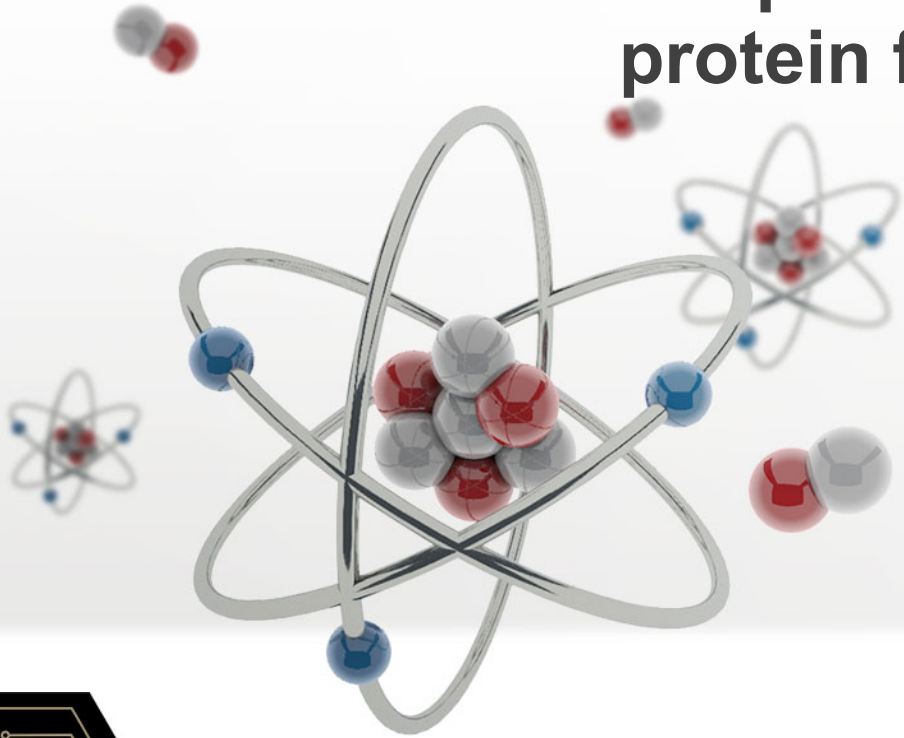
New deep learning architectures for protein function prediction and design

