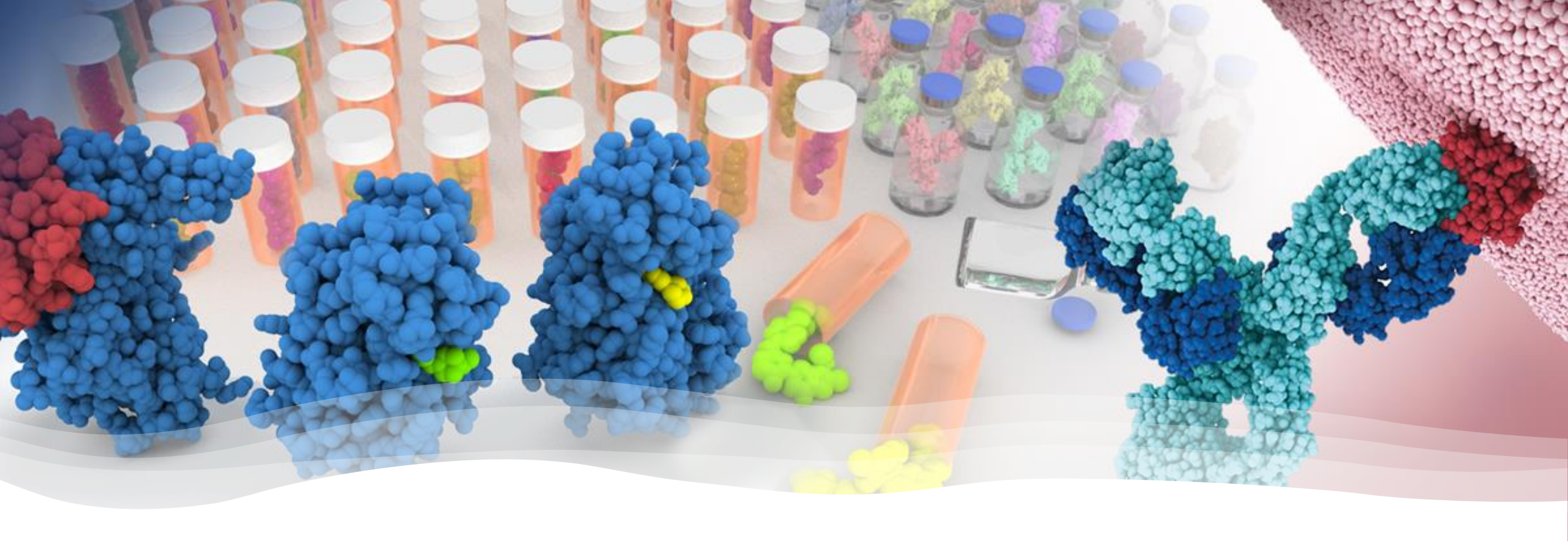


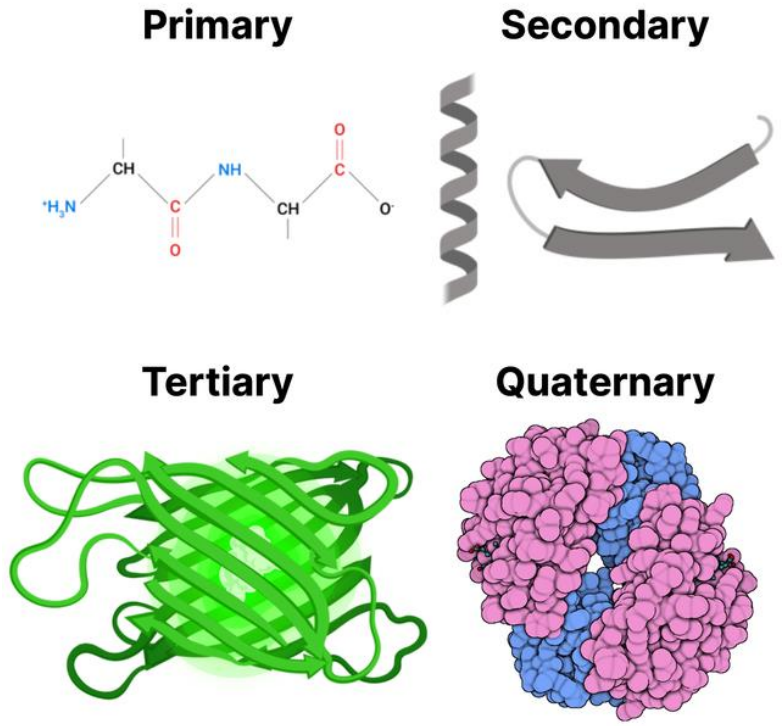
Proteins in Drug Discovery:
Protein structures and how to modulate protein function(s)

Outline

1. Protein structures
2. Scope and complexity of the protein world
3. Modulating protein functions
4. Chemical space: finding chemicals that modulate protein functions
5. Case studies of protein complexity and nuance

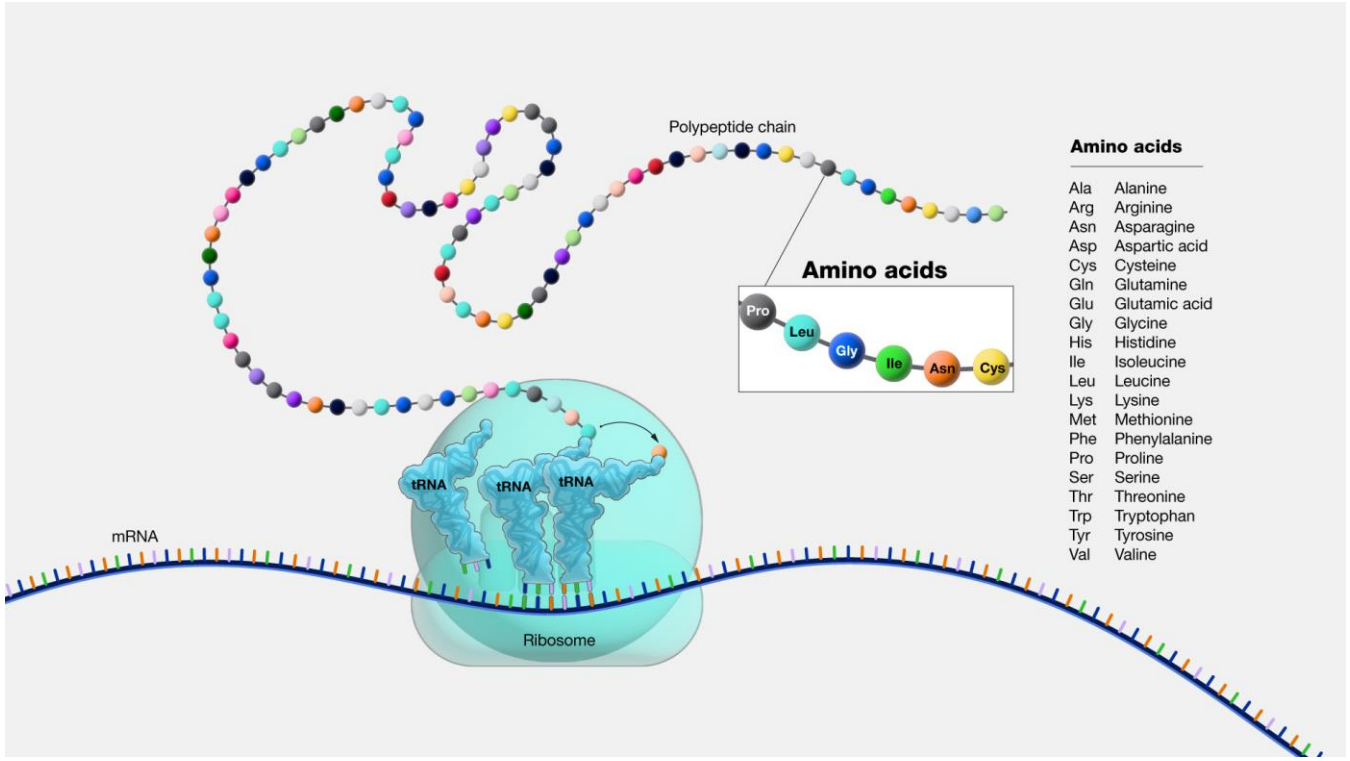
Protein structure

- Matters for our ability to modulate protein function (drugs)
- Classical levels of organization
  - Primary to quaternary



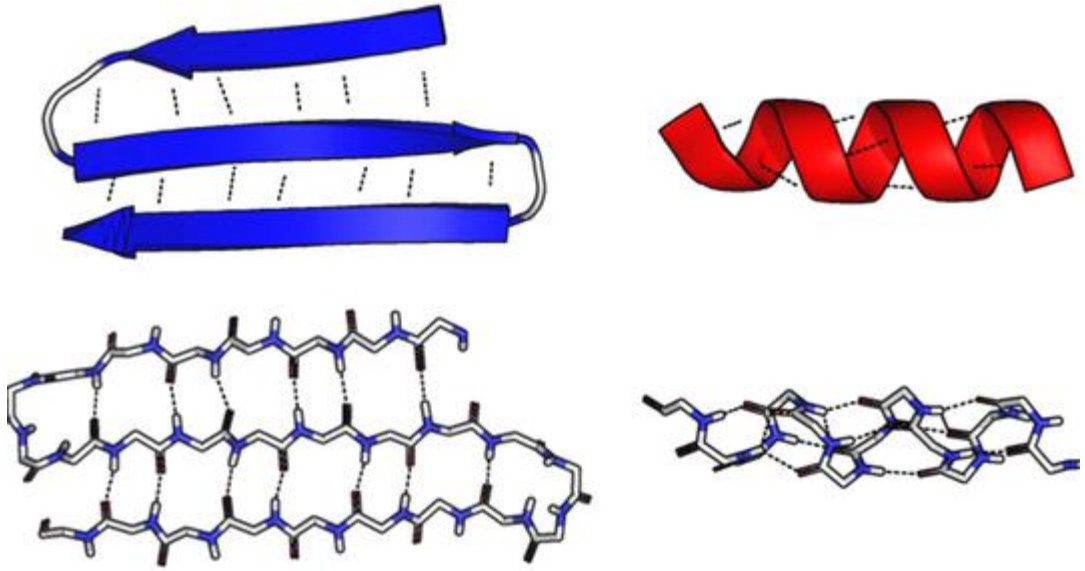
Proteins: primary amino acid (AA) sequence

We can consider amino acid sequence as the **input space**

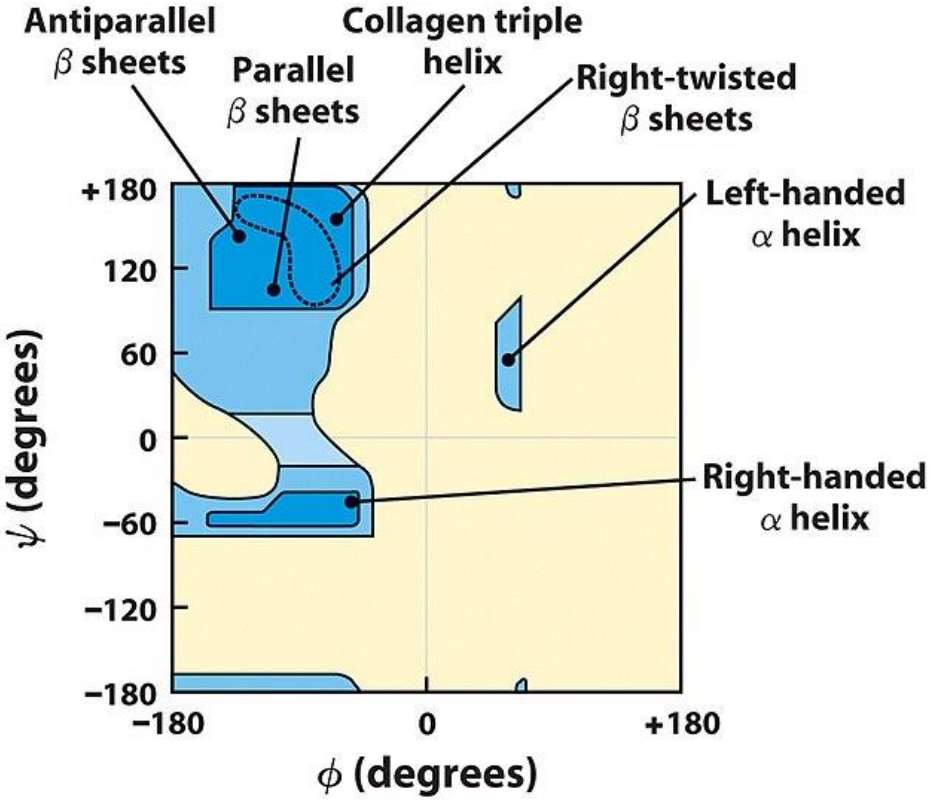


Proteins: secondary structure

- Local folding patterns
- Determined by AA sequence and stabilized by hydrogen bonds between the protein backbones



Different amino acids have different conformational properties due to angles (phi, psi) in the amino acids



Images: Chemistrytalk

Proteins: tertiary structure

- 3D structure of all of the amino acid residues in the protein
- AA's that are far away in the primary sequence may be close together in the tertiary sequence

