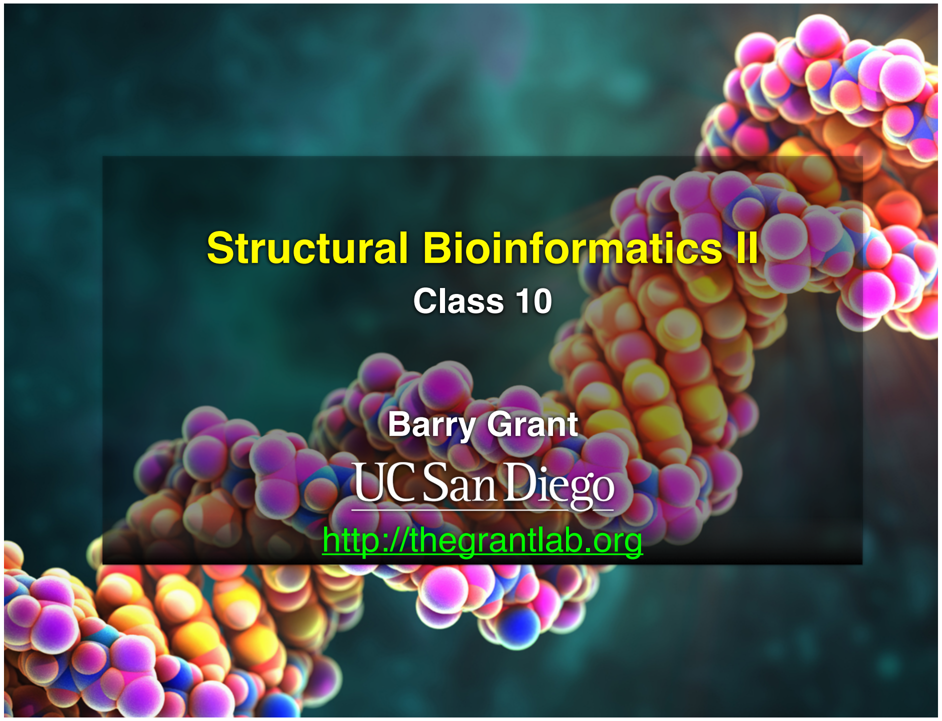


Structural Bioinformatics II

Class 10

Barry Grant

UC San Diego

http://thegrantlab.org

Today's Menu

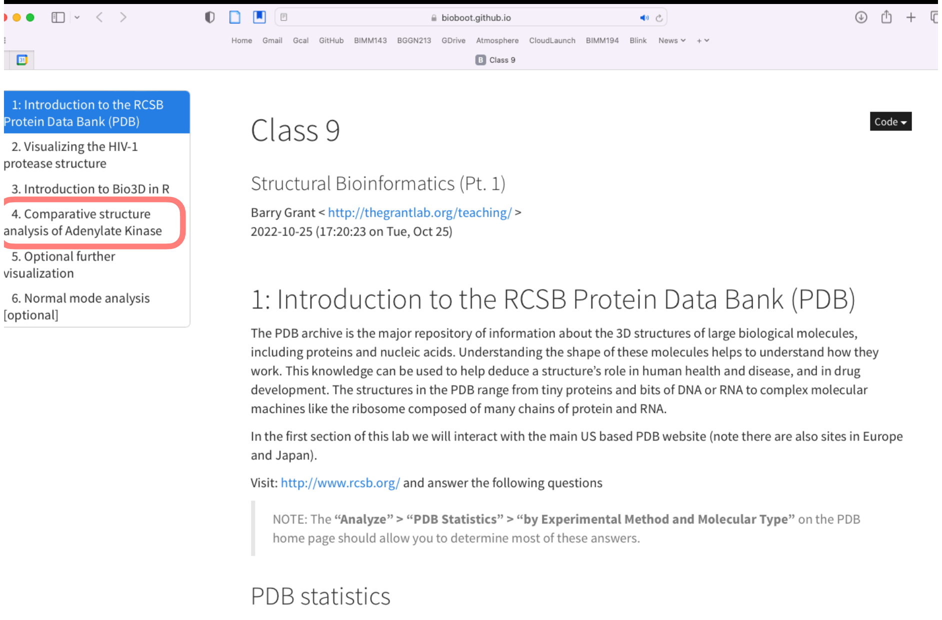
Overview of structural bioinformatics

- Motivations, goals and challenges

Representing, interpreting & modeling protein structure

- Visualizing & interpreting protein structures
- Analyzing protein structures
- Modeling protein structure

Finish last days Lab 09



1: Introduction to the RCSB Protein Data Bank (PDB)
2. Visualizing the HIV-1 protease structure
3. Introduction to Bio3D in R
4. Comparative structure analysis of Adenylate Kinase
5. Optional further visualization
6. Normal mode analysis [optional]

Class 9

Structural Bioinformatics (Pt. 1)

Barry Grant < http://thegrantlab.org/teaching/ >
2022-10-25 (17:20:23 on Tue, Oct 25)

1: Introduction to the RCSB Protein Data Bank (PDB)

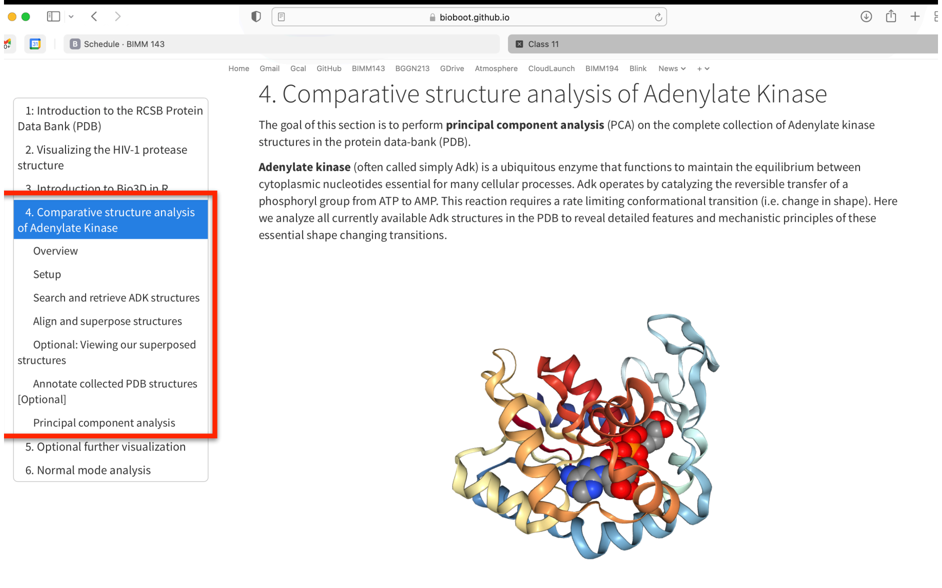
The PDB archive is the major repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. Understanding the shape of these molecules helps to understand how they work. This knowledge can be used to help deduce a structure's role in human health and disease, and in drug development. The structures in the PDB range from tiny proteins and bits of DNA or RNA to complex molecular machines like the ribosome composed of many chains of protein and RNA.

In the first section of this lab we will interact with the main US based PDB website (note there are also sites in Europe and Japan).

Visit: http://www.rcsb.org/ and answer the following questions

> NOTE: The **"Analyze"** > **"PDB Statistics"** > **"by Experimental Method and Molecular Type"** on the PDB home page should allow you to determine most of these answers.

PDB statistics

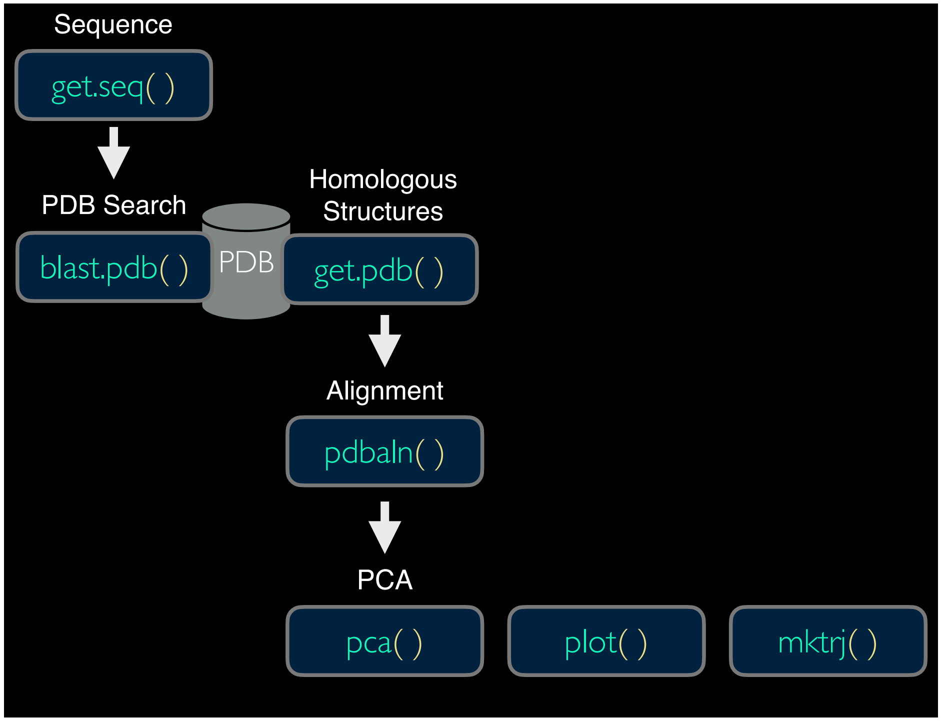


1: Introduction to the RCSB Protein Data Bank (PDB)
2. Visualizing the HIV-1 protease structure
3. Introduction to Bio3D in R
4. Comparative structure analysis of Adenylate Kinase
Overview
Setup
Search and retrieve ADK structures
Align and superpose structures
Optional: Viewing our superposed structures
Annotate collected PDB structures [Optional]
Principal component analysis
5. Optional further visualization
6. Normal mode analysis

4. Comparative structure analysis of Adenylate Kinase

The goal of this section is to perform **principal component analysis** (PCA) on the complete collection of Adenylate kinase structures in the protein data-bank (PDB).

Adenylate kinase (often called simply Adk) is a ubiquitous enzyme that functions to maintain the equilibrium between cytoplasmic nucleotides essential for many cellular processes. Adk operates by catalyzing the reversible transfer of a phosphoryl group from ATP to AMP. This reaction requires a rate limiting conformational transition (i.e. change in shape). Here we analyze all currently available Adk structures in the PDB to reveal detailed features and mechanistic principles of these essential shape changing transitions.



Sequence

get.seq( )

PDB Search

blast.pdb( )

PDB

Homologous Structures

get.pdb( )

Alignment

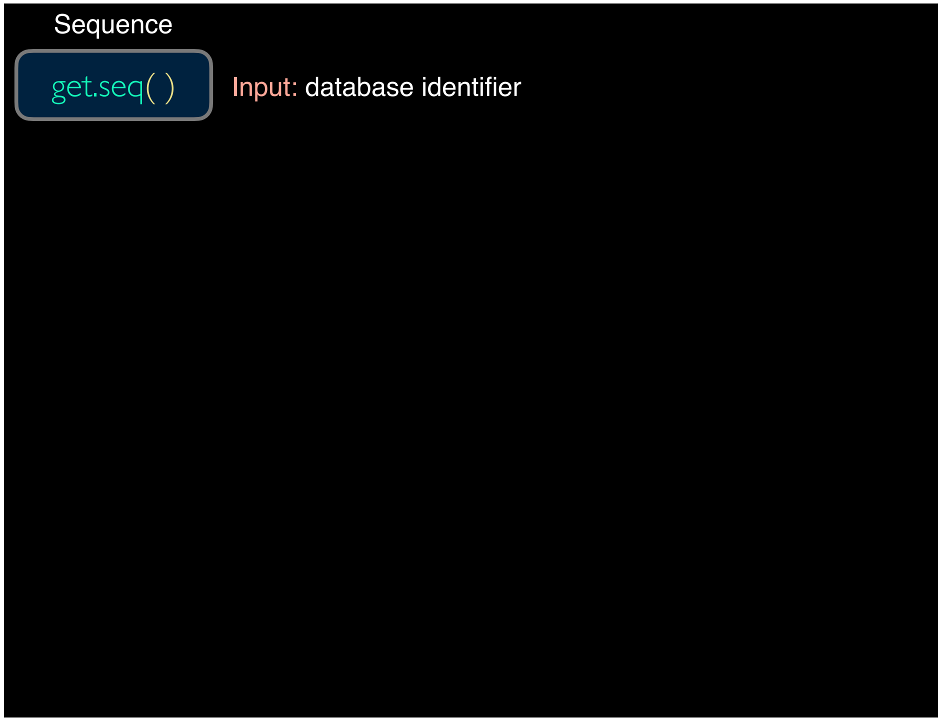
pdbaln( )

PCA

pca( )

plot( )

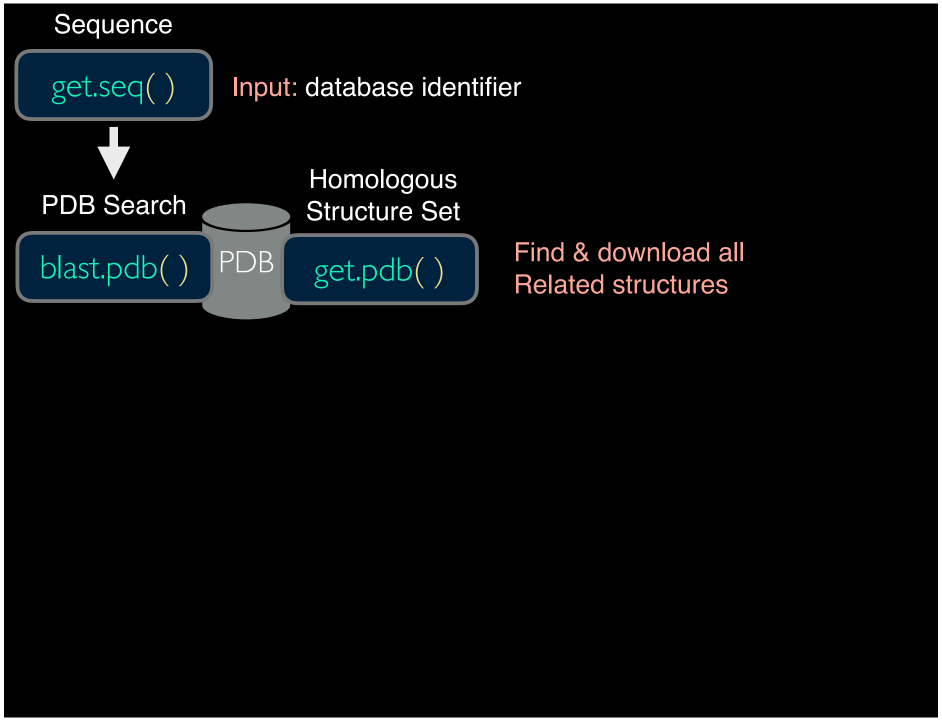
mktrj( )



