Going Deep – The Past, Present, and Future of Neural Networks in Structural Biology

Jens Meiler

UNIVERSITÄT LEIPZIG
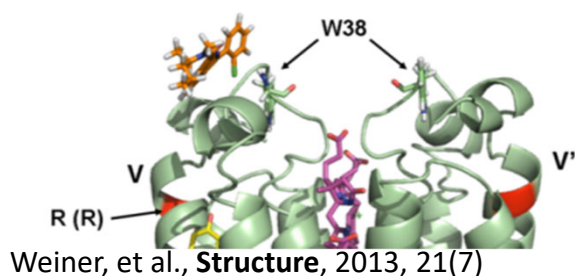
and
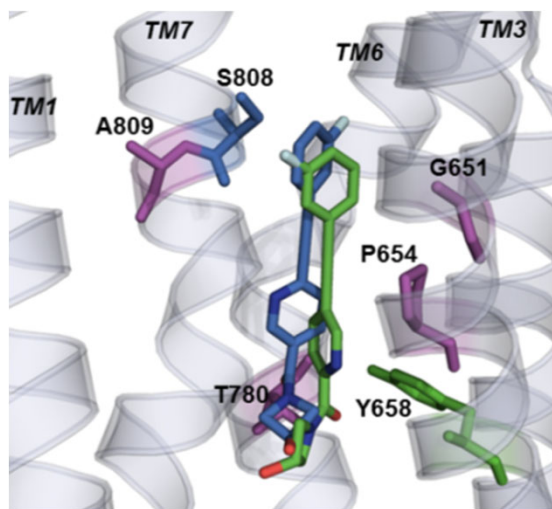
VANDERBILT UNIVERSITY

are partner universities: www.leipzig.vanderbilt.edu

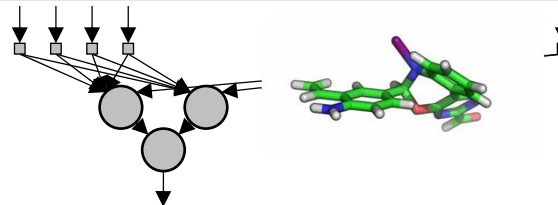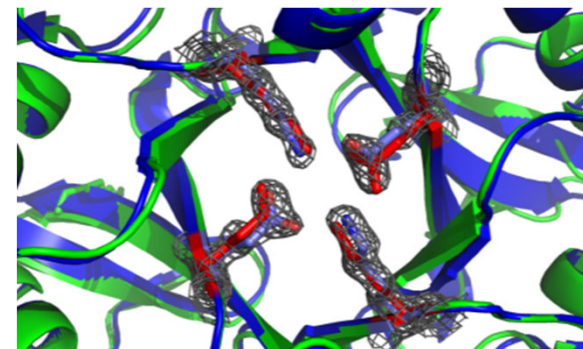# Computational Structural and Chemical Biology in the Meiler Lab

Weiner, et al., **Structure**, 2013, 21(7)

Protein structure prediction de novo and from limited experimental data

Dong, et al., **PLoS ONE**, 2013, 8(7)

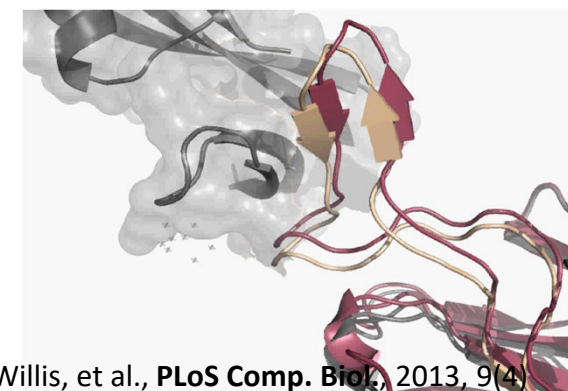*Merging ligand- and structure-based computer-aided drug discovery*

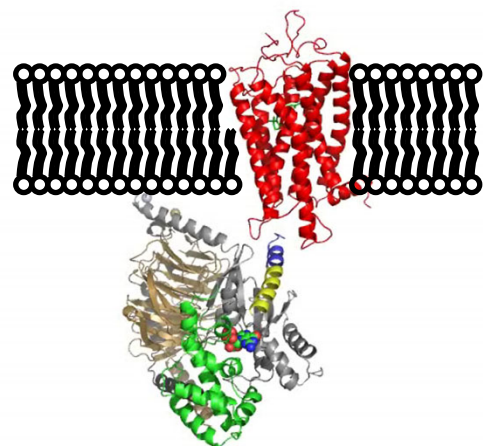Muller, et al., **ChemMedChem**, 2012, 7(3)

Fortenberry, et al., **JACS**, 2011, 133(45)

Design of large protein scaffolds, antibodies, and protein/ligand interfaces
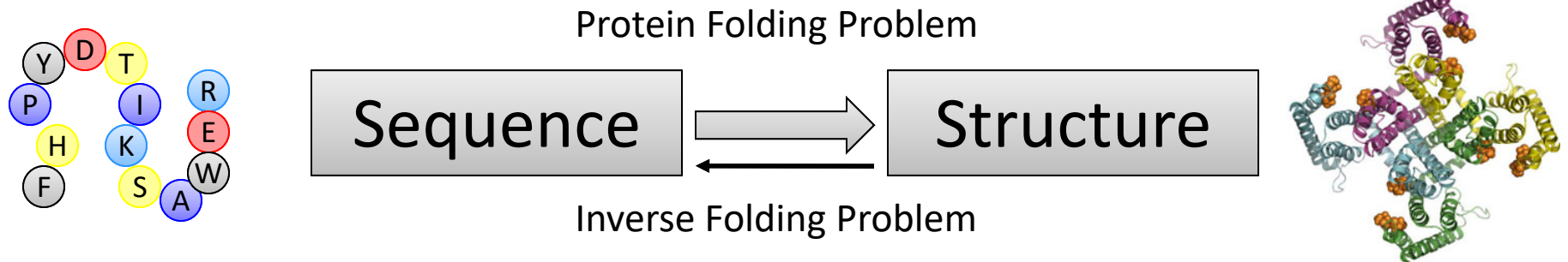

Willis, et al., **PLoS Comp. Biol.**, 2013, 9(4)

Alexander et al., **NSMB**, 2014, 21(1)

The (Inverse) Protein Folding Problem
Holy Grail of Comp. Structural Biology

Protein Folding Problem

| Sequence | ⇒ | Structure |

Inverse Folding Problem

- Given a protein's AA sequence, what is its 3-dimensional fold , and how does it get there?

- Assume 100 conformations for each amino acid in a 100 amino acid protein ⇨ 10^{200} possible conformations!

- Cyrus Levinthal's paradox of protein folding, 1968.
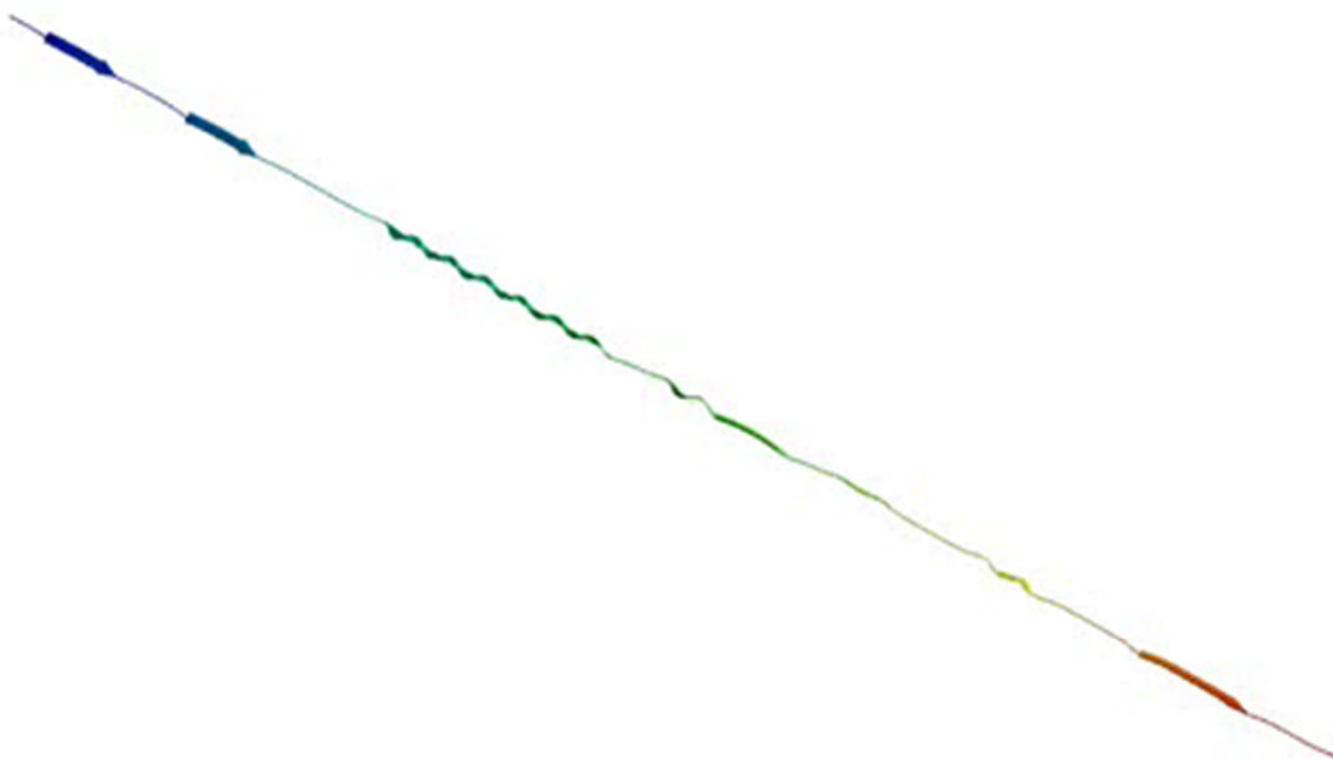
Rosetta: A Unified Framework for Tackling Molecular Modeling
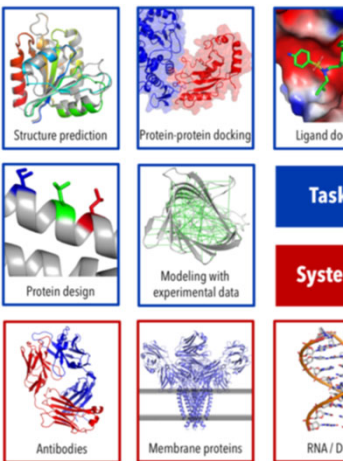
A. Leaver-Fay, et al.; "ROSETTA3: an object-oriented software suite …"; Methods Enzymol; 2011; Vol. 487 p. 545-74.
J. K. Leman, et al.; "Macromolecular modeling and design in Rosetta: recent methods and frameworks"; Nat Methods; 2020; Vol. 17 (7): p. 665-680.

Rosetta: A Unified Framework for Tackling Molecular Modeling
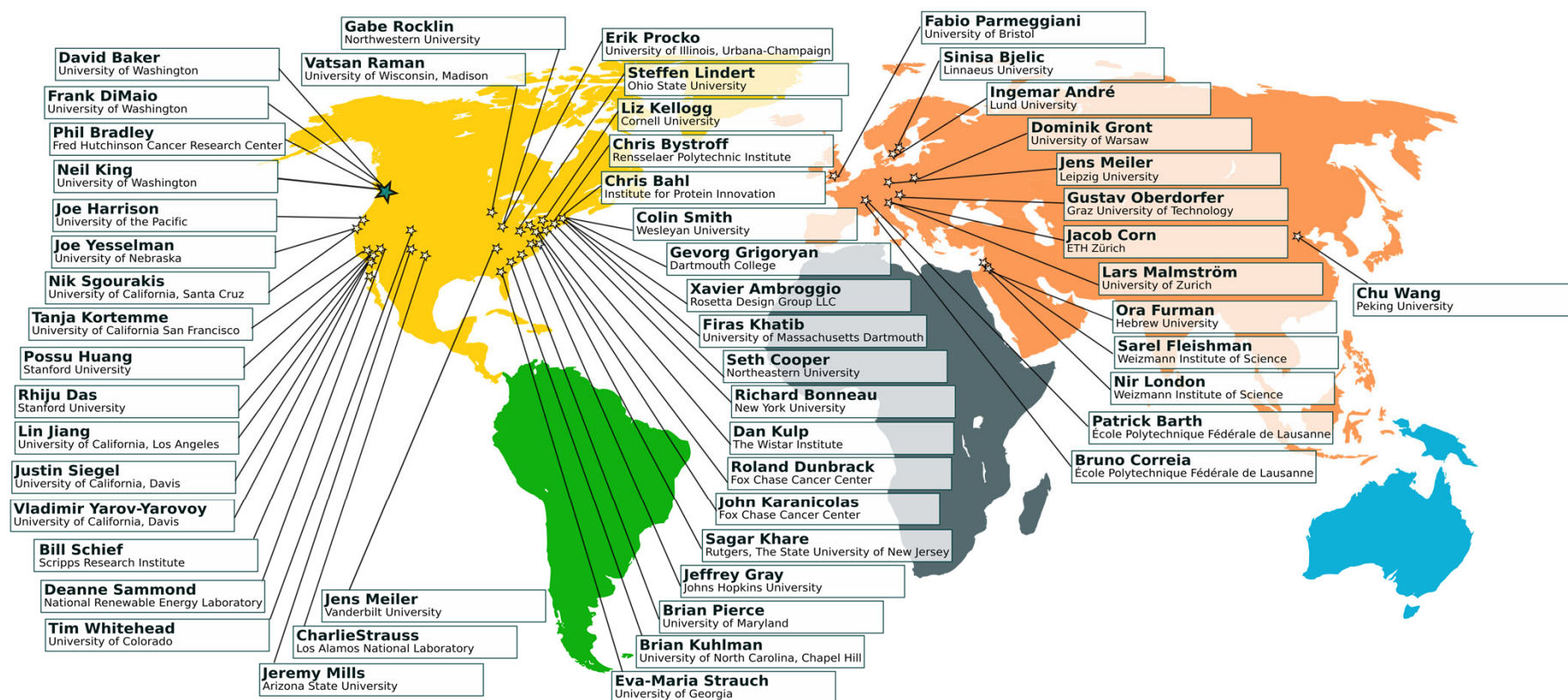
A. Leaver-Fay, et al.; "ROSETTA3: an object-oriented software suite …"; Methods Enzymol; 2011; Vol. 487 p. 545-74.
J. K. Leman, et al.; "Macromolecular modeling and design in Rosetta: recent methods and frameworks"; Nat Methods; 2020; Vol. 17 (7): p. 665-680.
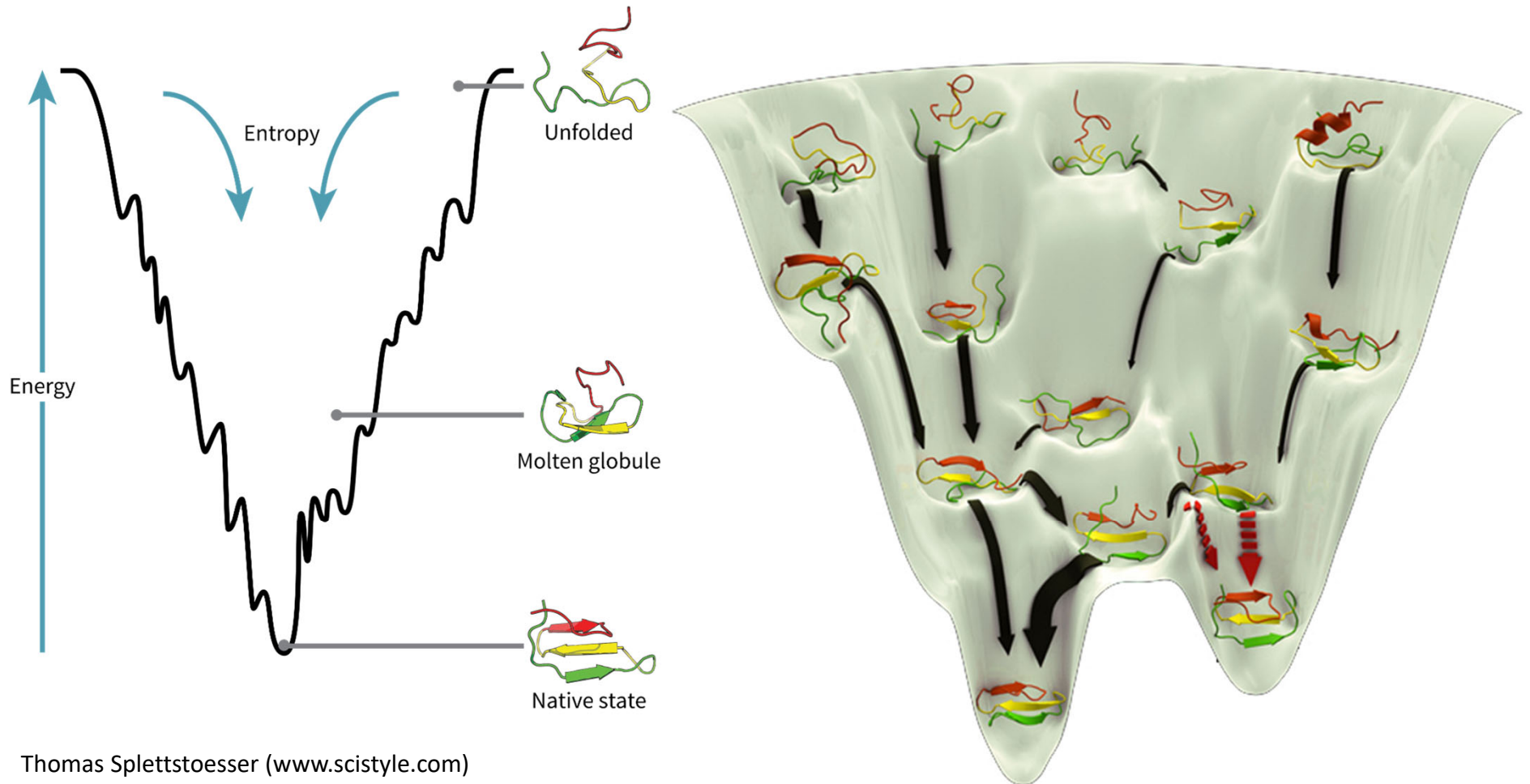
Peptide Bond Formation and Folding of Protein Tertiary Structure
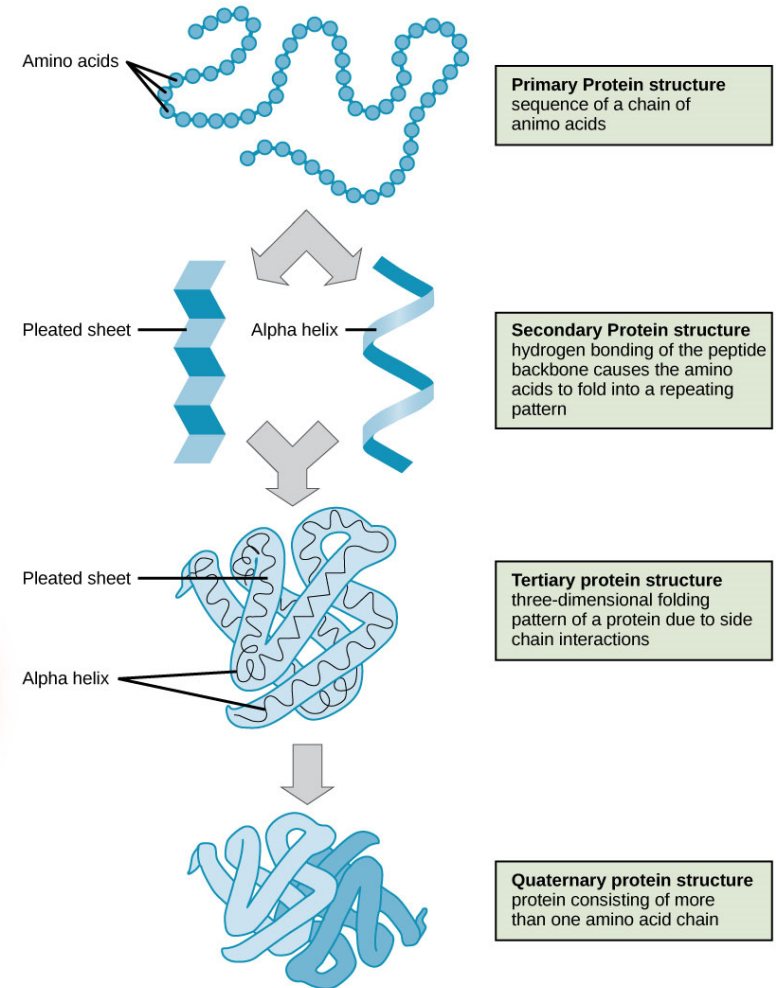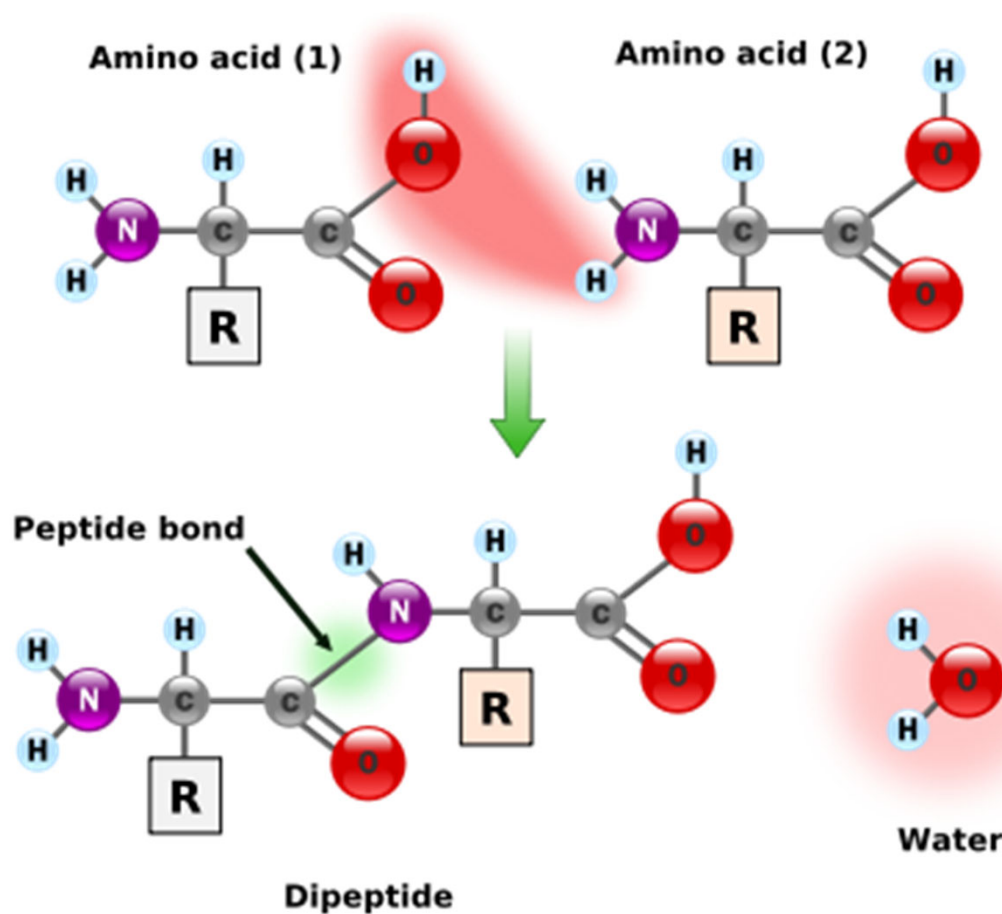
Protein Folding is Driven by the Minimization of Free Energy

Thomas Splettstoesser (www.scistyle.com)

Protein Tertiary Structure is Tied to Function
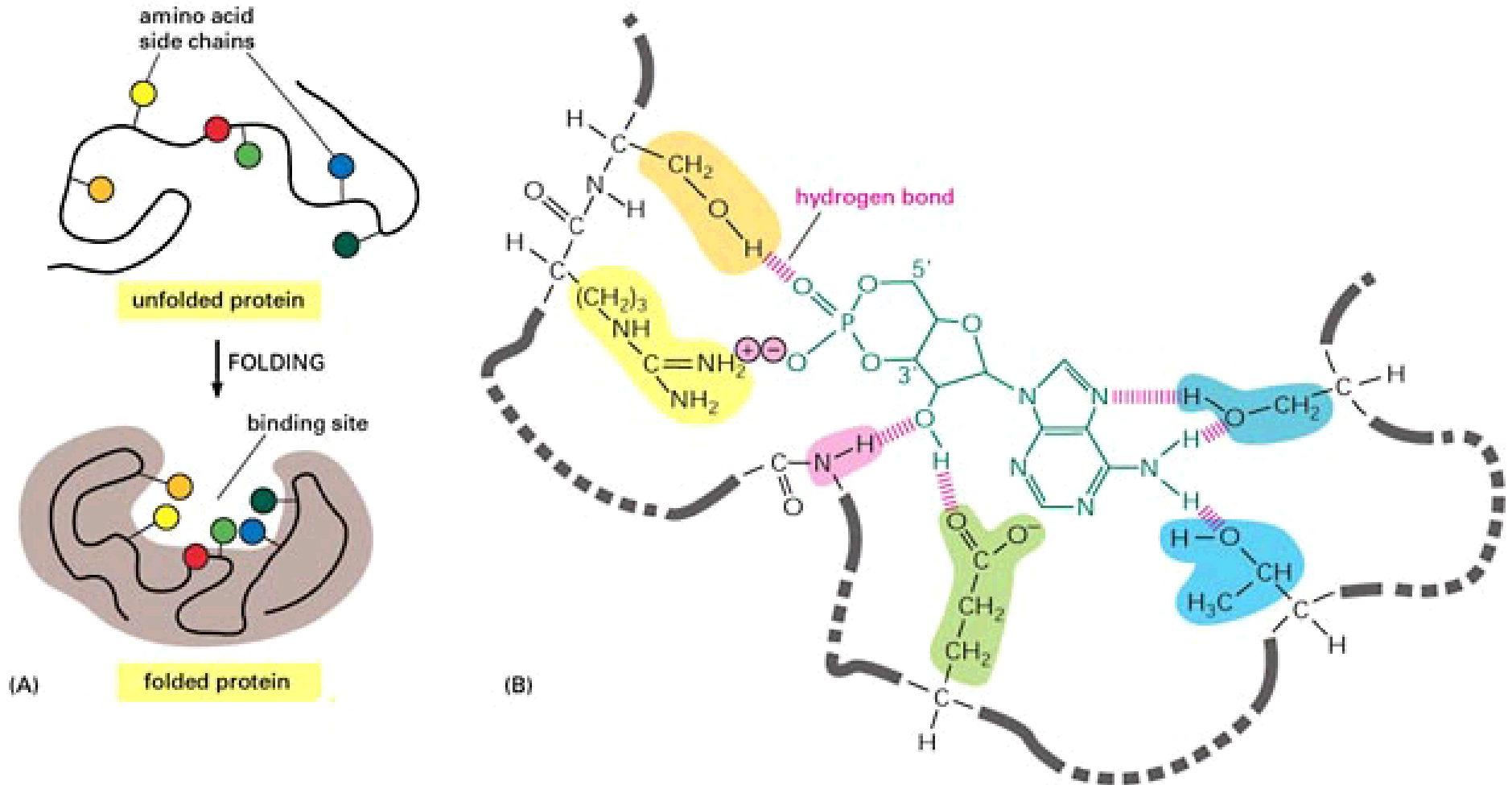
Did AlphaFold "solve" the Protein Folding Problem?

PROTEINFALTUNG VORHERSAGBAR?

Künstliche Intelligenz macht ernst im Biolabor

VON JOACHIM MÜLLER-JUNG - AKTUALISIERT AM 01.12.2020 - 16:42

Lebenswissenschaftler verneigen sich. Doch hat DeepMind mit seiner lernenden Maschine „AlphaFold" wirklich ein Jahrzehnte altes Problem der Biologie gelöst, wie behauptet wird? Eine Umfrage unter unabhängigen Experten.

Critical Assessment of protein Structure Prediction (CASP) – established 1994

- CASP is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994.
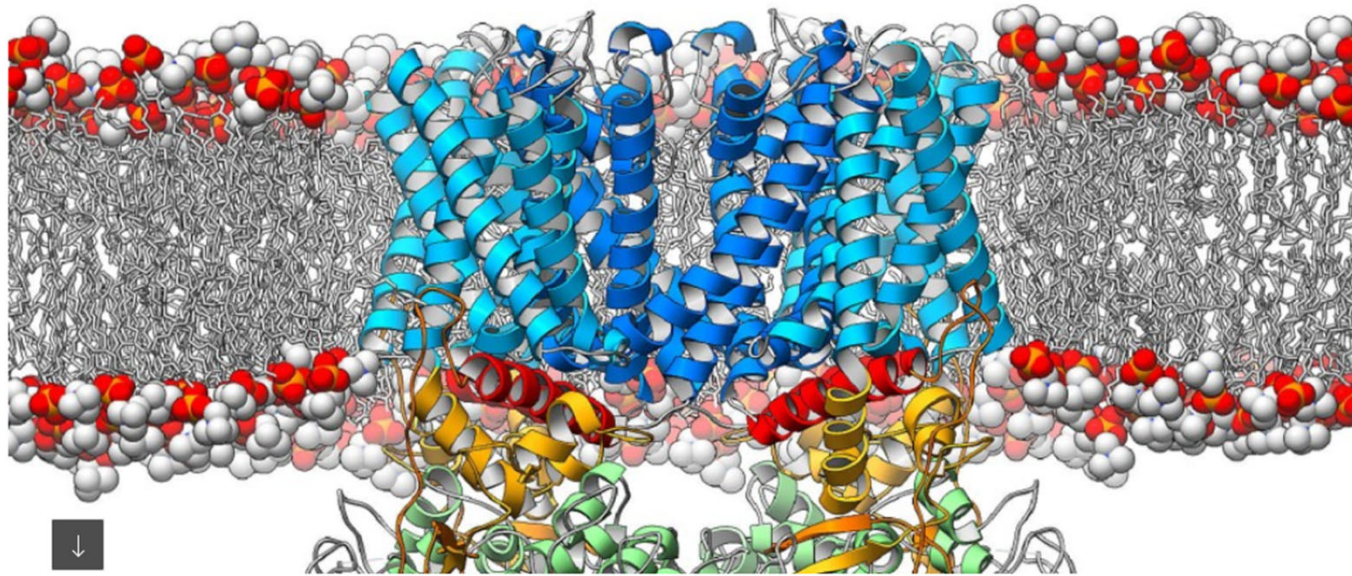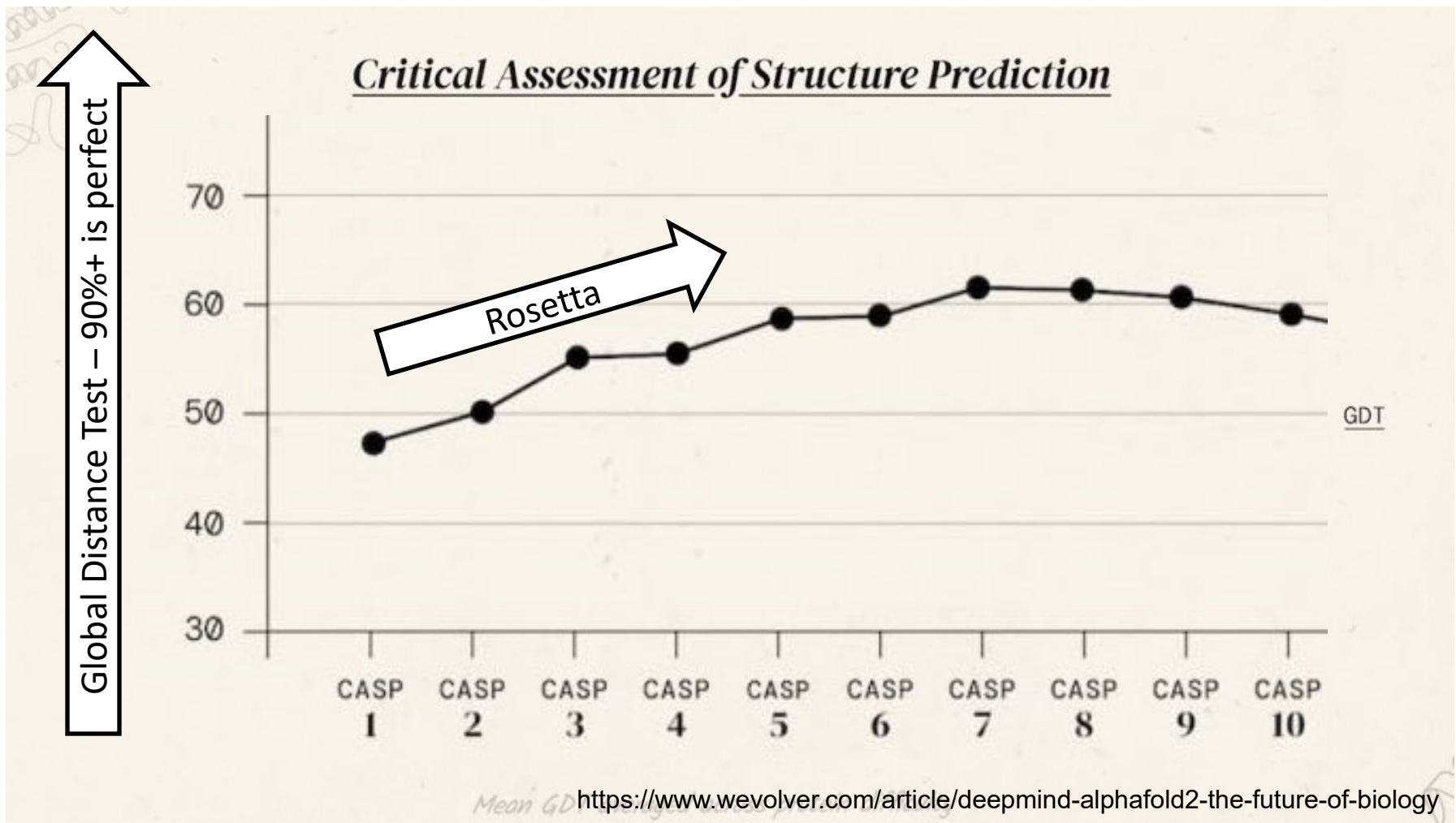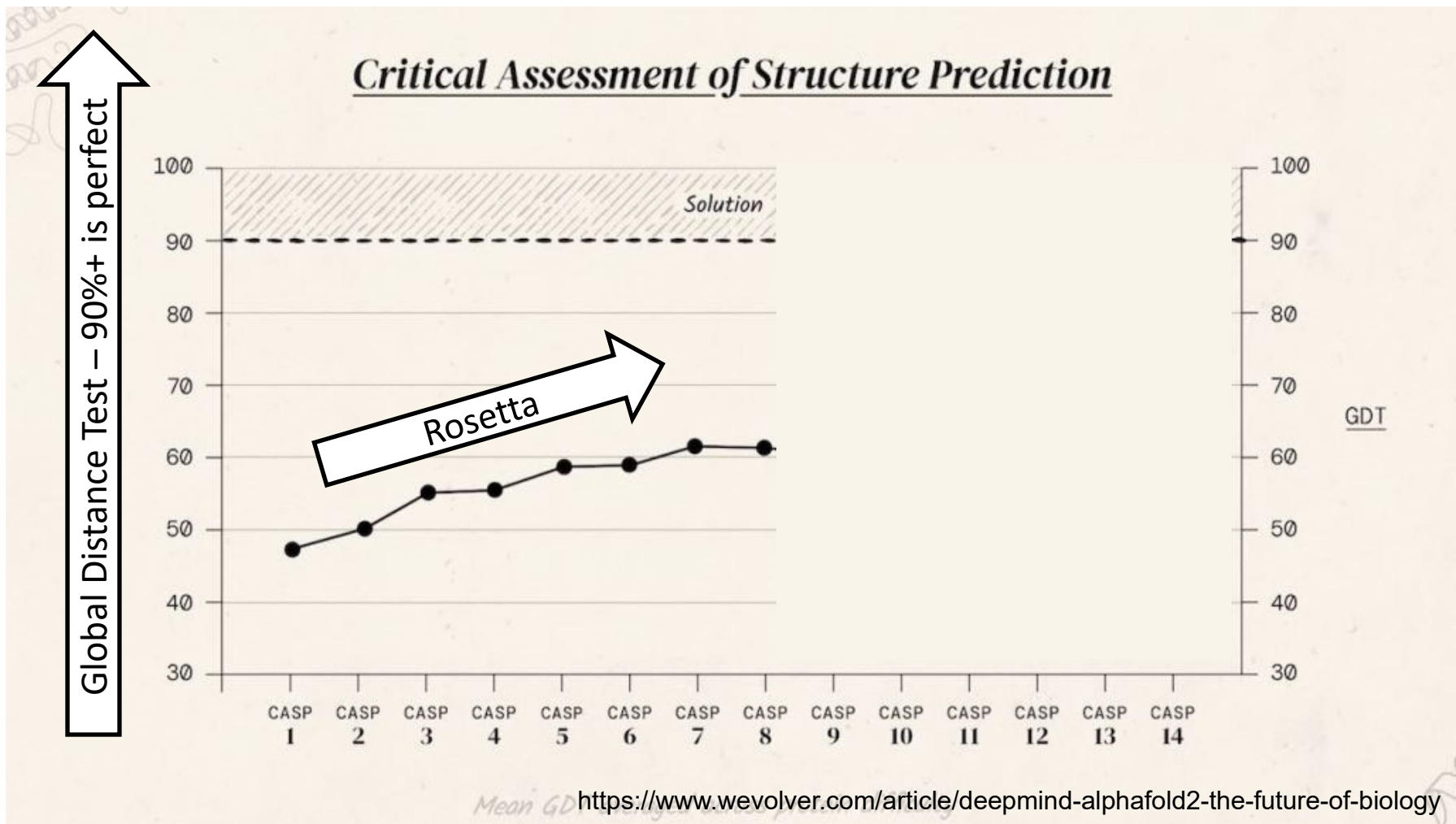
- CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users.

