DeepMind

Highly accurate protein structure prediction with AlphaFold

Tim Green

Learning to Discover, Paris

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

[1]DeepMind, London, UK, [2]Seoul National University, South Korea

* Equal contribution † Corresponding authors: John Jumper (jumper@deepmind.com), Demis Hassabis (dhcontact@deepmind.com)

Introduction

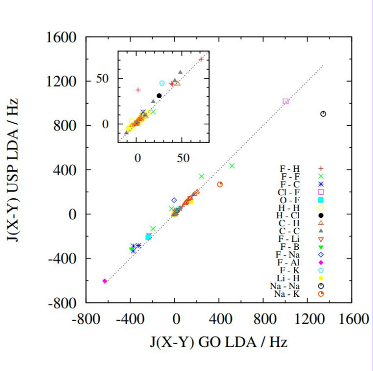
About me

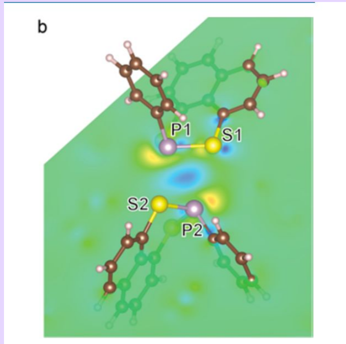
Tim Green – tfgg@deepmind.com

- 2006-2010 MSci at Cambridge Physics
- 2010-2014 DPhil at Oxford Materials (MML)
- 2014-2015 Postdoc at same
- 2016-present DeepMind

Thesis “Prediction of NMR J-coupling in condensed matter” developed DFT predictions with relativistic effects, disorder and temperature.

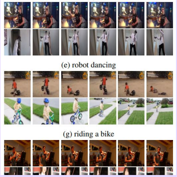
Work at DeepMind: *Computer Vision, Population Based Training, Deep RL, ML infrastructure* and *protein structure prediction*

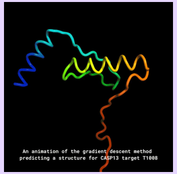


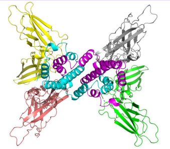


CASTEP

Sonnet




An animation of the gradient descent method predicting a structure for CASP13 target T1008



DeepMind and protein folding

A central part of DeepMind's mission is to solve fundamental scientific problems with AI

Predicting the 3D structure of a protein from its amino acid sequence is one such challenge

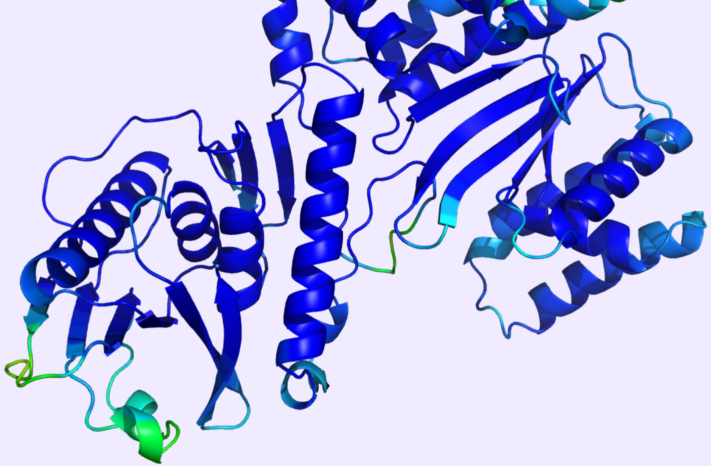
AlphaFold is our model that aims to solve this problem

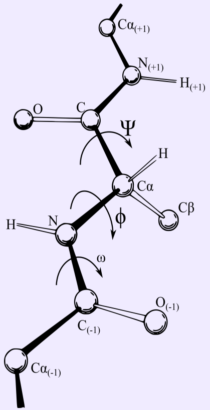
IAKFGDWINDEVERNVNEDGEPLLIQDVRQDSSKHYFFILKNGER
FDLLTREFDSFTSPDLTNEIKEITDQLSYYIYNKHFSSDFEQVEG
AKLNIQNEISQFVKEGKAPVQAAYNKLQDPDIKDLLDYYDNIEKH
SDEFESEIVKFFSEKKLIIKDAELEDVTQEGLNEGLQGGDLVQAF
EKNSKDNATANVKLMLSFLPKIDNLTGEPALGDYLNKPVFRSFDS
IHSELLEVLSDITTLHVQGEVLDVFSSMYNKIKELADFKKSFKPL
LEILDTIDEQKKTEFVQAFYLSKINFYTTTIETLETEDQNNTLTT
FKVQNVSNANNPISSKLTEYYTNFKYKILPGGKLNKGKLKDLQST
VTSLLEKTRKENNPKYKSDSDFYEVFEEGVVELMQVFEDLGVDSI
TFEAMDIFLKQFRFDLPENNAYKIMYQQYQGKLTNLNNLLKDIQS
NKINPYKINPFKNYSNLIFNSLAEAENYFIENNNESTIFSNGKTY
WNFARPSYISNRINTFKNNPGVLRQLLNTSYGQSSLWAKHLLGEE
KNVTGDFVLAGNARESASENRLKSLELSIFNSLQEKDKGAEGNDN
GSISIVDQLADKLNKVLRGGTKNGTSIYSTVTPGDKSTLHEIKID
HFIPETISSFSNGTMIFNDKIVNAFTDHFVSEVNRMKEAYQELET
LPESKRVVHYHTDARGNVMKDGKLAGNAFKSGHILSELSFDQITQ

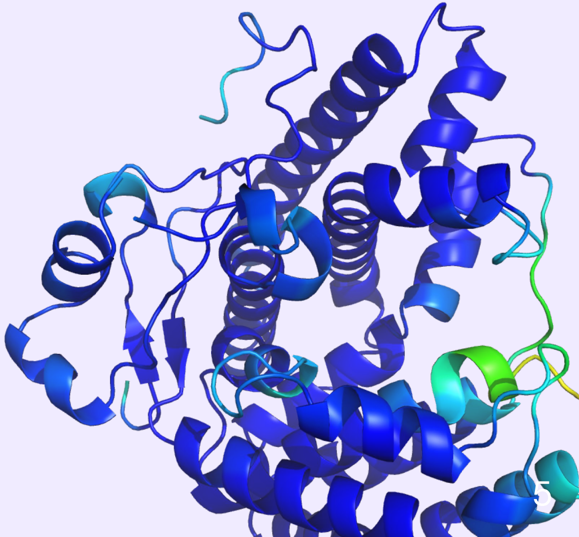
DDNEMLKLYNEDGSPINPKGAVS[illegible]ILIKQTINKVLNQRIKEN
IRYFKDQGLVIDTVNKDGNK[illegible]LDKSIMSEYTDDIQLTEFD
ISHVVSDFTLNSILASIEY[illegible]DPANYKNMVDFFKRVPATYT
NGTNLRLGLEANDHLFDVAVLENIVKPSAYLKEIGESLKLSDLSE
AEKKYILEAYEDVNQTDAQAWITPKRWAFLISRTGKWNSKYQSVY
NKILKSESLDASEMKLAAQPLKGVYFGLVNNTPTYLKYSQAVLLP
QLVAGTQLQSLADAMNKQDIGESIVLDGVKVGATTPNIVTDENGD
ILKSISLNPLTLSNADWKLQQDLPVKTIKPTLLGSQIQKNIYSSL
TDEATYTIENEAFNGSGMFQAINDTVSAMSNLSIAGLSSE[illegible]KDS
EGKIDKRKLYDMLEREMLDKGSAINLLKSIQKNLPIEAM[illegible]DK
LYNIVFSKINSAAVKLKTNGGSFIQLSNFGLDKQTADAKGITWLV

What are proteins?

- Proteins are molecular machines that are **essential to life**
- They have **many functions**: from our hair to our immune system
- Consist of **chains of amino acids** that fold into a 3D structure
- The exact **3D shape** is important for a protein’s function
- Understanding protein structures is a **fundamental problem in biology**



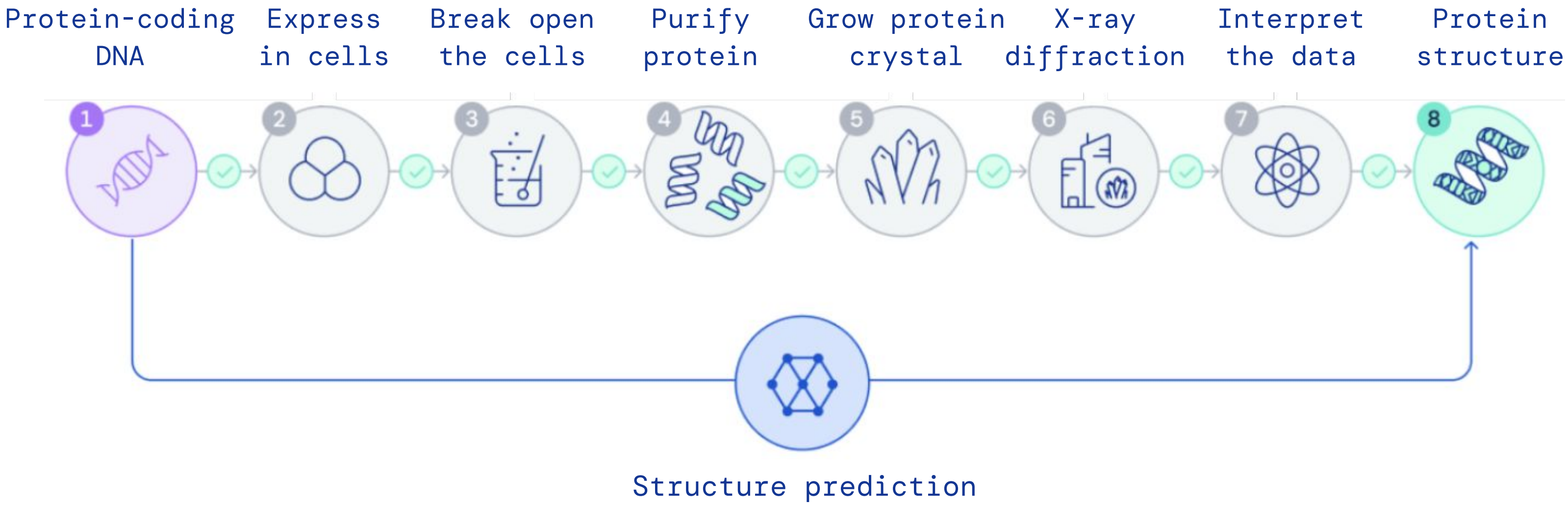




Why predict protein structures?

