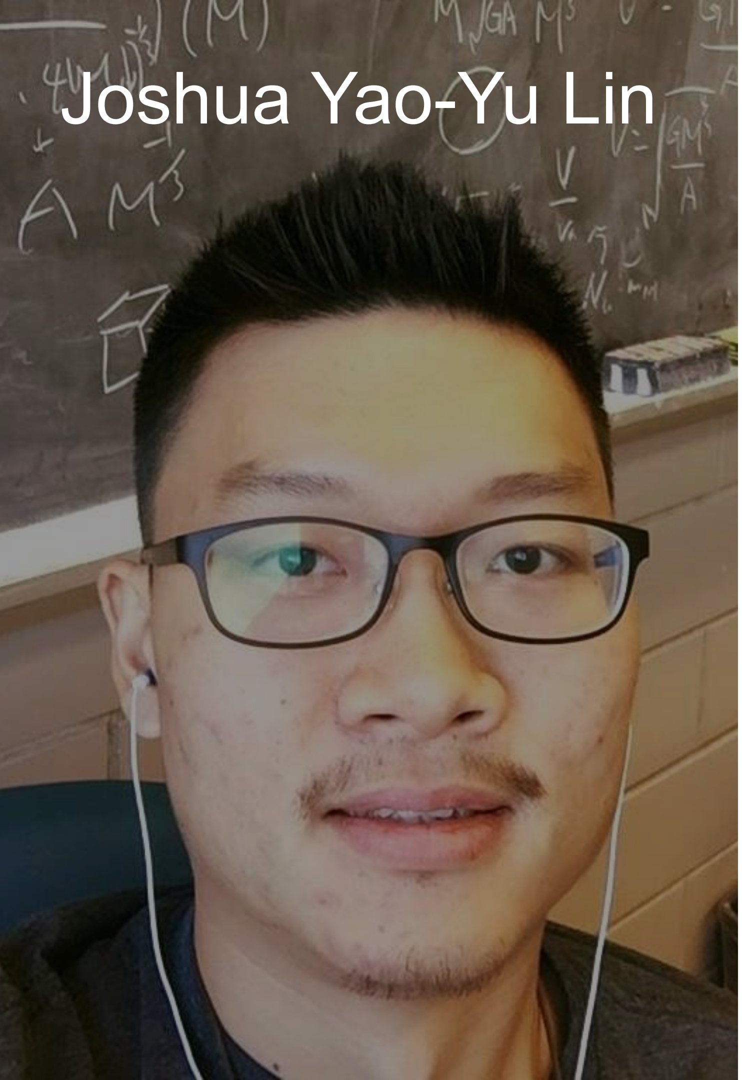
From Astronomy
to the world of Proteins with Machine Learning

Joshua Yao-Yu Lin 林曜宇 (Prescient Design/Genentech)

Astronomers Turned Data Scientists (ATDS) 2023 Meeting



Joshua Yao-Yu Lin

- Currently a Machine Learning Postdoc at Prescient Design Team in Genentech/Roche (Mentor: **Kyunghyun Cho**)

- UIUC Physics Ph.D. (2016-2022), MS at NTU, and BS at NTHU in Taiwan.

- My past research spans a wide range of Machine Learning application for astrophysics, including black hole image and dark matter/strong lensing

- ML Research Interest: **ML for Science**, **ML for drug discovery**, Self-Supervised Learning, ML interpretability

- ML intern experience: Simons Foundation/Flatiron Institute (CCA), Google Research (2021)

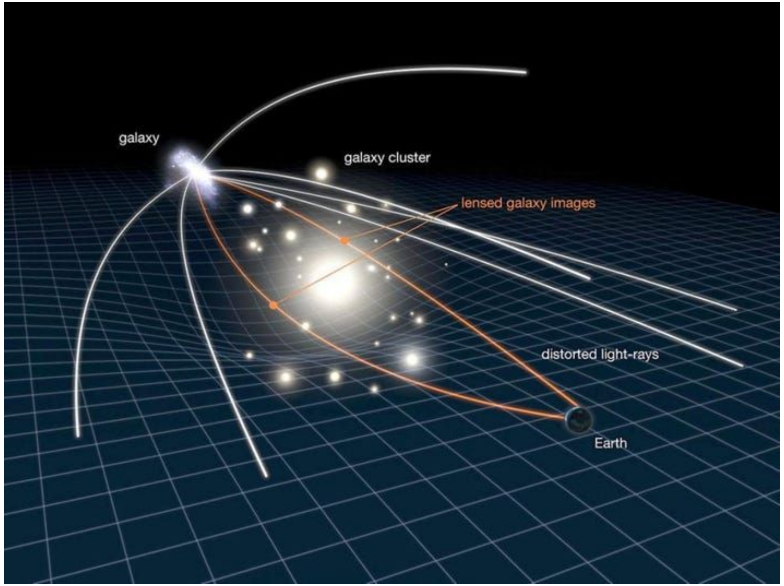
- I like: Traveling, Jazz, Bouldering/Climbing, Brewing hard cider

My previous research

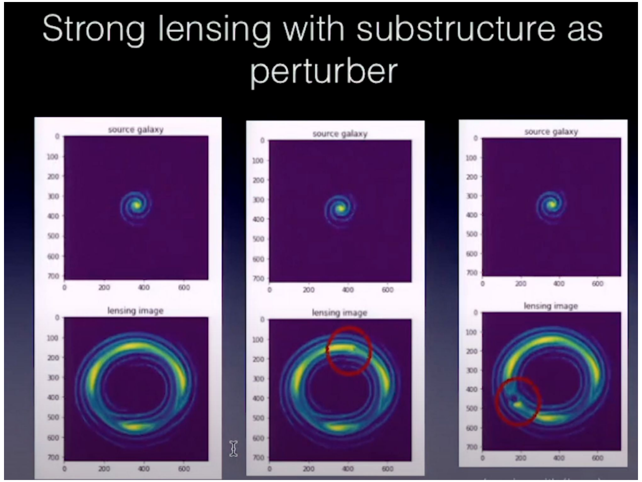
1) ML for Black hole image

2) Dark Matter and Strong gravitational lensing

3) Machine Learning application for astrophysics

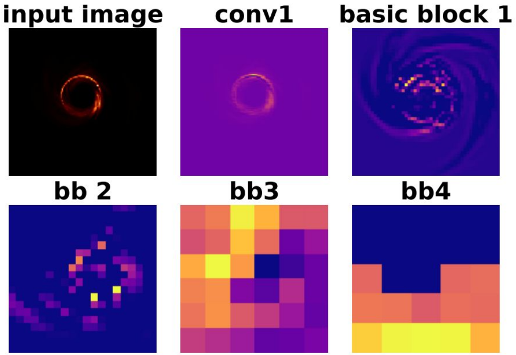


Machine Learning x astrophysics projects

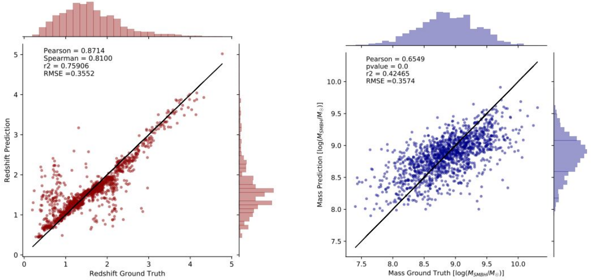


Strong lensing with substructure as perturber

Hunting for dark matter substructures with neural networks (NeurIPS workshop 2019)



