



SAPIENZA
UNIVERSITÀ DI ROMA

Protein response in equilibrium and out of equilibrium conditions

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

Bignozzi Enrico

ID number 1855163

Advisor

Fabio Cecconi

Co-Advisor

Academic Year 2025

Thesis not yet defended

Protein response in equilibrium and out of equilibrium conditions
Sapienza University of Rome

© Bignozzi Enrico. All rights reserved

This thesis has been typeset by L^AT_EX and the Sapthesis class.

Author's email: bignozzi.1855163@studenti.uniroma1.it

Contents

Indice	2
Introduction	7
0.1 Structure of amino acids	8
0.2 Protein structure	9
0.2.1 Primary structure	9
0.2.2 Secondary structure	9
0.2.3 Tertiary structure	9
0.2.4 Quaternary structure	9
0.3 Allostericity	11
0.3.1 QUESTA PARTE LA METTO? LA SUA FINALITA E' FARE CAPIRE IN PRATICA AL LETTORE COSI' CHE SIA MENO ASTRATTA QUESTA PRIMA PARTE, DIMMI TU FABIO	13
0.3.2 Practical exemple of the mechanism of allostericity	13
0.3.3 Applications of understanding protein function	13
0.4 PDZ Domain	15
1 Modeling the Interaction between Amino Acids: GNM	18
1.1 Interpretation of the Taylor Expansion	20
1.2 Gaussian Network Model (GNM) and characteristics	23
2 Introduction to causality and causality in allosteric mechanisms of proteins (Qui è molto sottile il discorso e probabilmente non dovrei toccare alcuni punti)	28
2.0.1 Correlation Between Fluctuations	32
2.1 Linear Response Theorem	33
3 Stochastic Processes	40
3.1 General Framework	40
3.1.1 Dynamics in Vectorial Form	40
3.2 Introduction to Normal Modes and Their Physical Significance in Proteins	42
3.2.1 Normal Modes as Eigenvectors of the Hessian Matrix	42
3.2.2 Equations of Motion in Terms of Normal Modes	42
3.2.3 Physical Interpretation of Normal Modes in Proteins	43
3.3 Analytical Solution of stochastic process with normal modes	44

3.4	Solving the Dynamics Using Normal Modes	44
3.5	DA QUI IN POI NON GUARDARE	44
3.5.1	Decomposition in Terms of Normal Modes	44
3.5.2	Solution of the Equation for $r_k(t)$	45
3.5.3	Mean in Terms of Normal Modes	45
3.5.4	Covariance in Terms of Normal Modes (calcoli sbagliati non guardare)	45
3.5.5	Response Function	46
3.5.6	Transfer Entropy	47
4	Conclusions Statical aprt	69
5	Conclusions	70

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.

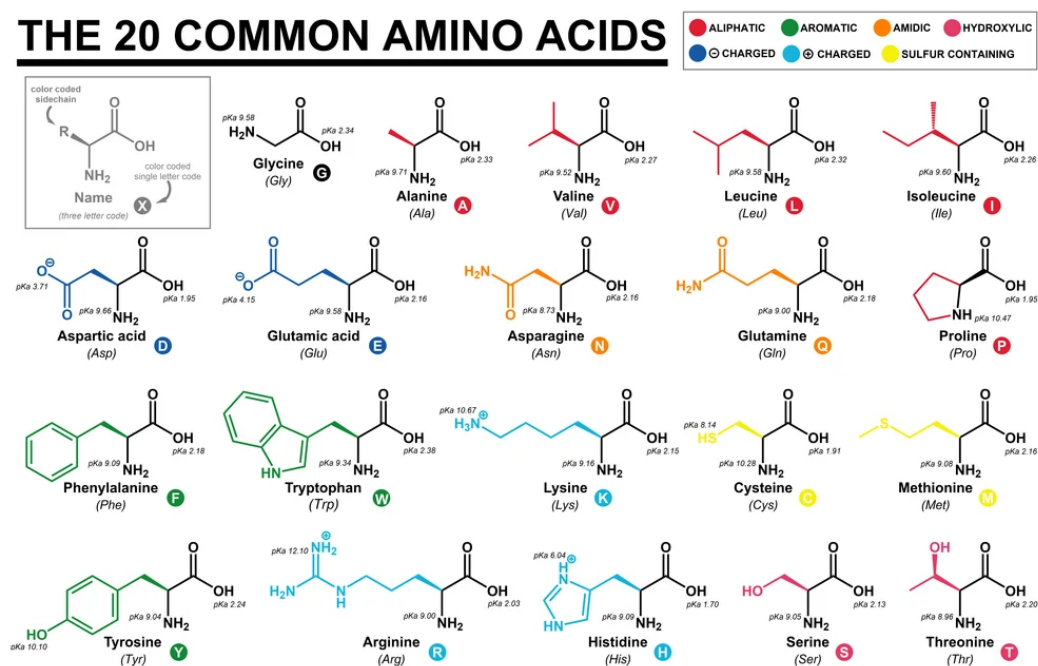


Figure 0.1. Amino Acids

0.2 Protein structure

The structure of the protein is important because it determines its function. The structure of a protein is divided into four levels:[\[4\]](#)

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.

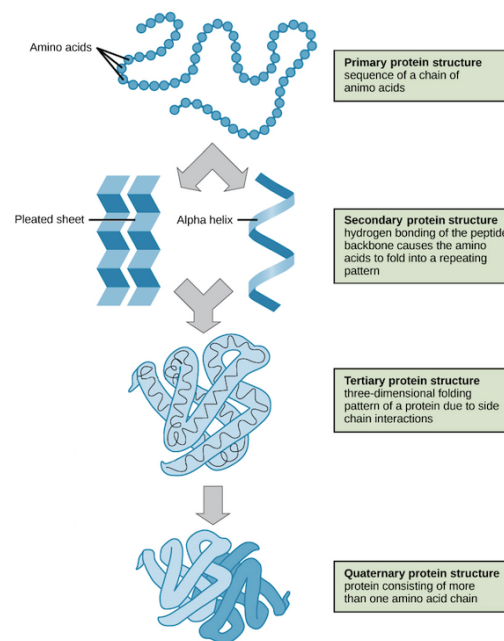


Figure 0.2. Structures

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, from the Greek *allos* (other) and *stereos* (structure), is the phenomenon of the change of the protein structure by the transmission of a signal from one site to another.[5] It 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. This avoids the protein to do various tasks in a regulated way. Additionally, allostericity 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.[6]

Mathematically, the phenomenon can be represented as a variation in the population of microstates $\{s_i\}$, described by a weighted Boltzmann distribution:[6]

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

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:[6]

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

where $[T]$ and $[R]$ represent the relative concentrations of the two conformational states, where the presence of an effector modifies L , stabilizing one state over the other. Therefore, proteins are not rigid structures; instead, they are ensembles of conformational states that fluctuate over temporal and spatial scales.

So, allosteric effectors 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:[6]

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

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 is useful for the study of the molecular basis of life and for the development of new

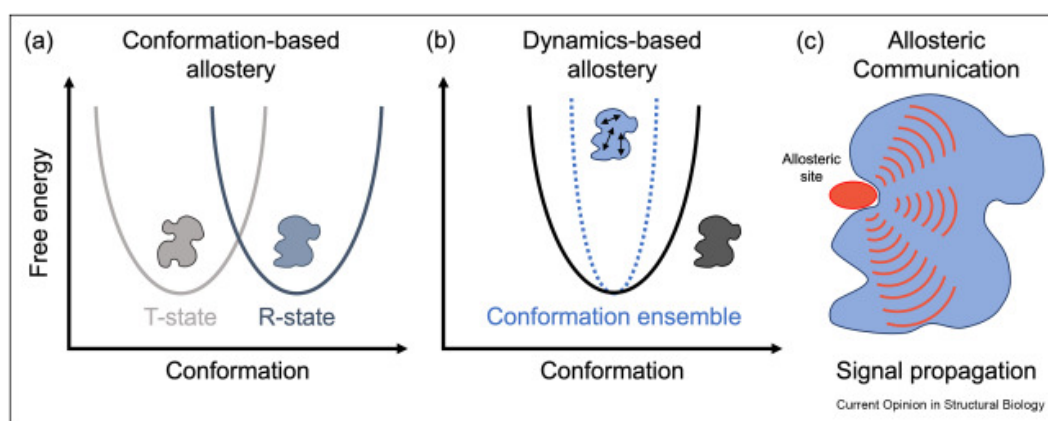


Figure 0.3. Schematic representation of allostericity: the interaction with an allosteric effector (site A) induces a conformational change that affects a distant site (site B).

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.

0.3.1 QUESTA PARTE LA METTO? LA SUA FINALITA E' FARE CAPIRE IN PRATICA AL LETTORE COSI' CHE SIA MENO ASTRATTA QUESTA PRIMA PARTE, DIMMI TU FABIO

0.3.2 Practical exemple of the mechanism of allostericity

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.

Hemoglobin exists in two main states:[7]

- **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.

0.3.3 Applications of understanding protein function

Exploring how proteins function in dynamic ways has numerous practical uses in areas like medicine, technology, and diagnostics:

- **Improving health treatments:** Discovering new ways to interact with specific protein sites can open doors to advanced therapies for various illnesses. For instance, it could help create medicines that manage oxygen levels in the body or develop drugs to target diseases like cancer or infections.
- **Designing synthetic proteins:** Insights into protein behavior are used to create innovative molecules, such as artificial blood substitutes or enzymes tailored for specific industrial or medical needs.
- **Advancing disease detection:** Certain protein characteristics make them valuable markers for identifying and tracking health conditions, from diabetes to tissue damage, offering more precise diagnostic tools.

- **Innovating drug development:** Strategies that focus on unique protein features can enhance drug effectiveness while minimizing unintended effects, paving the way for more targeted treatments in areas like cancer and heart disease.
- **Enhancing oxygen delivery:** Studying proteins involved in oxygen transport can lead to improved therapies for people with limited oxygen levels, whether due to medical conditions or high-performance demands.
- **Supporting agriculture and environment:** Understanding proteins in plants and microorganisms can boost food production or aid environmental cleanup, such as by improving nutrient efficiency or reducing carbon emissions.

0.4 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.[8]

Domains are essential because they serve as modular units that recur and combine to form complex proteins with distinct functions.[8]

Proteins often consist of multiple domains arranged in various combinations, creating multifunctional macromolecules capable of complex interactions. This domain architecture is essential for processes such as cell signaling, immune response, and metabolic regulation.

Moreover domains are not just structural units; they are also functional hubs that dictate the interactions and roles of proteins in cellular environments. [8]

Definition of Domain

A protein domain is generally defined by three key characteristics:[8]

1. **A structurally independent unit:** It is a portion of the protein that can fold into a stable three-dimensional structure on its own. This stability ensures that the domain can function even if other parts of the protein are absent or denatured.

The ability of domains to independently fold is crucial for their versatility and functionality.

2. **A functionally independent unit:** It can perform specific tasks, such as binding to a molecule or interacting with other proteins. For instance, the SH2 domain recognizes phosphorylated tyrosine residues, enabling signal transduction pathways.

