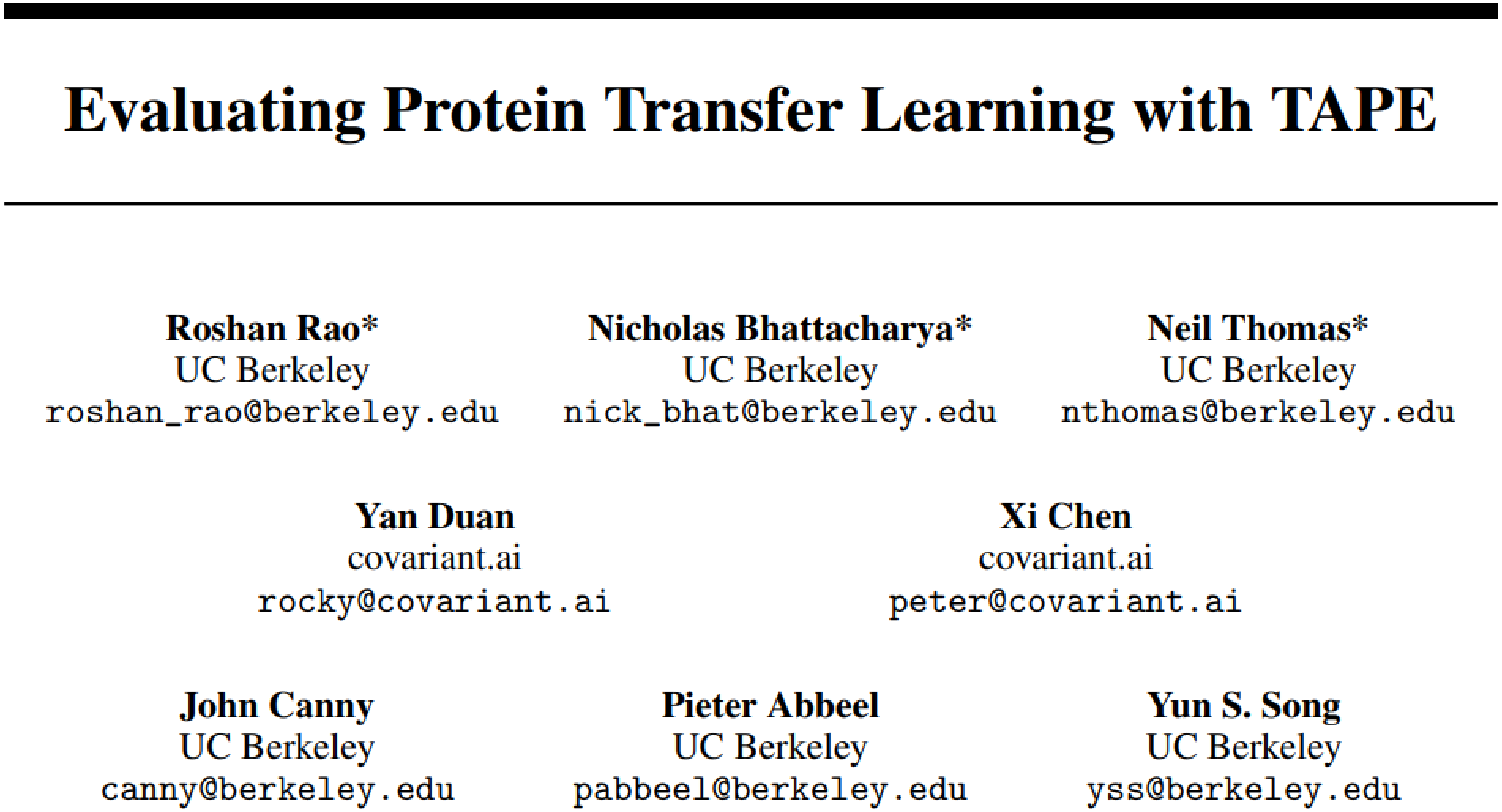


Evaluating Protein Transfer Learning with TAPE

Roshan Rao*
UC Berkeley
roshan_rao@berkeley.edu

Nicholas Bhattacharya*
UC Berkeley
nick_bhat@berkeley.edu

Neil Thomas*
UC Berkeley
nthomas@berkeley.edu

Yan Duan
covariant.ai
rocky@covariant.ai

Xi Chen
covariant.ai
peter@covariant.ai

John Canny
UC Berkeley
canny@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@berkeley.edu

