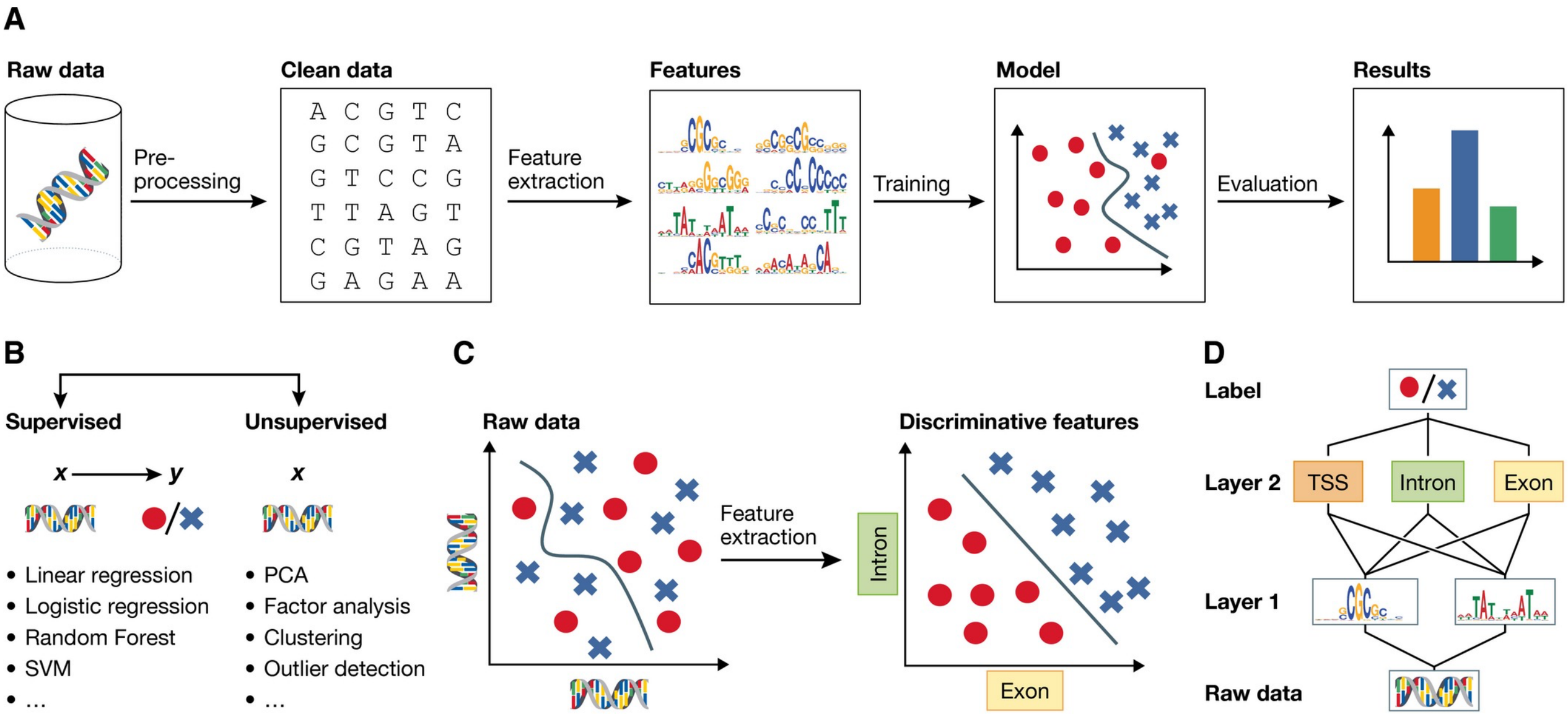
CS 6824: Deep Learning for Molecular Modeling

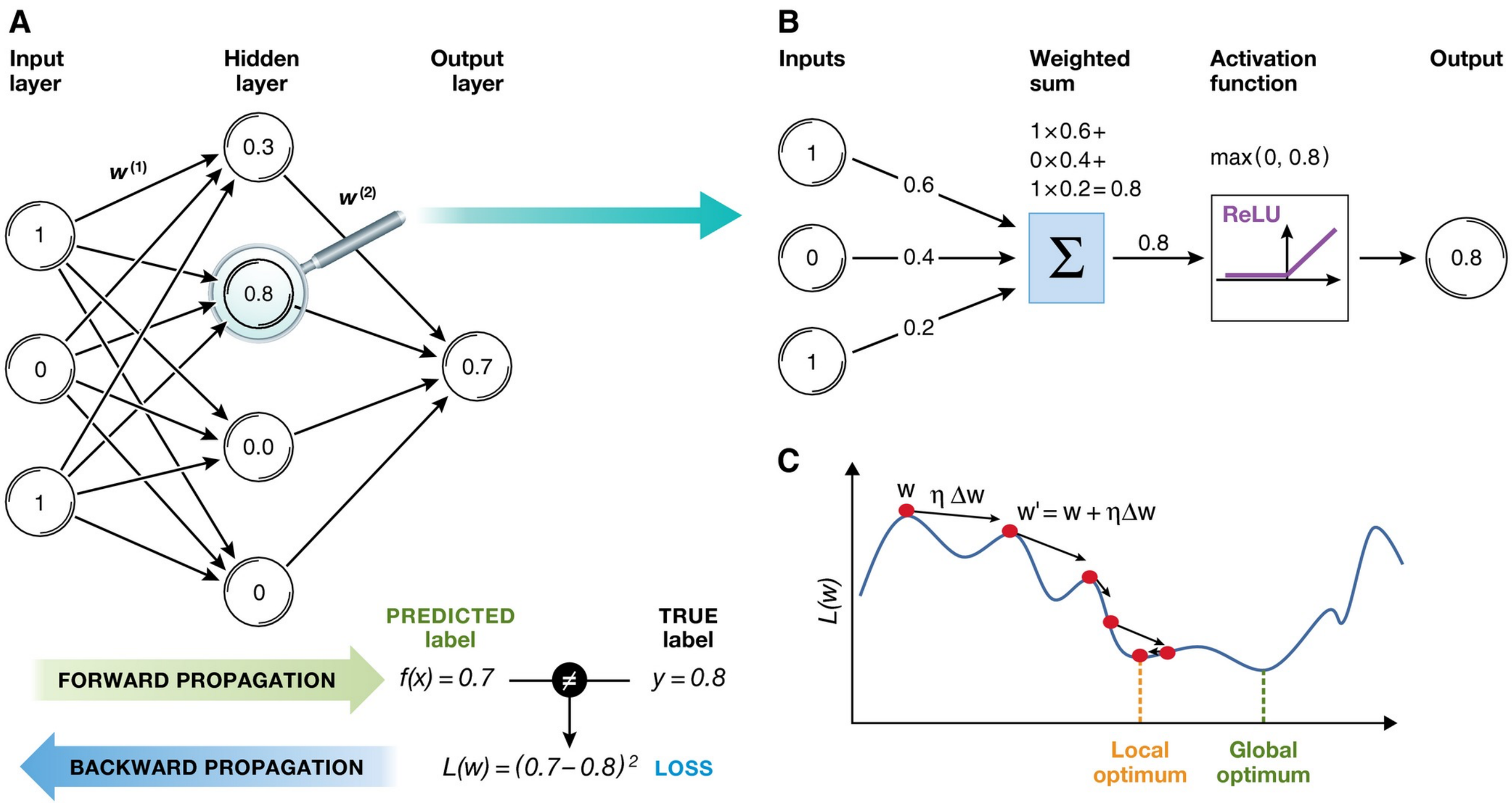
Acknowledgement:
Many of the images in the slides are derived from images.google.com or other publicly available sources.

Machine learning and representation learning



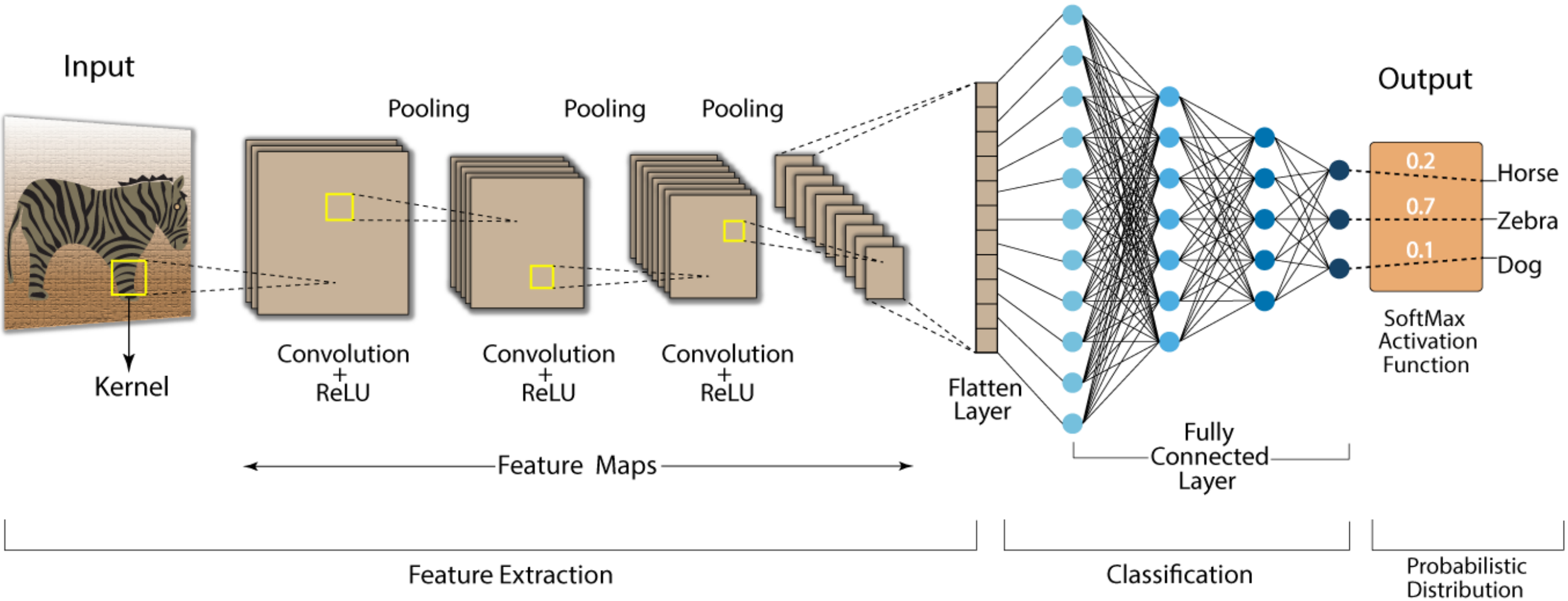
https://www.embopress.org/doi/full/10.15252/msb.20156651

