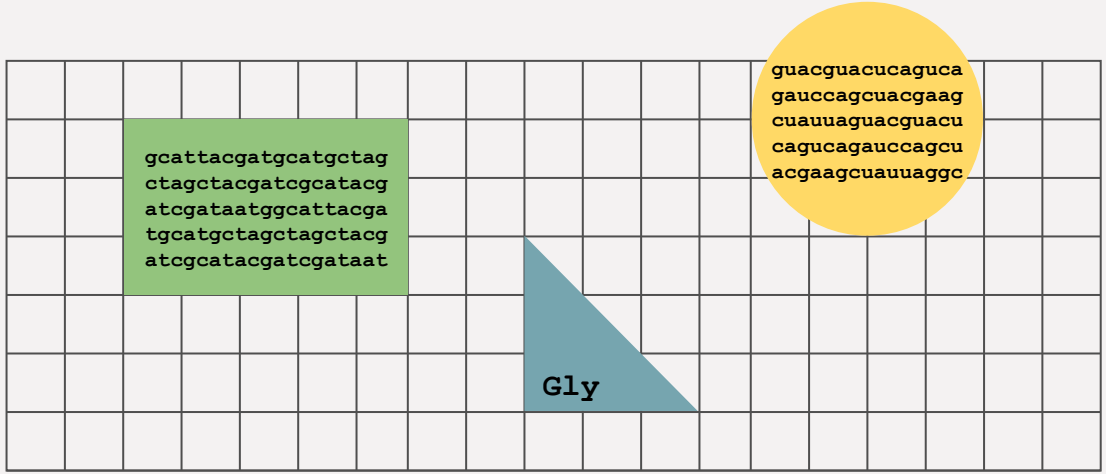
EVOLUTIONARY ALGORITHMS SIMULATING MOLECULAR EVOLUTION

(Yeah, it's kind of a mouthful.)

(Just call it "**EASME**.")



Presented by
James Satterthwaite Larkin Browning
+
María Victoria Liendro Dalinger

About James

- Bachelor's in CS from **UAB** 🐲 in 2020
  - Minor in neurobiology
- Master's in CSSE from **Auburn** 🐯 in 2022
- Doctorate in CSSE now ongoing
- Research focus on evolutionary algorithms (EAs) and their application to biology

Fun Facts

- I have two lizards, a ball python, and a tortoise.
- Insufferable Letterboxd guy.

About Victoria

- Bachelor's in biomedical engineering from **Universidad Nacional de Córdoba** in 2022
- Master's in data science from **Auburn** now ongoing
- Doctorate in CSSE coming up
- Research focus on bioinformatics applications of data science and machine learning

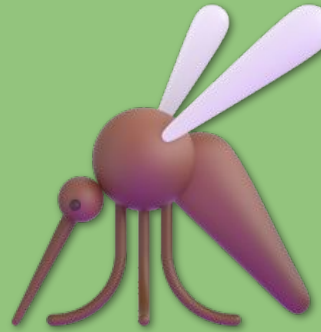
Fun Facts

- Only ever watches a movie once.
- Also watches them in two halves.

What’s on the agenda?

Bio 101: A Quick Refresh

I’m sure it’s been a while.



Where it All Began

The very first EASME project.

What We’re Up to Now

I promise we’re not creating bioweapons.

1. Biology 101

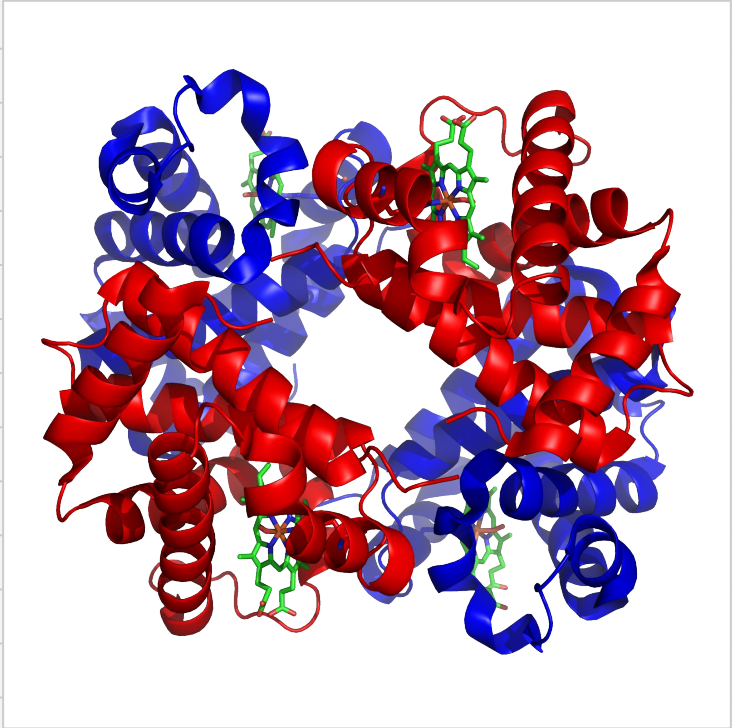
(Just the Important Stuff)

The idea that information flows from DNA to RNA to proteins is the “central dogma” of biology.

- Most of your cells contain deoxyribonucleic acid, or **DNA**, made up of four base pairs: adenine (**A**), thymine (**T**), guanine (**G**), and cytosine (**C**).
- **Transcription** converts **double-helix** DNA into **single-helix RNA**, which replaces thymine (T) with uracil (**U**).
- Ribosomes in your cells **translate** RNA into a string of **amino acids**, which then folds into a unique shape — this folded structure is a **protein**.

Three base pairs of RNA are translated into **one** amino acid. There are 64 possible RNA “codons,” which map to 20 amino acids — so it’s a hash table!

Proteins in More Detail



Hemoglobin

Evolution

(The Old-Fashioned Way)

Trillions of mutations happen inside your body every single day. Luckily, your cells automatically fix most of them — just hope "most" is enough!

- Evolutionary computing was inspired by the natural process of evolution.
- In nature, factors like radiation can cause one base pair in a DNA molecule to **mutate** into another.
  - Transition (A↔G, C↔T) is more common than transversion.
- As in evolutionary computing, some mutations are harmful, some are beneficial, and some are completely neutral.
  - As multiple, often similar codons map to the same amino acid, *most* mutations in biology are neutral.