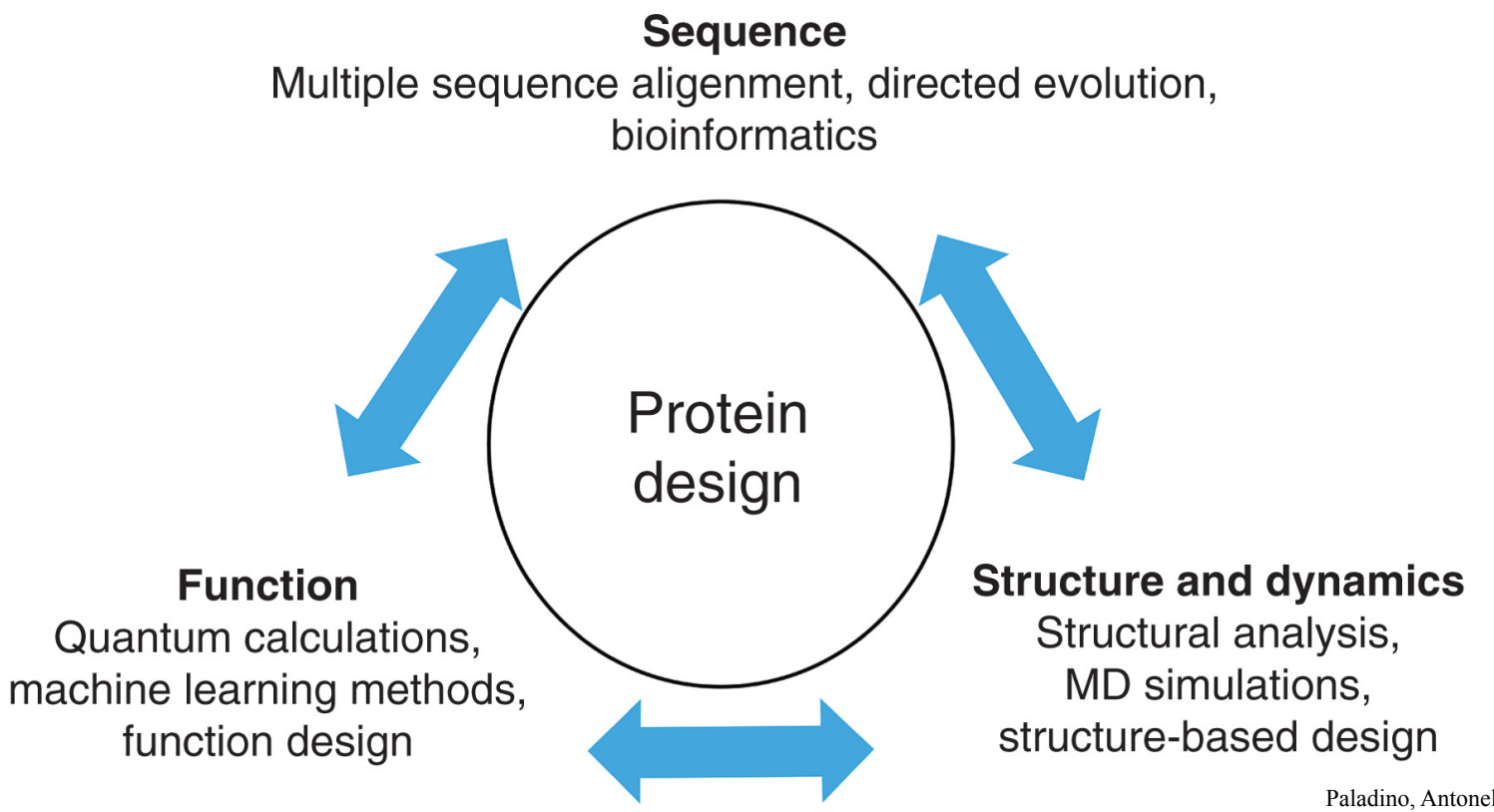


Xinchun Ran

Advisor: Prof. Zhongyue (John) Yang
Department of Chemistry
Vanderbilt University

