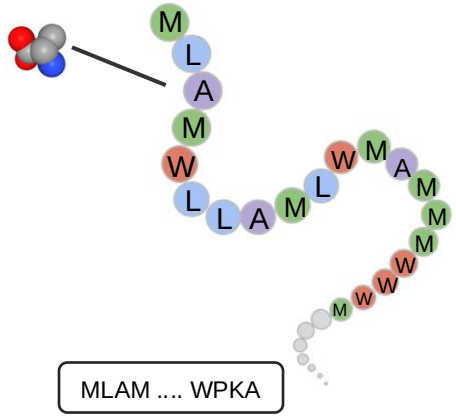
Heinrich Heine
Universität Düsseldorf

What are proteins?

Proteins

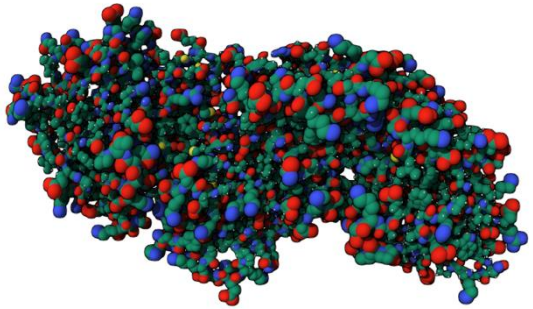
- Responsible for a vast number of functions within all living organism
- Proteins consist of amino acid sequences that fold into 3D structures

20 different amino acids; each amino acid is a small molecule



Folding into

3D structure



Small molecule image: pubchem.ncbi.nlm.nih.gov; Protein image: https://www.rcsb.org/3d-view

Protein folding

Every protein is made up of a sequence of amino acids bonded together

These amino acids interact locally to form shapes like helices and sheets

These shapes fold up on larger scales to form the full three-dimensional protein structure

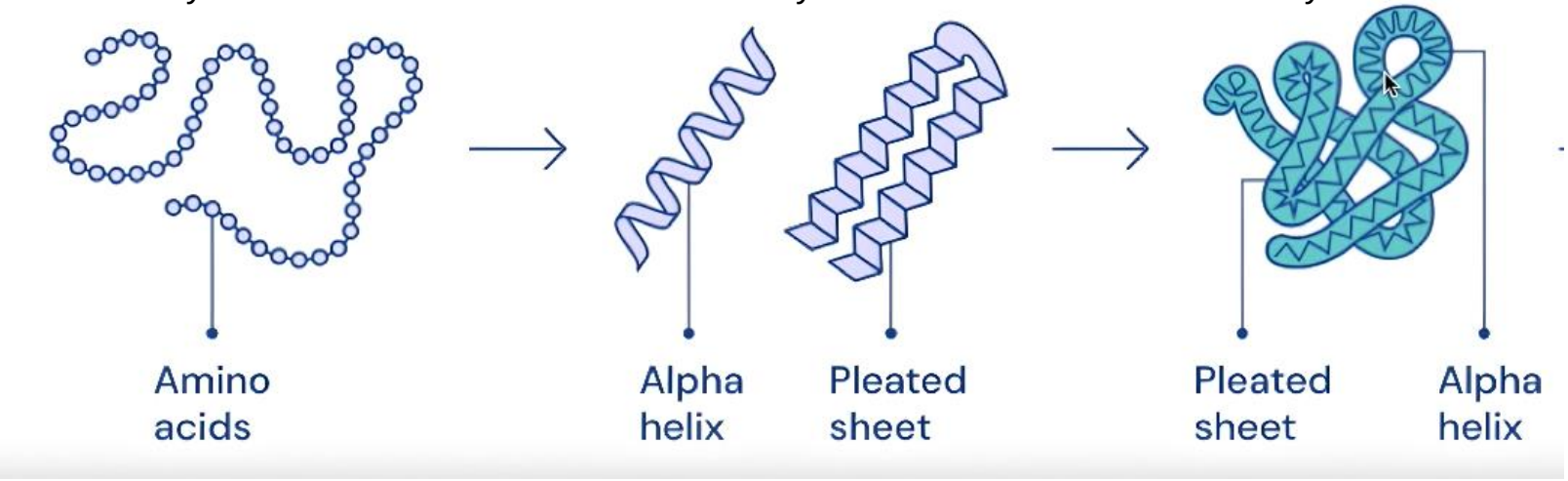
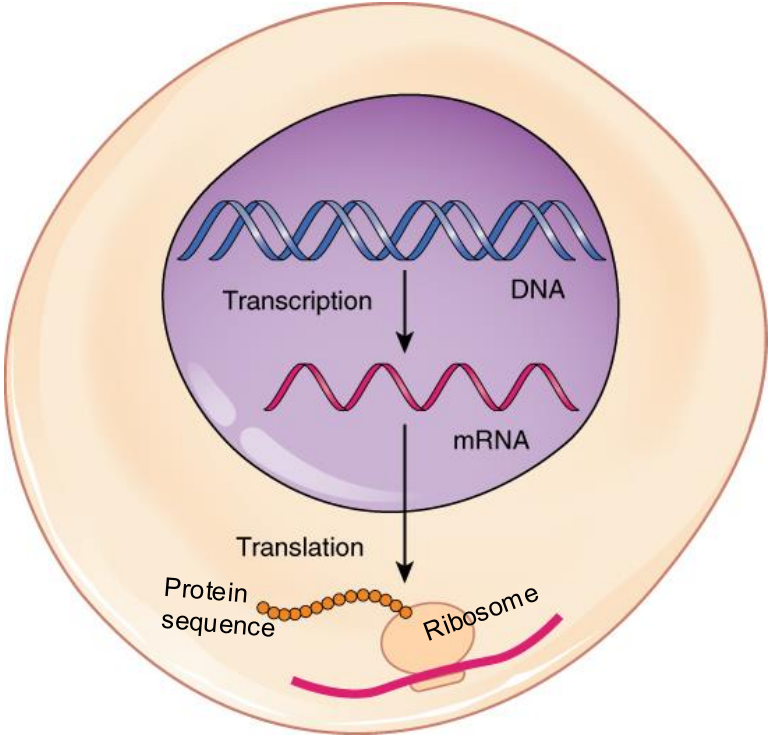


