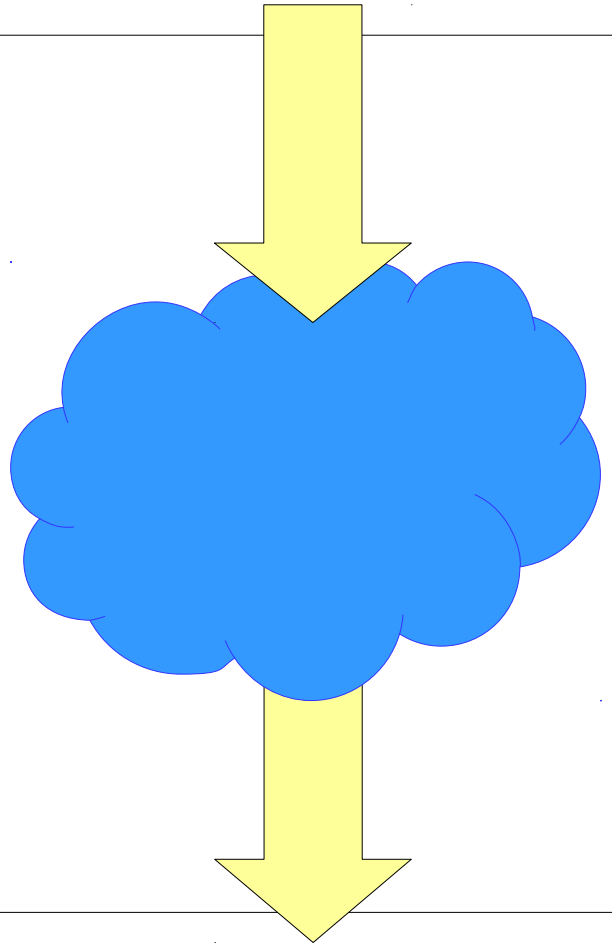
Beyond end-to-end learning

David Menéndez Hurtado

Inputs



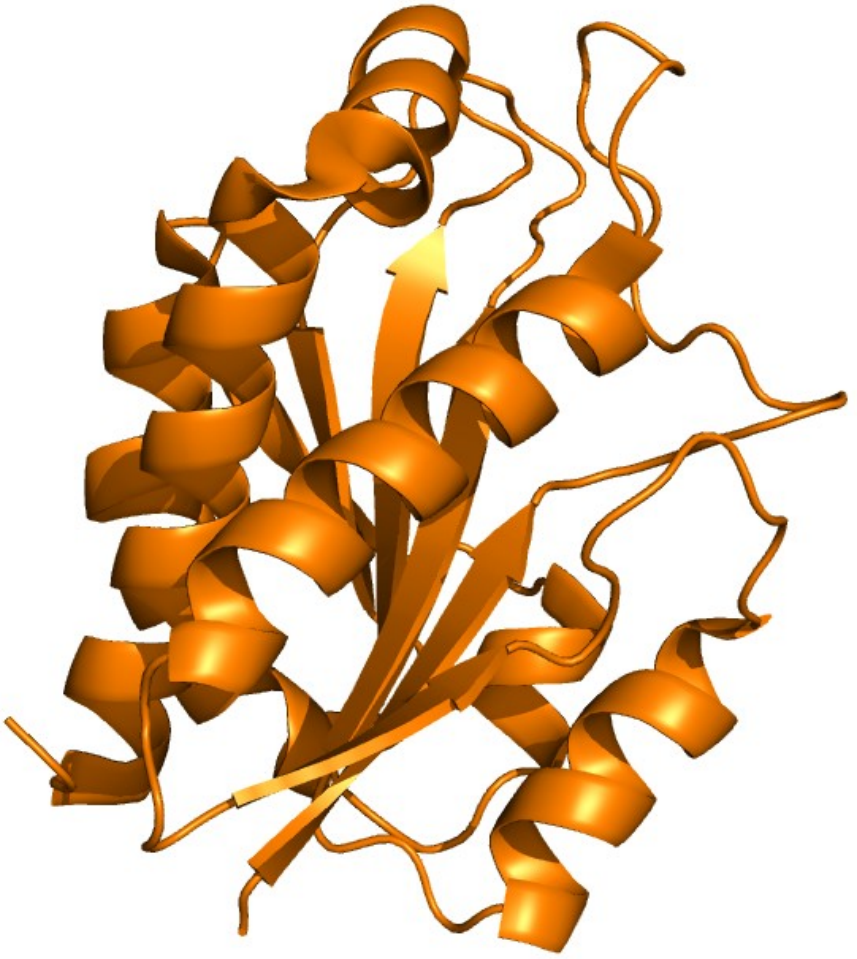
Outputs

Background I

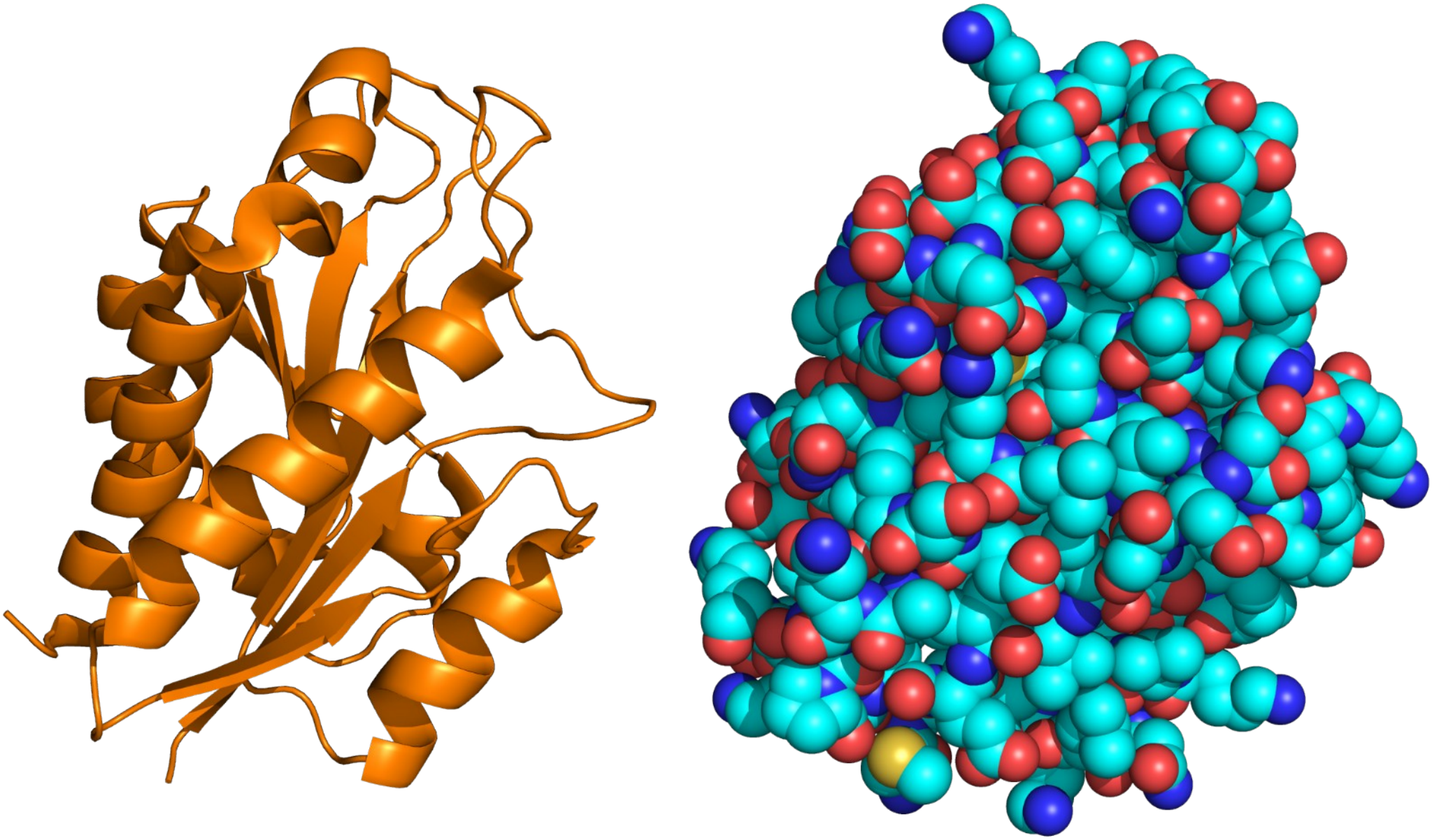
Protein structures and assessment of models

Protein structure prediction

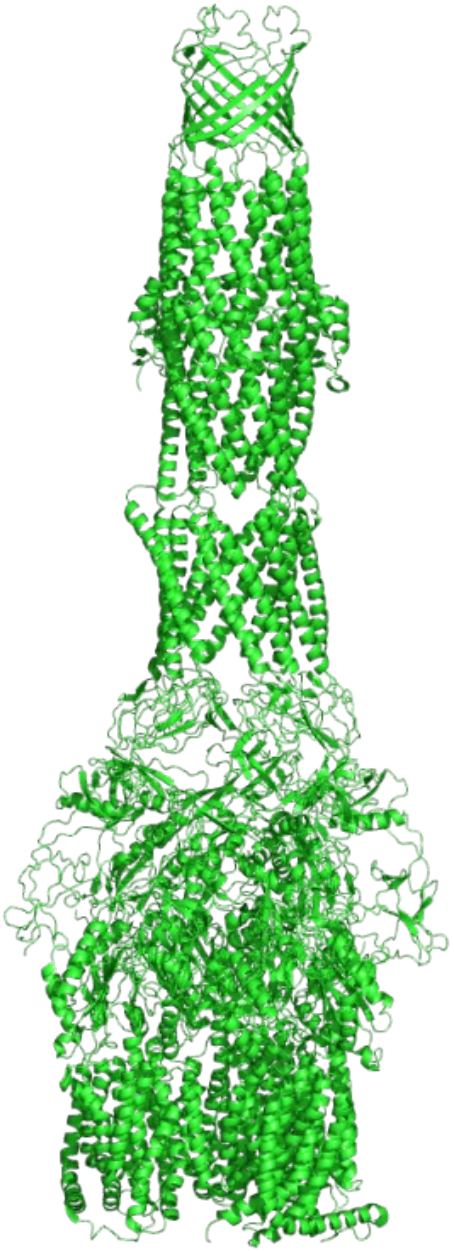
```
> N-acetylmuramoyl-L-alanine amidase
MKVVVIDAGHGAKDSGAVGISRKNYEKTFNLAMALKVESI
LKQNPKLEVVLTRSDDTFLELKQRVKVAENLKANVFVSIH
ANSSGSSASNGTETYYQRSASKAFANVMHKYFAPATGLTD
RGIRYGNFHVIRETTMPAVLLEVGYLSNAKEEATLFDEDF
QNRVAQGIADGITEYLDVK
```

