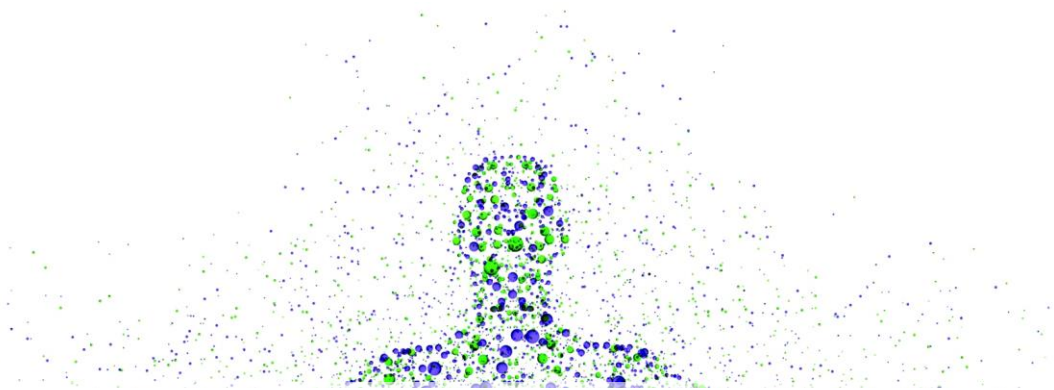
AI, Data and their impact on Life Sciences research

Pistoia Meeting
London
April 2024

Janet Thornton

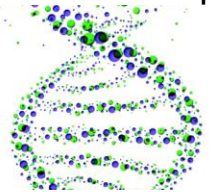
Director Emeritus

EMBL-EBI

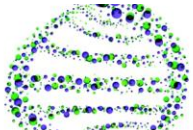


In the last 50 years, we have been living through SEVERAL revolutions.

DNA sequencing: Imaging methods: Computational power: AI methods

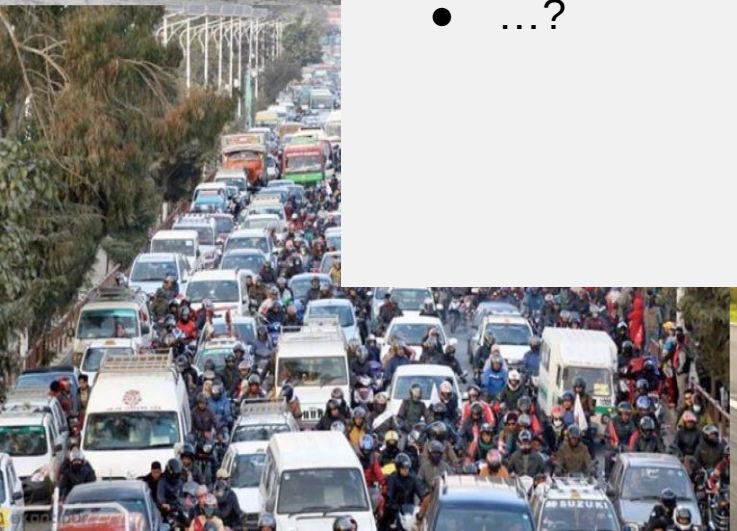


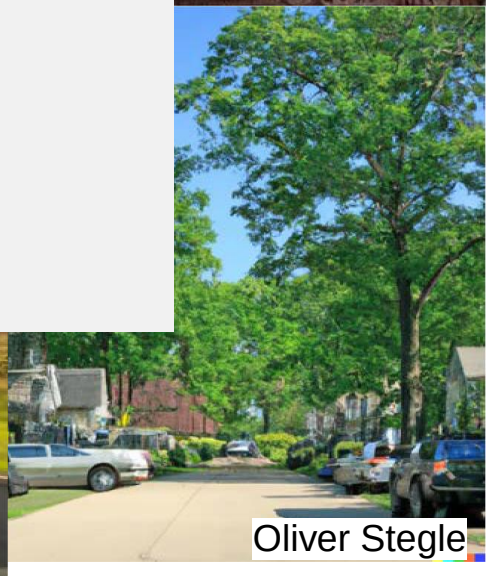
The impact on The Life Sciences is already huge



What did the invention of cars mean ~1900 for

- a horse-carriage driver
- a farrier; a blacksmith
- a bakery shop owner
- the city council of Cambridge/Heidelberg/Rome/…
- seaside resorts
- urban planners and developers
- the army
- …?





Oliver Stegle

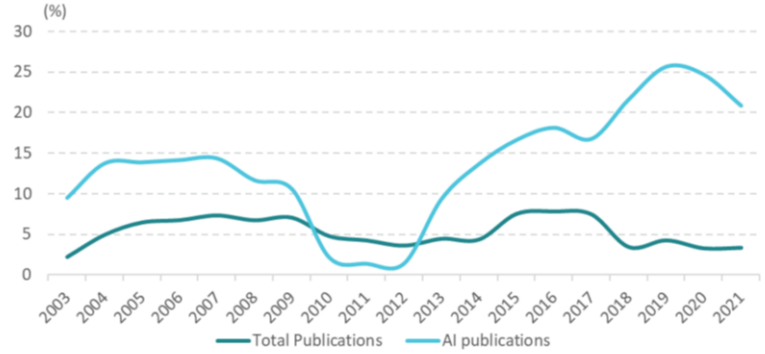
What AI “thinks” how AI will change laboratories

Generated using Dallee-E

Oliver Stegle

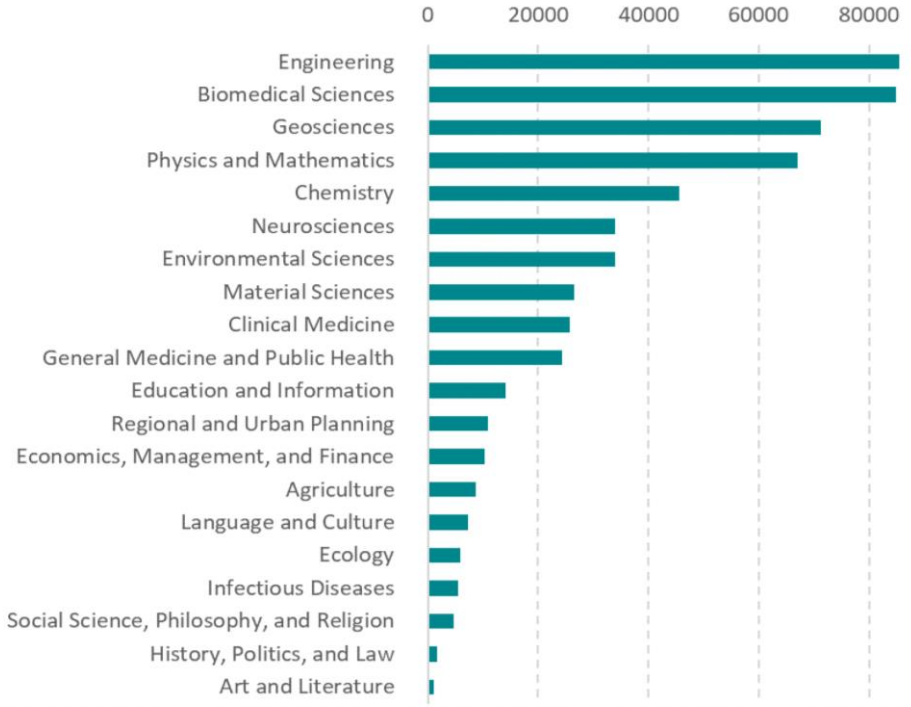
Use of AI is growing rapidly in all fields of Research

Growth in Scientific Activity – 3year average



Source: European Commission, DG Research & Innovation, calculations based on Web of Science data. Annual growth calculated as a 3-year rolling average.

Number of AI publications (2017-2021) per Scientific Domain



Overview

- AI & AlphaFold
  - A solution to a long-standing challenge
- Biological Data & Infrastructure
  - EMBL-EBI
  - Community Actions
- Impact on Life Sciences Research
  - Impact in Biology
  - Impact in Medical Sciences

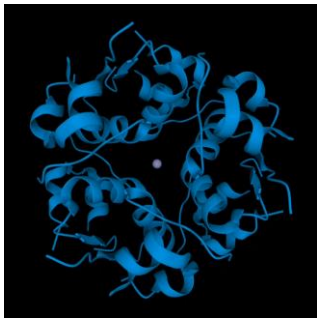
EMBL-EBI

The Challenge: Protein Structure Prediction

Protein Sequence

AGLYFE……

3D structure



The unique 3D structure of a protein is determined by its sequence.

Can we predict structure (ie coordinates) from sequence?

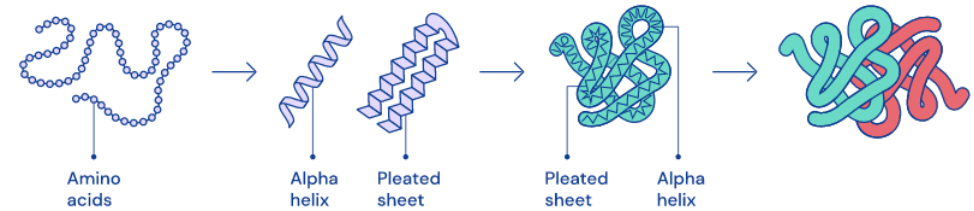
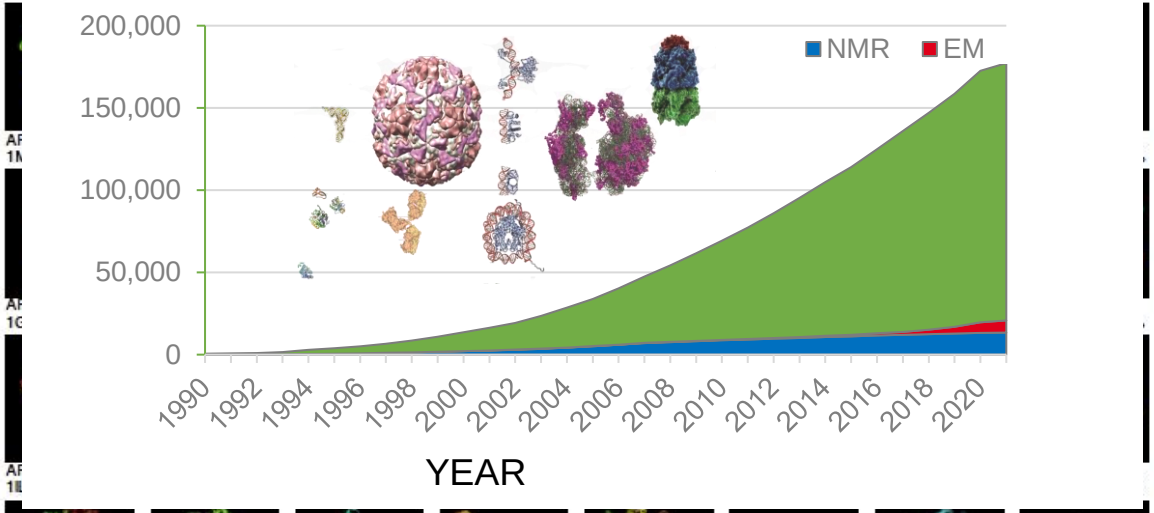


FIGURE 1: COMPLEX 3D SHAPES EMERGE FROM A STRING OF AMINO ACIDS.

