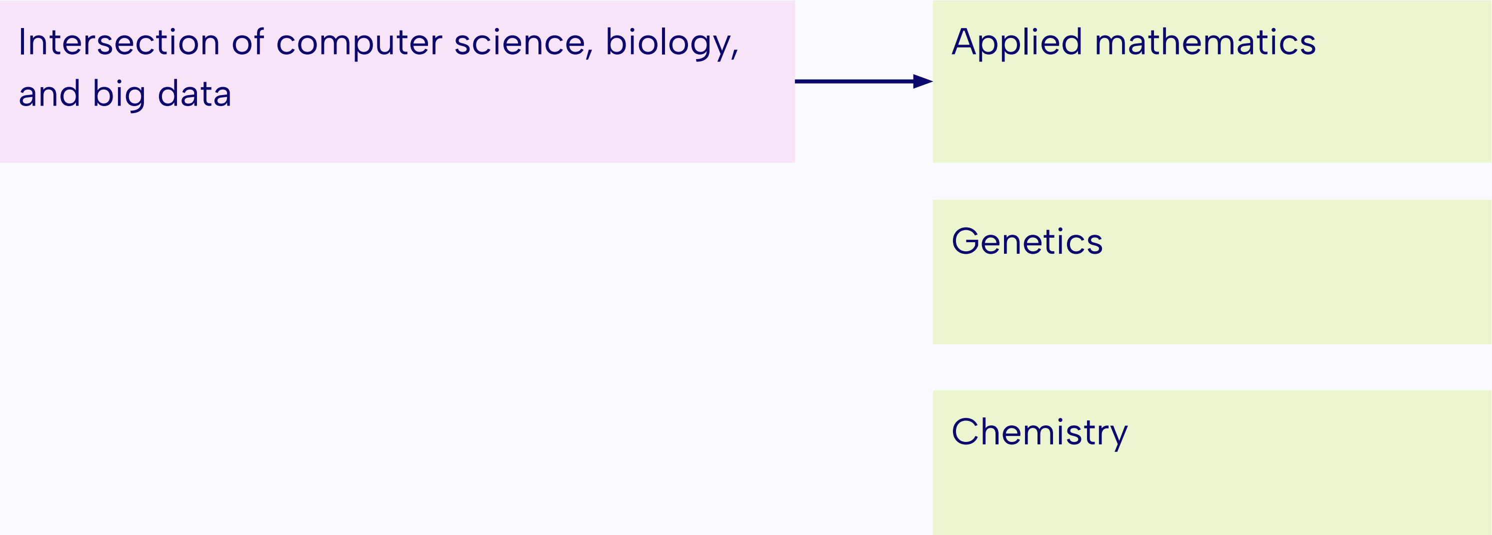
Generative and Geometric Deep Learning in Computational Biology

Darin Tsui

Computational biology



Intersection of computer science, biology, and big data

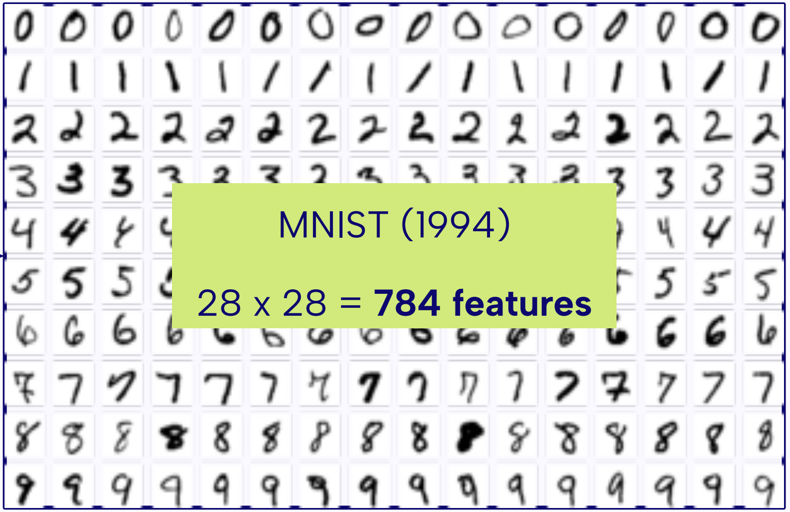
Applied mathematics

Genetics

Chemistry

Why machine learning?

Images:
high-dimensional data



MNIST (1994)

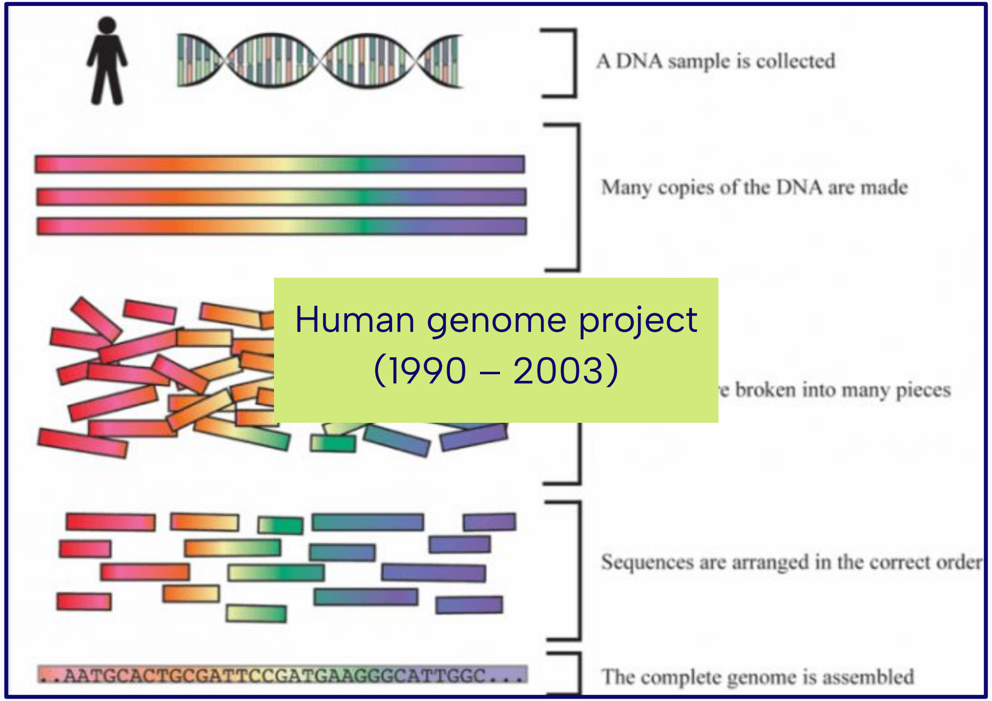
28 x 28 = **784 features**

Why machine learning?

3,200,000,000 nucleotides

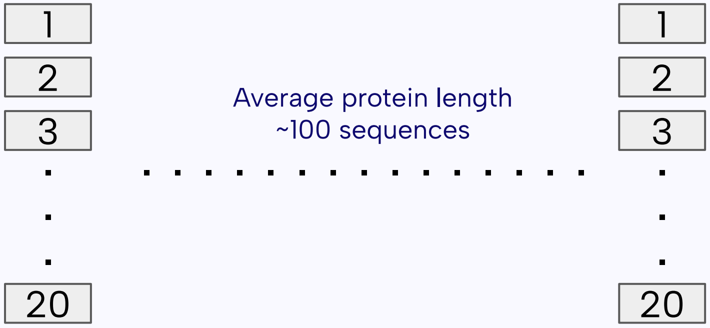
100,000 different proteins

20,000 – 25,000 genes


A DNA sample is collected
Many copies of the DNA are made
e broken into many pieces
Sequences are arranged in the correct order
..AATGCACTGCGATTCCGATGAAGGGCATTGGC...
The complete genome is assembled

Human genome project (1990 – 2003)

How many features?



