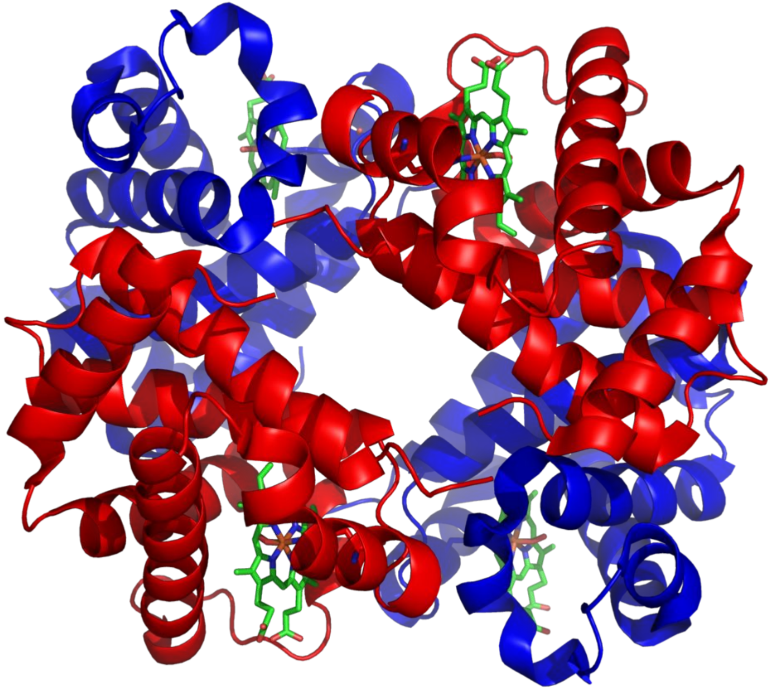
Evaluating Protein Transfer Learning with TAPE

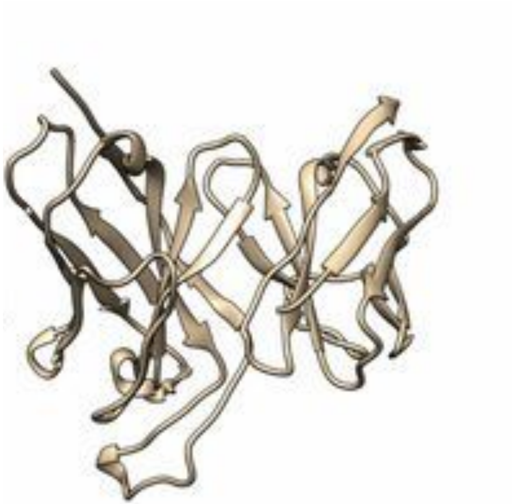
Roshan Rao*, Nicholas Bhattacharya*, Neil Thomas*,
Yan Duan, Xi Chen,
John Canny, Pieter Abbeel, Yun S. Song

Why care about proteins?

