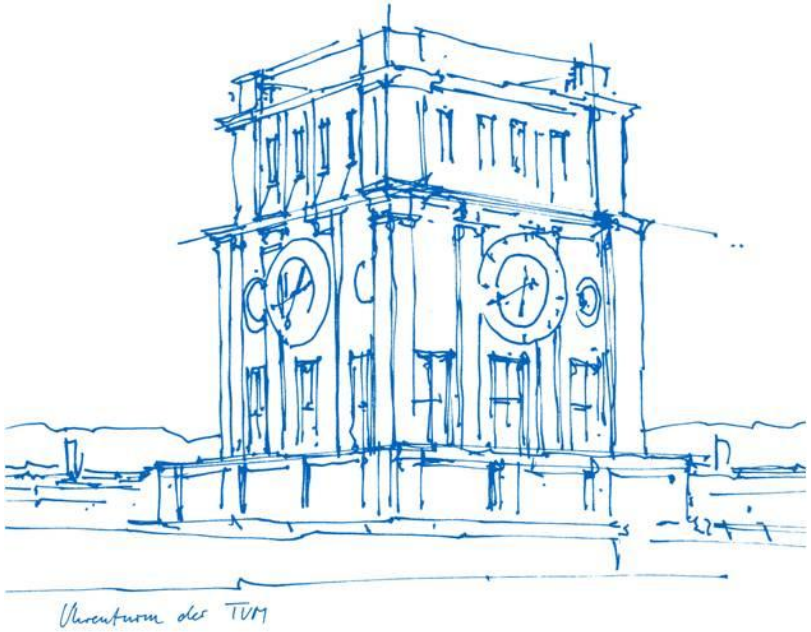
AlphaFold: Protein Structure Prediction - Distance Prediction

Technische Universität München

Faculty of Informatics

Chair of Computer Vision & Artificial Intelligence

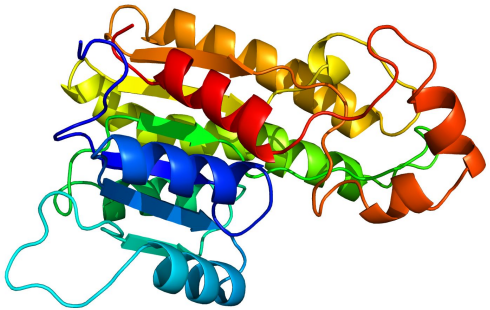
Ghalia Rehawi, Mayuran Surendran, Shubham Khatri
Supervisor: Christian Tomani
Date: 27.05.2020



Motivation

What is protein structure prediction?

- Inference of the 3D structure of a protein
  - Folding in 3D space
  - Tertiary structure: Three-dimensional atomic representation



Why do we do protein structure prediction?

- Inference the function of a protein
  - 3D structure determines the function
  - Medical fields: Drug design

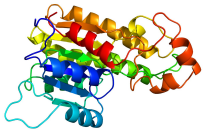
ACFRYVL →  → Function

Image source: https://en.wikipedia.org/wiki/Homology_modeling#/media/File:DHRS7B_homology_model.png

Dataset and Features

- ProteinNet : A standardized data set for machine learning of protein structure
  - Training, validation and test sets
  - Exists both in text and TFRecords format

- Features extracted from TFRecords:
  - Primary: Sequences of amino acids
  - Tertiary: Three-dimensional cartesian coordinates of each amino acid (ground truth)
  - Masks: One-bit indicators of whether the atomic coordinates for a protein residue are present.

ProteinNet: AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics* 20, 311 (2019). https://doi.org/10.1186/s12859-019-2932-0

Dataset and Features

- Features extracted from TFRecords:
  - Evolutionary: Positions Specific Scoring Matrix PSSM
    - Information about the conservation of amino acids in the protein sequence
    - Extracted from Multiple Sequence Alignment techniques (PSI-Blast)

