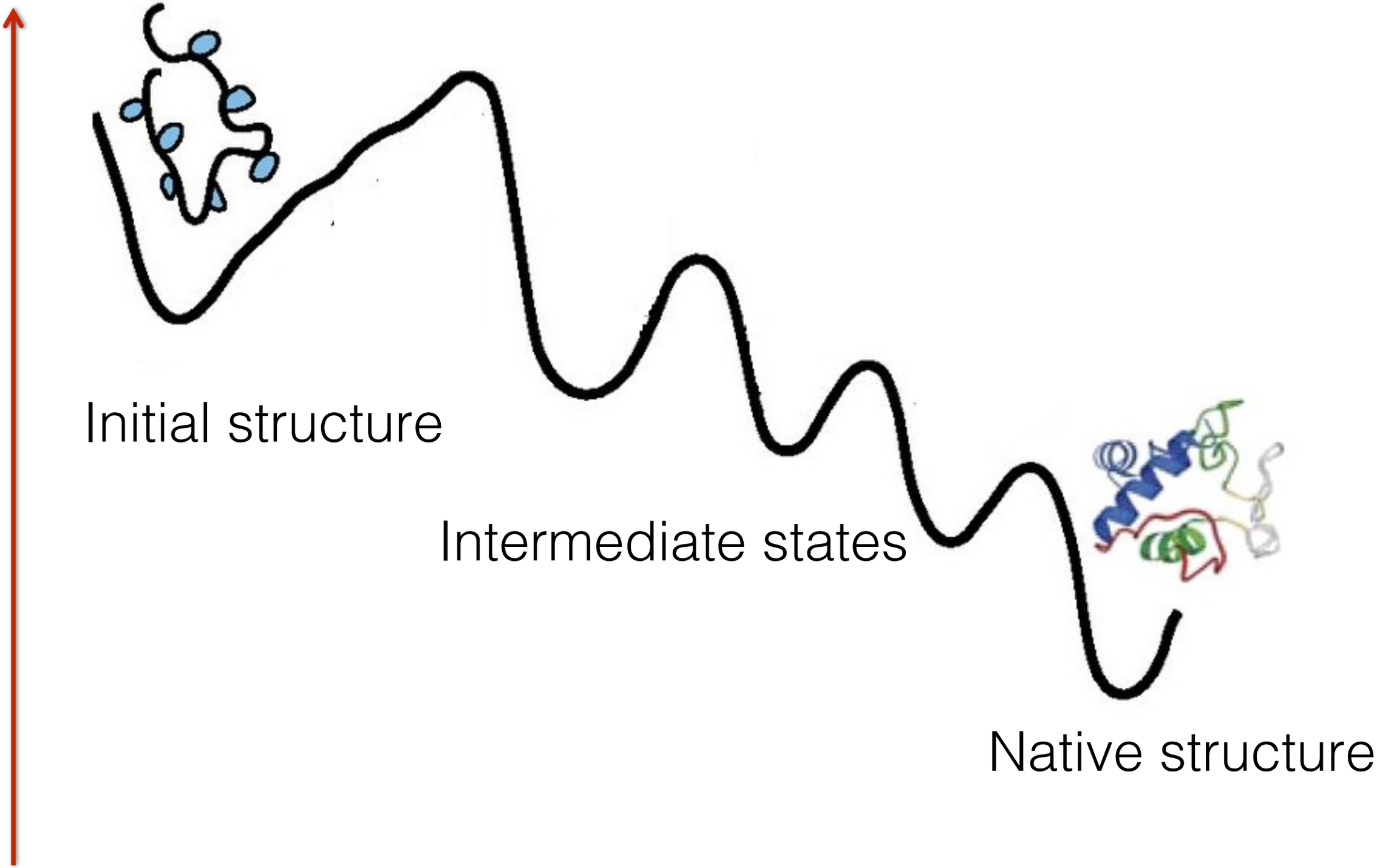
Machine learning algorithms for structural and functional genomics

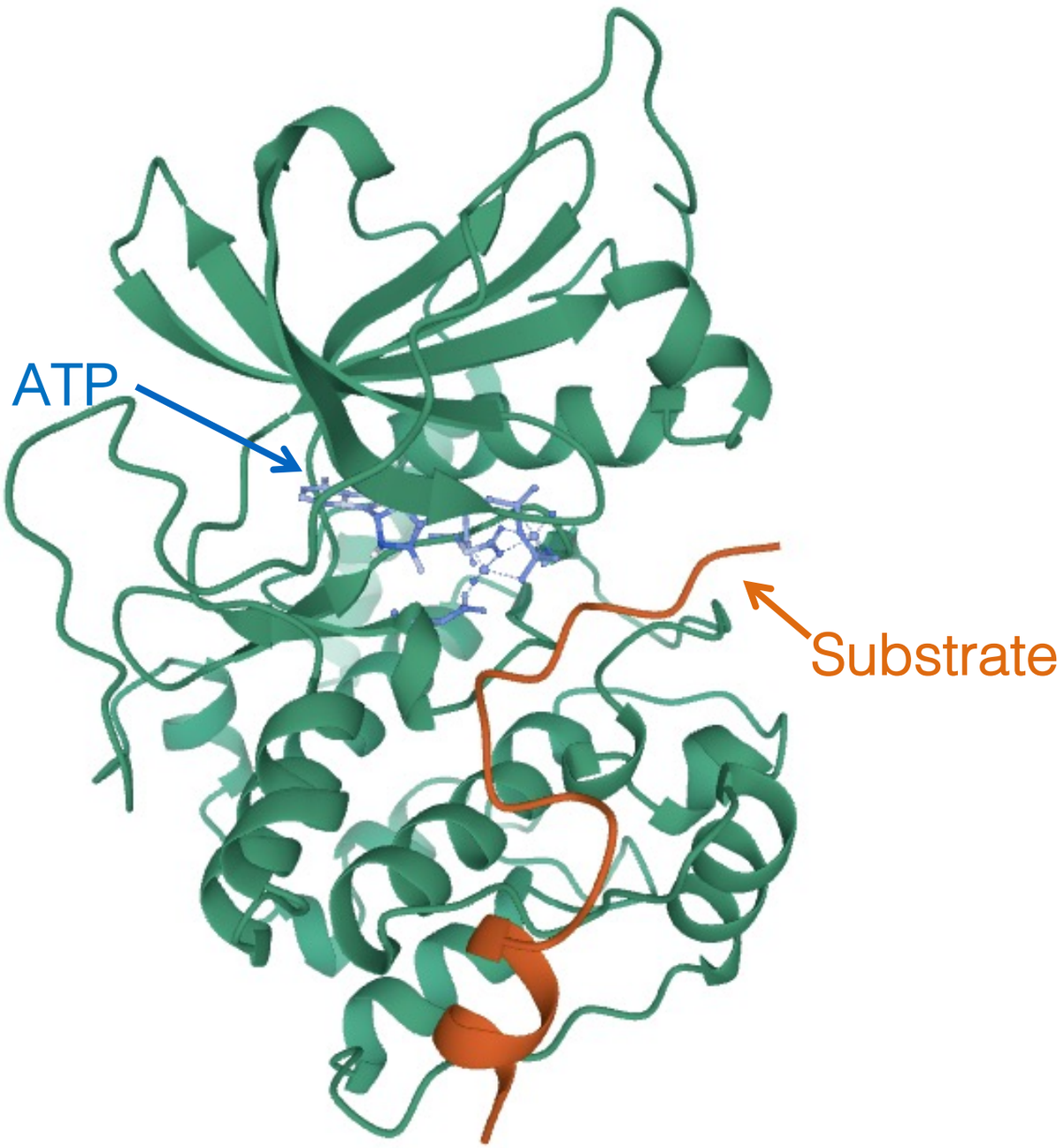
Jian Peng
Department of Computer Science
College of Medicine, Institute of Genomic Biology
University of Illinois at Urbana-Champaign