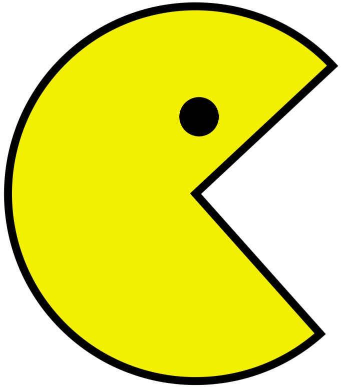
Hemoglobin

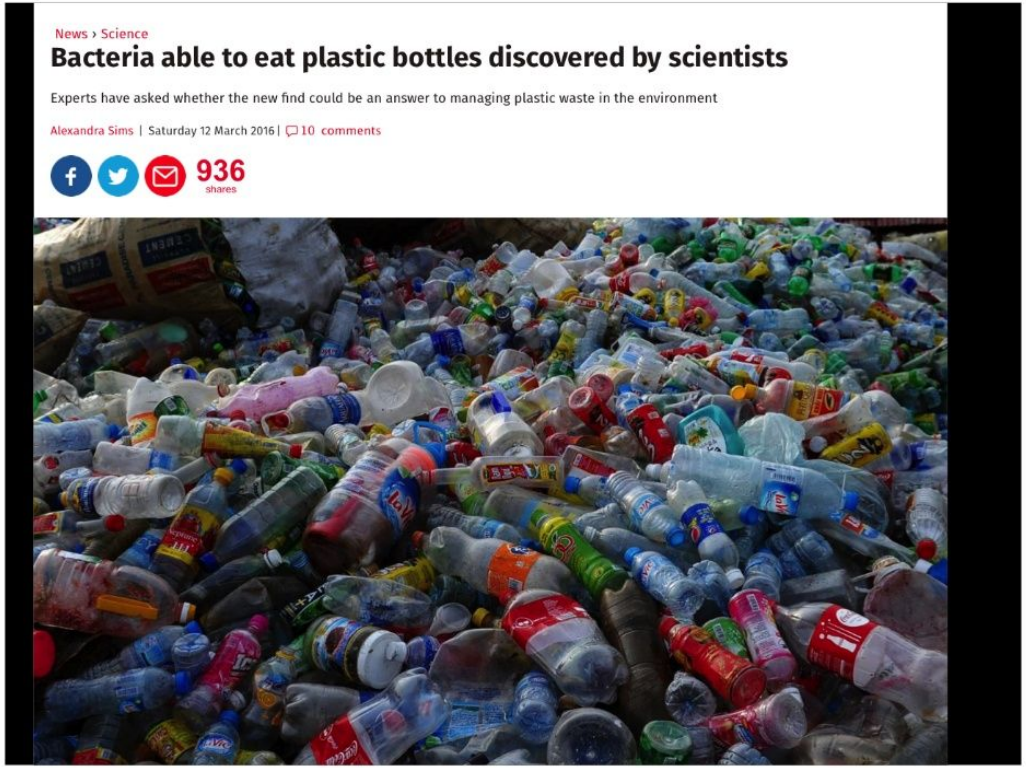


Antibodies



Tired of eating plastic?
Call 1-800-PROTEIN





News › Science

Bacteria able to eat plastic bottles discovered by scientists

Experts have asked whether the new find could be an answer to managing plastic waste in the environment

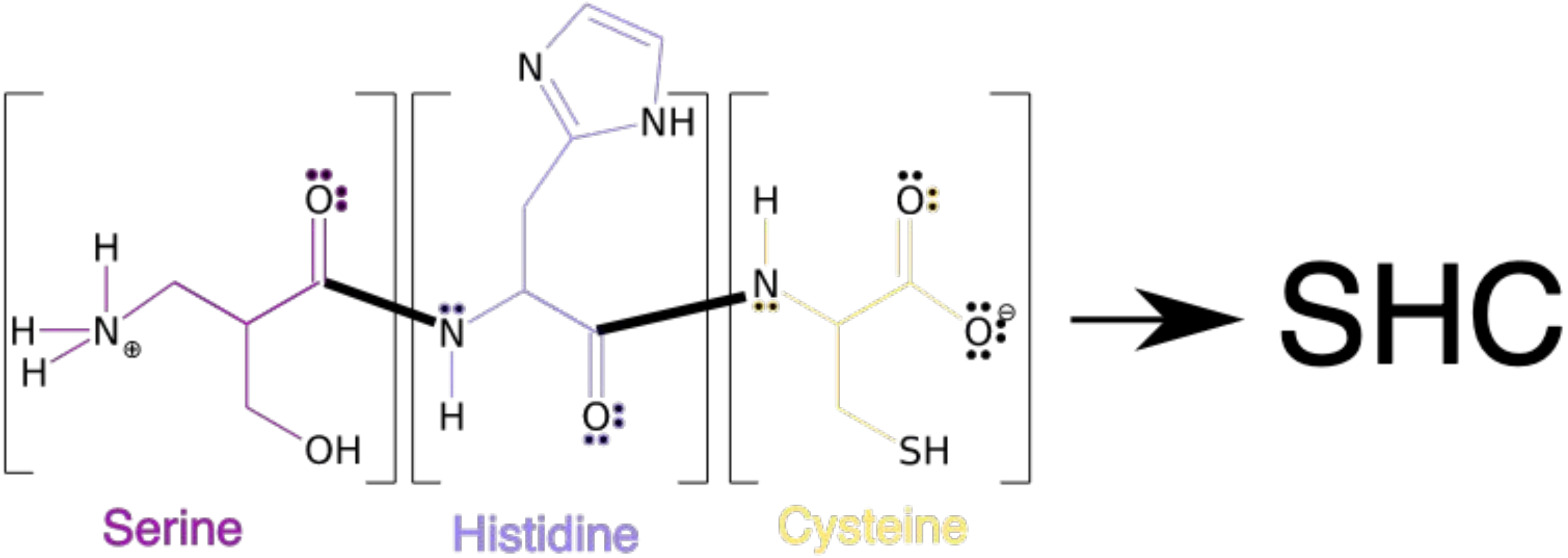
Alexandra Sims | Saturday 12 March 2016 | 10 comments

