Accelerating scientific discovery with AI

Demis Hassabis

Google DeepMind

Nobel Lecture - 8th Dec 2024

Thinking about thinking

Games AI as a critical stepping stone

DeepMind was founded in 2010 as a research lab to build AGI

Initially we used games as the perfect proving ground for AI

Go is the most complex game ever - the pinnacle for games AI

10^{170} possible positions > atoms in the universe!
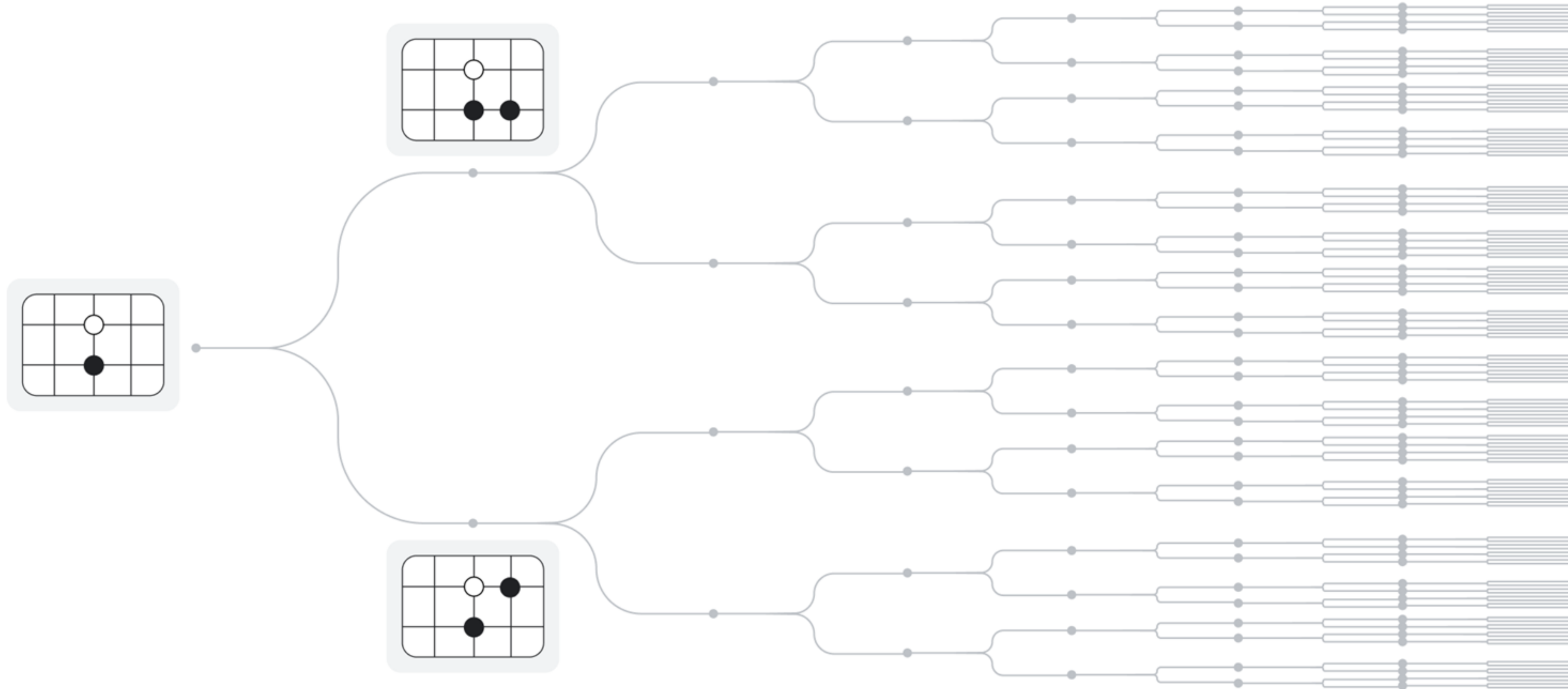
In 2016 AlphaGo beat Lee Sedol 4-1 in a famous match in Seoul

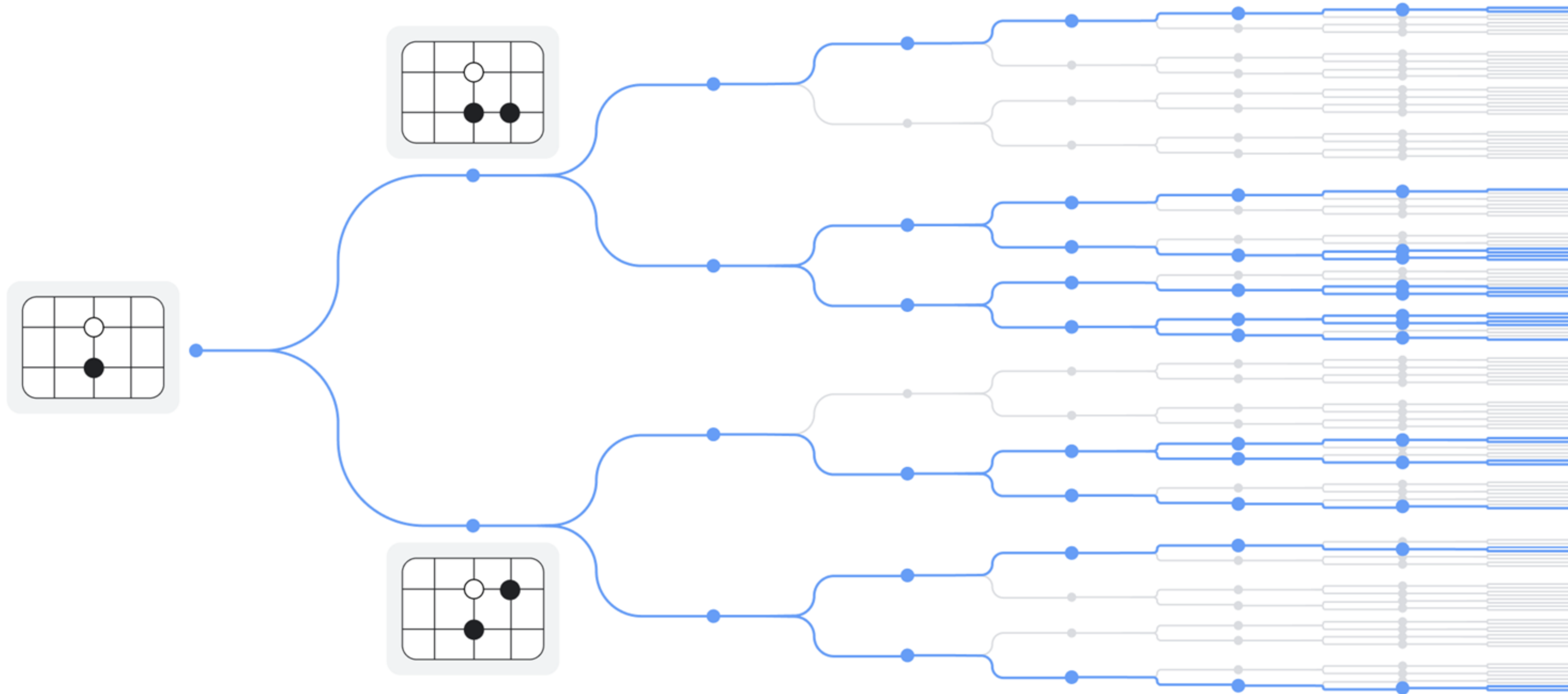But didn't just win, came up with novel creative strategies (Move 37)

AlphaGo mastered Go through self-learning: playing against itself and learning from its mistakes to build a useful model of Go
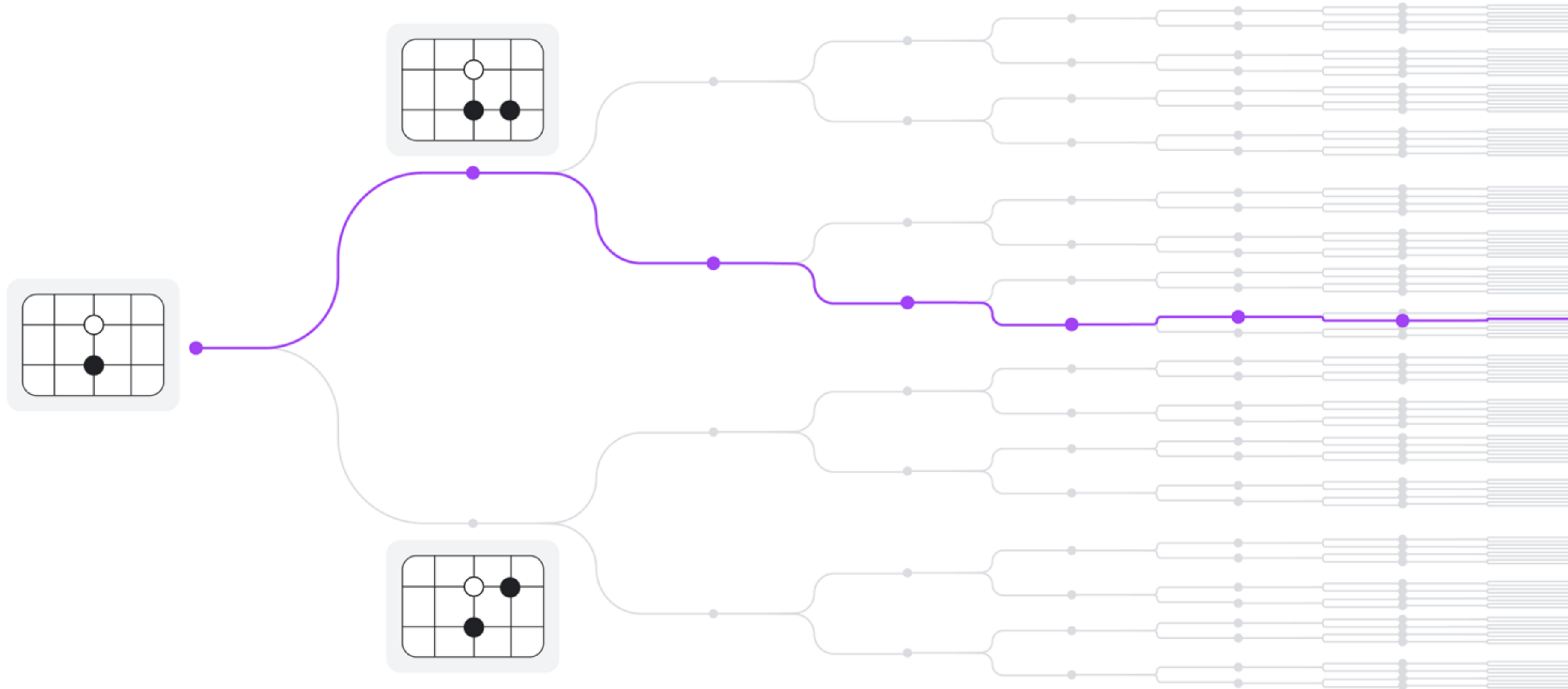