The protein folding problem

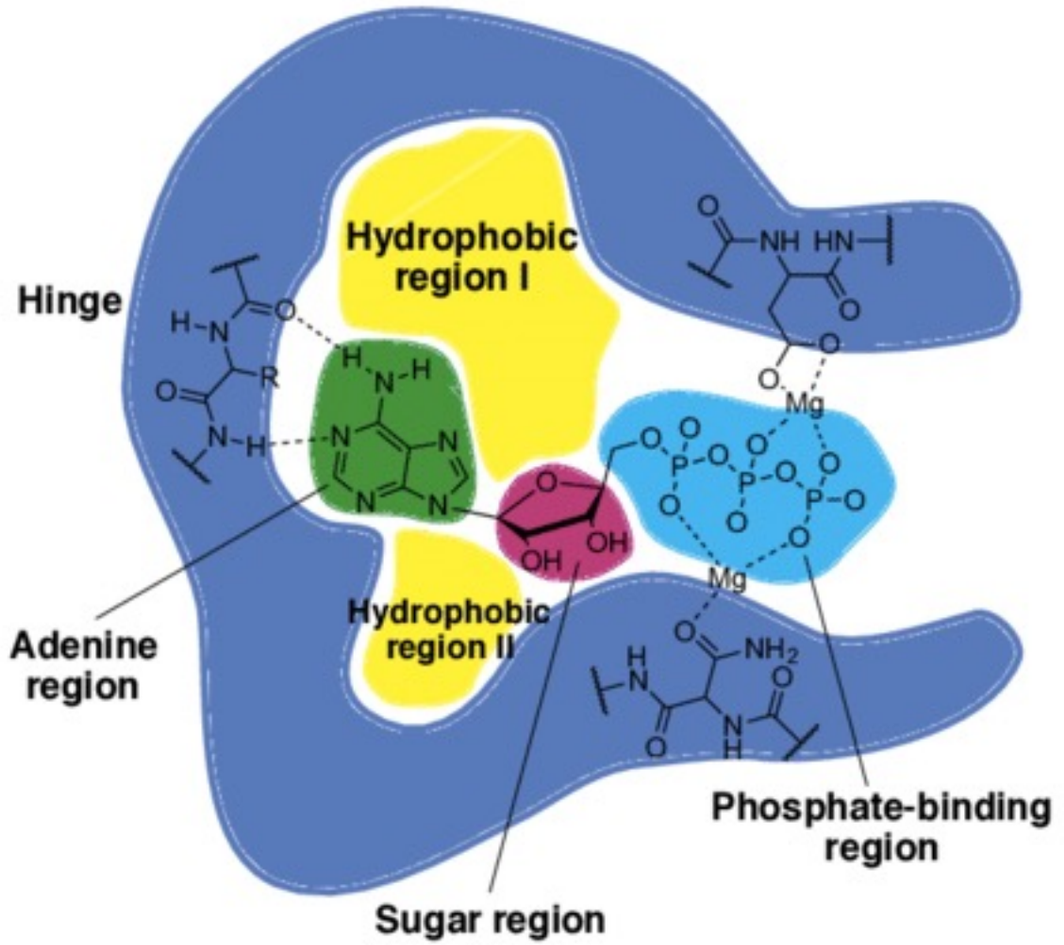


Englander, Mayne, PNAS, 2014

Structure provides insights on function



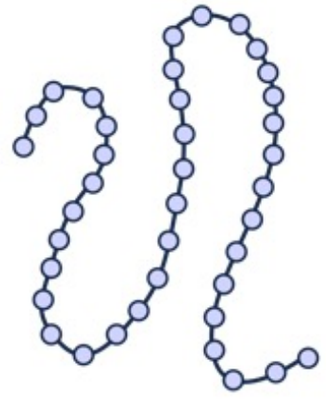
Protein Kinase

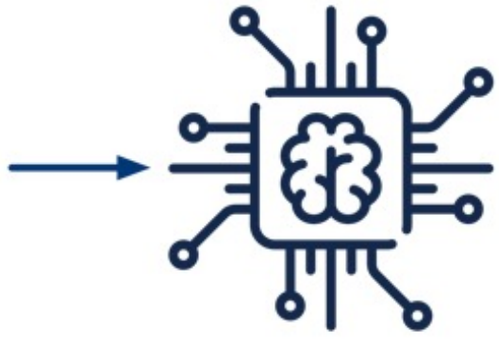


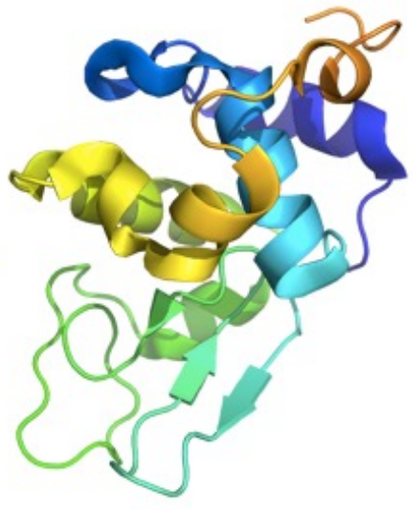
Images from rcsb.org

Protein structure prediction

Amino acid sequence







3D structure

Successful prediction algorithms

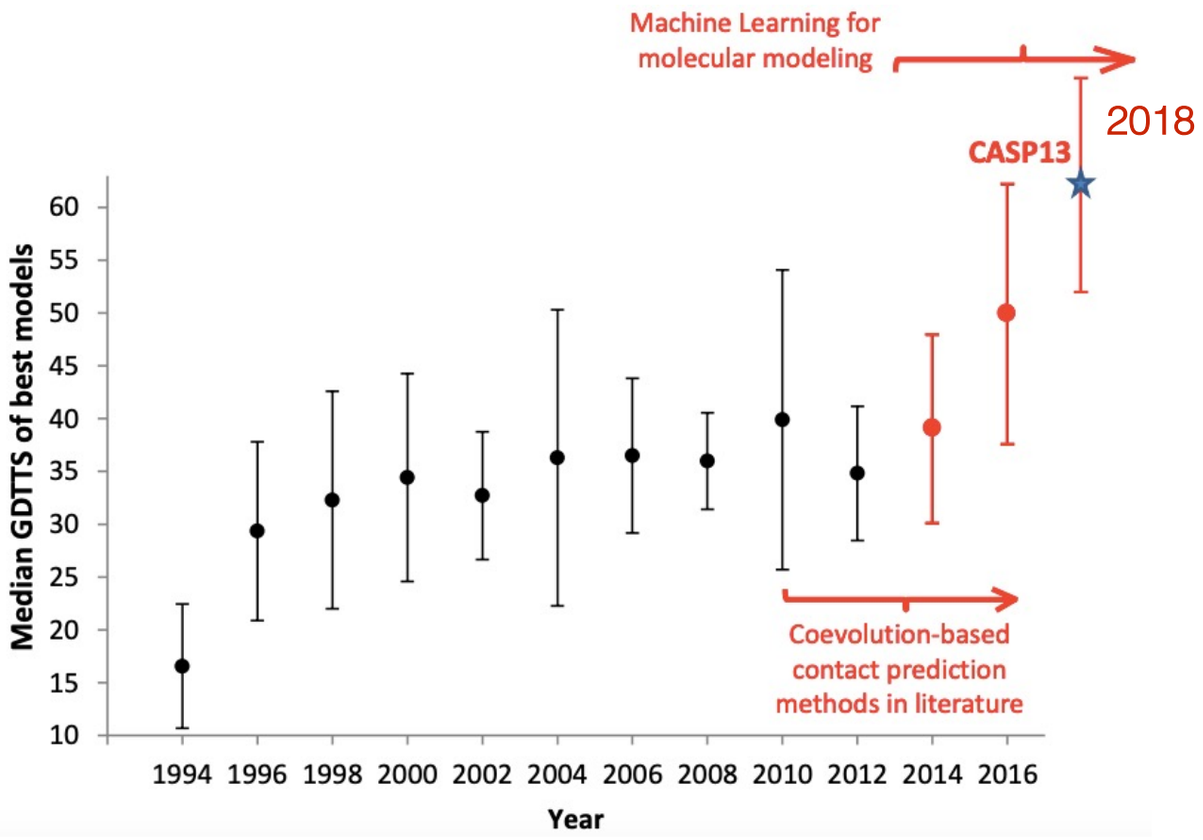
Rosetta

RaptorX

I-TASSER
Protein Structure & Function Predictions

ALPHAFOLD

Progress in recent CASPs

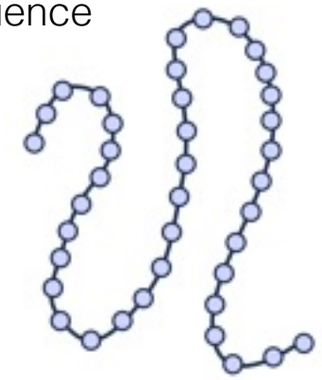


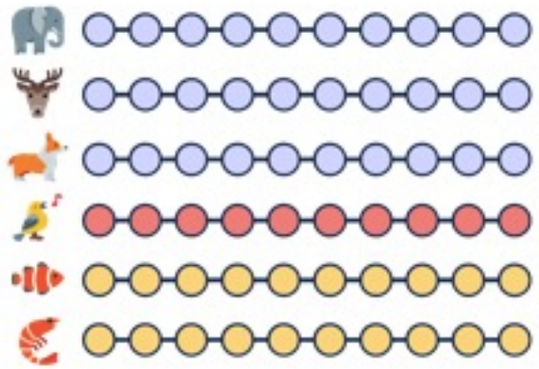
From CASP13 evaluation by Abriata and Dal Peraro

Structure prediction

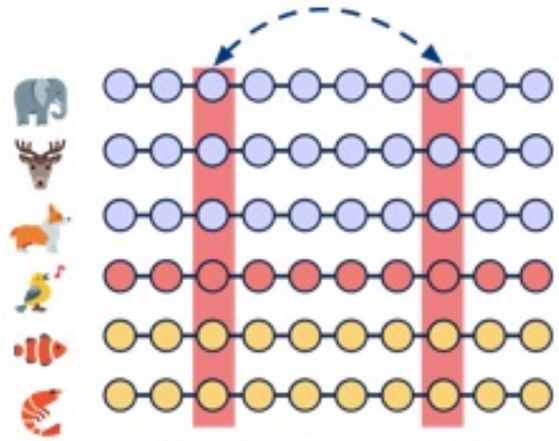
Current status: Protein structure prediction

sequence



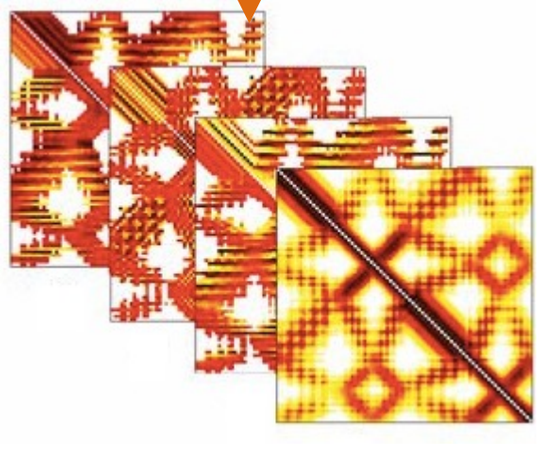


Alignment of homologs



Co-evolutionary analysis

Deep learning models



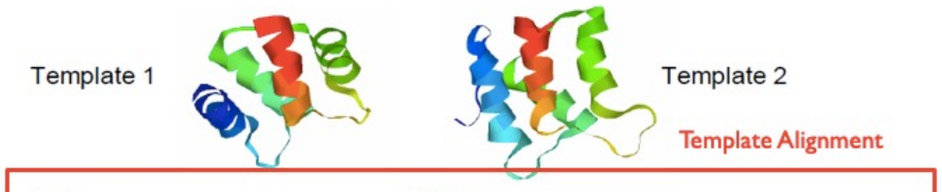
Residue-residue constraints

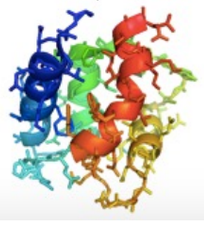
Backbone Optimization

Template-based modeling

PDB

Identify templates

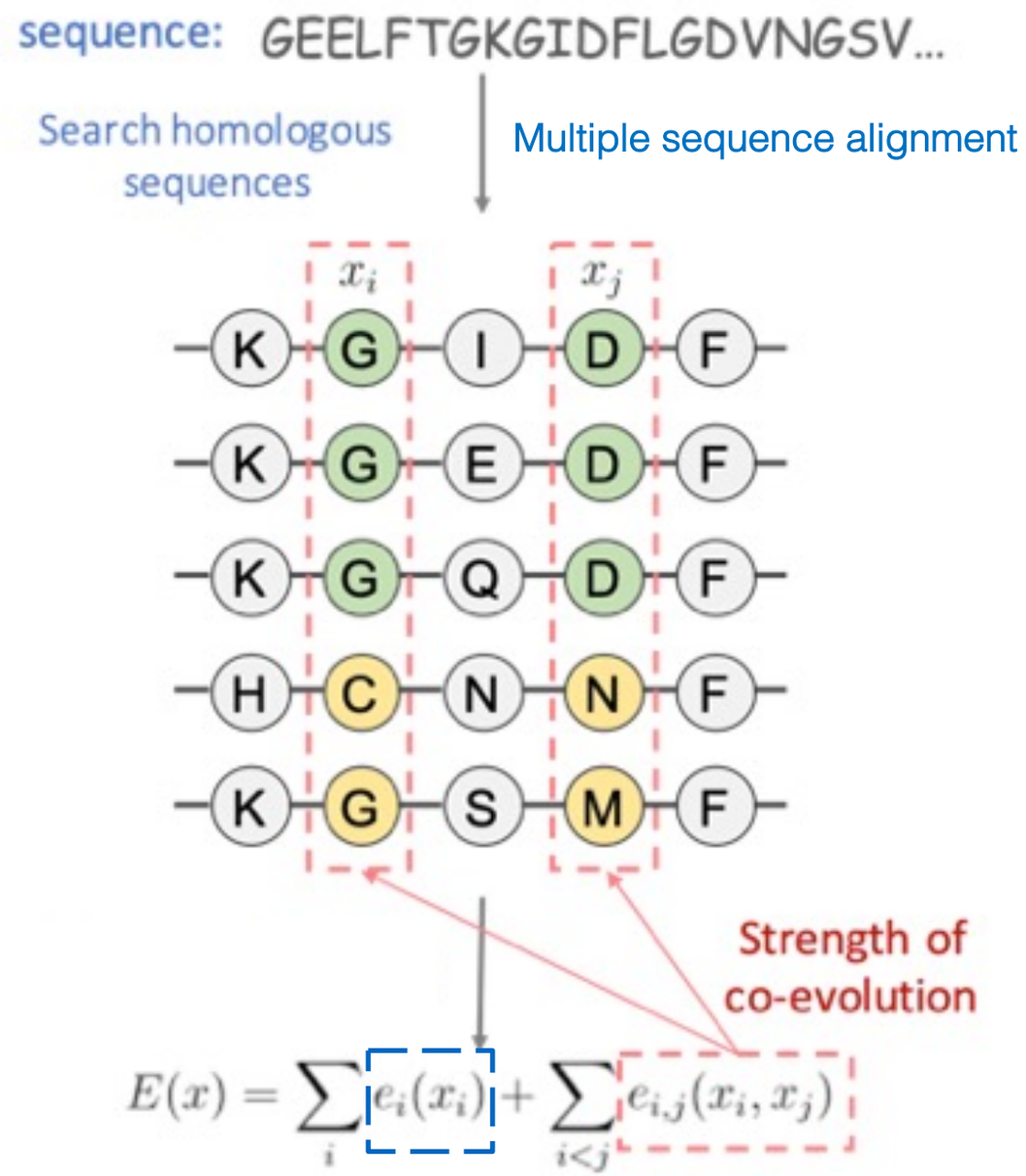


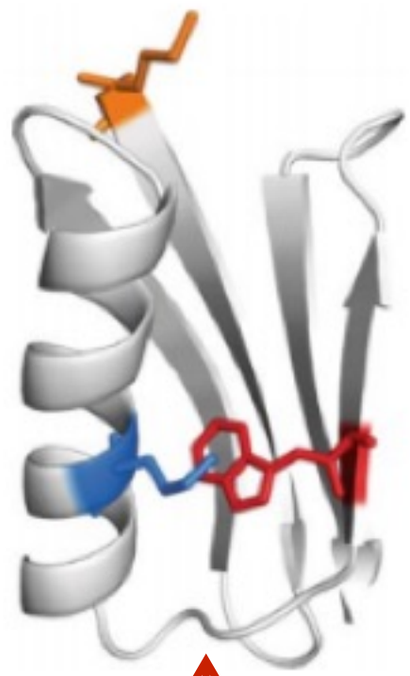


Side-chain construction

Contact-assisted modeling

Exploiting co-evolution for contact prediction





Residue contact

Evolutionary and structural constraints

Learning couplings from protein alignment

Capture independent sites

$$P(\mathbf{x}) = p_1(x_1)p_2(x_2)\cdots p_L(x_L)$$

or equivalently

$$P(\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_i e_i(x_i)\right)$$

Single potentials

Local preference

Amino acid i

x_i

Multiple sequence alignment

```
AQKLYLTHIDAEVDGD
ADTLYMTKIHHQFQGD
ADRLFITEVKQVFEGD
ADRLYMTKIHHTFEGD
ADKLYCTLIHNSFDGD
ADRLYMTKIHHEFEGD
ADTLYLTMIHQKFQAD
TDTLYITHIDETFQGD
ADTLYLTQIRNKFQGD
TSRMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD
```

Learning couplings from protein alignment

Capture pairwise interactions

Multiple sequence alignment

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_i e_i(x_i) + \sum_{i \neq j} e_{ij}(x_i, x_j)\right)$$

Single potentials

Local preference

Pairwise potentials

Co-evolution strength

Markov random field
Ising (Potts) model
Undirected graphical model

Learning with Markov Random Fields

$$L(e) = \prod_{n=1}^{N} \frac{1}{Z_e^{(n)}} \prod_{i}^{L} \exp[e_i(x_i^n) + \sum_{j \neq i} e_{i,j}(x_i^n, x_j^n)]$$