Yun S. Song
UC Berkeley
yss@berkeley.edu

Presented by: Rahmatullah Roche

Motivation

- Database of protein sequence growing exponentially
- Total number of sequence doubling each year
- Unlabeled sequence contains significant evolutionary information
- Can NLP extract the information?

Contribution: benchmarking

- Self supervised learning from unlabeled dataset
- Task assessing protein embedding (TAPE)
- Systematically evaluated semi-supervised learning on protein sequences
- 5 biologically relevant supervised task
- Hypothesis: multiple tasks are required to accurately benchmark any method
- Performance assessment of
  - Recurrent-based model
  - Convolution-based model
  - Attention-based model
  - Semi-supervised models

Background

- Protein terminology
  - (x1, x2, x3, …., xL) fixed alphabet for amino acids
- Protein sequence alignments
  - Query → Database→ Alignment
- Semi-supervised learning
  - Leverage information from both labeled and unlabeled data

Related Works

- Kernel-based pretraining for homology detection
- NLP-based techniques for transfer learning
- VAE to predict functional impact in mutations
- Transfer learning in protein ss and contact prediction
- Not rigorously benchmarked to assess the comparisons

Dataset

- Unlabeled sequence dataset
  - Pfam database of 31M protein domains
  - Training and test dataset split: 95%/5%

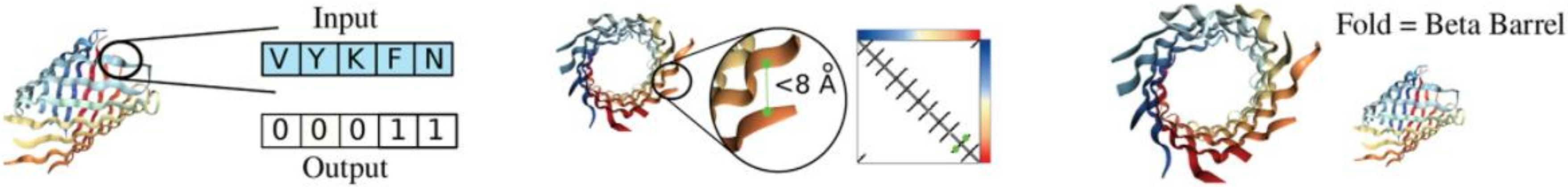
- Supervised datasets
  - Five biologically relevant downstream tasks
  - Dataset ranges in size 8k-50k for training

Tasks

- Self-supervised:
  - Next token prediction
  - Mask token prediction
- Downstream tasks:
  - Protein SS
  - Protein contact map
  - Protein homology detection
  - Fluorescence
  - Stability

