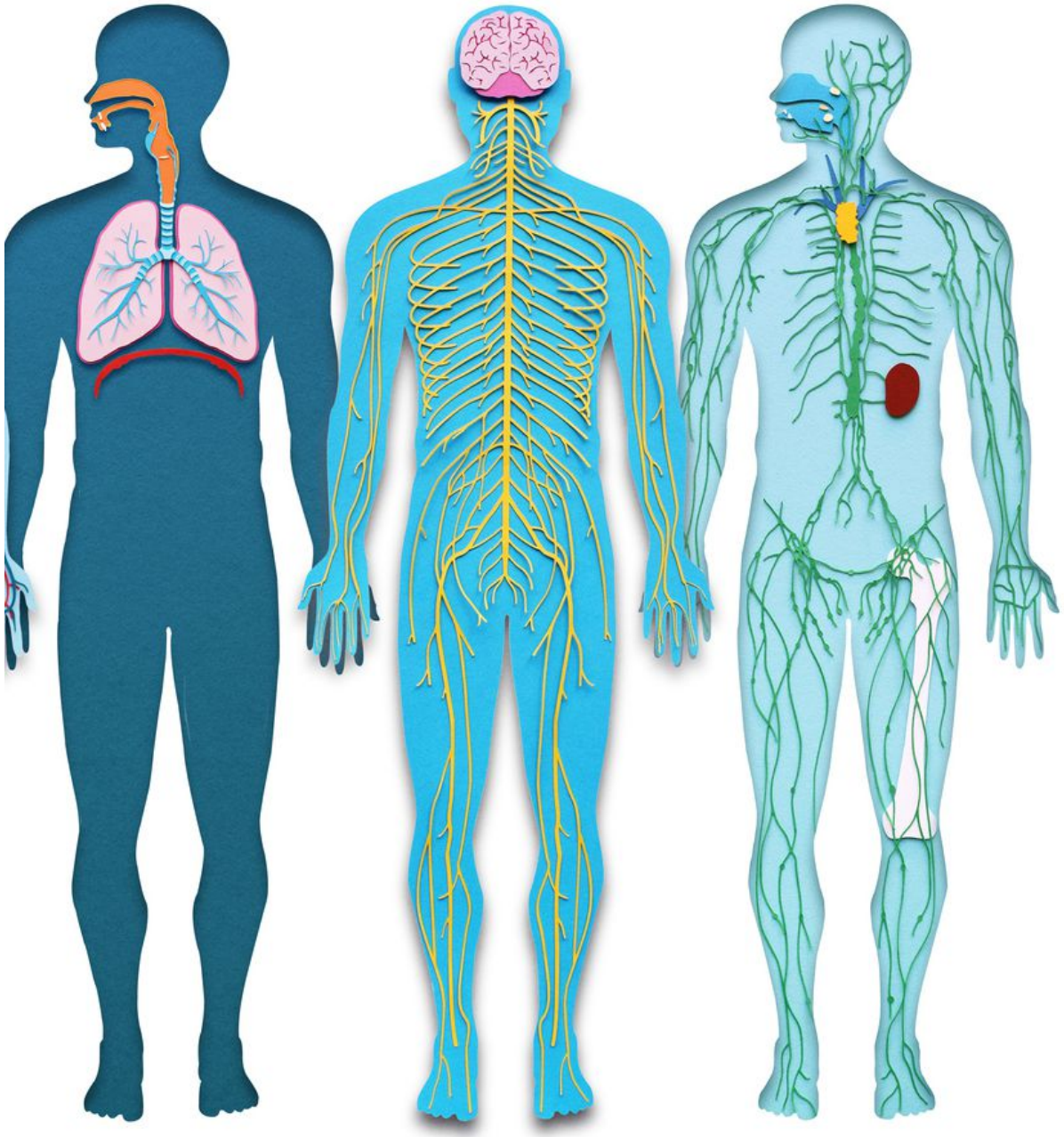
Deep Learning in Structural Biology and Protein Design: Where, How, and Why

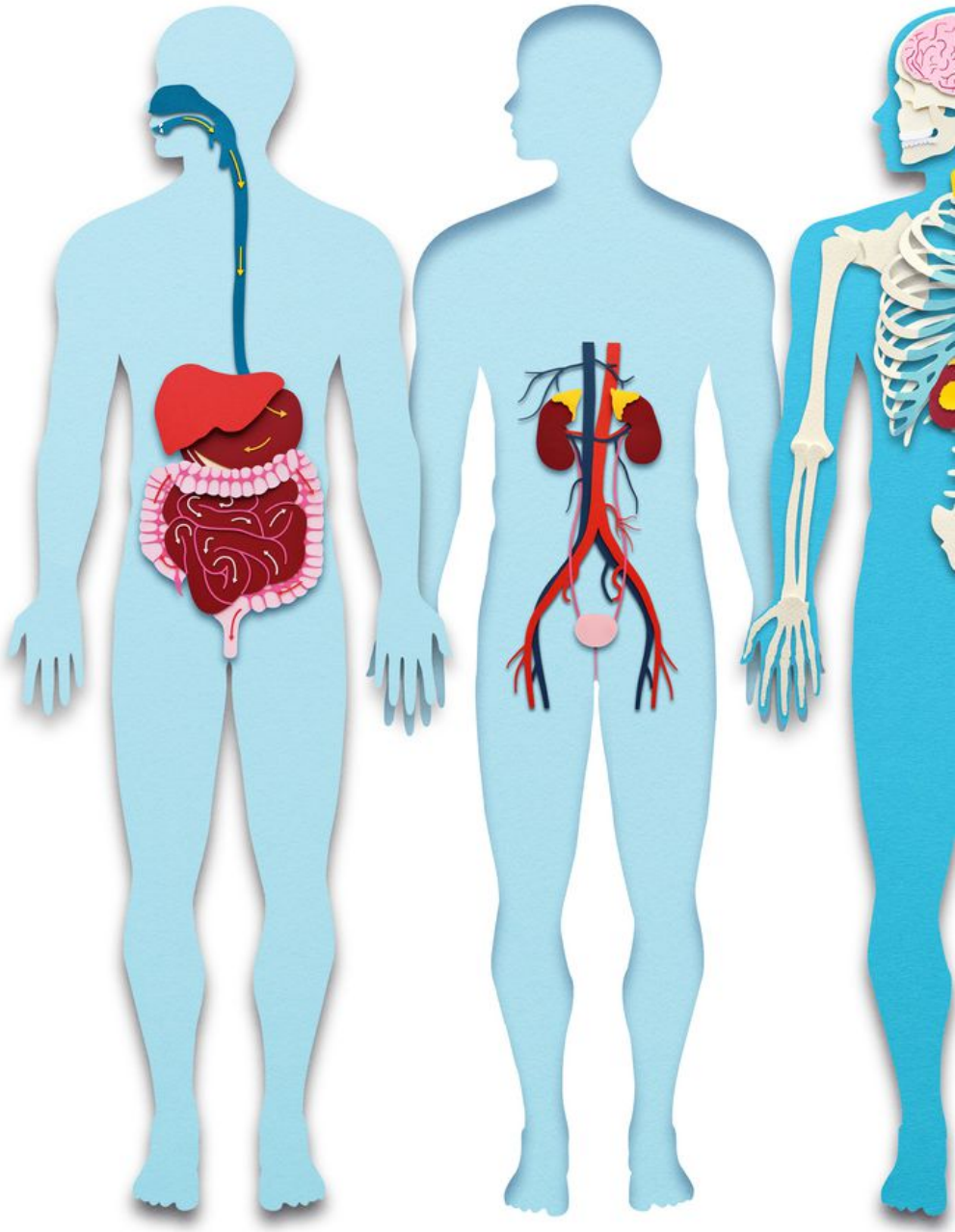
Deep Learning Theory Workshop and Summer School

