prepare.ai

UMSL | Department of Mathematics & Computer Science

Deep Learning Shines New Hopes on Solving the Half-a-Century-Old Problem of Protein Folding

Machine Learning & Protein Structure Prediction Lab

Badri Adhikari
adhikarib @ umsl.edu

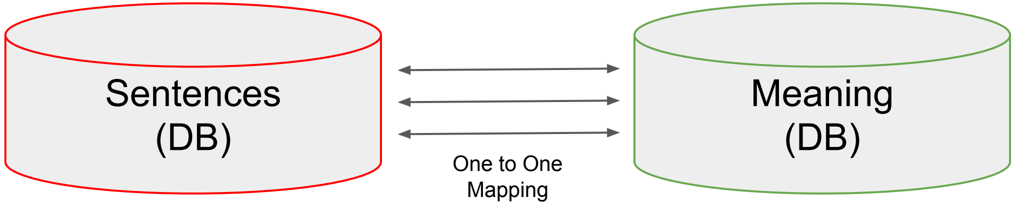
Assistant Professor of CS
Department of Mathematics & Computer Science
University of Missouri-St. Louis

Topics

- What is protein contact prediction?
- Early approaches (Feedforward NNs)
- Recent methods (ConvNets)
- How Deep Learning Contributes
- Conclusions

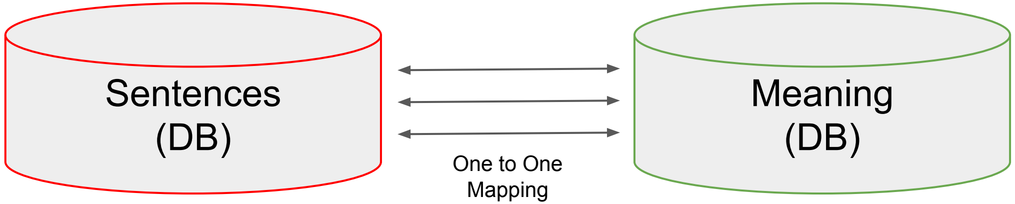
A Hypothetical Problem

- We would like to predict the meaning of a sentence …
  - Using machine learning



A Hypothetical Problem

- We would like to predict the meaning of a sentence …
    - Using machine learning



- How to represent meanings?

Sentence	Meaning
"Everyone should learn how to program because it teaches you how to think."	?

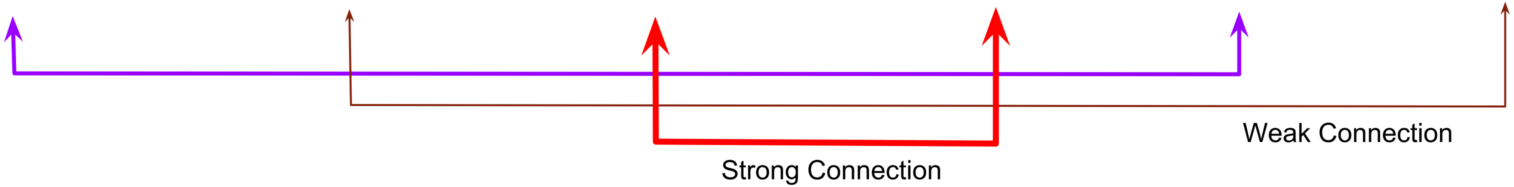
How to Represent the Meaning of a Sentence?

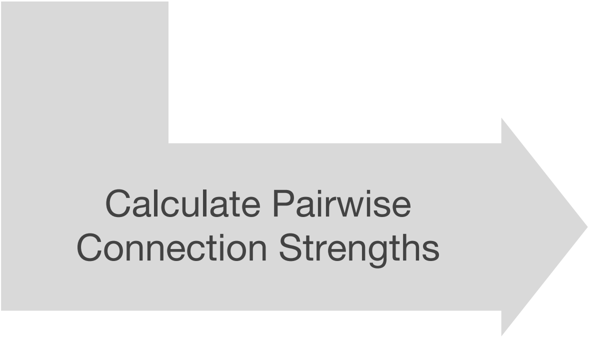
"Everyone should learn how to program because it teaches you how to think."

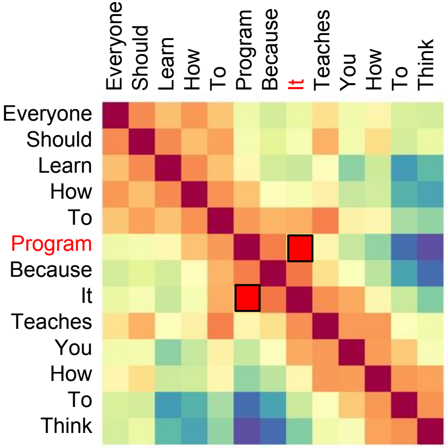
Strong Connection

How to Represent the Meaning of a Sentence?

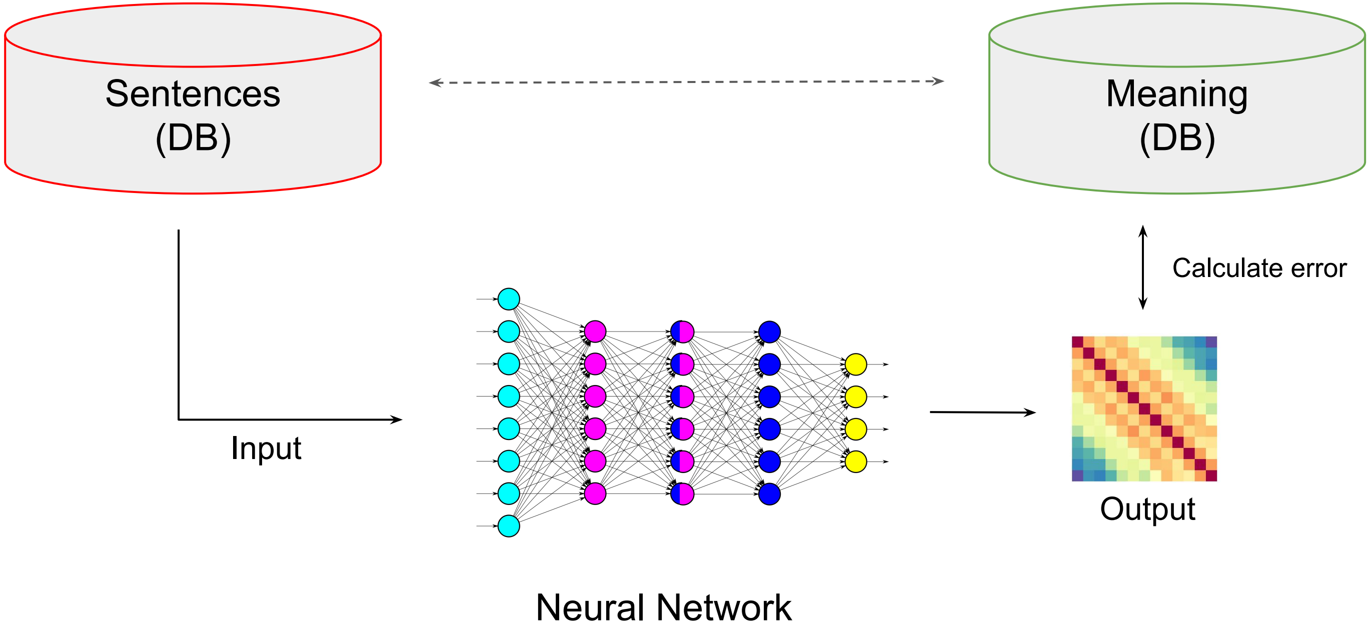
"Everyone should learn how to program because it teaches you how to think."