Sequence

get.seq( )

Input: database identifier



Sequence

get.seq( )

Input: database identifier

PDB Search

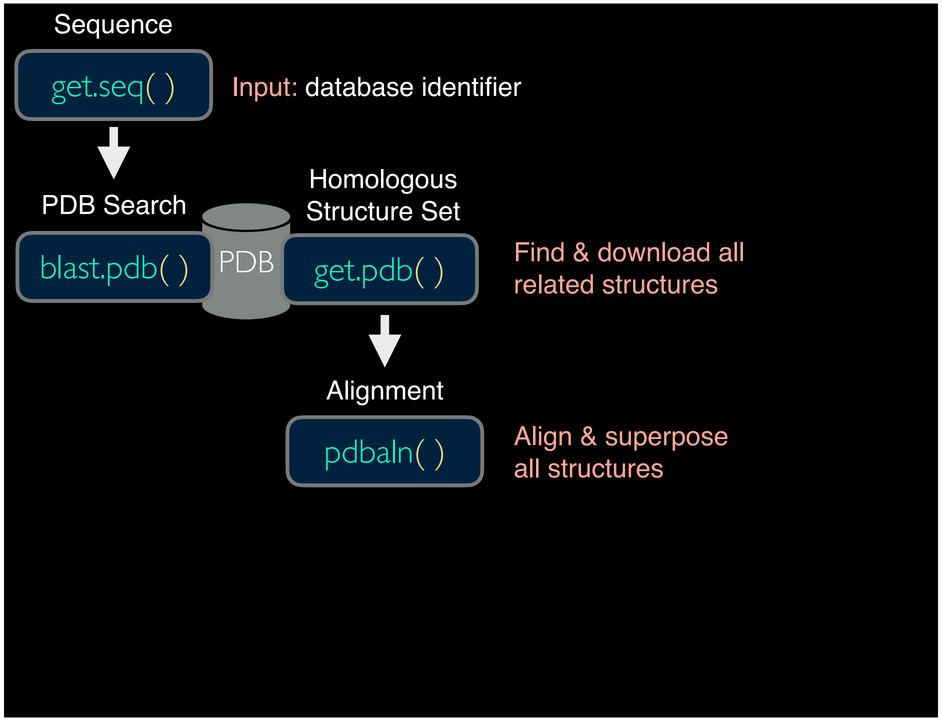
blast.pdb( )

PDB

Homologous Structure Set

get.pdb( )

Find & download all Related structures



Sequence

get.seq( )

Input: database identifier

PDB Search

blast.pdb( )

PDB

Homologous Structure Set

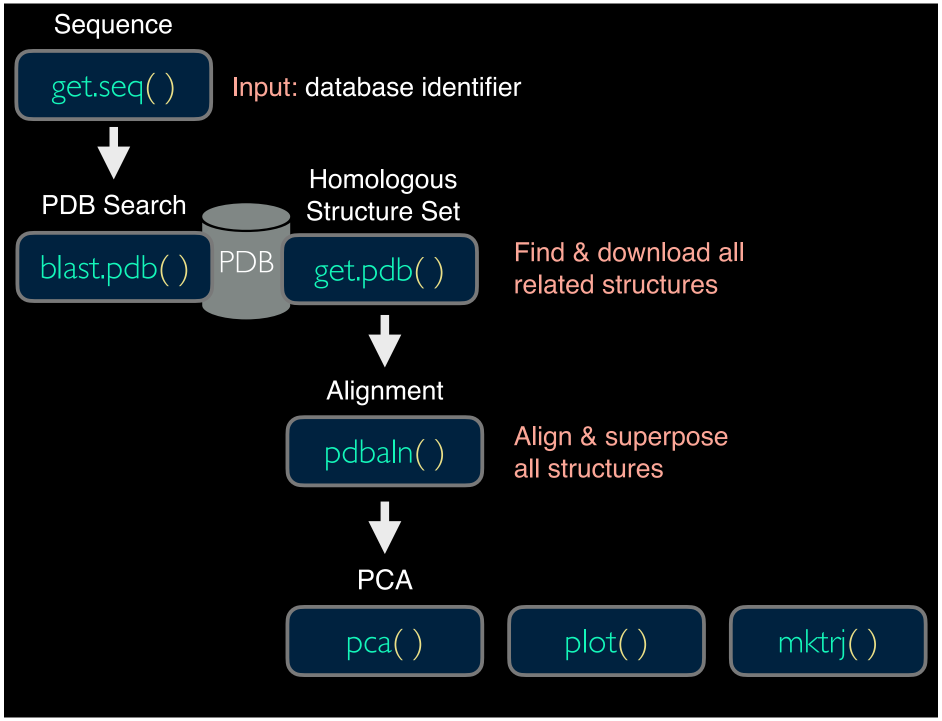
get.pdb( )

Find & download all related structures

Alignment

pdbaln( )

Align & superpose all structures



Sequence

get.seq( )

Input: database identifier

PDB Search

blast.pdb( )

PDB

Homologous Structure Set

get.pdb( )

Find & download all related structures

Alignment

pdbaln( )

Align & superpose all structures

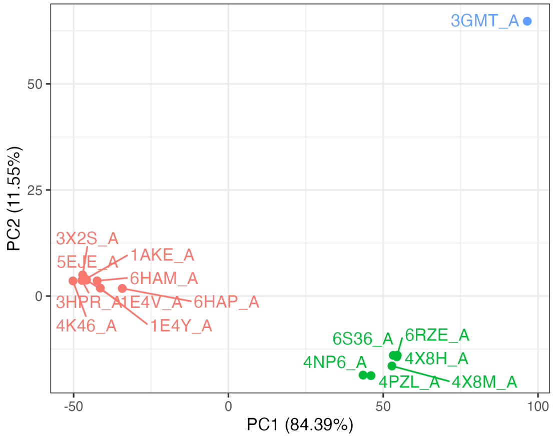
PCA

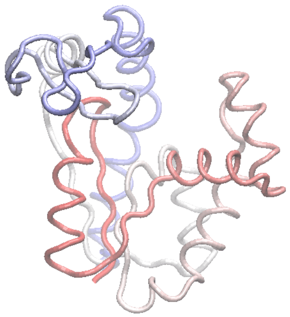
pca( )

plot( )

mktrj( )

PCA Results





Today’s Menu

Overview of structural bioinformatics

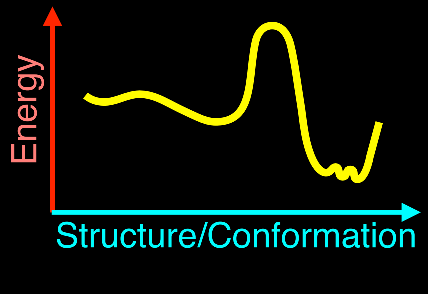
- Motivations, goals and challenges

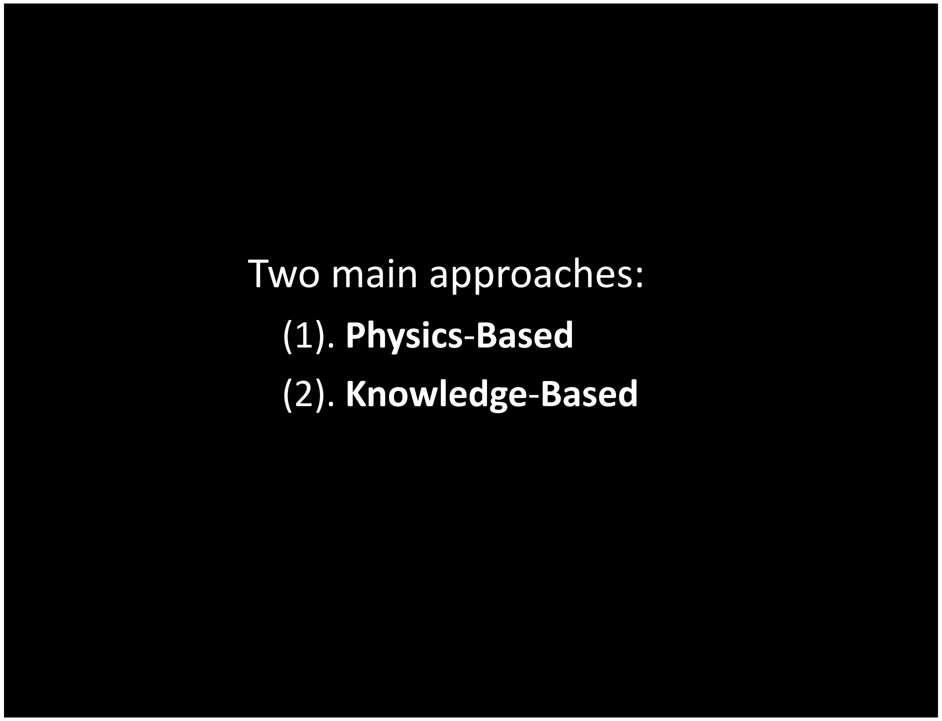
Representing, interpreting & modeling protein structure

- Visualizing & interpreting protein structures
- Analyzing protein structures
- Modeling protein structure
    - Physics based approaches
    - Knowledge based approaches
    - Structure prediction and drug discovery

Key concept:

Potential functions describe a systems energy as a function of its structure

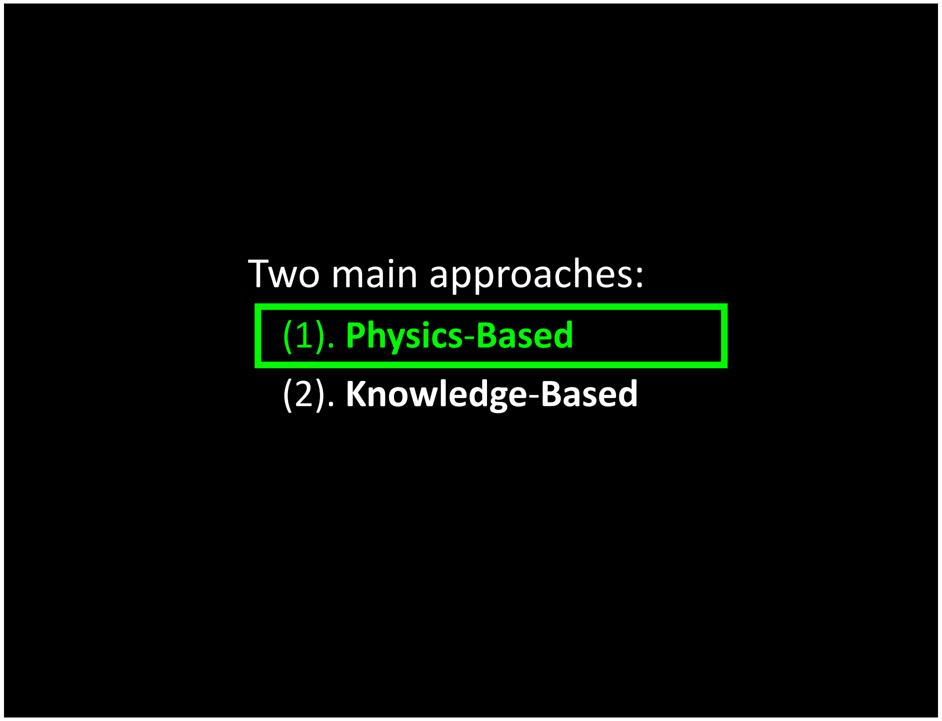




Two main approaches:

(1). **Physics-Based**

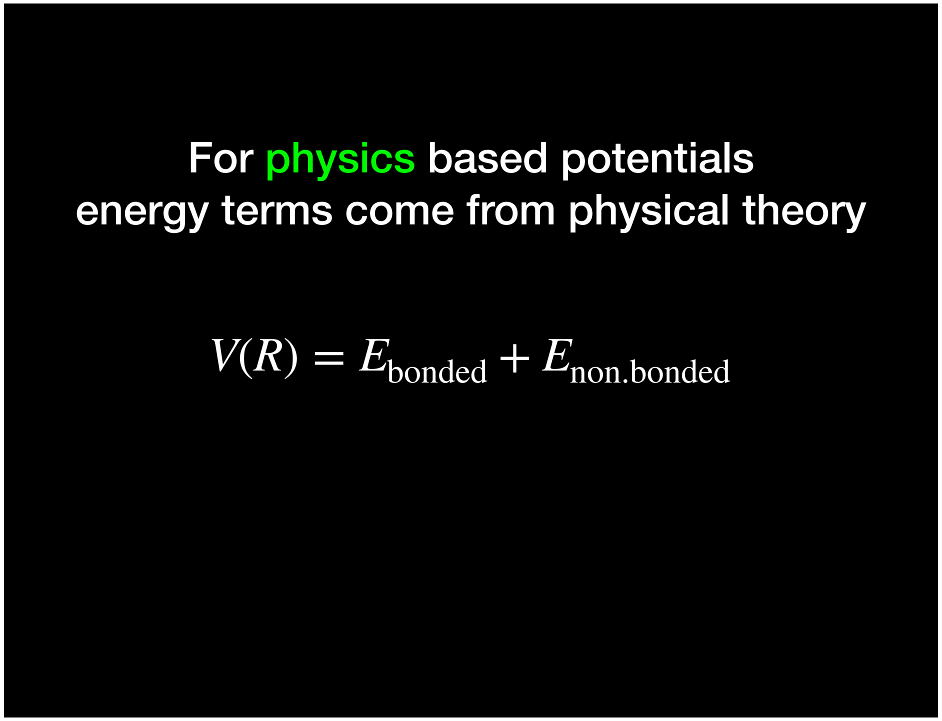
(2). **Knowledge-Based**



Two main approaches:

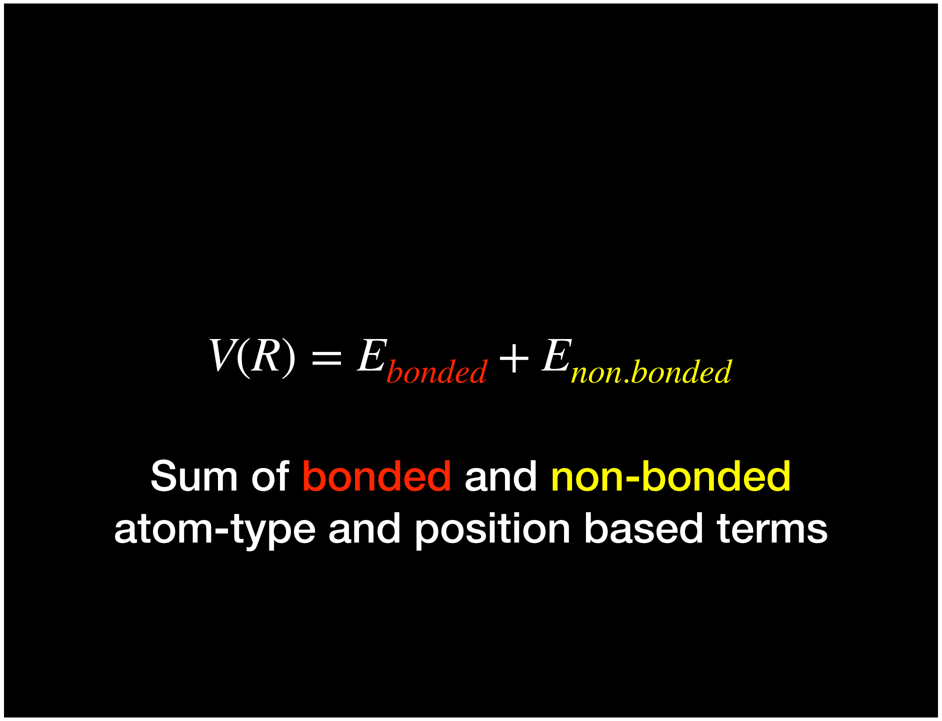
(1). **Physics-Based**

(2). **Knowledge-Based**



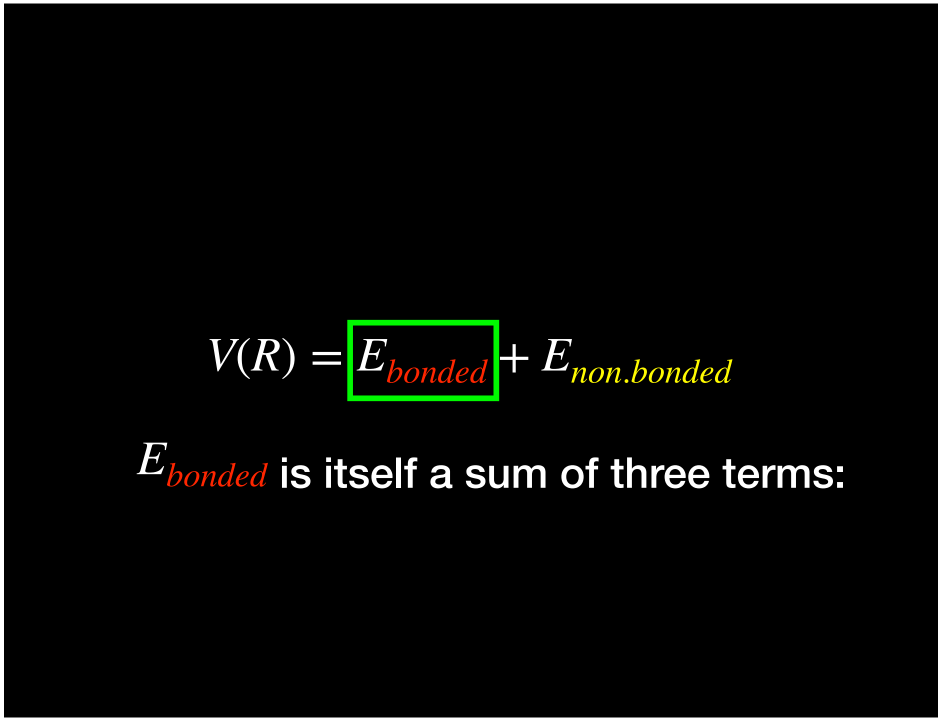
For physics based potentials
energy terms come from physical theory

$$V(R) = E_{\text{bonded}} + E_{\text{non.bonded}}$$



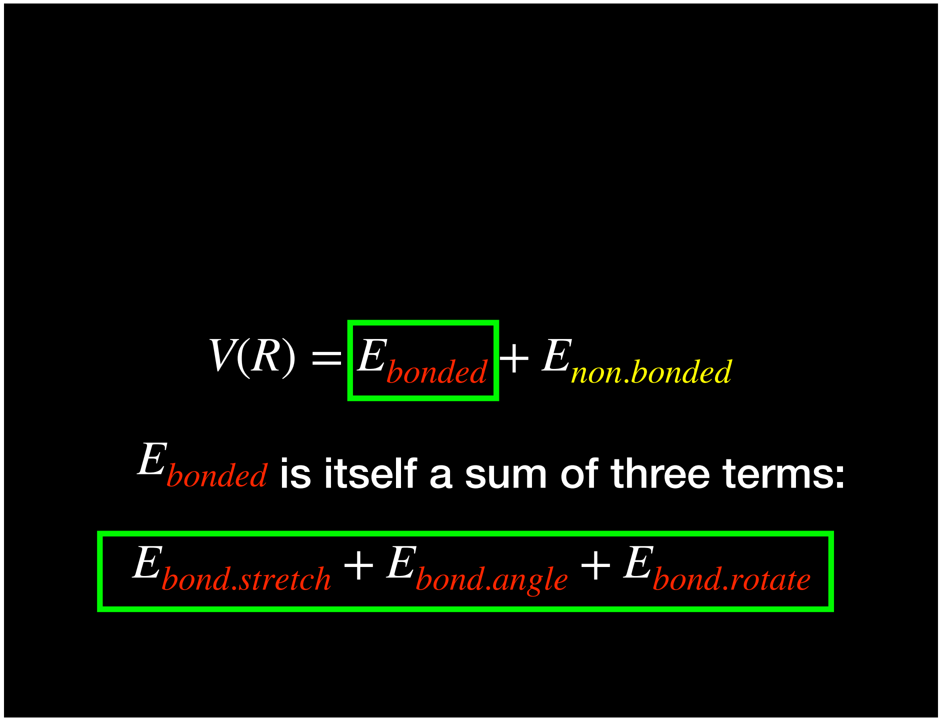
$$V(R) = E_{bonded} + E_{non.bonded}$$

Sum of bonded and non-bonded
atom-type and position based terms



$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

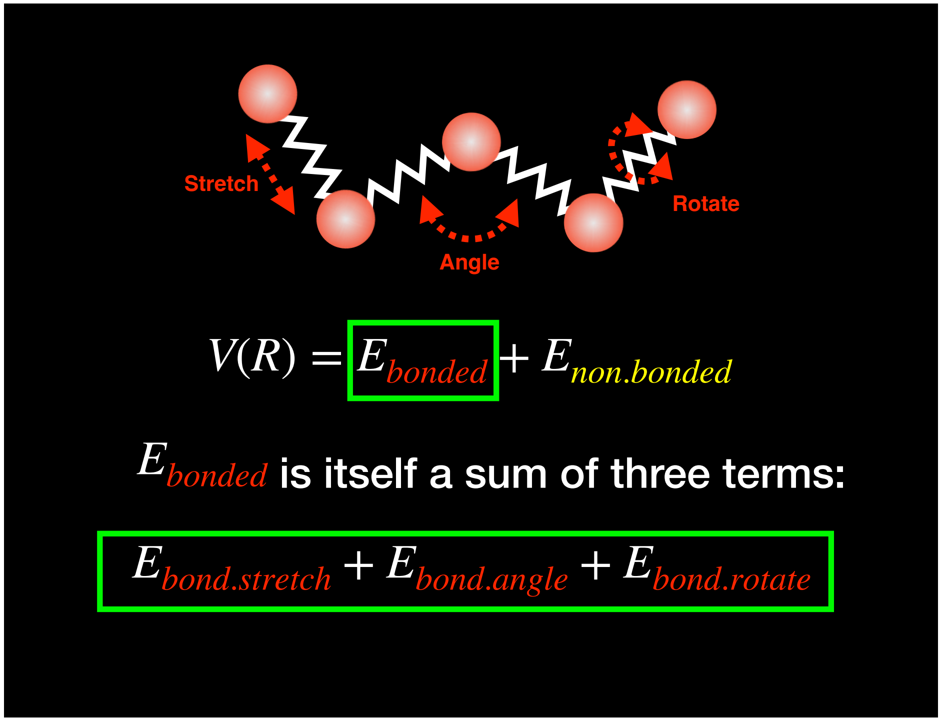
E_{bonded} is itself a sum of three terms:



$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

E_{bonded} is itself a sum of three terms:

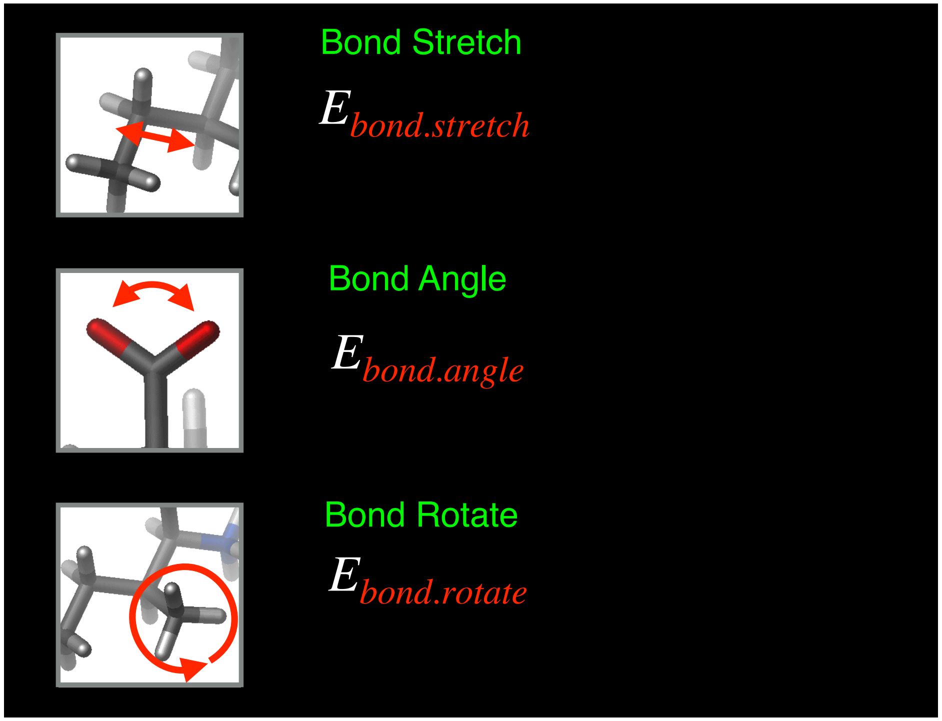
$$\boxed{E_{bond.stretch} + E_{bond.angle} + E_{bond.rotate}}$$



$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

E_{bonded} is itself a sum of three terms:

$$\boxed{E_{bond.stretch} + E_{bond.angle} + E_{bond.rotate}}$$



Bond Stretch

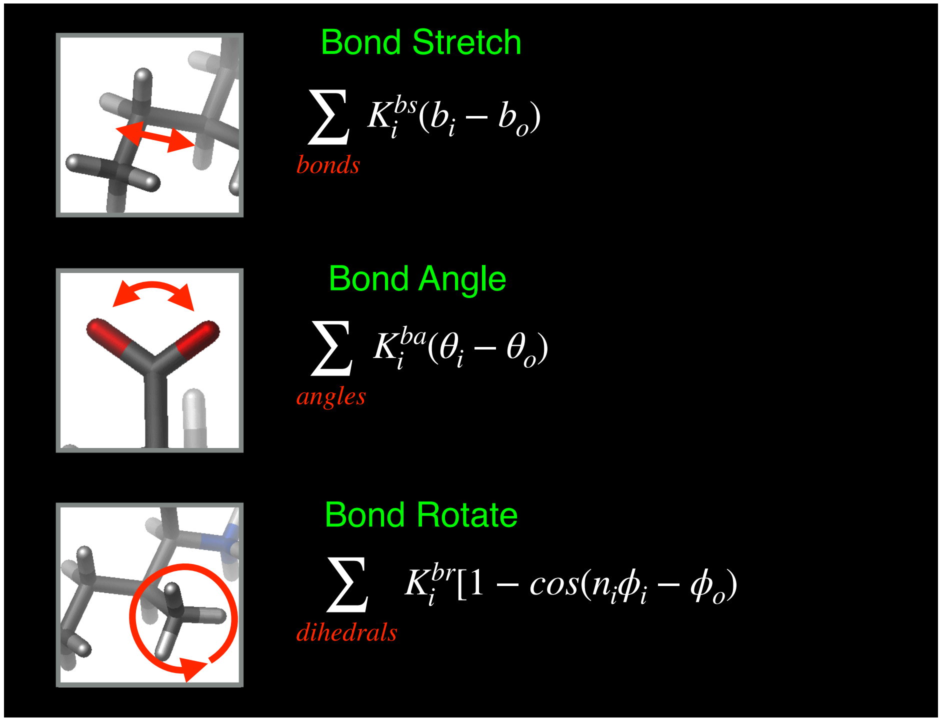
$E_{bond.stretch}$

Bond Angle

$E_{bond.angle}$

Bond Rotate

$E_{bond.rotate}$



Bond Stretch

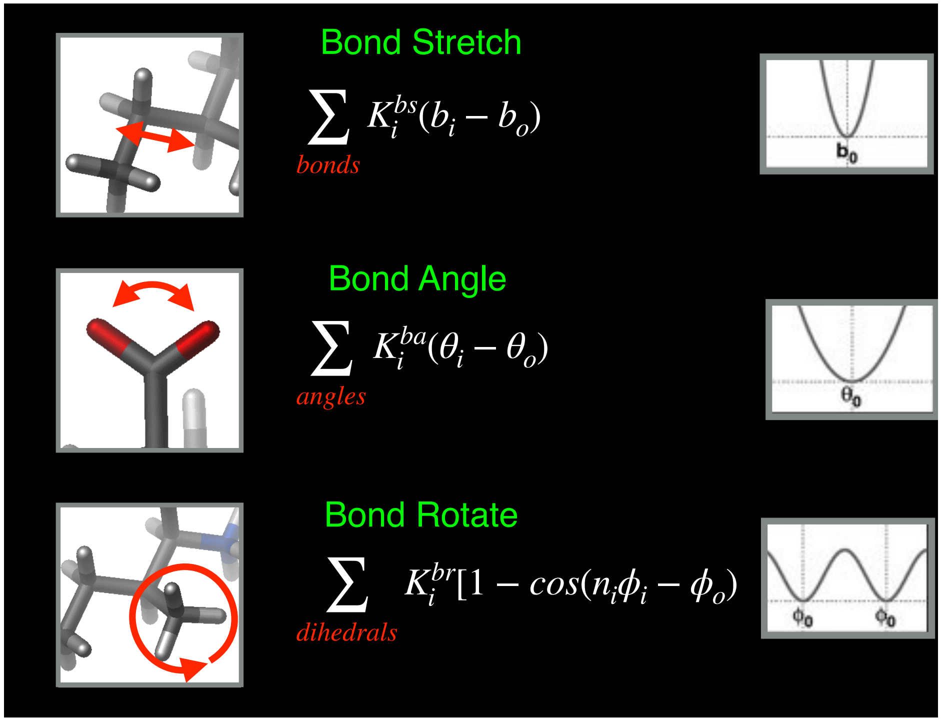
$$\sum_{bonds} K_i^{bs}(b_i - b_o)$$

Bond Angle

$$\sum_{angles} K_i^{ba}(\theta_i - \theta_o)$$

Bond Rotate

$$\sum_{dihedrals} K_i^{br}[1 - cos(n_i\phi_i - \phi_o)]$$



Bond Stretch

$$\sum_{bonds} K_i^{bs}(b_i - b_o)$$

Bond Angle

$$\sum_{angles} K_i^{ba}(\theta_i - \theta_o)$$

Bond Rotate

$$\sum_{dihedrals} K_i^{br}[1 - cos(n_i\phi_i - \phi_o)]$$

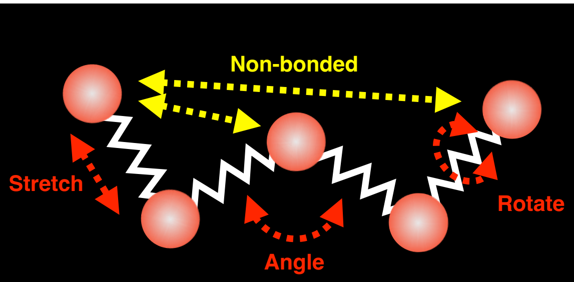
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

