CS 189/289

Some applications of AI in biology:

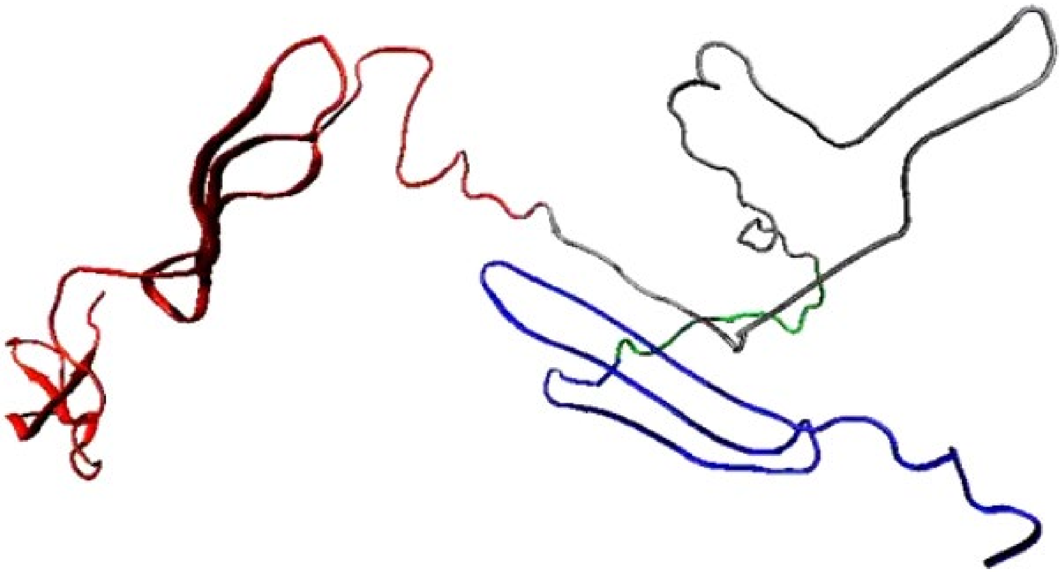
1. protein structure prediction
2. protein design

CS 189/289

Some applications of AI in biology:

1. protein structure prediction
2. protein design

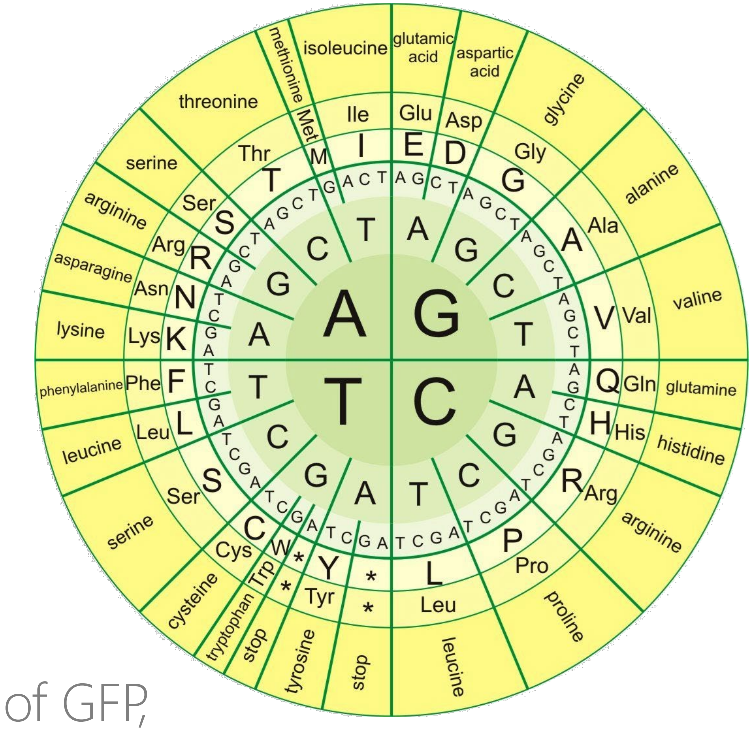
Proteins are strings of nucleotides

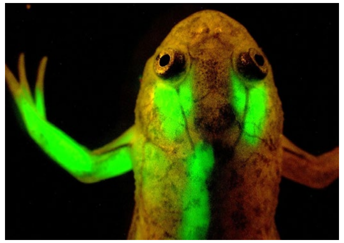


Green fluorescent protein (GFP) folding itself

238 length amino acid sequence:

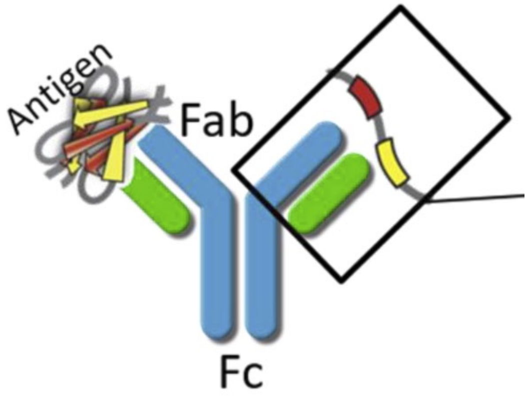
MSKGEELFTGVVPILVELDGDVNGHKFSVSG
EDFFKS…NSHNVYIMADKQKNGIKVNFKIRH



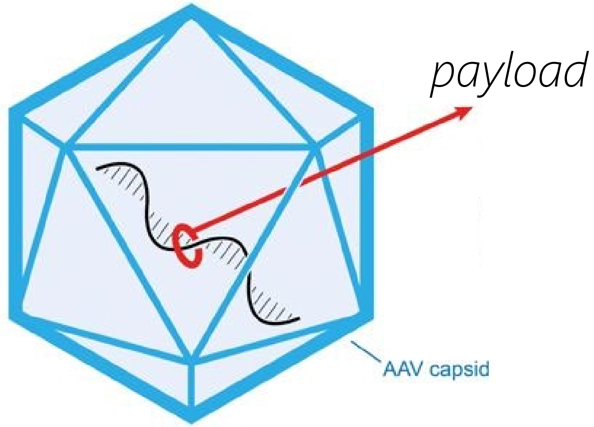


[2008 Nobel in chemistry for discovery and development of GFP,
Osamu Shimomura, Martin Chalfie and Roger Y. Tsien]

Protein engineering: therapeutics, environment, *etc.*

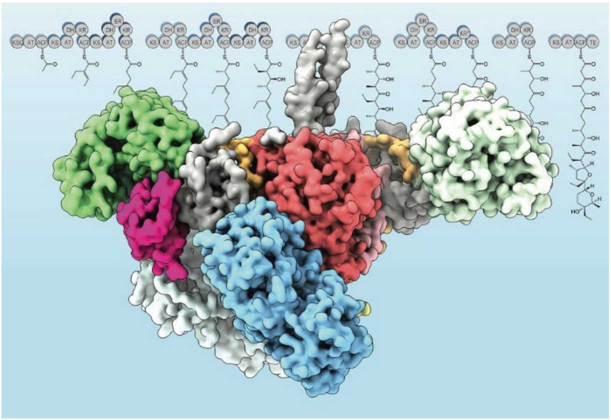


antibody therapeutics

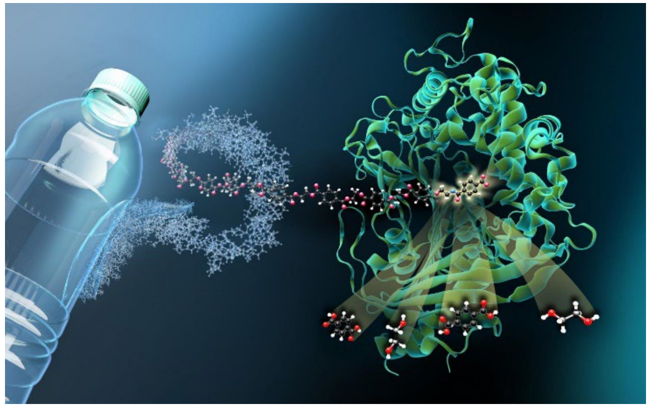


gene therapy virus delivery (AAV)

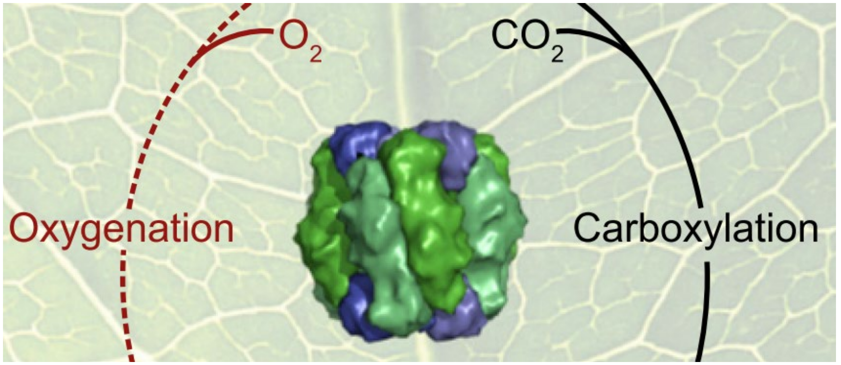
gene editing (CRISPR/Cas9)



antibiotics & biofuel production (PKS)



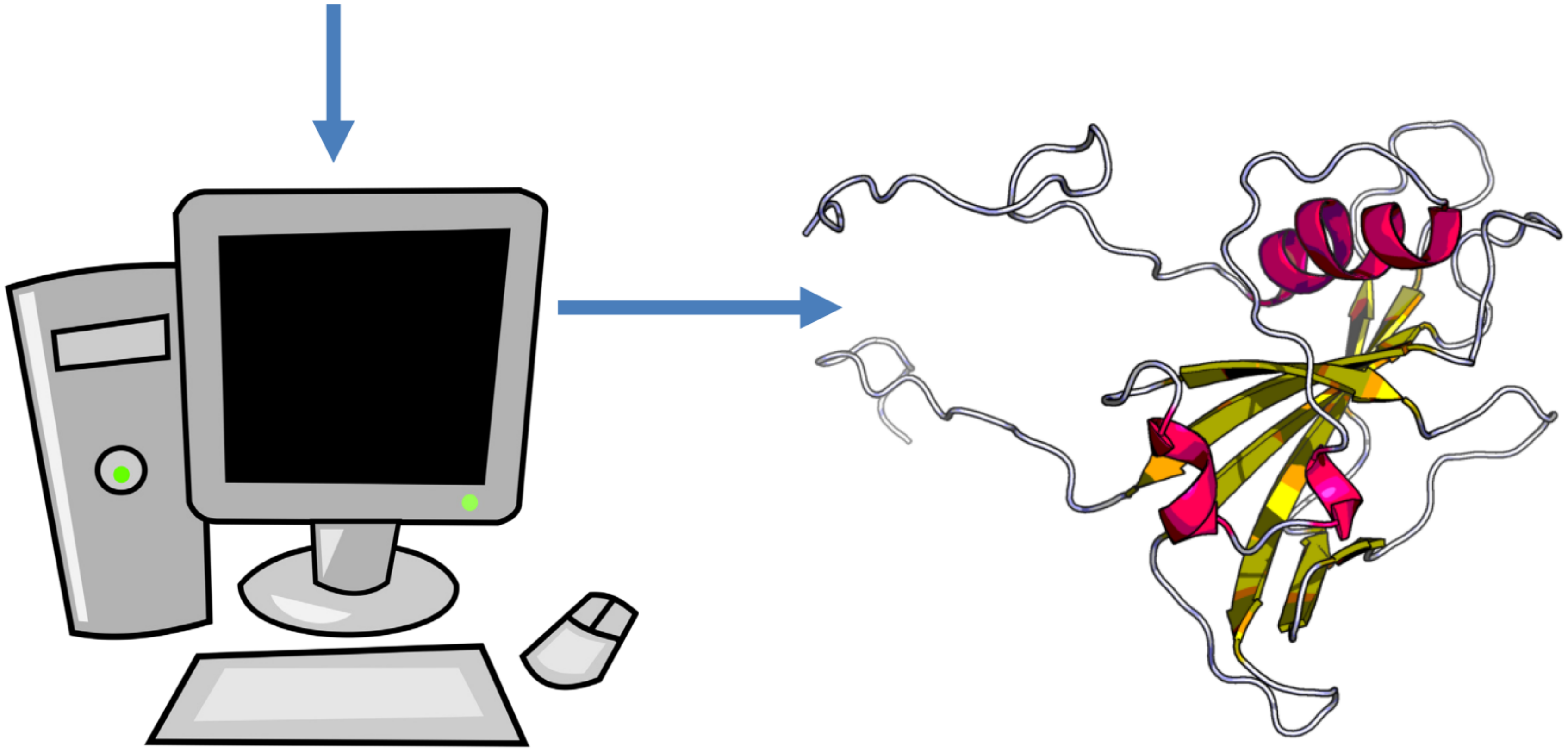
plastic recycling (PETase)



CO_2 biosequestration (RuBisCO)

Protein Structure Prediction

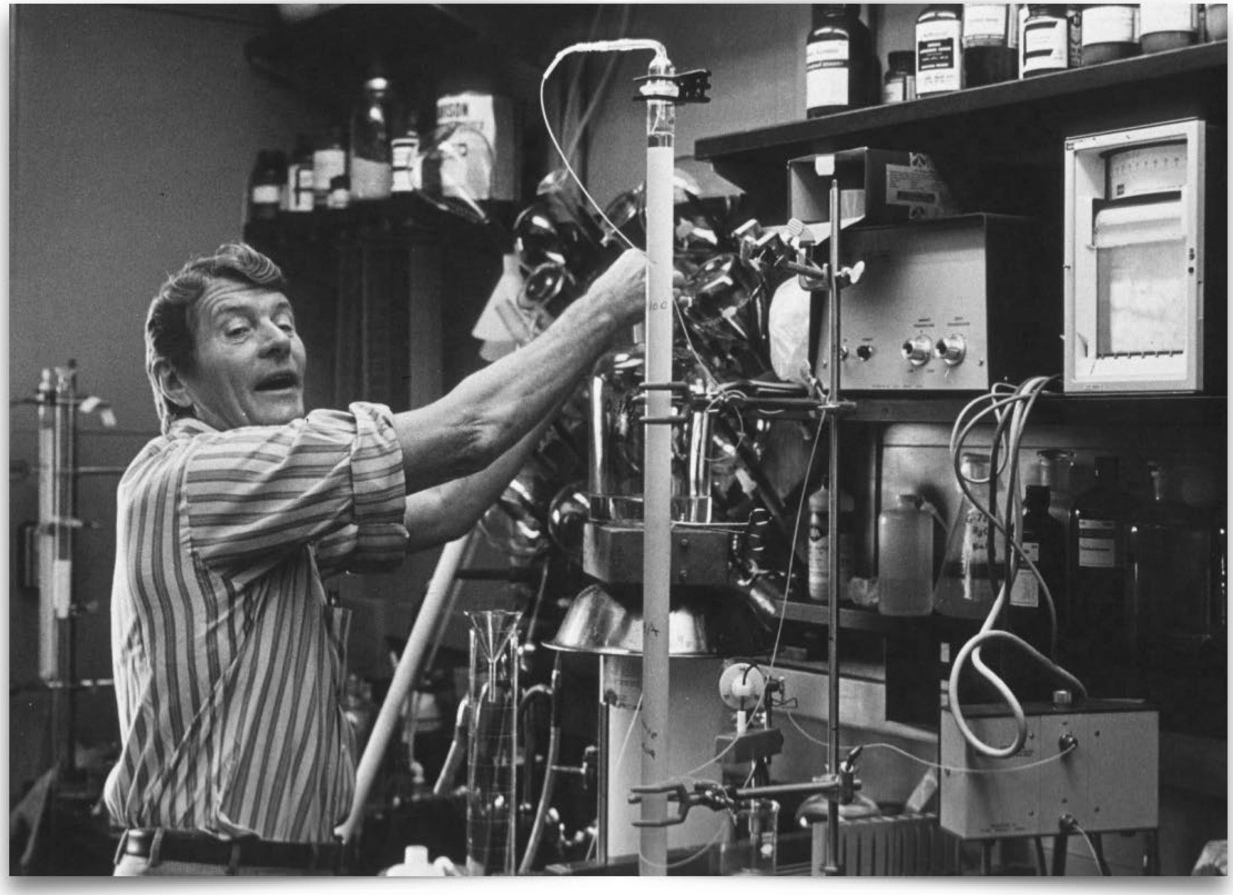
MEKVNFLKNGVLRLPPGFRFRPTDEELVVQYLKRKVFSFPLPASIIPEVEVYKSDPWDLPGDMEQEKYFFSTK
EVKYPNGNRSNRATNSGYWKATGIDKQIILRGRQQQQLIGLKKTLVFYRGKSPHGCRTNWIMHEYRLAN
LESNYHPIQGNWVICRIFLKKRGNTKNKEENMTTHDEVRNREIDKNSPVVSVKMSSRDSEALASANSELKK



Has been studied several decades

[slide from Jinbo Xu, TTI]

