Ab Initio Protein Structure Prediction: AlphaFold

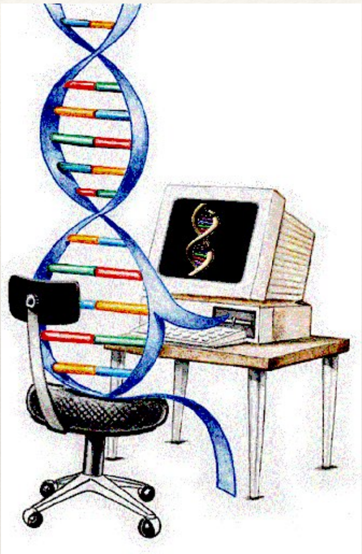
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2



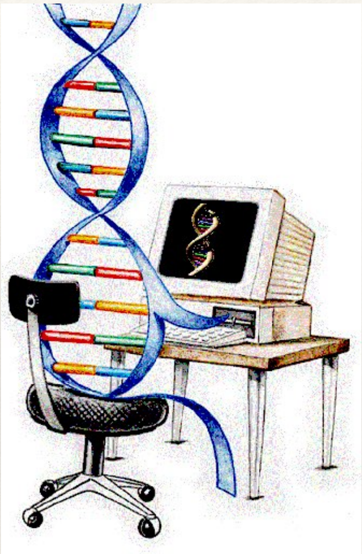
Ab initio Protein Structure Prediction

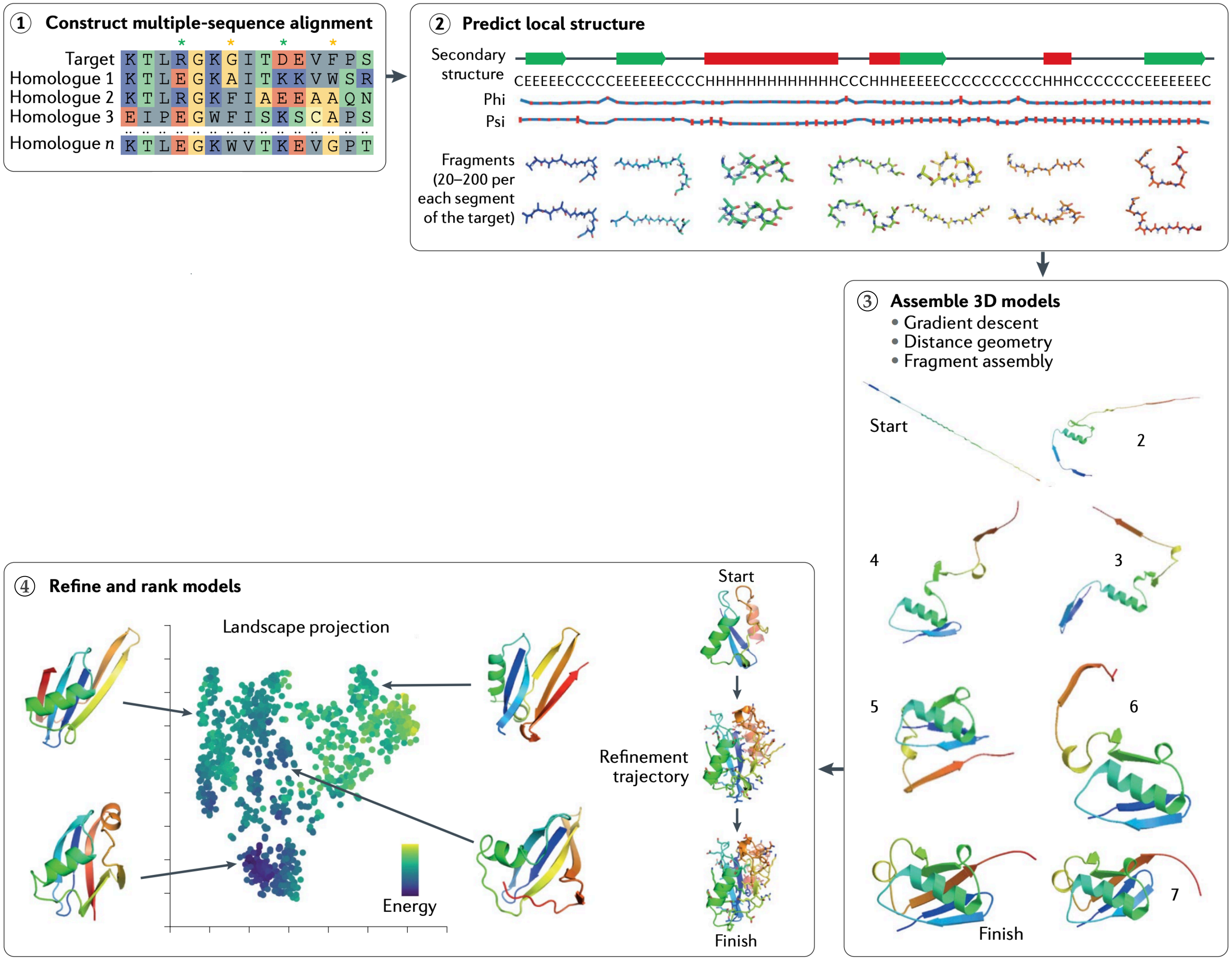
Ab initio prediction before AlphaFold

Ab initio prediction: Predicting Contacts

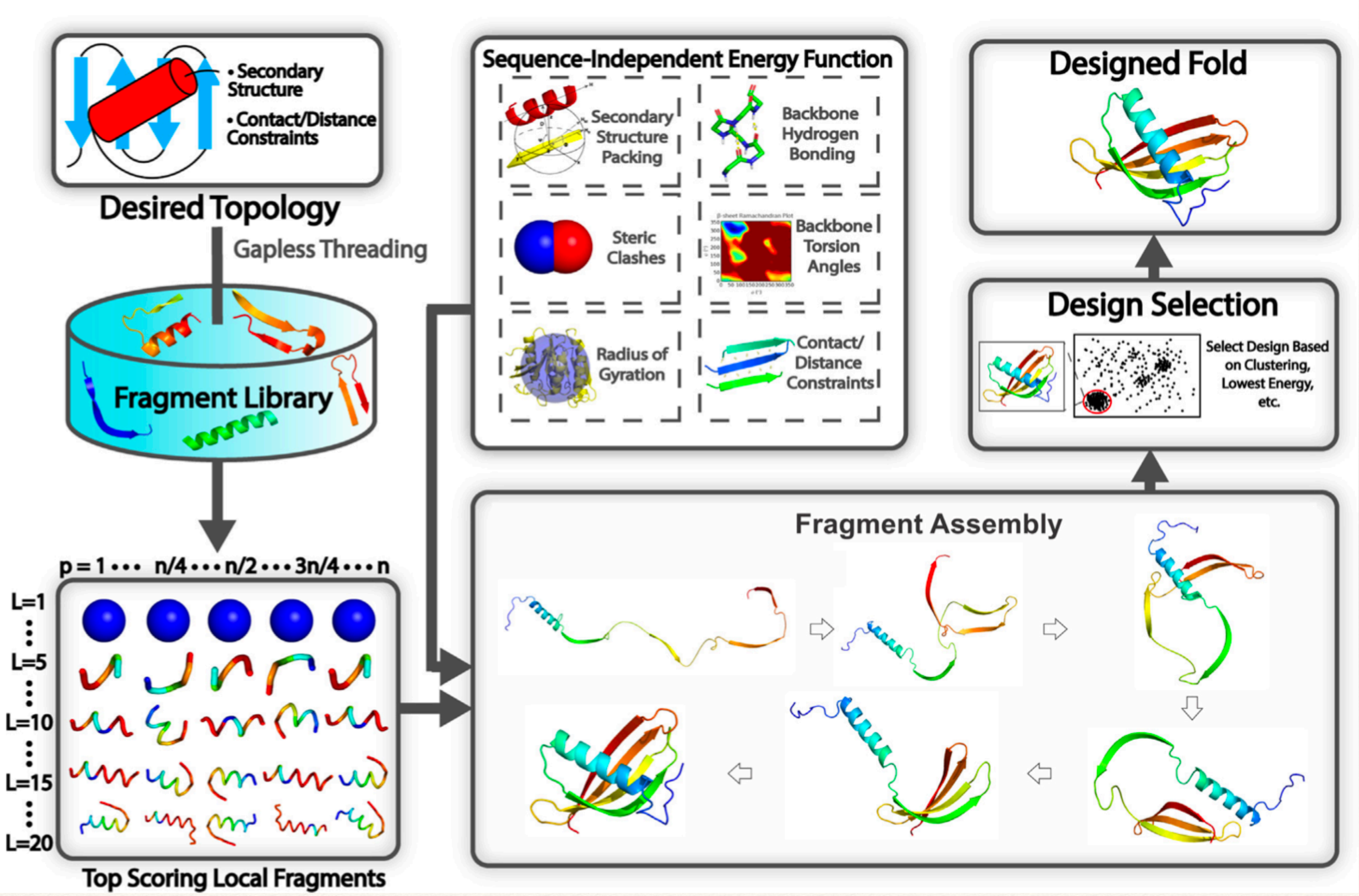
AlphaFold 1

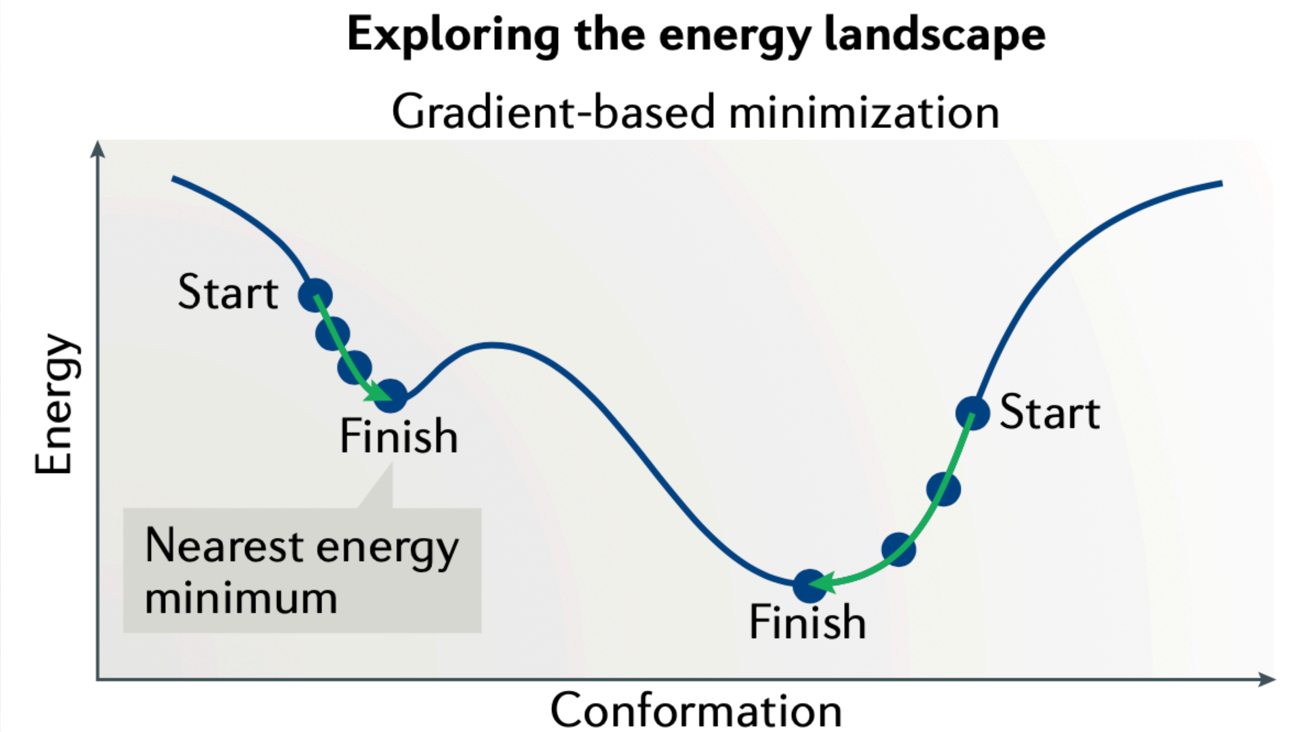
AlphaFold 2




① Construct multiple-sequence alignment
* * * *
Target K T L R G K G I T D E V F P S
Homologue 1 K T L E G K A I T K K V W S R
Homologue 2 K T L R G K F I A E E A A Q N
Homologue 3 E I P E G W F I S K S C A P S
Homologue n K T L E G K W V T K E V G P T
② Predict local structure
Secondary structure
CEEEEECCCCCEEEEEECCCCHHHHHHHHHHHHHHCCHHHEEEEECCCCCCCCCCHHHCCCCCCCCEEEEEEEC
Phi
Psi
Fragments (20–200 per each segment of the target)
③ Assemble 3D models
• Gradient descent
• Distance geometry
• Fragment assembly
Start
2
4
3
5
6
Finish
7
④ Refine and rank models
Landscape projection
Energy
Start
Refinement trajectory
Finish

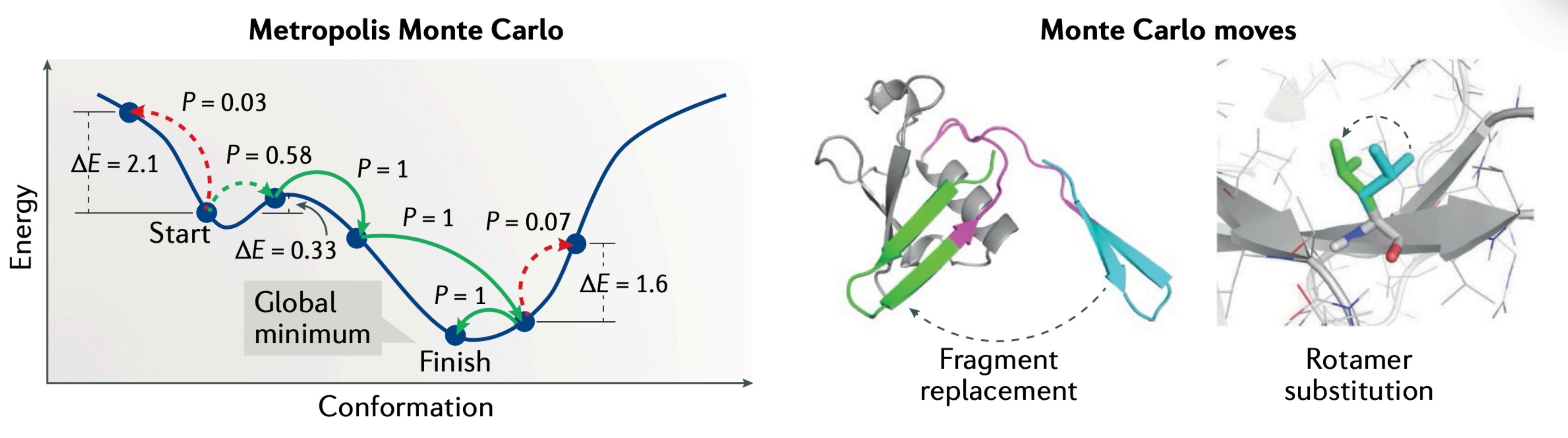
Fragment based methods


• Secondary Structure
• Contact/Distance Constraints
Desired Topology
Gapless Threading
Fragment Library
p = 1 ••• n/4 ••• n/2 ••• 3n/4 ••• n
L=1
L=5
L=10
L=15
L=20
Top Scoring Local Fragments
Sequence-Independent Energy Function
Secondary Structure Packing
