Structural Bioinformatics:

MSc Bioinformatics Cranfield

Dr Elliot Drew

e.drew@qmul.ac.uk

What is structural bioinformatics

It is the branch of bioinformatics related to the

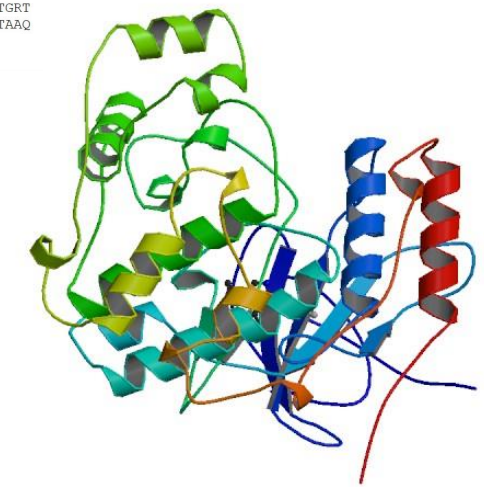
analysis and **prediction**

of the 3D structures of biological macromolecules

What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

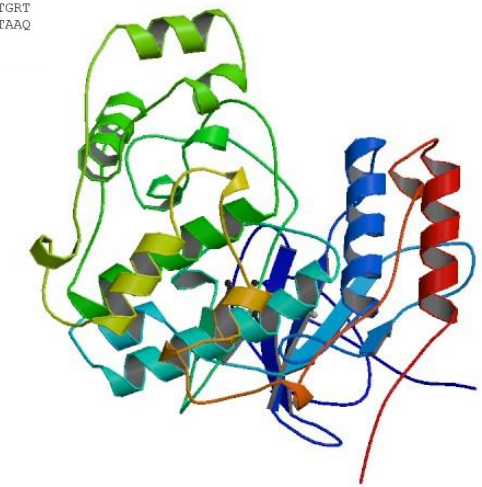
Predict Structure

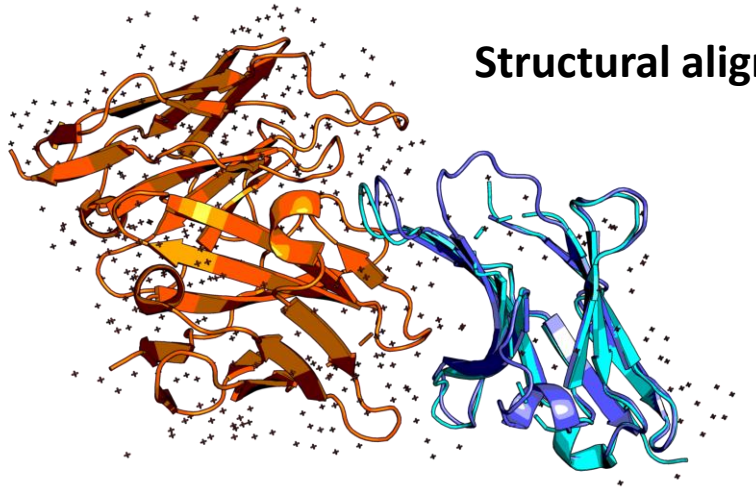


What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

Predict Structure



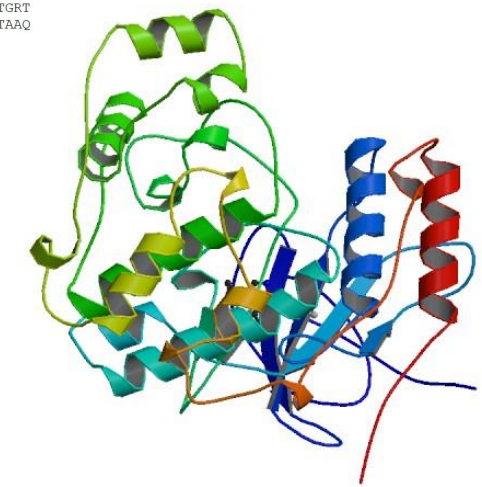


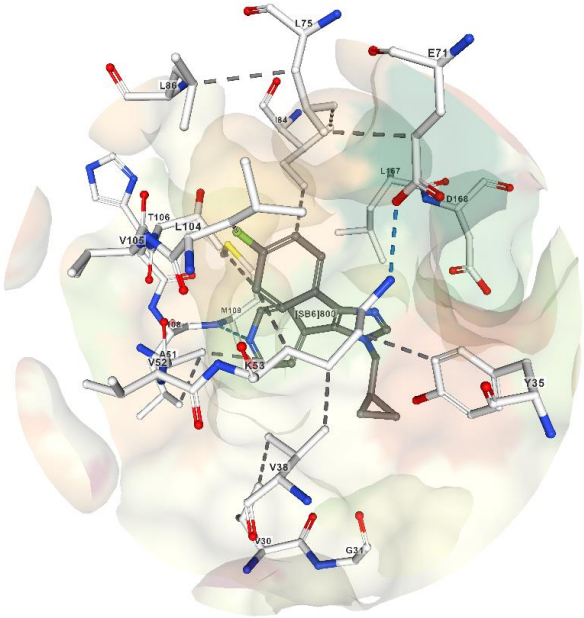
Structural alignment

What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

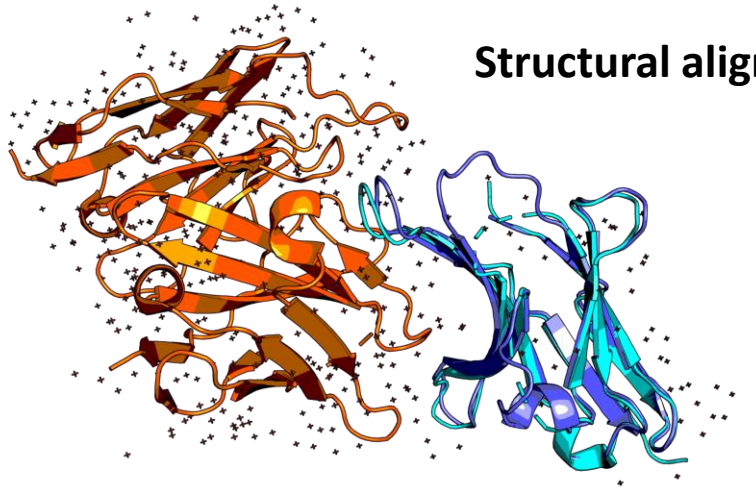
Predict Structure





Predict Interactions

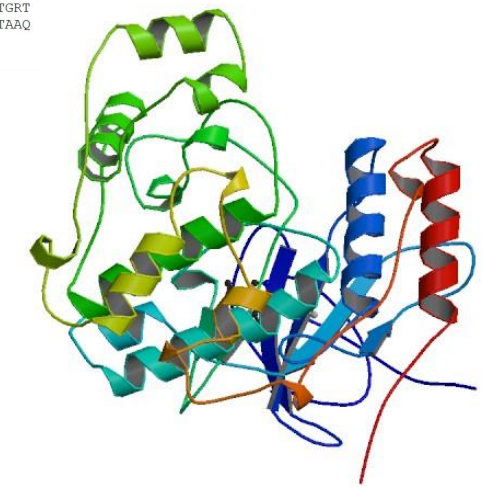
Structural alignment

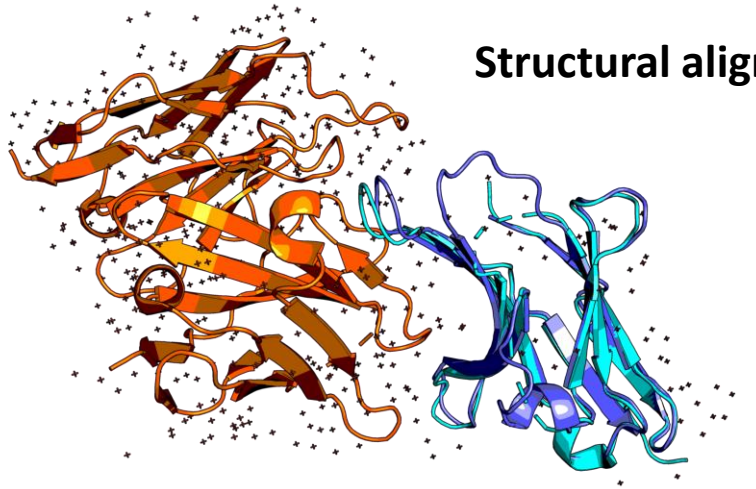


What is structural bioinformatics

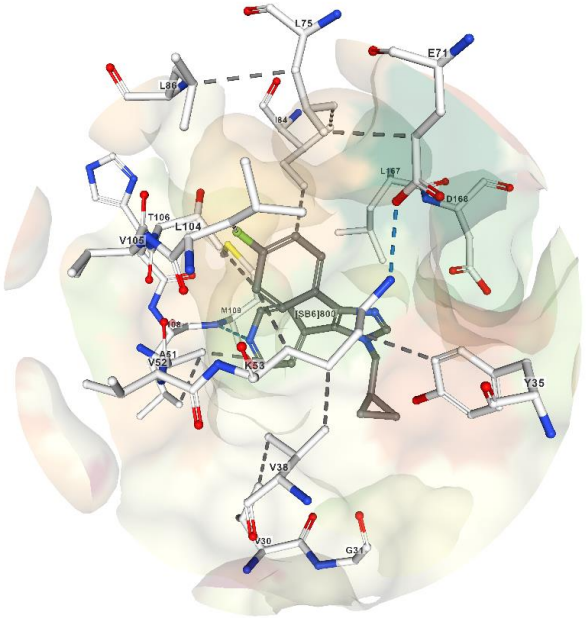
```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

Predict Structure

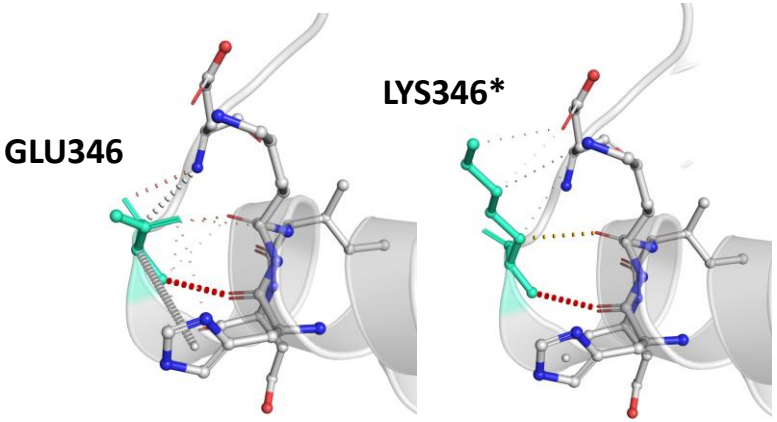




Structural alignment



Predict Interactions

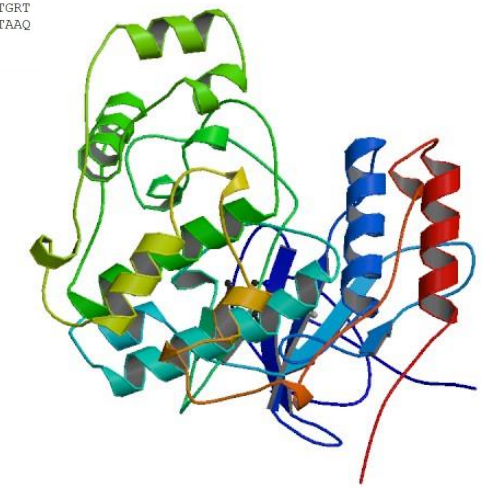


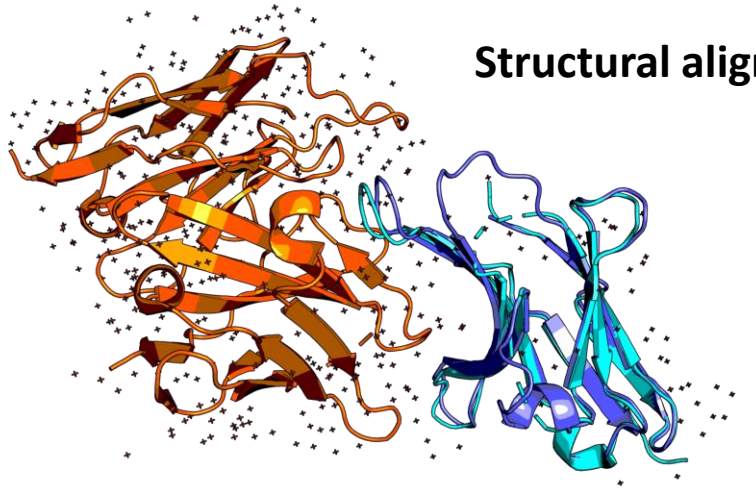
Computational Mutagenesis

What is structural bioinformatics

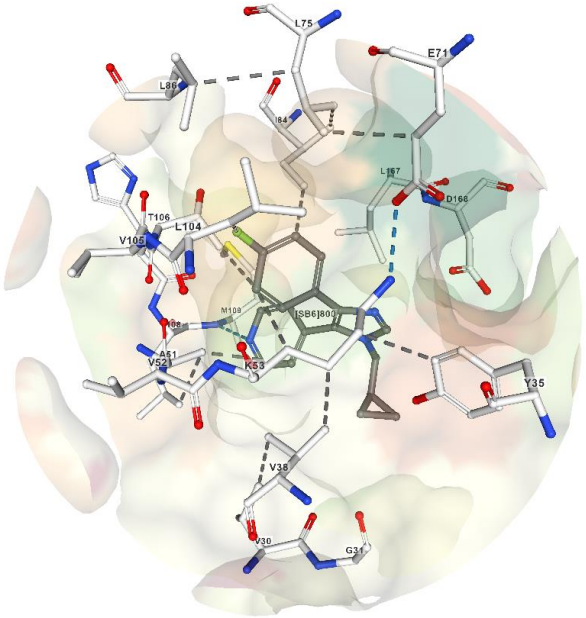
```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

Predict Structure

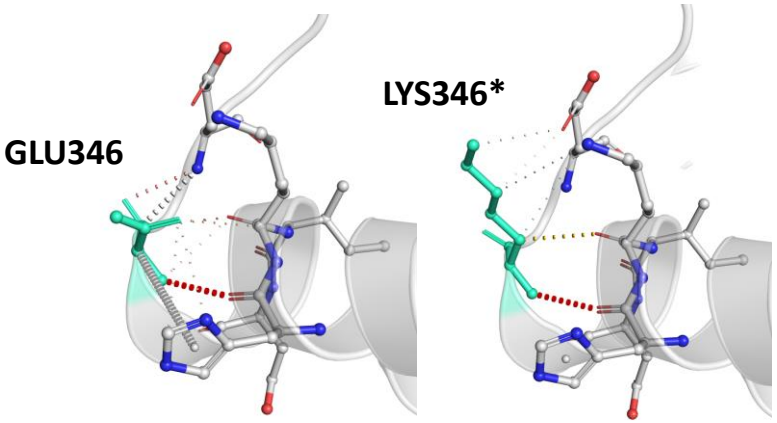




Structural alignment

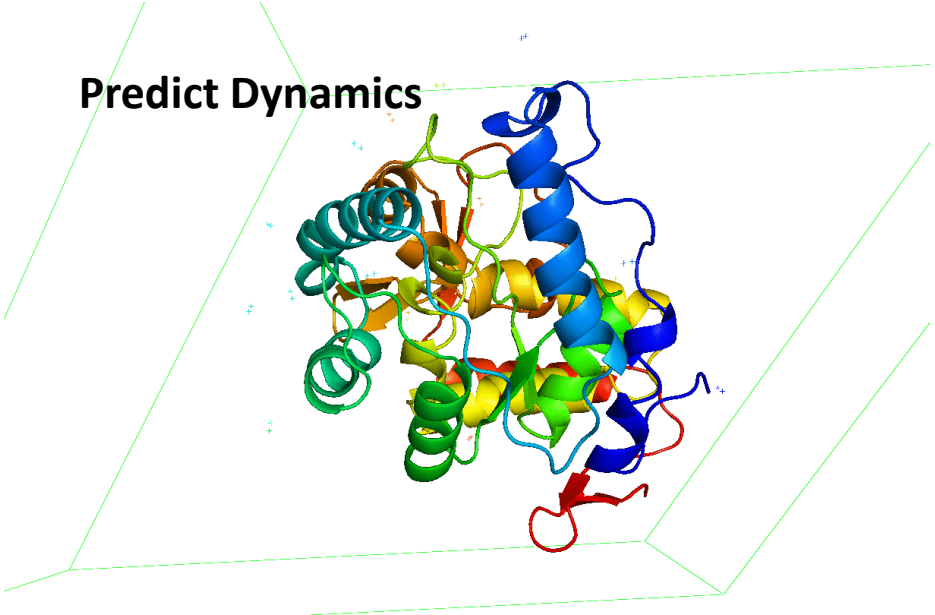


Predict Interactions



Computational Mutagenesis

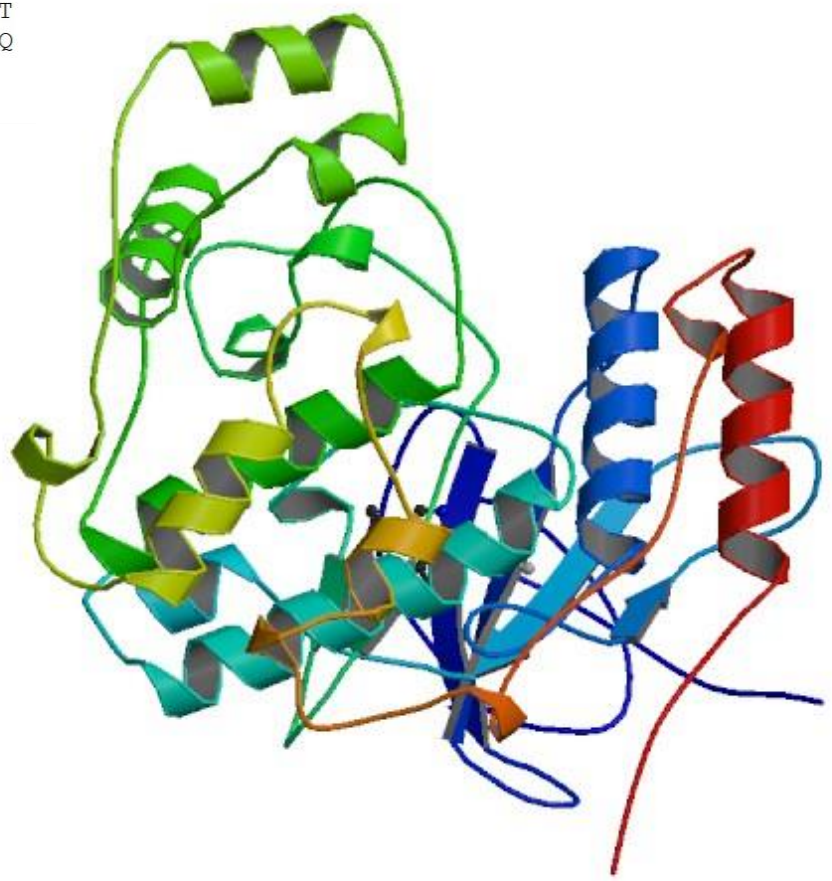
Predict Dynamics



What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

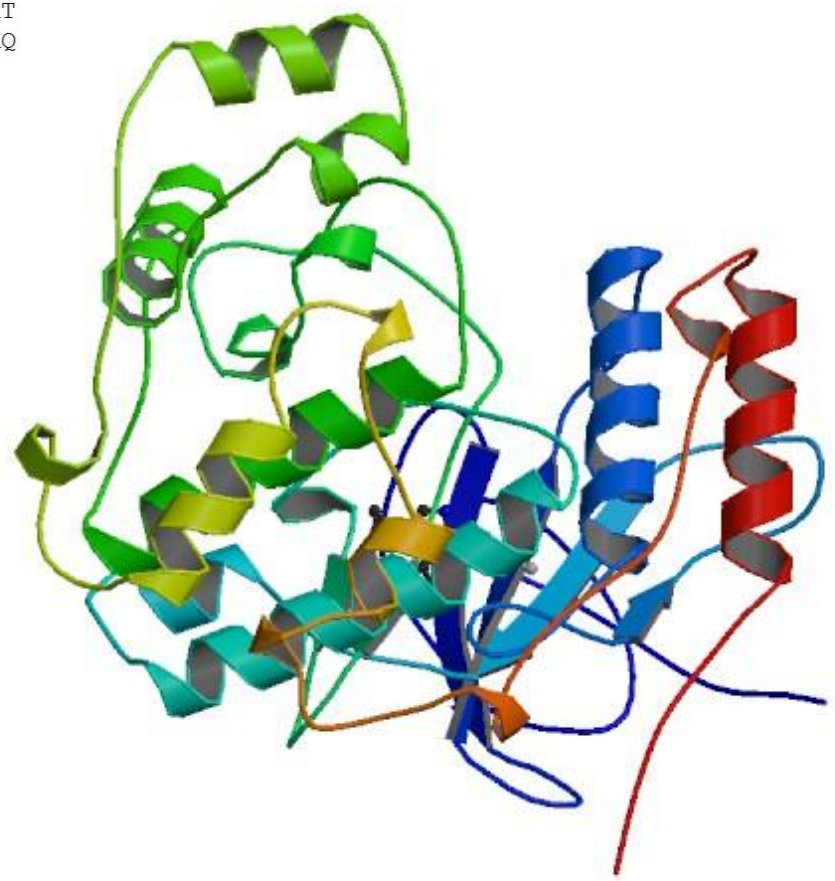
Predict Structure



What is structural bioinformatics

```
>1BL6:A|PDBID|CHAIN|SEQUENCE
GSSHHHHHHSSGLVPRGSHMSQERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQS
IIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMGADLNNIVKCQKLTDDHVQFLIYQILRGLKYI
HSADIIHRDLKPSNLAVNEDCELKILDFGLARHTDDEMTGYVATRWYRAPEIMLNWMHYNQTVDIWSVGCIMAELLTGRT
LFPGTDHIDQLKLILRLVGTPGAELLKKISSESARNYIQSLTQMPKMNFANVFIGANPLAVDLLEKMLVLDSDKRITAAQ
ALAHAYFAQYHDPDDEPVADPYDQSFESRDLLIDEWKSLTYDEVISFVPPPLDQEEMES
```

