Protein structure prediction

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

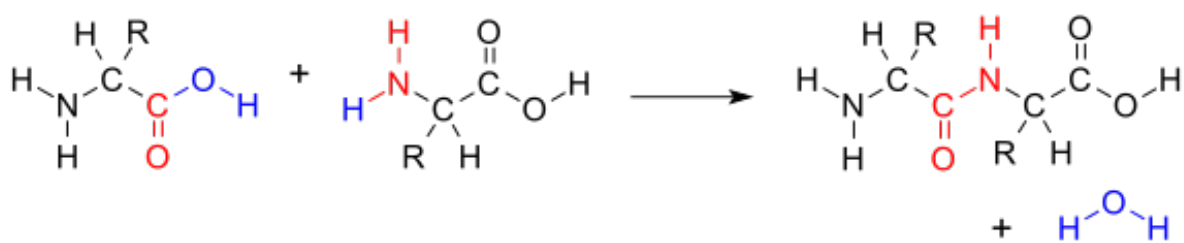
http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

Overview

- Primary structure and higher levels of protein description
  - what is it? what is protein folding?
  - why is protein structure improtant?
  - why do we need its computational prediction?
- experimental determination of protein structure
  - e.g. X-ray crystallography, expensive and time consuming,
- computational protein folding prediction methods
  - templates, contact predictions,
- breakthrough in 2018 and 2020
  - AlphaFold, DeepMind, deep neural networks,
  - Nature journal, the result verified in CASP competition,
  - the consequences for molecular biology and medicine.

Protein architecture

- Protein = a large macromolecule comprised of one or more long chains of amino acid residues
  - the residues linked be peptide bonds (strong covalent bonds),
- The structure of amino acids
  - a central carbon atom (α-carbon), an amino group (NH_2),
  - a carboxyl group (COOH), a side chain (R),
- Side chains distinguish between amino acids
  - amino acid properties such as polarity or charge.



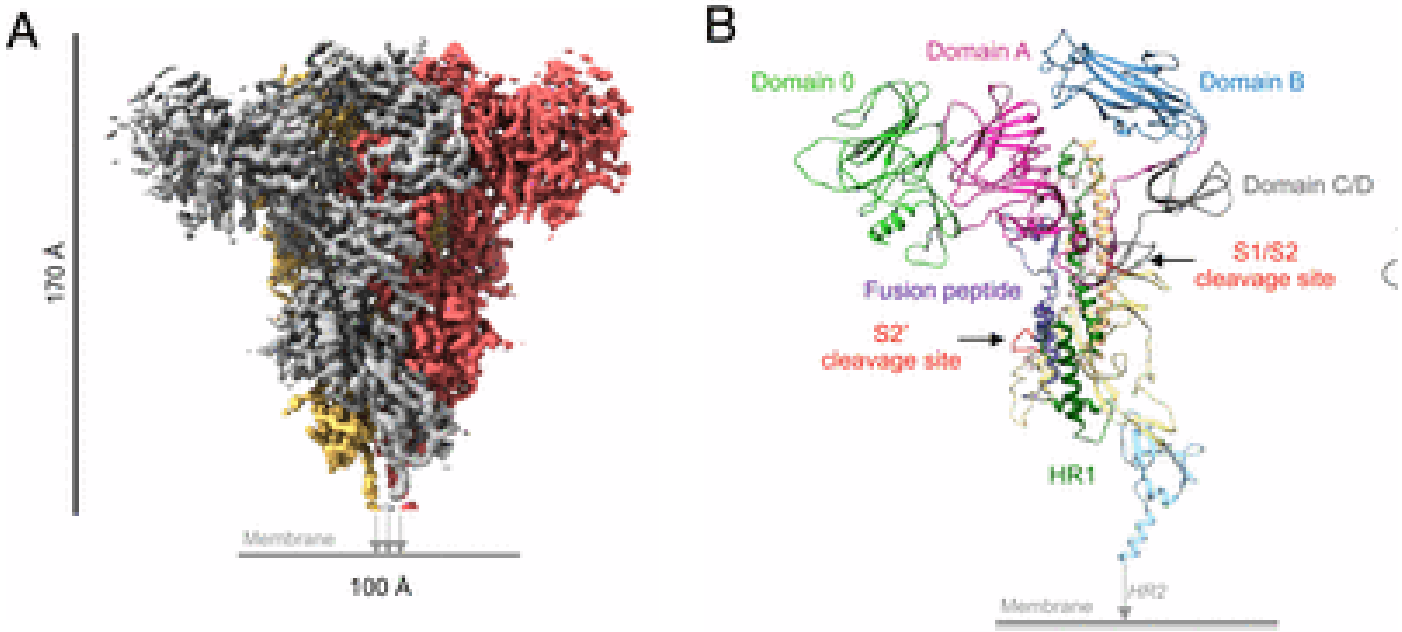
Peptide bond, Wikipedia.