The codons CGA, CGC, CGG, *and* CGU all map to the amino acid arginine (**R**).

AUG is the **start** codon, and UAA, UAG, and UGA are the **stop** codons.

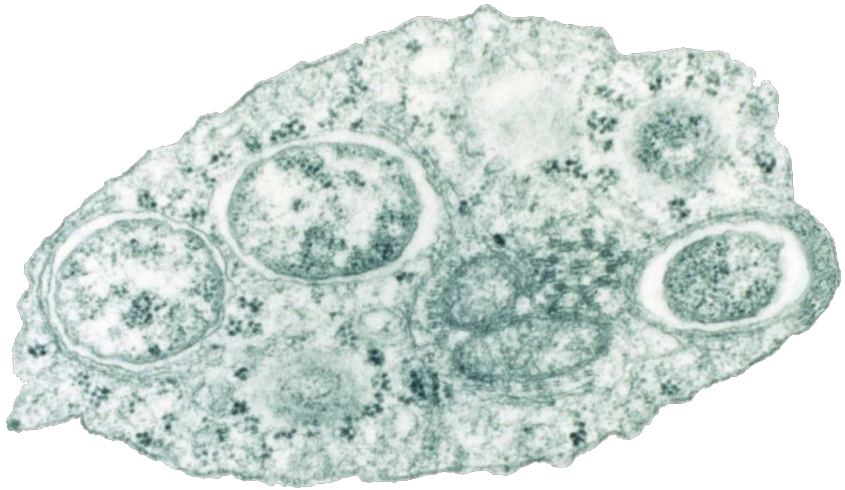
So what would happen if a CGA mutated to UGA…?

2. The First Project

feat. *Wolbachia pipientis*

The first EASME project, developed by Auburn faculty member John Beckmann and Dr. T, used an EA to simulate the evolution of cytoplasmic incompatibility in *Wolbachia*-infected mosquitoes.

That's a lot of words! But what do they *mean?*



Wolbachia in an infected host cell

Cytoplasmic Incompatibility Explained

	Uninfected Female	Infected Female
Uninfected Male	Uninfected Offspring	Infected Offspring
Infected Male	No Offspring ☠️	Infected Offspring

- Inside the sperm of an infected male, the *Wolbachia* bacteria will produce a toxin that will kill the offspring if left unattended.
- Inside the egg of an infected *female*, however, that very same bacteria will produce an antidote to that toxin!
- Through this simple toxin-antidote system, Wolbachia has become the single most common reproductive parasite in Earth's biosphere.

Which came first, the 🐔 or the 🥚?

Which evolved first, the toxin or the antidote?

Modeling the Emergence of TA Protein Functions

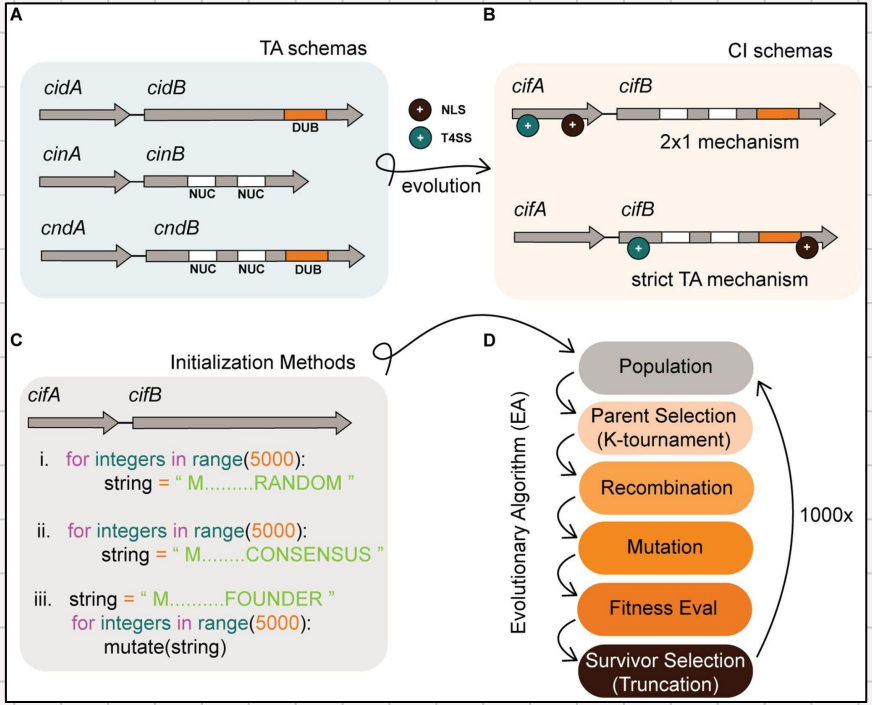
Rather than evolve some abstract representation, this project directly encoded, evolved, and evaluated DNA strings (using the A, T, G, C primitives).

The toxin-antidote system is controlled by a “lock and key” mechanism between the two proteins, which is easy to model.

Algorithm initialized a population of random sequences, then attempted to evolve them towards the modern wild-type sequences.