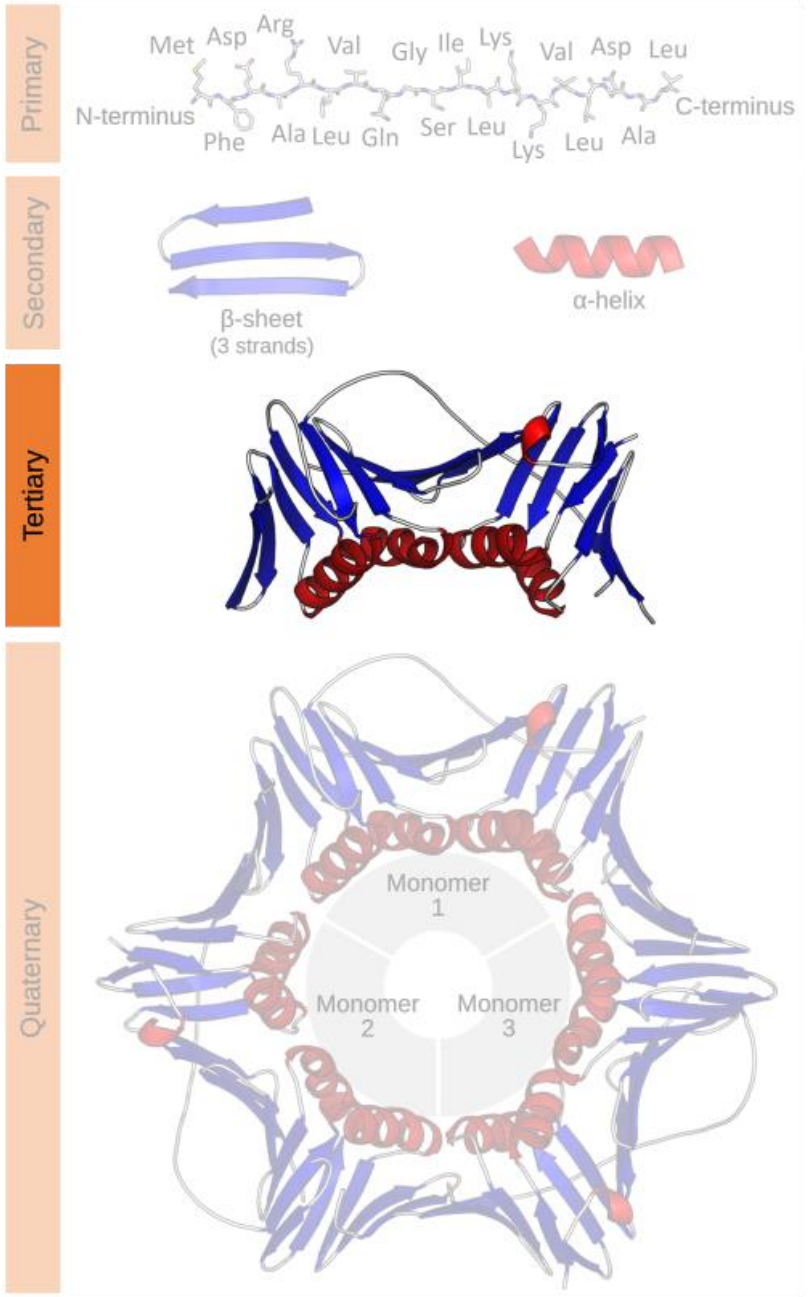


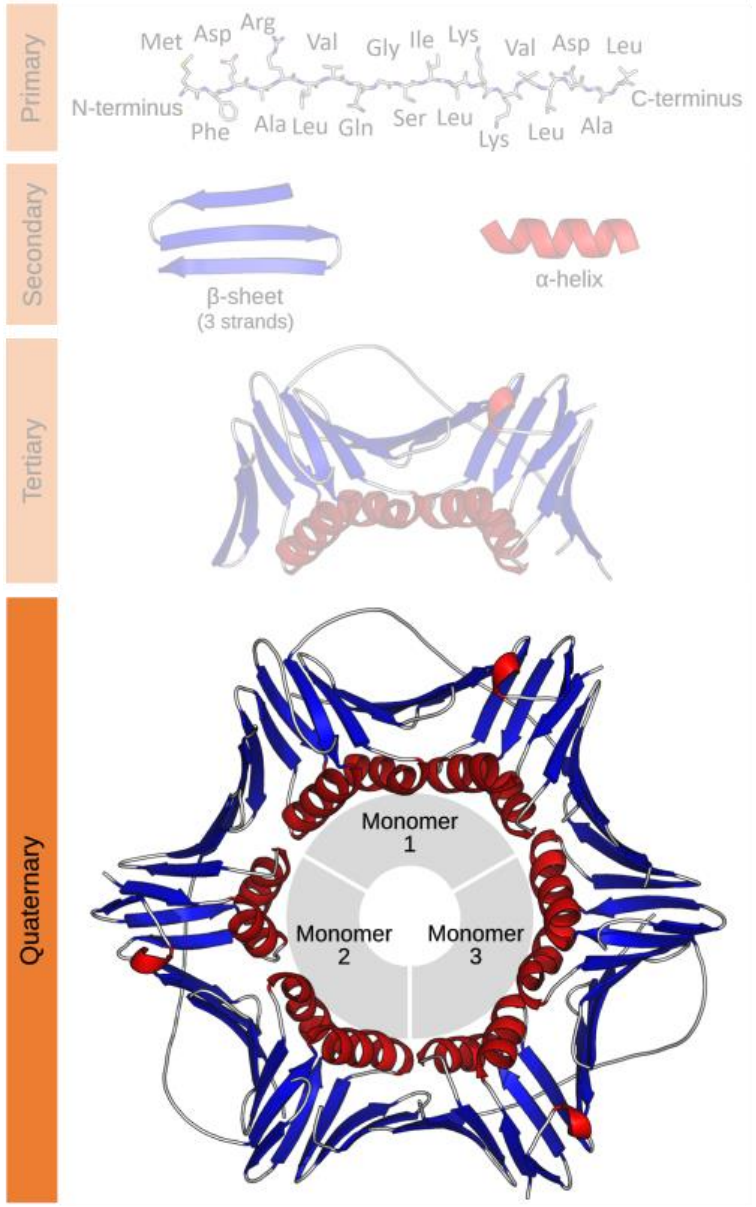
Image: wikipedia

Proteins: quaternary structure

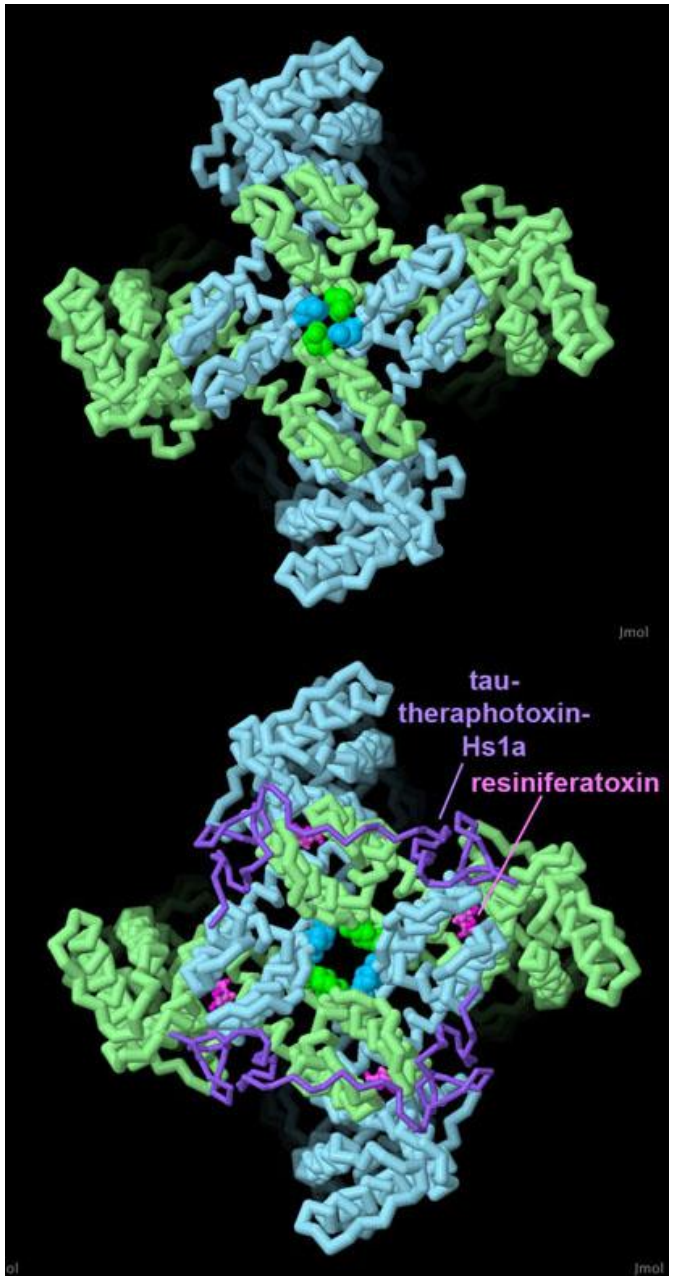
- Proteins composed of multiple subunits
  - multiple smaller protein chains
- Function of complex different than single subunits

Examples:

- PCNA
  - Proliferating cell nuclear antigen
- TRPV1 channel
  - Ion channel, sensitive to chilli peppers etc



PCNA (proliferating cell nuclear antigen)

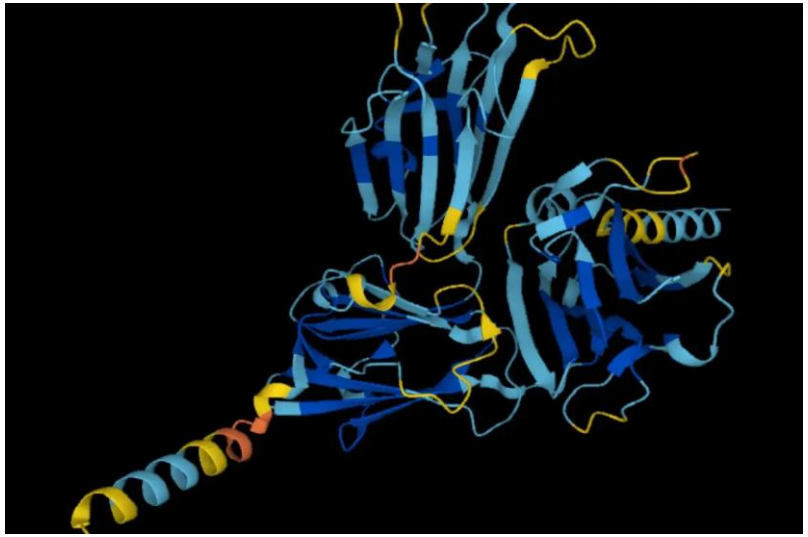


TRPV1 ion channel in closed and open forms

The protein folding problem

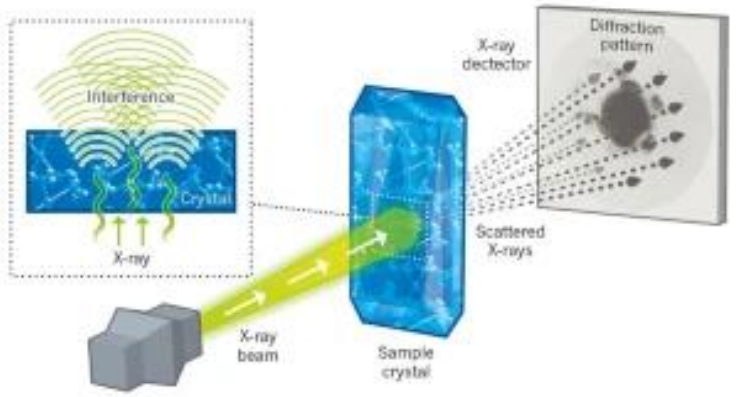
- Primary AA sequence determines 3D structure
  - This is what AlphaFold attempts to predict – see Module 1 and Dr. Maddison's lectures

- In cells, 3D structure of proteins is influenced by:
  - Local pH and local ionic strength
  - Chaperones, transporters and binding proteins
  - Coupling between protein production (translation) and protein folding
  - Redox environment (disulfide bonds)
  - Macromolecular crowding (cytoplasm is up to 40% macromolecules by volume)
  - Post-translational modifications
    - Small tags (phosphate, acetyl, etc) added to protein by cellular enzymes to regulate function
- Misfolded proteins → disease



Experimental aspects of protein structure determinations

- Early protein measurements:
  X-ray crystallography

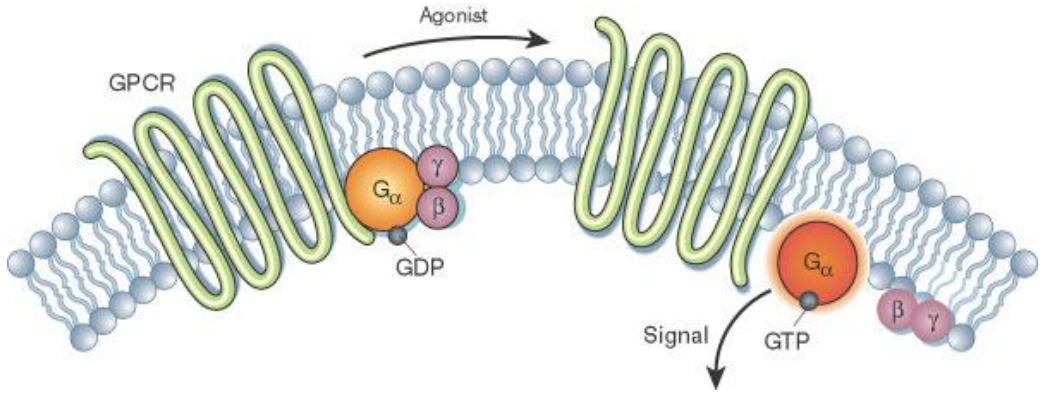


- For protein X-ray crystallography, require a crystal
- Membrane proteins very difficult to crystallize
- Membrane proteins are:
  - 25% of proteome
  - Disproportionately important in pharmacology (60% of clinical drugs)
  - < 5% of high-resolution protein structures in databases

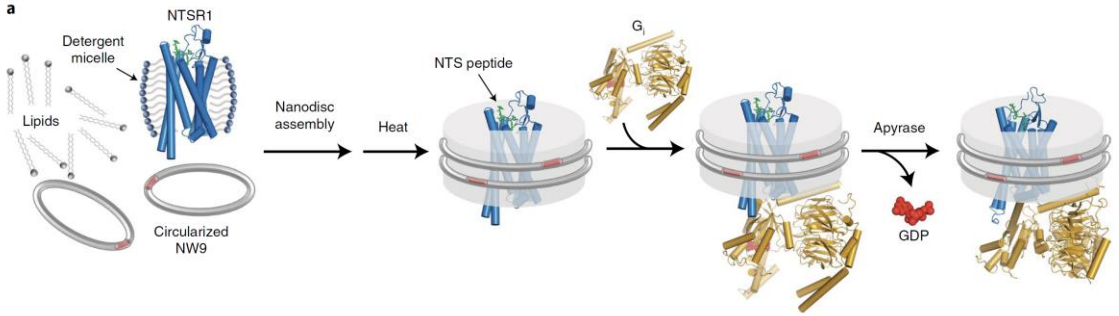
Image: Nature

GPCRs: key membrane proteins highly relevant for drug discovery

- G protein-linked receptors (GPCRs)
- 30% of FDA-approved drugs
  - Opioids, antihistamines, migraine drugs
- Important roles in cell signaling and physiology



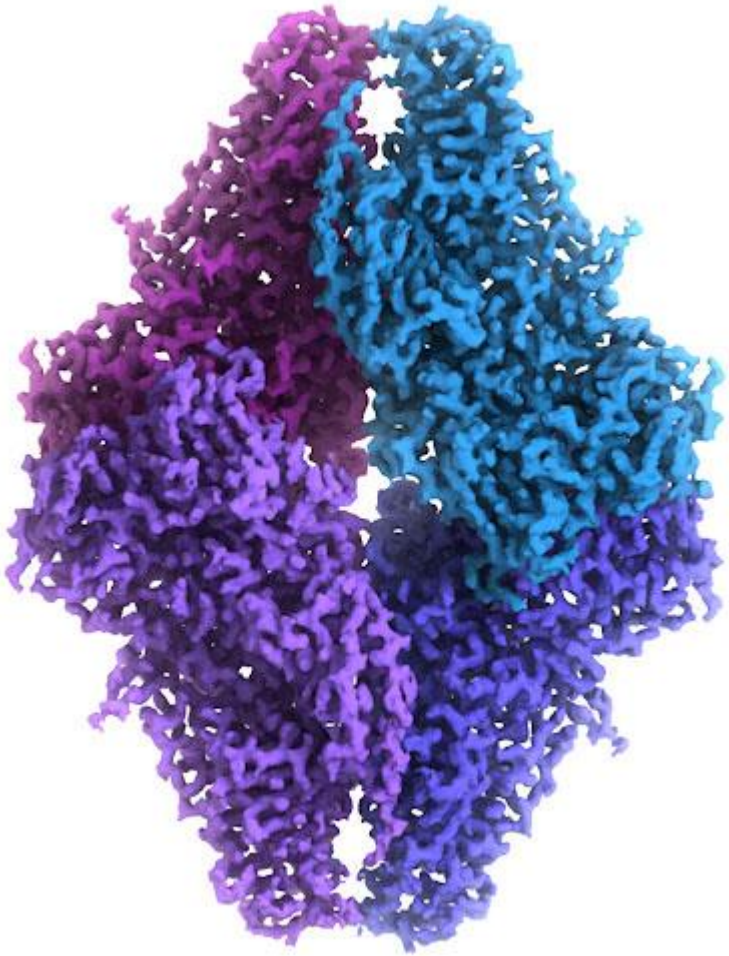
- Structural experiments aided by lipid nanodiscs to stabilize the GPCR:
- Structures are still under-characterized, especially relative to their DD importance



Zhang et al 2021, Nat Struc Mol Biol

Cryo-electron microscopy (Cryo EM)

- Helpful for structures of large macromolecular complexes
- many thousands of different single particles preserved in a thin layer of non-crystalline ice
  - (cryo-EM).
- If these views show the molecule in myriad different orientations..
- a computational approach (similar to CAT scans) will yield a 3D mass density map.
- Example protein: beta galactosidase (bacterial enzyme)
  - cryoEM data at EMDataBank entry EMD-2984
  - the atomic coordinates are in PDB entry 5a1a.



Technique	Typical Resolution	What Is Actually Measured	Key Physical / Experimental Notes	Primary Data Outputs	File Formats & Data Architecture
X-ray Crystallography	~0.8–3.0 Å (atomic resolution common)	X-ray diffraction intensities from protein crystals	Requires well-ordered crystals	Electron density map + refined atomic model	Raw: reflection intensities (HKL)
			averages over many molecules		Processed: electron density grids
			hydrogen atoms often not visible		Model: atomic coordinates (PDB / mmCIF)
Cryo-Electron Microscopy (**Cryo-EM**)	~2–4 Å routinely; best <2 Å	2D projection images of single particles embedded in vitreous ice	No crystallization	3D Coulomb potential (density) map + fitted atomic model	Raw: 2D images (movies)
			proteins frozen in near-native state		Map: 3D voxel grid (MRC/CCP4)
			structural heterogeneity possible		Model: atomic coordinates (PDB/mmCIF)
NMR Spectroscopy	~2–4 Å (ensemble, not single structure)	Nuclear spin interactions (chemical shifts, NOEs, J-couplings)	Solution-state; captures dynamics;	Structural restraints + ensemble of conformations	Restraints: distance/angle constraints
			size-limited (~<50 kDa typically)		Models: multiple coordinate sets (PDB with ensembles)
Small-Angle X-ray Scattering (SAXS)	~10–30 Å (low resolution)	Scattering intensity vs angle in solution	Shape-level info; no atomic detail	Global shape descriptors	Data: 1D intensity curves
					Models: low-res envelopes

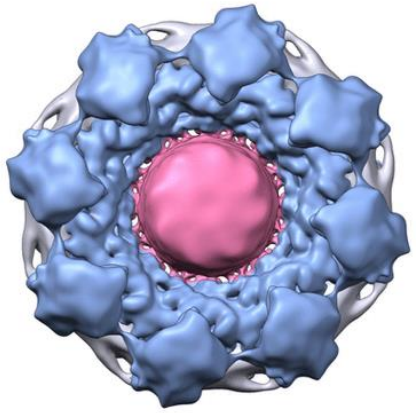
Takeaways
from data about protein structure experiments

- X-ray & Cryo-EM don’t measure atoms
  - they measure diffraction or images.
- Atomic coordinates are inferred models
- Resolution differs between techniques

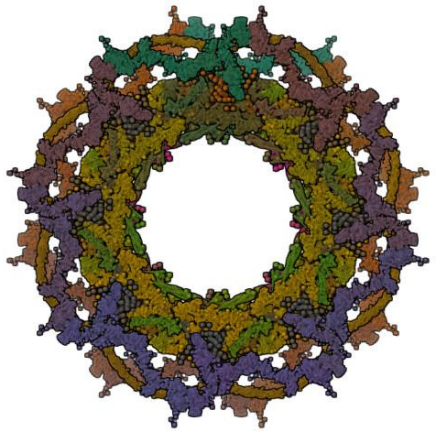
- ML models usually train on **processed interpretations**, not primary experimental signals.

