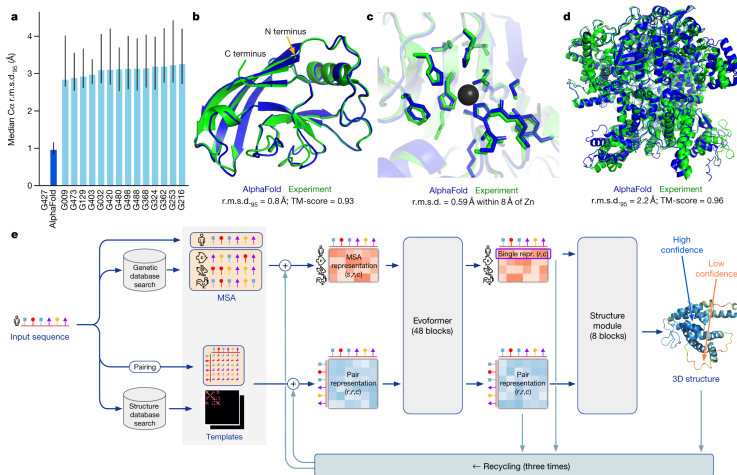


- **Complexity:** DL excels with complex data (images).
- **Data Size:** Takes advantage of big data (Uniprot  $10^8$  sequences).
- **Automatic Features:** No manual feature crafting needed.
- **Superiority:** Outperforms ML in image and text generation.
- **End-to-End:** Direct input-output, e.g., AlphaFold for protein folding.



(generated with midjourney)

## AlphaFold, AI to predict the fold of globular proteins at the experimental accuracy

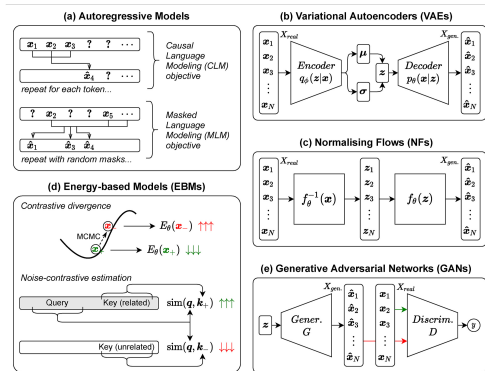


J. Jumper *et al*, 2020, Nature (And > 50 years of data collection)

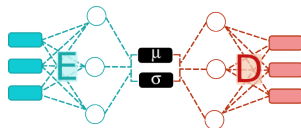
## **Generative models to sample sequences**

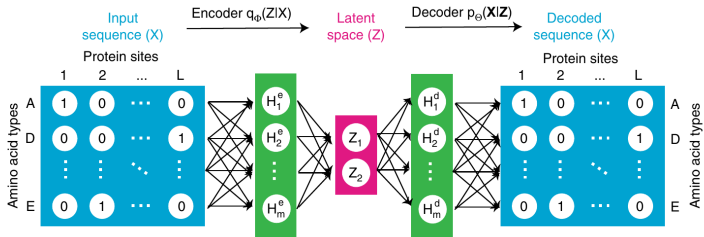
# Deep Generative Models

- **VAEs**: Probabilistic; encoder-decoder structure.
- **GANs**: Generator vs. discriminator competition.
- **RBMs**: Energy-based with visible/hidden layers.
- **Normalizing Flows**: Complex distribution transformations.
- **Autoregressive**: Sequence prediction.
- **Energy-Based Models (EBMs)**: Learn energy functions.



- VAEs: Generative models that learn to encode and decode data.
- Difference from standard autoencoders: Introduces probabilistic encoding.
- Application: Generating functional protein sequences efficiently.





- Linear dense NN:  $H = W \times S + b$
- Activation function:  $ReLU(H_i) = \max(0, H_i)$
- Output function:  $Softmax(X) = \hat{S} = \frac{\exp\{X_i\}}{\sum_j \exp X_j}$
- $Z \sim \mathcal{N}(\mu = 0, \sigma^2 = Id)$

$$ELBO(\theta, \phi) = \sum_Z q_{\phi}(Z|X) \log p_{\theta}(X|Z) - \sum_Z q_{\phi}(Z|X) \log \frac{q_{\phi}(Z|X)}{p_{\theta}(Z)} \quad (1)$$

$$ELBO(\theta, \phi) = \langle \mathcal{L}(\hat{S}) \rangle - D_{KL}(Encoder(S) || \mathcal{N}(Z|\mu, \sigma^2))$$