Partition function

Singleton potentials

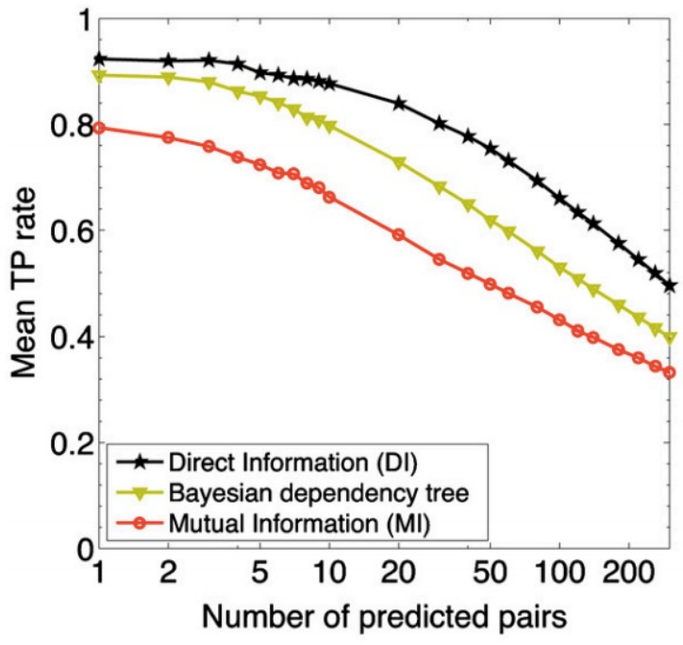
Pairwise potentials

Local AA preference

Pairwise AA couplings

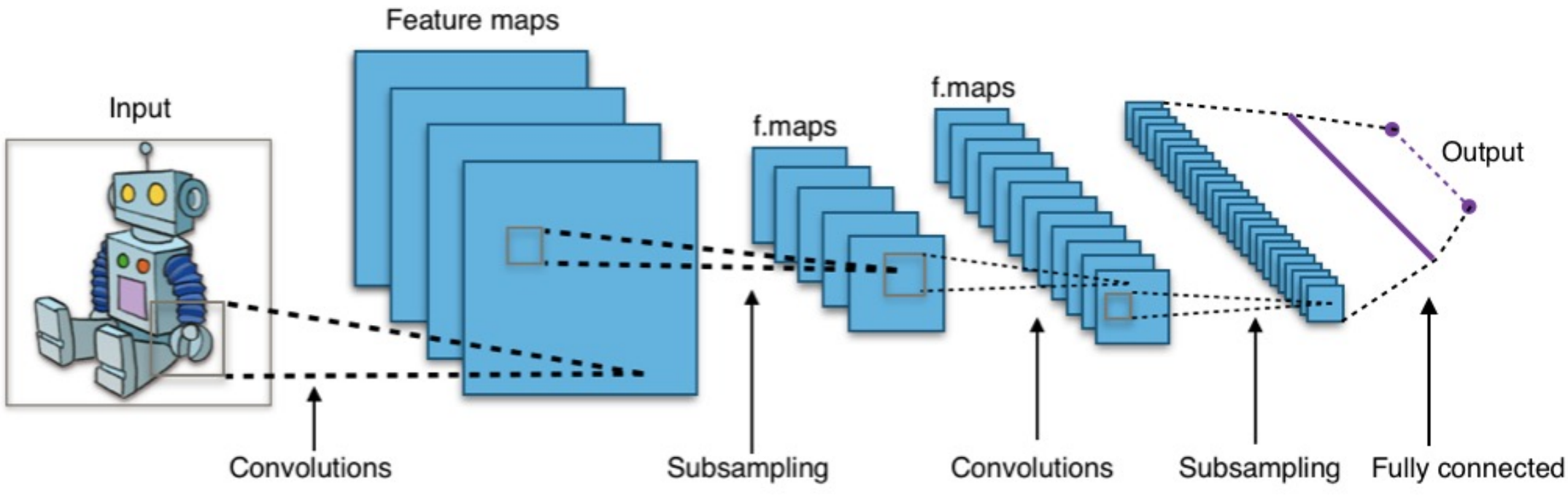
Learning algorithms:

- Mean fields approximation: EVFold, DirectInfo
- Gaussian approximation: PSICOV
- Pseudolikelihood: GREMLIN, CCMpred



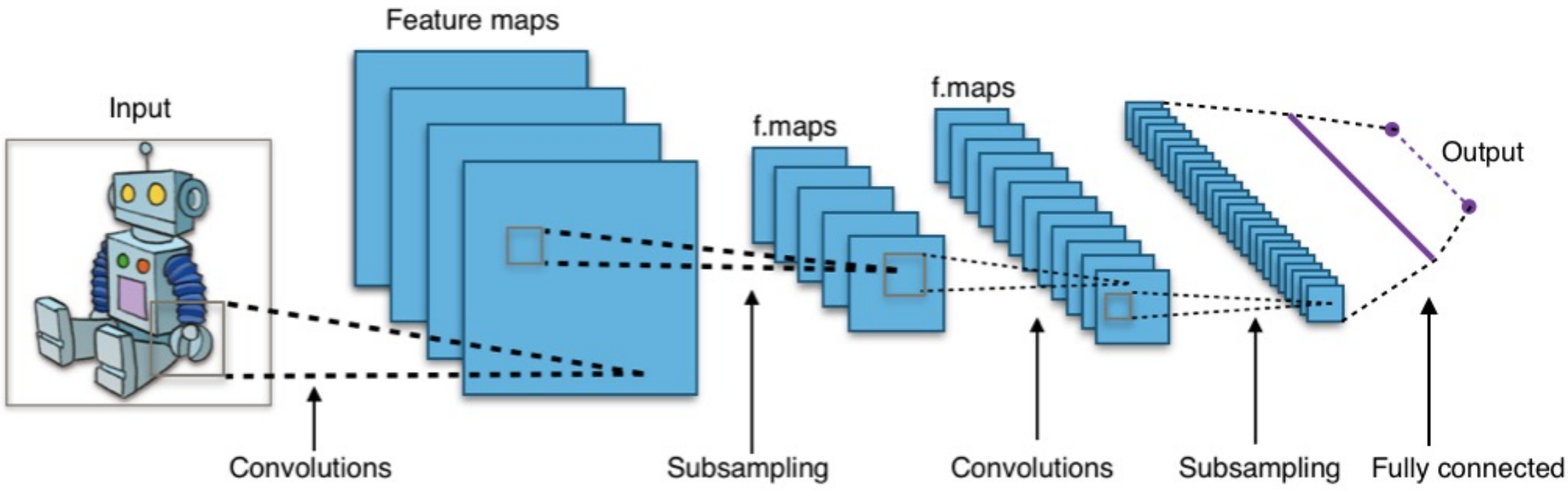
Balakrishnan, et al. Proteins 2008; Marks, et al. Plos one 2011; Morcos, et al PNAS, 2011; Jones, et al. Bioinformatics, 2011. Seemayer, et al. Bioinformatics 2014

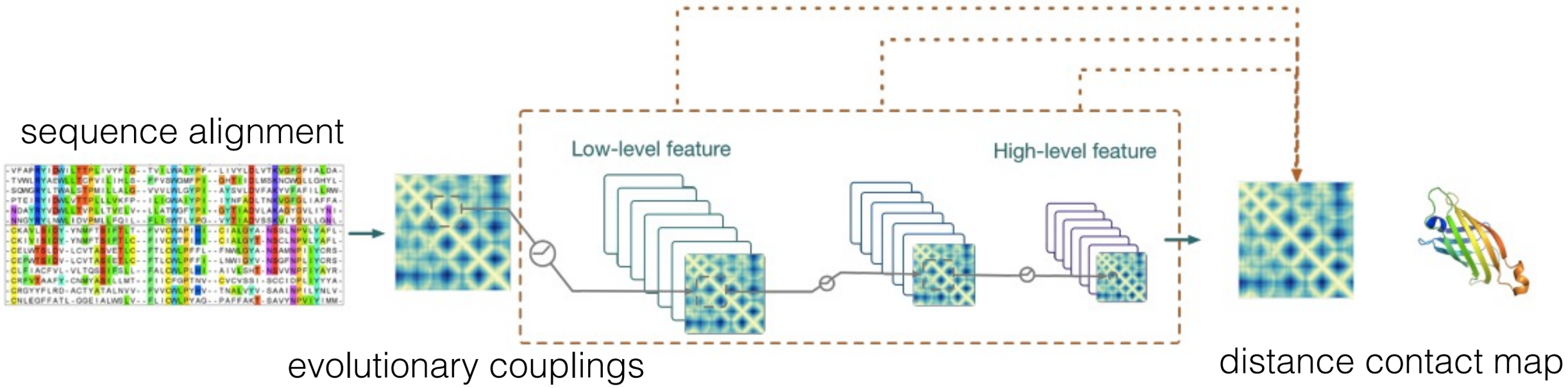
Deep convolutional NNs recognize image patterns



Lecun, Bengio, Hinton. Nature 2015

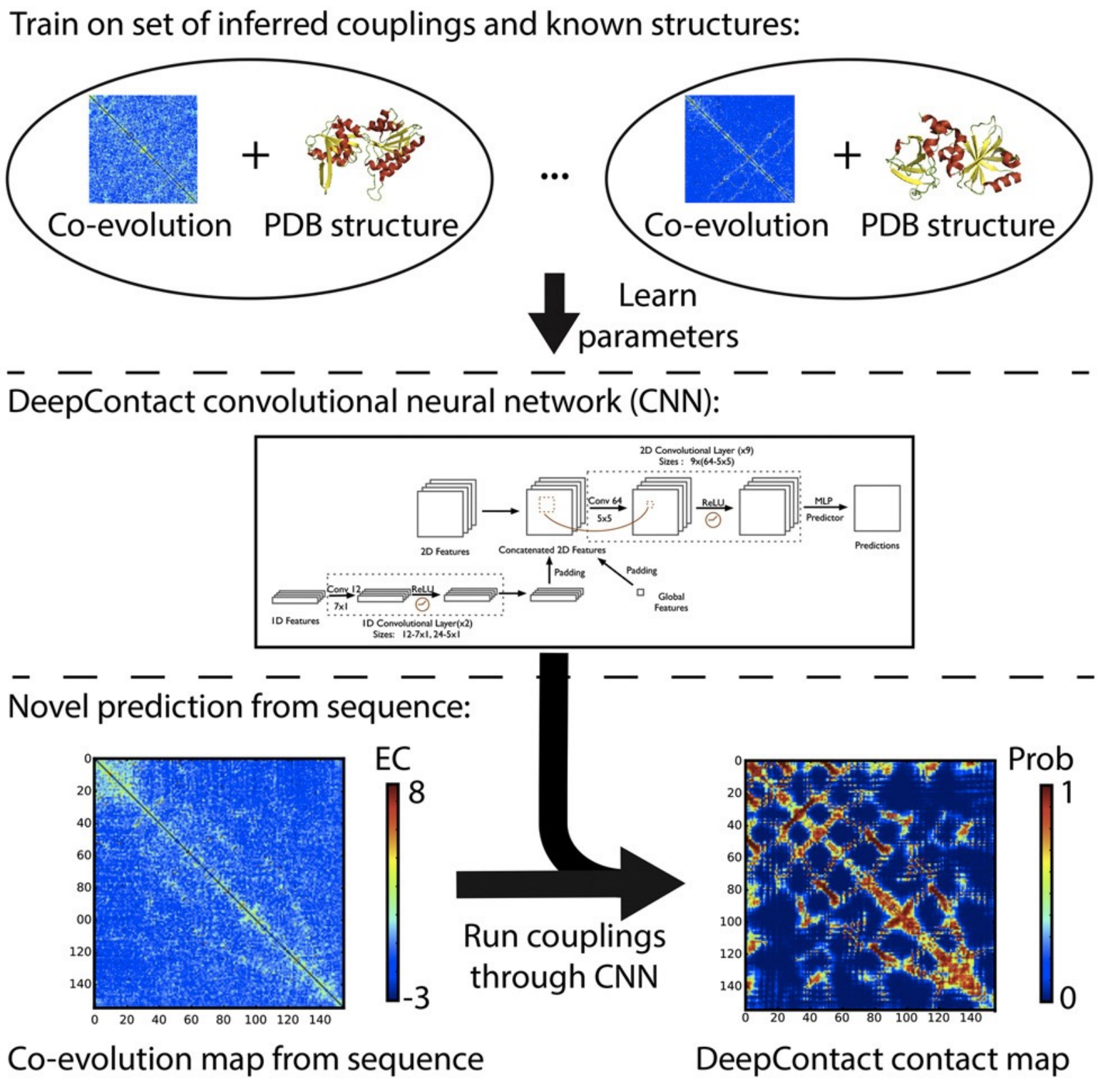
Deep convolutional NNs recognize coevolutionary patterns





