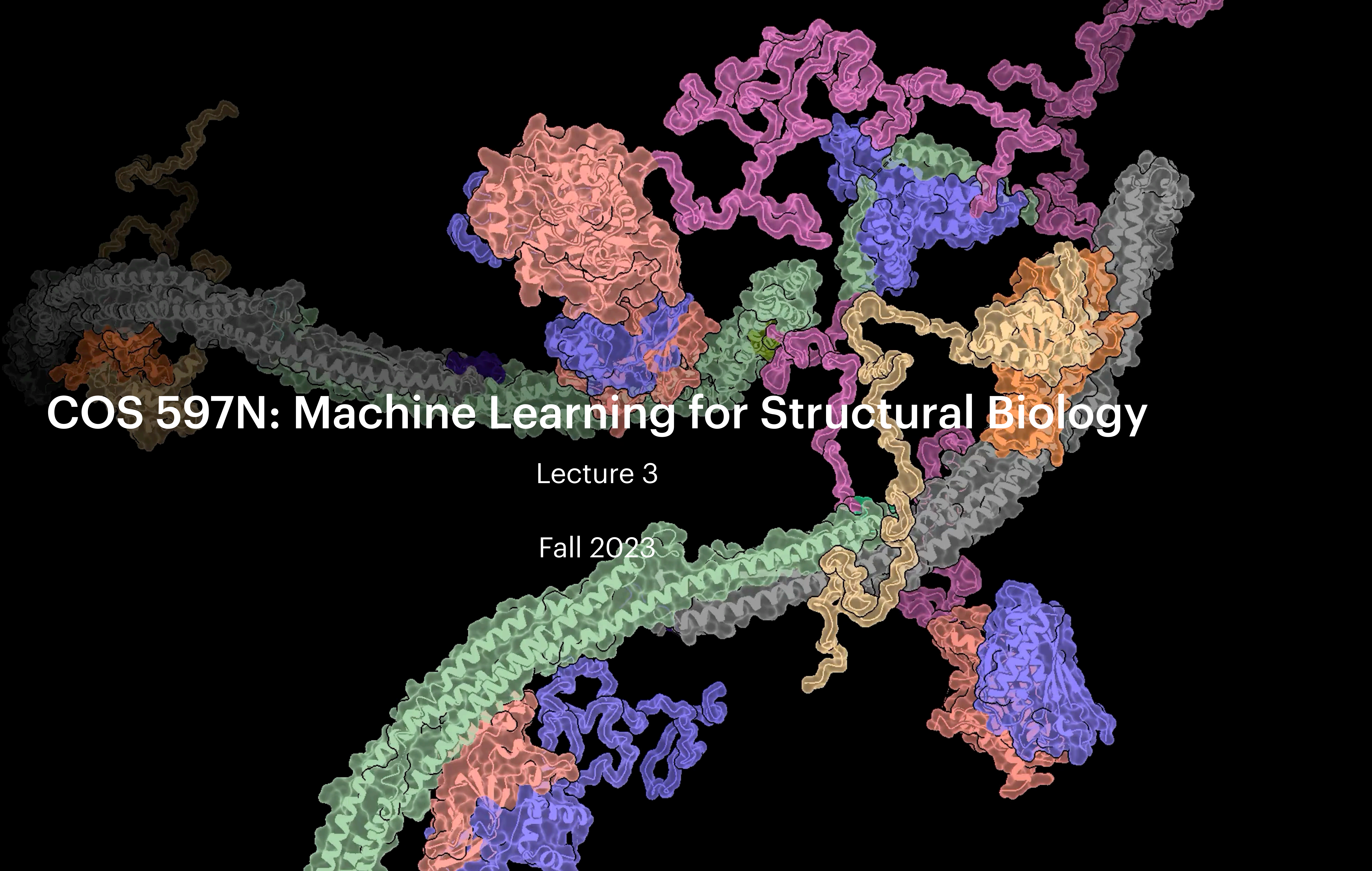


COS 597N: Machine Learning for Structural Biology

Lecture 3

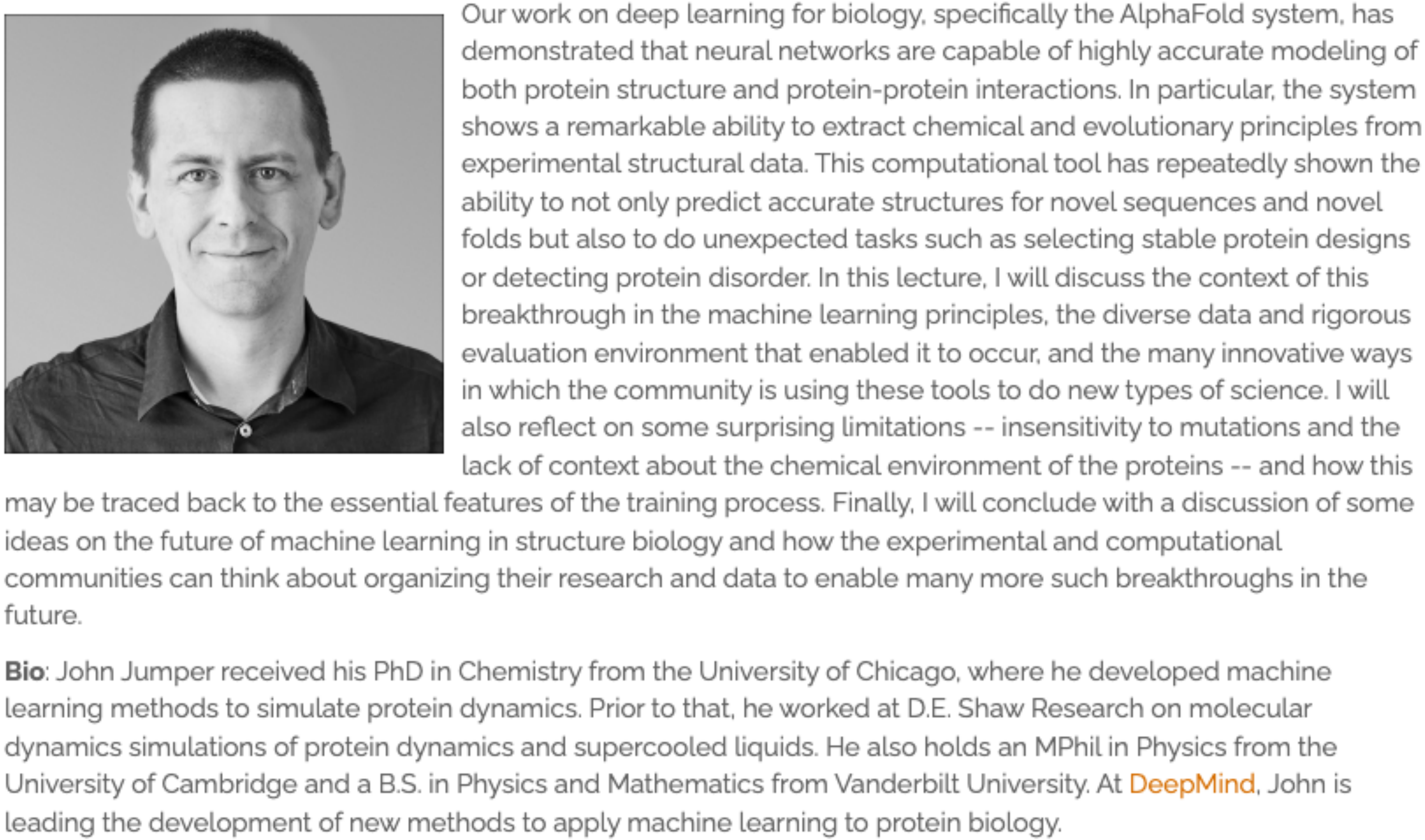
Fall 2023

Course Logistics

- Optional student-only “precept”, Tuesdays at 4:30p in CS 401.
- Course website: Schedule up for the next three weeks
- Today:
  - Viola Chen and Xiaxin Shen will present the AlphaFold2 paper
  - Feedback: Andy Zhang, Brendan Wang
- Next week 9/28:
  - I will present protein structure determination and cryo-EM reconstruction
- Oct 5: Protein language modeling — Guest instructor, and there will be a sign up next week for 5-min flash talks
- Oct 12: Protein design
- Oct 19 (fall break): Project proposal due (1 paragraph description of a final project idea)

