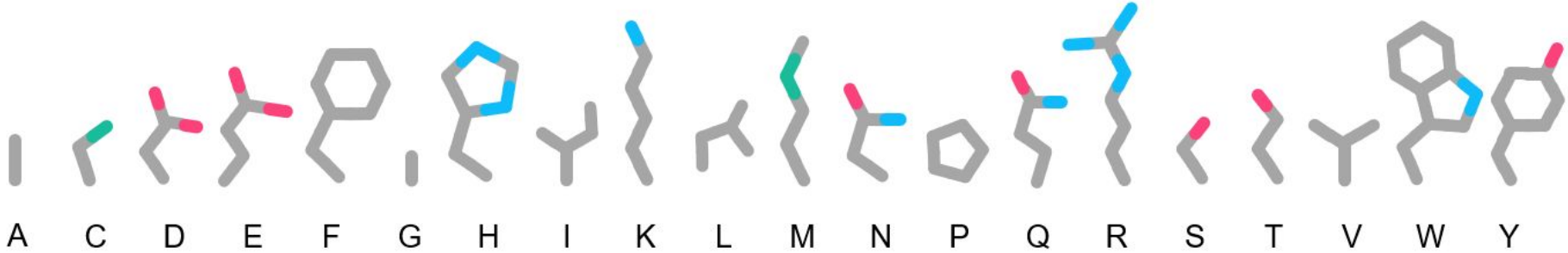
CSE7850/CX4803 Machine Learning in Computational Biology

Lecture 15: Learning from Structure Data

Yunan Luo

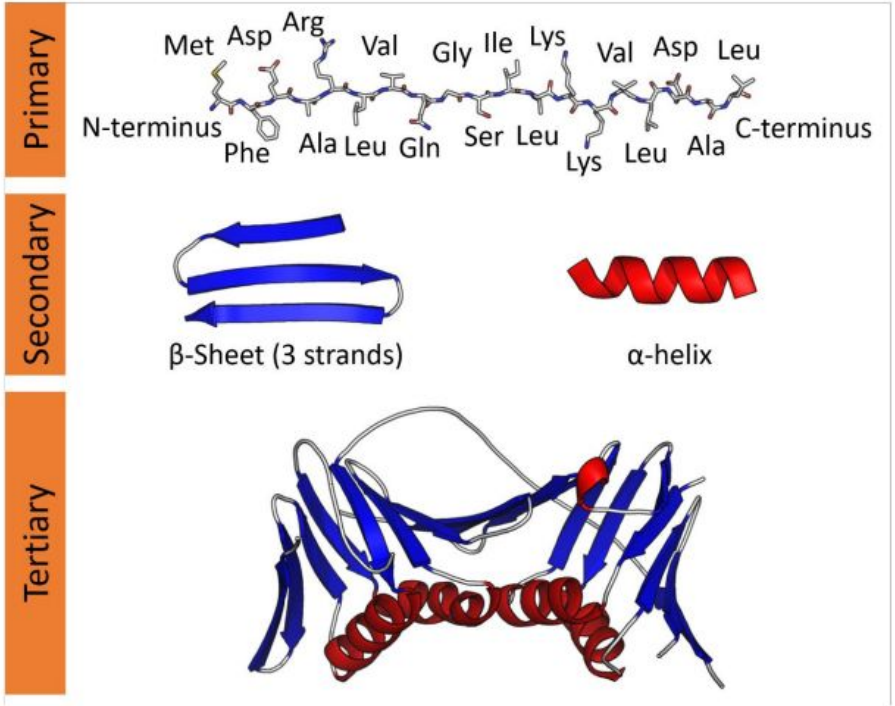
Amino acids are the building blocks of proteins



Amino acids vary in

- Size
- Shape
- Polarity
- Charg

Protein structure hierarchy



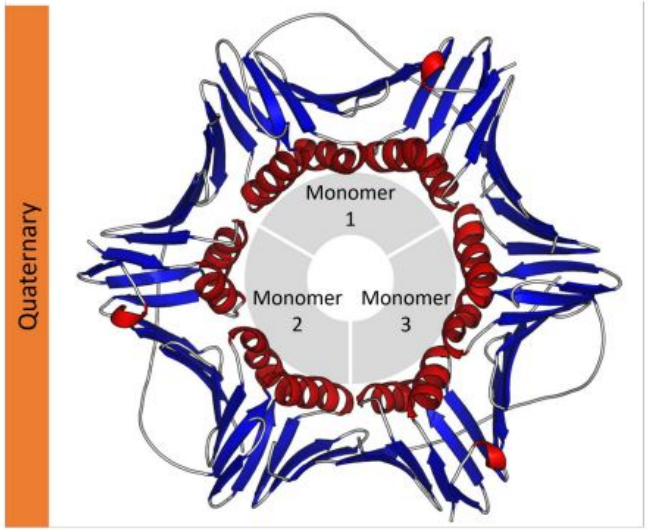
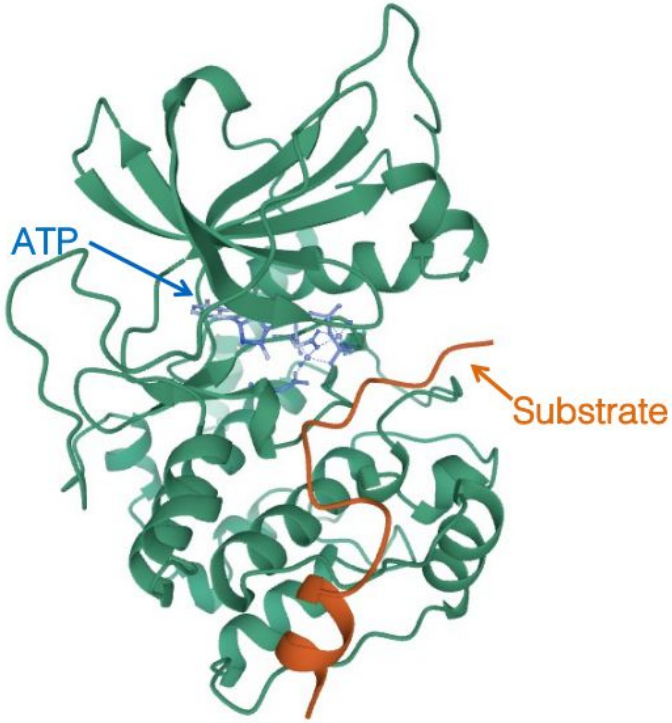
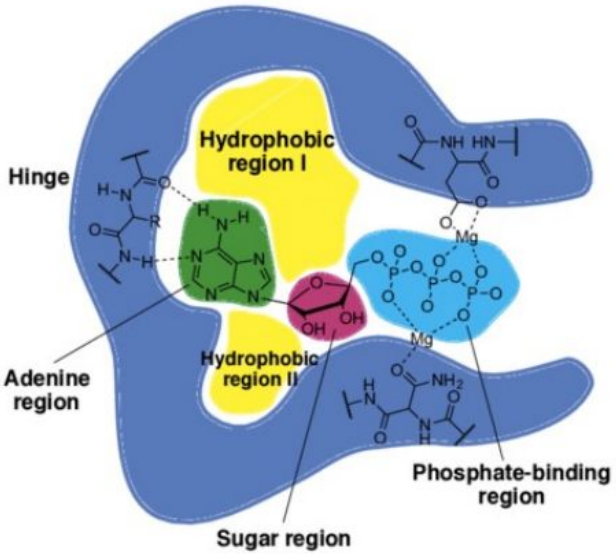


