

# Unleashing the power of protein engineering with artificial intelligence

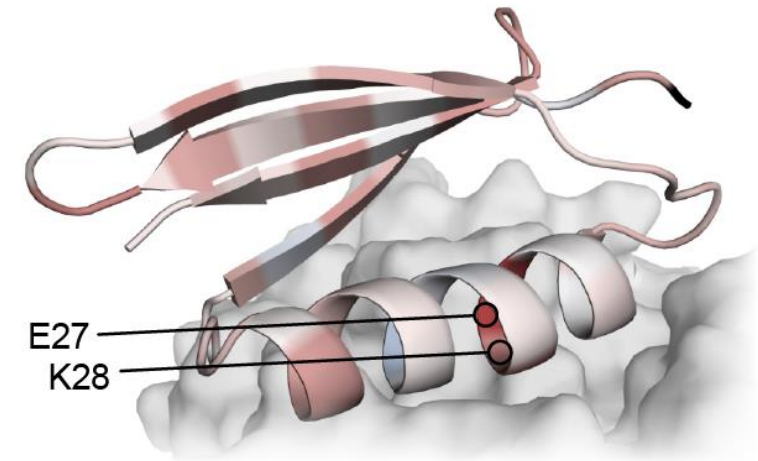
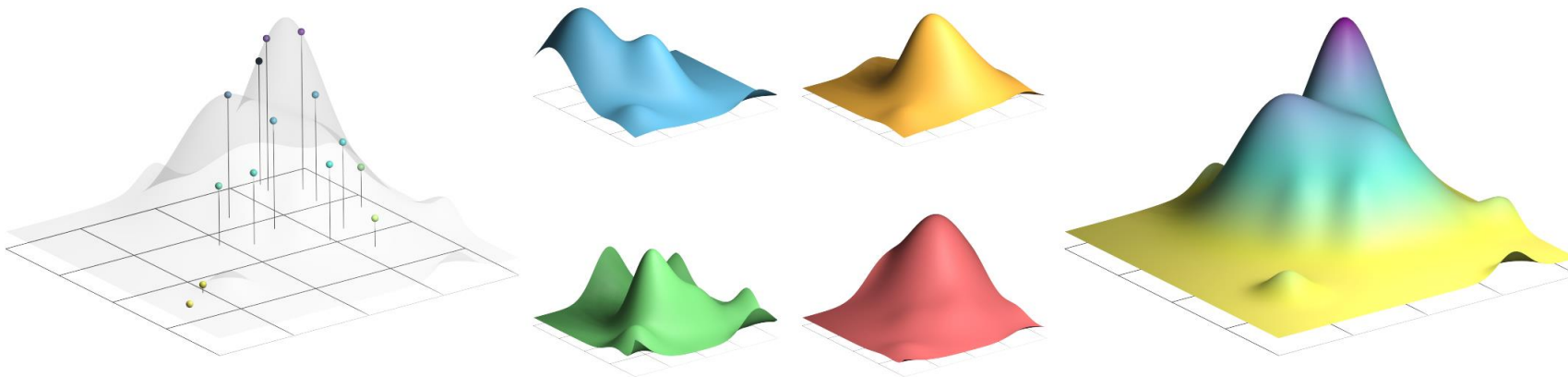
**Anthony Gitter**

Biostatistics and Medical Informatics, UW-Madison

Morgridge Institute for Research

Throughput Computing 2024

July 11th, 2024



✉ [gitter@biostat.wisc.edu](mailto:gitter@biostat.wisc.edu)

🏠 [www.biostat.wisc.edu/~gitter](http://www.biostat.wisc.edu/~gitter)

🐦 @anthonygitter



**“Our ability to perceive quality in nature begins, as in art, with the pretty. It expands through successive stages of the beautiful to values as yet uncaptured by language.”**