Amino acid sequence determines protein 3D structure

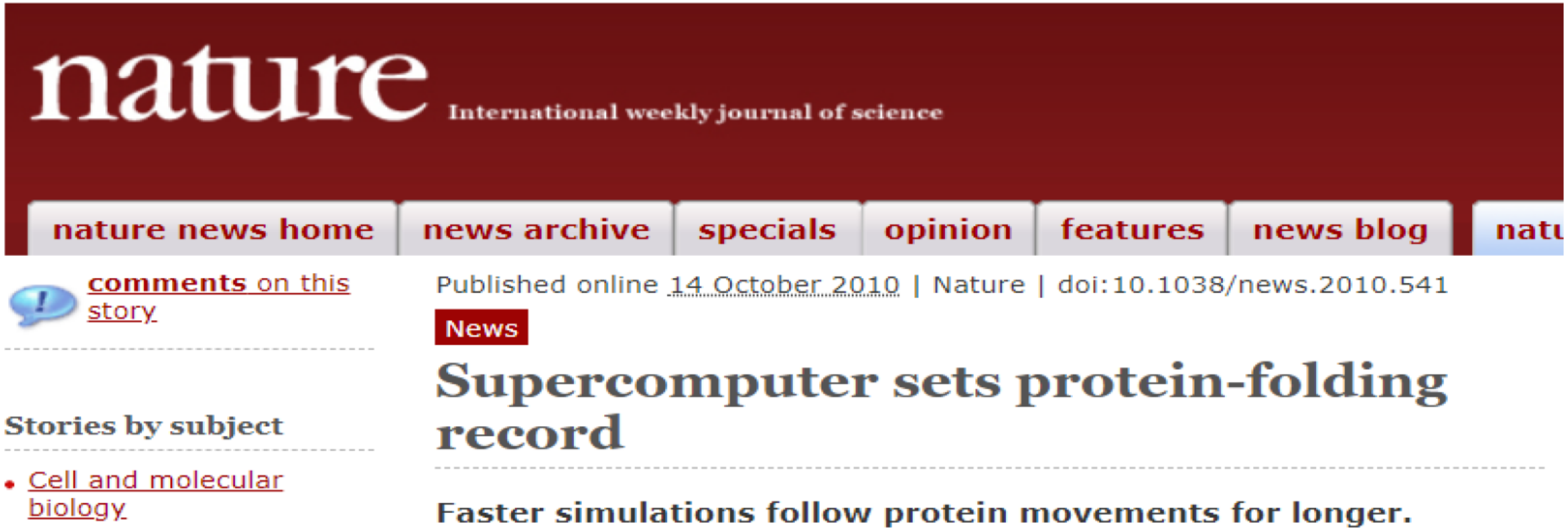


Christian Anfinsen
Nobel Prize in Chemistry 1972

Protein Structure Prediction

State of the Art Until 2015

- A lot of computing power needed
- Success rate is low even for small proteins



Published online 14 October 2010 | Nature | doi:10.1038/news.2010.541

News

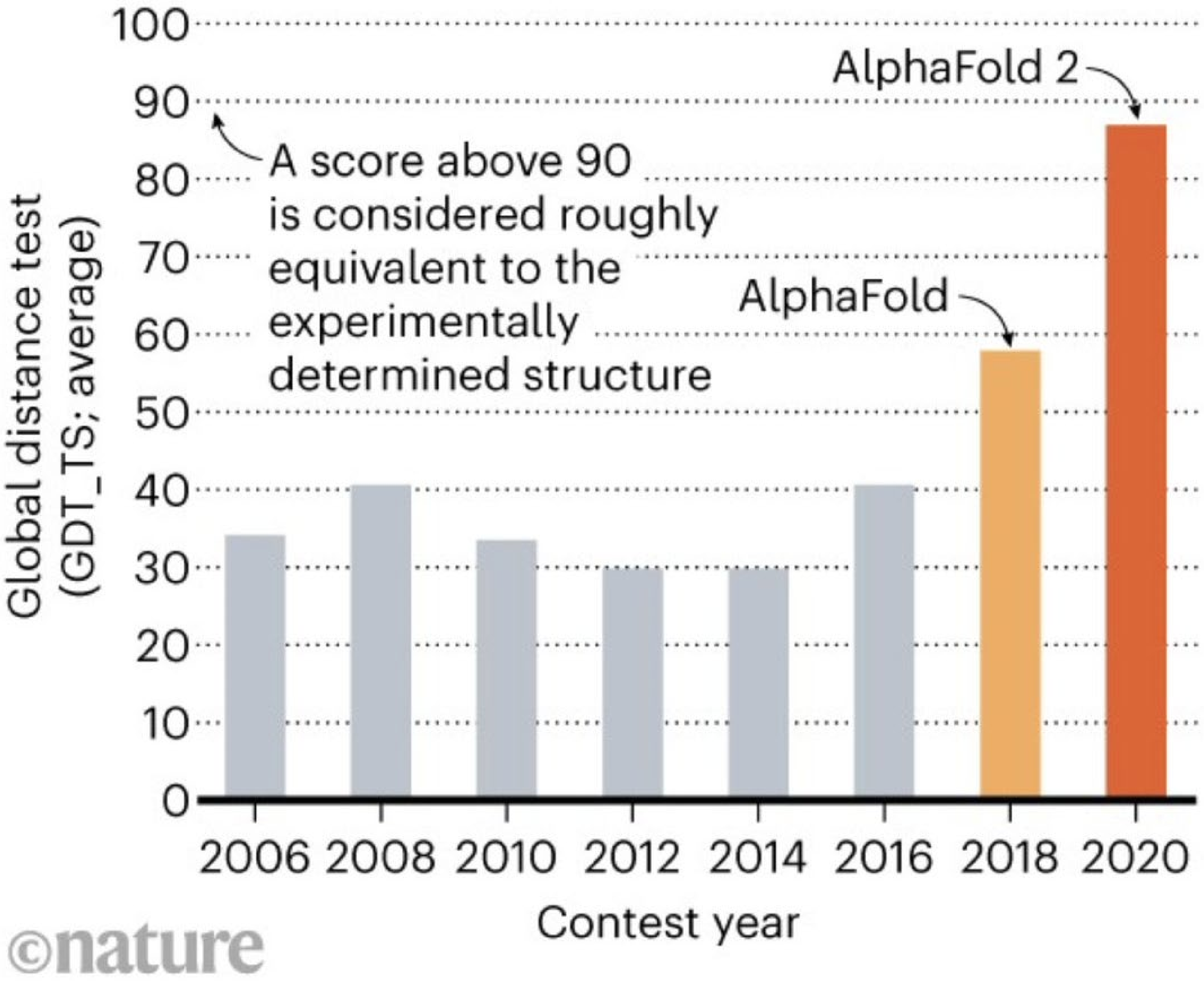
Supercomputer sets protein-folding record

Faster simulations follow protein movements for longer.

[slide from Jinbo Xu, TTI]

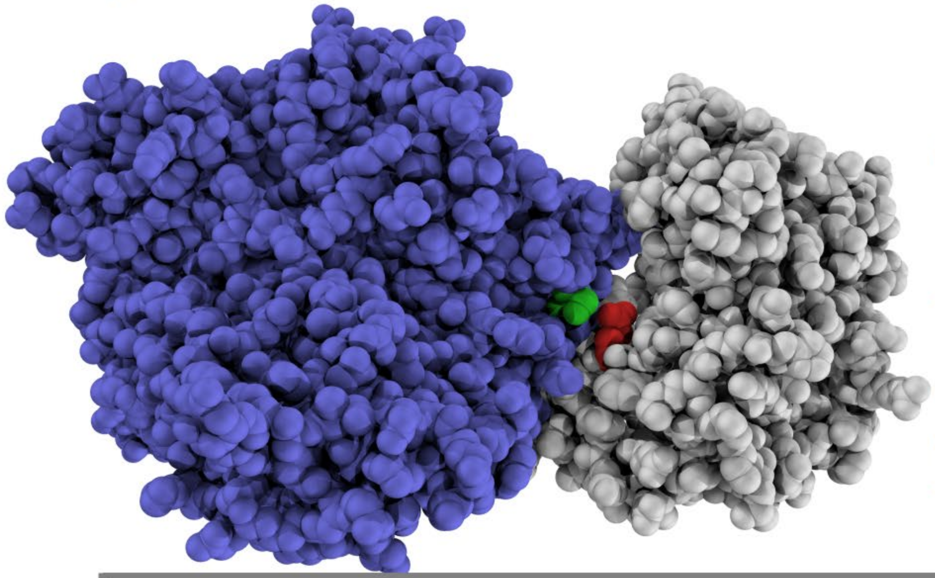
2020

State-of-the-art is deep learning based:



AlphaFold2 relies on previous key insights

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



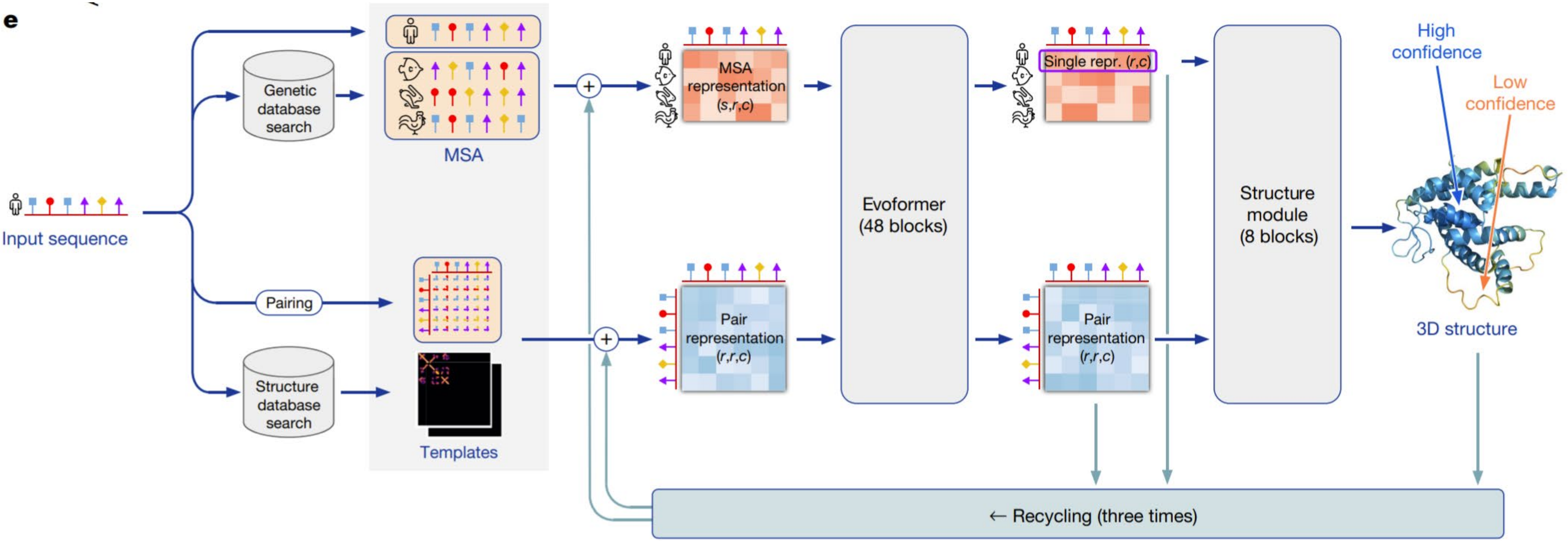
For example, a mutation that causes one amino acid to get bigger is more likely to preserve protein structure and function (and thus survive) if another amino acid gets smaller to make space

```
...GANPMHGRDQSGAVASLTSVA...
...GANPMHGRDQEGAVASLTSVA...
...GANPMHGRDEKGAVASLTSVG...
...GANPMHGRDSHGWLASCLSVA...
...GANPMNGRDVKGFVAAGASVA...
...GANPMHGRDRDGAVASLTSVA...
...GANPMHGRDQVGAVASLTSVA...
...GANPMHGRDQEGAVASLTSVA...
```

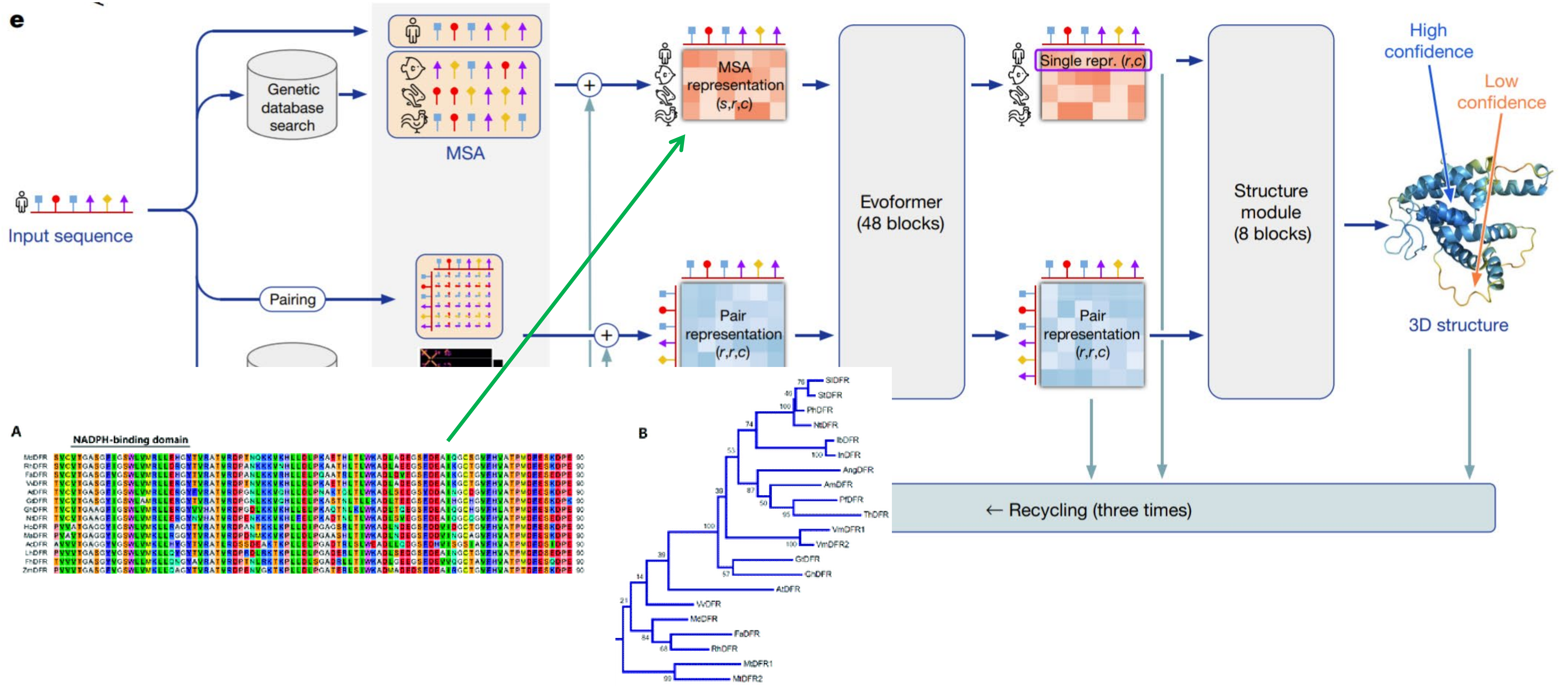
```
...VEDLMKEVVTYRHFMNASGG...
...VEALMARVLSYRHFMNASGG...
...VATVMKQVMTYRHYLRATGG...
...VARAMREIGKYAQVLKISRG...
...VPELMQDLTSYRHFMNASGG...
...ADHVLRRLSDFVPALLPLGG...
...FERARTALEAYAAPLRAMGG...
...VPEVMKKVMSYRHYLKATGG...
```

[slide from Jinbo Xu, TTI]

