# Learning protein sequence-function relationships from high-throughput molecular simulations

Sam Gelman · May 25, 2022

sgelman2@wisc.edu

# HTCondor, two ways

## Molecular simulations - CPUs

Rosetta Commons

RCSB PDB PROTEIN DATA BANK

Open Science Grid

intel inside

## Machine learning - GPUs

PyTorch

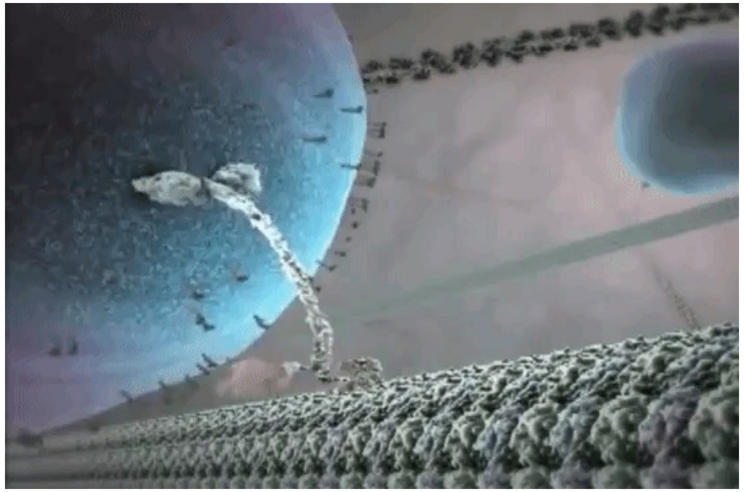
W&B

NVIDIA CUDA

PyTorch Lightning

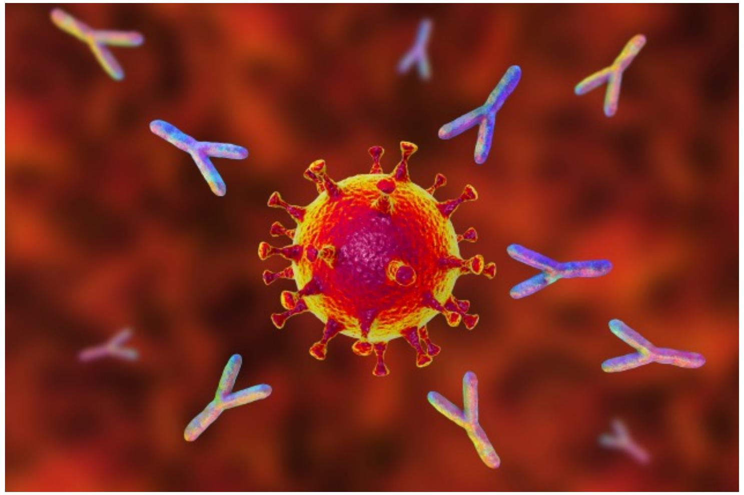
# Proteins

## Functional biomolecules composed of amino acids



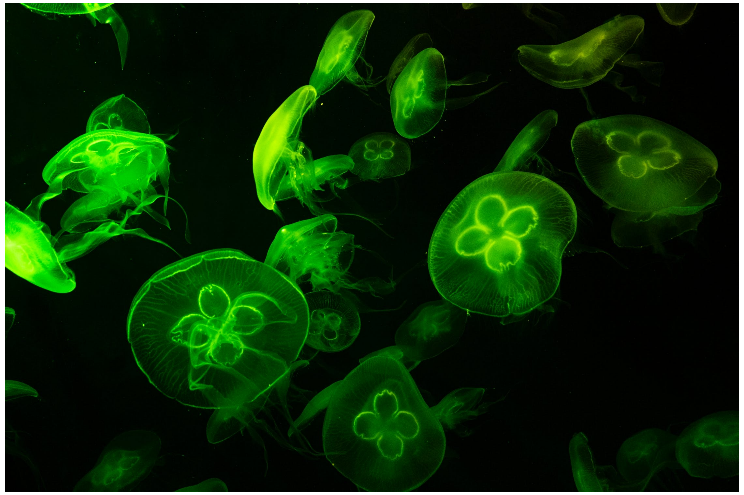
Kinesin

Transport protein



Antibodies

And a suspicious spike protein…



Green fluorescent protein

In jellyfish