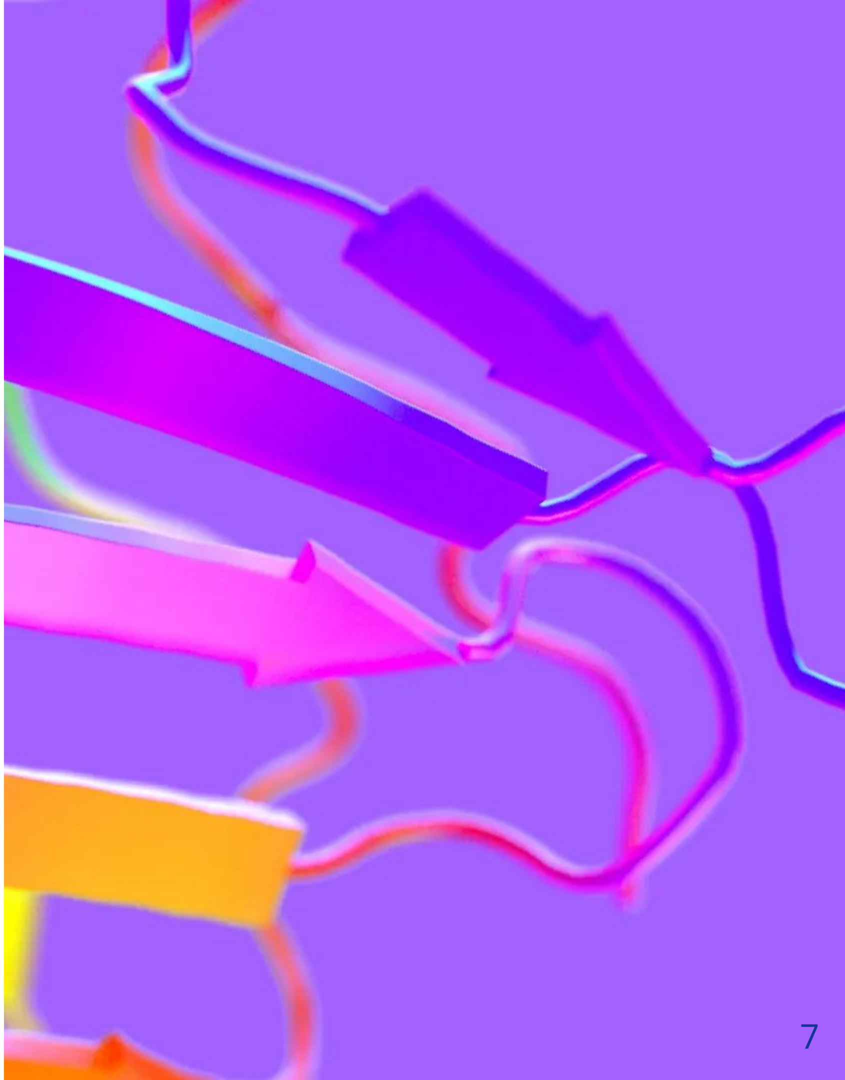
Experimental structure determination takes months to years.

Structure prediction can provide actionable information faster.



Agenda

- Introduction
- AlphaFold and CASP
- How AlphaFold works
- How AlphaFold builds protein structures
- AlphaFold impact
- AlphaFold-Multimer



AlphaFold and CASP

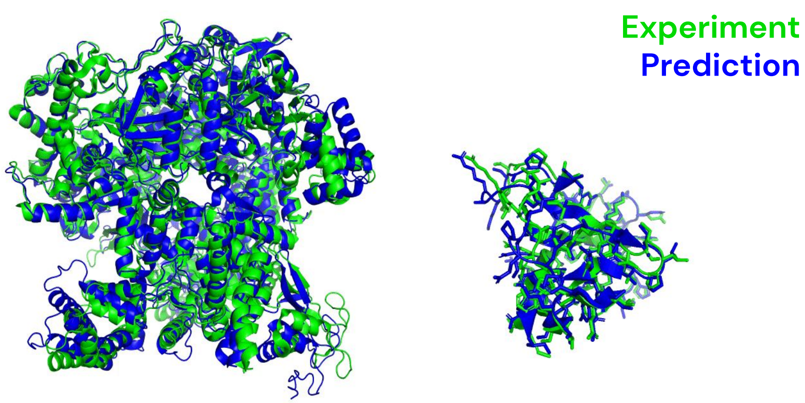
AlphaFold at CASP

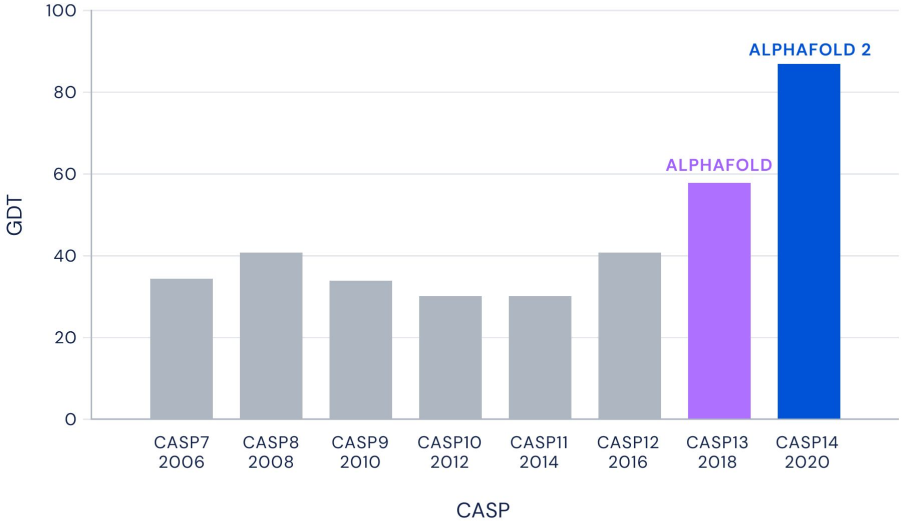
When working on these problems, a clear success metric is crucial

Fortunately, the protein structure prediction community had established CASP

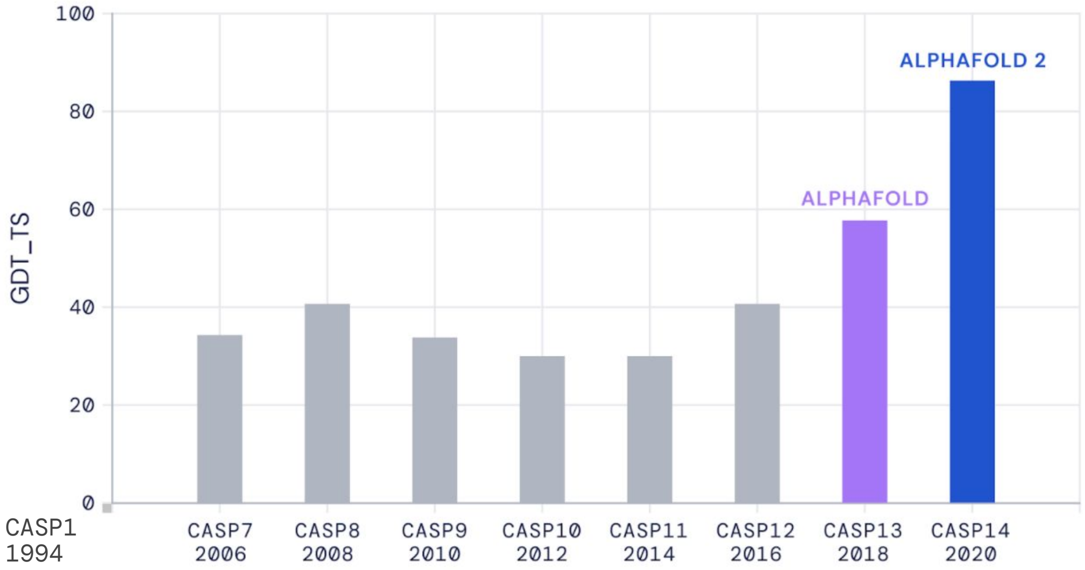
The CASP assessment involves predicting recently solved structures that aren’t yet public

At CASP14, AlphaFold was the top ranked method achieving consistently high accuracy



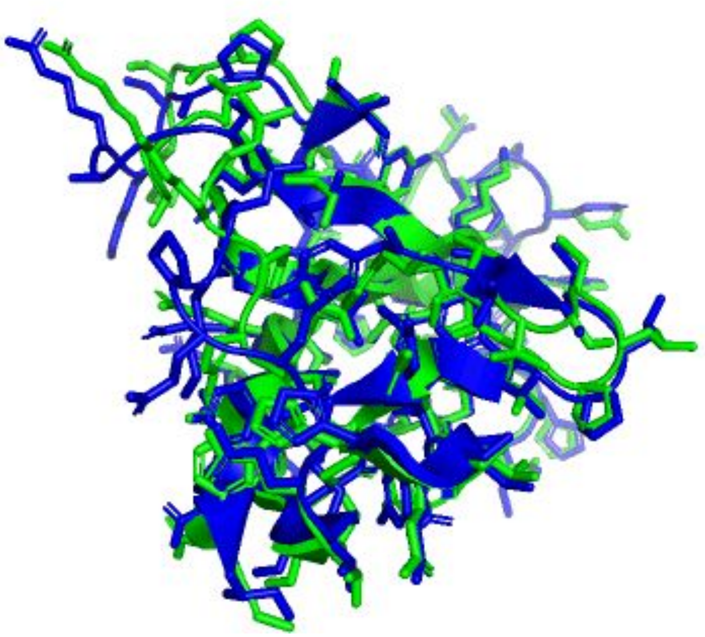


CASP: historical perspective



- CASP has provided a biennial blind assessment of structure prediction methods over the last 25 years
- AlphaFold 2 achieved a **median accuracy of 92.4 GDT** over all targets in CASP14
- In response, **AlphaFold was recognised as a solution to the structure prediction problem** by the CASP organizers

Protein example: T1064 (ORF8)

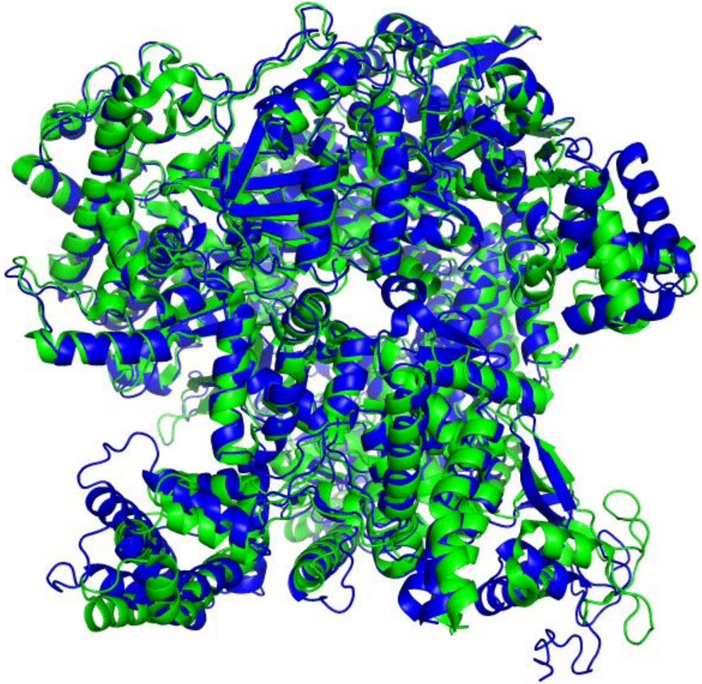


T1064 / 7jtl
87.0 GDT
(ORF8, SARS-CoV-2)

Ground truth
Prediction

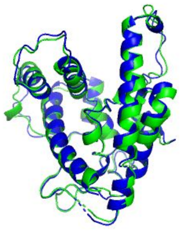
7JTL: Flower, T.G., et al. (2020) Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. Biorxiv.