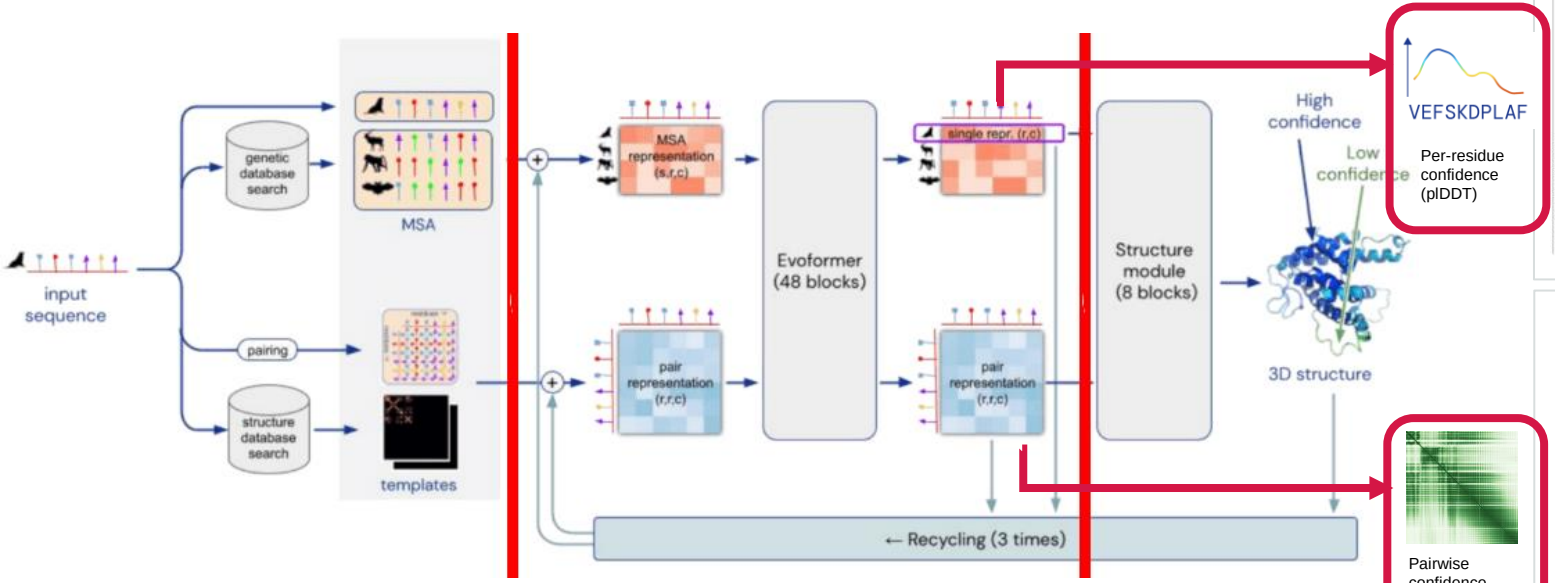
More & More Structures



Only 0.027% of UniProt accessions are directly mapped to the PDB (55,000+)

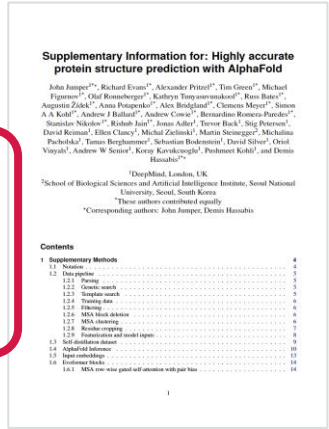
.-EBI

DeepMind’s AlphaFold – core method




Article

Highly accurate protein structure prediction with AlphaFold


Supplementary Information for: Highly accurate protein structure prediction with AlphaFold

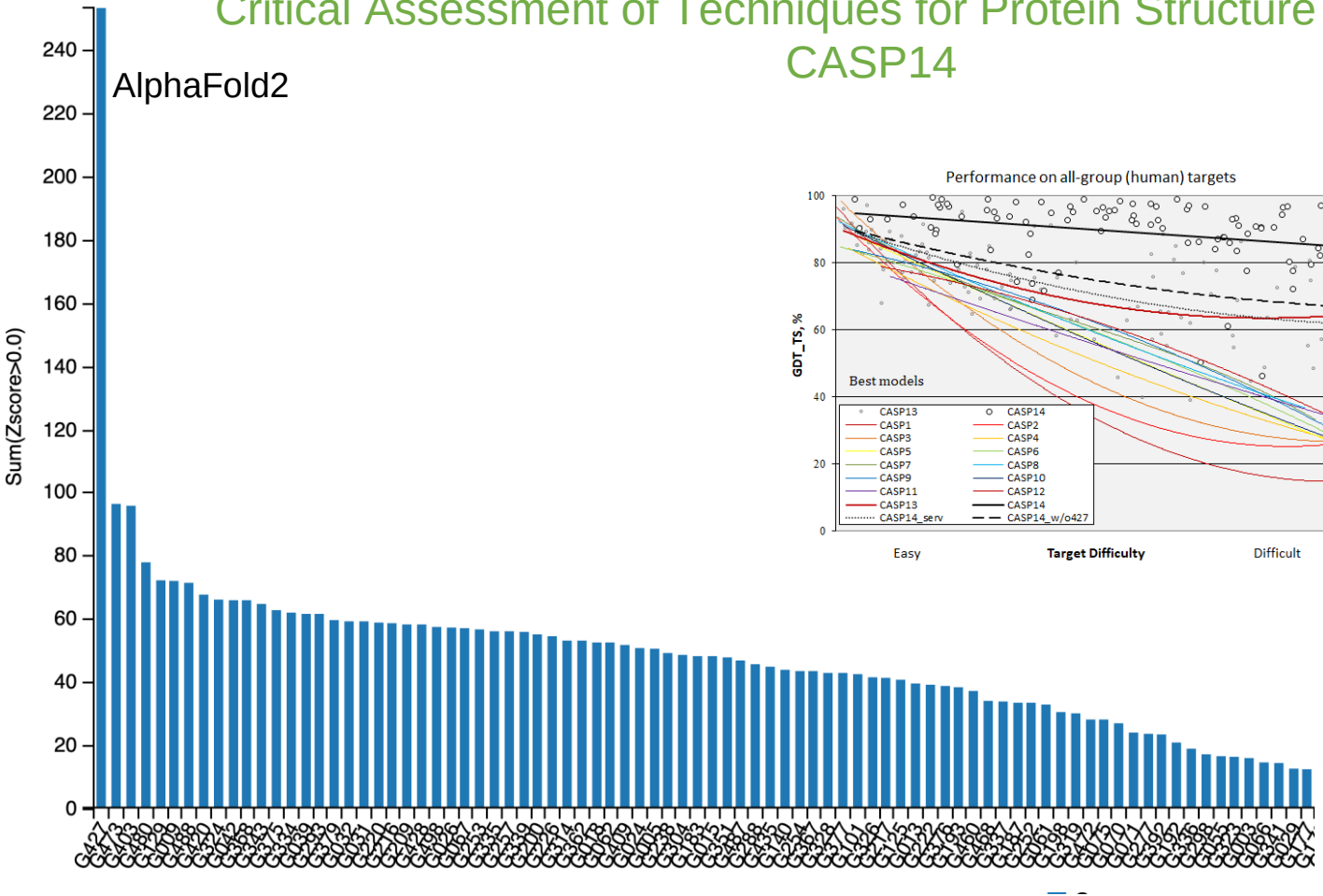
62 Page Supplementary Information, Google colab and open GitHub repository
See Jumper et al. 2021 (especially the SI) for details

This is the first time a serious scientific problem has been solved by AI. - John Moult

DeepMind EMBL-EBI

Critical Assessment of Techniques for Protein Structure Prediction
CASP14

AlphaFold2



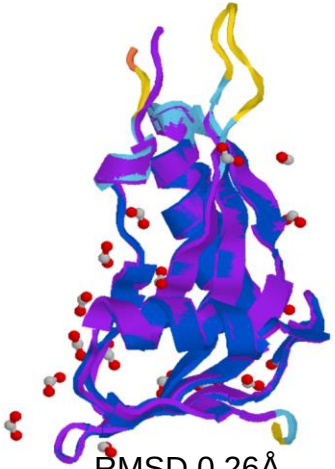
John Moult

Institute for Bioscience and Biotechnology Research (IBBR), University of Maryland, Washington, US

EMBL-EBI

What do the AlphaFold models look like?

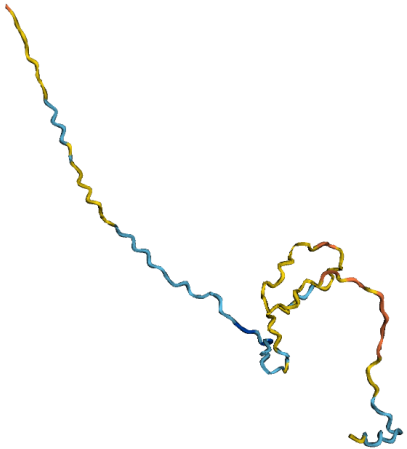
Q9NRX4 –
14 kDa phosphohistidine phosphatase



RMSD 0.26Å

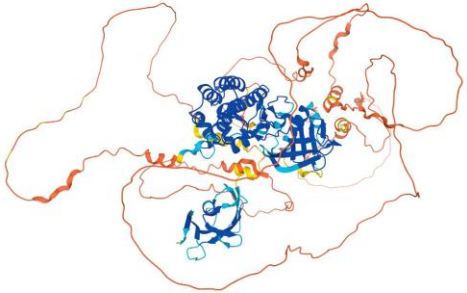
Average confidence score: 90.1
76.8% of residues score above 90

O60927
E3 ubiquitin-protein ligase PPP1R11



Average confidence score: 66.4
3.1% of residues score above 90

Q99558
MAP3 Kinase K14



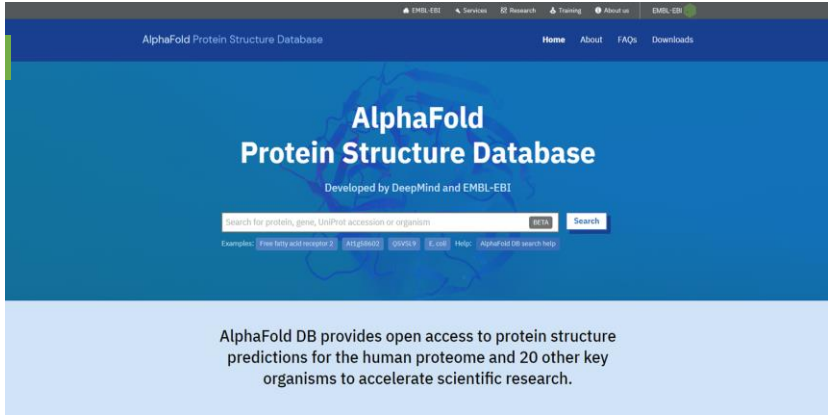
3 domains well predicted,
but connections unknown

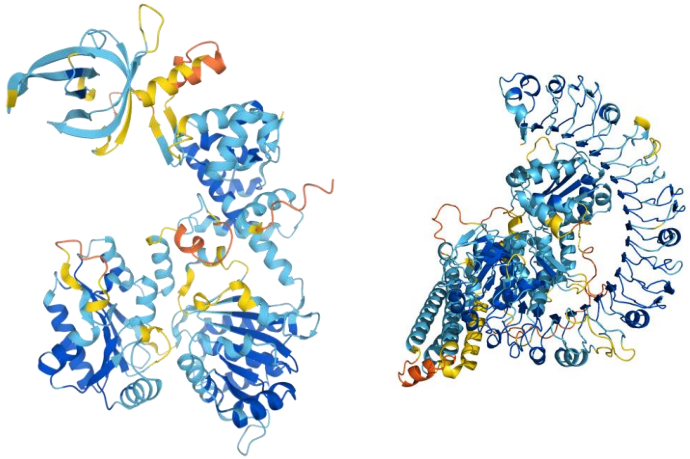
Confidence Scores are essential

EMBL-EBI

AlphaFold Database @ EMBL-EBI

- Launched on 22nd July 2021
  - Collaboration between EMBL and DeepMind
  - Open access (CC-BY-4.0 license)
  - Data available for bulk download via FTP (ftp.ebi.ac.uk/pub/database/alphafold)
  - Structure for every known protein in UniProt database are available or can be modelled
  - 3-D structures for virtually all (98.5%) of the human proteome
  - C.f. Only 17% human proteome in PDB





EMBL-EBI

Protein Structure Prediction – The perfect problem for AI

- Long standing important problem
  - All life is based on proteins;
  - Sequencing DNA is rapid and cheap; Sequencing proteins is much more difficult and expensive;
  - Structure determines biological function; Structure is critical for drug design
- Lots of data – 170K Protein structure entries in PDB
  - Data are curated, clean, accurate and freely available (sequences and structures)
  - Problem is well defined: - given a protein sequence, predict the 3D structure (coordinates)
  - Good metrics to measure success
- Evolution
  - Most proteins belong to a relatively small number of families
  - Conservation of protein structure during evolution and in different species facilitates prediction
- Many helpful scientific and technical developments in last 30 years
- CASP – Comparative Assessment of Structure Prediction: independent test of quality of prediction.