Classical neural network

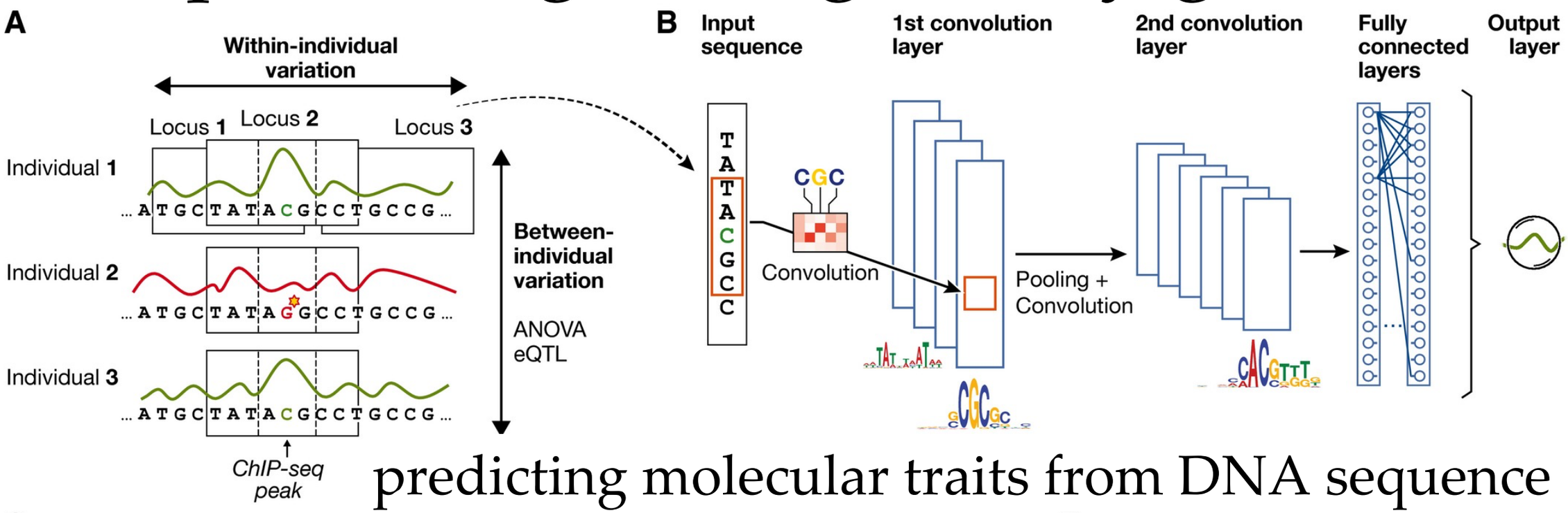


https://www.embopress.org/doi/full/10.15252/msb.20156651

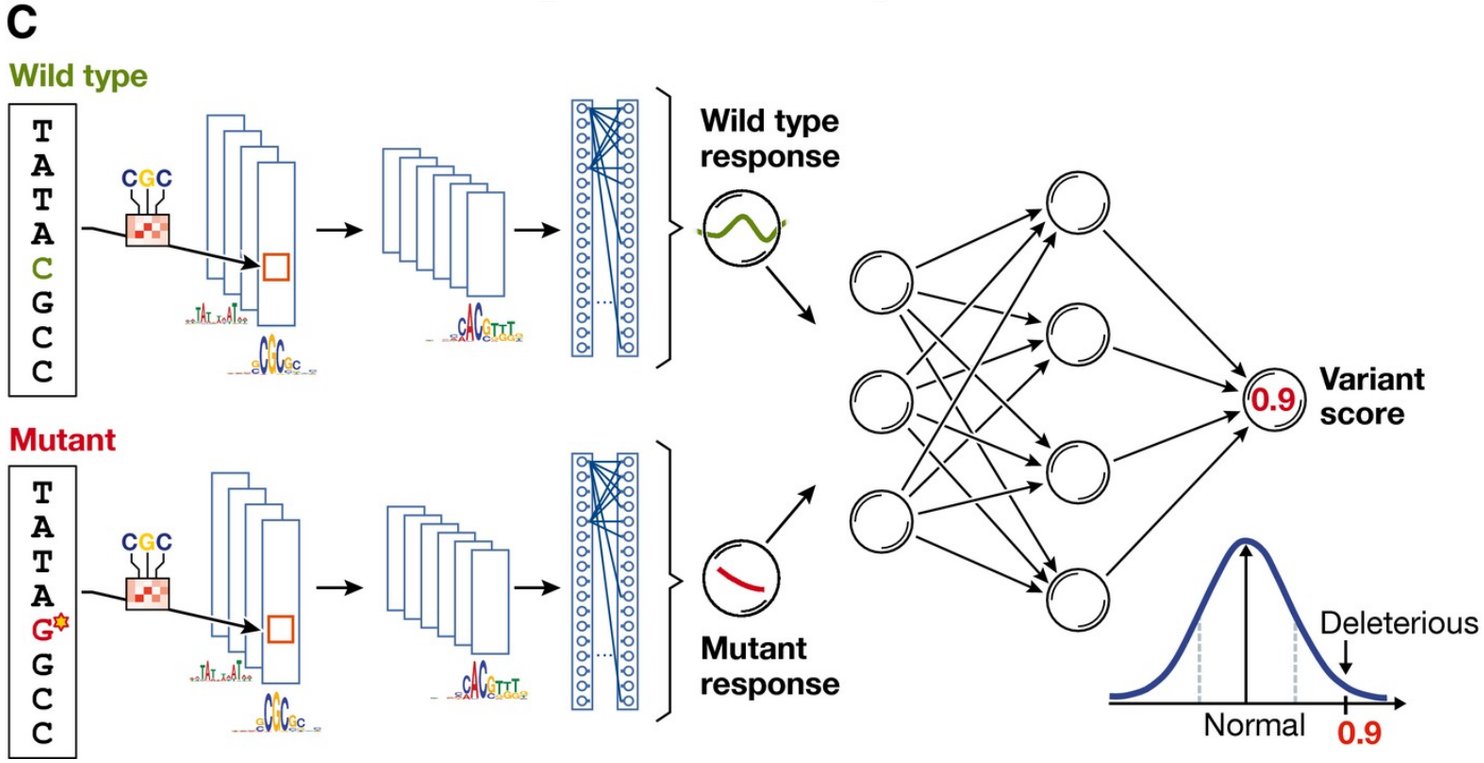
Convolutional neural network

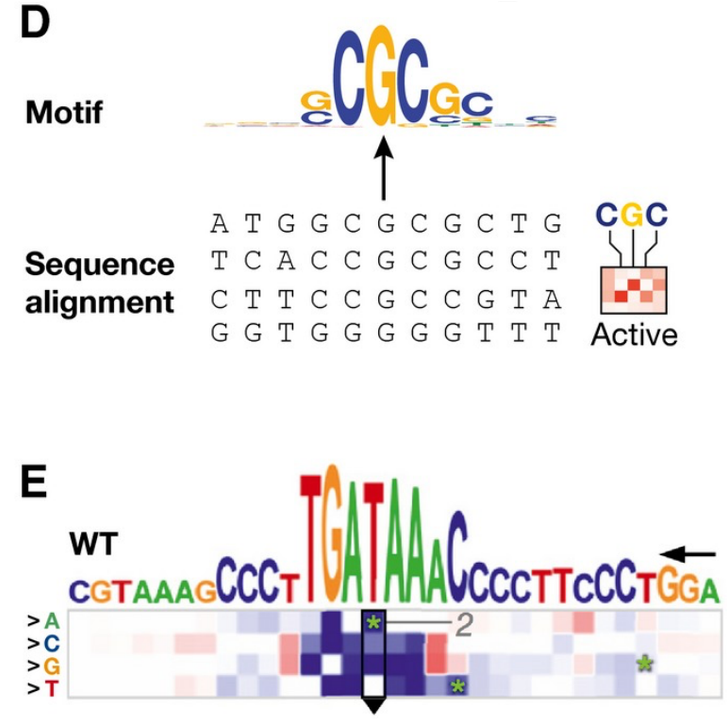


Deep learning for regulatory genomics



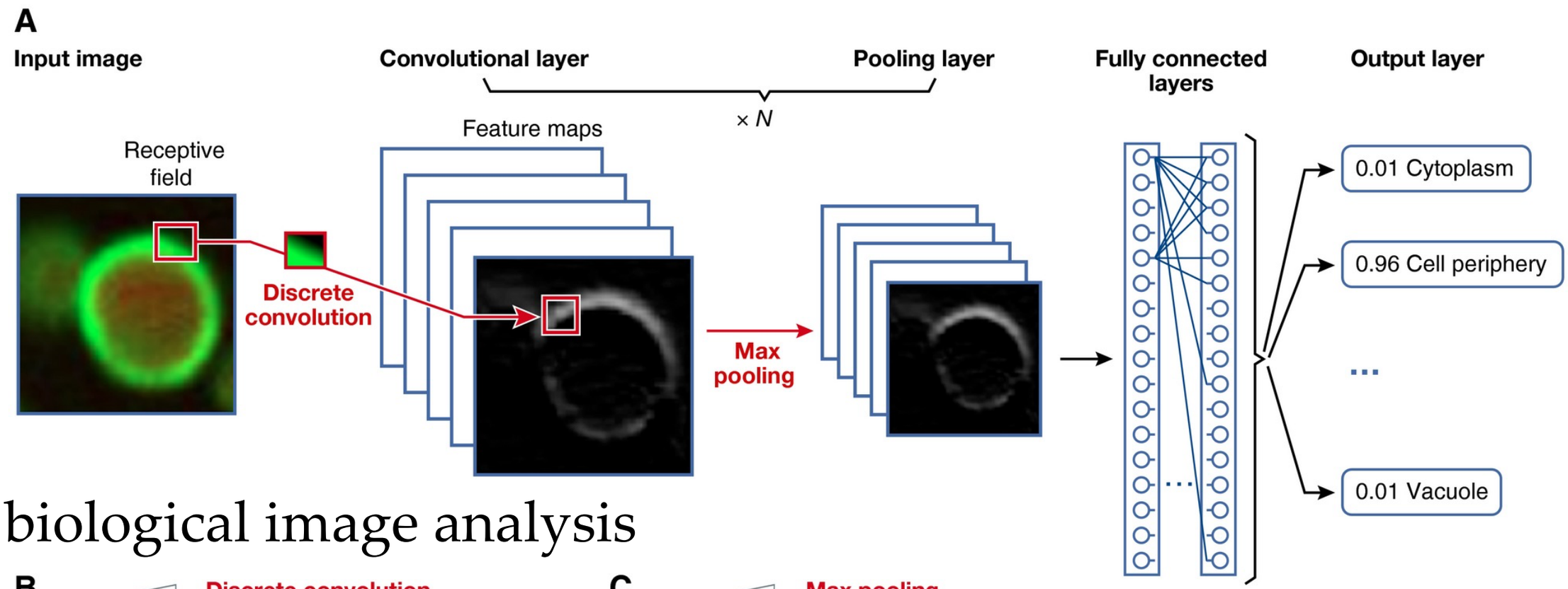
predicting molecular traits from DNA sequence



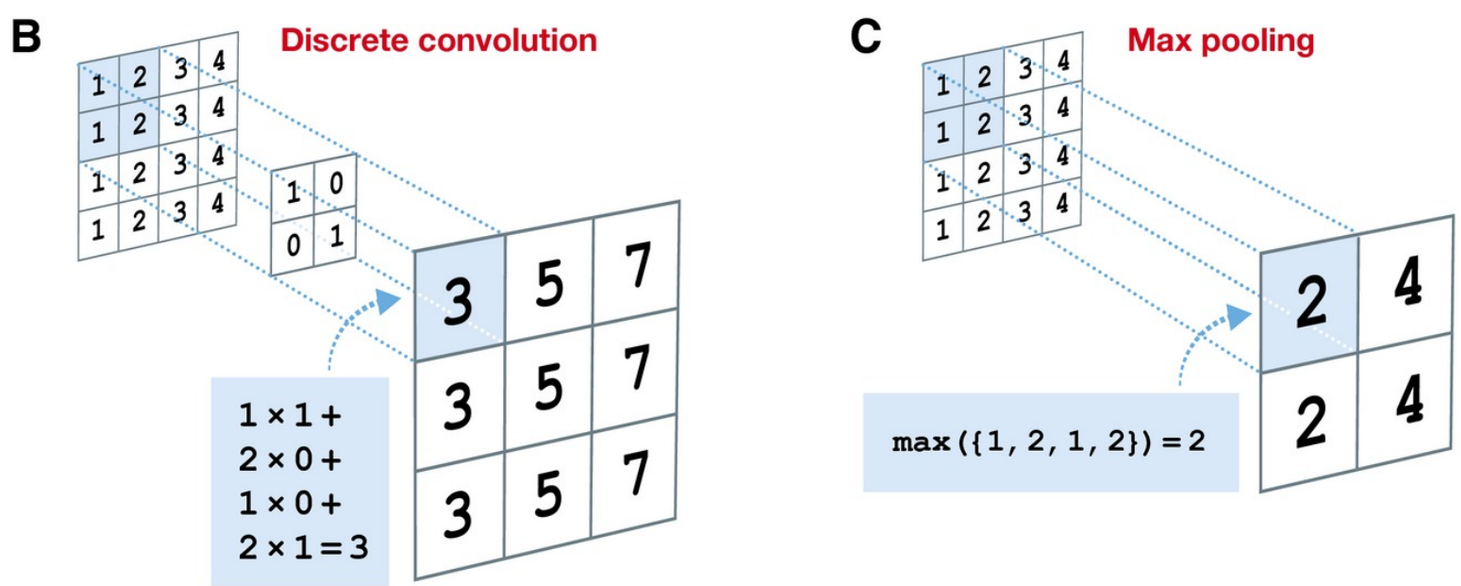


