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

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

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

Contents

Introduction

In this thesis we will study the allosteric mechanism of a protein in and out of equilibrium conditions.
Allostery is the phenomenon by which the binding of an effector molecule to a specific site on a protein, known as the allosteric site, induces a conformational change that affects the functional activity of a distant site, typically the active site. This interaction does not occur through direct binding between the sites but rather through a series of internal network changes within the protein structure, which transmits the signal and ultimately modulates its function. We want to understand it in a quantitative way, using causal indicators for studying the propagation of the signal along the structure of the protein.

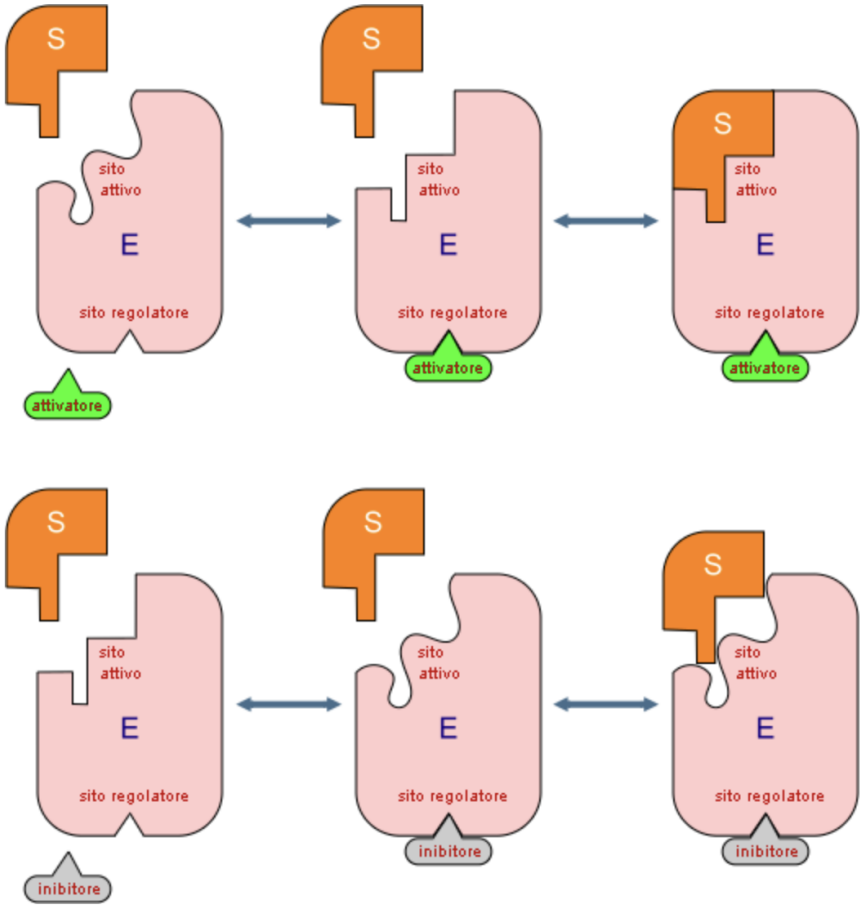


Figure 0.1. Schema of allosteric mechanism with active and allosteric site

Chapter 1: Protein Structure In the first chapter we will explain what a protein is and the basic concept about it.

Chapter 2: Relation structure-function-dynamic in proteins In the second chapter we will expose the importance relation structure-function-dynamic in proteins. It is very important the geometry structure of the protein and one of my aim is to show how much of the protein's behavior can be explained only from its geometric structure. In fact it comes from years and years of evolutionary optimization process.

Chapter 3: Relation structure-function-dynamic in proteins In the third chapter we will study the normal modes and the guassian network model as a model to trehat the protein as a network of interactions between atoms. This is a very useful model to study the protein's behavior and in particular the allosteric mechanism, in which model the geometric structure is the only important factor to explain the protein's behavior. Finaly we will introduct the allosteric mechanism and we will see specifically in 3LNX protein, where we expect that a signal from an allosteric

site can propagate to the hydrophobic pocket, that is the active site of the protein, avoiding the protein to catch a ligand.

Chapter 4: Causality indicators We will study also the causality indicators, that are a powerful tool to study the propagation of the signal along the protein's structure and to undestrand the allosteric mechanism. We will present the covariace, the response and the transfer entropy as causal indicators.

Chapter 5: Causality indicators We will show the result and how good this model explain the allosteric behavior of the protein.

Chapter 1

Proteins

1.1 Structures of Amino Acids and Proteins

Our objective is to explore how the amino acids making up the protein interact with each other.
Specifically we aim to comprehend allsoteric mechanisms . First we need to understand what is a protein.
Proteins are composed of a sequence of amino acids that fold into a three-dimensional structure which determines their function.
Some of the functions include: accelerating chemical reactions by lowering the activation energy (Catalysis); regulating signal transduction and gene expression (Regulation and Signaling); the transport of molecules through cell membranes (Transportation); and the providing of mechanical support and structural integrity to tissues (Structural roles).
But what is an amino acid?
Amino acids are the building blocks of proteins, they are organic molecules which contain a central carbon atom (α-carbon) bonded to an amino group (-NH2), a carboxyl group (-COOH), a (R-Group), that changes from amino acid to amino acid, and finally to a hidrogen atom.
The R-group is particularly important because it determines the chemical properties of the amino acid, such as whether it is hydrophilic, hydrophobic, acidic, or basic.

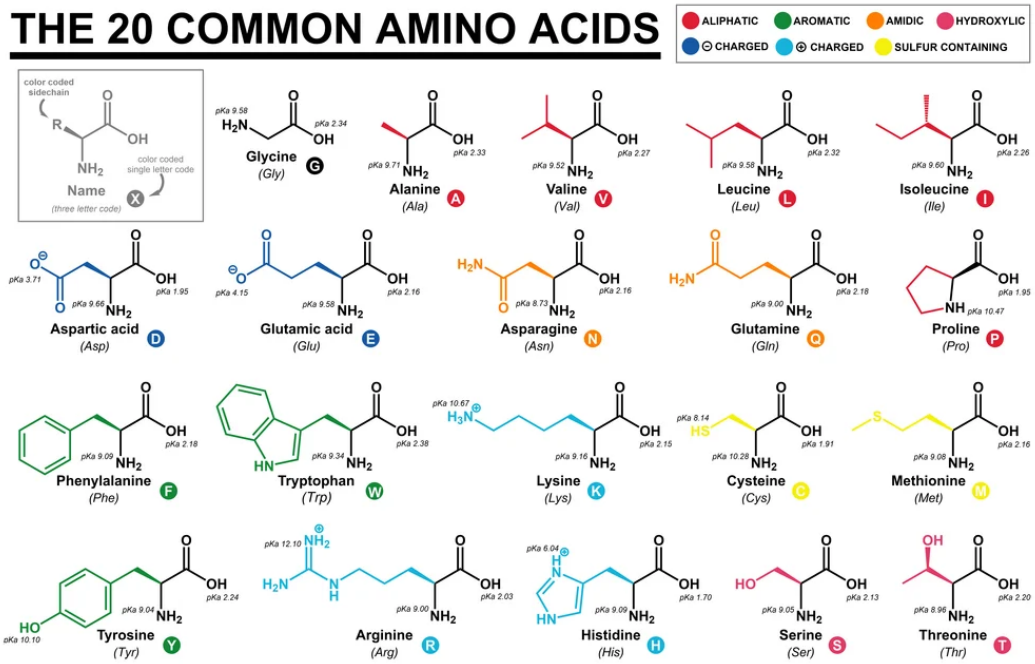


Figure 1.1. Amino Acids

The structure of a protein is divided into four levels which are very important for understanding its function.[4]
The primary structure is a linear sequence of amino acids in a protein chain, held together by peptide bonds.
Instead the secondary structure represents the local folding of the protein chain in patterns. In this work we will focus in the following secondary structures in particular on alpha-helix, beta-sheet and loop regiones.
The first two are ordered patterns in which every amino acid is connected with the other amino acid in a specific way, the loop instead is a disordered region of the protein.
In particular alpha-helix is a right-handed spiral structure stabilized by hydrogen bonds between the NH and CO groups of amino acids separated by four residues, while beta-sheet is a structure in which the amino acids are connected by hydrogen bonds in a zigzag (parallel or antiparallel) pattern and finally the loop is a region of the polypeptide chain that connect other secondary structures.
The tertiary structure describes the three-dimensional folding of the entire protein molecule and quaternary structure consist of more than one polypeptide chain that are arranged and interact with each other to form the functional protein complex.
Obviously all these structures arise from the interactions between the amino acids that compose the protein. Following we will see our model of interactions between the amino acids.

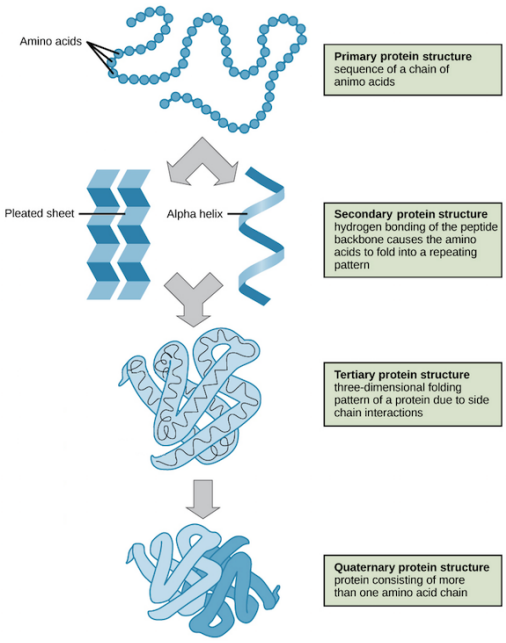


Figure 1.2. Structures

Chapter 2

Structure-Function-Dynamic Relationship in Proteins

2.1 Relation between Structure, Function, and Dynamics in Proteins

The fundamental idea is that a protein's structure dictates its function. This structure governs the interactions between amino acids within the protein. If proteins were random aggregates of amino acids, they would not be able to perform their specific functions.
In fact the evolution optimizes the protein's utility function, creating a structured arrangement that is not random. Finally this structure determines how signals propagate along the protein.

Proteins rely heavily on their ability to change shape in order to carry out their functions effectively; this involves more, than movement but also includes interactions between distant areas within the protein structure that are crucial for important biological processes, like attaching to other molecules or binding with ligands.
Proteins can undergo variations by either moving their acids around a central point or interacting with other molecules to alter their shape.
Therefore we propose (which is the assumption of this study) that the structural shape of the protein itself reveals a lot, about how the allosteric process works.
So we're going to create a model that relies completely on structure and try to clarify every detail. Obviously, there are many additional factors that we must consider to accurately model the behavior of proteins.
Obviously, there are many additional factors that we must consider to accurately model the behavior of proteins.

2.2 Allostery

Allostery, derived from the Greek *allos* (other) and *stereos* (structure), is the phenomenon of a change in protein structure caused by the transmission of a signal from one site to another.[5] More precisely, allostery occurs when the interaction of a molecule (effector) with a specific site on a protein, known as the allosteric site, induces a conformational change that influences the functional activity of another site, usually the active site. This process does not involve direct interactions between the two sites but occurs through changes in the protein's internal network.

There are two key types of allostery that govern how proteins regulate their functions: conformational allostery and dynamic allostery. While both are based on the transmission of a signal through the protein's structure, they differ in the mechanism of action. Conformational allostery involves a structural change in the protein that occurs when an effector molecule binds to an allosteric site. This binding induces a distinct, often visible, shift in the protein's shape, such as a transition from a less active (relaxed) state to a more active (tense) state. This structural shift propagates through the protein, modifying its function. Mathematically, conformational allostery can be described in terms of the population of microstates of the protein, as well as the equilibrium constant between different states, such as:

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

where $[T]$ and $[x]$ represent the relative concentrations of the two conformational states, relaxed and tense, and L describes the equilibrium constant between them.

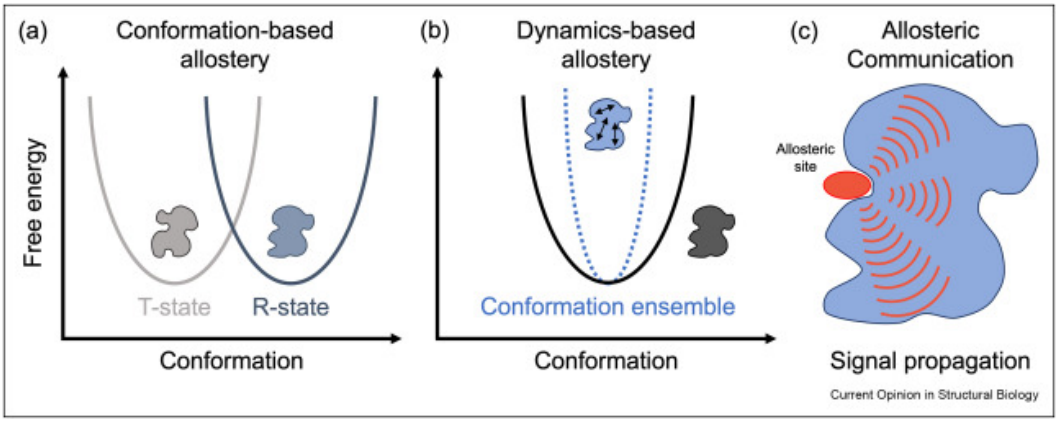


Figure 2.1. Schematic representation of allostery: the interaction with an allosteric effector (site A) induces a conformational change that affects a distant site (site B).

In contrast, dynamic allostery is based on fluctuations and flexibility within the protein's structure. Instead of a fixed structural change, dynamic allostery involves the modulation of the protein's internal dynamics, where small oscillations or fluctuations in the protein's atomic structure propagate across its network of interactions. These dynamic changes influence the protein's ability to bind ligands or interact with other biomolecules.

Dynamic allostery often does not result in obvious structural shifts but involves

subtle changes in the protein's flexibility. These changes are essential for the protein's function, as they allow the protein to respond to environmental signals without requiring large-scale conformational changes.

Both conformational and dynamic allostery rely on causal mechanisms—specifically, how an effector molecule's binding event at one site causes changes at a distant site. However, the distinction lies in how these signals are transmitted.
In conformational allostery, this is typically achieved through a direct structural change, while in dynamic allostery, it occurs via shifts in the protein's internal network dynamics, often without a distinct change in overall structure.
To fully grasp how proteins achieve their functional structure, it is crucial to explore the causal mechanisms behind these conformational and dynamic changes. Understanding how the signal propagates between amino acids and how localized binding events can influence distant regions of the protein is essential for studying allosteric mechanisms. This propagation of information is key to explaining how localized interactions can regulate the overall function of the protein, thus facilitating complex biological processes. This understanding is especially important in drug development, where manipulating these allosteric sites and their causal mechanisms can lead to the creation of more effective therapeutic strategies.