Protein example: T1044 (RNA Polymerase)

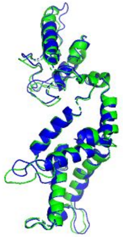


→ Folding as a single long chain

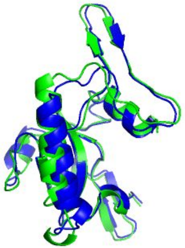
Individual domains



T1041



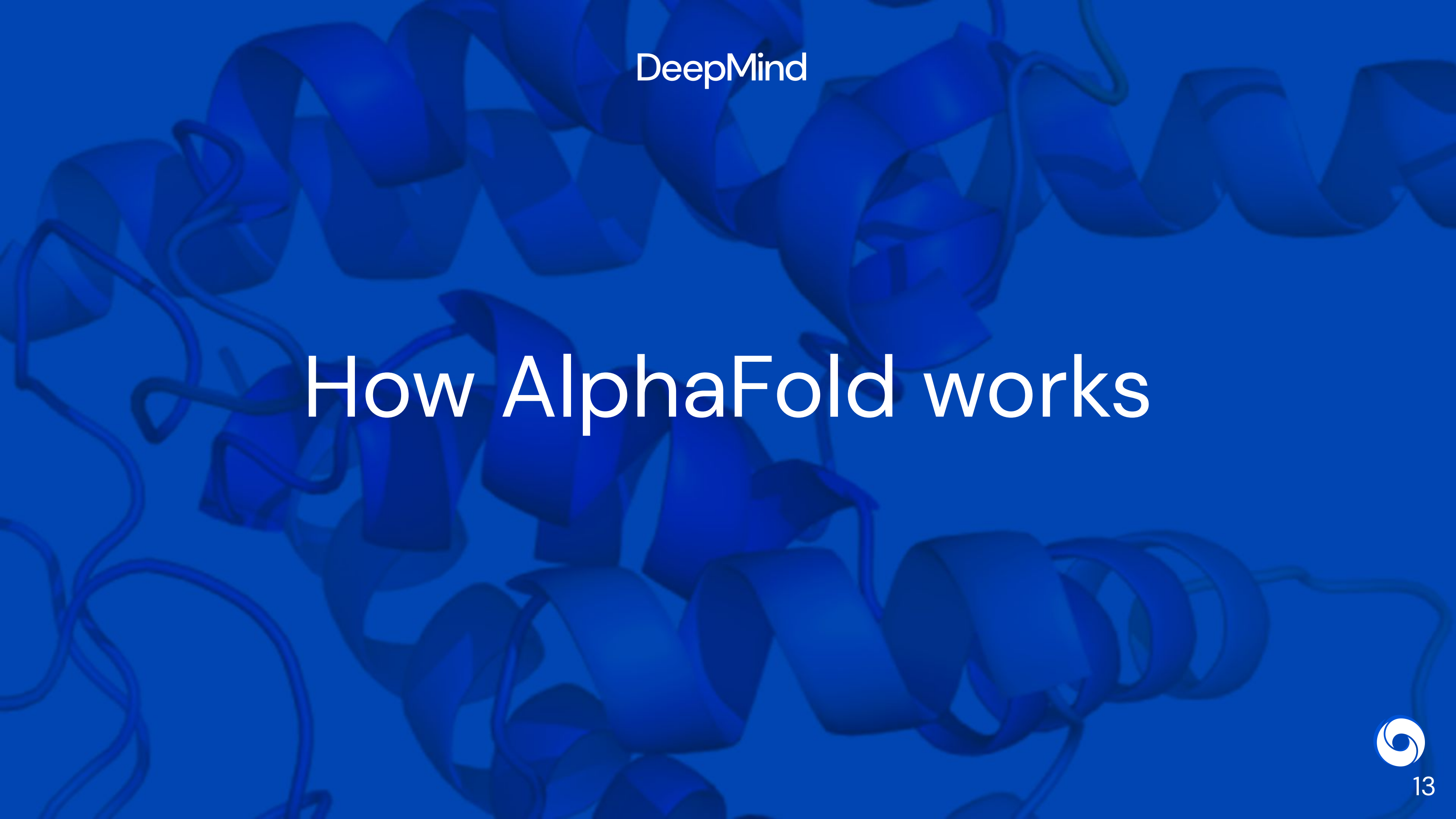
T1042



T1043

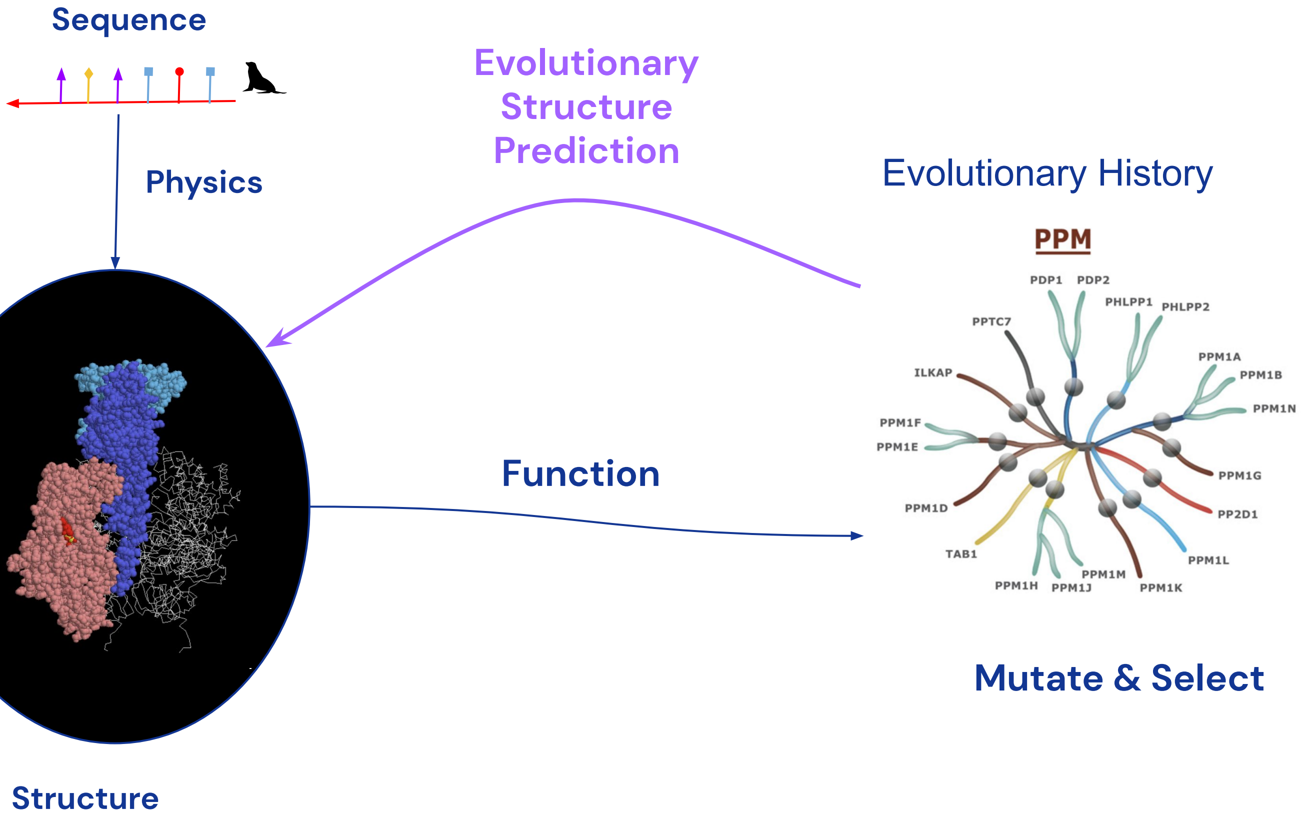
6VR4: Drobysheva, A.V., et al. Structure and function of virion RNA polymerase of a crAss-like phage. Nature (2020). (CASP14 target T1044)

Ground truth
Prediction



How AlphaFold works

Determining Structure from Evolution - Intuition



Sequence

Physics

Evolutionary Structure Prediction

Evolutionary History

Function

Mutate & Select

Structure

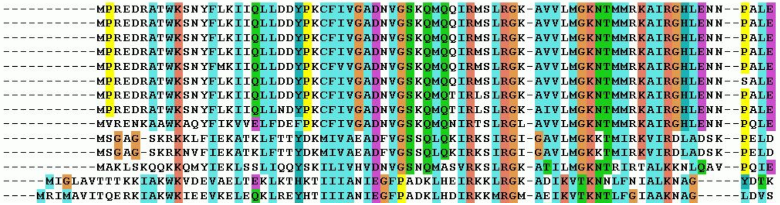
Model Inputs and Outputs

Inputs

- Amino acid sequence (residues) for the protein
- Evolutionary-related sequences (sequences that fold to the same structure, but whose amino acid sequence has diverged due to mutations)

SQETRKKCTEMKKKFKNCEVRCDESNHCVEVRCSDTKYTLC

Multiple Sequence Alignment (MSA)



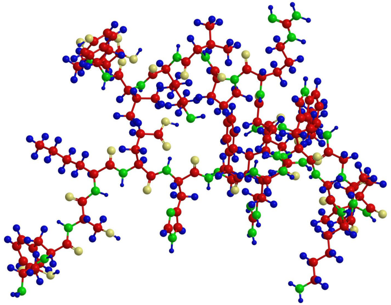
```
-----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
-----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
-----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
-----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
-----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
-----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
-----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
-----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
```

Training Data

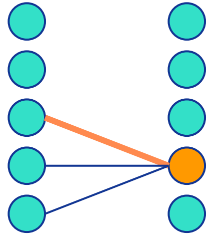
- **Labelled Data** – 170k structures, 40k after deduplicating
- **Unlabelled Data** – 350k deduplicated sequences

Output

- 3D position of every atom in the protein (300 – 50,000 atoms)

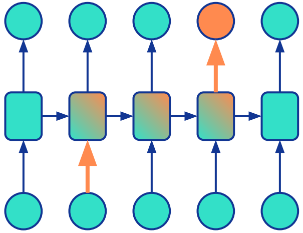


Inductive Bias for Deep Learning Models



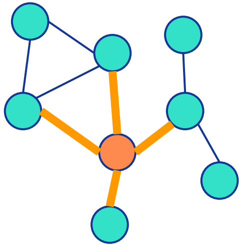
Convolutional Networks (e.g. computer vision)