Aldo Leopold, *A Sand County Almanac*







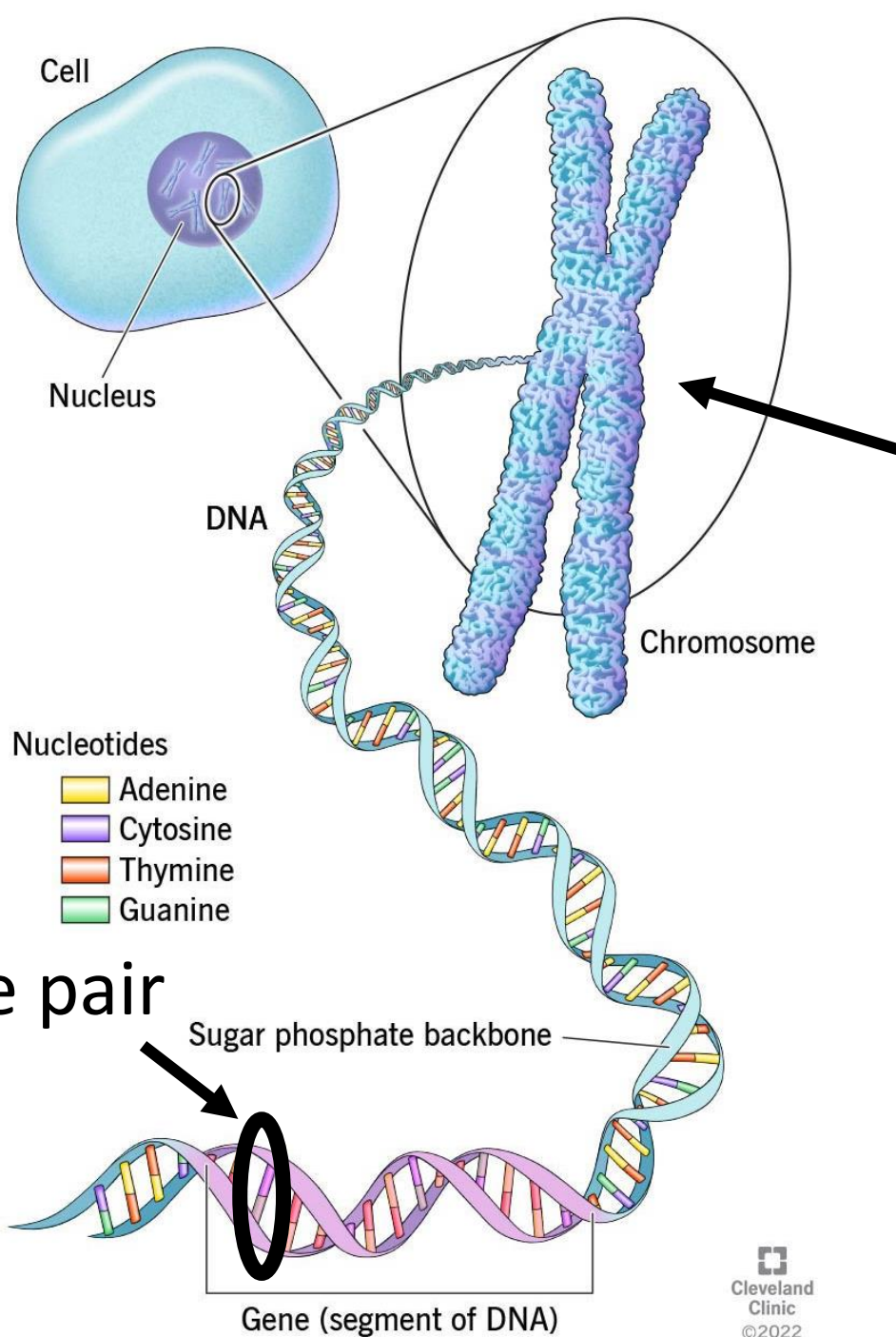




# Human

23 pairs of  
chromosomes

3.1 billion  
base pairs



# Polar bear

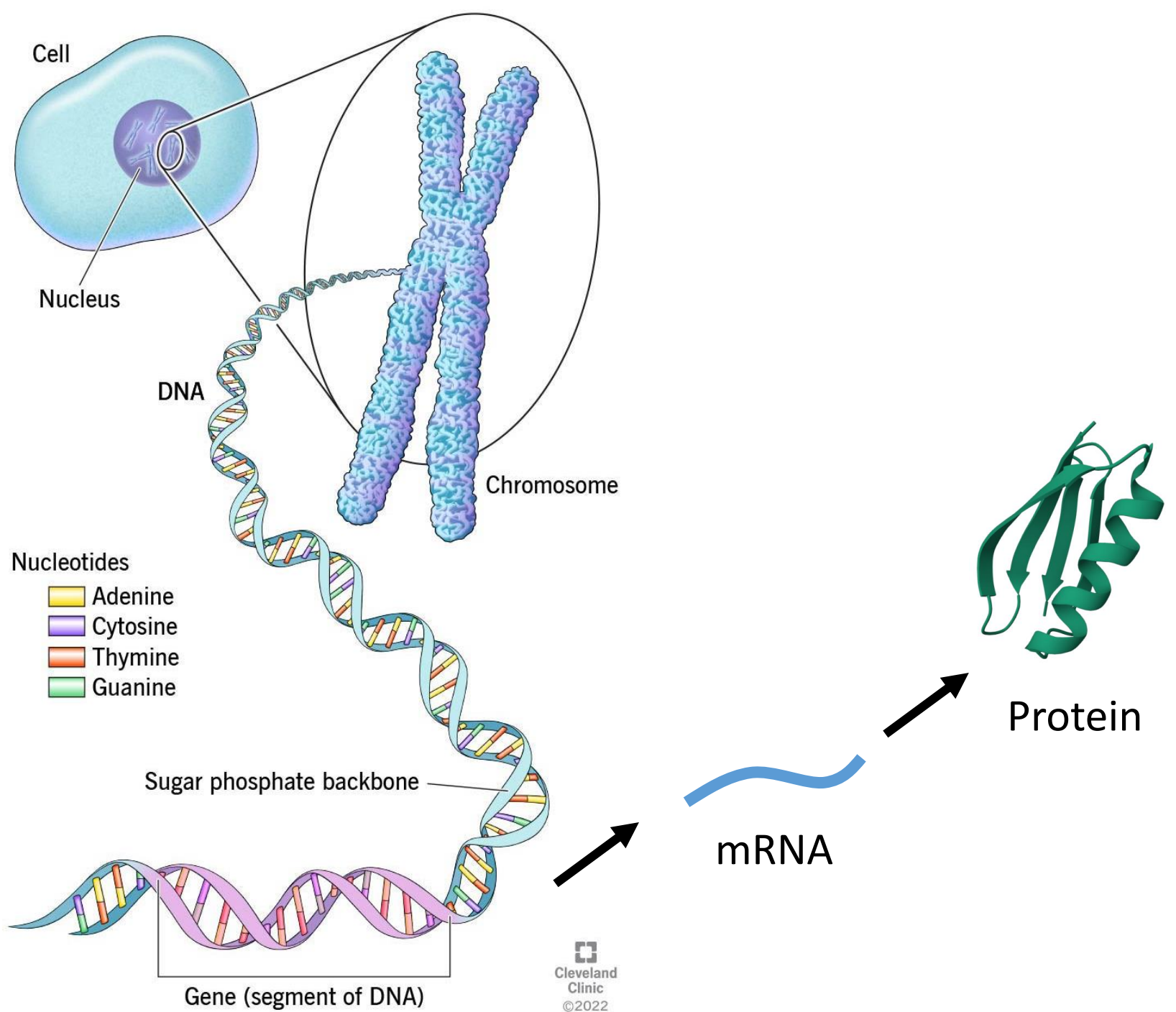
37 pairs of  
chromosomes

3.5 billion  
base pairs



**ATCCGACGA**

DNA characters

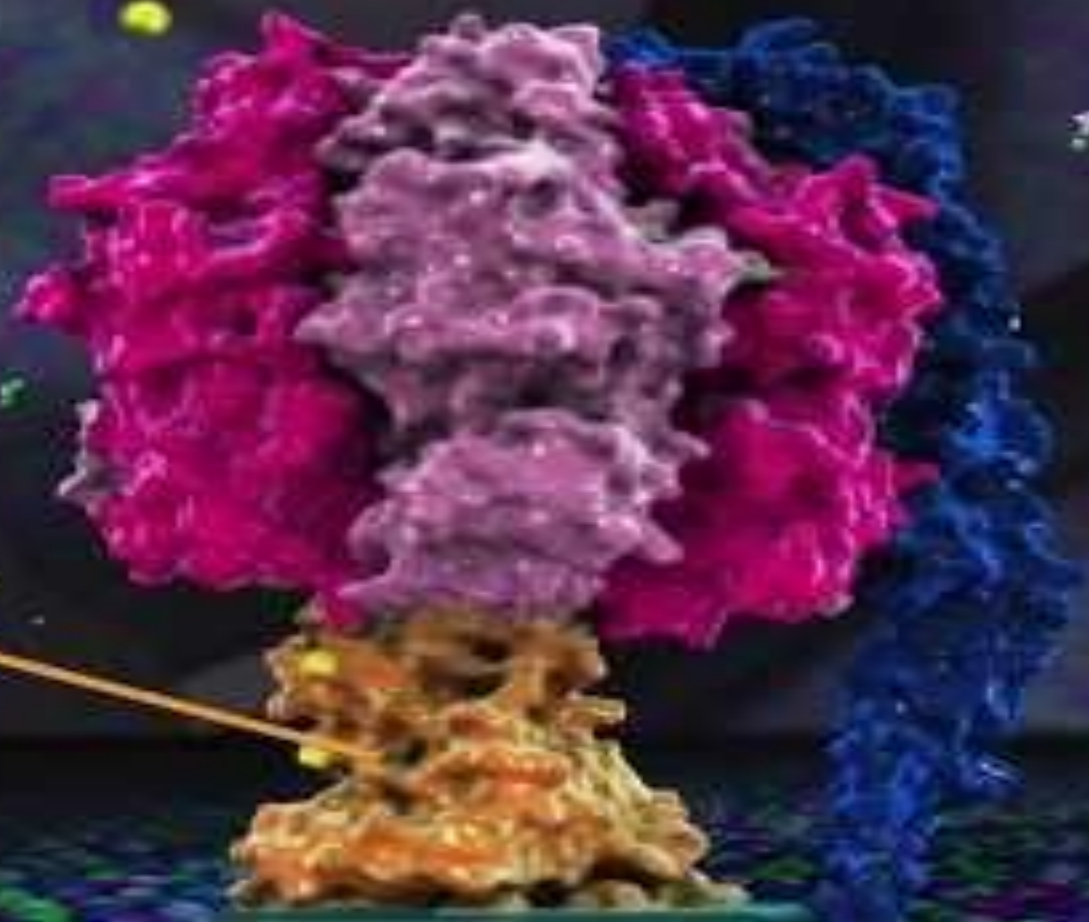








$\gamma$  and  $\epsilon$



H

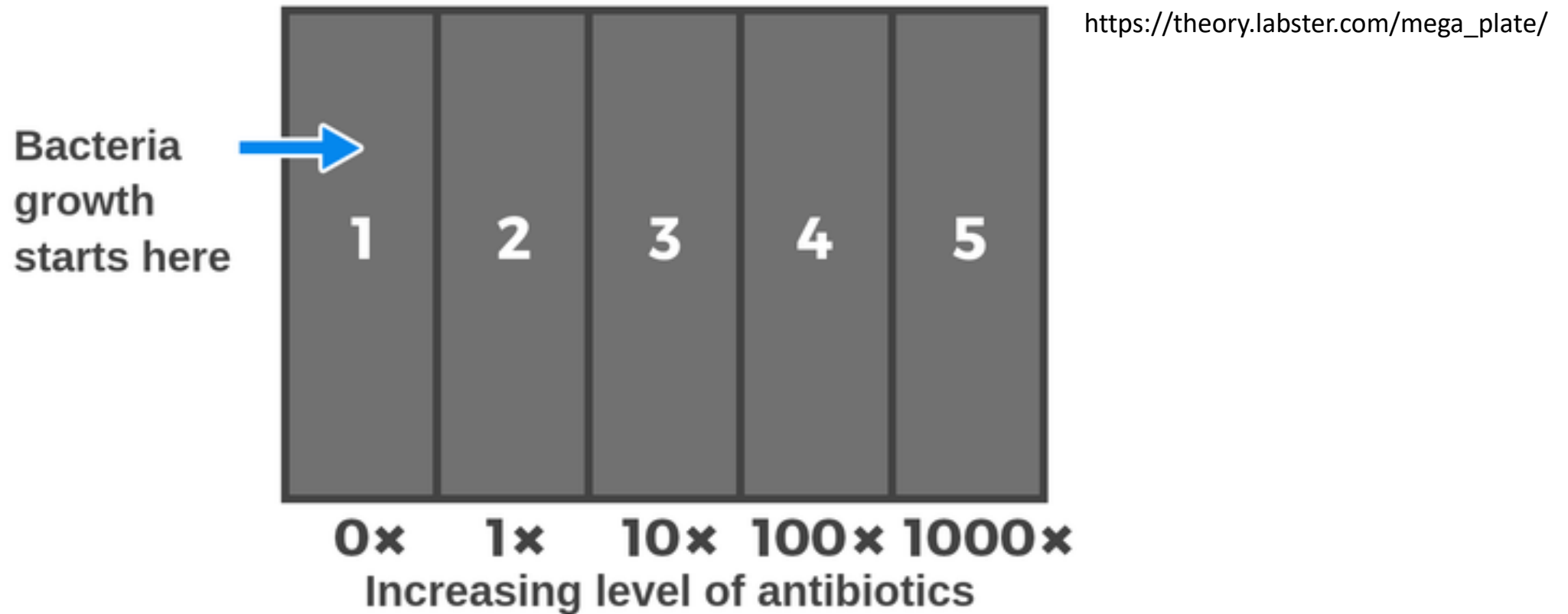
Evolutionary



# Synthetic biology



# Microbial Evolution and Growth Arena (MEGA) plate



Observing rapid bacterial evolution in the lab



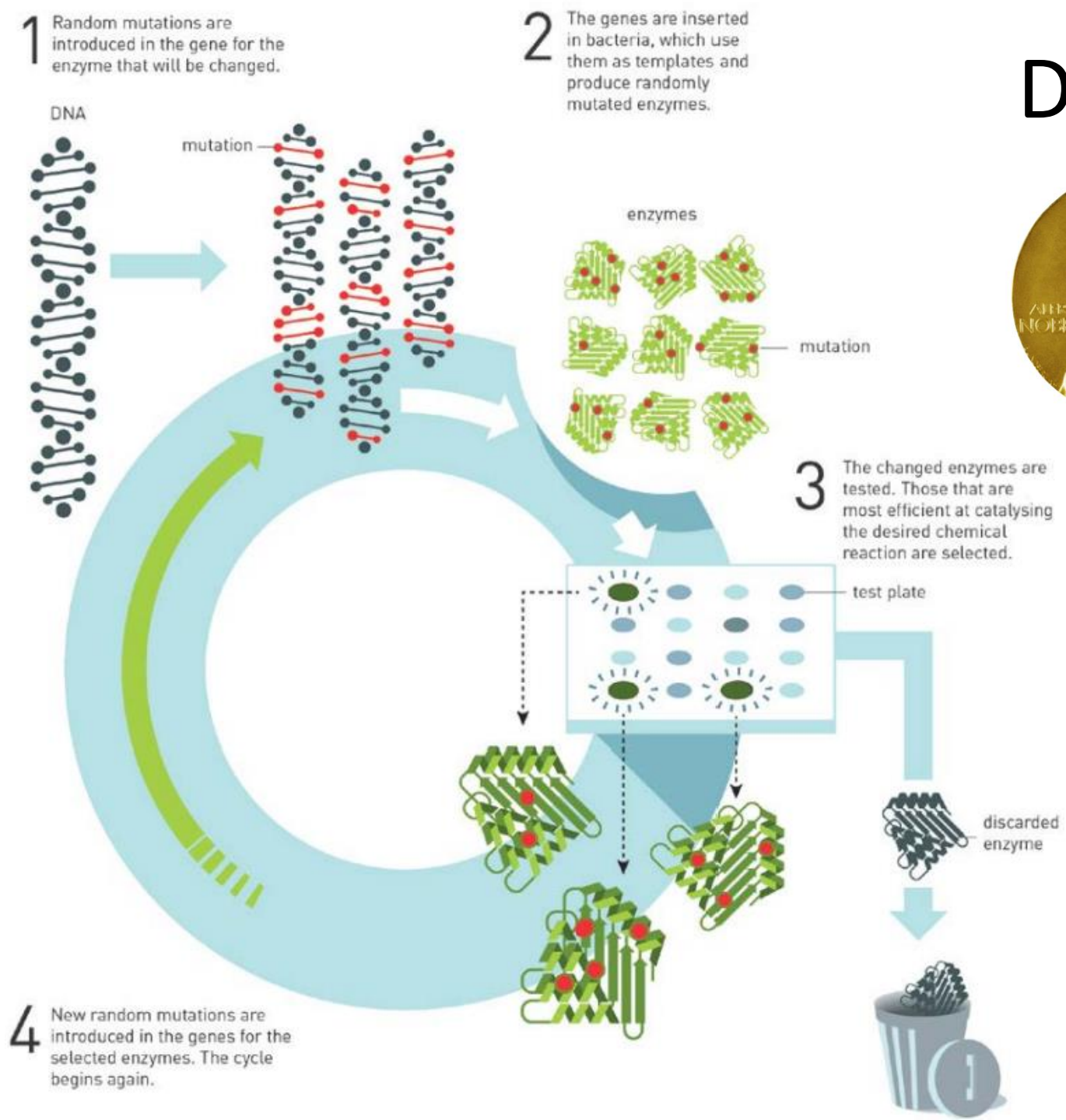




# Directed evolution



Frances H. Arnold  
awarded 1/2 the  
Nobel Prize in  
Chemistry 2018



1 Random mutations are introduced in the gene for the enzyme that will be changed.

2 The genes are inserted in bacteria, which use them as templates and produce randomly mutated enzymes.

3 The changed enzymes are tested. Those that are most efficient at catalysing the desired chemical reaction are selected.