Predict Structure



195 million sequences in TrEMBL

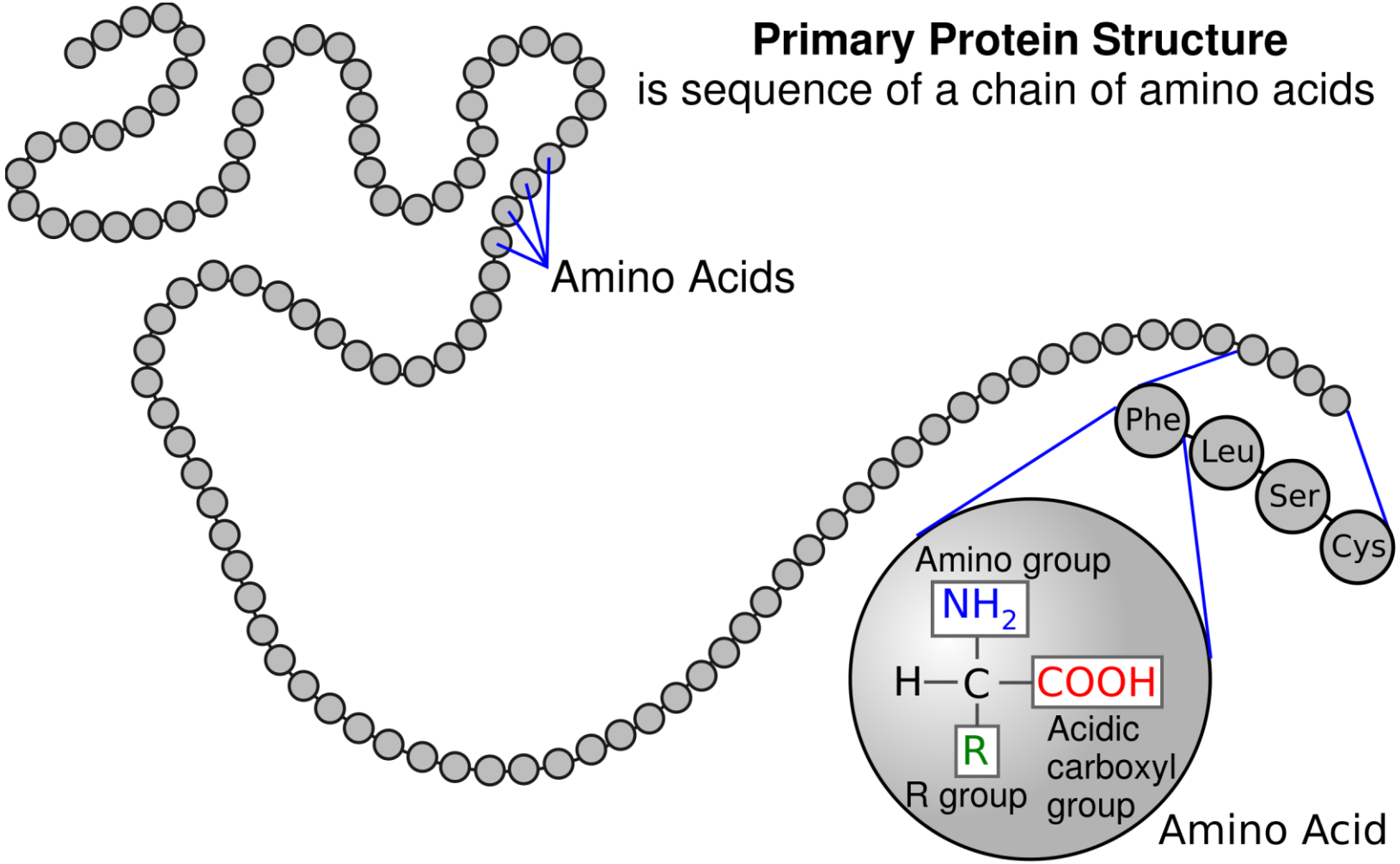
vs

170,000 structures in Protein Data Bank

Protein Structure

Four levels of Protein structure:

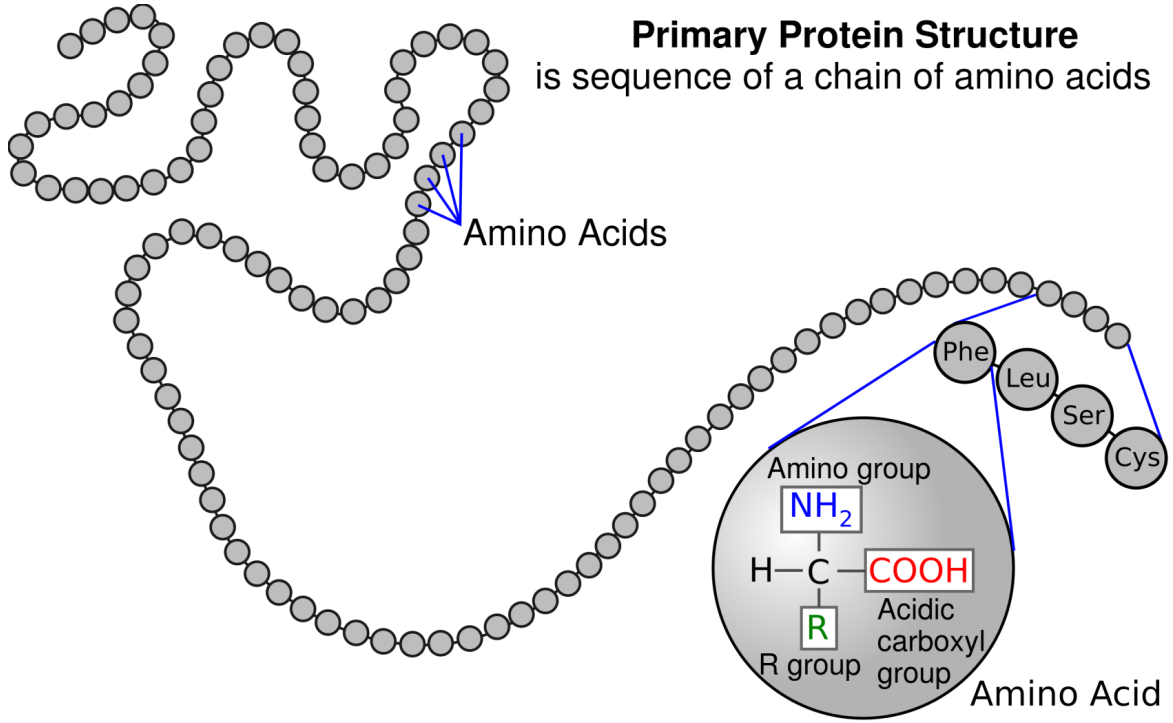
1. Primary

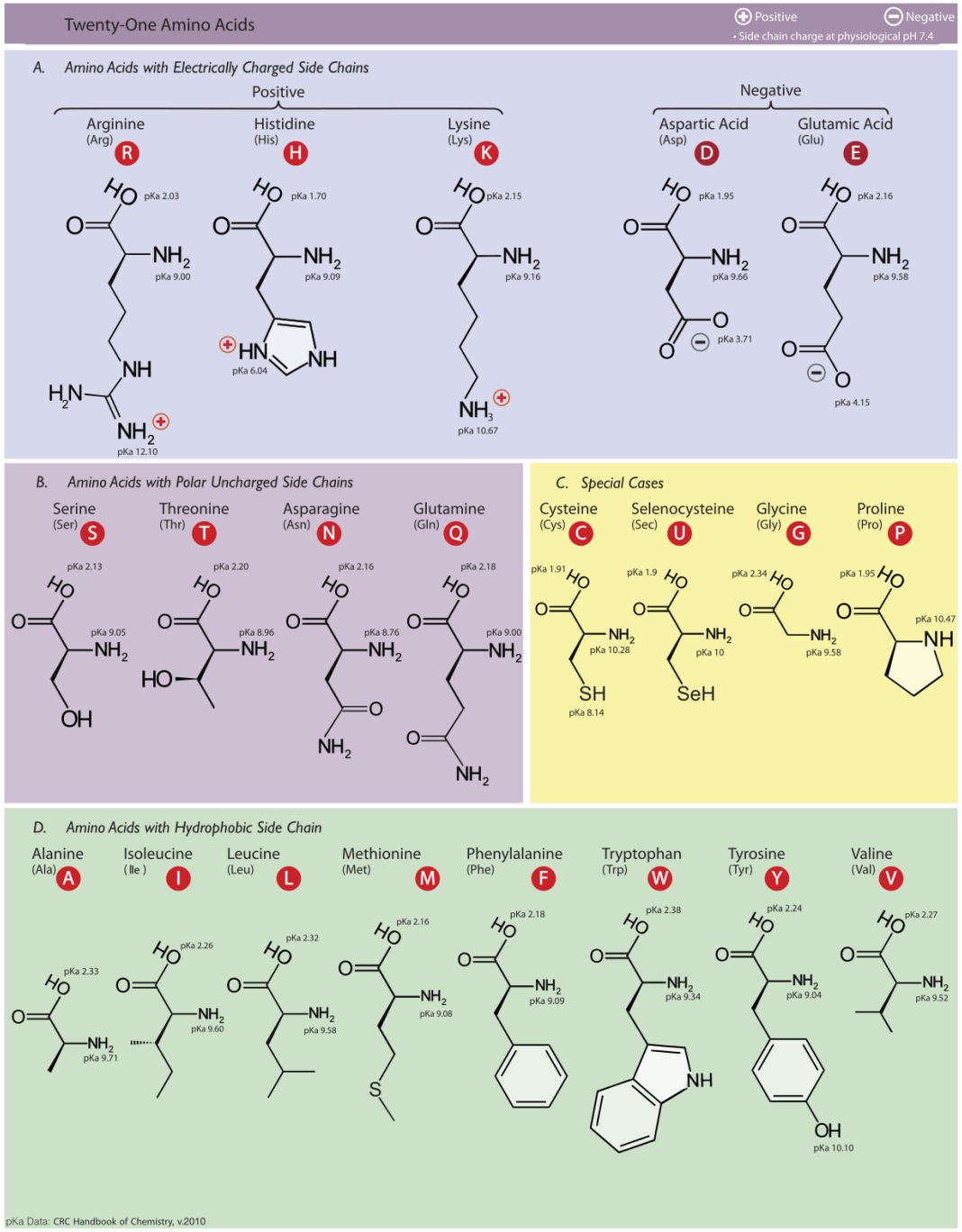


Protein Structure

Four levels of Protein structure:

1. Primary

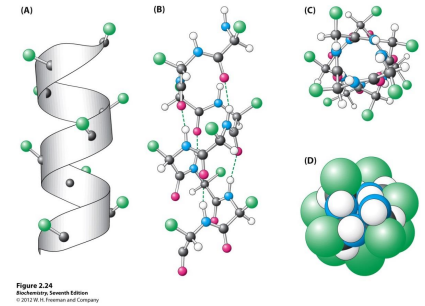


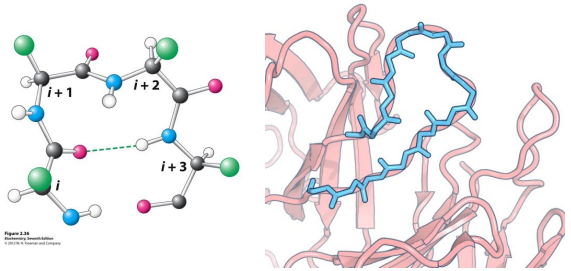


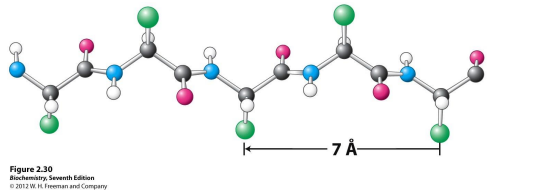
Protein Structure

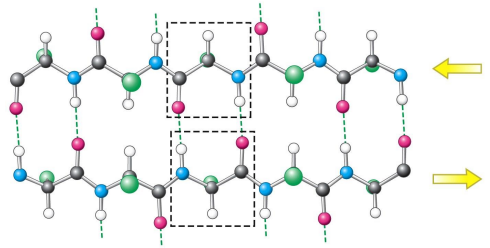
Four levels of Protein structure:

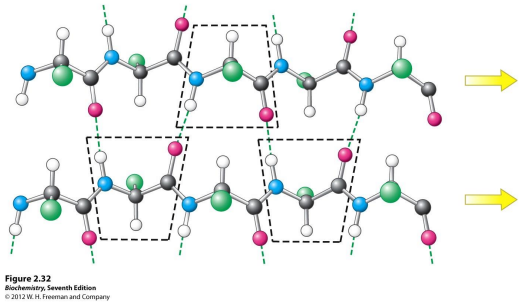
1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles
   - “3 State” – H E O