Backbone Hydrogen Bonding
Steric Clashes
Backbone Torsion Angles
Radius of Gyration
Contact/ Distance Constraints
Fragment Assembly
Designed Fold
Design Selection
Select Design Based on Clustering, Lowest Energy, etc.



Exploring the energy landscape

Gradient-based minimization


Metropolis Monte Carlo
Energy
Conformation
P = 0.03
ΔE = 2.1
P = 0.58
P = 1
Start
ΔE = 0.33
P = 1
P = 0.07
ΔE = 1.6
Global minimum
P = 1
Finish
Monte Carlo moves
Fragment replacement
Rotamer substitution

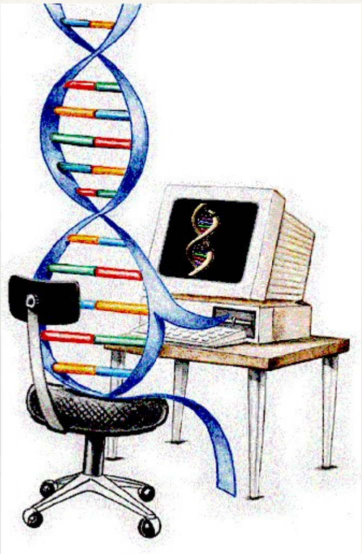
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

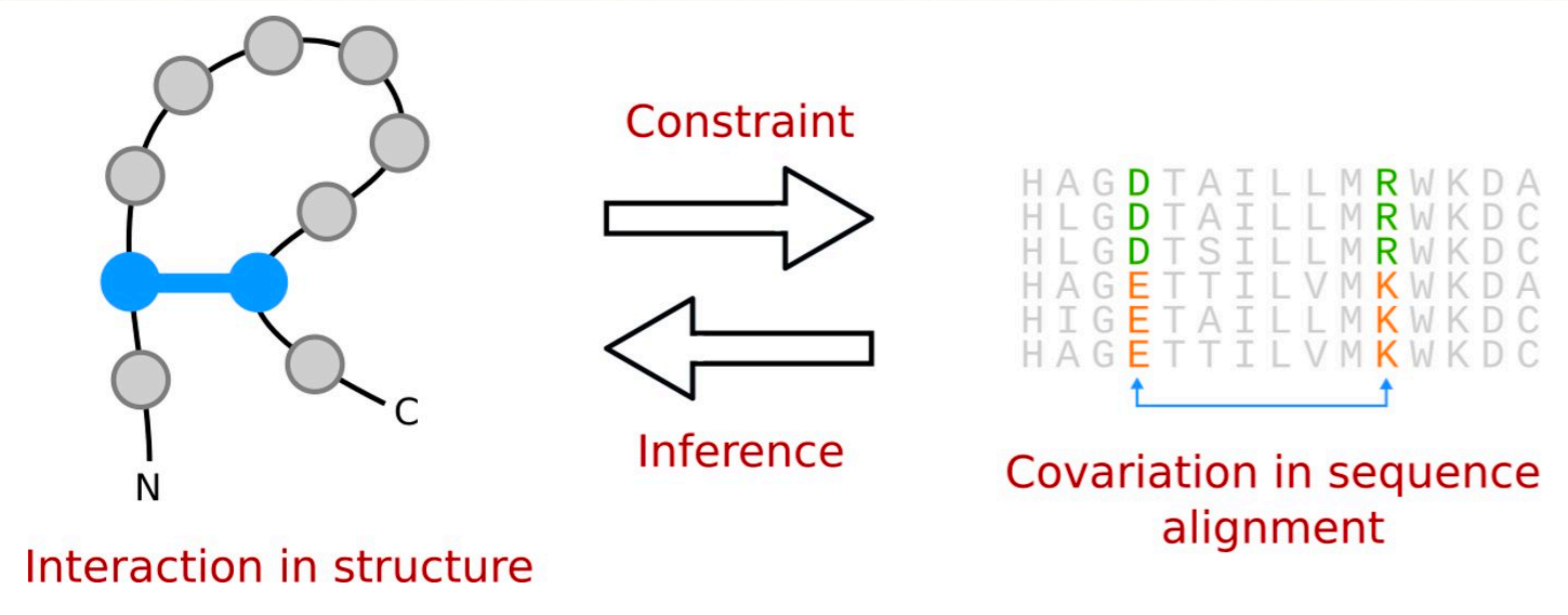
Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2



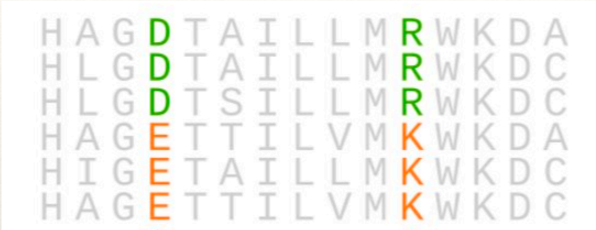
Predicting residue contacts



```
HAGDTAILLMRWKDA
HLGDTAILLMRWKDC
HLGDTSILLMRWKDC
HAGETTILVMKWKDA
HIGETAILLMKWKDC
HAGETTILVMKWKDC
```

Predicting residue contacts

1. Given a multiple sequence alignment (MSA):



X_1

```
HAGDTAILLMRWKDA
HLGDTAILLMRWKDC
HLGDTSILLMRWKDC
HAGETTILVMKWKDA
HIGETAILLMKWKDC
HAGETTILVMKWKDC
```

X_N

2. Compute "mean" sequence and covariance matrix:

$$\overline{\mathbf{X}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{X}_n$$

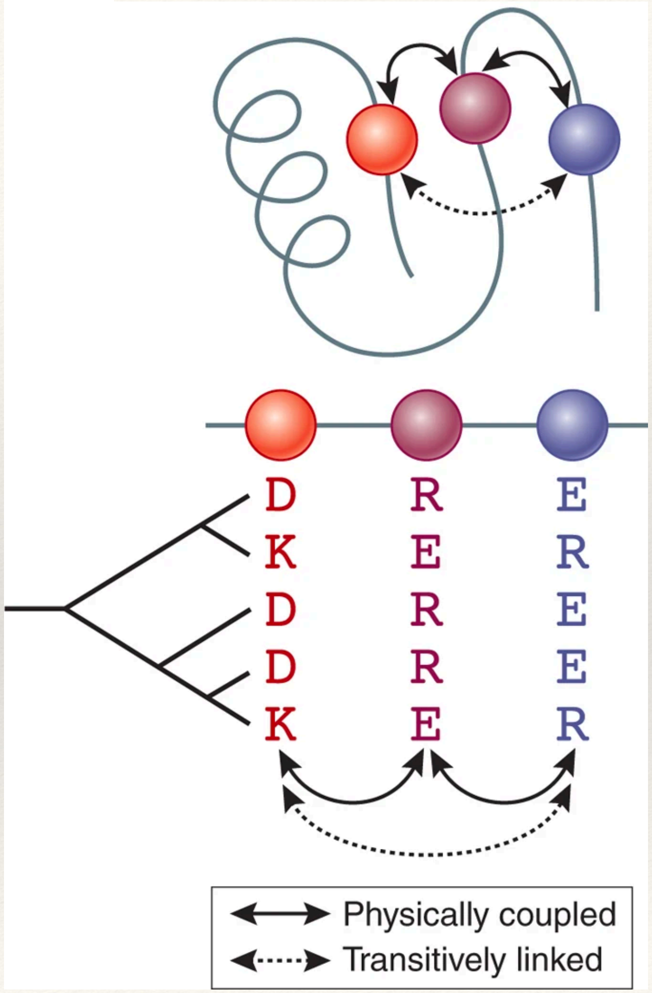
$$\overline{C} = C(MSA, \overline{\mathbf{X}}) = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{X}_n - \overline{\mathbf{X}})^T(\mathbf{X}_n - \overline{\mathbf{X}})$$