Some domains exhibit functional independence.

3. **An evolutionary modular unit:** Domains are often reused in different proteins during evolution, allowing for the development of new functions. This modularity is evident in protein families where similar domains are observed across diverse organisms, reflecting a shared evolutionary origin.

Domains are classified based on their structure, sequence, or function, with databases such as Pfam and SMART cataloging thousands of examples.[?] These resources are invaluable for researchers aiming to predict protein function, identify evolutionary relationships, and design novel biomolecules with tailored functionalities. Structural studies and computational analyses continue to expand our understanding of domain architectures, revealing the intricate interplay between form and function in proteins.

The PDZ Domain

The **PDZ domain** is a well-studied example of a protein domain. It is named after the three proteins in which it was first identified:[8]

- PSD-95 (Postsynaptic density protein 95),

- **DlgA** (Drosophila discs large protein),
- **ZO-1** (Zona Occludens 1).

The PDZ domain exemplifies the modularity of protein domains, as it appears in a wide variety of proteins across eukaryotes, highlighting its evolutionary significance. PDZ domains organize and stabilize large multiprotein complexes. This prevalence underscores their critical role in maintaining cellular architecture and facilitating biochemical signaling pathways.

Characteristics of the PDZ Domain

- **Length:** Typically consists of 80–90 amino acids, making it a compact and efficient structural module. Despite its small size, the PDZ domain has remarkable functional versatility, binding to a diverse array of target peptides with high specificity.
- **Main Function:** Facilitates protein-protein interactions by recognizing specific peptide sequences, often located at the C-terminal region of target proteins. This ability to bind peptides with high specificity and affinity makes PDZ domains essential for organizing protein complexes. Such specificity arises from the conserved binding groove, which accommodates peptide motifs in a lock-and-key fashion.[?]
- **Biological Role:**
 - Assembling protein complexes at cellular membranes. For instance, PDZ domains play a pivotal role in assembling synaptic proteins, ensuring efficient signal transduction in neurons.
 - Cellular signaling, particularly in pathways regulating cell growth and differentiation. PDZ-mediated interactions are involved in key signaling cascades, including those controlled by receptor tyrosine kinases and G-protein-coupled receptors.
- **Structure:** The PDZ domain has a characteristic structure composed of an antiparallel beta-sheet and two alpha-helices. This conserved fold forms a groove that interacts with the target peptide, contributing to its specificity. The structural plasticity of PDZ domains allows them to adapt to various ligands, enhancing their functional diversity.[?]

The PDZ Domain in the 3LNX Protein

The 3LNX protein, on which we will work, provides a clear example of how a PDZ domain binds to a target peptide. This structure offers important insights into the functioning of these domains:

1. How the peptide binds:

In the 3LNX protein, the PDZ domain has a "hydrophobic pocket" that accommodates the peptide it binds to. This binding is very precise and stable,

thanks to connections such as hydrogen bonds and weak interactions between nearby molecules. Small changes in the sequence of the peptide can influence how well the PDZ domain binds.

2. Domain flexibility:

When the peptide binds, the PDZ domain slightly adjusts its shape. This shows that it is very flexible and capable of interacting with many different types of molecules. This flexibility is essential for its role in cellular communication.

3. Biological importance:

PDZ domains, like the one in the 3LNX protein, are fundamental in synapse organization, where they help assemble signaling complexes necessary for brain function. Knowledge of this protein has been used to design molecules that can regulate these interactions, with potential applications for treating diseases.

Chapter 1

Modeling the Interaction between Amino Acids: GNM

The interactions between the amino acids that compose the protein are crucial for the folding of the protein into a functional structure. Thus, we need a Hamiltonian to describe the system.

If the protein is at equilibrium, we expect that the Hamiltonian is a function of a potential that depends on the position of every atom constituting the PDZ-specific domain. Therefore, my Hamiltonian H can be written as $H = V(\mathbf{r})$, where \mathbf{r} is the vector containing the positions of every atom that constitutes the protein.

Since this is a very complex system, we cannot solve the problem exactly, so we must use an approximation. The second-order approximation of a function $V(\mathbf{r})$, around an equilibrium point \mathbf{r}_0 , can be expressed as:[\[9\]](#)

$$V(\mathbf{r}) \approx V(\mathbf{r}_0) + \nabla V(\mathbf{r}_0)^\top (\mathbf{r} - \mathbf{r}_0) + \frac{1}{2} (\mathbf{r} - \mathbf{r}_0)^\top \mathbf{H}_V (\mathbf{r} - \mathbf{r}_0), \quad (1.1)$$

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:

$$\nabla V(\mathbf{r}) = \left. \frac{\partial V}{\partial \mathbf{r}} \right|_{\mathbf{r}=\mathbf{r}_0}, \quad (1.2)$$

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

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

$$\mathbf{H}_V = \left. \frac{\partial^2 V}{\partial \mathbf{r}^2} \right|_{\mathbf{r}=\mathbf{r}_0}, \quad (1.3)$$

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

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).

Thus, apart from constants, the Hamiltonian can be approximated as:

$$V(\mathbf{r}) \approx \frac{1}{2}(\mathbf{r} - \mathbf{r}_0)^\top \mathbf{H}_V (\mathbf{r} - \mathbf{r}_0). \quad (1.4)$$

This modeling is called Gaussian Network Model (GNM) and is commonly used to study the dynamics of proteins.

To reduce the computational complexity of the problem while preserving essential structural features, we focus solely on the α -carbon atoms because:

- α -carbon atoms form the backbone of the protein, defining its overall structure and shape.
- Their positions are sufficient to reconstruct the global three-dimensional conformation of the protein.
- Focusing on α -carbons significantly reduces the complexity of the system, lowering the number of degrees of freedom and making calculations more efficient.
- The dynamics of α -carbons often capture the collective motions of the protein, 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, serving as reference points for modeling.

This model is overall extendable to all other atoms of the protein.

1.1 Interpretation of the Taylor Expansion

Equation (1.4) can be interpreted not only as the Hamiltonian of the harmonic oscillator (spring-system) but also as a network model, where every α -carbon atom is a node interacting with other nodes.[9]

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.[10]

Graphs can be directed or undirected, weighted or unweighted, depending on the nature of the relationships they represent.

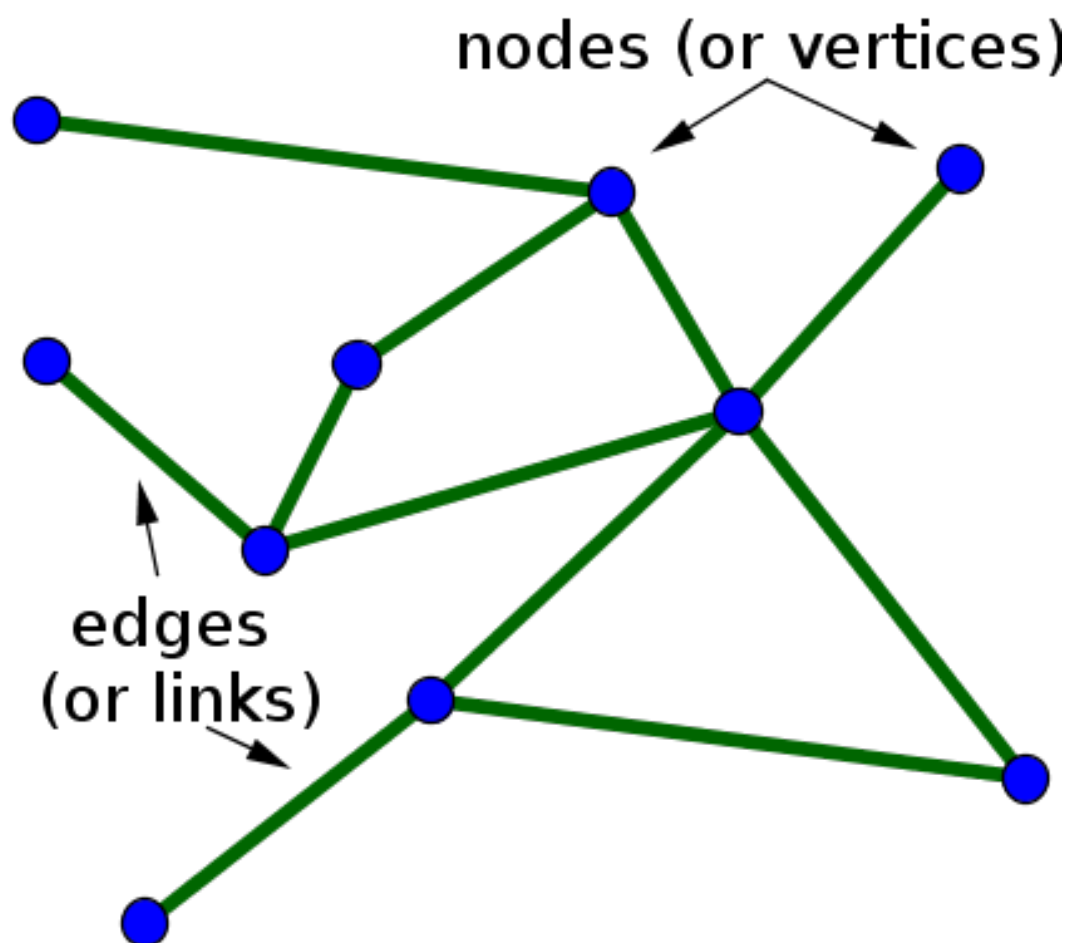


Figure 1.1. Schematic representation of a Network.

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:[10]

$$L = D - A \quad (1.5)$$

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

$$D_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \quad (1.6)$$

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:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases} \quad (1.7)$$

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:[10]

- **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:[10]

- **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:[10]

- **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. So it is the inverse of L in the subspace where L is invertible and minimize this difference $I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

1.2 Gaussian Network Model (GNM) and characteristics

The Hamiltonian of the system can be expressed, following (1.4), as:[11]

$$H(\mathbf{r}) \approx \frac{1}{2}(\mathbf{r} - \mathbf{r}_0)^\top \mathbf{K}(\mathbf{r} - \mathbf{r}_0), \quad (1.8)$$

where \mathbf{K} is the Laplacian matrix. For simplicity, we can rewrite it as:

$$H(\mathbf{r}) \approx \frac{1}{2}\mathbf{r}^\top \mathbf{K}\mathbf{r}, \quad (1.9)$$

or, in components:

$$H(\mathbf{r}) \approx \frac{1}{2} \sum_{i,j} r_i K_{ij} r_j. \quad (1.10)$$

Here, \mathbf{r} represents the displacement vector of atoms from their equilibrium positions. This formulation of the system leads to several important properties and characteristics.