https://www.youtube.com/watch?v=1usemtIYe_s

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles
   - “3 State” – H E O

Alpha Helix

Side chain

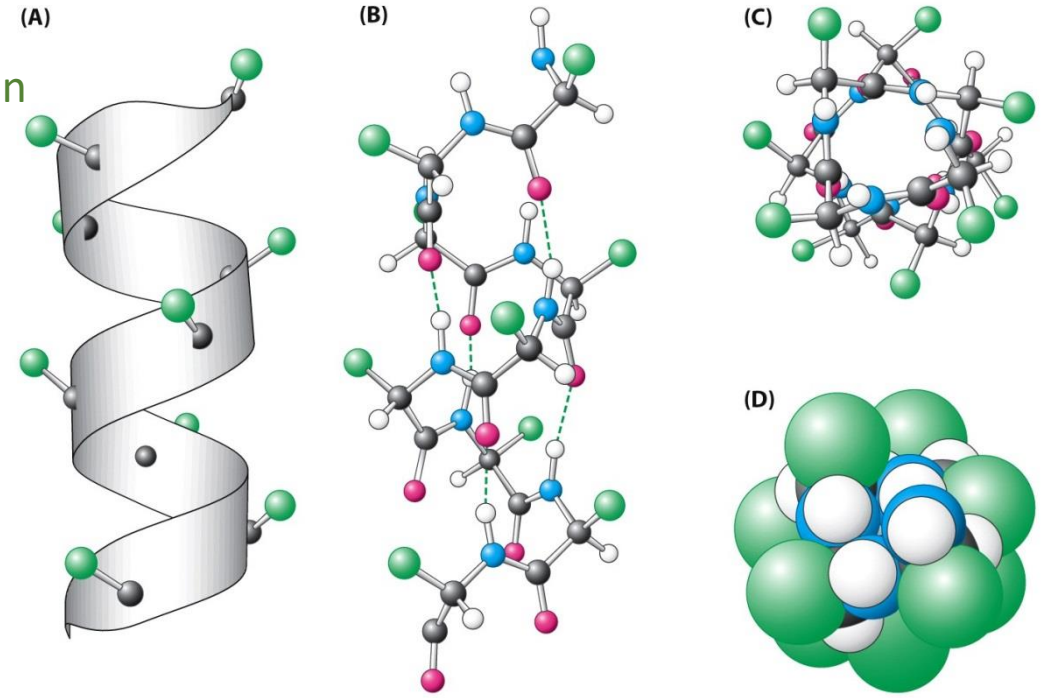


Figure 2.24
***Biochemistry*, Seventh Edition**

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles
   - “3 State” – H E O

Beta Strand

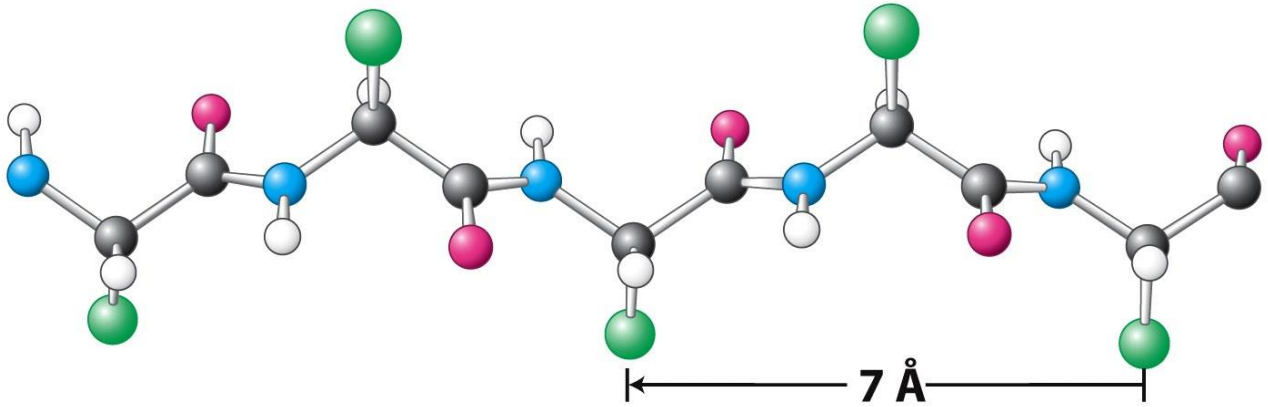


Figure 2.30
***Biochemistry*, Seventh Edition**

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles
   - “3 State” – H E O

Anti-Parallel Beta Sheet

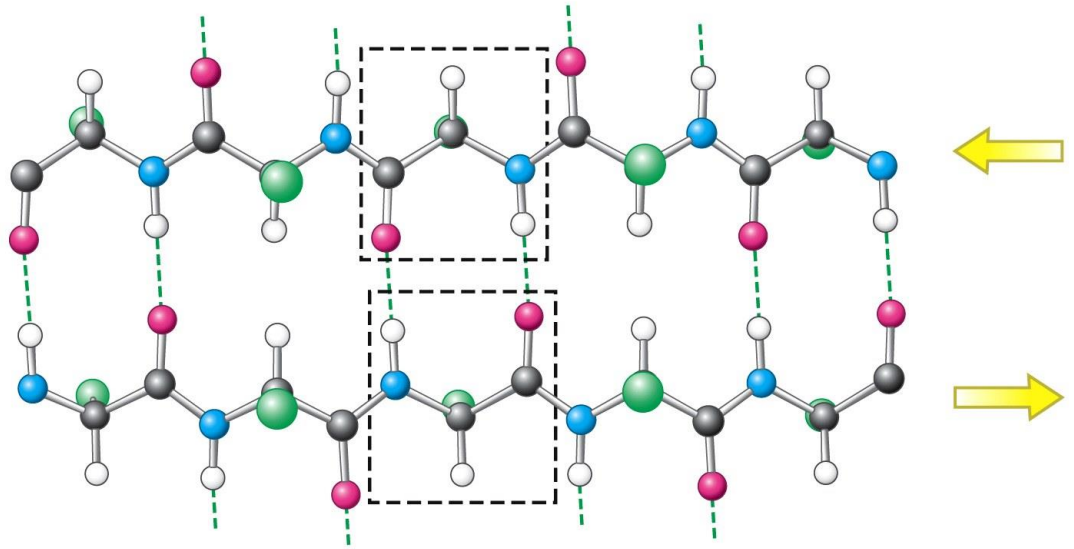


Figure 2.31
***Biochemistry*, Seventh Edition**

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles
   - “3 State” – H E O

Parallel Beta Sheet

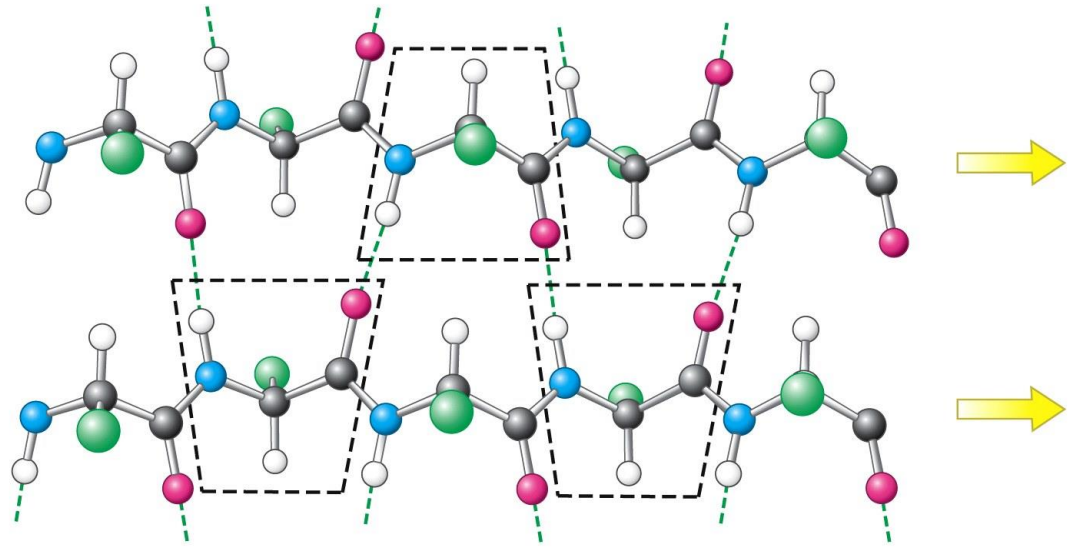


Figure 2.32
***Biochemistry*, Seventh Edition**

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles
   - “3 State” – H E O

Anti-Parallel Beta Sheet

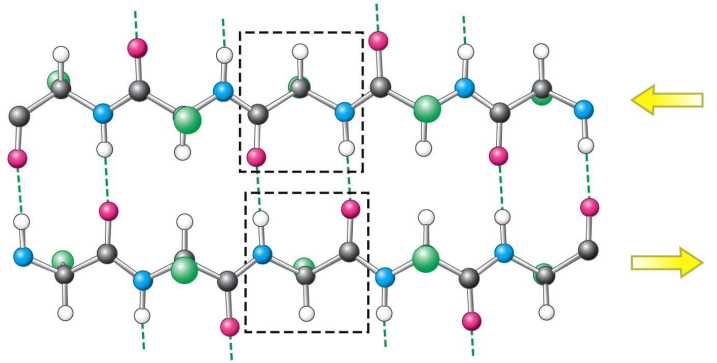


Figure 2.31
Biochemistry, Seventh Edition

Parallel Beta Sheet

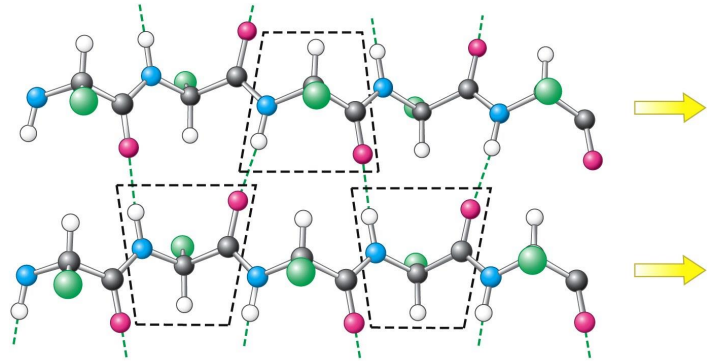
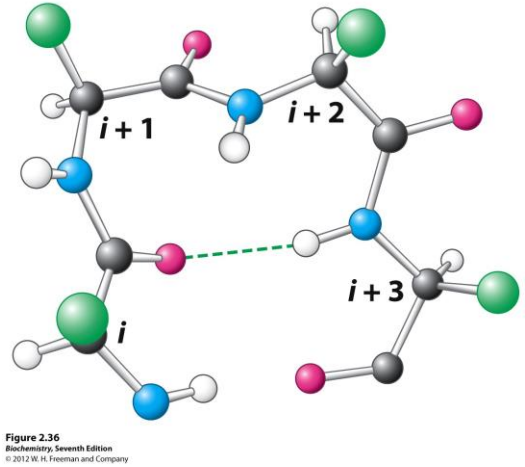


Figure 2.32
Biochemistry, Seventh Edition

Protein Structure

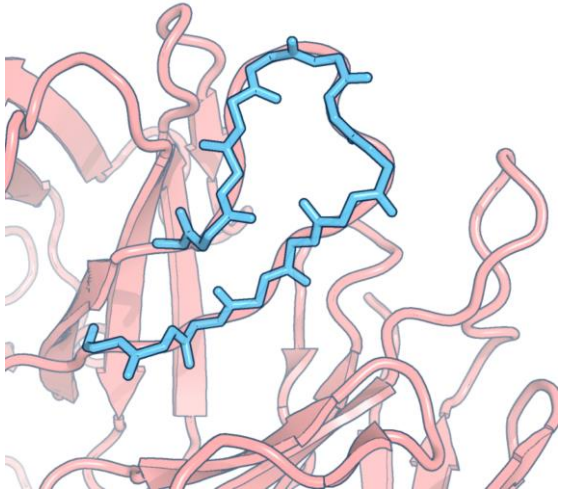
Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles



Beta-turn

"Other"



Loops

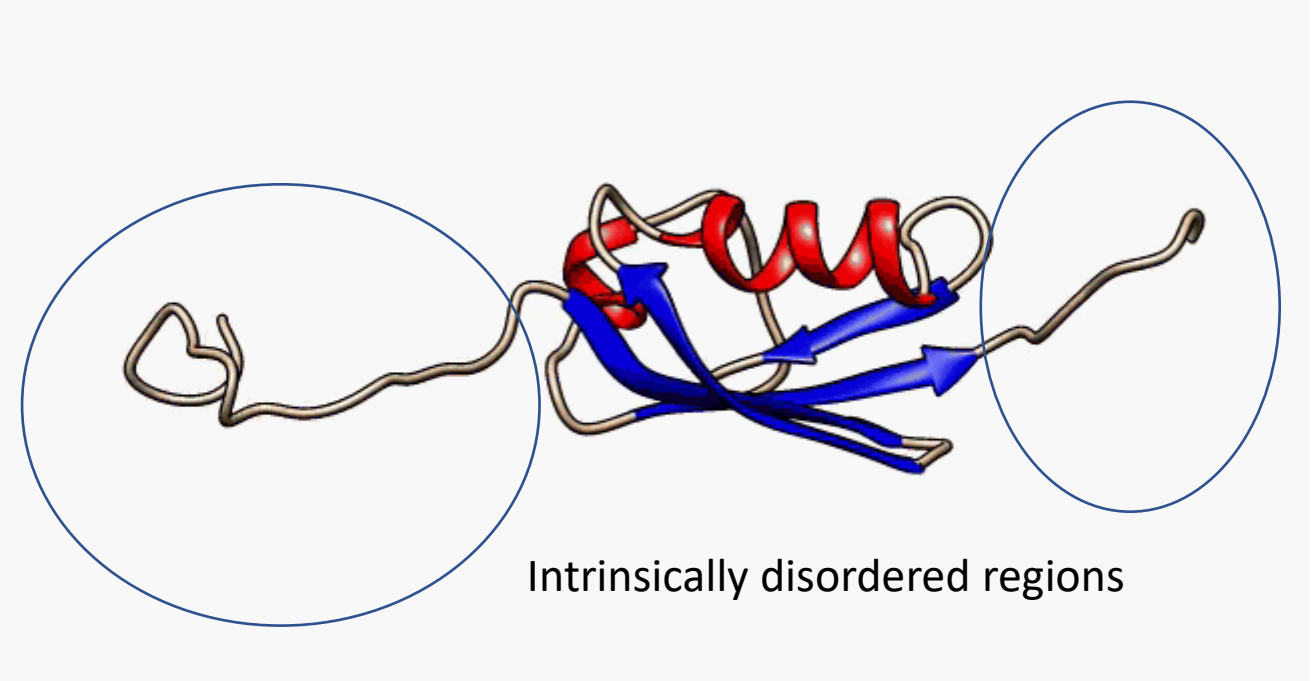
Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
   - 3D arrangement of local segments
   - Hydrogen Bonds
   - Phi/Psi angles

Intrinsic Disorder

No stable/fixed 3D structure

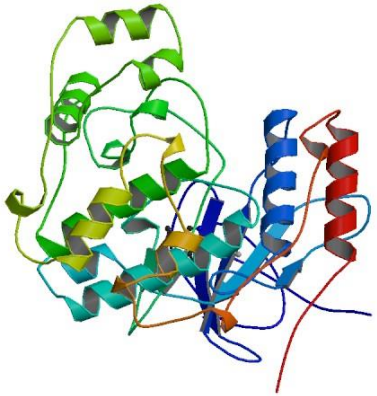


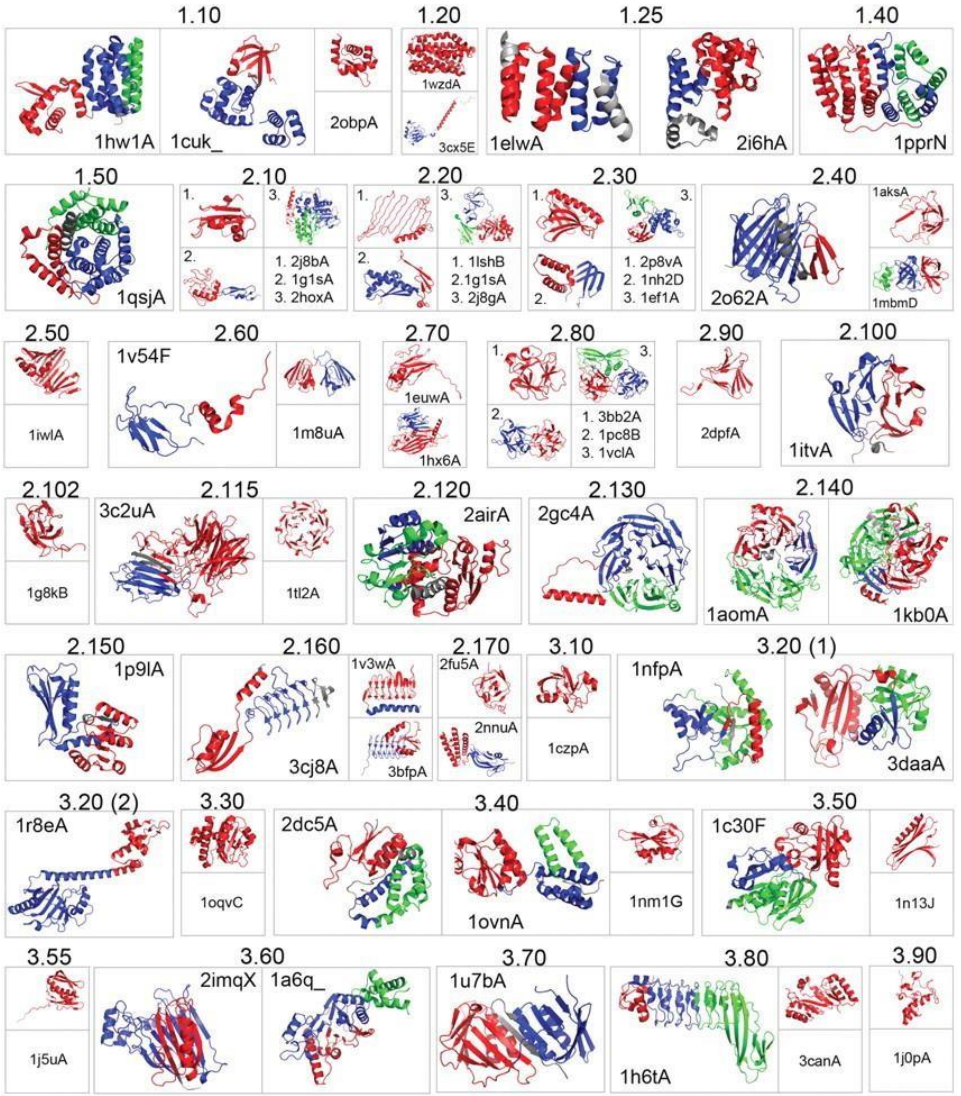
Intrinsically disordered regions

Protein Structure

Four levels of Protein structure:

1. Primary
2. Secondary
3. Tertiary
   - Protein **Fold**
   - A single chain



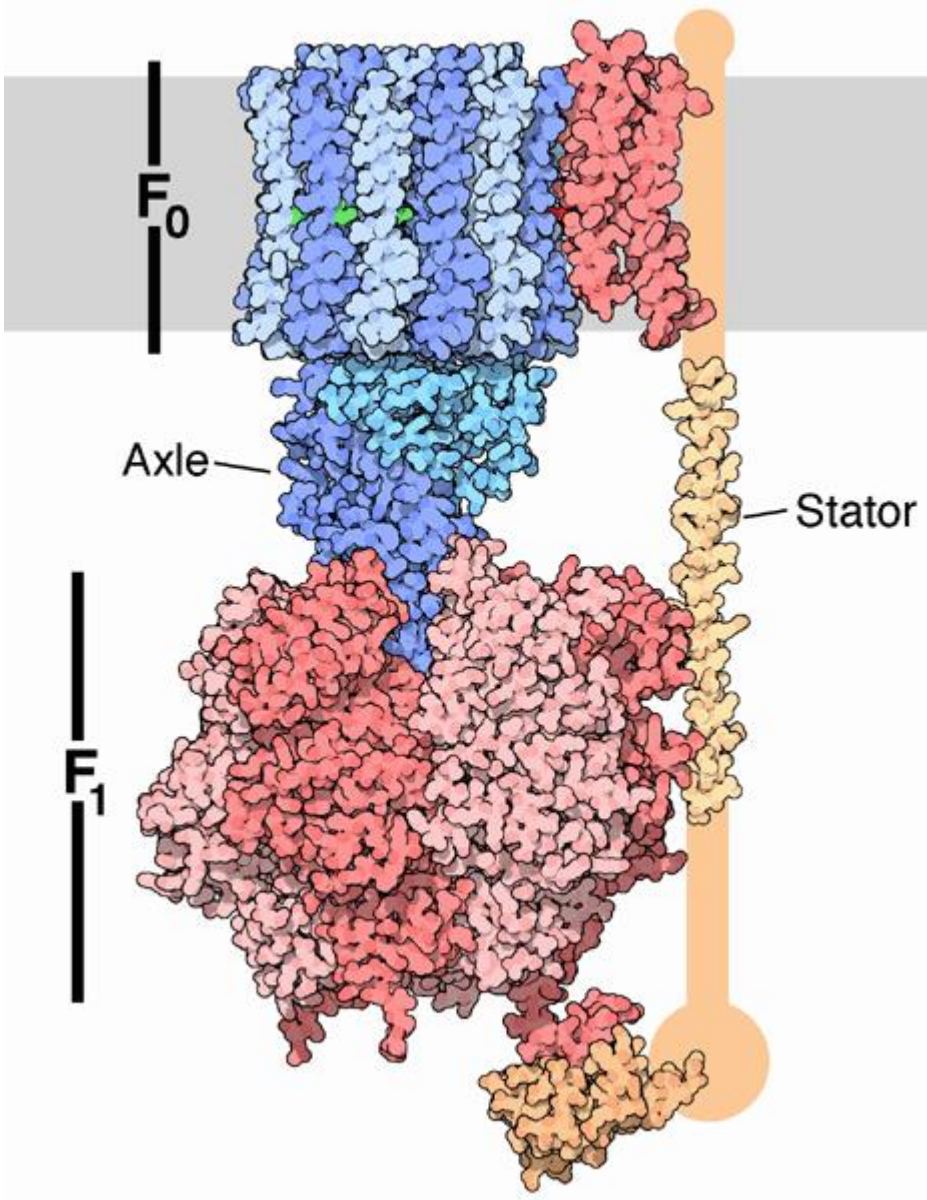


CATH/GENE3D

Protein Structure

Four levels of Protein structure:

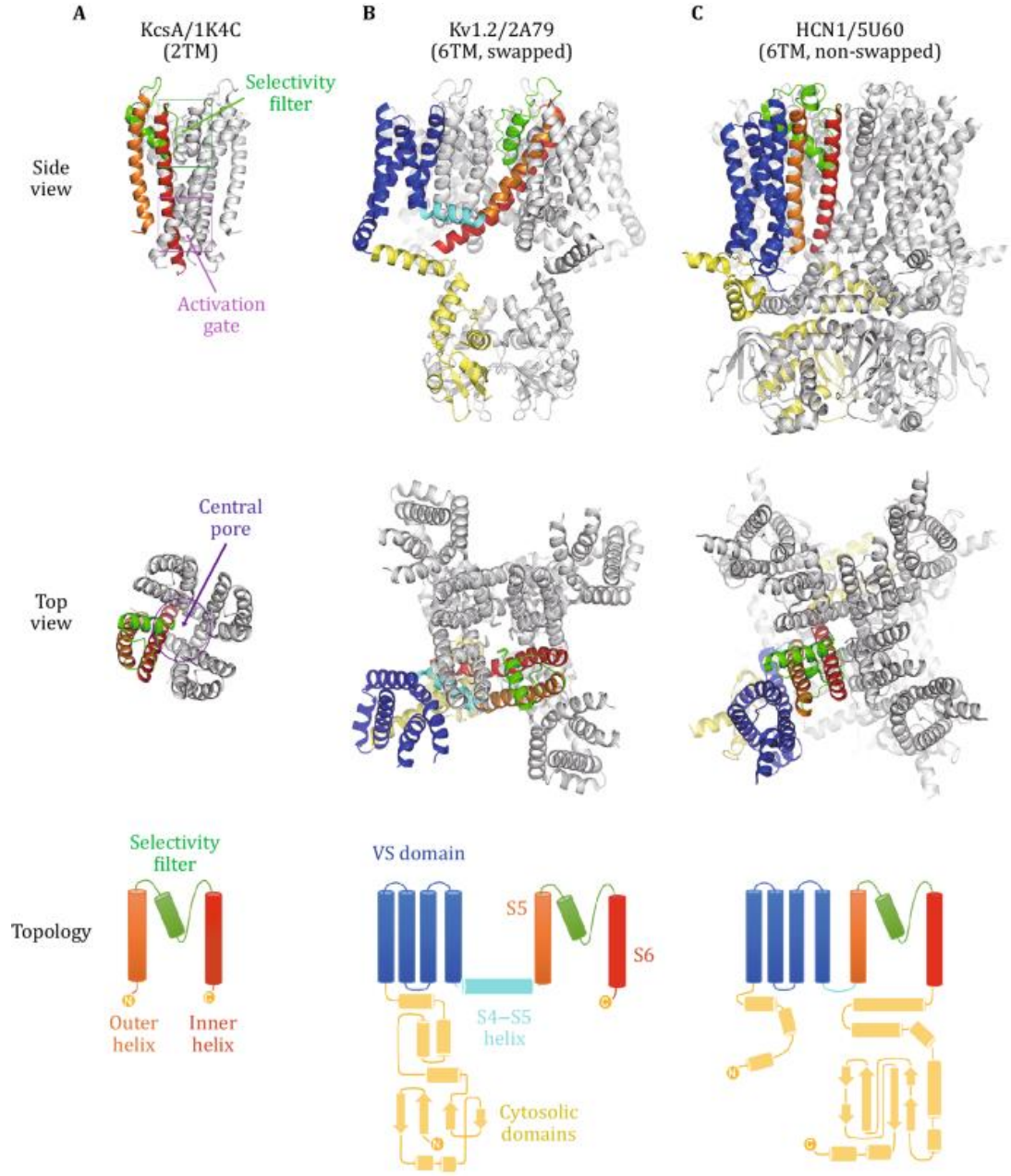
1. Primary
2. Secondary
3. Tertiary
4. Quaternary
   - Arrangement of multiple subunits
   - Protein complexes



Protein Structure

Four levels of Protein structure:

