NEAR EAST UNIVERSITY
EXPERIMENTAL HEALTH
SCIENCES RESEARCH CENTER

A Computational Path to Functional Analysis:
Structural Modeling, Ligand Binding, and Molecular Mechanics

Kerem Teralı, MRes (St And), PhD (Lond)

Department of Medical Biochemistry
Faculty of Medicine
Near East University

✉ kerem.terali@neu.edu.tr

Introduction

- Non-synonymous single nucleotide polymorphisms (nsSNPs) in human genes can result in phenotypes where the pathobiological basis may not be clear due to the lack of mutant protein structures.

- One of the unifying themes in protein science is that function correlates more highly with structure than with sequence.

- Protein structures can therefore be considered as **molecular phenotypes** potentially linking genetic variation to human disease.

Introduction (cont'd)

- Experimental investigation of the structure and function of missense mutants is a time-consuming and cost-intensive task.

- Such an investigation may well be facilitated by the employment of a diverse array of *in silico* tools which:

    i. allow for the modeling of amino acid substitutions in wild-type proteins;

    ii. help analyze the interactions of the mutant proteins with their binding partners including metal ions, small-molecule ligands, and other proteins.

Amino Acid Sequences

Where to Find Them?

The UniProt Knowledgebase

- **The UniProt Knowledgebase** is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.

- It is a part of UniProt, a collaboration between **the European Bioinformatics Institute**, **the Swiss Institute of Bioinformatics**, and **the Protein Information Resource**.

- The UniProtKB consists of two sections:

    i. reviewed (Swiss-Prot) – manually annotated records (*i.e.* records with information extracted from literature and curator-evaluated computational analysis);

    ii. unreviewed (TrEMBL) – computationally analyzed records (*i.e.* records that await full manual annotation).

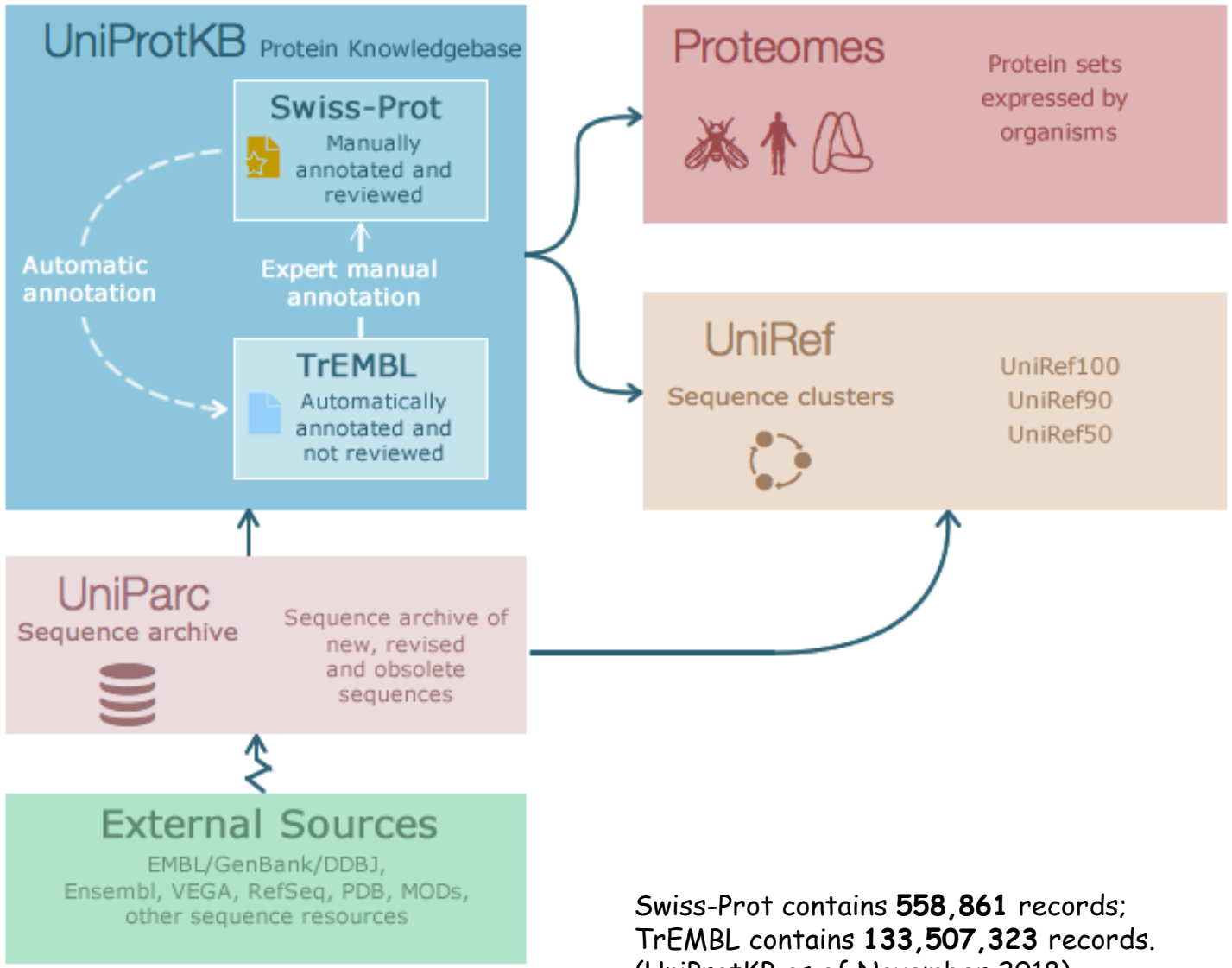
UniProtKB Data

Core Data

- ✓ Amino acid sequence
- ✓ Protein name or description
- ✓ Taxonomic data
- ✓ Citation information

Others

- ✓ Function
- ✓ Subcellular location
- ✓ Involvement in disease
- ✓ PTMs / Processing
- ✓ Expression
- ✓ Protein–protein interactions
- ✓ Structure
- ✓ Family and domains



Swiss-Prot contains **558,861** records;
TrEMBL contains **133,507,323** records.
(UniProtKB as of November 2018)

Atomic Coordinates

How to Download Them?

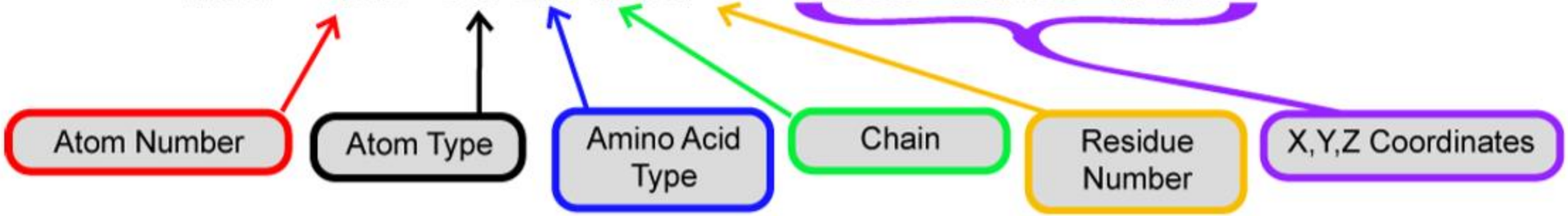
The Protein Data Bank

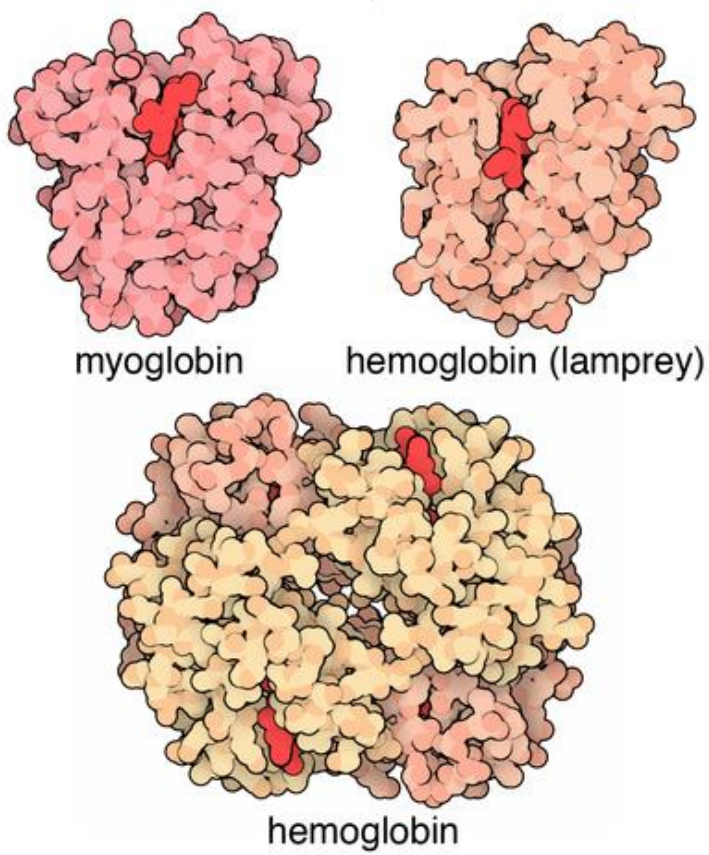
- Structural biology was born in **1958** with **John Kendrew**'s atomic structure of **myoglobin**, and in the following decade, the field grew rapidly.

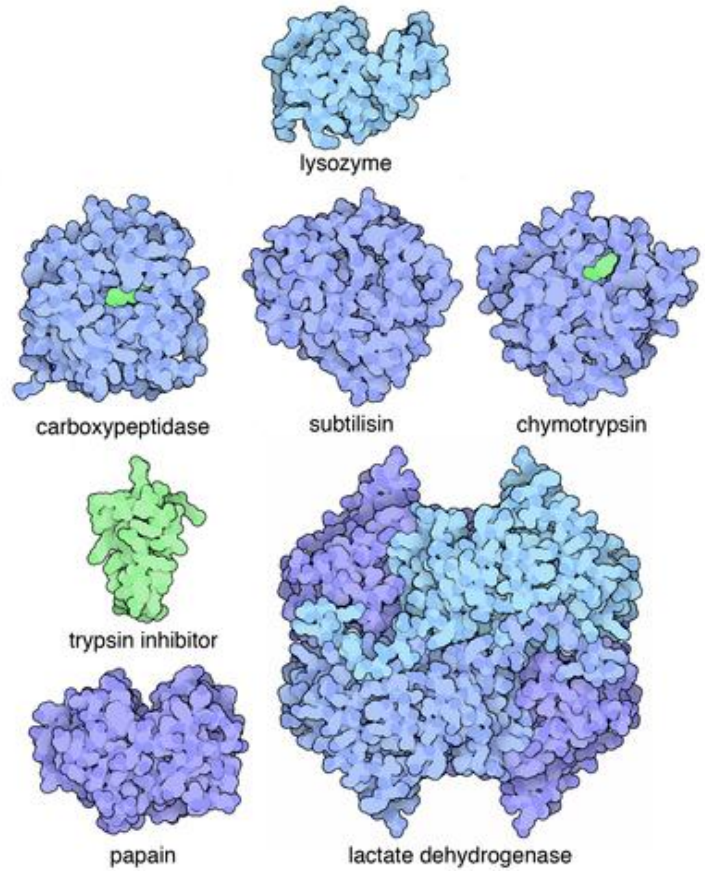
- By the early 1970's, there were **a dozen** atomic structures of proteins, and researchers were discovering that they had a goldmine of information.