3. Compute contact J(i,j)

$$J(i,j) = C(i,j)?$$

Predicting residue contacts

No! We need to pay attention to indirect effects:



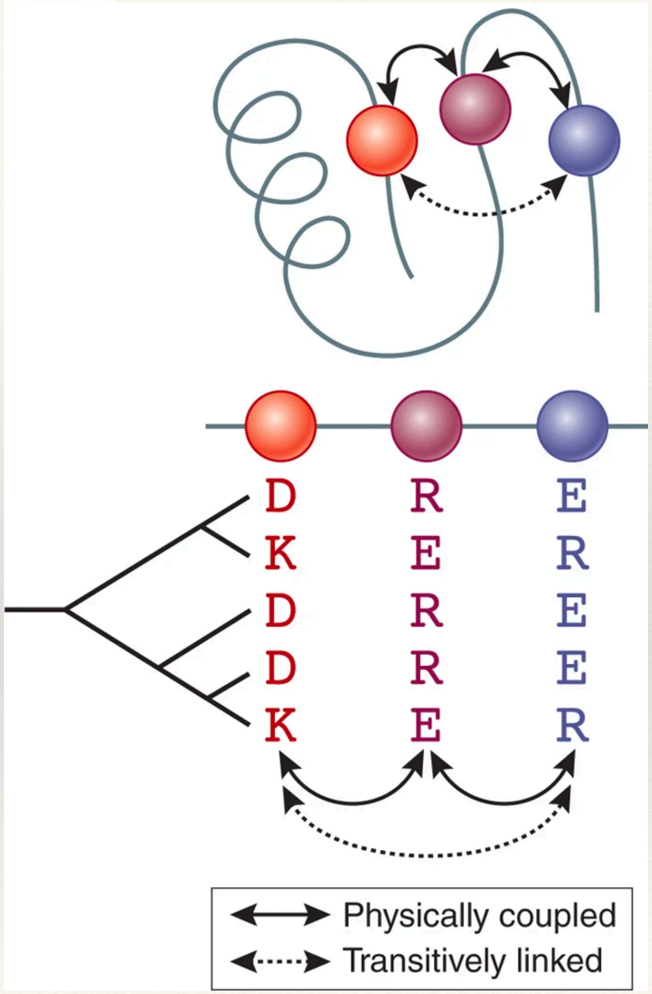
Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

Each sequence X_i in the MSA is drawn from a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, with the probability:

$$P(\mathbf{X}_n|\boldsymbol{\mu},\Sigma) = (2\pi)^{-\frac{Ls}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{X}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}_n-\boldsymbol{\mu})\right]$$



Predicting residue contacts

No! We need to pay attention to indirect effects:

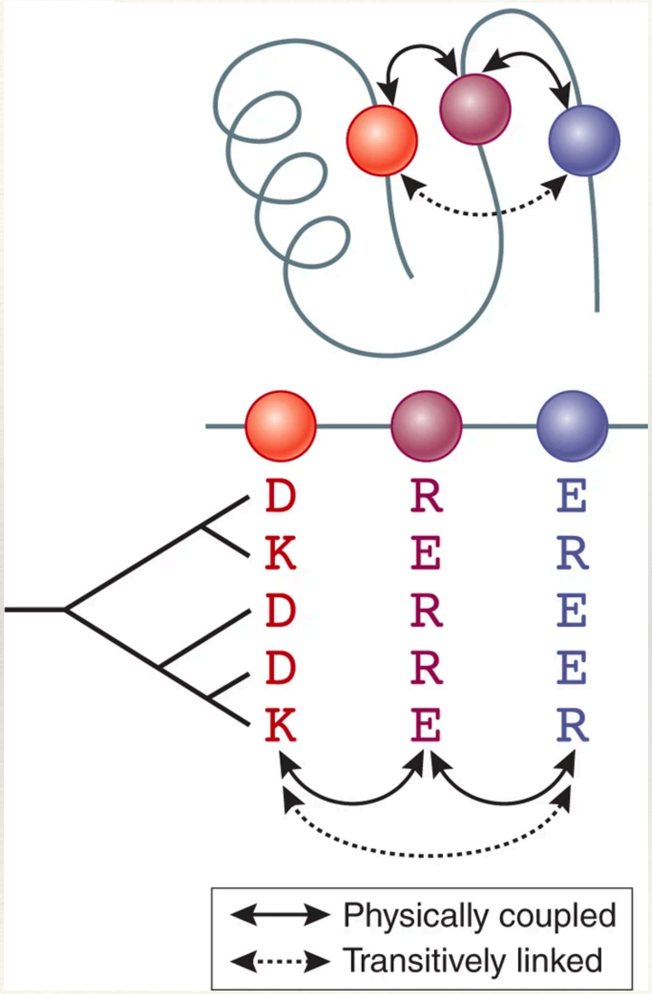
Gaussian model:

Each sequence X_i in the MSA is drawn from a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, with the probability:

$$P(\mathbf{X}_n \,|\, \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{Ls}{2}} \,|\,\Sigma\,|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{X}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}_n - \boldsymbol{\mu}) \right]$$

Assuming that the N sequences in the MSA are statistical independent, the probability, or likelihood of the data under this model is given by

$$P(MSA \,|\, \boldsymbol{\mu}, \Sigma) = \prod_{n=1}^{N} P(\mathbf{X}_n \,|\, \boldsymbol{\mu}, \Sigma)$$



Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

Each sequence X_i in the MSA is drawn from a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ, with the probability:

$$P(\mathbf{X}_n | \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{Ls}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{X}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}_n - \boldsymbol{\mu})\right]$$

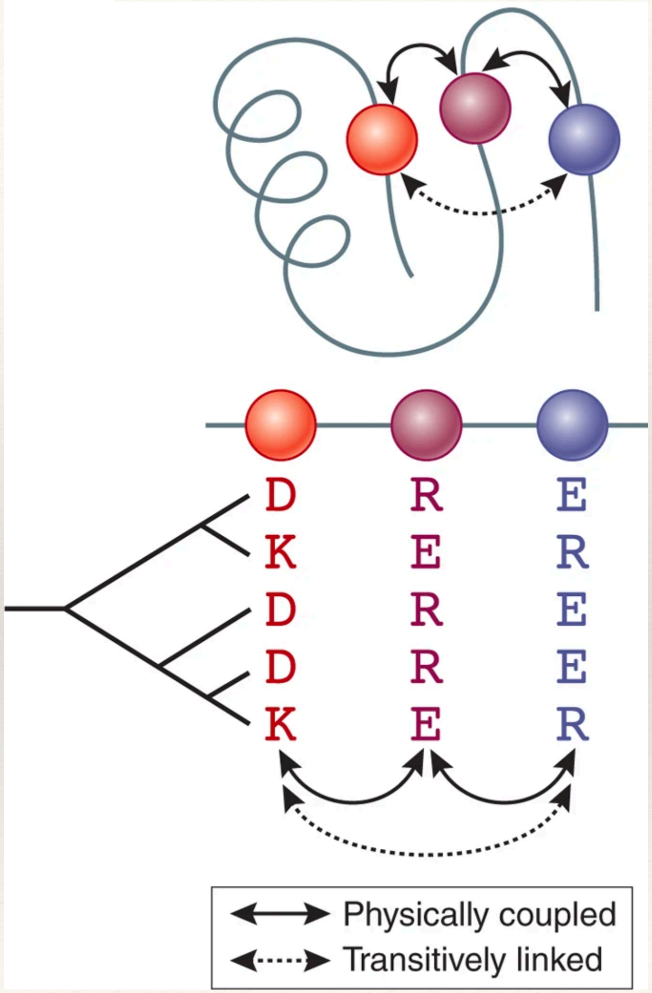
Assuming that the N sequences in the MSA are statistical independent, the probability, or likelihood of the data under this model is given by

$$P(MSA | \boldsymbol{\mu}, \Sigma) = \prod_{n=1}^{N} P(\mathbf{X}_n | \boldsymbol{\mu}, \Sigma)$$

Using the maximum likelihood estimator for this probability

$$\boldsymbol{\mu} = \overline{\mathbf{X}}$$

$$\Sigma = \overline{C} = C(MSA, \overline{\mathbf{X}})$$



Predicting residue contacts

No! We need to pay attention to indirect effects:

Gaussian model:

$$P(\mathbf{X}_n | \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{Ls}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{X}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}_n - \boldsymbol{\mu})\right]$$

$$\boldsymbol{\mu} = \overline{\mathbf{X}} \qquad \Sigma = \overline{C} = C(MSA, \overline{\mathbf{X}})$$

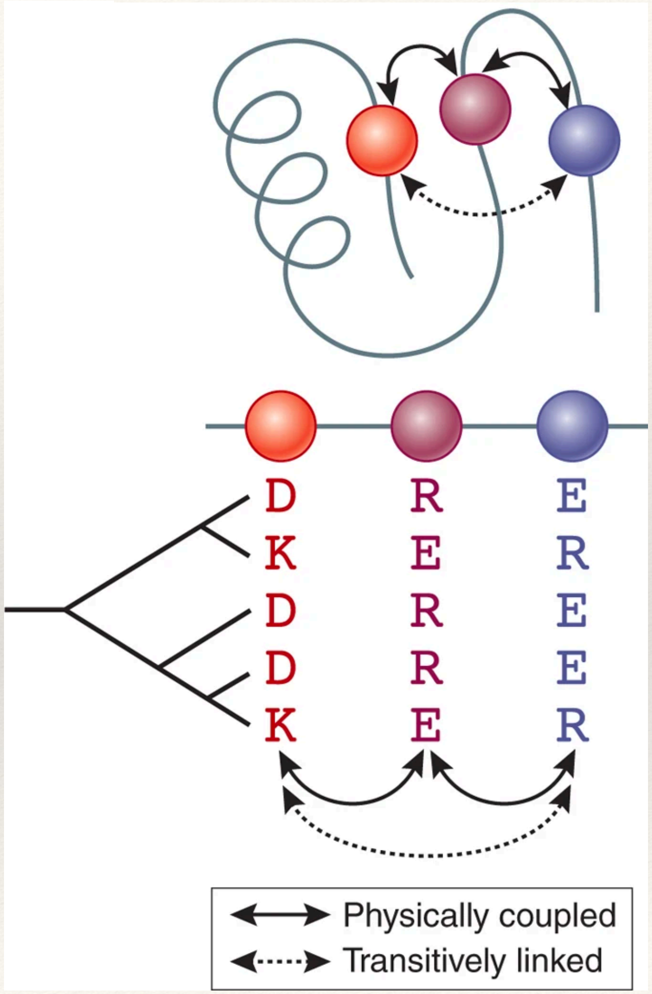
Note that:

$$(\mathbf{X}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}_n - \boldsymbol{\mu}) = \sum_{k=1}^{N} \sum_{l=1}^{N} (X_k - \mu_k)(\Sigma^{-1})(k,l)(X_l - \mu_l)$$