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



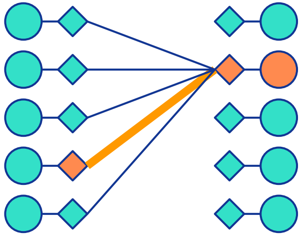
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



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

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

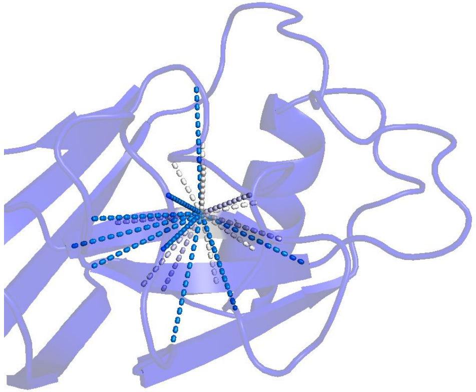


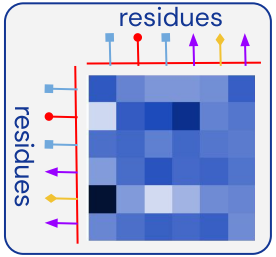
Attention Module (e.g. language)

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

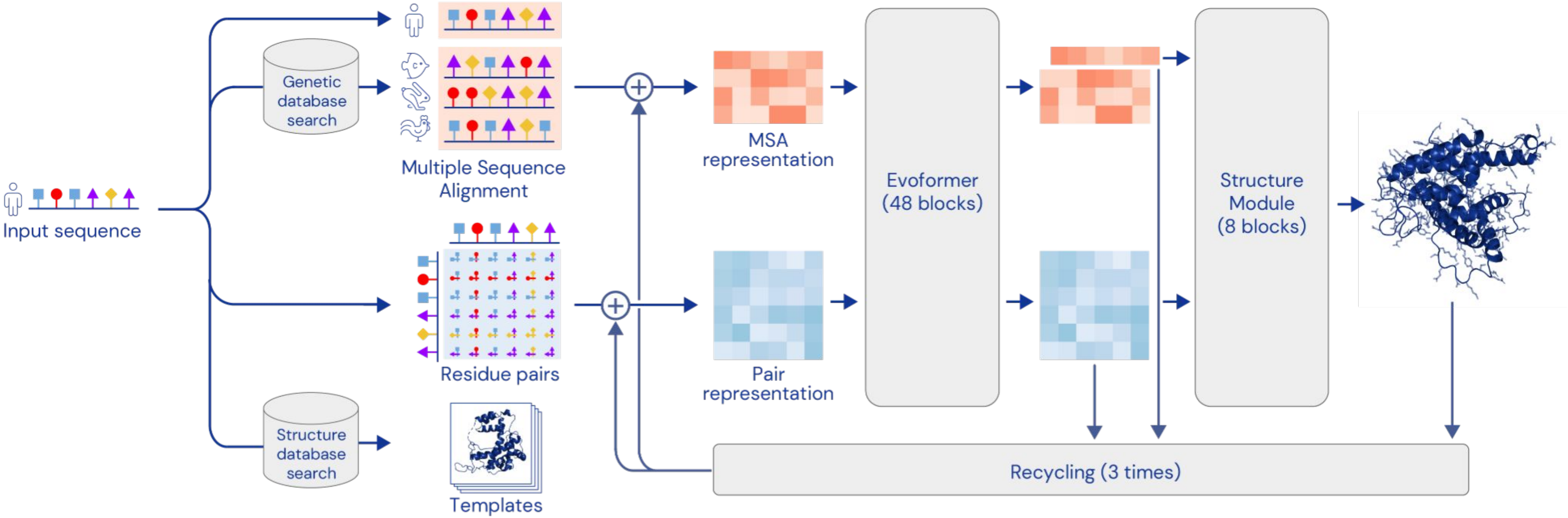
Putting our protein knowledge into the model

- Physical and geometric insights are built into the network structure, not just a process around it
- End-to-end system directly producing a structure instead of inter-residue distances
- Inductive biases reflect our knowledge of protein physics and geometry
    - The positions of residues in the sequence are de-emphasized
    - Instead residues that are close in the folded protein need to communicate
    - The network iteratively learns a graph of which residues are close, while reasoning over this implicit graph as it is being built





Network



Feeding certain outputs back through the network again improves performance

Structure module

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

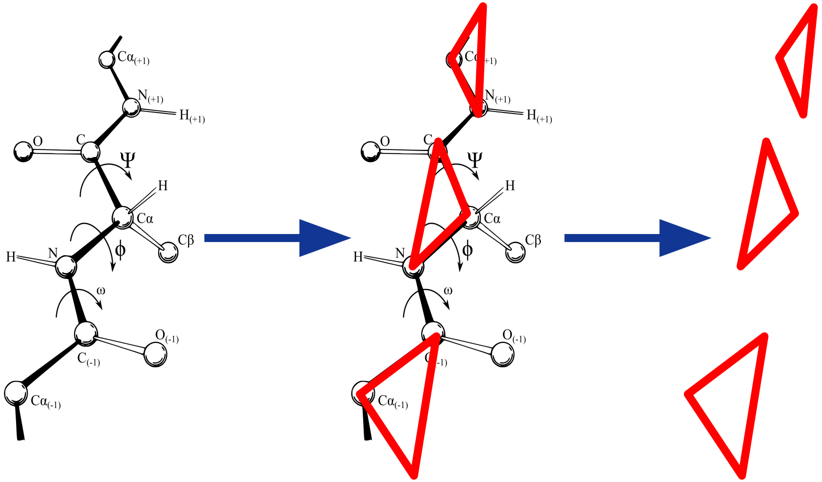
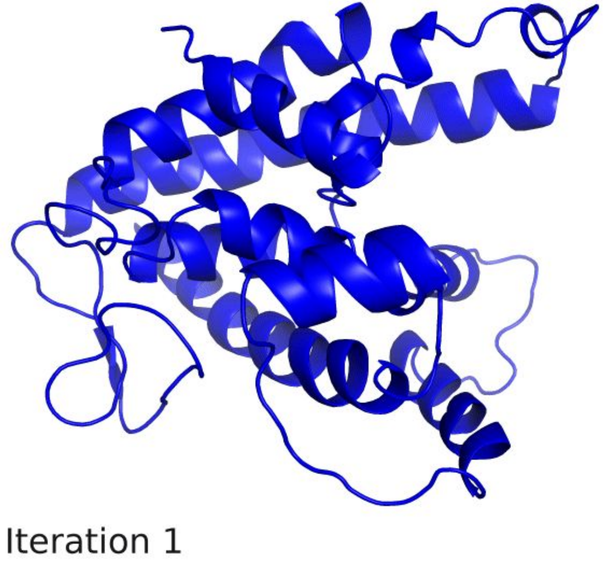


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

- **3-D equivariant transformer architecture** updates the rigid bodies / backbone
  - Also builds the side chains from torsion angles

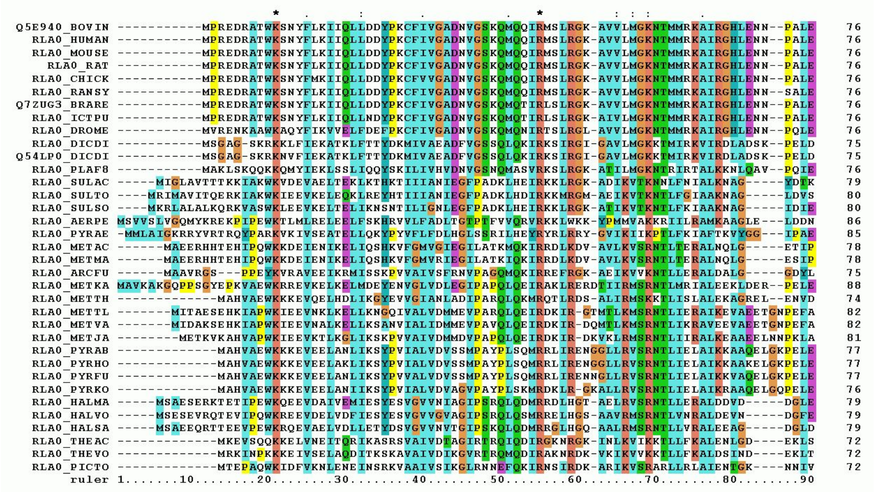


Target: T1041

Using unlabelled sequences in training

We know ~200k protein structures (Protein Data Bank) but several billion protein sequences.