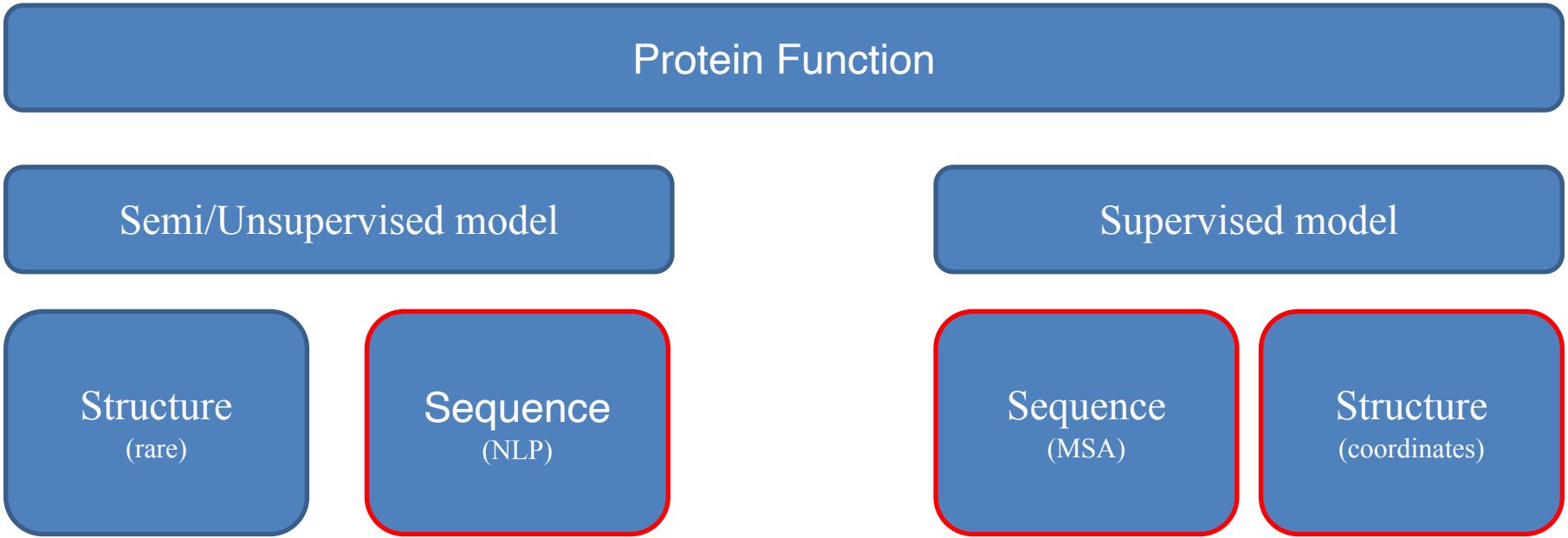
ZJYANG
GRP Vanderbilt

Introduction

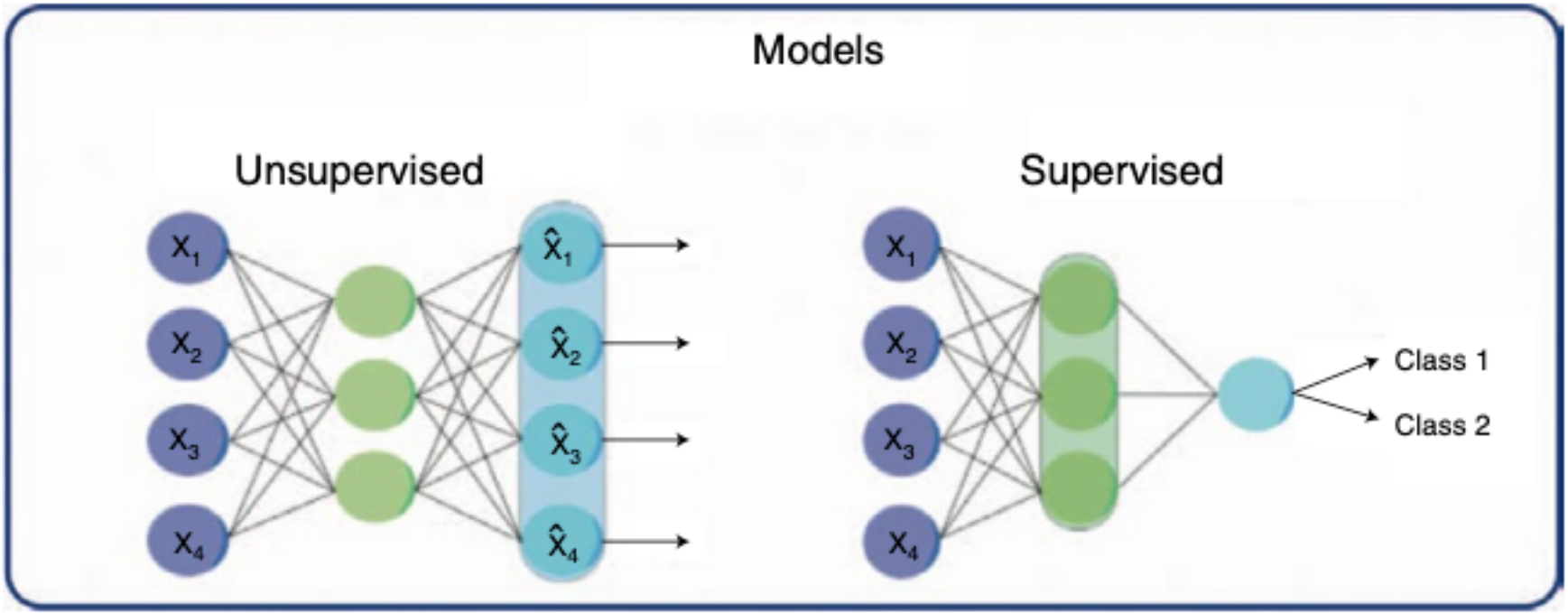


Paladino, Antonella, et al *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7.5 (2017): e1318.

Overview of Protein function



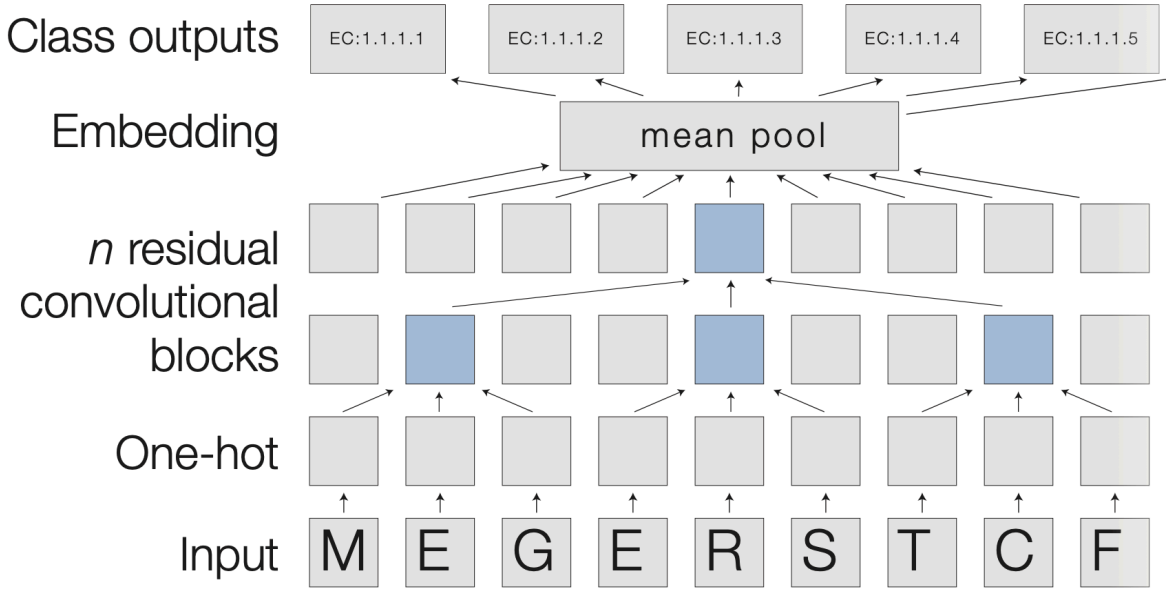
Supervised vs Unsupervised



Unsal, Serbulent, et al. "Learning functional properties of proteins with language models." *Nature Machine Intelligence* 4.3 (2022): 227-245.