- Even though the primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence many view the experiment more as a "world championship" in this field of science.

- More than 100 research groups from all over the world participate in CASP on a regular basis and it is not uncommon for entire groups to suspend their other research for months while they focus on getting their servers ready for the experiment and on performing the detailed predictions.

Critical Assessment of protein Structure Prediction (CASP) – established 1994

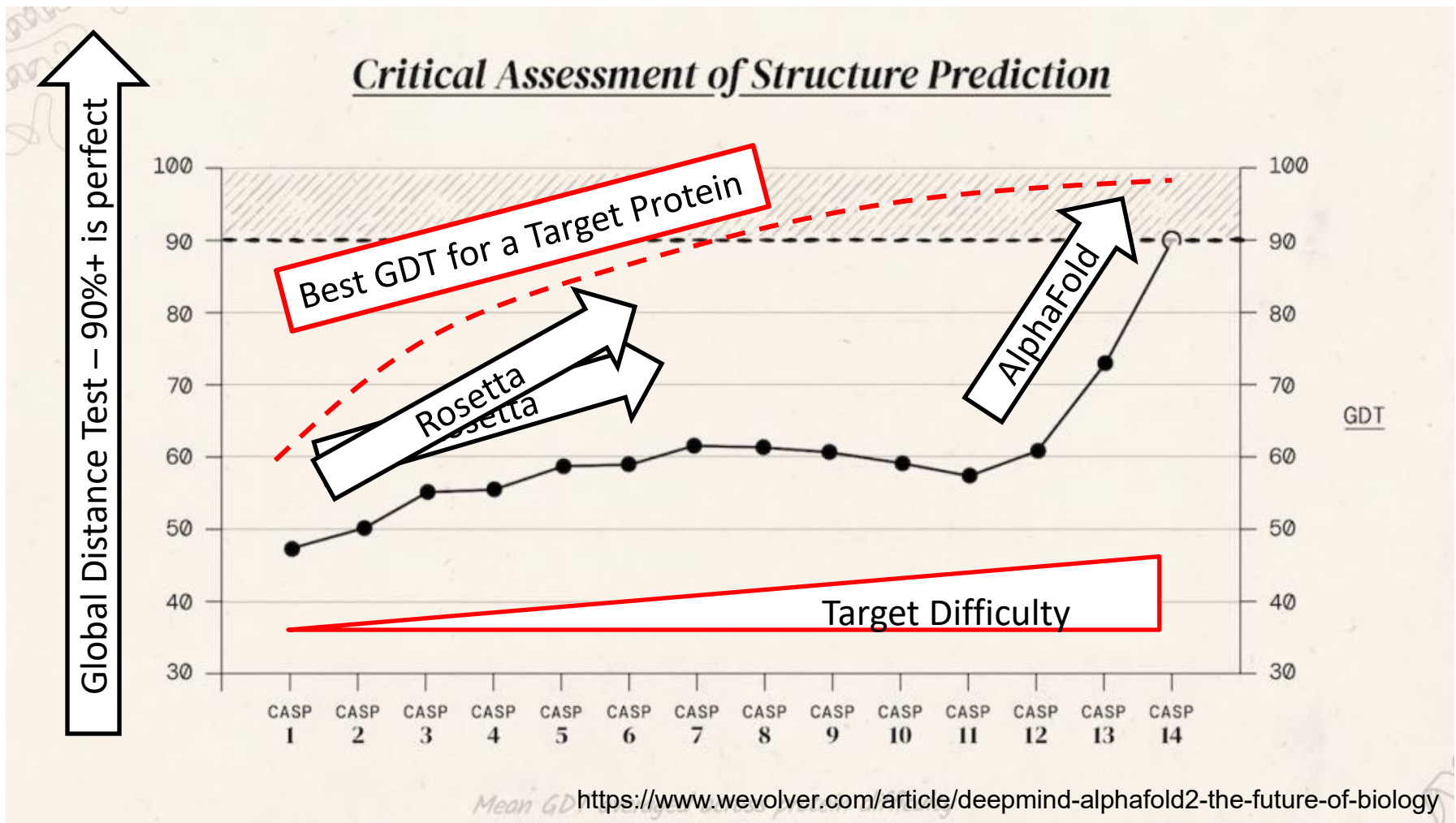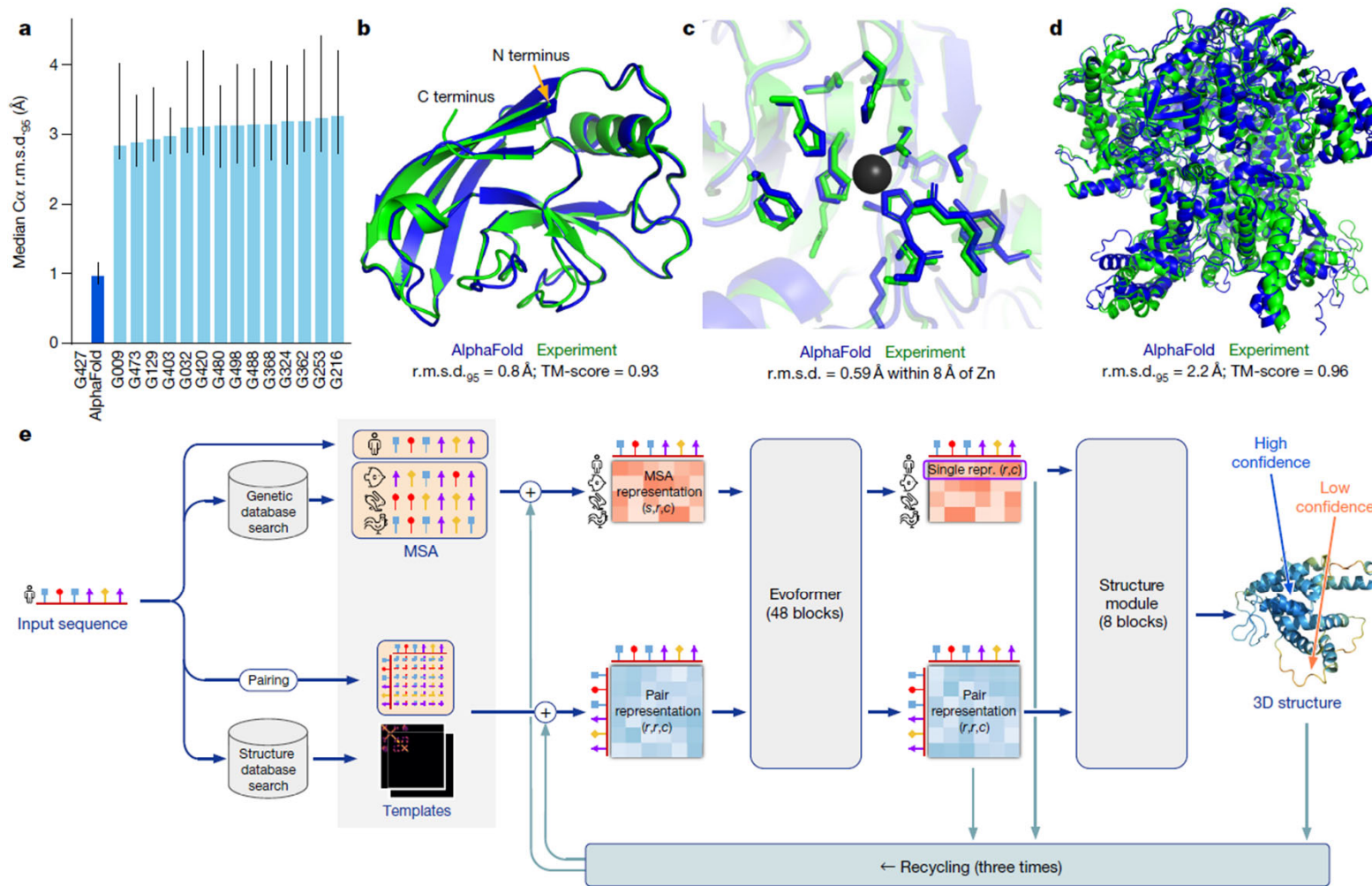# Critical Assessment of protein Structure Prediction (CASP) – established 1994

Critical Assessment of protein Structure Prediction (CASP) – established 1994
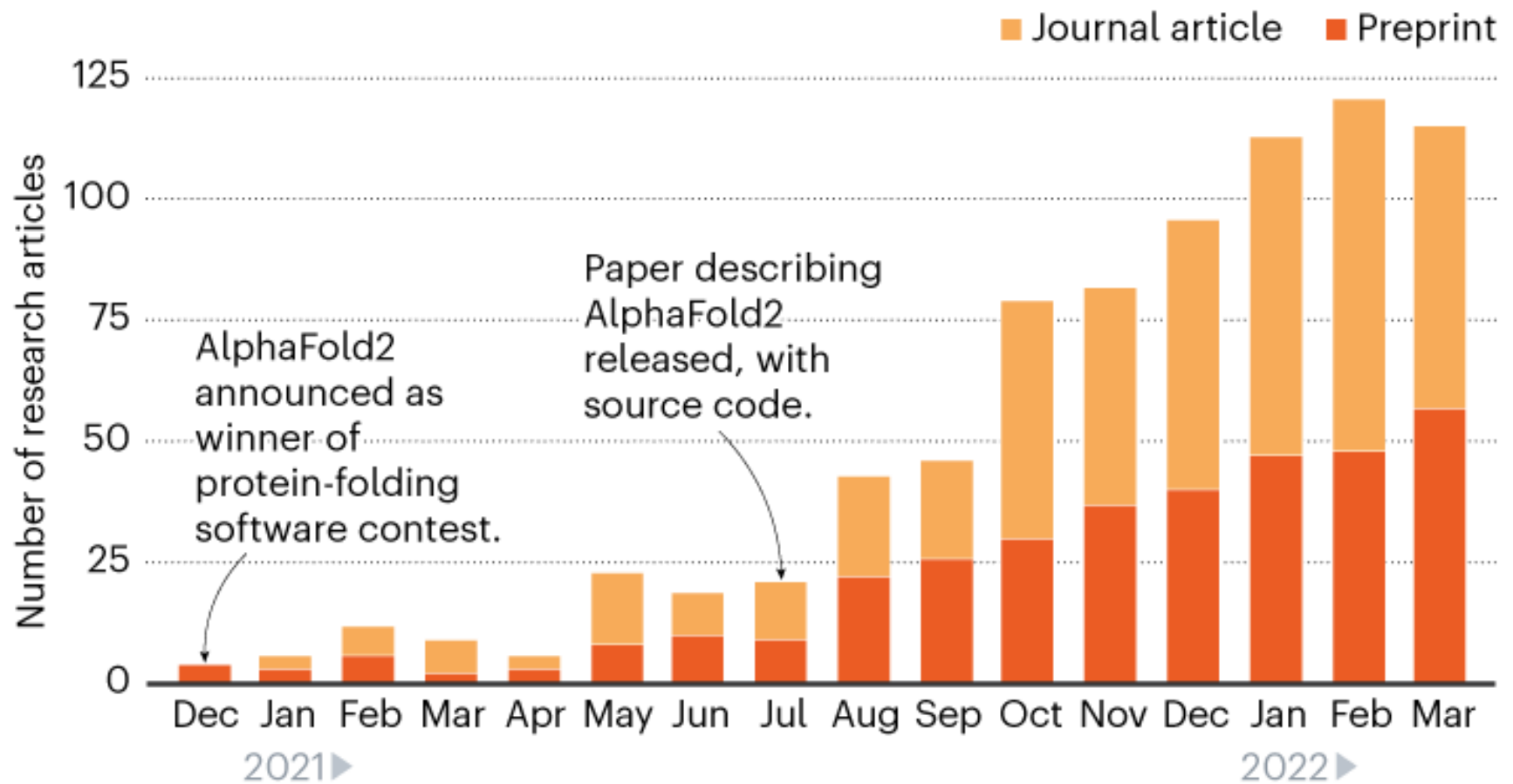

Critical Assessment of Structure Prediction

Global Distance Test – 90%+ is perfect

Best GDT for a Target Protein

Rosetta

AlphaFold

Target Difficulty

https://www.wevolver.com/article/deepmind-alphafold2-the-future-of-biology

Highly accurate protein structure prediction with AlphaFold

a

Median Cα r.m.s.d.$_{95}$ (Å)

G427 AlphaFold, G009, G473, G129, G403, G032, G420, G480, G498, G488, G368, G324, G362, G253, G216

b

N terminus
C terminus

AlphaFold   Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93

c

AlphaFold   Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

d

AlphaFold   Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

e

Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

J. Jumper, et al.; "Highly accurate protein structure prediction with AlphaFold"; *Nature*; **2021**; Vol. 596 (7873): p. 583-589.

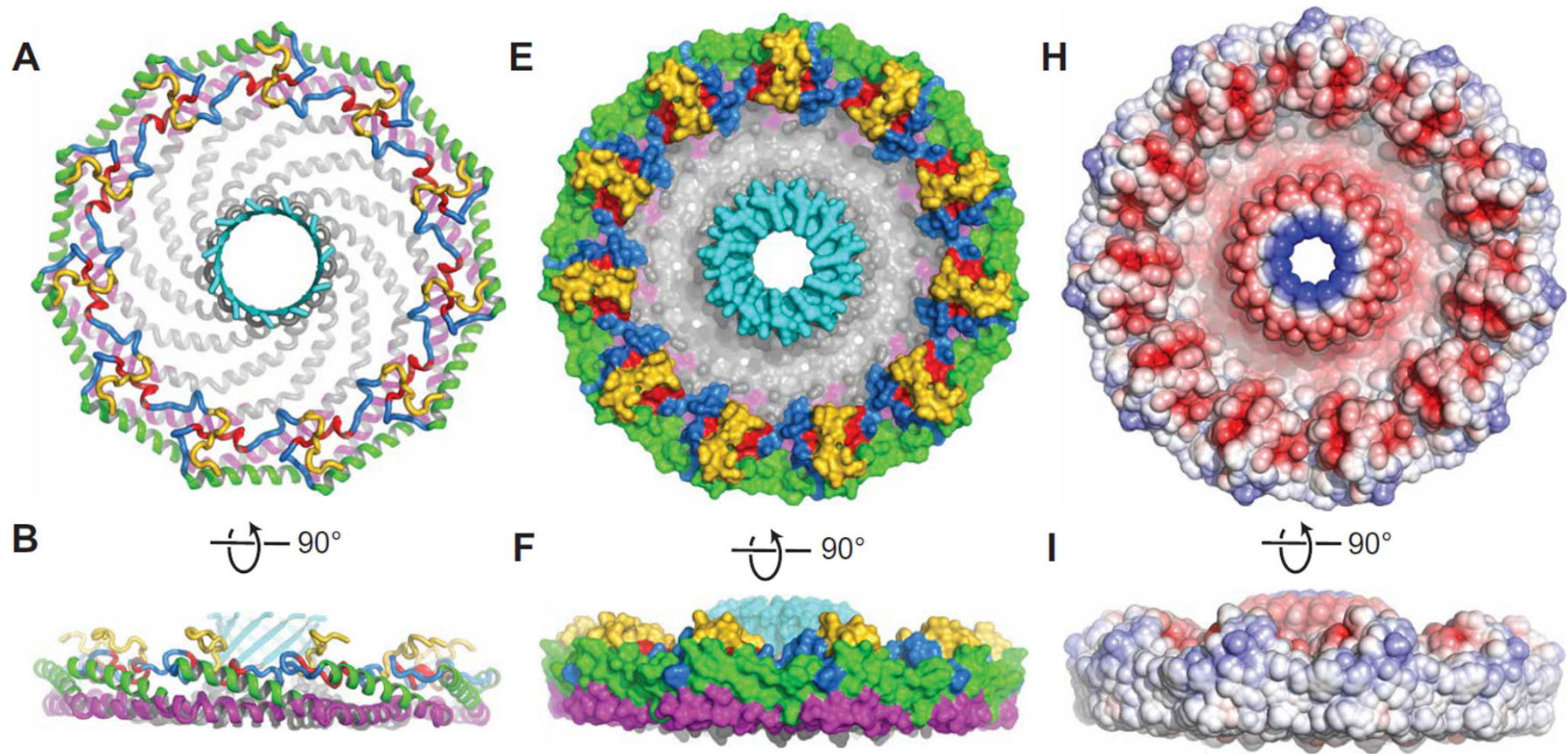# AlphaFoldMania – The number of research papers and preprints
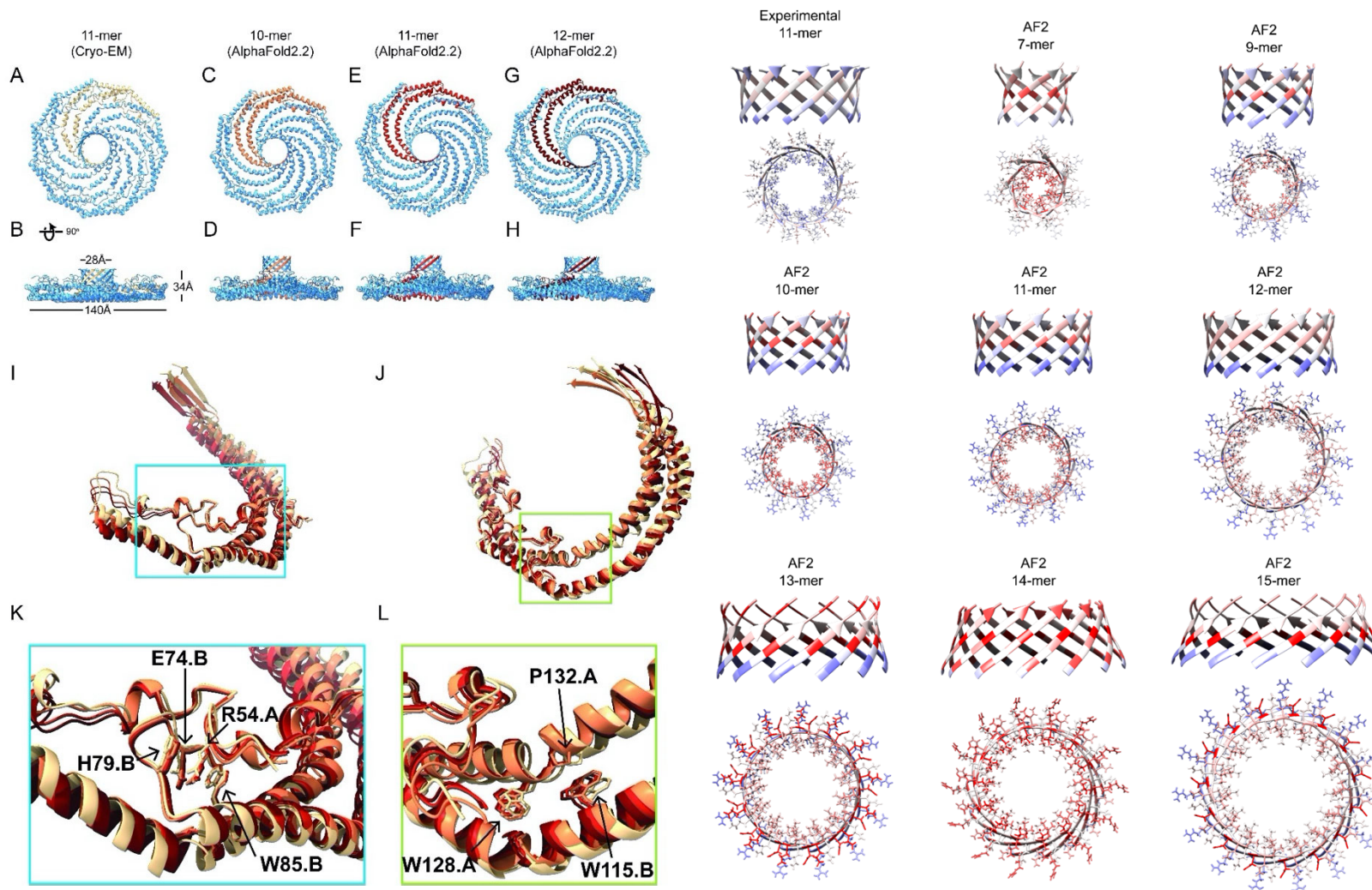

Nature, News Feature, 13 April 2022

Molecular Architecture of the Human Caveolin-1 Complex with AlphaFold2

Molecular Architecture of the Human Caveolin-1 Complex with AlphaFold2

A. Gulsevin, B. Han, J. C. Porta, H. S. McHaourab, J. Meiler and A. K. Kenworthy; "Template-free prediction of a new monotopic membrane protein fold and assembly by AlphaFold2"; *Biophys J; 2022; Vol. p.*

J. C. Porta, B. et al. "Molecular architecture of the human caveolin-1 complex"; *Science Advances*; **2022**; Vol. p.

Molecular Architecture of the Human Caveolin-1 Complex with AlphaFold2
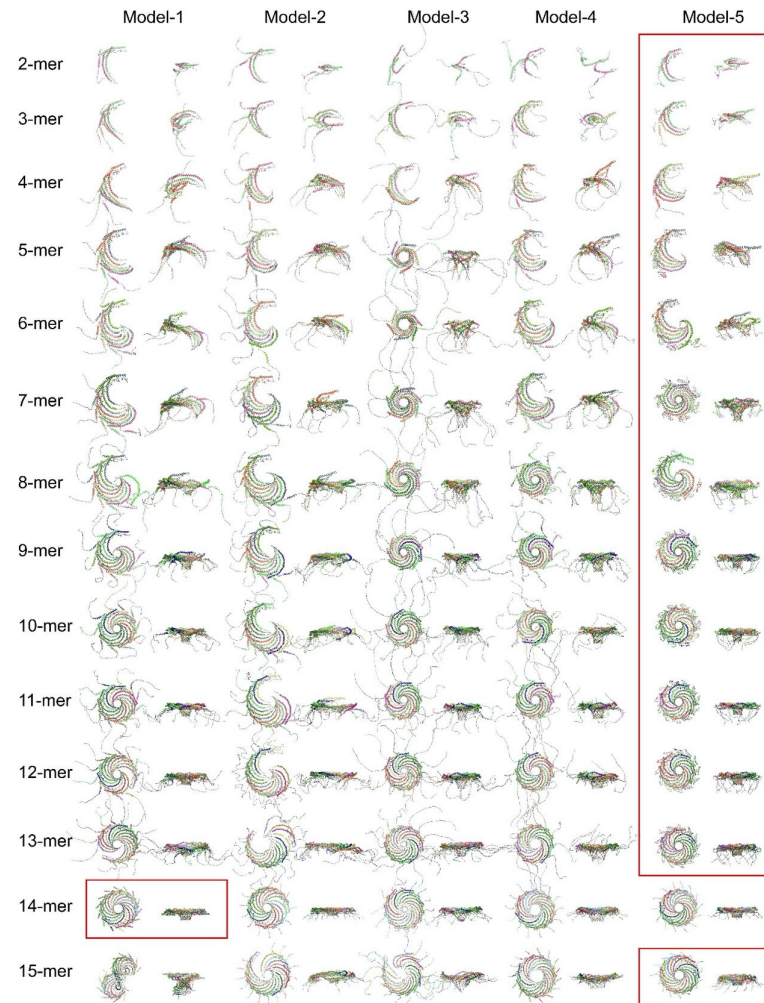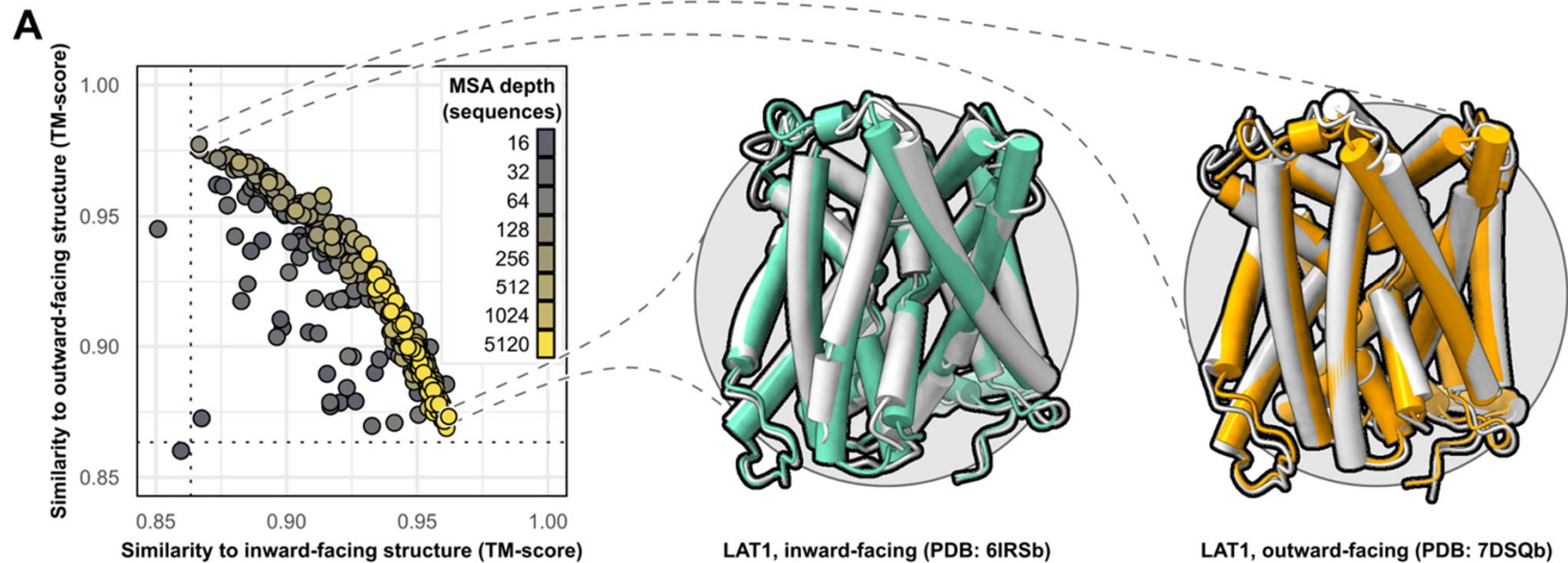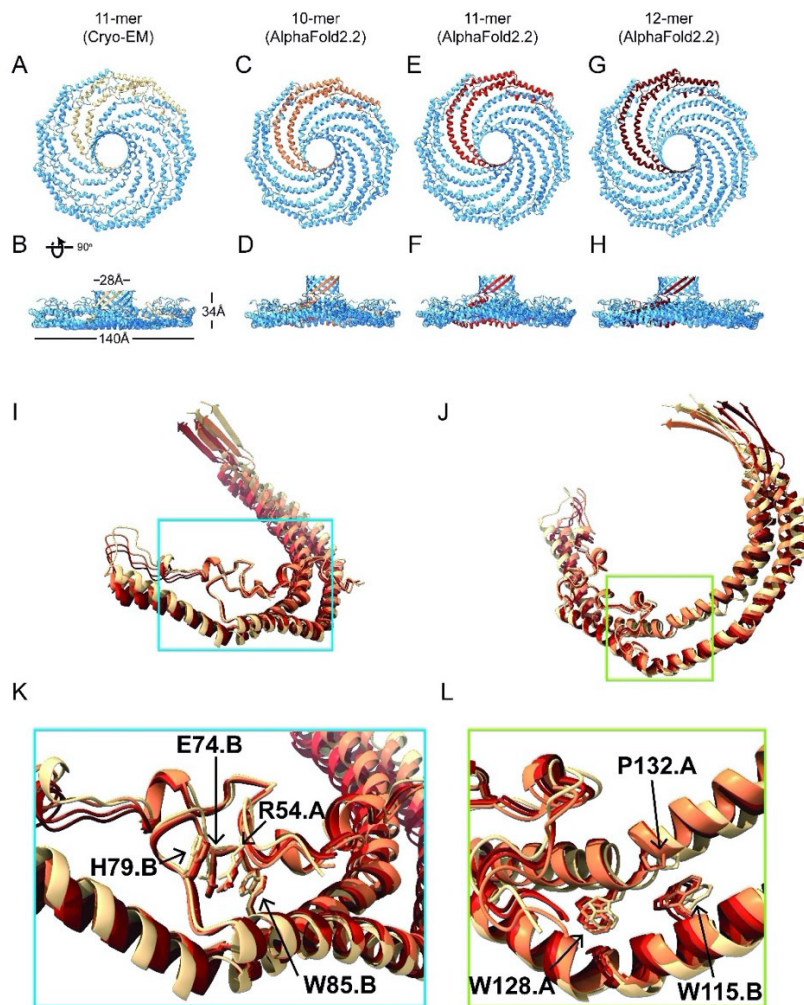

A. Gulsevin, B. Han, J. C. Porta, H. S. McHaourab, J. Meiler and A. K. Kenworthy; "Template-free prediction of a new monotopic membrane protein fold and assembly by AlphaFold2"; *Biophys J; 2022; Vol. p.*

J. C. Porta, B. et al. "Molecular architecture of the human caveolin-1 complex"; *Science Advances*; **2022**; Vol. p.

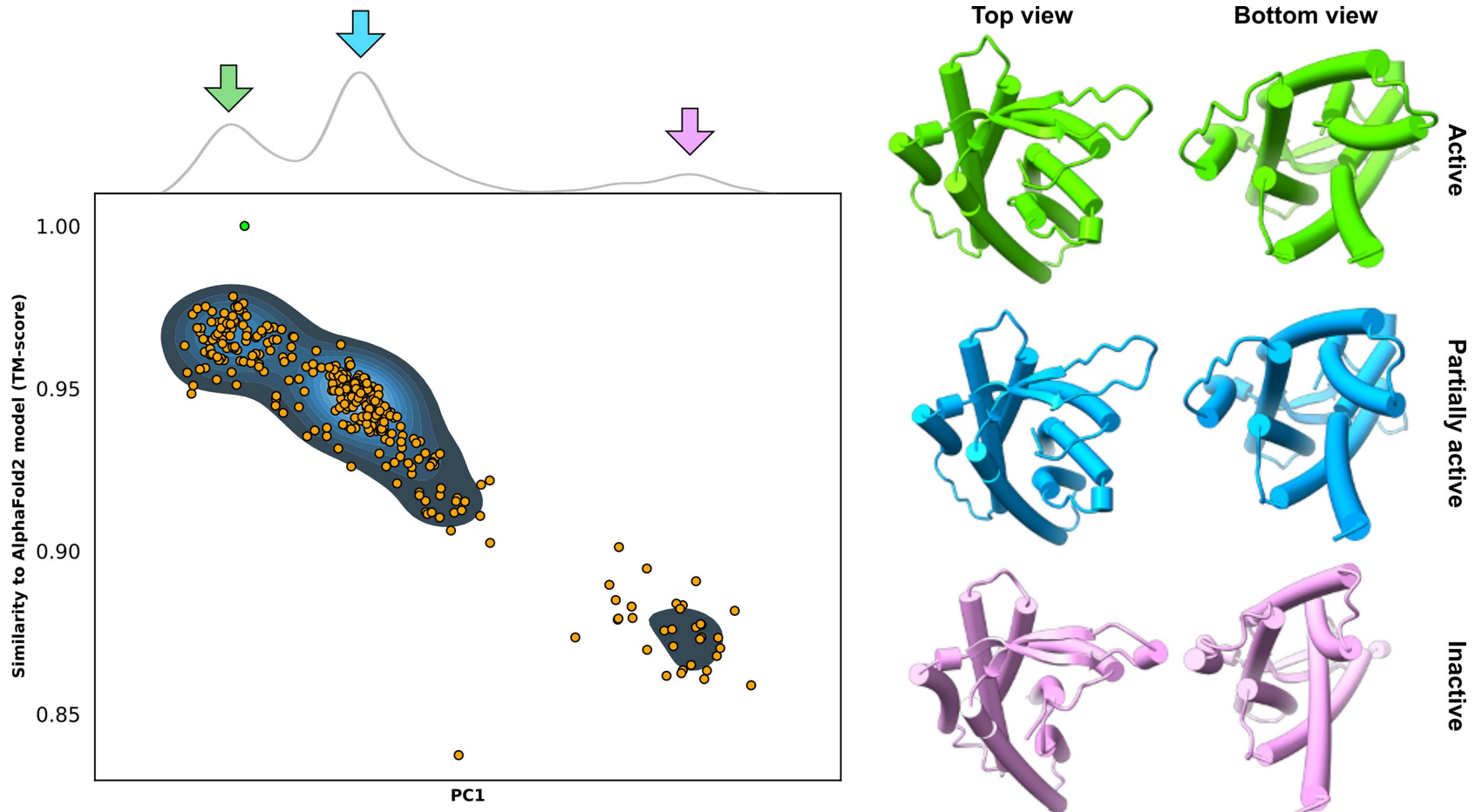# Sampling Alternative Conformational States with AlphaFold2

Figure 1. Alternative conformations of transporters and GPCRs can be predicted by AF2. (A) Representative models of the transporter LAT1 in IF and OF conformations. Experimental structures shown in gray and models shown in teal and orange.