Average protein length
~100 sequences

20 amino acids
found in proteins

20^{100} combinations

Over 98% of relevant protein variants still have unknown consequences*

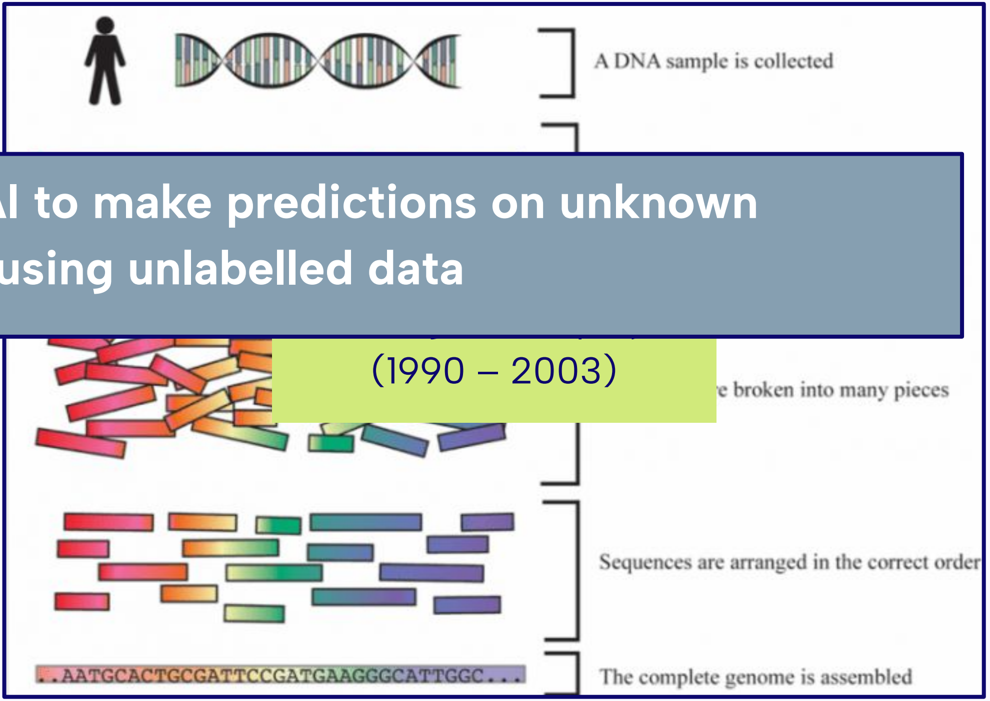
Frazer, J. *et al*. *Nature* (2021)

Why generative AI?

Labels are:

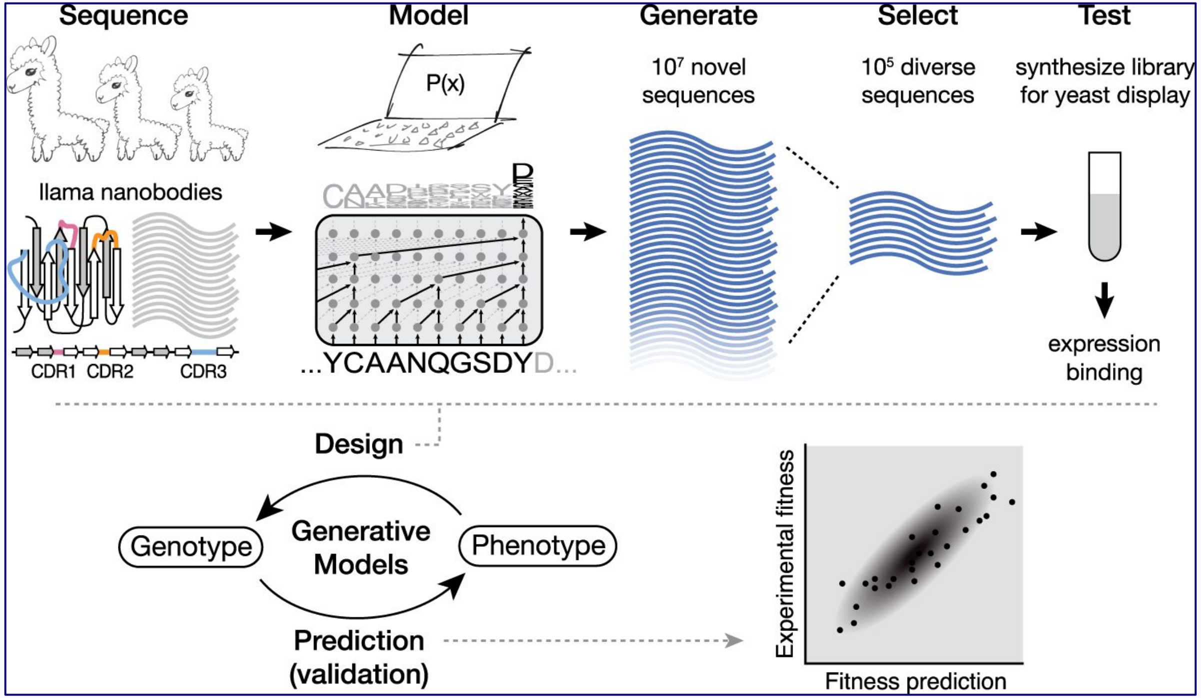
We can use generative AI to make predictions on unknown sequences using unlabelled data

- Biased
- Variable quality


A DNA sample is collected
e broken into many pieces
Sequences are arranged in the correct order
..AATGCACTGCGATTCCGATGAAGGGCATTGGC...
The complete genome is assembled

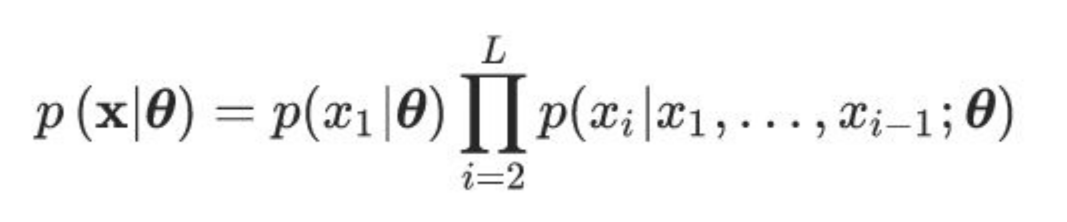
(1990 – 2003)

Autoregressive models for protein design



Shin, J. *et al. Nature Comm* (2021)

Autoregressive models for protein design



$$p\left(\mathbf{x}|\boldsymbol{\theta}\right) = p(x_1|\boldsymbol{\theta})\prod_{i=2}^{L} p(x_i|x_1,\ldots,x_{i-1};\boldsymbol{\theta})$$

x = sequence

θ = constraints for functional sequences

p(**x**|**θ**) = probability of sequence generation given evolution

Training set to predict masked values of amino acids

Shin, J. *et al. Nature Comm* (2021)

Autoregressive models for protein design

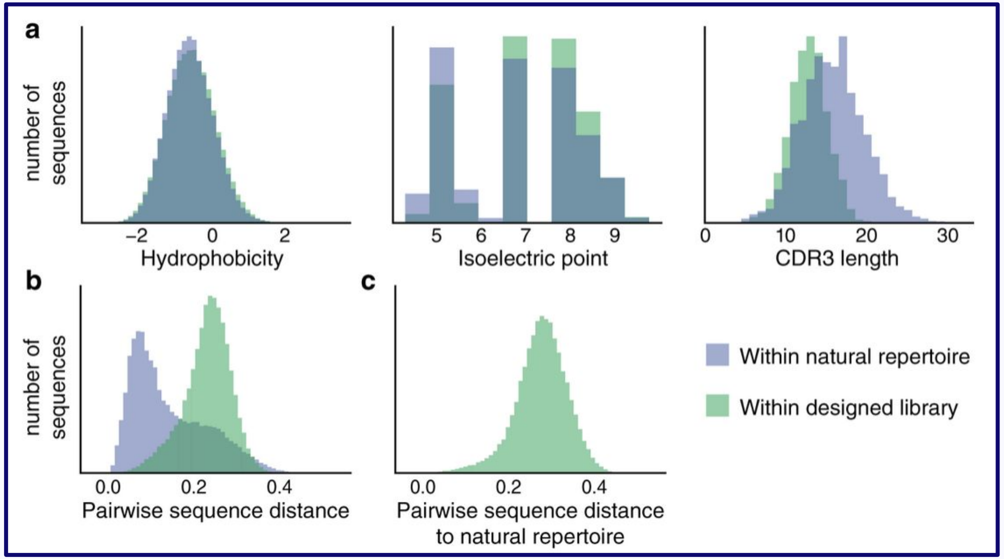
$$\log \frac{p(\mathbf{x}^{\text{Mutant}}|\boldsymbol{\theta})}{p(\mathbf{x}^{\text{Wild-type}}|\boldsymbol{\theta})}$$

