DeepMind

AlphaFold 2

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@google.com), Demis Hassabis (dhcontact@google.com)

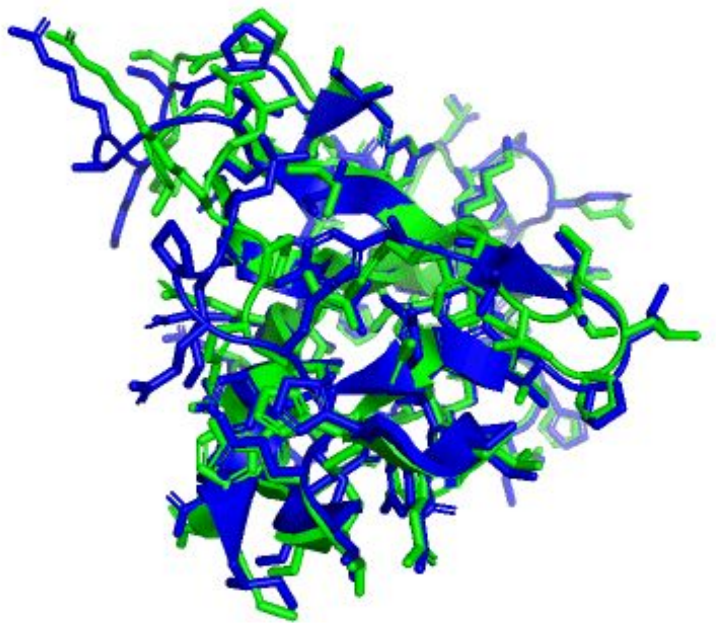
Protein folding at DeepMind

- DeepMind is on a long-term mission to advance scientific progress
- We're interested in solving fundamental scientific problems using AI
- Protein folding is such an important fundamental problem that is well-suited for AI
- We're thankful that CASP is providing such an ideal experimental setup to evaluate progress

Presenting the work of the AlphaFold team

+ Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Martin Steinegger, Michalina Pacholska, David Silver, Oriol Vinyals, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis
& with help from many others from across DeepMind

Protein example: T1064 (ORF8)

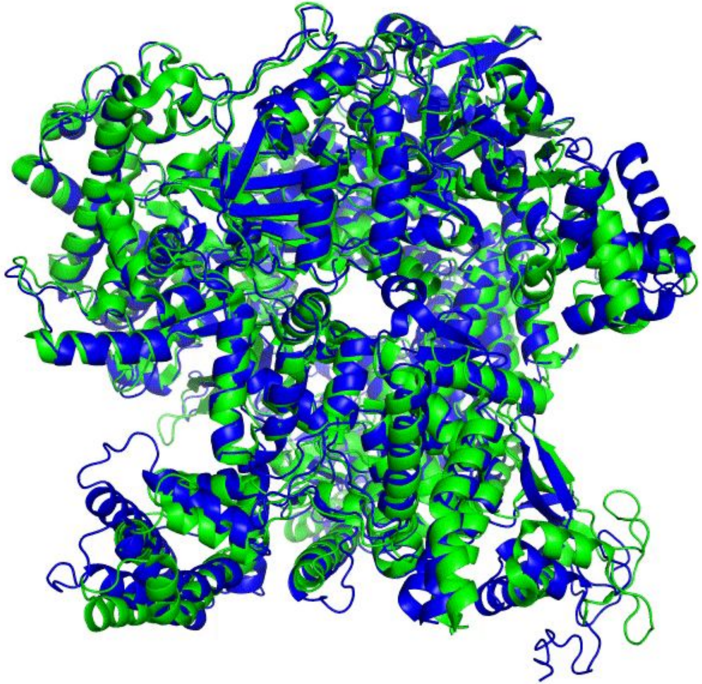


T1064 / 7jtl
87.0 GDT
(ORF8, SARS-CoV-2)

Ground truth
Prediction

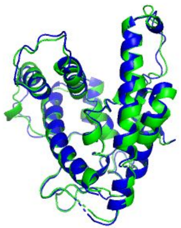
7JTL: Flower, T.G., et al. (2020) Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. Biorxiv.

Protein example: T1044 (RNA Polymerase)

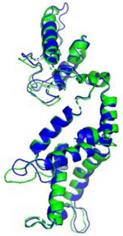


- Folding as a single long chain
- Long-chain-trained model trained after the submission

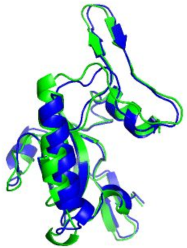
Individual domains



T1041



T1042

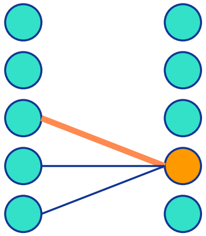


T1043

6VR4: Leiman, P.G., et al. Virion-packaged DNA-dependent RNA polymerase of crAss-like phage phi14:2 (CASP target). (To be published.)

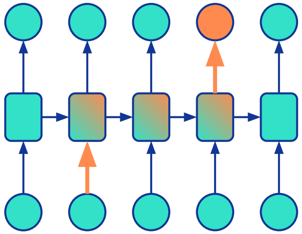
Ground truth
Prediction

Inductive Bias for Deep Learning Models



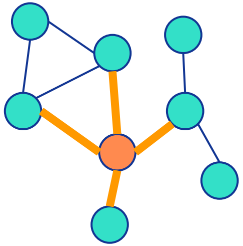
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours



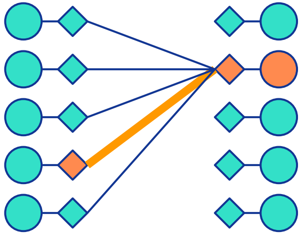
Recurrent Networks (e.g. language)

- 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

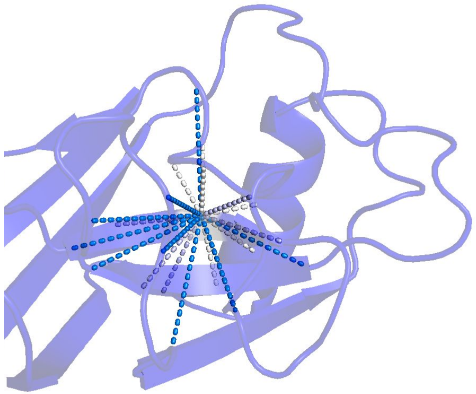


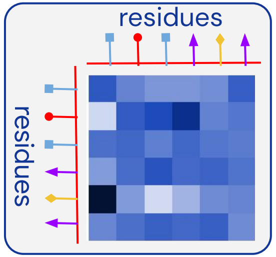
Attention Module (e.g. language)

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

Putting our protein knowledge into the model

- Physical 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





DeepMind

System Design

Inputs

Sequence databases

- UniRef90[6] (JackHMMER[3])
- BFD[5] (HHblits[4])
- MGnify clusters[2] (JackHMMER[3])

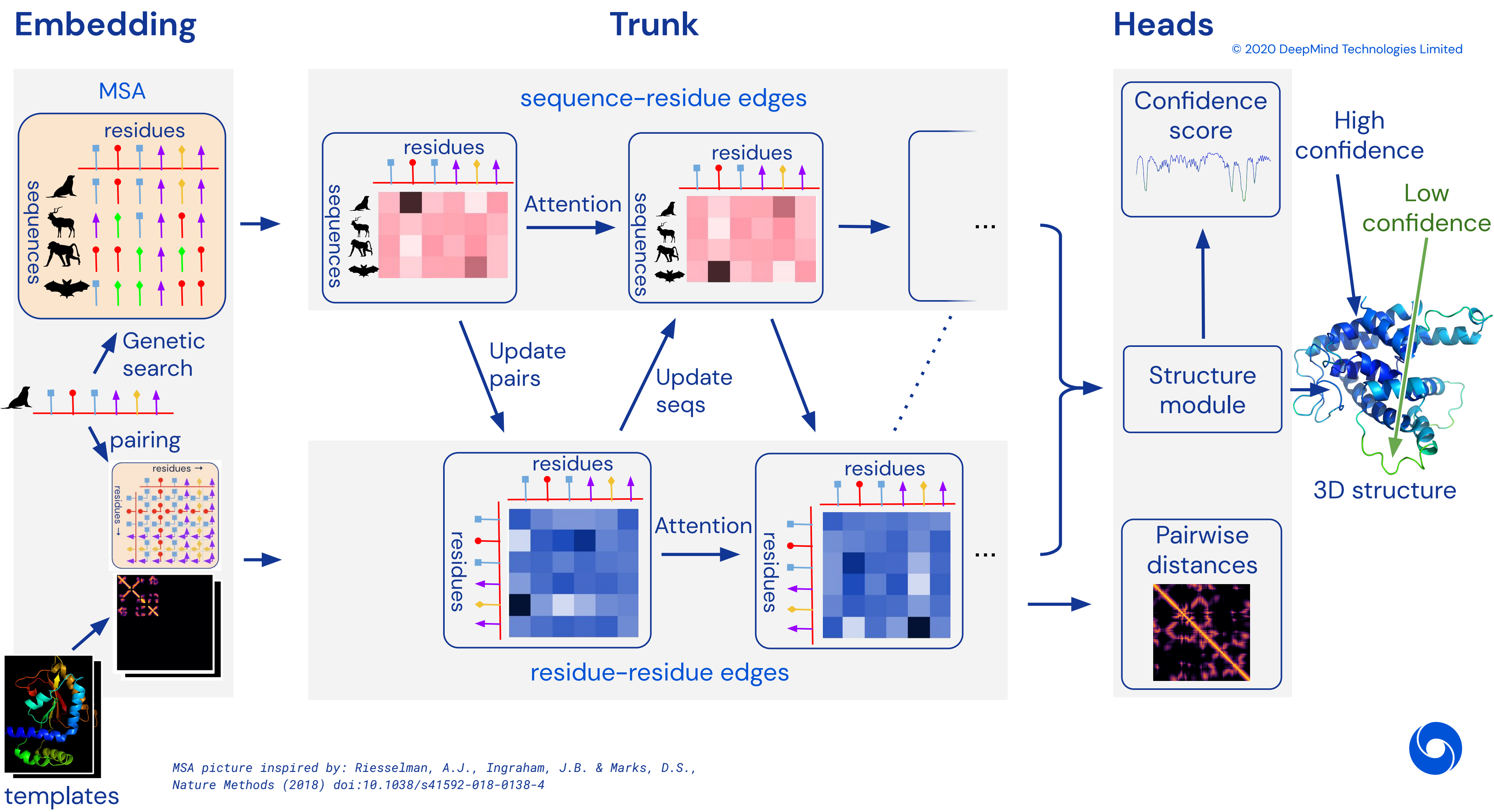
Structural databases

- PDB[1] (training)
- PDB70 clustering (hhsearch[4])

All publicly available data.