$E_{non.bonded}$ is a sum of two terms:

$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

$E_{non.bonded}$ is a sum of two terms:

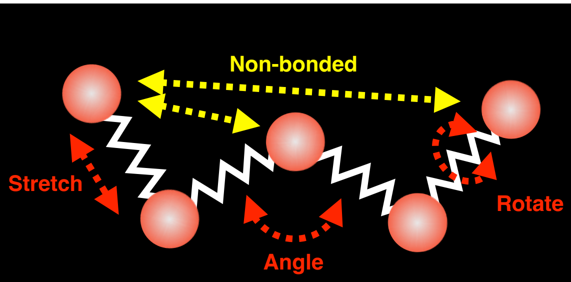
$$E_{van.der.Waals} + E_{electrostatic}$$



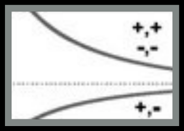
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

$E_{non.bonded}$ is a sum of two terms:

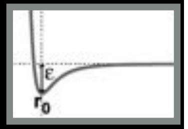
$$E_{van.der.Waals} + E_{electrostatic}$$



$$E_{electrostatic} = \sum_{pairs.i.j} \frac{q_i q_j}{\epsilon r_{ij}^2}$$



$$E_{van.der.Waals} = \sum_{pairs.i.j} \left[\epsilon_{ij}\left(\frac{r_{o.ij}}{r_{ij}}\right)^{12} - 2\epsilon_{ij}\left(\frac{r_{o.ij}}{r_{ij}}\right)^{6}\right]$$



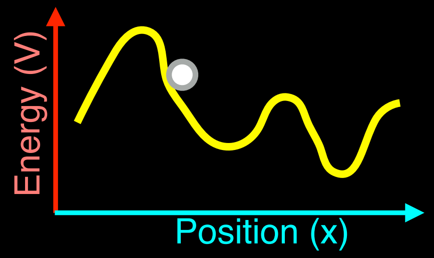
Total potential energy

The potential energy can be given as a sum of terms for: Bond stretching, Bond angles, Bond rotations, van der Walls and Electrostatic interactions between atom pairs

$$\begin{aligned} V(R) = {} & \left.\begin{array}{l} E_{bond.stretch} \\ +E_{bond.angle} \\ +E_{bond.rotate} \end{array}\right\} E_{bonded} \\ & \left.\begin{array}{l} +E_{van.der.Waals} \\ +E_{electrostatic} \end{array}\right\} E_{non.bonded} \end{aligned}$$

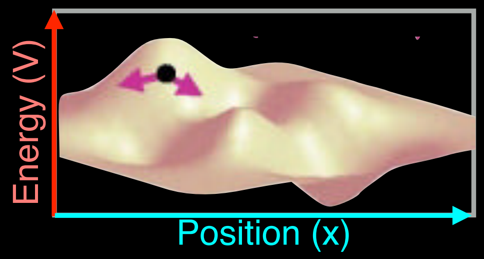
Potential energy surface

Now we can calculate the potential energy surface that fully describes the energy of a molecular system as a function of its geometry



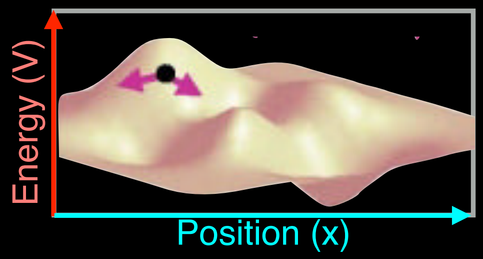
Potential energy surface

Now we can calculate the potential energy surface that fully describes the energy of a molecular system as a function of its geometry



Key concept:

Now we can calculate the potential energy surface that fully describes the energy of a molecular system as a function of its geometry



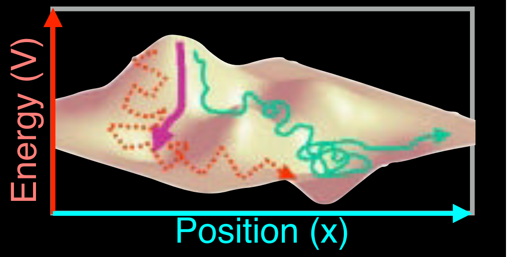
- The **forces** are the gradients of the energy

$F(x) = -dV/dx$

Moving Over The Energy Surface

- **Energy Minimization** drops into local minimum
- **Molecular Dynamics** uses thermal energy to move smoothly over surface
- **Monte Carlo Moves** are random. Accept with probability:

$exp(-\Delta V/dx)$



PHYSICS-ORIENTED APPROACHES

Weaknesses

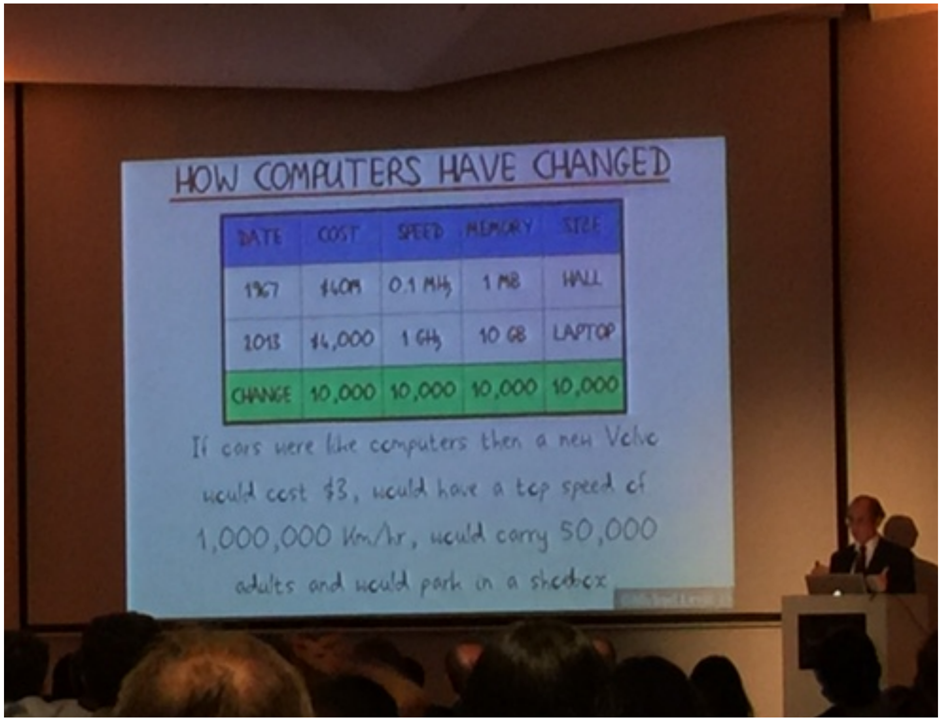
Fully physical detail becomes computationally intractable
Approximations are unavoidable
(Quantum effects approximated classically, water may be treated crudely)
Parameterization still required

Strengths

Interpretable, provides guides to design
Broadly applicable, in principle at least
Clear pathways to improving accuracy

Status

Useful, widely adopted but far from perfect
Multiple groups working on fewer, better approxs
Force fields, quantum
entropy, water effects
Moore's law: hardware improving

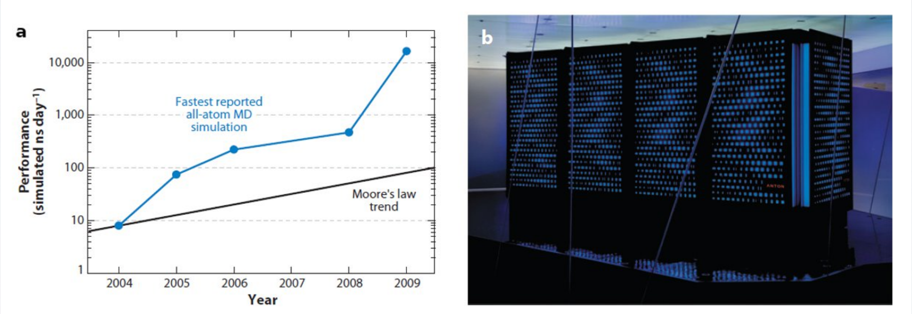


HOW COMPUTERS HAVE CHANGED

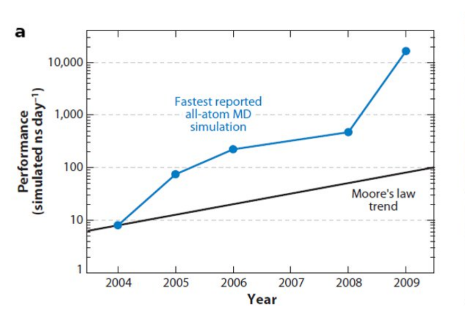
DATE	COST	SPEED	MEMORY	SIZE
1967	$40M	0.1 MHz	1 MB	HALL
2013	$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Volvo would cost $3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shoebox

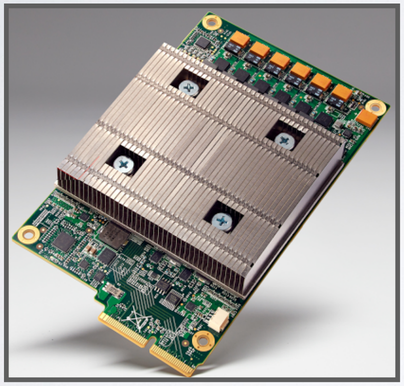
SIDE-NOTE: **ANTON** SUPERCOMPUTER



SIDE-NOTE: **GPUS** (GRAPHICAL PROCESSING UNITS)



SIDE-NOTE: **TPU** (TENSOR PROCESSING UNITS) FOR AI



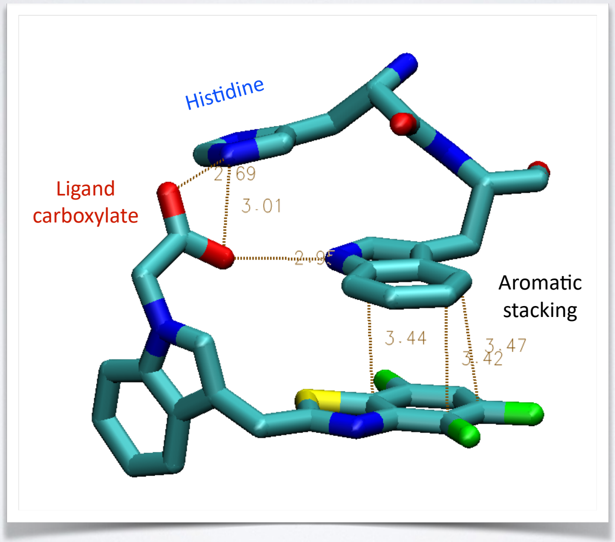
POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). **Physics-Based**

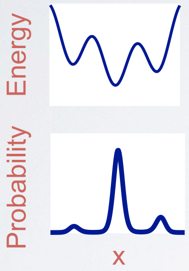
(2). **Knowledge-Based**

KNOWLEDGE-BASED DOCKING POTENTIALS



ENERGY DETERMINES **PROBABILITY** (STABILITY)

Basic idea: Use probability as a proxy for energy



Boltzmann:

$$p(r) \propto e^{-E(r)/RT}$$

Inverse Boltzmann:

$$E(r) = -RT \ln\left[p(r)\right]$$

Example: ligand carboxylate O to protein histidine N

Find all protein-ligand structures in the PDB with a ligand carboxylate O

1. For each structure, histogram the distances from O to every histidine N
2. Sum the histograms over all structures to obtain p($r_{O\text{-}N}$)
3. Compute E($r_{O\text{-}N}$) from p($r_{O\text{-}N}$)

KNOWLEDGE-BASED POTENTIALS

Weaknesses

Accuracy limited by availability of data

Strengths

Relatively easy to implement

Computationally fast

Status

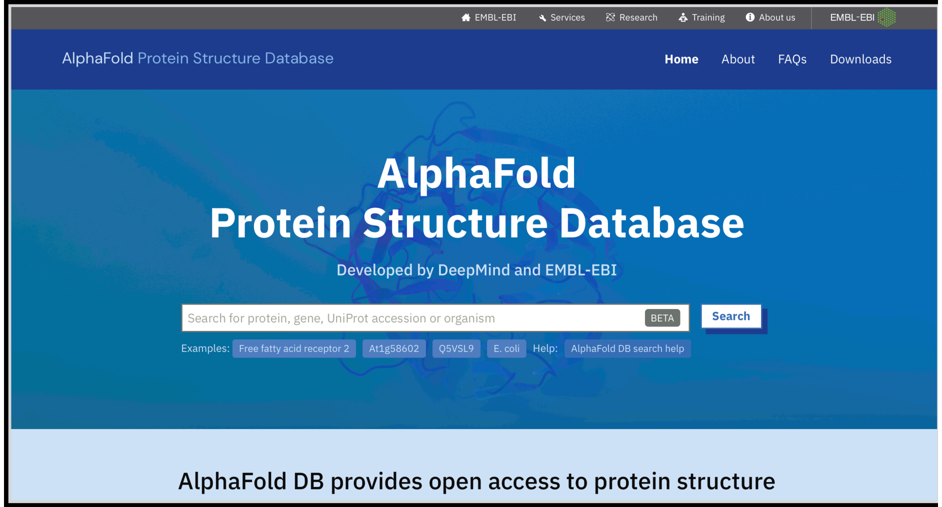
Useful, far from perfect

May be at point of diminishing returns

(not always clear how to make improvements)

- Break -

The future? Combining AI and Physics based approaches



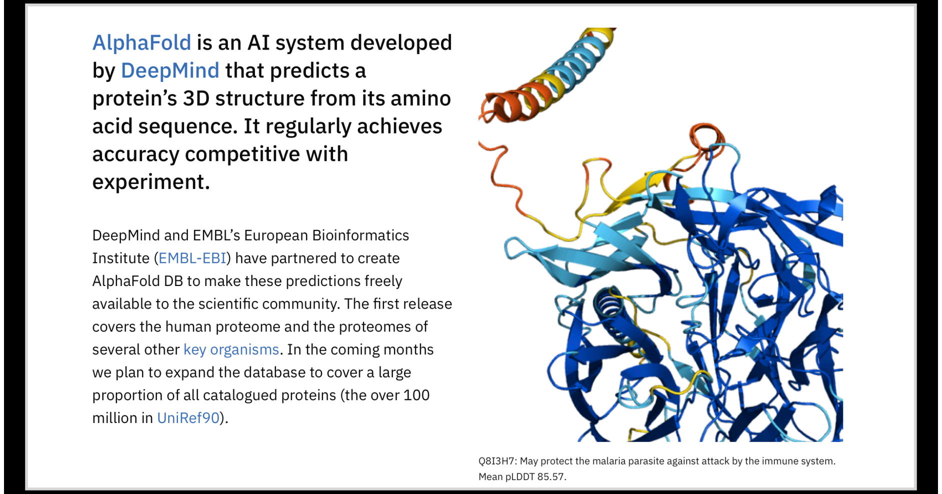
AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA **Search**

Examples: Free fatty acid receptor 2 | At1g58602 | Q5VSL9 | E. coli | Help: AlphaFold DB search help

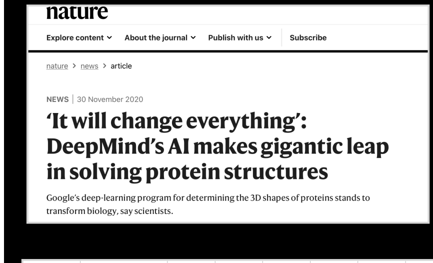
AlphaFold DB provides open access to protein structure



AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The first release covers the human proteome and the proteomes of several other key organisms. In the coming months we plan to expand the database to cover a large proportion of all catalogued proteins (the over 100 million in UniRef90).