936 shares

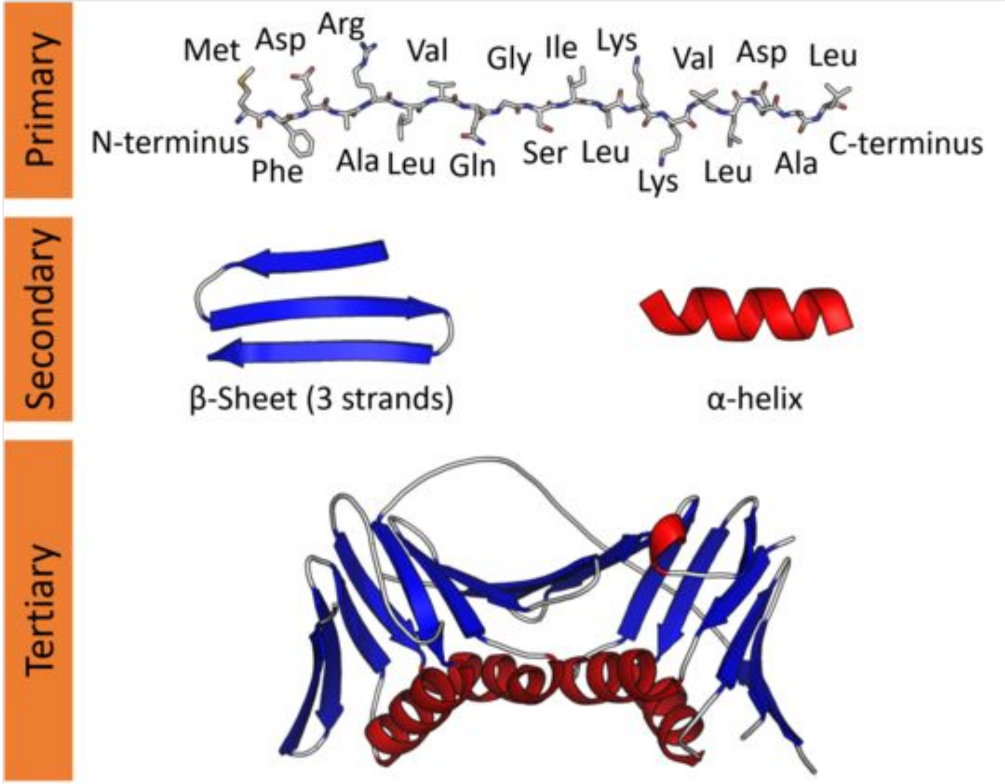
What is a protein?

A 3 slide crash course

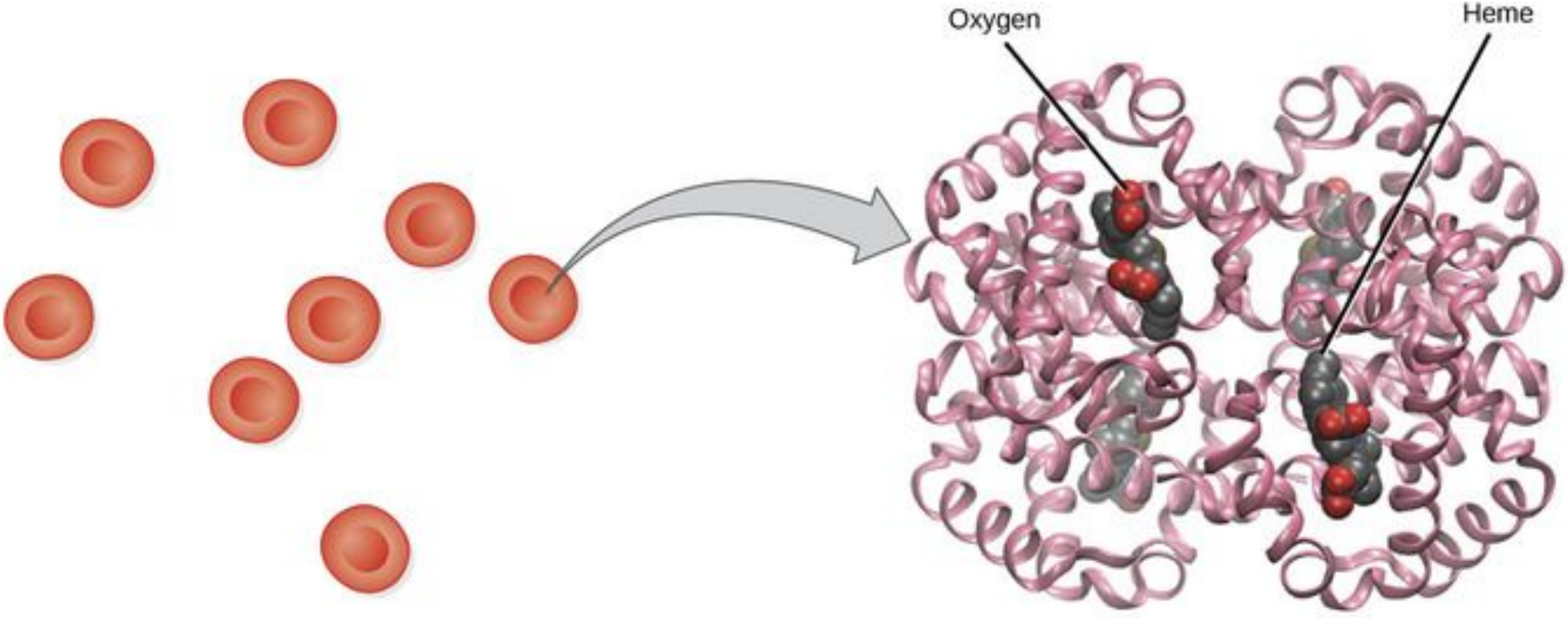
Sequence



Structure


Primary
Met
Asp
Arg
Val
Gly
Ile
Lys
Val
Asp
Leu
N-terminus
Phe
Ala
Leu
Gln
Ser
Leu
Lys
Leu
Ala
C-terminus
Secondary
β-Sheet (3 strands)
α-helix
Tertiary

Function



(a) Red blood cells

(b) Hemoglobin

How do we find new sequences?