Lee Sedol - 18 world titles

Game 2 - Move 37

Learning a neural network model to efficiently guide the search

Learning a neural network model to efficiently guide the search

Learning a neural network model to efficiently guide the search

What makes for a suitable problem for AI?

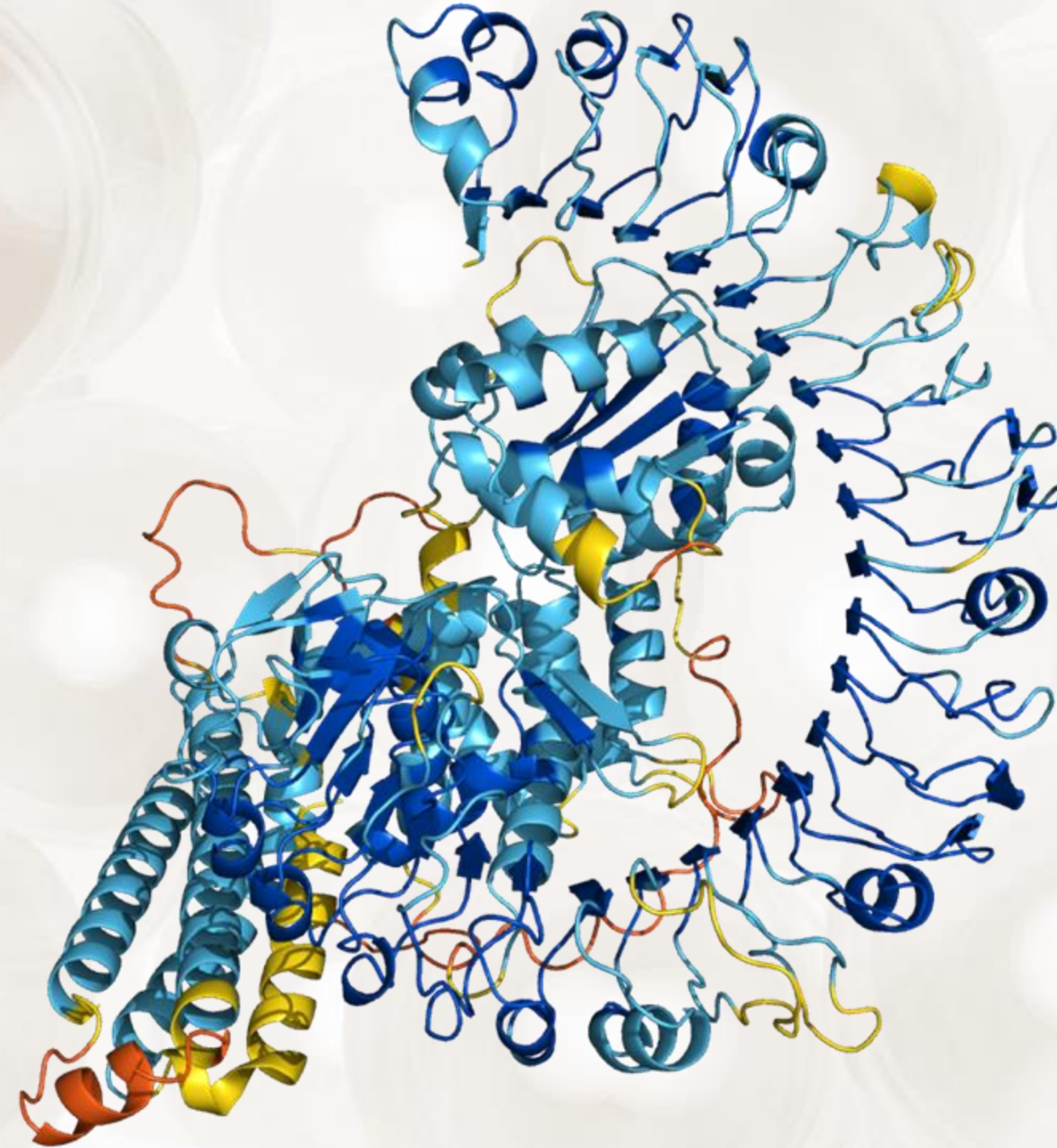**1** Massive combinatorial search space

2 Clear objective function (metric) to optimise against

3 Either lots of data *and/or* an accurate and efficient simulator

Proteins are the building blocks of life

The Protein Folding problem

Predicting the 3D structure of a protein directly from its 1D amino acid sequence

Amino Acid Sequence

MANKGRLGEQEAFAMTVPLLIDVDSSQQEALEAIALNDELVRVRRALYLDLGVPF
PGIHLRFNEGMGEGEYLISLQEVPVARGELKAGYLLVRESVSQLELLGIPYEKGEH
LLPDQETFWVSVEYEERLEKSQLEFFSHSQVLTWHLSHVLREYAEDFIGIQETRY
LLEQMEGGYGELIKEVQRIVPLQRMTEILQRLVGEDISIRNMRSILEAMVEWGQKE
KDVVQLTEYIRSSLKRYICYKYANGNNILPAYLFDQEVEEKIRSRVRQTSAGSYLA
LDPAVTESLLEQVRKTIGDLSQIQSKPVLIVSMDIRRYVRKLIESEYYGLPVLSYQE
LTQQINIQPLGRVCL
MGHHHHHHGNITLTKRQQEFLLLNGWLQLQCGHAERACILLDALLTLNPEHLAG
RRCRLVALLNNNQGERAEKEAQWLISHDPLQAGNWLCLSRAQQLNGDLDKARH
AYQHYLELKDHNESP
GAMGALPPDGHPVEPHLERLYPTAQSKRSLWDFASPGYTFHGLHRAQDYRREL
DTLQSLLTTSQSSELQAAAALLKCQQDDDRLLQIILNLLHKV
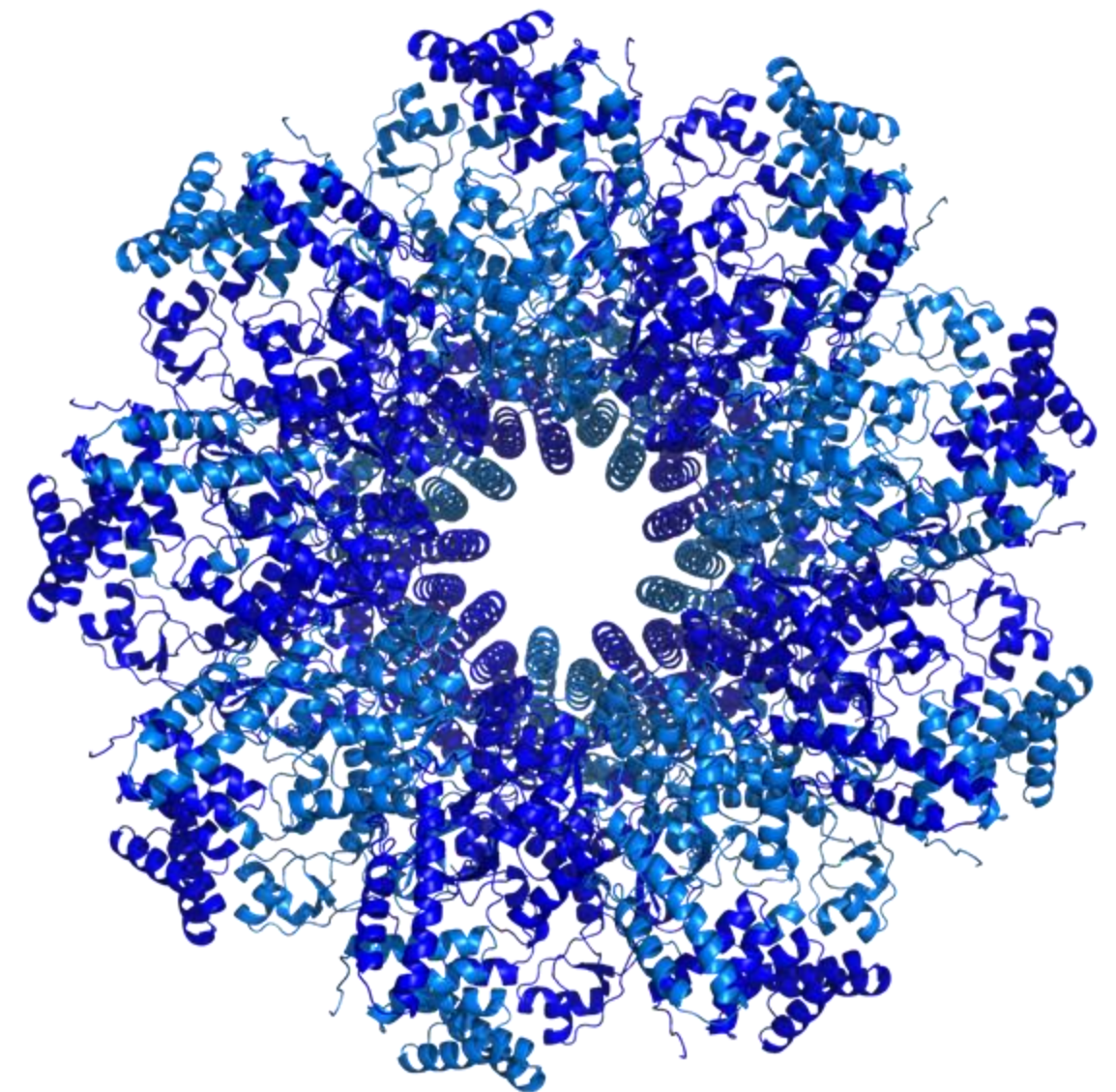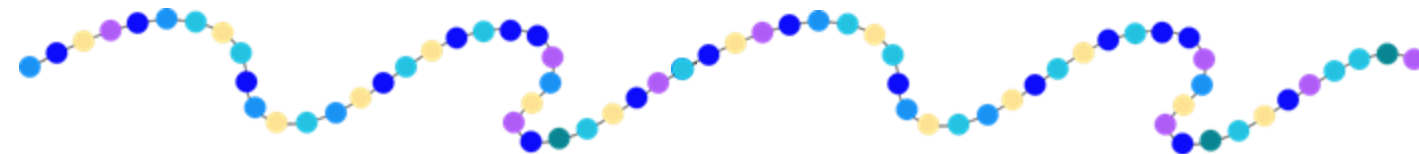
"thermodynamic hypothesis": a protein will assume the conformation that minimises the free energy in the system

