Deep Learning Applications in the Natural Sciences

P.Baldi

Department of Computer Science
Institute for Genomics and Bioinformatics
Center for Machine Learning and Intelligent Systems
University of California, Irvine

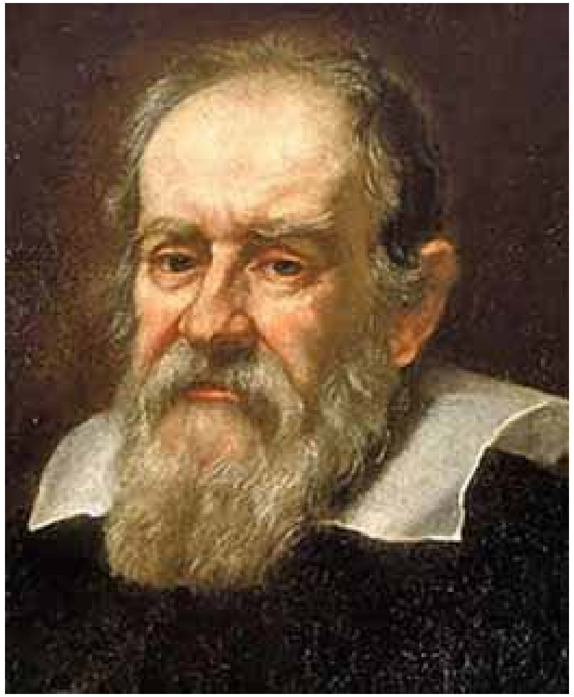
Scientific Discovery Drivers

1. Data (Sensors, Instruments, Data Bases, Internet, Storage…)

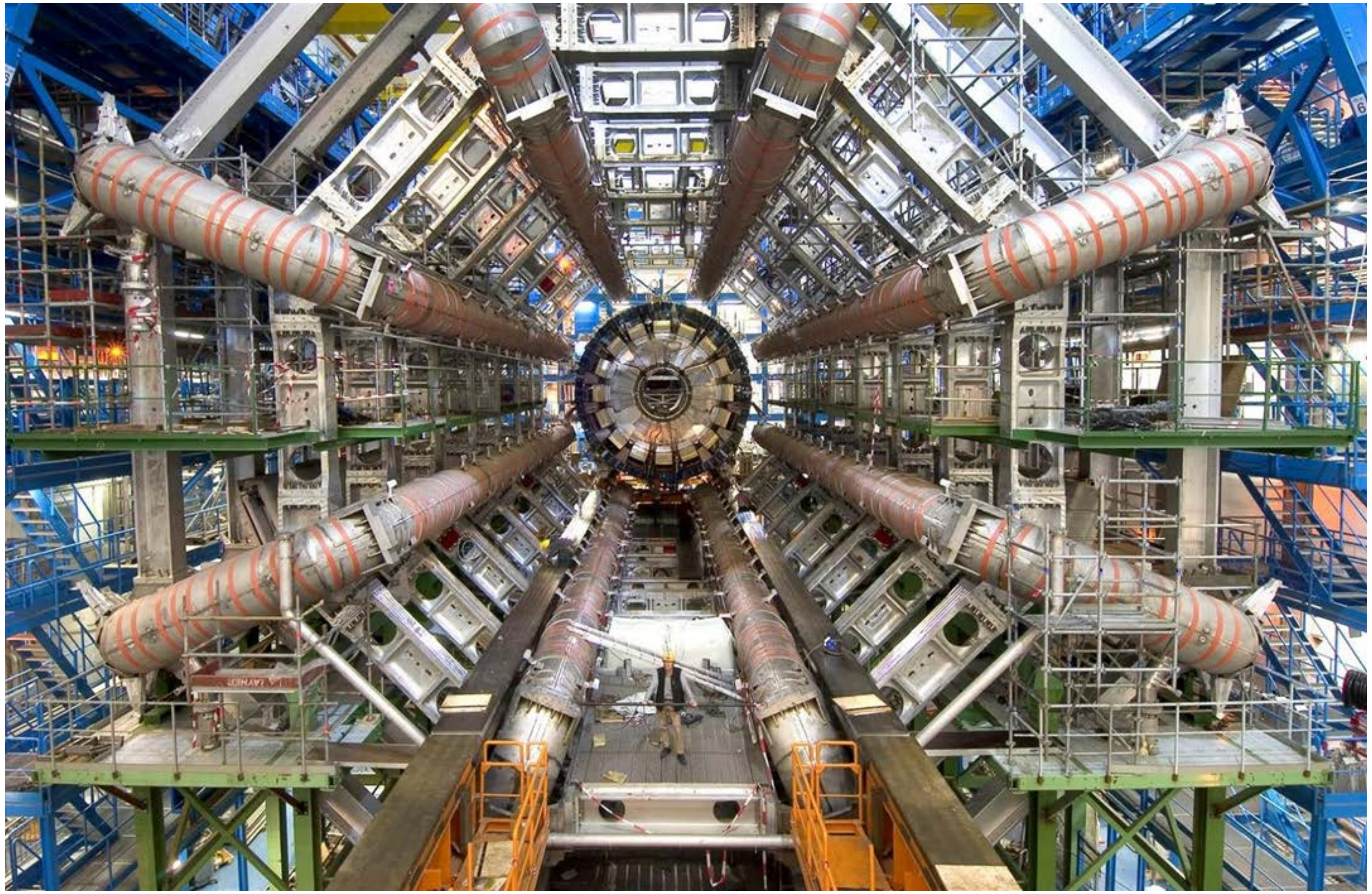
2. Computing (Clusters, Cloud, GPUs…)