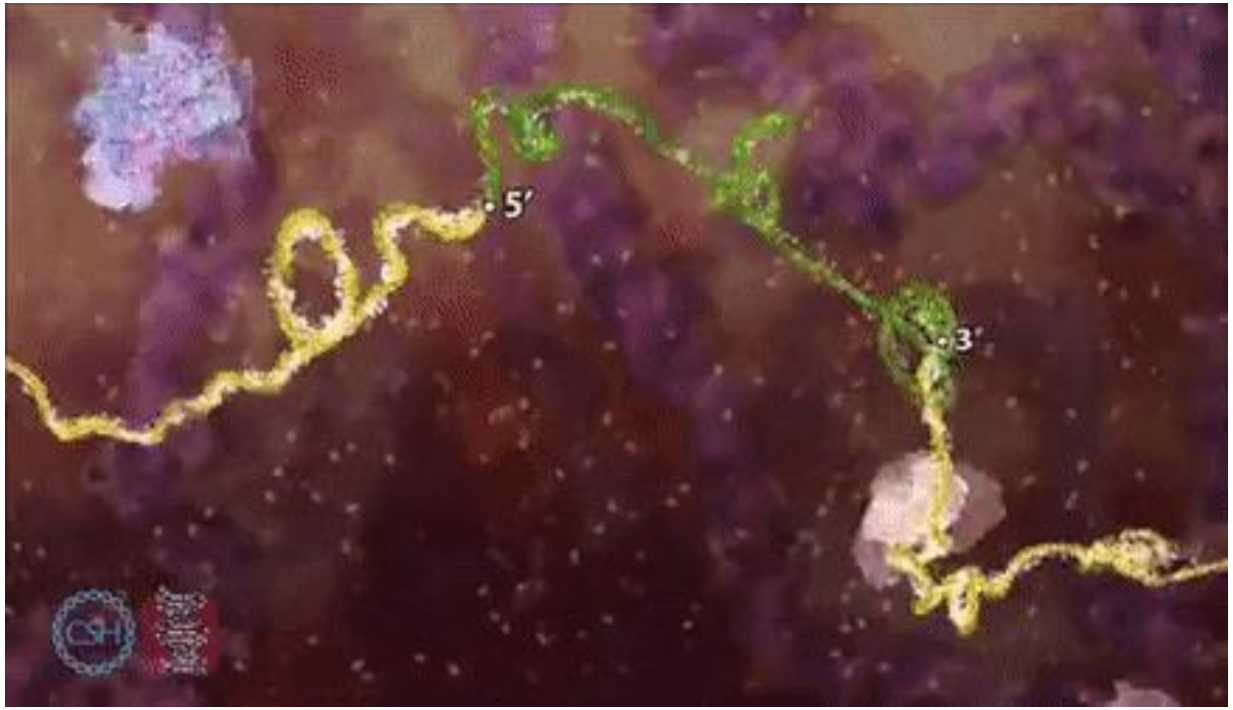
1. Primary
2. Secondary
3. Tertiary
4. Quaternary
   - Arrangement of multiple subunits
   - Protein complexes



Protein Structure

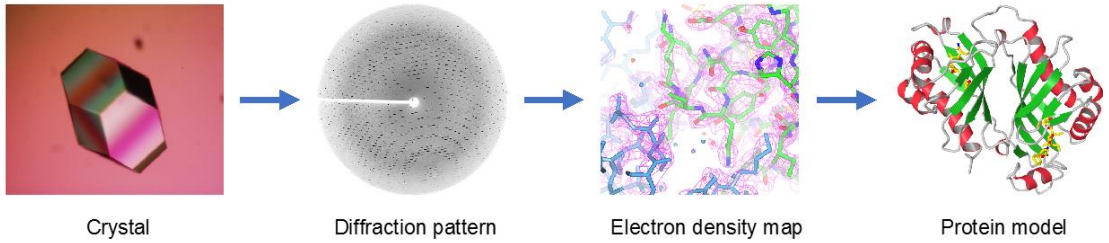
Four levels of Protein structure:

1. Primary
2. Secondary
3. Tertiary
4. Quaternary
   - Arrangement of multiple subunits
   - Protein complexes

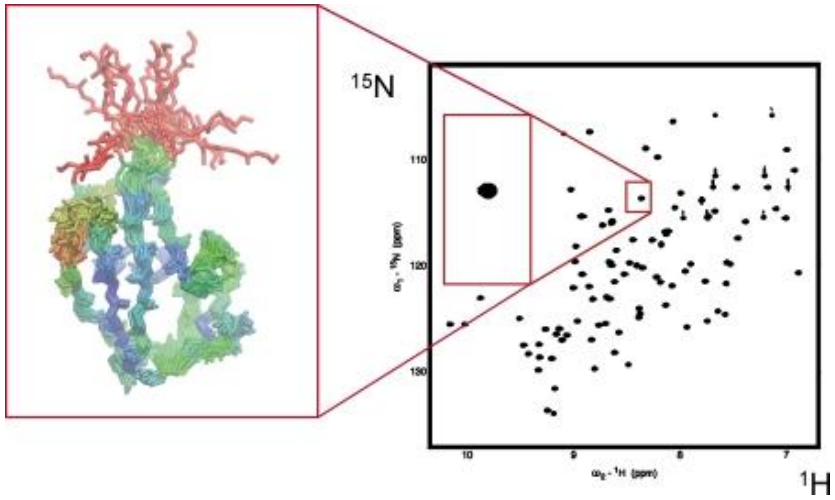


Protein Structure

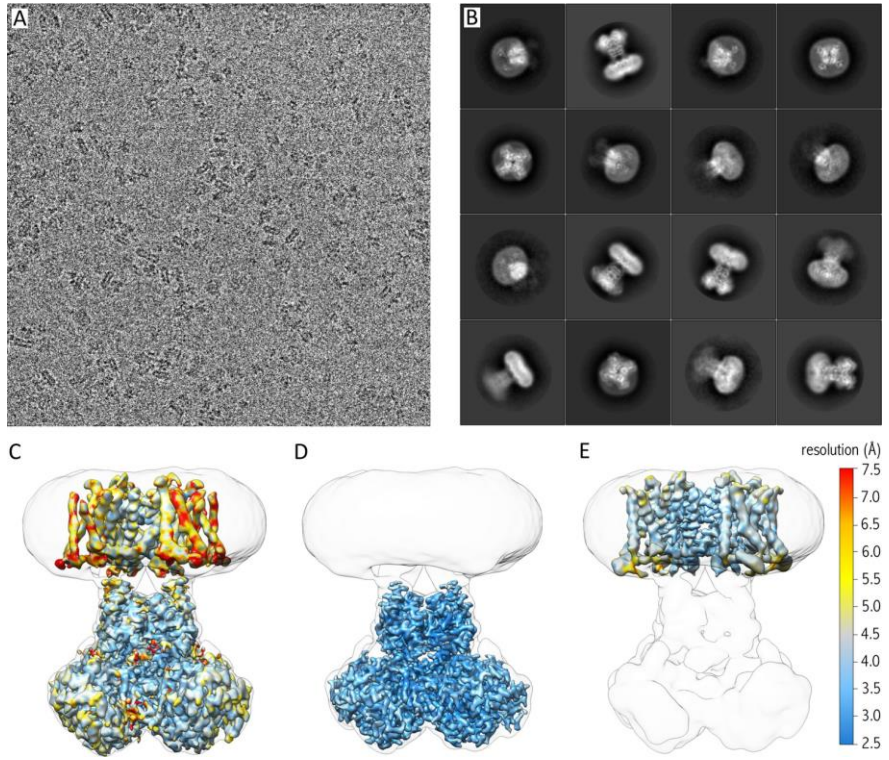
Experimental methods



X-ray crystallography



NMR spectroscopy



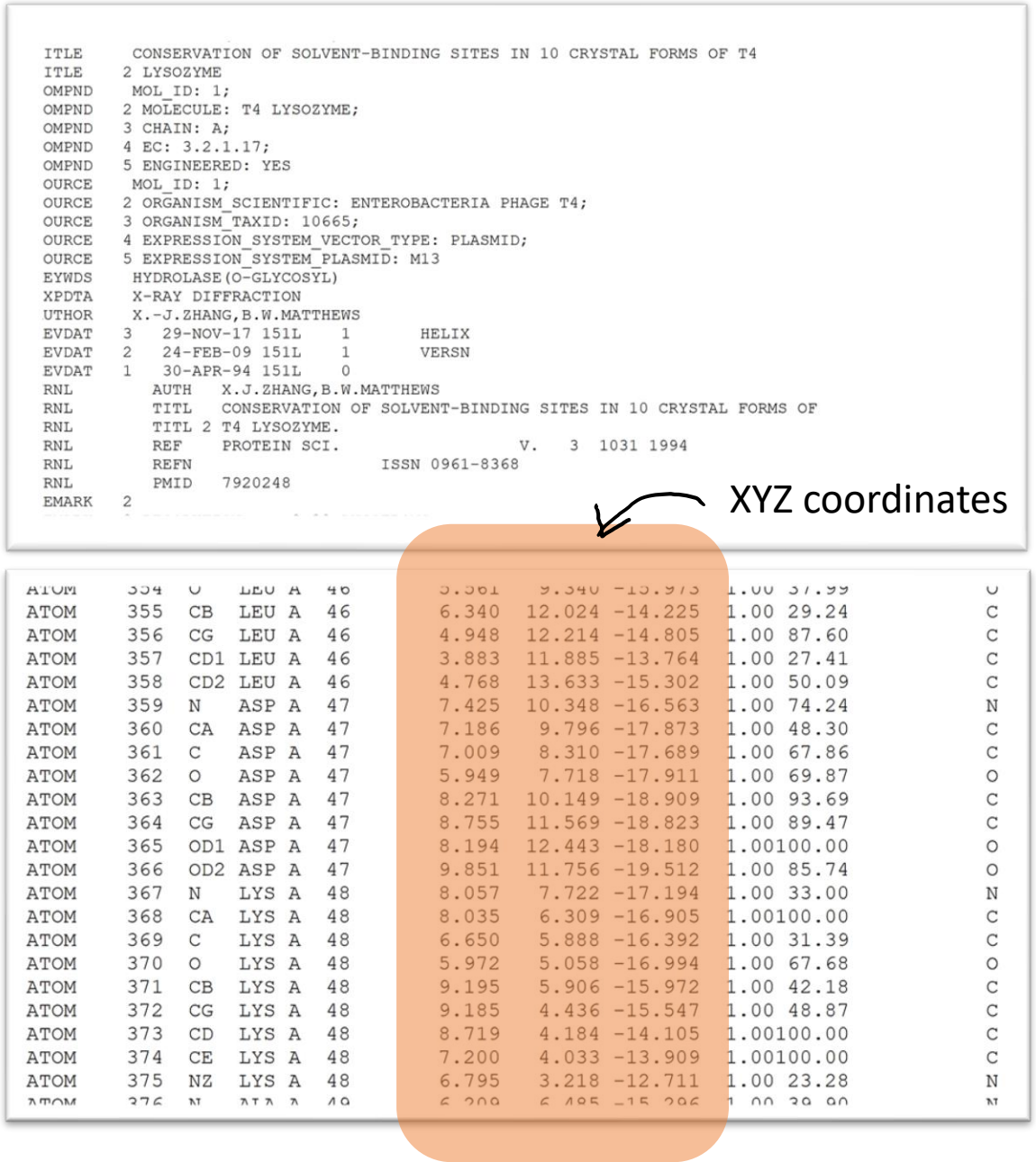
Cryo-EM microscopy

https://pdb101.rcsb.org/

Structural Data

File formats

- There are two main formats in use:
  - **PDB files**
    - Older format
    - More human readable
    - Common output of modelling servers



Structural Data

File formats

- There are two main formats in use:
  - PDB files
    - Older format
    - More human readable
    - Common output of modelling servers
  - **mmCIF files**
    - Newer format
    - More computer friendly – nightmare to read
    - **BioPython** is your friend

```
data_151L
#
_entry.id   151L
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    5.287
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
PDB   151L
WWPDB D_1000170119
#
_pdbx_database_status.status_code                     REL
_pdbx_database_status.entry_id                        151L
_pdbx_database_status.recvd_initial_deposition_date   1994-01-25
_pdbx_database_status.deposit_site                    ?
_pdbx_database_status.process_site                    BNL
_pdbx_database_status.SG_entry                        .
_pdbx_database_status.status_code_sf                  REL
_pdbx_database_status.pdb_format_compatible           Y
_pdbx_database_status.status_code_mr                  ?
_pdbx_database_status.status_code_cs                  ?
_pdbx_database_status.methods_development_category    ?
#
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Zhang, X.-J.'   1
'Matthews, B.W.' 2
#
_citation.id                        primary
```

Structural Data

File formats

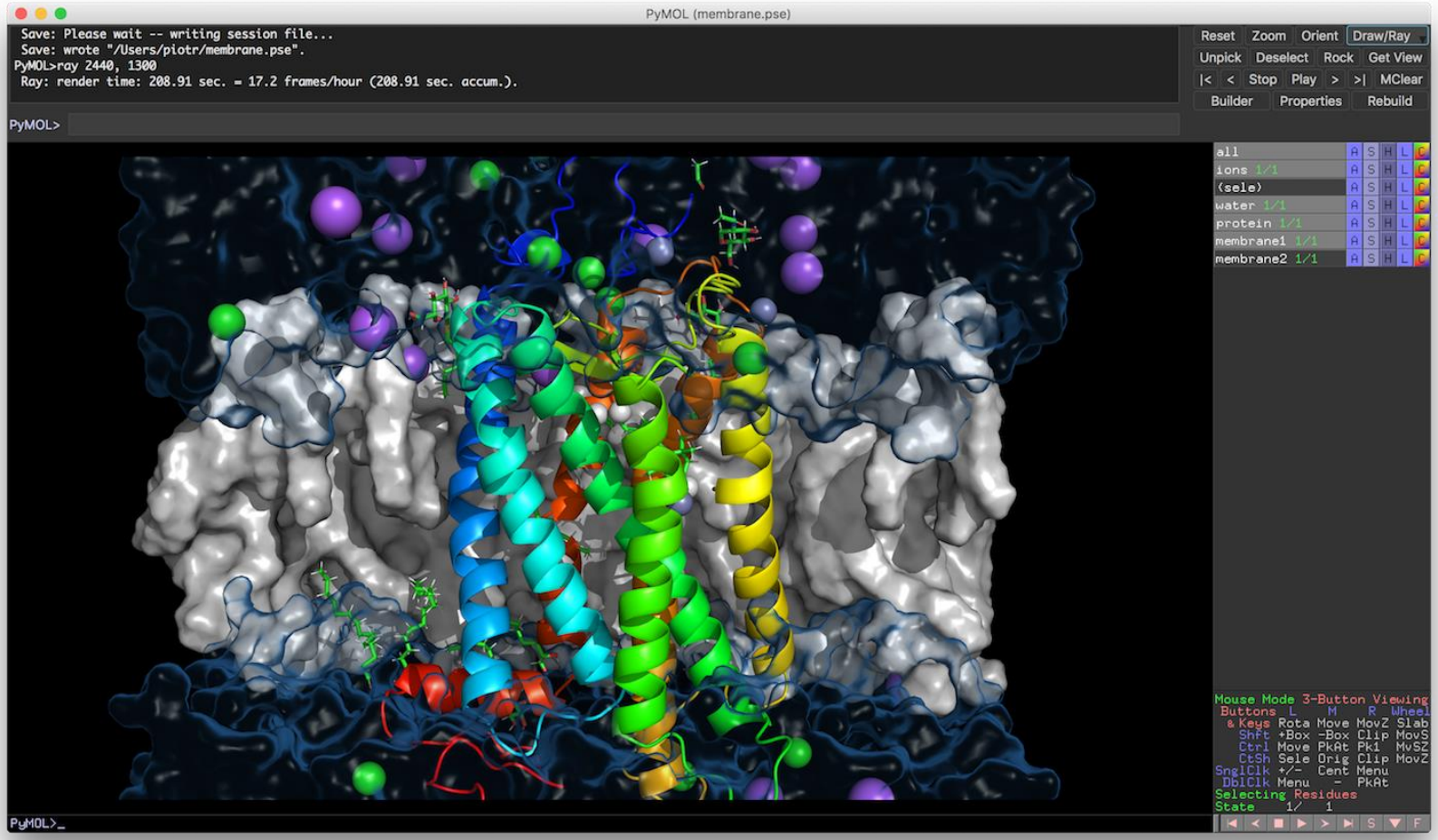
- The atomic coordinates of proteins are stored in the Protein Data Bank
- Experimentally determined
  - X-ray crystallography
  - Cryo-EM
  - NMR spectroscopy
- PDB IDS – “XXXX”, 4 letters and numbers

Protein Structure

https://pymol.org/edu/?q=educational/

File formats

- Can open both in protein visualisation software
- Or read them in a plain text editor

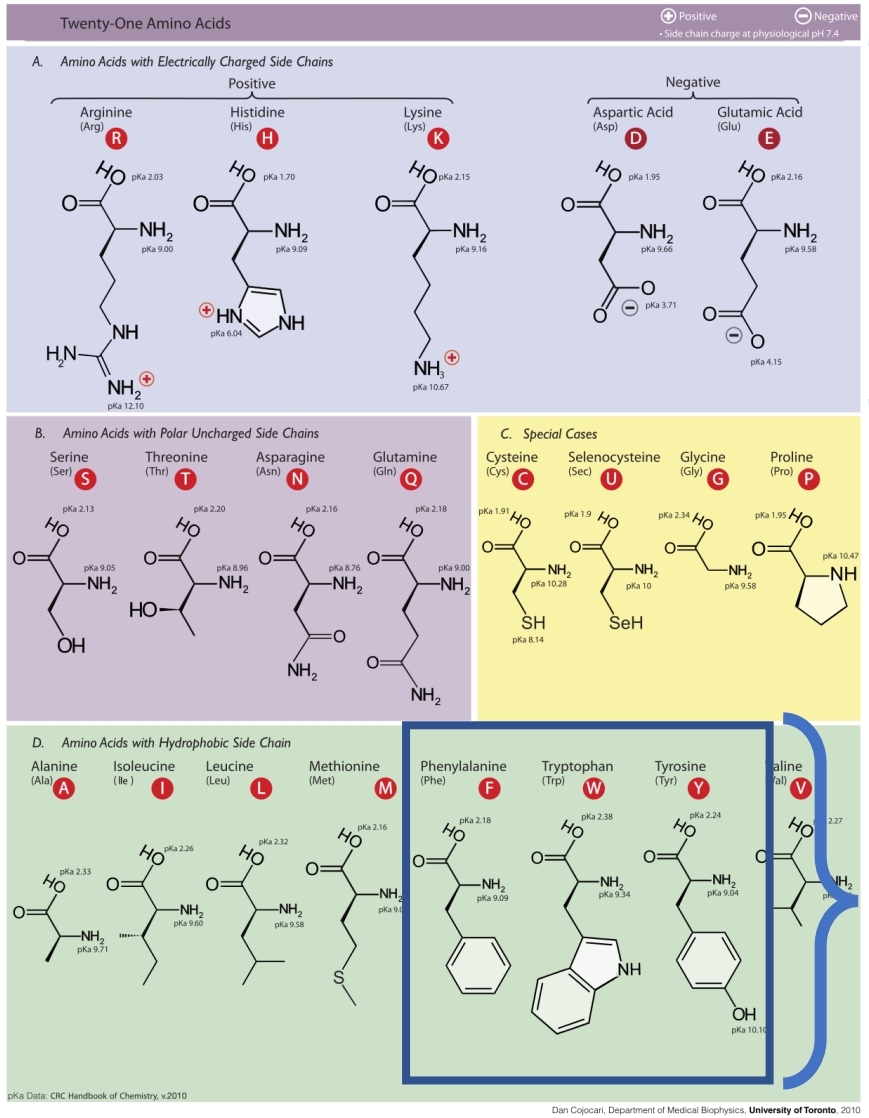


Seq to Structure

How to do we go from sequence to structure?

Sequence gives us:

1. **Type** of residue



Charged

Hydrophobic

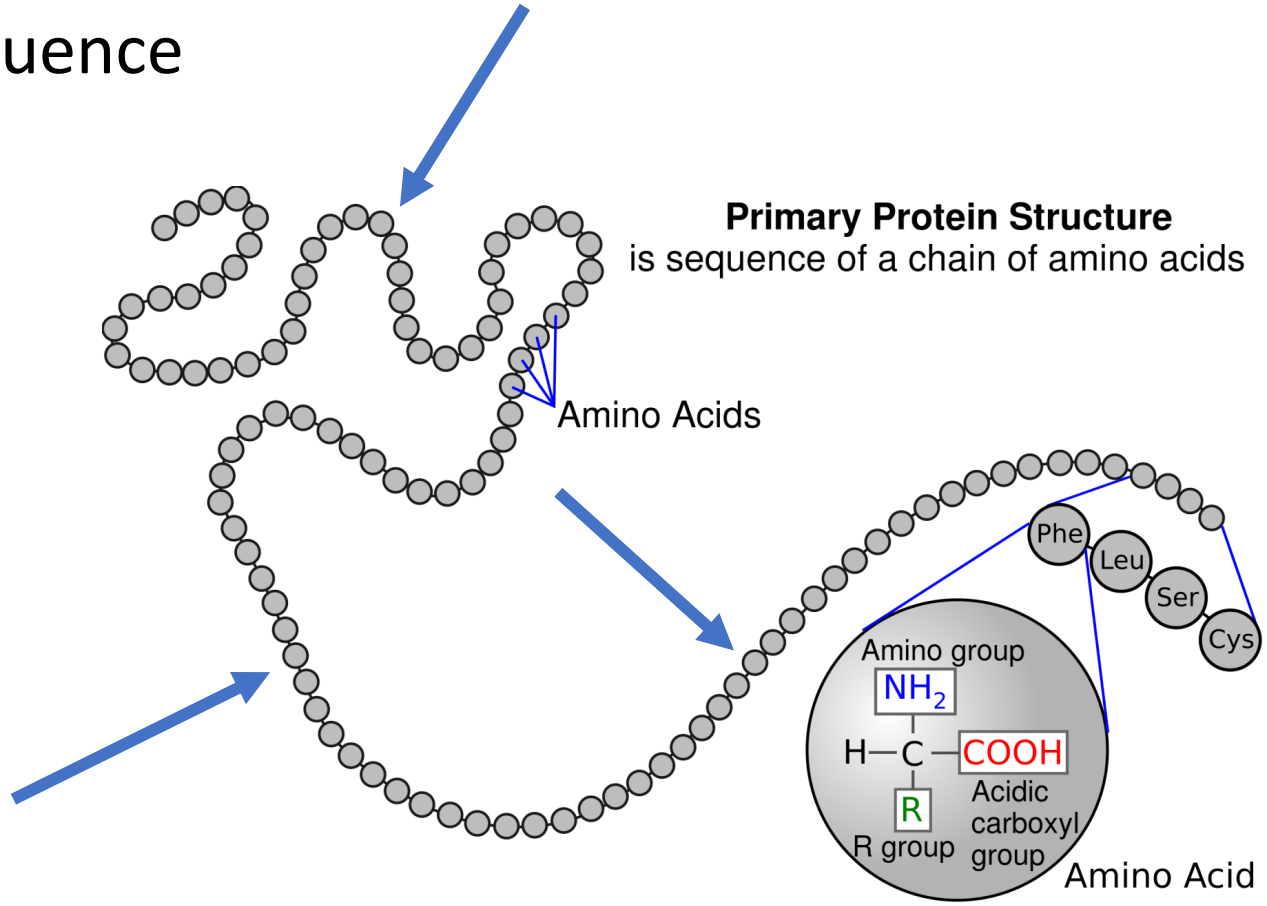
Aromatic

Seq to Structure

How to do we go from sequence to structure?