Opportunities in AI: Integrative Modeling

- Combination of methods:
  - structural biology methods
    - X-ray crystallography, NMR spectroscopy and electron microscopy
  - experimental methods
    - small angle solution scattering, Forster resonance energy transfer, chemical crosslinking, mass spectrometry, electron paramagnetic resonance spectroscopy, and other biophysical techniques
- Composite models may cover a variety of scales
- Example: Nuclear Pore Complex:



Cryo-EM



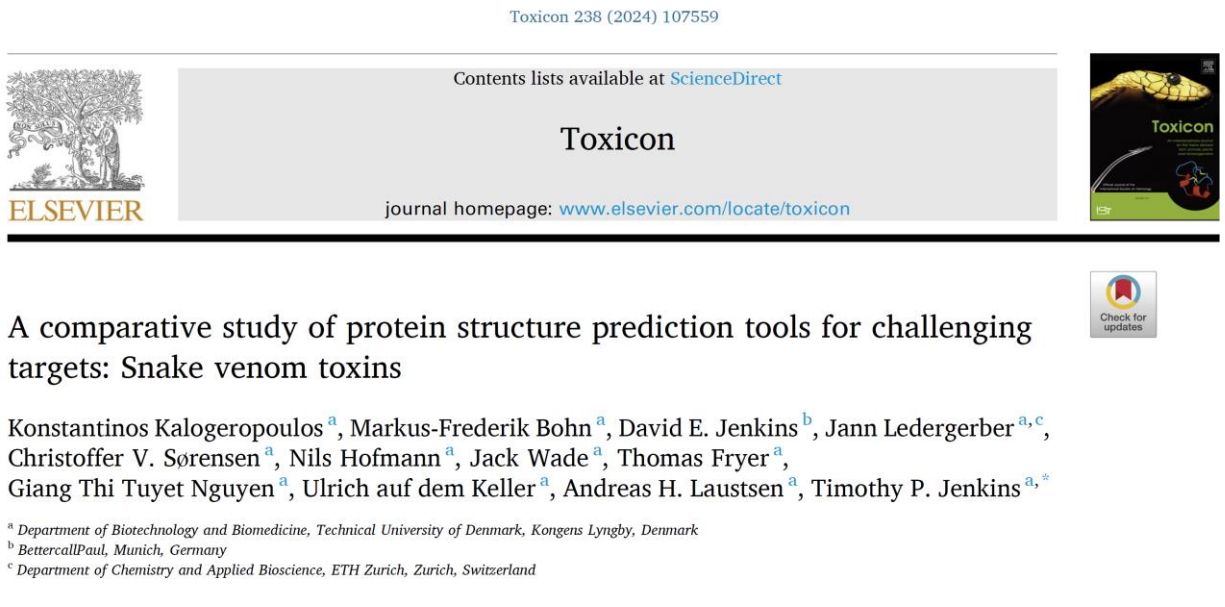
Integrative Modeling

"I believe that to be able to describe dynamic structures and predict structural changes in response to changing environmental conditions, it is important to incorporate different sources of data into the machine learning models, encompassing both simulations and experimental measurements, and to account for physics constraints and thermodynamic principles that need to be satisfied."

Cecilia Clementi, Frei Universitat of Berlin

Opportunities in AI for protein structure

- Complexity: understanding macromolecular dynamics and function, moving beyond the single structure frontier
- Complexity: modeling large, intricate, dynamic or transient complexes, particularly when conformational changes or weak interaction interfaces are involved
- Peptides and proteins as drugs: generative or design tools that allow engineering of new protein sequences (speed, success rate)
- Benchmarking: choice of structures



Toxicon 238 (2024) 107559

Contents lists available at ScienceDirect

Toxicon

journal homepage: www.elsevier.com/locate/toxicon

A comparative study of protein structure prediction tools for challenging targets: Snake venom toxins

Konstantinos Kalogeropoulos [a], Markus-Frederik Bohn [a], David E. Jenkins [b], Jann Ledergerber [a,c], Christoffer V. Sørensen [a], Nils Hofmann [a], Jack Wade [a], Thomas Fryer [a], Giang Thi Tuyet Nguyen [a], Ulrich auf dem Keller [a], Andreas H. Laustsen [a], Timothy P. Jenkins [a,*]

[a] Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby, Denmark
[b] BettercallPaul, Munich, Germany
[c] Department of Chemistry and Applied Bioscience, ETH Zurich, Zurich, Switzerland

Selectivity: and protein structure

Selectivity:

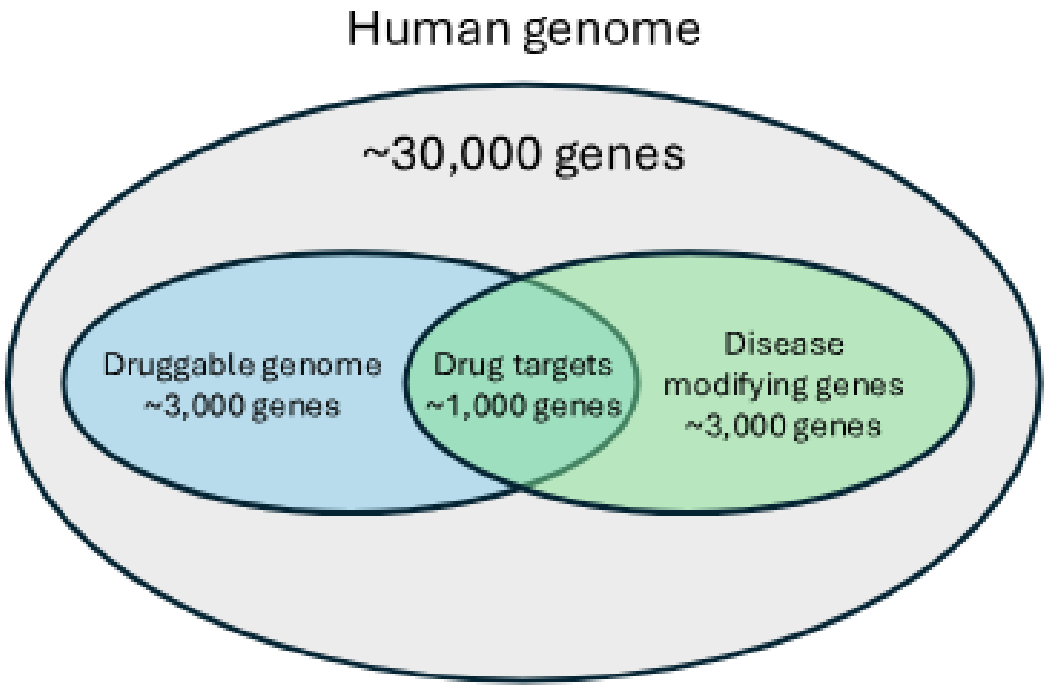
- your drug acts on intended protein and not on unintended off-target proteins
- Selectivity as a **structural** problem

- AI/ML work on protein structure →
  - impact on drug development
    - including selectivity

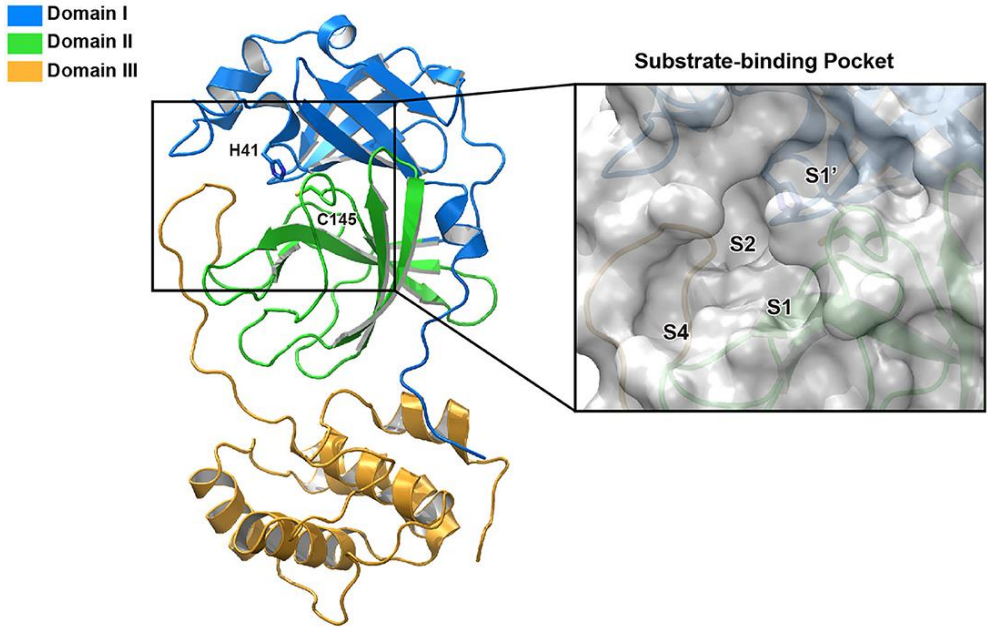
How do drugs modulate protein function?

Druggability

- A **druggable protein** can be functionally modified by a drug
- Has a 3D shape that is amenable to drug binding and action



SARS CoV2 protease binding pocket



Drug-relevant protein classes

- Protein classes relevant to drug discovery (~80% of drugs target these):
- Enzymes
  - Involved in doing “work” in cells, such as converting one chemical into another chemical
- Receptors
  - Involved in cellular communication, often found and cell surface and react to small molecules outside the cell
    - G protein-coupled receptors
    - Nuclear receptors
    - Ligand-gated ion channels

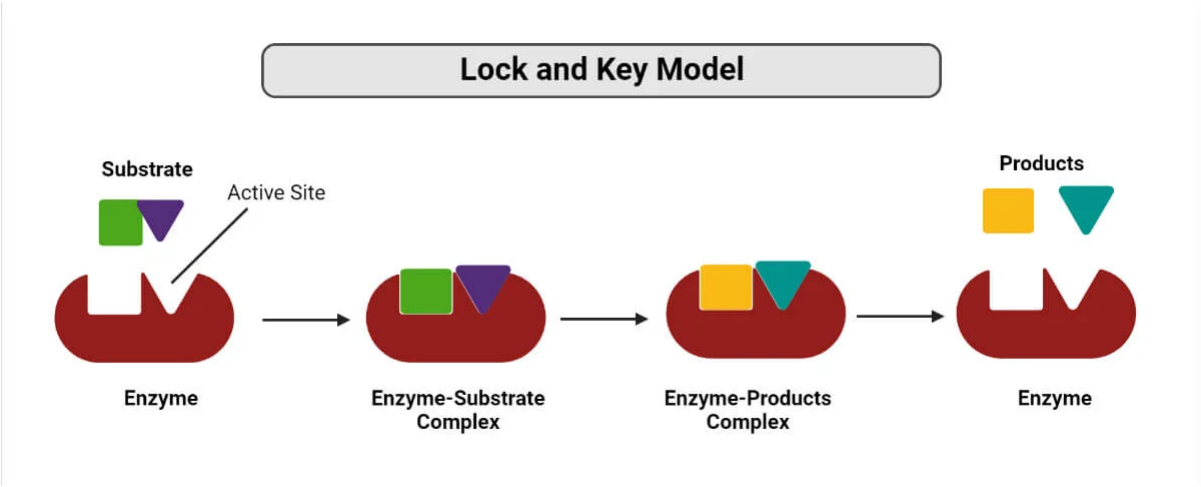
Drugs act on specific **functional states** of proteins

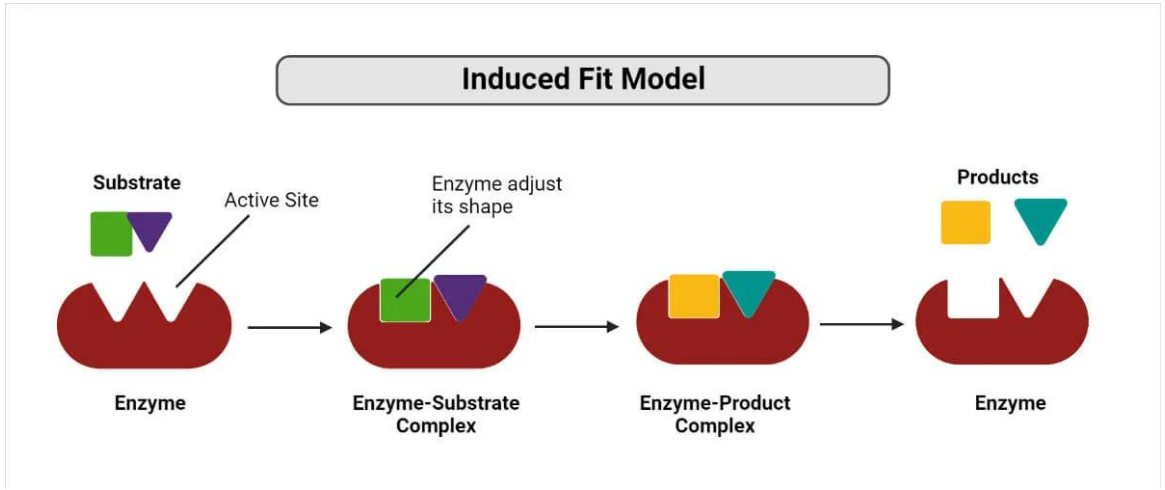
Modulating protein functions with **drugs**

- Turn **on** protein functions
  - (activation / agonist)
- Turn **off** protein functions
  - (inhibitor / antagonist)
- Change or modulate a protein function
  - Allosteric modulators
    - Positive
    - Negative

Modulating protein function: Enzymes

- Enzymes ~ half of all current drug targets
- Especially pathogen enzymes (virus, bacteria)





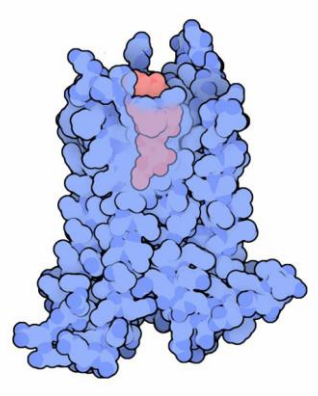
E-S complex: Transitional state

Images: BioRender

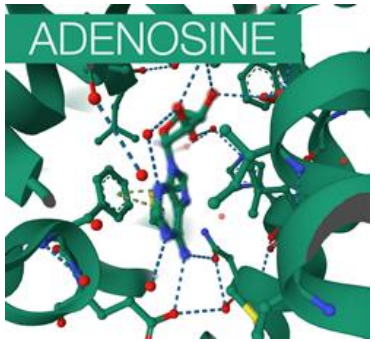
Modulating protein function: receptors

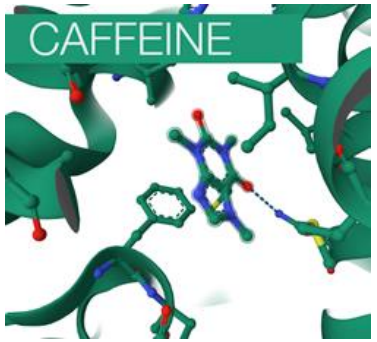
- Receptors: ~ half of all current drug targets
- Typically membrane proteins
- Bind to endogenous small molecules (“ligands”)
  - Eg: insulin receptor, estrogen receptor
- Binding → shape change (conformation) of protein
  - Initiates a chain reaction of biochemical events inside the cell
- Drug (or ligand) binding is generally reversible
- thus the concentration of the ligand or drug impacts the protein function

Example of receptor modulation



- Adenosine receptor – involved in sleepiness signals
- Ligand: adenosine (agonist)
- Agonists “turn on” receptors
- Caffeine is an **Antagonist** at the adenosine receptor
  - Blocks adenosine from acting on receptor
    - → less sleepiness
- https://pdb101.rcsb.org/learn/videos/caffeine-and-adenosine-antagonist-and-agonist





Drug-relevant protein classes

- Ion channels
  - move ions across membranes
- Transporters
  - Move small molecules across membranes

- Different protein classes have different structural constraints
  - Affects our interest in protein structures