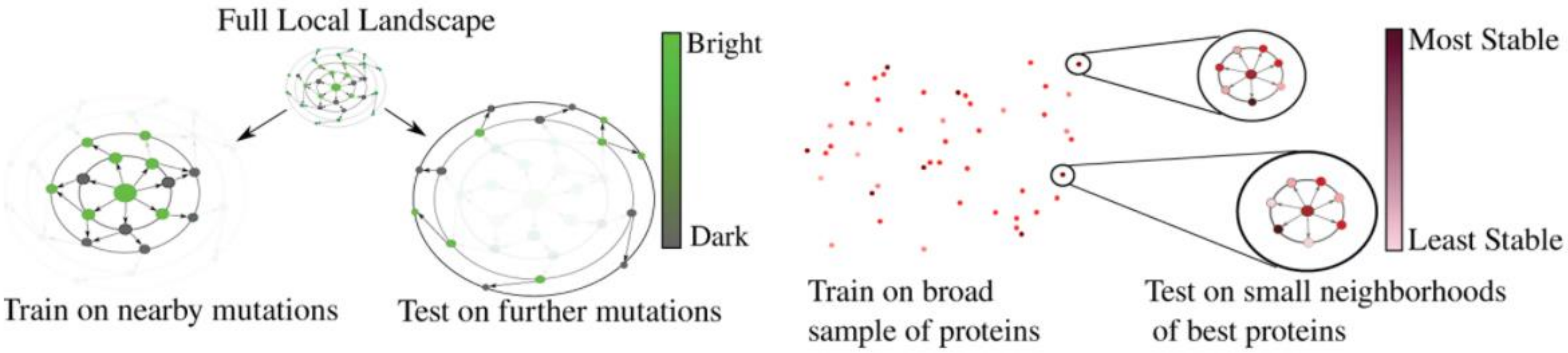
Tasks

Figure 1:



(a) Secondary Structure (b) Contact Prediction (c) Remote Homology

Figure 2:


Full Local Landscape
Bright
Dark
Train on nearby mutations
Test on further mutations
Most Stable
Least Stable
Train on broad sample of proteins
Test on small neighborhoods of best proteins

(a) Fluorescence (b) Stability

Losses

- Two self-supervised losses for NLP task
    - Next-token prediction
    - Masked-token prediction
- Protein specific loss
    - Further supervised pretraining of models
    - Supervised pretraining on contact prediction and remote homology detection can improve secondary structure prediction (Beplar et al)

Architectures for Downstream tasks

- LSTM
  - Two 3-layer LSTMs with 1024 hidden units corresponding to the forward and backward language models

- Transformer
  - 12-layer transformer
  - Each layer hidden size 512 units and 8 attention head
  - 38M parameters

- ResNet
  - 35 residual blocks
  - Each containing 2 conv. Layer with 256 filters
  - Kernel size 9, dilation rate 2

Architectures for Downstream tasks (cont.)

- Bepler et al.
  - Two 3-layer LSTMs with 512 hidden units corresponding to the forward and backward language models
- Alley et al.
  - Unidirectional mLSTM
  - 1900 hidden units

Architectures for Downstream tasks: baseline

- Secondary structure: NetSerf2.0
  - Two convolution layers followed by two bidirectional LSTM followed by a linear output layer

- Contact prediction architecture: Similar to RaptorX-contact
  - 30 residual blocks having 2 conv. Layers each
  - 64 filter for each conv. layer

- Remote homology and protein engineering architecture
  - Predict attention value for each position of sequence to compute attention-weighted mean embedding
  - Followed by 512 hidden unit dense layer
  - Followed by relu and linear activation