# Protein design

Modify proteins to have desired function

Industry

Example: laundry detergents

Medicine

Example: antibody treatments

Agriculture

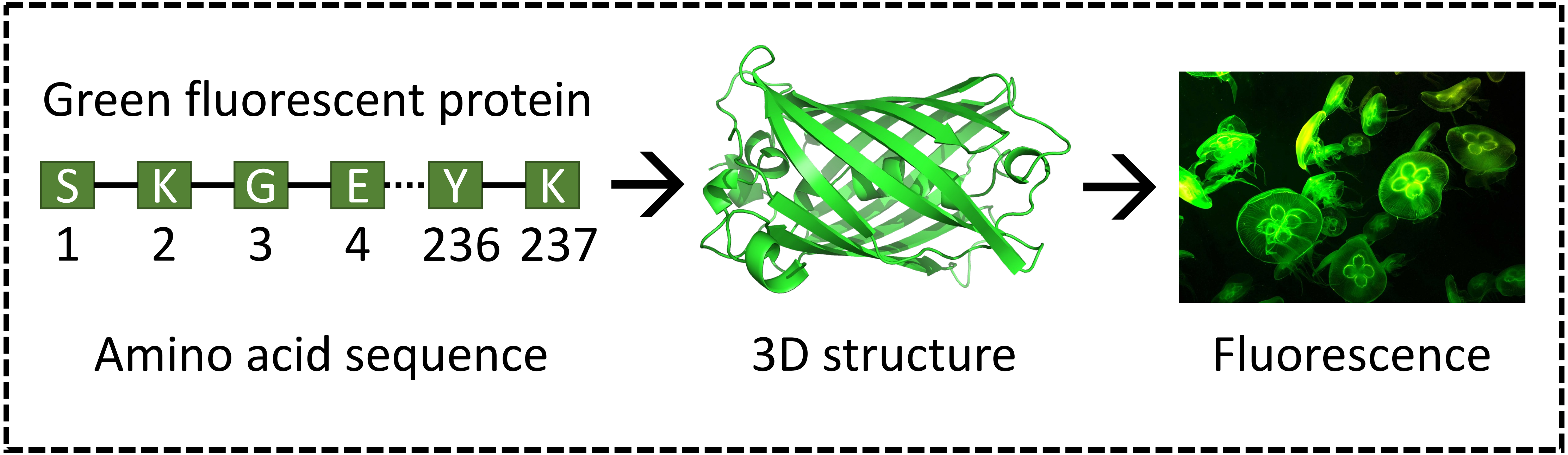
Example: herbicide resistance

Scientific Research

Example: marker proteins

# Sequence-function relationship

Amino acid sequence → 3D structure → function



Green fluorescent protein

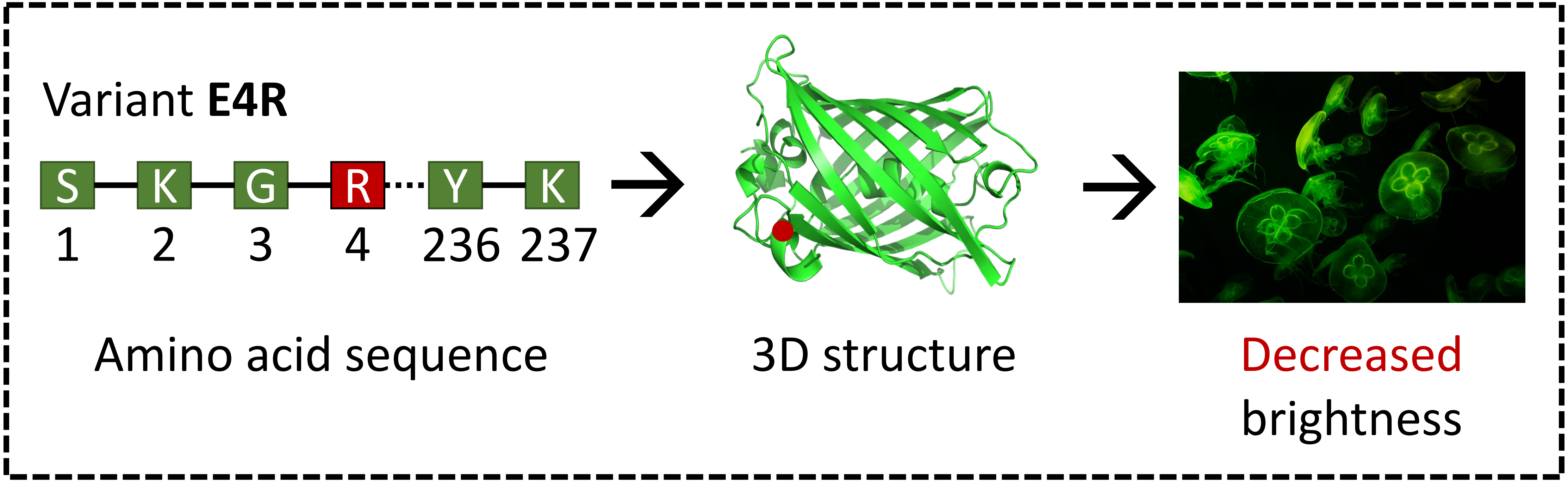
Amino acid sequence

3D structure

Fluorescence

# Sequence-function relationship

Amino acid sequence → 3D structure → function



Variant **E4R**

S K G R … Y K

1 2 3 4 236 237

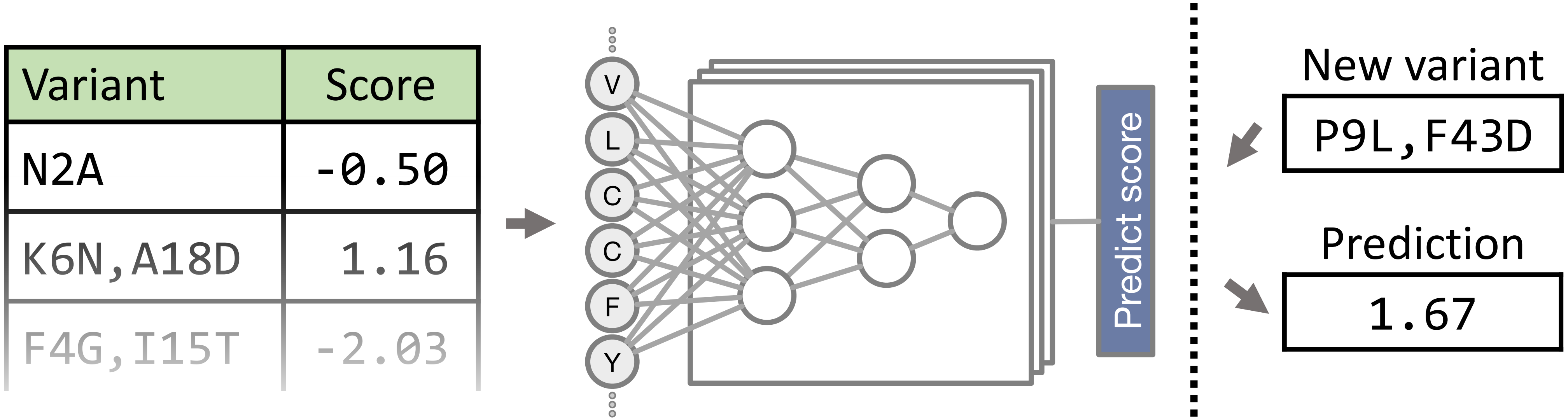
Amino acid sequence

3D structure

Decreased brightness

# Objective

Predict functional activity of protein variants



| Variant | Score |
|---|---|
| N2A | -0.50 |
| K6N,A18D | 1.16 |
| F4G,I15T | -2.03 |

Experimental data

Neural network

New variant

P9L,F43D

Prediction

1.67

Predict scores for new variants

# METL (Mutational Effect Transfer Learning)

## Transfer learning from molecular simulations

| Variant | Score |
|---|---|
| N2A | -0.50 |
| K6N,A18D | 1.16 |
| F4G,I15T | -2.03 |

Experimental target data

+

RCSB PDB PROTEIN DATA BANK

Rosetta Commons

Simulated source data

- ✓ Learn from fewer examples
- ✓ Extrapolate outside of training data

Better performance!