Initially, the probability density at equilibrium is given by:[11]

$$P(\mathbf{r}) = \frac{1}{Z} e^{-\beta H(\mathbf{r})}, \quad (1.11)$$

where the partition function Z is:

$$Z = \int e^{-\beta H(\mathbf{r})} d\mathbf{r}. \quad (1.12)$$

Mean Position

The mean position $\langle \mathbf{r} \rangle$ can be calculated as:[11]

$$\langle \mathbf{r} \rangle = \int \mathbf{r} P(\mathbf{r}) d\mathbf{r}, \quad (1.13)$$

where:

$$P(\mathbf{r}) = \frac{1}{Z} e^{-\frac{\beta}{2} \mathbf{r}^\top \mathbf{K} \mathbf{r}}, \quad (1.14)$$

and the partition function is:[11]

$$Z = \int e^{-\frac{\beta}{2} \mathbf{r}^\top \mathbf{K} \mathbf{r}} d\mathbf{r}. \quad (1.15)$$

Substituting $P(\mathbf{r})$ into the expression for $\langle \mathbf{r} \rangle$, we obtain:

$$\langle \mathbf{r} \rangle = \frac{1}{Z} \int \mathbf{r} e^{-\frac{\beta}{2} \mathbf{r}^\top \mathbf{K} \mathbf{r}} d\mathbf{r}. \quad (1.16)$$

Since the integrand $\mathbf{r} e^{-\frac{\beta}{2} \mathbf{r}^\top \mathbf{K} \mathbf{r}}$ is symmetric with respect to \mathbf{r} , and there are no linear terms in the Hamiltonian, the Gaussian distribution is centered at $\mathbf{r} = 0$. Therefore:[11]

$$\langle \mathbf{r} \rangle = 0. \quad (1.17)$$

Covariance

The covariance between r_i and r_j is defined as:[\[11\]](#)

$$\text{Cov}(r_i, r_j) = \langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle. \quad (1.18)$$

Since $\langle r_i \rangle = 0$, it simplifies to:

$$\text{Cov}(r_i, r_j) = \langle r_i r_j \rangle. \quad (1.19)$$

Using the Boltzmann distribution:

$$P(\mathbf{r}) = \frac{1}{Z} e^{-\frac{\beta}{2} \mathbf{r}^\top \mathbf{K} \mathbf{r}}, \quad (1.20)$$

we calculate:

$$\langle r_i r_j \rangle = \frac{1}{Z} \int r_i r_j e^{-\frac{\beta}{2} \mathbf{r}^\top \mathbf{K} \mathbf{r}} d\mathbf{r}. \quad (1.21)$$

For a Gaussian distribution, the covariance matrix is given by:

$$\Sigma = \beta^{-1} \mathbf{K}^{-1} = \langle r_i r_j \rangle, \quad (1.22)$$

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

$$\text{Cov}(r_i, r_j) = \frac{1}{\beta} (\mathbf{K}^{-1})_{ij}. \quad (1.23)$$

QUESTA PARTE IN CUI METTO LA DERIVAZIONE DEI BETA FACTORS LA METTO? Referenza materia condensata sapienza

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:

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

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 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.

Model Evaluation: R^2 and MAE QUESTA IDEM NON SO SE METTERLA O NO. SCELGO IN FUNZIONE SE METTO QUESTE METRICHE DI ERRORE NELLA TESI DOPO

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.

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.

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.

Chapter 2

Introduction to causality and causality in allosteric mechanisms of proteins (Qui è molto sottile il discorso e probabilmente non dovrei toccare alcuni punti)

Correlation not only hinders testing the quality of models through beta factors but also obstructs understanding interactions and the relations between nodes/residues. Our primary aim is to identify the allosteric site—the region from which a signal is transmitted to other sites of the protein, but as is 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).

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

In mathematical terms, causal relationships are often modeled as:[\[13\]](#)

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

where f describes how A and other factors jointly determine B .

Experimentally, causality is typically established through controlled interventions and the observation of resulting changes.

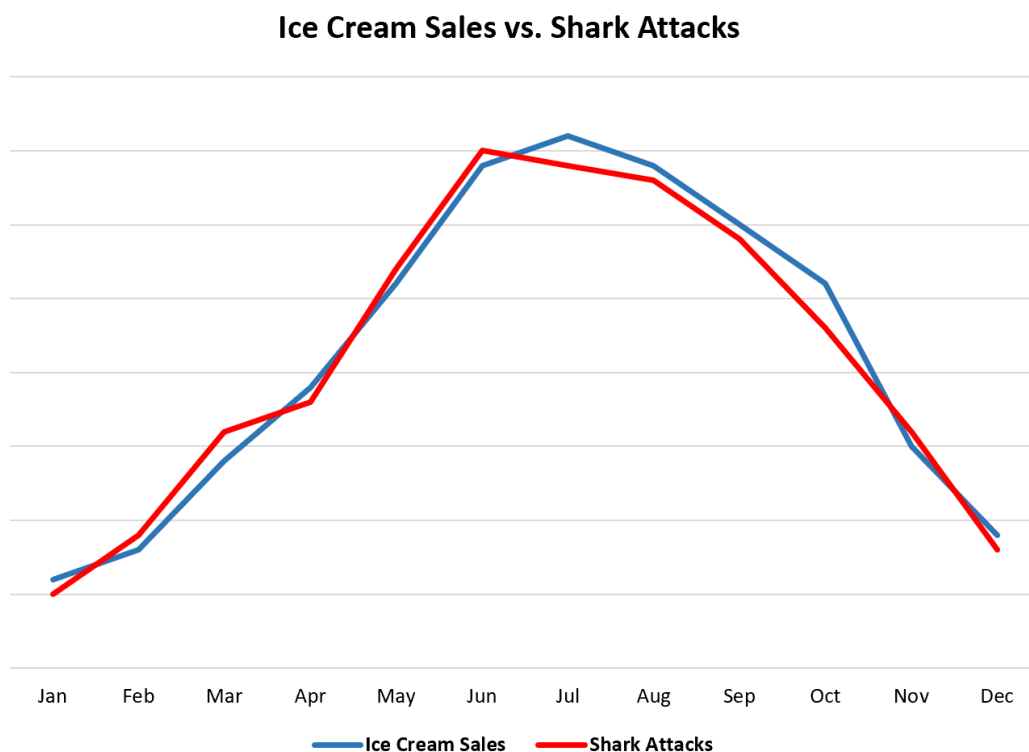


Figure 2.1. Correlation is not causation.

Deterministic and Stochastic Causality

Causality can be classified into two main types: deterministic and stochastic.

Deterministic Causality

Deterministic causality implies that a specific cause invariably leads to a specific effect. If an event A causes an event B , then B always occurs whenever A occurs. This relationship can be expressed as $A \implies B$, meaning A is both necessary and sufficient for B . Deterministic causality is observed in systems governed by precise physical laws, such as classical mechanics, where given initial conditions and forces, trajectories are entirely predictable.

Stochastic Causality

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 always follow A . This is expressed probabilistically as $P(B|A) > P(B|\neg A)$, indicating that A raises the probability of B . Stochastic causality is prevalent in complex systems like biology or social sciences, where multiple interacting factors and inherent randomness influence outcomes. For example, smoking increases the probability of lung cancer, though not all smokers develop the disease.

Both deterministic and stochastic causality are critical for understanding natural

phenomena. Deterministic causality offers precise predictive power, while stochastic causality is indispensable for modeling systems where randomness and variability play significant roles. Together, they provide a comprehensive framework for analyzing causal relationships.

Geometric Perspective of Causality and Shannon Entropy (questa è la parte piu' da togliere fra tutte probabilmente)

When all dimensions or variables of a system are fully known, uncertainty about future states vanishes, transitioning the system into a deterministic regime. This idea relates 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 $P(x_i) = \delta(x_i) \rightarrow H(X) = 0$, the system has no uncertainty, indicating complete predictability. So if my target variable has the distribution of the delta function, the entropy is zero, so the system is deterministic. In deterministic systems, knowing all dimensions eliminates ambiguity, enabling precise predictions of future states. However, in practice i don't know all the dimensions of my system. Thus, while $H(X) = 0$ theoretically implies determinism, real-world complexities and limitations must be accounted for.

To understand the causality of allosteric mechanisms in proteins, it is crucial to investigate the physical interactions and pathways that transmit signals between distant sites. For this reason, additional indicators are needed.

Correlation

2.0.1 Correlation Between Fluctuations

The correlation between two variables, such as the fluctuations in the positions of particles i and j , provides insight into how these components of the system interact. The correlation function is defined as:[\[12\]](#)

$$C_{ij} = \langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle, \quad (2.1)$$

where:

- $\delta \mathbf{r}_i = \mathbf{r}_i - \langle \mathbf{r}_i \rangle$ is the deviation of the position of particle i from its equilibrium value.
- $\delta \mathbf{r}_j = \mathbf{r}_j - \langle \mathbf{r}_j \rangle$ is the deviation of the position of particle j from its equilibrium value.
- $\langle \cdot \rangle$ denotes an ensemble average.

The correlation C_{ij} quantifies the degree to which the positions of particles i and j are linearly related at equilibrium.

Physical Interpretation of Correlations

The correlation C_{ij} reflects how strongly two particles are connected in the system:

- **Positive correlation:** A positive $C_{ij} > 0$ indicates that fluctuations in \mathbf{r}_i and \mathbf{r}_j tend to occur in the same direction, suggesting cooperative behavior or a direct connection.
- **Negative correlation:** A negative $C_{ij} < 0$ implies that fluctuations in \mathbf{r}_i and \mathbf{r}_j tend to occur in opposite directions, indicating opposing interactions or constraints between the particles.
- **Zero correlation:** A correlation of $C_{ij} = 0$ indicates no linear relationship between the fluctuations, suggesting either independence or non-linear interactions.

Why Correlation Alone is Not Causation

While correlations reveal interactions, they do not establish a causal relationship. The limitations include:

- **Symmetry:** Correlation is symmetric ($C_{ij} = C_{ji}$), whereas causality is asymmetric. The fact that i is correlated with j does not imply that i causes j or vice versa.
- **Confounding variables:** Correlation can arise due to a common cause influencing both i and j , rather than a direct causal link between them.
- **Indirect interactions:** Correlation can reflect indirect interactions mediated by other components of the system, rather than a direct causal pathway.

Using Correlations to Infer Interaction Networks

Despite its limitations, correlation is a powerful tool for mapping interaction networks in complex systems. For example:

- **Identifying connected components:** Strong correlations can indicate direct physical or functional connections between components.
- **Inferring collective behavior:** Patterns of correlations across multiple components can reveal collective modes of motion or global interactions in the system.
- **Complementing causal analysis:** Correlation provides a starting point for exploring causal relationships, which can be further refined using response functions and perturbation experiments.

In summary, correlation serves as a valuable indicator of interactions within a system. However, establishing causality requires additional analysis, such as studying response functions or applying controlled perturbations.

2.1 Linear Response Theorem

The linear response theorem describes how the average position of a particle i , denoted as $\langle \mathbf{r}_i \rangle$, responds to a perturbation applied to another particle j through an external force f_j . The response is given by:

$$R_{ij} = \frac{\partial \langle \mathbf{r}_i \rangle}{\partial f_j}, \quad (2.2)$$

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 .