D. Del Alamo, D. Sala, H. S. McHaourab and J. Meiler; "Sampling alternative conformational states of transporters and receptors with AlphaFold2"; *Elife*; **2022**; Vol. 11 p.
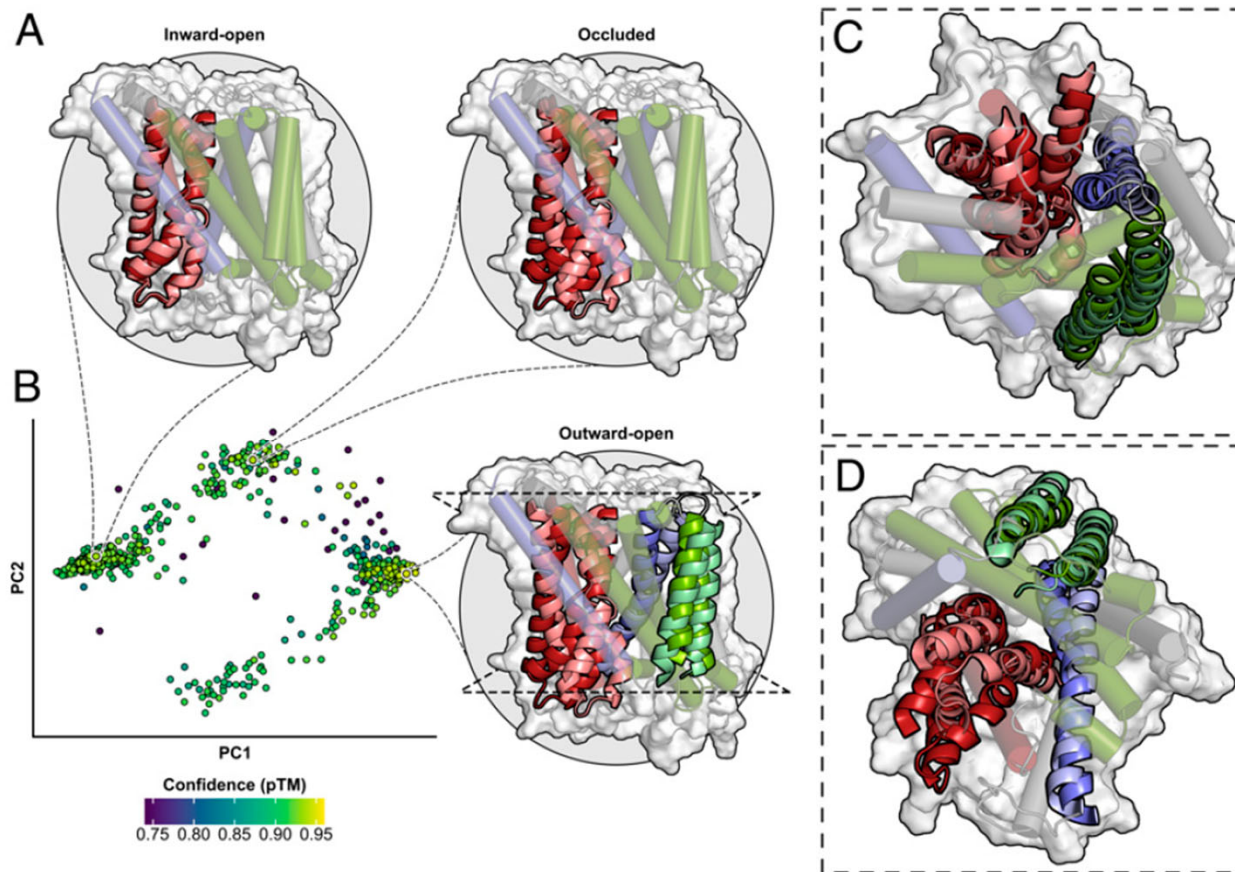
AF2 Predicted Conformations for the Adhesion GPCR ADGRG5/GPR114

D. Del Alamo, D. Sala, H. S. McHaourab and J. Meiler; "Sampling alternative conformational states of transporters and receptors with AlphaFold2"; *Elife*; **2022**; Vol. 11 p.

Integrating Limited Experimental Data: NMR, EPR, MassSpec, cryo-EM, …

D. Del Alamo, L. DeSousa, R. M. Nair, S. Rahman, J. Meiler and H. S. Mchaourab; "Integrated AlphaFold2 and DEER investigation of the conformational dynamics of a pH-dependent APC antiporter"; *Proc Natl Acad Sci U S A; 2022; Vol. 119 (34): p. e2206129119*

AlphaFold Protein Structure Database 200 Million Predicted Protein Structures

Since 1993(!) – Neural Networks for Chemists

Jure Zupan, Johann Gasteiger

Neural Networks for Chemists

An Introduction

VCH
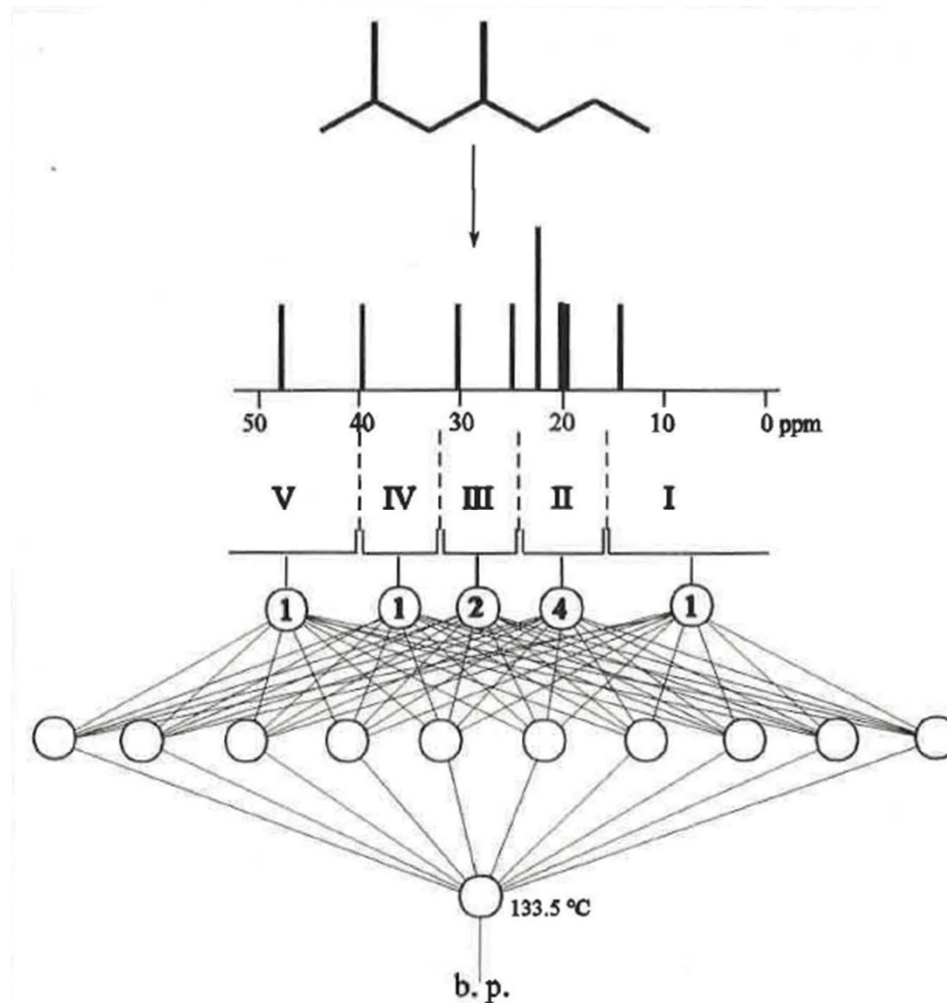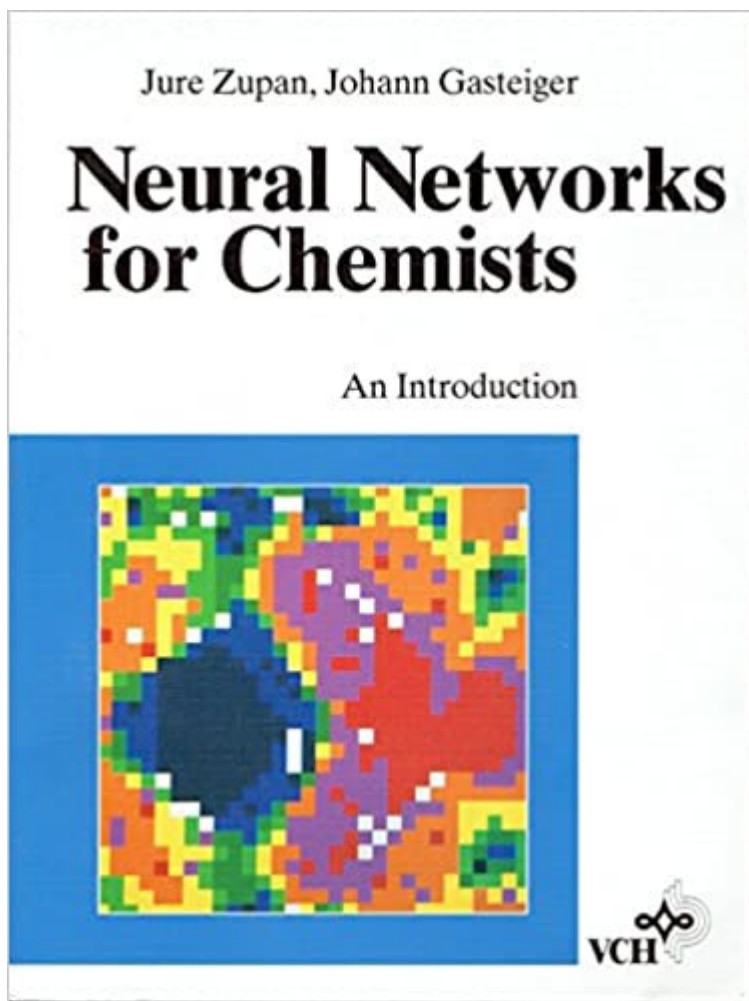
J. Meiler and R. Meusinger "Use of Neural Networks to Determine Properties of Alkanes from their 13C-NMR Spectra" in *Software - Entwicklung in der Chemie*; Gasteiger, J., Ed. Gesellschaft Deutscher Chemiker: Frankfurt am Main; **1995**; Vol. 10: p. 259-263.
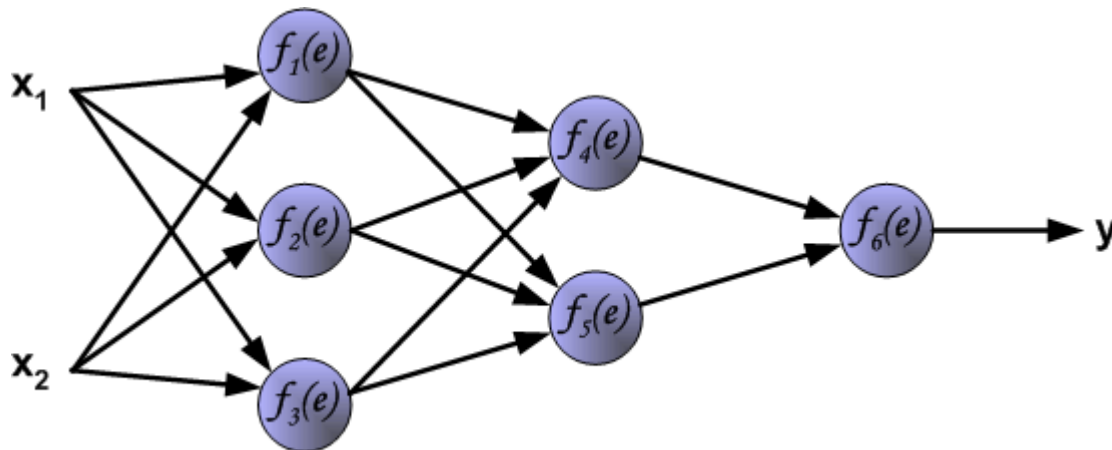
Back Propagation of Errors is the most popular Training Algorithm

- Teaching process of multi-layer neural network employing *backpropagation* algorithm. To illustrate this process, consider the three layer neural network with two inputs and one output:
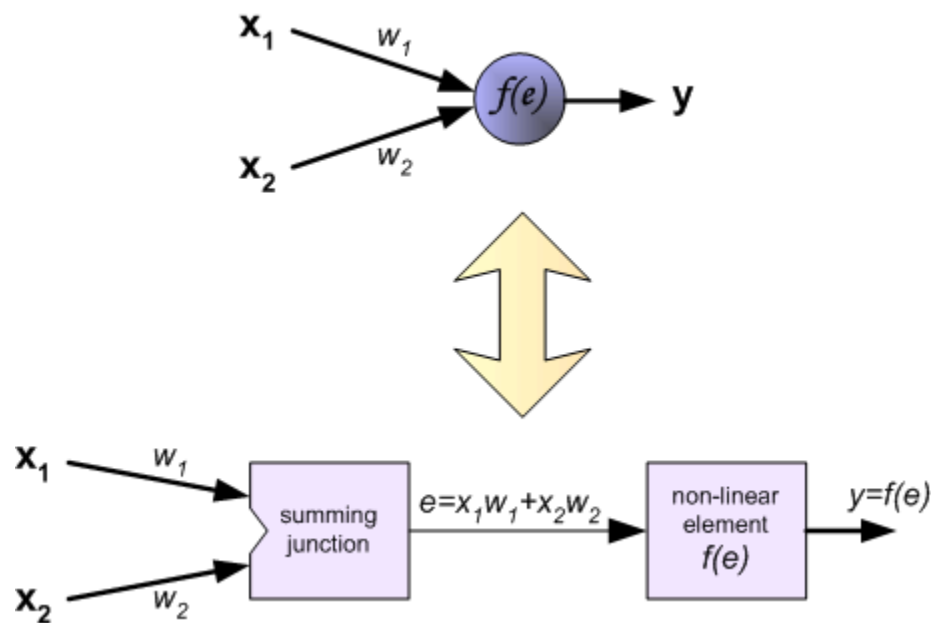
Back Propagation of Errors is the most popular Training Algorithm

- Each neuron is composed of two units. First unit adds products of weights coefficients and input signals. The second unit realizes nonlinear function, called neuron activation function. Signal e is summed weighted input signal, and $y = f(e)$ is output signal of nonlinear element. Signal y is also output signal of neuron:
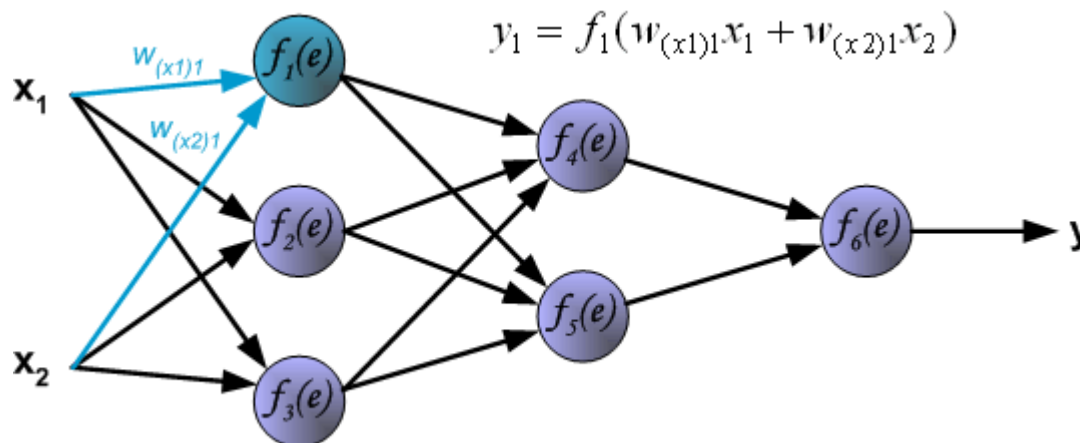

http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

Back Propagation of Errors is the most popular Training Algorithm

- To teach the neural network we need training data set. The training data set consists of input signals (x_1 and x_2) assigned with corresponding target (desired output) z. The network training is an iterative process. In each iteration weights coefficients of nodes are modified using new data from training data set. Modification is calculated using algorithm described below: Each teaching step starts with forcing both input signals from training set. After this stage we can determine output signals values for each neuron in each network layer. Pictures below illustrate how signal is propagating through the network, Symbols $w_{(xm)n}$ represent weights of connections between network input x_m and neuron n in input layer. Symbols y_n represents output signal of neuron n.

$$y_1 = f_1(w_{(x1)1}x_1 + w_{(x2)1}x_2)$$

Back Propagation of Errors is the most popular Training Algorithm

- To teach the neural network we need training data set. The training data set consists of input signals (x_1 and x_2) assigned with corresponding target (desired output) z. The network training is an iterative process. In each iteration weights coefficients of nodes are modified using new data from training data set. Modification is calculated using algorithm described below: Each teaching step starts with forcing both input signals from training set. After this stage we can determine output signals values for each neuron in each network layer. Pictures below illustrate how signal is propagating through the network, Symbols $w_{(xm)n}$ represent weights of connections between network input x_m and neuron n in input layer. Symbols y_n represents output signal of neuron n.
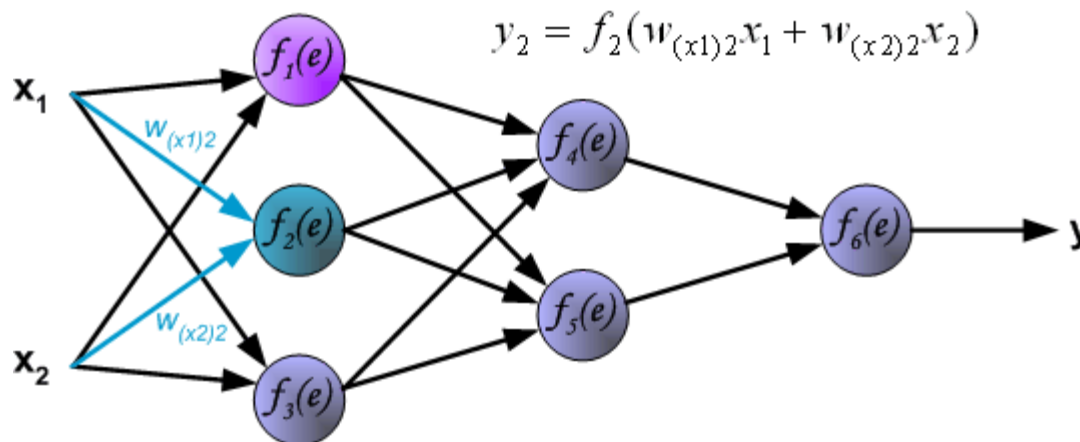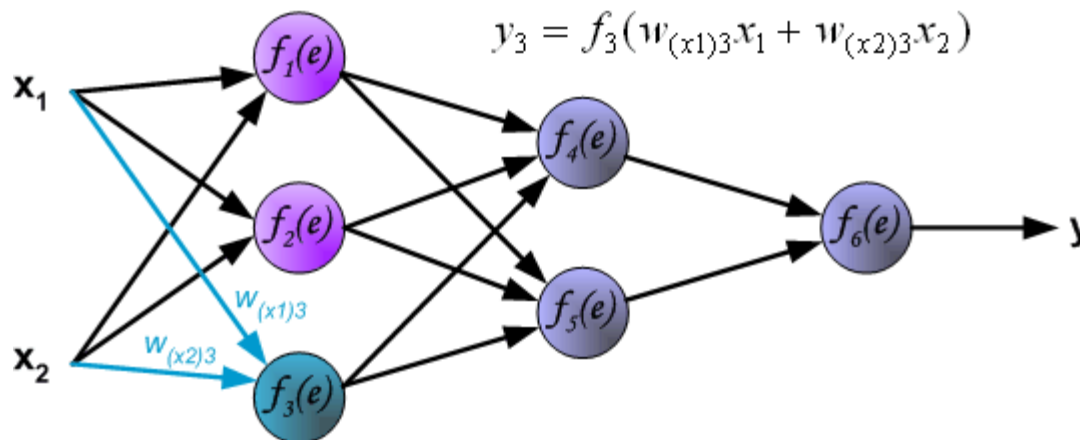
$$y_2 = f_2(w_{(x1)2}x_1 + w_{(x2)2}x_2)$$

Back Propagation of Errors is the most popular Training Algorithm

▪ To teach the neural network we need training data set. The training data set consists of input signals (x_1 and x_2) assigned with corresponding target (desired output) z. The network training is an iterative process. In each iteration weights coefficients of nodes are modified using new data from training data set. Modification is calculated using algorithm described below: Each teaching step starts with forcing both input signals from training set. After this stage we can determine output signals values for each neuron in each network layer. Pictures below illustrate how signal is propagating through the network, Symbols $w_{(xm)n}$ represent weights of connections between network input x_m and neuron n in input layer. Symbols y_n represents output signal of neuron n.
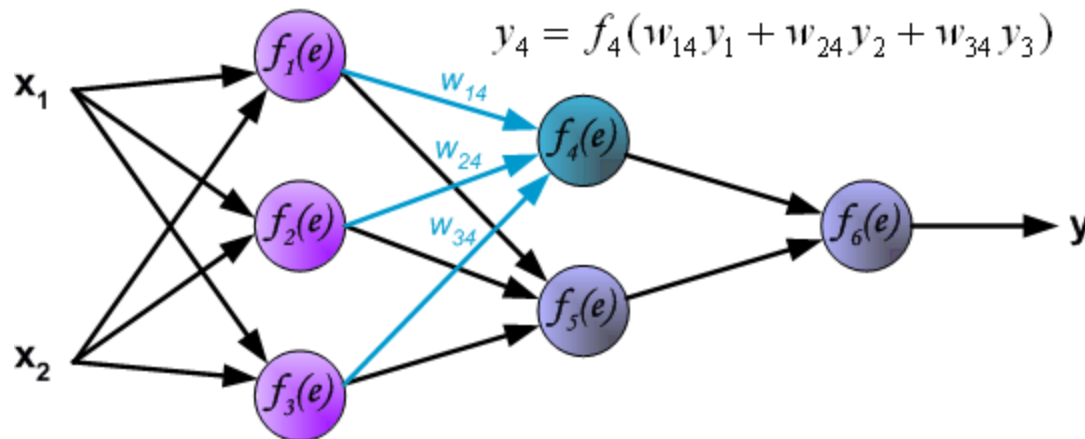
$$y_3 = f_3\left(w_{(x1)3}x_1 + w_{(x2)3}x_2\right)$$

Back Propagation of Errors is the most popular Training Algorithm

- Propagation of signals through the hidden layer. Symbols w_{mn} represent weights of connections between output of neuron m and input of neuron n in the next layer.

$$y_4 = f_4(w_{14}y_1 + w_{24}y_2 + w_{34}y_3)$$

Back Propagation of Errors is the most popular Training Algorithm

- Propagation of signals through the hidden layer. Symbols w_{mn} represent weights of connections between output of neuron m and input of neuron n in the next layer.
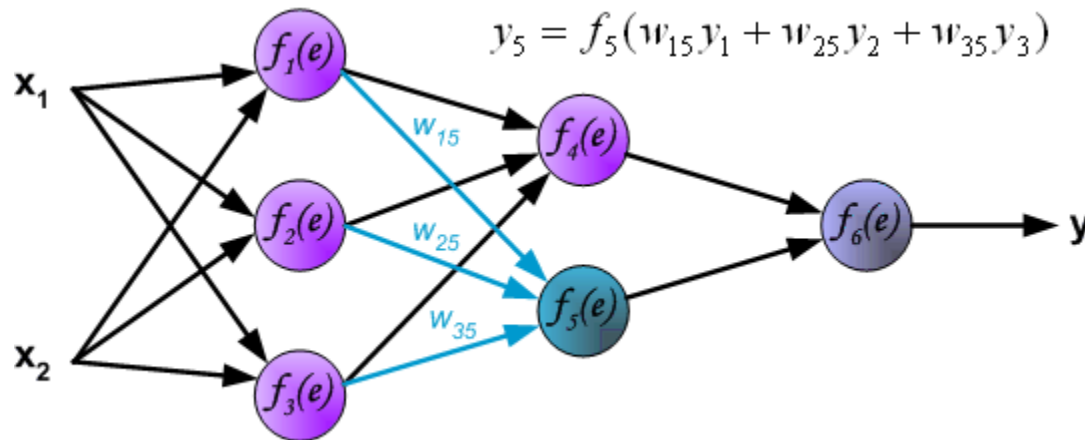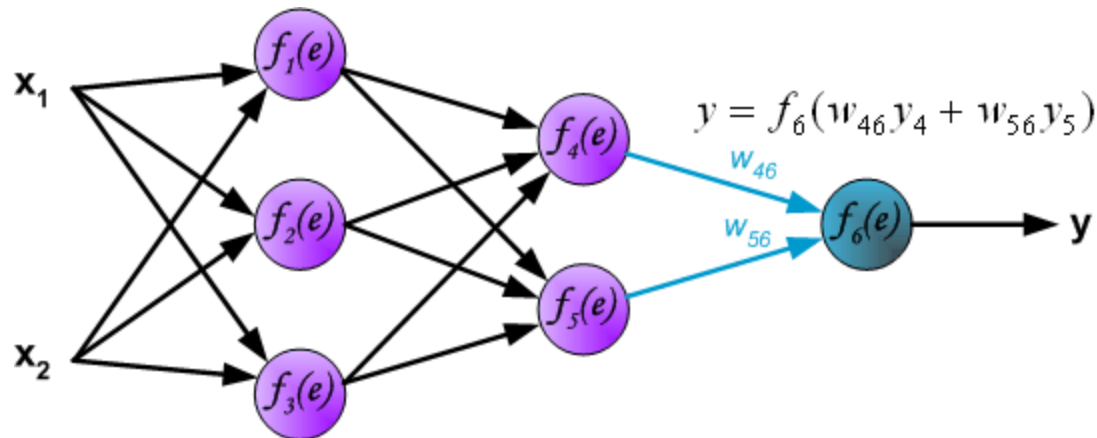
$$y_5 = f_5(w_{15}y_1 + w_{25}y_2 + w_{35}y_3)$$

Back Propagation of Errors is the most popular Training Algorithm
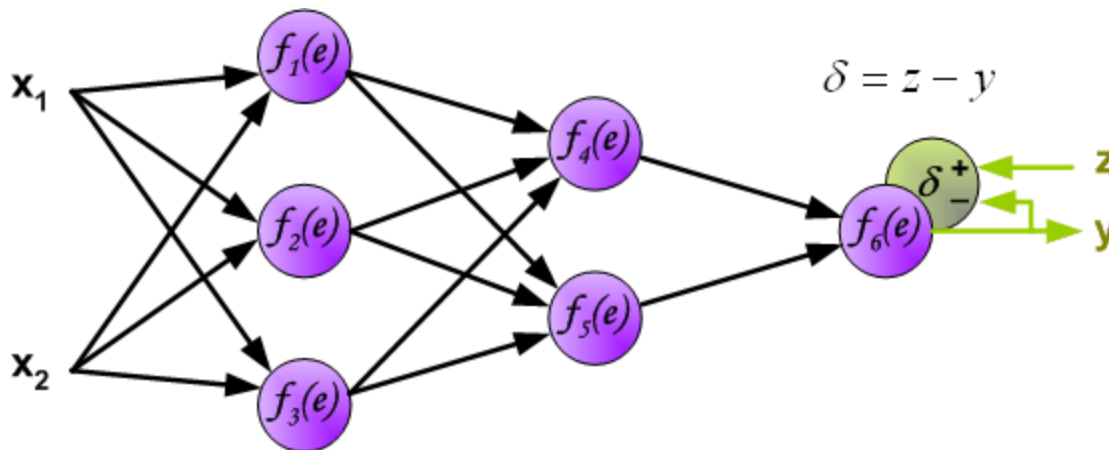
- Propagation of signals through the output layer.

$$y = f_6(w_{46}y_4 + w_{56}y_5)$$

Back Propagation of Errors is the most popular Training Algorithm

- In the next algorithm step the output signal of the network *y* is compared with the desired output value (the target), which is found in training data set. The difference is called error signal *d* of output layer neuron.
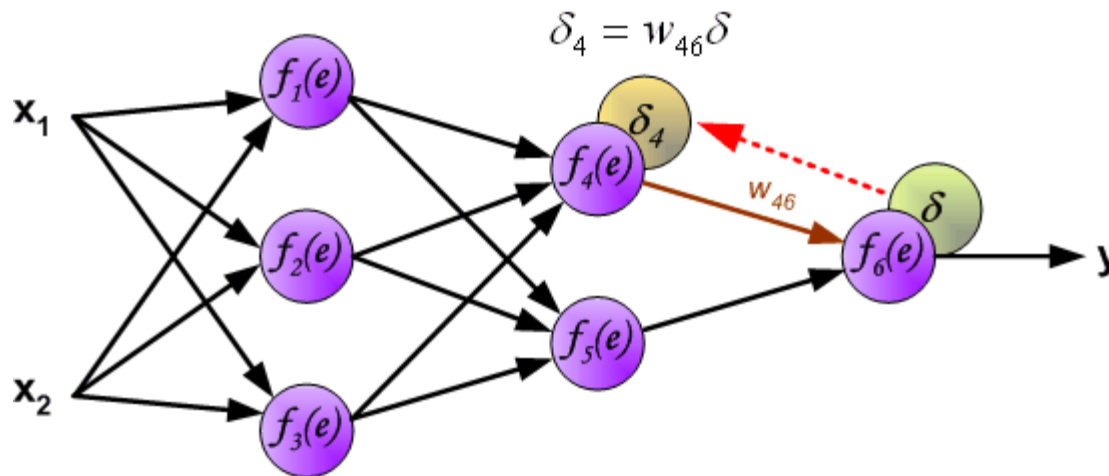
$$\delta = z - y$$