Image copied from: https://youtu.be/z5JZOOQFe0k?si=WrVfmnVR0VKIGuAl

Protein production and folding



- Ribosomes are macromolecules that produce the protein amino acid sequence stored in the genetic code of the cell

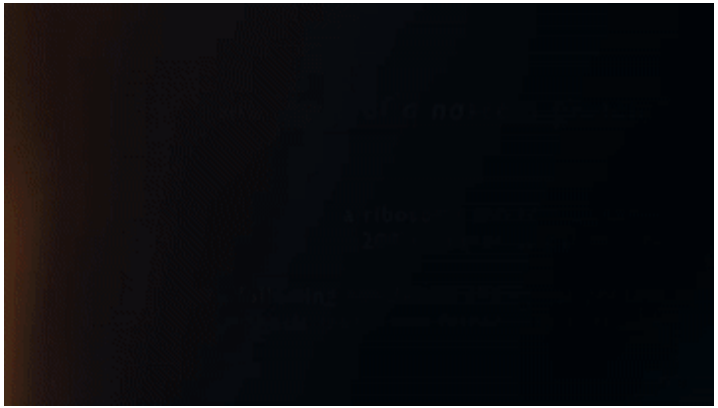
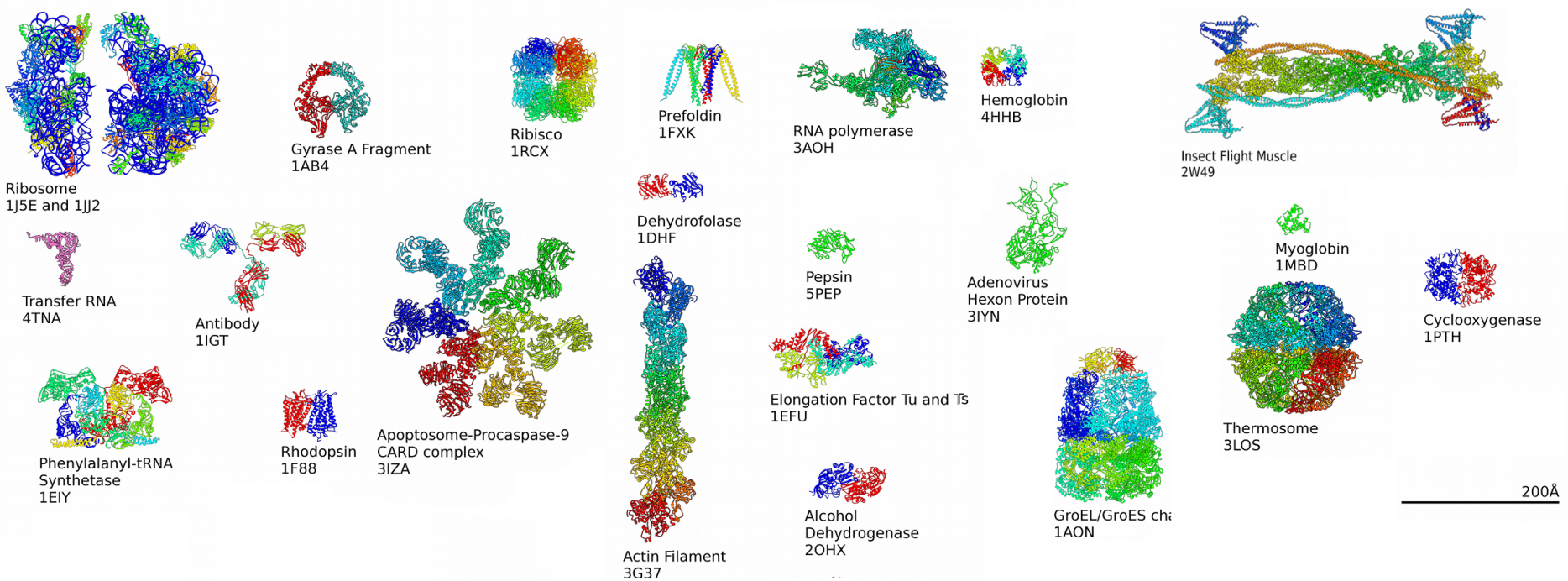


