Evaluating Protein Transfer Learning with TAPE

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

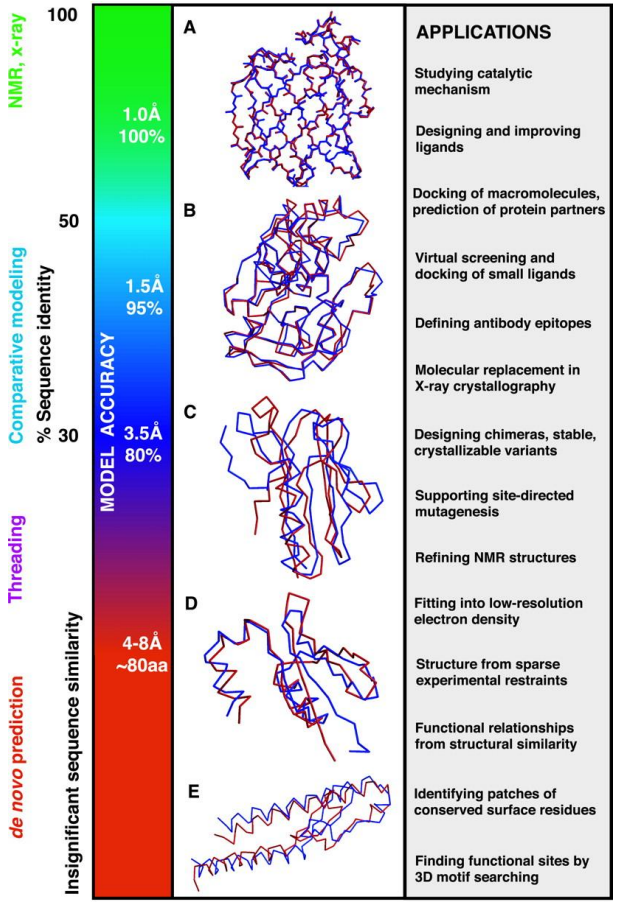
NeurIPS 2019

CSE 590C - 11/2/20

Nicasia Beebe-Wang, Pascal Sturmfels and Sheng Wang

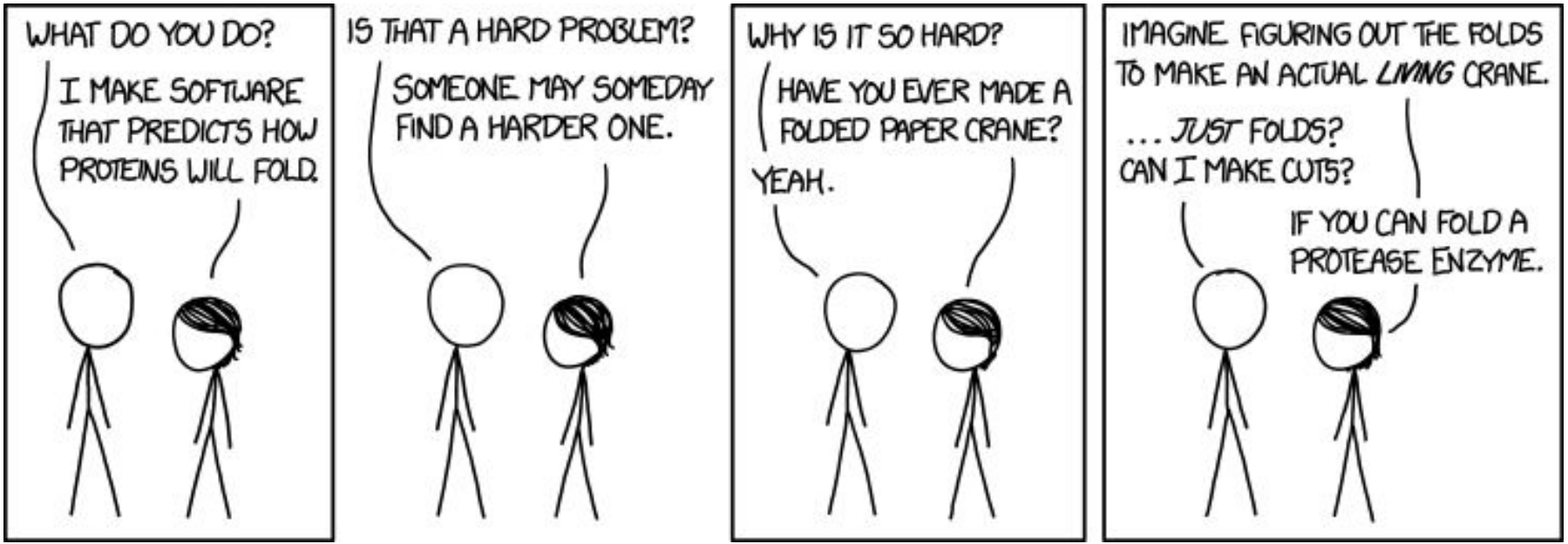
Background: Proteins

- Predicting structural and functional properties from protein sequences is a long-standing goal in computational biology
- Better prediction enables applications like antibiotic resistance prediction and drug engineering/discovery



Baker, David, and Andrej Sali. "Protein structure prediction and structural genomics." *Science* 294.5540 (2001): 93-96.