AlphaFold2 “almost end-to-end” neural network

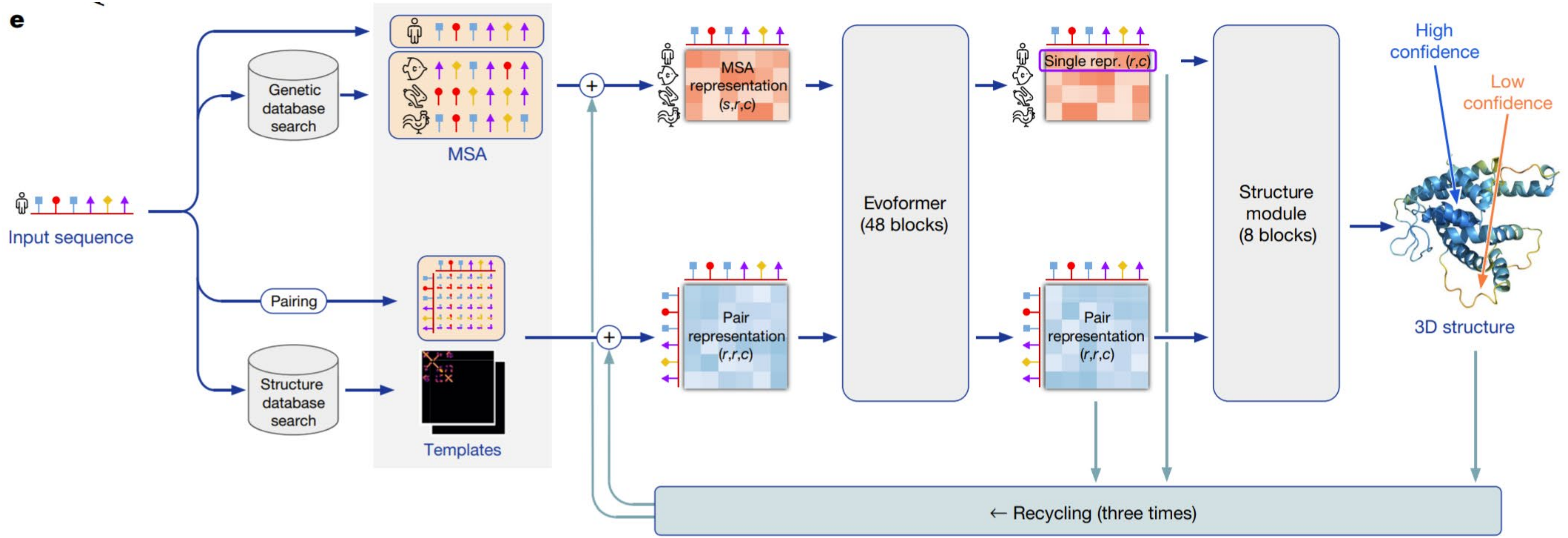

e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

AlphaFold2 “almost end-to-end” neural network



© 2020 DeepMind Technologies Limited

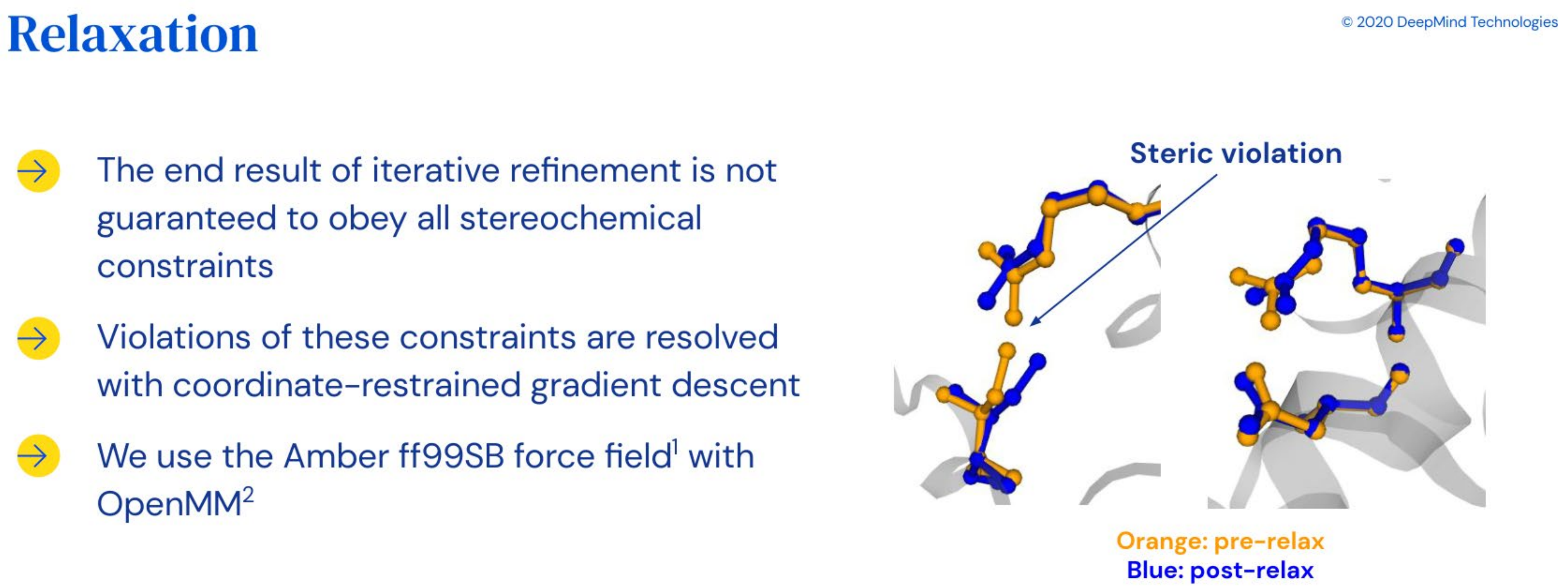
AlphaFold2 “almost end-to-end” neural network


e
Input sequence
Genetic database search
Pairing
Structure database search
MSA
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

(uses an equivariant attention architecture)

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.



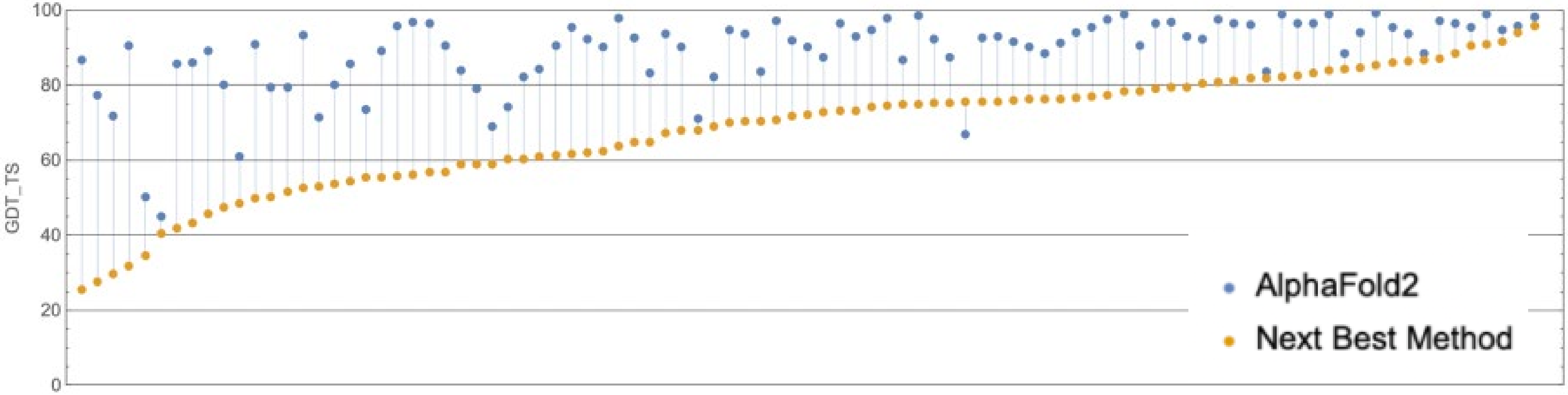
Relaxation

- 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]

Steric violation

Orange: pre-relax
Blue: post-relax

AlphaFold2 “almost end-to-end” neural network



From great blog by Mohamed Alquraishi:

https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/

Some thoughts on AlphaFold2

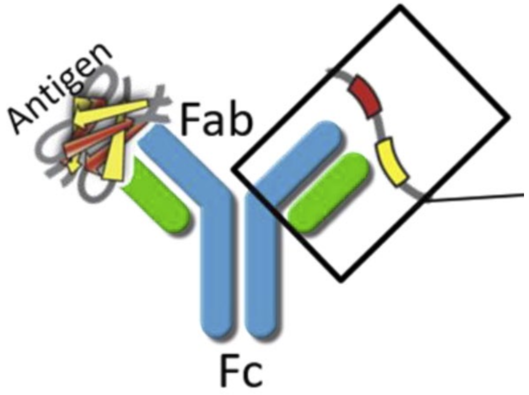
- DeepMind took on a long-tackled, well-defined problem, with clear data, clear benchmarks, and a clear way to demonstrate improvement.
- Expense of protein structure data used for AlphaFold2, conservatively estimated at ~US$20 billion (Burley et al., 2023).
- They relied **heavily** on years of prior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

CS 189/289

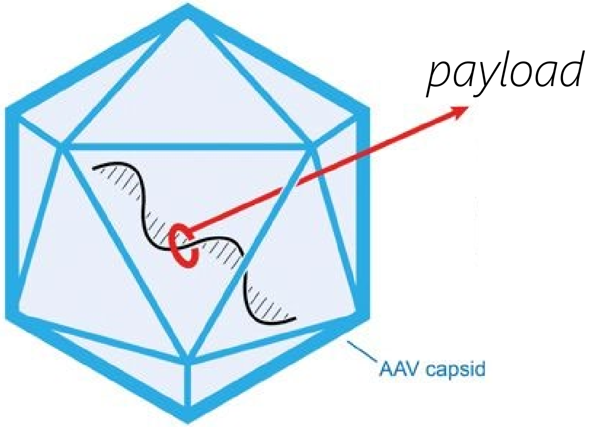
Some applications of AI in biology:

1. protein structure prediction
2. protein design

Protein engineering: therapeutics, environment, *etc.*

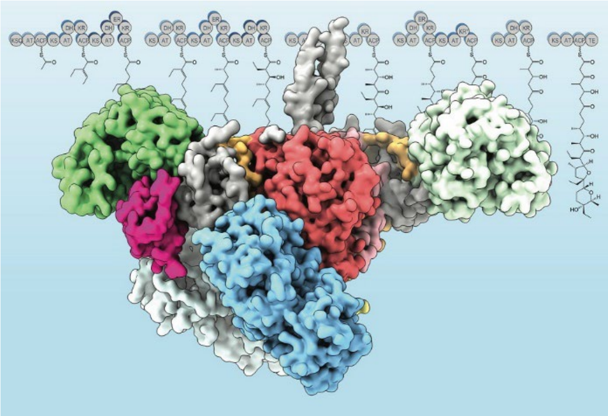


antibody therapeutics

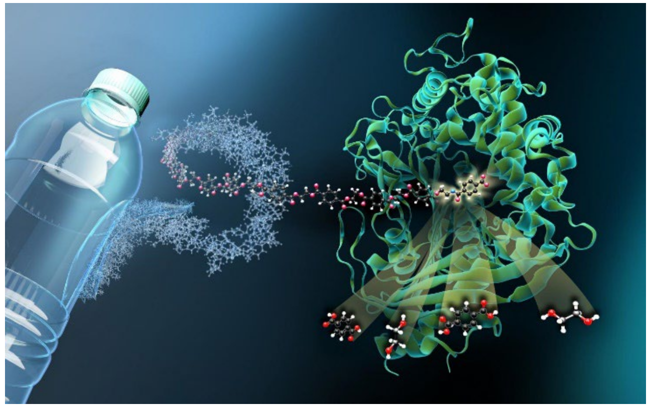


gene therapy virus delivery (AAV)

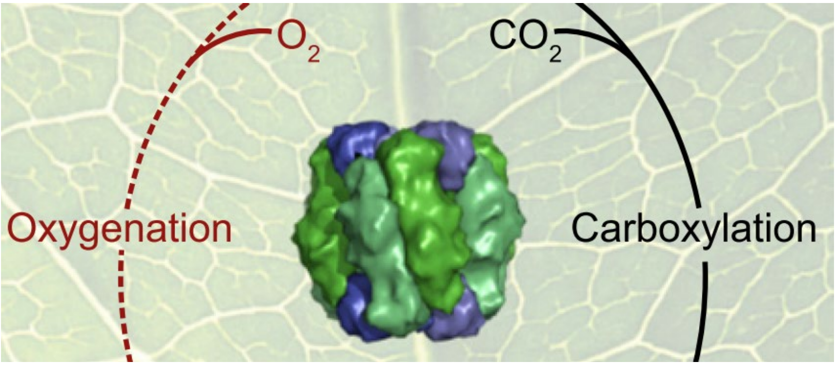
gene editing (CRISPR/Cas9)



antibiotics & biofuel production (PKS)



plastic recycling (PETase)



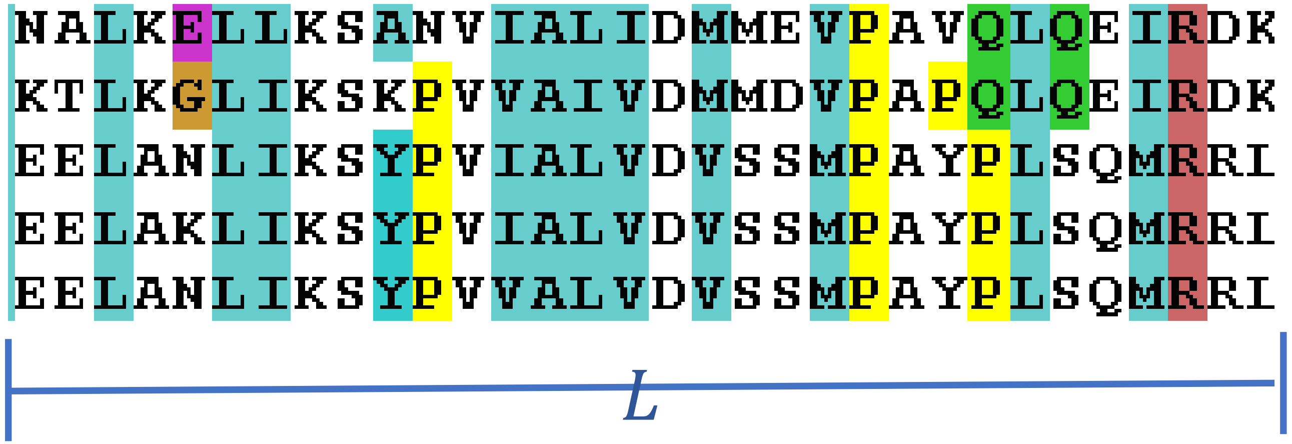
CO_2 biosequestration (RuBisCO)

Fundamental difficulty: design space is nearly infinite

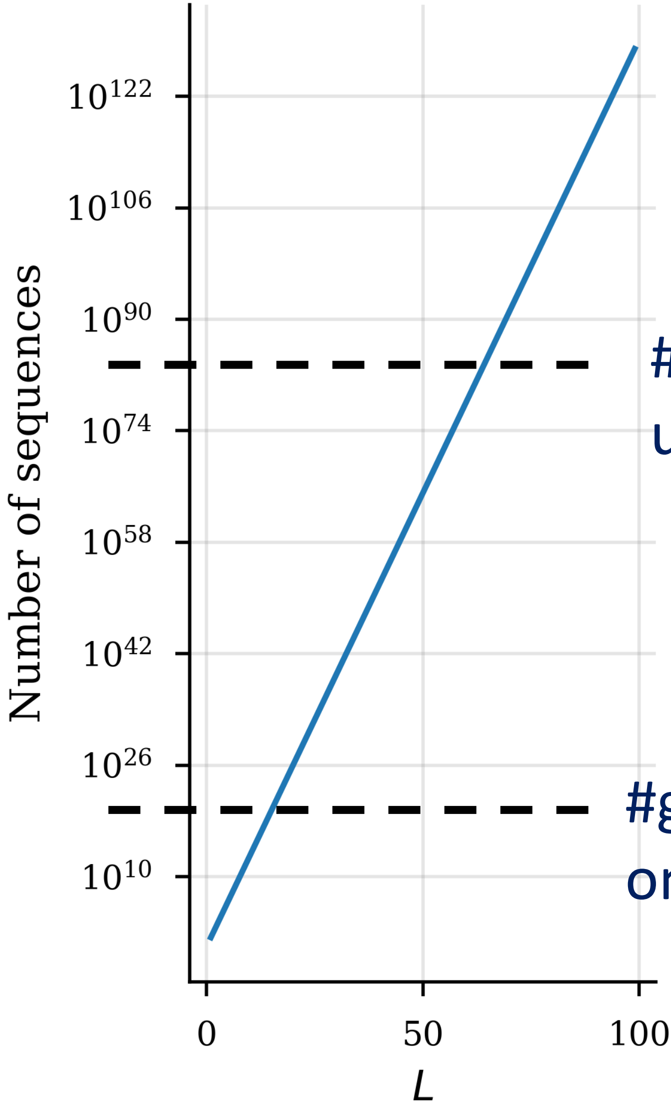
- Also highly rugged design space

$\Longrightarrow$ size scales as $\sim \mathbf{20}^{L}$

- Discrete search space (no gradients)



NALKELLKSANVIALIDMMEVPAVQLQEIRDK
KTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDK
EELANLIKSYPVIALVDVSSMPAYPLSQMRRI
EELAKLIKSYPVIALVDVSSMPAYPLSQMRRI
EELANLIKSYPVVALVDVSSMPAYPLSQMRRI

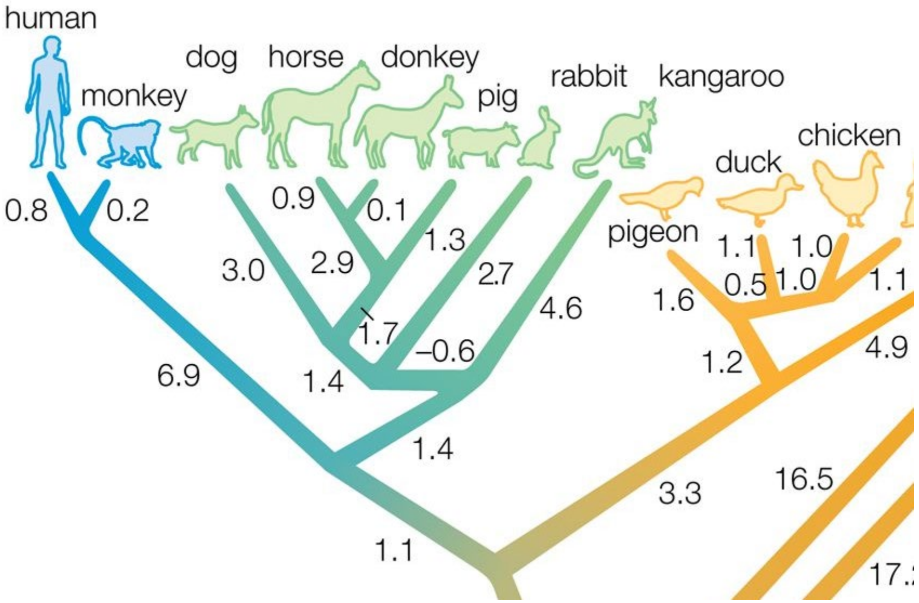


#atoms in universe $\sim 10^{80}$

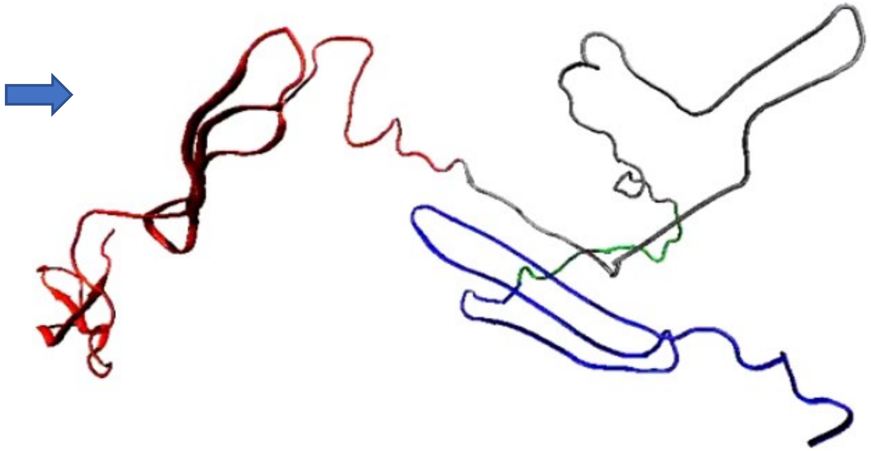
#grains of sand on earth $\sim 10^{18}$

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



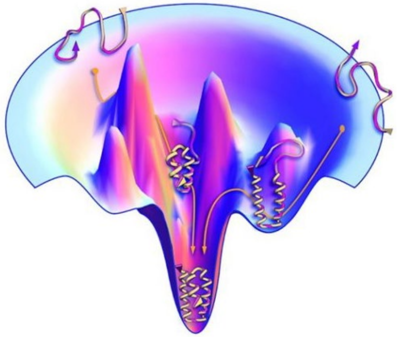
green fluorescent protein folding itself

Successes in navigating this complex space

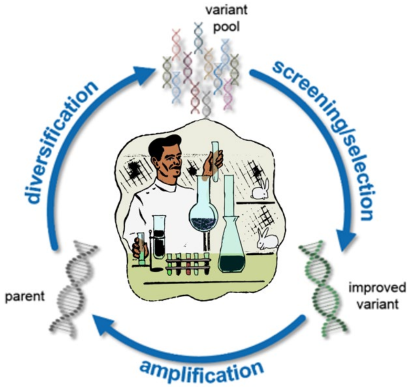
1. Nature: via evolution *over millions of years.*
2. Various protein engineering strategies.