Feature Extraction on Synthetic Black Hole Images (ICML workshop 2020)



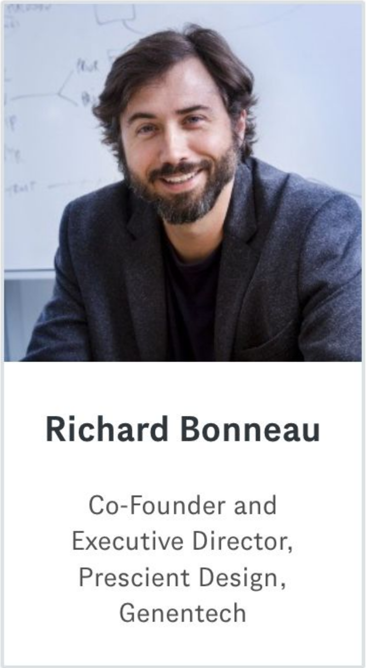
AGNet: Weighing Black Holes with Machine Learning (NeurIPS workshop 2020)

Prescient Design@Genentech/Roche

- Founded in Jan 2021, focusing on machine learning for **Protein Design**
- Acquired by **Genentech/Roche** ~ August 2021
- Around 50 people in the team (ML Scientist/ Engineer, Bio/Chem)

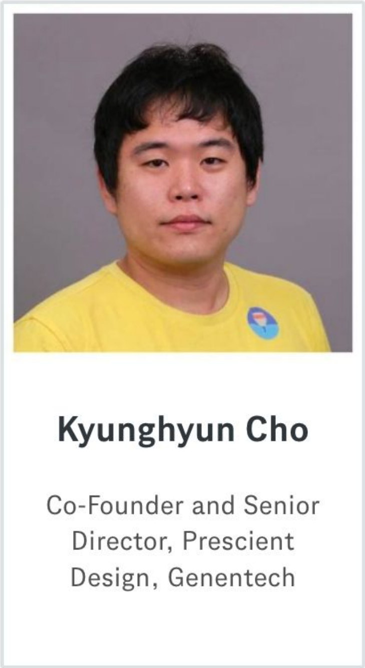
Vladimir Gligorijevic

Co-Founder and Senior Director, Prescient Design, Genentech



Richard Bonneau

Co-Founder and Executive Director, Prescient Design, Genentech



Kyunghyun Cho

Co-Founder and Senior Director, Prescient Design, Genentech

Prev. @Flatiron Institute/NYU

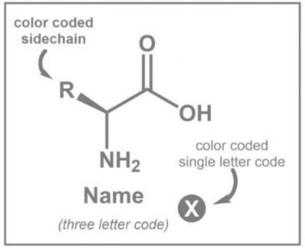
NYU CS/Data Science

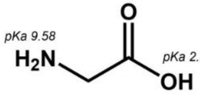
Prescient Design

A Genentech Accelerator

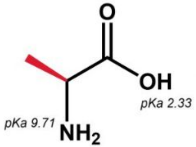
THE 20 COMMON AMINO ACIDS

● ALIPHATIC ● AROMATIC ● AMIDIC ● HYDROXYLIC
● ⊖ CHARGED ● ⊕ CHARGED ● SULFUR CONTAINING

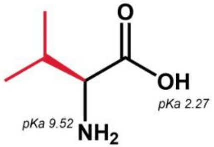




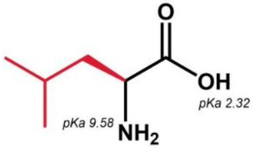
Glycine G
(Gly)



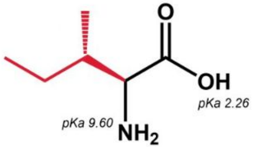
Alanine A
(Ala)



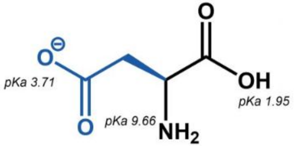
Valine V
(Val)



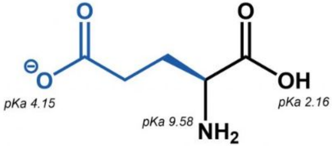
Leucine L
(Leu)



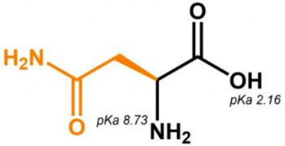
Isoleucine I
(Ile)



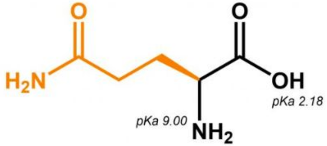
Aspartic acid D
(Asp)



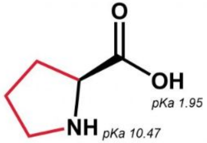
Glutamic acid E
(Glu)



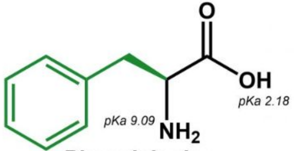
Asparagine N
(Asn)



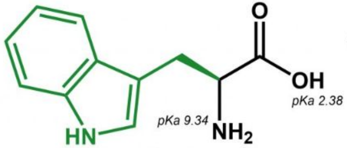
Glutamine Q
(Gln)



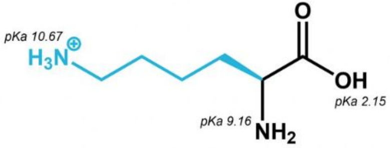
Proline P
(Pro)



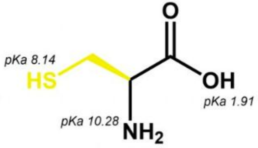
Phenylalanine F
(Phe)



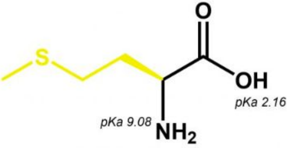
Tryptophan W
(Trp)



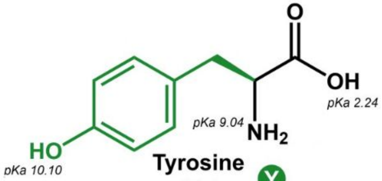
Lysine K
(Lys)



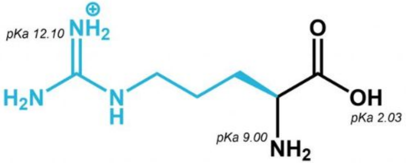
Cysteine C
(Cys)



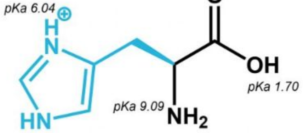
Methionine M
(Met)



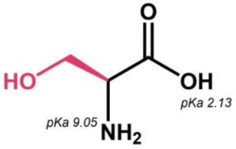
Tyrosine Y
(Tyr)



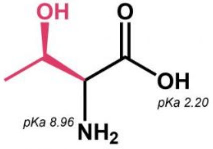
Arginine R
(Arg)



Histidine H
(His)

