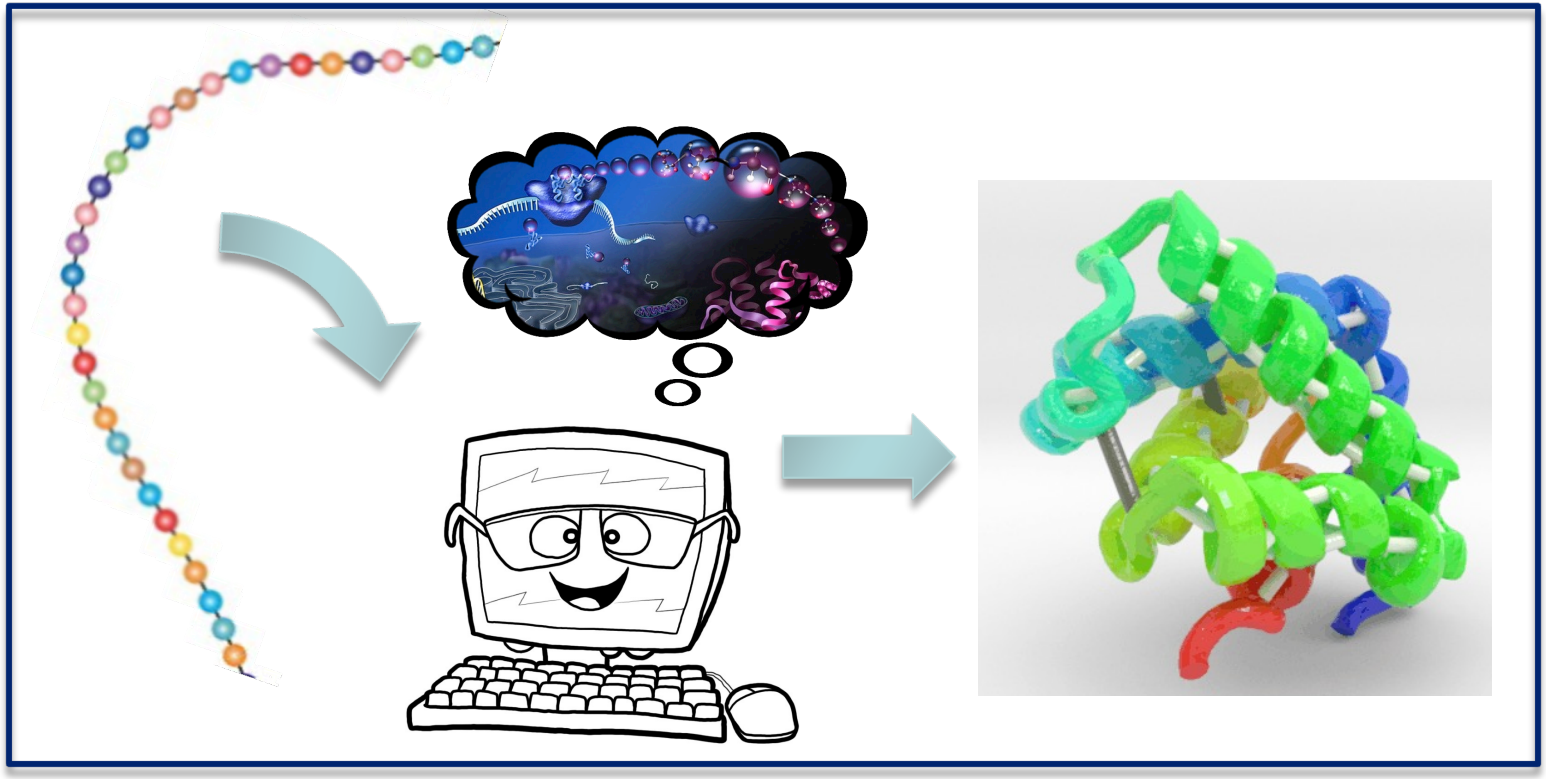
LOSCHMIDT LABORATORIES

Models of structures

3D structure prediction

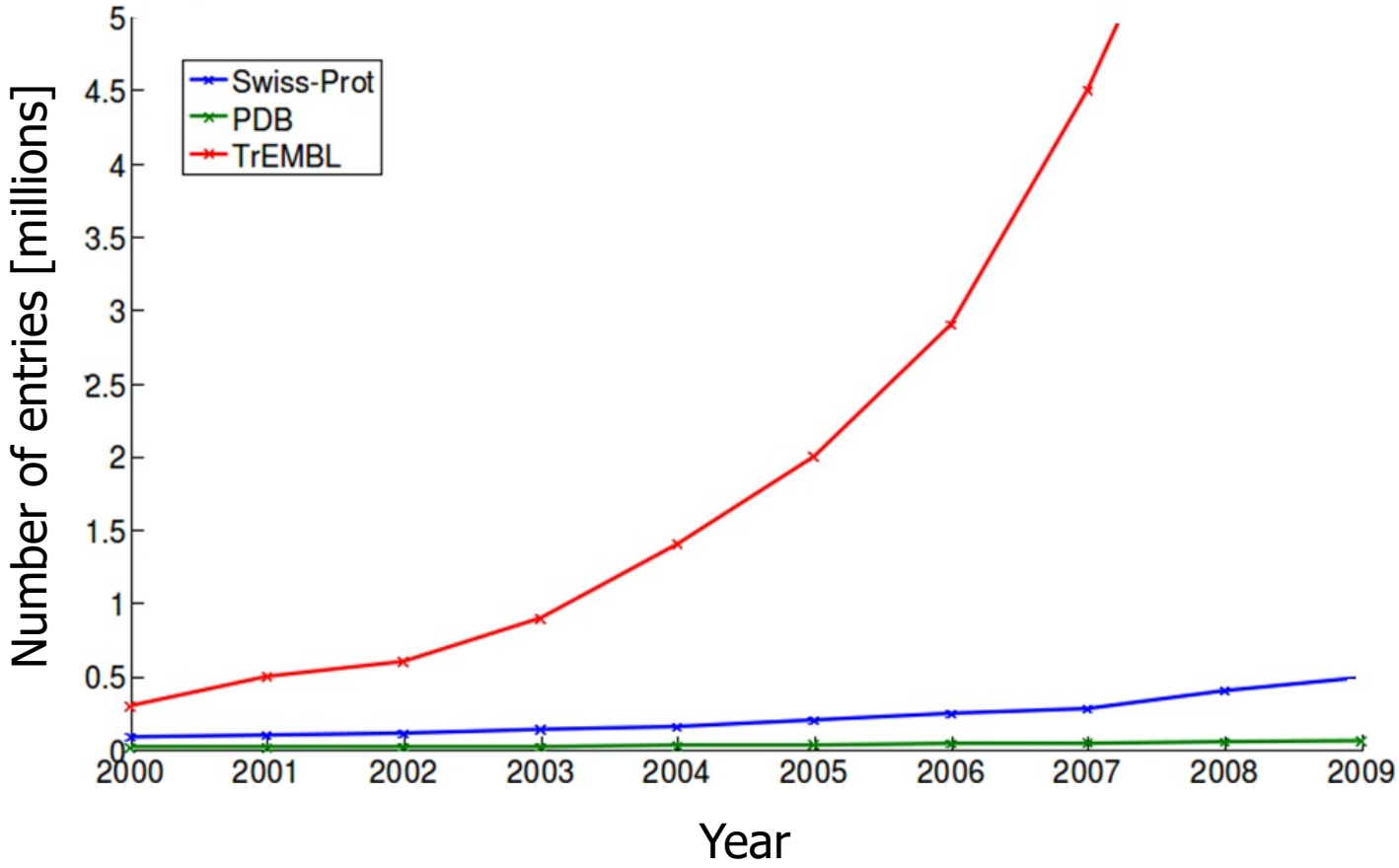


3D structure prediction

- homology modeling
- fold recognition
- *ab initio* prediction
- “hybrid” approaches
- Assesment
- databases of protein models

Importance of structure

- no experimental structure for most of the sequences

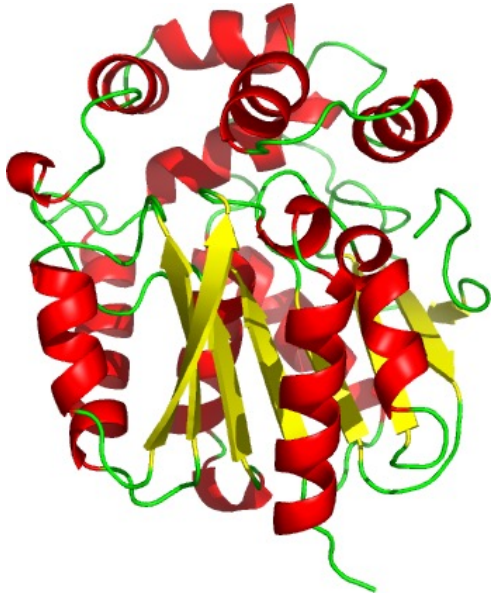


Homology modelling

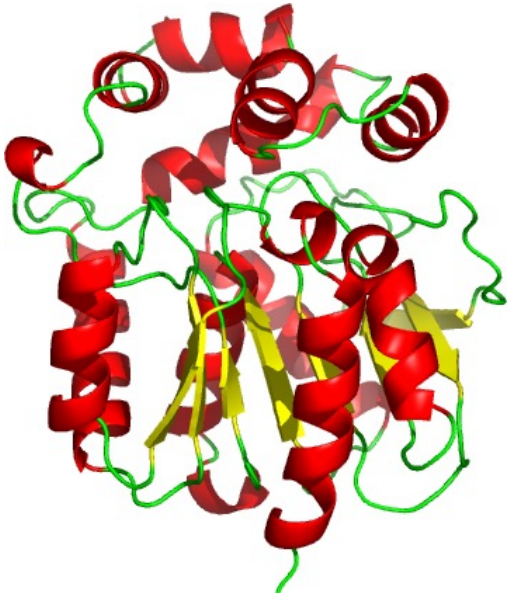
- basic principle – structure is more conserved than sequence

Homology modeling

- basic principle – structure is more conserved than sequence
  - similar sequences adopt practically identical structures



haloalkane dehalogenase
LinB (PDB-ID 1iz7)

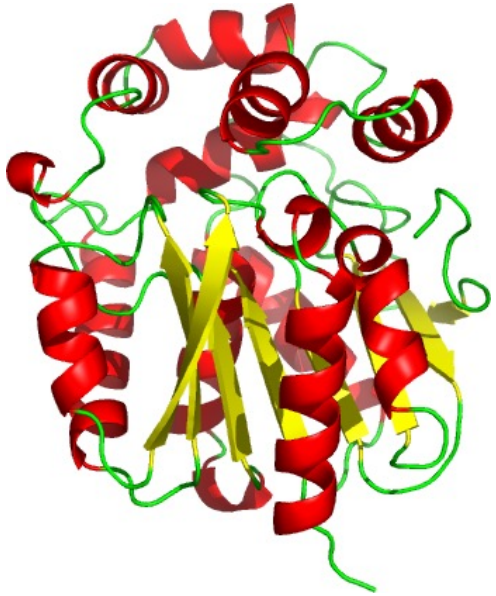


haloalkane dehalogenase
DhaA (PDB-ID 1cqw)