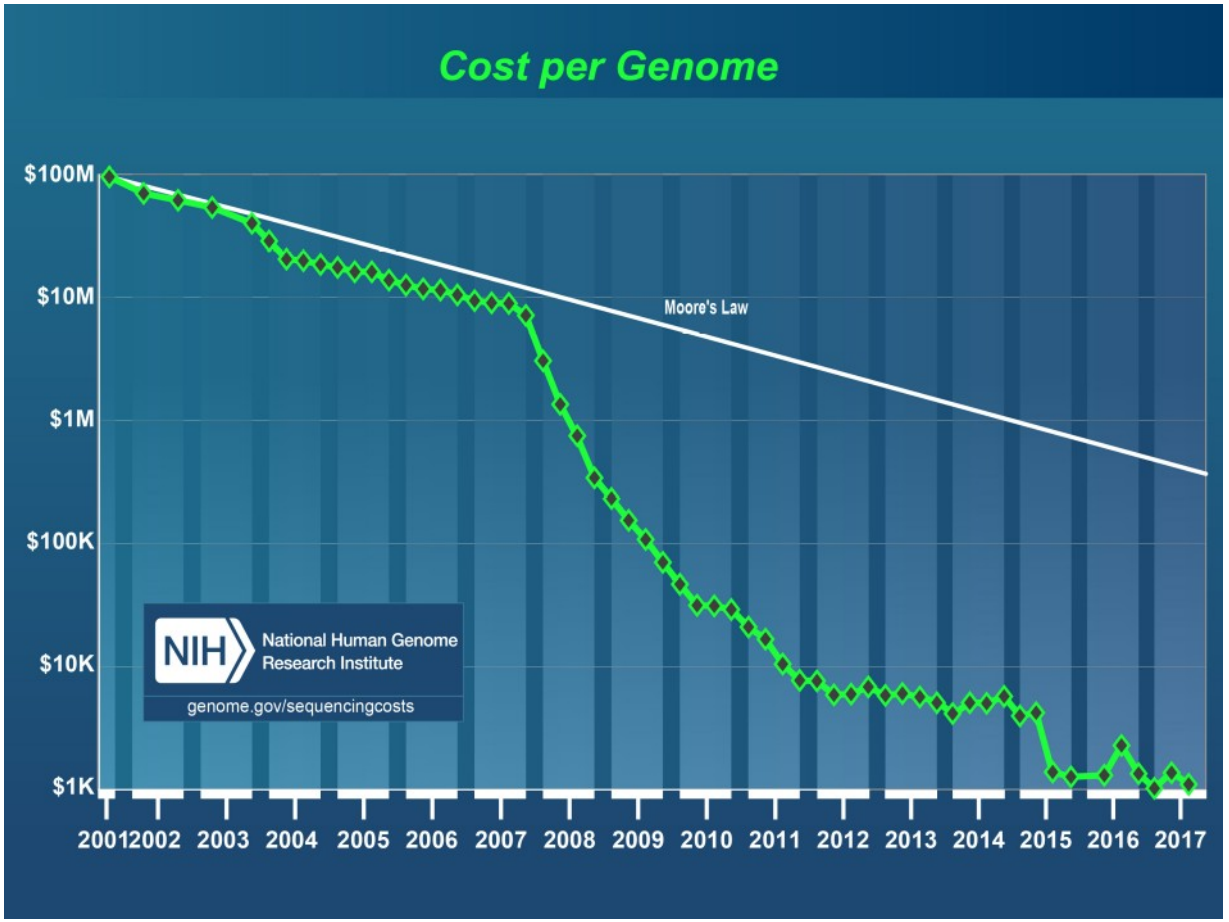


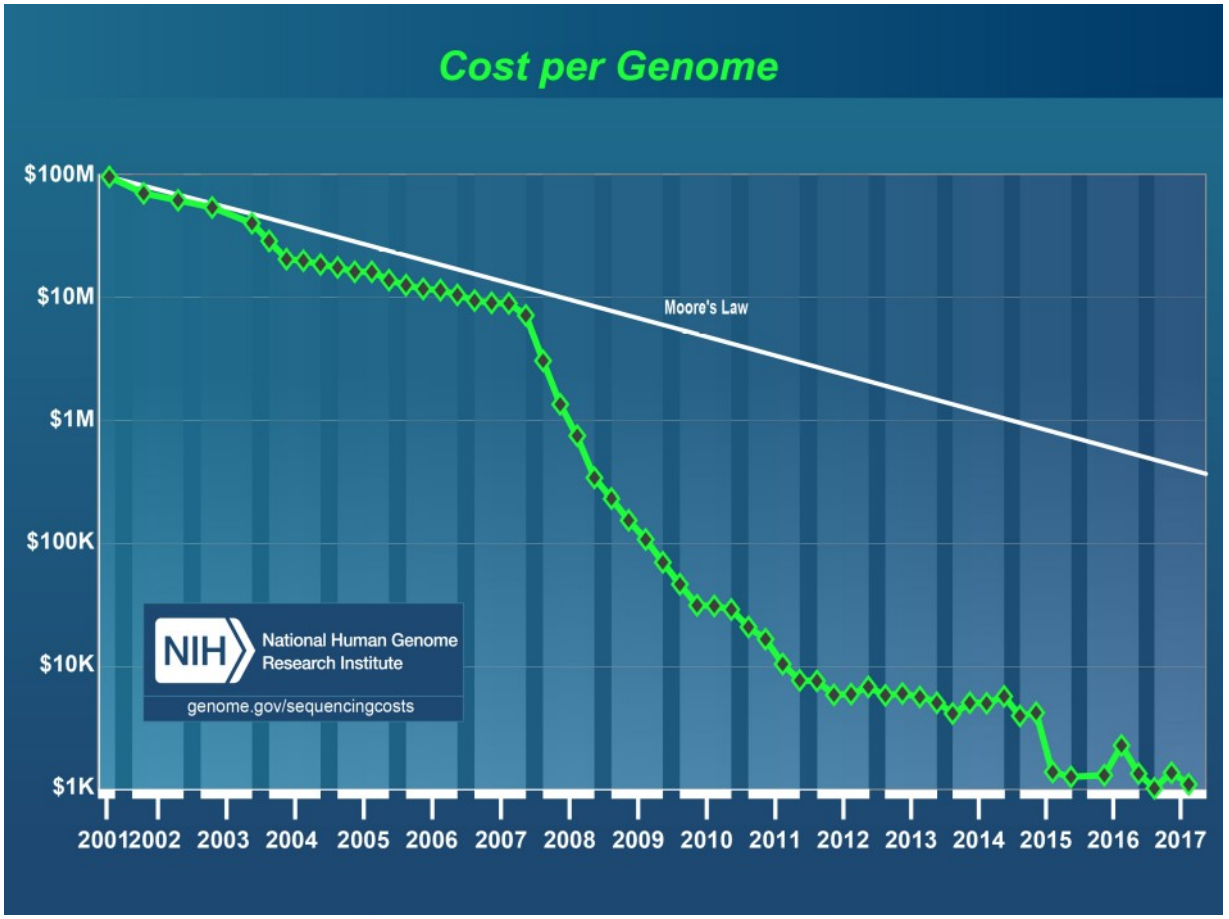
Amiga photo: Wikimedia Commons



Why?



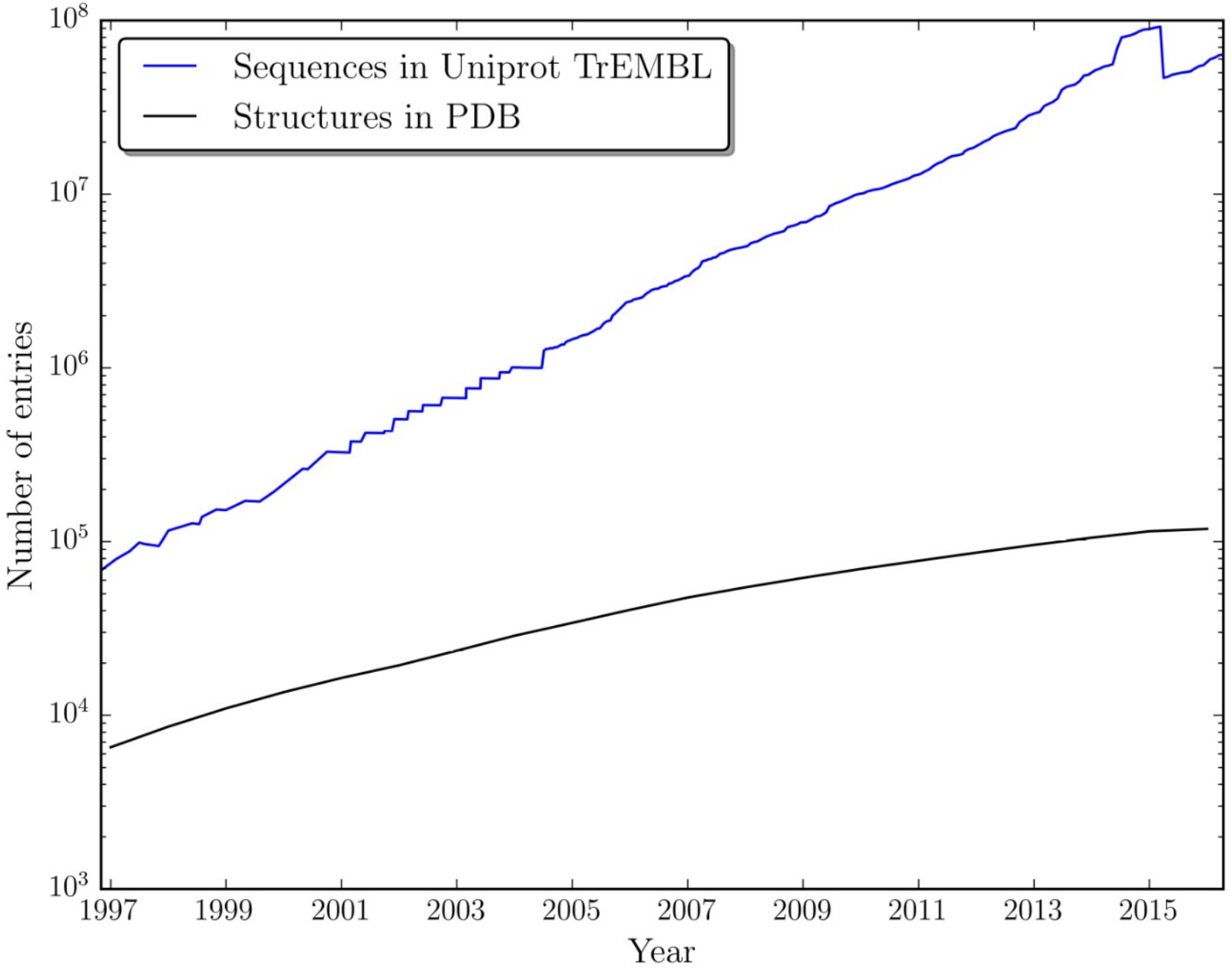

Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017



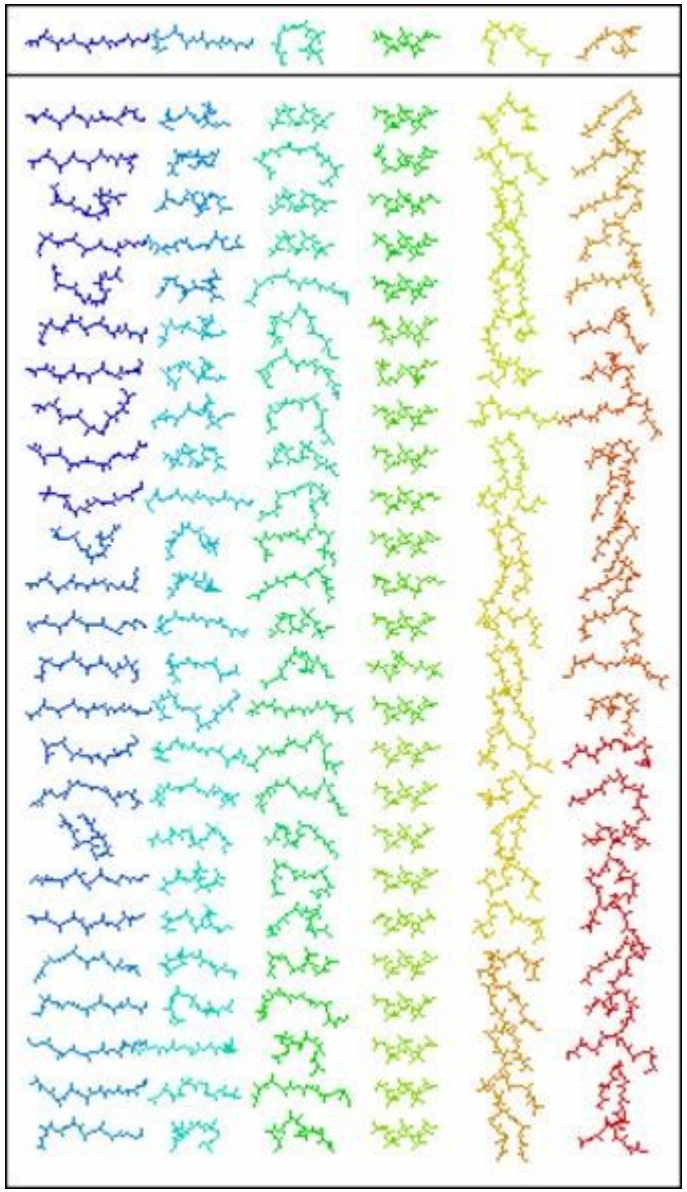
NOVEL DRUG TARGET	AVERAGE COST	SUCCESS RATE
Soluble bacterial targets	$140,000	35%
Soluble human protein (eg kinases, proteases, NHRs)	$450,000	35%
Bacterial membrane proteins	$1.5 million	10%
Human membrane proteins	$2.5 million	10%

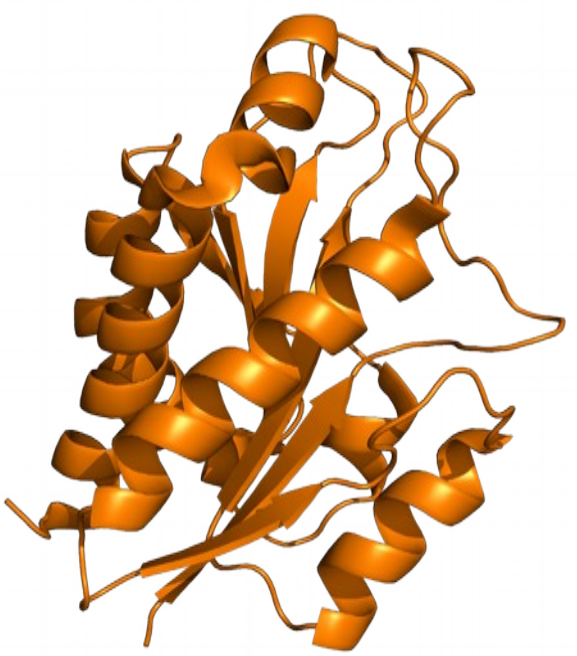
Raymond C. Stevens. The cost and value of three-dimensional protein structure. (2003)

Why?