Proeinfer:Semisupervised Sequence from Google

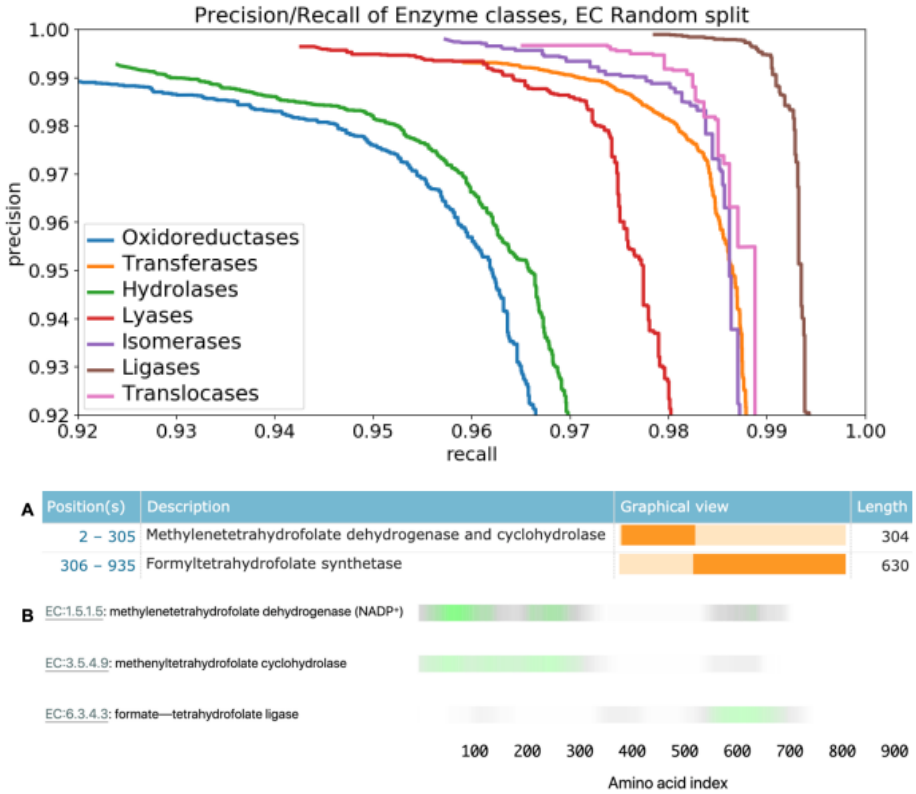
- Input: one-hot encoding
- Output: EC number
- Architecture: encoder + classifier
- The classifier trained with 20000 EC number labeled
- Optimizer: Adam
- Loss: cross-entropy



Sanderson, Theo, et al. "ProteInfer: deep networks for protein functional inference." *bioRxiv* (2021).

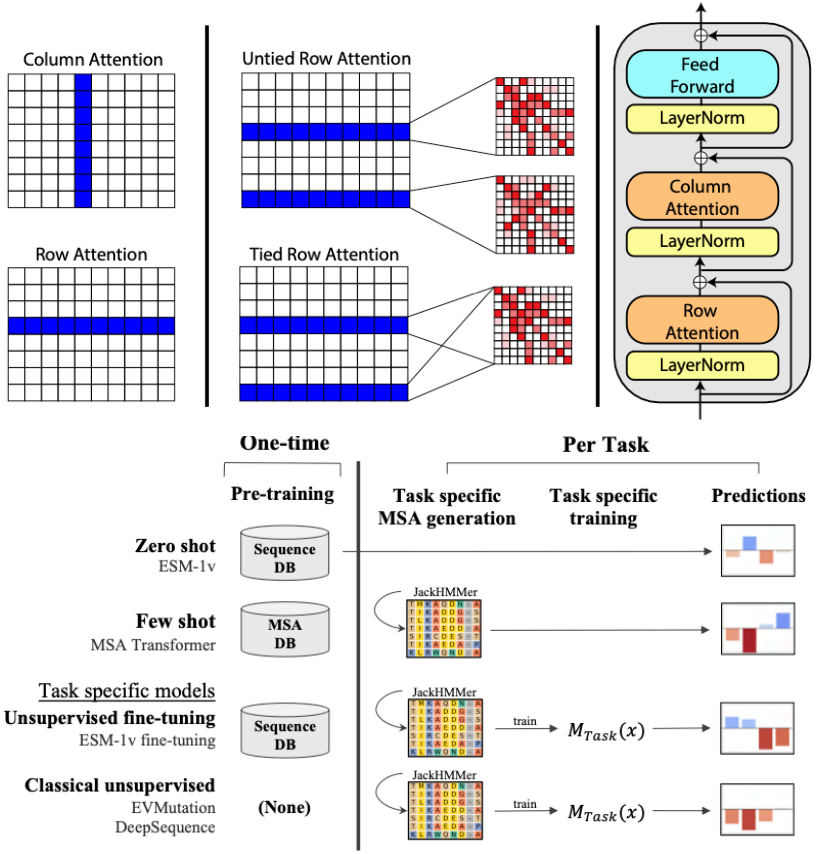
Proeinfer: Semisupervised Sequence Continued

