The rise of deep learning in drug discovery

Russ B Altman, MD, PhD
Stanford University

Machine Learning

- Machine learning is the discipline within computer science where algorithms are given data and learn how to categorize or classify it.

Two general types of machine learning:

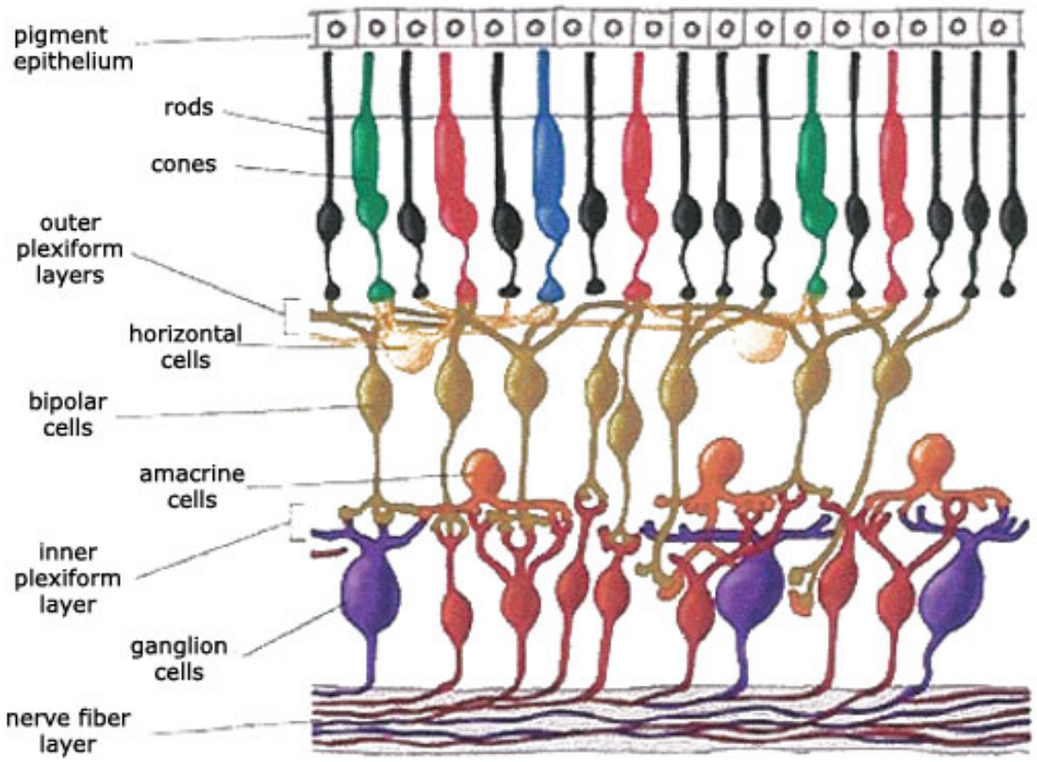
- **Supervised Machine Learning**: provides data as well as labels (the supervision), e.g. lots of data about drugs with labels of whether they are toxic or not

- **Unsupervised Machine Learning**: provides data but no labels, so groups similar objects together = clustering, e.g. lots of data about drugs that can be grouped together based on these data.

Deep Learning

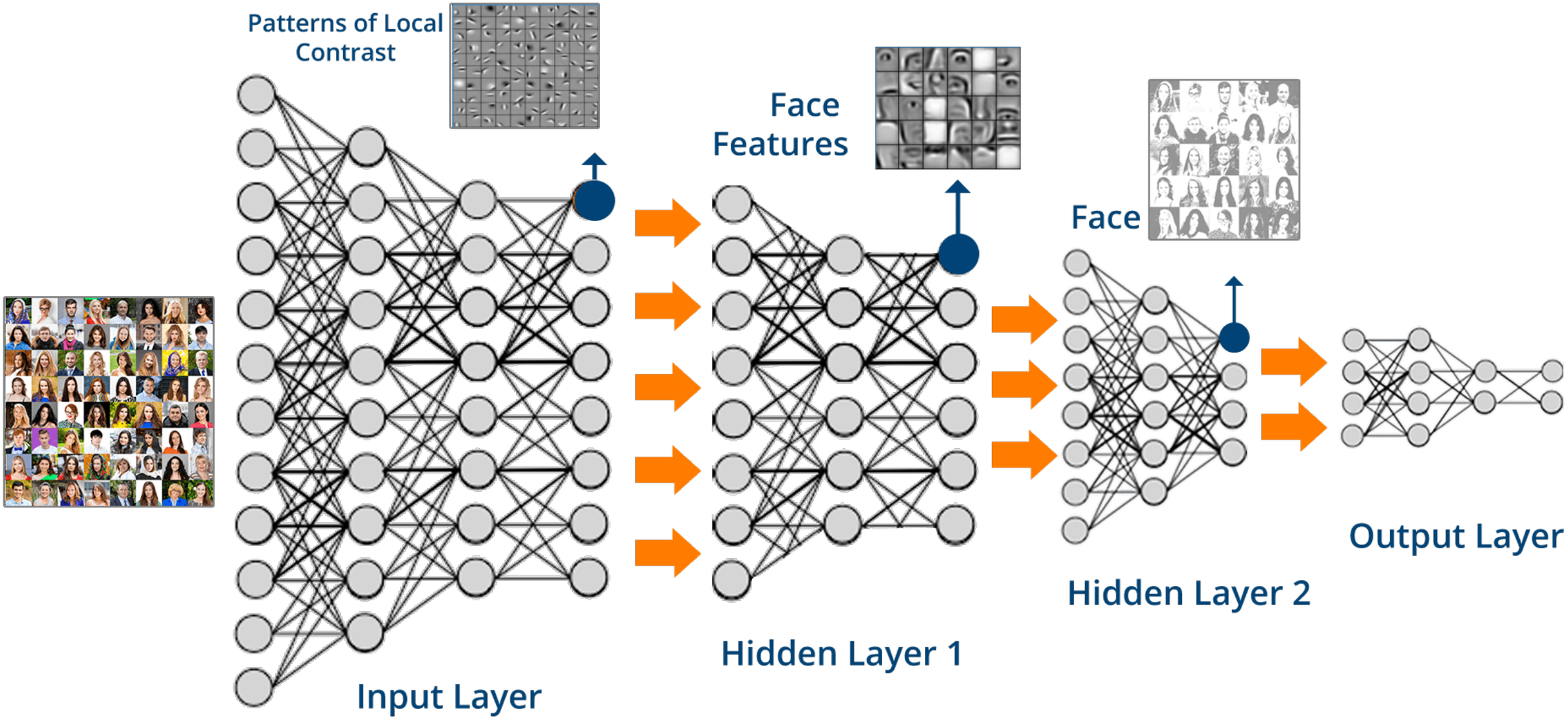
- Deep Learning is a sub-discipline within Machine Learning in which both supervised and unsupervised ML takes place using an analogy to neural processing = neural networks

- Cf. processing of light in the retina.



http://www.arn.org/docs/glicksman/eyw_041101.htm

Deep Learning



https://www.michaelchimenti.com/2017/11/deep-neural-nets-software-2-0/

Key features of Deep Learning

- Requires typically very large data sets (100K to millions of examples to learn).
  - Some emerging methods (e.g. transfer learning) to reduce the requirement for data
- (Surprisingly) Can learn the key features in the data and does not need human expert to do “feature engineering”
- Given sufficient data, able to learn complex non-linear functions of the input data to predict the output “label”
- If the data is available, the resulting classifiers have outpaced many traditional machine learning approaches by leaps and bounds.