Derivation and Relation to Correlations

To derive the connection between the response and the correlations of position fluctuations, consider the Hamiltonian of the system under a small external perturbation:

$$\mathcal{H} = \mathcal{H}_0 - \sum_j f_j \mathbf{r}_j, \quad (2.3)$$

where \mathcal{H}_0 is the unperturbed Hamiltonian and $-f_j \mathbf{r}_j$ is the interaction term between the external force and the position of particle j .

The average position of particle i in the perturbed system is given by:

$$\langle \mathbf{r}_i \rangle = \frac{\int \mathbf{r}_i e^{-\beta \mathcal{H}} d\mathbf{r}}{\int e^{-\beta \mathcal{H}} d\mathbf{r}}, \quad (2.4)$$

where $\beta = \frac{1}{k_B T}$ is the inverse thermal energy, k_B is the Boltzmann constant, and T is the temperature.

Expanding the exponential $e^{-\beta\mathcal{H}}$ to first order in the perturbation f_j :

$$e^{-\beta\mathcal{H}} \approx e^{-\beta\mathcal{H}_0} (1 + \beta f_j \mathbf{r}_j), \quad (2.5)$$

and substituting this expansion into the expression for $\langle \mathbf{r}_i \rangle$:

$$\langle \mathbf{r}_i \rangle \approx \langle \mathbf{r}_i \rangle_0 + \beta f_j \langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle, \quad (2.6)$$

where:

- $\langle \mathbf{r}_i \rangle_0$ is the average position of particle i in the absence of perturbation.
- $\delta \mathbf{r}_i = \mathbf{r}_i - \langle \mathbf{r}_i \rangle_0$ is the fluctuation of the position relative to its equilibrium value.
- $\langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle$ is the scalar product of position fluctuations between particles i and j .

Taking the derivative of $\langle \mathbf{r}_i \rangle$ with respect to f_j gives the response:

$$R_{ij} = \frac{\partial \langle \mathbf{r}_i \rangle}{\partial f_j} = \beta \langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle. \quad (2.7)$$

This result shows that the response R_{ij} is directly proportional to the equilibrium correlation between the fluctuations of the positions of particles i and j , a manifestation of the fluctuation-dissipation theorem.

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:[12]

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

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

Transfer Entropy (TE) is a powerful measure of directed information transfer between stochastic processes. For stationary Gaussian processes, TE can be computed analytically using the covariance matrices of the processes involved. The TE from x_j to x_i at a time lag t is given by:[12]

$$TE_{j \rightarrow i}(t) = -\frac{1}{2} \ln \left(1 - \frac{\alpha_{ij}(t)}{\beta_{ij}(t)} \right), \quad (2.8)$$

where:

$$\alpha_{ij}(t) = [C_{ii}(0)C_{ij}(t) - C_{ij}(0)C_{ii}(t)]^2, \quad (2.9)$$

$$\beta_{ij}(t) = [C_{ii}(0)C_{jj}(0) - C_{ij}^2(0)] [C_{ii}^2(0) - C_{ii}^2(t)]. \quad (2.10)$$

Here:

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

Proof Sketch for Gaussian TE

The derivation of Transfer Entropy (TE) for Gaussian systems leverages the fact that the entropy of a multivariate Gaussian distribution depends only on the determinant of its covariance matrix.

This section provides a step-by-step explanation of the derivation for $TE_{j \rightarrow i}(t)$. [12]

Consider a system described by the variables $x_i(t)$, $x_i(0)$, and $x_j(0)$. The joint covariance matrix of these variables is:

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

where:

- $C_{ii}(0)$ and $C_{jj}(0)$ are the variances of x_i and x_j , respectively,
- $C_{ii}(t)$ is the autocovariance of x_i at time lag t ,
- $C_{ij}(t)$ is the cross-covariance between $x_i(t)$ and $x_j(0)$,
- $C_{ij}(0)$ is the instantaneous cross-covariance between $x_i(0)$ and $x_j(0)$.

The entropy of a multivariate Gaussian distribution is given by:

$$H(\mathbf{x}) = \frac{1}{2} \ln \left([(2\pi e)^n \det(\Sigma)] \right),$$

where n is the dimensionality of \mathbf{x} , and Σ is its covariance matrix.

For the variables $[x_i(t), x_i(0), x_j(0)]$, the entropy is:

$$H(x_i(t), x_i(0), x_j(0)) = \frac{1}{2} \ln \left([(2\pi e)^3 \det(\Omega)] \right).$$

The conditional entropy of $x_i(t)$ given $[x_i(0), x_j(0)]$ is computed using the Schur complement. For a covariance matrix partitioned as:

$$\Omega = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix},$$

the Schur complement of C in Ω is:

$$\Sigma_{x_i(t)|x_i(0), x_j(0)} = A - BC^{-1}B^\top.$$

In our case, the conditional covariance matrix for $x_i(t)$ given $x_i(0)$ and $x_j(0)$ is:

$$\Sigma_{x_i(t)|x_i(0), x_j(0)} = C_{ii}(0) - \begin{bmatrix} C_{ii}(t) & C_{ij}(t) \end{bmatrix} \begin{bmatrix} C_{ii}(0) & C_{ij}(0) \\ C_{ij}(0) & C_{jj}(0) \end{bmatrix}^{-1} \begin{bmatrix} C_{ii}(t) \\ C_{ij}(t) \end{bmatrix}.$$

The entropy is then:

$$H(x_i(t)|x_i(0), x_j(0)) = \frac{1}{2} \ln \left([2\pi e \det(\Sigma_{x_i(t)|x_i(0), x_j(0)})] \right).$$

Similarly, the conditional covariance matrix for $x_i(t)$ given $x_i(0)$ is:

$$\Sigma_{x_i(t)|x_i(0)} = C_{ii}(0) - \frac{C_{ii}^2(t)}{C_{ii}(0)}.$$

The entropy is:

$$H(x_i(t)|x_i(0)) = \frac{1}{2} \ln \left([2\pi e \det(\Sigma_{x_i(t)|x_i(0)})] \right).$$

Transfer Entropy is defined as:

$$TE_{j \rightarrow i}(t) = H(x_i(t)|x_i(0)) - H(x_i(t)|x_i(0), x_j(0)).$$

Substituting the expressions for the conditional entropies:

$$TE_{j \rightarrow i}(t) = \frac{1}{2} \ln \left(\frac{\det(\Sigma_{x_i(t)|x_i(0)})}{\det(\Sigma_{x_i(t)|x_i(0), x_j(0)})} \right).$$

Using the determinant properties of Gaussian covariance matrices and after algebraic manipulation, this simplifies to:

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

where:

$$\begin{aligned} \alpha_{ij}(t) &= [C_{ii}(0)C_{ij}(t) - C_{ij}(0)C_{ii}(t)]^2, \\ \beta_{ij}(t) &= [C_{ii}(0)C_{jj}(0) - C_{ij}^2(0)] [C_{ii}^2(0) - C_{ii}^2(t)]. \end{aligned}$$

This is the analytical formula for Transfer Entropy in Gaussian systems.

The final expression reveals how the directed information transfer from x_j to x_i depends on:

1. The time-lagged cross-correlation $C_{ij}(t)$,
2. The variances $C_{ii}(0)$ and $C_{jj}(0)$,
3. The instantaneous cross-correlation $C_{ij}(0)$,
4. The autocovariance $C_{ii}(t)$.

This formula captures the influence of x_j on x_i while accounting for their shared history, providing a rigorous measure of causality.

Causality and Transfer Entropy

Transfer Entropy is widely regarded as an indicator of causality because it quantifies directed information flow between variables. The key reasons include:

- **Directionality:** Unlike correlation, TE is asymmetric ($TE_{j \rightarrow i} \neq TE_{i \rightarrow j}$). This asymmetry reflects the directional influence of x_j on x_i , aligning with the principle of causation.
- **Temporal dependence:** TE explicitly incorporates time-lagged variables, ensuring that the direction of information flow respects temporal precedence, a fundamental requirement for causal inference.
- **Beyond correlation:** While correlation measures symmetric associations, TE captures the dynamic influence of x_j on x_i , conditioned on their past states. This conditioning eliminates spurious correlations arising from common drivers or indirect interactions.

Relation to Correlation and Response Functions

The relationship between Transfer Entropy, correlation, and response functions highlights their complementary roles in understanding causality:

- **Correlation as a precursor:** Correlation measures the strength of association but is symmetric and does not imply causation. TE builds upon this by quantifying directed dependencies.
- **Response functions as a mechanism:** Linear response functions describe how perturbations in one variable influence another. They are closely related to correlations through the fluctuation-dissipation theorem and provide a mechanistic basis for TE.
- **TE as a probabilistic extension:** TE extends the concepts of correlation and response functions into a probabilistic framework. It quantifies not just direct influences but also the flow of information, incorporating both linear and non-linear dependencies.

By combining these tools, one can construct a comprehensive picture of causality in complex systems, where TE provides a probabilistic and directional perspective, correlation identifies initial associations, and response functions reveal the underlying mechanisms.

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 Processes

A stochastic process is a mathematical framework used to describe systems that evolve over time in a probabilistic manner.

By definition, a stochastic process associates a set of random variables $\{X_t\}_{t \geq 0}$ to time t , where each random variable X_t represents the state of the system at a specific time point.

Such processes are essential in modeling phenomena in physics, biology, finance, and many other fields, where uncertainty and noise play significant roles.

3.1 General Framework

The dynamics of a stochastic process are governed by probabilistic laws that combine deterministic and random components. In general, the evolution can be expressed as:

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

where:

- X_t represents the state of the system at time t ,
- $H(X_t)$ is the Hamiltonian function, which dictates the deterministic behavior,
- $\nabla H(X_t)$ is the gradient of the Hamiltonian, the force, describing the deterministic direction of motion,
- η_t is a stochastic noise term, modeled as Gaussian white noise with zero mean and variance σ^2 , i.e., $\langle \eta_t \eta_{t'} \rangle = \sigma^2 \delta(t - t')$.

3.1.1 Dynamics in Vectorial Form

Substituting the gradient into the general stochastic equation, using (1.4), the dynamics can be expressed in vectorial form as:

$$\frac{d\mathbf{r}_t}{dt} = -\mathbf{K}\mathbf{r}_t + \boldsymbol{\eta}_t, \quad (3.2)$$

where:

- \mathbf{r}_t is the state vector at time t ,
- $-\mathbf{K}\mathbf{r}_t$ is the deterministic term driving the system towards equilibrium,
- $\boldsymbol{\eta}_t$ is a vector of independent Gaussian noise components.

In component form, the dynamics for each element i are:

$$\frac{dr_{i,t}}{dt} = - \sum_j K_{ij} r_{j,t} + \eta_{i,t}. \quad (3.3)$$

The described dynamics correspond to a multidimensional Ornstein-Uhlenbeck process, which is the simplest continuous-time Gaussian process with a mean-reverting property.

These processes are used extensively in modeling systems with memory and noise.