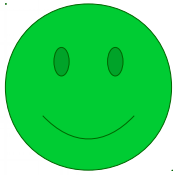

Sequences in Uniprot TrEMBL
Structures in PDB
Number of entries
Year
10^8
10^7
10^6
10^5
10^4
10^3
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015

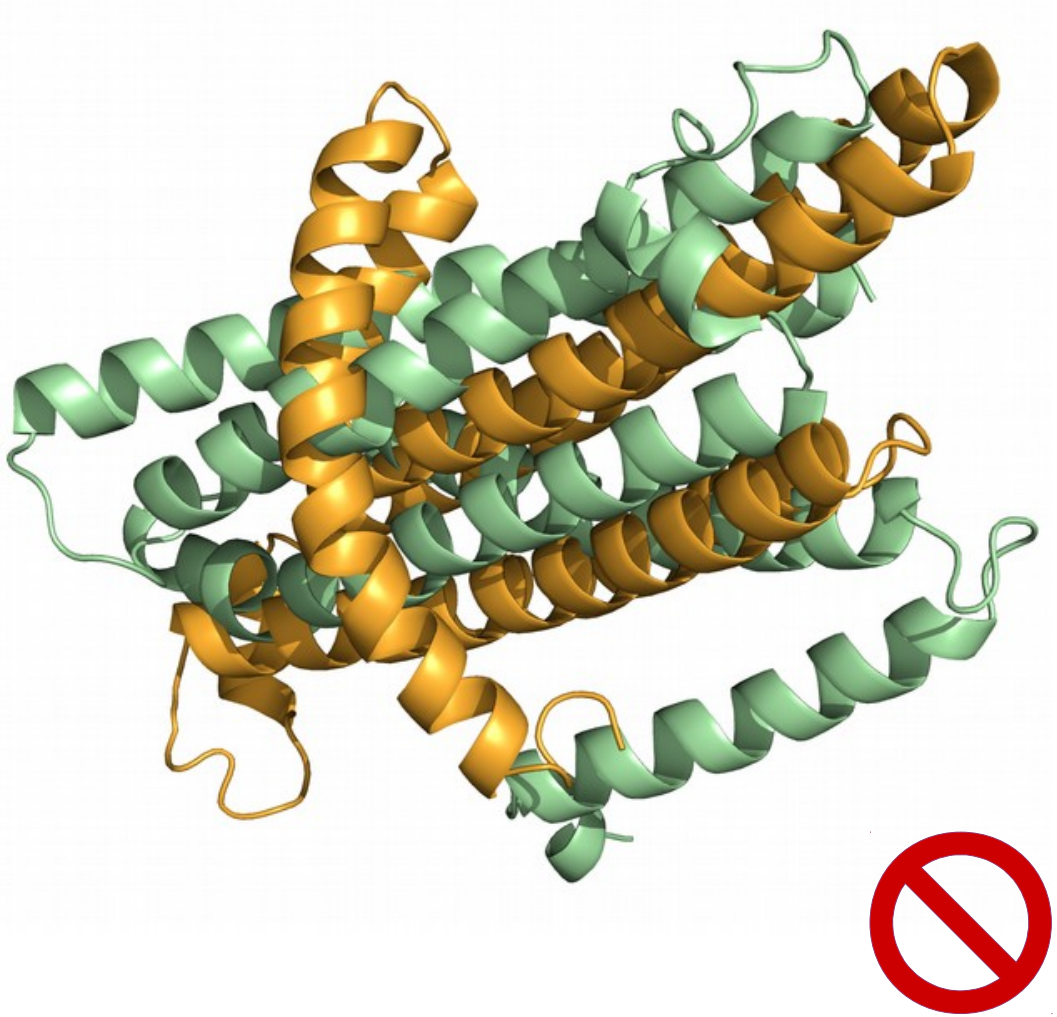
Solution:



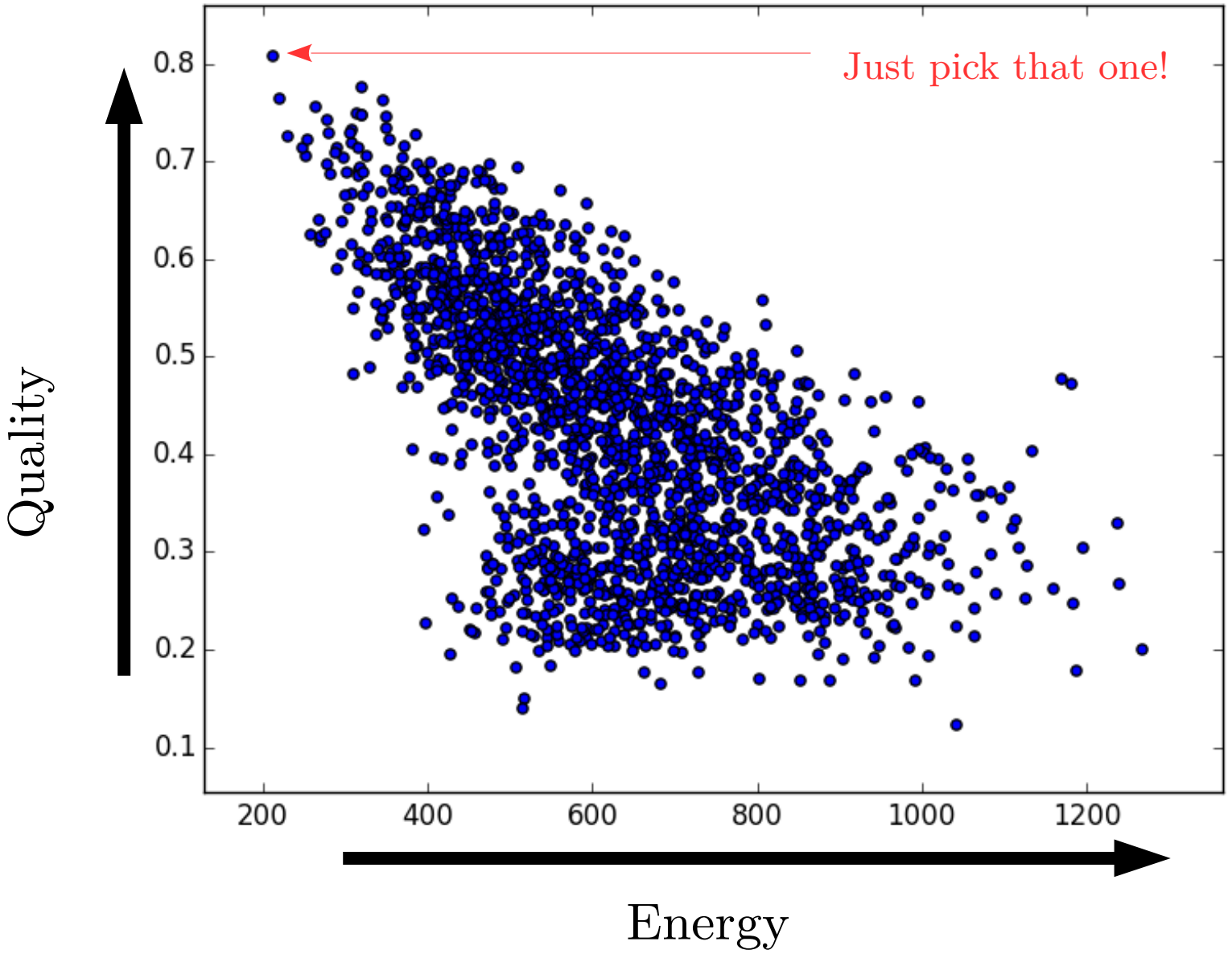


Model Quality Assessment

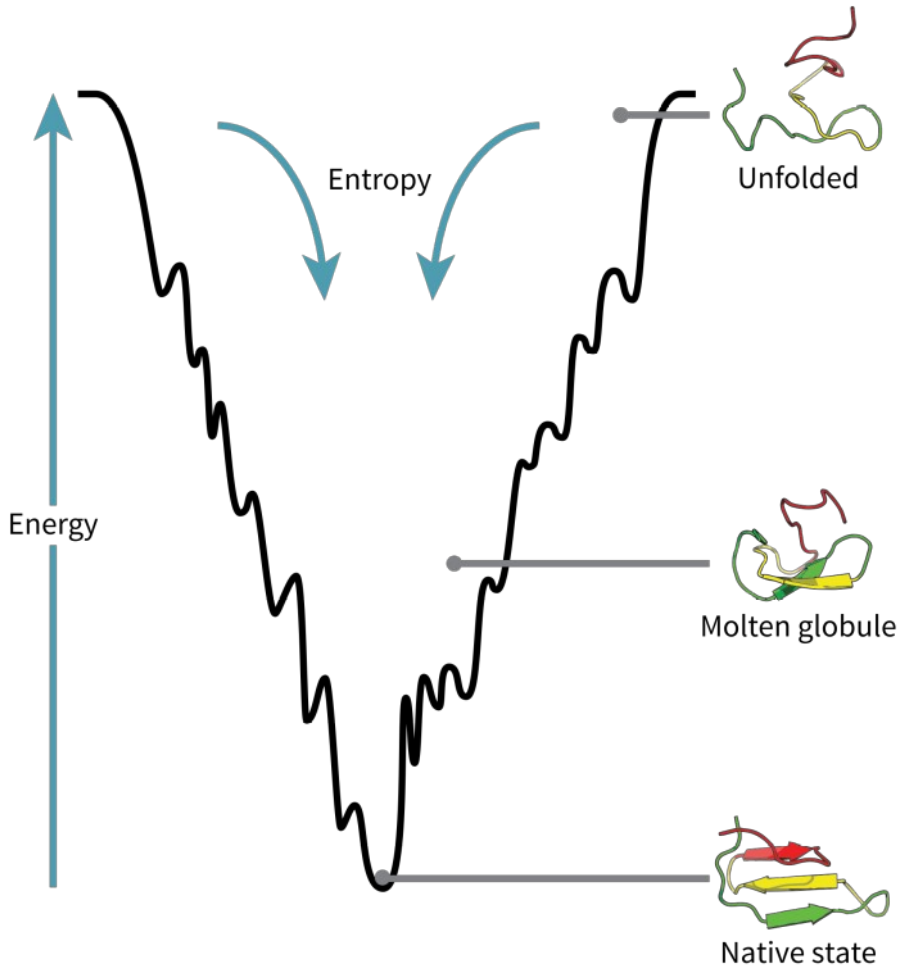




Why is this a thing?

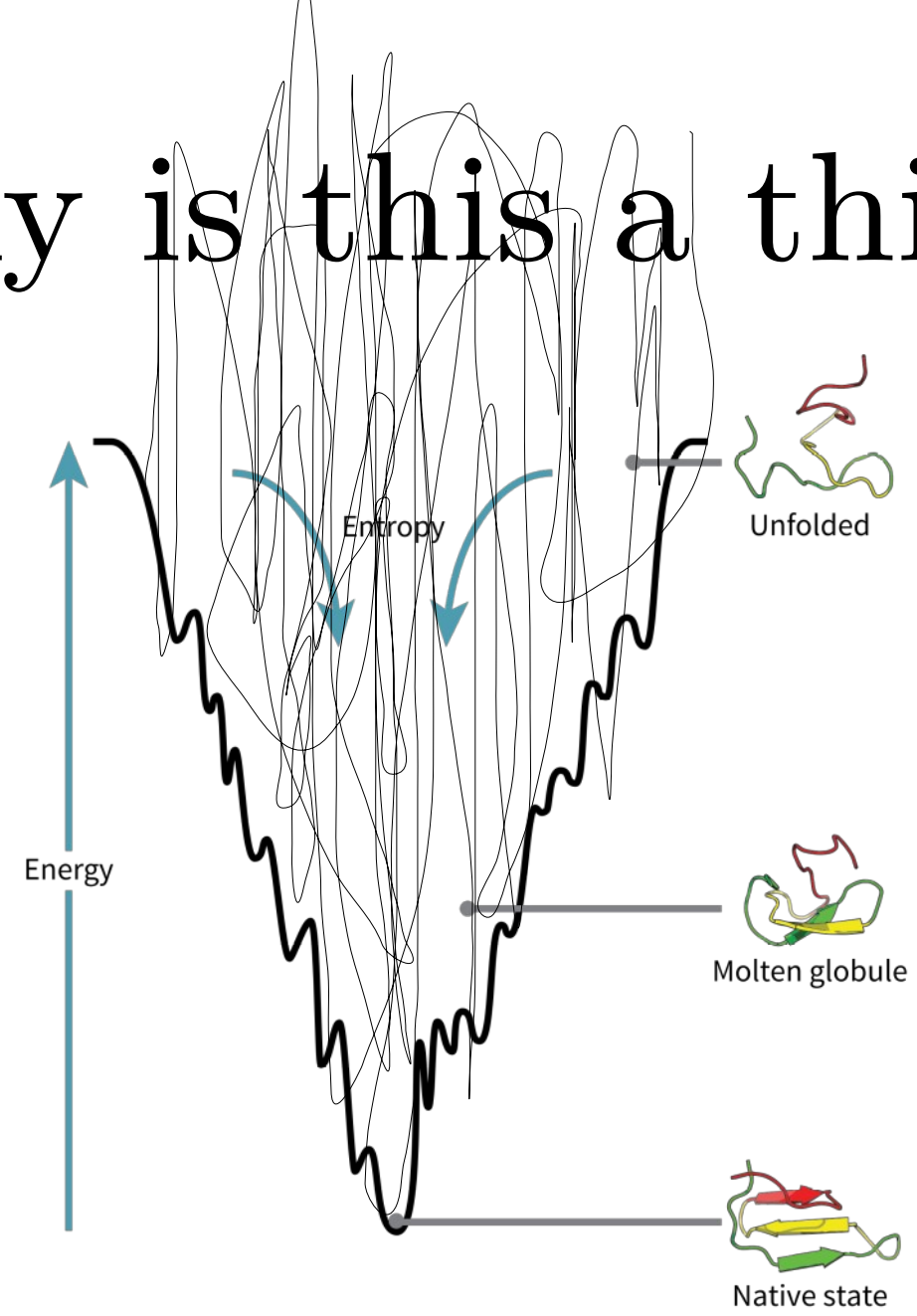


Why is this a thing?