Collecting unlabeled data

easy

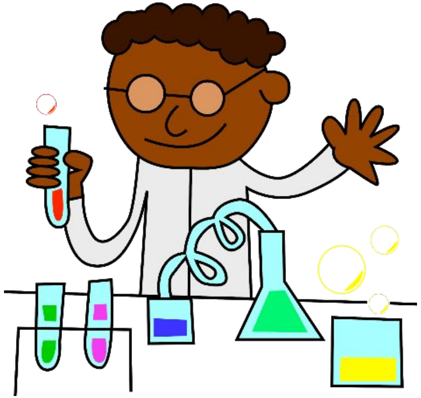
1. Put on protective equipment

2. Collect dirt

3. Throw it in the sequencer

4. Lots of Sequences (Genes)

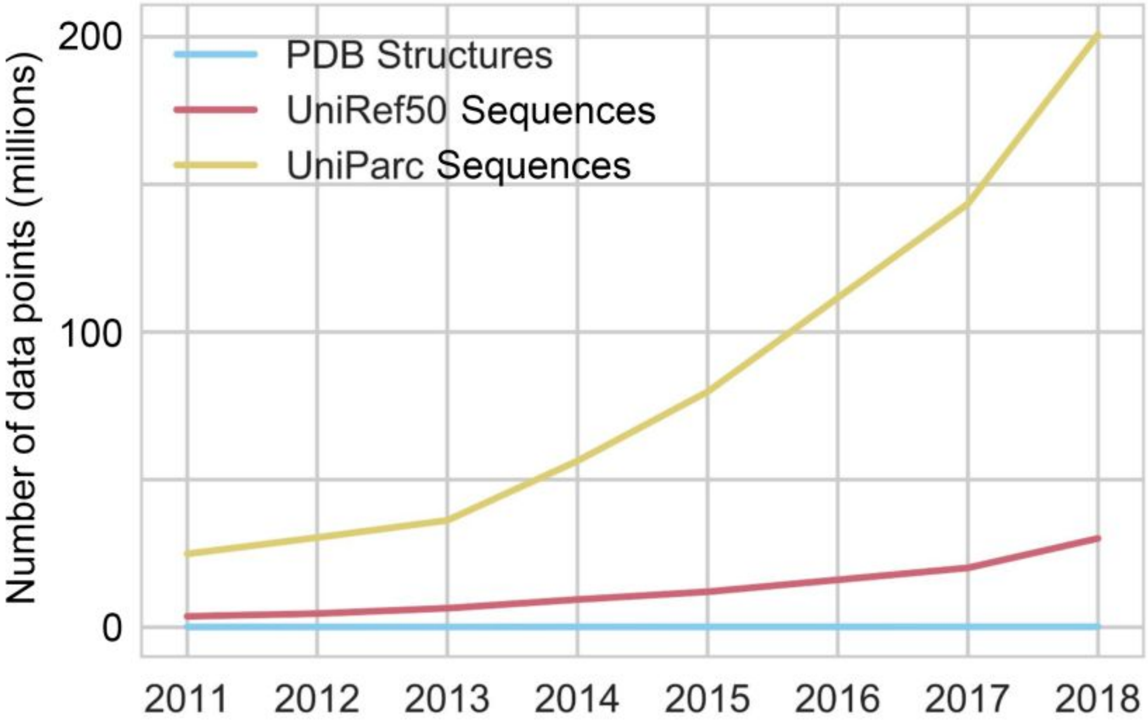
How do we determine new structures?



Collecting labeled data

hard

We cannot keep up with the sequence explosion



https://moalquraishi.wordpress.com/2019/04/01/the-future-of-protein-science-will-not-be-supervised/