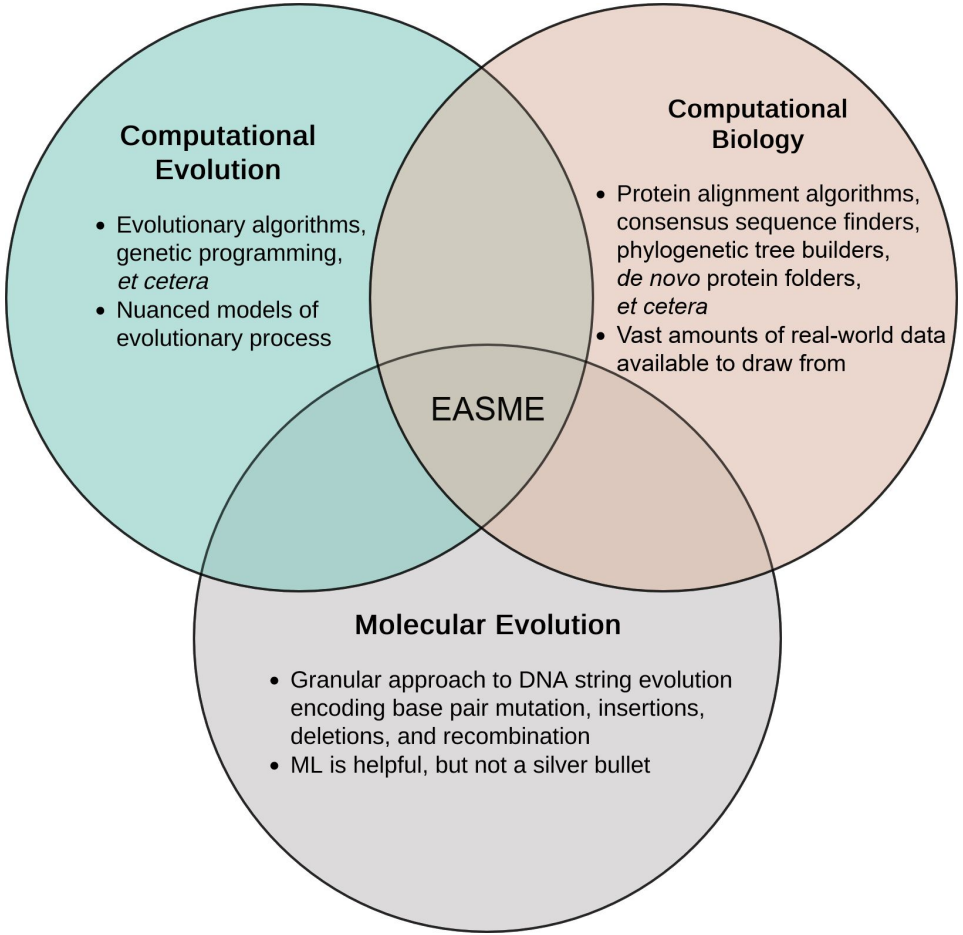
Fitness was determined by the presence of certain enzymatic domains, a “sliding window” search for binding residues, and the presence of NLS/T4SS.

Both wild-type mechanisms can be reached from the random initializations, but starting conditions will determine which one is reached first, and one is easier than the other.

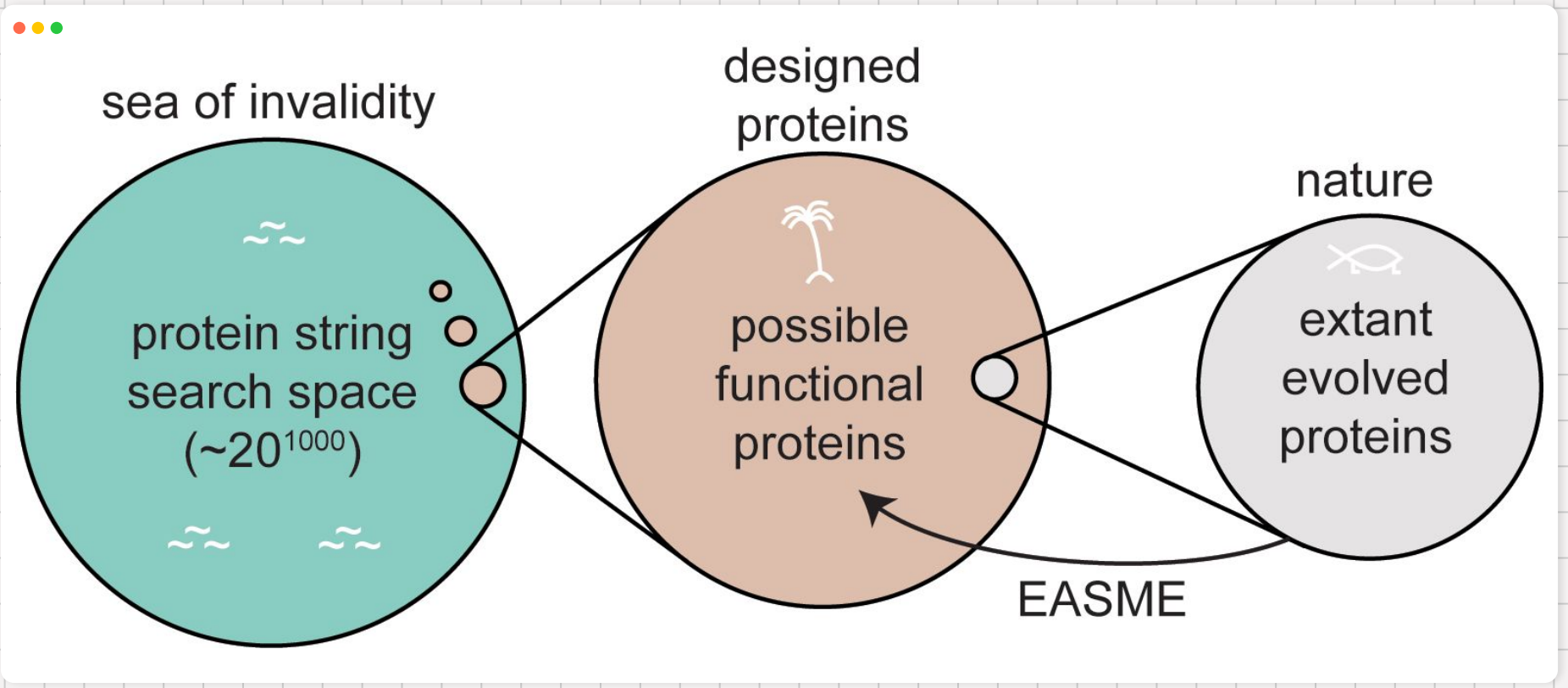


3. EASME

- EASME only became possible recently, due to the computational costs associated w/ many biological problems.
  - More sophisticated bioinformatics algorithms take better advantage of hardware.
- Wet lab experiments are a rarity in the CS field, and somewhat unique to EASME.



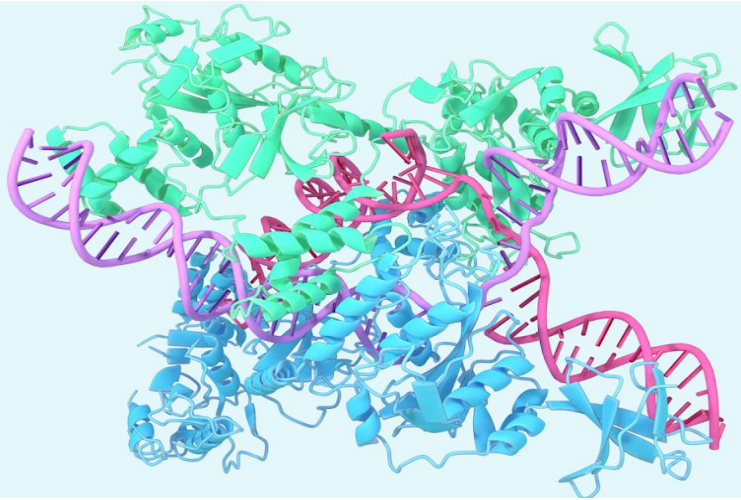
Search Space



Why Not ML?