A classcial exemple of allostery is the HIV protease.
The HIV protease is a key enzyme in the maturation of the HIV virus. After infecting a host cell, the virus produces a long chain of amino acids that must be cleaved into smaller peptides to form a functional, infectious protein. This process is regulated by allosteric mechanisms, where the binding of specific molecules to the protease induces conformational changes, affecting its ability to cleave the peptide chain and thus regulating viral maturation.

These cuts are essential for the virus to assemble the amino acids and mature into an infectious form. Therefore, inhibiting the protease's cutting ability can limit viral growth. This is exactly how protease inhibitors work in antiviral treatments.

This conformational change induced by allostery is a prime example of the importance of understanding allosteric mechanisms in proteins, as it enables us to intervene in processes crucial for viral replication and devise more effective treatments.

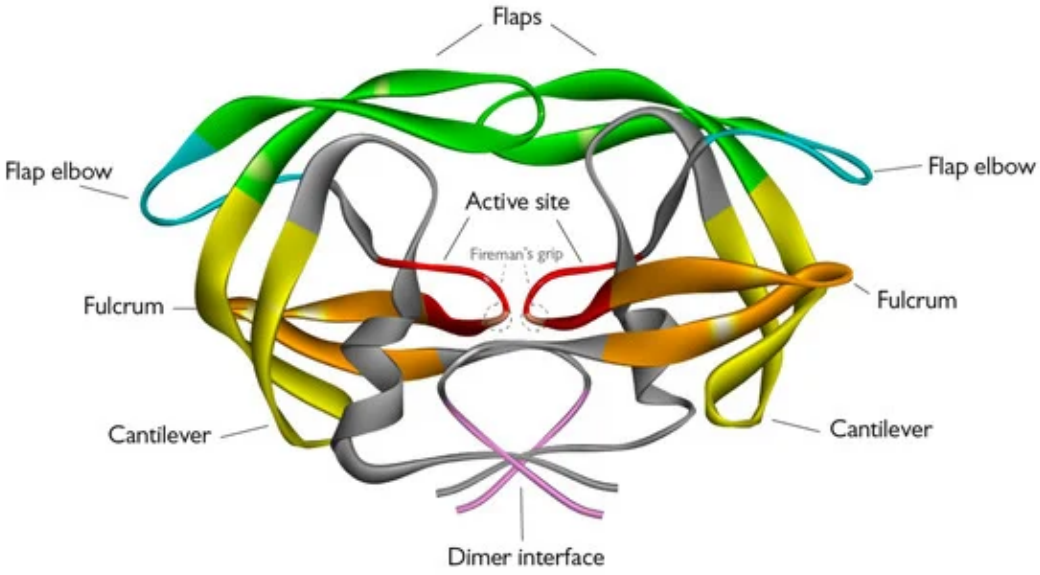


Figure 2.2. Structure of HIV Protease.

2.3 PDZ Domain and 3LNX

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]
They are tipycally characterized by two main characteristics; the first it is that the domain (as we said a portion of the protein) can fold into a three-dimensional structure on its own, the second is that it can play an indipendent role when it is unit with a protein. In addition, because it is a modular unit, it is used by other proteins to form more complex structures.[8]
Proteins often consist of multiple domains arranged in various combinations, creating multifunctional macromolecules capable of complex interactions.

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 (Postsynaptic density protein 95,Drosophila discs large protein, Zona Occludens), it is a modular domain, its function is to organize and stabilize large multiprotein complexes, maintaining cellular architecture and facilitating biochemical signaling pathways. [8]
The PDZ domain typically consists of 80–90 amino acid; facilitates protein-protein interactions by recognizing specific peptide sequences, often located at the C-terminal region of target proteins; assemble protein complexes at cellular membranes In this work we will see the 3LNX protein, which is a PDZ-specific domain.
The 3LNX protein offers important insights into the functioning of these domains. This protein has a hydrophobic pocket that accommodates a external peptide, that binds with the 3LNX.
It is constitute from three alpha helix, that we will call alpha-α, alpha-β and $alpha-\gamma$, and six different beta-sheet, that we will call beta-α, beta-β, $beta-\gamma$,$beta-\delta$, $beta-\epsilon$,$beta-\eta$.

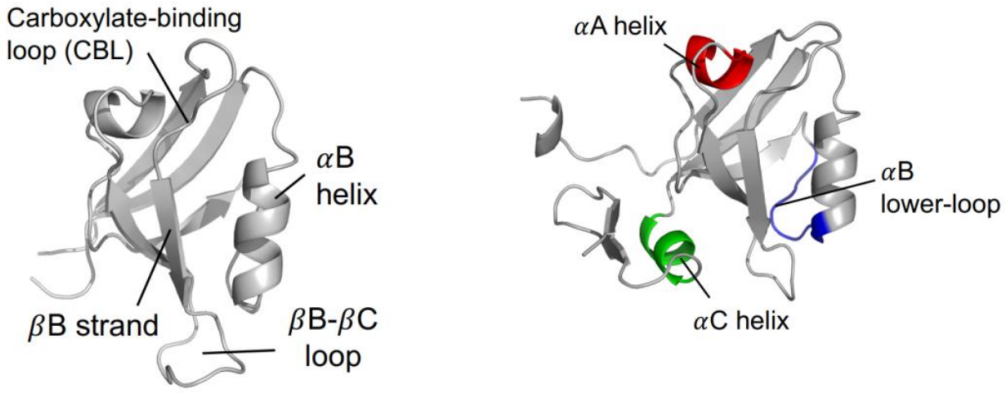
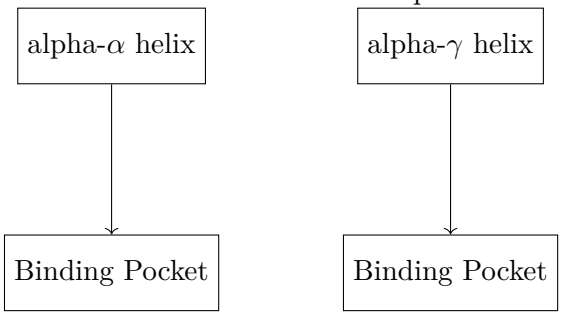


Figure 2.3. 3LNX protein

The allosteric mechanism works as follows: a perturbation at one allosteric site propagates a signal towards the hydrophobic pocket, located between the β-sheet and the α-helix.
This pocket can transition between open and closed states, which is the essence of allostery, preventing a ligand from binding or unbinding at the active site within the hydrophobic pocket.
These allosteric sites do not directly interact with the ligand but are crucial for transmitting structural changes within the protein upon ligand binding. In this study, we will investigate in detail two potential allosteric sites, identified by experimental biologists, located in the α-helix and the γ-helix, as suggested by previous research. [14].
So this is the schema of what we expect:



Chapter 3

Normal and functional mode Analysis

3.1 Introduction

We aim to study how the atoms comprising a protein oscillate around their equilibrium positions, thereby propagating signals and interacting with one another.
For this investigation, we utilize Normal Mode Analysis (NMA).
It is a technique that delineates a system in terms of its normal modes.
These modes provide a quantitative method to evaluate the vibration intensity of each atom in the system.
Arising from a harmonic (Gaussian) approximation, the normal modes can be collective movements at either low or high frequencies.
This technique precisely describes the vibrations of atoms at their equilibrium points.
Using this method, we can identify the modes and the most critical sites for allosteric mechanisms.
Thus, NMA establishes a connection between structure, function, and dynamics; indeed, by starting with the geometric structure of the protein, we can compute the normal modes and employ them to examine allostery in proteins. Hence, NMA serves as a theoretical bridge linking the structure and function of a protein.

3.2 Gaussian Network Model and Normal Mode Analysis

For studying the oscillation of the atoms around their equilibrium position we need to start defining the Hamiltonian of 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 and on bonding forces between atoms, electrostatic and Van der Waals interactions, solvent forces, and many additional factors.
As it clear, it is very complex to develop a real world model in this field, however, this is not our goal.
In fact our aim is to understand how much of the allosteric mechanism in proteins is

explainable only from its geometric factor.
Thus we represent the Hamiltonian of the system as H=V(r) where r is the vector containing the positions of all the atoms that make up the protein, so it depends only from this positions and not from other factors.
So we bet on the assumption that the protein's dynamics can be effectively described using only its geometric structure, specifically focusing on the positions of the backbone atoms (alpha-carbons), while neglecting finer details such as side-chain interactions or solvent effects.
In fact we believe that years of evolutionary optimization have led to a protein structure that is highly efficient and optimized for its function.
While this simplification allows for manageable computational complexity, it may overlook important local and long-range interactions.
Also in this way due to the complexity of the system, it is not feasible to solve it exactly.
Therefore, we rely on approximations, such as the second-order expansion of the potential energy around an equilibrium point, which provides a simplified representation of the protein's behavior.
However, this approximation neglects higher-order interactions that could influence the protein's dynamics, and as a result, it may fail to capture certain intricate behaviors, especially in regions where the protein undergoes significant conformational changes.
Mathamically the second-order approximation of a function $V(r)$, around an equilibrium point r_0, can be expressed as:[9]

$$V(r) \approx V(r_0) + \nabla V(r_0)^\top (r - r_0) + \frac{1}{2}(r - r_0)^\top H_V (r - r_0), \tag{3.1}$$

where:

- $V(r_0)$ is the value of the function at the equilibrium point r_0.
- $\nabla V(r_0)$ is the gradient of the function, defined as:

$$\nabla V(r) = \left.\frac{\partial V}{\partial r}\right|_{r=r_0}, \tag{3.2}$$

  which equals zero if r_0 represents a minimum point, corresponding to the equilibrium position.

- H_V is the Hessian matrix of $V(x)$, defined as:

$$H_V = \left.\frac{\partial^2 V}{\partial r^2}\right|_{x=x_0} = \begin{bmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_1 \partial x_{3N}} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2^2} & \cdots & \frac{\partial^2 V}{\partial x_2 \partial x_{3N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial x_{3N} \partial x_1} & \frac{\partial^2 V}{\partial x_{3N} \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_{3N}^2} \end{bmatrix}.$$

This matrix format ensures that all interactions between positional variables are considered for each pair of dimensions, crucial for accurately modeling the dynamics and for the normal modes.

Because working in high dimensions is complex, we will also introduce an isotropic approximation of the system (In the following lines, we will examine the drawbacks.). However, the system can be easily generalized to the anisotropic case as well.
Now we have a gaussian model which is a springs model, where every atom is connected with the other atoms with a spring.
If we want that only some residues are connected we can put a cutoff distance, so that the spring is connected only with the atoms that are in a certain distance.
In this way hamiltonian become:

$$H(x) = \frac{g}{2} \sum_{i,j} x_i K_{ij} x_j. \tag{3.4}$$

Where K is the Kirchhoff matrix which is the Hessian matrix of the system and g is a adjustable energy scale that can be set by matching the theoretical mean square displacement of the displacement from their native positions.
The Kirchhoff matrix is a matrix that represents the connections between nodes in a graph, without encoding information about its physical layout.
For a graph G with n vertices, the Kirchhoff matrix K is defined as:[10]

$$K = D - A \tag{3.5}$$

Where D is the degree matrix, a diagonal matrix defined as:

$$D_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \tag{3.6}$$

where $\deg(v_i)$ is the degree of vertex v_i, the number of edges connected to vertex (nodes) i; and A is the adjacency matrix, a square matrix defined as:

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

By the symmetry property of K, it follows that it is diagonalizable, it is positive semi-definite, and its smallest eigenvalue is always 0. Because it is positive semi-definite, K does not have a conventional inverse, but we can obtain the pseudo-inverse which is the matrix K^+ that minimizes the following equation:

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

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

This modeling is called Gaussian Network Model (GNM) and is commonly used to study the dynamic of proteins.
So to summarize this is a model which take in consideration only the geometric structure of the protein and no one of the other type of interactions.
In fact the GNM assumes that interactions between atoms in a protein can effectively be represented as elastic springs with uniform force constants. This linearity

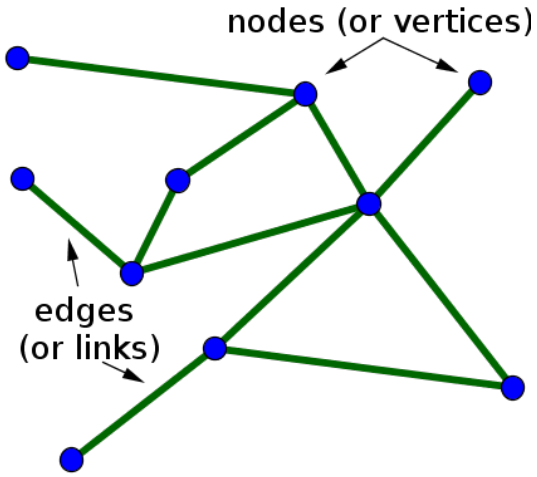


Figure 3.1. Schematic representation of a Network.

assumption implies that any deviation from equilibrium can be treated as a small, reversible perturbation that does not substantially alter the basic structure of the protein.
In addition it ignores the solvent effects and he model does not distinguish between different types of bonds, such as hydrogen bonds, van der Waals interactions, or electrostatic interactions, treating all interactions as if they had the same impact on protein dynamics.
Moreover, the GNM inherently assumes isotropy in the protein's interaction network, treating the protein structure as if it uniformly responds to perturbations in any direction. This isotropic assumption simplifies the mathematical treatment of protein dynamics by not requiring differential responses based on directional changes. While this makes the model computationally efficient, it overlooks the potential anisotropy in biological molecules, where directional dependencies are crucial for functions such as enzyme binding, signal transduction, and structural stability.
This isotropy assumption is reflected in the uniform force constants across all interactions, irrespective of their nature or directional orientation, further abstracting the model from the nuanced, directionally dependent behaviors seen in real protein dynamics. Such simplifications, while useful for broad analyses, might not capture critical aspects of protein functionality that depend on highly directional, anisotropic interactions. Therefore, while the GNM offers significant insights into protein behavior under the assumption of isotropic interactions, it might miss critical dynamics relevant in physiological conditions where anisotropy plays a key role.

Now for solving this system we need the normal modes, which are the eigeinvectors of the kirchoff matrix.

$$Kv_k = \lambda_k v_k. \tag{3.8}$$

In addition this modes are very usefull because they allow us to interpretate the motion of the atoms in the protein.
In fact now we can solve our motion's equation:

$$F = -\nabla V(x) = -Kx = M\ddot{x}. \tag{3.9}$$

where M is the diagonal mass matrix.
Now we normalize the coordinates with a change of variable: $x' = \sqrt{M}x$.
So substituting this transformation we obtain:

$$\ddot{x} = -Kx. \tag{3.10}$$

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

$$-\omega^2 v_k = -Kv_k. \tag{3.11}$$

Mathematically normal modes are the eigenvectors v_k of the Hessian matrix and we use them to solve the system. This equation demonstrates that λ_k, the k-th eigenvalue of the Hessian, is related to the angular frequency by, as equation 3.8:

$$\omega_k = \sqrt{\lambda_k}. \tag{3.12}$$

The interpretation of that is that the eigenvectors v_k tell us the directions of motion in the space of atomic coordinates and every mode represent a distinct oscillation of every atoms of the protein.
In addition the low intensity values of eigenvalues correspond to low frequency modes are usually in literature associate with long range effect and high frequency oscillationa re associated with local effects. Typically the long range modes are assocaited with functions like the opening and closing of ligand-binding sites which are allosteric transitions, We will use them to solve the system as we see in a few lines.

