CS 189/289

Some applications of AI in biology:

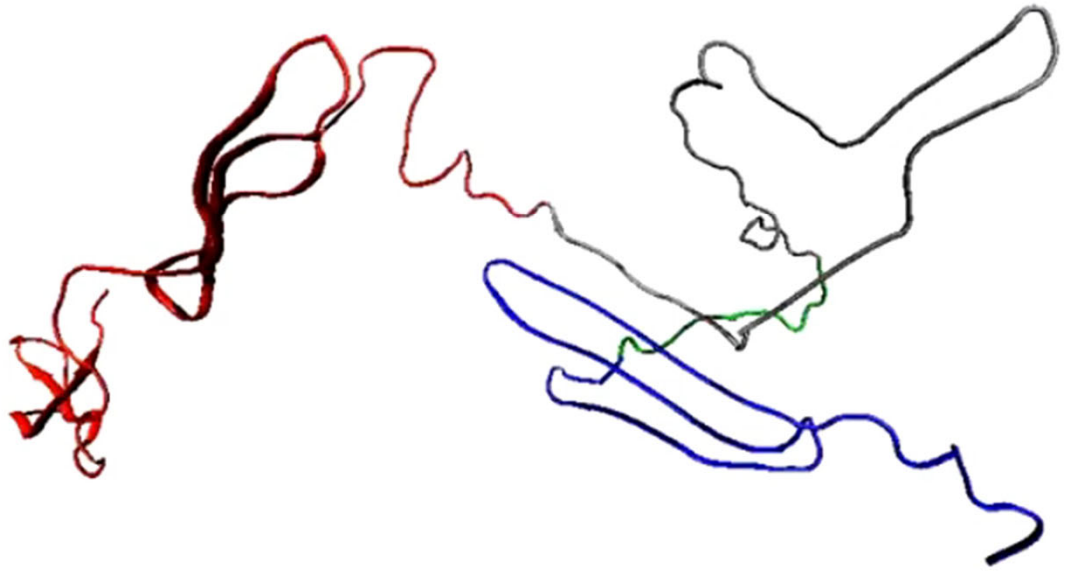
1. protein structure prediction
2. protein design

CS 189/289

Some applications of AI in biology:

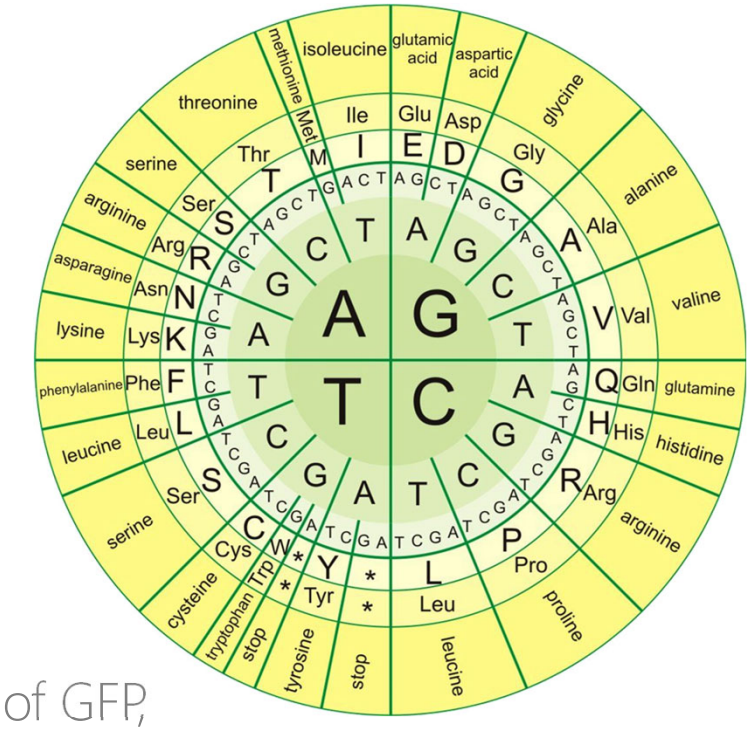
1. protein structure prediction
2. protein design

Proteins are strings of nucleotides

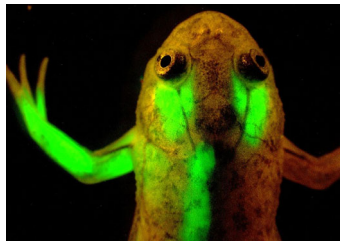


238 length amino acid sequence:

MSKGEELFTGVVPILVELDGDVNGHKFSVSG
EDFFKS…NSHNVYIMADKQKNGIKVNFKIRH

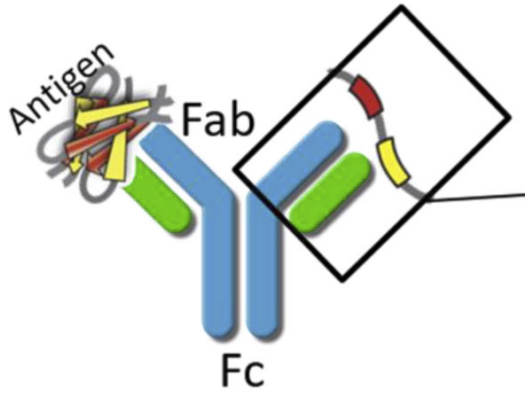


Green fluorescent protein (GFP) folding itself

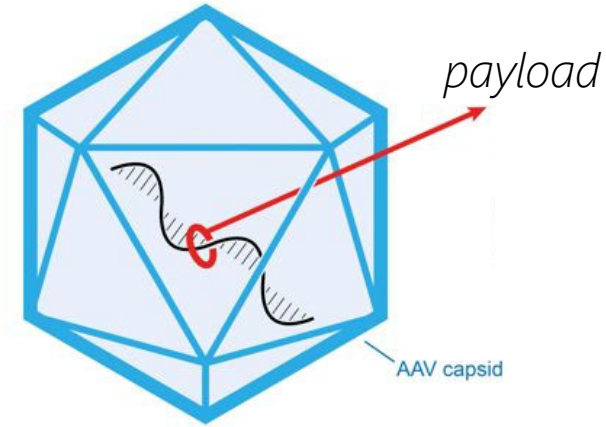


[2008 Nobel in chemistry for discovery and development of GFP, Osamu Shimomura, Martin Chalfie and Roger Y. Tsien]

Protein engineering: therapeutics, environment, *etc.*

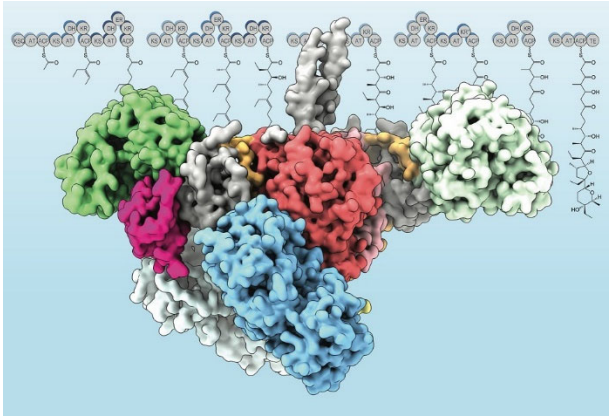


antibody therapeutics

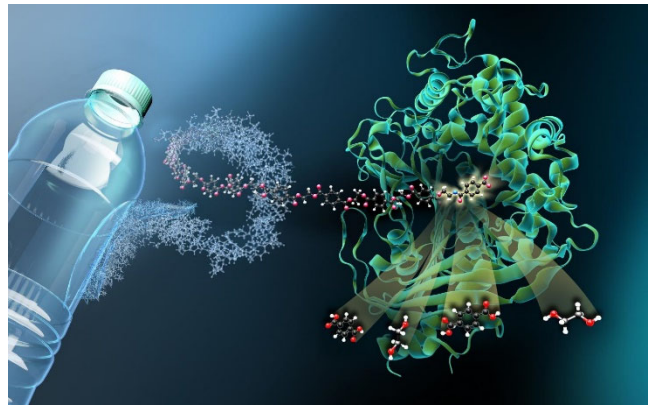


gene therapy virus delivery (AAV)

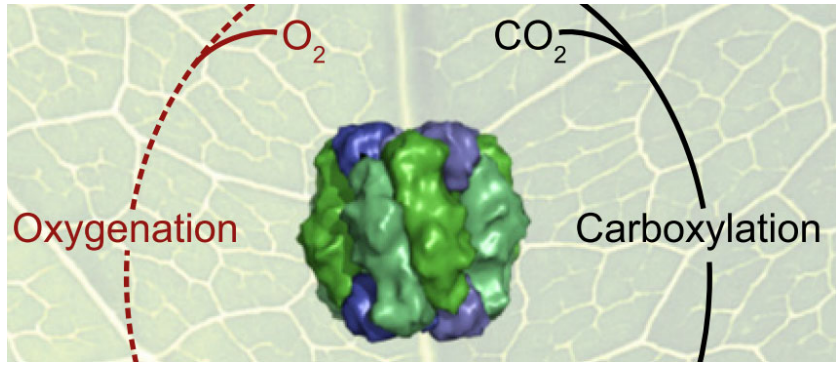
gene editing (CRISPR/Cas9)



antibiotics & biofuel production (PKS)



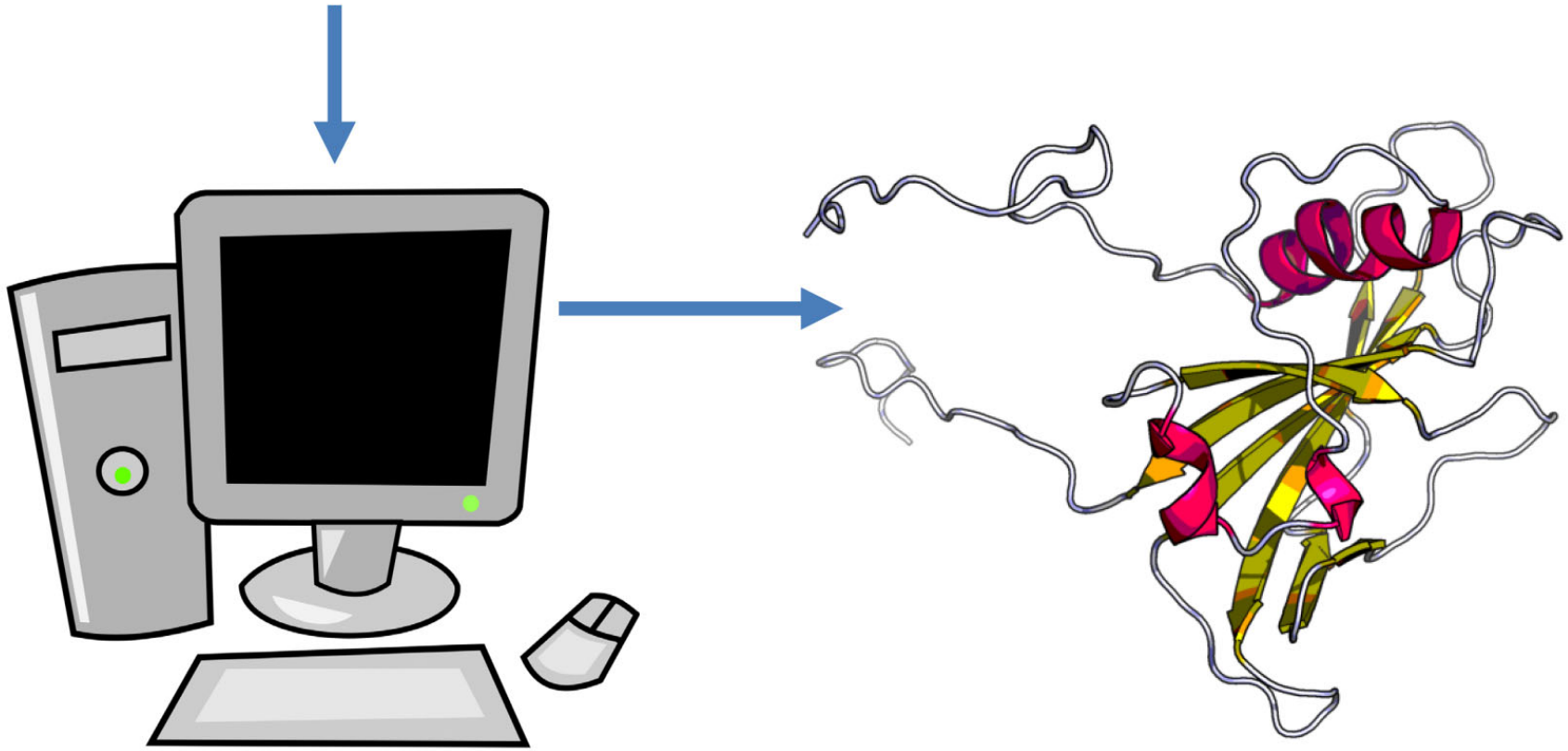
plastic recycling (PETase)



CO_2 biosequestration (RuBisCO)

Protein Structure Prediction

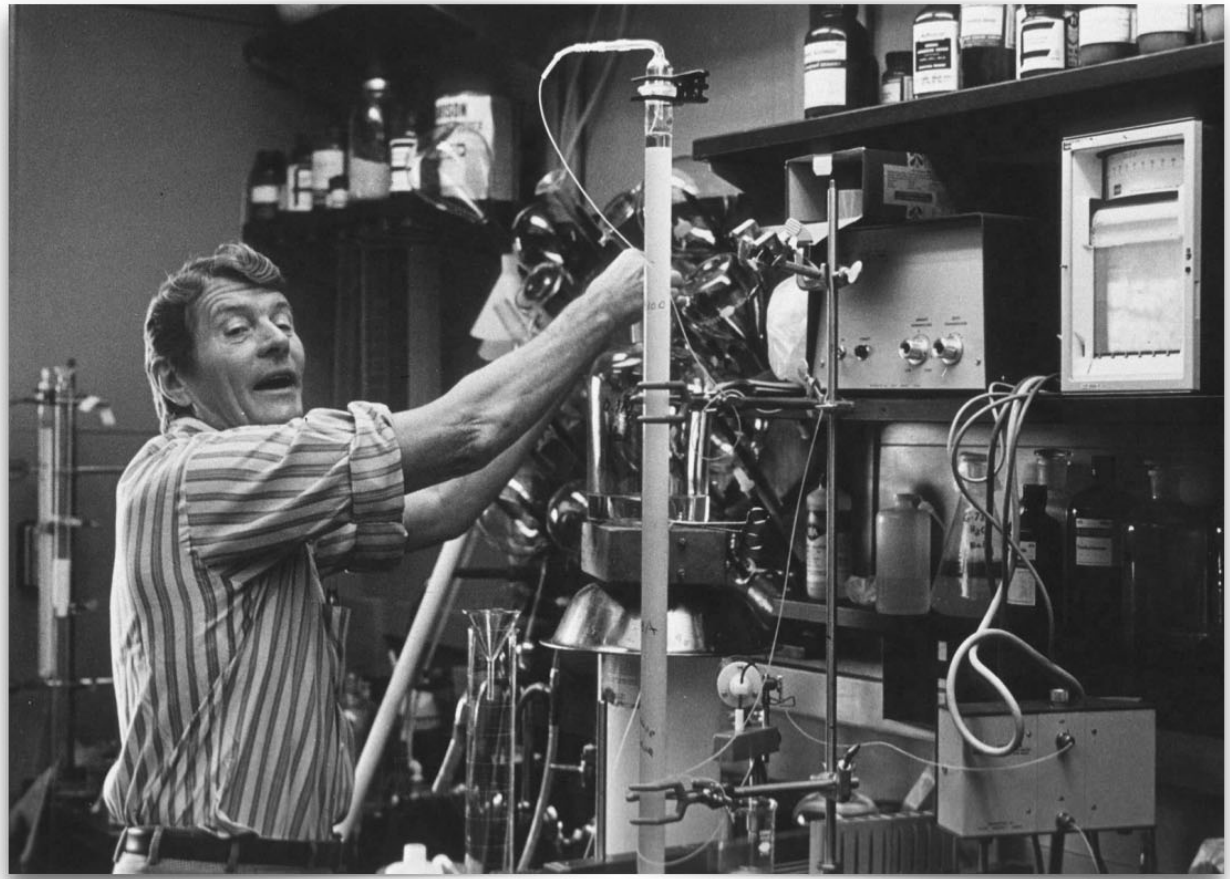
MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK



Has been studied several decades

[slide from Jinbo Xu, TTI]