Image from wiki

Structure provides insight on function

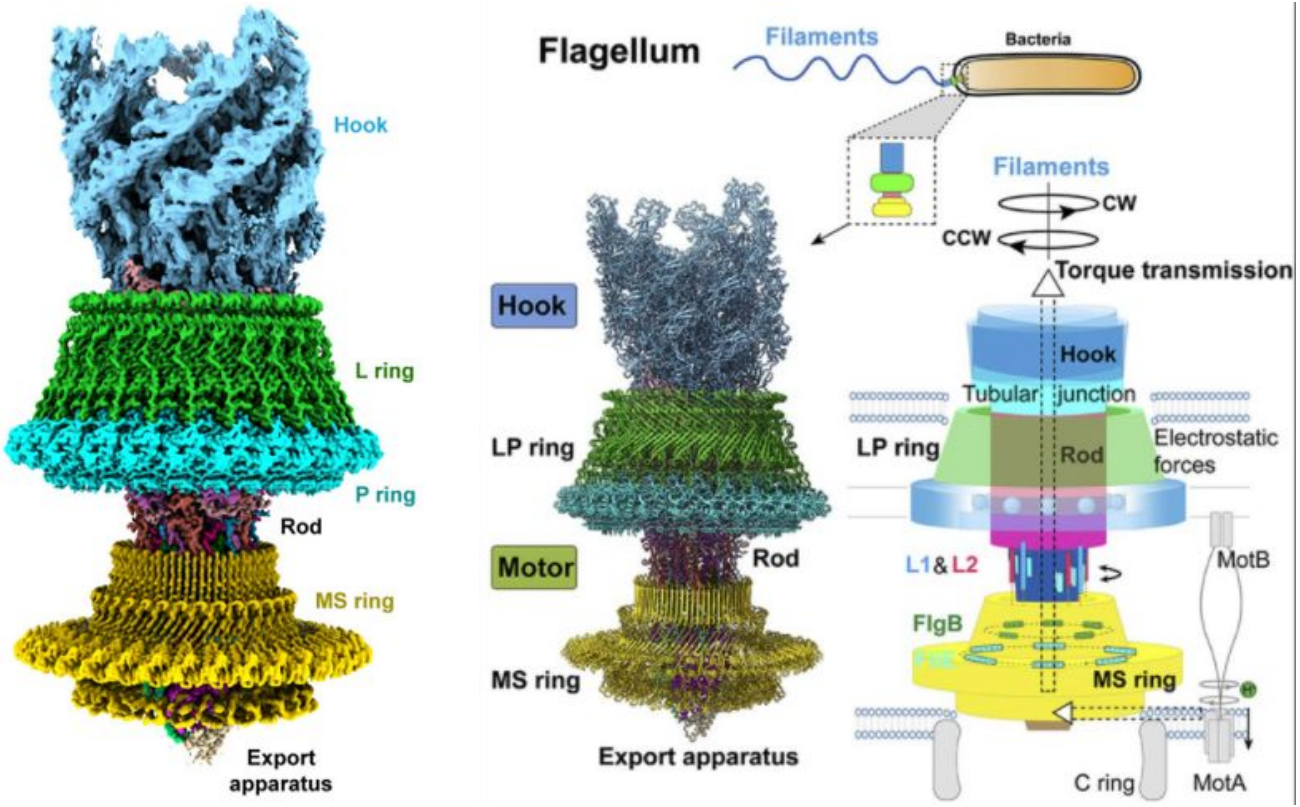


Protein Kinase

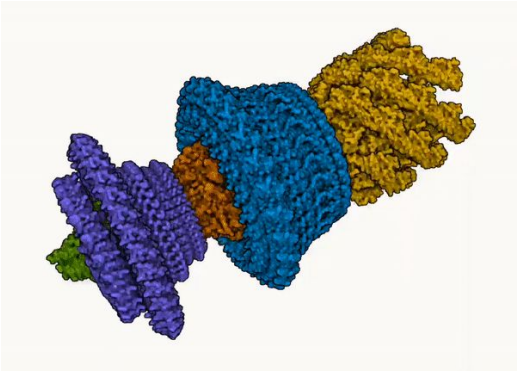


Slides credit: Jian Peng

Structure provides insight on function

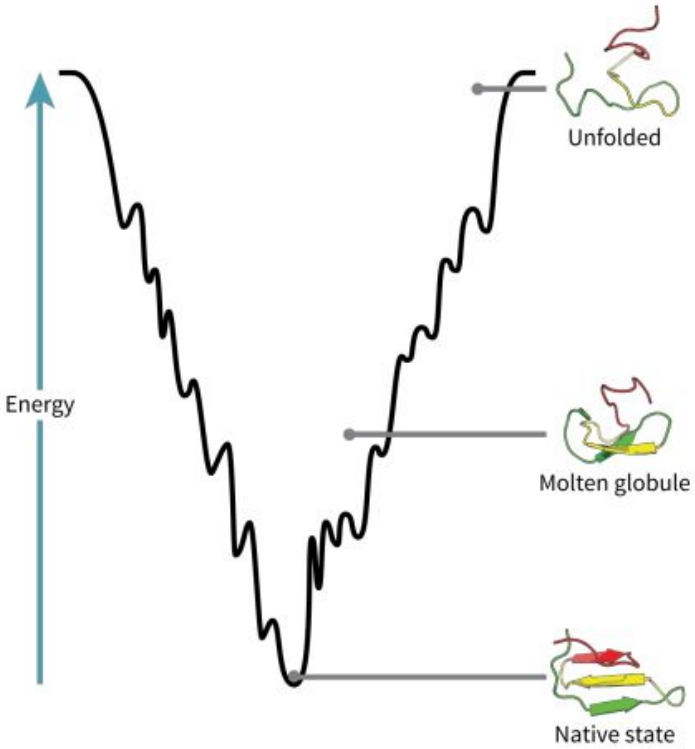


Bacterial flagellar motor (Tan et al., *Cell*, 2021)



Source: Twitter

The protein folding problem



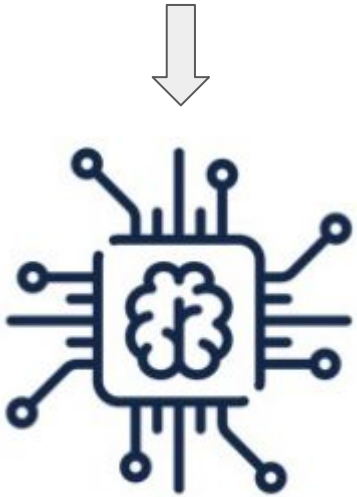
https://mit6874.github.io/assets/sp2021/recitations/r11.pdf

- The function of a protein is determined in large part by its 3D shape
- Can we predict the 3D structure of a protein given only its (1D) amino-acid sequence?

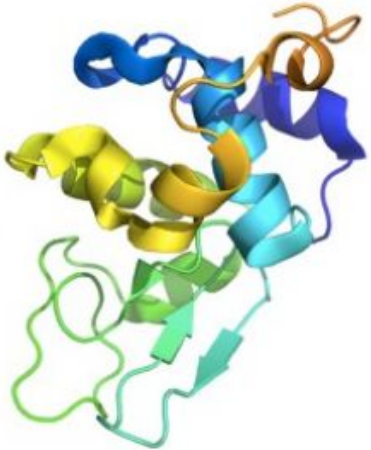
Protein structure prediction

Amino acid sequence

MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK



Algorithm / Model



Protein structure

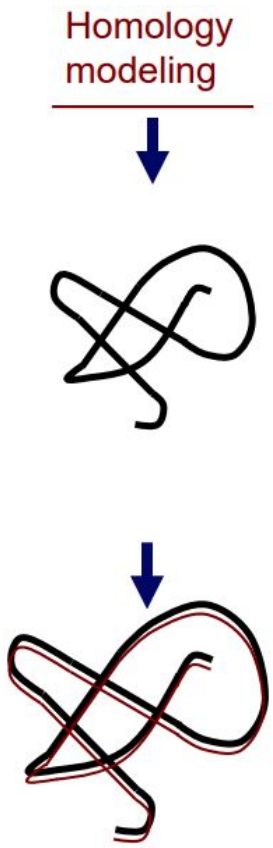
Slides credit: Jinbo Xu

Classical approaches for protein structure prediction

- Homology modeling
- Fold recognition (threading)
- Fragment assembly
- Molecular dynamics

Classical approaches for protein structure prediction

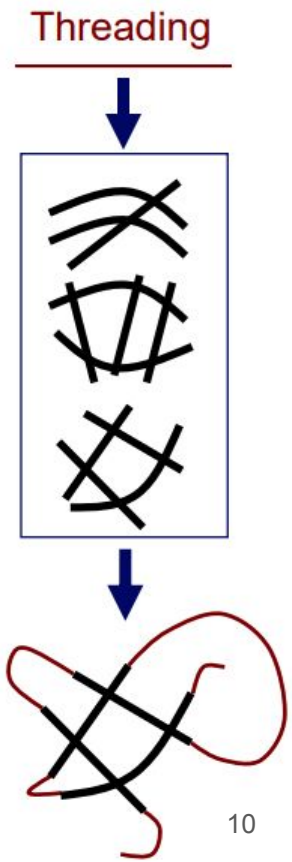
- **Homology modeling**
  - given a query sequence **Q**, a database of protein structures, do:
    - find protein **P** such that
      - structure of **P** is known
      - P has high sequence similarity to **Q**
    - return P's structure as an approximation to **Q**'s structure
- Fold recognition (threading)
- Fragment assembly
- Molecular dynamics



Adapted from CS776 by Anthony Gitter

Classical approaches for protein structure prediction

- Homology modeling
- **Fold recognition (threading)**
  - given a query sequence **Q**, a database of known folds, do:
    - find fold **F** such that **Q** can be aligned with **F** in a highly compatible manner
    - return **F** as an approximation to **Q**'s structure
- Fragment assembly
- Molecular dynamics

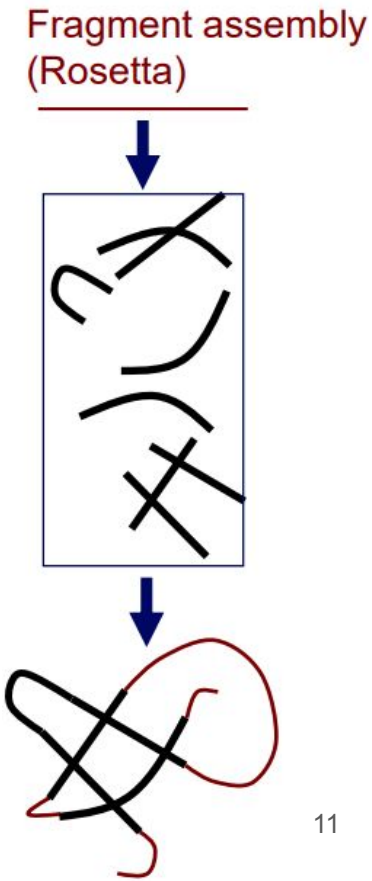


Fold: a description of the relative orientation of the secondary structure making up the tertiary structure.