[1] *Berman et al.,* Nature Structural Biology (2003) doi:10.1038/nsb1203-980
[2] *Mitchell et al.,* Nucleic Acids Research (2019) doi:10.1093/nar/gkz1035
[3] *Potter et al.,* Nucleic Acids Research (2018) doi:10.1093/nar/gky448
[4] *Steinegger et al.,* BMC Bioinformatics (2019) doi:10.1186/s12859-019-3019-7
[5] *Steinegger et al.,* Nature Methods (2019) doi:10.1038/s41592-019-0437-4
[6] *Suzek et al.,* Bioinformatics (2015) doi:10.1093/bioinformatics/btu739

Visualisations:
The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
AS Rose, et al., Bioinformatics (2018) doi:10.1093/bioinformatics/bty419

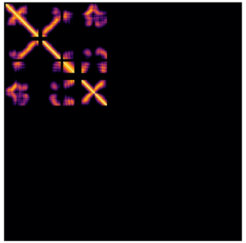


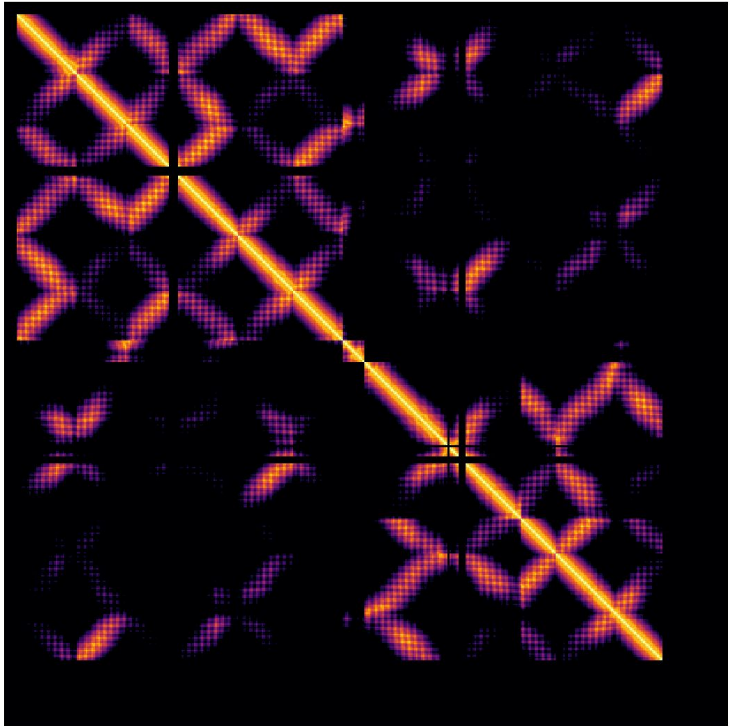
MSA picture inspired by: Riesselman, A.J., Ingraham, J.B. & Marks, D.S., Nature Methods (2018) doi:10.1038/s41592-018-0138-4

Template embedding

- 4 templates used (from PDB70 clusters, searched with HHsearch[1,2])
- Input features are sequences, side chains, and distograms
- Templates are processed in the same way as the residue-residue representation

Partial template:





[1] *Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9(2), 173-175.*
[2] *Steinegger, M. et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics, 20(1), 1-15.*

Structure module

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

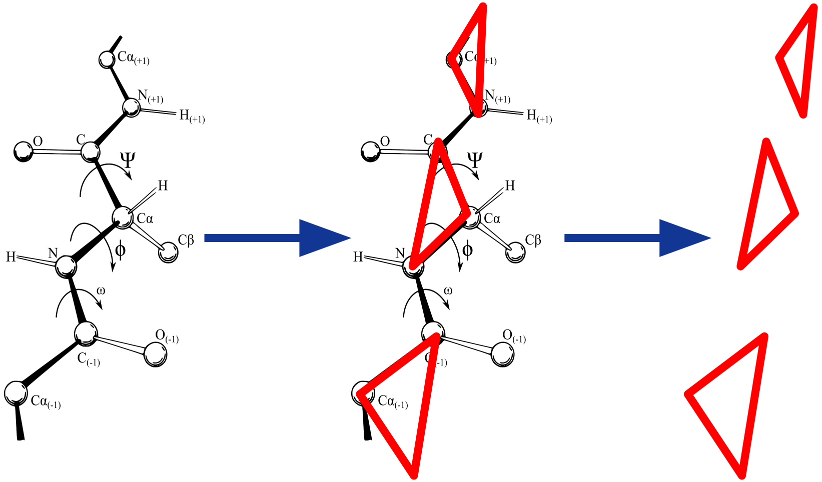
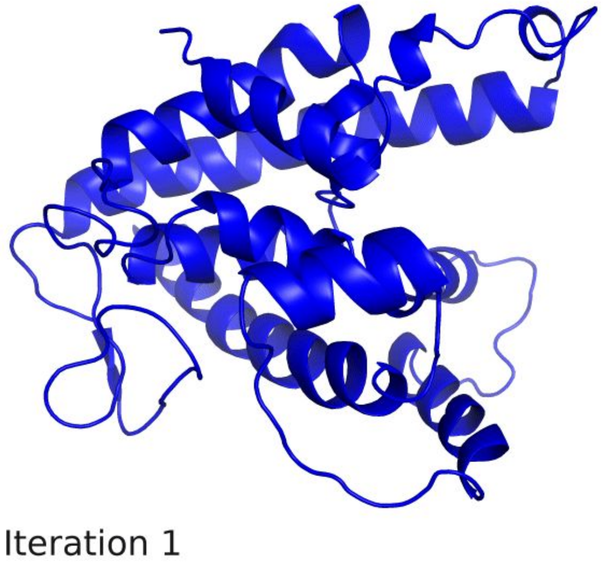


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

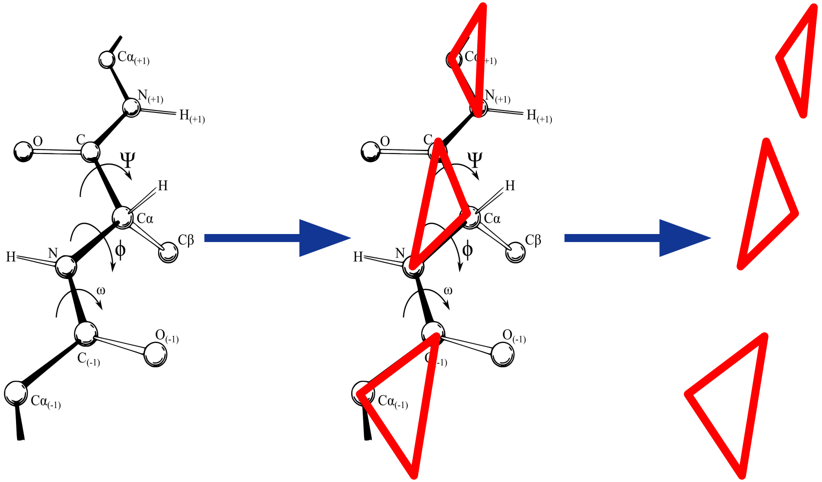
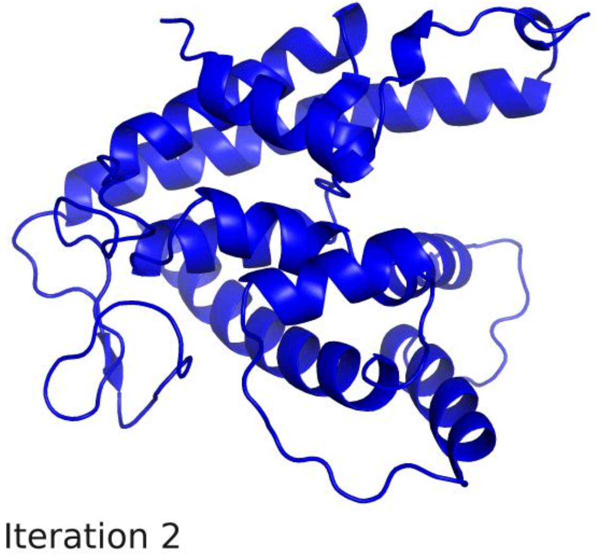


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

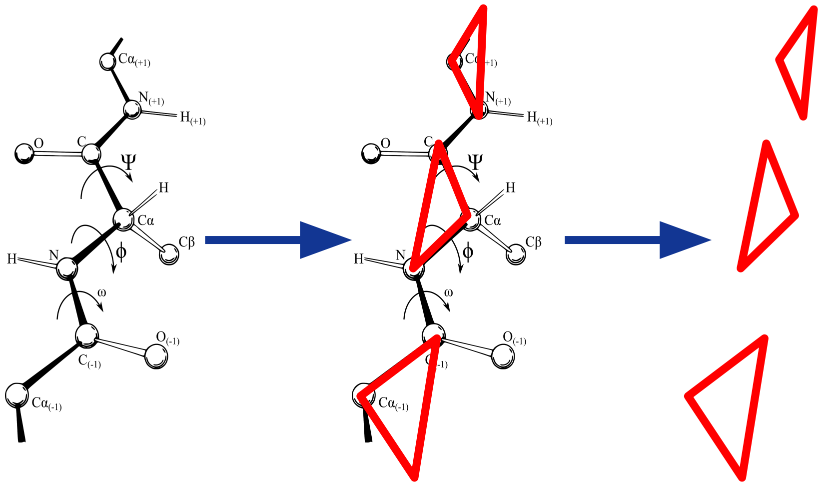
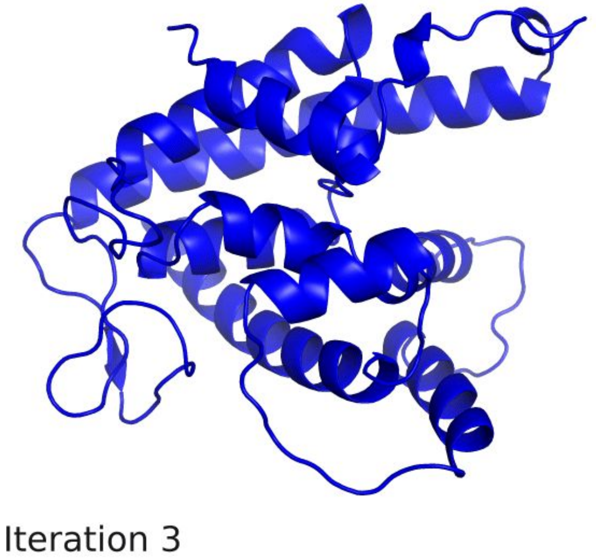


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

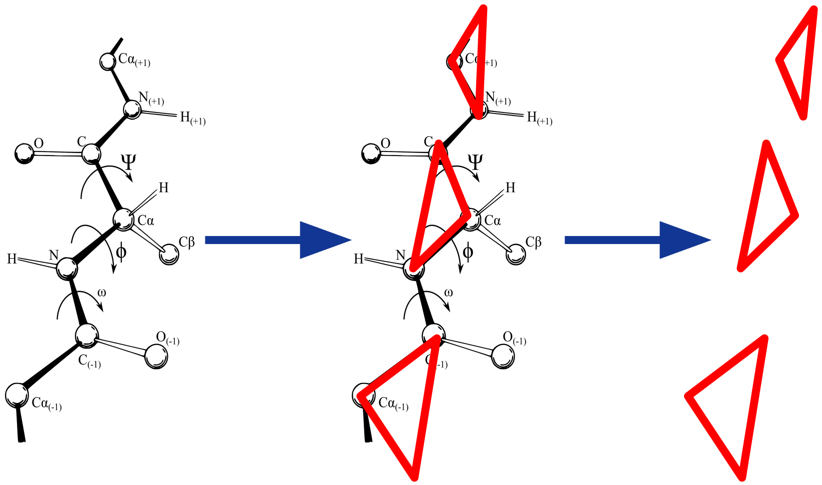
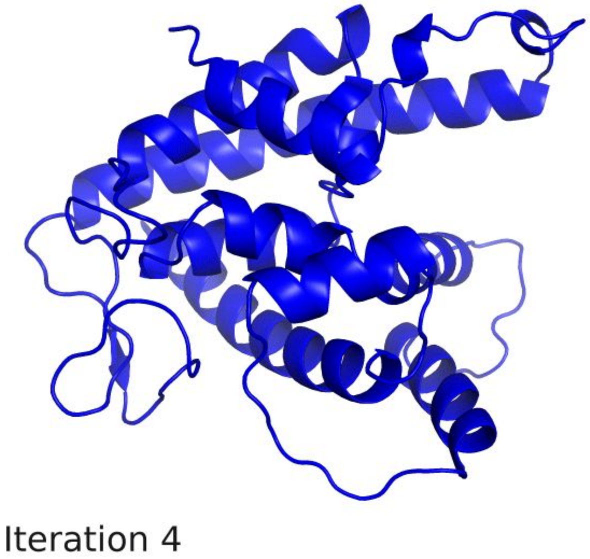