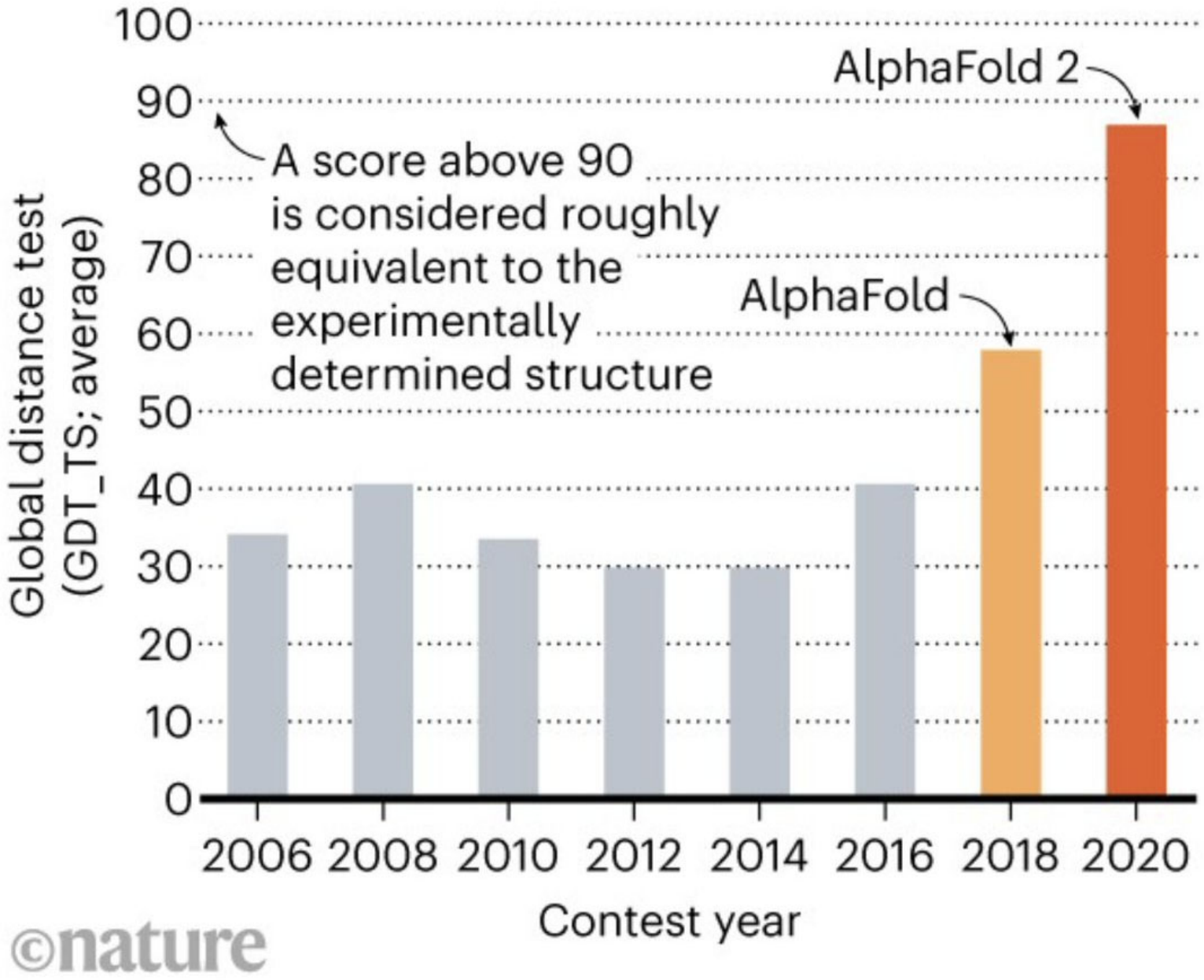
Amino acid sequence determines protein 3D structure



Christian Anfinsen
Nobel Prize in Chemistry 1972

2020

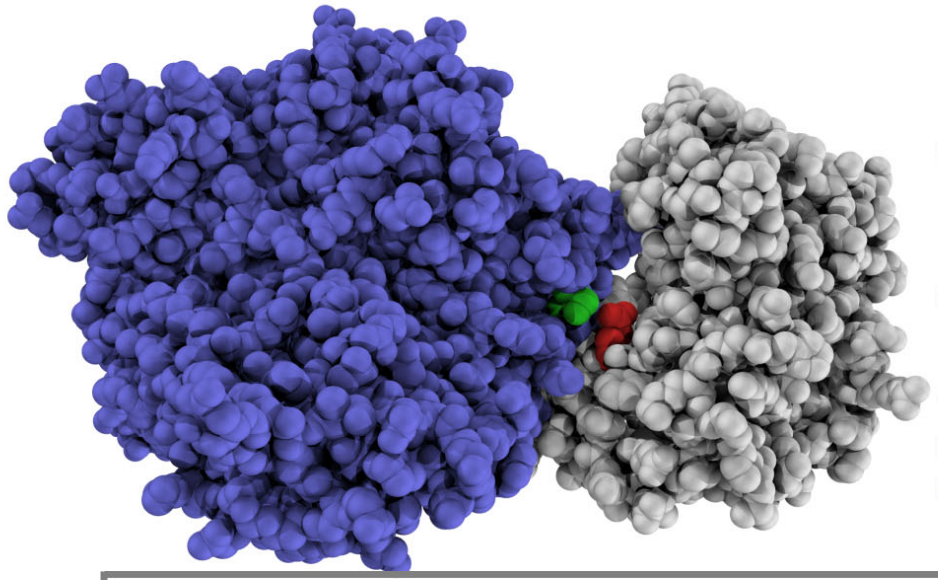
For the first time, state-of-the-art is deep learning based:



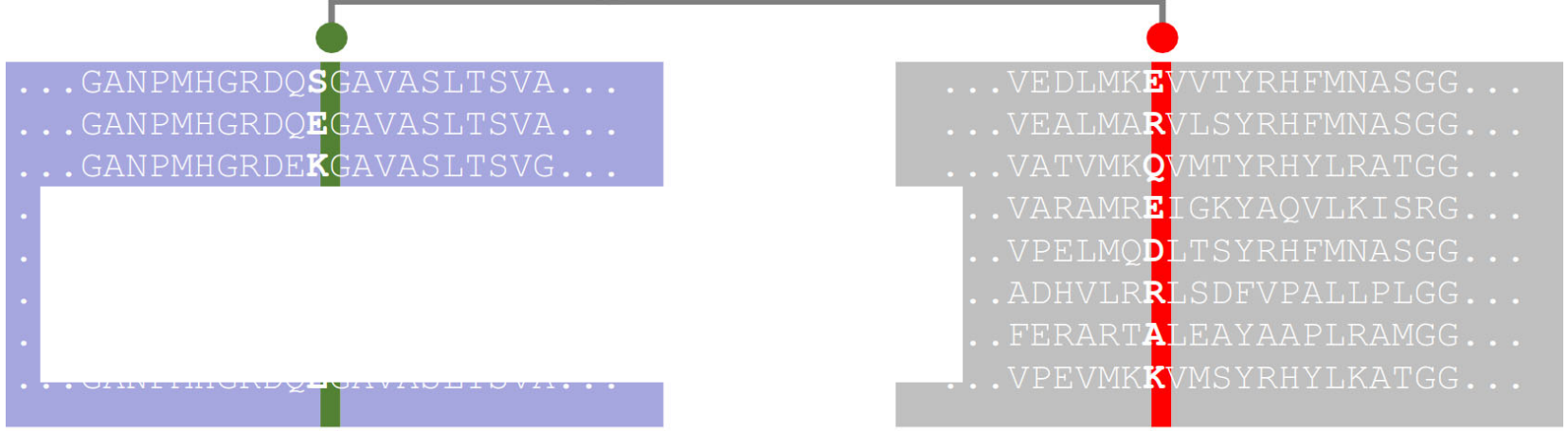
AlphaFold2 relies on previous key insight

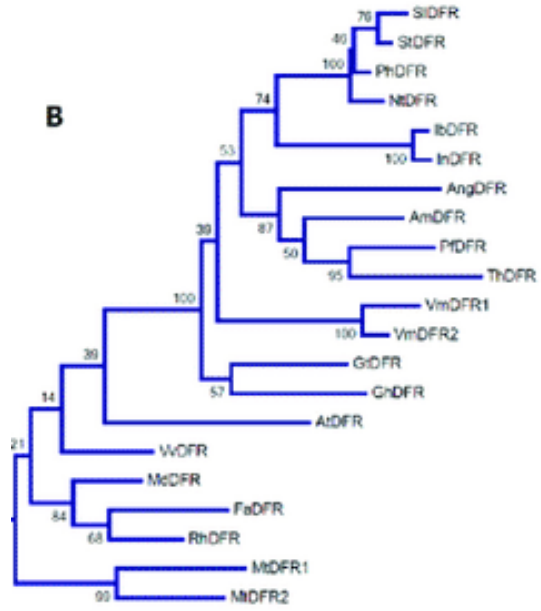
Amino acids in direct physical contact tend to covary or "coevolve" across related proteins

$$P(\sigma_1, \ldots, \sigma_N) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} h_i \sigma_i + \sum_{1 \le i < j \le N} J_{ij} \sigma_i \sigma_j\right)$$



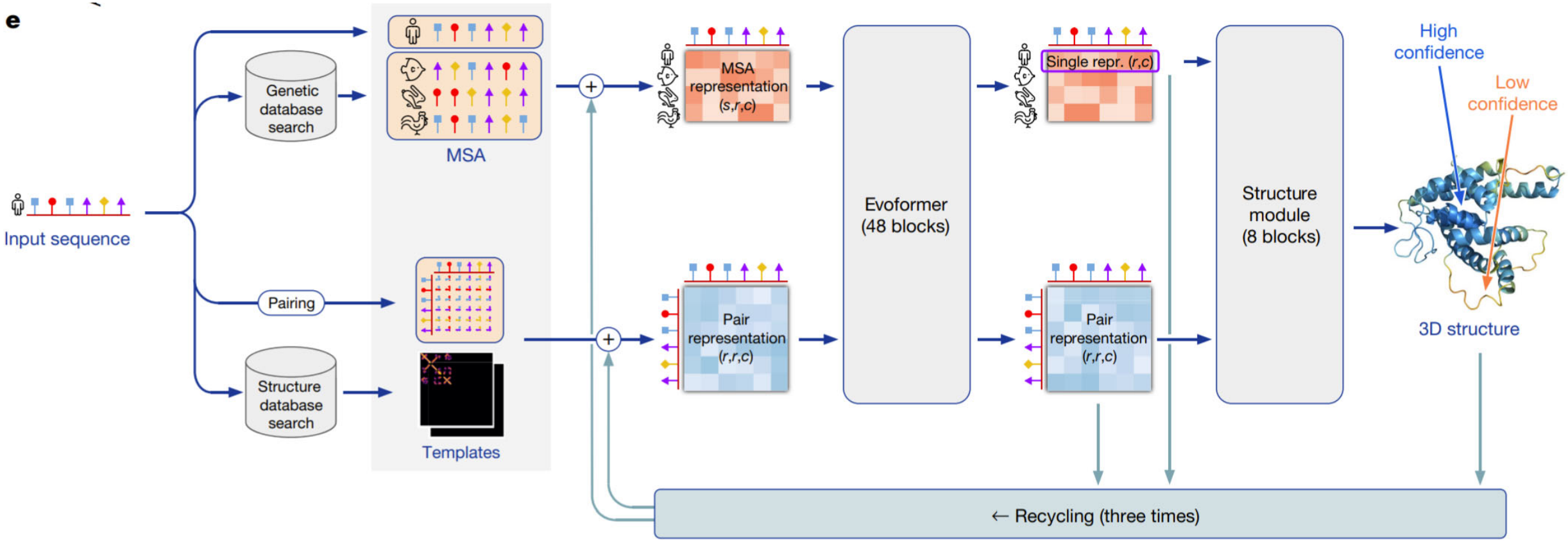
For example, a mutation that causes one amino acid to get bigger is more likely to preserve protein structure and function (and thus survive) if another amino acid gets smaller to make space



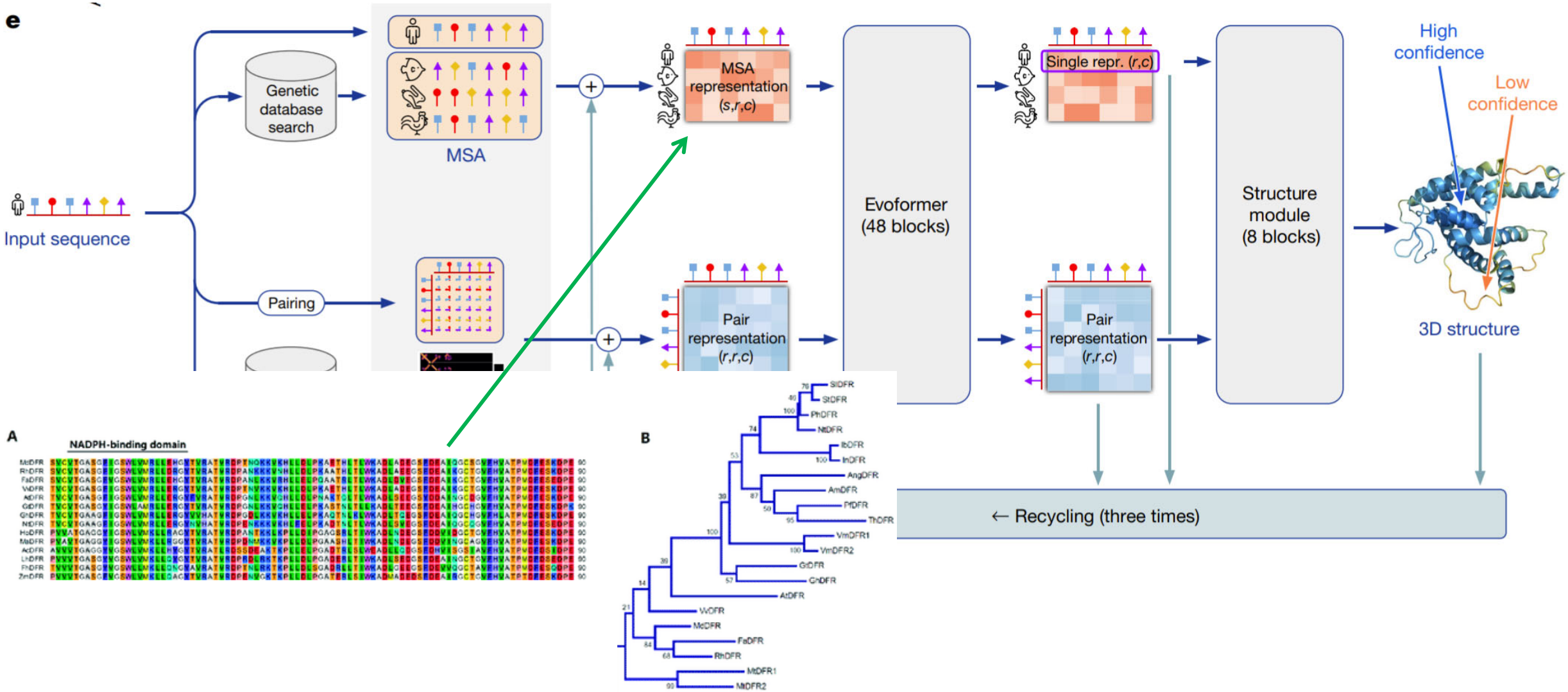


[slide from Jinbo Xu, TTI]