https://www.embopress.org/doi/full/10.15252/msb.20156651

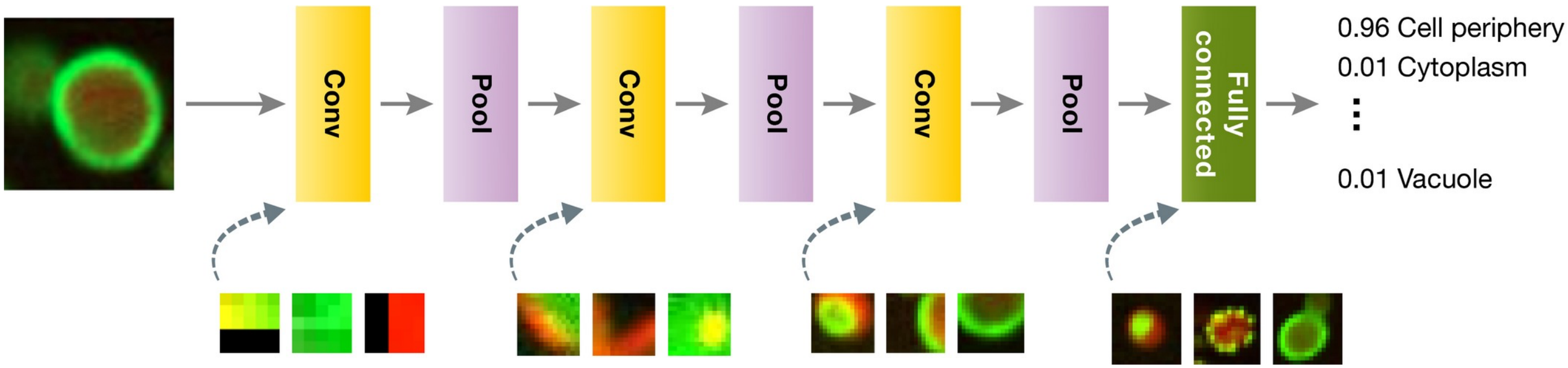
CNN for biological images



biological image analysis



Stacking



https://www.embopress.org/doi/full/10.15252/msb.20156651

Interpreting and visualizing convolutional networks

- Visualizing input weights:

First layer features

	In top left?	In top right?	…	In bottom right?
	0.21	0.24		0.01
	0.02	0.01		0.25
	0.01	0.03		0.19