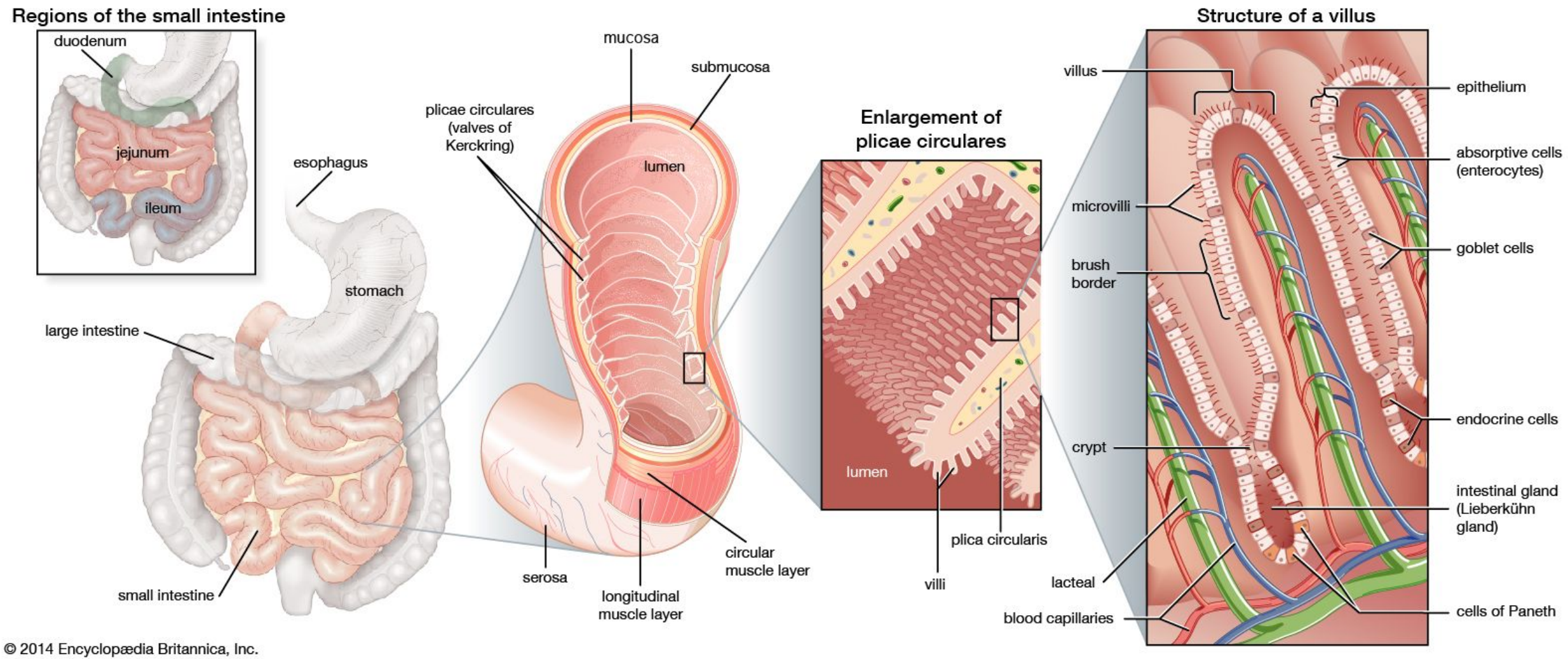
Aug 5, 2022 @ Simons Institute

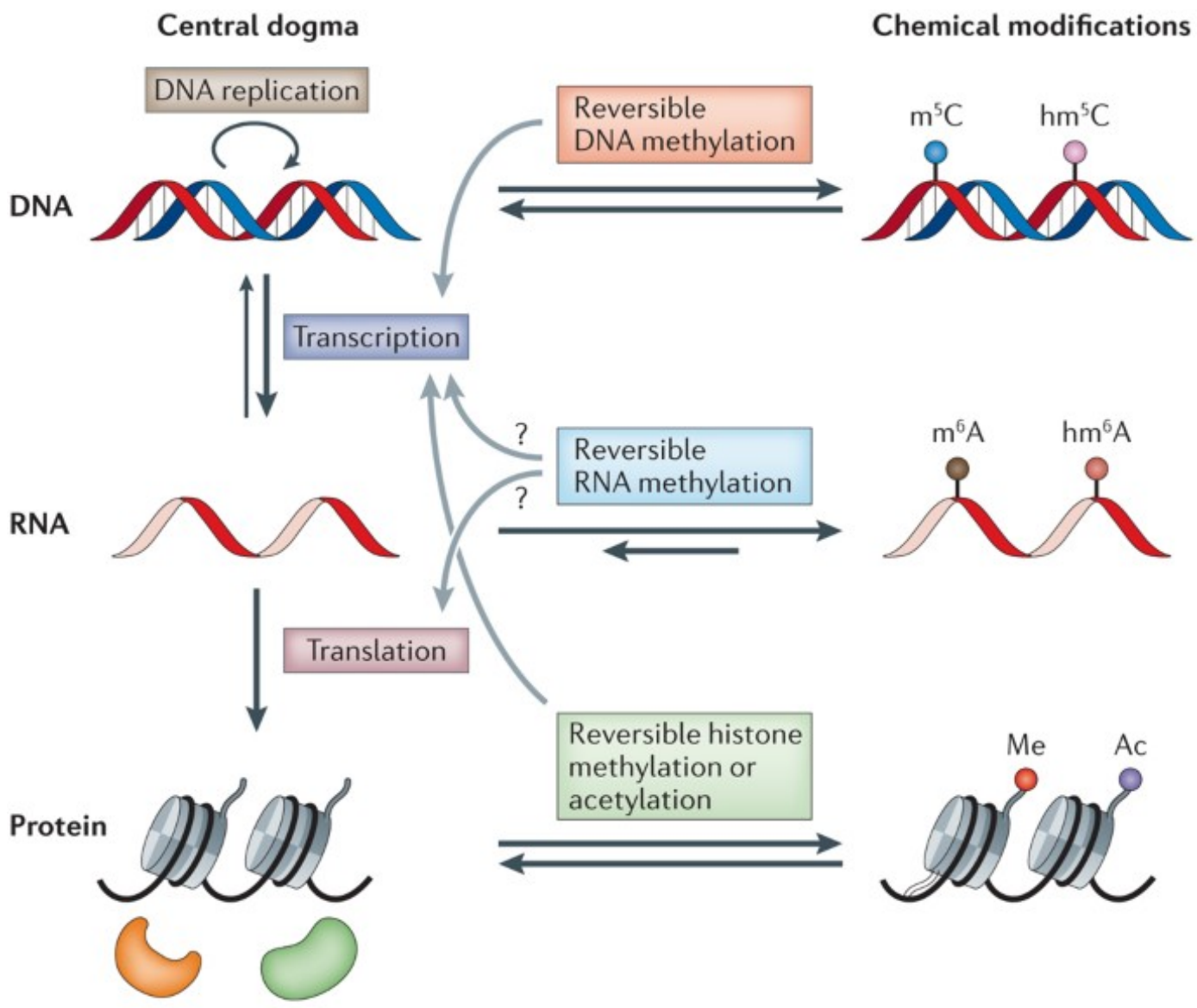
By Chloe Hsu (UC Berkeley)

Intro: A view of biology from living beings to molecules





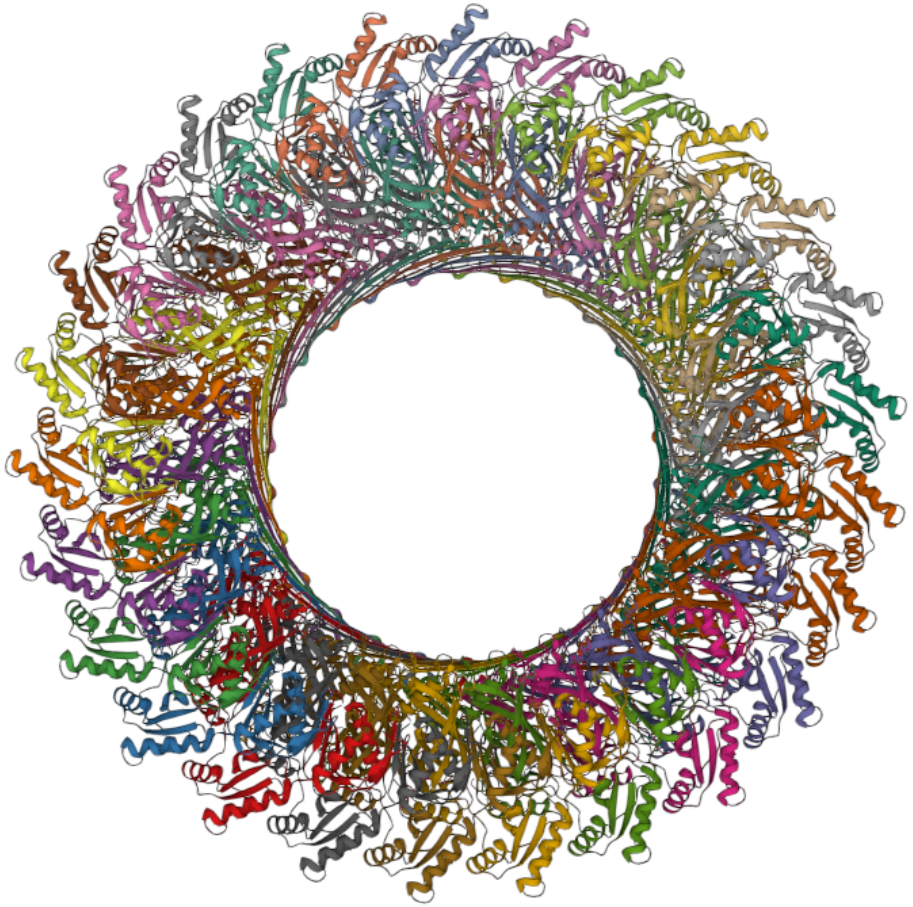




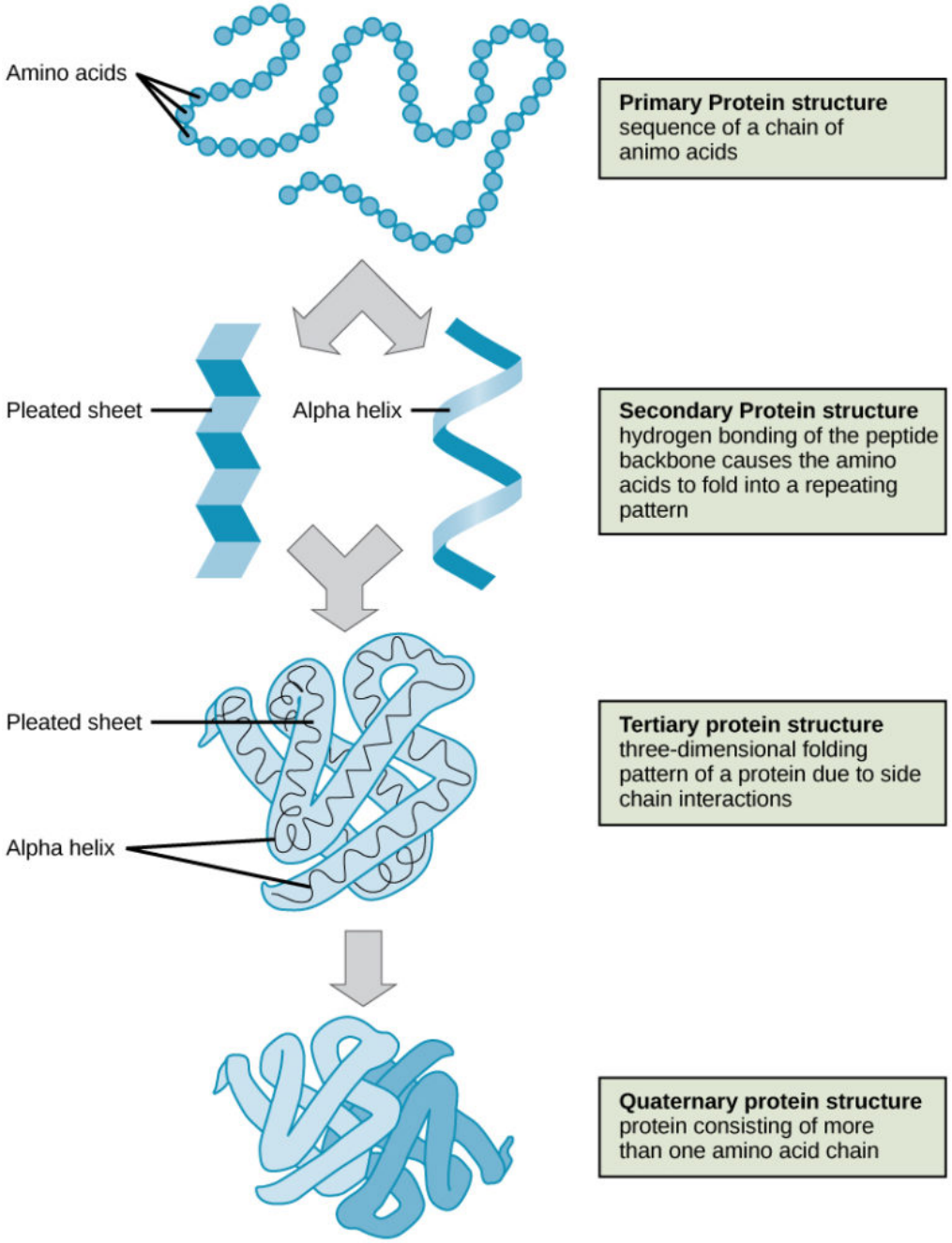
Nature Reviews | Genetics

“Where”: Advances in structural biology and protein design

Protein structures



Salmonella LP ring (It moves and swims!)



MOLECULAR MACHINERY: A Tour of the Protein Data Bank

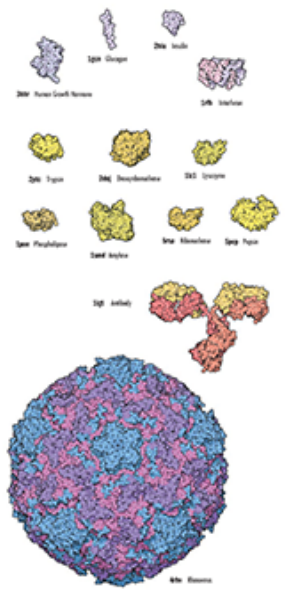
Living cells are filled with complex molecular machinery, a million times smaller than familiar machines like computers or automobiles. Cells use these tiny molecular machines to perform all of the jobs needed for life. Some are molecular scissors that cut food into cell-sized pieces. Some build new molecules when cells grow or when damaged tissues are repaired. Some are molecular bones and muscles that support cells and help them move and crawl. Some fight off attackers, defending against infection.

Researchers around the world are studying these molecules and determining their precise atomic structures. These structures are available on the Internet through the Protein Data Bank (http://www.pdb.org), the central storehouse of biomolecular structures. A few of the thousands of structures held in the Protein Data Bank are shown here. In these pictures, the molecules are all drawn at a magnification of 3,000,000 times, and each atom is shown as a small sphere. Many of these structures are composed of several subunits, which are indicated by different colors. An enormous range of sizes is shown here: the water molecule at the left has only three atoms and the ribosome shown below has hundreds of thousands.

Water
Glucose
Adenosine Triphosphate (ATP)

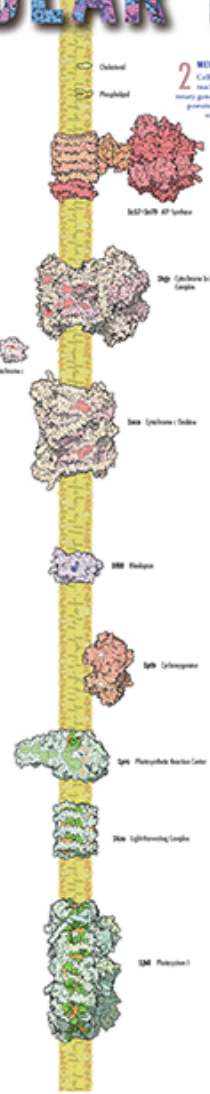
By David S. Goodsell, The Scripps Research Institute, La Jolla, California, USA

Graphic design by Eric W. Banister, San Diego Supercomputer Center



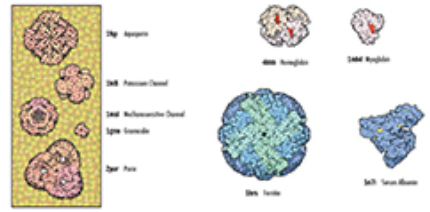
1 OUTSIDE THE CELL

Some molecular machines perform their jobs outside of cells. Many are compact, so that they can diffuse quickly to their site of action. This is true of the four hormones shown at the top: insulin and glucagon, which together regulate blood sugar levels, interferon, which carries signals in the immune system, and human growth hormone. The seven digestive enzymes (in yellow) are also small and very stable, so that they can survive the hostile environment in the digestive tract. Each of these enzymes has a small groove (oriented towards the top in each) that binds to a different target molecule and digests it. At the bottom is ribonuclease, the virus that causes the common cold, and an antibody, our major defense against viruses. Antibodies bind to viruses and prevent them from binding to cell surfaces, thus blocking infection.