Image: https://courses.lumenlearning.com/suny-biolabs1/chapter/making-proteins/; GIF: https://www.youtube.com/watch?v=2dV5s6v2v8Q (Spronk Studio)

Protein structure space

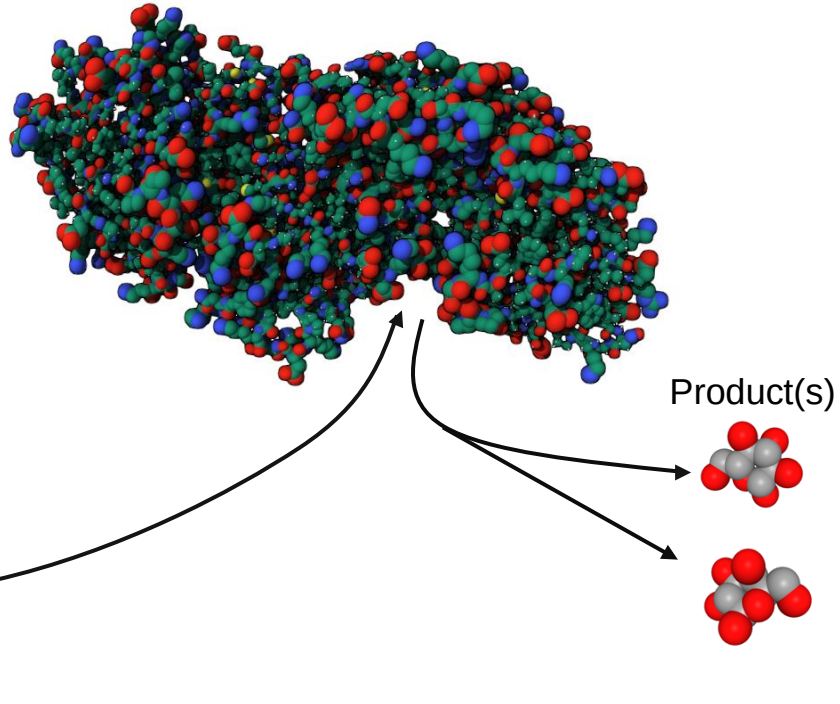


Modified image from: https://commons.wikimedia.org/wiki/File%3AProtein_structure_examples.png

Protein functions – Enzymes

- Enzymes are proteins that catalyze chemical reactions
- Lactase is an enzyme that converts lactose, the milk sugar, into smaller sugars, glucose and galactose

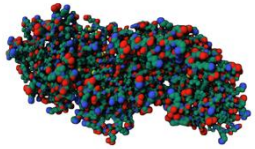
$$Lactose \xrightarrow{\textbf{Lactase}} Glucose + Galactose$$



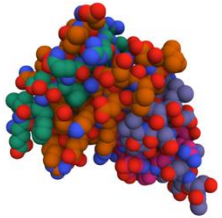
Small molecule images: pubchem.ncbi.nlm.nih.gov; Protein image: https://www.rcsb.org/3d-view