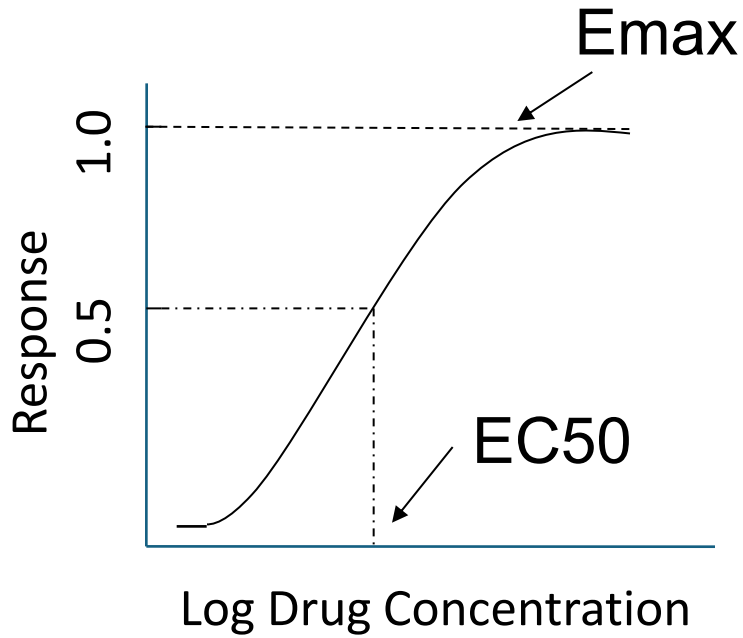
Shapes of dose-response curves

Emax – the maximal response achieved by an agonist

- referred to as *efficacy*

EC50 – drug concentration at 50% of Emax

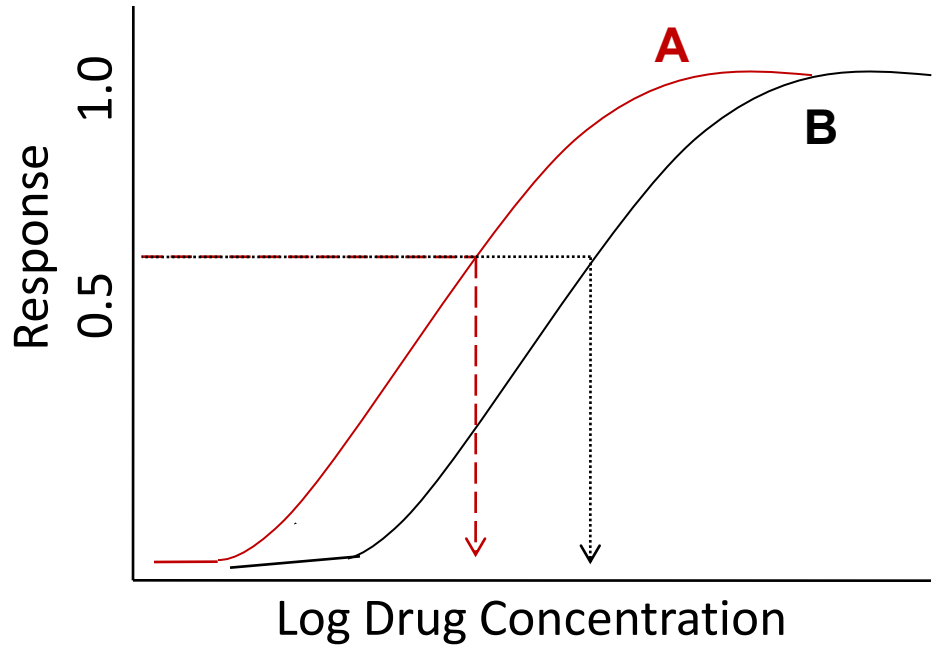
- referred to as *potency*



Example of “response” could be rise in intracellular signal due to receptor activation

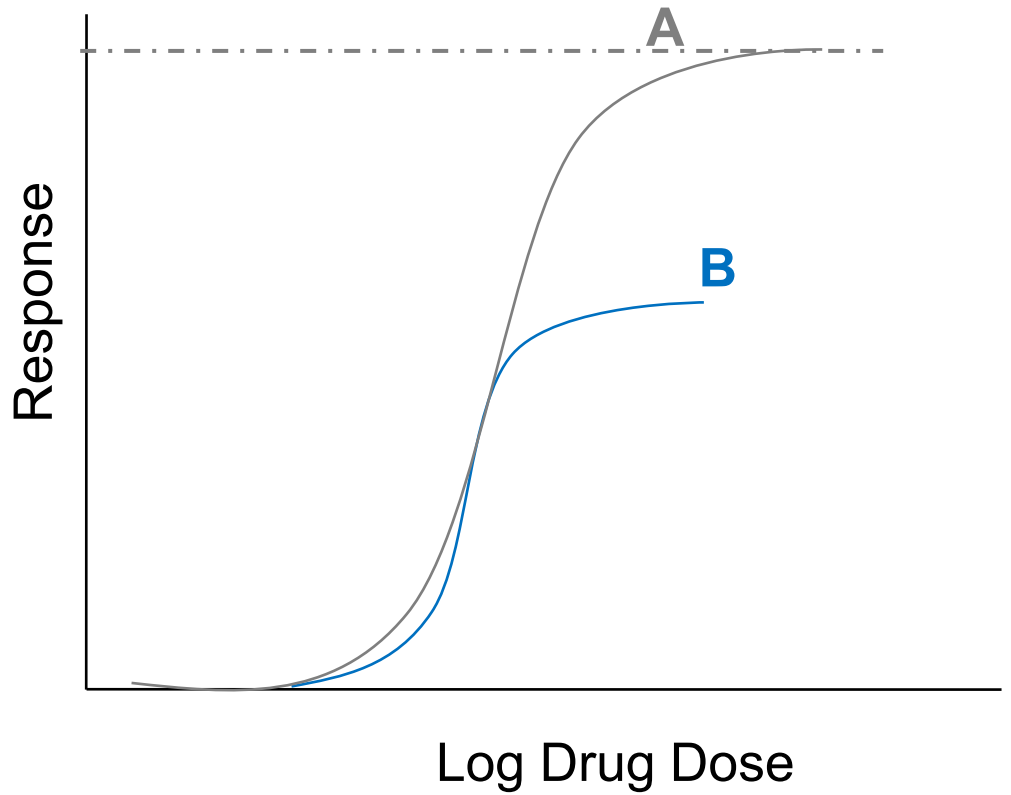
Potency

- Drug A is more potent than drug B
- In drug discovery, we are often seeking potent chemicals



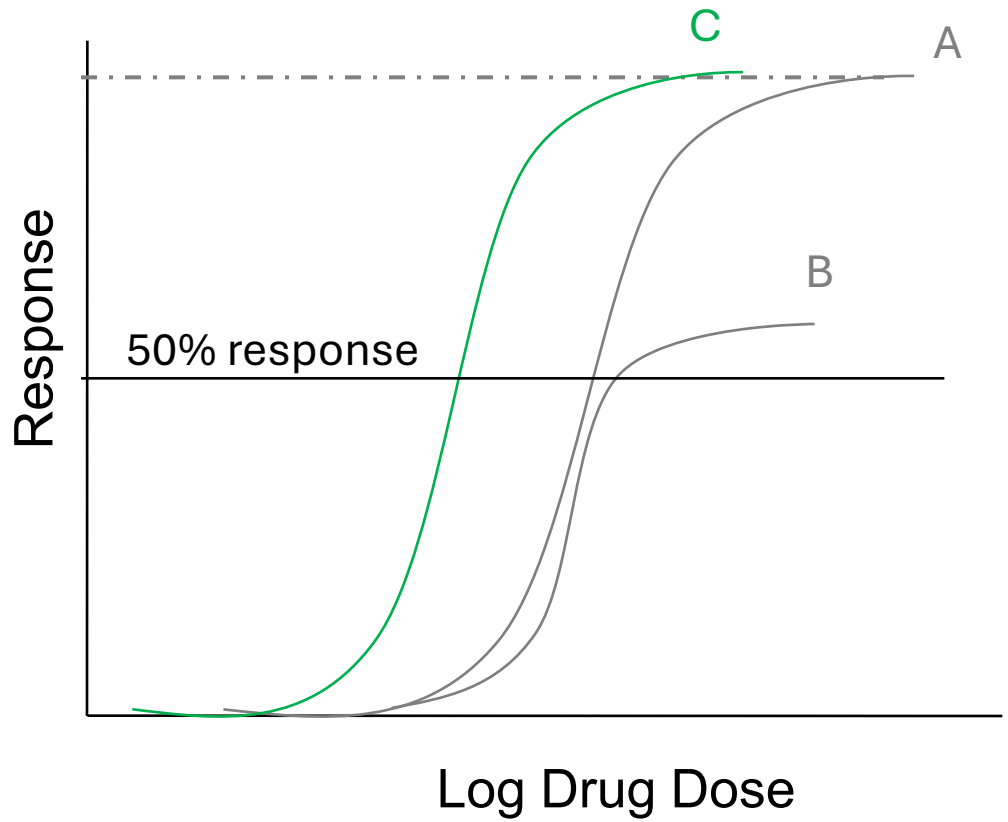
Agonists (turning receptors on)

- Full agonist (A)
- Partial Agonist (B)



Positive Allosteric Modulator (PAMs)

- Enhances effect of endogenous ligand
- ***Generally*** binds at site DISTINCT from endogenous ligand binds

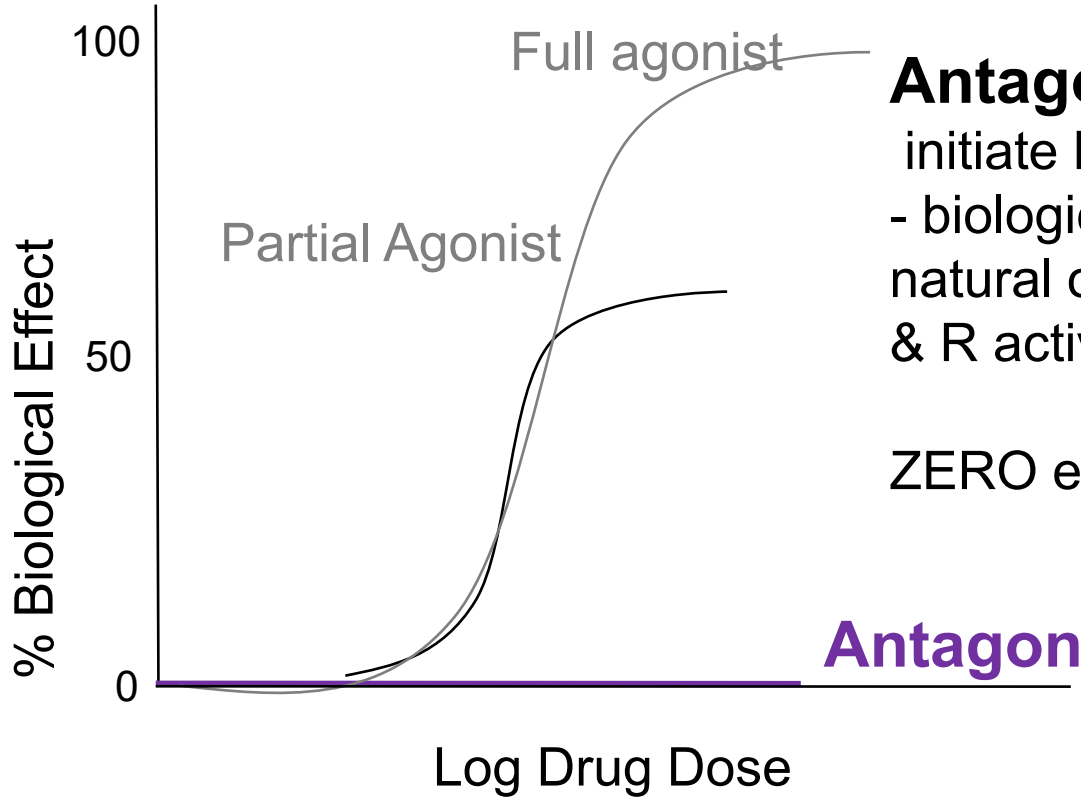


A= Full agonists - produce a full response when all Rs bound

B= Partial agonists – will **NOT** produce a full response EVEN when all Rs bound

C= Positive Allosteric Modulator– will shift agonist dose-response curve left

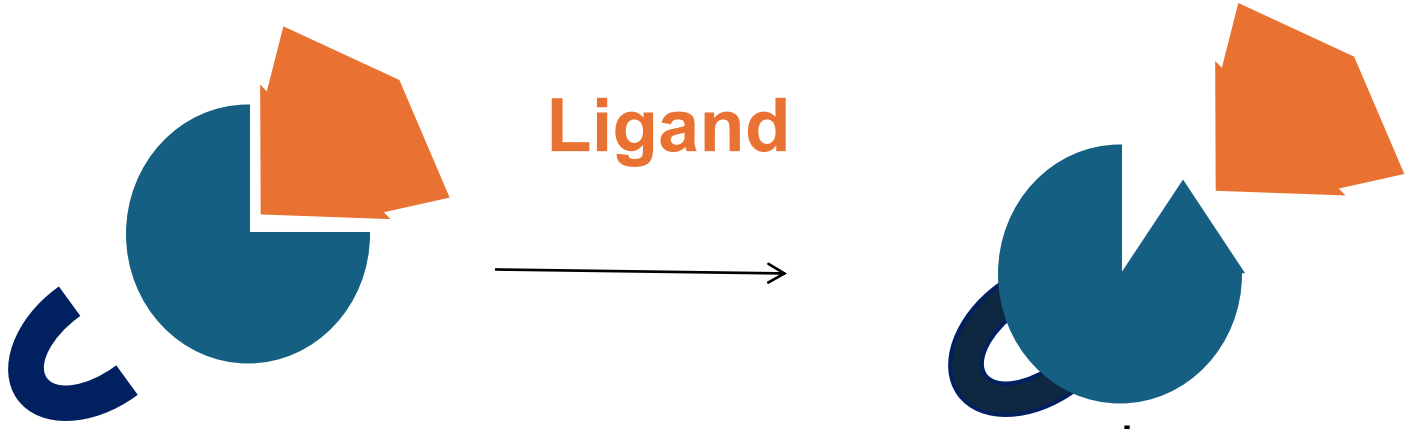
Antagonist (turning receptors off)



Antagonists- binds R but does not initiate biological activity/signaling cascade
- biological effects from "***preventing***" natural or endogenous agonist from binding & R activation

ZERO efficacy on its own

Negative Allosteric Modulators (NAMs)

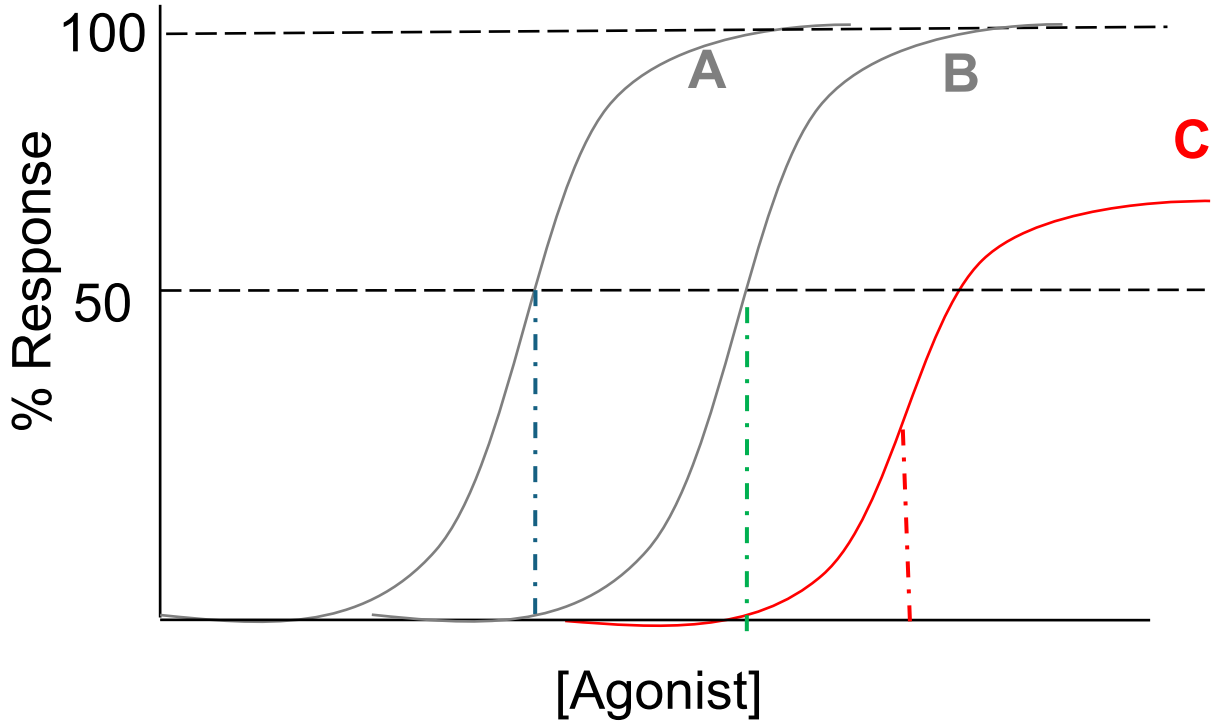


Negative Allosteric Modulator
~ non- competitive Ant

Signalling
Biological response

Antagonists: Non-Competitive Antagonists such as Negative Allosteric Modulators (NAMs)

- May not be at same site on R as agonist
  - **neg. allosteric modulator**
- Affects Ag potency and efficacy
  - (ie shifts D-R curve right and down)



B: Ag + Comp ANT

C: Ag + high conc.
non comp ANT
= reduced potency & efficacy

Balancing agonism and antagonism: new drug for Schizophrenia (2024)

- New drug target: cholinergic system (muscarinic receptors)
- Drug contains two chemicals
- Agonist at muscarinic receptors in brain and periphery
  - BUT side effects due to receptors in periphery
- Thus, added second chemical, a receptor antagonist that stays in periphery and doesn't cross blood-brain barrier
- Result: efficacy with minimized side effects