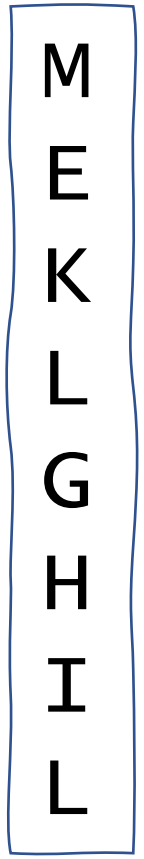
Sequence gives us:

1. **Type** of residue
2. **Position** of residue



Seq to Structure

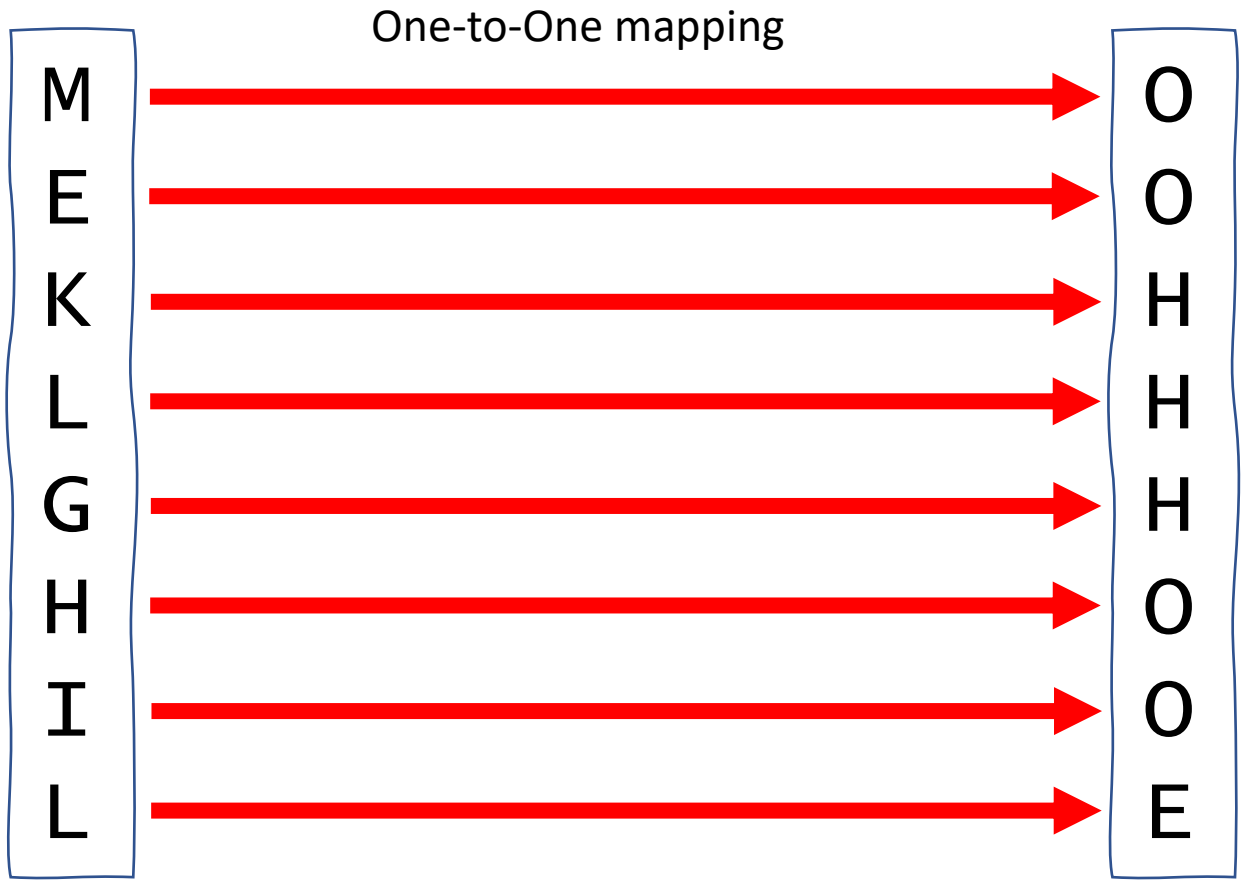
Predicting Secondary structure:



H Helix
E Beta Strand
O Other

Seq to Structure

Predicting Secondary structure:



Helix
Beta Strand
Other

Seq to Structure

Predicting Secondary structure:

- Simple version – AA propensities from protein structures
  - Poor accuracy
  - No position information
- Example - **Chou and Fasman**

Amino acid	Frequency	P_α	P_β	P_c
A—Ala	7.73	1.39	0.75	0.8
C—Cys	1.84	0.74	1.31	1.05
D—Asp	5.82	0.89	0.55	1.33
E—Glu	6.61	1.35	0.72	0.86
F—Phe	4.05	1.01	1.43	0.76
G—Gly	7.11	0.47	0.65	1.62
H—His	2.35	0.92	0.99	1.07
I—Ile	5.66	1.04	1.71	0.59
K—Lys	6.27	1.11	0.83	1
L—Leu	8.83	1.32	1.1	0.68
M—Met	2.08	1.21	0.99	0.83
N—Asn	4.5	0.77	0.62	1.39
P—Pro	4.52	0.5	0.44	1.72
Q—Gln	3.94	1.29	0.76	0.89
R—Arg	5.03	1.17	0.91	0.91
S—Ser	6.13	0.82	0.85	1.24
T—Thr	5.53	0.76	1.23	1.07
V—Val	6.91	0.89	1.86	0.64
W—Trp	1.51	1.06	1.3	0.79
Y—Tyr	3.54	0.95	1.5	0.78

Seq to Structure

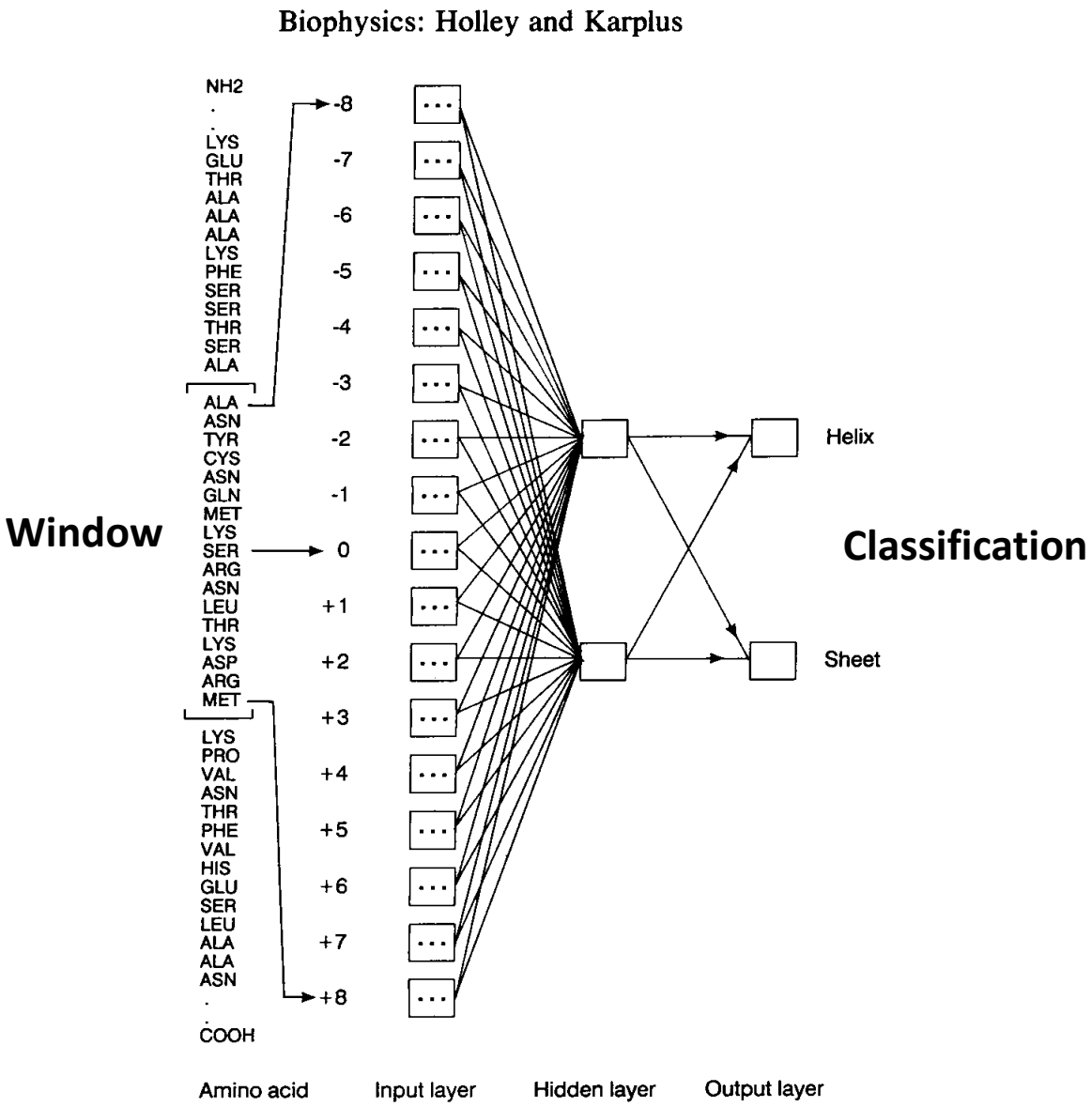
Predicting Secondary structure:

- Integration of position information – “sliding windows”
- Scoring matrices
- Example – **GOR algorithm**

Seq to Structure

Predicting Secondary structure:

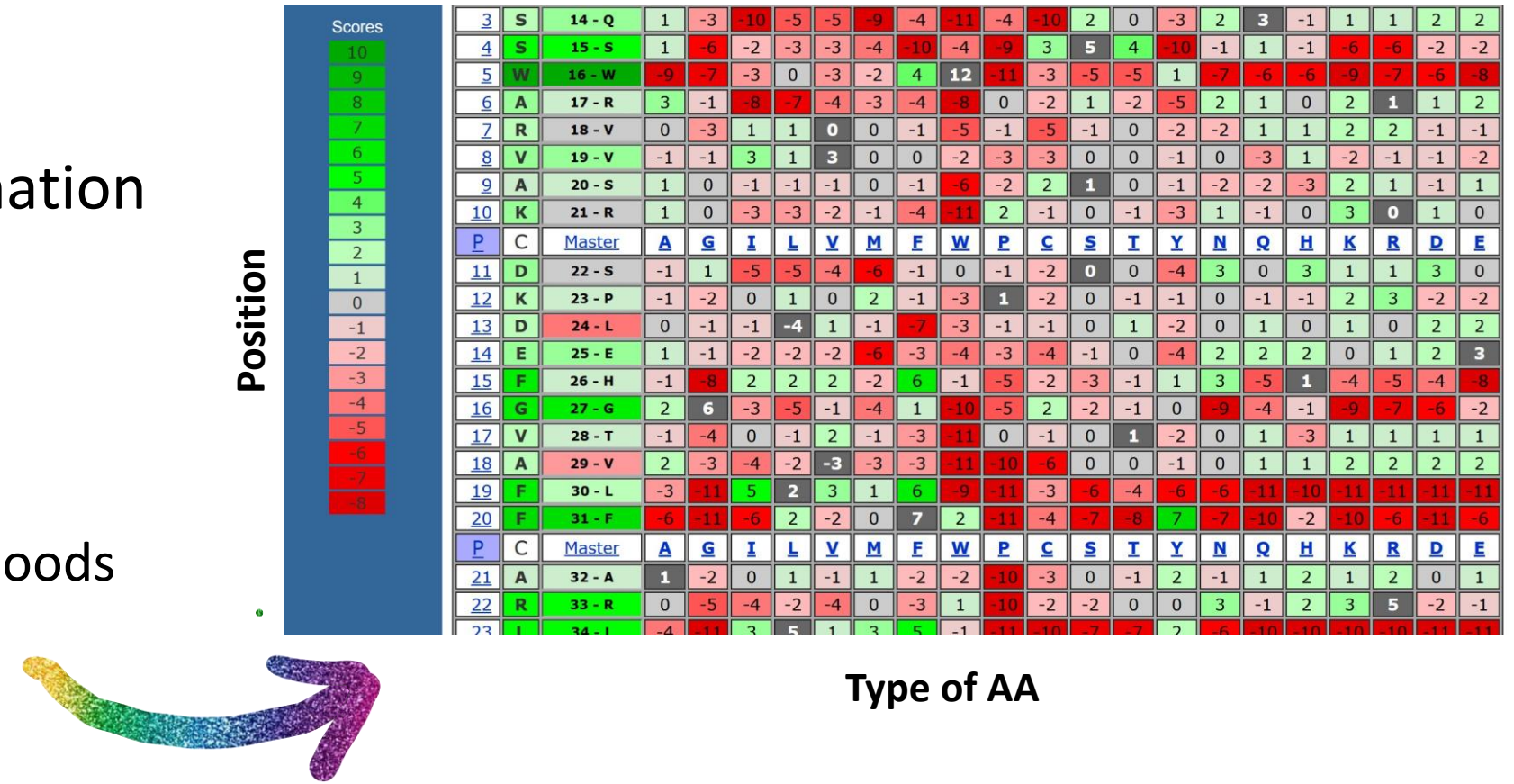
- Machine learning algorithms
  - Neural Networks
  - Support vector machines
  - Random Forest



Seq to Structure

Predicting Secondary structure:

- Evolutionary Information
  - Multiple sequence alignments
  - Profiles – probabilities/likelihoods
  - PSSMs – **PSIBLAST**



doi:10.1371/journal.pone.0168453.g003

Seq to Structure

Predicting Secondary structure:

- Nowadays, **Deep learning**
  - **Evolutionary information**
  - Prediction of contacts
  - Prediction of Phi/Psi angles
  - Prediction of SASA
- **~85% accuracy**

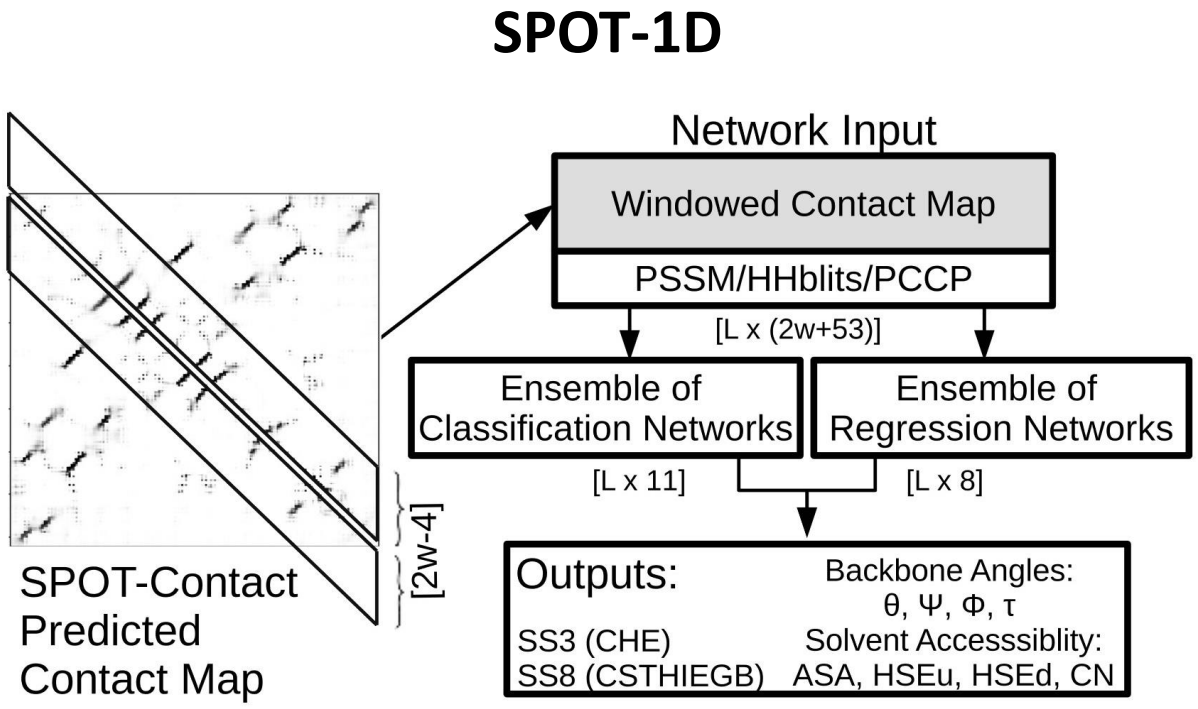


Figure S1: The overview of the SPOT-1D architecture for a target protein of length L and a window size of w.

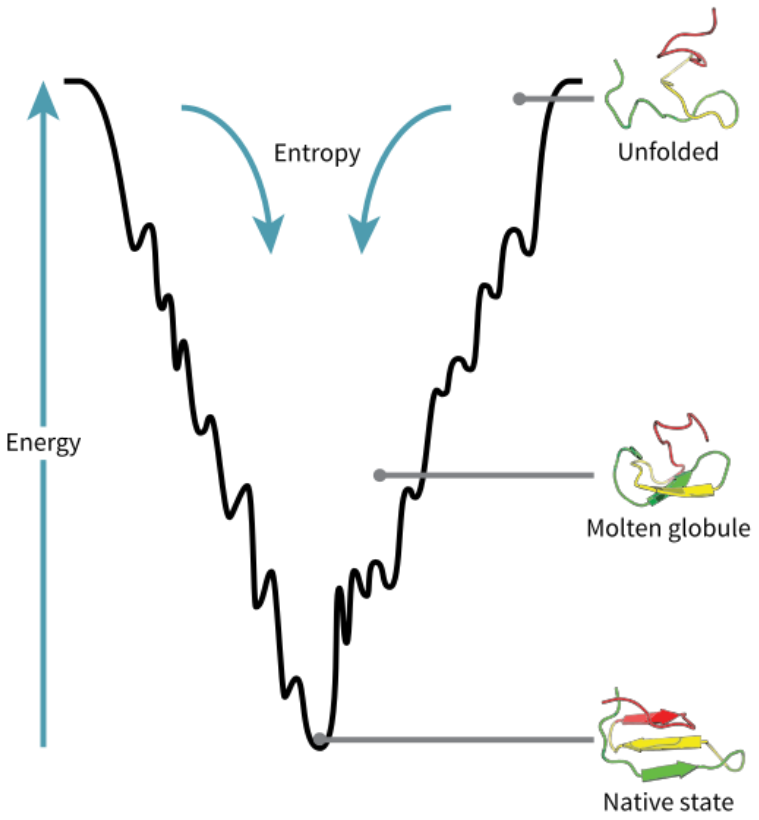
Hanson et al. 2019

Seq to Structure

- How do we predict **tertiary** structure or fold?