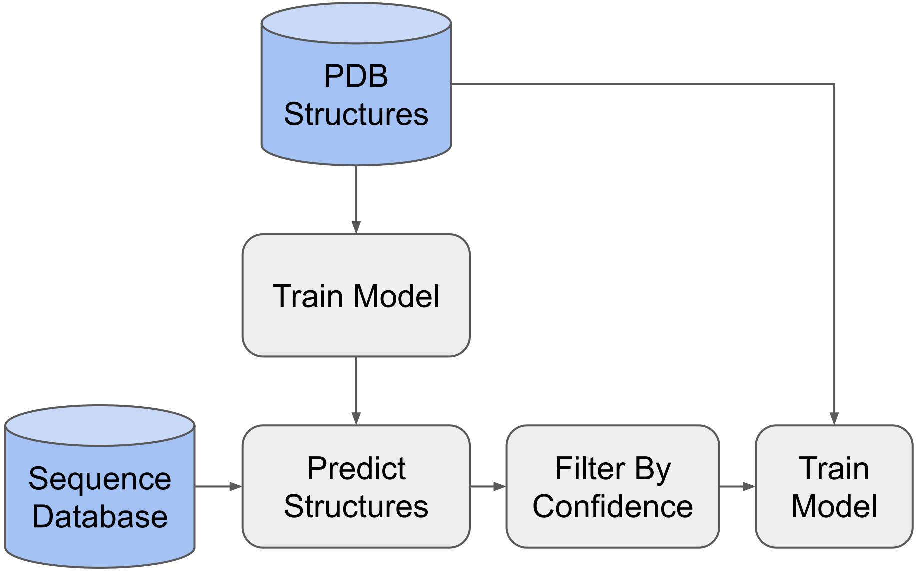
We use these data in two ways.



```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE    76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE    76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE    76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK    79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS    80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE    80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN    86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE    85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP    78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP    78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL    75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE    88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD    74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA    82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA    82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA    81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE    77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE    77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE    77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE    76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE    79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE    79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD    79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS    72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT    72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV    72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

MSA BERT

(train model to predict masked locations in MSA)



Noisy student self-distillation

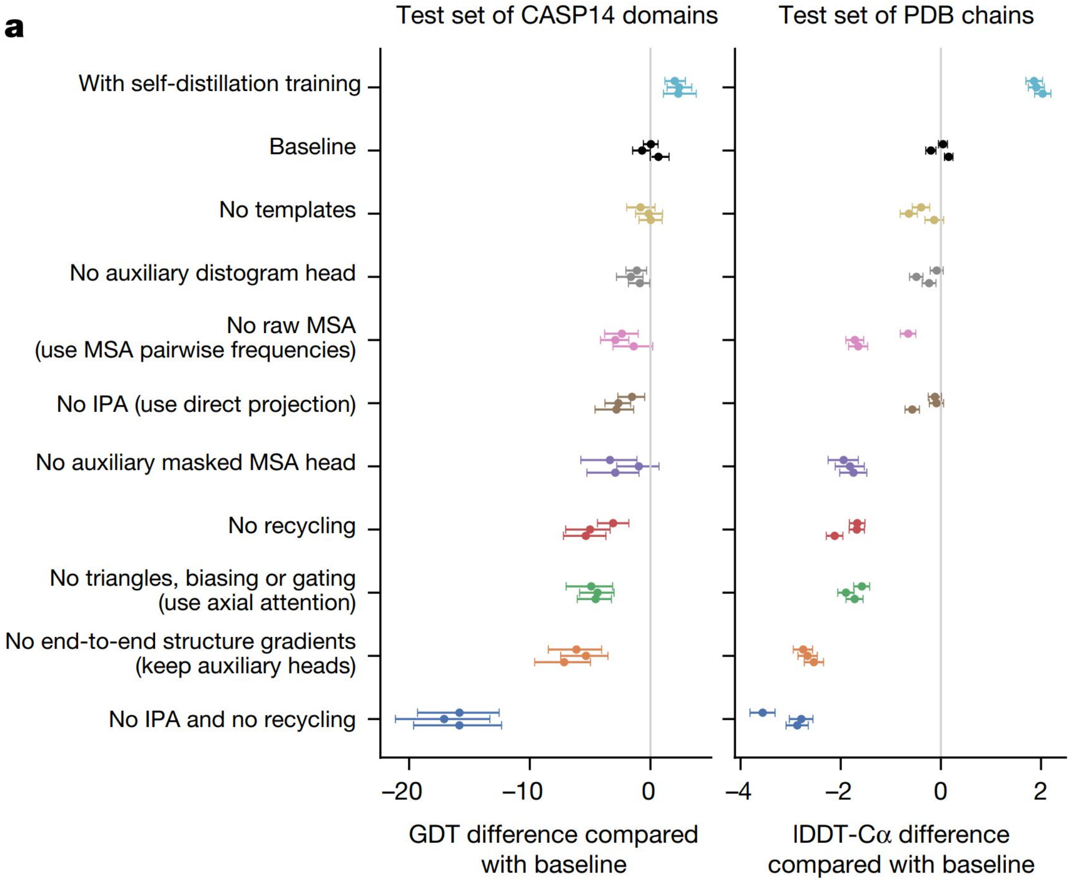
(train from predictions of same architecture)

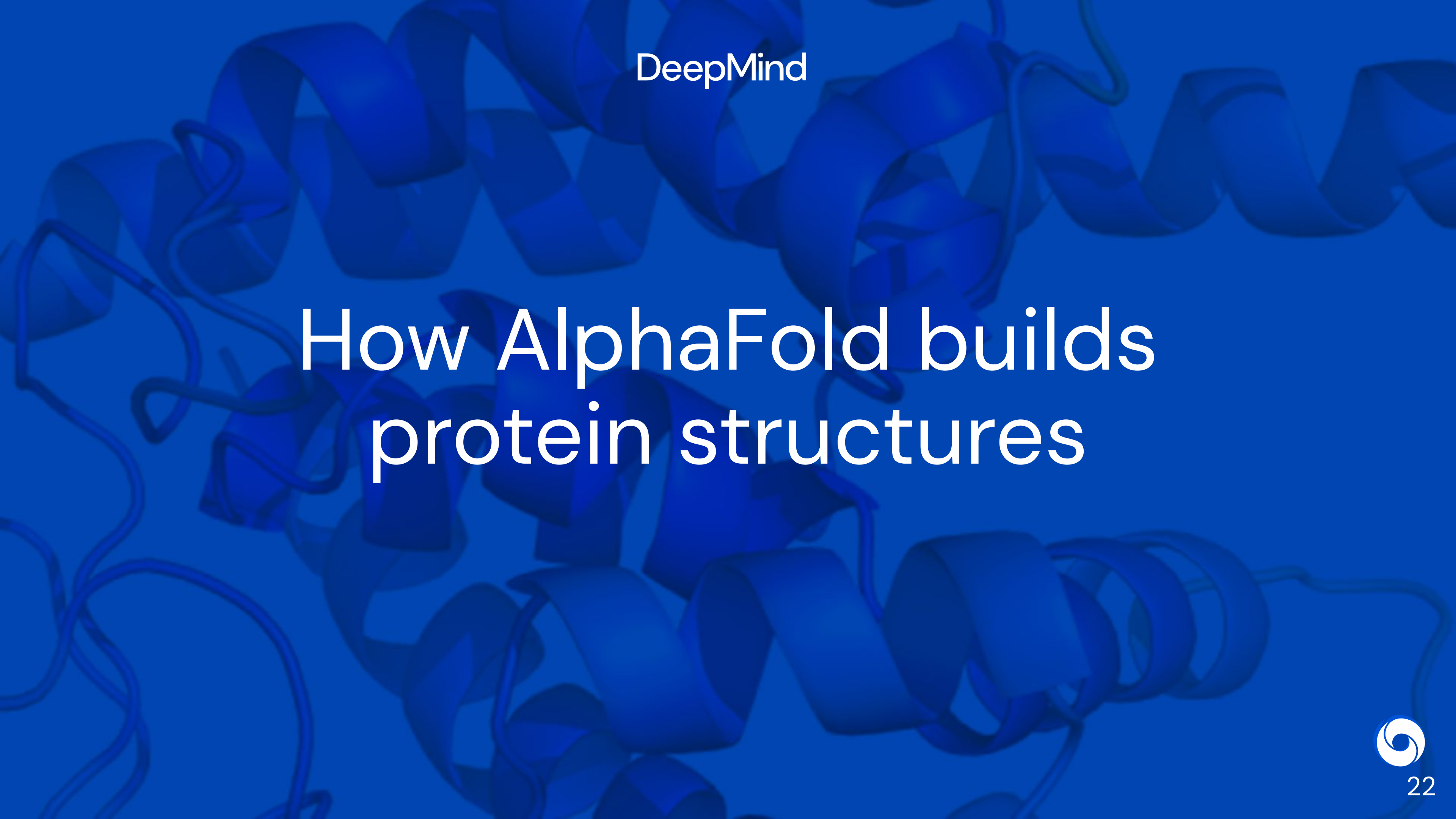
Which parts mattered? All of it

No single improvement is dominant

More important was the methodology of building the protein intuition into the model

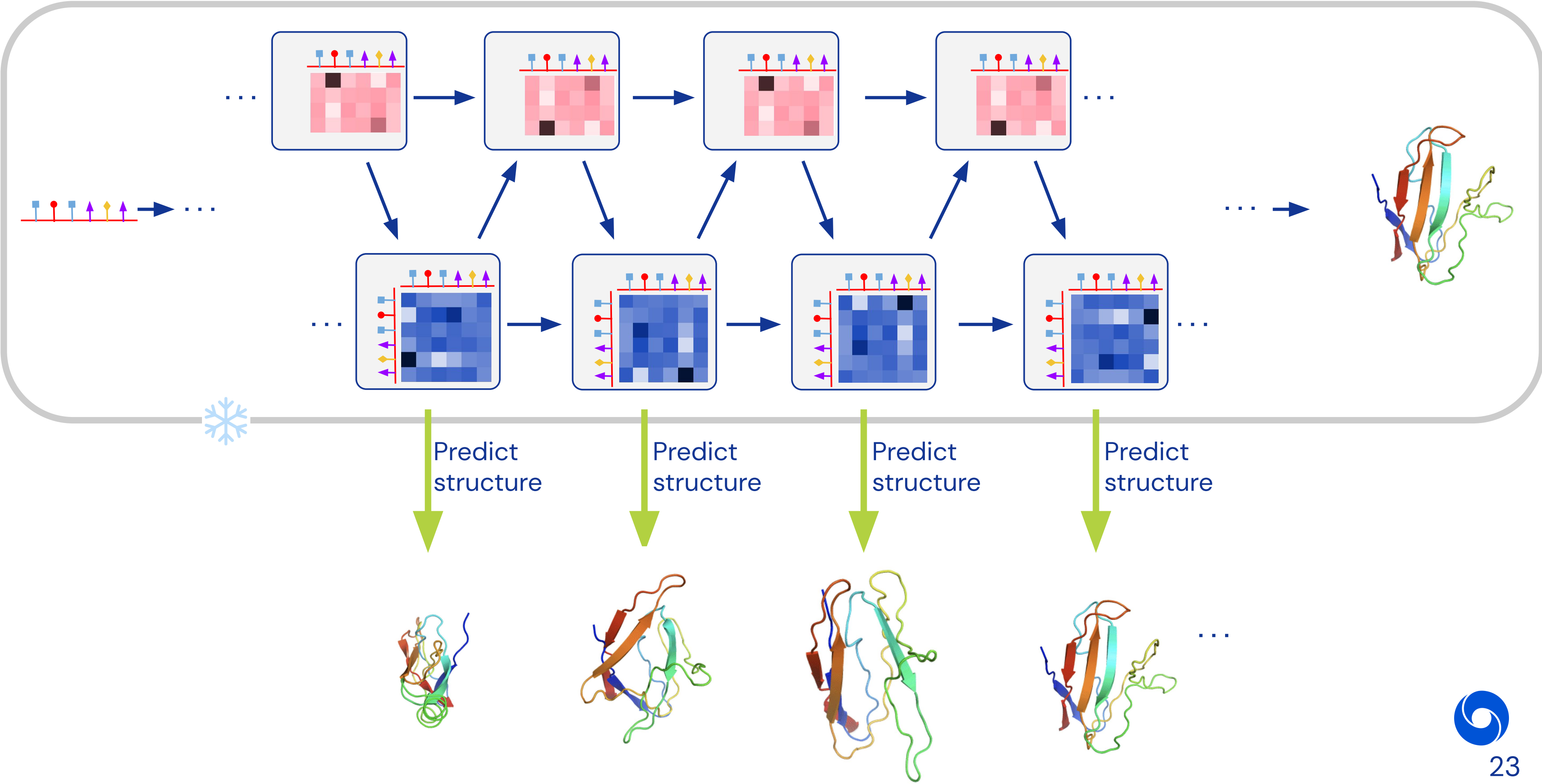
Multiple ablations suggest strong interactions between many of the components





How AlphaFold builds protein structures

Interrogating the Network



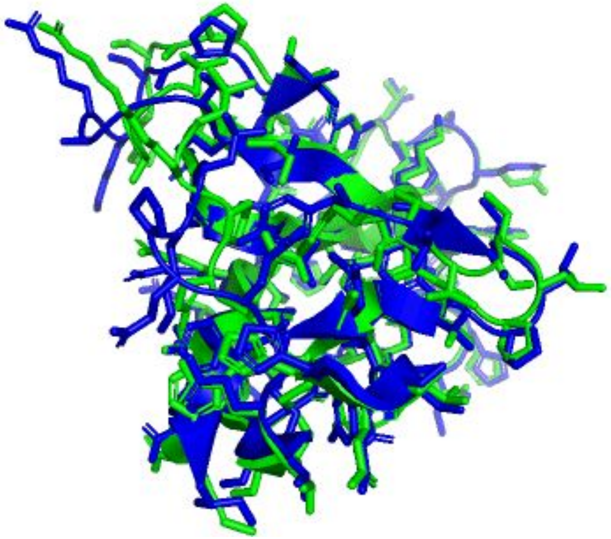
Predict structure

