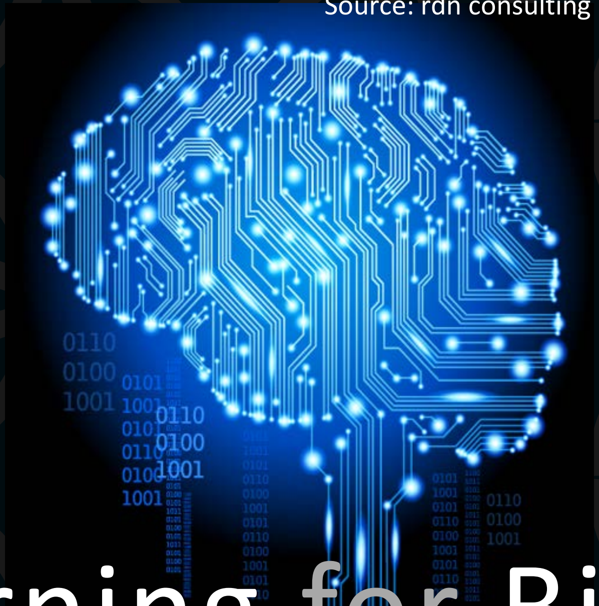
Source: rdn consulting



# Deep Learning for Biomedicine

## *Genomics and Drug Design*

**Truyen Tran**

Deakin University

truyen.tran@deakin.edu.au

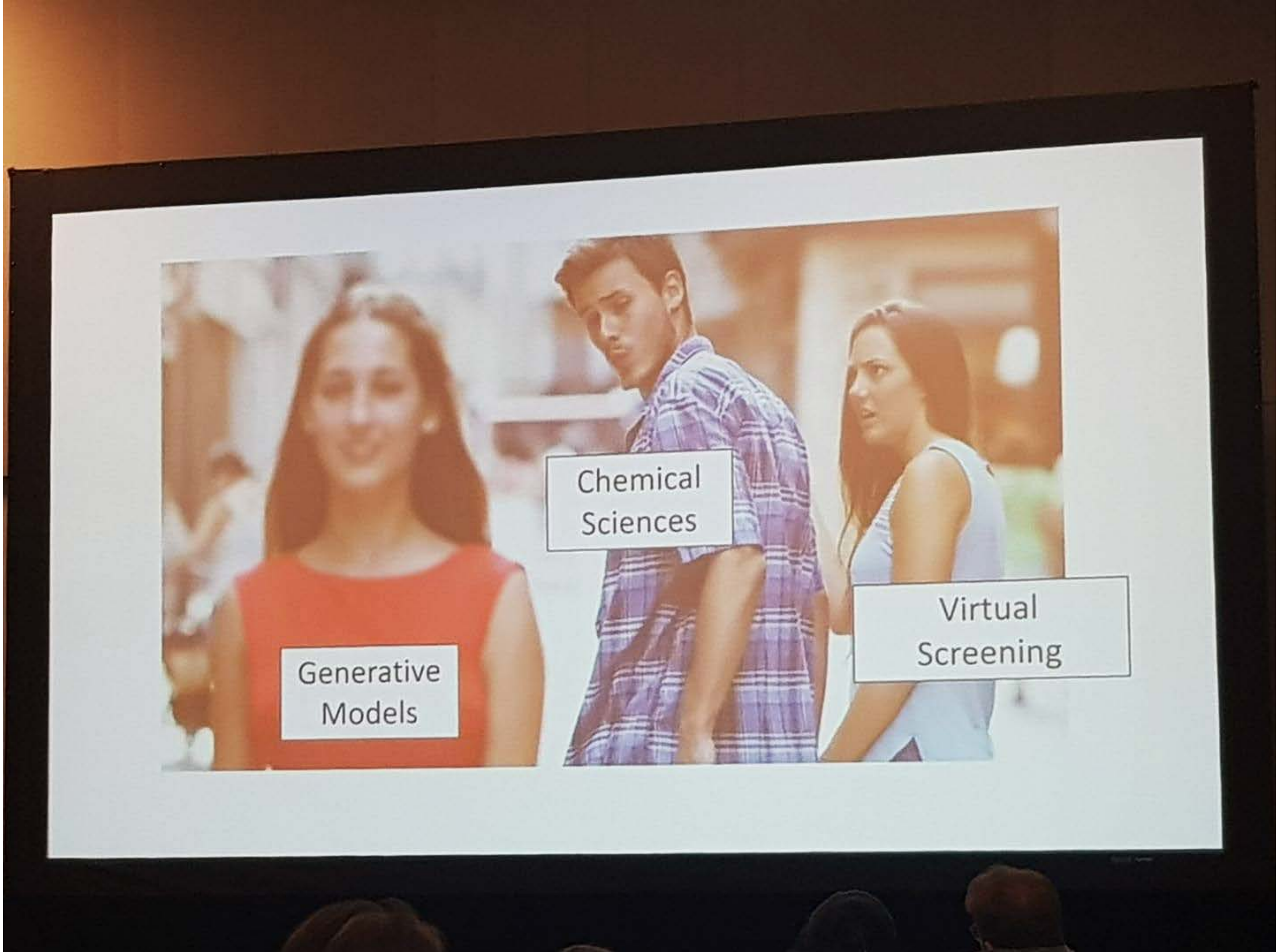
truyentran.github.io

@truyenoz

letdataspeak.blogspot.com

goo.gl/3jJ1O0

**Hanoi, Jan 2019**


Chemical Sciences
Virtual Screening
Generative Models

# Agenda

## Deep learning

- Neural architectures
- Generative models

## Genomics

- Nanopore sequencing
- Genomics modelling

## Drug design

- Bioactivity prediction
- Drug generation

## Future outlook

Epworth HealthCare

THE UNIVERSITY OF SYDNEY

GARVAN INSTITUTE

DEAKIN UNIVERSITY AUSTRALIA Worldly

Stanford University

Barwon Health

# Why now?

**High-impact** & **data-intensive**.

- Andrew Ng’s rule: impact on 100M+ people.
- Biomedicine is the only industry that will never shrink!

Ripe for innovations fuelled by deep learning techniques.

- Major recent advances and low hanging fruits are being picked.

Great **challenges**:

- High volume and high dimensional;
- Great privacy concerns;
- Need integrated approach to encompass great diversities.

It is the right time to join force with biomedical scientists!

healthpages.org

ase.edu

pharmacy.umaryland.edu

openi.nlm.nih.gov

marketingland.com

CTAAAGATGATCTTTAGTCCCGGTTCGAA
TCTTTAGTCCCGGTTGATAACACCAACC
GTAATACCAACCGGGACTAAAGATCCCG
GGGACTAAAGTCCCACCCCTATATATATG
TTCAAAATTTCTTCAAAAAAGAGGGGAG
GTGATTACATACAAATCGGAGGTGCCTA
TTTGTCATACTACATTTGCACCTATGTTTT
GTAAGTTGATGAGAGAGAAAATGTGTGT
SOCIAL MEDIA
BLOG
SMS
PubMed

**Big Rooms in Biomedicine**

# Machine learning = feature engineering = $$$

**$3M Prize, 3 years**

170K patients, 4 years worth of data

Predict length-of-stay next year

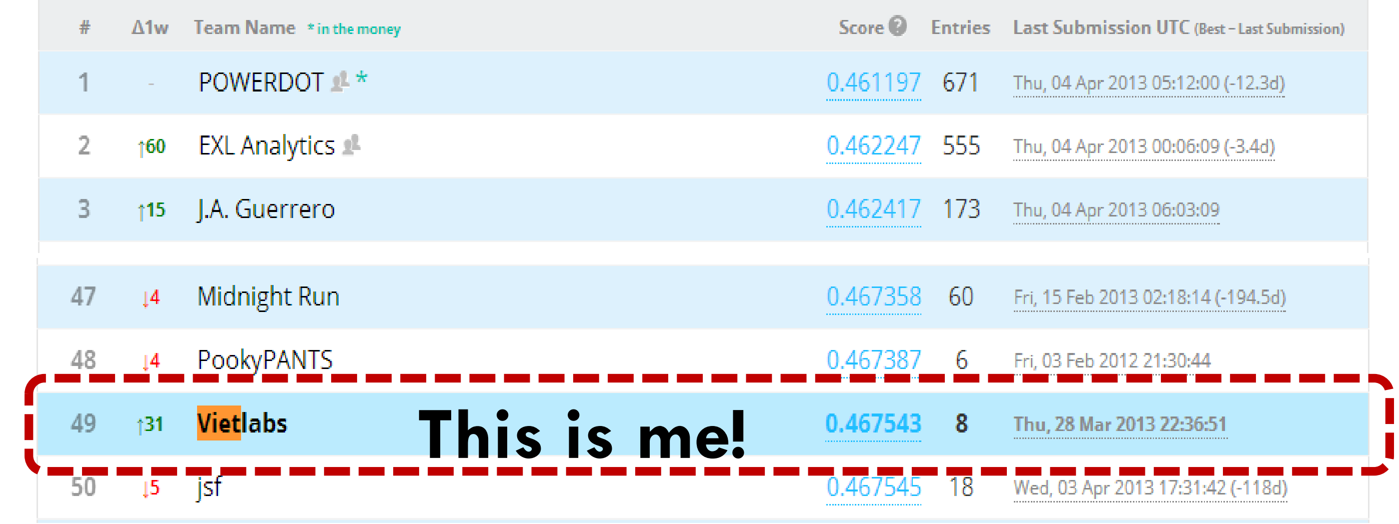
Not deep learning yet (early 2013), but strong ensemble needed → suggesting dropout/batch-norm

HERITAGE PROVIDER NETWORK HEALTH PRIZE

Truyen

Dashboard

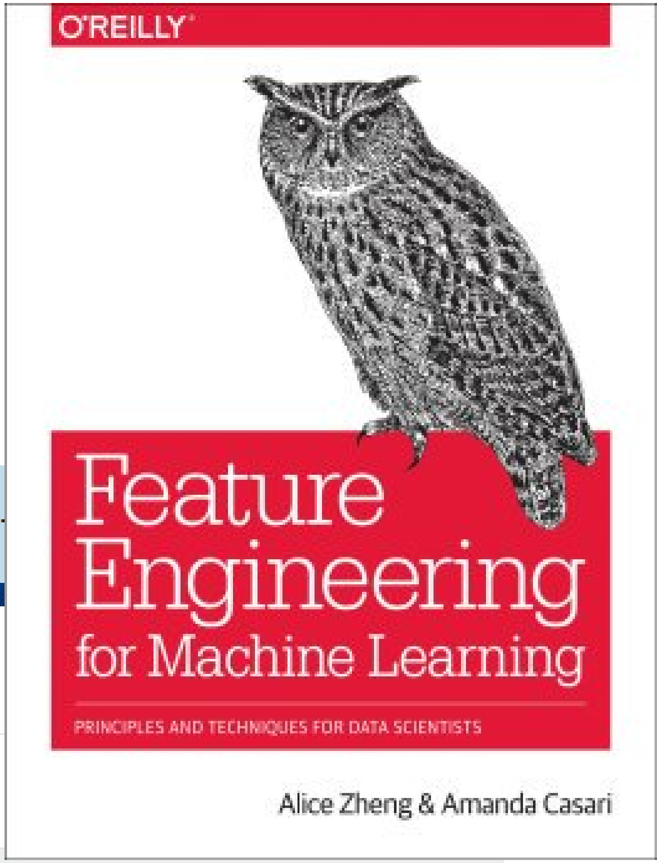
Leaderboard - Heritage Health Prize



This competition has completed. This leaderboard reflects the final standings.

| # | Δ1w | Team Name * in the money | Score | Entries | Last Submission UTC (Best – Last Submission) |
|---|---|---|---|---|---|
| 1 | - | POWERDOT * | 0.461197 | 671 | Thu, 04 Apr 2013 05:12:00 (-12.3d) |
| 2 | ↑60 | EXL Analytics | 0.462247 | 555 | Thu, 04 Apr 2013 00:06:09 (-3.4d) |
| 3 | ↑15 | J.A. Guerrero | 0.462417 | 173 | Thu, 04 Apr 2013 06:03:09 |
| 47 | ↓4 | Midnight Run | 0.467358 | 60 | Fri, 15 Feb 2013 02:18:14 (-194.5d) |
| 48 | ↓4 | PookyPANTS | 0.467387 | 6 | Fri, 03 Feb 2012 21:30:44 |
| 49 | ↑31 | Vietlabs | 0.467543 | 8 | Thu, 28 Mar 2013 22:36:51 |
| 50 | ↓5 | jsf | 0.467545 | 18 | Wed, 03 Apr 2013 17:31:42 (-118d) |

This is me!



O'REILLY

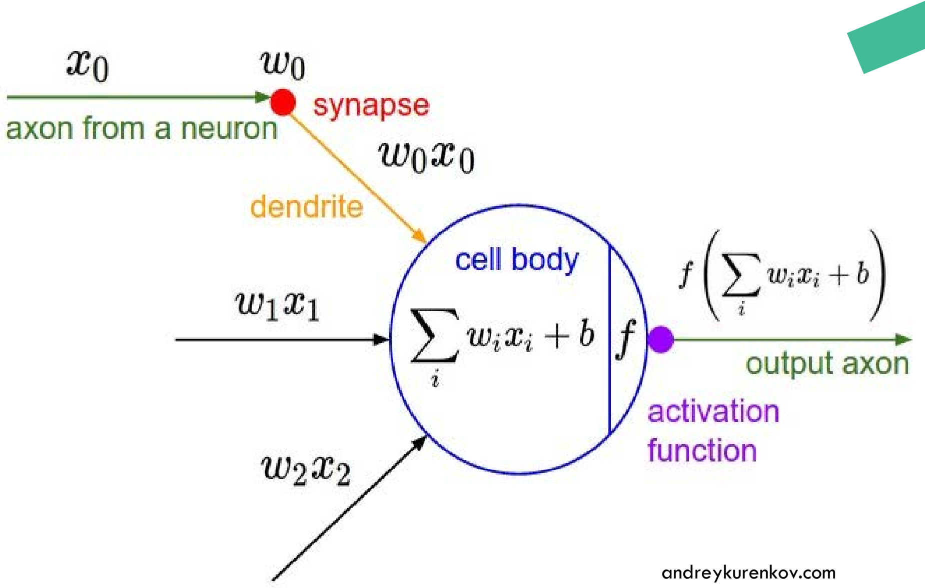
Feature Engineering for Machine Learning

PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS

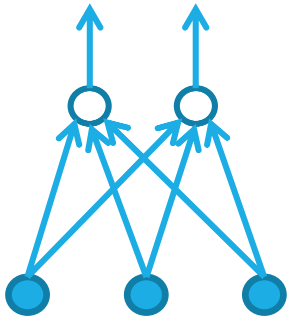
Alice Zheng & Amanda Casari

# Building block: Feature extractor

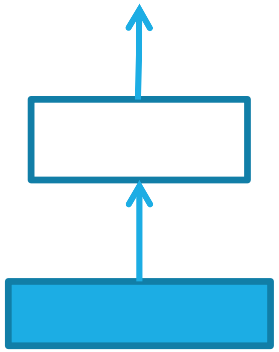
**Integrate-and-fire neuron**

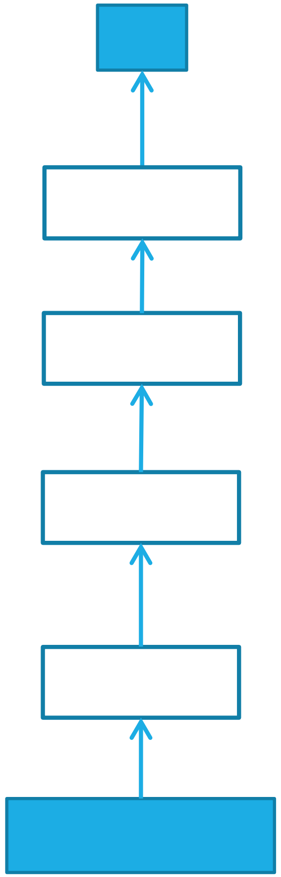


andreykurenkov.com



**Feature detector**





**Block representation**

# Building block: Recurrence

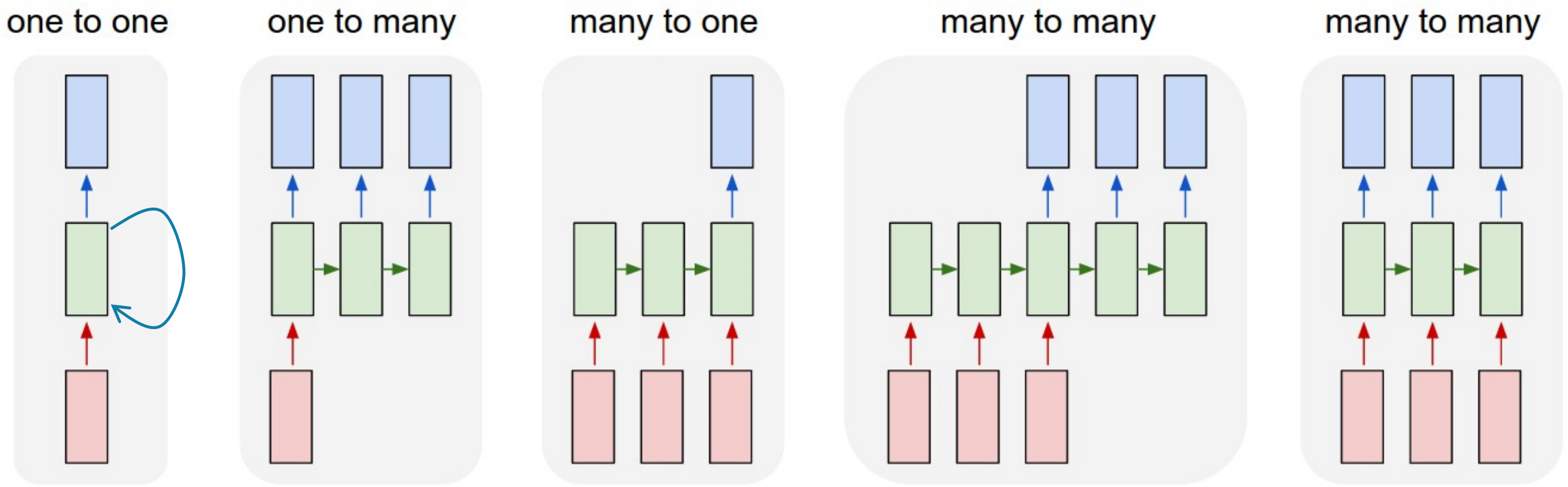
Classification

Image captioning

Sentence classification

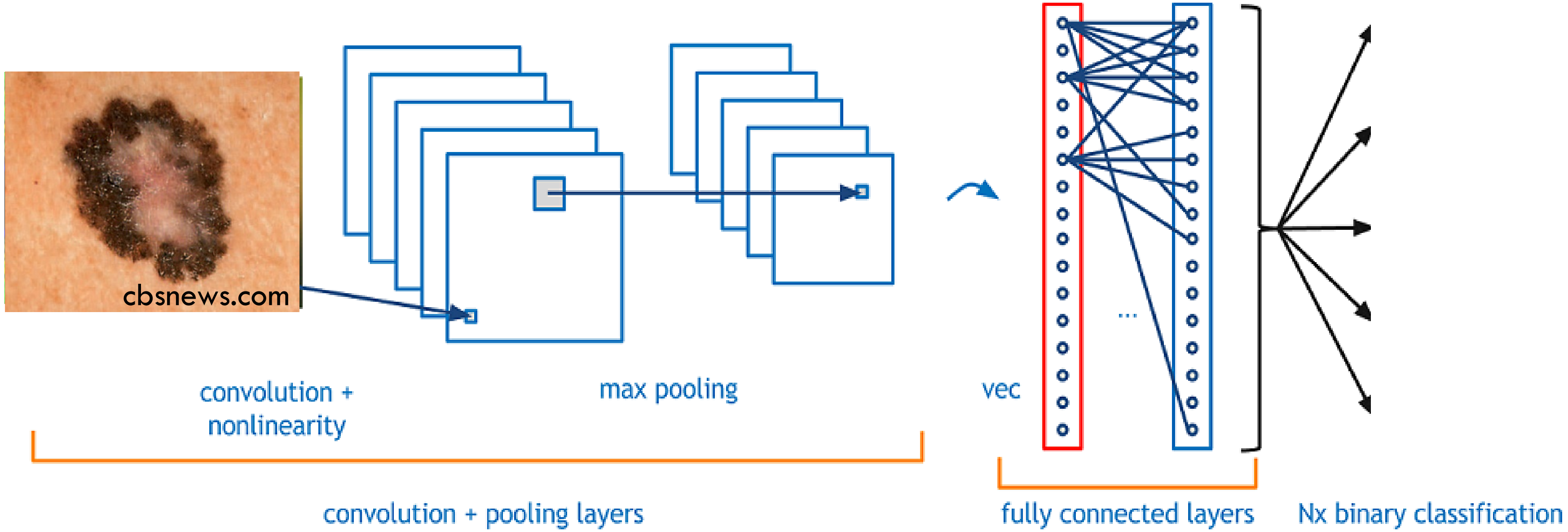
Neural machine translation

Sequence labelling



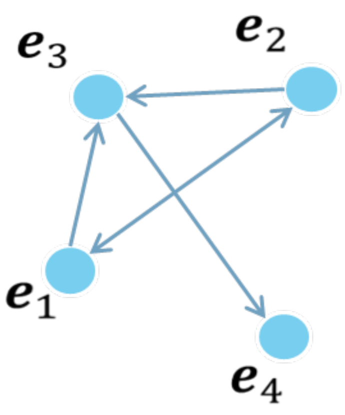
Source: http://karpathy.github.io/assets/rnn/diags.jpeg

# Building block: Convolution

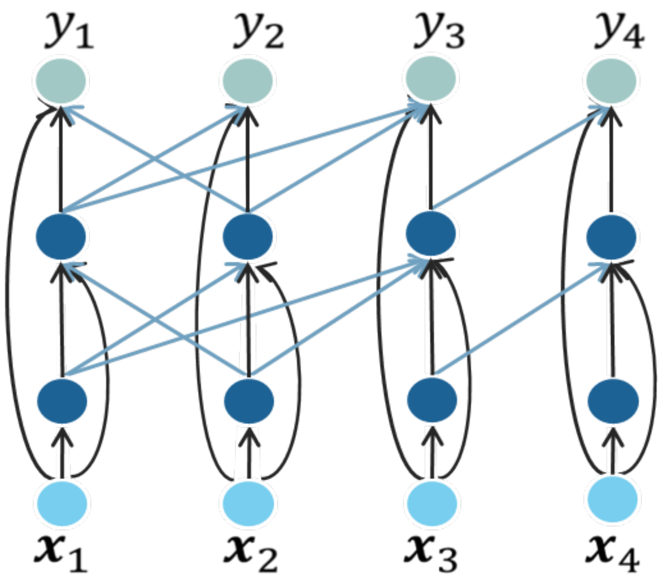


adeshpande3.github.io

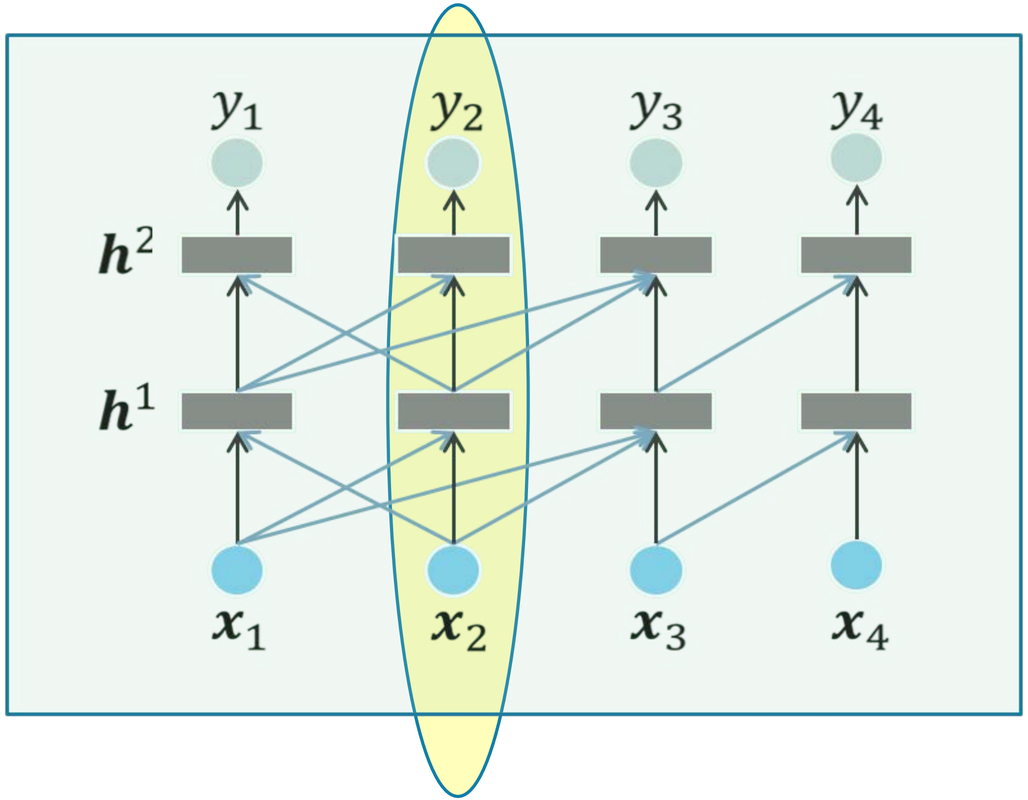
# Building block: Message passing



Relation graph



Stacked learning



Column nets

#REF: Pham, Trang, et al. "Column Networks for Collective Classification." *AAAI*. 2017.