Protein engineering strategies emerging

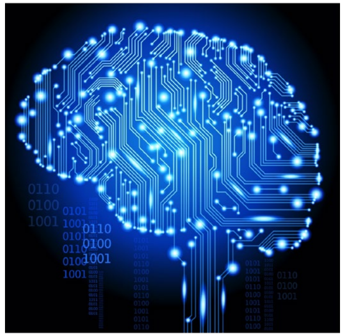
i. Computation ("data free"): **physics-based energy functions** (*e.g.*, Rosetta) to model **protein structure**, and protein binding. *~1997-2023'ish* (almost R.I.P.)



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



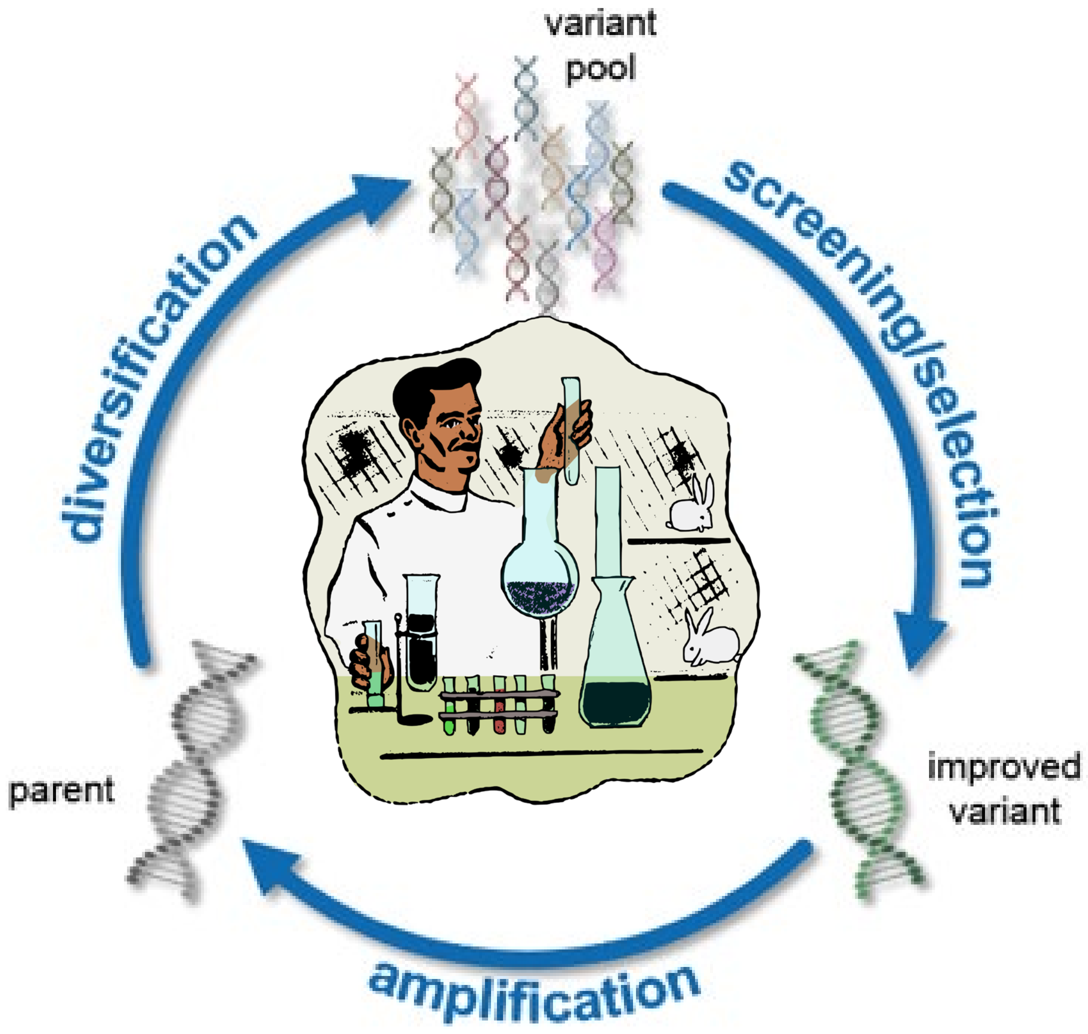
iii. Machine learning (augmented): generative models; function prediction; structure prediction, etc. *~2018(?)-present*



One strategy: ML-based Directed Evolution

Gregory Winter (left), Frances Arnold and George Smith share this year's Nobel Prize in Chemistry. Credit: L–R: Aga Machaj; Caltech; Univ. Missouri-Columbia

2018 Nobel Prize
in Chemistry



Goal: get same results with fewer measurements, and/or, get better result than pure DE.

1. Replace assay with predictive model.
2. Replace search with intelligent search.

https://www.bsse.ethz.ch/bpl/research/directed-evolution.html

Did AlphaFold2 "solve" protein engineering?

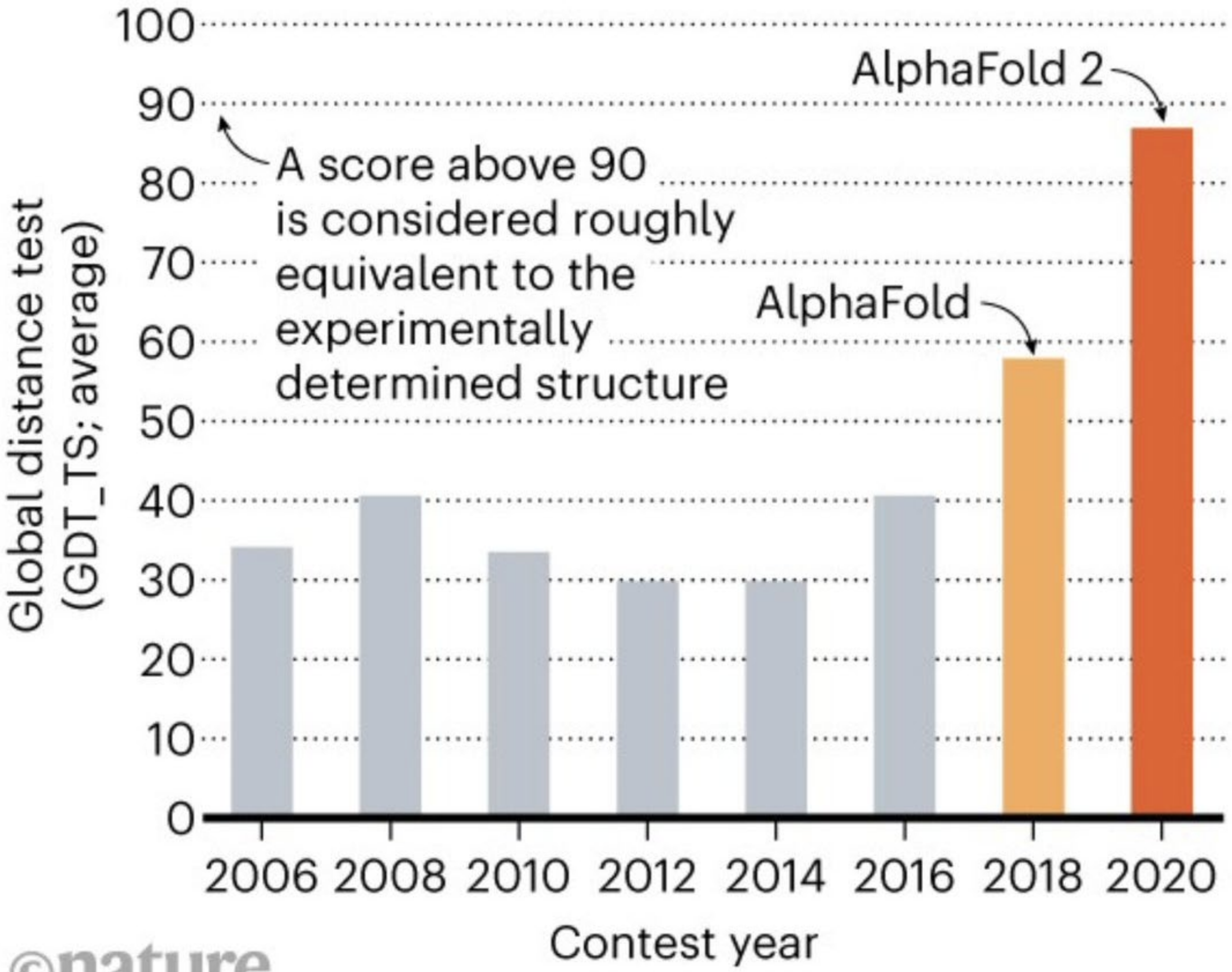
NEWS | 22 July 2021

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.

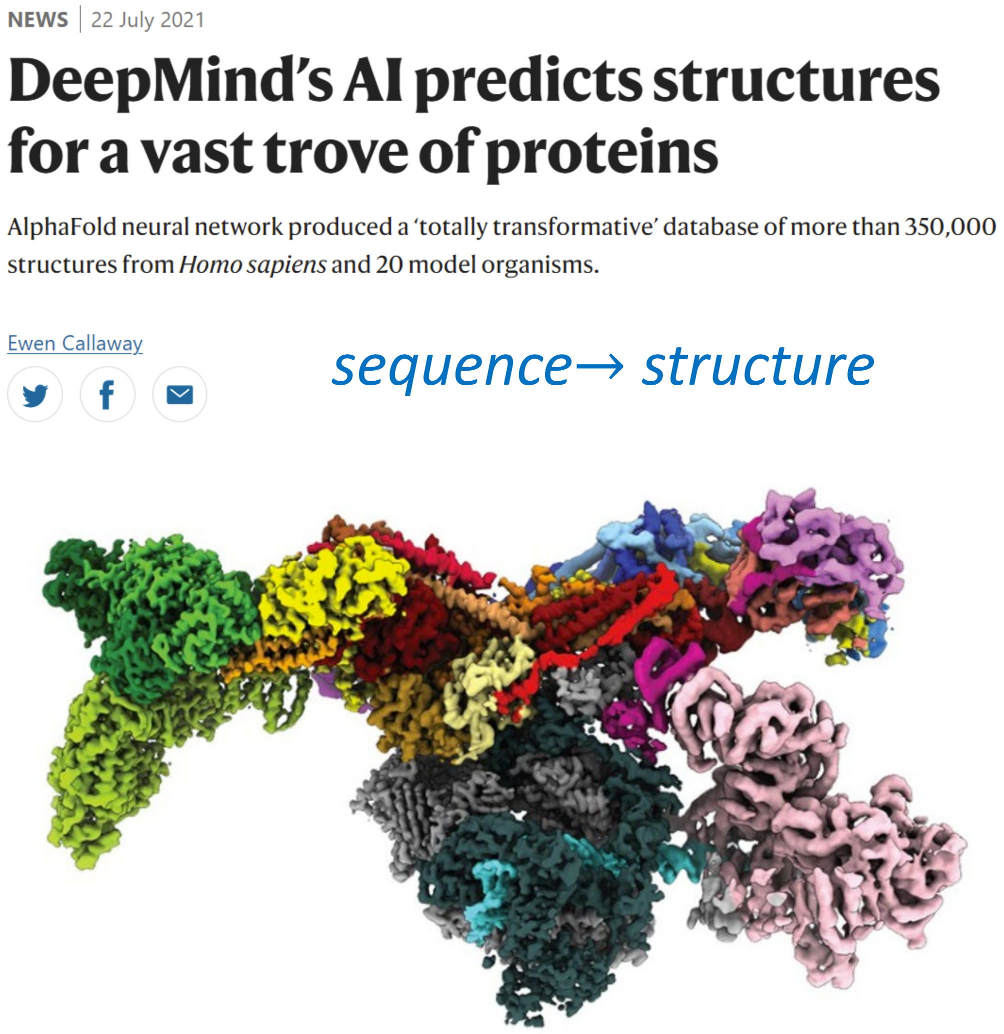
Ewen Callaway

sequence→ structure



©nature

Did AlphaFold2 “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

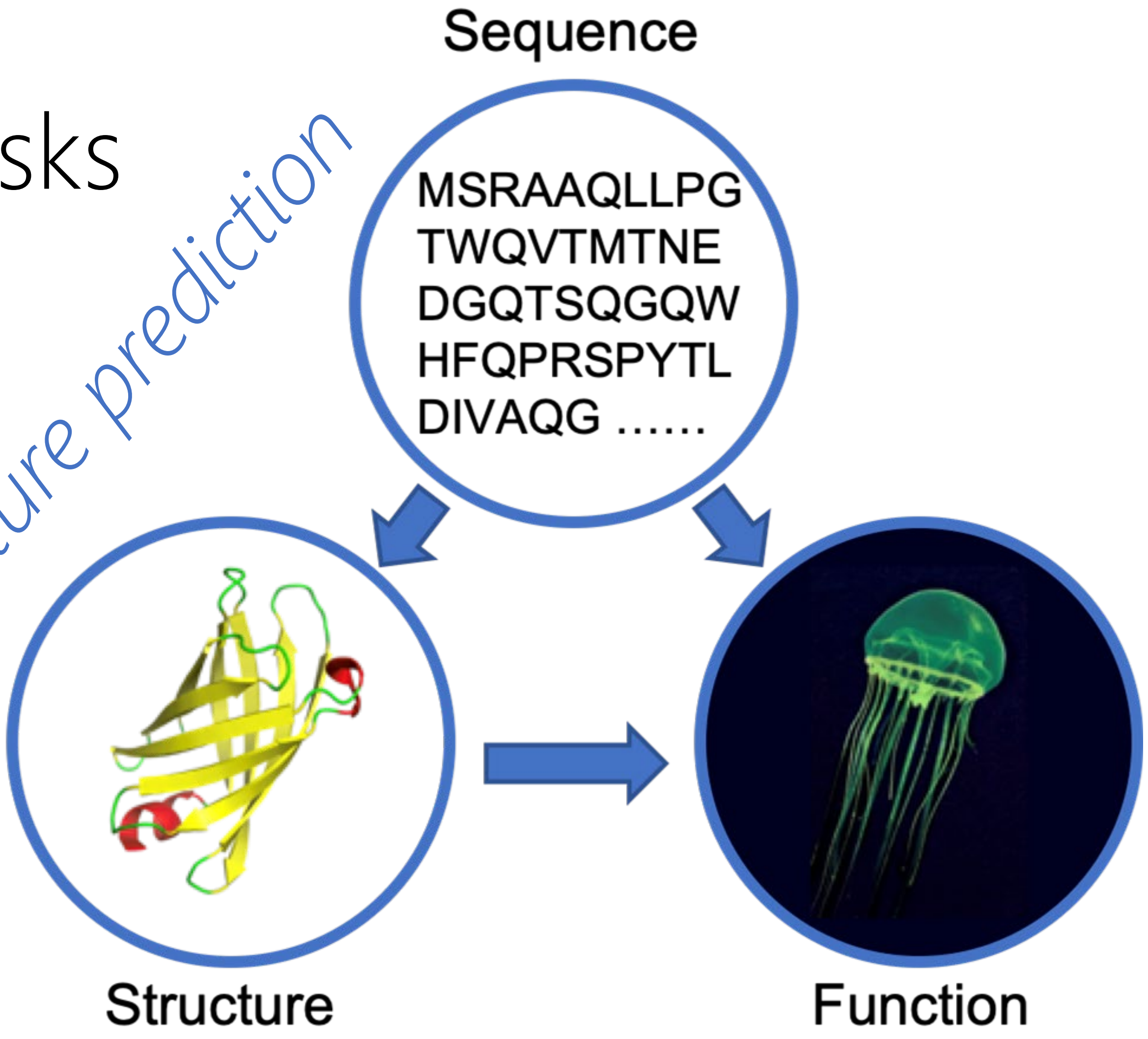
sequence→ structure

- No: don’t typically know which protein structures we need.
- If did, would need: *structure*→*sequence*. (decent ML solutions exist).
- Bottleneck challenge: predict which protein have the function we desire.
- AlphaFold2 *was* a breakthrough, and will surely be useful.