Now from the hamiltonian and from the normal modes we can compute two important quantities that are the probability density and the mean displacement between the atoms.
From this simple hamiltonian formulation of the system we can derive some statistical properties. The probability density at equilibrium is given by:[11]

$$P(x) = \frac{1}{Z} e^{-\beta H(x)}, \tag{3.13}$$

where the partition function Z is:

$$Z = \int e^{-\beta H(x)}\, dx. \tag{3.14}$$

Define it we can use it to compute a lot of important quantities, like for exemple the mean displacement of the atoms.
The mean position $\langle x \rangle$ can be calculated as:[11]

$$\langle x \rangle = \int xP(x)\, dx, \tag{3.15}$$

where:

$$P(x) = \frac{1}{Z} e^{-\frac{\beta}{2} x^\top gKx}, \tag{3.16}$$

and the partition function is:[11]

$$Z = \int e^{-\frac{\beta}{2} x^\top gKx} \, dx. \tag{3.17}$$

Substituting $P(x)$ into the expression for $\langle x \rangle$, we obtain:

$$\langle x \rangle = \frac{1}{Z} \int x e^{-\frac{\beta}{2} x^\top gKx} \, dx. \tag{3.18}$$

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

$$\langle x \rangle = 0. \tag{3.19}$$

3.3 Stochastic Processes

We use a stochastic process to model the oscillation around the equilibrium position of each atom that constitutes the protein.

To model the oscillation, we use the following equation:

$$m\frac{d^2\mathbf{x}}{dt^2} = -\gamma\frac{d\mathbf{x}}{dt} - \nabla H(\mathbf{x}) + \sqrt{2\gamma k_B T}\,\boldsymbol{\eta}(t) = -\gamma\frac{d\mathbf{x}}{dt} - g\mathbf{K}\mathbf{x} + \sqrt{2\gamma k_B T}\,\boldsymbol{\eta}(t),$$

where (x(t) is a vector representing the displacement of residues from their equilibrium positions; $H(\mathbf{x})$ is a scalar function (the Hamiltonian); $\boldsymbol{\eta}(t)$ is a vector of Gaussian white noise with zero mean and unit variance; m and γ are scalars representing the mass and the damping coefficient, respectively; $\mathbf{K}$ is a matrix of coupling constants between residues (symmetric and positive definite) and g is a scalar that normalizes $\mathbf{K}$.
In the overdamped regime ($\gamma \gg m$), the acceleration term can be neglected, simplifying the equation to:

$$\gamma\frac{d\mathbf{x}}{dt} = -g\mathbf{K}\mathbf{x} + \sqrt{2\gamma k_B T}\,\boldsymbol{\eta}(t).$$

To further simplify the equation, we introduce a new time scale:

$$\tau = \frac{t}{\gamma}, \quad dt = \gamma d\tau,$$

and normalize the Hamiltonian by working in units of $H_0 = 2k_B T$:

$$\tilde{H}(\mathbf{x}) = \frac{H(\mathbf{x})}{H_0}.$$

The equation then becomes:

$$\frac{d\mathbf{x}}{d\tau} = -\nabla\tilde{H}(\mathbf{x}) + \boldsymbol{\eta}(\tau) = -g\mathbf{K}\mathbf{x}(t) + \boldsymbol{\eta}(t),.$$

This dynamics corresponds to a multidimensional Ornstein-Uhlenbeck process, the simplest Gaussian process with mean-reverting properties.

The matrix $\mathbf{K}$ can be diagonalized:

$$\mathbf{K} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\dagger,$$

where $\mathbf{V}$ is the orthogonal matrix of eigenvectors ($\mathbf{V}^\dagger = \mathbf{V}^{-1}$) of $\mathbf{K}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues of $\mathbf{K}$.
Changing variables to the eigenbasis:

$$\tilde{\mathbf{x}}(t) = \mathbf{V}^\dagger\mathbf{x}(t),$$

the equation becomes:

$$\frac{d\tilde{\mathbf{x}}(t)}{dt} = -g\boldsymbol{\Lambda}\tilde{\mathbf{x}}(t) + \tilde{\boldsymbol{\eta}}(t),$$

where $\tilde{\boldsymbol{\eta}}(t) = \mathbf{V}^\dagger \boldsymbol{\eta}(t)$. Since the transformation is orthogonal, $\tilde{\boldsymbol{\eta}}(t)$ remains Gaussian white noise with the same properties as $\boldsymbol{\eta}(t)$. The diagonalization decouples the syssstem into independent equations for each mode k:

$$\frac{d\tilde{x}_k(t)}{dt} = -g\lambda_k \tilde{x}_k(t) + \tilde{\eta}_k(t),$$

where $\tilde{x}_k(t)$ is the k-th component of $\tilde{\mathbf{x}}(t)$, λ_k is the k-th eigenvalue of $\mathbf{K}$, and $\tilde{\eta}_k(t)$ is the k-th component of the transformed noise. The solution to the Ornstein-Uhlenbeck equation for each mode is:

$$\tilde{x}_k(t) = \tilde{x}_k(0)e^{-g\lambda_k t} + \int_0^t e^{-g\lambda_k(t-s)}\tilde{\eta}_k(s)\, ds.$$

This solution reveals the direction in the eigenvector space of residue fluctuation. The corresponding eigenvalue is proportional to the intensity of its oscillation, which is useful for understanding the role of each residue in the context of collective behavior. We will use this solution heavily in the following sections to analyze the allosteric propagation of signals in the protein.

Chapter 4

Introduction to causality and causality in allosteric mechanisms of proteins

Our primary aim, as we just said, is to identify the allosteric propagation of the signal from allosteric sites to other sites of the protein.
Generally causality refers to the relationship between causes and effects, where one event (the cause) directly influences or produces another event (the effect), so it implies a directional influence.
To understand the propagation of the signal we need to have a cause (the sorgent of the signal, in our case the allosteric site) and an effect (the dynamical changing in the protein).
To reveal causality we saw, for now, only one useful indicator: the covariance.
But we know that Covariance does not imply causation, so in the following pages we will discover new causal indicators.
While covariance is a starting point for identifying potential causal relationships, it primarily measures the strength of linear associations without providing directionality or confirming causation.
This limitation necessitates the introduction of more sophisticated indicators that can discern direct from indirect influences and clarify the pathways through which allosteric signals propagate.
To overcome the limitations of covariance, we will introduce additional causal indicators such as response functions and transfer entropy.
These indicators provide a more nuanced understanding of the causal relationships within proteins, allowing us to identify the specific residues and regions that play a crucial role in allosteric mechanisms.
Infact they are able to measure the casual relation in a different way from the covariance, in fact the response allows us to understand the direct relationship between residues and trasnfer entropy to catch the non linear interactions and so the directioanal flux of information.
These indicators not only measure the strength of associations but also help determine the direction and specificity of interactions, crucial for mapping the complex networks within proteins.

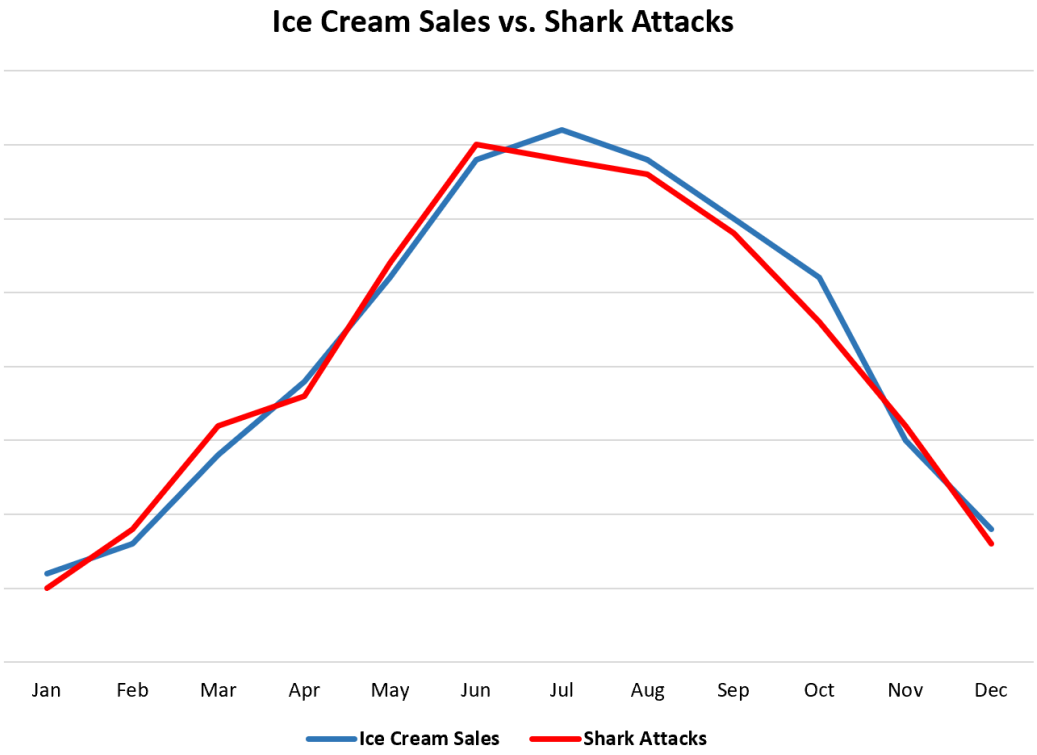


Figure 4.1. Covariance is not causation. The Covariance between two variables does not imply a causal relationship. This is an exemple of a spurious Covariance. Here the latent variable is the temperature.

Mathematically causal relationships are often modeled as:[13]

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

where f describes how A and other factors jointly determine B.
It is hard to understand causality in the real world, particularly in biological systems where the interactions between components are numerous and there are many factors that can influence the system.
Thus, testing if there is a causal relationship between two variables is a very challenging task, because my function depends on many other variables, making it difficult to discern whether changes in the output are related to changes in my specific variable or due to other changes.
Causality is very important because it allows us to understand which variables and dimensions are crucial for my system to infer a specific output. In a data-driven system for inferring causality, if we add many dimensions hoping to find patterns or covariance to predict the output, we encounter the problem of the curse of dimensionality, and thus our model will likely fail in its forecasting.
Therefore, understanding causality is very important for the inference of the system because it not only helps to better understand the phenomenon but also to build more effective inference models.
Today, there is no unique method to infer causality; instead, there are many methods that can be used for inference.

Each of these methods has its own advantages and disadvantages, but we cannot be certain that we have established true causality.
In our case, we are also dealing with a high-dimensional system, where moving a single residue results in collective motion along the protein, with various sections of the protein responding in very different ways.

Causality (and one of causal indicator which we will use) is connect to the concept of entropy and Shannon entropy.
Typically if we are forecasting a output X variable we can define a Shannon entropy associated with that output, defined as:

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

where X is a discrete random variable with possible states x_i, and $P(x_i)$ is the probability of x_i.
When $P(x_i) = \delta(x_i)$ -> $H(X) = 0$, 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.
So the distribution width of the x_i determines how much my system is predictable.
In general more the entropy is low, more the system is predictable, so more we know causal relation of the system.
The discussion on causality will be expanded to connect theoretical concepts with biological implications more directly.
This will include illustrations of how changes at an allosteric site can causally influence the active site's functionality, using specific biochemical pathways as examples.
This approach will help bridge the gap between abstract statistical measures and their tangible effects on protein behavior.
Now we will see some indicators to try to catch the allosteric causality mechanism in the system.
To further ground our discussion in practical terms, we will explore how allosteric mechanisms, often triggered by external stimuli or internal chemical changes, lead to observable changes in protein activity. This exploration will serve as a foundation to understand how subtle molecular changes can result in significant biological outcomes, emphasizing the importance of pinpointing the origin of these signals.

4.1 Covariance

The Covariance between two variables, such as the fluctuations in the positions of residues i and j, provides insight into how these components of the system interact. The covariance of the displacement $x(t)$ is defined as:[12]

$$C(t) = \langle x(t)x^\top(0)\rangle,$$

where $x(t) = V\tilde{x}(t)$. Substituting this into the definition:

$$C(t) = V\langle \tilde{x}(t)\tilde{x}^\top(0)\rangle V^\top.$$

Substituting the solution of $\tilde{x}$ into the covariance:

$$\langle \tilde{x}(t)\tilde{x}^\top(0)\rangle = \langle \tilde{x}(0)e^{-g\Lambda t}\tilde{x}^\top(0)\rangle + \left\langle \left(\int_0^t e^{-g\Lambda(t-s)}\tilde{\eta}(s)\,ds\right)\tilde{x}^\top(0)\right\rangle.$$

For the first term:

$$\langle \tilde{x}(0)e^{-g\Lambda t}\tilde{x}^\top(0)\rangle = e^{-g\Lambda t}\langle \tilde{x}(0)\tilde{x}^\top(0)\rangle.$$

In the stationary regime:

$$\langle \tilde{x}_k(0)\tilde{x}_l(0)\rangle = \frac{\delta_{kl}}{g\lambda_k}.$$

Thus:

$$\langle \tilde{x}(0)\tilde{x}^\top(0)\rangle = (g\Lambda)^{-1}.$$

Substitute this:

$$\langle \tilde{x}(0)e^{-g\Lambda t}\tilde{x}^\top(0)\rangle = e^{-g\Lambda t}\cdot(g\Lambda)^{-1}.$$

For the second term:

$$\left\langle \left(\int_0^t e^{-g\Lambda(t-s)}\tilde{\eta}(s)\,ds\right)\tilde{x}^\top(0)\right\rangle = 0,$$

because $\tilde{\eta}(s)$ is uncorrelated with $\tilde{x}(0)$. Thus, the stochastic contribution vanishes. The covariance in the eigenbasis is:

$$\langle \tilde{x}(t)\tilde{x}^\top(0)\rangle = e^{-g\Lambda t}\cdot(g\Lambda)^{-1}.$$

Transform back to the original basis:

$$C(t) = V\langle \tilde{x}(t)\tilde{x}^\top(0)\rangle V^\top.$$

Substitute the eigenbasis covariance:

$$C(t) = V\left(e^{-g\Lambda t}\cdot(g\Lambda)^{-1}\right)V^\top.$$

Or in components:

$$C_{ij}(t) = \sum_{k=1}^{N} v_{ik}v_{jk}\frac{e^{-g\lambda_k t}}{g\lambda_k},$$

Positive Covariances between residues i and j at time t indicate that the fluctuations in their positions are synchronized, while negative Covariances suggest that the fluctuations are anti-correlated.
Both positive and negative Covariances can reveal how residues interact and influence each other's dynamics.
As we just said it reveals a possible causal relationship between the residues, in addition we need a storng Covariance to find causality, but it is not sufficient condition to have causality.