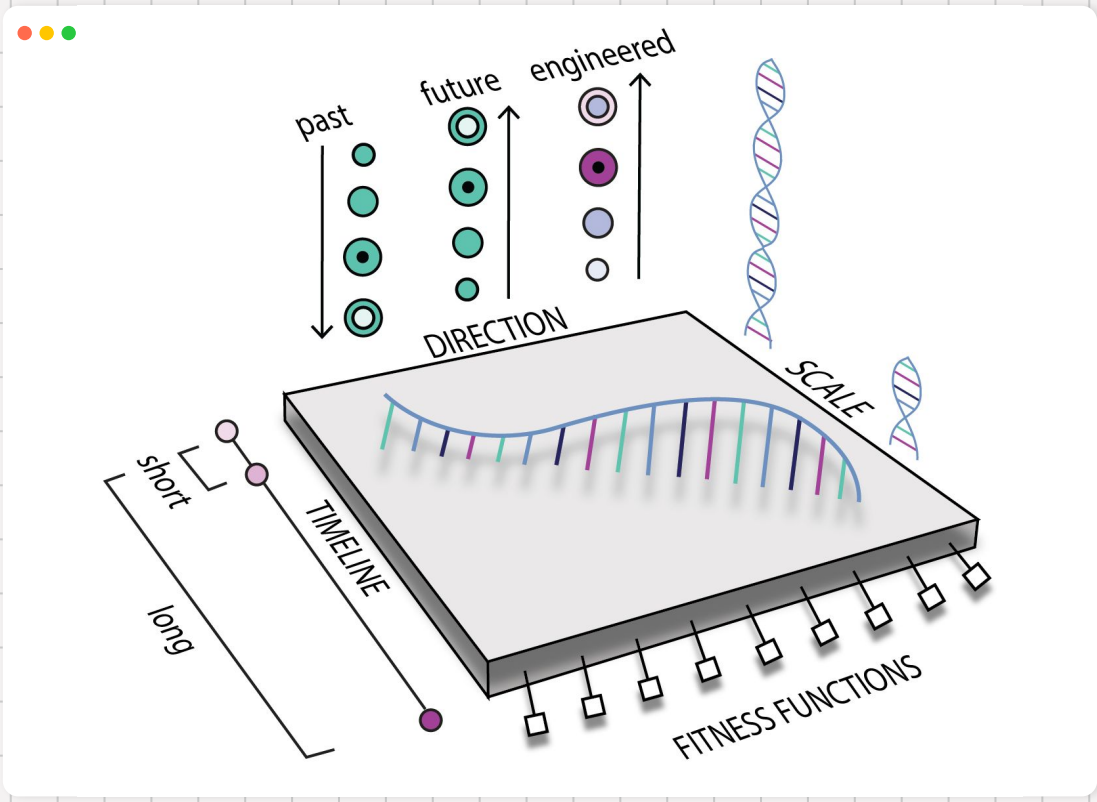
"Well, probably because this is a class on EC, not a class on ML..."

- Machine learning has been making a lot of waves in biology recently, especially with the release of Google's AlphaFold 3.
- However, EC is naturally a better fit for *evolving* new proteins, as it models the natural process that created existing proteins.
- ML models can be useful components of fitness functions, but would be somewhat limited by their own training data if tasked with designing entirely *new* proteins.
- Worth noting that EC arguably *is* a form of ML, just at the level of a population rather than individual.