Wikimedia Commons.

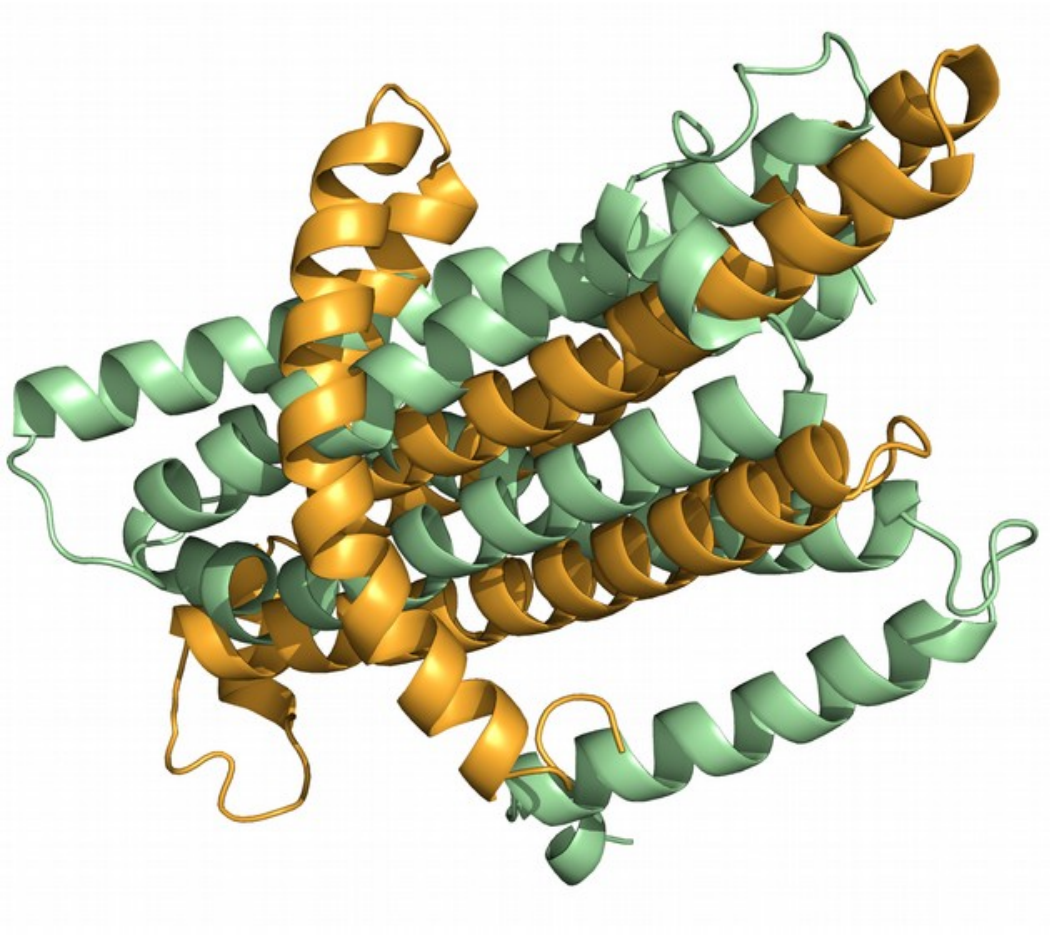
Why is this a thing?



Wikimedia Commons.

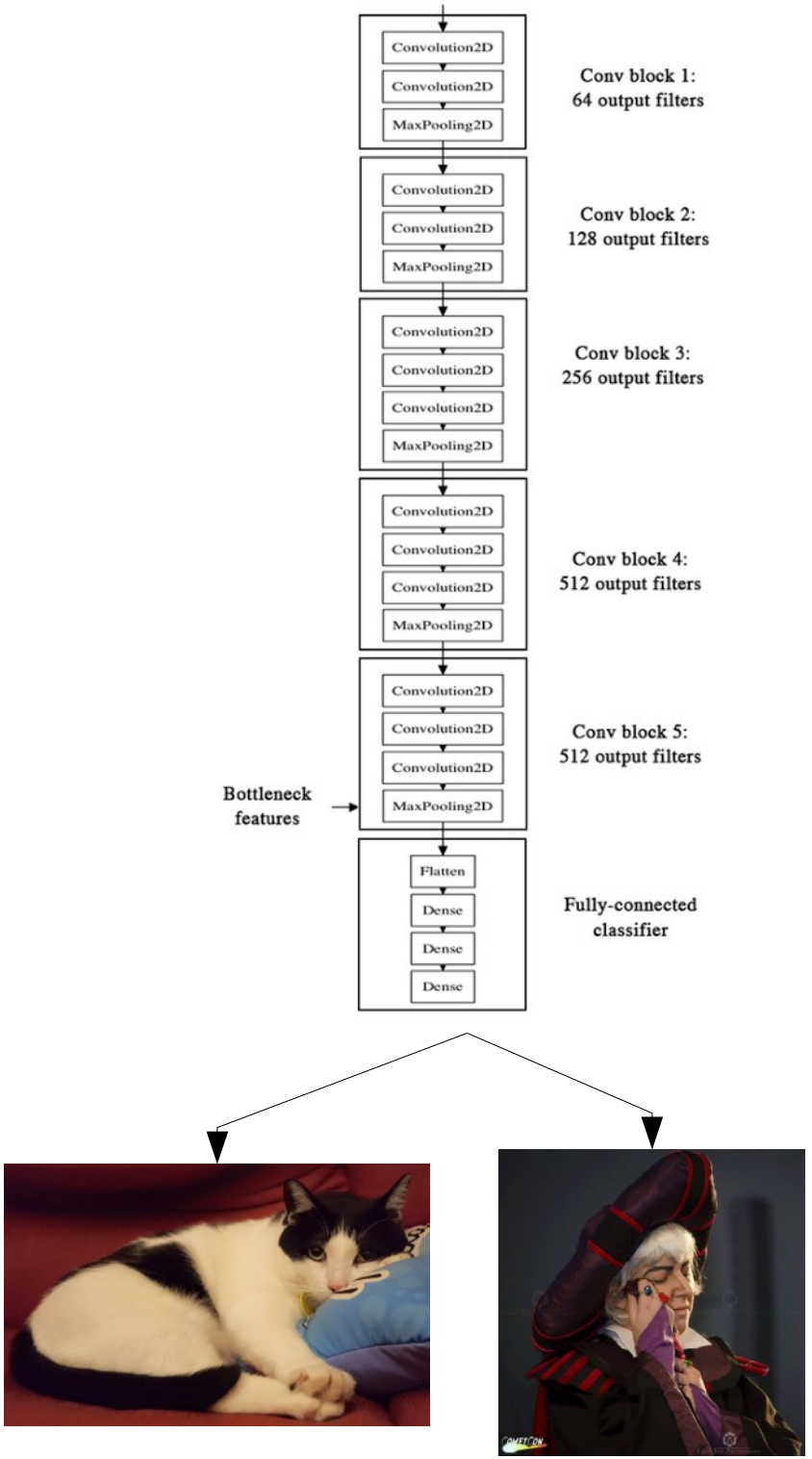
Why is this a thing?

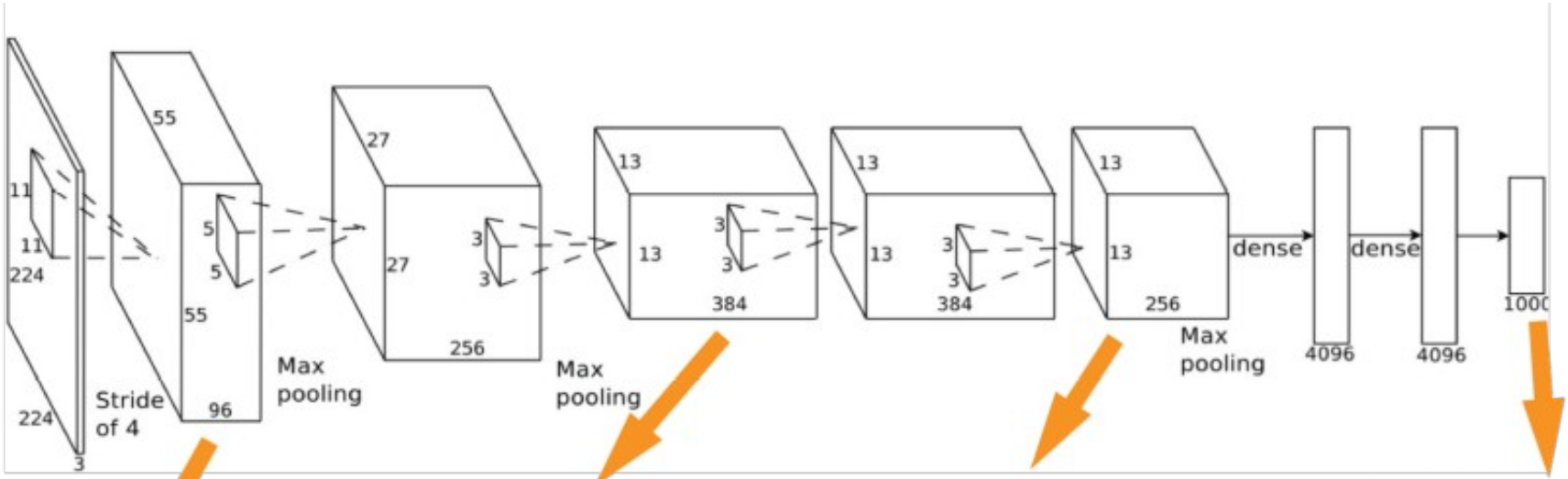
Global failures:

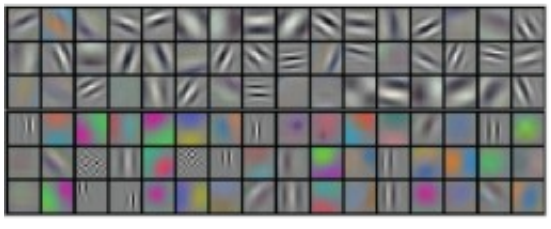


Background II

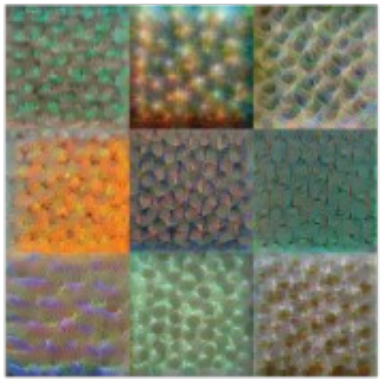
Deep transfer learning


Convolution2D
Convolution2D
MaxPooling2D
Conv block 1:
64 output filters
Convolution2D
Convolution2D
MaxPooling2D
Conv block 2:
128 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 3:
256 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 4:
512 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 5:
512 output filters
Bottleneck
features
Flatten
Dense
Dense
Dense
Fully-connected
classifier