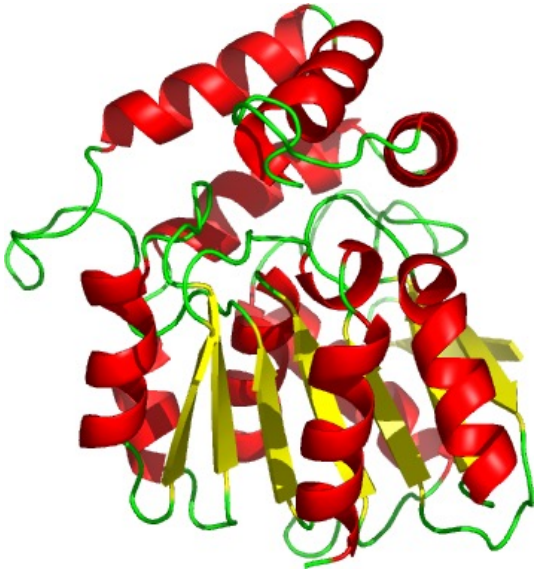
sequence identity: ~ 50 %

Homology modeling

- basic principle – structure is more conserved than sequence
  - distantly related sequences still fold into similar structures



haloalkane dehalogenase
LinB (PDB-ID 1iz7)

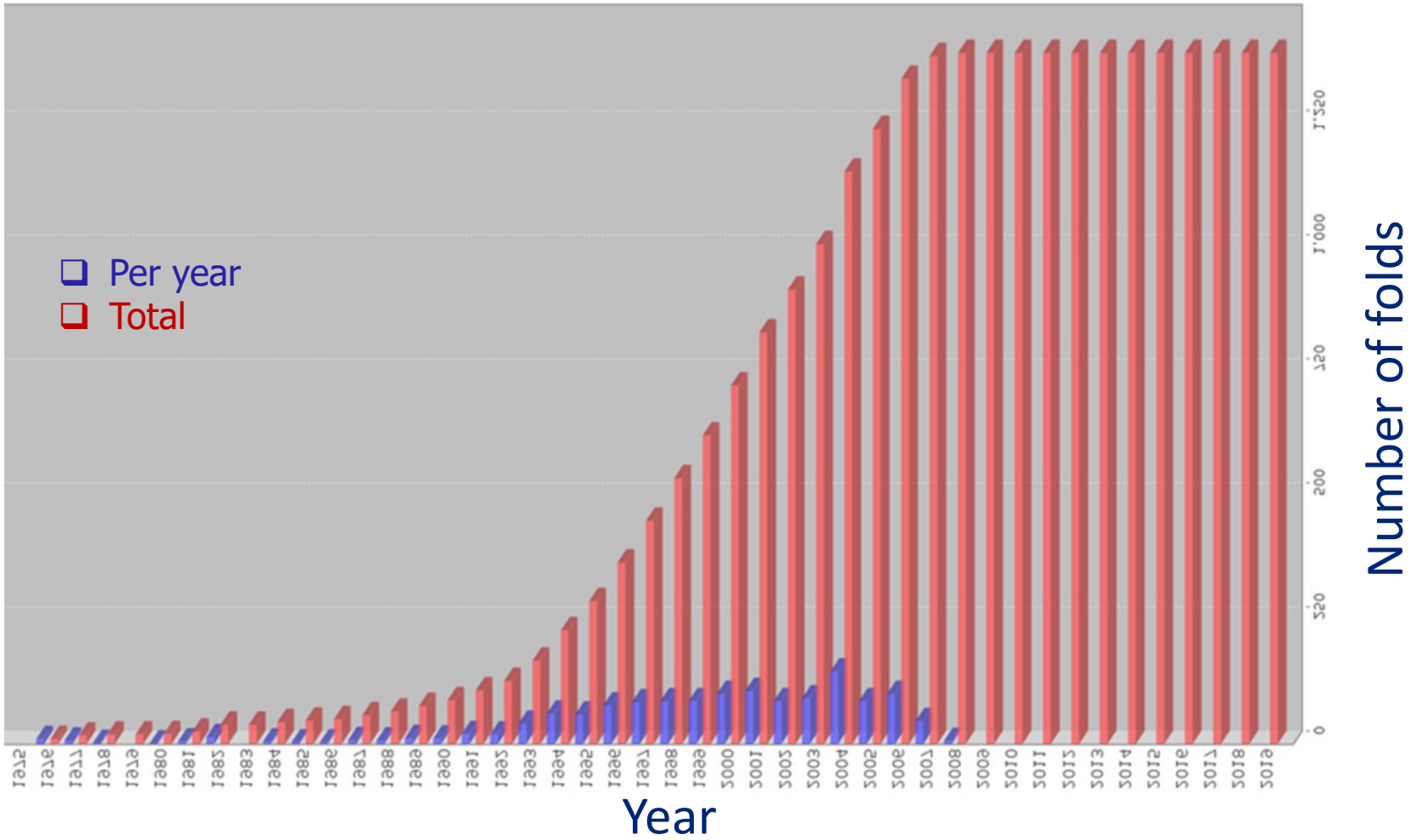


chloroperoxidase L
(PDB-ID 1a88)

sequence identity: ~ 15 %

Homology modeling

- number of folds in SCOP database

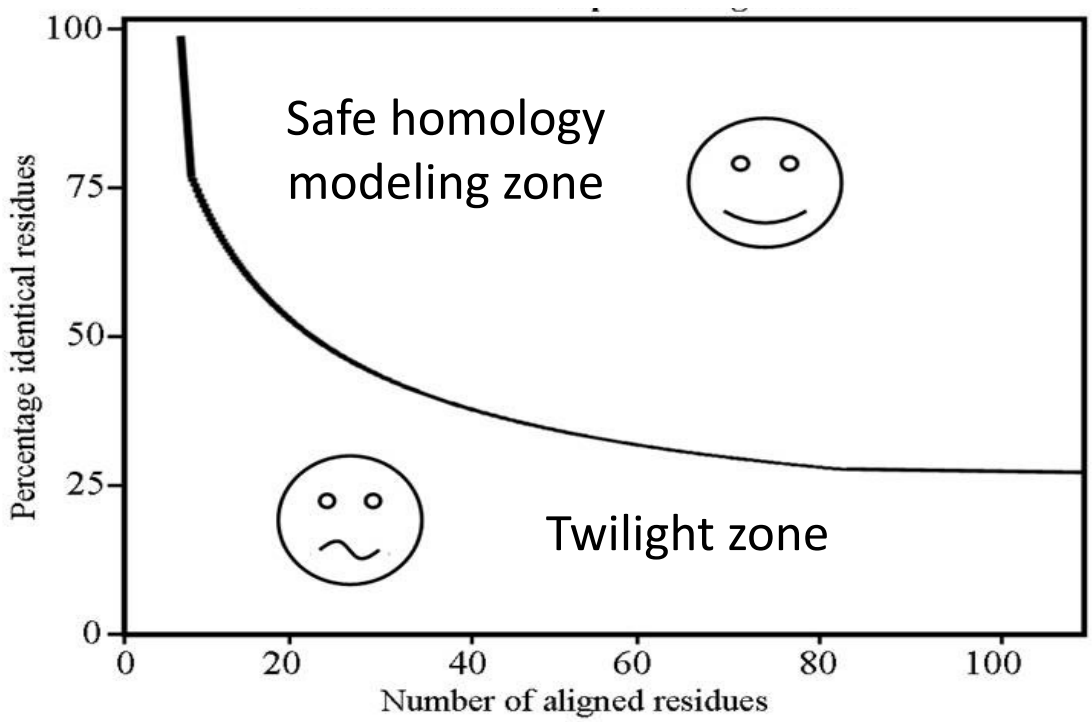


Homology modeling

- basic principle – structure is more conserved than sequence
  - similar sequences adopt practically identical structures
  - distantly related sequences still fold into similar structures
- builds an atomic-resolution model of the target protein based on the experimental 3D structure (template) of a homologous protein
- the most accurate 3D prediction approach
- if no reliable template is available → fold recognition or *ab initio* prediction

Homology modeling

- the quality of the model depends on the sequence identity /similarity between the target and template proteins
- For a standard length protein it should be > 25% / > 40%



Homology modelling – steps

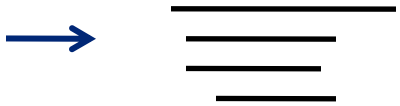
...MSLGAKPFGE...

target sequence

Homology modelling – steps

...MSLGAKPFGE...

target sequence



database search

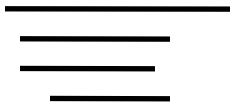
Database search

- standard sequence-similarity searches
  - comparison of the target sequence to all sequences with known 3D structures in the wwPDB database
  - BLAST, FASTA,...
- profile-based searches
  - more sensitive than standard sequence-similarity searches
  - PSI-BLAST, HHMER, HHblits, ...
- fold recognition methods
  - applied if no template can reliably be identified by the sequence or profile based methods (sequence identity < recommended 25 %)
  - FUGUE, GenTHREADER, pro-sp3-TASSER..

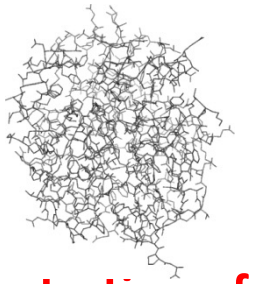
Homology modelling – steps

...MSLGAKPFGE...

target sequence



database search



selection of template

Selection of template

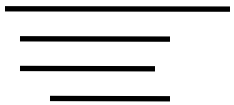
- wrong template = wrong model
- more than one possible template may be identified → a combination of different criteria to select the final template:
  - sequence identity between the template and target protein
  - coverage between the template and query sequences
  - the resolution of the template structure, number of errors
  - a portion of conserved residues in the region of interest (e.g., binding site residues)
  - ...
- multiple templates can be used to create a combined model

Homology modelling – steps

...MSLGAKPFGE...

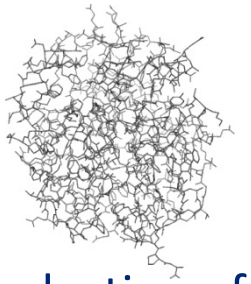
target sequence

→



database search

→



selection of template

→

...MSLGAKPFGE...
...MGV-AKTYGE...

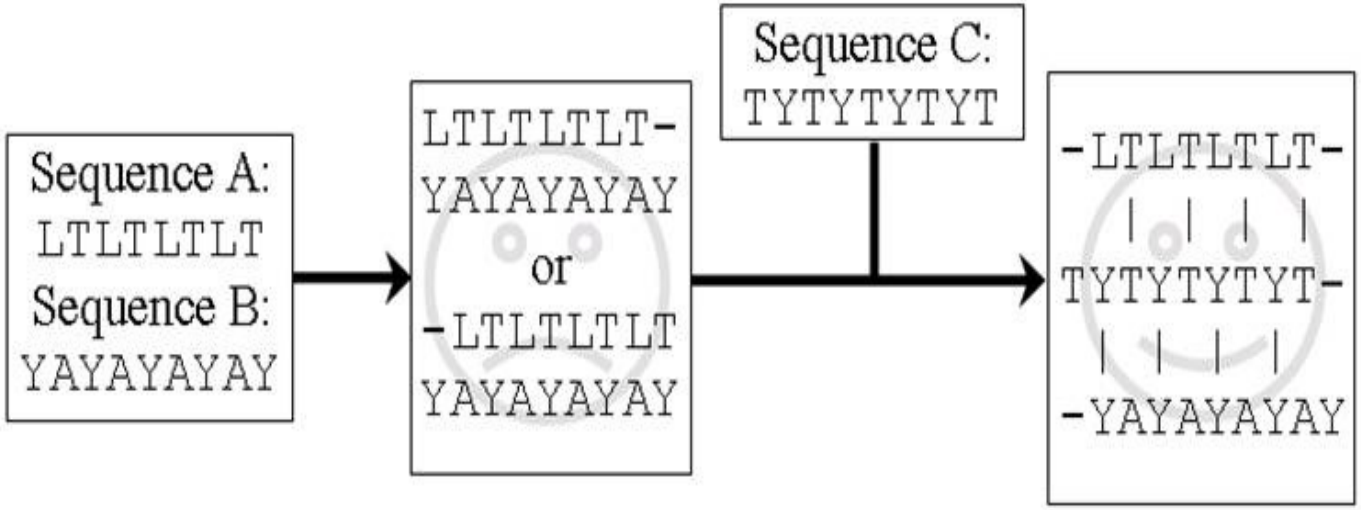
sequence alignment

Sequence alignments

- reliability of alignment decreases with decreasing similarity of the target and template sequences
- quality of alignment is crucial – it determines the quality of the final model
- the pairwise target-template alignment provided by the database search methods is almost guaranteed to contain errors → more sophisticated methods needed
  - multiple sequence alignment
  - Profile-driven alignments
  - correction of alignment based on the template structure

Sequence alignments

- multiple sequence alignment
  - works with more information than pairwise alignment → more reliable
  - MUSCLE, CLUSTAL Omega, T-Coffee

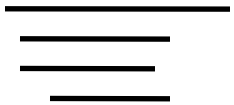


Homology modelling – steps

...MSLGAKPFGE...