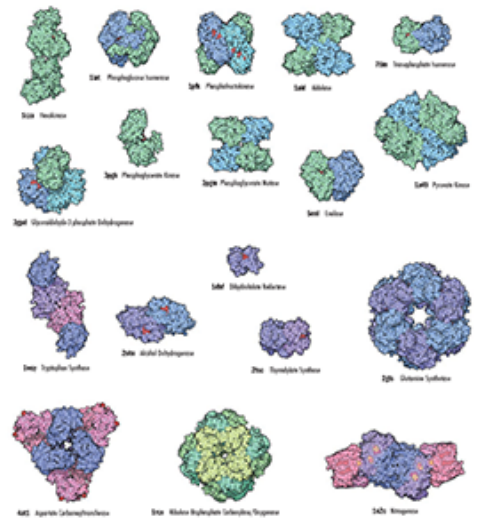
2 MEMBRANES

Cells are surrounded by a membrane made of lipids, like the phospholipid and cholesterol molecules shown at the top. Membranes keep the cellular machinery inside and unwanted material out. Many proteins are embedded in this membrane, performing a variety of essential tasks. ATP synthase is a rotary generator that produces ATP (adenosine triphosphate), the small molecule used for powering cells. The two large complexes below it charge a battery that powers ATP synthase, and the tiny protein cytochrome c shuttles electrons between them. Rhodopsin is found in membranes in the retina. The small retinal molecule inside of it changes shape when illuminated, causing the surrounding protein to send a signal to the brain. Cyclooxygenase builds one of the molecules used in signal pain—this cyclooxygenase molecule here, however, is blocked by two molecules of aspirin, shown inside in white. At the bottom are three molecules involved in photosynthesis, which capture energy from light and use it to power the synthesis of sugar in plant cells.



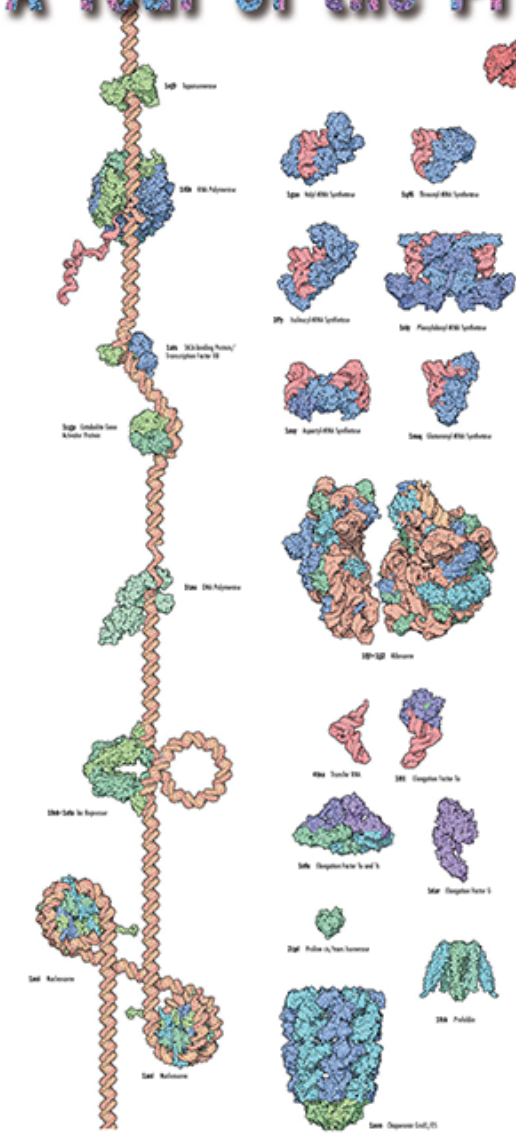
3 TRANSPORT AND STORAGE

Of course, a perfectly sealed membrane would be of little use to cells, because nutrients could not get in and wastes could not get out. The box shows a membrane looking face-on. Five proteins that form channels through the membrane are shown. To the right of the box are several soluble proteins involved in transport and storage of molecules. Hemoglobin and myoglobin carry oxygen. Ferritin forms a hollow shell that stores iron ions. Serum albumin carries many different molecules in the blood.



4 CHEMICAL FACTORIES

Cells build a bewildering variety of enzymes—proteins that perform chemical reactions. At the top are the ten enzymes that perform glycolysis, the breakdown of sugar to form ATP. Below that are several enzymes that perform different housekeeping reactions. Dihydrofolate reductase activates a key cofactor molecule and alcohol dehydrogenase breaks down alcohol. Ribulose bisphosphate carboxylase/oxygenase is the most common enzyme on the Earth, and performs a key step in the capture of carbon dioxide by plants to form sugar. The three synthases and the transferase make different building blocks for creating new molecules. Nitrogenase performs an essential role in the ecosystem by converting nitrogen gas into a form that living cells can use.

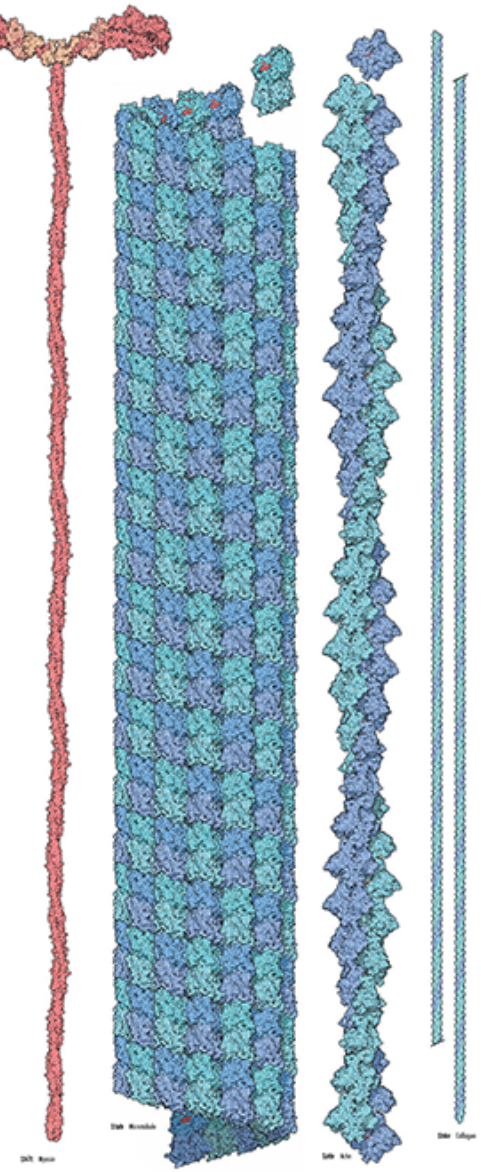


5 DNA

Genetic information is stored in the DNA double helix, seen running from top to bottom here. Many proteins are used to copy, read, and store this information. RNA polymerase copies the information into a strand of RNA that will be used to direct the construction of new proteins. It is assisted by topoisomerases, which relieve tension when the helix is wound and unwound, and guide to appropriate starting points by the two protein complexes below it. DNA polymerase replicates DNA strands—then, the polymerase is filling a gap in the double helix. Some proteins, like the lac repressor, grab DNA and bend it sharply, or even wrap it all the way around themselves, like the two nucleosomes at the bottom.

6 BUILDING NEW PROTEINS

New proteins are built by ribosomes—complex molecular factories that read the genetic code and use it to direct construction. Many molecular machines are needed to assist the process. Twenty different aminoacyl-tRNA synthases (six are shown here) load the building blocks onto tRNA, ready to be added to a growing protein chain. Several protein factors, shown below the ribosome, guide each tRNA into the proper spot. The three chaperonin proteins shown at the bottom help each new protein fold into its proper shape.



7 BEAMS AND GIRDERS

Cells are braced and supported by a complex infrastructure. This cytoskeleton is formed of sturdy filaments like actin and microtubules, composed of many subunits stacked like bricks. Myosin is a molecular motor that climbs along actin filaments, allowing the cell to move. Collagen, broken into two pieces here, is actually found outside of cells, where it forms connective tissue between cells.

February 2002

PROTEIN DATA BANK

http://www.pdb.org/ • info@rcsb.org

RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY
SAN DIEGO SUPERCOMPUTER CENTER
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

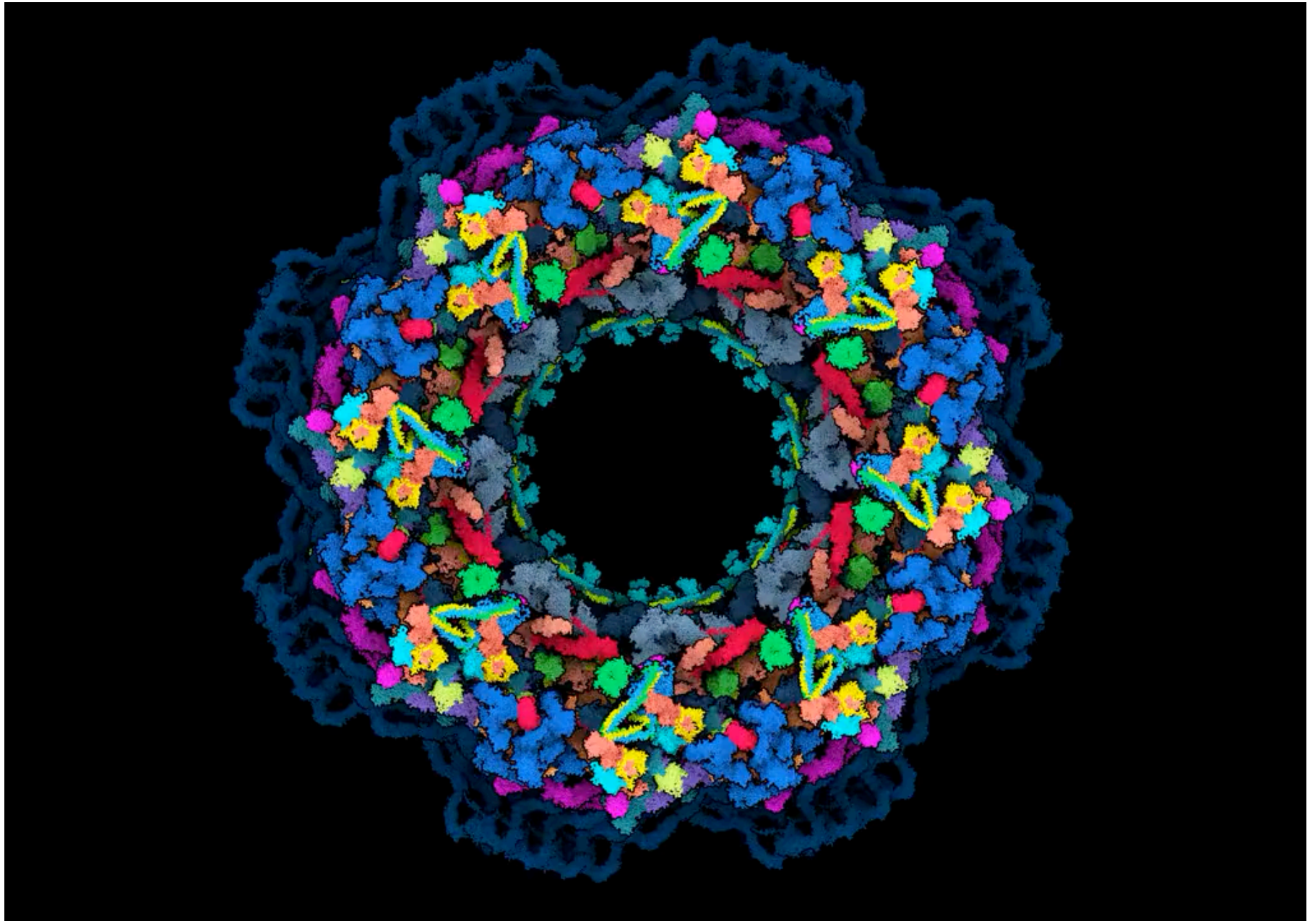
Structure prediction

Substantial improvement in structure prediction from AlphaFold2

Also notable in new capabilities for predicting protein-protein interactions (although not perfect yet)

Applications:

- **Interprets complex experimental data in structural biology at a new speed**
- Creates an unprecedented size of predicted structures for data mining & learning
- …



Mosalaganti, Shyamal, et al. "Artificial intelligence reveals nuclear pore complexity." BioRxiv (2021).