REVIEW ARTICLE | FOCUS
https://doi.org/10.1038/s41591-018-0300-7
nature medicine

High-performance medicine: the convergence of human and artificial intelligence

Eric J. Topol

The use of artificial intelligence, and the deep-learning subtype in particular, has been enabled by the use of labeled big data, along with markedly enhanced computing power and cloud storage, across all sectors. In medicine, this is beginning to have an impact at three levels: for clinicians, predominantly via rapid, accurate image interpretation; for health systems, by improving workflow and the potential for reducing medical errors; and for patients, by enabling them to process their own data to promote health. The current limitations, including bias, privacy and security, and lack of transparency, along with the future directions of these applications will be discussed in this article. Over time, marked improvements in accuracy, productivity, and workflow will likely be actualized, but whether that will be used to improve the patient–doctor relationship or facilitate its erosion remains to be seen.

Medicine is at the crossroad of two major trends. The first is a failed business model, with increasing expenditures and jobs allocated to healthcare, but with deteriorating key outcomes, including reduced life expectancy and high infant, childhood, and maternal mortality in the United States[1,2]. This exemplifies a paradox that is not at all confined to American medicine: investment of more human capital with worse human health outcomes. The second is the generation of data in massive quantities, from sources such as high-resolution medical imaging, biosensors

type of medical scan, with more than 2 billion performed worldwide per year. In one study, the accuracy of one algorithm, based on a 121-layer convolutional neural network, in detecting pneumonia in over 112,000 labeled frontal chest X-ray images was compared with that of four radiologists, and the conclusion was that the algorithm outperformed the radiologists. However, the algorithm's AUC of 0.76, although somewhat better than that for two previously tested DNN algorithms for chest X-ray interpretation[5], is far from optimal. In addition, the test used in this study is not necessarily comparable

Areas of Deep Learning Success

- **Imaging**: (the first application area where it emerged successful) Applications in radiology, pathology, dermatology. Images are a 2D array of data.

- **Text processing**: Application in mining the published literature, mining clinical notes, taking dictation, language translation. Text is a 1D array of data.

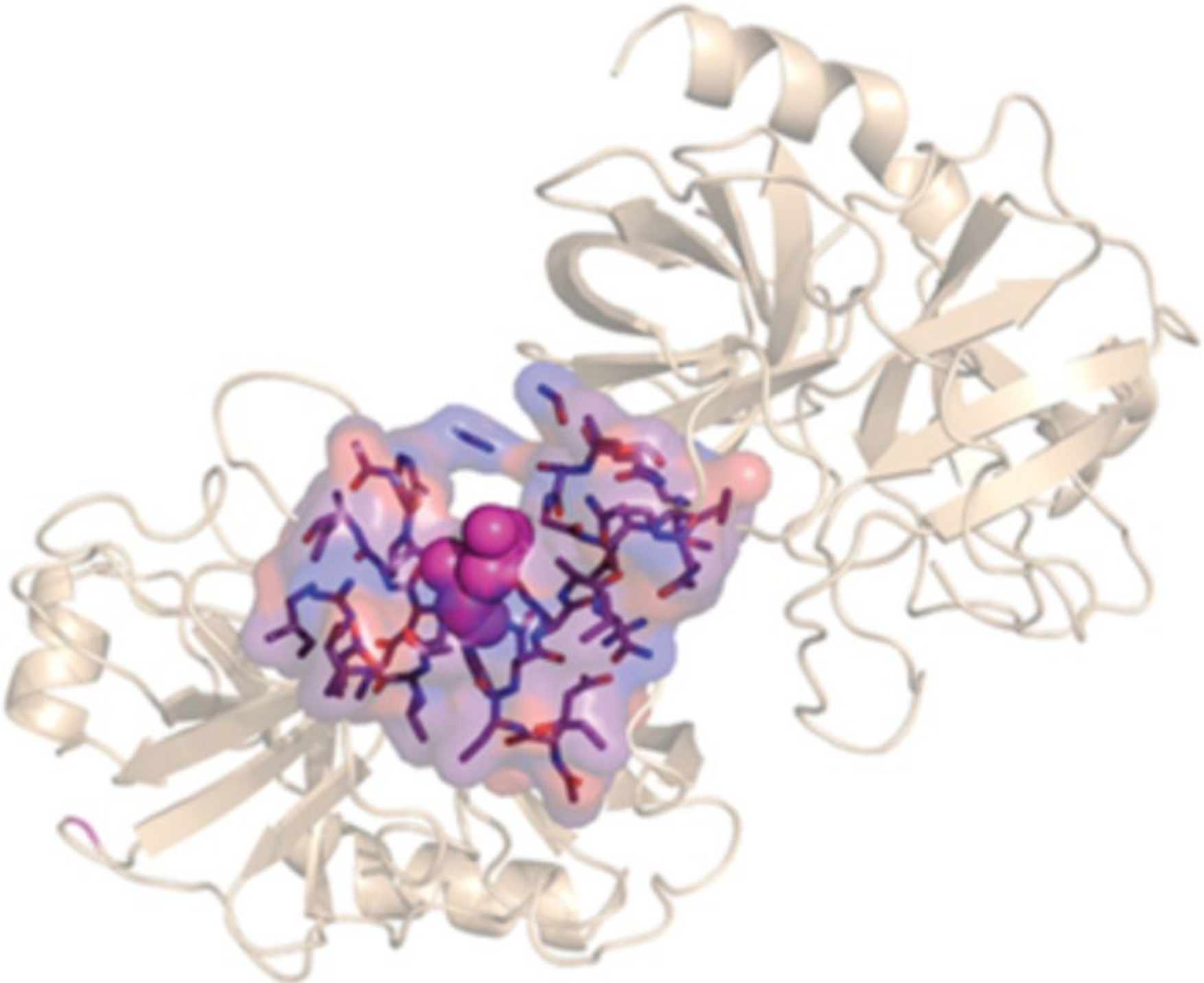
(Progress in DNA/Protein sequence analysis by analogy)

More challenging: integrating data that is not a regular array of homogenous data.