Conv 1: Edge+Blob



Conv 3: Texture

Numerical

Data-driven

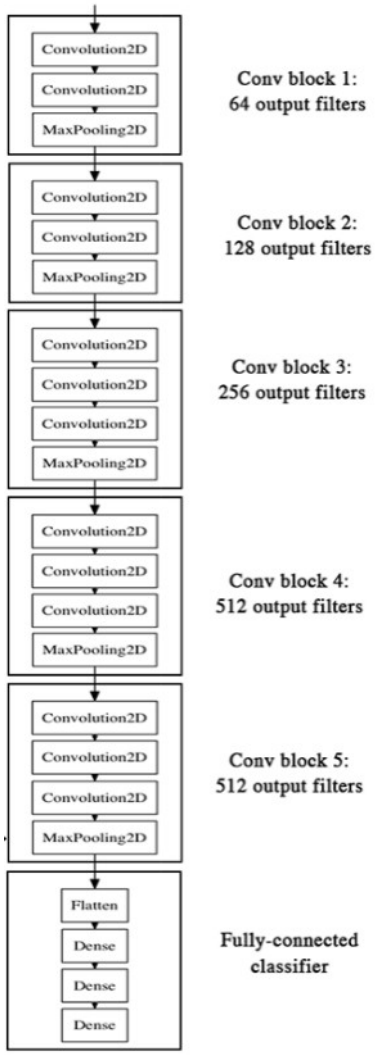
Conv 5: Object Parts

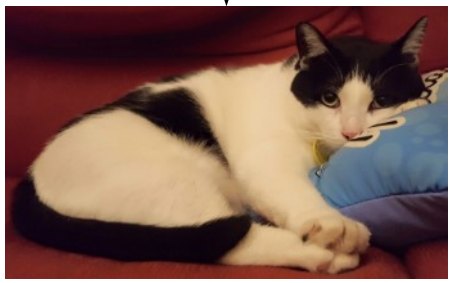
Fc8: Object Classes

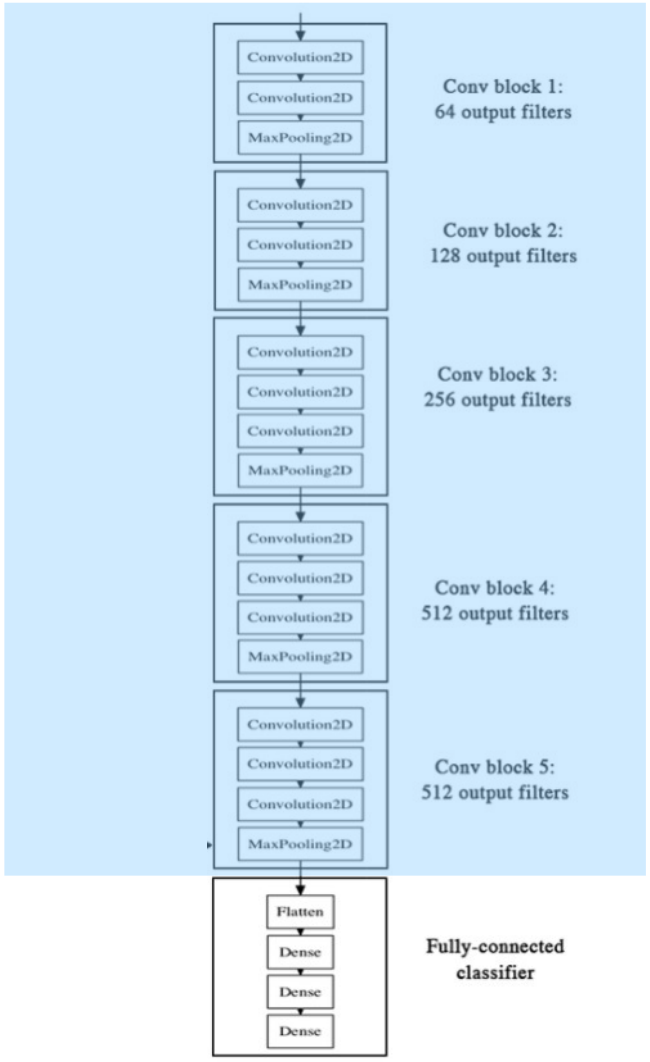
https://medium.com/nanonets/

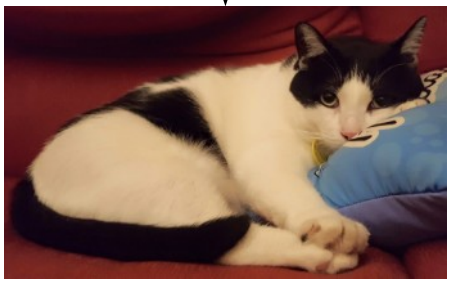
A more daunting task

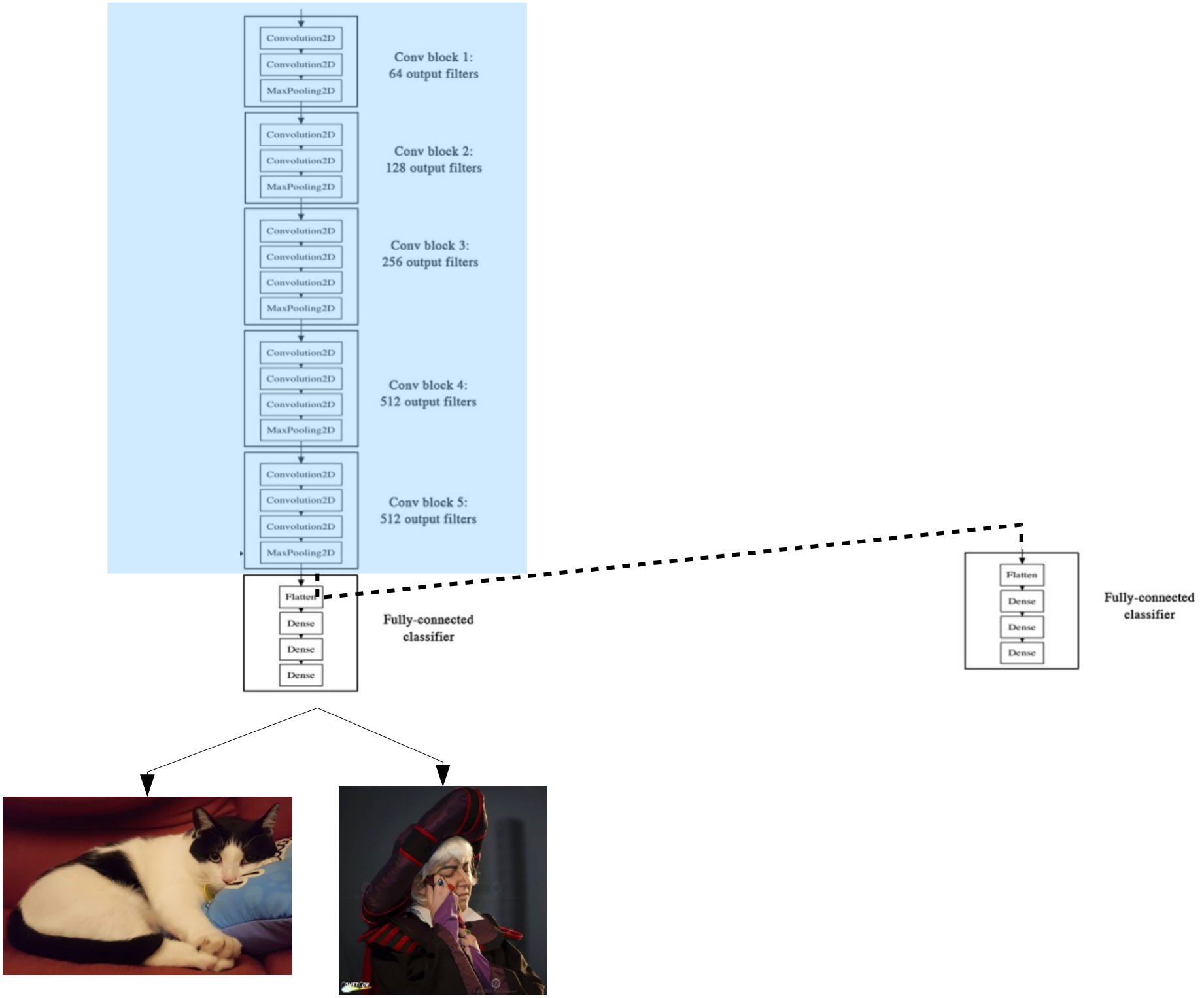
Wikimedia Commons.

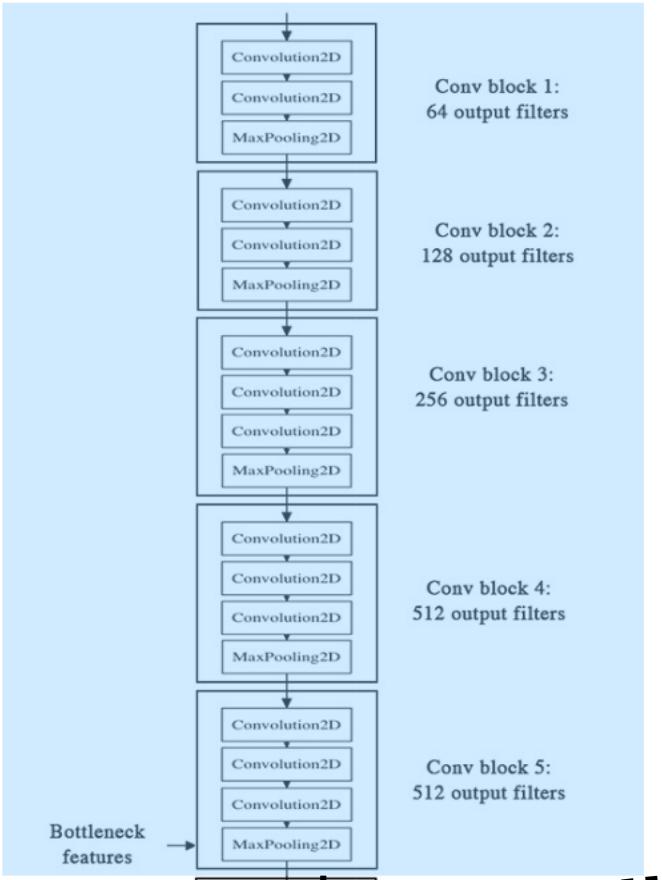

Convolution2D
Convolution2D
MaxPooling2D
Conv block 1:
64 output filters
Convolution2D
Convolution2D
MaxPooling2D
Conv block 2:
128 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 3:
256 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 4:
512 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 5:
512 output filters
Flatten
Dense
Dense
Dense
Fully-connected
classifier

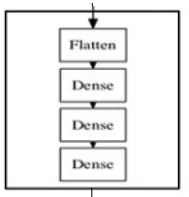



Convolution2D
Convolution2D
MaxPooling2D
Conv block 1:
64 output filters
Convolution2D
Convolution2D
MaxPooling2D
Conv block 2:
128 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 3:
256 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 4:
512 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 5:
512 output filters
Flatten
Dense
Dense
Dense
Fully-connected
classifier

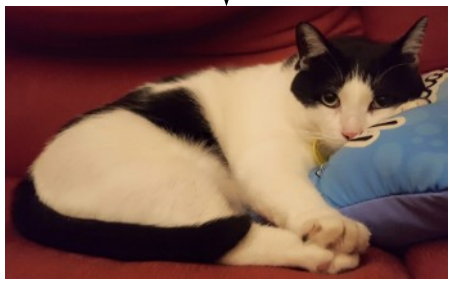


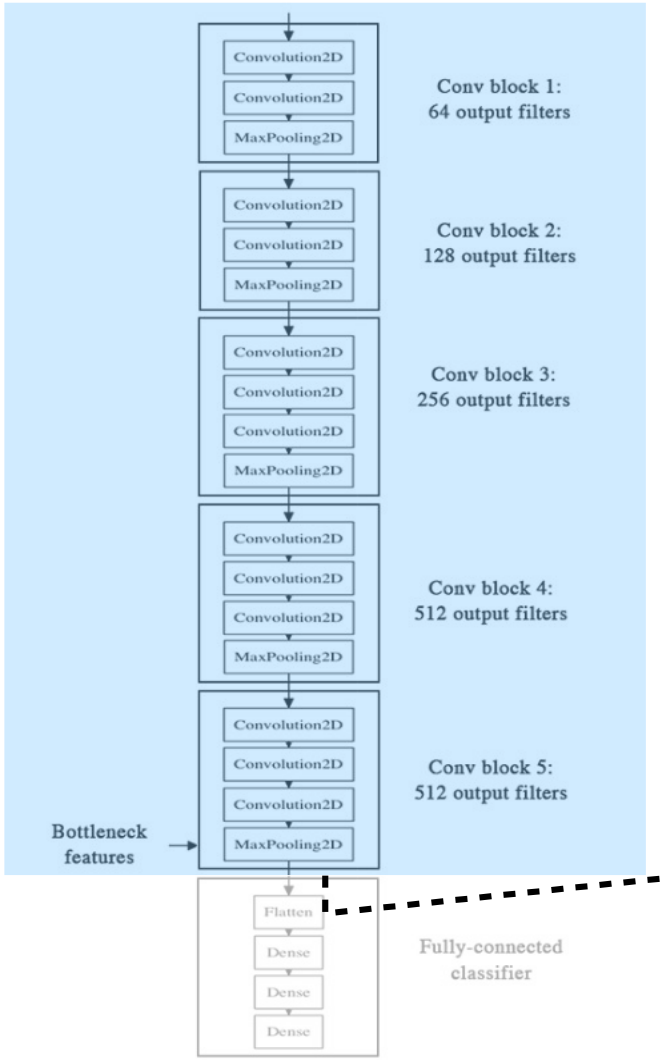

Convolution2D
Convolution2D
MaxPooling2D
Conv block 1:
64 output filters
Convolution2D
Convolution2D
MaxPooling2D
Conv block 2:
128 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 3:
256 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 4:
512 output filters
Convolution2D
Convolution2D
Convolution2D
MaxPooling2D
Conv block 5:
512 output filters
Flatten
Dense
Dense
Dense
Fully-connected
classifier
Flatten
Dense
Dense
Dense
Fully-connected
classifier