Protein conformation

- Protein does not have only its main chain
  - its backbone is rich in hydrogen-bonding potential,
  - residue carbonyl groups are good hydrogen-bond acceptors,
  - residue NH groups are good hydrogen-bond donors,
  - proteins further **fold**,
- the other than hydrogen bonds among amino acids that influence folding
  - ionic bonds, disulfide bonds,
  - van der Waals forces, volume constraints, hydrophobic interactions,
- in general, the amino acid sequence of a protein determines its 3D shape = conformation/folding,
  - Christian Anfinsen, The Nobel prize in chemistry in 1972,
  - exceptions: denaturation, disordered proteins, chaperons, phospohorylation.

Protein conformation

- Levinthal's paradox, 1969
  - a very large number of degrees of freedom in an unfolded polypeptide chain,
  - an astronomical number of possible conformations, estimate was 10^{300},
  - most small proteins fold spontaneously on a milli or microsecond time scale,
  - folding is sped up and guided by the rapid formation of local interactions.



Yang et al': Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans, 2020.

Levels of protein description

TABLE 3.2 A Summary of Protein Structure

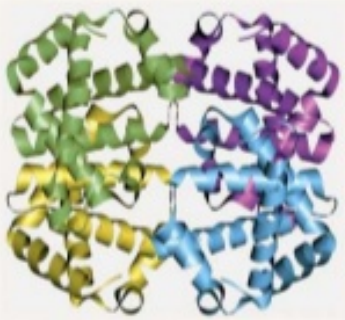
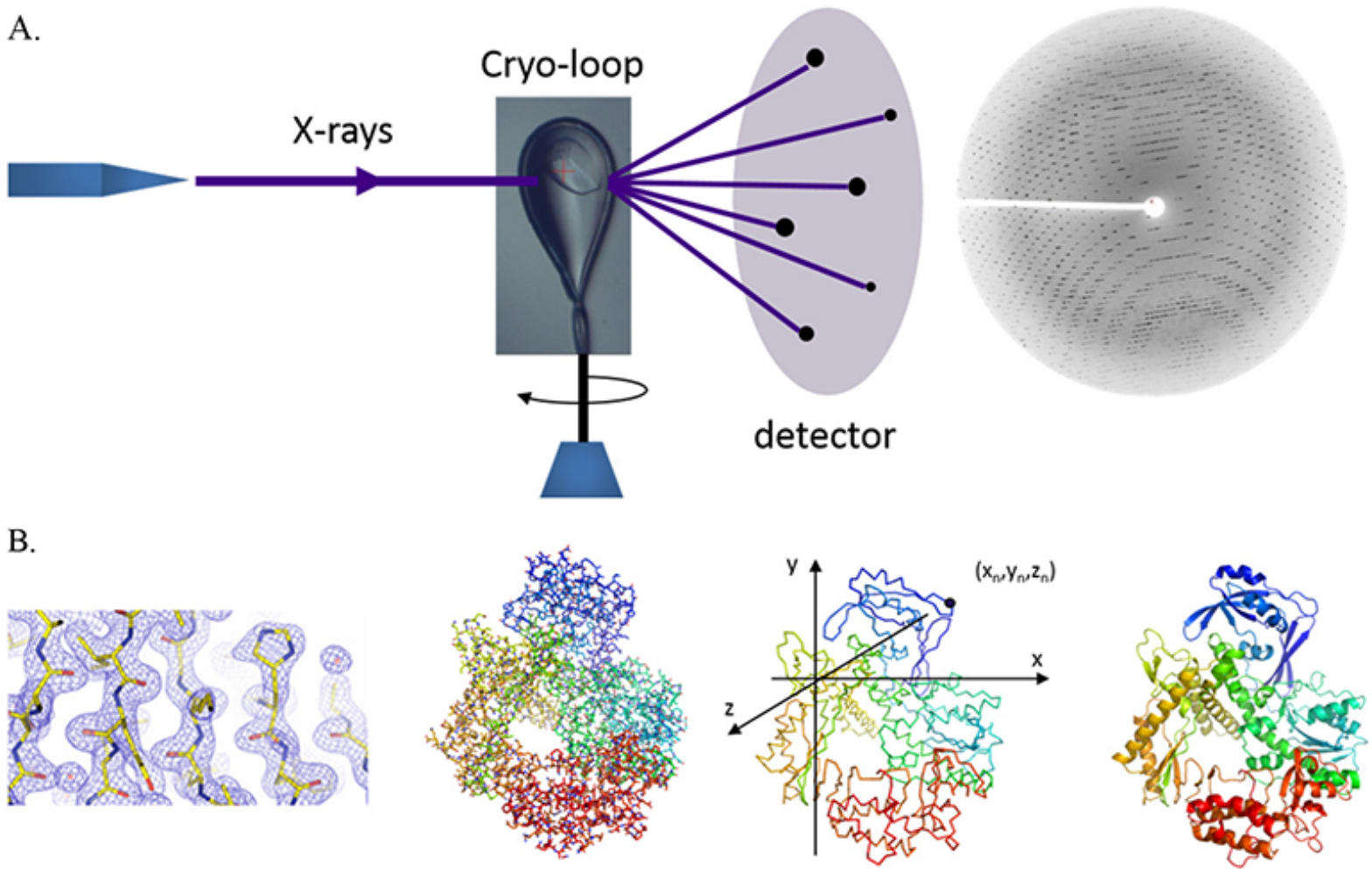
Level	Description	Stabilized by	Example: Hemoglobin
Primary	The sequence of amino acids in a polypeptide	Peptide bonds	
Secondary	Formation of α-helices and β-pleated sheets in a polypeptide	Hydrogen bonding between groups along the peptide-bonded backbone	
Tertiary	Overall three-dimensional shape of a polypeptide (The model on the right shows one of hemoglobin's subunits. The black and red atoms are in the heme group that carries oxygen; they are not part of the protein itself.)	Bonds and other interactions between R-groups, or between R-groups and the peptide-bonded backbone	
Quaternary	Shape produced by combinations of polypeptides. (The model on the right shows hemoglobin, which consists of four polypeptides.)	Bonds and other interactions between R-groups, and between peptide backbones of different polypeptides	

