Imperial College of Science, Technology and Medicine
Department of Chemistry

Approaches for studying allostery using network theory

Maxwell Hodges

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of Imperial College London

Copyright

Abstract

Allostery is the process whereby binding of a substrate at a site other than the active site modulates the function of a protein. Allostery is thus one of the myriad of biological processes that keeps cells under tight regulatory control, specifically one that acts at the level of the protein rather than through changes in gene transciption or translation of mRNA. Despite over 50 years of investigation, allostery has remained a difficult phenomenon to elucidate. Structural changes are often too subtle for many experimental methods to capture and it has become increasingly obvious that a range of timescales are involved, from extremely fast pico- to nanosecond local fluctuations all the way up to the millisecond or even second timescales over which the biological effects of allostery are observed. As a result, computational methods have arisen to become a powerful means of studying allostery, aided greatly by the staggering increases in computational power over the last 70 years.

A field that has experienced a surge in interest over the last 20 years or so is *network theory*, perhaps stimulated by the development of the internet and the Web, two examples of immensely important networks in our everyday life. One of the reasons for the popularity of networks in modelling is their comparative simplicity: a network consists of *nodes*, representing a set of objects in a system, and *edges*, that capture the relations between them.

In this thesis, we both apply existing ideas and methods from network theory and develop new computational network methods to study allostery in proteins. We attempt to tackle this problem in three distinct ways, each representing a protein using a different form of a network. Our initial work follows on logically from previous work in the group, representing proteins as *graphs* where atoms are nodes and bonds are energy weighted edges. In effect we disregard the 3-dimensional structure of the protein and instead focus on how the bond *connectivity* can be used to explain potential long range communication between allosteric and active sites in a multimeric protein. We then focus on a class of protein models known as *elastic network models*, in which our edges now correspond to mechanical Hooke springs between either atoms or residues, in order to attempt to understand the physical, mechanistic basis of allostery.

Declaration of Originality

I confirm that this thesis was produced by myself and the work presented herein is entirely my own unless referenced otherwise. Further, this thesis has not been submitted for any other degree or professional qualification.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Despite its critical importance to cellular regulation, allostery is still not a well understood phenomenon. Historically, theories about how allostery is governed have tended to be limited by the available experimental evidence. Initial explanations focused on thermodynamics descriptions of allostery, whereby binding of a ligand shifts the equilibrium of the protein to attenuate or enhance activity. Whilst such models[165, 166] are still useful, they are ultimately phenomenological and unable to make predictions about specific cases. Later, as high resolution crystal structures became available, the likes of Perutz[188] were able to begin to provide a structural basis for allostery based on comparisons of active and inactive structures. The concept of structural pathways began to emerge as an explanation as to how binding of a ligand at one site on the protein could cause an apparent functional change at the active site. Much later statistical studies on evolutionarily conserved residues by Lockless *et al*[142] also pointed towards pathways of residues in proteins. Here, a multiple sequence alignment (MSA) is used to find those pairs of residues that are statistically coupled, defined as the extent to which the type of the amino acid at one site changes in response to an alteration at a different site over the set of sequences. However, recent conceptions of allostery as a function of the free energy landscape, such as the ensemble allosteric model[168] have instead suggested that structural pathways are not necessary, and that instead allostery should be considered a property of an ensemble of proteins. Furthermore, earlier work by Cooper and Dryden[43] had raised the prospect of entropic contributions to allostery, casting further doubt on the structural view as the dominant driver.

The main aim of this thesis then is to elucidate whether long range structural perturbations from allosteric sites may be a plausible mechanism of allostery. There are currently no experimental methods that have the level of resolution required to observe this effect directly and as such, computational methods have been recruited towards this purpose. Perhaps the primary tool in this area is *molecular dynamics*, however even now the simulation of proteins on the time scale required for biological processes such as allostery (generally occurring from the millisecond[124] to the second range[92]) is extremely challenging. Here instead we draw on ideas from *network theory*, allowing us to develop efficient methods that probe various simplified network representations of proteins.

From a practical point of view, a deep understanding of allostery would offer great opportunities in the field of drug discovery[180]. The vast majority of current drugs target the active site of proteins, which often leads to problems with off-target effects given many proteins exhibit homology. Allosteric drugs by contrast have the potential to be both more specific and modulate protein function far more precisely and as such both a general understanding of allostery and a means to identify potential allosteric sites is of great theraputic interest.

1.2 Thesis outline

In Chapter 2 we introduce the requisite ideas and mathematics from network theory that we will use consistently throughout this thesis, with a particular focus on random walk dynamics and percolation which are both areas that deal with flow and communication on networks. Chapter 3 then introduces the biological aspects of the project. Allostery is discussed from two distinct but complementary viewpoints: the traditional *thermodynamic* explanation, which is supplemented with a more modern approach that considers the entire free energy landscape of the protein, and the *structural* view that considers how the actual mechanism of transition between protein states is triggered by binding of small ligands at some site on the protein surface.

Chapter 4 utilises a method called *bond-to-bond propensities* that was introduced by Amor *et al*[5] to find allosteric sites in proteins given only knowledge of the active site using a *graph* representation of the protein. Here, we use the method to model the effect of binding of an allosteric ligand to ATCase, a large allosteric protein and discover those parts of the protein that are particularly energetically coupled to the allosteric site.

We attempt to extend the principles of *bond-to-bond propensities* to elastic models of proteins in Chapter 5 where edges are now springs, developing a method called *elastic response* that models the effect of an ligand binding to a protein as the propagation of strain through the edges of the elastic model away from the allosteric site. We also use the same mechanical framework to develop a method called *interaction embeddedness* to find those edges that experience the highest average strain when the elastic network fluctuates randomly in a heat bath and link these high strain regions to protein function. In Chapter 6 we use a form of convex optimization called *semidefinite programming* to design *de novo* allosteric elastic networks that exhibit long range mechanical effects by optimizing the spring constants of the springs in the network.

Finally, we provide possible future directions for the methodology used in this thesis, in particular to Markov State models. There, each node in the network is a microstate of the protein derived from a molecular dynamics trajectory and the network thus takes the form of a Markov matrix that represents a discrete approximation to the free energy landscape.

1.3 Publications

The results presented in Chapter 4 were based on the work published in:

Hodges, M., Barahona, M. Yaliraki, S. N. *Allostery and cooperativity in multimeric proteins: bond-to-bond propensities in ATCase.* Sci. Rep. 8, 11079 (2018).

Chapter 2

Network Theory

> “*Why is network anatomy so important to characterize? Because structure always affects function.*”
>
> Steven Strogatz, *Exploring complex networks*

2.1 Networks: an overview

The last twenty years have seen an explosion of interest in the study of networks, undoubtedly due to the pervasiveness of such structures and the raft of data made available by the digital era. A full treatment of the extent of networks' appearances in apparently disparate areas of the literature is not possible here, but metabolism and gene regulation, citations, transport, ecology, the Web and the brain are just a small sample of those subjects that have harnessed (and further developed) the machinery of network science. Indeed, the development of the field of networks, under the broader umbrella of *complex systems*, across such a wide range of disciplines is perhaps not surprising when noting Steven Strogatz's remark above, that ultimately if we wish to understand how a system works, we must first elucidate how its various parts interact. There has then, somewhat recently, been an attempt to establish certain unifying principles that apply to networks across a range of different fields and as such, network theory has emerged as its own domain, though one that continues to be rooted firmly at the cross section of many other fields.

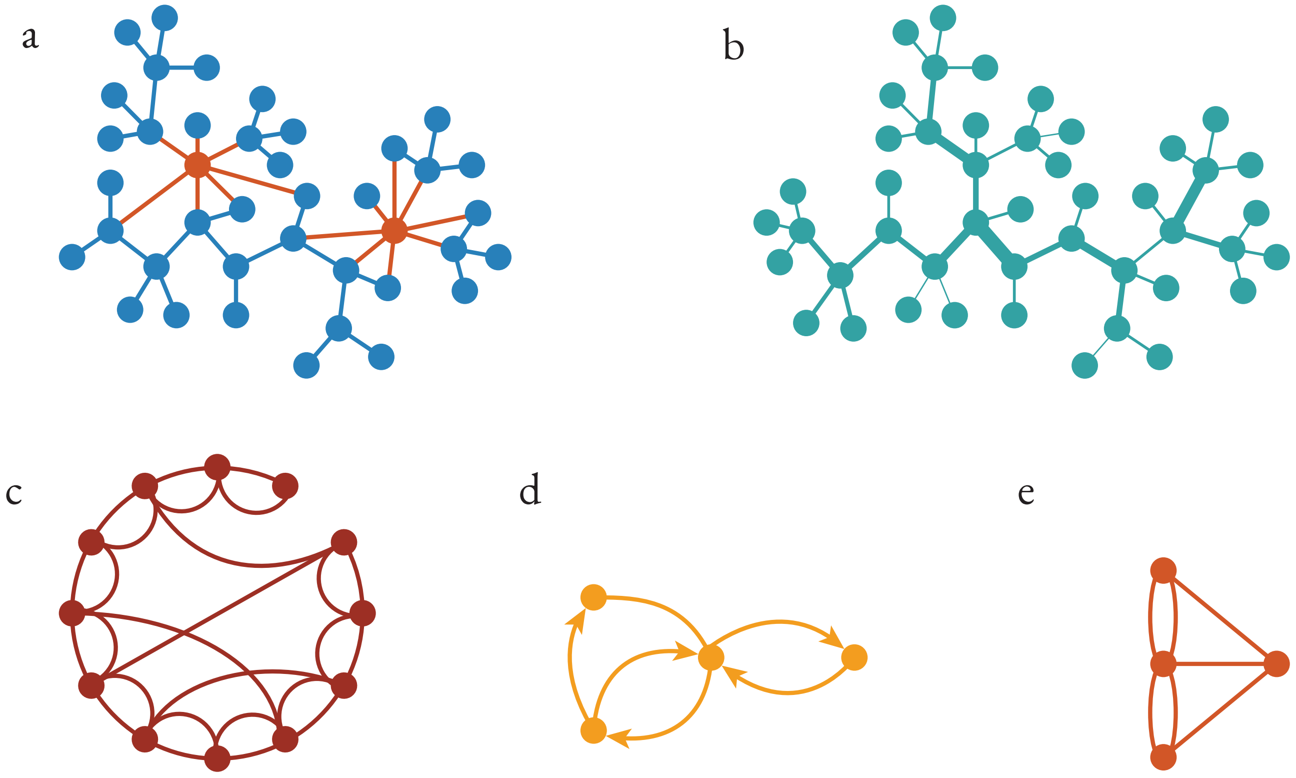


Figure 2.1: A number of commonly studied networks. a) A 'scale free' network[16], in which most nodes have small degree whilst a small number (in orange) have a very large number of connections such that the degree distribution follows a power law. b) A weighted network, in which the edges may take different values. The weight of an edge may indicate the strength of connection between two nodes for example. c) A small world network. By starting with a regular lattice, rewiring just a small number of edges significantly decreases the average shortest path length[240]. d) A directed network. e) The graph from the Bridges of Königsberg problem.