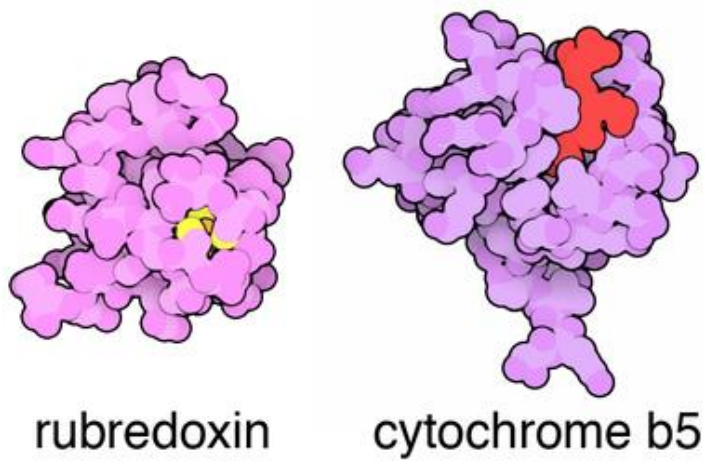
- However, the coordinate files for these structures are quite large, and in the days before the internet, it was difficult for individual researchers to share these large files with the growing number of interested structural biologists around the world.

```
ATOM   1132  NH1 ARG A 149      31.814 -31.597  16.995
ATOM   1133  NH2 ARG A 149      32.203 -32.934  18.816
ATOM   1134  N   ASN A 150      29.346 -24.359  18.812
ATOM   1135  CA  ASN A 150      28.480 -23.190  18.933
ATOM   1136  C   ASN A 150      28.606 -22.168  17.808
ATOM   1137  O   ASN A 150      27.803 -21.276  17.678
ATOM   1138  CB  ASN A 150      28.732 -22.524  20.282
ATOM   1139  CG  ASN A 150      28.284 -23.389  21.447
ATOM   1140  OD1 ASN A 150      27.205 -23.981  21.430
ATOM   1141  ND2 ASN A 150      29.110 -23.463  22.466
ATOM   1142  N   LEU A 151      29.629 -22.313  16.996
ATOM   1143  CA  LEU A 151      29.868 -21.415  15.894
ATOM   1144  C   LEU A 151      29.953 -22.205  14.597
ATOM   1145  O   LEU A 151      30.149 -23.422  14.614
ATOM   1146  CB  LEU A 151      31.208 -20.735  16.100
ATOM   1147  CG  LEU A 151      31.436 -19.884  17.337
ATOM   1148  CD1 LEU A 151      32.846 -19.333  17.256
```









Carrying oxygen

Enzyme active sites

Electron transport

The Protein Data Bank (cont'd)

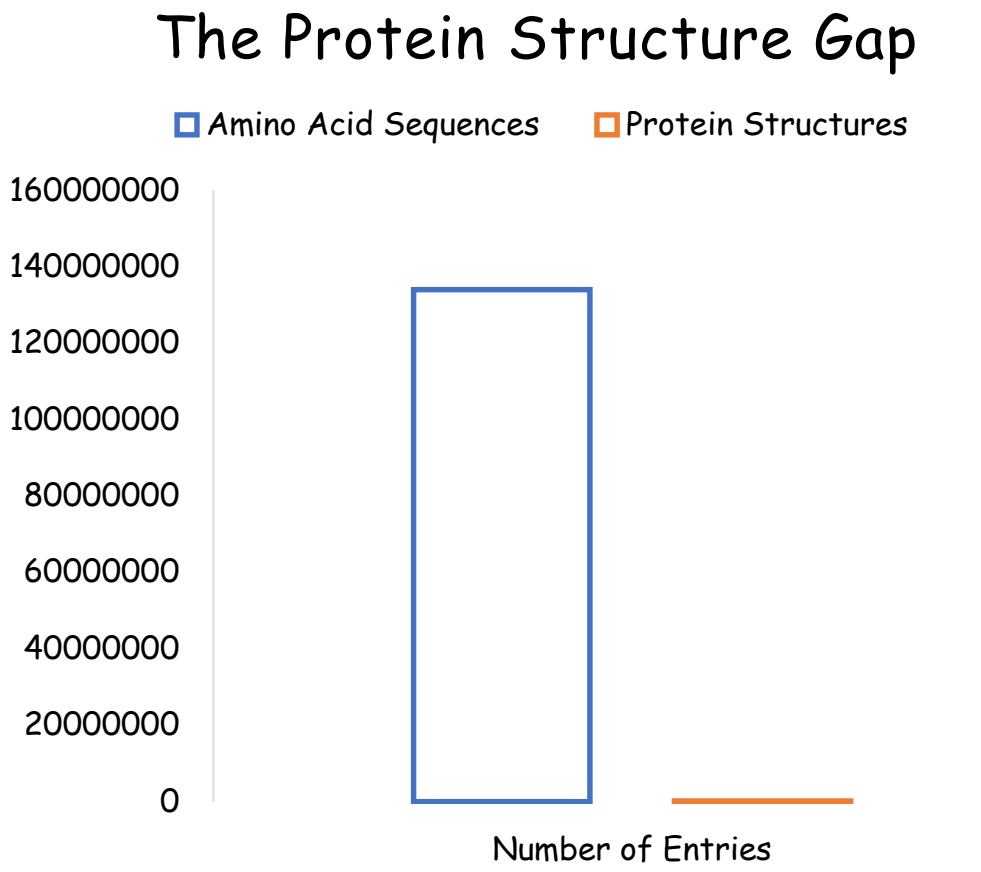
- **The Protein Data Bank** archive was created to solve this problem.
- Depositors would send their coordinates to the PDB, who would then mail them to interested users.
- In 1971, the PDB was jointly operated at **Brookhaven** and **the Cambridge Crystallographic Data Centre**.
- Nowadays, structures and experimental data are deposited at and processed by the **Worldwide PDB** (**wwPDB**) partner sites in America (RCSB PDB; http://rcsb.org), Europe (PDBe; http://pdbe.org), and Japan (PDBj; http://pdbj.org).

PDB Statistics

- In November 2018, the PDB holdings amounted to 146,093 structures.
- Of these, ~135,000 were protein structures solved primarily using **X-ray diffraction**, **nuclear magnetic resonance** (**NMR**) **spectroscopy**, and **electron microscopy**.
- The rest were mostly structures of nucleic acids and nucleic acid–protein complexes, and of some carbohydrates.
- It is worth mentioning that the total figure of ~135,000 protein structures mentioned above includes a considerable amount of **redundancy**.

Experimental Method	Proteins	Nucleic Acids	NA–Protein Complex	Other	Total
X-Ray	122,429	1,963	6,341	10	130,743
NMR	10,903	1,263	253	8	12,427
Electron Microscopy	1,841	31	657	0	2,529
Other	244	4	6	13	267
Multi-Method	119	5	2	1	127
Total	135,536	3,266	7,259	32	146,093

(RCSB PDB as of November 2018)


The Protein Structure Gap
Amino Acid Sequences
Protein Structures
160000000
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
Number of Entries


MIND THE GAP

Biomolecular Visualization Systems

- **Biomolecular visualization** deals with the graphical depiction of the structures of biomolecules and biomolecular complexes.

- It supports our understanding of the properties and interactions of elementary biological functional units that occur in cells.

- It also supports the «rational» design of new molecules such as pharmaceutically active compounds or customized biomolecules with specific properties.

- The types of visualization can be divided into showing **static geometry** or depicting **an animation**.

Biomolecular Visualization Systems (cont'd)

- In structural biology, many 3D models showing different attributes of the depicted biomolecule have been developed.