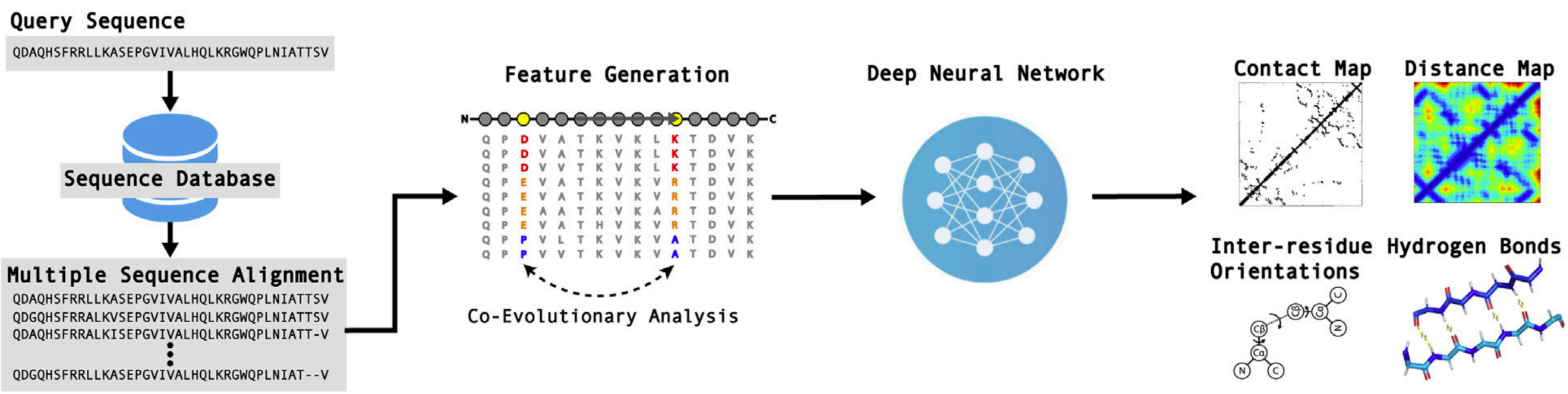
This shows that $(\Sigma^{-1})(k,l)$ serves as a coupling between positions k and l in the MSA.

Therefore:

$$J = \Sigma = (C(MSA, \overline{X}))^{-1}$$



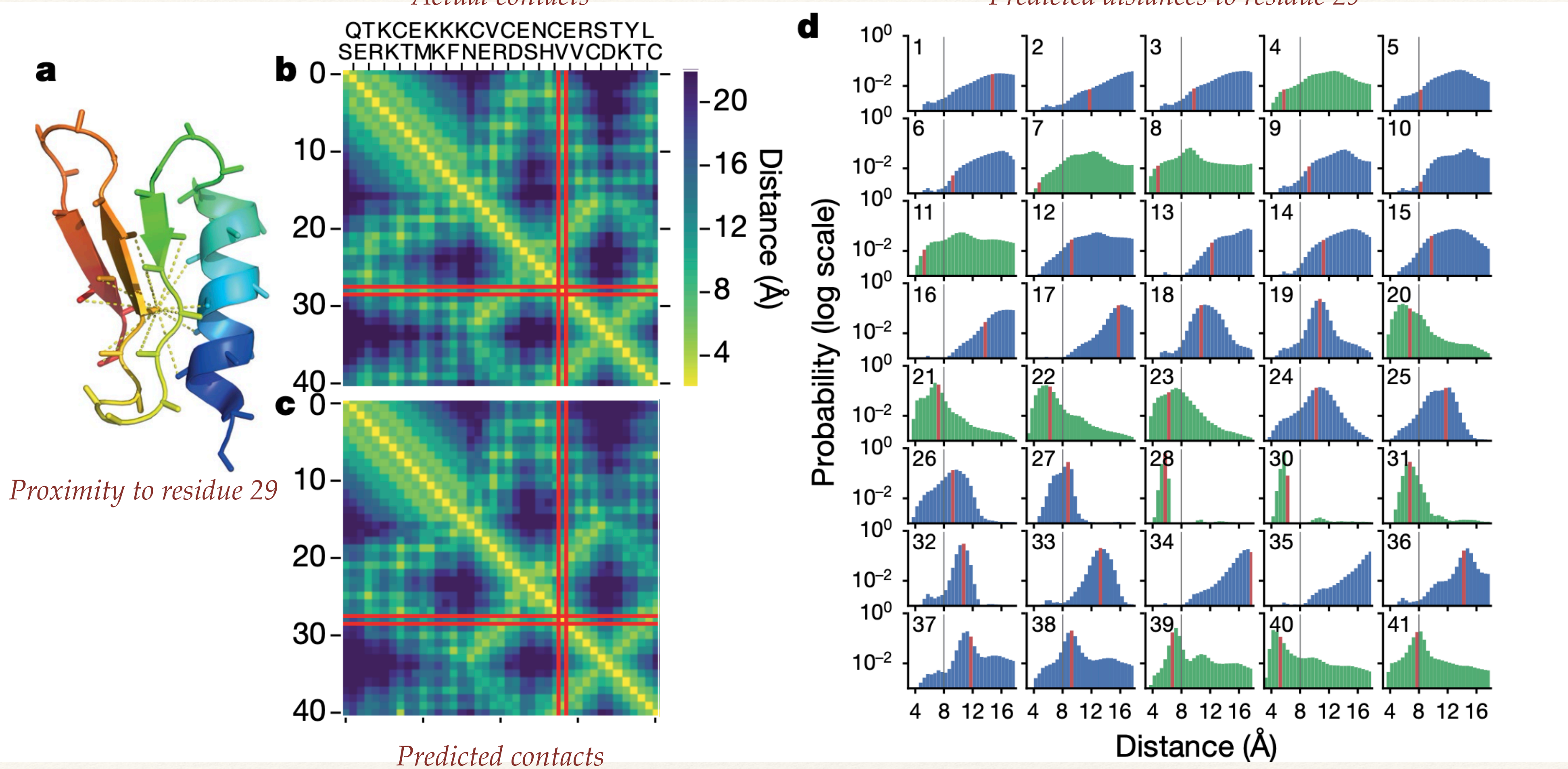
Predicting residue contacts


Query Sequence
QDAQHSFRRLLKASEPGVIVALHQLKRGWQPLNIATTSV
Sequence Database
Multiple Sequence Alignment
QDAQHSFRRLLKASEPGVIVALHQLKRGWQPLNIATTSV
QDGQHSFRRALKVSEPGVIVALHQLKRGWQPLNIATTSV
QDAQHSFRRALKISEPGVIVALHQLKRGWQPLNIATT-V
⋮
QDGQHSFRRLLKASEPGVIVALHQLKRGWQPLNIAT--V
Feature Generation
N
C
Q P D V A T K V K L K T D V K
Q P D V A T K V K L K T D V K
Q P D V V T K V K L K T D V K
Q P E V A T K V K V R T D V K
Q P E V A T K V K V R T D V K
Q P E A A T K V K A R T D V K
Q P E V A T H V K V R T D V K
Q P P V L T K V K V A T D V K
Q P P V V T K V K V A T D V K
Co-Evolutionary Analysis
Deep Neural Network
Contact Map
Distance Map
Inter-residue Orientations
Cβ
Cα
N
C
Hydrogen Bonds

Predicting residue contacts: How well does it work?

Actual contacts

Predicted distances to residue 29



Proximity to residue 29

Predicted contacts

CASP target T0995

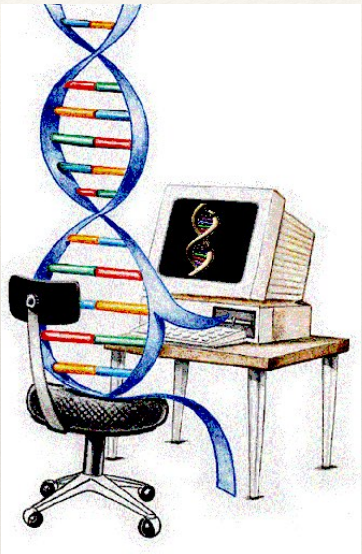
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

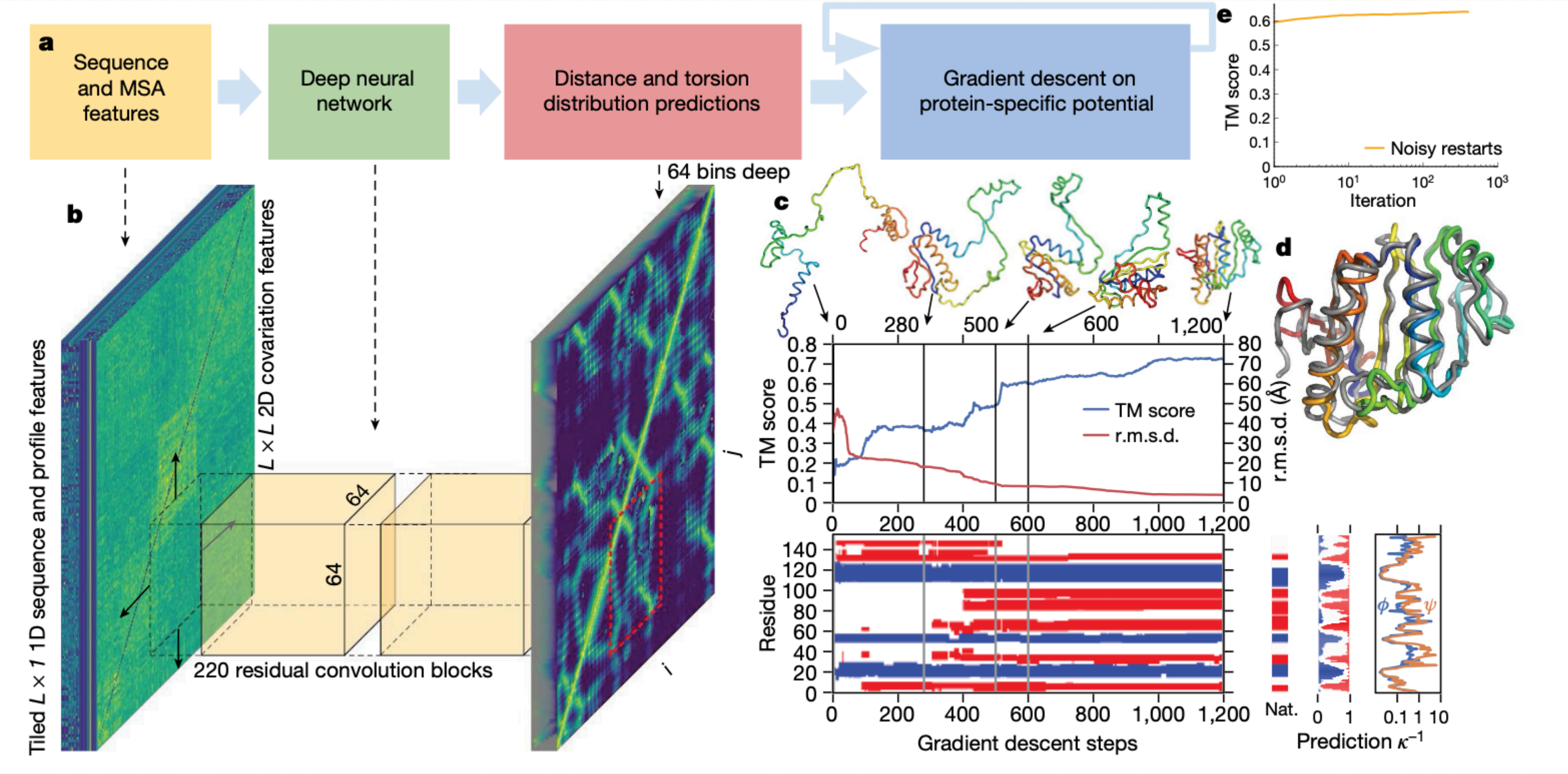
Ab initio prediction: Predicting Contacts