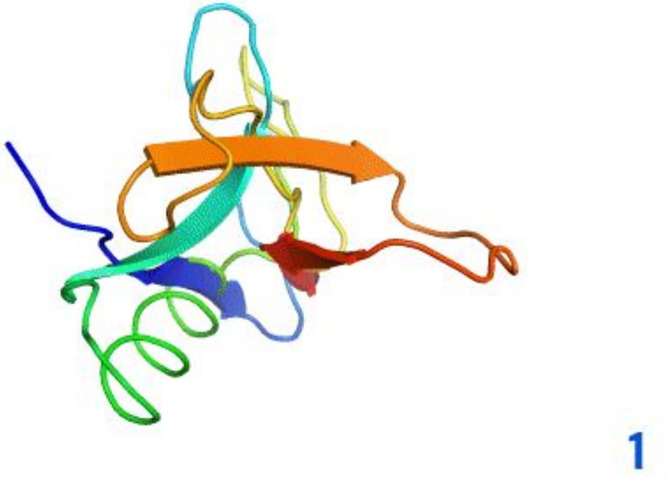
Predict structure

Predict structure

Predict structure

Model interpretability - ORF8 - Sars-Cov2

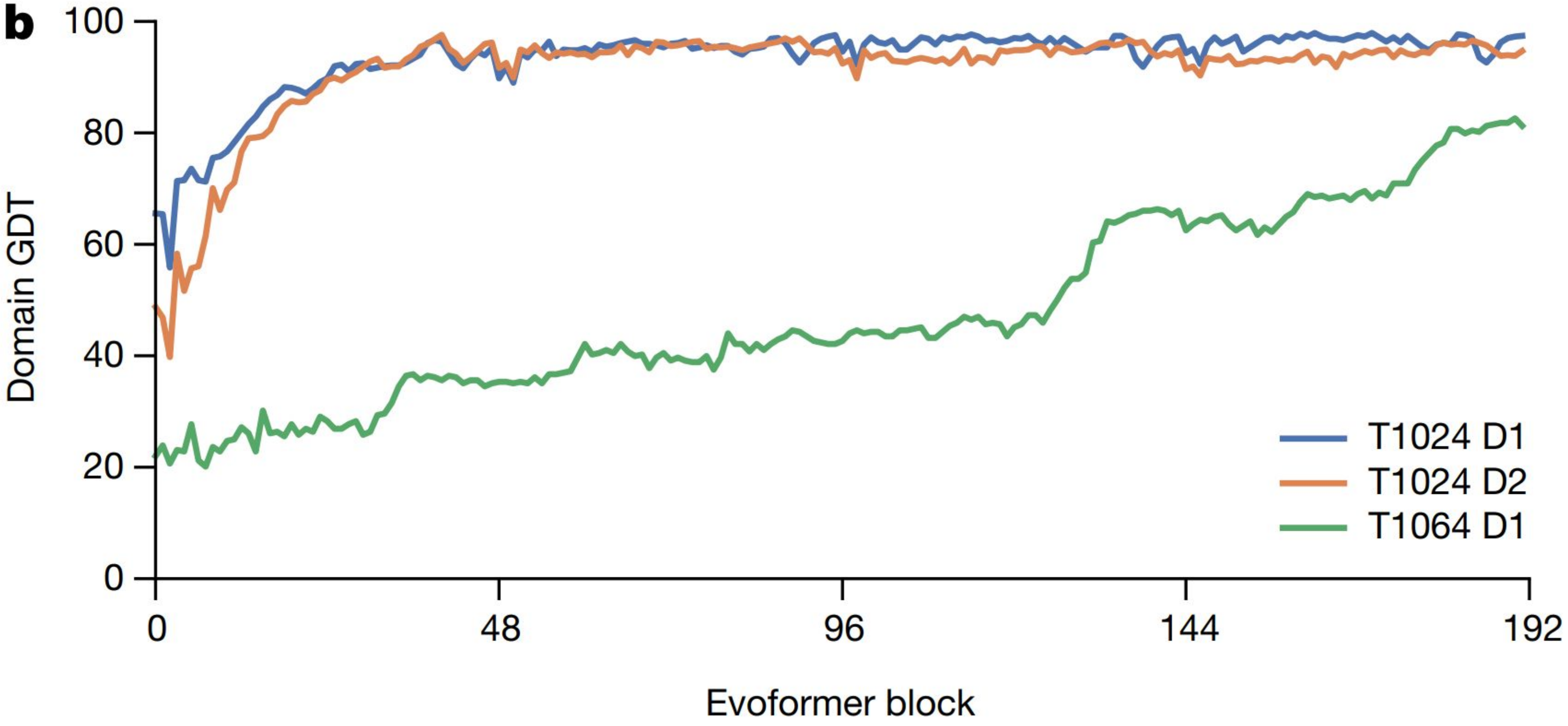




7JTL: Flower, T.G., et al. (2020) Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. Biorxiv.

Model interpretability - T1044

Model interpretability - Role of depth



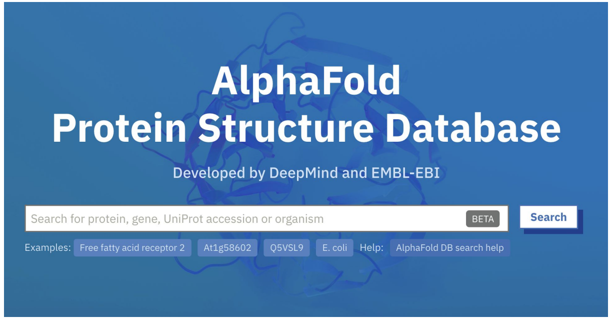
AlphaFold Impact

Open Source & AlphaFold Protein Structure Database

We open sourced the code and model weights to run AlphaFold – github.com/deepmind/alphafold (8.3k ⭐)

Also created AlphaFold Protein Structure Database

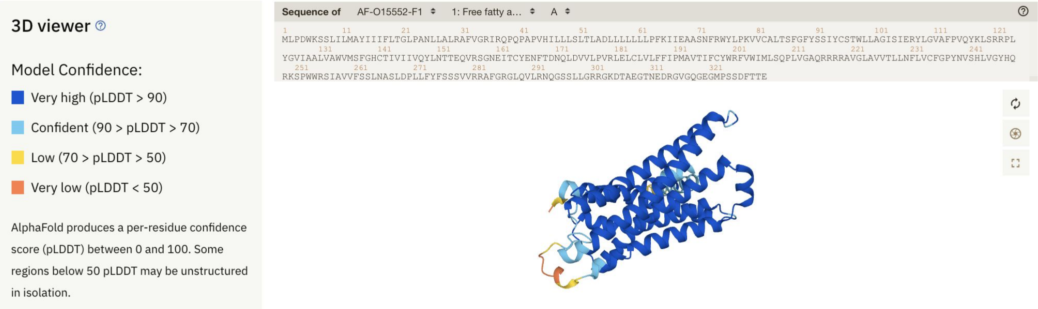
- Website developed and hosted by EMBL-EBI
- Contains pre-run predictions for **21 model organisms + SwissProt** (>800k structures)
- Plans to expand to **Uniref90** (~135M structures)



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



3D viewer

Model Confidence:

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

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

Sequence of AF-O15552-F1 · 1: Free fatty a... · A

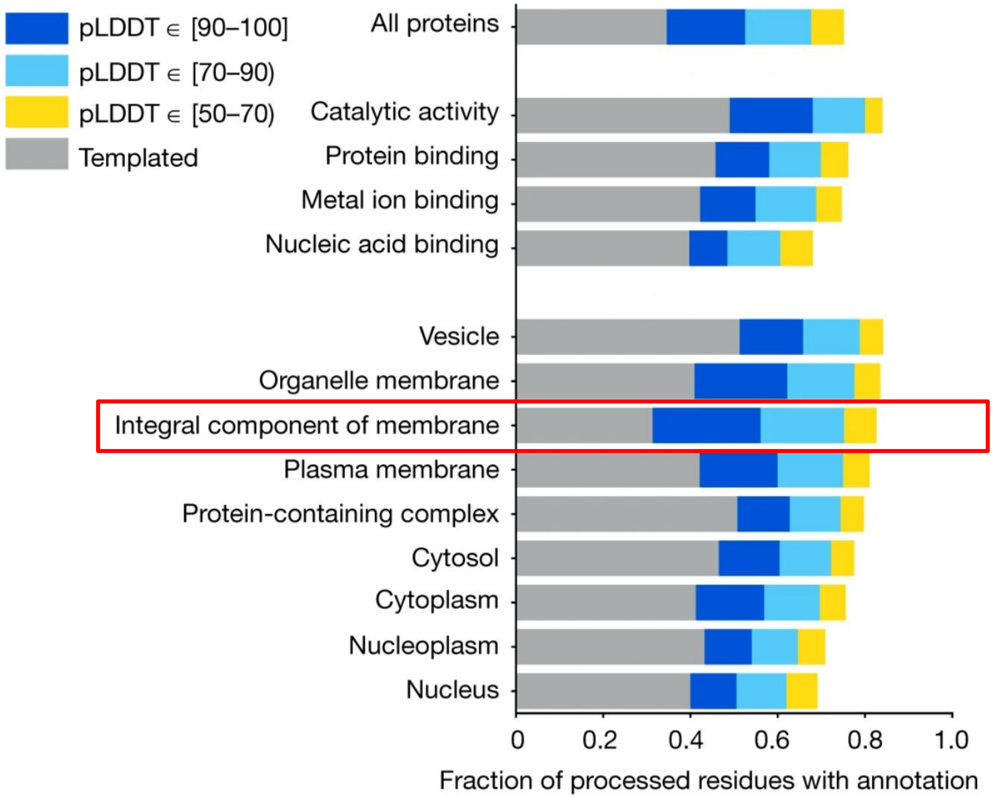
```
1         11        21        31        41        51        61        71        81        91        101       111       121
MLPDWKSSLIILMAYIIIFLTGLPANLLALRAFVGRIRQPQPAPVHILLLSLTLADLLLLLLLPFKIIEAASNFRWYLPKVVCALTSFGFYSSIYCSTWLLAGISIERYLGVAFPVQYKLSRRPL
      131       141       151       161       171       181       191       201       211       221       231       241
YGVIAALVAWVMSFGHCTIVIIVQYLNTTEQVRSGNEITCYENFTDNQLDVVLPVRLELCLVLFFIPMAVTIFCYWRFVWIMLSQPLVGAQRRRRAVGLAVVTLLNFLVCFGPYNVSHLVGYHQ
  251       261       271       281       291       301       311       321
RKSPWWRSIAVVFSSLNASLDPLLFYFSSSVVRRAFGRGLQVLRNQGSSLLGRRGKDTAEGTNEDRGVGQGEGMPSSDFTTE
```

EMBL

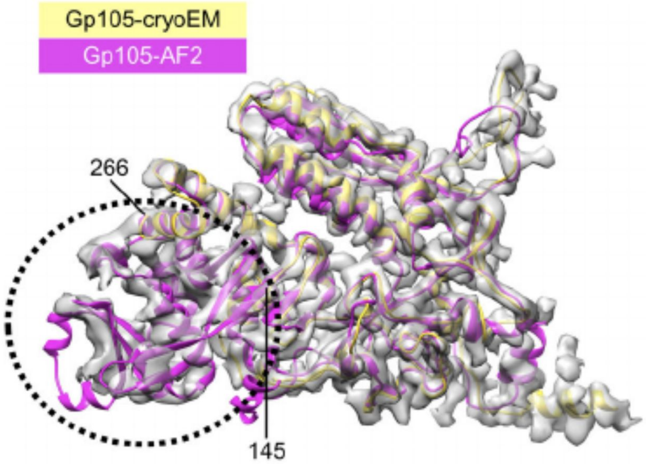
DeepMind

Increase in coverage of the human proteome

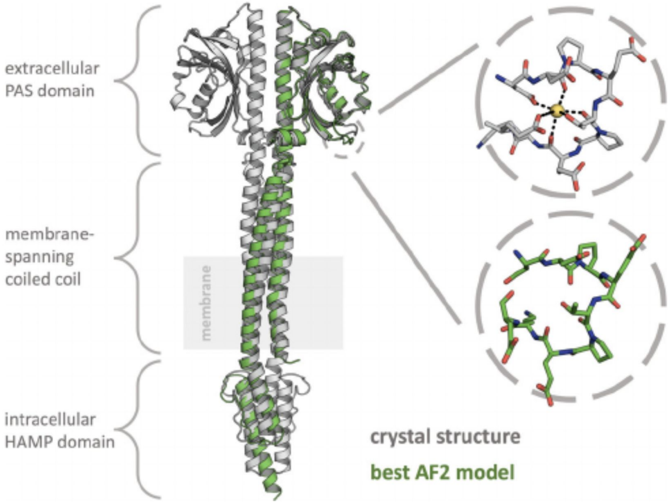
Even when accounting for template-modelling, AlphaFold greatly expands the high-accuracy structural coverage of the human proteome



AlphaFold as an aid to experimental structure determination



AlphaFold models of AR9 nvRNAP proteins fit the cryo-EM density nearly perfectly

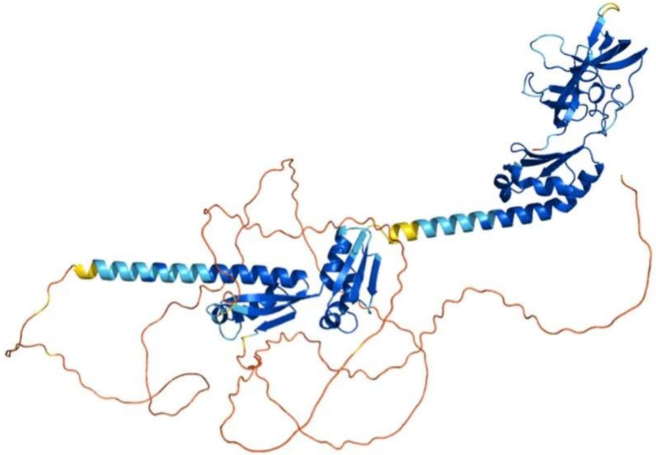


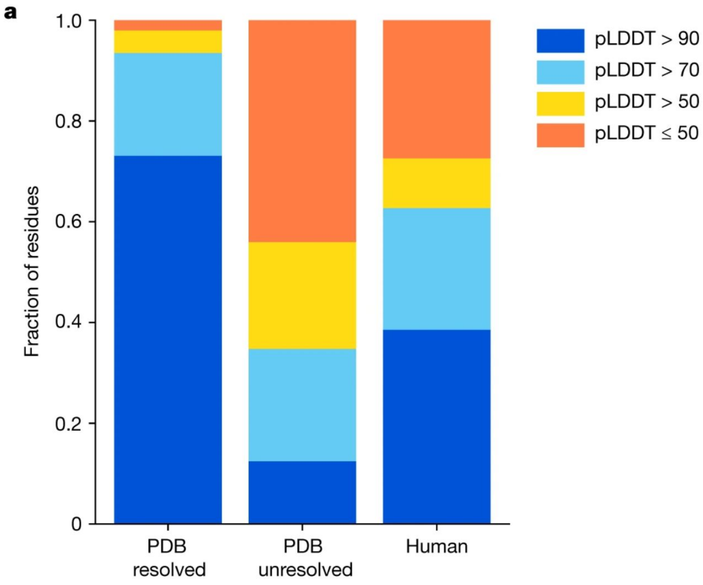
Crystal structure of dimeric Af1503