- The Precision for the EC classification is 99%
- The model was trained On 8 Telsa-GPUs
- The model has displayed with interactive
- The Proteinfer use few layers of the convolution neural network



Unsupervised sequence from Meta

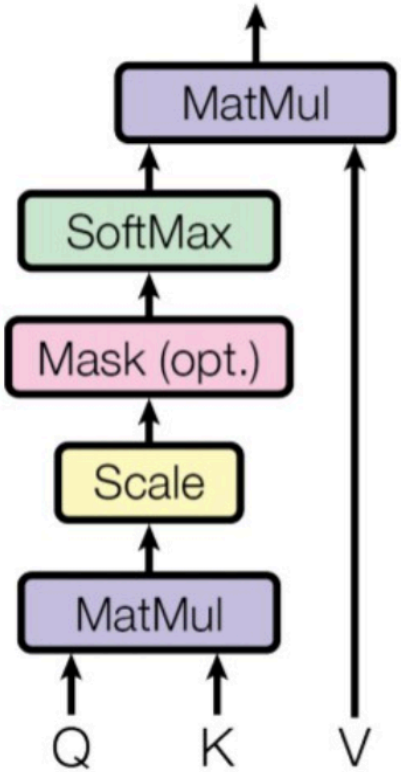
- ESMs (**Evolutionary Scale Modeling**) are series SOTA models
- Tasks: Stability, fluorescence
- Novel architecture: row attention; column attention
- Input: sequence
- Downstream task: DMS data



Meier, Joshua, et al. "Language models enable zero-shot prediction of the effects of mutations on protein function." *Advances in Neural Information Processing Systems* 34 (2021): 29287-29303.

Unsupervised sequence from Meta Continued

- Training: sequences dataset (250 million), 34 layers transformers
- Params: The params for the EMS-1b has 669 MB
- Time scale: 128 A100 GPU 4 weeks
- Derivatives: ESM1v, ESM-1b, MSA-transformer, inverse folding
  ESM2



Meier, Joshua, et al. "Language models enable zero-shot prediction of the effects of mutations on protein function." *Advances in Neural Information Processing Systems* 34 (2021): 29287-29303.

Recap

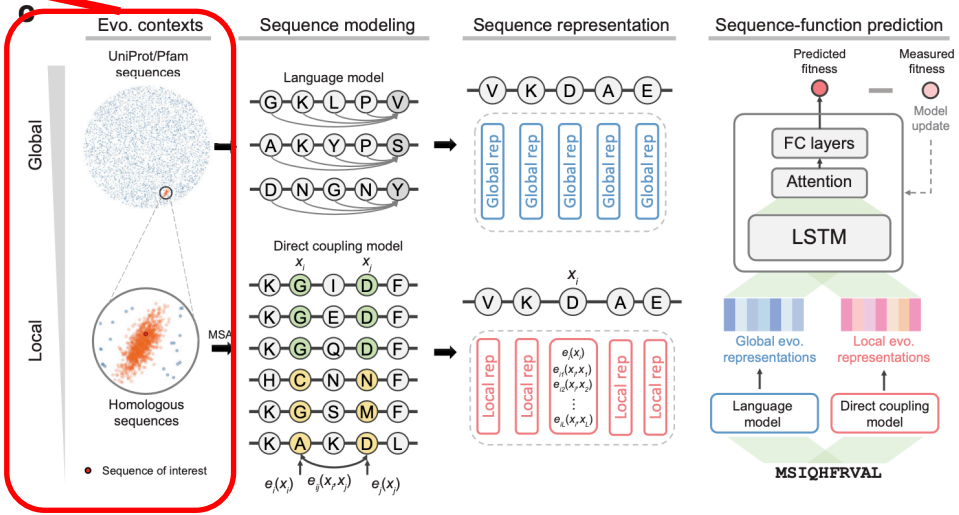
- Unsupervised language model: Transformer encoder/decoder-based model
- Input: sequences. treating protein as natural language
- Language vs Stat model: Large params, global information across the protein families.
- Promising solution for single sequence prediction

Task	Unsupervised contact prediction		
Test set	Large valid	CASP13-FM	CAMEO
Gremlin (Potts)	39.3	16.9	24.0
UniRep			
SeqVec			
TAPE	11.2	5.5	6.8
ProtBert-BFD	34.1	13.5	23.9
Prot-T5-XL-BFD	35.6	16.5	25.9
ESM-1	33.7	13.6	21.4

Meier, Joshua, et al. "Language models enable zero-shot prediction of the effects of mutations on protein function." *Advances in Neural Information Processing Systems* 34 (2021): 29287-29303.

