How do artificial neural networks learn to predict protein structures?

Badri Adhikari
adhikarib @ umsl.edu

Assistant Professor of CS
Department of Mathematics & Computer Science
University of Missouri-St. Louis

PRINCIPAL INVESTIGATOR

Badri Adhikari, PhD
Assistant Professor of Computer Science
Department of Mathematics and Computer Science
University of Missouri-St. Louis

312 Express Scripts Hall
St. Louis, MO 63121

Phone: 314-516-7393
Email: adhikarib@umsl.edu
Homepage: https://badriadhikari.github.io/

TEACHING

- Artificial Intelligence - 2018 Fall, 2019 Spring, 2019 Fall
- Deep Learning - 2019 Spring
- Programming and Data Structures - 2018 Spring
- Hands-on Deep Learning Workshops - 2018 Fall
- Advanced Data Structures and Algorithms - 2017 Fall

Hands-on Workshop on Deep Learning
Wednesday, November 14, 2018 - 3:30 PM to 6 PM *
Room 304, Express Scripts Hall
UMSL

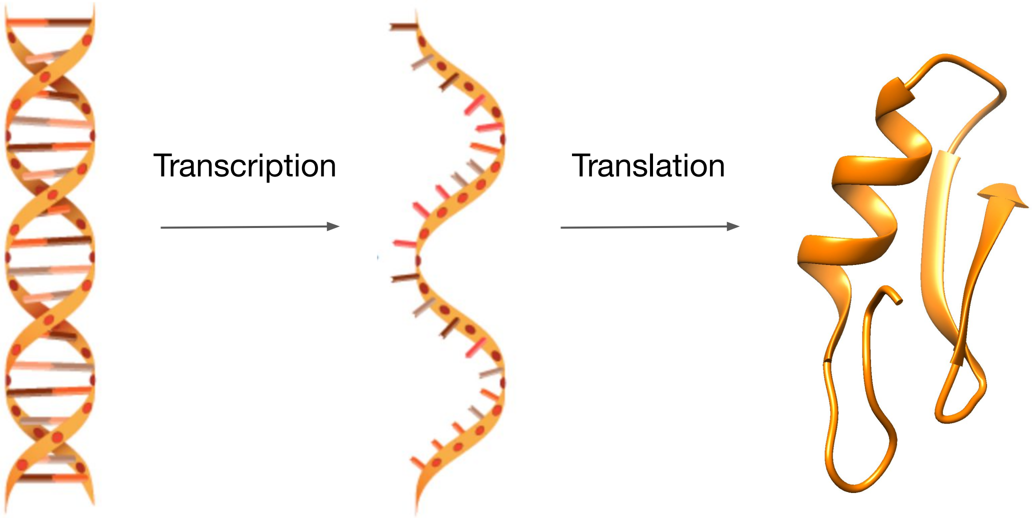
RESEARCH

- Investigate deep learning methods for protein structure prediction
- Collaboration with Robert Paul at MIMH (various mental health datasets)
- Collaboration with Lauren Salmimen at USC (UK biobank datasets, and various other datasets)
- Collaboration with researchers at International Maize and Wheat Development Center (CIMMYT)

How do artificial neural networks learn to predict protein structures?

The central dogma in molecular biology

- The fundamental process in life is the flow of information from DNA to proteins
  - How does this happen?



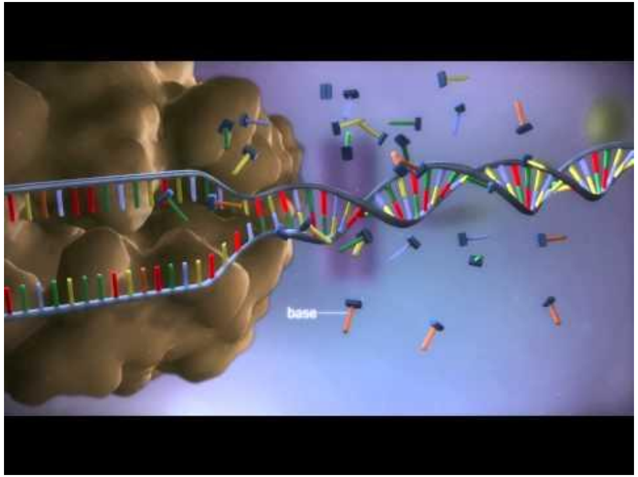
Transcription

Translation

DNA

RNA

Protein

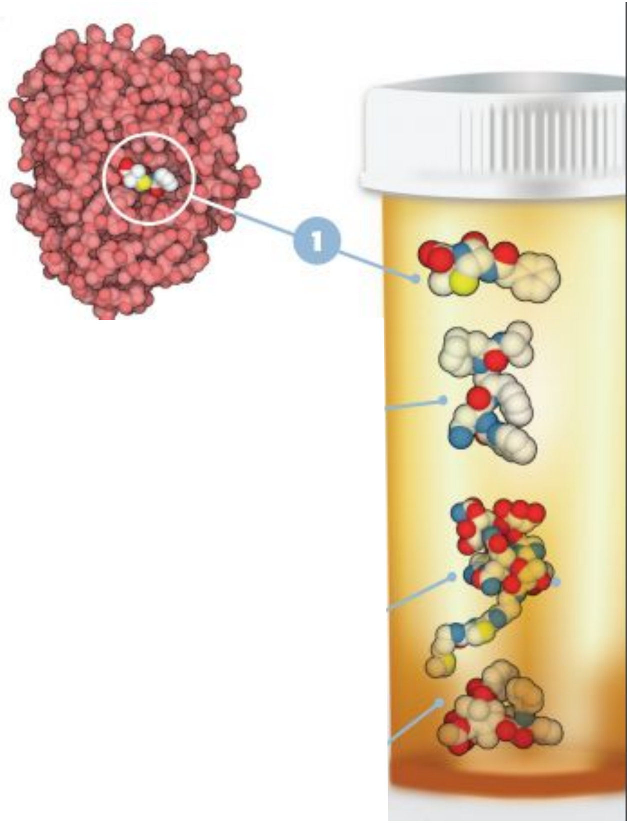


Protein folding is important

- Proteins are fundamental to understanding their role within the body
- Many diseases are believed to be caused by misfolded proteins
    - Alzheimer's, Parkinson's, Huntington's, cystic fibrosis, etc.
- One of the top 100 questions selected by the Science magazine

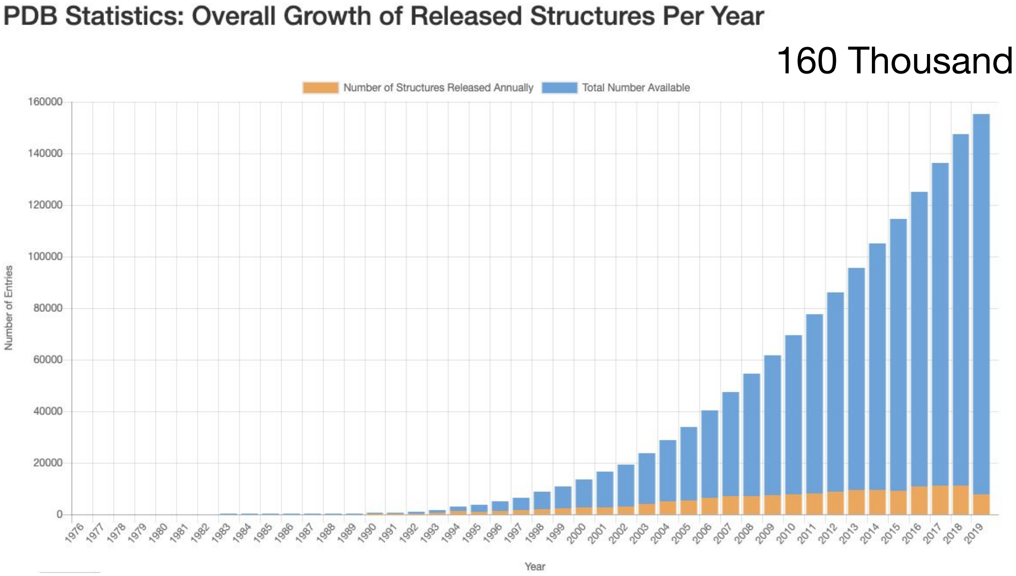
The need to obtain precise 3D structures

- Antibiotics need to kill pathogenic organisms like bacteria without poisoning the patient
- Often, these drugs attack proteins that are only found in the targeted bacterium which are crucial for their survival or multiplication
- For instance, penicillin attacks the enzyme that builds bacterial cell walls

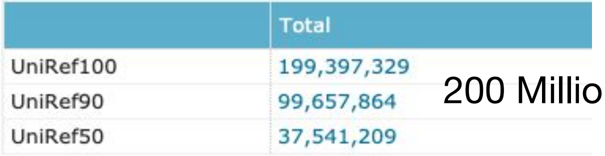


https://cdn.rcsb.org/pdb101/learn/resources/how-do-drugs-work-flyer.pdf

Only around 160K protein structures are solved so far

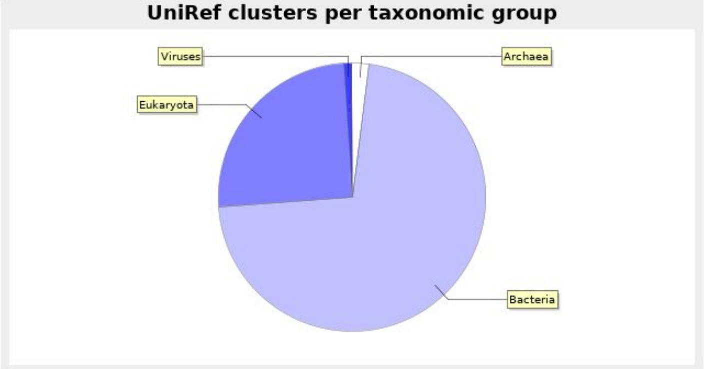


160 Thousand



	Total
UniRef100	199,397,329
UniRef90	99,657,864
UniRef50	37,541,209

200 Million



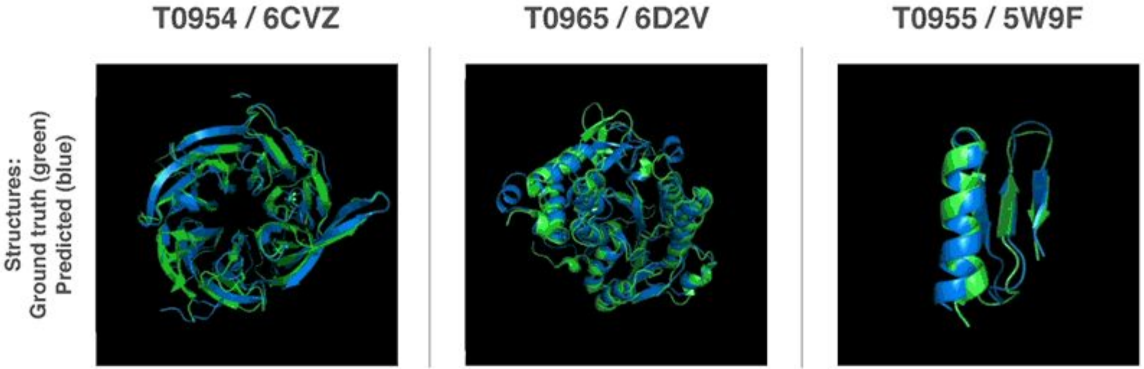
Can we predict protein structures accurately today?

