Protein variations. Effect prediction

Supervisor: Andrei Afansiev, iBinom
Ivan Sosin

The first goal

Investigate possibility to **discriminate between neutral and pathogenic non-synonymous single nucleotide variations** by means of Neural Networks.

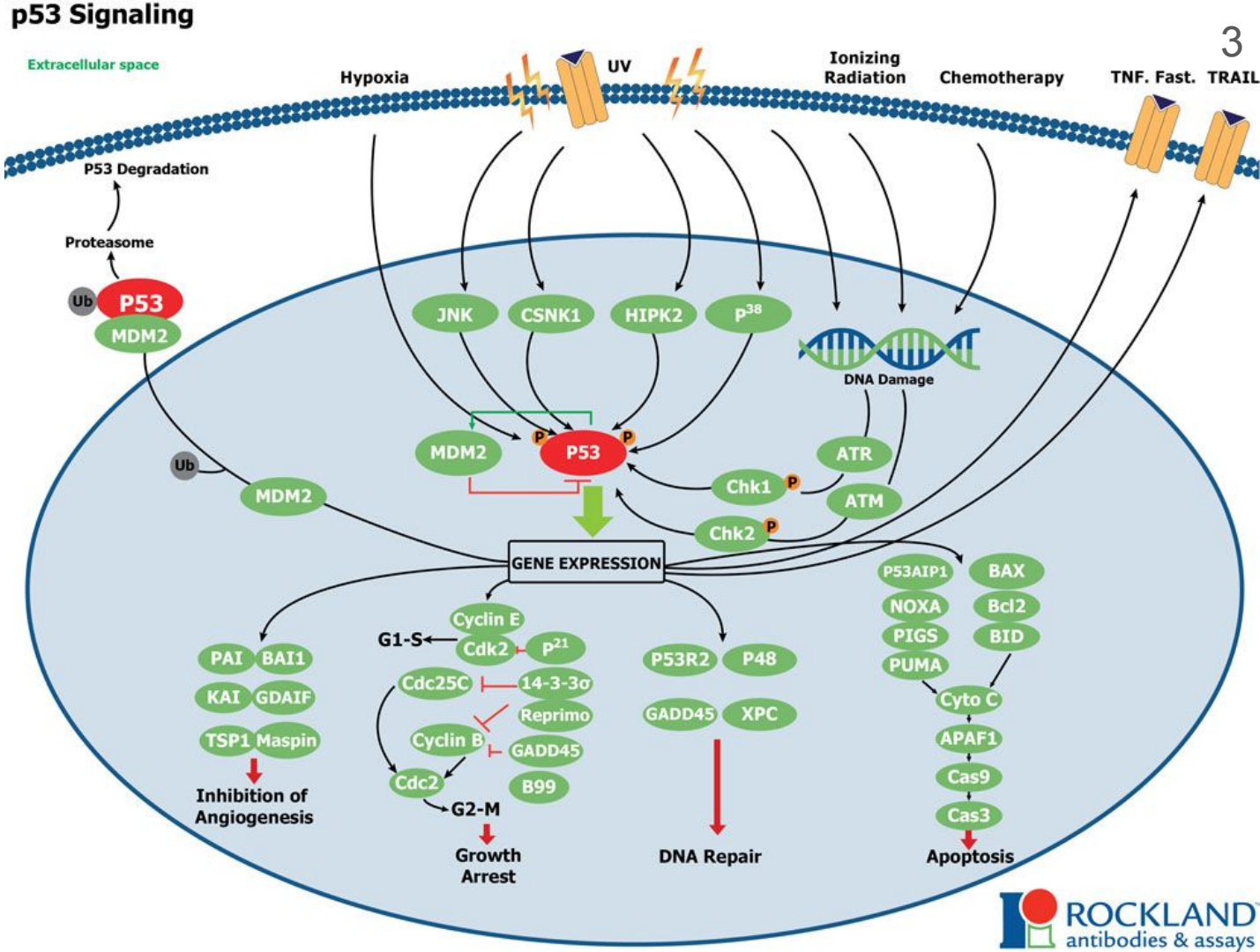
Importance: Knowing what is broken for any particular disease you can figure out what to cure.

The second goal

Investigate possibility to **predict post-translational modification sites**.

Importance: A lot of diseases are caused by malfunctioning regulatory pathways. Regulation is heavily based on PTMs.