Christian Anfinsen, Nobel lecture 1972

The Protein Folding problem: a 50-year grand challenge in biology

Why is this such a hard problem?

Experimentally determining even a single protein structure can often take years of painstaking work

Levinthal's Paradox: estimated ~10^{300} possible conformations that a typical protein can take

Enumerating all possibilities would take longer than the age of the universe

Yet somehow in nature proteins fold spontaneously within *milliseconds*

Must mean there's some topology in the energy landscape that guides the folding process
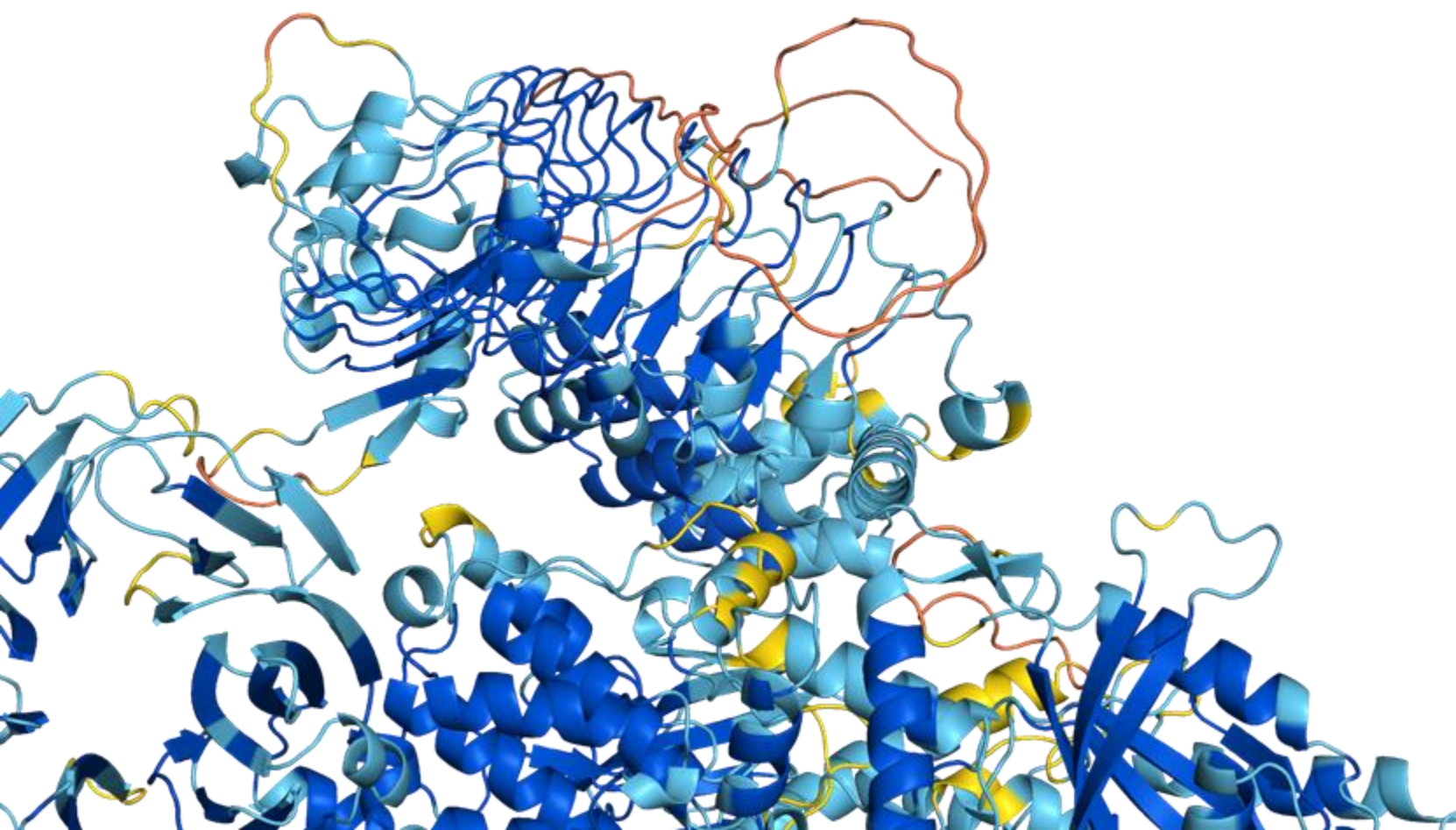
Data and benchmarks

After decades of experimental work ~170,000 structures had been determined and collated in Protein Data Bank (PDB), an incredible resource that we used as a starting point to train AlphaFold
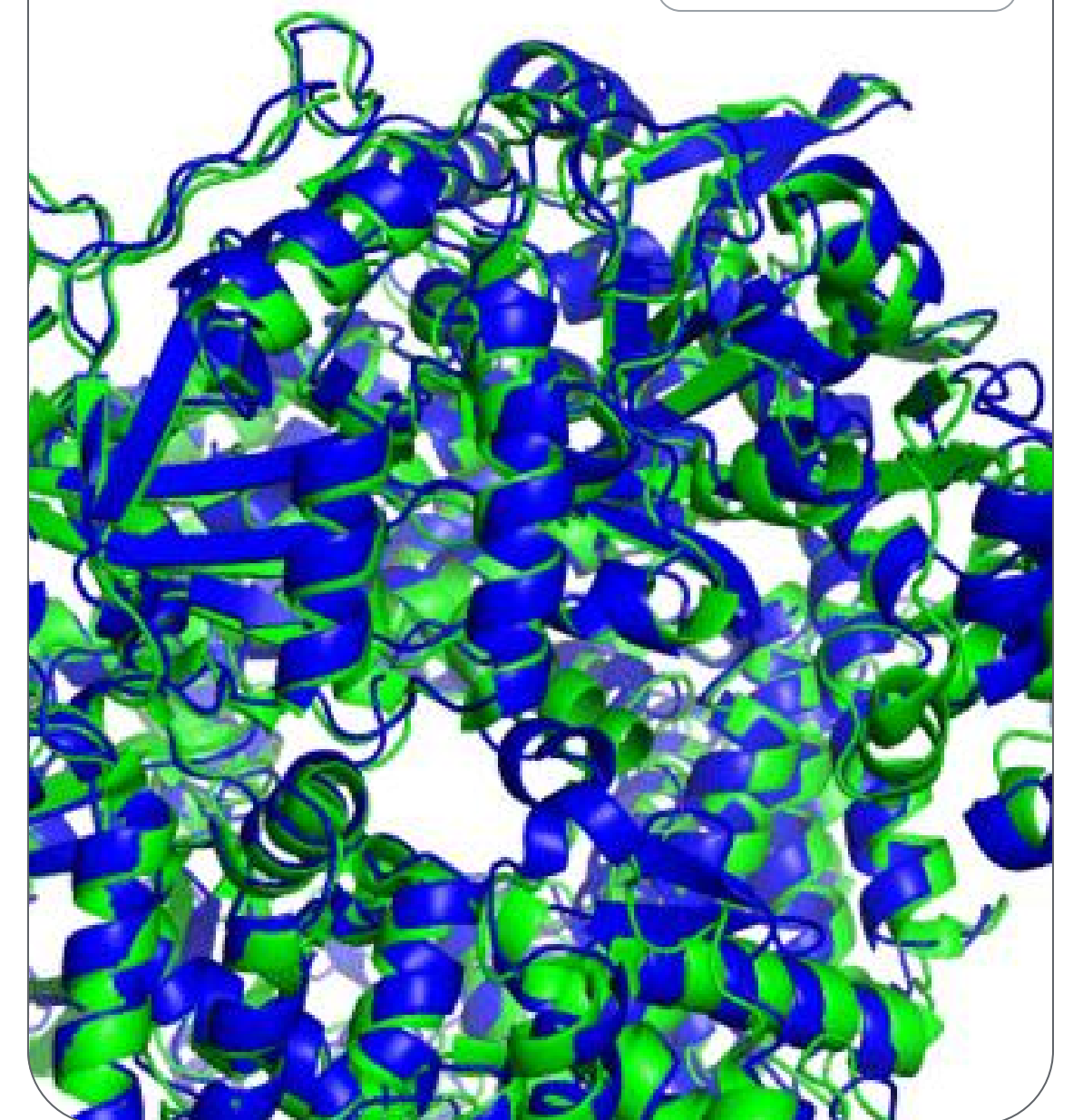
CASP competition: gold standard benchmark, run biennially since 1994 by John Moult, to test the best computational systems