Table 3-2 Biological Science, 2/e

Experimental determination of protein structure

- Very expensive and time consuming
  - there is a large sequence-structure gap (many more sequences than structures),
- key methods
  - X-ray crystallography,
  - cryo-EM (cryogenic electron microscopy),
  - NMR (nuclear magnetic resonance) spectroscopy,
  - mass spectrometry,
- not perfectly accurate too
  - around 95% of the structure reported correctly,
- the main question
  - **can we approach this accuracy with computational methods?**

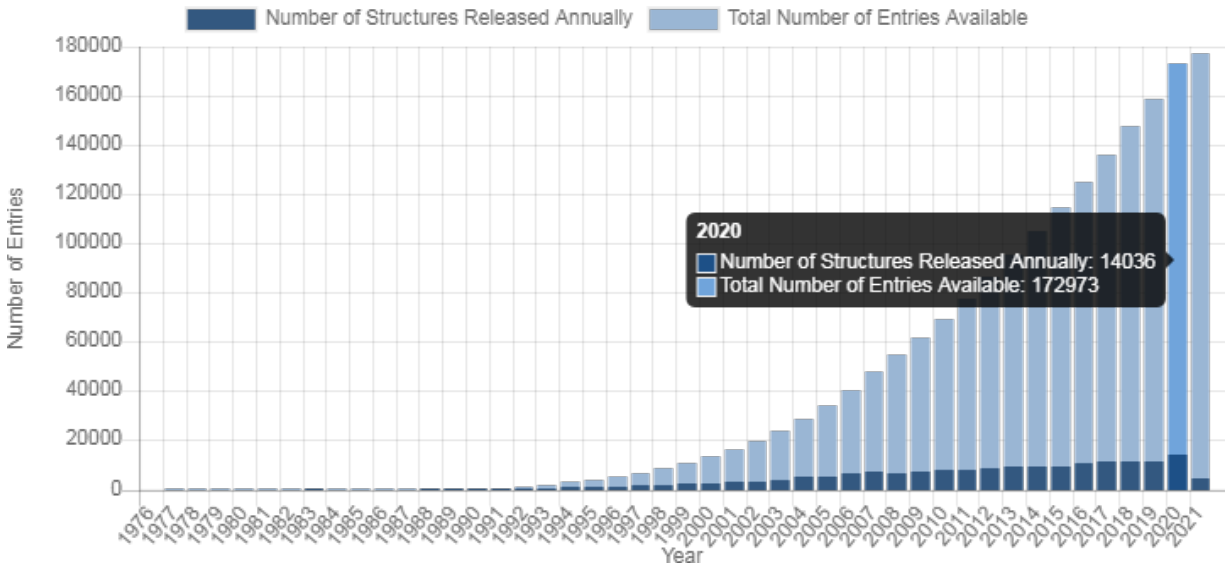
X-ray crystallography



Mayer: X-Ray Diffraction in Biology: How Can We See DNA and Proteins in Three Dimensions? 2017

Protein Data Bank (PDB)

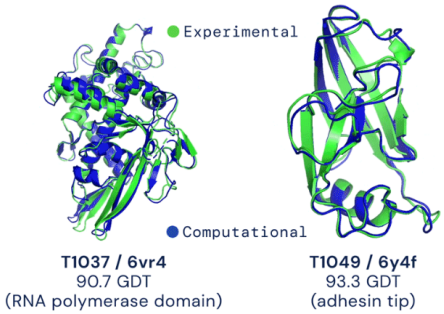
- PDB is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids,
- structures mostly obtained by the previously mentioned experimental methods,
- dedicated PDB and PDBML (XML) file formats, several free viewers exist,
- submission to the database required by most relevant journals when publishing.