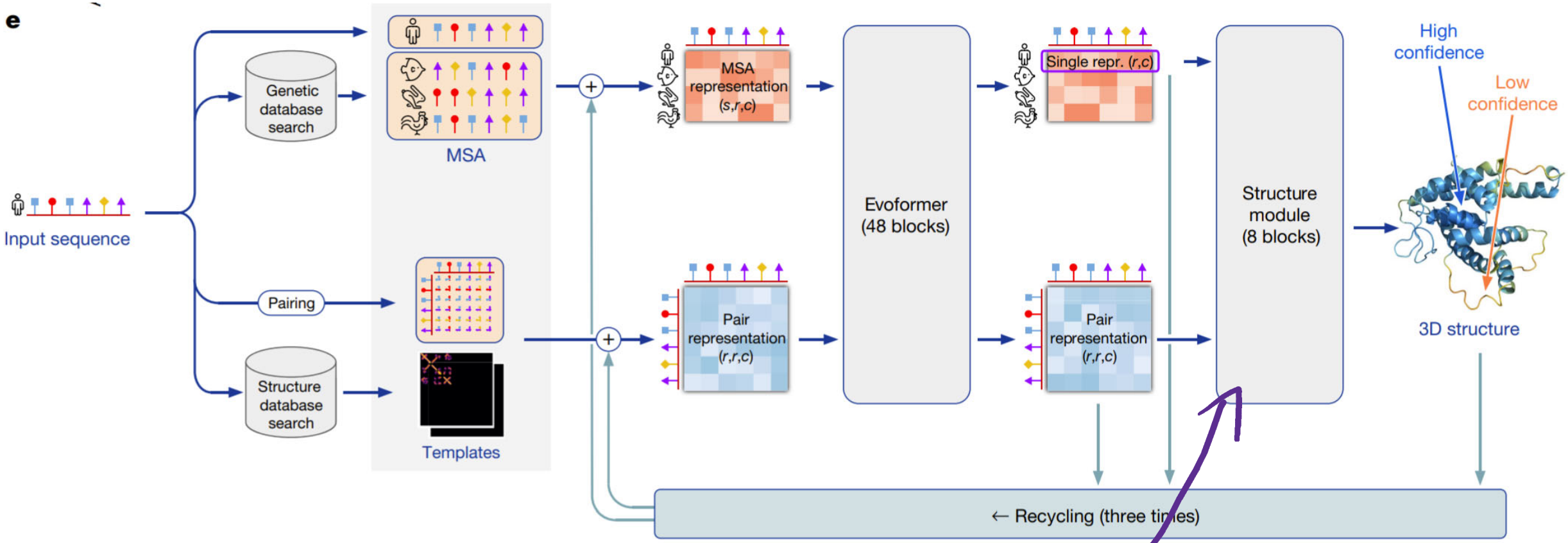
AlphaFold2 “almost end-to-end” neural network



AlphaFold2 “almost end-to-end” neural network



AlphaFold2 “almost end-to-end” neural network



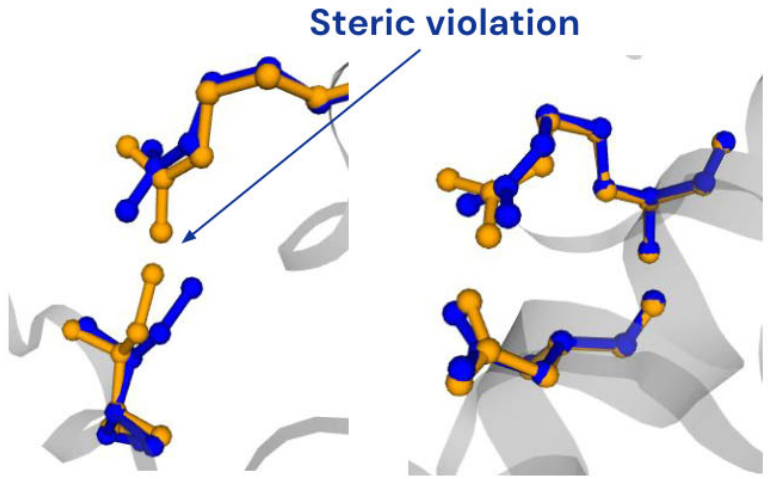
Uses a rotation equivariant attention architecture in structure module

AlphaFold2 “almost end-to-end” neural network

- Can end up with atom positions in violation of physics.
- Thus relies on old style energy-based approaches to refine the predicted 3D coordinates.

Relaxation

- The end result of iterative refinement is not guaranteed to obey all stereochemical constraints
- Violations of these constraints are resolved with coordinate-restrained gradient descent
- We use the Amber ff99SB force field[1] with OpenMM[2]



Orange: pre-relax
Blue: post-relax

Some thoughts on AlphaFold2

- DeepMind took on a long-tackled, well-defined problem, with clear data, clear benchmarks, and a clear way to demonstrate improvement.
- Expense of protein structure data used for AlphaFold2, conservatively estimated at ~US$20 billion (Burley et al., 2023).
- They relied **heavily** on years of prior work in protein folding research: “template-based modelling”, “evolutionary co-evolution modelling”, “contact prediction”, energy-functions.

AlphaFold3 (*Nature*, May 2024)

- Big scandal when paper came out because code wasn't released, results couldn't be reproduced, and public could not test it out. Just got released.
- Extends its predictive to include proteins, DNA, RNA, and ligands, ions, and interactions between them.
- No direct comparison of AF2 to AF3 for protein structure prediction, but seems better on some binding tasks than AF2-multimer.
- Now uses diffusion model at last stage to *sample* 3D structures (and throws away rotational equivariance in this module).

CS 189/289

Some applications of AI in biology:

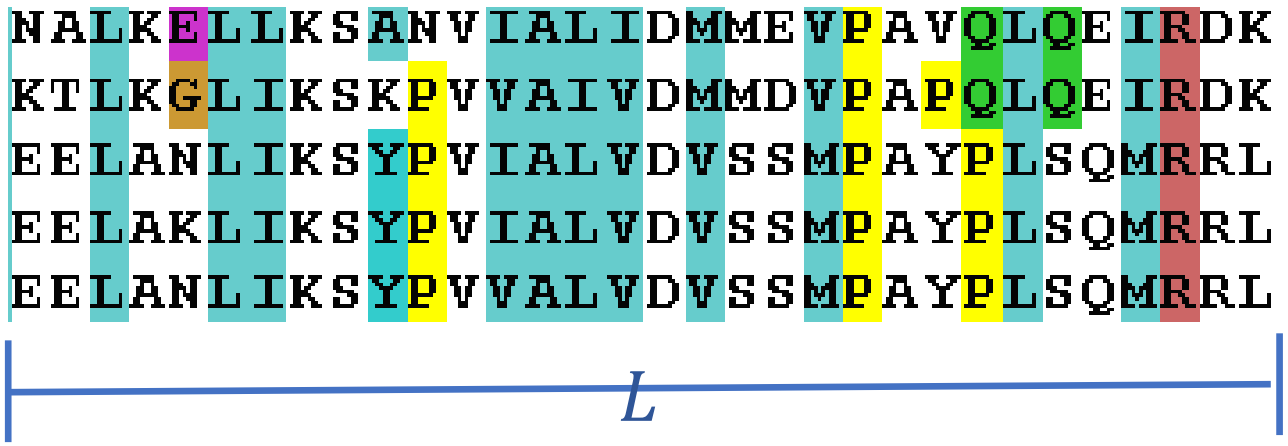
1. protein structure prediction
2. protein design

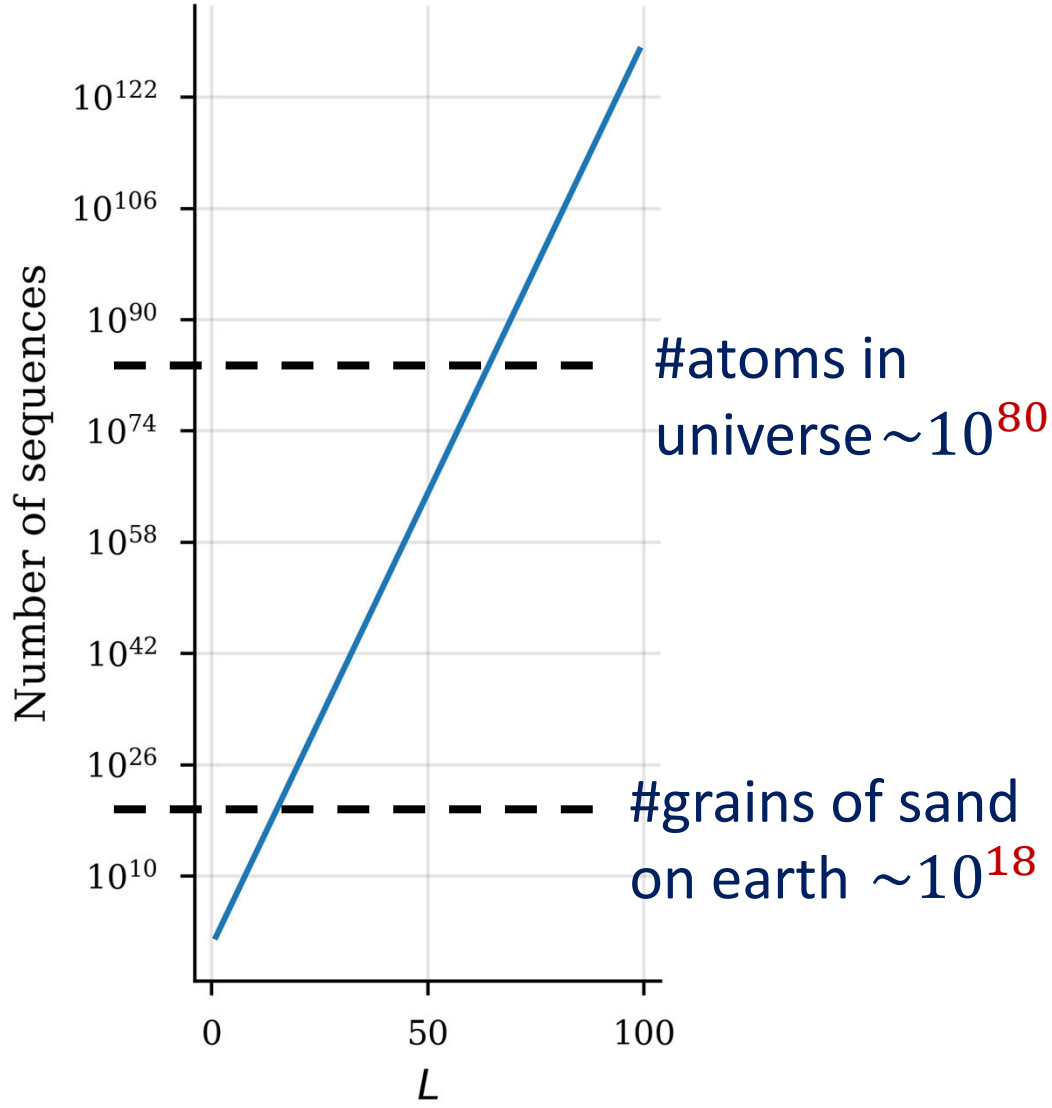
Fundamental difficulty: design space is nearly infinite

- Also highly rugged design space

$\Longrightarrow$ size scales as $\sim 20^L$

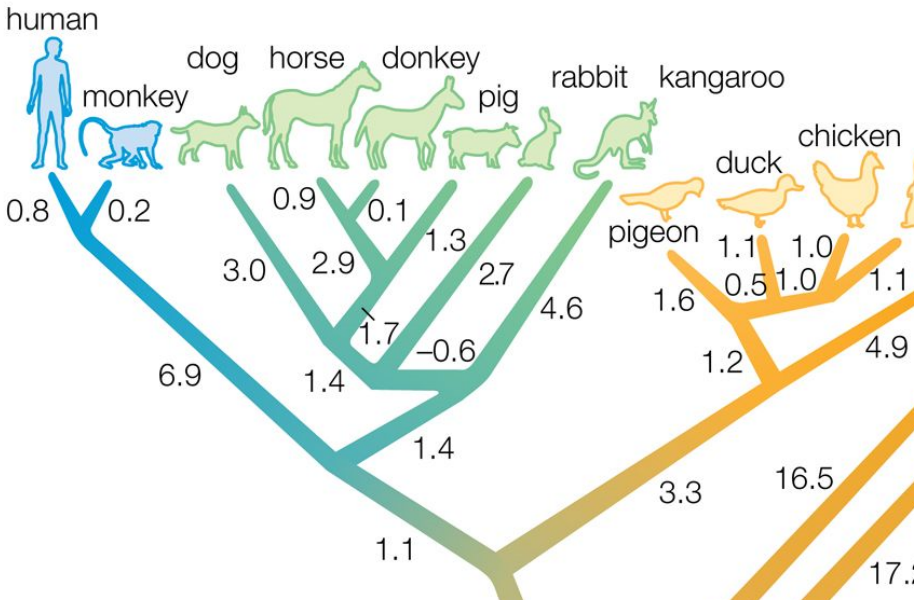
- Discrete search space (no gradients)



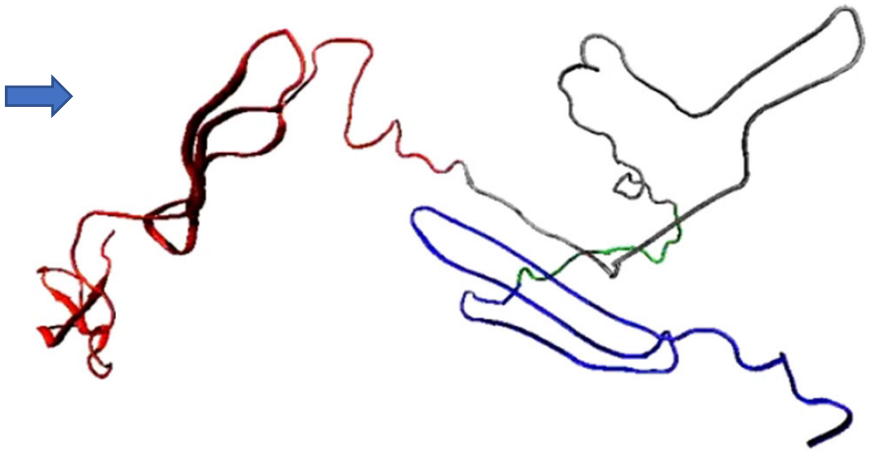


Successes in navigating this complex space

1. Nature: via evolution *over millions of years*.



MSKGEELFTGVVPILV
ELDGDVNGHKFSVSG
EGEGDATYGKLTLKFIC
TTGKLPVPWPTLVTTF
SYGVQCFSRYPDHMK
QHDFFKSAMPEGYVQ
ERTIFFKDDGNYKTRA
EVKFEGDTLVRIELKGI
DFKEDGNILGHKLEYN
YNSHNVYIMADKQKN
GIKVNFKIRHNIEDGSV
QLADYQQNTPIGDGPV
LLPDNHYLSTQSALSK
DPNEKRDHMVLLEFVT
AAGITHGMDELYK



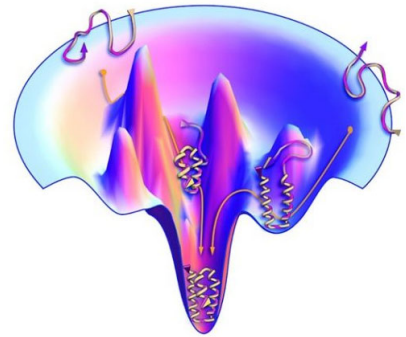
green fluorescent protein folding itself

Successes in navigating this complex space

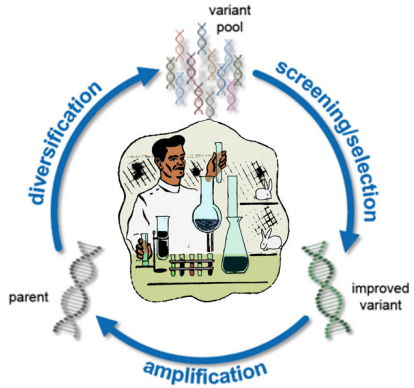
1. Nature: via evolution *over millions of years.*
2. Various protein engineering strategies.

Protein engineering strategies emerging

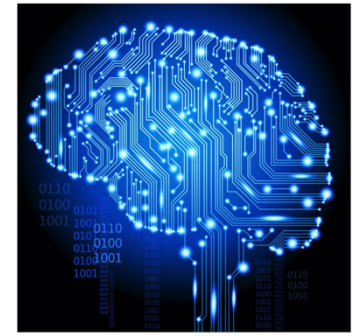
i. Computation ("data free"): **physics-based energy functions** (*e.g.*, Rosetta) to model **protein structure**, and protein binding.
~1997-2023'ish (almost R.I.P.) [2024 Nobel Prize]



ii. Wetlab: **directed evolution** to iteratively directly design property of interest.
~1993-present [2018 Nobel Prize]



iii. Machine learning (augmented): generative models; function prediction; structure prediction, etc. *~2018(?)-present*



Did AlphaFold2/3 "solve" protein engineering?