Adapted from CS776 by Anthony Gitter

Classical approaches for protein structure prediction

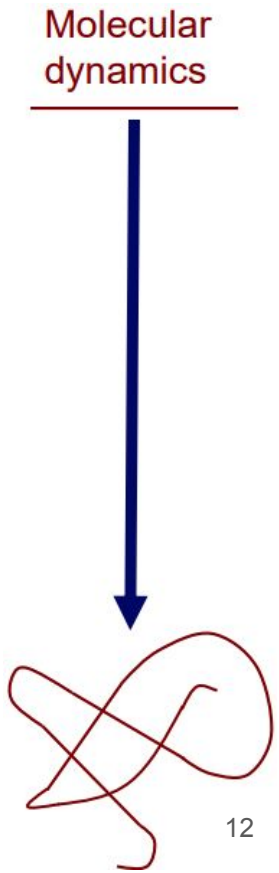
- Homology modeling
- Fold recognition (threading)
- **Fragment assembly (e.g., Rosetta)**
  - given a query sequence **Q**, a database of structure fragments, do
    - find a set of **fragments** that **Q** can be aligned with in a highly compatible manner
    - return fragment assembly as an approximation to **Q**'s structure
- Molecular dynamics



Adapted from CS776 by Anthony Gitter

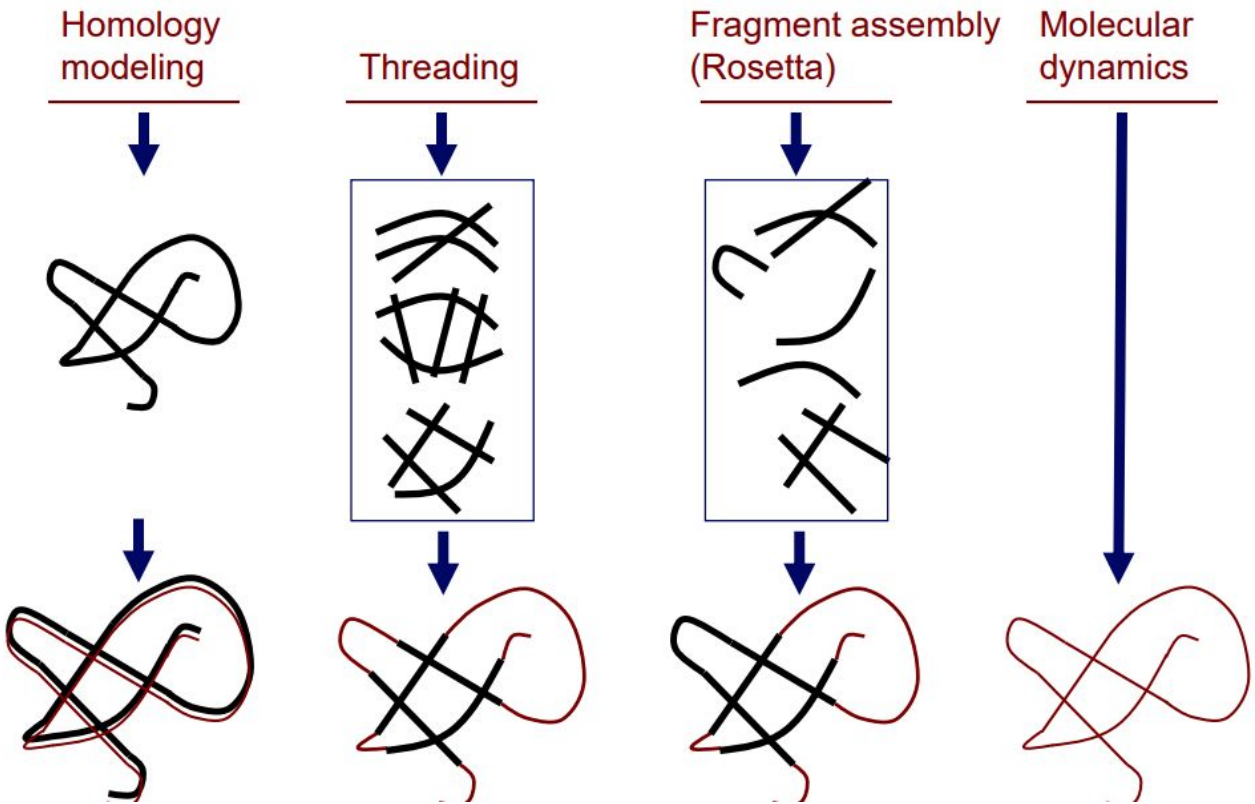
Classical approaches for protein structure prediction

- Homology modeling
- Fold recognition (threading)
- Fragment assembly (e.g., Rosetta)
- **Molecular dynamics**
  - given a query sequence Q
  - do: use laws of physics to simulate folding of Q



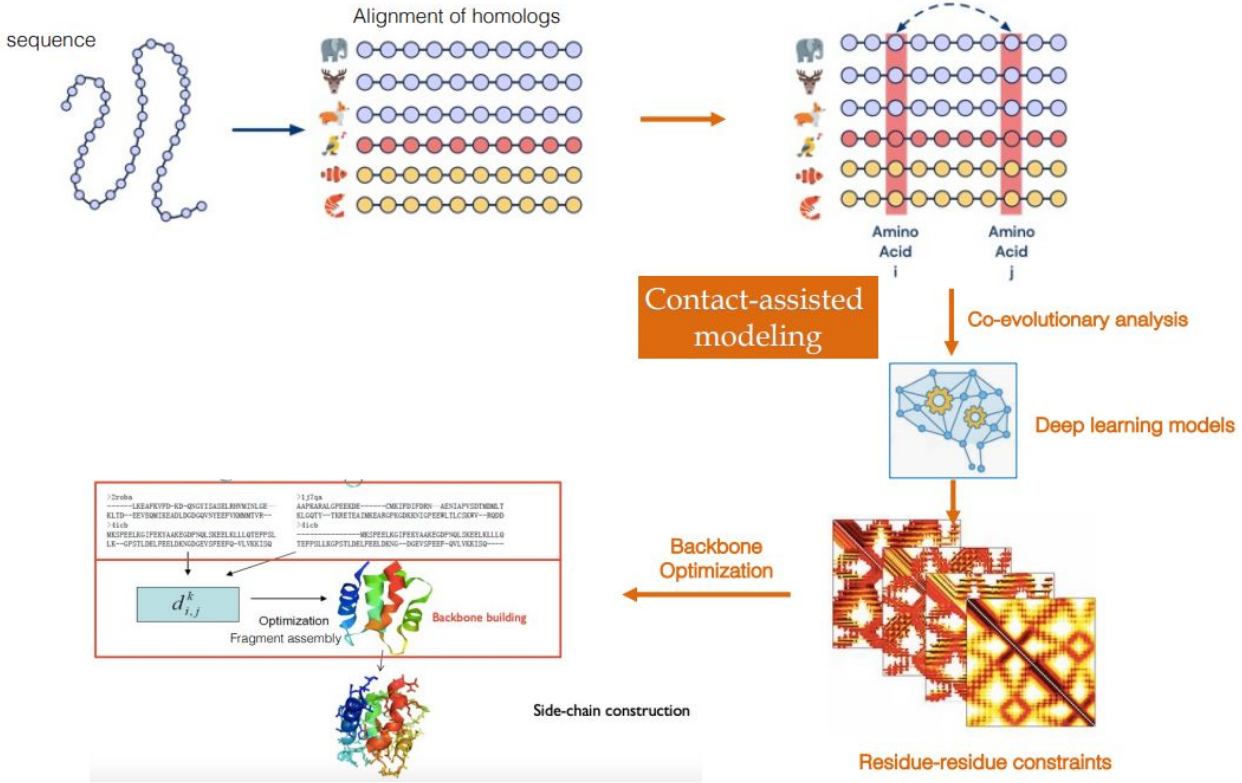
Adapted from CS776 by Anthony Gitter

Classical approaches for protein structure prediction



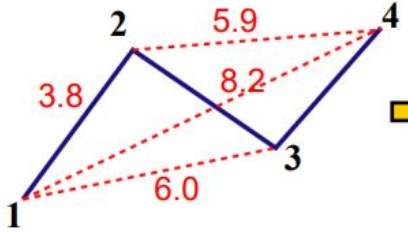
Adapted from CS776 by Anthony Gitter

New approach: Deep Learning



Slides credit: Jian Peng

Protein Distance & Contact Matrix



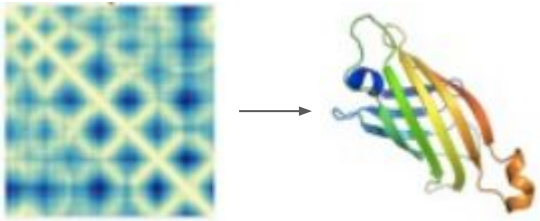
	1	2	3	4
1	0	3.8	6.0	8.2
2	3.8	0	3.8	5.9
3	6.0	3.8	0	3.8
4	8.2	5.9	3.8	0