4.2 Response Function

Another indicator of causality is the linear response function, which quantifies the influence of a perturbation on one residue on the behavior of another residue.
The response function $R_{ij}(t)$ is defined as:[12]

$$R_{ij}(t) = C_{ij}(t)C_{ij}^{-1}(0), \tag{4.1}$$

Where $C_{ij}(t)$ is the covariance between residues i and j at time t.
So it is completely determined by the covariance.
Substituting it we obtain:

$$R_{ij}(t) = \frac{1}{N} + \sum_{k=1}^{N} v_{ik}v_{kj}e^{-g\lambda_k t}.$$

The linear response describes how the average position of a residue i, denoted as $\langle x_i \rangle$, responds to a perturbation applied to another residue j through an external force f_j. The power of this indicator is that it tends to ignore the spurious Covariance beacuse it ignores the non direct interaction between the residues.
To let see why there is a perturbation and why it is connected with the covariance we need to see the fluctuation-dissipation theorem.
Infact if we consider the Hamiltonian of the system under a small external perturbation:

$$\mathcal{H} = \mathcal{H}_0 - \sum_j f_j x_j, \tag{4.2}$$

where $\mathcal{H}_0$ is the unperturbed Hamiltonian and $-f_j x_j$ is the interaction term between the external force and the displacements of residues j; we obtain that average position of dispalcement of residue i is:

$$\langle x_i \rangle = \frac{\int x_i e^{-\beta \mathcal{H}}\, dx}{\int e^{-\beta \mathcal{H}}\, dx}. \tag{4.3}$$

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

$$e^{-\beta \mathcal{H}} \approx e^{-\beta \mathcal{H}_0} \left(1 + \beta f_j x_j\right), \tag{4.4}$$

and substituting this expansion into the expression for $\langle x_i \rangle$:

$$\langle x_i \rangle \approx \langle x_i \rangle_0 + \beta f_j \langle \delta x_i \cdot \delta x_j \rangle, \tag{4.5}$$

Taking the derivative of $\langle x_i \rangle$ with respect to f_j gives the linear response:

$$R_{ij} = \frac{\partial \langle x_i \rangle}{\partial f_j} = \beta \langle \delta x_i \cdot \delta x_j \rangle. \tag{4.6}$$

This result shows that the response R_{ij} is directly proportional to the equilibrium covariance between the fluctuations of the displacement of residue i and j.

4.3 Transfer Entropy

The last causal indicator is the Transfer Entropy (TE), which quantifies the directional flow of information between two variables.
It add to the previous indicator the non linearity and directional information, so it is very useful to understand the causal relationship between the residues.
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]

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

where $H[a \mid b]$ is the conditional Shannon entropy of variable a given b, $x_i(t+\tau)$: State of x_i at time $t+\tau$, $x_i(\tau), x_j(\tau)$ are the state of x_i and x_j at time τ.

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]

$$TE_{j\to i}(t) = -\frac{1}{2}\ln\left(1 - \frac{\alpha_{ij}(t)}{\beta_{ij}(t)}\right) = TE_{i,j}, \tag{4.7}$$

where:

$$\alpha_{ij}(t) = \left[C_{ii}(0)C_{ij}(t) - C_{ij}(0)C_{ii}(t)\right]^2, \tag{4.8}$$

$$\beta_{ij}(t) = \left[C_{ii}(0)C_{jj}(0) - C_{ij}^2(0)\right]\left[C_{ii}^2(0) - C_{ii}^2(t)\right]. \tag{4.9}$$

So also it is completely determined by the covariance.
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 $TE_{j\to i}(t)$.[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:

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

where n is the dimensionality of 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:

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

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:

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

Transfer Entropy is defined as:

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

Substituting the expressions for the conditional entropies:

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

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

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

where:

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

Thus this formula captures the influence of x_j on x_i while accounting for their shared history, this conditioning eliminates spurious Covariances arising from common drivers or indirect interactions, providing a rigorous and directional measure of

causality. In the following trials to catch the allosteric causality mechanism in the protein we will use this indicator using the follwing procedure:
First we will calculcate the covariance to see if we have the necessary condition to have causality, then we will calculate the response function to see if we have a causal relationship between the residues and finally we will calculate the transfer entropy to see if we have a directional causal relationship between the residues.

Chapter 5

Results

This chapter presents detailed results from the computational analysis of the 3LNX protein, sourced from the Protein Data Bank.
Our study focuses on the dynamics of allosteric signal propagation within the protein structure, examining how these signals influence biological functions through conformational changes.
Specifically, we explore the impact of the connection radius on constructing the Kirchhoff matrix and its subsequent effect on the protein's covariance structures. These insights are pivotal for identifying allosteric sites and understanding the protein's dynamic mechanical properties.

5.1 Dataset 3LNX, Connection Radius, Kirchhoff Matrix and Covariance

The study utilizes the 3LNX dataset, focusing exclusively on alpha-carbon atoms to enhance the accuracy of our simulations.
We investigate how perturbations at allosteric sites propagate signals to the active sites, leading to significant conformational changes, notably the "opening" and "closing" of the protein's structural "hands" to interact with ligands.

	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 5.1. Dataset of 3LNX.

These dynamics are essential for understanding how allosteric sites are interconnected with the binding pockets involved in ligand binding. By perturbing these sites, particularly those in alpha-α regions, we observe targeted responses around the ligand site.
Choosing an appropriate connection radius is crucial for accurately capturing both local and long-range interactions between residues.

A too short radius may overlook long-range interactions, while a too long radius could obscure essential local interactions.
Based on literature and our findings, a connection radius of 8.0 nm was optimal, balancing the need to accurately reflect the protein's mechanical and dynamic properties.

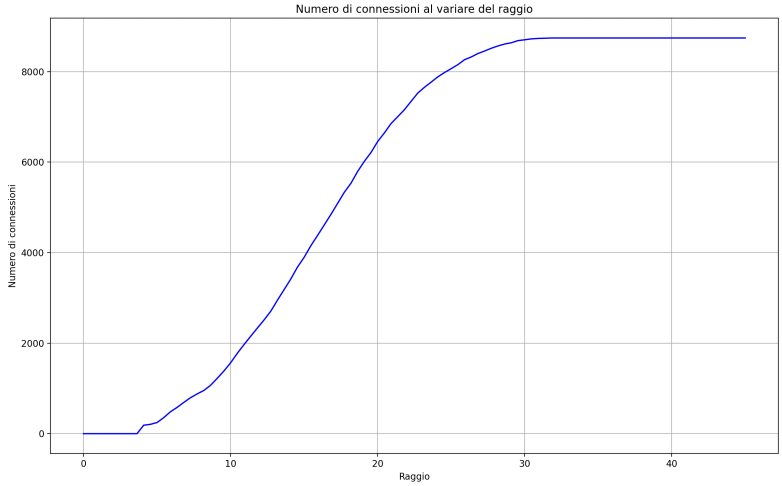


Figure 5.2. Number of links in function of the connection radius.

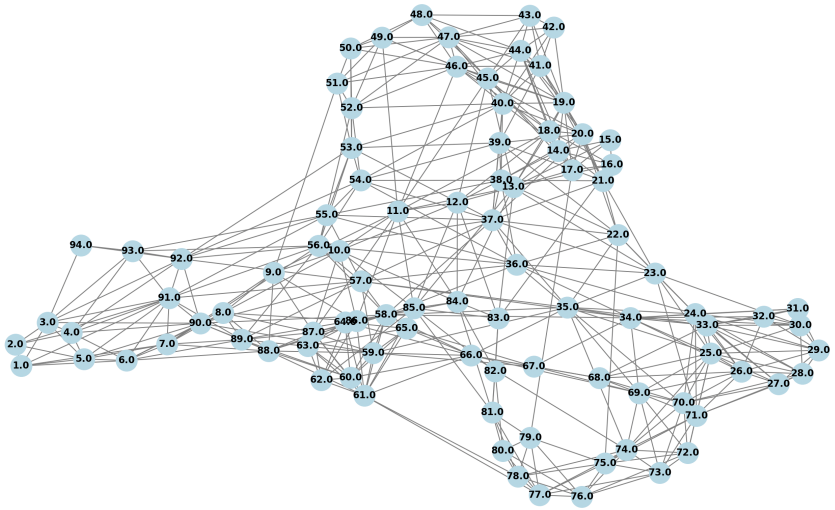


Figure 5.3. Bidirectional Graph obtained with radius of 8.0 nm.

Using this radius, we constructed a bidirectional graph, as shown, which captures the essential connectivity and dynamics within the protein structure.

By applying the Kirchhoff matrix formulation to the protein structure, we derive matrix K, which models the protein as a network of interactions. This matrix simplifies the complex web of protein interactions into pairwise connections, where

the link weights represent the coupling strength between residues. Such simplification helps to visualize and analyze the dynamic properties of the protein but omits more complex interactions like solvation effects or multi-body interactions, which could also influence the protein's behavior.

The weights of the links are set based on the proximity and type of interaction between residues: typically, near neighbors connected by strong covalent bonds or dense non-covalent interactions (such as van der Waals forces and hydrogen bonds) are assigned a weight of 20. These interactions are considerably stronger, approximately 20 times greater, than those between more distantly connected residues, which receive a weight of 1. This weighting scheme reflects the biological reality where closer residues have stronger interactions that significantly influence protein dynamics.

Formally, the protein's structure can be represented as a graph $G = (V, E)$, where V is the set of vertices (residues), and E is the set of edges (connections between residues). The weight $w(i, j)$ of the link between any two vertices i and j is defined as follows:

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

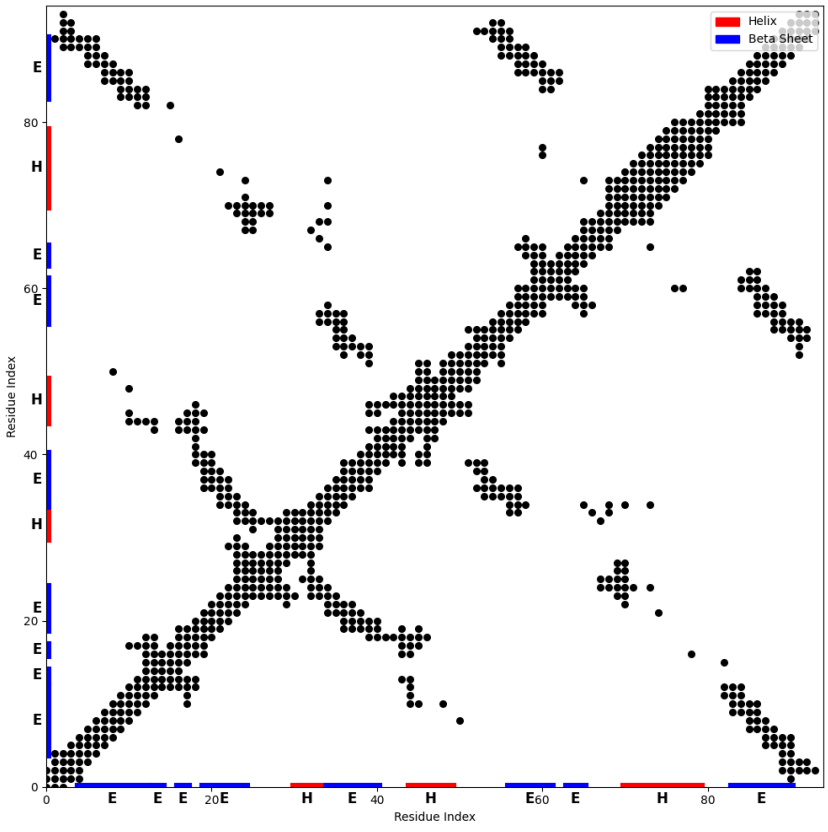


Figure 5.4. Kirchhoff Matrix of the protein structure.

This representation of the Kirchhoff matrix (5.4) visually encapsulates the

connectivity and interaction patterns within the protein, highlighting how structural dynamics are influenced by both local and long-range interactions.

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 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.

Now we have to understand deeply how these interacting residues influence each other's dynamics.
For studying it we have to calculate the Covariance matrix.
at time 0:
The Covariance matrix at time 0 is defined as:

$$C_{ij}(0) = \sum_{k=1}^{N} \frac{1}{\lambda_k} v_{ik} v_{jk}.$$

The two figures illustrate the covariance matrices obtained from the protein structure, one only plotting the positive covariance and the other only plotting the negative covariance.
In the following figures we plot also the secondary structure (Helices: red; Beta Sheets: blue) and the kirchoff matrix.

In these images we notice that we have positive Covariance where we have residues and negative Covariance where we don't have residues. More distant is a region from the residues more negative will be the Covariance.
We can see also a cluster of the most postive corerlation along the diagonal.
So the Kirchhoff matrix directly influences the Covariance matrices by encoding the connectivity between residues.

Now that we have the Covariance matrix at time 0, we can calculate the Beta Factors, wich are a critical measure of the protein's dynamic behavior.
They represent the atomic displacement and are useful for evaluating the protein's flexibility and stability and it also a tool for evaluating the theoretical model because we have experimental data of them.
Mathamically they are defined as:
$B_i = 8\pi^2 \langle C_{i,i} \rangle$

The plot above compares the experimental B-factors (blue curve), as we said in (4.9) $B_i = 8\pi^2 C_{ii}$, with the predicted B-factors (red curve) along the residue index. The B-factors represents the atomic displacement so they are a critical measure of the dynamic behavior of the protein structure.
They are also useful for testing the quality of the model.

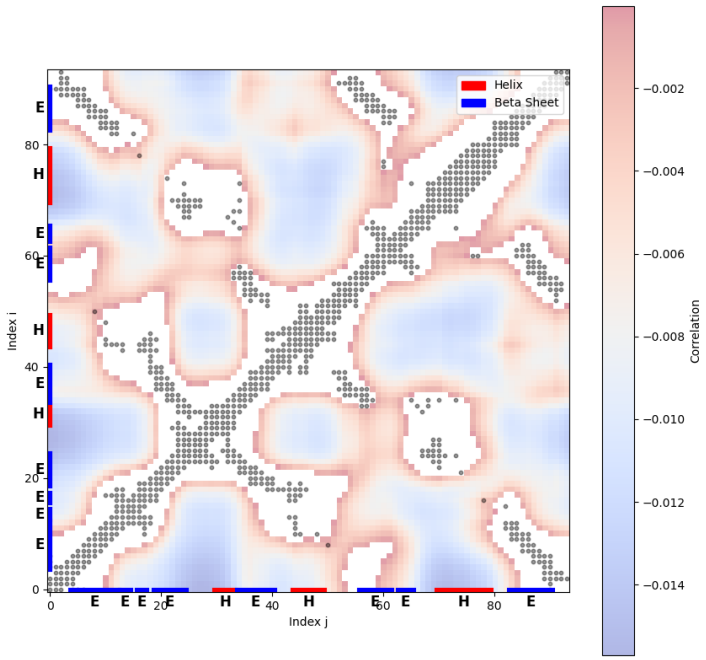


Figure 5.5. Negative covariance between residues at time 0.