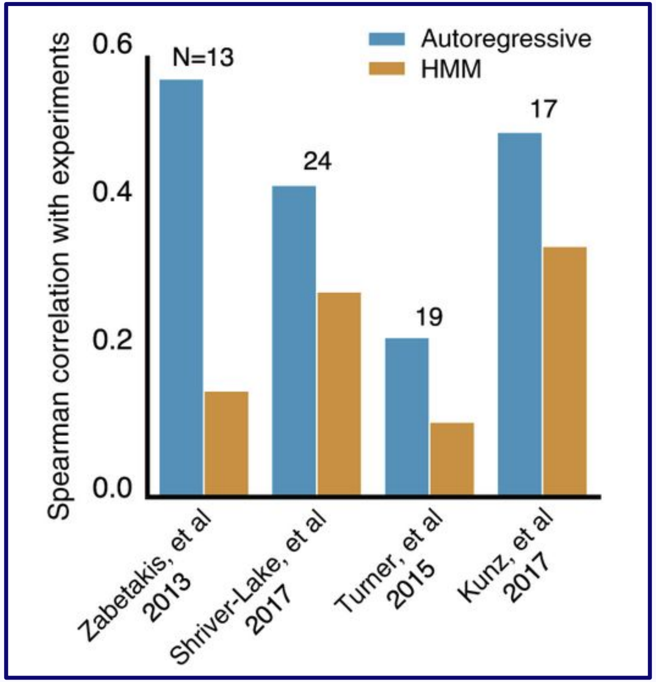
Predicting mutation effects using the **log-ratio of likelihoods**

- Summed cross-entropy between true vs predicted amino acids at each position, conditioned on previous amino acids

Shin, J. *et al. Nature Comm* (2021)

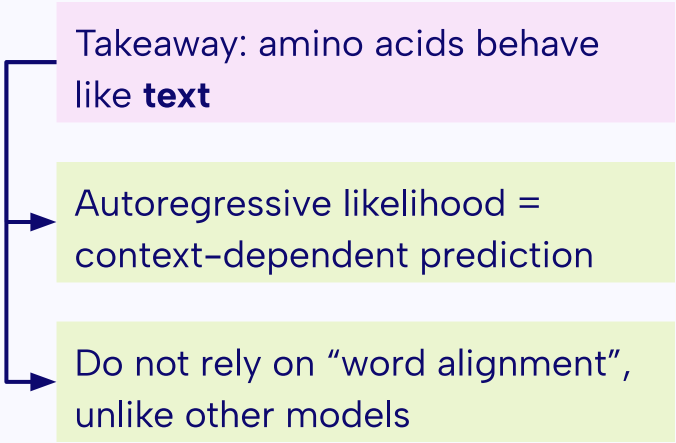
Autoregressive models for protein design





Shin, J. *et al*. *Nature Comm* (2021)

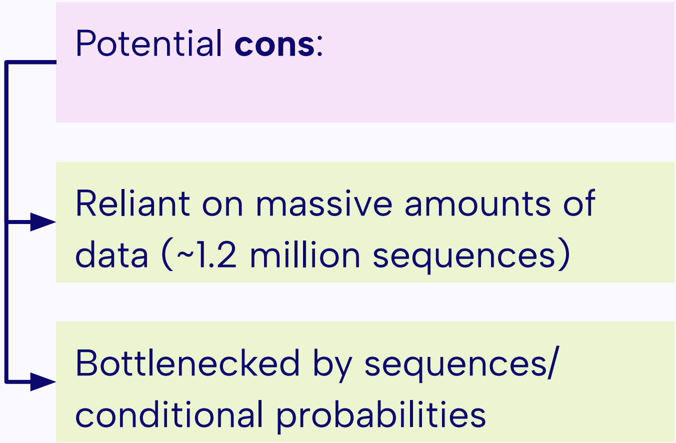
Autoregressive models for protein design



Takeaway: amino acids behave like **text**

Autoregressive likelihood = context-dependent prediction

Do not rely on “word alignment”, unlike other models



Potential **cons**:

Reliant on massive amounts of data (~1.2 million sequences)

Bottlenecked by sequences/ conditional probabilities

Shin, J. *et al. Nature Comm* (2021)

Alignment-based methods

Autoregressive Likelihood

Input sequences are of variable length

Raw sequences don’t capture evolutionary information

MTAIIKEIVSRNKRRYQED

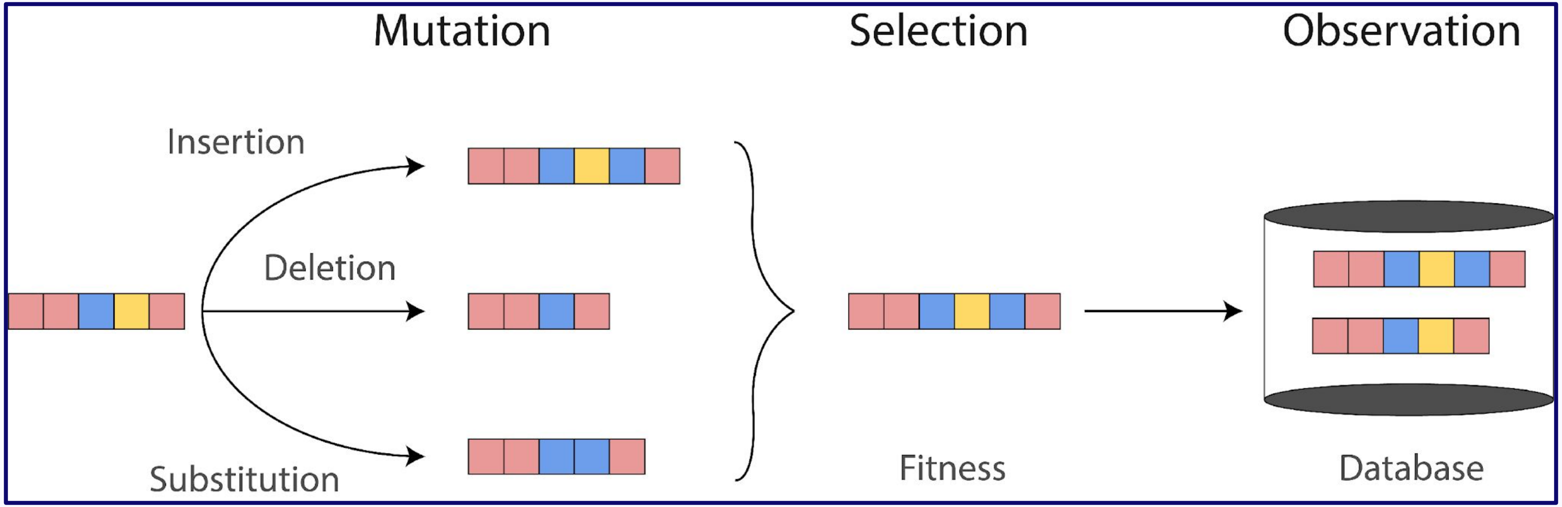
Multiple Sequence Alignment (MSA)

Fixed length

Sequences are clustered in “**protein families**”