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

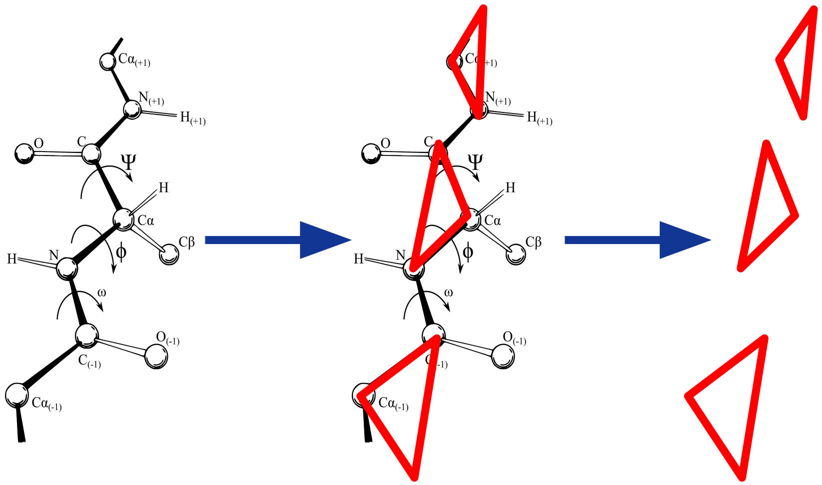
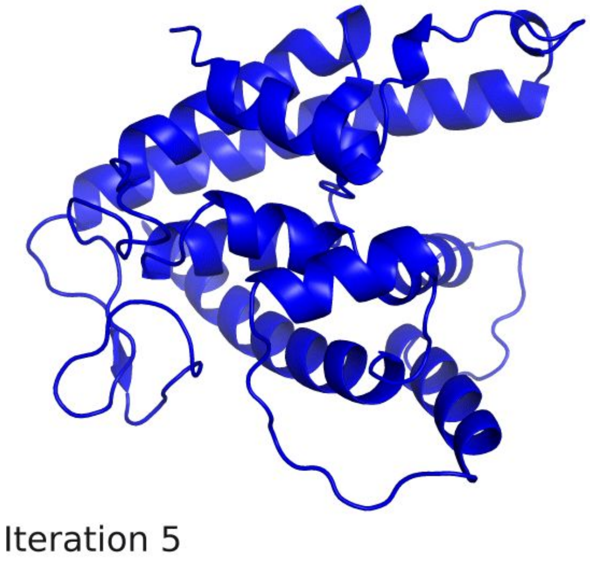


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

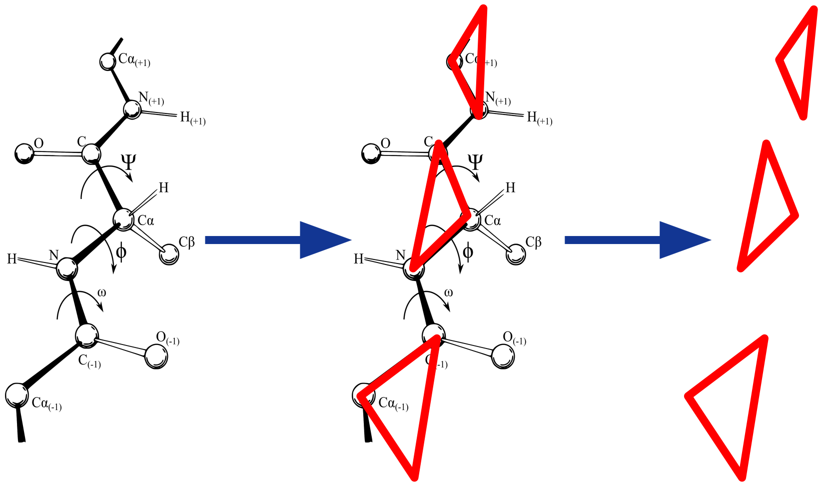
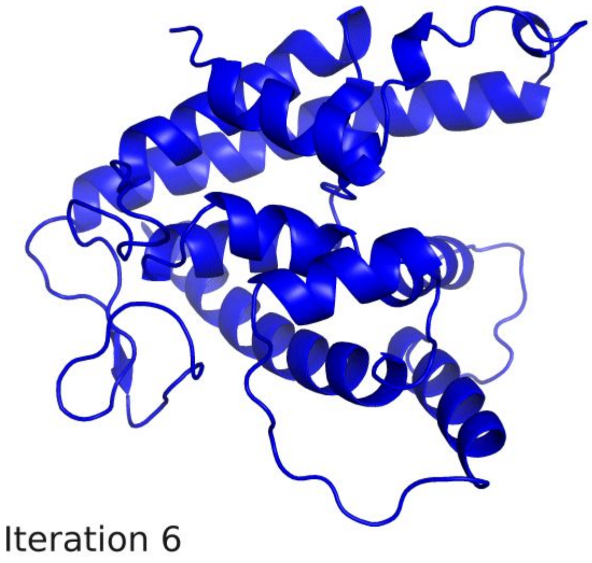


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

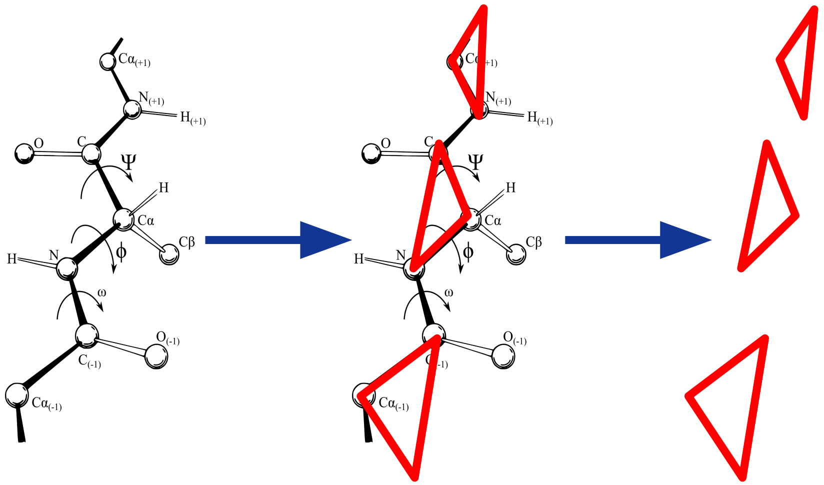
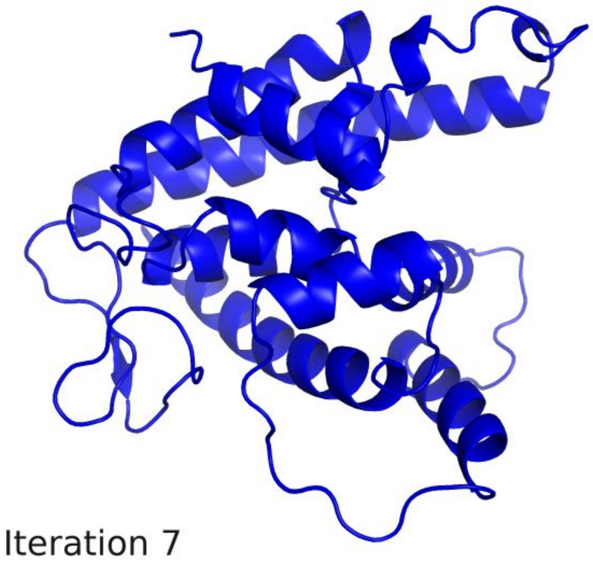


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

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



Target: T1041

Structure module

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

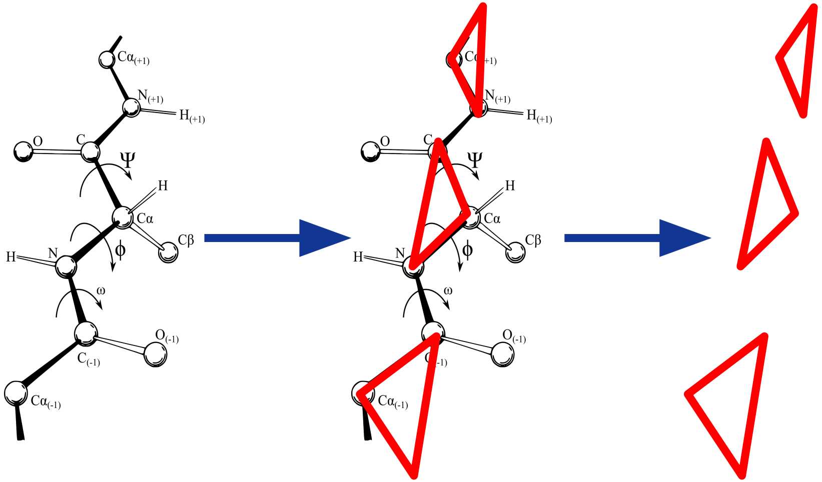
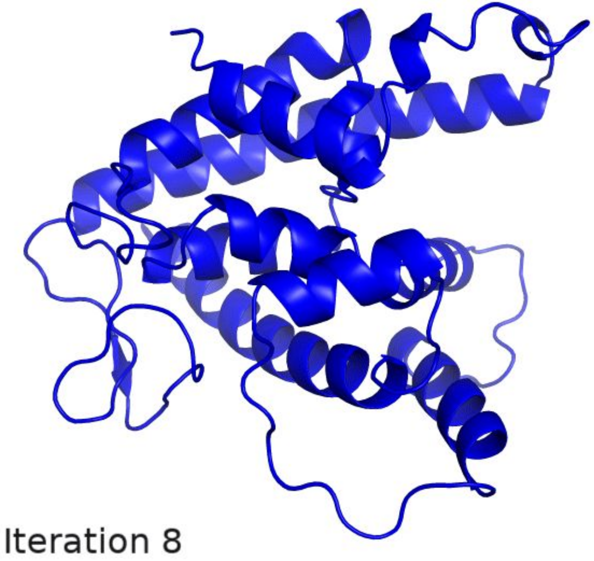


Image: Dcrjsr, vectorised Adam Rędzikowski (CC BY 3.0, Wikipedia)

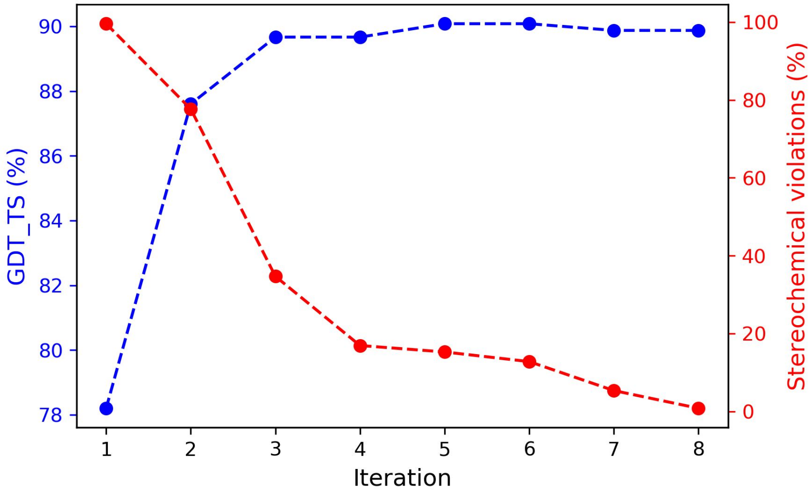
- **3-D equivariant transformer architecture** updates the rigid bodies / backbone
  - Also builds the side chains