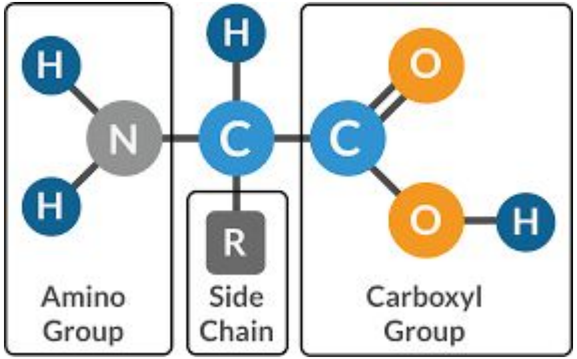


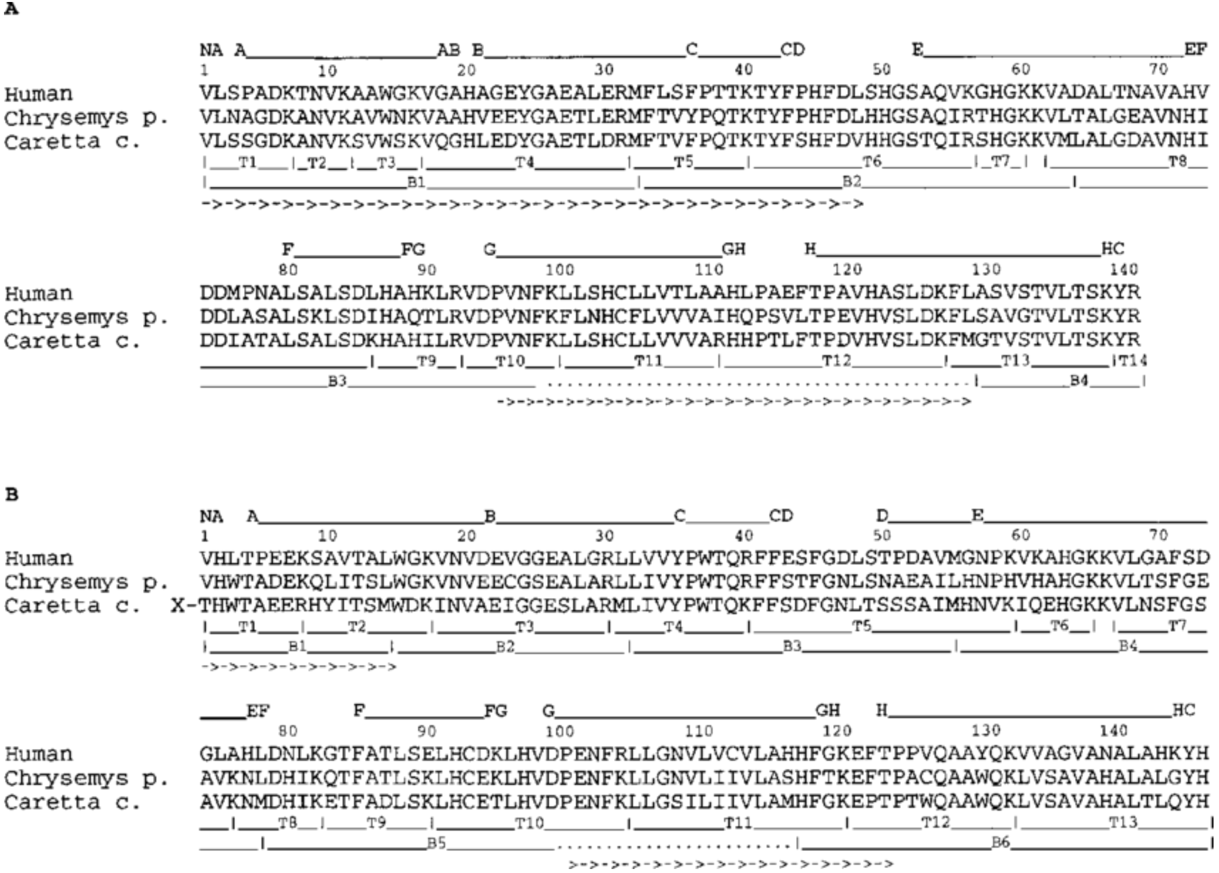
Serine S
(Ser)



Threonine T
(Thr)

Amino Acid Chain



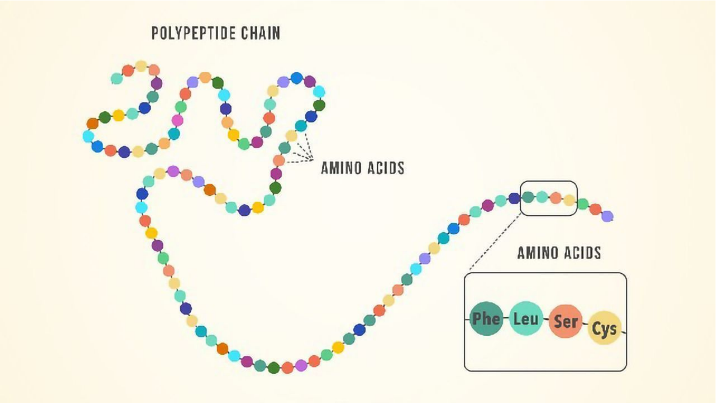


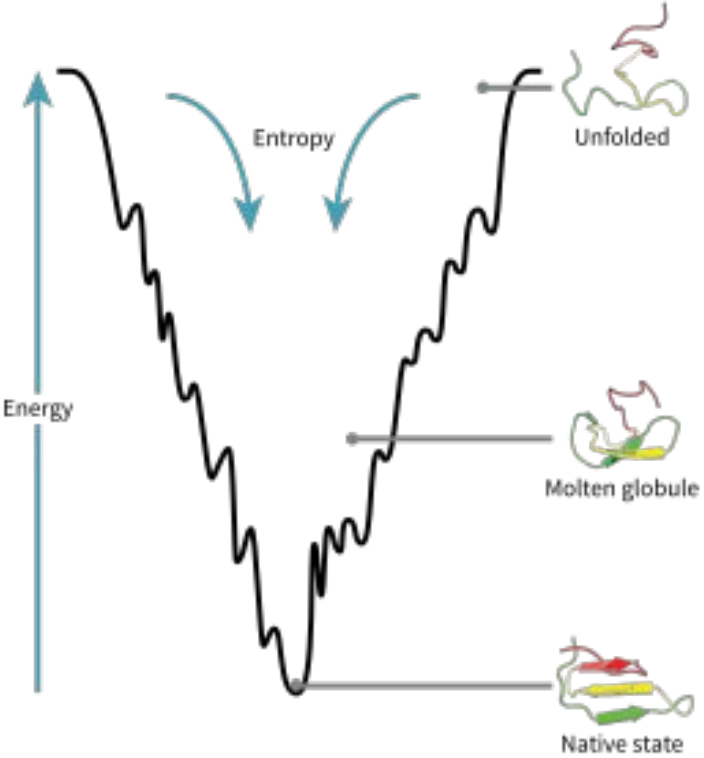
```
A

                NA A_______________AB B________________C_______CD          E____________________EF
                1        10          20          30          40          50          60          70
Human           VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHV
Chrysemys p.    VLNAGDKANVKAVWNKVAAHVEEYGAETLERMFTVYPQTKTYFPHFDLHHGSAQIRTHGKKVLTALGEAVNHI
Caretta c.      VLSSGDKANVKSVWSKVQGHLEDYGAETLDRMFTVFPQTKTYFSHFDVHHGSTQIRSHGKKVMLALGDAVNHI
                |__T1__|_T2_|_T3_|______T4______|___T5___|________T6________|_T7_| |________T8__
                |_____________B1_______________|_______________B2_____________|____________
                ->->->->->->->->->->->->->->->->->->->->->->->->->->->->->

                     F_________FG      G_________________GH      H____________________HC
                     80         90         100         110         120         130         140
Human           DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
Chrysemys p.    DDLASALSKLSDIHAQTLRVDPVNFKFLNHCFLVVVAIHQPSVLTPEVHVSLDKFLSAVGTVLTSKYR
Caretta c.      DDIATALSALSDKHAHILRVDPVNFKLLSHCLLVVVARHHPTLFTPDVHVSLDKFMGTVSTVLTSKYR
                ___________|___T9__|__T10___|_____T11_____|_______T12_______|_____T13______|T14
                ________B3____________  .........................................|________B4_____|
                                ->->->->->->->->->->->->->->->->->->->->->->->->->

B
                NA  A__________________B________________C______CD          D________E_______________
                1        10          20          30          40          50          60          70
Human           VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSD
Chrysemys p.    VHWTADEKQLITSLWGKVNVEECGSEALARLLIVYPWTQRFFSTFGNLSNAEAILHNPHVHAHGKKVLTSFGE
Caretta c.    X-THWTAEERHYITSMWDKINVAEIGGESLARMLIVYPWTQKFFSDFGNLTSSSAIMHNVKIQEHGKKVLNSFGS
                |__T1___|___T2_____|_______T3_______|____T4_____|__________T5__________|__T6__| |____T7__
                |_______B1_______|__________B2__________|____________B3____________|______________B4_______
                ->->->->->->->->->->->

                ____EF          F___________FG     G__________________________GH    H______________________HC
                      80          90          100          110          120          130          140
Human           GLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Chrysemys p.    AVKNLDHIKQTFATLSKLHCEKLHVDPENFKLLGNVLIIVLASHFTKEFTPACQAAWQKLVSAVAHALALGYH
Caretta c.      AVKNMDHIKETFADLSKLHCETLHVDPENFKLLGSILIIVLAMHFGKEPTPTWQAAWQKLVSAVAHALTLQYH
                ___|___T8__|____T9____|________T10________|_________T11_________|_______T12_____|_________T13________|
                ______|_______________B5______________.......................|__________________B6__________________|
                                >->->->->->->->->->->->->->->->->->->->->->
```