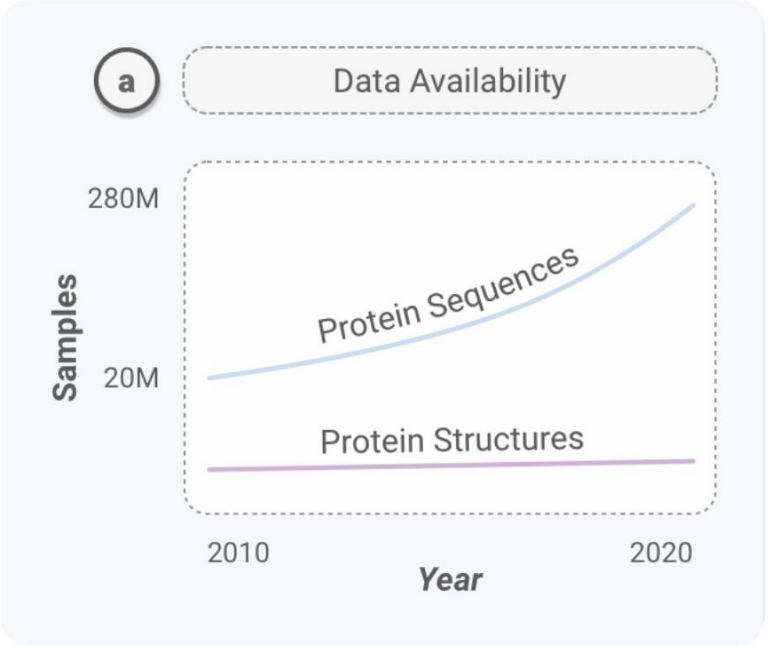
Background: Proteins



This slide was gratuitously stolen from the TAPE paper presentation

Background: Protein Databases

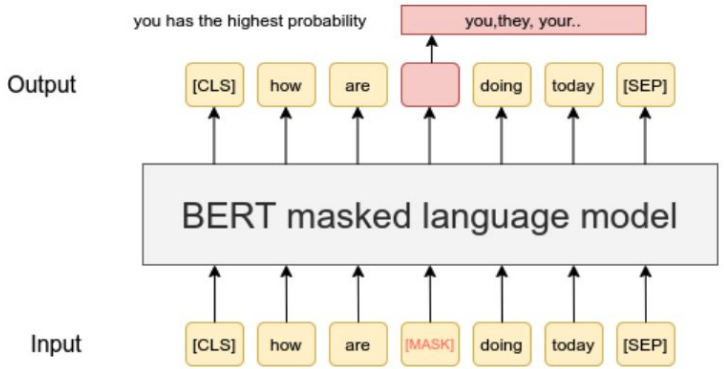
- Collecting labeled data is very expensive! Crystallography experiments can cost >$200,000
- Collecting unlabeled data (sequencing) is relatively cheap, meaning there is way more unlabeled data than labeled



Madani, Ali, et al. "ProGen: Language Modeling for Protein Generation." arXiv preprint arXiv:2004.03497 (2020).

Self-Supervised Learning

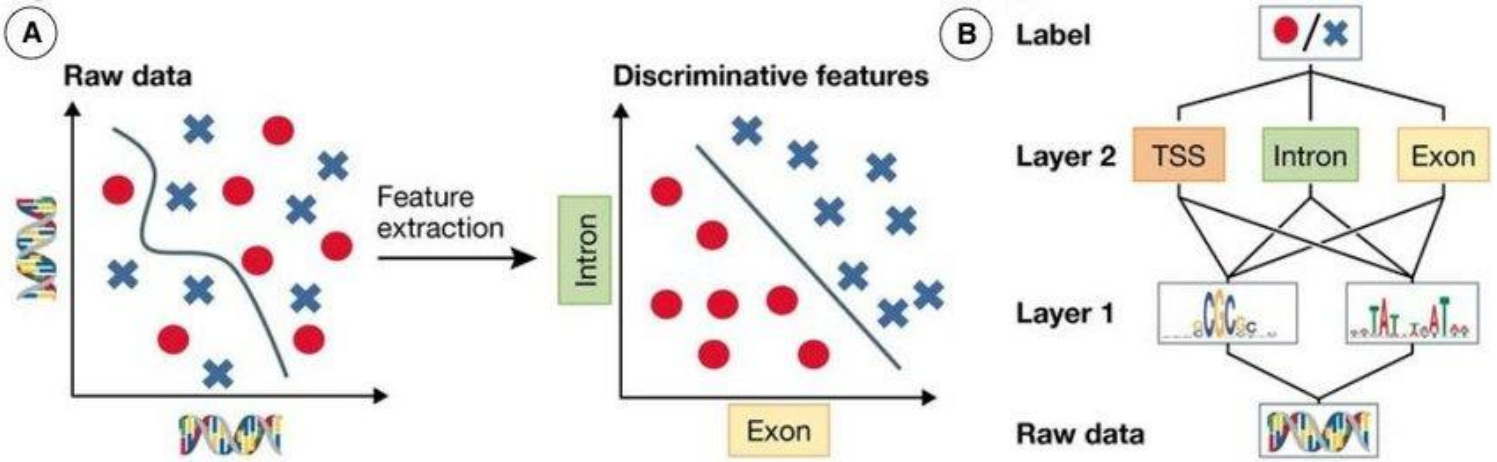
- How can we use unlabeled data to train better models?
- In the natural language processing domain, unlabeled data is leveraged through *self supervision:* pre-training on the unlabeled data via a proxy task that requires no labels
- Self-supervised models consistently outperform models trained from scratch



Jain, Abhilash. "Finnish Language Modeling with Deep Transformer Models." arXiv preprint arXiv:2003.11562 (2020).