3. Machine Learning (AI, Statistics, Data Mining, Algorithms…)

Scientific Discovery



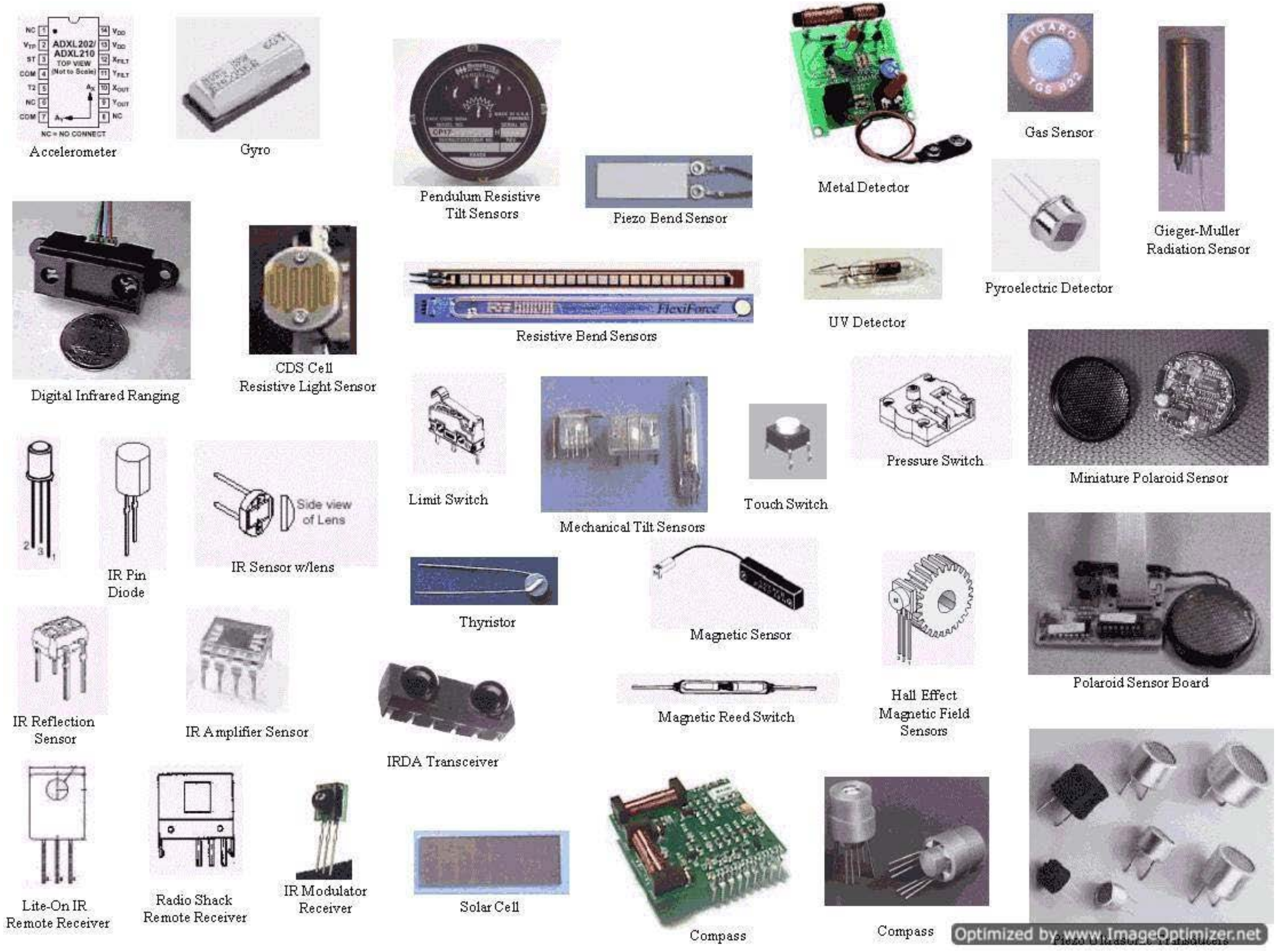
Cosmological frontier



Energy frontier

Genetic frontier

tggaagggctaattcactcccaacgaagacaagatatccttgatctgtggatctaccacacacaaggctacttccctgattagcagaactacacaccagggccagggatcagatatccactgacctttggatggtgctacaagctagtaccagttgagccagagaagttagaag
aagccaacaaaggagagaacaccagcttgttacaccctgtgagcctgcatggaatggatgacccggagagagaagtgttagagtggaggtttgacagccgcctagcatttcatcacatggcccgagagctgcatccggagtacttcaagaactgctgacatcgagcttgcta
caagggactttccgctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagatcctgcatataagcagctgctttttgcctgtactgggtctctctggttagaccagatctgagcctgggagctctctggctaactagggaacccactgcttaagc
ctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctcagaccctttagtcagtgtggaaaatctctagcagtggcgcccgaacagggacctgaaagcgaaagggaaaccagaggagctctctcgacgcaggactc
ggcttgctgaagcgcgcacggcaagaggcgaggggcggcgactggtgagtacgccaaaaattttgactagcggaggctagaaggagagagatgggtgcgagagcgtcagtattaagcgggggagaattagatcgatgggaaaaaattcggttaaggccagggggaa
agaaaaaatataaattaaaacatatagtatgggcaagcagggagctagaacgattcgcagttaatcctggcctgttagaaacatcagaaggctgtagacaaatactgggacagctacaaccatcccttcagacaggatcagaagaacttagatcattatataatacagtagca
accctctattgtgtgcatcaaaggatagagataaaagacaccaaggaagctttagacaagatagaggaagagcaaaacaaaagtaagaaaaaagcacagcaagcagcagctgacacaggacacagcaatcaggtcagccaaaattaccctatagtgcagaacatcc
aggggcaaatggtacatcaggccatatcacctagaactttaaatgcatgggtaaaagtagtagaagagaaggctttcagcccagaagtgatacccatgttttcagcattatcagaaggagccaccccacaagatttaaacaccatgctaaacacagtggggggacatcaagc
agccatgcaaatgttaaaagagaccatcaatgaggaagctgcagaatgggatagagtgcatccagtgcatgcagggcctattgcaccaggccagatgagagaaccaaggggaagtgacatagcaggaactactagtacccttcaggaacaaataggatggatgacaa
ataatccacctatcccagtaggagaaatttataaaagatggataatcctgggattaaataaaatagtaagaatgtatagccctaccagcattctggacataagacaaggaccaaaggaaccctttagagactatgtagaccggttctataaaactctaagagccgagcaagctt
cacaggaggtaaaaaattggatgacagaaaccttgttggtccaaaatgcgaacccagattgtaagactattttaaaagcattgggaccagcggctacactagaagaaatgatgacagcatgtcagggagtaggaggacccggccataaggcaagagttttggctgaagca
atgagccaagtaacaaattcagctaccataatgatgcagagaggcaattttaggaaccaaagaaagattgttaagtgtttcaattgtggcaaagaagggcacacagccagaaattgcagggcccctaggaaaaagggctgttggaaatgtggaaaggaaggacaccaaa
tgaaagattgtactgagagacaggctaattttttagggaagatctggccttcctacaagggaaggccagggaattttcttcagagcagaccagagccaacagccccaccagaagagagcttcaggtctggggtagagacaacaactccccctcagaagcaggagccgata
gacaaggaactgtatccttaacttccctcaggtcactctttggcaacgacccctcgtcacaataaagataggggggcaactaaaggaagctctattagatacaggagcagatgatacagtattagaagaaatgagtttgccaggaagatggaaaccaaaaatgatagggg
aattggaggttttatcaaagtaagacagtatgatcagatactcatagaaatctgtggacataaagctataggtacagtattagtaggacctacacctgtcaacataattggaagaaatctgttgactcagattggttgcactttaaattttcccattagccctattgagactgtaccagtaa
aattaaagccaggaatggatggcccaaaagttaaacaatggccattgacagaagaaaaaataaaagcattagtagaaatttgtacagagatggaaaaggaagggaaaatttcaaaaattgggcctgaaaatccatacaatactccagtatttgccataaagaaaaaagac
agtactaaatggagaaaattagtagatttcagagaacttaataagagaactcaagacttctgggaagttcaattaggaataccacatcccgcagggttaaaaaagaaaaaatcagtaacagtactggatgtgggtgatgcatatttttcagttcccttagatgaagacttcaggaa
gtatactgcatttaccatacctagtataaacaatgagacaccagggattagatatcagtacaatgtgcttccacagggatggaaaggatcaccagcaatattccaaagtagcatgacaaaaatcttagagccttttagaaaacaaaatccagacatagttatctatcaatacatgg
atgatttgtatgtaggatctgacttagaaatagggcagcatagaacaaaaatagaggagctgagacaacatctgttgaggtggggacttaccacaccagacaaaaaacatcagaaagaacctccattcctttggatgggttatgaactccatcctgataaatggacagtacagc
ctatagtgctgccagaaaaagacagctggactgtcaatgacatacagaagttagtggggaaattgaattgggcaagtcagatttacccagggattaaagtaaggcaattatgtaaactccttagaggaaccaaagcactaacagaagtaataccactaacagaagaagcag
agctagaactggcagaaaacagagagattctaaaagaaccagtacatggagtgtattatgacccatcaaaagacttaatagcagaaatacagaagcaggggcaaggccaatggacatatcaaatttatcaagagccatttaaaaatctgaaaacaggaaaatatgcaag
aatgaggggtgcccacactaatgatgtaaaacaattaacagaggcagtgcaaaaaataaccacagaaagcatagtaatatggggaaagactcctaaatttaaactgcccatacaaaaggaaacatgggaaacatggtggacagagtattggcaagccacctggattcct
gagtgggagtttgttaatacccctcccttagtgaaattatggtaccagttagagaaagaacccatagtaggagcagaaaccttctatgtagatggggcagctaacagggagactaaattaggaaaagcaggatatgttactaatagaggaagacaaaaagttgtcaccctaact
gacacaacaaatcagaagactgagttacaagcaatttatctagctttgcaggattcgggattagaagtaaacatagtaacagactcacaatatgcattaggaatcattcaagcacaaccagatcaaagtgaatcagagttagtcaatcaaataatagagcagttaataaaaaa
ggaaaaggtctatctggcatgggtaccagcacacaaaggaattggaggaaatgaacaagtagataaattagtcagtgctggaatcaggaaagtactatttttagatggaatagataaggcccaagatgaacatgagaaatatcacagtaattggagagcaatggctagtgat
tttaacctgccacctgtagtagcaaaagaaatagtagccagctgtgataaatgtcagctaaaaggagaagccatgcatggacaagtagactgtagtccaggaatatggcaactagattgtacacatttagaaggaaaagttatcctggtagcagttcatgtagccagtggatata
tagaagcagaagttattccagcagaaacagggcaggaaacagcatattttcttttaaaattagcaggaagatggccagtaaaaacaatacatactgacaatggcagcaatttcaccggtgctacggttagggccgcctgttggtgggcgggaatcaagcaggaatttggaattc
cctacaatccccaaagtcaaggagtagtagaatctatgaataaagaattaaagaaaattataggacaggtaagagatcaggctgaacatcttaagacagcagtacaaatggcagtattcatccacaattttaaaagaaaaggggggattggggggtacagtgcaggggaa
agaatagtagacataatagcaacagacatacaaactaaagaattacaaaaacaaattacaaaaattcaaaattttcgggtttattacagggacagcagaaatccactttggaaaggaccagcaaagctcctctggaaaggtgaaggggcagtagtaatacaagataatagt
gacataaaagtagtgccaagaagaaaagcaaagatcattagggattatggaaaacagatggcaggtgatgattgtgtggcaagtagacaggatgaggattagaacatggaaaagtttagtaaaacaccatatgtatgtttcagggaaagctaggggatggttttatagacatc
actatgaaagccctcatccaagaataagttcagaagtacacatcccactaggggatgctagattggtaataacaacatattggggtctgcatacaggagaaagagactggcatttgggtcagggagtctccatagaatggaggaaaaagagatatagcacacaagtagacc
ctgaactagcagaccaactaattcatctgtattactttgactgtttttcagactctgctataagaaaggccttattaggacacatagttagccctaggtgtgaatatcaagcaggacataacaaggtaggatctctacaatacttggcactagcagcattaataacaccaaaaaagata
aagccacctttgcctagtgttacgaaactgacagaggatagatggaacaagccccagaagaccaagggccacagagggagccacacaatgaatggacactagagcttttagaggagcttaagaatgaagctgttagacattttcctaggatttggctccatggcttagggca
acatatctatgaaacttatggggatacttgggcaggagtggaagccataataagaattctgcaacaactgctgtttatccattttcagaattgggtgtcgacatagcagaataggcgttactcgacagaggagagcaagaaatggagccagtagatcctagactagagccctgga
agcatccaggaagtcagcctaaaactgcttgtaccaattgctattgtaaaaagtgttgctttcattgccaagtttgtttcataacaaaagccttaggcatctcctatggcaggaagaagcggagacagcgacgaagagctcatcagaacagtcagactcatcaagcttctctatcaa
agcagtaagtagtacatgtaacgcaacctataccaatagtagcaatagtagcattagtagtagcaataataatagcaatagttgtgtggtccatagtaatcatagaatataggaaaatattaagacaaagaaaaatagacaggttaattgatagactaatagaaagagcagaa
gacagtggcaatgagagtgaaggagaaatatcagcacttgtggagatgggggtggagatggggcaccatgctccttgggatgttgatgatctgtagtgctacagaaaaattgtgggtcacagtctattatggggtacctgtgtggaaggaagcaaccaccactctattttgtgcat
cagatgctaaagcatatgatacagaggtacataatgtttgggccacacatgcctgtgtacccacagaccccaacccacaagaagtagtattggtaaatgtgacagaaaattttaacatgtggaaaaatgacatggtagaacagatgcatgaggatataatcagtttatgggatc
aaagcctaaagccatgtgtaaaattaaccccactctgtgttagtttaaagtgcactgatttgaagaatgatactaataccaatagtagtagcgggagaatgataatggagaaaggagagataaaaaactgctctttcaatatcagcacaagcataagaggtaaggtgcagaaa
gaatatgcattttttataaacttgatataataccaatagataatgatactaccagctataagttgacaagttgtaacacctcagtcattacacaggcctgtccaaaggtatcctttgagccaattcccatacattattgtgccccggctggttttgcgattctaaaatgtaataataagacgtt
caatggaacaggaccatgtacaaatgtcagcacagtacaatgtacacatggaattaggccagtagtatcaactcaactgctgttaaatggcagtctagcagaagaagaggtagtaattagatctgtcaatttcacggacaatgctaaaaccataatagtacagctgaacacatc
tgtagaaattaattgtacaagacccaacaacaatacaagaaaaagaatccgtatccagagaggaccagggagagcatttgttacaataggaaaaataggaaatatgagacaagcacattgtaacattagtagagcaaaatggaataacactttaaaacagatagctagca
aattaagagaacaatttggaaataataaaacaataatctttaagcaatcctcaggaggggacccagaaattgtaacgcacagttttaattgtggaggggaatttttctactgtaattcaacacaactgtttaatagtacttggtttaatagtacttggagtactgaagggtcaaataaca
ctgaaggaagtgacacaatcaccctcccatgcagaataaaacaaattataaacatgtggcagaaagtaggaaaagcaatgtatgcccctcccatcagtggacaaattagatgttcatcaaatattacagggctgctattaacaagagatggtggtaatagcaacaatgagtcc
gagatcttcagacctggaggaggagatatgagggacaattggagaagtgaattatataaatataaagtagtaaaaattgaaccattaggagtagcacccaccaaggcaaagagaagagtggtgcagagagaaaaaagagcagtgggaataggagctttgttccttgggtt
cttgggagcagcaggaagcactatgggcgcagcctcaatgacgctgacggtacaggccagacaattattgtctggtatagtgcagcagcagaacaatttgctgagggctattgaggcgcaacagcatctgttgcaactcacagtctggggcatcaagcagctccaggcaaga
atcctggctgtggaaagatacctaaaggatcaacagctcctggggatttggggttgctctggaaaactcatttgcaccactgctgtgccttggaatgctagttggagtaataaatctctggaacagatttggaatcacacgacctggatggagtgggacagagaaattaacaattac
acaagcttaatacactccttaattgaagaatcgcaaaaccagcaagaaaagaatgaacaagaattattggaattagataaatgggcaagtttgtggaattggtttaacataacaaattggctgtggtatataaaattattcataatgatagtaggaggcttggtaggtttaagaatag
tttttgctgtactttctatagtgaatagagttaggcagggatattcaccattatcgtttcagacccacctcccaaccccgaggggacccgacaggcccgaaggaatagaagaagaaggtggagagagagacagagacagatccattcgattagtgaacggatccttggcacttat
ctgggacgatctgcggagcctgtgcctcttcagctaccaccgcttgagagacttactcttgattgtaacgaggattgtggaacttctgggacgcagggggtgggaagccctcaaatattggtggaatctcctacagtattggagtcaggaactaaagaatagtgctgttagcttgctc
aatgccacagccatagcagtagctgaggggacagatagggttatagaagtagtacaaggagcttgtagagctattcgccacatacctagaagaataagacagggcttggaaaggattttgctataagatgggtggcaagtggtcaaaaagtagtgtgattggatggcctactg
taagggaaagaatgagacgagctgagccagcagcagatagggtgggagcagcatctcgagacctggaaaaacatggagcaatcacaagtagcaatacagcagctaccaatgctgcttgtgcctggctagaagcacaagaggaggaggaggtgggttttccagtcaca
cctcaggtacctttaagaccaatgacttacaaggcagctgtagatcttagccactttttaaaagaaaaggggggactggaagggctaattcactcccaaagaagacaagatatccttgatctgtggatctaccacacacaaggctacttccctgattagcagaactacacaccag
ggccaggggtcagatatccactgacctttggatggtgctacaagctagtaccagttgagccagataagatagaagaggccaataaaggagagaacaccagcttgttacaccctgtgagcctgcatgggatggatgacccggagagagaagtgttagagtggaggtttgaca
gccgcctagcatttcatcacgtggcccgagagctgcatccggagtacttcaagaactgctgacatcgagcttgctacaagggactttccgctggggactttccagggaggcgtggcctgggcgggactggggagtggcgagccctcagatcctgcatataagcagctgctttttg
cctgtactgggtctctctggttagaccagatctgagcctgggagctctctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtctgttgtgtgactctggtaactagagatccctcagacccttttagtcagtgtggaaaatct
ctagca


Accelerometer
Gyro
Pendulum Resistive Tilt Sensors
Piezo Bend Sensor
Metal Detector
Gas Sensor
Gieger-Muller Radiation Sensor
Pyroelectric Detector
UV Detector
Resistive Bend Sensors
FlexiForce
Digital Infrared Ranging
CDS Cell Resistive Light Sensor
Limit Switch
Mechanical Tilt Sensors
Touch Switch
Pressure Switch
Miniature Polaroid Sensor
IR Pin Diode
Side view of Lens
IR Sensor w/lens
Thyristor
Magnetic Sensor
Hall Effect Magnetic Field Sensors
Polaroid Sensor Board
IR Reflection Sensor
IR Amplifier Sensor
IRDA Transceiver
Magnetic Reed Switch
Lite-On IR Remote Receiver
Radio Shack Remote Receiver
IR Modulator Receiver
Solar Cell
Compass
Compass

Scientific Discovery Drivers

1. Data (Sensors, Instruments, Data Bases, Internet, Storage…)
   - -Exponential Growth
   - -Uneveness across Fields

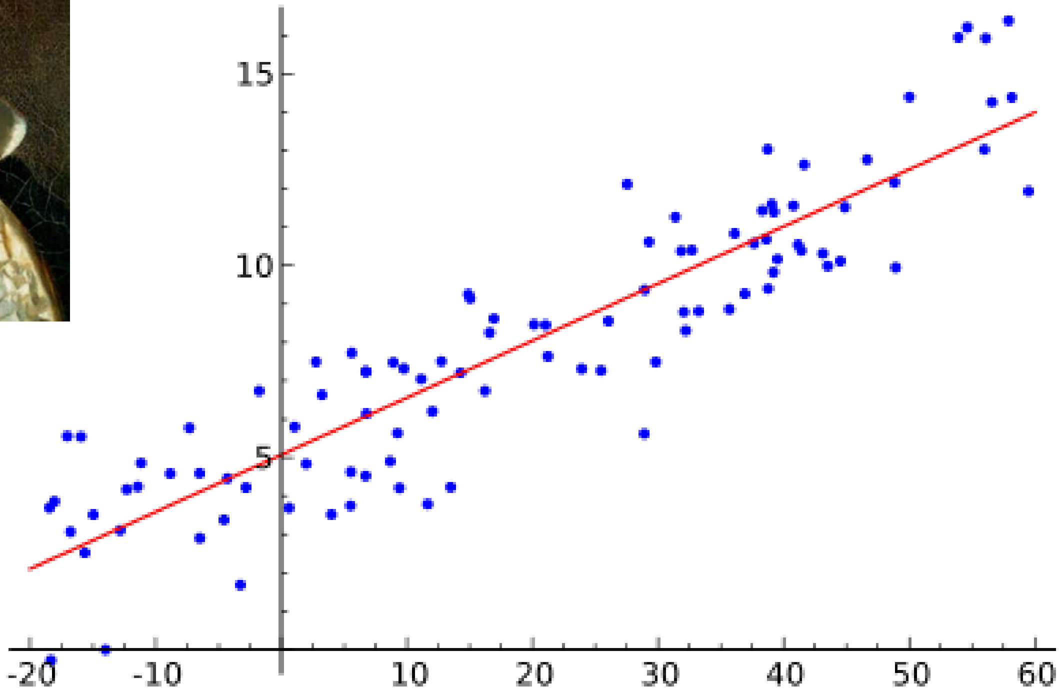
2. **Computing** (Clusters, Cloud, GPUs…)

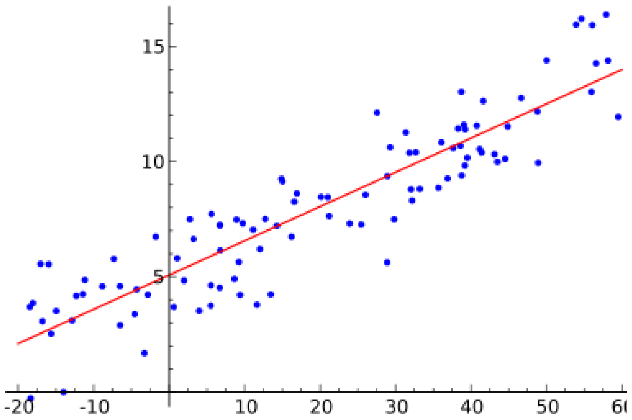
Scientific Discovery Drivers

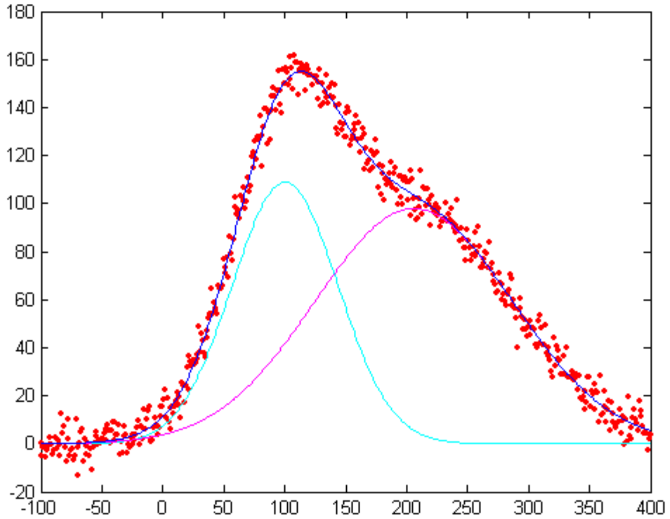
1. Data (Sensors, Instruments, Data Bases, Internet, Storage…)

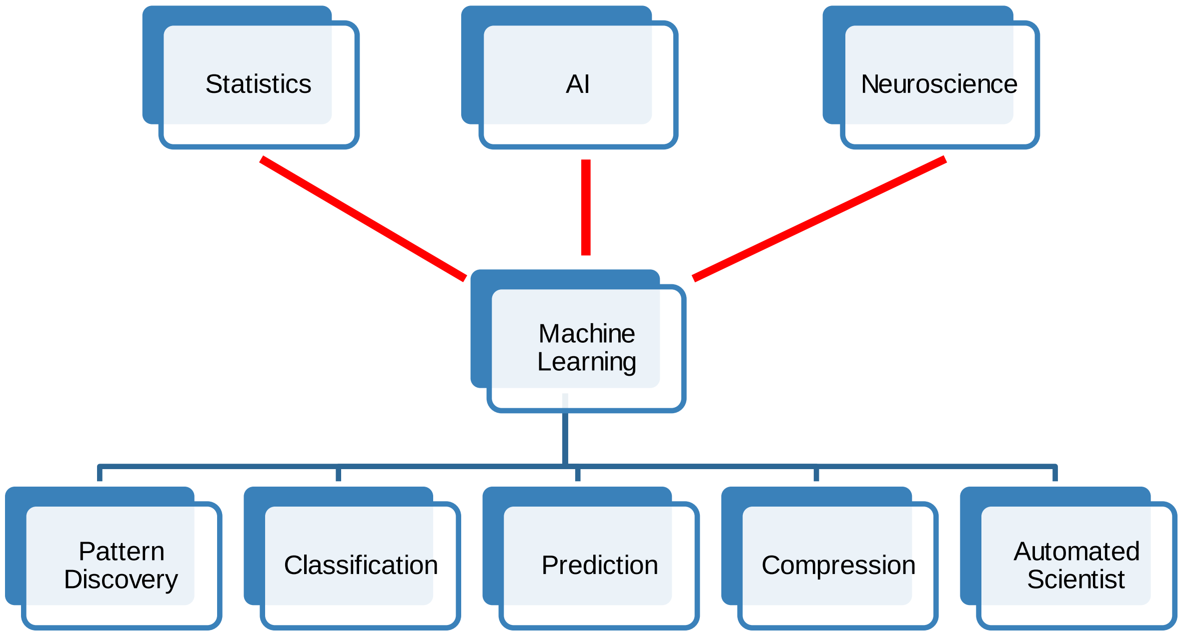
2. Computing (Clusters, Cloud, GPUs…)

3. **Machine Learning** (AI, Statistics, Data Mining, Algorithms…)