Critical threshold: predictions had to be accurate down to the width of an atom (<1.0Å average error) to be of use to experimentalists
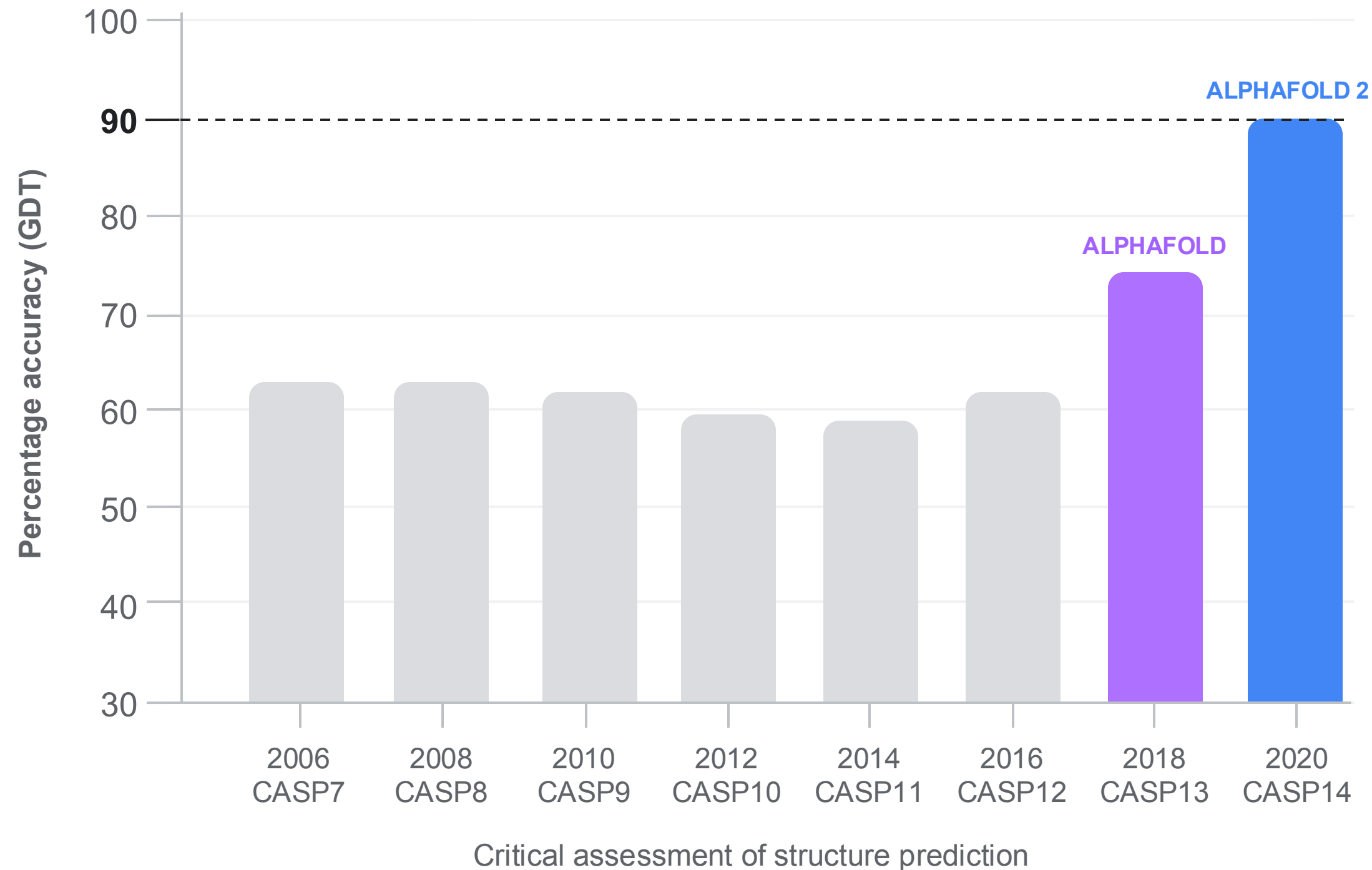
<1 atom width apart

● **Ground truth**
● **Prediction**

# AlphaFold2 achieved *atomic accuracy* at CASP14 (2020)

Mean GDT score for top team (avg across protein difficulty)

Not much progress made for a decade. Advanced the field with AlphaFold1 at CASP13 by introducing cutting edge ML as a main component for first time

Re-architected AlphaFold2 for CASP14 to reach *atomic accuracy* - leading the organisers to declare the problem solved!
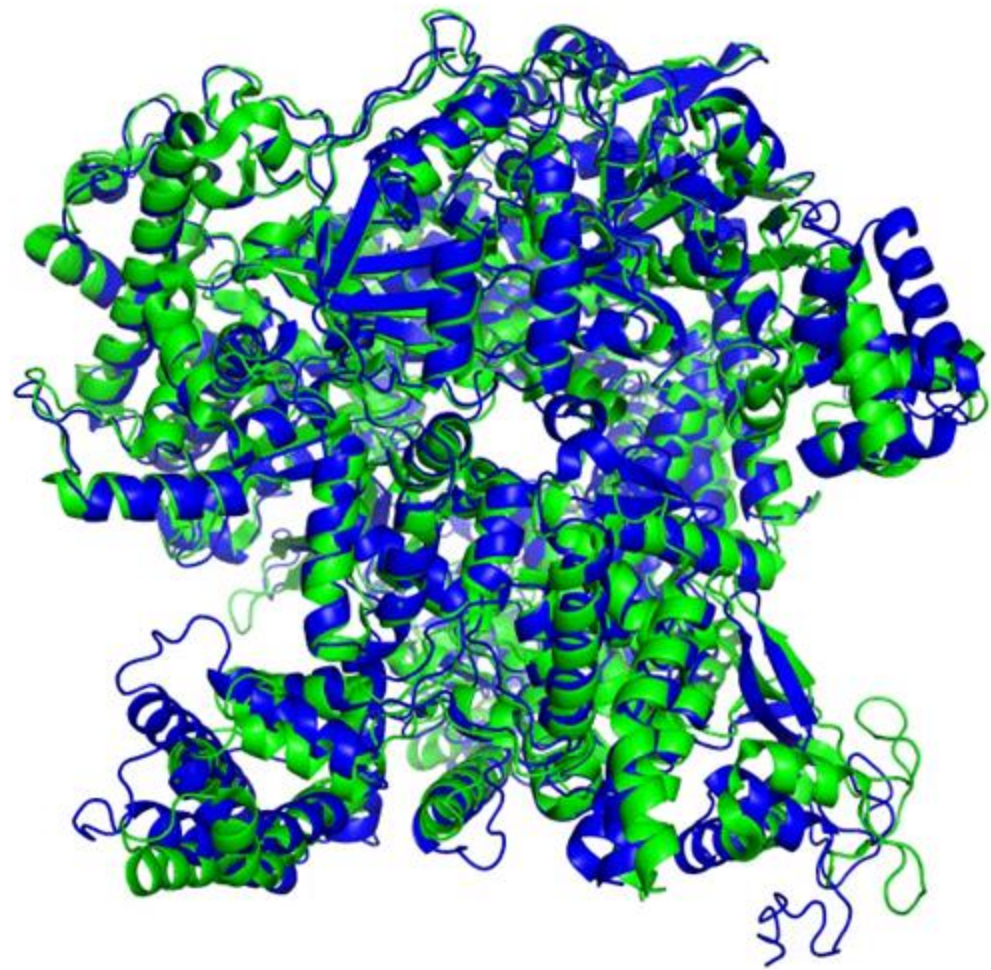
Innovate architecture of AlphaFold2: a hybrid system