# Training of a VAE

## Architecture chosen by 5-fold cross validation

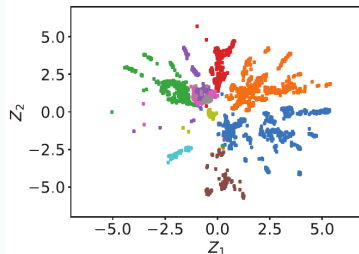
- $1 \times$  hidden (linear) layer fully connected (Encoder and Decoder)
- 512 units per hidden layer
- Latent space dimension = 10
- Parameter space (W) dense layer:  $L \times 21 \times 512 \rightarrow 10^6$

## Training parameters

- Re-weighting sequences (reduce redundancy and emphasize diversity)
- $10^4$  optimization steps (ADAM optimizer)
- Regularization

## Latent space representation of sequence space

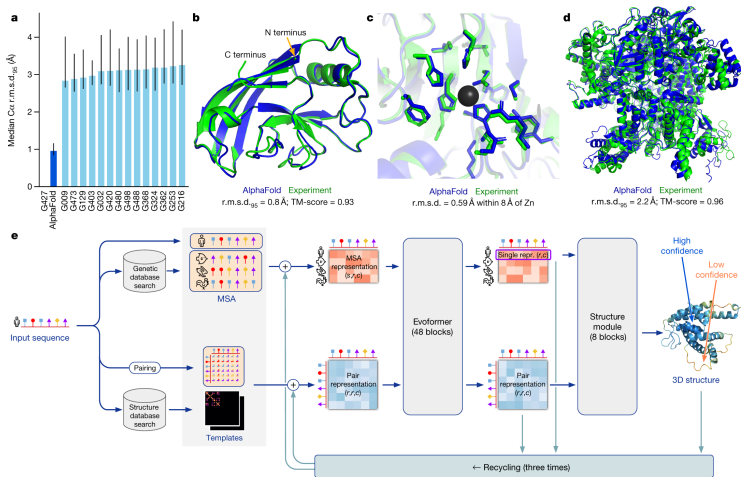
[Ding *et al*, 2020, Nat. Com.]



**AlphaFold is so good at predicting the structure, can't we just invert it? Yes, we can!**

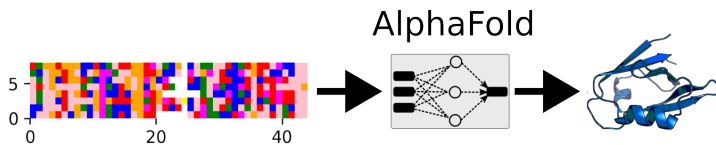


# AlphaFold structure prediction



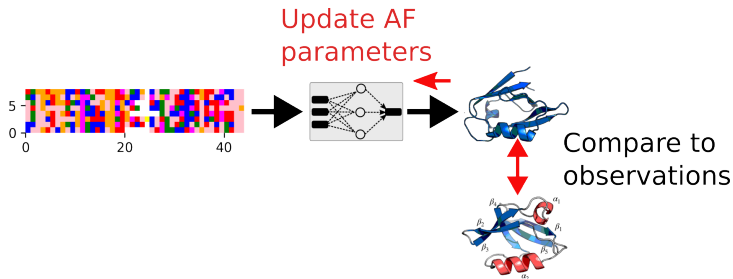
End-to-End prediction model accounting for evolutionary information as well as geometric information → Any parameter on the way is differentiable, even the input sequences.

Prediction of a structure's fold.

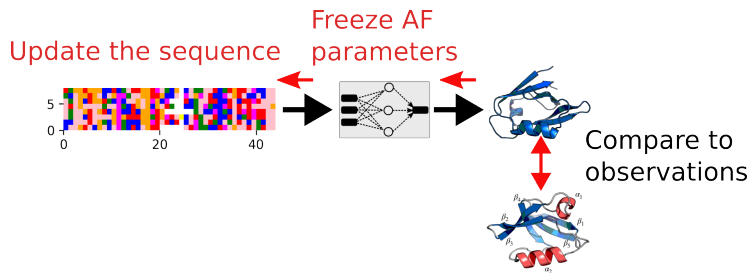


(J. Jumper *et al*, 2020, Nature)

Compare the prediction to the true structure and update accordingly the parameters of AlphaFold using the **gradient** of the loss function.