http://predictioncenter.org/
Critical Assessment of Protein Structure Prediction

World-wide competition
held every two years
(3 months long)

VS

Models predicted by DeepMind in CASP 13 (2018)



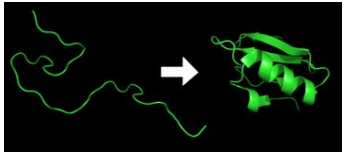
https://deepmind.com/blog/article/alphafold

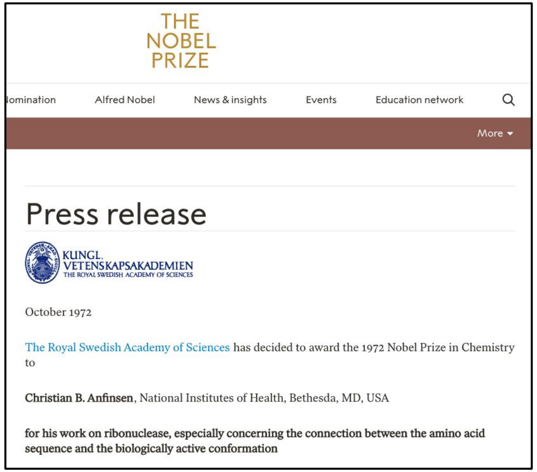
What does AI do exactly?

This is how computer scientists predict protein structures

Anfinsen's Dogma (1973)

Native structure is determined **only** by the protein's amino acid sequence





Press release

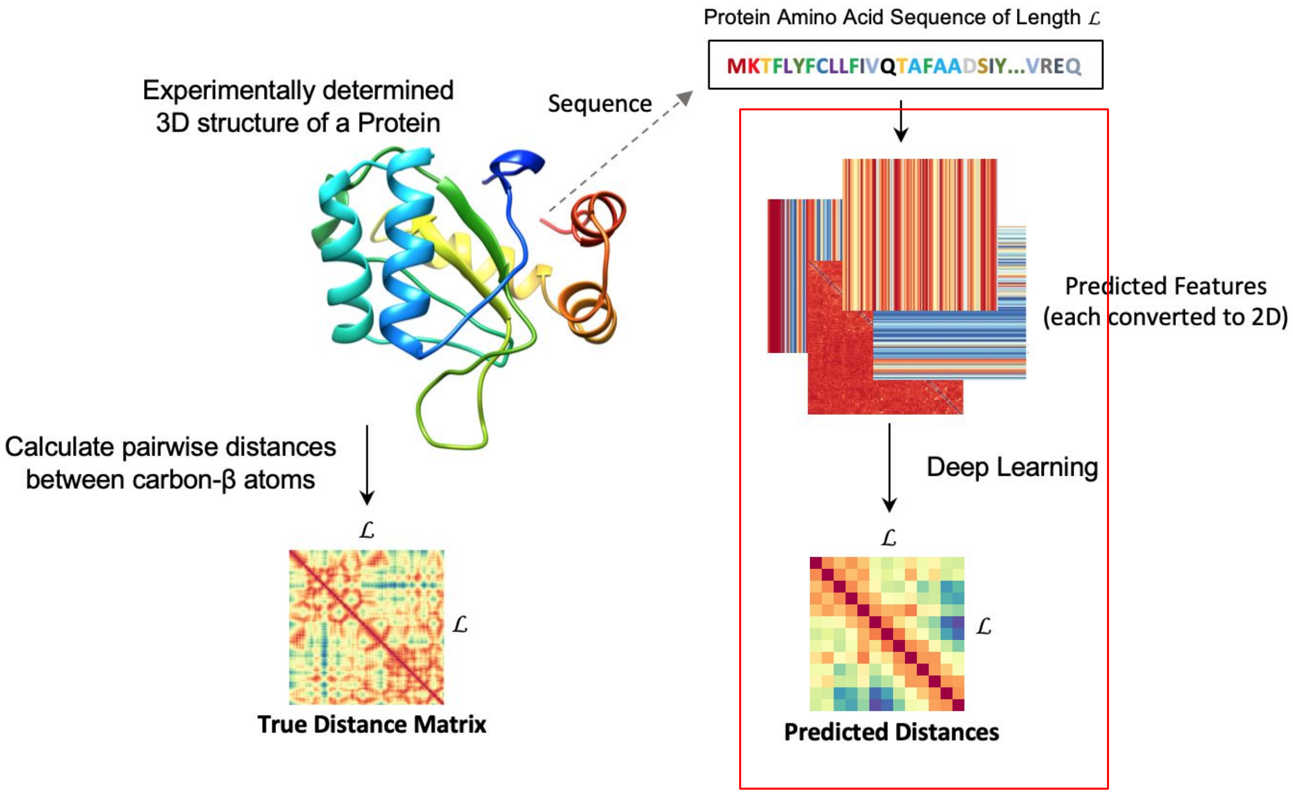
KUNGL. VETENSKAPSAKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

October 1972

The Royal Swedish Academy of Sciences has decided to award the 1972 Nobel Prize in Chemistry to

Christian B. Anfinsen, National Institutes of Health, Bethesda, MD, USA

for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation



Experimentally determined 3D structure of a Protein

Sequence

Protein Amino Acid Sequence of Length $\mathcal{L}$

MKTFLYFCLLFIVQTAFAADSIY...VREQ

Calculate pairwise distances between carbon-β atoms

Predicted Features (each converted to 2D)

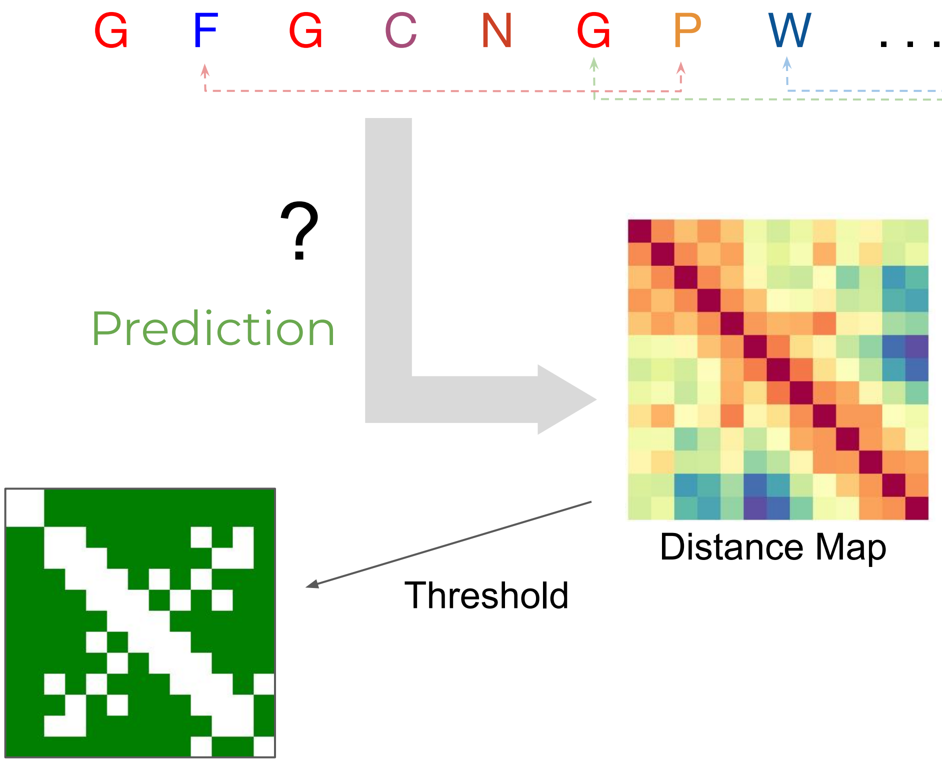
Deep Learning

True Distance Matrix

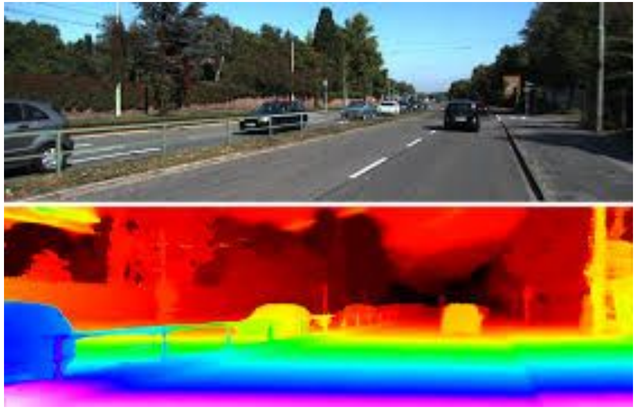
Predicted Distances

The distance prediction problem is an AI problem

A Protein's Amino Acid Sequence

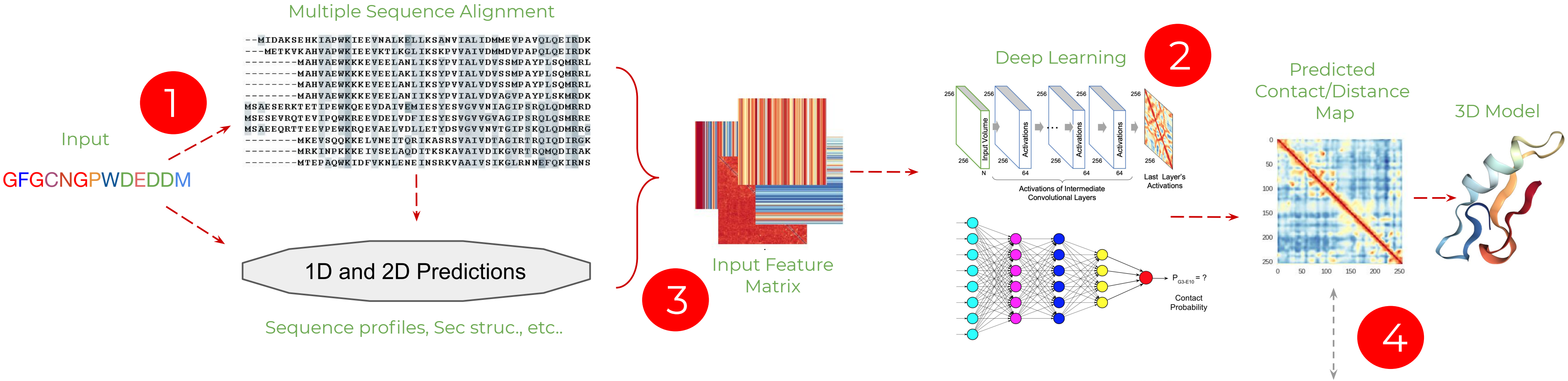


Pair	D
G_1 - F_2	3.5
G_1 - G_3	6.2
G_1 - C_4	8.1
...	
G_2 - E_9	8.0
...	