(Jumper et al., Nature 2021)
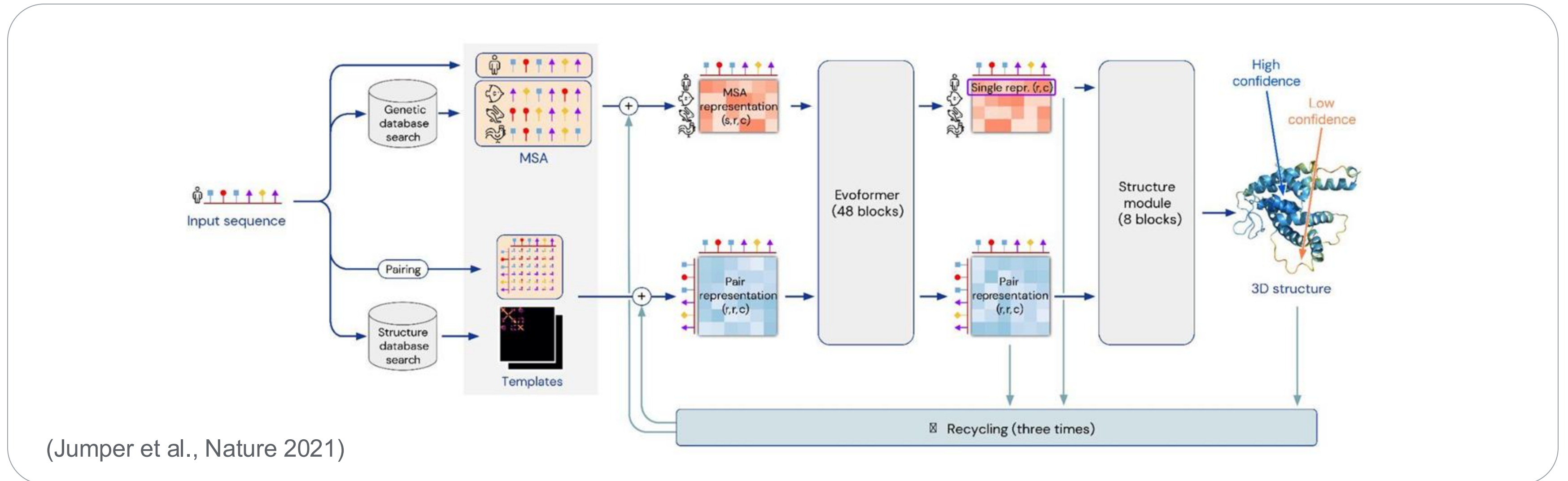
Built in evolutionary and physical constraints into the architecture without impacting the learning

Full end-to-end system with recycling stage that iteratively refines the structure prediction

AlphaFold's iterative steps toward a protein structure prediction

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

1

Impact of AlphaFold so far

Folded all 200M+ proteins known to science

Open-sourced AlphaFold and built the AlphaFold database with EMBL-EBI, to provide free and unrestricted access to every predicted structure

Consulted with 30+ biosecurity & bioethics experts to ensure the benefits outweighed any risks

2M+ researchers from 190 countries have made use of AlphaFold - 30000+ citations, standard tool

(Jumper et al., Nature 2021)

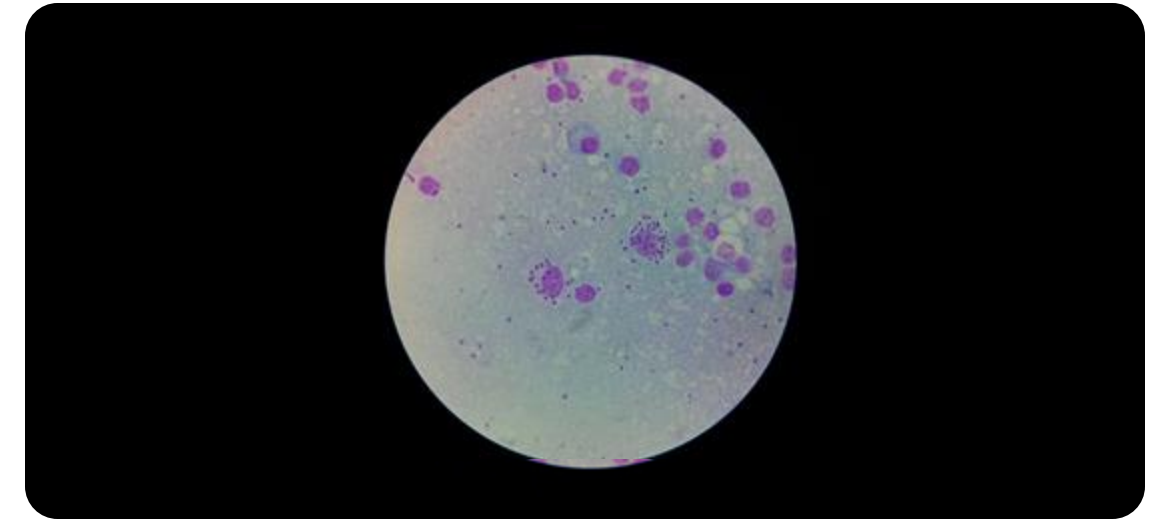# AlphaFold is accelerating progress on a huge range of problems

Plastic pollution

Designing enzymes that can digest plastic
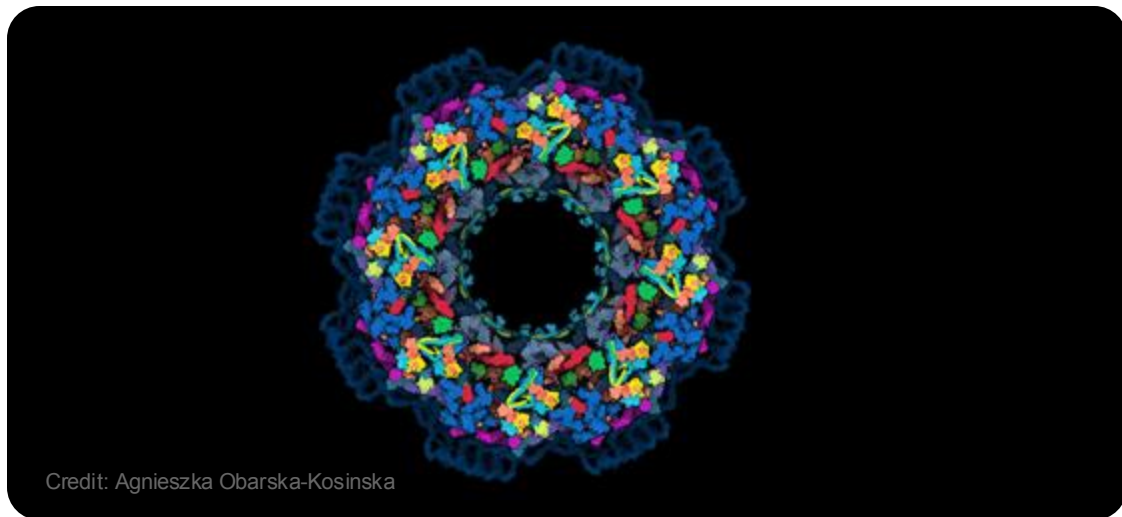
McGeehan et al. (Portsmouth)

Antibiotic resistance

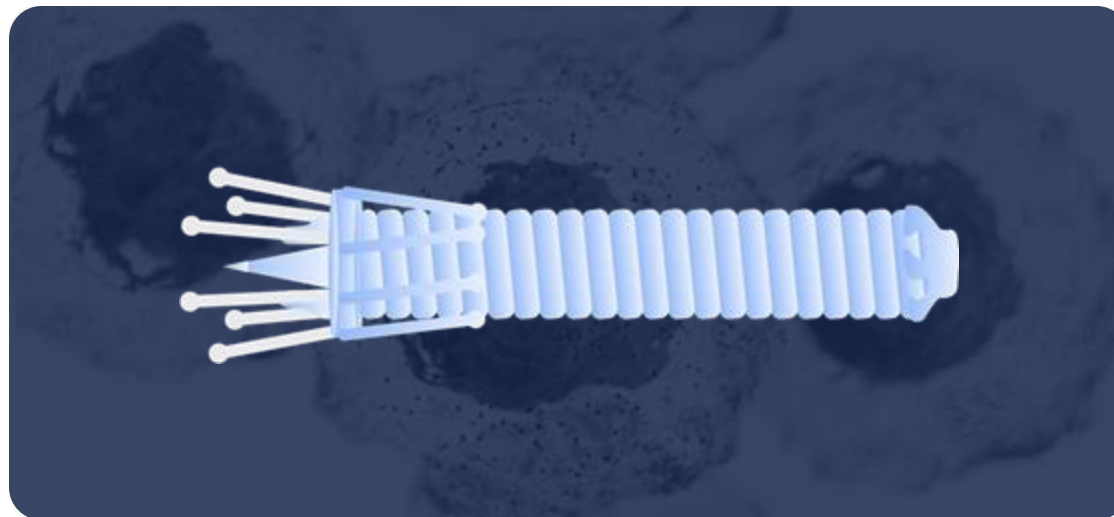Tackling the causes of antibiotic resistance

Sousa & Mitchell (Colorado)

Neglected diseases

Accelerating cures for underfunded diseases

Perry (DNDi), Higgins (Oxford)

Credit: Agnieszka Obarska-Kosinska

Structural biology

Determined structure of nuclear pore complex

Fontana et al., Mosalaganti et al. (Science)