200 Million sequences

150,000 structures

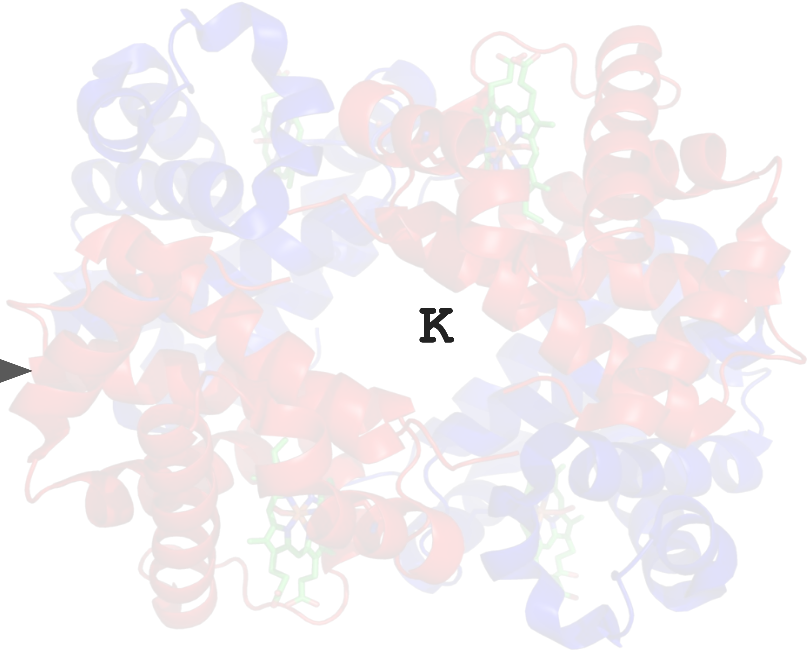
Can we take what we’ve learned from detailed experimental characterization and **generalize to unseen portions of sequence space**?

Pretrained models such as BERT make efficient use of labeled data

“Language Modeling” for Proteins

MVLSPADKTNV**[?]**AAWGKVG
AHAGEYGAEALE...

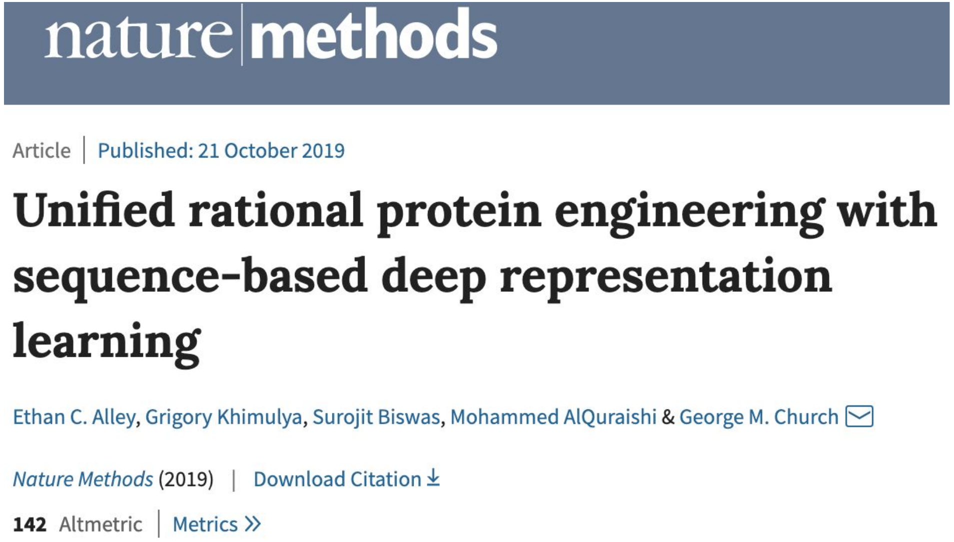
?



K

Which one is better?

- **Different downstream tasks**
- **Different pretraining corpuses**
  - 20 million sequences
  - 200 million sequences
- **Different compute budgets**
  - 1 GPU
  - 128 GPUs



nature methods

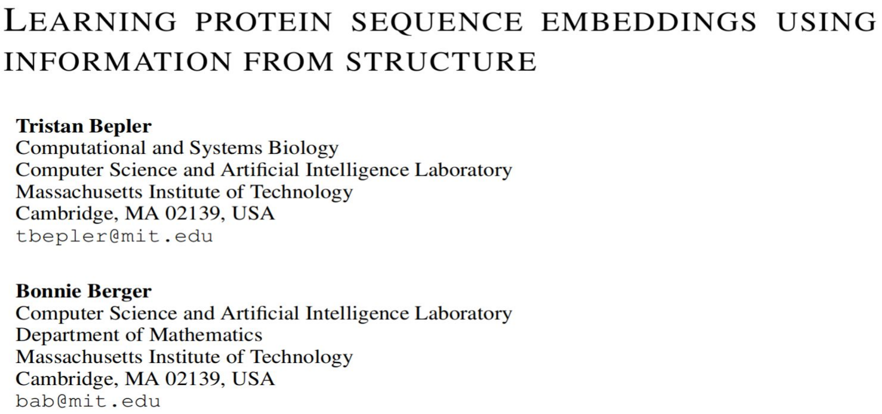
Article | Published: 21 October 2019

Unified rational protein engineering with sequence-based deep representation learning

Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi & George M. Church

Nature Methods (2019) | Download Citation

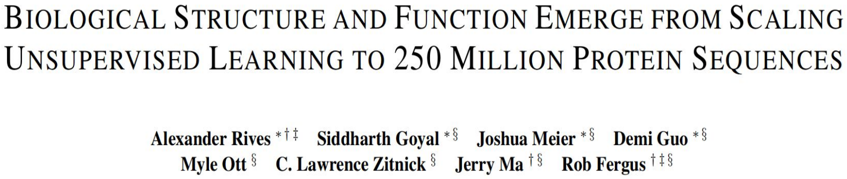
142 Altmetric | Metrics



LEARNING PROTEIN SEQUENCE EMBEDDINGS USING INFORMATION FROM STRUCTURE

Tristan Bepler
Computational and Systems Biology
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
tbepler@mit.edu