Self-Supervised Learning

- Why does this help? Self-supervised learning helps models learn a powerful internal representation of the input



Self-Supervised Protein Models

- Protein modeling and NLP have some similarities: discrete sequence input + large corpus of unlabeled data
- This has inspired many papers applying NLP models to protein sequences

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives [*1,2] Joshua Meier [*1] Tom Sercu [*1] Siddharth Goyal [*1] Zeming Lin [2] Demi Guo [3†] Myle Ott [1] C. Lawrence Zitnick [1] Jerry Ma [4†] Rob Fergus [2]

ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing

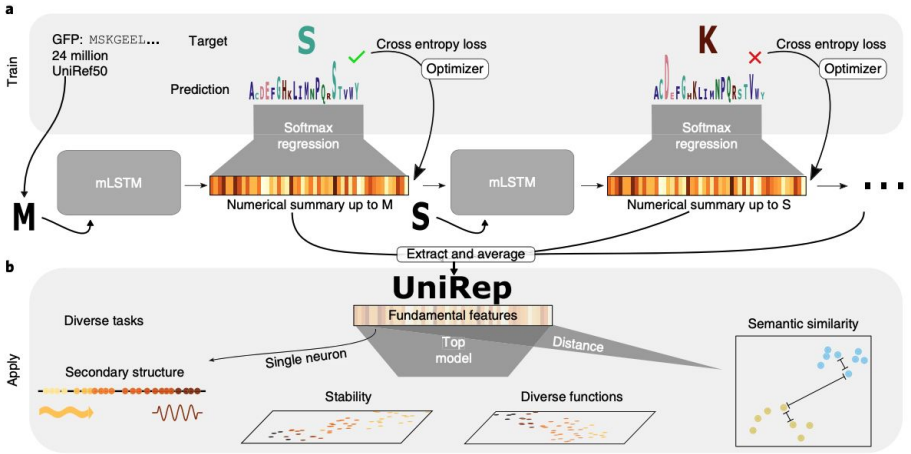
Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik and Burkhard Rost

ProGen: Language Modeling for Protein Generation

Ali Madani [1] Bryan McCann [1] Nikhil Naik [1] Nitish Shirish Keskar [1] Namrata Anand [2] Raphael R. Eguchi [2] Po-Ssu Huang [2] Richard Socher [1]

Unified rational protein engineering with sequence-based deep representation learning

Ethan C. Alley[1,2,6], Grigory Khimulya[6,7], Surojit Biswas[1,3,6], Mohammed AlQuraishi[4] and George M. Church[1,5*]



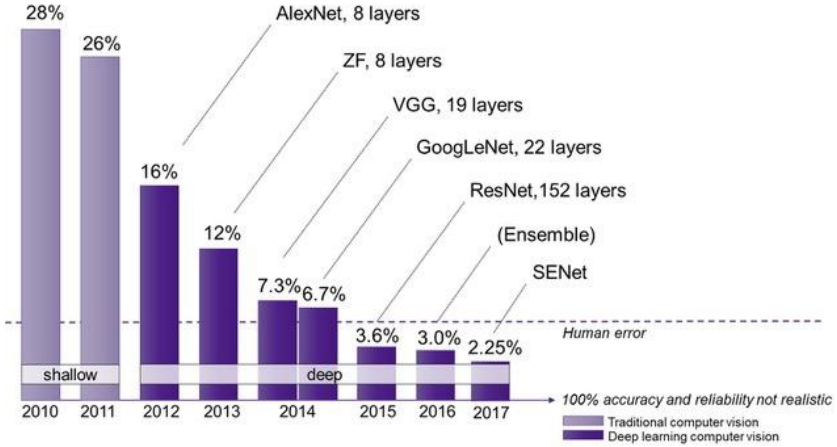
Alley, Ethan C., et al. "Unified rational protein engineering with sequence-based deep representation learning." *Nature methods* 16.12 (2019): 1315-1322.

How do we compare different models?

- Although there is growing interest in applying deep ML models to protein sequences, there is an issue: everyone gets data from a slightly different source, or pre-processes data in slightly different ways!
- Although there is some effort to standardize datasets (like CASP) we want to be able to separate gains from *pre-processing* from gains from *modeling*

The TAPE Benchmark

- The TAPE benchmark is a solution to this problem: it introduces a pre-training dataset and five downstream task datasets
- All of the datasets are pre-processed in the same way, making model comparisons easy
- Similar benchmarks in NLP (GLUE) and vision (ImageNet) have rapidly driven progress in the last 5 years



Notable Protein Benchmarks

- Critical Assessment of protein Structure Prediction (CASP)
- ProteinNet
- Others?