Distance matrix

Discretize

	1	2	3	4
1	0	1	1	0
2	1	0	1	1
3	1	1	0	1
4	0	1	1	0

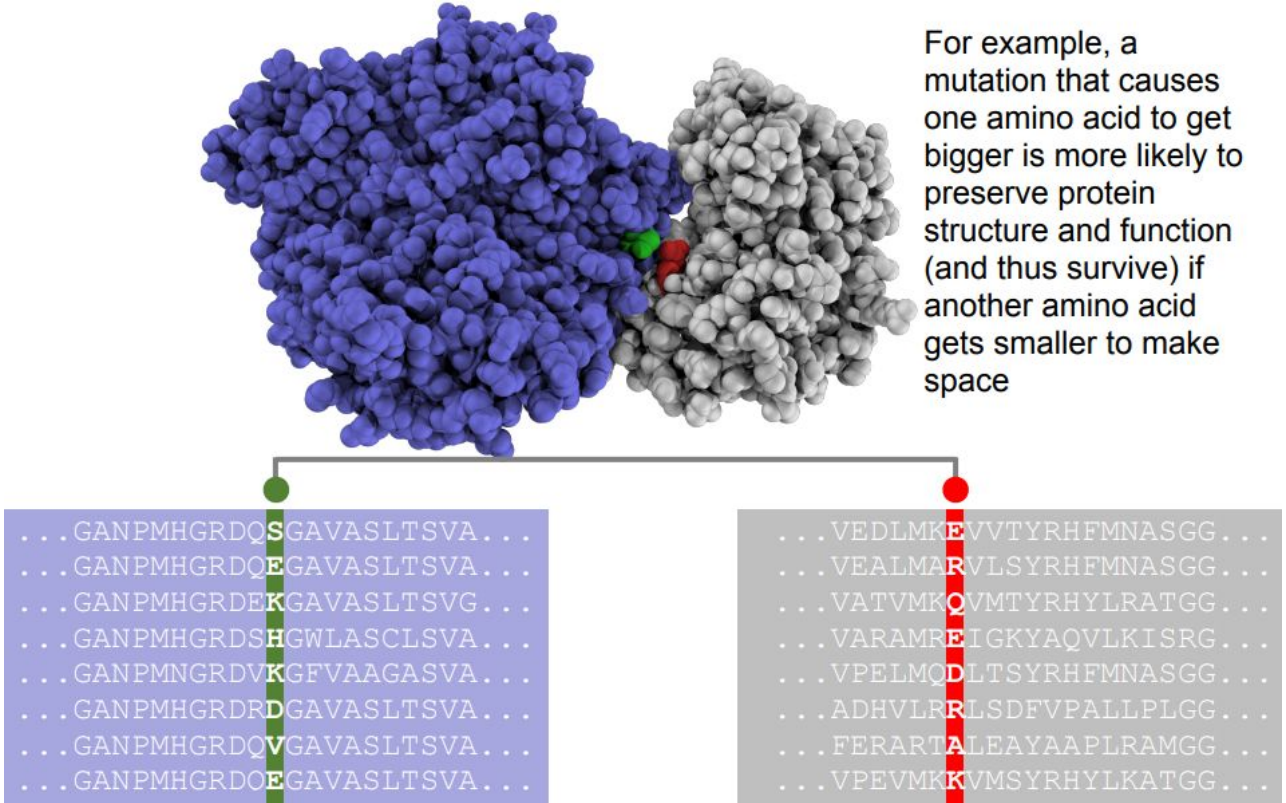
Contact matrix



Liu, Palmedo, et al. Cell Systems 2018

Slides credit: Jinbo Xu

Amino acids in direct physical contact tend to covary or “coevolve” across related proteins



Source: CS279 by Ron Dror

Learning co-evolution from multiple sequence alignment

Multiple sequence alignment

Amino acid i — x_i

Amino acid j — x_j

```
AQKLYLTHIDAEVDGD
ADTLYMTKIHHQFQGD
ADRLFITEVKQVFEGD
ADRLYMTKIHHTFEGD
ADKLYCTLIHNSFDGD
ADRLYMTKIHHEFEGD
ADTLYLTMIHQKFQAD
TDTLYITHIDETFQGD
ADTLYLTQIRNKFQGD
TSRMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD
```

Co-evolution

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

Single potentials — Local preference

Pairwise potentials — Co-evolution strength

Markov random field
Ising (Potts) model
Undirected graphical model

Learning with Markov Random Fields (MRF)

$$L(e) = \prod_{n=1}^{N} \frac{1}{Z_e^{(n)}} \prod_{i}^{L} \exp[e_i(x_i^n) + \sum_{j \neq i} e_{i,j}(x_i^n, x_j^n)]$$

Partition function

Singleton potentials

Local AA preference

Pairwise potentials

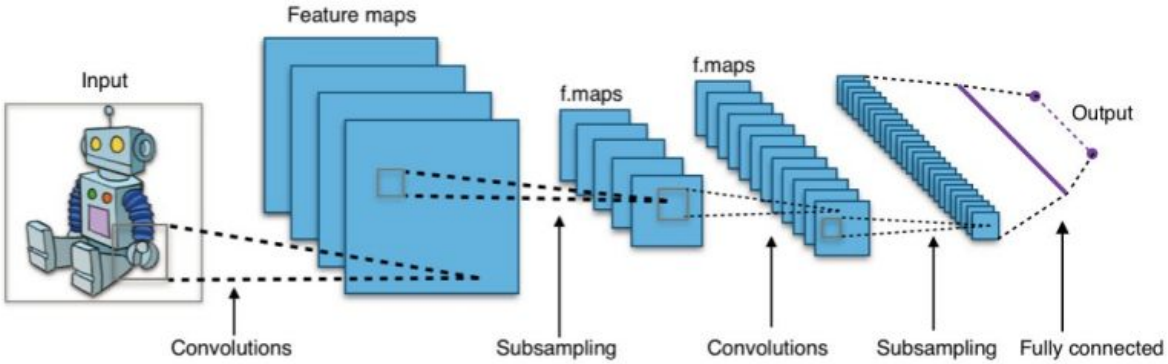
Pairwise AA couplings

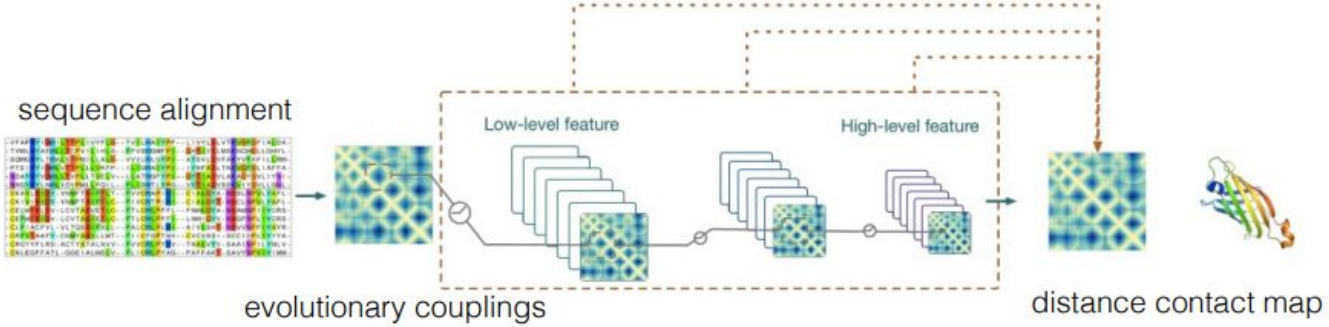
Learning algorithms:

- Mean fields approximation: EVFold, DirectInfo
- Gaussian approximation: PSICOV
- Pseudolikelihood: GREMLIN, CCMpred

Balakrishnan, et al. Proteins 2008; Marks, et al. Plos one 2011; Morcos, et al PNAS, 2011; Jones, et al. Bioinformatics, 2011. Seemayer, et al. Bioinformatics 2014

Deep convolutional NNs recognize coevolutionary patterns

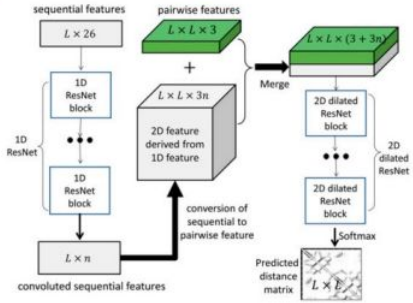




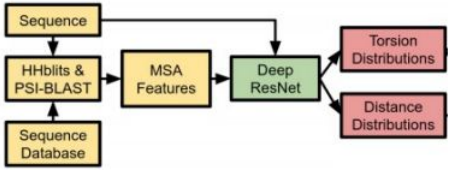
Liu, Palmedo, et al. Cell Systems 2018

Recent developments go beyond contact prediction

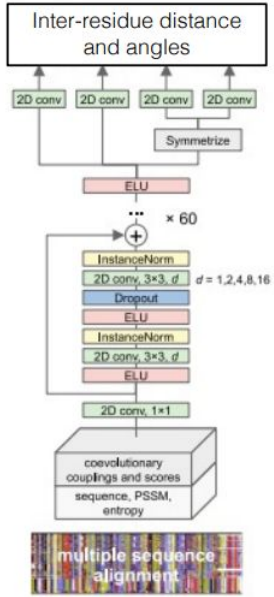
RaptorX



ALPHAFOLD



Rosetta



Improved protein structure prediction using predicted interresidue orientations

Jianyi Yang[a,1], Ivan Anishchenko[b,c,1], Hahnbeom Park[b,c], Zhenling Peng[d], Sergey Ovchinnikov[e], and David Baker[b,c,f,2]

PNAS, 2020

Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7
Received: 2 April 2019
Accepted: 10 December 2019

Andrew W. Senior[1,*], Richard Evans[1], John Jumper[1], James Kirkpatrick[1], Laurent Sifre[1], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]

Nature, 2020 (AlphaFold 1)

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4✉], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4✉]

Nature, 2021 (AlphaFold 2)