AlphaFold Structures – Fundamental limitations

AlphaFold does not predict protein folding pathway

AlphaFold not trained or validated for predicting the effect of mutations

The original AlphaFold2 did not consider multimers

Ligands are not included -

Low reliability regions have poor stereochemistry

Predictions may (or may not) lead to hypotheses about protein function – any hypotheses have to be tested by further experimentation

Access to good data were critical for the success of AlphaFold!

EMBL-EBI

- Biological Data & Infrastructure
  - EMBL-EBI
  - Community Efforts

EMBL-EBI

The data challenges in the life sciences

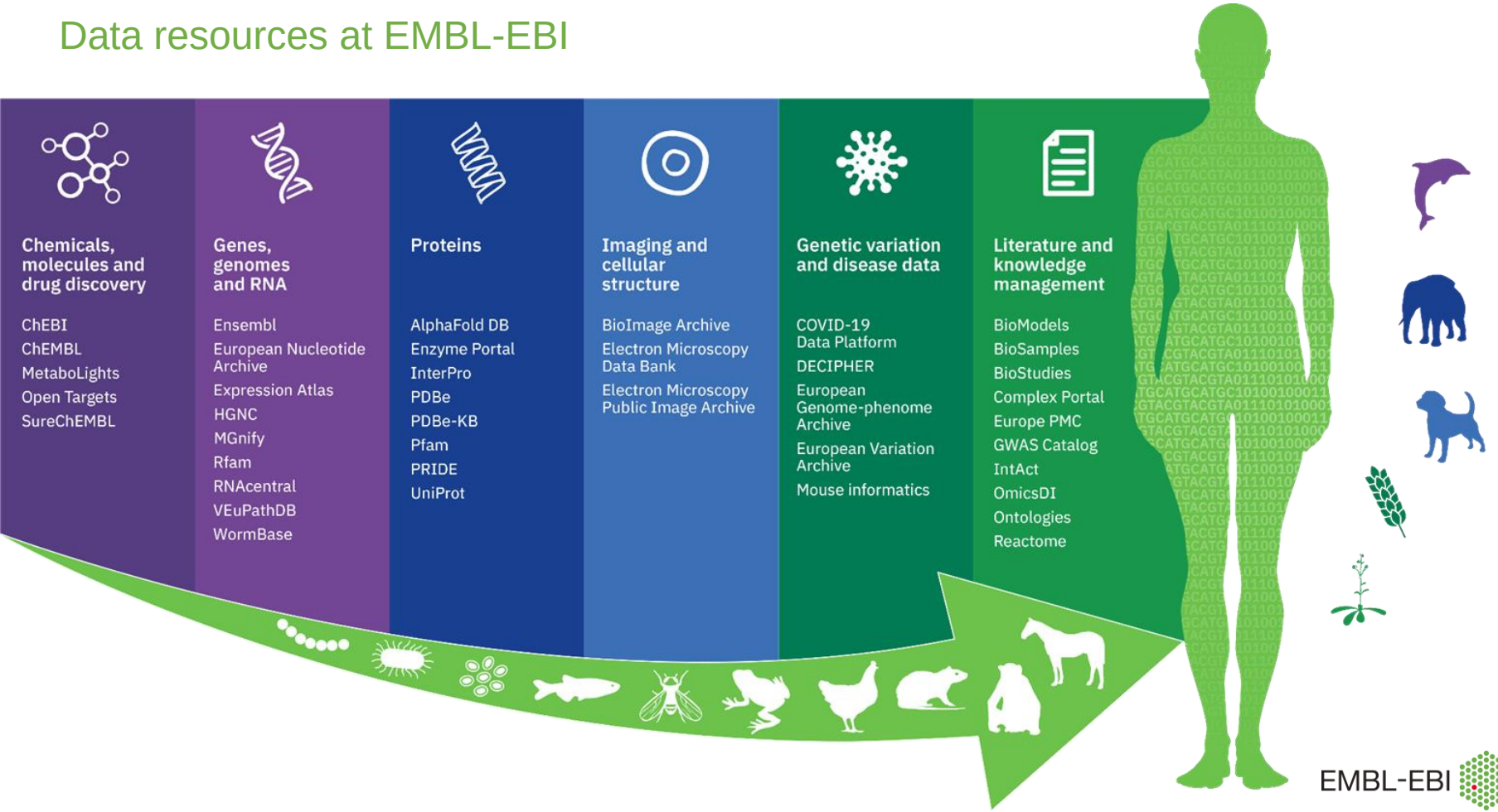
Data growth

Diversity of data

Geographic spread

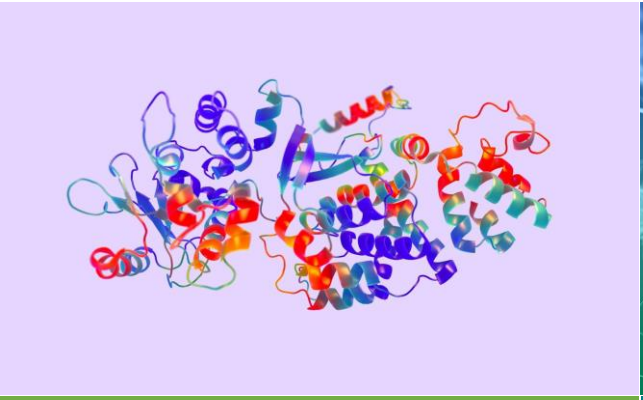
Sensitive data

Data resources at EMBL-EBI



EMBL-EBI

Impact of EMBL-EBI data resources



Over 40

open data resources

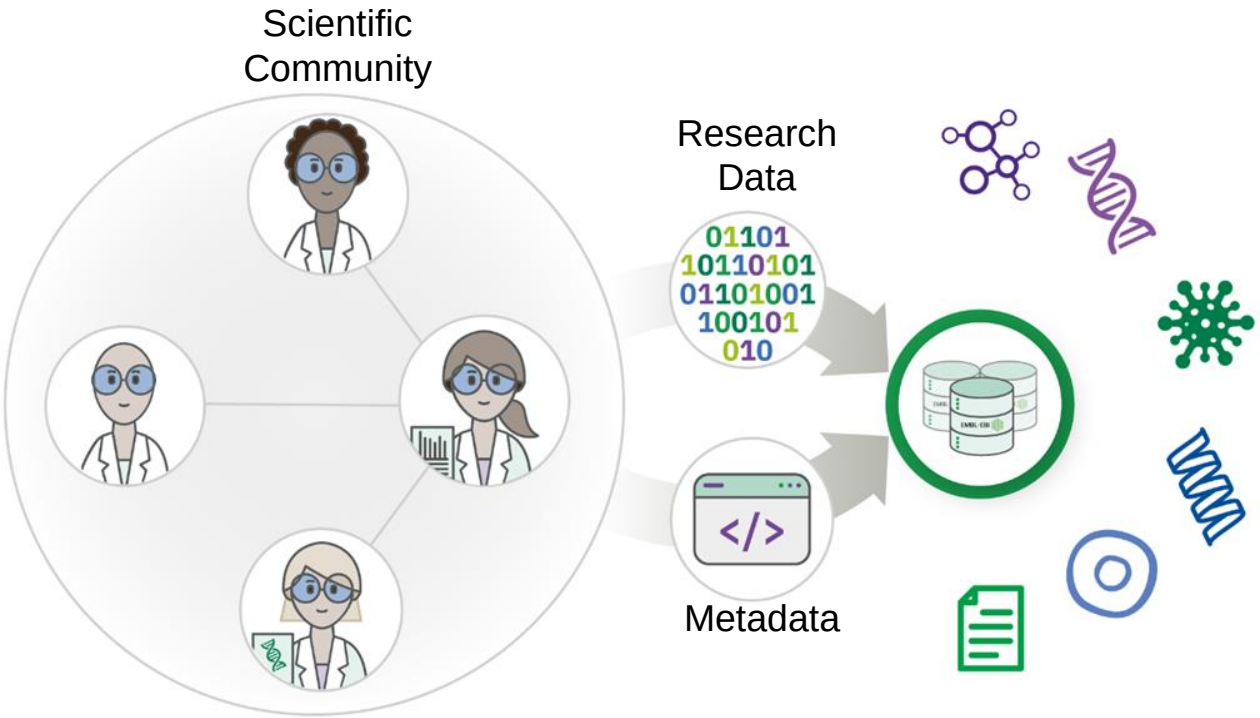
Over 100 million

requests to our websites on an average day

Our data resources are referenced in scientific papers 200 times every day

EMBL-EBI

AI and data science at EBI



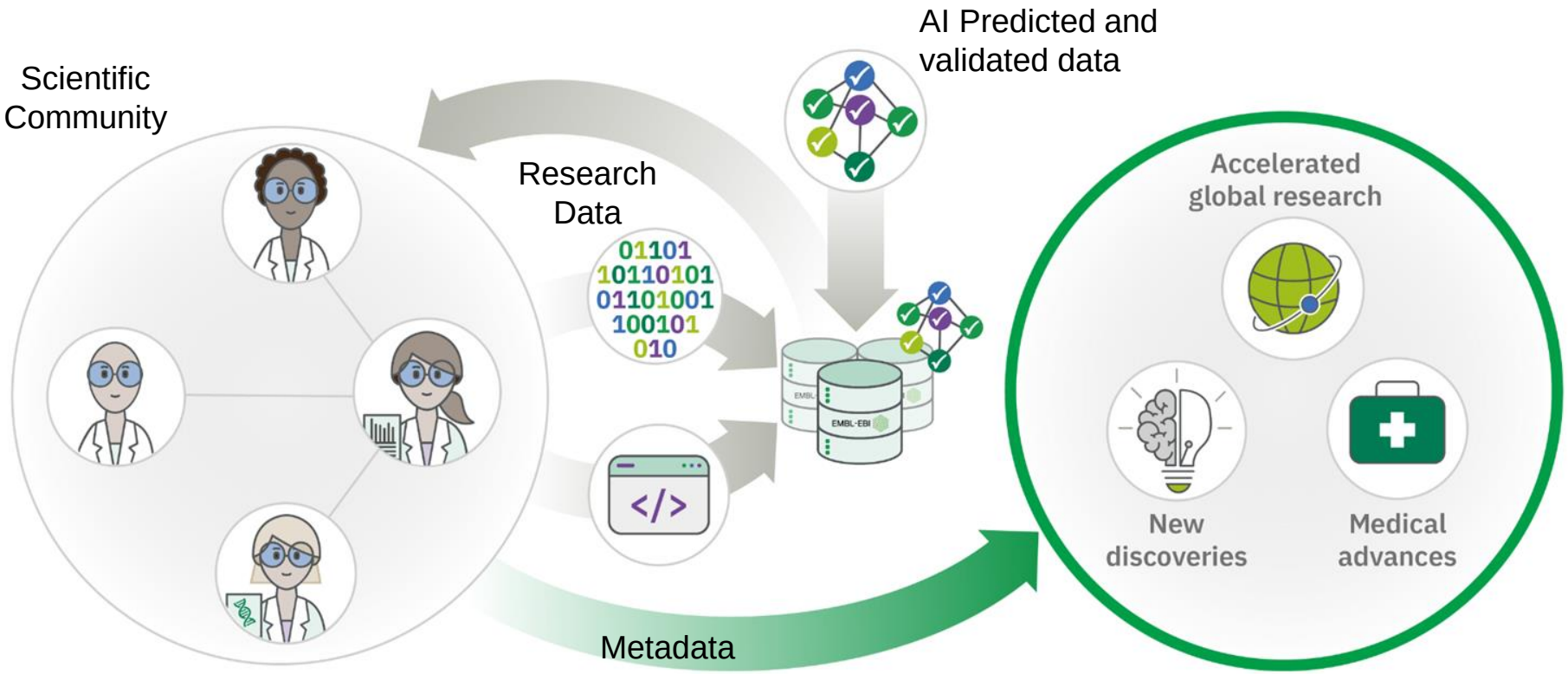
EMBL-EBI data resources:

- Curation
- Annotation
- Classification
- Enrichment
- Analysis

Our 40+ data resources span genetics, genomics, proteins, chemistry, literature data and more.

High quality OPEN data

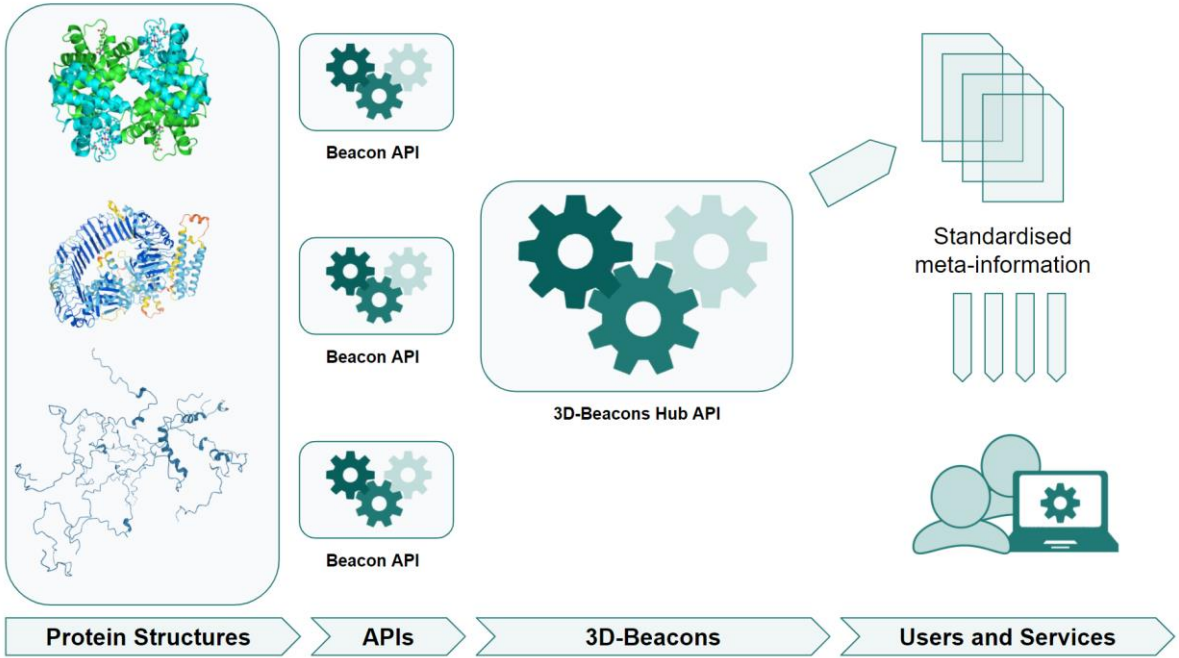
EMBL-EBI


Scientific Community
Research Data
AI Predicted and validated data
01101
10110101
01101001
100101
010
</>
EMBL-EBI
Accelerated global research
New discoveries
Medical advances
Metadata

EMBL-EBI

Community Efforts: 3D-Beacons – an infrastructure to make all protein structure models available, including AlphaFold

- Create a common API specification to access both experimental and predicted structures and annotations
- Provide quality assessment data for all structure models

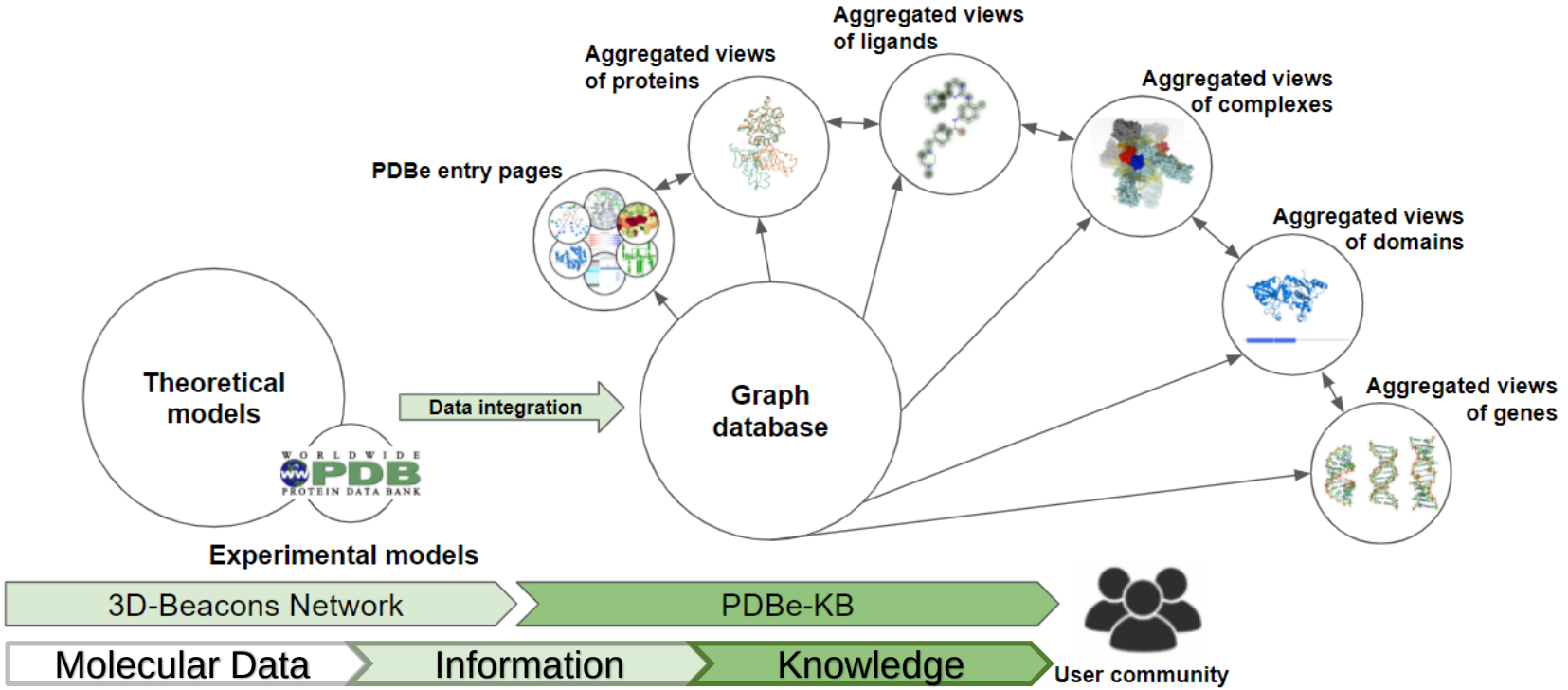


https://github.com/3D-Beacons/3D-Beacons

elixir

EMBL-EBI

Infrastructure for Structural Biology
The PDB Knowledge Base - from data to knowledge PDBe-KB