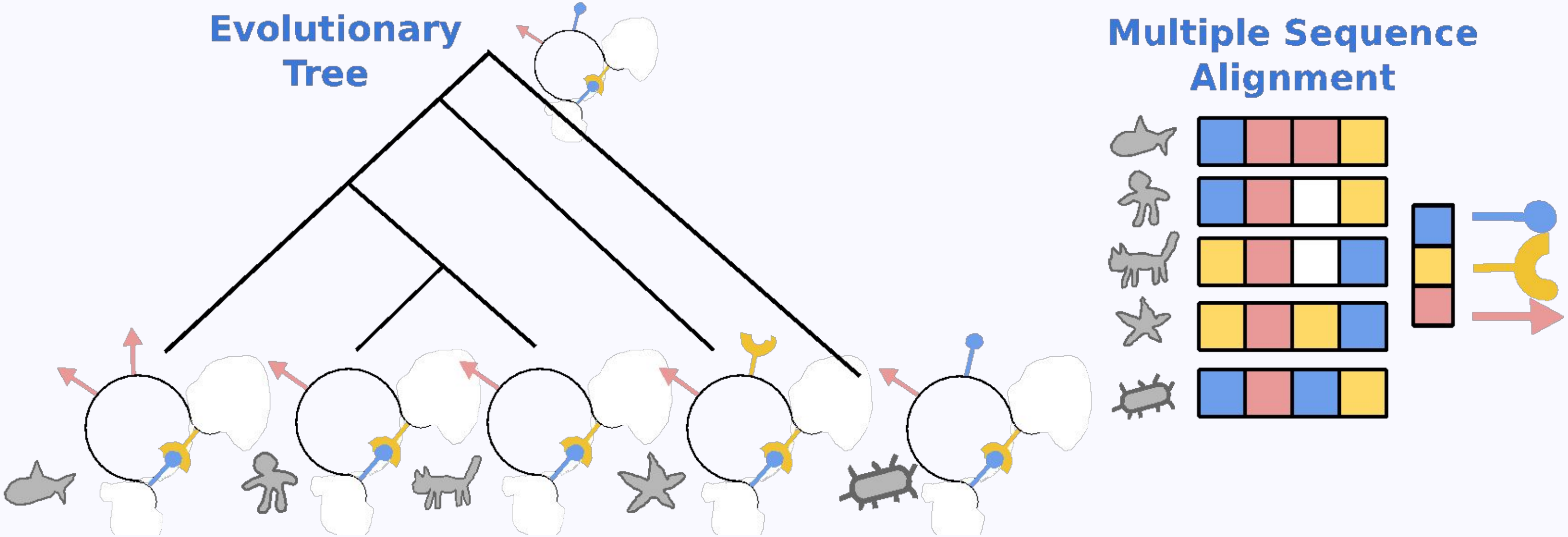
MTAIIKEIVSRNKRRYQED
MTAIIKEIVTTNKRRTQED
MTBIIKEIVSCNKRRTQED

Multiple Sequence Alignment (MSA)



Rao, R. (2021)

Multiple Sequence Alignment (MSA)

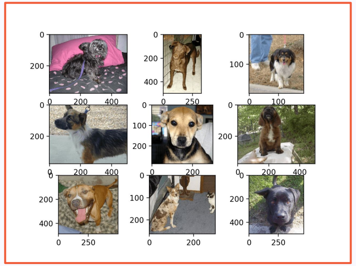


Query protein database for related sequences

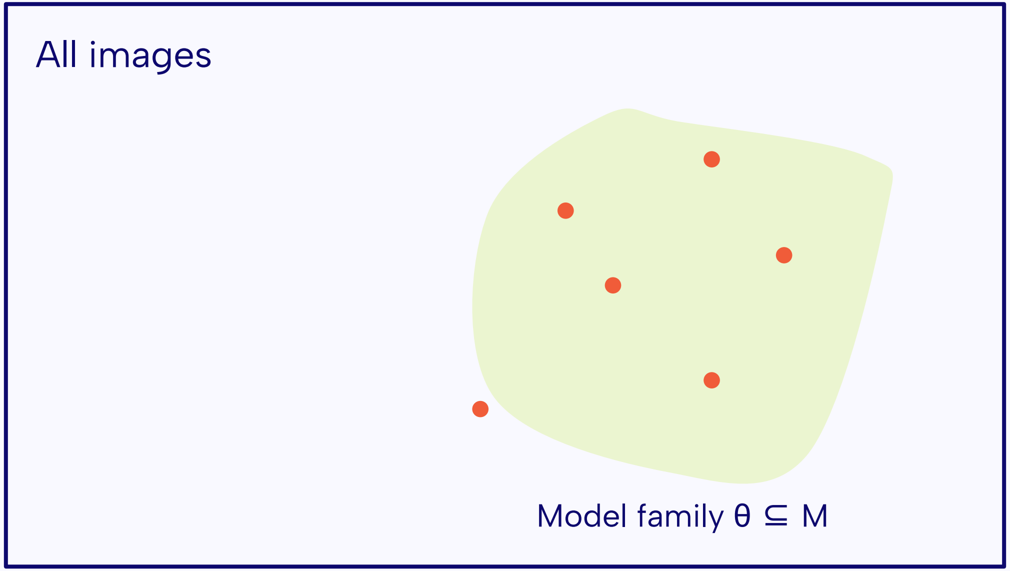
Align with heuristic based on edit distance

Rao, R. (2021)