We see that 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; these deviations may be due to limitations in the Kirchhoff matrix approximation or oversimplified assumptions in the modeling process.
To evaluate the model performance we have to define the following metrics RMSE e MAE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

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

Where y_i are observed values, $\hat{y}_i$ are predicted values. The main difference from RMSE and MAE is that RMSE gives more weight to large errors, while the MAE gives equal weight to all errors.
Now finaly with these quantitative metrics we can evaluate the model performance:
This table shows the RMSE and MAE values for the model and show the model is good to predict the beta factors behavior.
It is clear that we can improve the model and it is possible to obtain more accurate predictions.

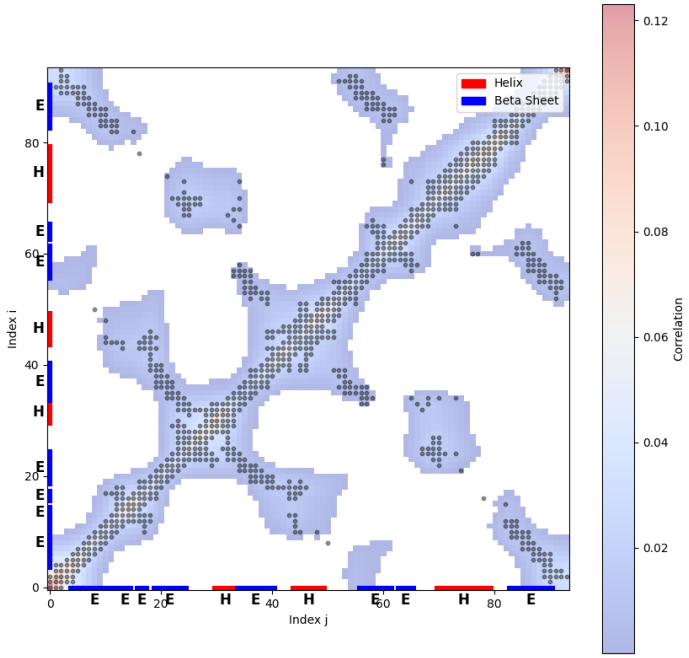


Figure 5.6. Positive covariance between residues at time 0.

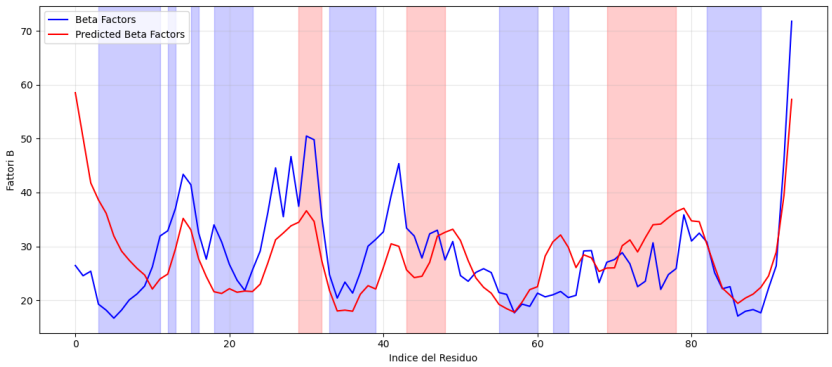


Figure 5.7. Beta factors.

Metric	Value
RMSE	8.5200
MAE	6.4146

Table 5.1. Model Performance Metrics

Now let's valuate if the model can catch the allosteric dynamics of the protein.

5.2 Causal indicators in time

In this section we will focus on the causal indicators in time and between residues. Understanding the dynamic interactions within the protein structure requires analyzing how causal indicators like Covariance and Response behave over time, supplemented by Transfer Entropy to gauge information flow.
In this way we have deeply view of allosteric mechanism of the protein.
Let's start with the Covariance in time.
The Covariance between residues at time t is defined as:

$$C_{ij}(t) = \sum_{k=1}^{N} v_{ik} v_{jk} \frac{e^{-g\lambda_k t}}{g\lambda_k},$$

This behavior is coherent with what expected, the Covariance decay over time,

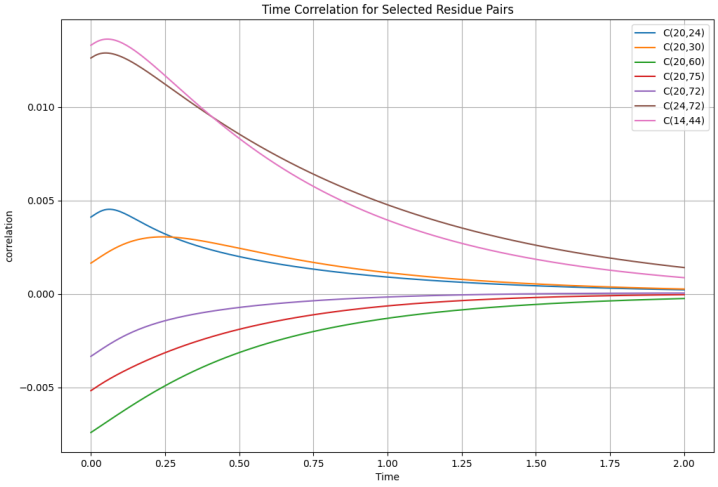


Figure 5.8. Covariances in time.

indicating a loss of direct dynamic influence as time progresses.
Moreover Covariance values vary significantly between residue pairs, reflecting differences in their initial dynamic coupling.
Finally residue pairs closer in space tend to show higher initial Covariances compared to more distant pairs.
The decay in Covariance is largely attributable to the intrinsic thermal motions within the protein structure.
Each residue in a protein is subject to thermal vibrations and random collisions with surrounding molecules, which contribute to the diffusion of any initial localized perturbations across the protein structure.
Over time, these random motions cause the initial correlations in residue movements to dissipate, leading to a decrease in Covariance values.
From a biological perspective, the decay of Covariance over time can be indicative of the protein's ability to return to a stable state after undergoing a conformational change.

This is crucial for protein functionality, as proteins often need to respond dynamically to cellular signals or environmental changes and then return to their baseline state to be ready for subsequent interactions.
The rate and pattern of Covariance decay can therefore provide insights into the resilience and stability of protein structures, as well as their capability to undergo conformational changes necessary for biological activity.
Moreover, understanding the decay patterns of Covariance can help in identifying regions within the protein that are particularly flexible or rigid. Regions with slower decay rates may indicate domains of the protein that maintain their conformational relationships over longer periods, which could be critical for maintaining the structural integrity necessary for specific biological functions. Conversely, areas with rapid Covariance decay might be more flexible, allowing the protein to adapt its shape for different functional interactions.

Now we can analyze the response in time.
The response metric is crucial for understanding how protein residues react over time to specific perturbations.
Unlike covariance, which can fluctuate between positive and negative values reflecting the strength and direction of coupling, the response is inherently positive, emphasizing the accumulation of effects from the initial perturbation:

$$R_{ij}(t) = \frac{1}{N} + \sum_{k=1}^{N} v_{ik} v_{kj} e^{-g\lambda_k t}.$$

This mathematical representation highlights how the response function quantifies the influence of dynamic changes across the protein structure as they decay over time. This behavior is coherent with what expected, the response decays over time,

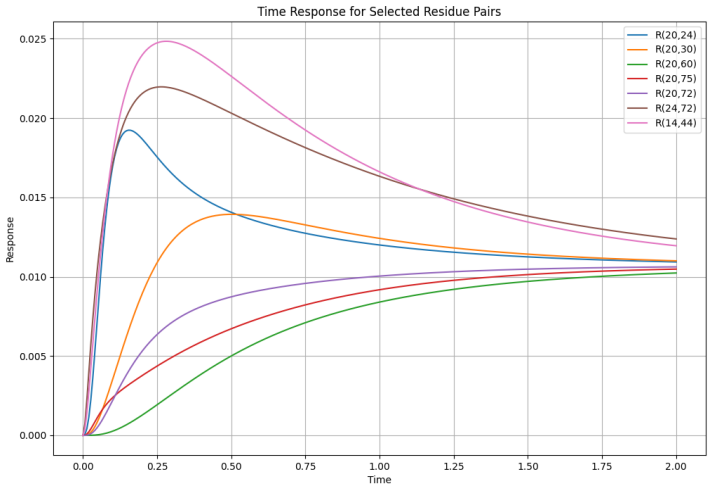


Figure 5.9. Responses in time, illustrating a sharp initial rise and gradual decay, reflecting dynamic adaptations within the protein structure.

indicating a loss of direct dynamic influence as time progresses.
Also, we notice that when covariance can be positive or negative, the response is always positive, starting from 0 and reaching the value of 1/N.
The responses exhibit a sharp rise initially, which is then followed by a gradual decay.
Residue pairs closer to the source of perturbation (spatially proximal) show a more pronounced and faster response.
This proximity effect highlights the importance of spatial relationships in protein dynamics, where residues near the source of perturbation are more immediately and strongly affected.
A peak is observed at an intermediate time point for most residue pairs, indicating a time of maximum information transfer.
This peak represents the critical moment when the effects of the perturbation are most acutely felt throughout the protein, suggesting key points of allosteric control

or regulatory interaction within the protein structure (We will soon explore this in detail).
Understanding these peaks can provide insights into the timing and extent of protein responses that are crucial for biological functions such as enzyme activation or signaling pathways.

Now we can analyze the transfer entropy in time:

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

where:

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

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

This formulation quantifies the directional non linear information flow from residue j to i over time, reflecting how changes in covariance impact the dynamics of information transfer within the protein.

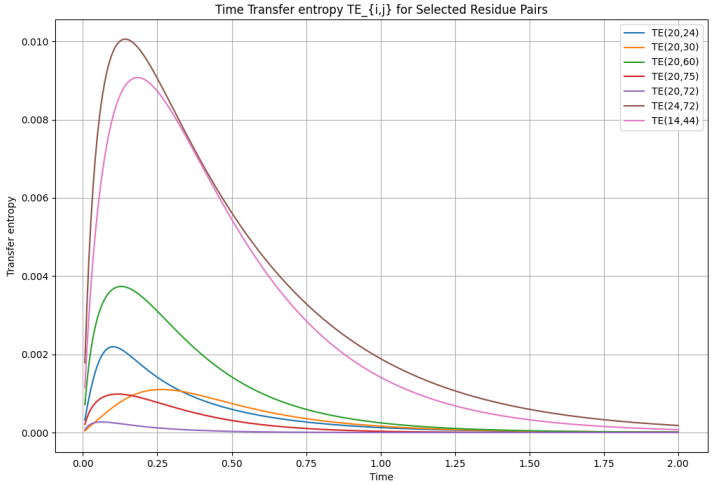


Figure 5.10. Transfer entropy over time, highlighting critical moments when the information flow peaks, indicating significant regulatory interactions within the protein structure.

In this analysis, we observe that a peak is seen at an intermediate time point for most residue pairs, signaling a time of maximum information transfer.
This peak is crucial as it may indicate a pivotal moment when one residue exerts a strong influence over another, potentially triggering a significant functional response or regulatory mechanism.
Moreover, the magnitude of transfer entropy varies, reflecting differences in the strength of directional coupling between pairs. Residue pairs with high initial connectivity, which are typically spatially closer within the protein structure, exhibit stronger peaks compared to those more distant.
This pattern underscores how structural proximity within a protein can significantly enhance the efficiency and impact of informational transfer, aligning with known biological principles that spatially close residues often participate in more direct and influential interactions.

Now that we understood the casual indicators in time we can finaly analyze the causal indicators between residues to identify the allosetric mechanism.
This is the main section of our work in which we will try to understand the allosteric mechanism of the protein and understand for real if the model explain in a good way the protein dynamics.
As we said before if the graph of the proteins were random we will expected random interaction.
For seeing if our model is in according with the real world results if we perturb the allosteric sites, situated in according with paper [14] in all the alpha-α and alpha-γ helices, we will expected that the signal will propagate along the protein structure for reaching the active sites, around beta-β and Alpha-α.
To analyze our system, it is essential to focus on timescales around the characteristic time τ.
In fact if the time t, in which we analyze the system, is too short so we don't see in a right way the propagation, in other hand if the time is too long we don't see the relaxation of the system.
So to determine the characteristic time τ of the system, we have to analyze the autocorrelation normalized function $C_{i,i}(t)/C_{i,i}(0)$ between all residues for all relevant times t.
The idea is that the mean of typical decay time is the characteristic time τ of my system.
Mathematically for each index i will take the autocorrelation normalized $C_{i,i}(t)/C_{i,i}(0)$ when they are equal to e^{-1} of their initial values. These times are the t_i for every residues.
Now i have an histogram of times and taking the mean of this vlaues i will obtain my characteristic time τ.
We estimated a time τ of 0.1842 ns.

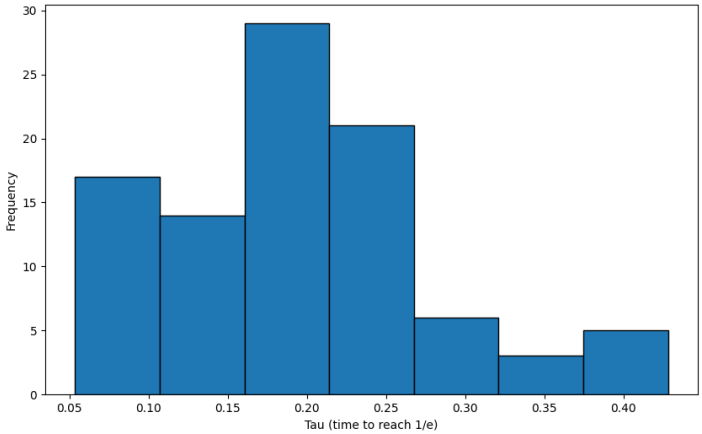


Figure 5.11. Hisotgram of t_i.

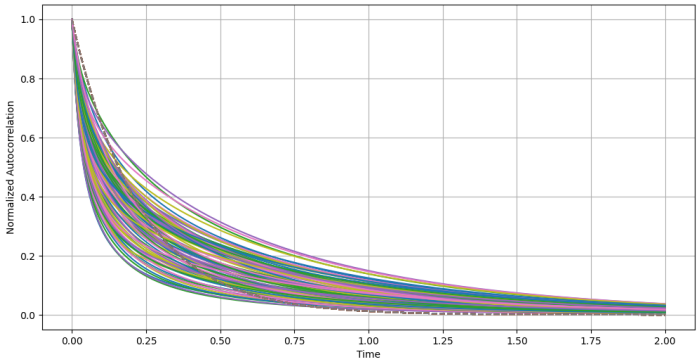


Figure 5.12. Normalized autocorrelations of all residues with also a sketched plot of the autocorrlation with the fitted tau.

5.3 Causal indicators between residues

Now we can analyze all the causal indicators between the allosteric sites and all other residues at the three times at [τ-0.5*τ, τ, τ+0.5*τ].
Our goal is to catch a link between the allosteric sites in the secondary structure alpha-α and alpha-γ and the binding pocket between beta-β and alpha-α.
To do it we have to compute the covariance, the response and the trasnfer entropy between residues.