Results: language modeling

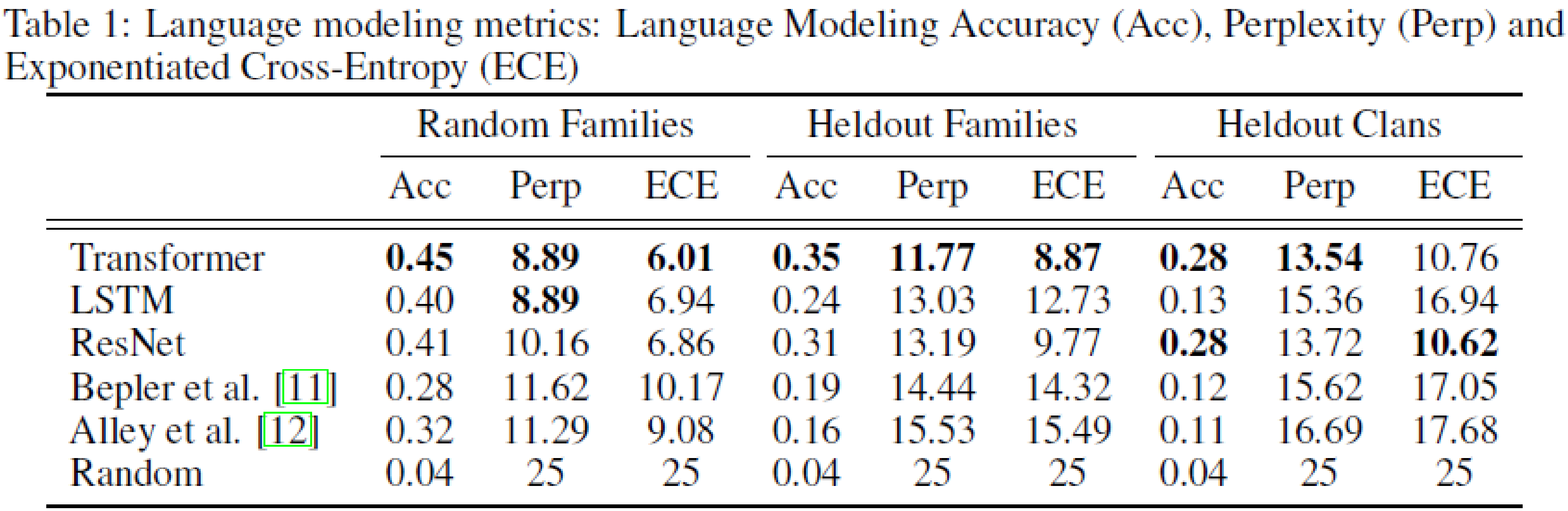


Table 1: Language modeling metrics: Language Modeling Accuracy (Acc), Perplexity (Perp) and Exponentiated Cross-Entropy (ECE)

	Random Families			Heldout Families			Heldout Clans		
	Acc	Perp	ECE	Acc	Perp	ECE	Acc	Perp	ECE
Transformer	**0.45**	**8.89**	**6.01**	**0.35**	**11.77**	**8.87**	**0.28**	**13.54**	10.76
LSTM	0.40	**8.89**	6.94	0.24	13.03	12.73	0.13	15.36	16.94
ResNet	0.41	10.16	6.86	0.31	13.19	9.77	**0.28**	13.72	**10.62**
Bepler et al. [11]	0.28	11.62	10.17	0.19	14.44	14.32	0.12	15.62	17.05
Alley et al. [12]	0.32	11.29	9.08	0.16	15.53	15.49	0.11	16.69	17.68
Random	0.04	25	25	0.04	25	25	0.04	25	25

