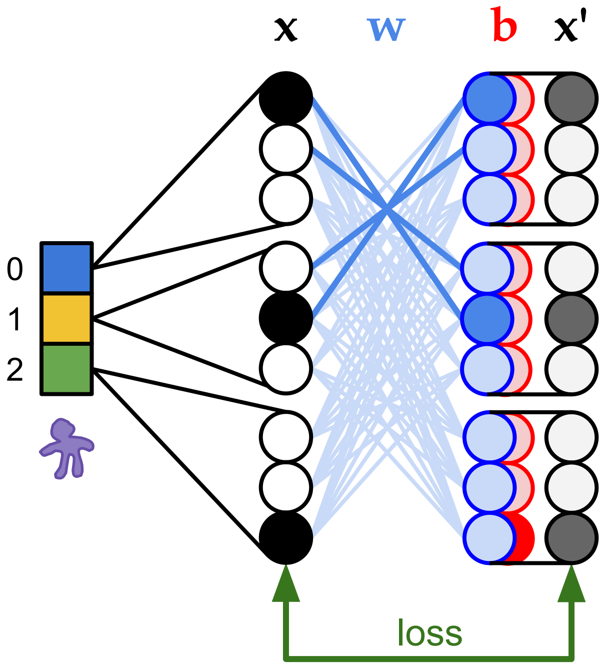


$\mathbf{x}' = \mathbf{x} @ \mathbf{w} @ \mathbf{w}_T$

loss = $(\mathbf{x}' - \mathbf{x})^2$

$\mathbf{w}$ = PCA($\mathbf{x}$).components_

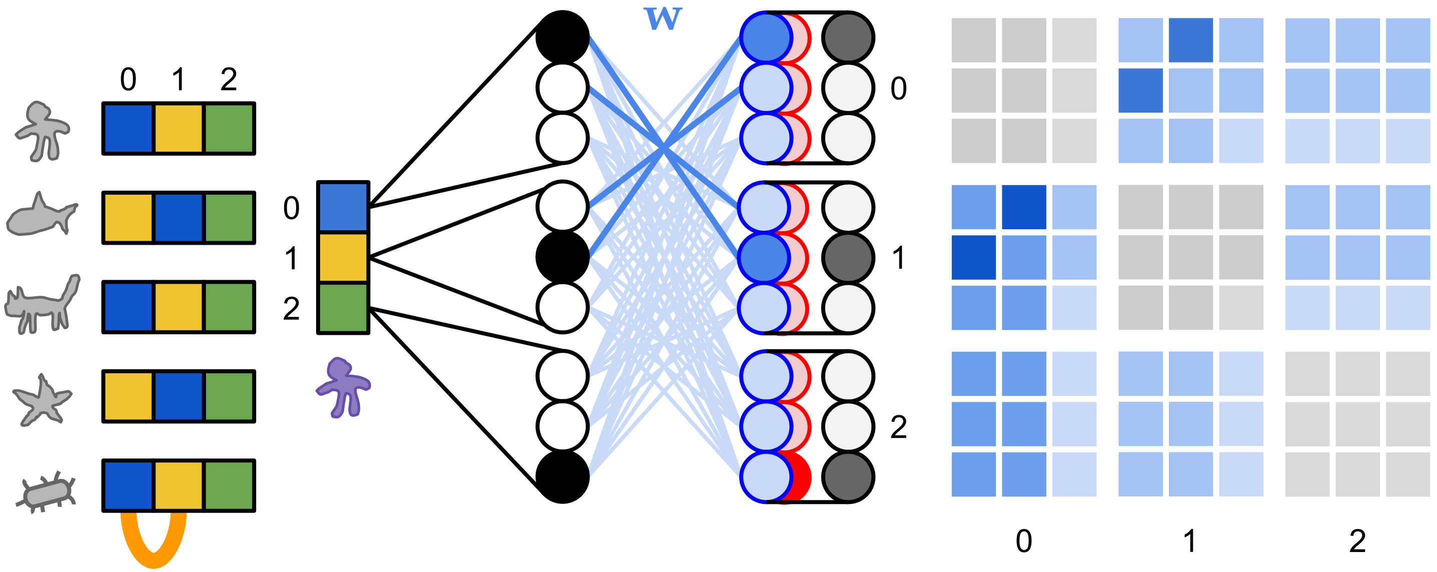
GREMLIN (Generative REgularized ModeLs of proteINs)

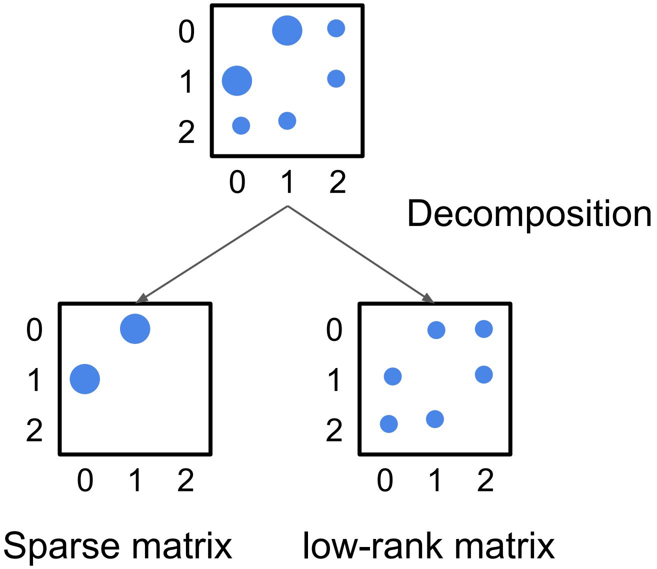


$\mathbf{x}' = \mathrm{softmax}(\mathbf{x} @ \mathbf{w} + \mathbf{b})$

$\mathrm{loss} = -\mathbf{x} * \log(\mathbf{x}') + \lambda \mathbf{b}^2 + \lambda \mathbf{w}^2$