# Supervised deep learning: steps

Step 0: Collect LOTS of high-quality data

- Corollary: Spend LOTS of time, $$ and compute power

Step 1: Specify the **computational graph** Y = F(X; W)

Step 2: Specify the loss L(W; D) for data D = {(X1,Y1), (X2,Y2), … }

Step 3: Differentiate the loss w.r.t. W (now mostly automated)

Step 4: Optimize the loss (a lot of tools available)

# Generative models

**Many applications:**

- Text to speech
- **Simulate data that are hard to obtain/share in real life (e.g., healthcare)**
- Generate meaningful sentences conditioned on some input (foreign language, image, video)
- Semi-supervised learning
- Planning

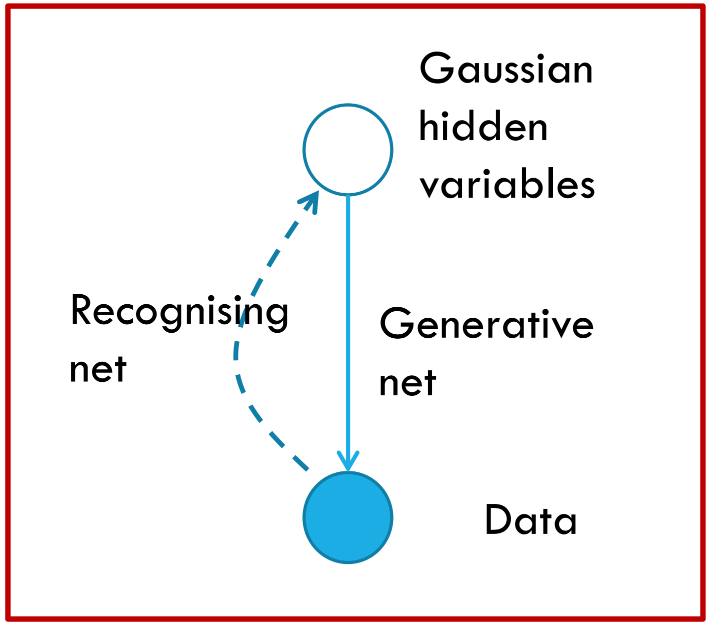
$$\mathbf{v} \sim P_{model}(\mathbf{v})$$

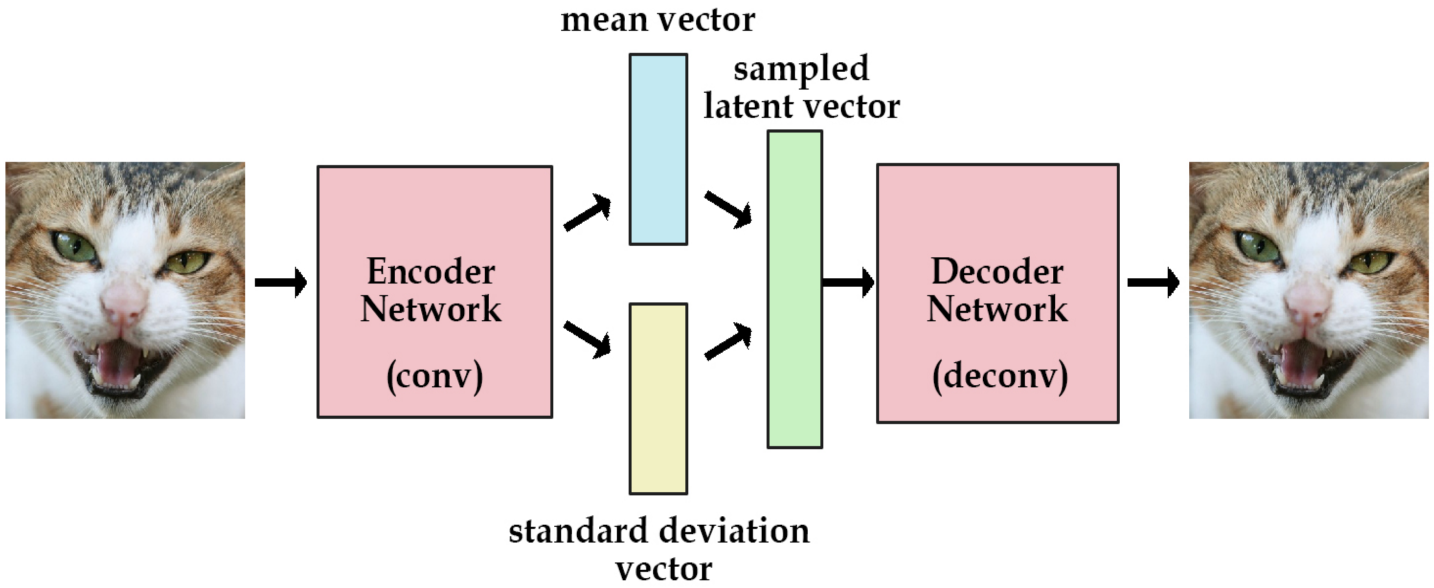
$$P_{model}(\mathbf{v}) \approx P_{data}(\mathbf{v})$$

# Variational Autoencoder

(Kingma & Welling, 2014)

Two separate processes: generative (hidden → visible) versus recognition (visible → hidden)

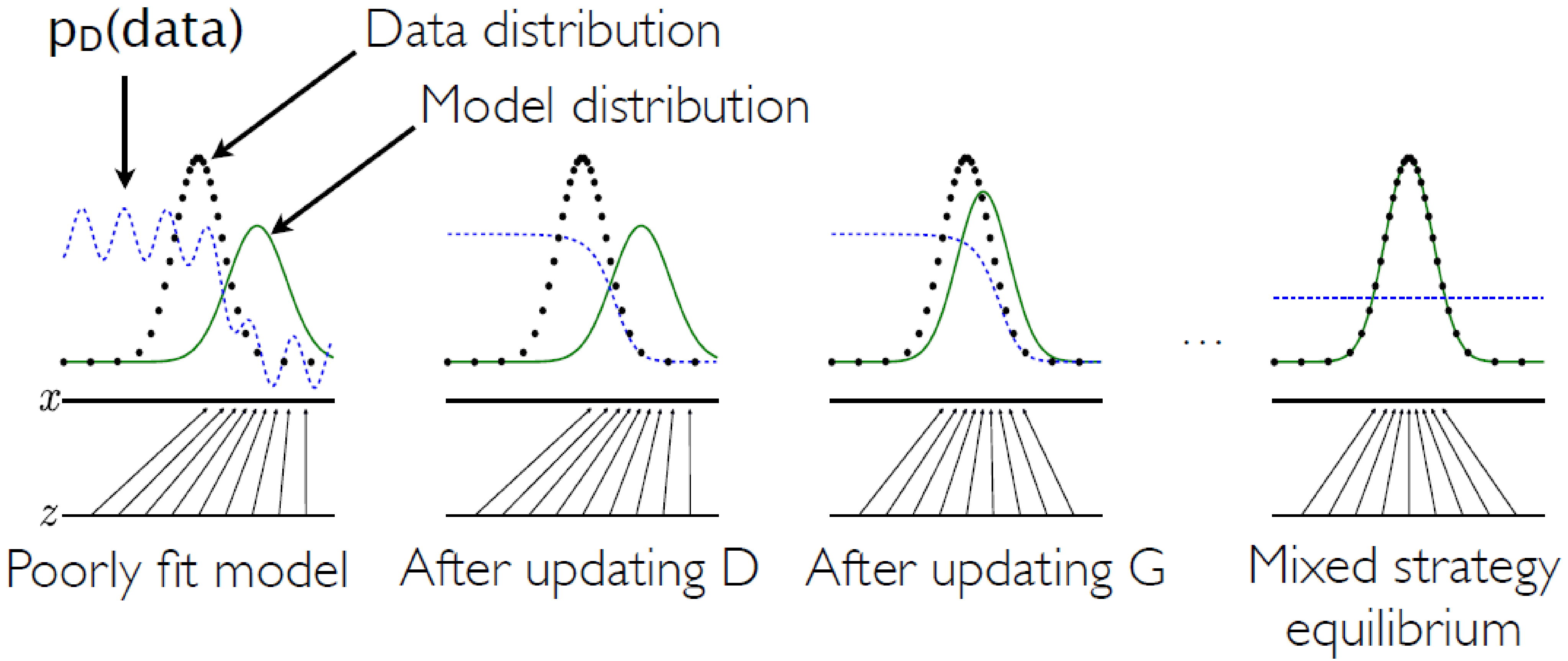




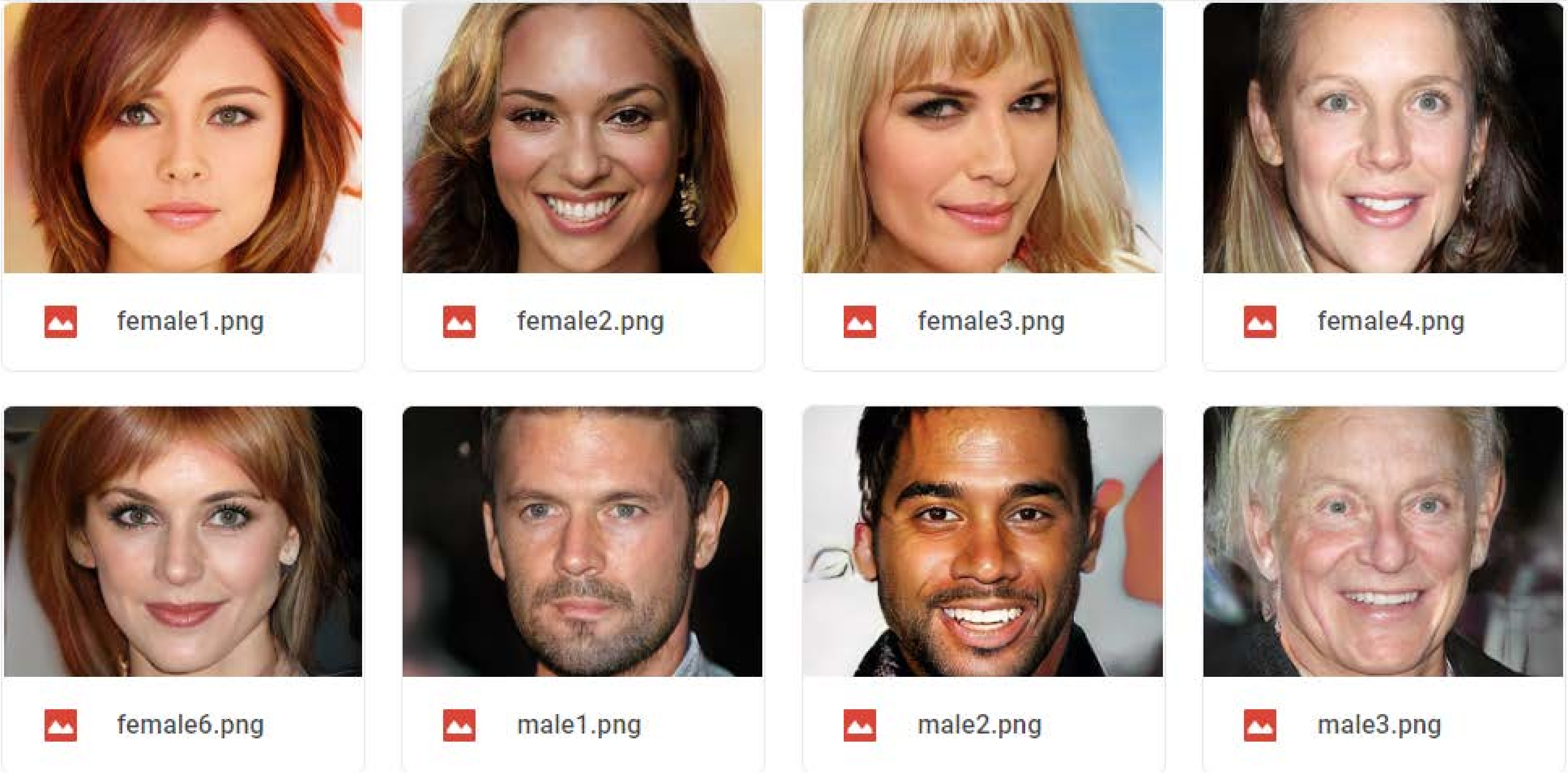
http://kvfrans.com/variational-autoencoders-explained/

# Generative adversarial networks

## (Adapted from Goodfellow’s, NIPS 2014)



# Progressive GAN: Generated images



Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196.*

# Agenda

## Deep learning

- Neural architectures
- Generative models

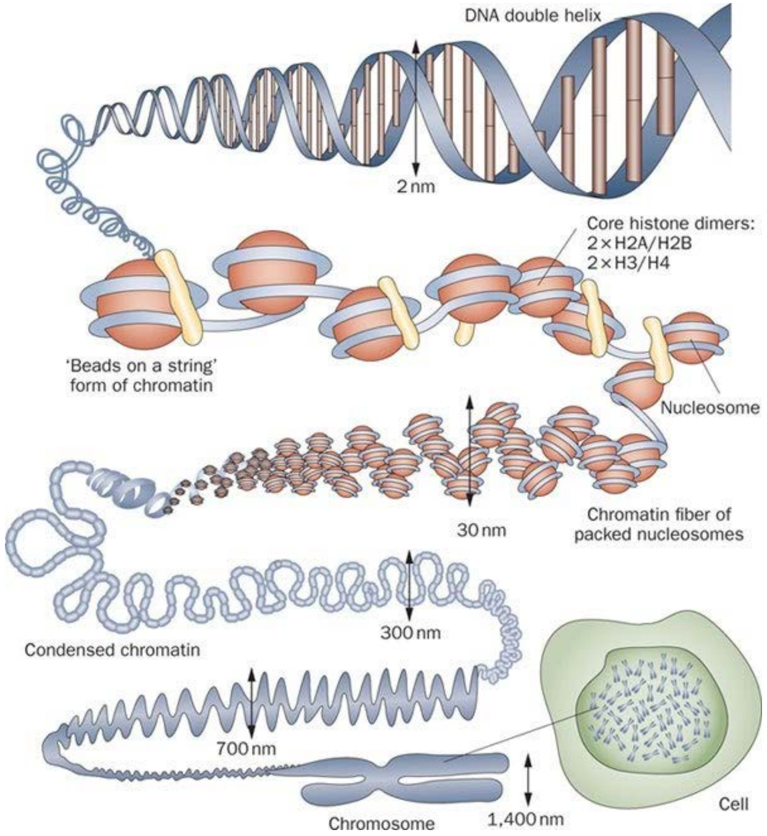
## Genomics

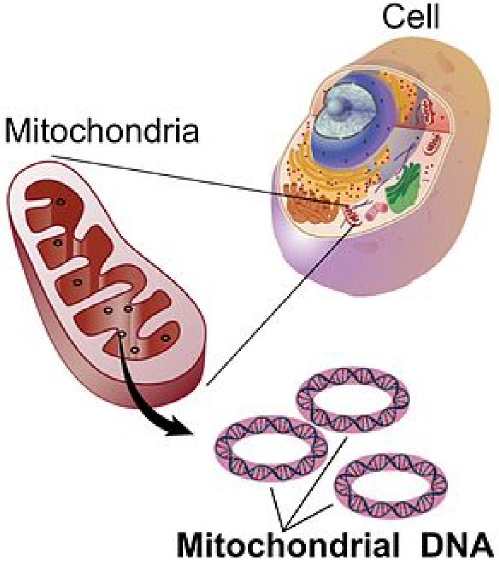
- Nanopore sequencing
- Genomics modelling

## Drug design

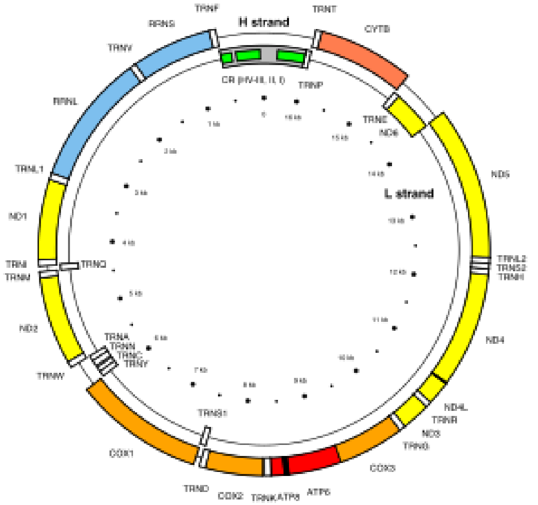
- Bioactivity prediction
- Drug generation

## Future outlook





MtDNA ring



https://qph.ec.quoracdn.net/main-qimg-2c39fede406d71fb534bbae6cc9b8aad-c
https://en.wikipedia.org/wiki/Mitochondrial_DNA

# Human genome

3 billion base-pairs (characters), 20K genes, 98% non-coding regions

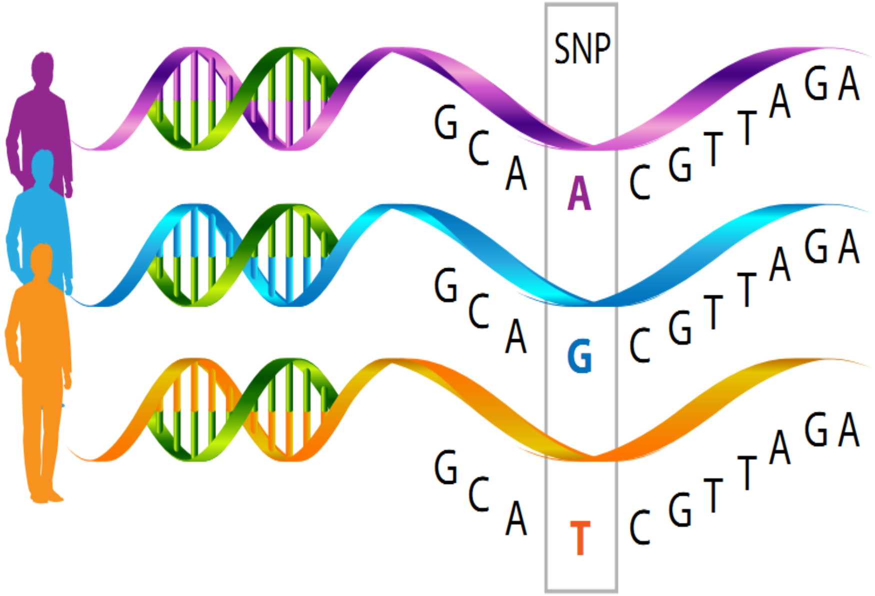
Any two random persons share 99.9% genome

The 0.1% difference is thought to account for all variations between us

- Appearance: Height (80% heritable), BMI, hair, skin colors
- IQ, education levels
- Genetic disorders such as cancers, bipolar, schizophrenia, autism, diabetes, etc.

Any two random persons share about 60% variations (SNV/SNP)

As we age, there are small mutations within our cells



https://neuroendoimmune.files.wordpress.com

# Sequencing

The first step is to read (sequence) the DNA/MtDNA, and represent the information as string of characters (A,C,G,T) in computer.

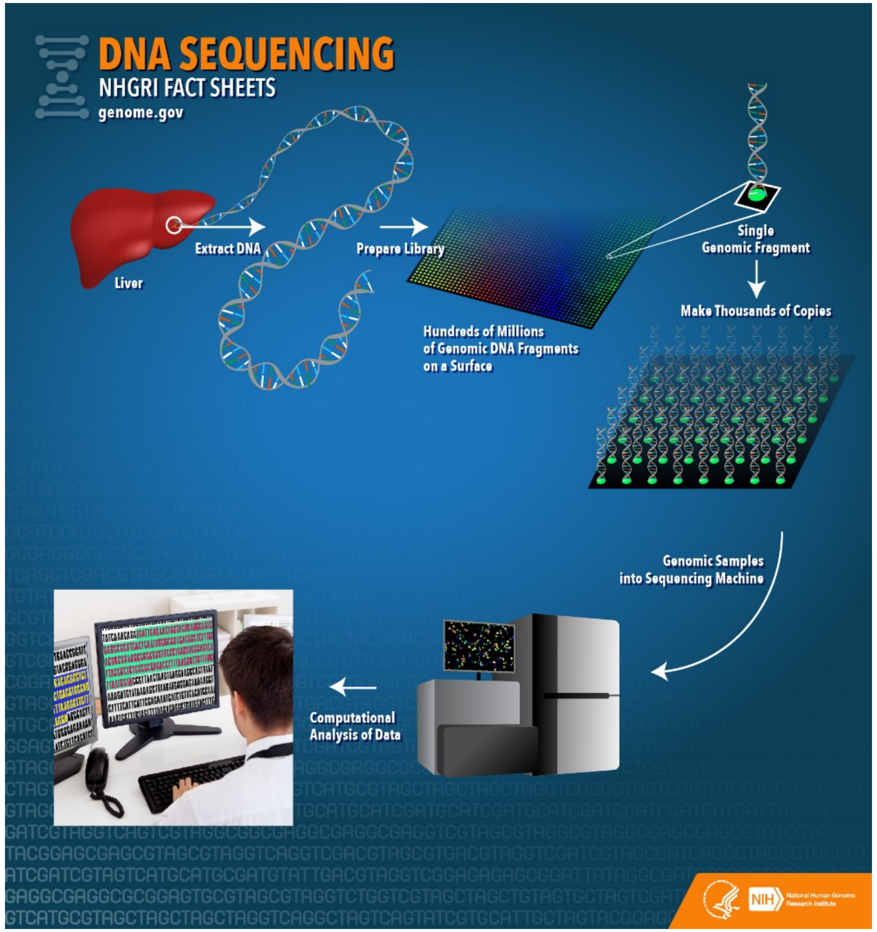
The most popular technique these days read short sequences (hundreds of characters), and align.

Each position is read typically at least 30 times to get enough confidence → Huge storage!!!

String alignment is then the key to final sequence → Need super-computer to do this fast.