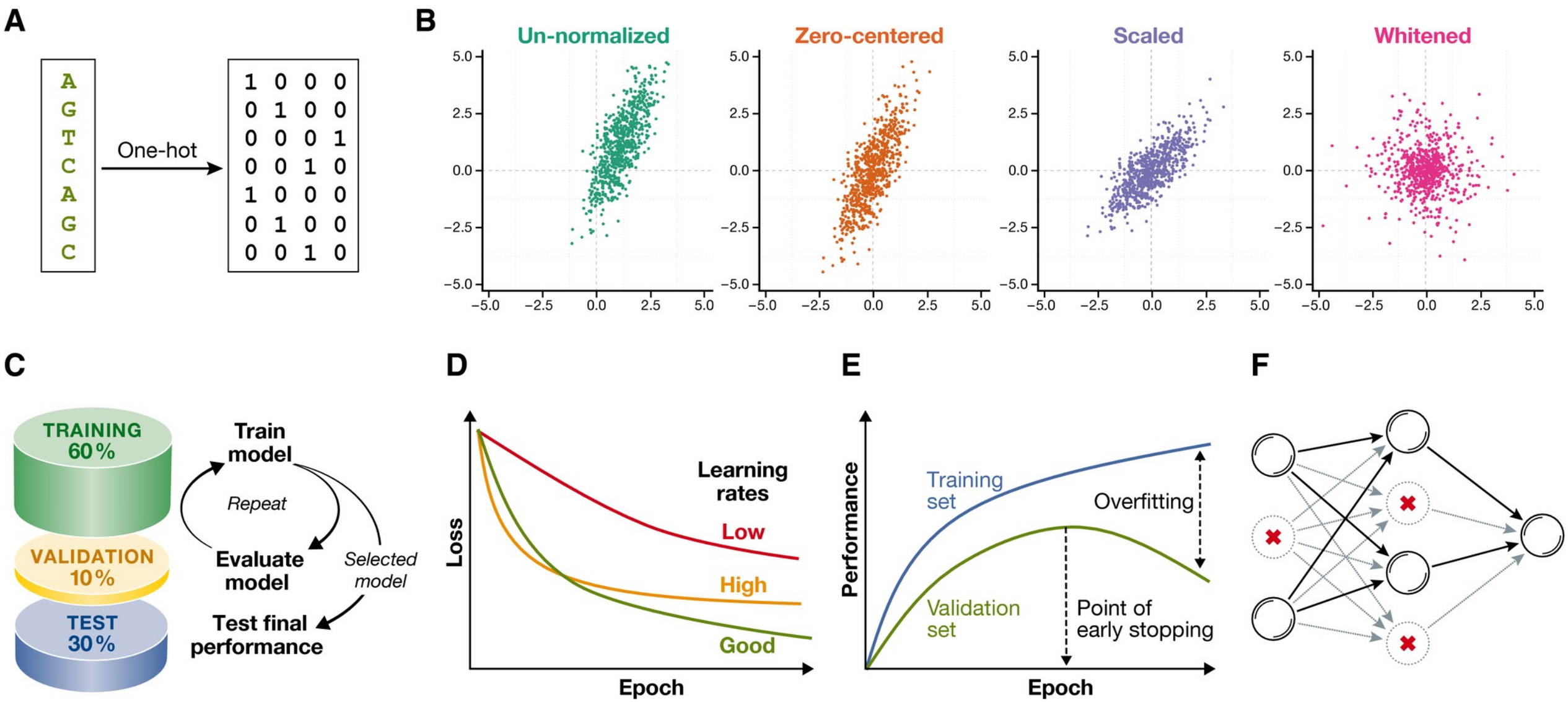
Third layer features

	In left?	In right?	…	In bottom?
	2.51	0.02		2.92
	0.03	0.01		0.02
	0.02	0.01		0.01

- Finding images that maximize neuron activity
- Hiding important image parts
- Visualizing similar inputs in two dimensions

https://www.embopress.org/doi/full/10.15252/msb.20156651

Data pre-processing for deep neural networks



https://www.embopress.org/doi/full/10.15252/msb.20156651

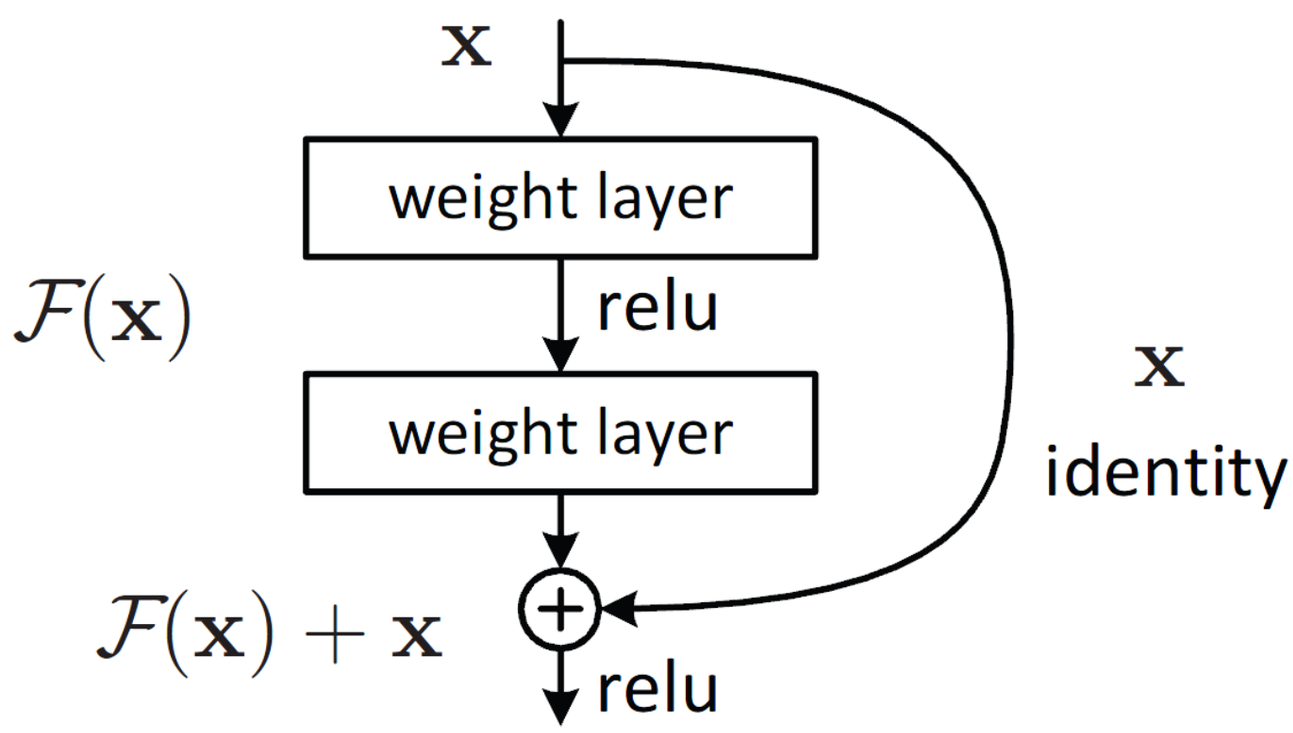
Training deep neural networks

- Choice of model architecture
- Determining the number of neurons in a network
- Optimization: Stochastic gradient descent
- Parameter initialization
- Learning rate and batch size
- Learning rate decay
- Batch normalization
- Analyzing the learning curve
- Monitoring training and validation performance