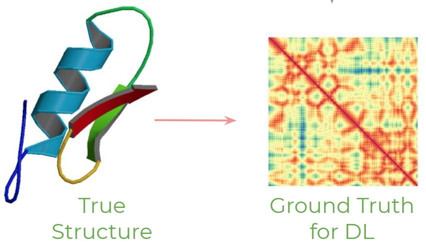


Depth Estimation / Image Segmentation

Attacking the protein distance prediction problem



```
--MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDK
---METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDK
---------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRL
---------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRL
---------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRL
---------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDK
MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRD
MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRE
MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRG
---------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGK
---------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAK
---------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNS
```

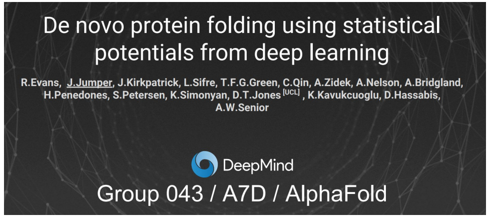


- 1 Generating Large and **High-quality MSAs**

DL
- 2 The Right Deep Learning **Architecture**
- 3 **Feature Engineering** (encoding, 1D to 2D, covariance/precision matrix)
- 4 Appropriate **Loss Function** (besides binary_crossentropy, MSE, etc.)

All top methods predict distances using residual neural networks

Top Methods in the most recent CASP Competition

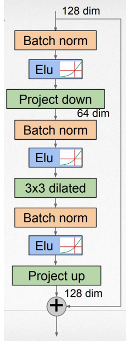


De novo protein folding using statistical potentials from deep learning

R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, D.T.Jones [UCL] , K.Kavukcuoglu, D.Hassabis, A.W.Senior

DeepMind

Group 043 / A7D / AlphaFold



DeepMetaPSICOV (DMP) in CASP13

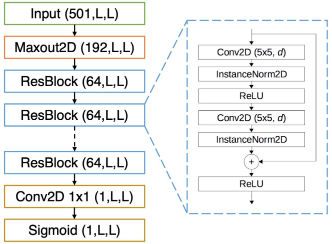
Shaun M Kandathil
University College London
&
The Francis Crick Institute

DeepMetaPSICOV model architecture

Deep, fully convolutional residual network

Total of 18 residual blocks plus one Maxout layer

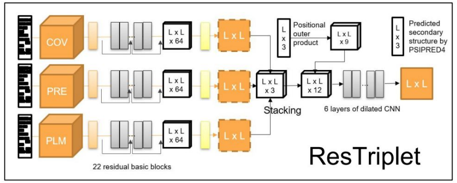
Dilated convolutions



ResTriplet/TripletRes:
Learning contact-maps from a triplet of coevolutionary matrices

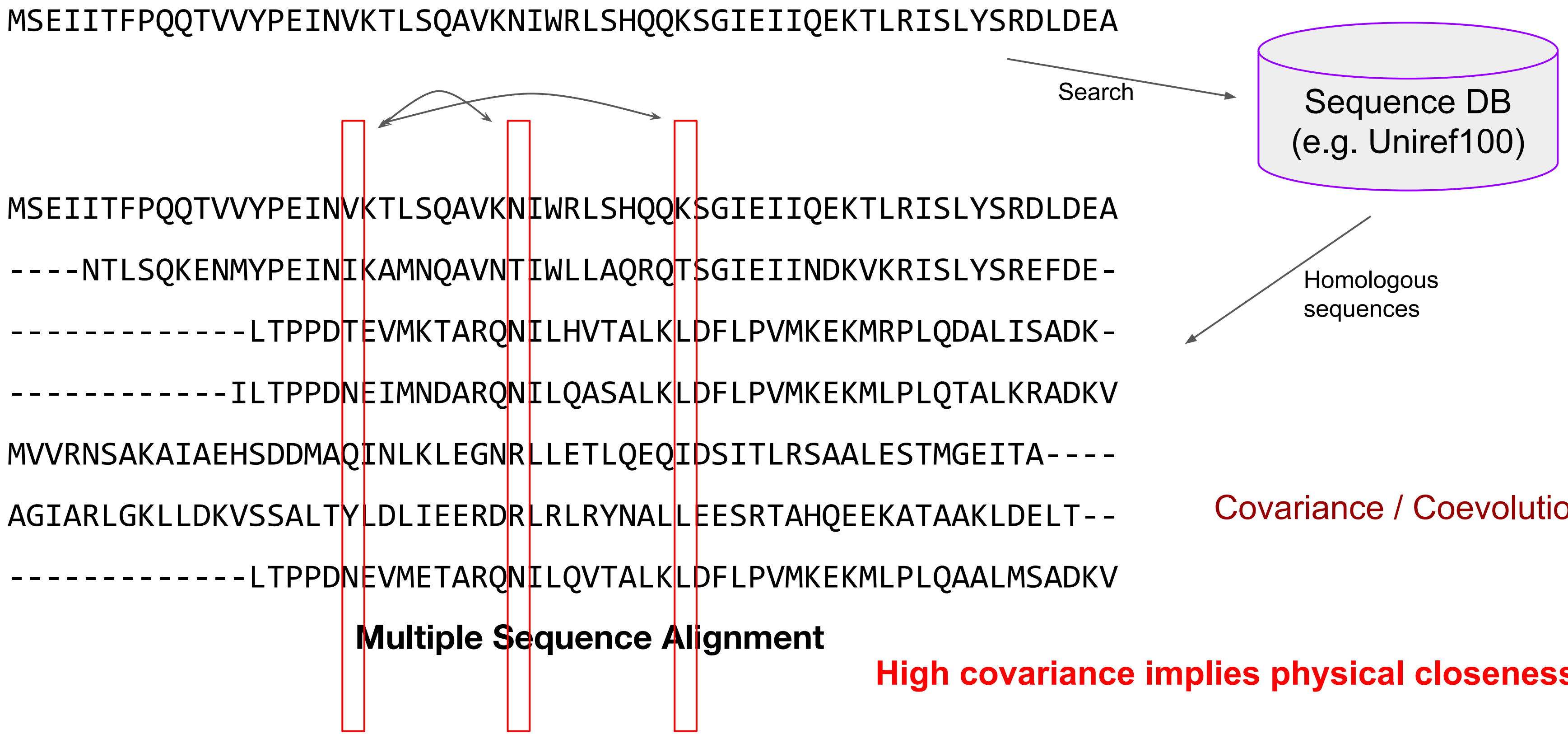
Eric W. Bell, Yang Li, Chengxin Zhang,
Dong-Jun Yu, Yang Zhang

Department of Computational Medicine and Bioinformatics,
University of Michigan - Ann Arbor



All these results show that residual networks are best architectures (for this problem)

Can We Learn to Predict Contacts WITHOUT ‘True’ Contacts?



```
MSEIITFPQQTVVYPEINVKTLSQAVKNIWRLSHQQKSGIEIIQEKTLRISLYSRDLDEA
```

Search

Sequence DB
(e.g. Uniref100)

Homologous sequences

```
MSEIITFPQQTVVYPEINVKTLSQAVKNIWRLSHQQKSGIEIIQEKTLRISLYSRDLDEA
----NTLSQKENMYPEINIKAMNQAVNTIWLLAQRQTSGIEIINDKVKRISLYSREFDE-
-------------LTPPDTEVMKTARQNILHVTALKLDFLPVMKEKMRPLQDALISADK-
------------ILTPPDNEIMNDARQNILQASALKLDFLPVMKEKMLPLQTALKRADKV
MVVRNSAKAIAEHSDDMAQINLKLEGNRLLETLQEQIDSITLRSAALESTMGEITA----
AGIARLGKLLDKVSSALTYLDLIEERDRLRLRYNALLEESRTAHQEEKATAAKLDELT--
-------------LTPPDNEVMETARQNILQVTALKLDFLPVMKEKMLPLQAALMSADKV
```

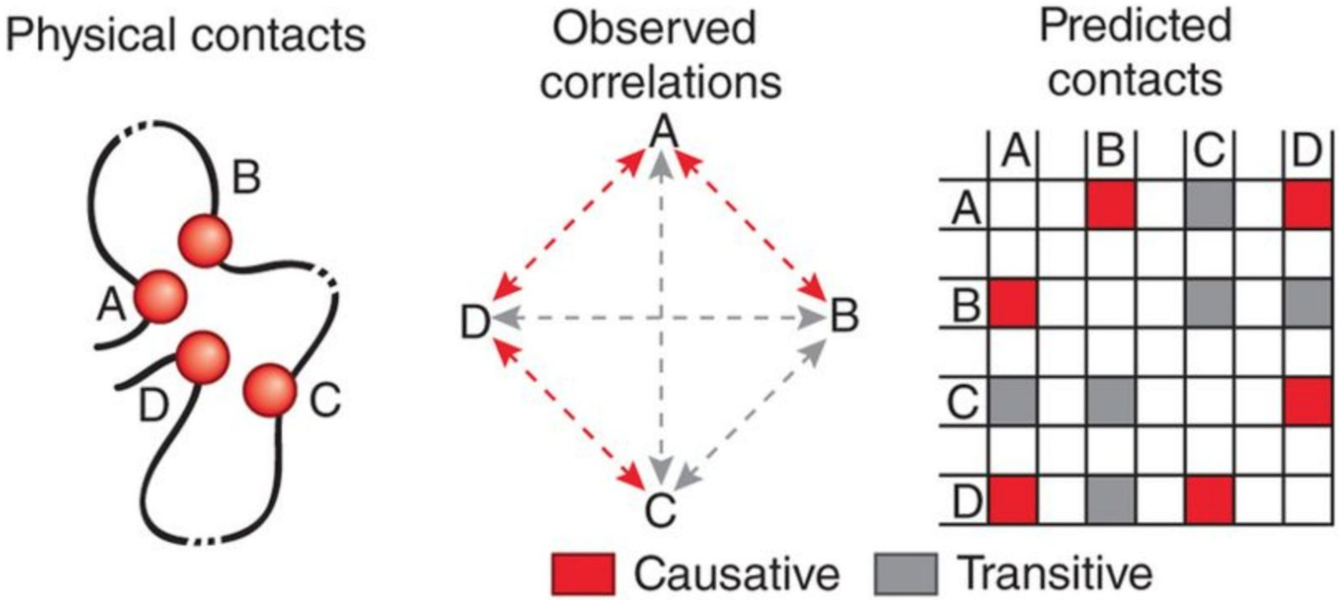
Covariance / Coevolution

Multiple Sequence Alignment

High covariance implies physical closeness!