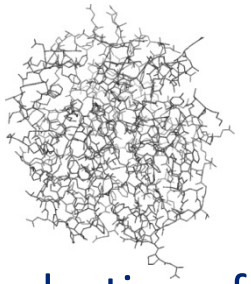
target sequence

→



database search

→



selection of template

→

...MSLGAKPFGE...
...MGV-AKTYGE...

sequence alignment

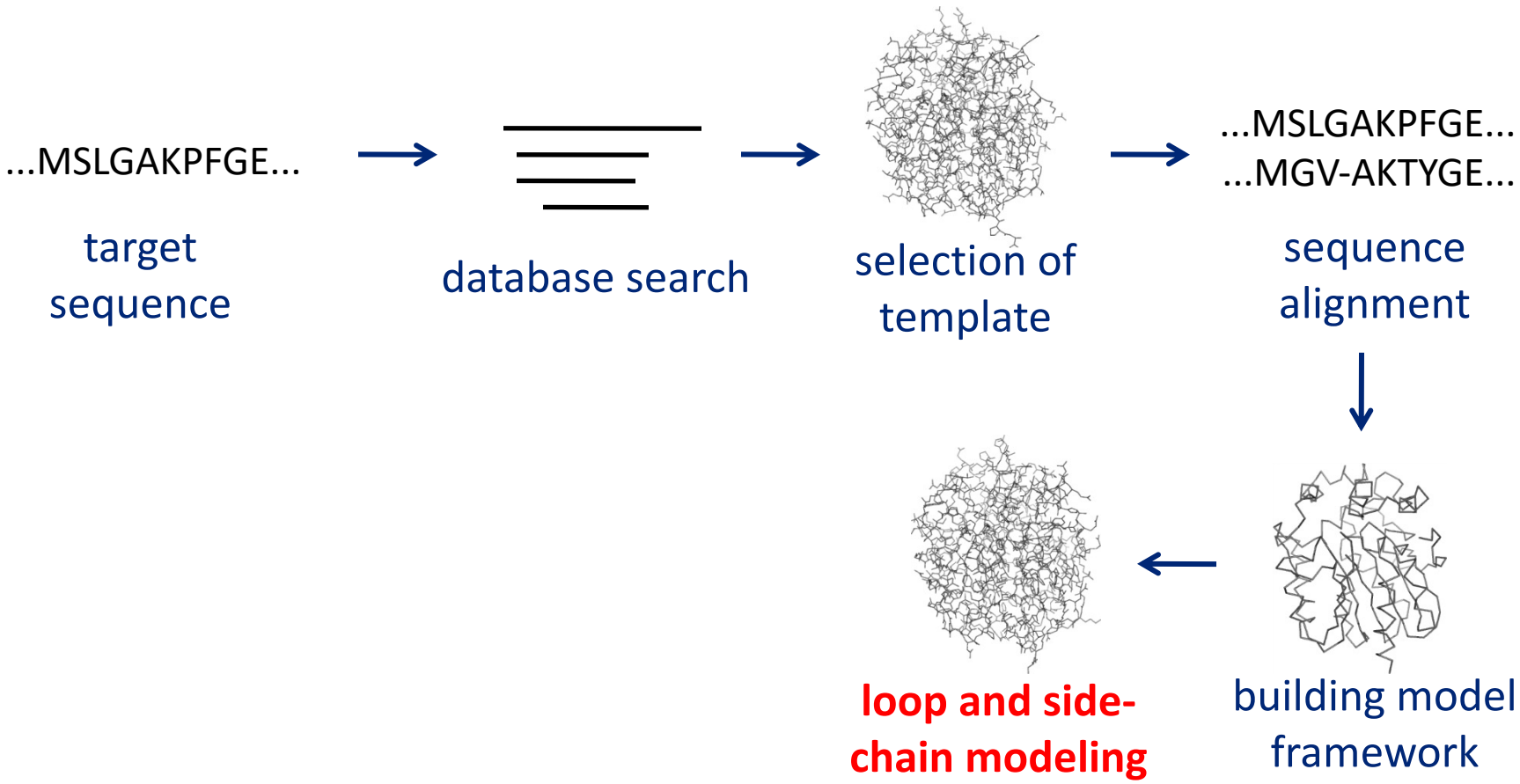
↓

building model framework

Building model framework

- copying the basic shape of the template to the model
  - if the two aligned residues differ, the backbone coordinates for N, Cα, C and O, and often also Cβ can be copied
  - conserved residues can be copied completely to provide an initial guess
  - residues that are not present in the target (because the target can have less residues than the template) are not copied

Homology modelling – steps



Loop modelling

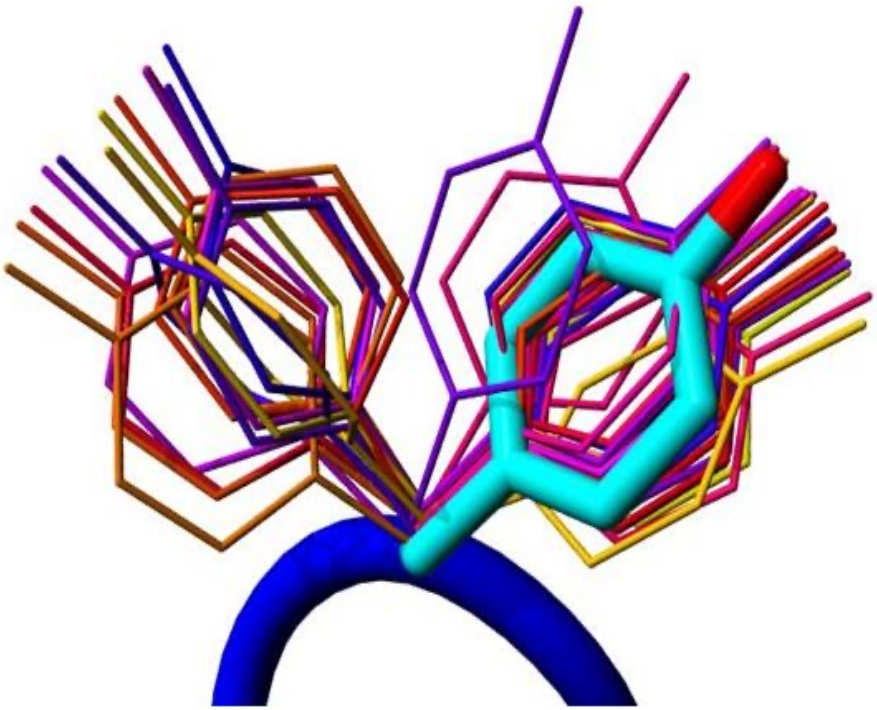
- inserting missing residues into the continuous backbone
- prediction of loop conformation is a difficult task (especially for loops > 5-8 residues long
  - knowledge based prediction – use of libraries of possible loop conformations known from experimentally determined structures with the same local sequence
  - *ab initio* prediction – use of energy functions to find the most optimal conformation, followed by minimization of the structure
  - hybrid approach – the loop is divided into small fragments that are all separately compared to known structures

Side-chain modelling

- adding side-chains of amino acids to the model backbone
  - rotamer libraries – common side-chain conformations (rotamers) extracted from high-resolution X-ray structures → possible rotamers explored and scored based on energy function
  - backbone-dependent rotamer libraries – the optimal conformation of the side chain depends on the local backbone conformation (5 - 9 neighboring residues) → explored only possible rotamers corresponding to the best backbone matches – greatly reduces conformational search space

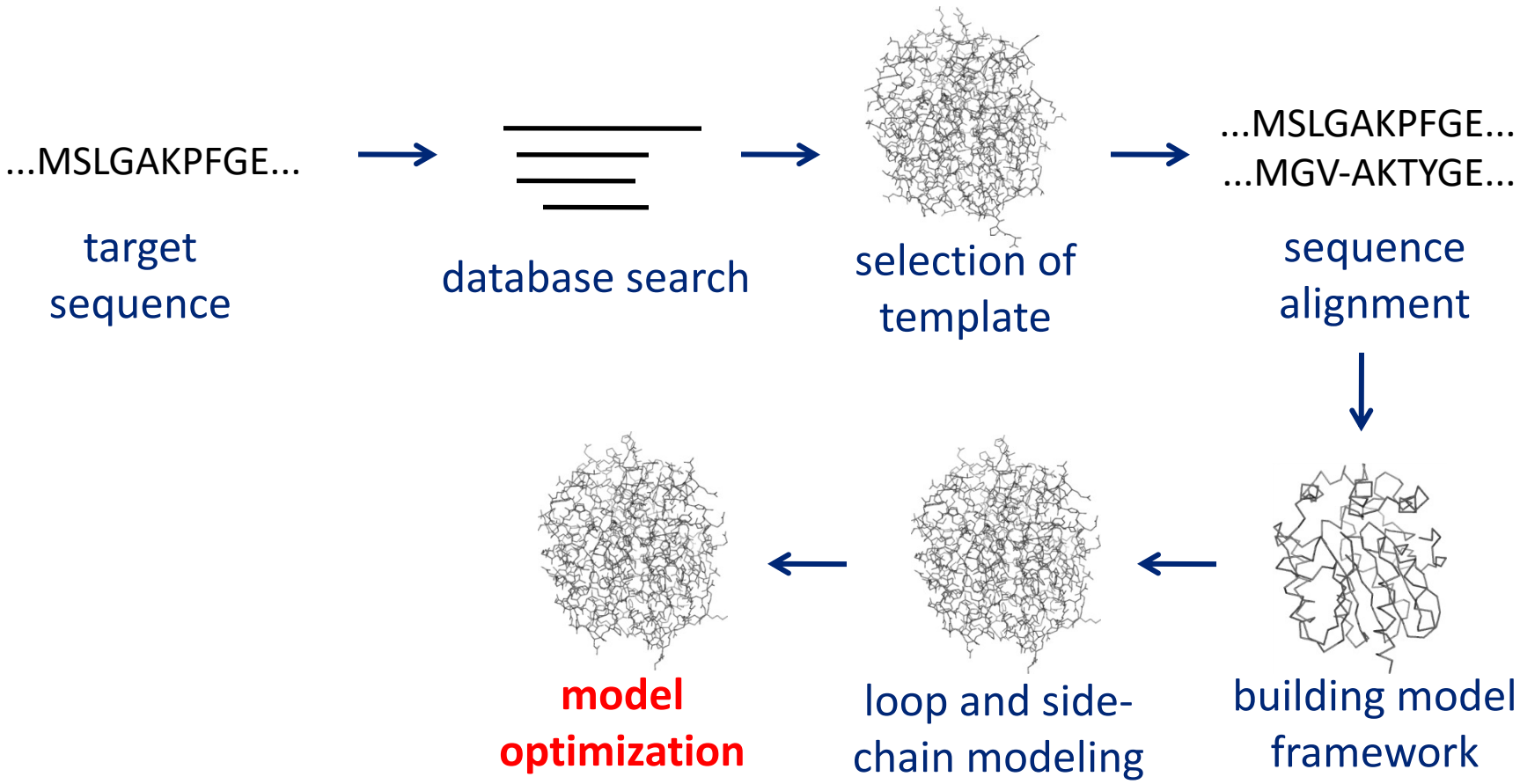
Side-chain modelling

- backbone-dependent rotamer library



According to the backbone-dependent rotamer library, the backbone favors two different conformations for Tyrosine which appear about equally often in the database .