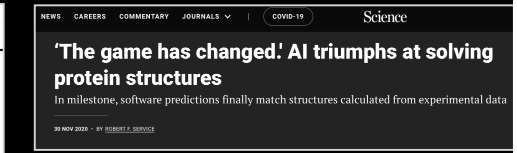
Q8I3H7: May protect the malaria parasite against attack by the immune system.
Mean pLDDT 85.57.



NEWS | 30 November 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

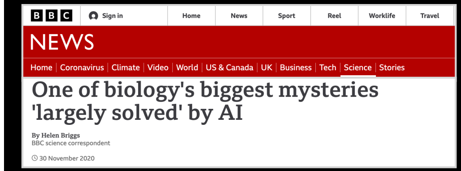
Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.



'The game has changed.' AI triumphs at solving protein structures

In milestone, software predictions finally match structures calculated from experimental data

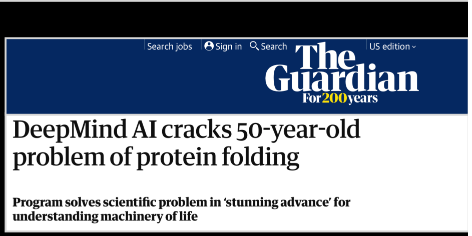
30 NOV 2020 · BY ROBERT F. SERVICE



One of biology's biggest mysteries 'largely solved' by AI

By Helen Briggs
BBC science correspondent

30 November 2020

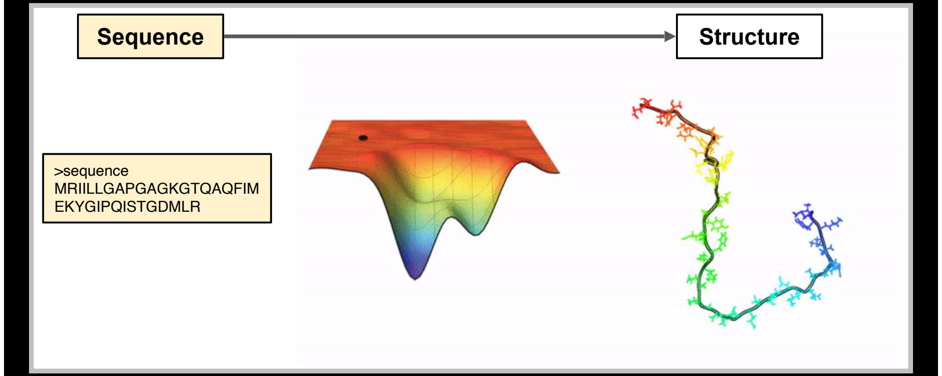


DeepMind AI cracks 50-year-old problem of protein folding

Program solves scientific problem in 'stunning advance' for understanding machinery of life

Protein Folding Problem

For a given sequence, find structure with lowest free energy

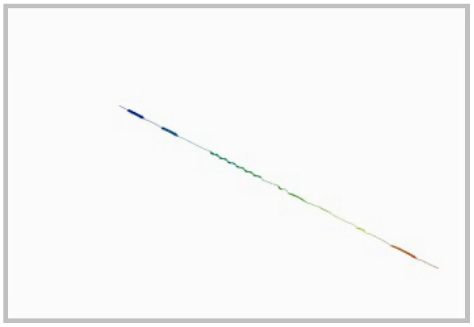


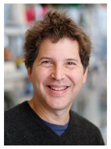
>sequence
MRIILLGAPGAGKGTQAQFIM
EKYGIPQISTGDMLR

[Video credit: C. Fennell]

Dill, K.A. and MacCallum, J.L., 2012. The protein-folding problem, 50 years on. *science*, *338*(6110), pp.1042-1046.

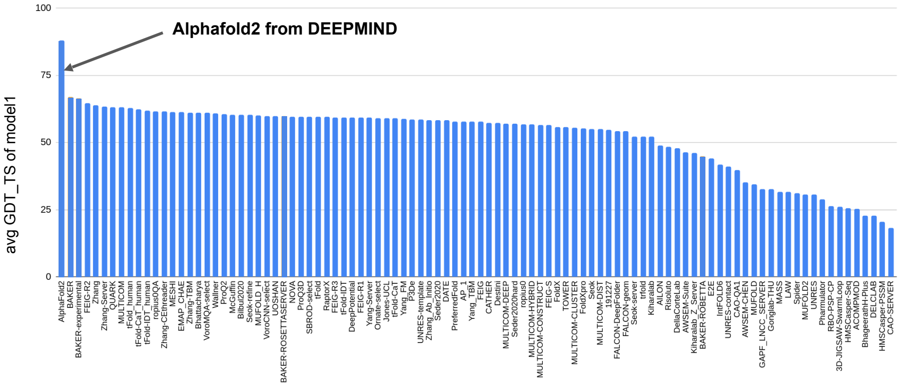
Rosetta - Protein "folding" with Energy function + fragments recombination





David
Baker

Results from CASP14
(Critical Assessment of protein Structure Prediction)

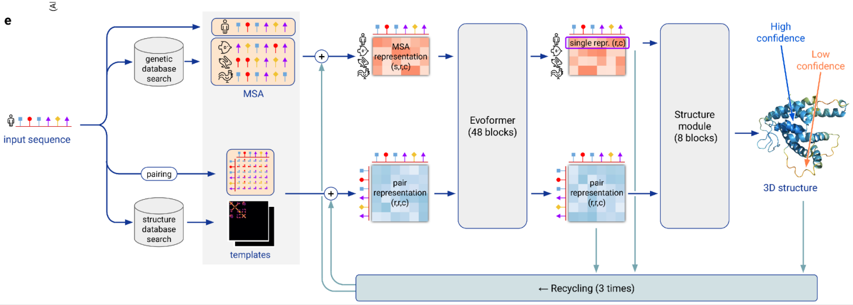


nature

https://doi.org/10.1038/s41586-021-03819-2

Accelerated Article Preview

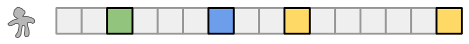
Highly accurate protein structure prediction with AlphaFold



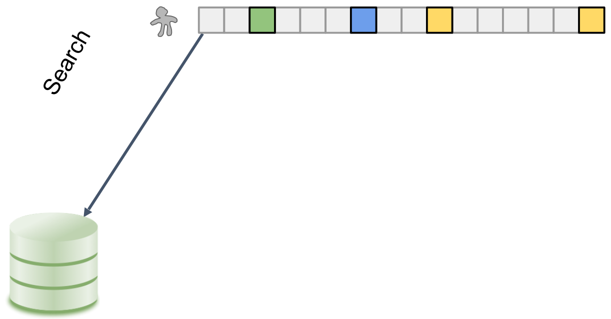
John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain,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

Multiple Sequence Alignments (MSAs)
are key inputs to these winning methods
(alphafold2 and RoseTTAFold)

Start with a single sequence

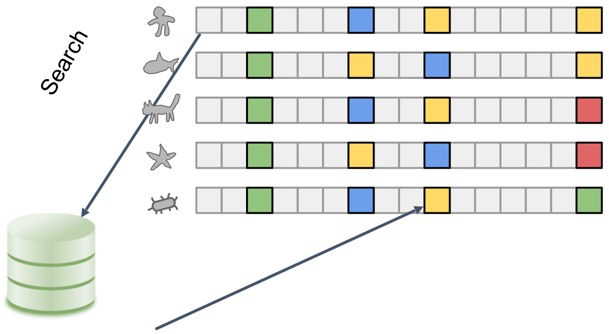


Search against a database of sequences



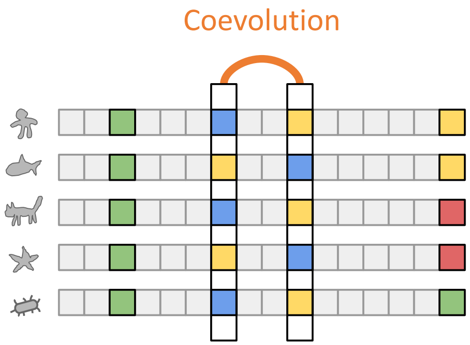
Database of
sequences

Generate a multiple sequence alignment

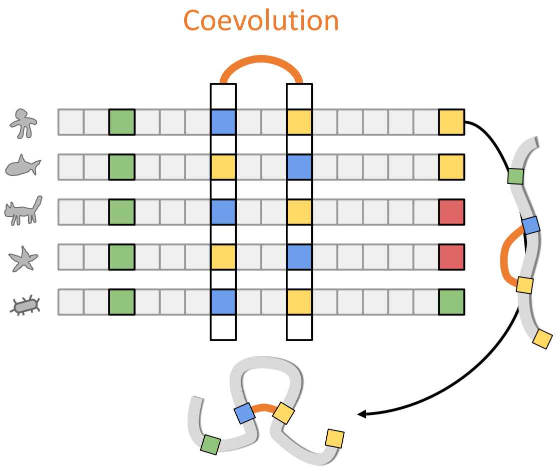


Database of
sequences

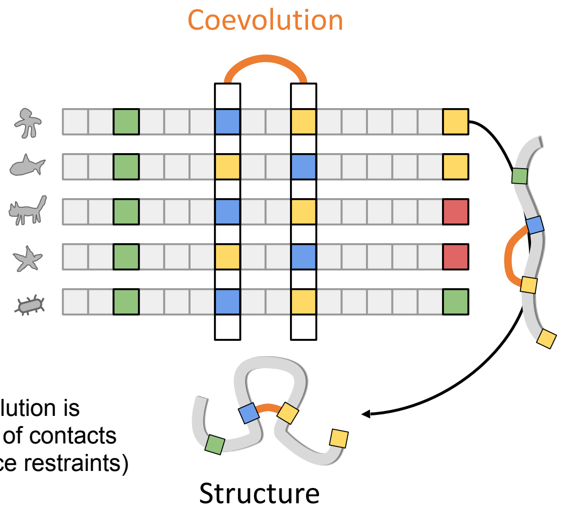
Analyze the MSA for coevolution



Use coevolution as restraints in folding simulations!



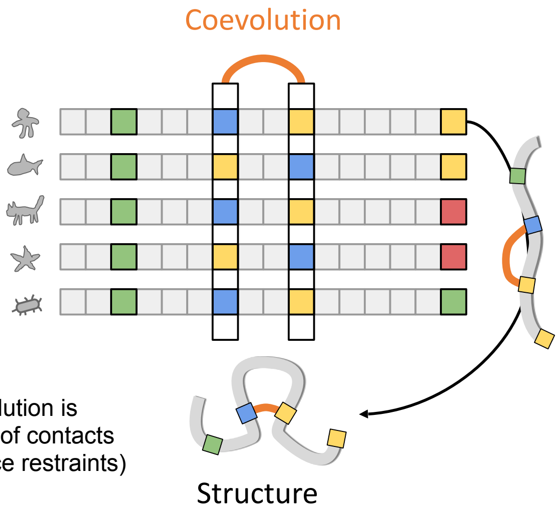
Use coevolution as restraints in folding simulations!



Coevolution is predictive of contacts (i.e. distance restraints)

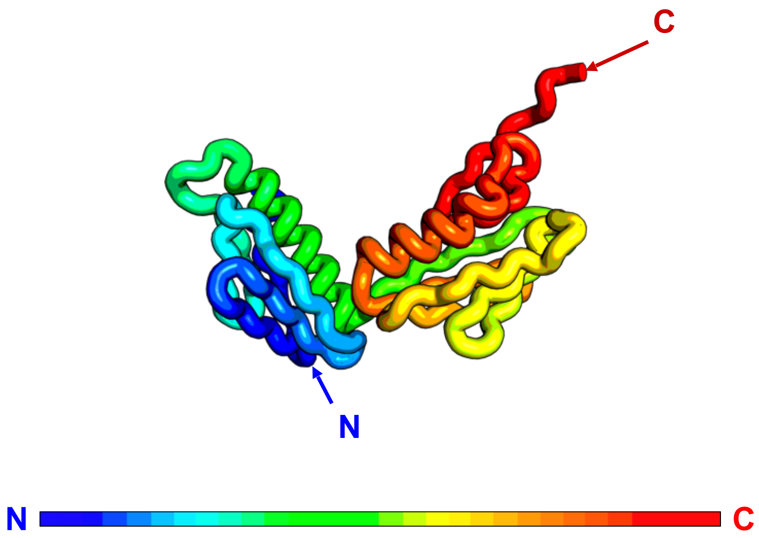
By measuring coevolution, we can infer contacts!