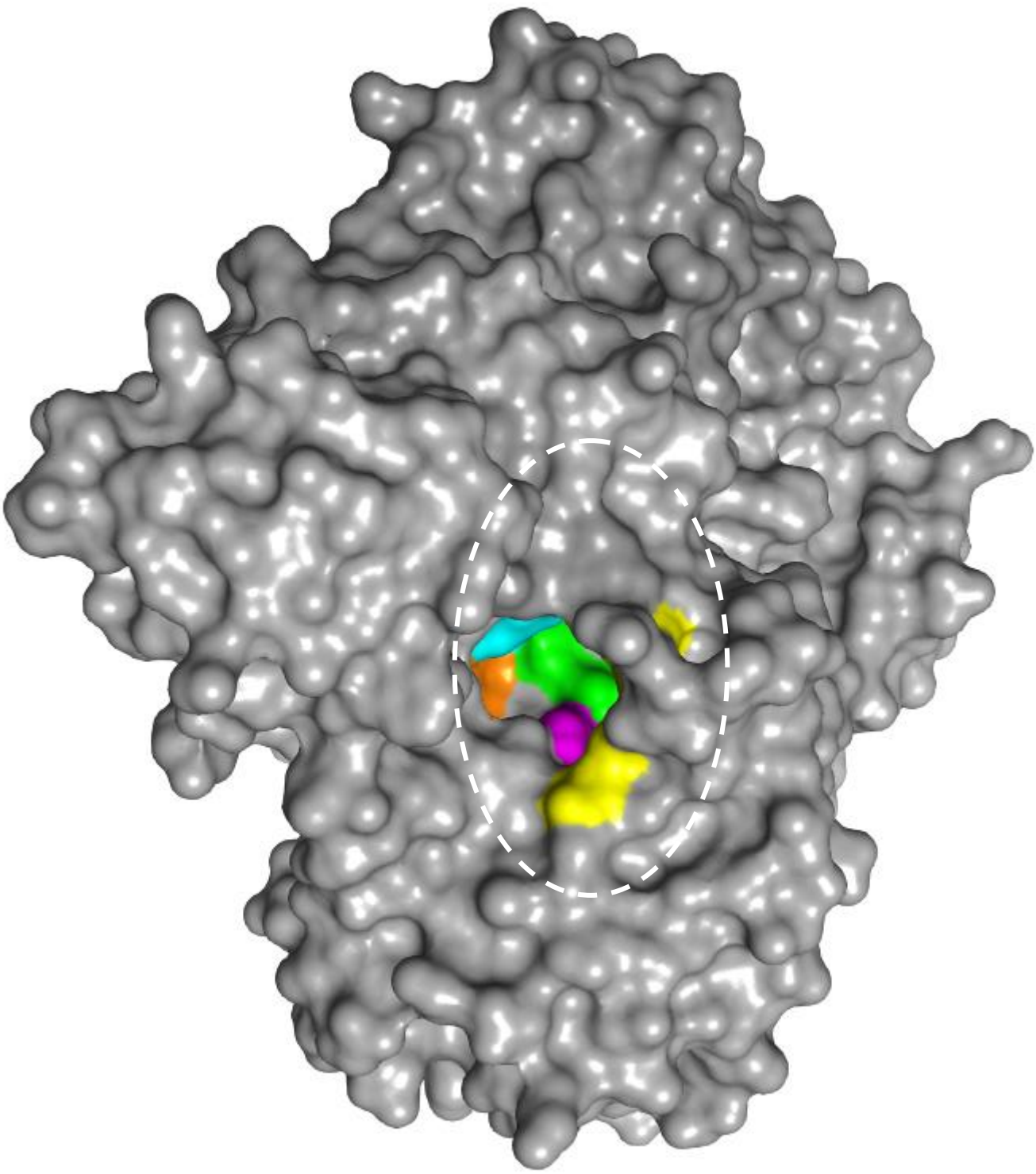
- The choice of representation depends purely on the intended analysis.

- 3D models can be classified into two categories:

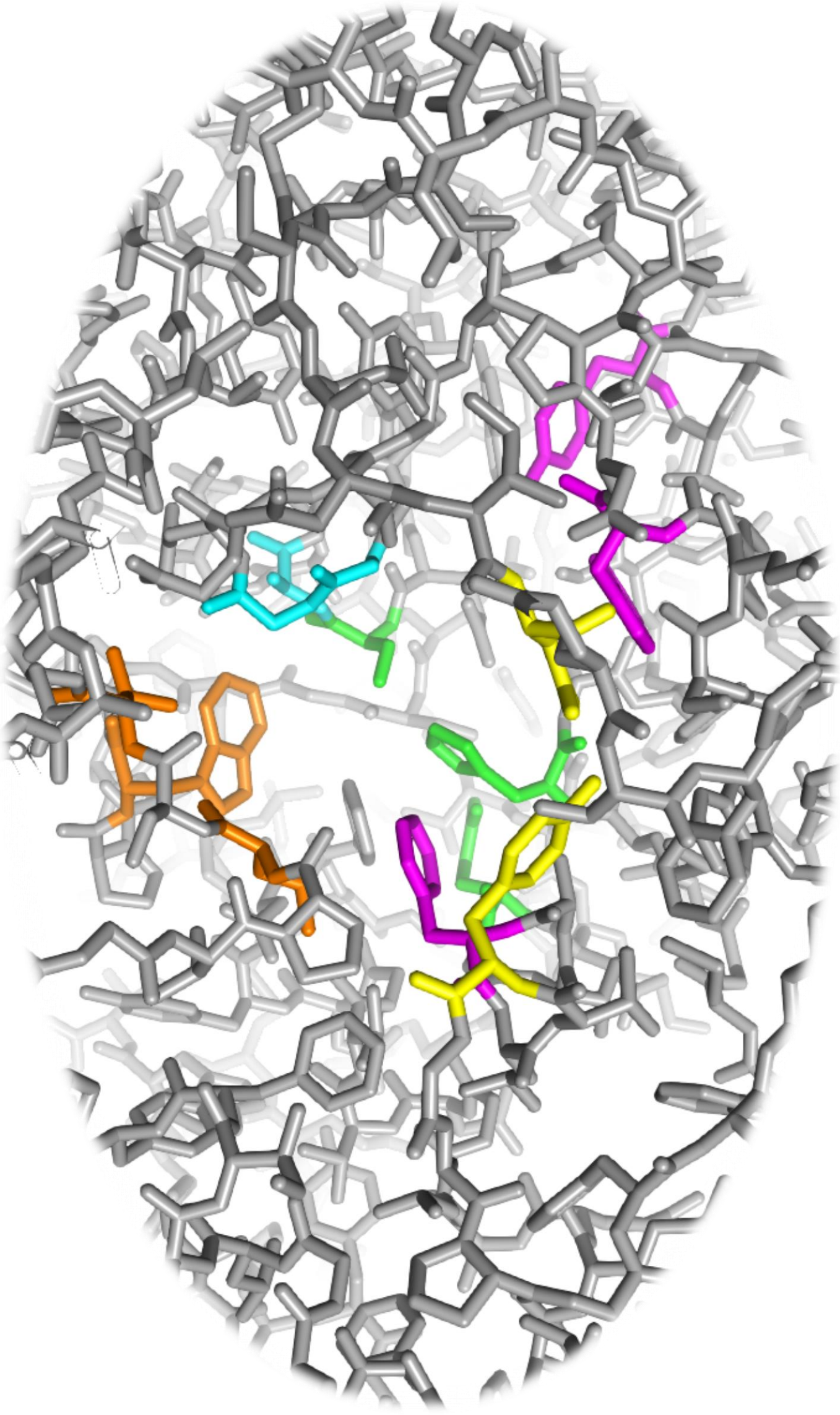
    i. atomistic – models that directly depict the atoms of a biomolecule (*e.g.* bond-centric models and solvent-excluded surface models);

    ii. abstract – models that illustrate a biomolecule's overall shape or a special feature of it (*e.g.* cartoon models).

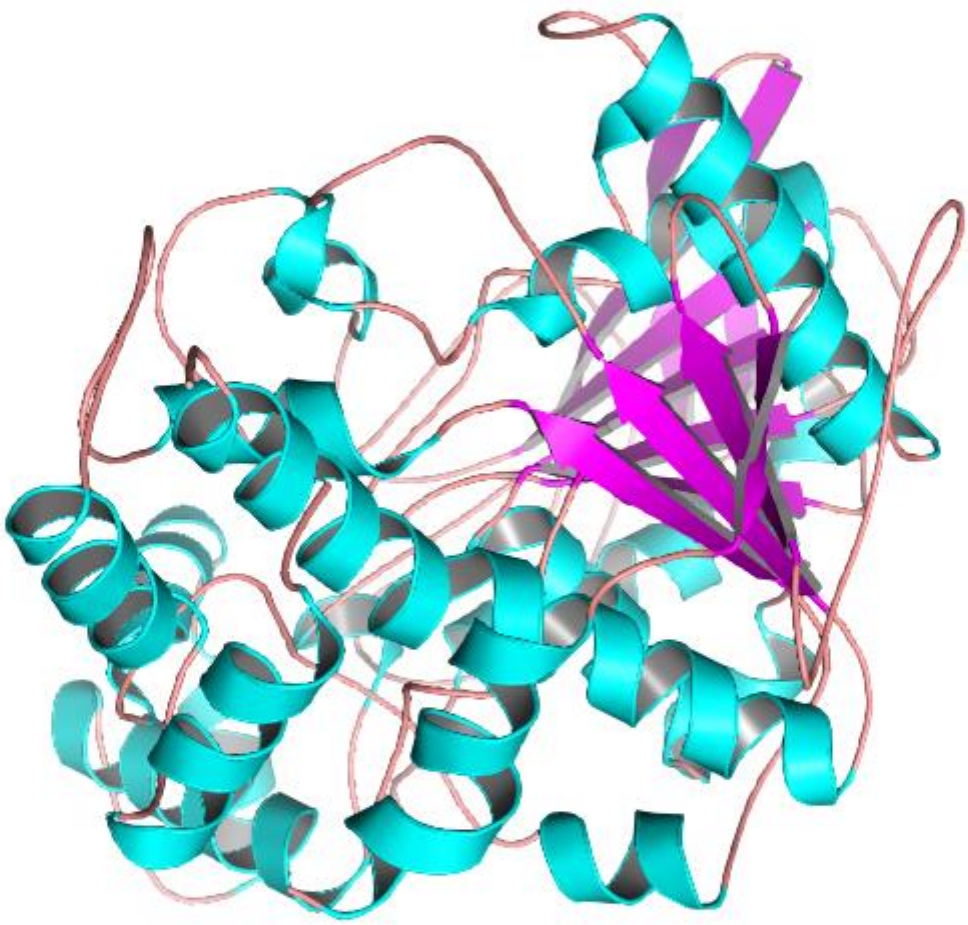
Surface representation (solvent-excluded surface model) of human BChE showing the active-site gorge (PDB ID: 4BDS).



Stick representation (bond-centric model) of human BChE showing the active-site residues (PDB ID: 4BDS).



Cartoon representation (abstract model) of human BChE showing the secondary structural elements (PDB ID: 4BDS).