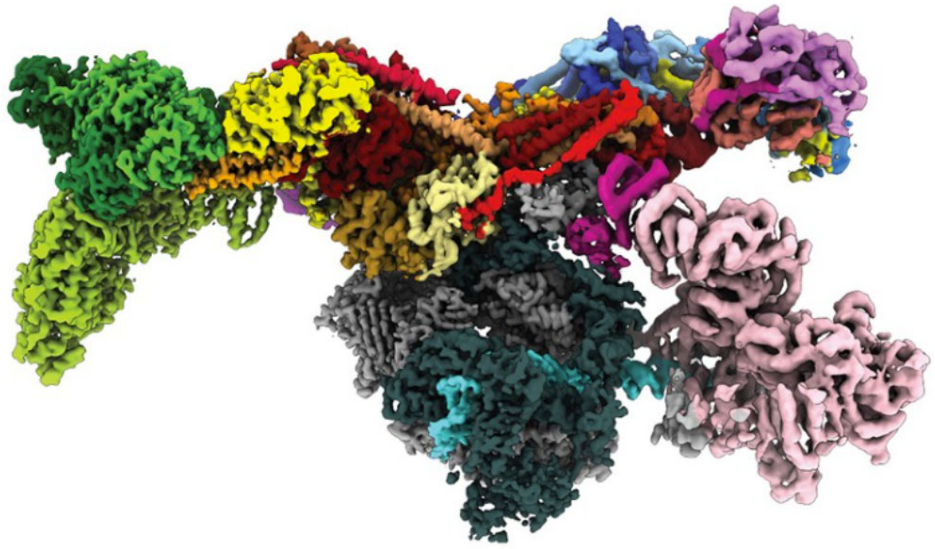
NEWS | 22 July 2021

DeepMind's AI predicts structures for a vast trove of proteins

AlphaFold neural network produced a 'totally transformative' database of more than 350,000 structures from *Homo sapiens* and 20 model organisms.

Ewen Callaway

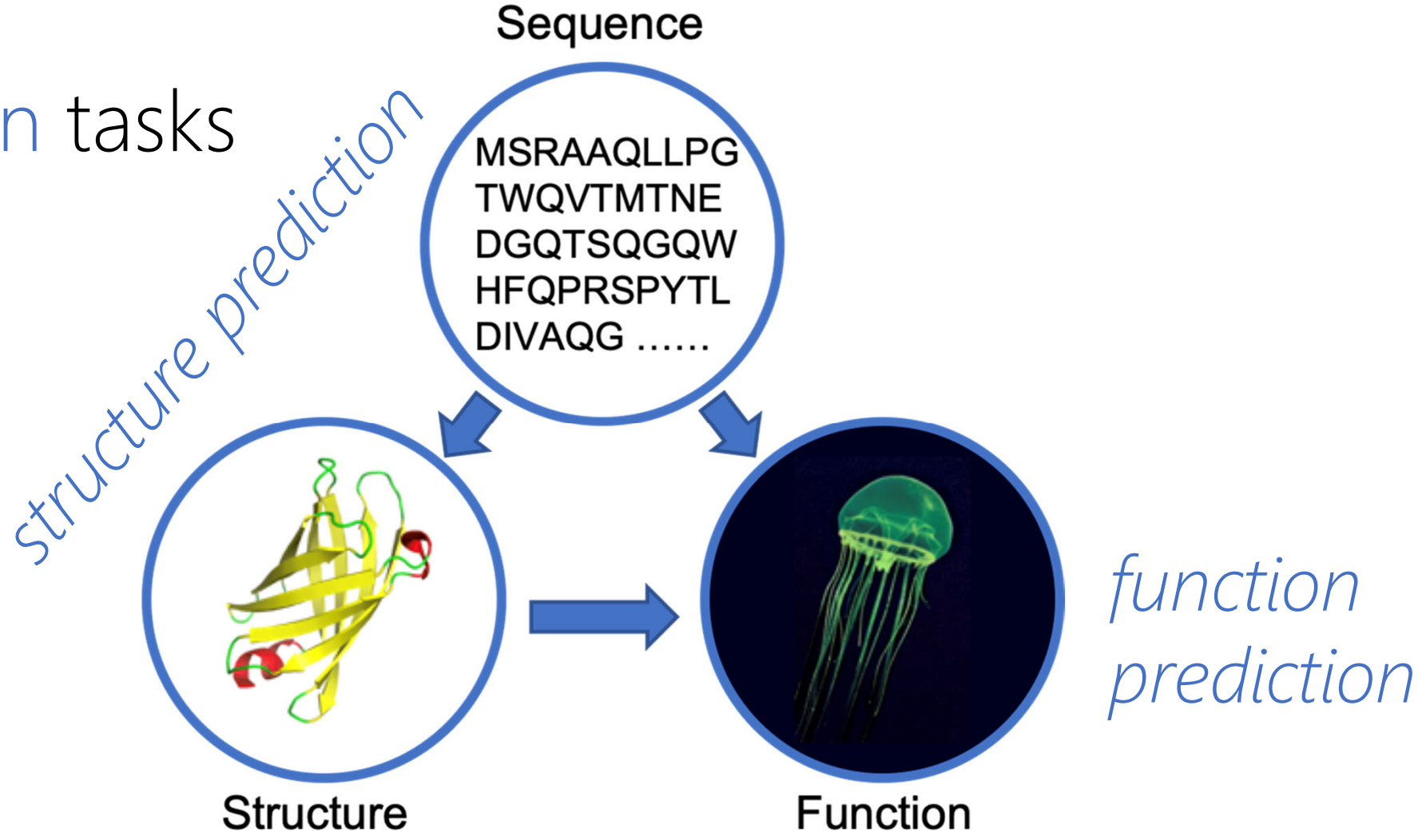
sequence→ structure



- No: don't typically know which protein structures we need.
- If did, would need: *structure→sequence*. (decent ML solutions exist).
- Bottleneck challenge: predict which proteinsjj have the function we desire—often extrapolatively.
- AlphaFold2 *was* a breakthrough, and is already useful.

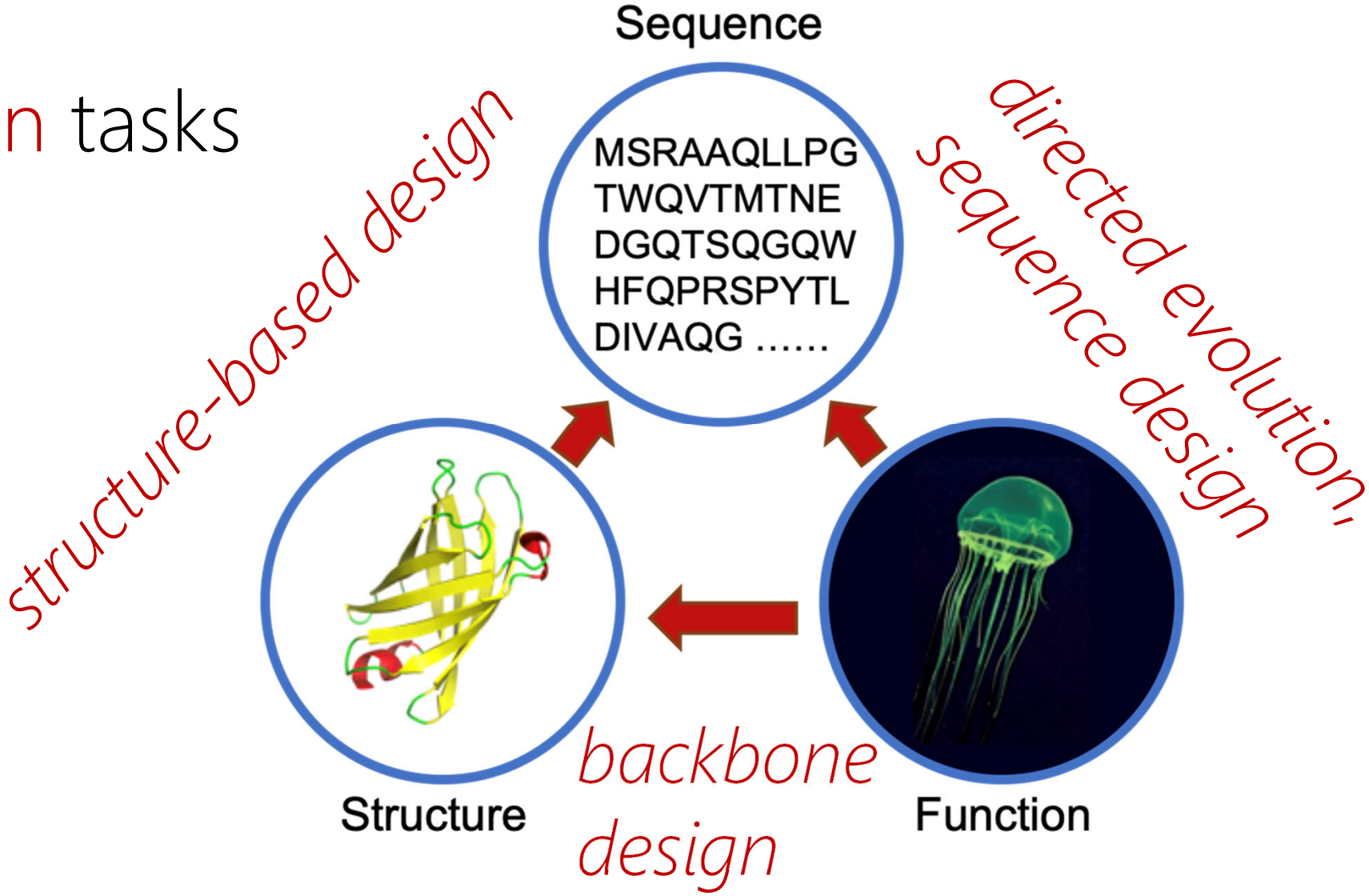
A suite of ML protein engineering problems

Prediction tasks



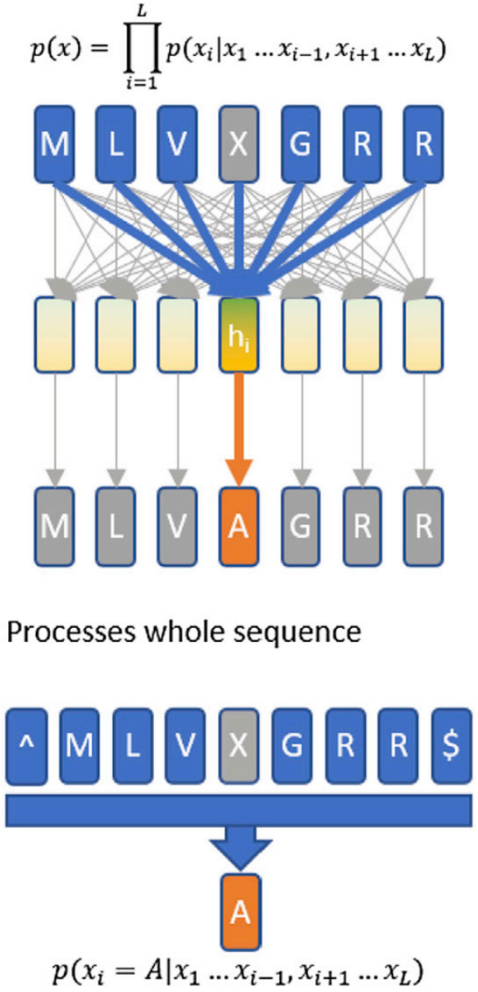
A suite of ML protein engineering problems

Design tasks



Some trends in ML + protein engineering

1. Representation learning: *un(self)supervised learning on* large-scale databases (millions of natural proteins, with *e.g.*, Transformers), or families.



- This is (approx.) *density estimation,* $p_\theta(sequence)$ through a bottleneck.

[Bepler *et al.*, *Cell Systems* 2021]

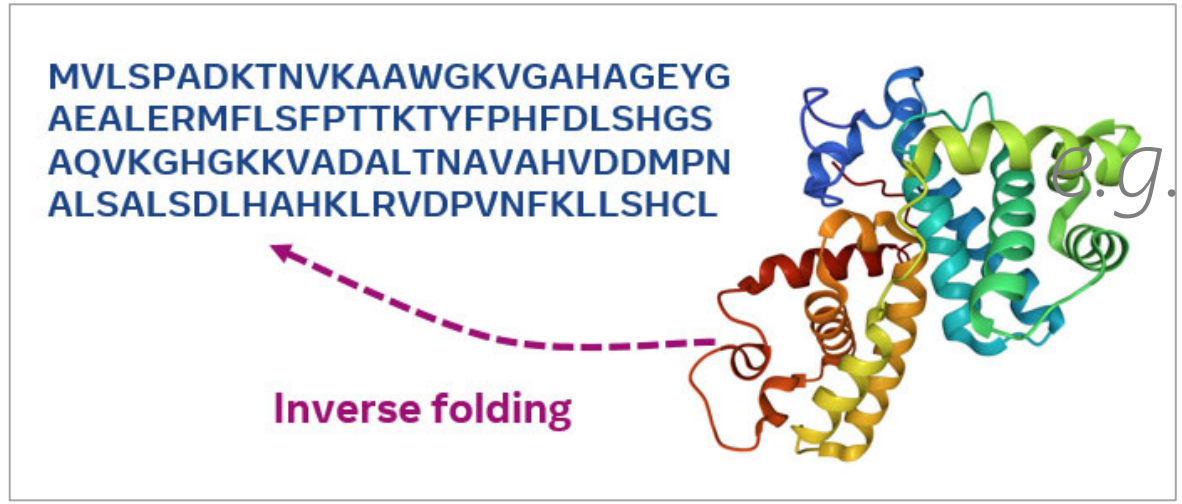
Some trends in ML + protein engineering

2. (Conditional) generative models for sequences.

This is (conditional) density estimation, $\boldsymbol{p_\theta(\text{sequence}|\mathbf{C})}$, (*e.g.* auto-regressive Transformer, Potts/VAE).

a) structure-conditioned,
   aka "inverse folding"

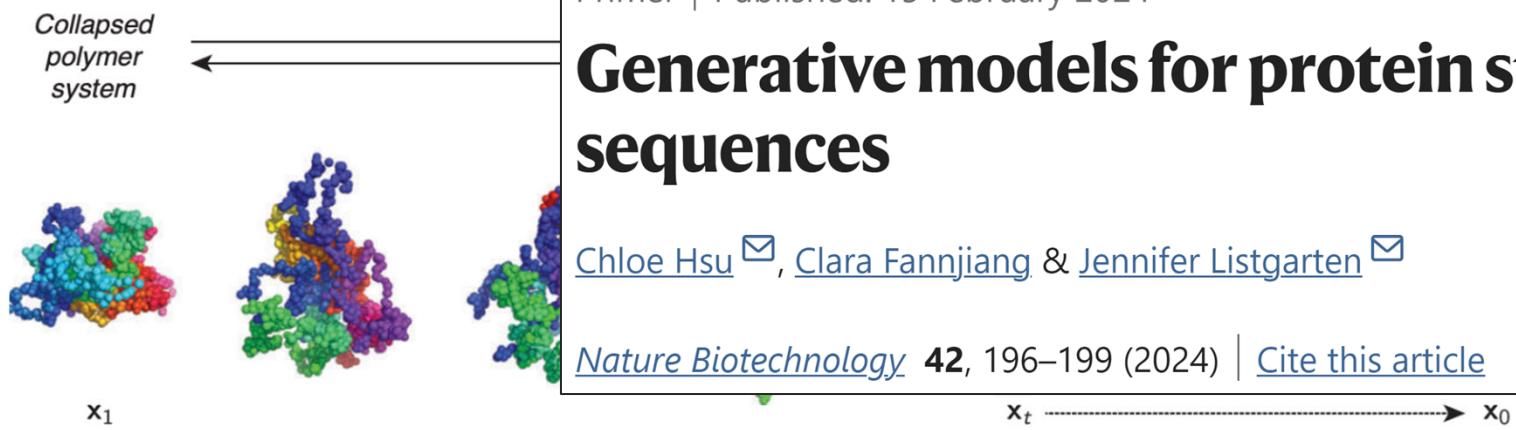
b) "control tag" conditioned,
   protein family



Some trends in ML + protein engineering

3. (Conditional) generative models for structure.