The common starting point for any discussion of networks is the solution by Euler in 1736 to the Seven Bridges of Königsberg problem, which is usually remarked to be the creation of the subject of graph theory. The terms graph theory and network theory are often used interchangeably and the difference is perhaps more one of emphasis, with network theory describing the application of mathematical methods to real world systems, rather than the study of networks or graphs for their own sake. In any case, Euler did solve a real world problem using the central abstraction of networks: that a system may be represented as a set of *nodes* (or vertices) that are joined together by a collection of *edges* (or links) as shown in Figure 2.1e. By doing so, Euler proved the problem of visiting all islands (the nodes) whilst crossing each bridge (the edges) only once had no solution: either zero or two nodes can have an odd number of edges joining them to act as the end points of the walk, but every other node must have an even number of edges to allow a walker to arrive, then depart. The Königsberg problem has 4 nodes with odd numbers of edges so is not traversable without revisiting edges.

Whilst many facets of networks are of great interest[172, 160], we focus here on those that will be used throughout this report. A network is said to be *unweighted* if its edges take on values of either 1 (presence of an edge) or 0 (no edge). *Weighted* edges may take on any value, though we shall only encounter those from the real numbers $\mathbb{R}$. An edge between two nodes can be *undirected* or it may possess a specific direction, in which case it is *directed*, though it should be noted that an undirected edge is equivalent to two directed edges of equal weight pointing in opposite directions. The *degree* of a node is the sum of the weights of the edges connected to that node (hence is simply the number of joined edges in the unweighted case). For directed networks, we often differentiate between the *in-degree* and the *out-degree*, which as might be expected, refer to the total weights of edges into and out of a node.

A network of N nodes may be represented mathematically by an $N \times N$ *adjacency matrix*, A, such that the entry A_{ij} is equal to the weight of the edge between nodes i and j in a network containing N nodes. If the network is undirected, then the adjacency matrix is *symmetric*. It is possible for nodes to possess *self loops* (i.e. an edge from the node back to itself), which show up as entries along the diagonal of the adjacency matrix. If the degrees of each of the nodes are compiled into an $N \times 1$ vector d, we may then define a diagonal matrix of node degrees: $D = \text{diag}(d)$. From this, we can define a representation of the network that we will use more commonly in this report, the Laplacian matrix: $L = D - A$. The entries of the Laplacian are thus:

$$L_{ij} = \begin{cases} -w_{ij}, & i \neq j \\ \sum_i w_{ij}, & i = j \end{cases} \tag{2.1}$$

with w_{ij} the weight of an edge from nodes i to j. Another useful description of a network is given by the *incidence matrix* that maps pairs of nodes to the edges that link them. Representing the incidence matrix B as an $E \times N$ matrix, each row corresponds to an edge with an entry of 1 at the index of node i and -1 at index j if the edge joins nodes i and j. The importance of the permutation of the signs depends on the system in question; for some networks it does not matter which way round it is (i.e. it would not matter if the entry of 1 was at entry j and -1 at i) but where direction is important (such as the spring models of Chapter 5), or for directed networks, a consideration of the meaning of the signs is crucial. The definition of the incidence matrix allows us to provide an alternate (but equivalent) construction of the Laplacian matrix:

$$L = B^T G B \tag{2.2}$$

where G is a diagonal matrix containing the edge weights of the network. The undirected Laplacian is positive semidefinite, which can be seen using Eq.(2.2). For any vector v:

$$\begin{aligned} v^T L v &= v^T B^T G B v \\ &= (Bv)^T G (Bv) \geq 0 \end{aligned} \tag{2.3}$$

which is just the sum of a set of squares, each multiplied by a positive number, which must be positive. If the network is connected, which for an undirected network consequently means that any node is reachable from any other node, then the Laplacian has a nullspace of dimension 1, corresponding to the zero eigenvalue. The nullspace is spanned by (any scalar mulitple of) the vector $v = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \end{pmatrix}$, which in electrical networks is physically represented by *Kirchoff's Voltage Law* that says the potential difference around any closed loop must be zero.

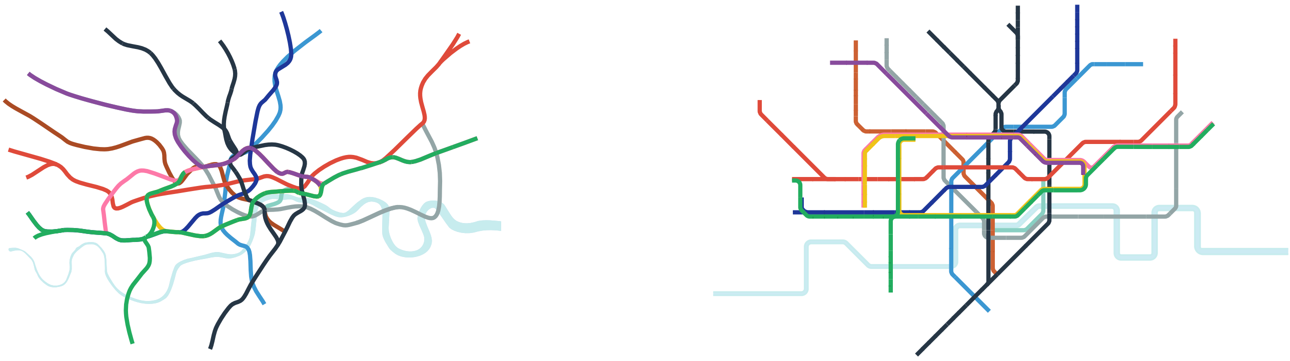


Figure 2.2: Left: a geographically accurate representation of the London Underground. Right: the familiar *topological* form of the map designed by Harry Beck in 1931 who, taking inspiration from electric diagrams, realised the *connectivity* of the network was more important to passengers than their actual location.

2.2 Random walks on networks

A particular aspect of networks we focus on in this report is that of a *random walk* process occurring on the network. An excellent review of simple random walks on networks is provided by Masuda *et al*[151] but we again provide some relevant details here.

We first point out the relationship between a random walk on a network and a *Markov chain*[230, 177]:

if we consider the outgoing probability of a walker leaving a node along a particular edge, as the weight of that edge divided by the total weight of edges exiting the node, then we have defined a Markov chain. For a *discrete time* random walk, we can make this relationship explicit:

$$M = D^{-1}A \tag{2.4}$$

M is the *Markov matrix* that contains the probability of a random walker jumping from node i to node j in the entry M_{ij}. For directed networks, we replace D with D_{out}, the diagonal matrix of the weights of edges leaving each of the nodes. If we then consider the *probability distribution* of the random walk over the N nodes contained in an $N \times 1$ vector p:

$$p_{t+1} = p_t M \tag{2.5}$$

By induction and defining some starting distribution p_0, we can see that the probability distribution on the network at some positive integer timestep t is simply:

$$p_t = p_0 M^t \tag{2.6}$$

These last two relationships rely on the defining property of Markov chains, that they are *memoryless*. In words, this simply means that when a random walker reaches a particular node i, the probability of where it goes next depends only on its current position, not on where it came from previously:

$$\Pr\left(X_{n+1} = x |\, X_1 = x_1,\ X_2 = x_2, \cdots,\ X_n = x_n\right) \tag{2.7}$$

$$= \Pr\left(X_{n+1} = x | X_n = x_n\right) \tag{2.8}$$

Random walks that do have memory are possible[200, 207], including the creatively named elephant random walk that retains entire memory of where it has been[210]. We restrict ourselves here however only to those that are *Markovian* and have no knowledge of their past trajectory. Note that M is a *stochastic matrix* so that its rows sum to 1 (in fact in the undirected case, M is doubly stochastic as its columns also sum to 1), capturing the fact that once a random walker reaches a node, it must leave along an edge (though a self loop would of course take it back to itself).

In order to generalise the above process to *continuous time*, we consider the dynamics of the walker on the network as being represented by a first order differential equation:

$$\frac{dp}{dt} = -Lp \tag{2.9}$$

where p is an $N \times 1$ vector whose i^{th} entry is the probability of a walker being on node i in a network and thus the entries of p sum to 1. L is the *discrete Laplacian*, so named because of its correspondence with the continuous Laplacian operator. If we write the *heat equation*[19] for continuous space:

$$\frac{dp}{dt} = -\nabla^2 p \tag{2.10}$$

We see the discrete Laplacian takes the place of the Laplace operator, now acting on the discrete space defined by our network. As it appears in Eq. (2.9), we refer to L as the *combinatorial* Laplacian. By considering the *waiting time* of a random walker at a node, we may also construct a different type of random walk in which the waiting time at each node is identical[130]:

$$\frac{dp}{dt} = -D^{-1}Lp \tag{2.11}$$

where D is a diagonal matrix containing the total weight of the edges attached to each node. The operator $D^{-1}L$ is similar (that is it has an identical set of eigenvalues) to the *normalised* Laplacian $L = D^{-1/2}LD^{-1/2}$ and thus we refer to it from now on as the normalised Laplacian. In the combinatorial case (Eq. (2.9)), the waiting time for a walker on a particular node will depend on the total weight of edges out of that node; if the total weight is twice as high, the expected waiting time is half as long. By dividing through the row i of L by the node degree d_i, the matrix D^{-1} has the effect of normalising the expected waiting time a random walker spends on each node (so that a walker will spend as much time on average on a node with 3 edges, as one with 2 edges), hence the name. There are in fact a number of alternatives for of the dynamics of the random walk, for example if we replace D^{-1} by D, we have a model for a *web surfer* who spends *more time* on a node with a greater number of edges[130].

The solution to equation (2.11) is:

$$p_t = p_0 \exp\left(-D^{-1}Lt\right) \tag{2.12}$$

In certain cases, for example for systems with very large numbers of nodes, the numerical calculation of the exponential term can be very expensive. We can linearise by taking only the first two terms of the Taylor series of the matrix exponential:

$$\begin{aligned} p_t &\approx p_0\,[I - D^{-1}L]^t && (2.13)\\ &= p_0\,D^{-1}[D - L]^t && (2.14)\\ &= p_0\,M^t && (2.15) \end{aligned}$$

where we have used the relationships $L = D - A$ and $M = D^{-1}A$. By restricting (2.13) to integer time steps, we can see that we have recovered the discrete time random walk dictated by the Markov matrix.

An important property of Markov chains are their *stationary distributions*, which are guaranteed to exist when the Markov chain is *ergodic* (any node can be reached from any other, so we also assume that the network is *connected*) and *reversible* and all the networks in this thesis satisfy these two properties. The stationary distribution then is the long time distribution that is unchanged under application of the transfer matrix and can be calculated as the leading left eigenvector of the matrix:

$$\pi = \pi T \tag{2.16}$$

where T refers to either $\exp\left(-D^{-1}L\right)$ or M if we have linearised and π is the stationary distribution, which is the same whether we use the full or linearised expression. The stationary distribution can therefore often be thought of as an *equilibrium* state of the network. Even if we construct a directed network where we can move between any two nodes, we do not necessarily guarantee stationarity at long time scale. The additional constraint is that the random walk must be reversible or satisfy *detailed balance*:

$$\pi_i T_{ij} = \pi_j T_{ji} \tag{2.17}$$

or by defining the symmetric matrix Π as diag (π):

$$\Pi T = (\Pi T)^T = T^T \Pi \tag{2.18}$$

The stationary distribution has been commonly used rank nodes, perhaps most famously in Google's PageRank[184] and the closely related HITS algorithm[121] for webpages, both of which are variants of eigenvector centrality[21].

2.3 Percolation

Percolation theory[221] is often utilised when discussing transport or communication through structures, including within proteins[132]. The principle of percolation theory is, imagine if we had a grid of connected sites which could take one of two states: filled or empty. The question is then, what fraction of sites needs to be filled, on average, in order to create a continuous path from one side of the network to the other. More specifically, this describes *site percolation*. If instead we fix the sites to all be filled, then we can instead ask what proportion of the *edges* need to be present to allow for a flow to occur through the network; this is *bond percolation*[24].