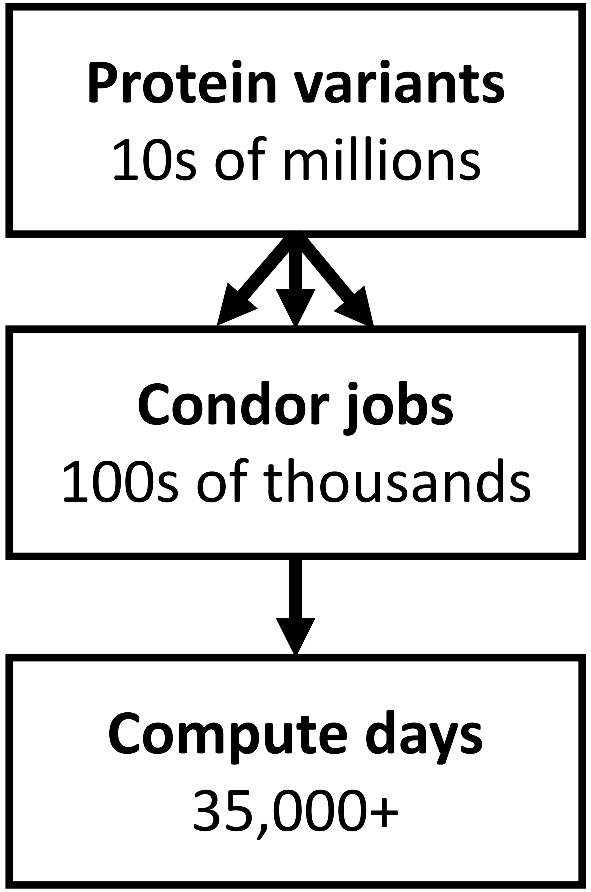
# METL (Mutational Effect Transfer Learning)

Run molecular simulations
→ **High-throughput CPU**

Train neural nets on molecular simulations
→ **Long-running GPU**

Transfer and finetune models on experimental data
→ **High-throughput GPU**

# Running molecular simulations



**Protein variants**
10s of millions

**Condor jobs**
100s of thousands

**Compute days**
35,000+

**Why HTC**

- Variants run independently

**Strategies**

- 5-10 hours per job
- Auto-retry & auto-release

**What went well**

- Capacity (especially with OSG!)