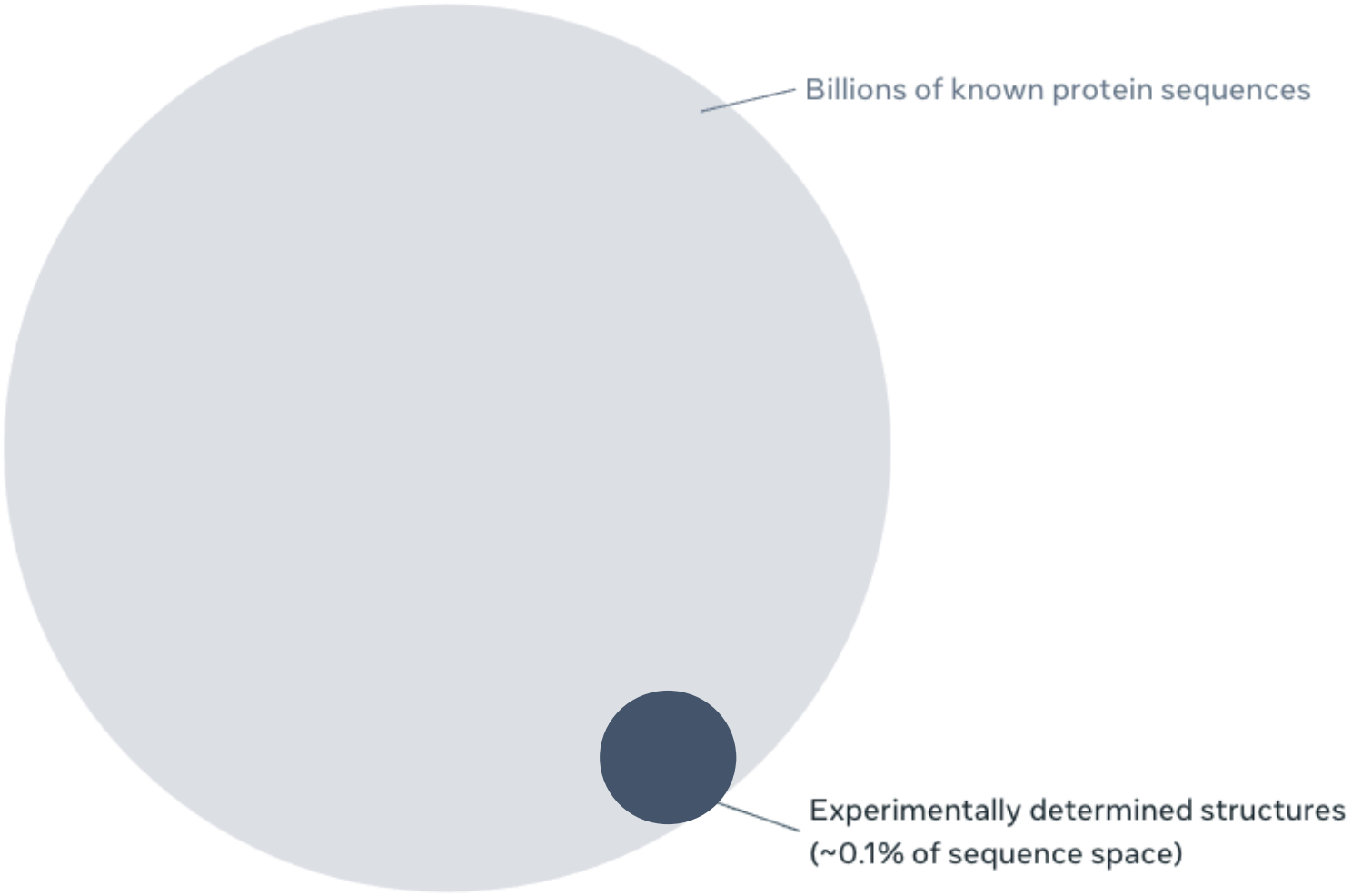
Structure prediction

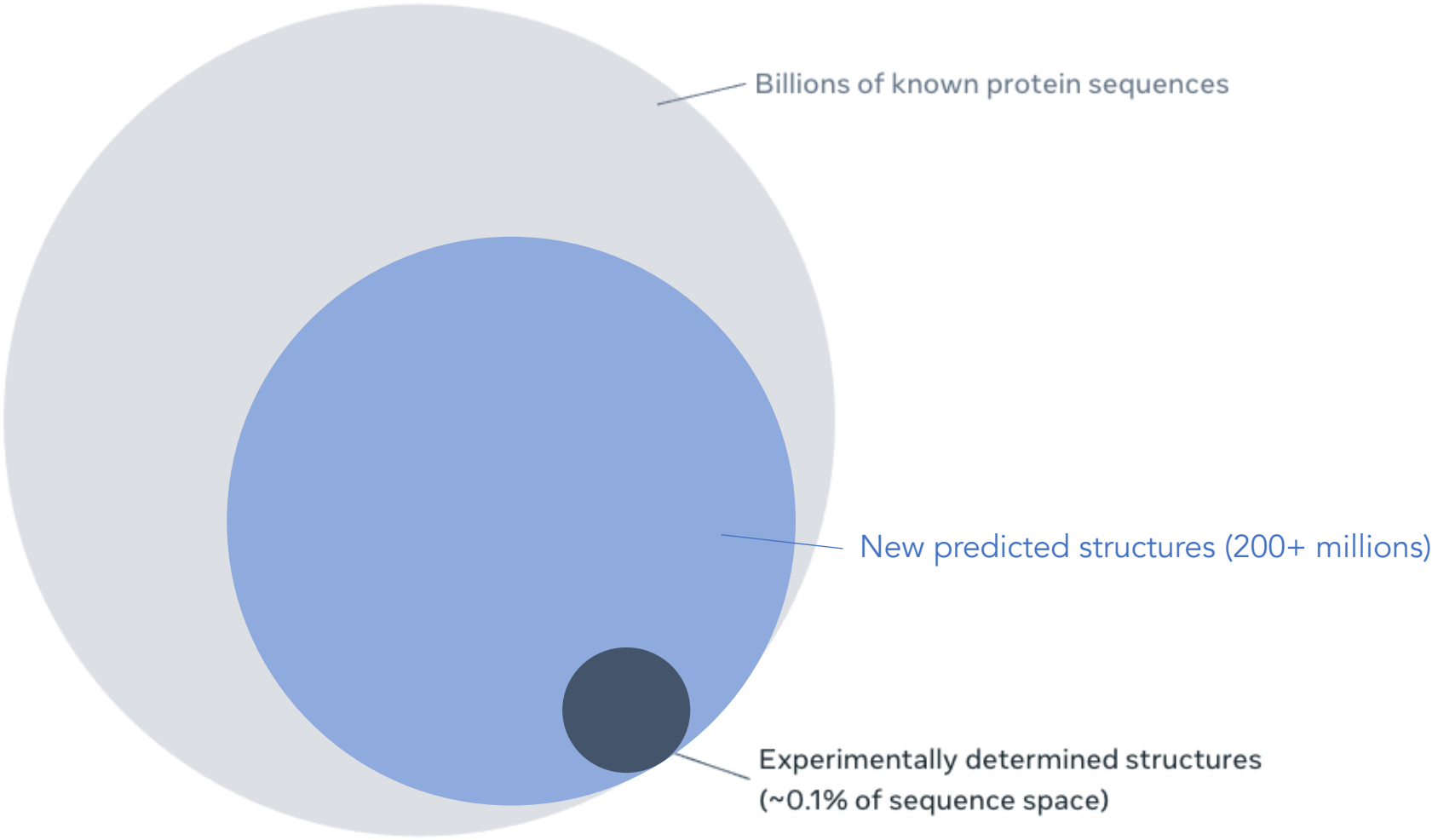
Substantial improvement in structure prediction from AlphaFold2

Also notable in new capabilities for predicting protein-protein interactions (although not perfect yet)

Applications:

- Interprets complex experimental data in structural biology at a new speed
- **Creates an unprecedented size of predicted structures for data mining & learning**
- …


Billions of known protein sequences
Experimentally determined structures
(~0.1% of sequence space)

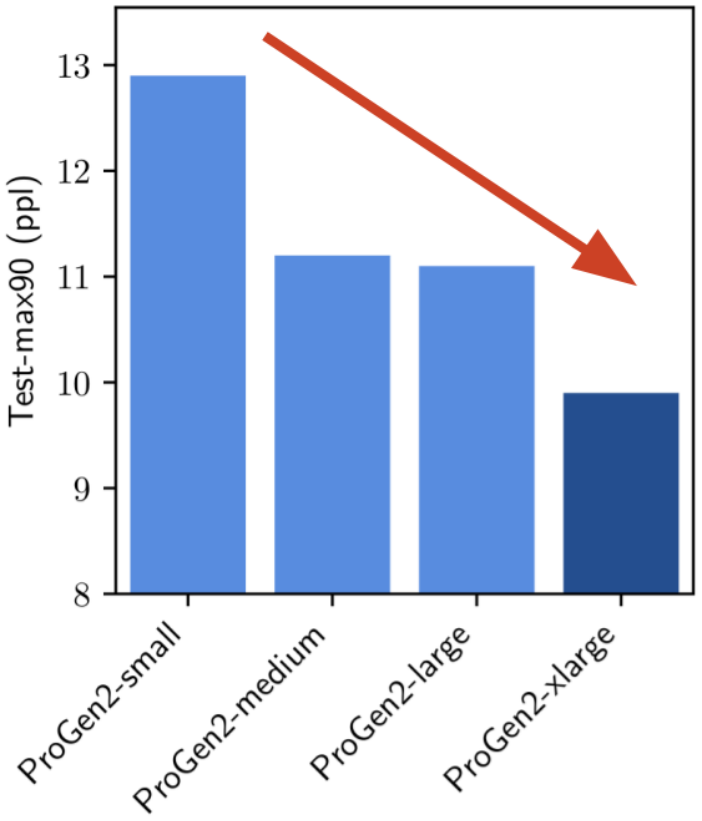

Billions of known protein sequences
New predicted structures (200+ millions)
Experimentally determined structures
(~0.1% of sequence space)

Protein sequence models

- Modeling the distribution of protein sequences through language models or other density models

- Bigger is better??