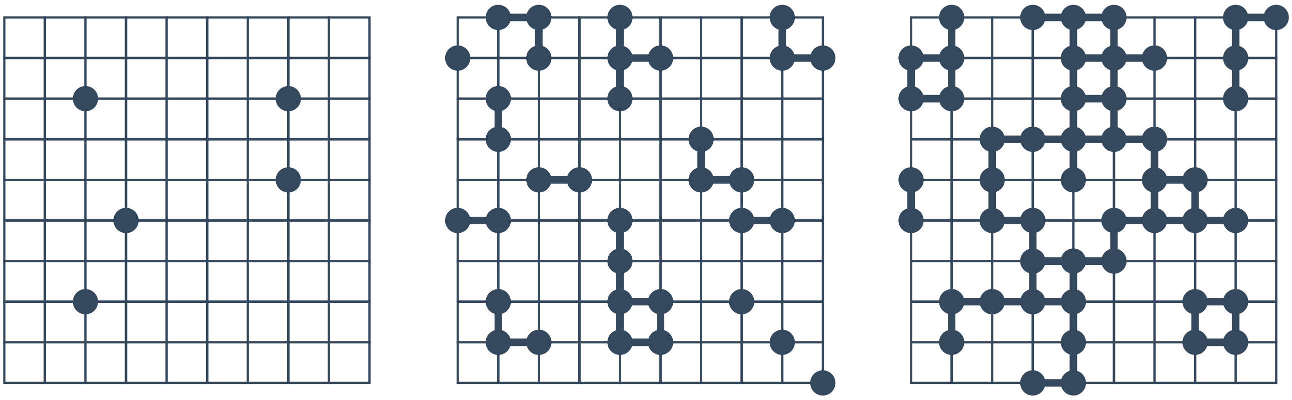


Figure 2.3: Site percolation on a square lattice. Two sites are said to be joined if they are adjacent to each other horizontally or vertically and are both filled. As the fraction of filled sites is increases from left to right, the critical value is surpassed and the inifite cluster forms.

One of the reasons for the intense interest in percolation models is because they are one of the simplest systems to display a *phase transition*. In both site percolation and bond percolation, above a certain *critical value* p_c of the fraction of filled sites or bonds p, a significant number of the sites are connected together to form a *giant cluster*. In the 2-dimensional square lattice for example, p_c for bond percolation is exactly $\frac{1}{2}$[117], though it should be noted that this result (and all analytic results for critical values) strictly hold in the limit as the network size tends to infinity. Most critical points do not have exact values, and indeed the site percolation critical value for the same lattice type is approximately 0.5927[51, 68, 173].

One physical realisation of a percolation problem is a *random resistor network*[119], where current is injected at one node and extracted at another on the other side of the network. Current will thus only flow in the network when there is a contiguous path of edges between the two nodes. As the fraction of edges is increased, the current remains zero until the critical point, when the giant cluster forms, so at that point current can flow. Above the critical point, the resistance of the network continues to drop as a greater number of alternate pathways are available for current to traverse. In elastic network representations of proteins (as explored in Chapter 5), a three dimensional model of the protein is constructed where the network nodes are balls (which are usually atoms or residues) and the edges are Hooke springs (representing chemical bonds or coarse-grained residue-residue interactions). The equivalent percolation problem is then: choosing two nodes at either end of the network and applying a force, in opposite directions, along the straight line between them, is there any resistance to that force?

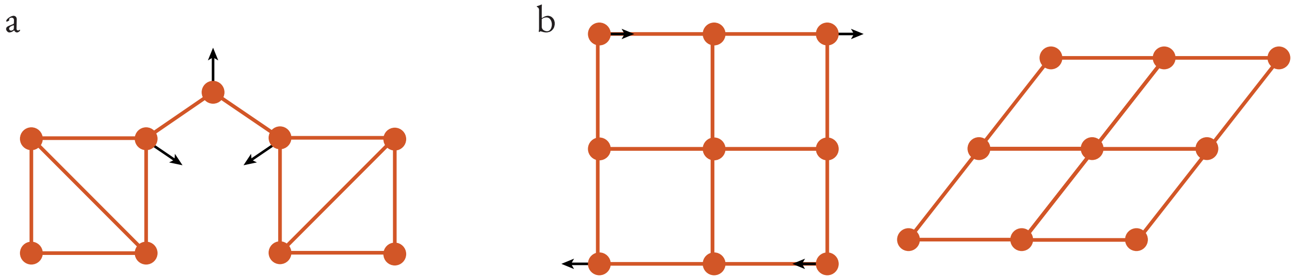


Figure 2.4: a) Two rigid subparts of a network are connected by a single node that has a *floppy mode* and thus rigidity cannot percolate between the two subparts. In the equivalent connectivity percolation problem however, current could flow across the bridging node. b) The square lattice exhibits no resistance to *shear stress*. Each of the bonds stays the same length after applying opposing forces to the top and bottom of the lattice and thus the square lattice can be *deformed* at *zero energy* cost.

It was not initially realised[50] but this *rigidity percolation* problem is in fact distinct from the *connectivity percolation* problem represented by the random resistor network. Feng and Sen[67] note that the rigidity percolation problem involves the propagation of a *vector* through the network (i.e. a 2-vector for displacement in 2D or a 3-vector for 3D), whilst the connectivity problem concerns the transfer of a *scalar* over the network, for example charge in the random resistor model. As a result, when we consider model containing only *central forces*, that is Hooke springs that join pairs of nodes, a rigid cluster must consist of a contiguous set of edge-sharing triangles in 2D or a set of face-sharing tetrahedra in 3D. Those nodes within the network that are not part of a rigid cluster are then free to move whilst admitting no change in the length of their attached springs (as in Fig.2.4a), thus there is a *zero energy* cost to their motion. These nodes are said to have one (or more) *floppy modes*, and the network as a whole can be divided into *rigid* and *floppy* regions.

One way to determine these floppy modes is to construct the $dN \times dN$ *stiffness matrix* for the elastic network (where d is the spatial dimension) and compute the matrix eigendecomposition. Every network structure has a number of *rigid motions*: in 2D there are 2 rigid translations and a rotation in the place, whilst in 3D there are 3 translations and 3 rotations. These motions correspond to eigenvectors of the stiffness matrix with *zero eigenvalues*. Any additional zero eigenvalues equate to floppy modes of the network, from which we can group together rigid and floppy nodes (with a caveat that there are no special symmetries in the network as discussed in Chapter 5). However, particular for larger networks, this approach can be computationally expensive. An alternative approach developed by Thorpe[231, 189] is based on *constraint counting*, an idea that dates back to Maxwell[153]. Each node in the system has d degrees of freedom (d again being the spatial dimension), and the addition of each edge results in the total degrees of freedom being reduced by one, so long as the constraint is *redundant*. A very simple case is that of two free nodes in 2-dimensional space: each node has two degrees of freedom, summing to a total of four. By adding an edge between them, we may still move one (arbitrarily chosen) node wherever we like but the second is now confined to the 1-dimensional line forming a circle around the first node (Fig. 2.5a). Thus our system now has only three degrees of freedom (assuming our edge is rigid, and thus cannot vibrate) corresponding to rigid motions.

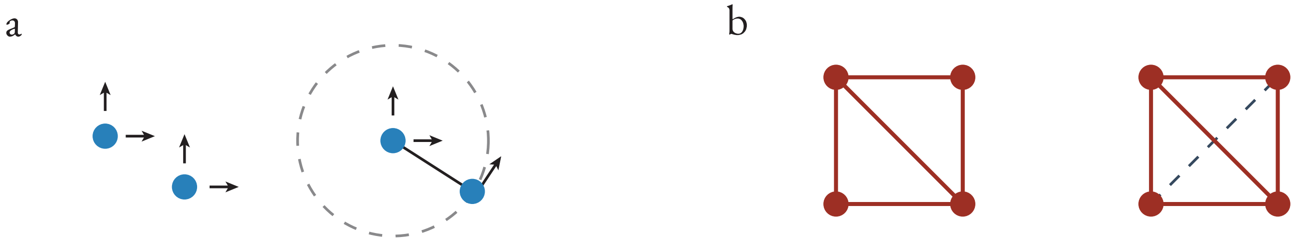


Figure 2.5: a) Whilst two individual nodes in 2-dimensions each have two degrees of freedom (chosen arbitrarily to be the vertical and horizontal directions), the addition of a rigid constraint between them means that relative to the first node, the second node can only move along a 1-dimensional line. b) A rigid system of four nodes in 2-dimensions. The addition of an additional constraint (shown as a dotted line) has no effect on the total degrees of freedom of the system (here simply the three rigid motions) and thus is defined as being *redundant*.

In two dimensions, this idea can be used to find the rigid components of the network using a method called the Pebble game[99], that assigns two "pebbles" (representing the two degrees of freedom) to each node. When an edge is placed between two nodes, one of the pebbles must cover the edge, representing the loss of a degree of freedom. Redundant edges are then those where there are no pebbles left to add to the edge. The pebble game has been extended to 3-dimensions [102, 101] under the name Floppy Inclusion and Rigid Substructure Toplogy (FIRST) in order to study the flexibility of proteins[100, 191]. However the additional caveat is that *angle constraints* must be included in the 3-dimensional case - this is in fact realistic for many bonds, such as covalent bonds, that resist changes to their angles as well as their lengths. The flexible motions of the protein are thus determined by dihedral angles in the FIRST framework.

Chapter 3

Protein dynamics

> “ *Biological function is ultimately rooted in the physical motions of biomolecules.* ”
>
> Katherine Henzler-Wildman & Dorothee Kern, *Dynamic personalities of proteins*

3.1 Molecular Machines

Proteins are the uncomplaining workhorses of nature, shuffling around the crowded environment of the cell to perform their exquisitely specialised tasks, driven only by fundamental physical laws. It is perhaps only relatively recently, however, that the "jigglings and wigglings"[70] of proteins have been pushed to the forefront of the mission to understand how proteins function[92]. By considering the motions of the protein, and by extension the ensemble of states it may exist in, the study of proteins has returned to the realm of statistical physics, aided by a plethora of modern experimental methods that probe the protein at minute timescales. However this is no trivial task, made particularly difficult by the huge range of timescales at which proteins motions occur, from picosecond vibrations of bonds to biological processes such as allostery or protein folding taking place on even the second timescale for larger structures.

Whilst a dizzying array of high resolution X-ray crystal structures populate the Protein Data Bank[17], we must be conscious of the fact that these structures are only a snapshot of a protein. It has been discovered that crystal structures fail to capture some substates of proteins[69] and even within a unit cell of a crystal there may be more than one stable state[190], whilst the structure itself can be affected by the conditions in which it was crystallised[138]. The dynamical equilibrium behaviour of proteins is sometimes said to be captured by the B-factors (also called Debye-Waller factors), but lattice disorder is encapsulated in the value in addition to true atomic fluctuations and so is not a true representation of the ensemble[92]. Furthermore, B-factors may also simply reflect crystallographic errors rather than any sort of intrinsic disorder, such that weak electron density in a part of the structure does not actually correspond to large amplitude motions[103].