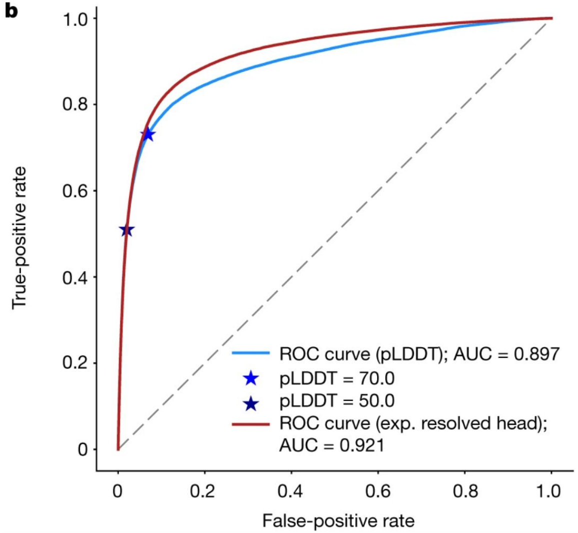
Figures taken from: Kryshtafovych, Andriy, et al. Computational models in the service of X-ray and cryo-EM structure determination. Proteins: Structure, Function, and Bioinformatics (2021).

AlphaFold confidence and disorder

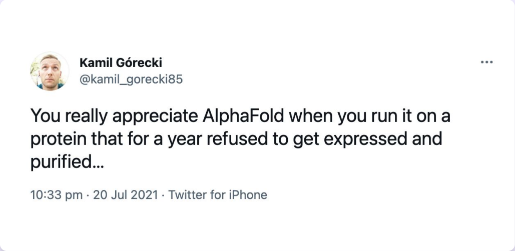
- In IDRs, there is no fixed structure for AlphaFold to identify
- So AlphaFold produces a boring ribbon *and* reports very low confidence (pLDDT)
- This make very low confidence a strong signal of disorder or chains that are unstructured in isolation







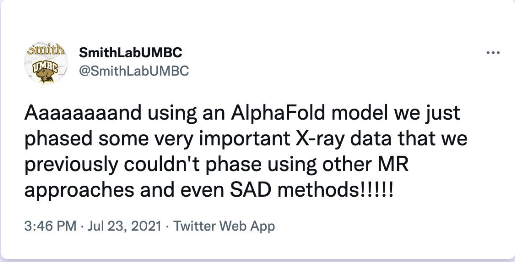
AlphaFold accuracy on CAID disorder benchmark



Kamil Górecki
@kamil_gorecki85

You really appreciate AlphaFold when you run it on a protein that for a year refused to get expressed and purified...

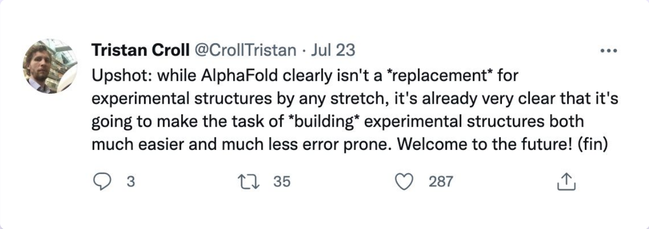
10:33 pm · 20 Jul 2021 · Twitter for iPhone



SmithLabUMBC
@SmithLabUMBC

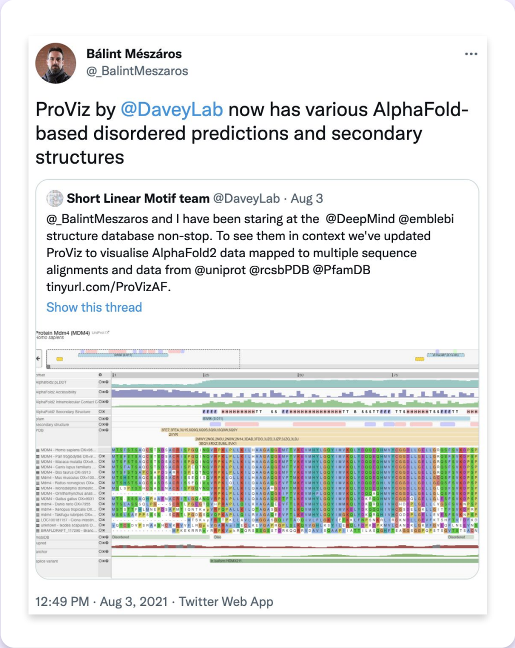
Aaaaaaaand using an AlphaFold model we just phased some very important X-ray data that we previously couldn't phase using other MR approaches and even SAD methods!!!!!

3:46 PM · Jul 23, 2021 · Twitter Web App



Tristan Croll @CrollTristan · Jul 23

Upshot: while AlphaFold clearly isn't a *replacement* for experimental structures by any stretch, it's already very clear that it's going to make the task of *building* experimental structures both much easier and much less error prone. Welcome to the future! (fin)



Bálint Mészáros
@_BalintMeszaros

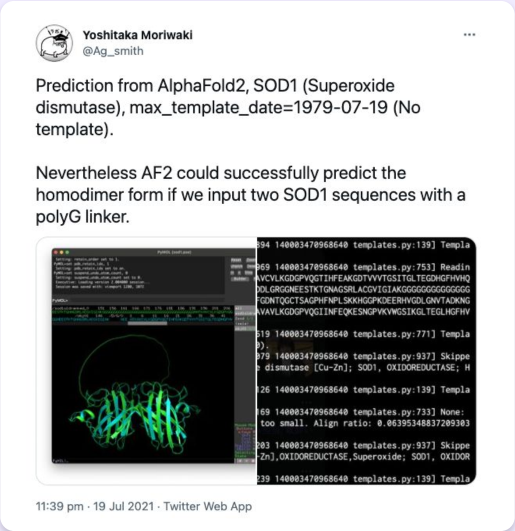
ProViz by @DaveyLab now has various AlphaFold-based disordered predictions and secondary structures

Short Linear Motif team @DaveyLab · Aug 3

@_BalintMeszaros and I have been staring at the @DeepMind @emblebi structure database non-stop. To see them in context we've updated ProViz to visualise AlphaFold2 data mapped to multiple sequence alignments and data from @uniprot @rcsbPDB @PfamDB tinyurl.com/ProVizAF.

Show this thread

12:49 PM · Aug 3, 2021 · Twitter Web App



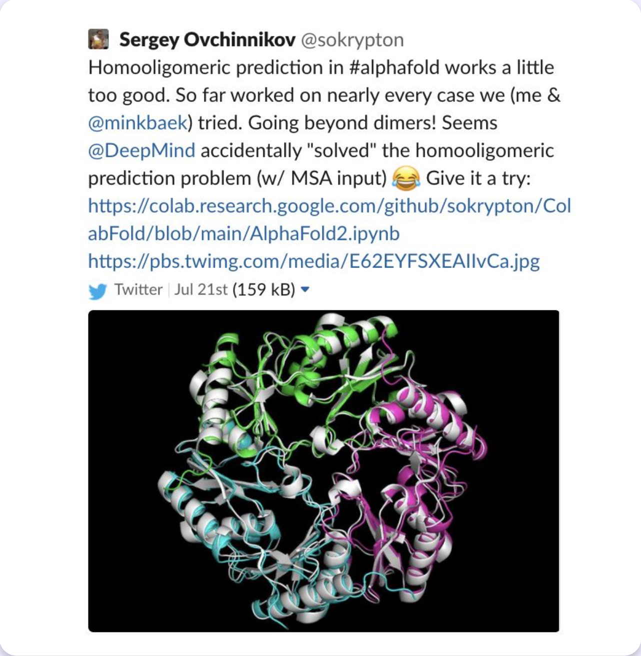
Yoshitaka Moriwaki
@Ag_smith

Prediction from AlphaFold2, SOD1 (Superoxide dismutase), max_template_date=1979-07-19 (No template).

Nevertheless AF2 could successfully predict the homodimer form if we input two SOD1 sequences with a polyG linker.