Low rank correction often required



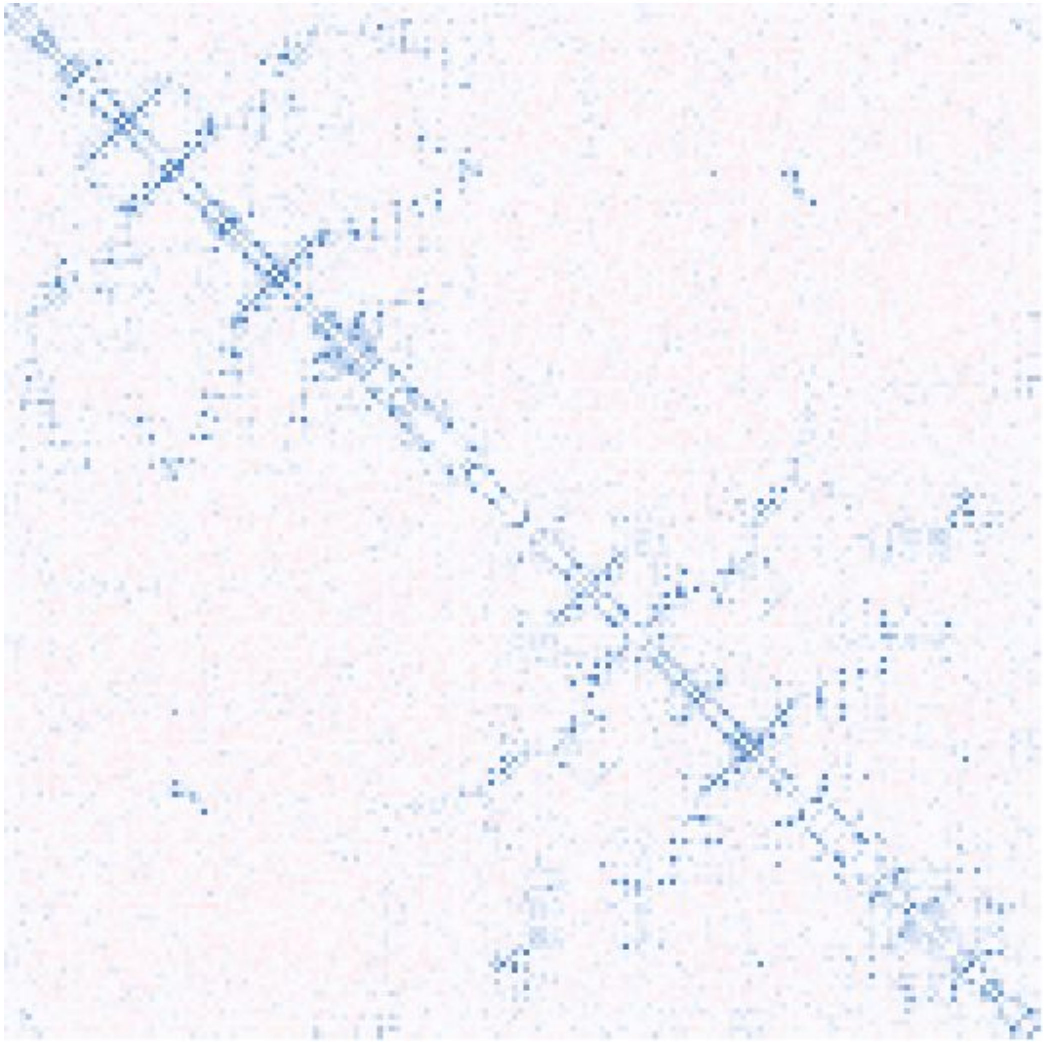


Decomposition

Sparse matrix

low-rank matrix

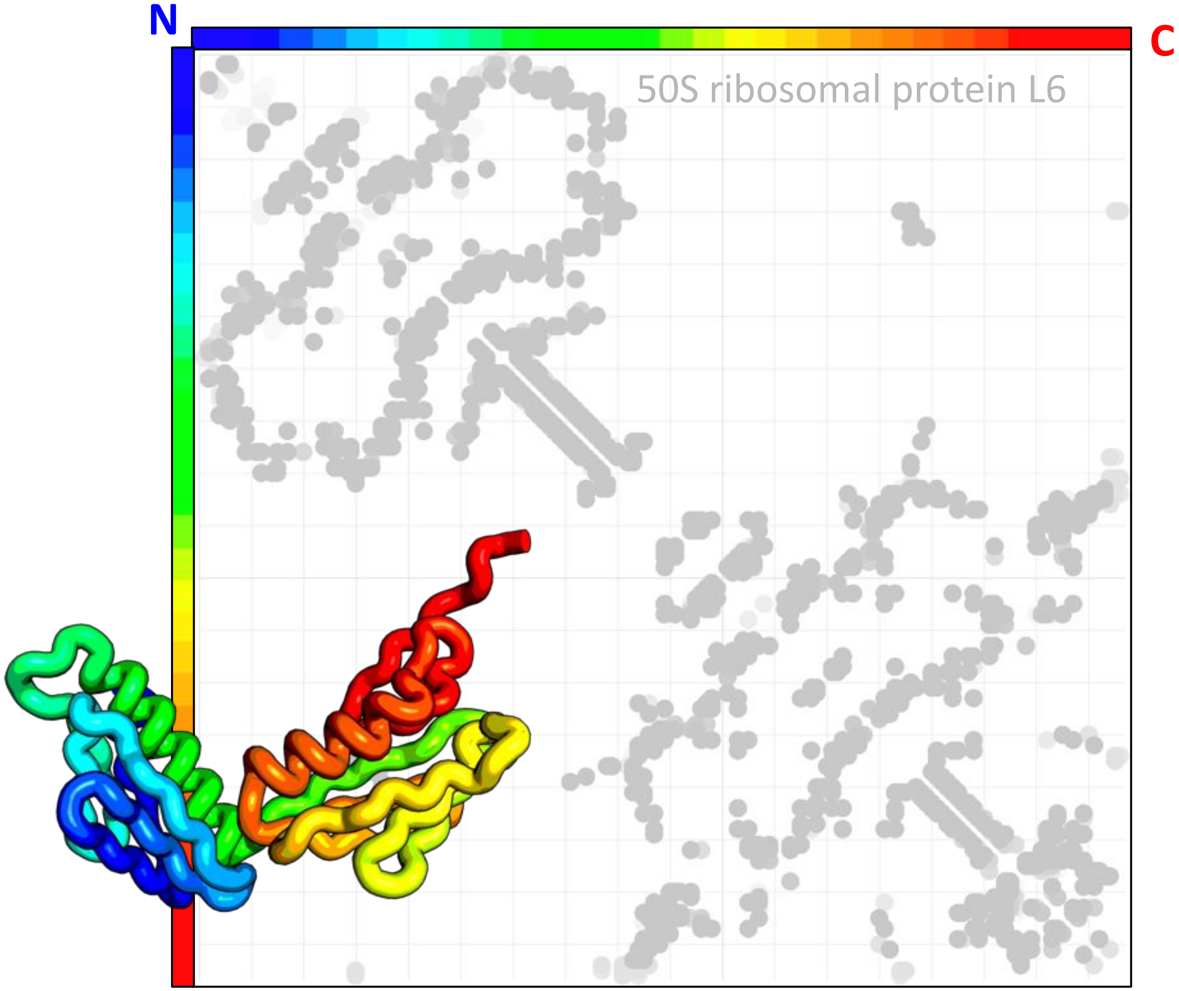
APC = raw - AP



$$AP = \frac{\sum(i,:) * \sum(:,j)}{\sum(:,:)}$$

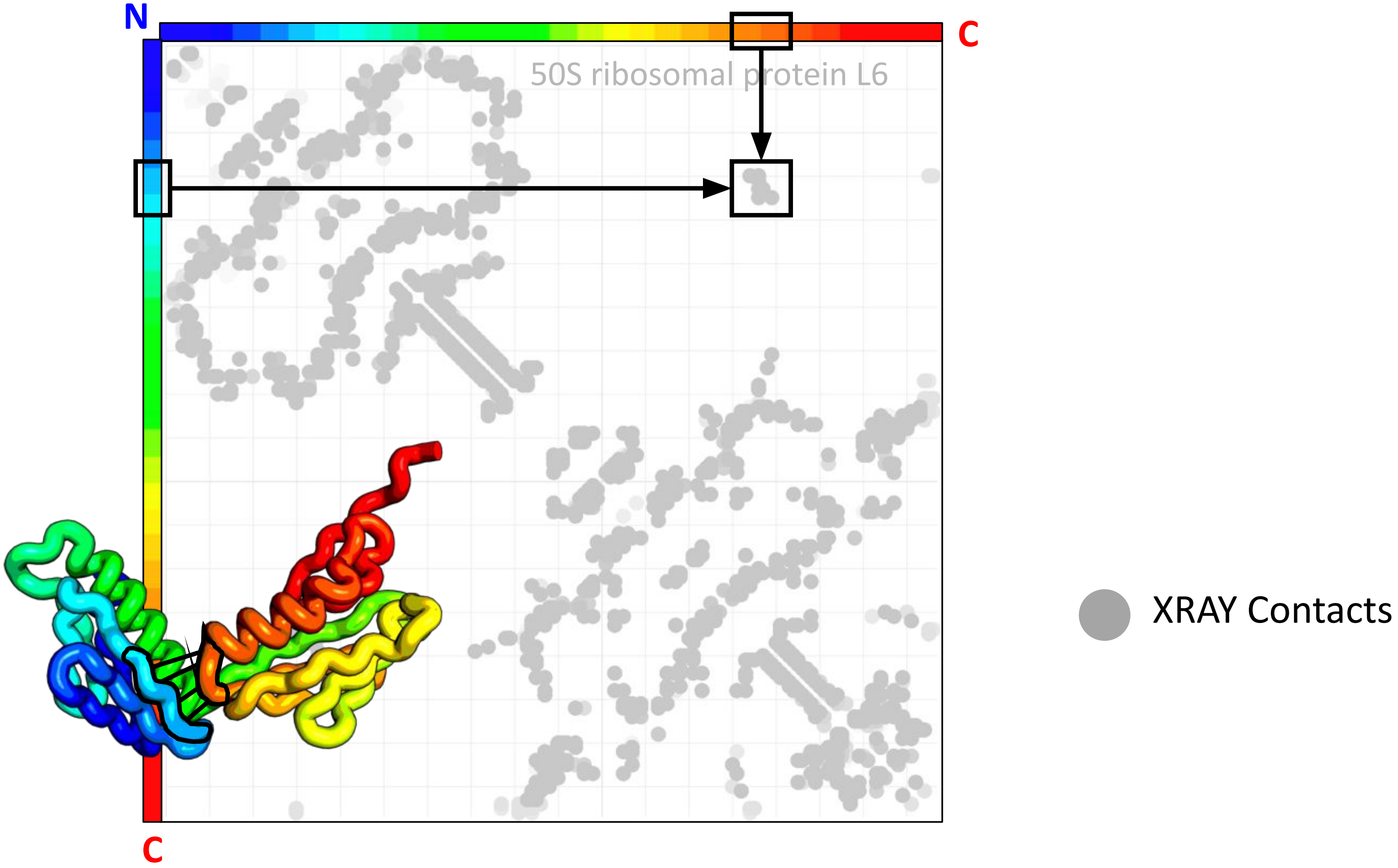
Dunn et al. 2008
Mutual information without the **influence of phylogeny or entropy** dramatically improves residue contact prediction.