Does this mean we can write an algorithm to predict contacts?

Can We Learn to Predict Contacts WITHOUT ‘True’ Contacts?



	A	B	C	D
A		🟥	▨	🟥
B	🟥		▨	▨
C	▨	▨		🟥
D	🟥	▨	🟥	

🟥 Causative ▨ Transitive

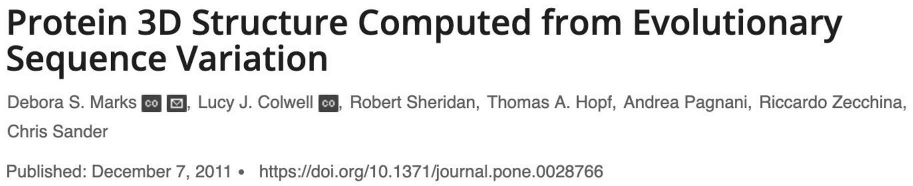
Perspective | Published: 08 November 2012

Protein structure prediction from sequence variation

Debora S Marks, Thomas A Hopf & Chris Sander

Nature Biotechnology **30**, 1072–1080 (2012) | Download Citation

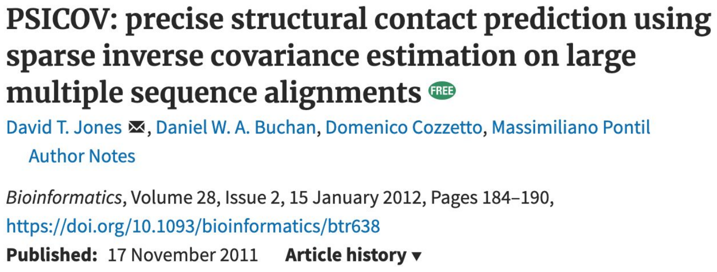
Can We Write Algorithms to Remove Transitive Noise?



Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, Chris Sander

Published: December 7, 2011 • https://doi.org/10.1371/journal.pone.0028766



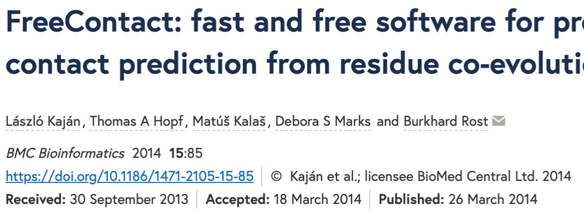
PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments FREE

David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, Massimiliano Pontil

Author Notes

Bioinformatics, Volume 28, Issue 2, 15 January 2012, Pages 184–190, https://doi.org/10.1093/bioinformatics/btr638

Published: 17 November 2011 **Article history** ▾



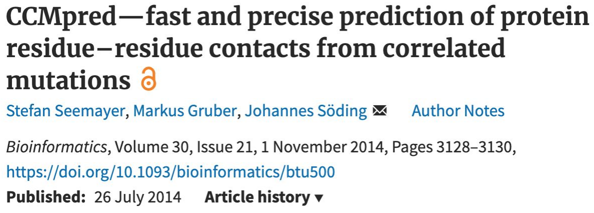
FreeContact: fast and free software for protein contact prediction from residue co-evolution

László Kaján, Thomas A Hopf, Matúš Kalaš, Debora S Marks and Burkhard Rost

BMC Bioinformatics 2014 **15**:85

https://doi.org/10.1186/1471-2105-15-85 | © Kaján et al.; licensee BioMed Central Ltd. 2014

Received: 30 September 2013 | **Accepted:** 18 March 2014 | **Published:** 26 March 2014



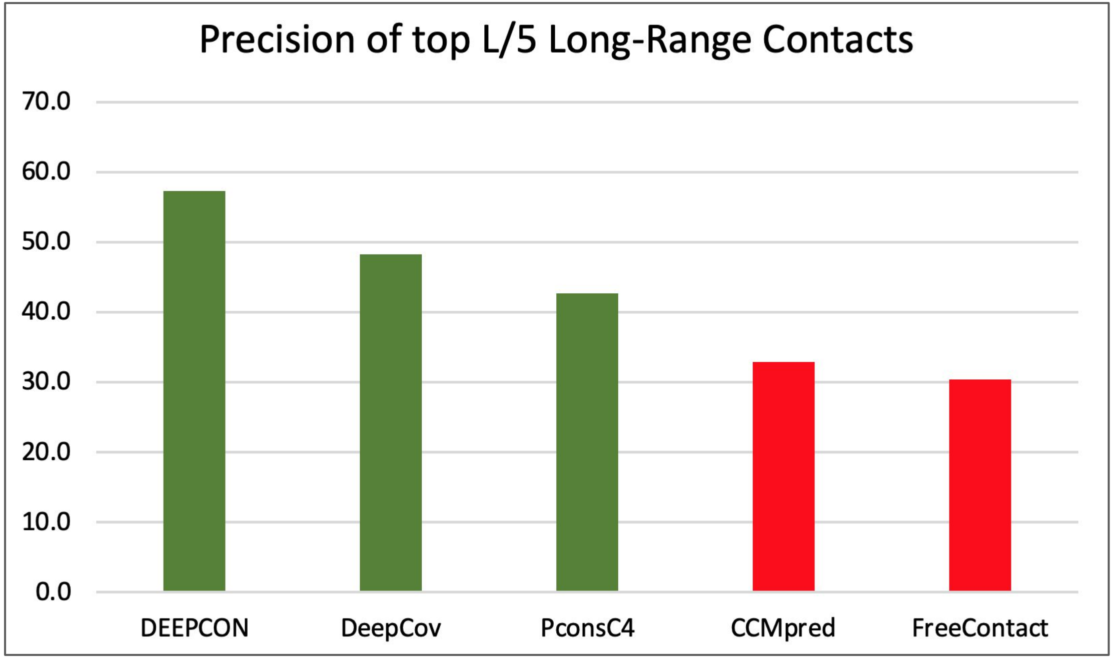
CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations

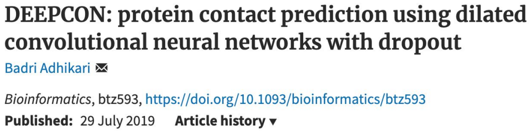
Stefan Seemayer, Markus Gruber, Johannes Söding Author Notes

Bioinformatics, Volume 30, Issue 21, 1 November 2014, Pages 3128–3130, https://doi.org/10.1093/bioinformatics/btu500

Published: 26 July 2014 **Article history** ▾

Can Deep Learning Remove Transitive Noise?




DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout

Badri Adhikari ✉

Bioinformatics, btz593, https://doi.org/10.1093/bioinformatics/btz593

Published: 29 July 2019 **Article history** ▾

What does AI do exactly?

It learns from input/output pairs and can outperform an algorithm based on theories!

How does AI work?

Neural plasticity

In 1953, Professor Theodor Erismann devised an experiment

- performing it upon his assistant and student, Ivo Kohler

He made Kohler wear a pair of hand-engineered goggles

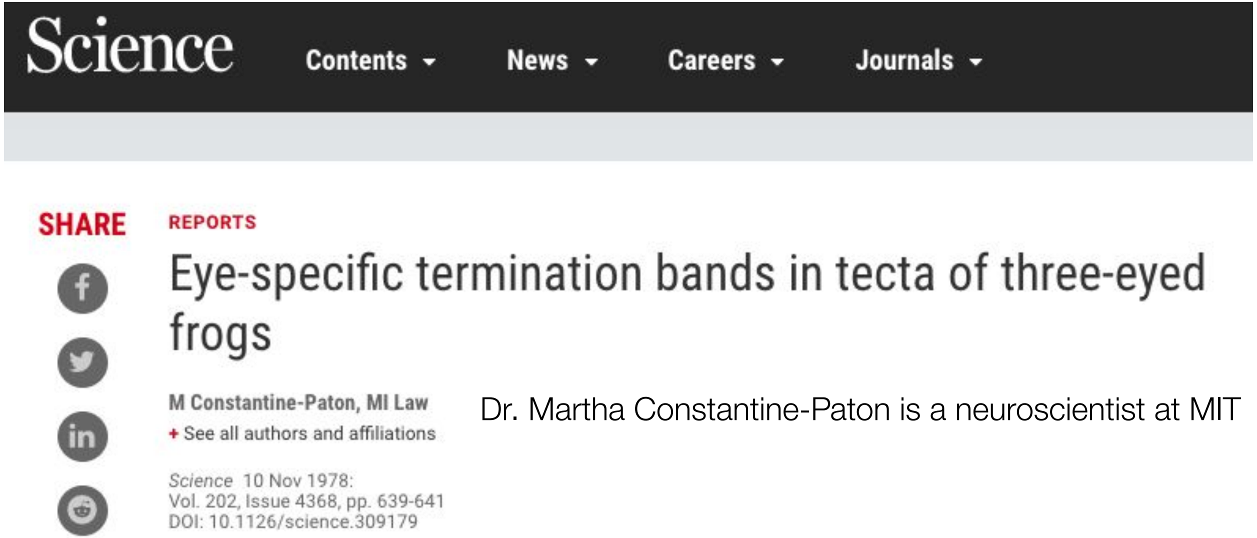
- Specially arranged mirrors flipped the light that would reach eyes, top becoming bottom, and bottom top.