RESEARCH ARTICLE | PROTEIN FOLDING

Accurate prediction of protein structures and interactions using a three-track neural network

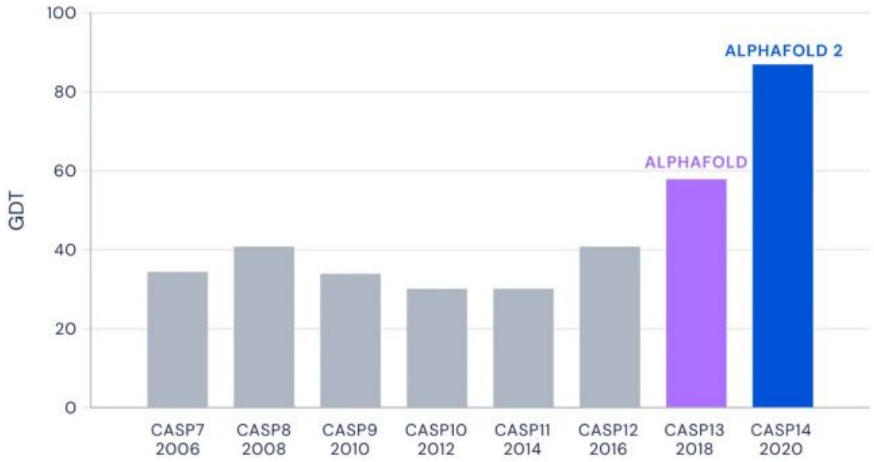
MINKYUNG BAEK, FRANK DIMAIO, IVAN ANISHCHENKO, JUSTAS DAUPARAS, SERGEY OVCHINNIKOV, GYU RIE LEE, JUE WANG, QIAN CONG, LISA N. KINCH, [...] DAVID BAKER +23 authors Authors Info & Affiliations

Science, 2021

Critical Assessment of protein Structure Prediction (CASP)

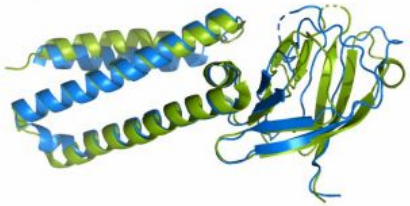
- Since 1994, every two years a contest is held to see who can best predict protein structures from peptide sequences
- Targets structures are held from publication until results are in

CASP14: DeepMind's AlphaFold 2



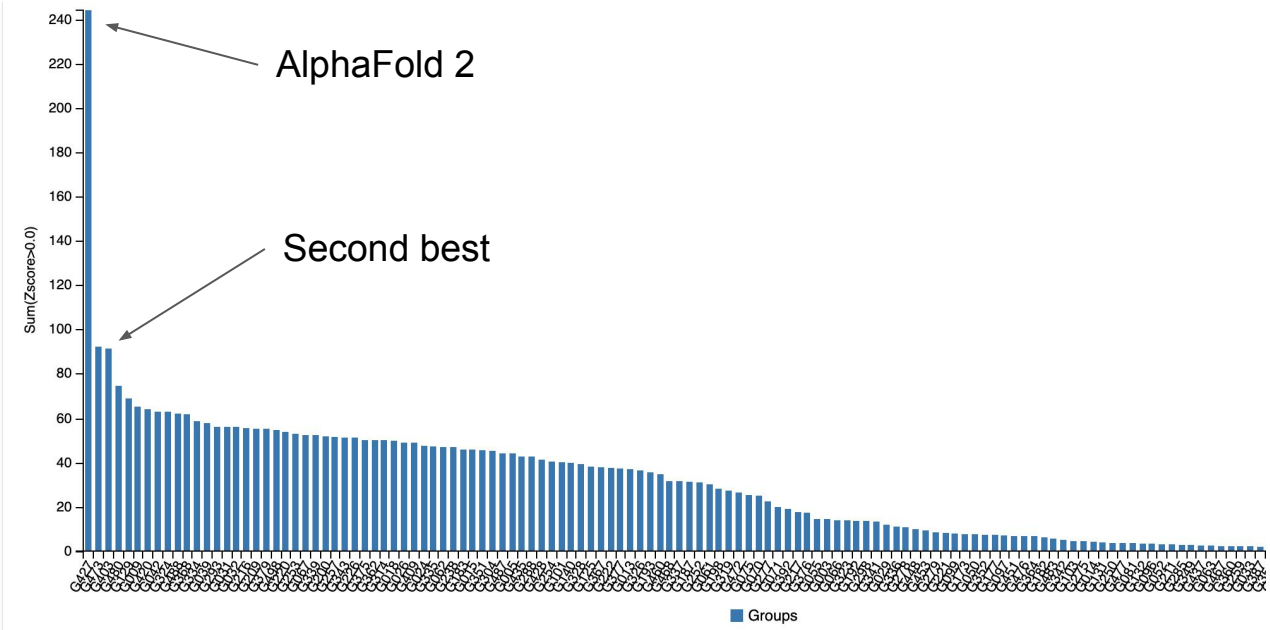
Blue: Predicted
Green: Actual

ORF8



ORF3a

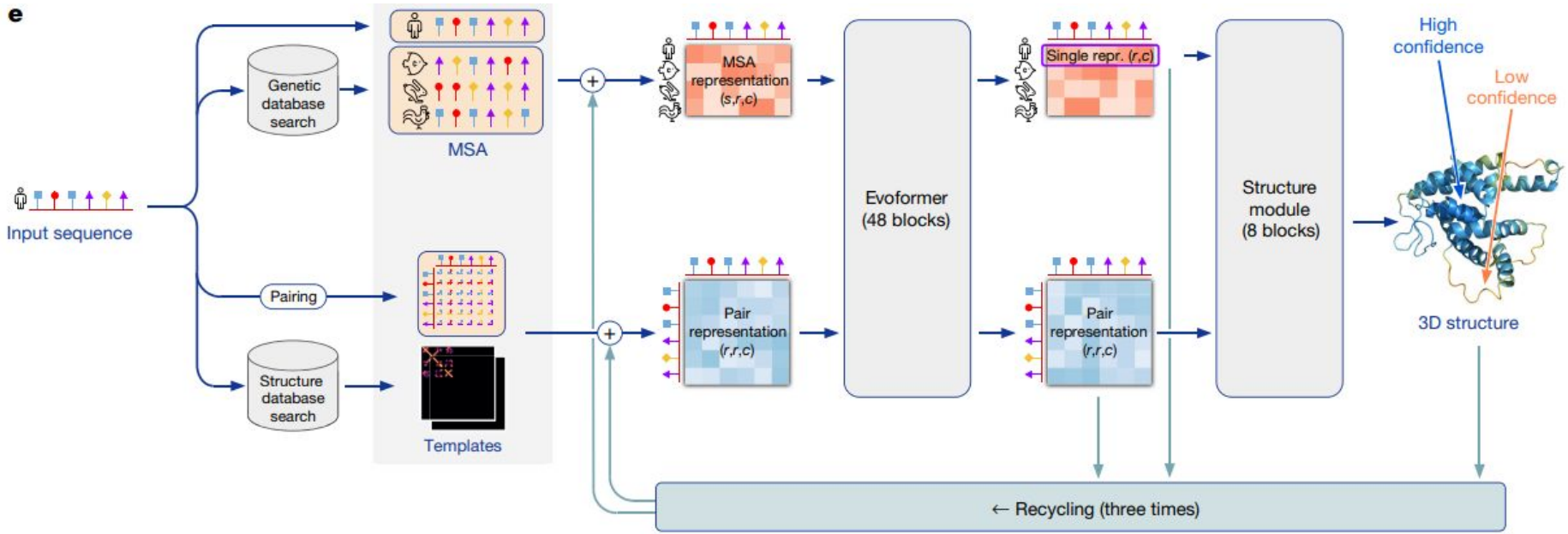
CASP14: DeepMind’s AlphaFold 2



#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2
3	403	BAKER-experimental	92	88.9672	3	0.9670	3	91.4731	3
4	480	FEIG-R2	92	72.5351	4	0.7884	4	74.5627	4