3d-beacons.org pdbe.org pdbe-kb.org

Reaching out to Industry: EMBL-EBI Industry Programme today

EMBL's mission to disseminate cutting-edge technologies to industry.

- Explore emerging areas for R&D
- Network of experts
- Solve big data challenges
- Developing data standards
- Efficiency savings

- Quarterly strategy meetings
- Knowledge-exchange workshops

Pharmaceutical and diagnostic

abbvie, astellas, astex pharmaceuticals, AstraZeneca, Bayer, Benevolent AI, Biogen, Boehringer Ingelheim, Bristol Myers Squibb, Daiichi-Sankyo, Eisai, Exscientia, GSK, Johnson&Johnson Innovative Medicine, Lilly, Lonza, Merck, moderna, MSD, Novartis, novo nordisk, Pfizer, Pierre Fabre, Roche, sanofi, Takeda, UCB

Agrofood and consumer goods

Bayer, syngenta, Unilever

AI is dominating computational approaches in the Life Sciences

EMBL-EBI

EMBL-EBI Industry Programme Workshops: 2024 schedule

1. Chemical spaces and synthesis automation informatics | UK : March 13-14 - EMBL-EBI
2. Recent developments in large language models for biology | US : April 9-10; BMS, Boston
3. **Advances in machine learning for protein design | UK : May 21-22 - EMBL-EBI**
4. **Data-driven research I.T. exchange | UK : June 12-13 - EMBL-EBI**
5. **Data Literacy in R&D | online June 2024 & In-person Oct 9-10 EMBL-EBI**
6. Cell-cell communication analysis | US : June 25-26 - BMS, Boston
7. Preclinical immunogenicity assessment of biologics | UK : Sept 11-12 - EMBL-EBI
8. **Advances in machine learning for protein design II | US Edition: Sept 18-19 - Sanofi, Boston**
9. Genetic Biomarkers in pharmaceutical development | US : October 29-28 - EMD Serono, Boston
10. Genetic and Genomic risk stratification in clinical trial design | US : Nov 13-14 - Genentech, CA
11. **Multi-omics concepts and applications in biotechnology research | UK : Nov 20-21 - EMBL-EBI**

EMBL-EBI

AI & Impact on Life Sciences Research

- Impact in Structural Biology
- Impact in Medical Sciences

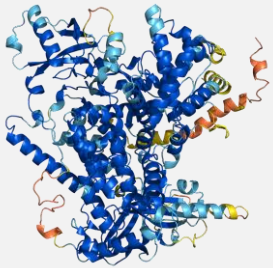
See Jones, David T., and Janet M. Thornton. "The impact of AlphaFold2 one year on." *Nature methods* 19.1 (2022): 15-20.