Contacts in proteins are evolutionarily conserved and encoded in a **MSA** due to **coevolution**

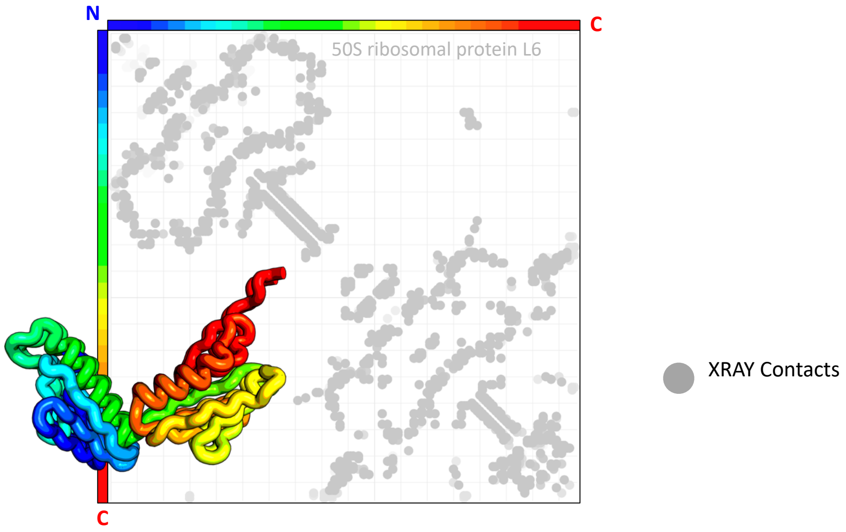


Coevolution is predictive of contacts (i.e. distance restraints)

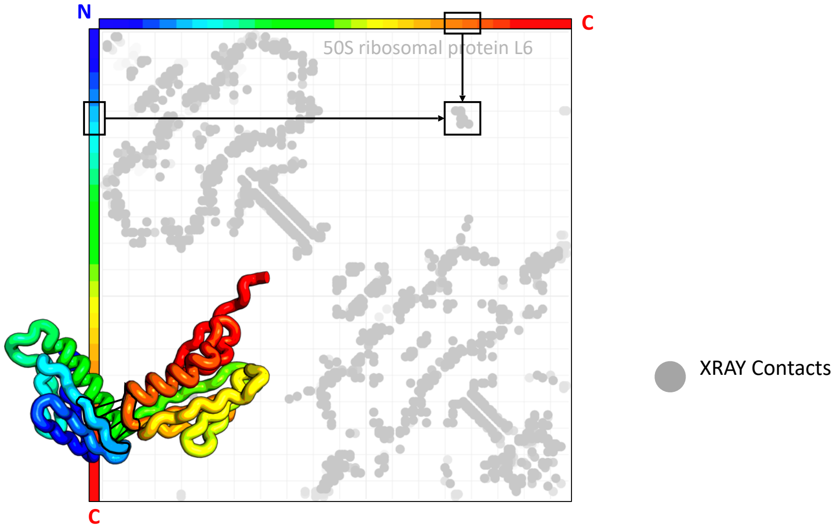
Review - How to read a contact/distance matrix?



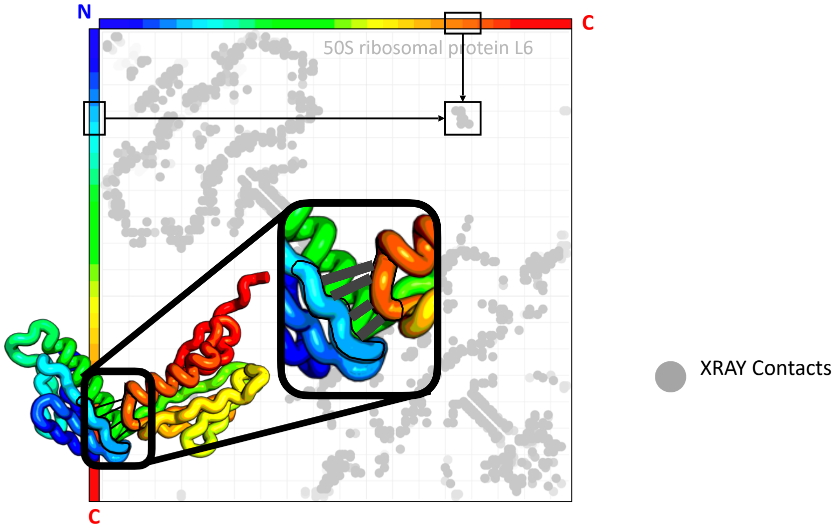
Contact map



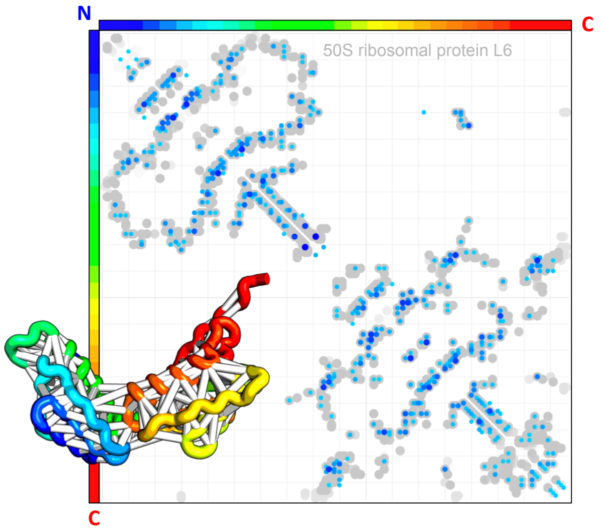
How to read a contact map



How to read a contact map



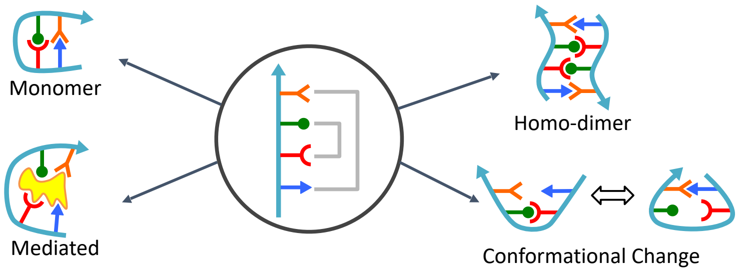
Overlay of predicted contacts on real contacts



● XRAY Contacts

● Predicted Contacts

The origin of contacts

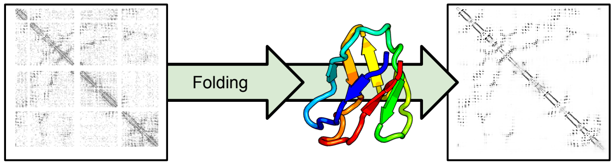


Anishchenko, I., **Ovchinnikov, S.**, Kamisetty, H. and Baker, D., 2017. Origins of coevolution between residues distant in protein 3D structures. *PNAS*, *114*(34), pp.9122-9127.

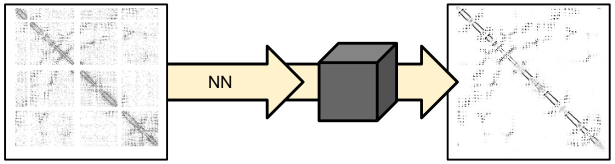
Slide Credit: Sergey Ovchinnikov (@sokrypton)

How to solve this problem?

- Enumerate folds and see which matches contacts best
- Try different number (or combination) of restraints
- Lots of sampling with ambiguous restraints

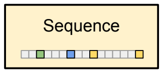


- Use NN to filter/enhance contacts before trying to fold

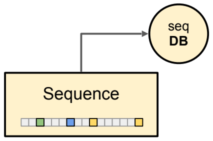


citations: bit.ly/3Mr8351

Alphafold2



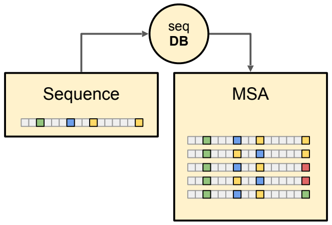
Alphafold2



seq DB

Sequence

Alphafold2

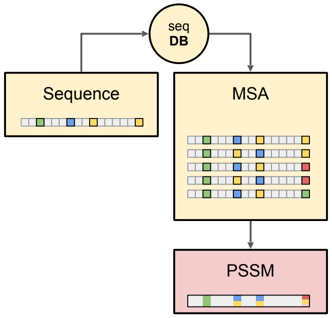


seq DB

Sequence

MSA

Alphafold2



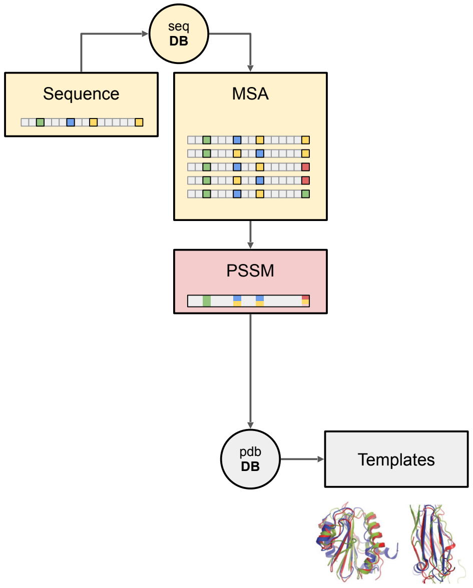
seq DB

Sequence

MSA

PSSM

Alphafold2



seq DB

Sequence

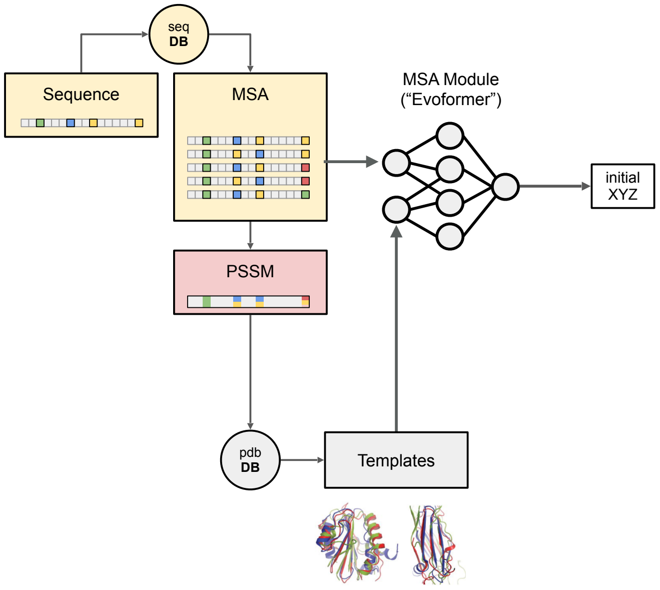
MSA

PSSM

pdb DB

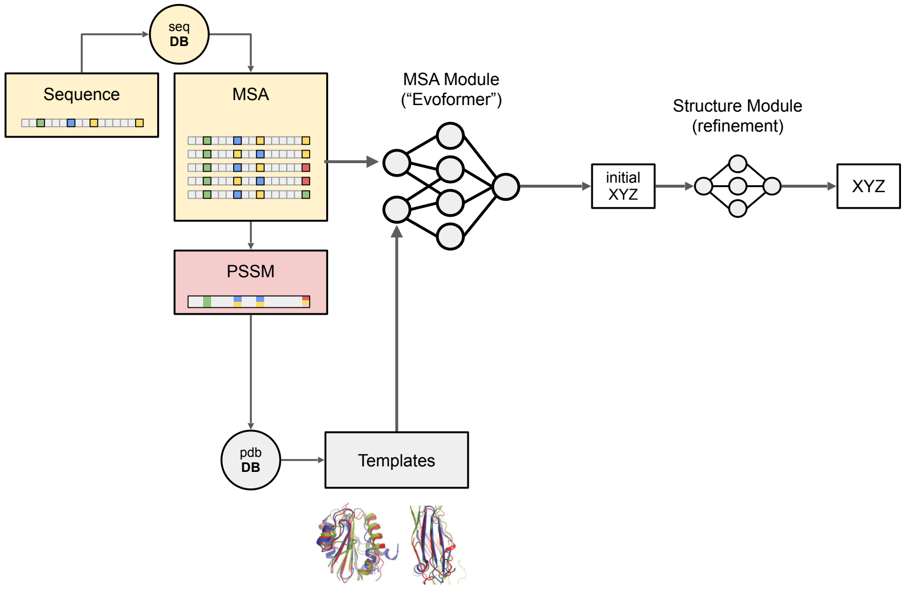
Templates

Alphafold2



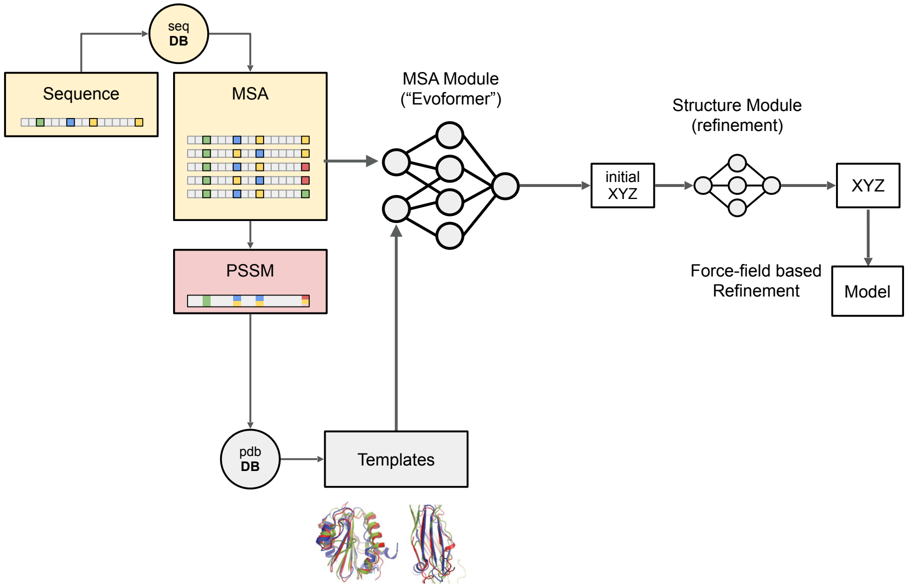
seq DB
Sequence
MSA
PSSM
pdb DB
Templates
MSA Module
(“Evoformer”)
initial XYZ