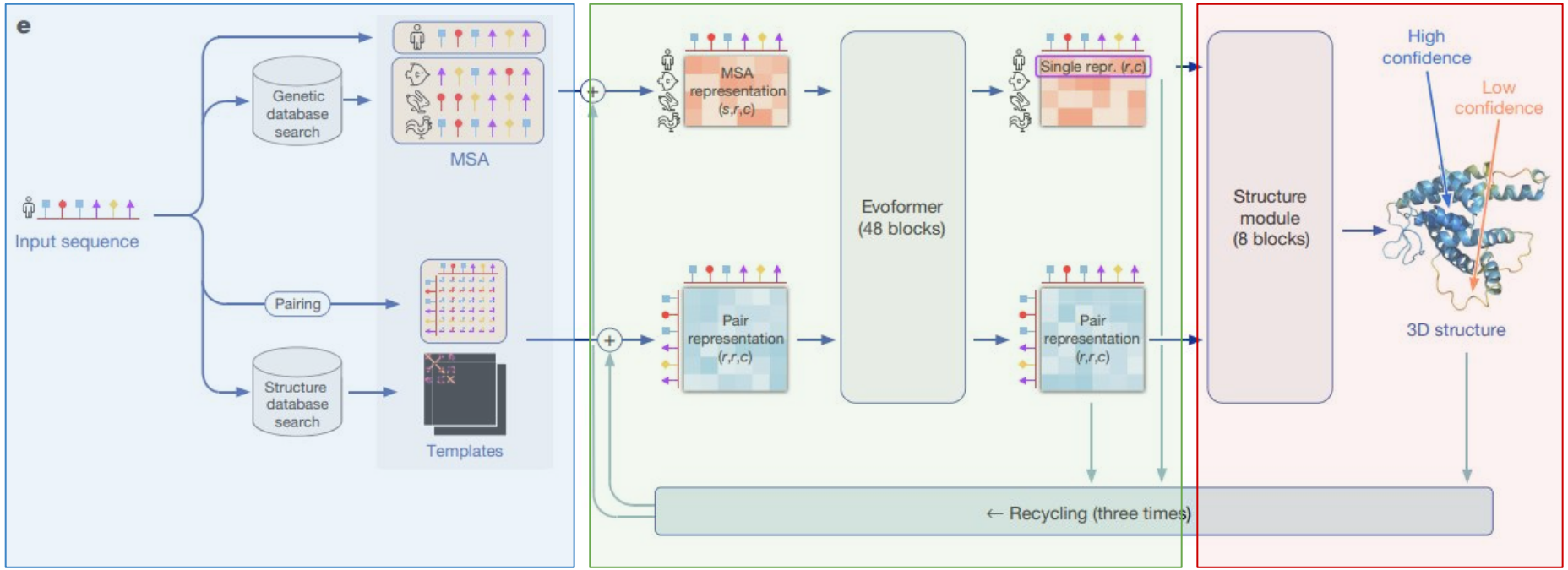
AlphaFold

AlphaFold model



Jumper et al., *Nature*, 2021

AlphaFold model



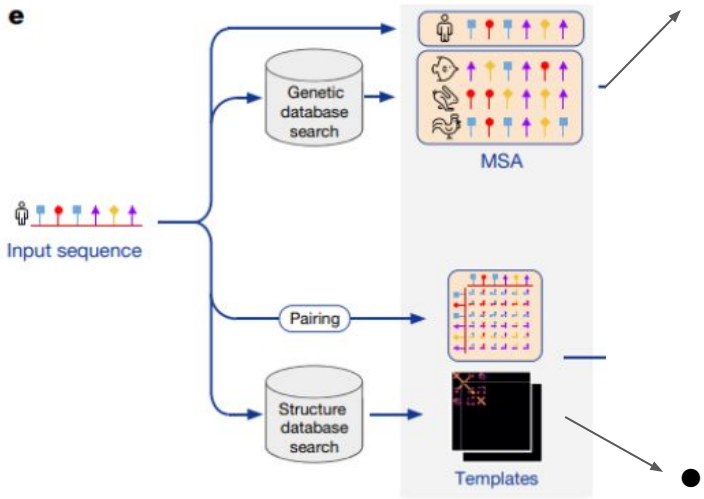
Input module

Embedding module (Evoformer)

Structure module

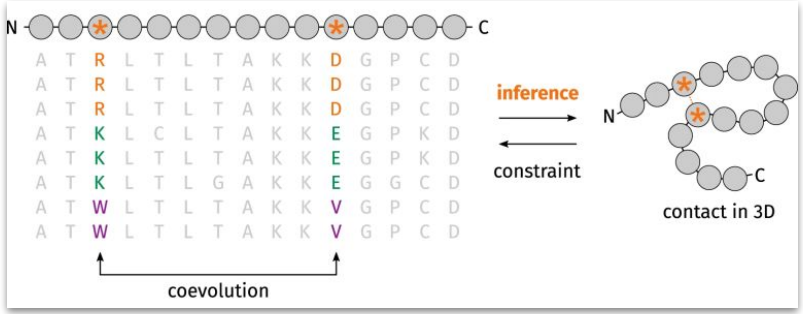
Jumper et al., *Nature*, 2021

Input module



Two types of input

- Multiple sequence alignment (MSA)
  - Residues in contact tend to coevolve

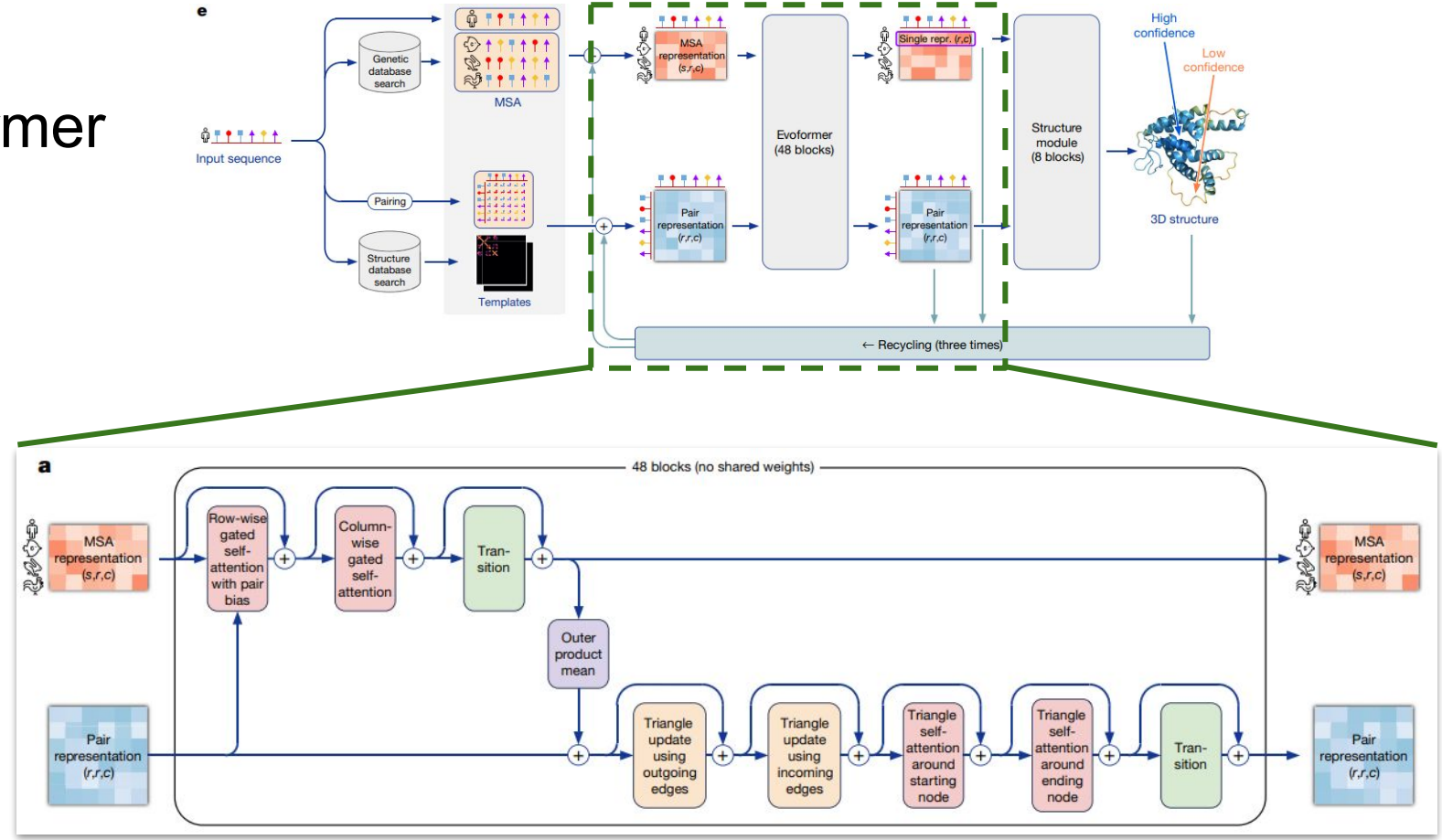


(Image source)

- Template structure
  - Structures are more conserved than sequence
  - Use conserved fragment to guide structure prediction

Jumper et al., *Nature*, 2021

Evoformer



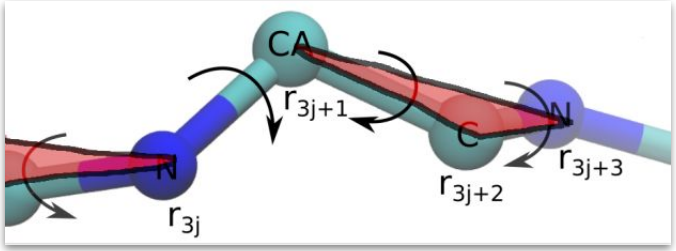
Key idea: learn MSA Representations and Pair Representations and iteratively exchange information between them

