# UniRep

Nature Methods 19

Representation

- Multiplicative LSTM
- Embedding: hidden units of LSTM
- Top model: can be applied to diverse tasks
- Rich information: evolutionary, functional and chemical property, secondary structure
- Tested on ~10 benchmarks of structural and functional properties of protein

mLSTM Top model

## Unified rational protein engineering with sequence-based deep representation learning

Rational protein engineering requires a holistic understanding of protein function. Here, we apply deep learning to unlabeled amino-acid sequences to distill the fundamental features of a protein into a statistical representation that is semantically rich and structurally, evolutionarily and biophysically grounded. We show that the simplest models built on top of this unified representation (UniRep) are broadly applicable and generalize to unseen regions of sequence space. Our data-driven approach predicts the stability of natural and de novo designed proteins, and the quantitative function of molecularly diverse mutants, competitively with the state-of-the-art methods. UniRep further enables two orders of magnitude efficiency improvement in a protein engineering task. UniRep is a versatile summary of fundamental protein features that can be applied across protein engineering informatics.

# eUniRep

Nature Methods 21

Engineering Representation

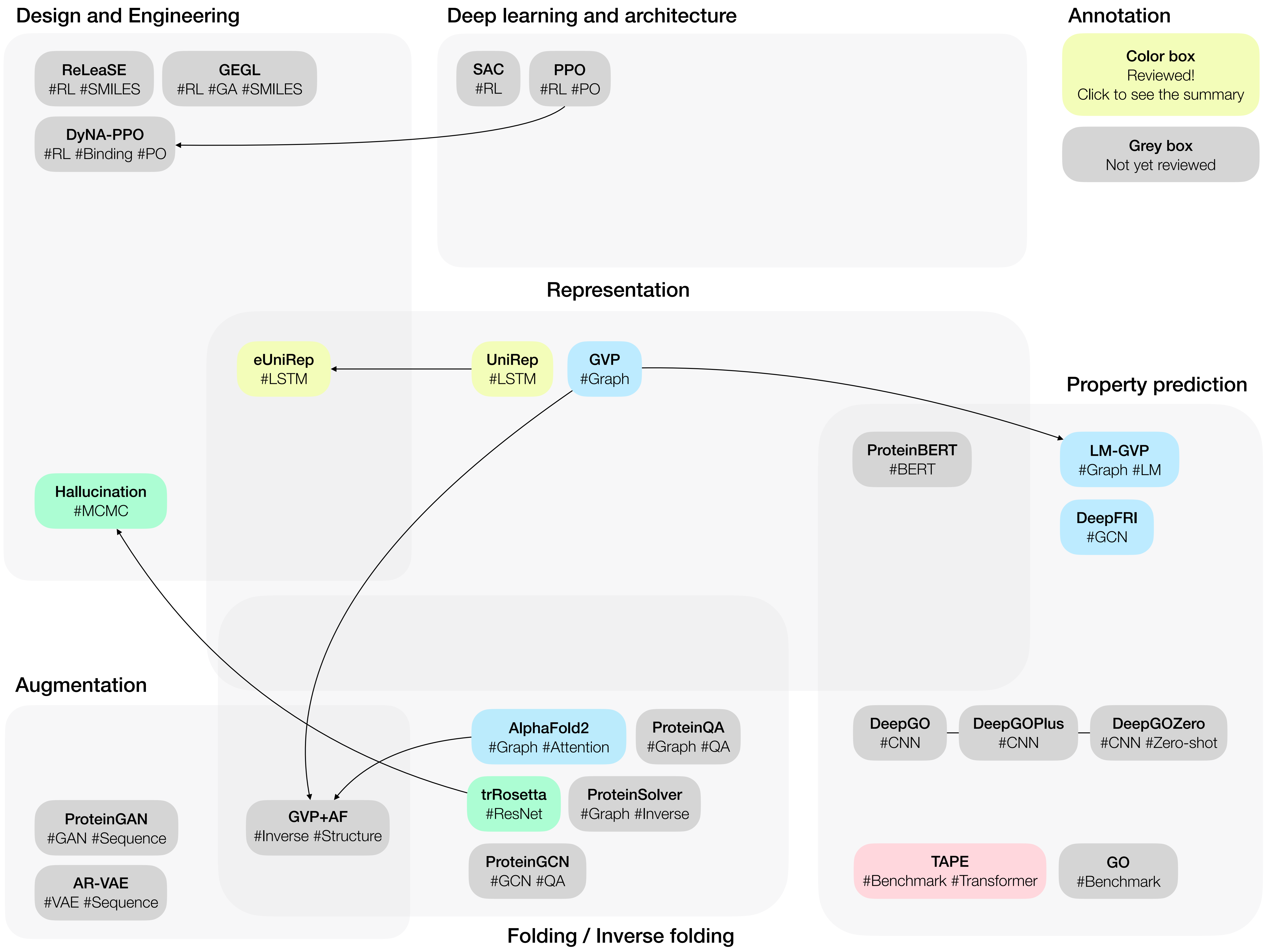
- UniRep: pre-trained supervisely on large data
- eUniRep: fine-tune UniRep to local sequence neighborhood of target protein
- Directed evolution: MCMC
- 24-to-24 design: fine-tune on 24 proteins and 24 designs is sufficient to observe >WT
- Diverse and novel designs

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.



## 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