- This is (conditional) density estimation, $p_\theta(\mathbf{backbone}|\mathbf{F})$, (*e.g.* "Diffusion" models latest trend).
- Only as good as function prediction, $p(F|backbone)$.
- Paired with inverse-folding to get sequence



[Ingraham *et al. Nature* 20223

Primer | Published: 15 February 2024

Generative models for protein structures and sequences

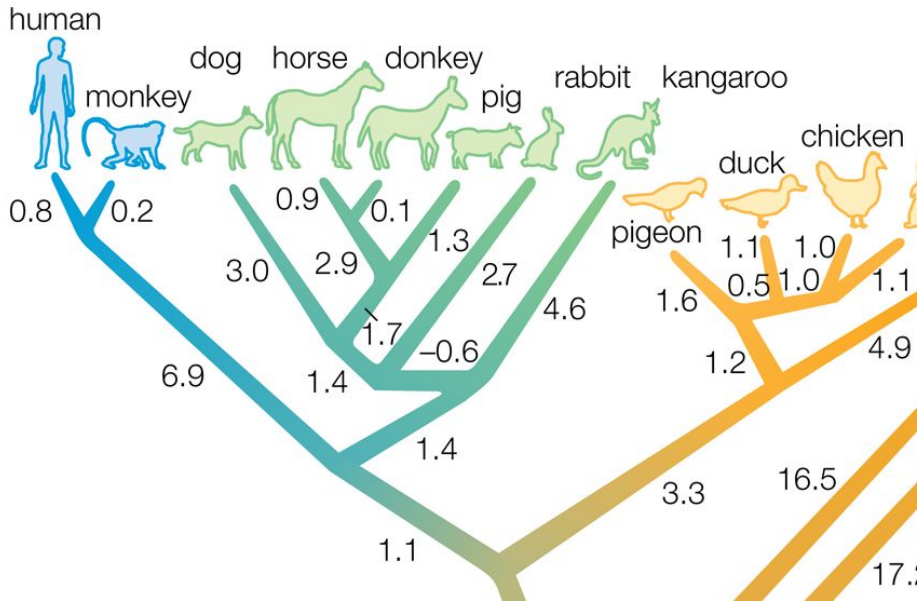
Chloe Hsu, Clara Fannjiang & Jennifer Listgarten

Nature Biotechnology **42**, 196–199 (2024) | Cite this article

Some trends in ML + protein engineering

4. ML to estimate function from sequence and/or function:

- *e.g.*, $p_\theta(F|sequence)$.
- Few or no labelled data.
- *Leverage evolutionary information**, or large unsupervised models on pan-proteomic database.

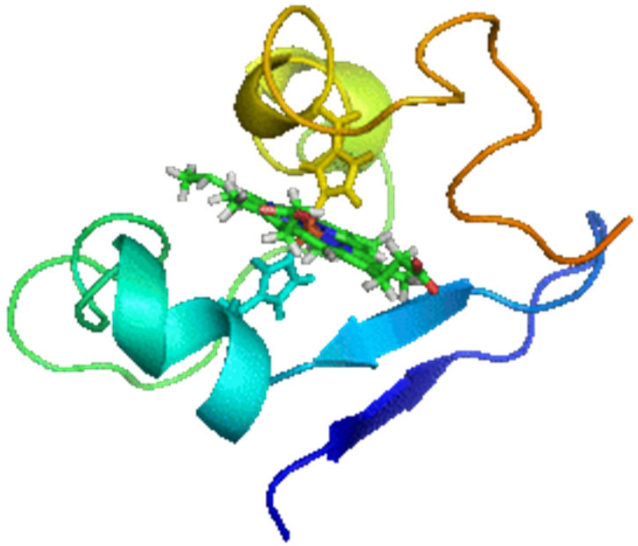


*key part of AlphaFold2/3

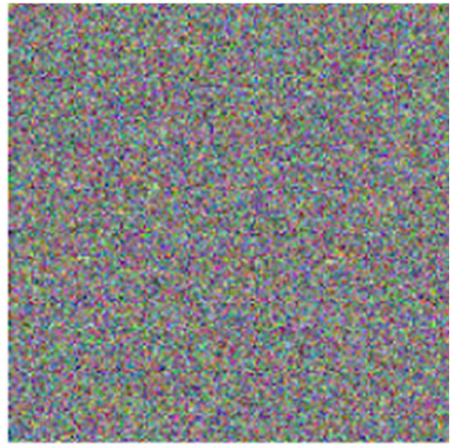
Some trends in ML + protein engineering

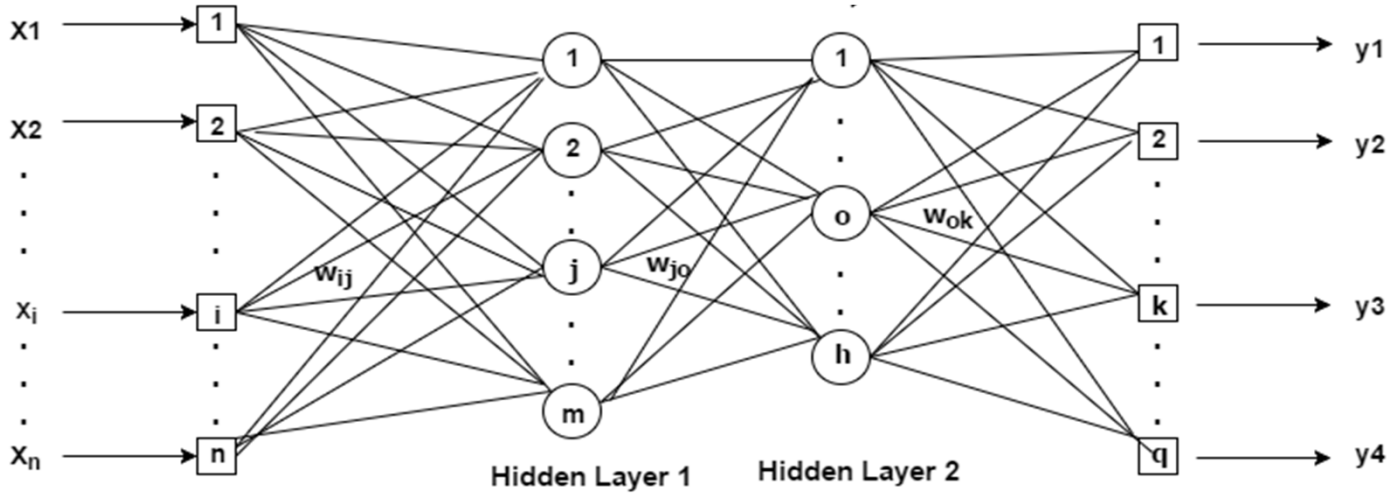
5. Structure prediction: filling the gaps left by AlphaFold2

- Orphan proteins (with *no/few homologs*).
- Protein-protein/DNA/RNA/small molecule binding.
- Protein dynamics and conformational distributions.



Analogy: can we trust “banana” design?

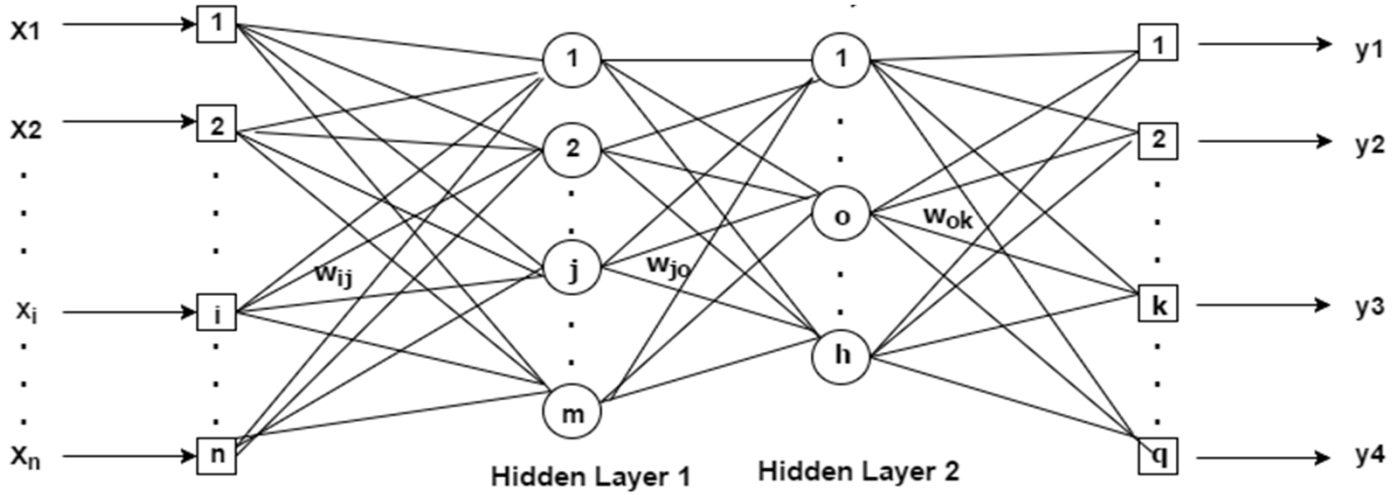




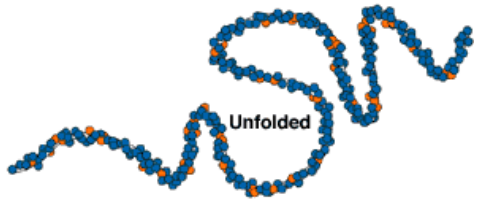
desired property

catalytic efficiency ↑

Naïve design yields abstract art ("pathology-finding").



desired property

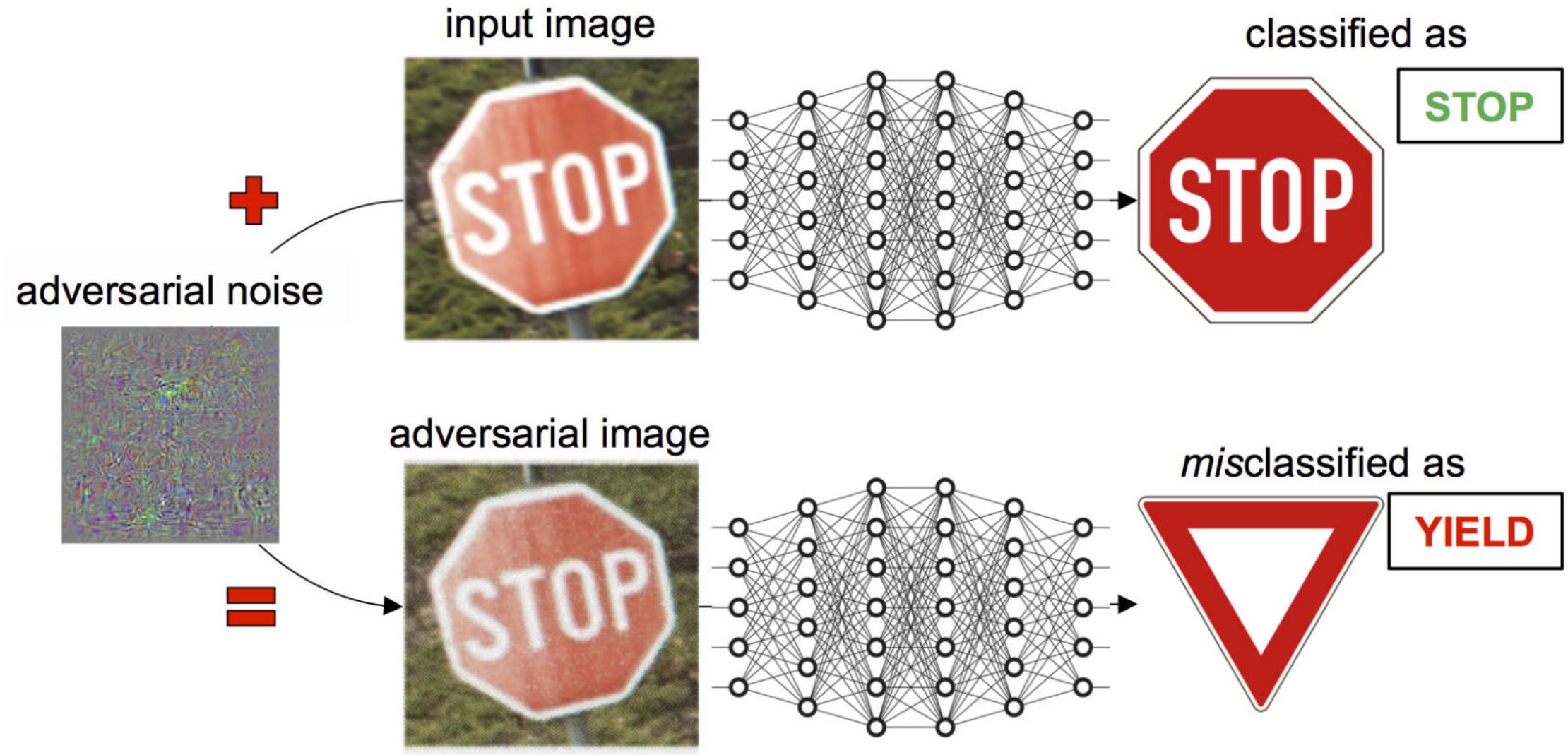


non-folding protein

catalytic efficiency ↑

1. Brookes *et al ICLM* 2019 (CbAS)
2. Fannjiang *et al NeurIPS* 2020 (autofocus)

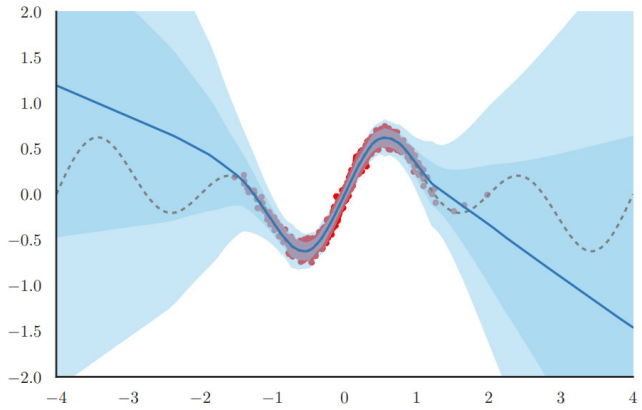
Pathologies of DNNs: in design, we're the adversary



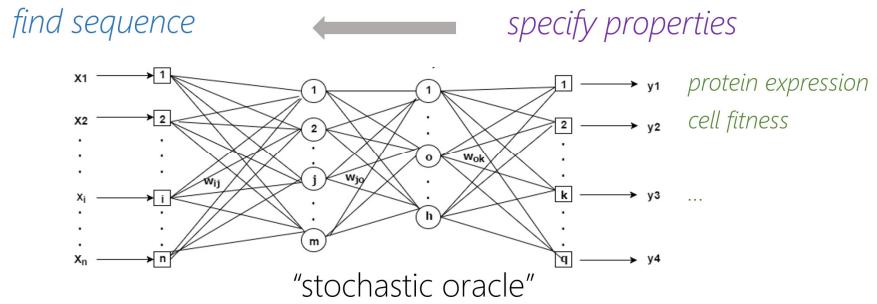
https://www.pluribus-one.it/research/sec-ml/wild-patterns

ML-based design challenges tackled in our group

1. A natural tension between leveraging the trained model for extrapolation, vs knowing that the model is not trustworthy in many areas of protein space (related to causality) [1,2].

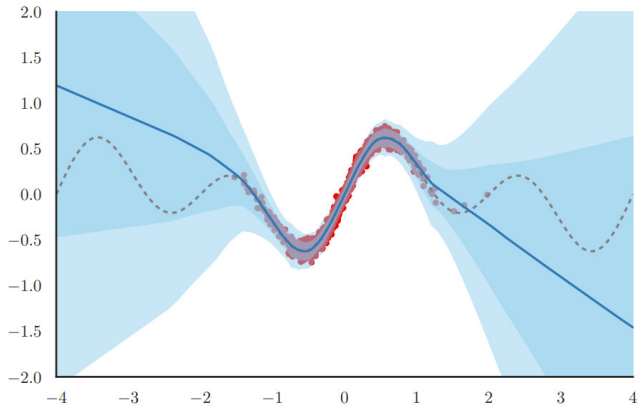


1. Brookes *et al* ICLM 2019 (CbAS)
2. Fannjiang *et al* NeurIPS 2020 (autofocus)

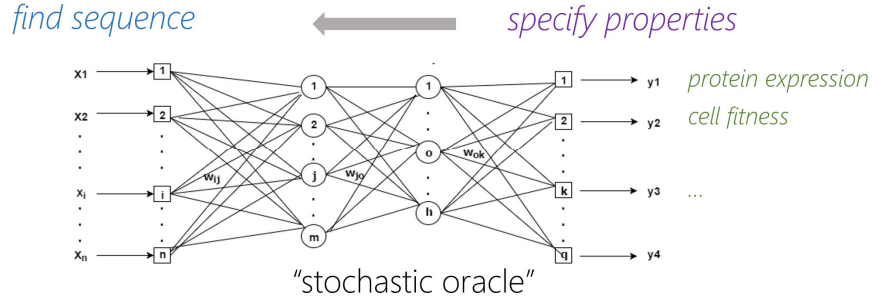


ML-based design challenges tackled in our group

1. A natural tension between leveraging the trained model for extrapolation, vs knowing that the model is not trustworthy in many areas of protein space (related to causality) [1,2].
2. Also related to estimation of *epistemic* uncertainty (whereas we typically think mostly of *aleotoric*) uncertainty [3, 4].