Jumper et al., *Nature*, 2021

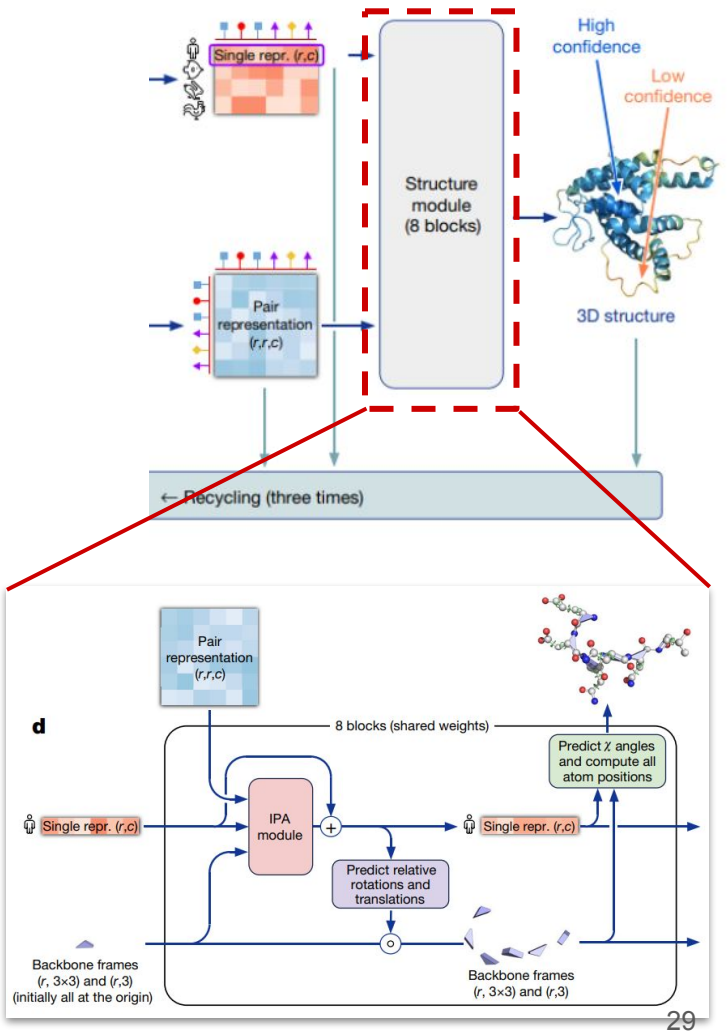
Structure module

- Get structure from the MSA and pair representations
- Each residue is parameterized by a tuple representing rotation and translation

$$T_i := (R_i, \vec{\mathbf{t}}_i)$$



(Image source)



- Iterative update to predict angle and atom positions

Readings

Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]

AlphaFold (*Nature*, 2021)

https://www.nature.com/articles/s41586-021-03819-2

RESEARCH

RESEARCH ARTICLE

PROTEIN FOLDING

Accurate prediction of protein structures and interactions using a three-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10,11], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,7,21], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

RoseTTAFold (*Science*, 2021)

https://www.science.org/doi/10.1126/science.abj8754

- Two deep learning papers on protein structure prediction published on the same day
- Both in the presentation paper list of this course

“Citizen science”: Volunteer/distributed computing

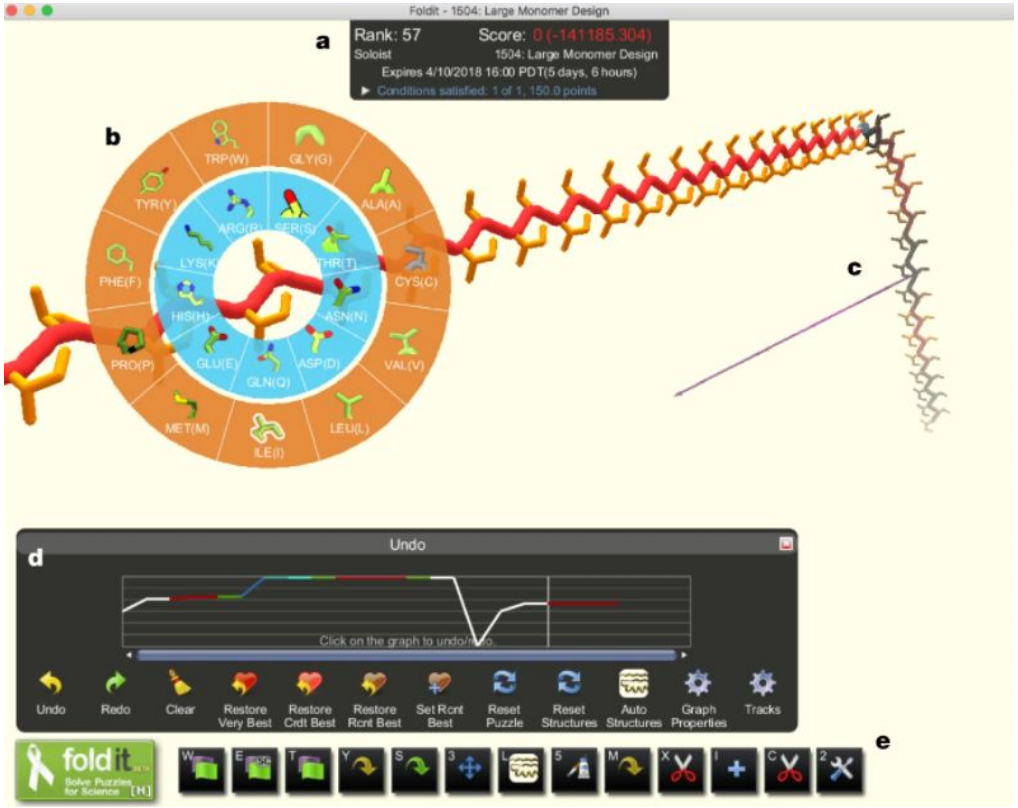
Folding@Home

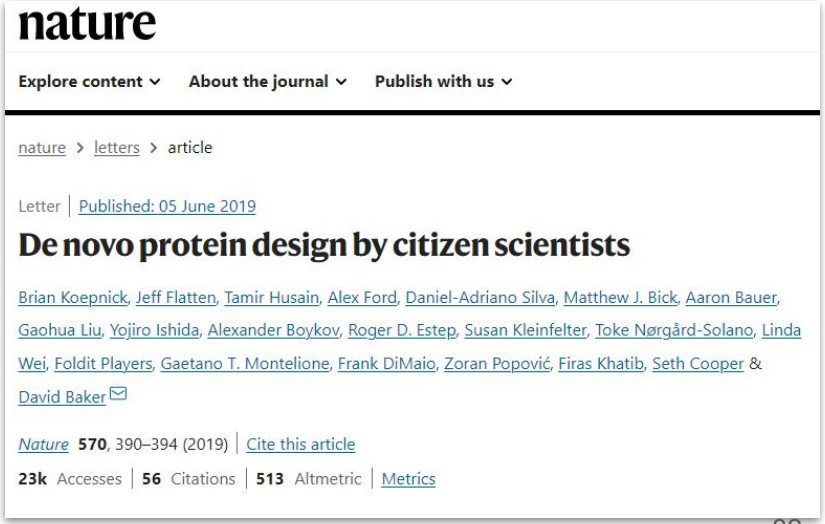
- https://foldingathome.org/home/
- Molecular dynamics (MD) simulations

Rosetta@home

- http://boinc.bakerlab.org
- Structure prediction

Foldit





nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › letters › article

Letter | Published: 05 June 2019

De novo protein design by citizen scientists

Brian Koepnick, Jeff Flatten, Tamir Husain, Alex Ford, Daniel-Adriano Silva, Matthew J. Bick, Aaron Bauer, Gaohua Liu, Yojiro Ishida, Alexander Boykov, Roger D. Estep, Susan Kleinfelter, Toke Nørgård-Solano, Linda Wei, Foldit Players, Gaetano T. Montelione, Frank DiMaio, Zoran Popović, Firas Khatib, Seth Cooper & David Baker

Nature **570**, 390–394 (2019) | Cite this article

23k Accesses | **56** Citations | **513** Altmetric | Metrics

Protein Folding & Protein Design

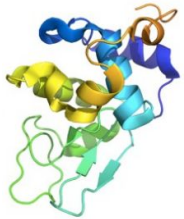
- **Protein folding** (protein structure prediction)
  - Sequence -> Structure

Amino acid sequence

MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK

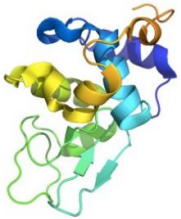
Algorithm / Model

Protein structure



- **Protein design**
  - Structure -> Sequence

Protein structure



Algorithm / Model

Amino acid sequence

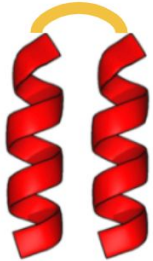
MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK

Protein Design by Hand!

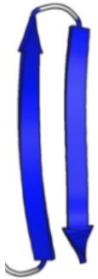
Design sequences that folds into the following structures:

Make a helix!

Make two helices



Make beta sheets?