15
10
5
-20
-10
10
20
30
40
50
60

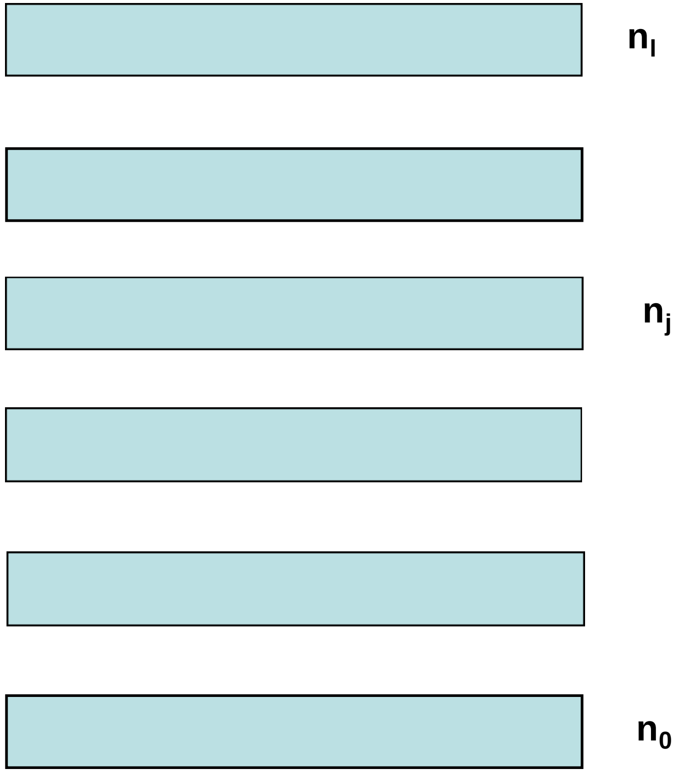

15
10
5
-20
-10
10
20
30
40
50
60


180
160
140
120
100
80
60
40
20
0
-20
-100
-50
0
50
100
150
200
250
300
350
400

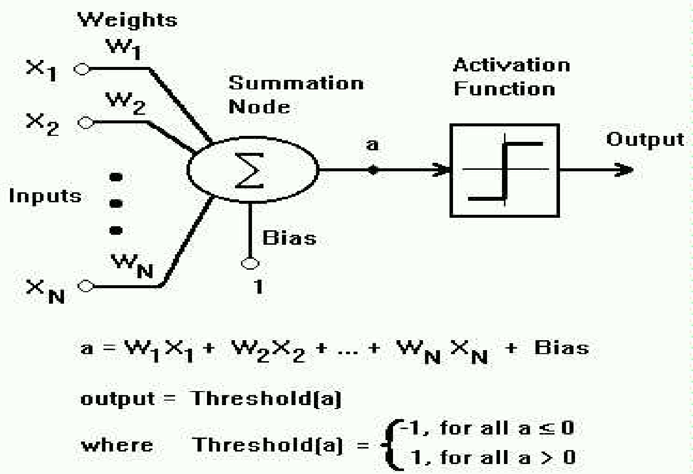


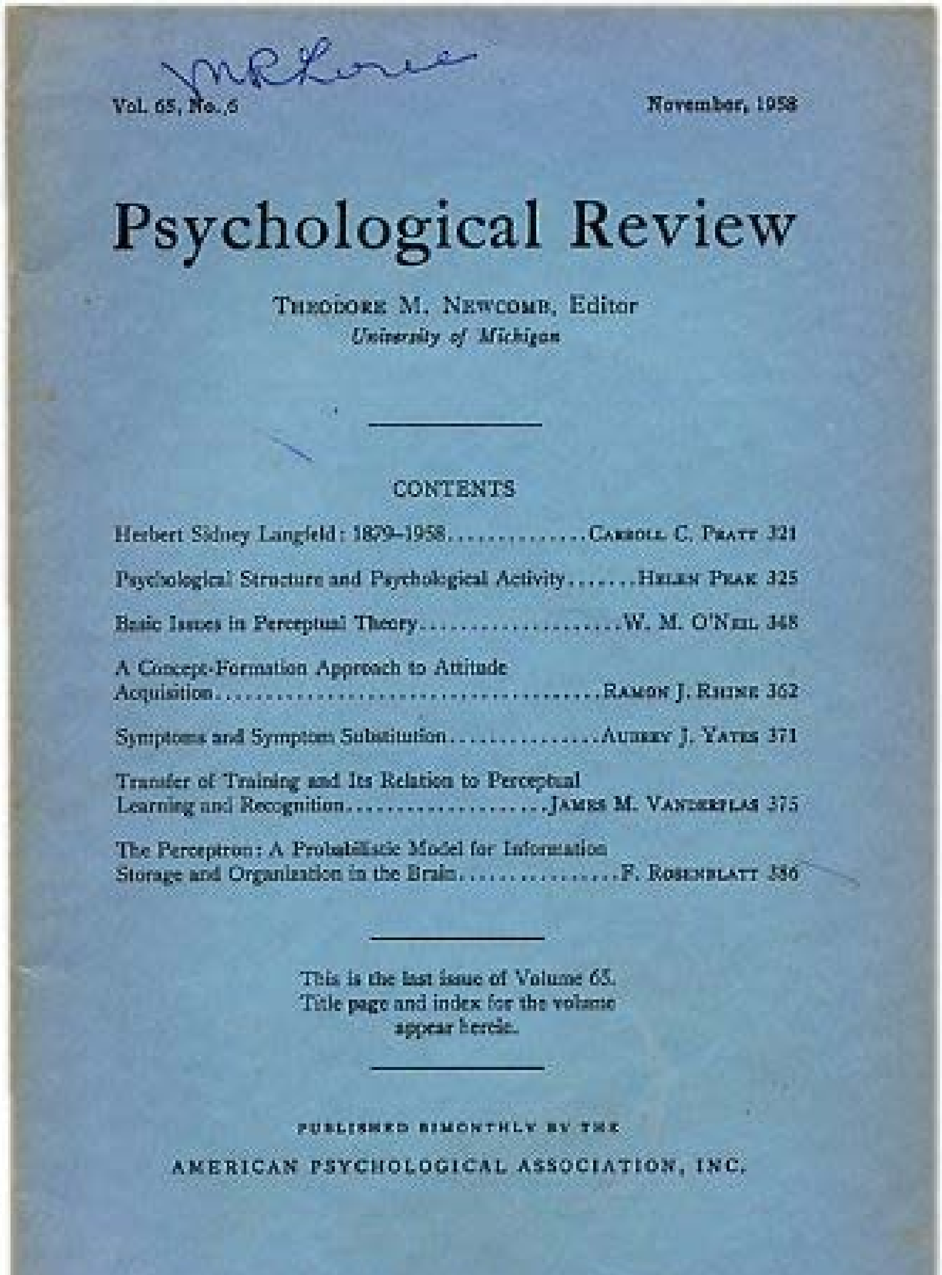
Statistics
AI
Neuroscience
Machine Learning
Pattern Discovery
Classification
Prediction
Compression
Automated Scientist

Deep Learning



Training set: (x_i,y_i)
or (x_i,x_i)
$i=1, . . ., m$





Vol. 65, No. 6 November, 1958

Psychological Review

THEODORE M. NEWCOMB, Editor
University of Michigan

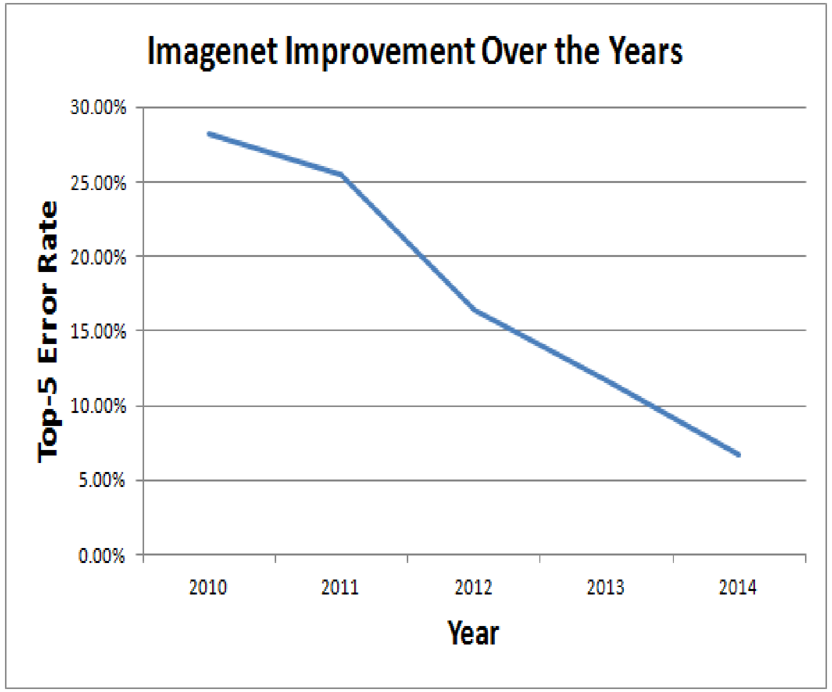
CONTENTS

This is the last issue of Volume 65.
Title page and index for the volume appear herein.

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Computer Vision - Image Classification

- **Imagenet**
- Over 1 million images, 1000 classes, different sizes, avg 482x415, color
- 16.42% Deep CNN dropout in 2012
- 6.66% 22 layer CNN (GoogLeNet) in 2014
- 4.9% (Google, Microsoft) super-human performance in 2015



Sources: Krizhevsky et al ImageNet Classification with Deep Convolutional Neural Networks, Lee et al Deeply supervised nets 2014, Szegedy et al, Going Deeper with convolutions, ILSVRC2014, Sanchez & Perronnin CVPR 2011, http://www.clarifai.com/ Benenson, http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html

DL Theory

- **1-Layer Networks**
  - Perceptron theorem
  - Linear regression; Logistic regression;
  - Statistical theory and design (top layer)

- **1.5-Layer Networks**
  - Bottom layer = random (Johnson-Lindenstrauss)
  - Bottom layer = similarities (dot products or kernels)→ SVM

DL Theory

- **2-Layers Networks**
  - Universal approximation
  - Autoencoders (compressive, expansive)
  - Linear autoencoders (PCA and landscape)
  - Non-linear autoencoders (Boolean autoencoder, clustering, NP-completeness)

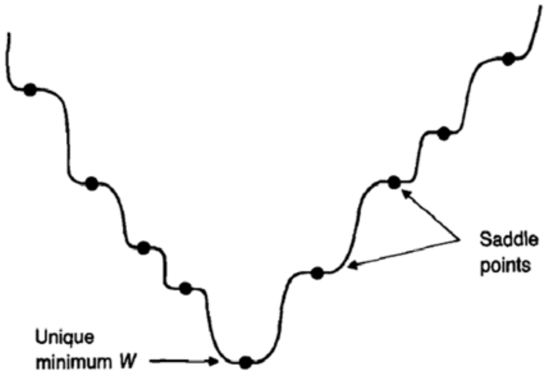


FIGURE 2. The landscape of E.

DL Theory

- **L-Layer Networks**
  - Linear
  - Boolean unrestricted
  - Local learning and its limitations (generalization of Hebb)
  - Optimality of Backpropagation
  - Design (Weight sharing, Compression and Dark Knowledge, etc)
  - Dropout, Initialization, Learning rates, hyperparameter optimization

- **Recurrent Networks**
  - Hopfield Model and Boltzmann machines
  - Design (DAGs, LSTMs, etc)

DL Theory

- **Importance of Group Theory**
  - Learning permutations
  - Permutations of the units
  - Symmetries of learning rules
  - Invariant recognition (Lie Groups)

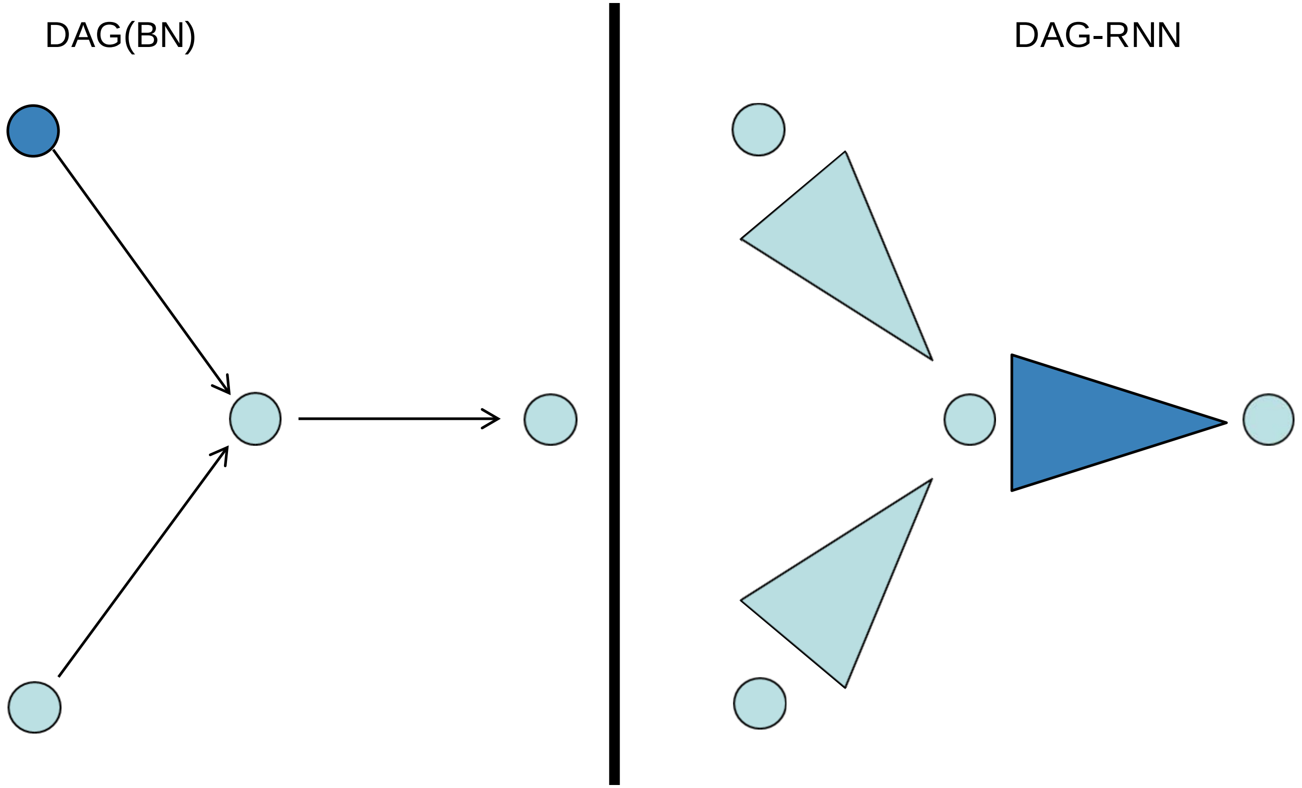
- **The Black-Box Problem…..**

Two Kinds of Problems and Architectures

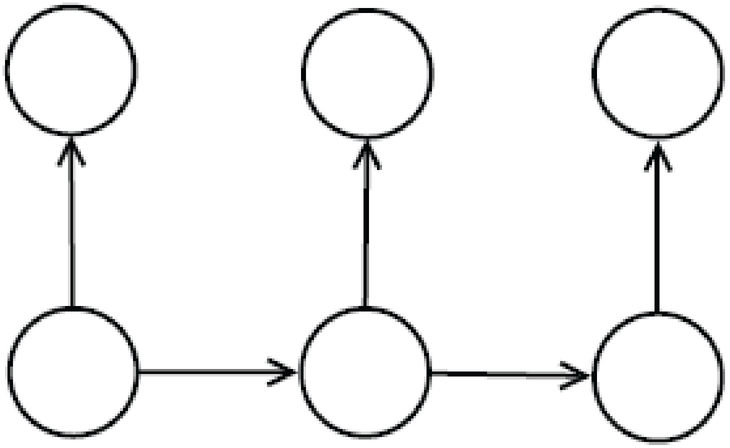
1. Input: vectors of fixed size
(e.g. images in computer vision).
Typical architectures: feedforward neural networks.

2. Input: structured objects of variable size
(e.g. sequences in NLP or bioinformatics, graphs in chemistry).
Typical architectures: recursive neural networks.