A suite of ML protein engineering problems

Prediction tasks

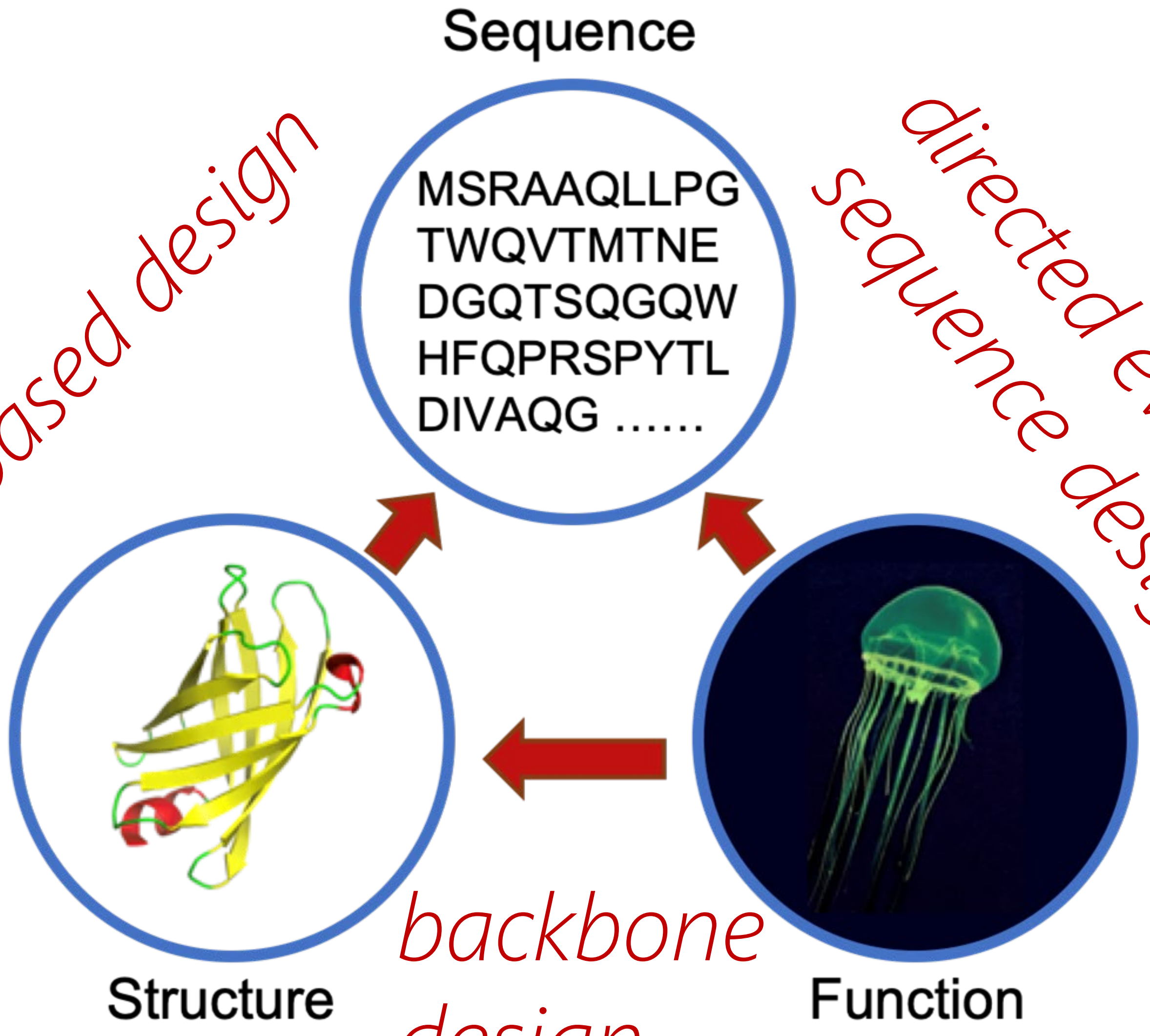
structure prediction



function prediction

A suite of ML protein engineering problems

Design tasks



structure-based design

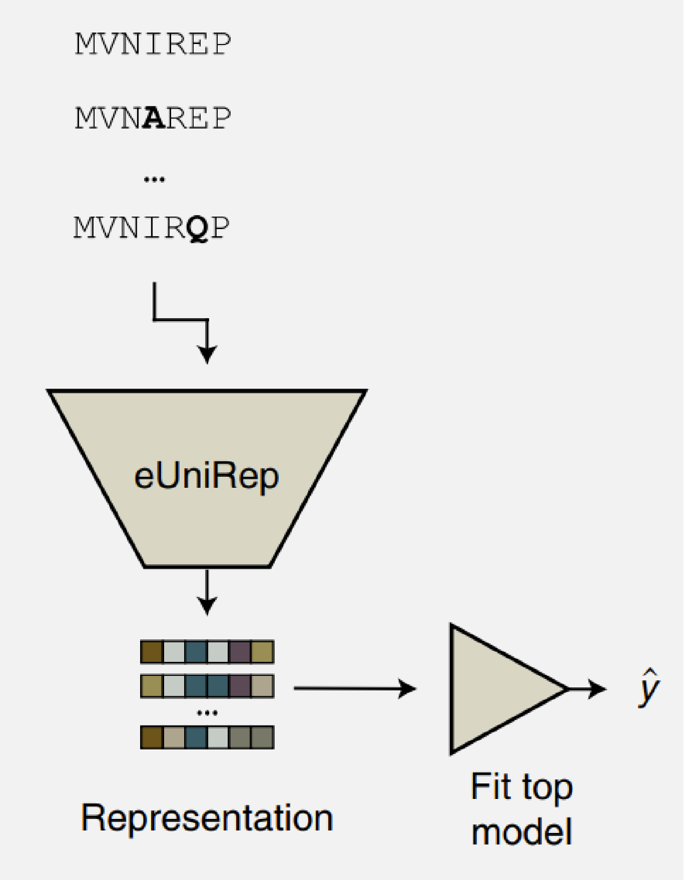
directed evolution, sequence design

backbone design

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



[Biswas *et al.*, *Nat. Meth.* 2021]

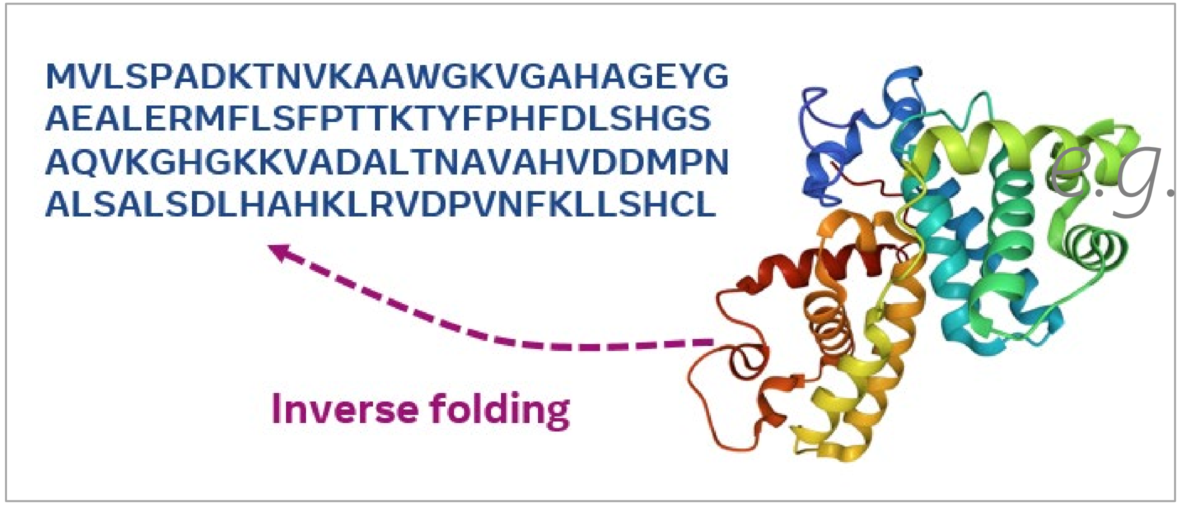
Some trends in ML + protein engineering

2. (Conditional) generative models for sequences.

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

a) structure-conditioned, aka "inverse folding"

b) "control tag" conditioned, protein family

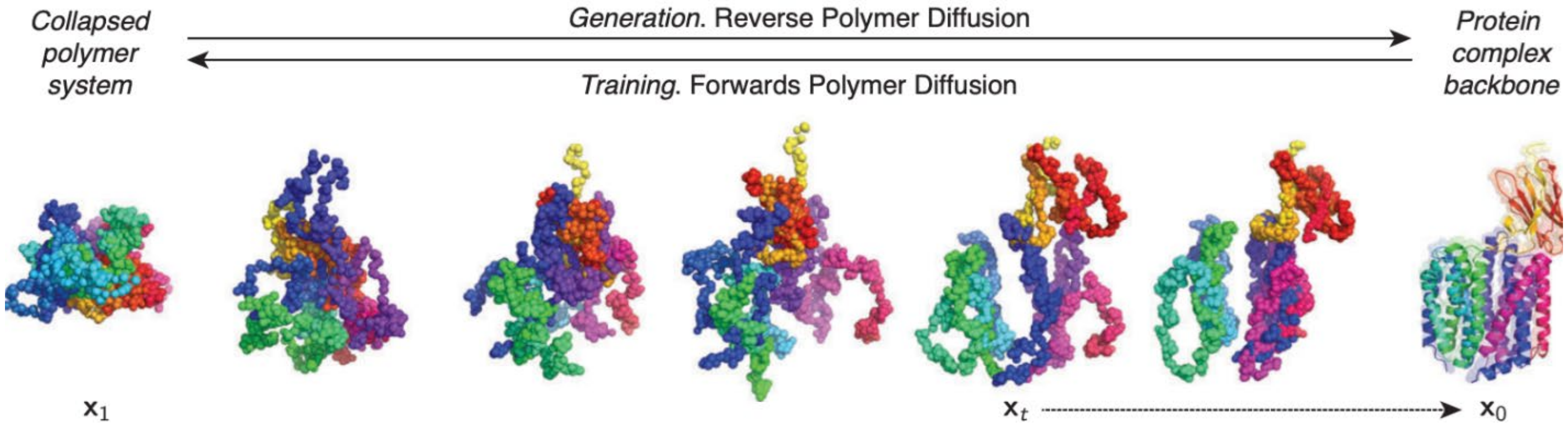

MVLSPADKTNVKAAWGKVGAHAGEYG
AEALERMFLSFPTTKTYFPHFDLSHGS
AQVKGHGKKVADALTNAVAHVDDMPN
ALSALSDLHAHKLRVDPVNFKLLSHCL
Inverse folding

e.g.,

Some trends in ML + protein engineering

3. (Conditional) generative models for structure.

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

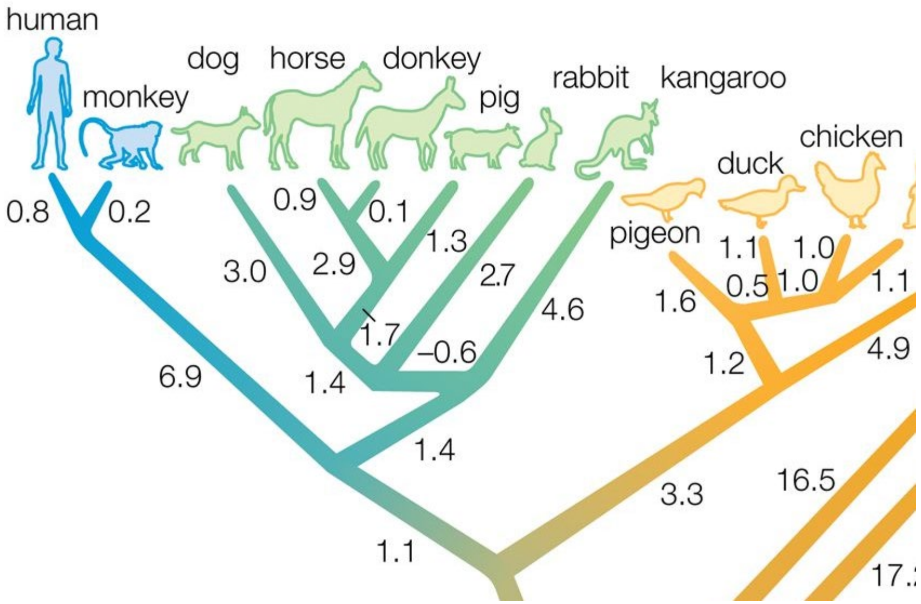

Collapsed polymer system
Generation. Reverse Polymer Diffusion
Training. Forwards Polymer Diffusion
Protein complex backbone
$\mathbf{x}_1$
$\mathbf{x}_t$
$\mathbf{x}_0$

[Ingraham *et al. bioRxiv* 2022]

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.

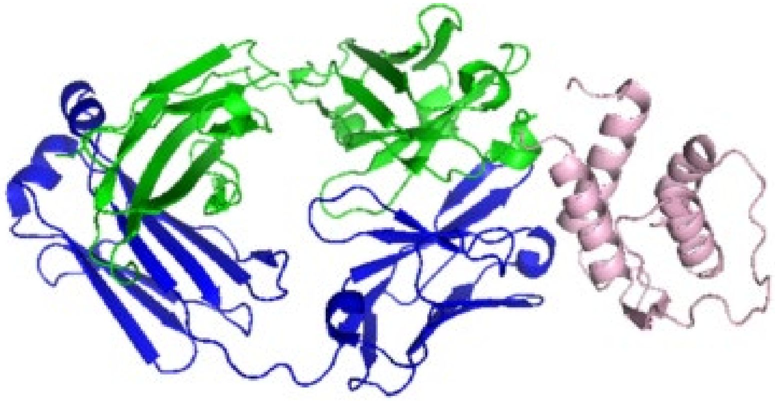


**key part of AlphaFold2*

Some trends in ML + protein engineering

5. Structure prediction: filling the gaps left by AlphaFold2

- Orphan proteins (with *no/few homologs*).
- Proteins *in bound form*.
- Protein dynamics and conformational distributions.
- Protein-protein binding.
- Protein-DNA/RNA binding



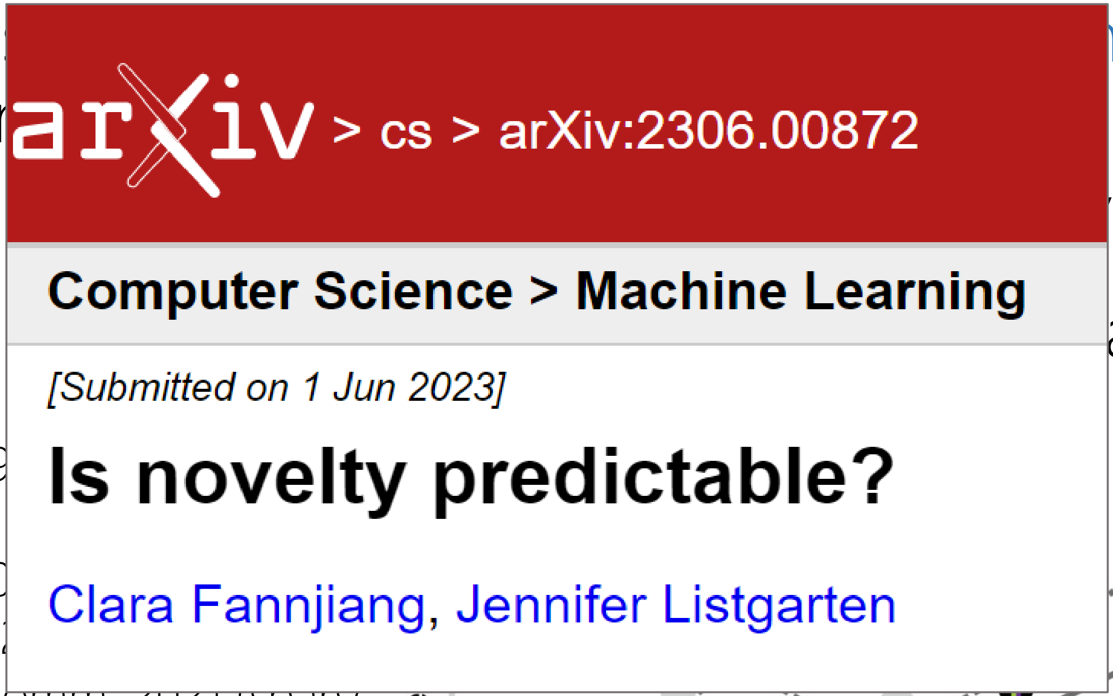
ML focus of my group: “ML-based design”:

A. Natural tension between extrapolation vs. trustworthiness. [1-4].

B. Related to cau ncertainty (whereas we typically think r

C. Suitable protei NLP) [4-7].

D. Design of distr al sequences [1,2,8,9].



arXiv > cs > arXiv:2306.00872

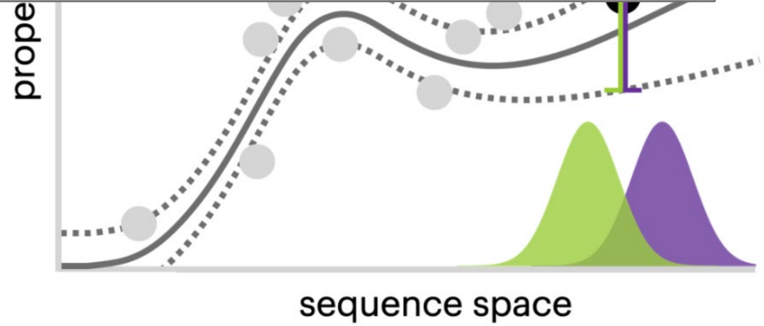
Computer Science > Machine Learning

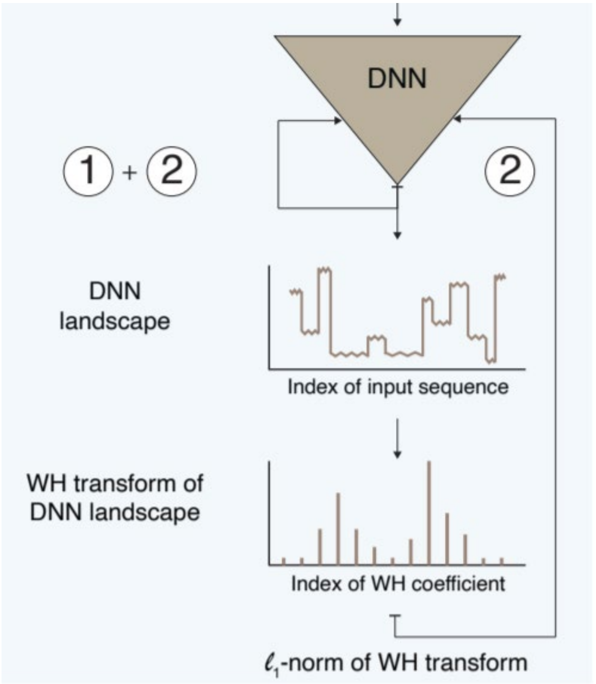
[Submitted on 1 Jun 2023]

Is novelty predictable?