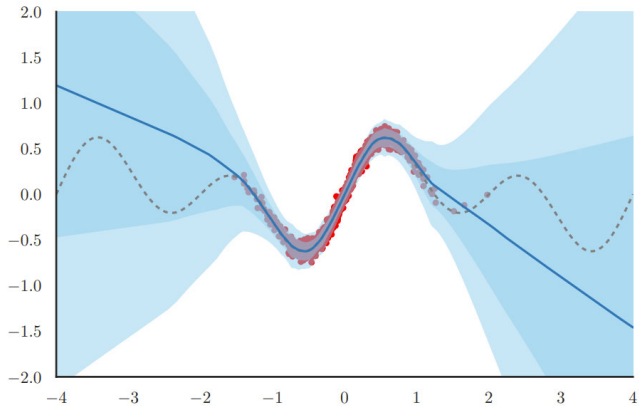


1. Brookes *et al* ICLM 2019 (CbAS)
2. Fannjiang *et al* NeurIPS 2020 (autofocus)
3. Nisinoff *et al* ACS Synth Bio 2023 (fv-BNN)
4. Fannjiang *et al* PNAS 2023(conformal)

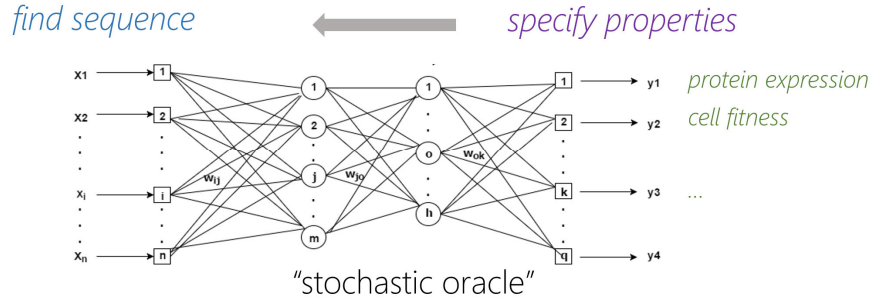


ML-based design challenges tackled in our group

1. A natural tension between leveraging the trained model for extrapolation, vs knowing that the model is not trustworthy in many areas of protein space (related to causality) [1,2].
2. Also related to estimation of *epistemic* uncertainty (whereas we typically think mostly of *aleotoric*) uncertainty [3,4].
3. Suitable protein inductive biases when using neural networks [3,5,6,7].



1. Brookes *et al* ICLM 2019 (CbAS)
2. Fannjiang *et al* NeurIPS 2020 (autofocus)
3. Nisinoff *et al* ACS Synth Bio 2023 (fv-BNN)
4. Fannjiang *et al* PNAS 2023(conformal)
5. Aghazadeh *et al* Nat. Comm. 2021
6. Brookes *et al* PNAS 2022
7. Hsu *et al* Nat. Biotech. 2022



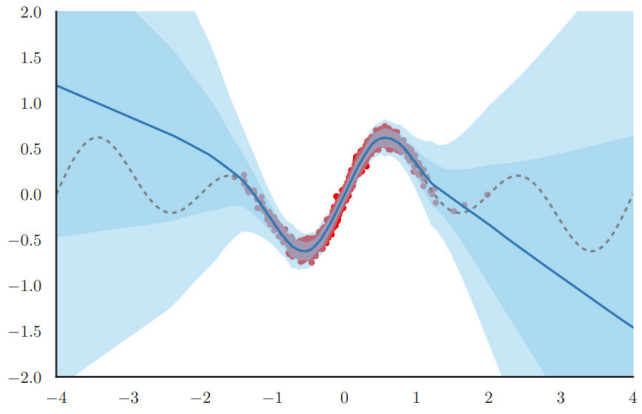
ML-based design challenges tackled in our group

1. A natural tension between leveraging the trained model for extrapolation, vs knowing that the model is not trustworthy in many areas of p
2. Also relate we typically t
3. Suitable p [3,5,6,7].
4. Design of distributions instead of individual sequences [1,2,8].

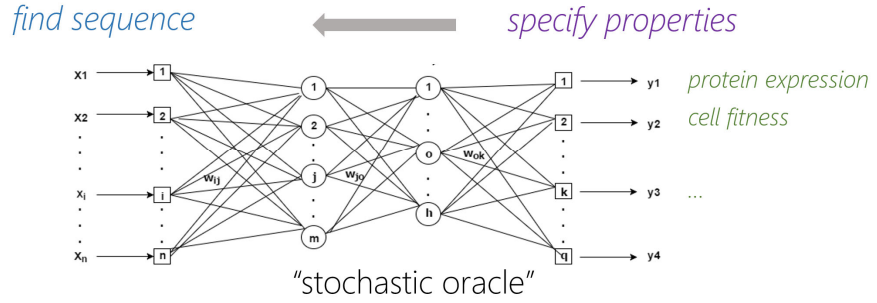
Is Novelty Predictable?

Clara Fannjiang and Jennifer Listgarten

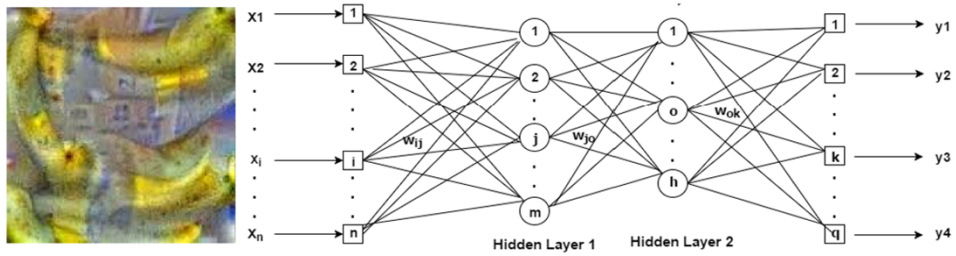
Cold Spring Harb Perspect Biol ***2023***



1. Brookes *et al* ICLM 2019 (CbAS)
2. Fannjiang *et al* NeurIPS 2020 (autofocus)
3. Nisinoff *et al* ACS Synth Bio 2023 (fv-BNN)
4. Fannjiang *et al* PNAS 2023(conformal)
5. Aghazadeh *et al* Nat. Comm. 2021
6. Brookes *et al* PNAS 2022
7. Hsu *et al* Nat. Biotech. 2022
8. Zhu, Brookes *et al* Science Advances 2024



Conditioning by Adaptive Sampling for Robust Design (CbAS)

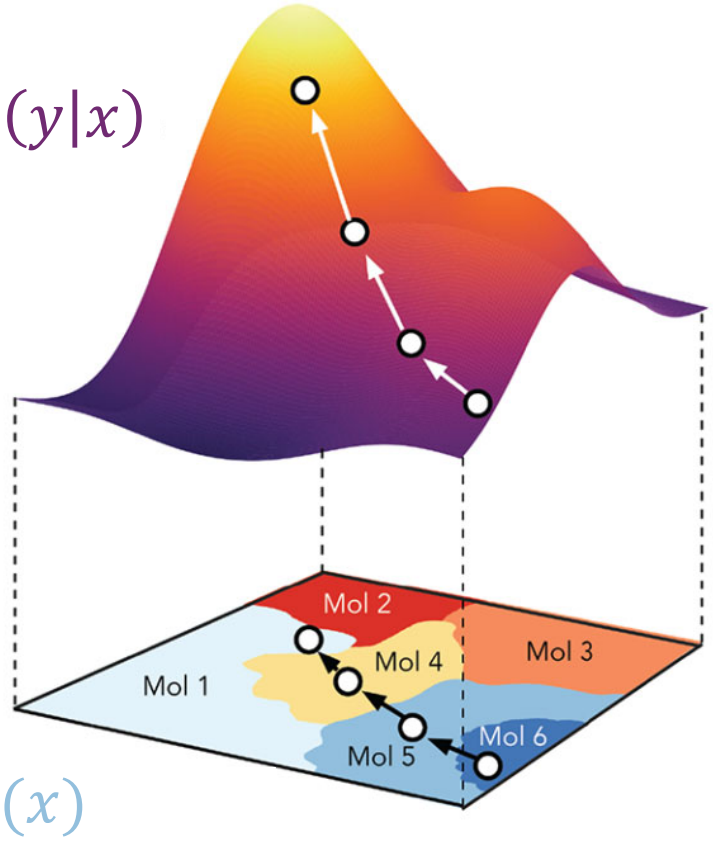


desired property

How to handle a pathology in design?

Leverage prior knowledge, $\boldsymbol{p(x)}$, by modeling:

1. Where training data lie.
2. "Protein-likeness", e.g. stability via biophysics, or implicitly via large pan-proteome unsupervised models.



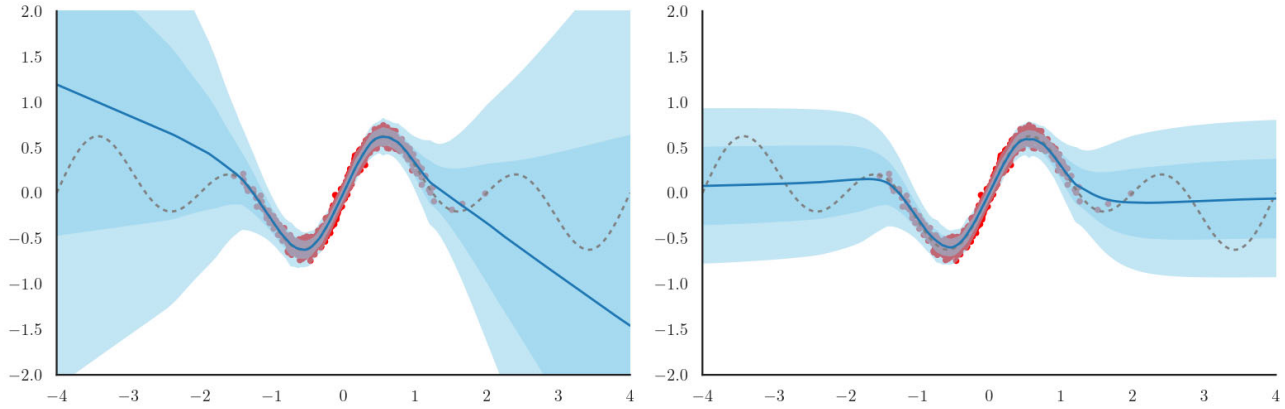
[Gomez-Bombarelli, ACS Cent. Sci. 2018.]

Brookes, Park & Listgarten *ICML* 2019

Augmenting Neural Networks with Priors on Functional Values

Coherent blending of function value prior information, such as biophysical models, to Bayesian Neural Networks (BNN).

Easy to implement, zero added cost.



regular BNN

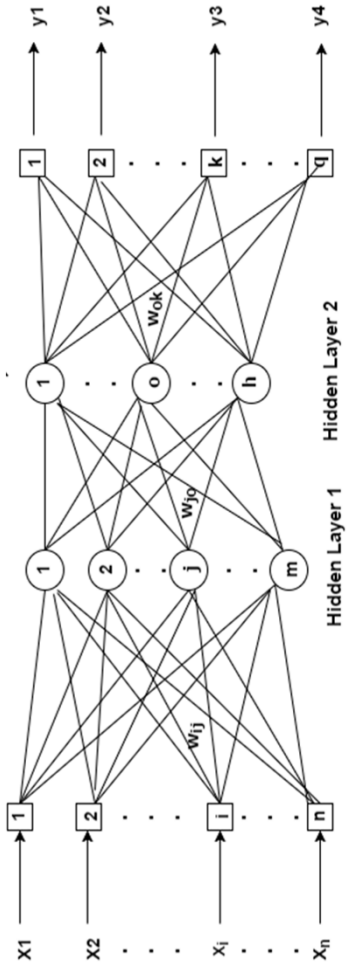
function-value augmented BNN

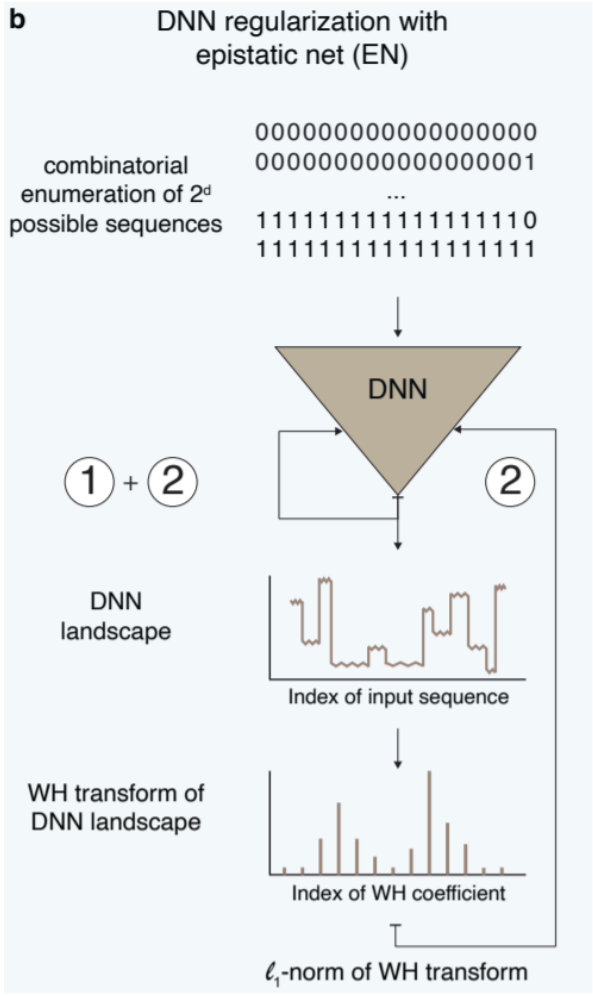
METHOD	LOG-LIKELIHOOD
NN	-8.33 ± 0.66
BNN	-5.73 ± 0.18
STACKING: BNN+NON-FUNCTIONAL PRIOR	-8.63 ± 0.33
STACKING: BNN+STABILITY PRIOR	-8.61 ± 0.34
fv-BNN (NON-FUNCTIONAL PRIOR)	-1.82 ± 0.00
fv-BNN (STABILITY PRIOR)	-1.53 ± 0.00

Hunter Nisonoff

Nisonoff, Wang, Listgarten, *bioRxiv*

Sparse Epistatic Networks



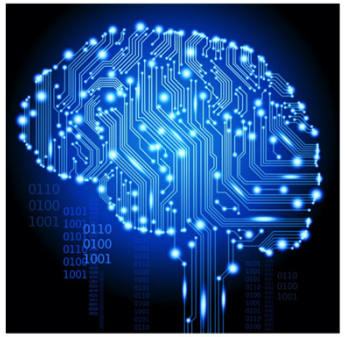


- Inject suitable inductive biases for protein sequence functions.
- *i.e.* sparsity in "epistatic" terms (aka Walsh-Hadamard basis of features).

Aghazadeh*, Nisonoff* *et al*, Nat Comm 2020

The real deal: testing+developing our ideas with wetlab collaborators

- David Schaffer (UC Berkeley; AAV for gene therapy)
- David Savage (UC Berkeley; CRISPR-Cas9 system)
- Chris Garcia (Stanford, protein-protein interactions)
- Phil Romero (U Wisconsin; enzymes for plastic degradation)
- Secure and Robust Biosystems Design Group (LL National Labs, Columbia University, University of Maryland, University of Minnesota)



+

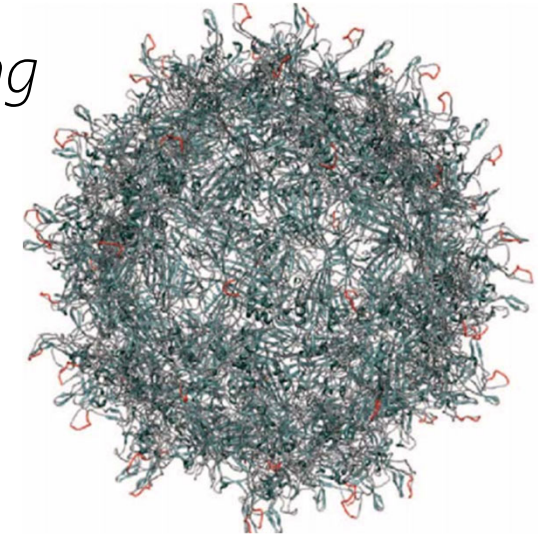
Engineering AAV for gene therapy delivery

The Adeno-associated virus (AAV) is a non-pathogenic virus that shows promise for delivering gene therapies (*e.g.* deliver blindness therapy to outer retina).