After 10 days, Kohler had grown accustomed to the invariably upside-down world

- everything seemed to him normal, rightside-up
- He could do everyday activities in public perfectly well: walk along a crowded sidewalk, even ride a bicycle

https://www.theguardian.com/education/2012/nov/12/improbable-research-seeing-upside-down

Neural plasticity



REPORTS

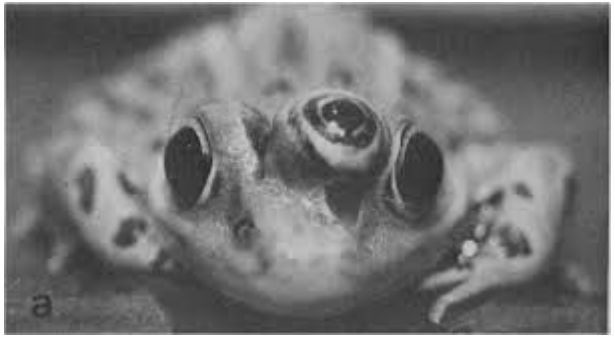
Eye-specific termination bands in tecta of three-eyed frogs

M Constantine-Paton, MI Law

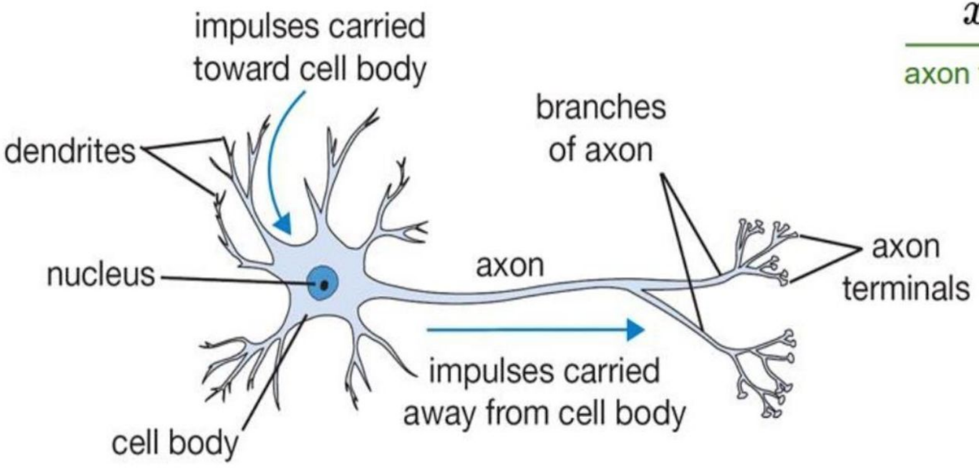
+ See all authors and affiliations

Science 10 Nov 1978:
Vol. 202, Issue 4368, pp. 639-641
DOI: 10.1126/science.309179

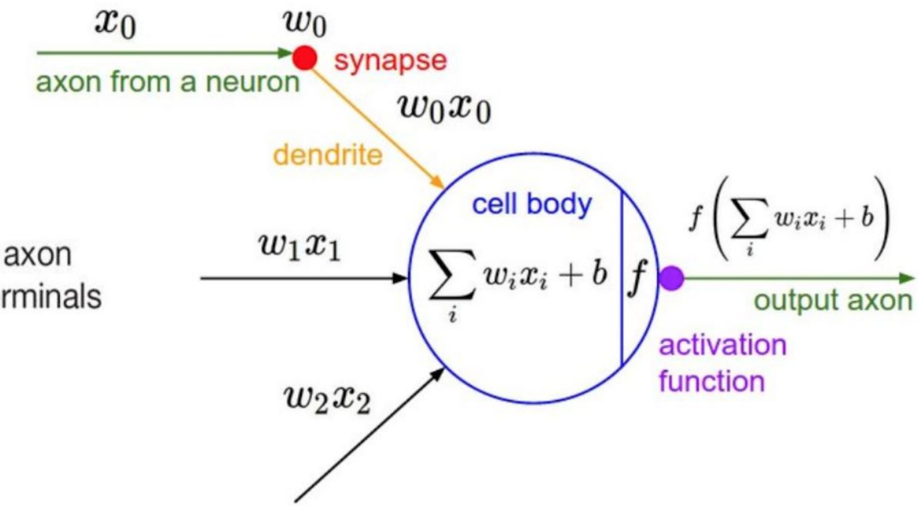
Dr. Martha Constantine-Paton is a neuroscientist at MIT

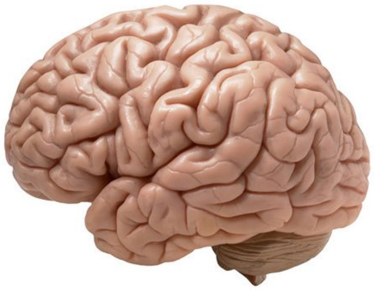


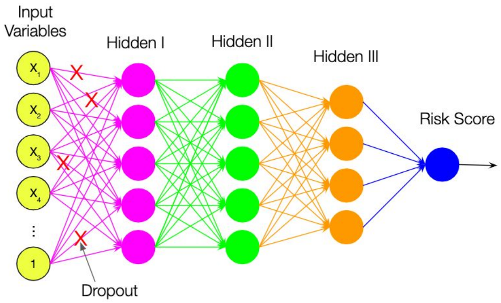
Biological vs. artificial neurons



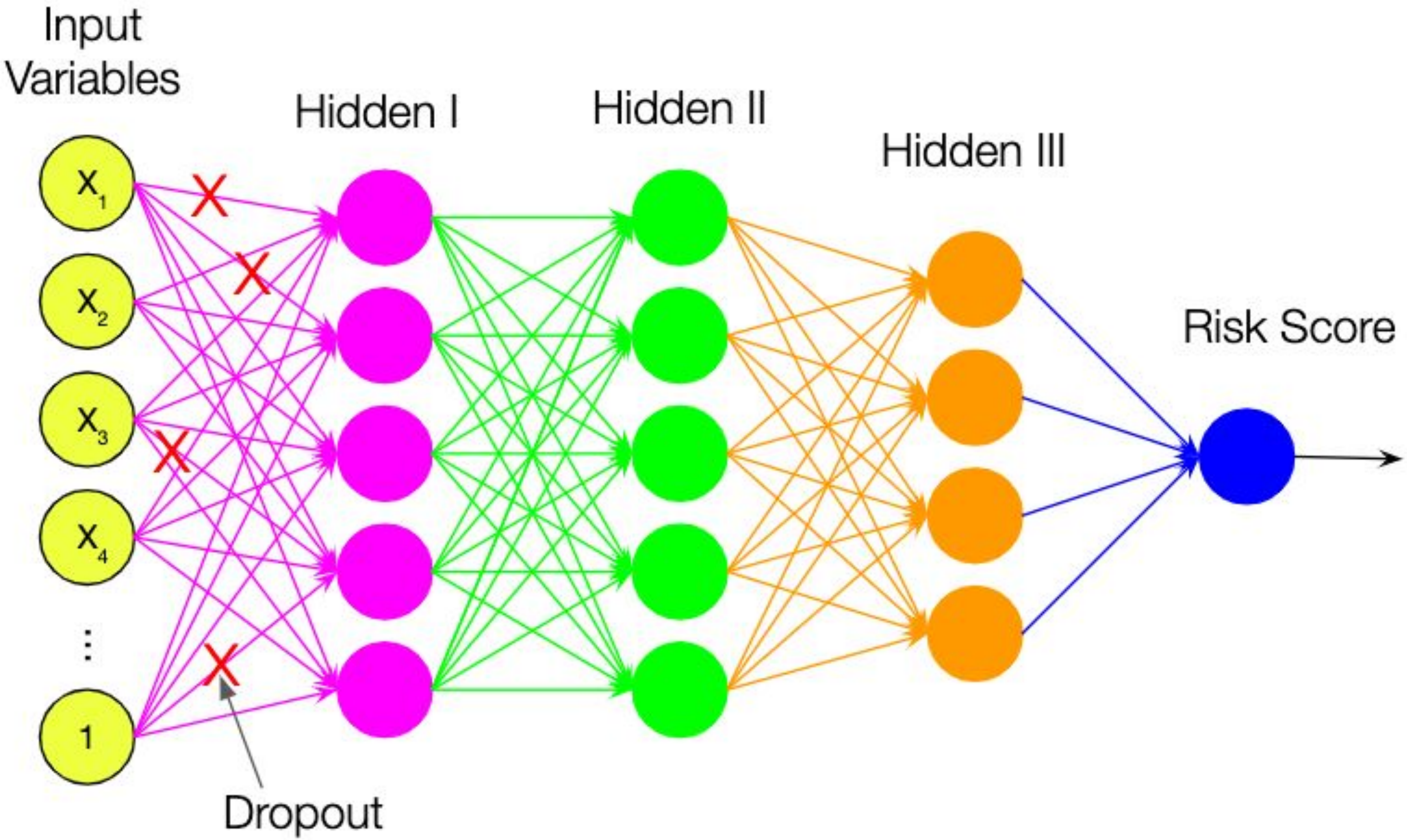
impulses carried toward cell body
dendrites
branches of axon
nucleus
axon
axon terminals
impulses carried away from cell body
cell body