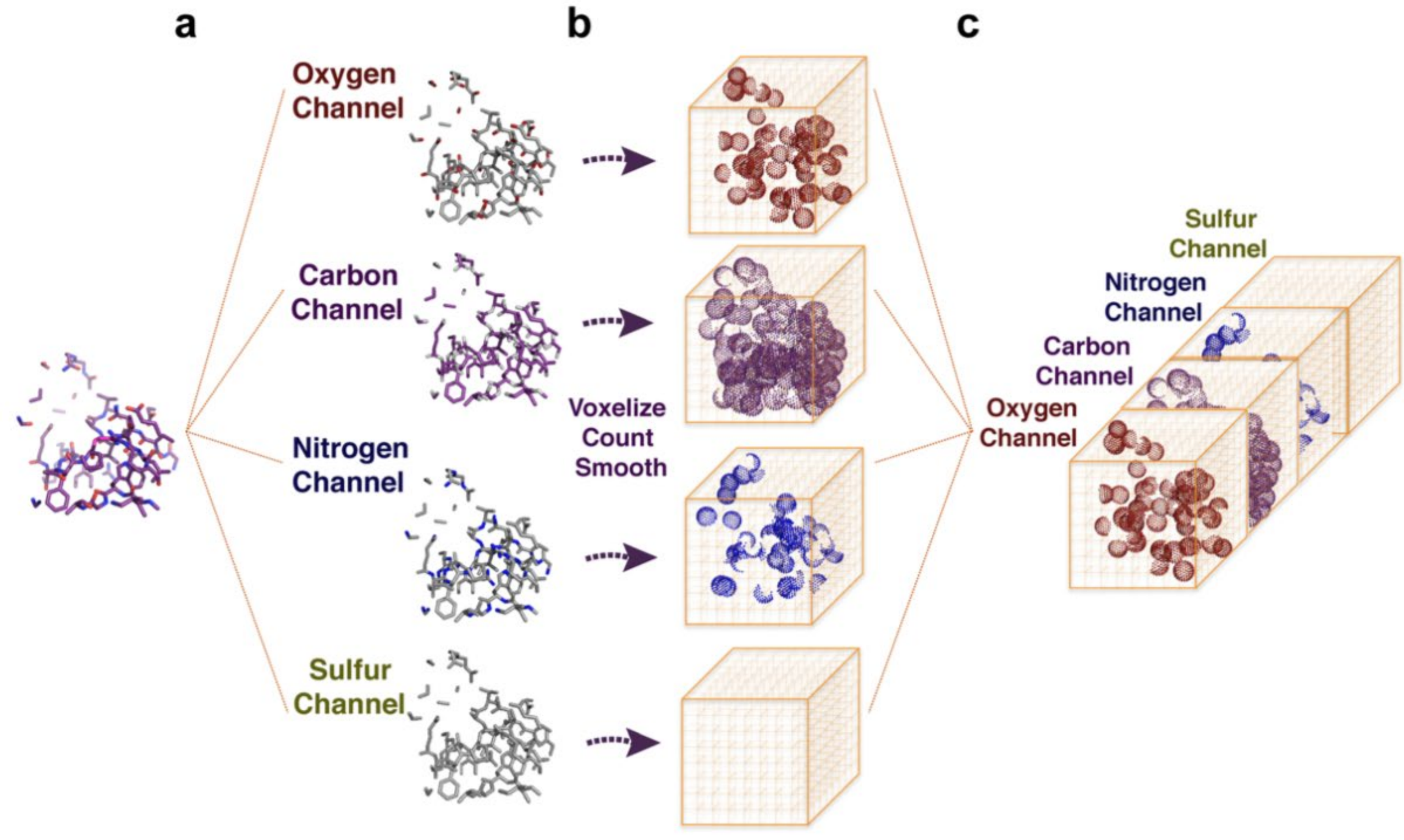
Sample applications in drug discovery

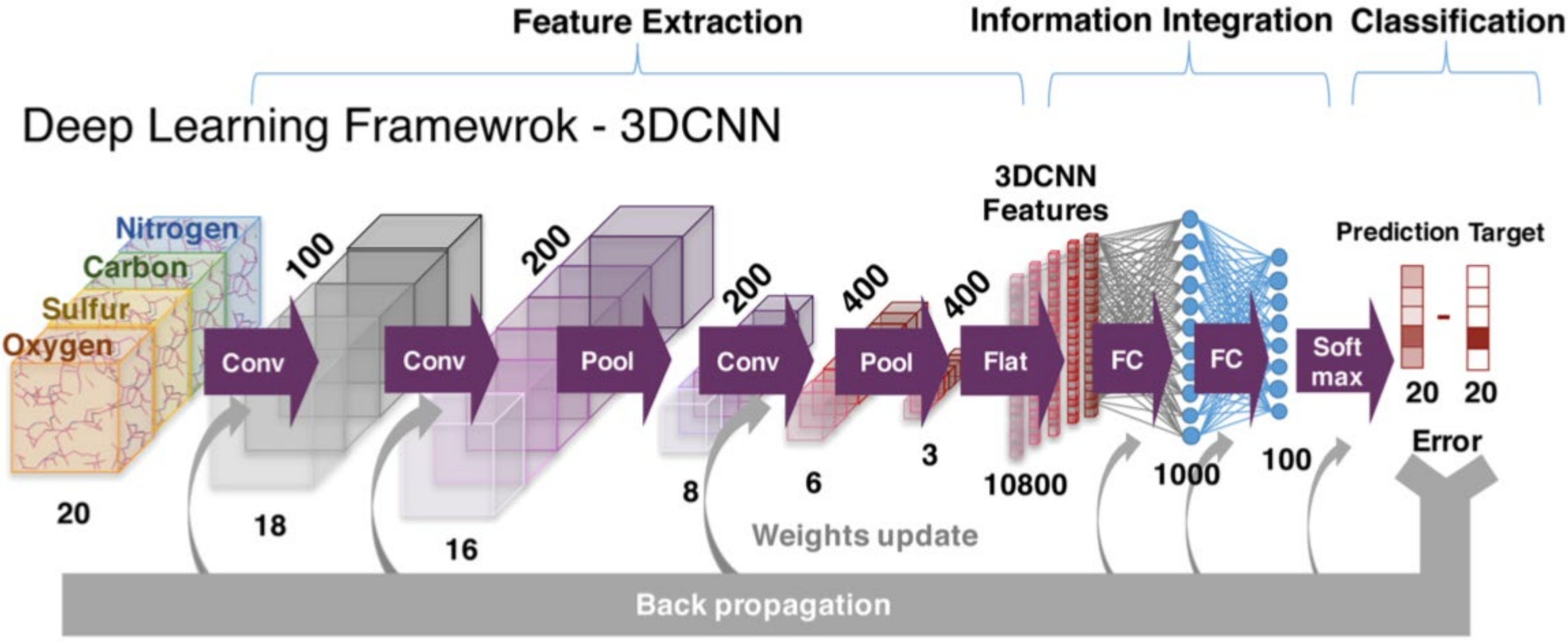
- **3D structure-based discovery/screening of ligands**
- **Cellular networks to find targets**
- **Gene-drug-phenotype networks to predict toxicity**
- **Predicting gene function (phenotype) from genotype**

Deep Networks for Understanding Molecular Structure



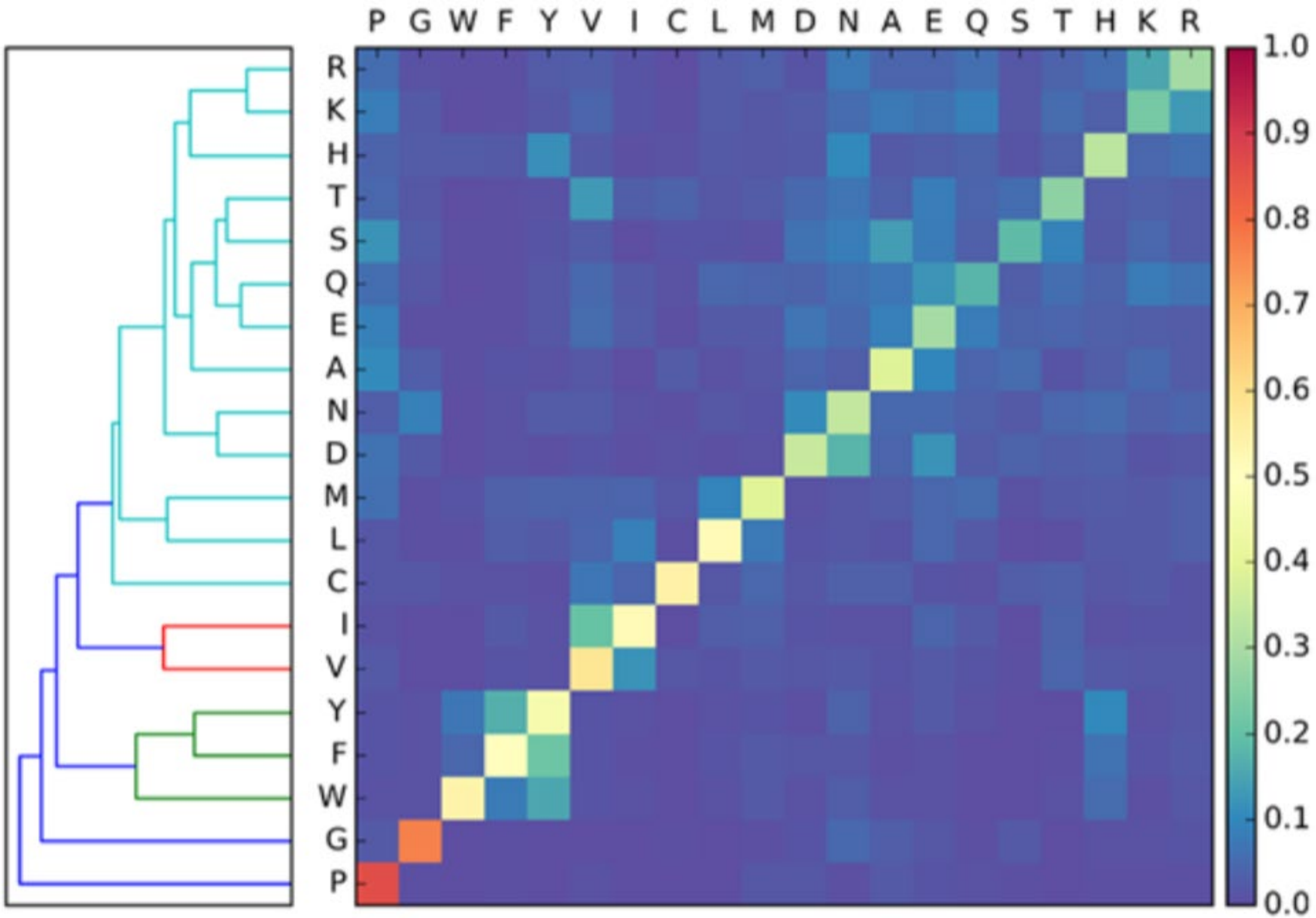
Torng W, Altman RB BMC Bioinformatics. 2017 Jun 14;18(1):302. PMID: 28615003