https://www.embopress.org/doi/full/10.15252/msb.20156651

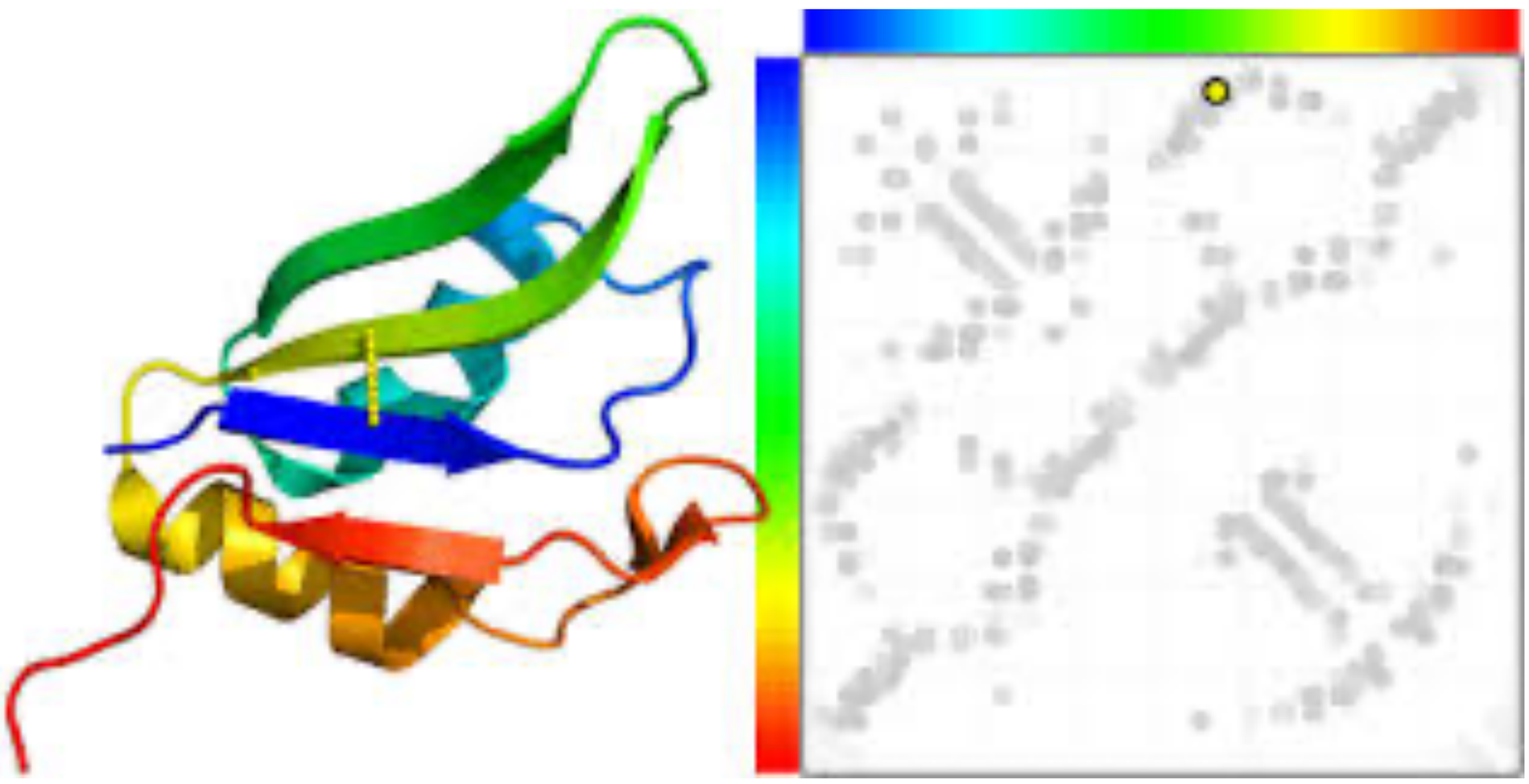
Residual Neural Network

- Instead of directly predicting y from x, predict y-x from x
- Add x and predicted y-x to estimate y
- y-x is the residual
- Equivalent to a shortcut connection between x and y
- Enables the training of deep networks by stacking many layers



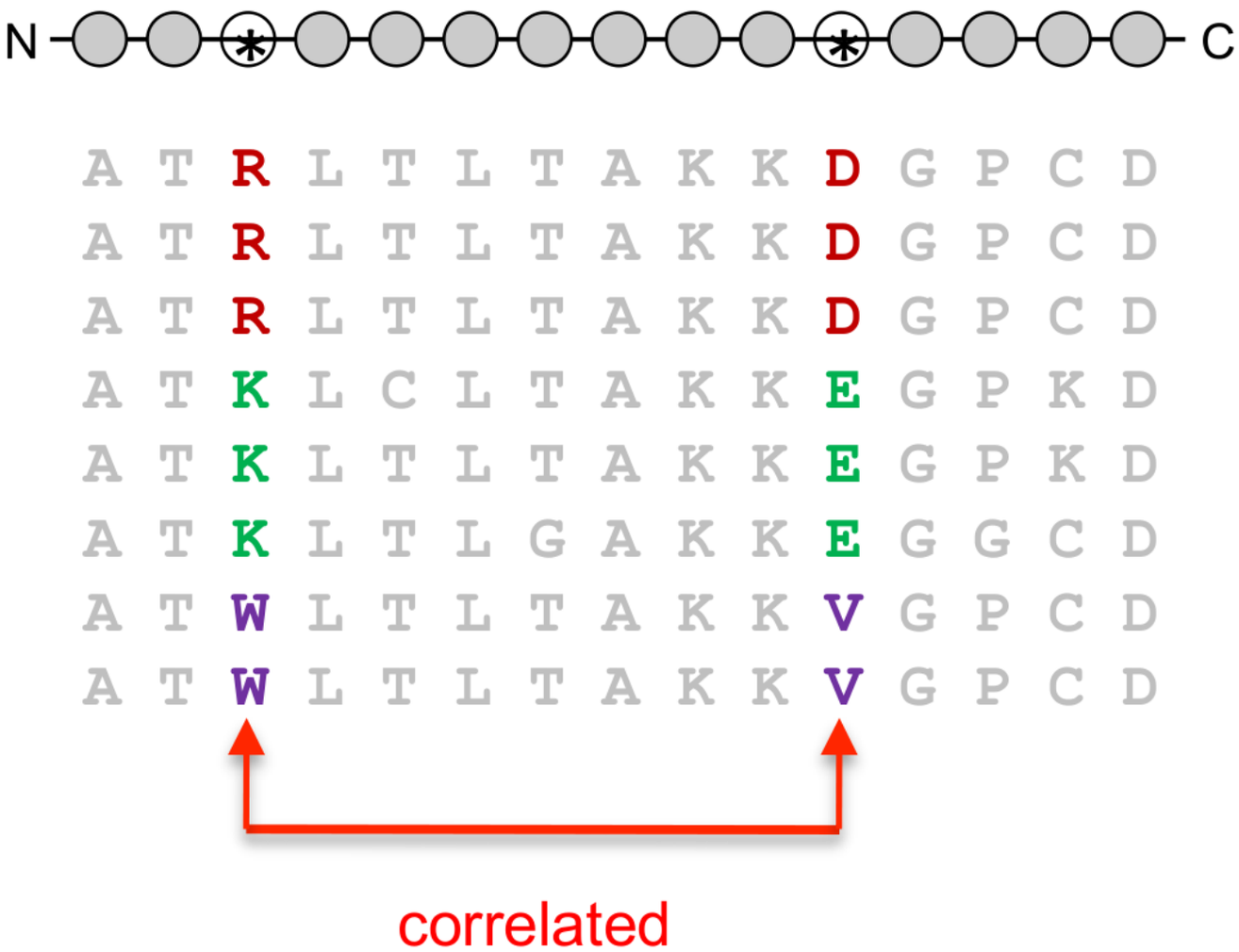
Protein contact map

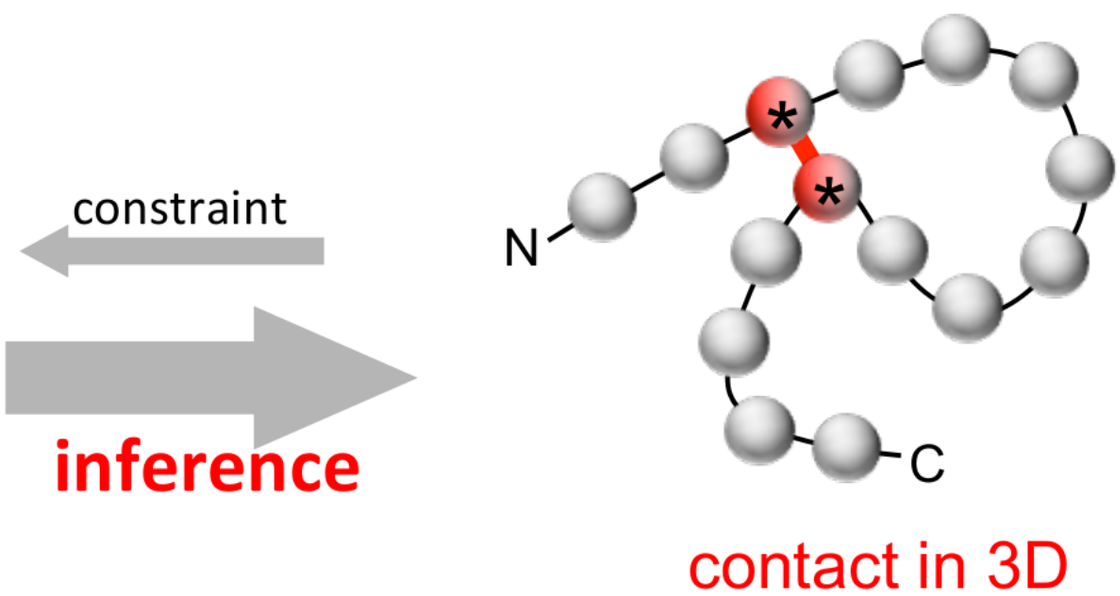
Protein contact map is a binary symmetric matrix capturing inter-residue interactions below a predefined distance threshold



Correlated mutation or co-evolution

Co-evolution patters can be analyzed to infer contacts

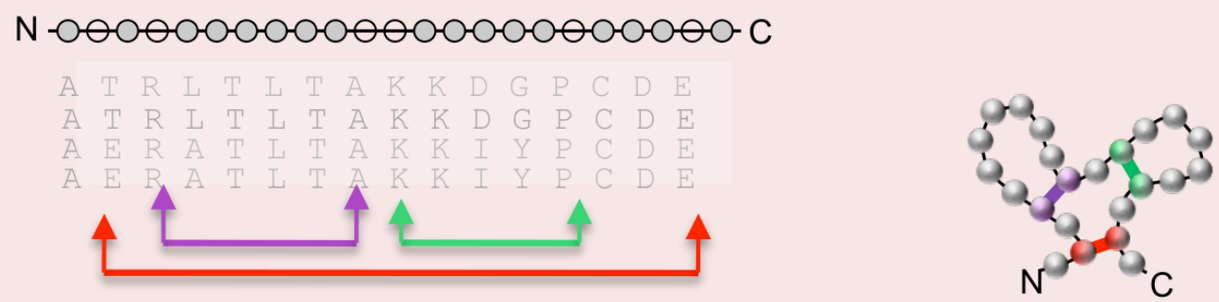




https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

Correlated mutation or co-evolution

Align evolutionary diverged sequences



Calculate covariance matrix for each pair of sequence positions for all pairs of amino acids (A,B)