3.2 Introduction to Normal Modes and Their Physical Significance in Proteins

Normal modes are a powerful tool for understanding protein dynamics. They are derived from the diagonalization of the Hessian matrix, \mathbf{H} , which encapsulates the curvature of the potential energy surface near a local minimum. The Hessian matrix is defined as the second derivative of the potential energy V with respect to the atomic coordinates $\{x_i\}$:

$$H_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j}, \quad (3.4)$$

where i and j span all degrees of freedom in the system. The Hessian matrix quantifies how changes in one coordinate affect the energy and how different degrees of freedom are coupled.

3.2.1 Normal Modes as Eigenvectors of the Hessian Matrix

Normal modes are the eigenvectors \mathbf{v}_k of the Hessian matrix. These eigenvectors correspond to collective motions of the system, and the associated eigenvalues λ_k are proportional to the squared frequencies of these motions:

$$\mathbf{H}\mathbf{v}_k = \lambda_k \mathbf{v}_k. \quad (3.5)$$

To understand why this relation holds, consider a Taylor expansion of the potential energy $V(\mathbf{x})$ around a local minimum at \mathbf{x}_0 :

$$V(\mathbf{x}) \approx V_0 + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x}, \quad (3.6)$$

where V_0 is the potential energy at the minimum, and \mathbf{x} represents the displacement from equilibrium. This quadratic approximation assumes small displacements, making the system effectively harmonic.

3.2.2 Equations of Motion in Terms of Normal Modes

From classical mechanics, the force \mathbf{F} is related to the gradient of the potential energy:

$$\mathbf{F} = -\nabla V(\mathbf{x}) = -\mathbf{H}\mathbf{x}. \quad (3.7)$$

Newton's second law gives the equation of motion:

$$M\ddot{\mathbf{x}} = -\mathbf{H}\mathbf{x}, \quad (3.8)$$

where M is the diagonal mass matrix. To simplify, we normalize the coordinates by the square root of the masses: $\mathbf{x}' = \sqrt{M}\mathbf{x}$. Substituting this transformation, the equation of motion becomes:

$$\ddot{\mathbf{x}} = -\mathbf{H}\mathbf{x}. \quad (3.9)$$

3.2. INTRODUCTION TO NORMAL MODES AND THEIR PHYSICAL SIGNIFICANCE IN PROTEINS

Now, assume solutions of the form $\mathbf{x}(t) = \mathbf{v}_k e^{i\omega_k t}$, where ω_k is the angular frequency of the k -th mode and \mathbf{v}_k is the corresponding eigenvector. Substituting this into the equation of motion:

$$-\omega_k^2 \mathbf{v}_k = -\mathbf{H} \mathbf{v}_k. \quad (3.10)$$

This equation demonstrates that λ_k , the eigenvalue of the Hessian, is related to the angular frequency by:

$$\omega_k^2 = \lambda_k. \quad (3.11)$$

Taking the square root, the normal mode frequencies are given by:

$$\omega_k = \sqrt{\lambda_k}. \quad (3.12)$$

For a protein consisting of N atoms, there are $3N$ degrees of freedom corresponding to the x , y , and z coordinates of each atom. However, six degrees of freedom (three translations and three rotations) are associated with rigid-body motions, leaving $3N - 6$ internal vibrational modes. For linear molecules, five degrees of freedom correspond to rigid-body motions, leaving $3N - 5$ modes.

The eigenvectors \mathbf{v}_k describe the directions of motion in the high-dimensional space of atomic coordinates. Each mode corresponds to a distinct collective motion of the protein, providing insight into its dynamic properties.

3.2.3 Physical Interpretation of Normal Modes in Proteins

In proteins, normal modes describe collective oscillations of atoms that are often highly correlated. These oscillations can be visualized as "breathing," twisting, or bending motions of the protein structure. The low-frequency modes, corresponding to small eigenvalues, are particularly important because they involve large-scale, global movements of the protein, which are typically associated with biological functions such as:

- Opening and closing of ligand-binding sites.
- Allosteric transitions, where a change in one part of the protein affects another distant site.
- Elastic deformations that stabilize the protein under mechanical stress.

The relationship between these motions and protein function is grounded in the energy landscape theory, which posits that the functional motions of a protein are encoded in its native structure and the curvature of its potential energy surface.

3.3 Analytical Solution of stochastic process with normal modes

3.4 Solving the Dynamics Using Normal Modes

The dynamics of the system is governed by the equation:

$$\frac{dr_{i,t}}{dt} = -\sum_j K_{ij}r_{j,t} + \eta_{i,t}, \quad (3.13)$$

or equivalently in vector form:

$$\frac{d\mathbf{r}_t}{dt} = -\mathbf{K}\mathbf{r}_t + \boldsymbol{\eta}_t, \quad (3.14)$$

where: - \mathbf{r}_t is the vector of the system variables, - \mathbf{K} is the Kirchhoff matrix, symmetric and diagonalizable, - $\boldsymbol{\eta}_t$ represents stochastic noise.

3.5 DA QUI IN POI NON GUARDARE

3.5.1 Decomposition in Terms of Normal Modes

The Kirchhoff matrix \mathbf{K} has eigenvalues λ_k and orthonormal eigenvectors \mathbf{v}_k such that:

$$\mathbf{K}\mathbf{v}_k = \lambda_k\mathbf{v}_k, \quad (3.15)$$

with $\lambda_k \geq 0$ and $\lambda_1 = 0$ for the zero mode. To express \mathbf{r}_t in the eigenvector basis, we write:

$$\mathbf{r}_t = \sum_k (\mathbf{v}_k^T \mathbf{r}_t) \mathbf{v}_k, \quad (3.16)$$

where $\mathbf{v}_k^T \mathbf{r}_t$ is the projection of \mathbf{r}_t onto the eigenvector \mathbf{v}_k . Substituting this into the equation of motion, we have:

$$\frac{d}{dt} \left(\sum_k (\mathbf{v}_k^T \mathbf{r}_t) \mathbf{v}_k \right) = -\mathbf{K} \left(\sum_k (\mathbf{v}_k^T \mathbf{r}_t) \mathbf{v}_k \right) + \boldsymbol{\eta}_t. \quad (3.17)$$

Using the orthonormality of the eigenvectors and the property of eigenvalues, we project this equation onto each \mathbf{v}_k :

$$\frac{d(\mathbf{v}_k^T \mathbf{r}_t)}{dt} = -\lambda_k (\mathbf{v}_k^T \mathbf{r}_t) + \mathbf{v}_k^T \boldsymbol{\eta}_t. \quad (3.18)$$

Let $v_k^T \mathbf{r}_t \equiv r_k(t)$, and rewrite the equation as:

$$\frac{dr_k(t)}{dt} = -\lambda_k r_k(t) + \eta_k(t), \quad (3.19)$$

where $\eta_k(t) = \mathbf{v}_k^T \boldsymbol{\eta}_t$ is the noise projected onto the k -th eigenvector.

3.5.2 Solution of the Equation for $r_k(t)$

For $\lambda_k > 0$, the solution to the differential equation is:

$$r_k(t) = r_k(0)e^{-\lambda_k t} + \int_0^t e^{-\lambda_k(t-t')} \eta_k(t') dt'. \quad (3.20)$$

3.5.3 Mean in Terms of Normal Modes

The mean of \mathbf{r}_t is defined as the expected value:

$$\langle \mathbf{r}_t \rangle = \mathbb{E}[\mathbf{r}_t], \quad (3.21)$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator. Using the decomposition in terms of normal modes:

$$\mathbf{r}_t = \sum_k r_k(t) \mathbf{v}_k, \quad (3.22)$$

the mean becomes:

$$\langle \mathbf{r}_t \rangle = \sum_k \langle r_k(t) \rangle \mathbf{v}_k. \quad (3.23)$$

From the solution of $r_k(t)$:

$$r_k(t) = r_k(0)e^{-\lambda_k t} + \int_0^t e^{-\lambda_k(t-t')} \eta_k(t') dt', \quad (3.24)$$

taking the expectation and assuming the noise $\eta_k(t)$ has zero mean ($\mathbb{E}[\eta_k(t)] = 0$), we get:

$$\langle r_k(t) \rangle = r_k(0)e^{-\lambda_k t}. \quad (3.25)$$

This shows that the mean dynamics is determined by the initial conditions and the decay rates of the modes.

3.5.4 Covariance in Terms of Normal Modes (calcoli sbagliati non guardare)

The covariance of \mathbf{r}_t is defined as:

$$\mathbf{C}(t) = \mathbb{E}[(\mathbf{r}_t - \langle \mathbf{r}_t \rangle)(\mathbf{r}_t - \langle \mathbf{r}_t \rangle)^T]. \quad (3.26)$$

Substituting the decomposition $\mathbf{r}_t = \sum_k r_k(t) \mathbf{v}_k$ and using the fact that the modes are independent, the covariance simplifies to:

$$\mathbf{C}(t) = \sum_k \mathbb{E}[(r_k(t) - \langle r_k(t) \rangle)^2] \mathbf{v}_k \mathbf{v}_k^T. \quad (3.27)$$

The variance of each mode $r_k(t)$ is given by:

$$\text{Var}[r_k(t)] = \mathbb{E}[r_k(t)^2] - \langle r_k(t) \rangle^2. \quad (3.28)$$

Using the solution of $r_k(t)$:

$$r_k(t) = r_k(0)e^{-\lambda_k t} + \int_0^t e^{-\lambda_k(t-t')} \eta_k(t') dt', \quad (3.29)$$

and assuming the noise $\eta_k(t)$ is white with covariance:

$$\mathbb{E}[\eta_k(t)\eta_k(t')] = 2D\delta(t-t'), \quad (3.30)$$

we can calculate the variance of $r_k(t)$. The expectation of the squared integral is:

$$\mathbb{E} \left[\left(\int_0^t e^{-\lambda_k(t-t')} \eta_k(t') dt' \right)^2 \right] = \frac{D}{\lambda_k} (1 - e^{-2\lambda_k t}). \quad (3.31)$$

Thus, the variance of $r_k(t)$ is:

$$\text{Var}[r_k(t)] = \frac{D}{\lambda_k} (1 - e^{-2\lambda_k t}). \quad (3.32)$$

Substituting this into the covariance matrix:

$$\mathbf{C}(t) = \sum_k \frac{D}{\lambda_k} (1 - e^{-2\lambda_k t}) \mathbf{v}_k \mathbf{v}_k^T. \quad (3.33)$$

For the zero mode ($\lambda_1 = 0$), the variance grows linearly with time, as expected for a diffusive process.

$$\text{Var}[r_1(t)] = 2Dt. \quad (3.34)$$

Thus, the full covariance matrix is:

$$\mathbf{C}(t) = 2Dt \mathbf{v}_1 \mathbf{v}_1^T + \sum_{k>1} \frac{D}{\lambda_k} (1 - e^{-2\lambda_k t}) \mathbf{v}_k \mathbf{v}_k^T. \quad (3.35)$$

This shows that the covariance is dominated by the zero mode at long times, while higher modes contribute transiently with exponentially decaying terms.