A DNA sequence is compared against the reference genome. Only the difference (0.1%) need to be stored.

- This does not usually apply for MtDNA, as each cell has as many as 500 MtDNAs, they are slightly different! More different as we age.


DNA SEQUENCING
NHGRI FACT SHEETS
genome.gov
Liver
Extract DNA
Prepare Library
Hundreds of Millions of Genomic DNA Fragments on a Surface
Single Genomic Fragment
Make Thousands of Copies
Genomic Samples into Sequencing Machine
Computational Analysis of Data
NIH National Human Genome Research Institute

Source: https://www.genome.gov

Biologist

Bioinformatician

Physician

Health informatician

AI/ML/DL

# How does deep learning work for biomedicine?

Discovery

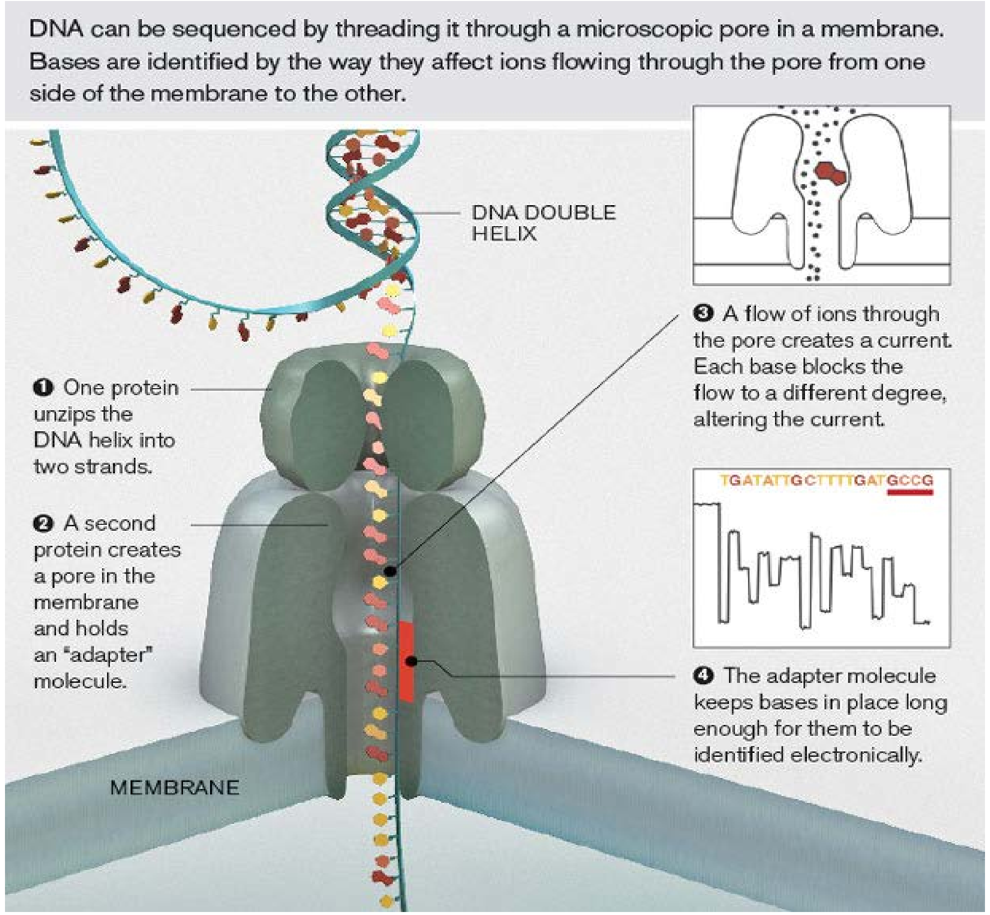
Diagnosis

Prognosis

Efficiency

http://hubpages.com/education/Top-Medical-Inventions-of-The-1950s
http://www.ctrr.net/journal
https://cdn1.iconfinder.com

# Nanopore sequencing ( electrical signals → A|C|G|T)



DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

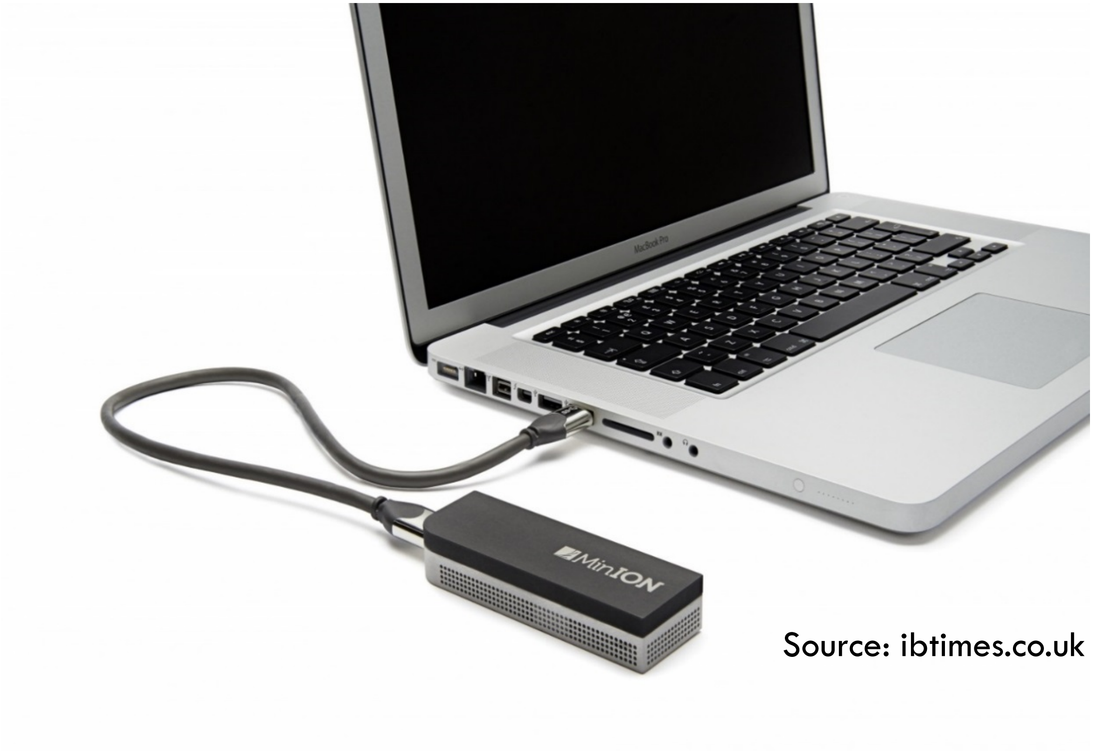
❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

Source: technologyreview.com



Source: ibtimes.co.uk

**Continuous segmentation & labelling**

# Deep architectures for nanopore sequencing

Aimed at real time recognition

**The setting is similar to speech recognition!**

- → The early days used HMMs. Now LSTMs.

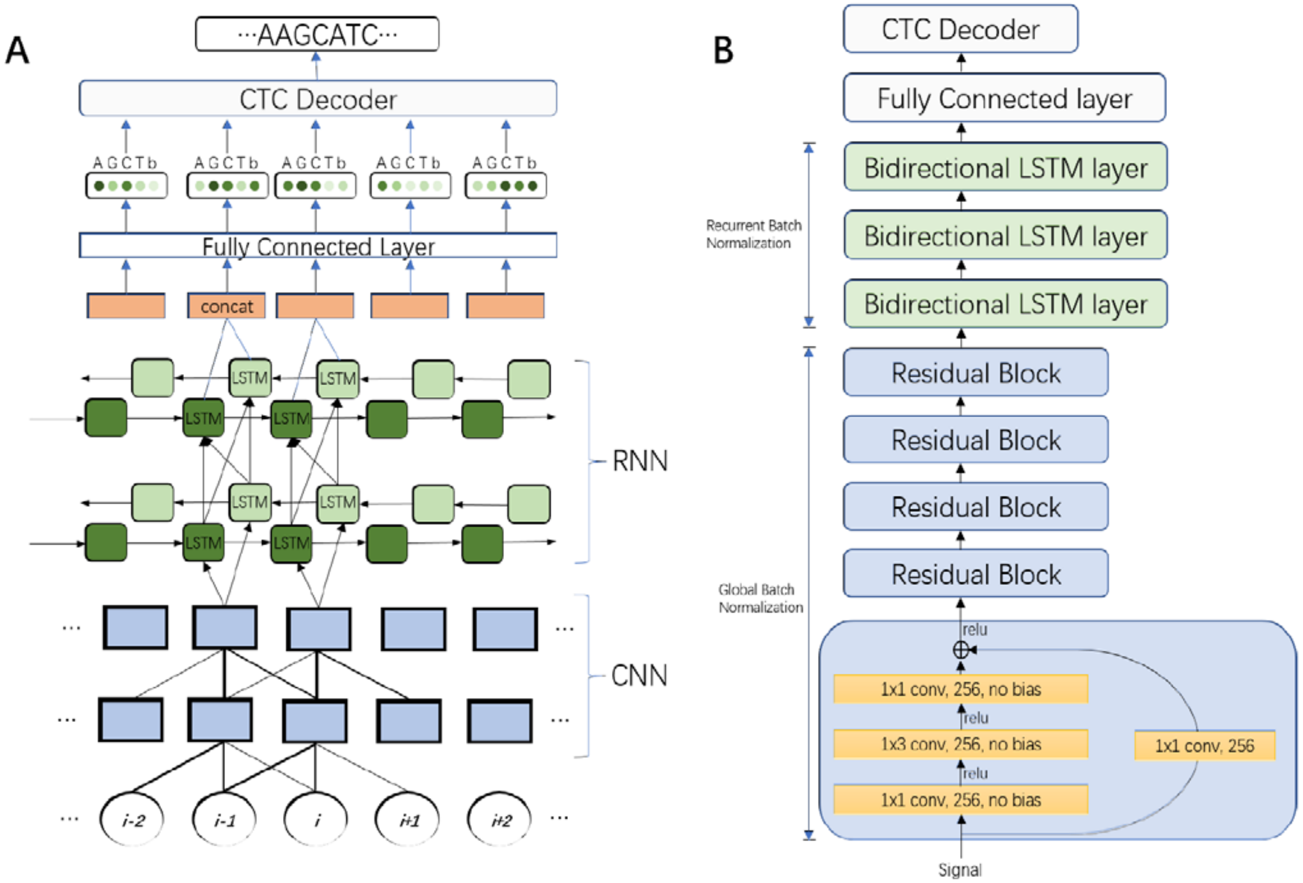
We will briefly review the latest:

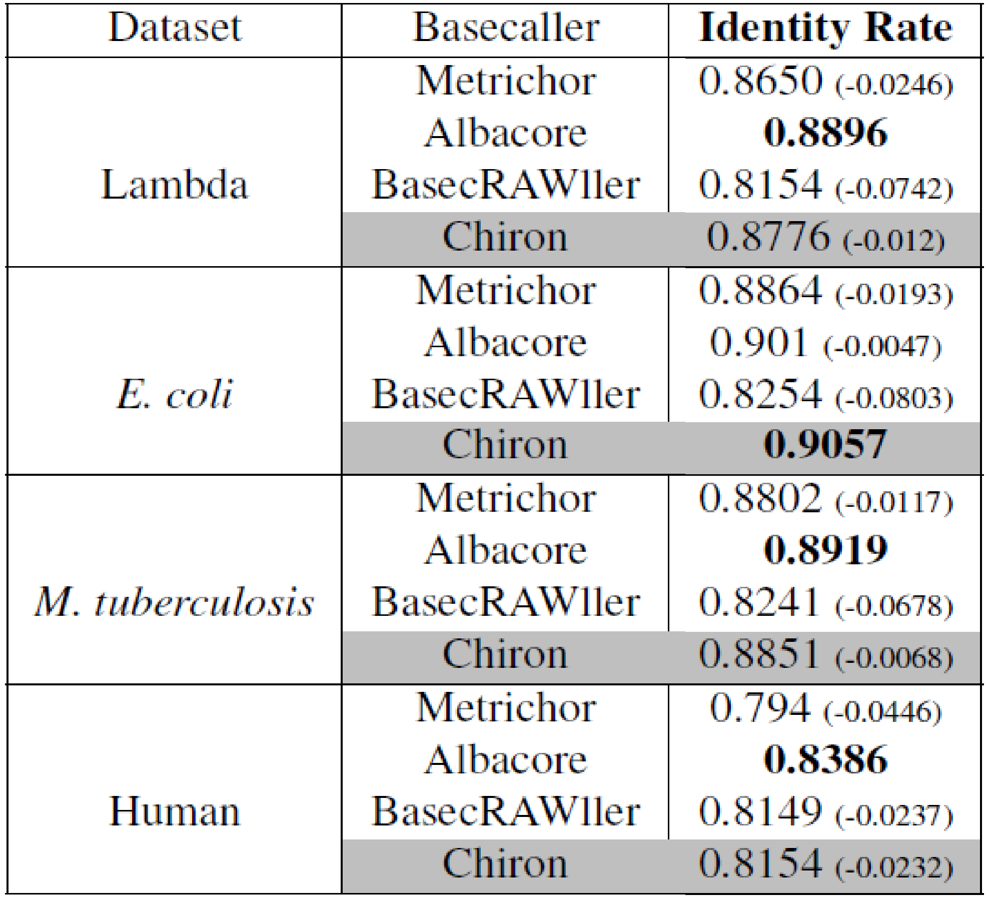
- **Chiron** (Teng et al., May 2018, UQ, Australia)

Other GRU/LSTM variants

- Nanonet (Oxford Nanopore Technologies, 2016)
- BasecRAWller (Stoiber & Brown, May 2017)
- **DeepNano** (Boza et al., June 2017, Comenius University in Bratislava, Slovakia)

# Chiron





| Dataset | Basecaller | **Identity Rate** |
|---|---|---|
| Lambda | Metrichor | 0.8650 (-0.0246) |
| | Albacore | **0.8896** |
| | BasecRAWller | 0.8154 (-0.0742) |
| | Chiron | 0.8776 (-0.012) |
| *E. coli* | Metrichor | 0.8864 (-0.0193) |
| | Albacore | 0.901 (-0.0047) |
| | BasecRAWller | 0.8254 (-0.0803) |
| | Chiron | **0.9057** |
| *M. tuberculosis* | Metrichor | 0.8802 (-0.0117) |
| | Albacore | **0.8919** |
| | BasecRAWller | 0.8241 (-0.0678) |
| | Chiron | 0.8851 (-0.0068) |
| Human | Metrichor | 0.794 (-0.0446) |
| | Albacore | **0.8386** |
| | BasecRAWller | 0.8149 (-0.0237) |
| | Chiron | 0.8154 (-0.0232) |

#REF: Teng, Haotien, et al. "Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning", GigaScience, Volume 7, Issue 5, 1 May 2018, giy037.

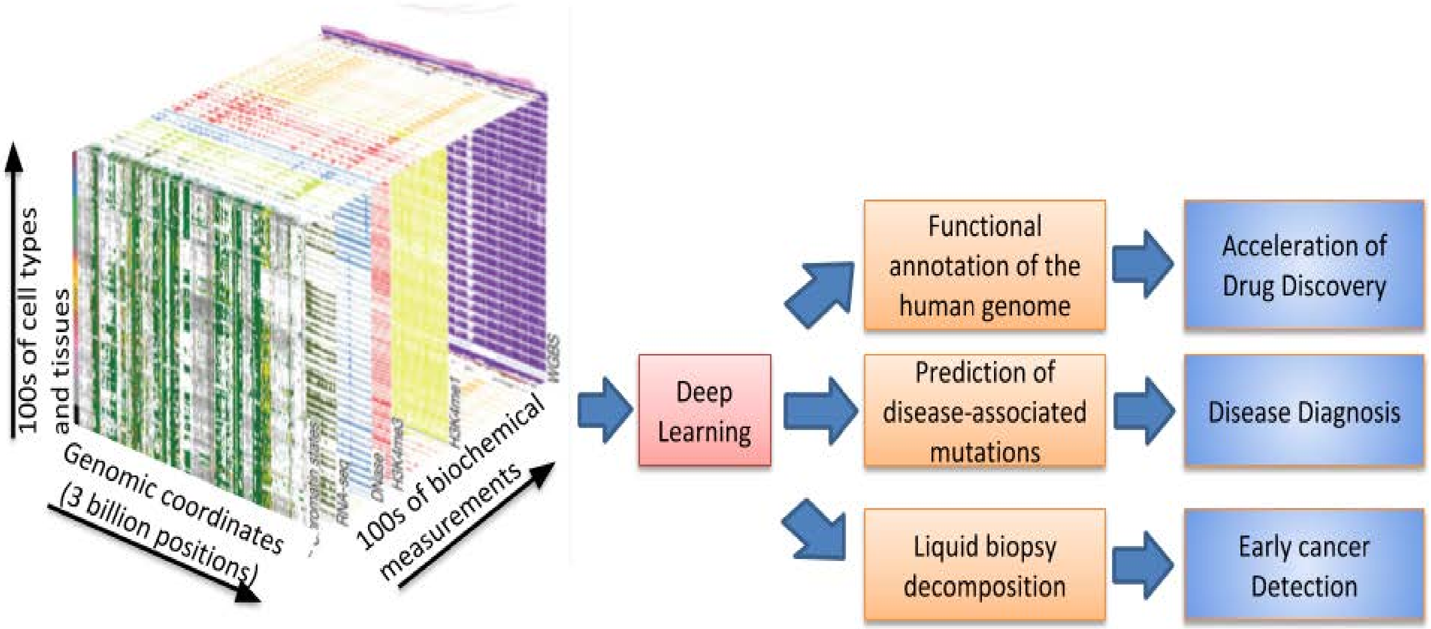
# Other recent works

Li, Yu, et al. "DeepSimulator: a deep simulator for Nanopore sequencing." *Bioinformatics* 1 (2018): 10.

Wick, Ryan R., Louise M. Judd, and Kathryn E. Holt. "Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks." *PLoS computational biology* 14.11 (2018): e1006583.

Wang, Sheng, et al. "WaveNano: a signal-level nanopore base-caller via simultaneous prediction of nucleotide labels and move labels through bi-directional WaveNets." *Quantitative Biology* 6.4 (2018): 359-368.

# Opportunities for Deep Learning in Genomics


100s of cell types and tissues
Genomic coordinates (3 billion positions)
100s of biochemical measurements
RNA-seq
DNase
H3K4me3
H3K4me1
WGBS

Deep Learning

Functional annotation of the human genome

Prediction of disease-associated mutations

Liquid biopsy decomposition

Acceleration of Drug Discovery

Disease Diagnosis

Early cancer Detection

Genetic diagnostics

Refining drug targets

Pharmaceutical development

Personalized medicine

Better health insurance

Synthetic biology

https://towardsdatascience.com/opportunities-and-obstacles-for-deep-learning-in-biology-and-medicine-6ec914fe18c2
https://www.oreilly.com/ideas/deep-learning-meets-genome-biology

# Some AI problems

DNA is a book, easy to read (costs less than $1K to sequence), extreme difficult to comprehend.

- It has 3B characters (A,C,T,G), 46 volumes (chromosomes), 20K chapters.
- The longest book has less than 10M characters, 13 volumes (“A la recherche du temps perdu" (In Search of Lost Time), by Marcel Proust, 2012) – as recognized by Guinness World Records.

Short sequences (100 chars) are predictive of protein binding, also gene start/end.

Proteins are big 3D graphs interacting with the 1D-2D strings (DNA, RNA), and other proteins & drugs (which are graphs themselves).

Long chains of influence, from SNP to cell, tissue and organ functions.

Viruses can be generated/edited on computer, hence discrete sequence generation problem.

# Filling the genotypes → phenotypes gap

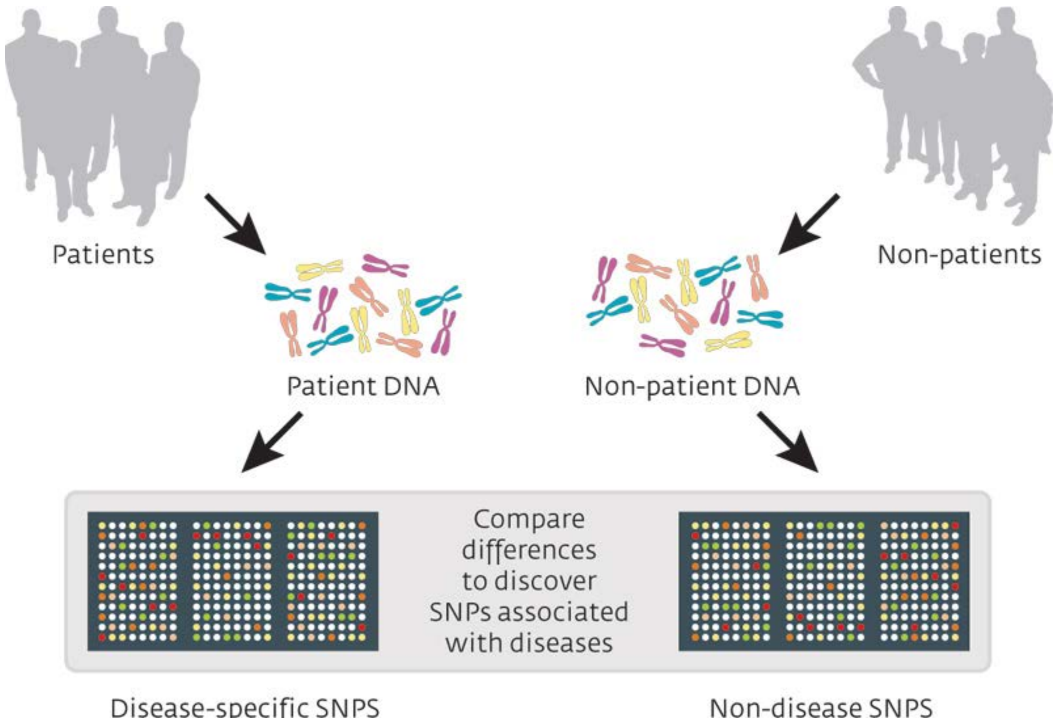
## Ultimate goals:

- Estimating explained variance in inheritability
- Discover risk factors
- Predicting individual phenotypes: Height, Glucose, BMI, IQ, Edu, Mental, Cancers…

## Some paths under investigation

- Predicting the bio of the cells, DNA + MtDNA, and more
- Statistical modeling of genetic architectures, e.g., Bayesian, mixed linear models, Gaussian Processes.
- Motif modeling with DNA/RNA/protein, e.g., predict binding sites
- Developing data-efficient techniques for genomics
- Integrating multimodalities

# GWAS: Genome-Wide Association Study



Source: http://vignette4.wikia.nocookie.net

Setting:

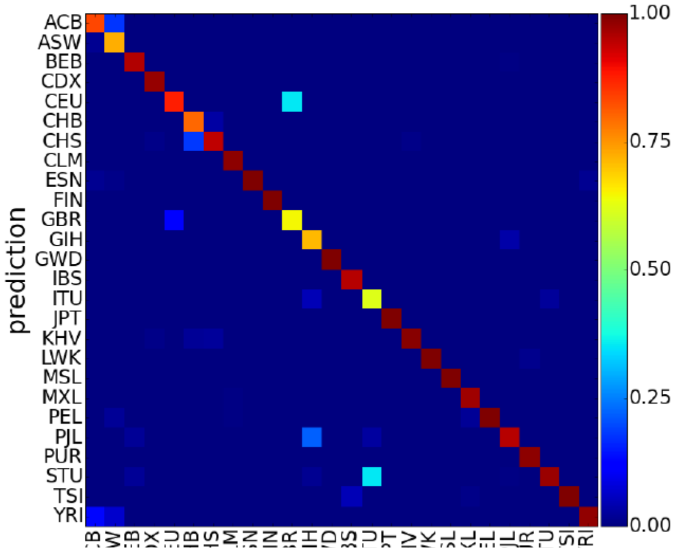
- For each DNA, only differences from a reference genome are recorded.
- The differences are SNPs, one per dimension.