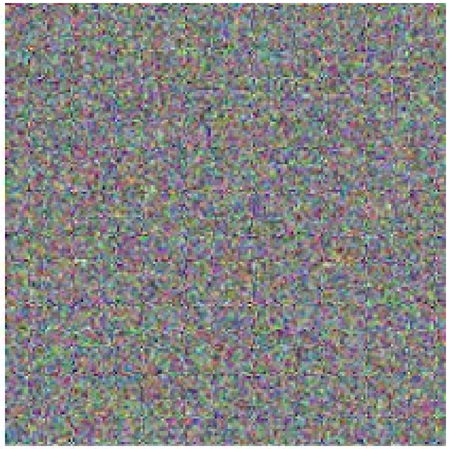
Clara Fannjiang, Jennifer Listgarten

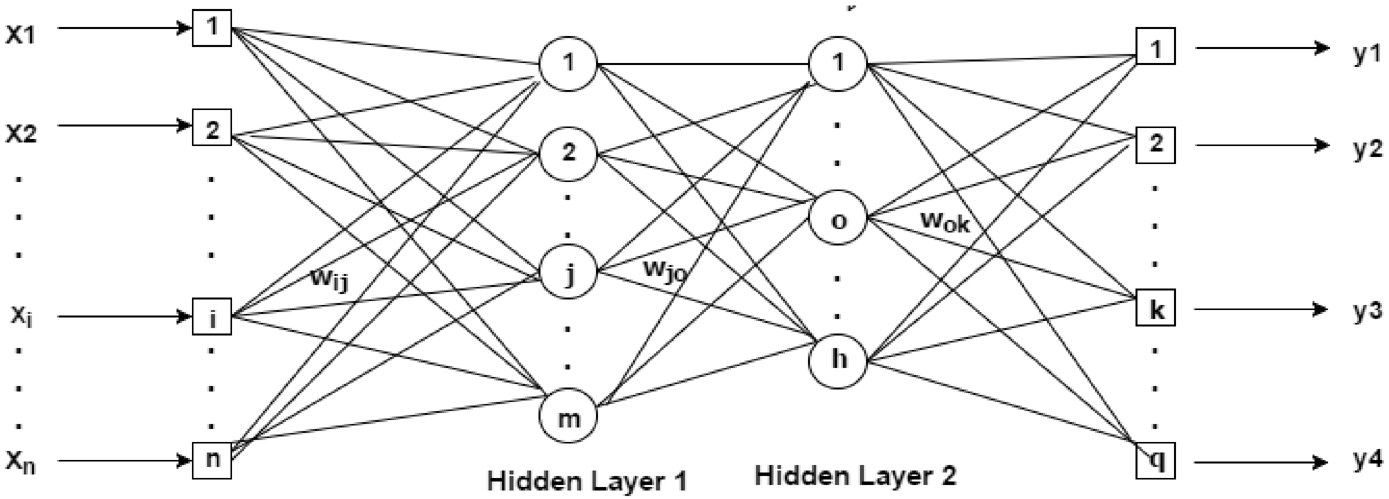
1. Brookes *et al ICLM* 2019
2. Fannjiang *et al NeurIPS*
3. Fannjiang *et al PNAS* 20
4. Nisonoff *et al* arXiv 202
5. Aghazadeh *et al* Nat. Comm. 2021 (sparse
6. Brookes *et al PNAS* 2022 (funct. pred.)
7. Hsu *et al Nat. Biotech*. 2022 (function prec
8. Zhu, Brookes, *et al* ,*bioRxiv*. (opt. design)
9. Busia & Listgarten, *bioRxiv* (log enrichmen
10. Fannjiang & Listgarten, *arXiv* (overview)



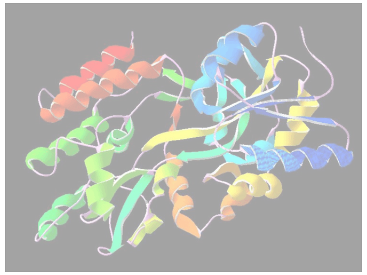


Analogy: can we trust "banana" design?



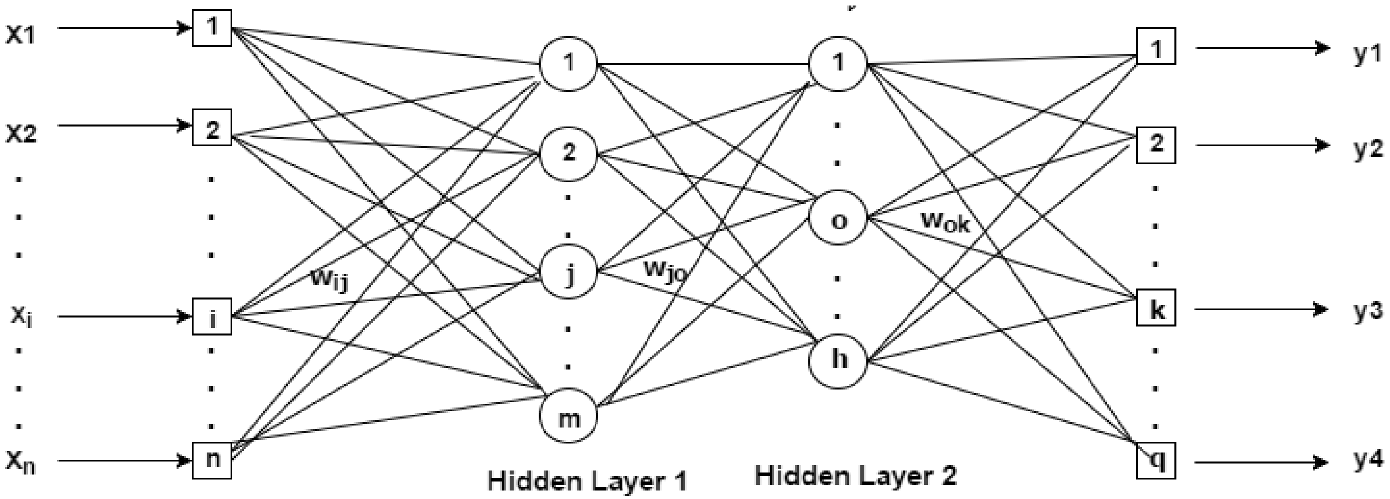


desired property

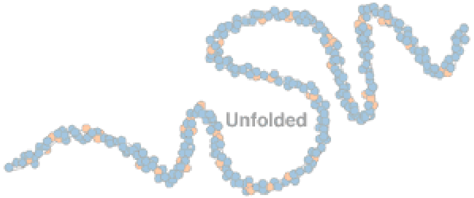


catalytic efficiency

Naïve design yields abstract art.



desired property

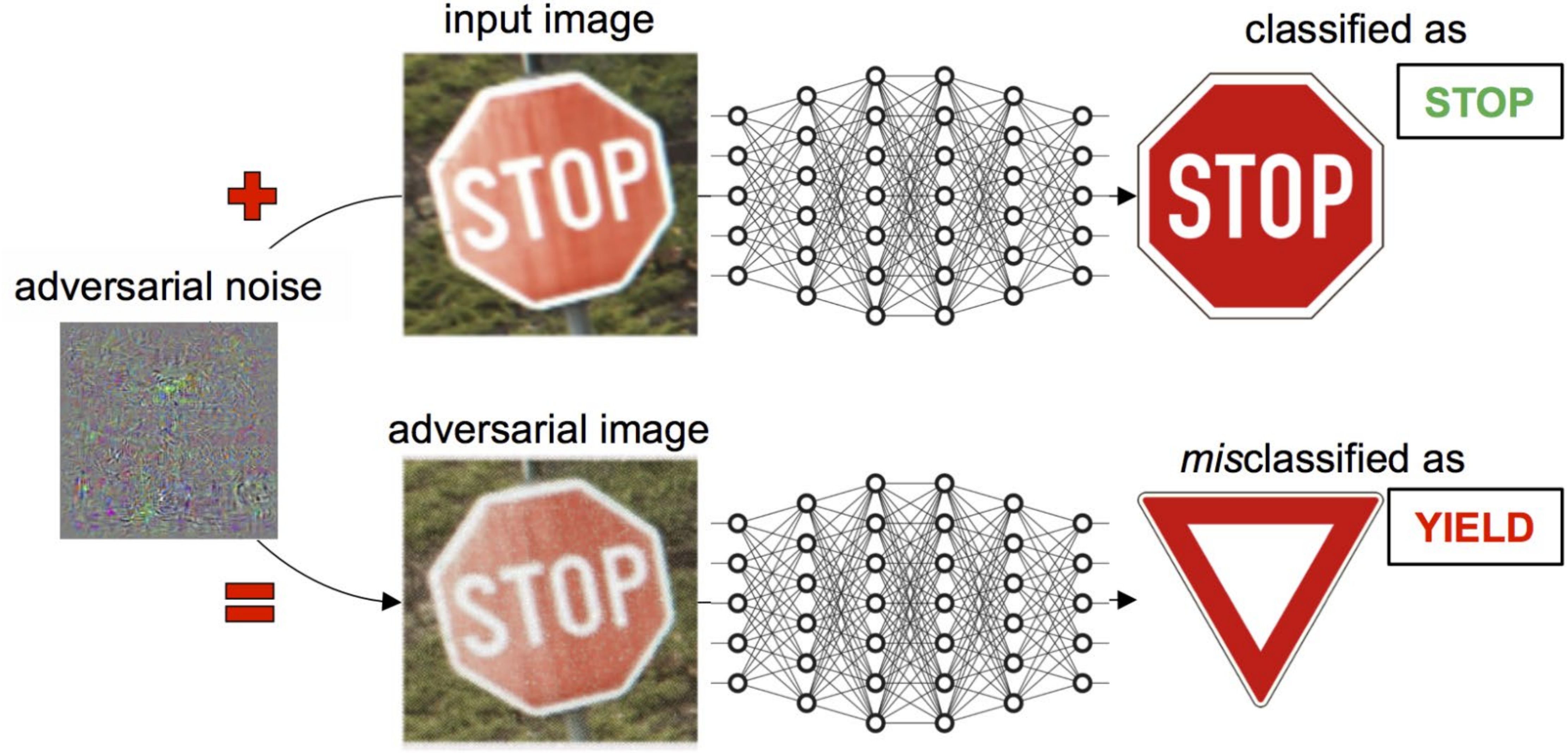


catalytic
efficiency

non-folding protein

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

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



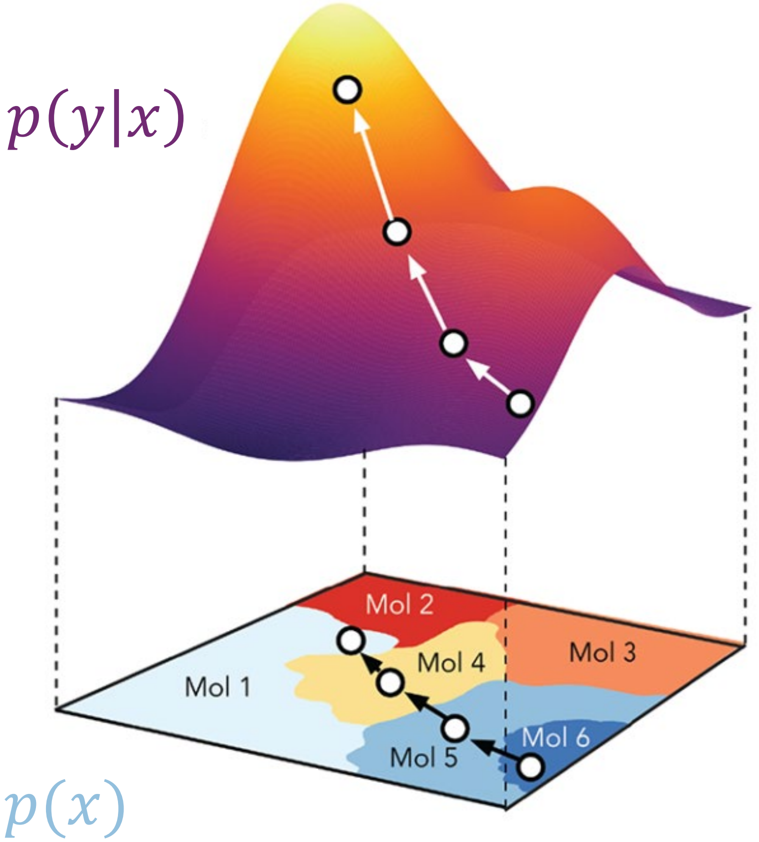
https://www.pluribus-one.it/research/sec-ml/wild-patterns

Conditioning by Adaptive Sampling for Robust Design (CbAS)

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.]

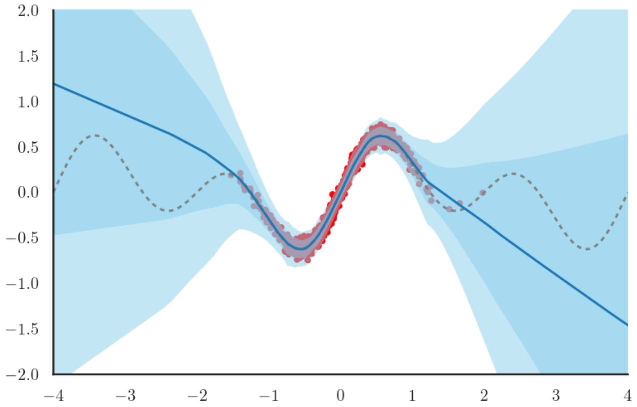
Brookes, Park & Listgarten *ICML* 2019

David Brookes

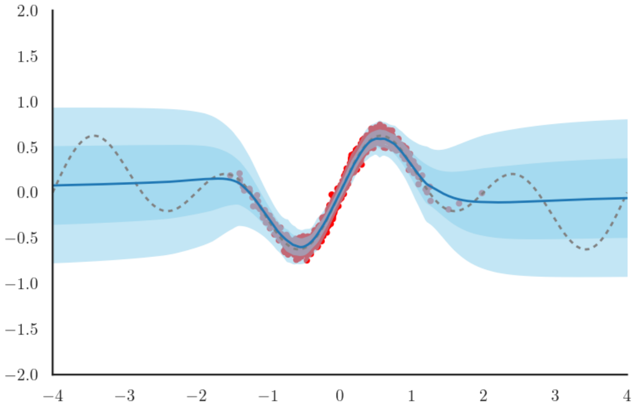
Augmenting Neural Networks with Priors on Functional Values

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

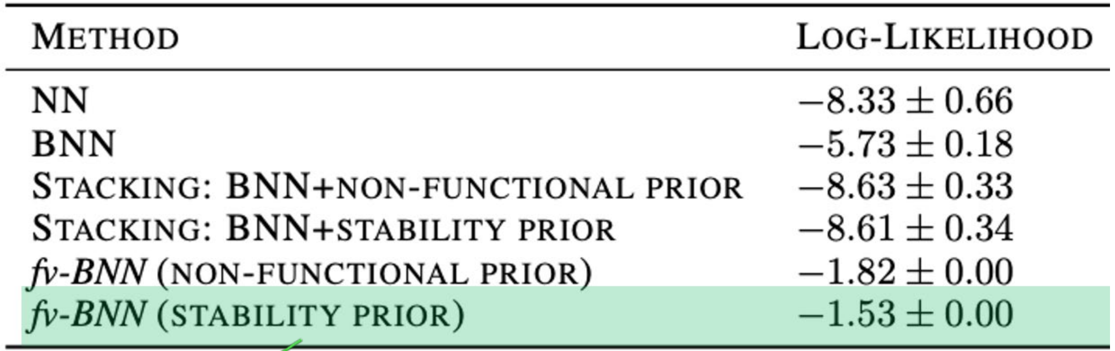
Easy to implement, zero added cost.



regular BNN



function-value augmented BNN



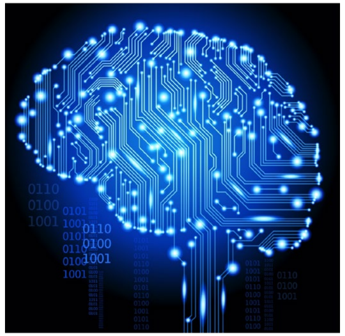
METHOD	LOG-LIKELIHOOD
NN	-8.33 ± 0.66
BNN	-5.73 ± 0.18
STACKING: BNN+NON-FUNCTIONAL PRIOR	-8.63 ± 0.33
STACKING: BNN+STABILITY PRIOR	-8.61 ± 0.34
fv-BNN (NON-FUNCTIONAL PRIOR)	-1.82 ± 0.00
fv-BNN (STABILITY PRIOR)	-1.53 ± 0.00

Hunter Nisonoff

<u>Nisonoff</u>, Wang, Listgarten, *bioRxiv*

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)


+

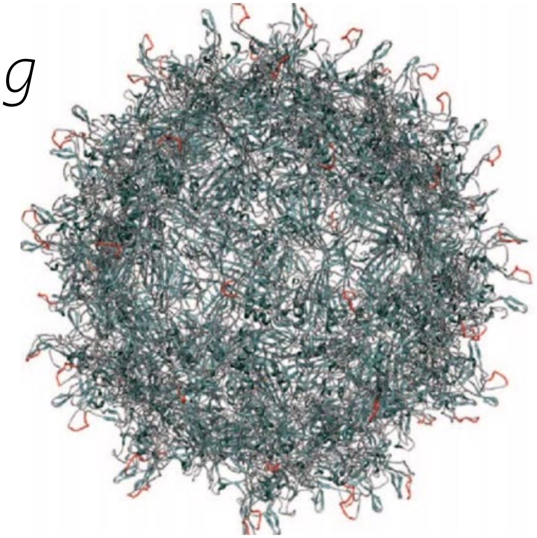
Engineering AAV for gene therapy delivery

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 on job market!