3.5.5 Response Function

The response function $R_{ij}(t)$ is defined, following (2.7), as: [12]

$$R_{ij}(t) = \frac{C_{ij}(t)}{C_{ij}(0)}, \quad (3.36)$$

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

$$C_{ij}(t) = \int_0^t \int_0^t e^{-\mathbf{K}(t-s)} \mathbf{Q} e^{-\mathbf{K}^\top(t-u)} ds du, \quad (3.37)$$

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

$$C_{ij}(0) = \langle \mathbf{r}_{i,0} \mathbf{r}_{j,0} \rangle. \quad (3.38)$$

Thus, the response function becomes:

$$R_{ij}(t) = \frac{\int_0^t \int_0^t e^{-\mathbf{K}(t-s)} \mathbf{Q} e^{-\mathbf{K}^\top(t-u)} ds du}{C_{ij}(0)}. \quad (3.39)$$

3.5.6 Transfer Entropy

The transfer entropy $TE_{j \rightarrow i}(t)$ is defined, following (2.8), as:

$$TE_{j \rightarrow i}(t) = -\frac{1}{2} \ln \left(1 - \frac{\alpha_{ij}(t)}{\beta_{ij}(t)} \right), \quad (3.40)$$

So it is completely determined by the correlation.

Experimental Set-up and Results

Detailed Presentation of Dataset 3LNX

The Protein Data Bank entry ****3LNX**** corresponds to a photoswitchable PDZ domain engineered to undergo light-induced conformational changes. This protein system incorporates an azobenzene-based photoswitch, enabling a reversible transition between distinct conformational states upon specific illumination conditions. Such light-triggered isomerization events alter the protein's fold, side-chain orientations, and overall dynamic properties, ultimately influencing the protein's binding capabilities and functional behavior.

The structure of 3LNX was determined using solution-phase NMR spectroscopy, providing a detailed ensemble of conformations that represent the domain under defined conditions. These data capture the subtle variations within the structural ensemble, highlighting how the presence of the azobenzene moiety modulates local and long-range interactions. In particular, the modifications in the ligand-binding groove—an essential feature of PDZ domains—can be correlated with changes in ligand accessibility, partner selection, and interaction specificity.

The conformational states observed in the 3LNX dataset span a range of structural compactness and side-chain packing arrangements. These variations influence the intrinsic flexibility and stability of the PDZ domain, as reflected in the B-factor data. By analyzing these fluctuations, researchers gain insight into the dynamic landscape that underpins the protein's functional transitions and understand how external stimuli (in this case, light) can guide the conformational equilibrium.

Overall, the 3LNX dataset serves as a fundamental reference point for studying light-responsive protein conformational changes. It provides a framework not only for understanding the molecular basis of photoswitching in engineered protein systems but also for guiding the rational design of future photoreactive proteins with tailored functionalities.

Structural Highlights:

- **Overall Structure:** The PDZ domain of 3LNX contains multiple β -strands and α -helices that create a canonical PDZ fold, complete with a characteristic ligand-binding groove.
- **Atomic Representation:** The azobenzene derivative is covalently attached, acting as a molecular photoswitch. Its isomerization influences the positioning of surrounding residues and the backbone conformation.
- **Functional Elements:** The ligand-binding groove is clearly visible, providing



Figure 3.1. Representation of the 3LNX dataset highlighting the photoswitchable PDZ domain. Atoms are colored by element: O (red), N (blue), C (beige), and H (white). The yellow feature indicates the azobenzene moiety.

insight into how photoswitching might modulate binding events by altering groove shape, depth, and accessibility.

Photoswitching Properties: The azobenzene moiety in 3LNX responds to light stimuli by switching between two isomeric states (often denoted as *cis* and *trans*). Each state imposes a different set of constraints on the protein's backbone and side-chain orientations. As a result, the local environment around the groove and adjacent secondary structure elements shifts, enabling researchers to study how structural rearrangements translate into functional changes.




figures/3LNX_graph.png

Figure 3.2. Graphical analysis of the 3LNX ensemble, showing variations in structural parameters, such as backbone dihedral angles and B-factors, across the conformational ensemble.

Connection Radius

A critical aspect of analyzing 3LNX’s structural data is the choice of the connection radius when constructing the Kirchhoff matrix. This matrix forms the basis of elastic network models (ENMs) and is widely used to probe the collective motions within proteins. By selecting an appropriate connection radius, one captures both local and long-range interactions between residues, ensuring that the resulting network accurately reflects the protein’s mechanical and dynamic properties.

In the case of 3LNX, setting the connection radius allows for an unbiased exploration of how the azobenzene switch affects communication pathways throughout the structure. Perturbations caused by the switch may propagate along the protein fold, influencing distant regions and modulating conformational states and functional outcomes.



figures/3LNX_Kirchhoff.png

Figure 3.3. Visualization of the Kirchhoff matrix derived from the 3LNX dataset. Different connection radii alter the interaction networks considered in the model, affecting the predicted collective motions.

Kirchhoff Matrix

Applying (1.5) 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:

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

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

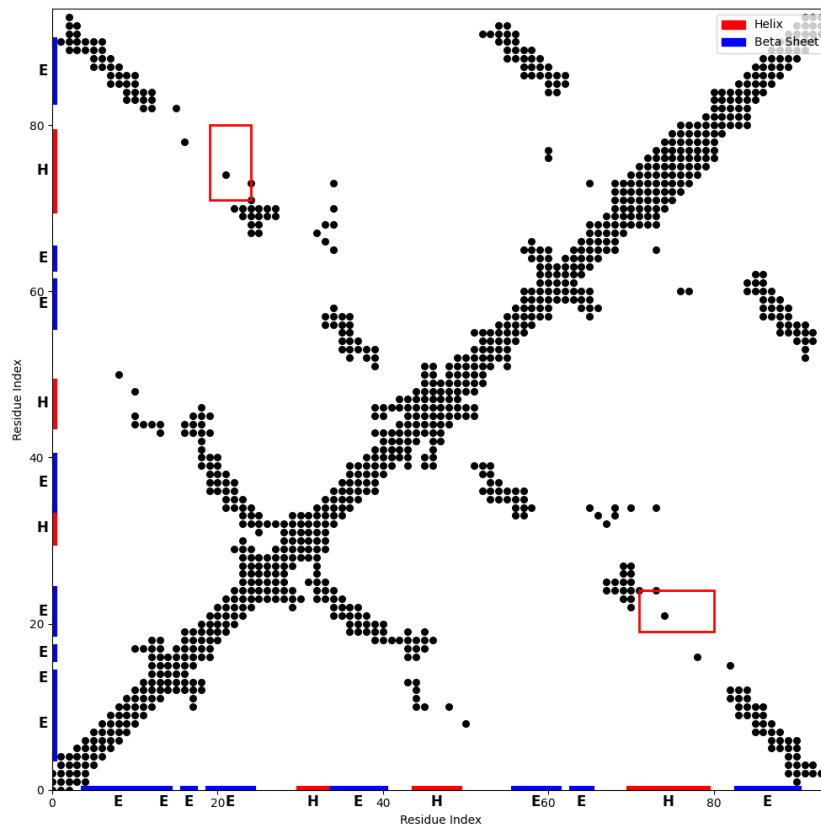


Figure 3.4. Kirchhoff Matrix.

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 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 correlation 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.

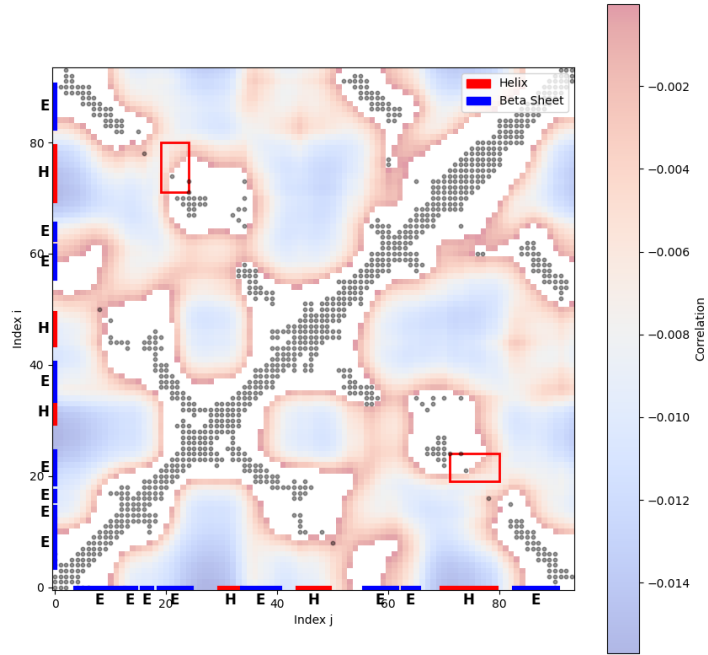


Figure 3.5. Dataset of 2M0Z.

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.
 - 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):**

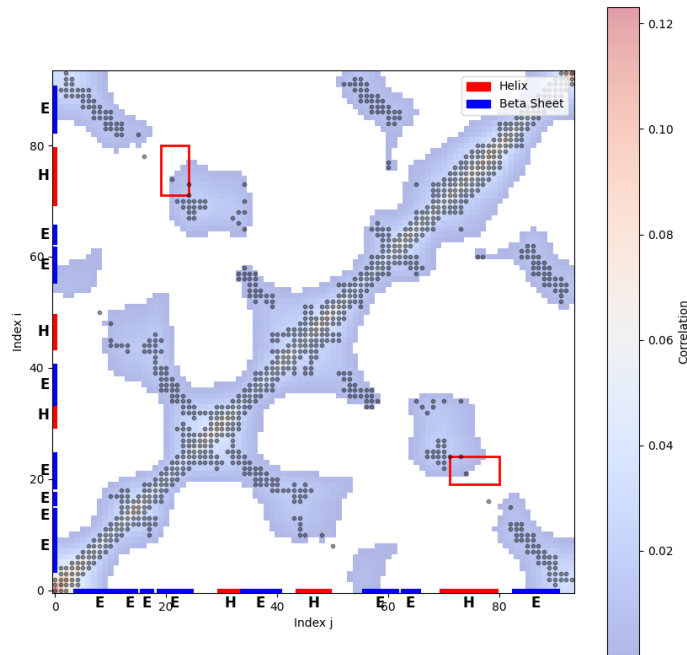


Figure 3.6. Dataset of 2M0Z.

- 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.
 - **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.

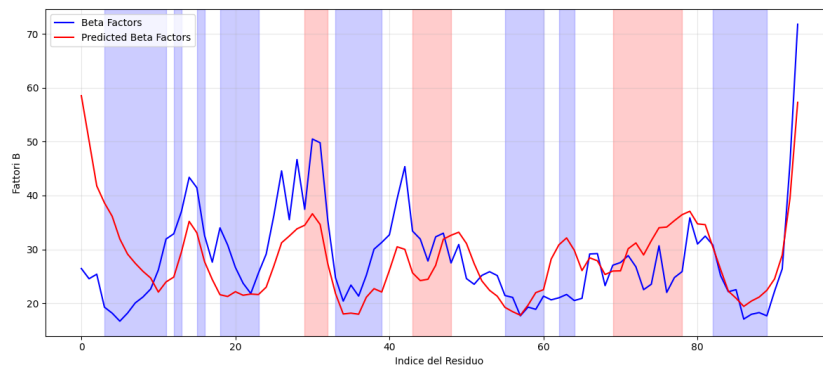


Figure 3.7. Beta factors.