a
b
c
Oxygen Channel
Carbon Channel
Nitrogen Channel
Sulfur Channel
Voxelize
Count
Smooth
Sulfur Channel
Nitrogen Channel
Carbon Channel
Oxygen Channel

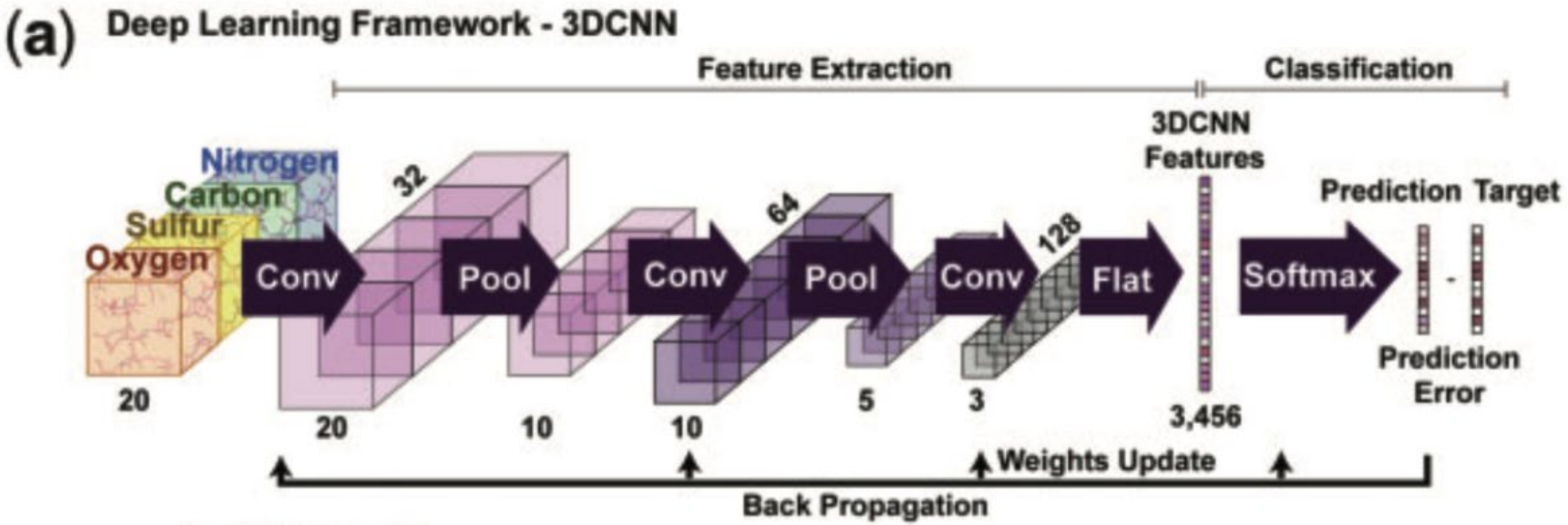

Feature Extraction
Information Integration
Classification
Deep Learning Framewrok - 3DCNN
Nitrogen
Carbon
Sulfur
Oxygen
20
Conv
100
18
Conv
200
16
Pool
200
8
Conv
400
6
Pool
400
3
Flat
3DCNN Features
10800
FC
1000
FC
100
Soft max
Prediction Target
20
20
Error
Weights update
Back propagation

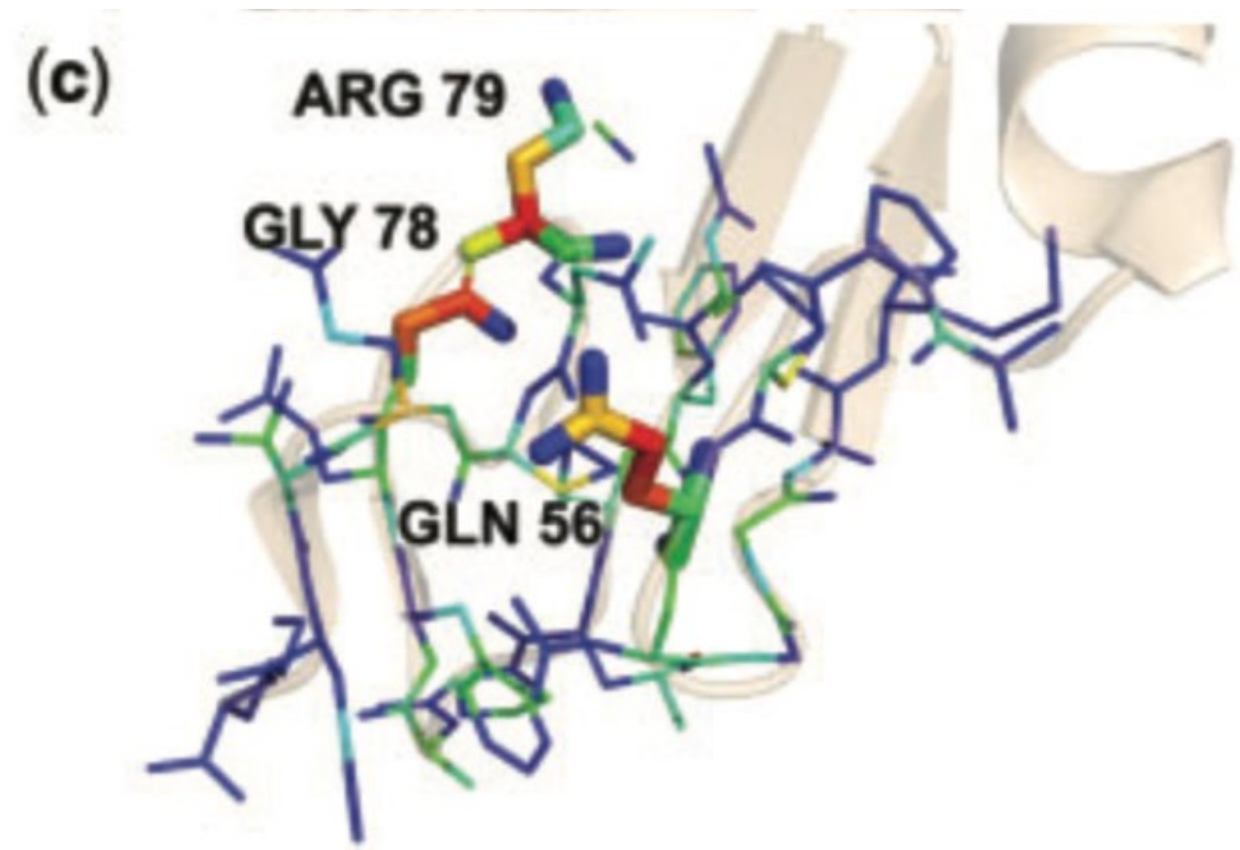
b

3DCNN-Test


P G W F Y V I C L M D N A E Q S T H K R
R K H T S Q E A N D M L C I V Y F W G P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Recognizing Epidermal Growth Factor Binding Site