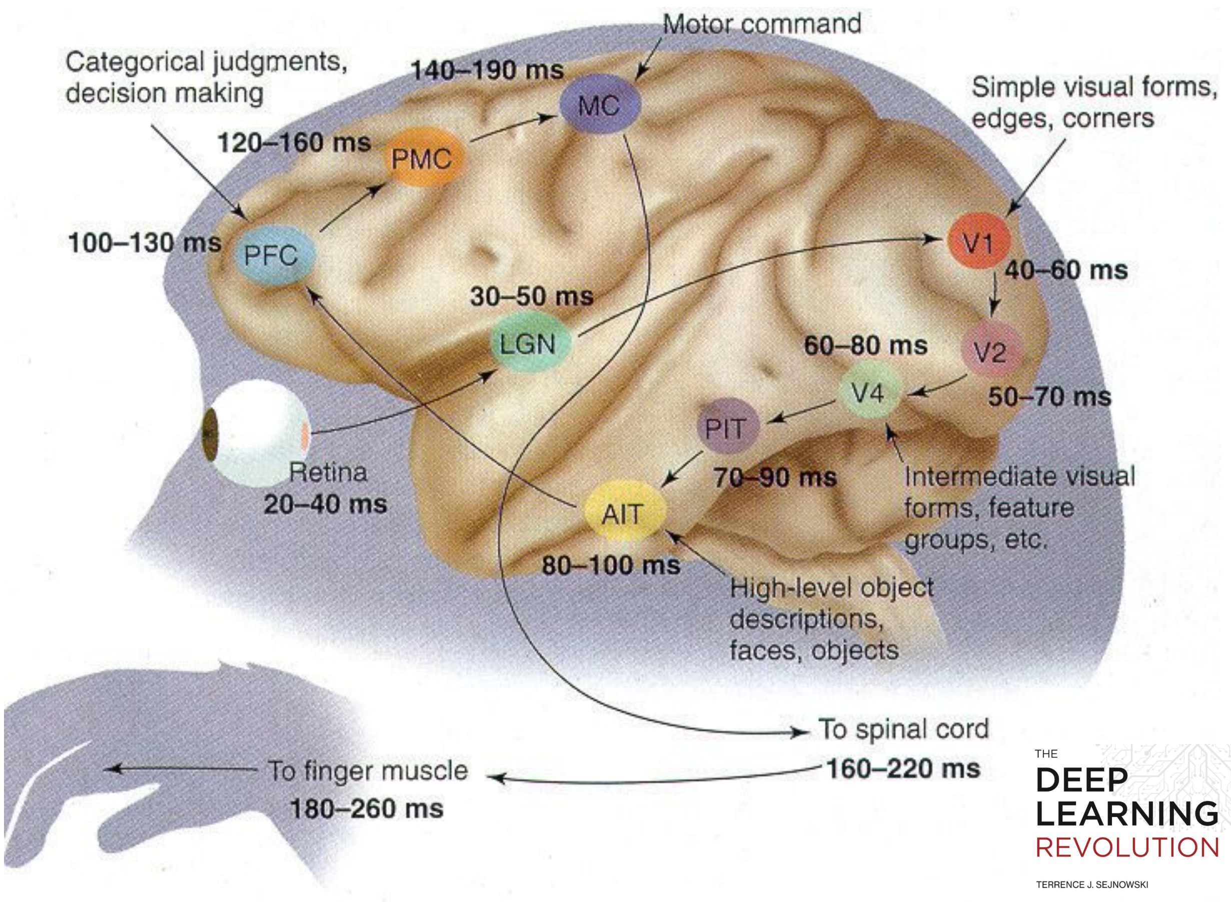




Feed-forward neural networks are very successful

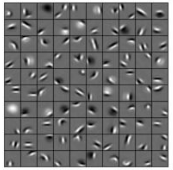


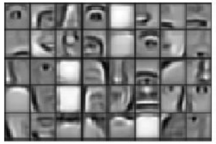
Human visual cortex is hierarchical



THE DEEP LEARNING REVOLUTION

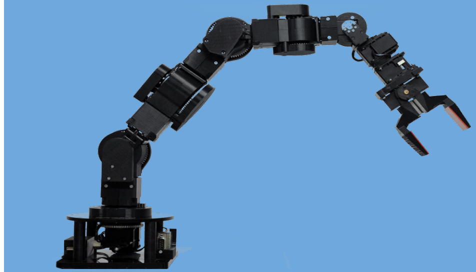
TERRENCE J. SEJNOWSKI





LGN: lateral geniculate nucleus
V1: primary visual cortex
V2: secondary visual cortex
V4: visual area 4
AIT and PIT: anterior and posterior inferotemporal cortex
PFC: prefrontal cortex
PMC: premotor cortex
MC: motor cortex

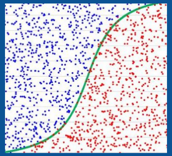
Artificial Intelligence vs. Machine Learning vs. Deep Learning



ARTIFICIAL INTELLIGENCE

a very broad field

1950s



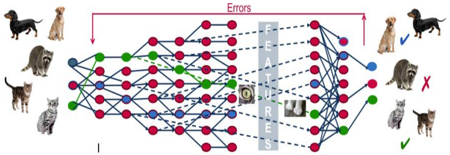
MACHINE LEARNING

Fundamentals of "learning from data"

1980s

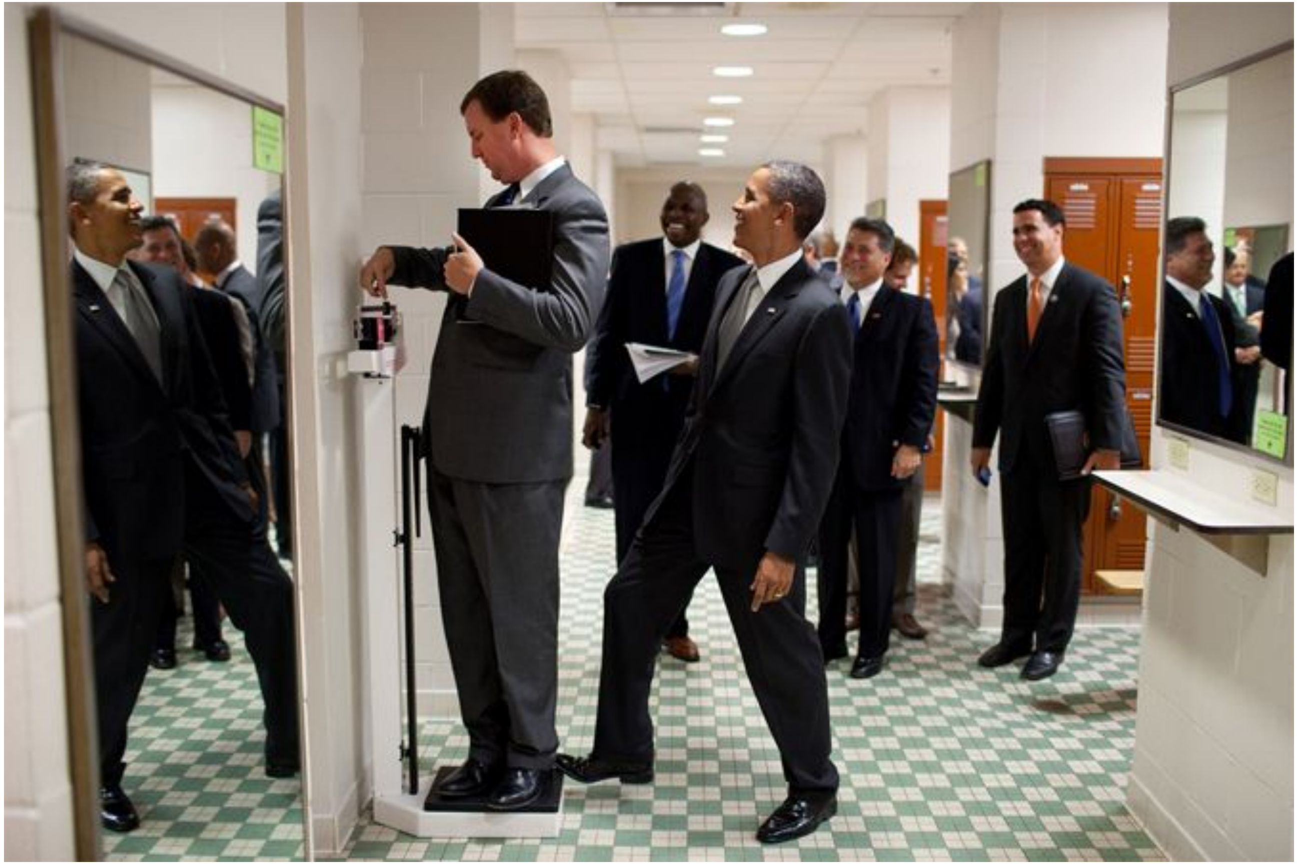
Deep Learning

Powerful trending ML methods



2010s

The state of computer vision and AI: we are really, really far away



http://karpathy.github.io/2012/10/22/state-of-computer-vision/

Some things "we" understand easily

There are **3 mirrors** in the scene so some of those people are **"fake" replicas** from different viewpoints

Recognize Obama from the few pixels that make up his face

You recognize that there's a person **standing on a scale**, even though the scale occupies only very few white pixels that blend with the background

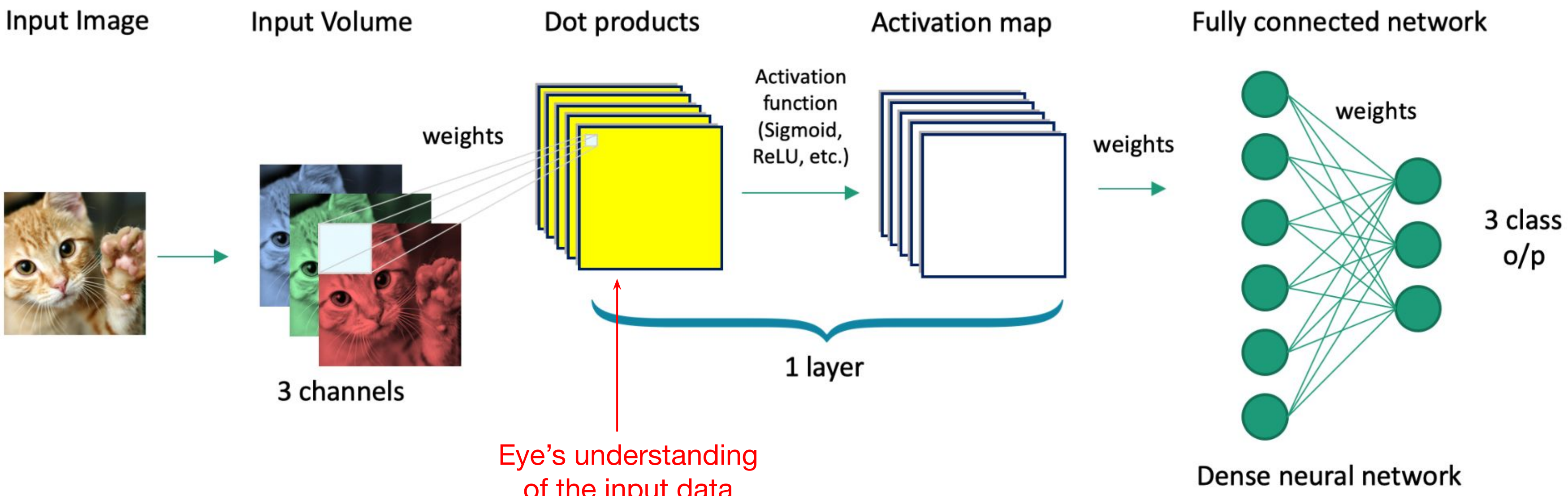
Obama has his **foot positioned just slightly** on top of the scale

Working physics - **Obama is leaning in on the scale**, which applies a force on it. Scale measures force that is applied on it, that's how it works => it will over-estimate the weight of the person standing on it.