4 New random mutations are introduced in the genes for the selected enzymes. The cycle begins again.

Scientific Background on the  
Nobel Prize in Chemistry 2018

# CRISPR/Cas9 genome editing



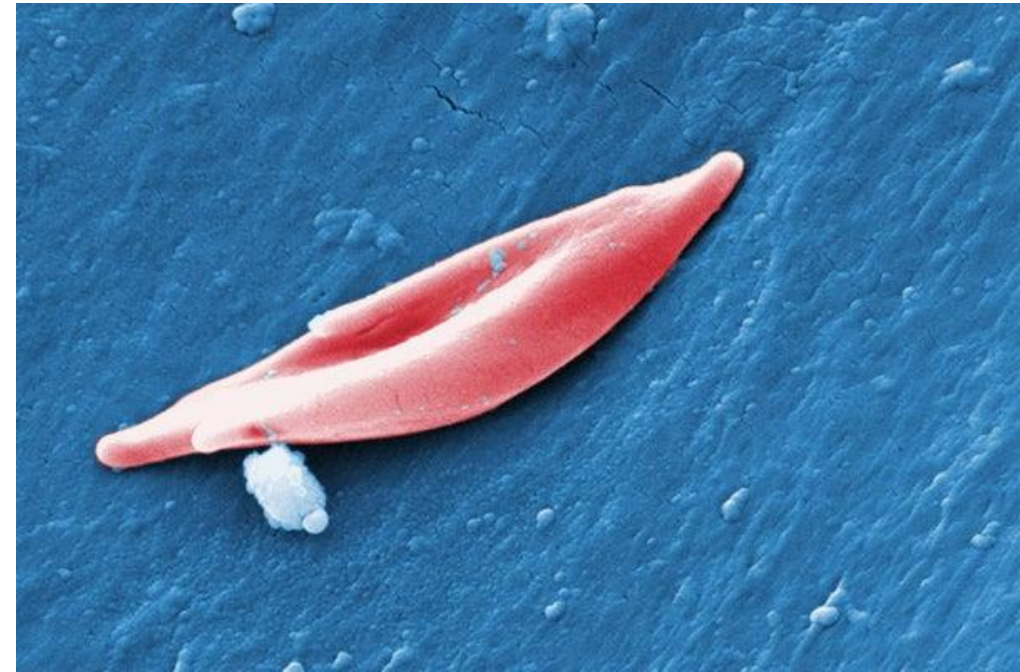
Emmanuelle Charpentier  
and Jennifer A. Doudna  
awarded the Nobel Prize  
in Chemistry 2020

BSIP SA/Alamy Stock Photo

Emmanuelle Charpentier and Jennifer A. Doudna. © Nobel Media. Ill. Niklas Elmehed

First CRISPR therapy FDA approved  
in December 2023

Casgevy for sickle cell disease





# AI-guided synthetic biology and protein engineering

# Why is protein engineering important?

Proteins can be modified to have biomedical applications



Proteins for  
herbicide  
tolerance

PDB 7M00



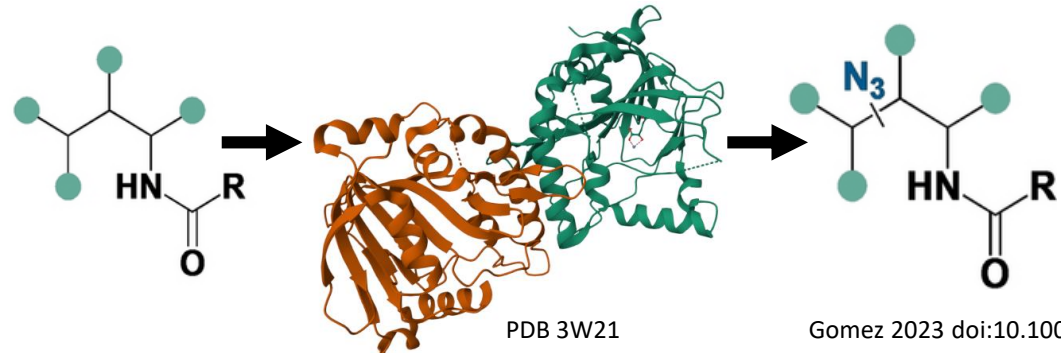
Proteins as  
biological  
factories



Proteins as  
medicines



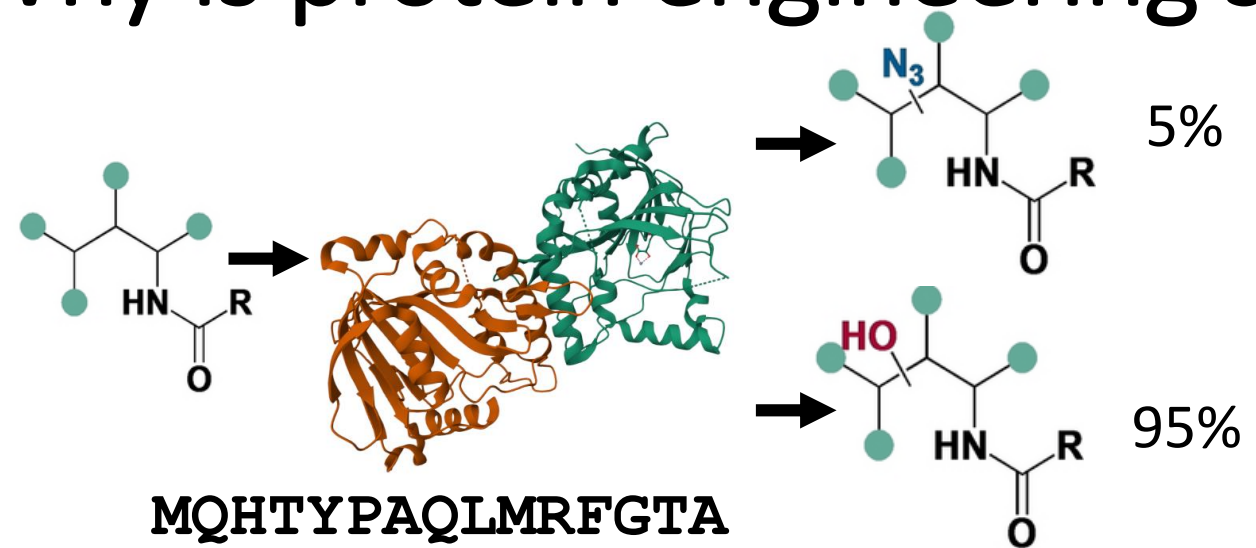
PDB 2QMT



Gomez 2023 doi:10.1002/anie.202301370



# Why is protein engineering so hard?



Natural proteins do not initially do the job we want



Need to engineer (modify) them by changing their sequence

**MQHTYPAQLMRFGTA**  
Amino acid sequence

Where do we modify the sequence? How many changes? Which changes?

MQHTYPAQLMRFGTAARAEHMTIAAAIHALDADEADAI VMDIVPDGERDAWDDDEGFSSSPFTKNAHHAGIVATSVTLGQLQREQGDKLVSKAAEYFGIACRVNDGLRTRFRVRLFSDALDAKPLTI GHDYEV EFL LATRRV  
YEPFEAPFNFAPH CDDVSYGRDTVNWPLKRSFPRQLGGFLT IQGADNDAGMVMWDNRPE SRAALDEMHA EYRETGAIAALERA AKIMLKPQPGQLTLFQSKNLHAI ERCTSTRRTMGLFLIHTEDGWRMFD

**MQHTYPA**S**LMRF**D**TA**

**MQHT**K**PAQLMRC**G**TA**

**MQ**P**PYPAQLMRF**G**PA**

**M**N**HTY**W**AQLMRF**G**T**E****





# Supervised learning to predict protein function

Sequence-function  
examples

Variant	Score
D138N, K140E	-2.35
N127A	-4.00
K180N, A182D	-4.14
D74G, I126T	1.20

10s-100s of thousands of  
protein variants characterized  
by **deep mutational scanning**

Gelman *et al.* *PNAS* 2021 doi:10.1073/pnas.2104878118  
<https://github.com/gitter-lab/nn4dms>

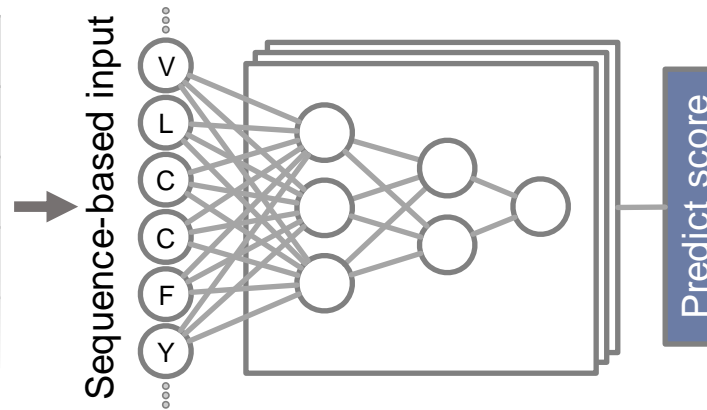


Sam Gelman

# Supervised learning to predict protein function

Sequence-function examples

Variant	Score
D138N, K140E	-2.35
N127A	-4.00
K180N, A182D	-4.14
D74G, I126T	1.20



10s-100s of thousands of protein variants characterized by **deep mutational scanning**

Tested **linear regression** and **fully connected, sequence convolutional**, and **graph convolutional** neural networks

Gelman et al. PNAS 2021 DOI:10.1073/pnas.2104878118  
<https://github.com/gitter-lab/nn4dms>



Sam Gelman

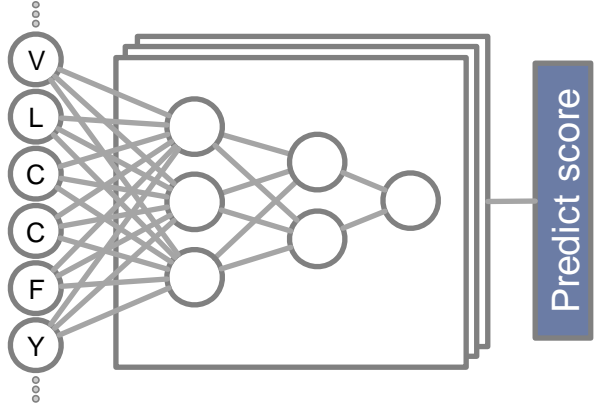


# Supervised learning to predict protein function

Sequence-function examples

Variant	Score
D138N, K140E	-2.35
N127A	-4.00
K180N, A182D	-4.14
D74G, I126T	1.20

Sequence-based input



Predict scores for new variants

Variant	Score
G177L, M189T	?????
L142K	?????

Variant	Score
G177L, M189T	0.003
L142K	-0.421

10s-100s of thousands of protein variants characterized by **deep mutational scanning**

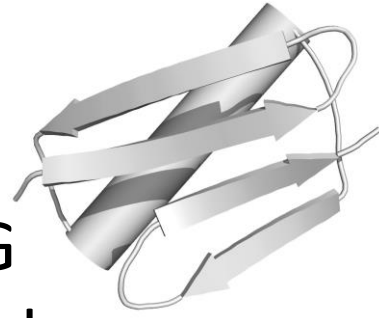
Tested **linear regression** and **fully connected, sequence convolutional**, and **graph convolutional** neural networks

Gelman et al. PNAS 2021 DOI:10.1073/pnas.2104878118  
<https://github.com/gitter-lab/nn4dms>



