SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea Magistrale in Fisica

ID number

Advisor

Co-Advisor

Academic Year

Thesis not yet defended

Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email:

Contents

Contents

Introduction

Proteins are essential components of life and key to the functionality of living organisms. [1] They are composed of a sequence of amino acids, a chain of amino acids, that fold into a three-dimensional structure that determines their function. The study of proteins is crucial for understanding the molecular basis of life and for the development of new drugs. The determination of the three-dimensional structure of a protein is a fundamental step in the study of its function.

Among the main functions of proteins are:[1]

- **Enzymatic catalysis**: Enzymatic proteins accelerate chemical reactions by reducing activation energy and regulating cellular metabolic processes.
- **Structural roles**: Some proteins, such as collagen or keratin, provide mechanical support and structural integrity to tissues.
- **Transport**: Molecules like hemoglobin facilitate oxygen transport, while other proteins transport nutrients and ions across cellular membranes.
- **Regulation and signaling**: Proteins play key roles in cellular signal transduction and the regulation of gene expression.

Our goal is to study the interactions between the amino acids that compose the protein, in order to understand how they fold into a functional structure.

0.1 Structure of Amino Acids

Each amino acid consists of a central carbon atom (α-carbon) bonded to four groups:[2]

- **Amino Group ($-NH_2$)**: A basic functional group.
- **Carboxyl Group ($-COOH$)**: An acidic functional group.
- **Side Chain (R-Group)**: This varies for each amino acid, giving it unique properties.
- **Hydrogen Atom**: A single hydrogen atom completes the structure.

The R-group, or side chain, determines the chemical properties of the amino acid, such as whether it is hydrophilic, hydrophobic, acidic, or basic.

0.1.1 Formation of Proteins

Proteins are formed through a process called polymerization, where amino acids are linked together by peptide bonds. This process involves a condensation reaction, where the carboxyl group of one amino acid reacts with the amino group of another, releasing a molecule of water.[0]

- **Primary Structure**: The linear sequence of amino acids, held together by peptide bonds.
- **Secondary Structure**: Local folding into structures such as α-helices and β-sheets, stabilized by hydrogen bonds.
- **Tertiary Structure**: The overall three-dimensional structure of a single polypeptide chain, determined by interactions such as ionic bonds, disulfide bonds, and hydrophobic interactions.
- **Quaternary Structure**: The arrangement of multiple polypeptide chains into a functional protein complex.

0.1.2 Significance of Amino Acids in Proteins

The specific sequence and properties of the amino acids determine the final structure and function of the protein. Proteins are involved in numerous biological functions, including catalysis (enzymes), signaling (hormones), and structural support (collagen).

Understanding how amino acids combine and fold into functional proteins is essential for exploring their role in biochemistry and molecular biology.

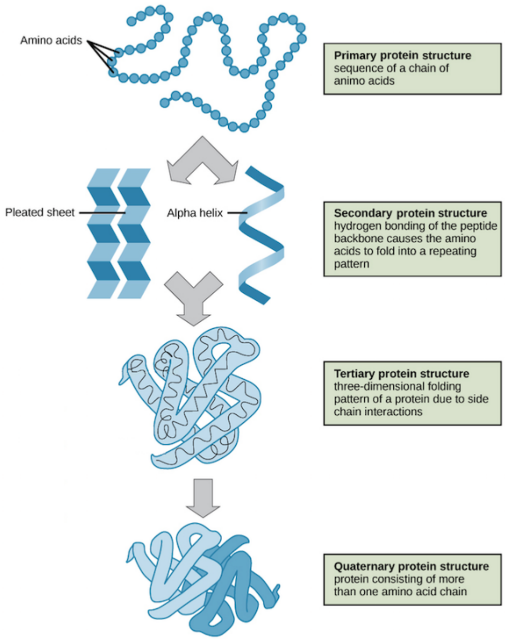

Amino acids
Primary protein structure
sequence of a chain of
animo acids
Pleated sheet
Alpha helix
Secondary protein structure
hydrogen bonding of the peptide
backbone causes the amino
acids to fold into a repeating
pattern
Tertiary protein structure
three-dimensional folding
pattern of a protein due to side
chain interactions
Quaternary protein structure
protein consisting of more
than one amino acid chain

Figure 0.1. Secondary strucutre

0.2 Protein structure

The scructure of the protein is important because it determines its function. The structure of a protein is divided into four levels:[0]

0.2.1 Primary Structure

The primary structure refers to the linear sequence of amino acids in a protein chain, held together by peptide bonds. This sequence is unique for each protein and dictates the higher levels of structure.

0.2.2 Secondary Structure

The secondary structure refers to the local folding of the protein chain into regular patterns such as alpha-helices and beta-pleated sheets. These structures are stabilized by hydrogen bonds between the backbone atoms of the amino acids.

0.2.3 Tertiary Structure

The tertiary structure describes the three-dimensional folding of the entire protein molecule, including all its side chains. This level of structure is stabilized by interactions such as hydrogen bonds, ionic bonds, hydrophobic interactions, and disulfide bridges.

0.2.4 Quaternary Structure

The quaternary structure applies to proteins that consist of more than one polypeptide chain. It describes how these chains are arranged and interact with each other to form the functional protein complex.

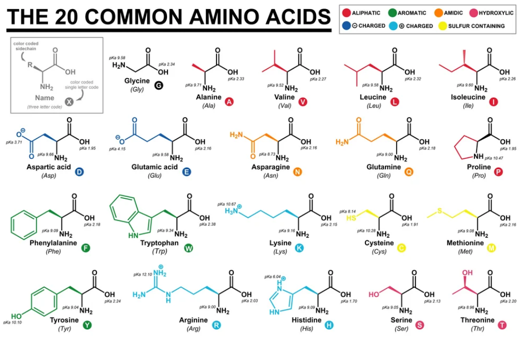

THE 20 COMMON AMINO ACIDS
● ALIPHATIC ● AROMATIC ● AMIDIC ● HYDROXYLIC
● ⊖ CHARGED ● ⊕ CHARGED ● SULFUR CONTAINING
color coded sidechain
color coded single letter code
Name (three letter code) X
Glycine (Gly) G pKa 9.58 pKa 2.34
Alanine (Ala) A pKa 9.71 pKa 2.33
Valine (Val) V pKa 9.52 pKa 2.27
Leucine (Leu) L pKa 9.58 pKa 2.32
Isoleucine (Ile) I pKa 9.60 pKa 2.26
Aspartic acid (Asp) D pKa 3.71 pKa 9.66 pKa 1.95
Glutamic acid (Glu) E pKa 4.15 pKa 9.58 pKa 2.16
Asparagine (Asn) N pKa 8.73 pKa 2.16
Glutamine (Gln) Q pKa 9.00 pKa 2.18
Proline (Pro) P pKa 10.47 pKa 1.95
Phenylalanine (Phe) F pKa 9.09 pKa 2.18
Tryptophan (Trp) W pKa 9.34 pKa 2.38
Lysine (Lys) K pKa 10.67 pKa 9.16 pKa 2.15
Cysteine (Cys) C pKa 8.14 pKa 10.28 pKa 1.91
Methionine (Met) M pKa 9.08 pKa 2.16
Tyrosine (Tyr) Y pKa 10.10 pKa 9.04 pKa 2.24
Arginine (Arg) R pKa 12.10 pKa 9.00 pKa 2.03
Histidine (His) H pKa 6.04 pKa 9.09 pKa 1.70
Serine (Ser) S pKa 9.05 pKa 2.13
Threonine (Thr) T pKa 8.96 pKa 2.20

Figure 0.2. Seocndaria

This structure are due to the interactions between the amino acids that compose the protein. Following we will see our model for the interaction between the amino acids.

0.3 Allostericity

Allostericity is the phenomenon of the change of the protein structure by the trasmission of a signal from one site to another. This avoid the protein to do various tasks in a regulated way. Allostericity, from the Greek *allos* (other) and *stereos* (structure), refers to a phenomenon in which the interaction of a molecule (effector) with a specific site of a protein, known as the allosteric site, induces a conformational change that influences the functional activity of another site, usually the active site. [0] This process does not occur through direct interactions between the two sites but rather via changes in the network of intramolecular interactions that regulate the structure and dynamics of the protein.

From a thermodynamic perspective, allostericity can be described as a modulation of the distribution of a protein's energy microstates. In other words, the interaction with an allosteric effector alters the set of available conformational states, facilitating or inhibiting access to functional configurations.[0]

Mathematically, the phenomenon can be represented as a variation in the population of microstates $\{s_i\}$, described by a weighted Boltzmann distribution:[0]

$$P(s_i) = \frac{e^{-\Delta F(s_i)/k_B T}}{\sum\limits_j e^{-\Delta F(s_j)/k_B T}}$$

where $\Delta F(s_i)$ represents the free energy associated with microstate s_i, k_B is the Boltzmann constant, and T is the absolute temperature.

The allosteric effect can also be quantified by considering the change in free energy associated with the effector interaction. For a protein with two principal states, R (relaxed) and T (tense), the allosteric equilibrium can be expressed in terms of an equilibrium constant:[0]

$$L = \frac{[T]}{[R]}$$

where $[T]$ and $[R]$ represent the relative concentrations of the two conformational states. The presence of an effector modifies L, stabilizing one state over the other.

Proteins are not rigid structures; instead, they are ensembles of conformational states that fluctuate over temporal and spatial scales. Allosteric effectors influence this dynamic network, remodeling energy pathways and connecting various functional states. Proteins can be represented as an ensemble of conformational states $\{C_i\}$, where i denotes a specific conformational state. The distribution of populations across states is described by a partition function $Z = \sum_i e^{-\beta E_i}$, where E_i is the energy associated with state C_i and $\beta = \frac{1}{k_B T}$ is the thermodynamic factor.

An allosteric effector can be represented as a perturbation that modifies the energy landscape of the conformational states $\{E_i\}$, remodeling energy pathways and altering the probability P_i of each state, where:[0]

$$P_i = \frac{e^{-\beta E_i}}{Z}, \quad \text{with } Z = \sum_i e^{-\beta E_i}.$$

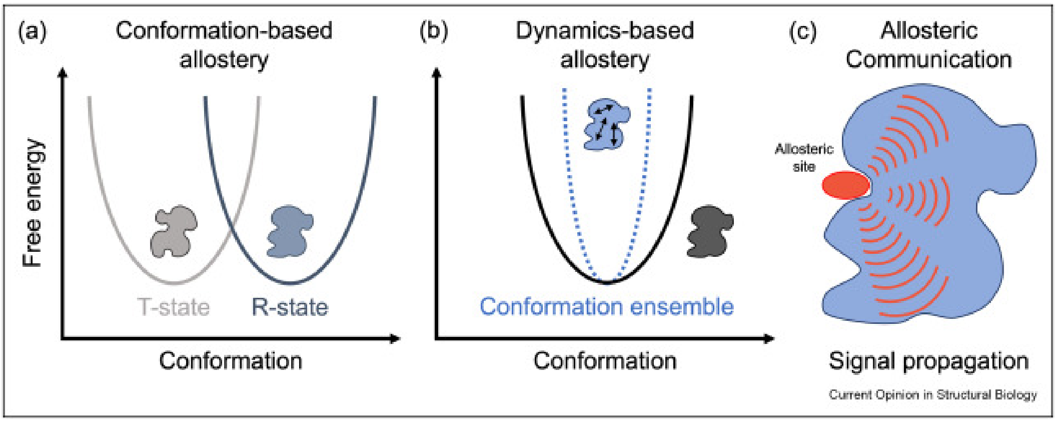


Figure 0.3. Schematic representation of allostericity: the interaction with an allosteric effector (site A) induces a conformational change that affects a distant site (site B).

This dynamic view integrates well with computational approaches such as molecular dynamics and interaction residue networks, which allow for the identification of specific pathways through which allosteric information propagates within the protein.

Understanding how and why the proteins fold into a functional structure it is useful for the study of the molecular basis of life and for the development of new drugs. To study it we need to understand the causal mechanism and the propagation of the signal between the amino acids that compose the protein.

0.4 Example of Allostericity in Hemoglobin and Red Blood Cells for Oxygen Transport

Hemoglobin, the oxygen-transporting protein within red blood cells, is a classic example of an allosteric protein. Its function is to bind oxygen in the lungs and release it efficiently in tissues where oxygen is needed. This process is regulated by its allosteric properties, which involve conformational changes and cooperative binding.

0.4.1 Mechanism of Allostericity in Hemoglobin

Hemoglobin exists in two main states:[0]

- **T-state (Tense):** This state has a lower affinity for oxygen and is stabilized in tissues where oxygen levels are low.
- **R-state (Relaxed):** This state has a higher affinity for oxygen and is stabilized in the oxygen-rich environment of the lungs.

When the first oxygen molecule binds to one of hemoglobin's four subunits, it induces a conformational change that increases the affinity of the remaining subunits for oxygen. This phenomenon is called *cooperative binding* and is a hallmark of allosteric regulation.