MSA information to learn data distributions



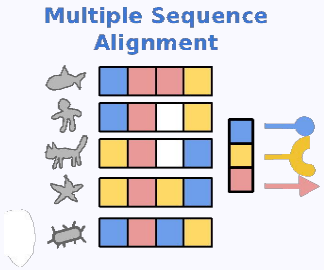
$x_i = P_{data}$
$i = 1, 2, \dots n$



All images

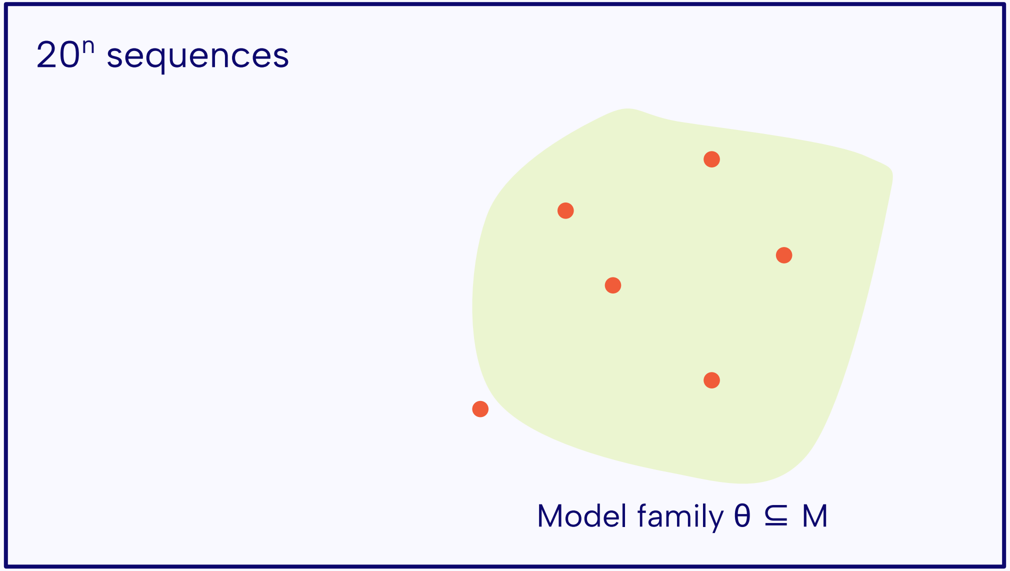
Model family θ ⊆ M

MSA information to learn data distributions



$x_i = P_{data}$

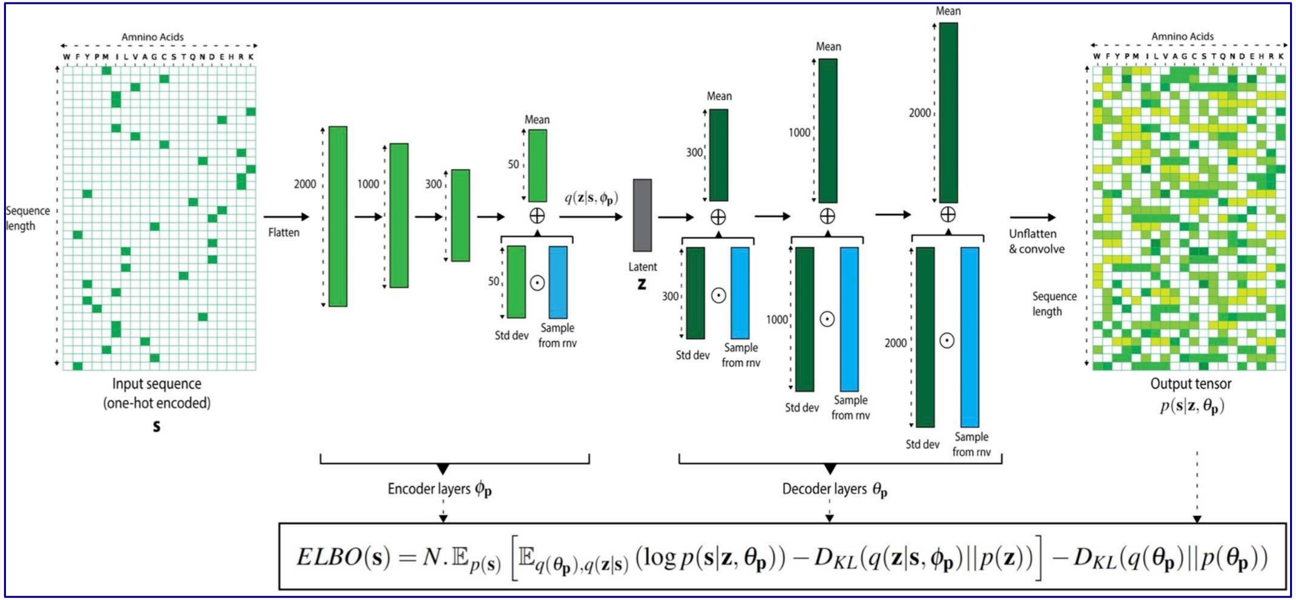
$i = 1, 2, \ldots n$


20^n sequences

Model family $\theta \subseteq M$

Rao, R. (2021)

Variational Autoencoders (VAEs) for variant prediction

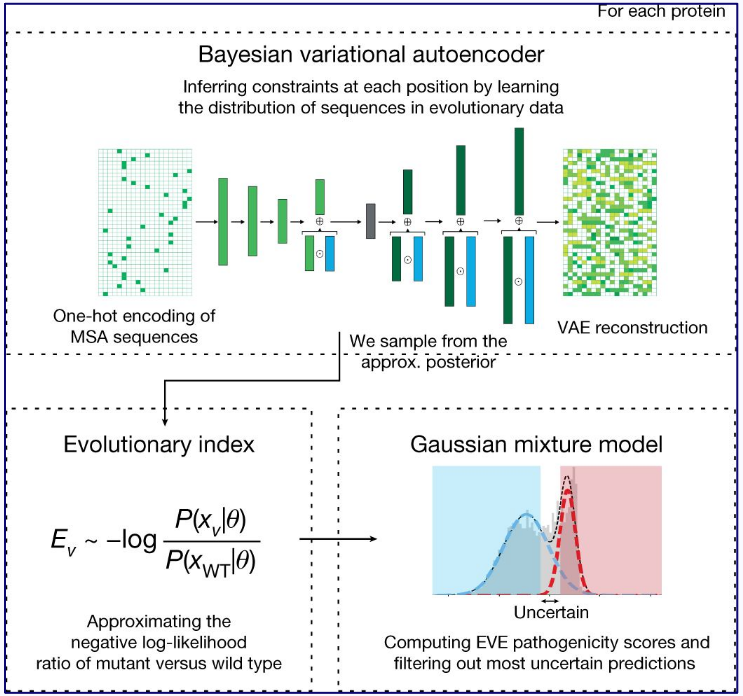


$$ELBO(\mathbf{s}) = N.\mathbb{E}_{p(\mathbf{s})}\left[\mathbb{E}_{q(\theta_\mathbf{p}),q(\mathbf{z}|\mathbf{s})}\left(\log p(\mathbf{s}|\mathbf{z},\theta_\mathbf{p})\right) - D_{KL}(q(\mathbf{z}|\mathbf{s},\phi_\mathbf{p})||p(\mathbf{z}))\right] - D_{KL}(q(\theta_\mathbf{p})||p(\theta_\mathbf{p}))$$

VAE learns a **protein family** distribution from MSA data

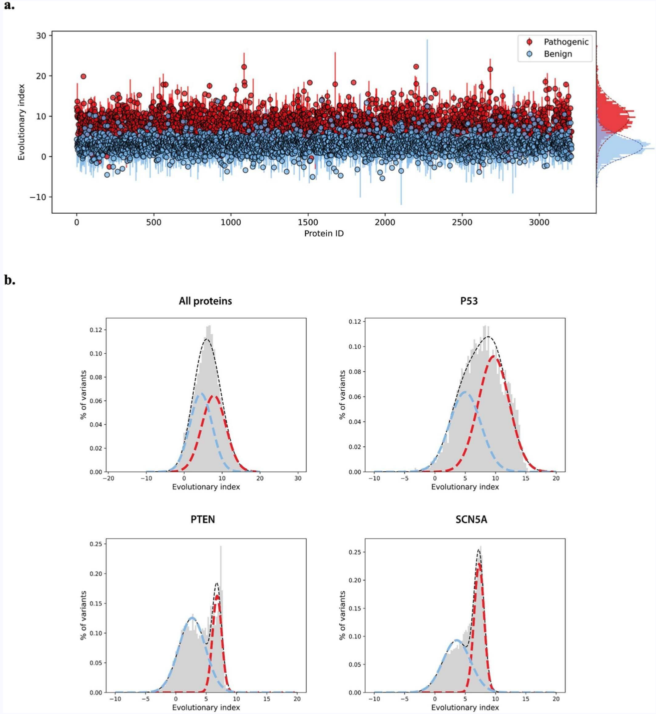
Frazer, J. *et al. Nature* (2021)

Variational Autoencoders (VAEs) for variant prediction


For each protein
Bayesian variational autoencoder
Inferring constraints at each position by learning the distribution of sequences in evolutionary data
One-hot encoding of MSA sequences
We sample from the approx. posterior
VAE reconstruction
Evolutionary index
$E_v \sim -\log \frac{P(x_v|\theta)}{P(x_{WT}|\theta)}$
Approximating the negative log-likelihood ratio of mutant versus wild type
Gaussian mixture model
Uncertain
Computing EVE pathogenicity scores and filtering out most uncertain predictions

Frazer, J. *et al*. *Nature* (2021)

Variational Autoencoders (VAEs) for variant prediction



Evolutionary Index (EI) = ELBO(**w**) - ELBO(**s**)

- **w** = wild-type (protein family “parent”)
- **s** = mutated sample

Frazer, J. *et al*. *Nature* (2021)

Variational Autoencoders (VAEs) for variant prediction

Takeaway: MSA information can be learned as a **distribution**

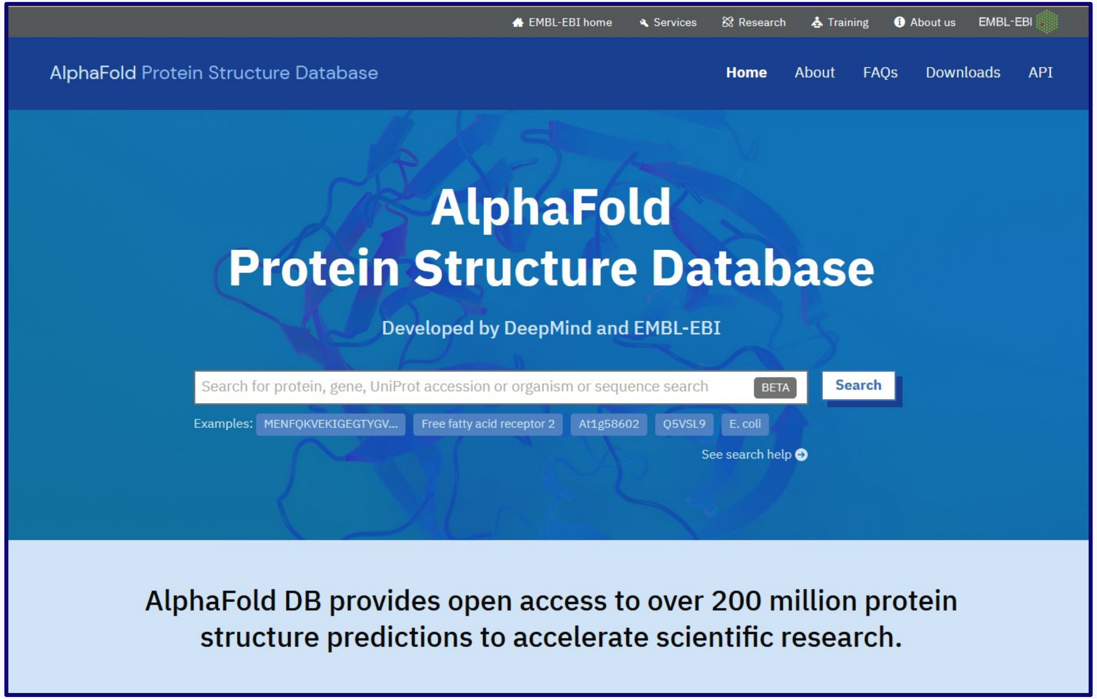
- MSA information generally implies better generalization
- **Significantly** less input data than in autoregressive models

Potential **cons**:

- Unclear what the VAE is exactly learning (how to validate?)
- How *precise* is the VAE with the limited data its given?