This is the plot of the Covariance between the specific allosteric sites at time τ and all the residues:

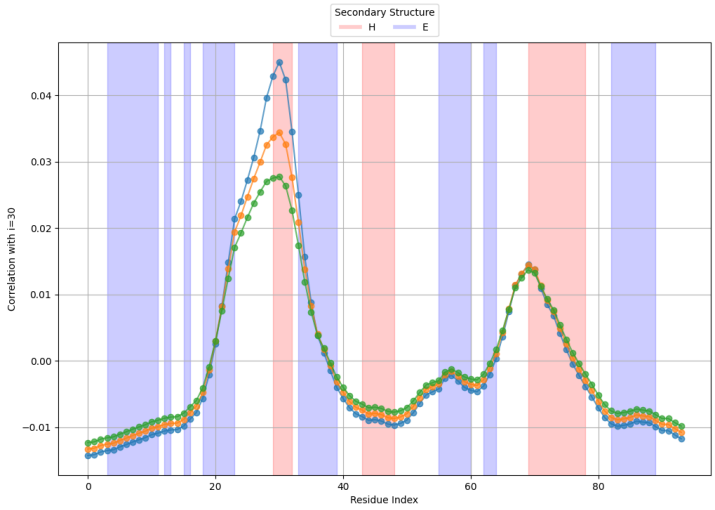


Figure 5.13. Covariance of 31-th residue at τ, showing localized interaction peaks indicative of dynamic communication pathways.

In the Covariance of residue 31 5.13 there are a high Covariance observed near the binding pocket,as we expected, particularly between the alpha-α and beta-β regions, and another allosteric site in the alpha-γ helix, highlights regions of significant allosteric communication.
Moreover we see also strong values of covariance in the head and in the tail of the protein, probability caused by spourious interactions. We will use following the others causal indicators to understand it better.

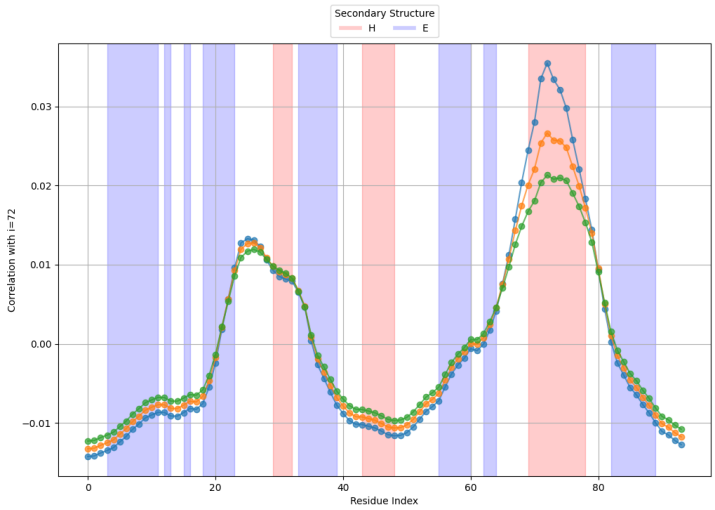


Figure 5.14. Covariance of 73-rd residue at τ, emphasizing the inter-domain communication crucial for protein functionality.

Now we analyze the Covariance of residue 73 5.14.
We notice a distinct peak around the 73-rd residue, due to the locatility, and other strong signals especially in regions like the beta-β-strand and alpha-α-helix near the binding pocket.
Also here we see strong values of covariance in the head and in the tail of the protein.

For both the analysis we see that the Covariance is high in the binding pocket, between the beta-β and alpha-α regions, and in the allosteric sites in the alpha-α and alpha-γ helices.
This is coherent with what we expected, the signal propagate from the allosteric sites to the binding pocket.

Now to do a deeper analysis we can calculate the response of the protein to the perturbation of the allosteric sites.
When perturbing the 31-th residue, a significant response is observed across various regions of the protein. This is visualized in the response data presented below:

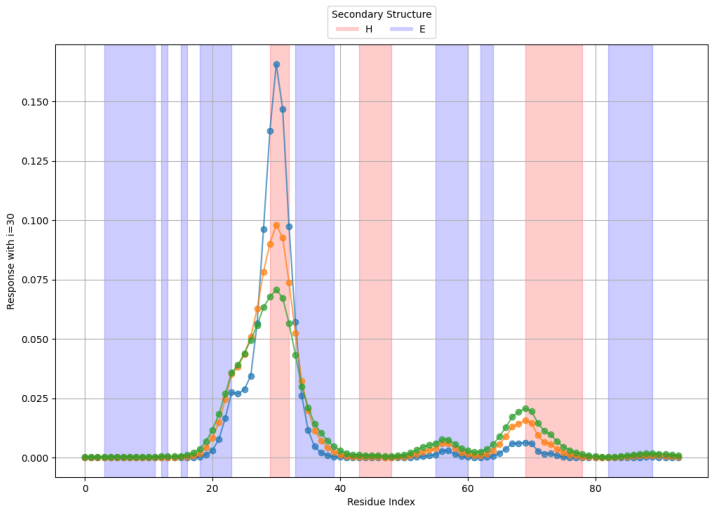
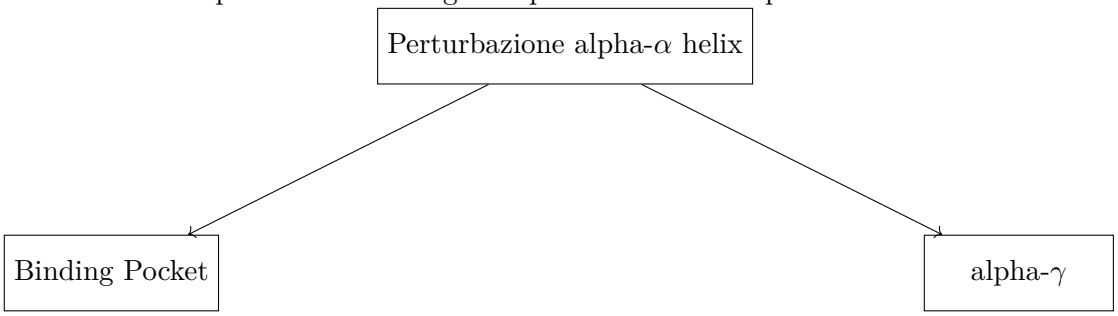


Figure 5.15. Dynamic response observed from perturbating the 31-th residue, illustrating signal propagation through alpha-*alpha* helical regions.

From the response 5.15 it is clear a response in the beta-β sheet, in the alpha-α helix and at the start of alpha-γ helix; indicating a strong propagation of the signal from the allosteric site directly to the binding pocket.
Moreover this result is particularly informative as it helps refute previous hypotheses suggesting significant signaling activities in the head and tail regions of the protein, where high covariance was initially observed.
In fact while those regions exhibit covariance, they do not necessarily contribute to causal interactions related to the protein's active functions.
But we have to note that while we observe these specific responses they do not alone confirm causality, however, the strong correlation between site-specific perturbation and localized response does support the hypothesis of these regions being critically involved in the allosteric mechanism.
With these insights, we can now present a new representation of the causal interactions within the protein eliminating the spurious relationship.



Now we analyze the response of the protein to the perturbation of the 73-rd residue: Perturbing the 73-rd residue we note a response in the binding pocket and other significant interactions in the alpha-β and moderate responses in the beta-γ

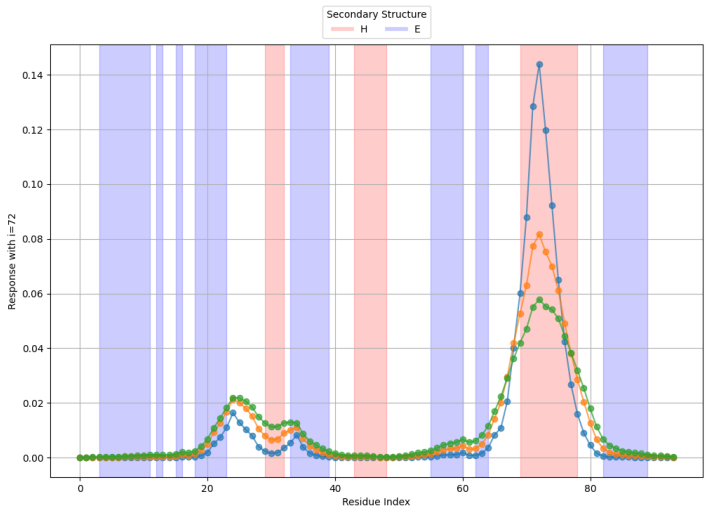
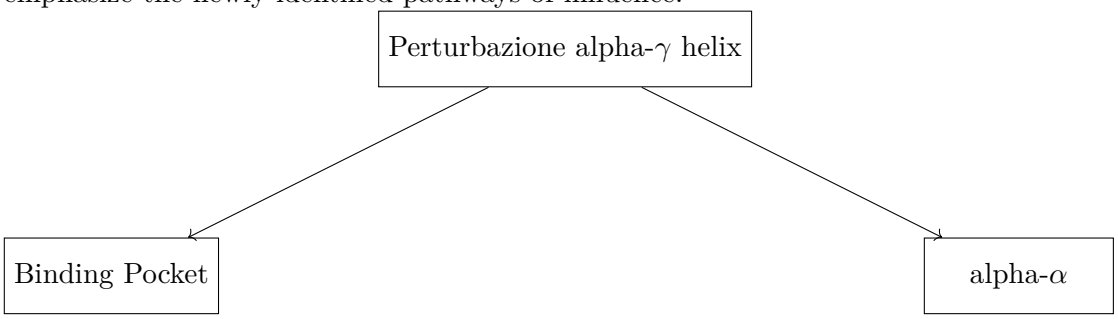


Figure 5.16. Dynamic response observed from perturbing the 73-rd residue, highlighting the interaction with critical functional regions of the protein.

regions.
That suggest the existence of the allosteric mechanism which we expect.
The alpha-β and beta-γ regions, showing varied levels of response, are likely involved in stabilizing the protein upon ligand binding or release.
To better represent these findings, we update our causal interaction diagram to emphasize the newly identified pathways of influence.



To have a deeper point of view of causality mechanism and to undestand the directional flux of information within the protein we have now to study the transfer entropy. We begin by examining the active transfer entropy from the allosteric sites to the other residues, represented as $T_{\text{allosteric site},j}$. Analyzing the images 5.17

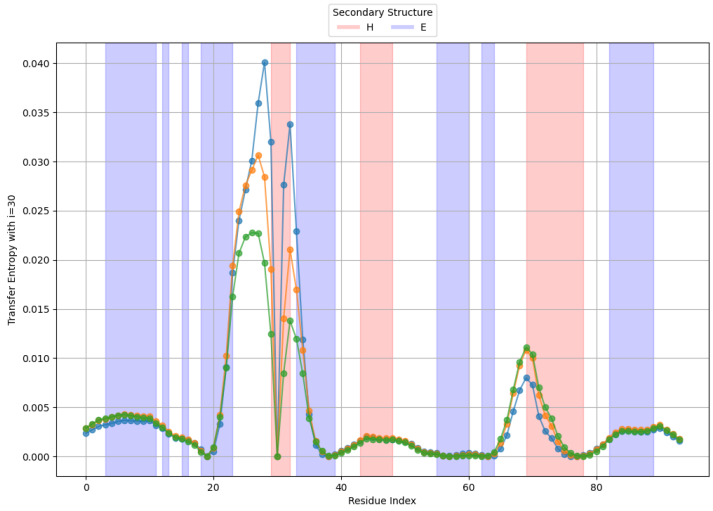
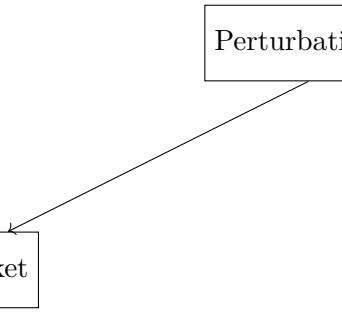


Figure 5.17. Transfer Entropy j->31 ($T_{31,j}$).

we observe that signals are notably propagated within the binding pocket and the alpha-γ helix.
The pattern revealed is quite similar to the response data.
To visualize these findings, we maintain our current graph of causal relations, which



clearly shows the directional influence between identified regions:

Similar findings are noted for the 73rd residue: This analysis confirms a consistent pattern of signal propagation from the alpha-γ helix to the binding pocket and back to the alpha-α region, solidifying our understanding of the interaction dynamics within these critical areas of the protein.

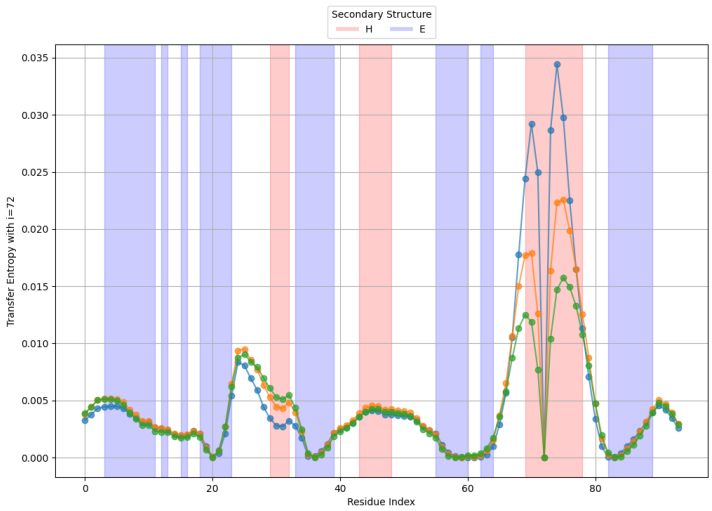


Figure 5.18. Transfer Entropy j->73 ($T_{73,j}$).

Our final step of the analysis is to study the passive transfer entropy, denoted as $T_{j,\text{allosteric site}}$, to examin the influence that allosteric sites recive to the other residues.

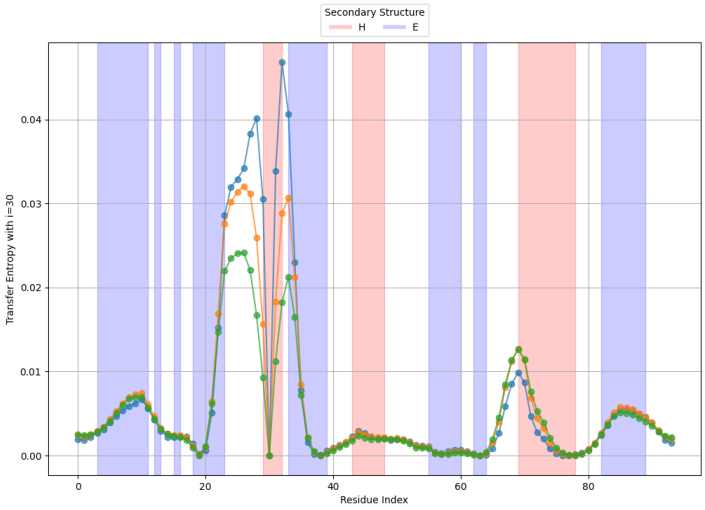


Figure 5.19. Transfer Entropy 31->i ($T_{i,31}$).