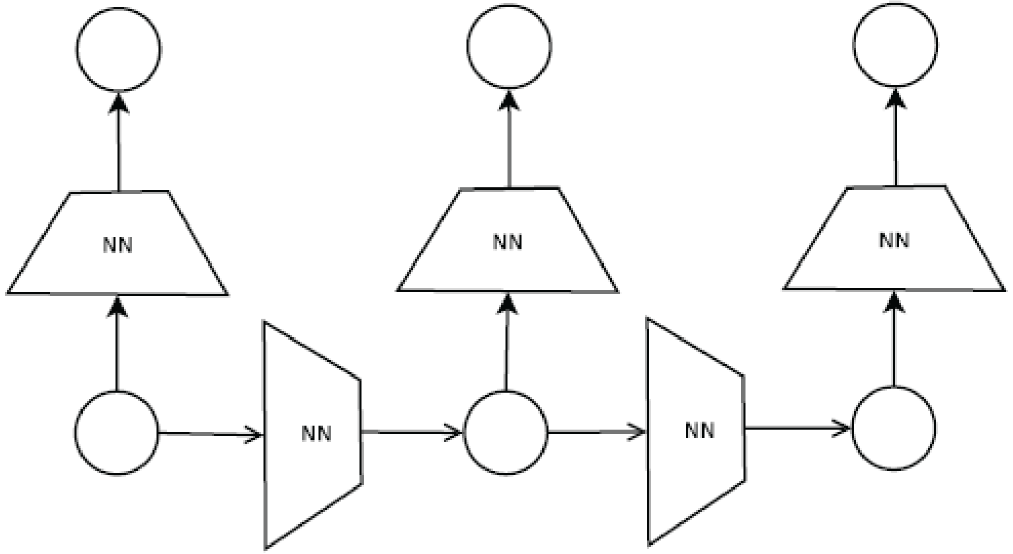
Design (Recursive Architectures)



Graphical Models: DAG-HMM



Recursive Neural Networks: DAG-RNN



Deep Learning in the Natural Sciences

- Physics

    -HEP: Identification of Exotic Particles (<1Å)
    -[Cosmology: Identification of Quasars (10^{26}m)]

- Chemistry

    -Prediction of Molecular Properties and Chemical Reactions (~1-10^{2}Å)

- Biology

    -Prediction of Protein Structures and Structural Features (10^{2}-10^{4}Å)

- Many more

Deep Learning in the Natural Sciences

- Physics

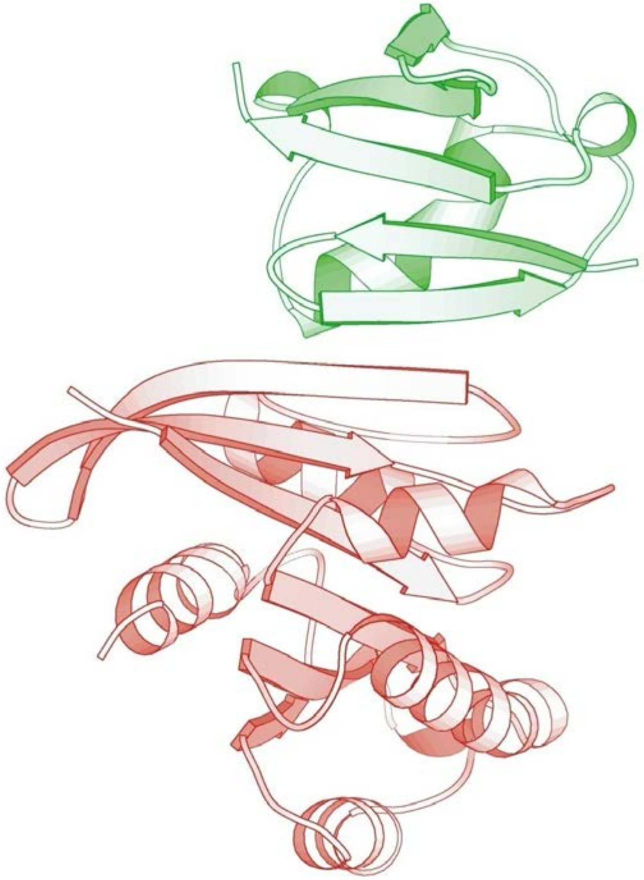
  -HEP: Identification of Exotic Particles (<1Ǻ)

- Chemistry

  -Prediction of Molecular Properties and Chemical Reactions (~1-102Ǻ)

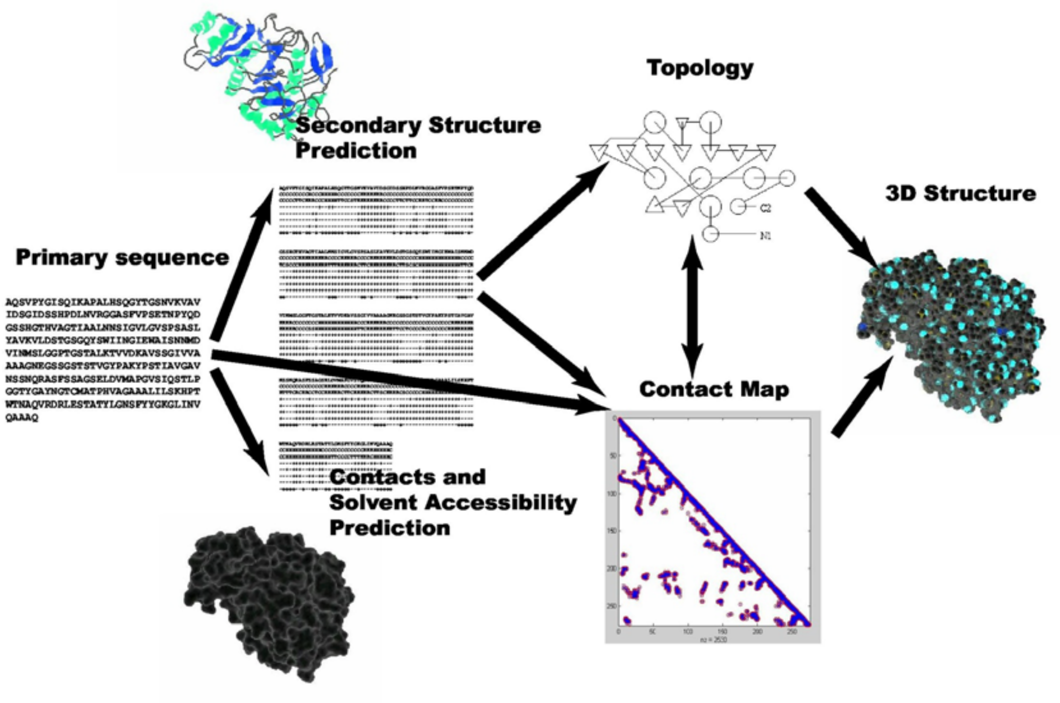
- **Biology**

  -Prediction of Protein Structures and Structural Features (10^2-10^4Ǻ)

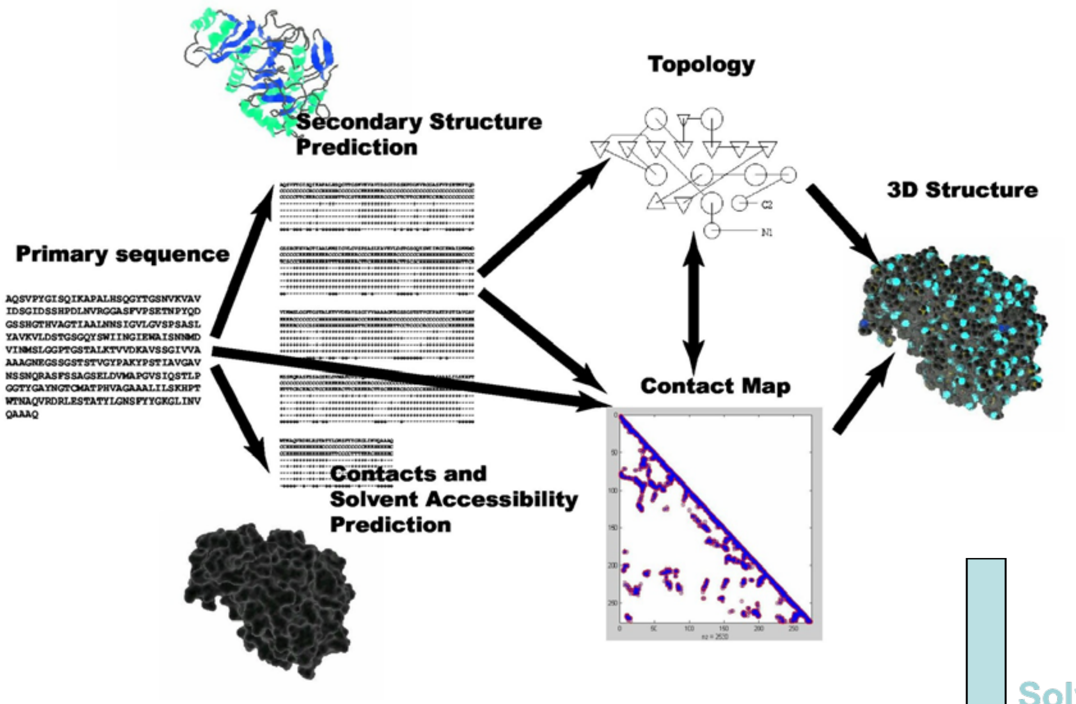


An important, complex, multi-faceted, somewhat ill defined problem.

Deep Learning in Biology

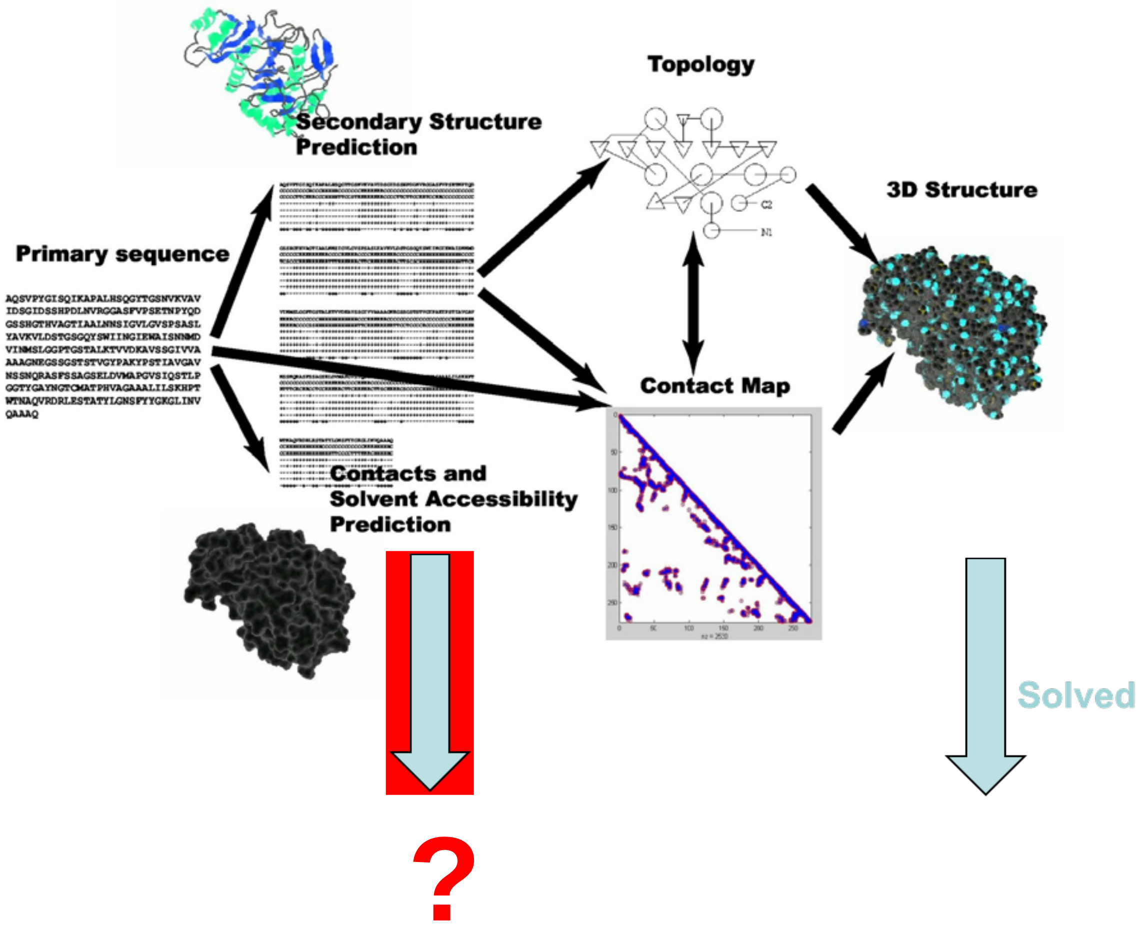


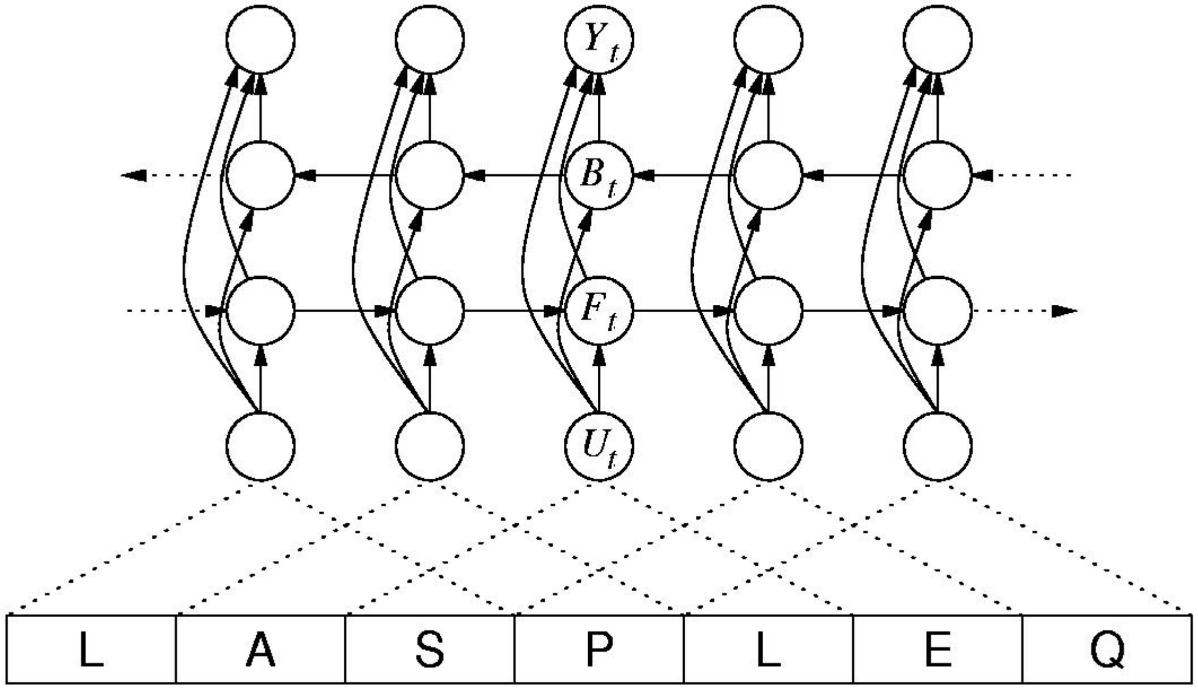
Deep Learning in Biology: Mining Omic Data

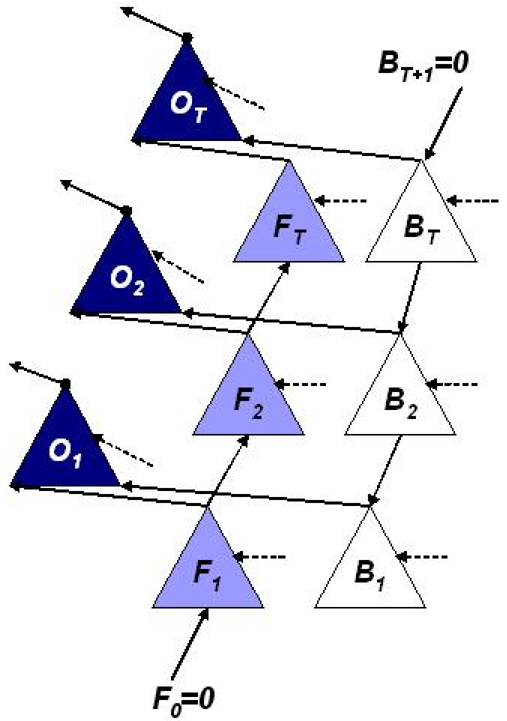


Solved

Deep Learning in Biology: Mining Omic Data




Y_t
B_t
F_t
U_t
L
A
S
P
L
E
Q

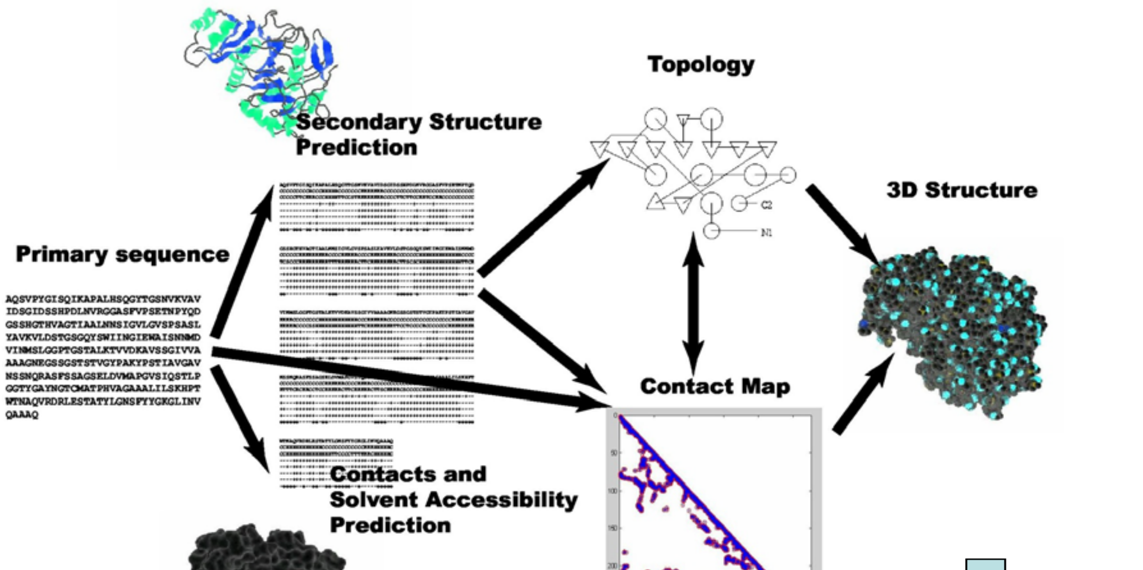


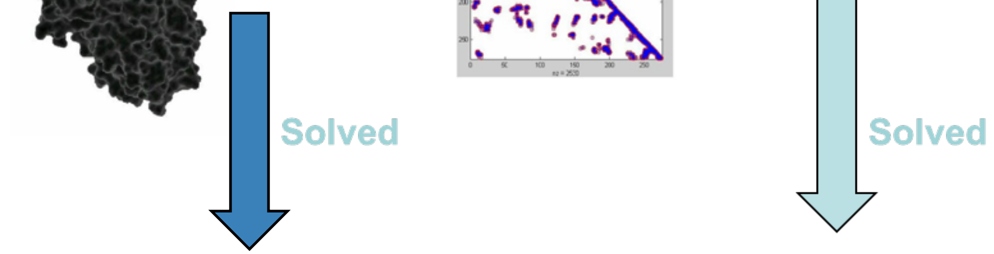
$B_{T+1}=0$
O_T
F_T
B_T
O_2
F_2
B_2
O_1
F_1
B_1
$F_0=0$