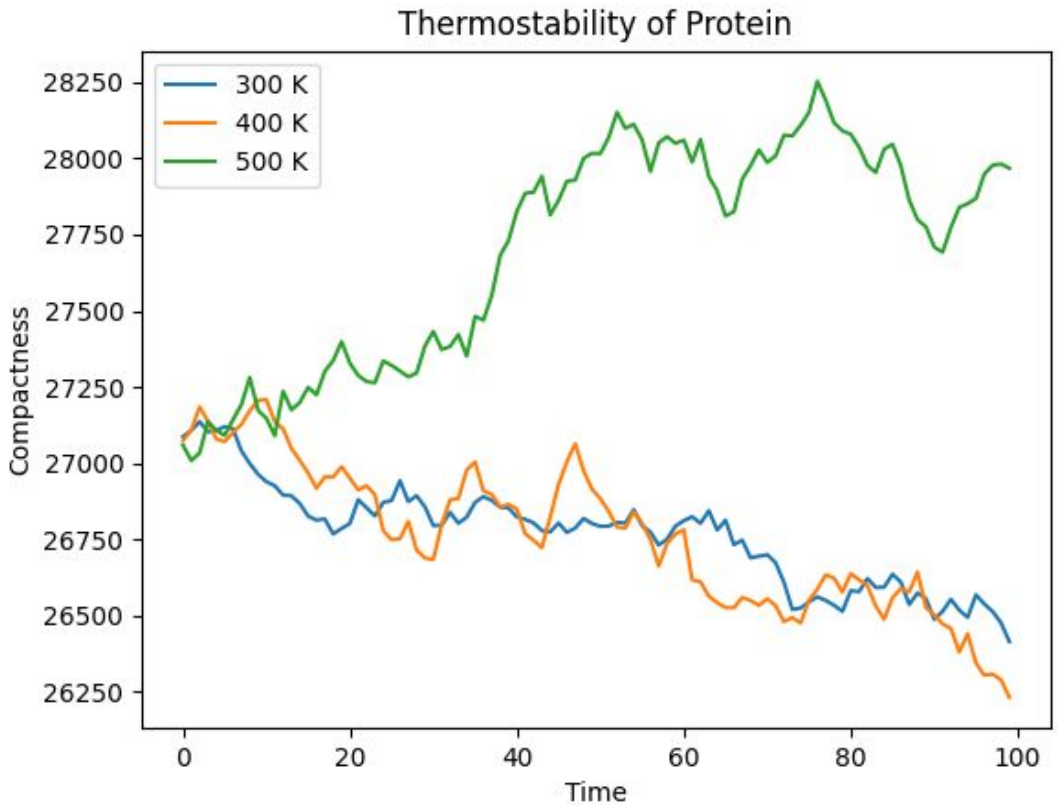
A protein folded by AlphaFold 3

Categorizing Projects

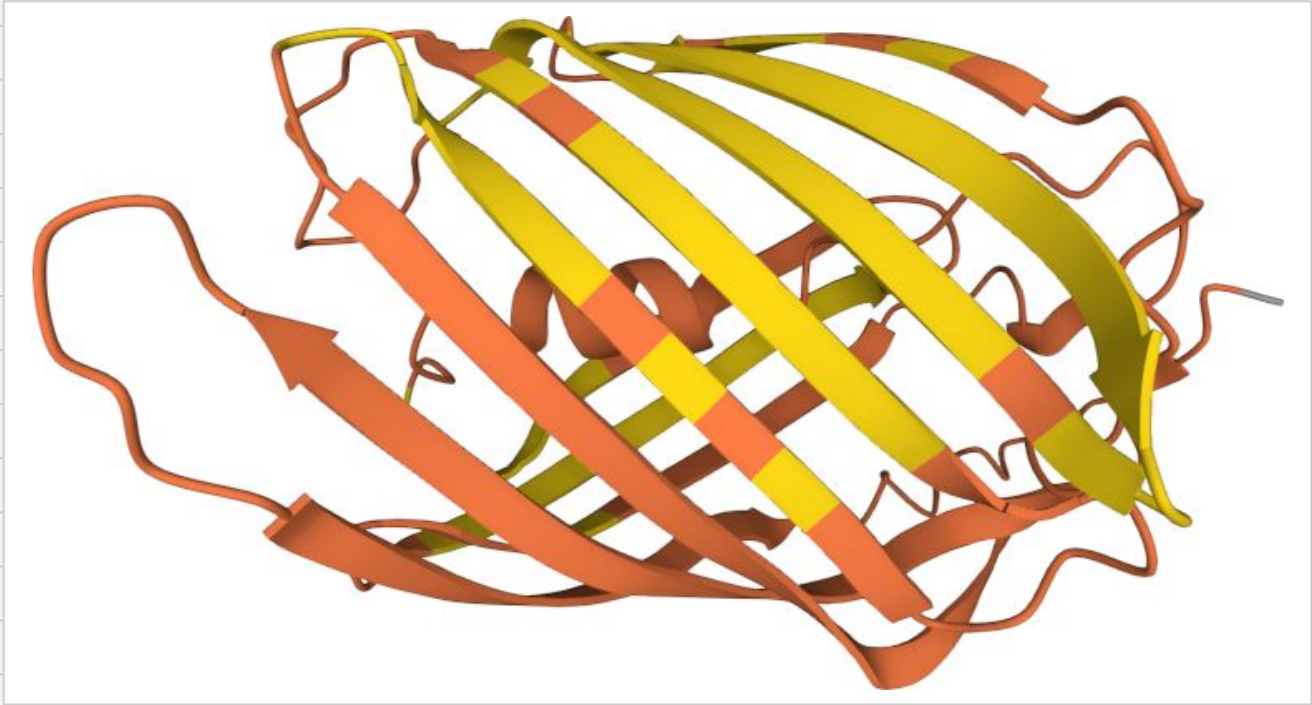


The Big Question:
Fitness

- How does one calculate the fitness of a protein?
- Protein folding is one of the biggest unsolved problems in science, but machine learning models can help us approximate structures.
- Approximated foldings and simple simulations can serve as rudimentary fitness functions.