Similarly, the release of oxygen in tissues is facilitated by other effectors, such as:

- **Carbon dioxide** (CO_2) and **protons** (H^+), which stabilize the T-state (Bohr effect).
- **2,3-Bisphosphoglycerate (2,3-BPG)**, which reduces oxygen affinity to promote oxygen release in tissues.

0.4.2 Practical Applications of Understanding Allostericity

The study of allostericity in proteins, particularly hemoglobin, has several practical applications, including:

- **Medical Interventions:** Developing drugs to modulate hemoglobin's oxygen affinity could help treat conditions like sickle cell anemia, anemia, or hypoxia in high-altitude environments.
- **Artificial Blood Substitutes:** Designing synthetic hemoglobin mimics with tunable oxygen-binding properties for use in emergencies or surgeries.
- **Diagnostics:** Allosteric changes in hemoglobin can serve as biomarkers for detecting diseases such as diabetes (glycated hemoglobin levels) or metabolic disorders.
- **Enhanced Oxygen Delivery:** Creating treatments to improve oxygen delivery to tissues, potentially benefiting athletes or individuals with cardiovascular or pulmonary diseases.

0.4.3 Summary of Potential Applications

Application	**Description**
Medical Interventions	Drugs to adjust oxygen affinity in diseases like anemia or sickle cell anemia.
Artificial Blood Substitutes	Synthetic hemoglobin for transfusions in emergencies.
Biomarkers	Using hemoglobin changes to detect diseases like diabetes.
Enhanced Oxygen Delivery	Improving oxygen transport for athletes or patients with breathing issues.

Understanding the principles of allostericity allows researchers to develop novel strategies to improve human health and tackle diseases more effectively.

Chapter 1

PDZ Domain

Definition of Protein Domains and the PDZ Domain

In molecular biology, a **domain** is a specific region of a protein that can perform a structural or functional role independently from the rest of the protein. Domains are essential because they serve as modular units that recur and combine to form complex proteins with distinct functions.

Definition of Domain

A protein domain is:

1. **A structurally independent unit**: It is a portion of the protein that can fold into a stable three-dimensional structure on its own.

2. **A functionally independent unit**: It can perform specific tasks, such as binding to a molecule or interacting with other proteins.

3. **An evolutionary modular unit**: Domains are often reused in different proteins during evolution, allowing for the development of new functions.

The PDZ Domain

The **PDZ domain** is an example of a protein domain. The name "PDZ" is derived from the three proteins in which it was first identified:

- **P**SD-95 (Protein Discs Large 4),

- **D**lgA (Drosophila discs large protein),

- **Z**O-1 (Zona Occludens 1).

Characteristics of the PDZ Domain

- **Length**: Typically consists of 80–90 amino acids.

- **Main Function**: Facilitates protein-protein interactions by recognizing specific peptide sequences, often located at the C-terminal region of other proteins.

- **Biological Role**:
  - Assembling protein complexes at cellular membranes.
  - Cellular signaling.
  - Maintaining cell polarity.
- **Structure**: The PDZ domain has a characteristic structure composed of an antiparallel beta-sheet and two alpha-helices.

Biological Importance

The PDZ domain is essential for:

- The assembly of multiprotein complexes, such as those involved in signal transduction.
- Proper spatial localization of proteins, for instance, in neuronal synapses.
- Maintaining membrane integrity and intercellular communication.

Chapter 2

Modelization of the Interaction between Amino Acids: GNM

It's clear that the interactions between the amino acids that compose the protein are crucial for the folding of the protein into a functional structure. So we need an hamiltonian to describe the system. If the protein is at the equilibrium i will expect that the hamiltonian is a function of a potential which depends from the position every atom that constituted the PDZ specific domain. So my hamiltonian H will be $H = V(\mathbf{r})$, where $\mathbf{r}$ is the vector that contains the position of every atom that constitute the protein. Because this is a complex system we can't solve the problem exactly, so we need to use an approximation.

The second-order approximation of a function $V(\mathbf{r})$, around an equilibrium point $\mathbf{r}_0$, can be written as:

$$V(\mathbf{r}) \approx V(\mathbf{r}_0) + \nabla V(\mathbf{r}_0)^\top(\mathbf{r} - \mathbf{r}_0) + \frac{1}{2}(\mathbf{r} - \mathbf{r}_0)^\top \mathbf{H}_\mathbf{V}(\mathbf{r} - \mathbf{r}_0) \tag{2.1}$$

Where:

- $V(\mathbf{r}_0)$ is the value of the function at the equilibrium point $\mathbf{r}_0$.
- $\nabla V(\mathbf{r}_0)$ is the gradient of the function, defined as:

$$\nabla V(\mathbf{r}) = \left.\frac{\partial V}{\partial \mathbf{r}}\right|_{\mathbf{r}=\mathbf{r}_0} \tag{2.2}$$

  which equals zero if $\mathbf{r}_0$ represents a minimum point, so the equilibrium position.

- $\mathbf{H}_\mathbf{V}$ is the **Hessian matrix** of $V(\mathbf{r})$, defined as:

$$\mathbf{H}_\mathbf{V} = \left.\frac{\partial^2 V}{\partial \mathbf{r}^2}\right|_{\mathbf{r}=\mathbf{r}_0} \tag{2.3}$$

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_1 \partial x_{3N}} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2^2} & \cdots & \frac{\partial^2 V}{\partial x_2 \partial x_{3N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial x_{3N} \partial x_1} & \frac{\partial^2 V}{\partial x_{3N} \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_{3N}^2} \end{bmatrix}.$$

and it is a symmetric matrix that describes the curvature of $V(\mathbf{r})$ around $\mathbf{r}_0$.

It is important to note that the **Hessian matrix $\mathbf{H_V}$** should not be confused with the **Hamiltonian**, which represents the total energy of a system in terms of conjugate coordinates (positions and momenta).

So apart from the constants we can write the hamiltonian as:

$$V(\mathbf{r}) \approx \frac{1}{2}(\mathbf{r}-\mathbf{r}_0)^\top \mathbf{H_V}(\mathbf{r}-\mathbf{r}_0) \tag{2.4}$$

For simplicity, we can consider only the α-carbon atoms because:

- The α-carbon atoms form the backbone of the protein, which defines its overall structure and shape.
- The positions of α-carbon atoms are sufficient to reconstruct the global three-dimensional conformation of the protein.
- Considering only the α-carbons significantly reduces the complexity of the system, lowering the number of degrees of freedom and making calculations more computationally efficient.
- The dynamics of α-carbons often capture the collective motions of the protein, which are crucial for understanding its function (e.g., domain movements, opening of active sites).
- Experimental techniques such as X-ray crystallography and NMR frequently provide high-resolution data specifically for α-carbon atoms, which serve as reference points for modeling (See section). This model is overlall extendaible to all other atoms of the protein.

2.1 Interpretion of the Taylor Expansion: The Gaussian Network Model

I can interpretate equation (2.4) not only as hamiltonian of the armonic oscillator (springs-system) but also like a Network Model where every α-carbon atom is a node that interacts with others node.

Definition of a Graph

A graph G is a mathematical structure used to model pairwise relations between objects. It consists of a set of vertices V (also called nodes) and a set of edges E, where each edge connects a pair of vertices. Formally, a graph is denoted as $G = (V, E)$. Graphs can be directed or undirected, weighted or unweighted, depending on the nature of the relationships they represent.

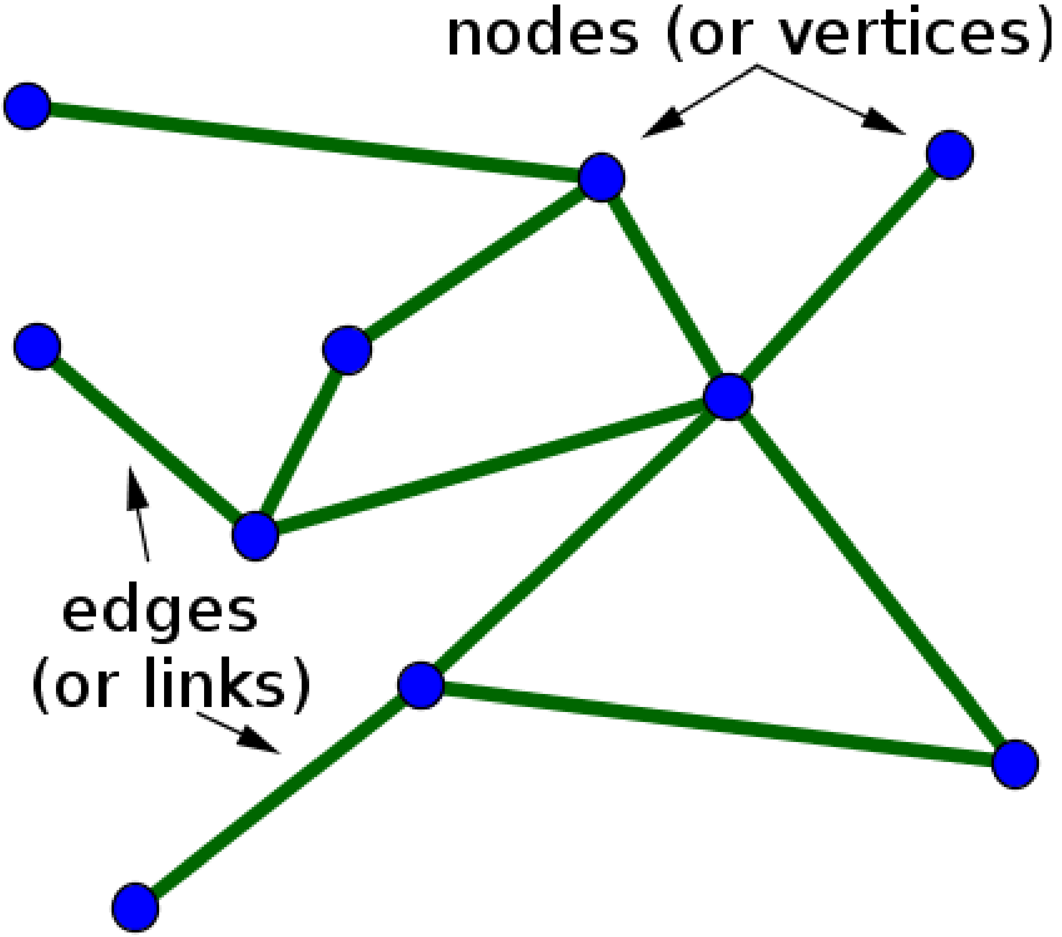


Figure 2.1. Schematic representation of a Network.

Kirchhoff Matrix (Laplacian Matrix)

The Kirchhoff matrix, also known as the graph Laplacian, is a matrix representation of a graph that encodes information about its structure. For a graph G with n vertices, the Laplacian matrix L is defined as:

$$L = D - A \tag{2.5}$$

Here: - D is the degree matrix, a diagonal matrix defined as:

$$D_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{2.6}$$

where $\deg(v_i)$ is the degree of vertex v_i, i.e., the number of edges connected to vertex i.

- A is the adjacency matrix, a square matrix defined as:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

The Laplacian matrix L, derived from D and A, captures essential properties of the graph, such as connectivity and flow. It is widely used in spectral graph theory, network analysis, and machine learning.

Properties of the Degree Matrix

The degree matrix D has the following properties:

- **Diagonal Matrix:** D is diagonal, meaning all off-diagonal elements are zero.
- **Non-Negative Entries:** Each diagonal element d_{ii} represents the degree of vertex i, which is always non-negative.
- **Relation to Graph Order:** For simple graphs, the sum of all diagonal elements of D equals twice the number of edges, i.e., $\text{trace}(D) = 2|E|$.

Properties of the Adjacency Matrix

The adjacency matrix A encodes the edge structure of the graph and has the following properties:

- **Symmetry:** For undirected graphs, A is symmetric, i.e., $a_{ij} = a_{ji}$.
- **Binary Entries:** For unweighted graphs, each element $a_{ij} \in \{0, 1\}$, where 1 indicates the presence of an edge between i and j, and 0 indicates absence.
- **Weighted Graphs:** For weighted graphs, a_{ij} takes the weight of the edge between vertices i and j.
- **Self-Loops:** Diagonal entries a_{ii} indicate self-loops. For simple graphs, $a_{ii} = 0$.
- **Spectral Properties:** The eigenvalues of A provide insights into graph connectivity and other structural properties, such as bipartiteness and clustering.

Mathematical Properties of the Network

The graph Laplacian has several important mathematical properties:

- **Symmetry:** For undirected graphs, L is symmetric and thus diagonalizable.
- **Positive Semi-definiteness:** The Laplacian matrix L is positive semi-definite, meaning that all its eigenvalues are non-negative.
- **Eigenvalues:** The smallest eigenvalue of L is always 0, corresponding to the eigenvector $\mathbf{1}$ (a vector of all ones). The multiplicity of the zero eigenvalue indicates the number of connected components in the graph.
- **Combinatorial Interpretation:** The determinant of the reduced Laplacian matrix (obtained by removing one row and one column) gives the number of spanning trees in the graph.