Bonnie Berger
Computer Science and Artificial Intelligence Laboratory
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
bab@mit.edu



BIOLOGICAL STRUCTURE AND FUNCTION EMERGE FROM SCALING UNSUPERVISED LEARNING TO 250 MILLION PROTEIN SEQUENCES

Alexander Rives [*†‡] **Siddharth Goyal** [*§] **Joshua Meier** [*§] **Demi Guo** [*§]
Myle Ott [§] **C. Lawrence Zitnick** [§] **Jerry Ma** [†§] **Rob Fergus** [†‡§]

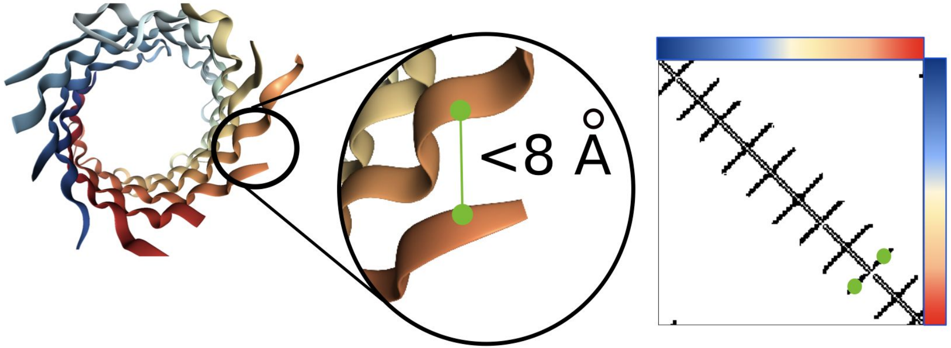
TAPE

Tasks **A**ssessing **P**rotein **E**mbeddings

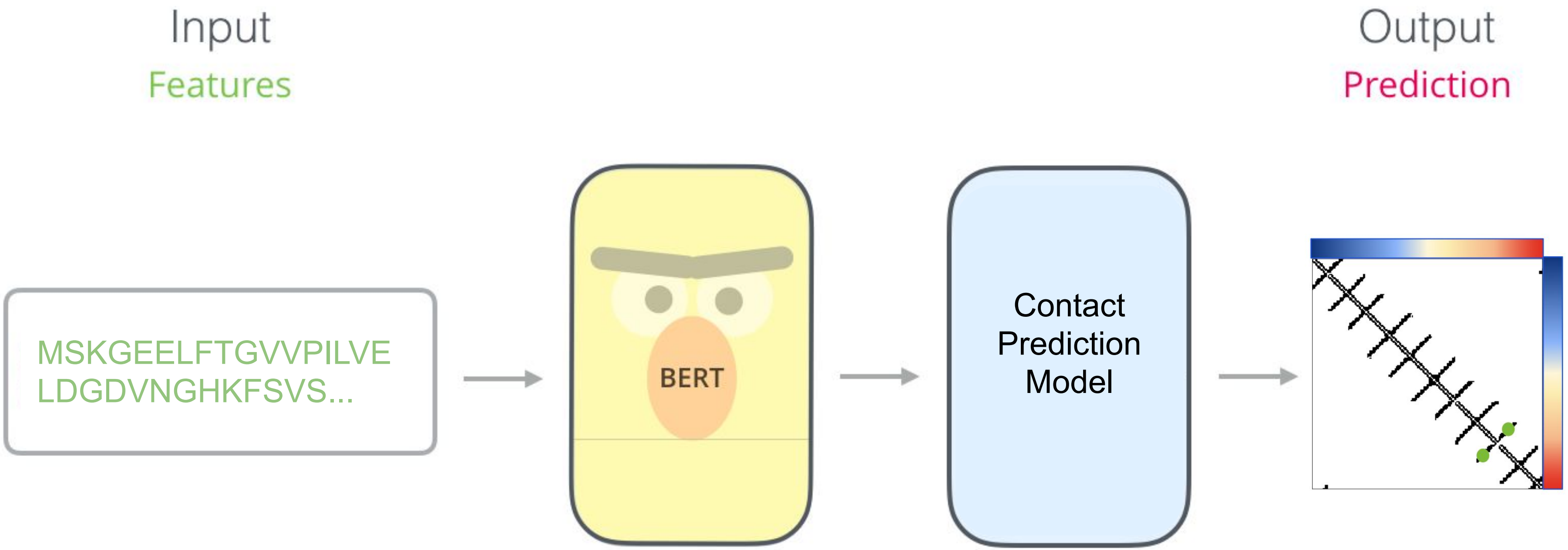
- **Fixed 5 downstream tasks** from different domains of protein biology testing meaningful generalization
- Pretrained 5 different models with:
  - **Fixed corpus**: 30 million sequences of protein domains
  - **Fixed budget**: 1 week on 4 NVIDIA V100s

TAPE

Contact Prediction



TAPE



Input

Features

MSKGEELFTGVVPILVE
LDGDVNGHKFSVS...