Unsupervised model + supervised statistic model

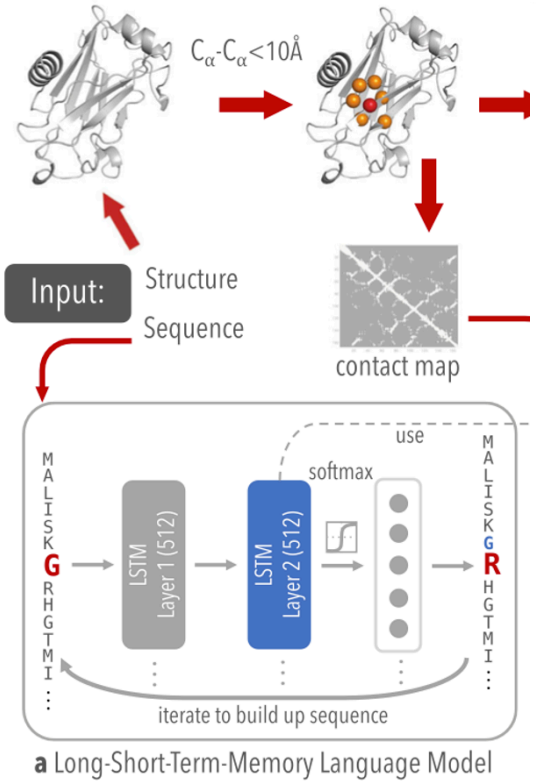
- Sequences embedding UniRep
- Evolutionary coupling CCMPred
- Local + Global sequence information
- Architecture: (DCA, Unirep) -> LSTM -> fitness regressor
- The mutation dataset with 20000 double mutation
- Mutation fitness: Spearnman 0.75



Luo, Yunan, et al. "ECNet is an evolutionary context-integrated deep learning framework for protein engineering." *Nature communications* 12.1 (2021): 1-14.

DeepFRI – Semisupervised learning Encoder

- Sequences embedding 512 layers LSTM
- Generate Contact map features from PDB structs
- Architecture: sequence LSTM embedding
- 10 million unlabeled sequence

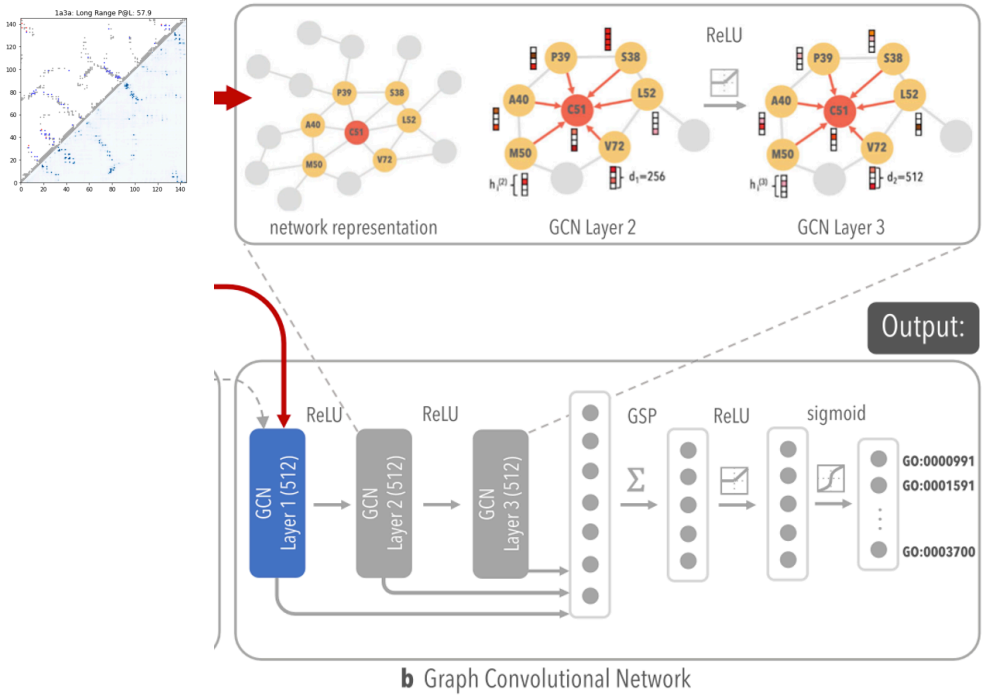


Gligorijević, Vladimir, et al. "Structure-based protein function prediction using graph convolutional networks." *Nature communications* 12.1 (2021): 1-14.