Inverse of Weighted Laplacian (Pseudo-Inverse)

For a connected graph, the weighted Laplacian L is not invertible due to the zero eigenvalue. However, the Moore-Penrose pseudo-inverse L^+ can be computed. The pseudo-inverse is used in various applications such as electrical network analysis, random walks, and graph-based optimization. It satisfies:

$$L^+L = LL^+ = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

where n is the number of nodes and $\mathbf{1}$ is a vector of all ones.

2.1.1 Gaussian Network Model (GNM)

For what we said before, now we can write the hamiltonian as:

$$H(\mathbf{r}) \approx \frac{1}{2}(\mathbf{r} - \mathbf{r}_0)^\top \mathbf{K}(\mathbf{r} - \mathbf{r}_0) \tag{2.8}$$

Dove $\mathbf{K}$ è la matrice laplaciana. Per semplicità la scriveremo nel seguent modo:

$$H(\mathbf{r}) \approx \frac{1}{2}\mathbf{r}^\top \mathbf{K}\mathbf{r} \tag{2.9}$$

o in componenti:

$$H(\mathbf{r}) \approx \frac{1}{2}\sum_{i_j} \mathbf{r_i}\mathbf{K_{ij}}\mathbf{r_j} \tag{2.10}$$

Dove r è il vettore spostamente rispetto alla posizione di equilibrio di ciascun atomo. Da questa modellizzazione del sistema seguono diverse proprietà e caratteristiche del sistema. Initialy we can write the Probability density at the equilibrium as:

$$P(\mathbf{r}) = \frac{1}{Z}e^{-\beta H(\mathbf{r})} \tag{2.11}$$

$$Z = \int e^{-\beta H(\mathbf{r})}\, d\mathbf{r} \tag{2.12}$$

At this point we can calculate the mean value of the position of the atoms:

Mean Position

The mean position $\langle r \rangle$ is given by:

$$\langle r \rangle = \int r P(r)\, dr, \tag{2.13}$$

where

$$P(r) = \frac{1}{Z}e^{-\frac{\beta}{2} r^\top \mathbf{K} r}, \tag{2.14}$$

and the partition function is:

$$Z = \int e^{-\frac{\beta}{2} r^\top \mathbf{K} r}\, dr. \tag{2.15}$$

Substituting $P(r)$ into the expression for $\langle r \rangle$, we have:

$$\langle r \rangle = \frac{1}{Z} \int r e^{-\dfrac{\beta}{2} r^\top \mathbf{K} r} \, dr. \tag{2.16}$$

The integrand $re^{-\dfrac{\beta}{2} r^\top \mathbf{K} r}$ is symmetric with respect to r, and since there are no linear terms in the Hamiltonian, the Gaussian distribution is centered at $r = 0$. Therefore:

$$\langle r \rangle = 0. \tag{2.17}$$

Covariance

The covariance between r_i and r_j is given by:

$$\mathrm{Cov}(r_i, r_j) = \langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle. \tag{2.18}$$

Since $\langle r_i \rangle = 0$, we have:

$$\mathrm{Cov}(r_i, r_j) = \langle r_i r_j \rangle. \tag{2.19}$$

Using the Boltzmann distribution:

$$P(r) = \frac{1}{Z} e^{-\dfrac{\beta}{2} r^\top \mathbf{K} r}, \tag{2.20}$$

we compute:

$$\langle r_i r_j \rangle = \frac{1}{Z} \int r_i r_j e^{-\dfrac{\beta}{2} r^\top \mathbf{K} r} \, dr. \tag{2.21}$$

For a Gaussian distribution, the covariance matrix is:

$$\mathit{\Sigma} = \beta^{-1} \mathbf{K}^{-1}, \tag{2.22}$$

where $(\mathbf{K}^{-1})_{ij}$ is the (i, j)-th element of $\mathbf{K}^{-1}$. Therefore:

$$\mathrm{Cov}(r_i, r_j) = \frac{1}{\beta} (\mathbf{K}^{-1})_{ij}. \tag{2.23}$$

Derivation of Diffraction Intensity with Atomic Vibrations

The diffraction intensity in crystallography is influenced by the atomic vibrations around their mean positions. These vibrations are characterized by the mean squared displacement, denoted as $\langle u^2 \rangle$. To derive the expression for intensity, we start with the electron density modulated by atomic displacement.

Let the position of an atom be $\mathbf{r}$, with its displacement modeled as a Gaussian distribution around its mean position $\mathbf{r_0}$. The thermal motion leads to a modification of the scattering factor:

$$f(\mathbf{q}) = f_0(\mathbf{q}) \cdot e^{-2\pi i \mathbf{q} \cdot \mathbf{u}}$$

where:

- $f_0(\mathbf{q})$: scattering factor for a stationary atom,
- $\mathbf{q}$: scattering vector,
- $\mathbf{u}$: displacement from the mean position.

The observed intensity $I(\mathbf{q})$ is proportional to the squared magnitude of the structure factor $F(\mathbf{q})$:

$$I(\mathbf{q}) \propto |F(\mathbf{q})|^2,$$

where $F(\mathbf{q})$ is the sum over all atomic contributions:

$$F(\mathbf{q}) = \sum_j f_j(\mathbf{q}) e^{2\pi i \mathbf{q}\cdot\mathbf{r}_j}.$$

Now consider the thermal vibrations. Averaging over the Gaussian distribution of displacements yields a damping factor:

$$\langle e^{-2\pi i \mathbf{q}\cdot\mathbf{u}} \rangle = e^{-2\pi^2 \langle (\mathbf{q}\cdot\mathbf{u})^2 \rangle}.$$

Assuming isotropic vibrations, this simplifies to:

$$\langle (\mathbf{q}\cdot\mathbf{u})^2 \rangle = |\mathbf{q}|^2 \langle u^2 \rangle,$$

where $\langle u^2 \rangle$ is the mean squared displacement. Substituting this back, the intensity is:

$$I(\mathbf{q}) \propto |F(\mathbf{q})|^2 e^{-8\pi^2 \langle u^2 \rangle |\mathbf{q}|^2}.$$

The factor $B = 8\pi^2 \langle u^2 \rangle$ emerges naturally, leading to:

$$I(\mathbf{q}) \propto |F(\mathbf{q})|^2 e^{-B|\mathbf{q}|^2}.$$

Importance of the Beta Factor as a Metric of Model Accuracy

The beta factor (B) is a crucial metric in crystallographic modeling for several reasons:

1. Representation of Atomic Dynamics

The beta factor reflects the thermal vibrations and dynamic behavior of atoms within the crystal. High B values indicate greater atomic mobility or disorder, while low B values suggest rigidity.

2. Assessment of Structural Flexibility

Regions with elevated beta factors often correspond to flexible or disordered parts of the molecule, such as loops, termini, or unstructured regions. These provide insights into the functional dynamics of the macromolecule.

3. Indicator of Model Quality

Unrealistically high or low beta factors across the structure can indicate modeling errors or issues with data quality, such as:

- Poor resolution of the diffraction data.
- Incorrect assignment of atomic positions.
- Misrepresentation of isotropic vs. anisotropic motion.

4. Validation of Refinement

During model refinement, the beta factors are optimized to fit the experimental data. If the refinement process yields consistent and reasonable beta factors, it supports the accuracy of the atomic model.

Conclusion

The beta factor (B) serves as both a physical descriptor of atomic motion and a diagnostic tool for evaluating the quality of crystallographic models. Monitoring B values provides insights into the structure's flexibility and ensures that the model faithfully represents the experimental data.

Model Evaluation: R^2 and MAE

R^2: Coefficient of Determination

The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:

- y_i: observed values,
- $\hat{y}_i$: predicted values,
- $\bar{y}$: mean of the observed values,
- n: total number of data points.

MAE: Mean Absolute Error

The formula for MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where:

- y_i: observed values,
- $\hat{y}_i$: predicted values,
- n: total number of data points.

Evaluation of Beta Factors and Model Quality

In crystallography, the Beta factors (B) are an important metric to assess the quality of a structural model. The value of B is given by:

$$B = 8\pi^2 \langle u^2 \rangle = 8\pi^2 \frac{k_B T}{k_{\text{eff}}}$$

Where:

- $\langle u^2 \rangle$: mean squared atomic displacement,
- k_B: Boltzmann constant,
- T: absolute temperature,
- k_{eff}: effective spring constant (describing atomic rigidity).

Importance of Beta Factors in Model Quality

1. **Thermal Motion Representation**: Beta factors capture the vibrational dynamics of atoms. Consistent B values across the model indicate reliable representation of the thermal motion.

2. **Detection of Disordered Regions**: High B values typically correspond to disordered or flexible regions in the structure (e.g., loops or termini).

3. **Model Refinement**: During refinement, fitting experimental data should yield realistic B values. Overly high or low B values may indicate incorrect atomic positions or errors in data processing.

4. **Outlier Detection**: A plot of B values versus residue number can reveal anomalies or poorly modeled regions in the structure.

Combining Metrics for Model Assessment

To fully evaluate the quality of a model:

- Use R^2 or a similar metric to evaluate how well the predicted data fits the observed data (e.g., intensities in crystallography).
- Use MAE or RMSD to assess the absolute error between predictions and observations.
- Analyze B factors to validate structural consistency and detect regions of disorder or flexibility.

Conclusion

Combining statistical measures such as R^2 and MAE with physical metrics like Beta factors provides a comprehensive evaluation of the model's accuracy and reliability.

2.2 Understanding and Causality of allosteric mechanisms in proteins

The correlation not only avoid to test the quality of the model through the beta factors but also avoid to understand the interactions between nodes. Our goal is to understand what is the allosteric site, that is the site from wich the signal is transmitted to the other sites of the protein. As it often said, correlation does not imply causation. Causality refers to the relationship between causes and effects, where one event (the cause) directly influences or produces another event (the effect). Formally, causality can be defined as follows:

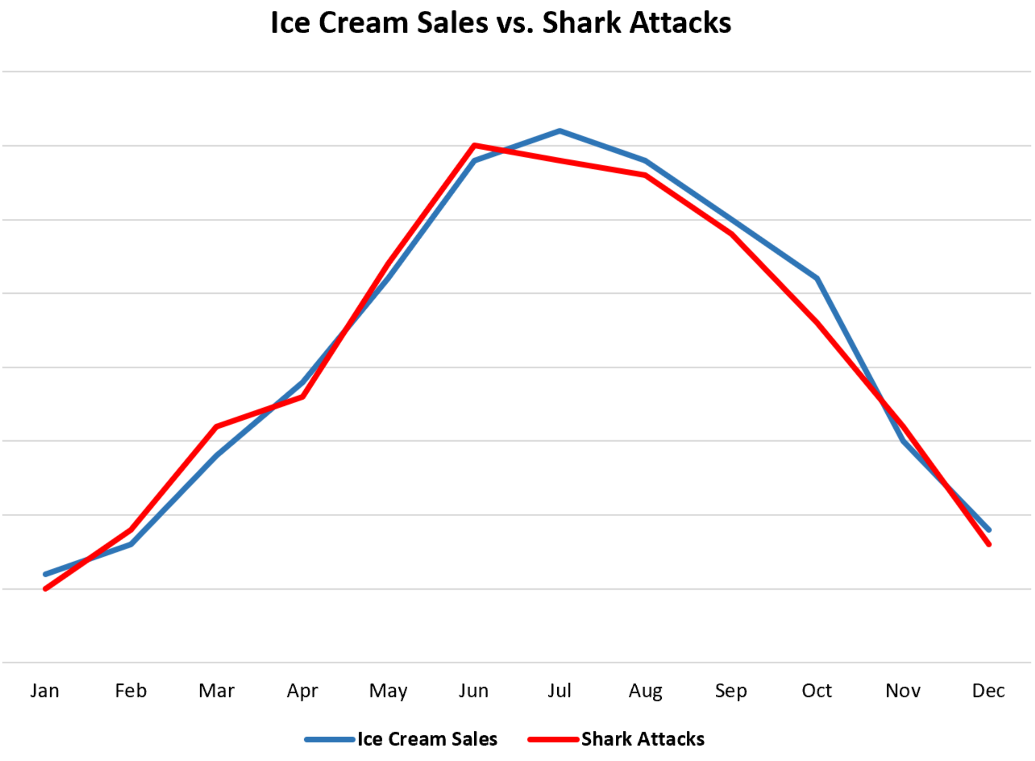


Figure 2.2. Correlation is not causation.

> A phenomenon A is said to be the cause of phenomenon B if the occurrence or change in A leads to a corresponding occurrence or change in B, provided that all other conditions remain constant.

Causality is a foundational concept in various fields, including physics, philosophy, and statistics. It implies a directional influence, where the cause precedes the effect in time and is necessary, sufficient, or contributory for the effect to occur.