BERT

Contact
Prediction
Model

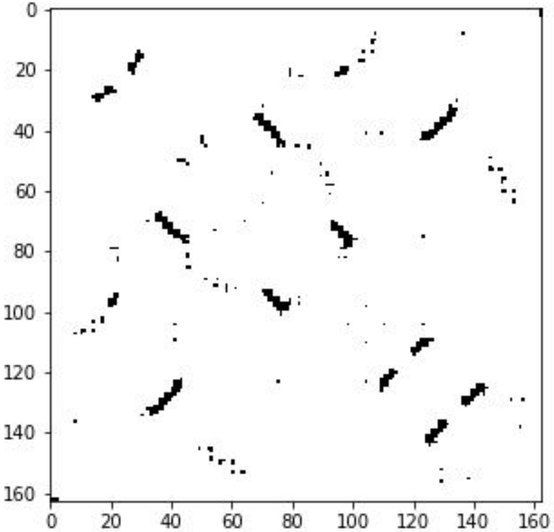
Output

Prediction

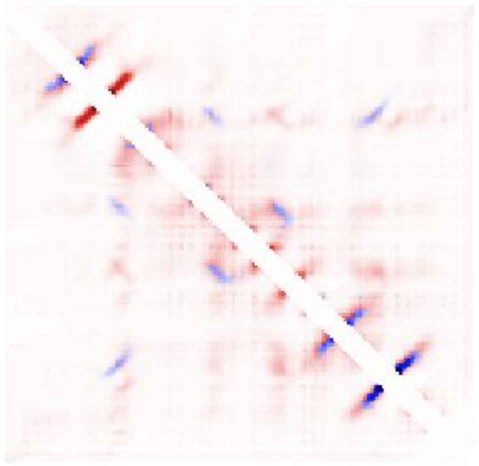
Pretraining Helps

🟦 True positive

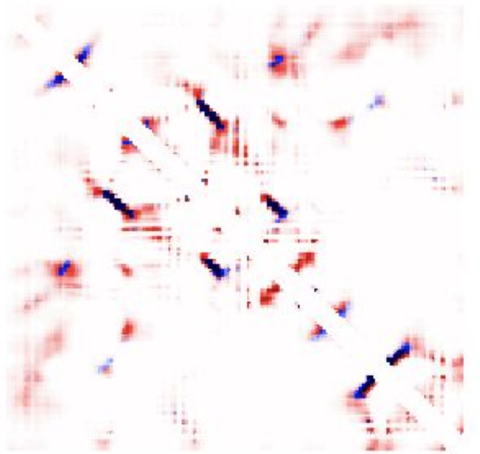
🟥 False positive



Ground Truth



No Pretraining (LSTM)

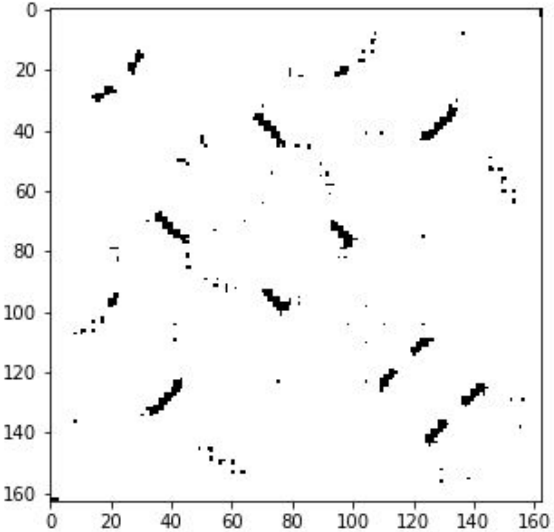


Pretrained (LSTM)

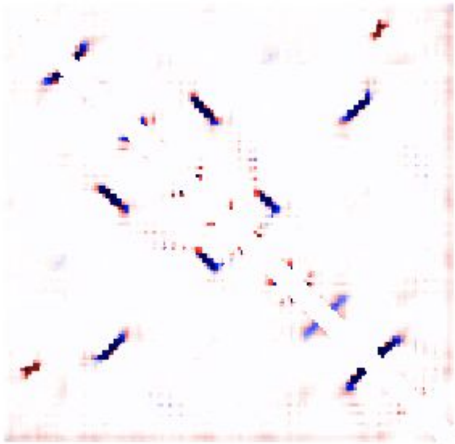
Unused signal remains!