CS Distinguished Colloquium Speaker: John Jumper

Monday Sept 25, 4:30p, Friend 101



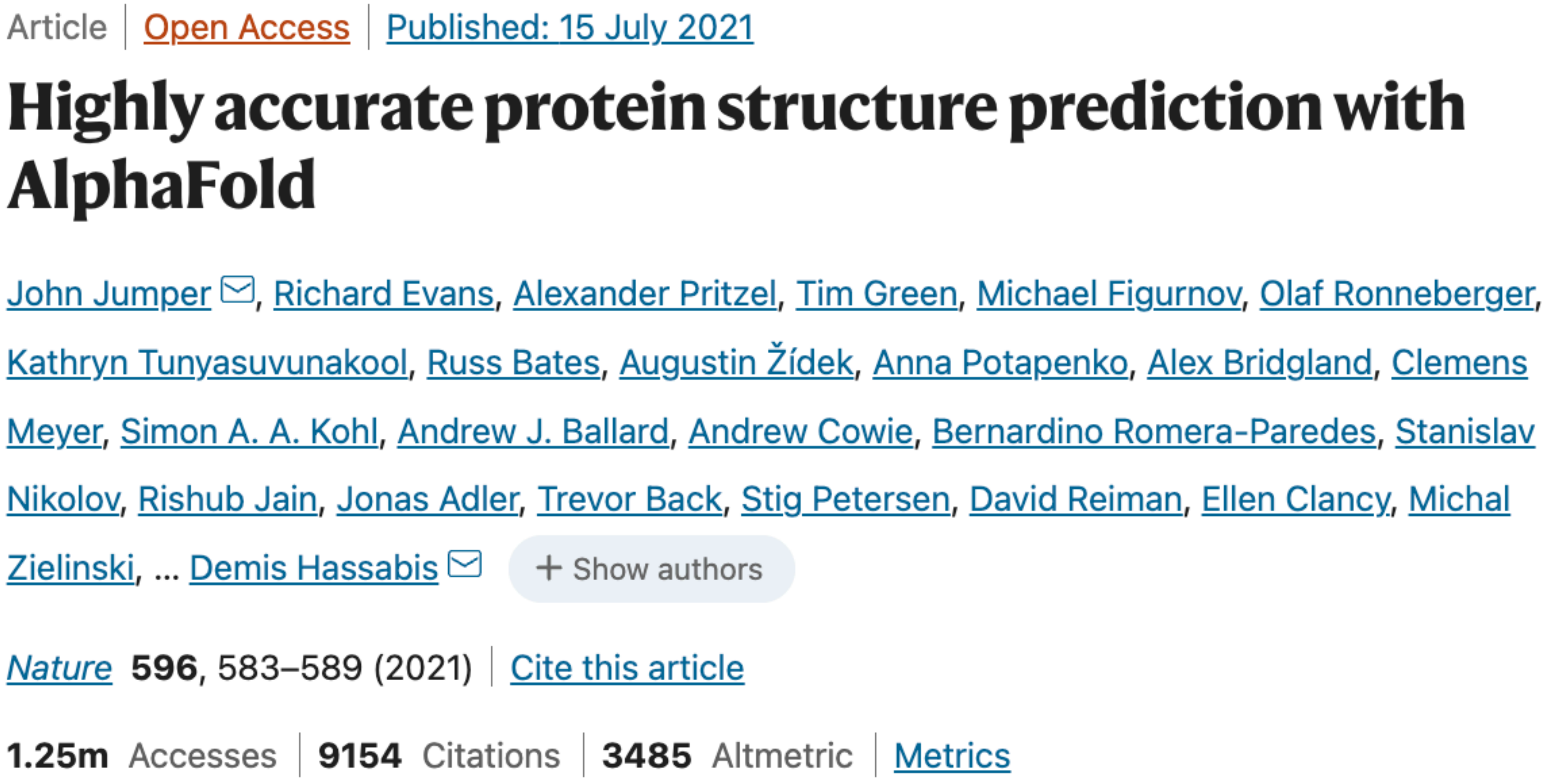
Our work on deep learning for biology, specifically the AlphaFold system, has demonstrated that neural networks are capable of highly accurate modeling of both protein structure and protein-protein interactions. In particular, the system shows a remarkable ability to extract chemical and evolutionary principles from experimental structural data. This computational tool has repeatedly shown the ability to not only predict accurate structures for novel sequences and novel folds but also to do unexpected tasks such as selecting stable protein designs or detecting protein disorder. In this lecture, I will discuss the context of this breakthrough in the machine learning principles, the diverse data and rigorous evaluation environment that enabled it to occur, and the many innovative ways in which the community is using these tools to do new types of science. I will also reflect on some surprising limitations -- insensitivity to mutations and the lack of context about the chemical environment of the proteins -- and how this may be traced back to the essential features of the training process. Finally, I will conclude with a discussion of some ideas on the future of machine learning in structure biology and how the experimental and computational communities can think about organizing their research and data to enable many more such breakthroughs in the future.

Bio: John Jumper received his PhD in Chemistry from the University of Chicago, where he developed machine learning methods to simulate protein dynamics. Prior to that, he worked at D.E. Shaw Research on molecular dynamics simulations of protein dynamics and supercooled liquids. He also holds an MPhil in Physics from the University of Cambridge and a B.S. in Physics and Mathematics from Vanderbilt University. At DeepMind, John is leading the development of new methods to apply machine learning to protein biology.

This lecture

AlphaFold2!

- AlphaFold2 presentation by Viola Chen and Xiaxin Shen
- AlphaFoldDB figures
- Question:
  - What did you think of the papers?
  - Blog post?
  - AlphaFold2 vs. AFDB?
  - Any other thoughts/reflections?