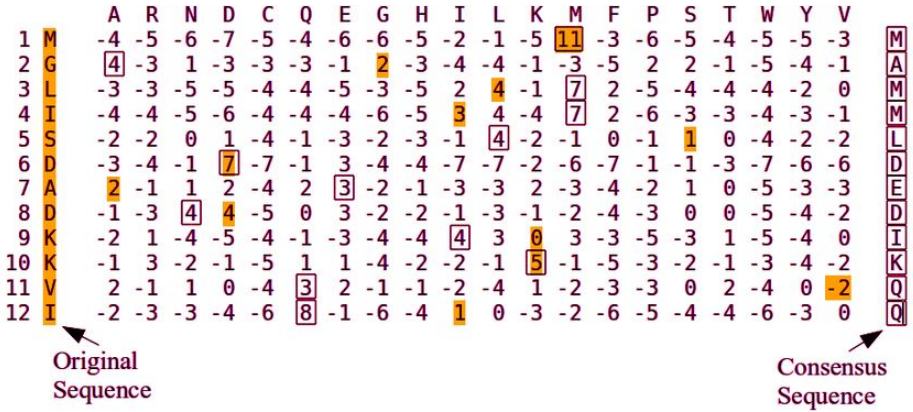
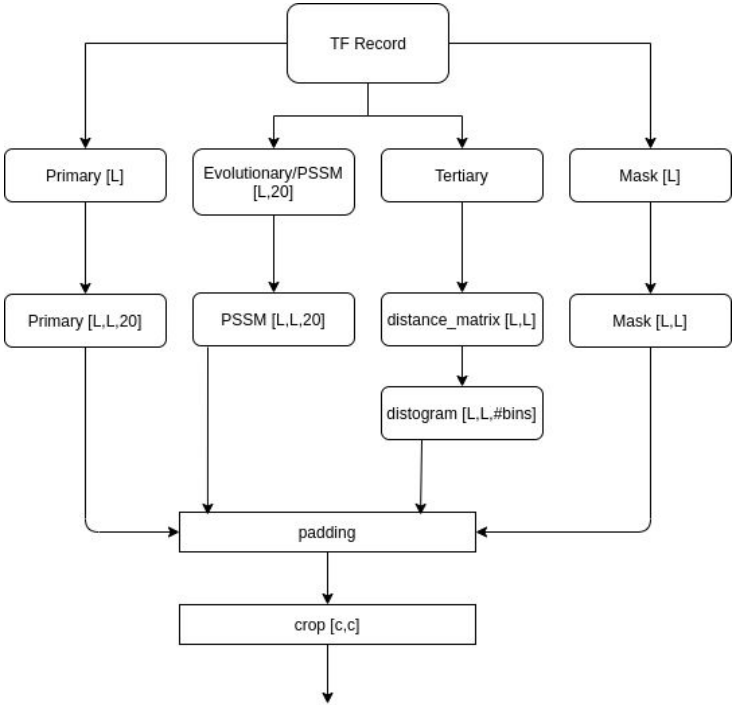
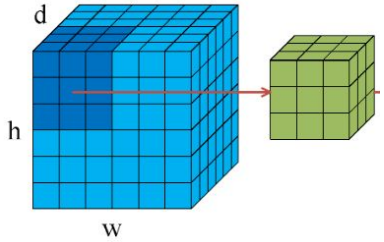


Image source: Enhanced Feature Extraction from Evolutionary Profiles for Protein Fold Recognition - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Extracting-the-consensus-sequence-from-a-PSSM-file-This-figure-shows-an-excerpt-from-a_fig1_314080284 [accessed 26 May, 2020]

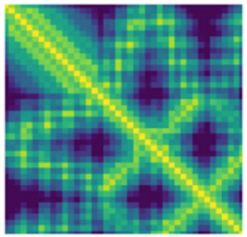
Feature Preprocessing



Model input is L x L x #Features



Model output is a discretized distance map (distogram)



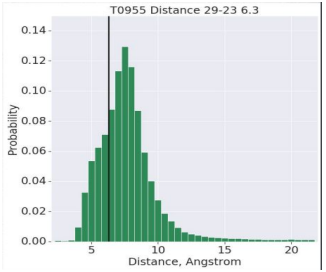
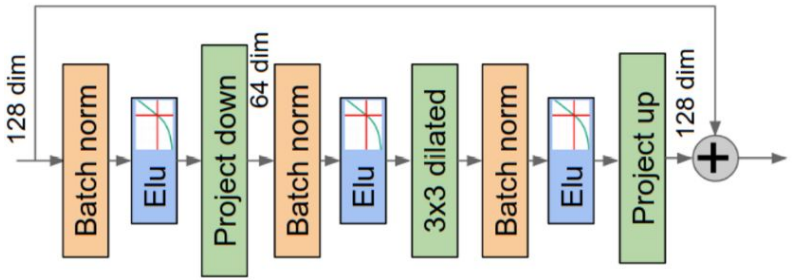


Image source: Dong, Hongwei, and Lamei Zhang. "PolSAR Image Classification with Lightweight 3D Convolutional Networks." *Remote Sensing* 12.3 (2020): 396.

Model Architecture

- Architecture is a deep two-dimensional dilated convolutional network with a variable number of residual block groups (a group always includes four residual blocks)



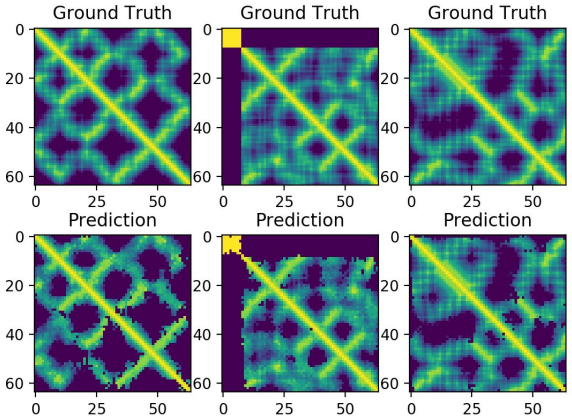
- Successive residual blocks in a group cycle through dilations of 1, 2, 4, 8 pixels to allow propagation of information quickly across the cropped region

- **AlphaFold**: 7 groups of 4 blocks (256 channels) + 48 groups of 4 blocks (128 channels)

Image source: Senior, A.W., Evans, R., Jumper, J. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706–710 (2020). https://doi.org/10.1038/s41586-019-1923-7

Experiments and Approaches

- Overfit the model on small number of samples and see if it can learn to reproduce the samples



- Tried different approaches to include masking into the loss function such as passing mask tensor as sample weight to categorical cross entropy loss function provided by TF Keras or implementing custom loss function

Current Challenges

- How to include mask matrix into the Keras categorical cross entropy function?
- How to access the shapes of tensors with variable sizes inside the mapping function in the data pipeline?

Next Steps

- Incorporate mask tensors in training
- Improve model architecture and scale up training of model using the implemented data loader
- Include PSSM features generated from Multiple Sequence Alignment

Thanks for listening
Questions!