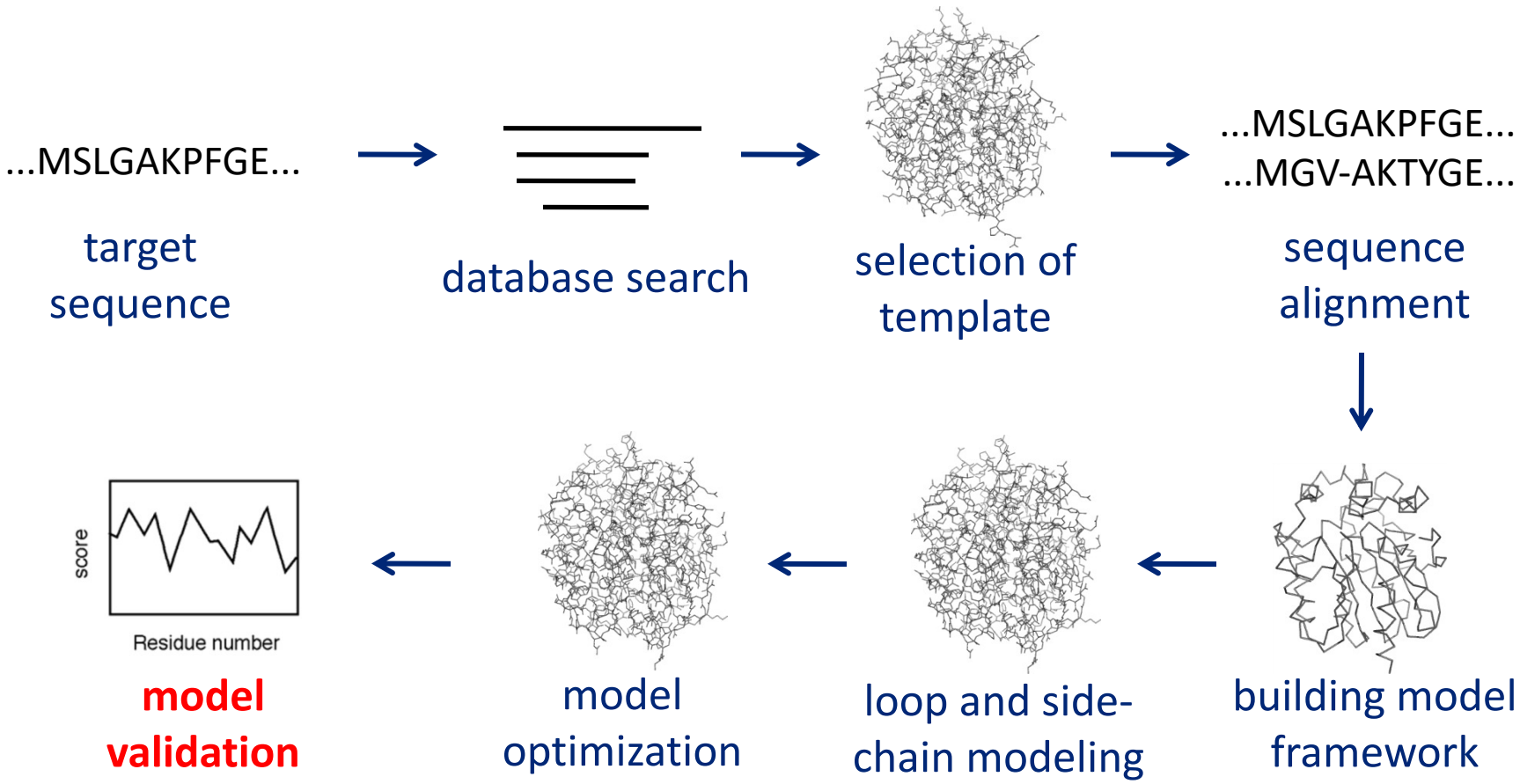
Homology modelling – steps



Model optimization

- energy minimization – may introduce many errors moving the model away from its correct structure → must be used carefully
- molecular dynamics simulation – follows the motions of the protein and mimics the folding process