Progress in Accuracy

- 1978: Chou and Fasman ~60% (statistical rules)
- 1988: Qian and Sejnowski ~64% (NNs)
- 1994: Rost and Sander ~74% (NNs)
- 1999: B. et al. ~78% (1DBRNNs)
- …..
- 2014: Magnan and B. ~95% (1DBRNNs + homology)

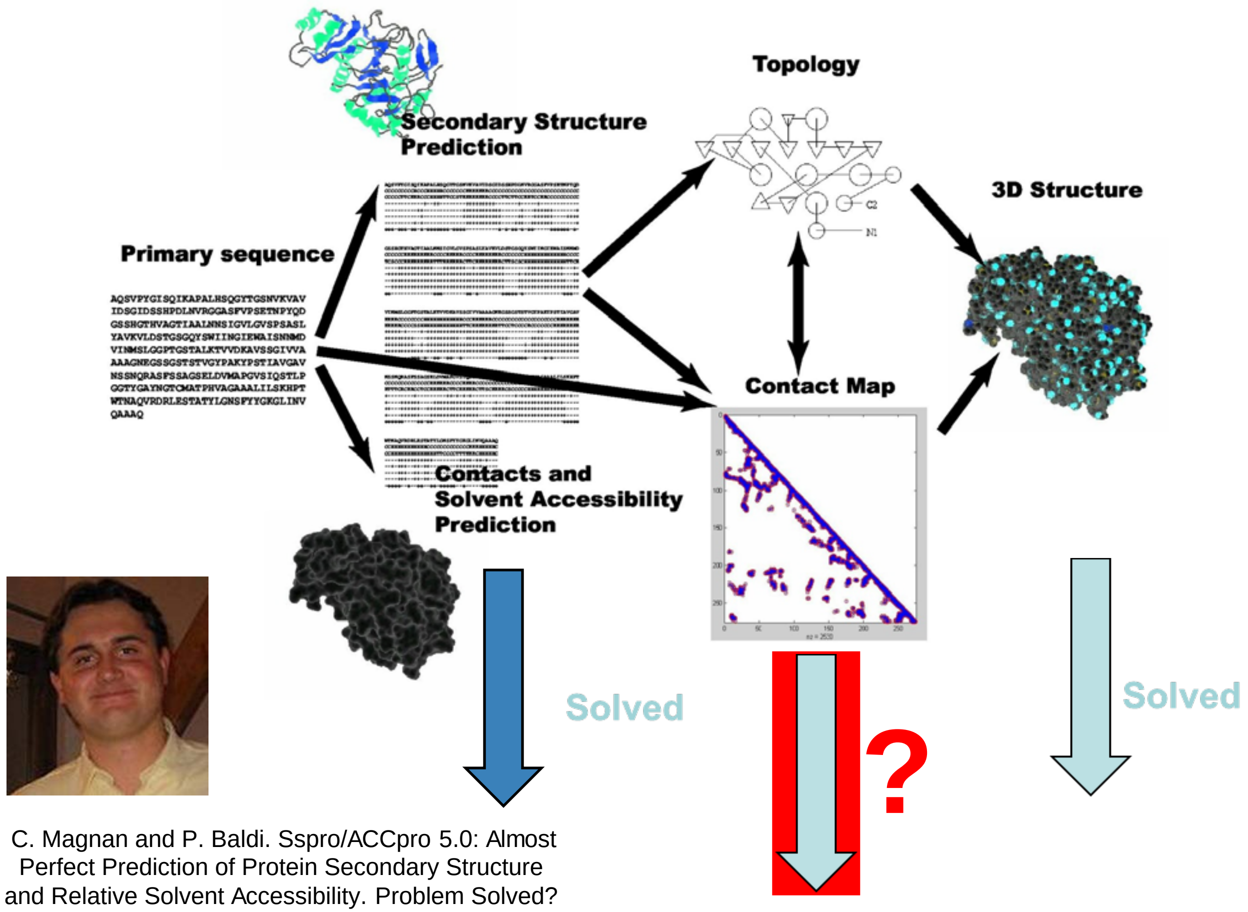
Deep Learning in Biology: Mining Omic Data



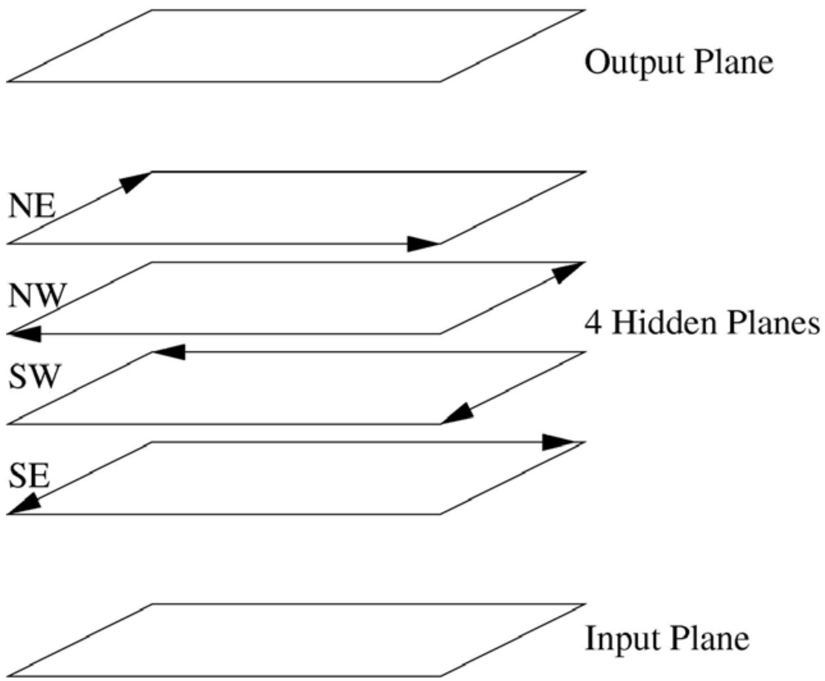


C. Magnan and P. Baldi. Sspro/ACCpro 5.0: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility. Problem Solved? *Bioinformatics*, (advance access June 18), (2014).

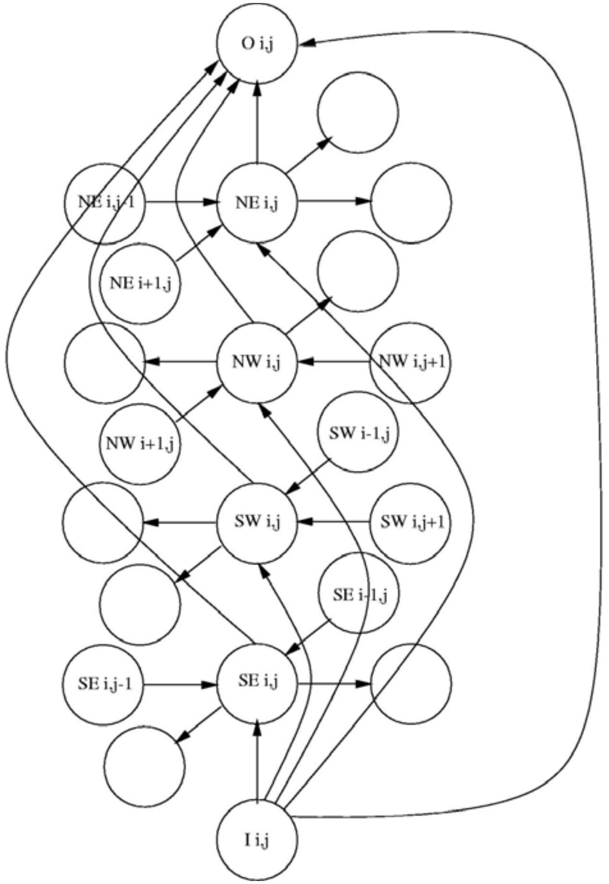
Deep Learning in Biology: Mining Omic Data



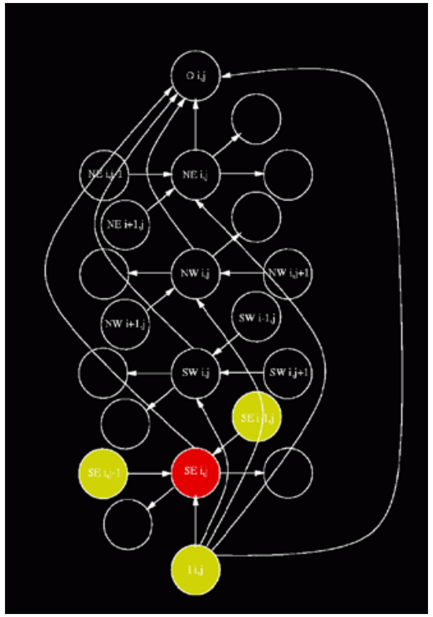
C. Magnan and P. Baldi. Sspro/ACCpro 5.0: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility. Problem Solved? *Bioinformatics*, (advance access June 18), (2014).



(In 3D, 8 hidden cubes, etc……)


O i,j
NE i,j-1
NE i,j
NE i+1,j
NW i,j
NW i,j+1
NW i+1,j
SW i-1,j
SW i,j
SW i,j+1
SE i-1,j
SE i,j-1
SE i,j
I i,j

2D RNNs



$$\begin{cases} O_{ij} = \mathcal{N}_O(I_{ij}, H^{NW}_{i,j}, H^{NE}_{i,j}, H^{SW}_{i,j}, H^{SE}_{i,j,}) \\ H^{NE}_{i,j} = \mathcal{N}_{NE}(I_{i,j}, H^{NE}_{i-1,j}, H^{NE}_{i,j-1}) \\ H^{NW}_{i,j} = \mathcal{N}_{NW}(I_{i,j}, H^{NW}_{i+1,j}, H^{NW}_{i,j-1}) \\ H^{SW}_{i,j} = \mathcal{N}_{SW}(I_{i,j}, H^{SW}_{i+1,j}, H^{SW}_{i,j+1}) \\ H^{SE}_{i,j} = \mathcal{N}_{SE}(I_{i,j}, H^{SE}_{i-1,j}, H^{SE}_{i,j+1}) \end{cases}$$

P. Baldi and G. Pollastri. The Principled Design of Large-Scale Recursive Neural Network Architectures—DAG-RNNs and the Protein Structure Prediction Problem. *Journal of Machine Learning Research*, 4, 575-602, (2003).

RR Results Summary - CASP10

10th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

RR Analysis

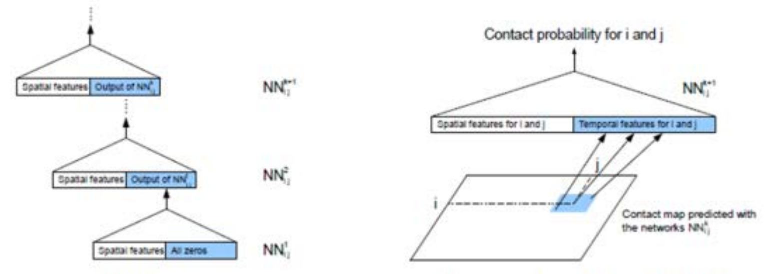
The table summarizes the evaluation of predictions in 'RR' category.
The analysis was performed at per domains basis; only predictions for domains classified as "FM", "TBM/FM", "TBM hard" were considered.
The groups were ranked according to sum of average Z-scores for two measures Acc and Xd.
The per target Z-scores were recalculated from the "cleaned" distributions, where the outlier predictions (below mean - 2 std dev) were eliminated.

- **Domain classification:**
  - FM
  - TBM/FM
  - TBM hard (max gdt_ts < 50)
- **Contact Range:** long
- **List Size:** L/5

#	GR#	GR Name	Count domains	AVG Acc	AVG Zscore Acc	AVG Xd	AVG Zscore Xd	Zscore Acc + Zscore Xd
1.	222 s	MULTICOM-CONSTRUCT	14	19.41	0.58	12.08	0.77	1.35
2.	305 s	IGBteam	15	19.22	0.72	10.19	0.58	1.30
3.	424 s	MULTICOM-NOVEL	14	20.39	0.50	10.32	0.72	1.22
4.	125 s	MULTICOM-REFINE	14	21.35	0.51	10.29	0.70	1.21
5.	413 s	ZHOU-SPARKS-X	12	12.26	0.62	8.26	0.59	1.21
6.	113 s	SAM-T08-server	11	16.13	0.72	9.44	0.47	1.19
7.	358 s	RaptorX-Roll	8	12.07	0.58	8.23	0.55	1.13
8.	314 s	ProC_S4	14	17.91	0.59	9.76	0.47	1.05
9.	087 s	Distill_roll	15	13.97	0.60	8.57	0.36	0.96
10.	489	MULTICOM	14	12.96	0.43	8.19	0.40	0.83
11.	184 s	ICOS	14	17.03	0.40	9.72	0.39	0.78
12.	396 s	ProC_S5	14	16.51	0.36	9.10	0.36	0.72
13.	381 s	SAM-T06-server	10	10.98	0.37	7.94	0.31	0.68
14.	332 s	PLCT	15	11.07	0.30	7.53	0.31	0.61
15.	139 s	CONSIP	9	10.99	0.32	7.35	0.27	0.59
16.	334 s	RBO-CON	3	8.93	0.22	7.54	0.26	0.49
17.	112 s	samcha-server	13	13.34	0.17	7.19	0.28	0.45
18.	257 s	ProC_S3	15	14.02	0.18	7.58	0.21	0.39
19.	081 s	MULTICOM-CLUSTER	11	8.27	0.08	7.31	0.29	0.37
20.	072 s	Distill	15	10.47	0.13	6.72	0.17	0.31

Deep Learning

P. Di Lena, K. Nagata, and P. Baldi.
Deep Architectures for Protein Contact Map Prediction.
Bioinformatics, 28, 2449-2457.



(a) DST-NN architecture

(b) Temporal input features for NN^k_{ij}

Figure 1: DST-NN architecture. (a) Overview. Each NN^k_{ij} represents a feed-forward neural network trainable by back-propagation. (b) For a pair of residues (i, j), the temporal inputs into NN^{k+1}_{ij} consist of the contact probabilities produced by the network at the previous level over a neighborhood of (i, j).