Biomolecular Visualization Systems (cont'd)

- The most robust and popular biomolecular visualization tools are as follows:

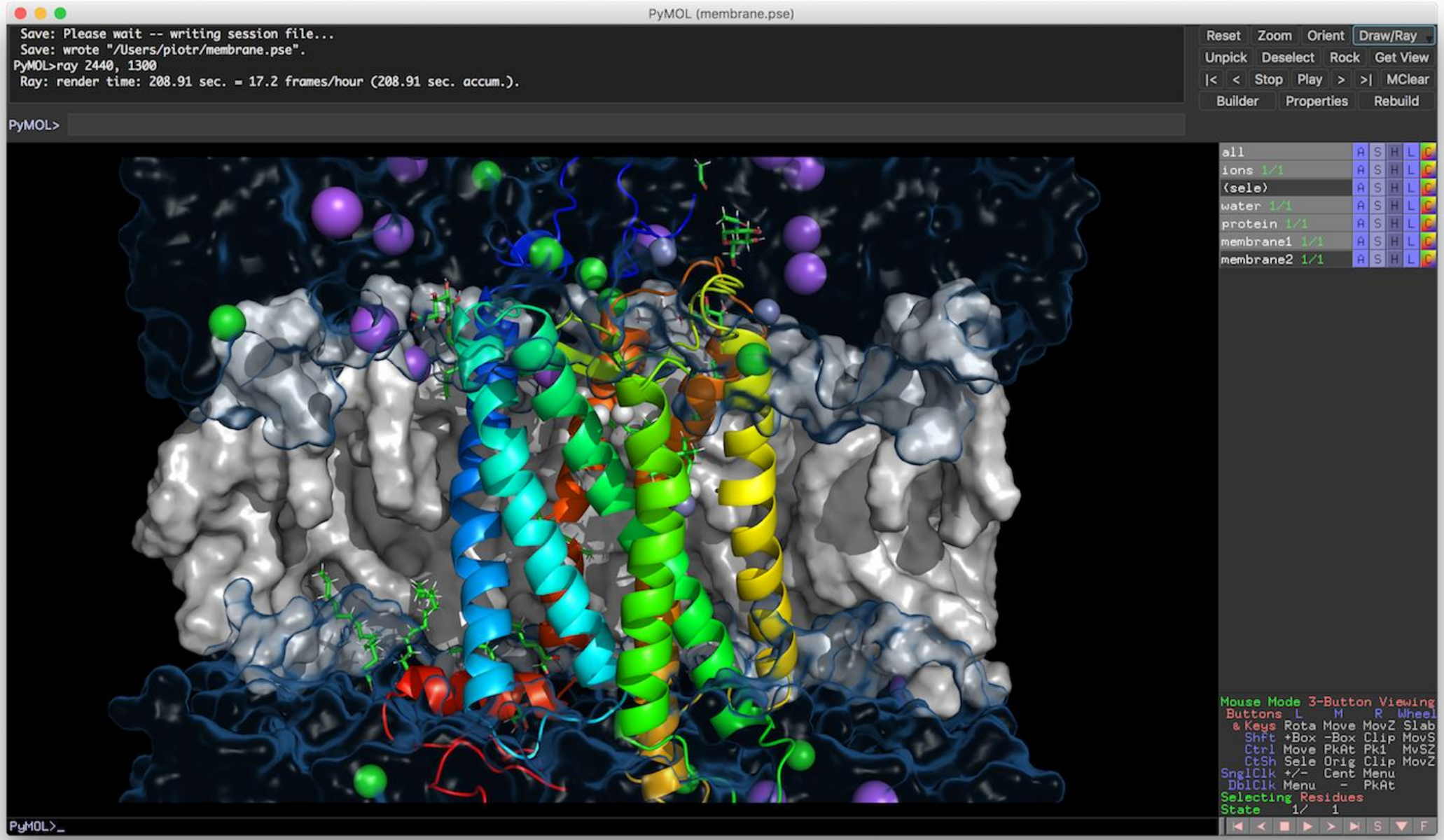
  i. VMD

  ii. PyMOL 👍

  iii. UCSF Chimera

  iv. YASARA View

  v. CAVER Analyst



PyMOL>rotate x, 90

PyMOL>zoom center, 12

PyMOL>create merged, membrane1 | membrane2

Biomolecular Modeling

Why to Build a Model?

Historical Concepts

The «physics» concept:

The native conformation of a protein corresponds to a global free energy minimum of the protein / solvent system. To identify the correct fold, some form of energy calculation should be used to evaluate compatibility of the protein sequence with a structural conformation.

Ludwig Boltzmann
(1844–1906)

The «biology» perspective:

Homologous proteins have evolved by molecular evolution from a common ancestor over millions of years. If we can establish homology to a known protein, we can predict aspects of structure and function of a new protein by similarity.

Charles Darwin
(1809–1882)

Protein Structure Prediction

- **Protein structure prediction** offers a theoretical alternative to experimental determination of 3D models.
- It is an efficient means of obtaining structural information when experimental techniques fail.
- It can also be employed to avoid the cost and time involved in X-ray diffraction, NMR spectroscopy, etc.
- Computational methods for protein structure prediction are divided into three categories: homology (comparative) modeling; threading (fold recognition); and *ab initio* modeling.

Homology Modeling

- **Homology**, or **comparative**, **modeling**, which is the most accurate method, derives models from the available structural information contained in close (*i.e.* having over 30% sequence identity) homologs.

- It involves an elaborate procedure of template selection, template–target alignment, main-chain prediction, loop modeling, side-chain packing, model refinement, and model evaluation.

- Among these steps, sequence alignment is the most important step, and loop modeling is the most difficult and error-prone step.

Homology Modeling (cont'd)

- Some of the most popular homology modeling tools are as follows:

  i. MODELLER

  ii. SWISS-MODEL 👍

  iii. CPHmodels

  iv. RosettaCM

  v. YASARA Structure

Threading

- **Threading** or **fold recognition** searches for a best-fitting structure in a structural fold library through matching secondary structure and energy criteria.

- It is used when no suitable templates can be found for homology modeling.

- The limitation is that this approach does not generate a biologically realistic model, but provides an essentially correct fold for the protein of interest.

Threading (cont'd)

- Some of the most popular threading tools are as follows:

  i. Phyre2

  ii. I-TASSER

  iii. RaptorX

Ab Initio Modeling

- ***Ab initio* modeling** attepmts to generate a structure without relying on templates, but by using physical rules only.

- It is used when neither homology modeling nor threading can be applied.

- However, the *ab initio* approach so far has limited success in predicting accurate protein structures.

- An objective evaluation platform, **CASP**, has been established to allow program developers to test the effectiveness of their prediction algorithms.

- Continued progress in *ab initio* modeling will be key to further refine homology models (and remote homology models) to higher accuracy.

Energy Minimization

How to Optimize Protein Structure?

Effects of Disease-Causing Missense Mutations on Proteins

- Pathogenic missense mutations can affect the function of a protein in various ways, including:

    i. altering protein stability (*i.e.* destabilizing or stabilizing the wild-type protein fold);

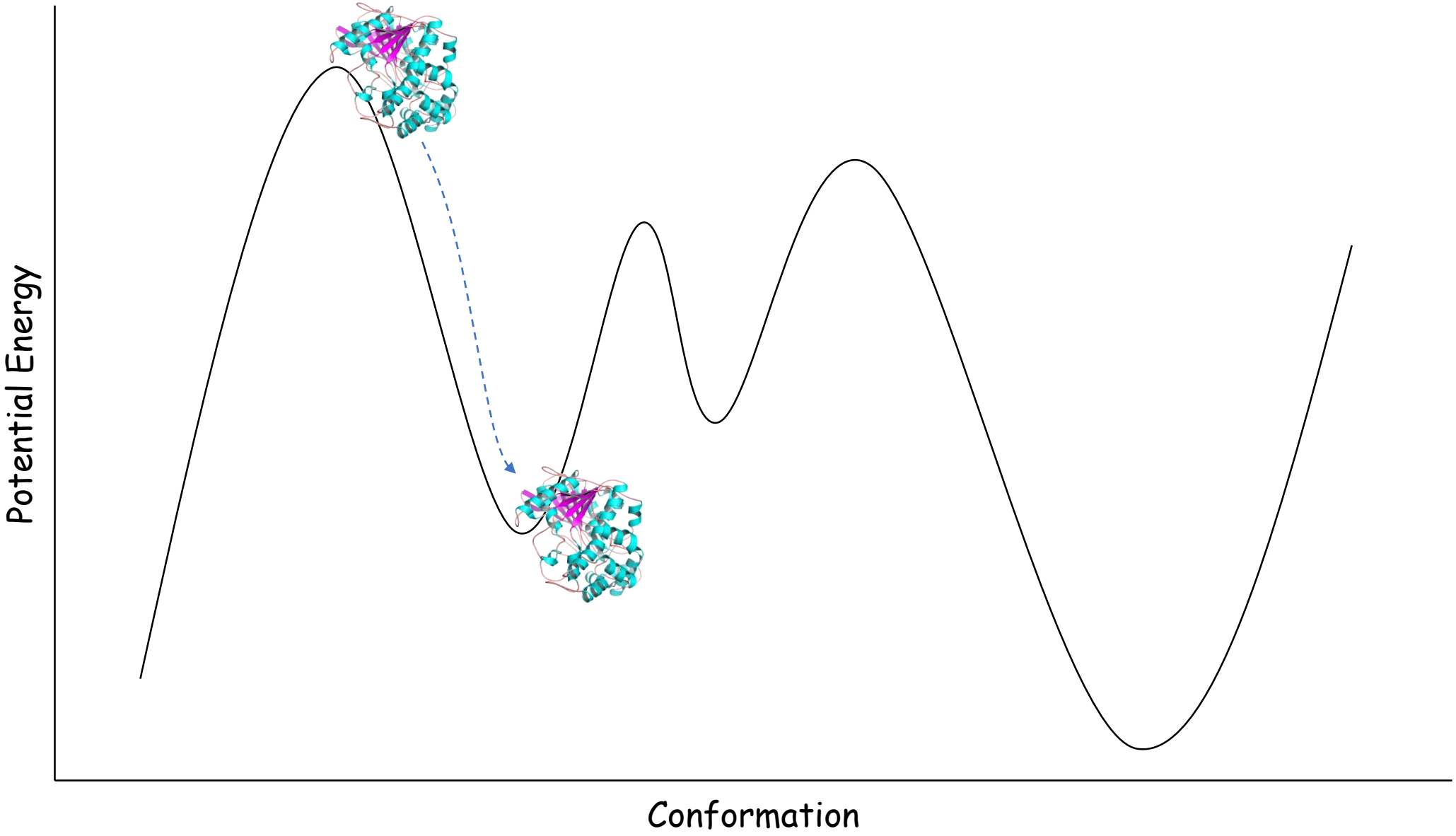
    ii. altering protein–ligand or protein–protein interactions;

    iii. altering H-bonding network;

    iv. many others.

Molecular Mechanics

- The structural consequences of missense mutations can be studied using *in silico* mutagenesis.
- Typically, *in silico* mutagenesis is followed by structure optimization which improves physical realism, stereochemistry and side-chain accuracy.
- For instance, **YASARA Energy Minimization Server** performs energy minimization of the mutant protein model in the presecence of a solvent (H_2O) shell.
- At the energy-minimized point, the configuration will ideally be in local potential energy minimum.
- Energy-minimizing protein receptors prior to docking is also a useful strategy for target preparation.


Potential Energy
Conformation

Molecular Docking

When to Use It?