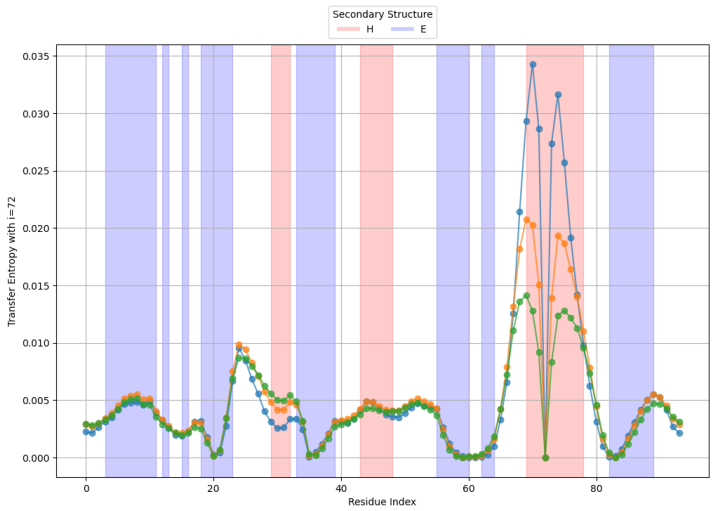


Figure 5.20. Transfer Entropy 73->i ($T_{i,73}$).

In our analysis, we observed that we don't have a great difference in patterns between the active and passive transfer entropy.
Thus it suggest a bi-directional mechanism of action, so allosteric sites both influence and are influenced by the binding pocket within the protein.
Thus this dual role highlights the sophisticated nature of allosteric modulation, where sites are not merely passive receivers of signals but also react and comuniate with allsoteric sites.

5.4 Conclusions of equlibrium stochastic process

The findings presented in this chapter contribute significantly to our understanding of allosteric dynamics in proteins through the application of Gaussian network models and normal mode analysis.
The covariance matrices and beta factors revealed interaction patterns and mobility within protein structures that are crucial for biological function.
Notably, the identification of specific peaks in the beta factors corresponding to the binding pocket and allosteric sites suggests these areas could be critical for regulating protein activities.
Moreover, the comparison between experimental data and model predictions showed good agreement, indicating that the utilized models can effectively capture essential aspects of protein dynamics.
The behavior of the covariance matrices and response analysis highlighted how dynamical changes in one part of the protein can significantly influence other distance regions, suggesting allosteric behavior.
In addition the causal indicators between residues provided insights into the directional flow of information within the protein, revealing critical pathways of allosteric communication according to expected results.
These results align with the theoretical models discussed in previous chapters and provide empirical validation of our research hypotheses.
Thus these observations shed light on the functional interplay between the main secondary structural regions: the binding pocket and the alpha helixs.
The analysis demonstrates that the geometric strucutrue of the protein is so important that is sufficent to explain mathematically the allosteric propagation of the protein.
So this method can be used to predict the allosteric sites of a protein and to understand the mechanism of the protein, helping the experimental biologysts to study allostery.
In addition this analysis suggest that not only there is a comunication between allosteric sites and active sites in protein, but also there is a comunication between allosteric sites.
Finaly this comunication between the allosteric sites and the binding pocket at equilibrium is reciprocal.
It would obviously be interesting to study the protein in out-of-equilibrium conditions to understand if the cause-effect behavior and the reciprocal communication between the allosteric sites and the binding pocket is still valid or if one will start to predominate over the other.

Chapter 6

Out-of-Equilibrium Stochastic Processes Induced by Heat Gradients

The first out-of-equilibrium condition for studying the protein is induced by a heat gradient.
In practice, what we will do is place the residues at two different temperatures based on their relative positions in the protein.
Infact it is imaginable that atoms inside the protein are more strongly bound and thus fluctuate less.
We introduced this heat gradient because we believe that allostery is often thought to emerge predominantly under non-equilibrium conditions.

6.1 Stochastic Dynamics Under a Heat Gradient

In the equilibrium phase, as discussed in Chapter 3.3, the evolution of a system can be described by the following stochastic differential equation:

$$\gamma \frac{d\mathbf{x}}{dt} = -g\mathbf{K}\mathbf{x} + \sqrt{2\gamma k_B T}\boldsymbol{\eta}(t). \tag{6.1}$$

In a system subjected to a heat gradient, the temperature varies spatially across the protein structure, with each residue experiencing a local temperature $T(x)$ that depends on its position. This introduces a non-equilibrium condition driven by spatially dependent thermal fluctuations.
Consequently, the motion equation is modified to account for this temperature gradient:

$$\gamma \frac{d\mathbf{x}(\mathbf{t})}{dt} = -g\mathbf{K}\mathbf{x}(\mathbf{t}) + \mathbf{B}\boldsymbol{\eta}(t). \tag{6.2}$$

where B is a diagonal matrix that incorporates the temperature of each residue. Specifically, $B = \sqrt{2\gamma k_B T(x)}$, making the noise term position-dependent.
Applying a similar sostituion of the previous chapter, but this time writing the Hamiltonian in unit of k_B, the equation simplifies to:

$$\frac{d\mathbf{x}(\mathbf{t})}{dt} = -g\mathbf{K}\mathbf{x}(\mathbf{t}) + \mathbf{B}\boldsymbol{\eta}(t). \tag{6.3}$$

where B is a diagonal matrix further refined to account for the discrete temperature values by incorporating a Kronecker delta function:

$$B_{i,j}(x) = \delta_{i,j} \cdot \sqrt{T(x)}. \tag{6.4}$$

This model enables a detailed exploration of the stochastic dynamics of proteins under non-equilibrium conditions.
Now we want to solve in the prevoius equation with the normal mode analysis so we can obtain the covariance function to analyze how the residues interact each other and we can obtain also the transfer entropy.
Substituting $X(t) = \mathbf{V}\mathbf{Q}(t)$ in the motion equation, proceeding in the same way of the equilibrium case, we obtain:

$$\frac{d(\mathbf{V}\mathbf{Q}(t))}{dt} = -g\mathbf{K}(\mathbf{V}\mathbf{Q}(t)) + \mathbf{B}_\mathbf{J}(t).$$

Substituting $\mathbf{K} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$, the term $-\mathbf{K}(\mathbf{V}\mathbf{Q}(t))$ becomes:

$$-\mathbf{K}(\mathbf{V}\mathbf{Q}(t)) = -\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top(\mathbf{V}\mathbf{Q}(t)) = -\mathbf{V}\boldsymbol{\Lambda}\mathbf{Q}(t)..$$

Thus the equation becomes:

$$\mathbf{V}\frac{d\mathbf{Q}(t)}{dt} = -g\mathbf{V}\boldsymbol{\Lambda}\mathbf{Q}(t) + \mathbf{B}_\mathbf{J}(t).$$

Multiplying both sides by $\mathbf{V}^\top$ (to return to the eigenvector space):

$$\frac{d\mathbf{Q}(t)}{dt} = -g\boldsymbol{\Lambda}\mathbf{Q}(t) + \mathbf{V}^\top\mathbf{B}_\mathbf{J}(t).$$

So for each component i, we have:

$$\frac{dQ_i(t)}{dt} = -g\lambda_i Q_i(t) + \sum_j v_{ji}^\top B_{jj}\eta_j(t).$$

As in the equilibrium case, the deterministic term $-\lambda_i Q_i(t)$ describes the evolution of the components $Q_i(t)$ under the influence of the eigenvalue λ_i, instead the stochastic term $\sum_j v_{ji}^\top B_{jj}\eta_j(t)$ represents the effect of noise transformed in the eigenvector space.
The solution is:

$$Q_i(t) = \int_{-\infty}^{t} e^{-g\lambda_i(t-\tau)} \sum_j v_{ji}^\top B_{jj}\eta_j(\tau)\, d\tau.$$

We now compute the Covariance between normal modes:

$$\langle Q_i(t)Q_j(s)\rangle = \left\langle \int_{-\infty}^{t} e^{-g\lambda_i(t-\tau)} \sum_k v_{ki}^\top B_{kk}\eta_k(\tau)\, d\tau \int_{-\infty}^{s} e^{-g\lambda_j(s-\tau')} \sum_l v_{lj}^\top B_{ll}\eta_l(\tau')\, d\tau' \right\rangle.$$

Expanding, we get:

$$\langle Q_i(t)Q_j(s)\rangle = \int_{-\infty}^{t}\int_{-\infty}^{s} e^{-g\lambda_i(t-\tau)}e^{-g\lambda_j(s-\tau')} \sum_k\sum_l v_{ki}^\top B_{kk}v_{lj}^\top B_{ll}\langle\eta_k(\tau')\eta_l(\tau')\rangle\, d\tau\, d\tau'.$$

The noise term $\langle\eta_k(\tau)\eta_l(\tau)\rangle$ satisfies:

$$\langle\eta_k(\tau)\eta_l(\tau')\rangle = \delta_{kl}\delta(\tau-\tau'),$$

which simplifies the double integral:

$$\langle Q_i(t)Q_j(s)\rangle = \sum_k (v_{ki}^\top B_{kk})(v_{kj}^\top B_{kk}) \int_{-\infty}^{min(t,s)} e^{-g\lambda_i(t-\tau)}e^{-g\lambda_j(s-\tau)}\, d\tau.$$

Evaluating the integral:

$$\int_{-\infty}^{min(t,s)} e^{-g(\lambda_i t+\lambda_j s)}e^{g(\lambda_i+\lambda_j)\tau}\, d\tau = \frac{e^{-g(\lambda_i t+\lambda_j s)}}{g(\lambda_i+\lambda_j)}\left[e^{g(\lambda_i+\lambda_j)min(t,s)}\right].$$

So finally we obtain:

$$\langle Q_i(t)Q_j(s)\rangle = \sum_k \frac{(v_{ki}^\top B_{kk})(v_{kj}^\top B_{kk})}{g(\lambda_i+\lambda_j)} e^{-g\lambda_i|t-s|}.$$

Now we can caompute teh covariance easily in the original space:

$$\langle X_i(t)X_j^\top(s)\rangle = \sum_k\sum_p\sum_m \frac{v_{ik}v_{jp}(v_{mk}^\top B_{mm}^2 v_{mp}^\top)}{g(\lambda_k+\lambda_p)} e^{-g\lambda_k|t-s|}.$$

The formula describes the temporal Covariance between the components $X_i(t)$ and $X_j(s)$ in a stochastic system under the influence of noise and a gradient of temperature.
Analyzing the formula we have the eigenvectors (v_{ik} and v_{jp}). Their role is to write

the Covariance in the eigenbasis of the matrix K. Instead the diagonal matrix B^2_{mm} is the fundamental term becasue it introduces the dependence from the temperature and thus the non equilibrium. Finally the term $e^{-g\lambda_k|t-s|}$ encapsulates the temporal decay of the Covariance, governed by the eigenvalue λ_k, more the eigenvalues are larger, faster the covariance decays.
The non equilibrium is also highlighted by the fact that the covariance is no more simmetric in time, suggesting a precise information direction.
The denominator ensures that the contributions from eigenmodes are properly weighted by their respective damping rates λ_k and λ_p. As before the transfer entropy it is completely determined by the Covariance. Otherwise define the linear response out of equilibrium is not easy, however is not our aim.

6.2 Temperature determination

As we said, we set two distinct temperature regimes.
One temperature is fixed at $T = 1$, representing a high-energy state, while the other temperature varies between $T = 0$ and $T = 1$ for each experiment, so it is $T = 1 - \epsilon$,, where $\epsilon = \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0\}$ for a total of 11 experiments (Note that when $epsilon = 0$ we are in the equilibrium case as before, so the results must be the same).
This configuration allows us to drive the system out of equilibrium and study the allosteric behavior under non-equilibrium conditions.
The temperature assigned to a given residue is determined based on its connectivity within the protein structure. Specifically, if a residue has five or more connections, it is assigned a temperature of $T = 1 - \epsilon$, (varying temperature); otherwise, it is assigned $T = 1$. This methodology links the temperature directly to the residue's position and role within the protein network.
Residues located near the center of the protein typically exhibit a higher number of connections due to their central role in maintaining structural integrity.
These residues tend to fluctuate less and are therefore associated with lower effective temperatures.
Conversely, residues with fewer connections, often found on the periphery, are more flexible and are assigned higher temperatures.
In the images 6.1 we have a represention of the different temperatures of every

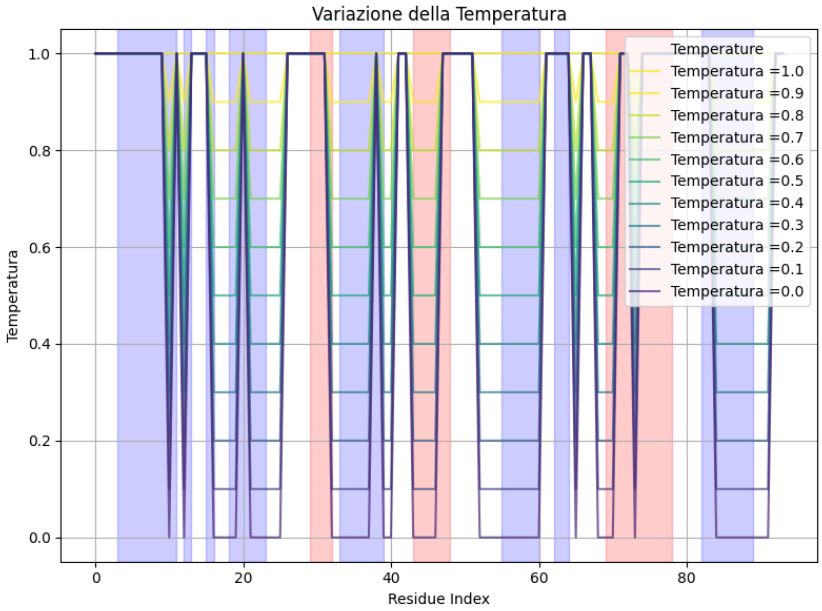


Figure 6.1. Temperature vs residue number

residues.
What we can see is that in the beta sheet regions residues tend to have higher connectivity due to their compact and stabilizing arrangement in the protein's core, otherwhise in the Alpha helix regions typically exhibit intermediate levels of

connectivity.
Finally residues located in loops or on the protein's surface generally have fewer connections, as these regions are more exposed and less structurally integrated.
now that we hace a deeper understanding of the non equilibrium condition we are ready to compute and anlyze the allostery through the causal indicators.

Chapter 7

Results of the non equilibrium dynamic

In this chapter we will expose the results obtain for the non-equilibrium stochastic process describing the oscilaltions of atoms in the protein under a heat gradient.
We will compare the results with the equilibrium case to see how the heat gradient affects the protein.
For the connection radius and for the Kirchhoff matrix are valid the precedence arguments in section 5.1.
Moreover we used the same characteristic time finded in the equilibrium case.
First we will describe the covariance and the beta factors to evaluate the quality of the model and after we will check how the heat gradient affects the protein's allosteric mechanism.

7.1 Covariance Matrix and Beta Factors

Let's start by analyzing the Covariance matrices obtained from the non-equilibrium stochastic process compared to the equlibrium case.