PDB growth: https://www.rcsb.org/stats/growth/growth-released-structures.

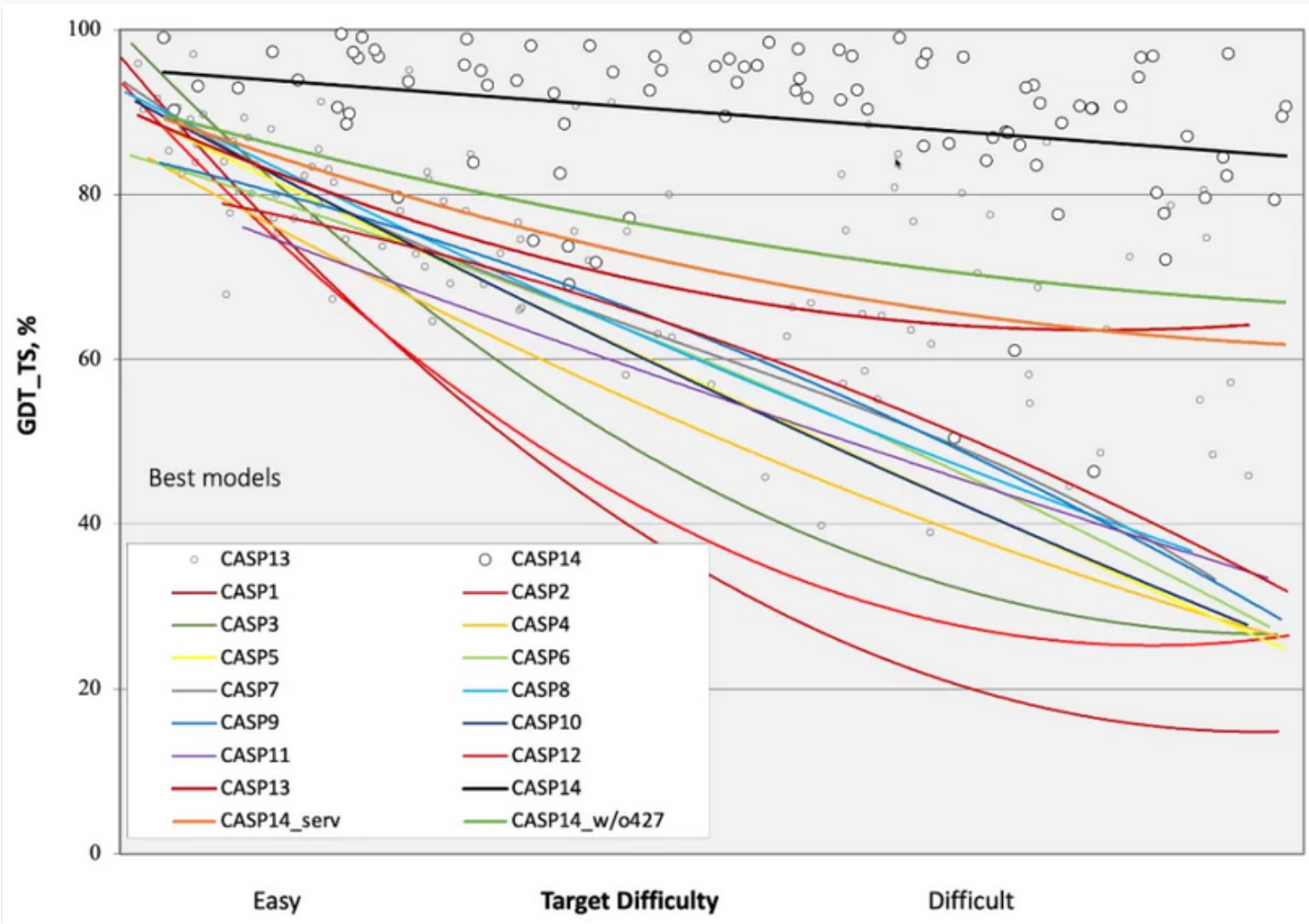
CASP

- CASP = Critical Assessment of protein Structure Prediction
  - informal olympics in computational protein folding,
  - competition + conference, https://predictioncenter.org/
  - independent assessment of methods of protein structure modeling,
  - taking place every two years since 1994,
  - structures that have just been solved and are kept on hold by PDB,
  - in 2020, CASP14
    * 120 (150) structures, 100 teams, 3 company teams,
    * AlphaFold2 (DeepMind, Google) reached global distance test scores above 90 out of 100 for about two-thirds of the CASP14 proteins.