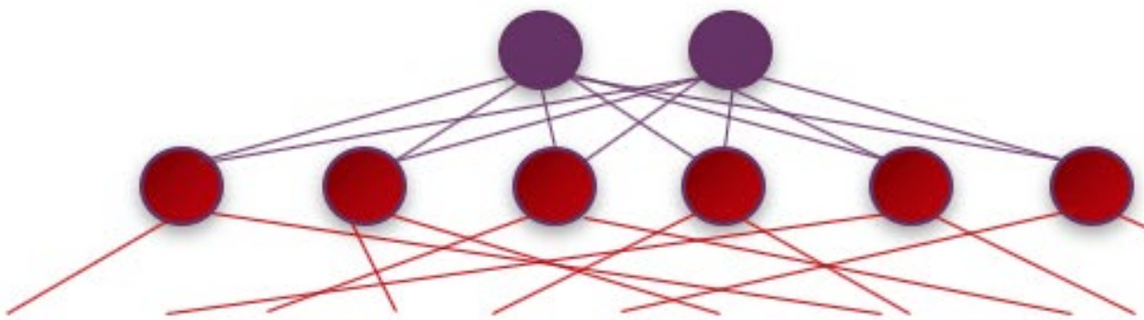




A Supervised Classifier for Binding Prediction

Binding Prediction

Softmax Classifier

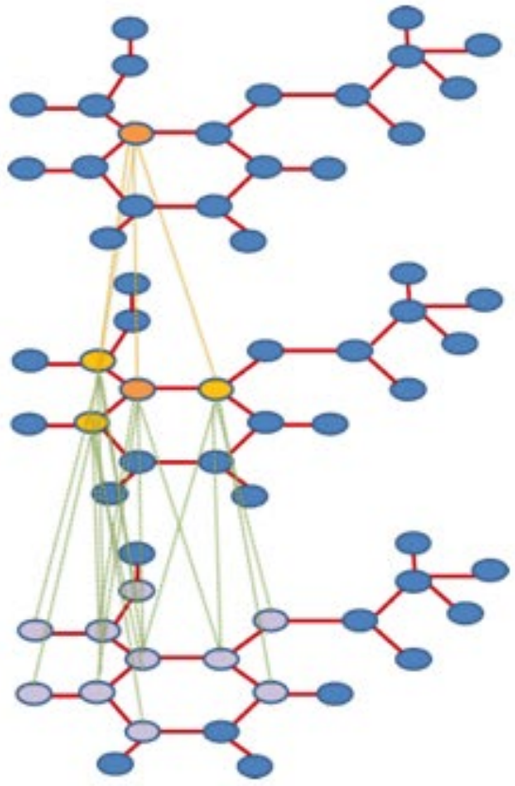


Interactions between pockets and molecules

SMALL MOLECULE REPRESENTATION

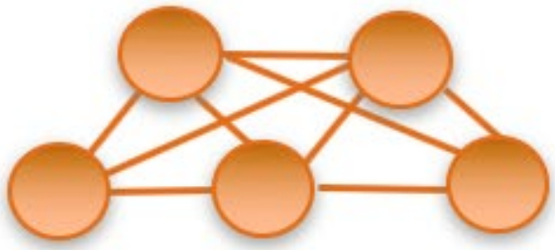
PROTEIN POCKET REPRESENTATION

Graph Convolution

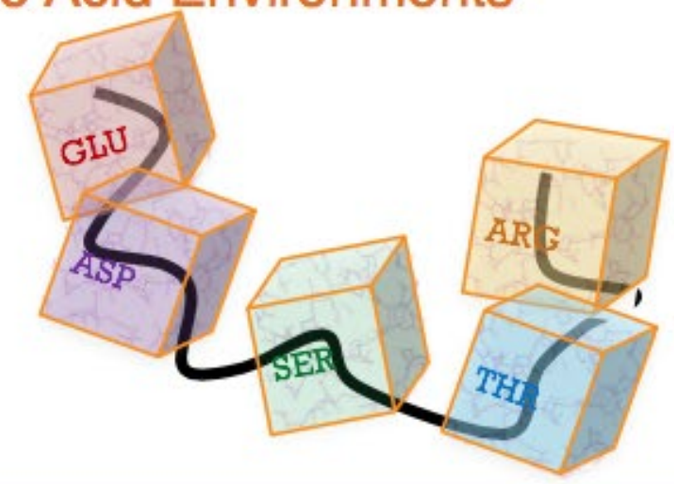


Duvenaud, David K., et al. 2015

Graph Convolution



Amino Acid Environments

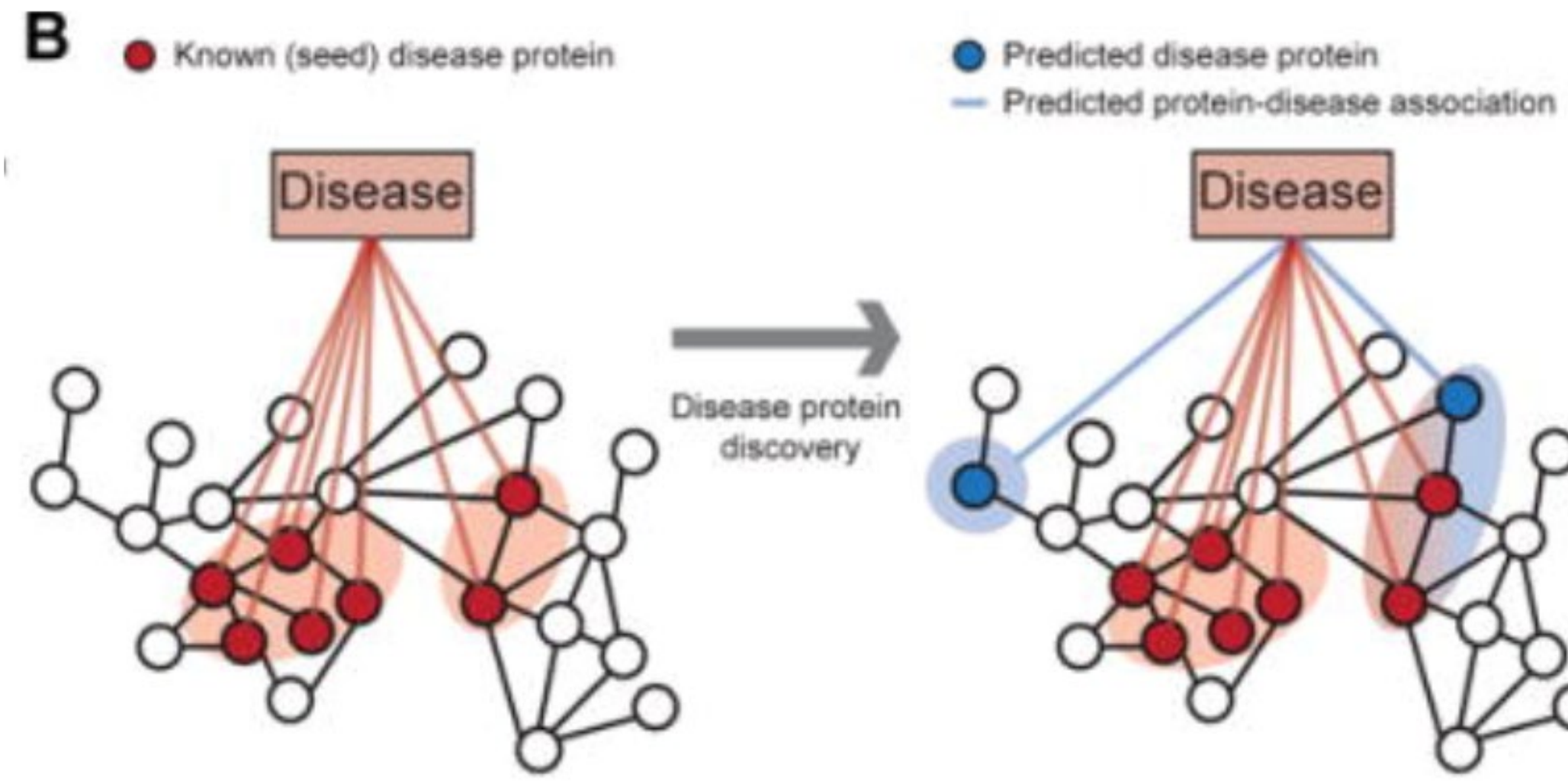