Alphafold2



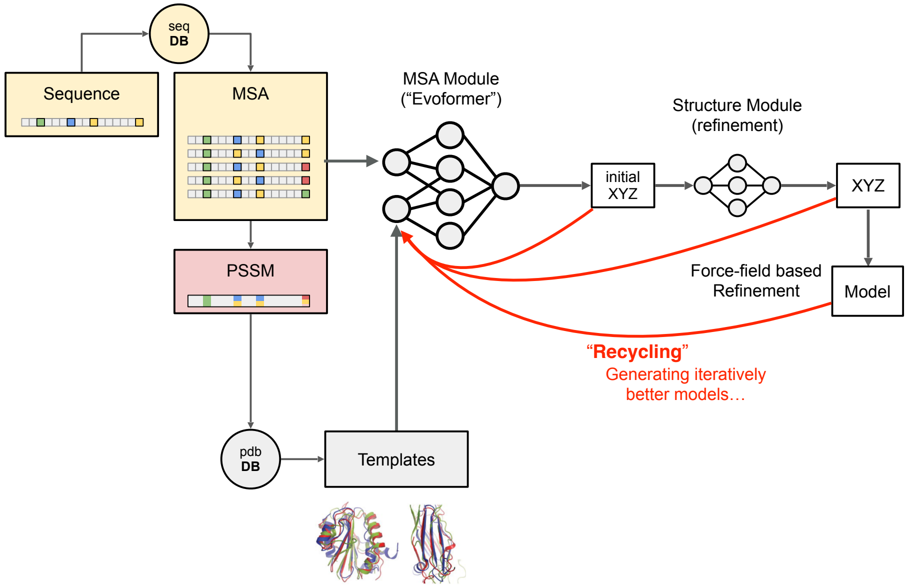
seq DB
Sequence
MSA
PSSM
pdb DB
Templates
MSA Module
(“Evoformer”)
initial XYZ
Structure Module
(refinement)
XYZ

Alphafold2

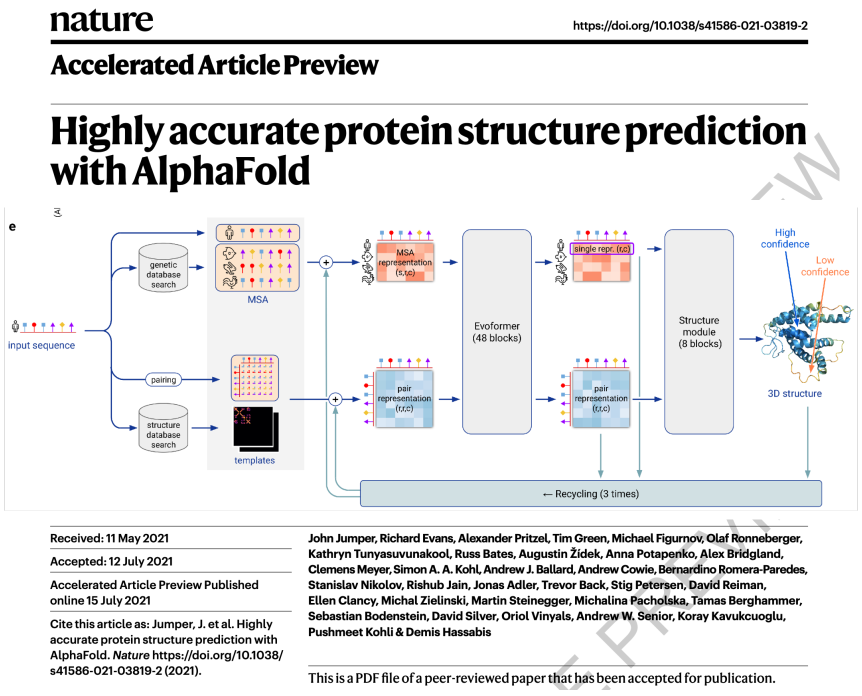


seq DB
Sequence
MSA
PSSM
pdb DB
Templates
MSA Module
(“Evoformer”)
initial XYZ
Structure Module
(refinement)
XYZ
Force-field based
Refinement
Model

Alphafold2 - New Critical detail Recycling



seq DB
Sequence
MSA
PSSM
pdb DB
Templates
MSA Module
(“Evoformer”)
initial XYZ
Structure Module
(refinement)
XYZ
Force-field based
Refinement
Model
“Recycling”
Generating iteratively
better models…



nature https://doi.org/10.1038/s41586-021-03819-2

Accelerated Article Preview

Highly accurate protein structure prediction with AlphaFold

Received: 11 May 2021

Accepted: 12 July 2021

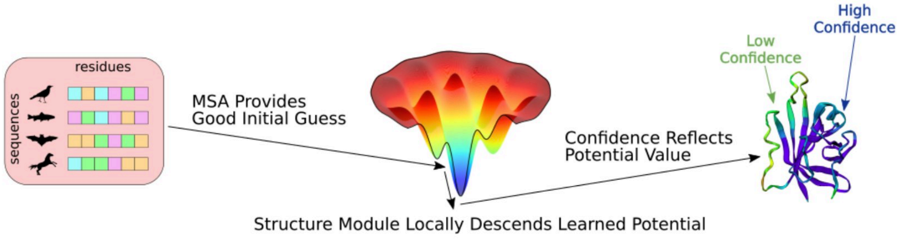
Accelerated Article Preview Published online 15 July 2021

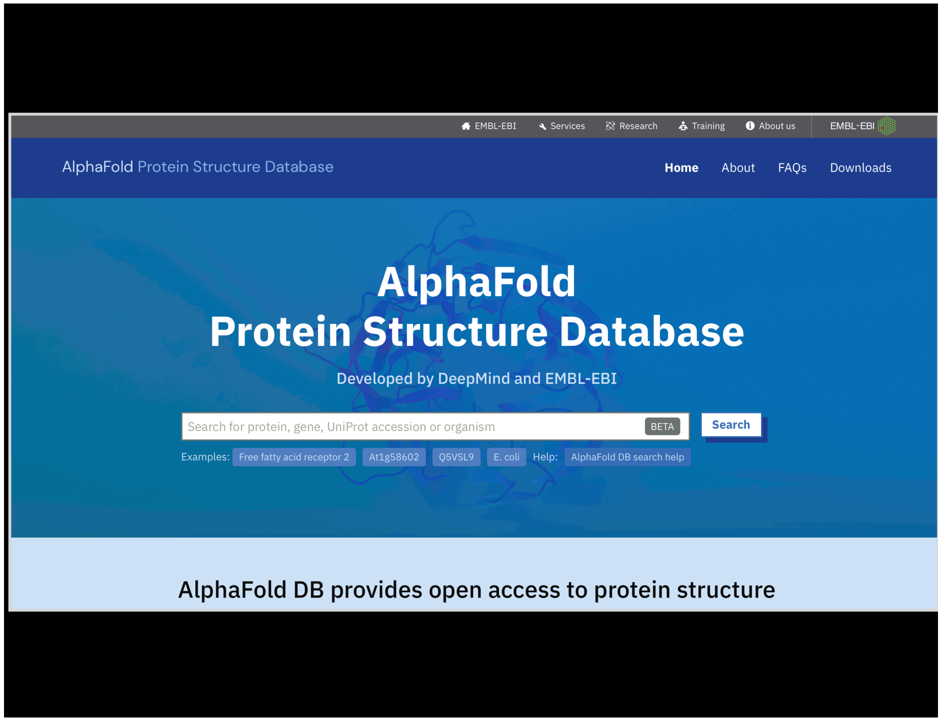
Cite this article as: Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).

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

This is a PDF file of a peer-reviewed paper that has been accepted for publication.

Hypothesis:
AlphaFold uses input MSA/Templates to "solve" the global search problem. The rest of the model refines the structure using the learned energy potential.



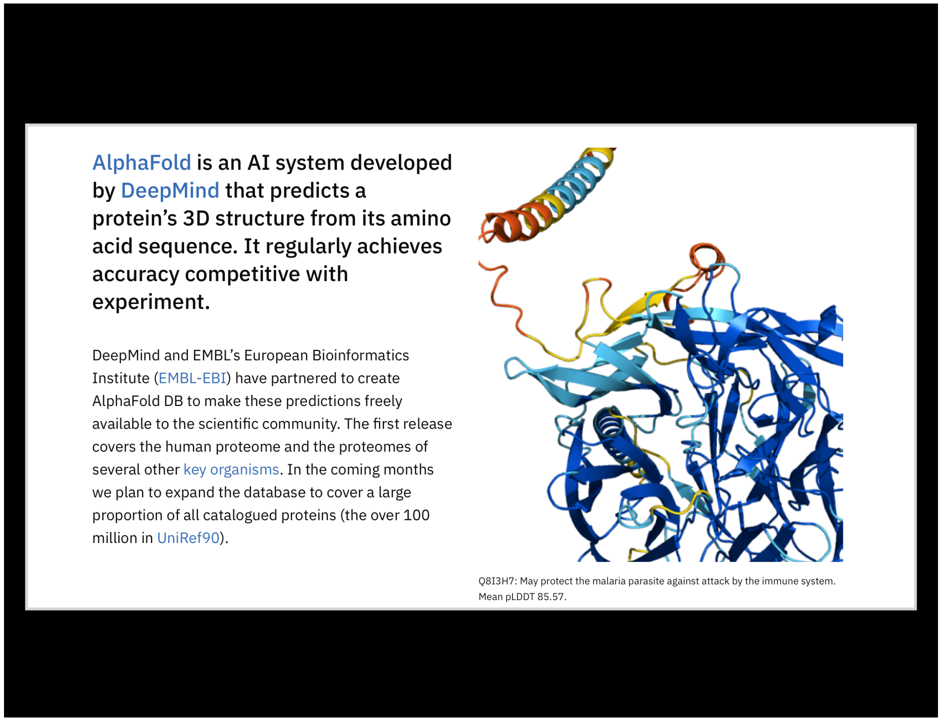


AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

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

AlphaFold DB provides open access to protein structure



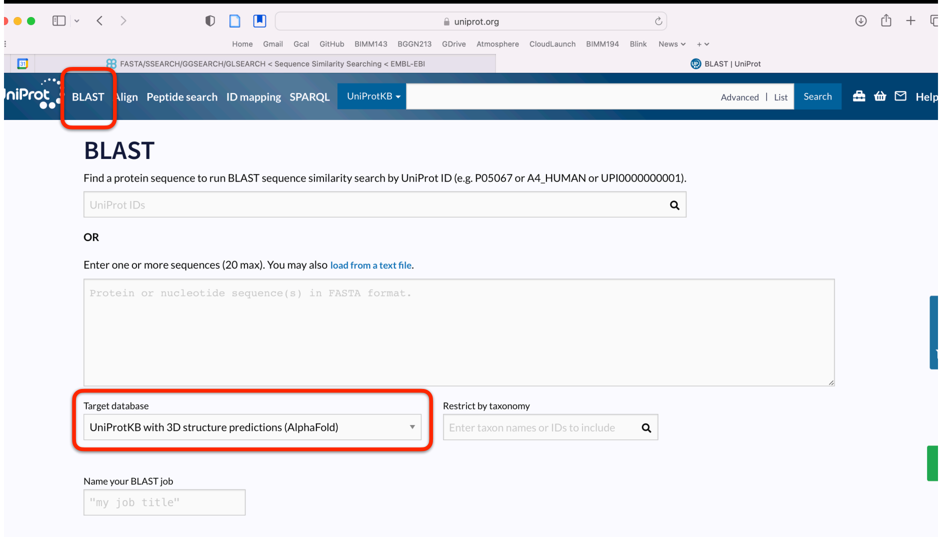
AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The first release covers the human proteome and the proteomes of several other key organisms. In the coming months we plan to expand the database to cover a large proportion of all catalogued proteins (the over 100 million in UniRef90).

Q8I3H7: May protect the malaria parasite against attack by the immune system.
Mean pLDDT 85.57.

Seach UniProt with AlphaFold

https://www.uniprot.org/blast


uniprot.org
Home Gmail Gcal GitHub BIMM143 BGGN213 GDrive Atmosphere CloudLaunch BIMM194 Blink News
FASTA/SSEARCH/GGSEARCH/GLSEARCH < Sequence Similarity Searching < EMBL-EBI
BLAST | UniProt
BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search Help

BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001).

UniProt IDs

OR

Enter one or more sequences (20 max). You may also load from a text file.

Protein or nucleotide sequence(s) in FASTA format.

Target database: UniProtKB with 3D structure predictions (AlphaFold)

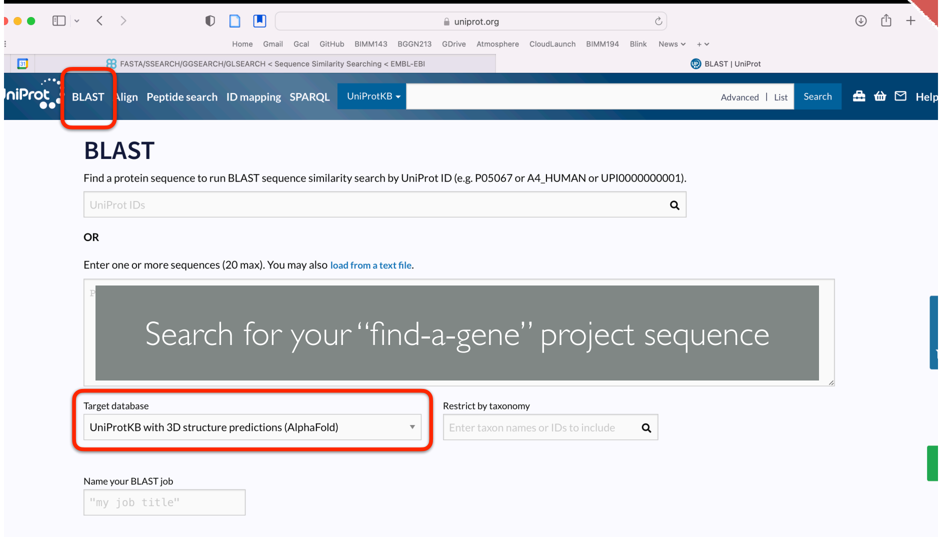
Restrict by taxonomy: Enter taxon names or IDs to include

Name your BLAST job: "my job title"

Seach UniProt with AlphaFold

Do it Yourself!

https://www.uniprot.org/blast


uniprot.org
Home Gmail Gcal GitHub BIMM143 BGGN213 GDrive Atmosphere CloudLaunch BIMM194 Blink News
FASTA/SSEARCH/GGSEARCH/GLSEARCH < Sequence Similarity Searching < EMBL-EBI
BLAST | UniProt
BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search Help

BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001).

UniProt IDs

OR

Enter one or more sequences (20 max). You may also load from a text file.