Protein terminology

- Represent a protein *x* of length *L* as a ***sequence of discrete amino acid*** characters ($x_1, x_2, ..., x_L$) in an alphabet of 25 letters (20 standard amino acids, 2 non-standard amino acids, 2 ambiguous amino acids, 1 unknown)

- Each protein has a ***3D structure***
  - Primary (amino acid sequence) → secondary (local features) → tertiary (global features)
  - Proteins often have a few large ***protein domains*** - evolutionary conserved well-defined sub-structures

- ***Homologs:*** two proteins that share a common evolutionary ancestor, but may have very different sequences if they diverged in the distant past

- Quantifying evolutionary relationships is important for avoiding contamination of test sets. In this paper, they mainly rely on ***sequence identity*** (exact amino acid matches)

Modeling Evolutionary Relationships with Sequence Alignments

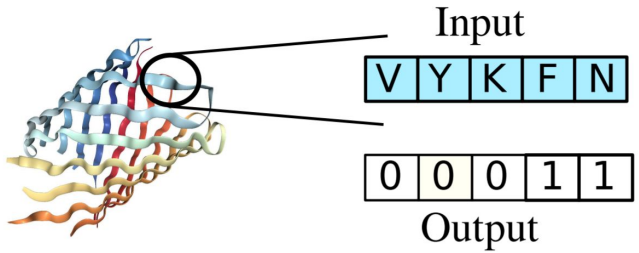
- Querying a protein:
  - An alignment based method uses a scoring system or HMM to align a query protein against proteins in a database
  - Can provide information about local perturbations, which may then be useful for understanding changes to structure/function
- Multiple alignment:
  - For a group of proteins, can construct a profile to summarize frequencies of amino acids → useful representation for downstream tasks

Datasets

- Goal: curate standardized benchmarking datasets with specific training, validation and test splits
- Pre-training corpus: large unlabeled sequence dataset
  - **Pfam** - database of 31M protein domains
  - Sequences are clustered into evolutionarily-related groups called **families**.
  - 1% of families are fully held out as a test set, and the remaining families are separated into a 95/5% training/validation split
  - Uniform random split test performance → in-distribution generalization
  - Heldout families test set performance → out-of-distribution generalization
- Supervised datasets - Different for each task, varying between 8 thousand and 50 thousand training examples.

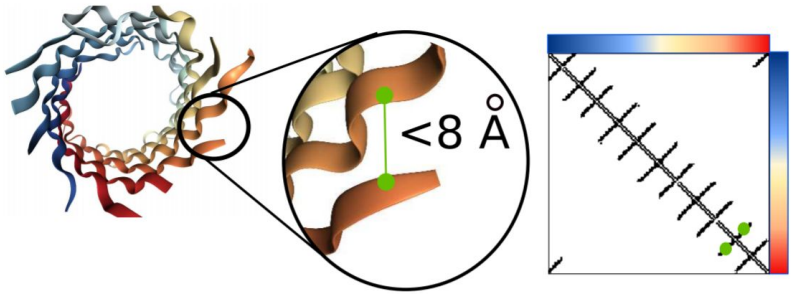
Task 1: Secondary Structure (SS) Prediction

- **Definition**: sequence-to-sequence task; Each input amino acid x_i is mapped to a label $y_i \in$ {Helix(*H*), Strand(*E*), Other(*C*)}.
- **Impact**:
  - Important feature for understanding the function of a protein
  - SS prediction tools are commonly used to create richer input features for higher-level models
- **Generalization**: tests the degree to which models learn ***local*** structure.
- **Metrics + Dataset:**
  - Trained with Klausen et al., 2019 (~11K sequences)
  - Measured (per-amino acid) test set ***accuracy*** on CB513 dataset (~500 sequences)
  - Data splits are filtered at 25% sequence identity to test for broad generalization.



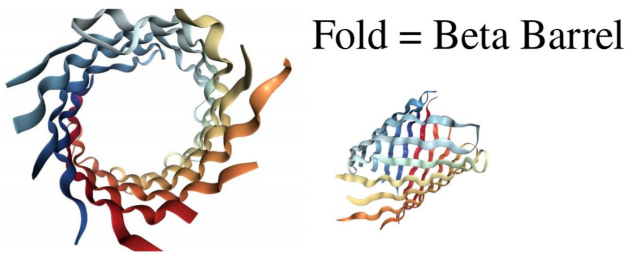
Task 2: Contact Prediction

- **Definition**: pairwise amino acid task; Each pair x_i, x_j of amino acids is labeled $y_{ij} \in \{0, 1\}$ indicating whether the amino acids are “in contact” (< 8Å apart)
- **Impact**: Powerful global information; robust modeling of full 3D protein structure
- **Generalization**: tests the model’s understanding of global protein context.
- **Metrics + Dataset:**
  - ProteinNet dataset (test set from the CASP12 competition) (~26K samples)
  - Data splits are filtered at 30% sequence identity.
  - Precision of the L/5 most likely contacts for medium- and long-range contacts on the ProteinNet CASP12 test set