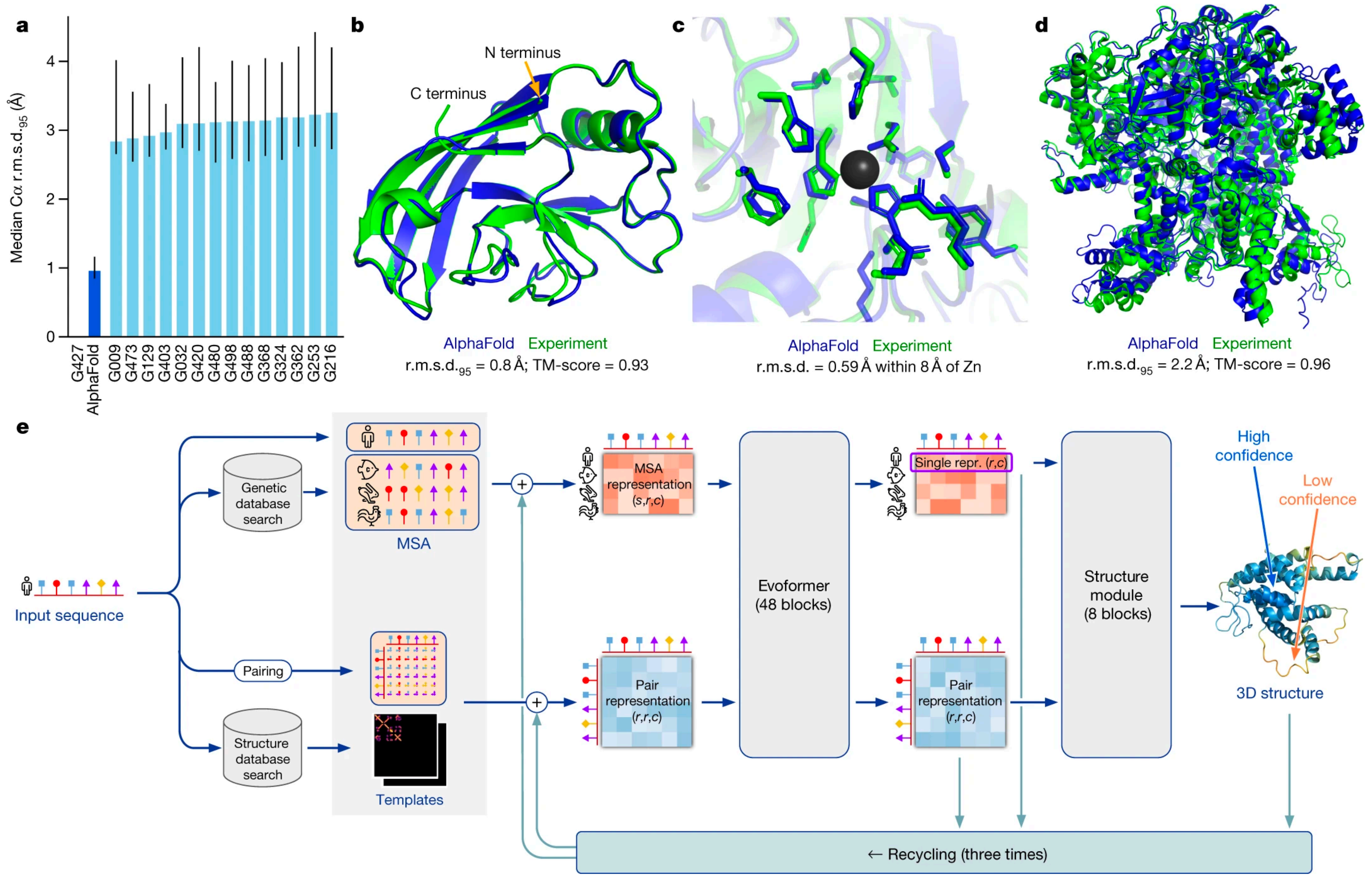
Article | Open Access | Published: 15 July 2021

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, ... Demis Hassabis + Show authors

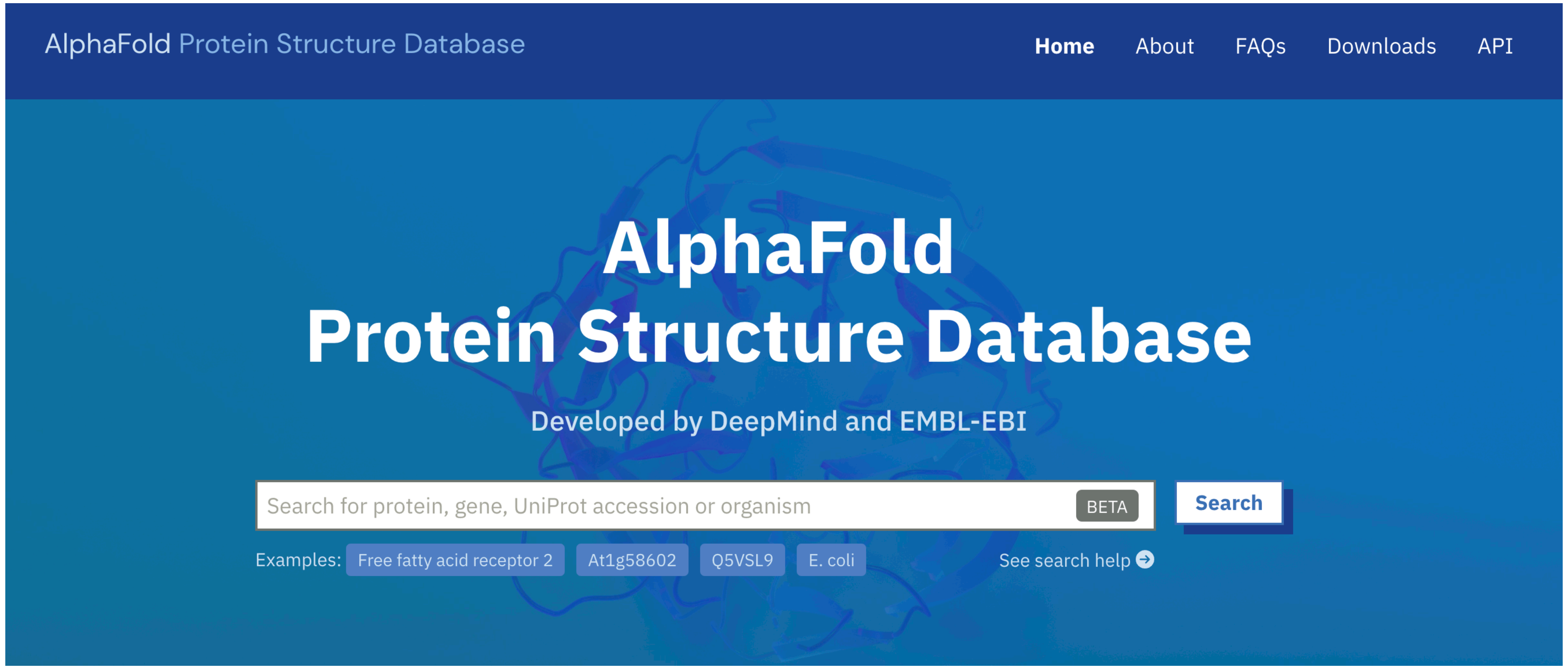
Nature **596**, 583–589 (2021) | Cite this article