Frazer, J. *et al*. *Nature* (2021)

Geometric deep learning for protein structure prediction



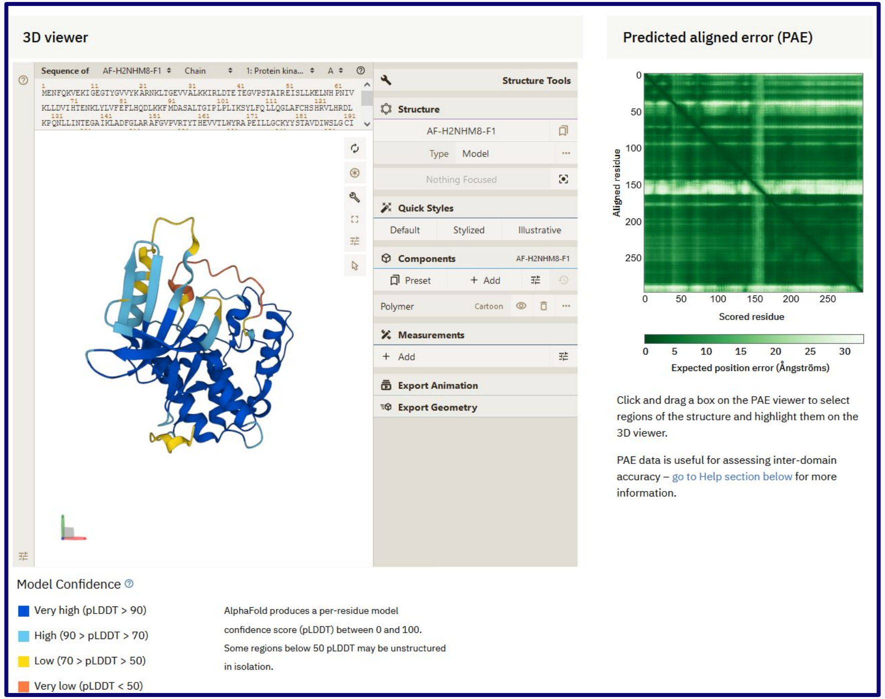
AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search
BETA
Search
Examples:
MENFQKVEKIGEGTYGV...
Free fatty acid receptor 2
At1g58602
Q5VSL9
E. coli
See search help

AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

Geometric deep learning for protein structure prediction



```
1         11        21        31        41        51        61
MENFQKVEKI GEGTYGVVYK ARNKLTGEVV ALKKIRLDTE TEGVPSTAIR EISLLKELNH PNIV
   71        81        91        101       111       121
KLLDVI HTENKLYLVF EFLHQDLKKF MDASALTGIP LPLIKSYLFQ LLQGLAFCHS HRVLHRDL
 131       141       151       161       171       181       191
KPQNLLINTEGA IKLADFGLAR AFGVPVRTYT HEVVTLWYRA PEILLGCKYY STAVDIWSLG CI
```

Click and drag a box on the PAE viewer to select regions of the structure and highlight them on the 3D viewer.

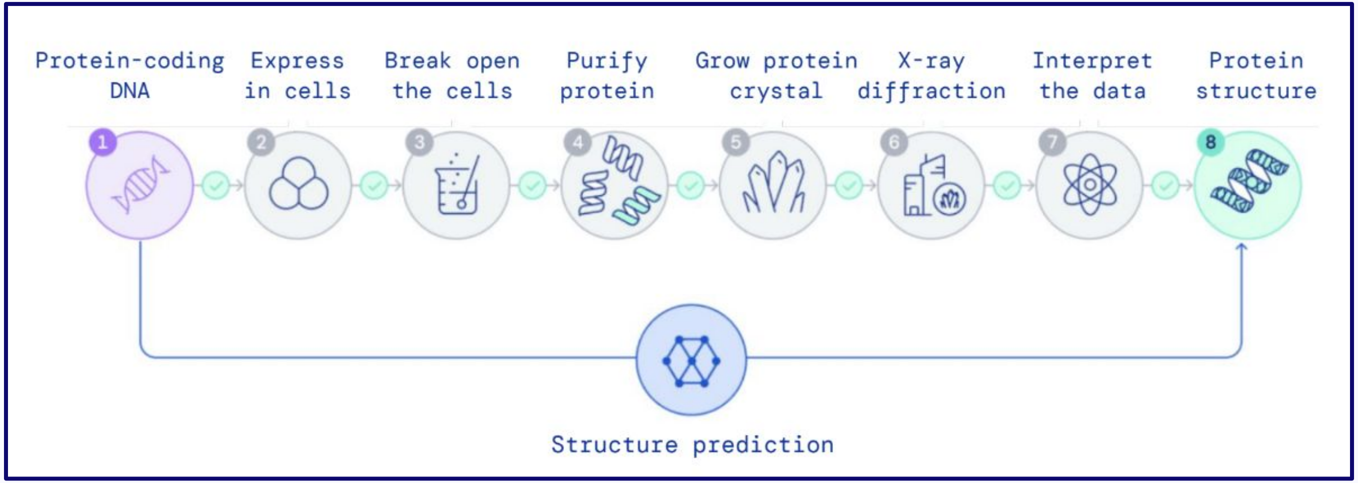
PAE data is useful for assessing inter-domain accuracy – go to Help section below for more information.

Model Confidence

- 🟦 Very high (pLDDT > 90)
- 🟦 High (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

AlphaFold produces a per-residue model confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

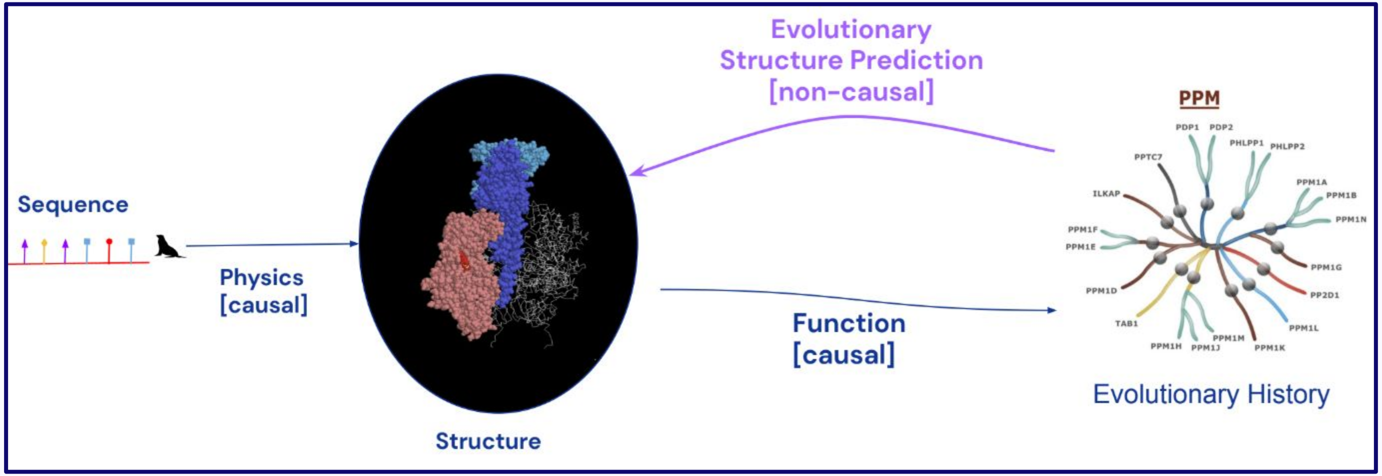
Why predict protein structures?