- Difficulties:
  - Search space is **MASSIVE**
    - Levinthal's Paradox (1960s) – time to sample >>>> age of universe – **POWER LAW**
  - Incomplete understanding of physics of protein folding
    - Protein-solvent interactions exhibit quantum-wavelike behaviour
  - Computational costs

- Need to reduce search space



Seq to Structure

Structure is **MORE CONSERVED** than sequence

```
DYDLKFGMNAGTSSNEYKAAEMFAKEVKEKSQGKIEISLYPSSQLGDDRAMLKQLKDGSLDFTFAESARF  2cex_1
VTWRLASSFPKSLDTIFGGAEVLSKXLSEATDGNFQIQVFSAGELVPGLQAADAVTEGTVECCHTVGYYY  2hzk_2


QLFYPEAAVFALPYVISNYNVAQKALFDTEFGKDLIKKMDKDLGVTLLSQAYNGTRQTTSNRAINSIADM  2cex_1
WGKDPTFALAAAVPFSLSARGINAWHYHGGGIDLYNEFLSQHNIVAFPGGNTGVQXGGWFRREINTVADX  2hzk_2


KGLKLRVPNAATNLAYAKYVGASPTPMAFSEVYLALQTNAVDGQENPLAAVQAQKFYEVQKFLAMTNHIL  2cex_1
QGLKXRVGGFAGKVXERLGVVPQQIAGGDIYPALEKGTIDATEWVGPYDDEKLGFFKVAPYYYYPGWWEG  2hzk_2


NDQLYLVSNETYKELPEDLQKVVKDAAENAAKYHTKLFVDGEKDLVTFFEKQGVKITHPDLVPFKESMKP  2cex_1
GPTVHFXFNKSAYEGLTPTYQSLLRTACHAADANXLQLYDWKNPTAIKSLVAQGTQLRPFSPEILQACFE  2hzk_2


YYAEFVKQTGQKGESALKQIE  2cex_1
AANEVYAEXEASNPAFKKIWD  2hzk_2
```

Seq to Structure

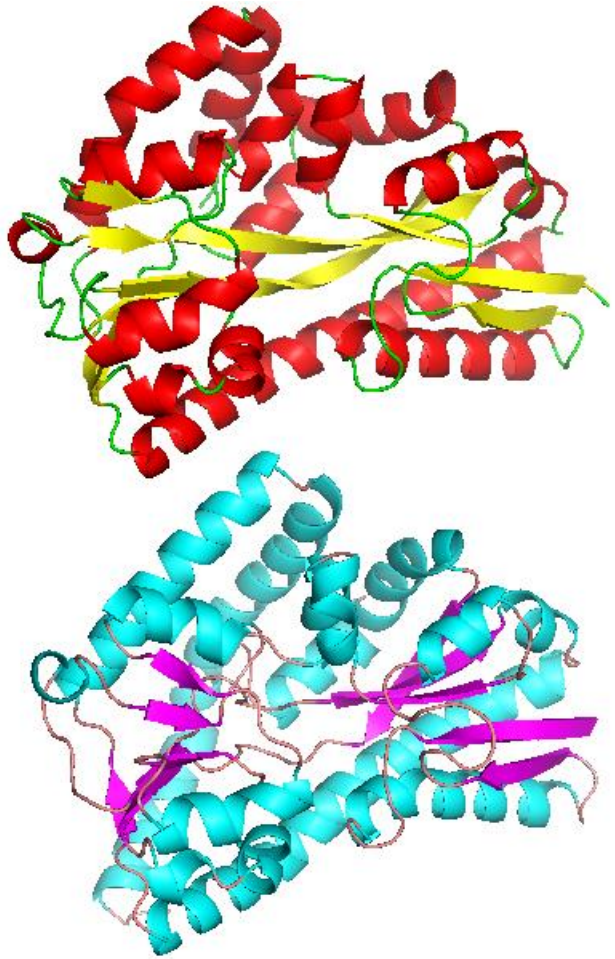
Structure is **MORE CONSERVED** than sequence

```
DYDLKFGMNAGTSSNEYKAAEMFAKEVKEKSQGKIEISLYPSSQLGDDRAMLKQLKDGSLDFTFAESARF  2cex_1
VTWRLASSFPKSLDTIFGGAEVLSKXLSEATDGNFQIQVFSAGELVPGLQAADAVTEGTVECCHTVGYYY  2hzk_2
                     **   *   *   *   *       *               *

QLFYPEAAVFALPYVISNYNVAQKALFDTEFGKDLIKKMDKDLGVTLLSQAYNGTRQTTSNRAINSIADM  2cex_1
WGKDPTFALAAAVPFSLSARGINAWHYHGGGIDLYNEFLSQHNIVAFPGGNTGVQXGGWFRREINTVADX  2hzk_2
    *   *   *                                *                    **  **

KGLKLRVPNAATNLAYAKYVGASPTPMAFSEVYLALQTNAVDGQENPLAAVQAQKFYEVQKFLAMTNHIL  2cex_1
QGLKXRVGGFAGKVXERLGVVPQQIAGGDIYPALEKGTIDATEWVGPYDDEKLGFFKVAPYYYYPGWWEG  2hzk_2
 ***  **    *         *                    *        *        *

NDQLYLVSNETYKELPEDLQKVVKDAAENAAKYHTKLFVDGEKDLVTFFEKQGVKITHPDLVPFKESMKP  2cex_1
GPTVHFXFNKSAYEGLTPTYQSLLRTACHAADANXLQLYDWKNPTAIKSLVAQGTQLRPFSPEILQACFE  2hzk_2
         *    *              *  **        *                    *

YYAEFVKQTGQKGESALKQIE  2cex_1
AANEVYAEXEASNPAFKKIWD  2hzk_2
   *              *
```

Exact matches = 35/ 301 = **11%**

Seq to Structure

Structure is **MORE CONSERVED** than sequence



2cex

Exact matches = 35/ 301 = **11%**

2hzk

Seq to Structure

Template-based structure prediction

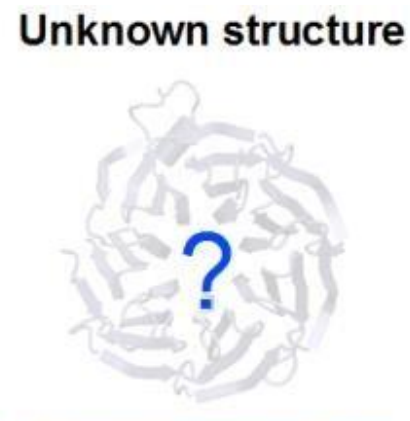
- Homology Modelling
- Protein Threading

De novo/Template-free structure prediction

- Physics based
- Knowledge-based

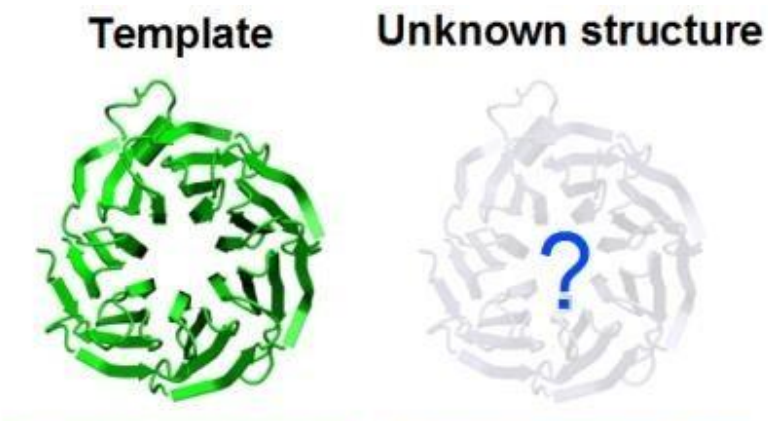
Template-based

- Homology modelling
  - Proteins that have similar sequence will probably have similar structures



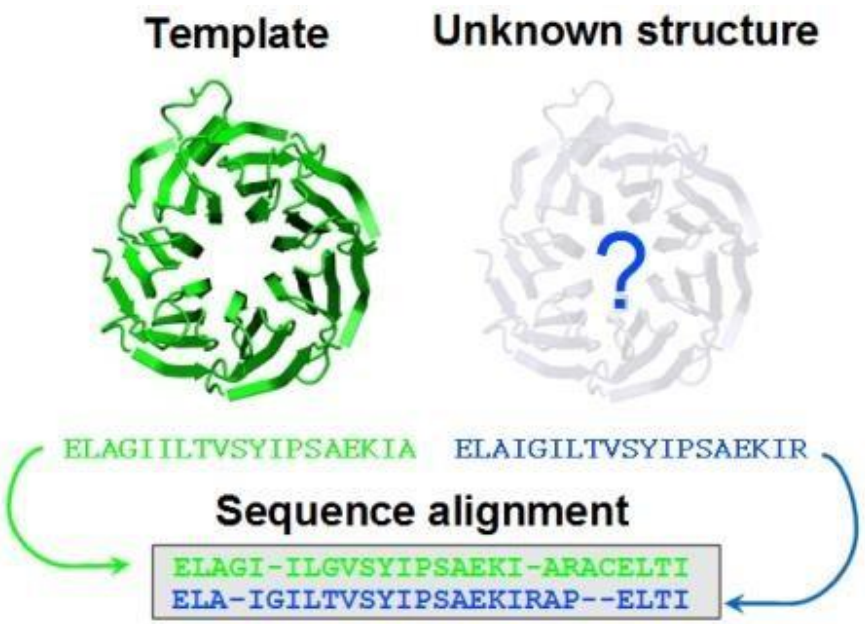
Template-based

- Homology modelling
  - Proteins that have similar sequence will probably have similar structures
  - Find most similar protein with a solved structure – **TEMPLATE**



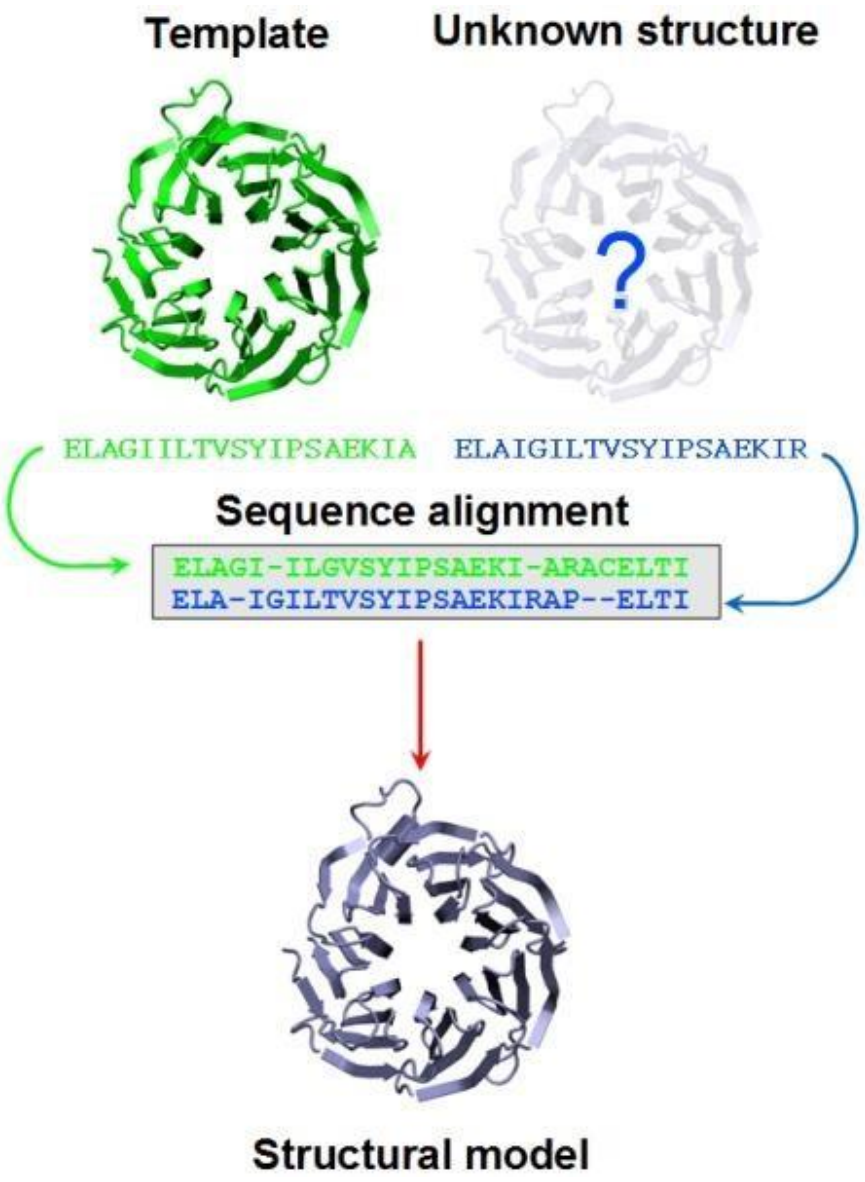
Template-based

- Homology modelling
  - Proteins that have similar sequence will probably have similar structures
  - Find most similar protein with a solved structure – **TEMPLATE**
  - **Sequence-Sequence** alignment



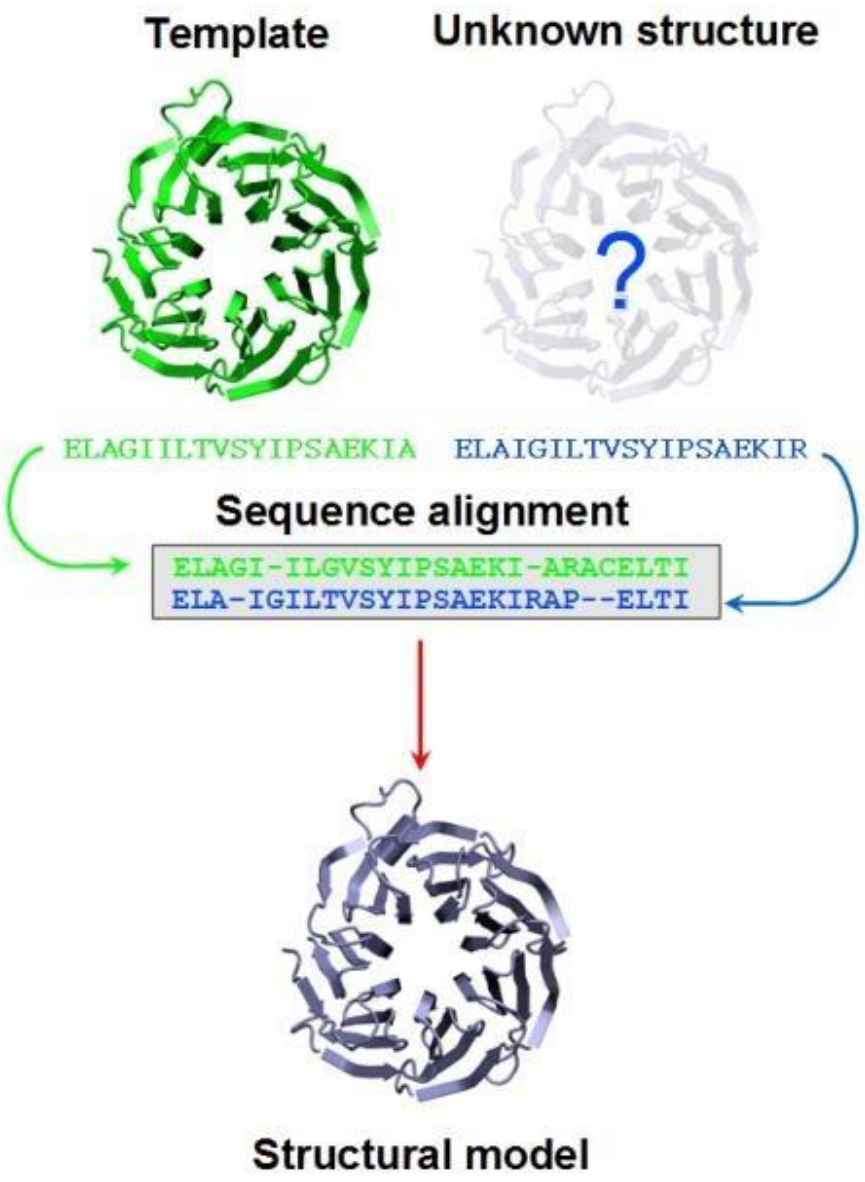
Template-based

- Homology modelling
  - Proteins that have similar sequence will probably have similar structures
  - Find most similar protein with a solved structure – **TEMPLATE**
  - **Sequence-Sequence** alignment
  - **Replace residues in template with those from target**



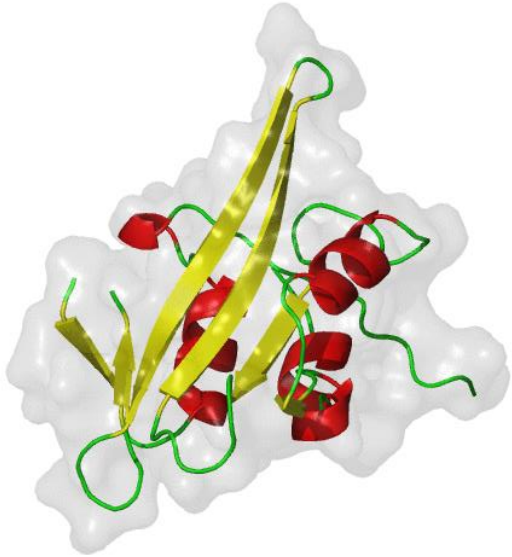
Template-based

- Homology modelling
  - Proteins that have similar sequence will probably have similar structures
  - Find most similar protein with a solved structure – **TEMPLATE**
  - **Sequence-Sequence** alignment
  - **Replace residues in template with those from target**
  - **>30-40% sequence identity rule-of-thumb as minimum requirement**
  - Example – **SWISS-MODEL**



Model Refinement

- Models can be made more accurate
- Final “polish” to get even closer to native state



- Physics based methods e.g. **molecular dynamics**

Forcefield

$$U(R) = \sum_{bonds} k_r (r - r_{eq})^2 \quad \textit{bond}$$

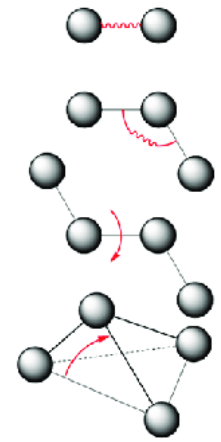
$$+ \sum_{angles} k_\theta (\theta - \theta_{eq})^2 \quad \textit{angle}$$

$$+ \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma]) \quad \textit{dihedral}$$

$$+ \sum_{impropers} k_\omega (\omega - \omega_{eq})^2 \quad \textit{improper}$$

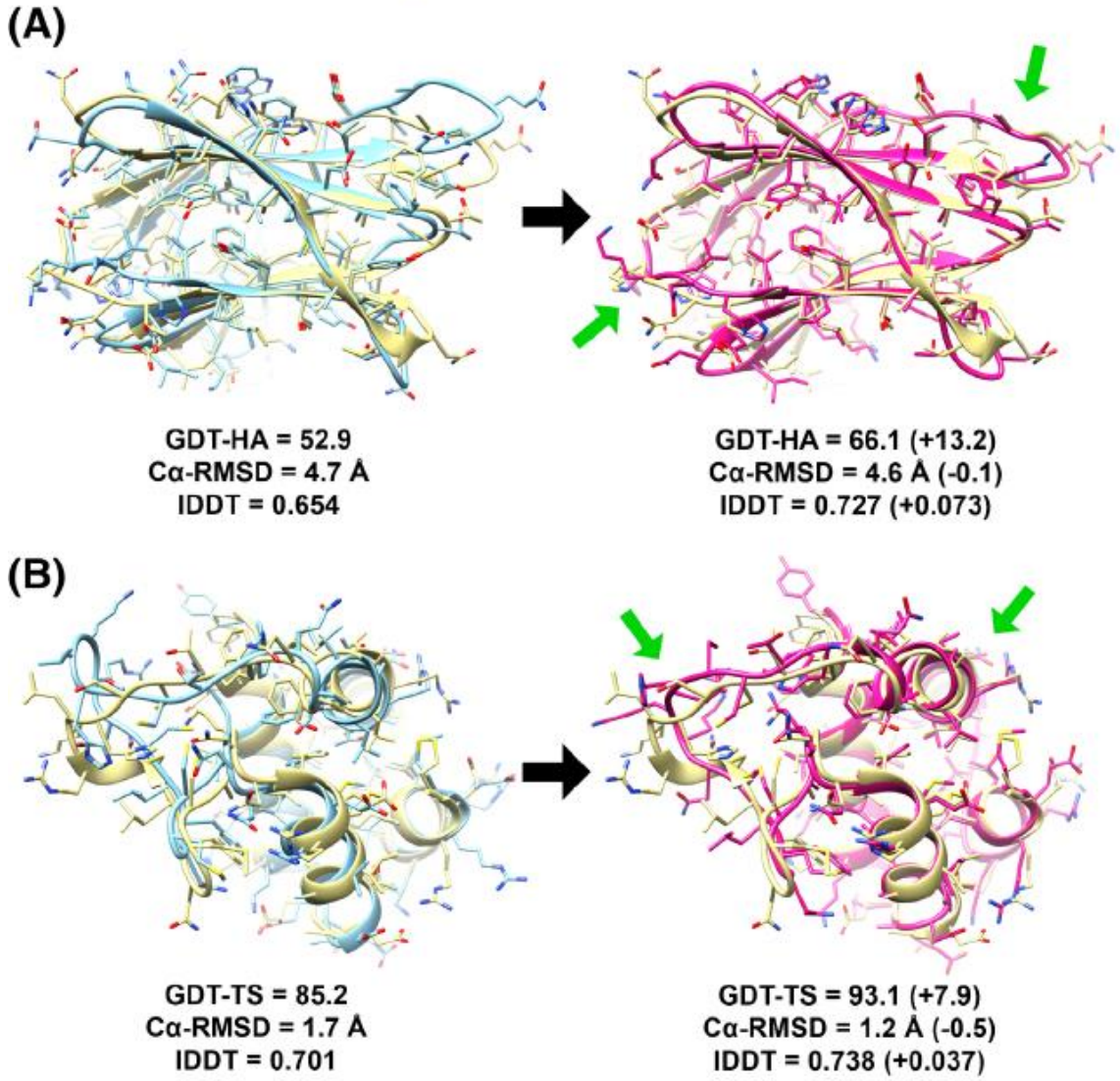
$$+ \sum_{i<j}^{atoms} \varepsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^{6} \right] \quad \textit{van der Waals}$$

$$+ \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad \textit{electrostatic}$$