Machine Learning to Predict the Meaning of a Sentence



What is Protein Contact Prediction?

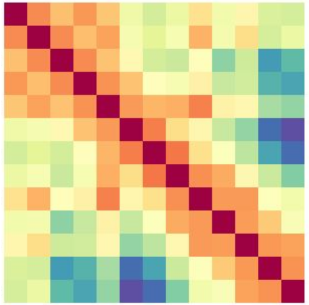
"Everyone should learn how to program because it teaches you how to think."

English Sentence

G F G C N G P W D E D D M

Protein Sequence

Predict which amino acids interact with which..



Distance Map

What is Protein Contact Prediction?

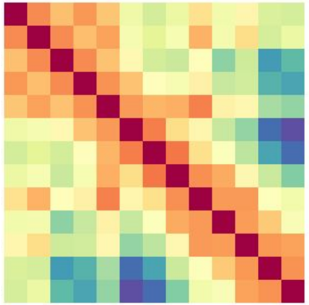
"Everyone should learn how to program because it teaches you how to think."

English Sentence

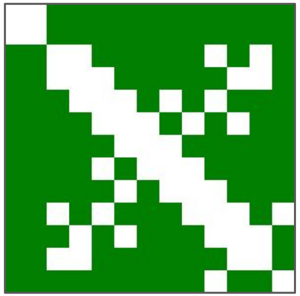
G F G C N G P W D E D D M

Protein Sequence

Predict which amino acids interact with which..

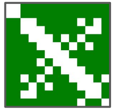


Distance Map



Contact Map

Why Predict Contacts?

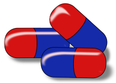
Precise protein contact prediction 

Leads to..

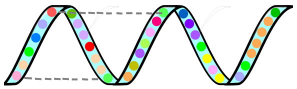
Accurate protein structure / function prediction 

Leads to..

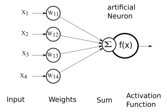
Curing diseases through drug design
(cancer, mental health diseases)



Better understanding of how life works
(through understanding of how proteins work)



Improvements in Machine / Deep Learning
(because contact prediction is a hard problem)



Early Approaches to Contact Prediction

Feature Engineering

"Everyone should learn how to program because it teaches you how to think."

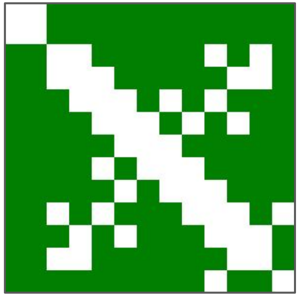
	Everyone	should	learn	how	to	program	because	it	teaches	you	how	to	think
Character Length	8	6	5	3	2	7	7	2	7	3	3	2	5
Noun (or not)	0	0	0	0	0	0	0	0	0	0	0	0	0
Pronoun (or not)	1	0	0	0	0	0	0	1	0	1	0	0	0
…	….												

	G	F	G	C	N	G	P	W	D	E	D	D	M
Polar (or not)	0	0	0	0	1	0	0	0	1	1	1	1	0
Positively charged (or not)	0	0	0	0	0	0	0	0	0	0	0	0	0
Hydrophobic (or not)	0	1	0	0	0	0	1	1	0	0	0	0	1
Helical (or not)	0	0	0	0	0	0	0	0	0	0	0	0	0
…	….												

Approach: Consider Each Contact as a Separate Problem

How likely are these two close?

	G	F	G	C	N	G	P	W	D	E	D	D	M
Polar (or not)	0	0	0	0	1	0	0	0	1	1	1	1	0
Positively charged (or not)	0	0	0	0	0	0	0	0	0	0	0	0	0
Hydrophobic (or not)	0	1	0	0	0	0	1	1	0	0	0	0	1
Helical (or not)	0	0	0	0	0	0	0	0	0	0	0	0	0
...	....												



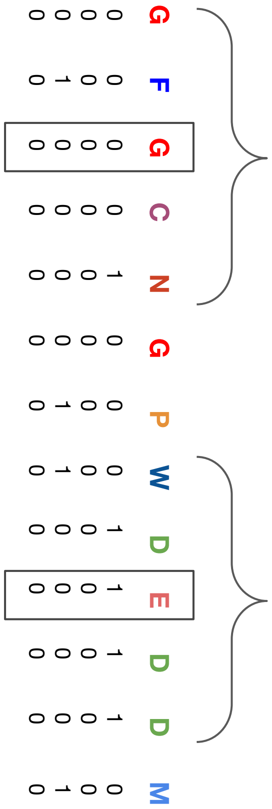
Predict Contact Map

Pair	In Contact?
G_1 - F_2	Y
G_1 - G_3	N
G_1 - C_4	N
...	
G_2 - E_{10}	Y
...	