**Problems**

- Very high dimensional (typically hundreds of thousands), low sample size (typically hundreds)
- Missing/unreliable data
- Typically very weak association
- Combating the False Discovery Rate (FDR) due to multiple parallel hypotheses: Individual *p*-value must be extremely small, e.g. 5×10e-8

# Diet networks for GWAS

#REF: Romero, Adriana, et al. "Diet Networks: Thin Parameters for Fat Genomic." *arXiv preprint arXiv:1611.09340* (2016).



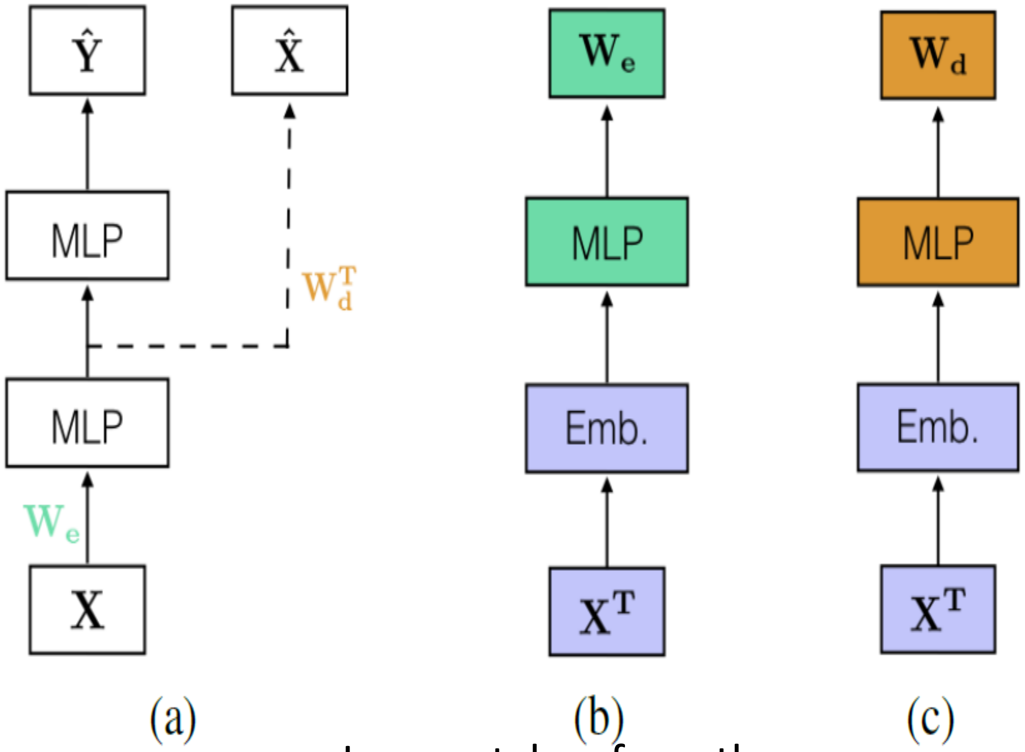
Use a “hypernet” to generate the main net.

Features are embedded (not data instance).

Unsupervised autoencoder as regularizer.

Works well on country prediction on the 1000 Genomes Project dataset.

- But this is a relatively easy problem. PCA, even random subspace can do quite well!



Images taken from the paper

# GWAS: Challenges

We are detecting rare events!!!

Results hard to replicate across studies.

- Model stability?

SNP → phenotypes seem impossible.

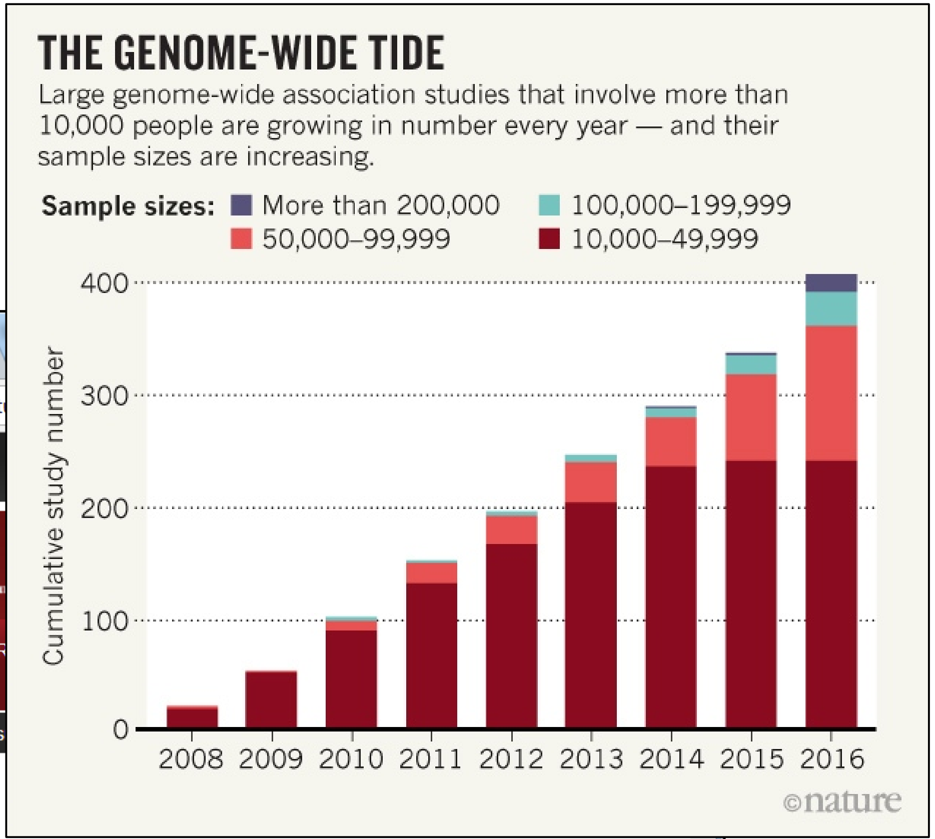
If it is (e.g., race prediction), little insights can be drawn upon.

The pathway is deep and complex

- Room for deep learning?

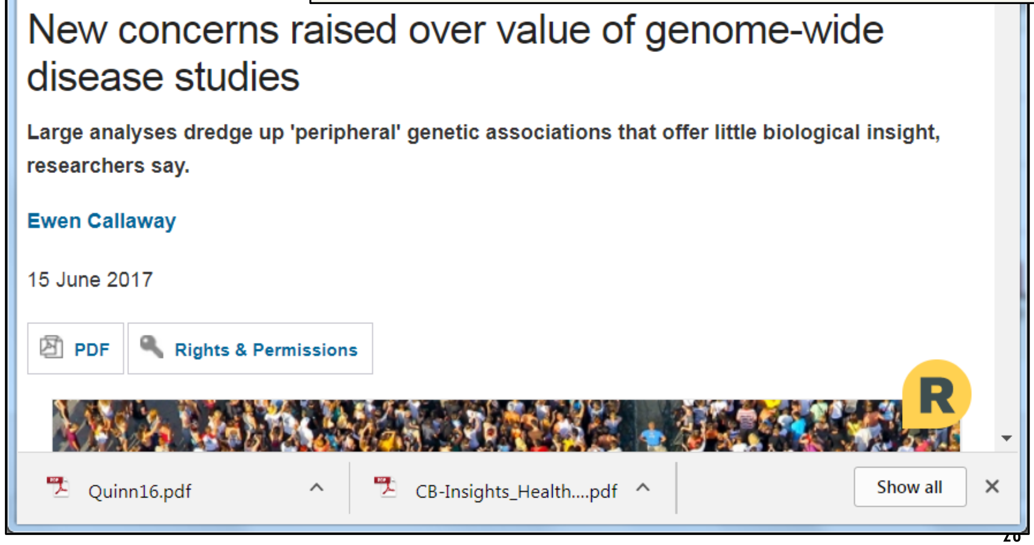
Room for structured models

- SNP annotations
- Spatial relationships
- Evolutionary trees



THE GENOME-WIDE TIDE

Large genome-wide association studies that involve more than 10,000 people are growing in number every year — and their sample sizes are increasing.



NATURE | NEWS

New concerns raised over value of genome-wide disease studies

Large analyses dredge up 'peripheral' genetic associations that offer little biological insight, researchers say.

Ewen Callaway

15 June 2017

# Rooms for deep learning

Bridge the genotype-phenotype gap

- Incorporating HUGE amount of data
- Modelling the multiple layers of complex biological processes in between.
- Starting from the DNA and its immediate functions, e.g., protein binding, gene start, alternative splicing, SNP annotations.

Deep learning has shown to work well in cognitive domains, where human can perform in less than a second.

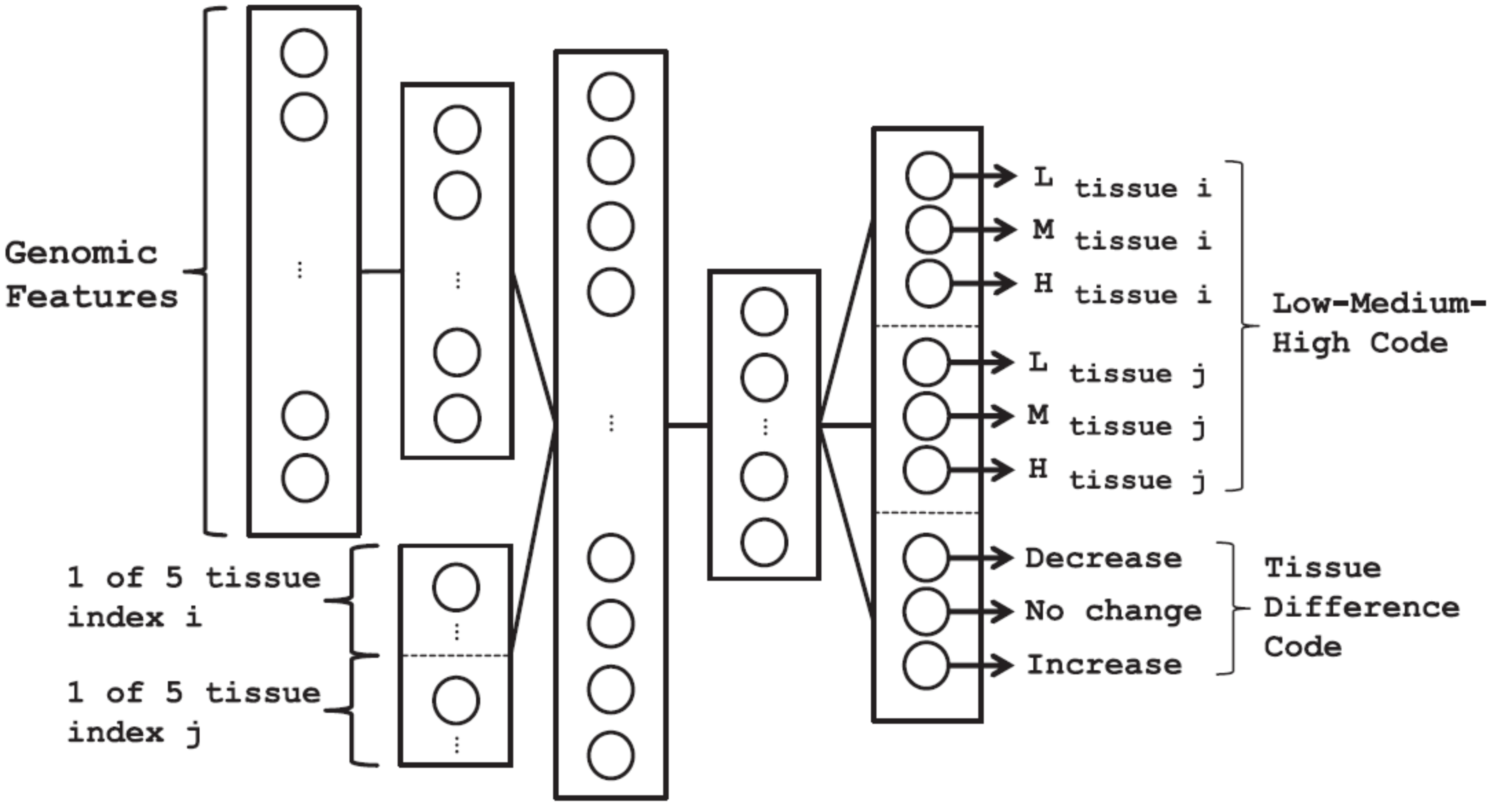
- We need to be super-human to bridge the gap.

New models for 2% of coding part, as well as 98% non-coding (probably having regulatory functions)

Incorporating biological understanding into model, not the black-box.

#Ref: https://www.oreilly.com/ideas/deep-learning-meets-genome-biology

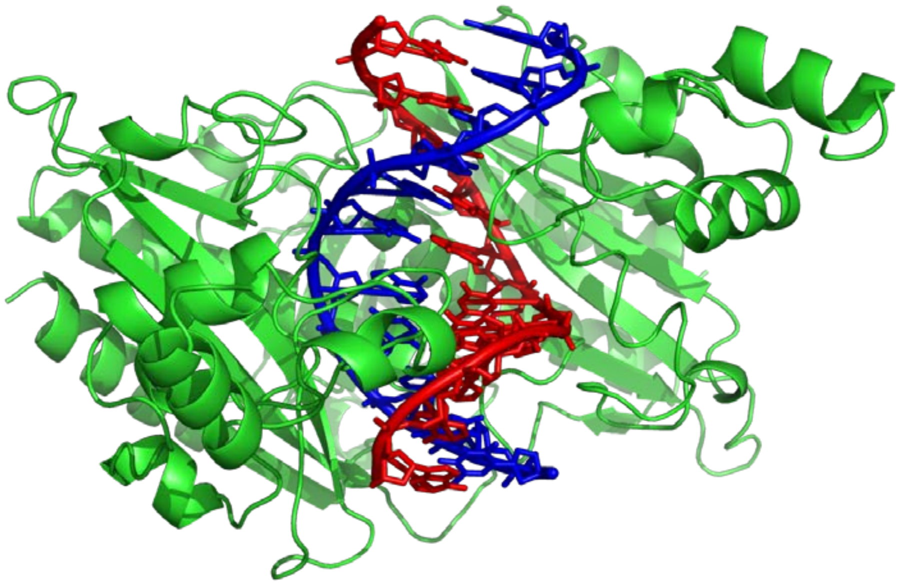
# Use of feedforward nets: Tissue-regulated splicing code



#REF: Leung, Michael KK, et al. "Deep learning of the tissue-regulated splicing code." *Bioinformatics* 30.12 (2014): i121-i129.

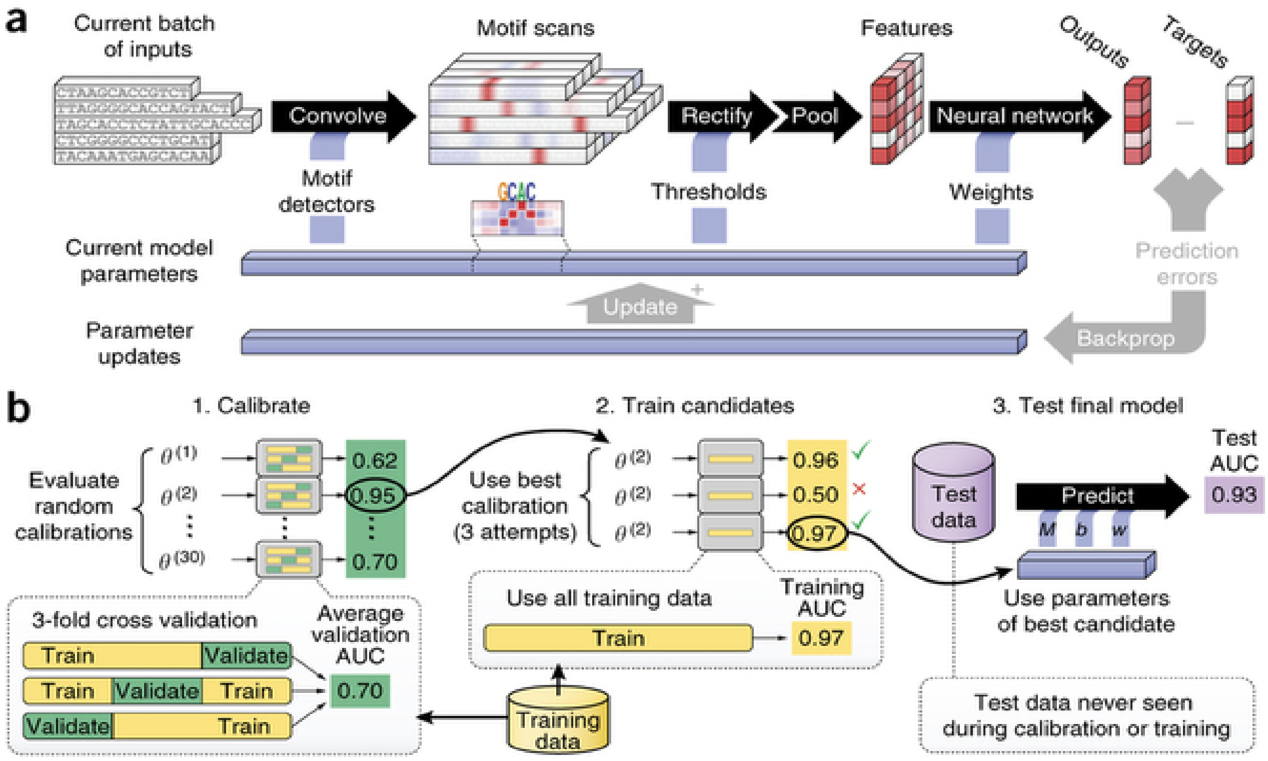
# Use of CNNs: Discovery of DNA motifs

**DeepBind** **(Alipanahi et al, Nature Biotech 2015)**



**The restriction enzyme EcoRV (green)**

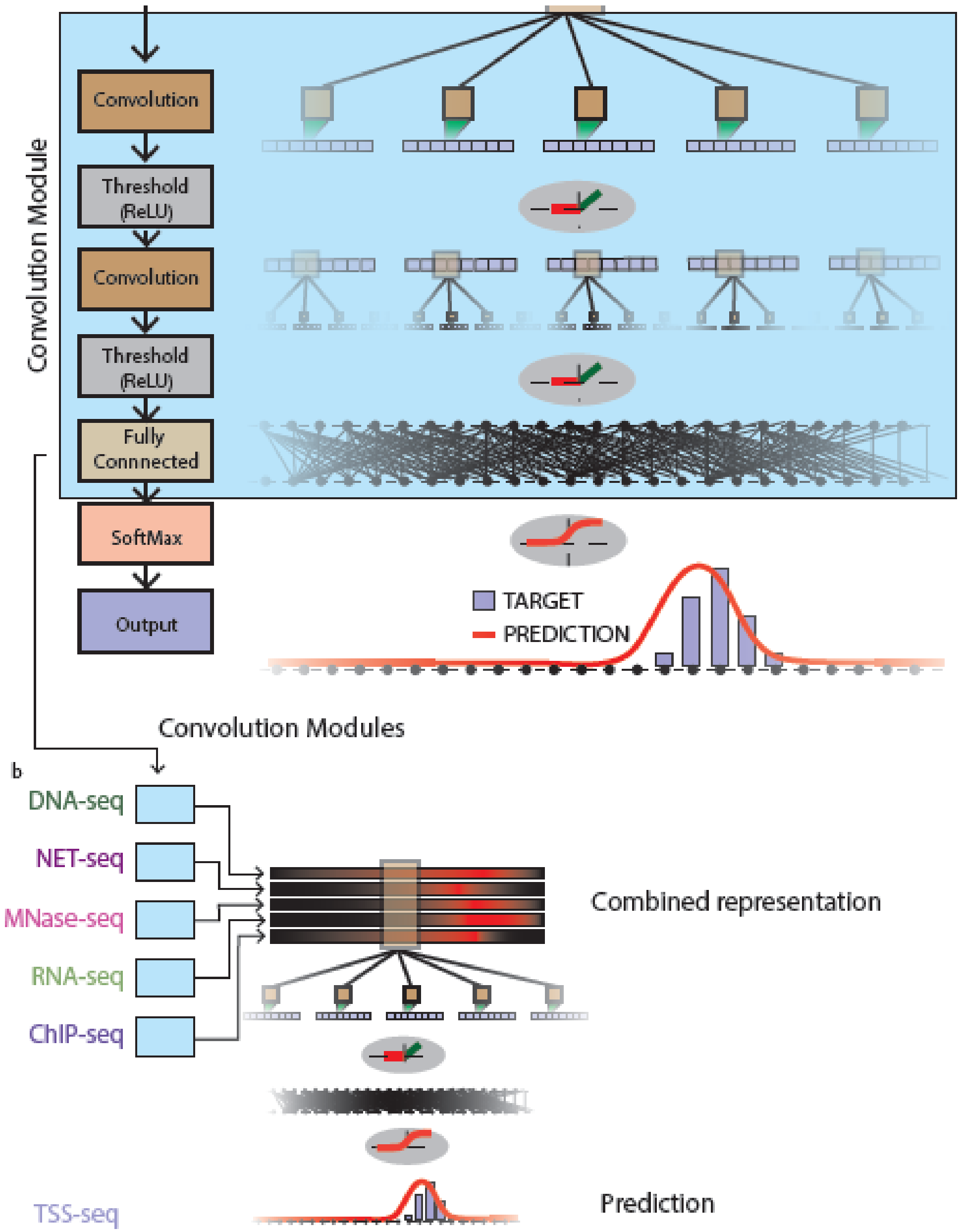
Source: wikipedia.org/wiki/DNA-binding_protein