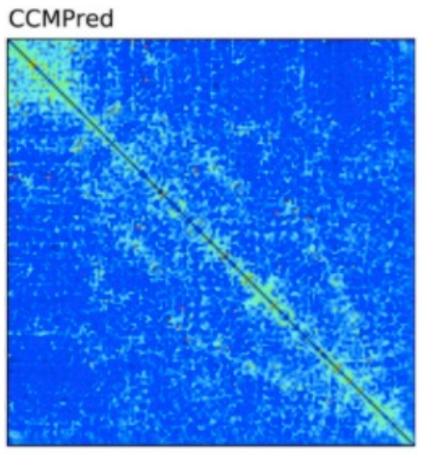
Liu, Palmedo, et al. Cell Systems 2018

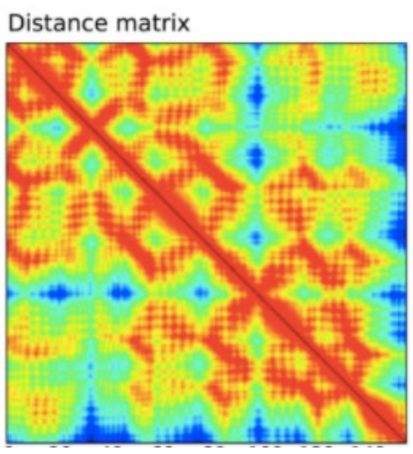
DeepContact for contact prediction

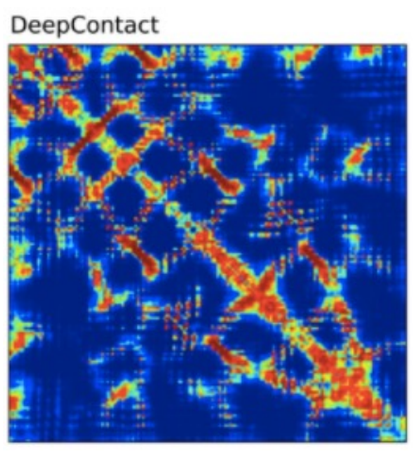


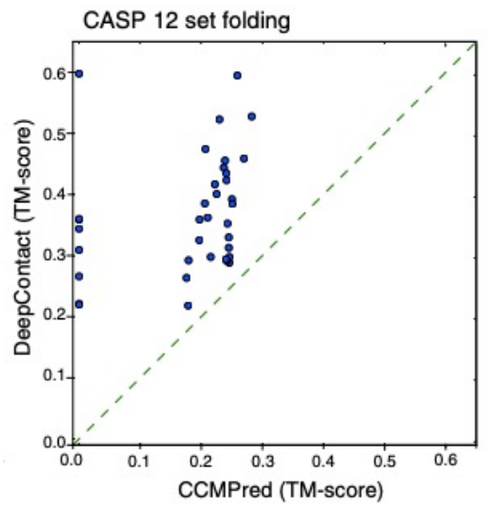
Liu, Palmedo, et al. Cell Systems 2018

Deep learning improves coevolution-based contact prediction





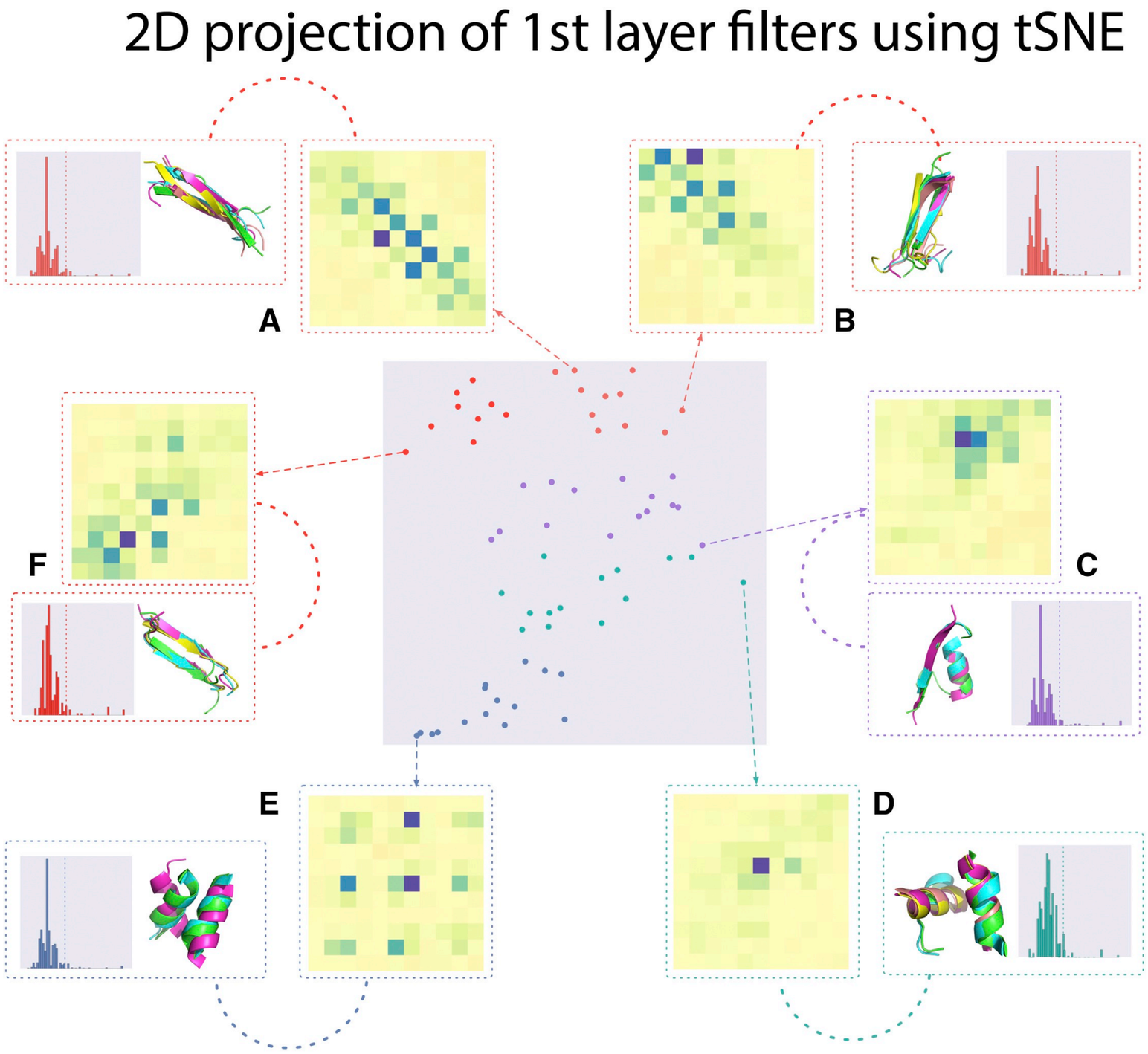




Ranked at the top in CASP12 in 2016 in Z-score ranking
on par with two other deep-learning based methods
(RaptorX-Contact and (Deep) MetaPSICOV) on other metrics

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)
1.	079	iFold_1	37	40.1644
2.	219	Deepfold-Contact	36	34.5989
3.	451	RaptorX-Contact	38	34.4778
4.	109	naive	36	27.8373
5.	431	Shen-Group	38	21.3752
6.	013	MetaPSICOV	38	17.6201
7.	287	MULTICOM-CLUSTER	38	15.1225
8.	345	MULTICOM-NOVEL	38	14.1530
9.	236	MULTICOM-CONSTRUCT	38	12.3499
10.	320	raghavagps	38	11.4462

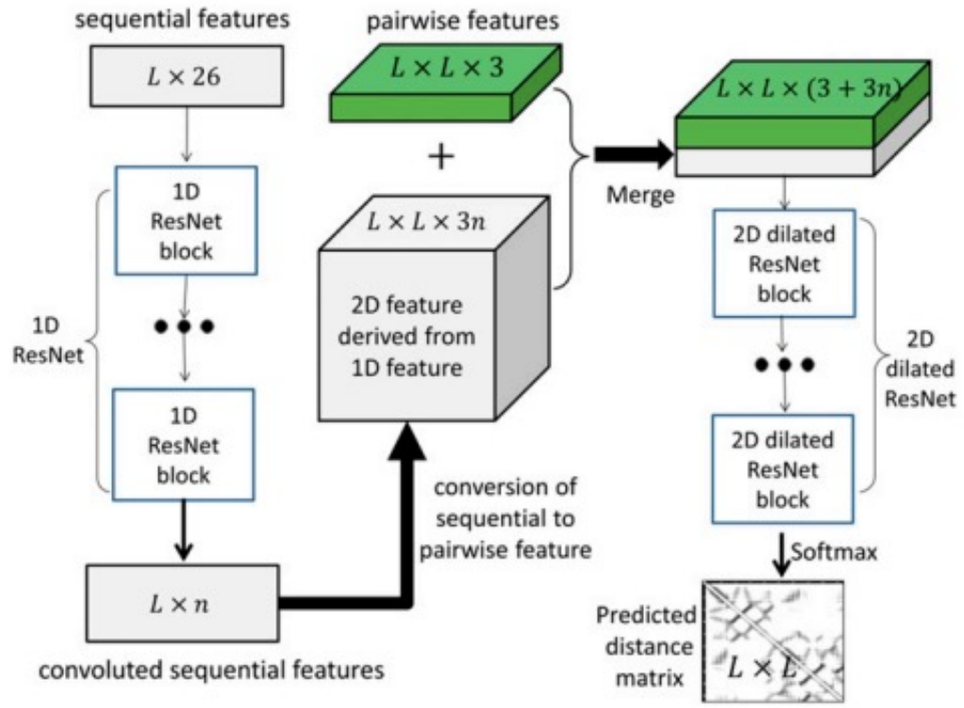
Why is deep learning effective?



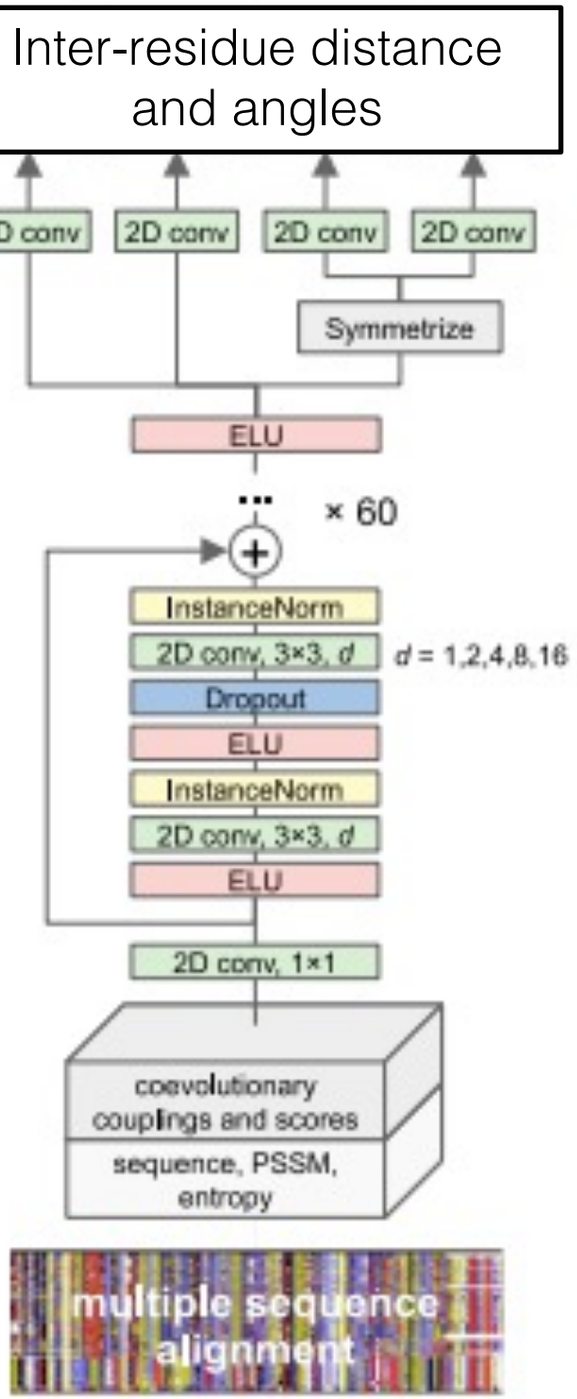
Deep neural network learns contact patterns