Extremely hard to determine experimentally. Only roughly about 200k protein structures available at the Protein Data Bank (PDB)

Bates, R. *et al. DeepMind* (2021)

Intuition meets machine learning



Evolutionary Structure Prediction [non-causal]

Sequence

Physics [causal]

Structure

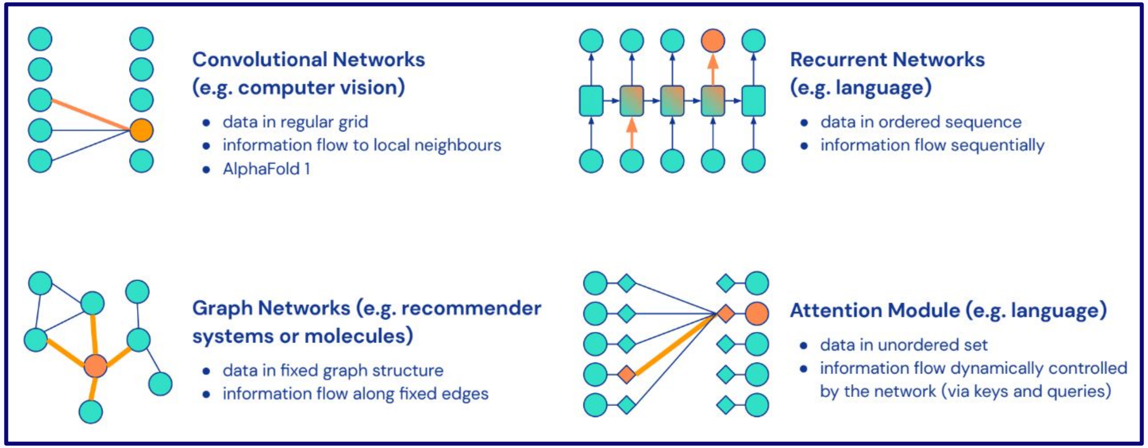
Function [causal]

Evolutionary History

Proteins conform to physics to go from a 2D sequence to a 3D structure. Evolutionary history gives us information about structure.

Bates, R. *et al. DeepMind* (2021)

Intuition meets machine learning



Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours
- AlphaFold 1

Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially

Graph Networks (e.g. recommender systems or molecules)

- data in fixed graph structure
- information flow along fixed edges

Attention Module (e.g. language)

- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)

Graph networks and **attention modules** best reflect our understanding of physics and geometry

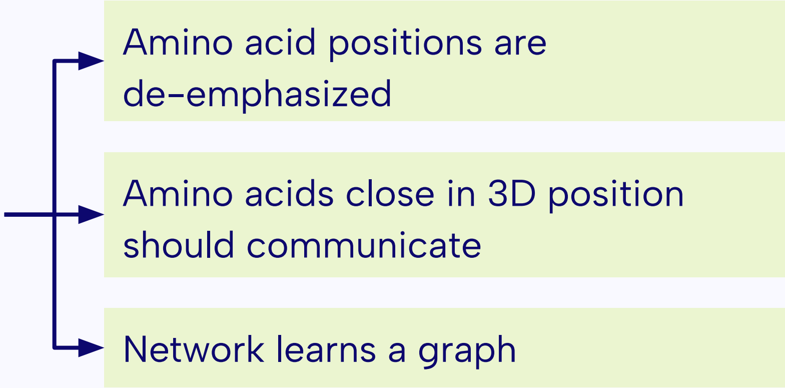
Bates, R. *et al. DeepMind* (2021)

Intuition meets machine learning

Input: Amino acid sequence

Output: 3D structure

Model should reflect our understanding of physics and geometry



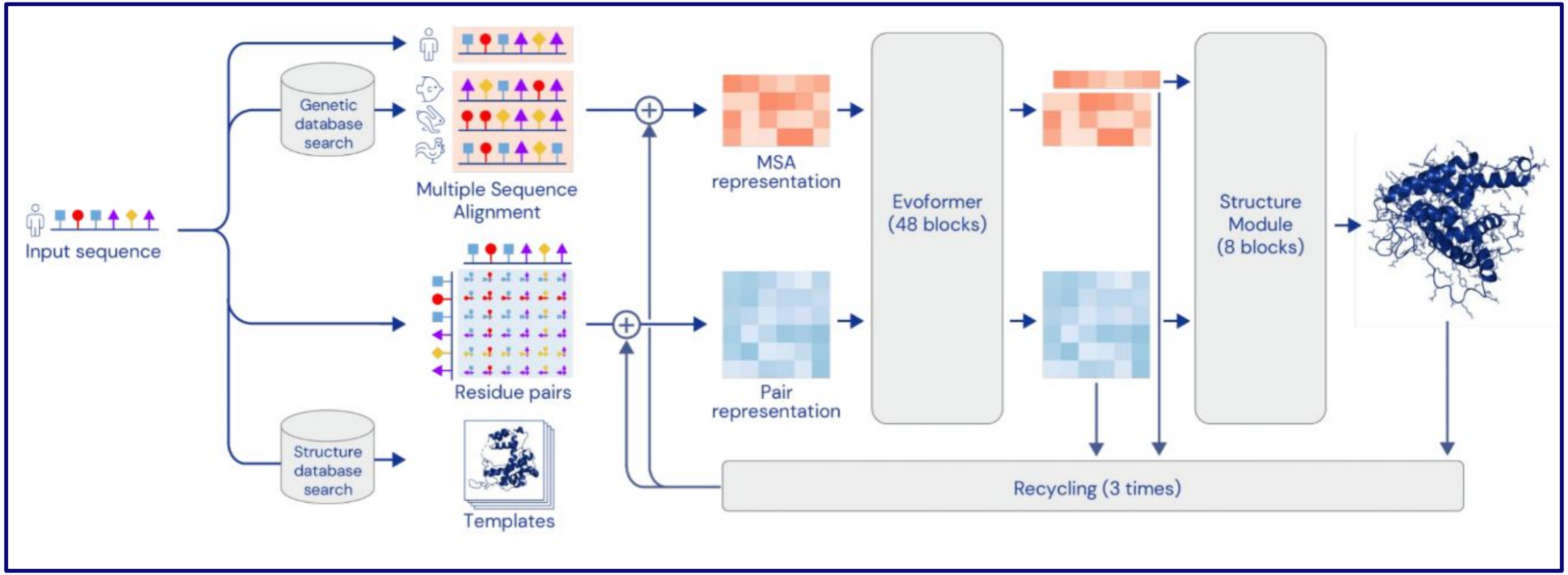
Amino acid positions are de-emphasized

Amino acids close in 3D position should communicate

Network learns a graph

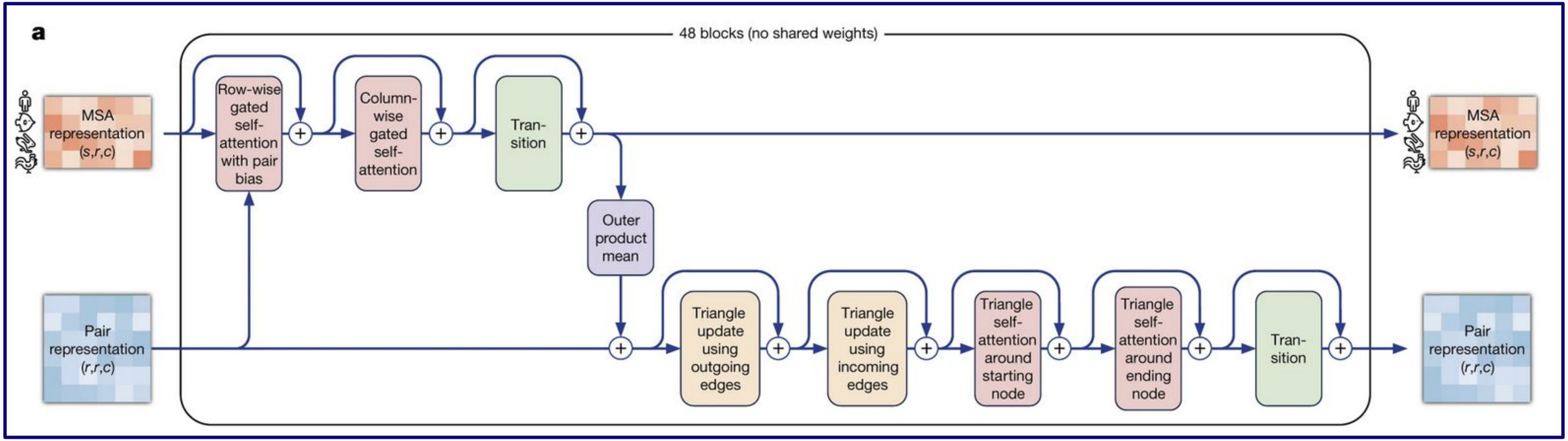
Bates, R. *et al. DeepMind* (2021)