3.2 Free energy landscapes of proteins

Traditionally proteins have been viewed as somewhat rigid structures once they have folded into their native state, partly as more rigid structures have tended to be easier to crystallise[248] and as a result, observed conformational changes in proteins even upon complex formation are quite often minor[222]. As such, it is common to read of "states" of a protein, perhaps the active state and the inactive state, such that each state is discrete. A more modern view, covered in depth in a number of reviews[241, 88] posits that instead proteins inhabit a spectrum of states as determined by the Boltzmann distribution. Changes in conditions, such as the addition of an allosteric ligand that increases the rate of a catalyst, are not then reasoned as the binary change from the inactive state to the active state, but instead as the alteration of the underlying landscape such that there is increased *sampling* of the active forms of the protein. Importantly then, it is the energy landscape of the protein that is subject to selection pressure, as modulation of the landscape determines protein function. Though the idea is not new, being introduced nearly 30 years ago by Frauenfelder *et al*[73], it has been made compelling by a range of experimental and computational techniques (many of which are discussed later in this chapter).

Indeed the concept of the energy landscape is hardly foreign to the study of proteins, being the basis of the protein folding problem[211] and the resolution of Levinthal's paradox[134] via the folding funnel. These recent developments then say that the bottom of the funnel is itself rich with structure at multiple resolutions and is what ultimately determines the function of the protein. It should be pointed out that in addition to the various timescales of motion, protein motions are highly directional as a result of various interactions between the constituent atoms. Furthermore, any energy landscape represents the distribution of states under specific conditions (pH, temperature etc) and so diagrams such as those in Fig 3.1, whilst helpful, are purely illustrative (though in some cases, by using a scalar coordinate such as the fraction of native contacts, a funnel shape can in fact be seen[182]).

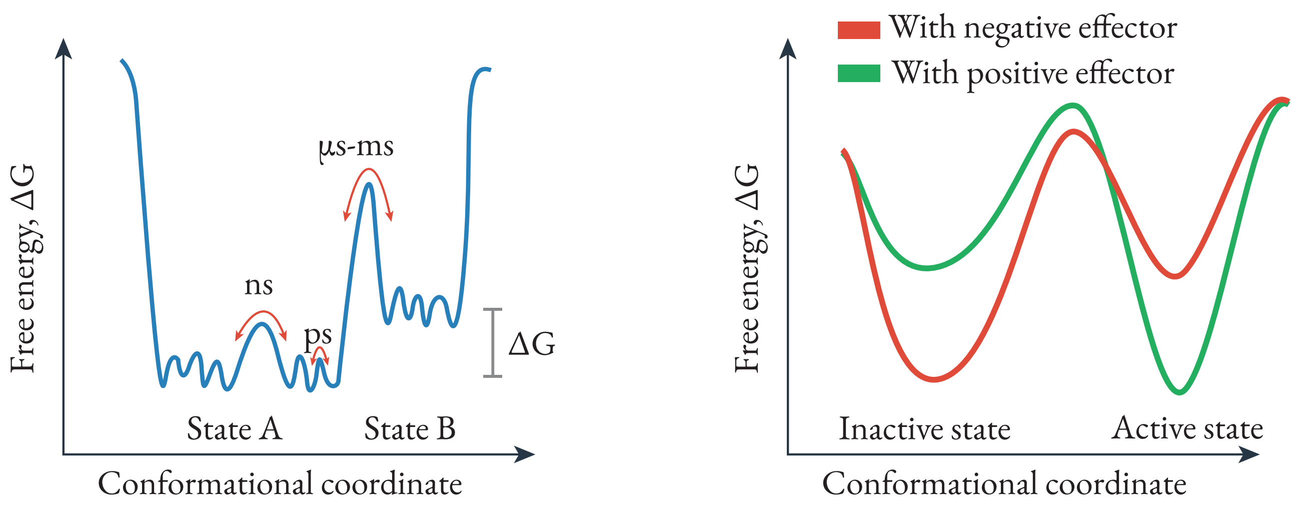


Figure 3.1: Ultimately, what drives protein function is the free energy landscape. Left, we see that protein dynamics occurs on huge range of timescales, spanning at least 10^{12} orders of magnitude, that are coupled together. Allostery can therefore be explained as the preferential stabilisation of the active state (or states) by a ligand so that it is sampled more frequently by the protein ensemble, as shown on the right.

3.3 Allostery

The thermodynamic view

Allostery is the process through which binding of a molecule distal to the active site of a protein causes an attenuation or an enhancement in the catalytic rate of that protein [179, 88, 196]. Despite being a crucial means of regulation within cells (being described as "the second secret of life" by Jacques Monod), the physical mechanisms underpinning this effect are still not well understood at the microscopic level, thus limiting the potential for chemical design and intervention. Most of the previous work on allostery has focused on thermodynamic models linking changes in catalytic rates to modifications in the conformation of the protein. Such an outlook led to the traditional models of allostery: the Monod-Wyman-Changeaux (MWC) model [165], whereby binding of allosteric substrates causes a *concerted* conformational shift of the protein subunits towards the active state, and the Koshland-Nemethy-Filmer (KNF) model [126], which proposed that binding of an allosteric substrate to a subunit drives the latter towards the active state and the overall transition to the full active state is *sequential*.

As noted by Guo and Zhou[88] however, there is often some confusion in the literature on the slightly subtle delineation between comparisons of the MWC and KNF models, and the two alternative allosteric mechanisms of *induced fit* and *conformational selection*, where the MWC and KNF models really concern how multimeric proteins transition between states: each subunit sequentially or all at once. Induced fit[125] describes the idea that the allosteric ligand binds at a protein site, causing a structural change at the binding pocket that destabilises the inactive state and drives a transition to the active state. Conformational selection[30, 146] (sometimes referred to as population shift) posits a pre-existing equilibrium between the active and inactive states, according to the Boltzmann distribution, that the allosteric effector alters in favour of the active state by preferential stabilising it. Thus, induced fit and conformational selection describe the *transition pathways* between inactive and active states. The two mechanisms are not necessarily orthogonal however and both may occur in the same system, meaning discrimination between the two within a catalytic system is not trivial - even the presence of a small proportion of the active state in the absence of ligand is not sufficient to prove that conformational selection dominates[85], particularly in cases where the binding pocket becomes closed in the active state[228]. Allostery must ultimately be a process where binding of a substrate leads to a change in free energy of the ensemble such that it is more favourable for the active state bound to the substrate to exist. More precisely, we must have $\Delta\Delta G = \Delta\Delta H - T\Delta\Delta S < 0$ for the binding of the substrate, where there are two possible contributions to $\Delta\Delta G$. One is that the binding of the allosteric ligand shifts the equilibrium of the active-inactive equilibrium over towards the active state as in Fig. 3.1. Thus the catalytic rate increases simply because there is more active catalyst for the reaction substrate to bind to. The other is that the $\Delta\Delta G$ of the binding of the reaction substrate decreases. We deal mainly with the first case in this thesis, in particular the effect that allosteric ligand binding has on the change in *enthalpy* of the active state, that is the case where ligand binding causes some rearrangement of (weak) bonds within the protein leading to a favourable enthalpy change. We do also, however, deal with the entropic description in Chapter 6 where we consider network structures where two sites are coupled for a normal mode of the protein, as in the "scissor" model of McLeish[159].

More recently, Hilser and coworkers proposed the ensemble allosteric model (EAM) [168], which rationalises allosteric outcomes according to the effect of the substrates on the entire conformational ensemble of the protein. They note that development of explanations for allostery have naturally been constrained by available experimental data, such that traditionally there has been an overemphasis on high resolution, but static X-ray crystal structures. There is now a growing appreciation of the role of dynamics in allostery [237], having initially been proposed by Dryden and Cooper[43] as a means of resolving how allostery can occur without structural change. Often described as a "broadening of the free energy basin"[88] of the active state, the result is that the average conformation of the protein in the active state is unchanged upon ligand addition, so that a solely static picture cannot explain functional changes. The intriguing example given by Cooper and Dryden is that of the binding of an allosteric substrate leading to a "stiffening" of the protein (by "freezing" out the lower frequency modes) and thus an increase in entropy (offset by the decrease in enthalpy from bond formation). Thus subsequent binding of the active site substrate leads to a smaller drop in entropy than there otherwise would be in the absence of the allosteric ligand and thus a *more negative* $\Delta\Delta G_{\text{binding}}$. An idealised structural example of this effect is the "scissor molecule"[159], which has a single degree of freedom allowing for a "scissor" motion with active and allosteric sites positioned at either end of the molecule. Despite the distance between the two sites, because the decrease of entropy occurs via freezing of the normal mode, the effect is *global*, thus linking the two sites.

Further computational methods have been developed to probe the role of entropy in allostery: Kalescky *et al*[110] performed *in-silico* mutatgenisis, scanning across all residues in molecular dynamics simulations the PDZ domain of human PTP1E protein. By applying rigid body constraints to each residue in turn, they were able to ascertain the contribution each residue made to the internal degrees of freedom - important residues were then those that showed a smaller entropy difference between unbound and bound states than the difference in the Wildtype case.

The importance of entropy was recently confirmed by experimentally by the design of protein switches [40], whereby a flexible linker was introduced between a effector binding domain and a catalytic domain. Initially, the isolated enzyme domain was dominated by the active state even in the absence of ligands, thus exhibiting no allostery. When the flexible linker was introduced, the additional conformational entropy reduced the effect of entropy, "flattening out" the energy landscape and allowing the ligand to then exhibit an allosteric effect.

Notably, the EAM does not require the existence of structural pathways between the allosteric sites and active sites within the protein (though it does not necessarily rule it out either). In particular, the discovery that intrinsically disordered proteins (IDPs)[244] exhibit allostery has generated particular intrigue, given their lack of the fixed tertiary structure that would seem necessary to transmit energetic changes between distal sites. An example is the Phd/Doc toxin-antitoxin system, which is an inhibitor of the ribosome A site and exhibits coupling between two regulatory sites and a DNA binding domain[79] that binds at PhD's own operon. As the concentration of Doc increases, it first acts to inhibit PhD before then reactivating transcription, in a process termed "conditional cooperativity".

The structural view

Whilst thermodynamic models of allostery provide understanding of the equilibrium effects of substrate binding, they are unable to provide a detailed description of how a signal may be transmitted between the allosteric binding site and the active site at the microscopic scale. The so called *structural view* of allostery posits that some form of propagation pathway between the allosteric and active sites exists as a condition for allostery, though the existence of such a pathway does not imply allosteric behaviour by itself. Tsai and Nussinov [235] argue that both a structural and a thermodynamic component is required for a complete description of allostery. The idea that structural pathways might exist in proteins began when Perutz[188] analysed the differences between haemoglobin with and without oxygen bound, identifying an important salt bridge in the inactive "T" state. Later, Szabo and Karplus[229] were able to incorporate this information, in addition to dependence on pH, into a quantitative model, marking the first attempt to make use of structural information to explain allostery.

The viability of long range transfer in proteins is often questioned, particularly when initiated by what appears to be very small perturbations at the allosteric site. However, as noted by Yu and Koshland[247], the remarkable specificity of enzymes should lead one to be less surprised that a 1Å change at a site can result in significant modulation, especially when one considered the constraint that changes are limited by the binding energy of the ligand. Furthermore, the authors point out that while earlier forms of proteins may have shown poorer communication, allosteric changes must have been optimized by evolution, bringing to mind the adage that "Nothing in biology makes sense except in the light of evolution"[58].