Recent developments go beyond contact prediction

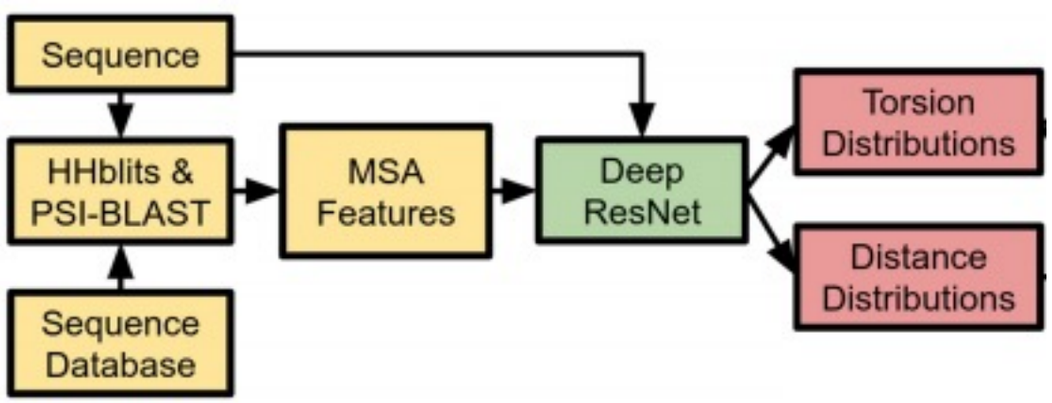
RaptorX



Rosetta

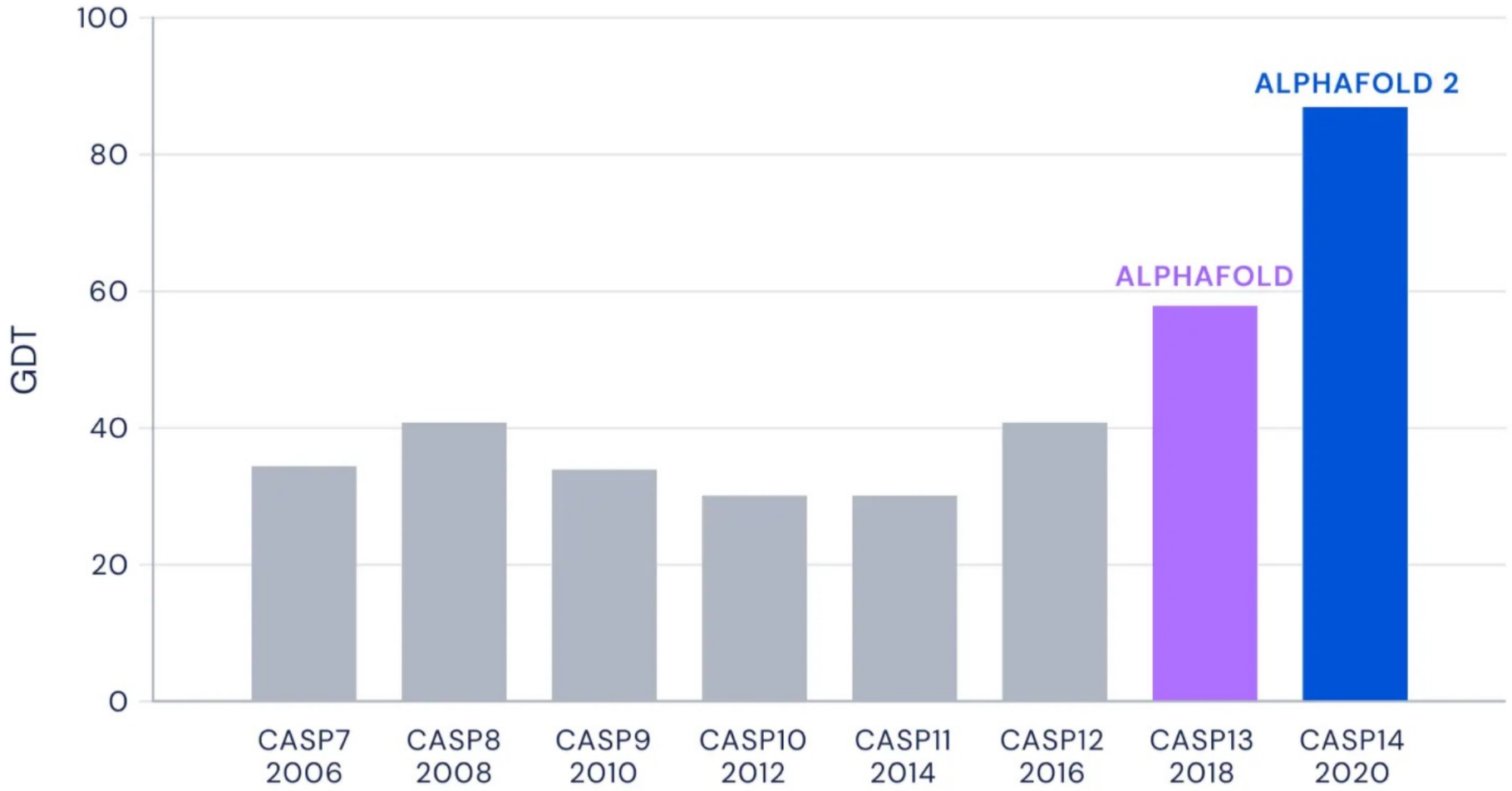


ALPHAFOLD



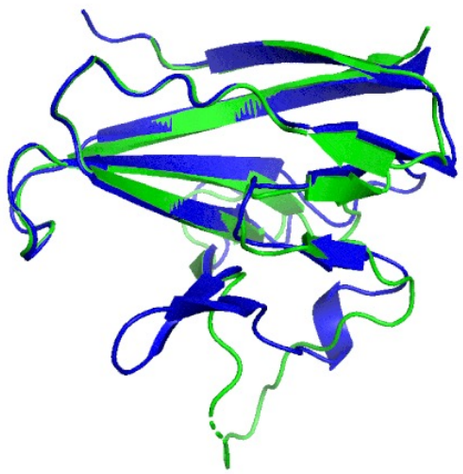
Xu. PNAS 2019; Senior, et al. Nature 2020; Yang et al. PNAS 2020

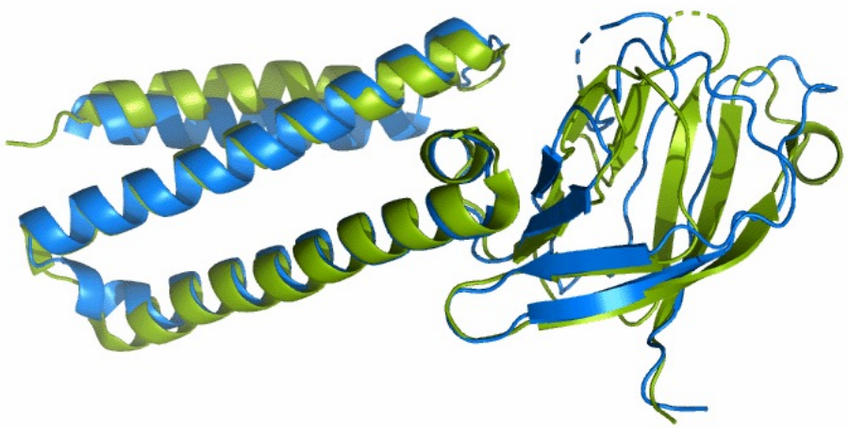
CASP14: DeepMind's AlphaFold 2



Blue: Predicted
Green: Actual

ORF8

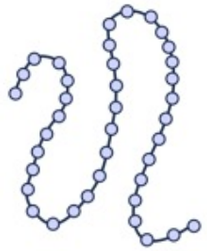




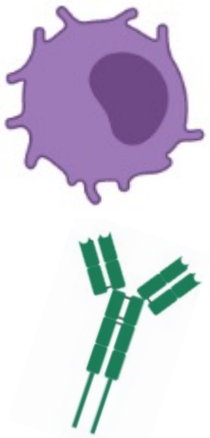
ORF3a

Function Prediction

Optimization of protein function

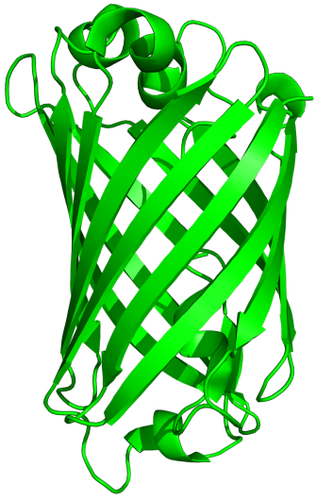


Sequence → Function



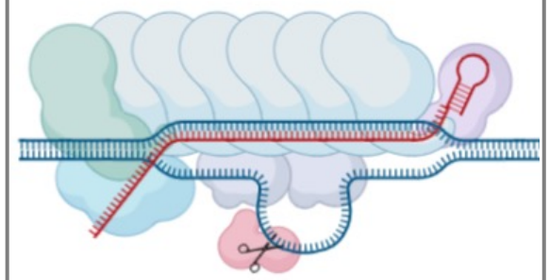
Antibody

Binding affinity



Fluorescent protein

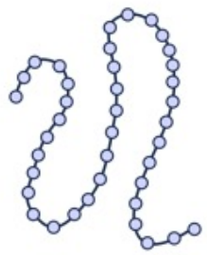
Fluorescence



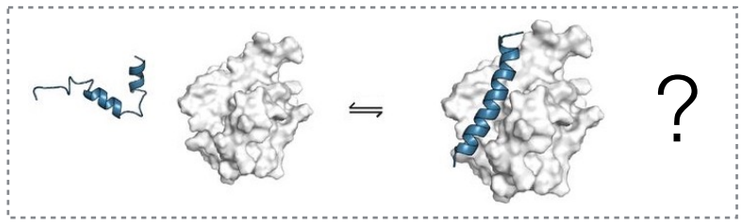
CRISPR/Cas9

Specificity

Sequence-to-function modeling



Sequence → Machine learning model → Function

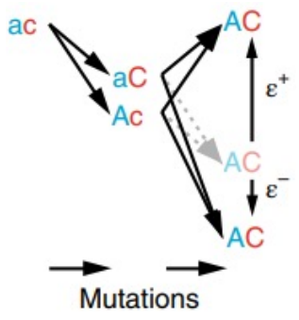


Need to differentiate function levels of closely related sequences

Sequence	Fitness
…DNGVDGEWTYDDATKTFTVTE	1.0
…DNG**C**DGEWTYDDATKTFTVTE	0.2
…DNGV**W**GEWTYDDATKTFTVTE	3.9
…DNGV**W**GEWTYDDATKTFT**F**TE	5.4
…DNGV**M**GEWTYDDATKTFT**D**TE	0.1

Need to model non-additive effect (epistasis)

$$f_{ij} = f_i + f_j + \varepsilon_{ij}$$



Need to generalize to unseen sequences/mutations

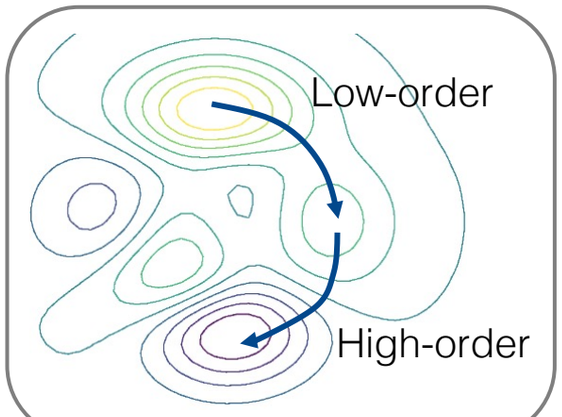
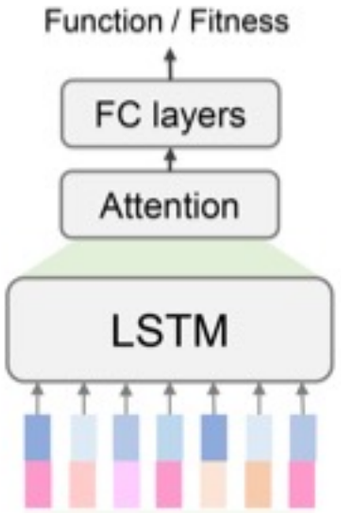
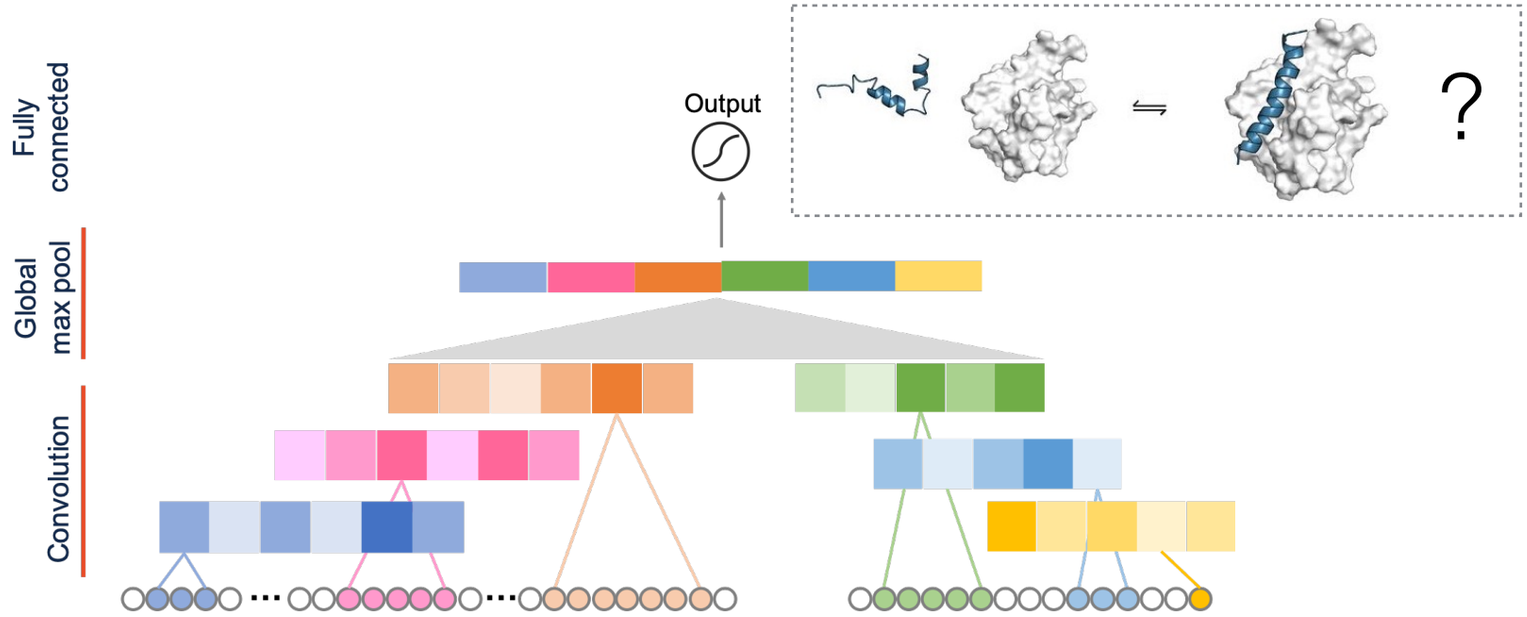


