LM-GVP

Scientific Reports 22

Property prediction

- Language model: amino acid information
- Graph neural network: structure information
- Hidden units of LM is concatenated to the scalar features of GVP
- Fine-tuning LM with LM-GVP architecture

GVP Transformer

Dual representation

An extensible sequence and structure informed DL framework for protein property prediction

Proteins perform many essential functions in biological systems and can be successfully developed as bio-therapeutics. It is invaluable to be able to predict their properties based on a proposed sequence and structure. In this study, we developed a novel generalizable deep learning framework, LM-GVP, composed of a protein Language Model (LM) and Graph Neural Network (GNN) to leverage information from both 1D amino acid sequences and 3D structures of proteins. Our approach outperformed the state-of-the-art protein LMs on a variety of property prediction tasks including fluorescence, protease stability, and protein functions from Gene Ontology (GO). We also illustrated insights into how a GNN prediction head can inform the fine-tuning of protein LMs to better leverage structural information. We envision that our deep learning framework will be generalizable to many protein property prediction problems to greatly accelerate protein engineering and drug development.

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.

GVP

ICLR 21

Representation

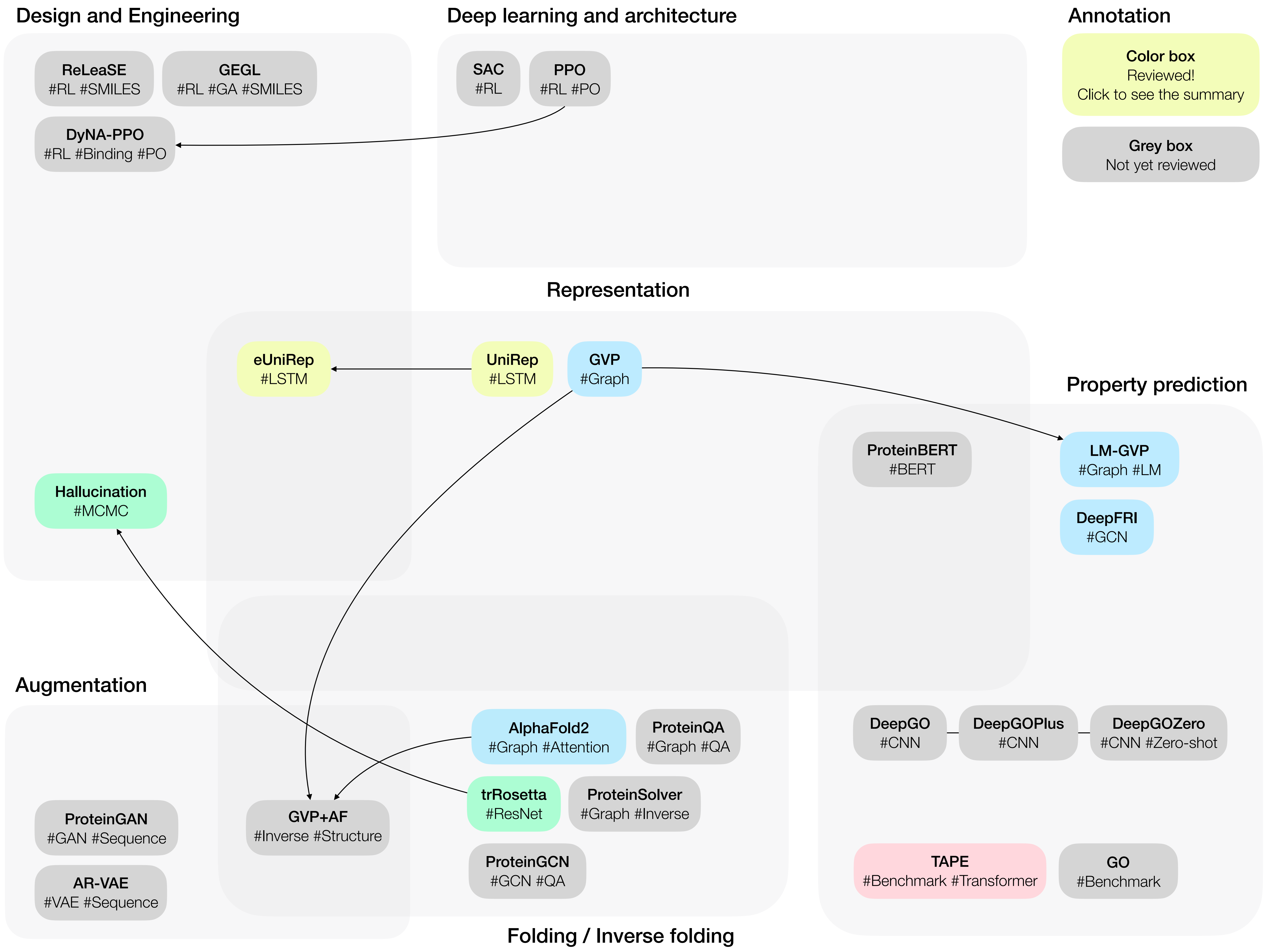
- Dual representation: scalar and vector
- Augment graph neural networks with geometric reasoning layers
- Learning vector-valued and scalar-valued functions over geometric vectors and scalars
- Can approximate any rotation-invariant and reflection-invariant function of vectors

Graph learning

Dual representation

Learning from protein structure with geometric vector perceptrons

Learning on 3D structures of large biomolecules is emerging as a distinct area in machine learning, but there has yet to emerge a unifying network architecture that simultaneously leverages the graph-structured and geometric aspects of the problem domain. To address this gap, we introduce geometric vector perceptrons, which extend standard dense layers to operate on collections of Euclidean vectors. Graph neural networks equipped with such layers are able to perform both geometric and relational reasoning on efficient and natural representations of macromolecular structure. We demonstrate our approach on two important problems in learning from protein structure: model quality assessment and computational protein design. Our approach improves over existing classes of architectures, including state-of-the-art graph-based and voxel-based methods.



Design and Engineering

ReLeaSE
#RL #SMILES

GEGL
#RL #GA #SMILES

DyNA-PPO
#RL #Binding #PO

Hallucination
#MCMC

Deep learning and architecture

SAC
#RL

PPO
#RL #PO

Annotation

Color box
Reviewed!
Click to see the summary

Grey box
Not yet reviewed

Representation

eUniRep
#LSTM

UniRep
#LSTM

GVP
#Graph

Property prediction

ProteinBERT
#BERT

LM-GVP
#Graph #LM

DeepFRI
#GCN

DeepGO
#CNN

DeepGOPlus
#CNN

DeepGOZero
#CNN #Zero-shot

TAPE
#Benchmark #Transformer

GO
#Benchmark

Augmentation

ProteinGAN
#GAN #Sequence

AR-VAE
#VAE #Sequence

GVP+AF
#Inverse #Structure

AlphaFold2
#Graph #Attention

ProteinQA
#Graph #QA

trRosetta
#ResNet

ProteinSolver
#Graph #Inverse

ProteinGCN
#GCN #QA

Folding / Inverse folding