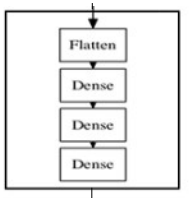




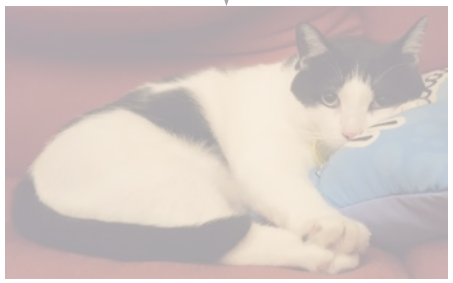
Fully-connected
classifier







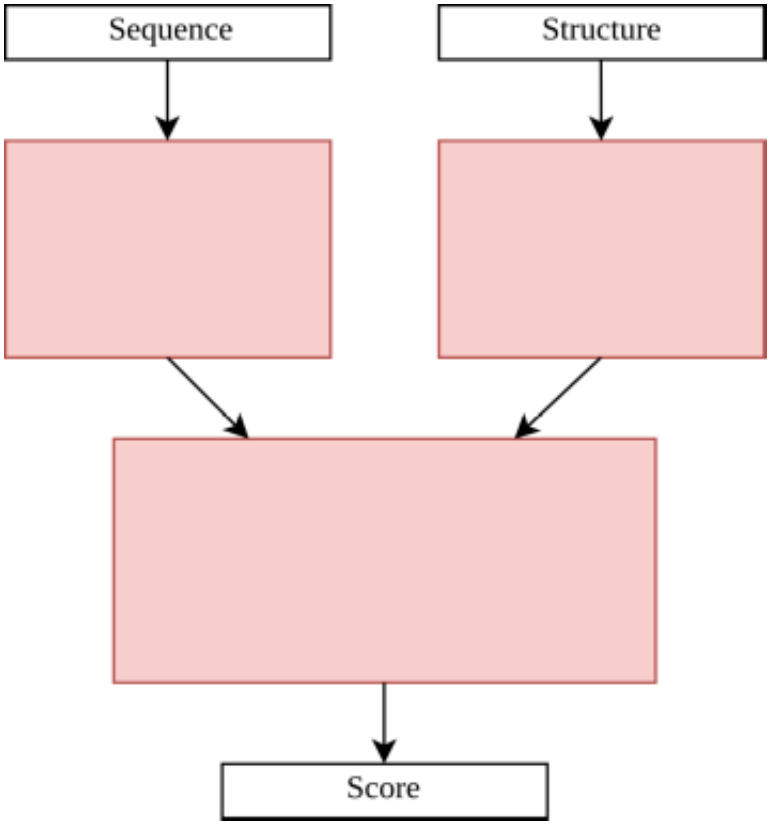
Fully-connected classifier

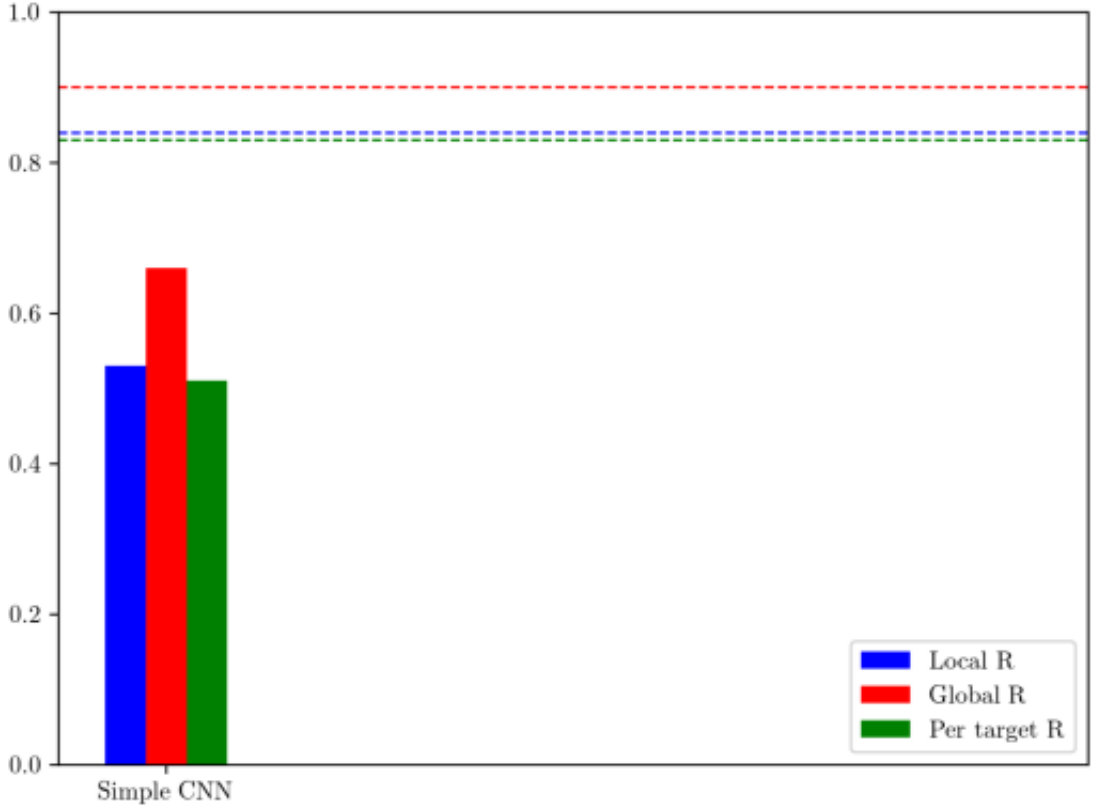


The body

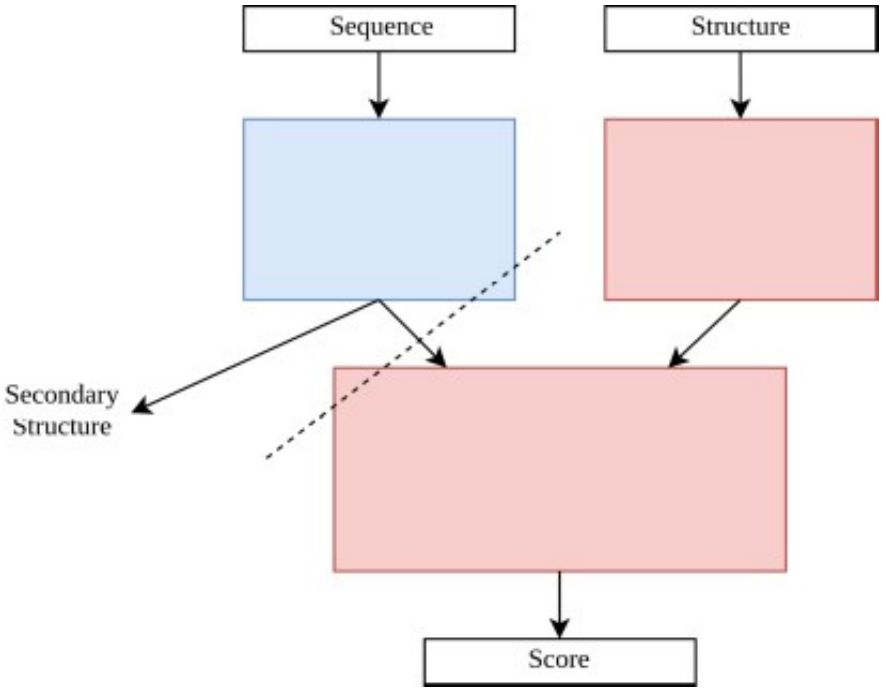
Where everything comes together

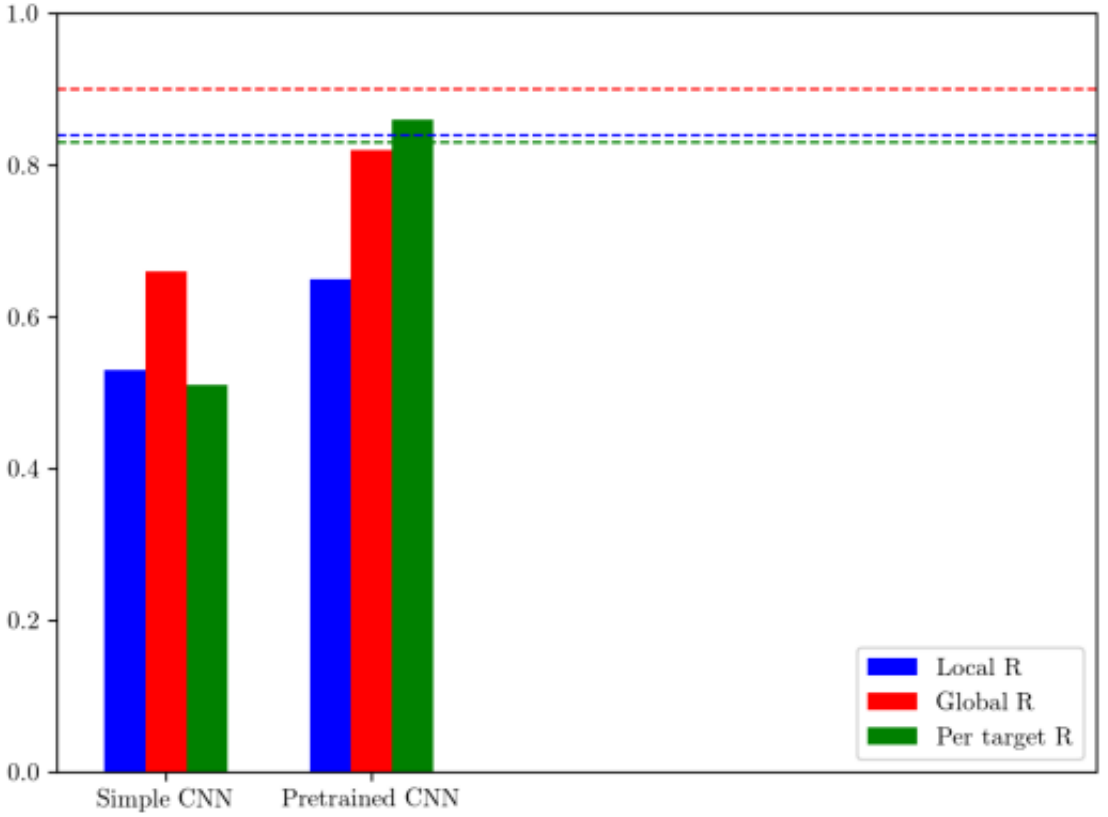
