TAPE

NIPS 19

Property prediction Benchmark

- 5 semi-supervised tasks
  1. Secondary structure: helix, strand, other
  2. Structure: each pair is in contact or not
  3. Remote homology detection
  4. Flourescence
  5. Stability
- Transformer > LSTM > ResNet > UniRep

Transformer mLSTM

Evaluating Protein Transfer Learning with Tasks Assessing Protein Embeddings (TAPE)

Protein modeling is an increasingly popular area of machine learning research. Semi-supervised learning has emerged as an important paradigm in protein modeling due to the high cost of acquiring supervised protein labels, but the current literature is fragmented when it comes to datasets and standardized evaluation techniques. To facilitate progress in this field, we introduce the Tasks Assessing Protein Embeddings (TAPE), a set of five biologically relevant semi-supervised learning tasks spread across different domains of protein biology. We curate tasks into specific training, validation, and test splits to ensure that each task tests biologically relevant generalization that transfers to real-life scenarios. We benchmark a range of approaches to semi-supervised protein representation learning, which span recent work as well as canonical sequence learning techniques. We find that self-supervised pretraining is helpful for almost all models on all tasks, more than doubling performance in some cases. Despite this increase, in several cases features learned by self-supervised pretraining still lag behind features extracted by state-of-the-art non-neural techniques. This gap in performance suggests a huge opportunity for innovative architecture design and improved modeling paradigms that better capture the signal in biological sequences. TAPE will help the machine learning community focus effort on scientifically relevant problems.

CARP

bioRxiv 22

Representation

- Convolutional masked language model
- ByteNet encoder
- Dilated convolution

Autoencoder CNN

Convolutions are competitive with transformers for protein sequence pretraining

Protein design aims to build new proteins from scratch thereby holding the potential to tackle many environmental and biomedical problems. Recent progress in the field of natural language processing (NLP) has enabled the implementation of ever-growing language models capable of understanding and generating text with human-like capabilities. Given the many similarities between human languages and protein sequences, the use of NLP models offers itself for predictive tasks in protein research. Motivated by the evident success of generative Transformer-based language models such as the GPT-x series, we developed ProtGPT2, a language model trained on protein space that generates *de novo* protein sequences that follow the principles of natural ones. In particular, the generated proteins display amino acid propensities which resemble natural proteins. Disorder and secondary structure prediction indicate that 88% of ProtGPT2-generated proteins are globular, in line with natural sequences. Sensitive sequence searches in protein databases show that ProtGPT2 sequences are distantly related to natural ones, and similarity networks further demonstrate that ProtGPT2 is sampling unexplored regions of protein space. AlphaFold prediction of ProtGPT2-sequences yielded well-folded non-idealized structures with embodiments as well as large loops and revealed new topologies not captured in current structure databases. ProtGPT2 has learned to speak the protein language. It has the potential to generate *de novo* proteins in a high throughput fashion in a matter of seconds. The model is easy-to-use and freely available.

Design and Engineering

ReLeaSE
#RL #SMILES

GEGL
#RL #GA #SMILES

DyNA-PPO
#RL #Binding #PO

ProGPT2
#GPT #De-novo

3D-MolGNN
#Graph #Actor-critic

Deep learning and architecture

SAC
#RL

PPO
#RL #PO

Decision Transformer
#RL #Transformer

Annotation

Color box
Reviewed!
Click to see the summary

Grey box
Not yet reviewed

Representation

eUniRep
#LSTM

UniRep
#LSTM

CARP
#CNN

GVP
#Graph

MSA2Prot
#Transformer

Hallucination
#MCMC

Property prediction

ProteinBERT
#BERT

LM-GVP
#Graph #LM

DeepFRI
#GCN

Augmentation

AlphaFold2
#Graph #Attention

ProteinQA
#Graph #QA

trRosetta
#ResNet

ProteinSolver
#Graph #Inverse

ProteinGAN
#GAN #Sequence

GVP+AF
#Inverse #Structure

ProteinGCN
#GCN #QA

AR-VAE
#VAE #Sequence

DeepGO
#CNN

DeepGOPlus
#CNN

DeepGOZero
#CNN #Zero-shot

TAPE
#Benchmark #Transformer

GO
#Benchmark

Folding / Inverse folding