Image from Schmiedel et al., 2019

Successful sequence-to-function models

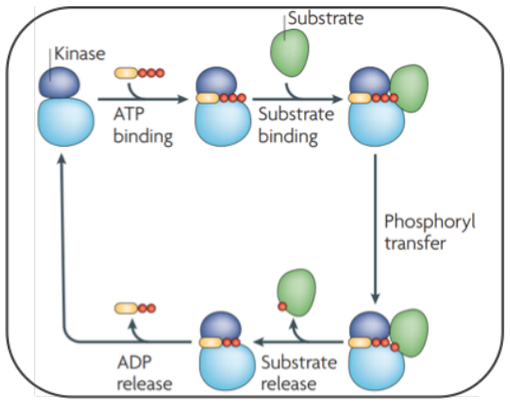


protein sequence



protein sequence

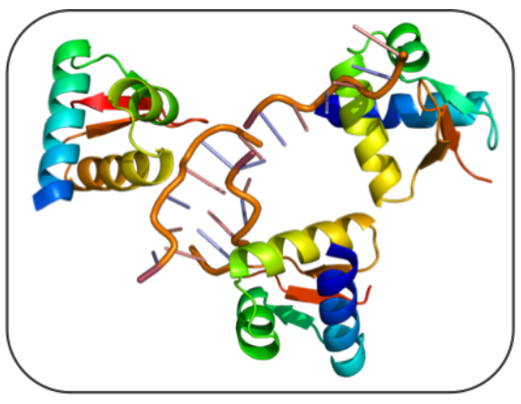
substrate representation



Kinase-peptide binding (protein phosphorylation)

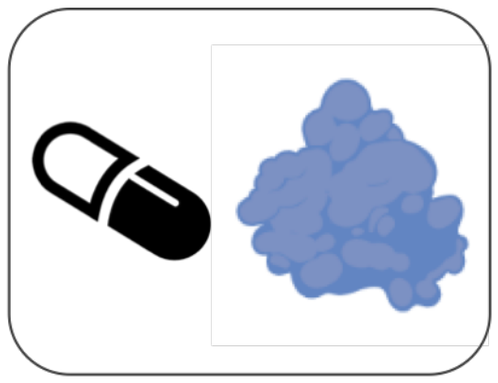
[Luo*, Ma* et al., *RECOMB,* 2018]

[Luo*, Ma* et al., *RECOMB,* 2019]



Protein-RNA binding

[Su*, Luo* et al, *PLOS CB*, 2019]

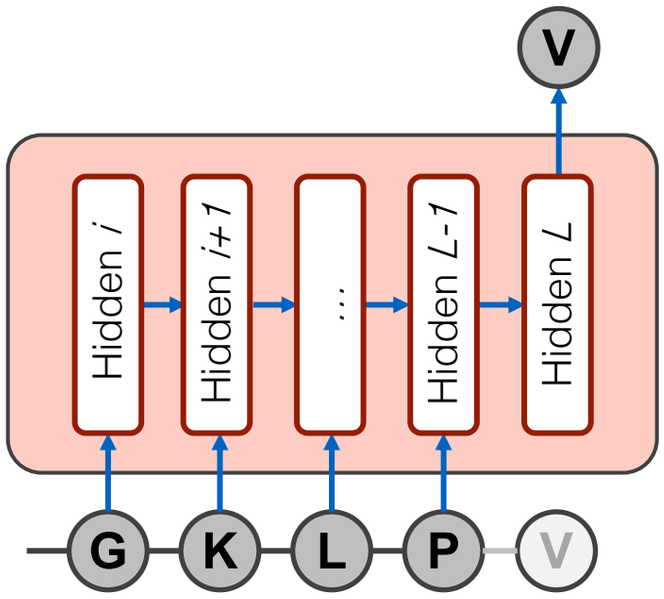


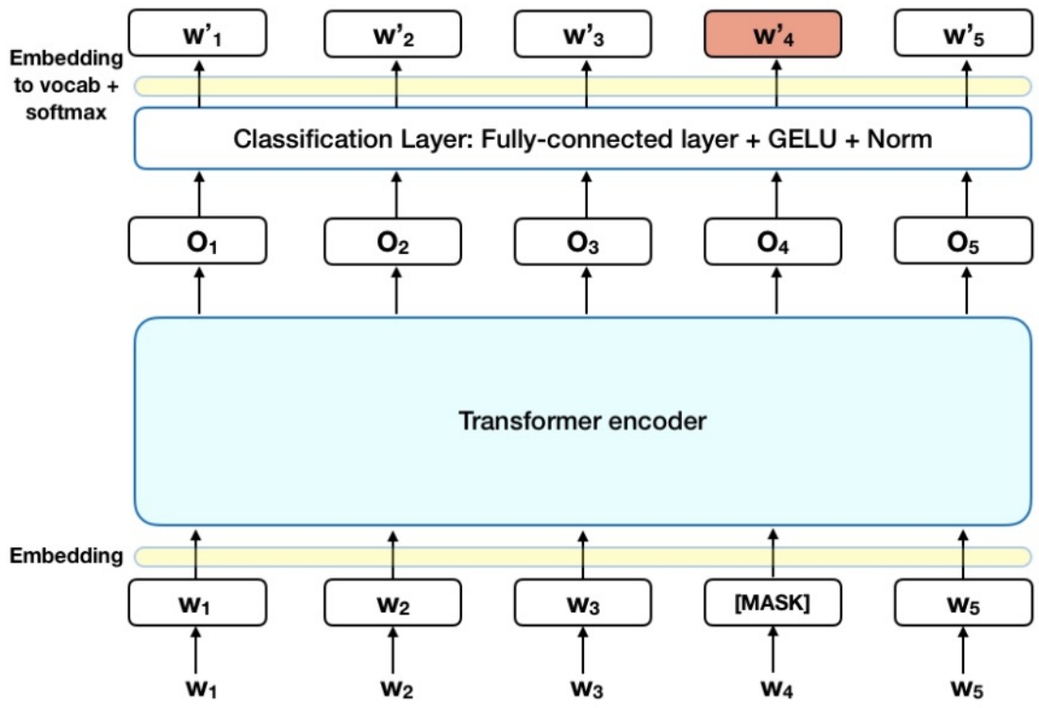
Kinase-drug binding

[Winner of IDG-DREAM Challenge, 2019]

Challenge: labeled data are expensive to get

Idea: unsupervised representation learning using language models using unlabeled data



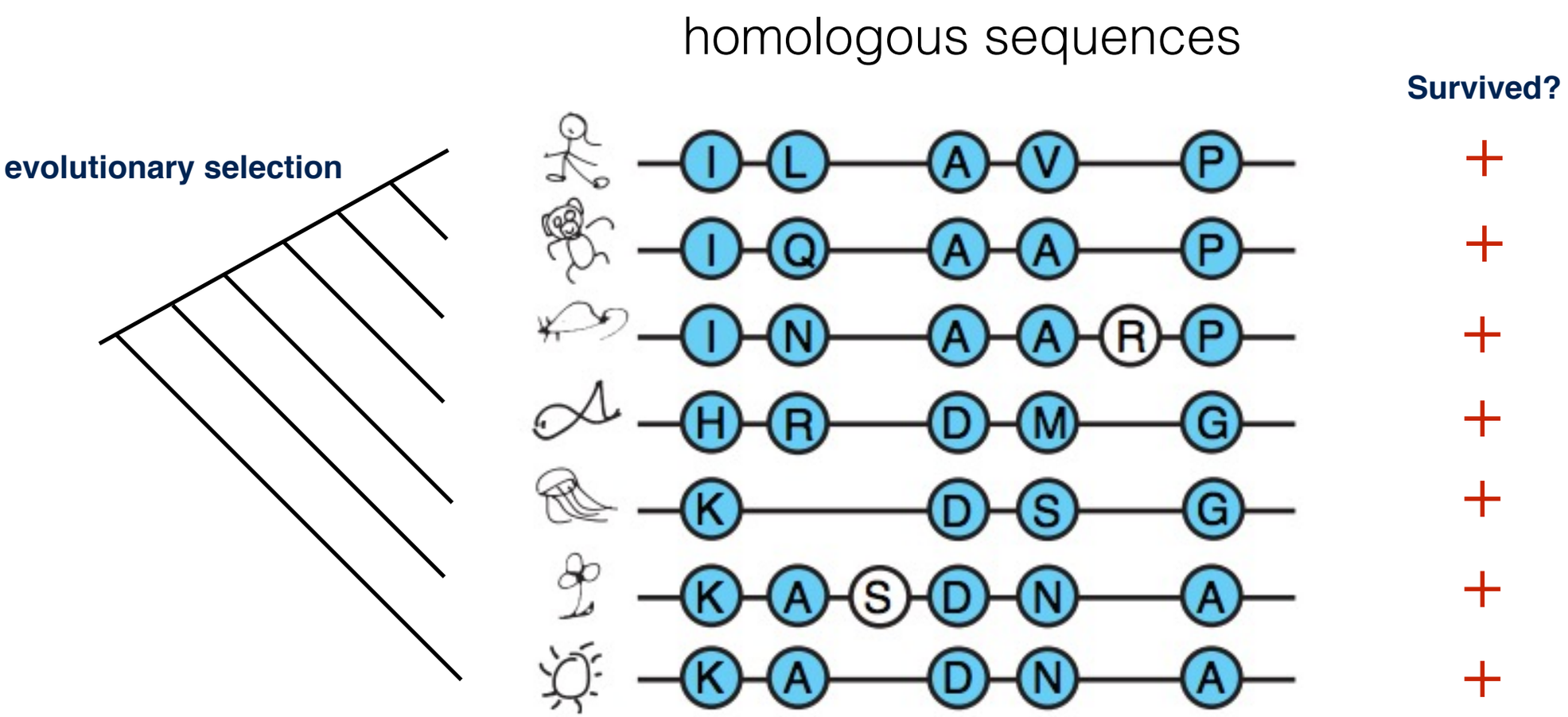


Trained on Pfam / UniProt database with unlabeled data

Too general, not specific/sensitive to a single or few changes in the sequence

Rives, Goyal, et al. bioRxiv 2019; Rao, Bhattacharya, et al. NeurIPS 2019; Yang et al. Nature Methods 2019; Elnaggar, et al. arxiv 2020

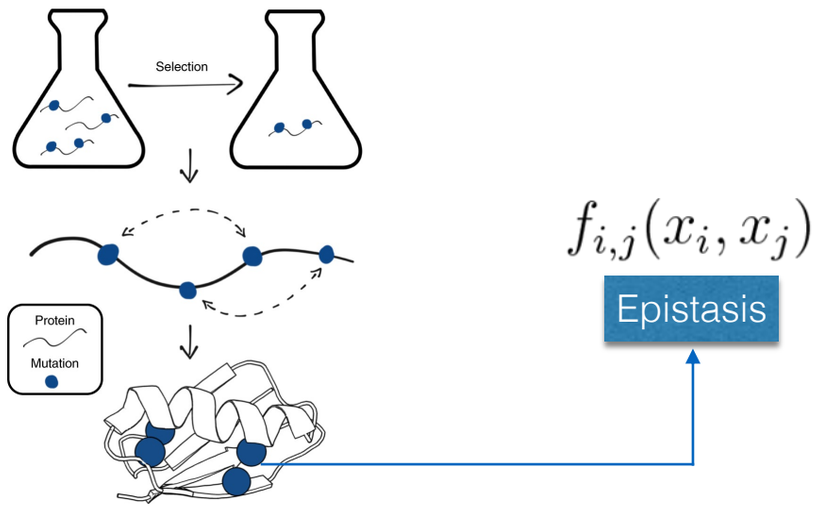
Another idea: Learning from natural evolution



Not enough for fitting language models but enough for getting good features

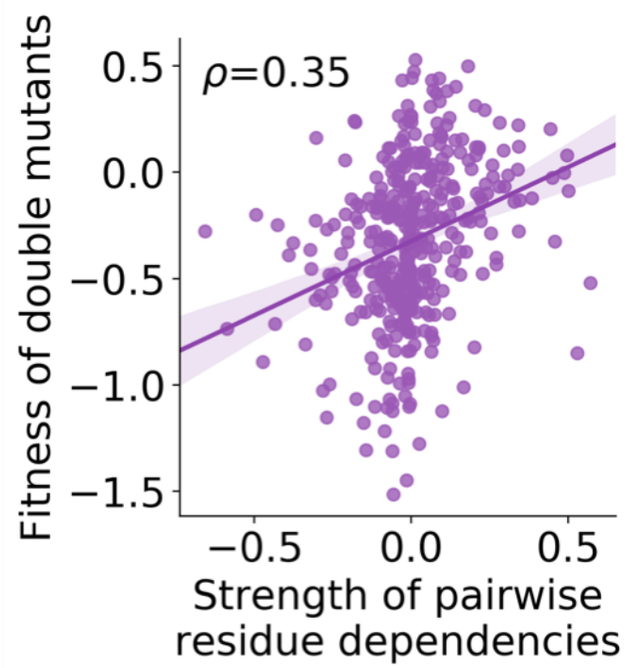
Image adopted from Hopf, et al. Nature Biotech 2017