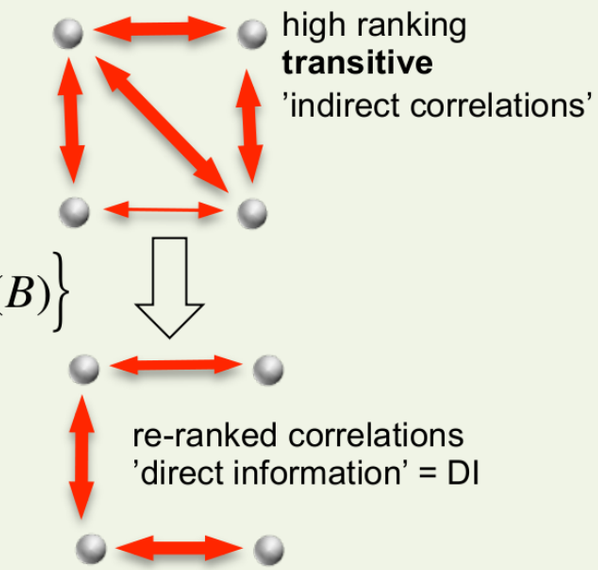
$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)P_j(B)$$

$$C_{ij}^{-1}(A,B) = -e_{ij}(A,B)_{i \neq j}$$

Identify maximally informative pair couplings using **statistical model** of entire protein to infer residue-residue co-evolution

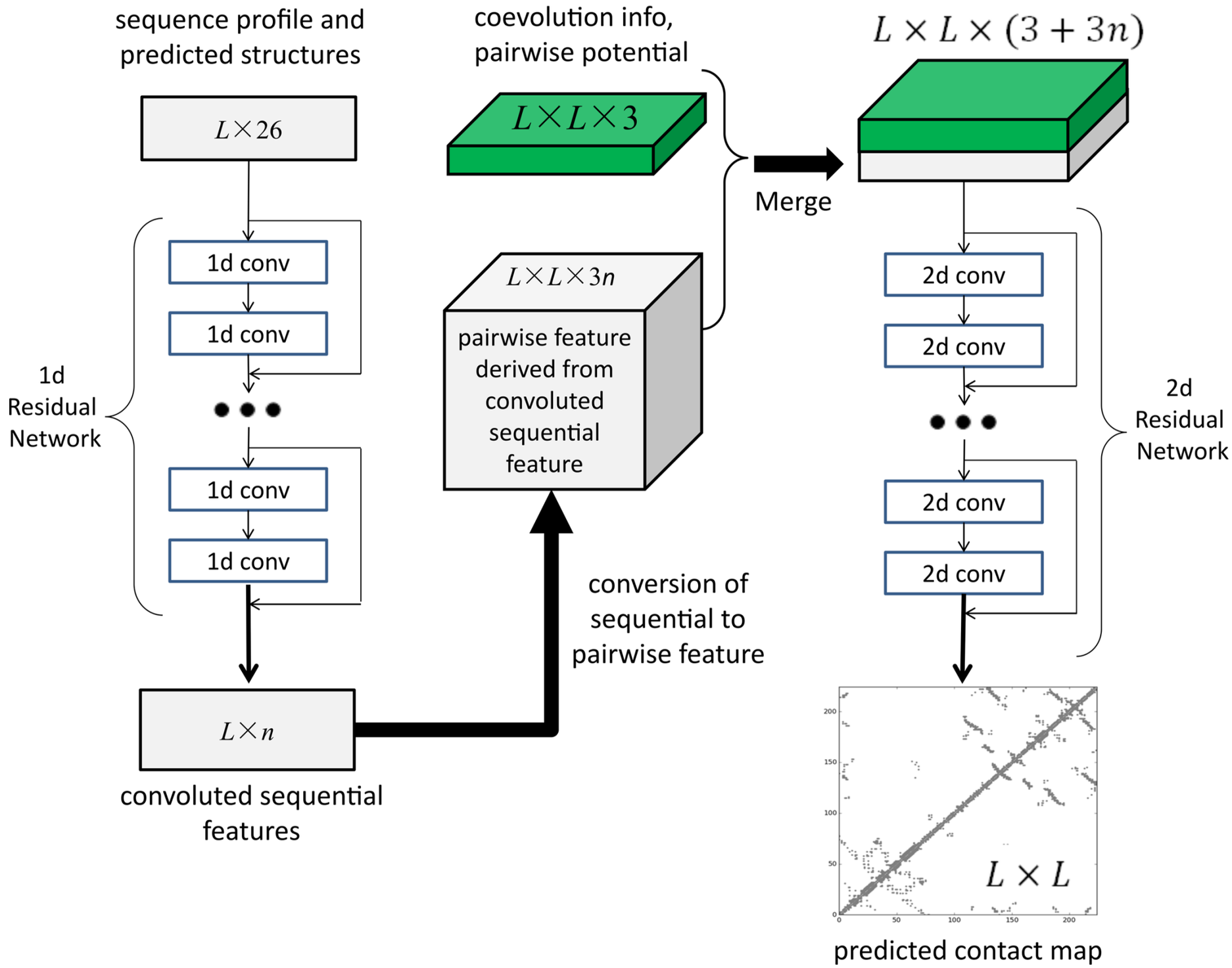
$$P_{ij}^{Dir}(A,B) = \frac{1}{Z}\exp\left\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}$$

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{Dir}(A,B)\ln\frac{P_{ij}^{Dir}(A,B)}{f_i(A)f_j(B)}$$



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766

ResNet for protein contact map prediction



https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324

ResNet accurately predicts protein contact map

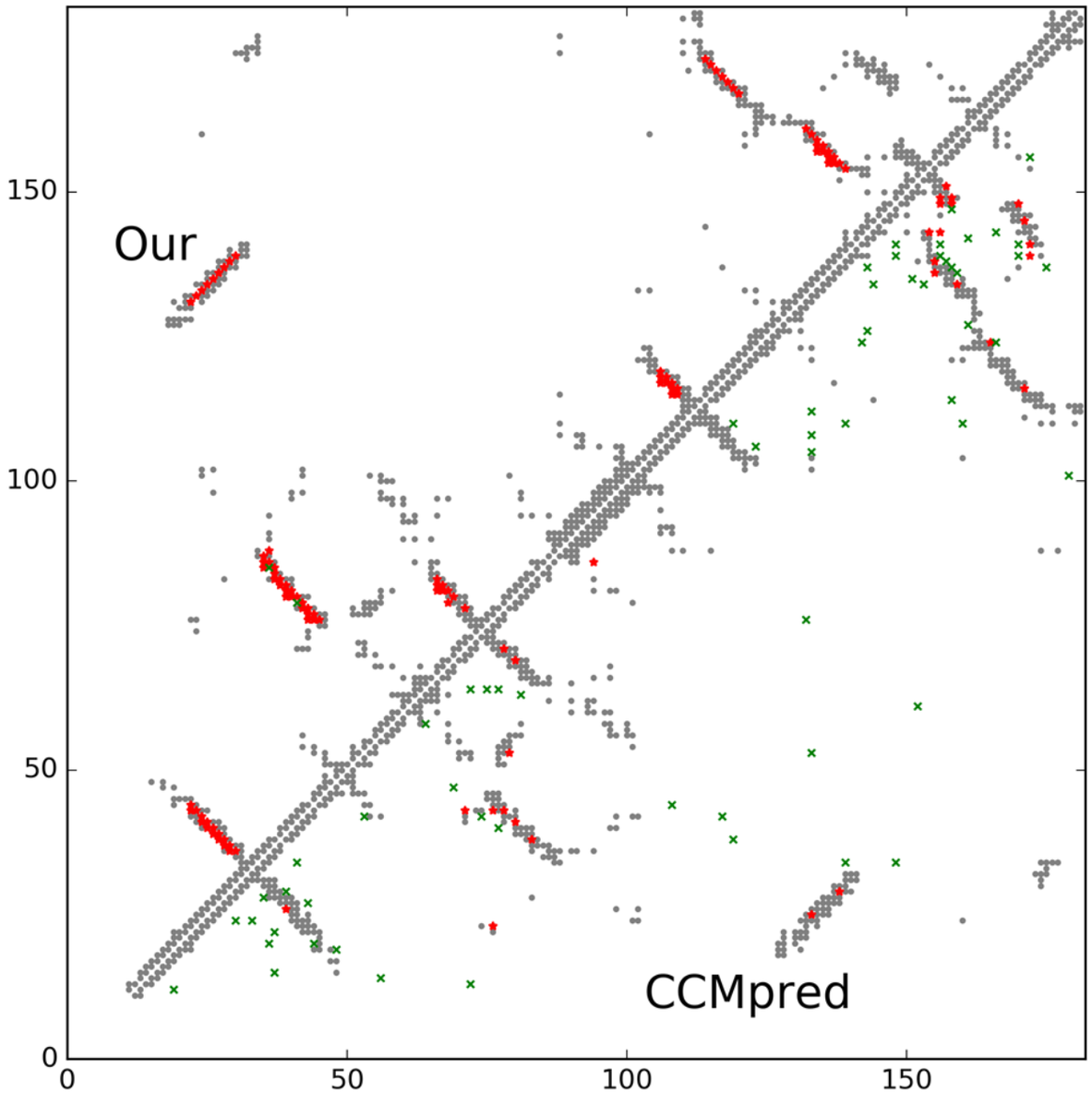
Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.50	0.40	0.26	0.17	0.64	0.52	0.34	0.22	0.74	0.68	0.53	0.39
PSICOV	0.58	0.43	0.26	0.17	0.65	0.51	0.32	0.20	0.77	0.70	0.52	0.37
CCMpred	0.65	0.50	0.29	0.19	0.73	0.60	0.37	0.23	0.82	0.76	0.62	0.45
plmDCA	0.66	0.50	0.29	0.19	0.72	0.60	0.36	0.22	0.81	0.76	0.61	0.44
Gremlin	0.66	0.51	0.30	0.19	0.74	0.60	0.37	0.23	0.82	0.76	0.63	0.46
MetaPSICOV	0.82	0.70	0.45	0.27	0.83	0.73	0.52	0.33	0.92	0.87	0.74	0.58
Our method	0.93	0.81	0.51	0.30	0.93	0.86	0.62	0.38	0.98	0.96	0.89	0.74