In mathematical terms, causal relationships can often be expressed using models such as:

$$B = f(A, \text{other factors})$$

where f is a functional relationship capturing how A and other factors jointly determine B. In experimental contexts, causality is typically established through controlled interventions and the observation of corresponding changes.

Deterministic and Stochastic Causality

Causality can be broadly classified into two types: deterministic and stochastic. In **deterministic causality**, a specific cause invariably leads to a specific effect. Mathematically, if an event A causes an event B, then B will always occur whenever A occurs. This relationship can be represented as $A \implies B$, meaning that A is both necessary and sufficient for B. Deterministic causality is commonly observed in systems governed by precise physical laws, such as the equations of motion in classical mechanics, where given initial conditions and forces, the trajectory of an object is entirely predictable.

In contrast, **stochastic causality** acknowledges that many causal relationships are probabilistic rather than absolute. Here, the occurrence of A increases the likelihood of B, but B does not necessarily follow A. This can be expressed probabilistically as $P(B|A) > P(B|\neg A)$, indicating that A raises the probability of B. Stochastic causality is particularly prevalent in complex systems, such as biology or social sciences, where multiple factors interact, and outcomes are influenced by inherent randomness or uncertainty. For example, smoking increases the probability of lung cancer, but not all smokers develop cancer.

Both deterministic and stochastic causality play crucial roles in understanding natural phenomena. While deterministic causality provides precise predictive power, stochastic causality is essential for modeling systems where randomness and variability are inherent. Together, these concepts offer a comprehensive framework for analyzing causal relationships across disciplines.

Geometric Perspective of Causality and Shannon Entropy

When all dimensions or variables of the system are fully known, the uncertainty about future states vanishes, and the system transitions into a deterministic regime.

This idea can be linked to Shannon entropy, defined as:

$$H(X) = -\sum_i P(x_i) \log(P)(x_i),$$

where X is a random variable with possible states x_i and $P(x_i)$ is the probability of x_i. When $H(X) = 0$, the system has no uncertainty, indicating complete predictability of the state X.

In a deterministic system, knowing all dimensions eliminates ambiguity, allowing the precise prediction of future states. However, in practice:

- Complex systems often exhibit chaotic or stochastic behavior, making predictions challenging even if the system is deterministic.
- Non-linear dynamics can amplify small uncertainties, leading to an apparent increase in unpredictability.

Thus, while $H(X) = 0$ implies determinism in theory, practical limitations and system complexities must be considered in real-world applications.

To understand the causality of allosteric mechanisms in proteins, we need to investigate the underlying physical interactions and pathways that transmit signals between distant sites. For theses reasons we need also another indicators:

2.2.1 Rensponse Function

Linear Response Theorem (Positions)

The linear response theorem, expressed in terms of the positions $\mathbf{r}_i$, describes how the average position of a particle i responds to a perturbation applied to parameter j. The response is given by:

$$R_{ij} = \frac{\partial \langle \mathbf{r}_i \rangle}{\partial f_j}$$

Where: - $\mathbf{r}_i$ is the position of particle i relative to its equilibrium position. - f_j is an external force applied to particle j. - $\langle \mathbf{r}_i \rangle$ is the average position of particle i.

Relation to Correlations

The response can also be expressed in terms of the correlations between the fluctuations in the positions of particles i and j:

$$R_{ij} = \beta \langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle$$

Where: - $\delta \mathbf{r}_i = \mathbf{r}_i - \langle \mathbf{r}_i \rangle$ is the fluctuation of the position from its average value. - $\beta = \frac{1}{k_B T}$ is the inverse thermal energy, with k_B being the Boltzmann constant and T the temperature. - $\langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle$ is the scalar product of the fluctuations in the positions of particles.

Physical Interpretation

This result states that the linear response of the system to a perturbation is proportional to the thermal correlation between the fluctuations of the particles' positions. It reflects the connection between the equilibrium properties of the system and its dynamic response.

Why is Linear Response an Indicator of Causality?

The linear response R_{ij} can be interpreted as an indicator of causality because it measures the direct influence of a change in parameter j (e.g., an external force) on variable i (e.g., position, velocity, or another observable). This connection arises for the following reasons:

- **Direct cause-effect relationship:** The response R_{ij} explicitly quantifies how a perturbation f_j applied to j influences the behavior of i. This reflects causality since the perturbation precedes the response.
- **Asymmetry of the response:** Causality implies asymmetry: if j causes a change in i, the reverse (i.e., i influencing j) does not necessarily occur. In the linear response framework, R_{ij} specifically measures the influence of j on i.
- **Relation to equilibrium correlations:** The proportionality $R_{ij} = \beta\langle\delta\mathbf{r}_i \cdot \delta\mathbf{r}_j\rangle$ implies that the equilibrium correlations encode information about the system's ability to propagate changes. This reinforces the causal interpretation.

In summary, the linear response is a quantitative framework to understand how perturbations propagate through a system, reflecting causal influences between different components.

Transfer Entropy: Formula, Proof, and Causality

Definition of Transfer Entropy

The *Transfer Entropy* (TE) measures the directional flow of information from a source variable x_j to a target variable x_i. It quantifies how much the knowledge of past states of x_j improves the prediction of future states of x_i, beyond what is already provided by the past states of x_i itself. The TE is defined as:

$$TE_{j\to i}(t) = H[x_i(t+\tau) \mid x_i(\tau)] - H[x_i(t+\tau) \mid x_i(\tau), x_j(\tau)]$$

Where:

- $H[a \mid b]$: Conditional Shannon entropy of variable a given b.
- $x_i(t+\tau)$: State of x_i at time $t+\tau$.
- $x_i(\tau), x_j(\tau)$: States of x_i and x_j at time τ.

Transfer Entropy for Gaussian Systems

For stationary Gaussian processes, the Transfer Entropy can be computed analytically. Using the covariance matrices of the processes, the TE is given by:

$$TE_{j\to i}(t) = -\frac{1}{2}\ln\Bigg(\,) \, (1 - \frac{\alpha_{ij}(t)}{\beta_{ij}(t)}\Bigg)$$

Where:

$$\alpha_{ij}(t) = [C_{ii}(0)C_{ij}(t) - C_{ij}(0)C_{ii}(t)]^2$$
$$\beta_{ij}(t) = [C_{ii}(0)C_{jj}(0) - C_{ij}^2(0)][C_{ii}^2(0) - C_{ii}^2(t)]$$

Here, $C_{ij}(t)$ represents the time-lagged cross-correlation between variables x_i and x_j, while $C_{ii}(0)$ and $C_{jj}(0)$ are the variances of x_i and x_j, respectively.

Proof Sketch for Gaussian TE

The derivation relies on the Gaussian property that conditional entropies can be expressed in terms of covariance matrices. For a system described by variables $x_i(t)$, $x_i(0)$, and $x_j(0)$, the covariance matrix is:

$$\mathit{\Omega} = \begin{bmatrix} C_{ii}(0) & C_{ii}(t) & C_{ij}(t) \\ C_{ii}(t) & C_{ii}(0) & C_{ij}(0) \\ C_{ij}(t) & C_{ij}(0) & C_{jj}(0) \end{bmatrix}$$

The conditional covariance matrices are computed to find $TE_{j\to i}$ based on the definition of conditional entropy. After algebraic manipulation, the formula for Gaussian TE above is obtained.

Causality and Transfer Entropy

Transfer Entropy is a powerful indicator of causality for the following reasons:

– **Directionality:** TE is asymmetric ($TE_{j\to i} \neq TE_{i\to j}$), capturing the directional flow of information from x_j to x_i.

– **Temporal dependence:** By using time-lagged variables, TE inherently respects the time ordering required for causation.

– **Beyond correlation:** While correlations measure symmetric associations, TE specifically quantifies the influence of one variable on another, conditioned on their histories.

Comparison with Correlation

- **Symmetry:** Correlation is symmetric ($C_{ij} = C_{ji}$), whereas TE is directional.
- **Causal inference:** Correlation measures only association, which can be spurious. TE identifies directed influences and distinguishes cause from effect.
- **Applications:** TE is particularly useful in systems with nonlinear dynamics, where correlations may fail to capture the true dependencies.

Chapter 3

Stochastic process

By definition we can associate a stochastic process to a set of random variables that evolve over time. The evolution of the process is governed by probabilistic laws, and each random variable represents the state of the system at a specific time point. In general we have:

$$\frac{dX_t}{dt} = -\nabla H(X_t) + \eta_t, \tag{3.1}$$

dove:

- X_t rappresenta lo stato del sistema al tempo t,
- $H(X_t)$ è la Hamiltoniana del sistema, che governa la dinamica deterministica,
- $-\nabla H(X_t)$ è il gradiente della Hamiltoniana (che fornisce la direzione di maggiore variazione),
- η_t è un termine di rumore stocastico (spesso modellato come un rumore bianco gaussiano con varianza σ^2).

So for our specific hamiltonian we can write:

$$H(\mathbf{r}) \approx \frac{1}{2} \sum_{i,j} \mathbf{r}_i \mathbf{K}_{ij} \mathbf{r}_j, \tag{3.2}$$

the gradient with respect to $\mathbf{r}$ is given by:

$$\nabla H(\mathbf{r}) = \mathbf{K}\mathbf{r}, \tag{3.3}$$

where $\mathbf{K}$ is the matrix of coefficients $\mathbf{K}_{ij}$.

Vectorial Form

The stochastic process dynamics in vectorial form are:

$$\frac{d\mathbf{r}_t}{dt} = -\mathbf{K}\mathbf{r}_t + \boldsymbol{\eta}_t, \tag{3.4}$$

where:

- $\mathbf{r}_t$ is the state vector at time t,
- $-\mathbf{K}\mathbf{r}_t$ is the deterministic term derived from the Hamiltonian,
- $\boldsymbol{\eta}_t$ is a stochastic noise vector (e.g., components are independent and Gaussian distributed).

In component form, the system is written as:

$$\frac{d\mathbf{r}_{i,t}}{dt} = -\sum_j \mathbf{K}_{ij}\mathbf{r}_{j,t} + \eta_{i,t}, \tag{3.5}$$

where:

- i denotes the index of the vector component,
- $\mathbf{K}_{ij}$ is the element of the matrix $\mathbf{K}$,
- $\mathbf{r}_{j,t}$ is the j-th component of the state vector at time t,
- $\eta_{i,t}$ is the noise term associated with the i-th component.

The analytical solution, assuming an initial condition $\mathbf{r}_0$, is:

$$\mathbf{r}_t = e^{-\mathbf{K}t}\mathbf{r}_0 + \int_0^t e^{-\mathbf{K}(t-s)}\boldsymbol{\eta}_s \, ds. \tag{3.6}$$

In this expression:

- $e^{-\mathbf{K}t}$ is the matrix exponential of $-\mathbf{K}t$,
- The first term, $e^{-\mathbf{K}t}\mathbf{r}_0$, represents the deterministic evolution,
- The second term, $\int_0^t e^{-\mathbf{K}(t-s)}\boldsymbol{\eta}_s \, ds$, accounts for the stochastic contributions from the noise.

The solution combines deterministic decay driven by $-\mathbf{K}\mathbf{r}_t$ and stochastic fluctuations due to $\boldsymbol{\eta}_t$.

The correlation function is defined as:

$$\langle \mathbf{r}_t \mathbf{r}_\tau^\top \rangle. \tag{3.7}$$

Substituting the solution and simplifying, we obtain:

$$\langle \mathbf{r}_t \mathbf{r}_\tau^\top \rangle = e^{-\mathbf{K}t} \langle \mathbf{r}_0 \mathbf{r}_0^\top \rangle e^{-\mathbf{K}^\top \tau} + \int_0^{\min\{(\}t,\tau)} e^{-\mathbf{K}(t-s)}\mathbf{Q}e^{-\mathbf{K}^\top(\tau-s)} \, ds. \tag{3.8}$$

In this expression:

- $e^{-\mathbf{K}t}$ is the matrix exponential,
- $\langle \mathbf{r}_0 \mathbf{r}_0^\top \rangle$ is the initial covariance matrix of $\mathbf{r}_0$,
- $\mathbf{Q}$ is the noise covariance matrix.