Deep Learning in the Natural Sciences

- Physics

  -HEP: Identification of Exotic Particles (<1Ǻ)

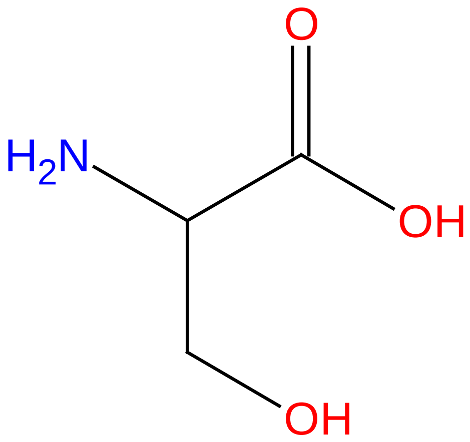
- **Chemistry**

  -Prediction of Molecular Properties and Chemical Reactions (~1-10^2Ǻ)

- Biology

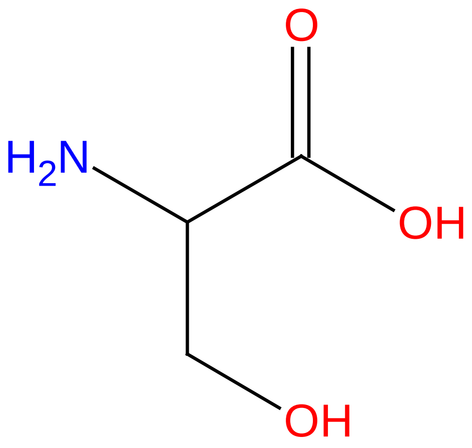
  -Prediction of Protein Structures and Structural Features (10^2-10^4Ǻ)

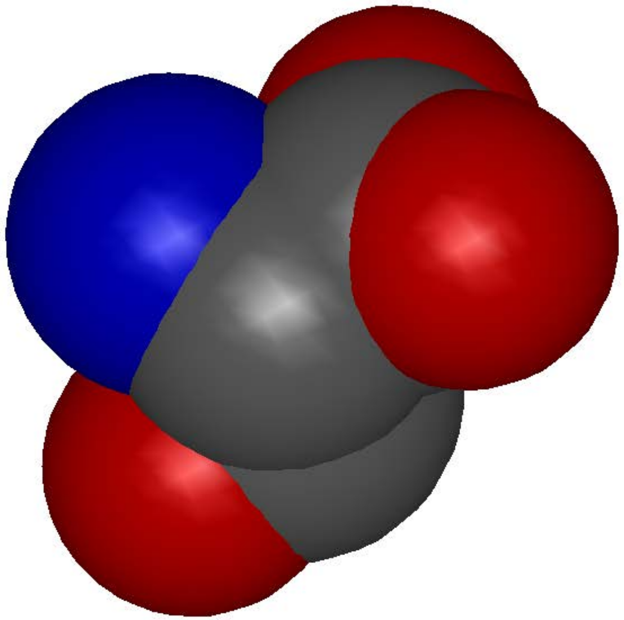
Prediction of Molecular Properties (Physical, Chemical, Biological)



Melting Temperature? Soluble? Toxic? etc

Data Representations

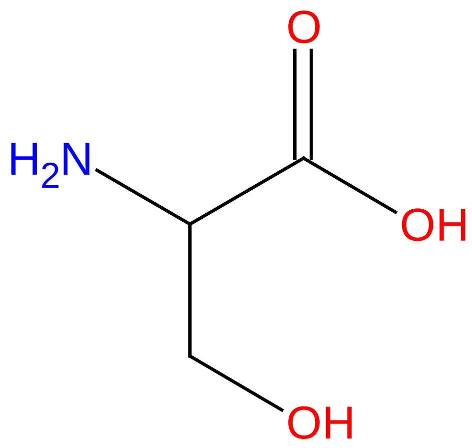




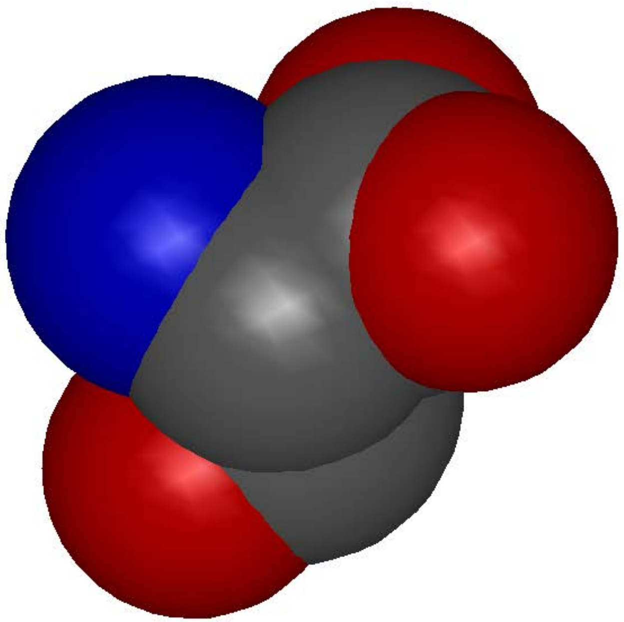
NC(CO)C(=O)O

0010001001010001

Data Representations

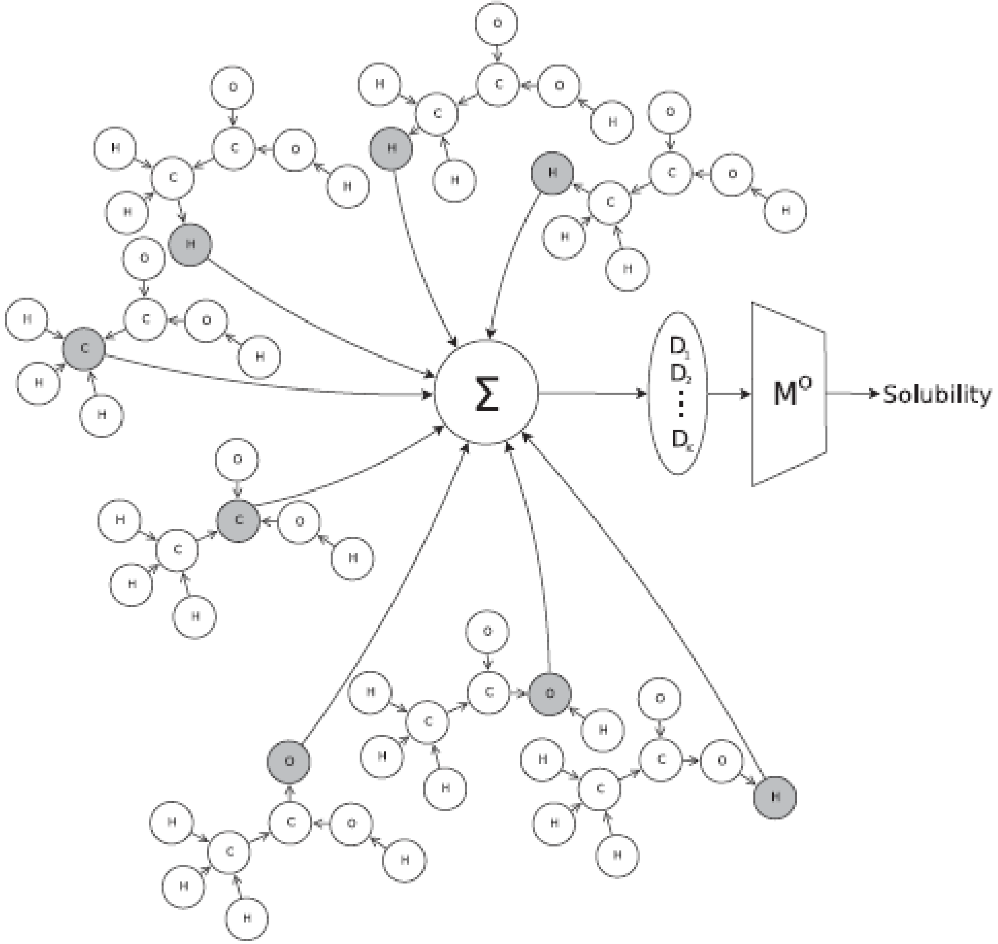


Problem: molecular graphs are undirected



NC(CO)C(=O)O

0010001001010001



A. Lusci, G. Pollastri, and P. Baldi. Deep Architectures and Deep Learning in Chemoinformatics: the Prediction of Aqueous Solubility for Drug-Like Molecules*. Journal of Chemical Information and Modeling*, **53, 7, 1563–1575, (2013).**

Deep Learning in Chemistry: Predicting Chemical Reactions

$RCH{=}CH_2 + HBr \rightarrow RCH(Br){-}CH_3$

- Many important applications (synthesis, retrosynthesis, etc)
- Three different approaches:
    1. QM
    2. Write a system of rules
    3. Learn the rules from big data

Writing a System of Rules: Reaction Explorer

Table 1. SMIRKS Transformation Rules Corresponding to a Simple Alkene Hydrobromination Reaction Model[a]

SMIRKS	description
[C:1]=[C:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition
[C+:1].[-:2]≫ [C+0:1][+0:2]	carbocation, anion addition

J. Chen and P. Baldi. No Electron Left-Behind: a Rule-Based Expert System to Predict Chemical Reactions and Reaction Mechanisms. *Journal of Chemical Information and Modeling*, 49, 9, 2034-2043, (2009).

- ReactionExplorer System has about 1800 rules
- Covers undergraduate organic chemistry curriculum
- Interactive educational system
- Licensed by Wiley and distributed world-wide

Problems

- Very tedious
- Non-scalable
- Undergraduate chemistry

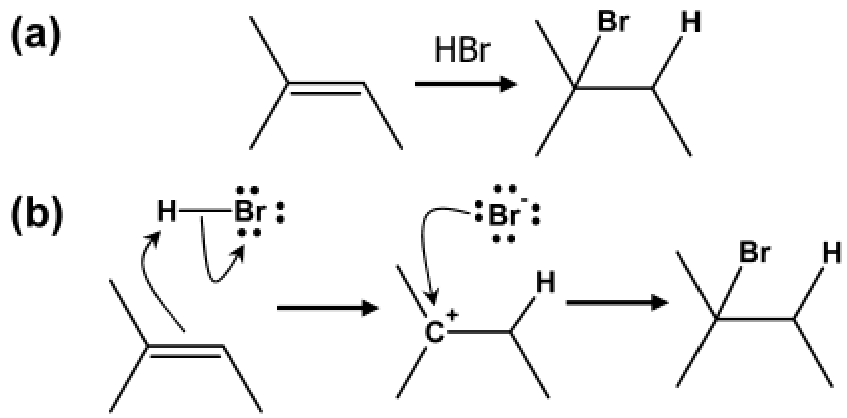
Table 3. Example of 10 Prioritized Transformation Rules, Relating to Alkene Hydrobromination Reactions, out of the 92 Rules Used in the Complete Robust HBr Reagent Model[a]

SMIRKS	description	electron flow	priority
[H:10][CH1:1][C+;!H0:2]≫ [C+:1][C+0:2][H:10]	carbocation, hydride shift from tertiary	1,10=10,2	10
[C:10][CH0:1][C+;!H0:2]≫ [C+:1][C+0:2][C:10]	carbocation, methyl shift from quaternary	1,10=10,2	9
[H:10][CH2:1][CH2+:2]≫ [C+:1][C+0:2][H:10]	carbocation, hydride shift from secondary	1,10=10,2	8
[C+:1].[-:2]≫ [C+0:1][+0:2]	carbocation, anion addition	2=1	7
[C:1]=[C;$(*O):2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, alkoxy	2,1=1,3;3,4=4	6
[C:1]=[C;$(*a):2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, benzyl	2,1=1,3;3,4=4	5
[C:1]=[C;$(**=*):2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, allyl	2,1=1,3;3,4=4	4
[C:1]=[CH0:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, tertiary	2,1=1,3;3,4=4	3
[C:1]=[CH1:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition, secondary	2,1=1,3;3,4=4	2
[C:1]=[C:2].[H:3][Cl,Br,I,$(OS=O):4]≫ [H:3][C:1][C+:2].[-:4]	alkene, protic acid addition	2,1=1,3;3,4=4	1

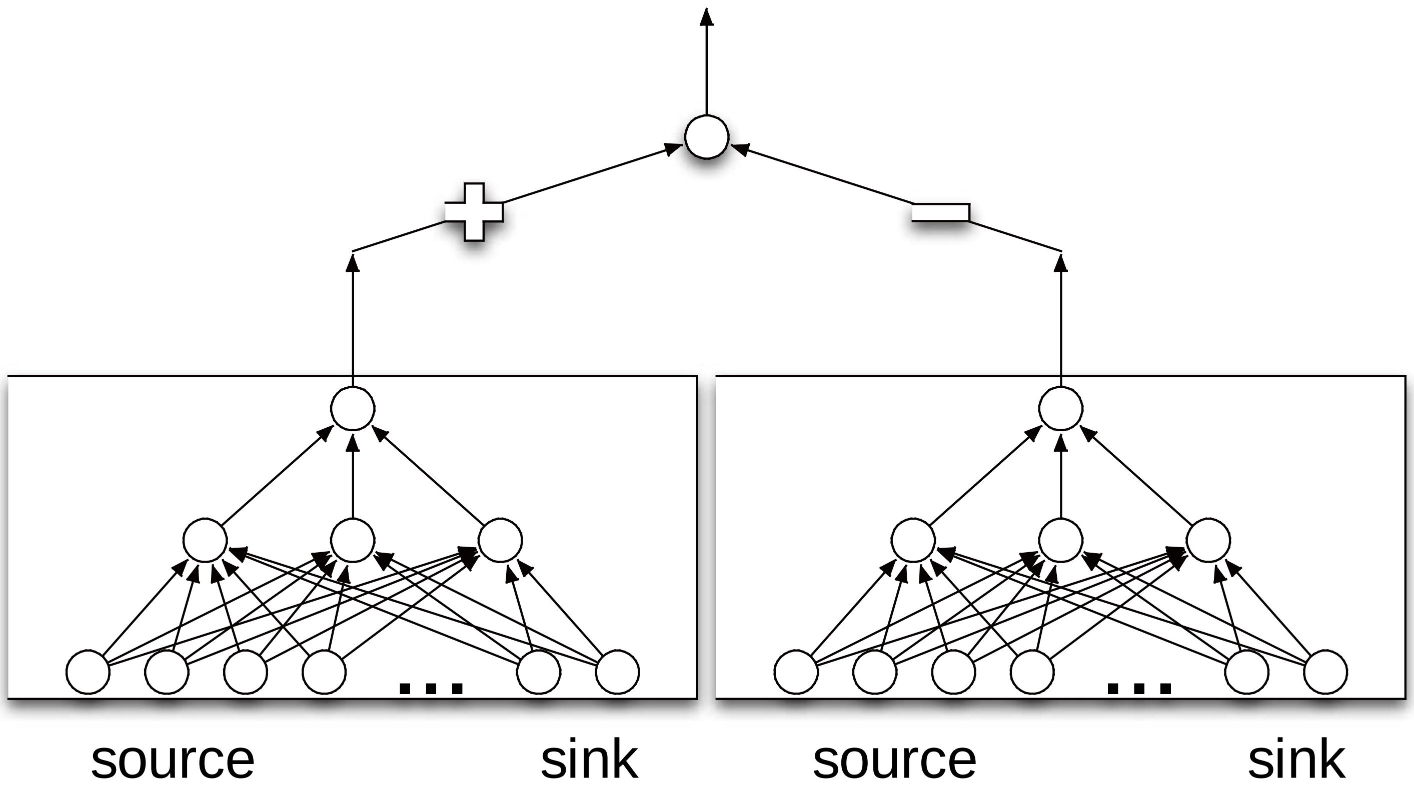
[a] Included for each transformation rule is not only the SMIRKS pattern and description but also a relative priority rank to indicate the order in which the rules should be attempted. The existence of several variants for similar rules and the customized priority ordering enables robust reaction predictions that address the issues noted in Figure 7. An electron flow specification accompanies each rule to support curved arrow mechanism diagrams.