The nature of this energy propagation is usually presumed to occur via the propagation of strain, that is generated at the allosteric site and travels through the protein towards the active site[53]. Leitner notes [132] that there are two alternate descriptions of this energy transfer: the traversing of energy from one residue to another along structural pathways (often utilised in discussions of energy dispersion after photoexcitation [127, 27]) or energy transfer between the normal modes of the protein [91, 118]. These two possibilities have been creatively described as the "domino model" and the "violin model"[124]. The domino model has been suggested to act as the propagation mechanism in PDK[142] and haemoglobin[143] based on conservation of residues forming a pathway within the protein. In the violin model, binding of a small molecule has an appreciable effect on one or more of the modes of the protein, much in the same way the pitch of a violin is altered by placing a finger at a certain point on the string. The authors argue that the domino model makes sense for much larger scale mechanical structures, at the scale of everyday life, but that at the molecular scale, thermal noise would lead to inadvertent activation of the "domino pathways", whilst the violin model

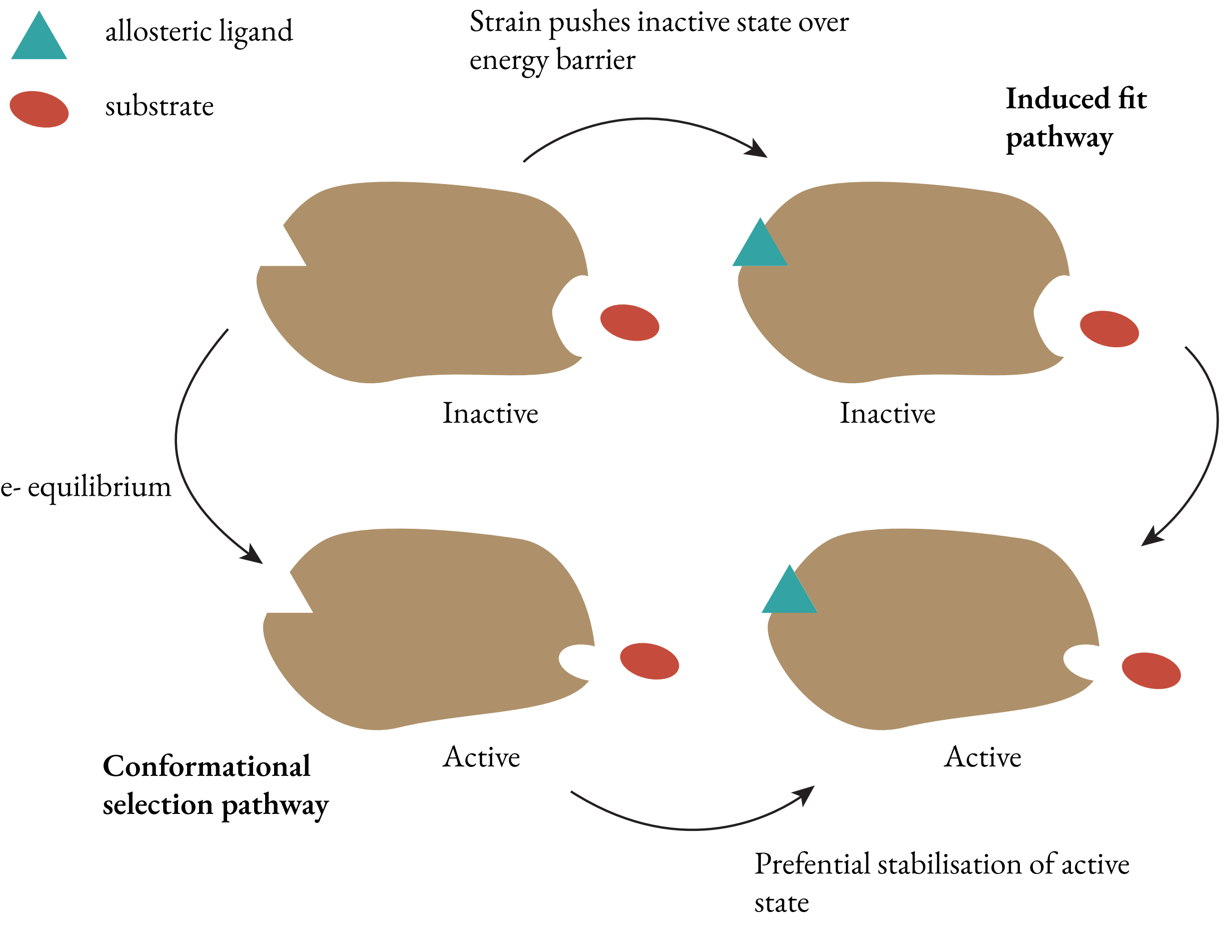

allosteric ligand
substrate
Strain pushes inactive state over energy barrier
Induced fit pathway
Inactive
Inactive
Pre- equilibrium
Conformational selection pathway
Active
Active
Prefential stabilisation of active state

Figure 3.2: The two proposed transition pathways for allostery: induced fit moving clockwise from the top left and conformational selection moving anticlockwise. The two alternative are not in fact orthogonal and in fact, induced fit may really be seen as a special case of conformational selection, as even if ligand binding drives the inactive state over the energy barrier, it must still also preferentially stabilise the active state.

actively relies on this equilibrium motion of the protein structure. Similarly, McLeish *et al*[158] argue that small structural changes tend to be localised by elastic inhomogeneities within the protein, as a particular example of Anderson localisation[205].

Energy flow is sometimes rationalised using concepts from *percolation theory*[193], in particular the percolation of node vibrations on a fractal object, which Alexander and Orbach found to be:

$$R^2 \sim t^{\alpha}, \quad \alpha = \bar{d}/D \tag{3.1}$$

with R^2 the mean square displacement, D the fractal dimension and $\bar{d}$ the spectral dimension. The fractal dimension results from the relationship (with mass M and length L): $M \sim L^D$, and D has been found to be roughly 2.54 for proteins[63, 8]. This value is very close to that of a percolation cluster in three dimensions[170], suggesting proteins may well be folded in such a way close to the percolation threshold so as to optimize vibrational flow through specific channels.

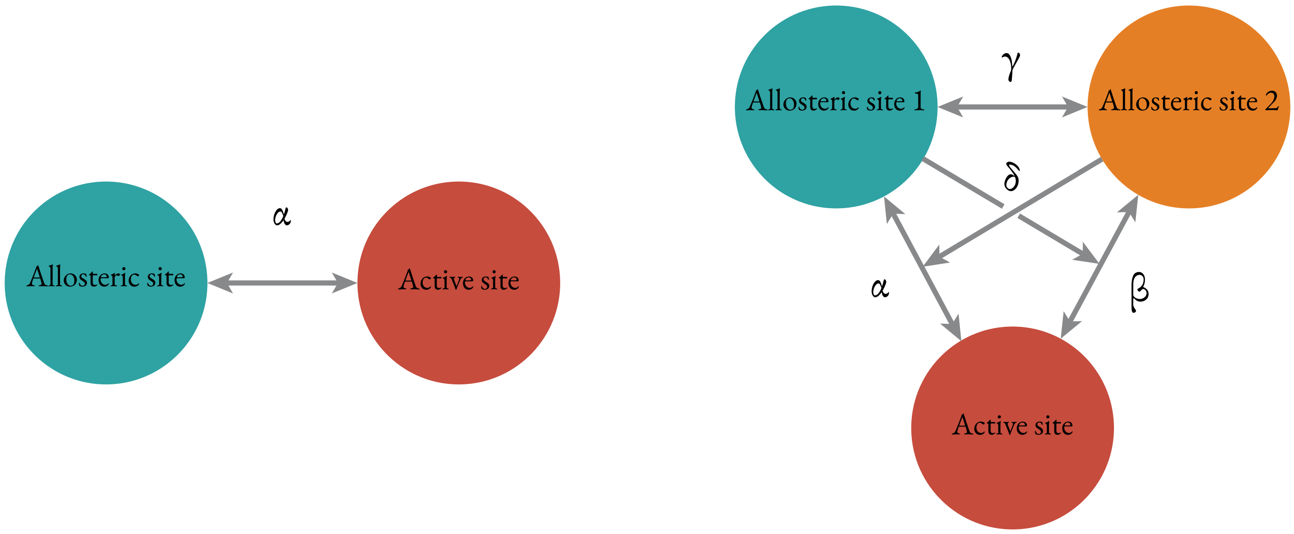


Figure 3.3: In the unified view, a coupling constant α is introduced that quantifies how stabilising the effect of ligand binding is at the active site. On the right, the idea can be extended to explain cooperative effects between allosteric sites, such that binding of a second ligand has a direct effect on the communication between the first ligand and the active, labelled as δ.

The unified view, of Tsai and Nussinov, fuses together the structural and thermodynamic explanations by considering allostery as a function of the energy landscape. If we imagine a protein consisting of a single regulatory and a single catalytic subunit, then the key parameter is the *allosteric efficacy*, α. The allosteric efficacy determines the differential stabilisation of the catalytic subunit between a ligand binding when the catalytic subunit is in either the active or inactive states. A positive effector is then one that preferentially stabilises the active state. In this framework, a pathway between the allosteric binding site and the catalytic subunit becomes a necessary but not sufficient condition for allostery; that is the pathway of residues itself is not what determines whether the protein is allosteric but instead how energetically favourable changes induced by the ligand are in the active state of the catalytic subunit relative to the inactive state. Thus a ligand may well stabilise the inactive state but so long as it stabilises the active state to a greater extent, it will have a positive allosteric effect.

3.4 Experimental approaches

Despite significant advances in single molecule experimental techniques, the study of energy propagation in individual proteins is far from trivial. Dyer *et al* [135] used ultrafast infrared spectroscopy to examine the flow of energy within albumin and found that the flow was ballistic and anisotropic rather than diffusive, supporting the idea that structural pathways exist within proteins that allow for efficient energy transfer between coupled sites. In fact, anisotropic energy flow appears to be common within proteins across a number of processes in addition to allostery, including diversion of energy from heated cofactors[127], photosensing[128] and photodissociation of ligands[203] as measured by Raman spectroscopy. There is evidence that the existence of such energy channels is essential in the preservation of the protein's folded state; substituting Zn into cytochrome c and exciting with ultraviolet light appears to lead to partial unfolding of the protein[131], suggesting proteins may use energy pathways to shunt away excess heat.

One of the challenges involved in experimental studies of allostery is that structural changes upon ligand binding can be subtle. A technique that is capable of measuring structural changes at an atomic level to an exquisite level of sensitivity is *nuclear magnetic resonance spectroscopy* (NMR)[201] and as a result, NMR has become a vital tool for probing allosteric changes. Additionally, NMR is suited to elucidating dynamics at a range of timescales[74], from the very fast *ps* changes implicated in entropically driven dynamic allostery[96, 104, 74] to slow *ms* timescales, close to to the timescale over which allostery is generally thought to take place. Allostery is often marked by very small structural changes, yet Falk *et al* [66] were able to exploit the high sensitivity of NMR chemical shifts to small structural changes and by using mutational studies to create singly bound thymidylate synthase dimers, they demonstrated that binding of the first allosteric effector *primes* the enzyme for the binding of the second effector, such that both effectors are required for the allosteric response. In an earlier study, Volkman *et al*[237] used NMR relaxation to study allostery in the signaling protein NtrC, a single domain protein. By observing the protein before and after phosphorylation, the authors discovered NtrC undergoes a population shift with a dynamical equilibrium between the two structures even in the absence of phosphate.