Sam Gelman

# Example: predicting GB1 IgG binding



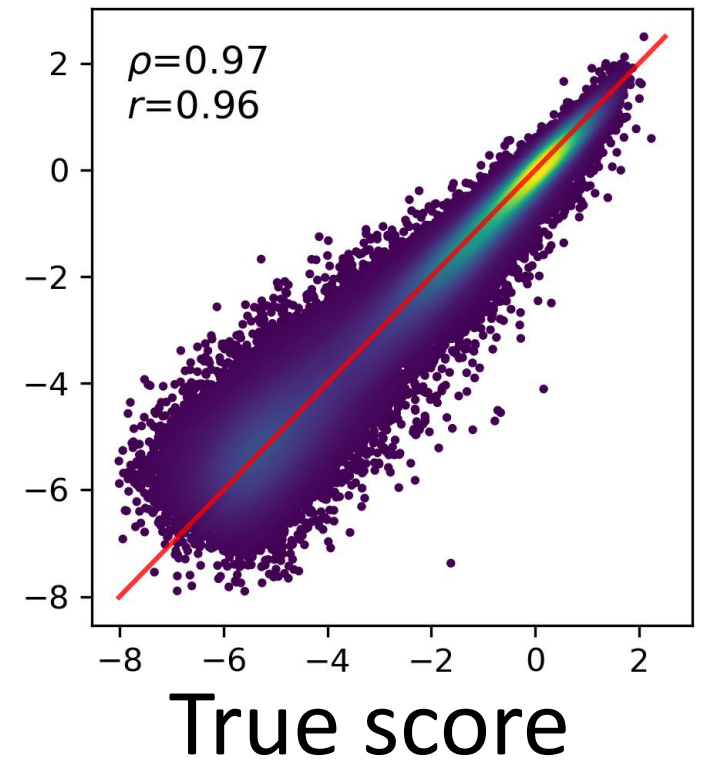
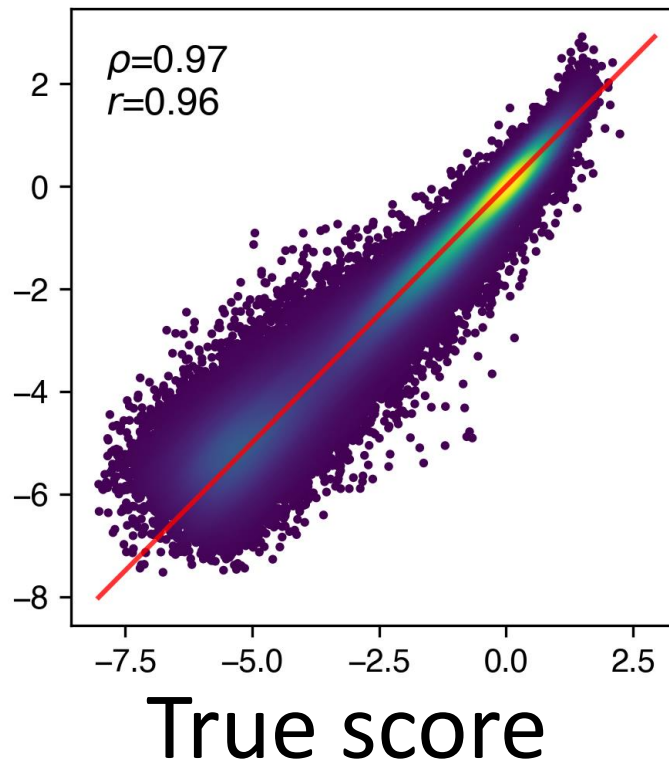
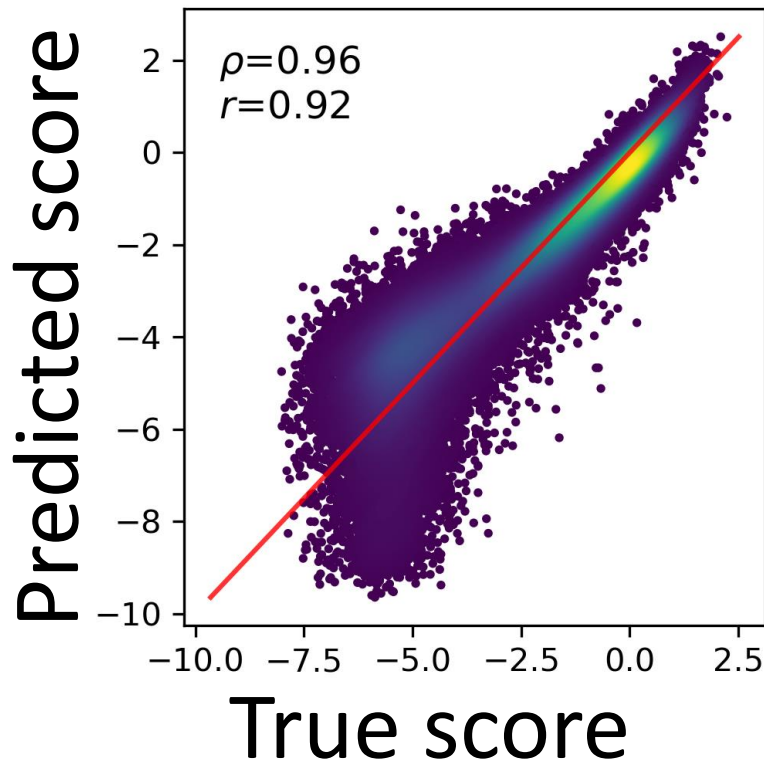
Small domain from streptococcal protein G that binds mammalian IgG