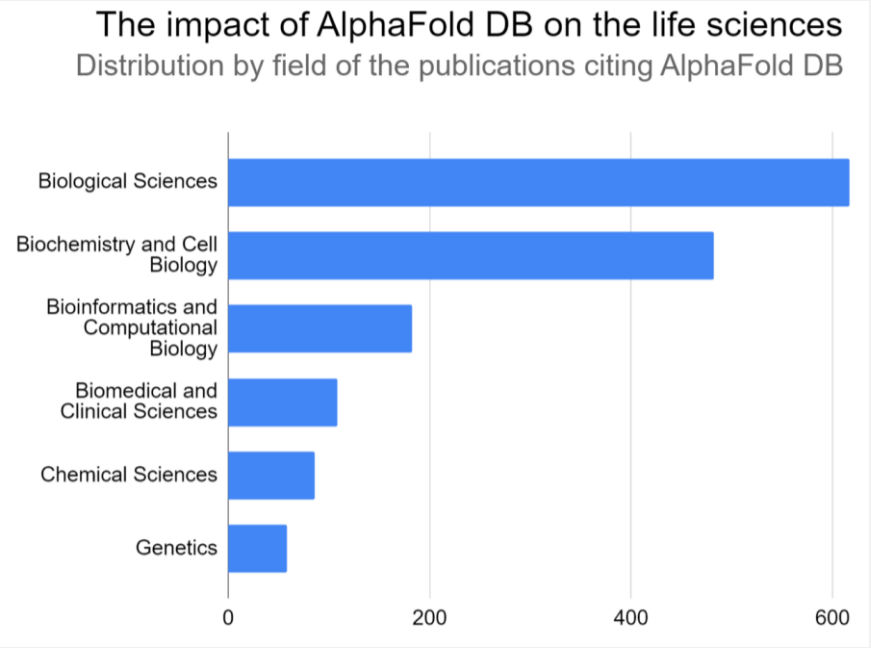
EMBL-EBI

AlphaFold Database Usage

- **>1.8 M unique users from every Continent**
- **>1B million API requests**
- **Users have downloaded 1.2M files**
- **>23 K bulk archives**

- **AF2 papers cited > 18,000 times**

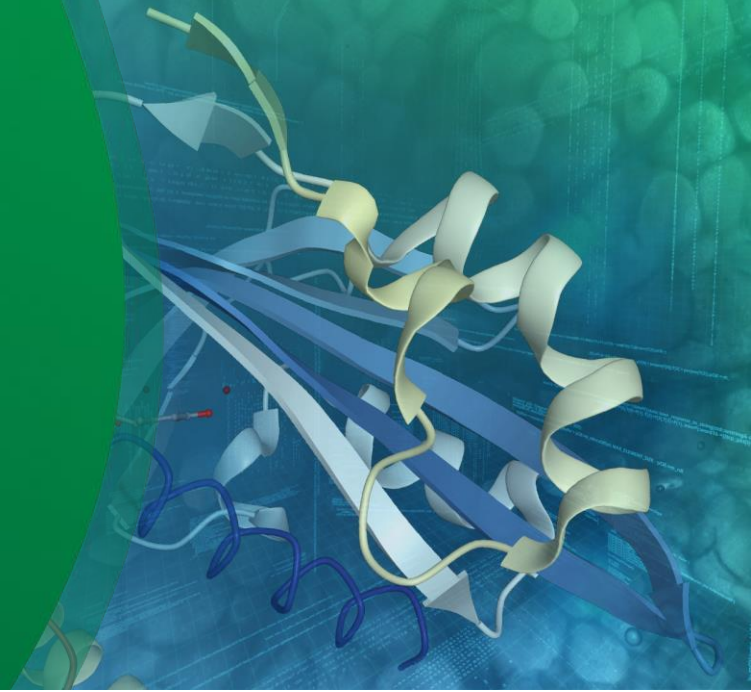




EMBL-EBI

Enabling discoveries using macromolecular structure data

Impact of structure prediction using AI methods



PDBe Protein Data Bank in Europe

PDBe-KB Protein Data Bank in Europe - Knowledge Base

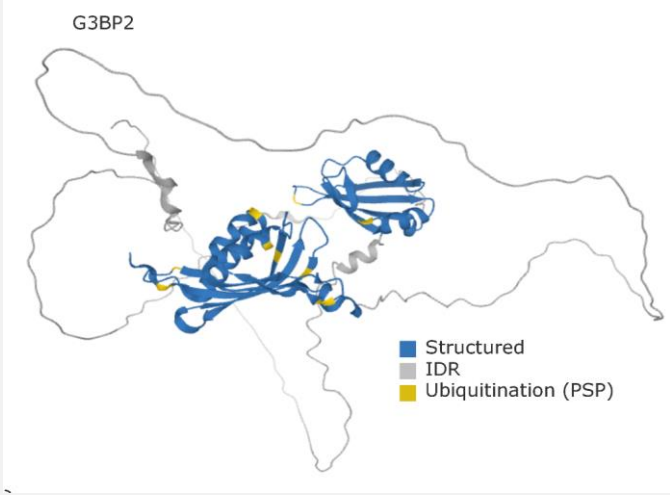
AlphaFold Protein Structure Database

pdbe.org | pdbe-kb.org | alphafold.ebi.ac.uk | 3d-beacons.org

EMBL-EBI

Sameer Velankar

Providing structural context



- Integrating large scale **PTM information** and predicted models
  - Phosphorylation, ubiquitination and acetylation
- **Chemical cross-linking studies** to identify and validate macromolecular complexes
- Combining **experimental data (cross-lining, co-fractionation) and structure prediction methods** to get insights into functional complexes and interactions

The structural context of posttranslational modifications at a proteome-wide scale

Isabell Bludau, Sander Willems, Wen-Feng Zeng, Maximilian T. Strauss, Fynn M. Hansen, Maria C. Tanzer, Ozge Karayel, Brenda A. Schulman, Matthias Mann

Published: May 16, 2022 • https://doi.org/10.1371/journal.pbio.3001636

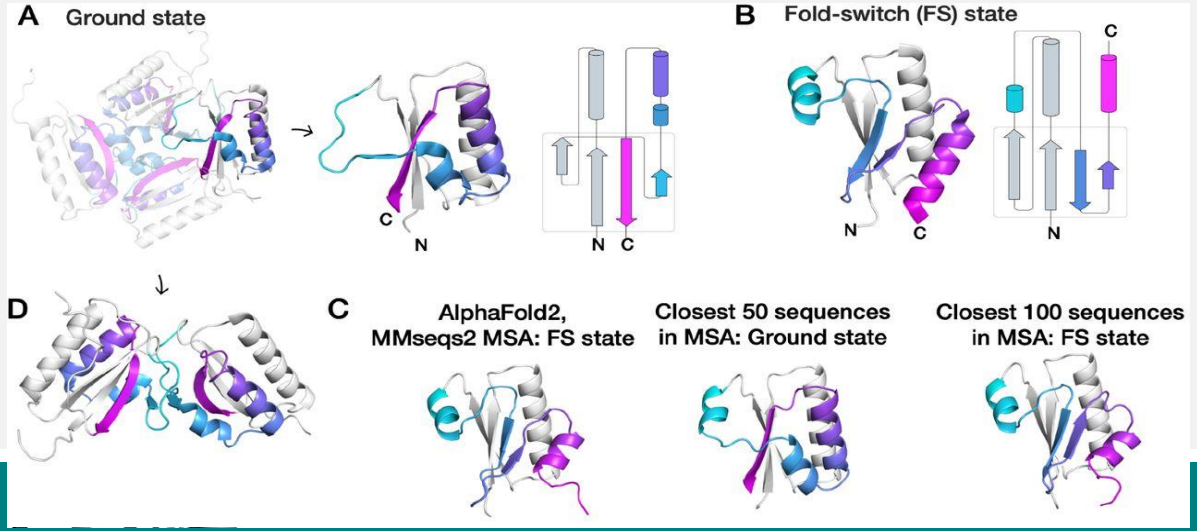
EMBL-EBI

Predicting multiple conformational states

Prediction of multiple conformational states by combining sequence clustering with AlphaFold2

Hannah K. Wayment-Steele, Sergey Ovchinnikov, Lucy Colwell, Dorothee Kern

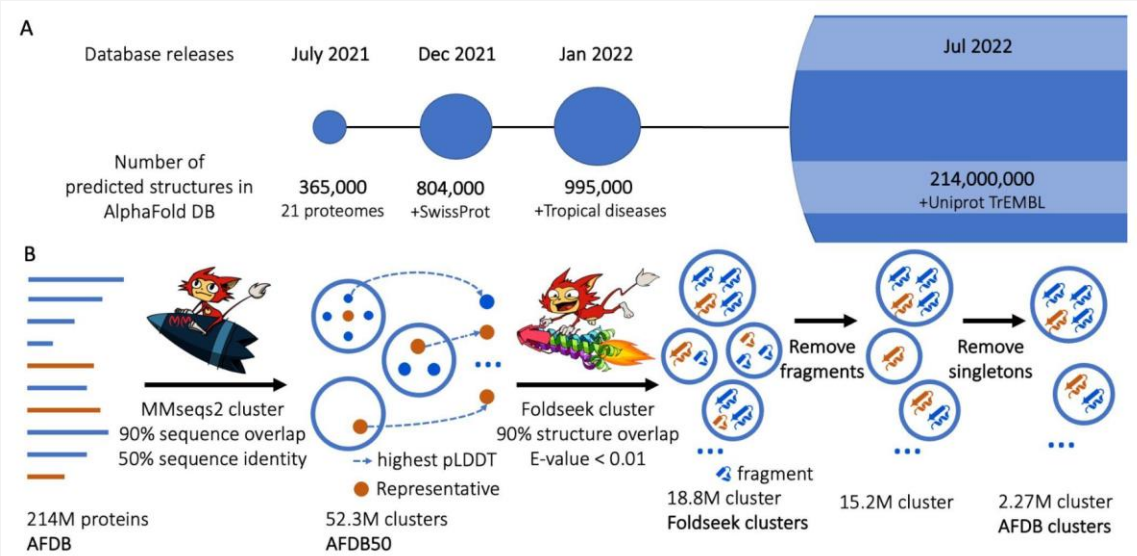
doi: https://doi.org/10.1101/2022.10.17.512570



Using only the closest 50 sequences by sequence distance returned from the MSA returns the ground state, but the closest 100 returns the fold-switch state.