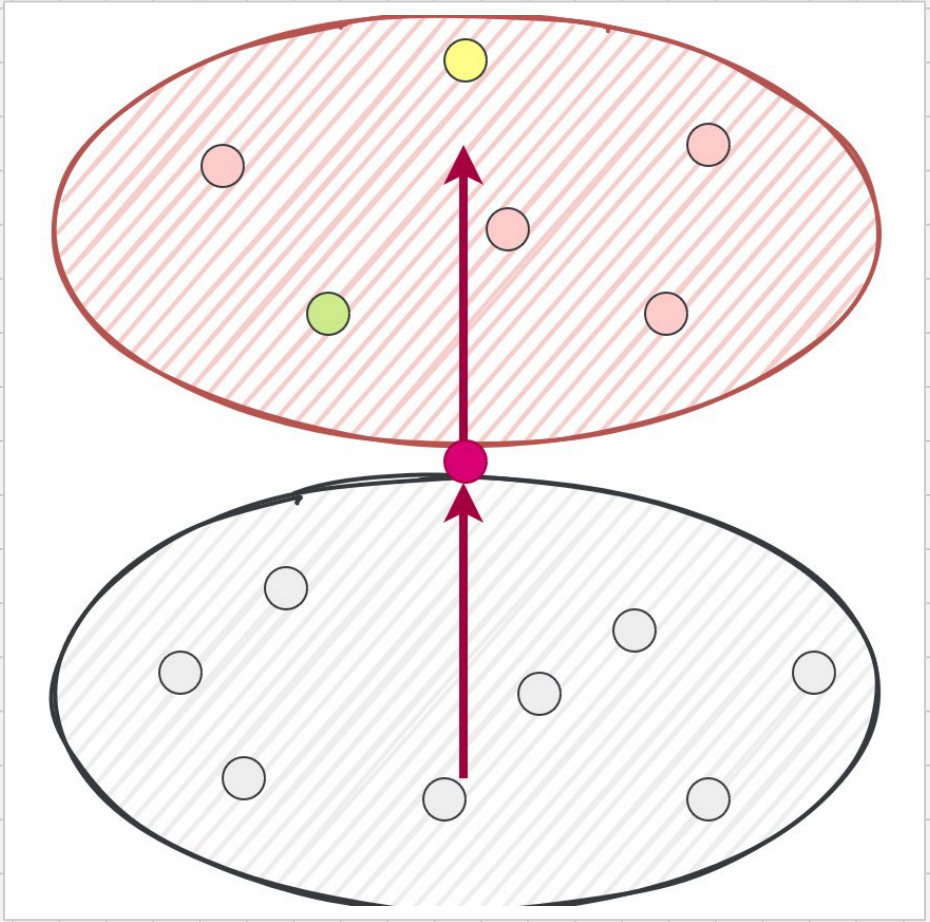
Next Step: Lab Testing

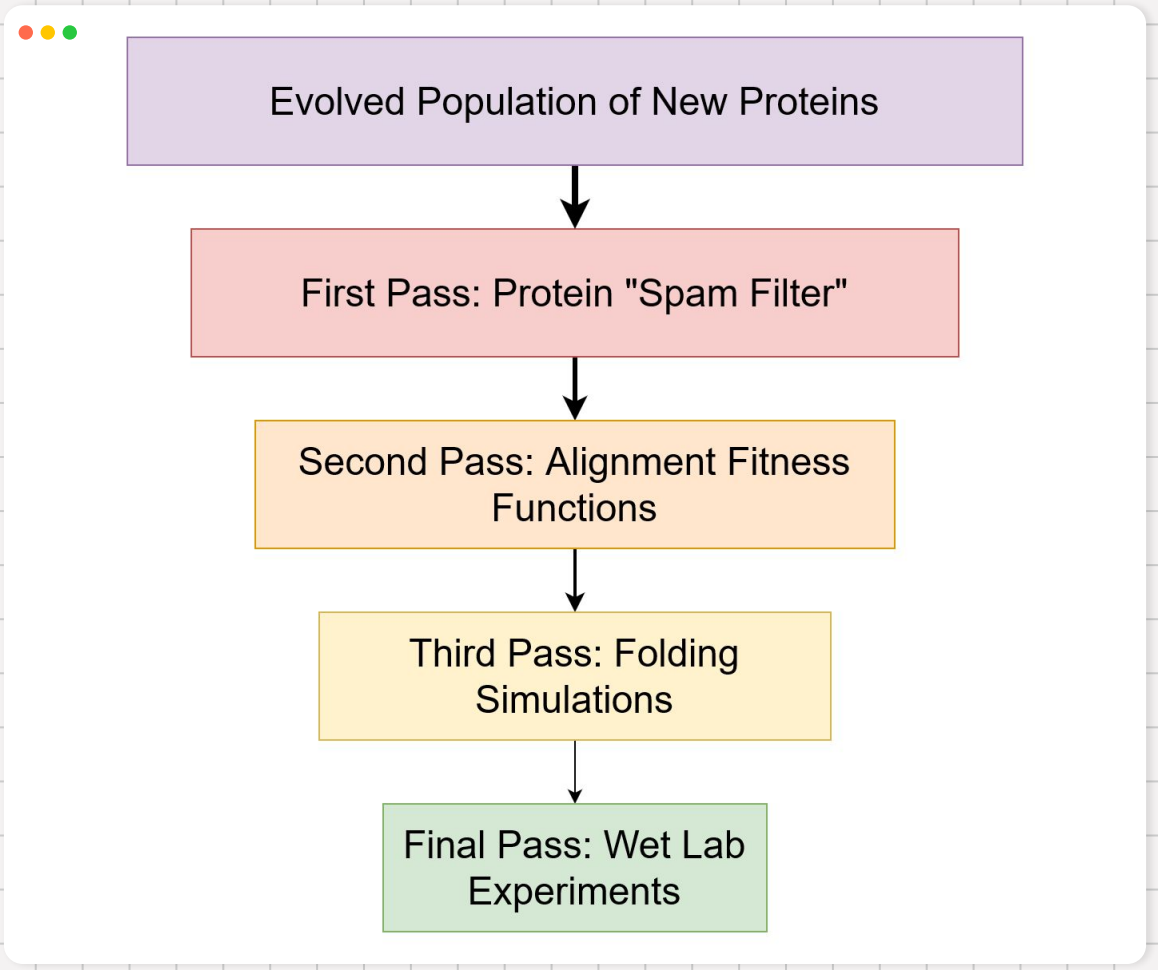


pMagenta

Fluorescent Space

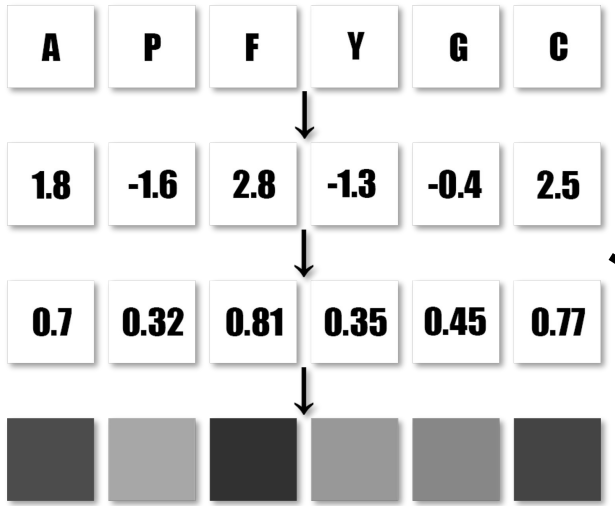
1. Download proteins similar to pMagenta from UniProt, some fluorescent, some not.
2. Group based on fluorescence.
3. Judge fitness of new variants by alignment to all specimens — homology to fluorescents is positive, homology to non-fluorescents is negative.
4. Ideally, new variants will move towards higher fluorescence in this high-dimensional space.



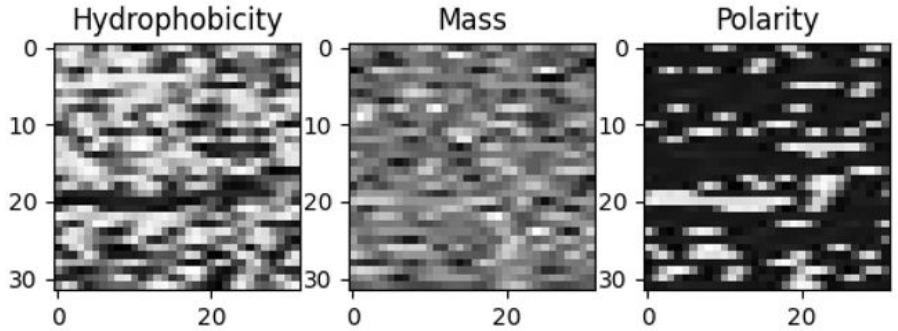

Evolved Population of New Proteins
First Pass: Protein "Spam Filter"
Second Pass: Alignment Fitness Functions
Third Pass: Folding Simulations
Final Pass: Wet Lab Experiments

Protein “Spam Filter”

- An amino acid sequence is used to generate a 32 x 32 x 3 matrix of noteworthy values:
  - Hydropathy of AAs
  - Charge of AAs
  - Mass of AAs
- Each unique amino acid has a distinct value for each of these traits — we scale them from `0` to `1`, then compress into a 32 x 32 "image."