Contact map

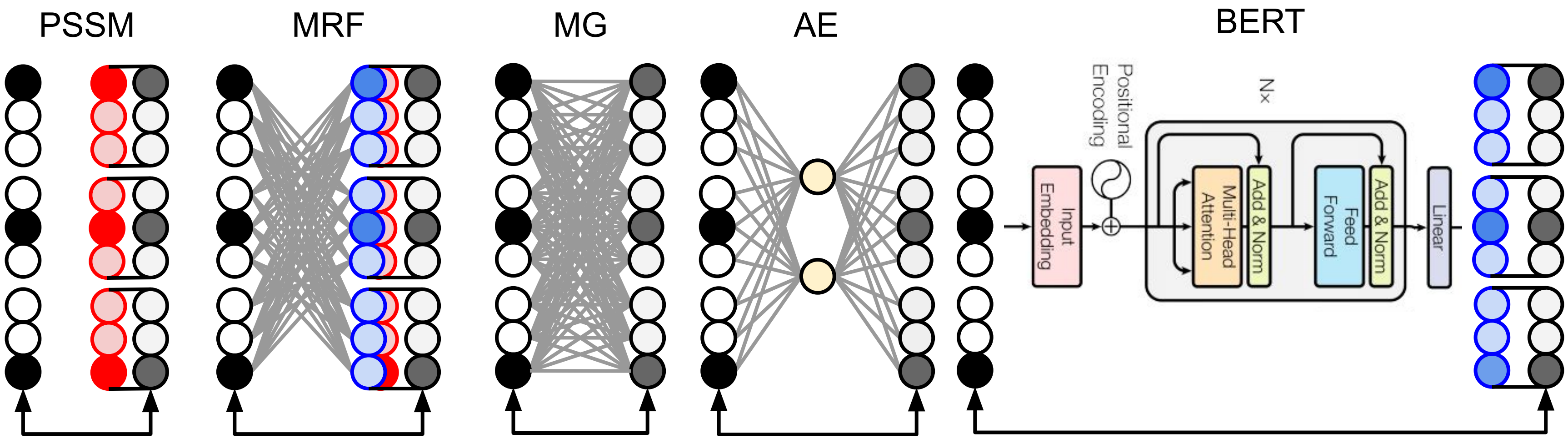


● XRAY Contacts

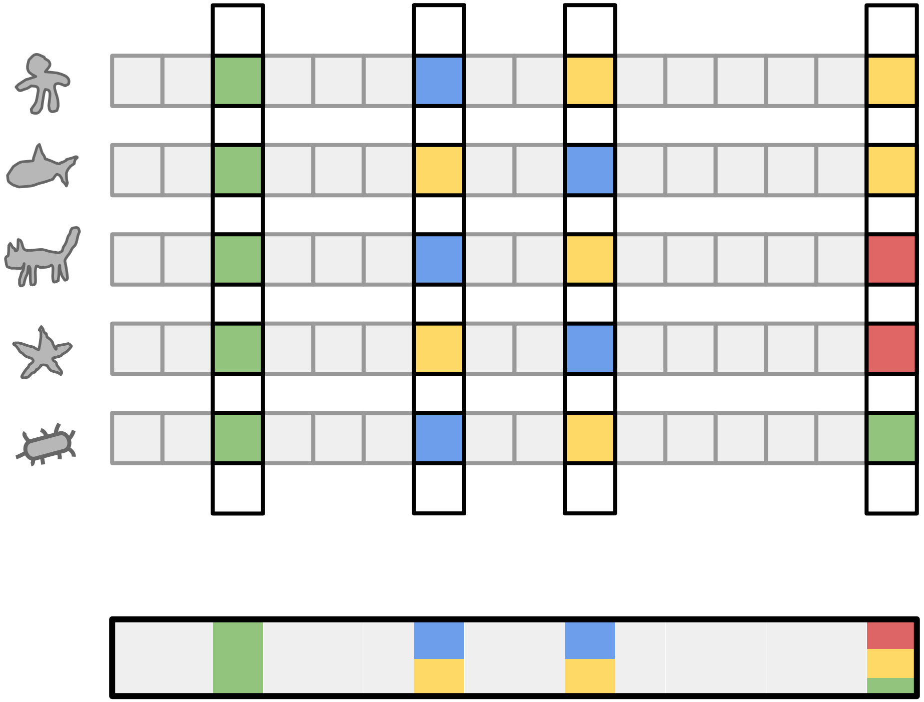
How to read a contact map



All these models can be expressed as “Autoencoders”



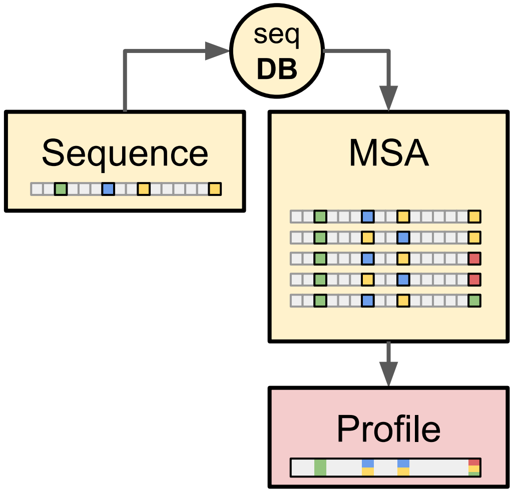
Analyze the MSA for conservation



Profile = Position-specific-scoring matrix

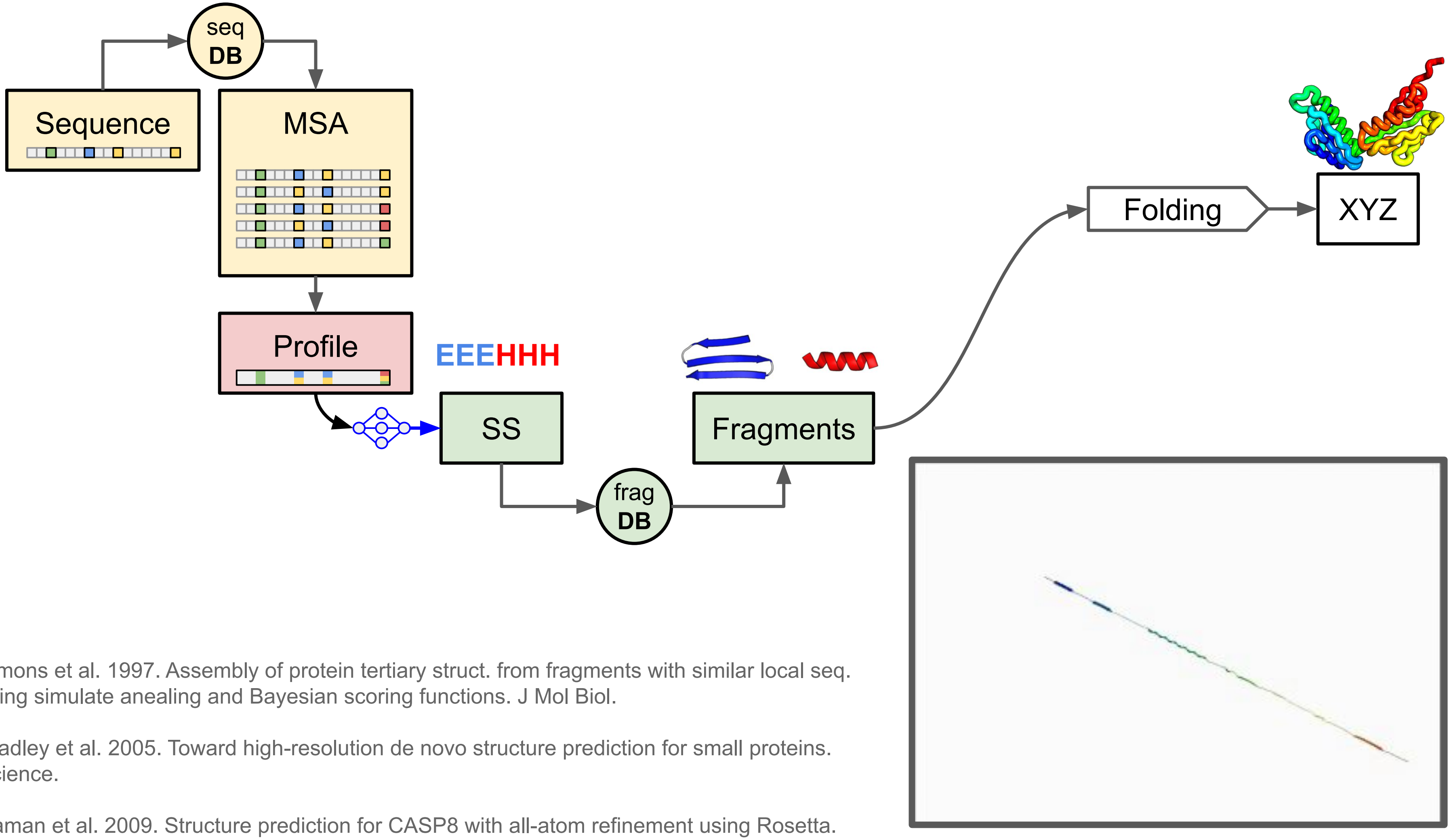
Typical structure prediction pipeline

MSA = multiple sequence alignment
DB = database
Profile = conservation



Typical structure prediction pipeline

MSA = multiple sequence alignment
DB = database
Profile = conservation
SS = secondary structure prediction
XYZ = coordinates

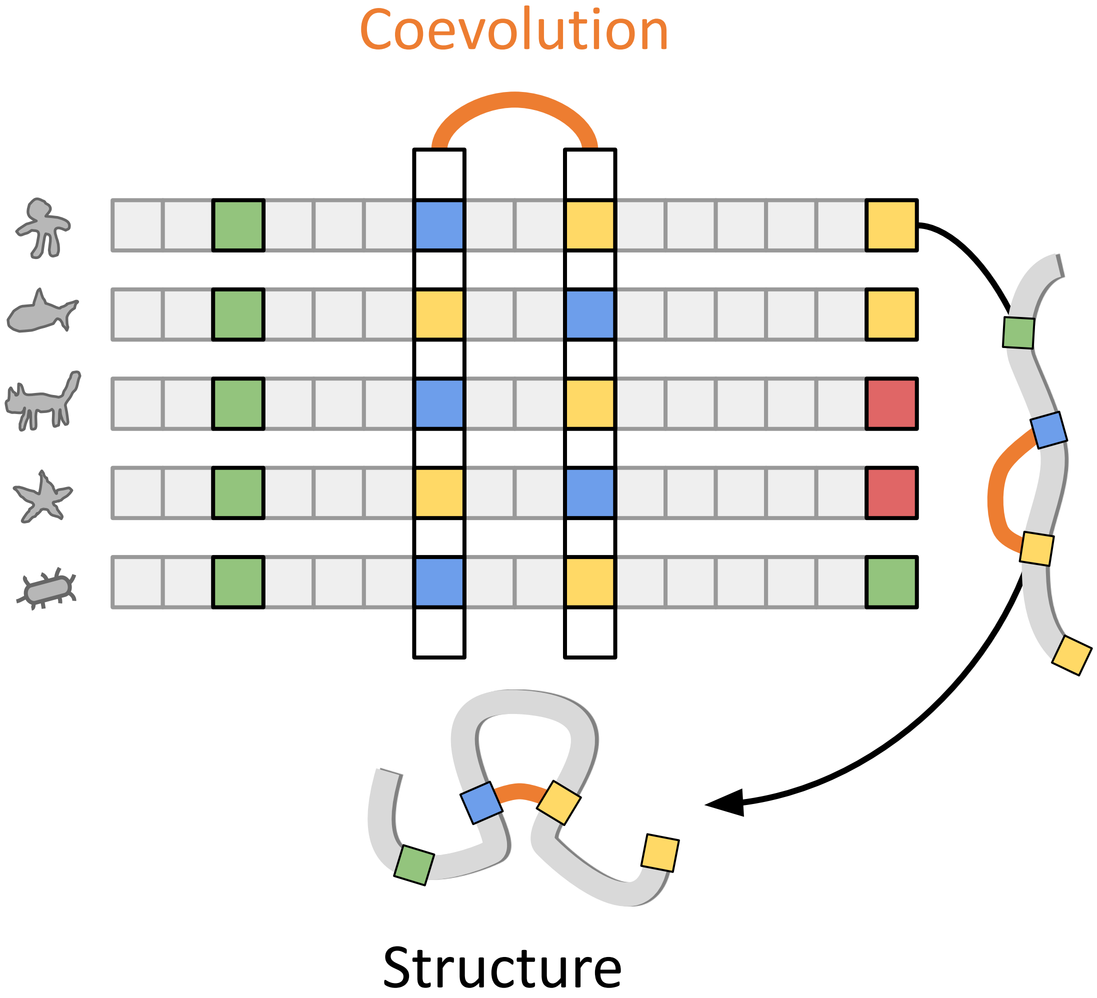


Simons et al. 1997. Assembly of protein tertiary struct. from fragments with similar local seq. using simulate anealing and Bayesian scoring functions. J Mol Biol.