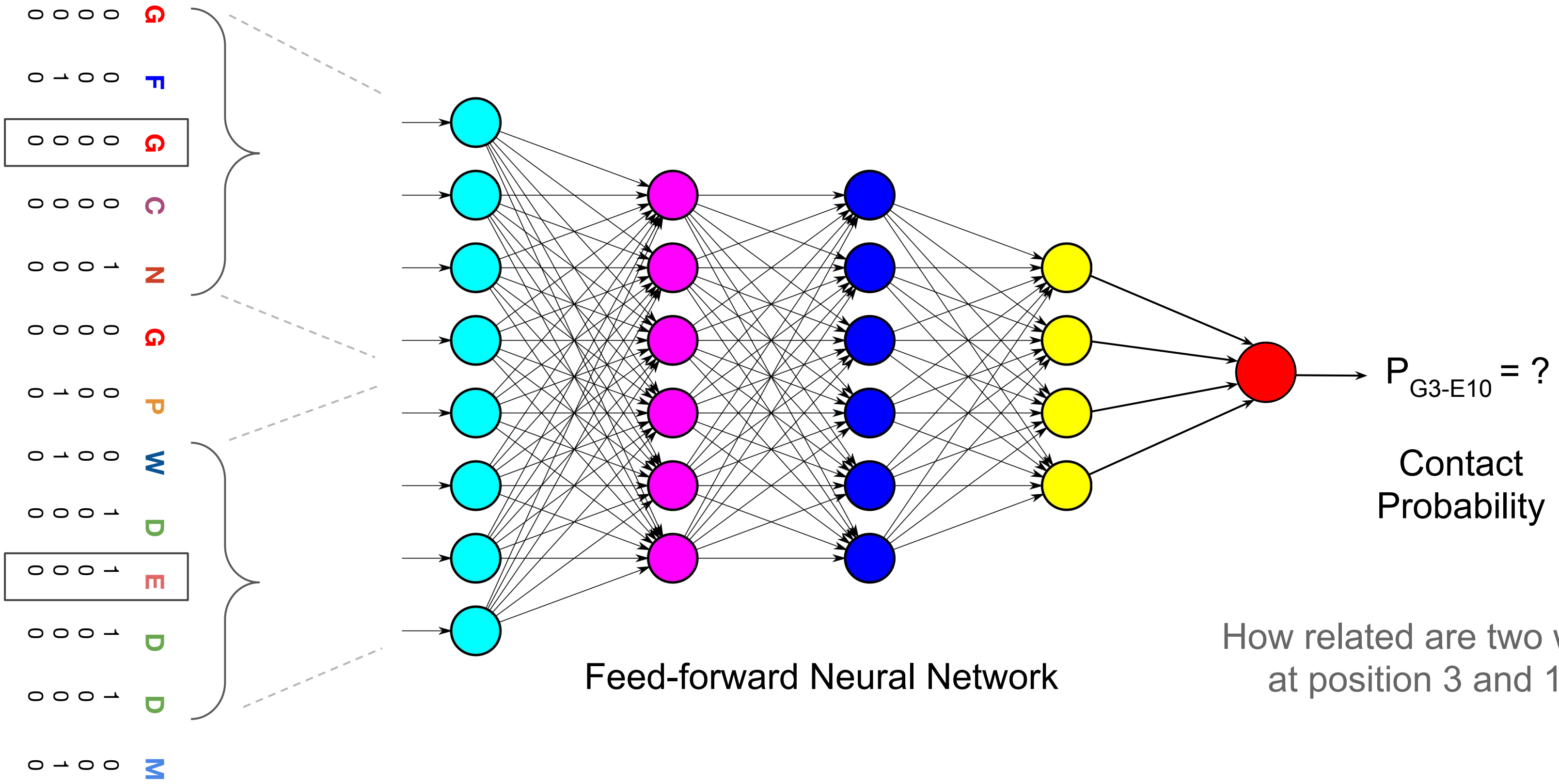
Predict Contacting Pairs
(instead of a full map)

Pros:
- Many subproblems!
- Feasible to train
- More data to train

Training and Testing



Training and Testing



How related are two words at position 3 and 10?

Early Approaches

2004

The predictor neural network is a **standard feed-forward network, with 56 inputs** as given above, **ten hidden units, and a single output**.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Protein contact prediction using patterns of correlation

Nicholas Hamilton, Kevin Burrage, Mark A. Ragan, Thomas Huber

First published: 14 May 2004 | https://doi.org/10.1002/prot.20160 | Cited by: 44

2005

PROFcon, a new method for predicting inter-residue contacts through a simple neural network..
We considered **information from two 'windows' around two residues i and j** for which the probability of a spatial contact was predicted... We used 738 input, **100 hidden** and 2 output units (contact/non-contact)

PROFcon: novel prediction of long-range contacts

Marco Punta, Burkhard Rost

Bioinformatics, Volume 21, Issue 13, , Pages 2960–2968, https://doi.org/10.1093/bioinformatics/bti454

Published: 12 May 2005 **Article history** ▾

We develop a new contact map predictor (SVMcon) that **uses support vector machines** to predict medium- and long-range contacts… SVMcon integrates profiles, secondary structure, relative solvent accessibility, contact potentials, and other useful features...

2007

Improved residue contact prediction using support vector machines and a large feature set

Jianlin Cheng and Pierre Baldi

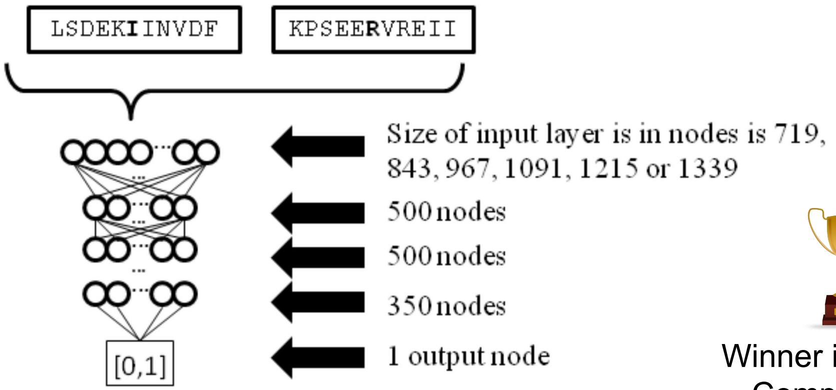
BMC Bioinformatics 2007 **8**:113

https://doi.org/10.1186/1471-2105-8-113 | © Cheng and Baldi; licensee BioMed Central Ltd. 2007

Received: 28 December 2006 | **Accepted:** 02 April 2007 | **Published:** 02 April 2007

Standard Feed-forward Neural Networks Worked for 2^4 Years

2012



Winner in CASP Competition (2012)

Predicting protein residue–residue contacts using deep networks and boosting

Jesse Eickholt, Jianlin Cheng Author Notes

Bioinformatics, Volume 28, Issue 23, 1 December 2012, Pages 3066–3072,
https://doi.org/10.1093/bioinformatics/bts598
Published: 09 October 2012 **Article history ▾**

2014

.. Classifiers are **classic feed-forward neural networks**, with **55 hidden units and a single output unit**..
.. To train these very large networks, alternate rounds of offline and online training are carried out until no further improvement in accuracy is obtained..