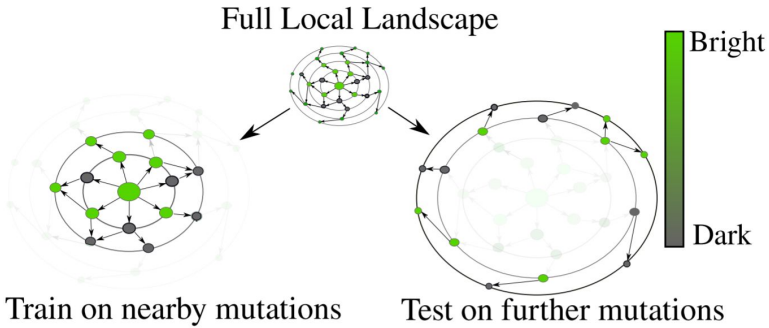
Task 3: Remote Homology Detection

- **Definition**: sequence classification task; Each sequence *x* is mapped to a label $y \in \{1,..., 1195\}$ representing different possible protein folds
- **Impact**: Of interest in microbiology and medicine; e.g.; detecting emerging antibiotic resistant genes
- **Generalization**: tests model's ability to detect structural similarity across distantly related inputs
- **Metrics + Dataset:**
  - Hou et al., 2017 dataset derived from the SCOP 1.75 database of hierarchically classified protein domains
  - Held out entire evolutionary groups from the training set, forcing models to generalize across large evolutionary gaps.
  - Report test classification accuracy.



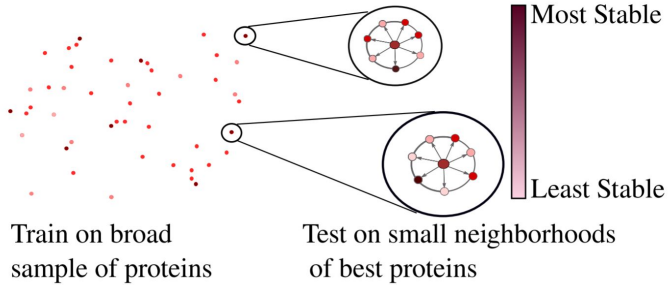
Task 4: Fluorescence Landscape Prediction

- **Definition**: regression task; Each input protein $x \rightarrow y \in R$, corresponding to the log-fluorescence intensity of x
- **Impact**: Would allow more efficient exploration of the landscape
- **Generalization**: tests model's ability to:
  - distinguish between very similar inputs
  - generalize to unseen combinations of mutations
- **Metrics + Dataset:**
  - Data generated from Deep Mutational Scanning (Sarkisyan et al., 2016) - characterized small neighborhoods of parent proteins through mutagenesis of avGFP protein
  - Train+Val: Hamming distance 3 neighborhood; Test: Hamming distance 4-5 neighborhood
  - Report Spearman's ρ on the test set.



Task 5: Stability Landscape Prediction

- **Definition**: regression task; each input protein $x \rightarrow y \in R$ measuring the folding stability (most extreme circumstances in which protein *x* maintains its fold above a protease concentration threshold)
- **Impact**: Would allow finding better refinements of top candidates of expensive protein engineering experiments
- **Generalization**: tests model's ability to generalize from a broad sampling of sequences and localize info in a neighborhood of a few sequences
- **Metrics + Dataset:**
  - Data from Rocklin et al., 2017 - Train/val sets come from 4 rounds of experimental design; test set contains Hamming distance-1 neighbors of top candidates
  - Report Spearman's ρ on the test set.



Experimental Overview

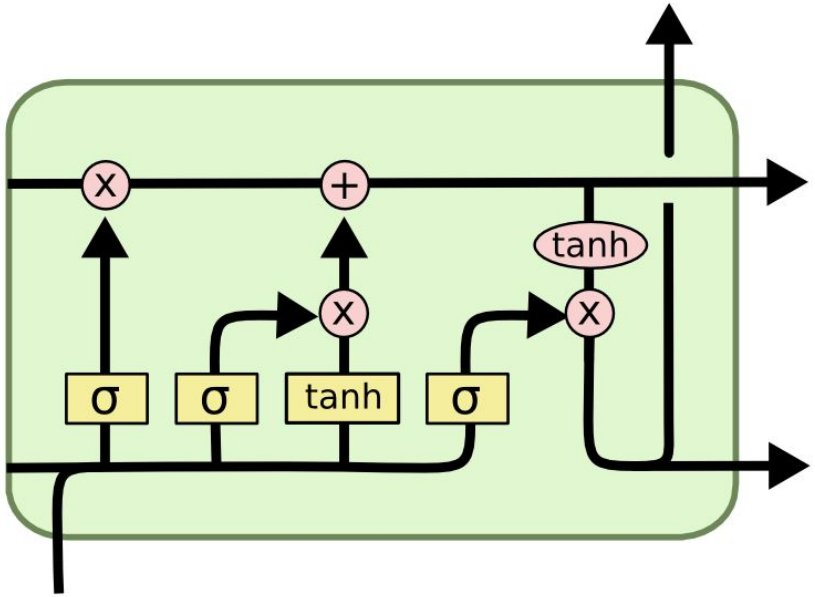
- The TAPE paper compares a host of models on the five downstream tasks:
  - A Transformer
  - An LSTM
  - A Residual Network
  - The CNN/LSTM from Bepler, Tristan, and Bonnie Berger. "Learning protein sequence embeddings using information from structure."
  - The LSTM from Alley, Ethan C., et al. "Unified rational protein engineering with sequence-based deep representation learning."
  - A one-hot and an alignment-based baseline