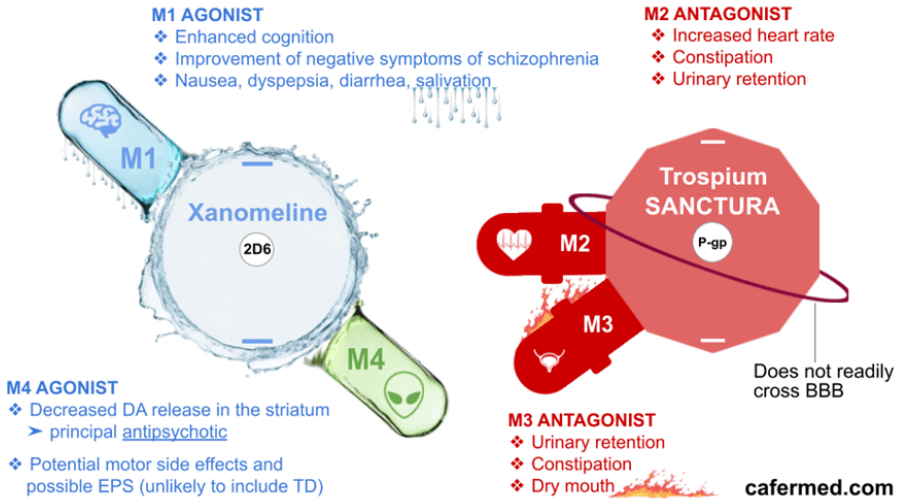
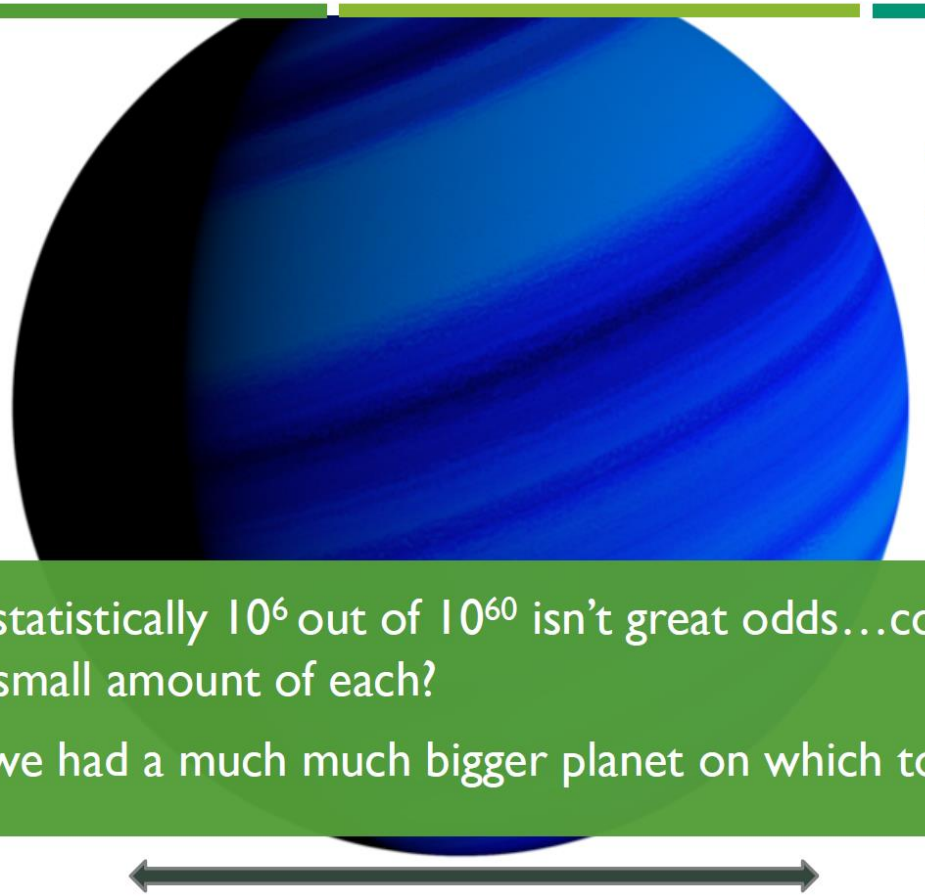


Image: hyperlab.hits.ai

Finding drugs to modulate proteins: searching in chemical space

So if we test 1 million compounds then we are bound to find a "hit" aren't we…?

Well, "drug-like" chemical space is very large…well, there may be 10^{60} compounds that *might* fulfil the requirements of "the impossible journey"



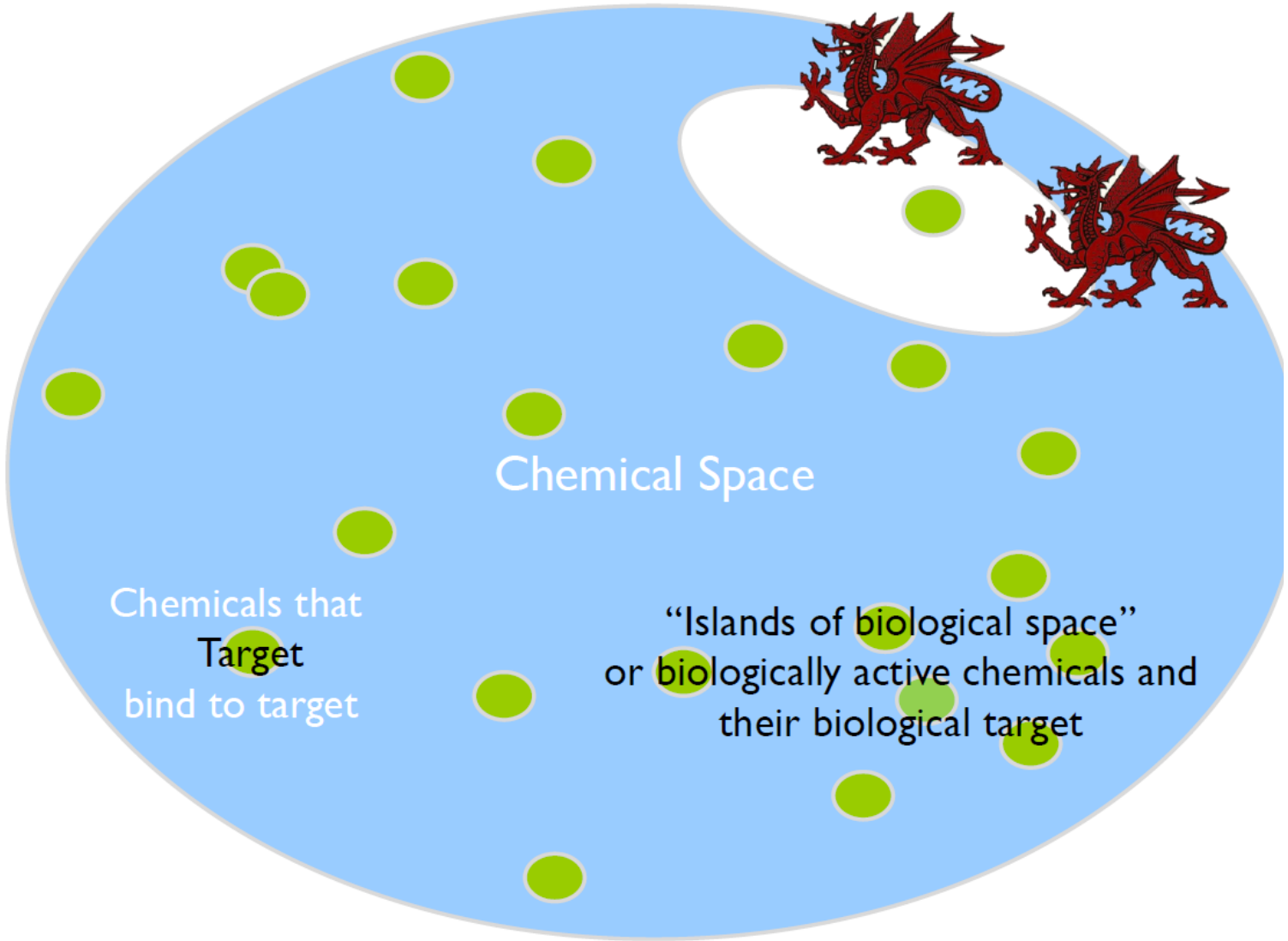
40 BILLION MILES …
AND THAT IS JUST FOR
1 MOLECULE OF EACH!

OK, so statistically 10^6 out of 10^{60} isn't great odds…couldn't we just make a really small amount of each?

Only if we had a much much bigger planet on which to store it!

But while drug-like chemical space can be regarded as a vast ocean (most of it unexplored), biological space can be regarded as tiny islands scattered throughout this ocean

…what are the chances of finding these islands?

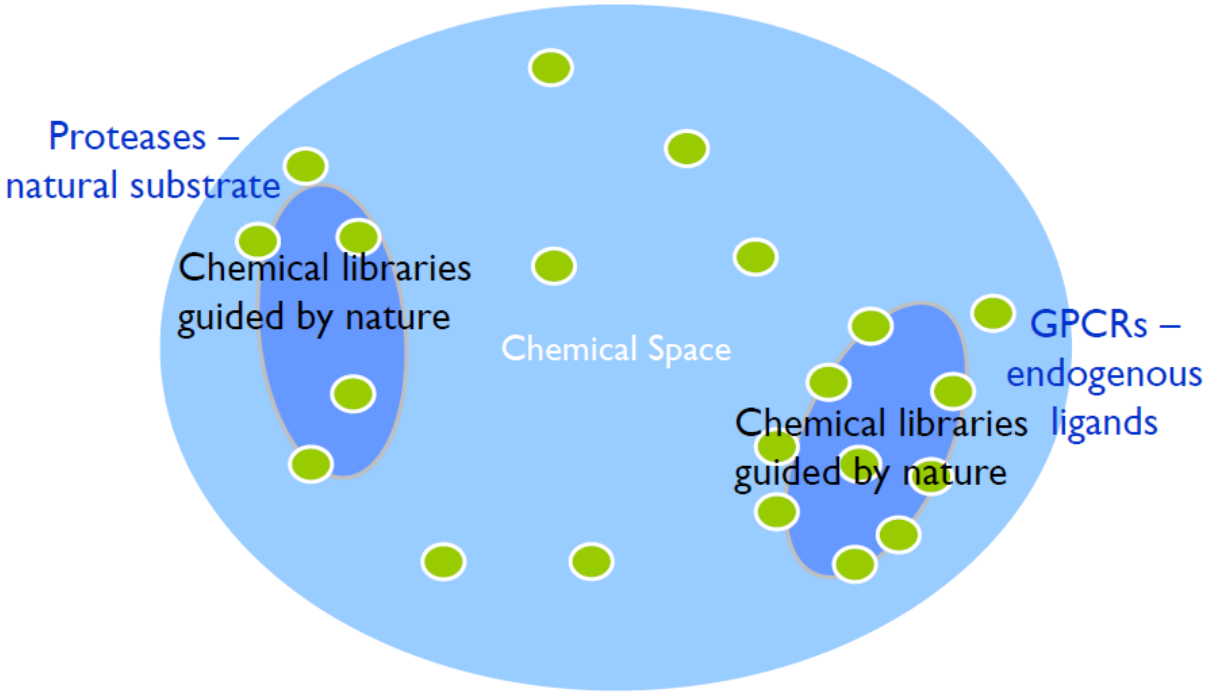


Chemical space and chemical libraries

…what are the chances of finding these islands?

…better if we have an idea of where to look.

Our libraries are based on the sort of thing that tends to work

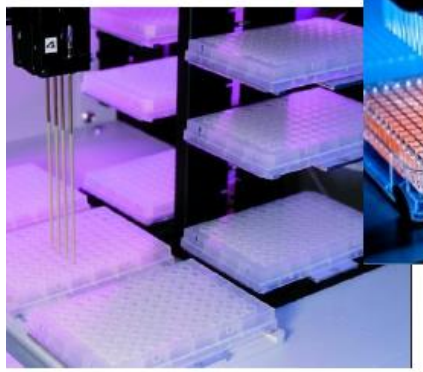


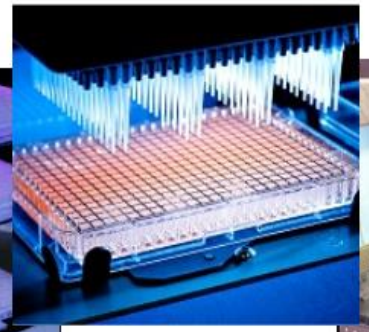
"The most fruitful basis for the discovery of a new drug is to start with an old drug"

Finding the overlap in chemical space and protein target space

Experimentally: High-throughput screening
Modeling: in silico HTS, other approaches

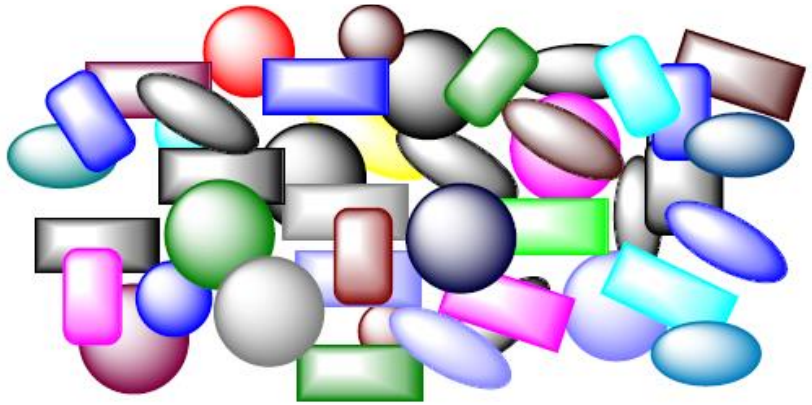
Chemical libraries (500 – 2M compounds) and robotic High-Throughput Screening





http://www.csrs.riken.jp/en/labs/cbuddp/images/fig1.jpgww.drugdiscovery.uc.edu/Images/CompoundLibrary.png

Opportunities in AI: interrogate larger or “better” slice of chemical space



Test 1 million compounds and you are bound to find one which does what you want…aren’t you?

Not necessarily .The haystack may be empty of needles

At present, only 10^8 compounds out of a possible 10^{60} exist, statistically no guarantee of finding something which works

Scale and **complexity** of the protein world

- **~20,000 protein-coding genes** in the human genome
- But “Protein” is not a single **class label**
- One gene ≠ one structure ≠ one function
- Relevant for drug discovery:
- **~100,000–1,000,000 distinct human protein species** (“human proteome”)
    - Alternative splicing (covered in section 3 of course)
    - Post-translational modifications (**PTM**s; additional “tags” on proteins like sugar, phosphate) - EXAMPLE
    - Proteolytic processing (cell cuts a larger into…
    - Context-specific expression

How many proteins are in a single human cell?

- ~10,000–12,000 different proteins expressed in a typical human cell
- **~1–3 billion** total protein molecules per cell
- Distribution is *extremely* skewed:
- Top ~100 proteins = majority of mass
- Thousands of proteins present at <100 copies/cell

- **ML analogy:**
- A massive long-tailed distribution with severe class imbalance

How many **residues** actually **matter**?

- Typical protein: 300–500 amino acids
- Binding site:
  - ~10–30 key residues
- Allosteric networks:
  - Often distributed across tens of residues

- **ML analogy**
  - High-dimensional input
  - Sparse, structured signal
  - Strong epistasis (nonlinear interactions)

Protein Data Bank (PDB)

- Database of 3-D structural information about proteins and nucleic acids
- Experimental measurements (cryo-EM, crystallography)
- More recently, also contains predicted structures (AlphaFold)
- Covered in Module 1 Collab

How many **conformations** does a protein have?

- Proteins exist as **ensembles**, not single structures
- Even “rigid” proteins sample dozens to hundreds of low-energy conformations
- Flexible proteins:
  - **Thousands+** of accessible states

PDB

- Usually captures **1–3 conformations**
- Often biased by:
  - Crystallization conditions
  - Ligands
  - Stabilizing mutations

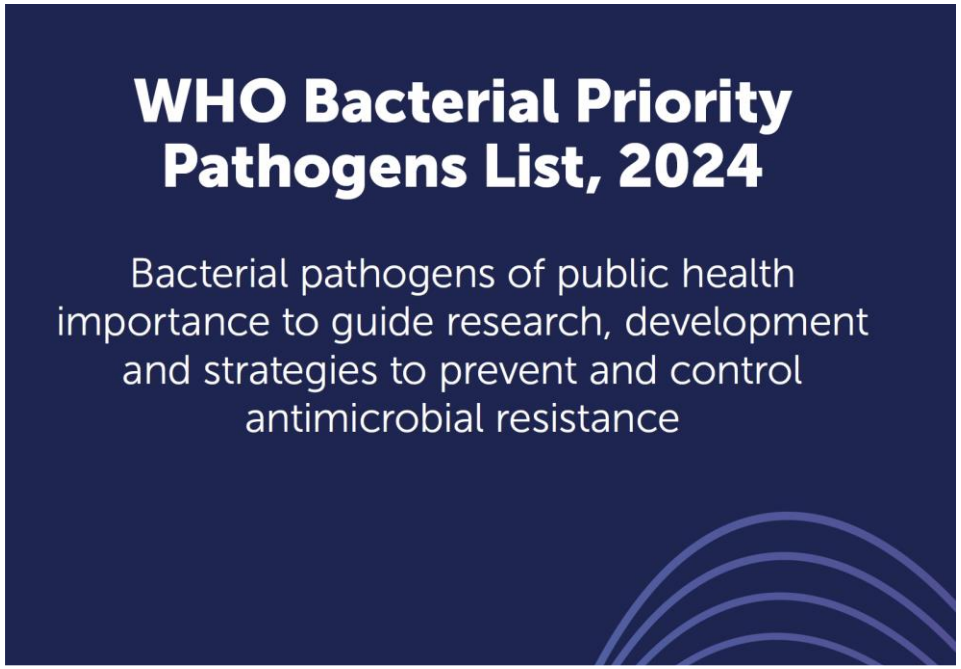
AlphaFold

- Predicts **one dominant fold**
- Does *not* represent:
  - Functional ensembles
  - Ligand-induced states
  - Signaling bias

- **ML takeaway:**
- You’re training on a projection of a high-dimensional distribution.