AlphaFold 1

AlphaFold 2



AlphaFold 1


a
Sequence and MSA features
Deep neural network
Distance and torsion distribution predictions
Gradient descent on protein-specific potential
b
Tiled L × 1 1D sequence and profile features
L × L 2D covariation features
64
64
220 residual convolution blocks
64 bins deep
j
i
c
0
280
500
600
1,200
TM score
r.m.s.d. (Å)
TM score
r.m.s.d.
Residue
Gradient descent steps
0
200
400
600
800
1,000
1,200
d
Nat.
Prediction κ⁻¹
0
1
0.1
1
10
ϕ
ψ
e
TM score
Noisy restarts
Iteration
10⁰
10¹
10²
10³

AlphaFold1

AlphaFold 1

Reminder:

To compare two sets of points (atoms) A={a_1, a_2, …a_N} and B={b_1, b_2, …,b_N}:

-Define a 1-to-1 correspondence between A and B

for example, a_i corresponds to b_i, for all i in [1,N]

-Compute RMS as:

$$RMS(A,B)=\sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i,b_i)^2}$$

Compute TM score:

$$TM(A,B)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\left(\frac{d(a_i,b_i)}{d_0(N)}\right)^2}$$

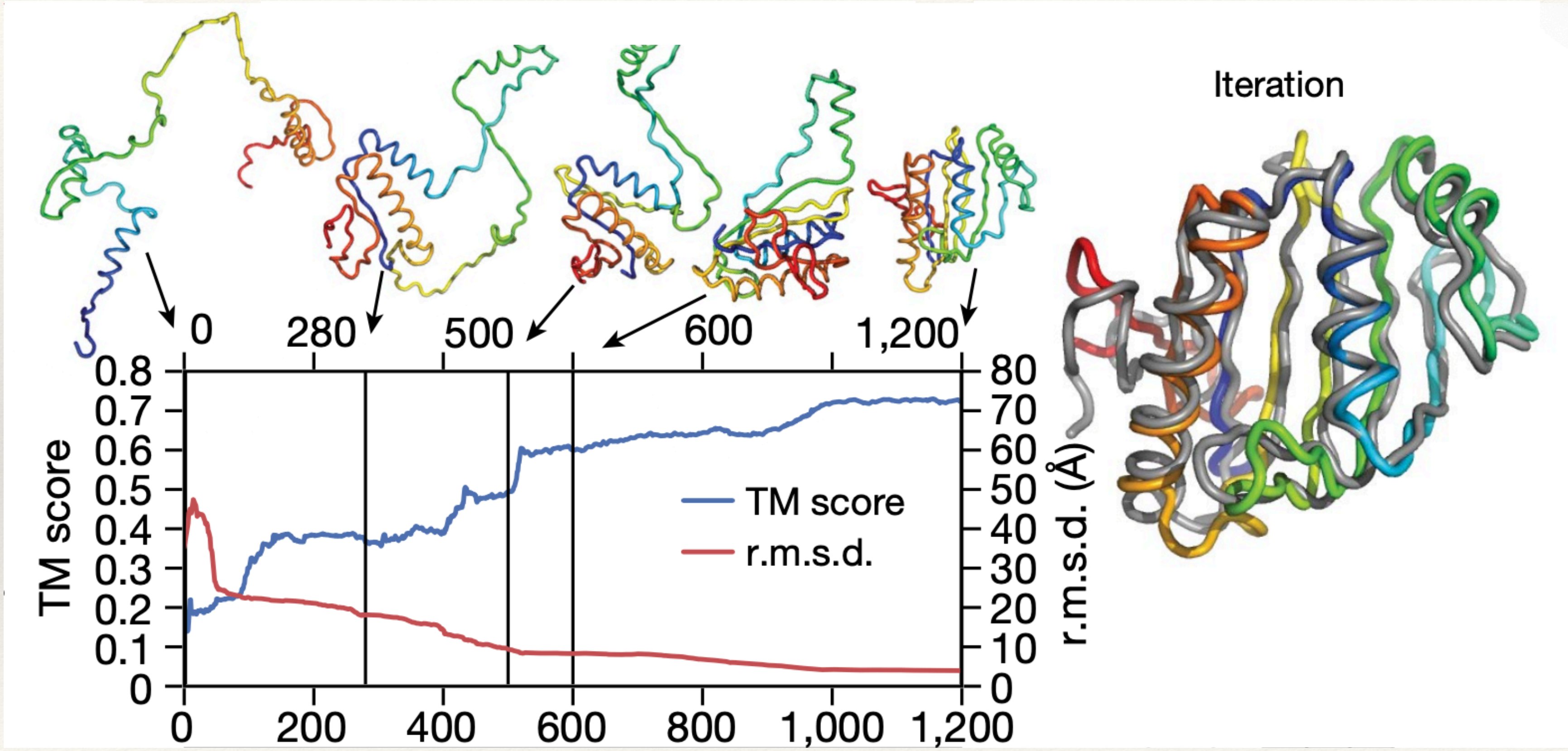
with $d_0(N) = 1.24\sqrt[3]{N-15} - 1.8$

$d(a_i,b_i)$ is the Euclidian distance between a_i and b_i after optimal alignment of B onto A

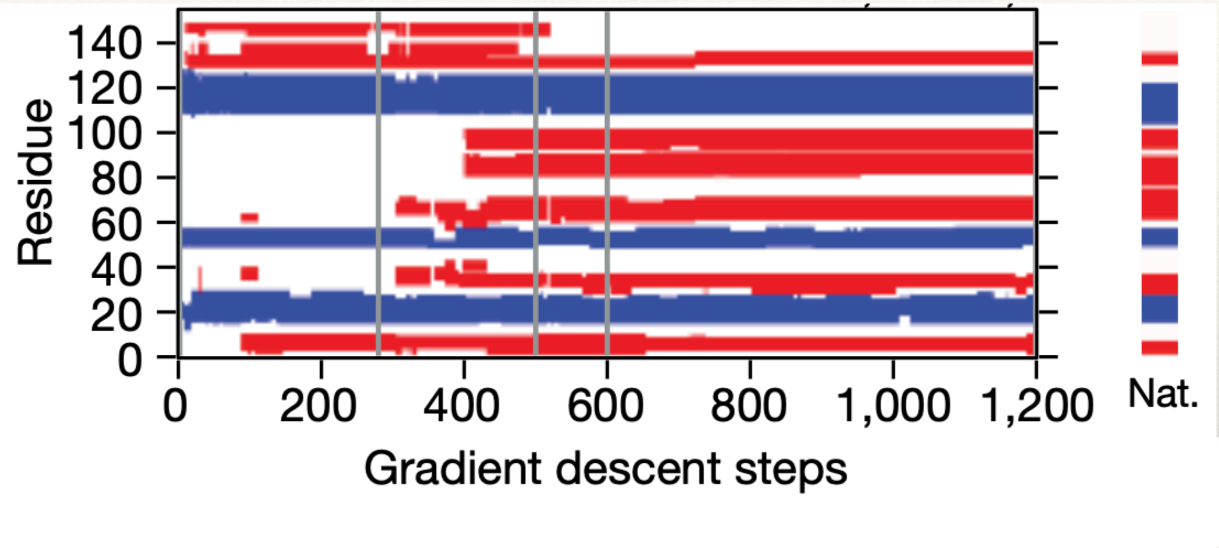
RMS: the lower, the better

TM: between [0,1]; the higher the better

AlphaFold 1

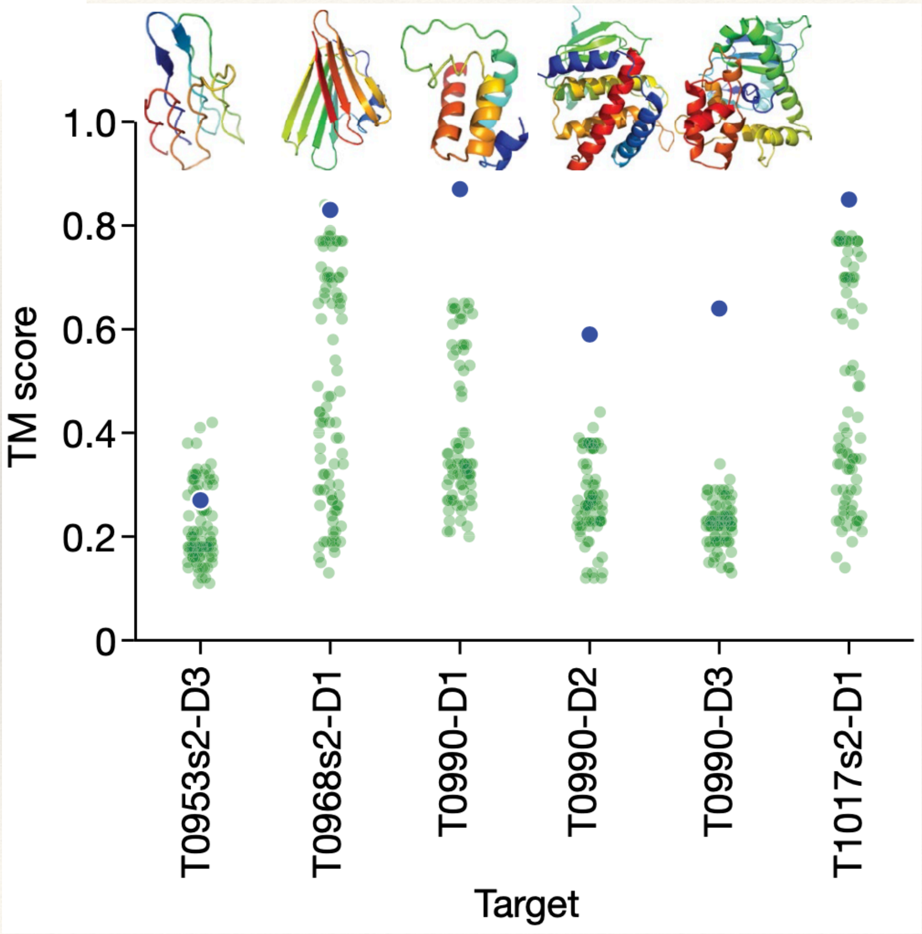

Iteration
0
280
500
600
1,200
TM score
r.m.s.d. (Å)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
80
70
60
50
40
30
20
10
0
0
200
400
600
800
1,000
1,200
TM score
r.m.s.d.