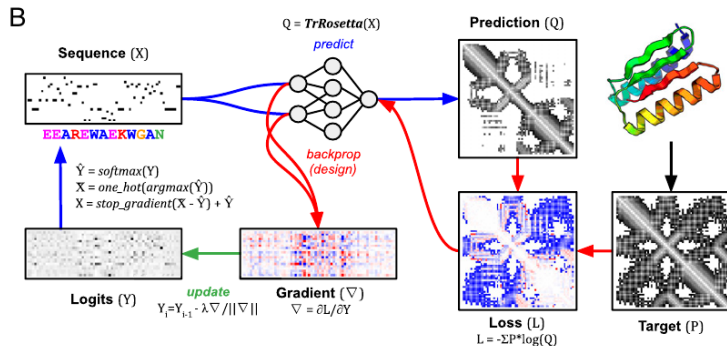


**To design:** use the gradient with respect to the input only to search for sequences that give the correct fold.



(Norm *et al*, 2021, PNAS)

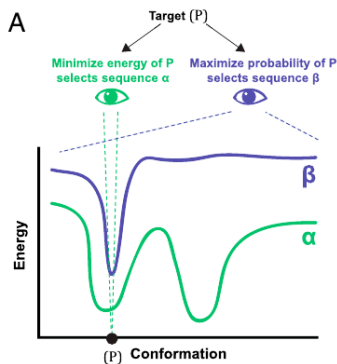
Encode the structure into distance matrices ( $D_{i,j}$  = distance between residues  $i$  and  $j$ ).



(Norm *et al*, 2021, PNAS)

## Why doing so?

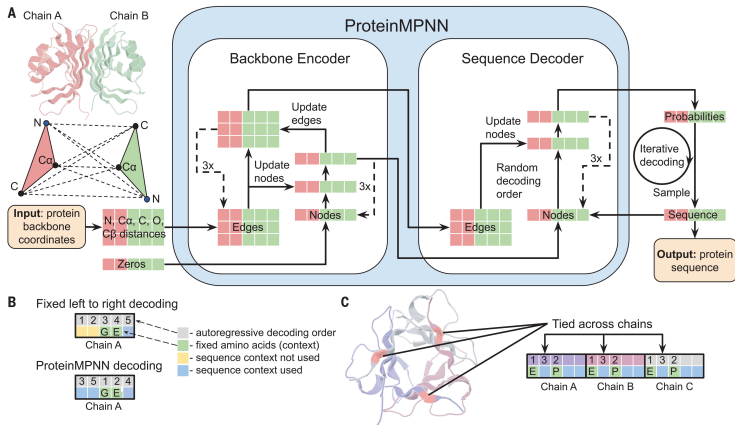
- Positive design: searching for sequences that fold into the target structure.
- Negative design: searching for sequences that fold **only** into the target structure.



(Norm *et al*, 2021, PNAS)

**Represent protein structures using graph  
neural networks** Yes, we can!

# Design proteins conditioned on the structure



[Dauparas *et al*, Science, 2022]



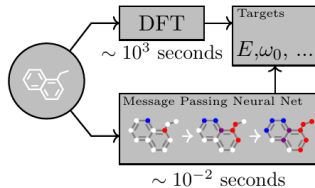
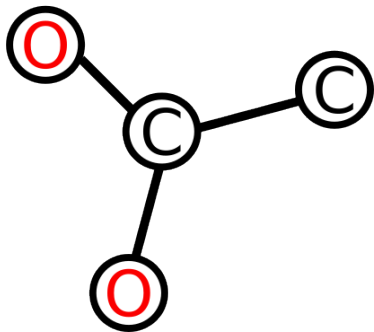


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]

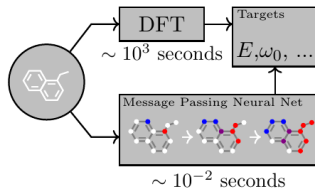
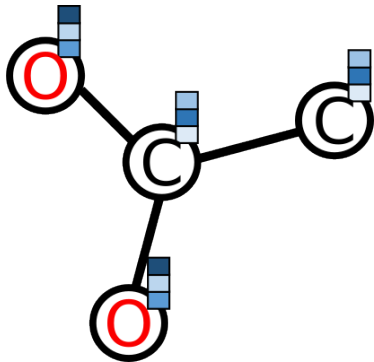


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]

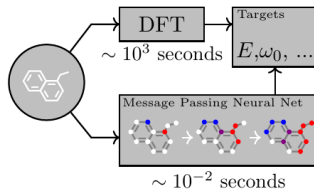
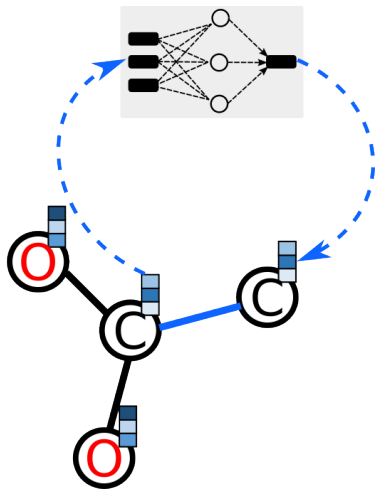