DeepFRI – Supervised learning part Predictor

- Architecture: sequence LSTM embedding -> contact map (GCN) -> classifier

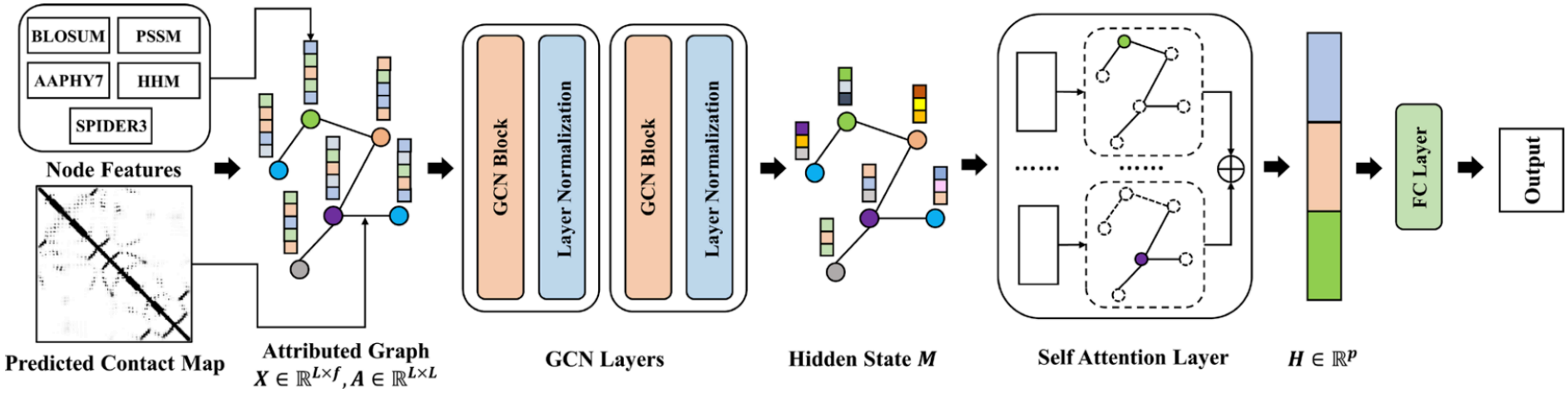
- Train: 700 PDB structs with GO labeled



First structure-based function prediction with convolution graph neural network

Gligorijević, Vladimir, et al. "Structure-based protein function prediction using graph convolutional networks." *Nature communications* 12.1 (2021): 1-14.

Supervised Structure based solubility prediction

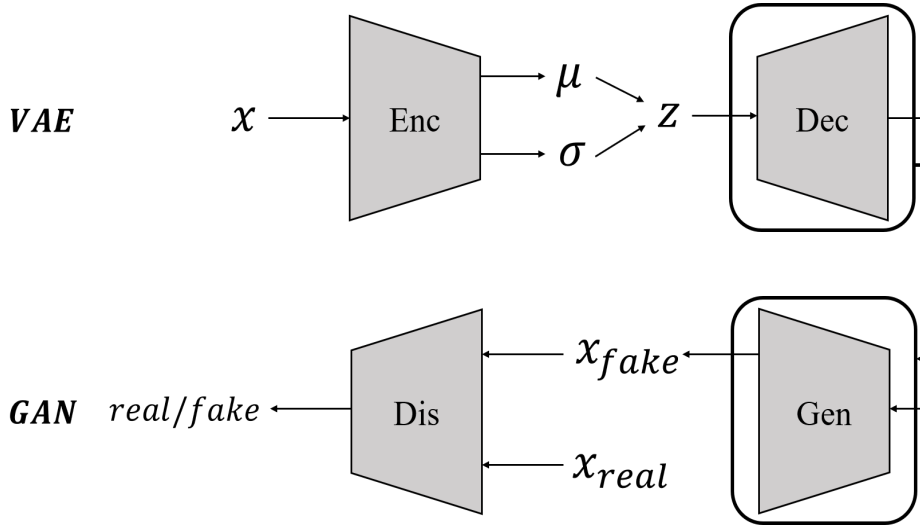


- 2 layers of LSTM as contact encoder, 3 layer of MLP as predictor to solubility score
- Train on 685 proteins; Accuracy: 0.78

Chen, Jianwen, et al. "Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map." Journal of cheminformatics 13.1 (2021): 1-10.

Introduction: Protein design

- Similar: Significant region protein contribute to function.
- Different: protein design try to search as much of combination of AA as possible
- Generally, Design model like to use KL divergence (Entropy) to measure Design NN model
- Encoder part using in design & function prediction can be the same



Overview: Protein function prediction & Design

Protein design

Generative model

Discriminative model

Structure
(coordinates, contact map)

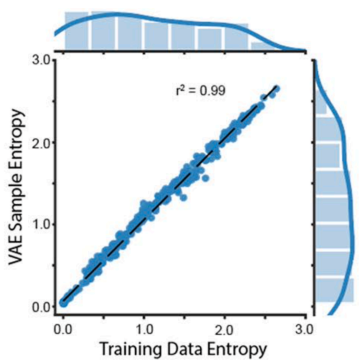
Sequence
(NLP)

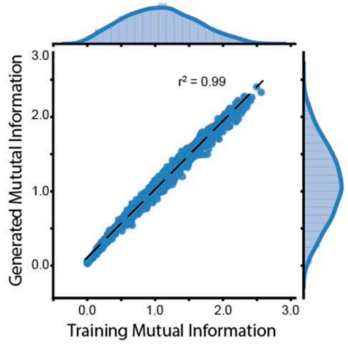
Sequence
(Sequences Analysis)