Architecture: simple CNN



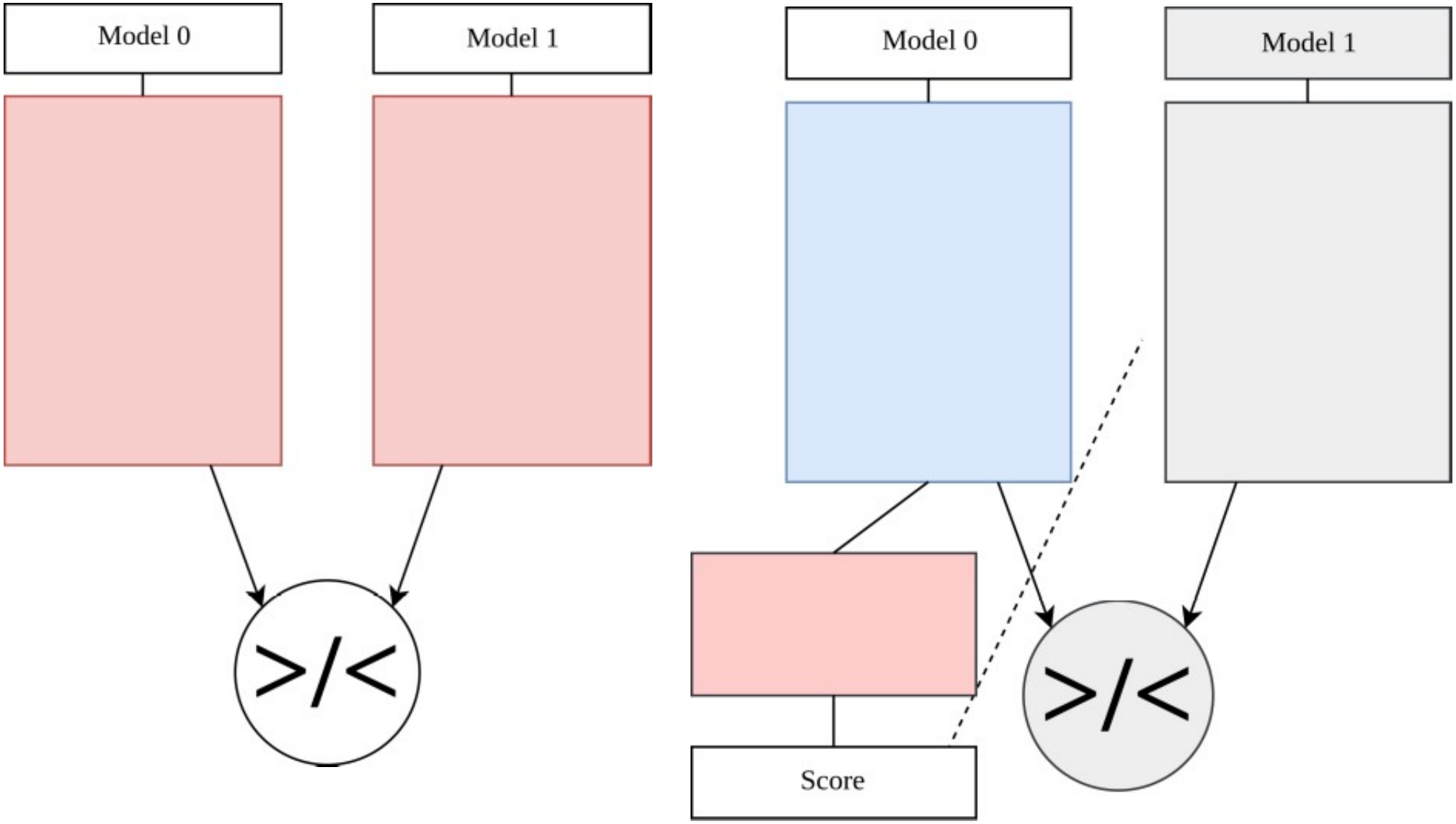


Architecture: pretrained CNN

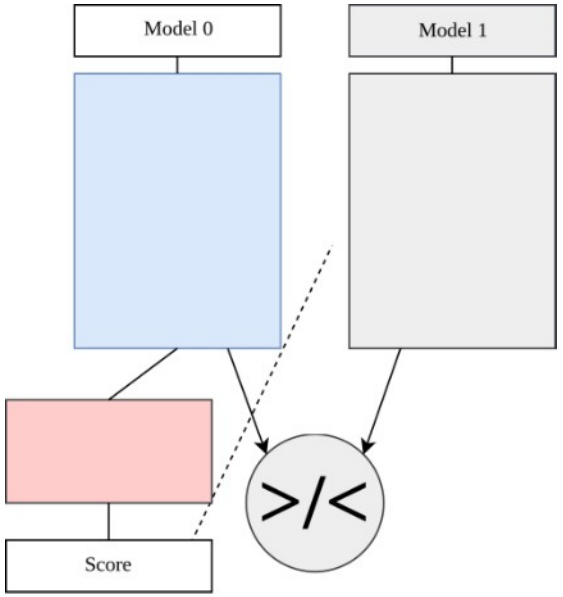


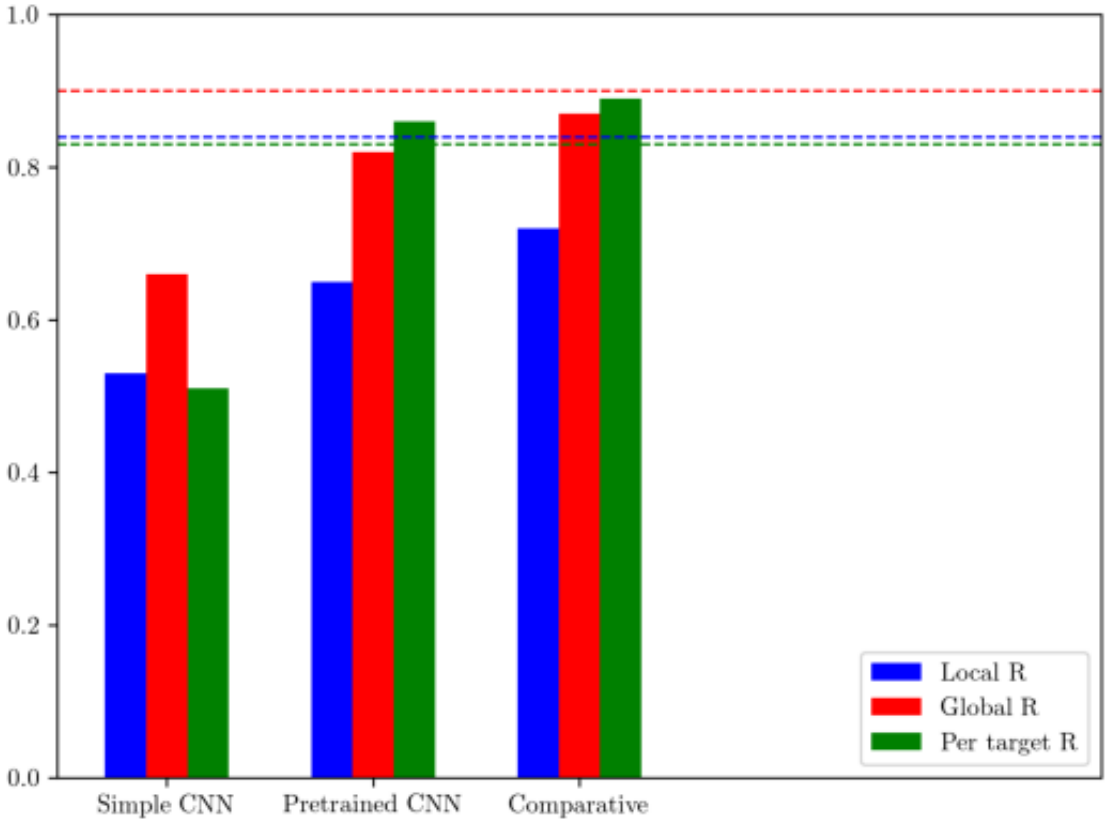


Architecture: comparative

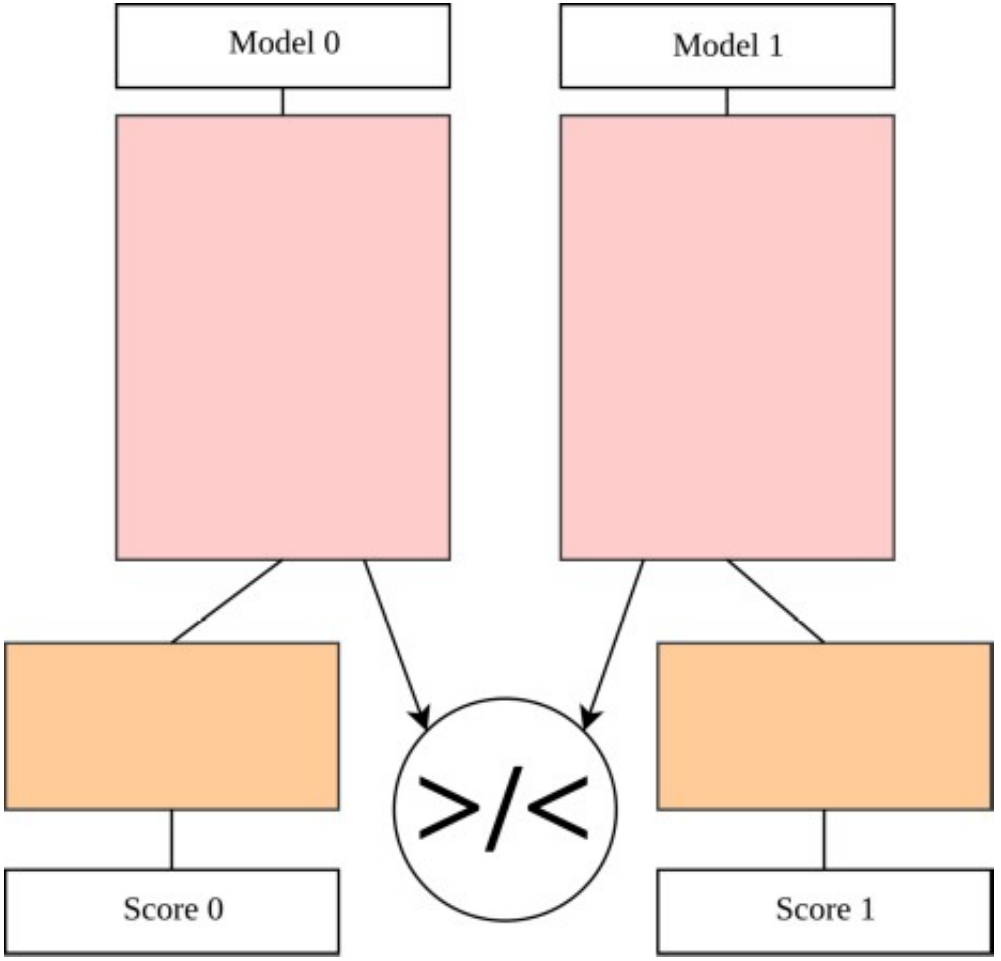


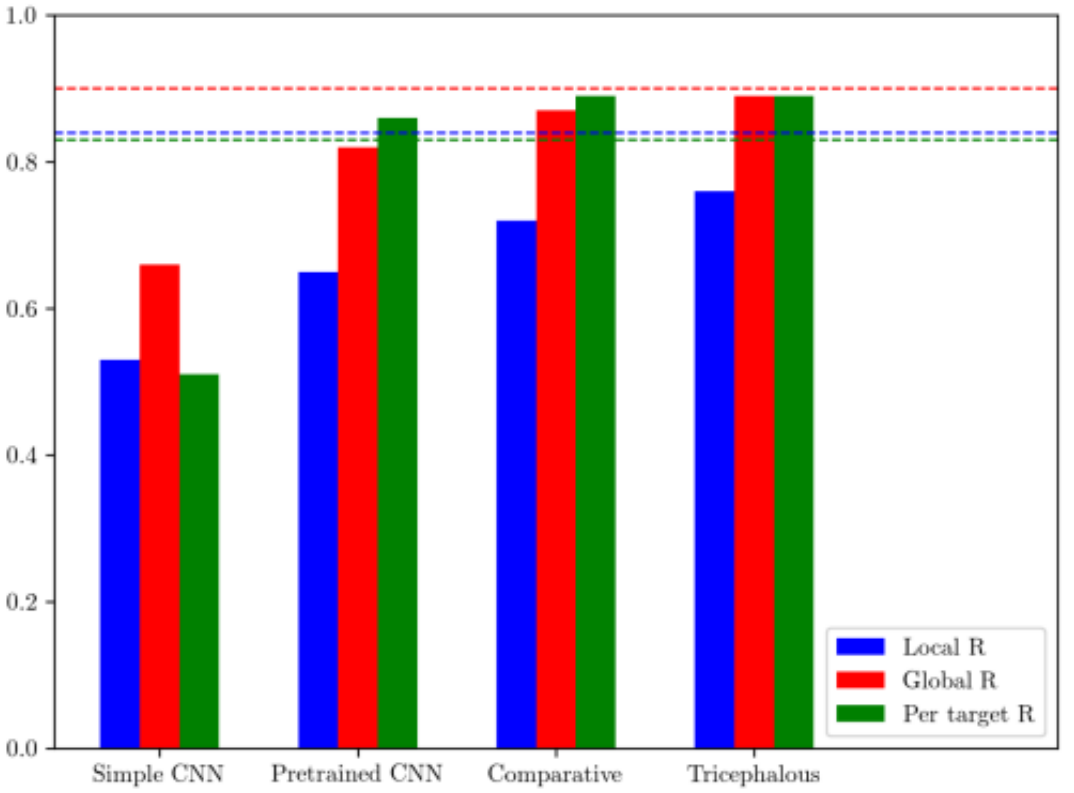
Architecture: comparative



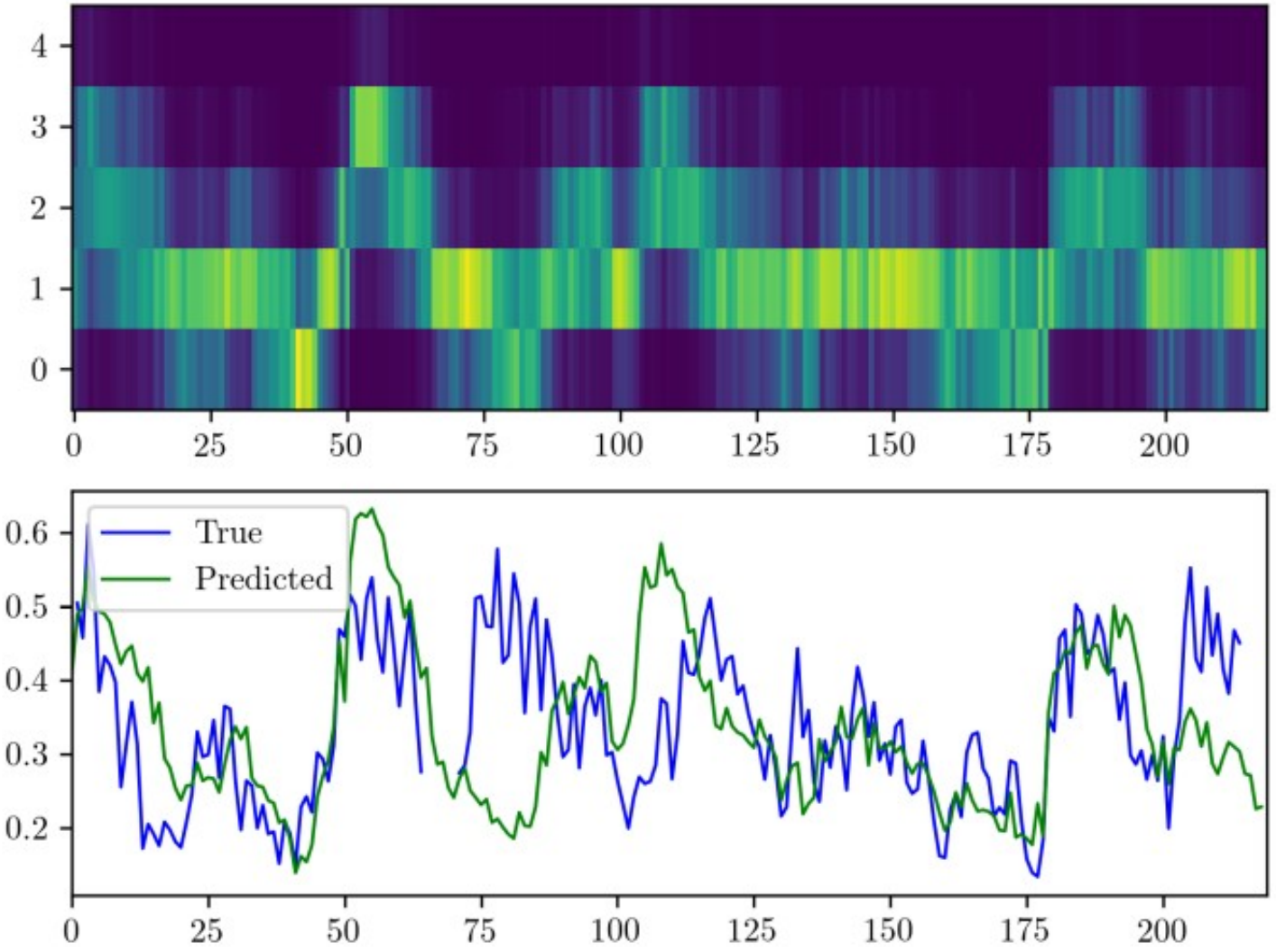