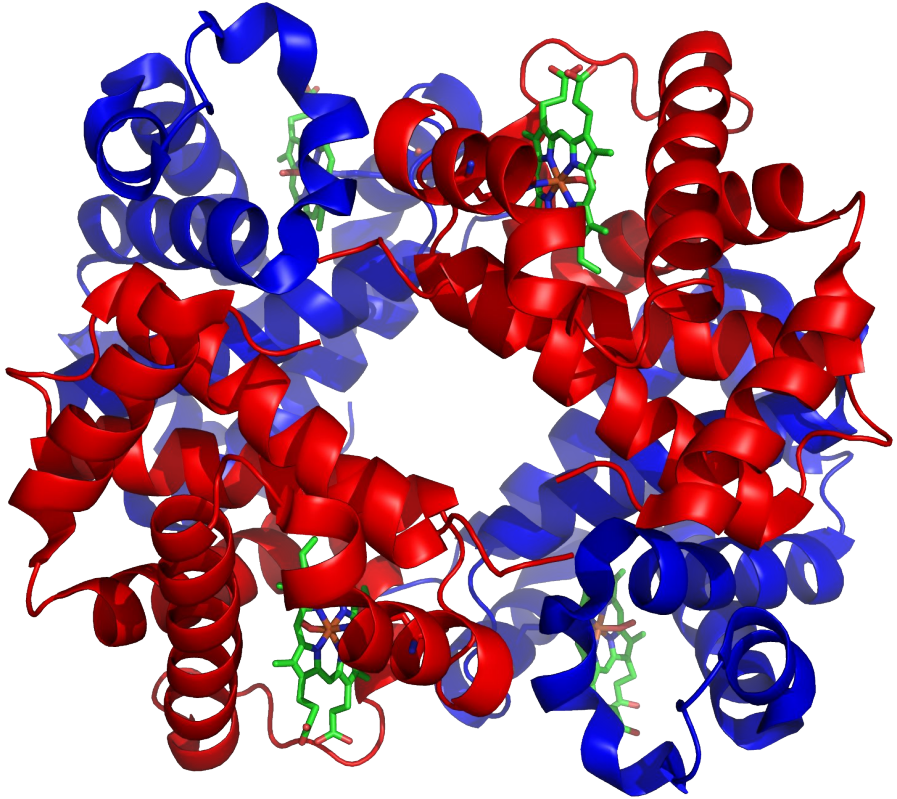
The conversion process from protein to “image”



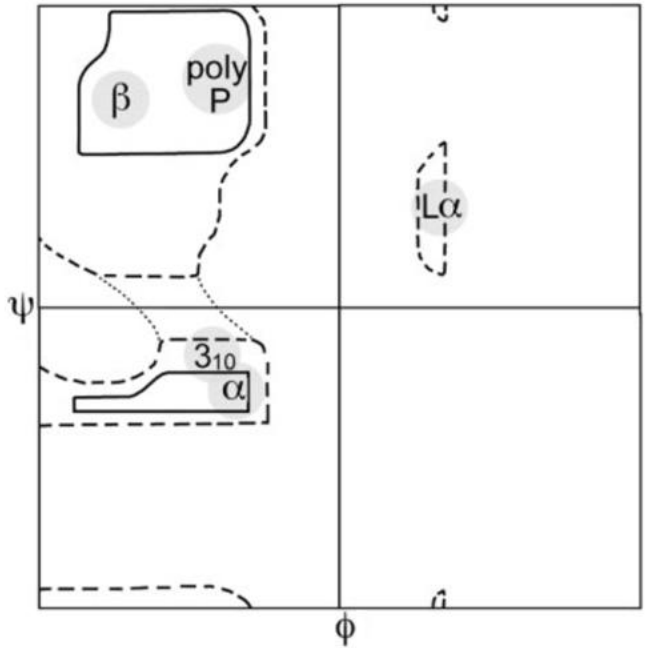
Several protein properties converted to machine-interpretable “images”

Training Process
Summary

- The **hemoglobin beta subunit** is a protein found in human red blood cells.
- We could generate 100 variants of this protein, each with a mutation rate from 1% to 100%, defining how likely each amino acid was to be randomly replaced.
    - (The 100% mutant is essentially a completely random protein.)
- Ideally, the mutation rate will *inversely* correlate with the predicted fitness, which will correlate with some "ground truth" of actual chemical stability.

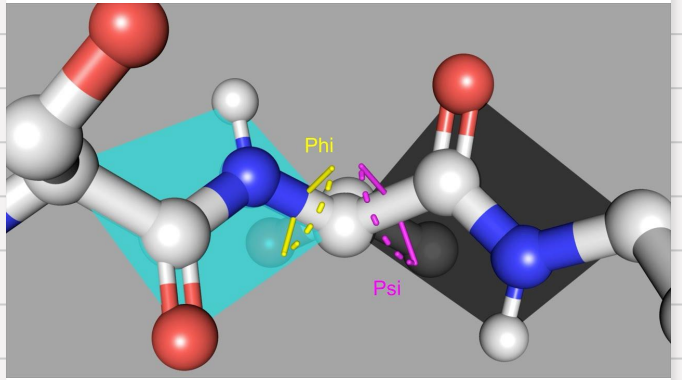


What is a **Ramachandran Plot?**



Original Ramachandran plot (meant to be universal, see Ramachandran *et al.* 1968)