https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

CASP competition results



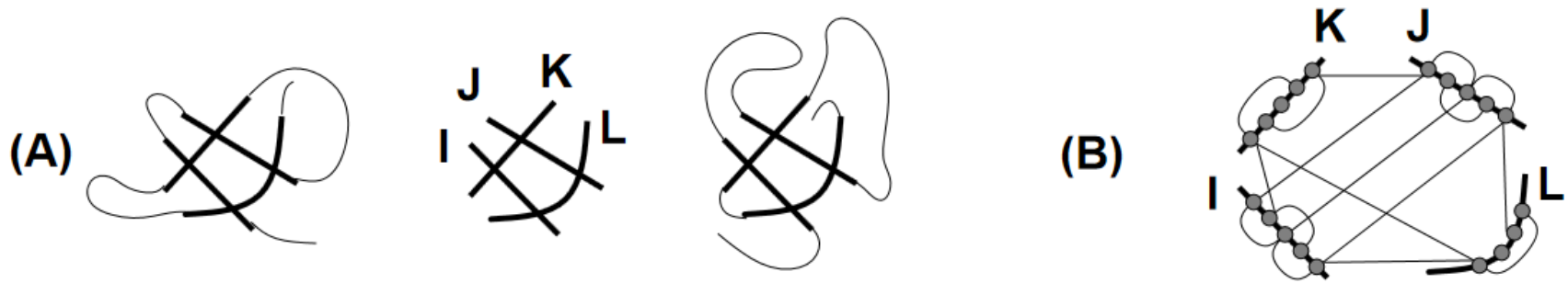
https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/.

Computational protein folding prediction methods

- Homology modeling
  - given: a query sequence Q and a database D of protein structures (PDB),
  - find: a protein P from D with high sequential similarity to Q,
  - return: P structure as an approximation of Q structure,
- molecular dynamics
  - given: a query sequence Q,
  - return: use laws of physics to simulate folding of Q,
- protein threading
  - given: a query sequence Q and a database D of known protein templates,
  - find: a template T from D that can be aligned with Q,
  - return: T as an approximation of Q structure.

Protein threading

- The key step
  - align a sequence to structure = template,
  - employ amino acid preferences for different structures,
  - an objective scoring function needed + search over space of alignments.
- Template
  - the core secondary structure segments (IJKL), loops unimportant,
  - different proteins map differently to a template (Fig A),
  - lines indicate interactions among AAs in template model (Fig B).



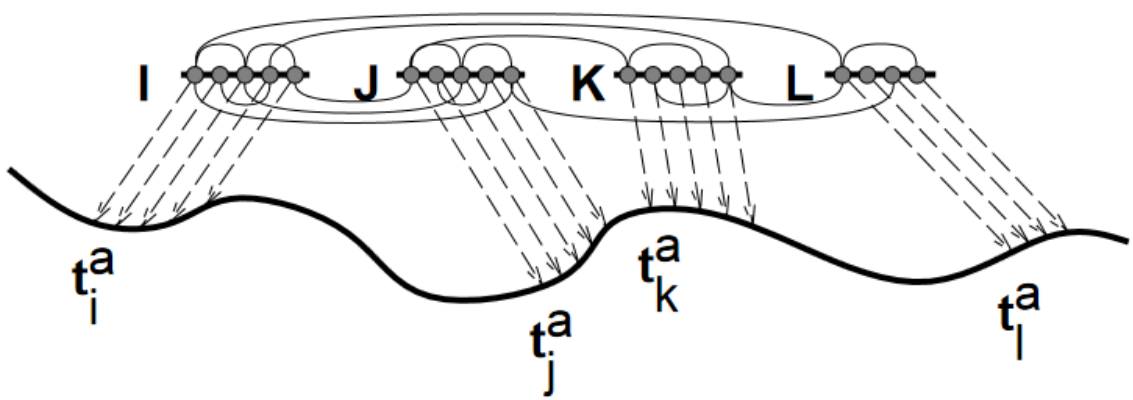
Lathrop et al.: Analysis and algorithms for protein sequence-structure alignment.