Studies on CheY cast doubt on a previously assumed[208] two state model of allostery. CheY is involved in the regulation of the flagella motors in *E. coli*, which drive bacterial motion[41]. Phosphorylation of CheY causes a conformational change at the active site that binds to the motor switch protein FLIM. Carr-Purcell-Meiboom-Gill (CPMG) dispersion experiments were used to follow the dynamics switching of the unphosphorylated form of CheY[156], finding that a two state concerted switch is not the mode of transition, but instead a pathway of residues switching asynchronously leads to the change between states. In a later study by the same authors[157], phosphorylation driven changes of the dynamics in the $ps-ns$ range occurred mainly within this set of residues, hinting at specific structural alterations modulating entropic contributions to allostery.

3.5 Computational approaches

Due to the challenges inherent to the experimental studies of allostery, a wide range of computational methods have been developed to model allosteric behaviour [196]. Whilst it is ultimately in experiments that we must put our trust, the freeing of practical constraints means that computational methods for studying protein dynamics have become immensely powerful, aided by the staggering growth in computer power in the last 50 years. There is currently no experimental method that can provide atomic level resolution for the dynamics of even moderately sized proteins and so it falls to computation methods to attempt to provide the detail at this level.

Molecular dynamics

Perhaps the gold standard in the computational study of protein dynamics, molecular dynamics (MD) works on the principle of numerically integrating Newton's equations of motions over the desired timescale given a defined *force field*[76, 194]. There are a number of force fields, including CHARMM[25], AMBER[45] and GROMACS[236] but the functional form is largely similar:

$$\begin{aligned} E_{\text{total}} = \sum_{\text{bonds}} \left(r - r_{eq}\right)^2 + \sum_{\text{angles}} K_\vartheta \left(\vartheta - \vartheta_{eq}\right)^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} \left[1 + \cos\left(n\phi - \gamma\right)\right] \\ + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} - \frac{q_i q_j}{R_{ij}}\right] \end{aligned} \tag{3.2}$$

That is, the force field is *classical*, though quantum mechanical computations are used to calculate the force parameters. The suitability of MD for studying protein ensembles derives from the *ergodic principle*, the same principle discussed in Chapter 2 in the context of random walks on networks, which in effect says that the *time averaged* properties of a single protein (over a long enough period) are the same as the snapshot *ensemble average* of the protein.

The history of MD[1] dates back to 1957, when small simulations on hard-sphere models were performed[3], whilst modern day methods are capable of reaching millisecond timescales for some proteins[215]. However even now, studying biological phenomena (often on the timescale of milliseconds, if not even seconds) is out of reach for moderately sized proteins, necessitating the use of various techniques to reduce the issue of structures becoming "stuck" in local energy minima. Specifically, the process of interest may be very fast, but infrequent. A constraint on how quickly MD simulations can explore the state space for a protein is that the time step for the integration algorithm must be shorter than the fastest process occurring within the protein, which is usually hydrogen bond vibrations that take place on the order of 10^{-15} seconds. Activated MD[154] first restricts a series of trajectories to be near the energy barrier between two states of interest in order to identify the barrier, before then running further conventional MD trajectories in the barrier's vicinity. Steered MD[97] uses additional forces to encourage trajectories along paths of particular interest, though this then imposes difficulties when applied to the study of equilibrium distributions. *Langevin dynamics* can be used to model the effect on motion of solvent:

$$M\frac{d^2r}{dt^2} = \Gamma\frac{dr}{dt} + F(r) + \phi(t)$$

where Γ is a diagonal matrix representing a diffusion term and $\phi(t)$ is $3N \times 1$ vector of Gaussian noise that acts as a heat bath, with $F(r)$ our internal force interactions and M the mass matrix. The random perturbations to the system can aid in crossing energy barriers, improving sampling[1]. When the viscosity of the surrounding the solvent is particularly high (Γ » M), we say the system is *overdamped* and set the acceleration term to zero:

$$\frac{dr}{dt} = -F(r) + \phi(t) \tag{3.3}$$

where we have also set Γ to 1. Equation (3.3) now describes *brownian dynamics* (BD), the diffusional counterpart of MD[64]. The damping of the solvent thus removes inertial effects and the protein performs a random walk over the state space determined by the force field. In the special case where the system is a 1-dimensional line of springs, such that the internal force terms are simply a set of spring constants, we can write:

$$\frac{dx}{dt} = -Lx + \phi(t) \tag{3.4}$$

and can see that we have obtained an equation of the same form as Eq. (2.10) from Chapter 2 (with the addition of the noise term), showing the correspondence between a random walk on a network and the scalar vibrations of an overdamped spring system, where the removal of inertia guarantees the *memorylessness* property.

MD simulations have found agreement with evolutionary conservation analysis; Ota and Agard[183] used a nonequilibrium MD method they called anisotropic thermal diffusion to simulate the protein PSD-95, a member of the PDZ family. By reducing the signal-to-noise ratio, the authors were able to discern long range dynamical correlations, which agreed closely with the residues highlighted by Lockless *et al*[142]. A subsequent study[212] on the same protein used a pump-probe MD method to identify energy transport channels and once again the same pathway was found, giving a physical basis to the evolutionary analysis. Ranganathan *et al* subsequently performed sequence analysis on other allosteric proteins[218, 227], including G-protein coupled receptors and haemoglobin and again discovered connected sequences of residues between distal functional sites.

Metadynamics is an MD technique that incorporates trajectory history into the force field, allowing it to "fill up" explored energy minima and more efficient explore the state space. Metadynamics was used by Palazzesi *et al*[185] to study binding of MLL to the KIX domain of CREB-binding protein (CBP). Previous NMR studies[26] had suggested that binding of MLL primes the KIX for binding of a second ligand, c-Myb, through a pathway linking the two binding sites. In the MD experiments, the MLL interaction with the $L_{12} - G_2$ region is transmitted through an $\alpha-$helix to a hydrophobic core, allowing the residue Ile657 to rotate into a favourable position for binding of the second ligand.

Markov State Models

Despite the rapid advances in processing power of modern computers, molecular dynamics simulations on even moderately sized proteins are time consuming, particularly if explicit solvent is used. The assembly of specialised hardware has allowed for trajectories running well into the millisecond timescale[213], but is currently prohibitively expensive. A more general approach however, is provided by Markov State Models (MSMs)[186, 39]. In an MSM, after a trajectory is collected, a set of *features* is selected, for example the set of dihedral angles. A dimensionality reduction step projects each of the snapshots of the trajectory into a reduced space using a method called *time-lagged independent component analysis* (tICA)[164] and snapshots that are close together within this space are clustered together to form a set of *microstates*, such that the dynamics within a microstate is fast enough that they may be considered a single conformation. Each of the microstates is then a node in a network. The edge weights can then be found empirically by counting the number of times the trajectory moves between each of the states, ultimately generating a discretized approximation to the free energy landscape of the protein. The power of MSMs is that they can also be built using a number of shorter MD runs, in order to more effectively sample the state space and avoid the problem of remaining trapped inside an initial energy well.

MSMs have been used to uncover "cryptic allosteric sites", that is functionally relevant sites that are unamenable to conventional methods[22]. MSMs were constructed for three proteins (TEM-1 β-lactamase, interleukin-2 (IL-2), and RNase H) in their apo, or unbound, states. By applying a pocket detection algorithm to a representative structure from each of the approximately 5000 microstates in their MSM, they were able to not only identify cryptic sites that appear infrequently, but to quantitatively assign a probability of the site being open. Furthermore, when the authors applied a clustering technique based on the mutual information between rotameric states of pairs of amino acids in $\beta-$lactamase, they found that the allosteric pocket and the active site were gathered into the same community. Malmstrom *et al*[149] built an MSM from both the active and inactive states of the cyclic nucleotide-binding domain of the regulatory subunit of protein kinase A.

Markov chain Monte Carlo

Famously introduced by Metropolis *et al*[161] having been initially devised by Stanislaw Ulum during the Manhattan project, Markov chain Monte Carlo (MCMC) based methods are an alternative approach to uncovering the potential energy landscape via sequentially generating configurations of the protein. By defining a suitable *acceptance rule*, it can be guaranteed that the set of samples converges to the correct (Boltzmann) distribution (though there are not necessarily any bounds on how long this may take). Using π as the (unnormalised) probability of a particular configuration at equilibrium:

$$\text{accept new configuration} = \begin{cases} 1, & \text{if} \pi_{\text{new}} > \pi_{\text{old}} \\ \frac{\pi_{\text{new}}}{\pi_{\text{old}}}, & \text{if} \pi_{\text{new}} < \pi_{\text{old}} \end{cases} \tag{3.5}$$

which is known as the Metropolis method and is a special case of a more general set of choices that must satisfy *detailed balance*. In particular, a choice must be made about how to generate the moves in the first place (the "Markov chain" part of MCMC) and if the choice is *symmetrical* (so that the probability of *choosing* a move from i to j is the same as from j to i) then the above equation holds. A fundamental issue, however, of MCMC methods is the difficulty of designing them to efficiently explore the state space: if the moves are too large, then there is only a very small probability that the move will be accepted, whilst generating only small moves will tend to lead to a random walk that takes a very long term to converge to the stationary distribution. These issues are a particular problem for large biomolecules[178].

A powerful combination of MD and MCMC methods is transition path sampling (TPS)[20], that efficiently samples trajectories between two pre-defined states. A set of trajectories between the two sites is generated using MD before using Monte Carlo techniques to accept or reject each of the paths, generating an ensemble of transitions with their associated probabilities. TPS was used to study the inactive-active transtion in CheY, thought to occur via a mechanism whereby a Thr87 residue moves away from the active upon phosphorylation, then allowing Tyr106 to occupy its active conformation inside the active site. Ma and Cui[147] examined the rotation of the Tyr106 residue, finding that rather than being dependent on the movement of Thr87, it is actually the formation of a hydrogen bond between the two residues that stabilises the active conformation. The finding highlights the necessity of an atomistic level of resolution to elucidate transition pathways, with individual bonds often having crucial effects on function.

Network models

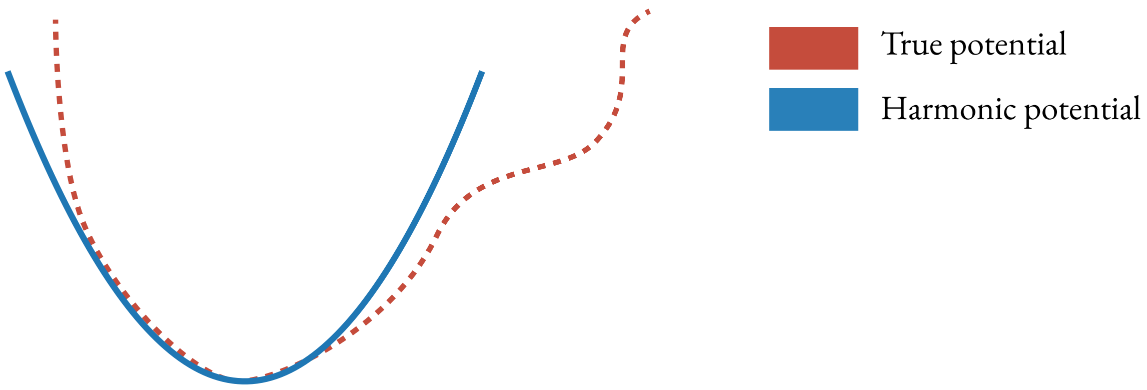


Figure 3.4: For a system at its energy minimum, a common approach is to model the potential energy surface as a quadratic. In many cases, this is a reasonable assumption under the condition that the system's displacement is small relative to its equilibrium position. In other words, a possibly complex interaction is modelled as a Hooke spring, which is exactly solvable.