Response Function and Transfer Entropy

Response Function

The response function $R_{ij}(t)$ is defined as:

$$R_{ij}(t) = \frac{C_{ij}(t)}{C_{ij}(0)}, \tag{3.9}$$

where the correlation $C_{ij}(t)$ is given by the analytical solution:

$$C_{ij}(t) = \int_0^t \int_0^t e^{-\mathbf{K}(t-s)} \mathbf{Q} e^{-\mathbf{K}^\top (t-u)} \, ds \, du, \tag{3.10}$$

and $C_{ij}(0)$ is the initial correlation:

$$C_{ij}(0) = \langle \mathbf{r}_{i,0} \mathbf{r}_{j,0} \rangle. \tag{3.11}$$

Thus, the response function becomes:

$$R_{ij}(t) = \frac{\int_0^t \int_0^t e^{-\mathbf{K}(t-s)} \mathbf{Q} e^{-\mathbf{K}^\top (t-u)} \, ds \, du}{C_{ij}(0)}. \tag{3.12}$$

Transfer Entropy

The transfer entropy $TE_{j \to i}(t)$ is defined as:

$$TE_{j \to i}(t) = -\frac{1}{2} \ln \left(\,) \, (1 - \frac{\alpha_{ij}(t)}{\beta_{ij}(t)} \right), \tag{3.13}$$

where:

$$\alpha_{ij}(t) = \left[C_{ii}(0) C_{ij}(t) - C_{ij}(0) C_{ii}(t) \right]^2,$$

$$\beta_{ij}(t) = \left[C_{ii}(0) C_{jj}(0) - C_{ij}^2(0) \right] \left[C_{ii}^2(0) - C_{ii}^2(t) \right].$$

Substituting Correlation Expressions

1. The diagonal correlations are:

$$C_{ii}(t) = \int_0^t \int_0^t e^{-\mathbf{K}(t-s)} \mathbf{Q}_{ii} e^{-\mathbf{K}^\top (t-u)} \, ds \, du. \tag{3.14}$$

2. The off-diagonal correlations are:

$$C_{ij}(t) = \int_0^t \int_0^t e^{-\mathbf{K}(t-s)} \mathbf{Q}_{ij} e^{-\mathbf{K}^\top (t-u)} \, ds \, du. \tag{3.15}$$

Substituting these into $\alpha_{ij}(t)$ and $\beta_{ij}(t)$:

$$\alpha_{ij}(t) = \Big[\langle \mathbf{r}_{i,0}^2 \rangle \cdot C_{ij}(t) - C_{ij}(0) \cdot C_{ii}(t)\Big]^2 ,$$

$$\beta_{ij}(t) = \Big[\langle \mathbf{r}_{i,0}^2 \rangle \langle \mathbf{r}_{j,0}^2 \rangle - C_{ij}^2(0)\Big] \cdot \Big[\langle \mathbf{r}_{i,0}^2 \rangle^2 - C_{ii}^2(t)\Big] .$$

Finally, substituting these into the transfer entropy expression:

$$TE_{j \to i}(t) = -\frac{1}{2} \ln \left( )\,(1 - \frac{\Big[\langle \mathbf{r}_{i,0}^2 \rangle \cdot C_{ij}(t) - C_{ij}(0) \cdot C_{ii}(t)\Big]^2}{\Big[\langle \mathbf{r}_{i,0}^2 \rangle \langle \mathbf{r}_{j,0}^2 \rangle - C_{ij}^2(0)\Big] \cdot \Big[\langle \mathbf{r}_{i,0}^2 \rangle^2 - C_{ii}^2(t)\Big]} \right) . \quad (3.16)$$

Detailed Presentation of Dataset 3LNX

The Protein Data Bank entry **2M0Z** represents the *cis* conformation of the same photoswitchable PDZ domain as in 2M10. The engineered domain incorporates the same mutations and azobenzene derivative, enabling light-dependent switching between the *cis* and *trans* states. This conformational state is achieved following illumination with UV light, which induces a structural rearrangement by altering the isomeric state of the azobenzene moiety. The result is a significant shift in the protein's structural and functional properties.

Like 2M10, the structure of 2M0Z was determined using solution-phase NMR spectroscopy. The dataset provides a set of conformations reflecting the structural ensemble of the protein in the *cis* state. By comparing this dataset to 2M10, researchers can identify specific structural changes induced by the photoswitching mechanism. This includes alterations in the binding groove, changes in side-chain orientations, and the dynamics of the azobenzene-protein interaction.

The *cis* state is characterized by a more compact structure compared to the *trans* state, with potential implications for protein-protein interactions and ligand binding. The B-factor data included in the dataset further elucidate regions of structural flexibility and rigidity, providing a comprehensive view of the dynamic behavior of the domain. Together with 2M10, this dataset serves as a cornerstone for understanding light-induced conformational dynamics and their functional implications in engineered proteins.

Alpha-carbon backbone of 2M0Z with B-factors visualized, highlighting the structural properties of the photoswitchable PDZ domain in its *trans* state. Structural Highlights: - **Overall Structure**: The PDZ domain is composed of a core structure containing several β-strands (visualized as flat ribbons) and α-helices (visualized as curved ribbons). This topology forms the typical PDZ fold, which includes a ligand-binding groove that is crucial for protein-protein interactions. - **Atomic Representation**: - The atoms in the structure are colored according to their type: oxygen (red), nitrogen (blue), carbon

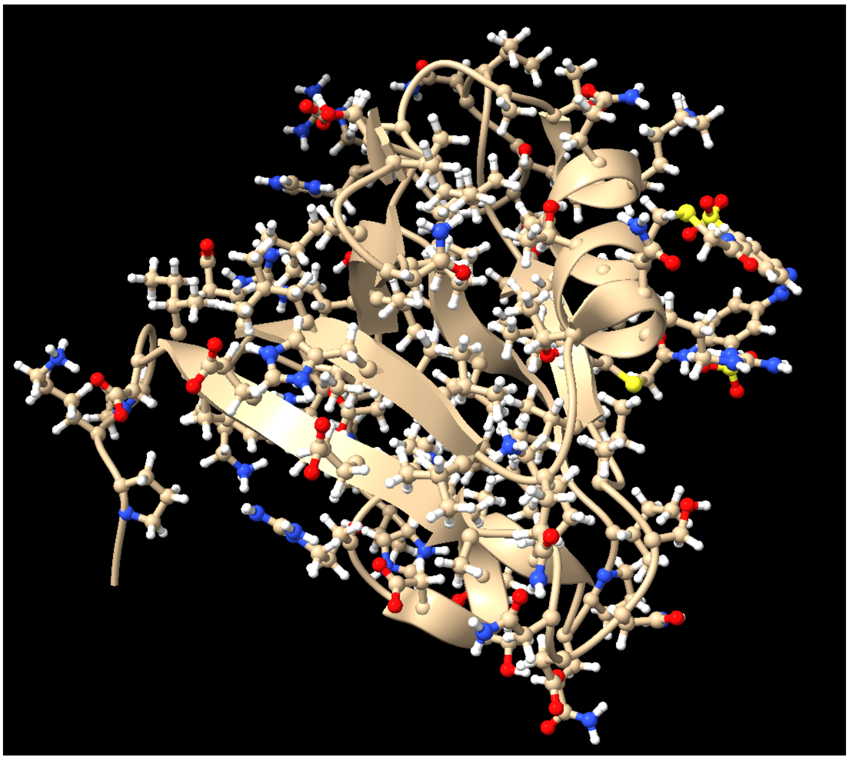
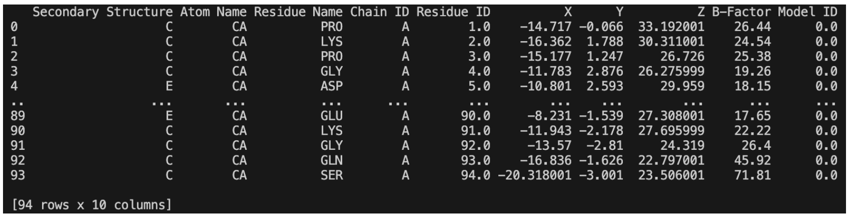


Figure 3.1



```
   Secondary Structure Atom Name Residue Name Chain ID Residue ID          X      Y          Z B-Factor Model ID
0                    C        CA          PRO        A        1.0    -14.717 -0.066  33.192001    26.44      0.0
1                    C        CA          LYS        A        2.0    -16.362  1.788  30.311001    24.54      0.0
2                    C        CA          PRO        A        3.0    -15.177  1.247     26.726    25.38      0.0
3                    C        CA          GLY        A        4.0    -11.783  2.876  26.275999    19.26      0.0
4                    E        CA          ASP        A        5.0    -10.801  2.593     29.959    18.15      0.0
..                 ...       ...          ...      ...        ...        ...    ...        ...      ...      ...
89                   E        CA          GLU        A       90.0     -8.231 -1.539  27.308001    17.65      0.0
90                   C        CA          LYS        A       91.0    -11.943 -2.178  27.695999    22.22      0.0
91                   C        CA          GLY        A       92.0     -13.57  -2.81     24.319     26.4      0.0
92                   C        CA          GLN        A       93.0    -16.836 -1.626  22.797001    45.92      0.0
93                   C        CA          SER        A       94.0 -20.318001 -3.001  23.506001    71.81      0.0

[94 rows x 10 columns]
```

Figure 3.2. Dataset of 3LNX.

(beige), and hydrogen (white). The yellow feature represents a functional group or modification, likely the azobenzene derivative responsible for the photoswitching behavior of this protein. - The molecular backbone (carbon-alpha trace) is visualized as a beige ribbon, representing the connectivity of the primary structure. - **Functional Elements**: The ligand-binding groove is clearly visible in this conformation, highlighting the open state of the protein that allows potential interaction with partner molecules.

Photoswitching Properties: The domain includes a covalently attached azobenzene derivative that acts as a molecular photoswitch. In this figure, the domain is locked in its *trans* state, representing the protein's conformation under standard conditions. The azobenzene modification plays a crucial role in mediating light-induced conformational changes between *trans* and *cis* states, altering the structural properties of the binding groove.

Visualization Purpose: This figure provides a clear view of the protein's structure, with all residues shown in atomic detail. The visualization can be used to analyze: - The structural stability and flexibility of different regions of the protein, particularly in conjunction with B-factor data. - The interaction sites for ligands and how their accessibility changes upon photoswitching.

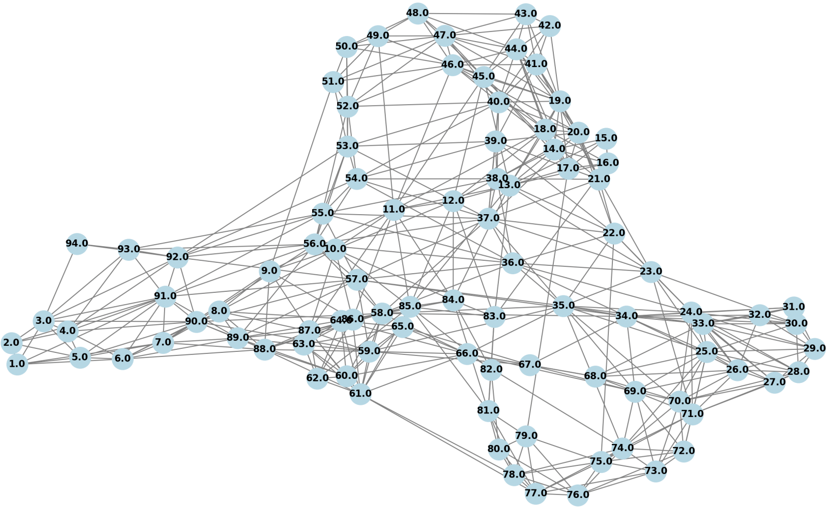


Figure 3.3. Graph of 3LNX.

Connection Radius

The connection radius is a critical parameter in the construction of the Kirchhoff matrix, as it determines the range of interactions considered between residues. By setting an appropriate connection radius, we can capture both local interactions between neighboring residues and long-range interactions that contribute to the protein's stability and dynamics.

Kirchhoff Matrix

Applying (2.5) to the protein structure, we obtain the Kirchhoff matrix $\mathbf{K}$ for the system. This matrix captures the connectivity between residues and provides insights into the protein's structural dynamics. We choose to put the value of the link between two native pair (near neighbor), equal to 20, otherwise the link is equal to 1. So: Let $G = (V, E)$ be a graph where V is the set of vertices and E is the set of edges. Define the weight of the link between two vertices i and j as $w(i, j)$.

We assign weights as follows:

$$w(i,j) = \begin{cases} 20, & \text{if } i \text{ and } j \text{ are near neighbors,} \\ 1, & \text{otherwise.} \end{cases}$$

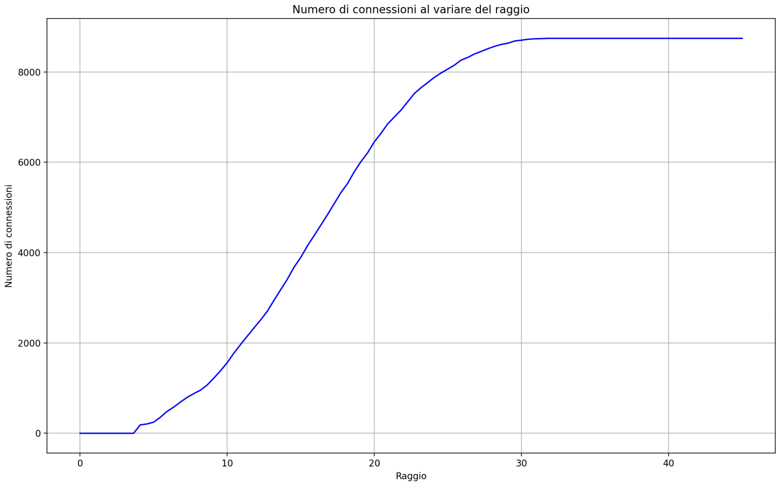


Figure 3.4. Dataset of 2M0Z.