Deep Learning Chemical Reactions

RCH=CH2 + HBr → RCH(Br)–CH3



Siamese Architecture



Deep Learning Chemical Reactions

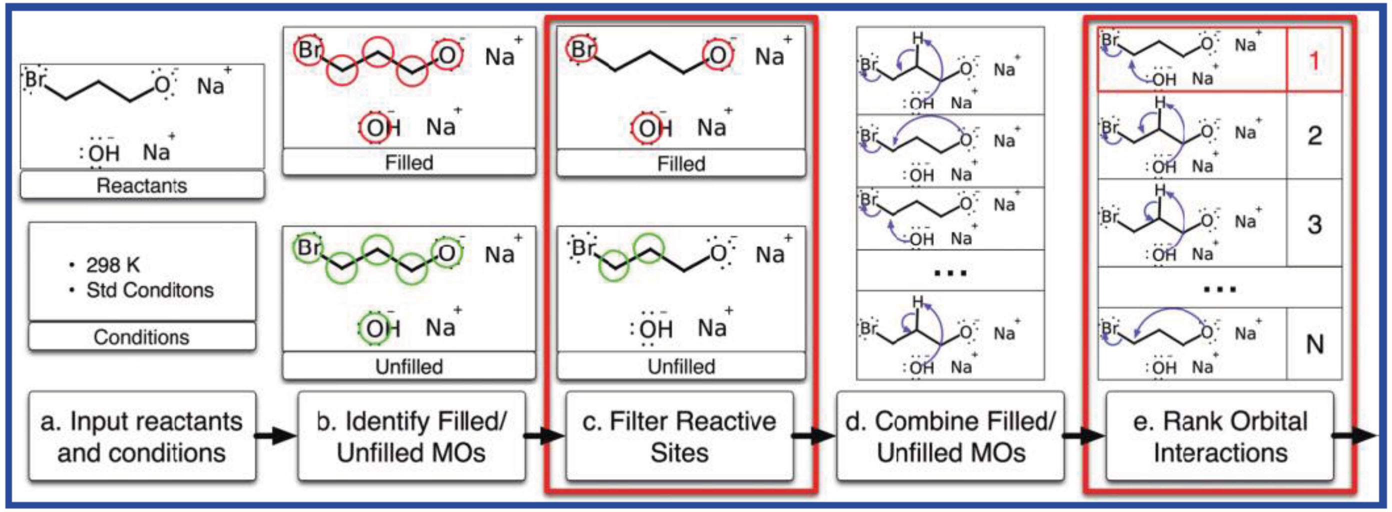
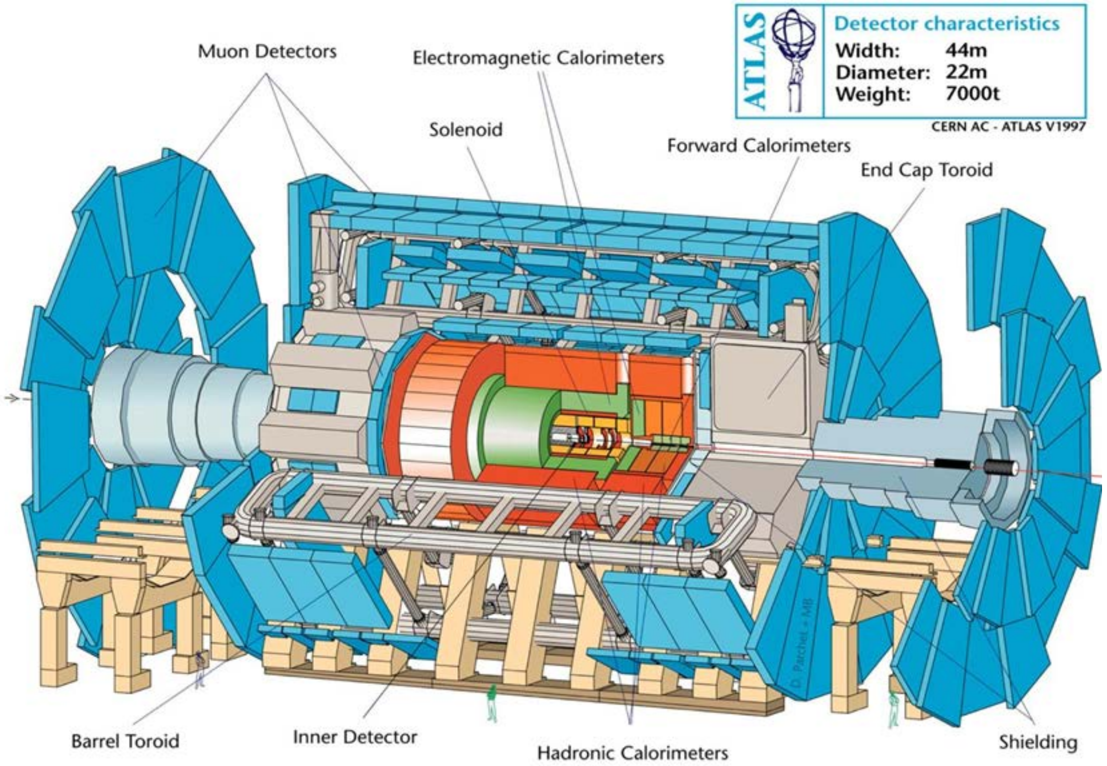


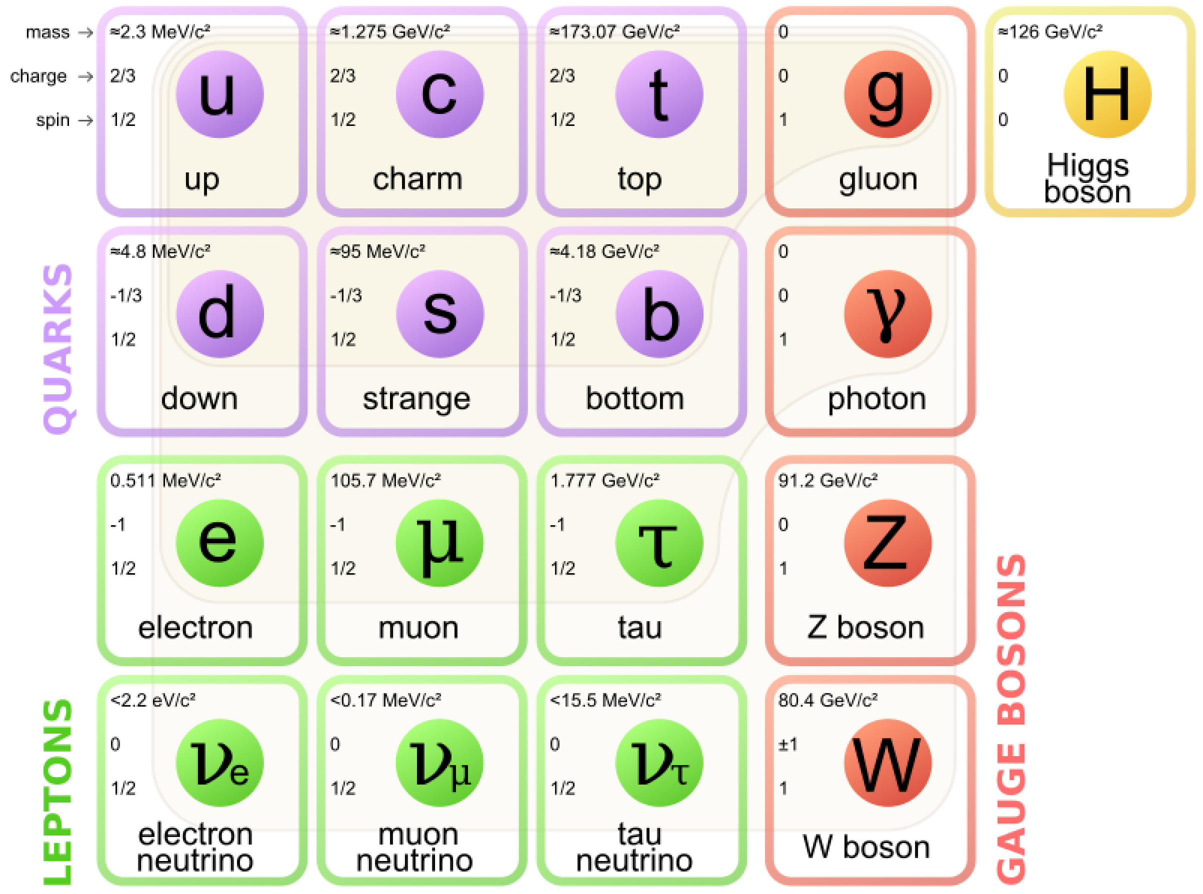
Figure 2. Overall reaction prediction framework: (a) A user inputs the reactants and conditions. (b) We identify potential electron donors and acceptors using coarse approximations of electron-filled and -unfilled MOs. (c) Highly sensitive reactive site classifiers are trained and used to filter out the vast majority of unreactive sites, pruning the space of potential reactions. (d) Reactions are enumerated by pairing filled and unfilled MOs. (e) A ranking model is trained and used to order the reactions, where the best ranking one or few represent the major products. The top-ranked product can be recursively chained to a new instance of the framework for multistep reaction prediction.

M. Kayala, C. Azencott, J. Chen, and P. Baldi. Learning to Predict Chemical Reactions. *Journal of Chemical Information and Modeling*, 51, 9, 2209–2222, (2011).

M. Kayala and P. Baldi. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *Journal of Chemical Information and Modeling*, 52, 10, 2526–2540, (2012).

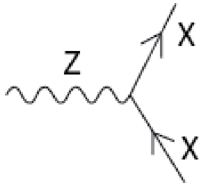
Deep Learning in HEP



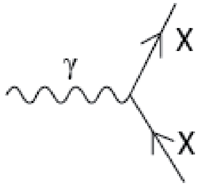


mass →
charge →
spin →
≈2.3 MeV/c²
2/3
1/2
u
up
≈1.275 GeV/c²
2/3
1/2
c
charm
≈173.07 GeV/c²
2/3
1/2
t
top
0
0
1
g
gluon
≈126 GeV/c²
0
0
H
Higgs boson
≈4.8 MeV/c²
-1/3
1/2
d
down
≈95 MeV/c²
-1/3
1/2
s
strange
≈4.18 GeV/c²
-1/3
1/2
b
bottom
0
0
1
γ
photon
0.511 MeV/c²
-1
1/2
e
electron
105.7 MeV/c²
-1
1/2
μ
muon
1.777 GeV/c²
-1
1/2
τ
tau
91.2 GeV/c²
0
1
Z
Z boson
<2.2 eV/c²
0
1/2
νe
electron neutrino
<0.17 MeV/c²
0
1/2
νμ
muon neutrino
<15.5 MeV/c²
0
1/2
ντ
tau neutrino
80.4 GeV/c²
±1
1
W
W boson
QUARKS
LEPTONS
GAUGE BOSONS

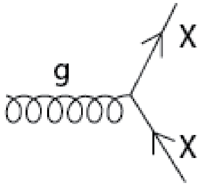
Standard Model Interactions
(Forces Mediated by Gauge Bosons)



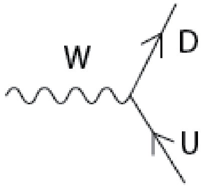
X is any fermion in
the Standard Model.



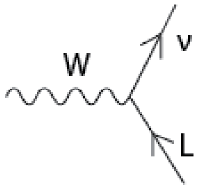
X is electrically charged.



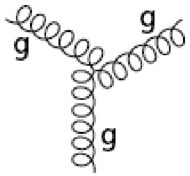
X is any quark.

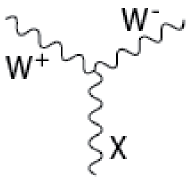


U is a up-type quark;
D is a down-type quark.

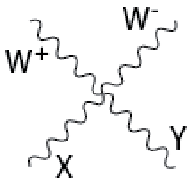


L is a lepton and ν is the
corresponding neutrino.

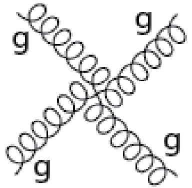




X is a photon or Z-boson.



X and Y are any two
electroweak bosons such
that charge is conserved.



Deep Learning in HEP

- Higgs Boson Detection (Nature Communications, 2014)
- Supersymmetry (Nature Communications, 2014)
- Higgs Decay (NIPS 2015, Phys. Rev. Let. 2015)
- Dark Knowledge, DarkMatter (JMLR C&P 2015)
- Jet substructure, jet tagging, parameterized classifiers

- Common Features and Results:

  -dozens of features: raw + human-derived

  -millions of examples

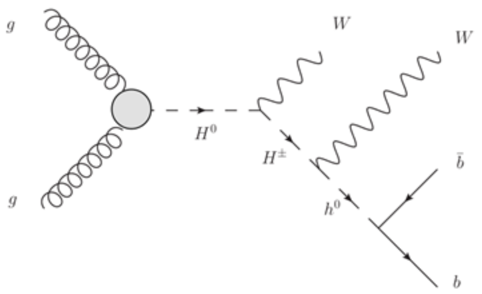
  -classification problems

  -deep learning outperforms current methods, with or without human-derived features

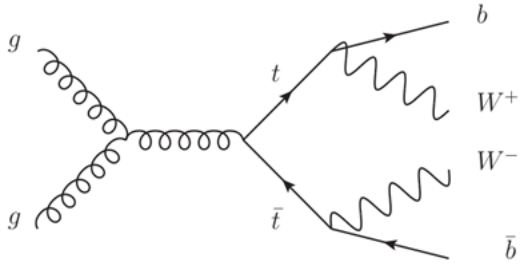
  - dark knowledge improves shallow architectures

Higgs Boson Detection

Higgs boson decay signal



Background process



Simulation tools:

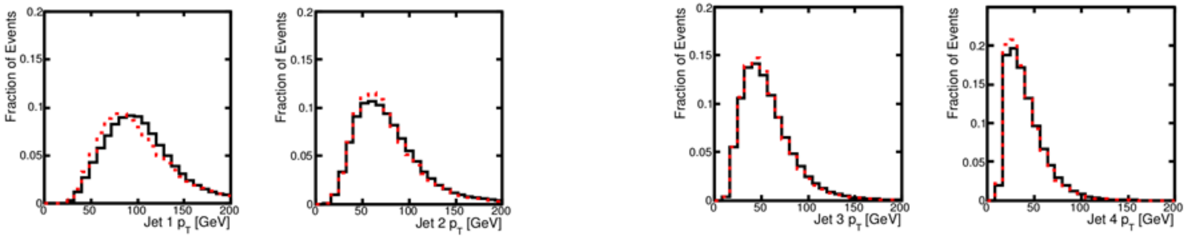
- MadGraph (collisions)
- PYTHIA (showering and hadronization)
- DELPHES (detector response)

11 M examples

Higgs Boson Detection

Supervised learning problem:

- Two classes
- 11 million training examples (roughly balanced)
- 28 features
  - o 21 low-level features (momenta of particles)
  - o 7 high-level features derived by physicists



Signal (black) vs background (red)

Data available at archive.ics.uci.edu/ml/datasets/HIGGS

Higgs Boson Detection

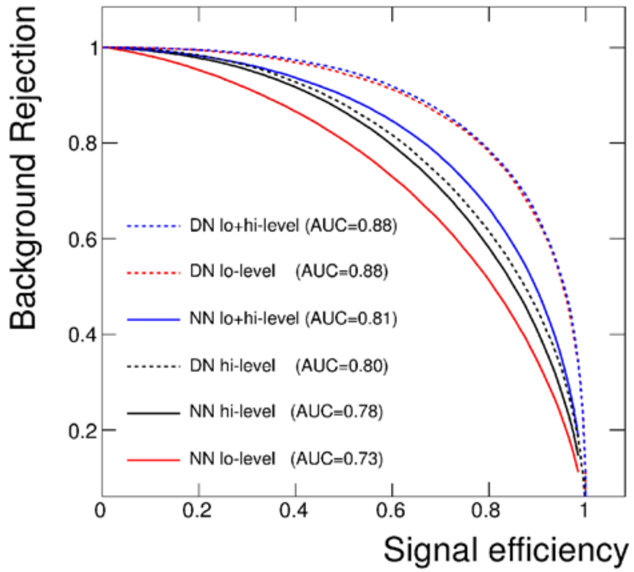
Tuning deep neural network architectures.

Hyper parameters	Choices
Depth	2,3,4,5,6 layers
Hidden units per layer	100,200,300,500
Learning rate	0.01, 0.05
Weight decay	0, 0.00001
Pre-training	none, autoencoder
	multi-task autoencoder
Input features	low-level, high-level
	complete set

Best:

- 5 hidden layers
- 300 neurons per layer
- Tanh hidden units, sigmoid output
- No pre-training
- Stochastic gradient descent
- Mini batches of 100
- Exponentially-decreasing learning rate
- Momentum increasing from .5 to .99 over 200 epochs
- Weight decay = 0.00001

Higgs Boson Detection



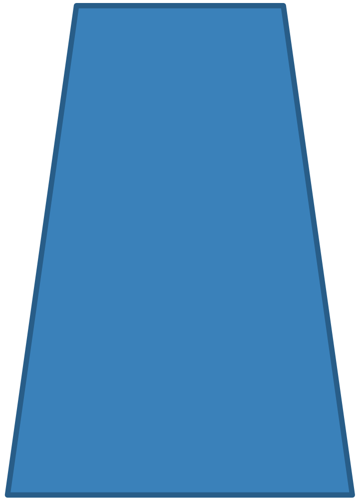
	AUC		
Technique	Low-level	High-level	Complete
BDT	0.73	0.78	0.81
NN	0.733 (0.007)	0.777 (0.001)	0.816 (0.004)
DN	0.880 (0.001)	0.800 ($<$ 0.001)	0.885 (0.002)

Deep network improves AUC by 8%

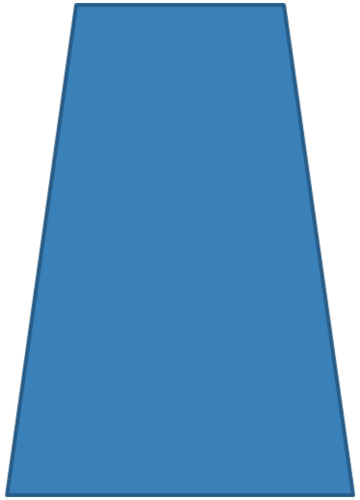
BDT= Boosted Decision Trees in TMVA package

Nature Communications, July 2014

Dark Knowledge



1. Train deep architecture using **binary targets** (multi-class classification)



2. For each training example retrieve the **soft targets** from the output of the trained deep architecture.

3. Use the soft targets (which contain the dark knowledge), to train a shallow architecture.

Dark Knowledge

Table 7: Performance of shallow networks trained with dark knowledge.