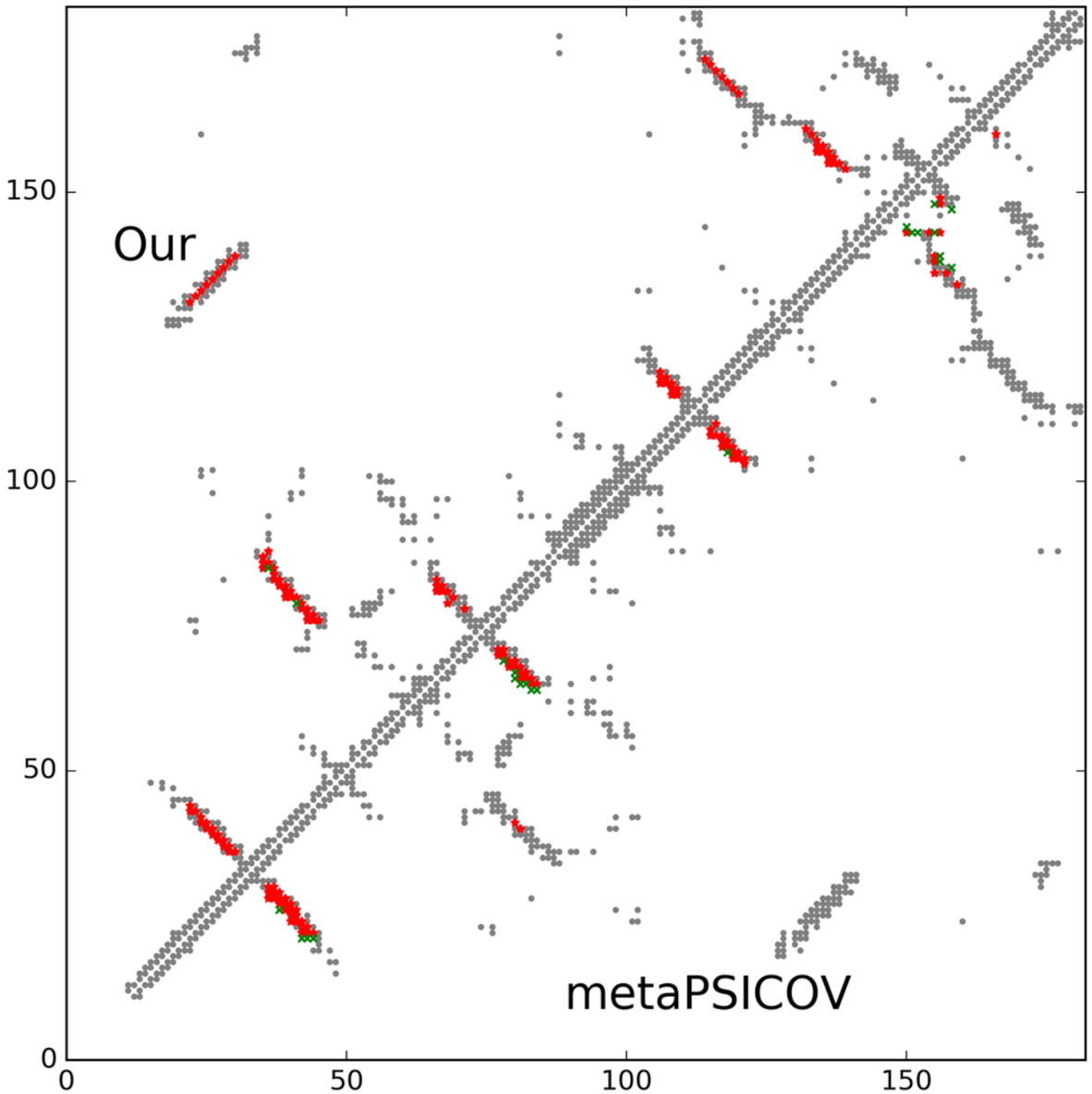
doi:10.1371/journal.pcbi.1005324.t001

Contact prediction accuracy on 150 Pfam families

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324

ResNet accurately predicts protein contact map

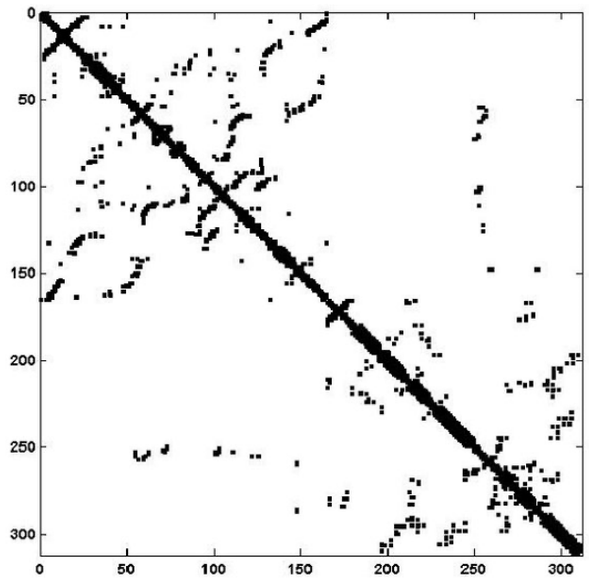


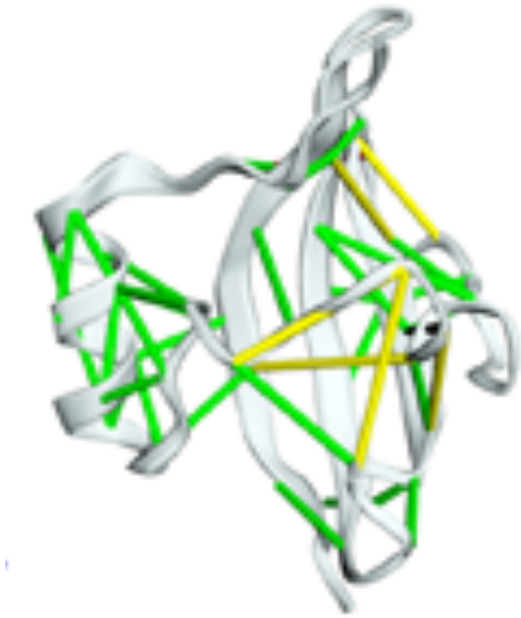


correct (red) and incorrect (green) predicted contacts on native contacts (gray)

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324

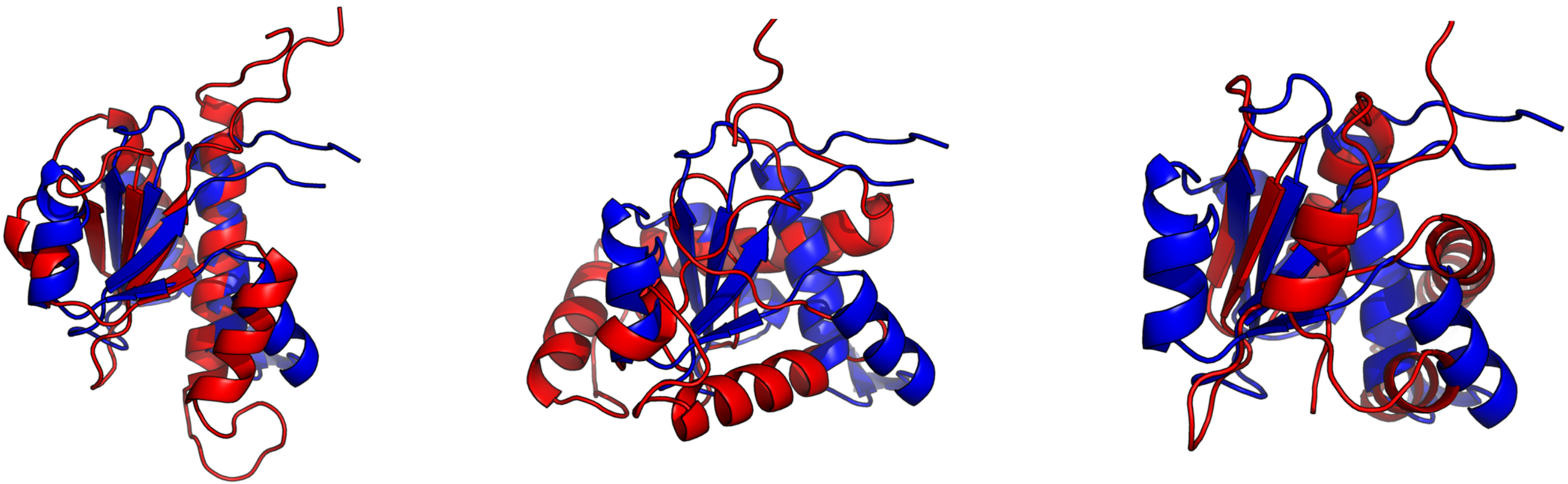
From contact maps to 3D structures





- Constraint satisfaction using
  - Distance geometry
  - Stochastic optimization