1.25m Accesses | **9154** Citations | **3485** Altmetric | Metrics


a
Median Cα r.m.s.d.$_{95}$ (Å)
0
1
2
3
4
G427
AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Highly accurate protein structure prediction for the human proteome

Tunyasuvunakool et al. Nature 2021



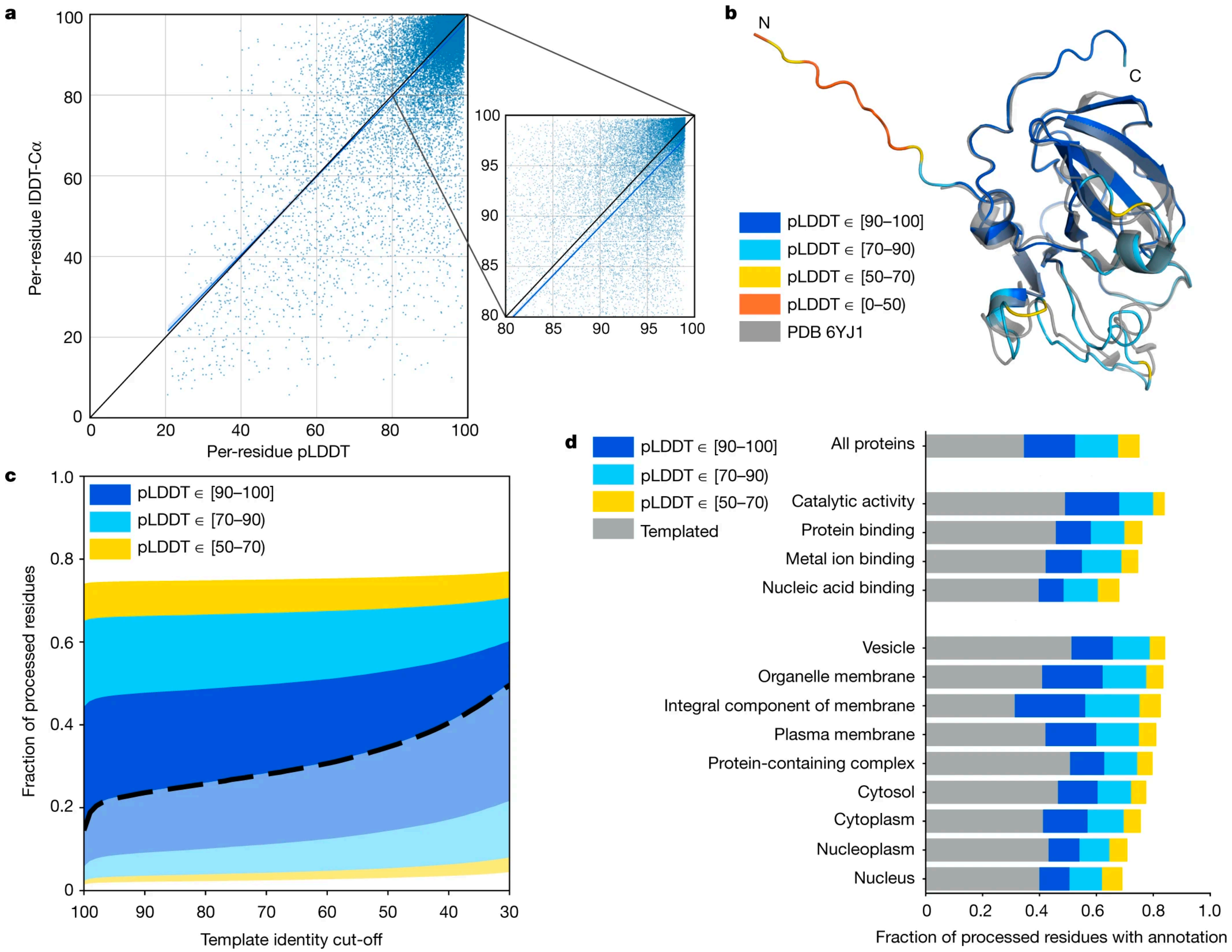
AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism

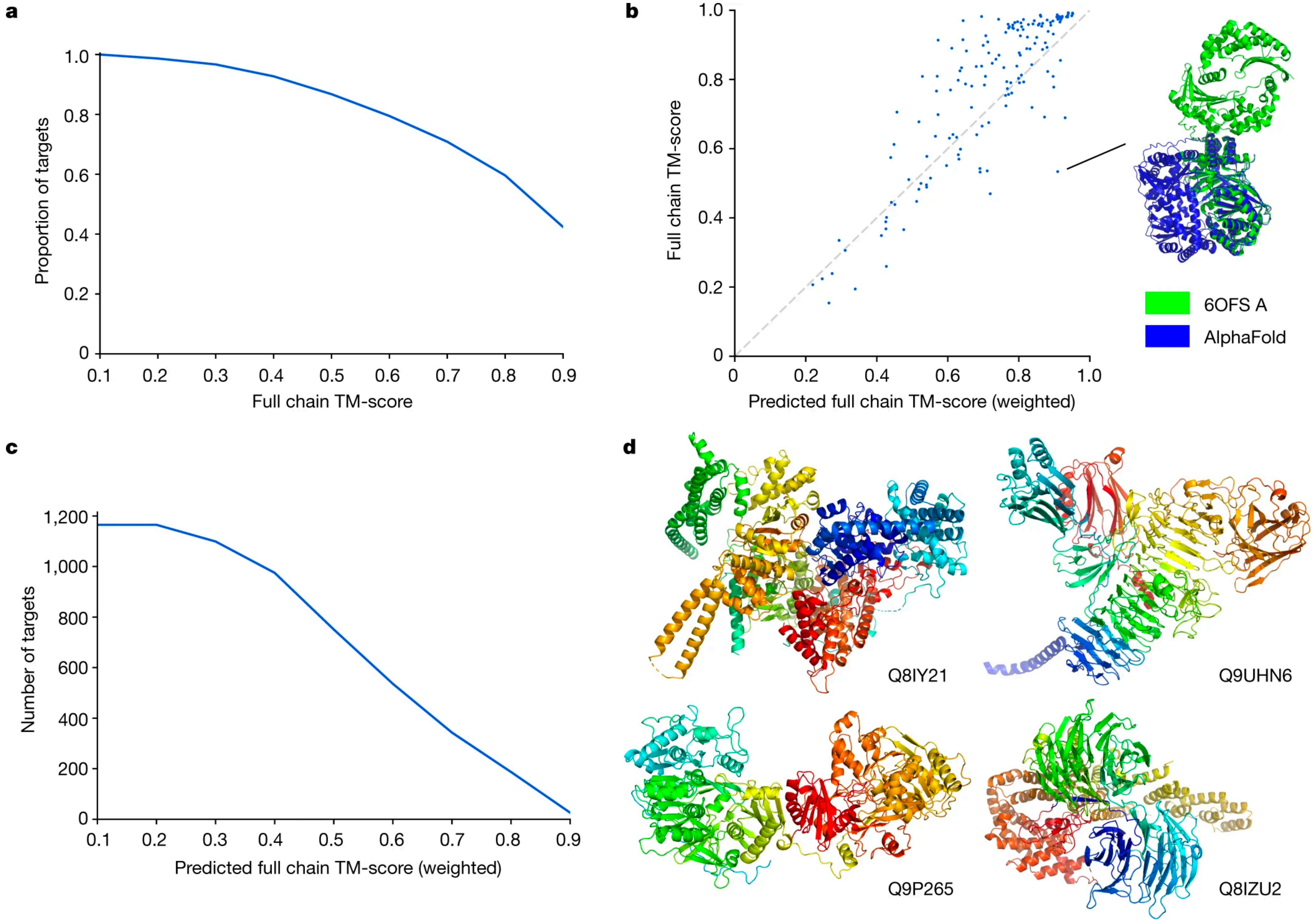
- The structures of around 100k unique proteins have been determined, but this represents a small fraction of the billions of known protein sequences
  - 17% of the total residues in human protein sequences are covered by an experimentally determined structure
- Here, they apply AlphaFold2 at proteome scales, covered 98.5% of the human proteins
  - 58% of residues have a confident prediction (pLDDT > 70)
  - 36% of residues have a highly confident prediction (pLDDT > 90)
- Identify strong multi-domain predictions, regions that are likely to be disordered
- Release of a database providing structural hypotheses