**Challenges**

- Bad servers or sites

**Mitigation**

- Block servers
- *on_exit_hold* and auto-release

Rosetta Commons

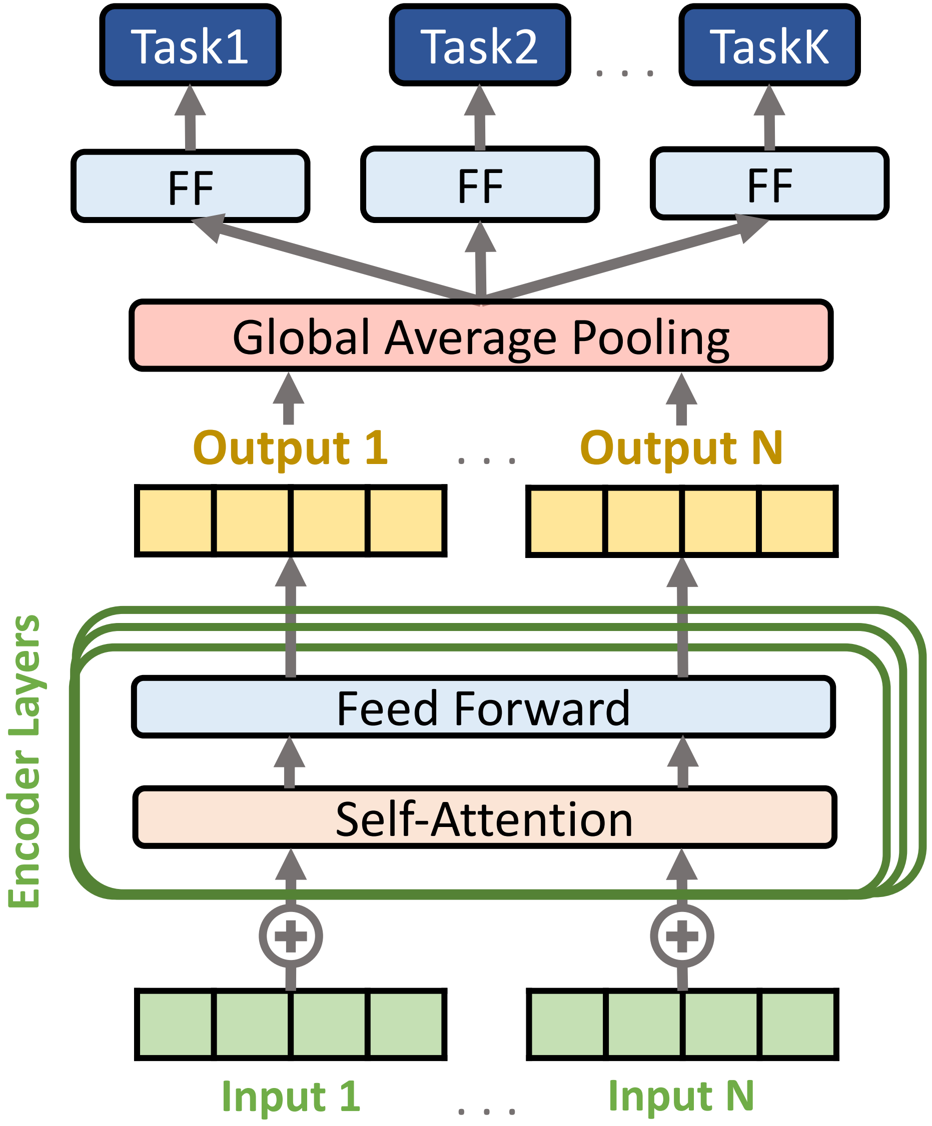
Open Science Grid

# Training neural nets

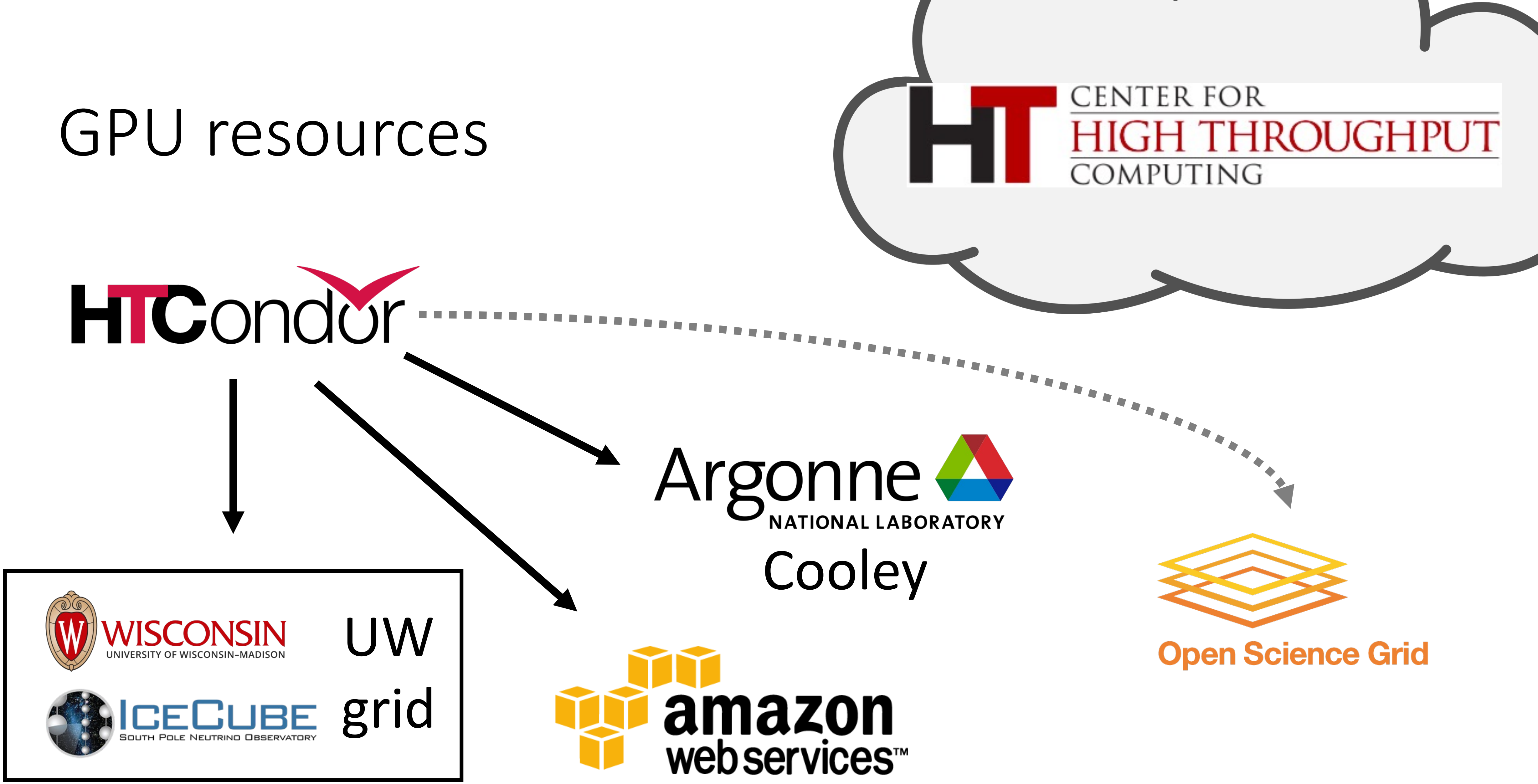
Neural networks need GPUs and can take a long time to train!