P53 signalling pathway



Resources

Databases:

HUMSAVAR (Variant-disease association data)

PhosphoSitePlus

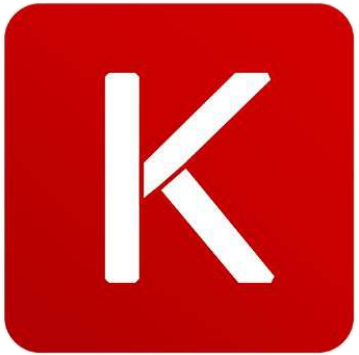
UniProt

Toolbox

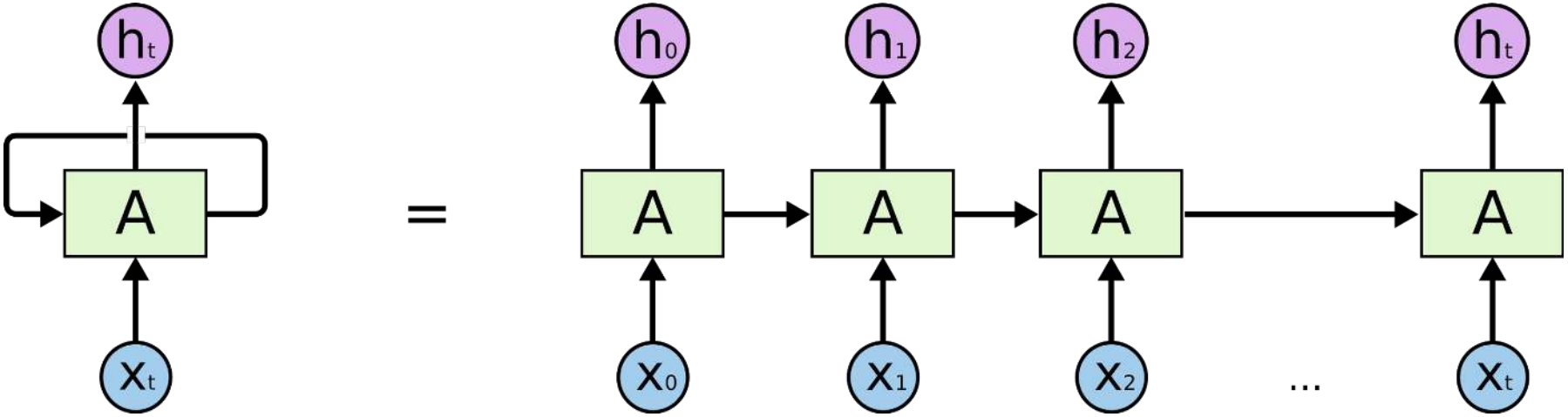
Theano

Lasagne

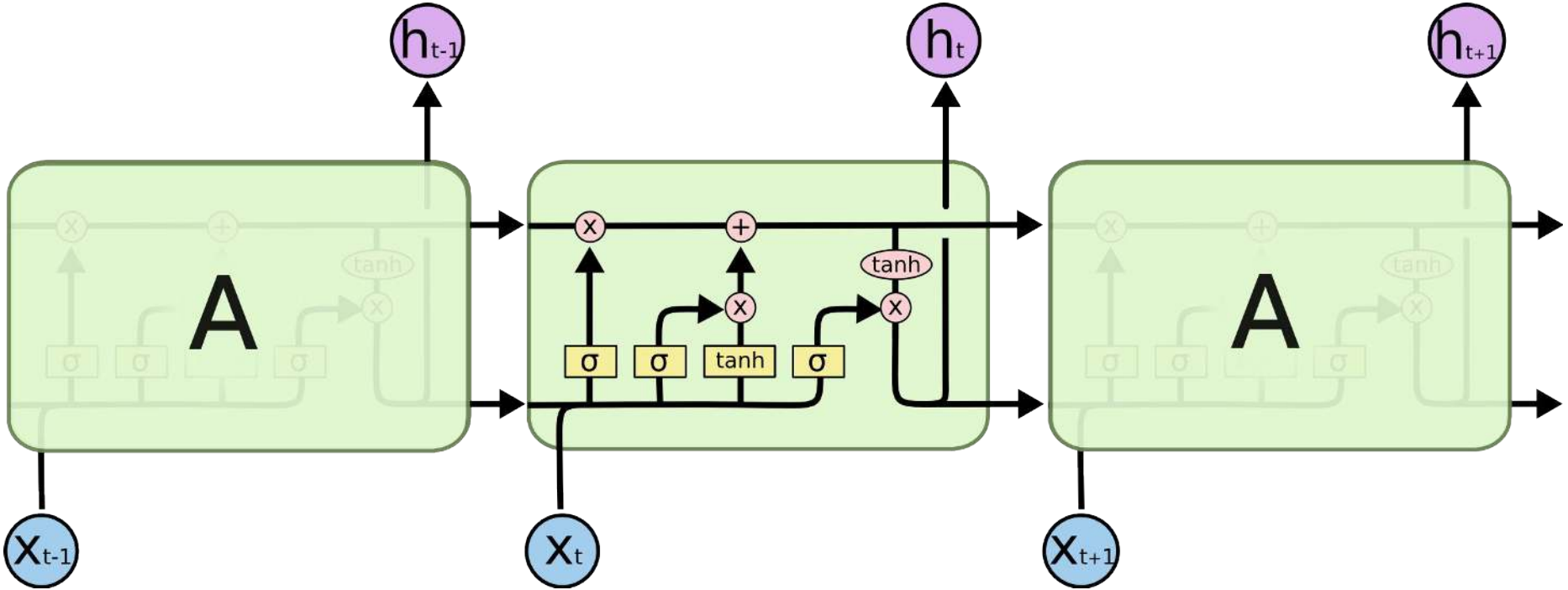
Keras



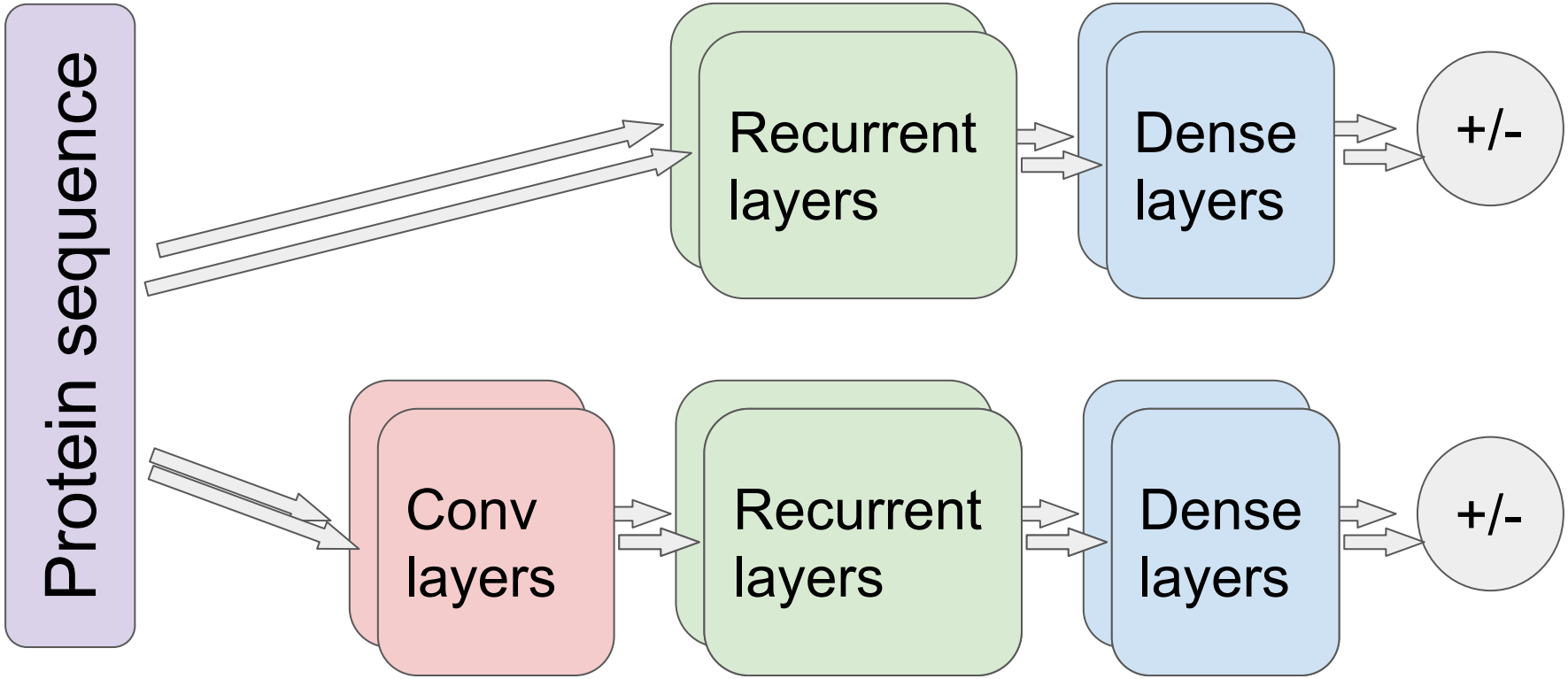
Core unit - Long short-term memory (LSTM) - 1