Winner in CASP Competition (2014)

MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins

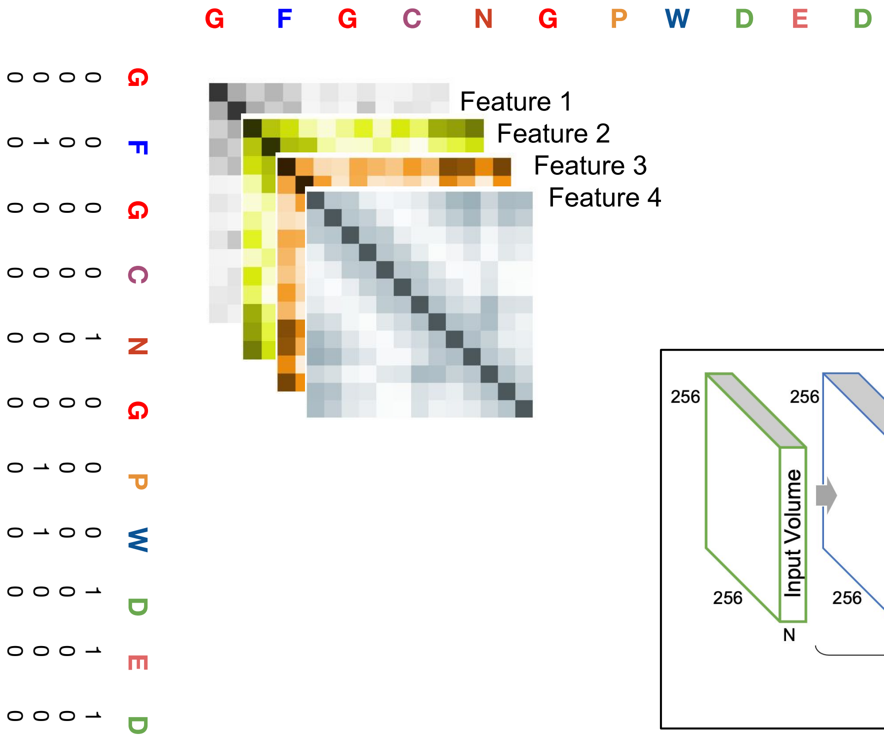
David T. Jones, Tanya Singh, Tomasz Kosciolek, Stuart Tetchner Author Notes

Bioinformatics, Volume 31, Issue 7, 1 April 2015, Pages 999–1006,
https://doi.org/10.1093/bioinformatics/btu791
Published: 26 November 2014 **Article history ▾**

Recent Methods

In 2016, Jinbo Xu's Group Tested ConvNets...

.. By stacking **multiple convolution layers**, the network can **learn information in a very large sequential context**..
.. test results suggest that **deep learning can revolutionize protein contact prediction**...



Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

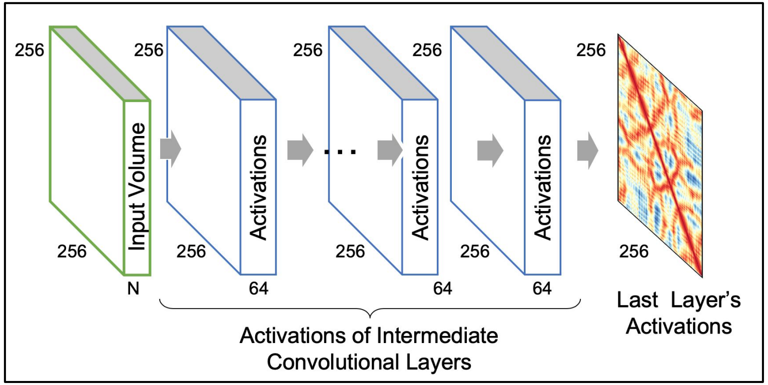
Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, Jinbo Xu

doi: https://doi.org/10.1101/073239

Now published in *PLOS Computational Biology* doi: 10.1371/journal.pcbi.1005324

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Winner in CASP Competition (2016 & 2018)



A single ConvNet is Much More Accurate than Feed-forward NN

..Using pair frequency data as the input features, **DeepCov is able to almost match the average performance of MetaPSICOV2**, which is quite impressive given the simplicity of the feature set..

High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features

David T Jones ✉, Shaun M Kandathil

Bioinformatics, Volume 34, Issue 19, 01 October 2018, Pages 3308–3315,
https://doi.org/10.1093/bioinformatics/bty341
Published: 26 April 2018 **Article history ▾**

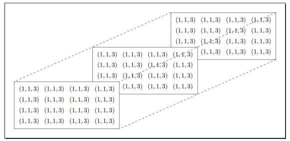
..With the same features as input, a CNN network trained with all contacts and non-contacts achieves a slightly better precision of 35.4% on top L/5 long-range contacts than DNCON 1.0. So, a **single CNN model performs better than a boosted and ensembled deep belief networks**, suggesting that the deep convolutional neural network (CNN) is more suitable for contact prediction than the deep belief network (DBN)..

DNCON2: improved protein contact prediction using two-level deep convolutional neural networks

Badri Adhikari, Jie Hou, Jianlin Cheng ✉

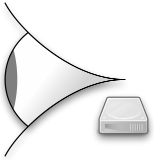
Bioinformatics, Volume 34, Issue 9, 01 May 2018, Pages 1466–1472,
https://doi.org/10.1093/bioinformatics/btx781
Published: 08 December 2017 **Article history ▾**

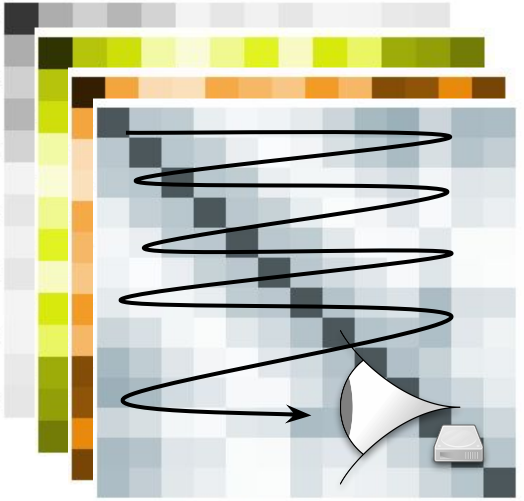
How & Why Do ConvNets Work?