Let’s talk about:

- Getting GPU resources
- Checkpointing
- Logging



# GPU resources



Cooley

UW

grid

# Checkpointing

**Required** for long-running models

PyTorch Lightning

Save and restore model checkpoints

```
ModelCheckpoint(every_n_epochs=1)
trainer.fit(ckpt_path=ckpt_path)
```

Integrate with HTCondor checkpointing

```
class CondorStopping(EarlyStopping)
CondorStopping(every_n_epochs=1)
```

# Logging

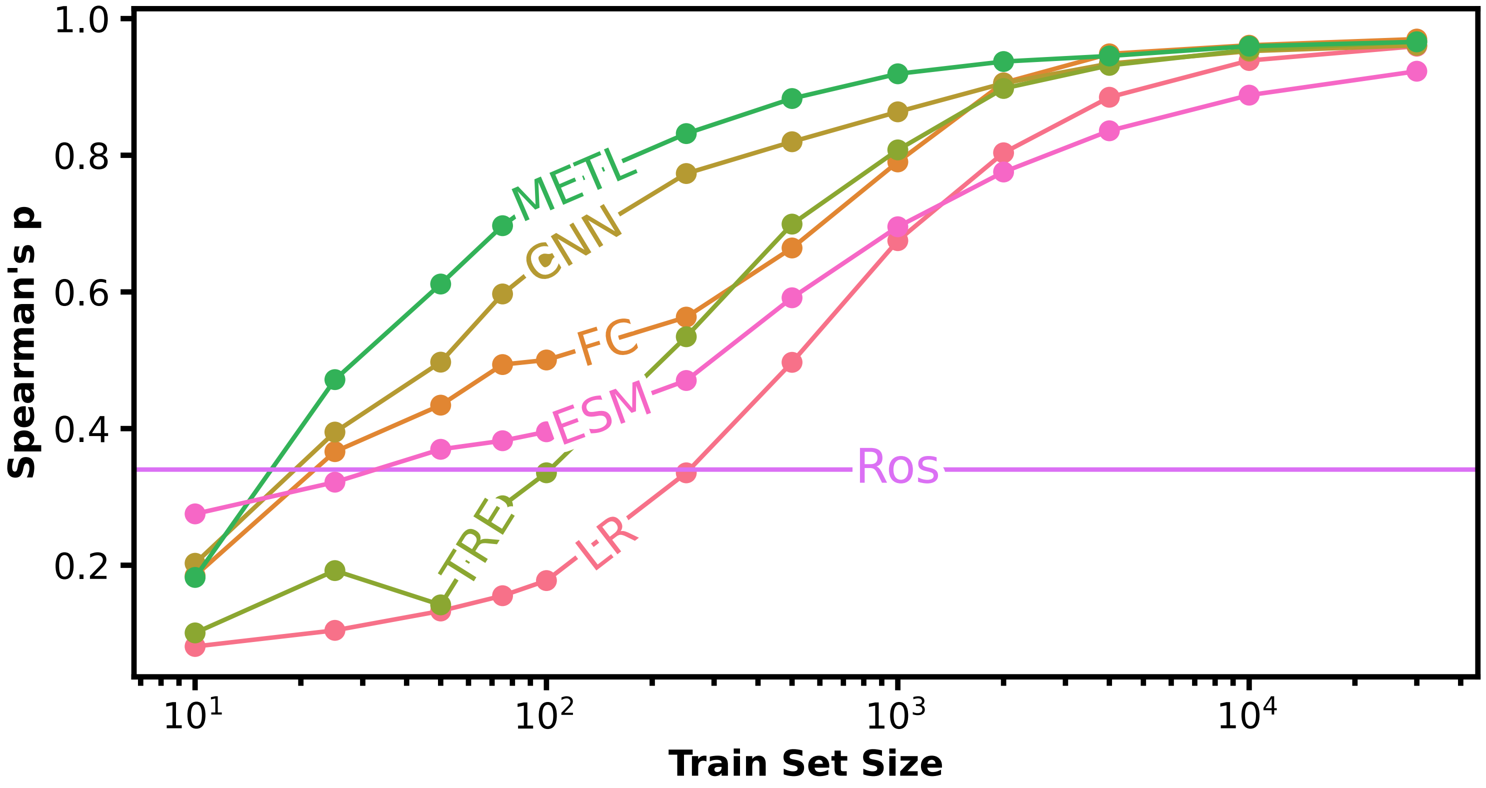
Understand training progress and results