The TAPE models

- Three models, all inspired by NLP models: a transformer, a residual network, and an LSTM
- They evaluate both pre-trained models and models trained from scratch on all five downstream tasks

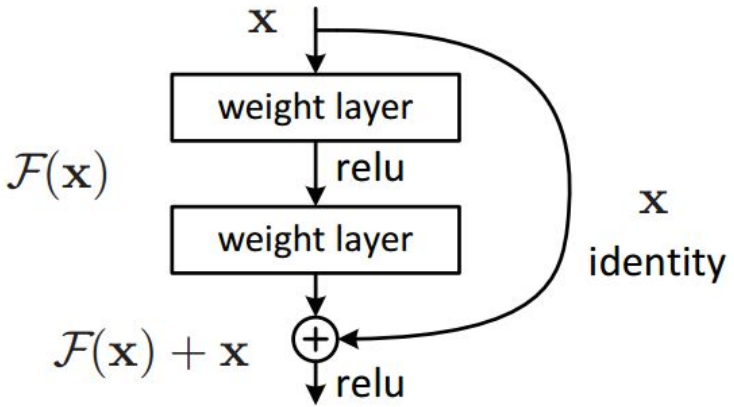
LSTM

- An LSTM is a variant of a recurrent neural network, and has been used for sequence learning for years
- Their LSTM is bidirectional, and has three layers of 1024 units



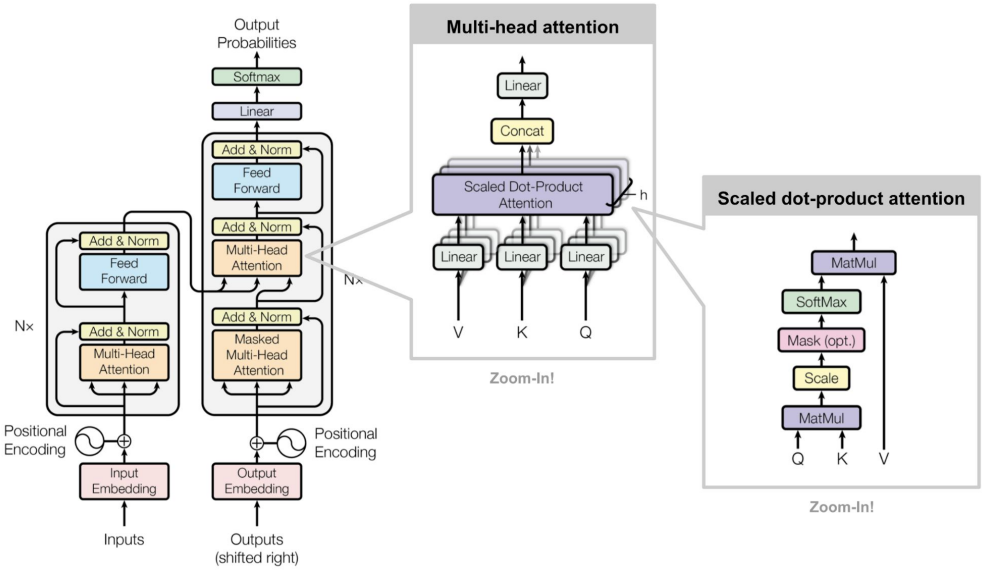
ResNet

- A residual network is a type of convolutional neural network, and was invented for vision tasks
- It has since been applied in a 1D sense for sequences
- Their ResNet has 35 residual blocks, each with two large convolutional layers (kernel size of 9, 256 filters, dilation 2)



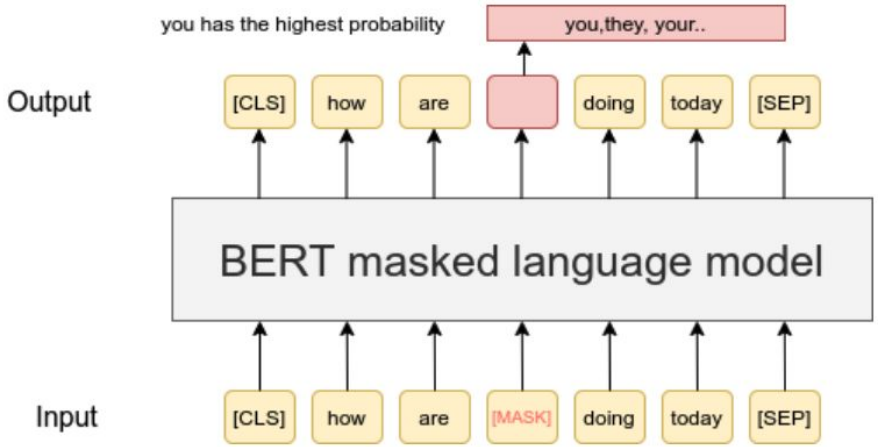
Transformer

- Their final model is a transformer, which is the current SOTA for NLP tasks
- It consists of alternating feed forward layers and self-attention layers, which compute and weight pairwise similarity between all pairs of inputs
- 12 layers deep, 12 attention heads



Pre-Training

- All models are pre-trained for a week using the pfam dataset
- They are trained with masked language (amino) modeling: predicting masked amino acids from surrounding context