Given the high computational demands of MD methods, there has been much attention paid to the development of simplified models of proteins that attempt to retain important characteristics of dynamics. Perhaps the most successful of these have been *elastic network models* (ENMs), in which the force field in Eq. (3.2) is replaced with a set of Hooke springs between pairs of atoms, usually determined by a distance cutoff. The seminal work in this area was carried out by Tirion[233] who used an atomistic ENM to perform *normal mode analysis* (NMA) on G-actin, a muscle protein. By Taylor expanding the potential energy surface about a minimum point with respect to the atom displacements:

$$V(r - r_0) = V_0 + J(r - r_0) + \frac{1}{2}(r - r_0)^T H(r - r_0) + O((r - r_0)^3) \tag{3.6}$$

and using the fact that at the bottom of the energy well, all the first derivatives must be zero (i.e. $J = 0$ where J is the Jacobian matrix) and the constant term V_0 is arbitrary as we are only interested in potential energy *changes* and so can be set to zero. All terms of order $(r - r_0)^3$ are then dropped so that $V(r - r_0) = \frac{1}{2}(r - r_0)^T H(r - r_0)$, with the Hessian, H, the matrix of mixed second order partial derivatives for all pairs of atoms. The normal modes then are those motions of the structure whereby the displacements line up with the internal forces. Using that force, F, is the derivative of potential energy $\frac{dV}{d(r-r_0)} = H(r - r_0)$, we wish to find those displacement that are in the same direction as the forces, or:

$$Hr_i = \lambda_i M r_i \tag{3.7}$$

so that the problem reduces to a generalised eigenvalue problem involving the Hessian matrix and the mass matrix M. Tirion found that despite the relative simplicity of the elastic model, the slowest frequency modes were in fact a good match to MD simulations of the same protein[234], an observation that has been repeated in numerous additional studies[13, 61].

NMA has been frequently used in techniques that aim to model allosteric binding. Balabin *et al*[15] modelled the perturbation induced by the binding of an allosteric ligand to an elastic network and calculated the effect on the motion along a normal mode. They were then able to construct a coupling matrix that describes those parts of the protein that show strong coupling to other sites. Mitternacht and Berezovsky[163] defined functional sites as those that undergo high strain when bound to a ligand, in the sense that the surrounding residues are moving in different directions, using a measure they termed 'binding leverage'. These sites thus allow coupling to a number of modes to modulate dynamics. Another method used a technique that defined important residues as those whose perturbation caused large changes to conformation[11]. In fact, there are now an extensive number of web servers that make use of normal mode calculations to predict allosteric sites on proteins, including Allosite[95], Allopred[84] and SPACER[82], aided by the speed of eigenvalue calculations on modern workstations.

ENMs have also been combined with MD simulations as a hybrid approach. Gur *et al*[89] calculated the normal modes of adenylate kinase to determine potential transition pathways between states, while simultaneously using an MD procedure to determine the energetics of the dynamics. Ligand induced large scale conformational change was investigated by Wang *et al* using umbrella sampling, a form of MD that aids in sampling parts of the energy landscape separated by large barriers. Again, NMA was used to generate plausible transitions between states. The authors found that across three proteins (adenylate kinase, calmodulin and p38α), a conformational selection mechanism was seen for the first two, but an induced fit appeared to be the means of transition in the third, once more highlighting that transition pathways are likely to be tuned for specific proteins, rather than following rigid, general principles.

Other network based methods

Network based approaches have become increasingly common, aided by the suite of techniques developed by the growing field of network theory for the general study of network topology. Daily *et al*[49] identified those residues whose interactions change upon binding by modelling proteins as residue-residue contact networks (that is, $N \times N$ graphs as distinct from $3N \times 3N$ elastic networks), finding that in 15 allosteric structures, a set of residues exhibiting large contact rearrangement formed a contiguous path between the allosteric and active sites, though no such path was present in the remaining 10 proteins. Del Sol et al[52] borrowed shortest path techniques from graph theory to find those residues that are particularly crucial to signalling within allosteric proteins finding that conserved residues were particularly important in maintaining a low characteristic path length (the average of the number of steps of the shortest paths between all pairs of nodes in the network) in these proteins. Ribeiro and Ortiz used MD trajectories to build up so called protein energy networks (PENs)[195]. In a separate paper, they discovered that when residue motion correlations are used to create the network, statistical errors render the results less accurate than when interaction energies are used, as a result of the high sensitivity of the signalling pathways to the network topology[?]. Network-theoretic machine learning tools have also been applied to fully atomistic protein graphs [54, 4] demonstrating that a wealth of information can be obtained from static structures, avoiding the time consuming calculations often involved in molecular dynamics or Monte Carlo approaches.

Chapter 4

Bond-to-bond propensity analysis of ATCase

4.1 Aspartate carbamoyltransferase: a model for allostery in multimeric proteins

Aspartate carbamoyltransferase, or ATCase, is a classic example of an allosteric enzyme whose catalytic rate is attenuated by the binding of various substrates, and has been the subject of intense study for over 50 years[33, 80]. Biologically, its role is to catalyse the initial step of the pyrimidine biosynthesis pathway; that is the conversion of L-aspartate and carbamoyl phosphate to N-carbamyl-L-aspartate and phosphate. Adopting a dodecameric structure, ATCase consists of six catalytic subunits and six regulatory subunits and notably does not follow Michaelis-Menten kinetics, as observed by the absence of a hyperbolic saturation curve where increases in rate slow down with increasing substrate concentration as the enzyme saturates. Instead, the enzyme exists in two distinct states: a biologically active "relaxed" state (or R state) and an inactive "tense" state (T state) that exist in a dynamic equilibrium [139], resulting in a sigmoidal curve, in which the rate accelerates with higher ligand concentration before flattening out at saturation[?].

ATCase displays both *homotropic* and *heterotropic* allostery. Binding of the reaction substrates to one of the active sites (in the multimeric structure) leads to a shift in equilibrium towards the active R state - we call this homotropic allostery. Heterotropic allostery then refers to the binding of ligands distinct from the reaction substrates to the protein, usually at a distal site, that cause a change in the reaction rate[165]. Both ATP (positive effector) and CTP (negative effector) bind to ATCase and it is this phenomenon that drives a negative feedback mechanism. CTP is a pyrimidine and thus high levels of pyrimidine biosythesis generates

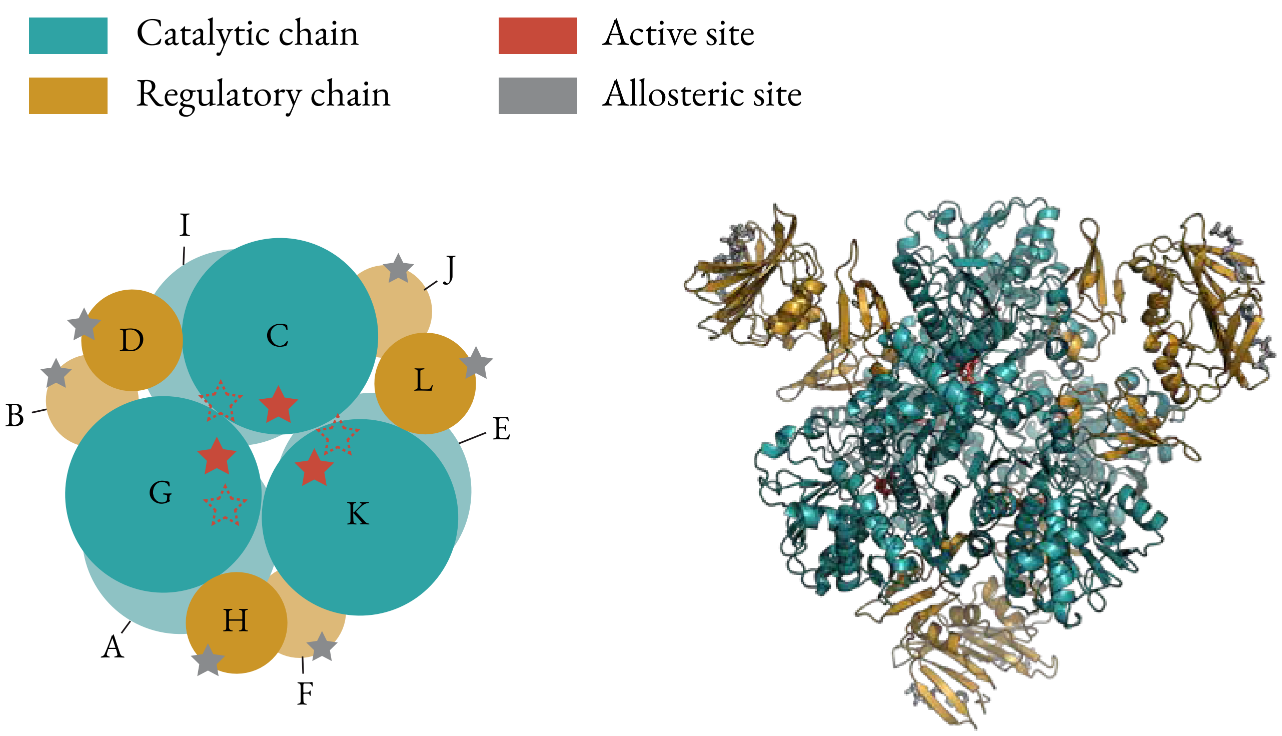


Figure 4.1: ATCase comprises of six catalytic and six regulatory subunits, shown in green and gold respectively, with more than 43000 atoms. PALA (red) is a bisubstrate analogue of the reaction substrates (carbamoyl phosphate and aspartate) and sits in the active site, while ATP and CTP bind to the regulatory subunits and are shown in silver.

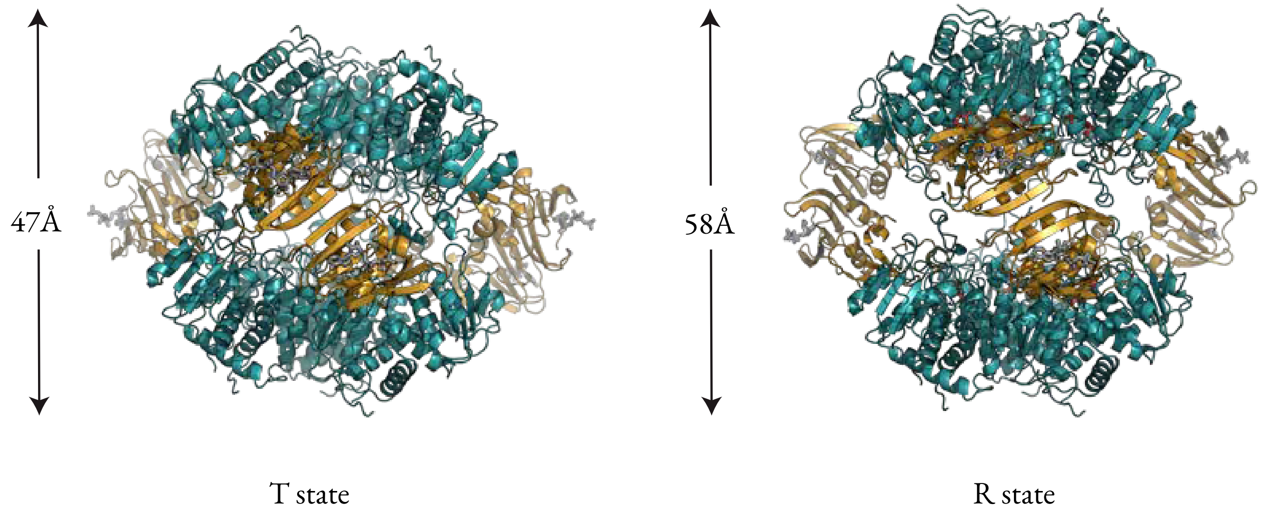


Figure 4.2: The inactive "tense" (T) state and active "relaxed" (R) states of ATCase. ATCase expands by 11Å along its 3-fold axis upon transitioning to the R state.

higher concentrations of CTP, that in turn attenuate the catalytic rate. Whilst both ATP and CTP bind to both the active and inactive state of ATCase and cause slight changes in the quaternary structure, the binding of ATP to the inactive T state and CTP to the active R state is not sufficient to cause a population shift to the opposite state[111].