- Applications
  - Predicts effects of genetic mutations
  - Guides sequence choices in protein engineering



151M params

6.4B params

Sequence density model
(language model)

Protein sequence → Sequence likelihood

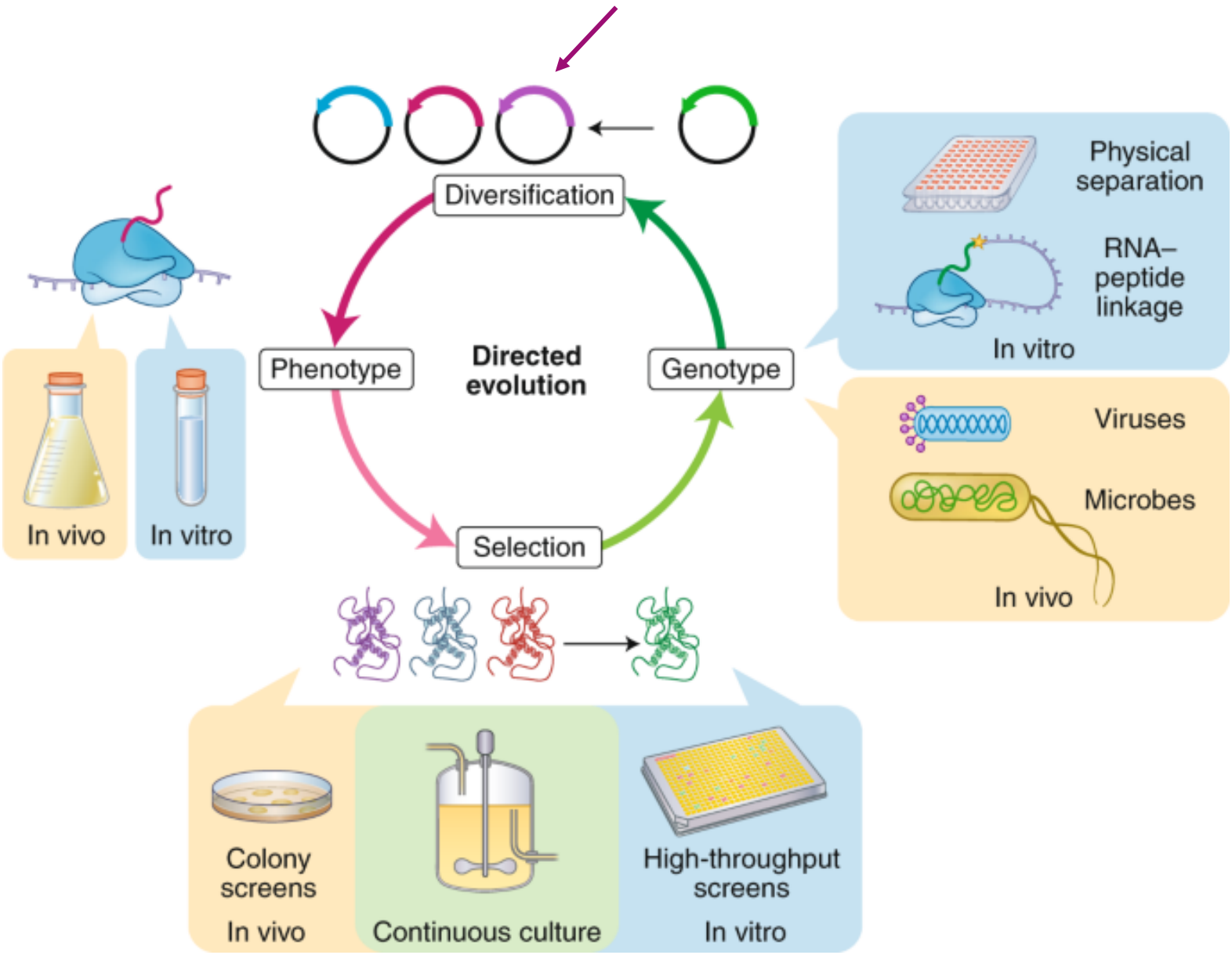
VALYDYEARTED — Neutral variation

I G — Disease variation

VALYDYEARTED — Neutral variation

R W — Enhanced variation

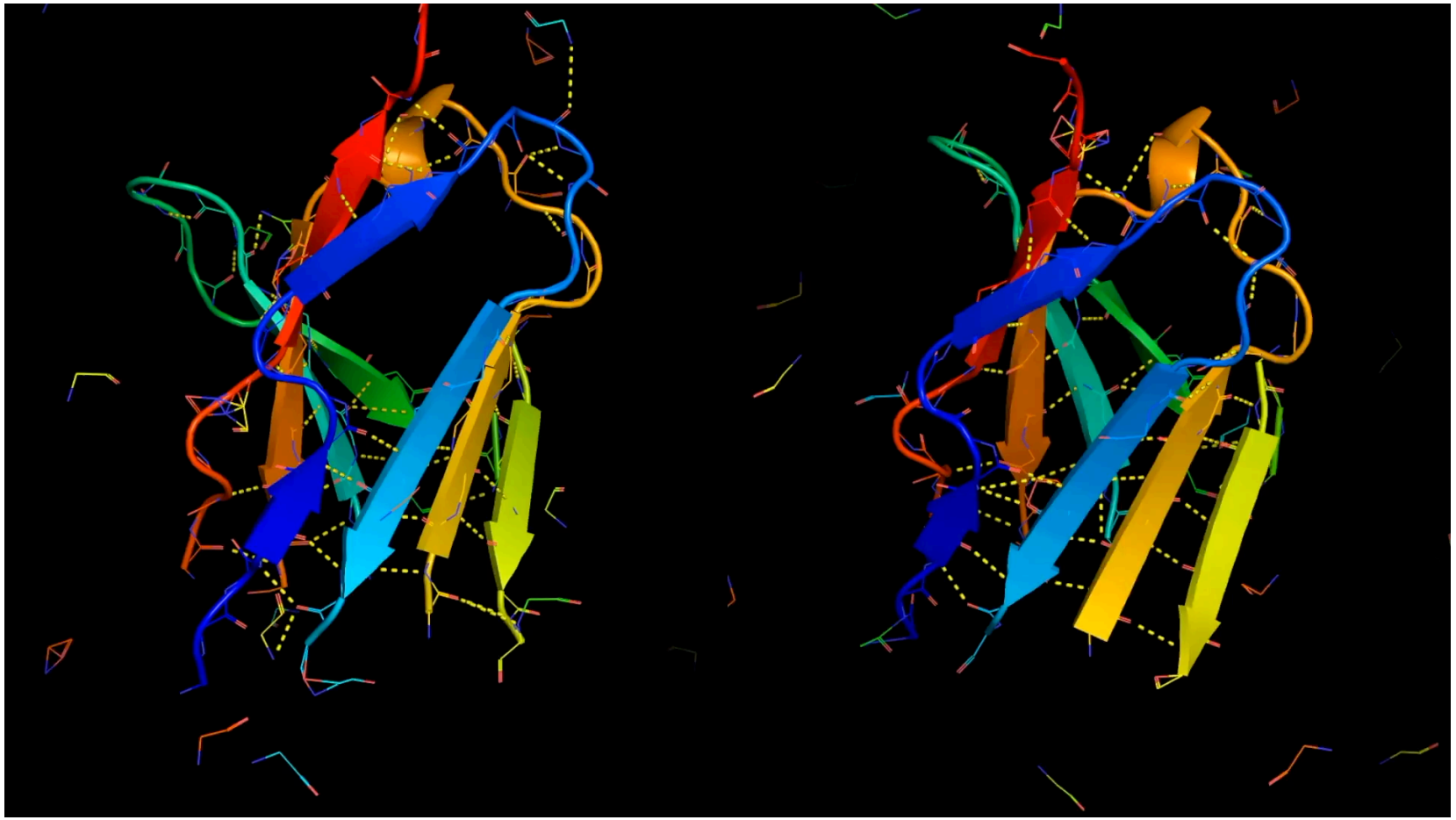
Machine learning-based library design



Morrison, M.S., Podracky, C.J. & Liu, D.R. The developing toolkit of continuous directed evolution. Nat Chem Biol 16, 610–619 (2020).

Generative models for protein design

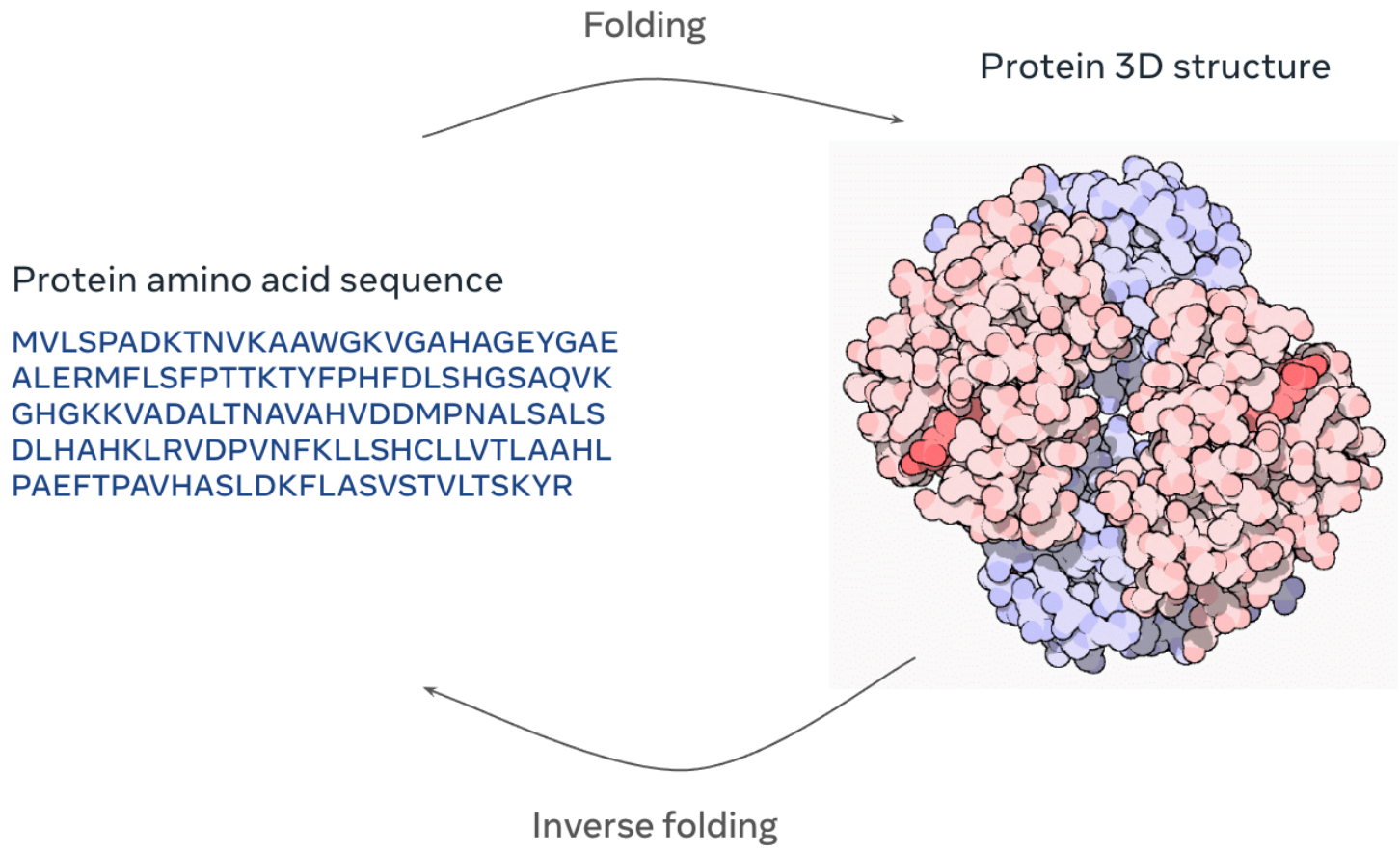
Denoising diffusion probabilistic model for protein structure and sequence



Anand, Namrata, and Tudor Achim. "Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models." arXiv preprint arXiv:2205.15019 (2022).