Back Propagation of Errors is the most popular Training Algorithm

- It is impossible to compute error signal for internal neurons directly, because output values of these neurons are unknown. The idea is to propagate error signal d (computed in single teaching step) back to all neurons, which output signals were input for discussed neuron.

$$\delta_4 = w_{46}\delta$$

Back Propagation of Errors is the most popular Training Algorithm

- It is impossible to compute error signal for internal neurons directly, because output values of these neurons are unknown. The idea is to propagate error signal d (computed in single teaching step) back to all neurons, which output signals were input for discussed neuron.
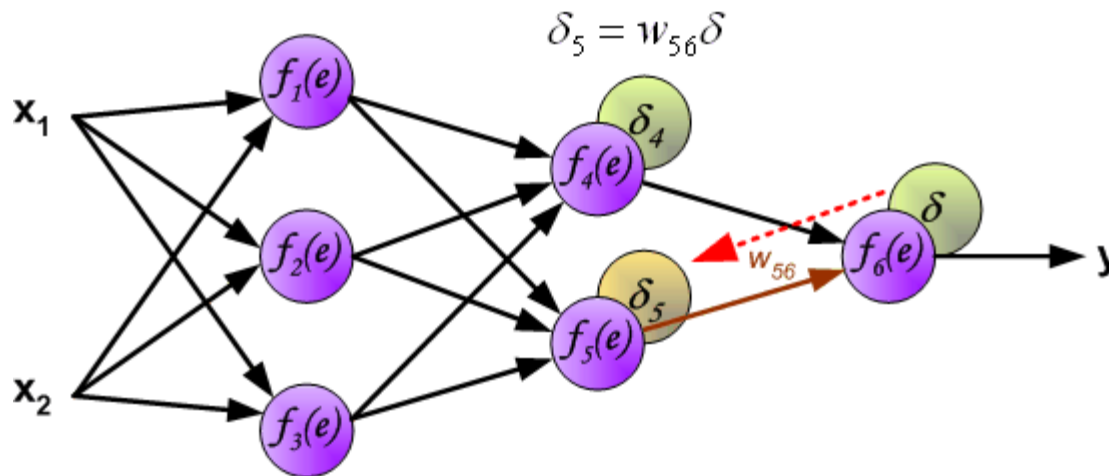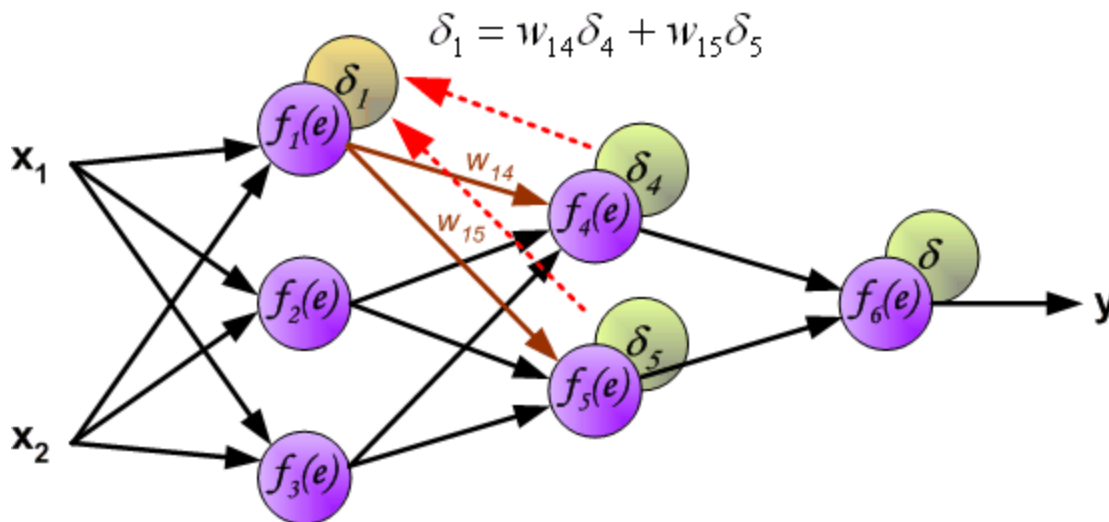
$$\delta_5 = w_{56}\delta$$

Back Propagation of Errors is the most popular Training Algorithm

- The weights' coefficients w_{mn} used to propagate errors back are equal to this used during computing output value. Only the direction of data flow is changed - signals are propagated from output to inputs one after the other. This technique is used for all network layers. If propagated errors came from few neurons they are added. The illustration is below:

$$\delta_1 = w_{14}\delta_4 + w_{15}\delta_5$$

Back Propagation of Errors is the most popular Training Algorithm

- The weights' coefficients w_{mn} used to propagate errors back are equal to this used during computing output value. Only the direction of data flow is changed - signals are propagated from output to inputs one after the other. This technique is used for all network layers. If propagated errors came from few neurons they are added. The illustration is below:
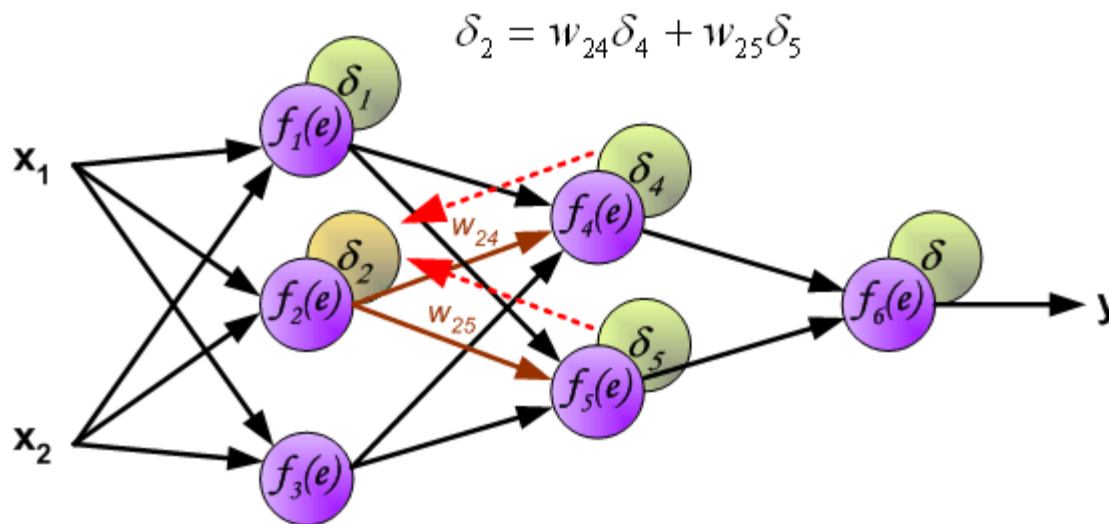

$$\delta_2 = w_{24}\delta_4 + w_{25}\delta_5$$

http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

Back Propagation of Errors is the most popular Training Algorithm

- The weights' coefficients w_{mn} used to propagate errors back are equal to this used during computing output value. Only the direction of data flow is changed - signals are propagated from output to inputs one after the other. This technique is used for all network layers. If propagated errors came from few neurons they are added. The illustration is below:
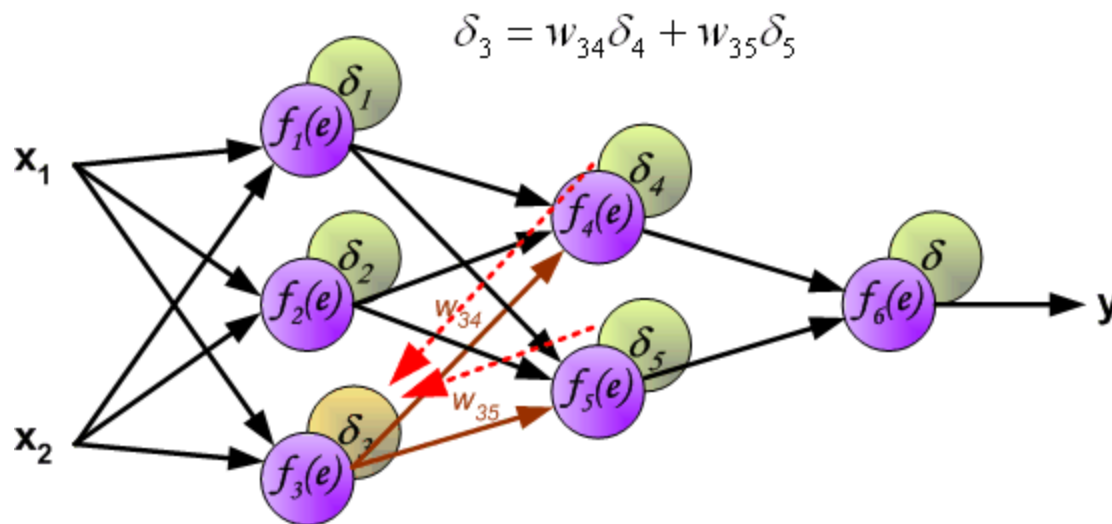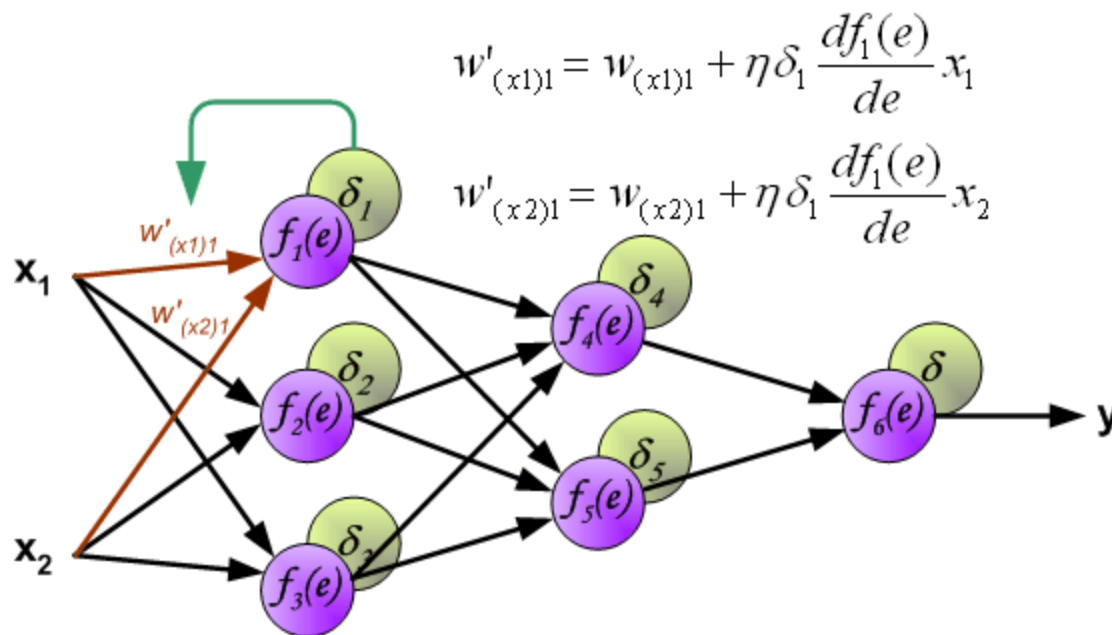

$$\delta_3 = w_{34}\delta_4 + w_{35}\delta_5$$

Back Propagation of Errors is the most popular Training Algorithm

- When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified. In formulas below *df(e)/de* represents derivative of neuron activation function (which weights are modified).

$$w'_{(x1)1} = w_{(x1)1} + \eta \delta_1 \frac{df_1(e)}{de} x_1$$

$$w'_{(x2)1} = w_{(x2)1} + \eta \delta_1 \frac{df_1(e)}{de} x_2$$

Back Propagation of Errors is the most popular Training Algorithm

- When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified. In formulas below *df(e)/de* represents derivative of neuron activation function (which weights are modified).
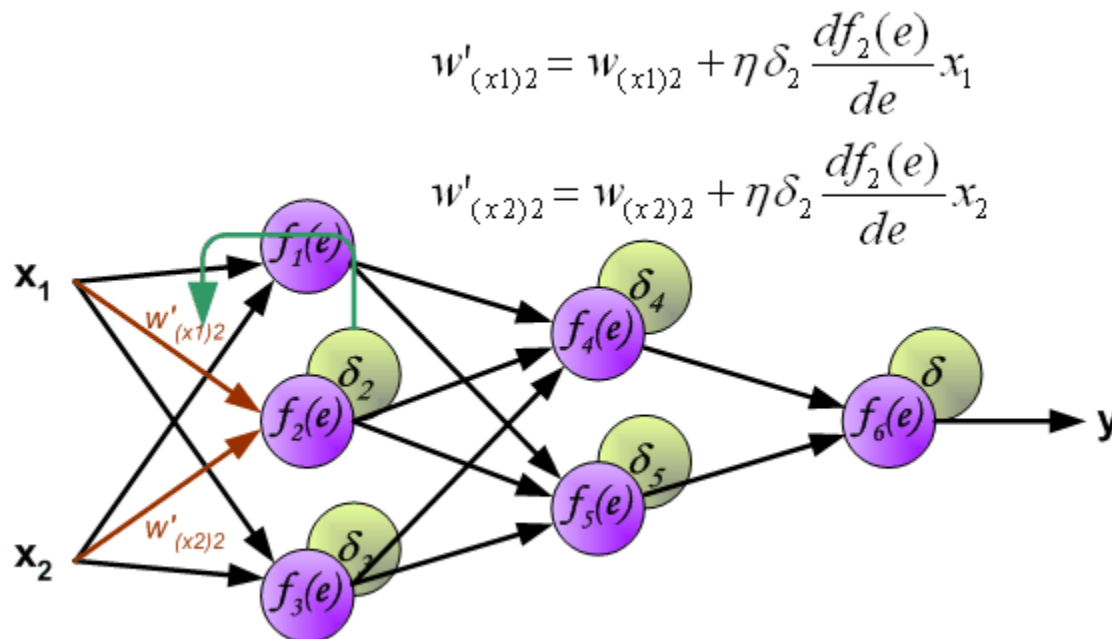
$$w'_{(x1)2} = w_{(x1)2} + \eta \delta_2 \frac{df_2(e)}{de} x_1$$

$$w'_{(x2)2} = w_{(x2)2} + \eta \delta_2 \frac{df_2(e)}{de} x_2$$

http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

Back Propagation of Errors is the most popular Training Algorithm

- When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified. In formulas below *df(e)/de* represents derivative of neuron activation function (which weights are modified).
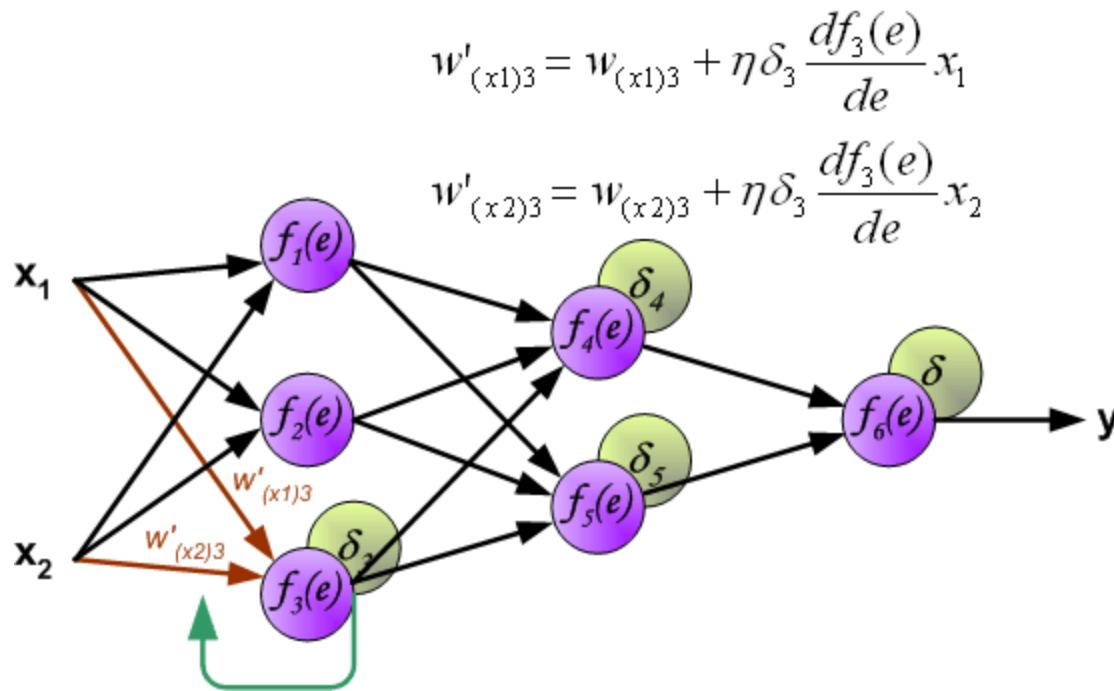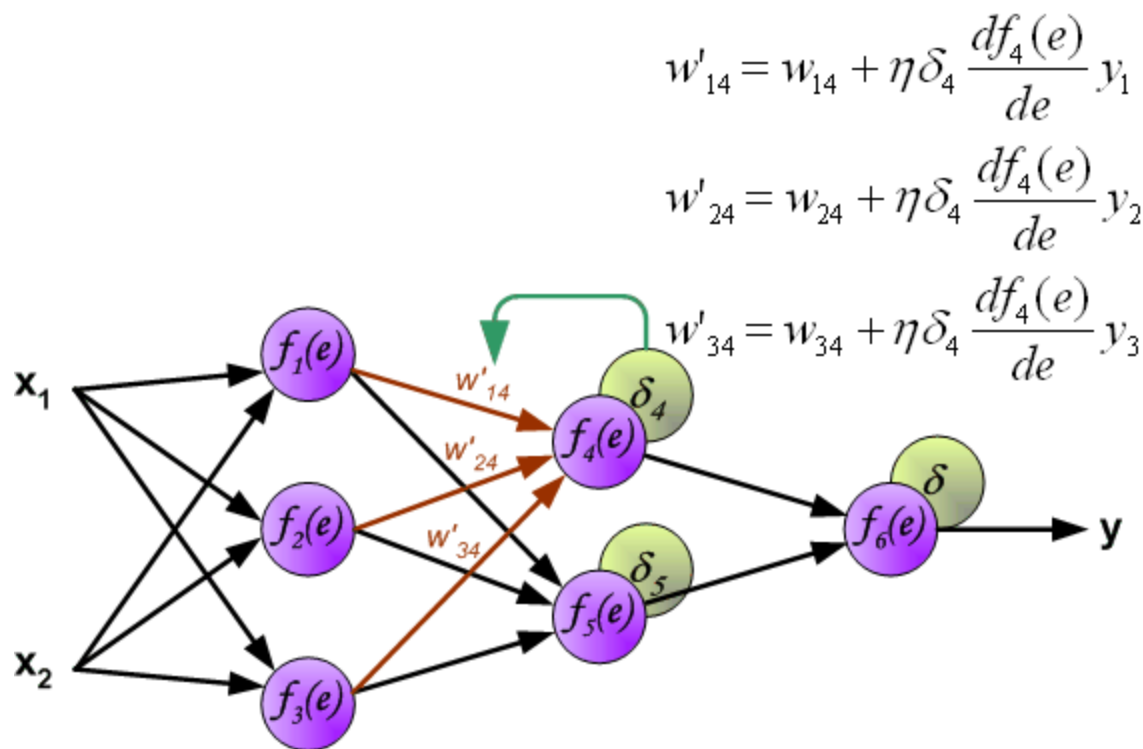
$$w'_{(x1)3} = w_{(x1)3} + \eta \delta_3 \frac{df_3(e)}{de} x_1$$

$$w'_{(x2)3} = w_{(x2)3} + \eta \delta_3 \frac{df_3(e)}{de} x_2$$

http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

Back Propagation of Errors is the most popular Training Algorithm

- When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified.

$$w'_{14} = w_{14} + \eta \delta_4 \frac{df_4(e)}{de} y_1$$

$$w'_{24} = w_{24} + \eta \delta_4 \frac{df_4(e)}{de} y_2$$

$$w'_{34} = w_{34} + \eta \delta_4 \frac{df_4(e)}{de} y_3$$

Back Propagation of Errors is the most popular Training Algorithm

- When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified.
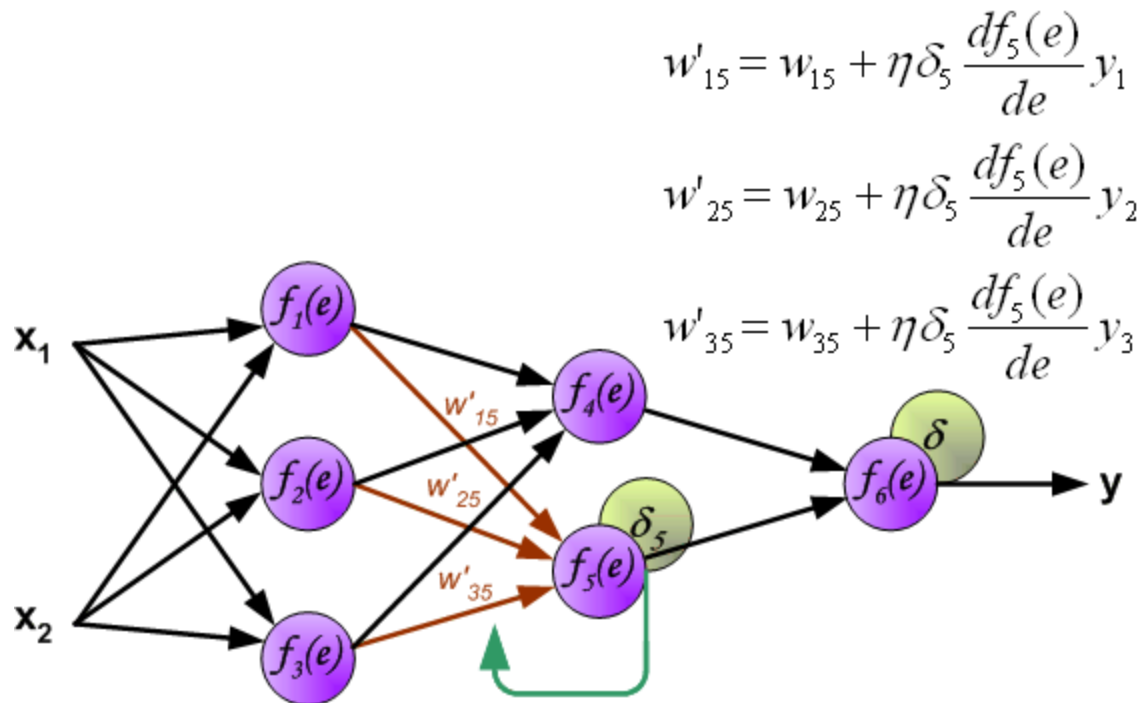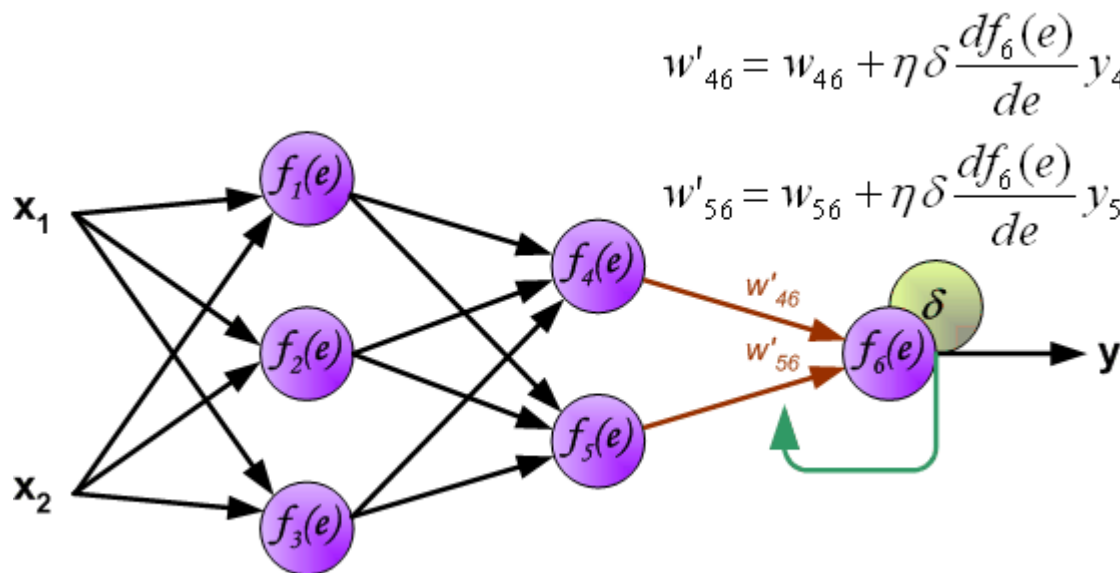
$$w'_{15} = w_{15} + \eta \delta_5 \frac{df_5(e)}{de} y_1$$

$$w'_{25} = w_{25} + \eta \delta_5 \frac{df_5(e)}{de} y_2$$

$$w'_{35} = w_{35} + \eta \delta_5 \frac{df_5(e)}{de} y_3$$

http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

Back Propagation of Errors is the most popular Training Algorithm

- When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified. In formulas below *df(e)/de* represents derivative of neuron activation function (which weights are modified).
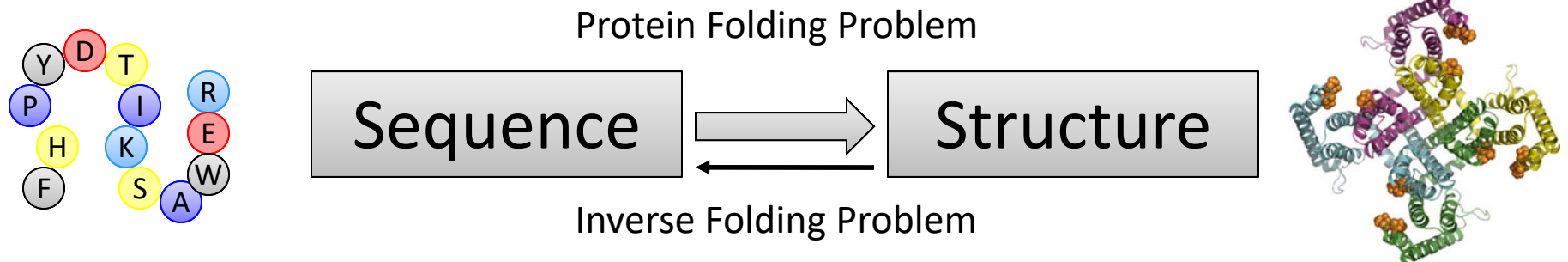
$$w'_{46} = w_{46} + \eta \delta \frac{df_6(e)}{de} y_4$$

$$w'_{56} = w_{56} + \eta \delta \frac{df_6(e)}{de} y_5$$

The (Inverse) Protein Folding Problem Holy Grail of Comp. Structural Biology

Protein Folding Problem

| Sequence | $\Rightarrow$ | Structure |

Inverse Folding Problem

- Given a protein's AA sequence, what is its 3-dimensional fold , and how does it get there?
- Assume 100 conformations for each amino acid in a 100 amino acid protein $\Rightarrow 10^{200}$ possible conformations!
- Exhaustive sampling is impossible – e.g. earth is less than 10^{10} years old.
- Cyrus Levinthal's paradox of protein folding, 1968.

# Protein Folding using Lattice Models and Grid Searches

- Arrange amino acids randomly on three-dimensional grid
- Define a simplified energy function that measures exposure (red=buried, blue exposed), etc.
- Search arrangements using Monte Carlo or Genetic algorithms
- Works only for very small proteins (<50AA)
- Popular in earlier days of protein structure prediction (1990-2000) for reduced computational requirements

R. Unger and J. Moult; "Genetic algorithms for protein folding simulations"; *J Mol Biol*; **1993**; Vol. 231 (1): p. 75-81.

A. Kolinski and J. Skolnick; "Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme"; *Proteins*; **1994**; Vol. 18 (4): p. 338-52.