AlphaFold Architecture



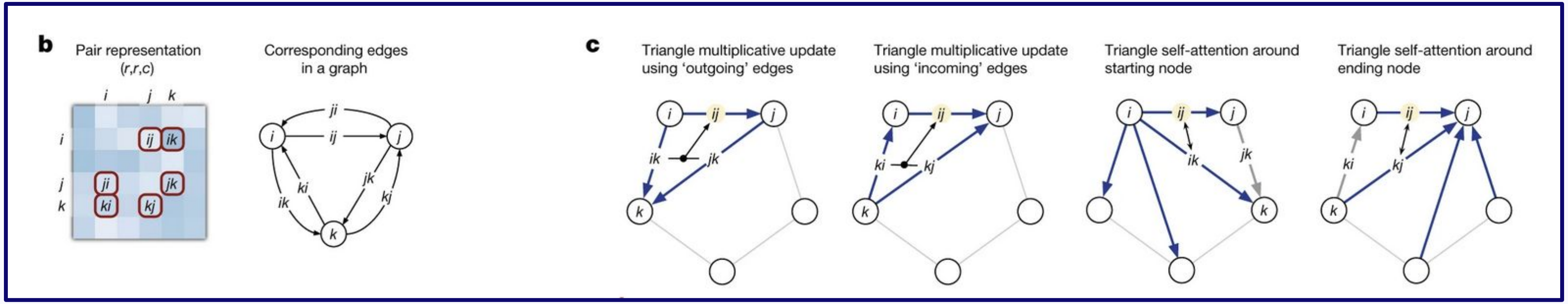
Bates, R. *et al. DeepMind* (2021)

Evoformer Architecture


a
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)

Jumper, J. *et al. Nature* (2021)

Evoformer - Triangular Attention

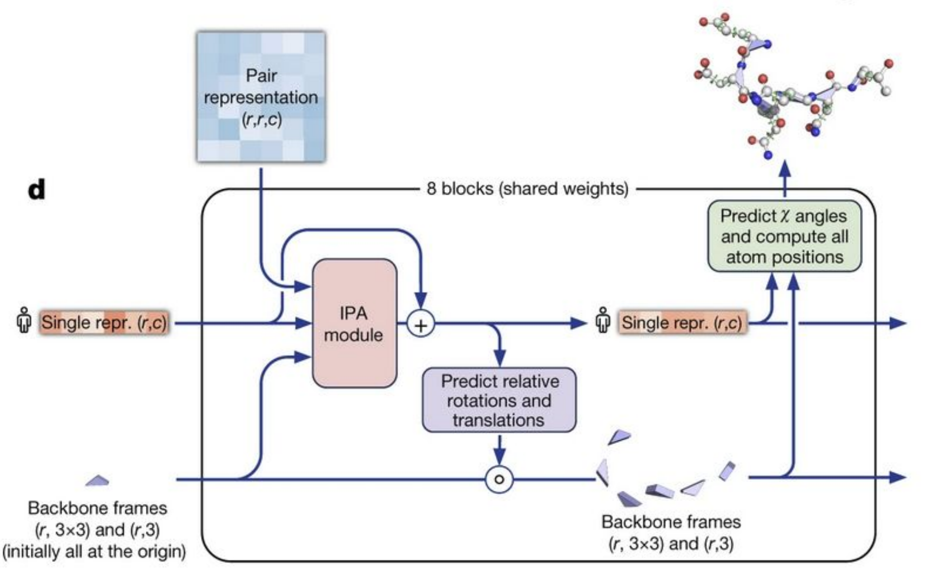


Purpose of pair representation is to encode distances between amino acids (i,j)

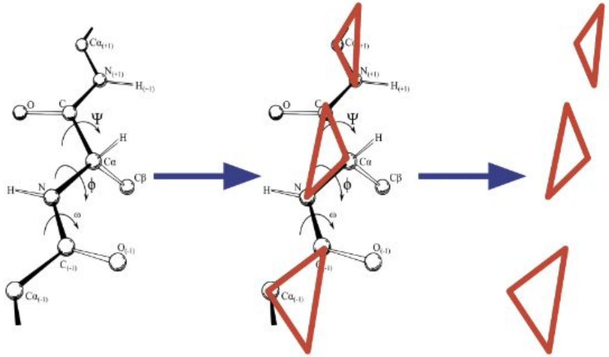
- Constrain these pairwise representations by the triangle inequality
- Allows for updating that's more consistent with 3D structure

Jumper, J. *et al.* *Nature* (2021)

Structure Module: Invariance and Equivariance


Pair representation (r,r,c)
d
8 blocks (shared weights)
Single repr. (r,c)
IPA module
Predict relative rotations and translations
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)

- **End-to-end folding** instead of gradient descent
- Protein backbone = gas of 3-D rigid bodies (chain is learned!)

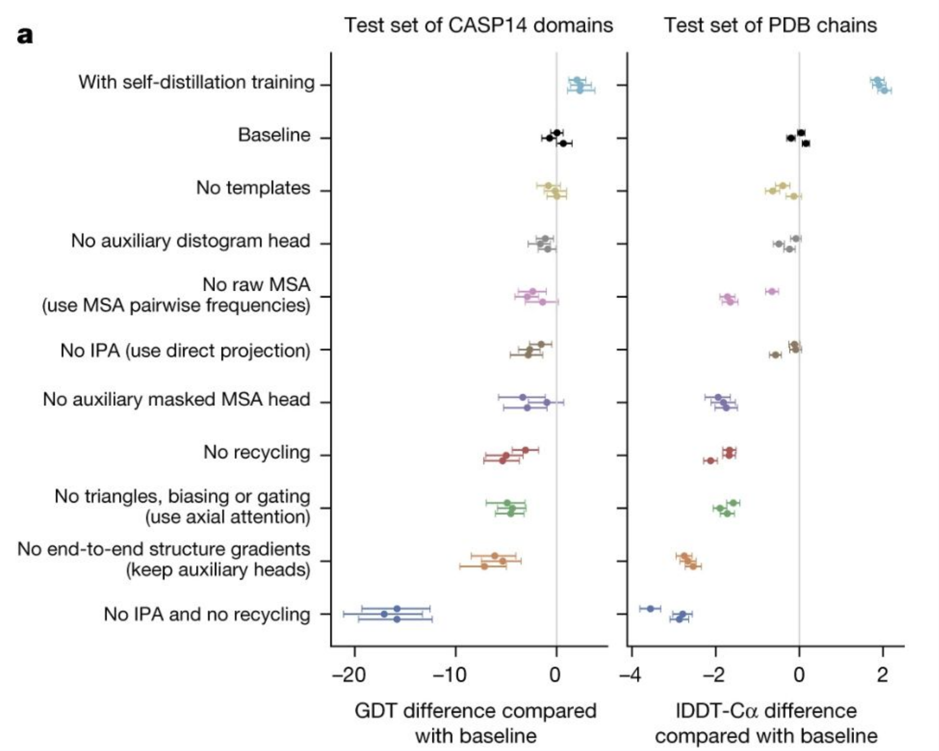


Invariant Point Attention (IPA): update seq representation without affecting 3D positions. Augments local frames of reference without affecting the global frame of reference

Equivariant update using updated sequence representation

Jumper, J. *et al. Nature* (2021)

A collection of innovations



Jumper, J. *et al. Nature* (2021)

AlphaFold for Structure Prediction

Takeaway: no one change made structure prediction possible

- Emphasis on geometric constraints
- A need for more “smarter” neural networks

Potential **cons**:

- Constrained by high-depth MSA data
- Slow inference time

Jumper, J. *et al*. *Nature* (2021)

What’s the future of computational biology?

Need for **interpretable** algorithms

Need for **scalable** algorithms

Need for **smarter** algorithms

Thank you!