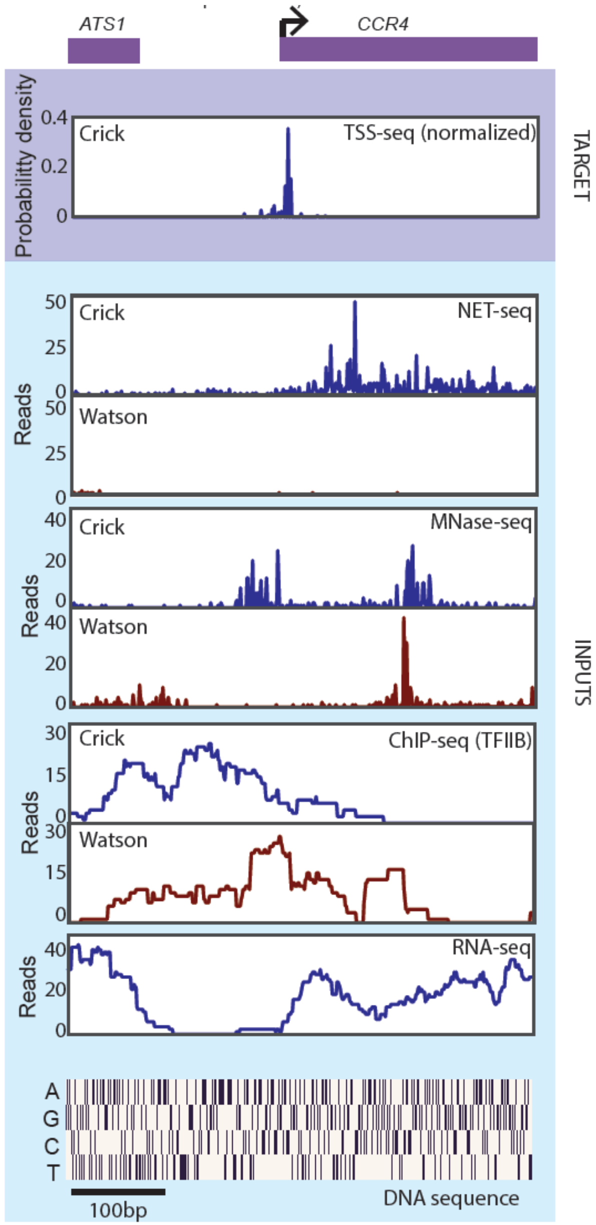


http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html

# Use of CNNs: FIDDLE

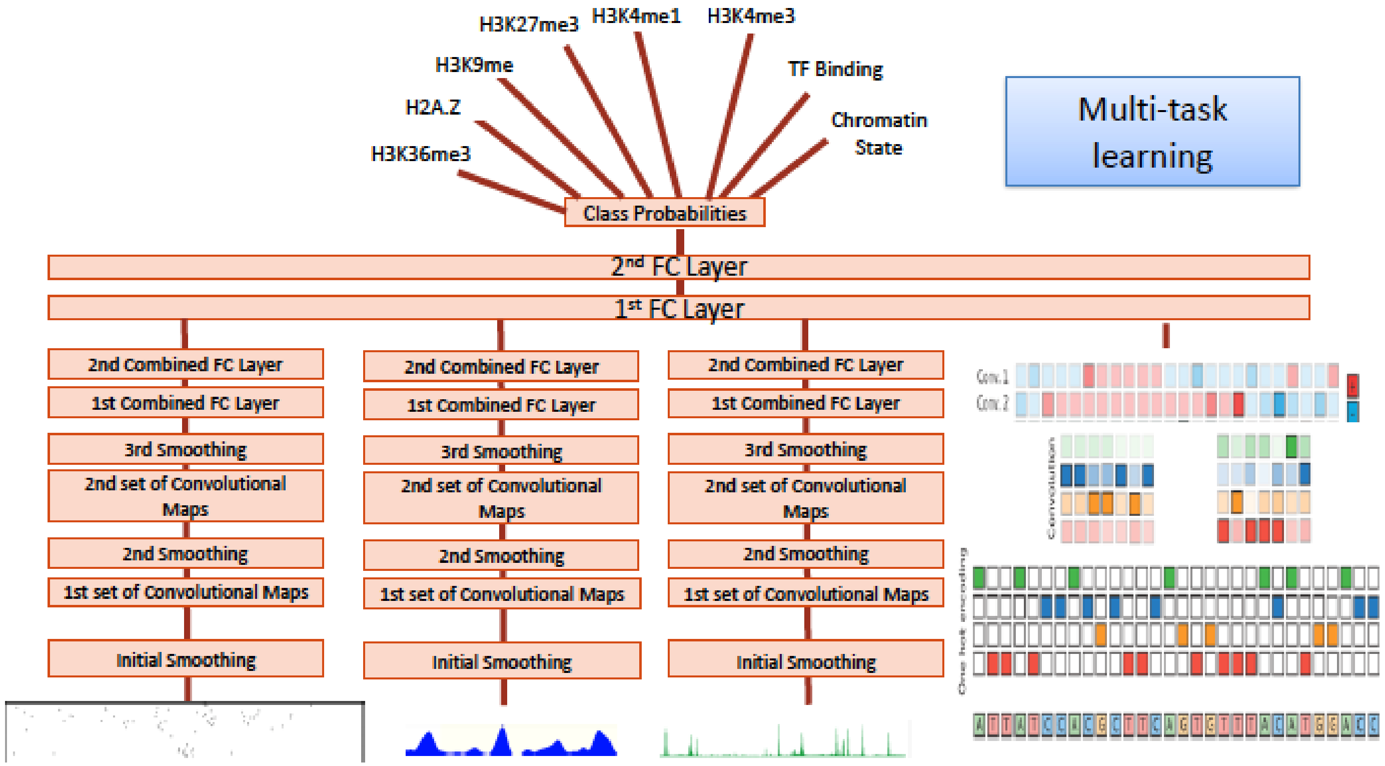
#REF: Eser, Umut, and L. Stirling Churchman. "FIDDLE: An integrative deep learning framework for functional genomic data inference." *bioRxiv* (2016): 081380.





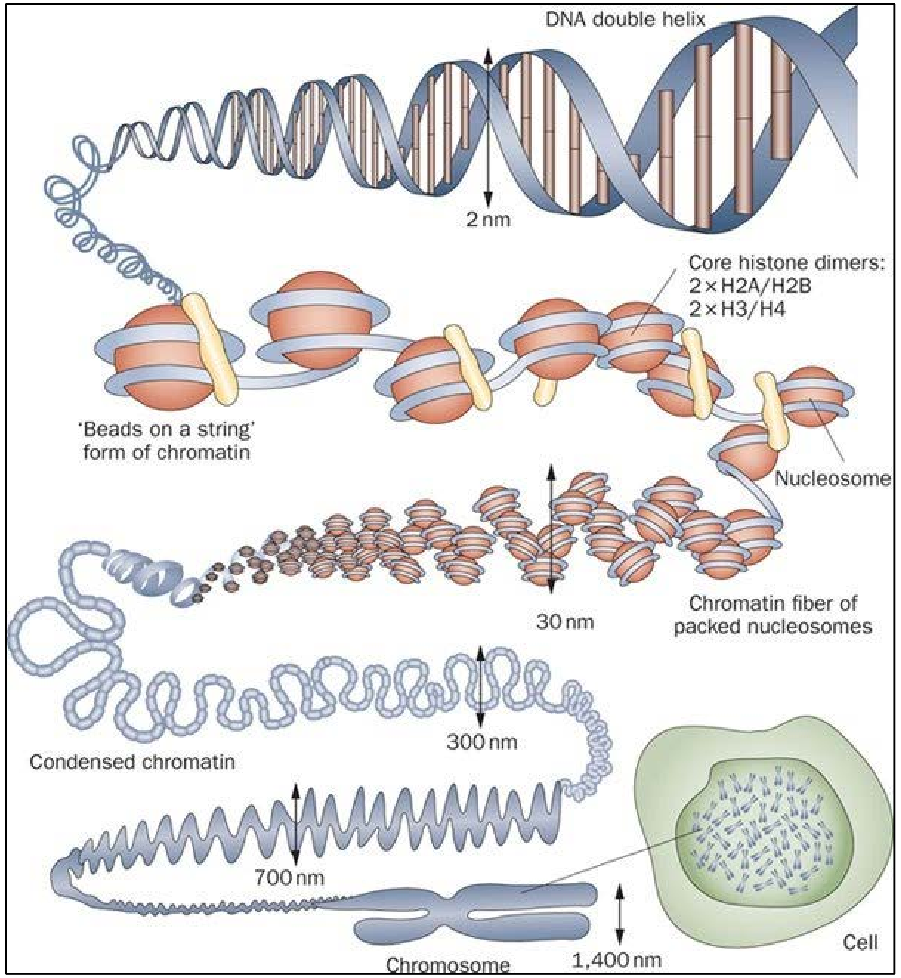
# THE CHROMPUTER

Integrating multiple inputs (1D, 2D signals, sequence) to simulatenously **predict multiple outputs**



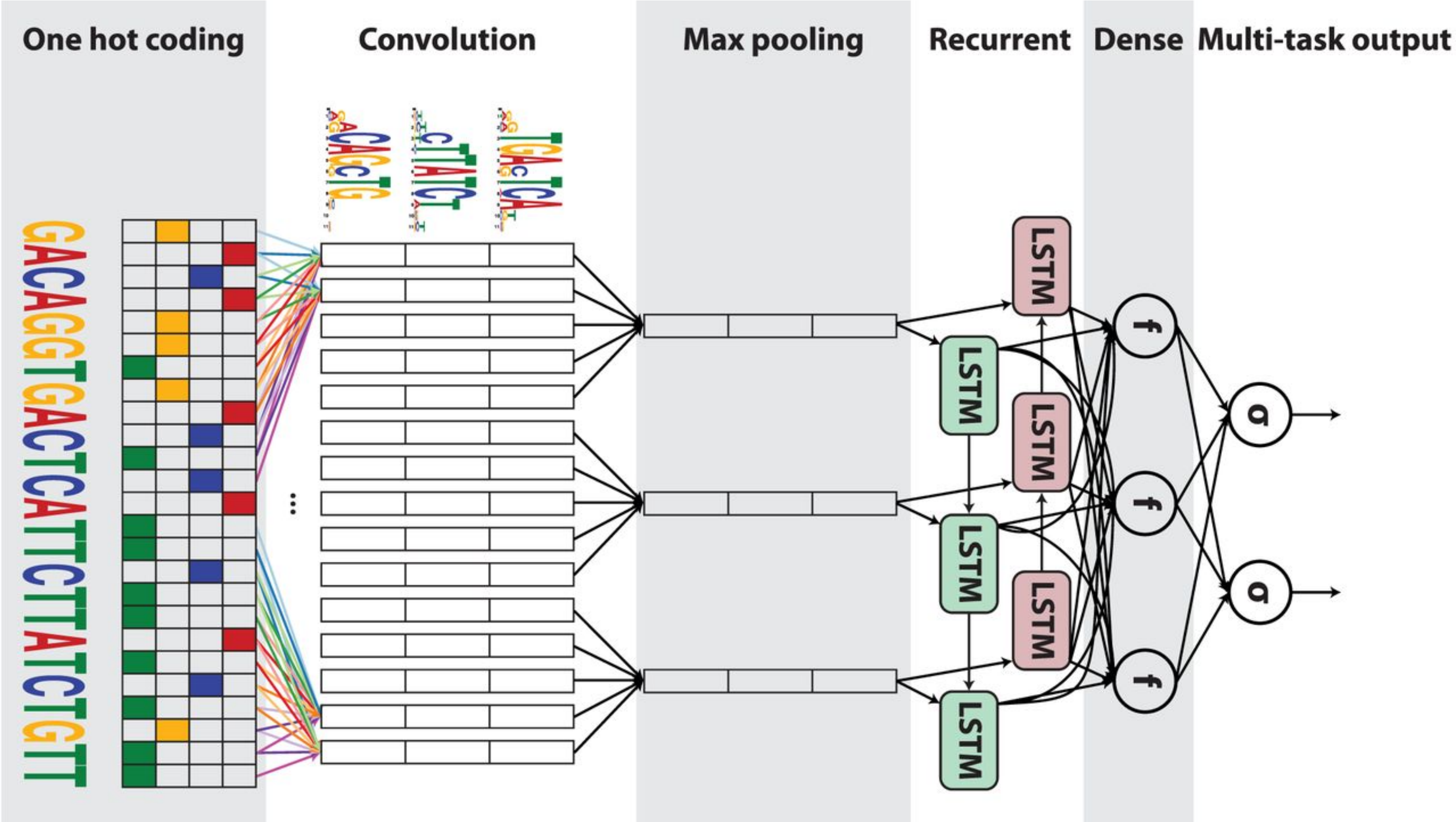
Source: https://simons.berkeley.edu/sites/default/files/docs/4575/2016-kundaje-simonsinstitute-deeplearning.pdf

# Chromatins



https://qph.ec.quoracdn.net

# User of CNN+RNNs: DanQ



#REF: Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic acids research* 44.11 (2016): e107-e107.

# More models/frameworks

DragoNN

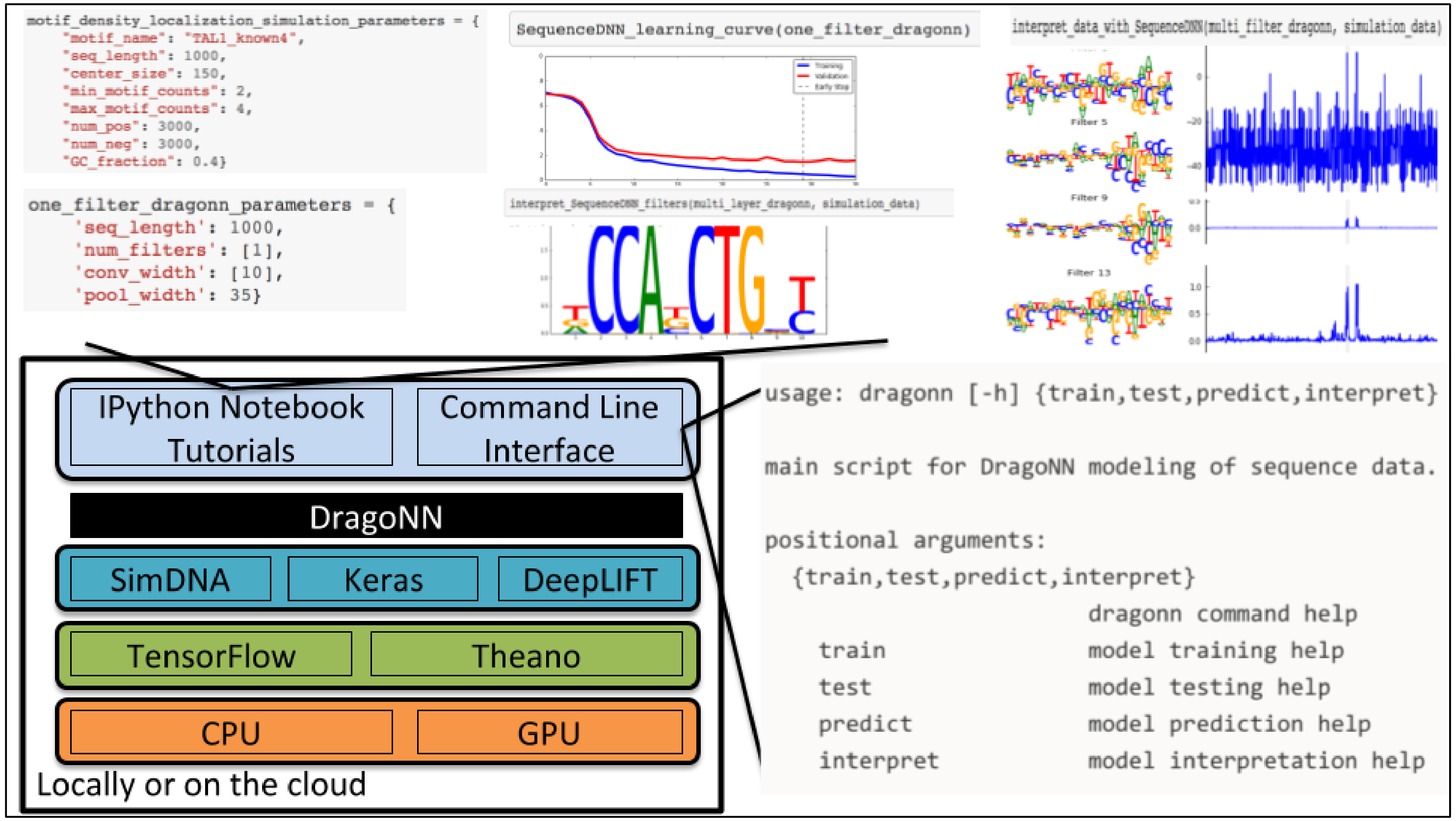
DeepChrome

DeepSEA

Basset

DeepBound

…



```
motif_density_localization_simulation_parameters = {
    "motif_name": "TAL1_known4",
    "seq_length": 1000,
    "center_size": 150,
    "min_motif_counts": 2,
    "max_motif_counts": 4,
    "num_pos": 3000,
    "num_neg": 3000,
    "GC_fraction": 0.4}
```

```
one_filter_dragonn_parameters = {
    'seq_length': 1000,
    'num_filters': [1],
    'conv_width': [10],
    'pool_width': 35}
```

```
SequenceDNN_learning_curve(one_filter_dragonn)
```

```
interpret_SequenceDNN_filters(multi_layer_dragonn, simulation_data)
```

```
interpret_data_with_SequenceDNN(multi_filter_dragonn, simulation_data)
```

```
usage: dragonn [-h] {train,test,predict,interpret}

main script for DragoNN modeling of sequence data.

positional arguments:
  {train,test,predict,interpret}
                        dragonn command help
    train               model training help
    test                model testing help
    predict             model prediction help
    interpret           model interpretation help
```

http://kundajelab.github.io/dragonn

# What make biomedicine hard for deep learning?

Great diversity but may be small in size

High uncertainty, low-quality/missing data

Reusable models do not usually exist

Human doesn’t know how to read biomedicine (Brendan Frey, U of Toronto)

Require deep thinking for a reasonable deep architecture

However, at the end of the day, we need only a few generic things:

- Vector → DNN (e.g., highway net) | Sequence → RNN (e.g., LSTM, GRU)
- Repeated motifs → CNN | Set → Attention
- Graphs → Conv graphs; Column Networks
- Generative models → VAE; GAN

# Agenda

## Deep learning

- Neural architectures
- Generative models

## Genomics

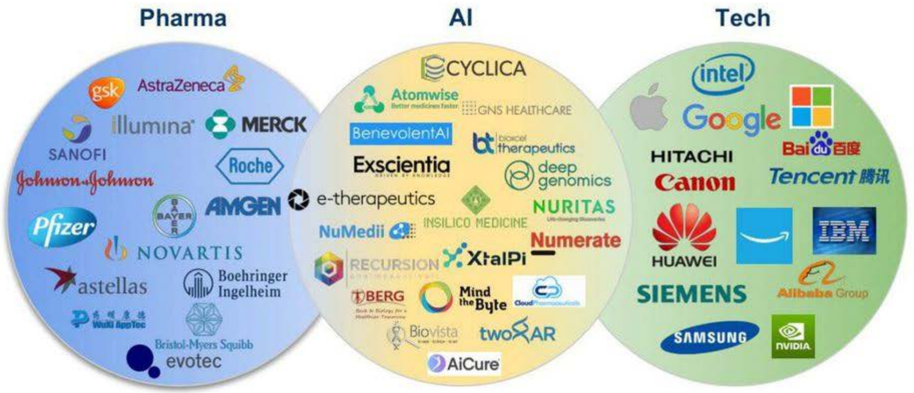
- Nanopore sequencing
- Genomics modelling

## Drug design

- Bioactivity prediction
- Drug generation

## Future outlook

Leading Companies
Advanced AI in Healthcare and Drug Discovery



| Drug Discovery | Pharma Corporations | Tech Corporations | Advanced R&D in Healthcare |
|---|---|---|---|
| INSILICO MEDICINE<br>Exscientia<br>CloudPharmaceuticals | gsk | Tencent 腾讯 | XtalPi<br>Atomwise |
| CYCLICA<br>Atomwise<br>Numerate | MERCK | Alibaba.com | XtalPi |
| XtalPi<br>Biovista<br>IBM Watson | Pfizer | amazon | XtalPi |
| Exscientia<br>BERG<br>RECURSION | SANOFI | Baidu 百度 | Atomwise |
| AiCure<br>Numerate<br>NuMedii | Boehringer Ingelheim | Google | XtalPi |
| Biovista<br>INSILICO MEDICINE | NOVARTIS | IBM | IBM Watson |
| Atomwise<br>CYCLICA | BAYER | nVIDIA | INSILICO MEDICINE |
| AiCure<br>GNS HEALTHCARE | Roche | Canon | |
| BenevolentAI | Johnson&Johnson | | |
| INSILICO MEDICINE | WuXi AppTec | HUAWEI | |
| BERG | AstraZeneca | intel | |

https://www.forbes.com/sites/yiannismouratidis/2018/12/16/the-rising-star-companies-in-ai-drug-development

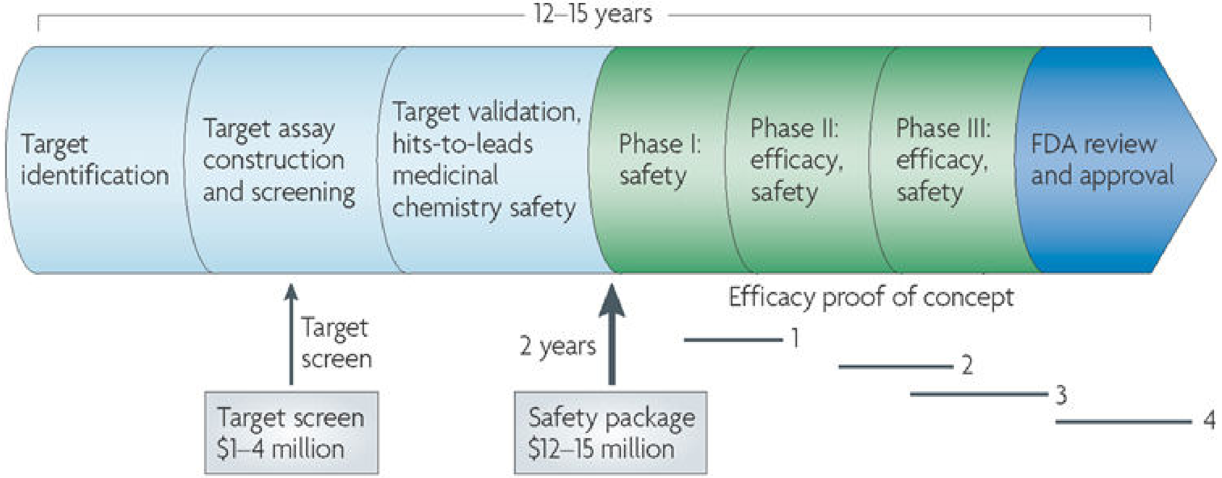
# Deep learning for drug discovery

Predicting bioactivities from molecules

Drug representation, unsupervised learning from graphs

Generate from bioactivities to molecular graphs

**$500M - $2B**


12–15 years
Target identification
Target assay construction and screening
Target validation, hits-to-leads medicinal chemistry safety
Phase I: safety
Phase II: efficacy, safety
Phase III: efficacy, safety
FDA review and approval
Target screen
2 years
Efficacy proof of concept
1
2
3
4
Target screen $1–4 million
Safety package $12–15 million

Nature Reviews | Drug Discovery

#REF: Roses, Allen D. "Pharmacogenetics in drug discovery and development: a translational perspective." *Nature reviews Drug discovery* 7.10 (2008): 807-817.

# Traditional method: Combinatorial chemistry

Generate variations on a template

Returns a list of molecules from this template that

- Bind to the pocket with good pharmacodynamics?
- Have good pharmacokinetics?
- Are synthetically accessible?

#REF: Talk by Chloé-Agathe Azencott titled "Machine learning for therapeutic research", 12/10/2017

# First step: Map molecule → drug properties (binding/acting)

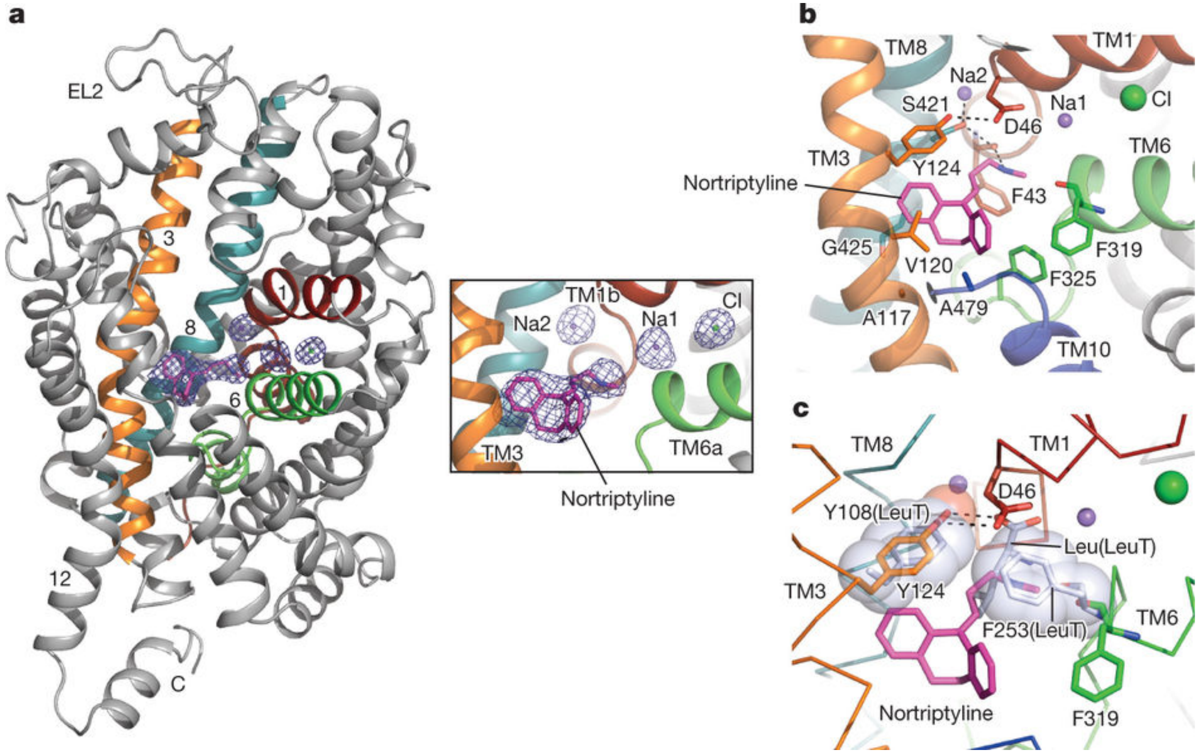
Drugs are small bio-molecules

Traditional techniques:

- Graph kernels (ML)
- Molecular fingerprints (Chemistry)

Modern techniques

- Molecule as graph: atoms as nodes, chemical bonds as edges



#REF: Penmatsa, Aravind, Kevin H. Wang, and Eric Gouaux. "X-ray structure of dopamine transporter elucidates antidepressant mechanism." *Nature* 503.7474 (2013): 85-90.