A. Sali, E. Shakhnovich and M. Karplus; "Kinetics of protein folding. A lattice model study of the requirements for folding to the native state"; *J Mol Biol*; **1994**; Vol. 235 (5): p. 1614-36.
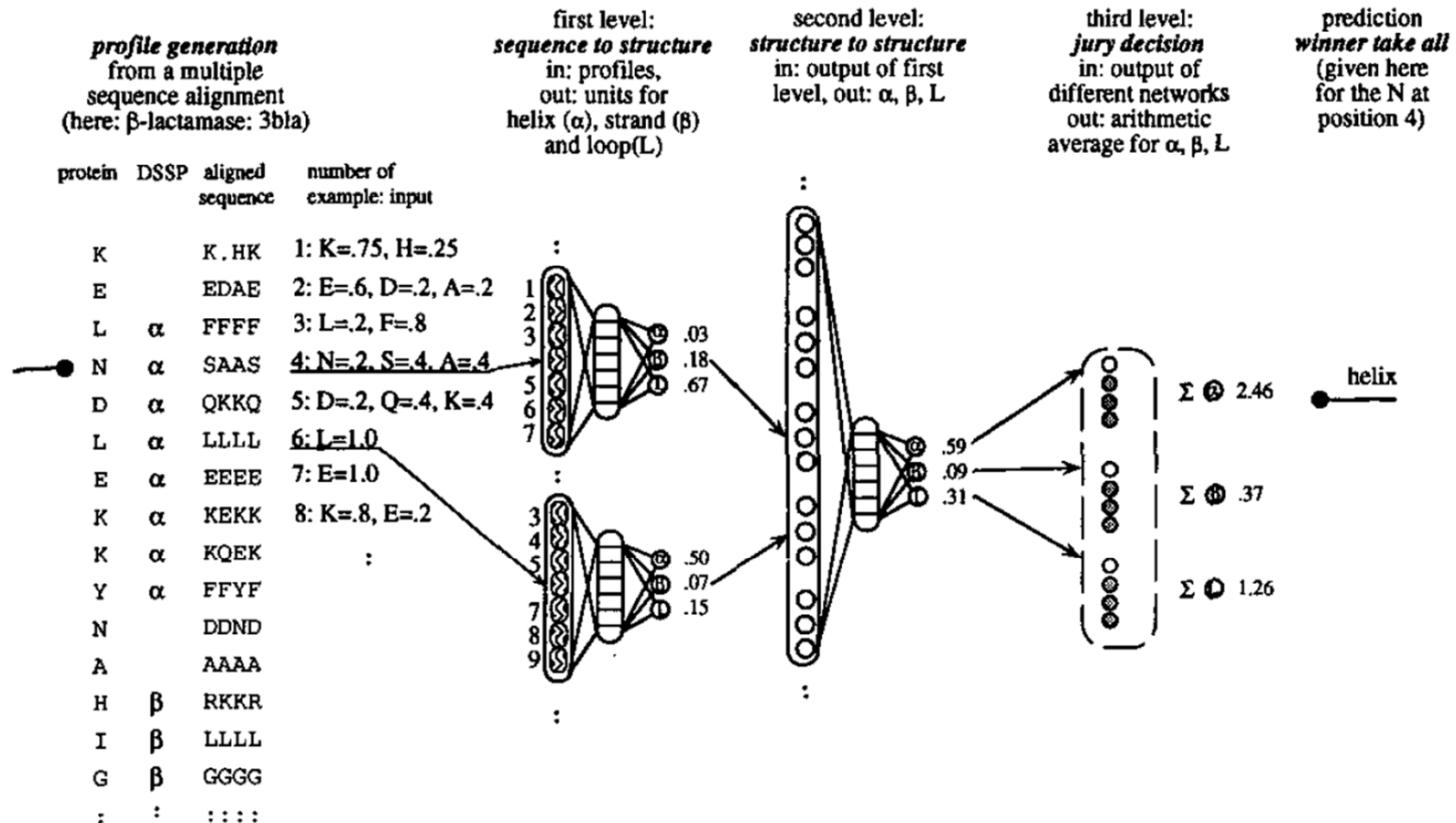
K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan; "Principles of protein folding--a perspective from simple exact models"; *Protein Sci*; **1995**; Vol. 4 (4): p. 561-602.
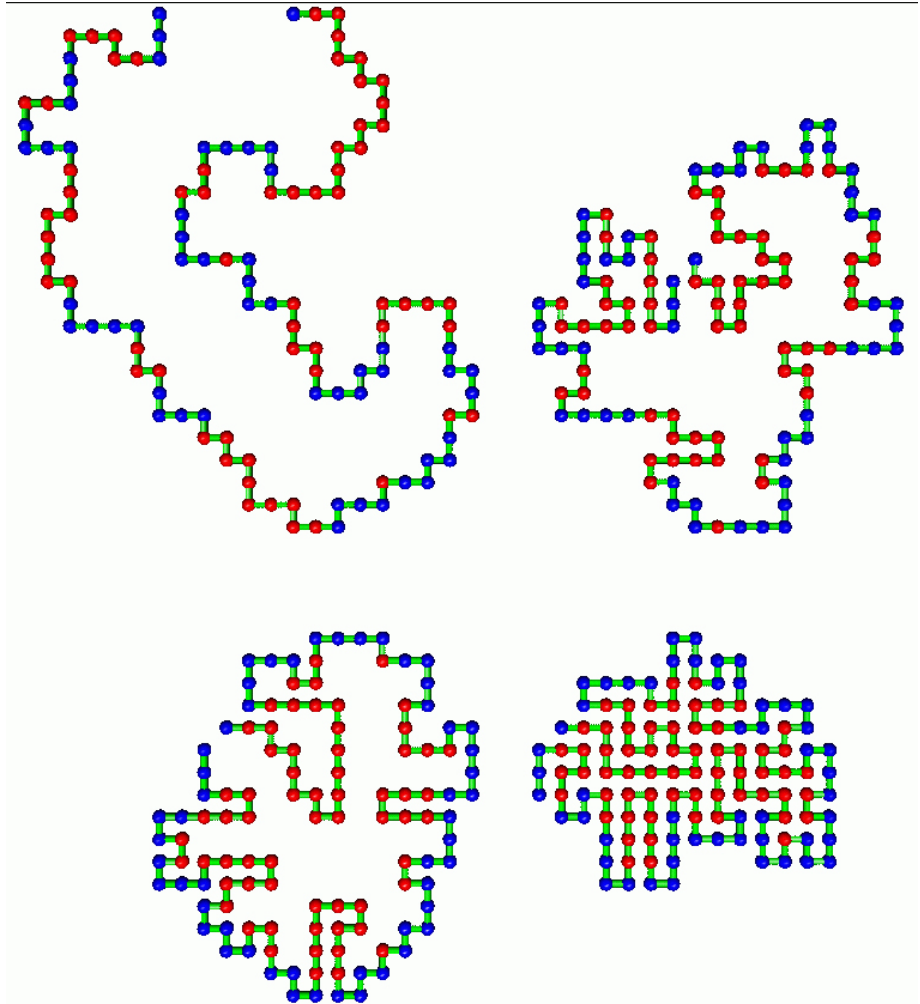
General Scheme of Protein Structure Prediction
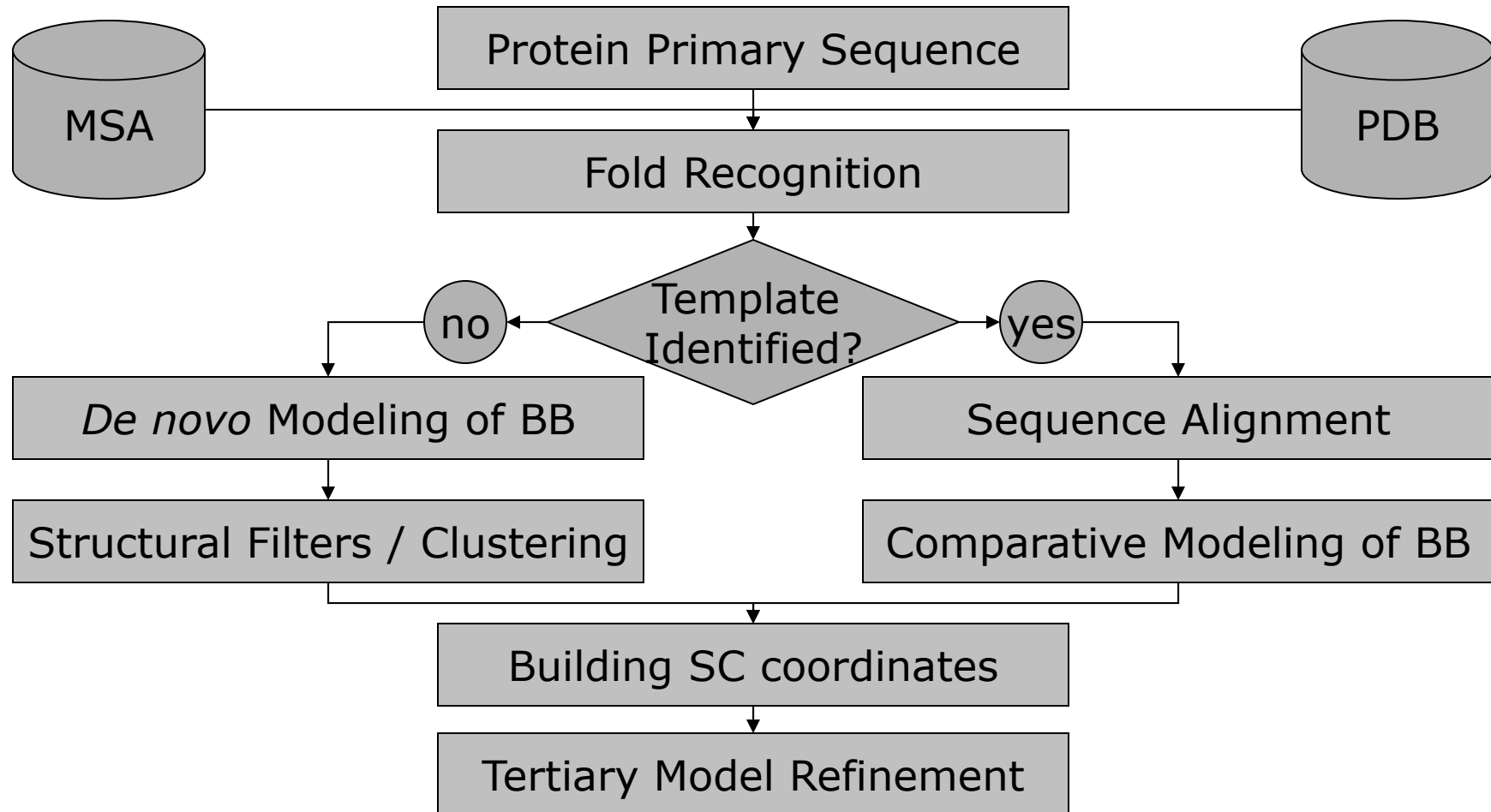
PhD - Prediction of protein secondary structure at better than 70% accuracy

B. Rost and C. Sander; "Prediction of protein secondary structure at better than 70% accuracy"; *J. Mol. Biol.*; **1993**; Vol. 232 (2): p. 584-99; J. Meiler, A. Zeidler, F. Schmaschke and M. Muller; "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks"; *Journal of Molecular Modeling*; **2001**; Vol. 7 (9): p. 360-369.

BCL::Jufo9D >70% correct 9-state prediction, >80% SS, >90% TM

========== P R E D I C T I O N ==========

	tm_C	tr_C	sol_C	tm_S	tr_S	sol_S	tm_H	tr_H	sol_H
tm_C	63.7	13.3	1.9	4.8	2.0	1.1	6.0	6.9	0.5
tr_C	4.8	64.7	3.4	1.9	7.8	0.5	0.9	15.3	0.8
sol_C	3.8	23.0	43.9	2.4	2.5	8.7	2.1	2.4	11.2
tm_S	6.2	0.9	0.9	78.6	7.5	5.1	0.0	0.3	0.4
tr_S	3.0	12.0	1.7	14.2	66.5	1.8	0.6	0.1	0.0
sol_S	3.7	6.2	11.1	7.0	4.9	55.9	4.0	4.2	3.1
tm_H	2.0	3.8	0.1	0.4	0.1	0.1	80.1	13.2	0.4
tr_H	2.0	16.9	0.2	0.2	1.1	0.1	12.0	66.0	1.6
sol_H	1.4	6.0	9.0	0.8	0.7	2.9	2.0	8.8	68.5

(Left axis: ===REAL STATE===)

TM   trans-membrane
TR   transition
SOL solution

H   helix
S   strand
C   coil

J. K. Leman, R. Mueller, M. Karakas, N. Woetzel and J. Meiler; "Simultaneous prediction of protein secondary structure and transmembrane spans"; *Proteins*; **2013**; Vol. 81 (7): p. 1127-40.

Example 1: Succinate dehydrogenase (1NEK)

- Secondary structure prediction
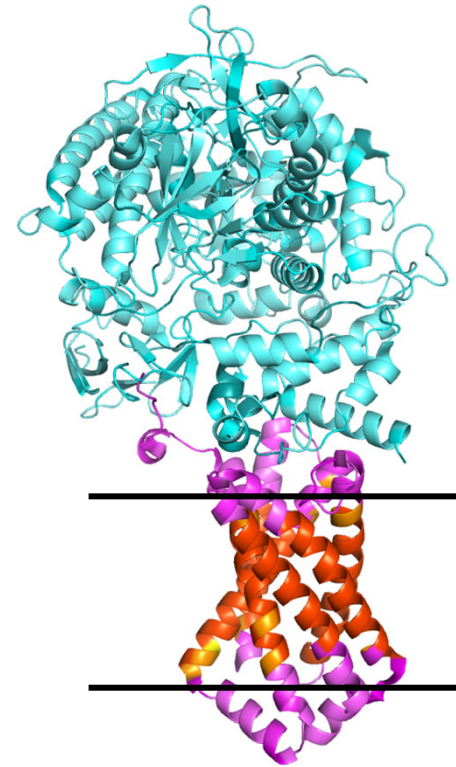
■ helix
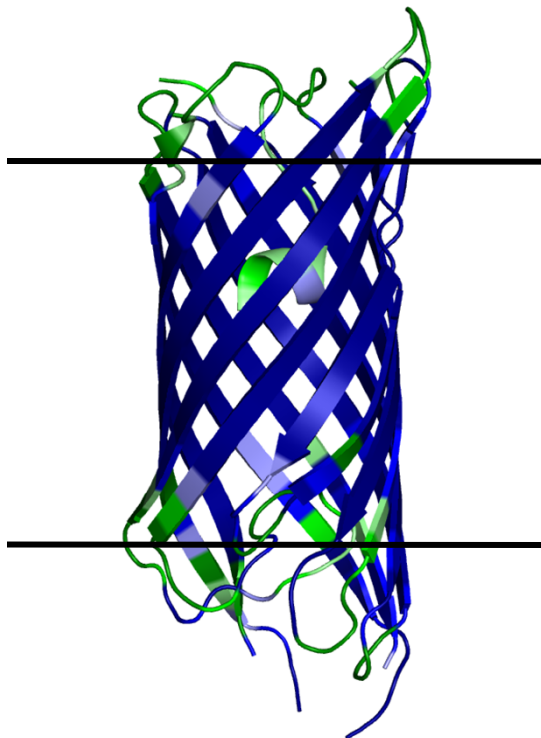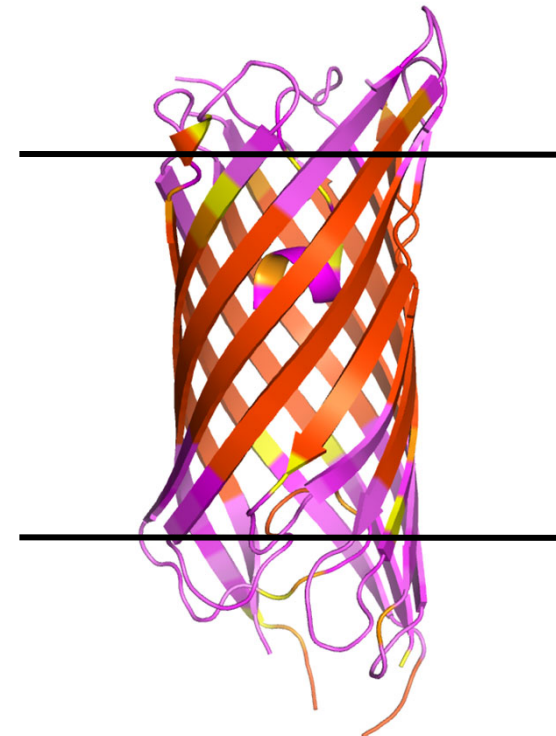■ strand
■ coil
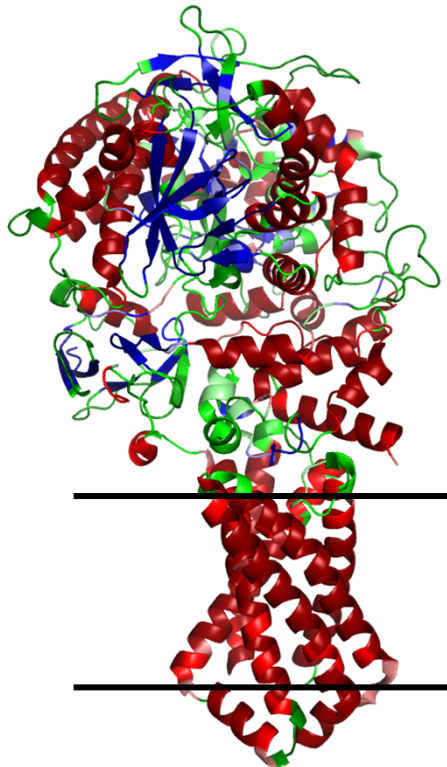
■ membrane
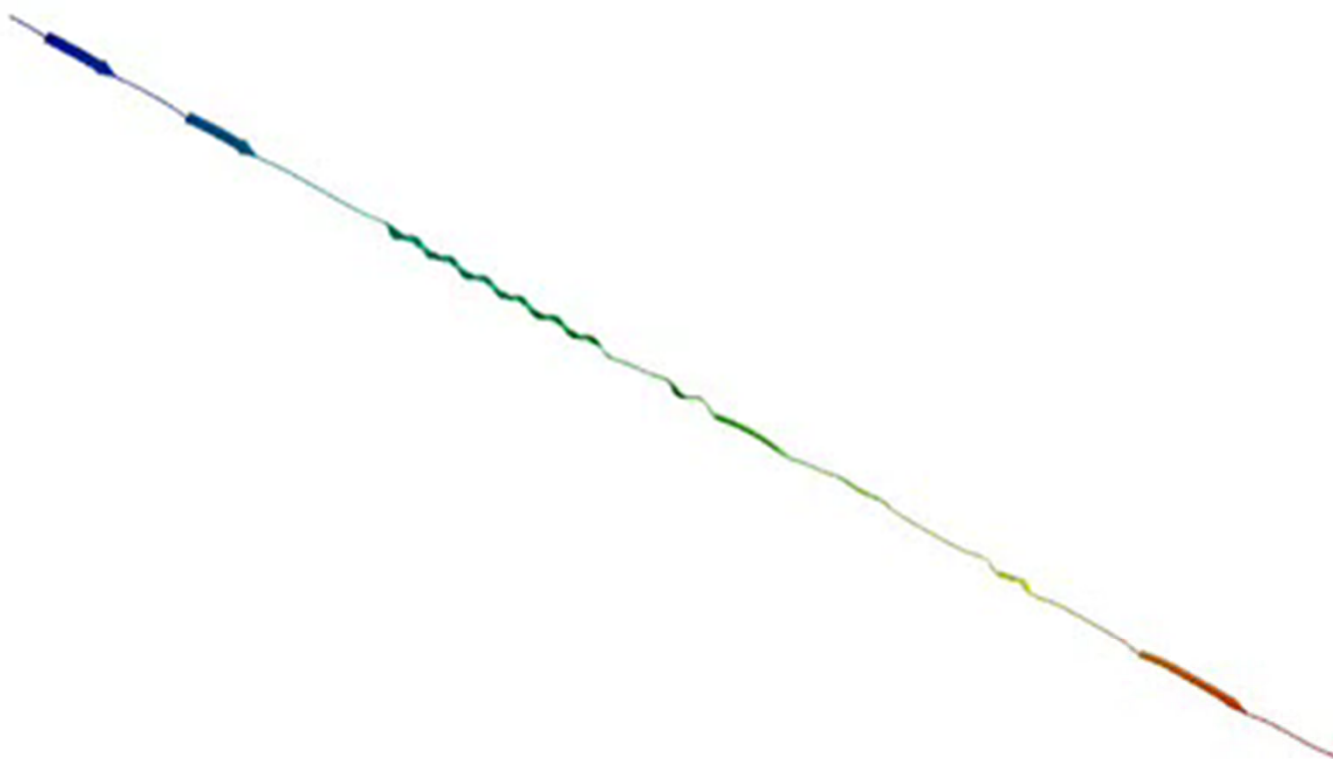■ transition
■ solution

- Trans-membrane span identification

**92%   81%
residues correct**

Example 2: EspP autotransporter beta-domain (2QOM)

- Secondary structure prediction

  ■ helix
  ■ strand
  ■ coil

  ■ membrane
  ■ transition
  ■ solution

- Trans-membrane span identification

80%   74%
residues correct

A NN–based Consensus Predictor that Improves Fold Recognition

During recent years many protein fold recognition methods have been developed, based on different algorithms and using various kinds of information. To examine the performance of these methods several evaluation experiments have been conducted. These include blind tests in CASP/CAFASP, large scale benchmarks, and long-term, continuous assessment with newly solved protein structures. These studies confirm the expectation that for different targets different methods produce the best predictions, and the final prediction accuracy could be improved if the available methods were combined in a perfect manner. In this article a neural-network-based consensus predictor, Pcons, is presented that attempts this task. Pcons attempts to select the best model out of those produced by six prediction servers, each using different methods. Pcons translates the confidence scores reported by each server into uniformly scaled values corresponding to the expected accuracy of each model. The translated scores as well as the similarity between models produced by different servers is used in the final selection. According to the analysis based on two unrelated sets of newly solved proteins, Pcons outperforms any single server by generating ~8%–10% more correct predictions. Furthermore, the specificity of Pcons is significantly higher than for any individual server. From analyzing different input data to Pcons it can be shown that the improvement is mainly attributable to measurement of the similarity between the different models. Pcons is freely accessible for the academic community through the protein structure-prediction metaserver at http://bioinfo.pl/meta/.
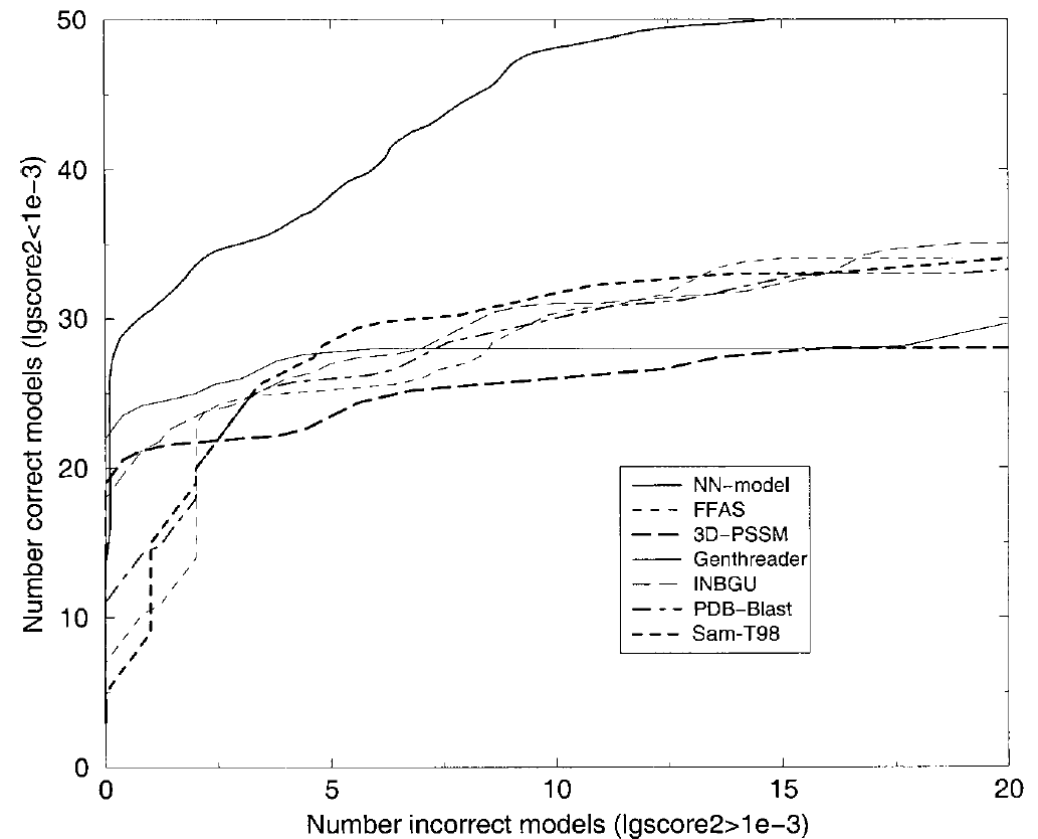
- Lundstroem, J.; Rychlewski, L.; Bujnicki, J.; Elofsson, A., Pcons: A neural-network –based consensus predictor that improves fold recognition. Protein Sci. 2001, 10, 2354-2362.

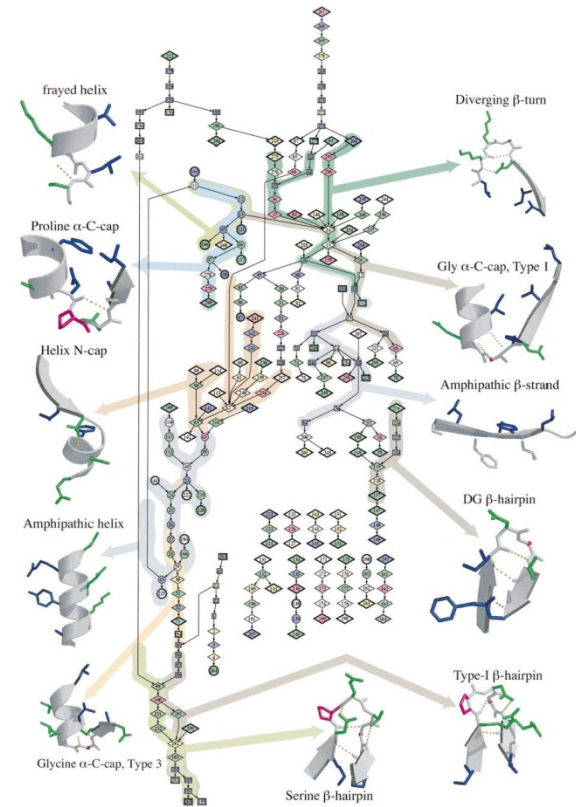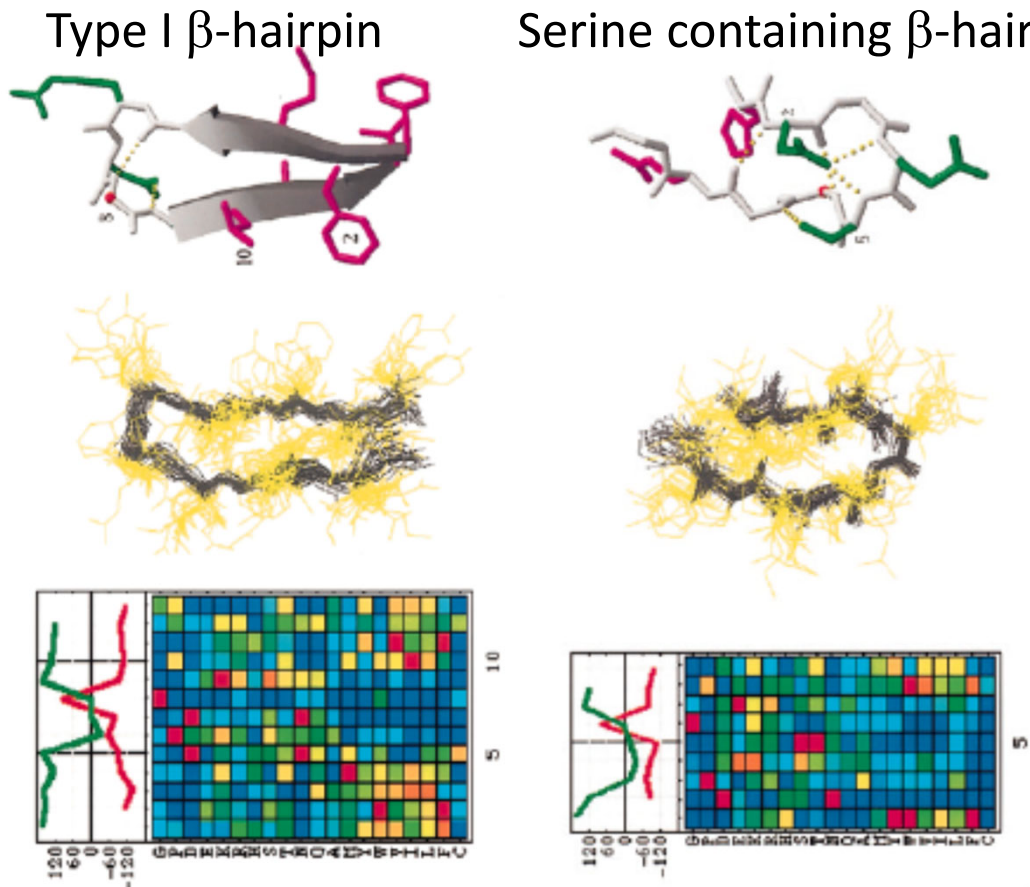# Hidden Markov Models Identify Local Structural Motives from Sequence

Type I β-hairpin **Serine containing β-hairpin**[178]

The Grammatical Structure of Protein Sequences

1. Bystroff, C.; Baker, D., Prediction of Local Structure in Proteins Using a Library of Sequence-Structure Motifs. *J. Mol. Biol.* **1998,** 281, 565-577. 2. Bystroff, C.; Thorsson, V.; Baker, D., HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins. *J. Mol. Biol.* **2000,** 301, 173-190.

ANN – Derived Contact Numbers Improve Membrane Protein Structure Prediction
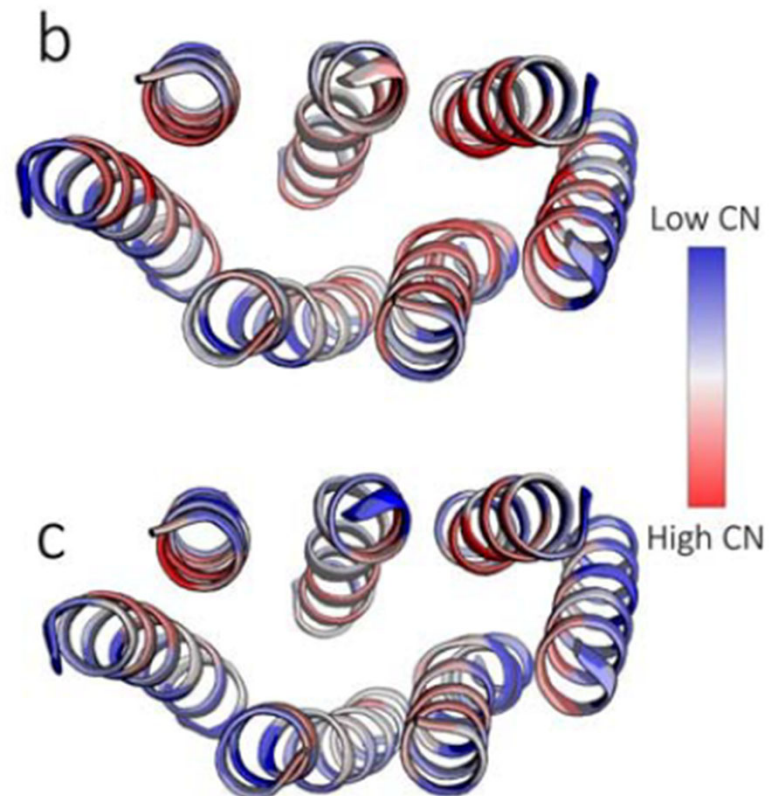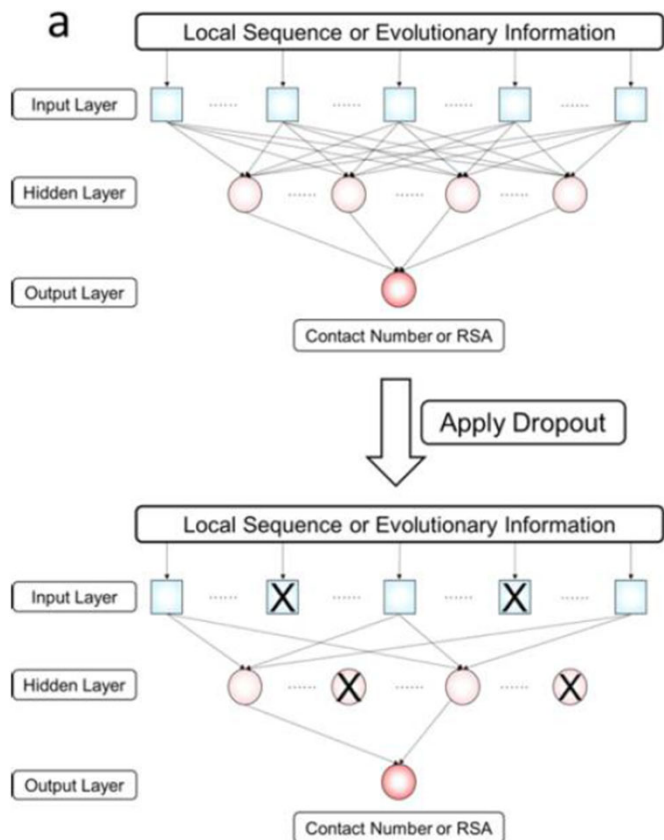
B. Li, J. Mendenhall, E. D. Nguyen, B. E. Weiner, A. W. Fischer and J. Meiler; "Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins"; *J Chem Inf Model*; **2016**; Vol. 56 (2): p. 423-34.
B. Li, J. Mendenhall, E. D. Nguyen, B. E. Weiner, A. W. Fischer and J. Meiler; "Improving prediction of helix-helix packing in membrane proteins using predicted contact numbers as restraints"; *Proteins*; **2017**; Vol. 85 (7): p. 1212-1221.