Deep Networks for predicting new drug targets

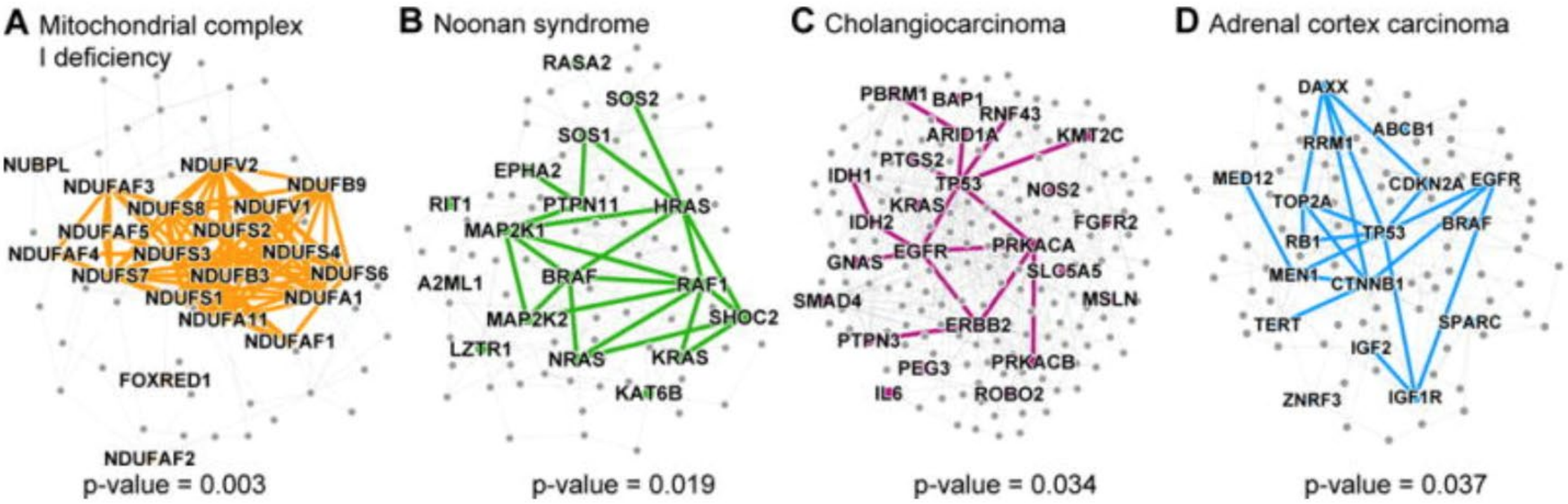
Same Network analyzed by two Deep Learning Methods.

Top shows neighborhoods based on connectivity of nodes in network

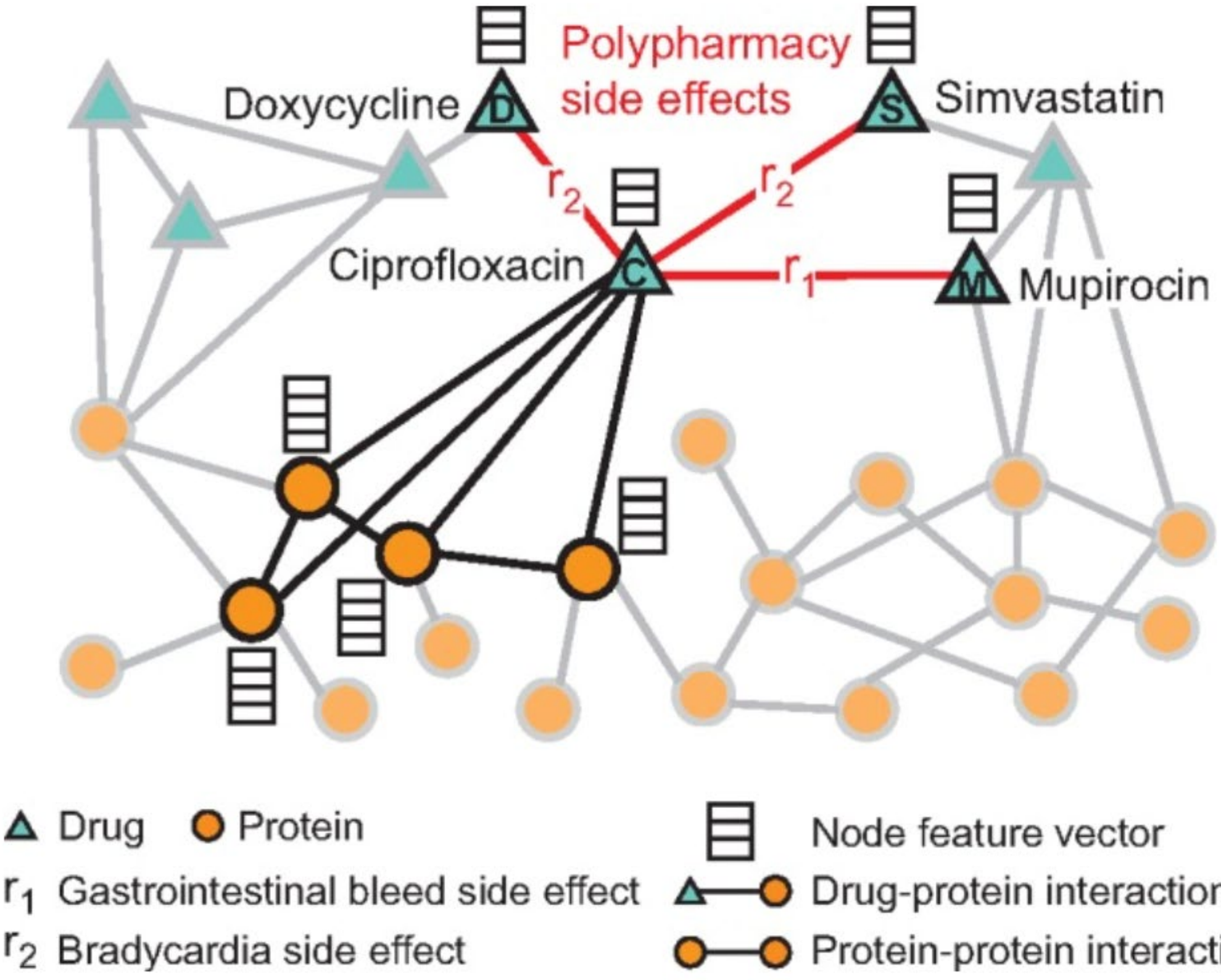
Bottom shows nodes that are similar based on the role that they play in the network



Agrawal M, Zitnik M, Leskovec J. Large-scale analysis of disease pathways in the human interactome. Pac Symp Biocomput. 2018;23:111-122. PMID: 29218874