Fig. 1: Model confidence and added coverage



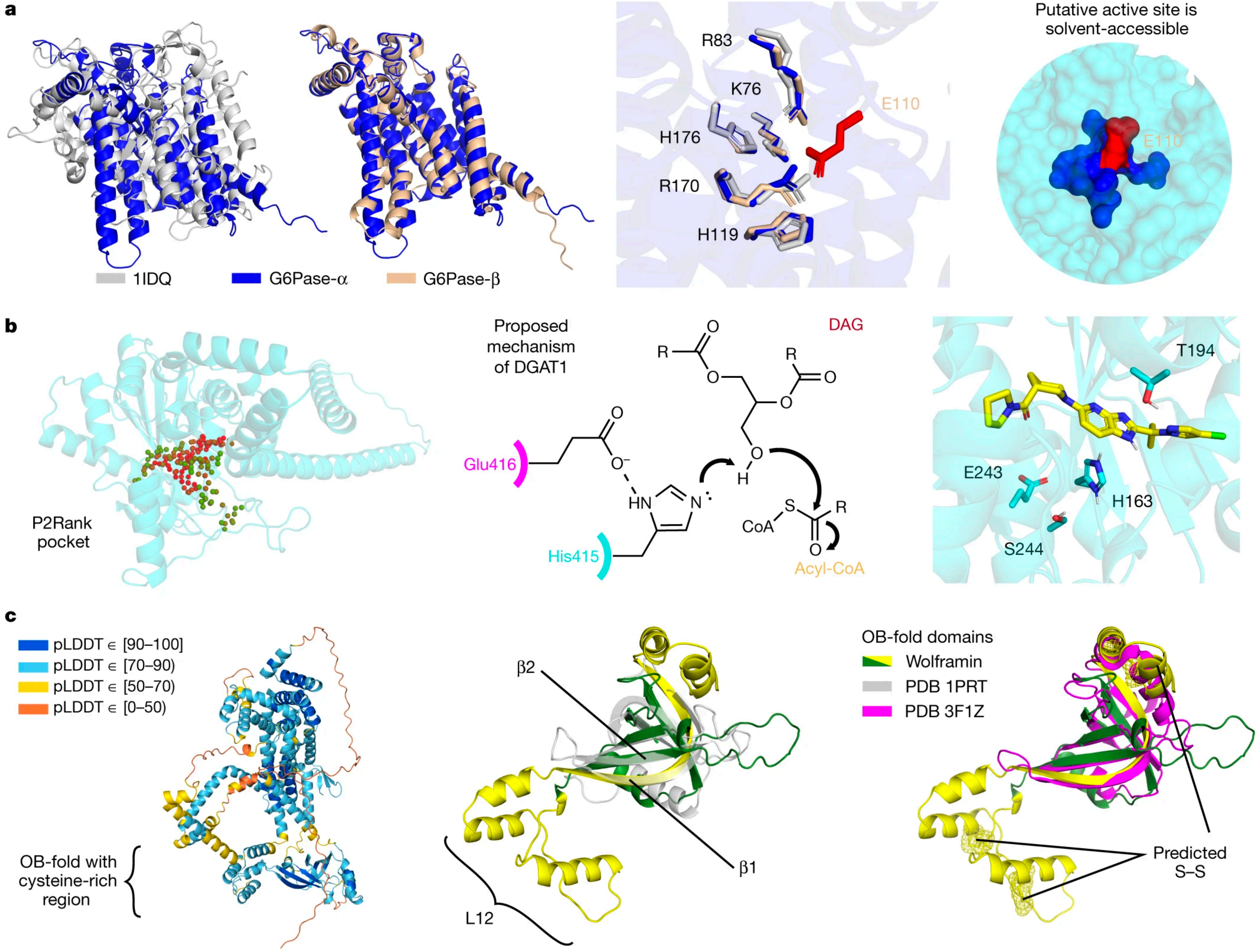
a, Correlation between per-residue pLDDT and lDDT-Cα. Data are based on a held-out set of recent PDB chains (Methods) filtered to those with a reported resolution of <3.5 Å (n = 10,215 chains and 2,756,569 residues). The scatterplot shows a subsample (1% of residues), with the blue line showing a least-squares linear fit and the shaded region a 95% confidence interval estimated with 1,000 bootstrap samples. The black line shows $x = y$, for comparison. The smaller plot is a magnified region of the larger one. On the full dataset, the Pearson's $r = 0.73$ and the least-squares linear fit is $y = (0.967 \pm 0.001) \times x + (1.9 \pm 0.1)$. **b**, AlphaFold prediction and experimental structure for a CASP14 target (PDB: 6YJ1)[64]. The prediction is coloured by model confidence band, and the N terminus is an expression tag included in CASP but unresolved in the PDB structure. **c**, AlphaFold model confidence on all residues for which a prediction was produced (n = 10,537,122 residues). Residues covered by a template at the specified identity level are shown in a lighter colour and a heavy dashed line separates these from residues without a template. **d**, Added residue-level coverage of the proteome for high-level GO terms, on top of residues covered by a template with sequence identity of more than 50%. Based on the same human proteome dataset as in **c** (n = 10,537,122 residues).