Protein–Ligand Interactions

What happens when a small-molecule ligand binds within a cavity on a protein?

a) It can activate the protein.

b) It can inhibit the protein.

c) It can be metabolized by the protein.

d) It can be transported by the protein.

e) All of the above.

Protein–Ligand Interactions (cont'd)

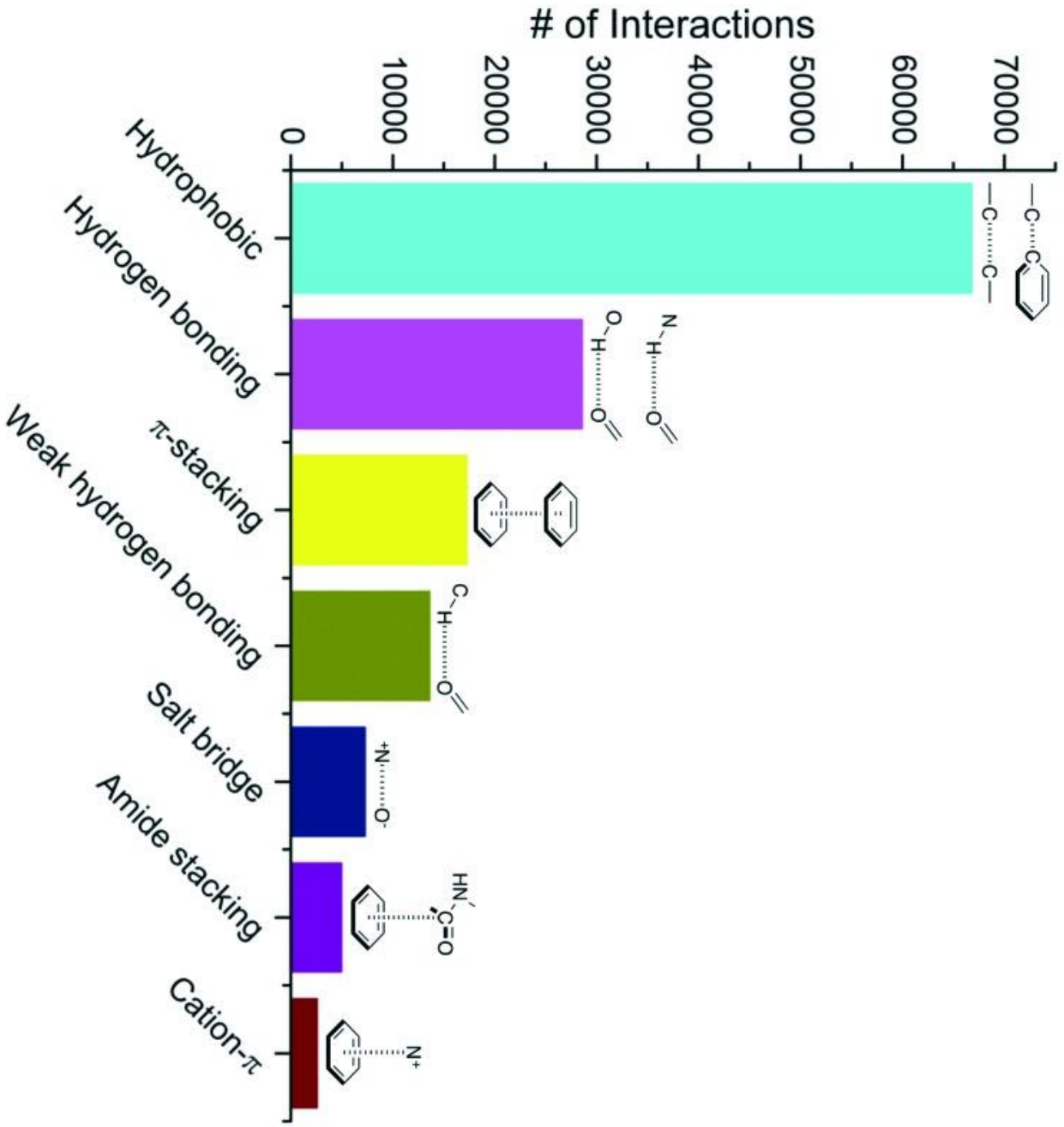
- Many biomolecules interact with small molecules, such as cofactors, metabolites, or drugs, collectively defined as «ligands».

- Biomolecular interactions including enzyme–substrate, receptor–signaling molecule, and antibody–antigen play crucial roles in numerous biological processes.

- These interactions are primarily due to complementary H-bonds, salt bridges, hydrophobic contacts, etc. between a protein and a ligand.

- Predicting ligands that bind with sufficient strength to a corresponding protein is a challenging task in biochemistry and has significant implications for drug discovery.

Frequency distribution of the most common non-covalent interactions observed in protein–ligand complexes extracted from the PDB (*Med. Chem. Commun.*, 2017, **8, 1970 –1981)**

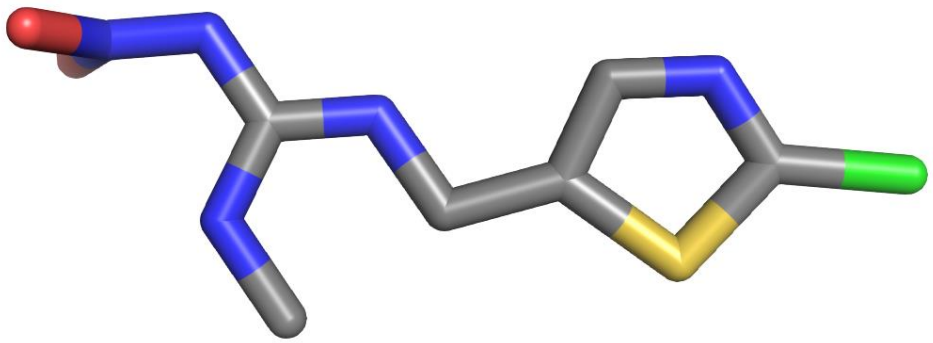


Principles of Molecular Docking

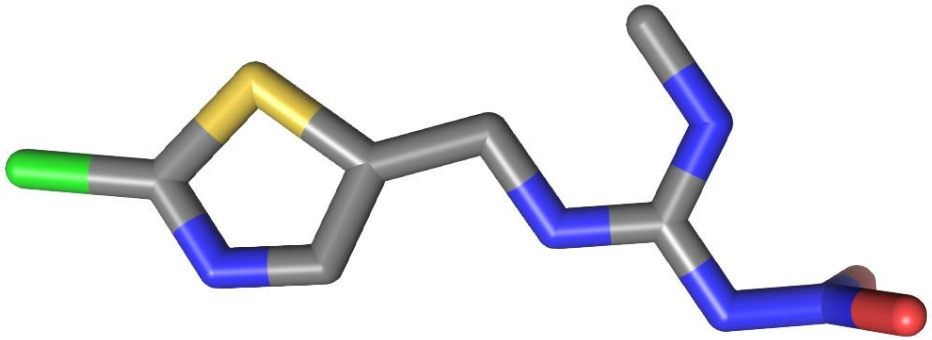
- Typically, the goals of **molecular docking** are the identification of a ligand that binds within a cavity on a receptor and the prediction of its preferred (*i.e.* energetically most favorable) binding pose.

- The term «binding pose» considers the **orientation** of a ligand relative to its receptor as well as the ligand's **conformation** and **position**.

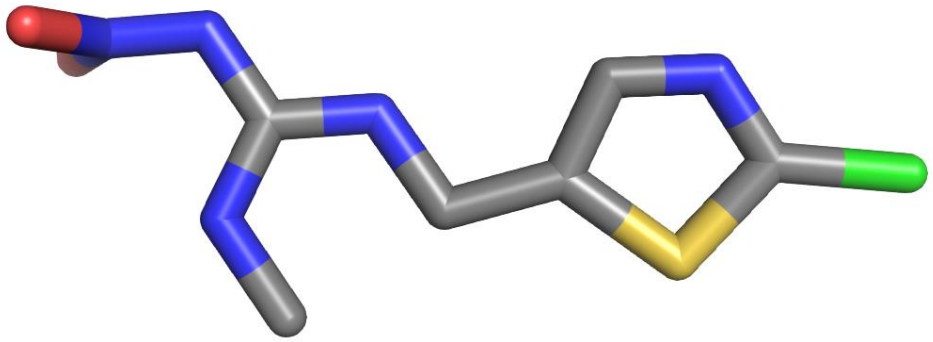
- In order to accomplish this task, molecular docking tools will generate a set of different ligand binding poses and use a scoring function to estimate binding affinities for the generated poses in order to determine the best binding mode.