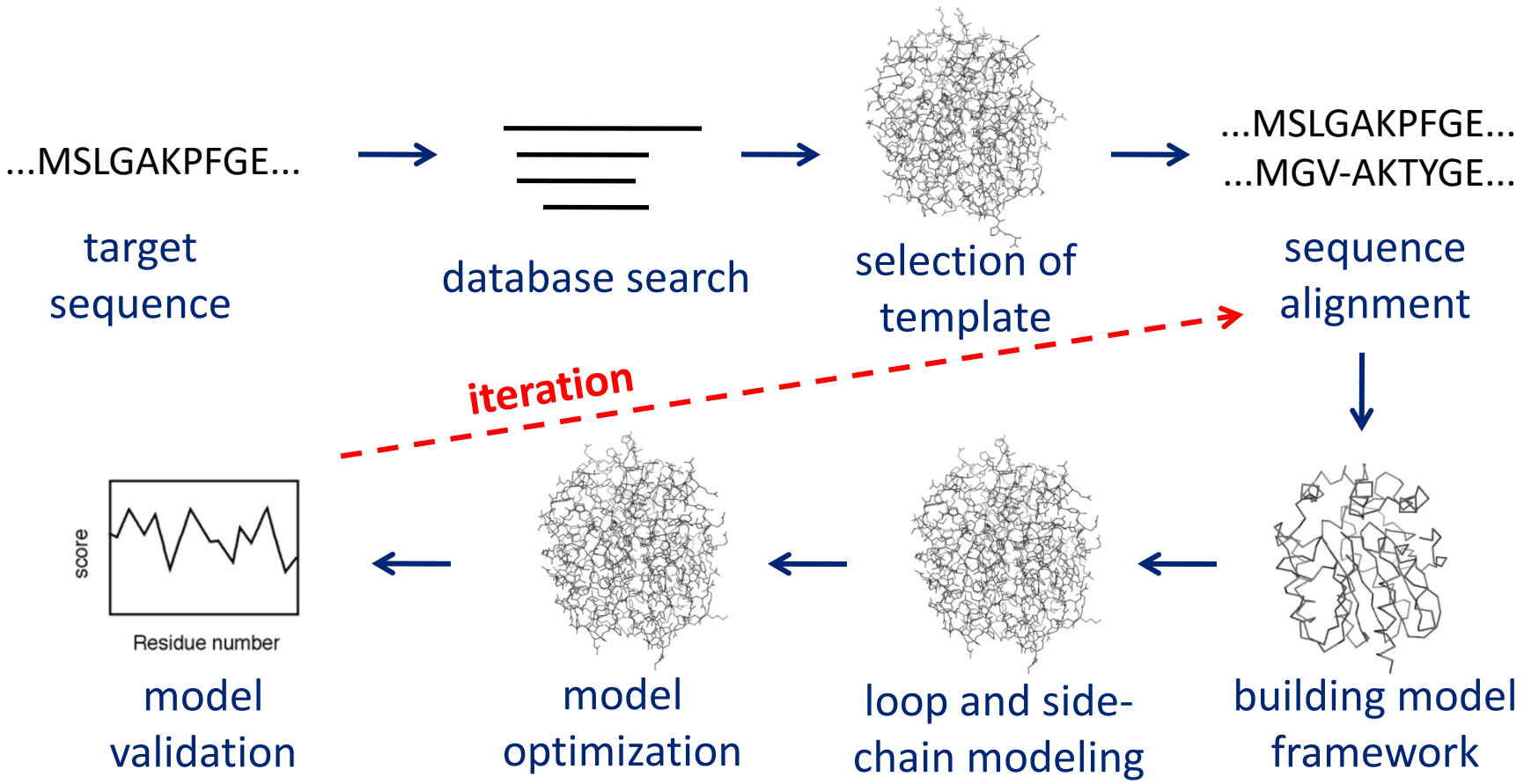
Homology modelling – steps



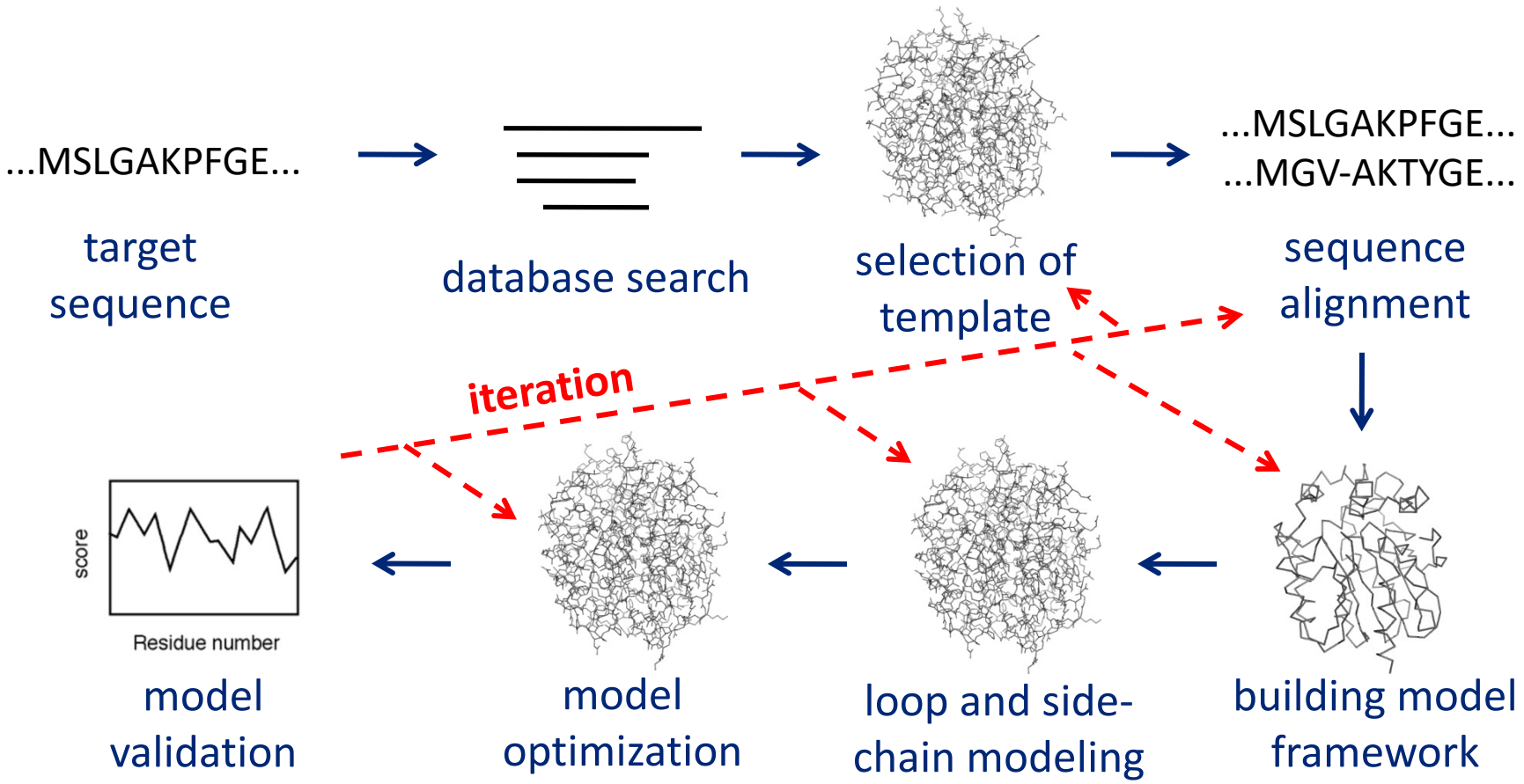
Model validation

- finished model contain errors (like any other structure) – the number of errors (for a given method) mainly depends on:
  - the percentage of sequence identity between template and target sequence, e.g., 90 %: the accuracy of the model comparable to X-ray structures; 50 %-90 %: larger local errors; identity < 25 %: often very large errors
  - the number of errors in the template structure
- problems that occur far from the site of interest may be ignored, others should be tackled

Homology modelling – steps



Homology modelling – steps



Iteration

- portions of the homology modeling process can be iterated to correct identified errors
  - small errors introduced during the optimization → running a shorter molecular dynamics simulation
  - error in a loop → choosing another loop conformation in the loop modeling step
  - large mistakes in the backbone conformation → repeating the whole process with another alignment or even different template
  - …

Homology modeling programs

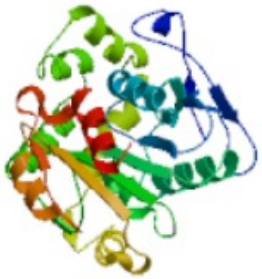
- MODELLER
  - http://salilab.org/modeller/
  - models built by satisfying the spatial restraints of the C α - C α bond lengths and angles, the dihedral angles of the side-chains, and van der Waals interactions
  - restraints calculated from the template structures
  - available as a web server at different sites, e.g., part of: ModWeb workflow https://modbase.compbio.ucsf.edu/modweb/, GeneSilico server https://genesilico.pl/toolkit/unimod?method=Modeller or Bioinformatics toolkit http://toolkit.lmb.uni-muenchen.de/modeller

Homology modeling programs

- SWISS-MODEL
  - http://swissmodel.expasy.org/
  - fully automated protein structure homology modeling server

Print/Save this pa

Model Summary



Model information:
Modelled residue range: 1 to 297
Based on template: [2xt0A] (1.90 Å)
Remark: No search for template was performed.
Only user specified template was used for modelling.
Sequence Identity [%]: 40.33
Evalue: 0.00e-1

Quality information: [details]
QMEAN Z-Score: -2.61

Quaternary structure information: [details]
Template (2xt0): MONOMER
Model built: SINGLE CHAIN

Ligand information: [details]
Ligands in the template: SO4: 2.
Ligands in the model: none.

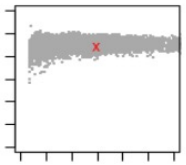
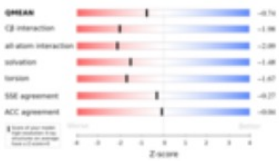
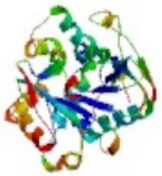
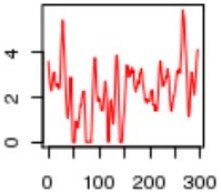
logs: [Templates] [Alignment] [Modelling]
display model: as [pdb] - as [DeepView project] - in [AstexViewer]
download model: as [pdb] - as [Deepview project] - as [text]

Model validation

- mostly the same principles as used for the validation of experimental structures
- always check both model and template
  - The model cannot improve the template if this is “bad” in regions
- checks of normality
  - inside/outside distributions of polar and apolar residues
  - bad contacts
  - evaluation of atom/residue environment
- energy-based checks
  - side-chain clashes
  - bond lengths and angles

Model validation programs

- QMEAN
  - https://swissmodel.expasy.org/qmean/
  - composite scoring function for the quality estimation of protein structure models; evaluates torsion angles, solvation and non-bonded interactions and the agreement between predicted and calculated secondary structure and solvent accessibility

Global scores				Local scores	
Model name	QMEAN score	Estimated absolute quality NEW	Z-scores of QMEAN terms NEW	Residue error <1Å >3.5Å	Residue error plot
modbase-model_6d51f947356cc91f0e1be73c6d7e11d2.pdb	0.705	 Z-score=-0.74 [plot 1] [plot 2]	 [png]	 [jpg] [pdb] Jmol	 [png] [table]

Model validation programs

- Verify3D
- ANOLEA
- PROCHECK
- WHATCHECK
- PROSA II
- …

Fold recognition (Threading)

- predicts the fold of a protein by fitting its sequence into a structural database and selecting the best fitting fold
- provides a rough approximation of the overall topology of the native structure → does not generate fully refined atomic models for the query sequence
- can be used when no suitable template structures available for homology modeling
- fails if the correct protein fold does not exist in the database
- high rates of false positives

Fold recognition (Threading)

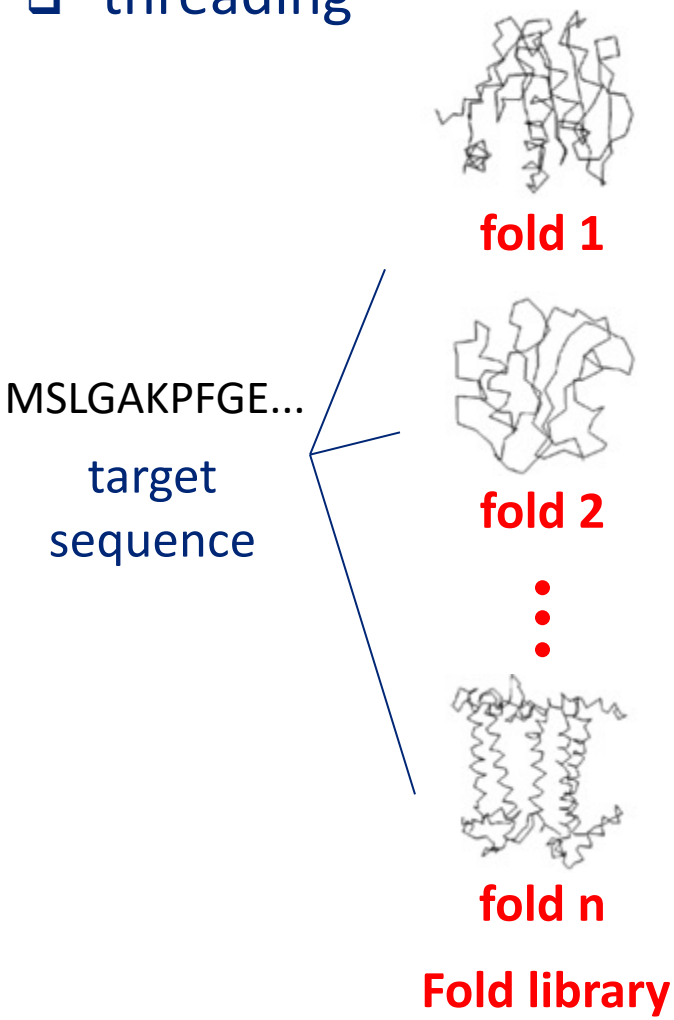
- threading

MSLGAKPFGE...

target sequence

Fold recognition (Threading)

- threading

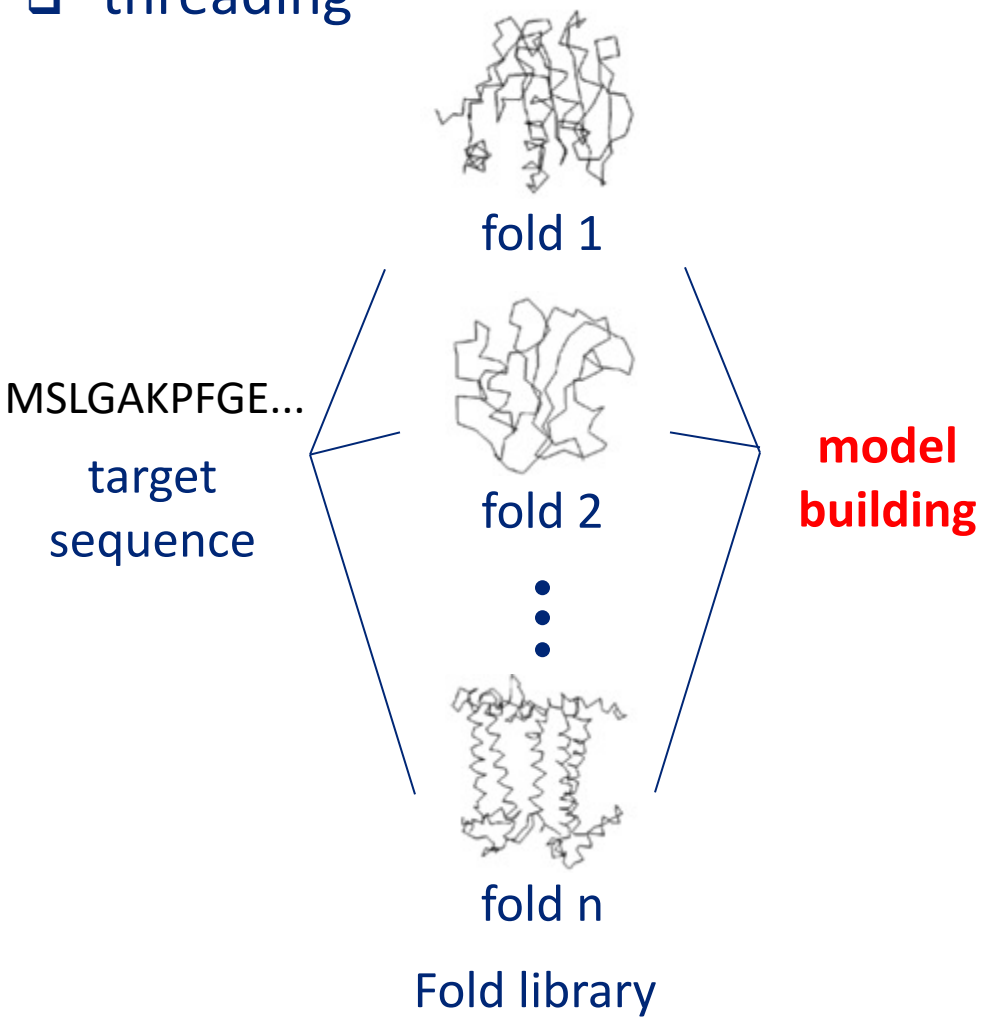


Fold recognition (Threading)

- pairwise energy-based methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using energy-based criteria
  1. alignment of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)

Fold recognition (Threading)