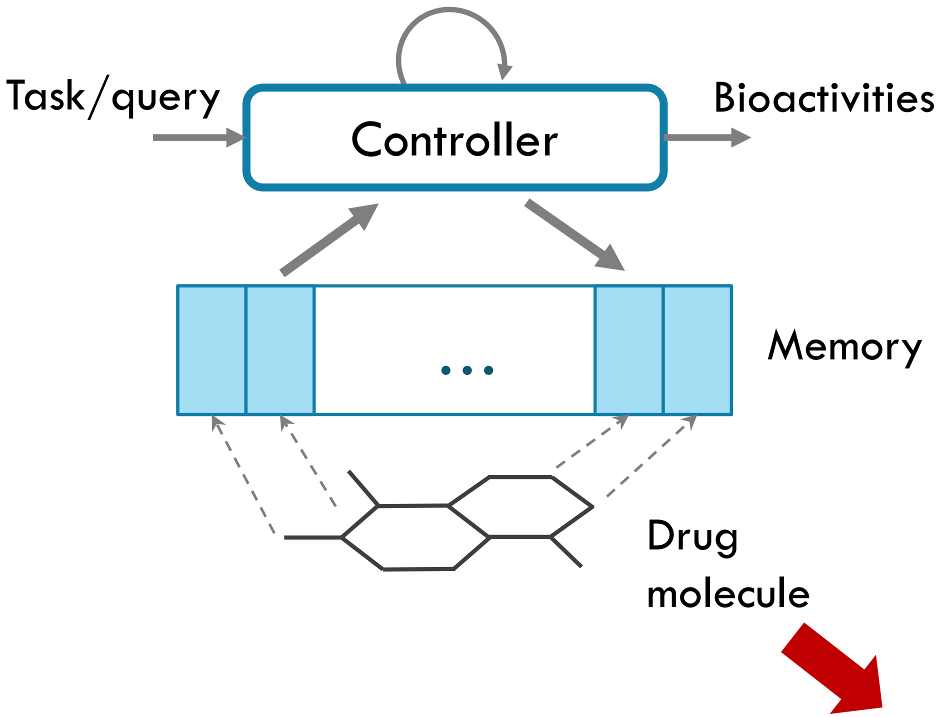
# 3 methods for bioactivity prediction

Graph memory networks (GMN) for drug bioactivity prediction

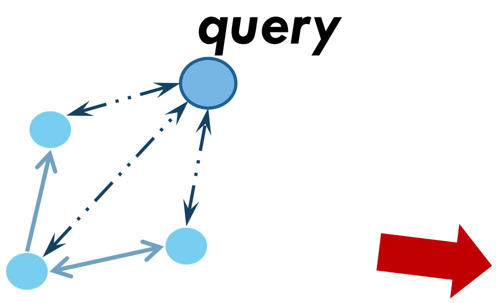
Graph attentional multi-label learning (GAML) for drug multi-target binding & repurposing

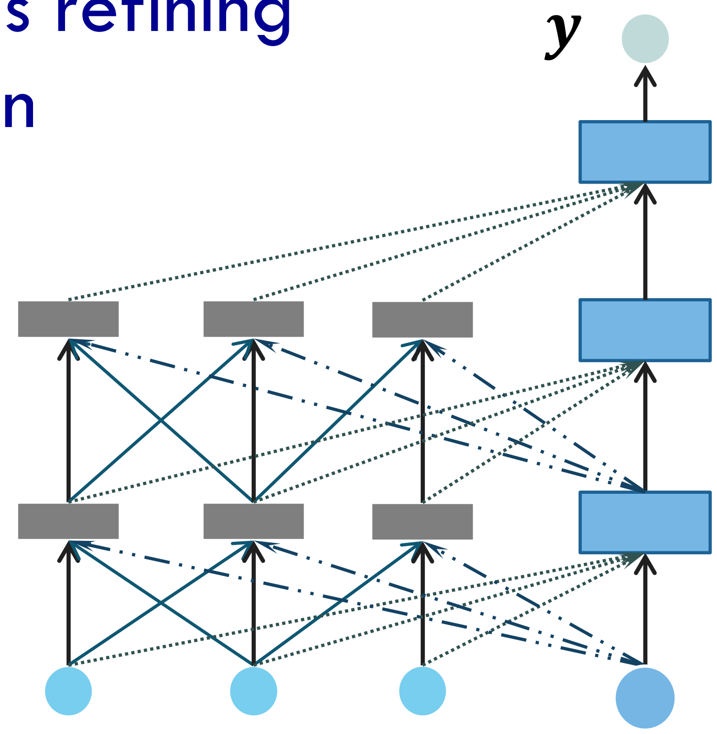
Relational dynamic memory networks (RDMNs) for drug-drug / drug-protein interaction

# Graph memory networks



Message passing as refining atom representation





#Ref: Pham, Trang, Truyen Tran, and Svetha Venkatesh. "Graph Memory Networks for Molecular Activity Prediction." *ICPR'18*.

# Graph memory networks: Results

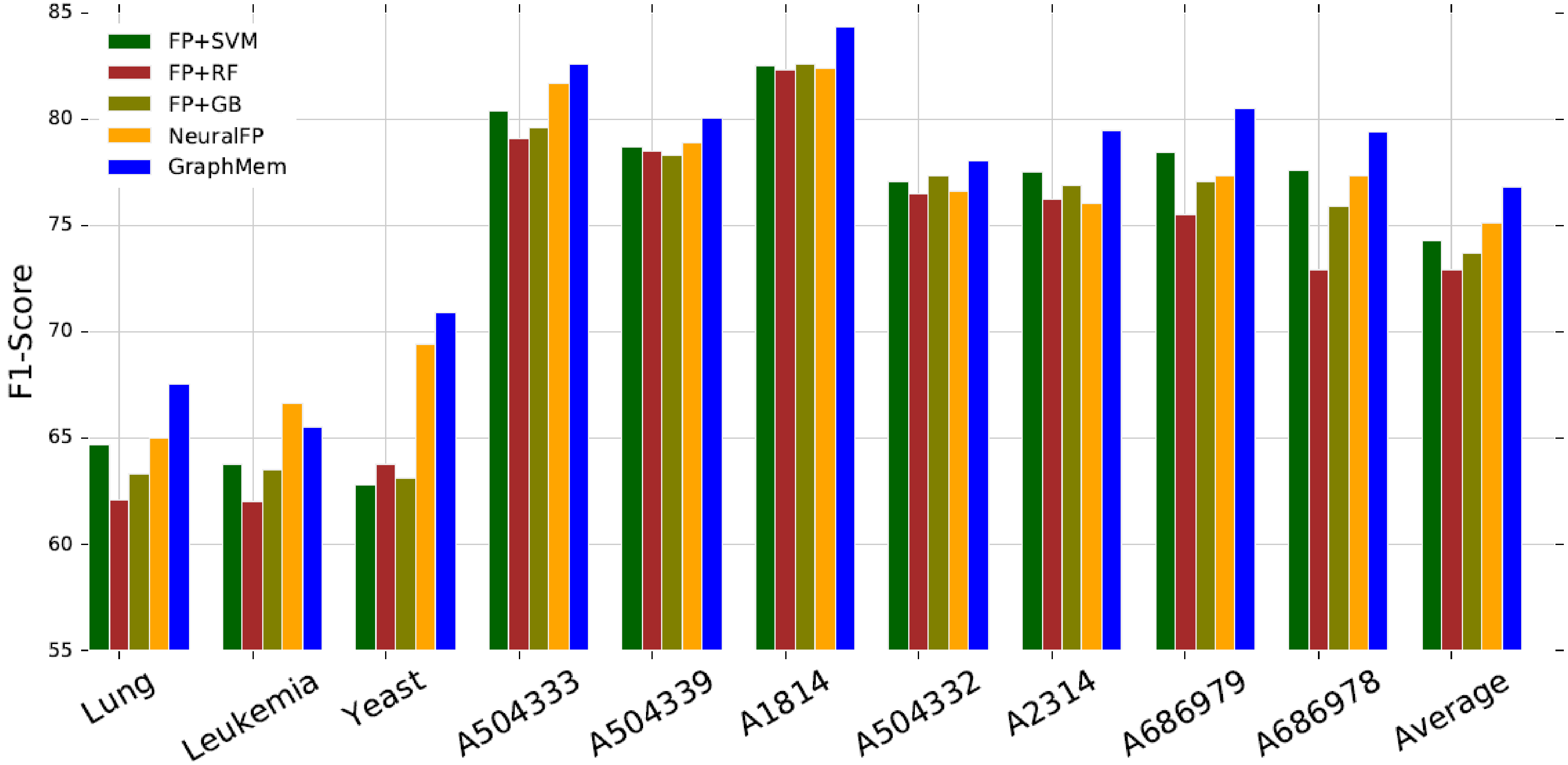
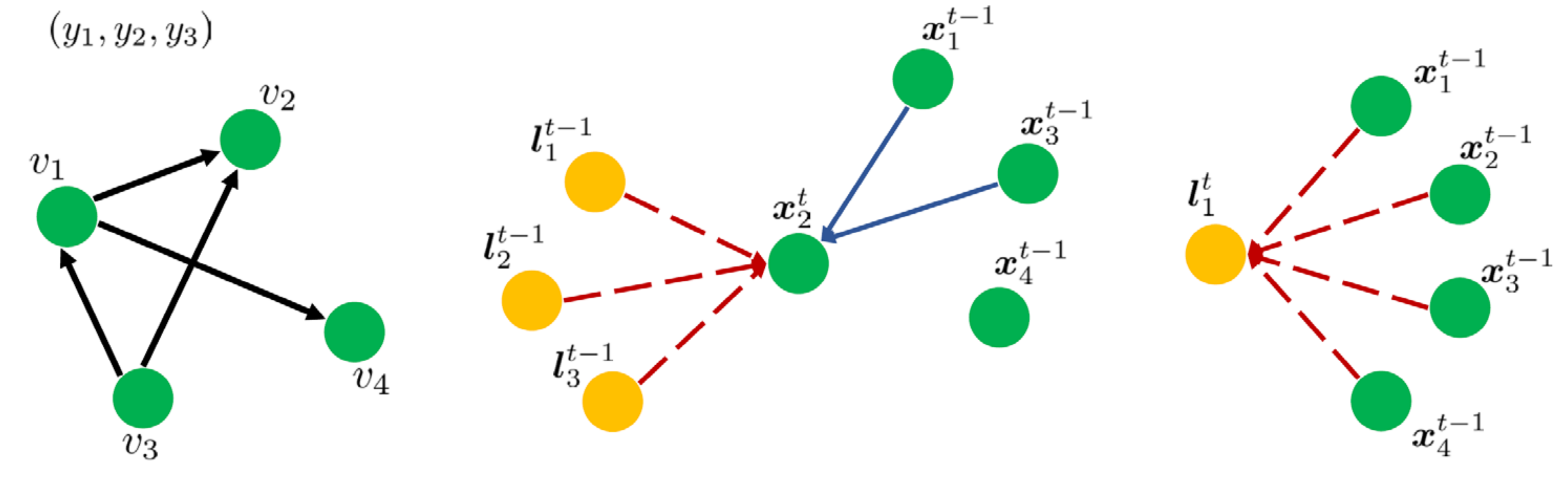


Figure 2: F1-score (%) for NCI datasets. FP = Fingerprint; RF = Random Forests; GBM = Gradient Boosting Machine. Best view in color.

# Multi-target binding for drug repurposing

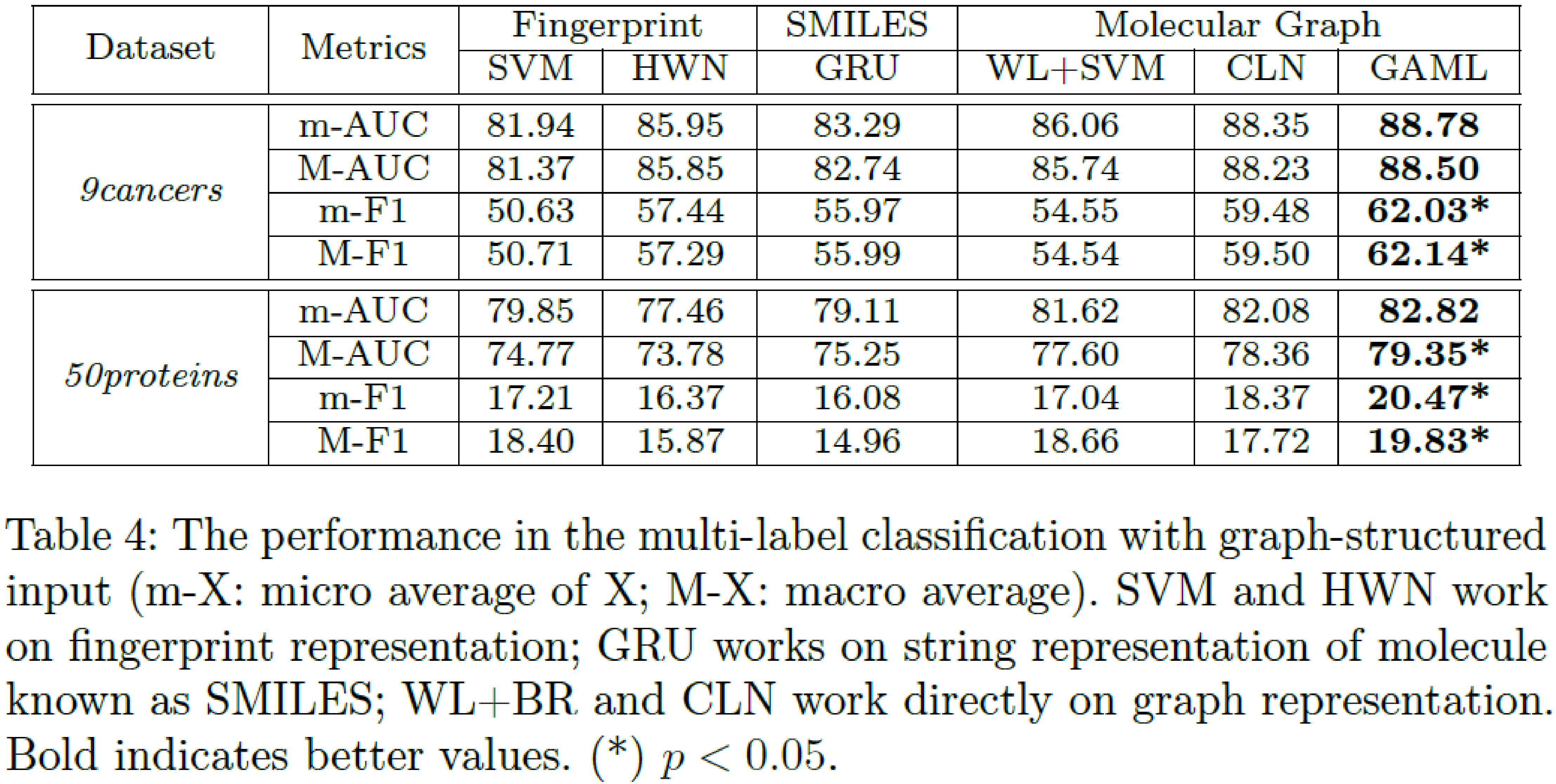


(a) A input graph with 4 nodes and 3 labels

(b) Input node update

(c) Label node update

#REF: Do, Kien, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *Machine Learning, 2019.*

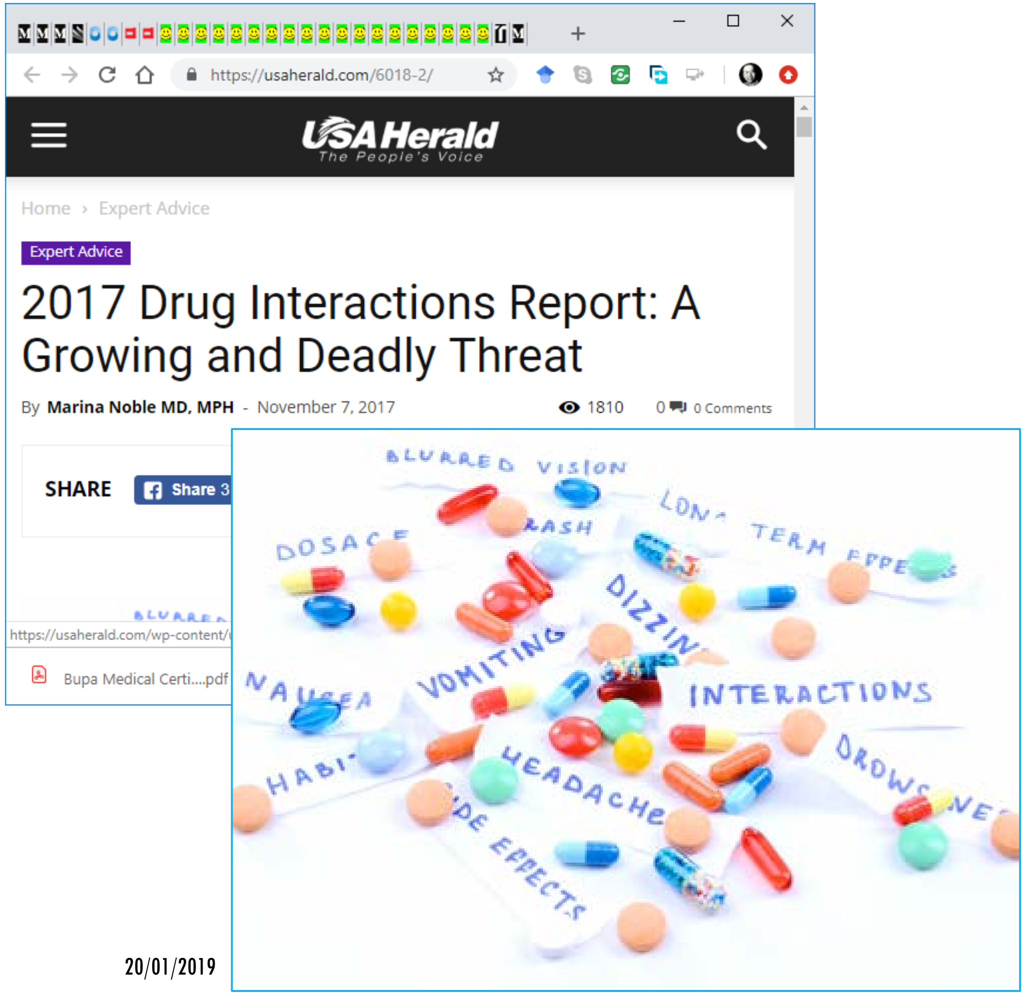


| Dataset | Metrics | Fingerprint | | SMILES | Molecular Graph | | |
|---|---|---|---|---|---|---|---|
| | | SVM | HWN | GRU | WL+SVM | CLN | GAML |
| *9cancers* | m-AUC | 81.94 | 85.95 | 83.29 | 86.06 | 88.35 | **88.78** |
| | M-AUC | 81.37 | 85.85 | 82.74 | 85.74 | 88.23 | **88.50** |
| | m-F1 | 50.63 | 57.44 | 55.97 | 54.55 | 59.48 | **62.03*** |
| | M-F1 | 50.71 | 57.29 | 55.99 | 54.54 | 59.50 | **62.14*** |
| *50proteins* | m-AUC | 79.85 | 77.46 | 79.11 | 81.62 | 82.08 | **82.82** |
| | M-AUC | 74.77 | 73.78 | 75.25 | 77.60 | 78.36 | **79.35*** |
| | m-F1 | 17.21 | 16.37 | 16.08 | 17.04 | 18.37 | **20.47*** |
| | M-F1 | 18.40 | 15.87 | 14.96 | 18.66 | 17.72 | **19.83*** |

Table 4: The performance in the multi-label classification with graph-structured input (m-X: micro average of X; M-X: macro average). SVM and HWN work on fingerprint representation; GRU works on string representation of molecule known as SMILES; WL+BR and CLN work directly on graph representation. Bold indicates better values. (*) $p < 0.05$.

#REF: Do, Kien, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *arXiv preprint arXiv:1804.00293*(2018).