Protein threading

- The best threading
  - a sequence-template mapping that minimizes the objective function,
  - NP-hard optimization task, can be solved heuristically or branch & bound.
- An example of objective function with pairwise interactions between segments

$$f(\vec{t}) = \sum_i g_1(i, t_i) + \sum_i \sum_{j>i} g_2(i, j, t_i, t_j)$$

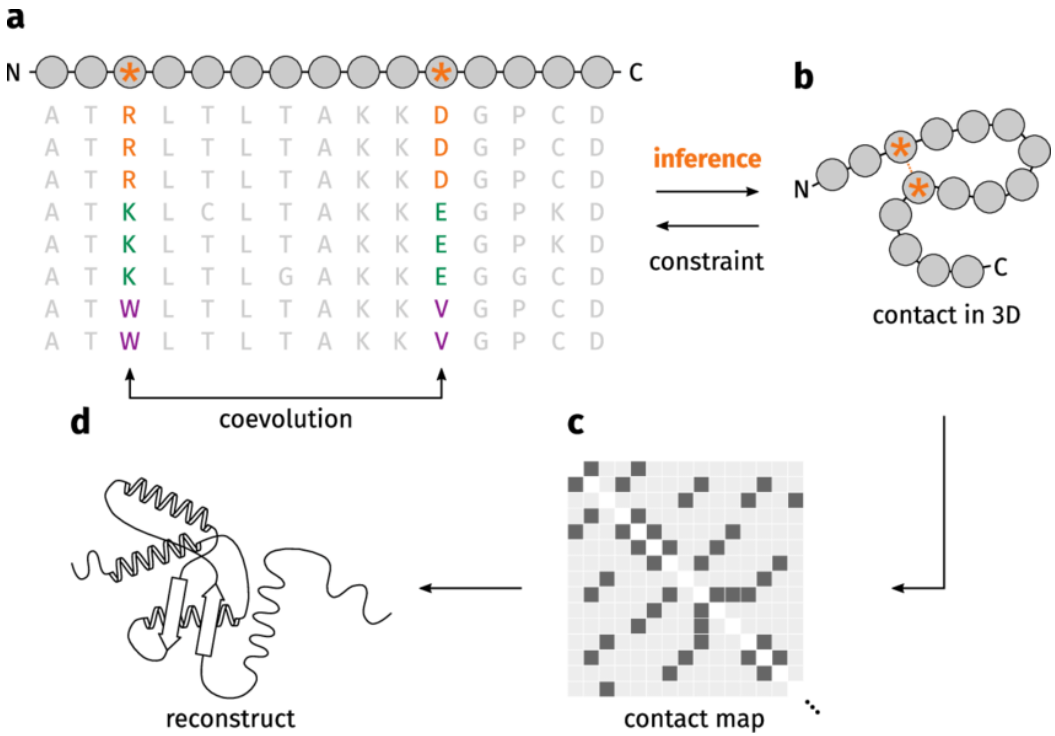
  - threading defined by a vector $\vec{t}$ whose elements indicate the indeces of amino acids placed in the first position of each core segment.



Lathrop et al.: Analysis and algorithms for protein sequence-structure alignment.

Protein structure prediction by co-evolution techniques

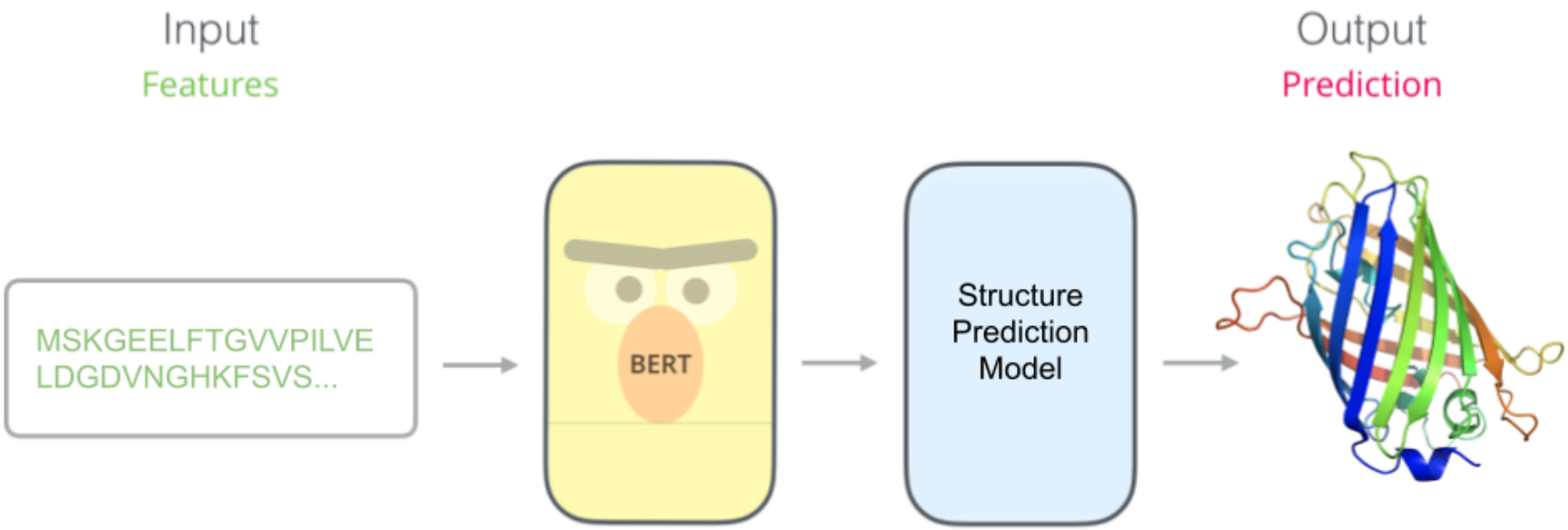
- Co-evolution techniques
  - two residues which mutate in a correlated fashion considered co-evolving,
  - co-evolution is interpreted as functional dependence and spatial proximity.