AlphaFold 1



Helix in blue, strand in red

AlphaFold 1: Success



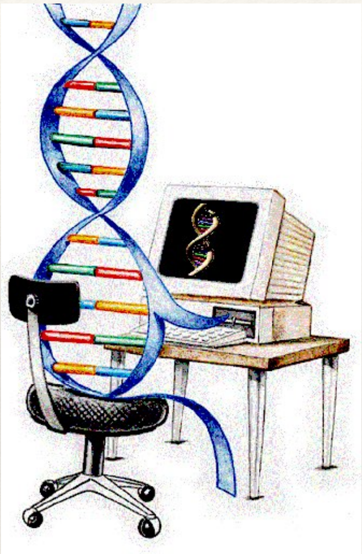
Ab initio Protein Structure Prediction

Ab initio prediction before AlphaFold

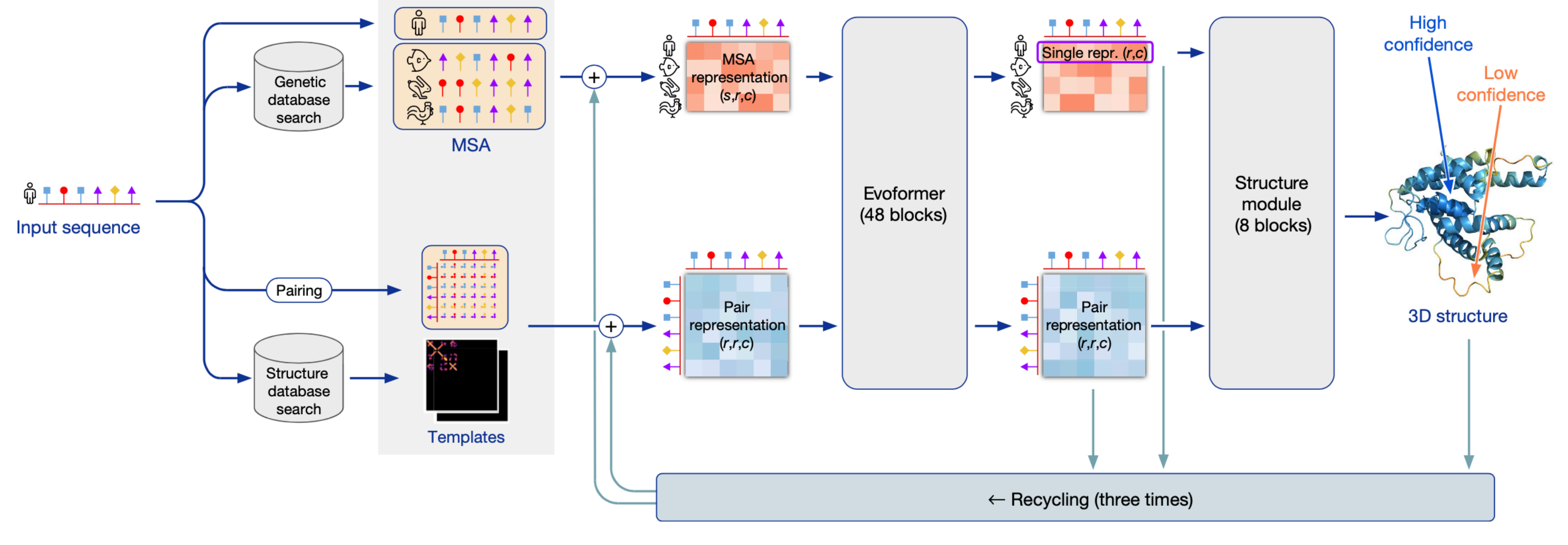
Ab initio prediction: Predicting Contacts

AlphaFold 1

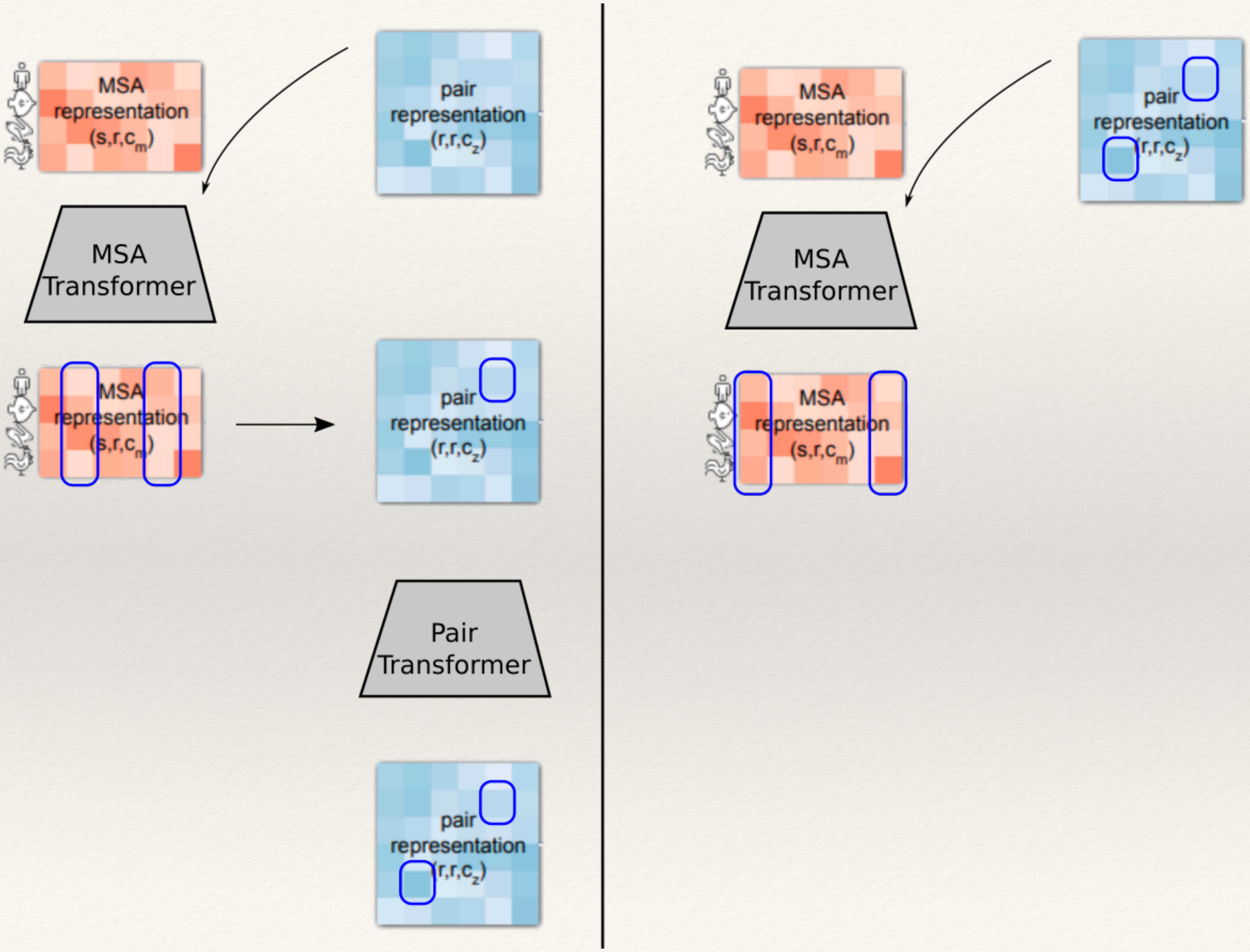
AlphaFold 2



AlphaFold 2

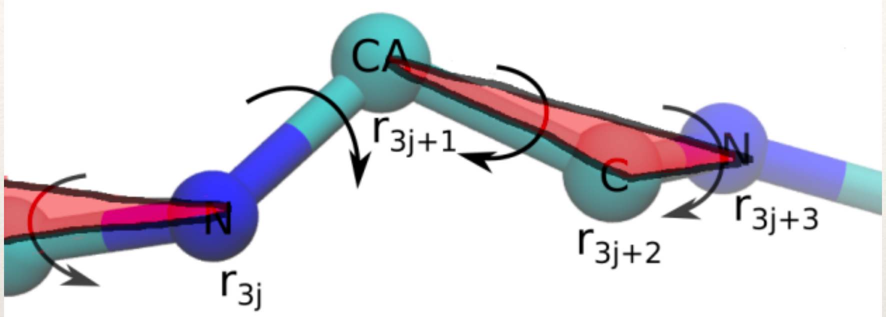

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three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

AlphaFold 2: some intuition


MSA representation (s,r,c$_m$)
pair representation (r,r,c$_z$)
MSA Transformer
MSA representation (s,r,c$_m$)
pair representation (r,r,c$_z$)
Pair Transformer
pair representation (r,r,c$_z$)
MSA representation (s,r,c$_m$)
pair representation (r,r,c$_z$)
MSA Transformer
MSA representation (s,r,c$_m$)

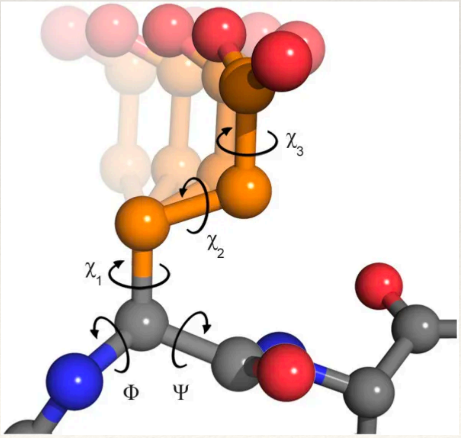
AlphaFold 2: the structure module

Predicting backbone:
the residues form a gas soup of triangles whose relative positions are characterized by affine transformation