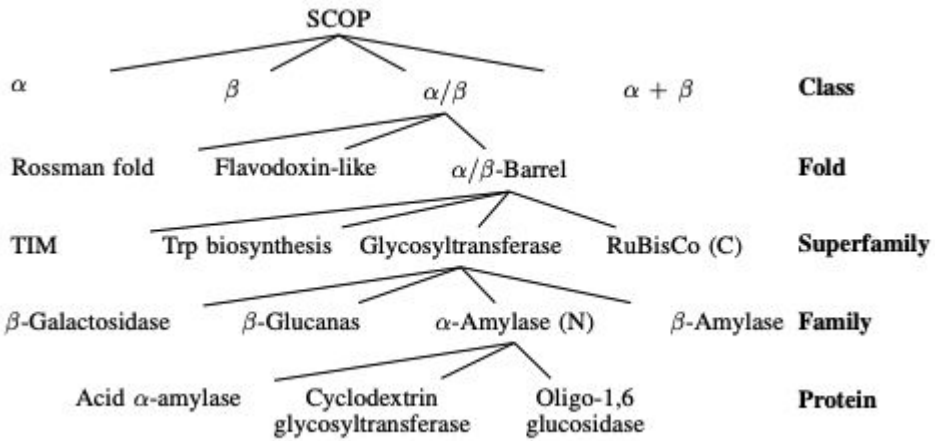


Baselines

- They compare against four baselines
  - The CNN/LSTM from Bepler, Tristan, and Bonnie Berger. "Learning protein sequence embeddings using information from structure."
  - The LSTM from Alley, Ethan C., et al. "Unified rational protein engineering with sequence-based deep representation learning."
  - A one-hot baseline
  - An alignment-based baseline

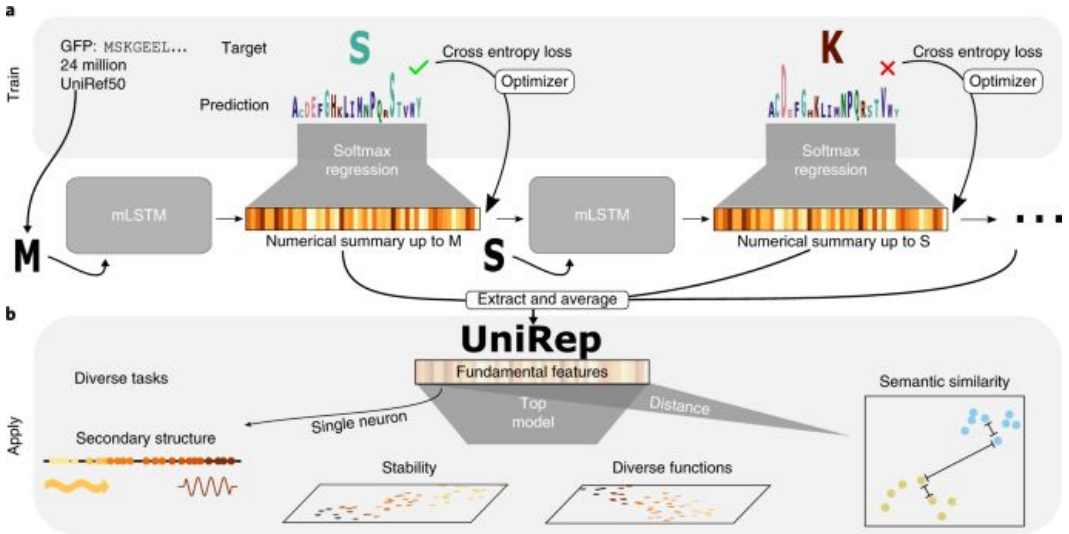
Learning protein sequence embeddings using information from structure

- This paper introduces a new pre-training task that predicts the SCOP similarity level of two proteins, based on their embeddings
- Uses a joint biLSTM+CNN architecture



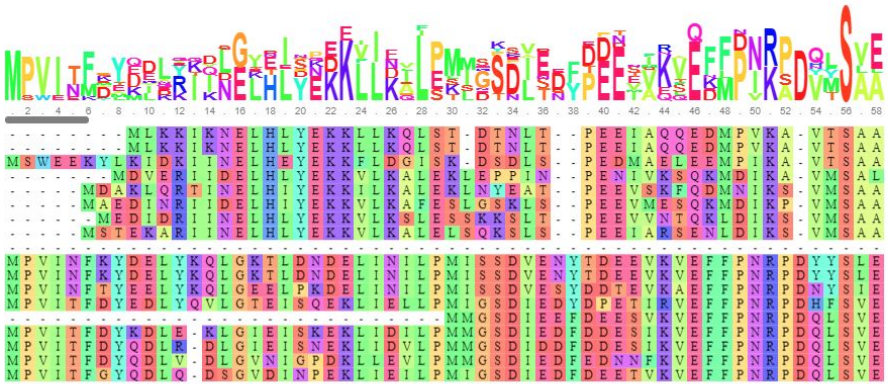
Unified rational protein engineering with sequence-based deep representation learning

- This paper introduces an LSTM that is trained to generate a protein left-to-right, one amino acid at a time, as opposed to masked language modeling



One-Hot and Alignment Baselines

- In addition, they compare against domain specific (CNN + LSTM architectures) models that use either one-hot featurization, or the standard HMM profile featurization derived from multiple sequence alignments
- The alignment-based baselines are Netsurfp2.0, RaptorX and DeepSF for secondary structure, contact prediction and remote homology respectively



