## 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

# AlphaFold2

Nature 21

Protein folding

- Consider folding as graph inference problem
- Pair representation: relation between residues
- Evoformer: MSA representation updates the pair representation through attention
- Structure module: predict relative positions and translations of backbone
- Iterative local refinement

Graph learning Evoformer

MSA Representation Attention

## Highly accurate protein structure prediction with AlphaFold

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort, the structures of around 100,000 unique proteins have been determined, but this represents a small fraction of the billions of known protein sequences. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'—has been an important open research problem for more than 50 years. Despite recent progress, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14), demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

# 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.

# 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.

# Hallucination

Nature 21

Engineering

- Structure prediction networks can be inverted to generate new protein sequences
- Random sequence → diffused distance map
- MCMC: repeat mutation and accept if it improved score → sharp distance map
- KL-divergence of trRosetta and background network, trRosetta without skip connection

trRosetta MCMC

Background network

## De novo protein design by deep network hallucination

There has been considerable recent progress in protein structure prediction using deep neural networks to predict inter-residue distances from amino acid sequences. Here we investigate whether the information captured by such networks is sufficiently rich to generate new folded proteins with sequences unrelated to those of the naturally occurring proteins used in training the models. We generate random amino acid sequences, and input them into the trRosetta structure prediction network to predict starting residue–residue distance maps, which, as expected, are quite featureless. We then carry out Monte Carlo sampling in amino acid sequence space, optimizing the contrast between the inter-residue distance distributions predicted by the network and background distributions averaged over all proteins. Optimization from different random starting points resulted in novel proteins spanning a wide range of sequences and predicted structures. We obtained synthetic genes encoding 129 of the network-hallucinated sequences, and expressed and purified the proteins in *E. coli;* 27 of the proteins yielded monodisperse species with circular dichroism spectra consistent with the hallucinated structures. We determined the three-dimensional structures of three of the hallucinated proteins, two by X-ray crystallography and one by NMR, and these closely matched the hallucinated models. Thus, deep networks trained to predict native protein structures from their sequences can be inverted to design new proteins, and such networks and methods should contribute alongside traditional physics-based models to the de novo design of proteins with new functions.

# 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.

# trRosetta

PNAS 20

Protein folding

- Consider folding as classification problem
- Classify 4 features: distance, dihedral angles
- Residual network with 2D convolution, InstanceNorm, ELU, and Dropout
- Loss: categorical cross entropy
- Used Rosetta functions for 3D structure prediction

ResNet Classification

MSA Representation

## Improved protein structure prediction using predicted interresidue orientations

The prediction of interresidue contacts and distances from coevolutionary data using deep learning has considerably advanced protein structure prediction. Here, we build on these advances by developing a deep residual network for predicting interresidue orientations, in addition to distances, and a Rosetta-constrained energy-minimization protocol for rapidly and accurately generating structure models guided by these restraints. In benchmark tests on 13th Community-Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP13)- and Continuous Automated Model Evaluation (CAMEO)-derived sets, the method outperforms all previously described structure-prediction methods. Although trained entirely on native proteins, the network consistently assigns higher probability to de novo-designed proteins, identifying the key fold-determining residues and providing an independent quantitative measure of the “ideality” of a protein structure. The method promises to be useful for a broad range of protein structure prediction and design problems.

# 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.

# **Paper Map** (Last update: 5/12)

Minji Lee