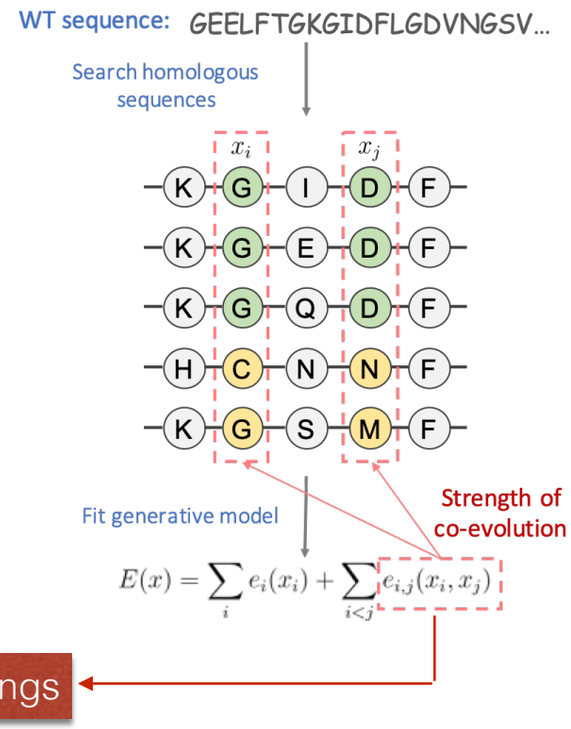
Co-evolution correlates with function



$f_{i,j}(x_i, x_j)$

Epistasis

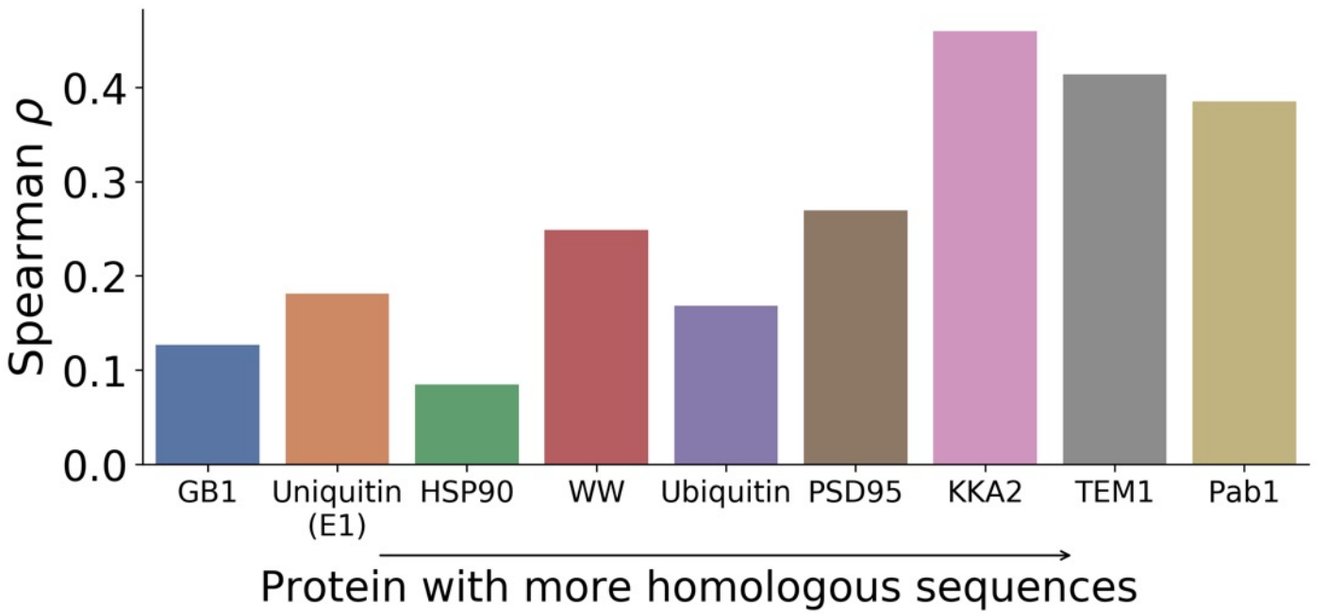




$E(x) = \sum_i e_i(x_i) + \sum_{i<j} e_{i,j}(x_i, x_j)$

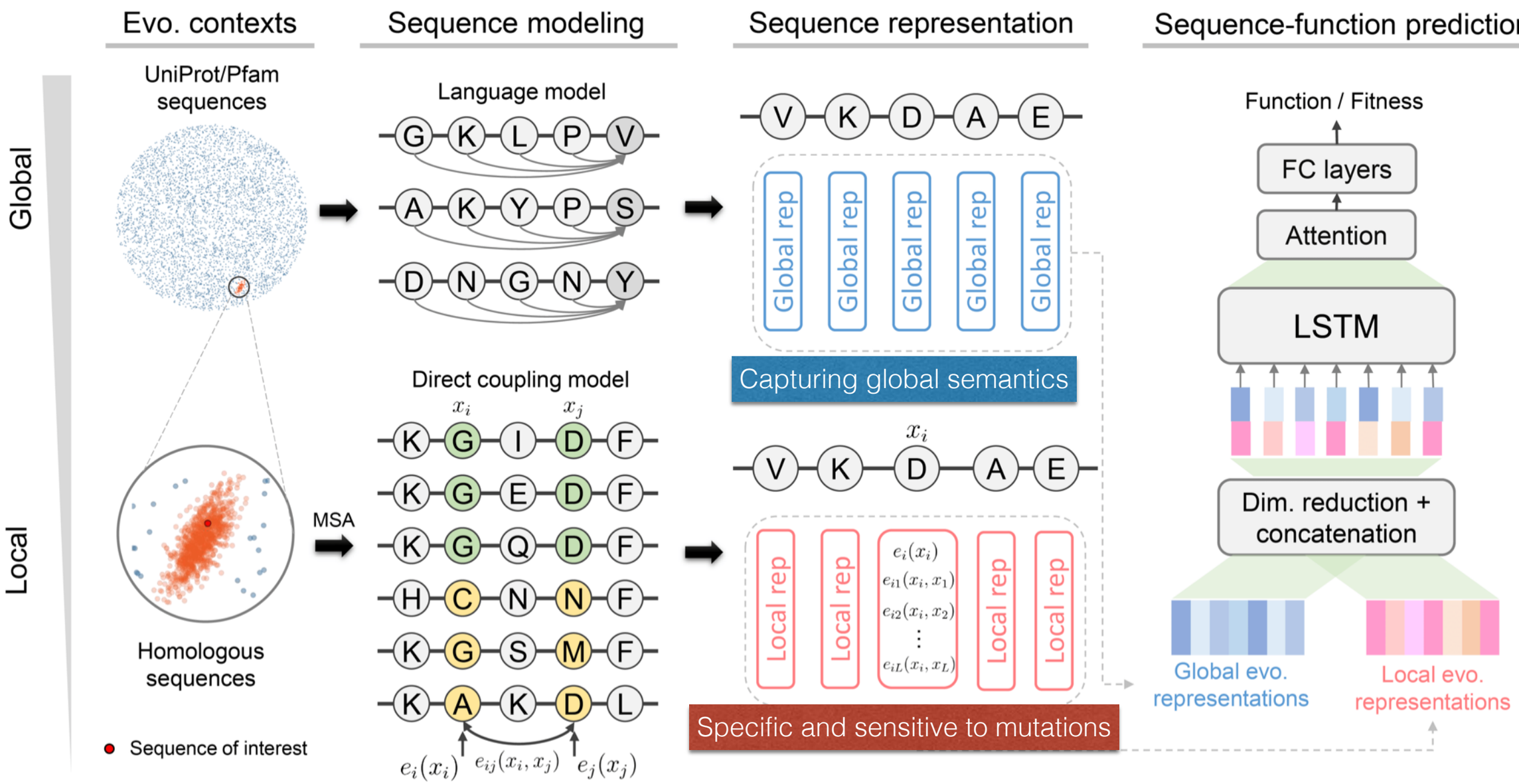
$e_{i,j}(x_i, x_j)$

Evolutionary couplings



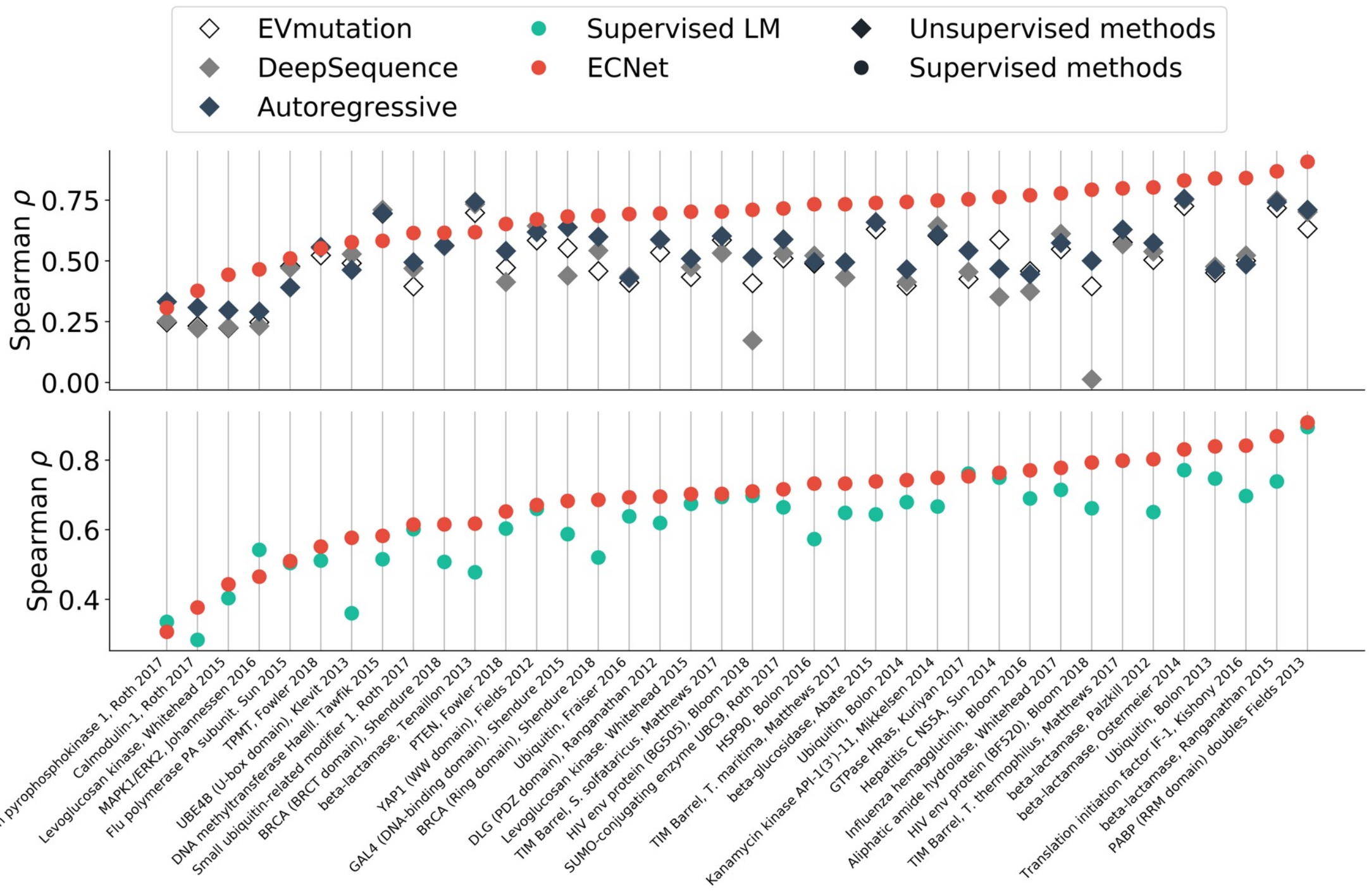
Hopf, et al. Nature Biotech 2017; Riesselman, et al. Nature Methods 2018; Luo, et al. RECOMB 2020