Bittrich et al.: StructureDistiller: Structural relevance scoring identifies the most informative entries of a contact map.

The role of pre-training in deep models

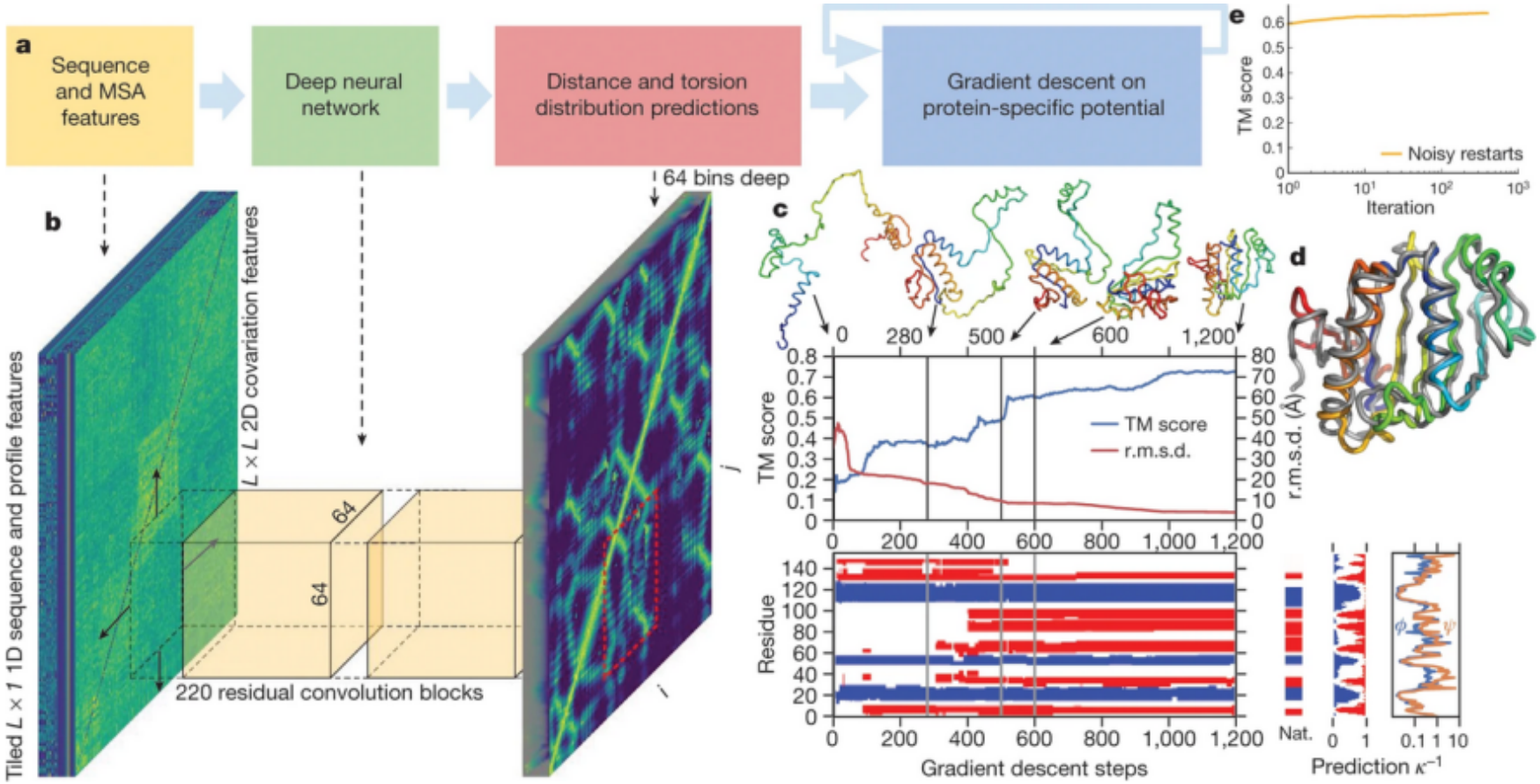
- Learn from protein sequences via self-supervision
  - there are hundreds of millions unlabeled protein sequences available,
  - do the same that has previously been done with large text corpora
    - train a model that can fill in a random part of masked protein sequence,
    - then use this model in a specific task (structure prediction in our case).