Protein folding problem

Protein’s amino acid sequence -> three-dimensional atomic structure prediction.

The notion of a folding “problem” first emerged around 1960, with the appearance of the first atomic-resolution protein structures





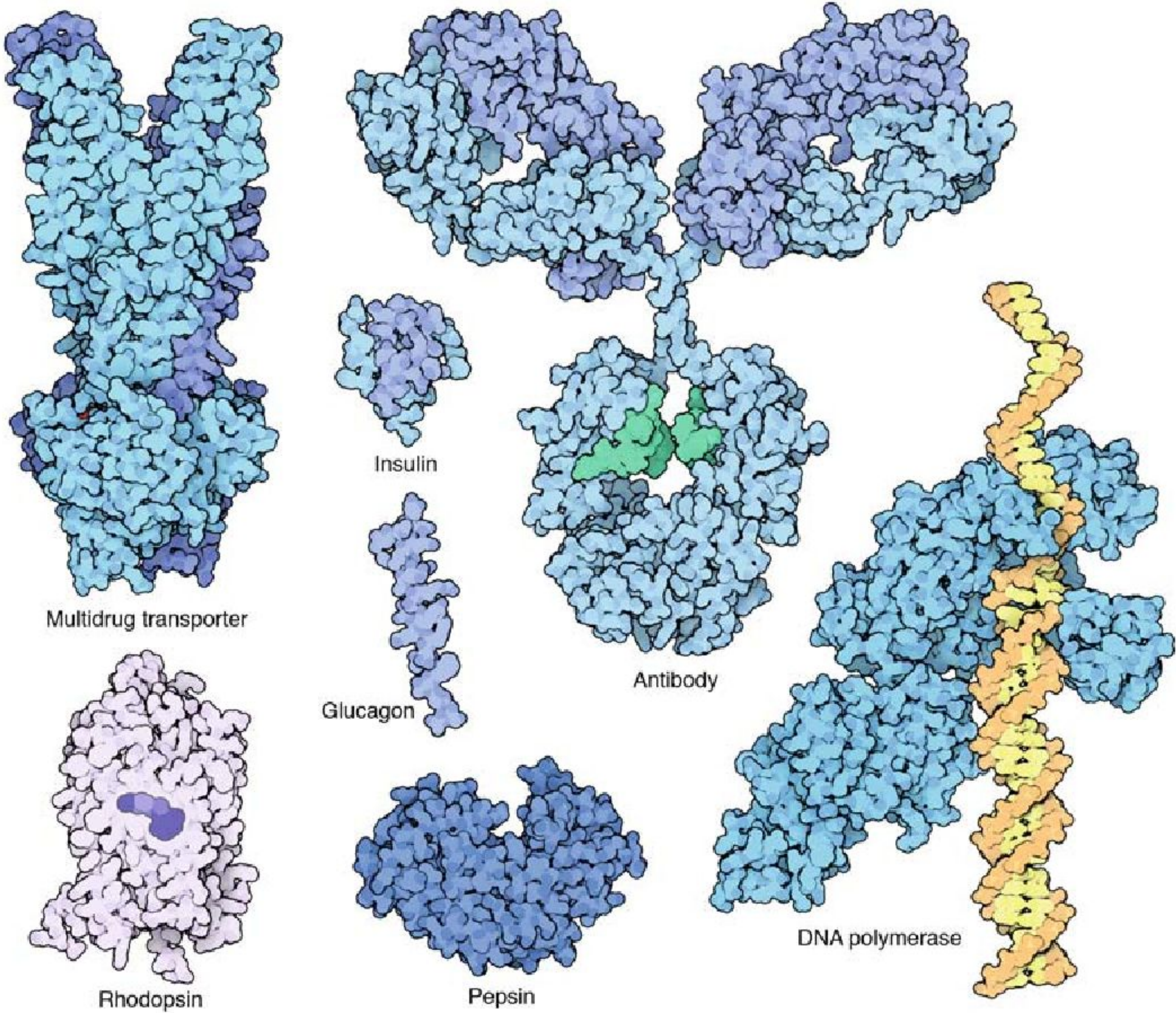
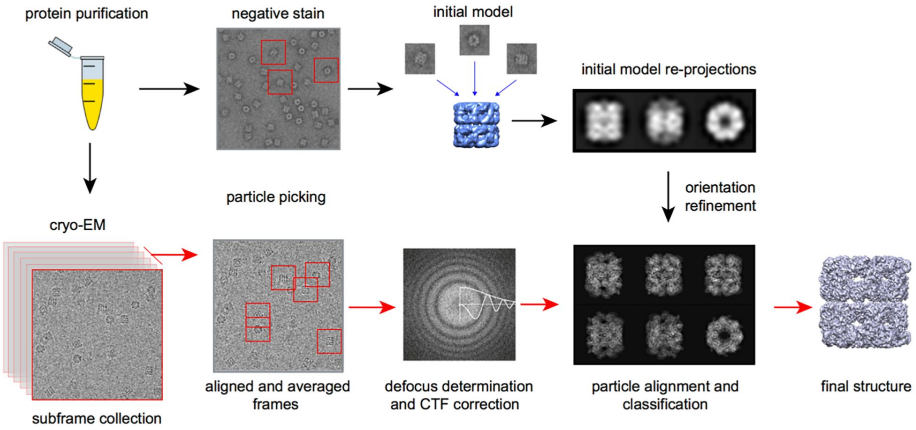
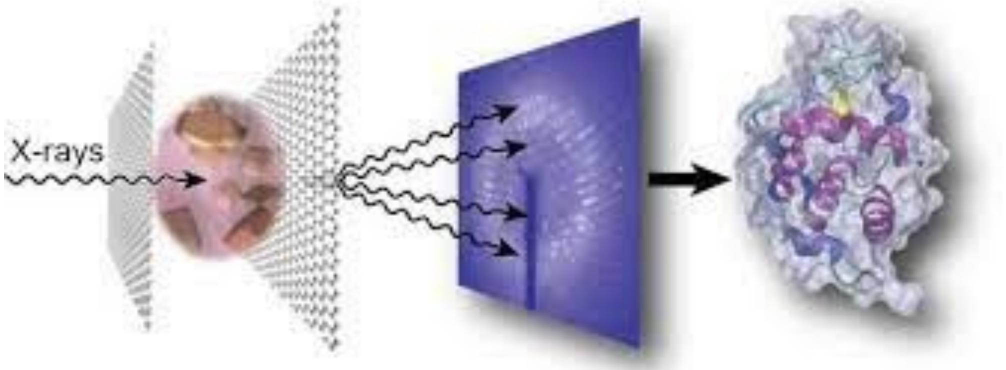