Justification of Weights

The chosen weights are based on the relative importance of interactions in the context of protein stability and structural dynamics:

Near-neighbor interactions (native pairs): Physical Basis: Near neighbors are typically connected by strong covalent bonds or stabilized by dense non-covalent interactions (e.g., van der Waals forces, hydrogen bonds). These interactions dominate the local stability and rigidity of the protein structure.

Quantitative Support: The energy contributions from covalent bonds (300–400 kJ/mol) and closely packed van der Waals interactions are significantly higher than those from distant or weaker interactions.

Weight Selection: By assigning a weight of 20, we prioritize the influence of these dominant interactions in the Kirchhoff matrix. This ensures that the matrix reflects the critical role of local connectivity in determining the protein's overall structural stability.

Distant interactions: Physical Basis: Residues that are not near neighbors may still interact through long-range forces (e.g., electrostatic or weak van der Waals forces). However, these interactions are generally less frequent and contribute less to the rigidity of the protein.

Weight Selection: Assigning a lower weight of 1 acknowledges the reduced importance of these interactions while still accounting for their presence in the protein network. We notice first that as expected the Kirchhoff matrix is symmetric and that the diagonal elements are all negative. This is consistent with the physical interpretation of the matrix, where the diagonal elements represent the sum of the coupling strengths between a residue and all other

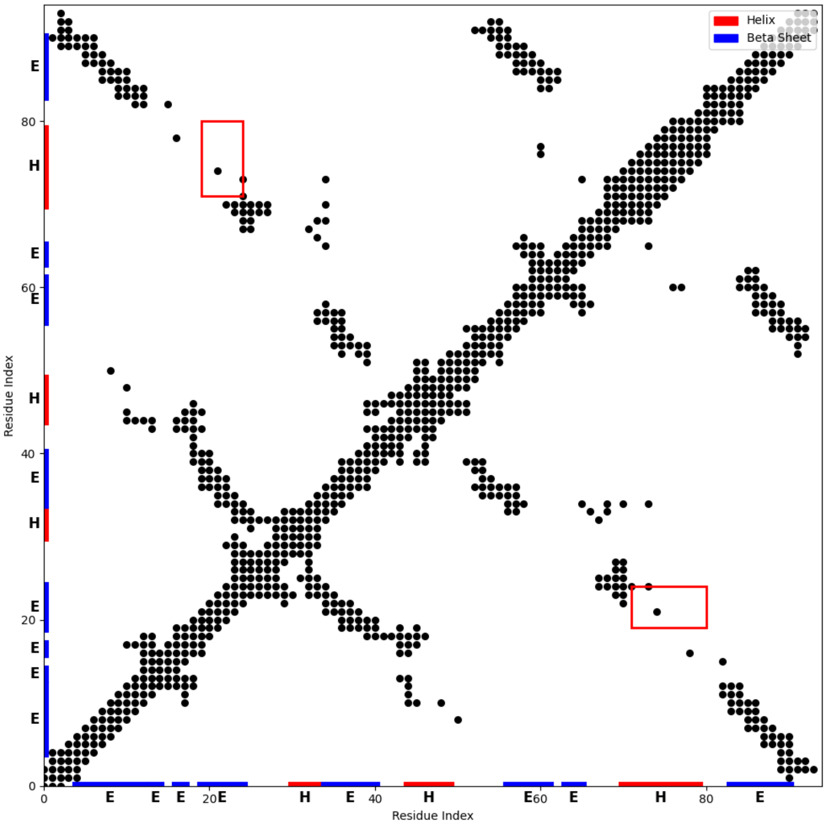


Figure 3.5. Kirchoff Matrix.

residues in the protein. The negative sign indicates that the interactions are stabilizing, as expected in a protein structure. The off-diagonal elements represent the coupling between pairs of residues, reflecting the network of interactions that stabilize the protein's structure. It is important to notice also that most of the links are between near residues, but we have also some cluster of links between distant residues. This is important because it means that the protein is not a simple chain of residues but it is a complex network of interactions.

Correlation Matrix

The two figures below illustrate the correla tion matrices obtained from the protein structure. The elements of the matrix represent correlations between residues i and j, derived from the dynamics and structure of the system. Overlaying the secondary structure (Helices: red; Beta Sheets: blue) helps to contextualize the correlations within structural motifs. The Kirchhoff matrix was used to compute these correlations.

Figure Analysis:

- **First Plot: Correlation Matrix (NaNs handled as False):**
  - The red box highlights a region of strong local correlations, likely associated with tightly coupled residues within a helix or beta sheet structure.
  - The light background suggests weak or negligible correlations outside the main diagonal and the highlighted regions, reflecting weaker interactions or dynamic decoupling.
  - The diagonal dominance indicates strong correlations between residues within the same structural domain.
- **Second Plot: Correlation Matrix (NaNs handled as True):**
  - The overall structure appears more diffuse, with non-local correlations extending further from the diagonal.
  - The red box again highlights a region of strong correlations, consistent with a structural motif (helix or beta sheet).
  - The increased blue intensity in some regions may suggest the inclusion of weak, long-range interactions due to the handling of missing values (NaNs as True).
  - Differences in the correlation intensity compared to the first plot reflect the impact of NaN handling on the computation of the correlation matrix.

Relation to the Kirchhoff Matrix: The Kirchhoff matrix directly influences the correlation matrices by encoding the connectivity between residues.

- The strong correlations in the highlighted regions correspond to residues with high connectivity in the Kirchhoff matrix (near neighbors with weights of 20).
- Weaker correlations in the off-diagonal regions are consistent with weaker weights (1) assigned to distant interactions in the Kirchhoff matrix.

Beta Factors

The plot above compares the experimental B-factors (blue curve) with the predicted B-factors (red curve) along the residue index. The B-factors, which represent the atomic displacement or flexibility, are a critical measure of the dynamic behavior of the protein structure.

Analysis:

- **Experimental B-factors (blue curve):**
  - The blue curve shows regions of higher flexibility (peaks) and rigidity (troughs) as derived from experimental data.

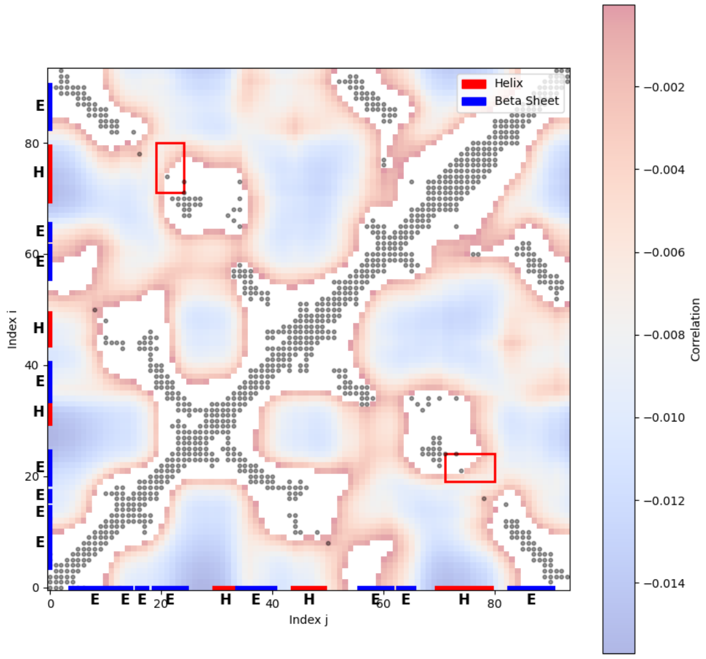


Figure 3.6. Dataset of 2M0Z.

  - ∗ Peaks in the B-factors align with regions of loops or exposed residues, typically more flexible regions.
  - ∗ Troughs align with structured regions such as helices or beta sheets, which are inherently more rigid.

- **Predicted B-factors (red curve):**
  - ∗ The predicted B-factors generally follow the experimental trend, demonstrating the effectiveness of the model in capturing the dynamics of the protein.
  - ∗ Discrepancies between the predicted and experimental curves are visible in certain regions, particularly near residues with extreme flexibility or rigidity. These deviations may be due to limitations in the Kirchhoff matrix approximation or oversimplified assumptions in the modeling process.

Secondary Structure Context:

- The background shading (red and blue bands) indicates secondary structure elements:
  - ∗ **Red regions:** Helices, characterized by lower B-factors and greater rigidity.

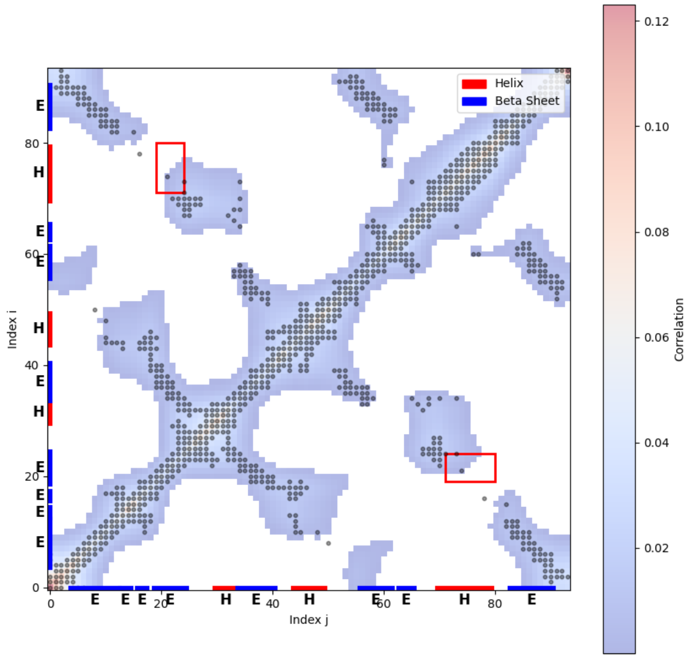


Figure 3.7. Dataset of 2M0Z.

  - ∗ **Blue regions:** Beta sheets, which also exhibit lower B-factors compared to loops but slightly higher flexibility than helices.
- The alignment of peaks and troughs in the B-factor curves with these shaded regions provides structural context to the observed dynamics.

Model Implications:

- The good agreement between experimental and predicted B-factors in structured regions (helices and beta sheets) supports the validity of the Kirchhoff matrix-based model.
- Discrepancies in loop regions suggest that additional factors, such as long-range interactions or local disorder, might need to be incorporated into the model for improved accuracy.

This analysis highlights the capability of the Kirchhoff matrix to predict protein flexibility while identifying areas for potential refinement.

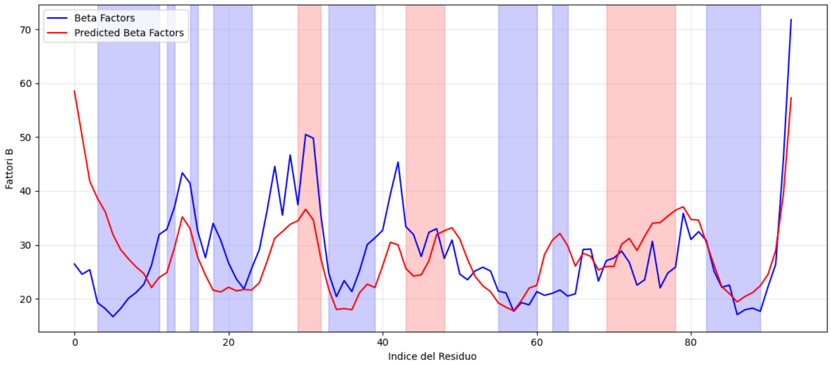


Figure 3.8. Beta factors.

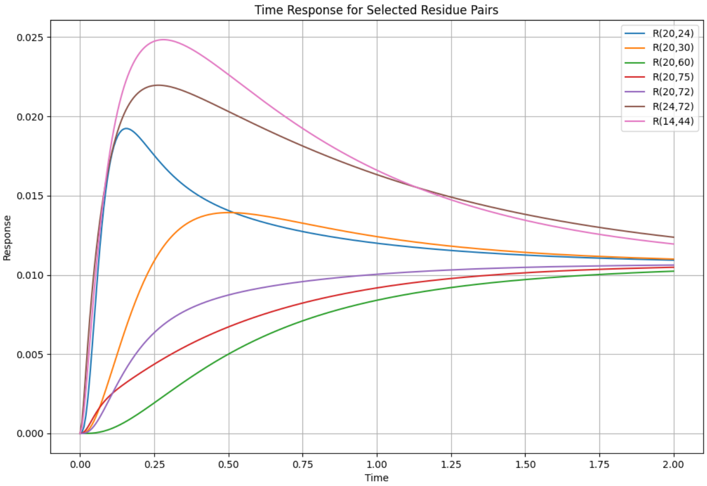


Figure 3.9. Responses in time.