Search for your "find-a-gene" project sequence

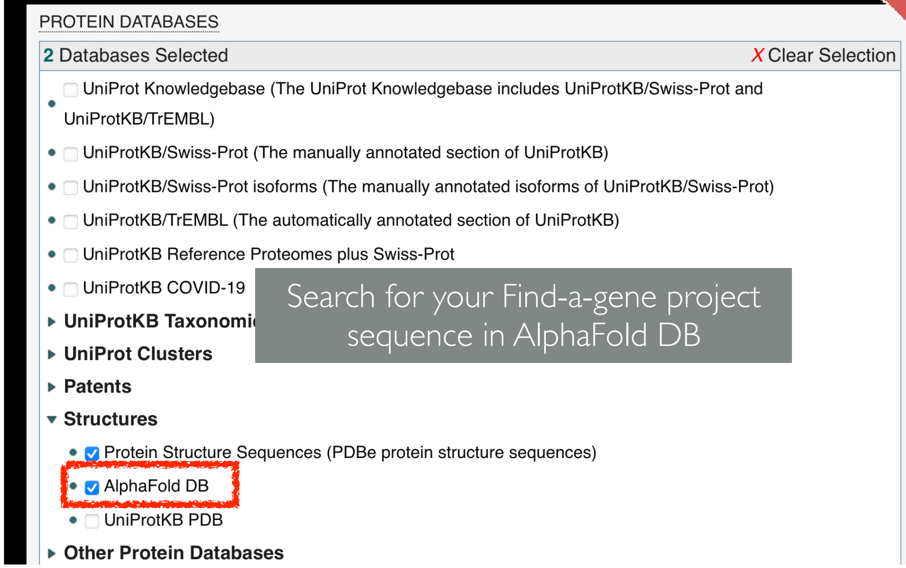
Target database: UniProtKB with 3D structure predictions (AlphaFold)

Restrict by taxonomy: Enter taxon names or IDs to include

Name your BLAST job: "my job title"

Do it Yourself!

https://www.ebi.ac.uk/Tools/sss/fasta/



PROTEIN DATABASES

2 Databases Selected X Clear Selection

- □ UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- □ UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- □ UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- □ UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- □ UniProtKB Reference Proteomes plus Swiss-Prot
- □ UniProtKB COVID-19

UniProtKB Taxonomi

UniProt Clusters

Patents

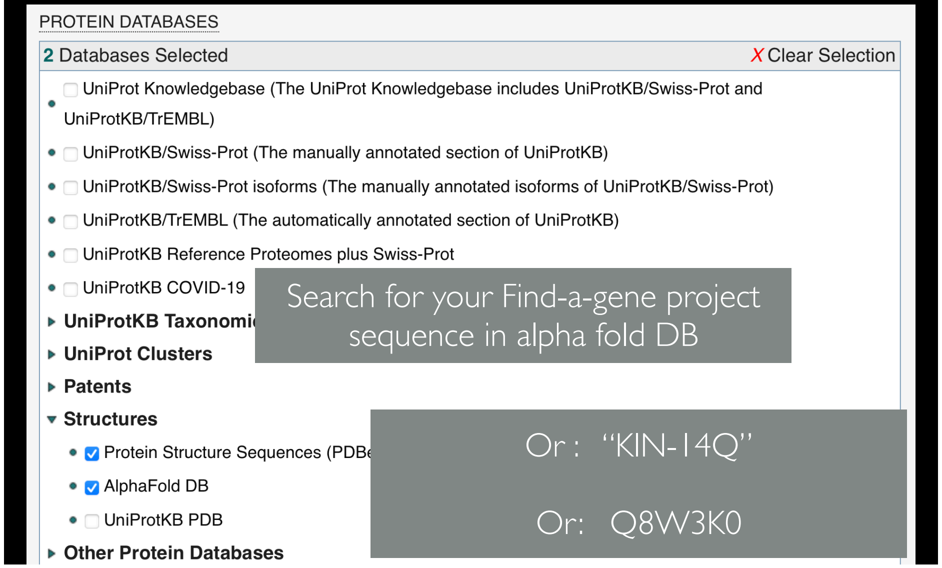
Structures

- ☑ Protein Structure Sequences (PDBe protein structure sequences)
- ☑ AlphaFold DB
- □ UniProtKB PDB

Other Protein Databases

Search for your Find-a-gene project sequence in AlphaFold DB

https://www.ebi.ac.uk/Tools/sss/fasta/



PROTEIN DATABASES

2 Databases Selected X Clear Selection

- □ UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- □ UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- □ UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- □ UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- □ UniProtKB Reference Proteomes plus Swiss-Prot
- □ UniProtKB COVID-19

UniProtKB Taxonomi

UniProt Clusters

Patents

Structures

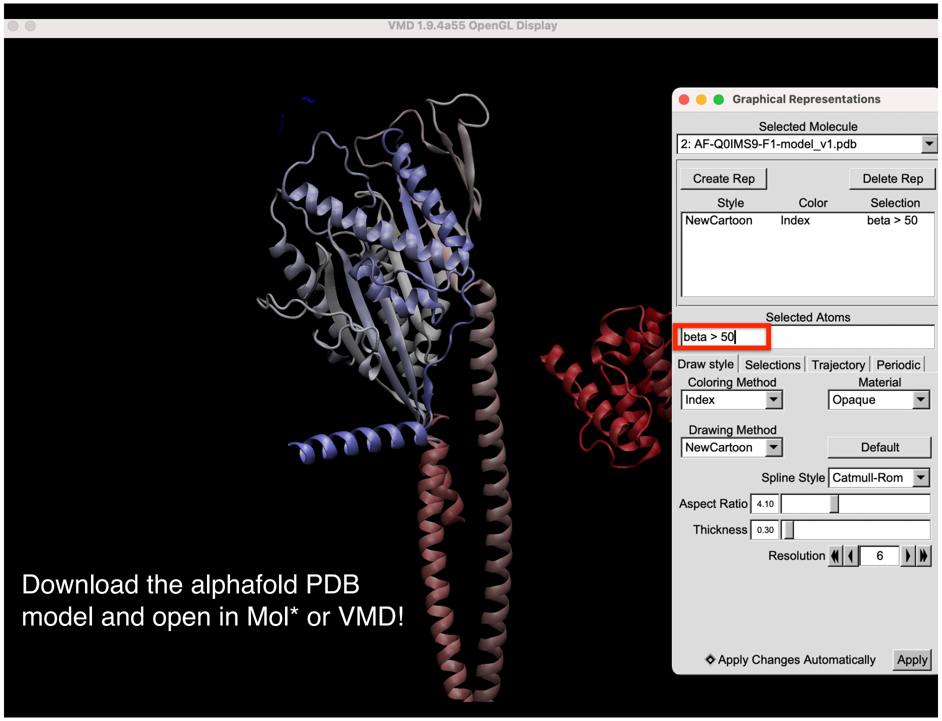
- ☑ Protein Structure Sequences (PDBe
- ☑ AlphaFold DB
- □ UniProtKB PDB

Other Protein Databases

Search for your Find-a-gene project sequence in alpha fold DB

Or : "KIN-14Q"

Or: Q8W3K0


VMD 1.9.4a55 OpenGL Display
Graphical Representations
Create Rep
Delete Rep
Draw style
Selections
Trajectory
Periodic
Default
◇ Apply Changes Automatically
Apply

Selected Molecule: 2: AF-Q0IMS9-F1-model_v1.pdb

Style	Color	Selection
NewCartoon	Index	beta > 50

Selected Atoms: beta > 50

Coloring Method: Index

Material: Opaque

Drawing Method: NewCartoon

Spline Style: Catmull-Rom

Aspect Ratio: 4.10

Thickness: 0.30

Resolution: 6

Download the alphafold PDB model and open in Mol* or VMD!

AlphaFold low confidence regions

- AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100 that is written to the B-factor column.
- To remove low confidence regions (with low pLDDT scores)

```
p <- read.pdb("AF-model.pdb")

# Find atoms with good confidence score (pLDDT)
atoms <- which( p$atom$b > 70 )

# Trim to selected atoms
p2 <- trim.pdb(p, as.select( atoms ) )
write.pdb(p2, file="high_confidence_model.pdb")
```

https://github.com/sokrypton/ColabFold

Evolutionary scale modeling (ESM)

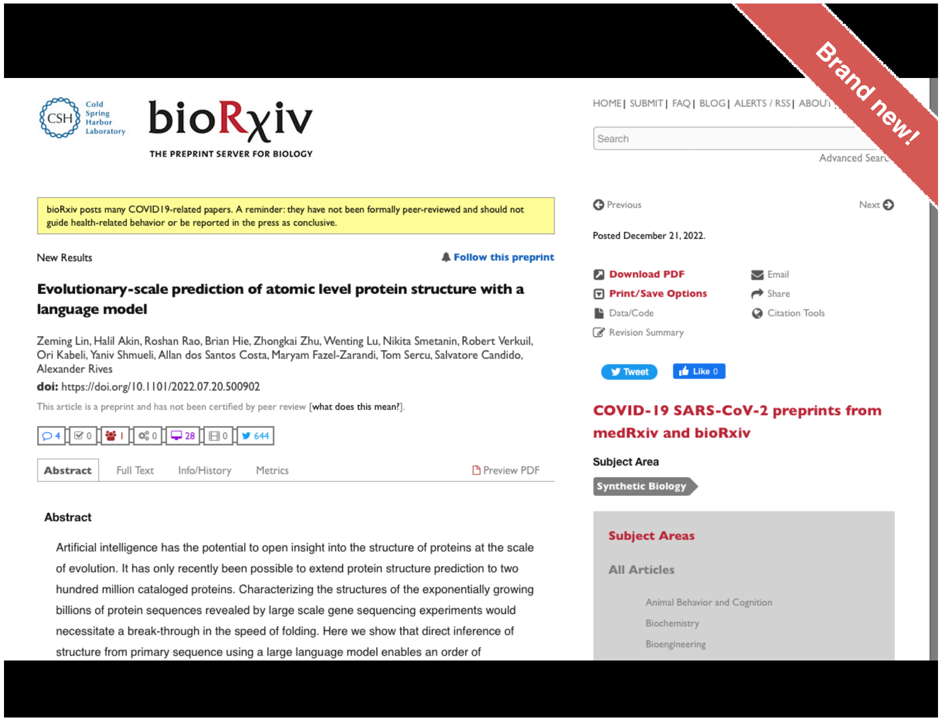
For short monomeric proteins (< 400 amino acids) consider using the new ESMFold

https://esmatlas.com/

[No need for GPU & comparatively fast]

Alternative: Language Models

- AlphaFold (and related methods) need to search through large protein databases to identify related sequences.
- They require a large group of evolutionarily related sequences as input so that they can extract the patterns that are linked to structure.
- ESM-fold uses a language model that learns these evolutionary patterns during its training on protein sequences, enabling faster structure prediction from a single sequence.



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Evolutionary-scale prediction of atomic level protein structure with a language model

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, Alexander Rives

doi: https://doi.org/10.1101/2022.07.20.500902

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

Artificial intelligence has the potential to open insight into the structure of proteins at the scale of evolution. It has only recently been possible to extend protein structure prediction to two hundred million cataloged proteins. Characterizing the structures of the exponentially growing billions of protein sequences revealed by large scale gene sequencing experiments would necessitate a break-through in the speed of folding. Here we show that direct inference of structure from primary sequence using a large language model enables an order of

Posted December 21, 2022.

COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

Subject Area

Synthetic Biology

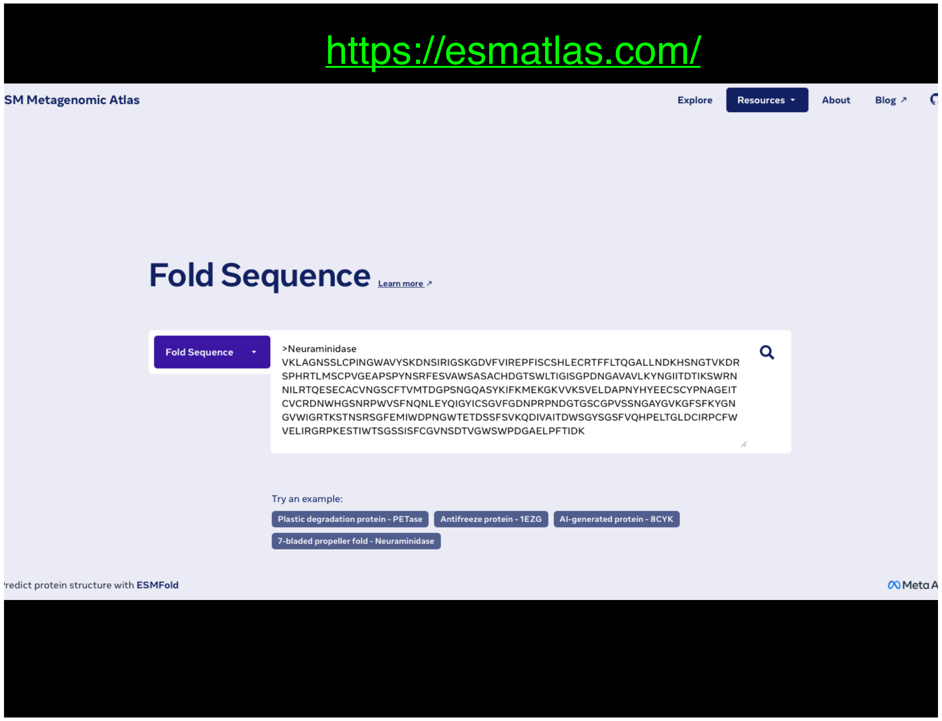
Subject Areas

All Articles

Animal Behavior and Cognition

Biochemistry

Bioengineering



https://esmatlas.com/

Fold Sequence

```
>Neuraminidase
VKLAGNSSLCPINGWAVYSKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDR
SPHRTLMSCPVGEAPSPYNSRFESVAWSASACHDGTSWLTIGISGPDNGAVAVLKYNGIITDTIKSWRN
NILRTQESECACVNGSCFTVMTDGPSNGQASYKIFKMEKGKVVKSVELDAPNYHYEECSCYPNAGEIT
CVCRDNWHGSNRPWVSFNQNLEYQIGYICSGVFGDNPRPNDGTGSCGPVSSNGAYGVKGFSFKYGN
GVWIGRTKSTNSRSGFEMIWDPNGWTETDSSFSVKQDIVAITDWSGYSGSFVQHPELTGLDCIRPCFW
VELIRGRPKESTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
```

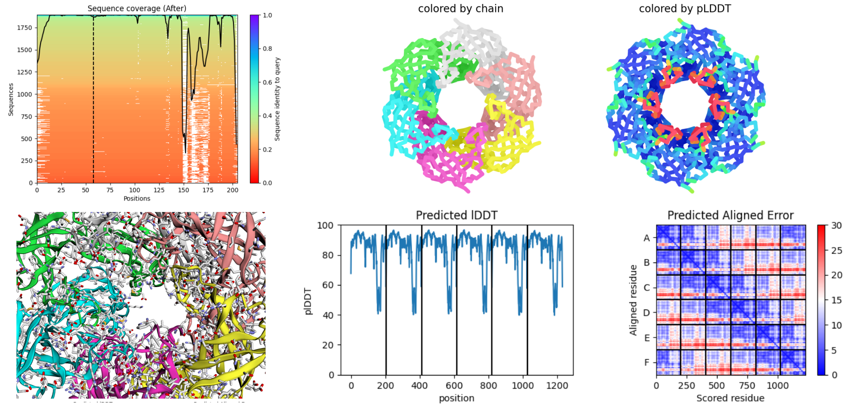
Try an example:

Plastic degradation protein - PETase
Antifreeze protein - 1EZG
AI-generated protein - 8CYK
7-bladed propeller fold - Neuraminidase

redict protein structure with ESMFold

ColabFold

Making Protein folding accessible via Google Colab



github.com/sokrypton/ColabFold

https://github.com/sokrypton/ColabFold