# Drug-drug interaction via *Relational Dynamic Memory Networks*



https://usaherald.com/6018-2/

USA Herald

The People's Voice

Home › Expert Advice

Expert Advice

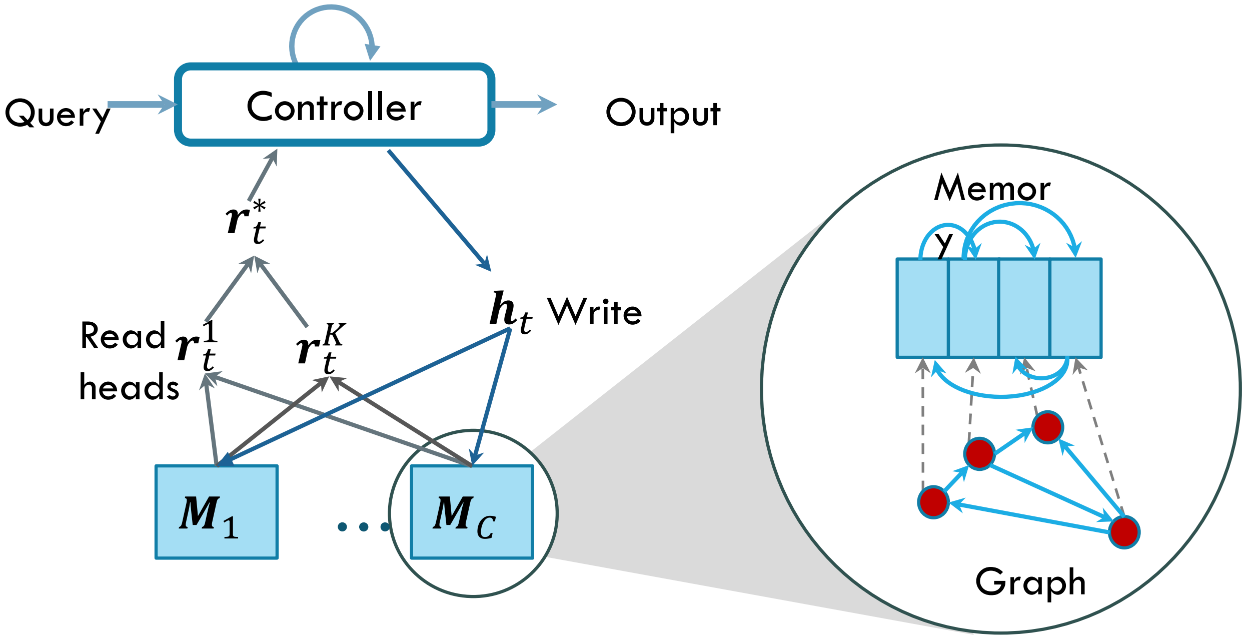
## 2017 Drug Interactions Report: A Growing and Deadly Threat

By **Marina Noble MD, MPH** - November 7, 2017 1810 0 0 Comments

SHARE Share 3

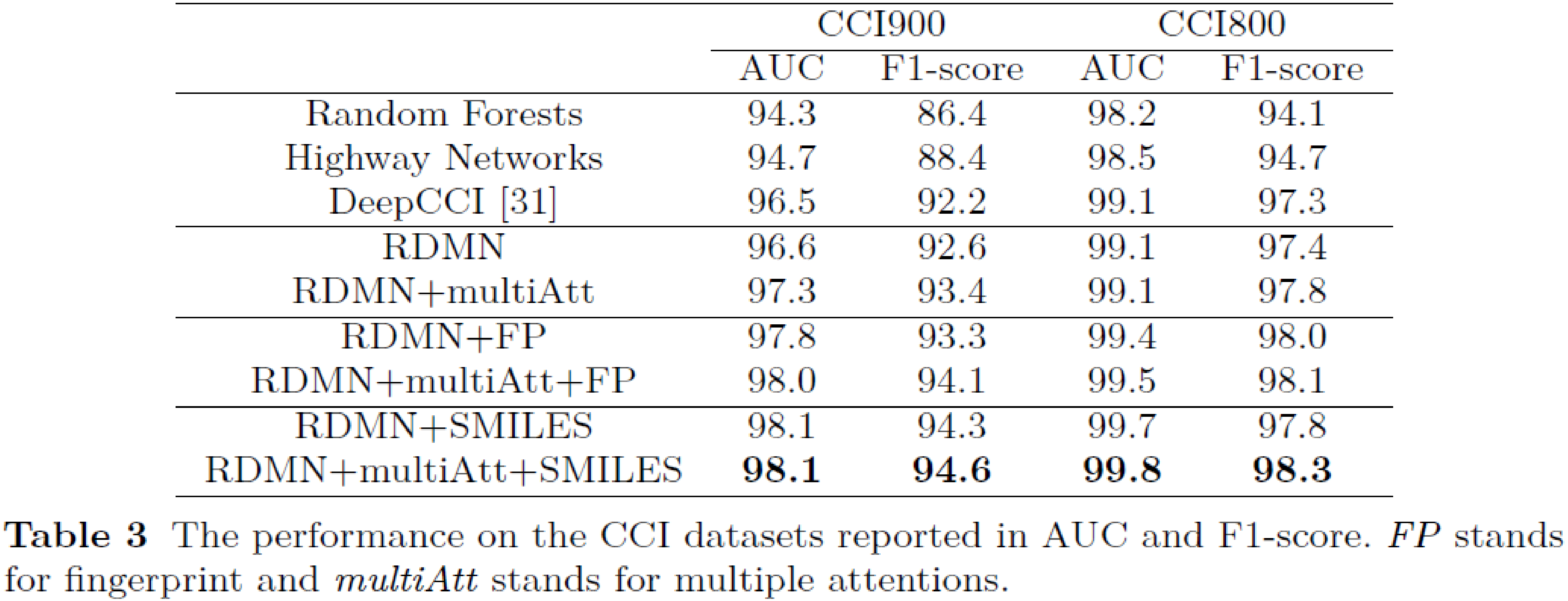
https://usaherald.com/wp-content/

Bupa Medical Certi....pdf



#REF: Pham, Trang, Truyen Tran, and Svetha Venkatesh. "Relational dynamic memory networks." *arXiv preprint arXiv:1808.04247*(2018).

# Results on STITCH database



| | CCI900 | | CCI800 | |
|---|---|---|---|---|
| | AUC | F1-score | AUC | F1-score |
| Random Forests | 94.3 | 86.4 | 98.2 | 94.1 |
| Highway Networks | 94.7 | 88.4 | 98.5 | 94.7 |
| DeepCCI [31] | 96.5 | 92.2 | 99.1 | 97.3 |
| RDMN | 96.6 | 92.6 | 99.1 | 97.4 |
| RDMN+multiAtt | 97.3 | 93.4 | 99.1 | 97.8 |
| RDMN+FP | 97.8 | 93.3 | 99.4 | 98.0 |
| RDMN+multiAtt+FP | 98.0 | 94.1 | 99.5 | 98.1 |
| RDMN+SMILES | 98.1 | 94.3 | 99.7 | 97.8 |
| RDMN+multiAtt+SMILES | **98.1** | **94.6** | **99.8** | **98.3** |

**Table 3** The performance on the CCI datasets reported in AUC and F1-score. *FP* stands for fingerprint and *multiAtt* stands for multiple attentions.

# Drug generation

We now have methods for compute bioactivties of a drug molecule

We need a reverse method to generate drug molecules from desirable bioactivities

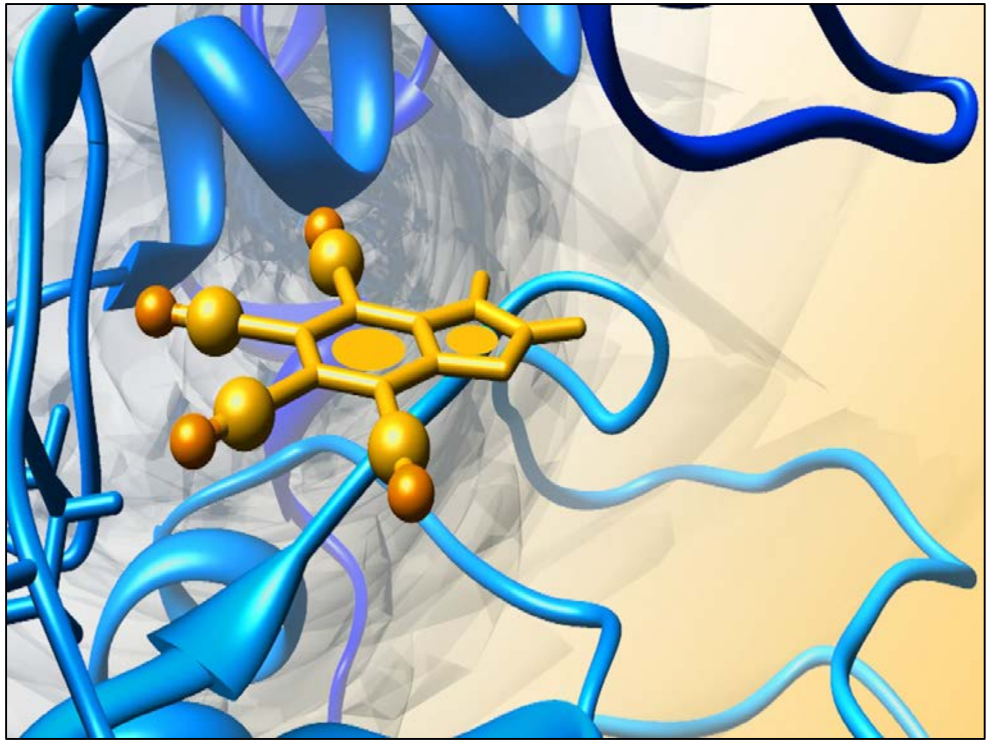
The space of drugs is estimated to be 1e+23 to 1e+60

- Only 1e+8 substances synthesized thus far.

It is impossible to model this space fully.

The current technologies are not mature for graph generations.

But approximate techniques do exist.



Source: pharmafactz.com

# Old and new methods

## Existing methods:

- Exhausted search through a fixed library
- Discrete local search: genetic algorithms, similar discrete interpolation
- The search space is still large.

## Deep learning methods:

- Faster, more efficient to find new drugs
- Able of generate molecules that are likely the good candidates

# Deep learning methods

Representing molecules using fingerprints

Representing graph as string, and use sequence VAEs or GANs.

Graph VAE & GAN

- Model nodes & interactions
- Model cliques

Sequences

- Iterative methods

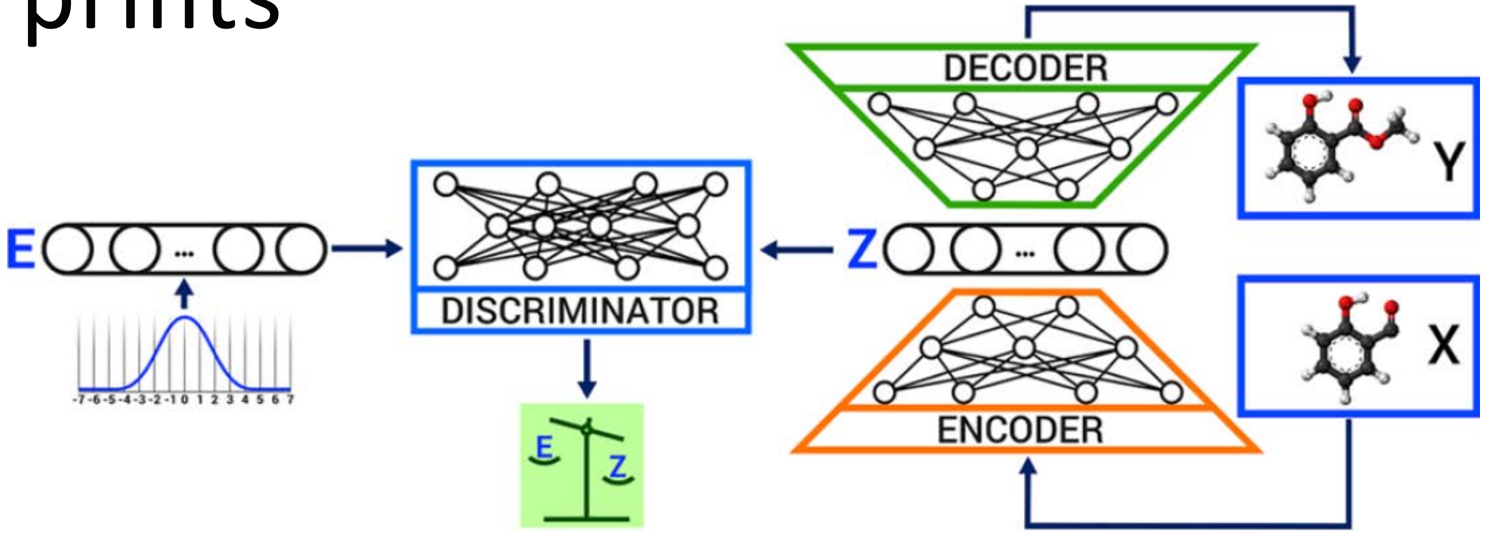
Reinforcement learning

- Discrete objectives

Any combination of these + memory.

Kadurin, Artur, et al. "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology." *Oncotarget* 8.7 (2017): 10883.

# Molecule → fingerprints



Input of the encoder : the fingerprint of a molecule

The decoder outputs the predicted fingerprint .

The generative model generates a vector E, which is then discriminated from the latent vector of the real molecule by the discriminator.

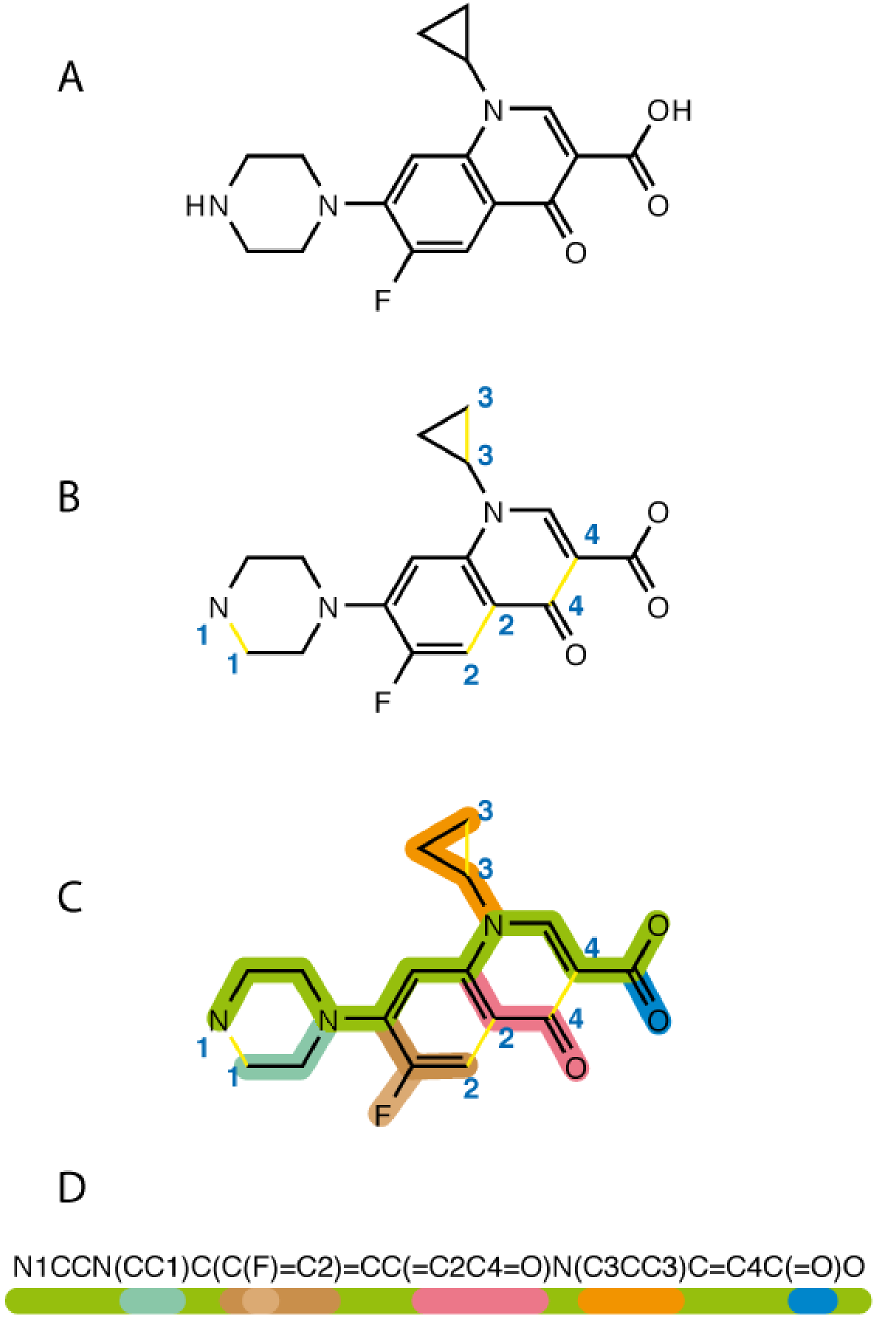
# Molecule → string

Using SMILES representation of drug, to convert a molecular graph into a string

- SMILES = Simplified Molecular-Input Line-Entry System

Then using sequence-to-sequence + VAE/GAN to model the continuous space that encodes/decodes SMILES strings

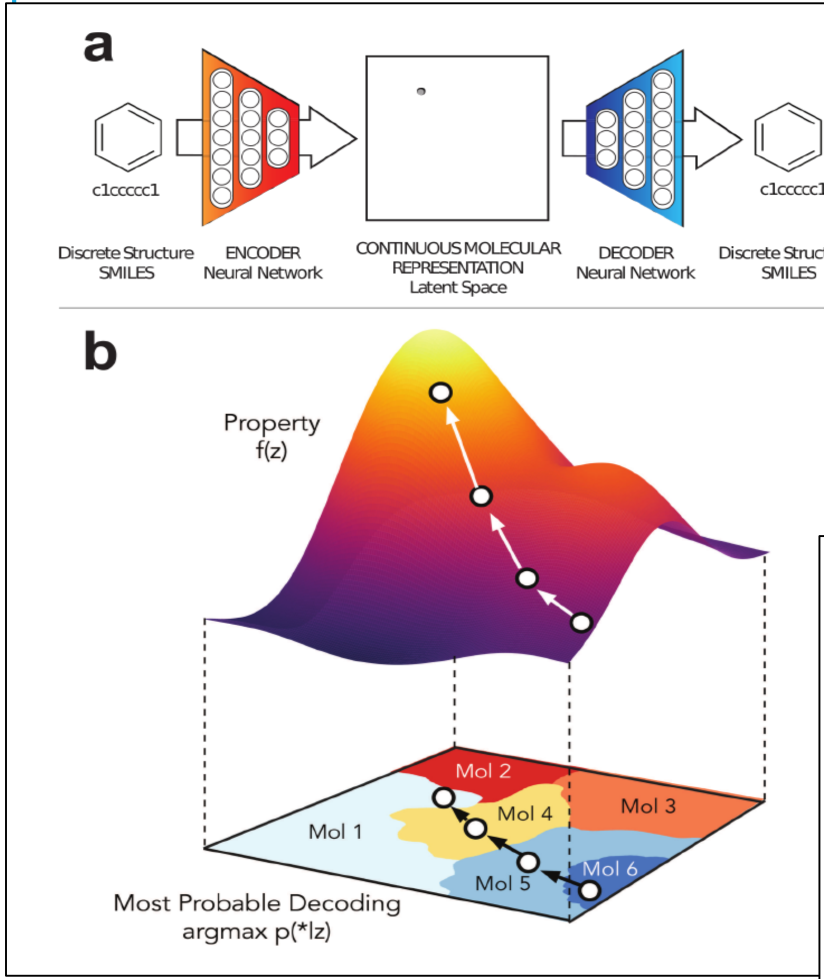
- Allow easy optimization on the continuous space



Source: wikipedia.org

#REF: Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *arXiv preprint arXiv:1610.02415* (2016).

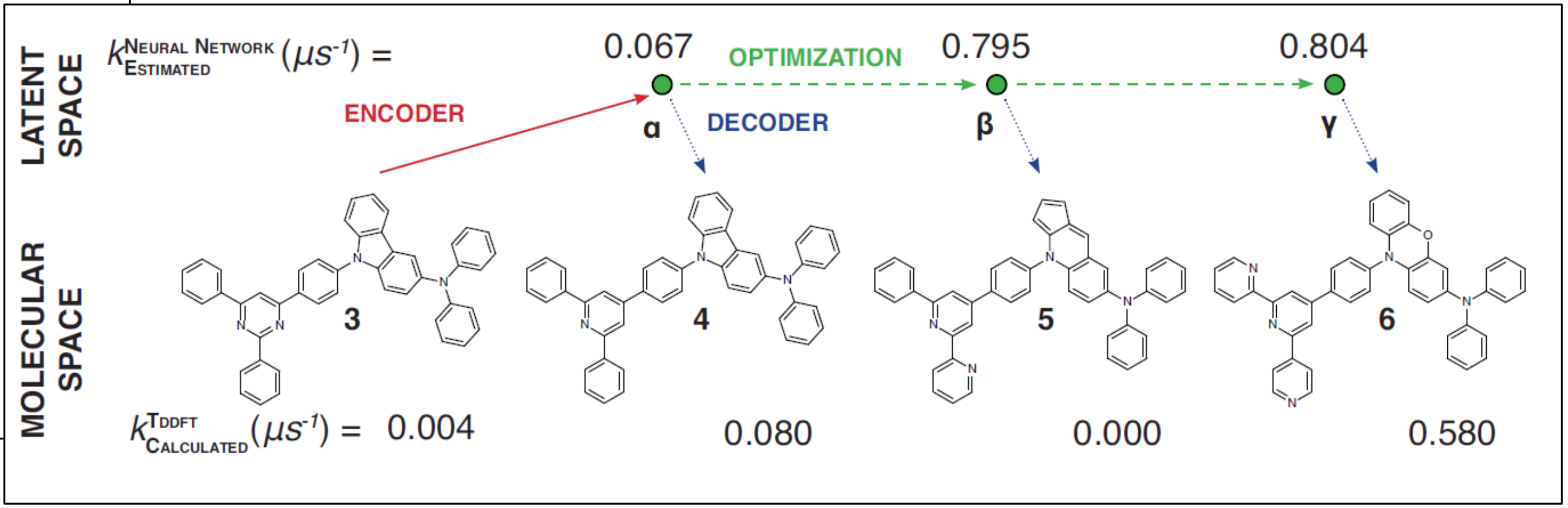
# VAE for drug space modelling



## Uses VAE for sequence-to-sequence.

#REF: Bowman, Samuel R., et al. "Generating sentences from a continuous space." *arXiv preprint arXiv:1511.06349* (2015).

Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS Central Science* (2016).


LATENT SPACE
MOLECULAR SPACE
$k^{\text{NEURAL NETWORK}}_{\text{ESTIMATED}}\ (\mu s^{-1}) =$
0.067
OPTIMIZATION
0.795
0.804
ENCODER
α
DECODER
β
γ
3
4
5
6
N
O
$k^{\text{TDDFT}}_{\text{CALCULATED}}\ (\mu s^{-1}) =$
0.004
0.080
0.000
0.580

# Drawbacks of string representation

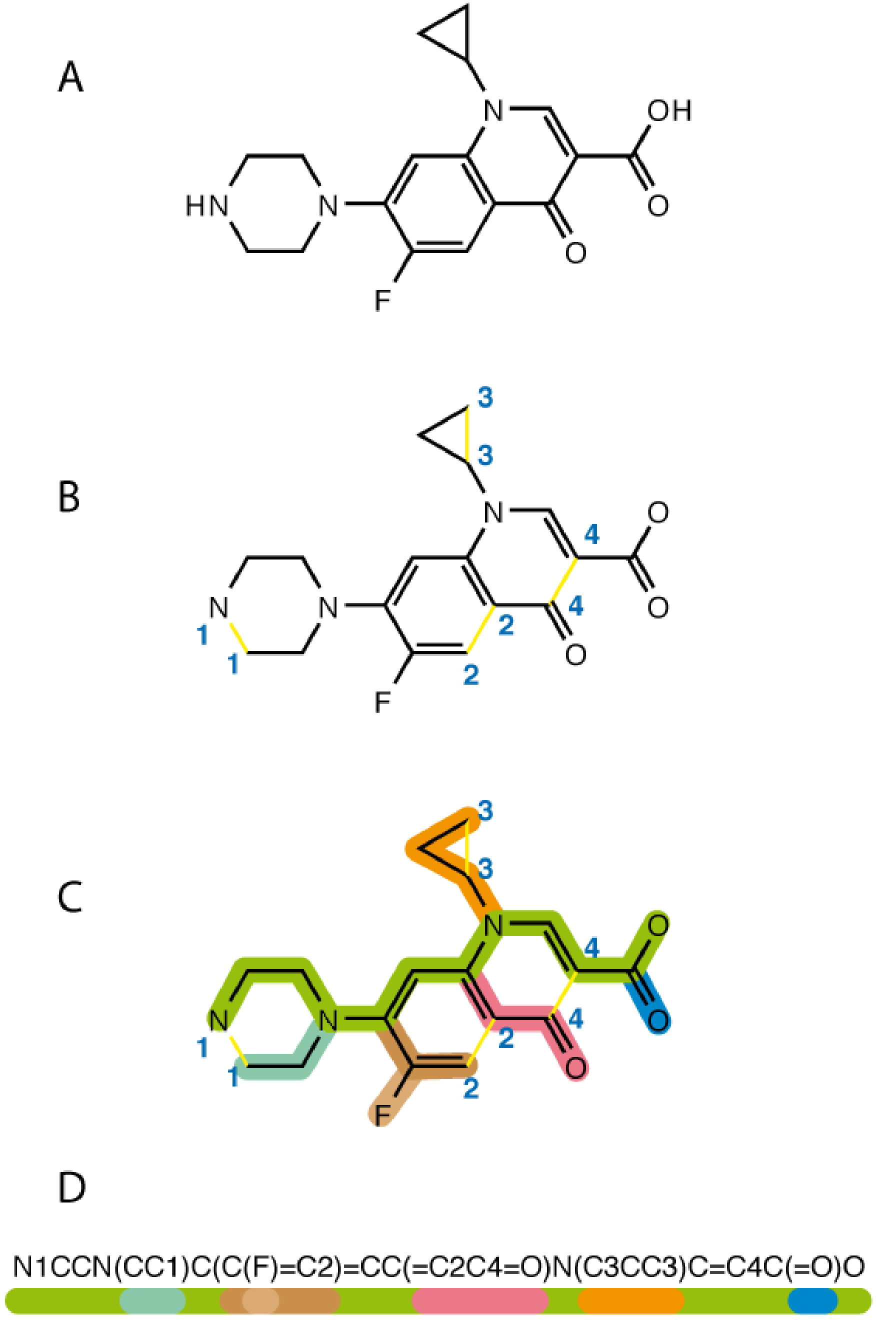
String → graphs is not unique!