Results: downstream tasks

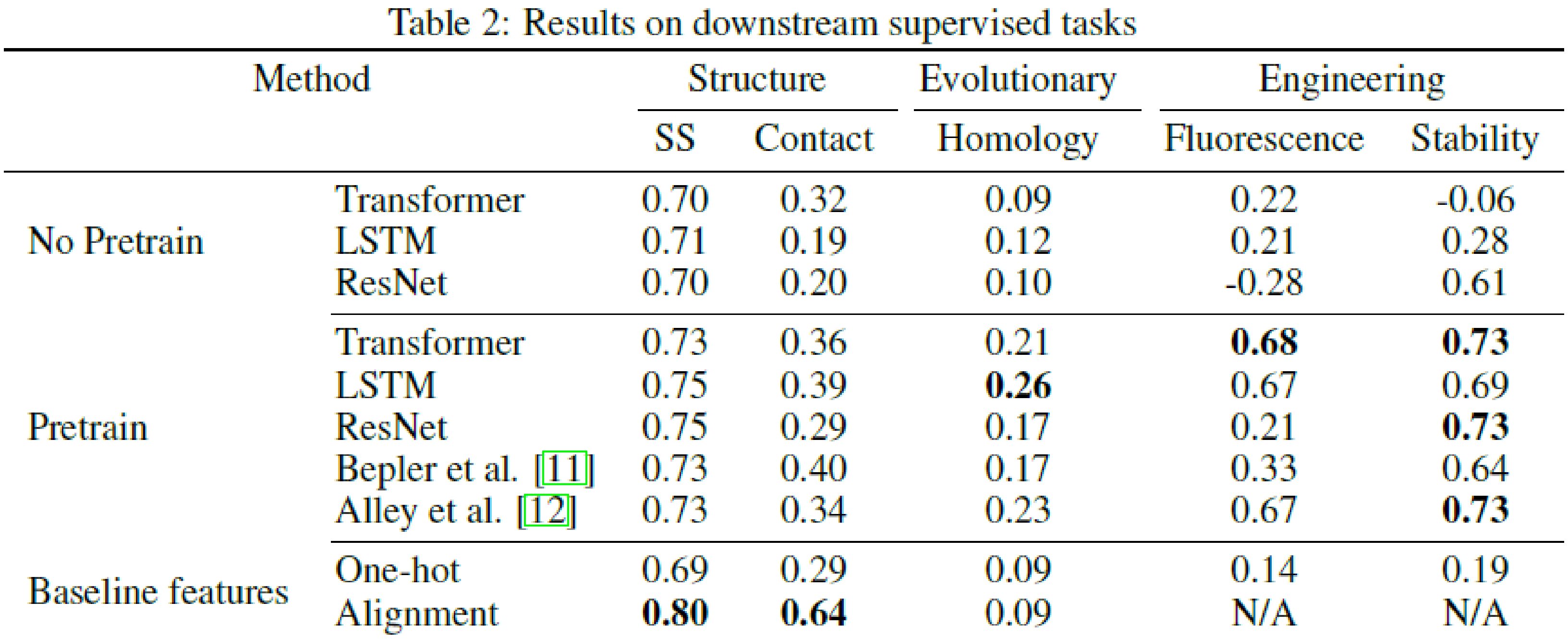
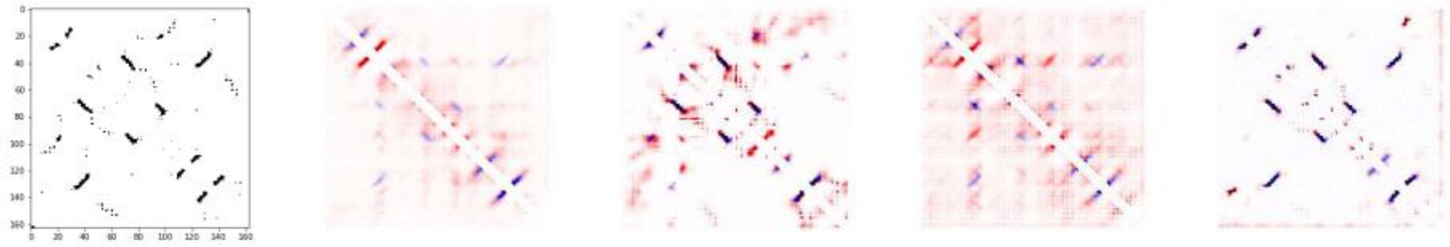


Table 2: Results on downstream supervised tasks

Method		Structure		Evolutionary	Engineering	
		SS	Contact	Homology	Fluorescence	Stability
No Pretrain	Transformer	0.70	0.32	0.09	0.22	-0.06
	LSTM	0.71	0.19	0.12	0.21	0.28
	ResNet	0.70	0.20	0.10	-0.28	0.61
Pretrain	Transformer	0.73	0.36	0.21	**0.68**	**0.73**
	LSTM	0.75	0.39	**0.26**	0.67	0.69
	ResNet	0.75	0.29	0.17	0.21	**0.73**
	Bepler et al. [11]	0.73	0.40	0.17	0.33	0.64
	Alley et al. [12]	0.73	0.34	0.23	0.67	**0.73**
Baseline features	One-hot	0.69	0.29	0.09	0.14	0.19
	Alignment	**0.80**	**0.64**	0.09	N/A	N/A



(a) True Contacts (b) LSTM (c) LSTM Pretrain (d) One Hot (e) Alignment

Discussions

- Alignment-based input currently outperforms self-supervised featurization
  - All state-of-the art methods use alignment-based input features
  - Can pertaining along with alignment-based input improve performance?
- Multiple tasks are required to appropriately benchmark a given model
  - Transformer performs worst in ss and contact prediction but best in fluorescence and stability tasks
- A challenge for future research in self-supervised learning
  - Create models for protein specific tasks or generalized tasks?