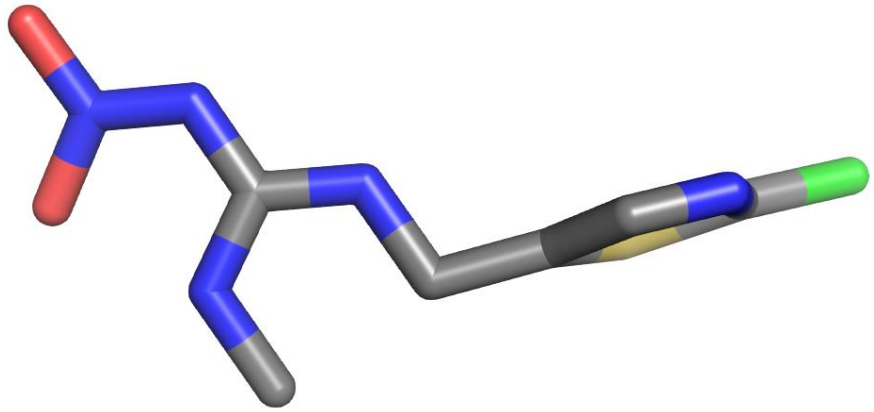
Orientation A



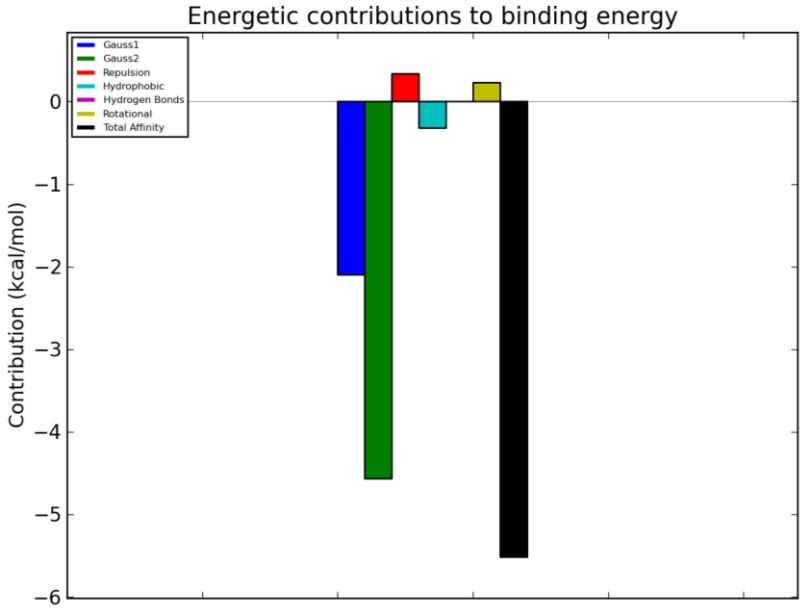
Orientation B

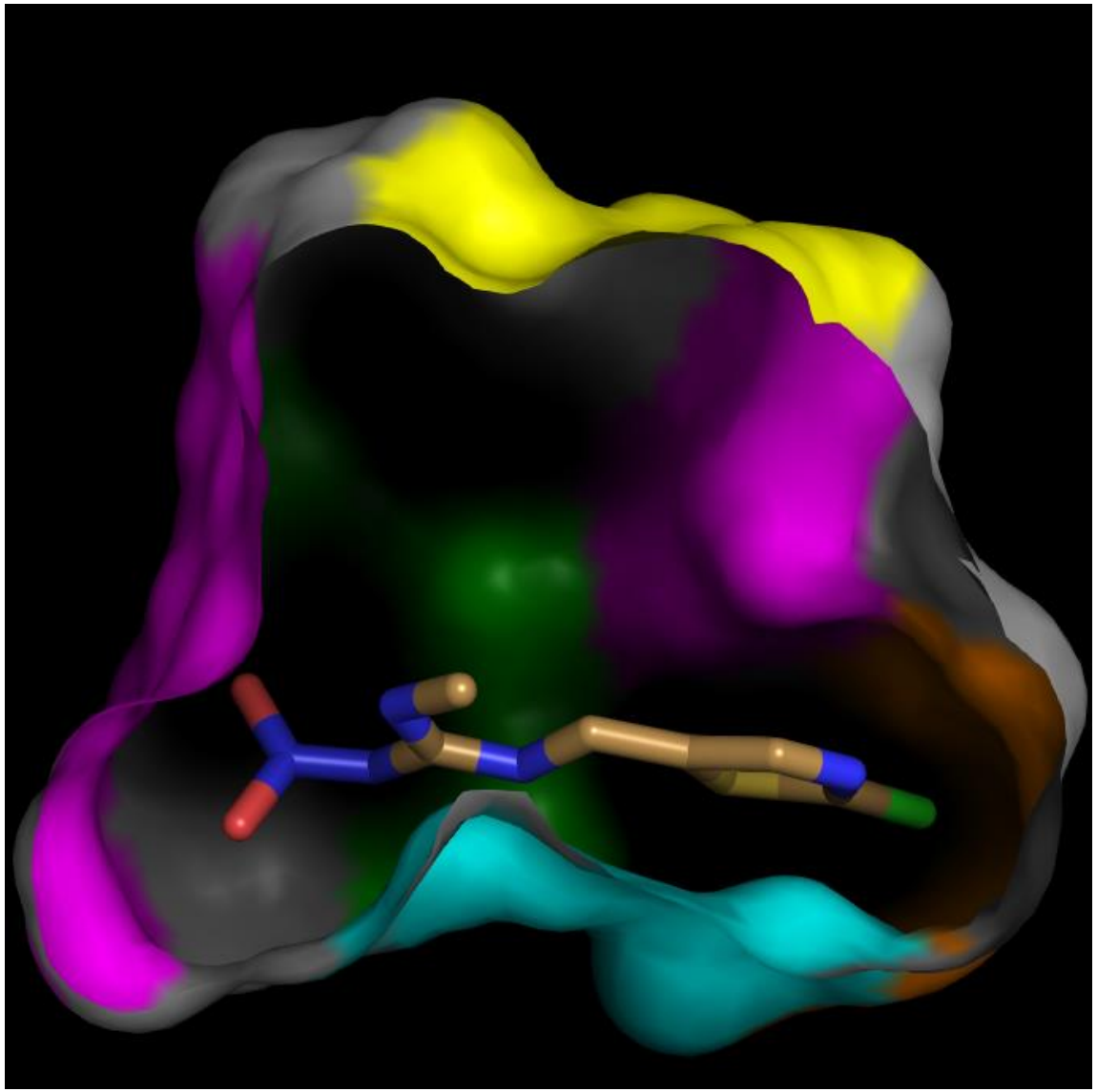


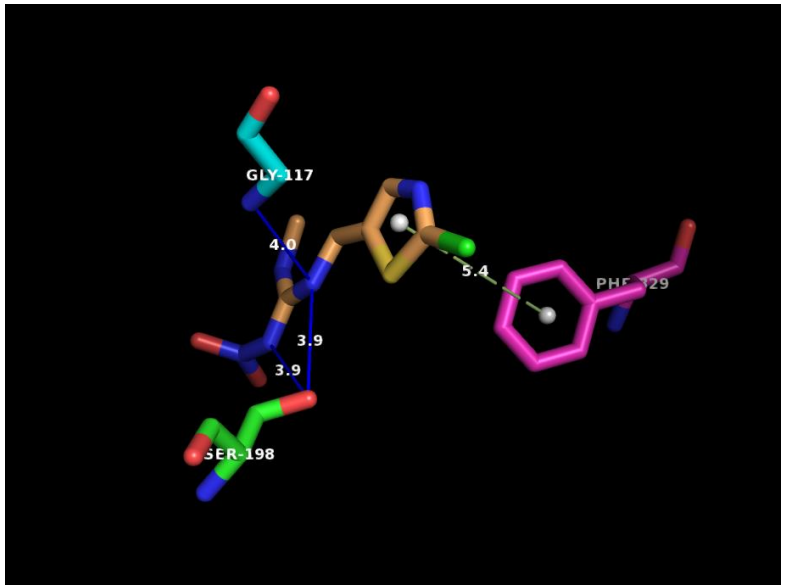
Conformation A



Conformation B


Energetic contributions to binding energy
Gauss1
Gauss2
Repulsion
Hydrophobic
Hydrogen Bonds
Rotational
Total Affinity
Contribution (kcal/mol)
0
−1
−2
−3
−4
−5
−6




GLY-117
4.0
5.4
3.9
3.9
SER-198

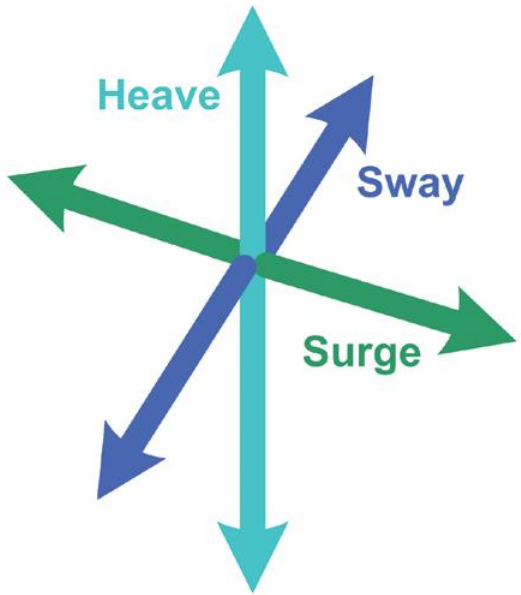
Principles of Molecular Docking (cont'd)

- Protein–ligand docking procedures are grouped into two categories:

  i. Rigid docking. This approximation treats both the ligand and the receptor as rigid and explores only six degrees of translational and rotational freedom, hence excluding any kind of flexibility.

  ii. Flexible docking. A more common approximation is to model the ligand flexibility while assuming a rigid protein receptor, thereby considering only the conformational space of the ligand.

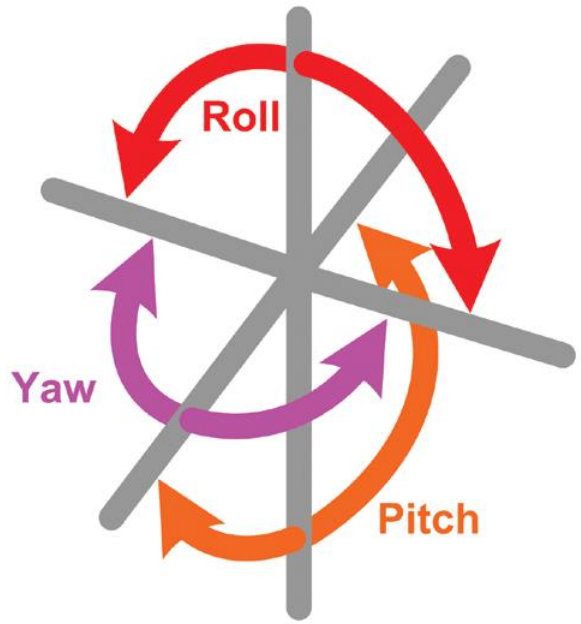
Translational Movement in Three Perpendicular Axes



Surge: Moving forward/backward
Heave: Moving up/down
Sway: Moving left/right

+

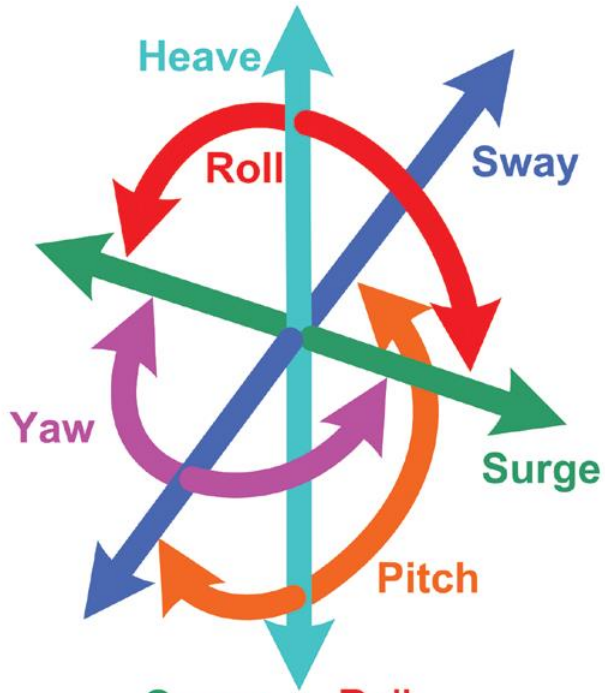
Rotational Movement about Three Perpendicular Axes