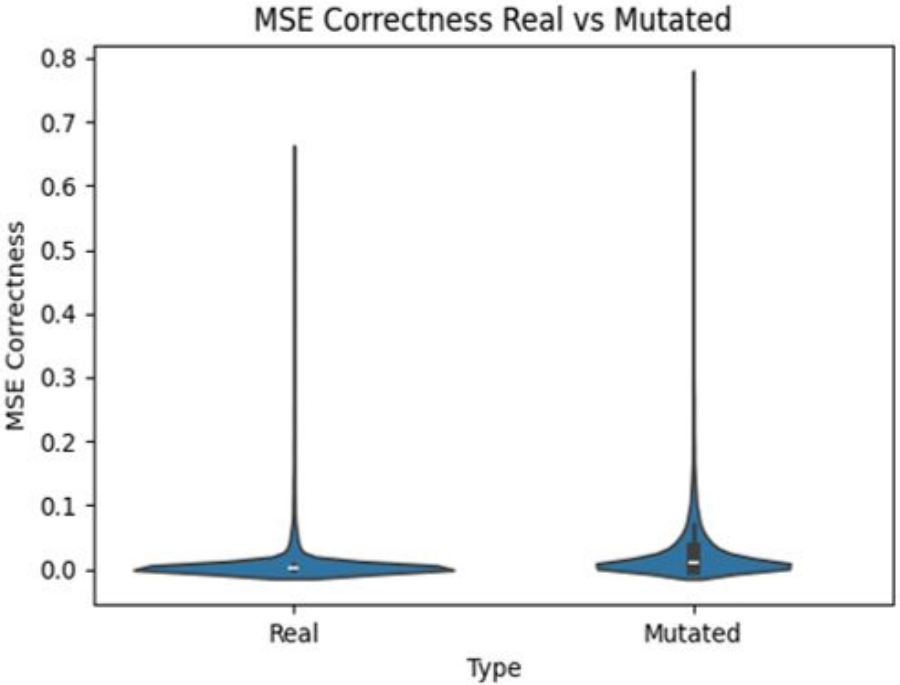
- A Ramachandran plot analyzes the angles between atoms in a protein backbone and determines which are “energetically allowed.”
    - I.e., which are likely to be stable and unstressed.
- Angles can be “favored,” “allowed,” or “disallowed.”

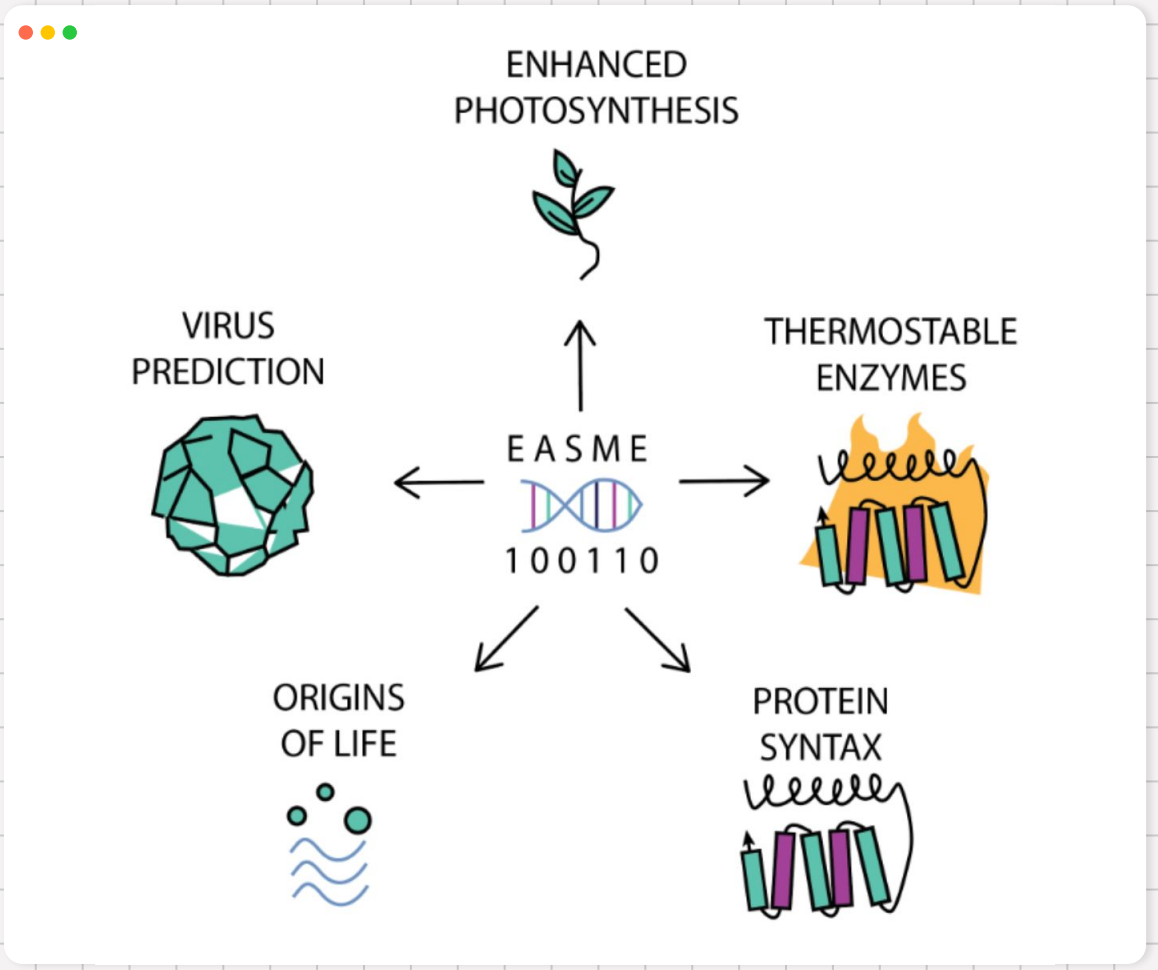


Angles that constitute a Ramachandran pair

What Does This **Mean?**

The spam filter trained off real proteins from model organisms, as well as randomly-mutated proteins (where the fitness was set at `1-MUT_RATE`), predicted fitness scores consistent with Ramachandran "ground truth" values.




ENHANCED
PHOTOSYNTHESIS
VIRUS
PREDICTION
THERMOSTABLE
ENZYMES
EASME
100110
ORIGINS
OF LIFE
PROTEIN
SYNTAX

easme.auburn.edu

Sources

- Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w
- Browning, J., Tauritz, D., Beckmann, J. “Evolutionary algorithms simulating molecular evolution: a new field proposal.” *Briefings in Bioinformatics* Vol. 25, Issue 5 (2024).
- Hemoglobin render: https://commons.wikimedia.org/wiki/File:1GZX_Haemoglobin.png
- “Modeling emergence of Wolbachia toxin-antidote protein functions with an evolutionary algorithm,” Beckmann *et al.* (2023). *Frontiers in Microbiology*, Vol. 14.
- Ramachandran *et al.*, 1968
- Image of *Wolbachia*: https://commons.wikimedia.org/wiki/File:Wolbachia.png