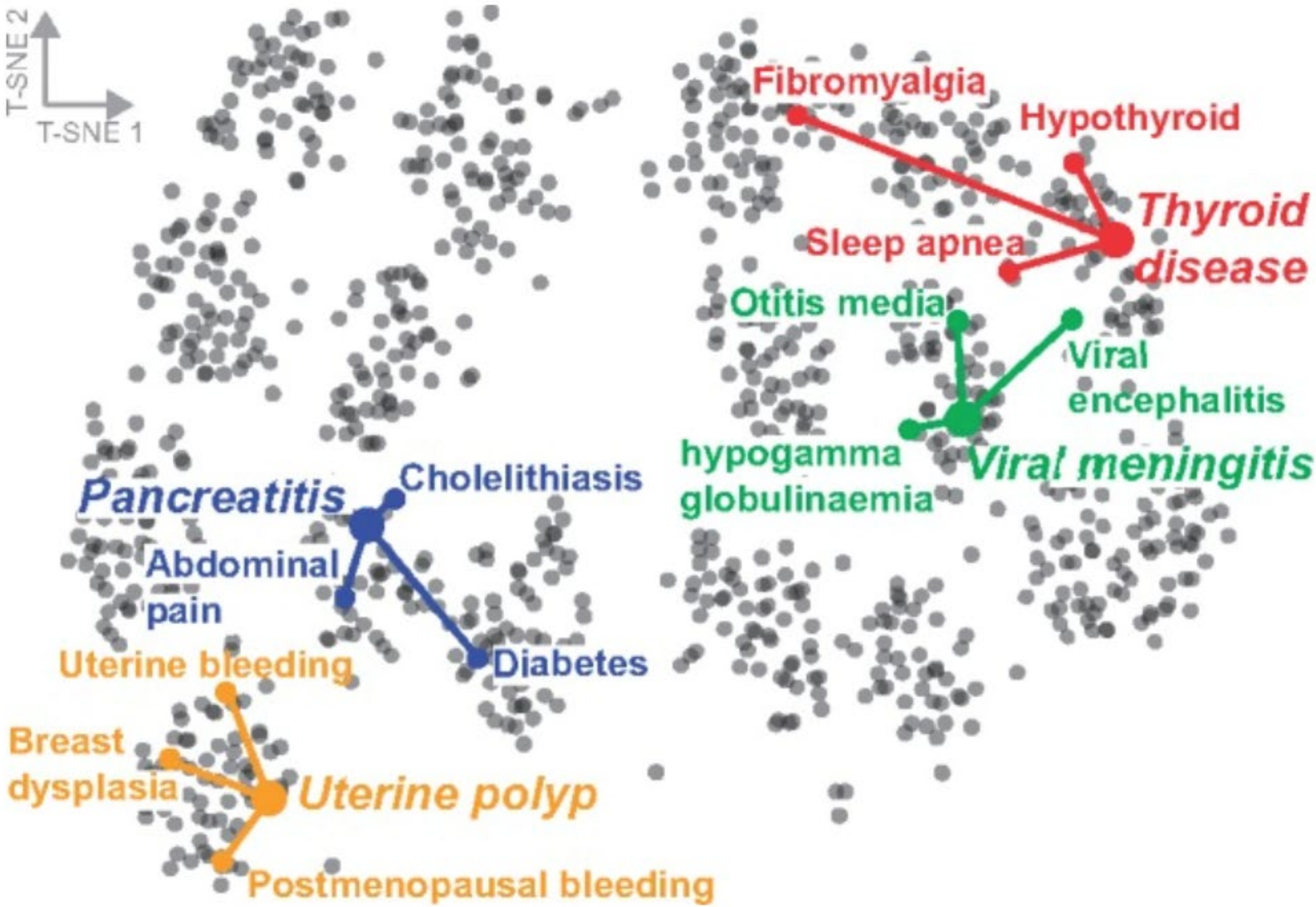

A Mitochondrial complex I deficiency
NUBPL
NDUFV2
NDUFAF3
NDUFB9
NDUFS8
NDUFV1
NDUFAF5
NDUFS2
NDUFAF4
NDUFS3
NDUFS4
NDUFS7
NDUFB3
NDUFS6
NDUFS1
NDUFA1
NDUFA11
NDUFAF1
FOXRED1
NDUFAF2
p-value = 0.003
B Noonan syndrome
RASA2
SOS2
SOS1
EPHA2
RIT1
PTPN11
HRAS
MAP2K1
A2ML1
BRAF
RAF1
MAP2K2
SHOC2
LZTR1
NRAS
KRAS
KAT6B
p-value = 0.019
C Cholangiocarcinoma
PBRM1
BAP1
RNF43
ARID1A
KMT2C
PTGS2
IDH1
TP53
NOS2
KRAS
IDH2
FGFR2
EGFR
PRKACA
GNAS
SLC5A5
SMAD4
MSLN
ERBB2
PTPN3
PEG3
PRKACB
IL6
ROBO2
p-value = 0.034
D Adrenal cortex carcinoma
DAXX
ABCB1
RRM1
MED12
CDKN2A
EGFR
TOP2A
RB1
TP53
BRAF
MEN1
CTNNB1
TERT
SPARC
IGF2
ZNRF3
IGF1R
p-value = 0.037

Deep Networks for predicting drug-drug interactions



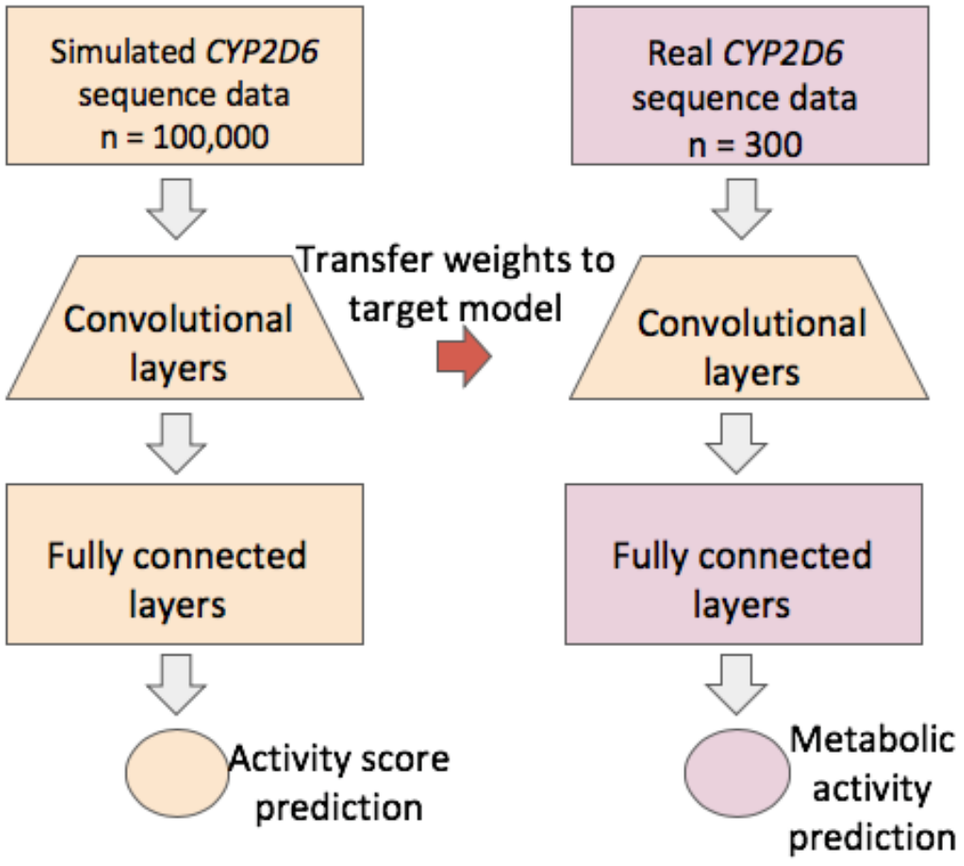
Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018 Jul 1;34(13):i457-i466. PMID: 29949996