Convolutional Neural Networks (CNN) can understand Different Levels of Resolution
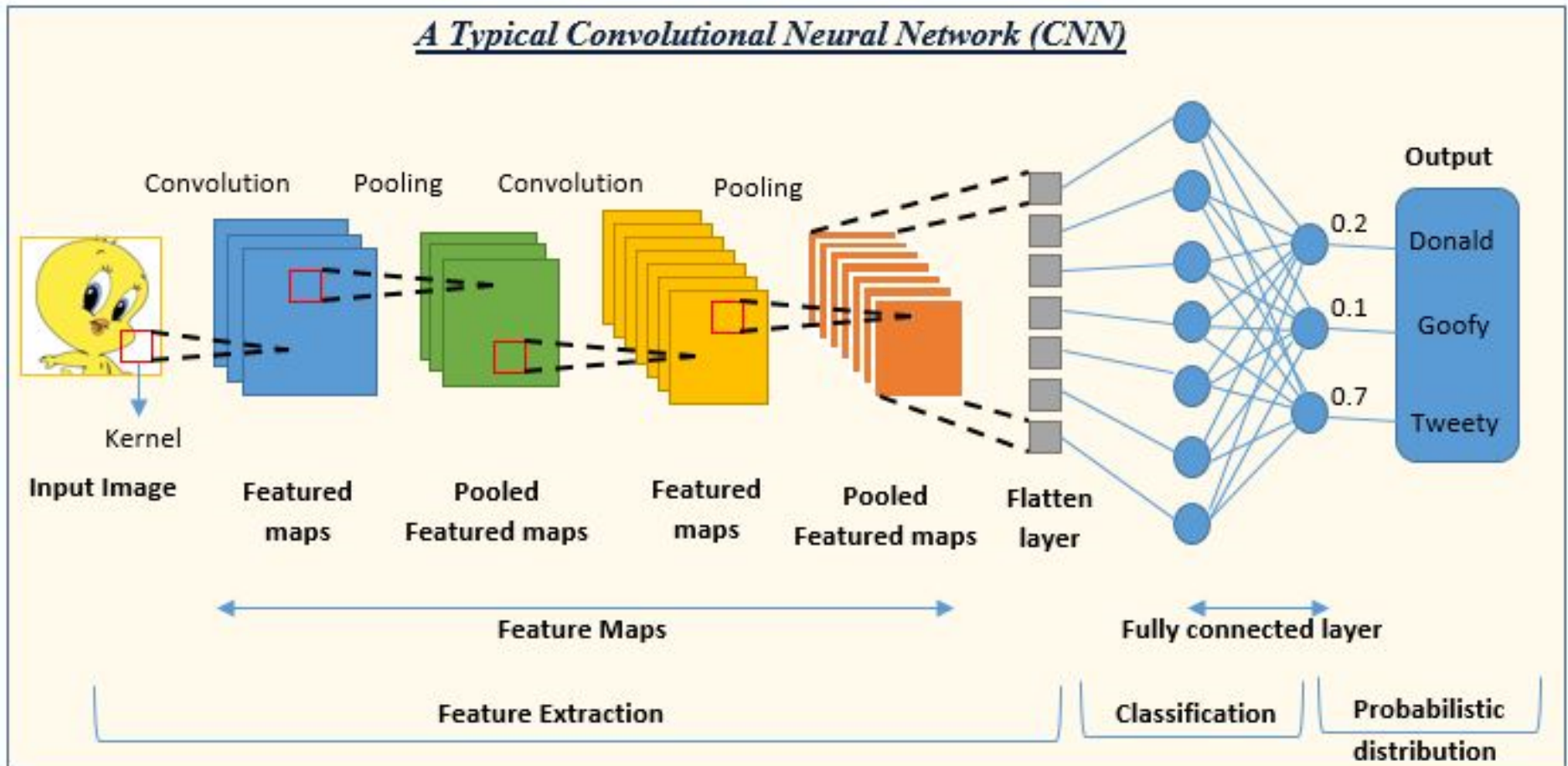
Paying Attention
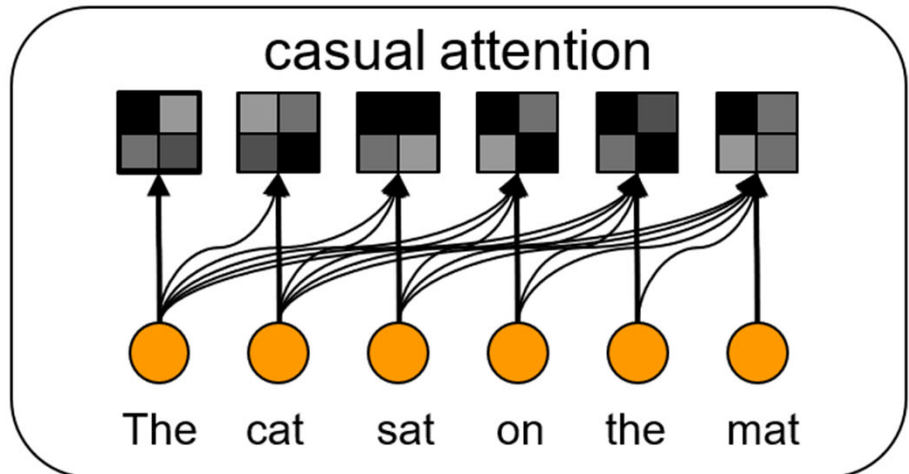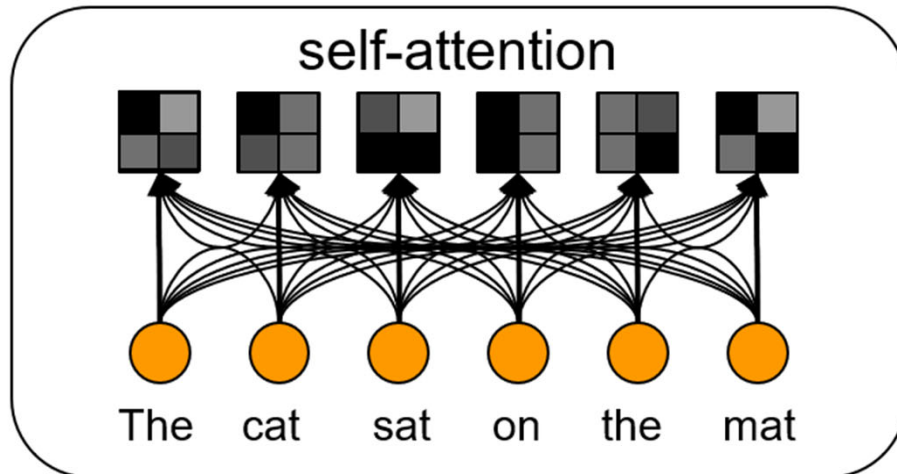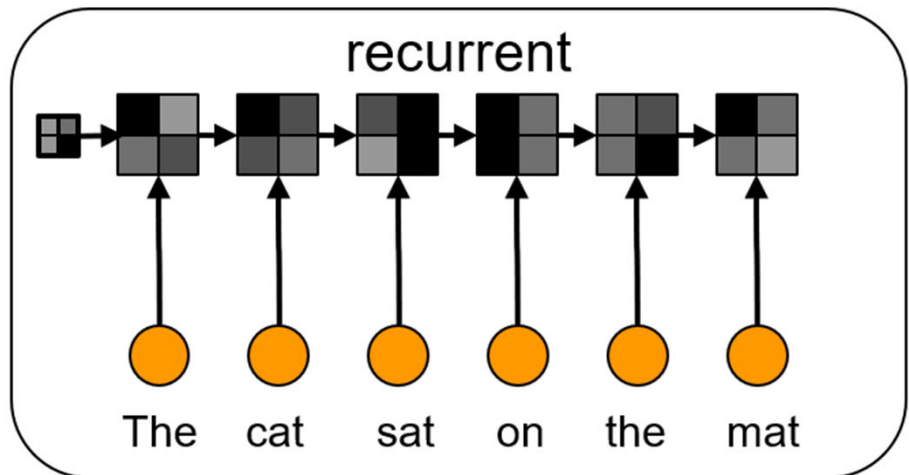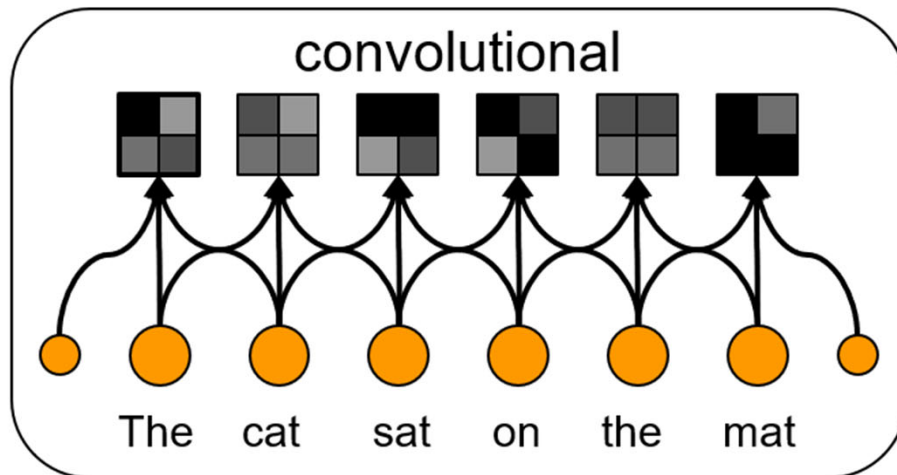
Transformers and Attention – let the Neural Network figure out Importance
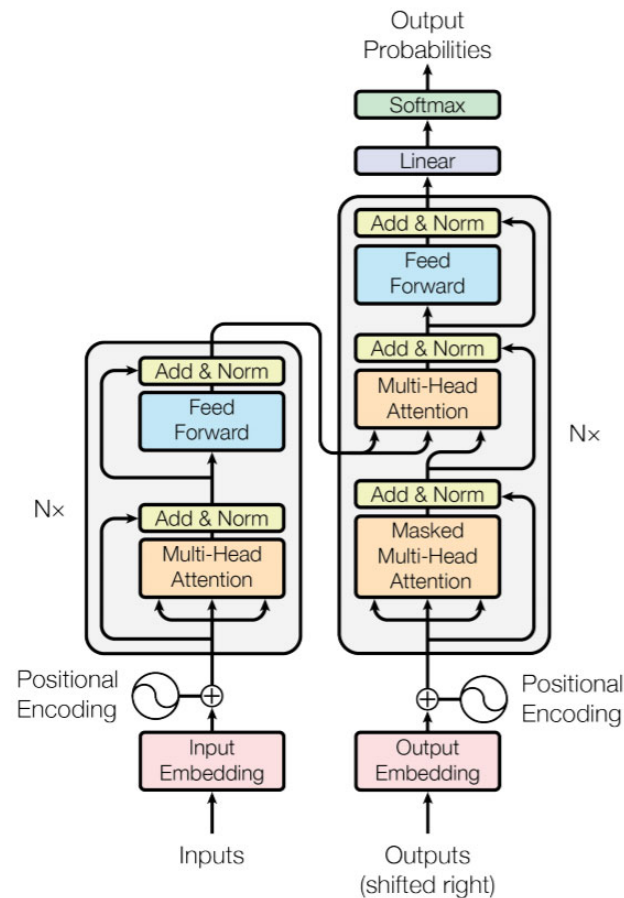
Figure 1: The Transformer - model architecture.

Transformers and Attention – let the Neural Network figure out Importance

Output Probabilities

Softmax

Linear

Residual connections and layer normalization

Add & Norm

Feed Forward

Add & Norm

Multi-Head Attention

Nx

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Add & Norm

Feed Forward

Add & Norm

Multi-Head Attention

Nx

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

Add & Norm

Masked Multi-Head Attention

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Positional Encoding

Positional Encoding

Input Embedding

Output Embedding

Inputs

Outputs (shifted right)

The Future of Artificial Neural Networks in Biomedical Research – Some Thesis

1. All problems that have infinite/near infinite data available for training will be smashed (think language processing, sequence problems in biochemistry, protein structure)

2. New architectures and structures of ANNs will emerge that will be parallel in size to or larger then the human brain (10^{14} connections) with substructures matching in complexity

3. The biggest challenge for biomedical research will emerge with limited datasets that forbid training of super-large ANNs; Expert Knowledge will Design the Optimal ANN

4. For the next Decade (at least), you need to be an expert in machine learning and structural/chemical biology to contribute to progress in a meaningful way

5. We will start an honest discussion on ethics of artificial intelligence as these systems will start to act human-like on many levels all the way to having self-awareness

Highly accurate protein structure prediction with AlphaFold
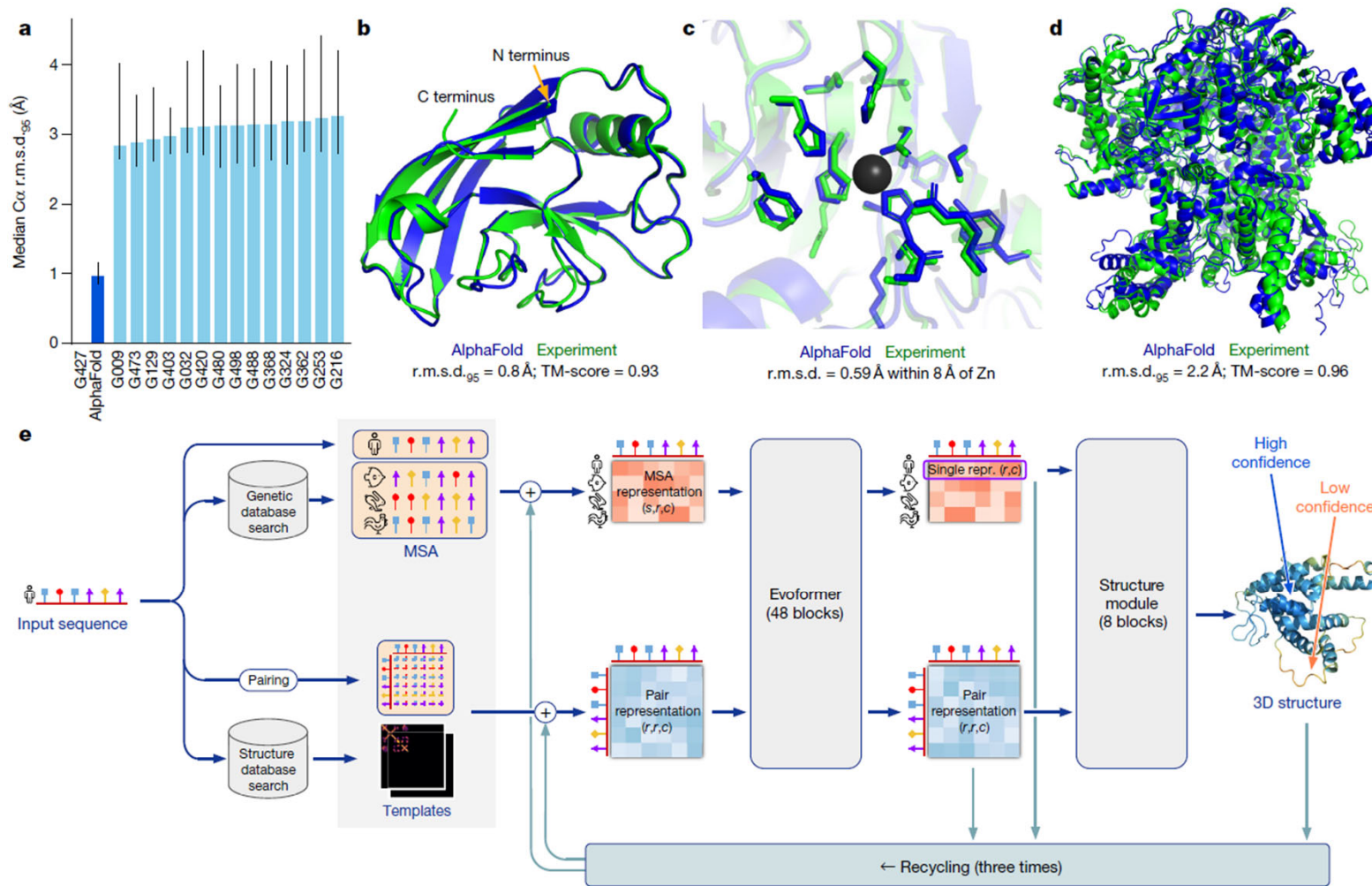
J. Jumper, et al.; "Highly accurate protein structure prediction with AlphaFold"; *Nature*; **2021**; Vol. 596 (7873): p. 583-589.

Highly accurate protein structure prediction with AlphaFold – Evoformer
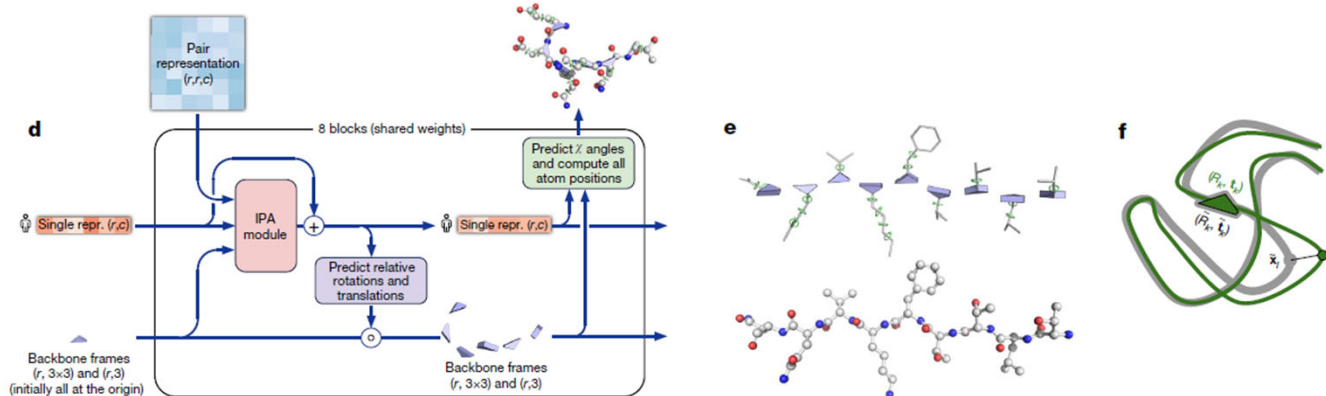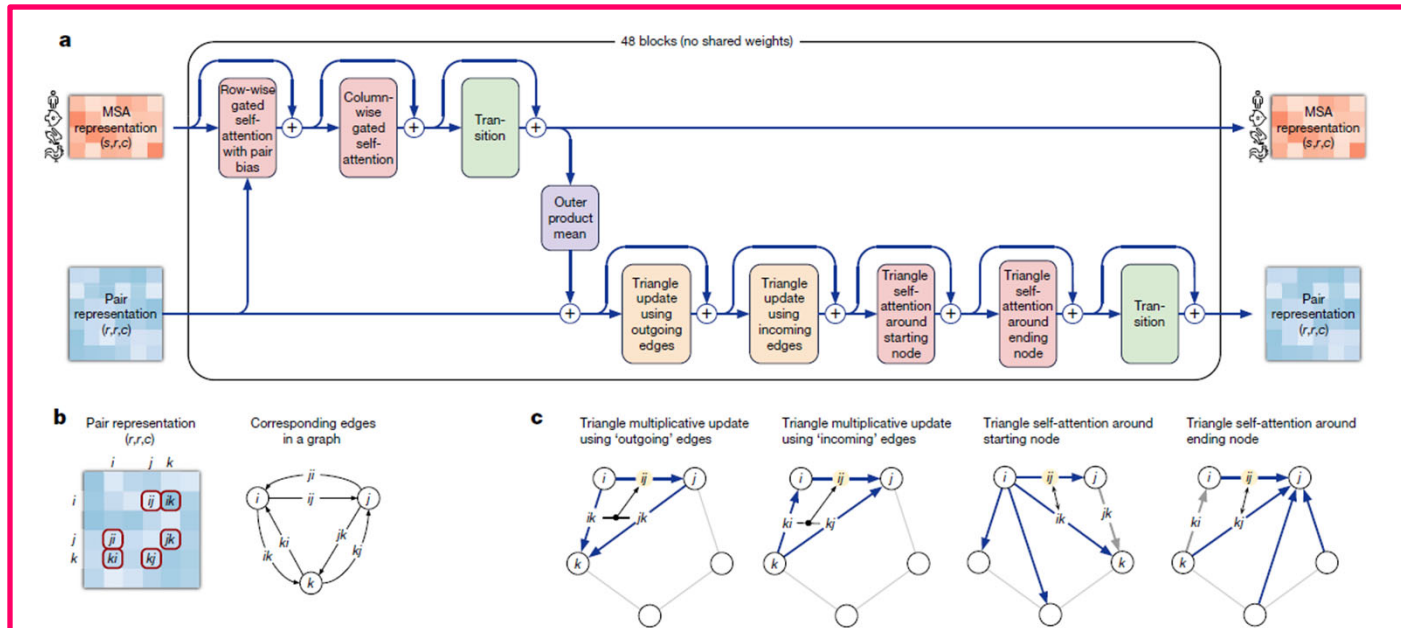
J. Jumper, et al.; "Highly accurate protein structure prediction with AlphaFold"; *Nature*; **2021**; Vol. 596 (7873): p. 583-589.

Highly accurate protein structure prediction with AlphaFold – Structure
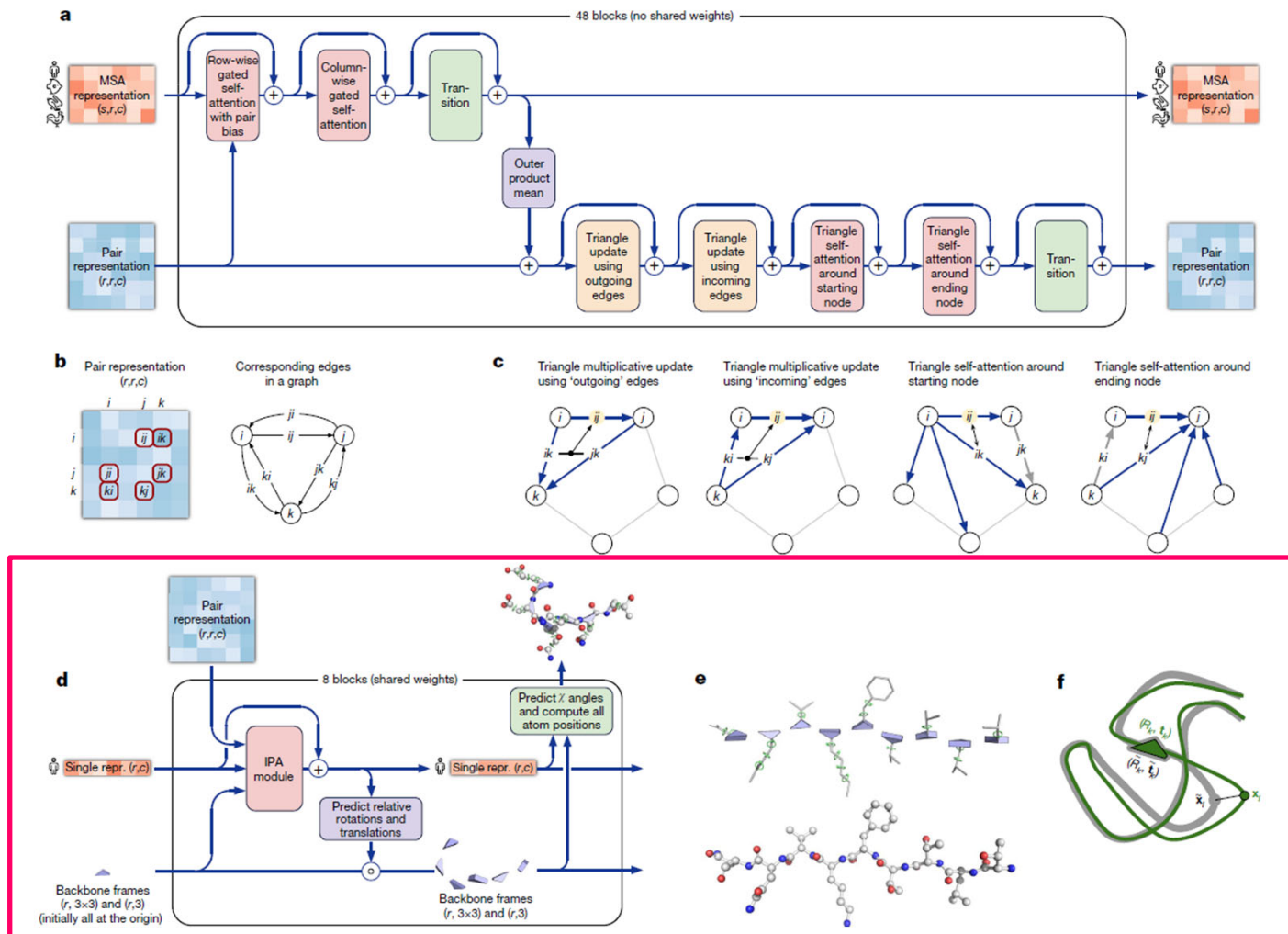
J. Jumper, et al.; "Highly accurate protein structure prediction with AlphaFold"; *Nature*; **2021**; Vol. 596 (7873): p. 583-589.

Highly accurate protein structure prediction with AlphaFold
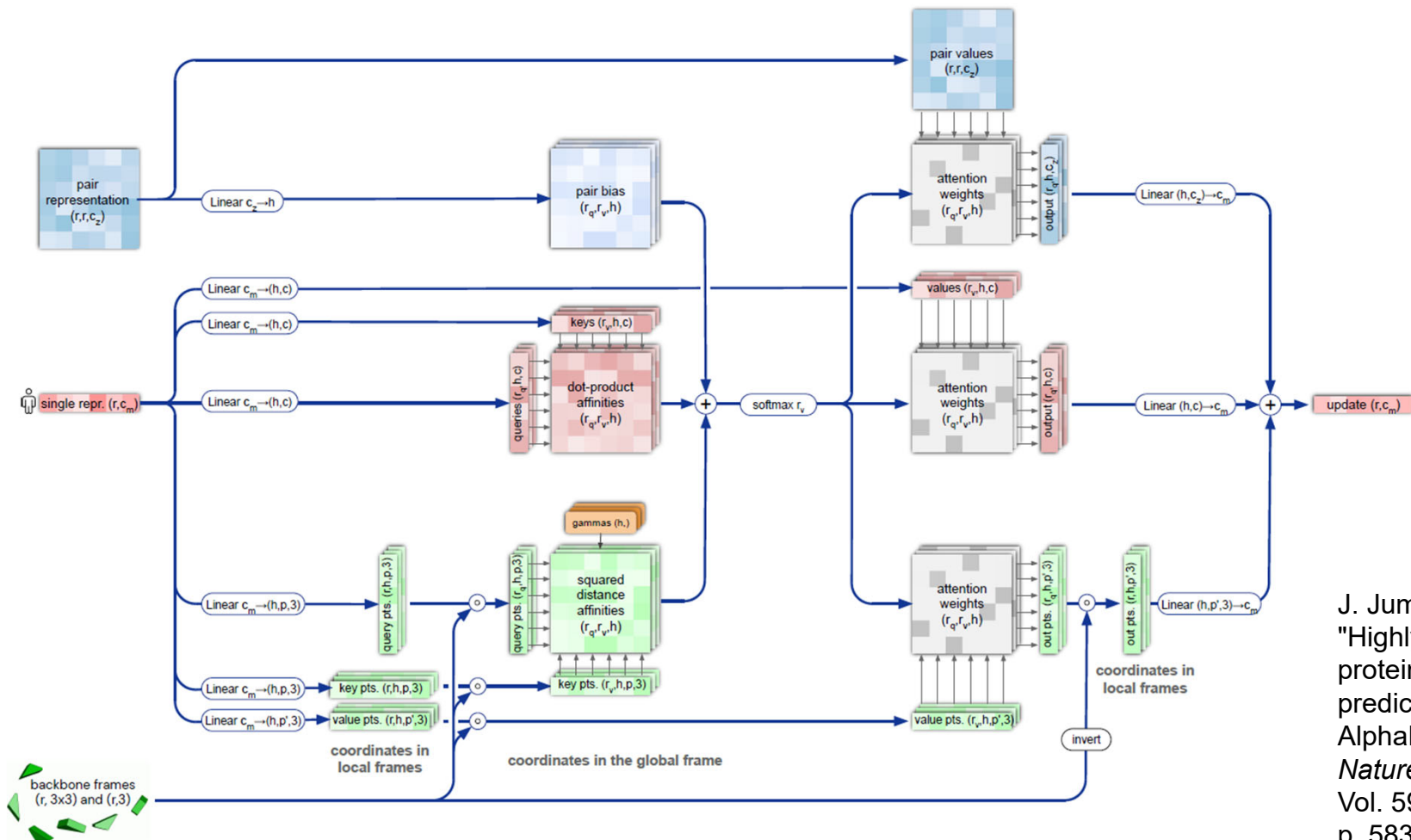
J. Jumper, et al.;
"Highly accurate protein structure prediction with AlphaFold";
Nature; **2021**;
Vol. 596 (7873): p. 583-589.

Sampling Alternative Conformational States with AlphaFold2
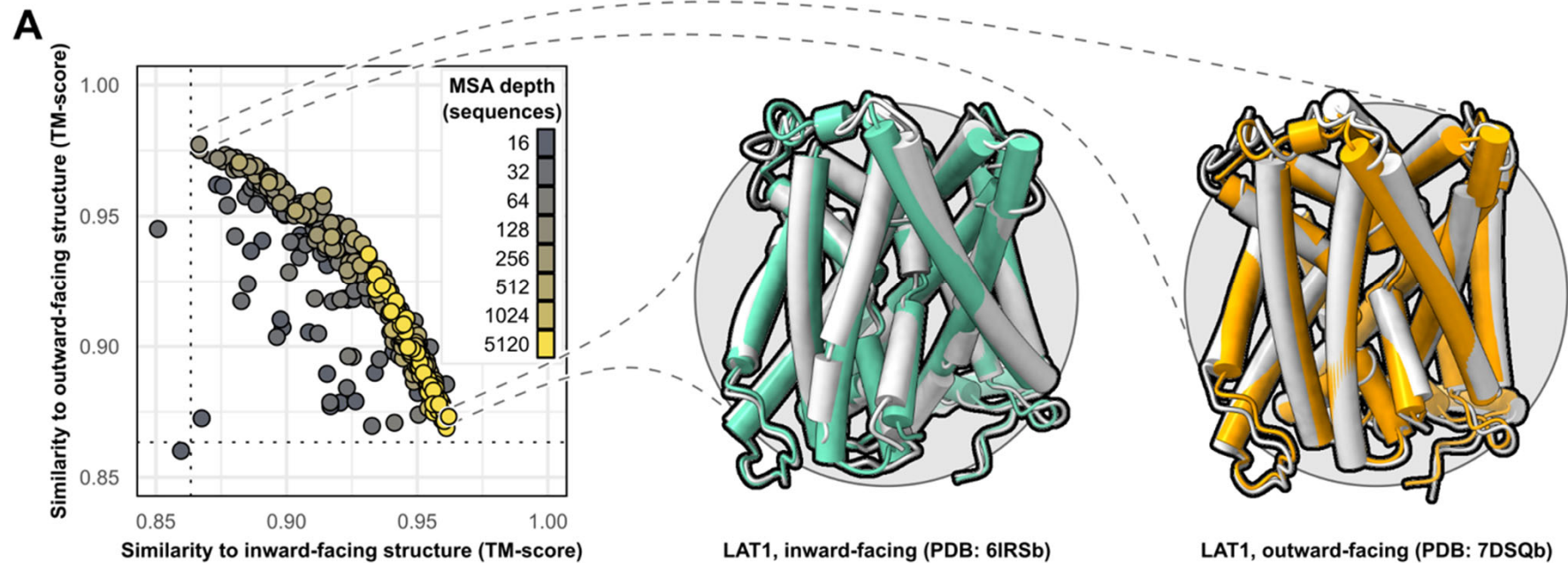
Figure 1. Alternative conformations of transporters and GPCRs can be predicted by AF2. (A) Representative models of the transporter LAT1 in IF and OF conformations. Experimental structures shown in gray and models shown in teal and orange.

D. Del Alamo, D. Sala, H. S. McHaourab and J. Meiler; "Sampling alternative conformational states of transporters and receptors with AlphaFold2"; *Elife*; **2022**; Vol. 11 p.