	AUC		
Architecture	Benchmark 1	Benchmark 2	Benchmark 3
NN	0.842	0.8786	0.797
NN w/ dark knowledge	0.850	0.8788	0.799
DNN	0.885	0.8790	0.802

Jet Substructure and Jet Tagging

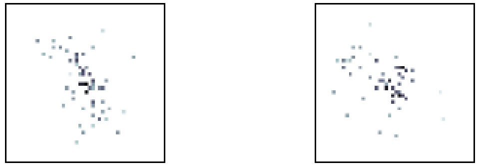


Figure 1: Typical jet images from class 1 (single jet) on the left, and class 2 (two overlapping jets) on the right, after preprocessing.

Technique	Performance Signal efficiency	AUC
BDT on derived features	90.4%	96.4%
Shallow NN (32x32 input)	86.7% (0.01%)	95.2% (0.01%)
Compressed deep NN (32x32 input)	90.2%	96.3%
Deep NN (32x32 input)	92.7% (0.03%)	97.0% (0.01%)
Deep NN (48x48 input)	93.0%	97.1%

Table 3: Signal efficiency at 90% background rejection and AUC for each method. The best shallow neural network and the best deep neural network trained on 32×32 pixel images were trained with three different random initializations, and we report the mean and standard deviation for each metric.

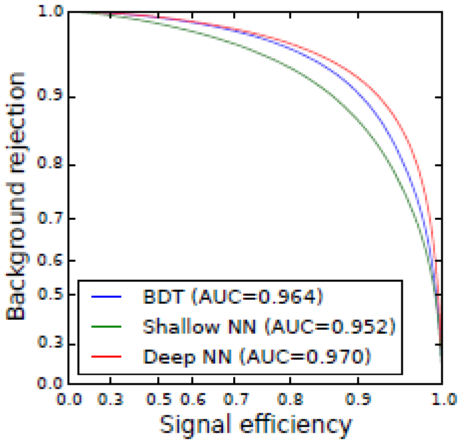
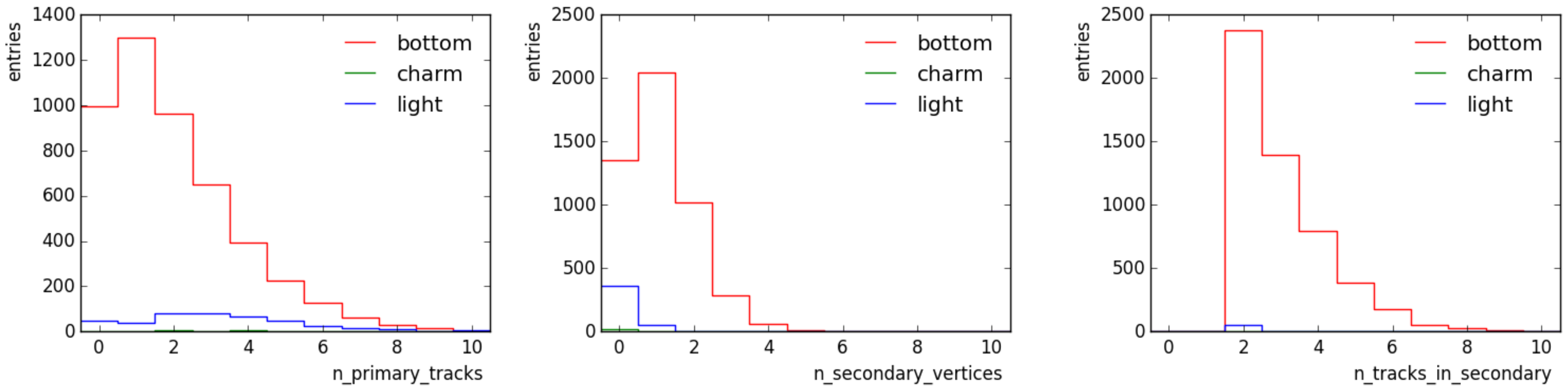


Figure 3: ROC curves comparing the performance of a deep convolutional network trained on the images, a shallow network trained on the images, and boosted decision trees trained on the designed features.

Jet Substructure

- **Use medium-level variables**
  - Primary vertex (5 variables)
  - Primary tracks (8 variables)
  - Secondary vertices (5 variables)
  - Secondary tracks (8 variables)



- **Use RNNS, CNNs, etc**

The Black-Box Problem

- It is a problem. But:

The Black-Box Problem

- It is a problem.

1. It is sometimes overblown:

The Black-Box Problem

- It is a problem.

1. It is sometimes overblown:
   - You do not fully understand your car.

The Black-Box Problem

- It is a problem.

1. It is sometimes overblown:
   - You do not fully understand your car.
   - You do not fully understand your brain.

The Black-Box Problem

- It is a problem.

1. It is sometimes overblown:
   - You do not fully understand your car.
   - You do not fully understand your brain.
   - The LHC is a collection of black boxes.

The Black-Box Problem

- It is a problem.

1. It is sometimes overblown:
   - You do not fully understand your car.
   - You do not fully understand your brain.
   - The LHC is a collection of black boxes. However the modularity is very important.

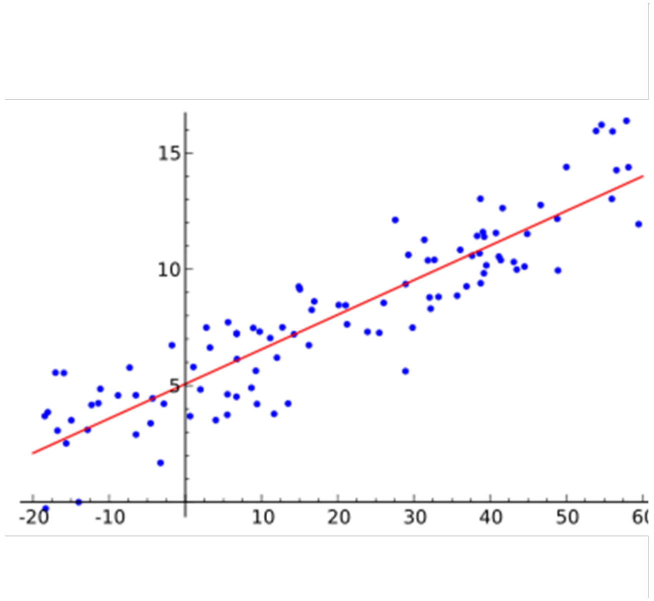
The Black-Box Problem

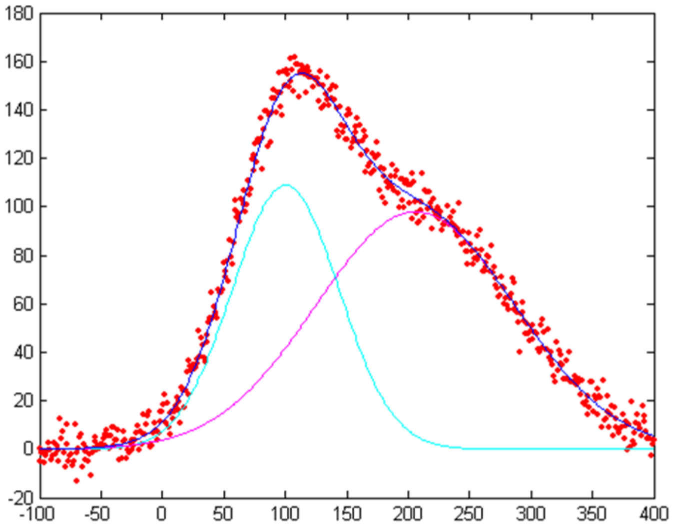
2. It is possible to open up the box and test the output of each neuron under multiple inputs—but very time consuming. Perhaps NNs to understand NNs will be developed.

3. However some degree of opacity is to be expected and inevitable. NNs are a fundamentally different model of computation, where processing and memory are completely interwined, rather than being separated as in current digital computers. In a neural network, data is not stored at a computer address, but rather shattered in each synaptic weight. This is already the case in linear regression.

The Black-Box Problem

4. What does it mean to be a Black Box? What does it mean to understand? To understand is to compress.





The Black-Box Problem

5. The importance of modularity:

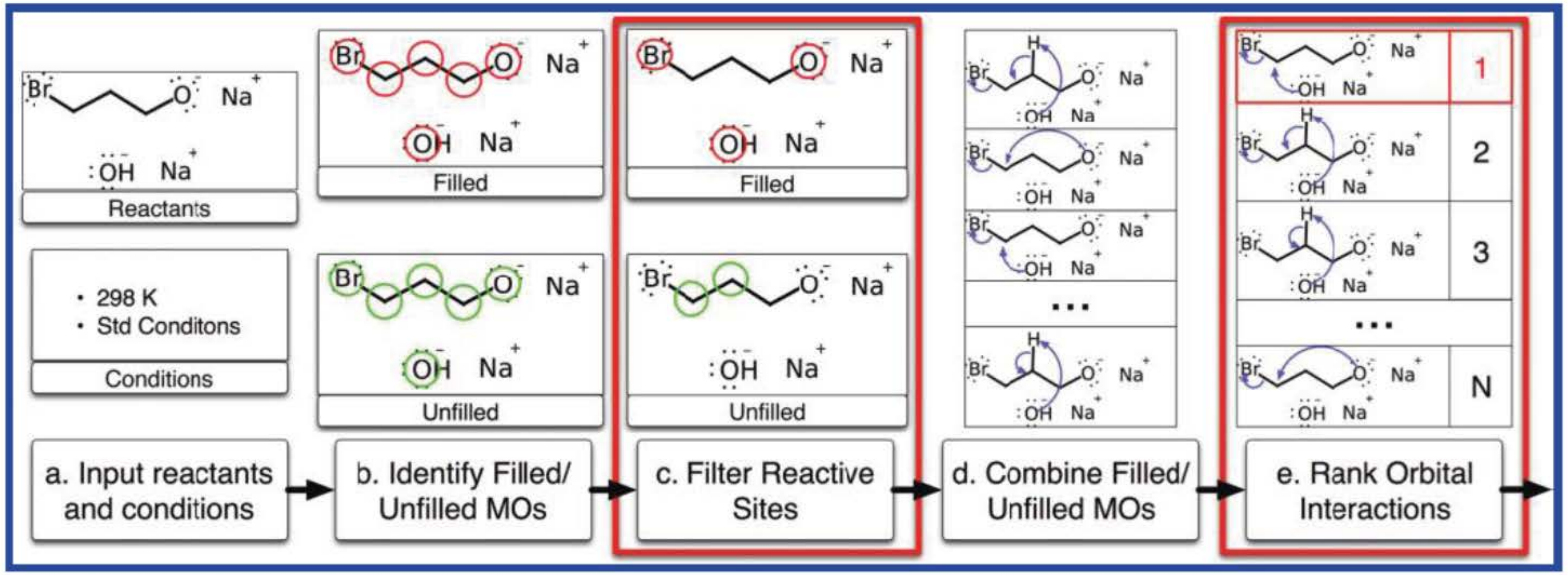
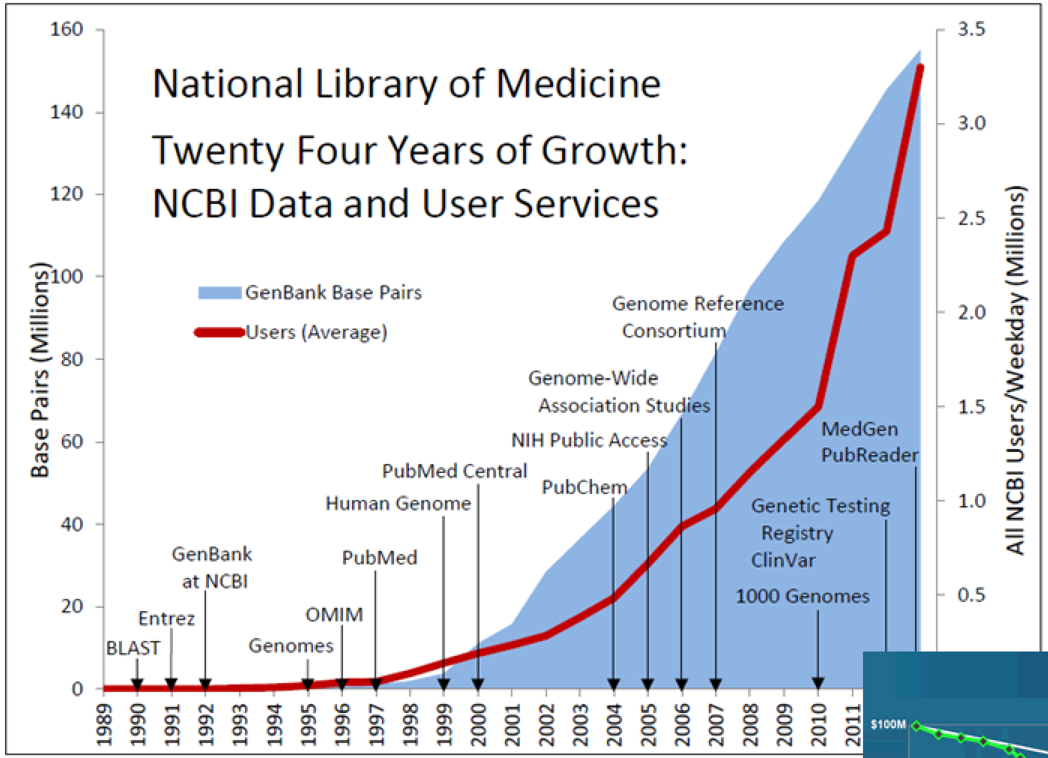


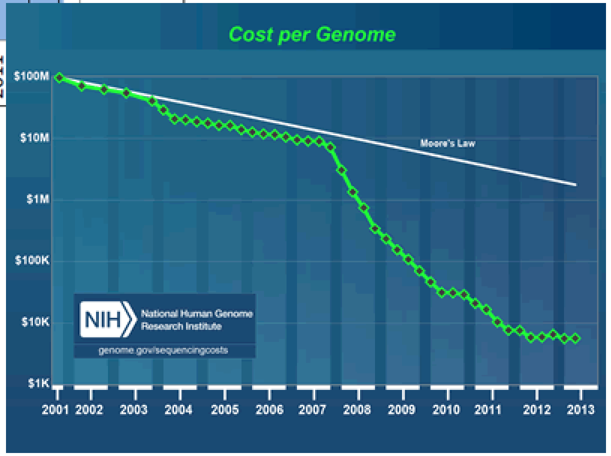
Figure 2. Overall reaction prediction framework: (a) A user inputs the reactants and conditions. (b) We identify potential electron donors and acceptors using coarse approximations of electron-filled and -unfilled MOs. (c) Highly sensitive reactive site classifiers are trained and used to filter out the vast majority of unreactive sites, pruning the space of potential reactions. (d) Reactions are enumerated by pairing filled and unfilled MOs. (e) A ranking model is trained and used to order the reactions, where the best ranking one or few represent the major products. The top-ranked product can be recursively chained to a new instance of the framework for multistep reaction prediction.

Conclusions

- DL is important but hardly a new idea.
- Examples of applications of deep learning in the natural sciences (Biology, Chemistry, and Physics).
- Natural sciences offer many other challenges and opportunities (QM, Earth Sciences, Astronomy, etc).
- Not only important scientific and technological applications, but also significant challenges, opportunities, and inspiration for machine learning and AI.
- DL yields state of the art performance in HEP and much more to come.

THANK YOU


National Library of Medicine
Twenty Four Years of Growth:
NCBI Data and User Services
GenBank Base Pairs
Users (Average)
Base Pairs (Millions)
All NCBI Users/Weekday (Millions)
160
140
120
100
80
60
40
20
0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
BLAST
Entrez
GenBank
at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide
Association Studies
Genome Reference
Consortium
MedGen
PubReader
Genetic Testing
Registry
ClinVar
1000 Genomes
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013