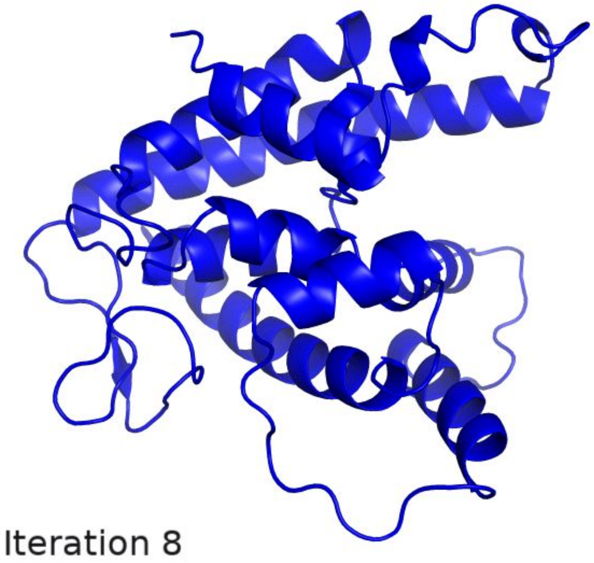
Target: T1041

Refinement in structure module

Improves both accuracy and stereochemical quality



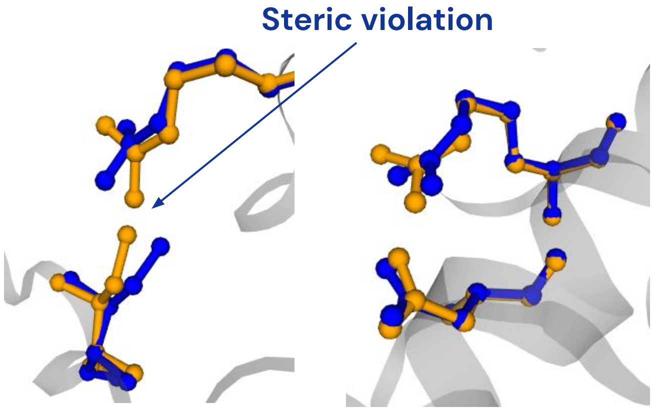
Target: T1041



Target: T1041

Relaxation

- The end result of iterative refinement is not guaranteed to obey all stereochemical constraints
- Violations of these constraints are resolved with coordinate-restrained gradient descent
- We use the Amber ff99SB force field[1] with OpenMM[2]

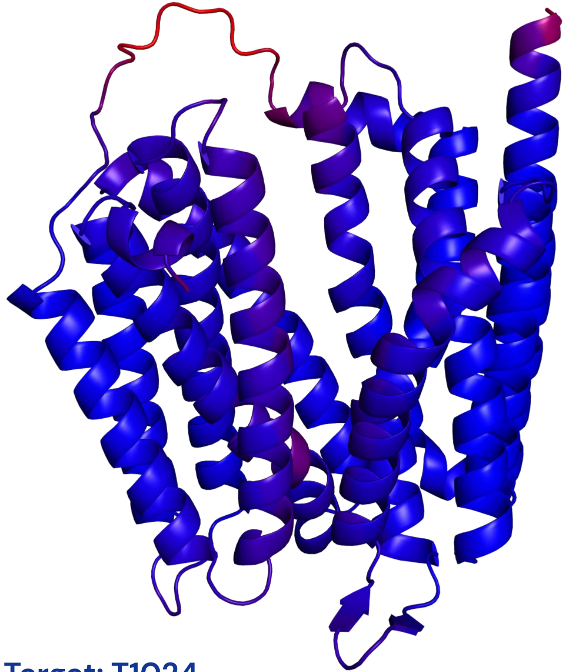


Orange: pre-relax
Blue: post-relax

[1] Hornak, V. et al. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins: Structure, Function, and Bioinformatics, 65(3), 712-725.
[2] Eastman, P. et al. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Computational Biology, 13(7), e1005659.

Knowing where we are right

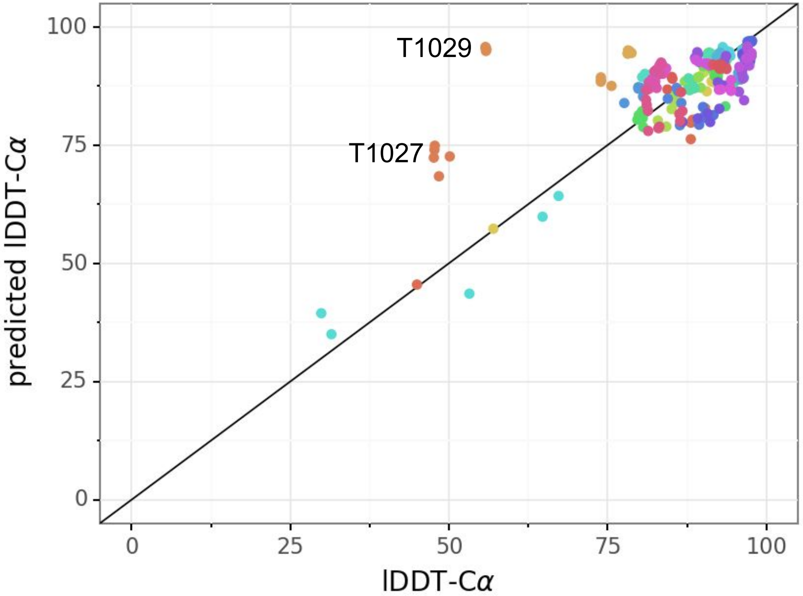
lDDT-Cα prediction from the last layer of the structure module



Target: T1024

Confidence calibration on CASP14 chains

Median absolute error: 3.3 LDDT-Cα



Five models per chain, coloured by chain
Excluding T1044 domains, T1088

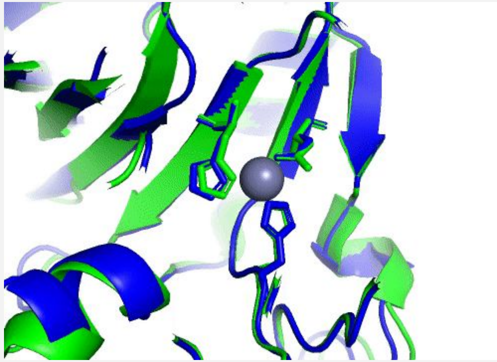
DeepMind

How AlphaFold understands proteins

Biological context

- Computational structure prediction is typically underspecified
    - Oligomeric state, ligands, DNA-binding, experimental conditions, multiple conformations etc.
- Our networks implicitly models the missing context
- Uses a variety of physical and evolutionary information (e.g. profile-only is still pretty accurate)

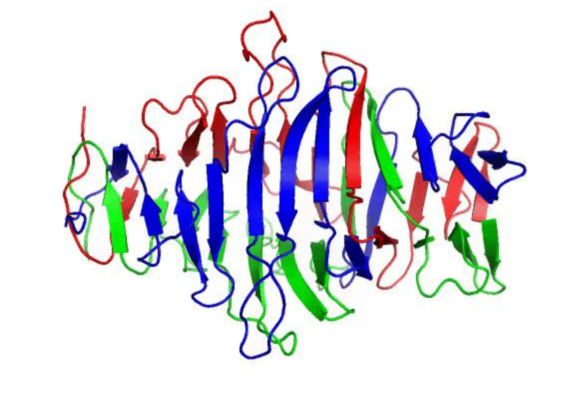
T1056 (zinc binding)



AlphaFold / Experiment

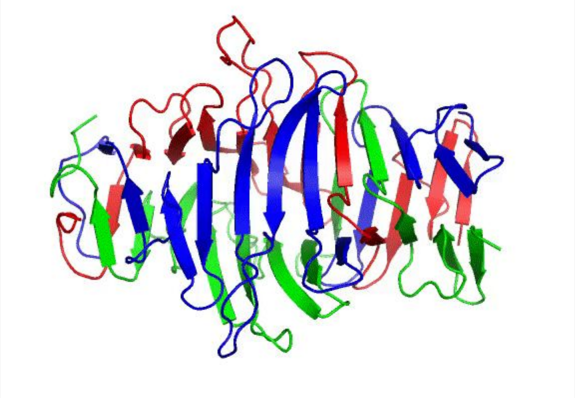
TBM-hard, 98.2 GDT

T1080 (trimer)



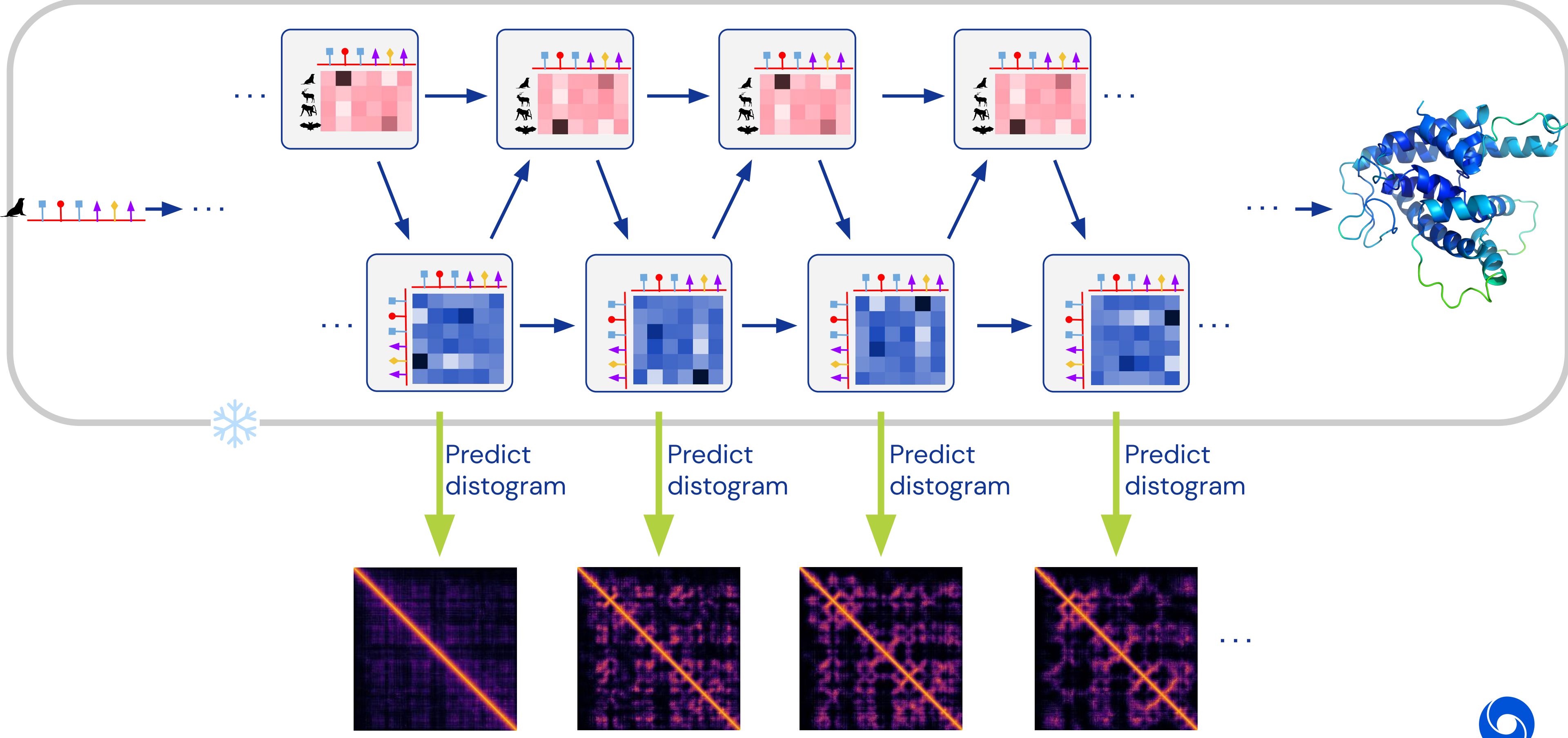
AlphaFold (monomer prediction x3)