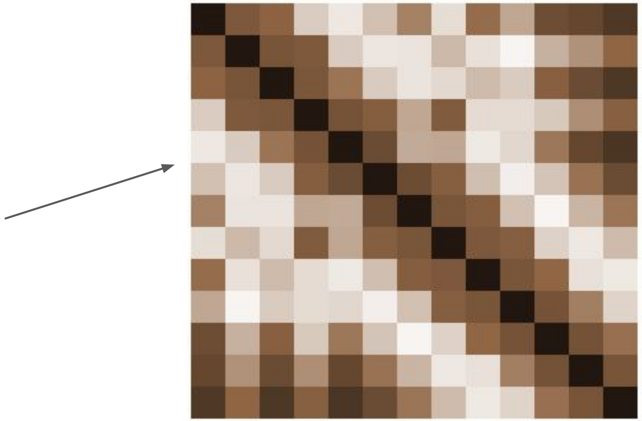


Artificial neurons are inspirations of biological neurons..

Convolutional neurons are like our “**eyes with memory**”..







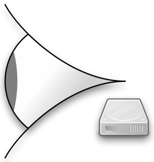
Eye’s understanding of the input data

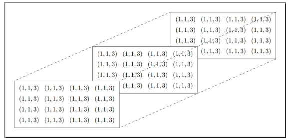
1st Layer of Conv. Neurons

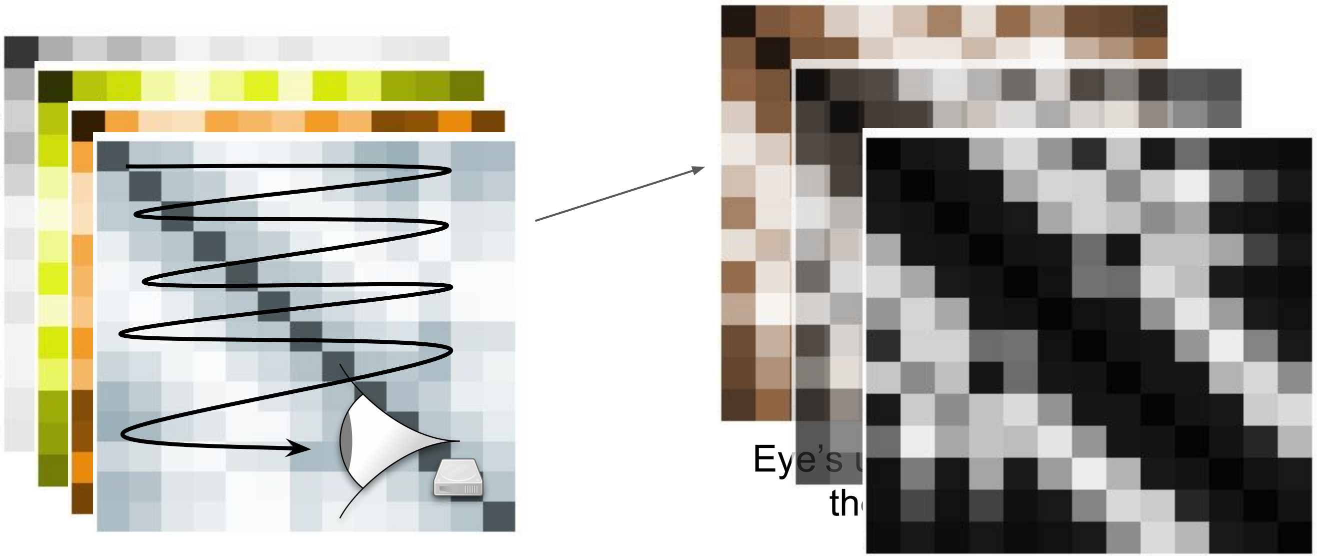
How & Why Do ConvNets Work?

Artificial neurons are inspirations of biological neurons..

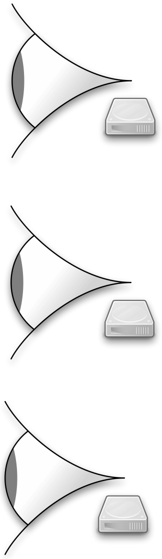
Convolutional neurons are like our “**eyes with memory**”..







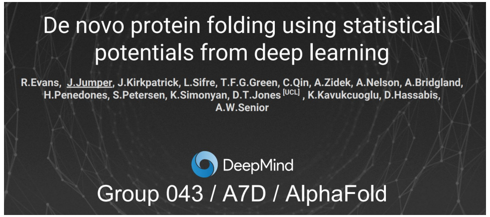
1st Layer of Conv. Neurons

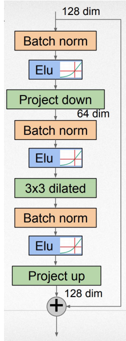


2nd Layer of Conv. Neurons

What ConvNet Architectures are Best Fit for This Problem?