Fig. 2: Full chain structure prediction



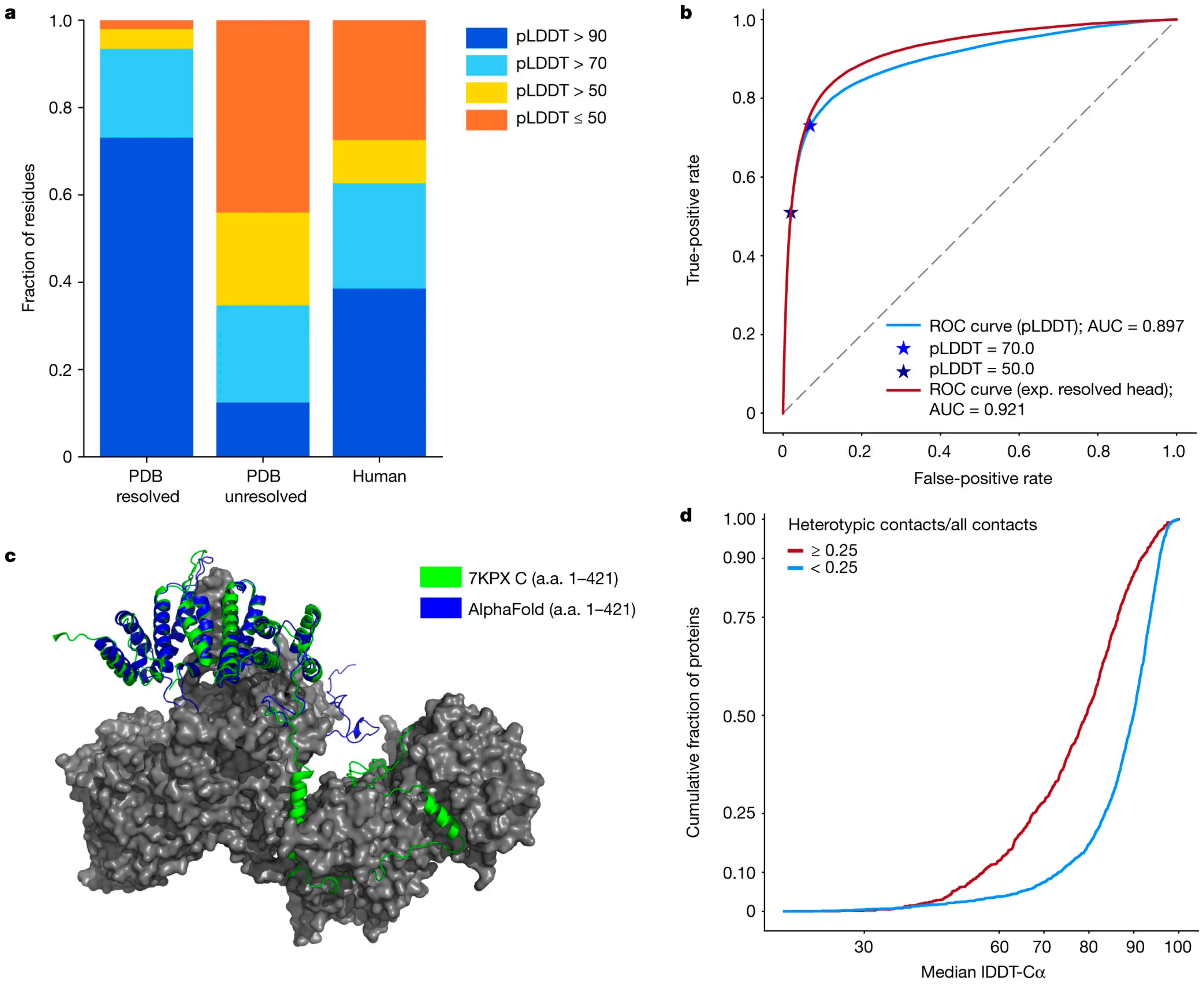
a, TM-score distribution for AlphaFold evaluated on a held-out set of template-filtered, long PDB chains (*n* = 151 chains). Includes recent PDB proteins with more than 800 resolved residues and best 50% coverage template below 30% identity. **b**, Correlation between full chain TM-score and pTM on the same set (*n* = 151 chains), Pearson's *r* = 0.84. The ground truth and predicted structure are shown for the most over-optimistic outlier (PDB: 6OFS, chain A). **c**, pTM distribution on a subset of the human proteome that we expect to be enriched for structurally novel multidomain proteins (*n* = 1,165 chains). Human proteome predictions comprise more than 600 confident residues (more than 50% coverage) and no proteins with 50% coverage templates. **d**, Four of the top hits from the set shown in **c**, filtering by pTM > 0.8 and sorting by number of confident residues. Proteins are labelled by their UniProt accession. For clarity, regions with pLDDT < 50 are hidden, as are isolated smaller regions that were left after this cropping.

Fig. 3: Highlighted structure predictions



a, Left, comparison of the active sites of two G6Pases (G6Pase-α and G6Pase-β) and a chloroperoxidase (PDB 1IDQ). The G6Pases are glucose-forming enzymes that contain a conserved, solvent-accessible glutamate (red; right) opposite the shared active-site residues (middle). **b**, Left, pocket prediction (P2Rank[65]) identifies a putative binding pocket for DGAT2, which is involved in body-fat synthesis. Red and green spheres represent the ligandability scores by P2Rank of 1 and 0, respectively. Middle, a proposed mechanism for DGAT1[51] activates the substrate with Glu416 and His415, which have analogous residues in the DGAT2 pocket. The docked inhibitor is well placed for polar interactions with His163 and Thr194 (right). The chemical structure (middle) is adapted from ref. [51]. **c**, Predicted structure of wolframin, mutations in which cause Wolfram syndrome. Although there are regions in wolframin with low pLDDT (left), we could identify an OB-fold region (green/yellow), with a comparable core to a prototypical OB-fold (grey; middle). However, the most similar PDB chain (magenta; right) lacks the conserved cysteine-rich region (yellow) of our prediction. This region forms the characteristic β1 strand and an extended L12 loop, and is predicted to contain three disulfide bridges (yellow mesh).

Fig. 4: Low-confidence regions



a, pLDDT distribution of the resolved parts of PDB sequences (n = 3,440,359 residues), the unresolved parts of PDB sequences (n = 589,079 residues) and the human proteome (n = 10,537,122 residues). **b**, Performance of pLDDT and the experimentally resolved head of AlphaFold as disorder predictors on the CAID Disprot-PDB benchmark dataset (n = 178,124 residues). **c**, An example low-confidence prediction aligned to the corresponding PDB submission (7KPX chain C)[66]. The globular domain is well-predicted but the extended interface exhibits low pLDDT and is incorrect apart from some of the secondary structure. a.a., amino acid. **d**, A high ratio of heterotypic contacts is associated with a lower AlphaFold accuracy on the recent PDB dataset, restricted to proteins with fewer than 40% of residues with template identity above 30% (n = 3,007 chains) (Methods). The ratio of heterotypic contacts is defined as: heterotypic/(intra-chain + homomeric + heterotypic).