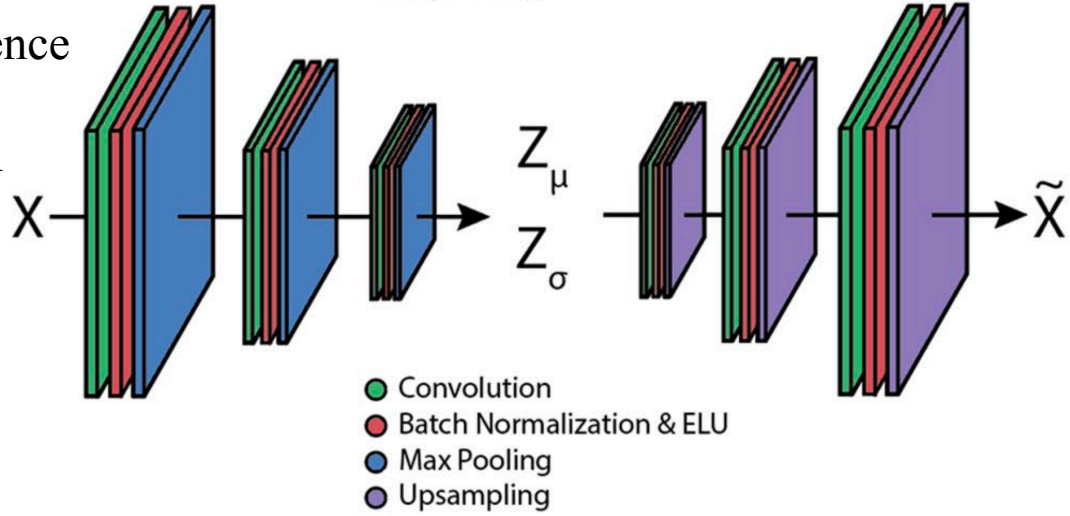
Structure
(binding pocket, property)

VAE -- Protein design Generative model

- VAE library for specific enzyme sequence space
- Architecture: Encoder + decoder both [3 layers CNN]
- Input: rSASA, one-hot encoding sequence
- ddG from Rosetta, DSSP as evaluation
- Data: 10^7 variant training data
- output: 87 Kcat/Km enhanced mutant



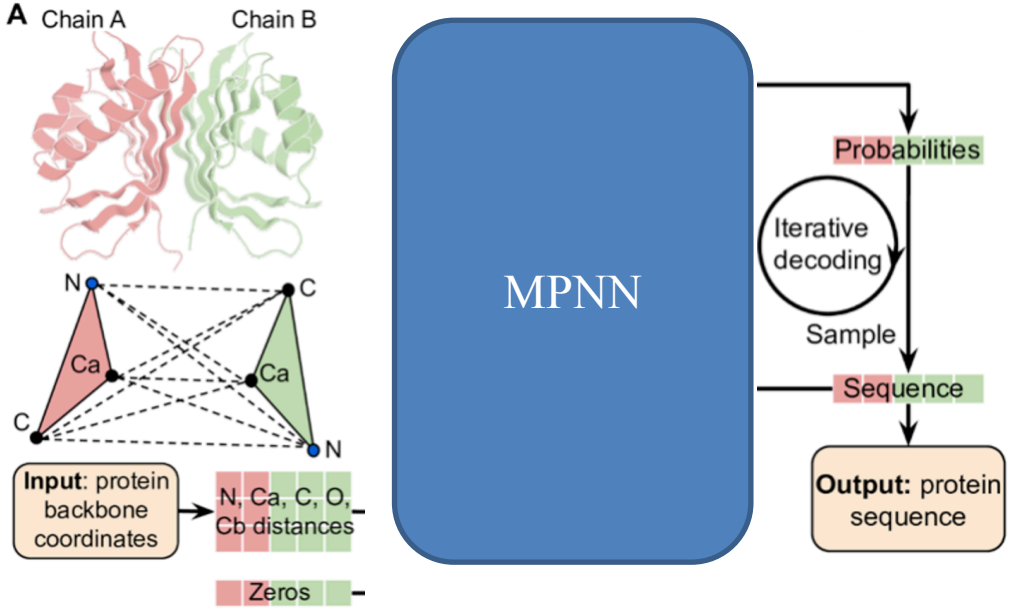




Giessel, Andrew, et al. "Therapeutic enzyme engineering using a generative neural network." *Scientific reports* 12.1 (2022): 1-17.

ProteinMPNN – Discriminative structure-based model

- Structural embedding with MPNN
- 3 layers of MPNN structures encoders ->
  2 layers of MPNN sequence/site decoders
- Input: Distance map
- 25136 unlabeled structures



Sequence Recovery Rate: 52.4%

Dauparas, Justas, et al. "Robust deep learning based protein sequence design using ProteinMPNN." *bioRxiv* (2022).

Recap

- Transformer and CNN are quite prevalent in sequence embedding.
- Input: sequences & structure coordinates
- Unsupervised learning in trend in protein function prediction and design
- Sequence embedding is important part in function prediction tasks

Thank you

I wish to thank Selina Wang who provide ideas in improving slides