🟦 True positive

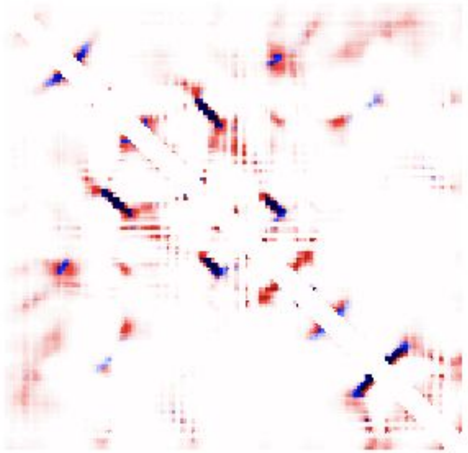
🟥 False positive



Ground Truth



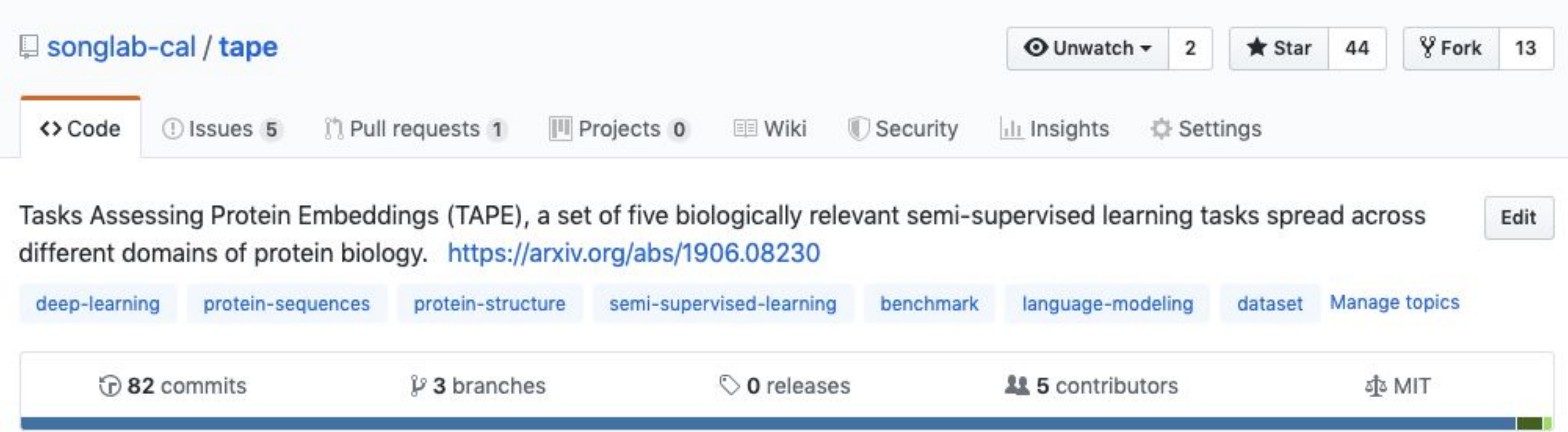
Non-neural features



Pretrained (LSTM)

Data/code for benchmark available

https://github.com/songlab-cal/tape



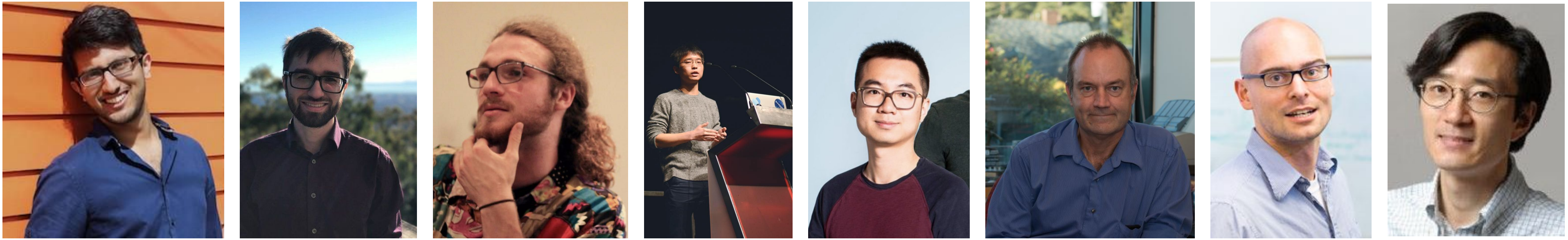
songlab-cal / tape

Tasks Assessing Protein Embeddings (TAPE), a set of five biologically relevant semi-supervised learning tasks spread across different domains of protein biology. https://arxiv.org/abs/1906.08230

deep-learning protein-sequences protein-structure semi-supervised-learning benchmark language-modeling dataset Manage topics

82 commits 3 branches 0 releases 5 contributors MIT

Come chat more at our poster! (#79)



Roshan Rao, Nick Bhattacharya, Neil Thomas, Rocky Duan, Peter Chen, John Canny, Pieter Abbeel, Yun S. Song

CHAN ZUCKERBERG BIOHUB

aws educate

Berkeley DeepDrive

NIH
National Institutes of Health

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Open Philanthropy Project