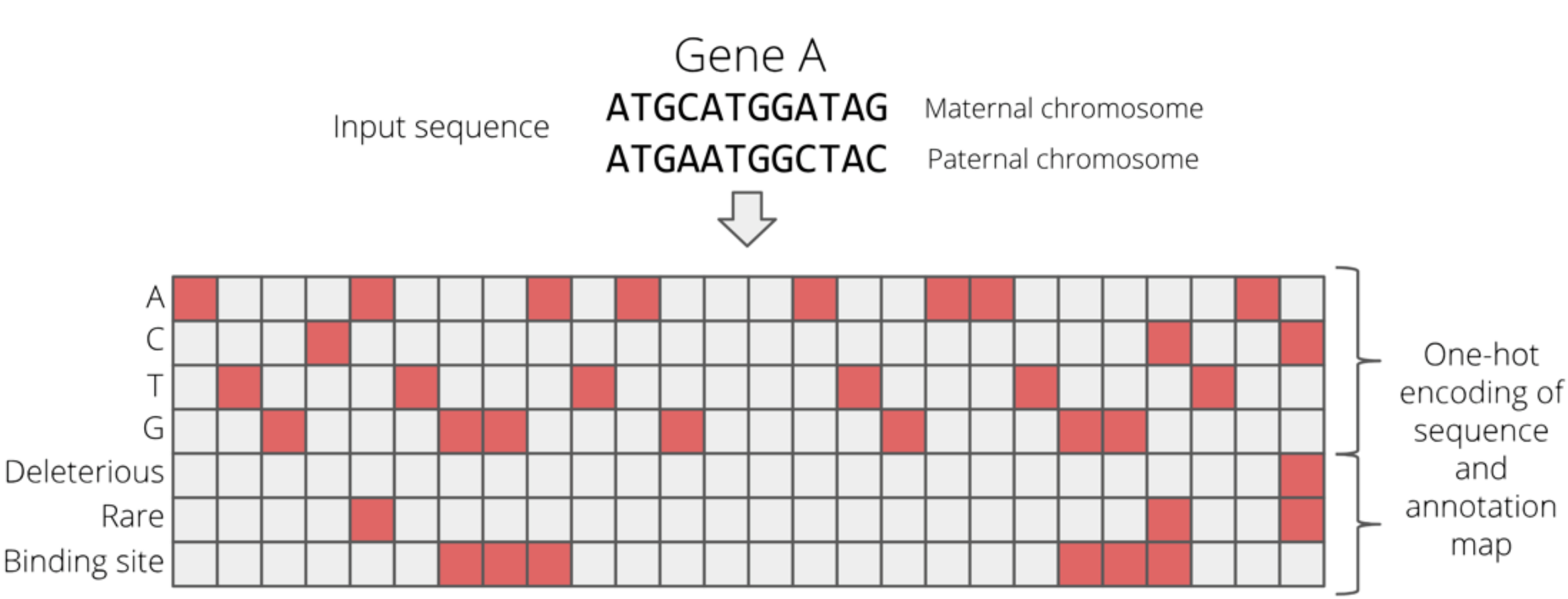
Co-occurring side effects in drug combinations

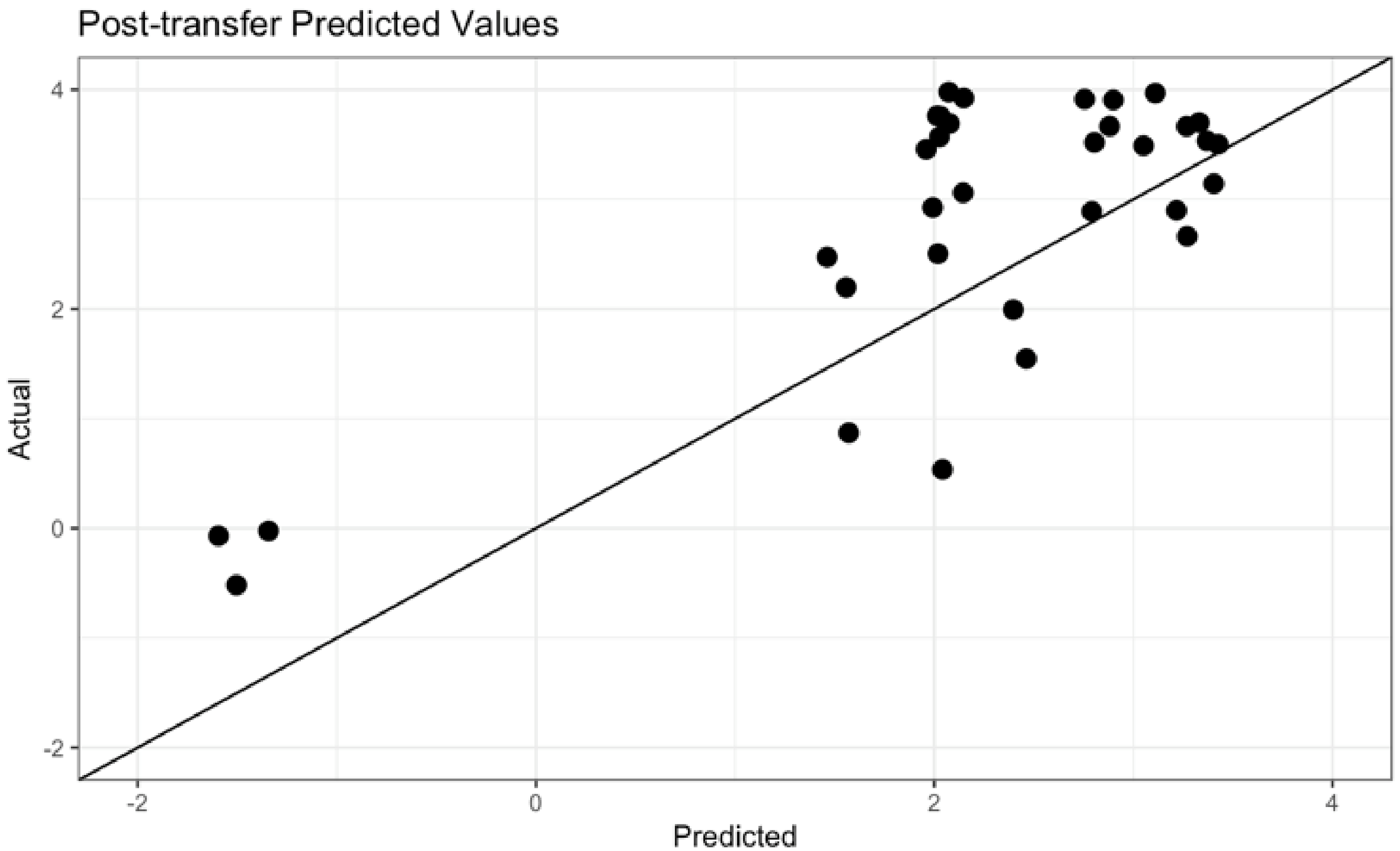


Deep Networks for predicting protein function

General strategy for CYP2D6 prediction and representation of DNA sequence data







Greg McInnes, unpublished results (with Erika Woodahl = U. Montana)

Summary

- Deep Learning is a type of machine learning
  - Heavy data requirements
  - Able to fit data extremely well
  - Need to validate rigorously with held-out data
- Early applications show promise for data-driven drug discovery, improving speed and quality of nominated drug targets.

Stanford University

Thanks!

russ.altman@stanford.edu

Support: NIH NIGMS, NIH NLM, NIH NCATS, FDA CERSI, Chan-Zuckerberg Biohub

CHAN ZUCKERBERG BIOHUB

FDA U.S. FOOD & DRUG ADMINISTRATION

NIH National Center for Advancing Translational Sciences

NIH NLM