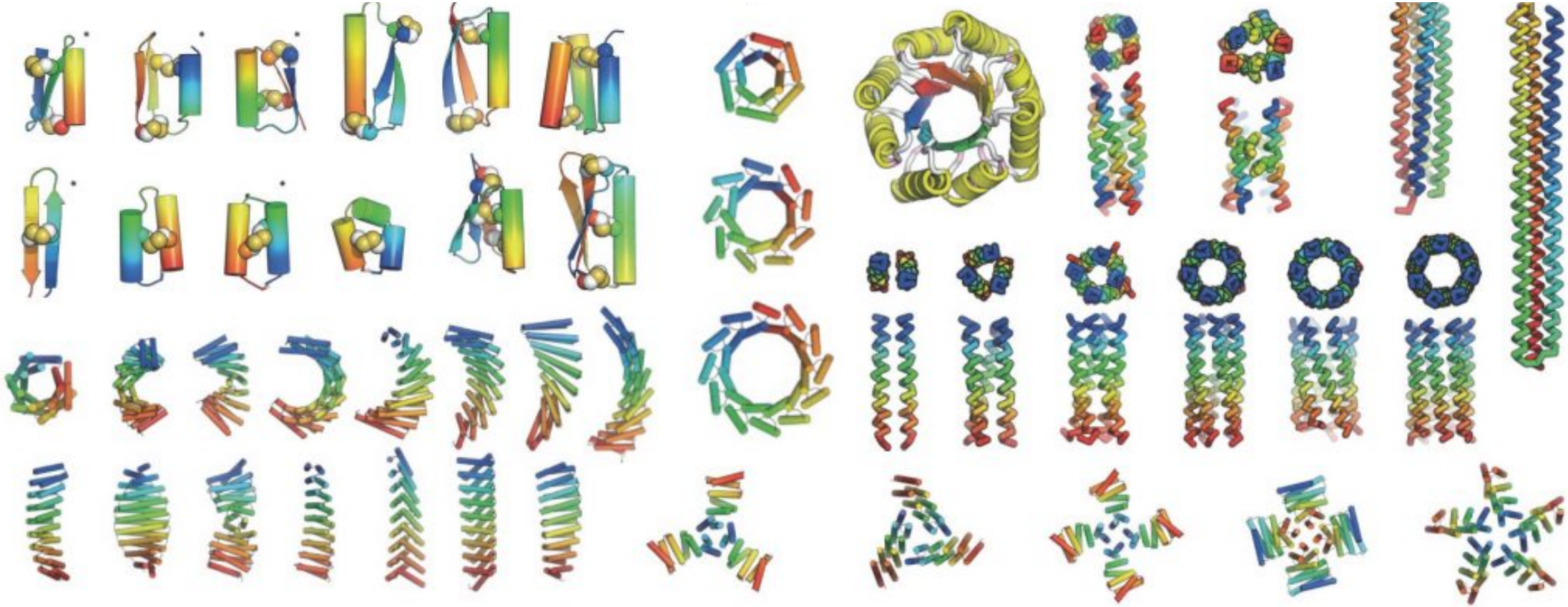
Generative models for protein design

Conditional generative models for protein design, e.g. inverse folding models



Hsu, C., Verkuil, R., ... & Rives, A. (2022). Learning inverse folding from millions of predicted structures. ICML 2022.

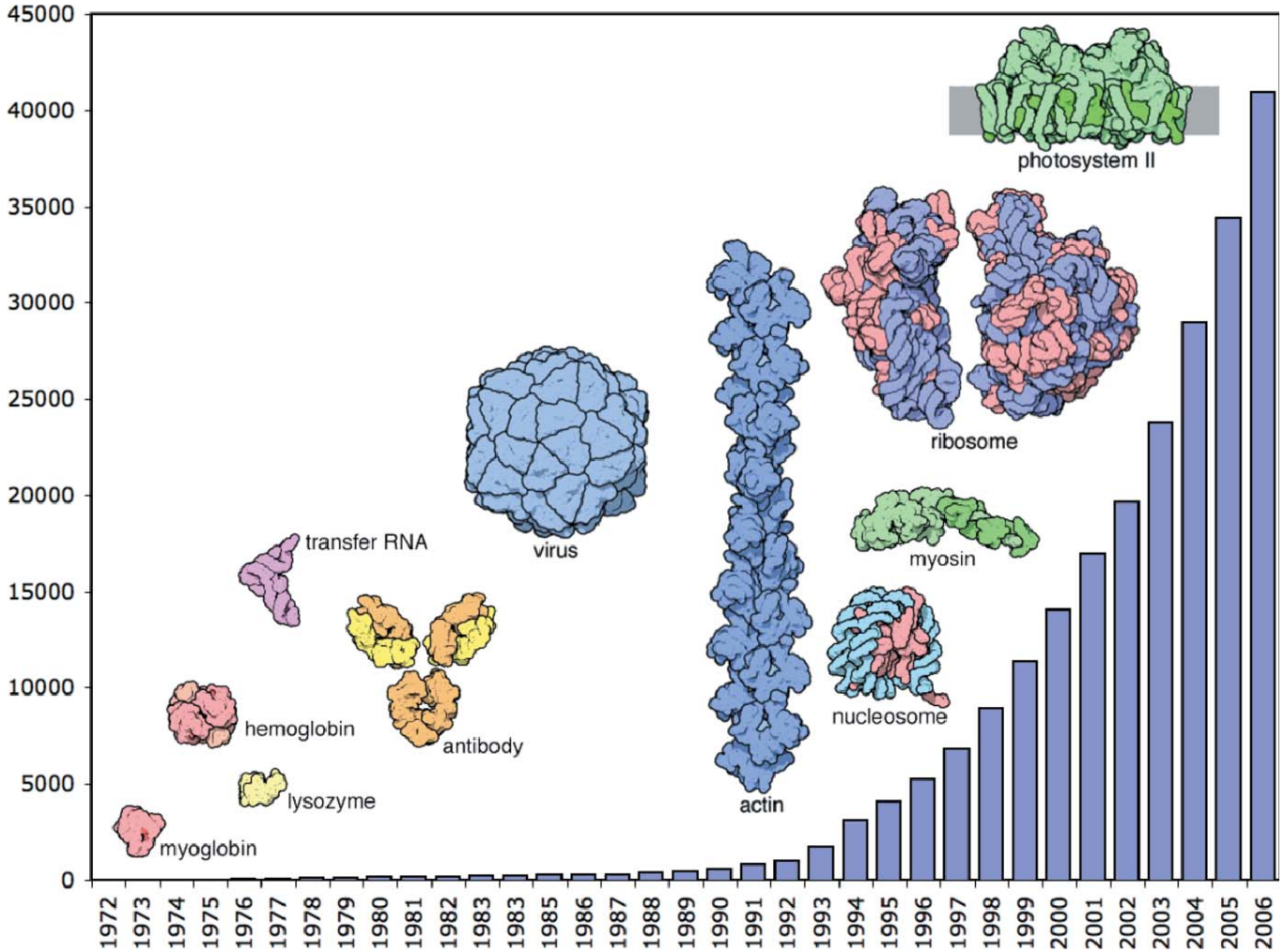
Beyond naturally existing proteins: De novo proteins



Huang, Po-Ssu, Scott E. Boyken, and David Baker. "The coming of age of de novo protein design." *Nature* 537.7620 (2016): 320-327.

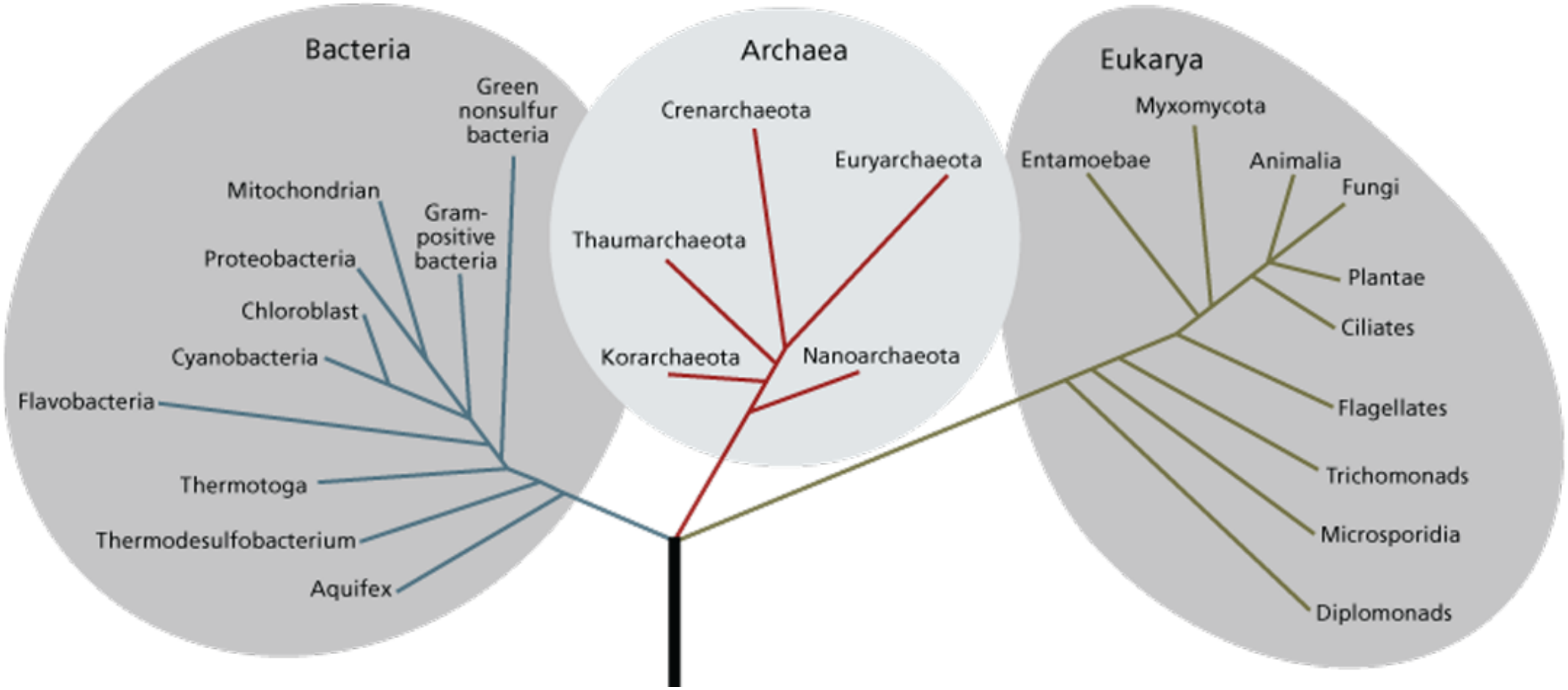
“How”: The data and the learning algorithms

Protein structures: Protein Data Bank



(Today over 100,000)

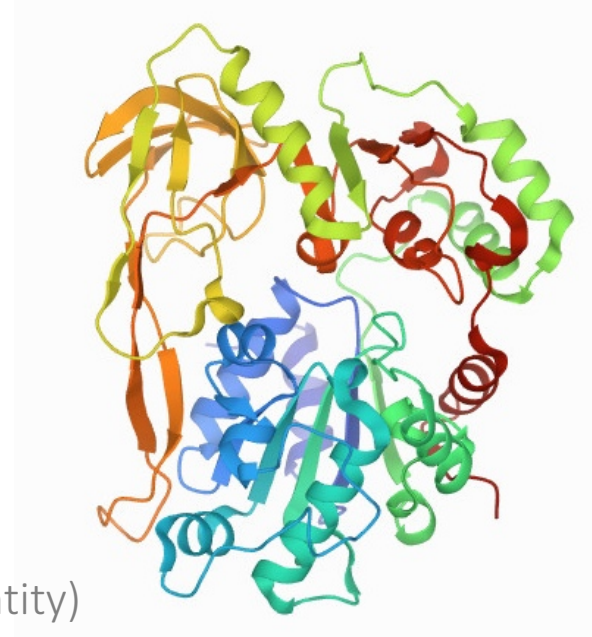
Evolutionary data: observing the products of evolution

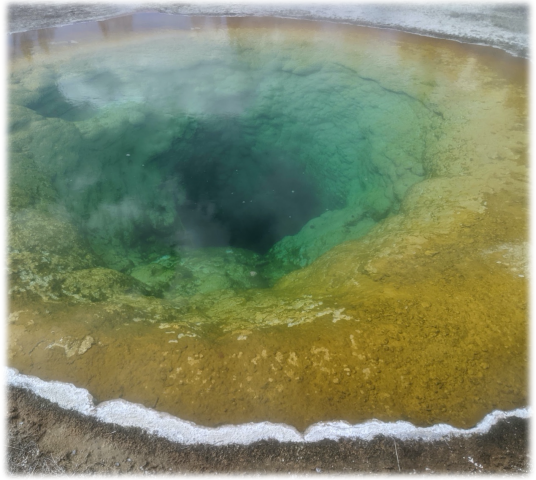


Evolutionary data: observing the products of evolution

Human PIF1 helicase (PDB: 6HPH)