UC Berkeley: Chem. & Bio. Engineering

David Schaffer

Bonnie Zhu



David Brookes
(now at Dyno)

Akosua Busia
(now at Dyno)

Zhu, Brookes, Busia,..., Nowakowski, Listgarten, Schaffer, *Science Advances 2024*

Promising AAV clinical trials

Recent clinical trial success:

Leber's congenital amaurosis (AAV)

Spinal muscular atrophy (AAV)

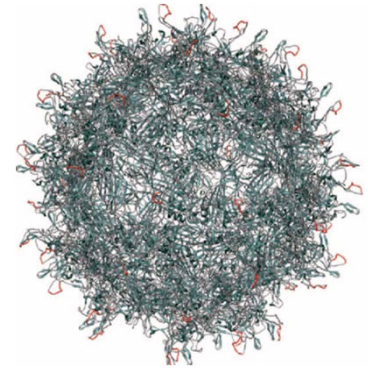
Hemophilia B (AAV)

Lipoprotein lipase deficiency (AAV)

Many diseases targets are still beyond the reach of current gene delivery technology

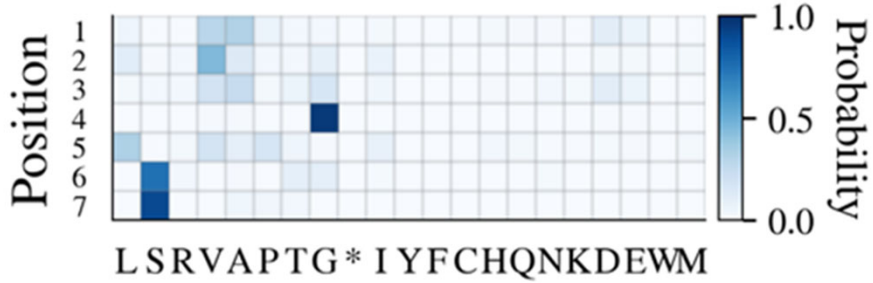
Ongoing challenges for AAV-based therapeutics

- Inefficient delivery to target tissues/cells.
- Non-specific delivery.
- Pre-existing immunological neutralization.
- Inefficient uptake into target cells.

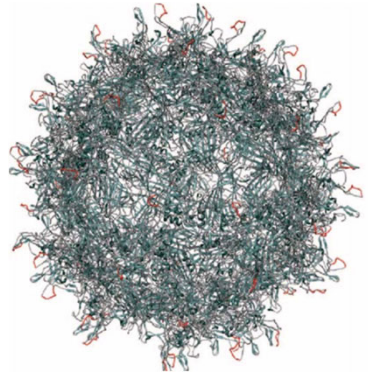


First AAV project goal, "library design":

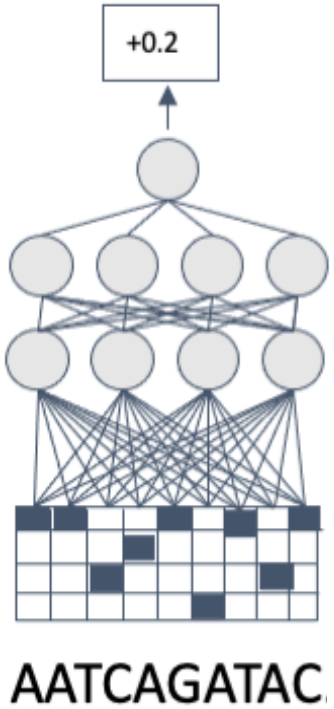
- Obtain optimal starting "library" for all these engineering goals.
- *i.e.*, large fraction of library gets wasted because doesn't *package*.

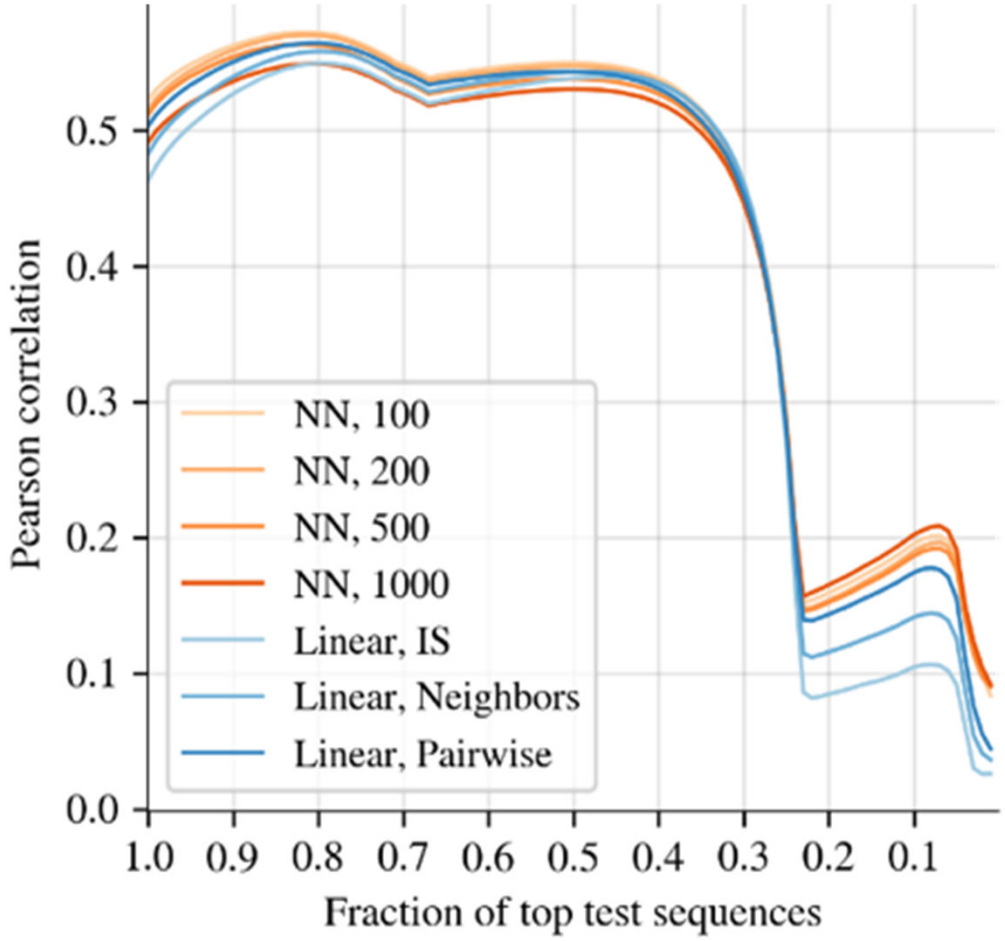


AAV library design

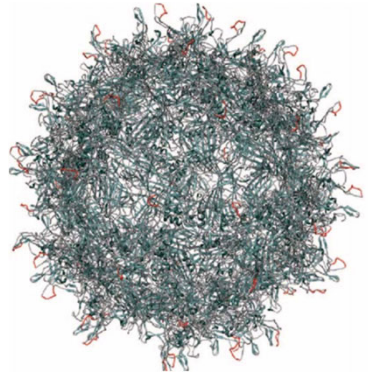


1. Build predictive model and test (*sequence→packaging* fitness).

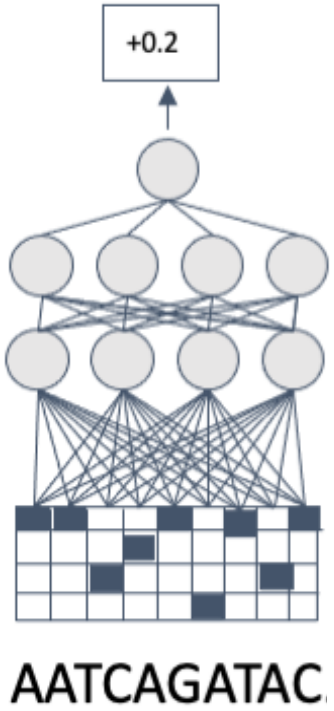


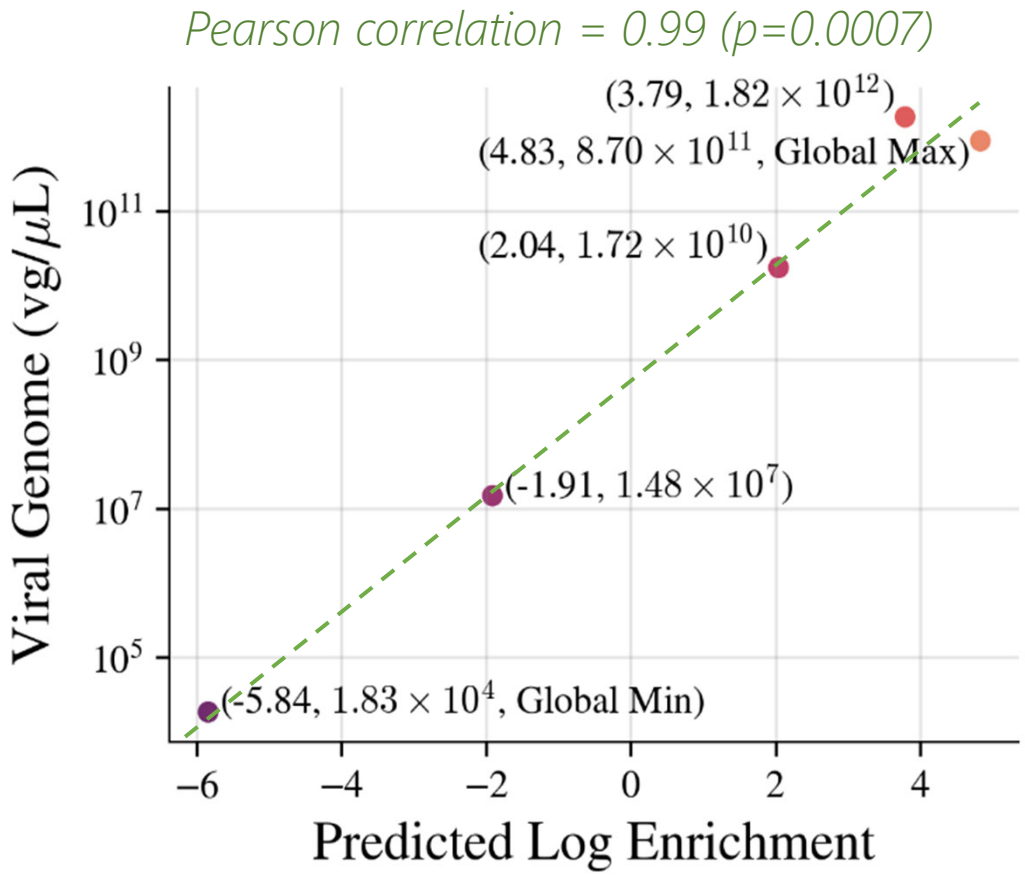


AAV library design



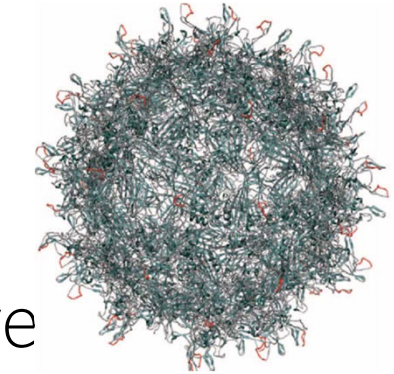
2. Wetlab validate model (measure titer directly)





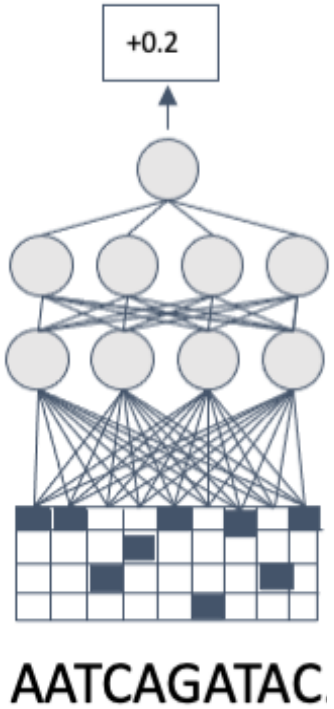
Sequences	Predicted Log Enrichment	Experimental Viral Titer (vg/μL)
LSSTTAA	4.834	8.70×10^{11}
DSRLSGT	3.793	1.82×10^{12}
LEPDAAL	2.044	1.72×10^{10}
IRWRATG	(-) 1.91	1.48×10^{7}
RWPRRVL	(-) 5.84	1.83×10^{4}

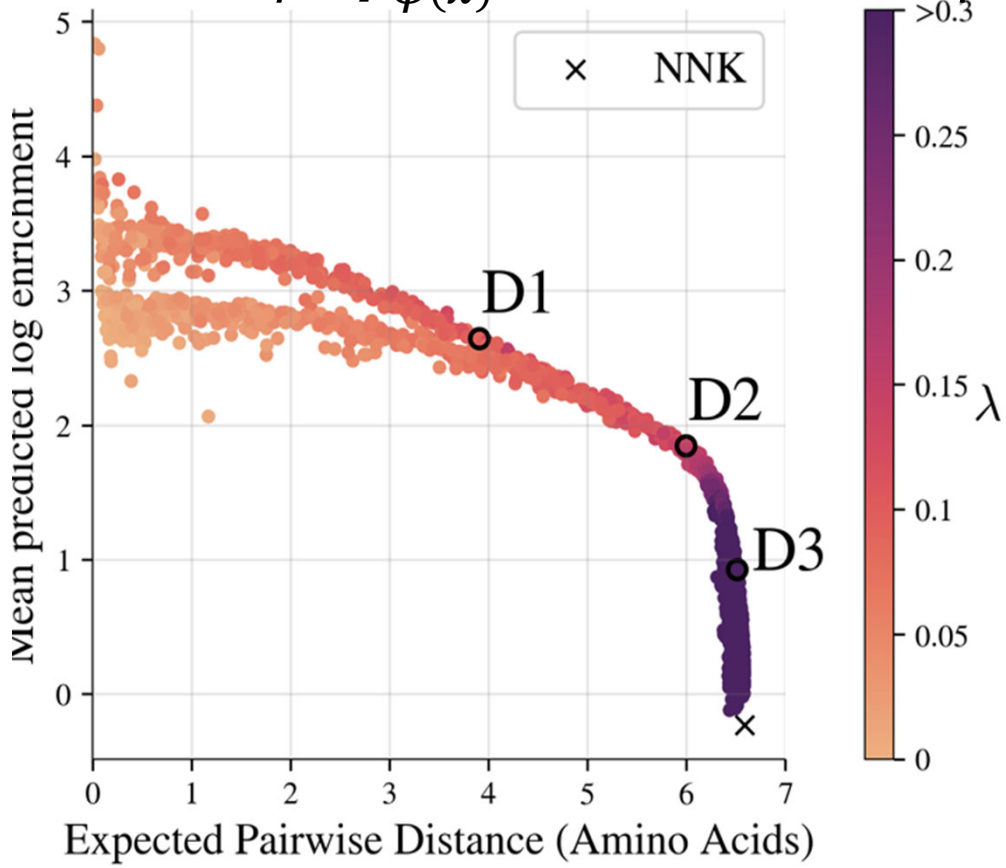
AAV library design



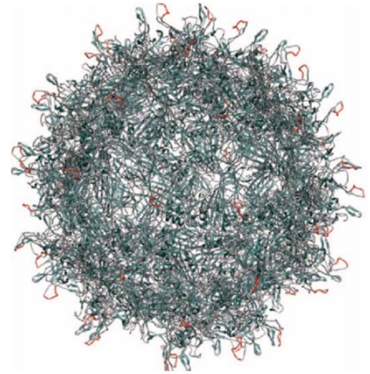
3. Invert ML predictive model to get diversity-fitness optimality curve