Image credit: The machinery of life
By by David S. Goodsell

How to get Protein Structures: Cryo-EM & Crystallography

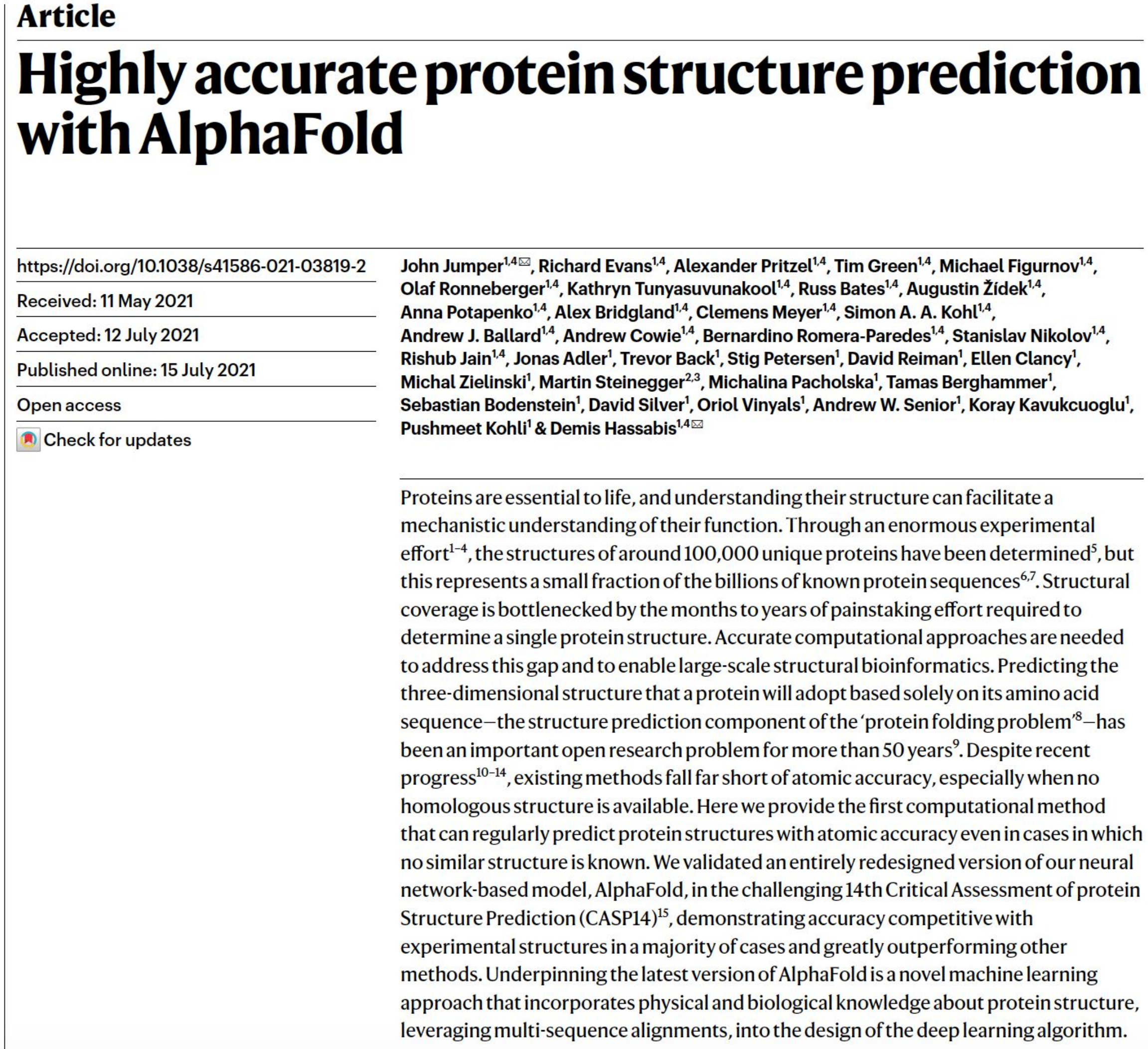

protein purification
negative stain
initial model
initial model re-projections
orientation refinement
particle picking
cryo-EM
subframe collection
aligned and averaged frames
defocus determination and CTF correction
particle alignment and classification
final structure

Time consuming and challenging for certain proteins



Alphafold II

John Jumper (DeepMind)