Respones in Time

Tau

Mean First time passage

Time indicators

Graphic Analysis of Secondary Structures

The figures provide a detailed view of the dynamic properties of the protein structure, focusing on the beta sheet around residue 20 and the alpha helix near residue 75. These secondary structures exhibit distinct dynamical behaviors,

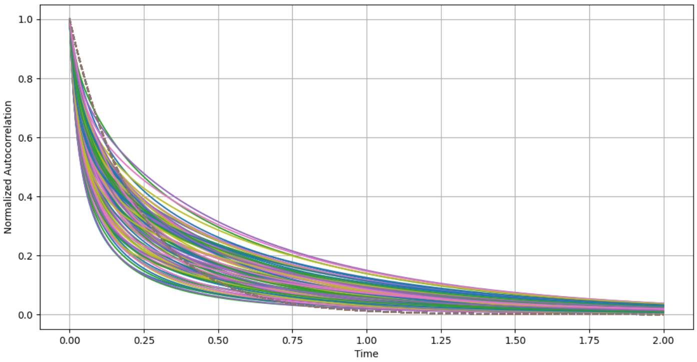


Figure 3.10. Responses in time.

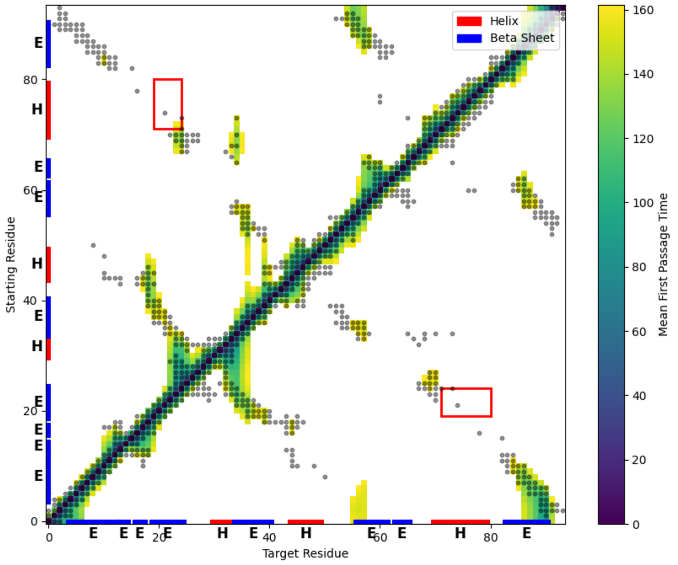


Figure 3.11. Responses in time.

which are quantified using metrics such as inverse first mean time, correlation, response, and transfer entropy.

Figure 1: Inverse First Mean Time

- **Beta Sheet Region (Residue 20):** The beta sheet exhibits a pronounced peak in the inverse first mean time. This indicates a region of lower flexibility and slower dynamics, as the stabilizing hydrogen bonding network within the beta sheet reduces the probability of transitions to other dynamic states. The peak signifies that this region serves as an anchor point in the protein's dynamic landscape, contributing to the

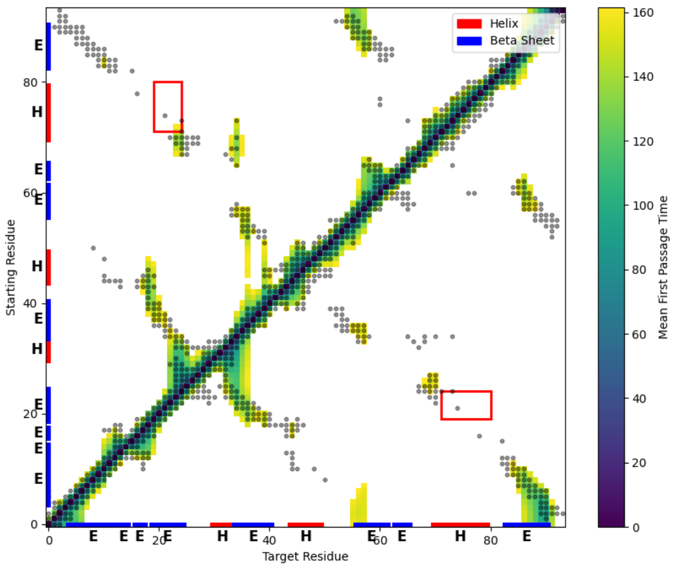


Figure 3.12. Responses Matrix.

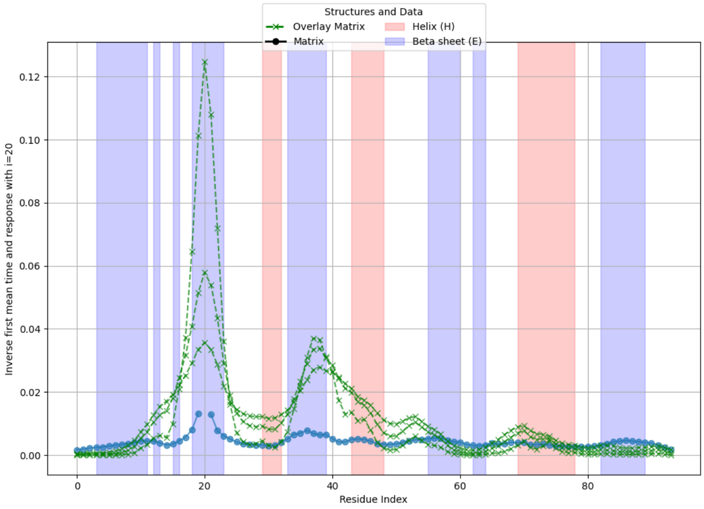


Figure 3.13. Inverse first mean time perturbating 21-th residue.

overall structural stability.

– **Alpha Helix Region (Residue 75):** The helix region has a relatively flat and lower inverse first mean time, reflecting faster dynamics and increased local adaptability. The helix's hydrogen bonds, while stabilizing, allow for greater flexibility along the helical axis, resulting in a quicker mean response time.

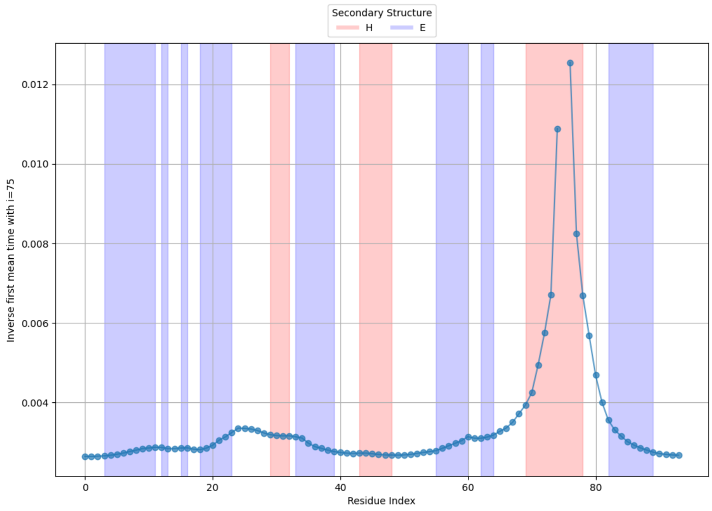


Figure 3.14. Inverse first mean time perturbating 76-th residue.

- **Interpretation:** The contrasting dynamics between the beta sheet and the helix are evident in this figure. The beta sheet is a rigid structural element, whereas the helix contributes flexibility and adaptability to the protein. This dynamic complementarity supports the protein's ability to maintain structural integrity while facilitating conformational changes required for its function.

Figure 2: Correlation with Residue 20

- **Beta Sheet Region (Residue 20):** Strong correlation values are observed in this region, reflecting the tightly coupled dynamics among the residues in the beta sheet. These high correlations are a direct consequence of the stabilizing hydrogen bonds that link the beta strands, enforcing collective motion within this region.
- **Alpha Helix Region (Residue 75):** Moderate correlation values in the helix region suggest that residues near residue 75 are influenced by, but less directly connected to, the beta sheet. The helix's intrinsic flexibility reduces the strength of its correlation with the beta sheet residues.
- **Interpretation:** This figure emphasizes the beta sheet's role as a hub of strong dynamic interactions, anchoring the protein's motion. The helix, while dynamically active, plays a more peripheral role in interacting with the beta sheet.

Figure 3: Correlation with Residue 75

- **Beta Sheet Region (Residue 20):** Lower correlation values between the beta sheet and residue 75 indicate that the beta sheet's rigidity limits

its dynamic communication with the helix. This suggests that the beta sheet's motion is more self-contained, reducing its coupling to distant regions like the helix.

- **Alpha Helix Region (Residue 75):** High correlation values around residue 75 highlight the helix's central role in propagating dynamic signals within the protein. These correlations reflect the helix's ability to communicate dynamically with both nearby residues and more distant regions.
- **Interpretation:** The helix serves as a dynamic mediator, interacting more strongly with distant residues compared to the beta sheet. This figure underscores the helix's flexibility and its role in bridging different structural regions of the protein.

Figure 4: Response with Residue 20

- **Beta Sheet Region (Residue 20):** The response values around residue 20 show a sharp peak, consistent with the beta sheet's stabilizing influence on the protein's dynamics. The high response values highlight the beta sheet's role in shaping the local dynamic environment.
- **Alpha Helix Region (Residue 75):** In the helix region, the response to residue 20 is weaker, indicating that the helix dynamics are less directly influenced by the beta sheet. The helix's intrinsic flexibility decouples it from the rigid motions of the beta sheet.
- **Interpretation:** The beta sheet acts as a localized stabilizing force, with its influence tapering off in regions like the helix. This behavior supports the protein's need for both localized rigidity and broader adaptability.

Figure 5: Response with Residue 75

- **Beta Sheet Region (Residue 20):** A weak response is observed in the beta sheet, consistent with its rigidity and lower susceptibility to dynamic signals originating from the helix.
- **Alpha Helix Region (Residue 75):** A strong response peak near residue 75 reflects the helix's active role in facilitating dynamic transitions and responding to perturbations.
- **Interpretation:** The helix is a key dynamic element, complementing the beta sheet's rigidity by enabling dynamic adaptability and communication across the protein.

Figure 6: Transfer Entropy with Residue 20

- **Beta Sheet Region (Residue 20):** High transfer entropy values in the beta sheet reflect strong localized information flow, emphasizing the beta sheet's role in stabilizing and coordinating motions within its region.

- **Alpha Helix Region (Residue 75):** The helix shows lower transfer entropy values, indicating that it is less dynamically influenced by residue 20. This highlights the beta sheet's role as a local stabilizer rather than a global signal mediator.
- **Interpretation:** The beta sheet governs local dynamics through strong information flow, while the helix remains relatively decoupled in terms of direct influence.

Figure 7: Transfer Entropy with Residue 75

- **Beta Sheet Region (Residue 20):** Transfer entropy values indicate weak information flow from the helix to the beta sheet, consistent with the beta sheet's structural rigidity and localized dynamics.
- **Alpha Helix Region (Residue 75):** Strong peaks in transfer entropy around residue 75 highlight its role in propagating dynamic information across the protein, making it a key mediator of long-range communication.
- **Interpretation:** The helix complements the beta sheet by actively facilitating dynamic interactions and communication throughout the protein.

Figure 8: Transfer Entropy Across Residues

- **Beta Sheet Region (Residue 20):** The beta sheet demonstrates strong localized information transfer, confirming its role as a stabilizing region within the protein.
- **Alpha Helix Region (Residue 75):** High transfer entropy values in the helix illustrate its central role in mediating long-range dynamic interactions, reinforcing its complementary function alongside the beta sheet.
- **Interpretation:** This figure synthesizes the overall dynamic behavior, showcasing the beta sheet as a stabilizing anchor and the helix as a mediator of dynamic adaptability and communication.

Conclusion

The dynamic interplay between the beta sheet and the helix is quantitatively evident across all figures. The beta sheet stabilizes the protein through localized rigidity and strong intra-regional dynamics, while the helix introduces flexibility and facilitates long-range communication. These complementary roles are essential for balancing the stability and functional adaptability of the protein structure.

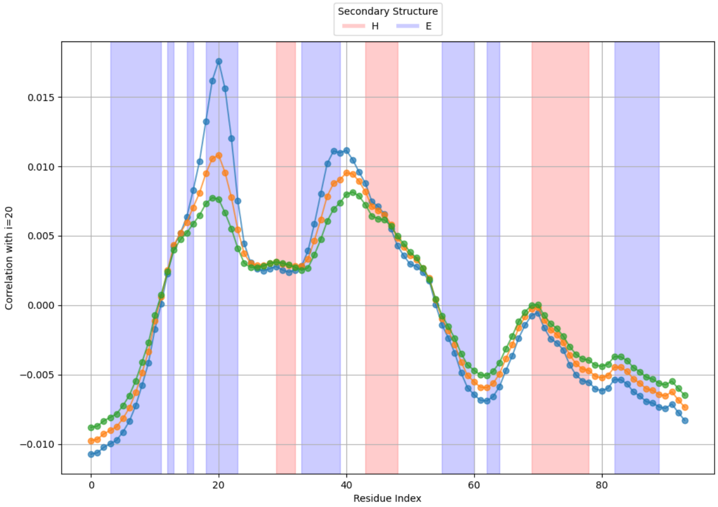


Figure 3.15. Correlation perturbating 21-th residue.