Model Refinement

- Models can be made more accurate
- Final “polish” to get even closer to native state
- Physics based methods e.g. **molecular dynamics**
- **Example - PREFMD**



Lim Heo, Michael Feig: High-accuracy protein structures by combining machine-learning with physics-based refinement Proteins (2020) 88, 637-642

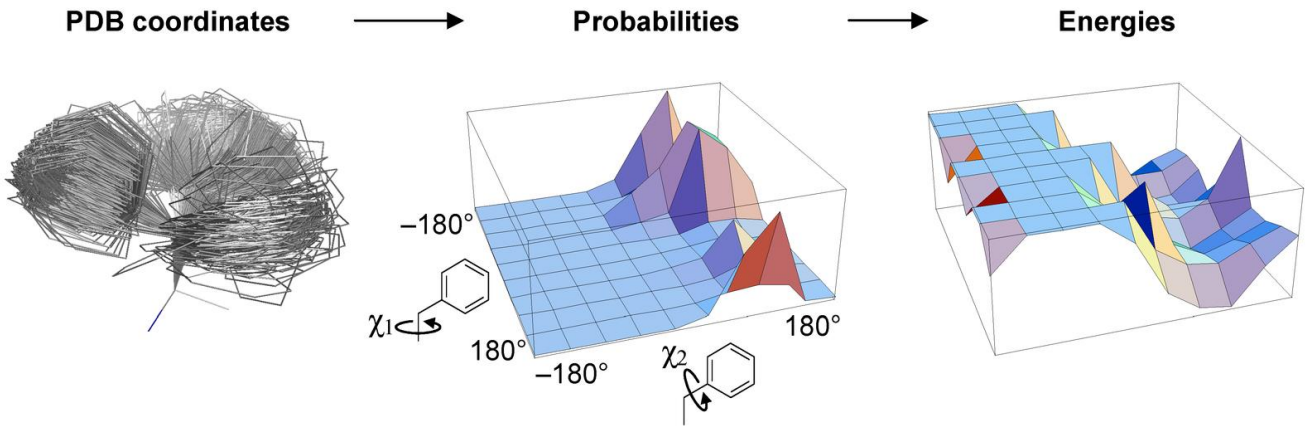
Model Refinement

- Models can be made more accurate
- Final “polish” to get even closer to native state

- **Knowledge-Based/Statistical Potentials**

- **Example - Rosetta**

“…scoring functions derived from an analysis of known protein structures in the Protein Data Bank (PDB).”



Chi-angle probability distribution for a sidechain – More probable states are probably lower energy states.

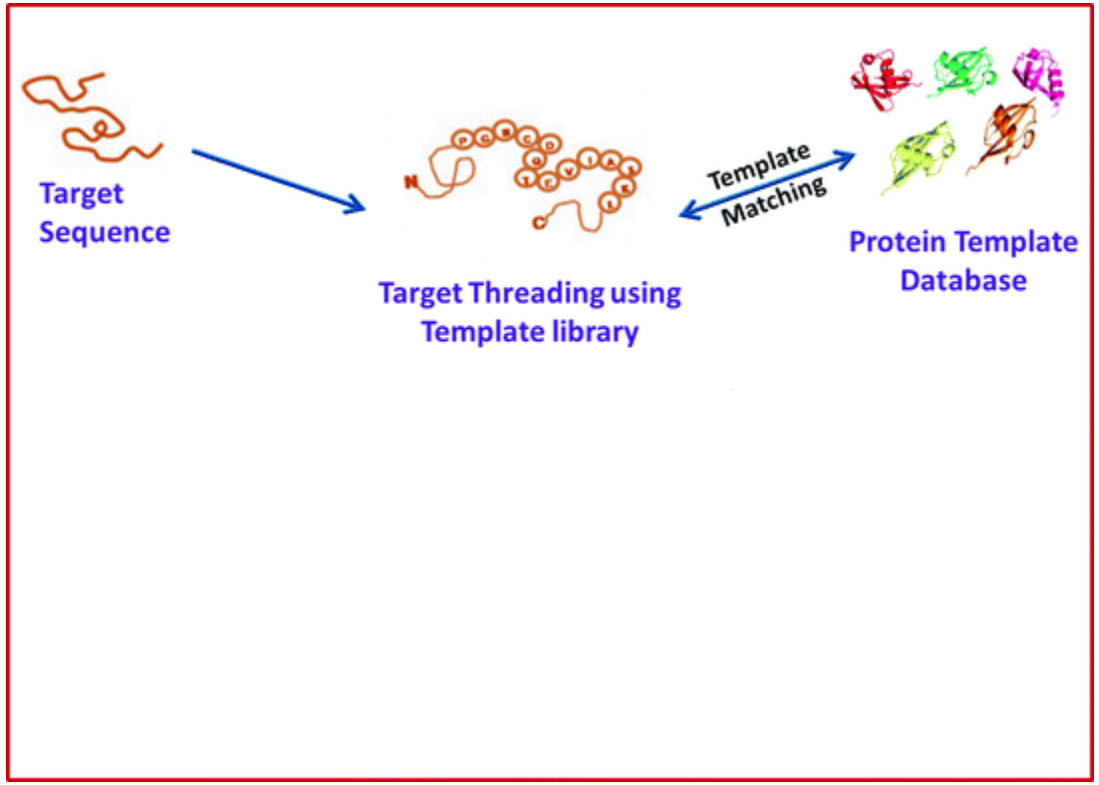
Template-based

- Protein Threading
  - **Sequence-fold alignment**
  - If close homologues don’t exist

Ref: Pulari and Gulati, 2016

Template-based

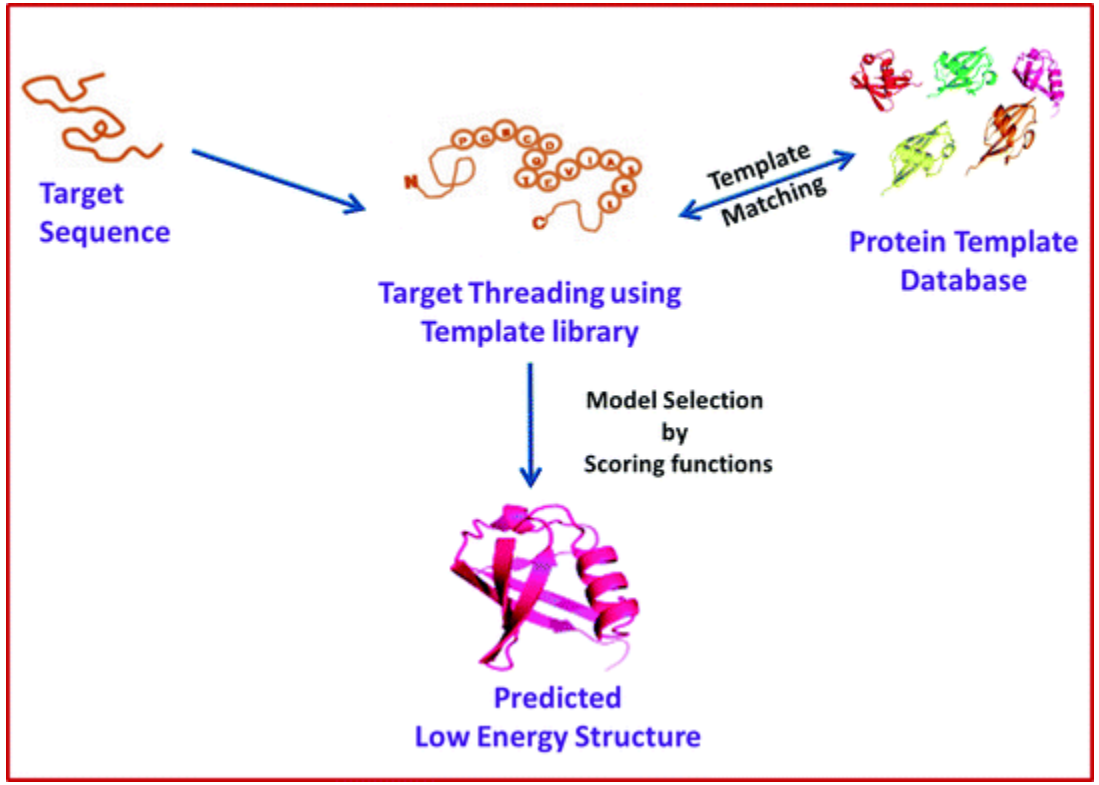
- Protein Threading
  - **Sequence-fold alignment**
  - If close homologues don't exist
  - **"thread"** input sequence onto known protein folds



Ref: Pulari and Gulati, 2016

Template-based

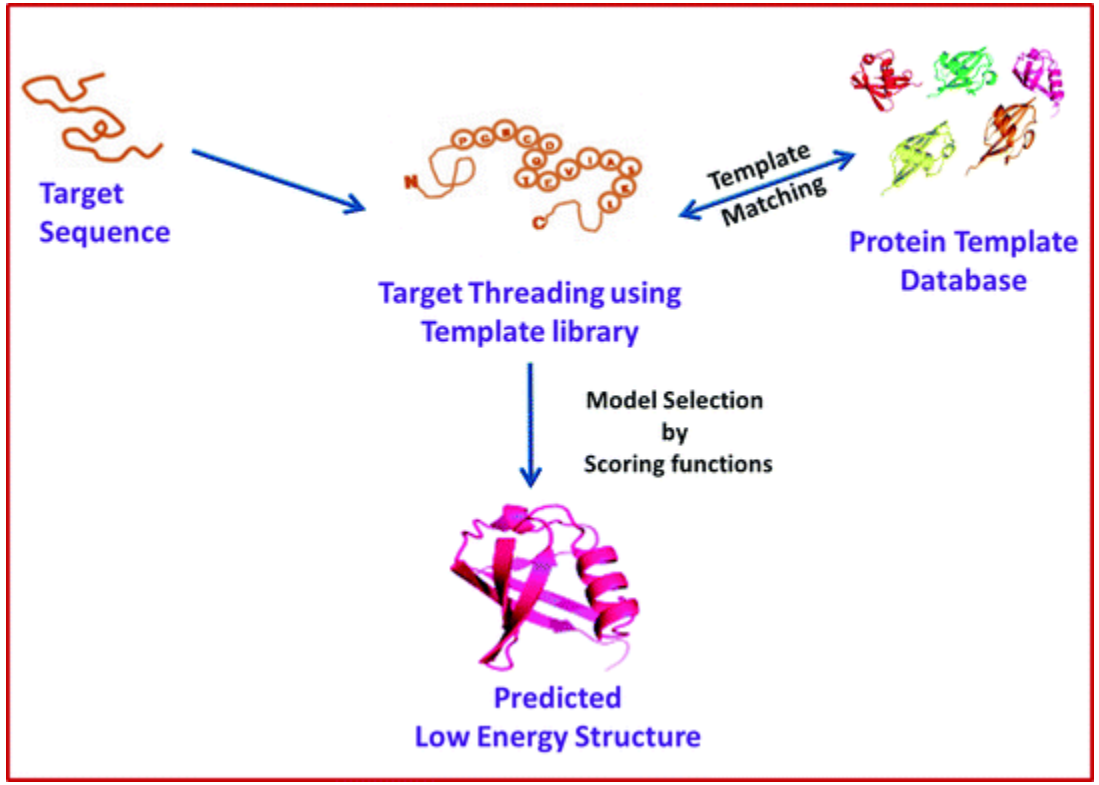
- Protein Threading
  - **Sequence-fold alignment**
  - If close homologues don't exist
  - **"thread"** input sequence onto known protein folds
  - Evaluate resulting models using **statistical or physics based scoring functions**



Ref: Pulari and Gulati, 2016

Template-based

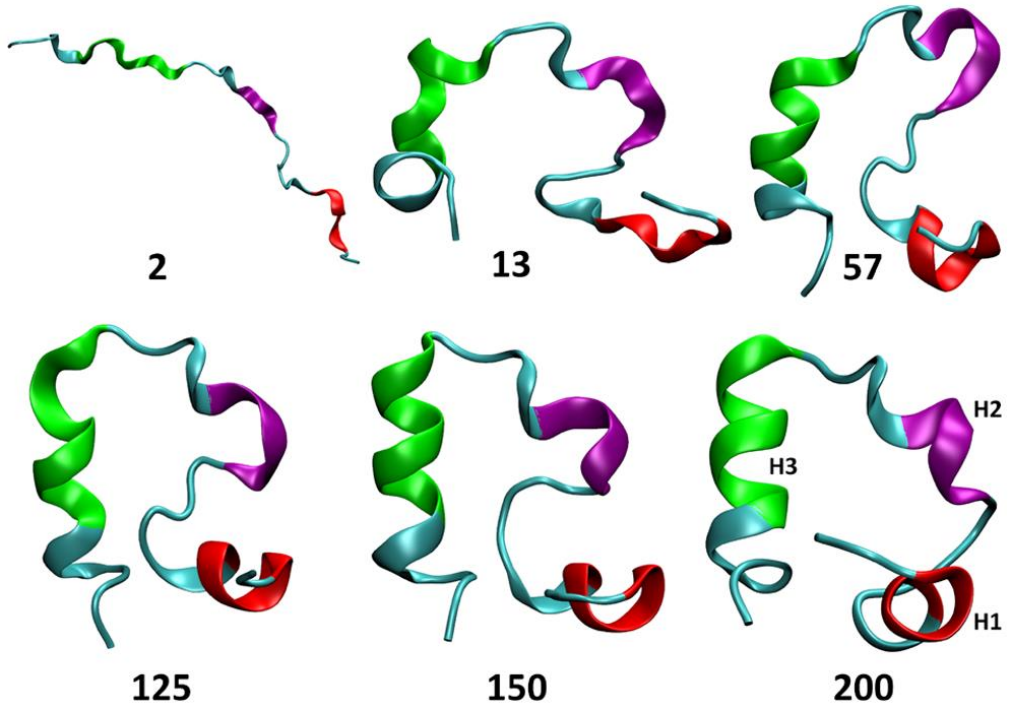
- Protein Threading
  - Works because **fold space is limited**
  - 90% of new structures have a fold that already exists in the PDB
- Example - ITASSER



Ref: Pulari and Gulati, 2016

Template-free

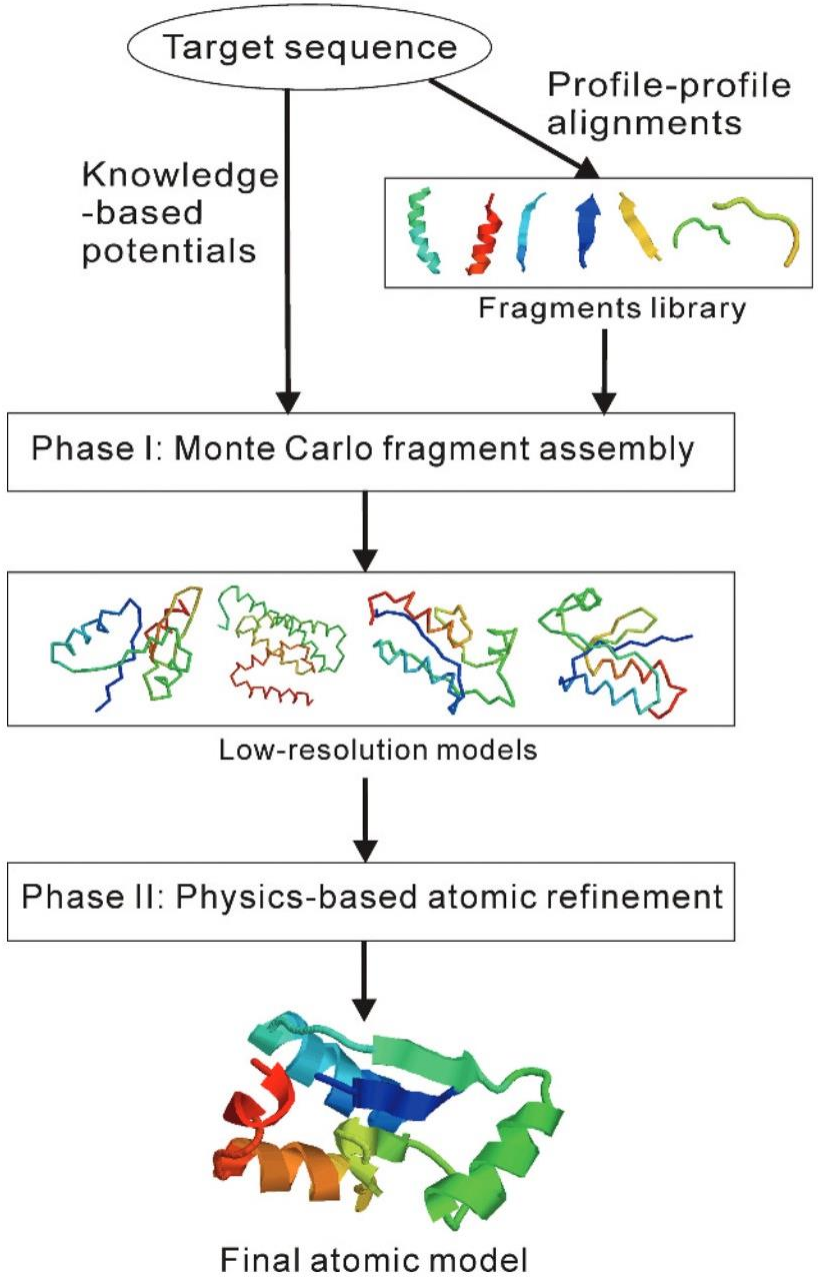
- *Ab Initio* or *“de novo”*
  - When you have no homologous/fold template
  - Physics based methods
    - E.g. Molecular Dynamics – Anton, D.E. Shaw; Folding@home



doi: https://doi.org/10.1371/journal.pone.0029628.g004

Template-free

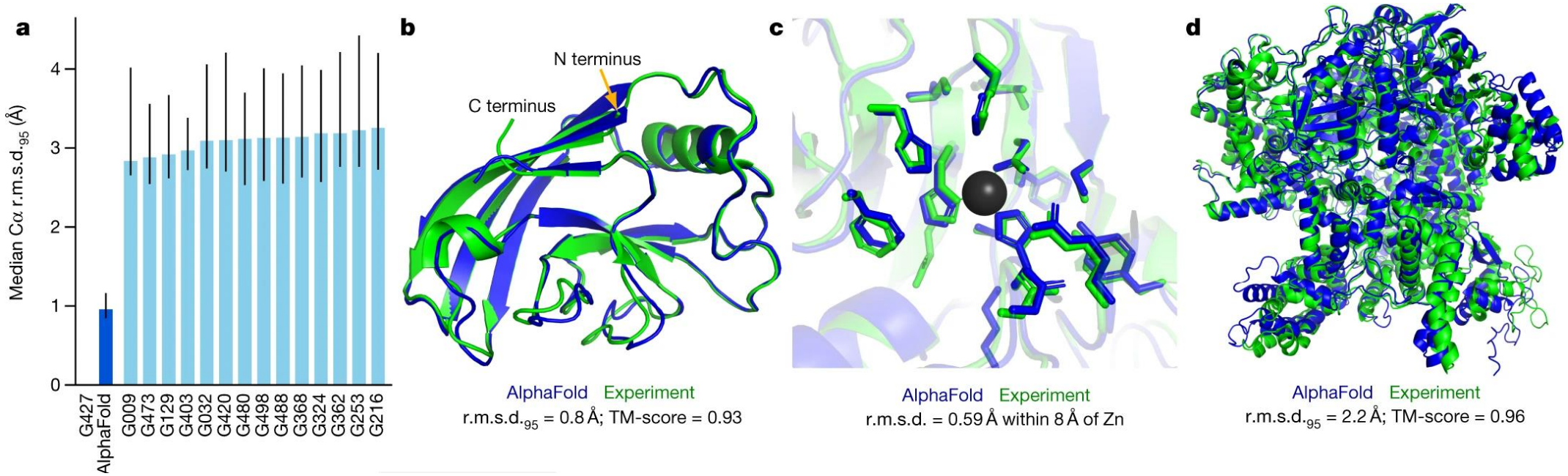
- *Ab Initio* or *"de novo"*
  - When you have no homologous/fold template
  - Physics based methods
    - E.g. Molecular Dynamics – Anton, D.E. Shaw; Folding@home
  - Fragment based methods
    - E.g. Rosetta
  - **Computationally Expensive!**