Zhu, Brookes, Busia,…, Nowakowski, Listgarten, Schaffer, *bioRxiv*

Promising AAV clinical trials

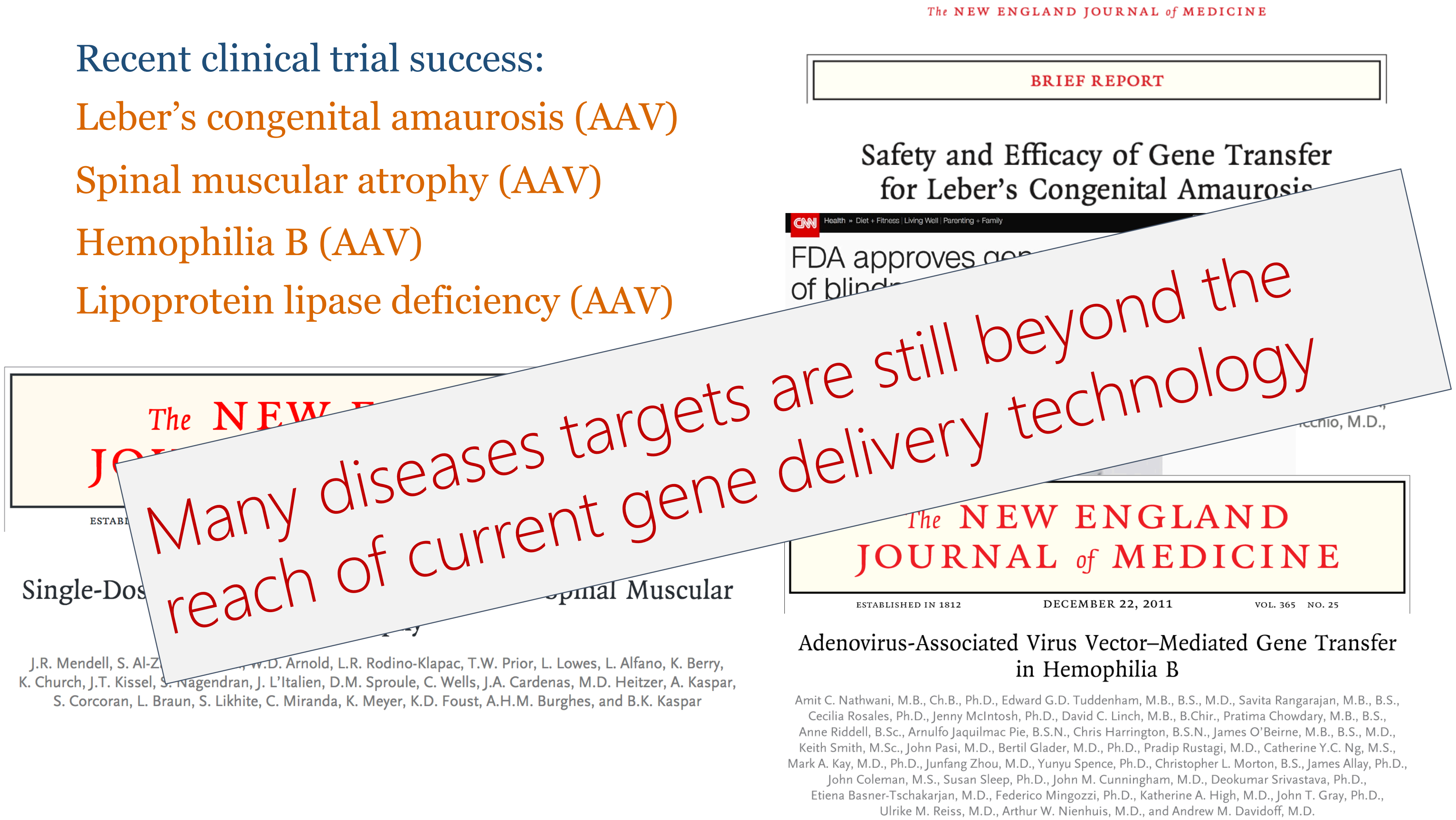
Recent clinical trial success:

Leber’s congenital amaurosis (AAV)

Spinal muscular atrophy (AAV)

Hemophilia B (AAV)

Lipoprotein lipase deficiency (AAV)



The NEW ENGLAND JOURNAL of MEDICINE

BRIEF REPORT

Safety and Efficacy of Gene Transfer for Leber’s Congenital Amaurosis

FDA approves ge
of blind

cchio, M.D.,

The NEW E
JO
ESTABL

Single-Dos pinal Muscular

J.R. Mendell, S. Al-Z W.D. Arnold, L.R. Rodino-Klapac, T.W. Prior, L. Lowes, L. Alfano, K. Berry, K. Church, J.T. Kissel, S. Nagendran, J. L’Italien, D.M. Sproule, C. Wells, J.A. Cardenas, M.D. Heitzer, A. Kaspar, S. Corcoran, L. Braun, S. Likhite, C. Miranda, K. Meyer, K.D. Foust, A.H.M. Burghes, and B.K. Kaspar

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 22, 2011 VOL. 365 NO. 25

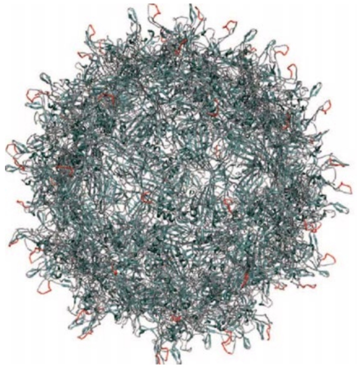
Adenovirus-Associated Virus Vector–Mediated Gene Transfer in Hemophilia B

Amit C. Nathwani, M.B., Ch.B., Ph.D., Edward G.D. Tuddenham, M.B., B.S., M.D., Savita Rangarajan, M.B., B.S., Cecilia Rosales, Ph.D., Jenny McIntosh, Ph.D., David C. Linch, M.B., B.Chir., Pratima Chowdary, M.B., B.S., Anne Riddell, B.Sc., Arnulfo Jaquilmac Pie, B.S.N., Chris Harrington, B.S.N., James O’Beirne, M.B., B.S., M.D., Keith Smith, M.Sc., John Pasi, M.D., Bertil Glader, M.D., Ph.D., Pradip Rustagi, M.D., Catherine Y.C. Ng, M.S., Mark A. Kay, M.D., Ph.D., Junfang Zhou, M.D., Yunyu Spence, Ph.D., Christopher L. Morton, B.S., James Allay, Ph.D., John Coleman, M.S., Susan Sleep, Ph.D., John M. Cunningham, M.D., Deokumar Srivastava, Ph.D., Etiena Basner-Tschakarjan, M.D., Federico Mingozzi, Ph.D., Katherine A. High, M.D., John T. Gray, Ph.D., Ulrike M. Reiss, M.D., Arthur W. Nienhuis, M.D., and Andrew M. Davidoff, M.D.

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

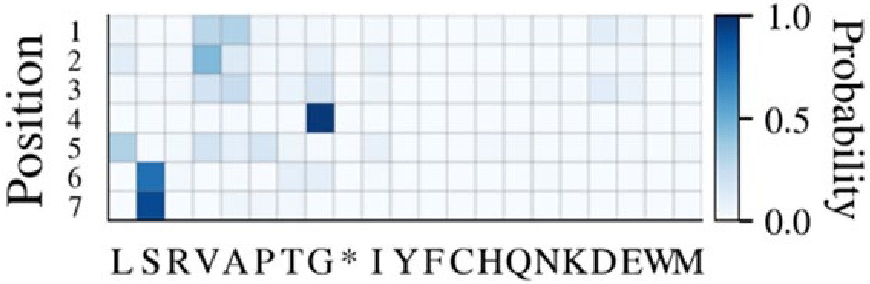
Ongoing challenges for AAV-based therapeutics

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

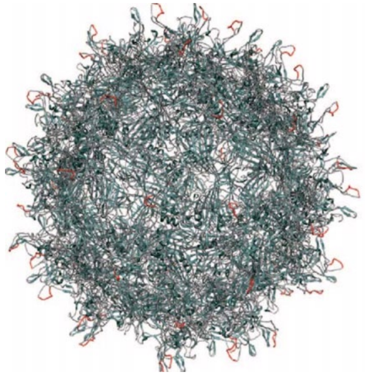


First AAV project goal, “library design”:

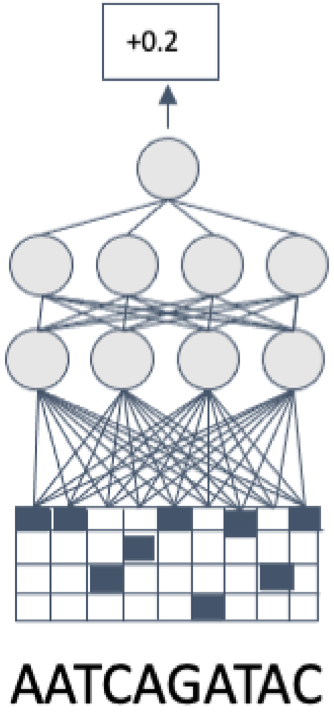
- Obtain <u>optimal starting “library</u>” for all these engineering goals.
- *i.e.*, fix the huge amount of library that gets wasted because doesn’t “<u>package</u>”.

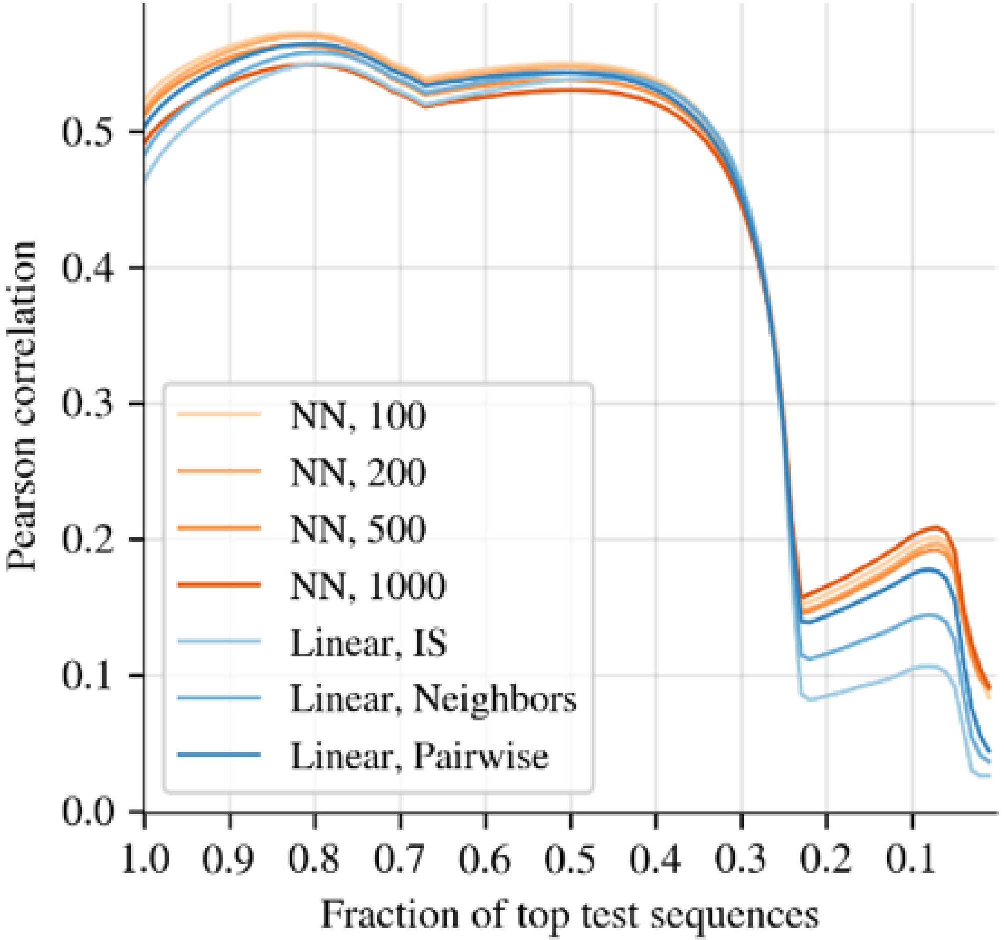


AAV library design

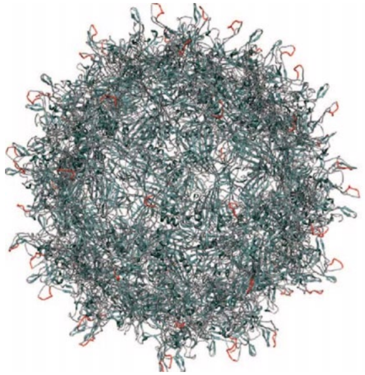


1. Build predictive model and test (*sequence→packaging* fitness).

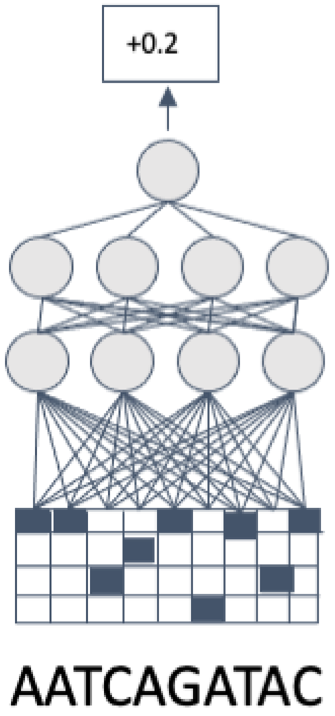


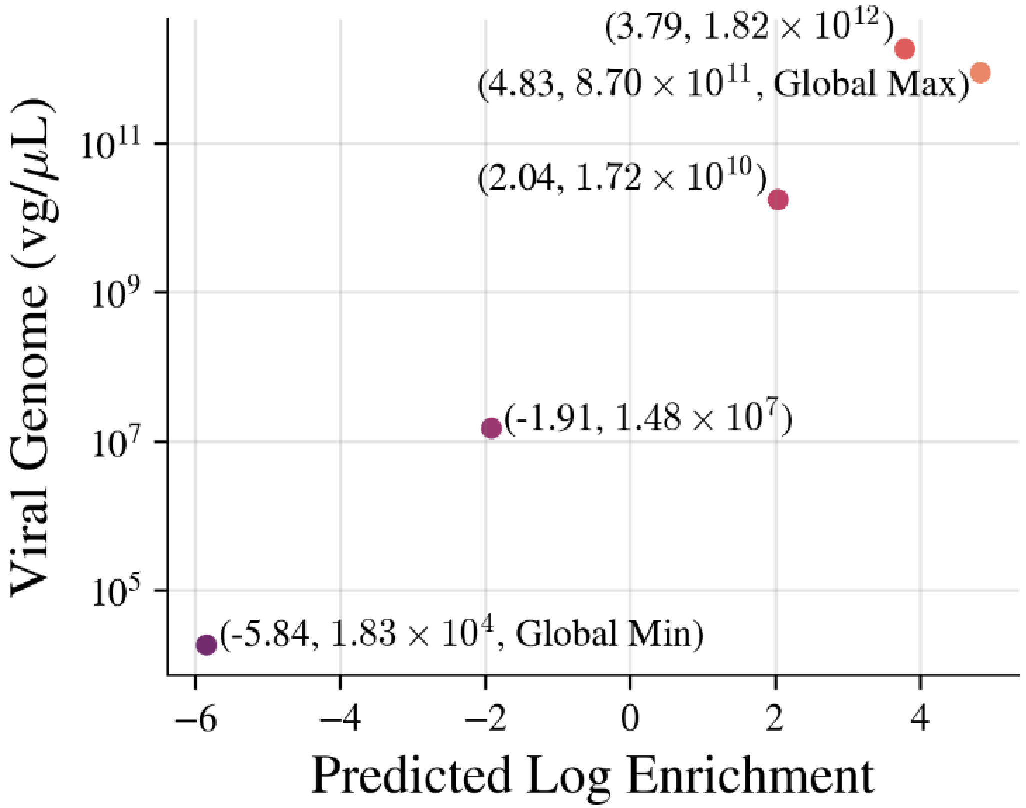


AAV library design



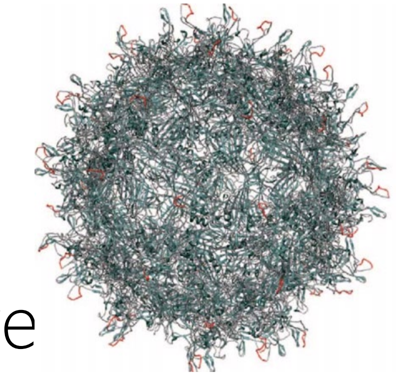
2. Wetlab validate model (measure titer directly)