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

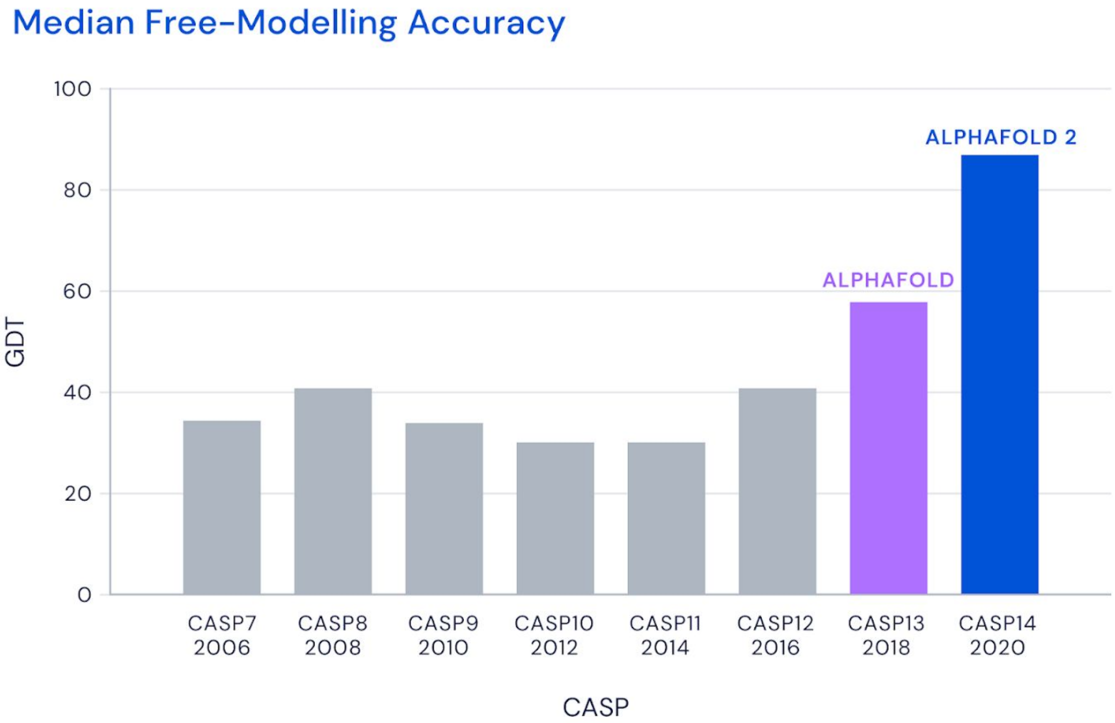
Open access

Check for updates

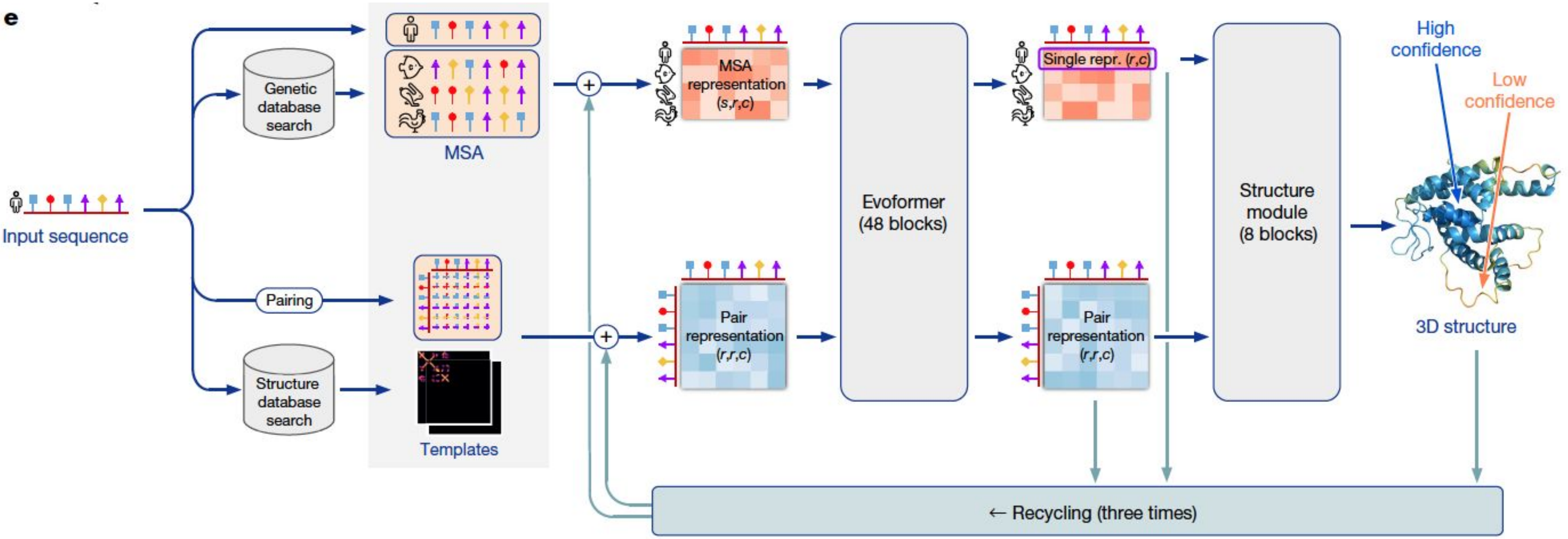
John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

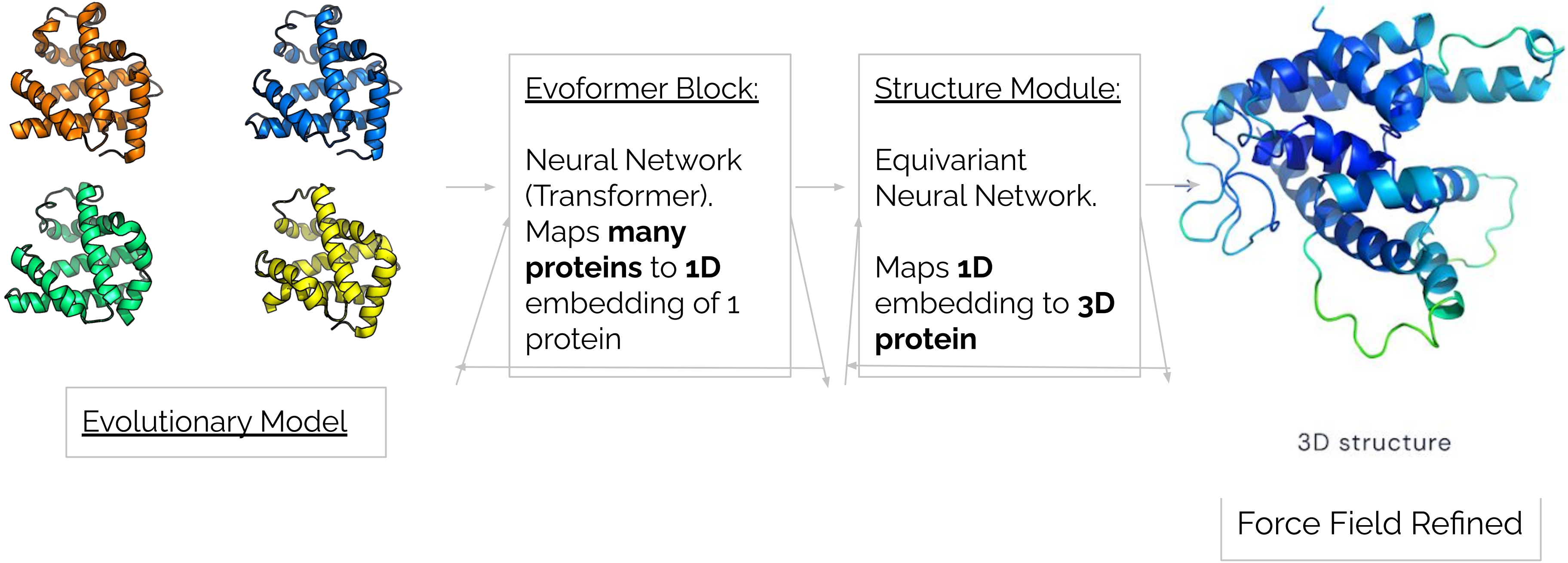
CAPS challenge



AlphaFold 2 (DeepMind)


e
Genetic database search
MSA
Input sequence
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)

Alphafold 2: We will understand this by the end



Evolutionary Model

Evoformer Block:

Neural Network (Transformer). Maps **many proteins** to **1D** embedding of 1 protein

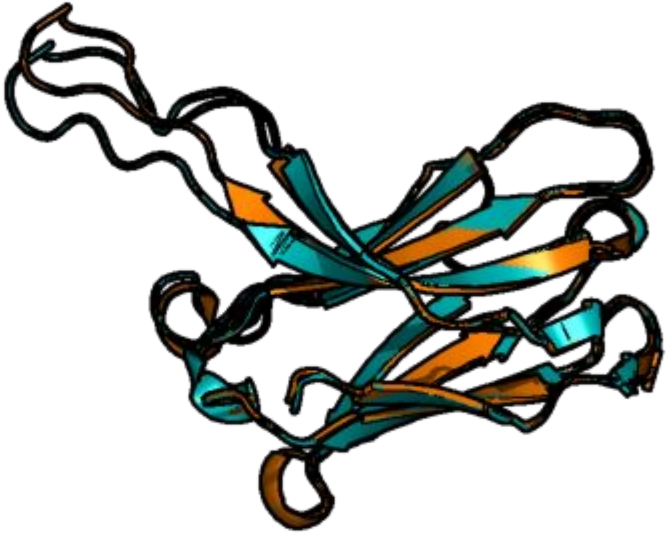
Structure Module:

Equivariant Neural Network.