Lots of string are invalid

Precise 3D information is lost

Short range in graph may become long range in string

A better way is to encode/decode graph directly.



Source: wikipedia.org

#REF: Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *arXiv preprint arXiv:1610.02415* (2016).

# Better approach: Generating molecular graphs directly

**No regular, fixed-size structures**

Graphs are ***permutation invariant***:

- #permutations are exponential function of #nodes
- The probability of a generated graph G need to be marginalized over all possible permutations

**Multiple objectives:**

- **Diversity** of generated graphs
- **Smoothness** of latent space
- Agreement with or optimization of multiple “**drug-like**” objectives

# GraphVAE

Handles irregular structures

- Predict the whole adjacency matrix, node types and edge types

Deals with variable size graph

- Bounded by the size of the largest graph in training data.

Handles permutation invariance

- Matching every pair of nodes in 2 graphs

Partially promotes diversity

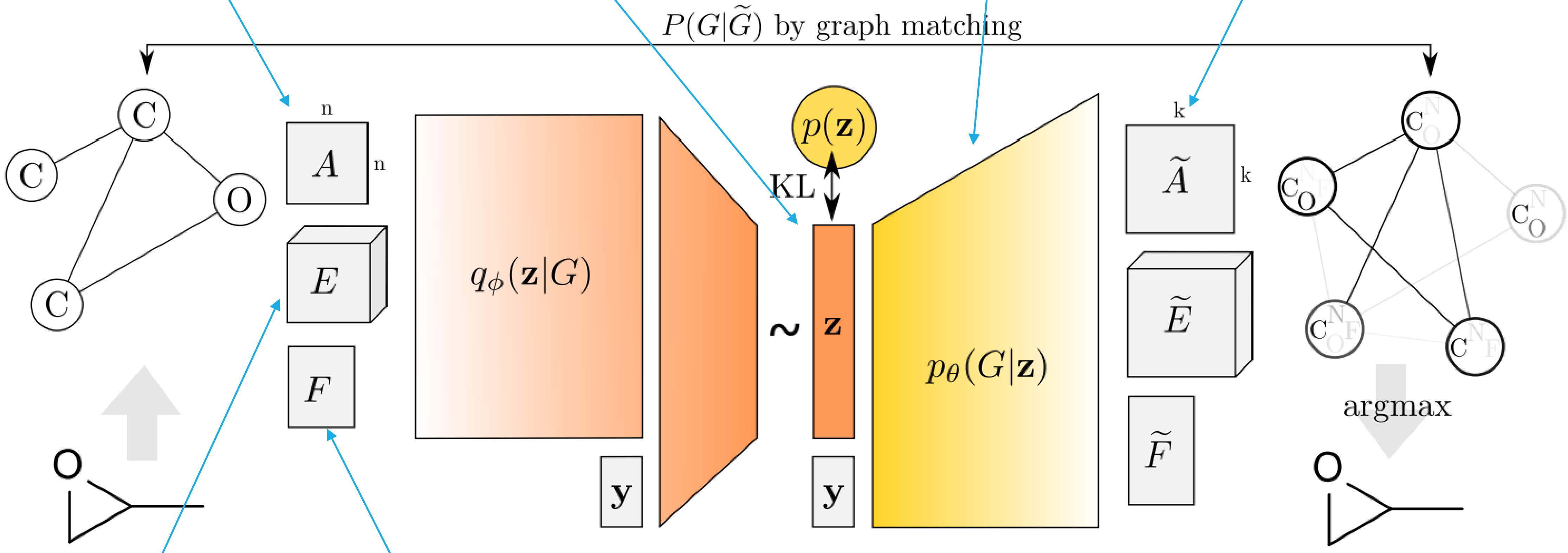
#REF: Simonovsky, M., & Komodakis, N. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *arXiv preprint arXiv:1802.03480.*

Adjacency matrix

Latent vector for whole graph

The graph size are bounded

k>n



Edge types

Node types

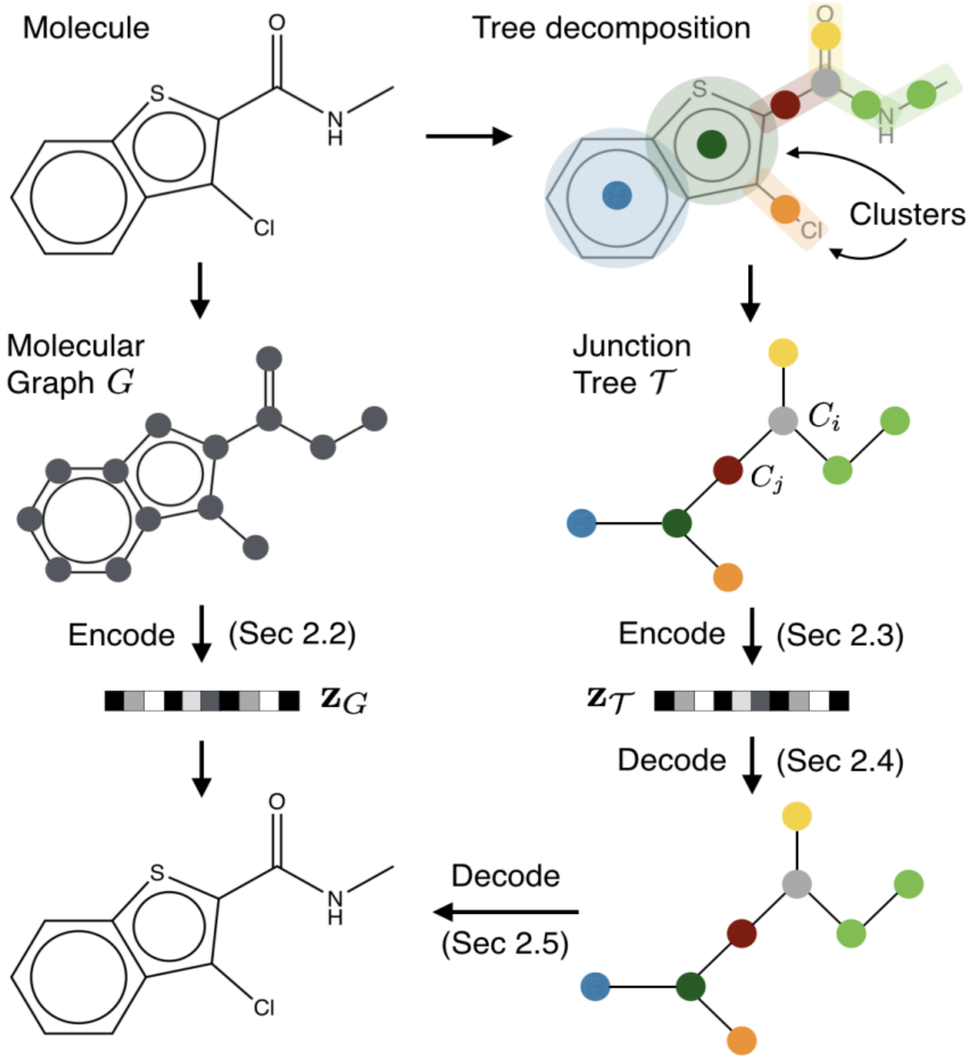
#REF: Simonovsky, M., & Komodakis, N. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *arXiv preprint arXiv:1802.03480.*

# Junction tree VAE

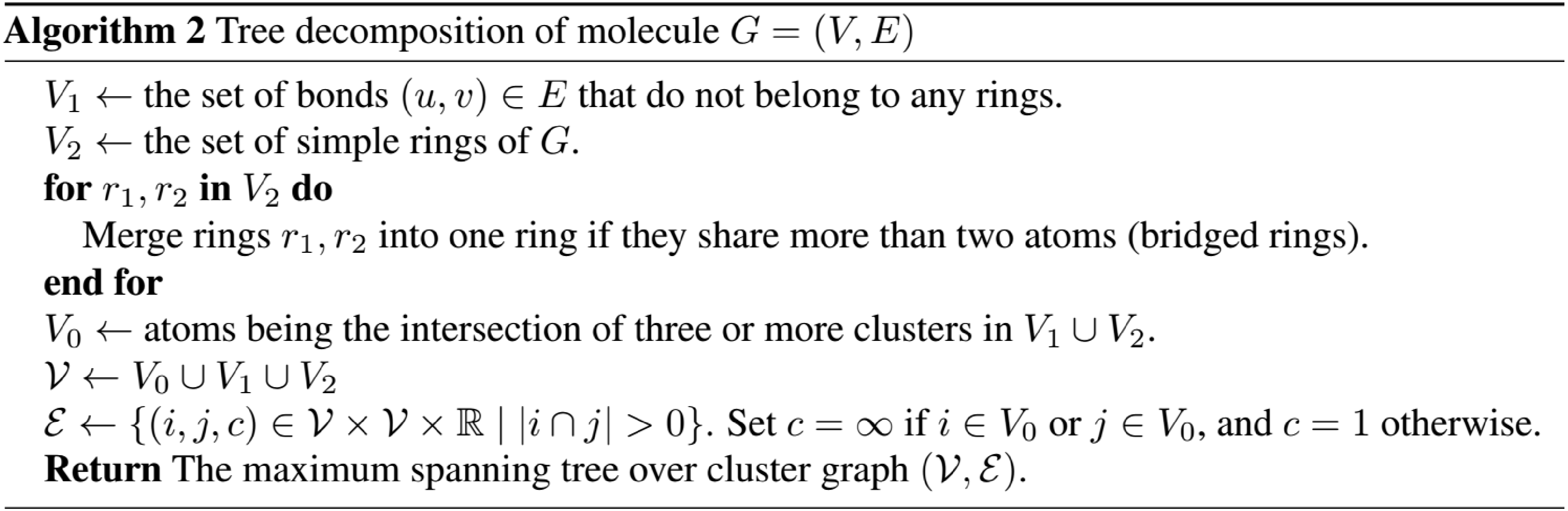
Junction tree is a way to build a “thick-tree” out of a graph

Cluster vocab:

- rings
- bonds
- atoms

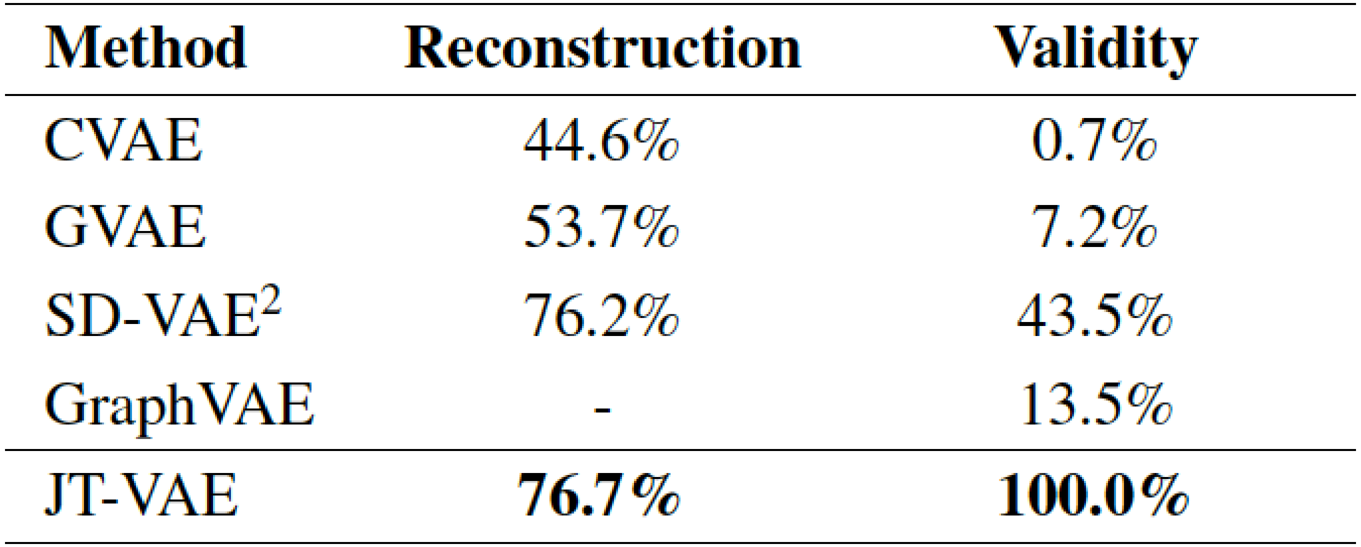


Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *ICML’18*.



**Algorithm 2** Tree decomposition of molecule $G = (V, E)$

```
$V_1 \leftarrow$ the set of bonds $(u, v) \in E$ that do not belong to any rings.
$V_2 \leftarrow$ the set of simple rings of $G$.
for $r_1, r_2$ in $V_2$ do
    Merge rings $r_1, r_2$ into one ring if they share more than two atoms (bridged rings).
end for
$V_0 \leftarrow$ atoms being the intersection of three or more clusters in $V_1 \cup V_2$.
$\mathcal{V} \leftarrow V_0 \cup V_1 \cup V_2$
$\mathcal{E} \leftarrow \{(i, j, c) \in \mathcal{V} \times \mathcal{V} \times \mathbb{R} \mid |i \cap j| > 0\}$. Set $c = \infty$ if $i \in V_0$ or $j \in V_0$, and $c = 1$ otherwise.
Return The maximum spanning tree over cluster graph $(\mathcal{V}, \mathcal{E})$.
```



| Method | Reconstruction | Validity |
|---|---|---|
| CVAE | 44.6% | 0.7% |
| GVAE | 53.7% | 7.2% |
| SD-VAE$^2$ | 76.2% | 43.5% |
| GraphVAE | - | 13.5% |
| JT-VAE | **76.7%** | **100.0%** |

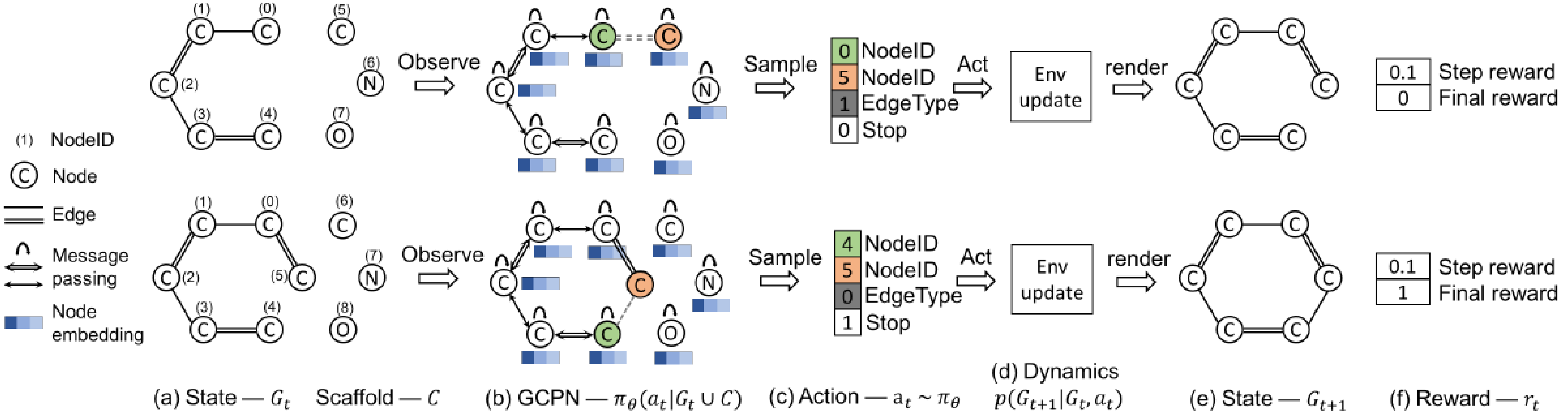
Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *ICML'18*.

# Graphs + Reinforcement learning

Generative graphs are very hard to get it right: The space is too large!

Reinforcement learning offers step-wise construction: one piece at a time

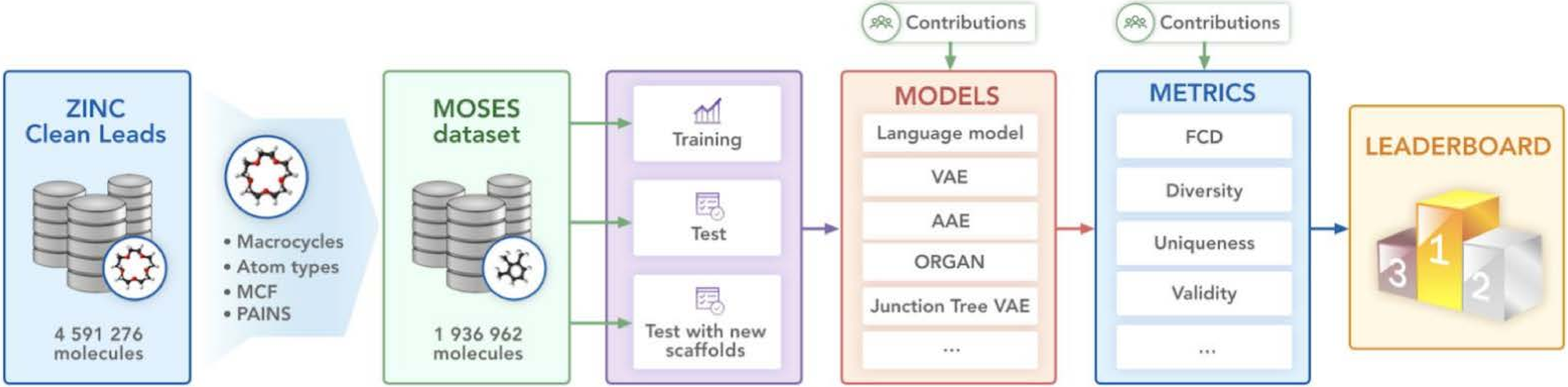
- A.k.a. Markov decision processes
- As before: Graphs offer properties estimation



(a) State — $G_t$ Scaffold — $C$ (b) GCPN — $\pi_\theta(a_t|G_t \cup C)$ (c) Action — $a_t \sim \pi_\theta$ (d) Dynamics $p(G_{t+1}|G_t, a_t)$ (e) State — $G_{t+1}$ (f) Reward — $r_t$

You, Jiaxuan, et al. "Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation." *NeurIPS* (2018).

# Play ground: MOSES



https://medium.com/neuromation-io-blog/moses-a-40-week-journey-to-the-promised-land-of-molecular-generation-78b29453f75c

# The outlook

Read an extremely long book of DNA and answer any queries about it

- Memory-augmented neural networks (MANN), and
- Multiple hierarchical attentions and grammars

Instead of read, write (DNA/viruses/RNA/proteins)

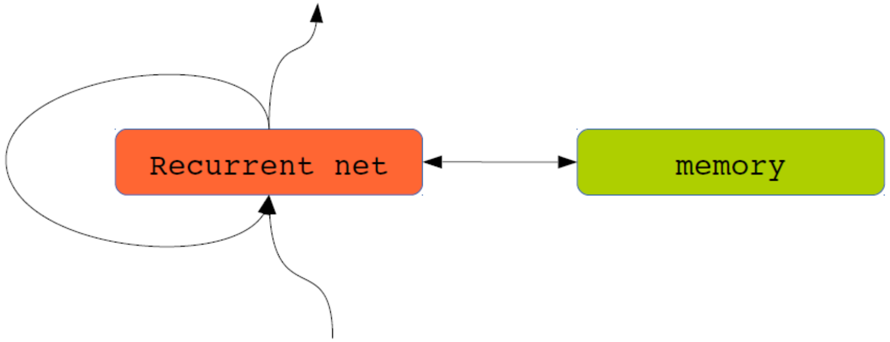
Supper-rich genome SNP annotation

The society of things (DNA/RNA/protein)

Transfer learning between cell types, tissues and diseases

Biology-driven deep nets (e.g., knowledge as memory)

Handling rare events (e.g., the role of memory)



(LeCun, 2015)

# References

Ching, Travers, et al. "Opportunities And Obstacles For Deep Learning In Biology And Medicine." *bioRxiv* (2018): 142760

Eser, Umut, and L. Stirling Churchman. "FIDDLE: An integrative deep learning framework for functional genomic data inference." *bioRxiv* (2016): 081380.

Leung, Michael KK, et al. "Deep learning of the tissue-regulated splicing code." *Bioinformatics* 30.12 (2014): i121-i129.

Lanchantin, Jack, Ritambhara Singh, and Yanjun Qi. "Memory Matching Networks for Genomic Sequence Classification." *arXiv preprint arXiv:1702.06760* (2017).

Pham, Trang, et al. "Column Networks for Collective Classification."*AAAI*. 2017

Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." *Nucleic acids research* 44.11 (2016): e107-e107.

Teng , Haotien, et al. "Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning", *GigaScience*, Volume 7, Issue 5, 1 May 2018, giy037.

Wagstaff, K. L. (2012, June). Machine learning that matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning* (pp. 1851-1856). Omnipress.

Altae-Tran, Han, et al. "Low Data Drug Discovery with One-Shot Learning." *ACS central science* 3.4 (2017): 283-293.

Angermueller, Christof, et al. "Deep learning for computational biology." *Molecular systems biology* 12.7 (2016): 878.

Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." *Advances in neural information processing systems*. 2015.

# References (cont.)

Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *arXiv preprint* arXiv:1610.02415 (2016).

Gupta, Anvita, et al. "Generative Recurrent Networks for De Novo Drug Design." *Molecular Informatics* (2017).

Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv preprint arXiv:1802.04364*.

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., & Zhavoronkov, A. (2017). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 8(7), 10883.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., & Zhavoronkov, A. (2017). druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics*, 14(9), 3098-3104.

Kien Do, et al. "Attentional Multilabel Learning over Graphs-A message passing approach." *arXiv preprint* arXiv:1804.00293(2018).

Kien Do, Truyen Tran, Svetha Venkatesh, "Learning Deep Matrix Representations", *arXiv preprint* arXiv:1703.01454

Kusner, Matt J., Brooks Paige, and José Miguel Hernández-Lobato. "Grammar Variational Autoencoder." *arXiv preprint* arXiv:1703.01925 (2017).

Penmatsa, Aravind, Kevin H. Wang, and Eric Gouaux. "X-ray structure of dopamine transporter elucidates antidepressant mechanism." *Nature* 503.7474 (2013): 85-90.

Pham, Trang, Truyen Tran, and Svetha Venkatesh. "Graph Memory Networks for Molecular Activity Prediction."*ICPR'18*.

Roses, Allen D. "Pharmacogenetics in drug discovery and development: a translational perspective." *Nature reviews Drug discovery* 7.10 (2008): 807-817.

Segler, Marwin HS, et al. "Generating focused molecule libraries for drug discovery with recurrent neural networks." *arXiv preprint* arXiv:1701.01329 (2017).

Simonovsky, M., & Komodakis, N. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *arXiv preprint arXiv:1802.03480*.