Drug delivery

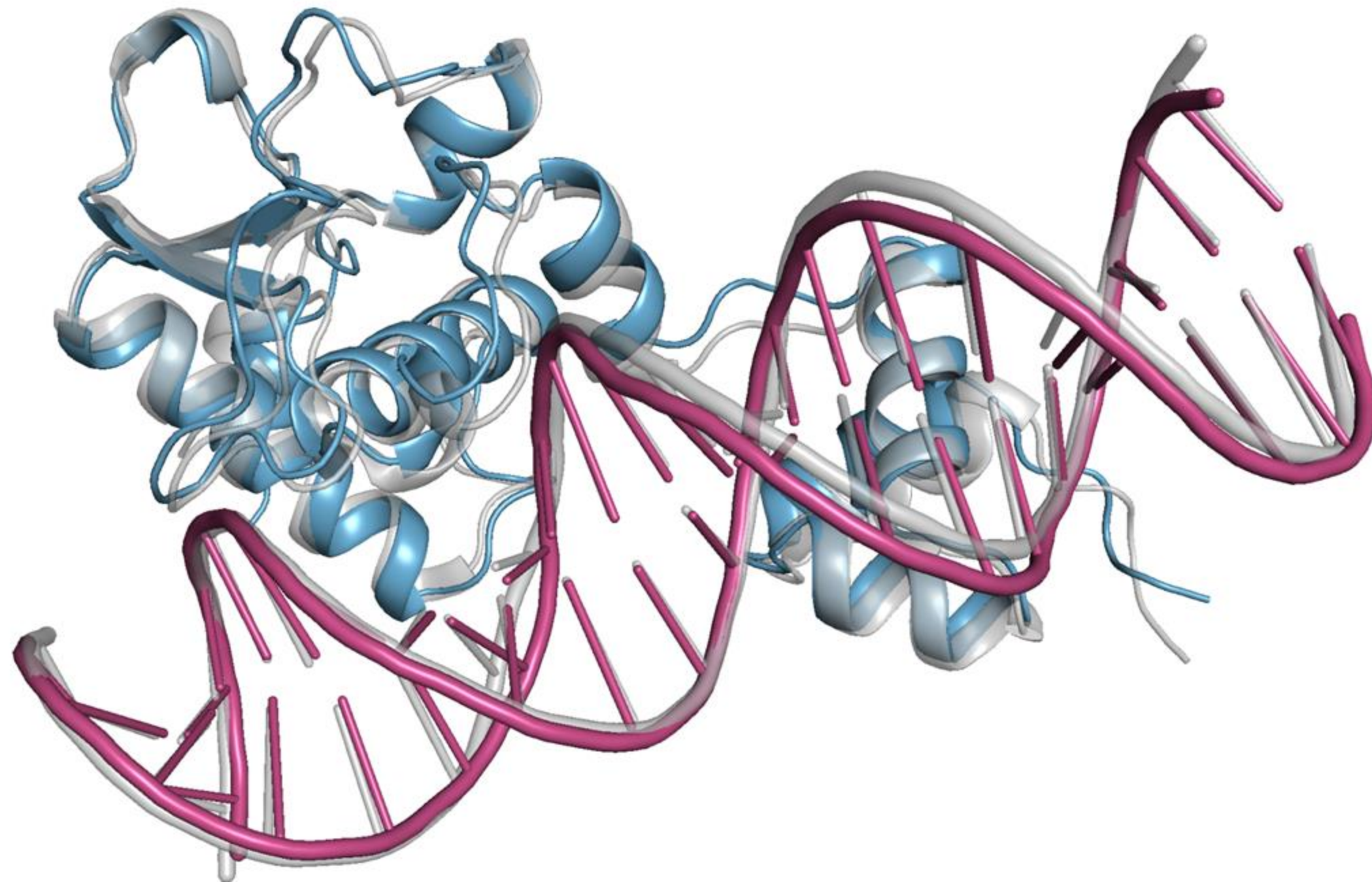Molecular protein syringe for payload delivery

Zhang et al. (Broad Institute)

Fertilisation mechanism

Discovery of mechanisms in reproduction

Pauli (Vienna BioCenter)

AlphaFold 3

Modelling all of life's biomolecules

(Abramson et al., Nature 2024)

Implications
for the bigger picture

Making search tractable

Taking a step back, what is the essence of what our systems are doing?

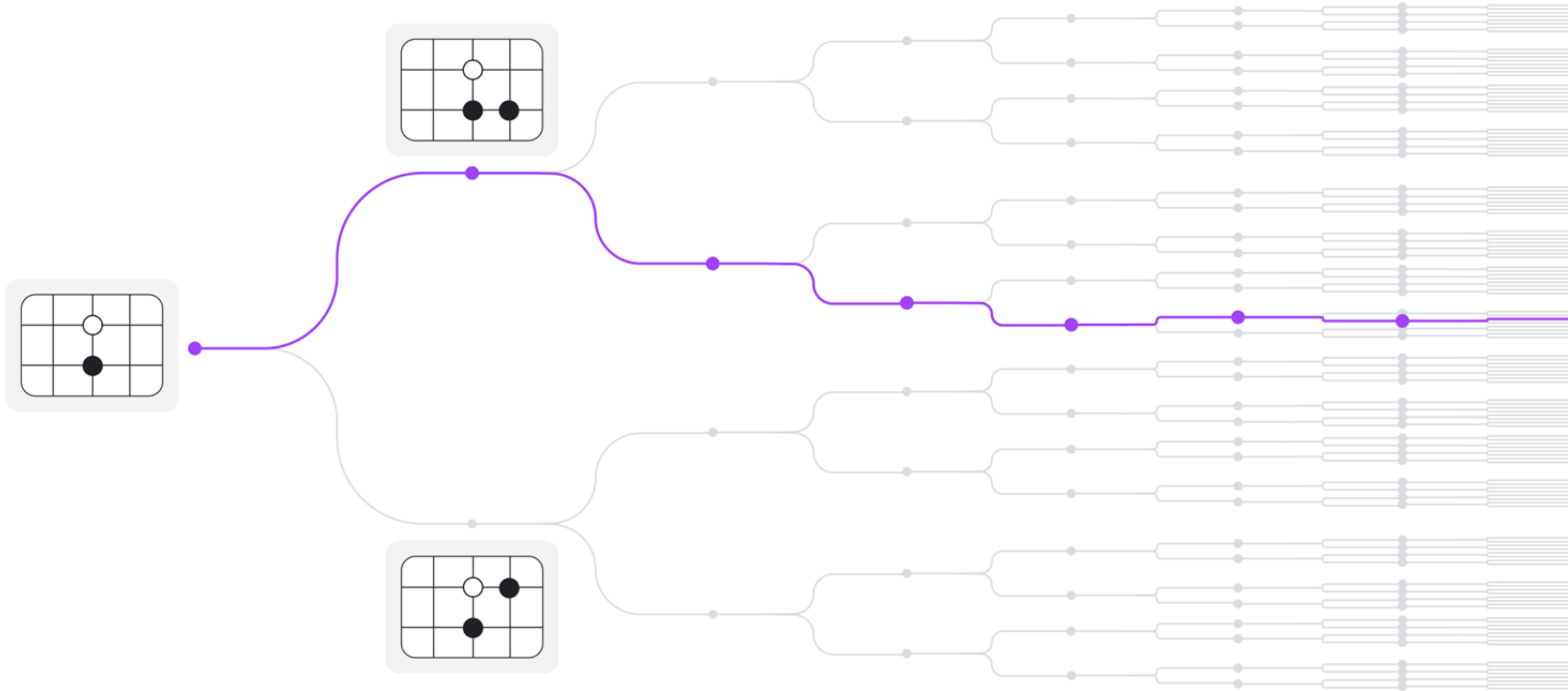Finding the optimal solution in an enormous combinatorial space

Learn a model of that environment (from data or simulation)