Molecular Architecture of the Human Caveolin-1 Complex with AlphaFold2

Molecular Architecture of the Human Caveolin-1 Complex with AlphaFold2
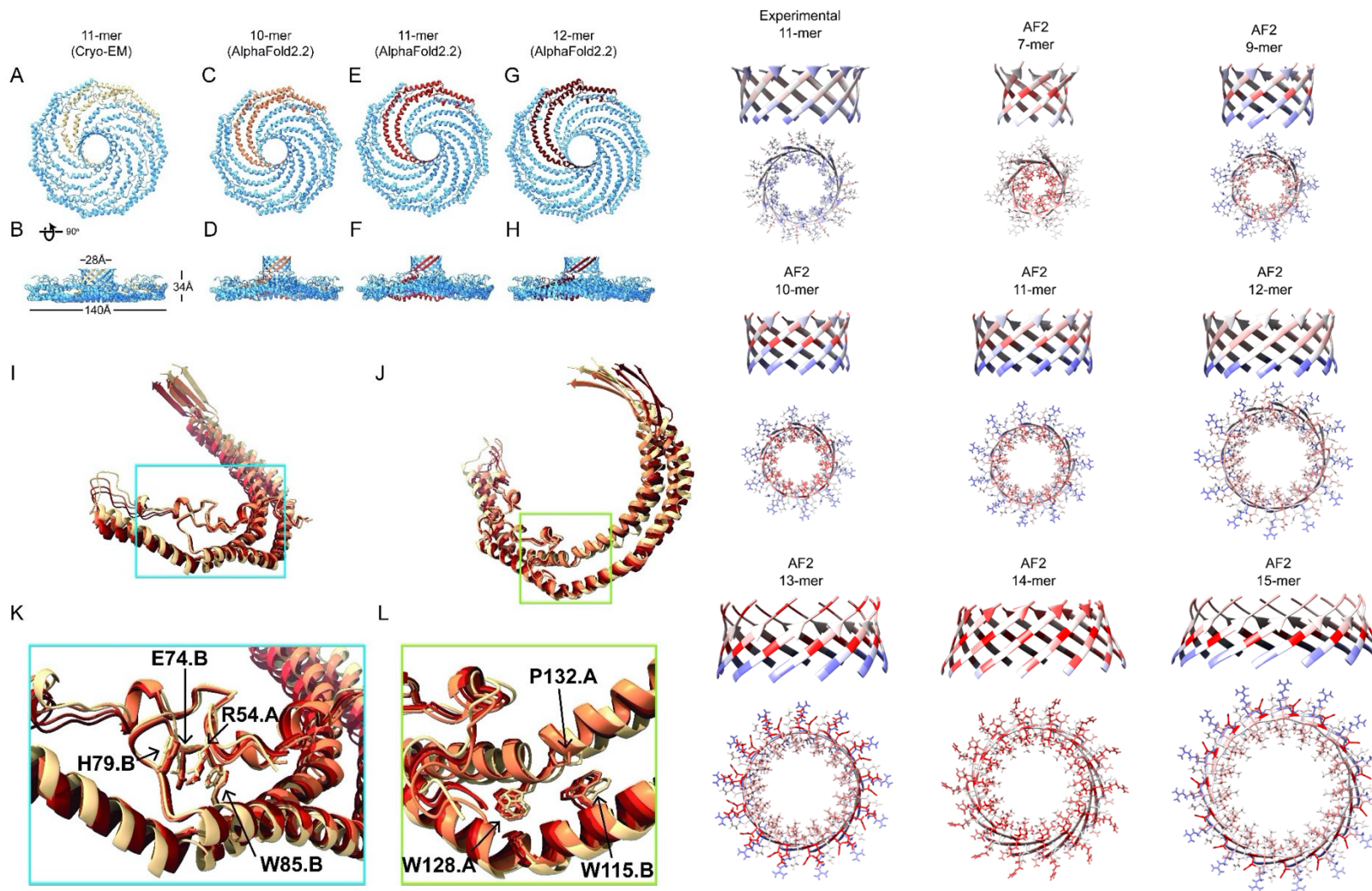
J. C. Porta, B. Han, A. Gulsevin, J. Chung, Y. Peskova, S. Connolly, H. S. Mchaourab, J. Meiler, E. Karakas, A. K. Kenworthy and M. D. Ohi; "Molecular architecture of the human caveolin-1 complex"; *Science Advances*; **2022**; Vol. p.
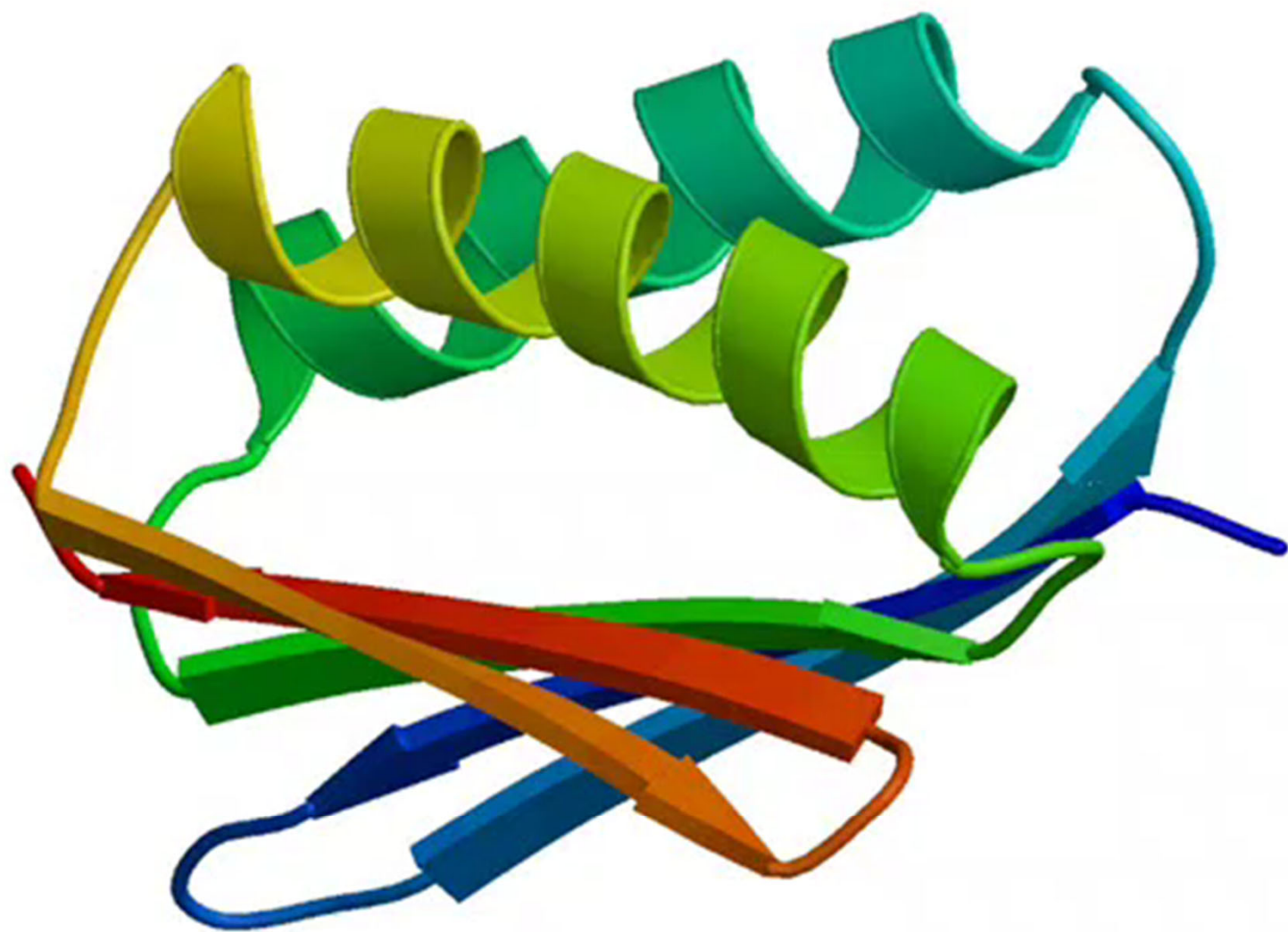
Antibody Diversity is Limited to 10^{11} Germline Antibodies

Locus	V Genes	D Genes	J Genes
IGH	38-46	23	6
IGK	34-38	0	5
IGL	29-33	0	4-5

V(D)J Recombination | **Junctional Diversity** | **Heavy/Light Chain Pairing**
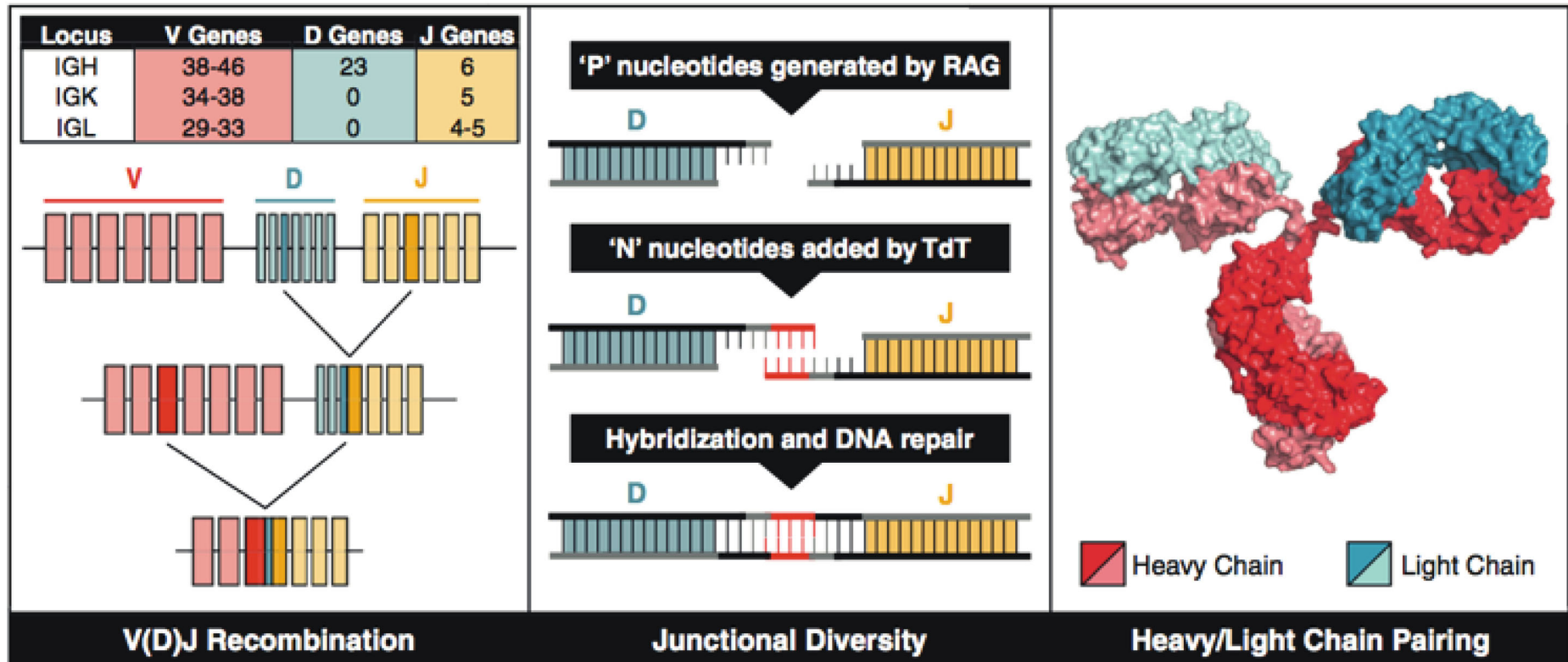
V-Gene || non-templated (N) Nucleotides || D-Gene || N-Nucleotides || J-Gene

Finn et al. *Curr Opin. Immunol* 2013

Complementarity Determining Regions (CDRs) recognize Antigens
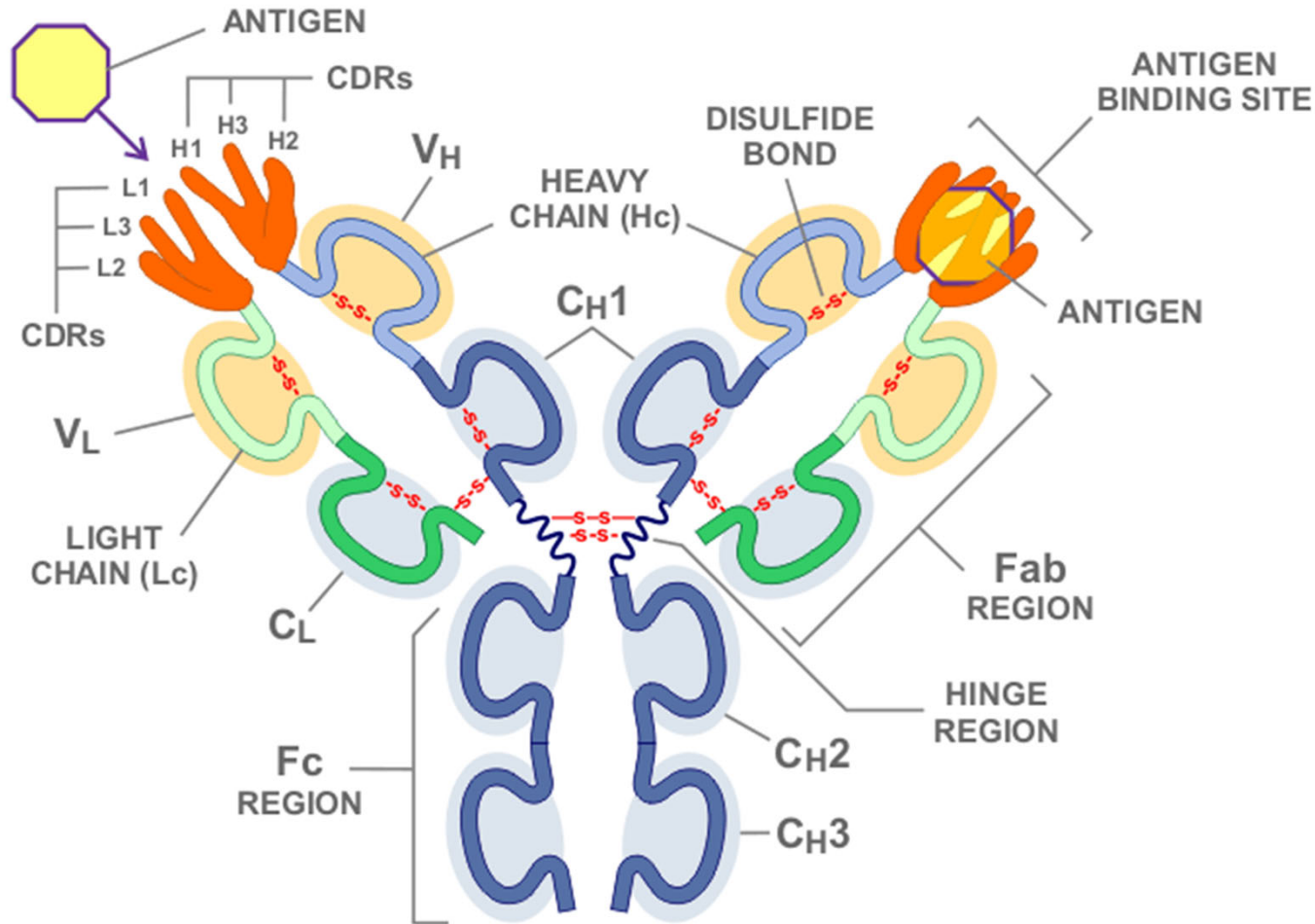
MSD of flexible proteins predicts sequences optimal for conformational change

Sauer, M. F., Sevy, A. M., Crowe, J. E., Jr., & Meiler, J. (2020). Multi-state design of flexible proteins predicts sequences optimal for conformational change. *PLoS Comput Biol, 16(2), e1007339.* https://doi.org/10.1371/journal.pcbi.1007339
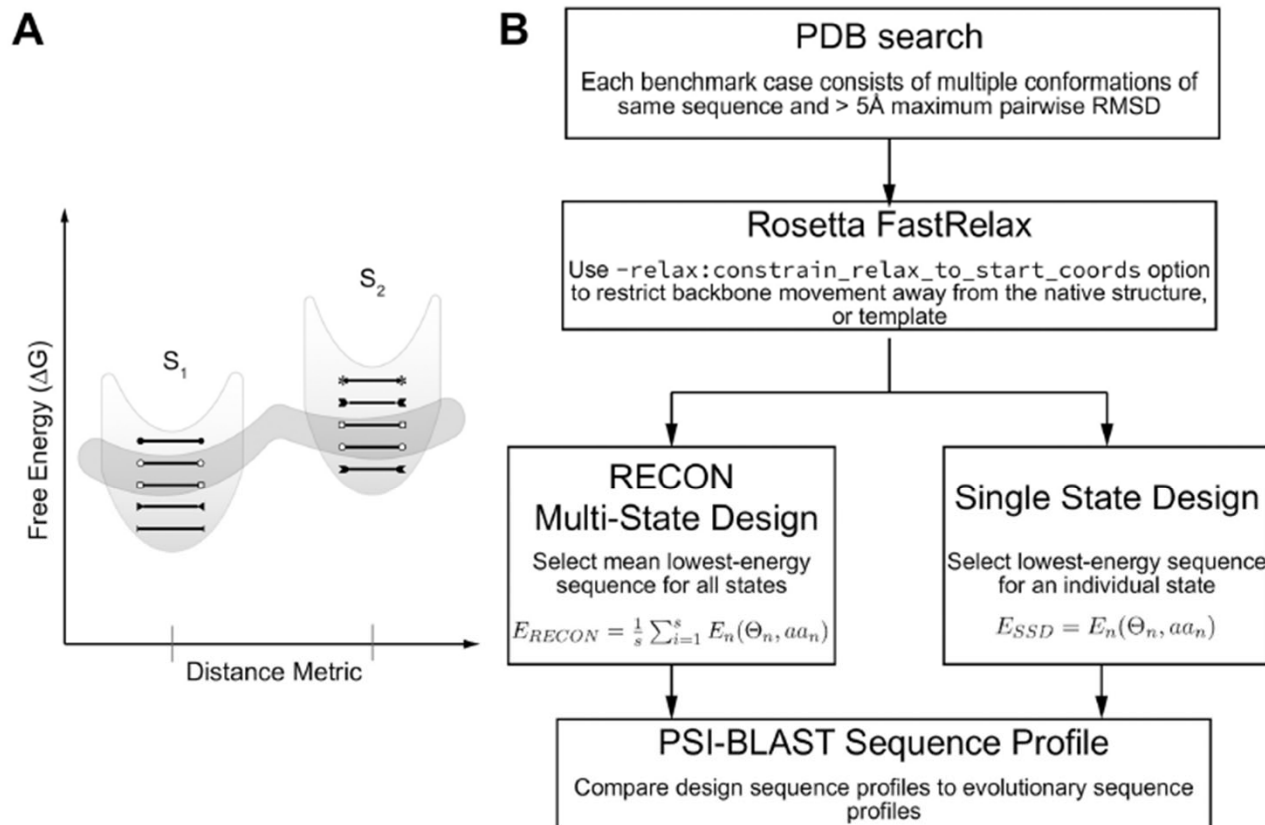
Fig 1. **Graphical representation of hypothesis and experimental design.** (A) Schematic of sequence space and the impact of flexibility on sequence tolerance. S_1 and S_2 represent two unique conformations of the same residue length separated by some RMSD that populate two local energy minima. Black lines with end caps represent unique sequences that are energetically most favorable for a single conformation. The dark shaded area encircles sequences that are energetically favorable for both conformations. Here we illustrate that by using multiple conformations during protein design, we identify sequences that are energetically suitable for conformational flexibility, yet are not necessarily the most stable sequence for any given conformation. Additionally, the requirement to adopt multiple conformations constrains the number of suitable sequences (B) Flow chart of benchmark design.

Simulating Antibody Affinity Maturation in the Computer

- Rapid Multi-State Design Algorithm for Rosetta

- SSD for Affinity Maturation
- MSD for Broad Neutralization

Table 6. Comparison of design-generated sequences to evolutionary sequence profiles of input proteins.

Benchmarkcase	Evolutionary sequence similarity (%)[a]		
	RECON FBB	RECON BBM	MPI_MSD
CheY	56.3	70.5	57.5
Elastase	60.3	70.7	65.9
FYN	87.0	87.0	96.0
PAPD	61.7	65.3	52.4
Ran	76.6	79.3	82.5
V_H1-69	90.6	91.7	32.0
V_H3-23	50.7	50.7	36.4
V_H5-51	69.0	67.0	30.4
Average	69.0	72.8	56.6

Designs produced by MPI_MSD or fixed backbone (FBB) or backbone minimized (BBM) RECON algorithms were compared to sequence profiles of evolutionarily related proteins at designed positions.
[a]Sequence similarity is computed as the Sandelin-Wasserman similarity, normalized as a percentage. See methods for details.
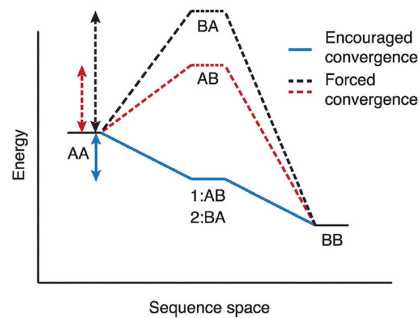
A. M. Sevy, T. M. Jacobs, J. E. Crowe, Jr. and J. Meiler; "Design of Protein Multi-specificity Using an Independent Sequence Search Reduces the Barrier to Low Energy Sequences"; *PLoS Comput Biol*; **2015**; Vol. 11 (7): p. e1004300.
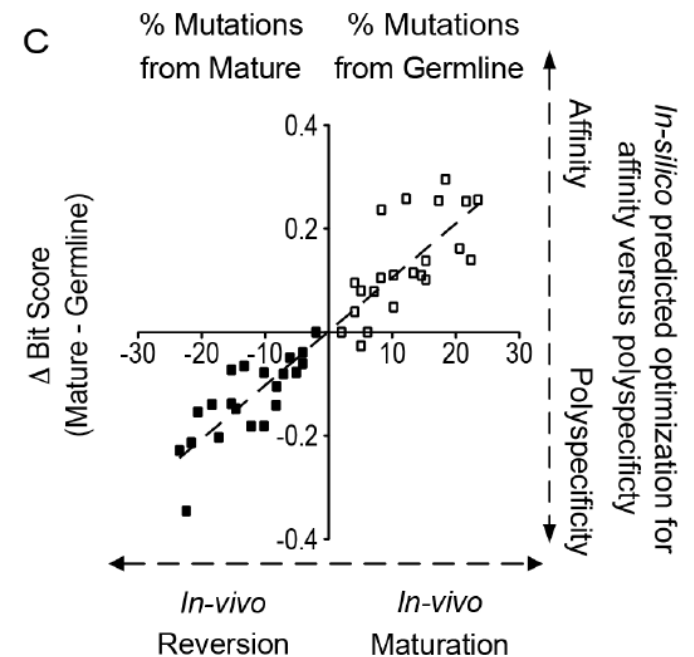
J. R. Willis, B. S. Briney, S. L. DeLuca, J. E. Crowe, Jr. and J. Meiler; "Human germline antibody gene segments encode polyspecific antibodies"; *PLoS Comput Biol*; **2013**; Vol. 9 (4): p. e1003045.

Redesign of PG9 Enhances Binding Potency and Breadth of Neutralization

- 30 Amino Acid HCDR3 but few somatic mutations

- N109Y Mutant is Predicted to Stabilize HCDR3 in Active Conformation

J. R. Willis, G. Sapparapu, S. Murrell, J. P. Julien, V. Singh, H. G. King, Y. Xia, J. A. Pickens, C. C. LaBranche, J. C. Slaughter, D. C. Montefiori, I. A. Wilson, J. Meiler and J. E. Crowe, Jr.; "Redesigned HIV antibodies exhibit enhanced neutralizing potency and breadth"; *J Clin Invest*; **2015**; Vol. 125 (6): p. 2523-31.

Position Specific Scoring Matrix for Screening of Candidate Antibodies

- Do HIV-Naïve Humans have PG9-like Antibodies?

- Screening of 25,000 HCDR3s using Rosetta-Inspired PSSMs
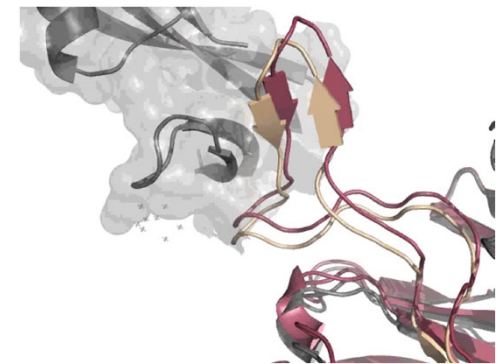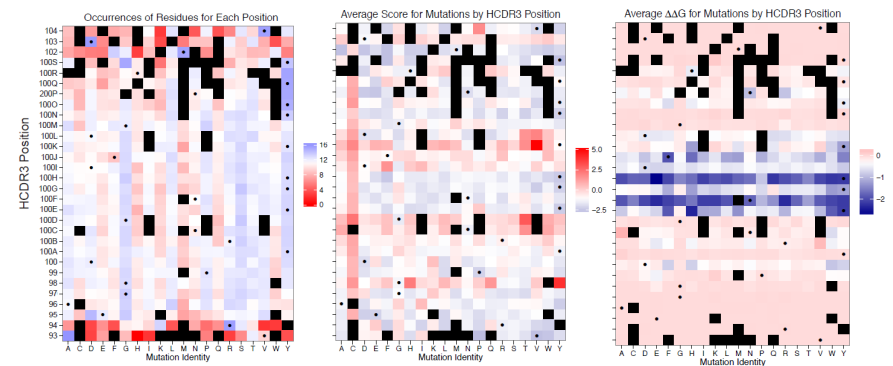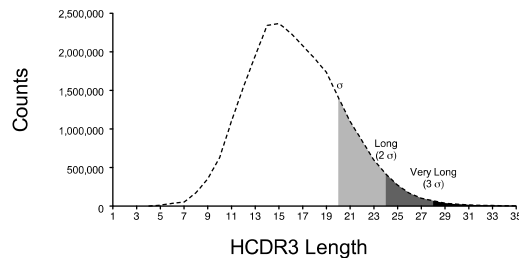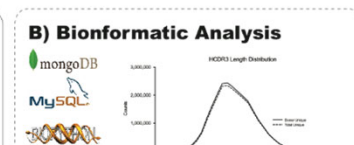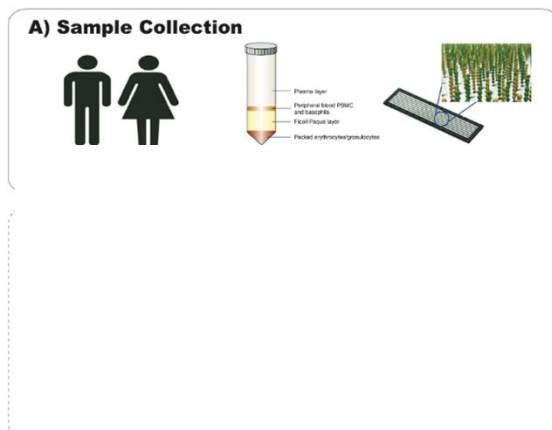
Finn, J. A., Dong, J., Sevy, A. M., Parrish, E., Gilchuk, I., Nargi, R., Scarlett-Jones, M., Reichard, W., Bombardi, R., Voss, T. G., Meiler, J., & Crowe, J. E., Jr. (2020). Identification of Structurally Related Antibodies in Antibody Sequence Databases Using Rosetta-Derived Position-Specific Scoring. *Structure, 28(10), 1124-1130 e1125. https://doi.org/10.1016/j.str.2020.07.012*
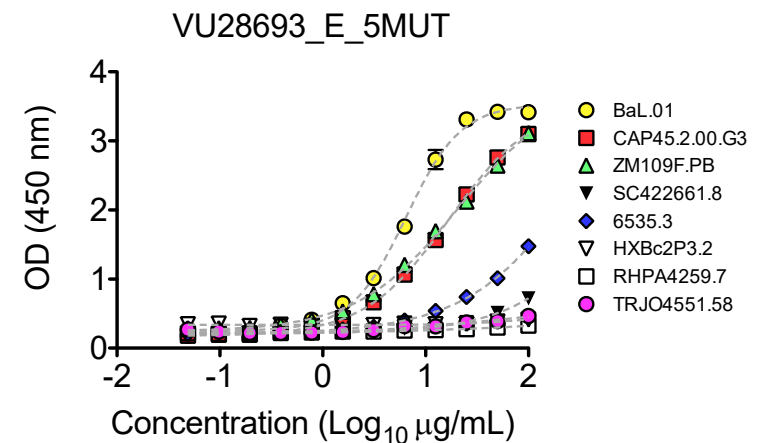
In silico Affinity Maturation of Candidate Antibody HCDR3s

- Rosetta Design
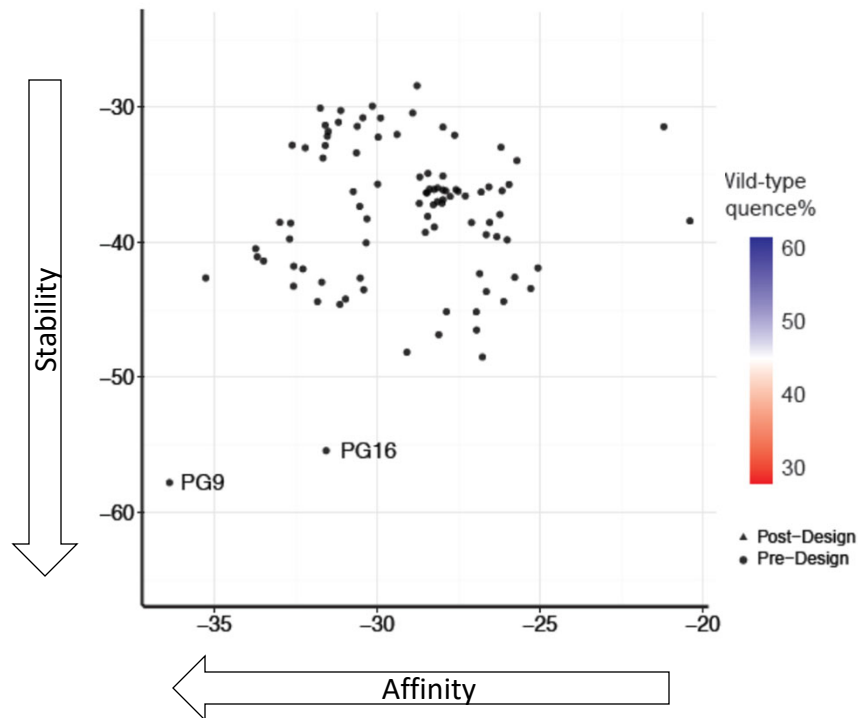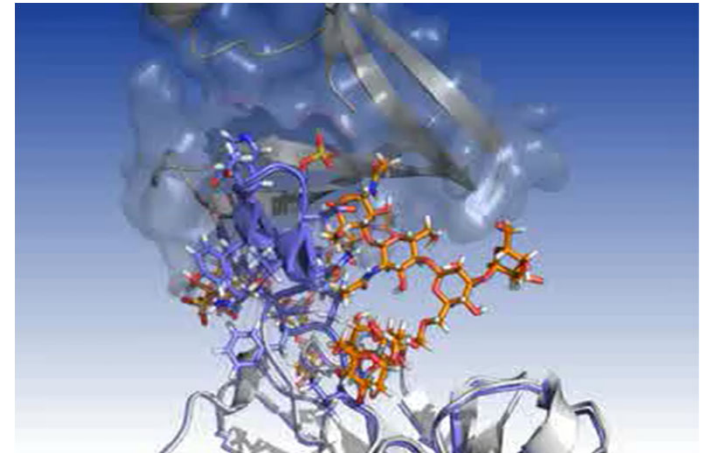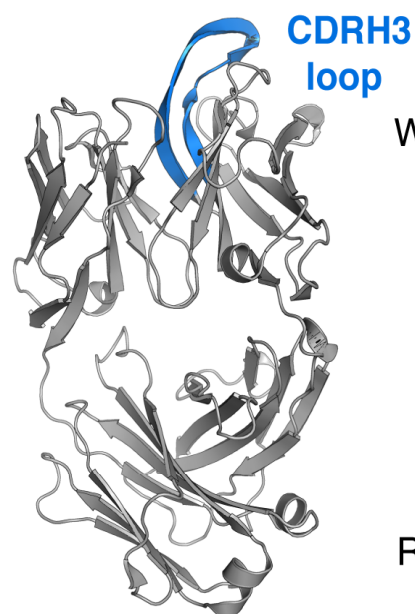

- Experimental Verification

J. R. Willis, ..., J. Meiler and J. E. Crowe, Jr.; "Long antibody HCDR3s from HIV-naive donors presented on a PG9 neutralizing antibody background mediate HIV neutralization"; Proc Natl Acad Sci U S A; 2016; Vol. 113 (16): p. 4446-51.
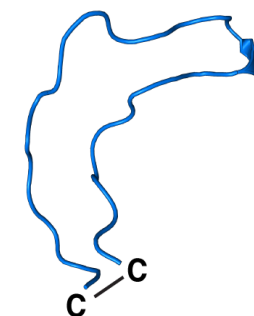
CDRH3-based cyclic peptides targeting influenza
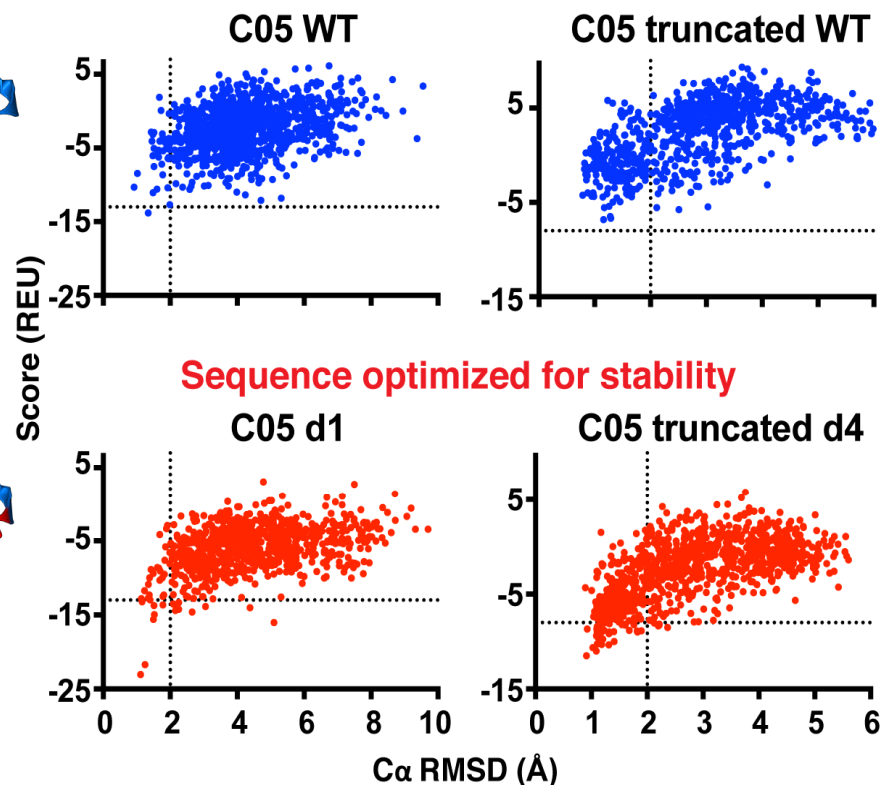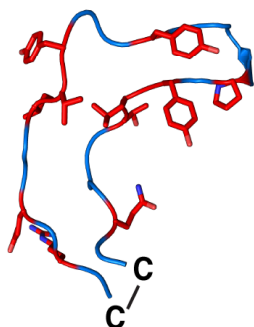
A. Influenza antibody with long CDRH3 loop

B. Folding simulations with RoSETTA

Alexander M Sevy, Iuliia M. Gilchuk, Rachel Nargi, Mattie Jensen, Jens Meiler, James E. Crowe; "Computationally designed cyclic peptides derived from an antibody loop increase breadth of binding for influenza variants; submitted

C05-based cyclic peptides have increased breadth of HA recognition

Group	Subtype	Strain	C05 d1	C05 d4	C05 IgG
1	H1N1	A/Solomon Islands/03/2006	+++	+++	++++
		A/Solomon Islands/03/2006 head domain	++	++	+++
		A/Brevig Mission/1/1918	-	-	-
		A/Tottori/YK012/2011	-	-	-
		A/mallard/Alberta/35/1976	-	-	++
		A/Puerto Rico/8/1934	+++	+++	-
		A/Texas/36/1991	-	-	-
		A/New Caledonia/20/1999	+++	+++	++
		A/California/04/2009	-	++++	-
	H2N2	A/Japan/305/1957	+++	++	++++
		A/Singapore/1/1957	+++	+++	++++
	H5N1	A/Vietnam/1203/2005	-	-	-
		A/Indonesia/5/2005	-	-	-
	H9N2	A/turkey/Wisconsin/1/1966	++++	+++	++
	H16N3	A/black-headed gull/Sweden/4/1999	-	-	-
2	H3N2	A/Hong Kong/1/68	+++	+++	++++
		A/Brisbane/10/2007	+++	+++	++++
		A/Perth/16/2009	+++	-	++++
		A/Panama/2007/1999	-	-	++++
		A/Bangkok/1/1979	-	-	-
	H4N6	A/duck/Czechoslovakia/1956	+++	+++	-
	H7N9	A/Shanghai/02/2013	+++	+++	-
		A/Netherlands/219/2003	-	-	-
	H15N8	A/shearwater/Western Australia/2576/1979	-	-	-

## Legend

++++	<10 nM	Gray: No change in breadth compared to IgG
+++	10-100 nM	Green: Gain of breadth compared to IgG
++	100-1,000 nM	Orange: Loss of breadth compared to IgG
+	>1,000 nM	
-	Binding not detected	
NT	Not tested	

Proof of principle for epitope-focused vaccine design: respiratory syncytial virus