Comparing and clustering structures at scale

- 2.27M non-singleton structural clusters.
  - >1.1M (50% of AFDB clusters) found to be, at least partially, similar to previously known structures in the PDB
  - 4% seem species specific
- 31% lack annotations representing likely novel structures
  - Tend to have few representatives covering only 4% of all proteins



Clustering predicted structures at the scale of the known protein universe

Inigo Barrio Hernandez[1]#, Jingi Yeo[2]#, Jürgen Jänes[3], Milot Mirdita[2], Cameron L.M. Gilchrist[2], Tanita Wein[4], Mihaly Varadi[1], Sameer Velankar[1], Pedro Beltrao[3,5]✉, Martin Steinegger[2,6,7]✉

What is hidden in the darkness?
Deep-learning assisted large-scale protein family curation uncovers novel protein families and folds

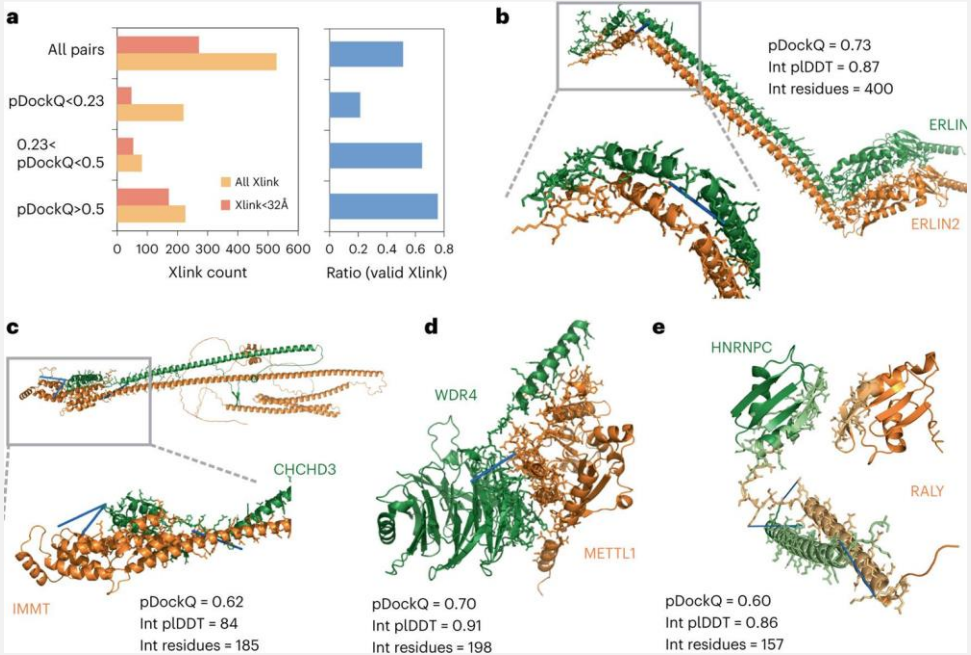
Janani Durairaj[1,2], Andrew Waterhouse[1,2], Toomas Mets[3], Tetiana Brodiazhenko[3], Minhal Abdullah[3], Mehmet Akdel[4], Antonina Andreeva[5], Alex Bateman[5], Vasili Hauryliuk[3,6,7], Tanel Tenson[3], Torsten Schwede[1,2], Joana Pereira[1,2]

EMBL-EBI

Large scale predictions of protein interactions

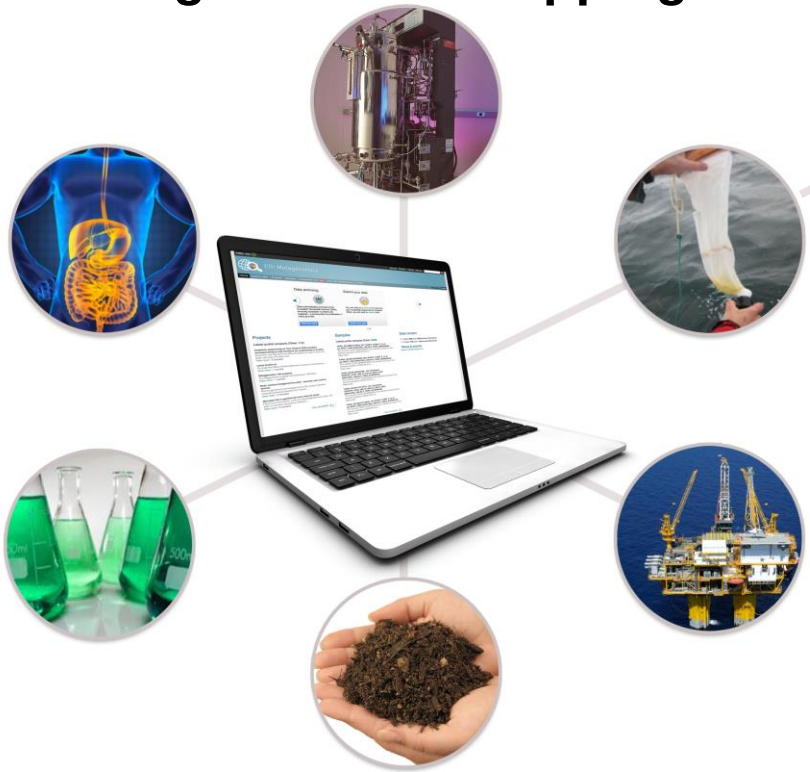
- Predicted the structures of **65,000** pairs of *human* proteins, obtained **3,000** high-confidence pairs
- **1,400** of the newly predicted high-confidence complexes lacked homology to existing structures

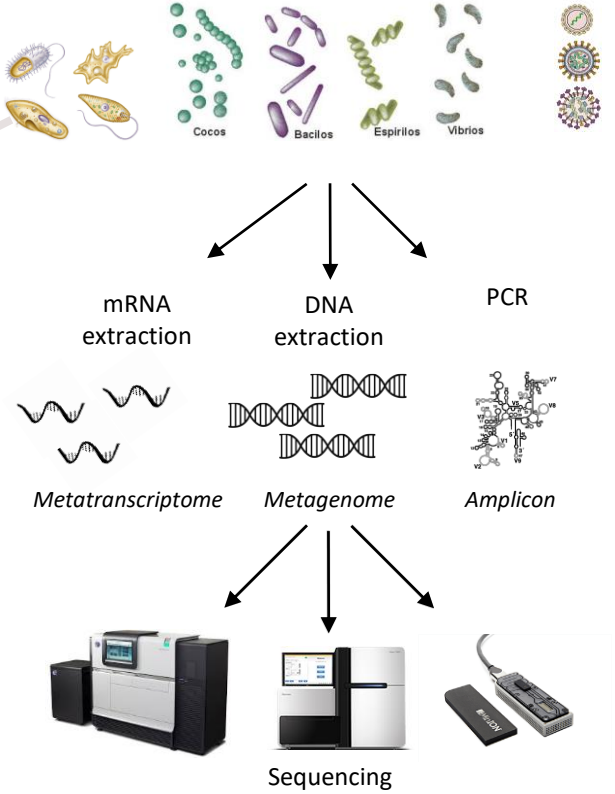
- <u>Cross-linking validation:</u>
  - Experimental evidence confirmed 171 out of 246 (**70%**) complexes predicted with high confidence



Burke, D. F. et al., Towards a structurally resolved human protein interaction network. Nat. Struct. Mol. Biol. (2023).

More and more protein sequences available: Metagenomics is tapping into the 99% of the unknown microbes





Case study - harnessing microbes for bioremediation

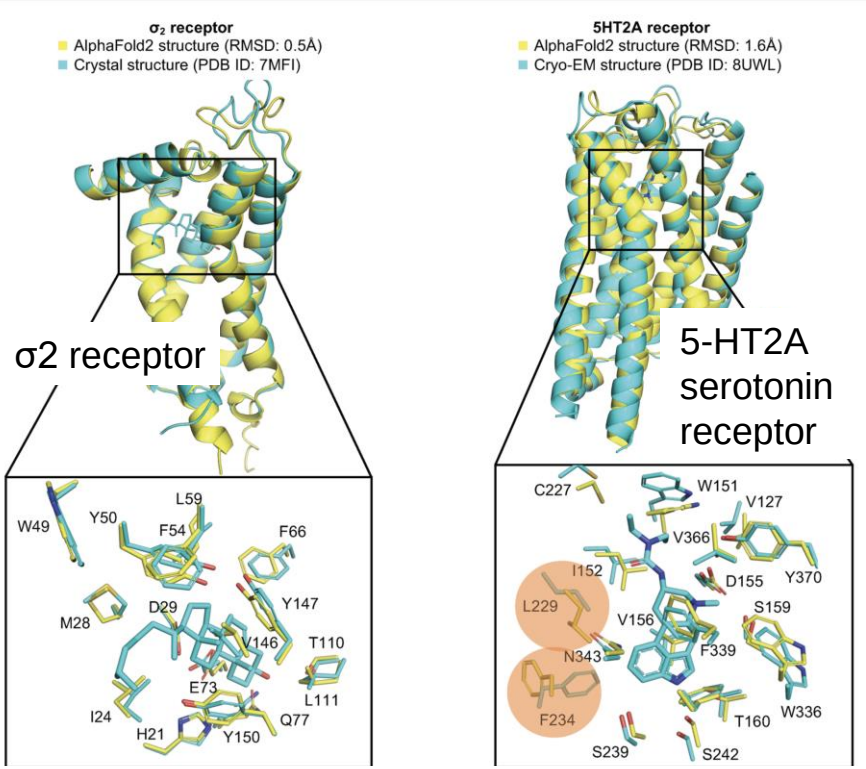
EMBL-EBI

Specific Applications in progress now

- Drug discovery
- Tackle neglected diseases
- Understand antibiotic resistance
- Fight plastic pollution
- Increase health of honey bees
- Understand how ice forms

Are AF2 models good for drug discovery?

- Retrospective simulation of structure-based ligand discovery may underestimate the ability of AF2 structures to template new ligand discovery prospectively
- The prospective large library docking campaigns against the AF2 models were no less effective than those against experimental structures
- The hit rates were high and were not significantly different between the modelled and experimental structures.
- Against the 5-HT2A serotonin receptor, the AF2 structure led, if anything, to more potent and selective compounds



AlphaFold2 structures template ligand discovery

Jiankun Lyu, Nicholas Kapolka, Ryan Gumpper, Assaf Alon, Liang Wang, Manish K. Jain, Ximena Barros-Álvarez,

New normal in structural biology

Improved AlphaFold modeling with implicit experimental information

Thomas C. Terwilliger[1,2*], Billy K. Poon[3], Pavel V. Afonine[3], Christopher J. Schlicksup[3], Tristan I. Croll[5], Claudia Millán[5], Jane. S. Richardson[6], Randy J. Read[5] and Paul D. Adams[3,5]

Accelerating crystal structure determination with iterative AlphaFold prediction

Thomas C. Terwilliger, Pavel V Afonine, Dorothee Liebschner, Tristan I Croll, A J McCoy, Robert D Oeffner, Christopher J Williams, Billy K Poon, Jane S Richardson, Randy J Read, Paul D. Adams

doi: https://doi.org/10.1101/2022.11.18.517112

- **Predicted model rebuilt in one cycle is used as a template for prediction in the next cycle.**
- **87% of the recent 215 structures yielded a model with at least 50% of Cα atoms matching those in the deposited models within 2Å.**

AlphaFold predictions: great hypotheses but no match for experiment

Thomas C. Terwilliger, Dorothy L Leibschner, Tristan Croll, Christopher J Williams, A J McCoy, Billy K Poon, Pavel Afonine, Robert D Oeffner, Jane Shelby Richardson, Randy J Read, Paul D. Adams

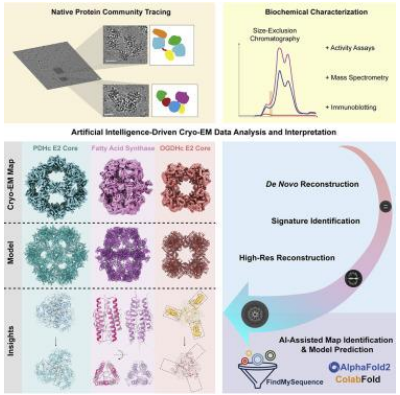
doi: https://doi.org/10.1101/2022.11.21.517405

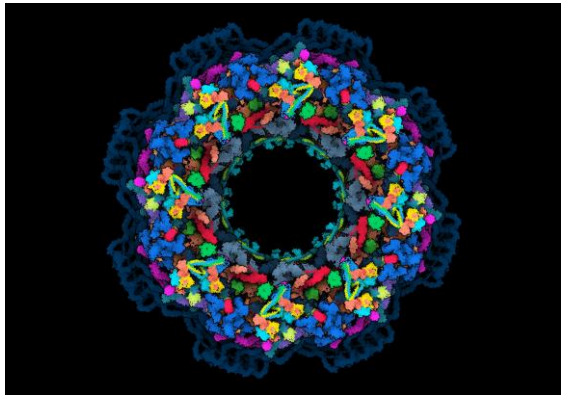
AlphaFold confidence (pLDDT)	Median prediction error (Å)	Percentage with error over 2 Å
>90	0.6	10
80 - 90	1.1	22
70 - 80	1.5	33
<70	3.5	77