FM/TBM, 85.9 GDT

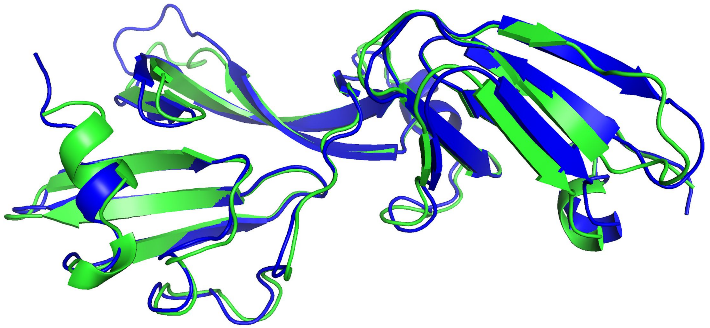


Experimental structure

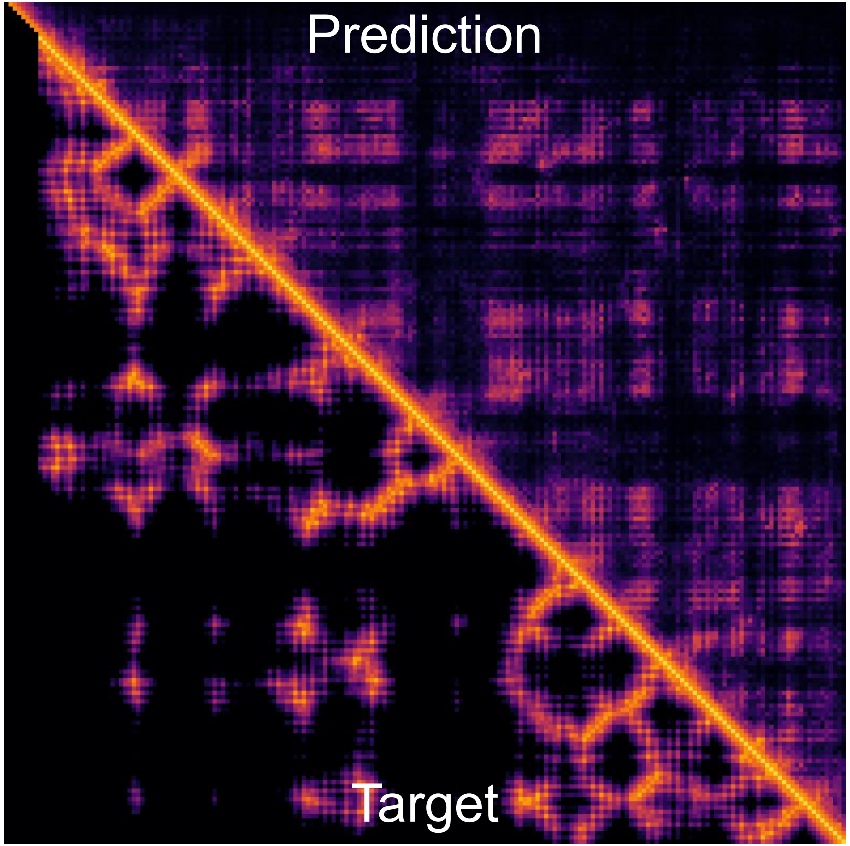
Interrogating the Network



Model interpretability - T1038

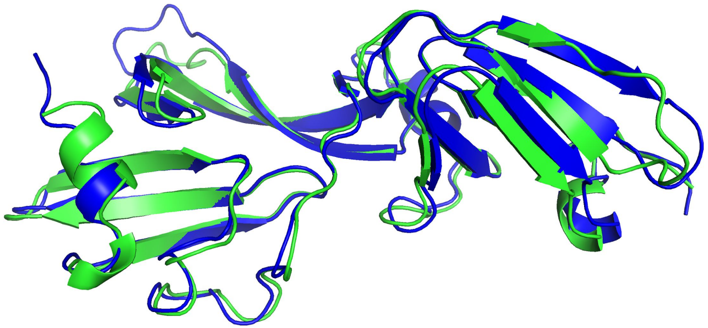


T1038

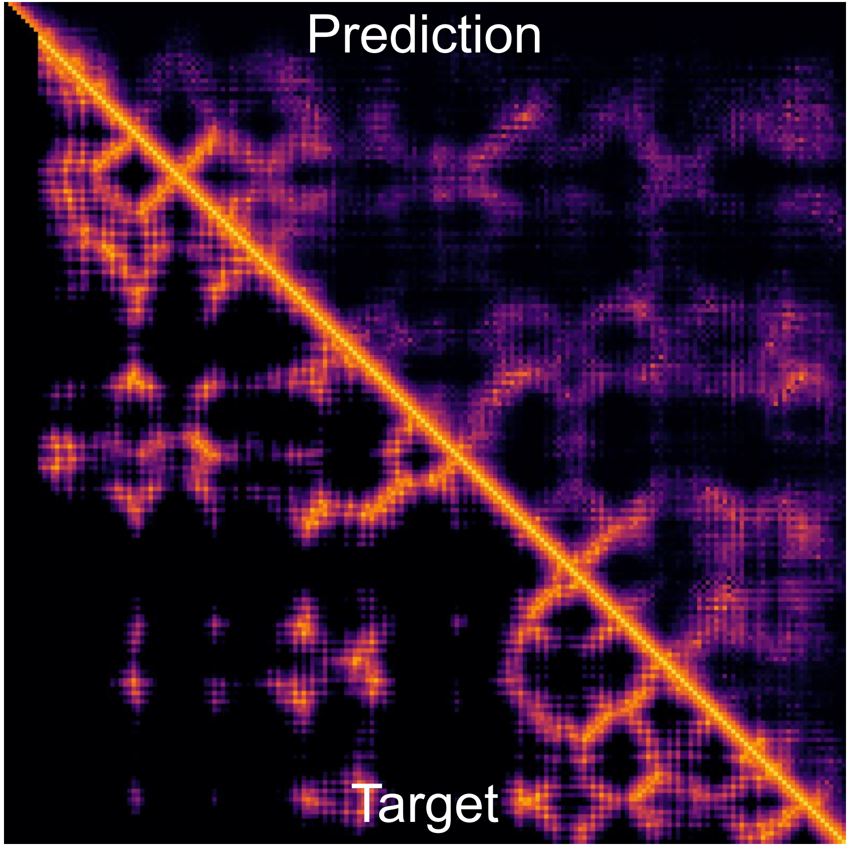


6YA2: Bahat, Y., et al. First structure of a glycoprotein from enveloped plant virus. (To be published.)

Model interpretability - T1038

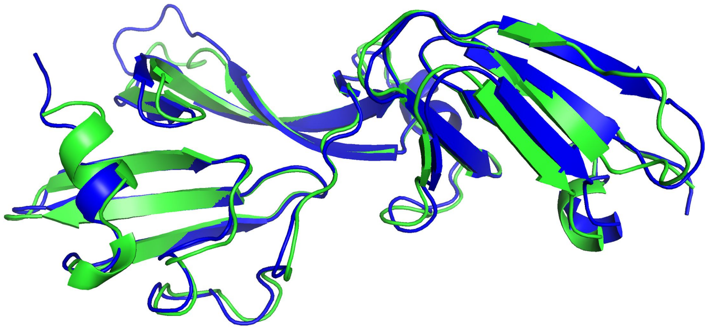


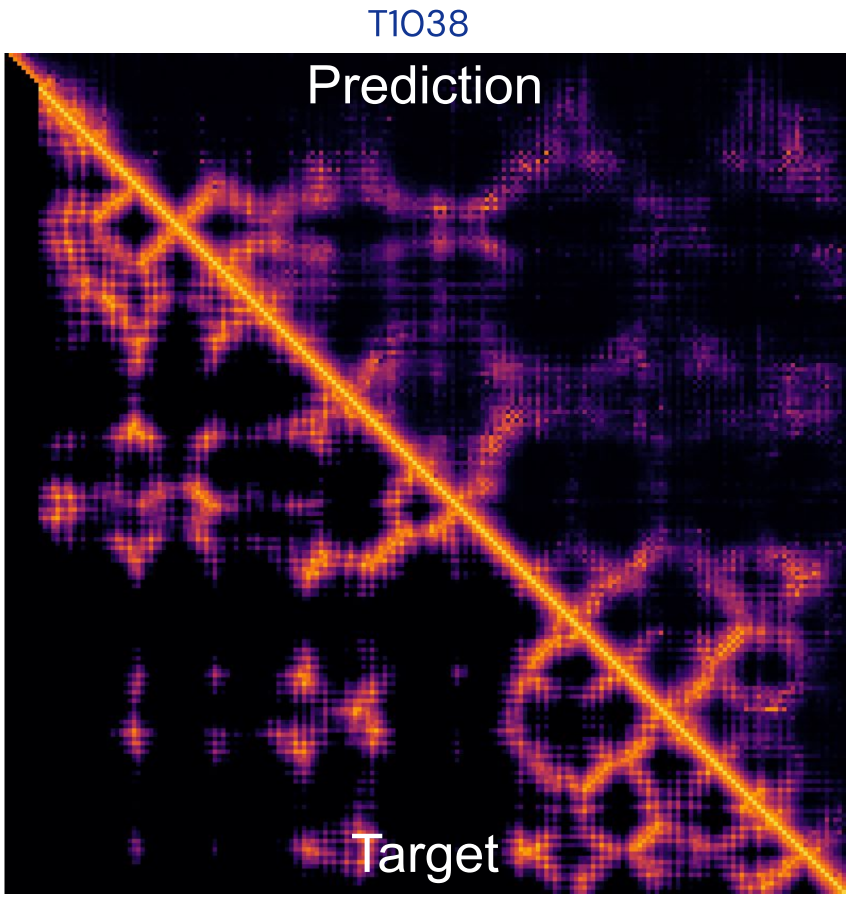
T1038



6YA2: Bahat, Y., et al. First structure of a glycoprotein from enveloped plant virus. (To be published.)

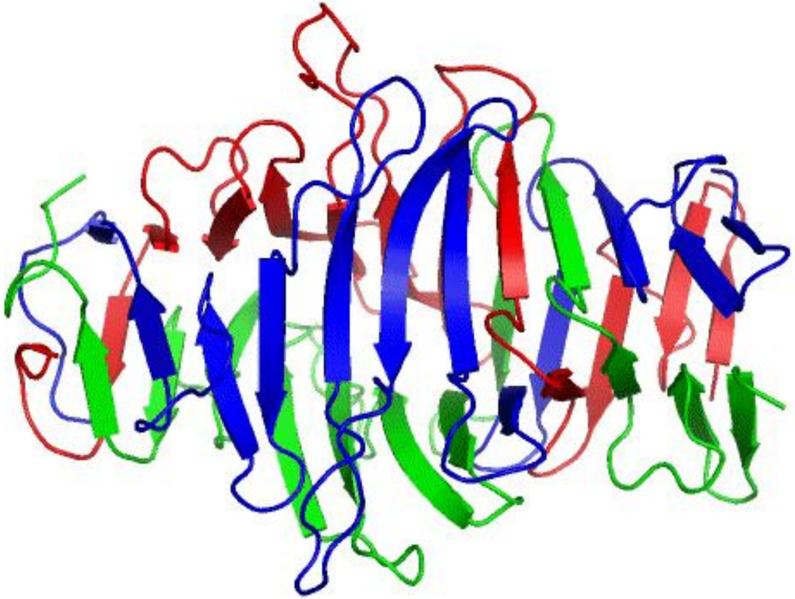
Model interpretability - T1038





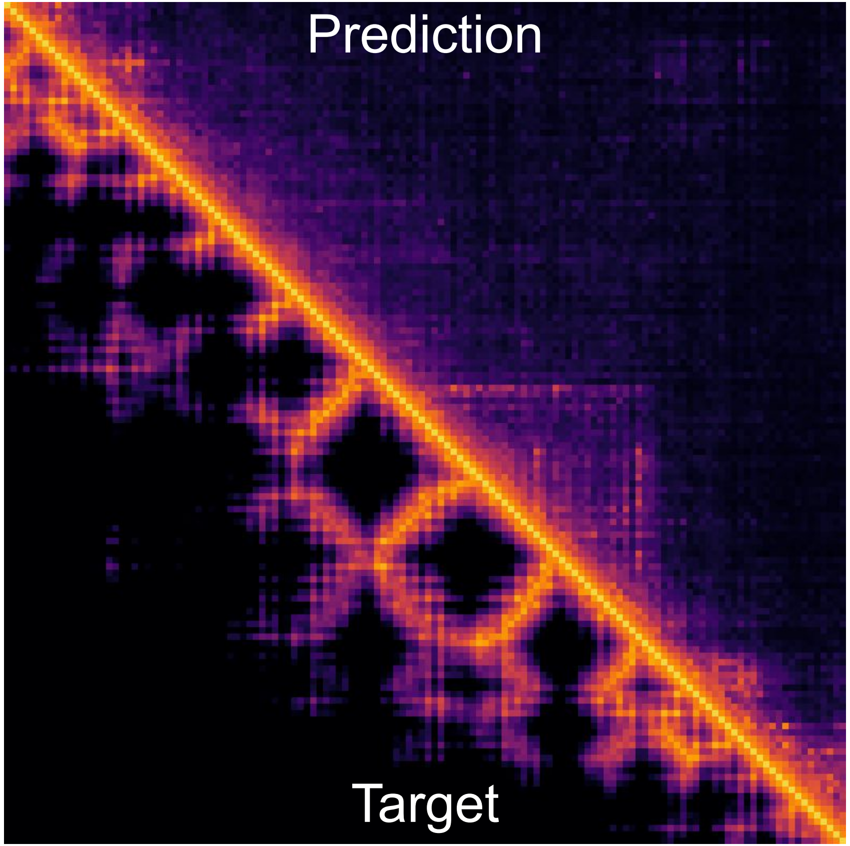
6YA2: Bahat, Y., et al. First structure of a glycoprotein from enveloped plant virus. (To be published.)