Top Methods in the most recent CASP Competition



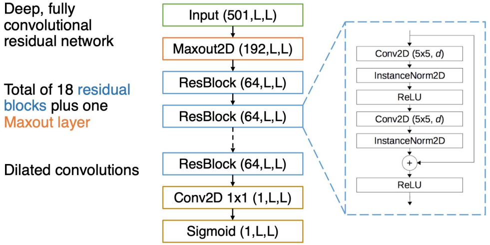


UCL

DeepMetaPSICOV (DMP) in CASP13

Shaun M Kandathil
University College London
&
The Francis Crick Institute

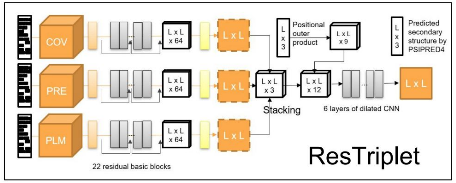
DeepMetaPSICOV model architecture



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

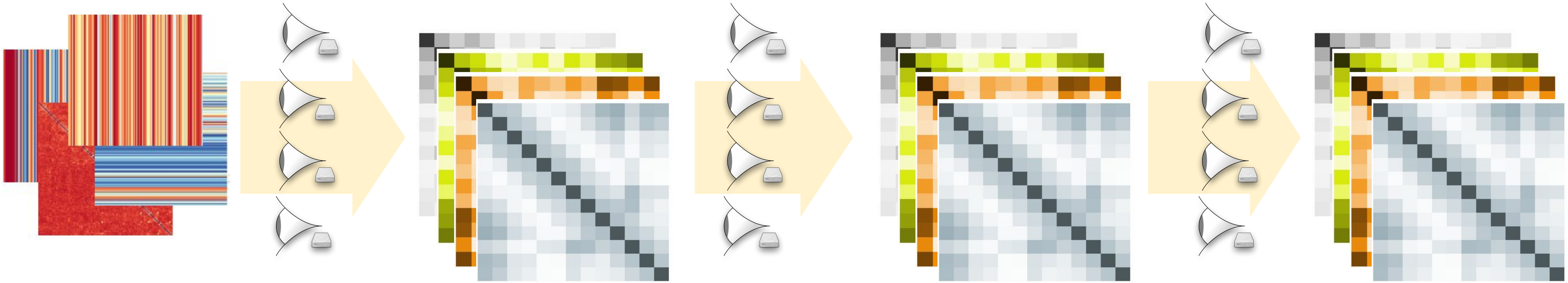
Eric W. Bell, Yang Li, Chengxin Zhang, Dong-Jun Yu, Yang Zhang

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



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

What are Residual Networks?



Input Features

1st Layer's Feature Mappings

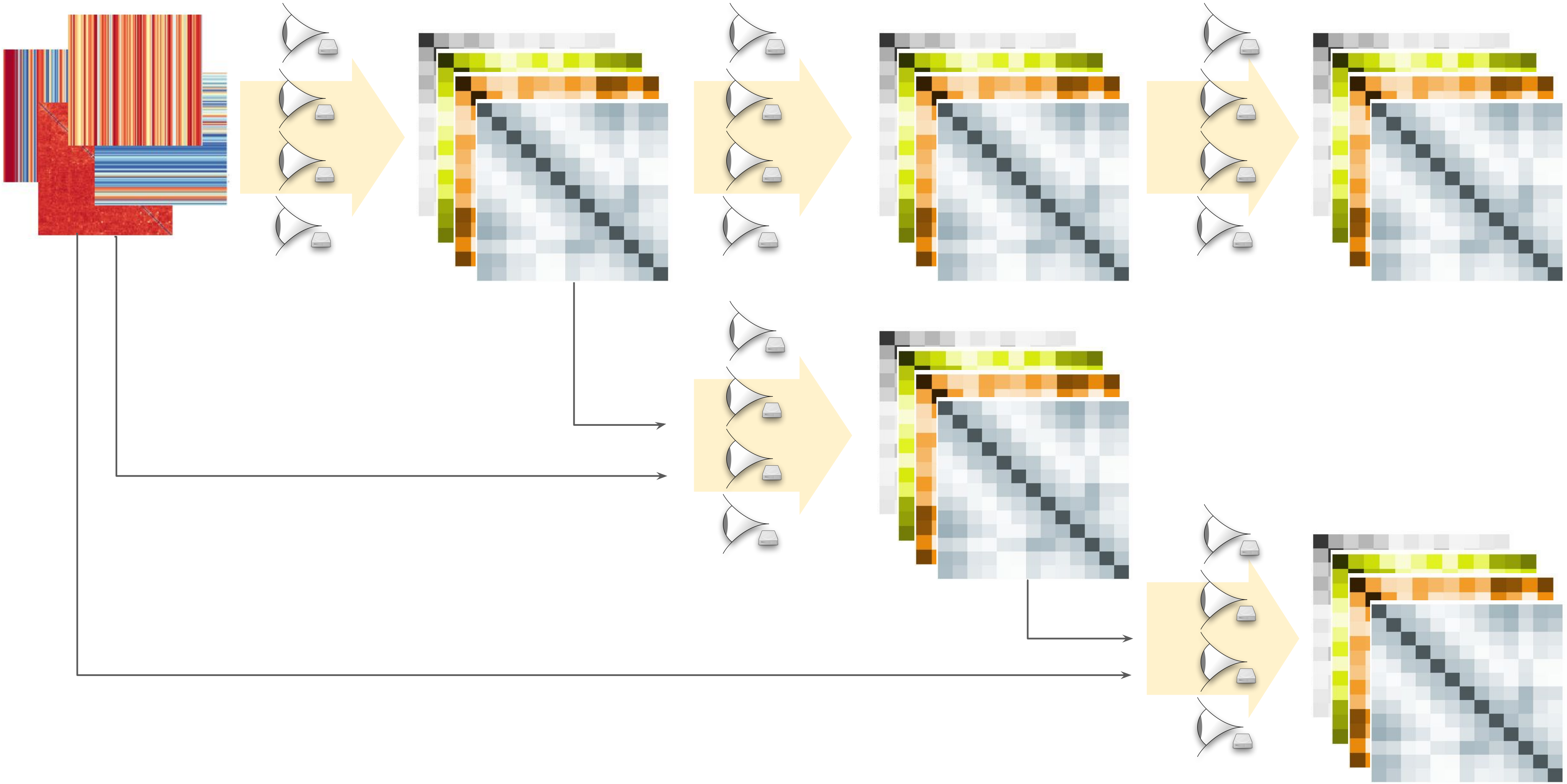
2nd Layer's FM

3rd Layer's FM

Information continues to be lost..

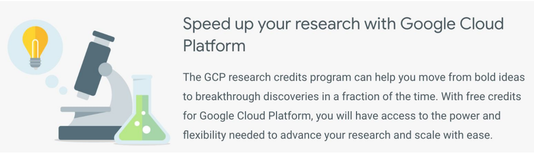
.. learning becomes difficult

What are Residual Networks?