Use that model to guide a search according to an objective function
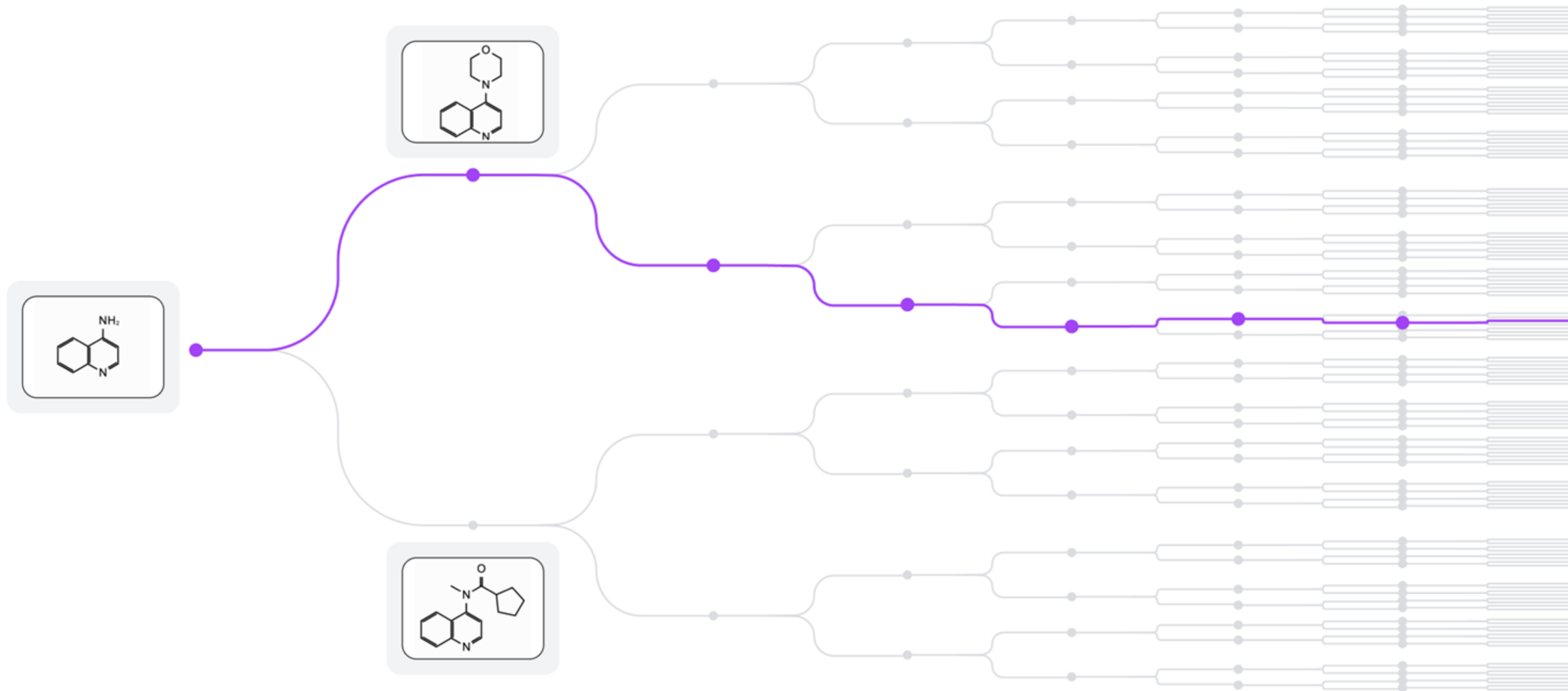
Turns out this is a very general solution and many problems fit this approach

Finding the best Go move

Finding the best molecule in chemical space

The new era of 'digital biology'

Biology at its most fundamental level can be thought of as an information processing system albeit a phenomenally complex and emergent one

AI is potentially the perfect description language for biology

AlphaFold is a proof point that could usher in a new era of 'digital biology'

Started **Isomorphic Labs** to build on AlphaFold and reimagine the drug discovery process from first principles with AI (years to months)

We sometimes think of this as doing Science at Digital Speed
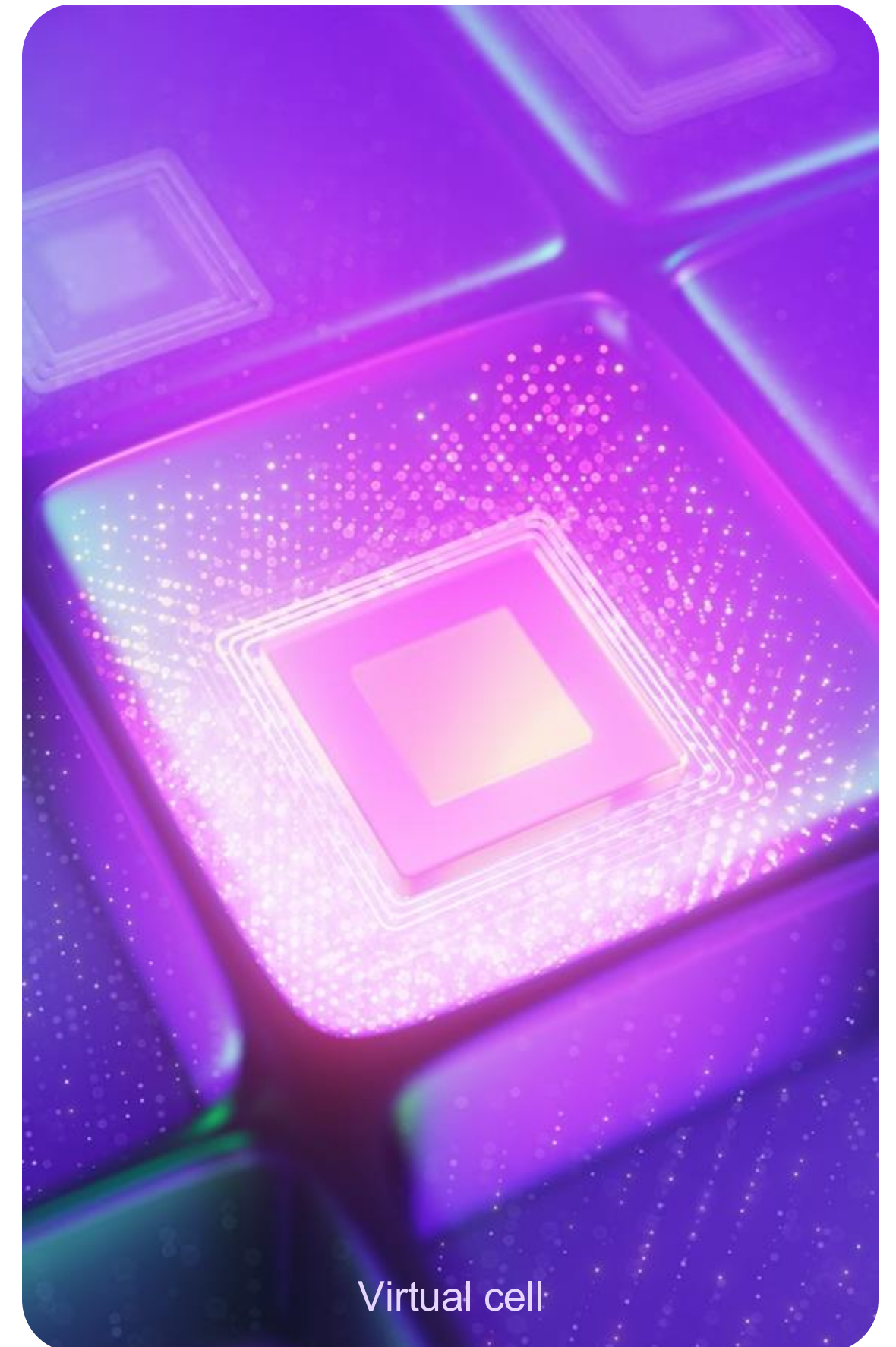
Virtual cell

What are the limits of classical systems?

Classical Turing Machines can do much more than we previously thought

By doing a massive amount of pre-compute upfront to develop a good model

Then use the model to efficiently explore a solution space in polynomial time

My Proposed Conjecture:

"Any pattern that can be generated or found in nature can be efficiently discovered and modelled by a classical learning algorithm"