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]

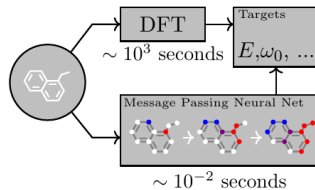
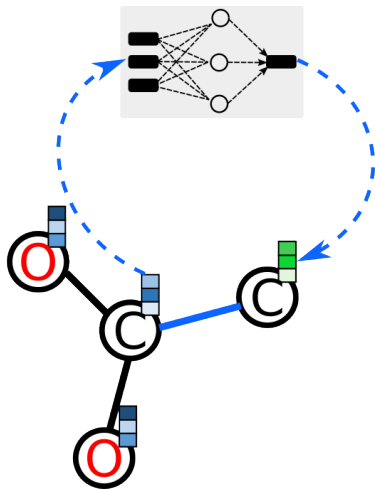


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]

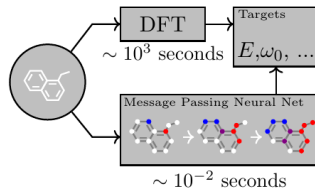
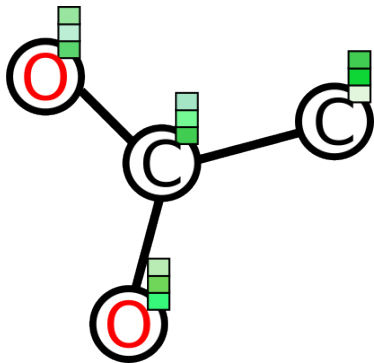


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]

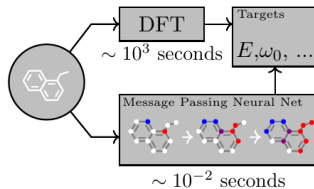
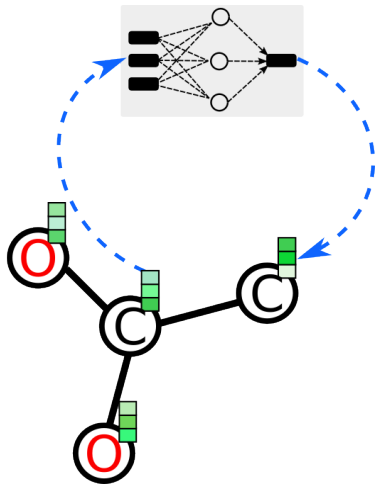
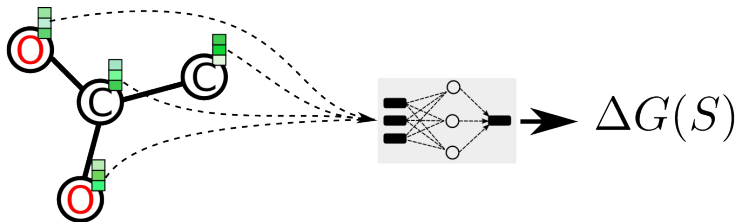
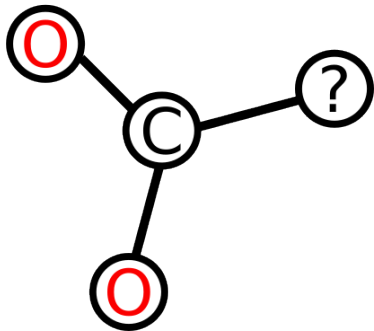


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]



Predict a molecular property using the contextualized node vectors, for example, the binding free energy to a protein target.



Mask the identity of one atom...

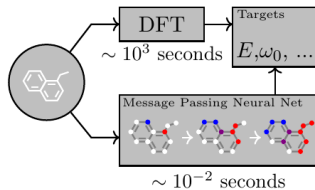
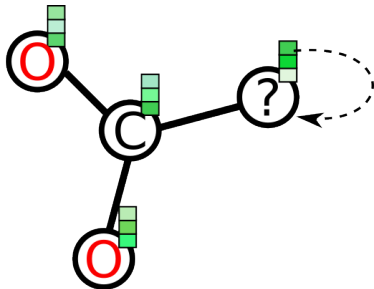


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]





... predict back the identity of the masked atom.

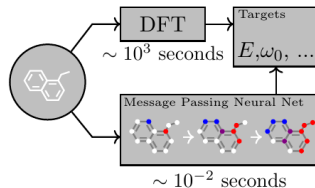


Figure 1. A Message Passing Neural Network predicts quantum properties of an organic molecule by modeling a computationally expensive DFT calculation.

[Gilmer *et al*, ICML, 2017]