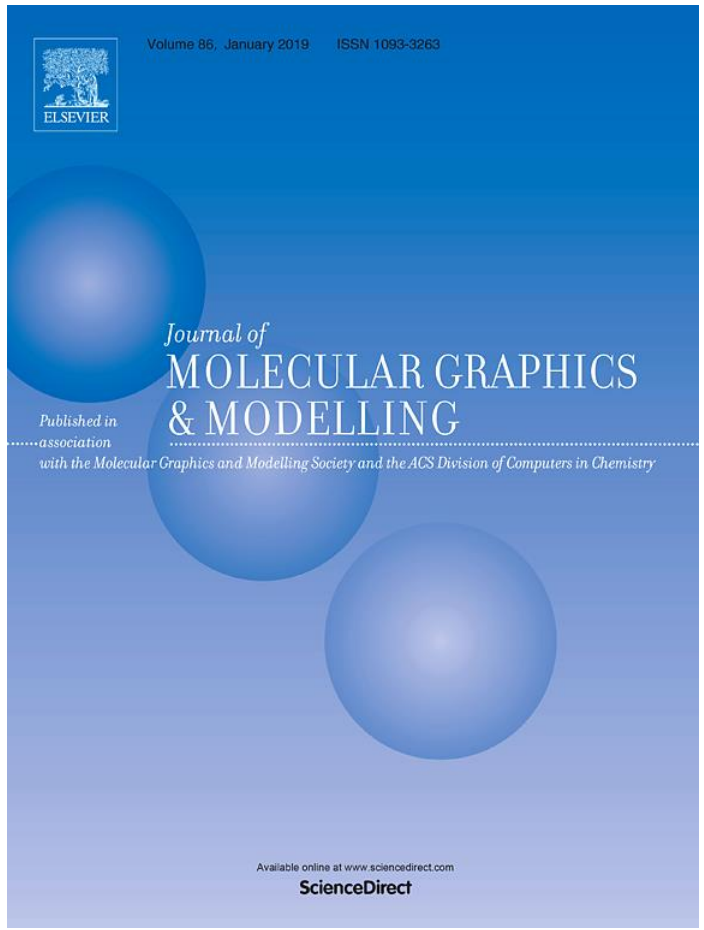
Roll: Tilting side to side
Pitch: Tilting forward and backward
Yaw: Turning left and right

=

Six Degrees of Freedom



Surge **Roll**
Heave **Pitch**
Sway **Yaw**



ELSEVIER

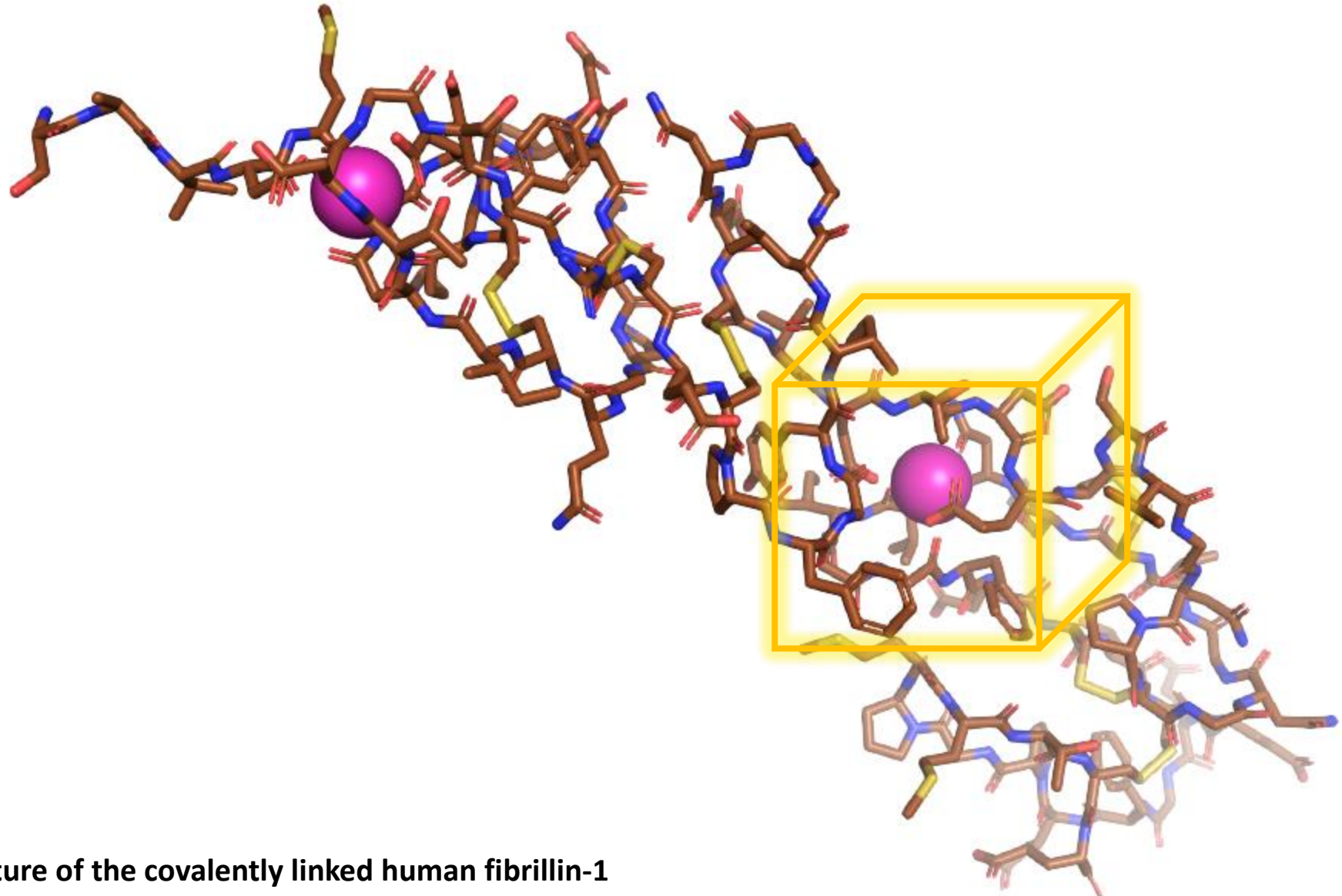
Kerem Teralı (2018): An evaluation of neonicotinoids' potential to inhibit human cholinesterases: protein–ligand docking and interaction profiling studies, *Journal of Molecular Graphics and Modelling*, 84, 54–63, DOI: 10.1016/j.jmgm.2018.06.013

Marfan Syndrome (MFS)

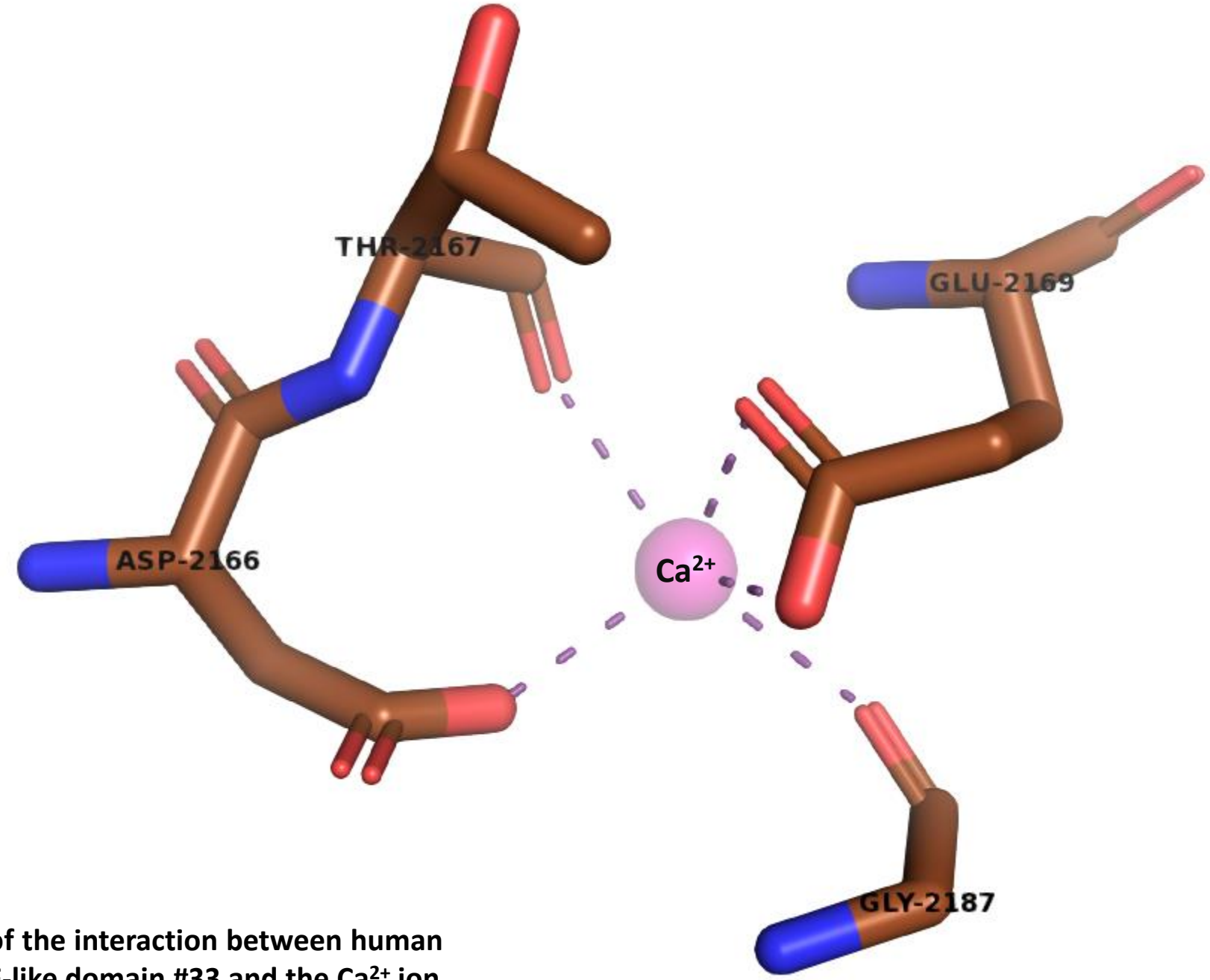
Mutation: c.7828G>C (p.E2610Q) in *FBN1*