Bradley et al. 2005. Toward high-resolution de novo structure prediction for small proteins. Science.

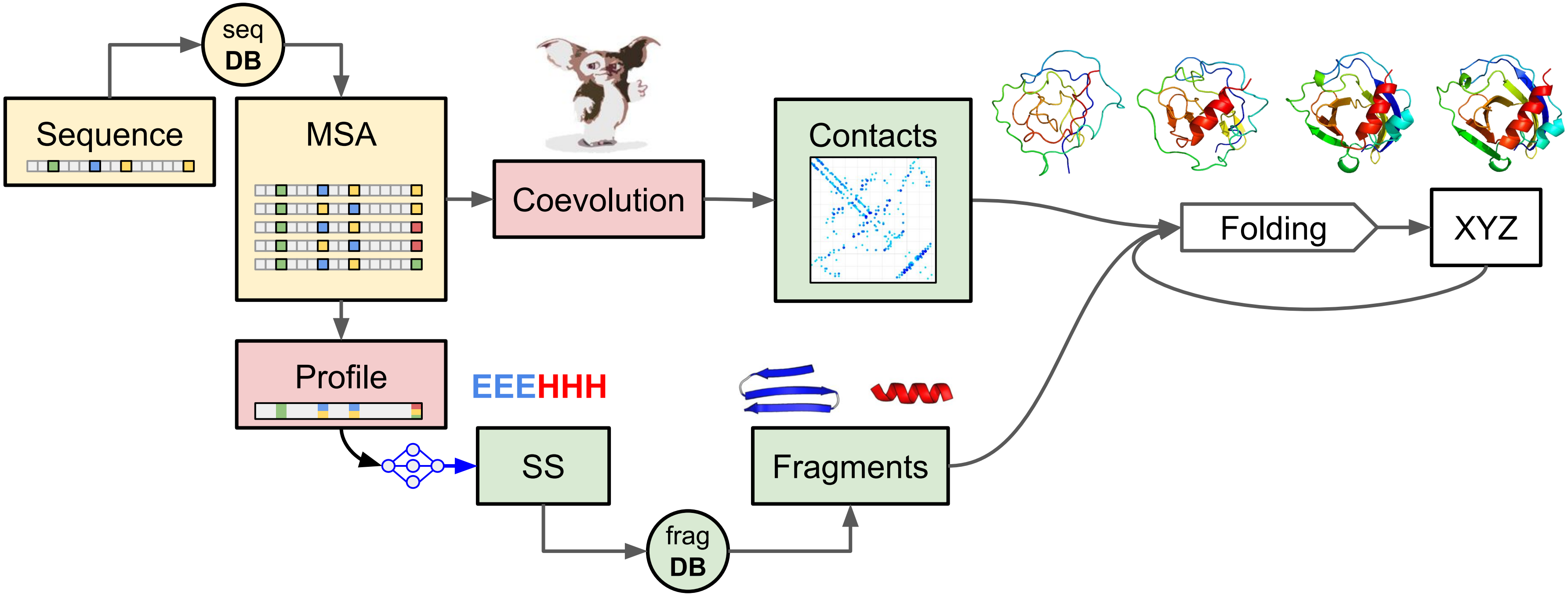
Raman et al. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins

Use the as restraints in folding simulations!



citations: tinyurl.com/coevopapers

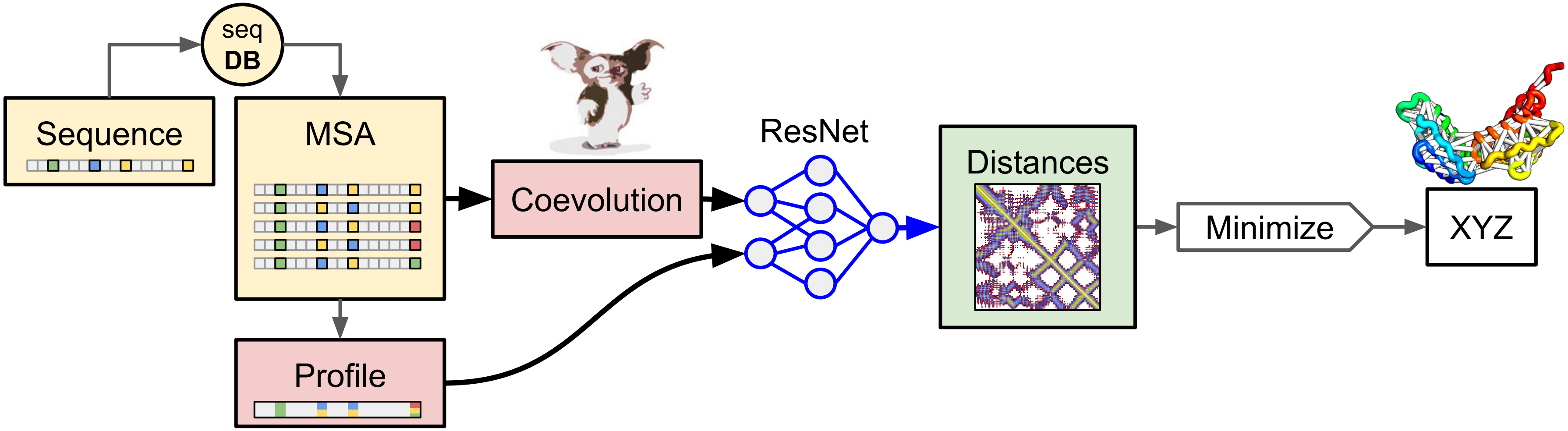
Though our pipeline worked great, it was too expensive to run (100K computers running for 2 weeks per prediction).



Ovchinnikov et al. 2015. Improved de novo struct. pred. in CASP11 by incorporating Co-evol. info. into rosetta. *Proteins*

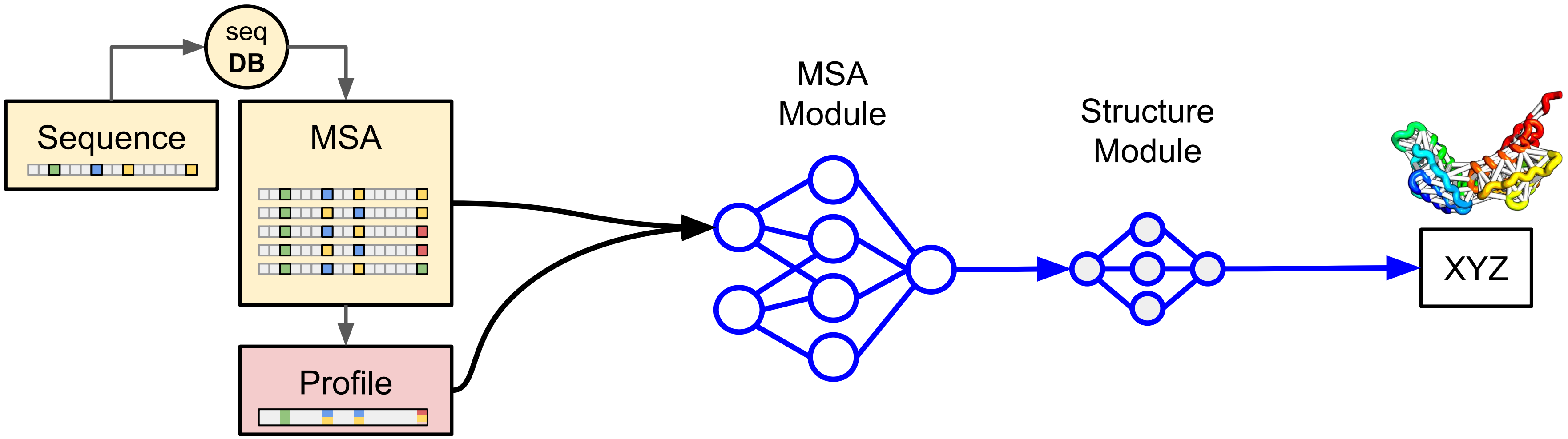
additional citations: tinyurl.com/coevopapers

AlphaFold1 - use Neural Networks extract constraints from raw coevolution features.