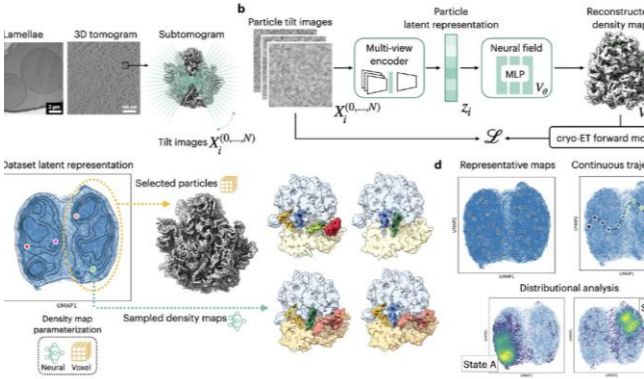
- Goal of structural biology is to shed light on e.g. biological function or mechanism of biological macromolecules/complexes, not “stamp collecting”
- Many structural studies can now start with a model as the null-hypothesis
- Design experiments based on models
  - Construct engineering to express individual domains without the “spaghetti”
  - Design and determine structures of mutants
  - Determine structures of complexes
  - Fragment/compound screens

EMBL-EBI

A(nother) Golden Age for Structural Biology: From Structural Inventories to Processes







High-resolution cryo-EM of cell extracts volume-based *de novo* identification and AI-assisted atomic modeling

Skalidis, Ioannis, et al. "Cryo-EM and artificial intelligence visualize endogenous protein community members." *Structure* 30.4 (2022): 575-589.

Total mass: 120 Mda

Old model: 35 Mda

New model: 70 MDa

Artificial intelligence reveals nuclear pore complexity

Shyamal Mosalaganti, Agnieszka Obarska-Kosinska, Marc Siggel, Beata Turonova, Christian E. Zimmerli, Katarzyna Buczak, Florian H. Schmidt, Erica Margiotta, Marie-Therese Mackmull, Wim Hagen, Gerhard Hummer, Martin Beck, Jan Kosinski

doi: https://doi.org/10.1101/2021.10.26.465776

Visualisation of structural landscape and dynamics

Deep reconstructing generative networks for visualizing dynamic biomolecules inside cells

Ramya Rangan, Sagar Khavnekar, Adam Lerer, Jake Johnston, Ron Kelley, Martin Obr, Abhay Kotecha, Ellen D. Zhong

doi: https://doi.org/10.1101/2023.08.18.553799

Electron microscopy/AI-assisted modelling

Integrative/hybrid methods

Structure dynamics in their biological context (visual proteomics)

EMBL-EBI

Visual Proteomics : A Proof of Concept with *Chlamydomonas reinhardtii*

PLANT: **The Chloroplast**

Thylakoids: Light Harvesting
Pyrenoid: Carbon Fixation

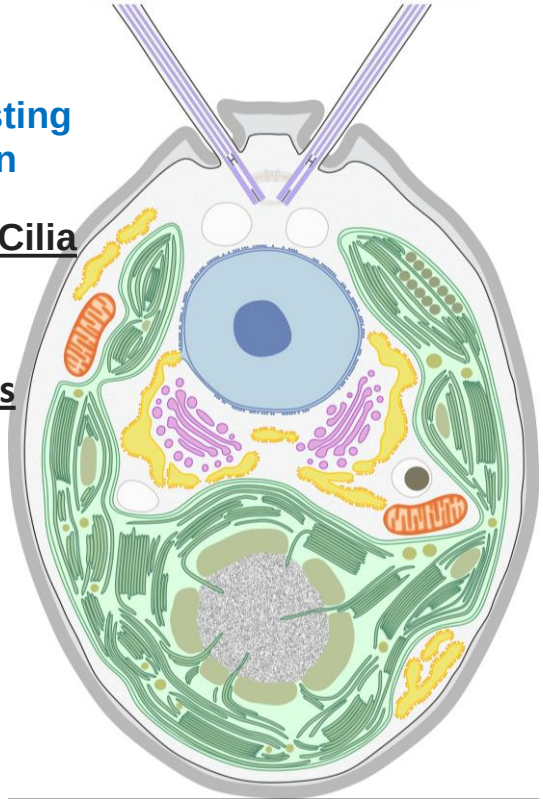
ANIMAL: **Centrioles and Cilia**

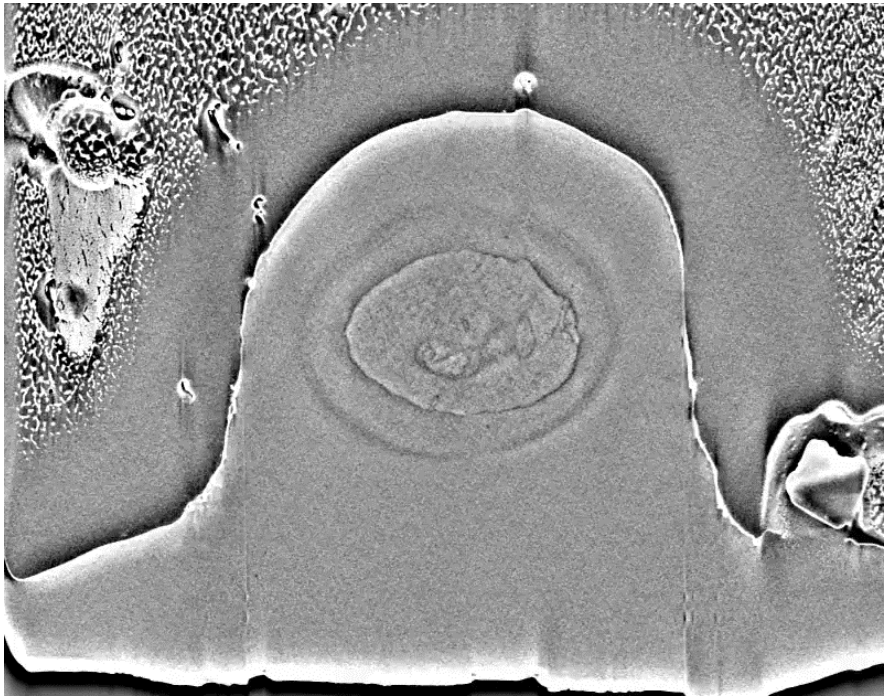
"Core" Eukaryotic Organelles

Nucleus

Endoplasmic Reticulum & Golgi Apparatus

Mitochondria

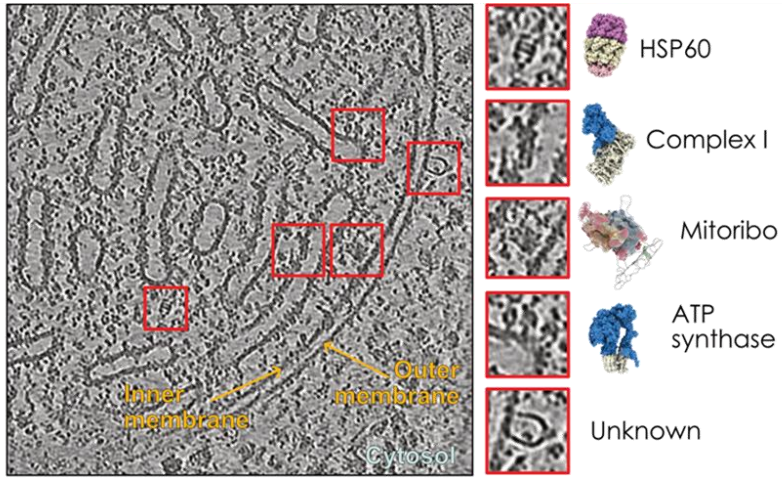


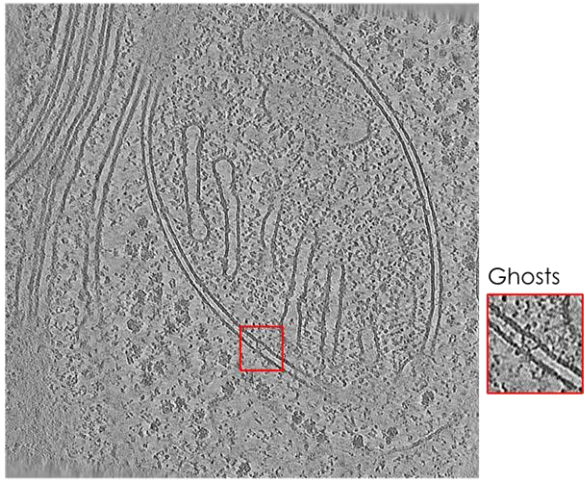


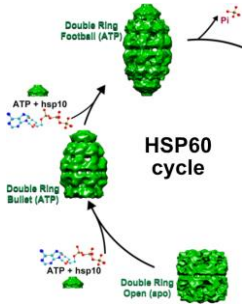
cryo slice and view of a single *Chlamy* cell

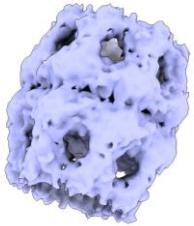
In situ structural cell biology of mitochondria

Need better identification of known and rare targets, *need better classification*

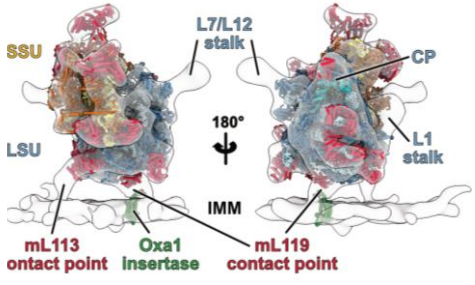






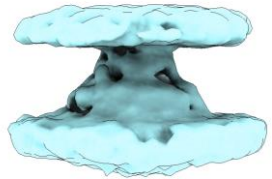


HSP60



Mitoribosomes

ATP Synthase



Something ?