Train supervised learning models on 16,384 sequence-function examples

Linear regression

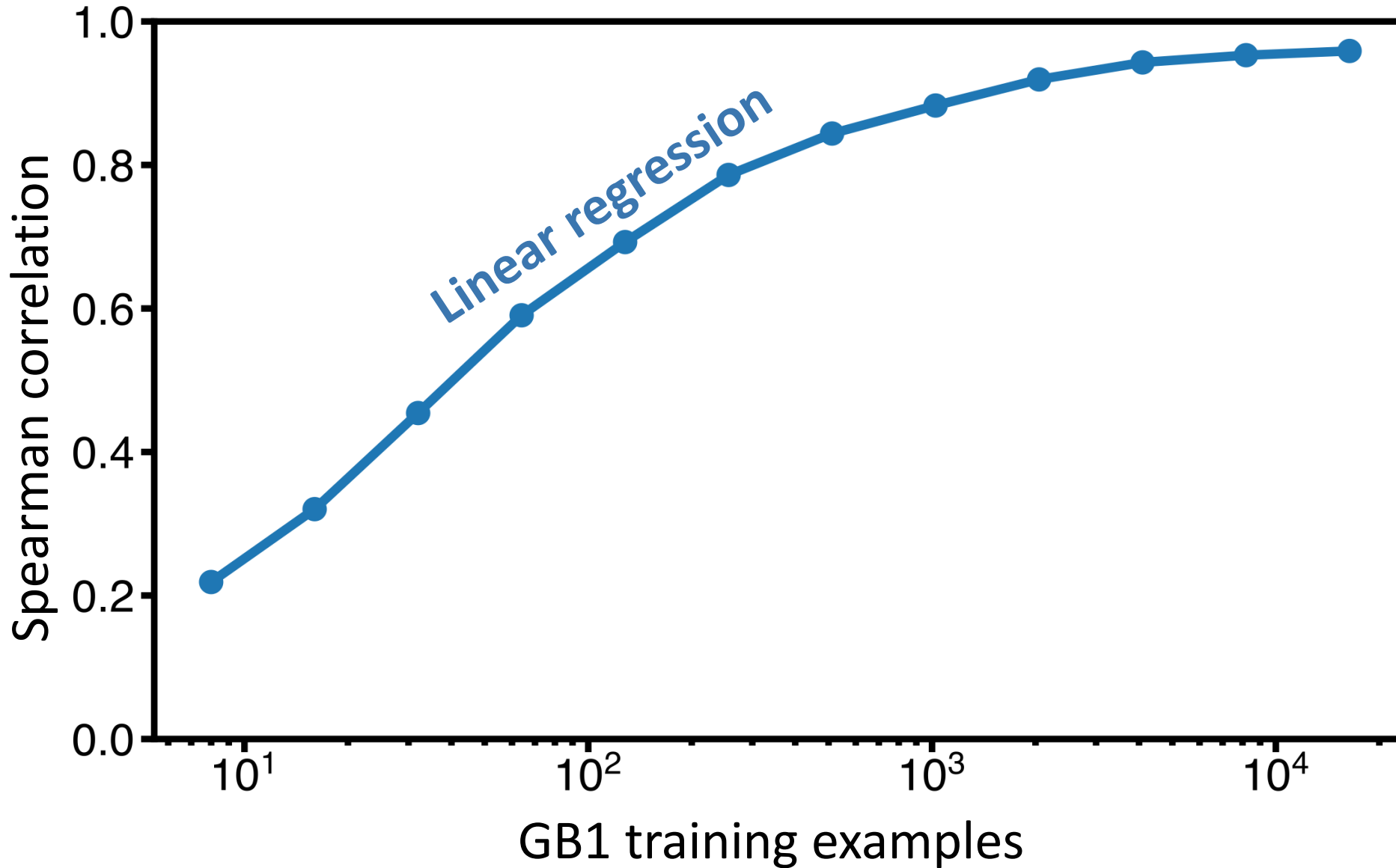
Fully connected

Sequence convolution

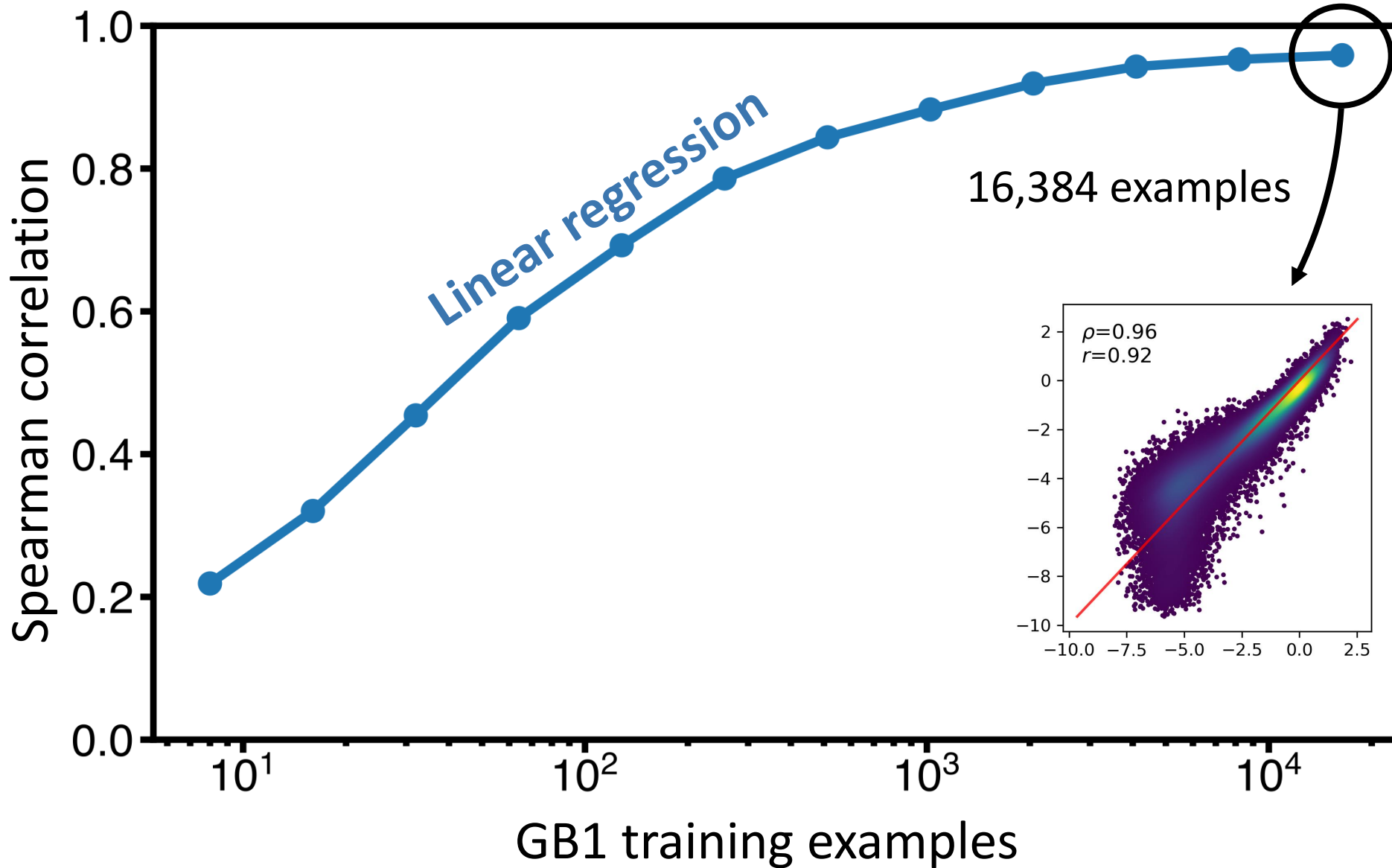




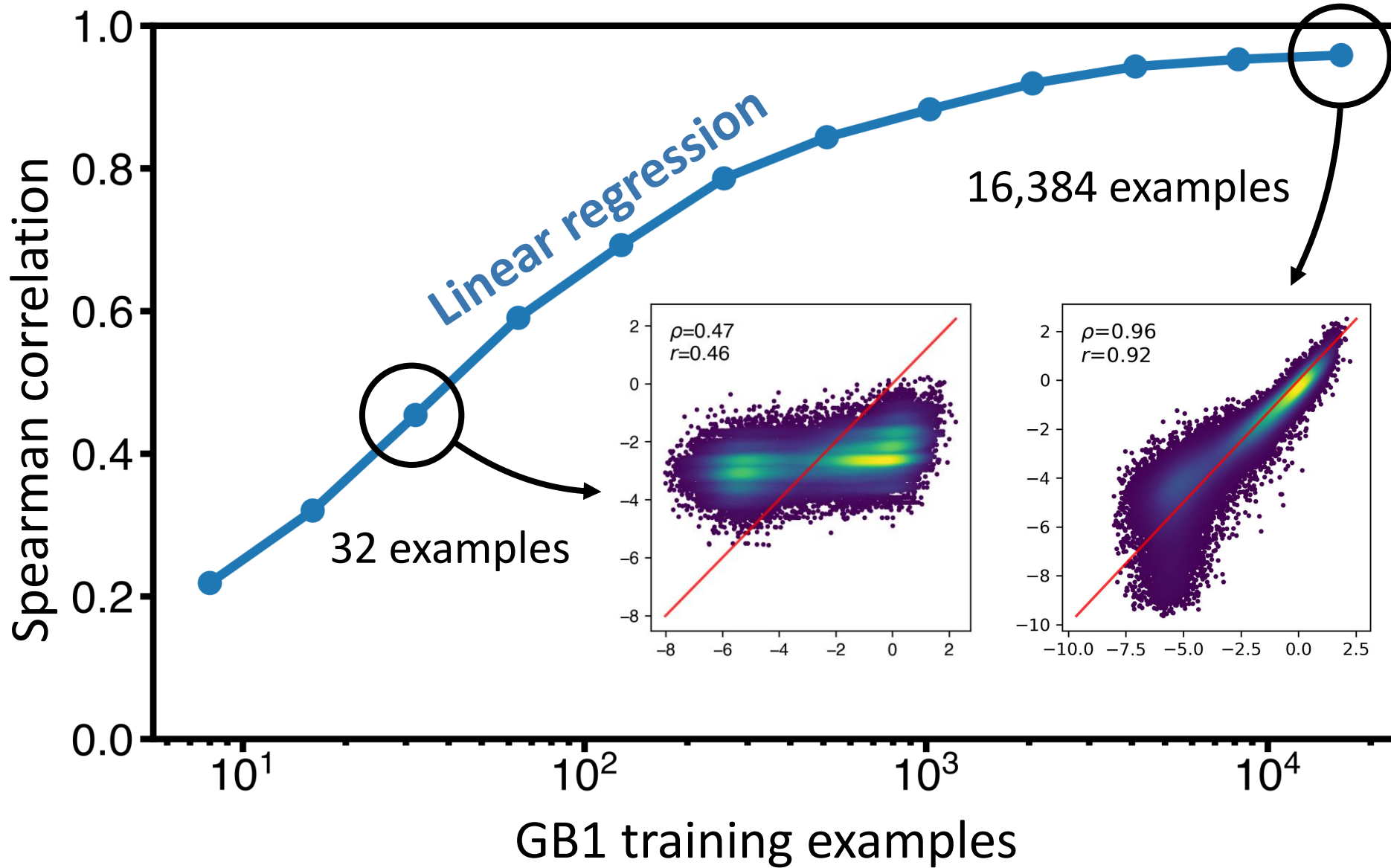
# Models struggle with less training data



# Models struggle with less training data



# Models struggle with less training data





# METL: Mutational Effect Transfer Learning

Transfer learning based on **biophysical simulations**

- 1** Simulate protein variants with Rosetta
- 2** Train a model to predict the Rosetta energies
- 3** Transfer representation to experimental data



Sam Gelman

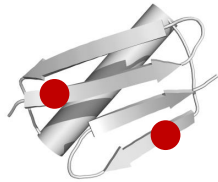
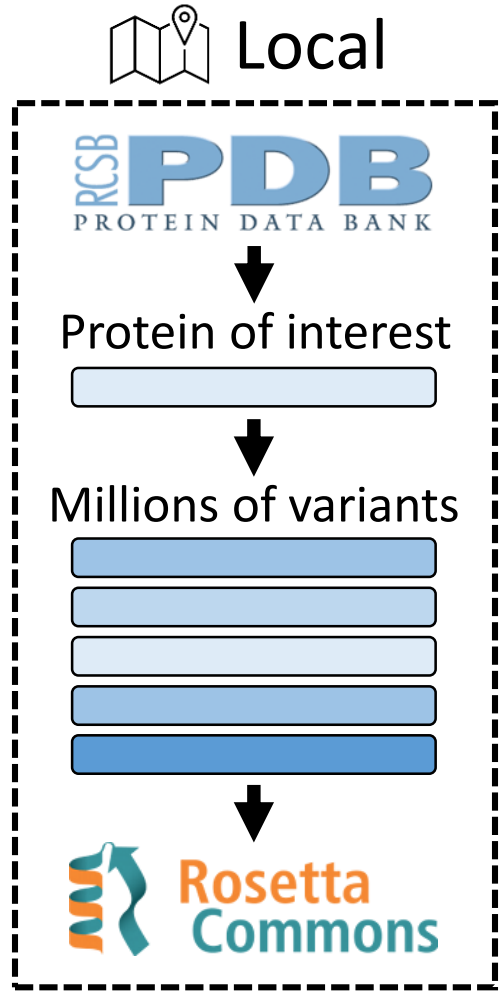


Bryce Johnson

# METL step 1. Generate biophysical simulations

**Local mode:**  
Generate  
biophysical  
simulations for  
one specific  
protein

GB1 example



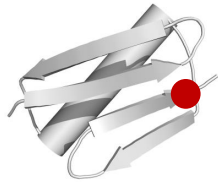
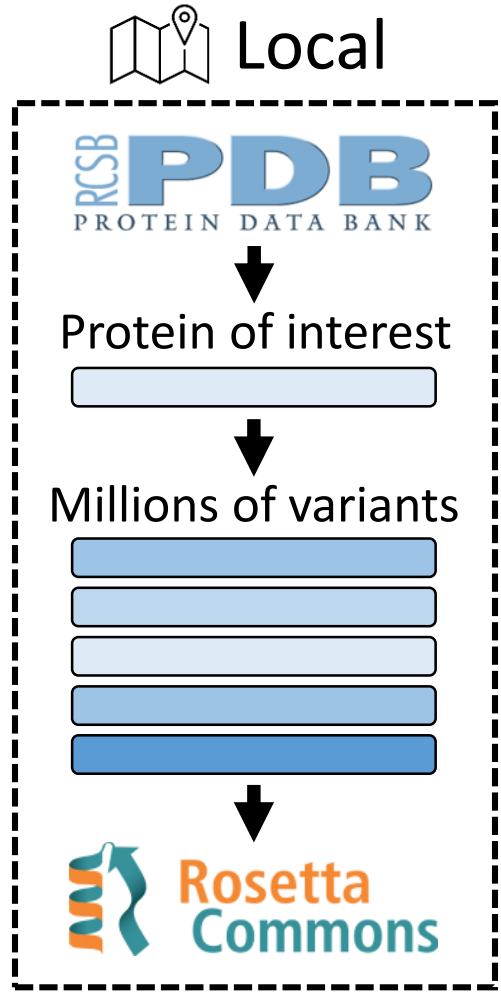
		Energies (55 total)			
PDB	Variant	total score	dslf fa13	fa atr	...
2QMT	E34R,L53M	-237.1	0.12	24.2	...

+ millions of simulated variant effects

# METL step 1. Generate biophysical simulations

**Local mode:**  
Generate biophysical simulations for one specific protein

GB1 example



		Energies (55 total)			
PDB	Variant	total score	dslf fa13	fa atr	...
2QMT	E34R,L53M	-237.1	0.12	24.2	...
2QMT	N2A	-222.8	0.34	24.3	...

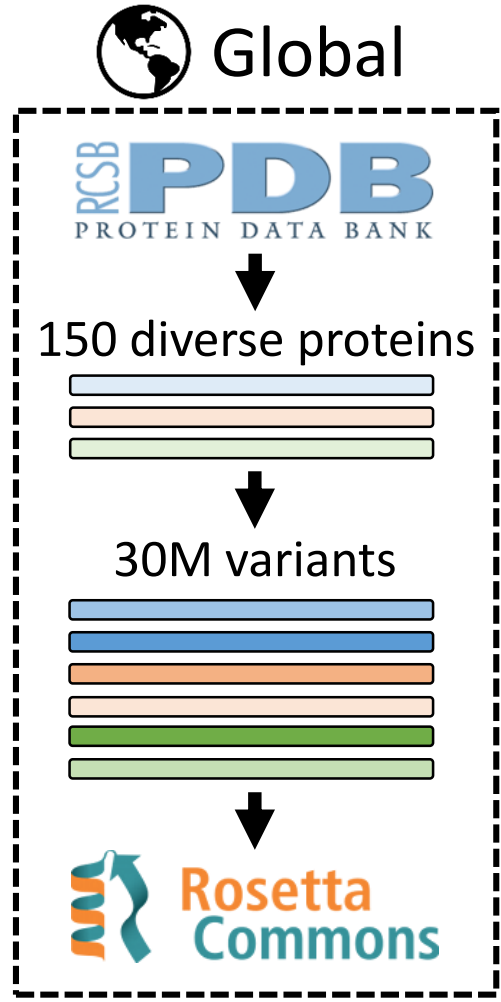
+ millions of simulated variant effects



# METL step 1. Generate biophysical simulations

**Global mode:**  
Generate  
biophysical  
simulations for  
diverse protein  
structures

Runtime **≈35,000**  
**compute-days**  
(powered by CHTC  
and OSG Consortium)



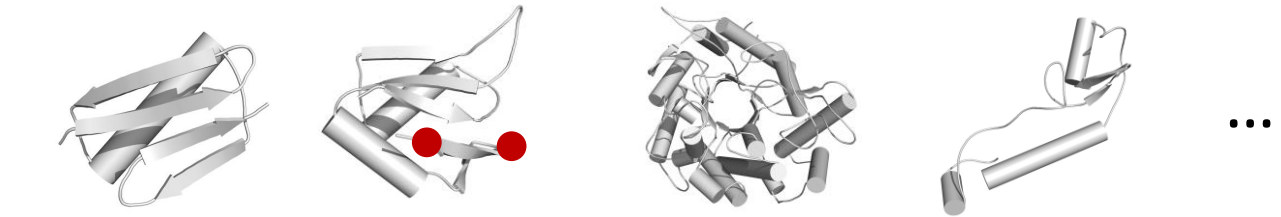
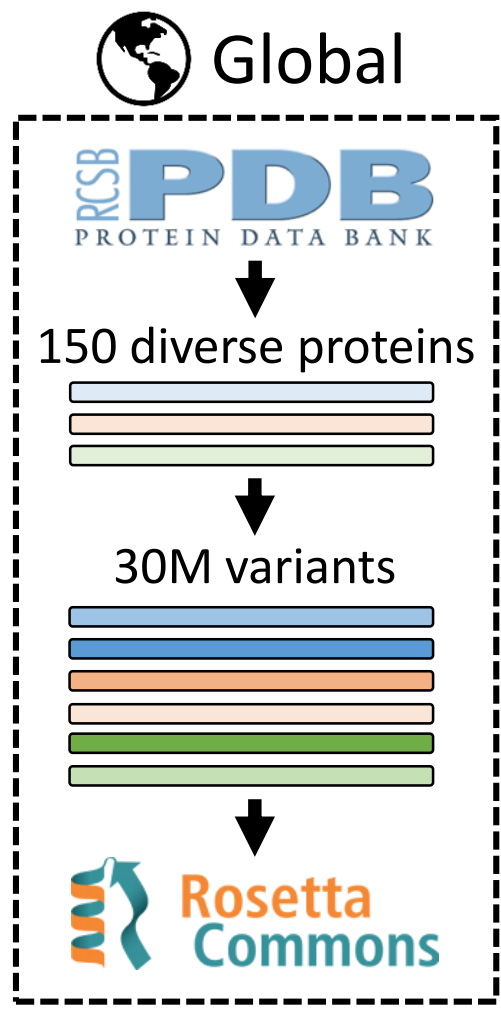
		Energies (55 total)			
PDB	Variant	total score	dslf fa13	fa atr	...
2QMT	E34R,L53M	-237.1	0.12	24.2	...
2QMT	N2A	-222.8	0.34	24.3	...