## Weights & Biases

- Stream progress to online dashboard
- Track metrics and system utilization
- Manage hundreds of models

# Results

Results shown for GB1 dataset



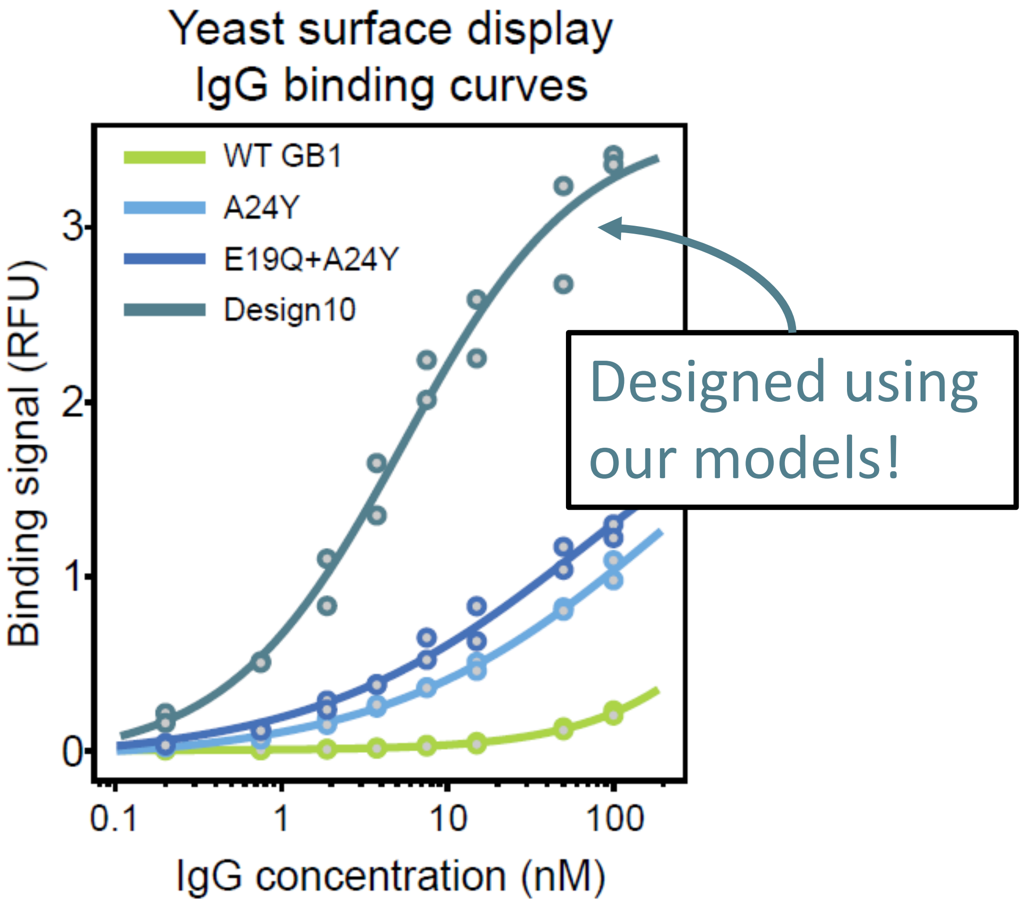
LR: linear regression · FC: fully connected · CNN: sequence convolutional · TRE: transformer encoder · METL: our approach · ESM: evolutionary scale modeling · Ros: Rosetta's total_score