- threading

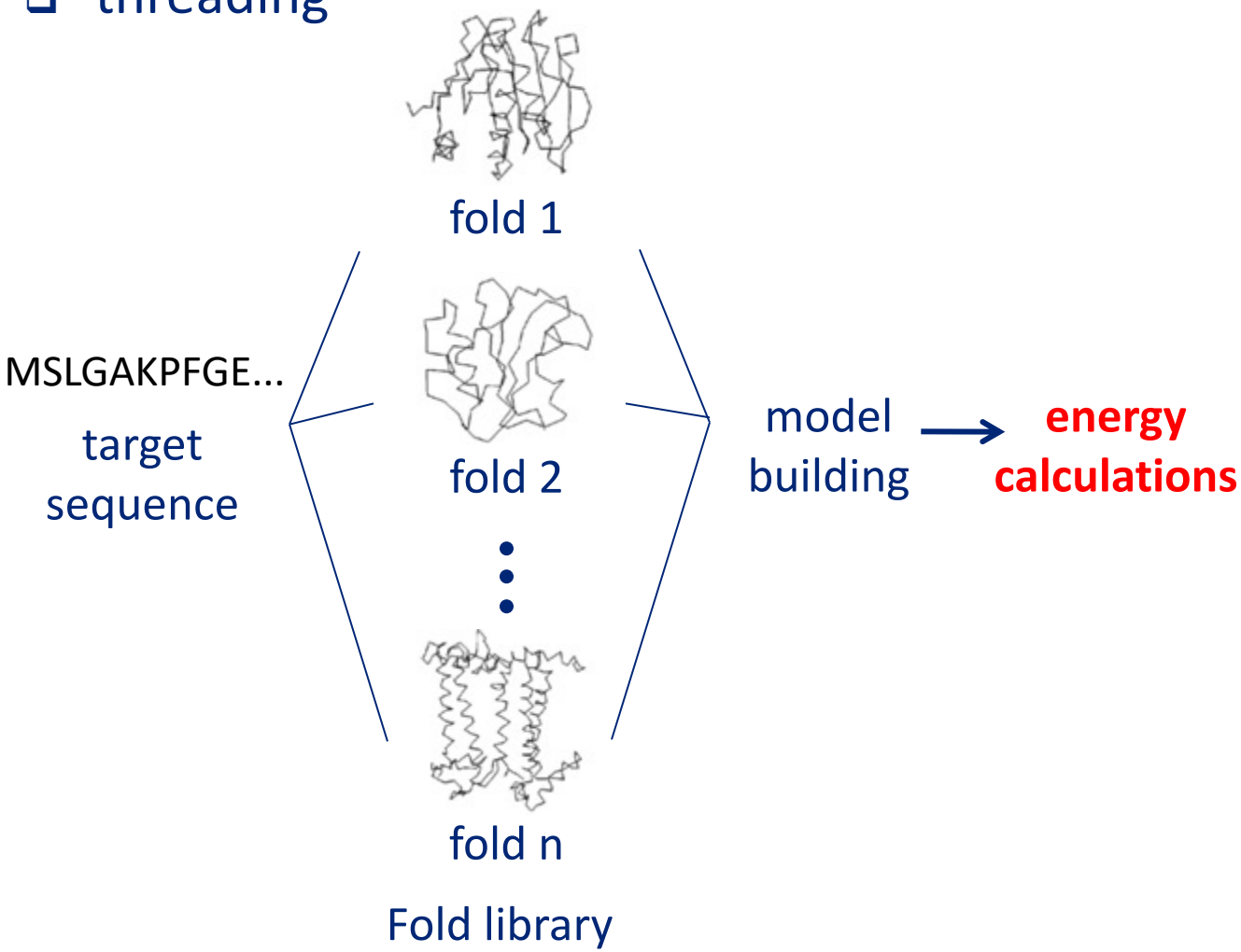


Fold recognition (Threading)

- pairwise energy-based methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using energy-based criteria
  1. alignment of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)
  2. building a crude model for the target sequence (replacing aligned residues in the template structure with the corresponding residues in the query)

Fold recognition (Threading)

- threading

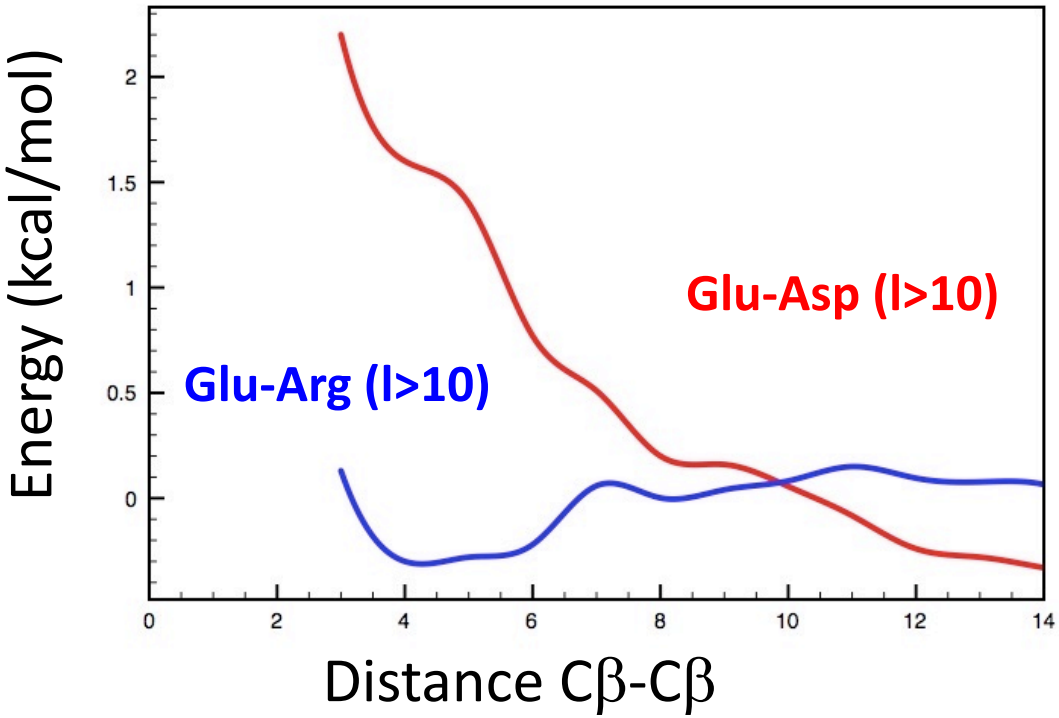


Fold recognition (Threading)

- pairwise energy-based methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using energy-based criteria
  1. alignment of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)
  2. building a crude model for the target sequence (replacing aligned residues in the template structure with the corresponding residues in the query)
  3. calculating energy of the raw model

Fold recognition (Threading)

- pairwise energy-based methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using energy-based criteria



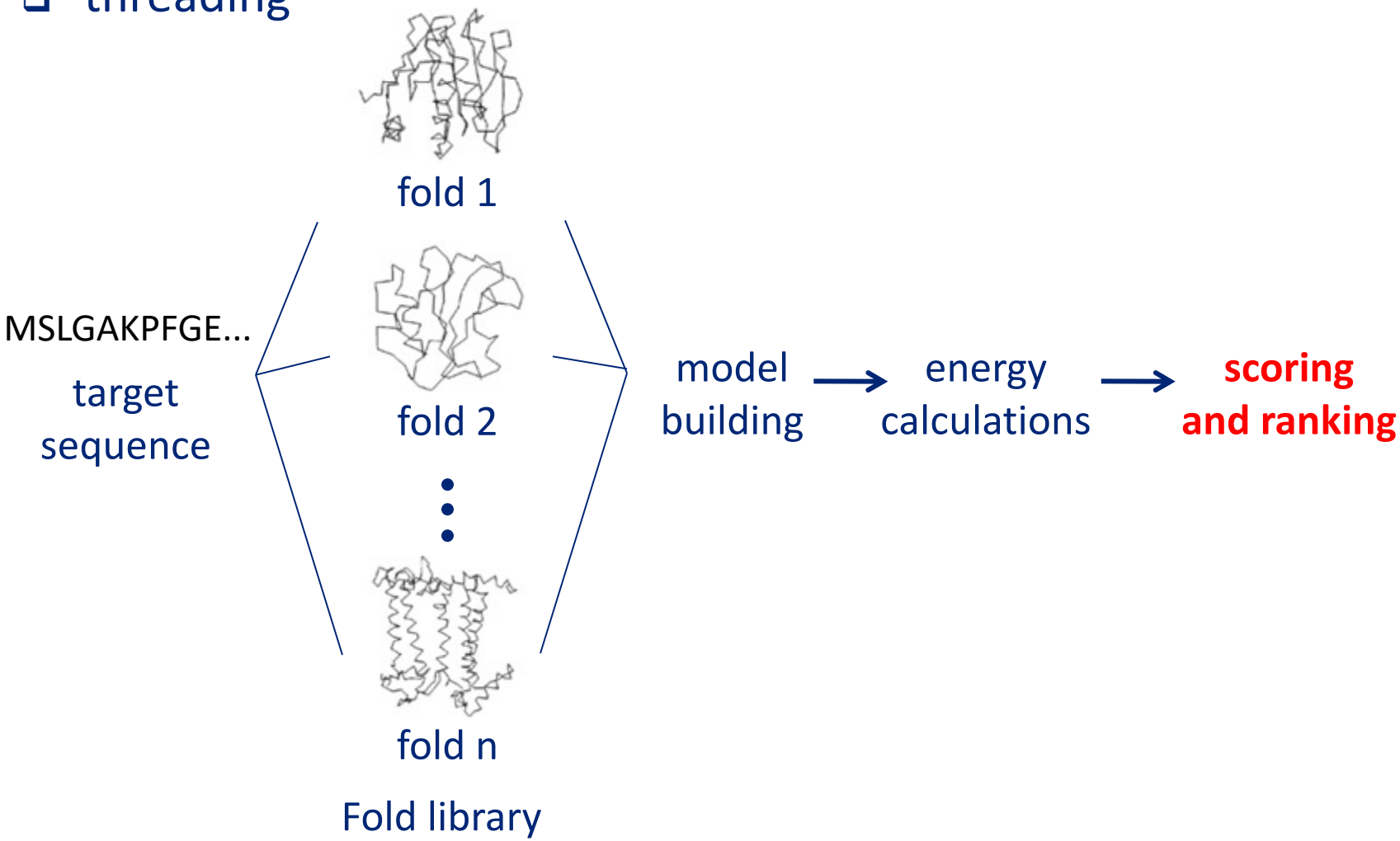
l is distance in sequence (density normalization required)

can be calculated from collections of known structures

5-3D Modelling -> Fold recognition (Threading)

Fold recognition (Threading)

- threading



Fold recognition (Threading)

- pairwise energy-based methods (threading) – protein sequence is searched for in a structural database to find the best matching structural fold using energy-based criteria
  1. alignment of the query sequence with each structural fold in the fold library (essentially performed at the sequence profile level)
  2. building a crude model for the target sequence (replacing aligned residues in the template structure with the corresponding residues in the query)
  3. calculating energy of the raw model
  4. ranking of the models based on the energetics – the lowest energy fold represents the structurally most compatible fold

Fold recognition (Profiles)

- profile methods

Fold recognition programs

- PHYRE
  - http://www.sbg.bio.ic.ac.uk/phyre2/
  - profile-based method
  - the highest scoring alignments are used to construct full 3D models of the query – missing or inserted regions are repaired using a loop library and reconstruction procedure, side-chains are placed using a fast graph-based algorithm