[Google Colab]

Cheatsheet:

Secondary structure propensity

	C	H	E
A	-0.41	0.46	-0.40
E	-0.27	0.42	-0.53
Q	-0.24	0.36	-0.39
K	-0.05	0.21	-0.32
R	-0.22	0.24	-0.14
M	-0.37	0.26	0.03
L	-0.64	0.35	0.13
W	-0.37	0.06	0.36
Y	-0.36	-0.03	0.46
F	-0.39	-0.03	0.48
I	-0.77	-0.03	0.74
V	-0.69	-0.25	0.89
C	-0.06	-0.25	0.41
T	0.14	-0.39	0.30
H	0.14	-0.15	0.01
S	0.32	-0.25	-0.17
N	0.52	-0.30	-0.65
D	0.50	-0.21	-0.78
G	0.82	-1.05	-0.56
P	0.90	-0.94	-1.06

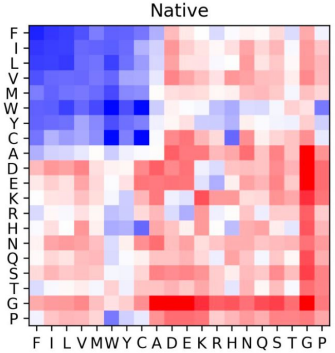
log(P(AA|SS) / P(AA))

(E) Sheets

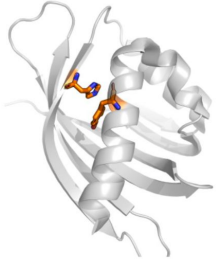
(H) Helix

(C) Random Coil (residues which are not in any of the above conformations)

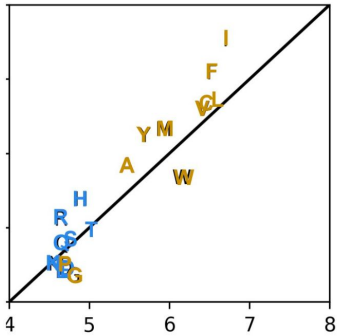
Pairwise potential



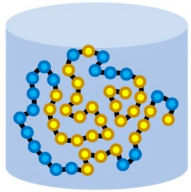
log(P(A,B|Contact) / P(A)*P(B))



Average depth per amino acid type

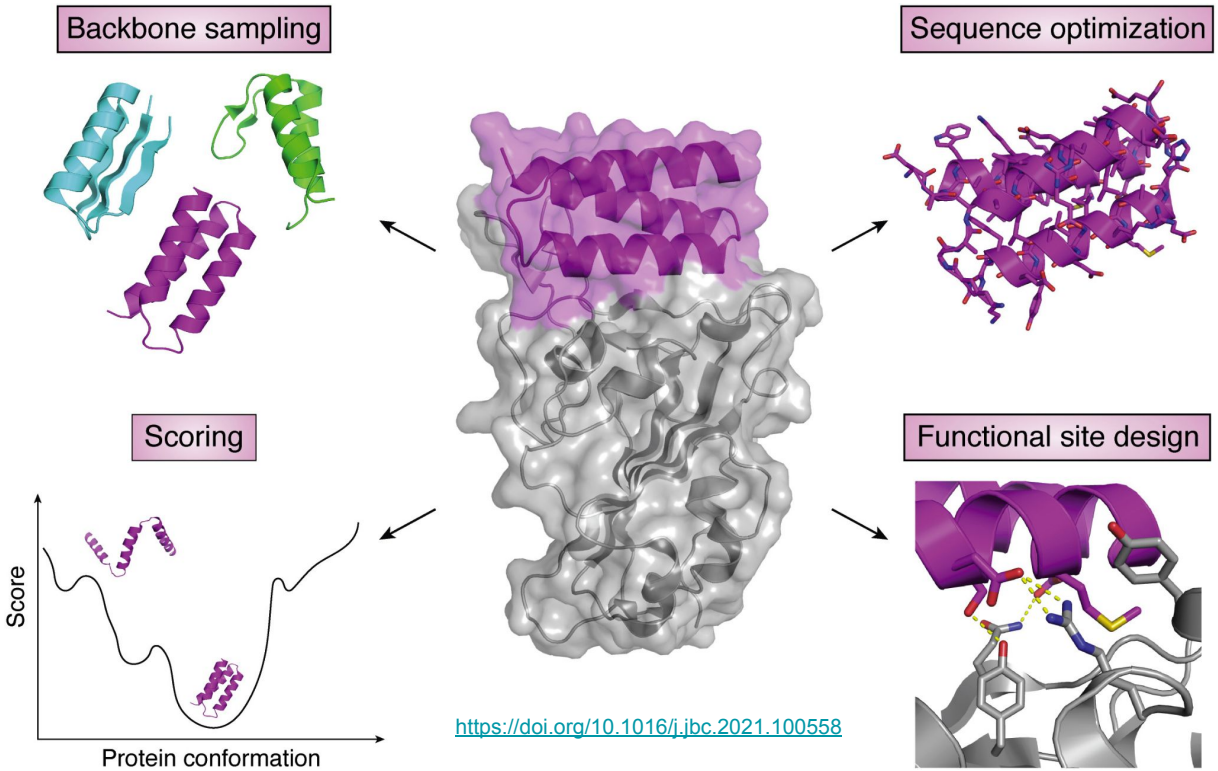


Depth = Distance from surface
aka. how buried is each amino acid



Slides credit: Sergey Ovchinnikov

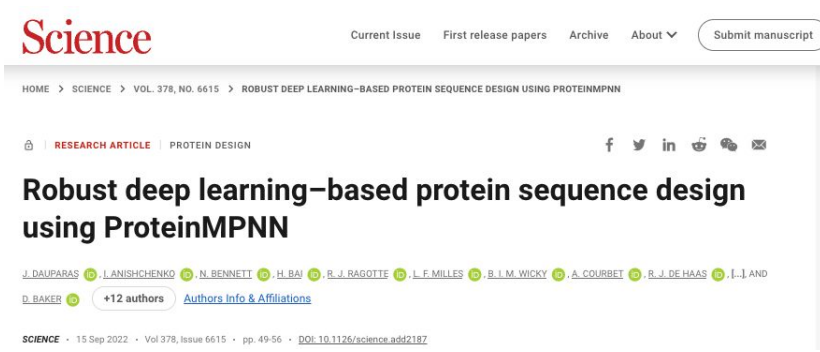
De novo Protein design

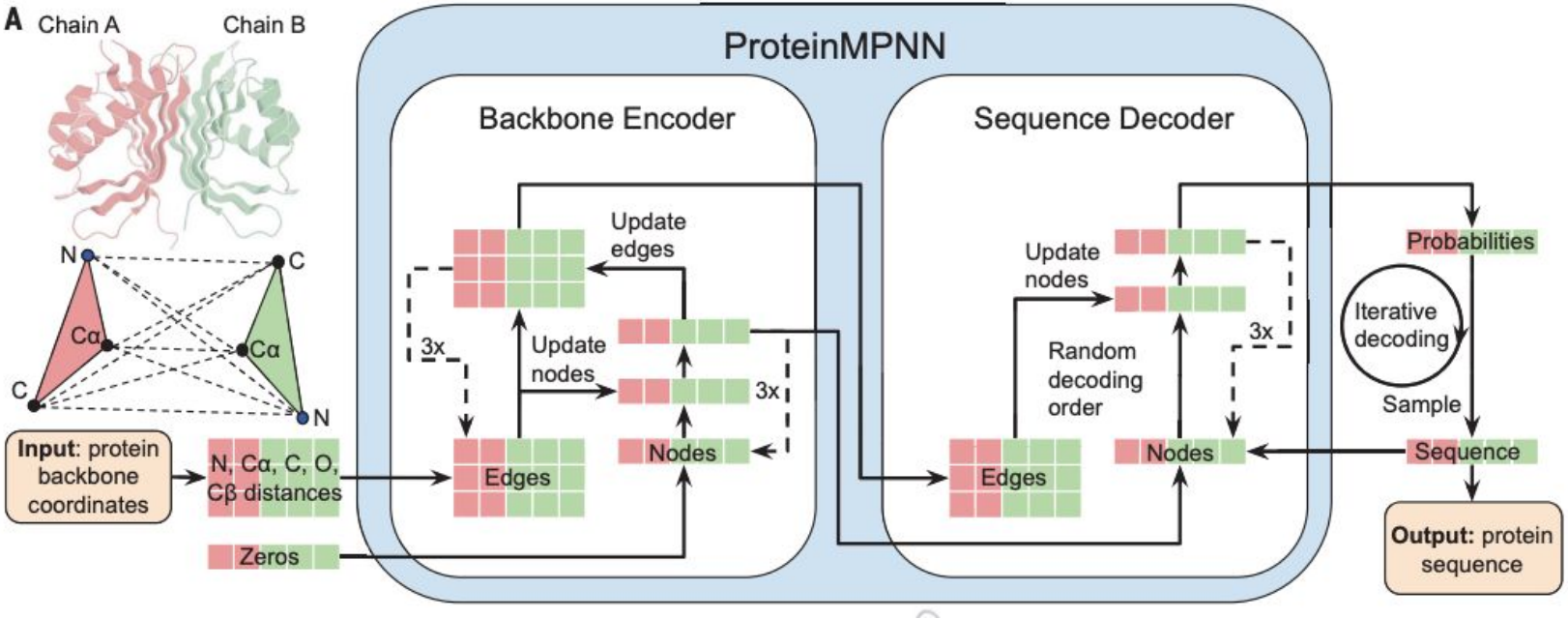


https://doi.org/10.1016/j.jbc.2021.100558

The design of a functional *de novo* protein, for example, a binder (*middle*, *magenta*) to a target protein (*middle*, *gray*), requires sampling of the backbone structure space to find a backbone compatible with the function, sequence optimization to stabilize the backbone, and designing the functional site interactions. A scoring function is necessary to select designs with desired properties, typically by identifying low-energy sequence–structure combinations.

Deep Learning for Protein Design





https://huggingface.co/spaces/simonduerr/ProteinMPNN

Conclusion

- Protein folding
  - Sequence -> Structure
- Protein design
  - Structure -> Sequence

Reminder:

- You are encouraged to ask a question during the Q&A time of a presentation
- Log your question in a survey form to receive the points (by 11:59pm, same day). Once we verified your asked question, you will receive 2 points.
- The question is expected to be in-depth and preferably can prompt discussion/debate
- clarification questions will not count, e.g.:
  - Did the paper compare their model with Random forest?
  - How did they process the input data?