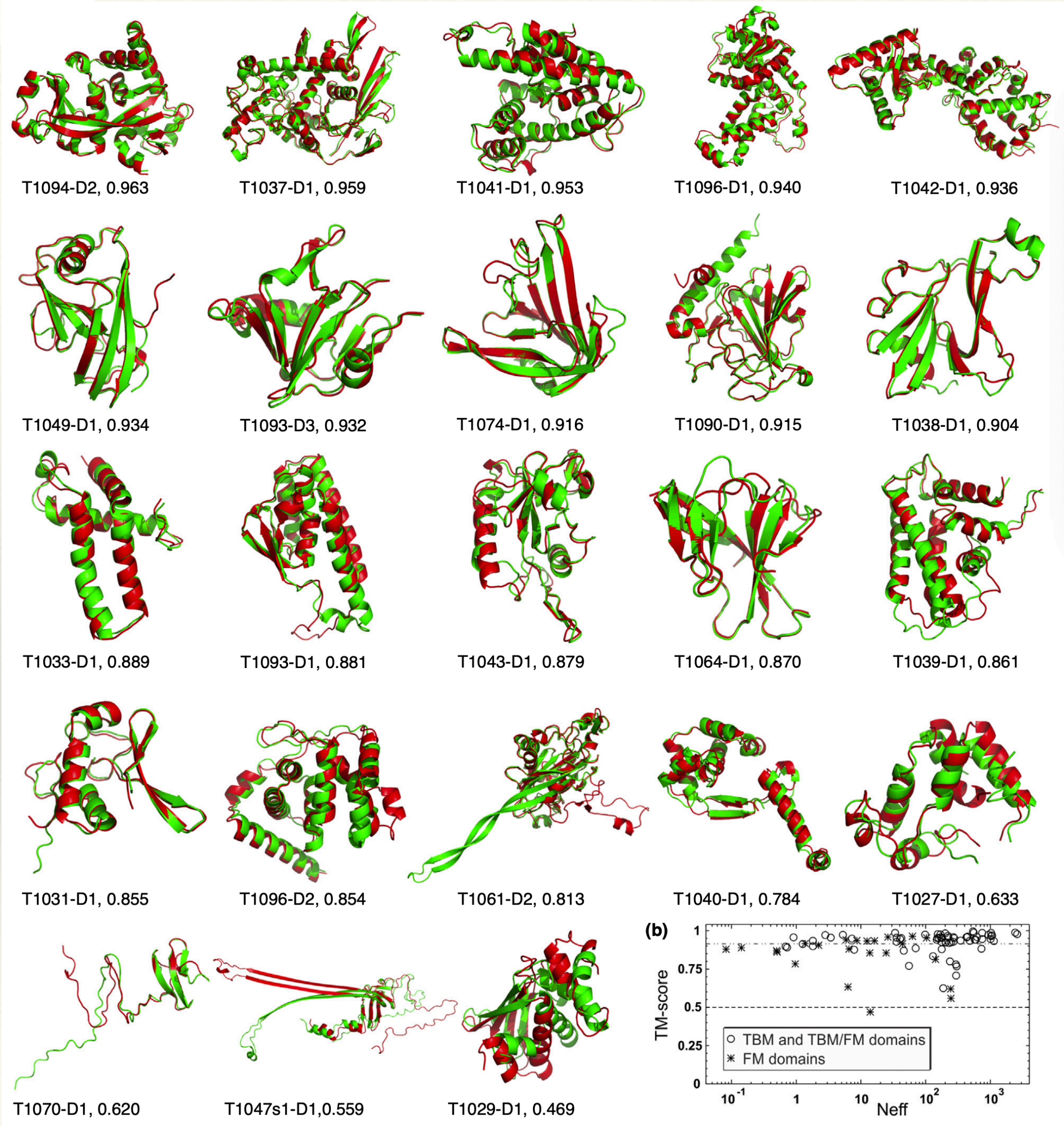


$$\mathbf{M} = \begin{pmatrix} \boxed{\begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix}} & \boxed{\begin{matrix} a_{14} \\ a_{24} \\ a_{34} \end{matrix}} \\ \begin{matrix} 0 & \quad 0 & \quad 0 \end{matrix} & 1 \end{pmatrix}$$

Predicting side chains:

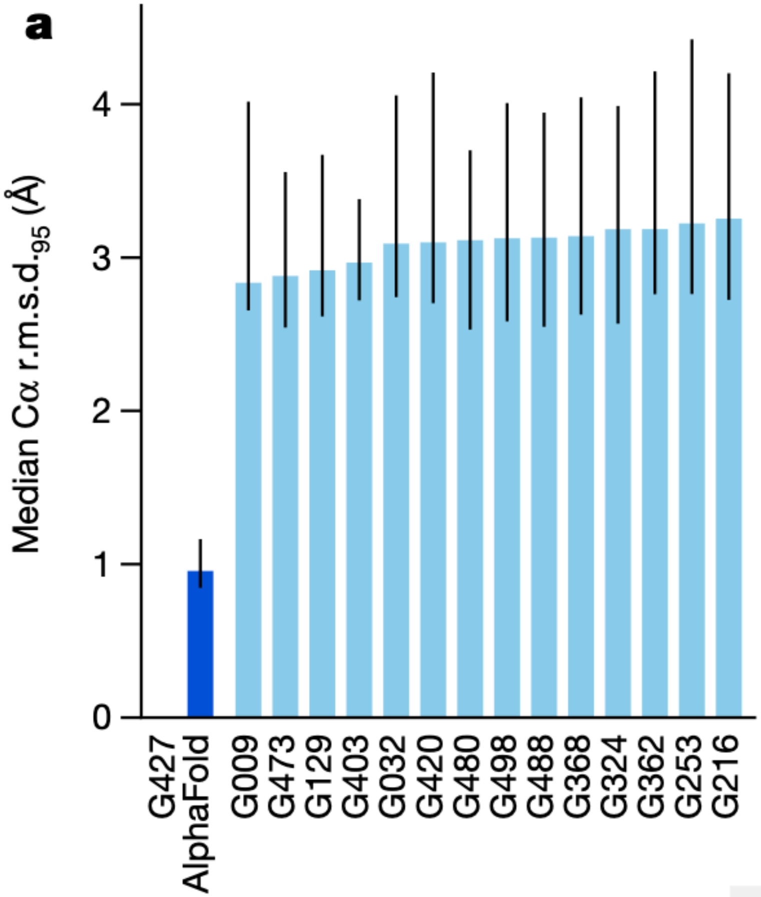


Successes at CASP14



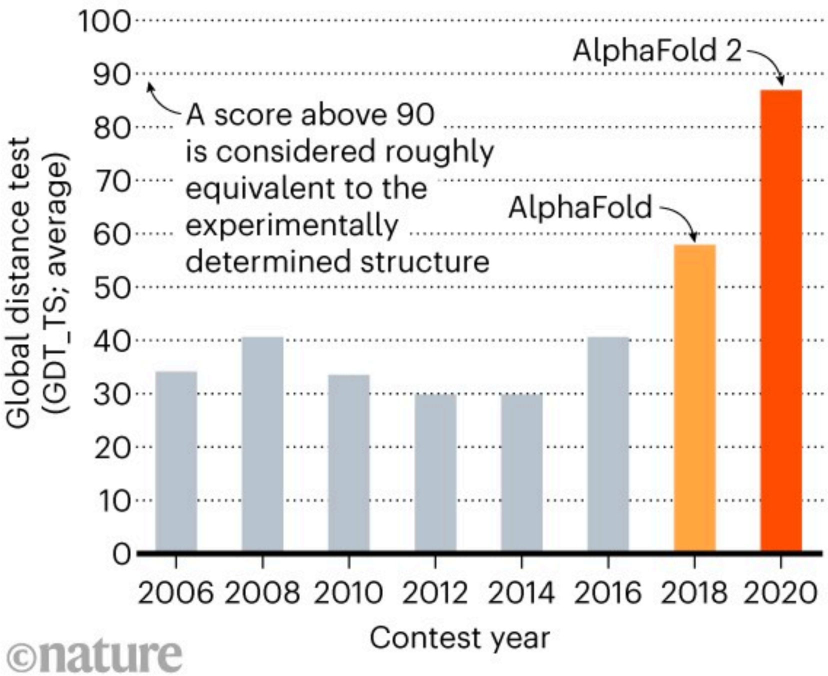
TBM: template-based modeling
FM: free modeling

Successes at CASP14



STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.

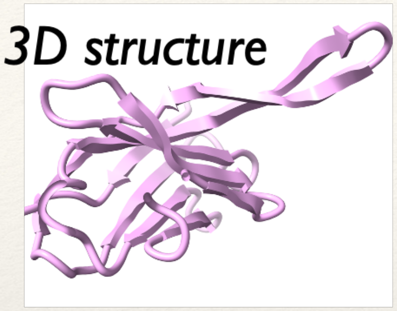


©nature

Training

- *Sequence*
- *Multiple sequence alignment*
- *3D structure*

}



→ **21 million parameters**

Prediction

- *Sequence*
- *Multiple sequence alignment*

```
EVQLVESGGGLVQPGGSLRLSCAASGFNIYSSSIHWVRQAPGKGLEWVAYI
............................F....M........Q........
..................K.........Y....L.......A.........
..............A.............................V......
....................A..............................
............................L....V....E............
.........A...................................Q.....
```

21 million parameters

Focus attention on important relationships

→

3D prediction

Confidence estimates

Credit: Tom Terwilliger, Los Alamos NL

Multiple sequence alignment

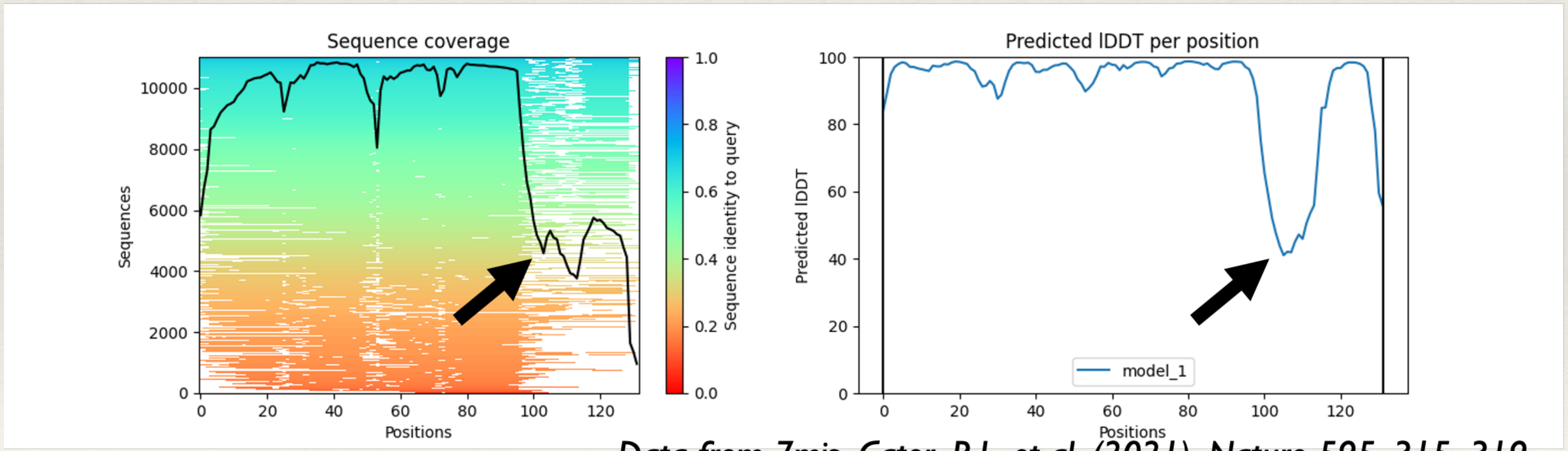
```
EVQLVESGGGLVQPGGSLRLSCAASGFNIYSSSIHWVRQAPGKGLEWVAYI
............................F....M........Q........
..................K.........Y....L.......A.........
..............A.............................V......
....................A..............................
............................L....V....E............
.........A...................................Q.....
```