In the image 7.3, we observe the negative part of the Covariance matrix at two different temperature settings: The left matrix ($T = 1 - \epsilon = 0.2$) shows localized and intense negative covariance clusters confronting to the equilibrium covariance matrix.
This finer detail indicates that at lower temperatures, the protein's structural elements are less mobile, leading to more pronounced and localized interactions between residues.
This seems coherent.
In contrast, the right matrix ($T = 1 - \epsilon = 1$ (equilibrium case)) exhibits smoother covariance patterns, which suggests increased mobility of the residues in a disordered way.
At this equilibrium temperature, the increased thermal energy allows for greater mobility of the protein's residues, smoothing out the localized negative covariance patterns seen at lower temperatures.
This enhanced flexibility can be advantageous for proteins that need to adapt to

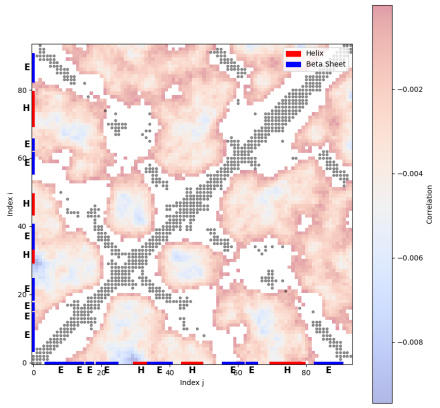


Figure 7.1. Negative covariance between residues at time 0 with $T = 1 - \epsilon = 0.2$.

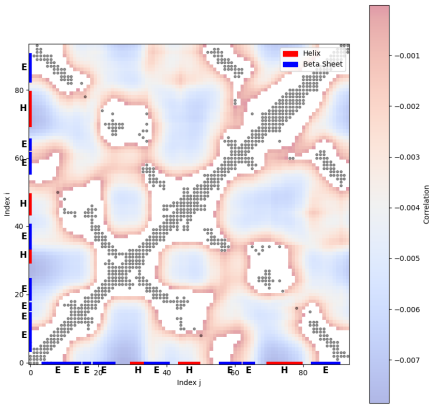


Figure 7.2. Negative covariance between residues at time 0 with $T = 1 - \epsilon = 1$ (equilibrium case).

Figure 7.3. Comparison of negative covariances between residues at time 0 for different T values.

different functional states or interact with various molecular partners.
Obviously at higher temperatures, the smooth patterns suggest that the residues are moving more freely, reflecting increased kinetic energy.
This flexibility allows proteins to adapt their shapes more readily, accommodating various biochemical interactions required for cellular processes.
The ability to adjust structure with temperature facilitates essential functions such as substrate binding in enzymes.

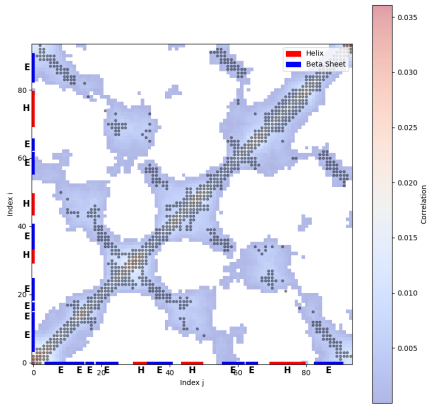


Figure 7.4. Positive covariance between residues at time 0 with $T = 1 - \epsilon = 0.2$.

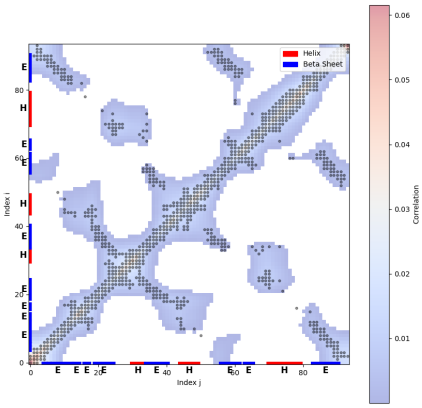


Figure 7.5. Positive covariance between residues at time 0 with $T = 1 - \epsilon = 1$ (equilibrium case).

Figure 7.6. Comparison of positive covariances between residues at time 0 for different T values.

Instead the image 7.6 represents the positive part of the Covariance matrix at two different temperatures and the analysis is the viceversa of the previous case: At $T = 1 - \epsilon = 0.2$, the Covariance patterns are more sparse and lower.
This localization suggests that at lower temperatures, protein structures are more rigid and interactions between residues are more specific, potentially stabilizing certain functional forms or interactions essential for biological activity in colder environments.
instead at $T = 1 - \epsilon = 1$(equilibrium case), the Covariances are more localized, suggesting increased flexibility.

Now that we have the covariance matrix we can calculate the beta factors to evaluate the quality of the model.
The plot above compares the experimental B-factors (blue curve), represented by the formula (4.9) $B_i = 8\pi^2 C_{ii}$, with the predicted B-factors (red curve) along the residue index.
These metrics provide a quantitative assessment of the model's accuracy in capturing the protein's dynamic behavior, revealing how well the theoretical models approximate realresidues vibrations.

ΔT	MAE	RMSE
$T = 1$	6.4146	8.5200
$T = 0.9$	6.3206	8.4244
$T = 0.8$	6.2239	8.3239
$T = 0.7$	6.1239	8.2191
$T = 0.6$	6.0212	8.1107
$T = 0.5$	5.9457	8.0004
$T = 0.4$	5.8701	7.8913
$T = 0.3$	5.8117	7.7885
$T = 0.2$	5.7542	7.7003
$T = 0.1$	5.7030	7.6413
$T = 0.0$	5.6775	7.6319

Table 7.1. Errors of MAE and RMSE for different values of ΔT

By analyzing the system in and out of equilibrium $\Delta T = 1 - \epsilon$, we observe that the system driven out of equilibrium ($\Delta T = 1$) fits better the beta factors, resulting in lower error metrics and better alignment between predicted and experimental B-factors.
This indicates that the system's response to a larger temperature gradient is more realistic and so also the real world protein exhibits different intensity in the fluctuations of its atom internally and externally.
In particular now we can catch in a better way the spikes around the $alpha - \alpha$ and around the end of $alpha - \beta$ and $alpha - \gamma$ helixs
In conclusion the out of equilibrium model is a better approximation of the real protein's behavior.

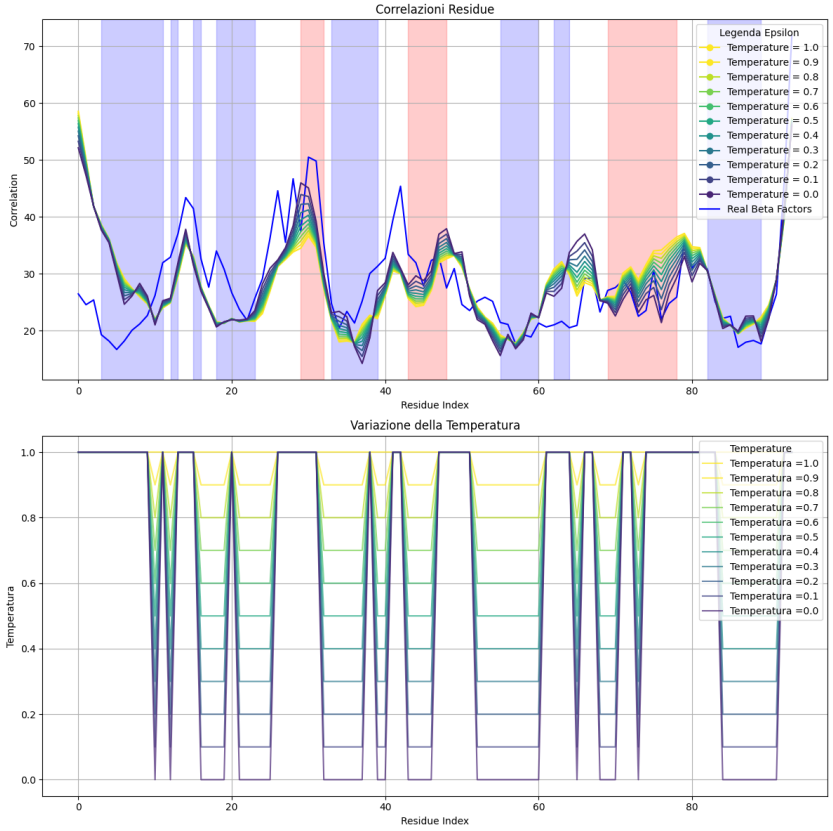


Figure 7.7. Beta factors with different ΔT.

7.2 Causal indicators between residues

In this section our aim is to understand how the out of equilibrium condition affects the causal allosteric relationship between residues.
We will anlyze the covariance and the transfer entropy between residues, confronting the results with the equilibrium condition.
It is interesting to see how the heat gradient affects the signal propagation in the protein and if the reciprocal mechanism of allostery find in the equilibrium case is still valid.
Let's start with the covariance between the residues: We want to see, as before, a high absolute value of covariance between the allosteric sites and the active sites in the binding pocket.

Analyzing figure 7.8 we notice that are valid the conclusion of the equilibrium part, as we said: There are a high Covariance observed near the binding pocket,as we expected, particularly between the alpha-α and beta-β regions, and another allosteric site in the alpha-γ helix, highlights regions of significant allosteric communication. Moreover we see also strong values of covariance in the head and in the tail of the protein, probability caused by spourious interactions.

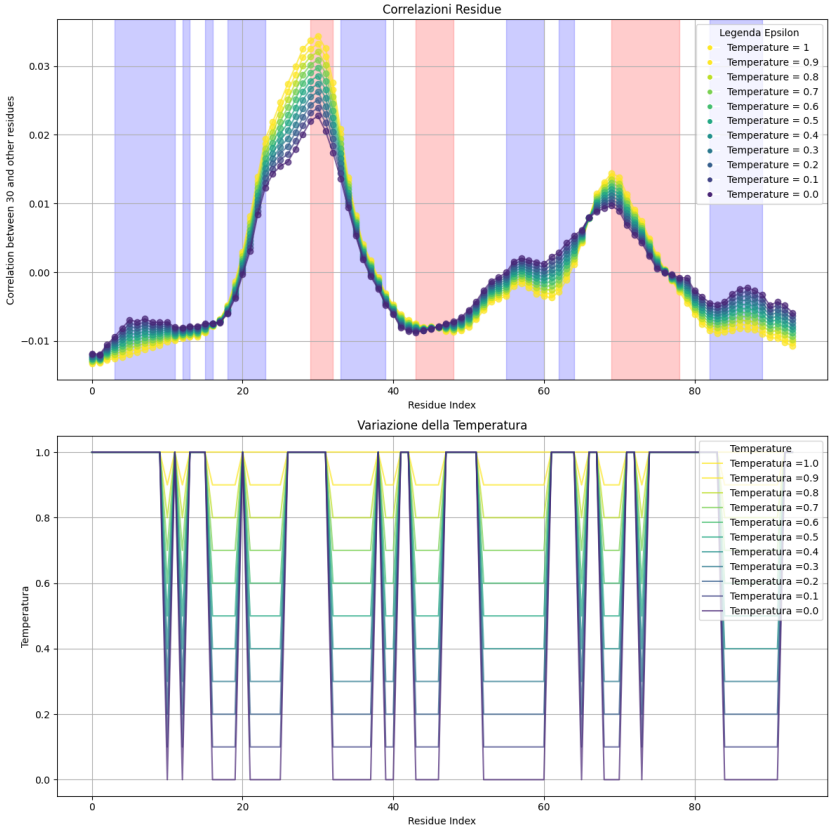


Figure 7.8. Dynamic covariance observed from perturbating the 31-th residue, illustrating signal propagation through alpha-*alpha* helical regions out of equilibrium.

Also for the Covariance of residue 73 7.9 the general conclusion is still the same of the equilibrium case.

Moreover for both the images when temperatures are higher we can see how more dispersed and higher in modulus the covariance is.
Instead at lower temperatures the covariance is more localized.
Especially we see a lot of variance of the correlation in function of the temperature around the alpha-α-helix.
We can also see that the covariance for some residues are completely indifferent to the heat gradient.
Finally we see also that the values of the covariance when we introduce the gradient not become all higher o lower but some become higher and some other lower.
In conclusion the allosteric behavior in the covariances of the protein is the same of the equilibrium case but the model explain in a better way the beta factors.

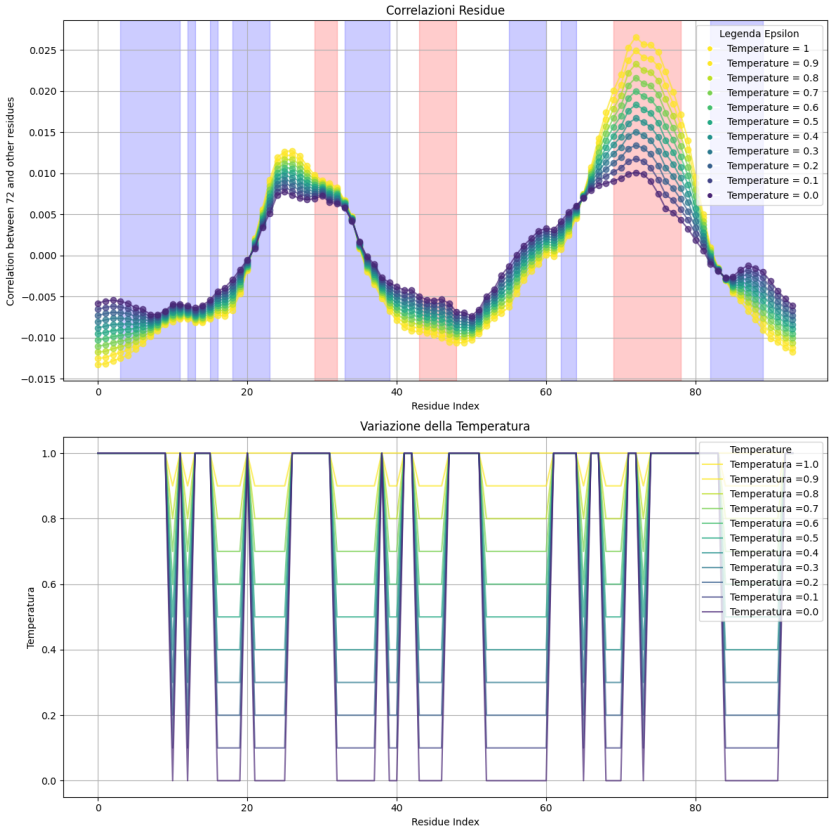


Figure 7.9. Dynamic covariance observed from perturbating the 73-rd residue, illustrating signal propagation through alpha-*gamma* helical regions out of equilibrium.

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, Allosteric regulation

[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] Covariance, response and entropy approaches to allosteric behaviors: a critical comparison on the ubiquitin case Fabio Cecconi, Giulio constantini, 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 Bueti, and Alessandro Laio

[14] Allosterism in the PDZ Family, Amy O. Stevens and Yi He.