Fold recognition programs

- PHYRE

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily
	d3aedha_ (length:145) **100% i.d.**		9.3e-20	100 %	0.90 Biotext	Globin-like	Globin-like
	c2bk9A_ (length:153) **23% i.d.**		7.7e-17	100 %	0.89 Biotext	**PDB header:**oxygen transport	**Chain:** A: **PDB Molecule:**cg9734-pa;

Fold recognition programs

- RaptorX
  - http://raptorx.uchicago.edu/
  - provides single-template threading, alignment quality prediction, and multiple-template threading

- GenTHREADER
  - http://bioinf.cs.ucl.ac.uk/psipred/
  - uses a hybrid of the profile and pairwise energy methods
  - multiple sequence alignment and secondary structure predictions derived for the query are used as input for threading
  - threading results are evaluated using neural networks

Ab initio prediction

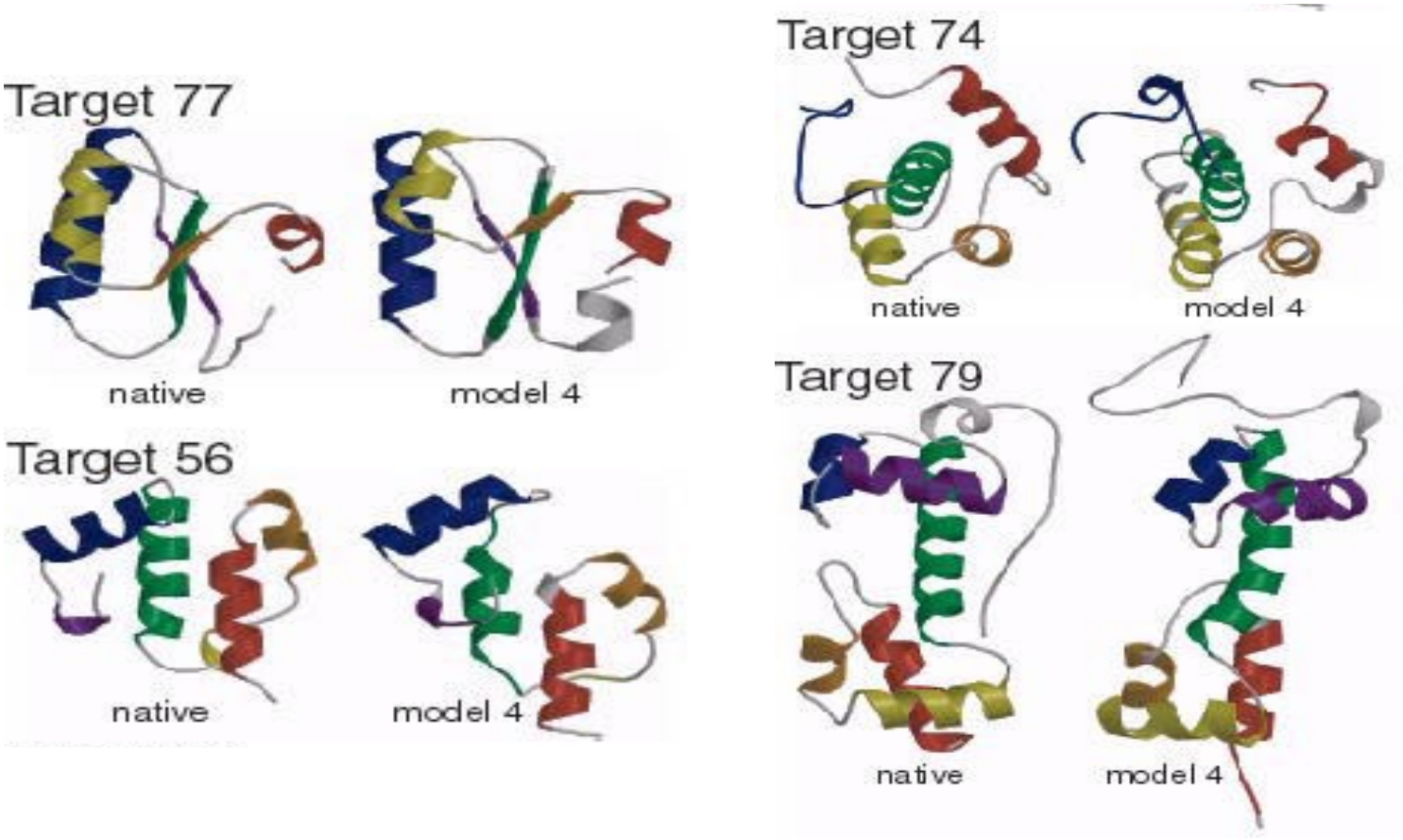
- attempts to generate a structure by using physicochemical principles only
- used when neither homology modeling nor fold recognition can be applied
- search for the structure in the global free-energy minimum
- so far still limited success in getting correct structures

Ab initio prediction programs

- Rosetta
  - http://www.rosettacommons.org/
  - software suite for predicting and designing protein structures, protein folding mechanisms, and protein-protein interactions

Ab initio prediction programs

- Rosetta

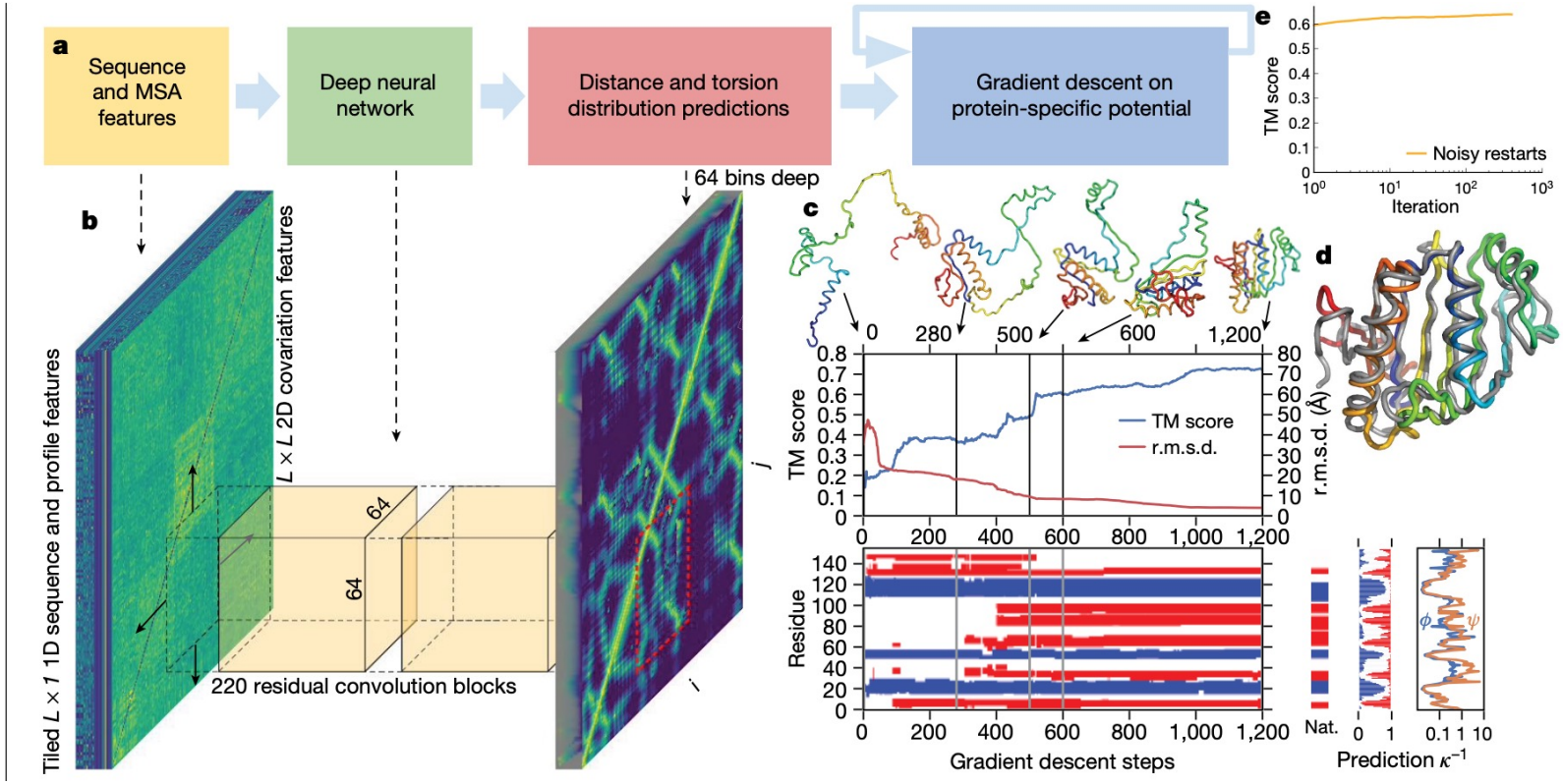


“Hybrid” 3D structure prediction programs

- I-TASSER
  - http://zhanglab.ccmb.med.umich.edu/I-TASSER/
  - combines homology modeling, threading and *ab initio* predictions
  - No. 1 server for protein structure prediction in previous CASP experiments

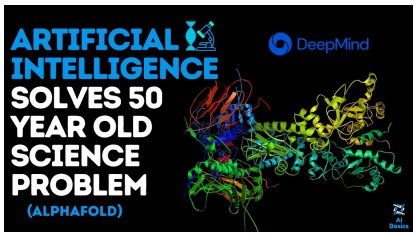
- Robetta
  - http://robetta.bakerlab.org/
  - combines homology modeling and *ab initio* predictions
  - implements ROSETTA software

AlphaFold1: ML-powered threading

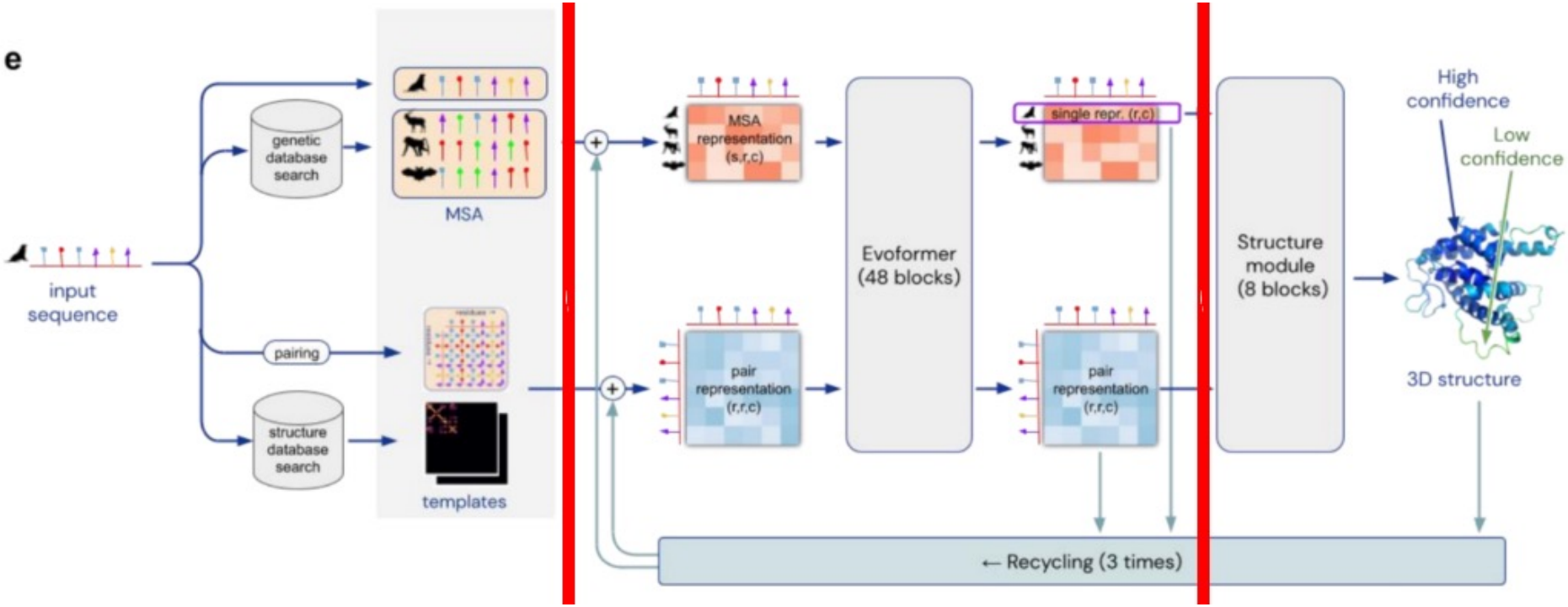


- Combines threading with ML
- No. 1 server for protein structure prediction in CASP13 (2018) experiment