Senior et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature*

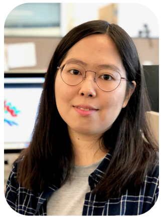
AlphaFold2/RoseTTAFold - Neural network everything



Jumper J. et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*

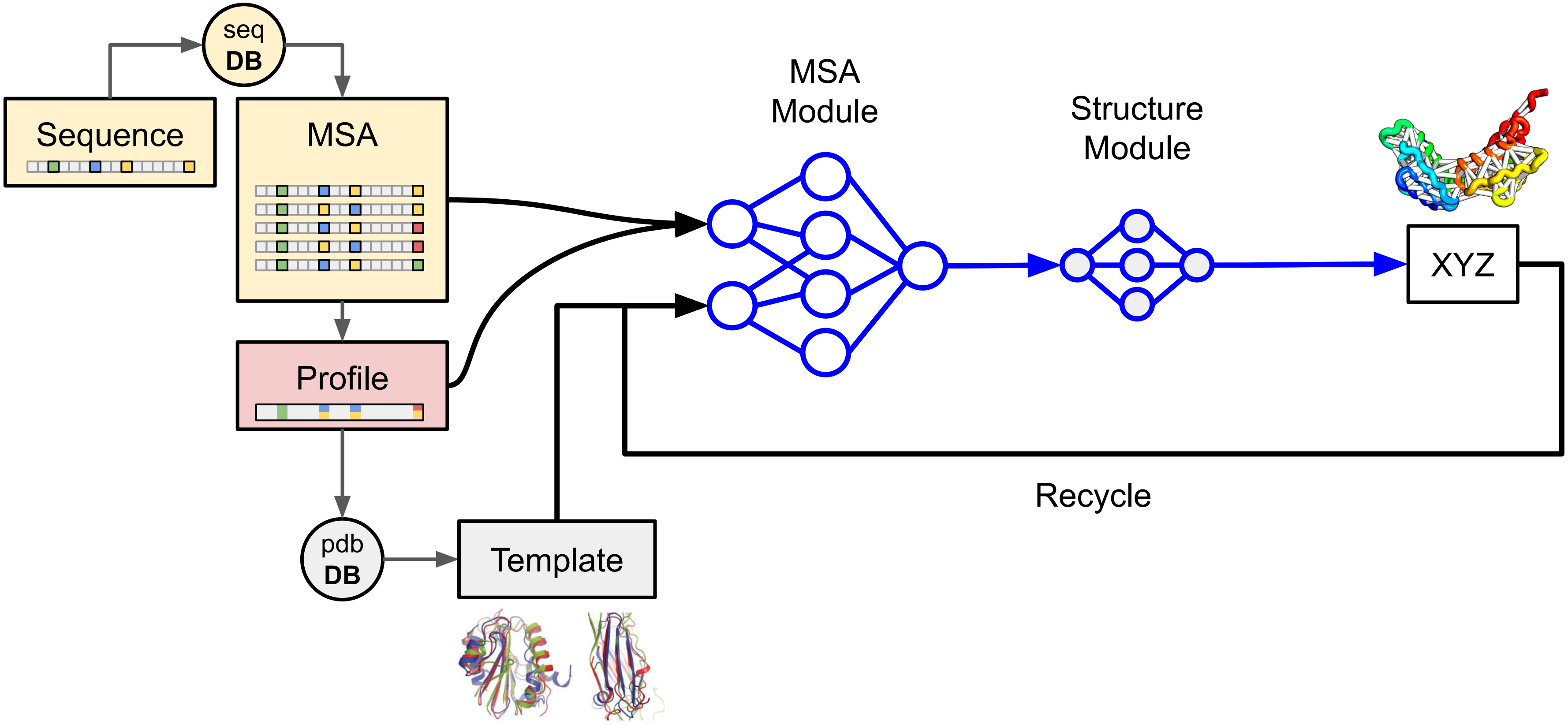
Baek M, DiMaio F, Anishchenko I, Dauparas J, **Ovchinnikov S,** …, Baker D. 2021. Acc. pred. of protein struct. and inter. using a 3-track NN. *Science*

John
Jumper

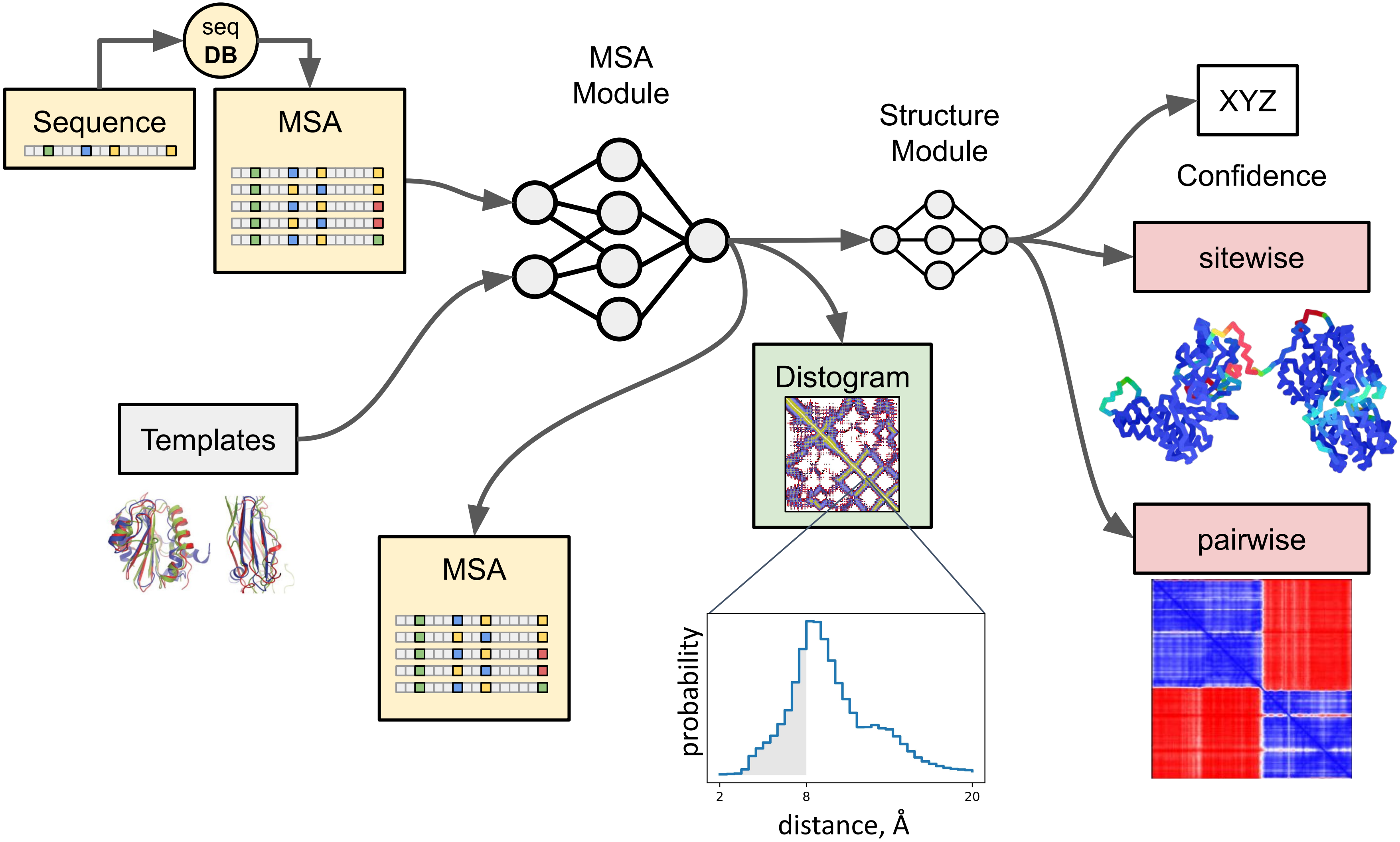


Minkyung
Baek

AF/RF - Use previously solved structures as templates

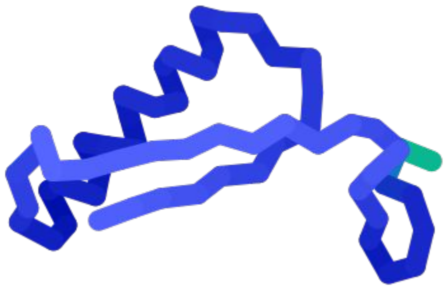


What else does AlphaFold return?

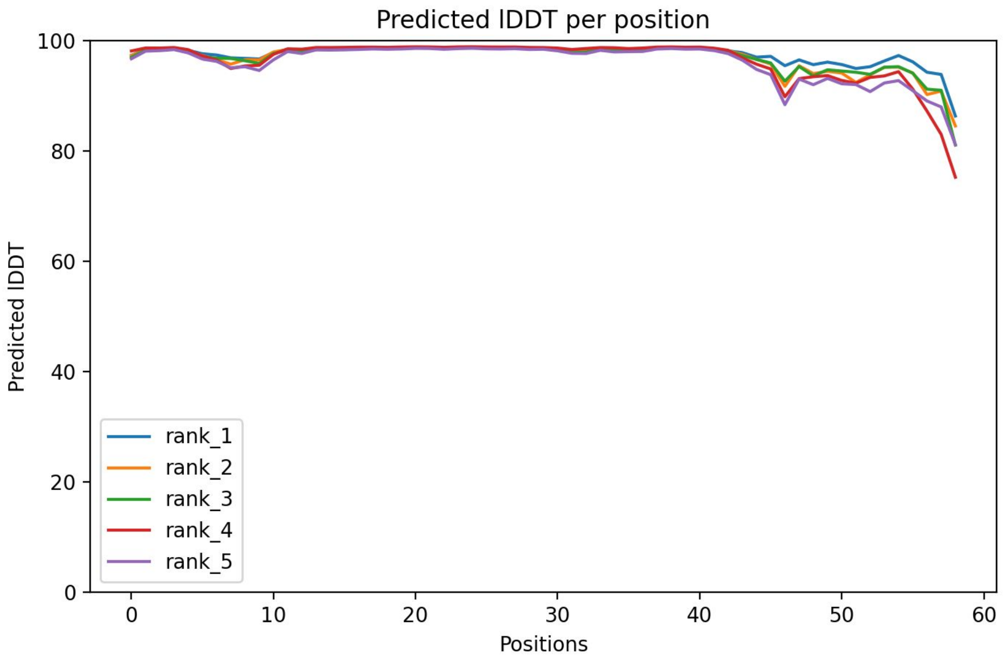


Confidence metrics

- **pLDDT** - "local" confidence per position
  - range 0 to 100 (higher better)
  - **Very low** (<50), **Low** (60), **OK** (70), **Confident** (80), **Very high** (>90)
  - Useful for deciding which local features (loops etc) are poorly modeled