Model interpretability - T1080

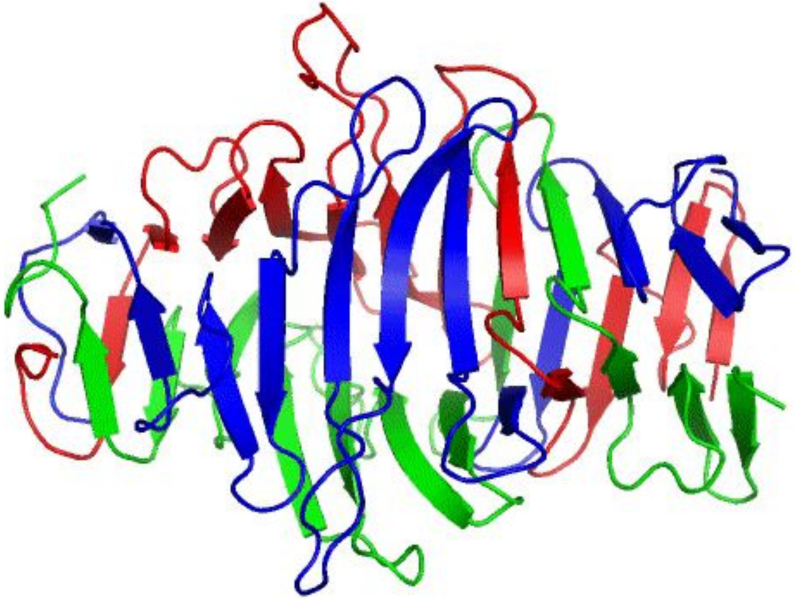


T1080: Not yet in PDB

T1080

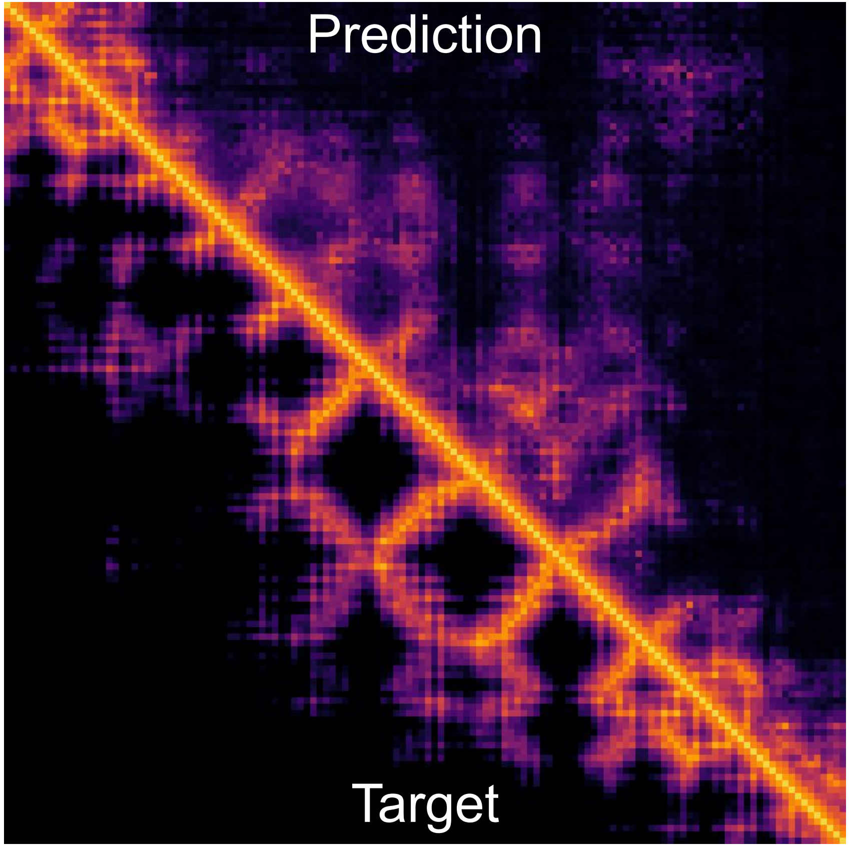


Model interpretability - T1080

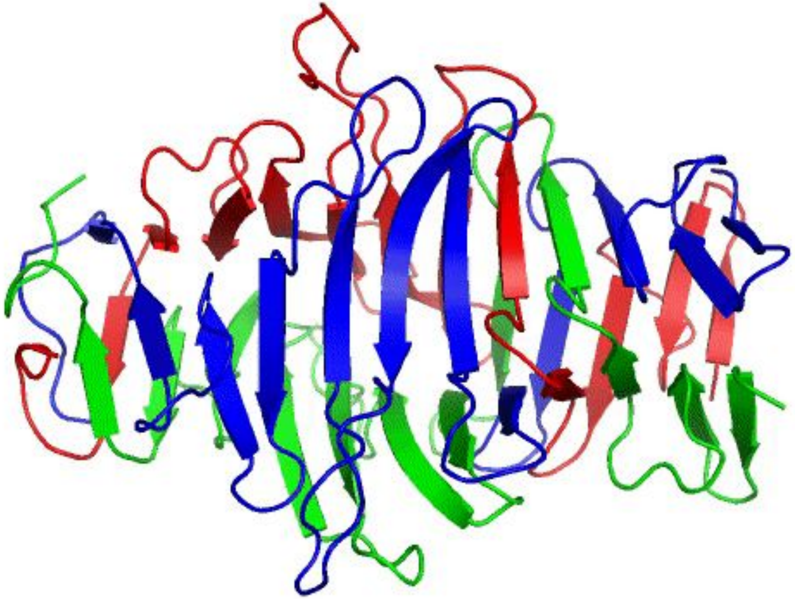


T1080: Not yet in PDB

T1080

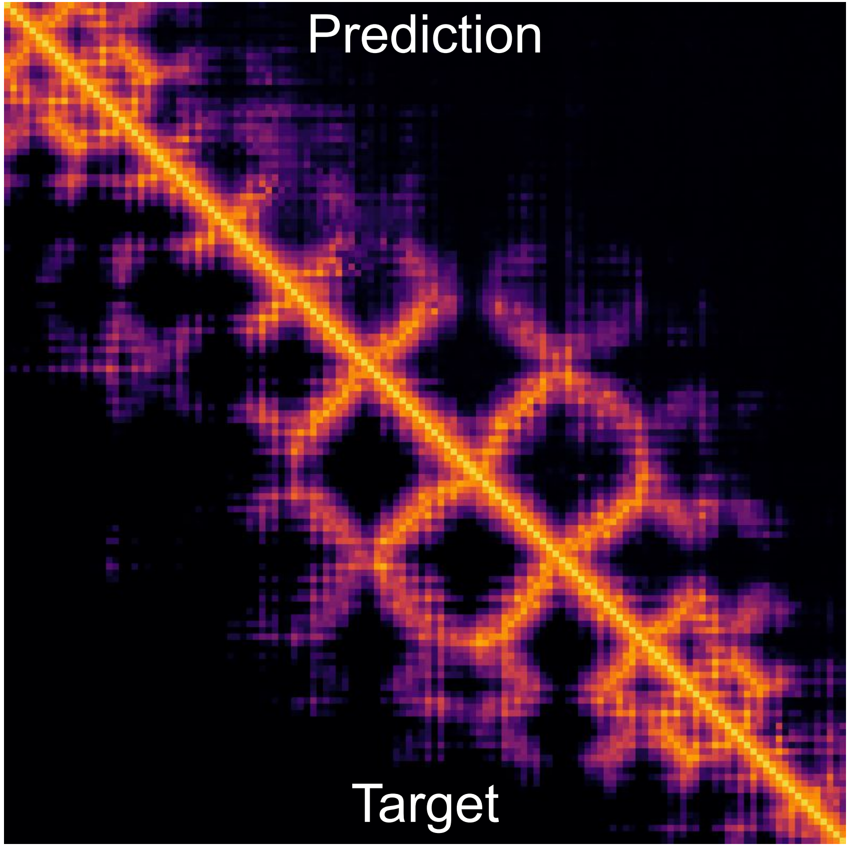


Model interpretability - T1080

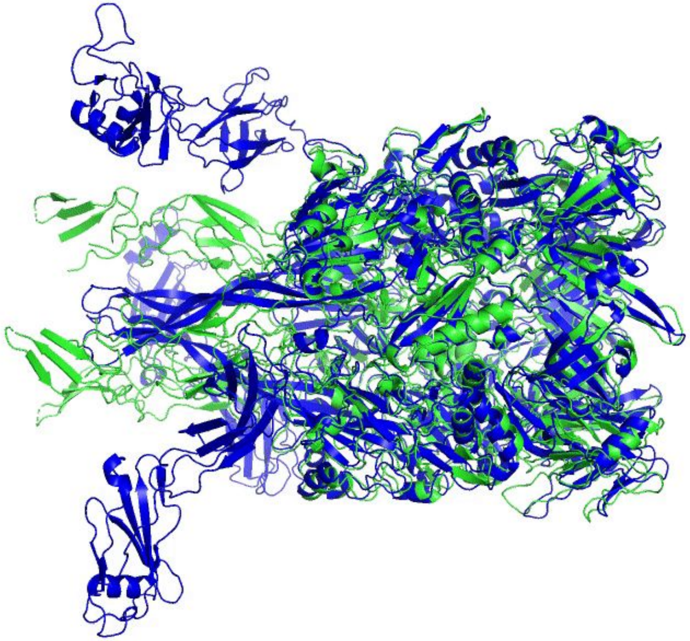


T1080: Not yet in PDB

T1080

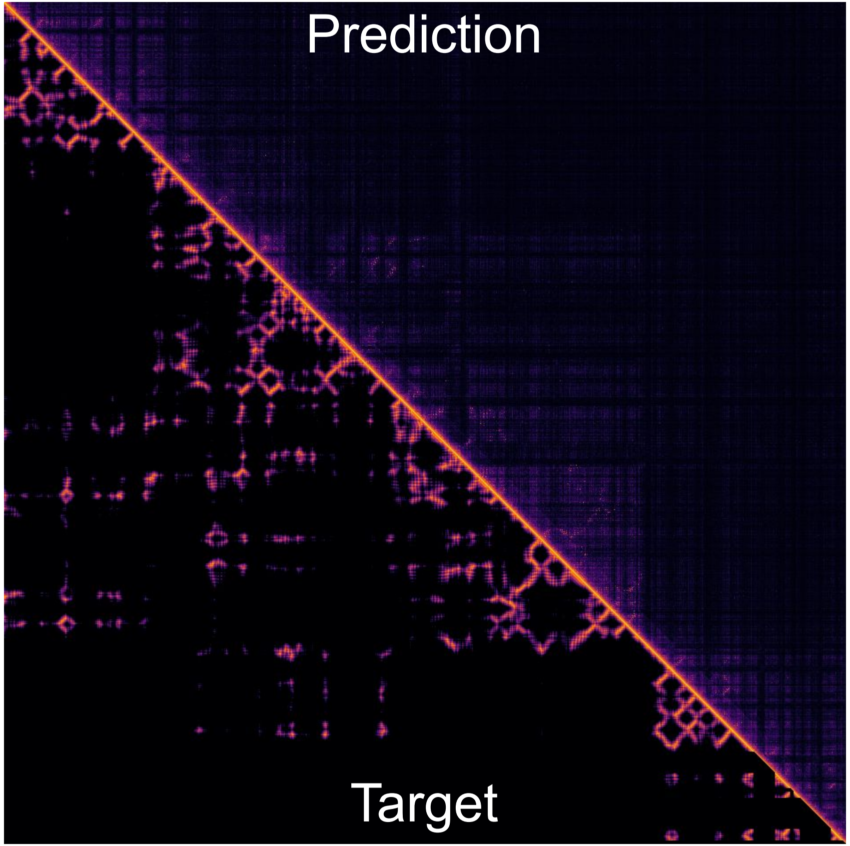


Model interpretability - T1061

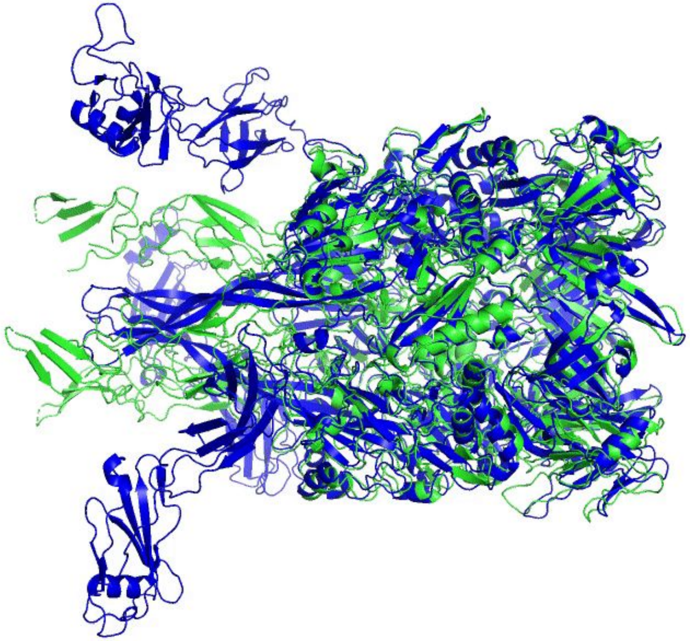


T1061: Not yet in PDB
3 copies of monomer prediction overlaid on crystal

T1061

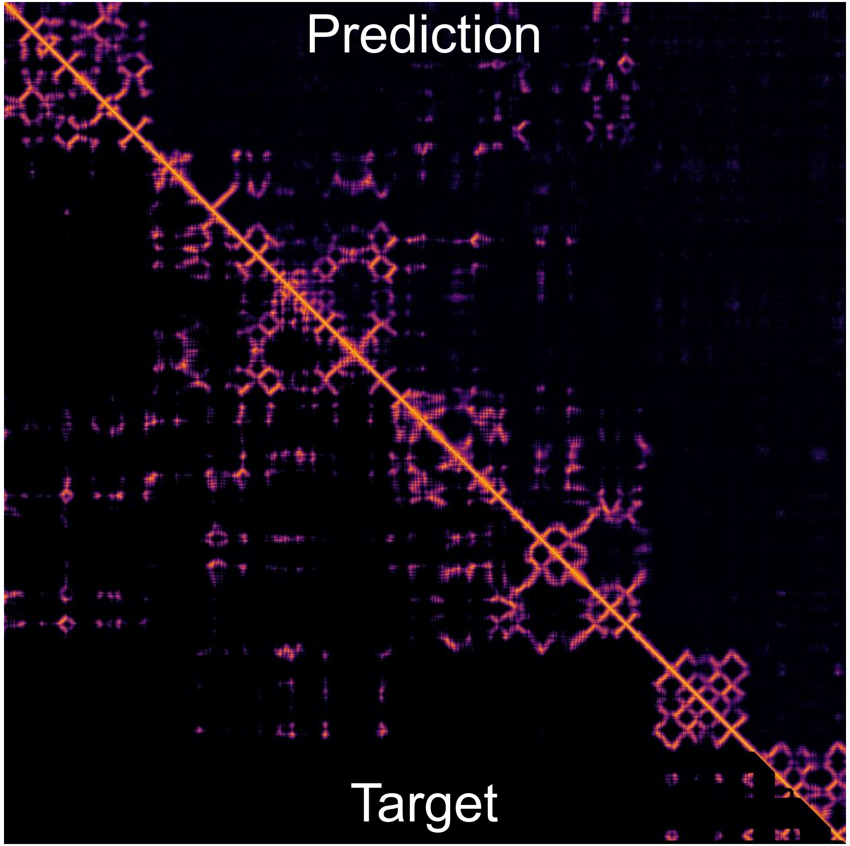


Model interpretability - T1061

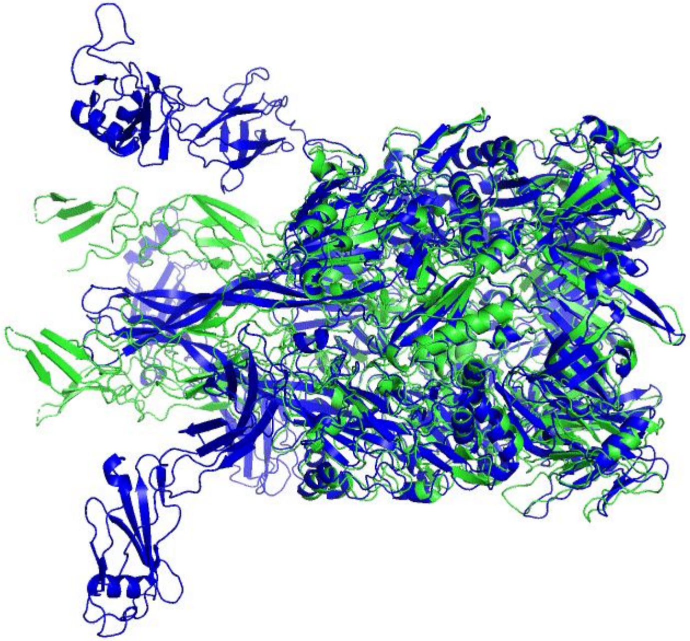


T1061: Not yet in PDB
3 copies of monomer prediction overlaid on crystal

T1061

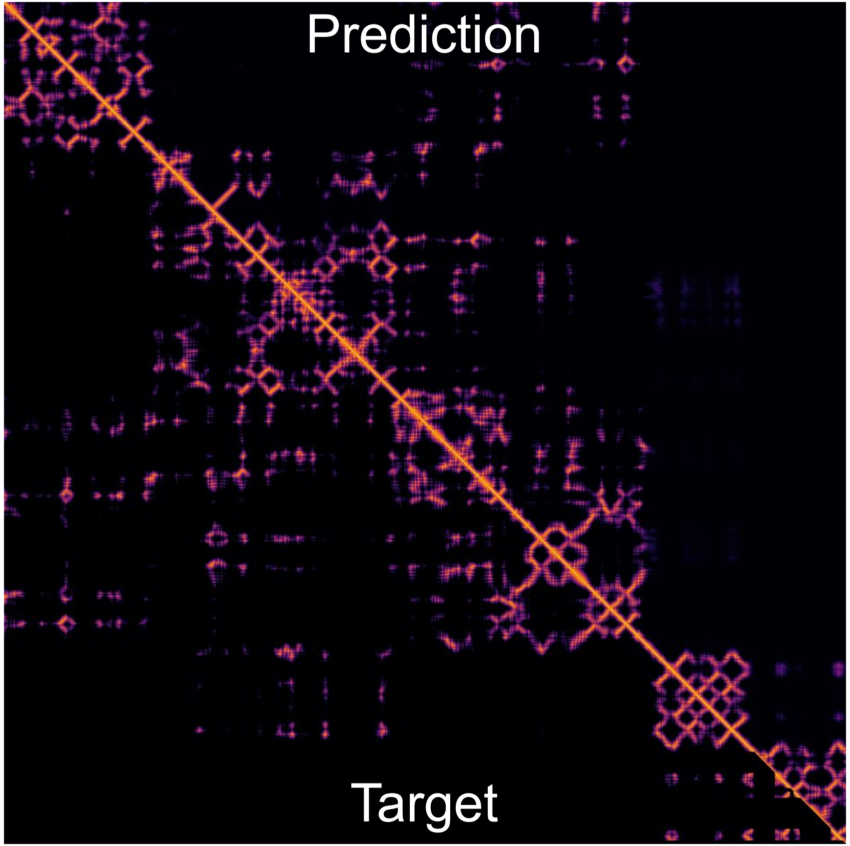


Model interpretability - T1061

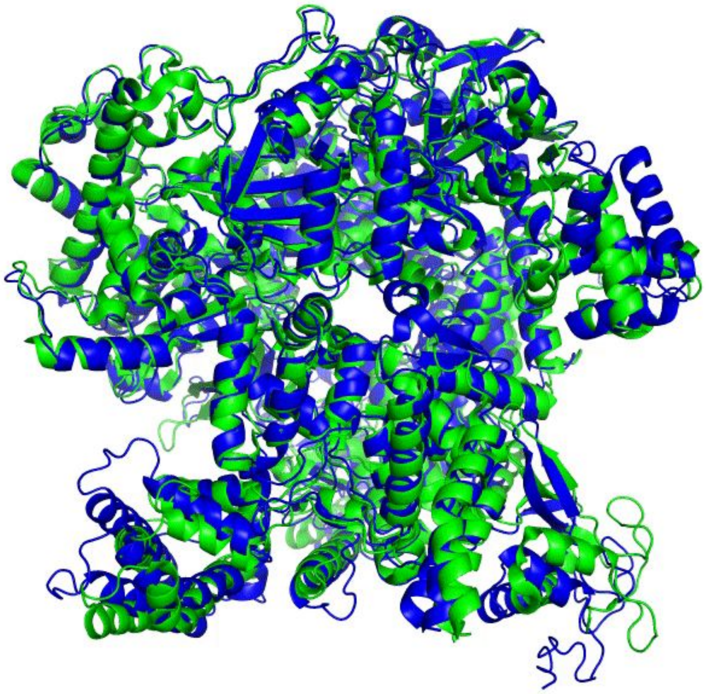


T1061: Not yet in PDB
3 copies of monomer prediction overlaid on crystal

T1061

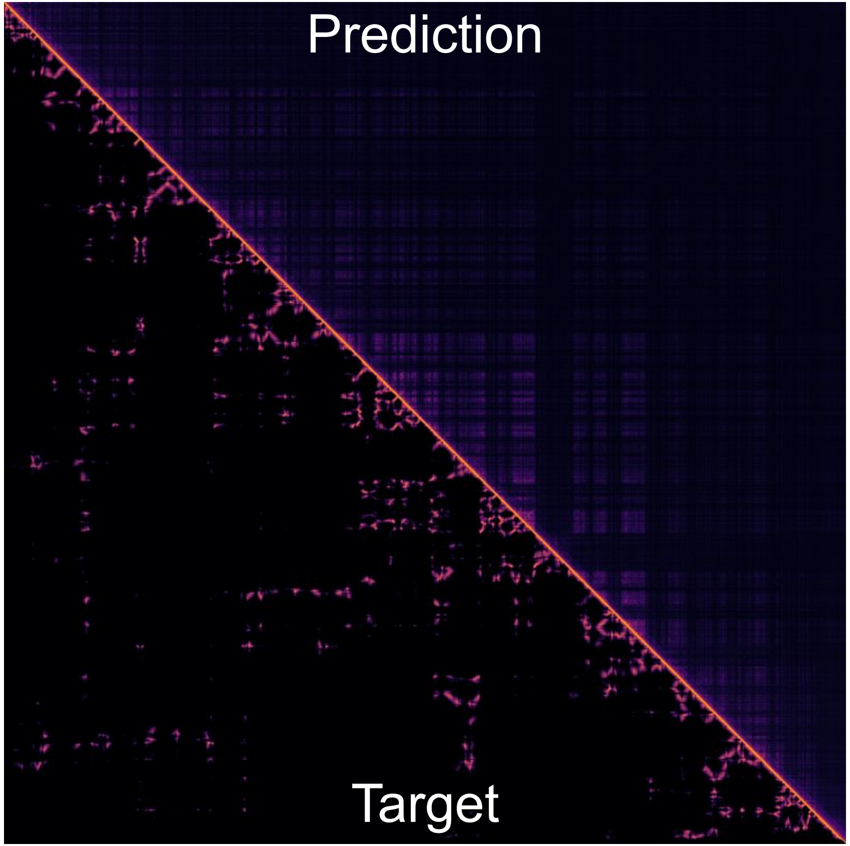


Model interpretability - T1044

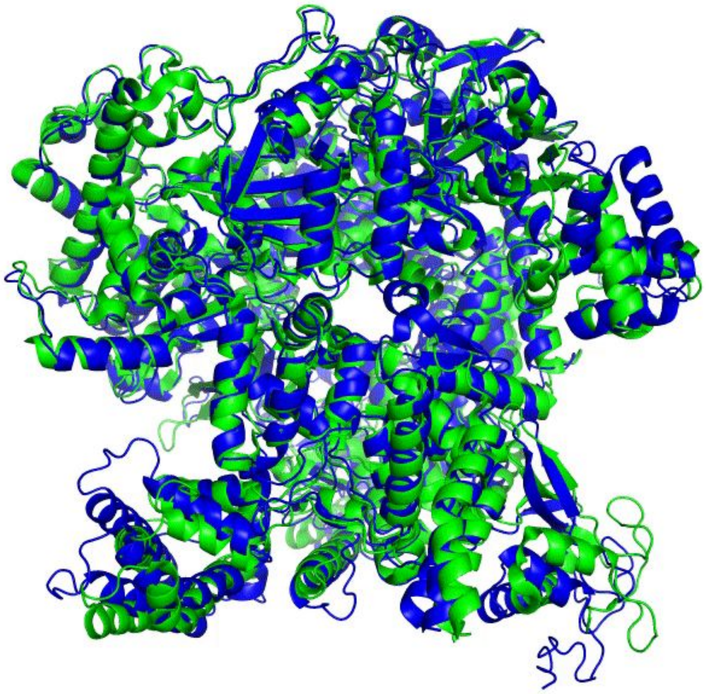


6VR4: Leiman, P.G., et al. Virion-packaged DNA-dependent RNA polymerase of crAss-like phage phi14:2 (CASP target). (To be published.)

T1044

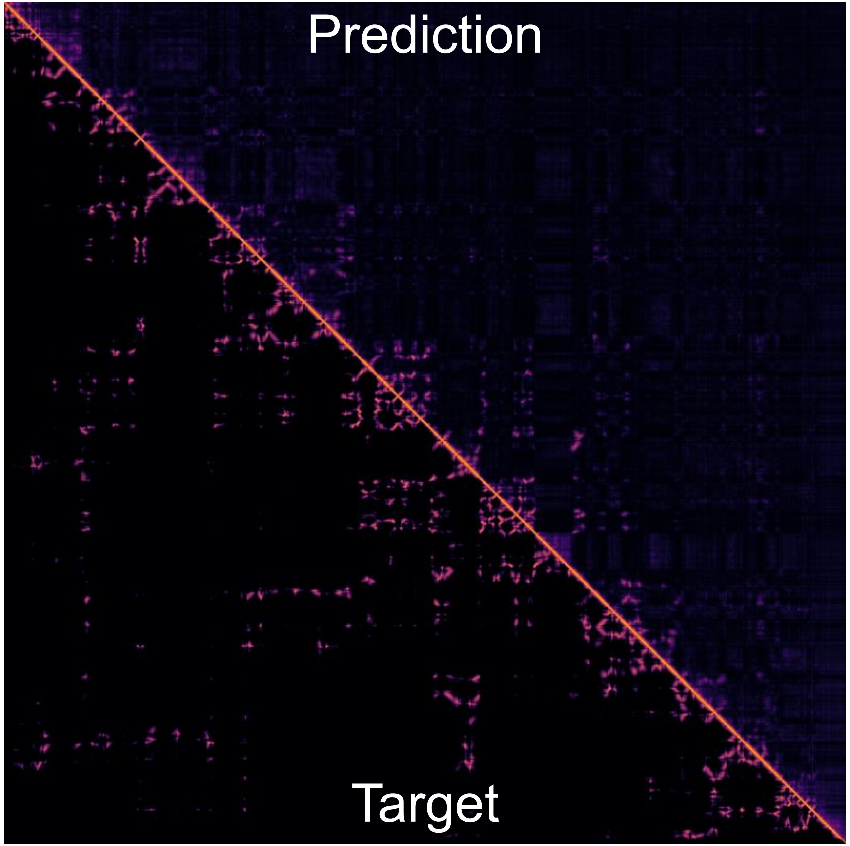


Model interpretability - T1044