+ millions of simulated variant effects

# METL step 1. Generate biophysical simulations

**Global mode:**  
Generate  
biophysical  
simulations for  
diverse protein  
structures

Runtime  $\approx$  **35,000**  
**compute-days**  
(powered by CHTC  
and OSG Consortium)

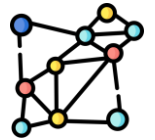


		Energies (55 total)			
PDB	Variant	total score	dslf fa13	fa atr	...
2QMT	E34R,L53M	-237.1	0.12	24.2	...
2QMT	N2A	-222.8	0.34	24.3	...
1CVJ	K6N,A18D	-119.2	0.11	50.1	...

+ millions of simulated variant effects

# METL step 2. Train model to predict energies

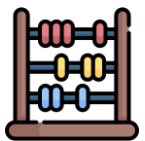
Model learns a representation based on the Rosetta energies



Transformer encoder



Relative position encoding based on protein 3D structure



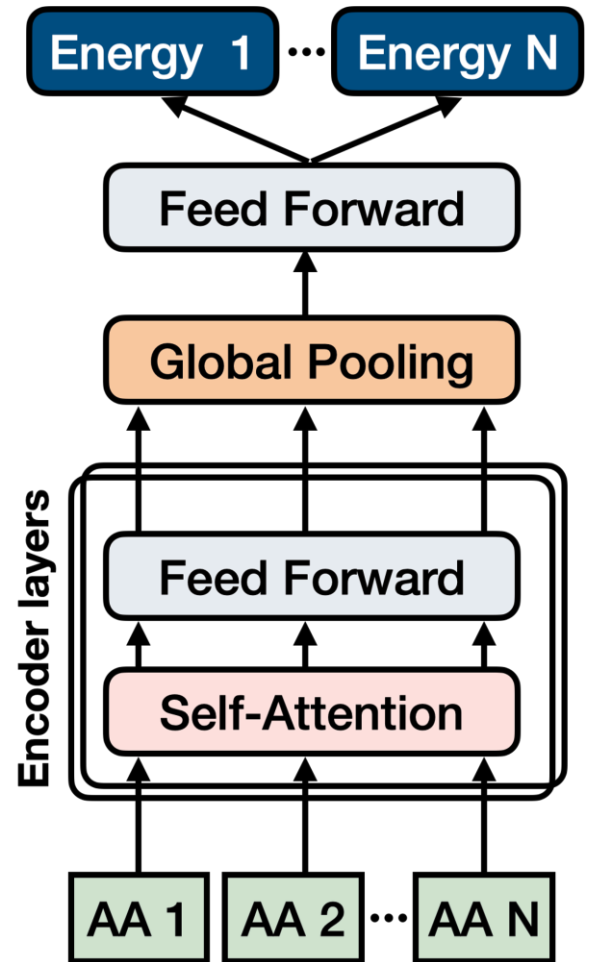
Local → 2M params  
Global → 20M params

Predicts outputs of biophysical simulations

Pooling supports different seq lengths

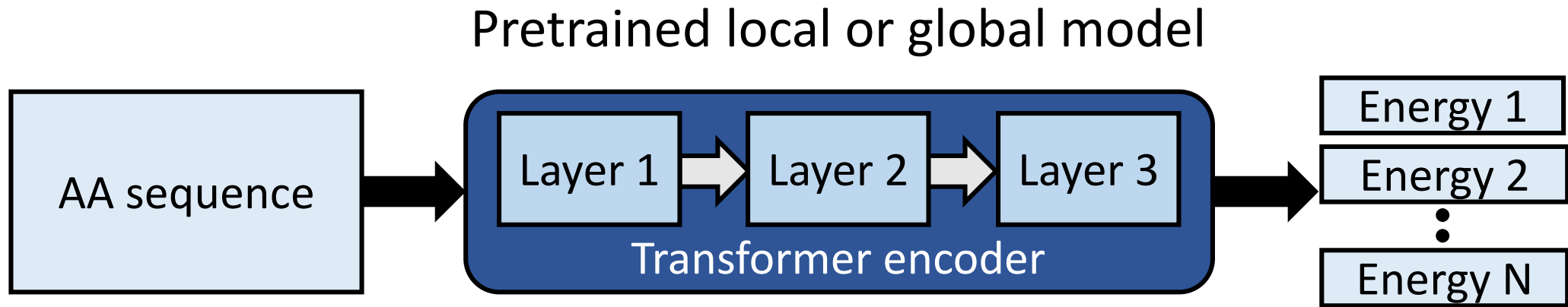
Based on attention mechanism

Amino acid sequence input

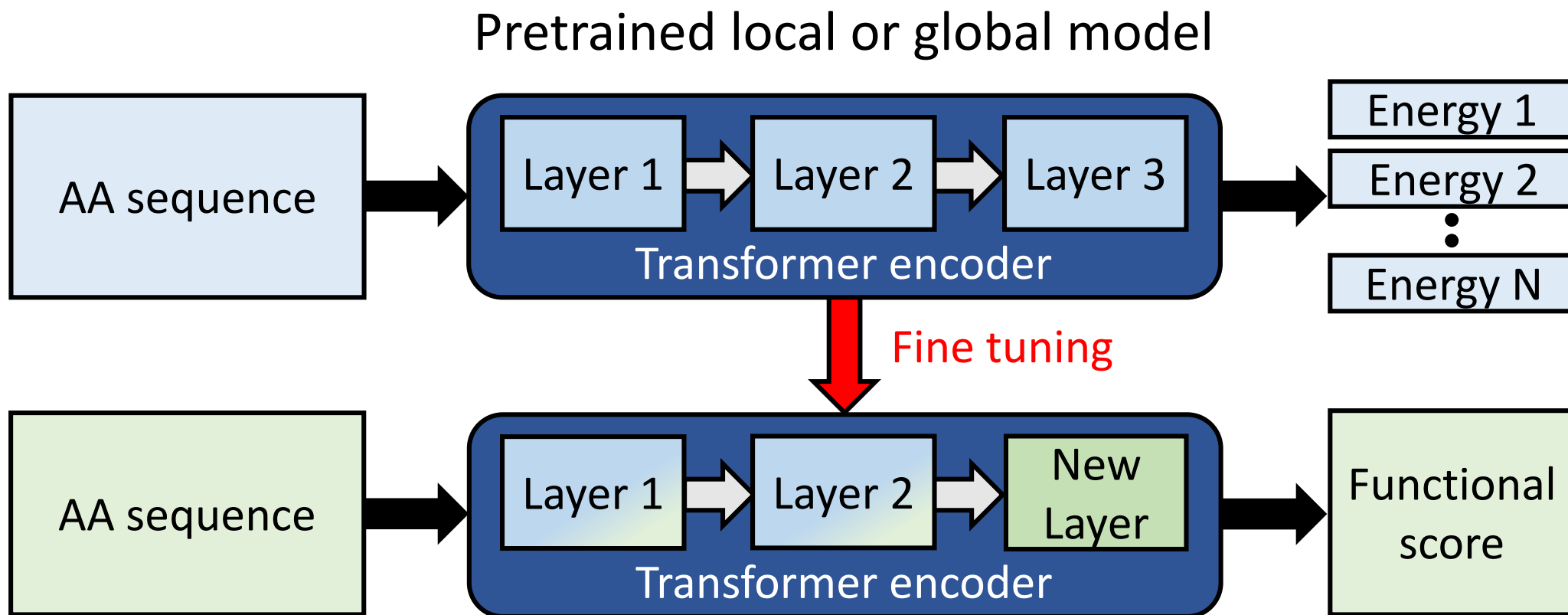




# METL step 3. Transfer to experimental data




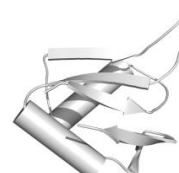
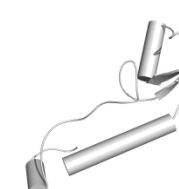


# METL step 3. Transfer to experimental data



10s to 1000s of experimental sequence-function examples

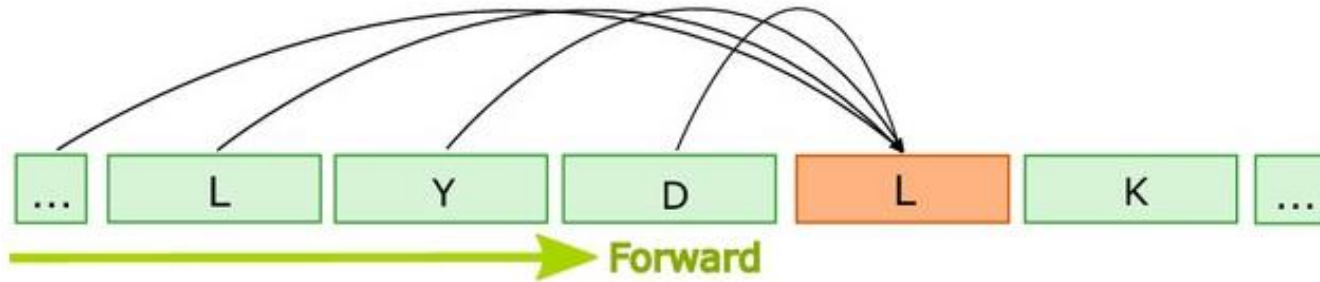
# Evaluating METL on diverse protein datasets

	<b>Dataset</b>	<b>Description</b>	<b>Assay</b>	<b>Len</b>	<b>Examples</b>
	avGFP	Green fluorescent protein	Brightness	237	51,714
	DLG4	Postsynaptic density protein 95 (PDZ3 domain)	CRIP binding	66	517,653
	GB1	Protein G (B1 domain)	IgG binding	56	536,084
	Pab1	Poly(A)-binding protein (RRM2 domain)	mRNA binding	75	37,710
	Ube4b	Ubiquitination factor E4B (U-box domain)	Ubiquitin ligase activity	102	88,375