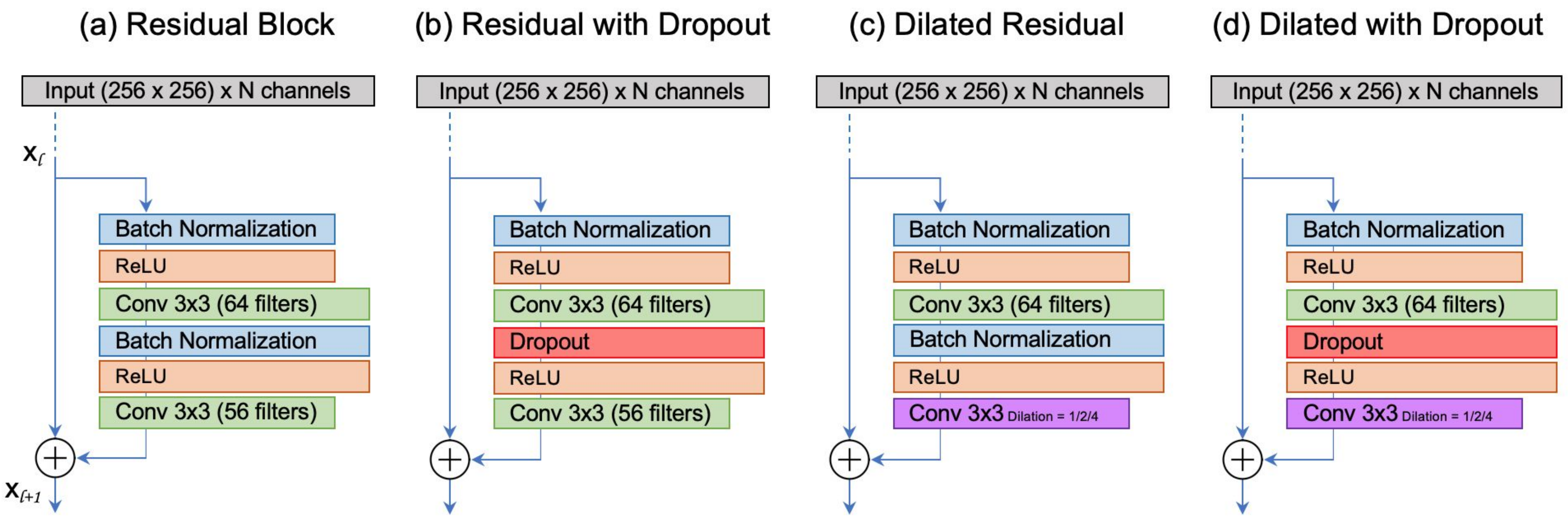
What Variations of Residual Architectures are Best Fit?

- To obtain an answer we have to try ‘almost’ all possible architectures
  - A lot of computing resources (GPUs)
- The input data for training is [2 GB to 200 GB+]
  - In one epoch (less than 20 minutes) we need to read 200 GB of data
  - On HPC clusters such as Lewis, training takes at least 10 days (with regular hard-drives)
    - Great GPUs (V100) but poor time limits (2 hours) & slow HDDs
  - We need SSDs (SATA & M.2)
- Applied to Google for resources
  - $5000 worth of Google Cloud Credits
  - Finished them in less than a week and requested more

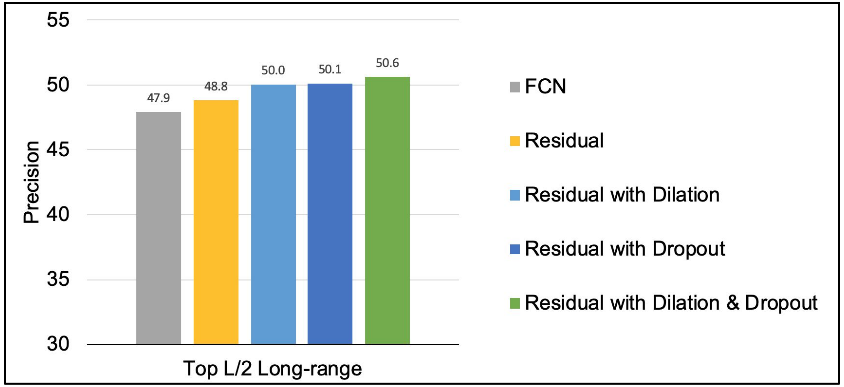


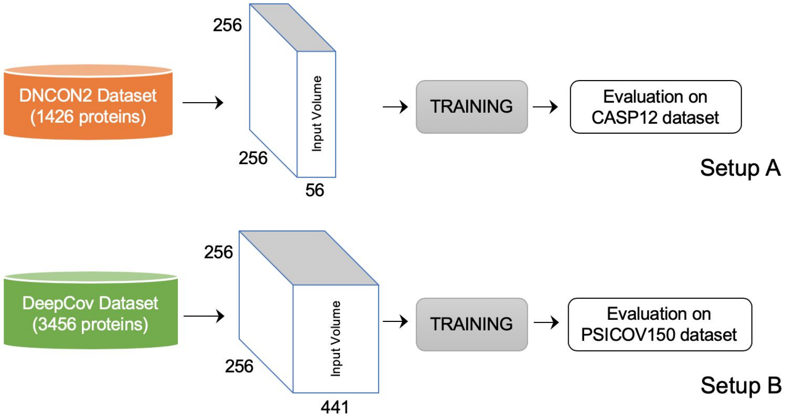
- Applied to NVIDIA for resources
  - Awarded a Quadro P6000 GPU (performs similar to V100s; extremely useful)

We Tested Various Residual Networks Architectures



Residual Networks with Dilation & Dropout Perform Best





..Here, we experiment with **two diverse datasets that use different input features**. When trained on the DeepCov dataset consisting of 3,456 proteins, using the same dataset for training and testing our method **achieves up to 6% and 15% higher precision** on the PSICOV150 protein dataset when top L/5 and L/2 long-range contacts are evaluated, respectively (L is protein length)..

DEEPCON: Protein Contact Prediction using Dilated Convolutional Neural Networks with Dropout

Badri Adhikari

doi: https://doi.org/10.1101/590455

This article is a preprint and has not been peer-reviewed [what does this mean?].

Cold Spring Harbor Laboratory bioRxiv THE PREPRINT SERVER FOR BIOLOGY

But.. Is There Room for Improvement? YES

- At the most recent CASP conference:

  "It was good to see Google DeepMind win this time..
  I was sick of seeing Rosetta win since almost two decades.."
  - a senior scientist at the conference

- Google plans to continue its 'fundamental' research
- We are still far from end-to-end deep learning

How Does Deep Learning Deliver Improved Performance?

Can We Learn to Predict Contacts WITHOUT 'True' Contacts?

