eUniRep

Nature Methods | 21

Engineering | Representation

- UniRep: pre-trained supervisely on large data
- eUniRep: fine-tune UniRep to local sequence neighborhood of target protein
- Directed evolution: MCMC
- 24-to-24 design: fine-tune on 24 proteins and 24 designs is sufficient to observe >WT
- Diverse and novel designs

mLSTM | MCMC

Evolutionary fine-tuning

Low-N protein engineering with data-efficient deep learning

Protein engineering has enormous academic and industrial potential. However, it is limited by the lack of experimental assays that are consistent with the design goal and sufficiently high throughput to find rare, enhanced variants. Here we introduce a machine learning-guided paradigm that can use as few as 24 functionally assayed mutant sequences to build an accurate virtual fitness landscape and screen ten million sequences via in silico directed evolution. As demonstrated in two dissimilar proteins, GFP from *Aequorea victoria* (avGFP) and *E. coli* strain TEM-1 β-lactamase, top candidates from a single round are diverse and as active as engineered mutants obtained from previous high-throughput efforts. By distilling information from natural protein sequence landscapes, our model learns a latent representation of 'unnaturalness', which helps to guide search away from nonfunctional sequence neighborhoods. Subsequent low-N supervision then identifies improvements to the activity of interest. In sum, our approach enables efficient use of resource-intensive high-fidelity assays without sacrificing throughput, and helps to accelerate engineered proteins into the fermenter, field and clinic.

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

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

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