# 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.

# 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.

## 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

# 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.