Results: Language modeling metrics

- Drop in out-of-distribution generalization ability; Held-out family accuracy is consistently lower than random-split
- Lower perplexity will not necessarily correspond with downstream prediction tasks

	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

Acc=Accuracy, Perp=Perplexity, ECE=Exponentiated Cross-Entropy

Results: supervised tasks

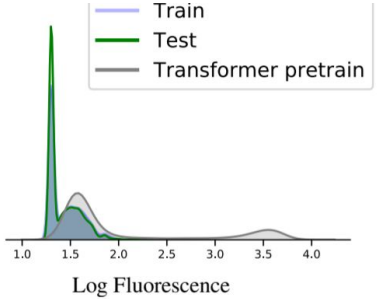
- Self-supervised pretraining almost always improves performance
- BUT, for structure tasks, NN methods do worse than alignment-based baselines

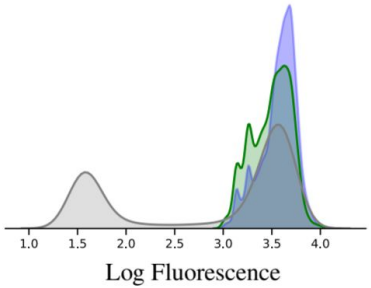
Method		Structure		Evolutionary	Engineering	
		SS	Contact	Homology	Fluorescence	Stability
No Pretrain	Transformer	0.70	0.32	0.09	0.22	-0.06
	LSTM	0.71	0.19	0.12	0.21	0.28
	ResNet	0.70	0.20	0.10	-0.28	0.61
Pretrain	Transformer	0.73	0.36	0.21	**0.68**	**0.73**
	LSTM	0.75	0.39	**0.26**	0.67	0.69
	ResNet	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
		accuracy	precision	accuracy	Spearman's ρ	

Protein engineering: Beneficial vs deleterious mutations

Fluorescence task

- bimodal distribution with dark and bright modes
- Important goal: distinguish between beneficial and deleterious mutations
- t-SNE of Pretrained transformer embeddings: Some successful clustering, but many proteins misclassified



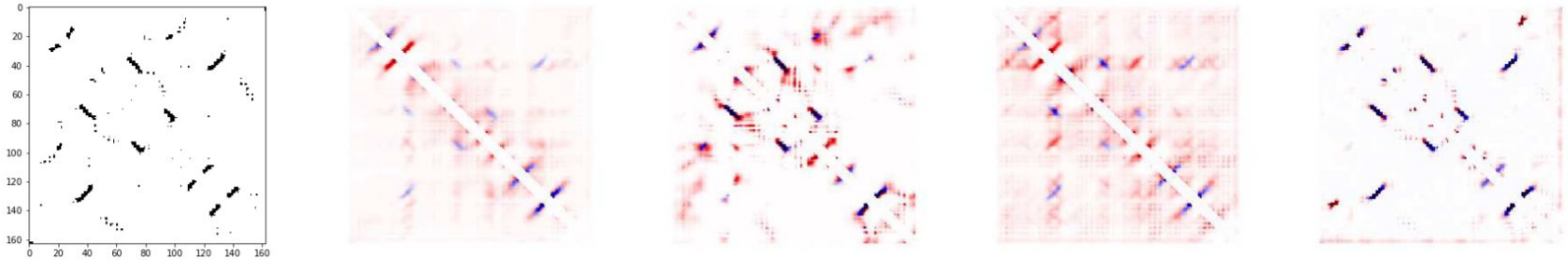


Stability task:

- Use parent protein as decision boundary and label mutation as beneficial or deleterious based on change in protein stability prediction
- Best pretrained: 70% accuracy; Best non-pretrained: 68% accuracy

Long range contact prediction

Blue: true positive
Red: false positive



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

Figure 4: Predicted contacts for chain 1A of a Bacterioferritin comigratory protein (pdbid: 3GKN).

- LSTM: pretraining helps the model capture more long-range info & improves overall resolution
- Hand-engineered alignment features lead to much better performance

Discussion

- Need for multiple benchmark tasks

- Self-supervised pretraining almost always improves performance
- Performance gap for structure tasks → opportunity for innovation (especially incorporating alignment-based representations)

- Datasets + benchmarks in TAPE: systematic model-evaluation framework for ML researchers to contribute to the field

Discussion questions

- Self-supervision: Do you think the standard language modeling as a task is enough? Should researchers create protein-specific tasks?
- Opportunity for multi-task learning?
- Do you think pre-training would be less useful if there are a lot of training samples available?
- Are there any missing tasks in this benchmark? What other protein prediction tasks do you think are important to include?
- Would you use this?

Extra slides

Table S1: Dataset sizes

Task	Train	Valid	Test
Language Modeling	32207059	N/A	2147130 (Random-split) / 44314 (Heldout families)
Secondary Structure	8678	2170	513 (CB513) / 115 (TS115) / 21 (CASP12)
Contact Prediction	25299	224	40 (CASP12)
Remote Homology	12312	736	718 (Fold) / 1254 (Superfamily) / 1272 (Family)
Fluorescence	21446	5362	27217
Stability	53679	2447	12839