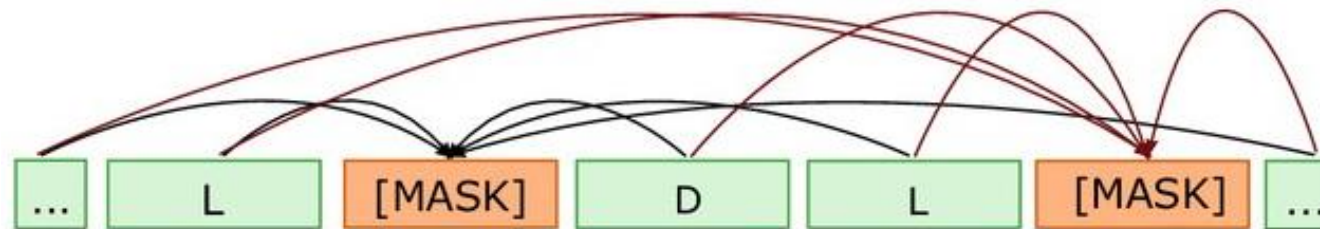


# Compare to evolutionary protein language models

Predict next amino acid in protein sequence



Predict hidden amino acids in protein sequence



Machine learning models trained on millions of natural protein sequences

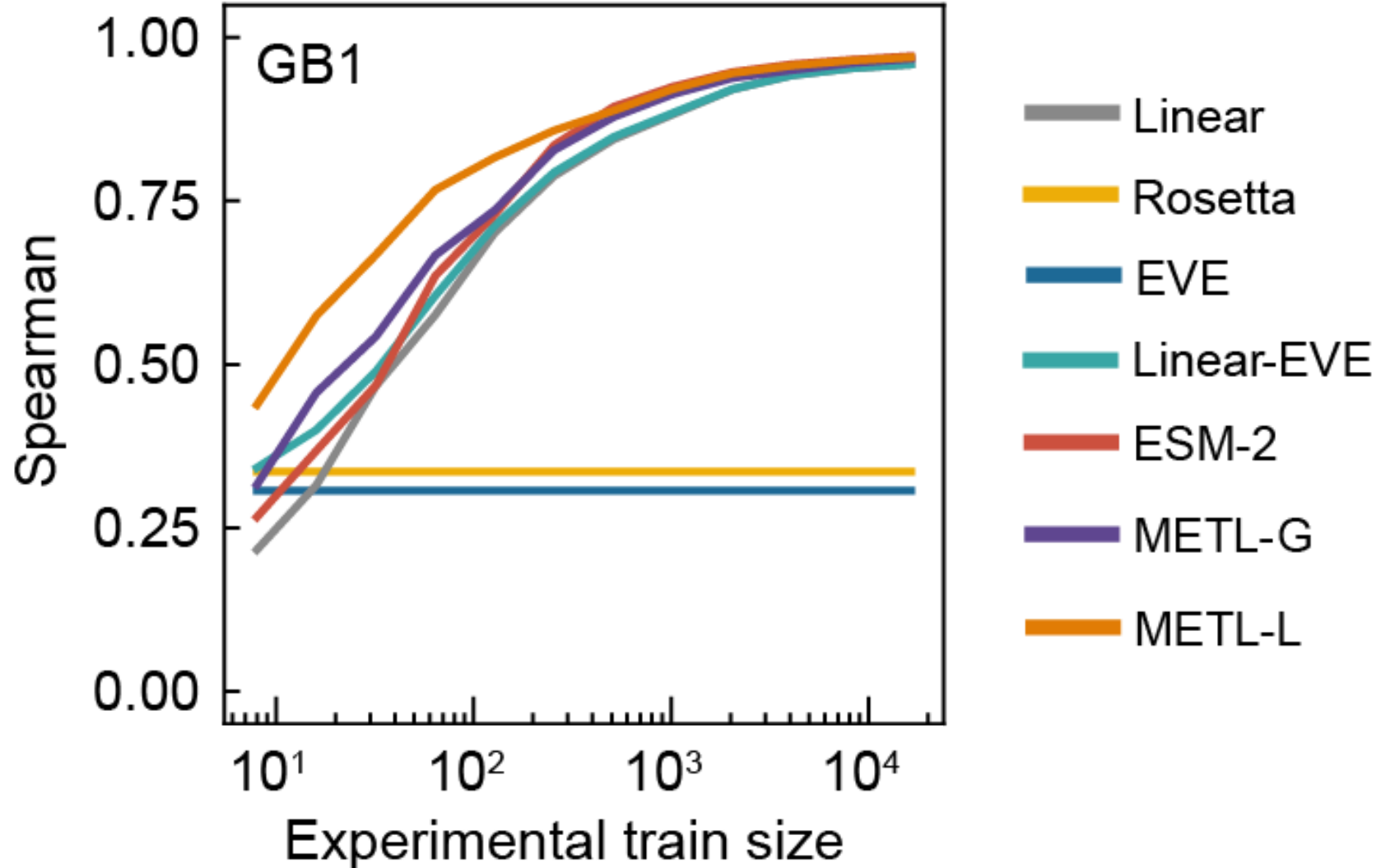
Learn evolutionary information

Useful for predicting:

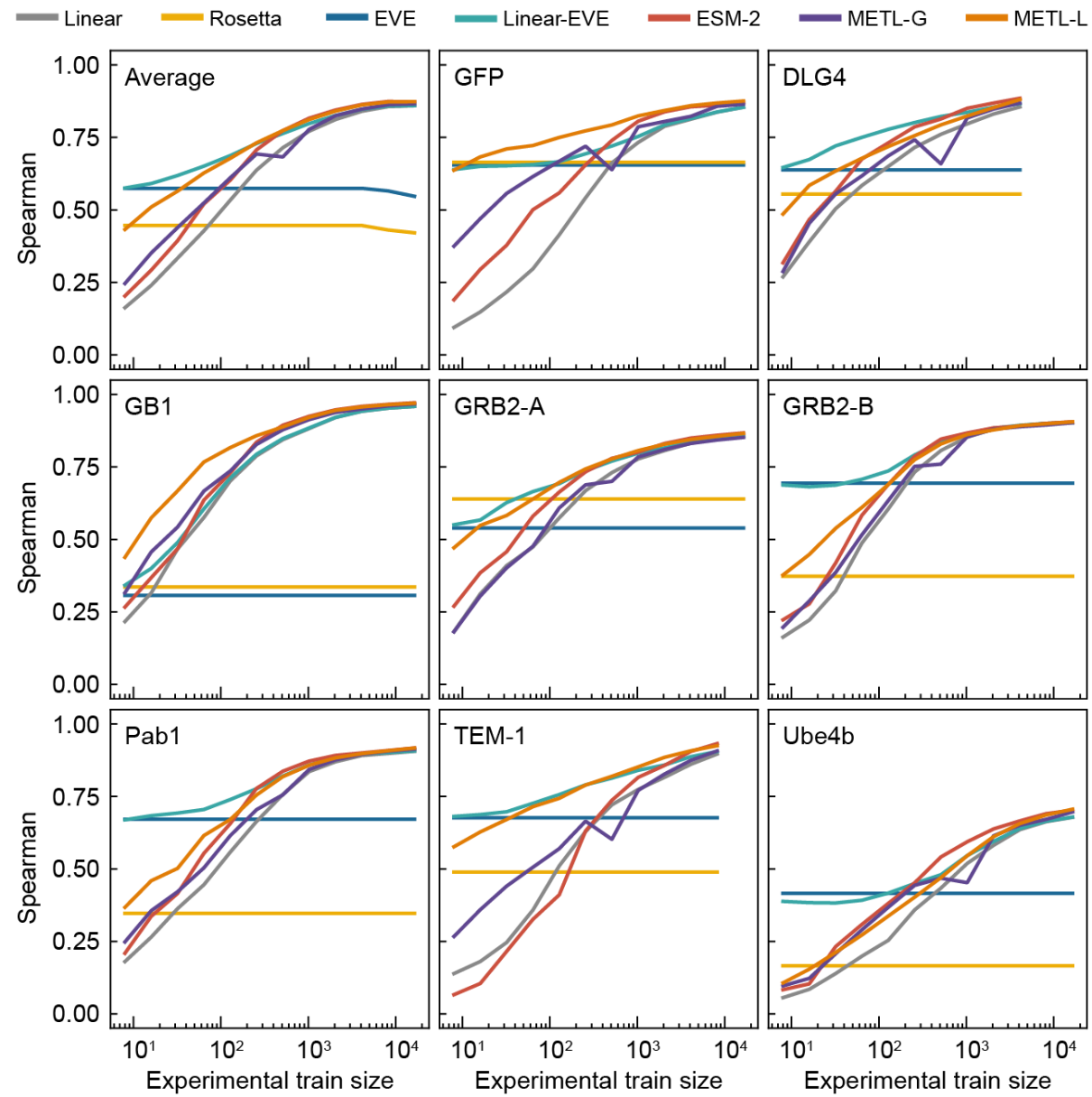
- amino acid properties
- protein function
- protein structure
- protein interactions

“Transformer-based deep learning for predicting protein properties in the life sciences”  
Chandra *et al.* *eLife* 2023 <https://doi.org/10.7554/eLife.82819>

# Pretraining on biophysical simulations can improve function prediction from limited data



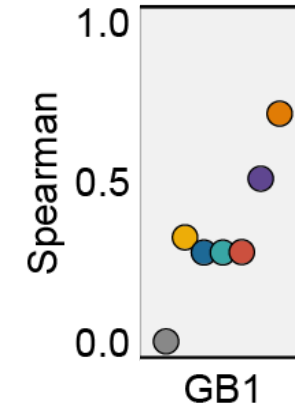
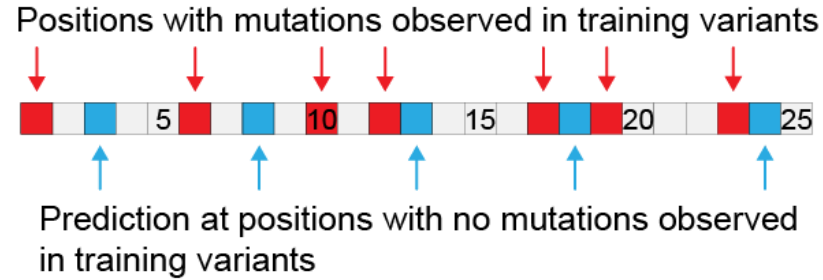
# Prediction performance is protein function dependent





# Biophysical pretraining improves protein fitness prediction in challenging settings

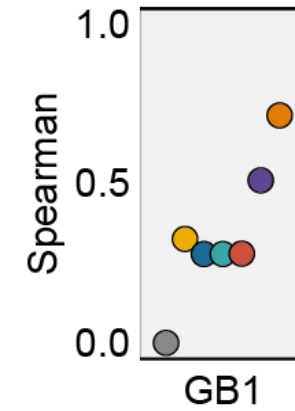
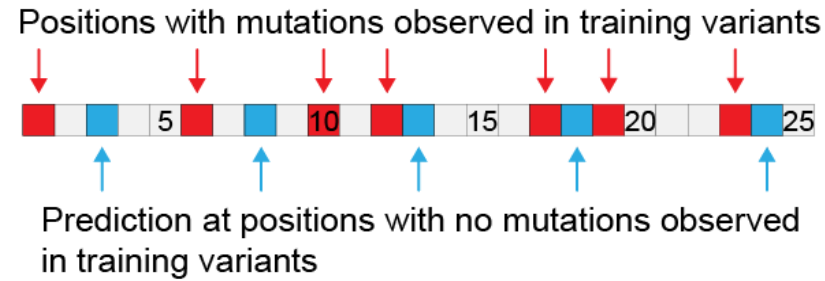
Position extrapolation: generalizing across sequence positions



Linear Rosetta EVE Linear-EVE ESM-2 METL-G METL-L

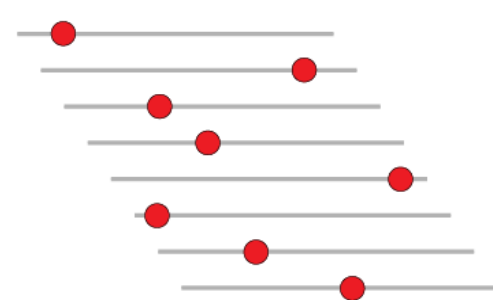
# Biophysical pretraining improves protein fitness prediction in challenging settings

## Position extrapolation: generalizing across sequence positions

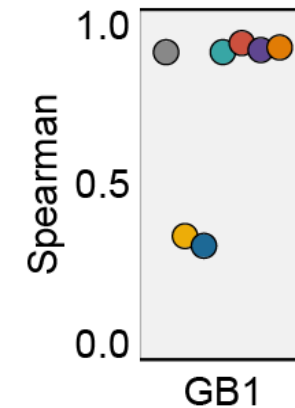
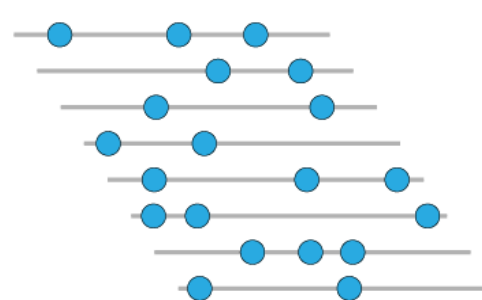