```
394 140003470968640 templates.py:139] Templa

969 140003470968640 templates.py:753] Readin
AVCVLKGDGPVQGTIHFEAKGDTVVVTGSITGLTEGDHGFHVHQ
DDLGRGGNEESTKTGNAGSRLACGVIGIAKGGGGGGGGGGGGGG
FGDNTQGCTSAGPHFNPLSKKHGGPKDEERHVGDLGNVTADKNG
AVAVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHV

519 140003470968640 templates.py:771] Templa
0).
079 140003470968640 templates.py:937] Skippe
e dismutase [Cu-Zn]; SOD1, OXIDOREDUCTASE; H

126 140003470968640 templates.py:139] Templa

169 140003470968640 templates.py:733] None:
 too small. Align ratio: 0.0639534883720930З

203 140003470968640 templates.py:937] Skippe
-Zn],OXIDOREDUCTASE,Superoxide; SOD1, OXIDOR

239 140003470968640 templates.py:139] Templa
```

11:39 pm · 19 Jul 2021 · Twitter Web App



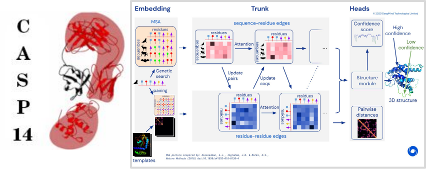
Sergey Ovchinnikov @sokrypton

Homooligomeric prediction in #alphafold works a little too good. So far worked on nearly every case we (me & @minkbaek) tried. Going beyond dimers! Seems @DeepMind accidentally "solved" the homooligomeric prediction problem (w/ MSA input) 😂 Give it a try: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
https://pbs.twimg.com/media/E62EYFSXEAIIvCa.jpg

Twitter | Jul 21st (159 kB) ▾

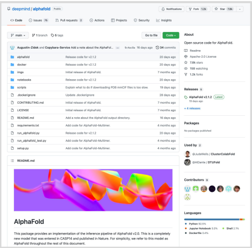
AlphaFold timeline

CASP14 conference



MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

AlphaFold paper & code release



Augustin-Zidek and Copybara-Service Add a note about the AlphaFol... 9c4ac8a 16 days ago 34 commits

alphafold	Release code for v2.1.2	20 days ago
docker	Release code for v2.1.2	20 days ago
imgs	Initial release of AlphaFold.	7 months ago
notebooks	Release code for v2.1.2	20 days ago
scripts	Explain what to do if downloading PDB mmCIF files is too slow.	19 days ago
.dockerignore	Update .dockerignore	28 days ago
CONTRIBUTING.md	Initial release of AlphaFold.	7 months ago
LICENSE	Initial release of AlphaFold.	7 months ago
README.md	Add a note about the AlphaFold output directory.	16 days ago
requirements.txt	Add code for AlphaFold-Multimer.	4 months ago
run_alphafold.py	Release code for v2.1.2	20 days ago
run_alphafold_test.py	Add code for AlphaFold-Multimer.	4 months ago
setup.py	Add code for AlphaFold-Multimer.	4 months ago

README.md

AlphaFold

This package provides an implementation of the inference pipeline of AlphaFold v2.0. This is a completely new model that was entered in CASP14 and published in Nature. For simplicity, we refer to this model as AlphaFold throughout the rest of this document.

About

Open source code for AlphaFold.

Readme
Apache-2.0 License
7.8k stars
158 watching
1.2k forks

Releases 5

AlphaFold v2.1.2 Latest
19 days ago
+ 4 releases

Packages

No packages published

Used by 2

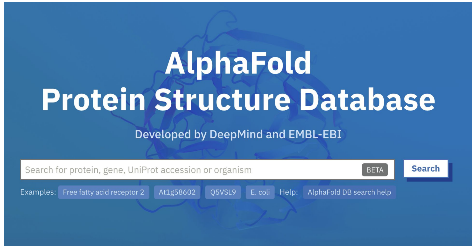
@JudeWells / ClusterColabFold
@AlDante / DTUFold

Contributors 9

Languages

- Python 92.5%
- Jupyter Notebook 5.0%
- Shell 2.1%
- Dockerfile 0.4%

AlphaFold DB & Proteome paper



AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA

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

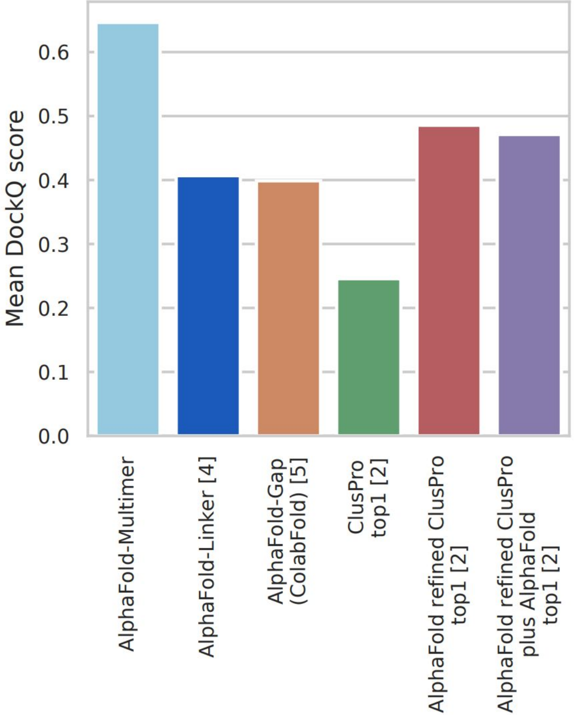
November 2020

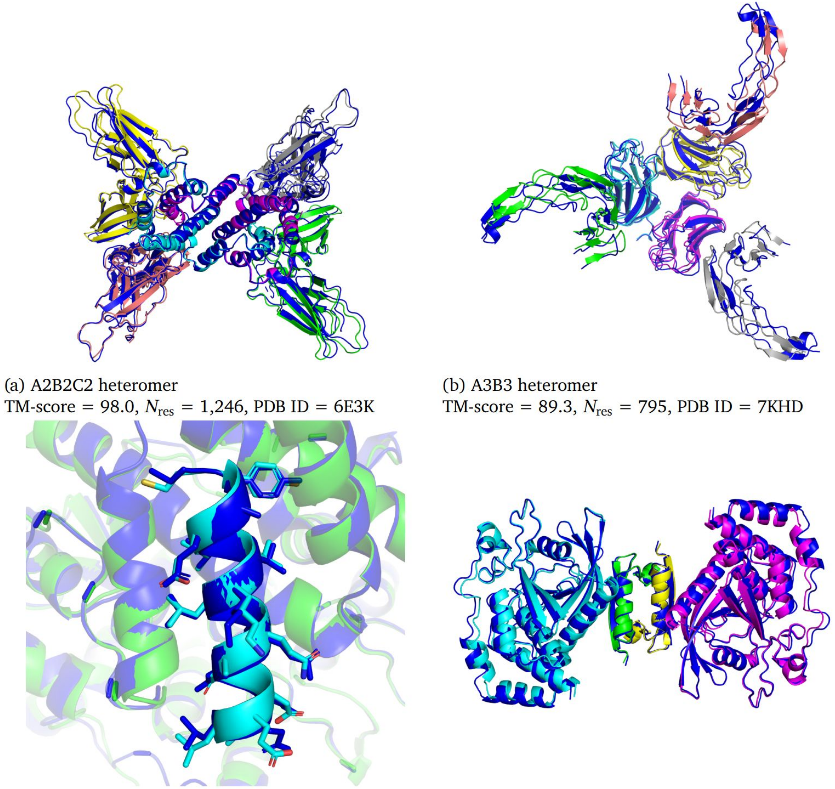
July 2021

AlphaFold-Multimer

Training AlphaFold to predict protein complexes (AlphaFold-Multimer)

Adapting the inputs, loss function, and training of AlphaFold to handle multimers and then training the model from scratch





(a) A2B2C2 heteromer
TM-score = 98.0, N_{res} = 1,246, PDB ID = 6E3K

(b) A3B3 heteromer
TM-score = 89.3, N_{res} = 795, PDB ID = 7KHD

(c) Protein-peptide complex
TM-score = 96.0, DockQ = 0.948, N_{res} = 385, PDB ID = 6JMT

(d) A2B2 heteromer
TM-score = 98.3, N_{res} = 716, PDB ID = 6IWD

Richard Evans et al. 2021 - https://doi.org/10.1101/2021.10.04.463034

Organizing

- Make it easy to measure your performance – one metric, one leaderboard
- Incrementally improve this metric
  - +0.5% per week adds up – enough quantitative change adds up to qualitative change!

- How to create research velocity:
  - Enable fast iteration – the more ideas you can test, the faster you progress
  - Allow people to build on each other’s work – you should always be improving SOTA
  - Test your code – avoid errors that damage progress

DeepMind work in science

- Protein structure prediction (AlphaFold)
- Quantum chemistry (QMC, DFT)
- Genomics
- Weather prediction
- Fusion reactor control
- Lattice QCD
- Glassy dynamics
- Mathematical discovery

Acknowledgements

Thank you to everyone who made AlphaFold possible!

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

The wider team at DeepMind and EMBL-EBI

DeepMind EMBL-EBI

CASP 14

YEARS OF Protein Data Bank

The CASP community **The experimental biology community**

Q&A