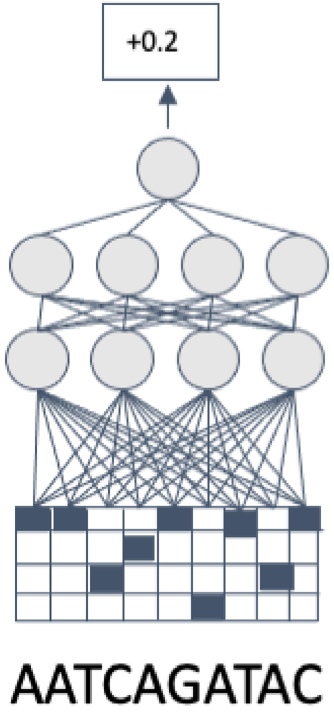
Sequences	Predicted Log Enrichment	Experimental Viral Titer (vg/μL)
LSSTTAA	4.834	8.70 × 10^{11}
DSRLSGT	3.793	1.82 × 10^{12}
LEPDAAL	2.044	1.72 × 10^{10}
IRWRATG	(-) 1.91	1.48 × 10^{7}
RWPRRVL	(-) 5.84	1.83 × 10^{4}

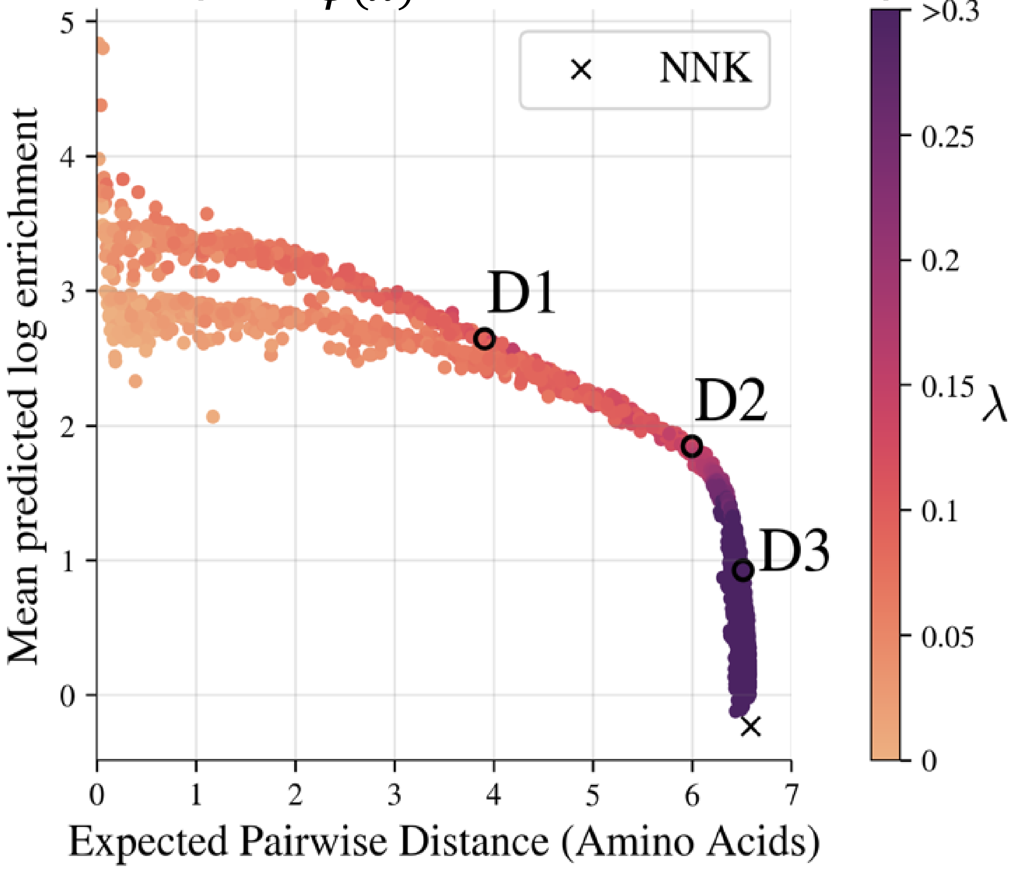
AAV library design



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

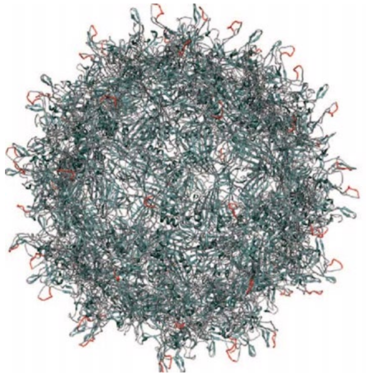
$$argmax_{\phi}\mathbb{E}_{p_{\phi(x)}}[f(x)] + \lambda H[p_{\phi}]$$



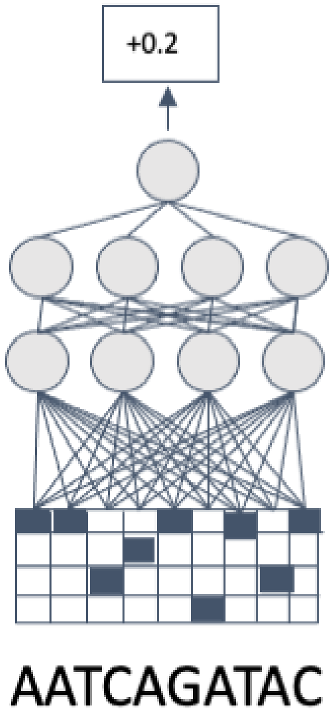


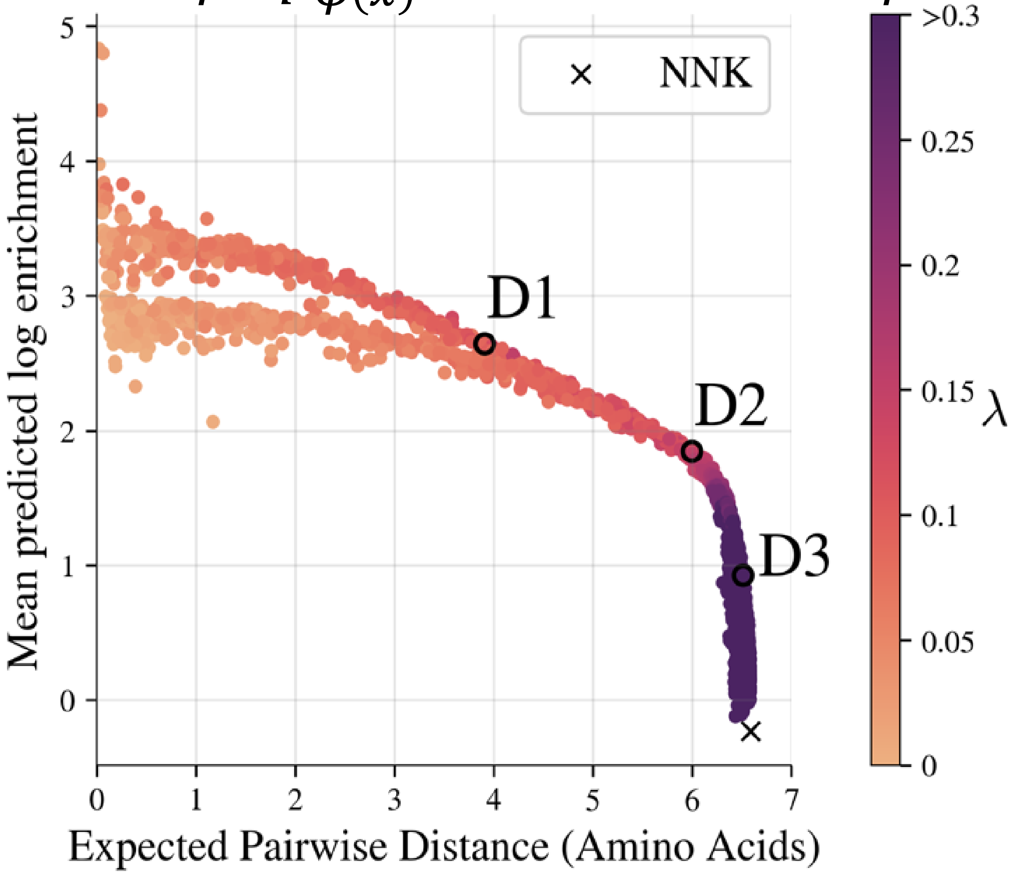
AAV library design

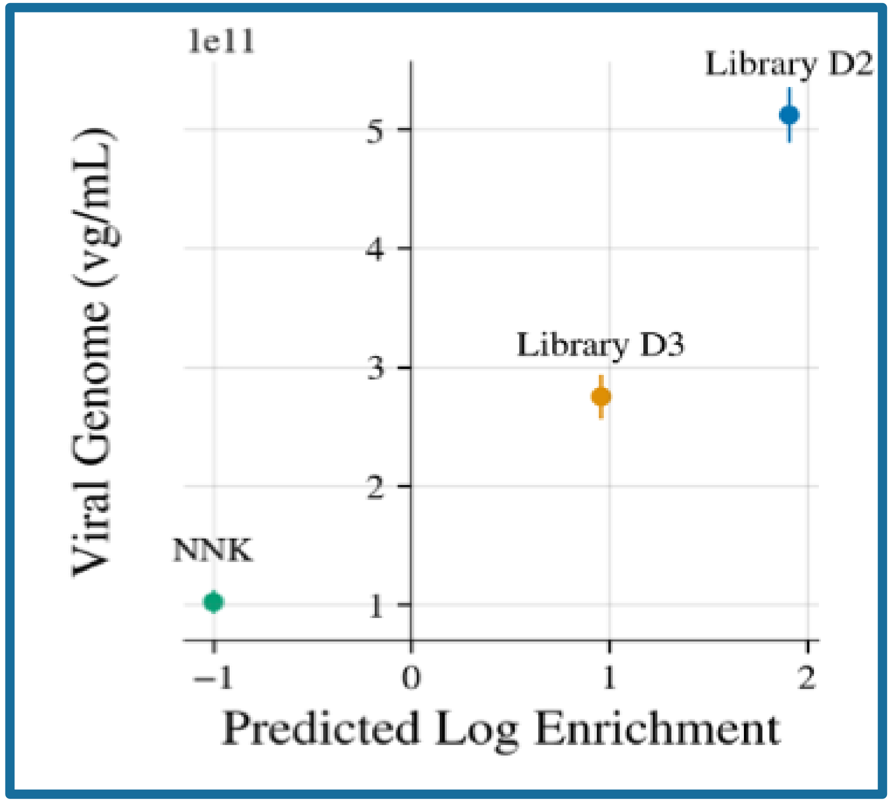
4. Validate in the lab.



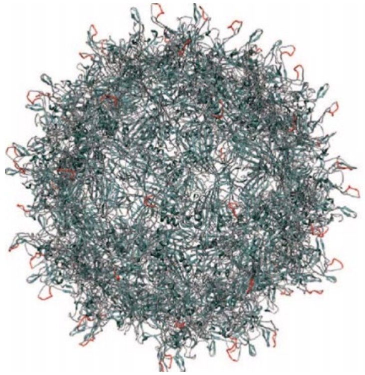
$$argmax_{\phi}\mathbb{E}_{p_{\phi(x)}}[f(x)] + \lambda H[p_{\phi}]$$



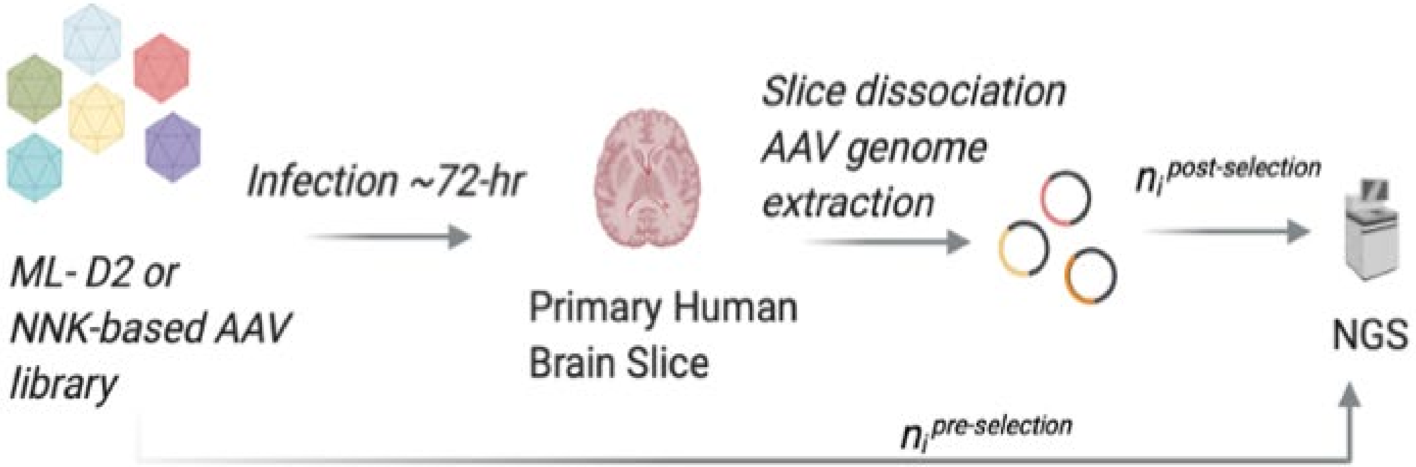




AAV library design



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



ML- D2 or NNK-based AAV library

Infection ~72-hr

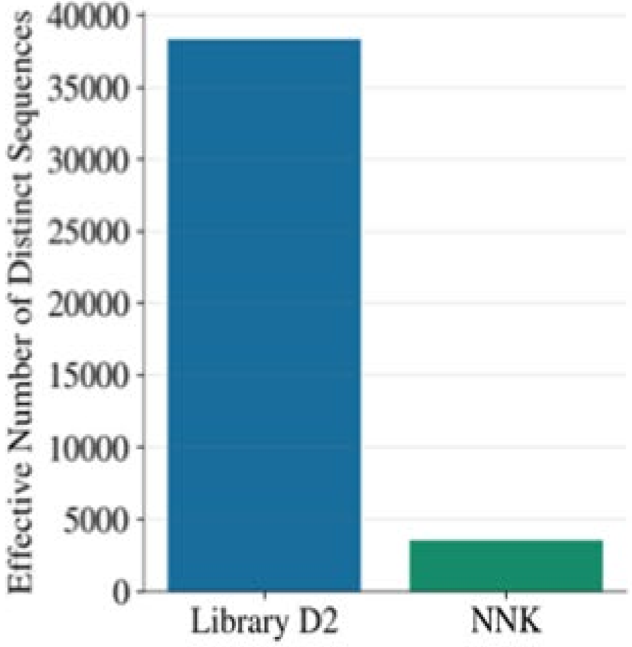
Primary Human Brain Slice

Slice dissociation AAV genome extraction

$n_i^{post\text{-}selection}$

NGS

$n_i^{pre\text{-}selection}$



ML library

currently used library

Parting thoughts: ML + protein engineering

1. Exciting times!
2. Are we close to ChatGPT4 for protein engineering? No.
3. Far less data than in text, vision—will need to be much more clever for the answers to “emerge” (unless same functions).
4. AlphaFold2 and progeny will help advance protein engineering.
5. Predicting function (generally) will remain difficult problem for a long time.
6. Whiplash---this field is moving quickly, hard to tell what is real/ useful.

The perpetual motion machine of AI-generated data and the distraction of "ChatGPT as scientist"

Jennifer Listgarten

EECS Department
University of California, Berkeley

Technical Report No. UCB/EECS-2023-239
November 30, 2023

http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-239.pdf

Since ChatGPT works so well, are we on the cusp of solving science with AI? Isn't AlphaFold2 suggestive that the potential of LLMs in biology and the sciences more broadly is limitless? Can we use AI itself to bridge the lack of data in the sciences in order to then train an AI? Herein we present a discussion of these topics.