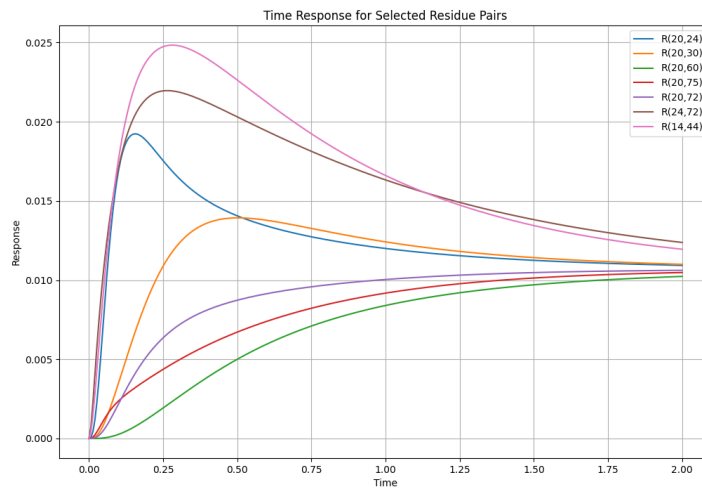
Responses in Time

Figure 3.8. Responses in time.

Tau

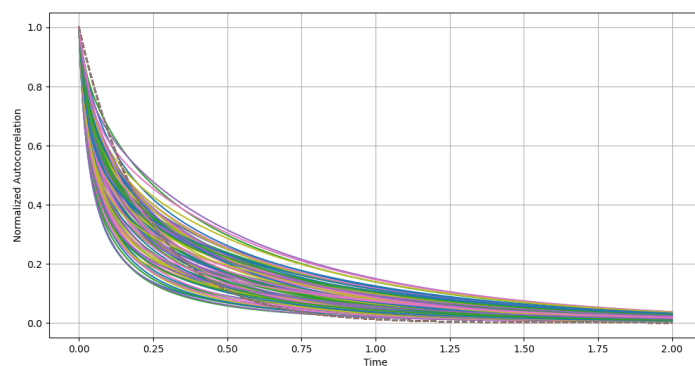


Figure 3.9. Responses in time.

Mean First time passage

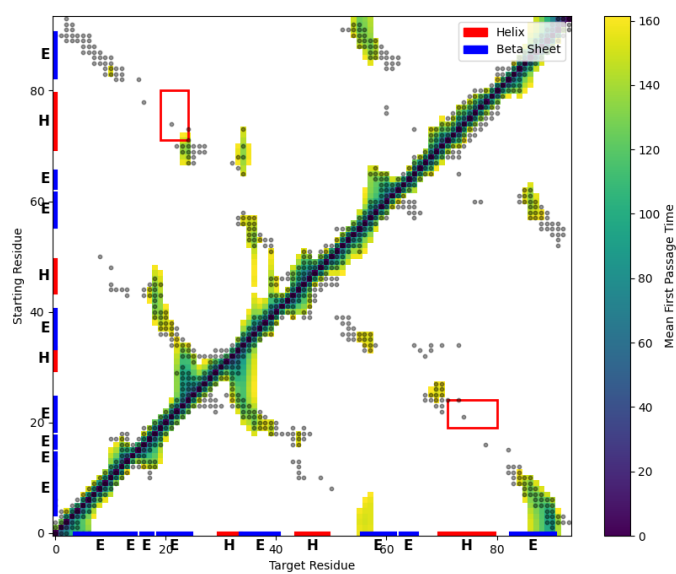


Figure 3.10. Responses in time.

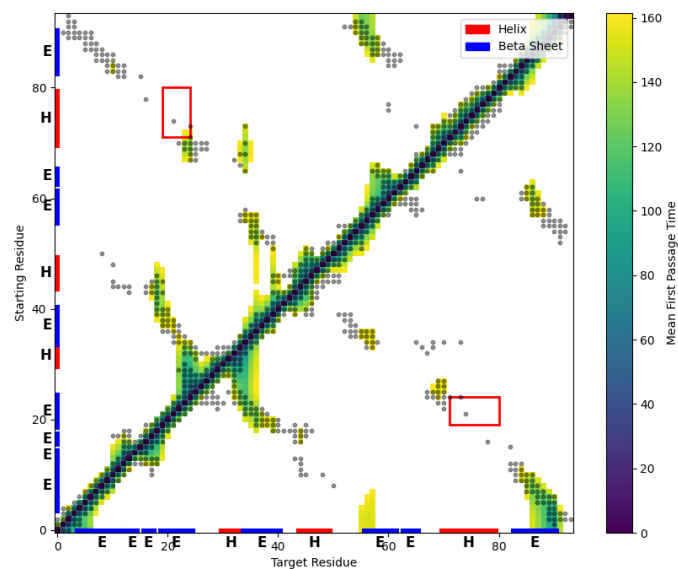


Figure 3.11. Responses Matrix.

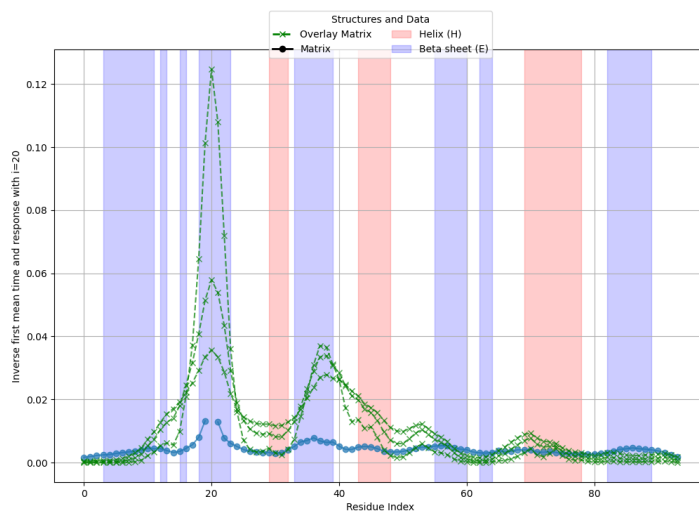


Figure 3.12. Inverse first mean time perturbing 21-th residue.

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, which

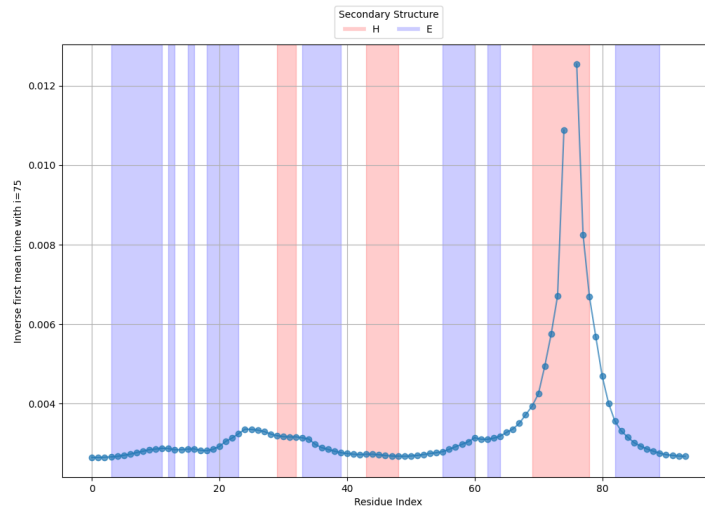


Figure 3.13. Inverse first mean time perturbing 76-th residue.

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 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.
- **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.

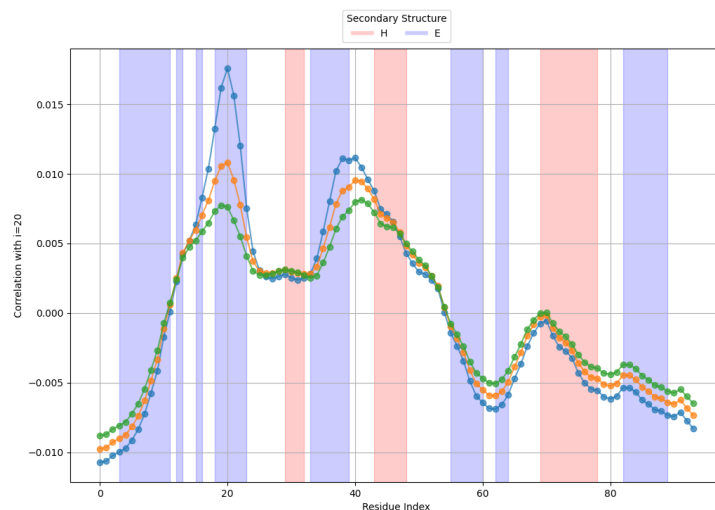


Figure 3.14. Correlation perturbing 21-th residue.

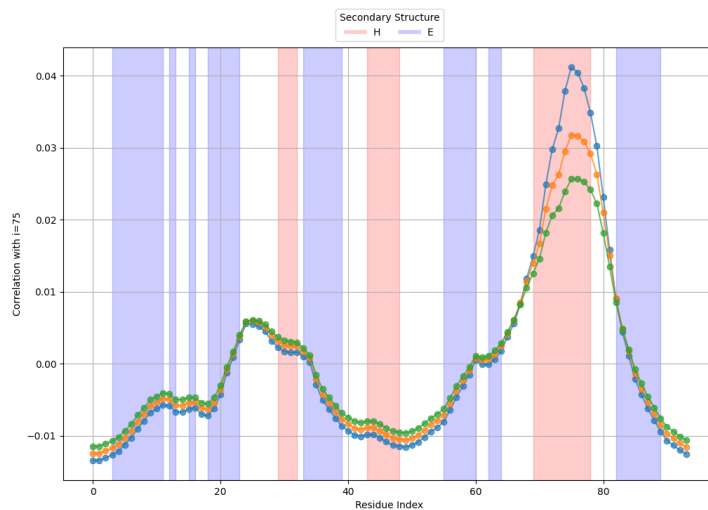


Figure 3.15. Correlation perturbing 76-th residue.

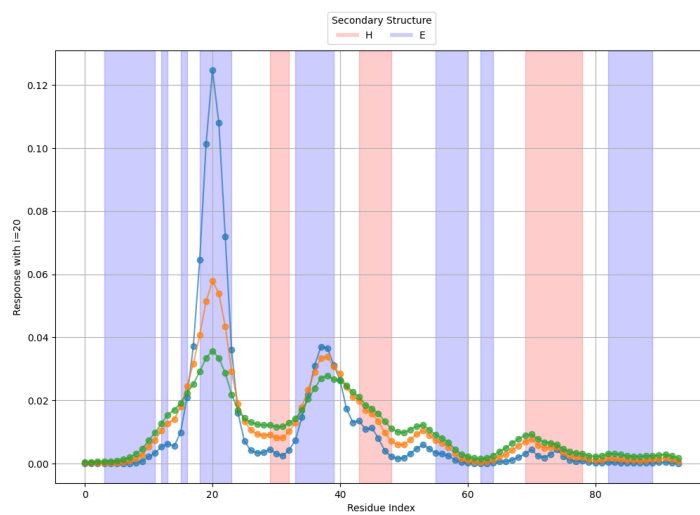


Figure 3.16. Response perturbing 21-th residue.