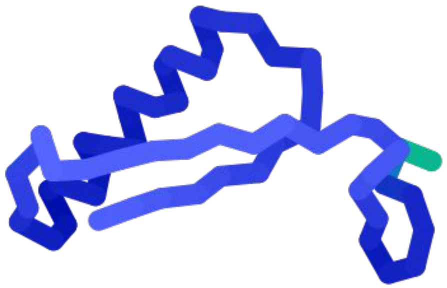


pLDDT 96.1



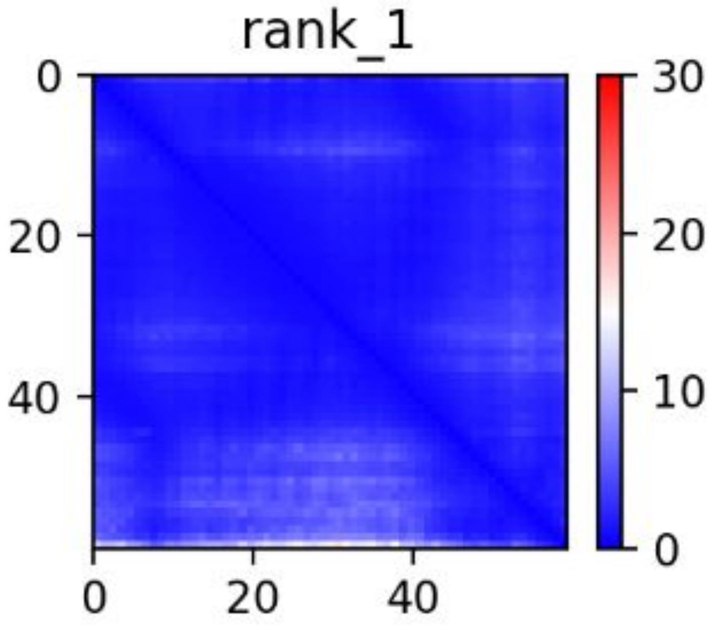
Confidence metrics

- pAE - confidence for every pair of positions
  - range 0 to 30 (lower better, in angstroms)
  - Useful for domain-domain or protein-protein interactions



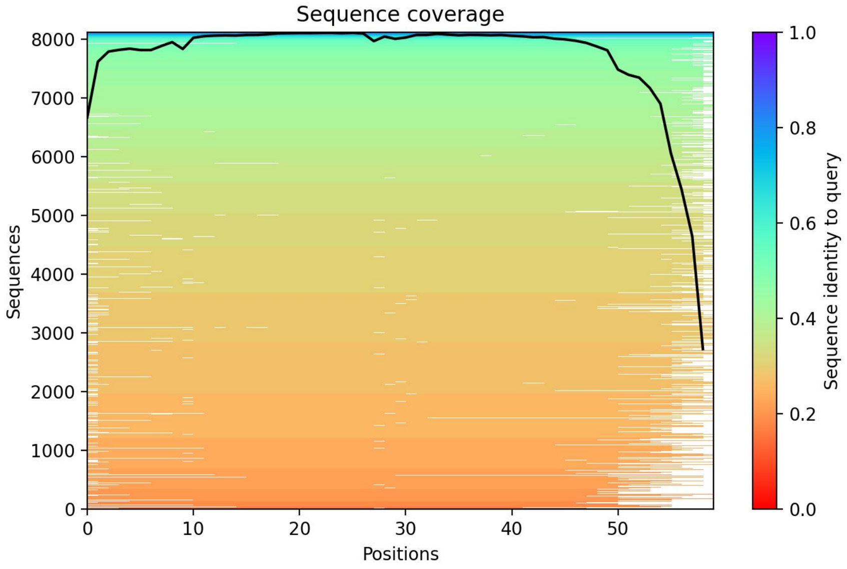
pLDDT 96.1

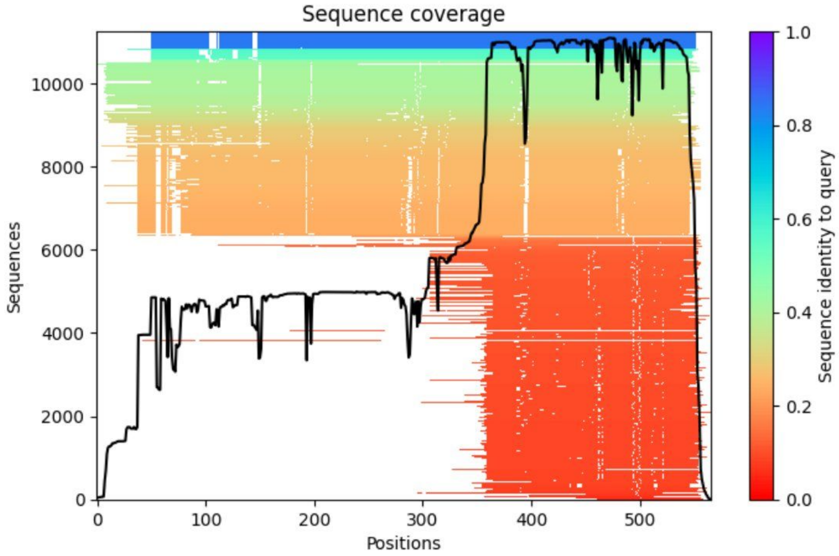
pTMscore **0.756**



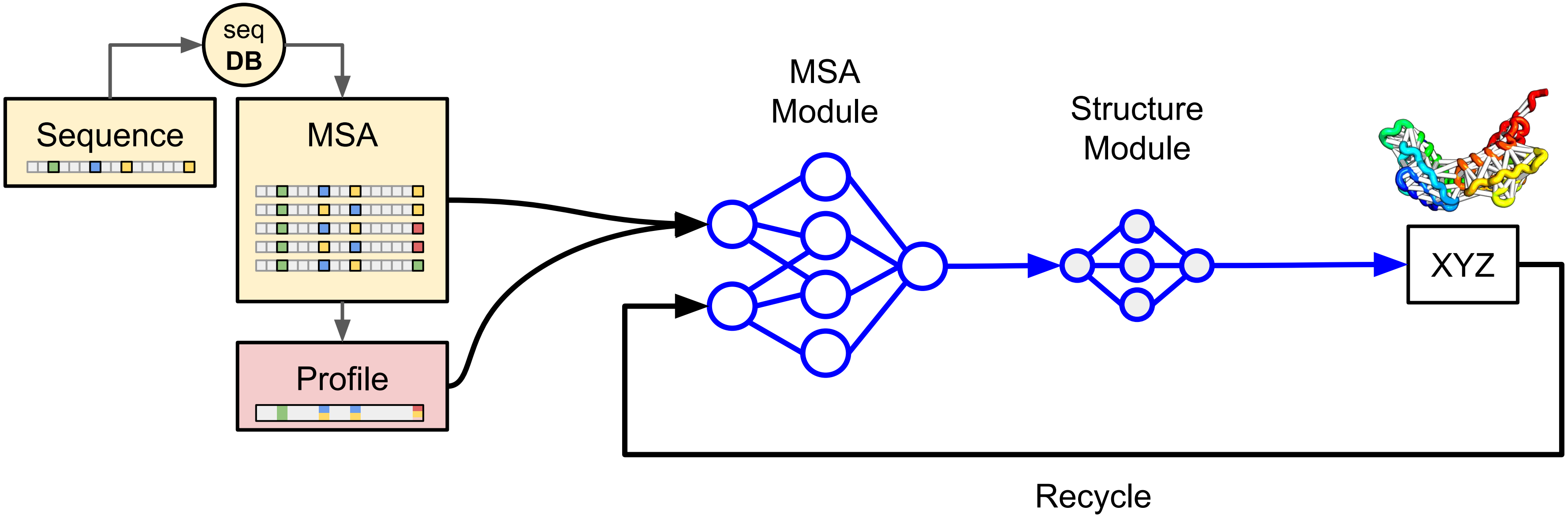
- pTM - predicted TMscore (integrates pAE values)
  - range 0 to 1 (higher better)
  - good as a single value to tell you how good the overall structure is.
  - recommend value for confident structure > 0.7

MSA plots very important to assess what info is available for the prediction!