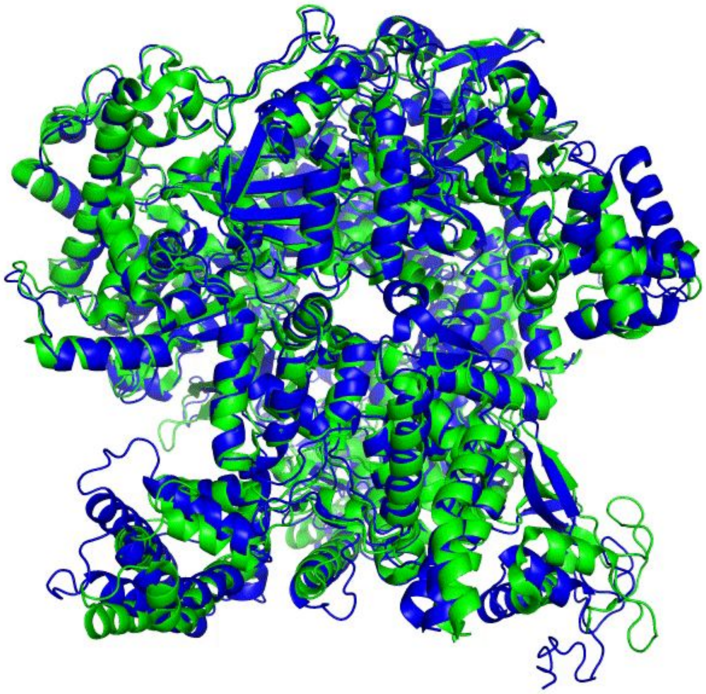


6VR4: Leiman, P.G., et al. Virion-packaged DNA-dependent RNA polymerase of crAss-like phage phi14:2 (CASP target). (To be published.)

T1044

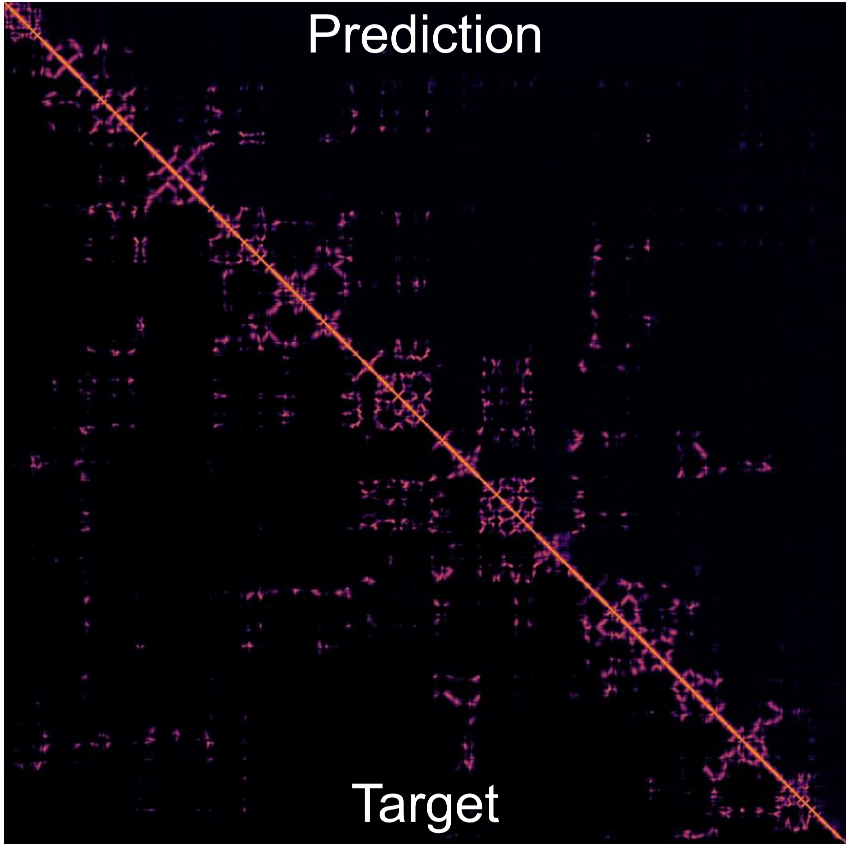


Model interpretability - T1044



6VR4: Leiman, P.G., et al. Virion-packaged DNA-dependent RNA polymerase of crAss-like phage phi14:2 (CASP target). (To be published.)

T1044



Manual interventions

We learned a lot during CASP14!

- **Domains arising from H1044 (RNA polymerase):**
  - Genetics search of full chain but folded in 4 parts
  - Resulting pieces were used as templates to build the full chain
  - Afterward, we fine-tuned our models to handle very long chains
  - Can now obtain this accuracy in a fully-automated way
- **T1064 (ORF8)**
  - Five additional sequences were added to the MSA using NCBI Protein BLAST
  - Tried more models to find a confident one
- **T1024 (Multidrug transporter)**
  - Clustered templates into different classes to get diversity of opening angle
- **Additional targets:**
  - Often the model diversity is low despite the error scores saying that there is error
  - We would try to put older models in later positions to increase diversity

What went badly

- Manual work required to get a very high-quality Orf8 prediction
- Genetics search works much better on full sequences than individual domains
- Final relaxation required to remove stereochemical violations

What went well

- Building the full pipeline as a single end-to-end deep learning system
- Building physical and geometric notions into the architecture instead of a search process
- Models that predict their own accuracy can be used for model-ranking
- Using model uncertainty as a signal to improve our methods (e.g. training new models to eliminate problems with long chains)

Wrap up & future outlook

- We have built a system that confidently predicts accurate structures for most proteins – and knows when it is wrong

- As for CASP13[1,2], we'll publish a peer-reviewed paper
- We're also working on providing broad access to our work
- Demis Hassabis will be giving a keynote on Friday about *Using AI to accelerate scientific discovery*

- Lots of exciting work ahead for the field: Complexes, conformational change etc
- Thanks again to the CASP organizers, experimentalists and everyone on whose work we're building

[1] Senior, A. W., et al. "Improved protein structure prediction using potentials from deep learning." Nature 577.7792 (2020): 706-710.
[2] Senior, A. W., et al. "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)." Proteins 87.12 (2019): 1141-1148.

DeepMind

End