The **person** measuring his weight **is not aware** of Obama doing this

There are **people in the back** who **find** the person's imminent confusion **funny**

Residual networks, used for distance prediction, are ConvNets



Input Image

Input Volume

Dot products

Activation map

Fully connected network

weights

3 channels

Activation function (Sigmoid, ReLU, etc.)

weights

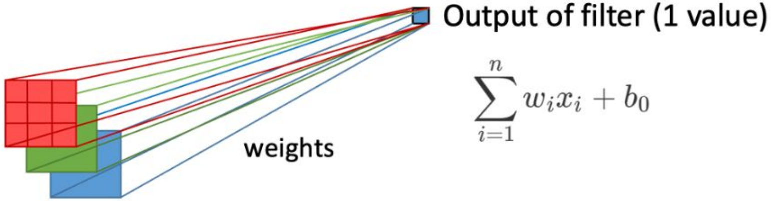
weights

1 layer

3 class o/p

Dense neural network

Eye's understanding of the input data

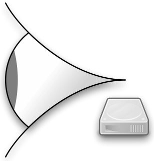


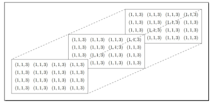
3 x 3 filter

Output of filter (1 value)

weights

$$\sum_{i=1}^{n} w_i x_i + b_0$$

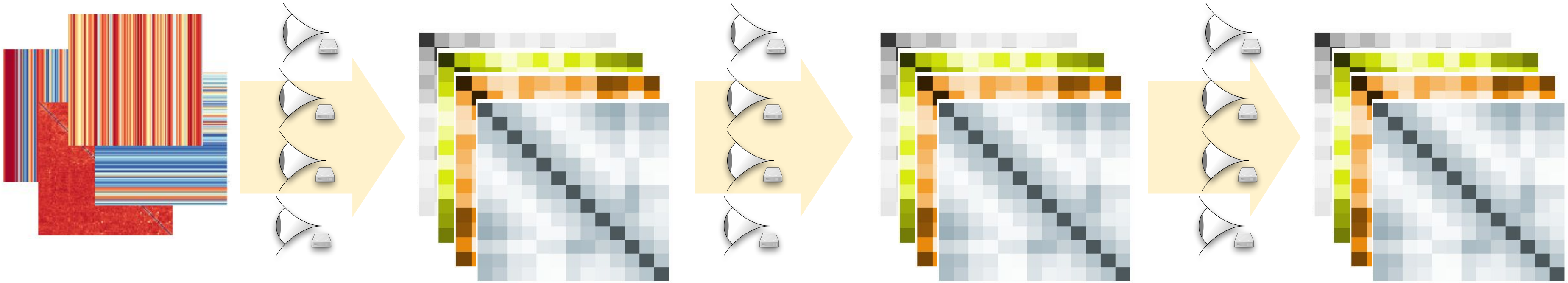




Mathematically,
It is simply a
FxFxC matrix

Convolutional neurons are like our "**eyes with memory**"..

Residual networks



Input Features

1st Layer’s Feature Mappings

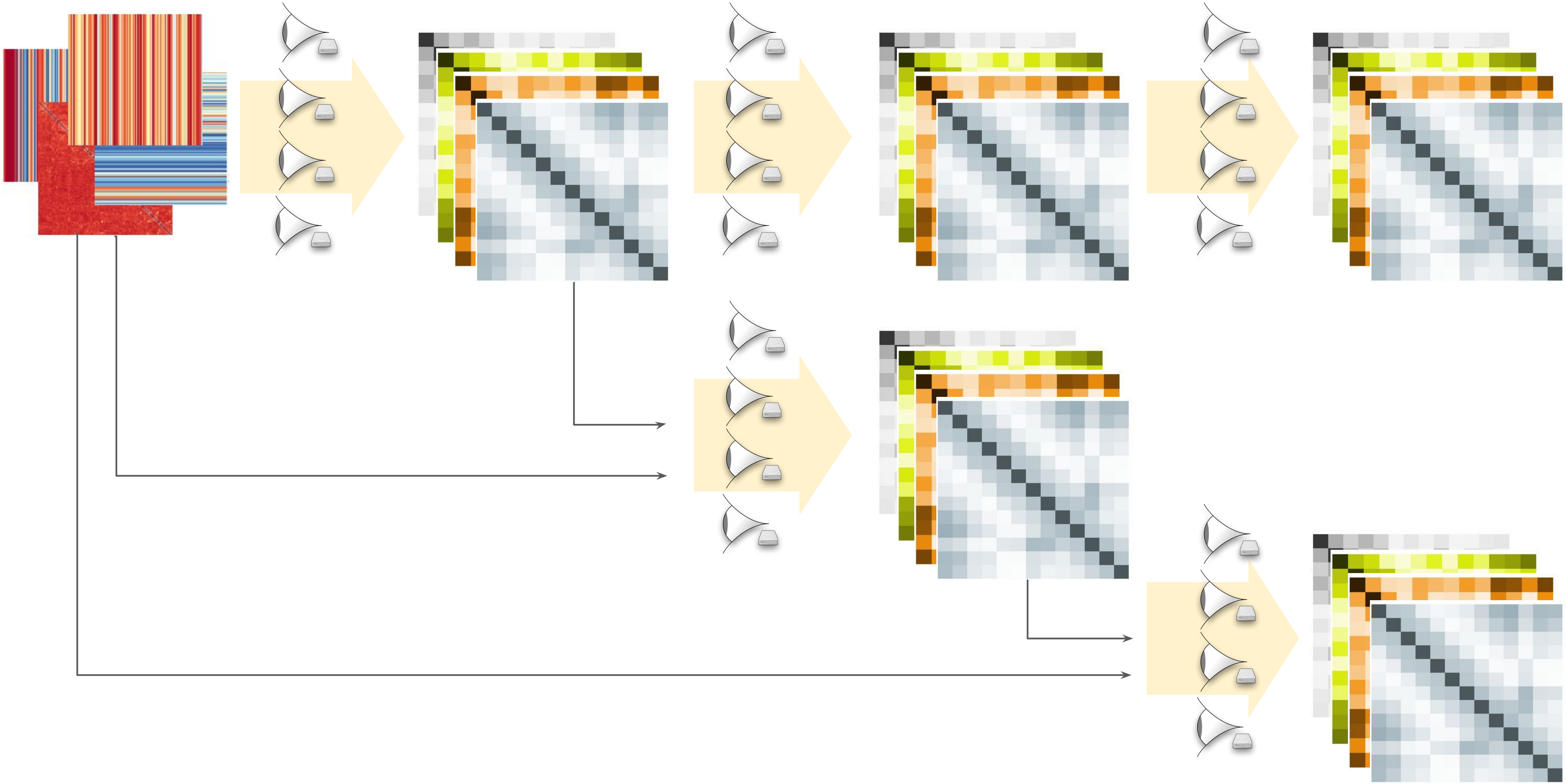
2nd Layer’s FM

3rd Layer’s FM

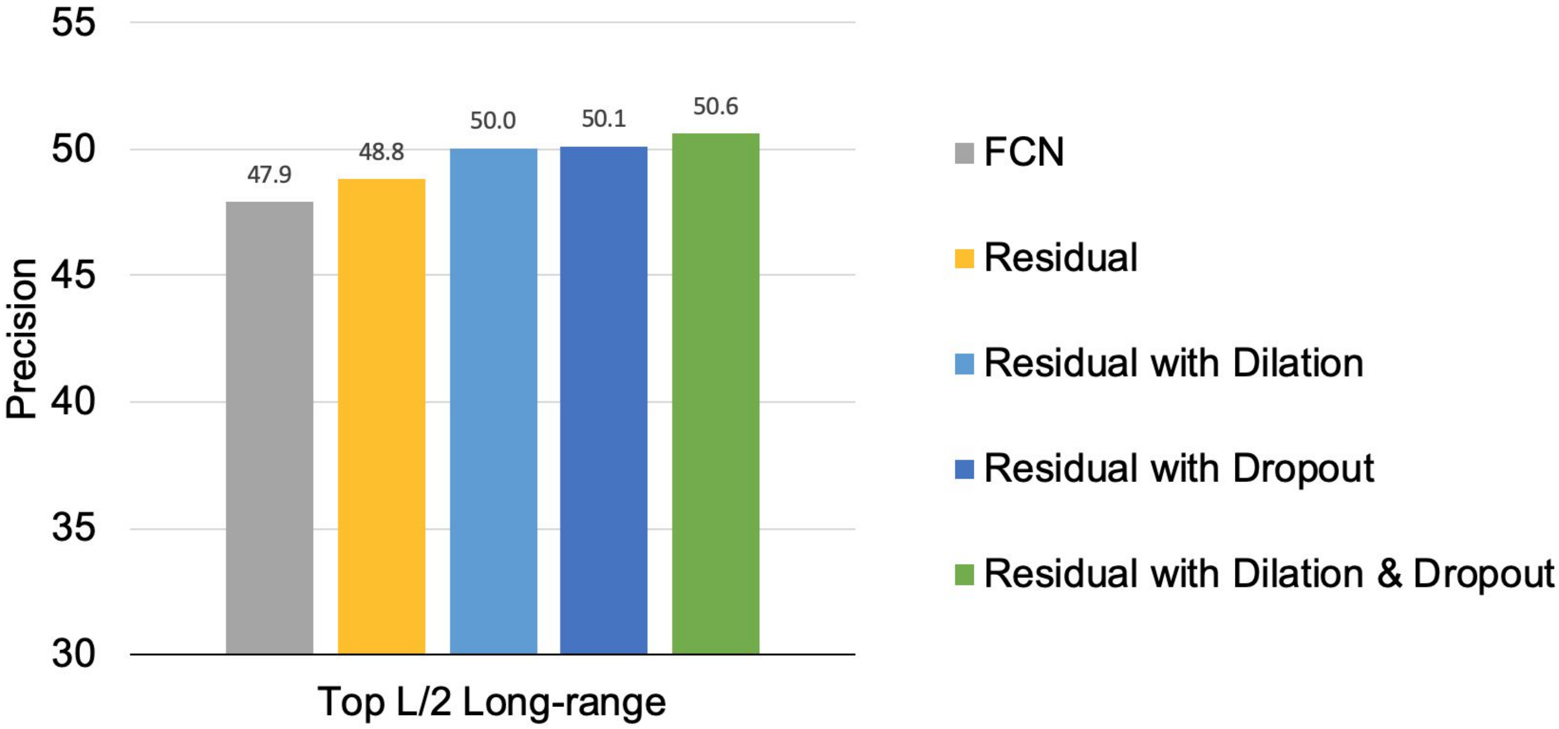
Information continues to lose ..