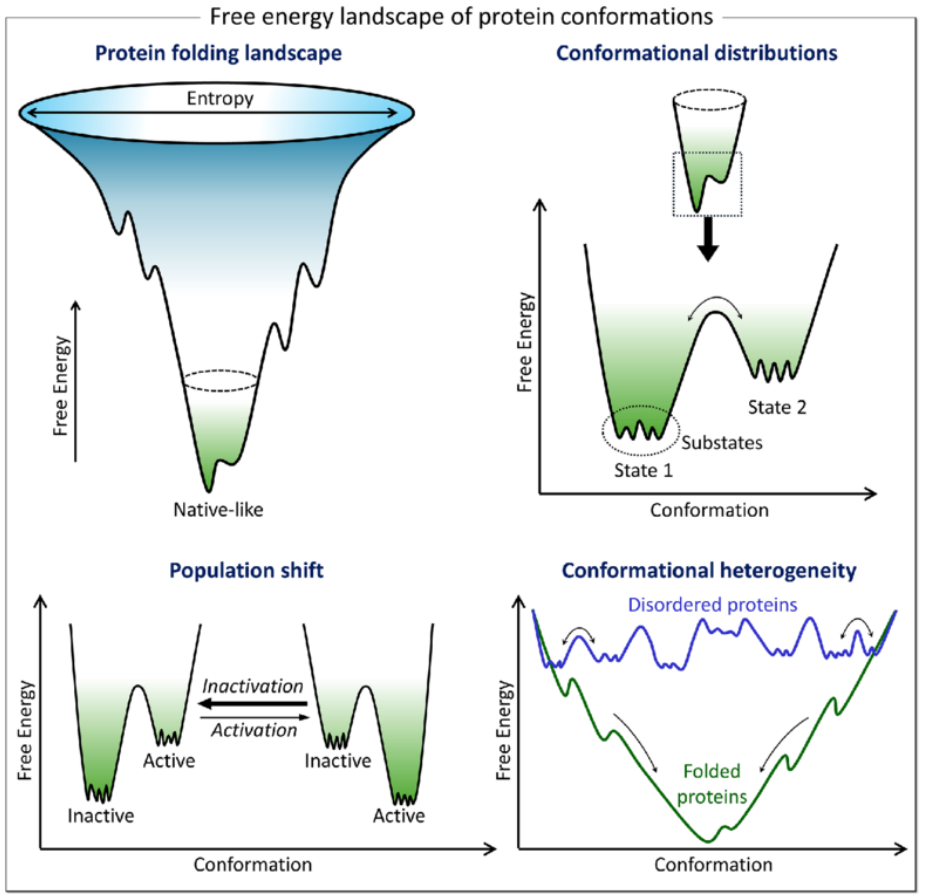
Pathogens (viruses, bacteria, etc)

- **Human proteins:**
  ~20,000 genes
  ~10^5–10^6 protein species
  ~10^5 interactions per cell
- **Pathogen proteins:**
  ~10^4–10^5 medically relevant proteins
  Low redundancy
  High essentiality
- **Conclusion:**
  Pathogen proteomes are smaller, sharper, and more learnable.



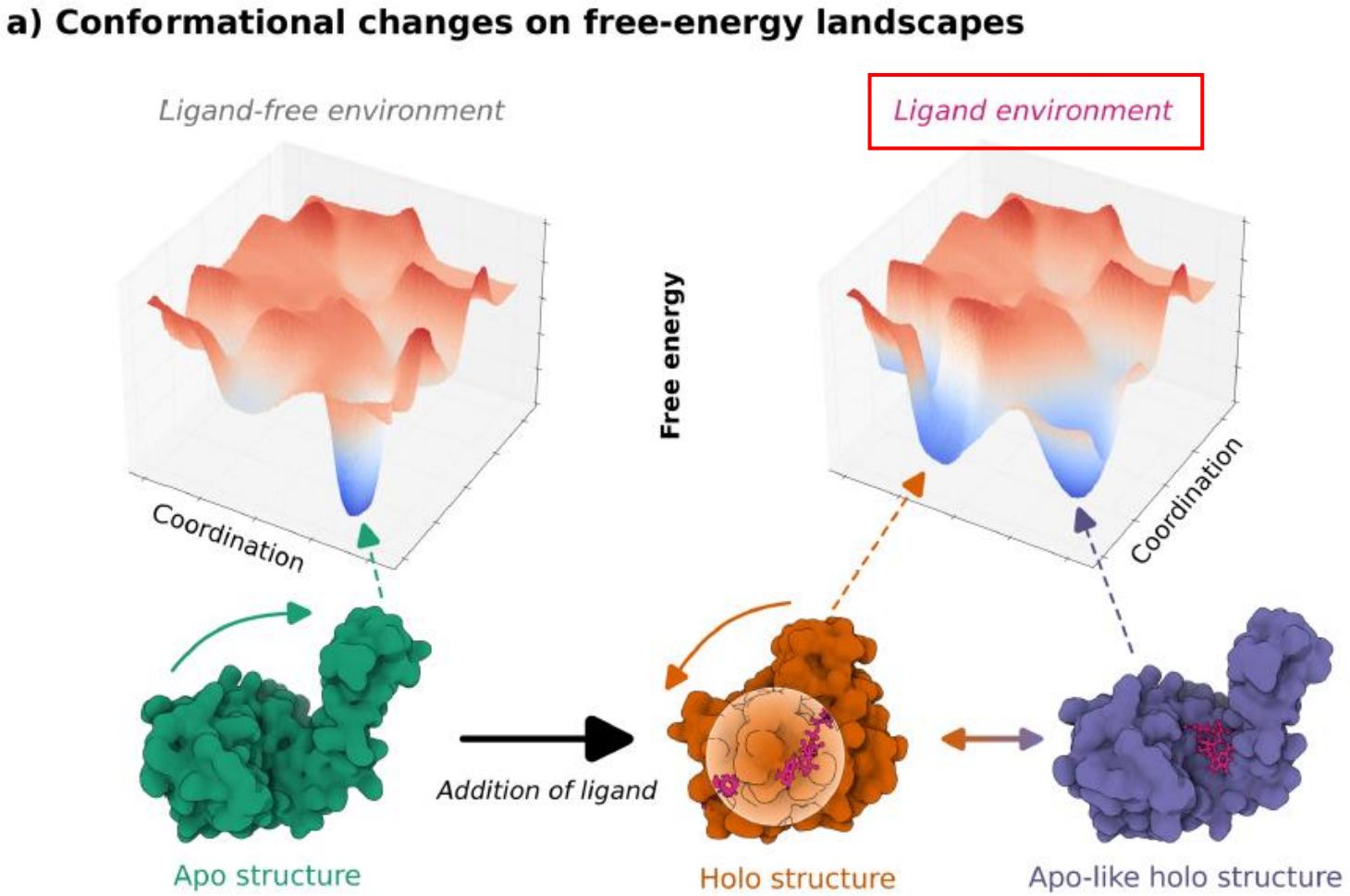
Modern structural biology: Focus on **Ensembles**

- Variety of structures for any given protein
- Ensembles are the sum or set of all of these variations
- Complexity = conformational heterogeneity



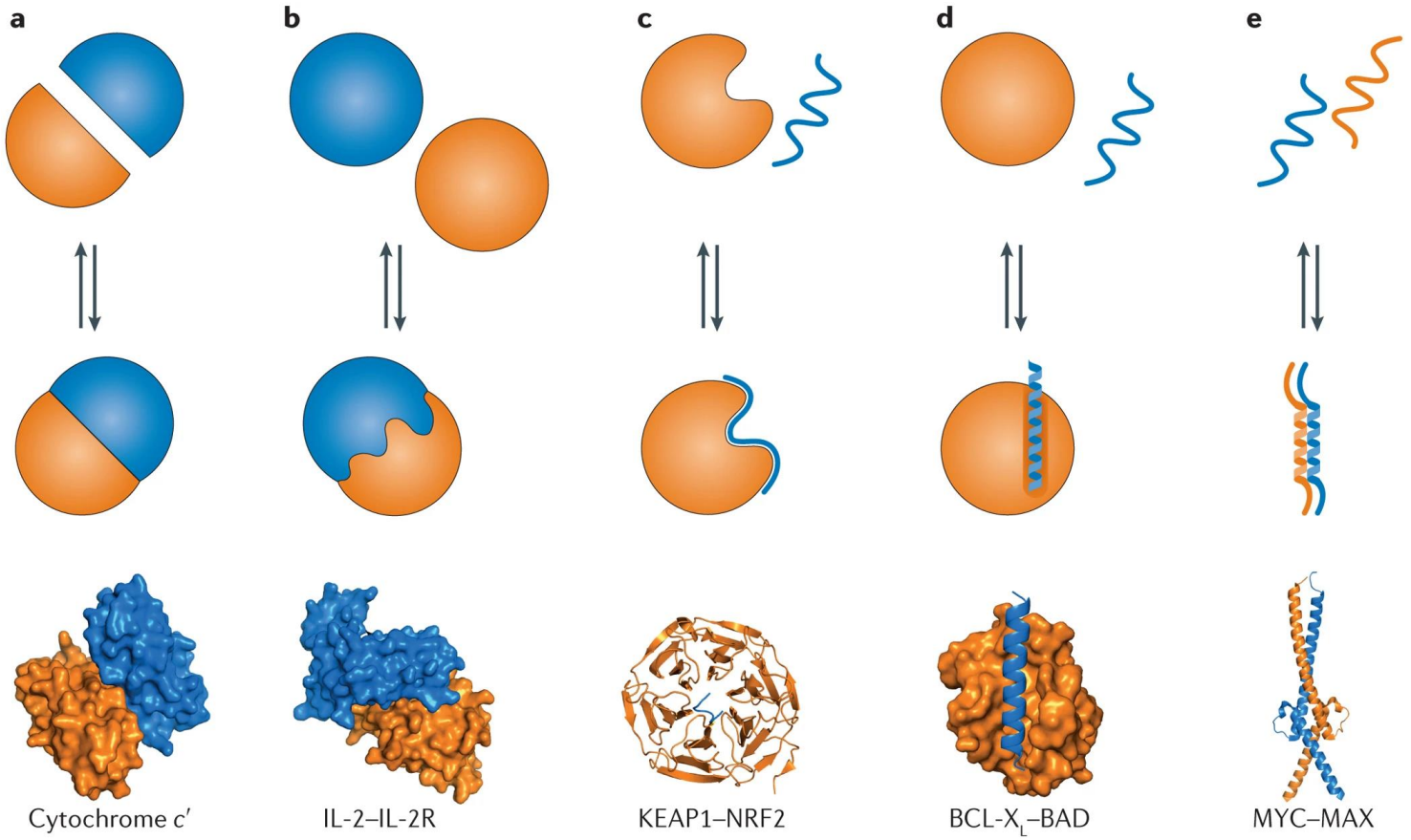
Nussinov et al., Biophysical Reviews 2025

Small molecules (ligands, drugs) can change the ensembles and likelihood of protein conformations



Ellaway et al., 2024 Protein Dynamics

Opportunities: AI for protein-protein interactions



Nature Reviews | Drug Discovery

Beyond “undruggable”: focus on **protein–protein interactions** (traditionally considered “undruggable”)

- The “interactome”
- Estimated **300,000 – 1,000,000 possible PPIs** in humans
- At any moment in a given cell:
  - ~50,000–200,000 active interactions
- Most PPIs are:
  - Transient
  - Context-dependent
  - Cell-type-specific
  - Condition-specific

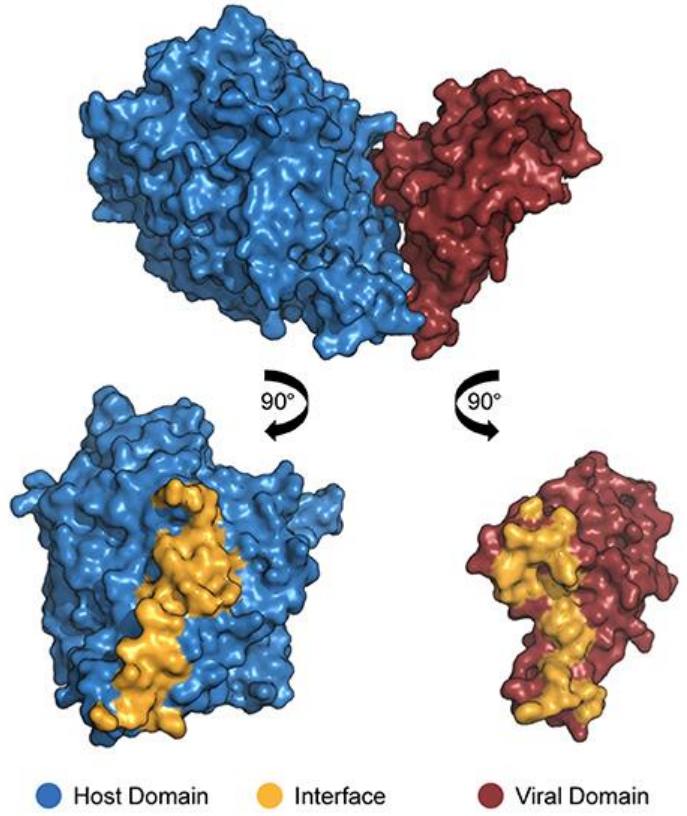
- **ML takeaway:**
- You are never modeling “the” interaction network—only a conditional slice.

How many **interaction interfaces** does one protein have?

- Many proteins have multiple, partially overlapping interfaces
- Interfaces can be:
  - Mutually exclusive
  - Allosterically coupled
  - State-dependent
- Typical numbers:
  - 1–5 distinct functional interfaces per protein
  - Often not present simultaneously

Implications for modeling:

- A structure labeled “protein A” may have:
  - Multiple tasks
  - Multiple binding modes



Case study 1: Cryptic pockets

nature chemistry

ARTICLES

https://doi.org/10.1038/s41557-021-00707-0

Check for updates

SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome

Maxwell I. Zimmerman[1,2], Justin R. Porter[1,2], Michael D. Ward[1,2], Sukrit Singh[1,2], Neha Vithani[1,2], Artur Meller[1,2], Upasana L. Mallimadugula[1,2], Catherine E. Kuhn[1,2], Jonathan H. Borowsky[1,2], Rafal P. Wiewiora[3,4], Matthew F. D. Hurley[5], Aoife M. Harbison[6], Carl A. Fogarty[6], Joseph E. Coffland[7], Elisa Fadda[6], Vincent A. Voelz[5], John D. Chodera[4] and Gregory R. Bowman[1,2]

SARS-CoV-2 has intricate mechanisms for initiating infection, immune evasion/suppression and replication that depend on the structure and dynamics of its constituent proteins. Many protein structures have been solved, but far less is known about their relevant conformational changes. To address this challenge, over a million citizen scientists banded together through the Folding@home distributed computing project to create the first exascale computer and simulate 0.1 seconds of the viral proteome. Our adaptive sampling simulations predict dramatic opening of the apo spike complex, far beyond that seen experimentally, explaining and predicting the existence of 'cryptic' epitopes. Different spike variants modulate the probabilities of open versus closed structures, balancing receptor binding and immune evasion. We also discover dramatic conformational changes across the proteome, which reveal over 50 'cryptic' pockets that expand targeting options for the design of antivirals. All data and models are freely available online, providing a quantitative structural atlas.