Lee, Wu and Zhang, 2009

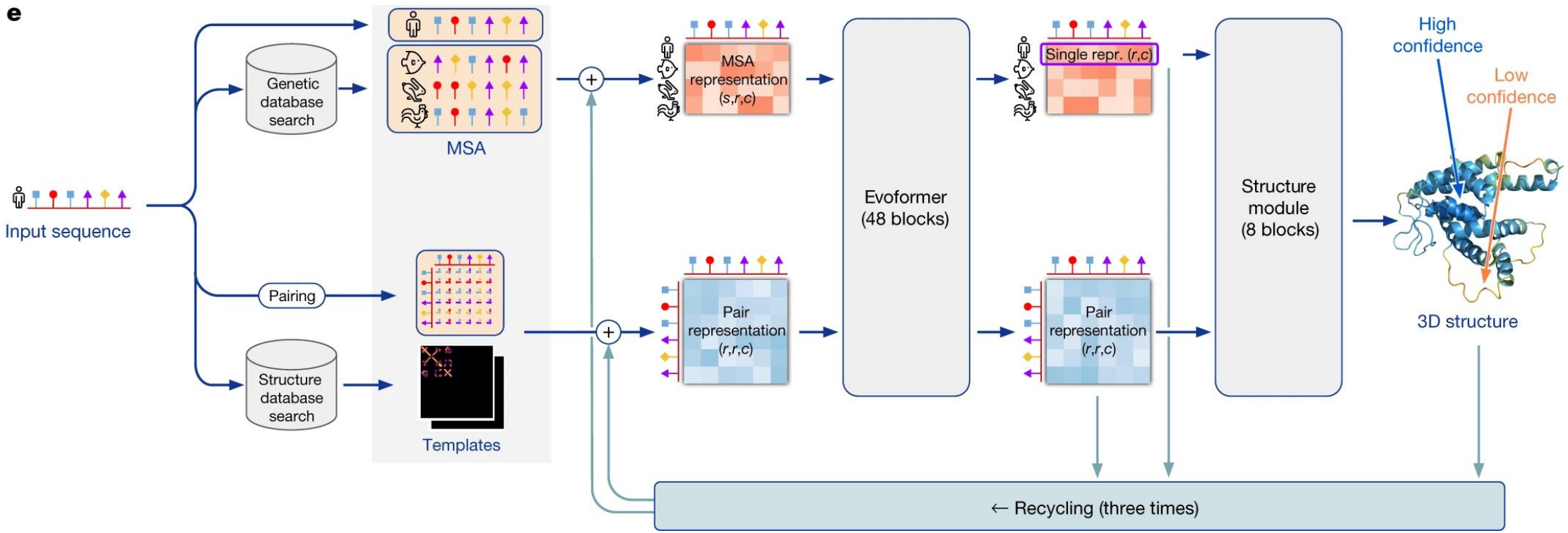
Cutting-Edge

- **AlphaFold2**
- Deepmind



Cutting-Edge

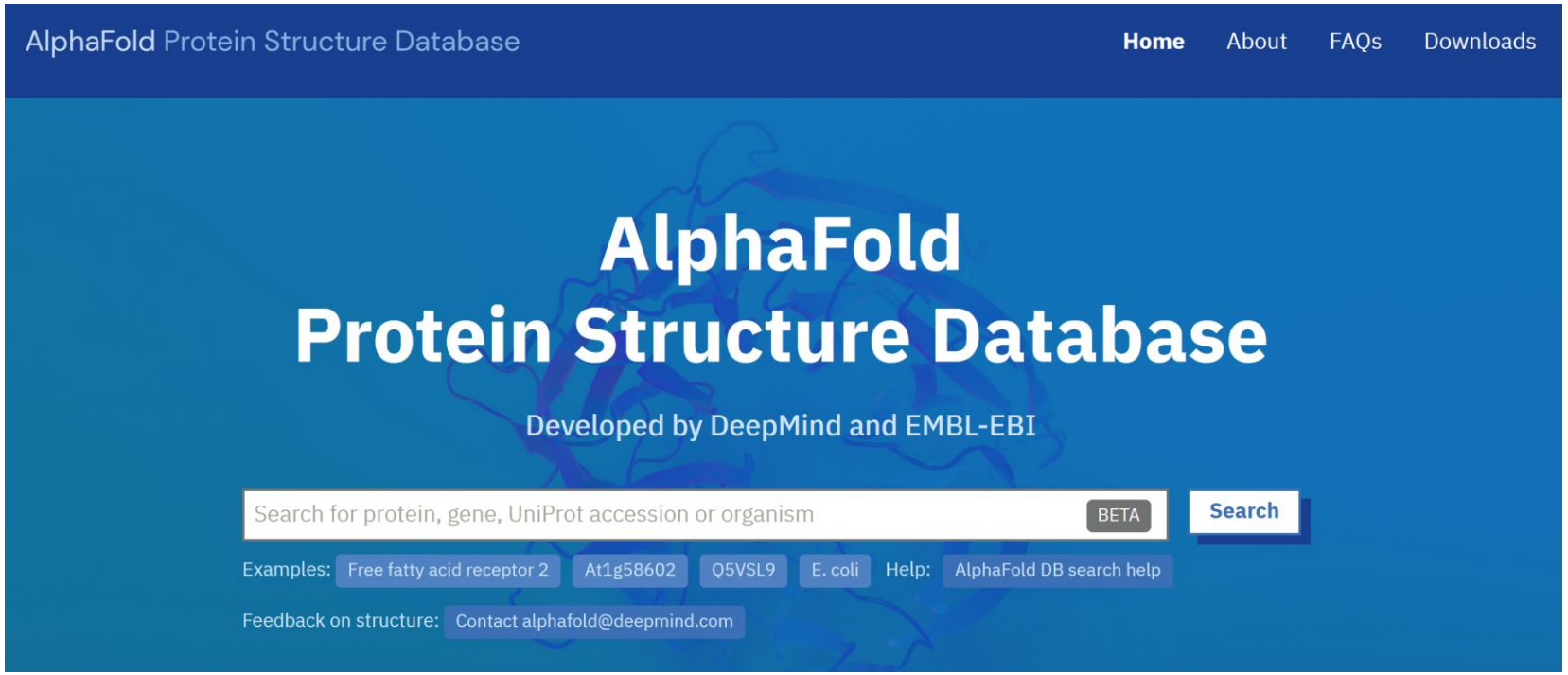
- "**End-to-End** differentiable"
- Can use templates
- New version can predict complexes!



Cutting-Edge

- **AlphaFold2**
- Deepmind

- Database in collaboration with the EBI



https://alphafold.ebi.ac.uk/

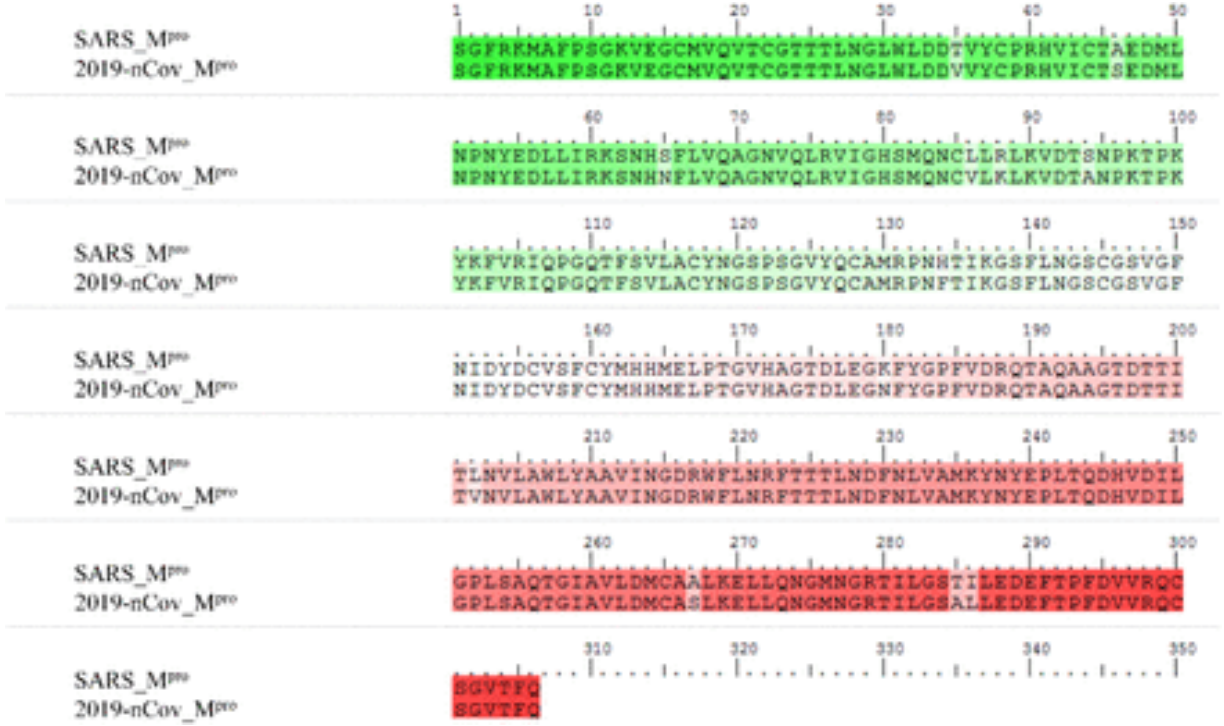
Case study

SARS-CoV-2 main protease

Xu et al. January 2020

- Try and identify drug treatments

High Sequence identity with SARS-CoV-1 proteins



https://www.biorxiv.org/content/10.1101/2020.01.27.921627v1.full

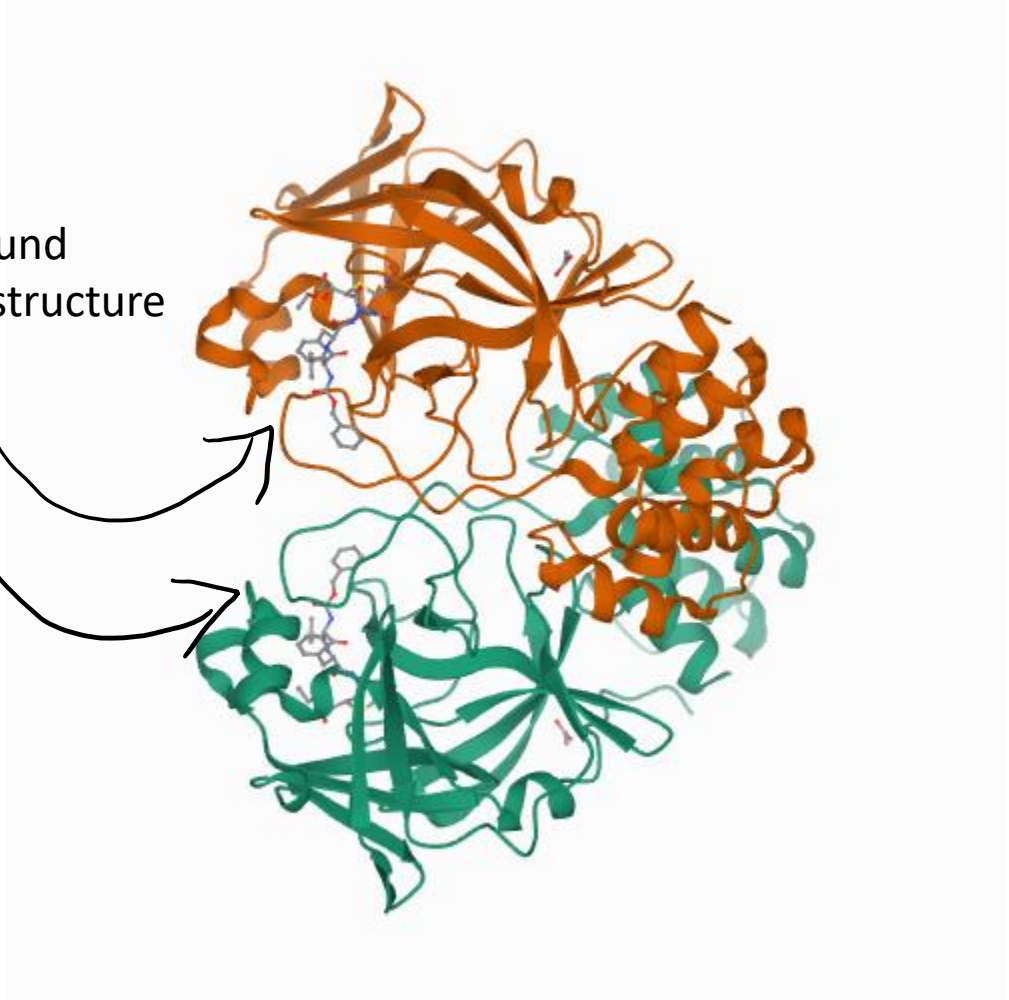
Case study

SARS-CoV-2 main protease

Xu et al. January 2020

- Try and identify drug treatments
- Homology modelling of main protease
  - SWISS-MODEL

2GTB - template



https://www.biorxiv.org/content/10.1101/2020.01.27.921627v1.full

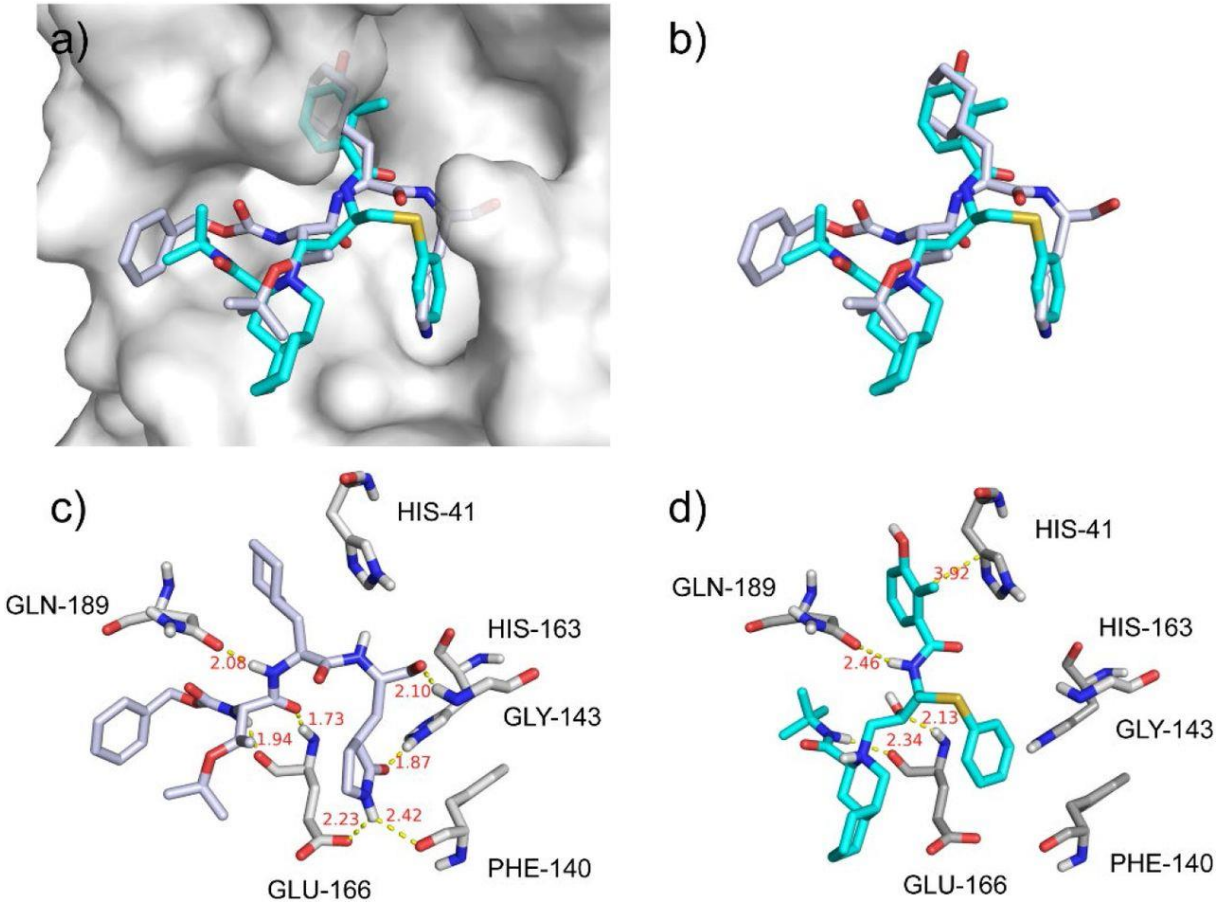
Case study

SARS-CoV-2 main protease

Xu et al. January 2020

- Try and identify drug treatments
- Homology modelling of main protease
  - SWISS-MODEL
- Virtual screening of ~2000 drugs to the models
  - Autodock Vina

Nelfinavir



Original Ligand **Bound Drug**

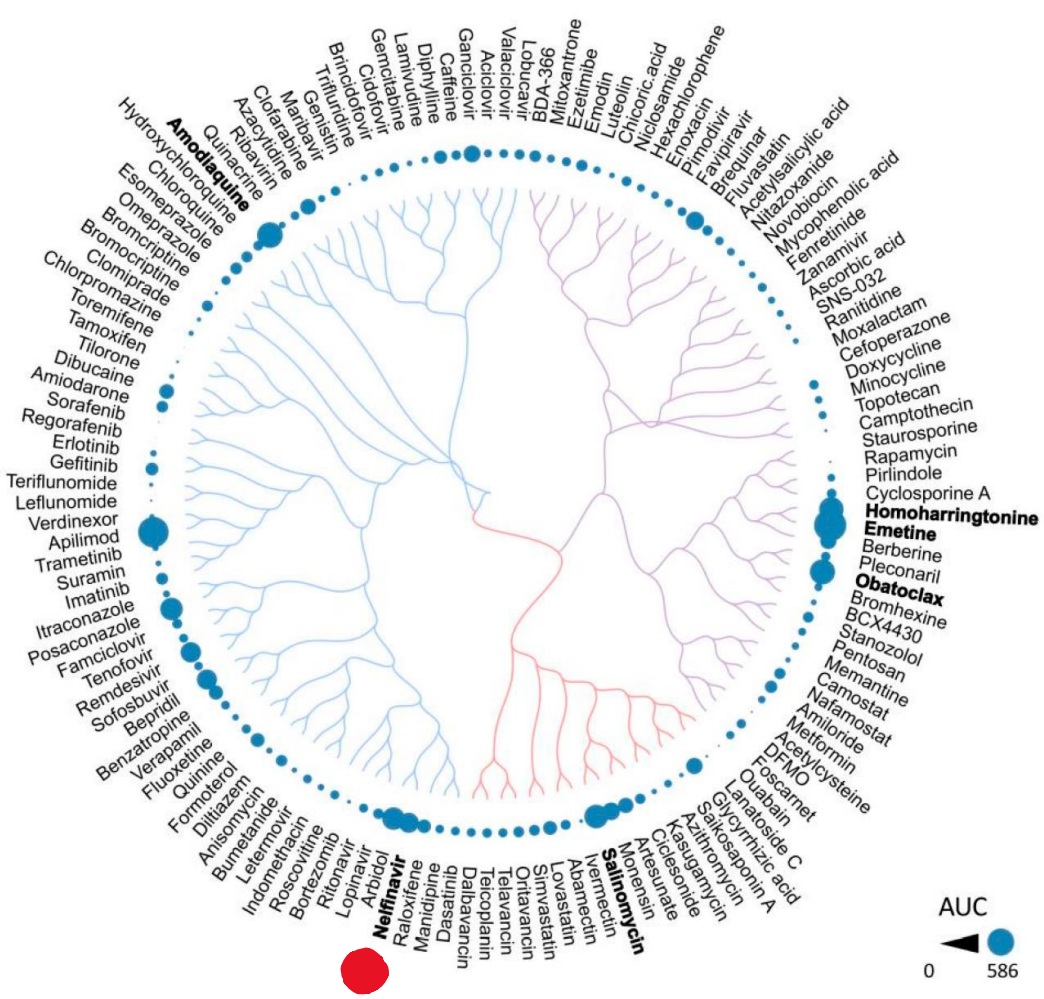
https://www.biorxiv.org/content/10.1101/2020.01.27.921627v1.full

Case study

SARS-CoV-2 main protease

Ianevski et al. June 2020

- Subsequent in vitro studies showed Nelfinavir inhibits viral infection in cell cultures
- As of earlier this year (2021), clinical trials in patients



https://www.mdpi.com/1999-4915/12/6/642#

Summary

- Structural bioinformatics is the analysis and prediction of the 3D structures of biological macromolecules
- We can predict both secondary and tertiary structure from sequence
  - Central dogma – sequence -> structure -> function
  - Evolutionary information key!
- Structure is more conserved than sequence
  - Fold space is limited
- Template-based modelling vs Template-Free modelling
- Cutting edge is Deep Learning – Alphafold 2
- Modelling combined with function prediction e.g. docking, molecular dynamics allows for fast hypothesis generation *in silico*