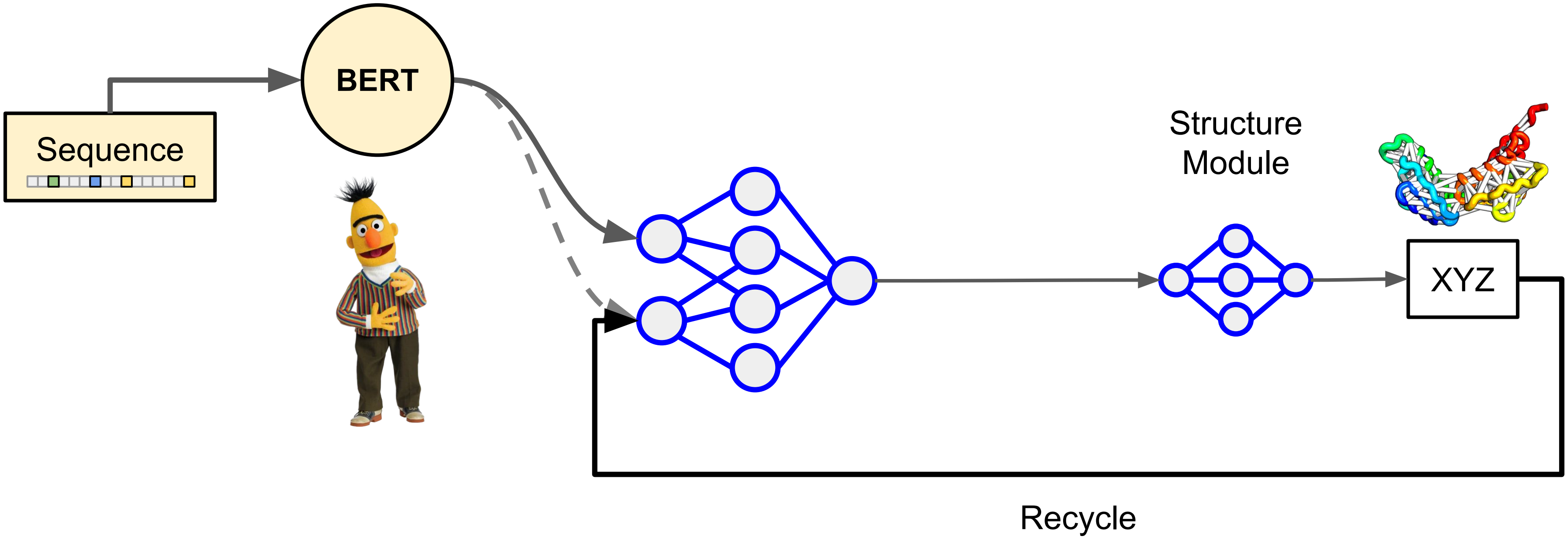




AlphaFold

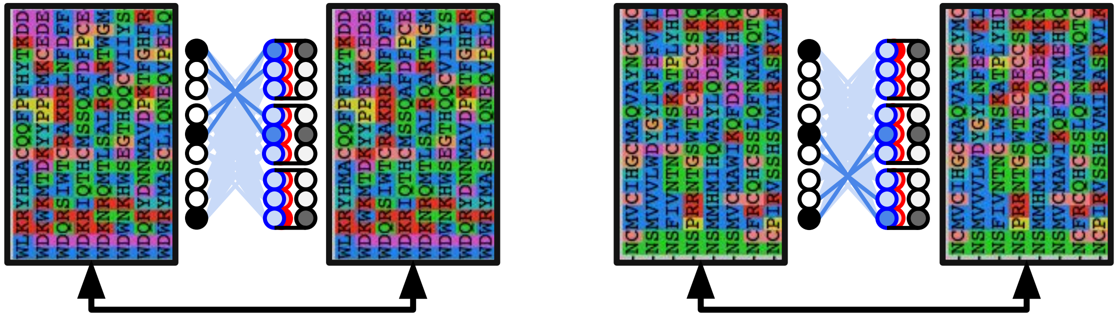

seq
DB
Sequence
MSA
Profile
MSA
Module
Structure
Module
XYZ
Recycle

ESMFold - Age of protein language models?

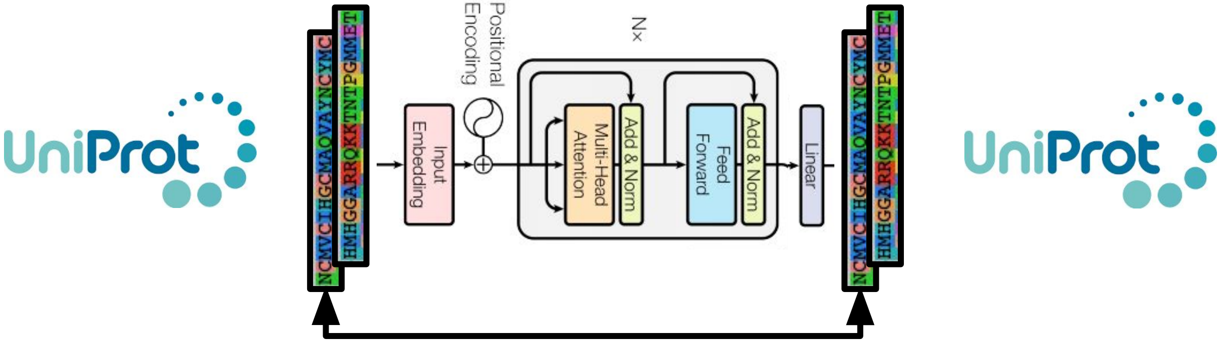


- Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B. and Ma, J., 2022. High-resolution de novo structure prediction from primary sequence. *BioRxiv*.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. and Rives, A., 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkare, A., Roye, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G.M. and Sorger, P.K., 2022. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, pp.1-7.

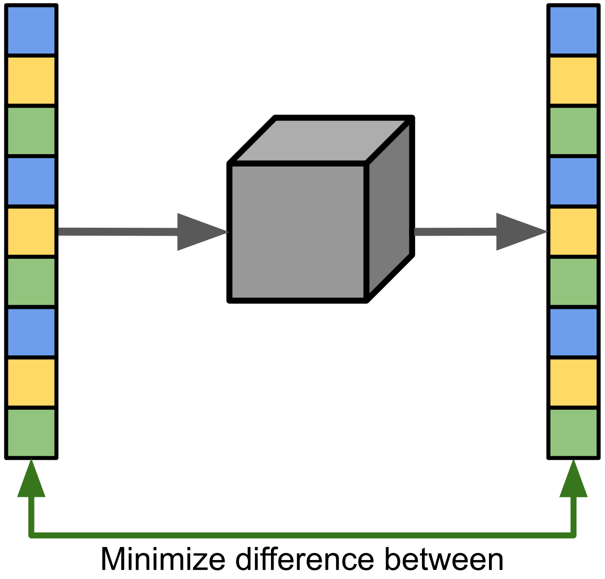
MRF - Different model for each MSA (or protein family)



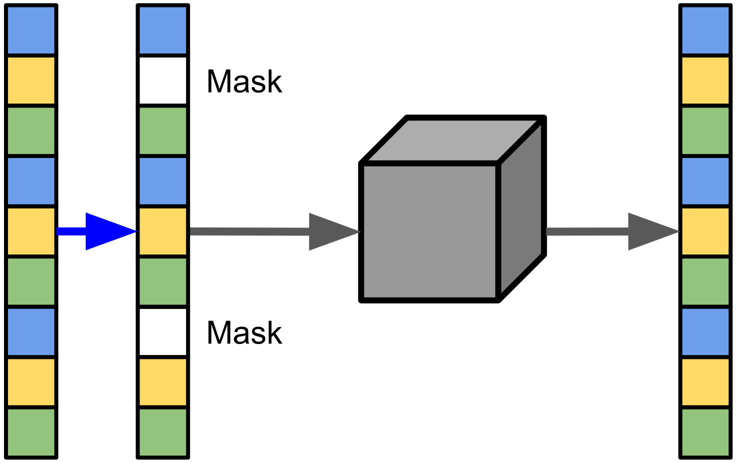
BERT - Same model for all sequences



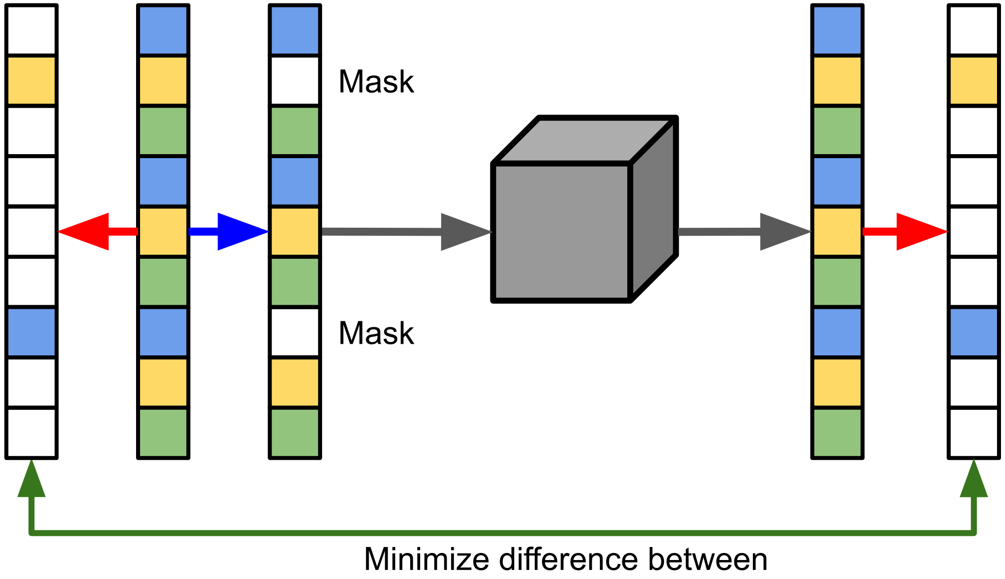
unsupervised



BERT (ESM1) - Masked language modeling (or self-supervised)