## Regime extrapolation: predicting how mutations combine

Train on single mutants

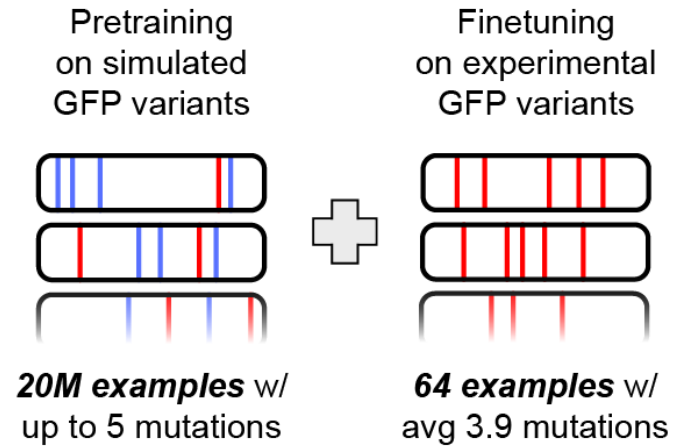
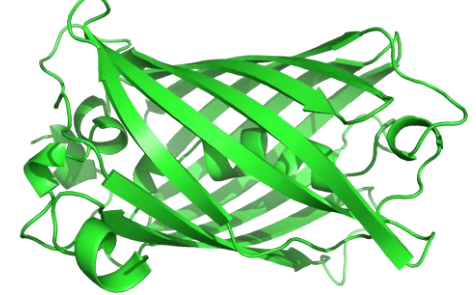


Test on multi-mutants



Linear Rosetta EVE Linear-EVE ESM-2 METL-G METL-L

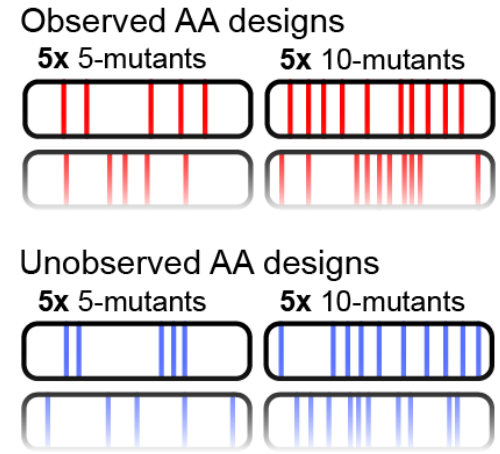
# METL for GFP engineering



Pretraining  
→  
Finetuning

METL-Local model for GFP brightness

Simulated annealing optimization  
→



Mutations observed in experimental training variants

Mutations *not* observed in experimental training variants

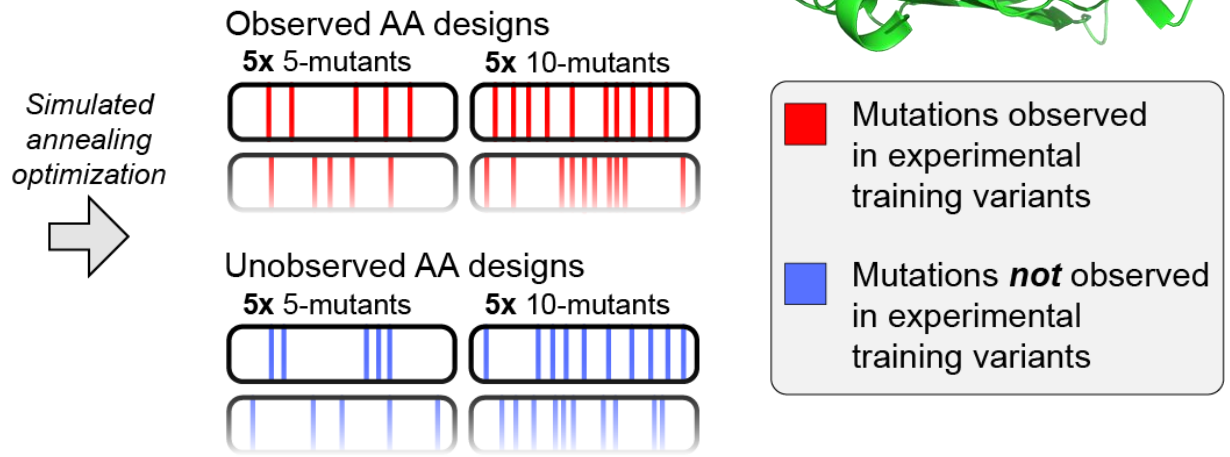
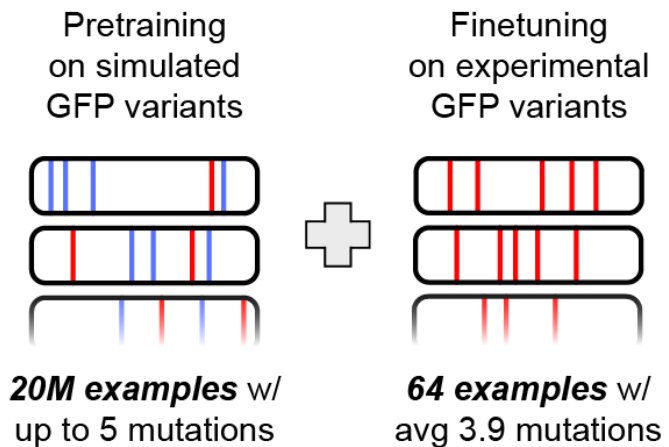
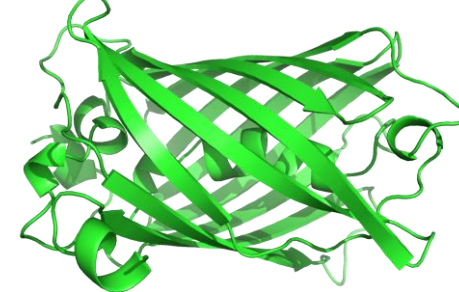


Chase Freschlin



Phil Romero

# METL for GFP engineering

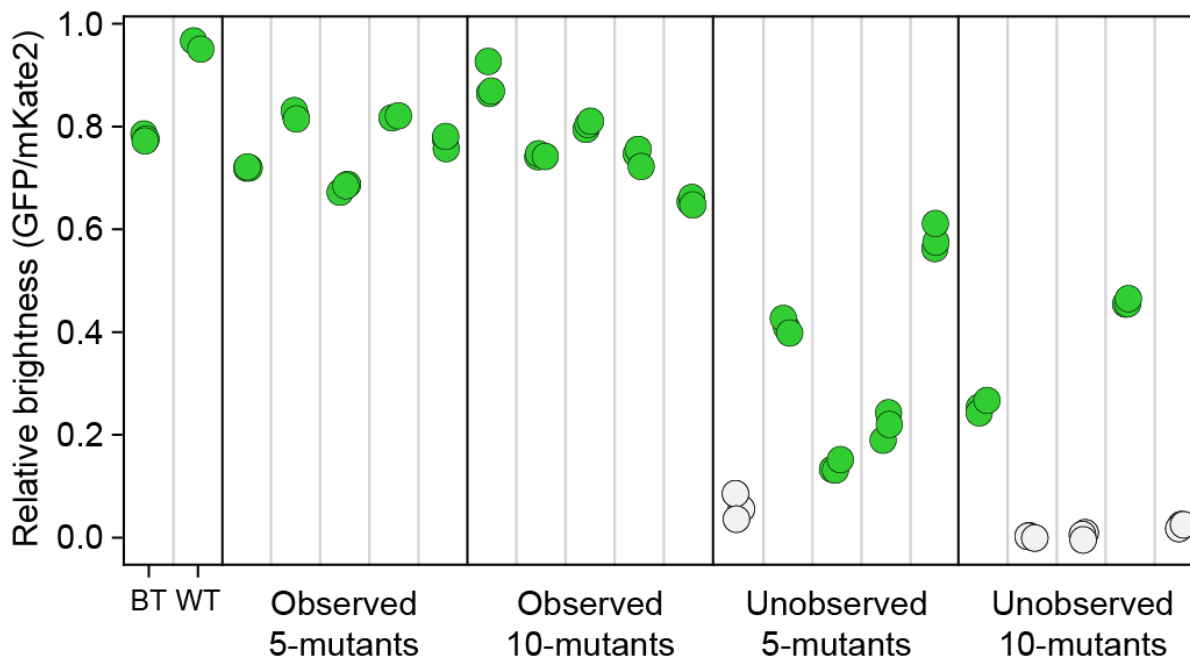


Chase Freschlin



Phil Romero

16 of 20 designs are functional in this challenging setting





# METL conclusions and questions





Simulations can help overcome experimental data scarcity

METL can guide wet lab protein design

What protein functions are more compatible with evolutionary versus biophysical modeling?

How can we better customize biophysical simulations?

**Biophysics-based protein language models for protein engineering**

 Sam Gelman, Bryce Johnson,  Chase Freschlin, Sameer D'Costa,  Anthony Gitter,  
 Philip A. Romero

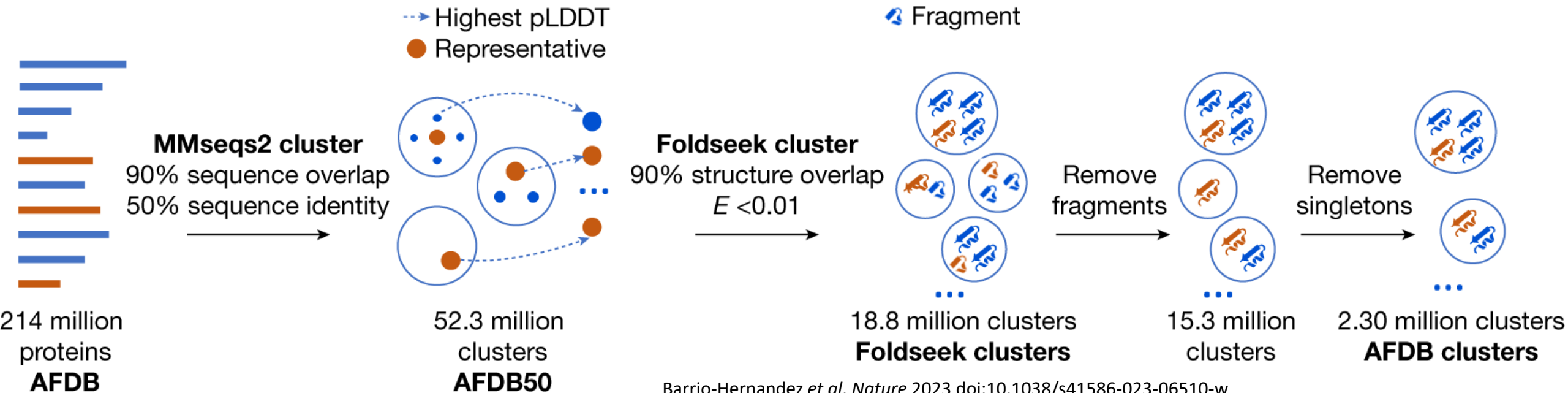
<https://doi.org/10.1101/2024.03.15.585128>

<https://github.com/gitter-lab/metl>

# METL future directions

Can we do active learning with the biophysical simulations?

Guide simulator as the model trains based on where it has poor generalization performance



Should we be doing this?

# Lessons from cinema: July 21, 2023

**Forbes**

**The 3 'Godfathers' Of AI  
Have Won The Prestigious  
\$1M Turing Prize**

“We shouldn’t be too surprised when the most powerful stereotype machine ever constructed spits out stereotypes”

- Michael Baym in a recent discussion about modern AI



# Lessons from cinema: July 21, 2023



# Benefits and harms of AI in synthetic biology

Community is most concerned with viruses and toxic agents

Risks of software and models versus the wet lab techniques that deploy them

AI researchers should not be the only experts making these decisions but must accept responsibility for their work

# Acknowledgements

## METL protein engineering

Sam Gelman  
Bryce Johnson  
Sameer D'Costa  
Chase Freschlin  
Phil Romero

## Lab members and close collaborators

Sam Gelman	Ryan Kassab	Justin Hiemstra
Bryce Johnson	Arnav Sharma	Chase Freschlin
Daniel McNeela	John Peters	Meg Taylor
Yifan Deng	Carol Sze	Amy Freitag
Neha Talluri	Daniel Nachreiner	
Nistha Panda	Sumedha Sanjeev	

## Funding and resources

NIH R01 GM135631  
NIH Commons Credits  
NSF 2226451  
UW-Madison Center for High-Throughput Computing  
OSG Consortium

Morgridge Institute for Research  
Jeanne M. Rowe Chair  
NVIDIA GPU hardware award  
Icons by Freepik (flaticon.com)