Thermophilic bacteria PIF1 helicase (PDB: 6S3E)





(~30% sequence identity)

Evolutionary data: observing the products of evolution

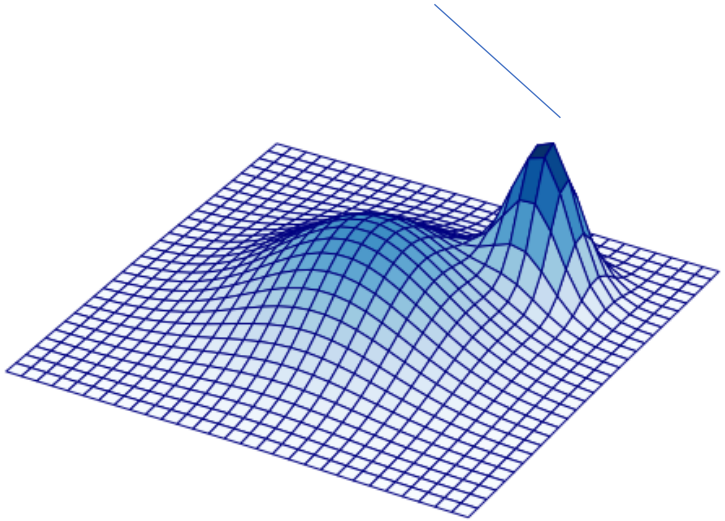
Multiple sequence alignment (MSA)

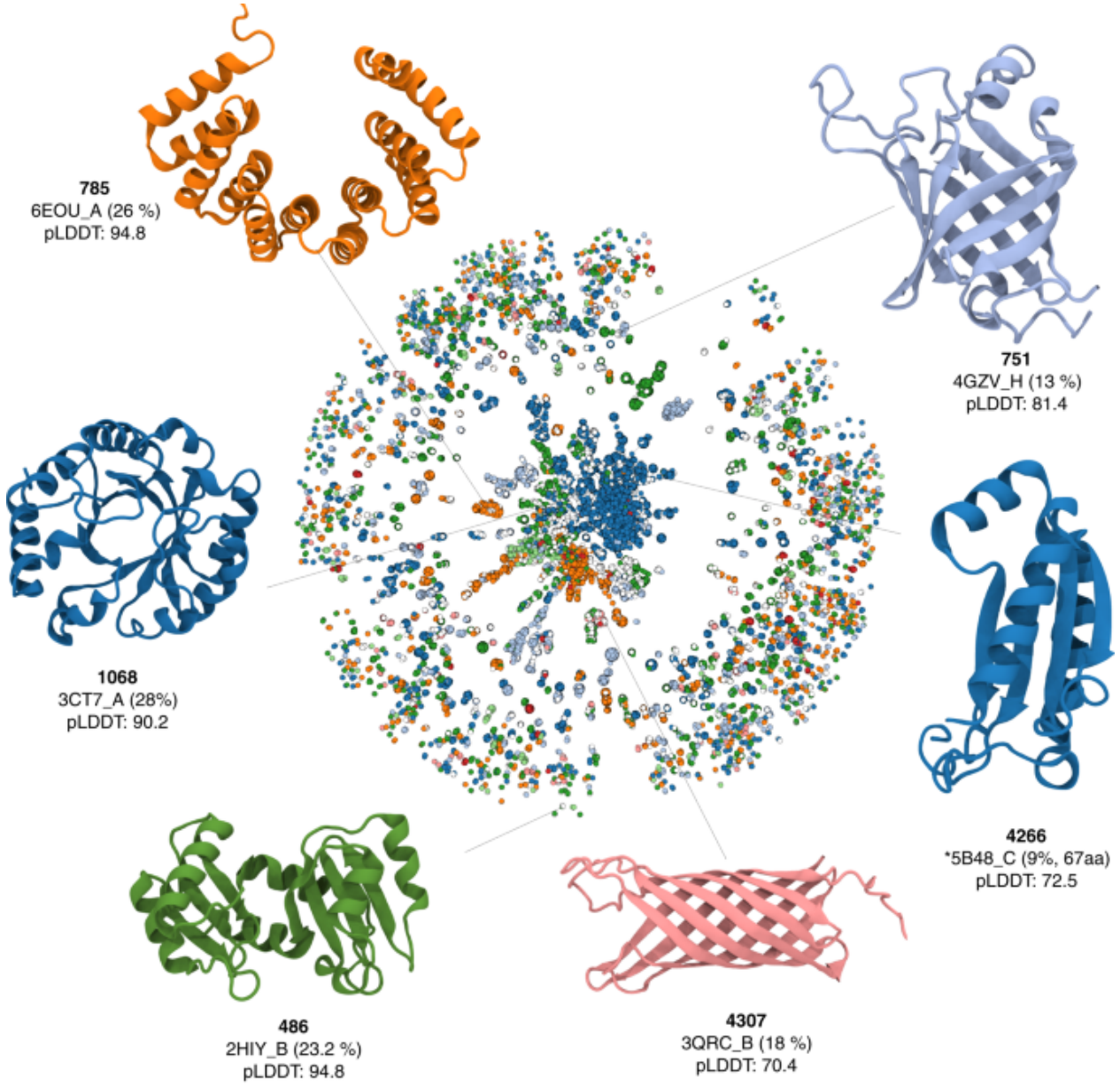
PIF1 helicase

```
WALRKTRKRLEEPFGGVKVLLLGDTRQLEPVVPGGEEALYIARTWGGPFFFQAHVWEE--  180
   R  R++  +PFGG+++++ GD  QL PV  G +            F FQ+  W+
AVARAVRQQ-NKPFGGIQLIICGDFLQLPPVTKGSQPP---------RFCFQSKSWKRCV  168

-VALRVHRLWESQRQREDPLFAELLKRLRQG--DPQALETLNRAAVRPDGGEEPGTLILT  237
 V L + ++W    ++ D  F  LL+ +R G    +    L   A    G +      L
PVTLELTKVW----RQADQTFISLLQAVRLGRCSDEVTRQLQATASHKVGRDGIVATRLC  224

PRRKEADALNLKRLEALPGKPLEYQAQVKG-EFAET---DFPTEAALTLKKGAQVILLRN  293
  + +    N +RL+ LPGK   ++A     E A T     P    L LK GAQV+L++N
THQDDVALTNERRLQELPGKVHRFEAMDSNPELASTLDAQCPVSQLLQLKLGAQVMLVKN  284

DPLGE-YFNGDLGWVEDLEAEALAVRLKR--NGRRVVIRPFVWEKIVYTYDSEREEIKPQ  350
  +     NG  G V   EAE   +   R   G   VI    W     T  +   ++  +
LSVSRGLVNGARGVVVGFEAEGRGLPQVRFLCGVTEVIHADRW-----TVQATGGQLLSR  339

VVGTFRQVPVRLAWALTVHKAQGLTLDKVHLELGRGLFAHGQLYVALTRVRRLQDL  406
     +Q+P++LAWA+++HK+QG+TLD V + LGR +FA GQ YVAL+R R LQ L
-----QQLPLQLAWAMSIHKSQGMTLDCVEISLGR-VFASGQAYVALSRARSLQGL  389
```

Evolutionary density model

High density area = likely functional proteins





Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* **13,** 4348 (2022).

Before deep learning: Simple models on evolutionary data

Hidden Markov Models (HMMs) only model the site-specific amino acid features.

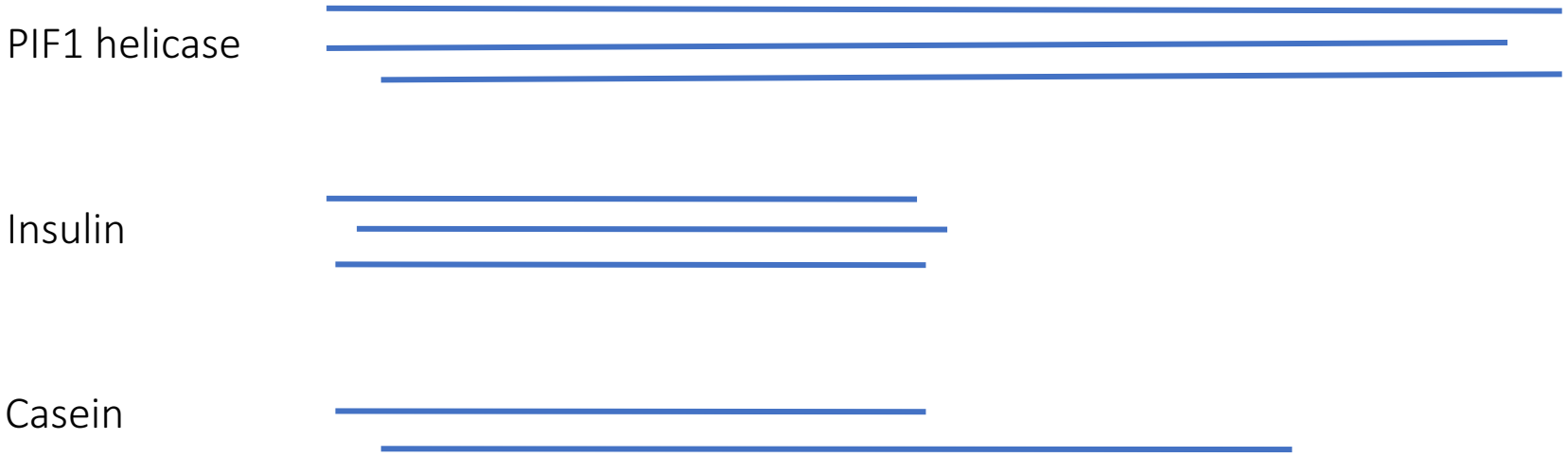
Potts models, also known as Markov Random Fields (MRFs), model the site-specific amino acid features and the pairwise interactions between sites.

They work quite well, too

(when used on specific protein families in combination with search algorithms)

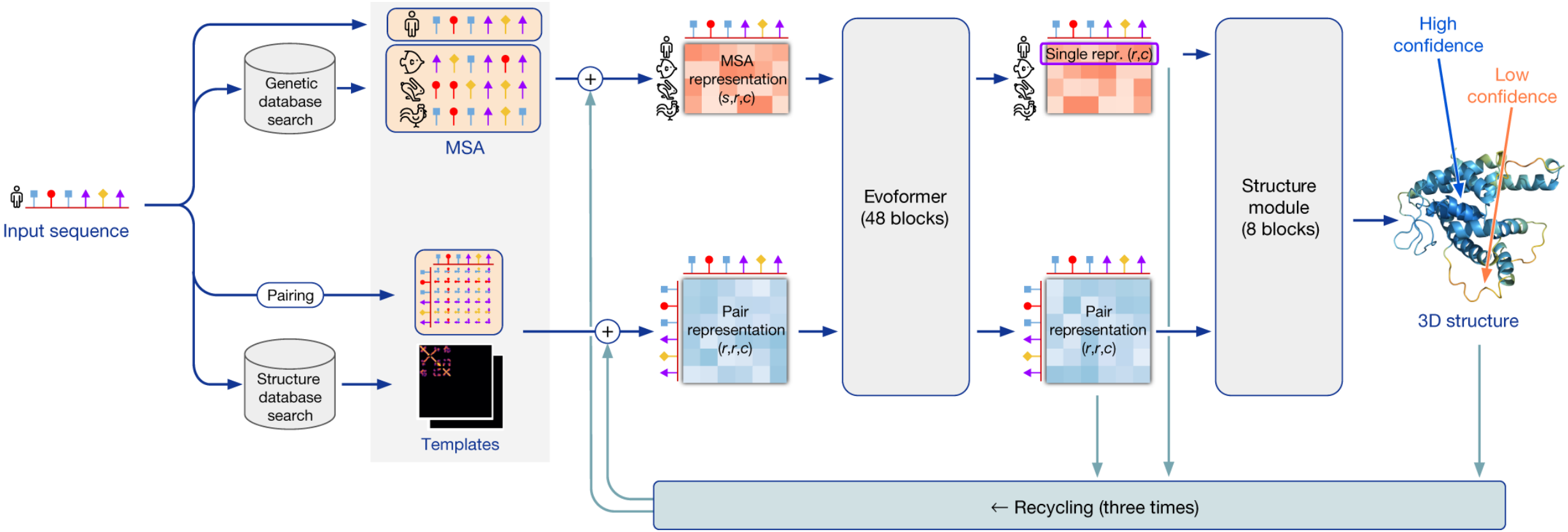
Replacing multiple sequence alignments with deep learning models