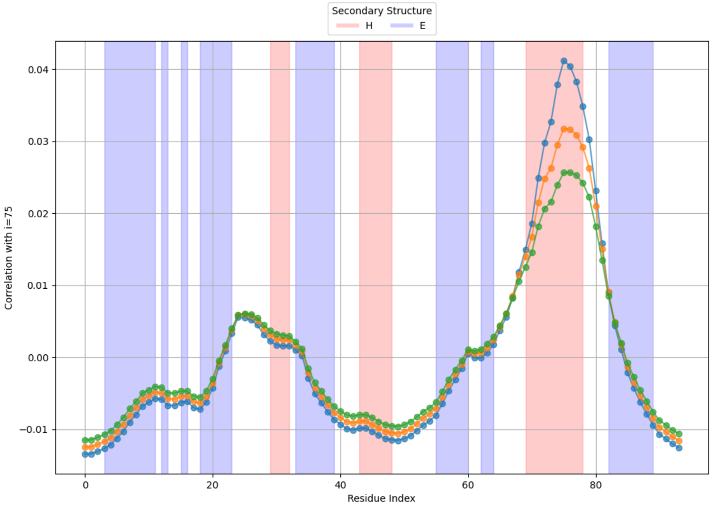


Figure 3.16. Correlation perturbating 76-th residue.

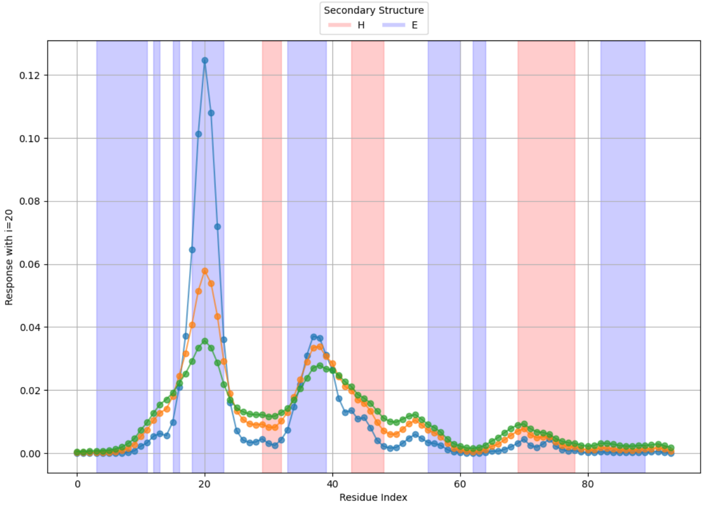


Figure 3.17. Response perturbating 21-th residue.

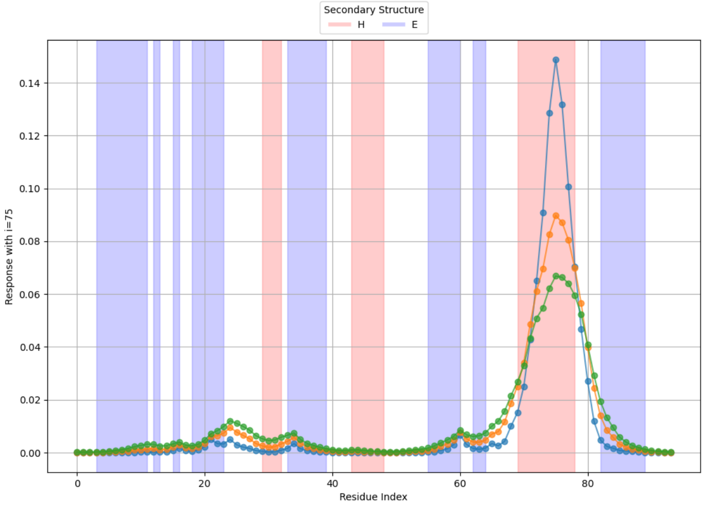


Figure 3.18. Response perturbating 76-th residue.

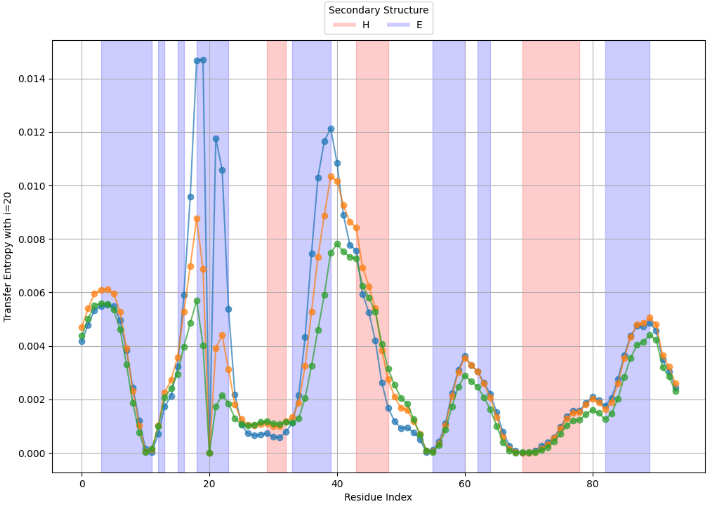


Figure 3.19. Transfer Entropy (21->j) perturbating 21-th residue.

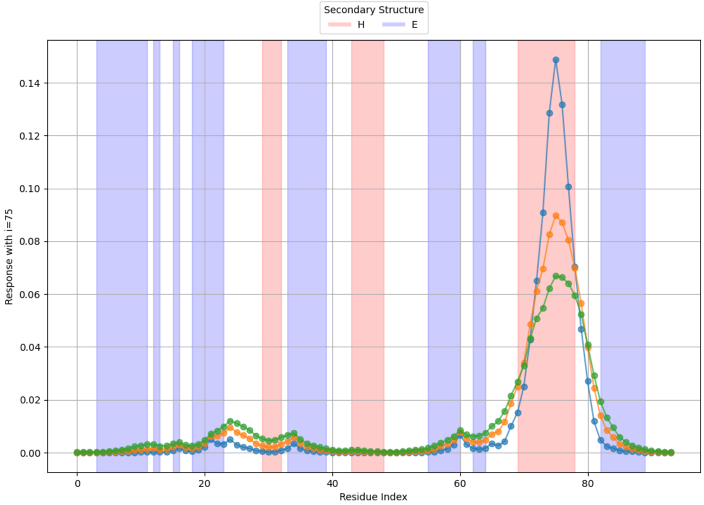


Figure 3.20. Transfer Entropy (76->j) perturbating 76-th residue.

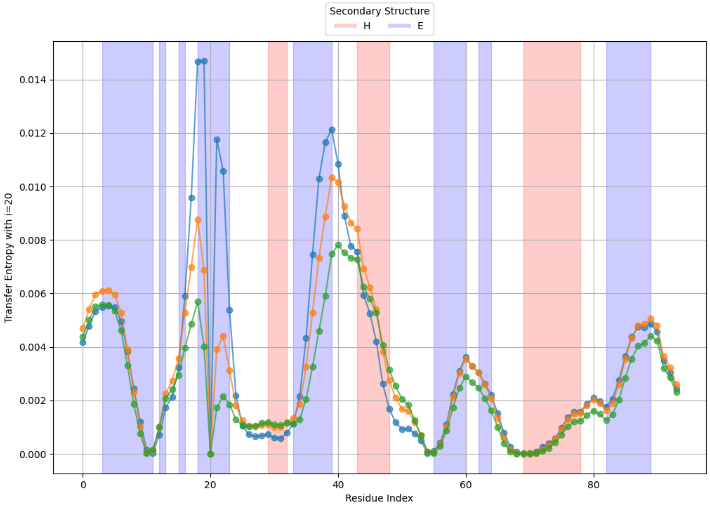


Figure 3.21. Transfer Entropy (21->j) perturbating 21-th residue.

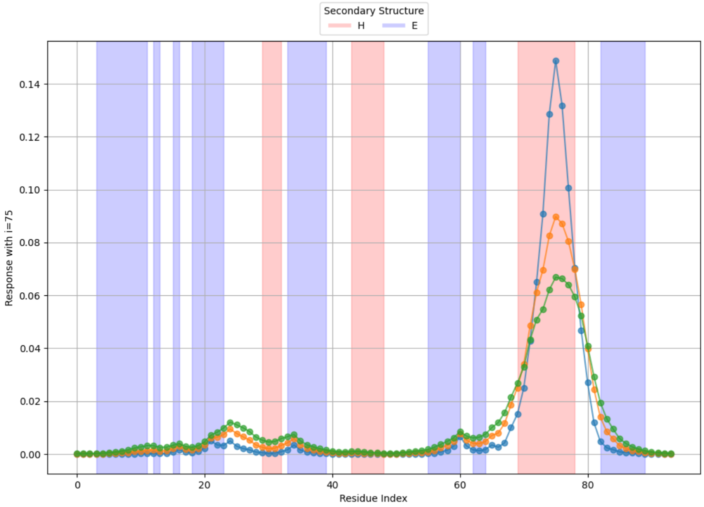


Figure 3.22. Transfer Entropy (76->j) perturbating 76-th residue.

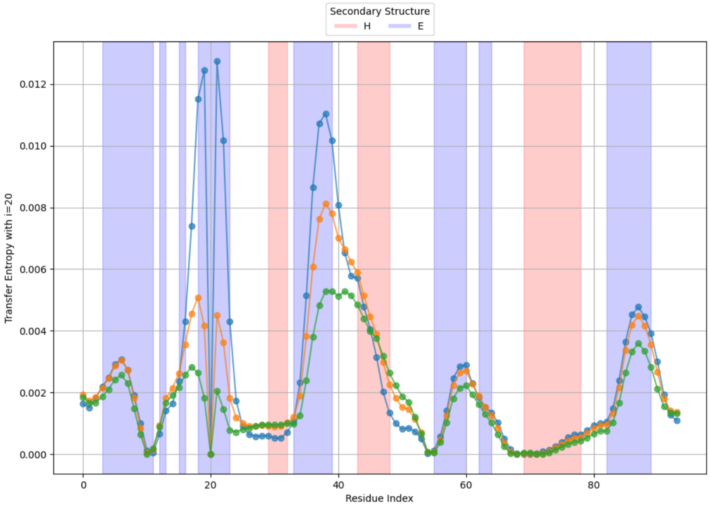


Figure 3.23. Transfer Entropy (j->21) perturbating 21-th residue.

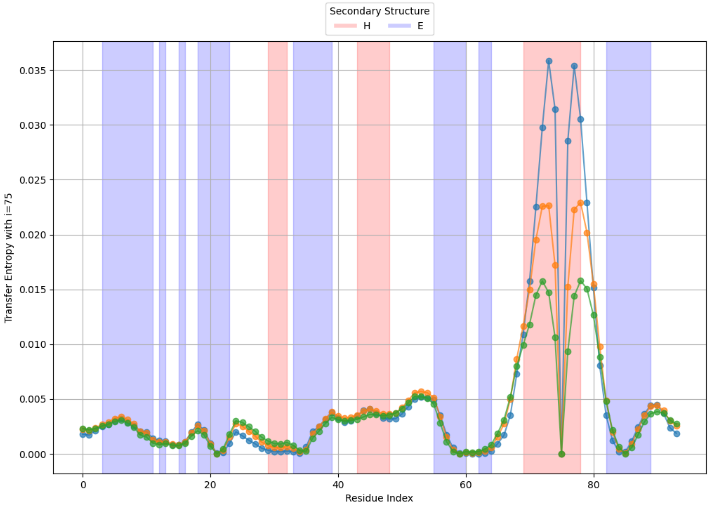


Figure 3.24. Transfer Entropy (j->76) perturbating 76-th residue.

Chapter 4

Conclusions Statical aprt

Chapter 5

Conclusions

Bibliography

[1] https://www.sciencelearn.org.nz/resources/209-role-of-proteins-in-the-body

[2] University$_{o}f_{C}alifornia_{D}avis, Introductory_{B}iology$

University$_{o}f_{C}alifornia_{D}avis, Introductory_{B}iology$

Nature, https://www.nature.com/scitable/topicpage/protein-structure-14122136/

Wikipedia, https://en.wikipedia.org/wiki/Allosteric_regulation?utm_source=chatgpt.com

Statistical Mechanics of Allosteric Enzymes

University$_{o}f_{C}alifornia_{D}avis, Introductory_{B}iology, Hemoglobinandallostericeffects$