Maps **1D** embedding to **3D protein**

Force Field Refined

High-level Impact



Timeline

- Dec 2018: Alphafold 1 wins CASP
    - CASP: Critical Assessment of protein Structure Prediction
- Jan 2020: Alphafold 1 Published
- Nov 2020: Alphafold 2 solves CASP
- Aug 2021: Alphafold 2 Published, 20K human proteins published

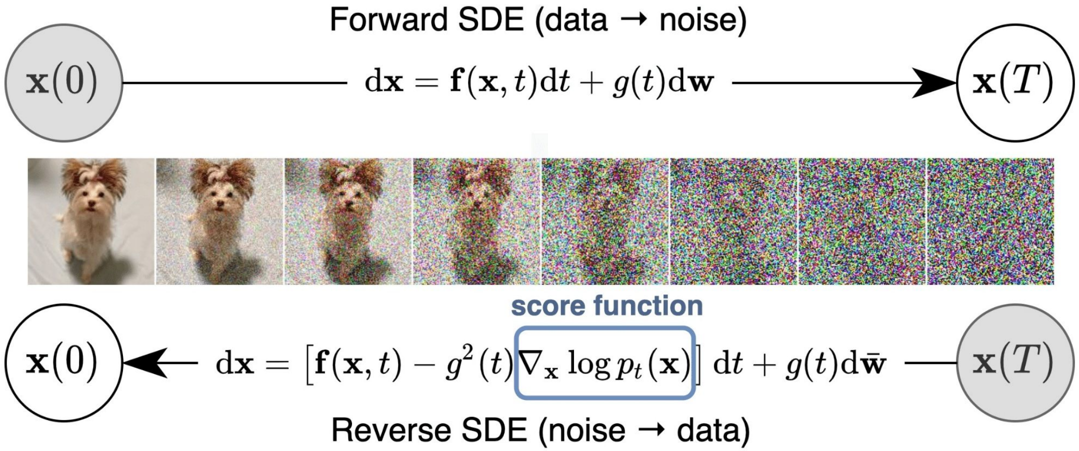
Impact on drug discovery

- Gives little functional information
- Most important proteins were already known
- **Universal end-to-end molecular drug discovery** now possible

Generative models for protein

- With sequence length ~ 100 and 20 options of amino acid it would take more than 10^30 years to generate all possible proteins.
- **Generative models** for proteins are needed - ML for **protein design**.

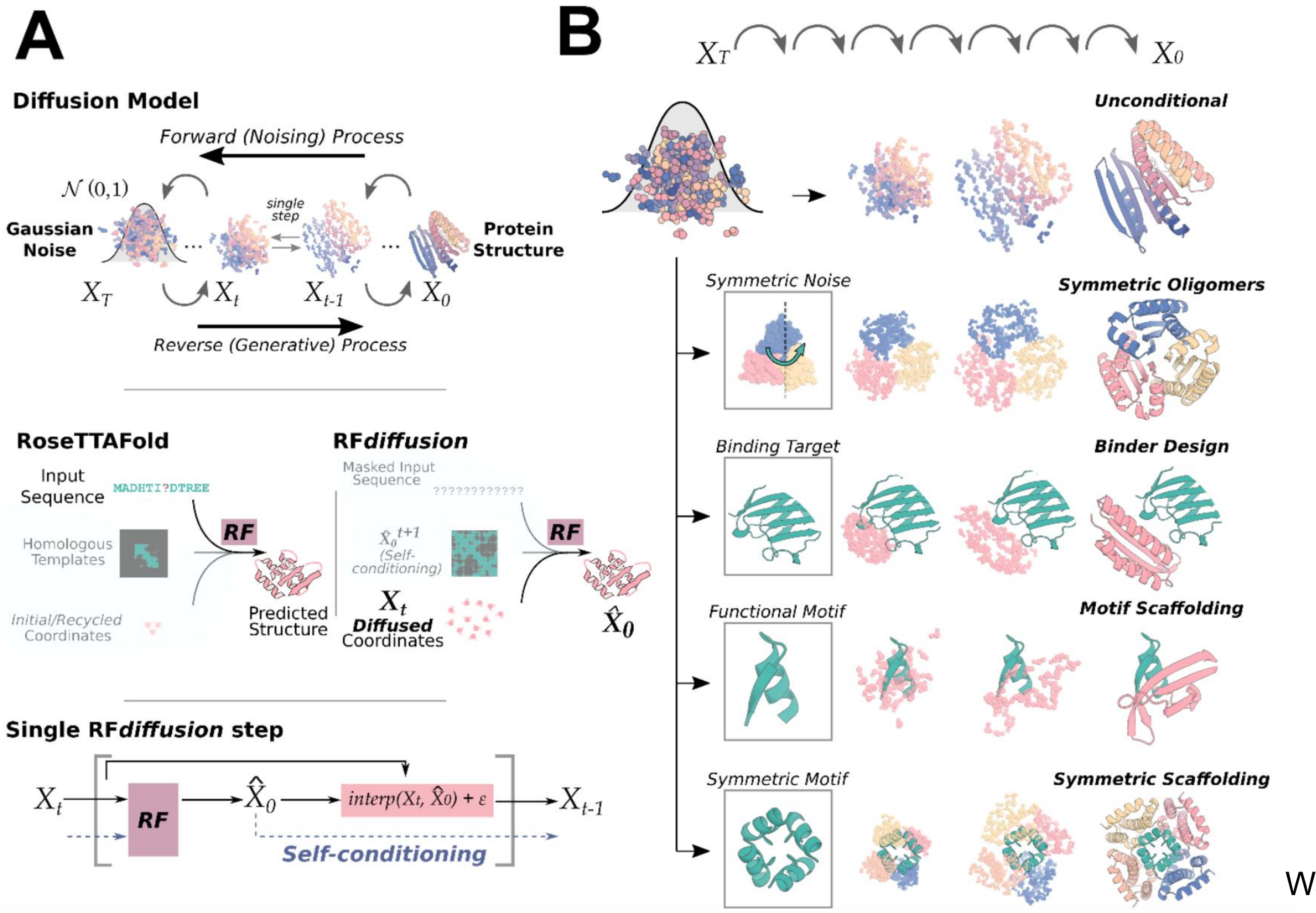
Diffusion Model (& Score matching)


Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_\mathbf{x} \log p_t(\mathbf{x})}\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)