MSEIITFPQQTVVYPEINVKTLSQAVKNIWRLSHQQKSGIEIIQEKTLRISLYSRDLDEA

Search

Protein
Sequence DB

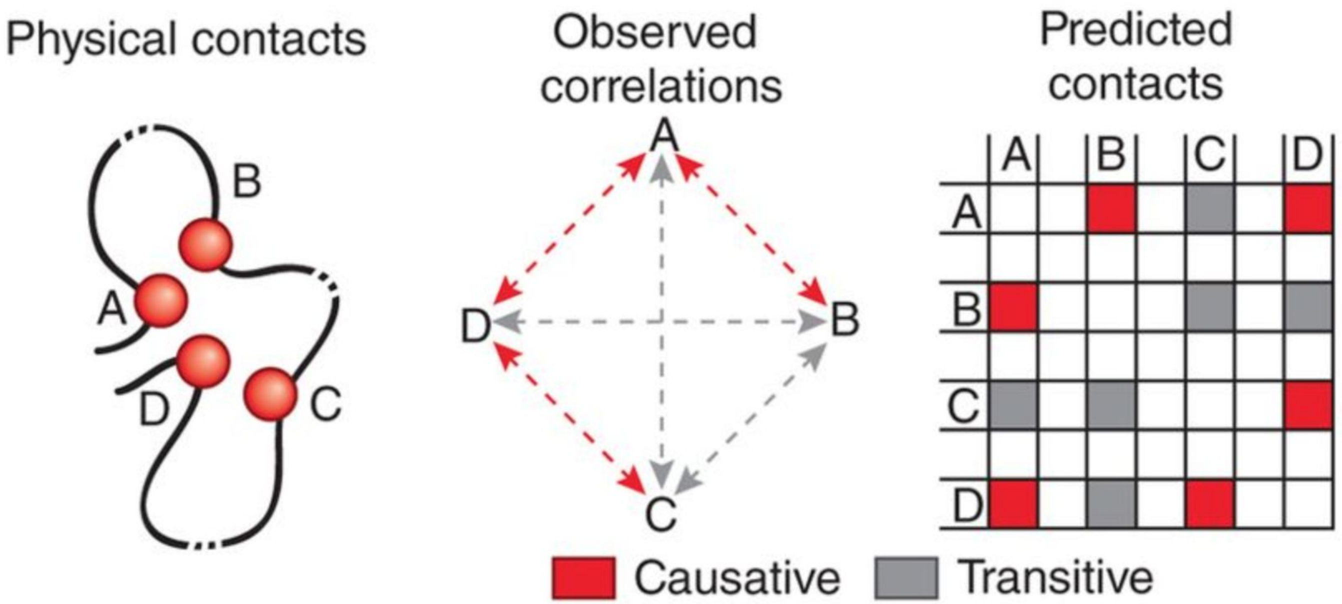
Homologous
sequences

```
----NTLSQKENMYPEINIKAMNQAVNTIWLLAQRQTSGIEIINDKVKRISLYSREFDE-
-------------LTPPDTEVMKTARQNILHVTALKLDFLPVMKEKMRPLQDALISADK-
------------ILTPPDNEIMNDARQNILQASALKLDFLPVMKEKMLPLQTALKRADKV
MVVRNSAKAIAEHSDDMAQINLKLEGNRLLETLQEQIDSITLRSAALESTMGEITA----
AGIARLGKLLDKVSSALTYLDLIEERDRLRLRYNALLEESRTAHQEEKATAAKLDELT--
-------------LTPPDNEVMETARQNILQVTALKLDFLPVMKEKMLPLQAALMSADKV
DVVRRSAQAIAAHSDDVAQINLTLESNRLLETLQQQIDSMTQRSATQESAMALIAA----
AGIARLGKLLDKVSSALNYLDLIEERDRLRYRYNALLDDSRTALREAQATREKLDELTAL
```

Covariance / Coevolution

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

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



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

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

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

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** ▾

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

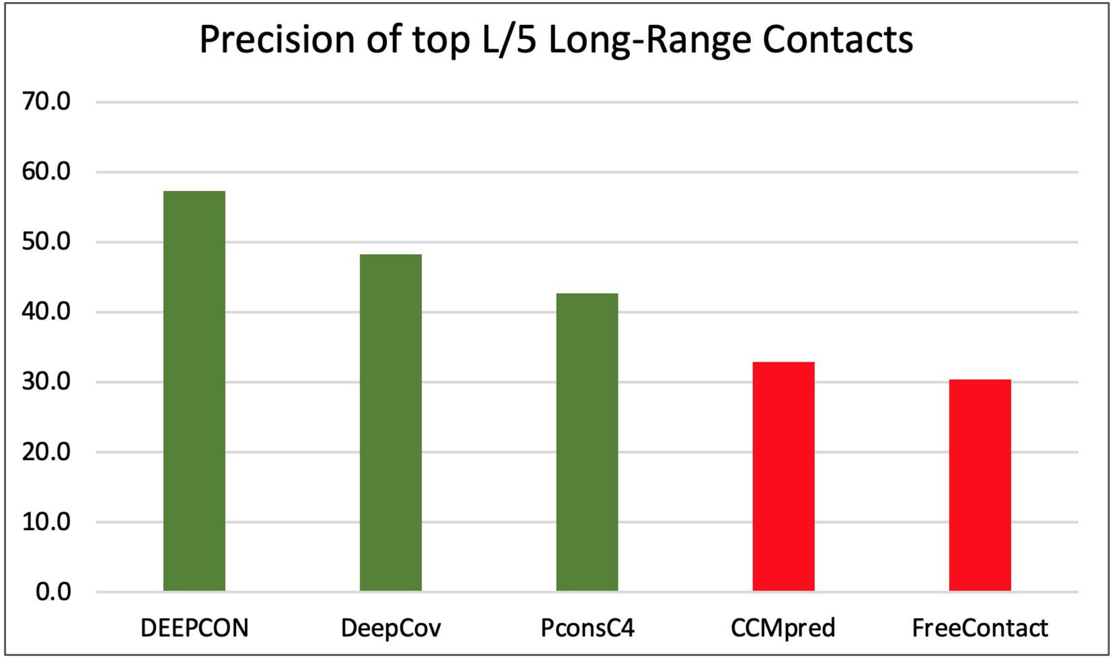
Stefan Seemayer, Markus Gruber, Johannes Söding Author Notes

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

doi: https://doi.org/10.1101/590455

What Do We Learn

Conclusions

1) Groups who were good at exploring ‘new flavors’ did well
   - Learn various deep learning methods, even when you don’t see a direct fit to your problem

2) Balanced efforts of ML experts and domain experts brought success
   - Do you have enough ML ‘breadth’ or team?

3) When end-to-end is not possible, correct feature engineering becomes important
   - Is feature engineering solved for your problem? If not, focus your research here!
   - For example, for standard images, we don’t need feature engineering

4) Using ‘a lot of data’ and ‘deep architectures’ improves performance
   - Have you tried using “all the data” and “a large architecture”?

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