→

Residues that ***co-vary*** *are probably close in 3D structure*

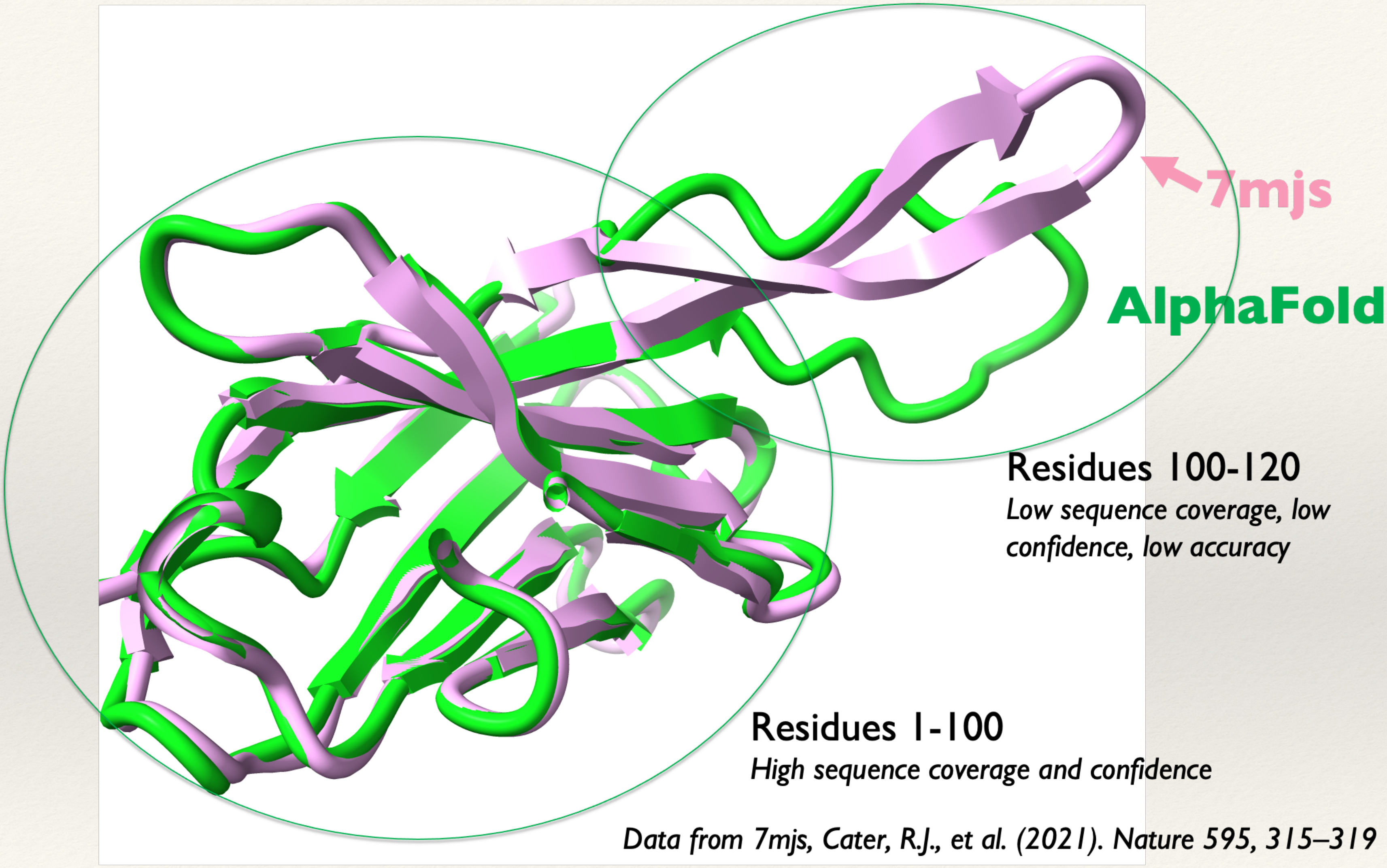
All sequences in alignment should be compatible with the right structure

Sequence coverage → Confidence



Data from 7mjs, Cater, R.J., et al. (2021). Nature 595, 315–319

Credit: Tom Terwilliger, Los Alamos NL



Residues 100-120

Low sequence coverage, low confidence, low accuracy

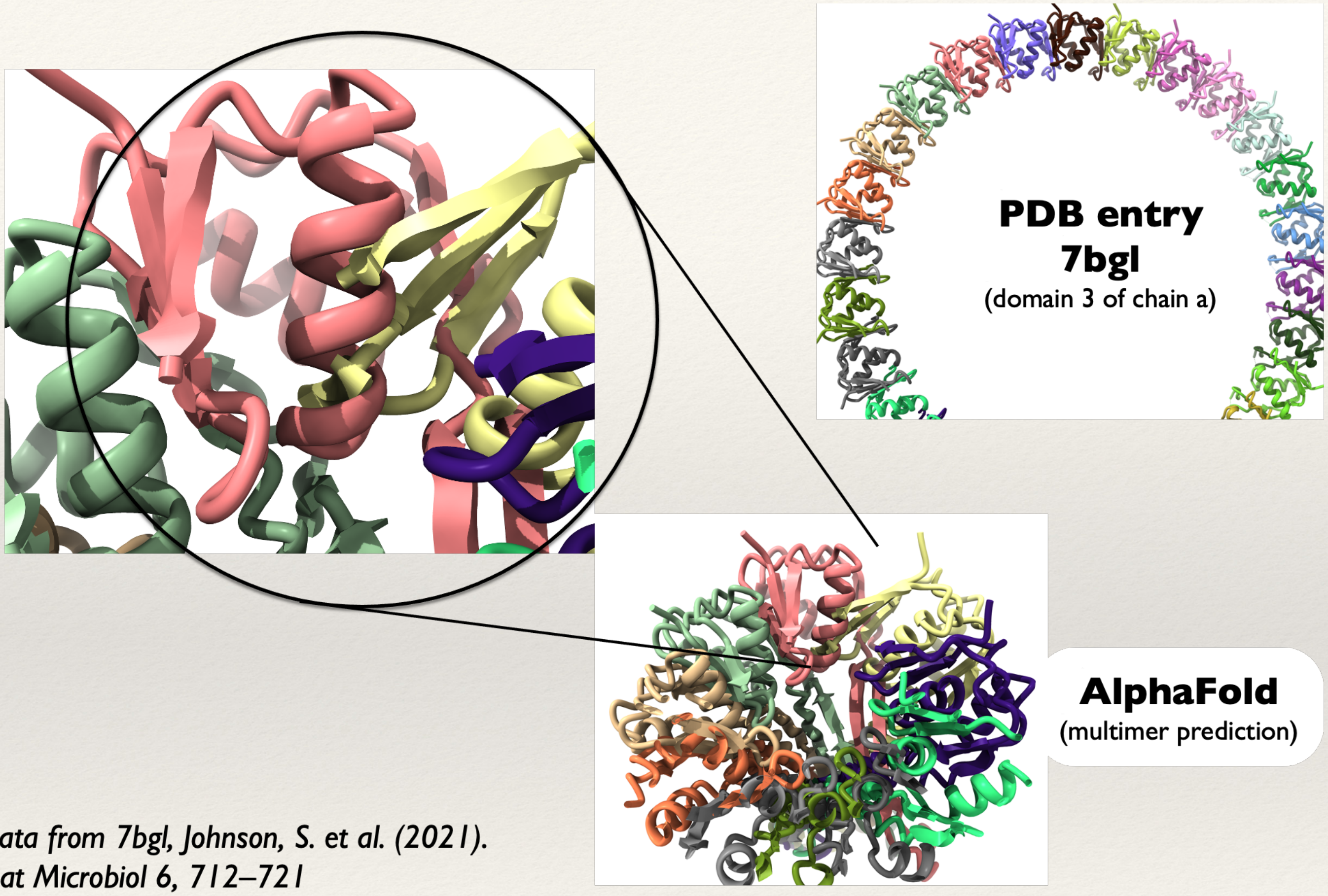
Residues 1-100

High sequence coverage and confidence

Data from 7mjs, Cater, R.J., et al. (2021). Nature 595, 315–319

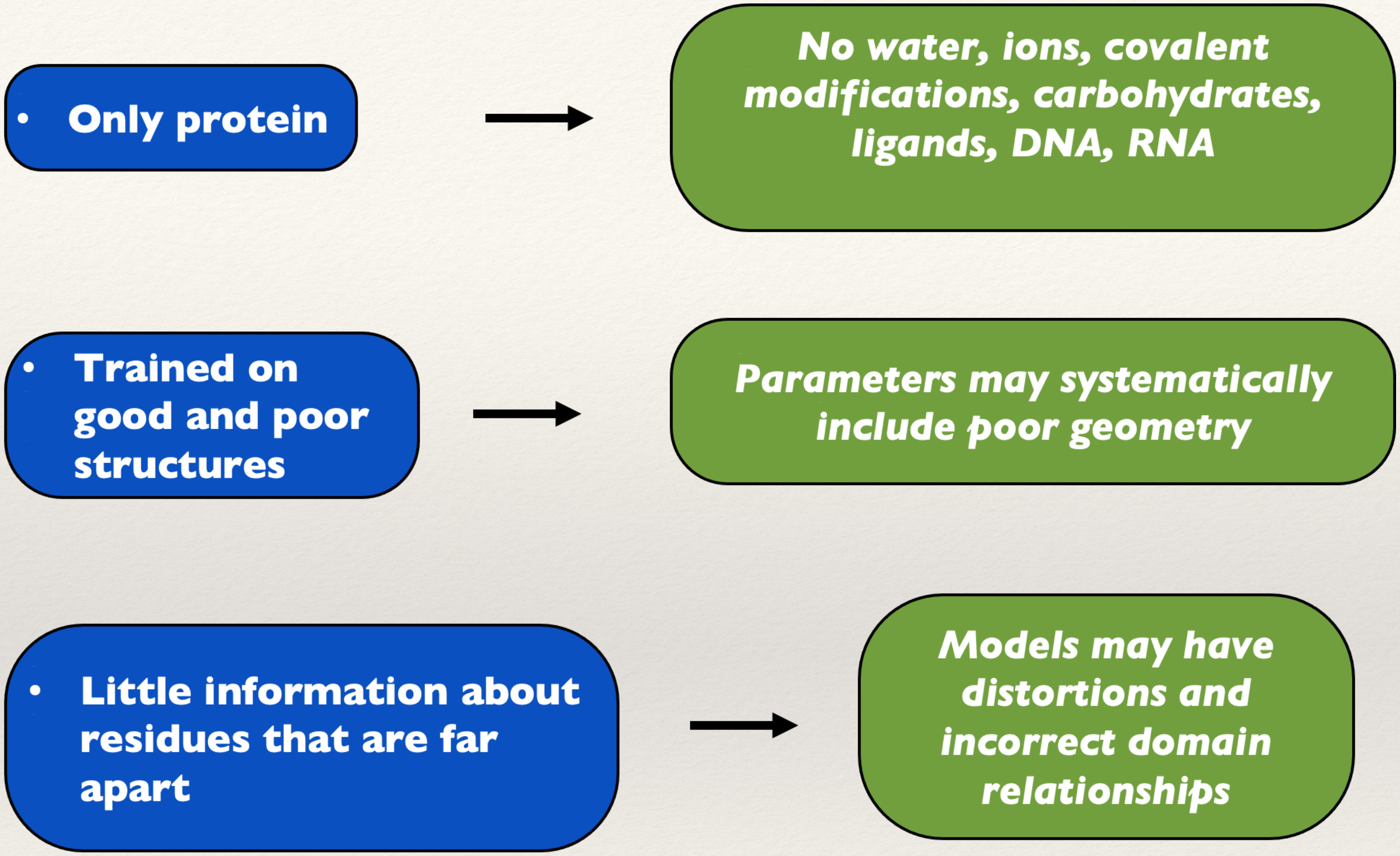
Credit: Tom Terwilliger, Los Alamos NL

Multimeric proteins



*Data from 7bgl, Johnson, S. et al. (2021).
Nat Microbiol 6, 712–721*

Credit: Tom Terwilliger, Los Alamos NL



- **Only protein** → ***No water, ions, covalent modifications, carbohydrates, ligands, DNA, RNA***
- **Trained on good and poor structures** → ***Parameters may systematically include poor geometry***
- **Little information about residues that are far apart** → ***Models may have distortions and incorrect domain relationships***

Credit: Tom Terwilliger, Los Alamos NL