Some main classes of proteins

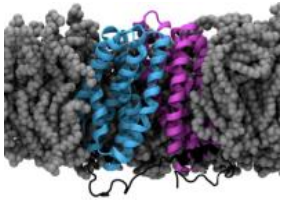
- Enzymes



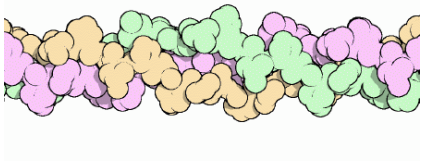
- Regulatory Proteins



- Transport Proteins



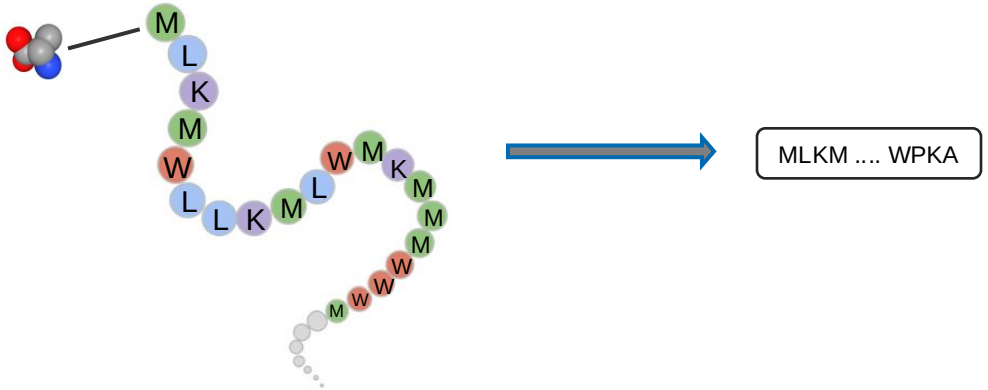
- Structural Proteins



Images copied from: https://philipwfowler.wordpress.com/; https://www.rcsb.org/3d-view; https://pdb101.rcsb.org/motm/4

Sequence representations of proteins

- Proteins can be represented through their amino acid sequence
- The amino acid sequence is readily available for most proteins (UniProt.org)



Alanine - A
Arginine - R
Asparagine - N
Aspartic acid - D
Cysteine - C
Glutamine - Q
Glutamic acid - E
Glycine - G
Histidine - H
Isoleucine - I
Leucine - L
Lysine - K
Methionine - M
Phenylalanine - F
Proline - P
Serine - S
Threonine - T
Tryptophan - W
Tyrosine - Y
Valine - V

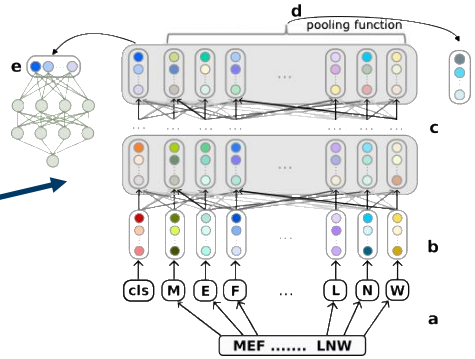
Small molecule image: pubchem.ncbi.nlm.nih.gov

FASTA files

- Protein amino acid sequences are typically stores in FASTA files
  - FASTA format is a text-based format
  - An entry begins with a greater-than character (">") followed by a description of the sequence (the same line)
  - Following lines contain protein sequence
- Example:

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

Transformer Network



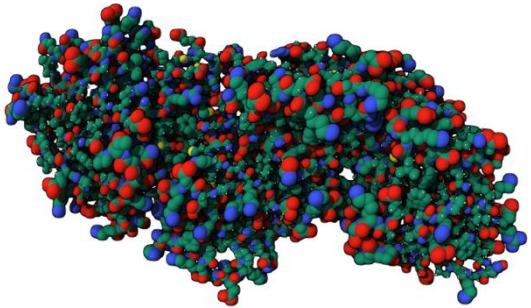
Sequence properties and statistics

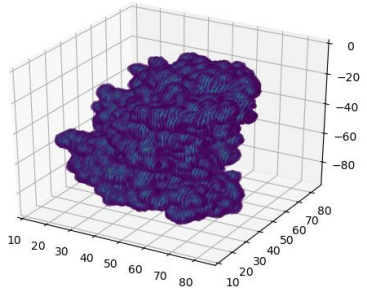
Example copied from: https://en.wikipedia.org/wiki/FASTA_format

3D-Representations of proteins

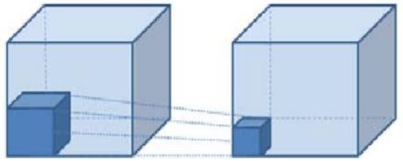
- Representing protein 3D structures: The PDB (Protein Data Bank) format
  - is a text file
  - includes spatial coordinates for each atom in the molecule

```
ATOM      1  N   VAL A   1      19.323  29.727  42.781
ATOM      2  CA  VAL A   1      20.141  30.469  42.414
ATOM      3  C   VAL A   1      21.664  29.857  42.548
ATOM      4  O   VAL A   1      21.985  29.541  43.704
ATOM      5  CB  VAL A   1      19.887  31.918  43.524
ATOM      6  CG1 VAL A   1      20.656  32.850  42.999
ATOM      7  CG2 VAL A   1      18.692  31.583  43.506
                    ...
```





3D Convolutional Neural Networks



Protein image: https://www.rcsb.org/3d-view; 3D conv image copied from: https://paperswithcode.com/method/3d-convolution