(Protein language models)

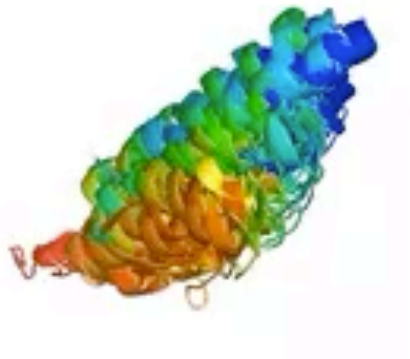


New ability to jointly model all protein families regardless of “alignments”

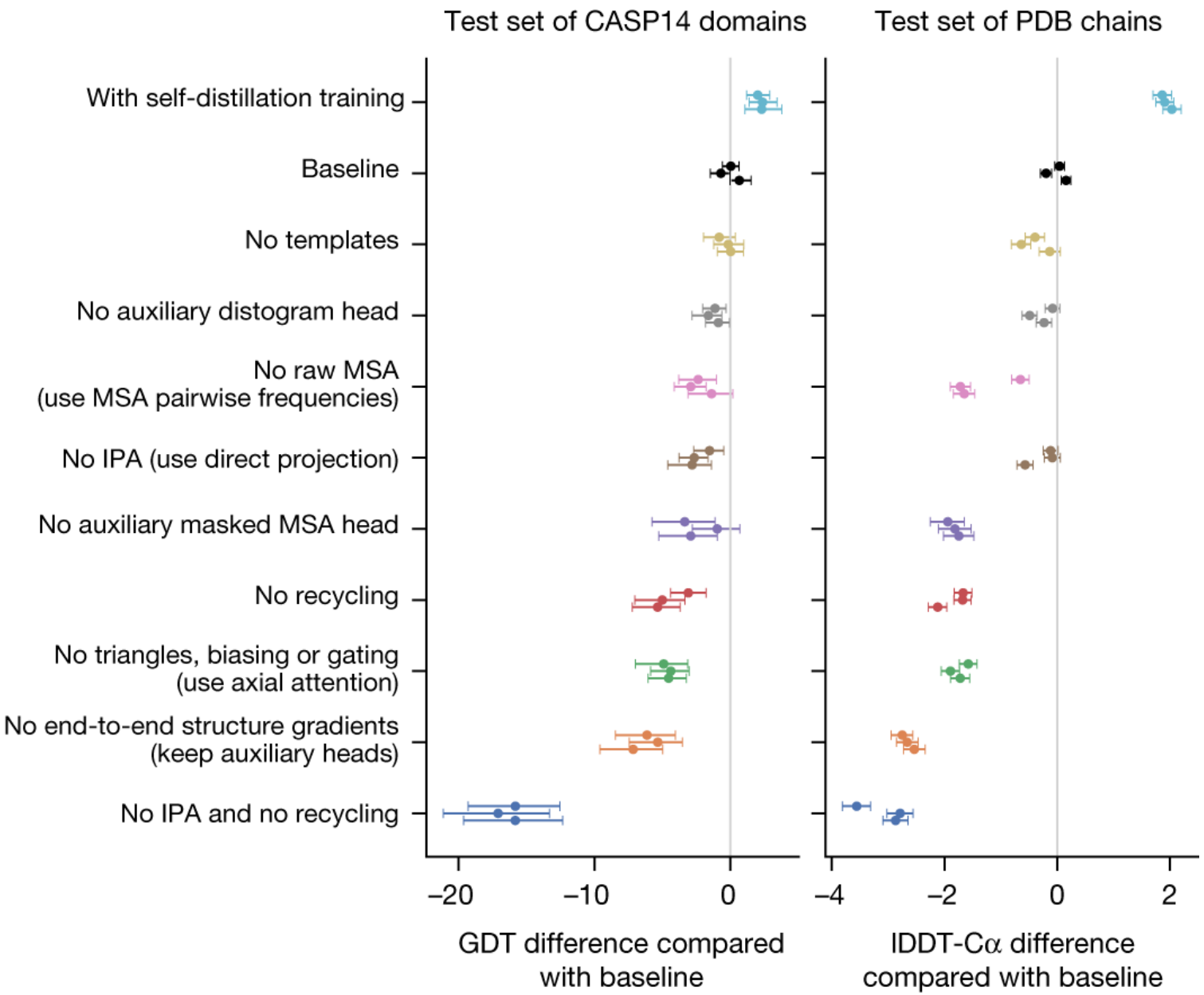
AlphaFold2 also relies on evolutionary data (multiple sequence alignments)



Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).



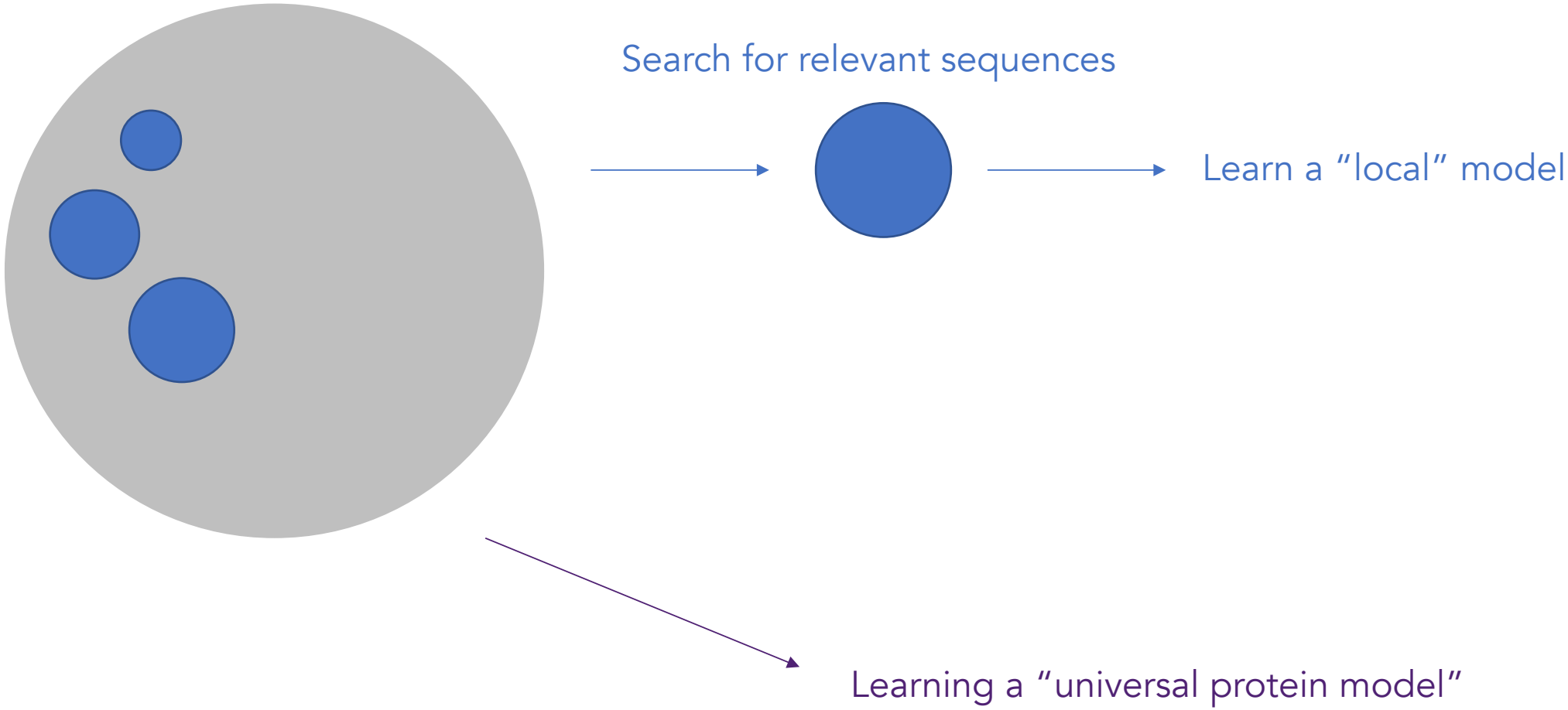
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction



Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

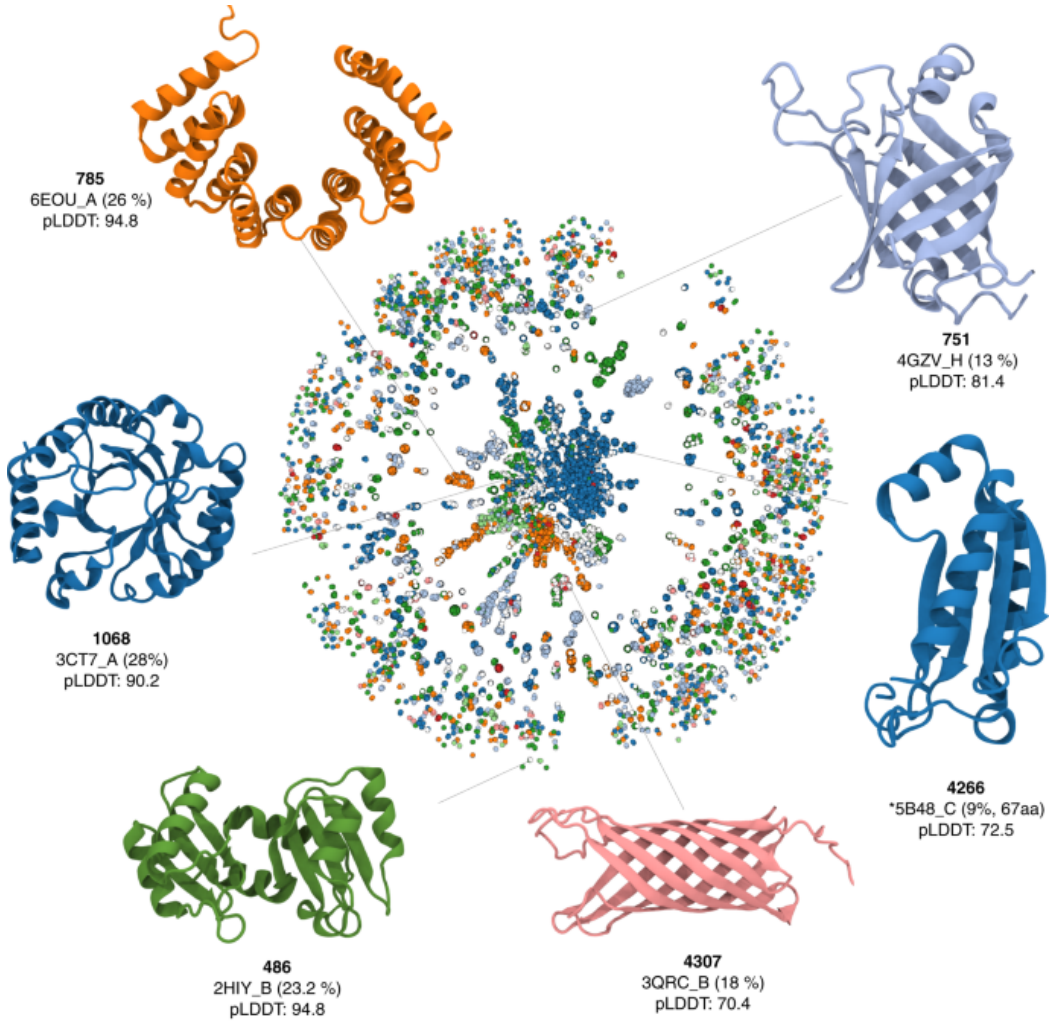
“Why”: Common themes and open questions

“Search & learn” vs. “universal learning”



“Generalization” vs “memorization”

Are language models / single-sequence folding models effectively memorizing the evolutionary database?



Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* **13,** 4348 (2022).