$$argmax_{\phi}\mathbb{E}_{p_{\phi(x)}}[f(x)] + \lambda H[p_{\phi}]$$

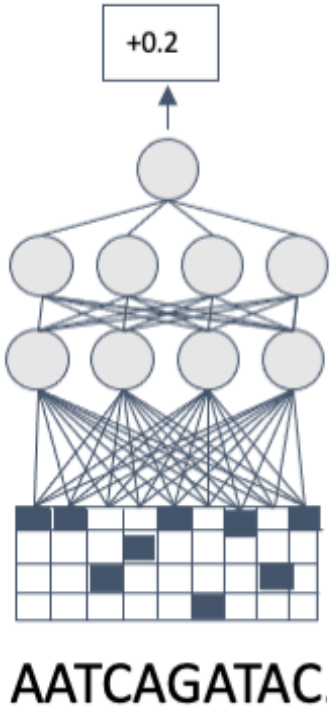




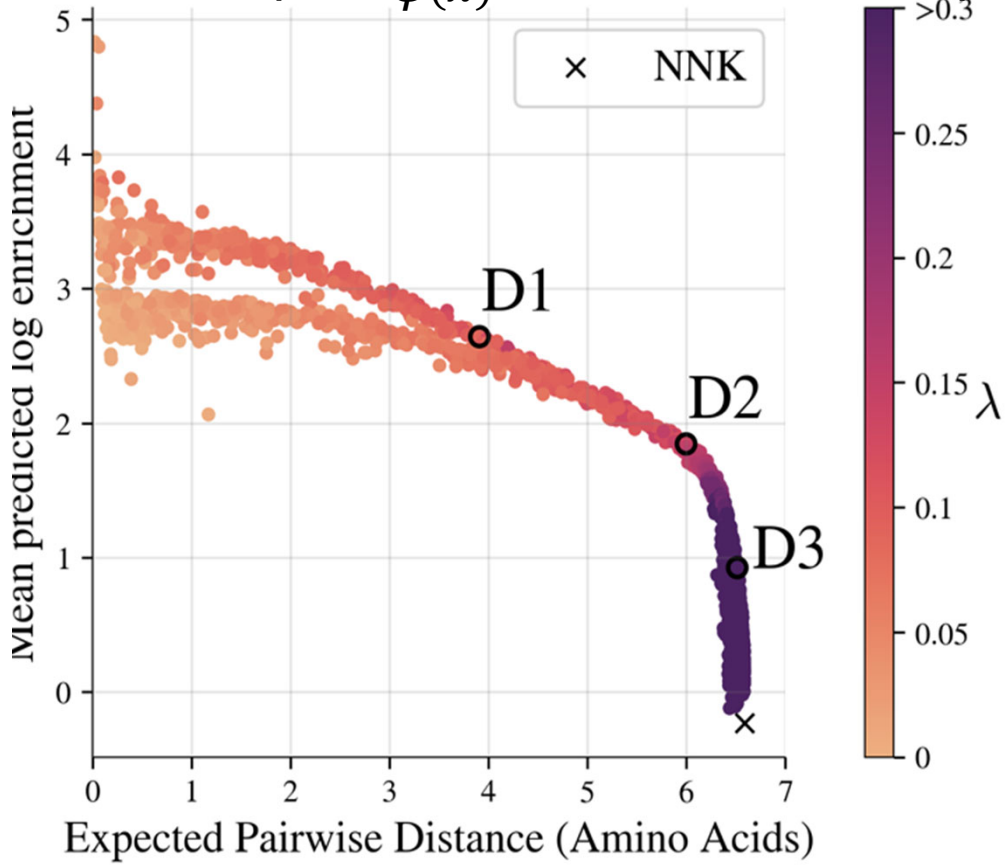
AAV library design



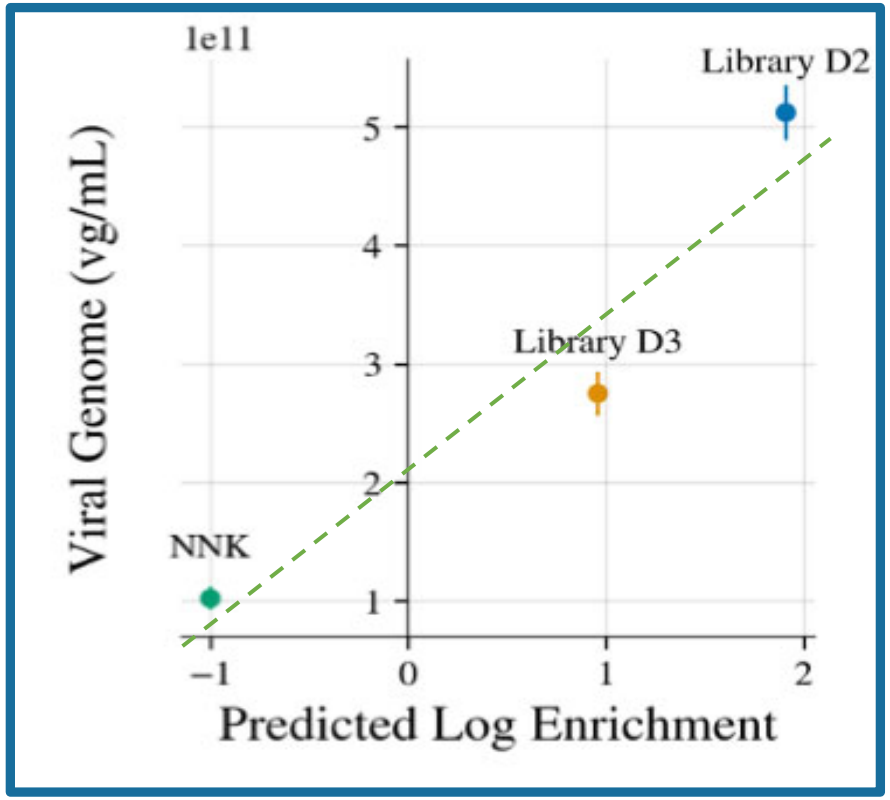
4. Validate in the lab.



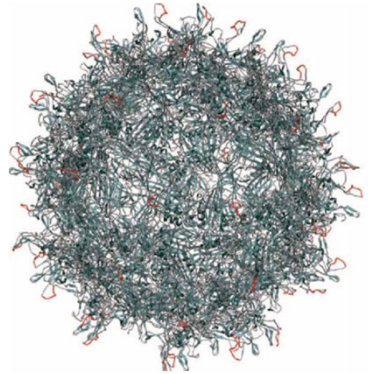
$$argmax_{\phi} \mathbb{E}_{p_{\phi(x)}}[f(x)] + \lambda H[p_{\phi}]$$



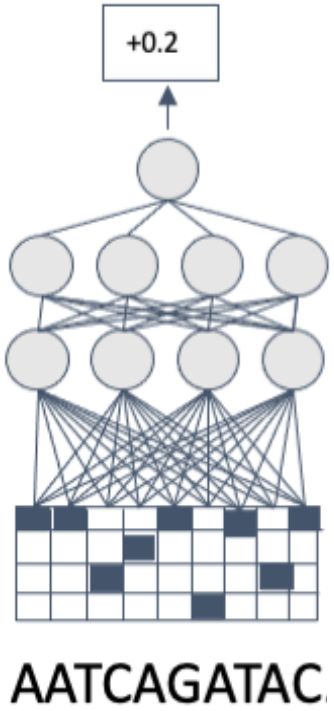
Pearson correlation = 0.96 (p=0.001)



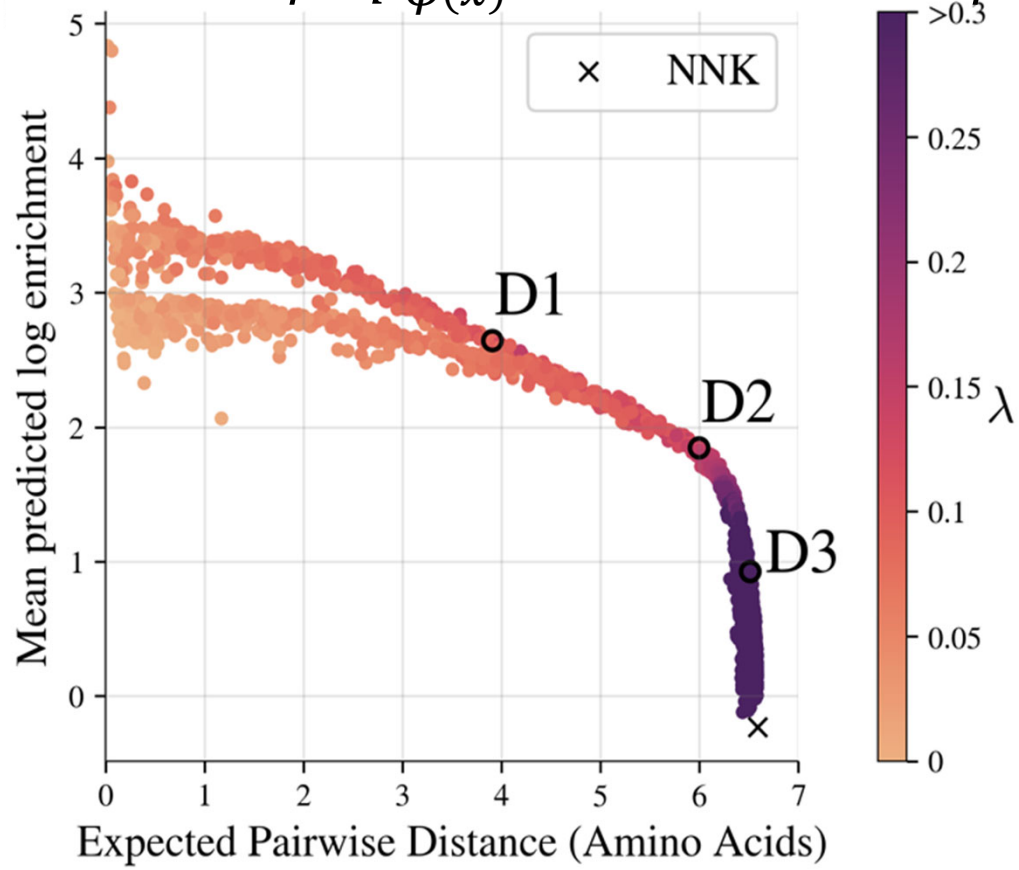
AAV library design



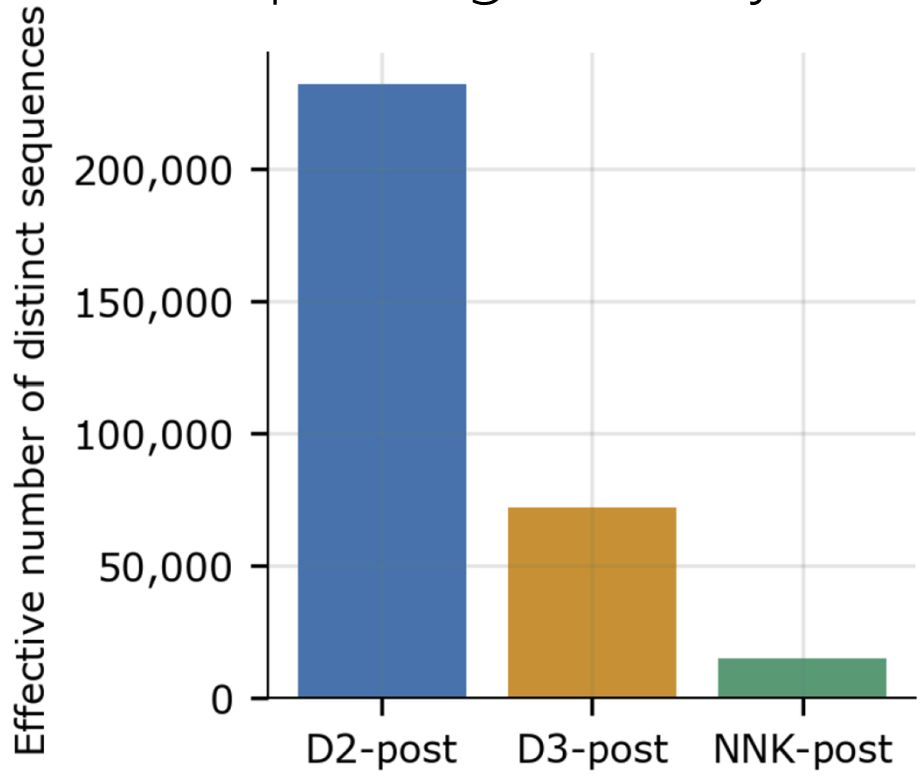
4. Validate in the lab.



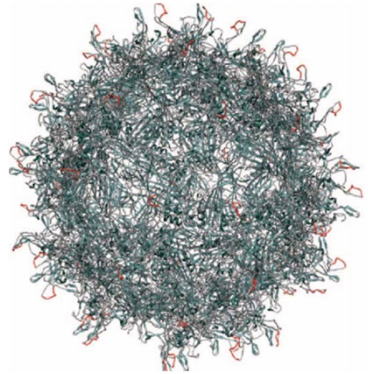
$$argmax_{\phi} \mathbb{E}_{p_{\phi(x)}}[f(x)] + \lambda H[p_{\phi}]$$



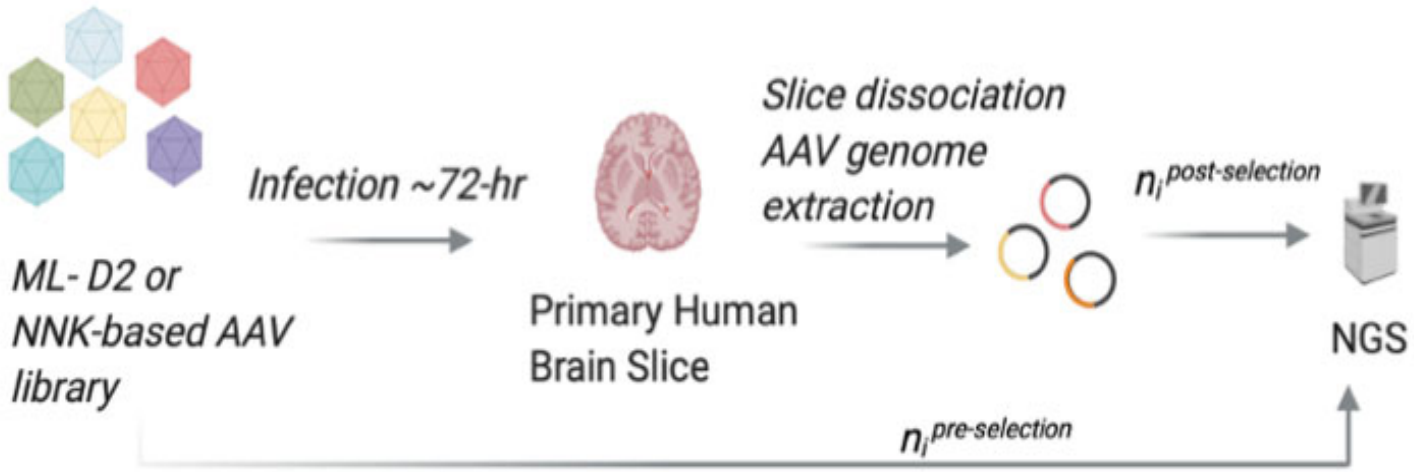
Post-packing diversity

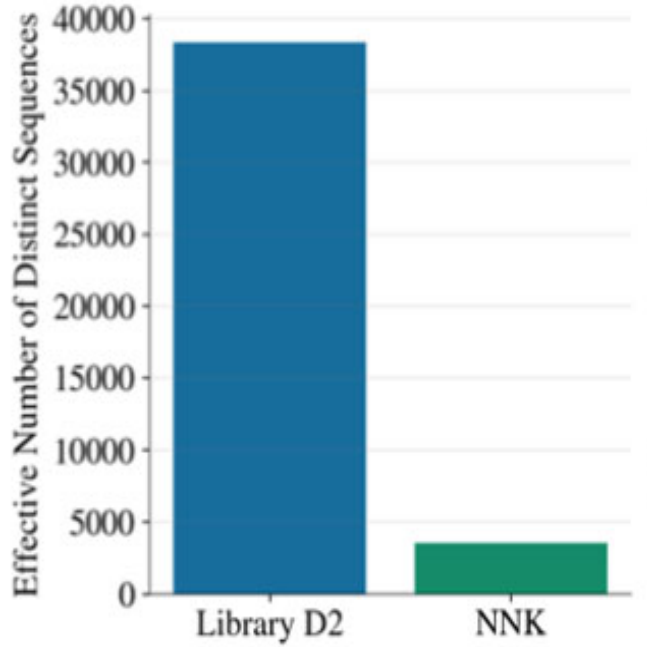


AAV library design



5. Demonstrate better downstream selection (human brain cell infectivity), that it *was not specifically designed for.*





ML library

currently used library