Diffusion Model & Score matching

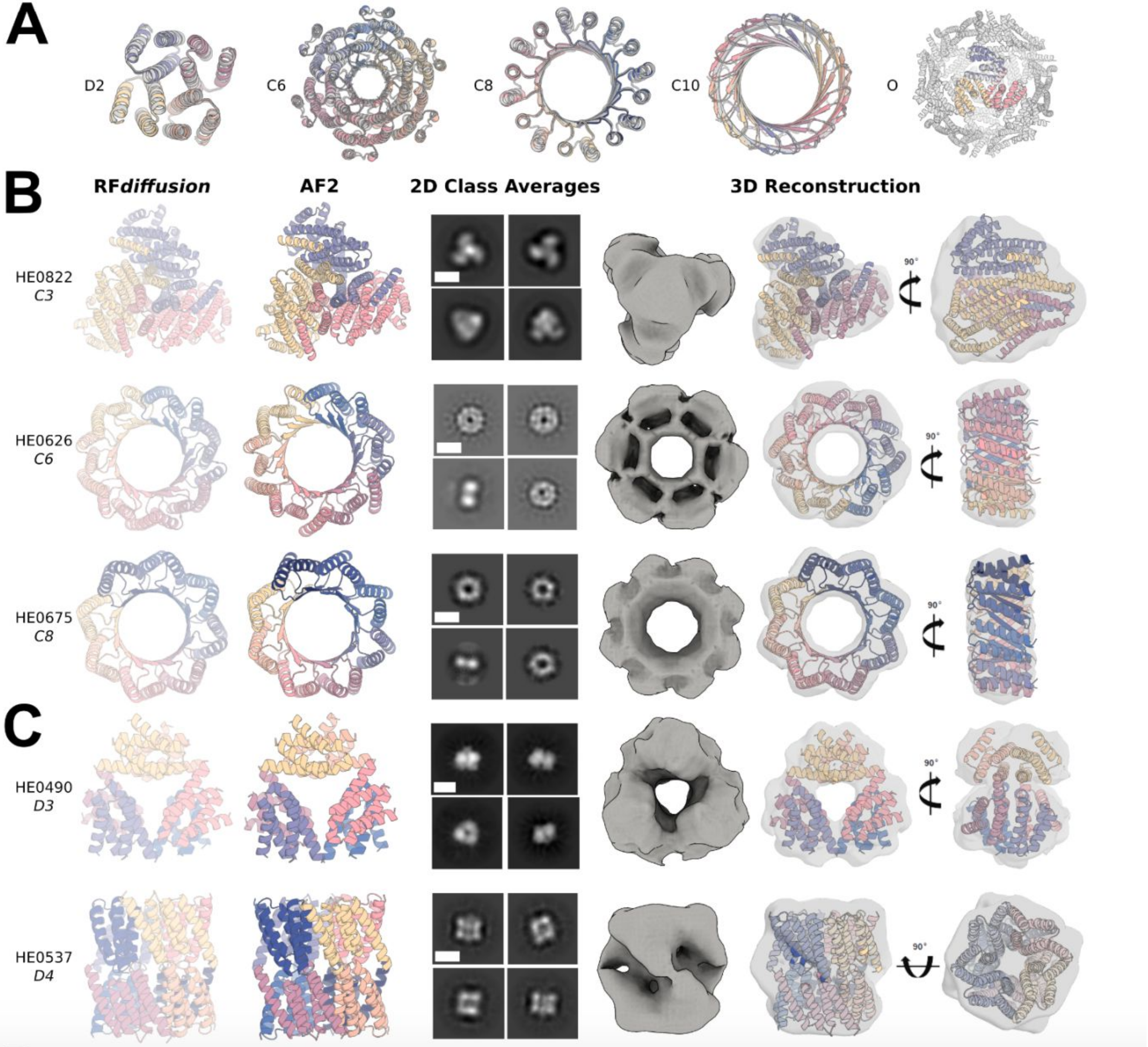
“A painting of a fox sitting in a field at sunrise in the style of Claude Monet”

DALLE 2 by OpenAI



Diffusion model for

Watson et. al. 2022



Watson et. al. 2022

My own journey

2016 - 2022 University of Illinois at Urbana-Champaign

2021 Spring Flatiron Institute/Simons Foundation (CCA)

2021 Summer Intern at Google Research

2022 - now Fall Machine Learning postdoc at Prescient Design

- Before I started my Ph.D. I barely know how to code/program and have 0 knowledge of machine learning
- I never thought I will be working on Bio-related job
- I applied for academic (astro) postdocs, jobs in tech/biotech. Decided to join the protein world and so far enjoy every moment of it.

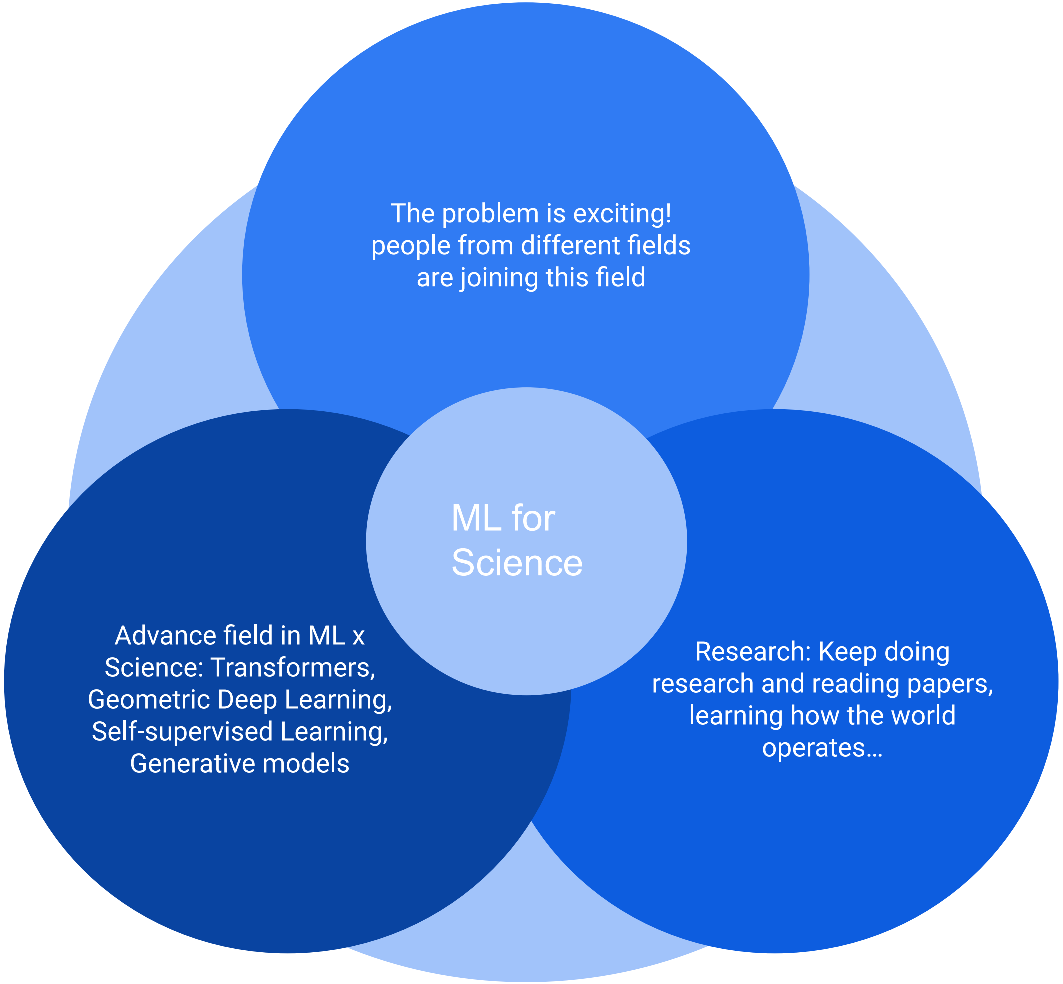
Why I switch from astrophysics to biotech?

SIMONS FOUNDATION

Google Research

Prescient Design

A Genentech Accelerator



The problem is exciting! people from different fields are joining this field

ML for Science

Advance field in ML x Science: Transformers, Geometric Deep Learning, Self-supervised Learning, Generative models

Research: Keep doing research and reading papers, learning how the world operates…

Thanks for having me!

- Follow me on twitter/LinkedIN: joshualin24, Joshua Yao-Yu Lin
- Email: Joshualin24@gmail.com
- If you are interested in chat/ or work on an ML + astro side project please feel free to reach out!
- We're hiring - please check the Genentech webpage!