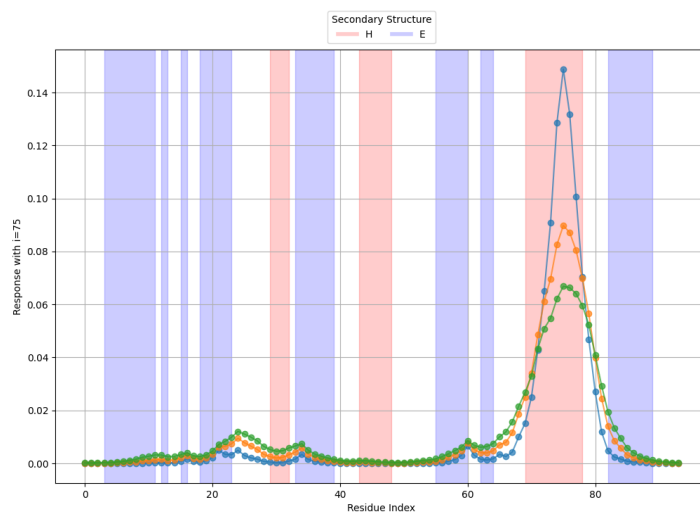


Figure 3.17. Response perturbing 76-th residue.

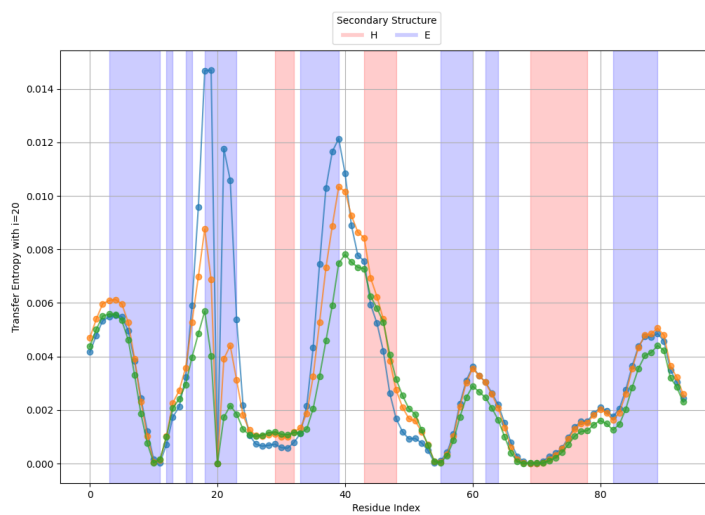


Figure 3.18. Transfer Entropy (21->j) perturbing 21-th residue.

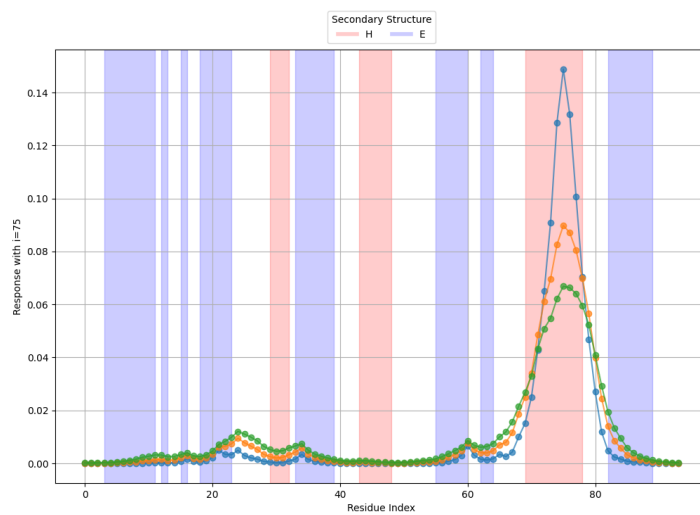


Figure 3.19. Transfer Entropy (76->j) perturbing 76-th residue.

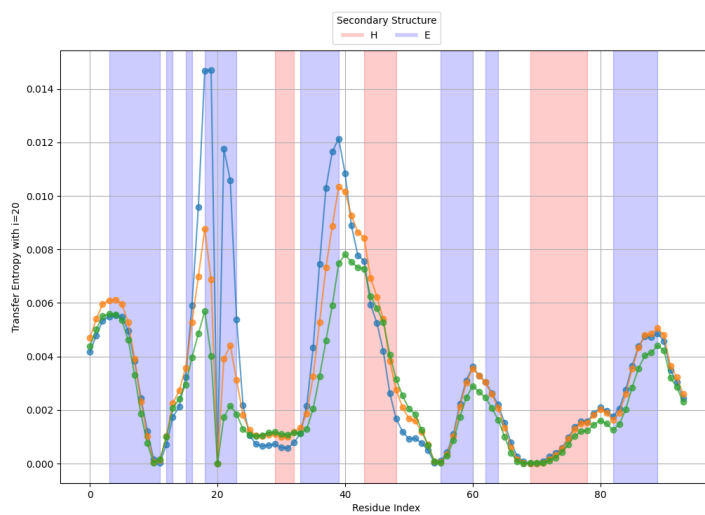


Figure 3.20. Transfer Entropy (21->j) perturbing 21-th residue.

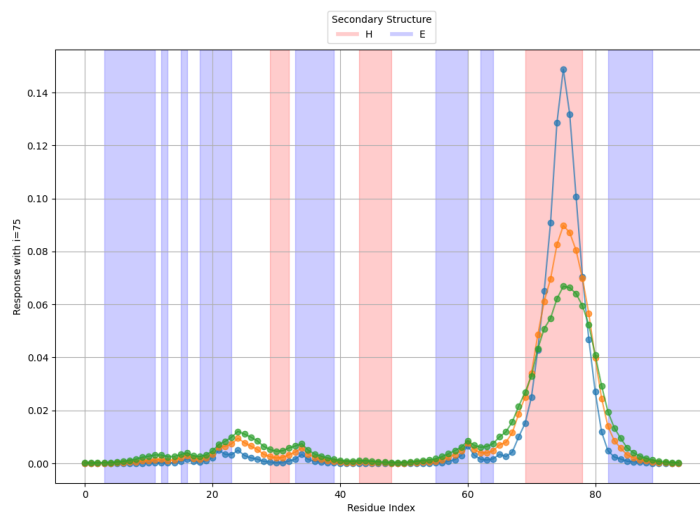


Figure 3.21. Transfer Entropy (76->j) perturbing 76-th residue.

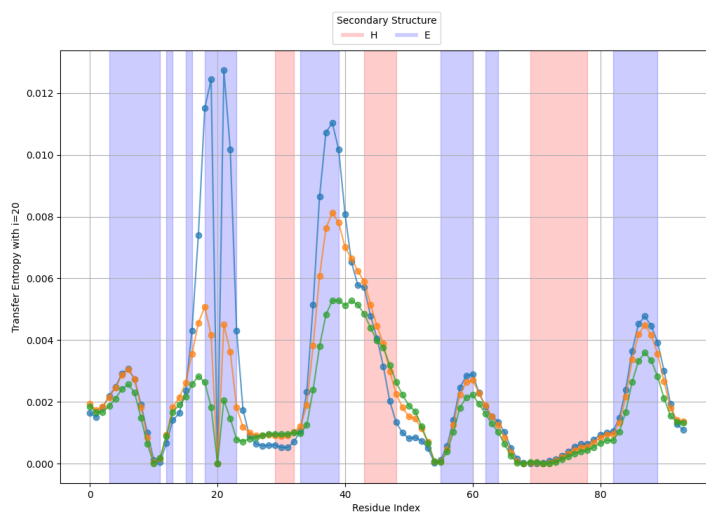


Figure 3.22. Transfer Entropy ($j > 21$) perturbing 21-th residue.

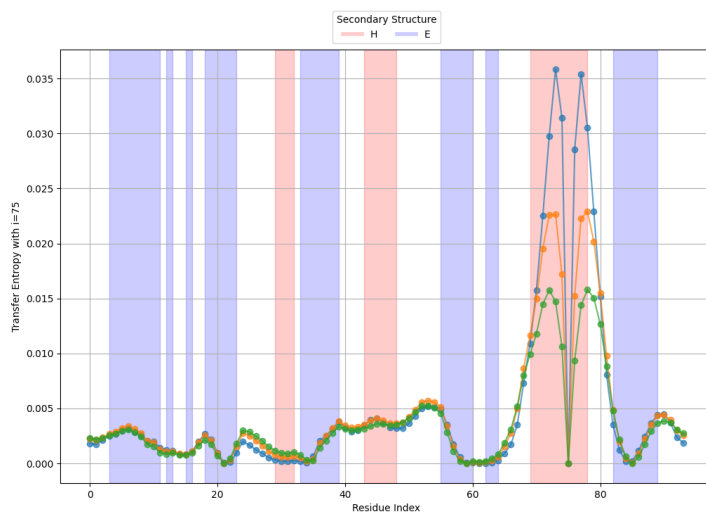


Figure 3.23. Transfer Entropy ($j > 76$) perturbing 76-th residue.

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 of California Davis, Introductory Biology*, https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_%28Easlon%29/Readings/04.3%3A_Amino_Acids
- [3] *University of California Davis, Introductory Biology*, https://bio.libretexts.org/Courses/University_of_California_Davis/BIS_2A%3A_Introductory_Biology_%28Britt%29/01%3A_Readings/1.17%3A_Protein_Structure
- [4] Nature, <https://www.nature.com/scitable/topicpage/protein-structure-14122136/>
- [5] Wikipedia, https://en.wikipedia.org/wiki/Allosteric_regulation?utm_source=chatgpt.com
- [6] Statistical Mechanics of Allosteric Enzymes
- [7] University of California Davis, Introductory Biology, Hemoglobin and allosteric effects
- [8] Allosterism in the PDZ Family, Amy O. Stevens and Yi He
- [9] Protein elastic network models and the ranges of cooperativity, Lei Yanga, Guang Songa, and Robert L. Jernigana
- [10] Introduzione alla Teoria dei Grafi, Vittorio Loreto, Francesca Tria.
- [11] Time- and ensemble-average statistical mechanics of the Gaussian network model, Alessio Lapolla, Maximilian Vossel, and Aljaz Godec
- [12] Correlation, response and entropy approaches to allosteric behaviors: a critical comparison on the ubiquitin case Fabio Cecconi, Giulio Costantini, Carlo Guardiani, Marco Baldovin and Angelo Vulpiani
- [13] Robust inference of causality in high-dimensional dynamical processes from the Information Imbalance of distance ranks Vittorio Del Tatto, Gianfranco Fortunato, Domenica Buetti, and Alessandro Laio