https://bair.berkeley.edu/blog/2019/11/04/proteins/

AlphaFold: the main ideas

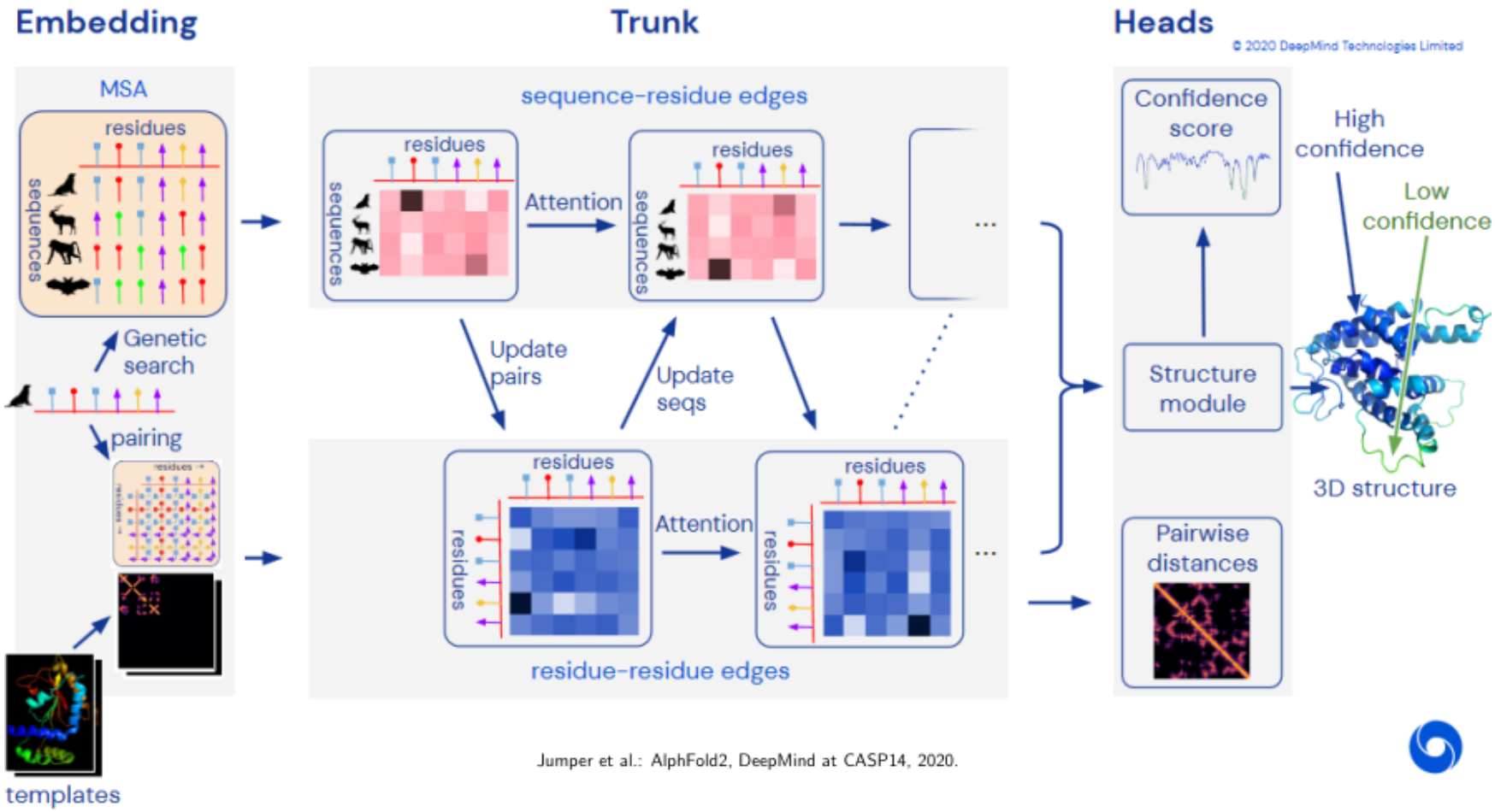
- Instead of contact map, predict full distance/torsion distributions with CNN,
- gradient descent to find the folding that fits the distance and torsion map.



Senior et al.: Improved protein structure prediction using potentials from deep learning, Nature, 2020.

AlphaFold2: the main ideas

- Different from previous that over-accounted for nearby residue interactions,
- sub-networks coupled together into a single differentiable end-to-end model.



Jumper et al.: AlphFold2, DeepMind at CASP14, 2020.

Summary

- A protein sequence of interest, what is its structure?
  - its structure available (in PDB)? simply use it, otherwise
  - BLAST against the protein sequences with available structures (in PDB),
  - search for protein domains with known structure,
  - apply computational methods, protein threading or other available,
  - energy minimization for very short proteins or finetuning,
- impact of computational predictive models
  - protein engineering for drug development.