Google DeepMind

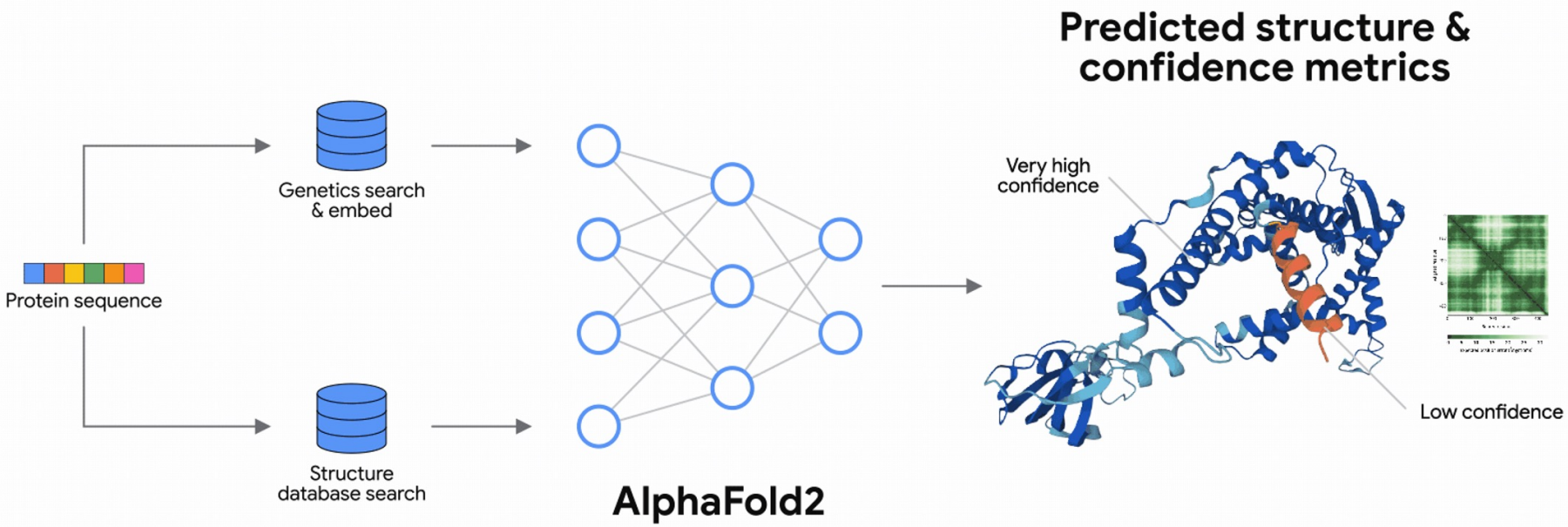


AlphaFold 2: ML revolution



- Two independent traks for sequence and structure, each ML-powered
- Attention layer at the structure module doing the “trick)
- AF multimer for protein interactions.
- No. 1 server for protein structure prediction in CASP14 (2020) experiment

AlphaFold 3: turning up another notch



- Simplified representations (sequence-based)
- Structural info only optional.
- Simplified network
- Models all sorts of biomolecules.

ML-powered (reverse) folding

Google DeepMind

AlphaFold

Meta

ESMFold

RosettaFold difussion

ProteinMPNN

- State of the art homology modelling approaches.
- Simple problems can be solved by simpler approaches.
- ESMFold: quality length-dependent
- Hallucinates new 3Ds from scratch (ML learned how structure looks like)
- Solves reverse problem: from 3D predict optimal sequence.

Assessment of prediction methods

- CASP (**C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction)
  - http://predictioncenter.org/
  - biannual international contest providing objective evaluation of the performance of individual prediction methods
  - evaluation based on a large number of blind predictions - contestants are given protein sequences whose structures have been solved, but not yet published - results of the predictions are compared with the newly determined structure
  - competition in several categories

Assessment of prediction methods

- CAMEO (**C**ontinuous **A**utomated **M**odel **E**valuati**O**n)
  - https://www.cameo3d.org/
  - weekly assessment of new structures in the PDB
  - registered prediction servers are sent weekly requests on not-so-easy new structures in the weekly PDB pre-release.
  - Multiple scores considered, normalized average (IDDT) reported
  - Categories:
    - 3D: Prediction of the 3D coordinates of a protein from sequence
    - QE: Model quality Estimation: Assessment of quality measures reported by participant servers

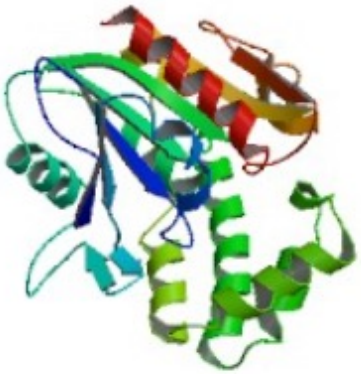
Databases of protein models

- ModBase
  - http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
  - database of annotated protein models generated by the automated pipeline including the MODELLER program
  - contains ~38 millions models for ~6.5 millions unique sequences

Quality criteria indicate whether the model is considered reliable (green) or unreliable (red).

Target Region	34-301
Protein Length	301
Template PDB Code	1r3dA
Template Region	4-262
Sequence Identity	12.00%
E-Value	2e-25
GA341	0.18
Dataset	nysgxrc_1r3d_3-06
ModPipe Version	ModPipe1.0
Model Date	2006-04-15

for all Models of this Sequence :

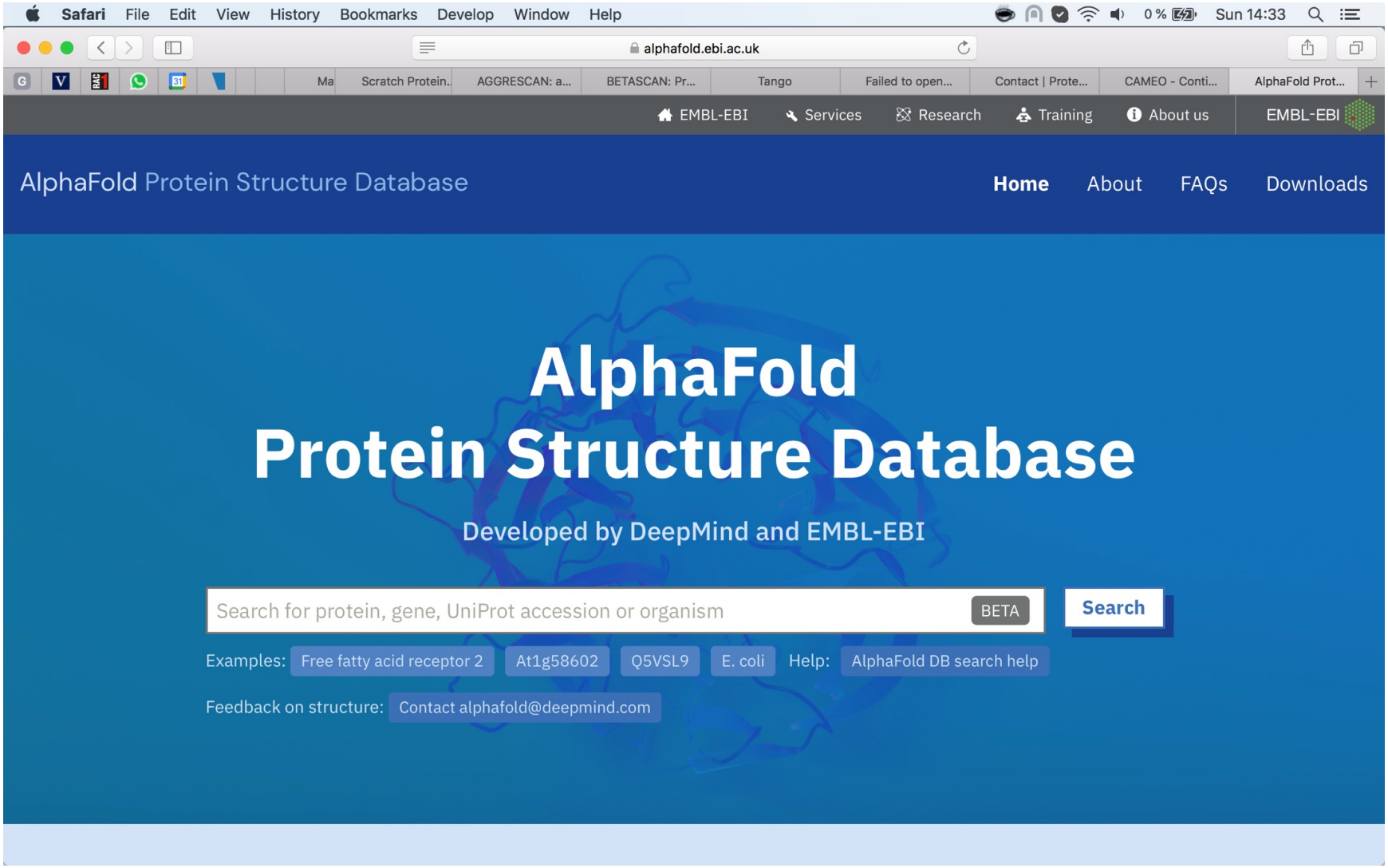


Databases of protein models

- SWISS-MODEL repository
  - http://swissmodel.expasy.org/repository/
  - database of annotated protein models generated by the automated homology-modeling pipeline SWISS-MODEL.
  - contains 2.2 millions models for UniProt sequences

- PMDB (**P**rotein **M**odel **D**ata**B**ase)
  - http://srv00.recas.ba.infn.it/PMDB/
  - contains manually built 3D protein models
  - users can download as well as submit models along with related supporting evidence

Databases of protein models



References

- Gu, J. & Bourne, P. E. (2009). **Structural Bioinformatics, 2nd Edition**, Wiley-Blackwell, Hoboken, p. 1067.
- Xiong, J. (2006). **Essential Bioinformatics**. Cambridge University Press, New York, p. 352.
- Schwede, T. & Peitsch, M. C. (2008). **Computational Structural Biology: Methods and Applications**, World Scientific Publishing Company, Singapore, p. 700.
- Shapiro, B. A. *et al.* (2007). Bridging the gap in RNA structure prediction. *Current opinion in structural biology* **17**: 157-165.