Leveraging computational power to model protein folding

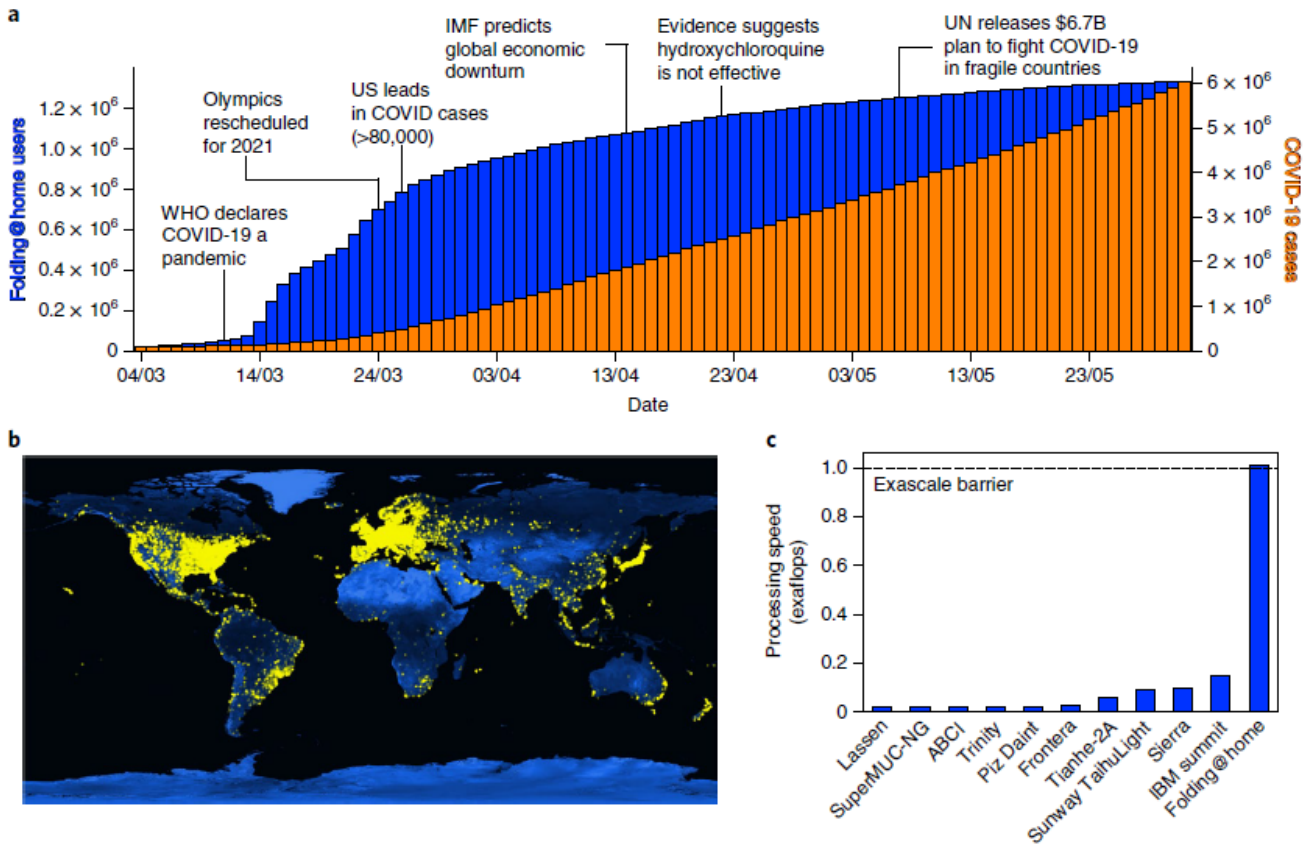
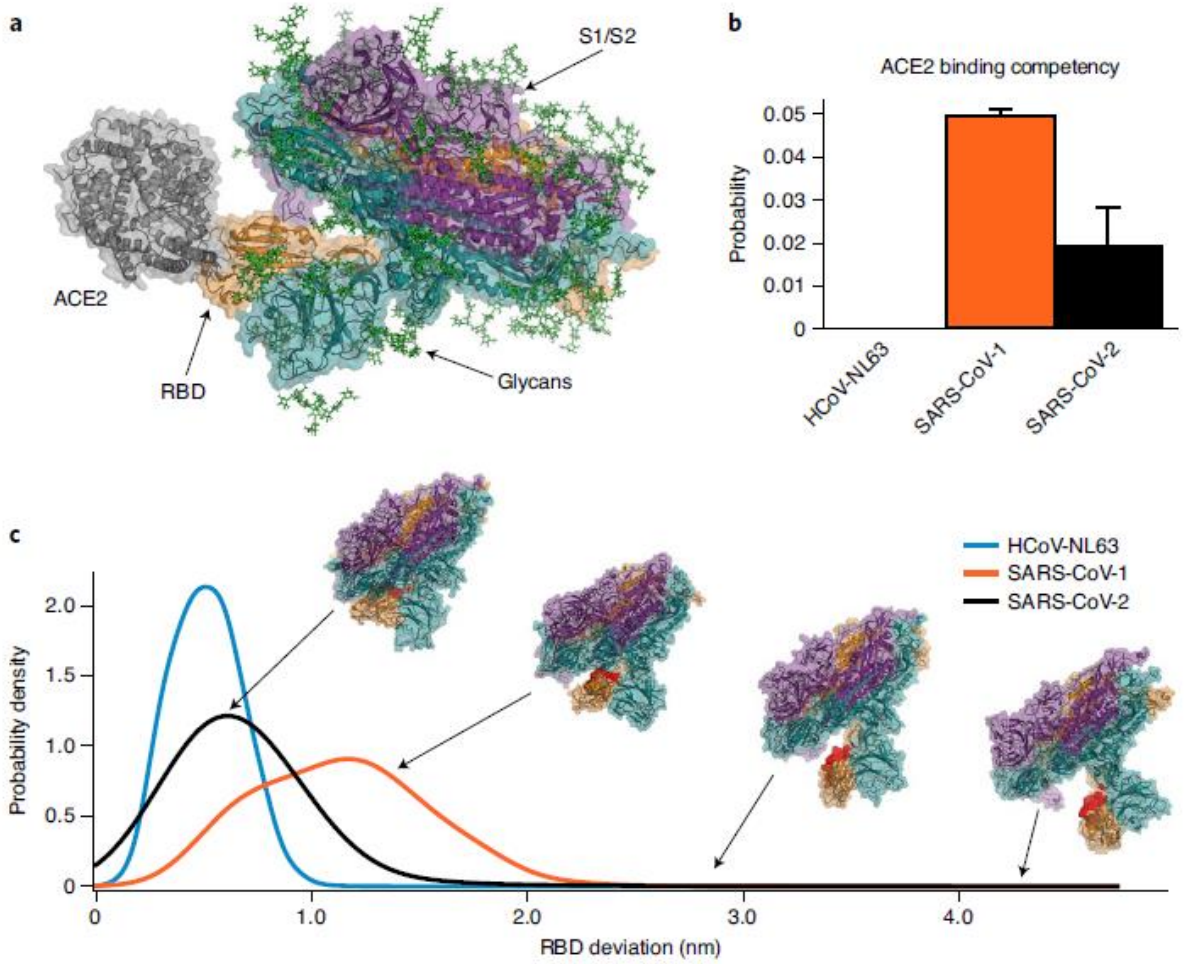


Fig. 1 | Summary of Folding@home's computational power. a, The growth of Folding@home in response to COVID-19. The cumulative number of users is shown in blue and COVID-19 cases are shown in orange. **b**, Global distribution of Folding@home users. Each yellow dot represents a unique IP address contributing to Folding@home. **c**, The processing speed of Folding@home and the next 10 fastest supercomputers in exaflops (one exaflop is 10^{18} floating point operations per second).

Structural characterization of spike opening and conformational masking

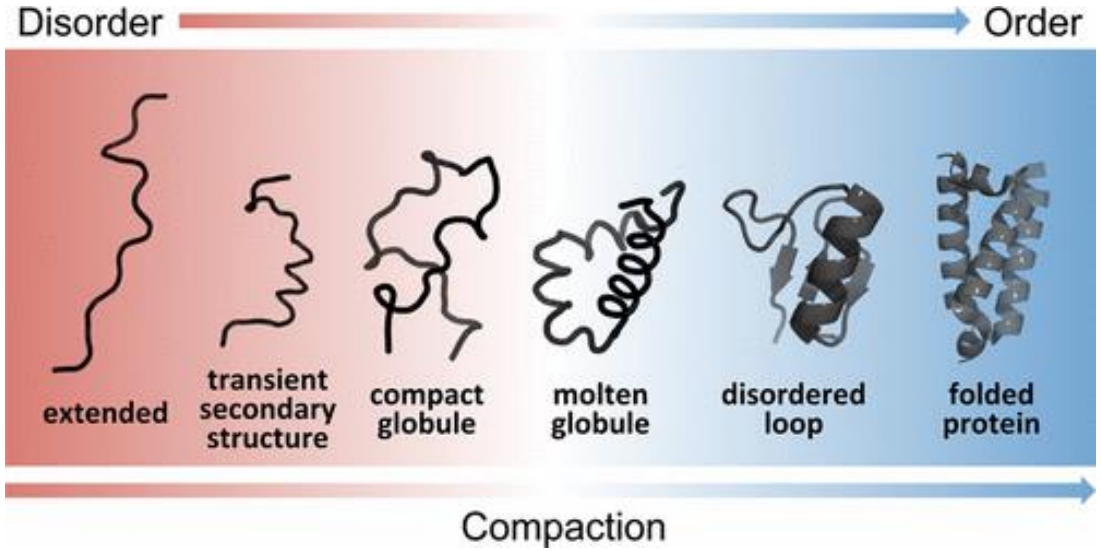
Green: post-translational modifications (glycans)



Note the cryptic **red** region that becomes exposed

Receptor Binding Domain deviation

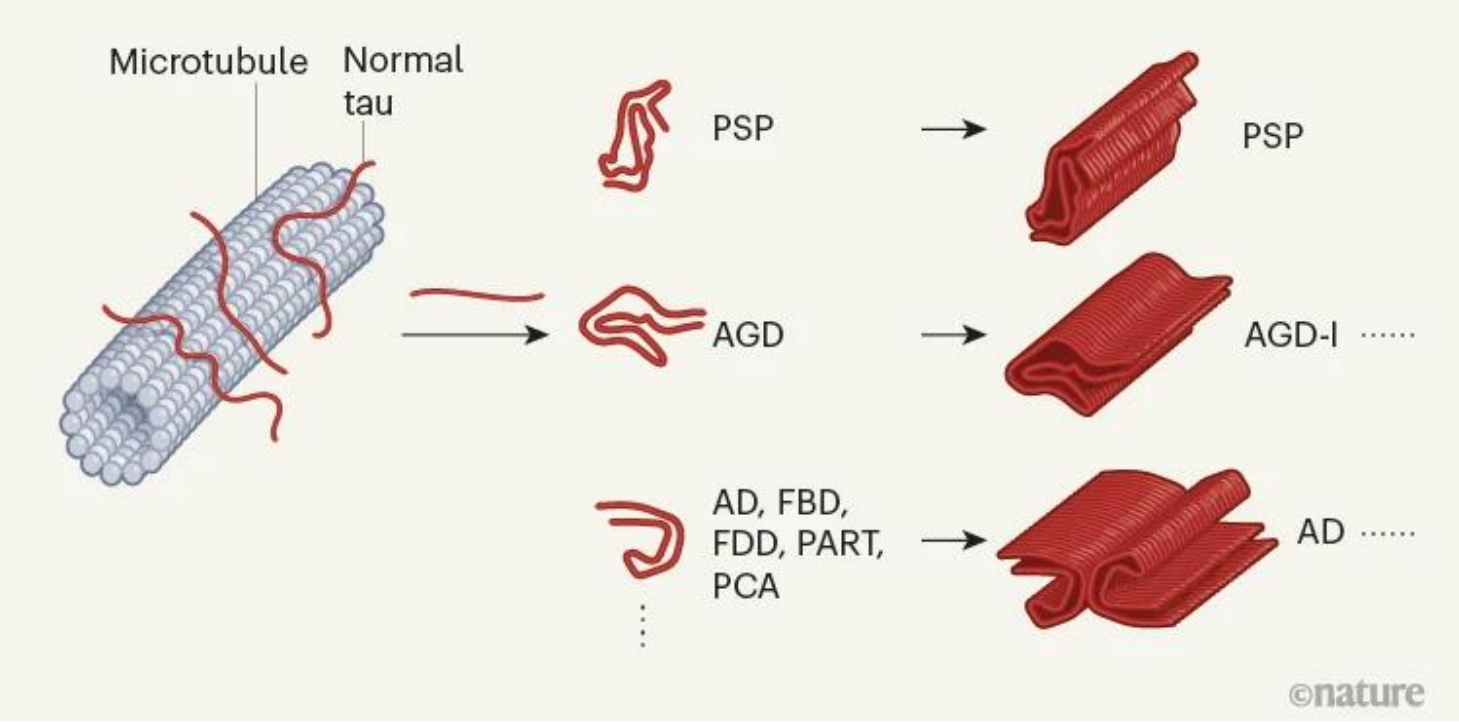
Case Study 2: Intrinsically Disordered Proteins (IDPs)



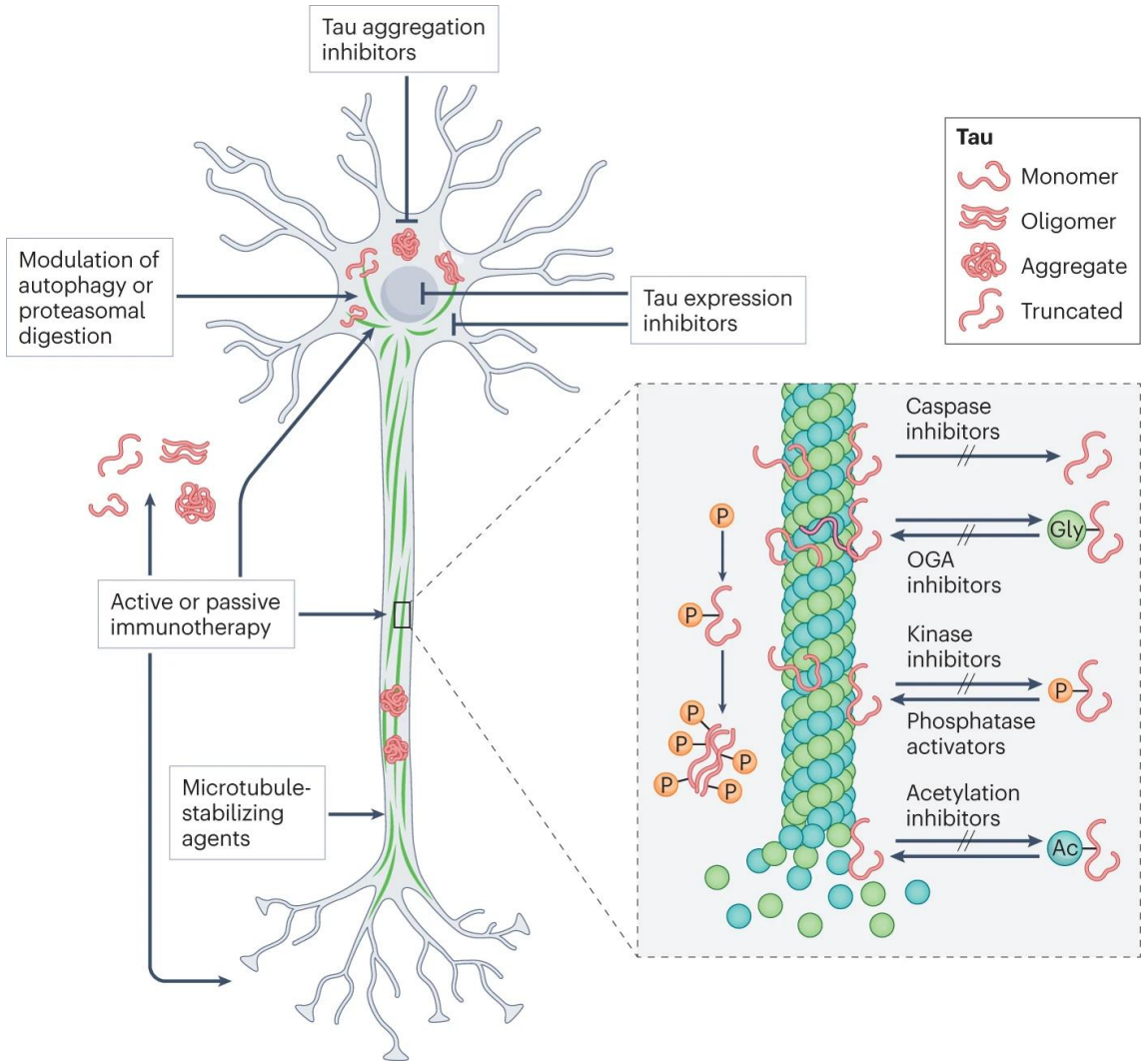
- IDPs are common in biology
  - (IDP regions in ~ 40% of human proteins, ~10% fully IDPs)
- roles in scaffolding and protein-protein interactions
- Ground truth data about IDPs – extremely spare and very hard to acquire
- Misfolding of IDPs is problematic and involved in several diseases
- https://treventis.com/technology/protein-misfolding-the-basics/