.. learning becomes difficult

Residual networks



Variants of residual networks can perform even better



DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout

Badri Adhikari ✉

Bioinformatics, btz593, https://doi.org/10.1093/bioinformatics/btz593

Published: 29 July 2019 **Article history ▾**

Hardware for protein distance prediction

Most current deep learning experiments are performed on sample datasets:

- 3 K representative proteins [200 GB]
- Special SSDs known as M2s that directly attach to the motherboard
- Powerful GPUs such as V100 and P6000 are required
- One experiment (training) takes about 24 hours

Full dataset:

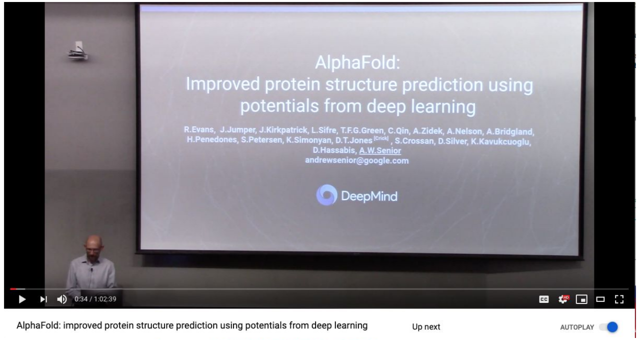
- 50 K proteins [around 10 TB]
- One experiment can take up to 10 days

Feature generation:

- 1000s of CPU time for a few days

Present and Future Research in Protein Folding

Increased interest from industry



AlphaFold:
Improved protein structure prediction using potentials from deep learning

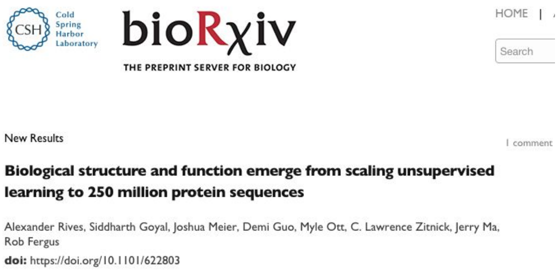
R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, D.T.Jones[Crick], S.Crossan, D.Silver, K.Kavukcuoglu, D.Hassabis, A.W.Senior
andrewsenior@google.com

DeepMind

AlphaFold: improved protein structure prediction using potentials from deep learning

INSTITUTE FOR
Protein Design
UNIVERSITY of WASHINGTON

Facebook AI Research



CSH Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
HOME |
Search

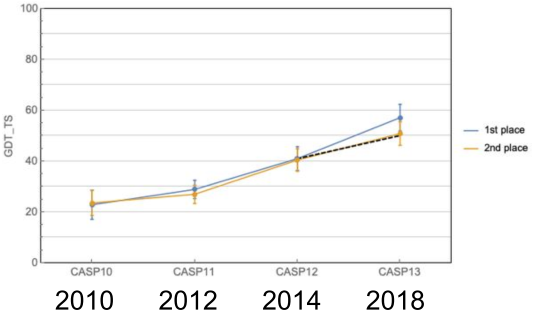
New Results

1 comment

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

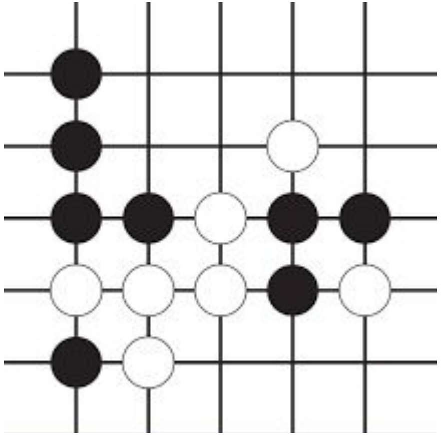
Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, Rob Fergus

doi: https://doi.org/10.1101/622803



Reinforcement learning.. a big hope..

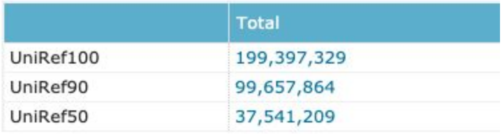
Chess



Game of Go

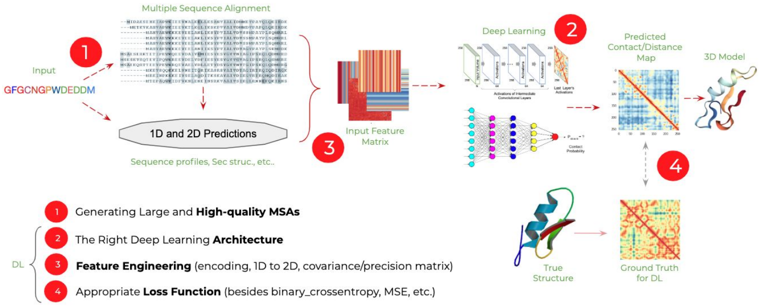
?

A lot of data + many powerful algorithms = much work to do



	Total
UniRef100	199,397,329
UniRef90	99,657,864
UniRef50	37,541,209

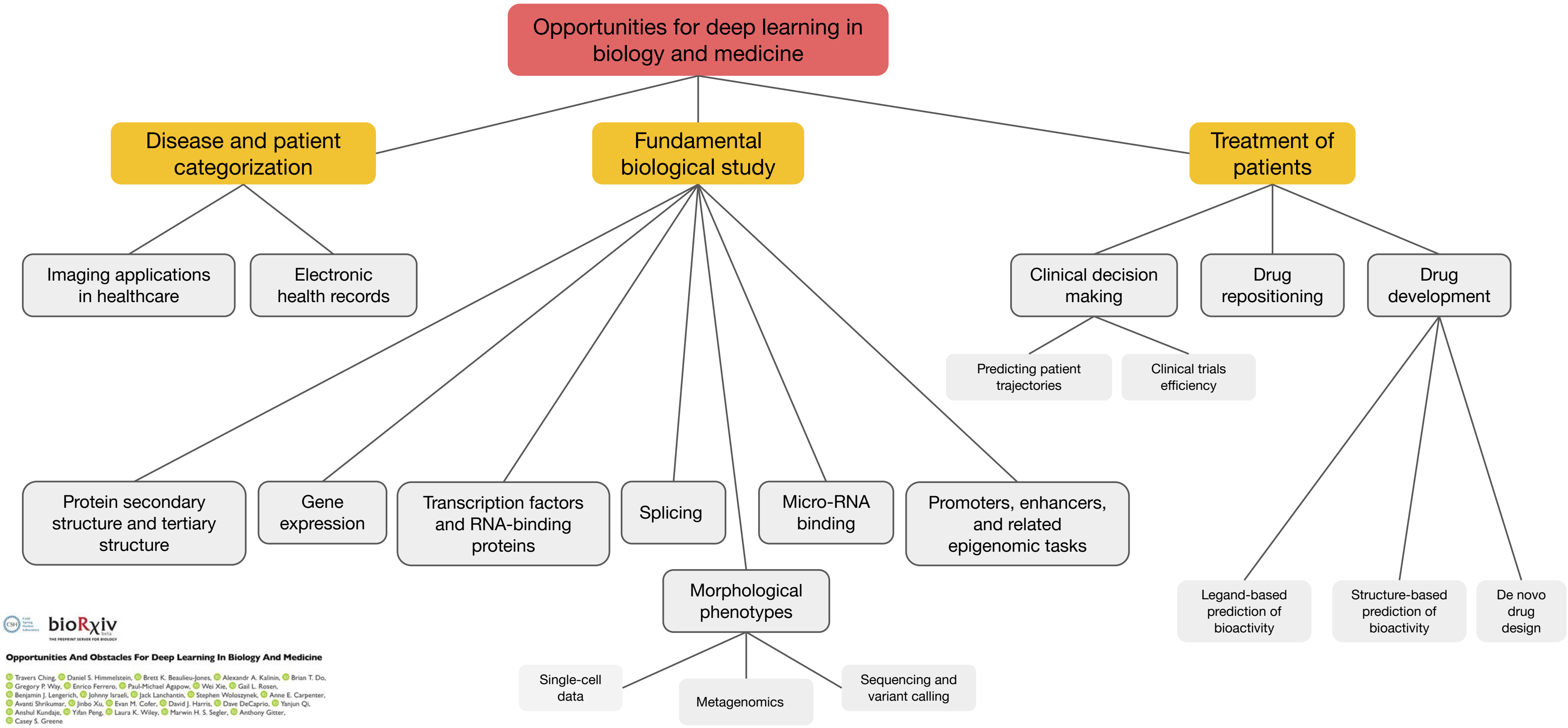
+ Deep Learning +



1. Generating Large and **High-quality MSAs**
2. The Right Deep Learning **Architecture**
3. **Feature Engineering** (encoding, 1D to 2D, covariance/precision matrix)
4. Appropriate **Loss Function** (besides binary_crossentropy, MSE, etc.)

A unique problem

Deep learning for biology and medicine


Opportunities for deep learning in biology and medicine
Disease and patient categorization
Fundamental biological study
Treatment of patients
Imaging applications in healthcare
Electronic health records
Clinical decision making
Drug repositioning
Drug development
Predicting patient trajectories
Clinical trials efficiency
Protein secondary structure and tertiary structure
Gene expression
Transcription factors and RNA-binding proteins
Splicing
Micro-RNA binding
Promoters, enhancers, and related epigenomic tasks
Morphological phenotypes
Single-cell data
Metagenomics
Sequencing and variant calling
Legand-based prediction of bioactivity
Structure-based prediction of bioactivity
De novo drug design

bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY

Opportunities And Obstacles For Deep Learning In Biology And Medicine

Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Anthony Gitter, Casey S. Greene

doi: https://doi.org/10.1101/142760

Conclusions

1) Deep learning methods are full of promise but also have a lot of limitations
2) A key component of the protein folding problem, distance prediction, is largely a deep learning problem
3) Solving the problem of protein folding requires expertise from both domains - deep learning and bioinformatics
4) Protein folding problem will potentially unravel the limitations of AI and DL

Acknowledgements

Research Support & Contribution

Cezary Janikow

Sharlee Climer

Cynthia Jobe

Anthony Ackah-Nyanzu

Patrick Kong

Sri Harsha Akurathi

IT Support

Philip Reiss

Kenneth Voss

Michael Remier

MU - Research Computing Support Services (RCSS)

Computing Resources

Google

NVIDIA®

University of Missouri System
COLUMBIA | KANSAS CITY | ROLLA | ST.LOUIS

THANK
YOU