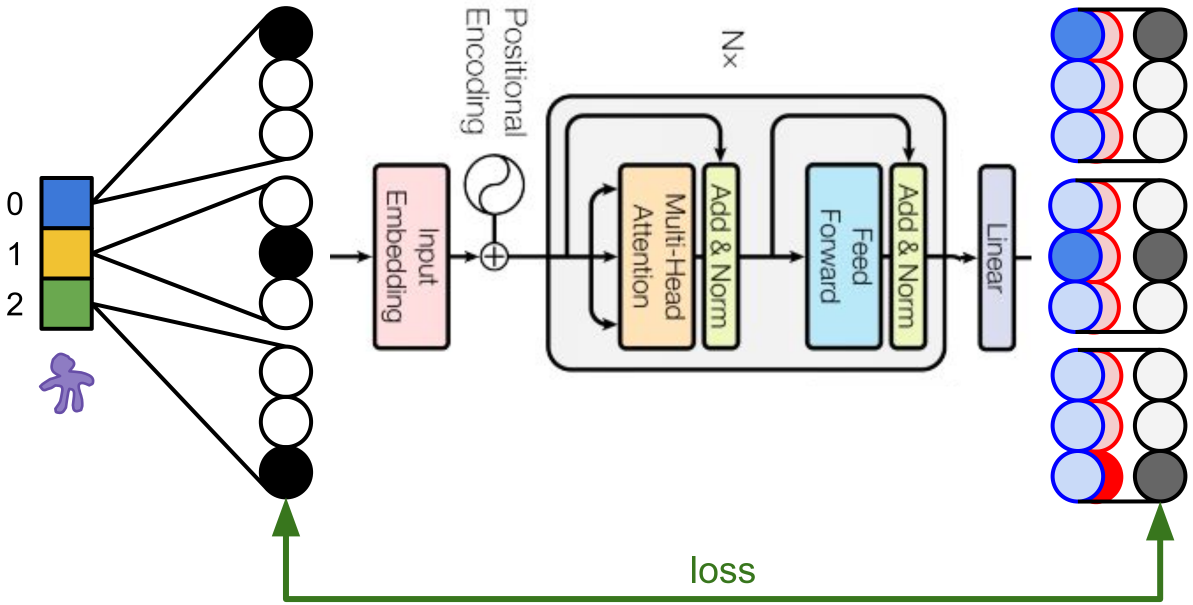
"Masked language modeling" is an approximation of "Pseudolikelihood"



$$\mathcal{L}_{PL}(\theta; x) = \sum_{i=1}^{L} \log p_\theta(x_i | x_{\setminus i})$$

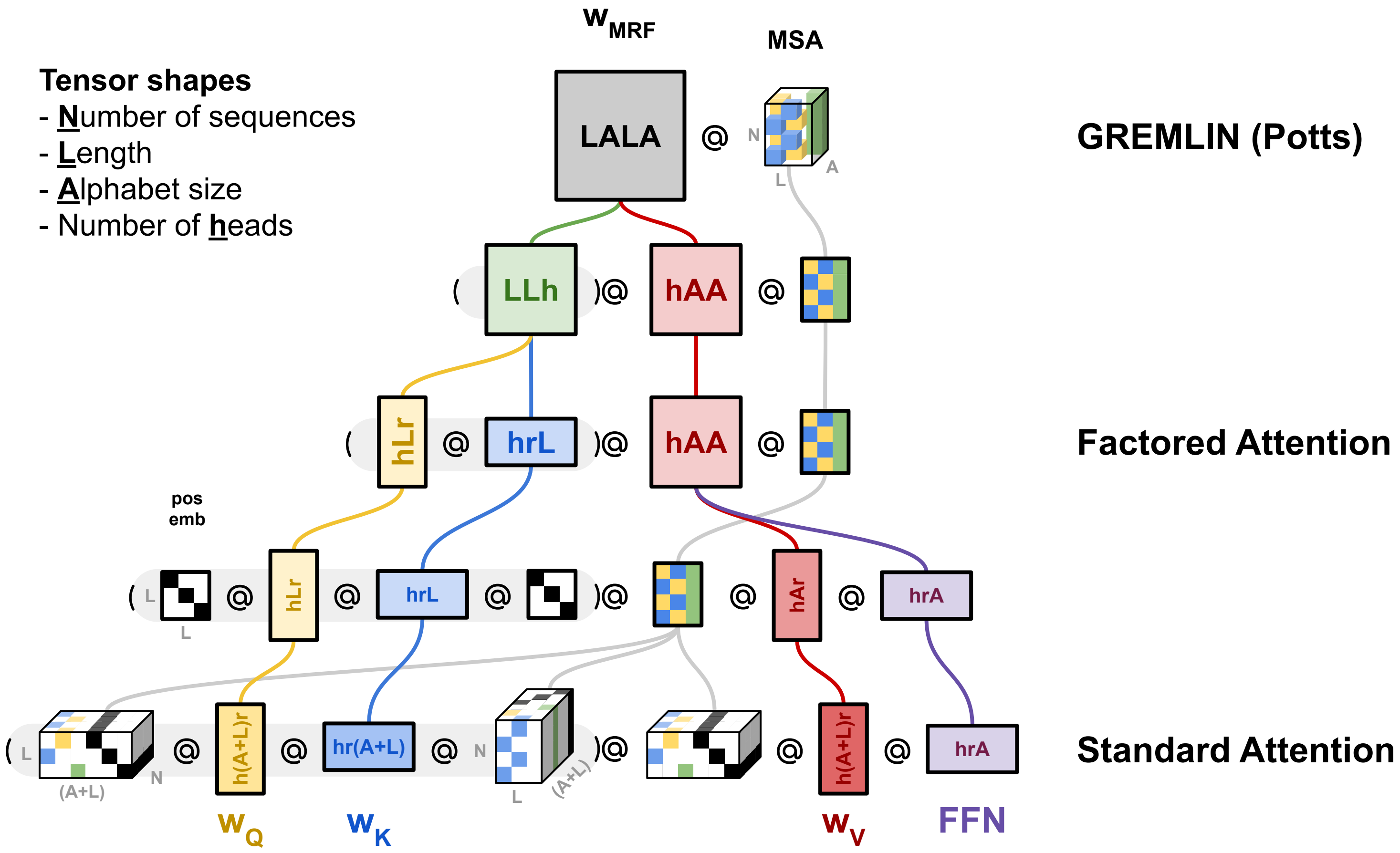
$$\mathcal{L}_{MLM}(\theta; x, M) = \sum_{i \in M} \log p_\theta(x_i | x_{\setminus M})$$

So where is it learning contacts?



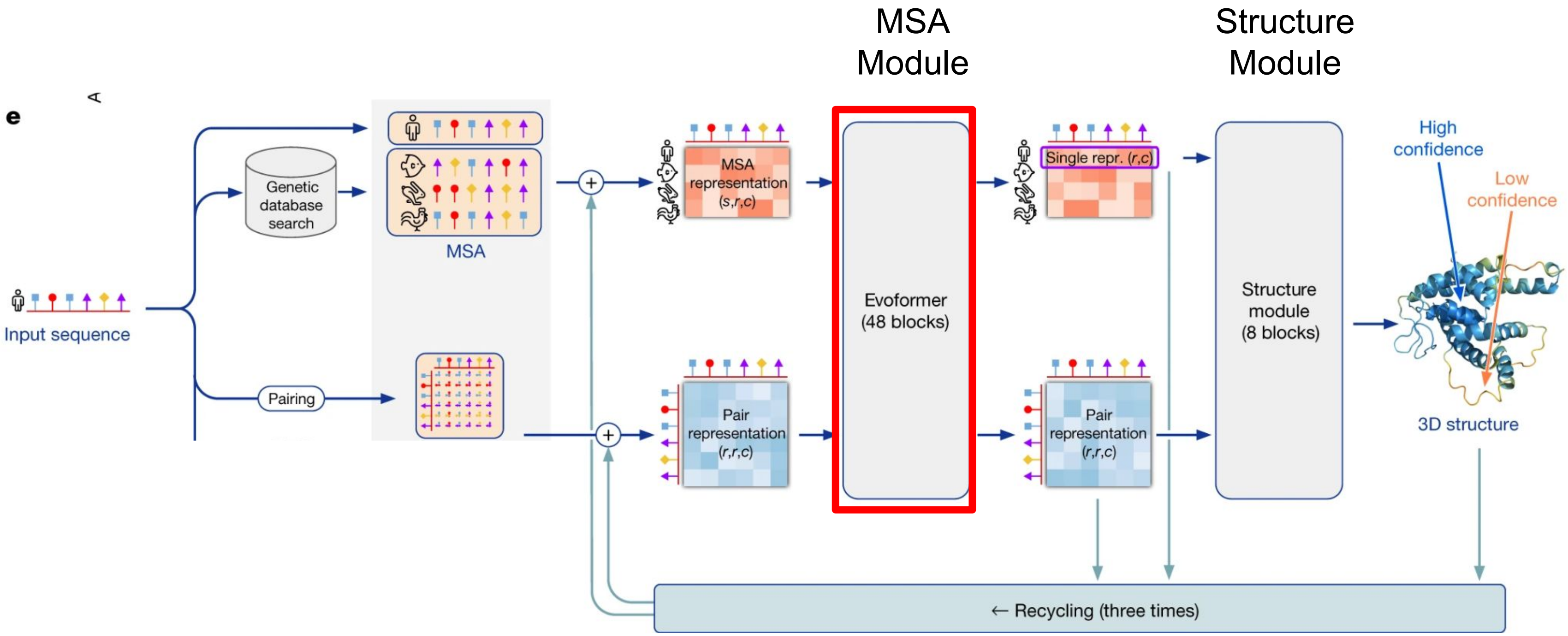
From GREMLIN to Standard Attention

Tensor shapes
- **N**umber of sequences
- **L**ength
- **A**lphabet size
- Number of **h**eads



https://pubmed.ncbi.nlm.nih.gov/34890134/

AlphaFold


MSA Module
Structure Module
e
Input sequence
Genetic database search
Pairing
MSA
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)