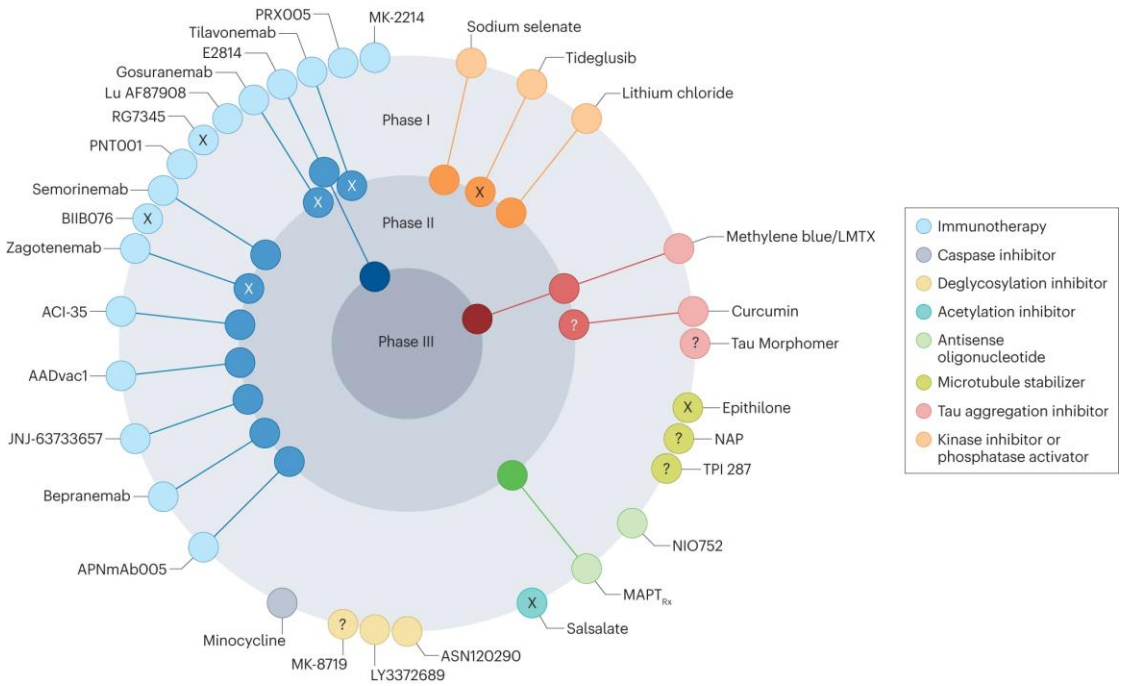
Specific example of Tau, an IDP

- Tau misfolding involved in neurodegeneratives diseases



Drug development pipeline: targeting Tau

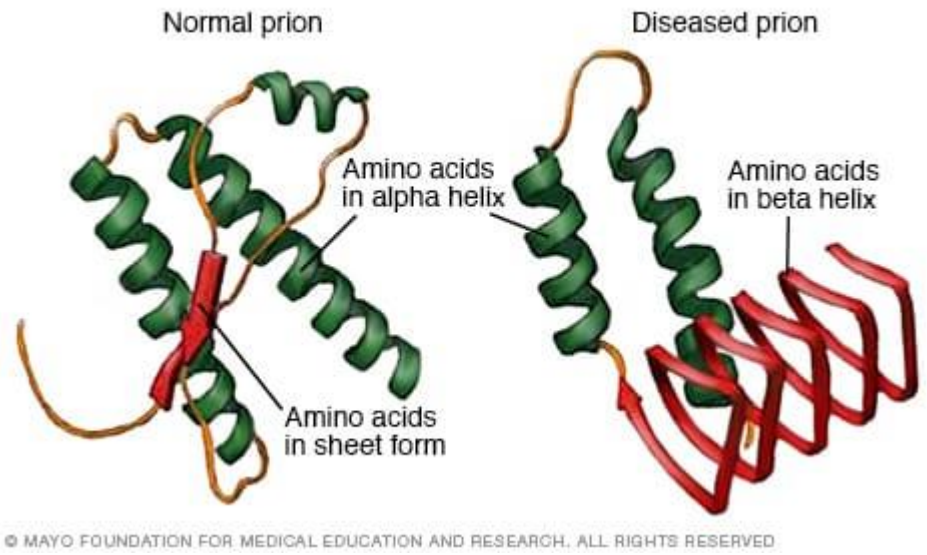




2023 Drug Dev for AD snapshot

Prions: a protein-only infectious disease

- Prions as another IDP example
- The misfolded protein will generate more misfolded protein by helping others to fold
  - Propagation
  - Creutzfeld-Jacob disease (mad cow)



Design of intrinsically disordered region protein binders

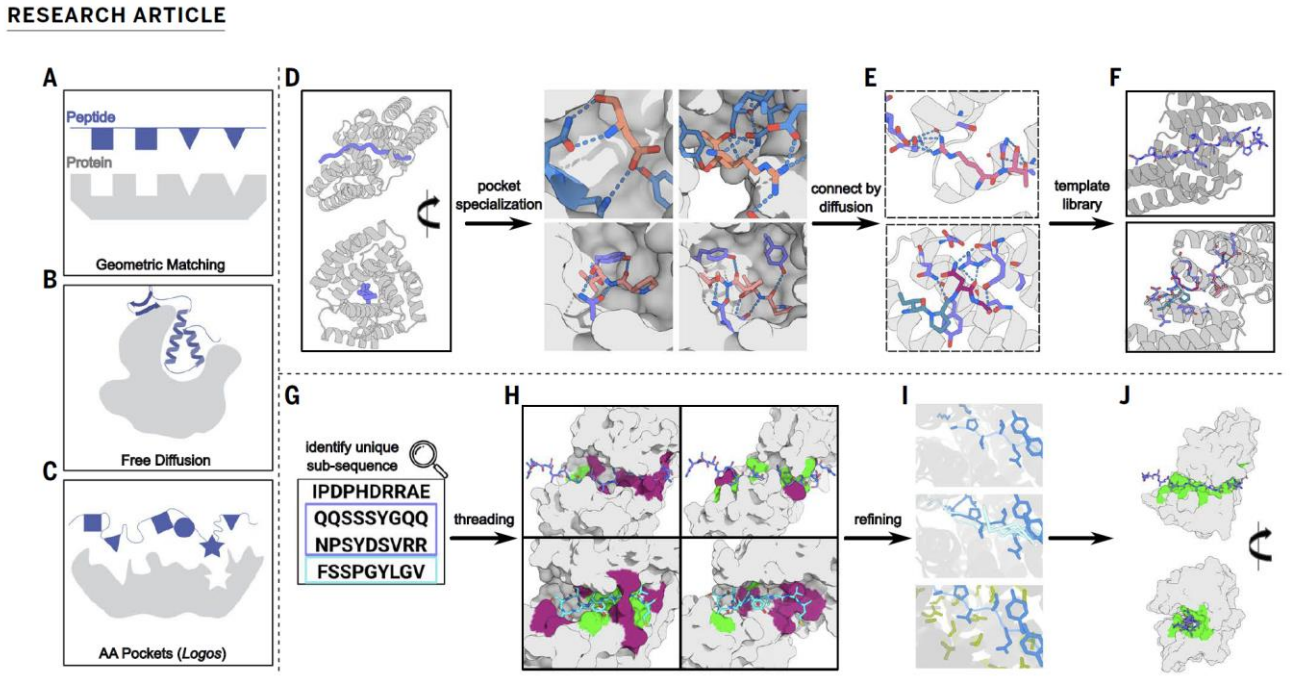


Fig. 1. Overview of IDR binder design protocol. (**A** to **C**) Design methods. (A) Repeat protein–based geometric matching approach requiring one-to-one matching between identically spaced repeat units on designed binder and target peptide. (B) Unconstrained free diffusion approach folds targets into structures frequently observed in the PDB training set, primarily helices but also strands. (C) The amino acid (AA) pockets approach explored here combines the designed pockets and extended scaffolds of the geometric matching approach with the ability of RFdiffusion to recombine and diversify the pockets to achieve more general recognition of nonrepeating sequences. (**D** to **F**) Template library construction. (D) (Left) Designed binder scaffolds wrap around extended peptide backbones, enabling contact with each target amino acid. (Right) Example binding pockets. (E) Binding pockets are connected using RFdiffusion (each peptide window is colored differently, as purple, pink, and blue) into templates for general sequence recognition. (F) Examples of two of the 1000 generated templates. (**G** to **J**) IDR binding pipeline. (G) Unique subsequences (purple and cyan) were identified through a protein sequence database search and (H) threaded through the template library to identify optimal matches between amino acid segments and binder pockets. Pocket matches are green and mismatches are dark red on the protein surface. (I) Matches are refined using "one-sided partial diffusion" (top), where only the binder is changed; "two-sided partial diffusion" (middle), where the target and the binder can be changed; "motif diffusion" (bottom), where key interacting motifs (target, blue; binder, green) are unchanged while the rest are noised, reconnected, diversified, and optimized. (J) Examples of resulting designs. [Panels (A) to (C) were created with BioRender.com]

https://github.com/RosettaCommons/RFdiffusion

Wu et al., Science 2025

Case study 3: where improving the affinity of drugs is not biologically helpful

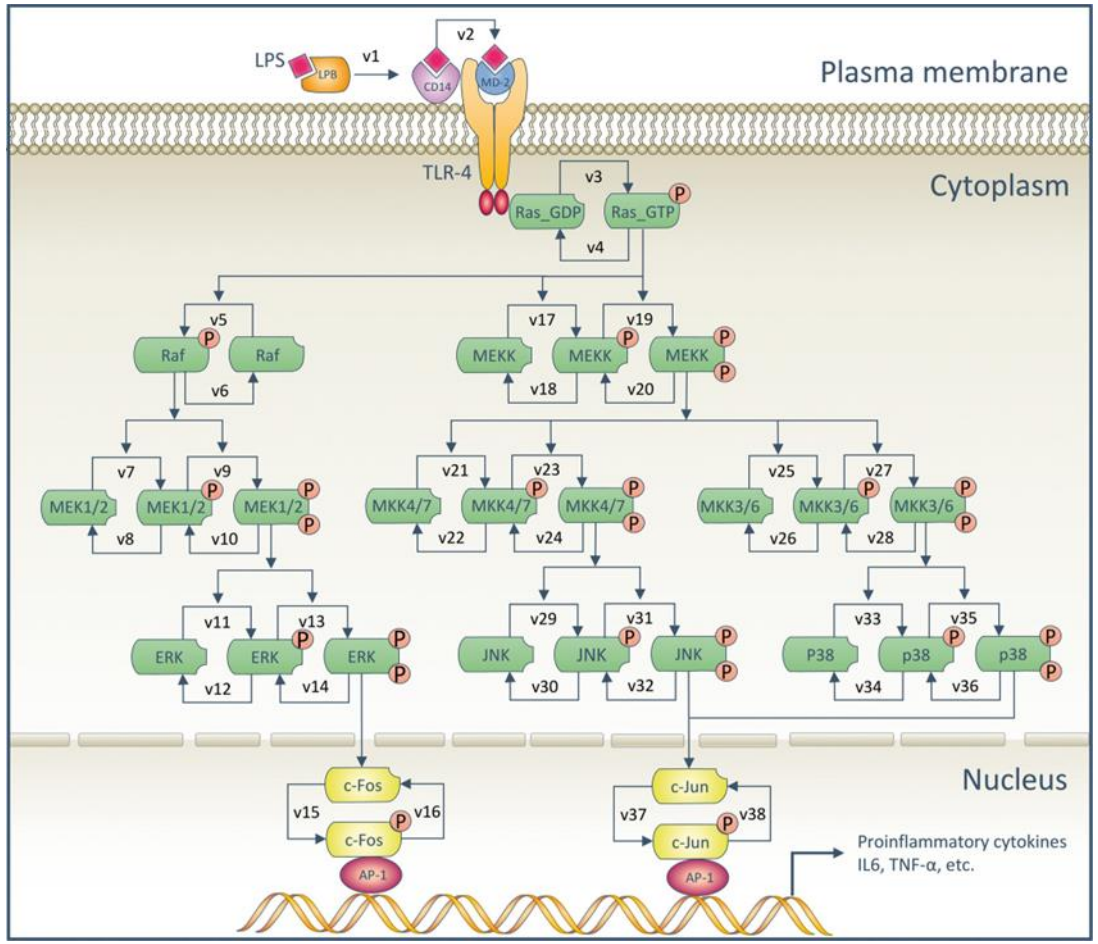
- Potency: how much of the drug is needed to have an effect
- Potency of “average” clinical drug: 20 nM
- Affinity: tightness of binding between drug and protein

Low-affinity interactions are especially important in:

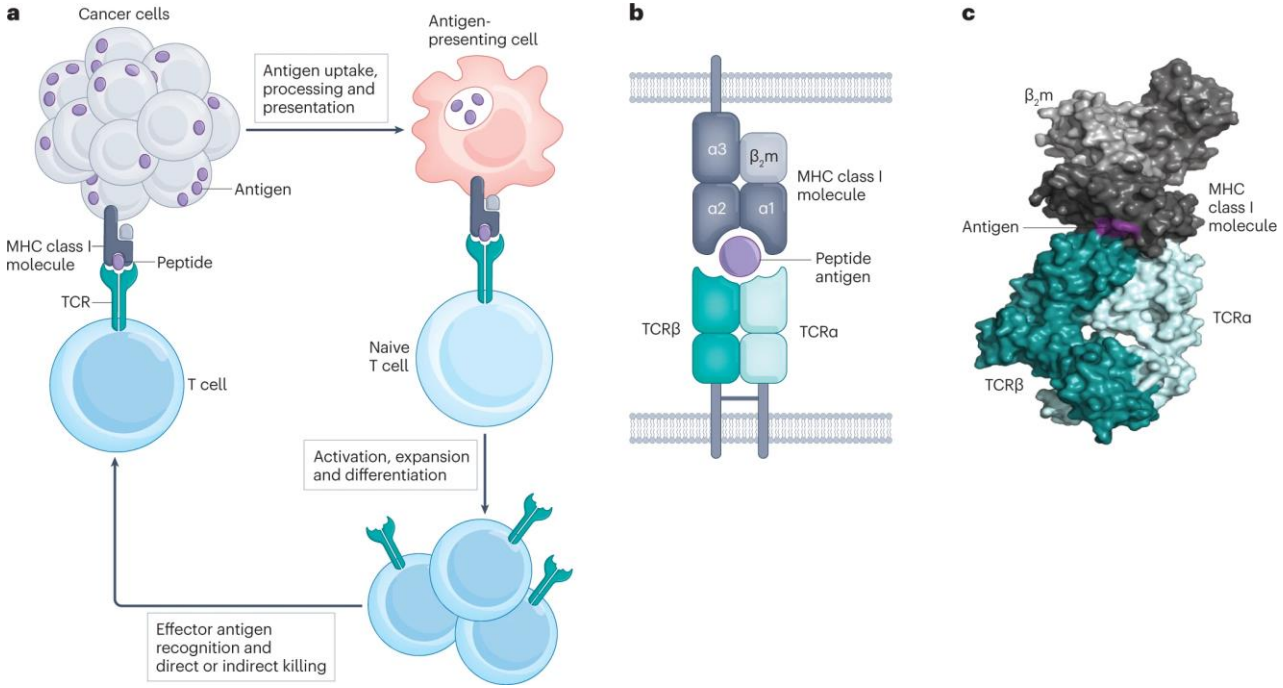
- Cell signaling cascades
- T-cell receptor function

Model of MAPKinase signaling cascade

- Weak binders are essential to amplify the signal
- Here: post-translational modifications (PTMs)
- Goal is a single enzymatic event (kinase activity) can have a cascading downstream effect to amplify the signal



T-cell receptors on the outside of T-cells play important roles for immune system function



T cells need to:

Distinguish self from non-self
Identify pathogens (antigens) and activate their destruction

Require low-affinity interactions to identify “self”

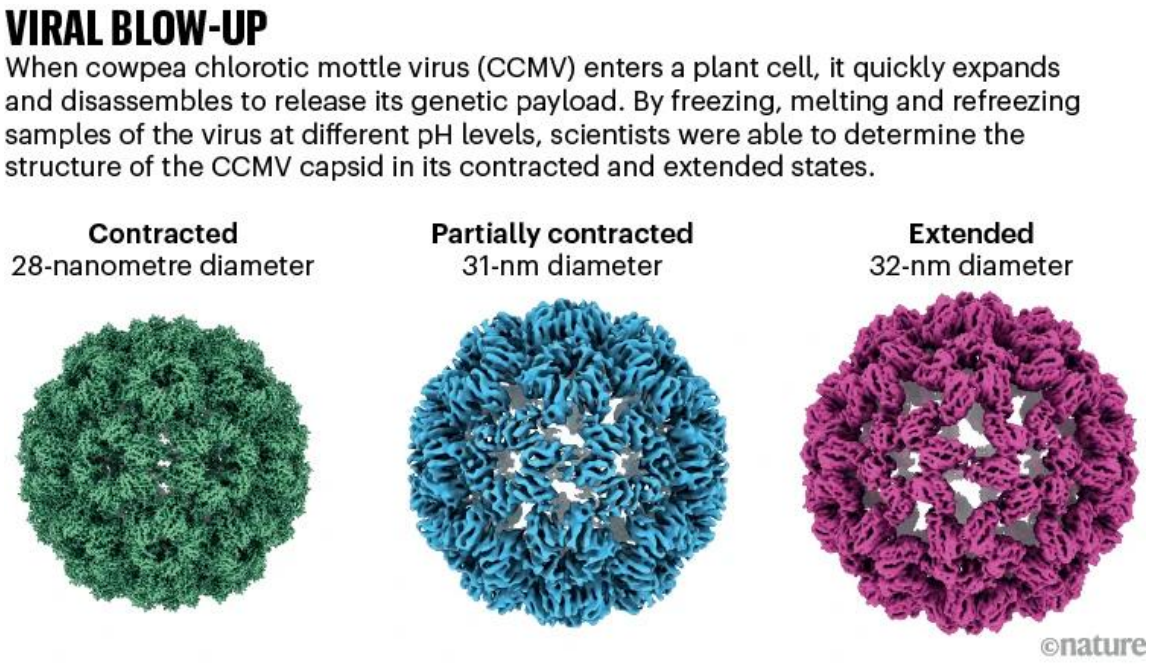
Drug development of high affinity drugs – not helpful for desired outcome

Drug discovery for T-cell receptors: an area of opportunity in ML

Hudson et al., 2023 Nat Rev Immunol

Opportunities:
Molecular dynamics and proteins in motion

- Protein motion is essential for the function of many proteins
- Fast and variable timescales of motion, with clinical implications
- Think movies, not frames. Time-resolved CryoEM addresses this



Opportunities for AI/ML:

- How might AI help us explore more complex aspects of protein biology and protein-chemical interactions?

- How might AI help us explore the “dark proteome”?