Collaboration with Uni. Basel

AI/Data in the Medical Sciences

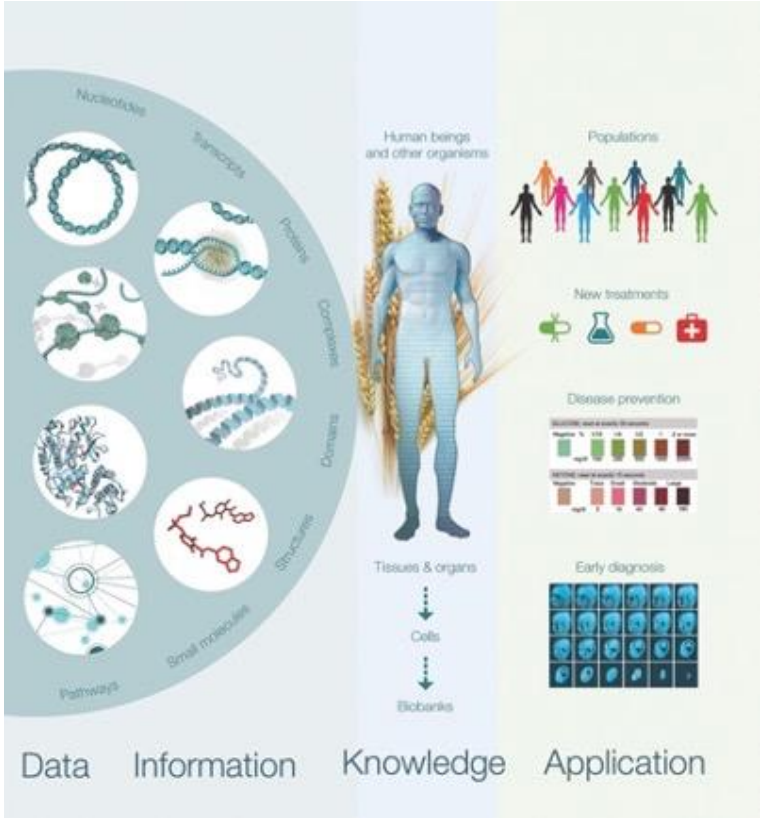
In many ways, medicine is an ideal arena for the application of AI:

- Lots of data
- Many expert clinicians
- Highly regulated profession

- Current approach is mainly based on previous experience

- Imaging is critically important

BUT



AI & Medicine: Diagnosis & Prognosis

- Long standing important problem – ✔
- Lots of data – ✔
  - Data are curated, clean, accurate and freely available ✕
  - Problem is well defined: ✕
  - Good metrics to measure success - ?? Maybe in extremis
- Evolution
  - Many diseases are genetic – BUT impact of environment – more complicated
  - Different species often behave differently eg mouse & human
- Many helpful technical developments in last 30 years - SOME, but underlying causes of diseases often unknown
- CASP ✖

Major challenges:- DATA quality: lack of standards: country specific: data confidentiality: fragmentation of medical ‘disciplines’; Ethics: Implementation etc etc;

Scale is enormous!

BEGIN WITH DATA

HDRUK
Health Data Research UK

AI & Data for Medicine

Mission is to accelerate trustworthy data use to enable discoveries that improve people's lives

25/04/2024

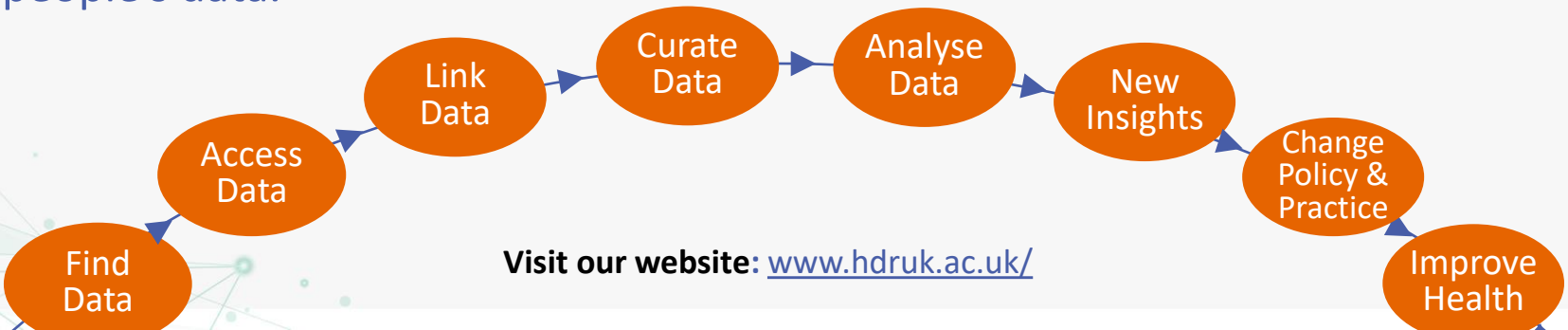
Professor Dame Janet Thornton

Member of HDR UK's Board of Trustees

Why is HDR UK needed?

HDRUK
Health Data Research UK

- In principle the NHS has cradle-to-grave records for 65 million people
- Access to this data for researchers is often a lengthy, fragmented process. It can take years to find, access and ‘wrangle’ data from different places into a state that researchers can use for analysis
- HDR UK is working with multiple partners to make it easier for researchers to find, access and work with the data they need to generate new discoveries, using streamlined systems that are designed to ensure the privacy and security of people’s data.



Visit our website: www.hdruk.ac.uk/

RETFound: Using retinal imaging to predict and detect disease

HDRUK
Health Data Research UK

- Detects eye diseases and predicts risk of health conditions

  e.g., Parkinson's disease, heart attack and strokes

- One of the first AI foundation models in healthcare and the first in ophthalmology
- Pre-trained on 1.6m retinal images
- Adaptable and validated in multiple disease detection tasks
- Freely available for use on GitHub

"If the UK can combine high quality clinical data from the NHS, with top computer science expertise from its universities, it has the true potential to be a world leader in AI-enabled healthcare. We believe that our work provides a template for how this can be done."

- Professor Pearse Keane

Learn more: Eye-scanning AI predicts and detects disease in world-first, study suggests - HDR UK

Foresight: Predicting future health of patients

HDRUK
Health Data Research UK

- Predicts medical events, aids decision making and informs clinical research
- Trained using NHS ECRs from over 811,000 people in the UK
- Precision rates of 68% to 88%
- Clinically validated

  Predictions were 93% relevant

"One of the main advantages of Foresight is that it can easily scale to more patients, hospitals or disorders with minimal or no modifications, and the more data it receives, the better it gets."

- Zeljko Kraljevic

Learn more: New AI tool may offer insights into patients' future health - HDR UK

The importance of high-quality, representative data for AI models

HDRUK
Health Data Research UK

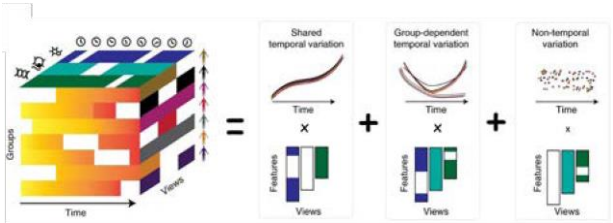
- Analysed ethnicity data from GP and hospital records of over 61m people
- Highlighted that ethnicity information was missing for almost 1 in 10 patients
- First phase of a three-phase project aiming to reduce bias in AI health prediction models

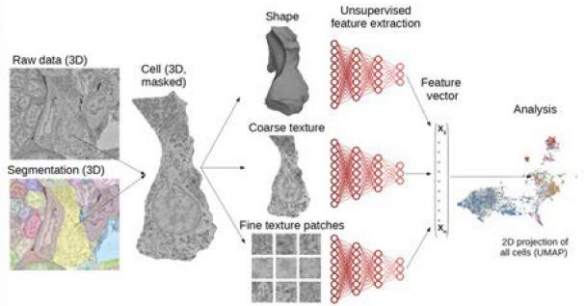
"Because AI-based healthcare technology depends on the data that is fed into it, a lack of representative data can lead to biased models that ultimately produce incorrect health assessments. Better data from real-world settings, such as the data we have collected, can lead to better technology and ultimately better health for all.'

- Professor Sara Khalid

Learn more: Ethnicity data from health records of over 61 million people studied in detail for first time - BHF Data Science Centre

Today: Opening up the world's Biodata for AI





Macromolecular Structures

Protein design & data generation informed by AI

Multi-modal integration

of multiple modalities across space and time (omics, imaging, text…)

Imaging

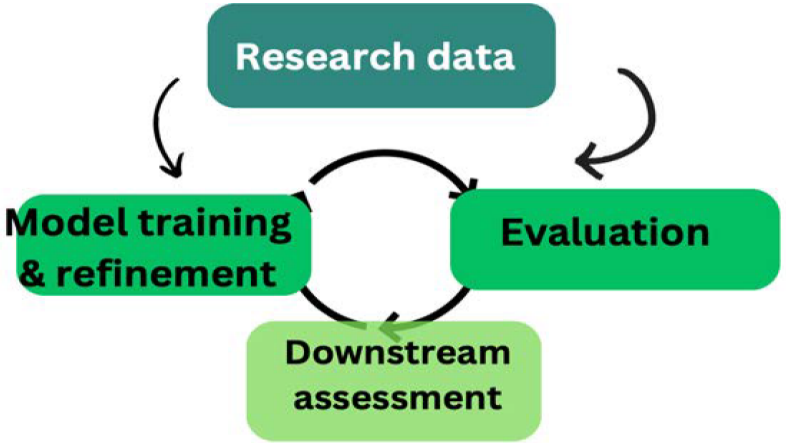
AI-driven cellular imaging models across scales

The Future?

HDRUK
Health Data Research UK

Ingredients of (today's) AI advances in biology

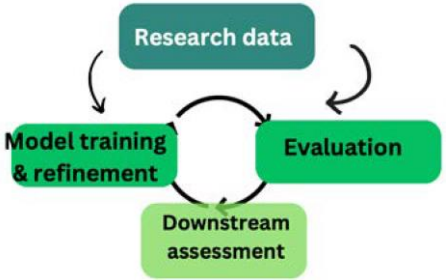
- Well-defined question & metrics
- Computable, high-quality data
- Creative AI developers
- Sufficient compute capacity
- Continuous model benchmarking
- Community assessment of model outputs in practice



=> Future needs the same for Medical Research

EMBL

The Evolution of AI Paradigms



Task-specific AI

Often building on ground truth labels. Incompasses AlphaFold and many EMBL AI tools today.

Foundation models

Task agnostic AI.
Trained on large volumes of data, typically label free.

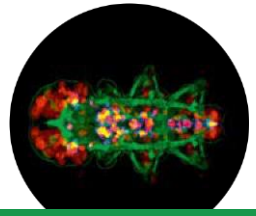
Lab-in-the-loop AI

Tie data generation and technology to AI.

future

EMBL-EBI

AI spans all aspects of any organisation (eg. EMBL)



Research

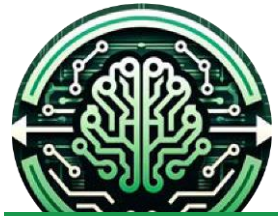
- Application of AI to scientific discovery
- Lab-in-the-loop AI
- Rigorous, explainable, causal AI modelling & theory

Services

- Enhance EMBL data resources through AI
- Deliver AI-ready data resources

European coordination

- Coordination of European AI networks
- Engage in European AI policy



Technology transfer

- Initiate new pre-competitive partnerships in BioAI
- Set examples for impactful collaborations with industry

Training & ethics

- Enhance training using AI
- Interdisciplinary training: AI & life sciences
- Guidelines and best practices

Operations

- Information management
- Building maintenance
- Space usage
- Controlling & legal

Biology Research and Medical Research

- It has taken the last 30 years to develop the data infrastructure for **biological research**, characterized by:
  - Data Sharing – culture and deposition tools
  - FAIR data
  - Agreed standards
  - A public data infrastructure – many different types of data
  - Developments of many new methods

Huge progress in understand- though still a long way to going of biological processes

- **Medical Research** is just starting on this journey
  - Data sharing remains challenging – in part because of patient confidentiality but also culture
  - Scale is at least ten times bigger
  - Data are even more siloed than biological data – by country/even by hospital
  - Data infrastructure just now beginning to develop – TREs, data wrangling etc
  - Potential commercial gains
  - Systems biology

BUT – this journey is underway and I have no doubt that in the next 10-20 years, progress will be enormous and affect every one of us both scientifically and personally

Acknowledgements

Sameer Velankar & Colleagues at EMBL

Scientists at DeepMind

PDBe & AlphaFold Database@EMBL-EBI

Health Data Research UK

Oliver Stegle (EMBL-HD)

My group

Sameer Velankar

DeepMind

wellcome trust

EMBL European Molecular Biology Laboratory

EMBL-EBI