ECNet: integrating evolutionary contexts for protein function prediction



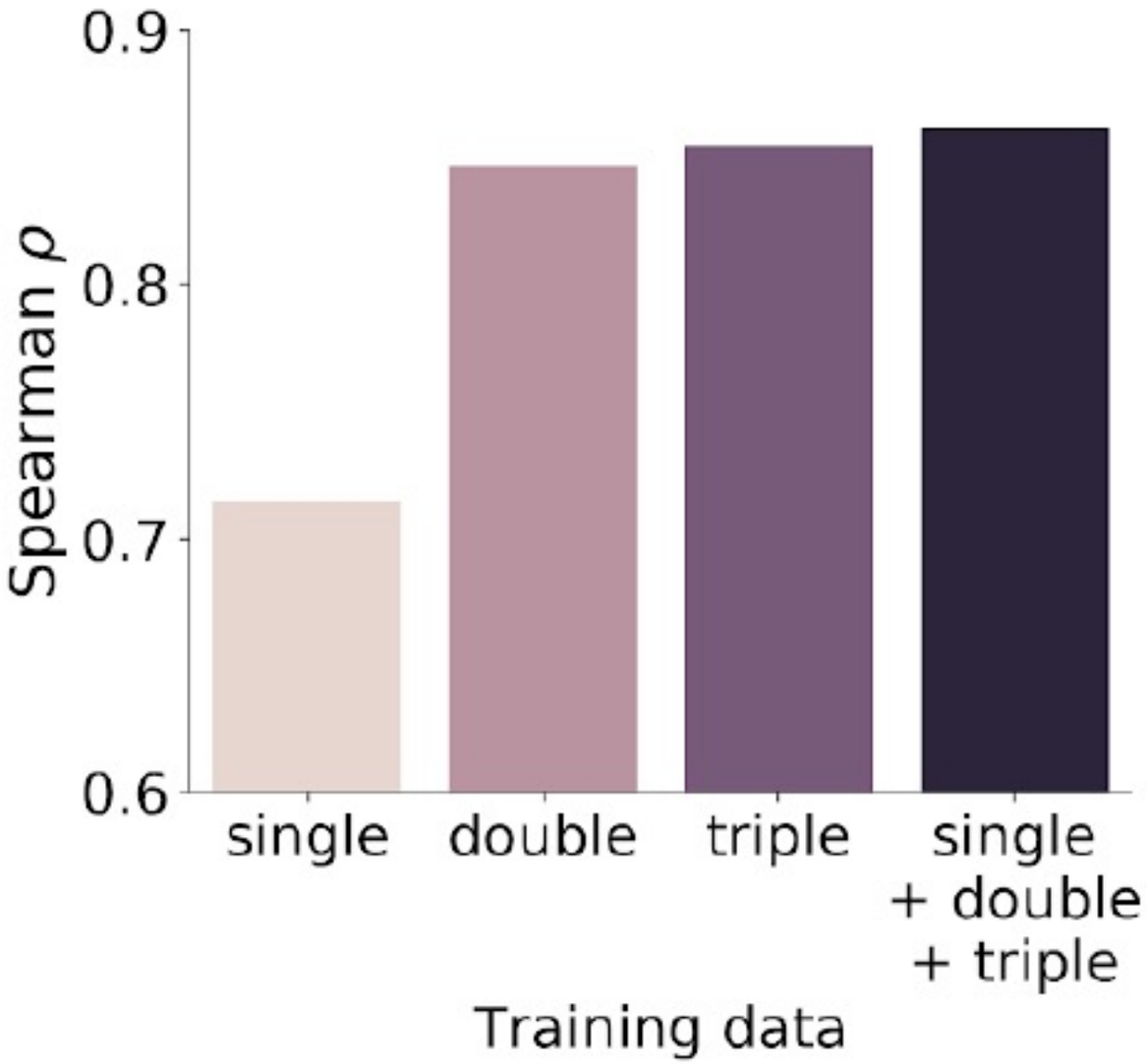
Luo, et al. RECOMB 2020

Evaluation on single-mutation datasets

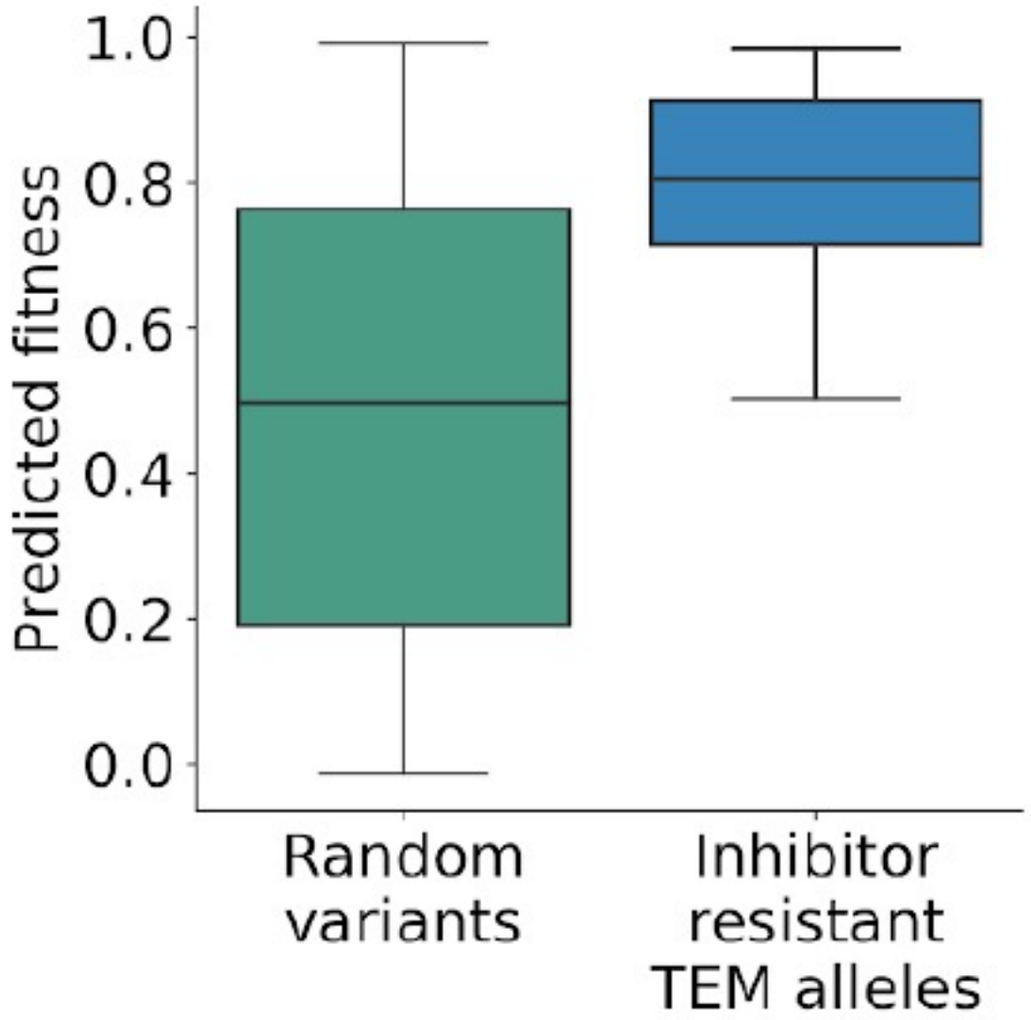


Generalization to high-order mutations

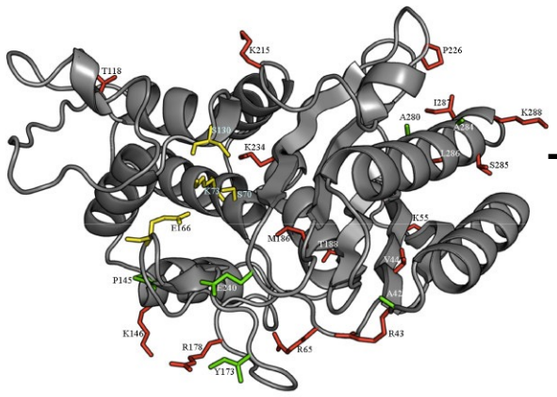
Test on high-order (4~11) GFP variants



High-order (3~15) resistant TEM



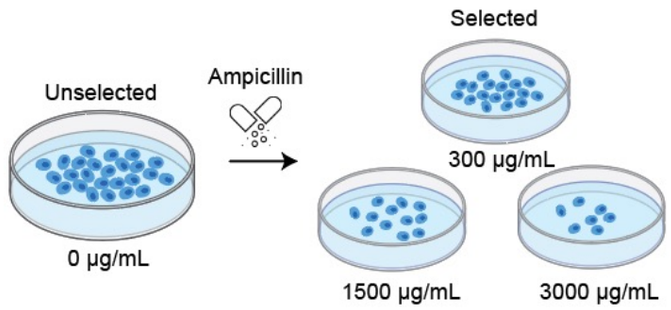
Engineering of inhibitor-resistant TEM1 beta-lactamase

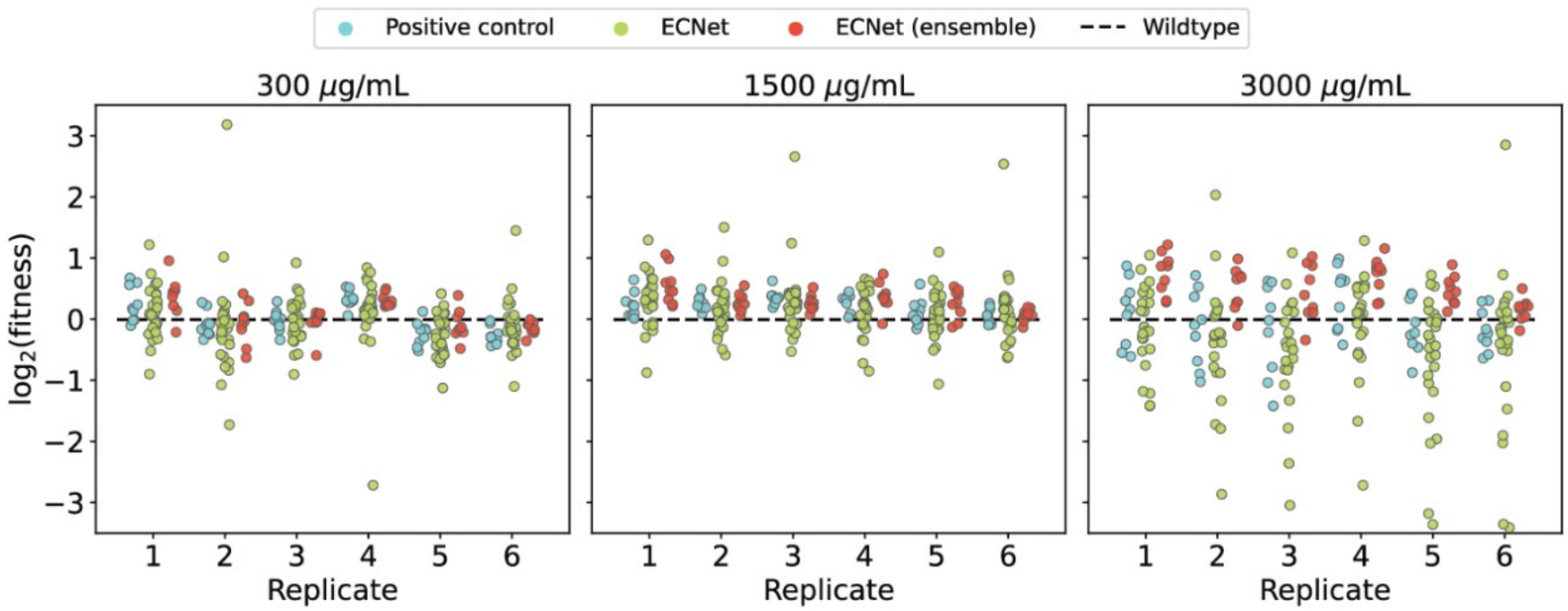


ECNet

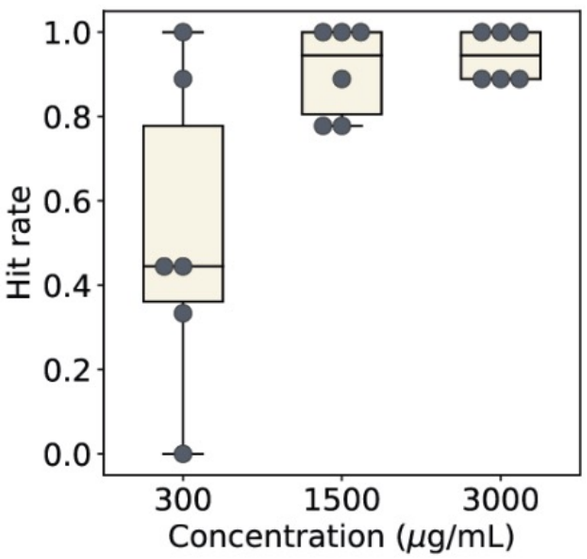
Experimental validation

Variants construction & fitness measurement (ampicillin resistance)





Fraction of improved predictions



Thank you & acknowledgements

Students: Sheng Wang, Yunan Luo, Hoon Cho, Nate Russell, Wesley Qian, Yang Liu, Yufeng Su, Palash Sashittal,

Collaborators: Bonnie Berger, Vik Khurana, Trey Ideker, Mikko Taipale, Jianzhu Ma, Jianyang Zeng, Qiang Liu, Mohammed El-Kebir, Huimin Zhao, Martin Burke, David Heckerman, ChengXiang Zhai, Jiawei Han, Aditya Parameswaran, Jadwiga Bienkowska, Lenore Cowen, Alex Schwing, Jonathan Carlson, Sumaiya Nazeen, Dina Zielinski

Alfred P. Sloan FOUNDATION

U.S. DEPARTMENT OF ENERGY

Microsoft