# Check out our publication (previous work)

***Neural networks to learn protein sequence-function relationships from deep mutational scanning data***

Sam Gelman, Sarah A Fahlberg, Pete Heinzelman, Philip A Romero$^+$, Anthony Gitter$^+$

*Proceedings of the National Academy of Sciences*, 118:48, 2021



Yeast surface display IgG binding curves

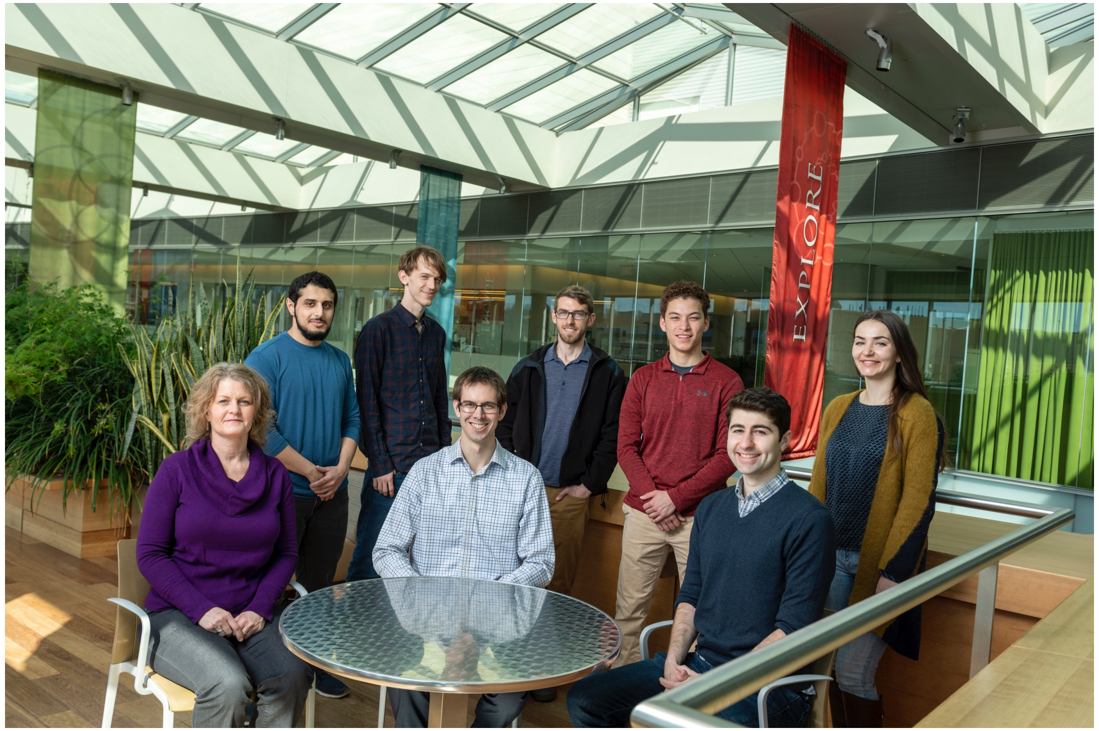
Designed using our models!

# Conclusion

Thanks to HTCondor, the Center for High-Throughput Computing, and Open Science Grid for making this research possible!

# Acknowledgements

- Morgridge Institute for Research
- Phil Romero and Romero Lab
- PhRMA Foundation
- Center for High Throughput Computing and Cooley
- NHGRI training grant to the Genomic Sciences Training Program T32 HG002760
- National Institutes of Health (NIH) Cloud Credits Model Pilot, a component of the NIH Big Data to Knowledge (BD2K) program
- GPU hardware from NVIDIA



The Gitter Lab (early 2020)

# Thank you

# Questions?

Feel free to reach out!
sgelman2@wisc.edu