MFS: Background

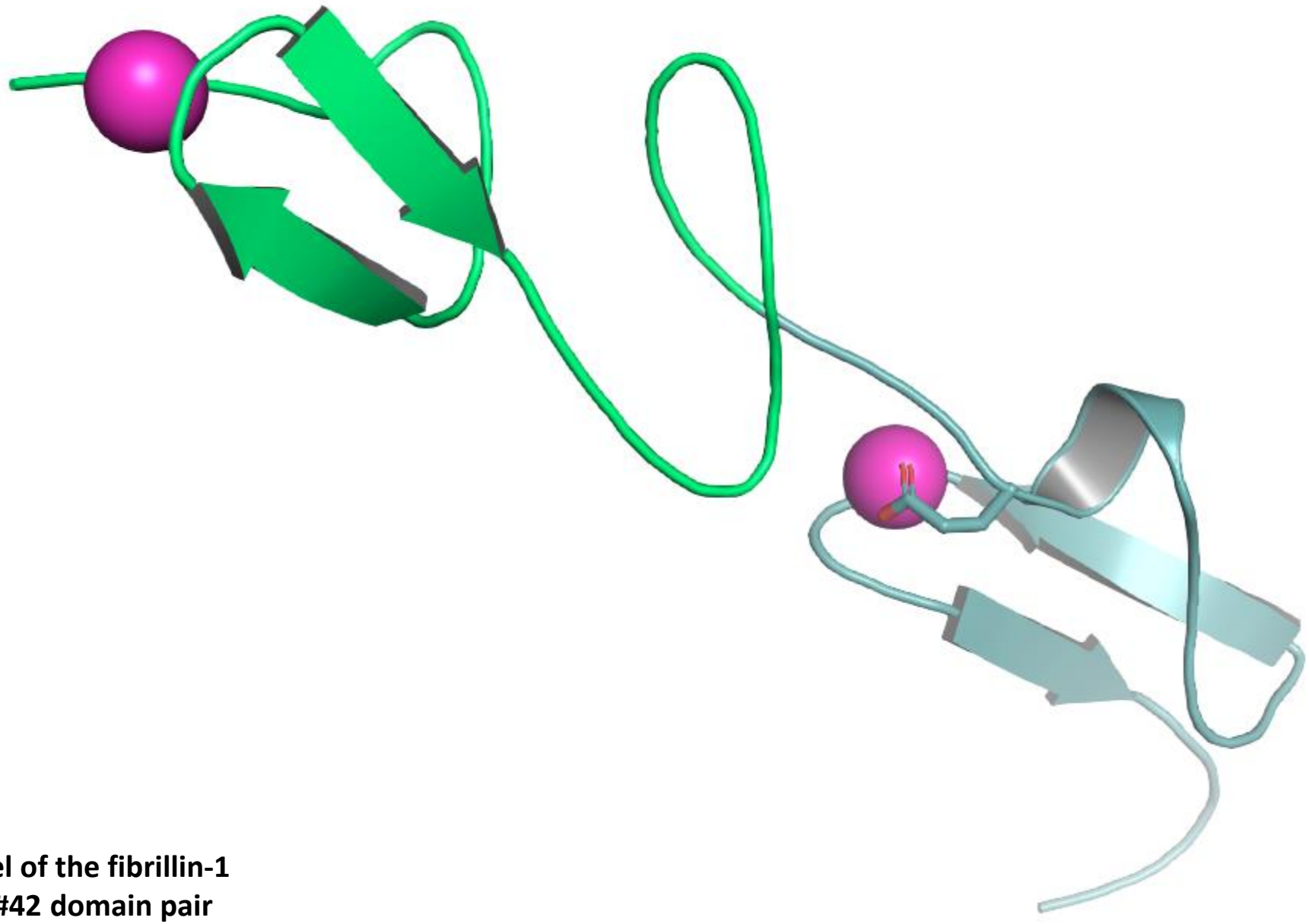
- MFS is a multisystem disorder with musculoskeletal, ocular and cardiovascular abnormalities.
- It results from mutations in *FBN1* that encodes fibrillin-1, a secreted 350-kDa ECM glycoprotein.
- Fibrillin-1 mostly consists of epidermal growth factor (EGF)-like domains.
- Of the 47 EGF-like domains in human fibrillin-1, 43 start with the conserved D-X-D/N-E motif which is involved in calcium binding (hence called cbEGF-like domains).
- Calcium binding to fibrillin-1 has been shown to provide structural stabilization, protection against proteolysis, and structural determinants for interaction with a number of elements of the ECM.



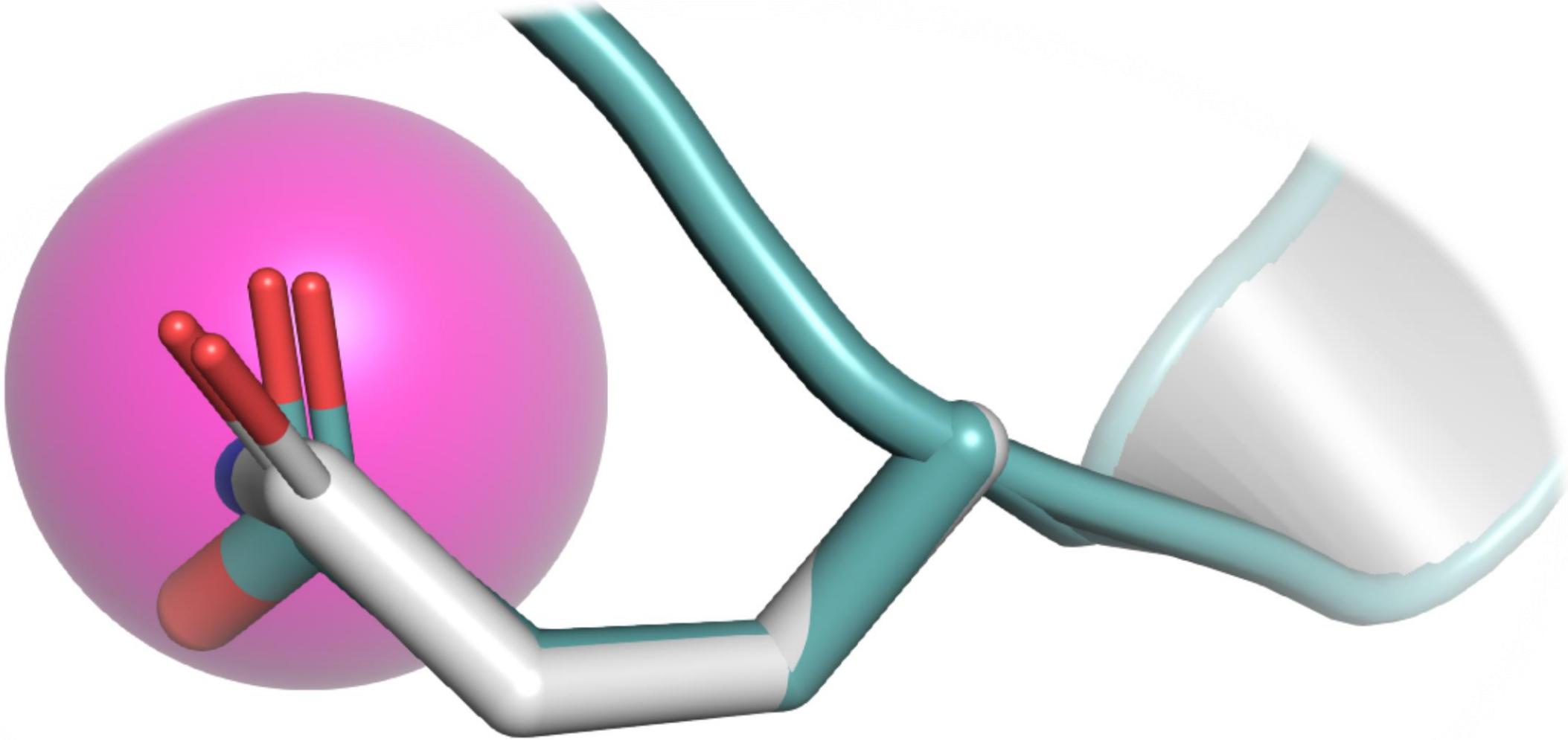
NMR structure of the covalently linked human fibrillin-1 cbEGF-like domains #32 and #33 (PDB ID: 1EMN)



Close-up view of the interaction between human fibrillin-1 cbEGF-like domain #33 and the Ca^{2+} ion



Homology model of the fibrillin-1 cbEGF-like #41–#42 domain pair



Close-up view of the *in silico* introduced p.E2610Q mutation

MFS: Conclusion

- The p.E2610Q substitution is likely to cause a minimal perturbation to the local protein structure because Glu and Gln are similar in hydrophilicity, α-helical propensity, and spatial requirements.
- However, the substitution is predicted to hinder Ca^{2+} binding due to the loss of one of the metal-coordinating oxygen atoms.
- We suggest that the p.E2610Q substitution impairs the ability of the C-terminal portion of fibrillin-1 to bind Ca^{2+}, possibly leaving the mutant protein vulnerable to proteases.
- Alternatively, it may disrupt the interplay between fibrillin-1 and its interacting partners or affect fibrillin-1 secretion.

Useful Web Sites

- UniProtKB (https://www.uniprot.org/uniprot/)
- RCSB PDB (https://www.rcsb.org/)
- Protein–Ligand Interaction Profiler (PLIP; https://projects.biotec.tu-dresden.de/plip-web/plip/index)
- CPHmodels (http://www.cbs.dtu.dk/services/CPHmodels/)
- Open-Source PyMOL (https://github.com/schrodinger/pymol-open-source)

Taylor & Francis
Taylor & Francis Group

Mahmut Cerkez Ergoren, Burcu Turkgenc, Kerem Teralı, Orhan Rodoplu, Aline Verstraeten, Lut Van Laer, Gamze Mocan, Bart Loeys, Omer Tetik & Sehime G. Temel (2018): Identification and characterization of a novel *FBN1* gene variant in an extended family with variable clinical phenotype of Marfan syndrome, *Connective Tissue Research*, DOI: 10.1080/03008207.2018.1472589

Achondroplasia (ACH) with Psychomotor Delay

Mutation: c.1138G>A (p.G380R) in *FGFR3*

Unpublished Data

Treacher–Collins Syndrome (TCS)

Mutation: c.299T>C (p.L100P) in *POLR1D*

Unpublished Data

The End

Any Questions?