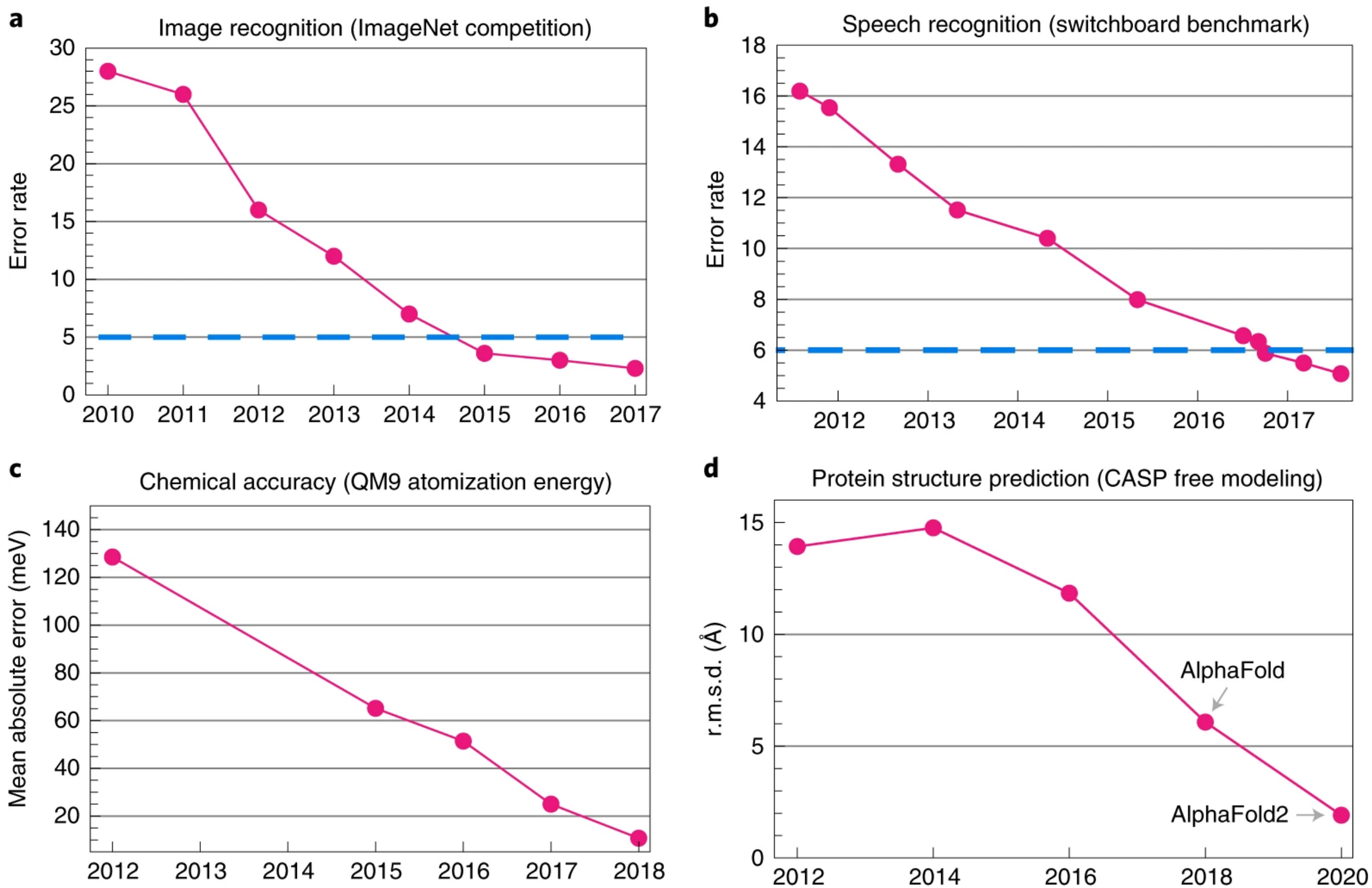
This often leads to accurate 3D structures



predicted models (red) and the native structure (blue)

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324

Deep learning revolution



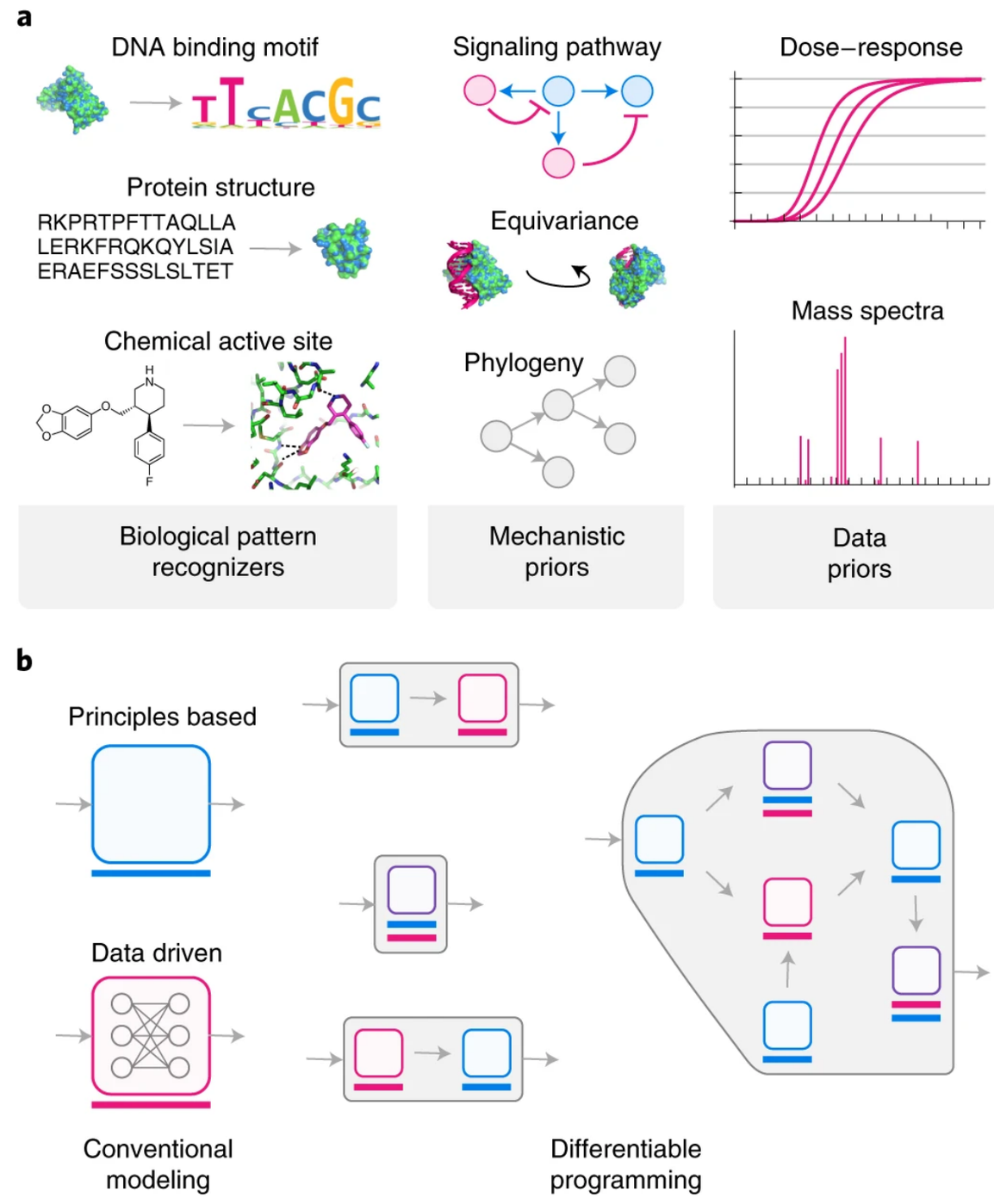
https://www.nature.com/articles/s41592-021-01283-4

Differentiable biology

- **Biological pattern recognizers**
  - 1D vectors comprising DNA/RNA sequences (DNA binding motif)
  - 2D grids with fixed dimensions (protein contact map prediction)
  - Generalizing 2D grids to higher dimensions, for example, by discretizing 3D space into equal-sized cubes (affinity of protein–drug complexes)

- **Mechanistic priors**
  - ML research in biology increasingly incorporates prior knowledge about structure, chemistry, and evolution

- **Data priors**
  - Biology involves analysis of incomplete, noisy and heterogeneous data
  - Incorporating priors that account for the data generation process is necessary to minimize the effects of error and fuse disparate data types



https://www.nature.com/articles/s41592-021-01283-4