If it turns out that classic systems can model certain types of quantum systems, it could potentially have big implications for complexity theory including P=NP, and maybe even fundamental physics!
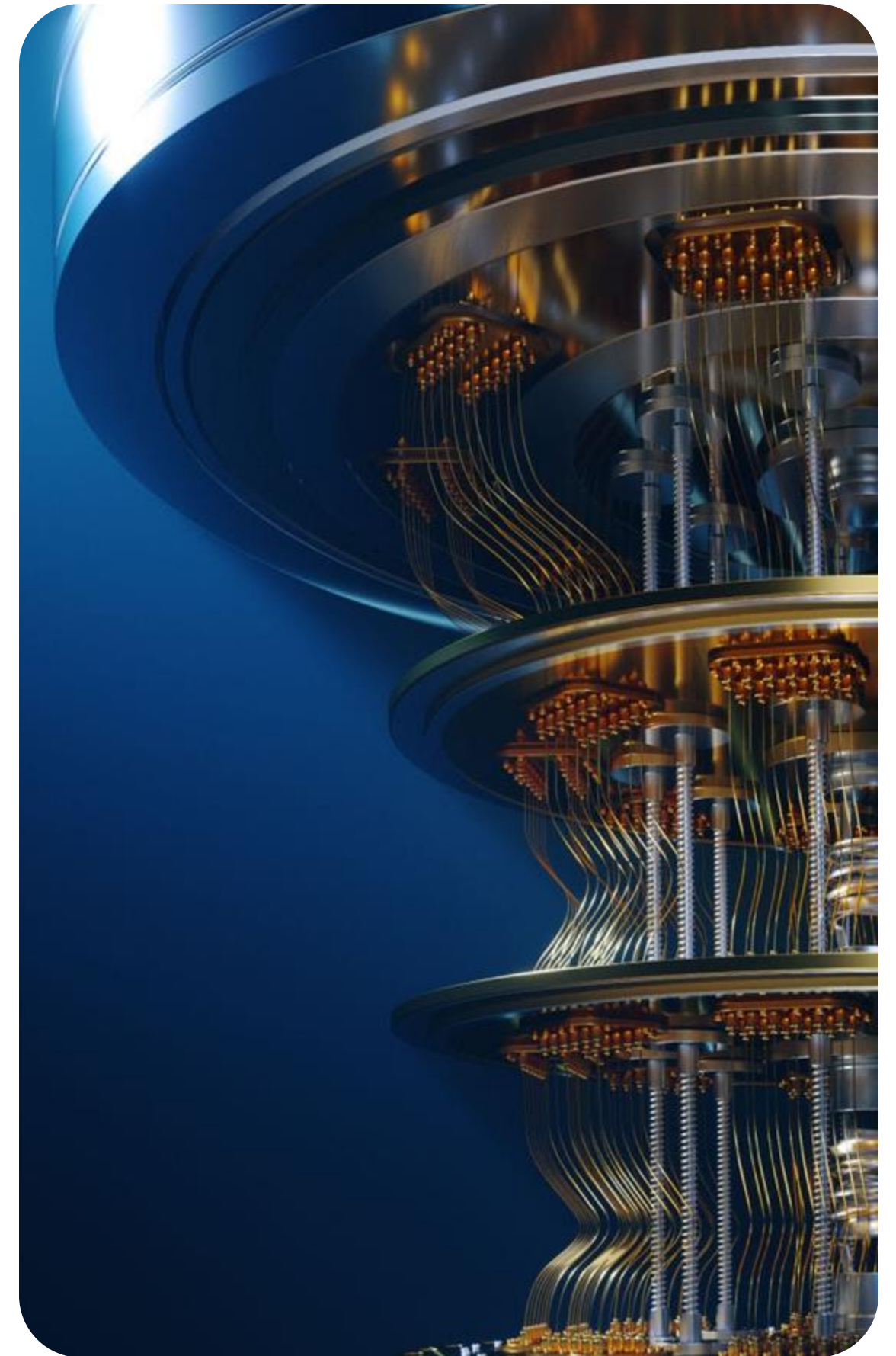
AI for Science, Medicine & Climate
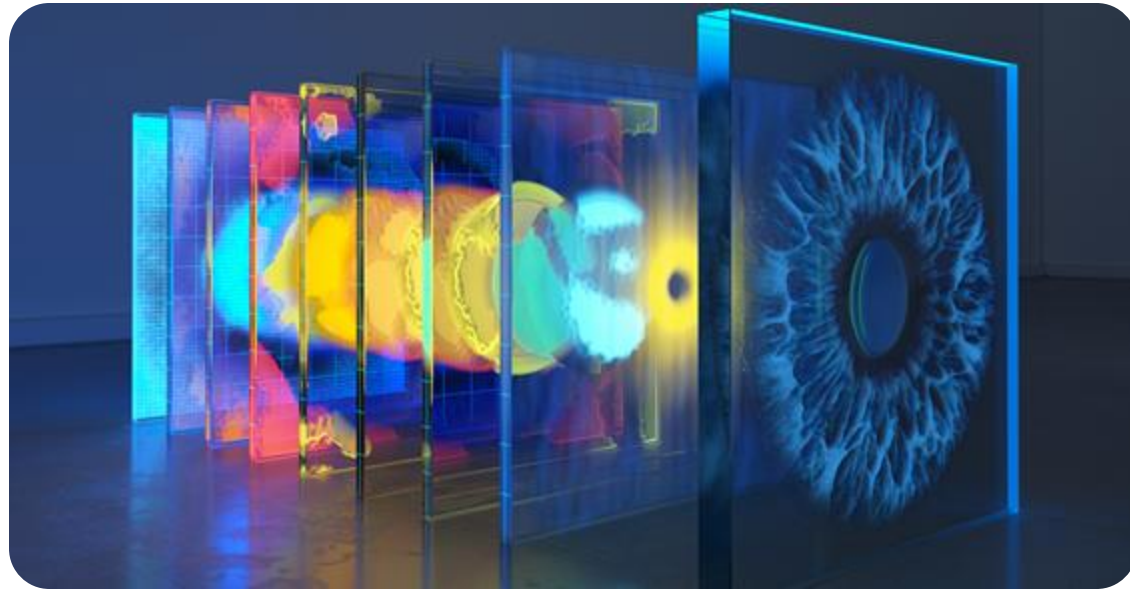
Identifying eye disease from retinal scans
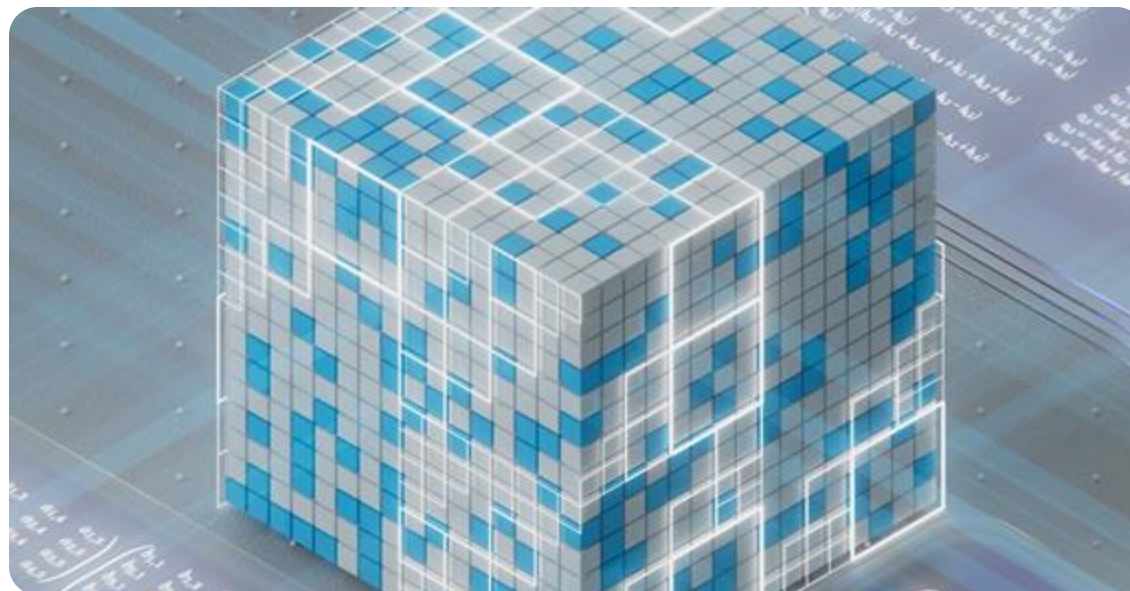
(De Fauw et al., Nature Medicine 2018)

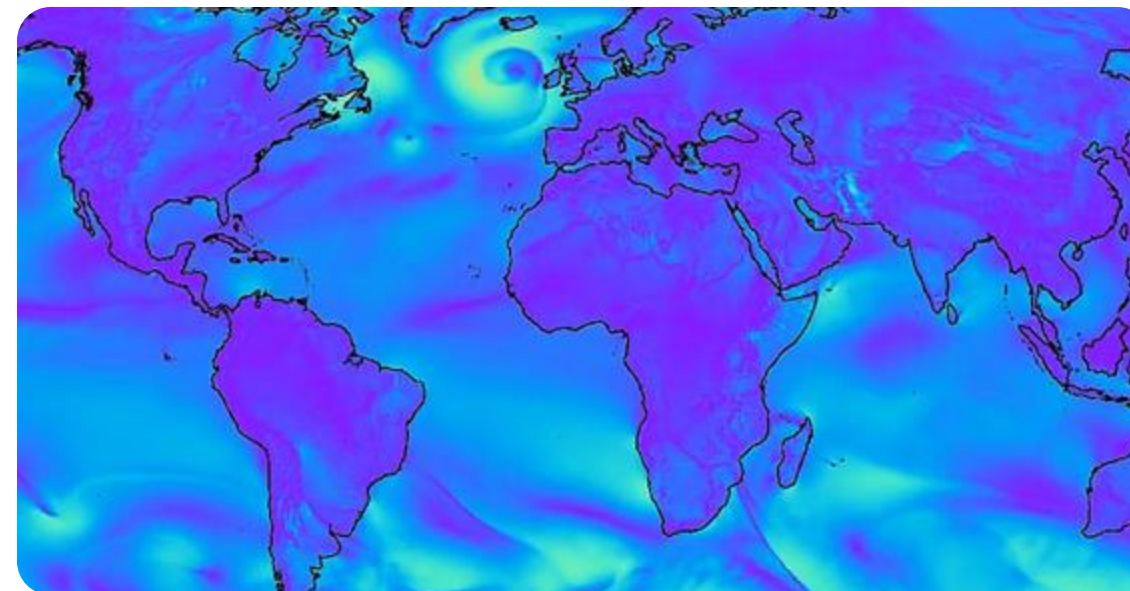**Genetic missense mutations**

(Cheng et al., Science 2023)

Fusion - plasma containment

(Degrave et al., Nature 2022)

Faster matrix multiplication

(Fawzi et al., Nature 2022)

SOTA weather forecasting

(Lam et al., Science 2023)

Discovery of new materials

(Merchant et al., Nature, 2023)

Advancing AI responsibly

AI has incredible potential to help with humanity's greatest challenges

But AI must be built responsibly and safely, and be used for the benefit of everyone

Critical to engage with a wide range of stakeholders from government, academia, and civil society

Technology as transformative as AGI requires exceptional care and foresight

AGI - the ultimate general-purpose tool
to help us understand the universe

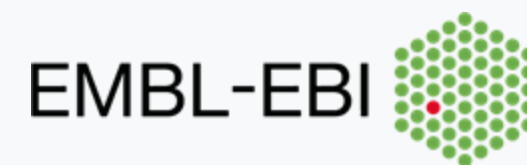# Thank you to the incredible *AlphaFold* team!

| AlphaGo and AlphaZero teams | The wider teams at DeepMind and Google | Our collaborators at EMBL-EBI | The CASP community | PDB & experimental biology community |