Vaccines prevent infectious disease largely by inducing protective neutralizing antibodies against vulnerable epitopes. Several major pathogens have resisted traditional vaccine development, although vulnerable epitopes targeted by neutralizing antibodies have been identified for several such cases. Hence, new vaccine design methods to induce epitope-specific neutralizing antibodies are needed. Here we show, with a neutralization epitope from respiratory syncytial virus, that computational protein design can generate small, thermally and conformationally stable protein scaffolds that accurately mimic the viral epitope structure and induce potent neutralizing antibodies. These scaffolds represent promising leads for the research and development of a human respiratory syncytial virus vaccine needed to protect infants, young children and the elderly. More generally, the results provide proof of principle for epitope-focused and scaffold-based vaccine design, and encourage the evaluation and further development of these strategies for a variety of other vaccine targets, including antigenically highly variable pathogens such as human immunodeficiency virus and influenza.
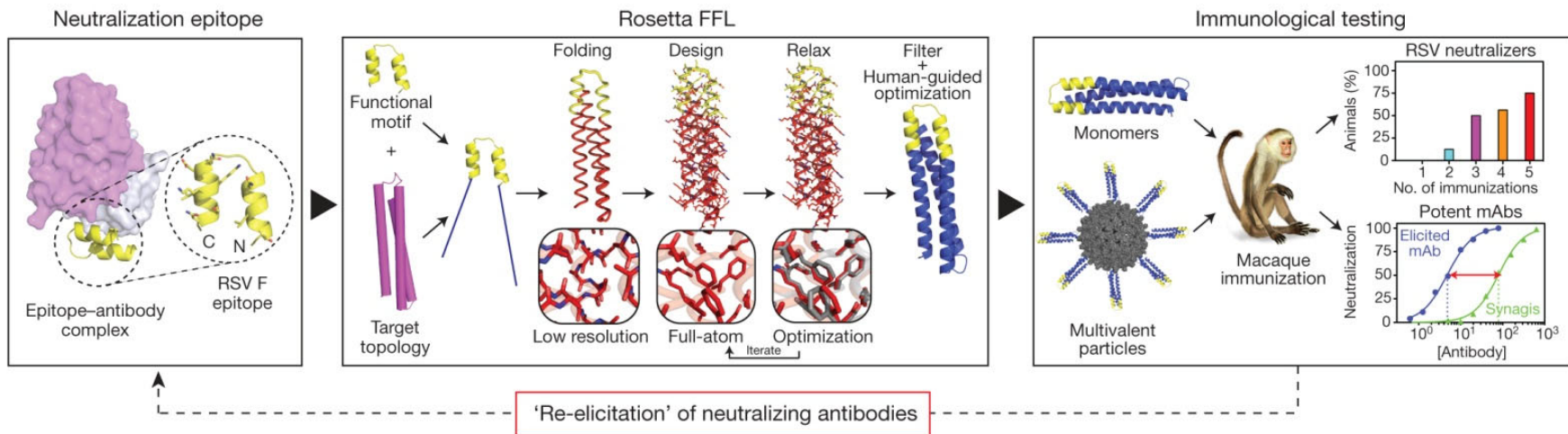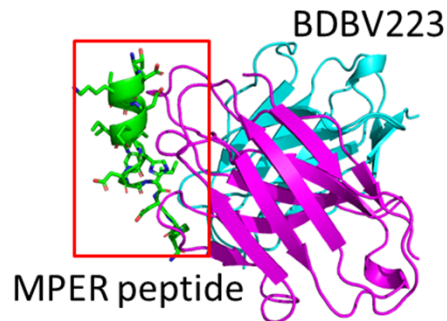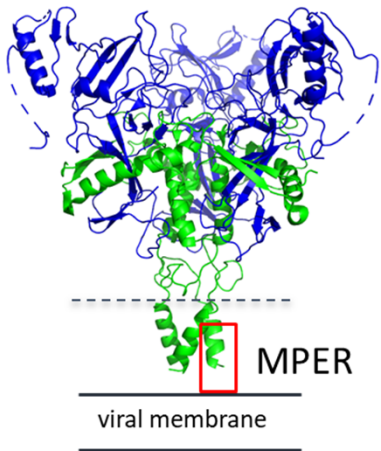
B. E. Correia, J. T. Bates, R. J. Loomis, G. Baneyx, C. Carrico, J. G. Jardine, P. Rupert, C. Correnti, O. Kalyuzhniy, V. Vittal, M. J. Connell, E. Stevens, A. Schroeter, M. Chen, S. Macpherson, A. M. Serra, Y. Adachi, M. A. Holmes, Y. Li, R. E. Klevit, B. S. Graham, R. T. Wyatt, D. Baker, R. K. Strong, J. E. Crowe, Jr., P. R. Johnson and W. R. Schief; "Proof of principle for epitope-focused vaccine design"; *Nature*; **2014**; Vol. 507 (7491): p. 201-6.
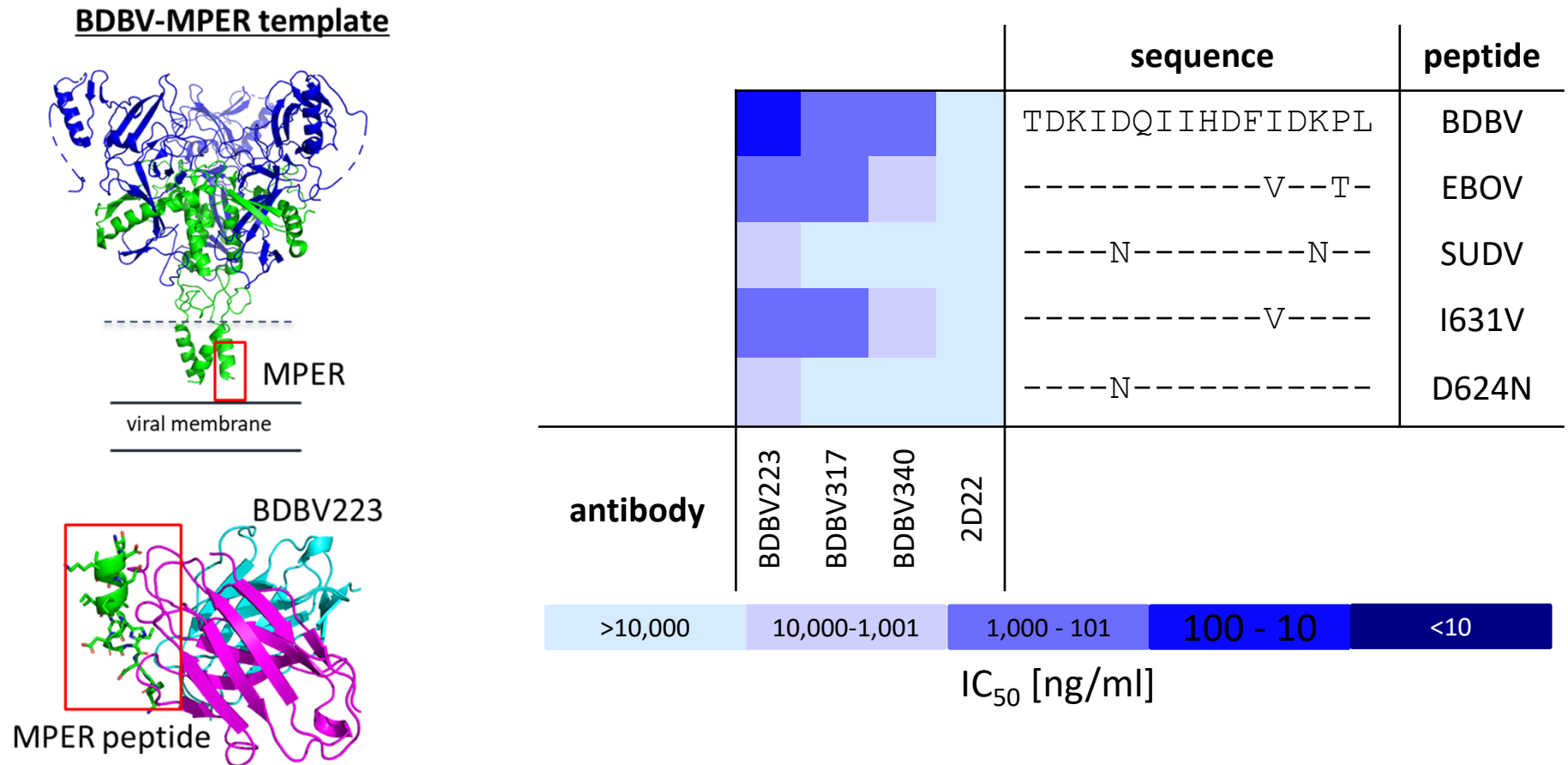
Epitope-focused vaccine design to elicit HR2/MPER antibodies

BDBV-MPER template

MPER

viral membrane

BDBV223

MPER peptide

Binding studies using the three BDBV-MPER targeting antibodies

BDBV-MPER template

MPER

viral membrane

BDBV223

MPER peptide

antibody	BDBV223	BDBV317	BDBV340	2D22	sequence	peptide
					TDKIDQIIHDFIDKPL	BDBV
					-----------V--T-	EBOV
					----N---------N--	SUDV
					-----------V----	I631V
					----N-----------	D624N

>10,000	10,000-1,001	1,000 - 101	100 - 10	<10

IC$_{50}$ [ng/ml]

Epitope-focused vaccine design to elicit HR2/MPER antibodies

BDBV-MPER template

MPER

viral membrane

BDBV223

MPER peptide

Rosetta immunogen design

Rosetta Commons
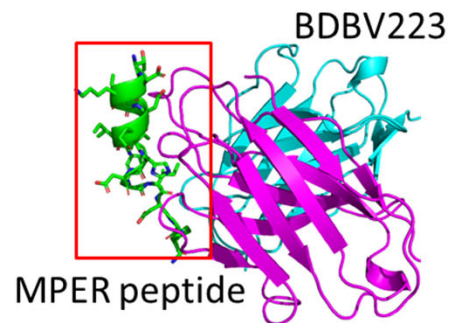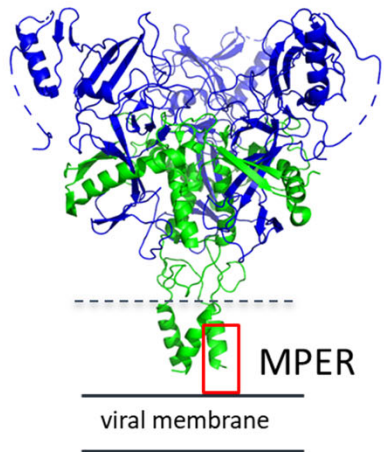
Sidechain Grafting
Backbone Grafting
FoldFromLoops
(1. Gen.)
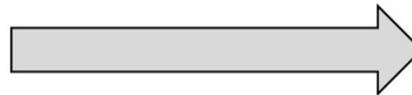
PDB crystal structures
< 2.5 A resolution
E. coli expressed

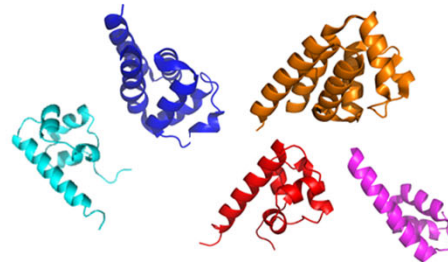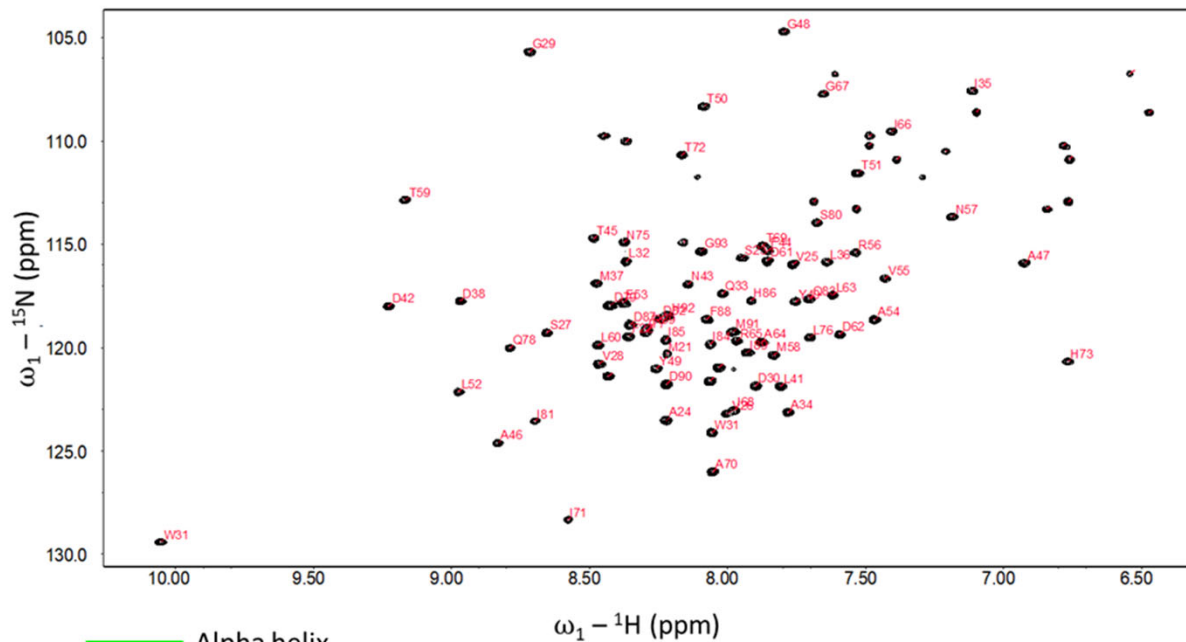# Epitope-focused vaccine design to elicit HR2/MPER antibodies
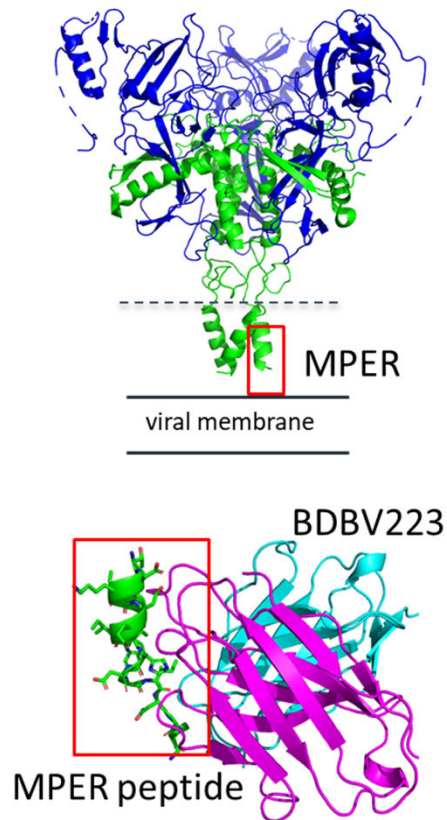
BDBV-MPER template

MPER

viral membrane

MPER peptide

BDBV223

Rosetta immunogen design

Rosetta Commons

Sidechain Grafting
Backbone Grafting
FoldFromLoops
(1. Gen.)
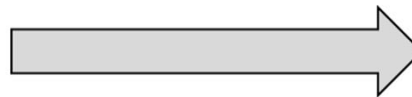
PDB crystal structures
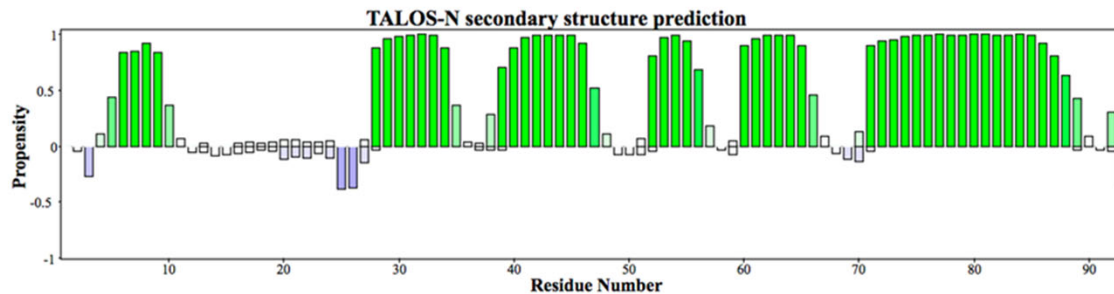< 2.5 A resolution
E. coli expressed

Immunogen testing

design 1
design 2
design 3
design 4
design 5
design 6

FoldFromLoops
backbone graft
sidechain graft
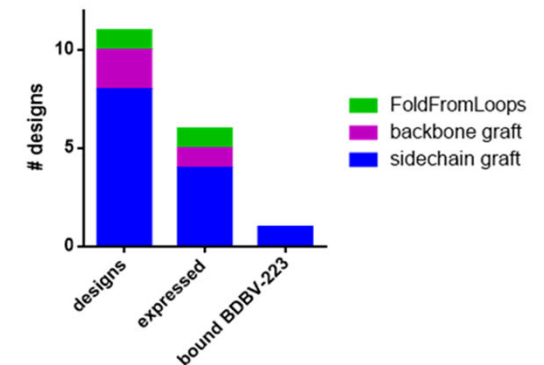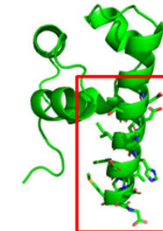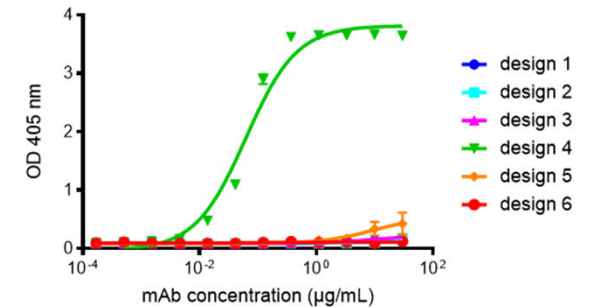
^1H-^{15}N HSQC spectrum showing the assignment for the MPER immunogen

Presentation on Self-Assembling Particle Platform

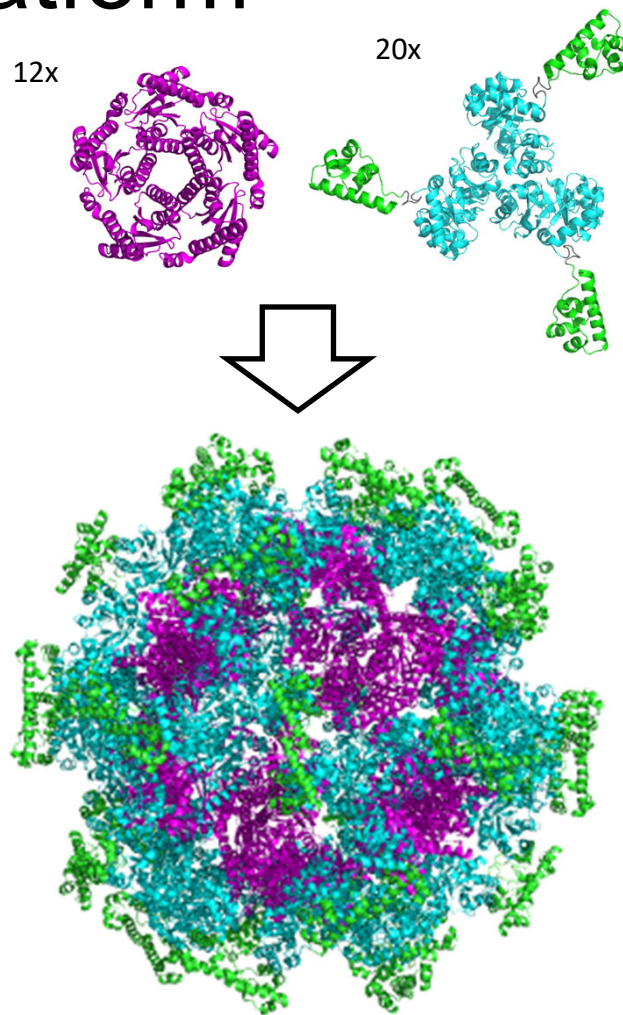12x   20x

Dynamic light scattering

ELISA binding

- BDBV-MPER immunogen
- empty assembly
- BDBV-MPER assembly
- BDBV-GP
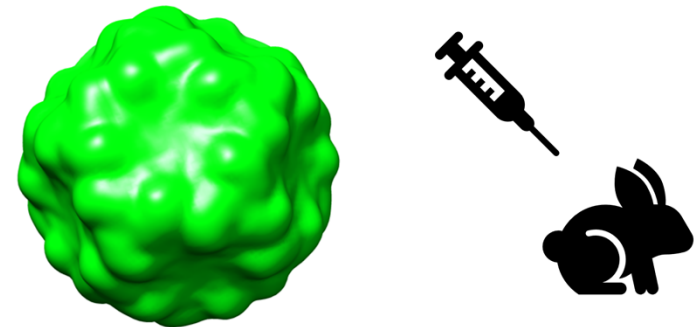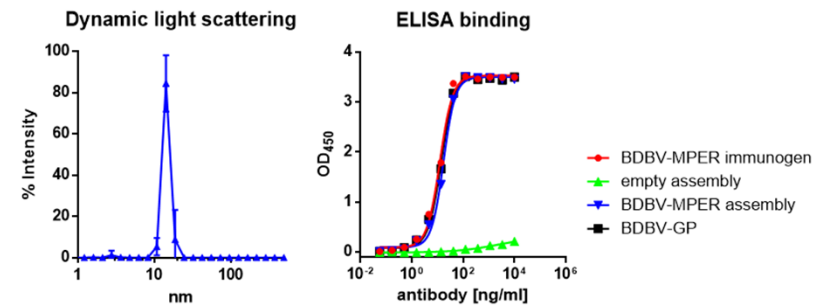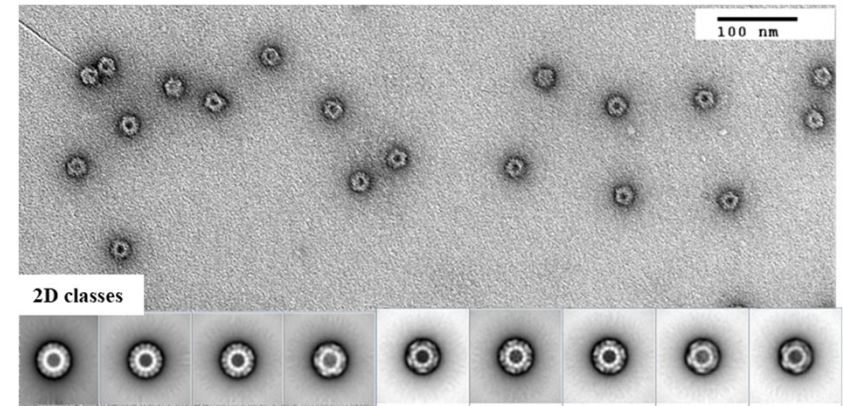
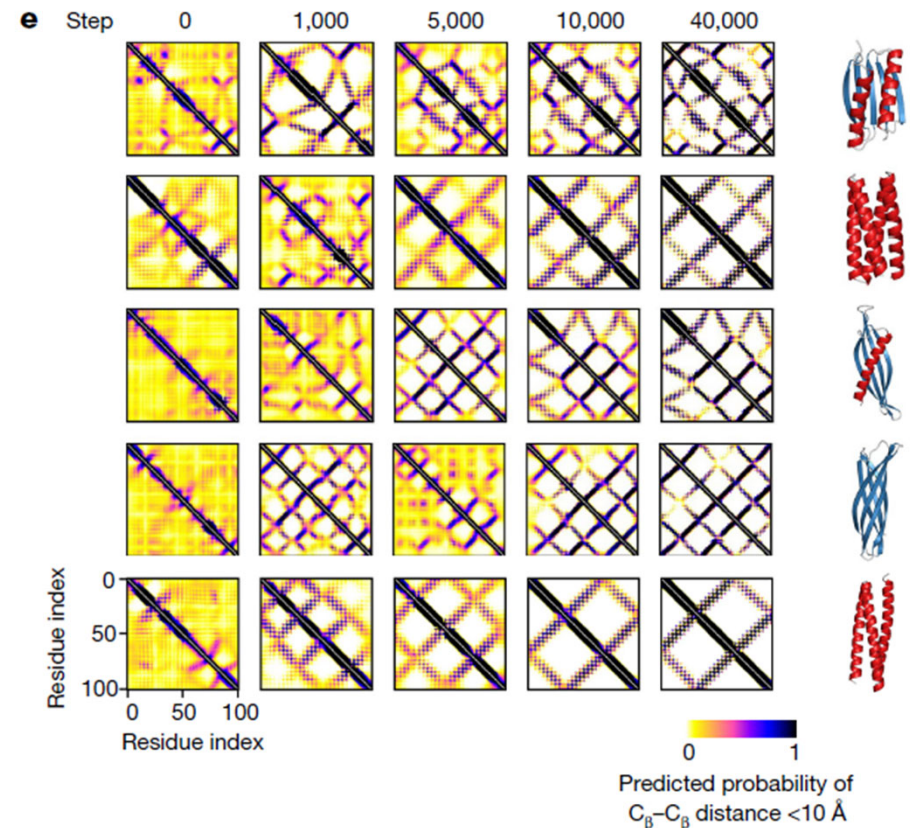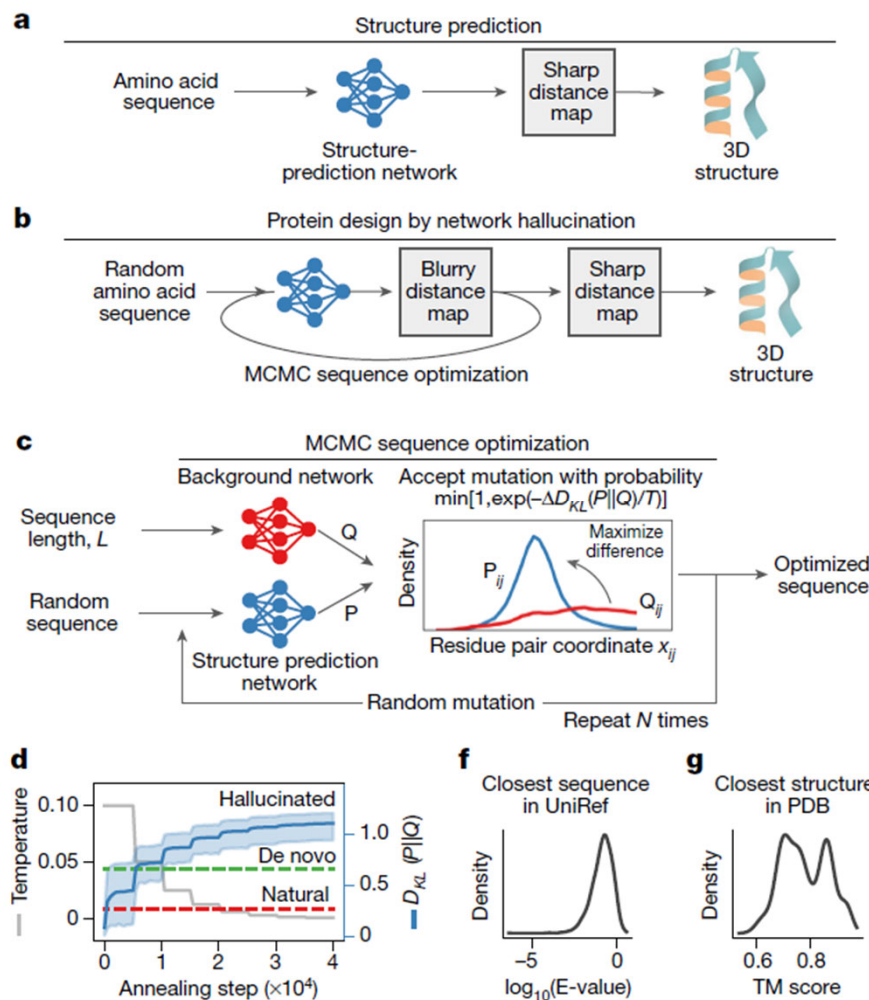# Rosetta Antibody Design biased to only create Human-Like Antibodies
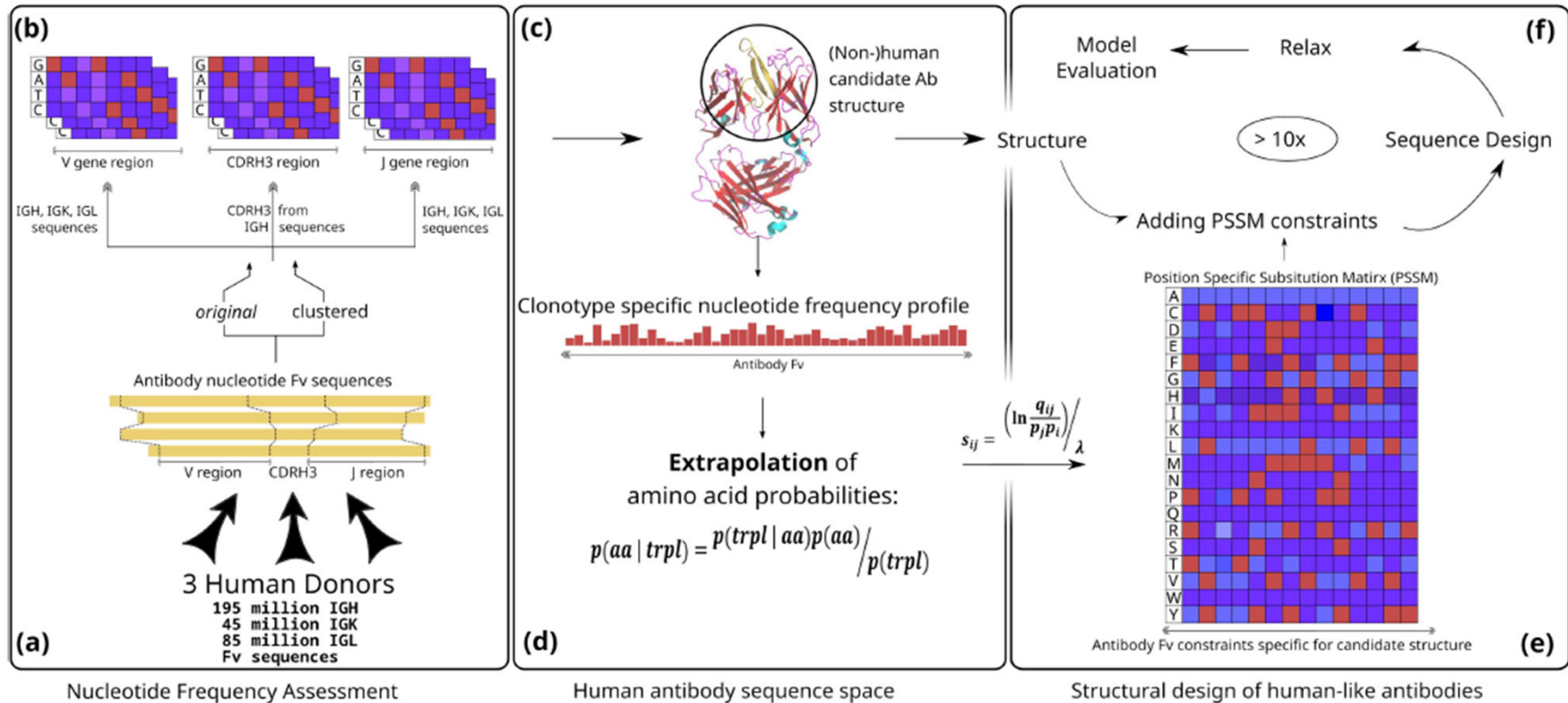
Schmitz, S., Schmitz, E. A., Crowe, J. E., Jr., & Meiler, J. (2022). The human antibody sequence space and structural design of the V, J regions, and CDRH3 with Rosetta. *MAbs, 14(1), 2068212. https://doi.org/10.1080/19420862.2022.2068212*

Robust deep learning-based protein sequence design using ProteinMPNN

De novo Protein Design by Deep Network Hallucination

I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione and D. Baker; "De novo protein design by deep network hallucination"; *Nature; 2021; Vol. 600 (7889): p. 547-552.*

Acknowledgements

Current Members:
Abdullah Al Mamun
Alexander Fürll
Alican Gulsevin
Anja Landsmann
Benjamin Brown
Brennica Marlow
Carie Fortenberry
Carissa Li
Chris Jurich
Chris Moth
Claiborne Tydings
Cristina Martina
Davide Sala
Eli McDonald
Elleansar Okwei
Fabian Liessmann
Felipe Engelberger
Gustavo Araiza
Hope Woods
Jannis De Riz
Kaitlyn V. Ledwitch

Katherine Larochelle
Kortney Melancon
Kristina Vogel
Lance Liu
Lunkas v. Bredow
Mateusz Sklodowski
Moritz Ertelt
Michael Pritchard
Nathan Bloodworth
Oanh Vu
Paul Eisenhuth
Qianzhen Shao
Robert Mann
Rocco Moretti
Shannon Smith
Souhrid Mukherjee
Taylor Jones
Thomas Scott
Tracy Tang
Victoria Most
Vivian Ehrlich
Yidan Tang

Selected Alumni:
Annalen Bleckmann
Alex Sevy
Axel Fischer
Bian Li
Brett Kroncke
Brian Bender
Brian Weiner
Clara Schoeder
David Nannemann
Diego del Alamo
Georg Kuenze
Jeff Mendenhall
Jordon Willis
Julia Koehler Leman
Liz Dong Nguyen
Mariusz Butkiewicz
Nathan Alexander
Nils Woetzel
Nina Bozhanova
Steffen Lindert
Will Lowe