Architecture: tricephalous

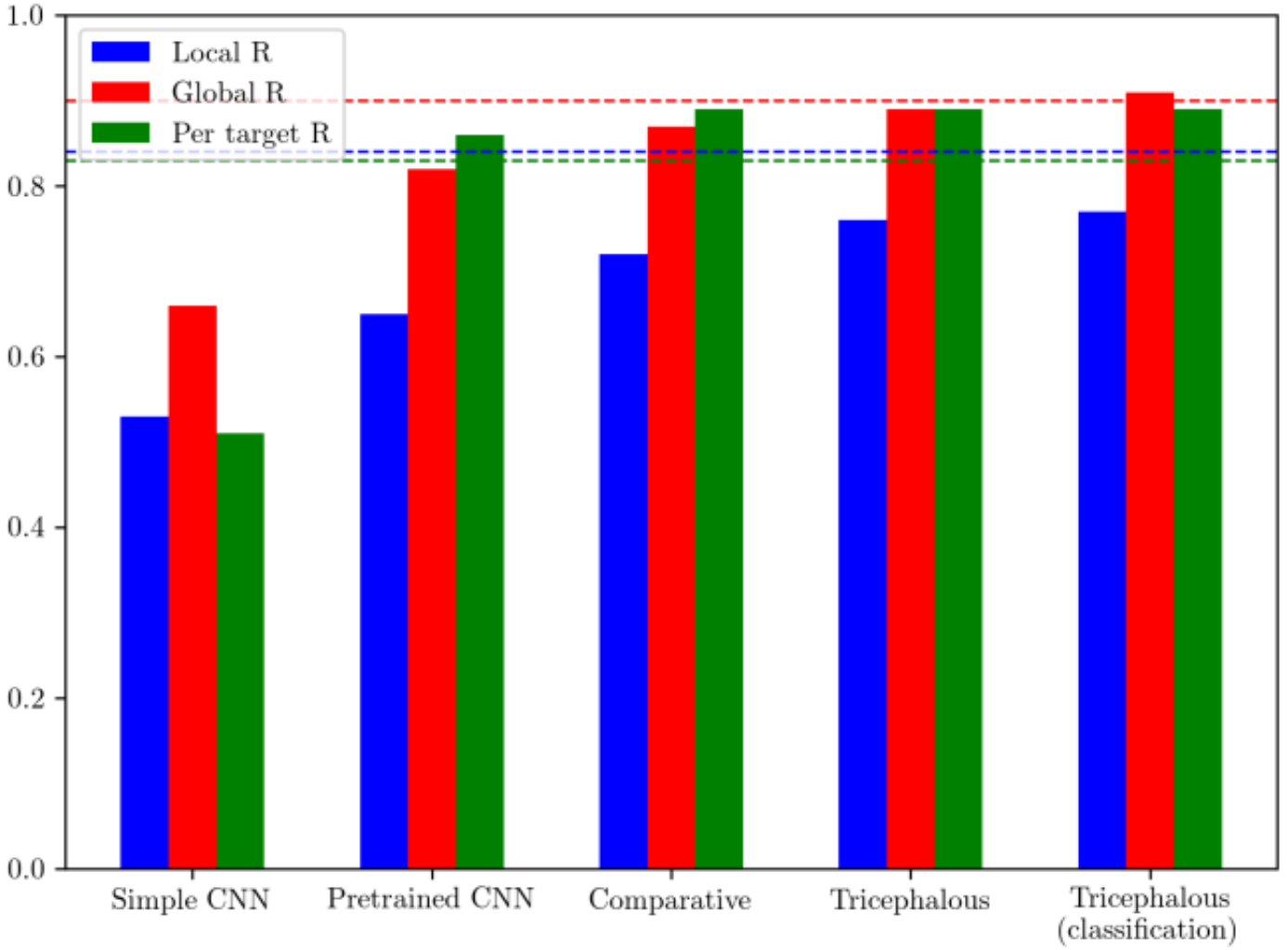




Regression as classification



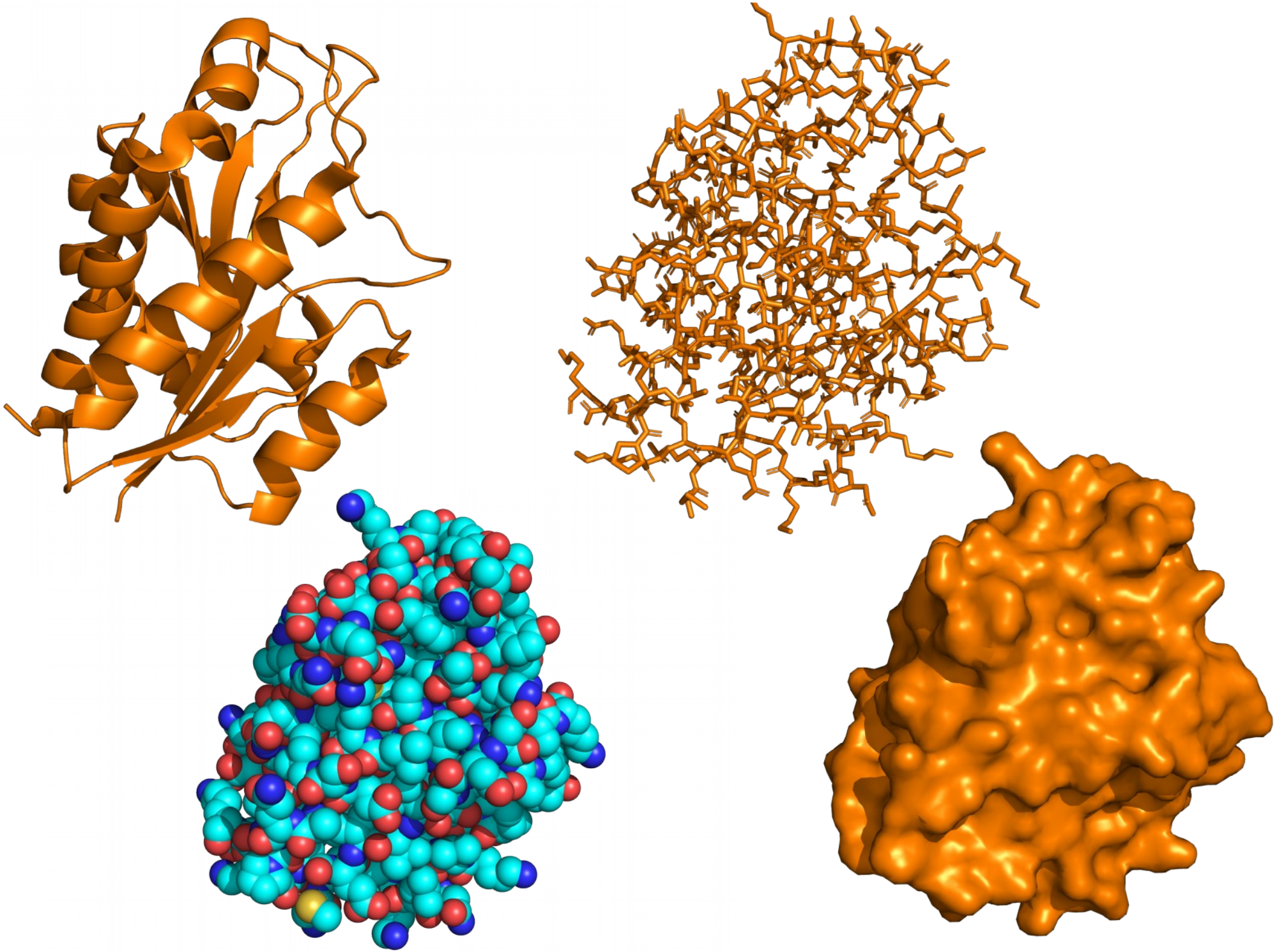
Regression as classification



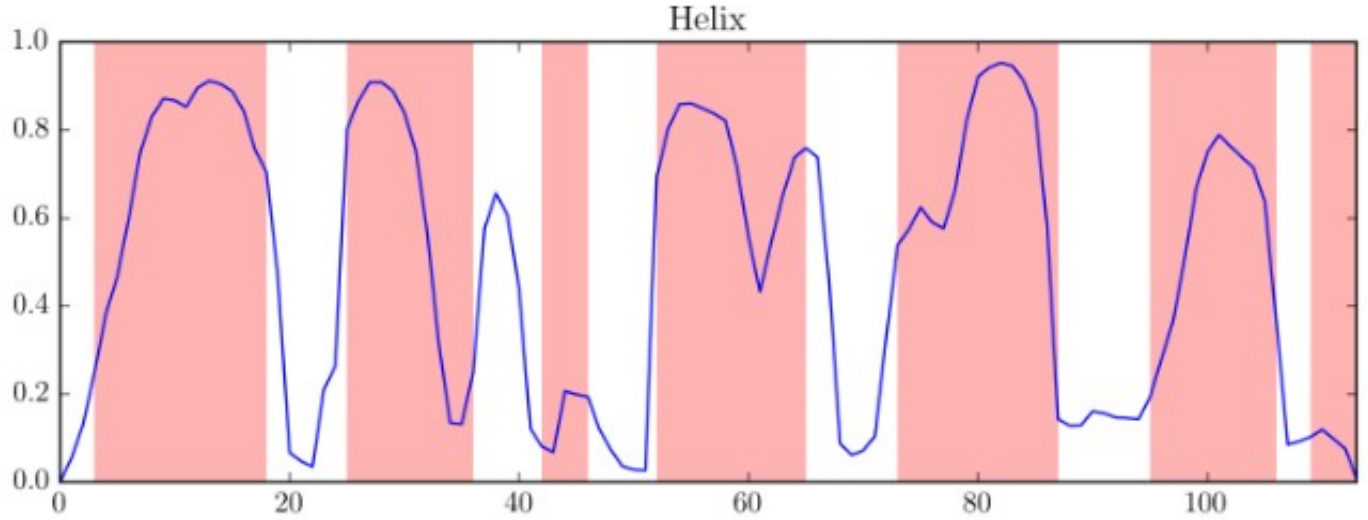
Conclusions

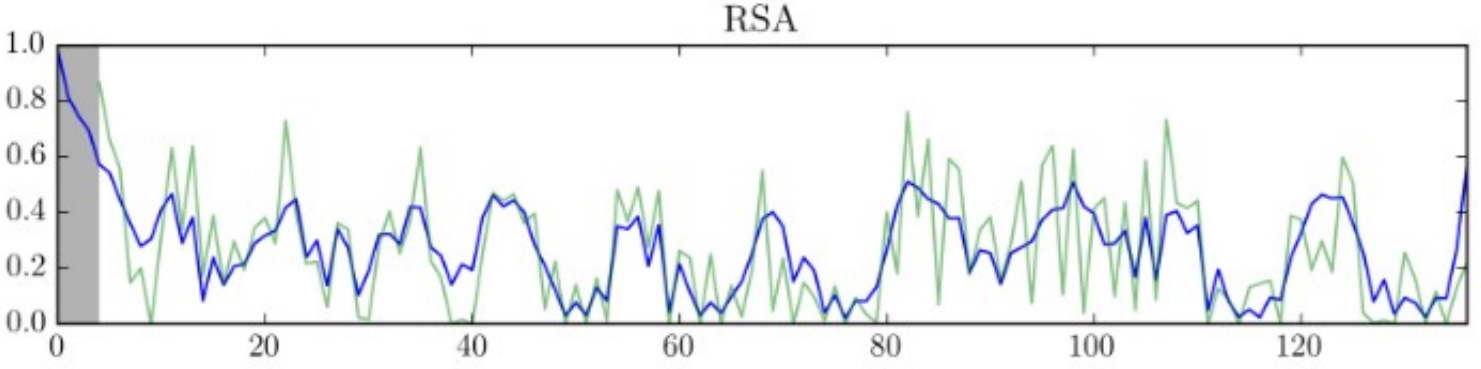
Mind your structures!

End.



Features: auxiliary predictors





Features: chemistry