Core unit - Long short-term memory (LSTM) - 2



General architecture



Protein structure representation

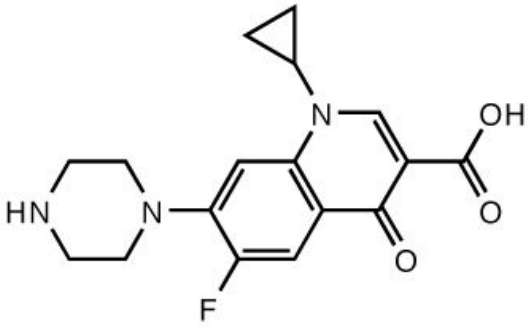
1) Amino acid -> ASCII code

2) Amino acid -> feature vector

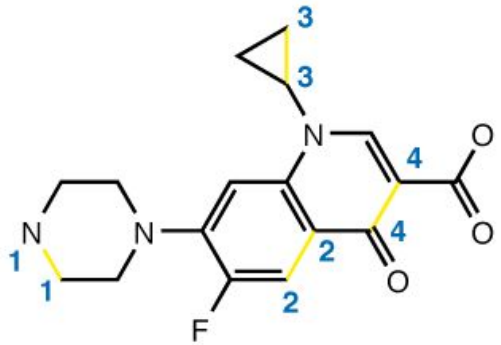
3) Amino acid -> vector embedding

Simplified molecular-input line-entry (Smiles)

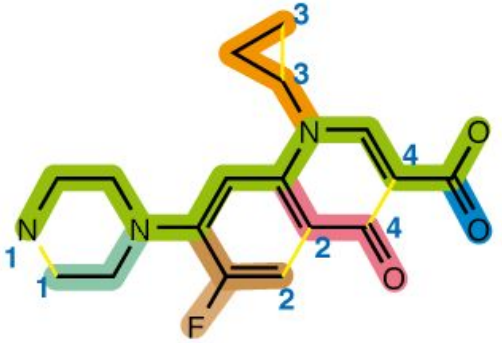
A



B



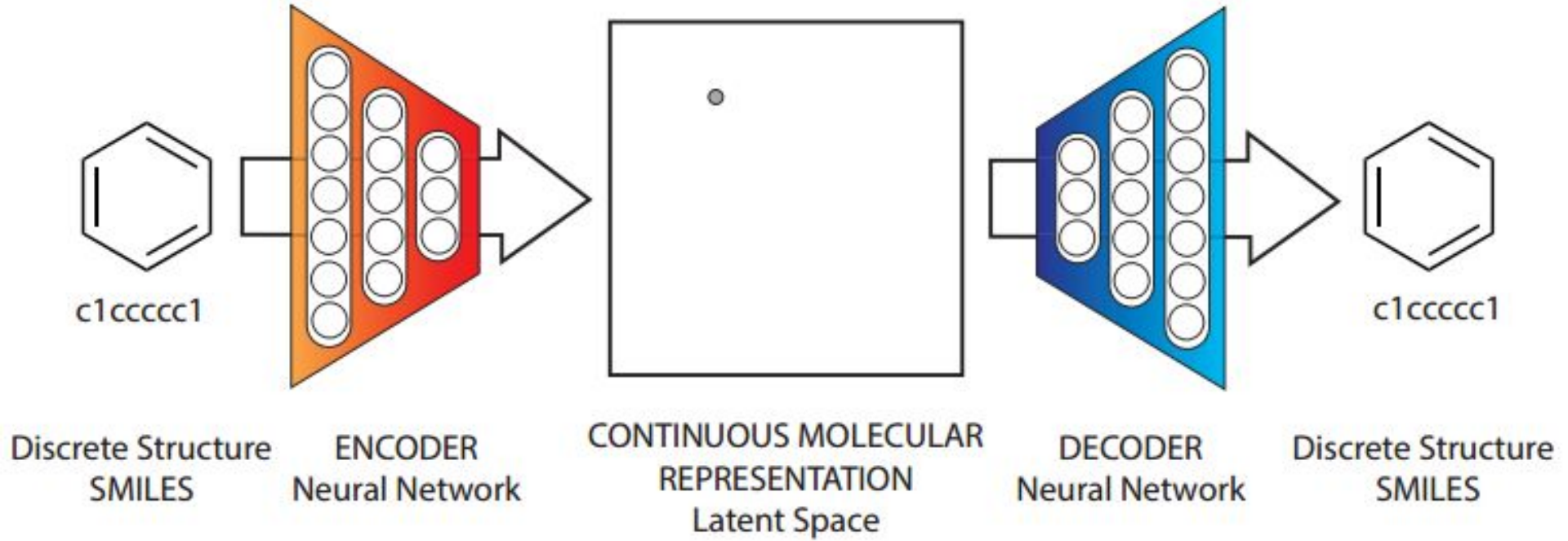
C



D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Smiles autoencoder



Achievements

Best achieved accuracy is … ~50%

Causes:

1) Overfitting
2) Poor protein representation
3) Nature of protein to protein interaction

Why don’t we make our own protein autoencoder?

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY

int[500]

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY

Achieved accuracy ~ 40%

Post-translational modification site prediction

Phosphorylation

Sliding window:

Best achieved accuracy ~77.5%

Thank for your attention!

Questions?