4.2 Application of bond-to-bond propensity

Bond-to-bond propensities was shown by Amor *et al*[5] to be able to predict *allosteric sites* in a large range of proteins through knowledge only of the active sites of those proteins. A number of features of the method stands out. Firstly, it uses an atomistic description of the protein when constructing the network so does not use any coarse-graining techniques to reduce the complexity of the protein structure. Despite this, the method is very computationally efficient; the calculations are carried out in almost linear time with respect to the number of edges as a result of recent work in algorithmic theory[220, 116]. Finally, in contrast to many graph theoretical approaches, bond-to-bond propensities is focused on the edges in a network, and thus in a biological system, the bonds. It is through bonds that energy transfer occurs in a protein upon binding of an allosteric ligand[?] and this appears to be the vital link between the mathematical basis of the method and the physical processes occurring in the protein.

The success of the method in identifying these sites motivates this chapter, in which we study the "reverse" process. That is, using ligands bound to the allosteric site on the protein as the source of a perturbation so as to replicate the actual physical process that occurs. From this, we are able to examine how that perturbation spreads throughout the protein structure and identify those residues that are particularly crucial to energy transport. By comparing the process on both the active and inactive states of ATCase, we aim to then explain how the different energy propagation processes may affect the equilibrium between the two states and thus the allosteric effect of altering the catalytic rate of the protein.

The formulation of bond-to-bond propensity was presented in detail in Ref [5] and thus is summarised here. The key matrix that defines bond-to-bond propensities is M , the $m \times m$ bond-to-bond transfer matrix, where m is the number of bonds. The element M_{ji} describes how a perturbation at bond i is transmitted to bond j via a propagation that includes the entire graph structure [206]. M is shown to be given by

$$M = \frac{1}{2} W B^T L^\dagger B \tag{4.1}$$

where B is the $n \times m$ incidence matrix for the graph with n nodes and m edges and $L^\dagger$ is the pseudo-inverse of the weighted Laplacian matrix L, which governs the diffusion dynamics on the energy-weighted graph [130]. The weighted Laplacian is given by:

$$L = \begin{cases} -\omega_{ij}, & i \neq j. \\ \sum_j \omega_{ij}, & i = j, \end{cases} \tag{4.2}$$

where ω_{ij} corresponds to the interaction energy between atoms i and j. More compactly, the Laplacian can be rewritten as $L = BWB^T$ where $W = \mathrm{diag}(\omega_{ij})$ is a $m \times m$ diagonal matrix that contains the average fluctuation energy of interactions of all edges on the diagonal: $M_{bb} = \frac{1}{2}\langle w_b y_b y_b \rangle = \frac{1}{2}\langle w_b \left(x_{\mathrm{head(b)}} - x_{\mathrm{tail(b)}}\right)^2 \rangle$.

To evaluate the effect of perturbations from a group of bonds b' (e.g., belonging to a ligand) on another bond b we select the corresponding columns of the matrix M and compute the sum of the absolute values in the b^{th} row of the selected columns:

$$\Pi_b^{\mathrm{raw}} = \sum_{b' \in \mathrm{ligand}} |M_{bb'}| \tag{4.3}$$

where b' includes all the weak bonds between the protein and the source (i.e., the ligand).

The *bond propensity* is then defined as:

$$\Pi_b = \frac{\Pi_b^{\mathrm{raw}}}{\sum_b \Pi_b^{\mathrm{raw}}}, \tag{4.4}$$

which is normalised by the total propensity score of all the bonds in the system.

The results presented in this chapter are often in the form of the *residue propensity*, which is calculated by summing over the normalised bond propensities of the bonds belonging to the residue R:

$$\Pi_R = \sum_{b \in R} \Pi_b. \tag{4.5}$$

Quantile regression

In general, the propensity of a bond within the protein decays away from the perturbation source. To detect significant effects in the protein structure, we need to compare bond propensities at a similar distance from the source, thus taking into account the expected effect of distance. This is achieved using *conditional quantile regression* (QR) [123], which allows us to identify high propensity bonds at the tail of the highly non-normal distribution [5].

The distance of a bond from the perturbation source is taken to be the minimum distance between that bond b and any of the bonds of the chosen source residues:

$$d_b = \min_{b' \in \text{source bonds}} |x_b - x_{b'}| \,, \tag{4.6}$$

where x_b holds the cartesian coordinates of the midpoint of bond b. Because propensity scores are seen to generally fall away exponentially with distance, the logarithm of the propensity is used to generate the parameters in the QR minimisation problem:

$$\hat{\beta}_b^{\text{prot}}(p) = \underset{(\beta_{b,0},\, \beta_{b,1})}{\operatorname{argmin}} \sum_{b}^{\text{protein}} \varrho_p \left(\log(\Pi_b) - (\beta_{b,0} + \beta_{b,1} d_b) \right) \tag{4.7}$$

where

$$\varrho_p(y) = \left| y \left(p - \mathbb{1}_{y<0} \right) \right| \tag{4.8}$$

is the QR loss function to be minimised for each quantile p and $\mathbb{1}$ denotes the indicator function. The result of this optimisation is the model $\hat{\beta}^{\text{prot}} = \left(\hat{\beta}_{b,0}^{\text{prot}}(p), \hat{\beta}_{b,1}^{\text{prot}}(p) \right)$ that describes the quantiles of the propensities for all bonds in the protein. In continuum elastic models, the response to an external perturbation scales as r^{1-d} with r the distance from the perturbation source and d the spatial dimension[133] but we see empirically across all the proteins studied thus far that the propensity values decay rapidly with distance from the source[5], a result concomitant with observations in granular materials where local response is heterogeneous and dependent on structure[140, 148]. Thus the more quickly decaying exponential function is a more suitable fit here.

The *bond quantile score* can then be calculated for each bond in the protein by finding the quantile ϱ_p such that:

$$p_b = \underset{p\in[0,1]}{\operatorname{argmin}} \left|\log(\Pi_b) - \left(\hat{\beta}^{\text{prot}}_{b,0}(p) + \hat{\beta}^{\text{prot}}_{b,1}(p)\, d_b\right)\right| \tag{4.9}$$

for bond b with propensity Π_b at a distance d_b from the source bonds. The corresponding *residue quantile score* (p_R) is similarly defined, instead using residue propensities and the minimum distance between the atoms of each residue and those of the source bonds:

$$\hat{\beta}^{\text{prot}}_{R}(p) = \underset{(\beta_{R,0},\ \beta_{R,1})}{\operatorname{argmin}} \sum_{R}^{\text{protein}} \varrho_p \left(\log(\Pi_R) - (\beta_{R,0} + \beta_{R,1} d_R)\right) \tag{4.10}$$

and

$$p_R = \underset{p\in[0,1]}{\operatorname{argmin}} \left|\log(\Pi_R) - \left(\hat{\beta}^{\text{prot}}_{R,0}(p) + \hat{\beta}^{\text{prot}}_{R,1}(p) d_R\right)\right| \tag{4.11}$$

We can then use this bond quantile score (and its corresponding residue analogue p_R) to establish which bonds (and residues) have significantly propensities once the distance effect has been regressed out. Our quantile regression calculations make use of the R library *quantreg* written by R. Koenker [122]. QR is discussed in further detail in the Appendix.

4.3 Structural data

The three X-ray crystal structures of ATCase used in this work were downloaded from the Protein Data Bank (PDB) [17]. We studied two active state structures: 4KGV, the R state bound to ATP (obtained at 1.2Å resolution [42]); and 1D09, the unligated active state (resolved at 2.1Å [107]). We also used one inactive structure: 5AT1, the T state bound to CTP (obtained at 2.6Å resolution [224]).

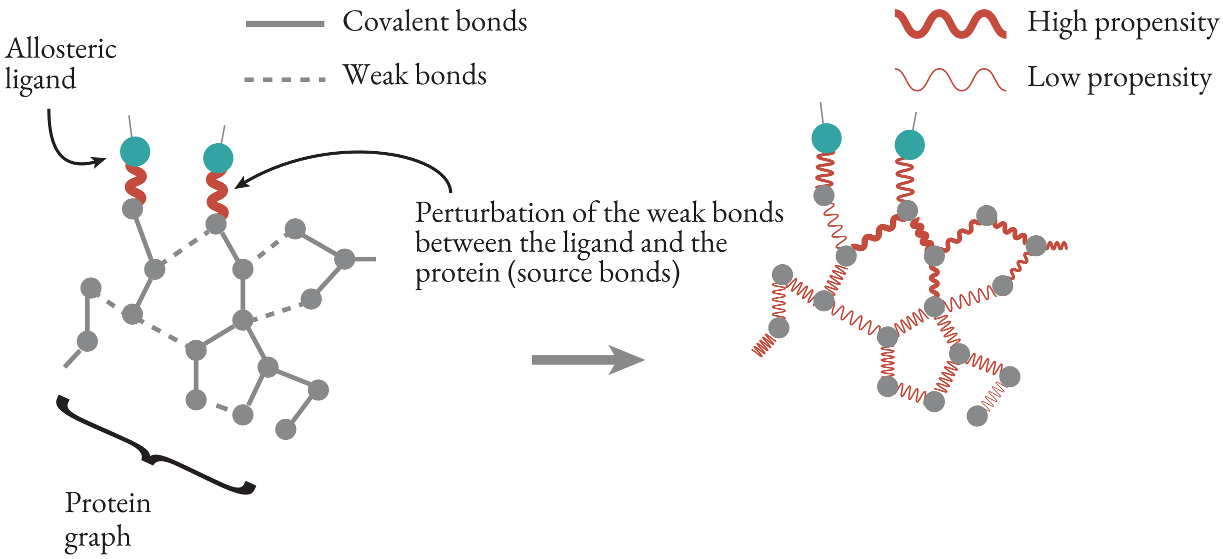


Figure 4.3: A perturbation is applied to the weak bonds between the allosteric ligand and the protein and the reponse across the entire set of bonds in the protein is calculated. Bonds that are more strongly coupled to the allosteric site exhibit a stronger response and are thus said to have a higher *propensity*. Then propensity scores are ranked according to their distance from the allosteric source by *quantile regression* to give their final score.

4.4 Construction of the protein graph

The initial step in the method is the conversion of the 3-dimensional coordinates of the atoms of the protein to a graph, that is a collection of nodes (here representing the atoms) and edges that link them. The weight of an edge between two nodes corresponds to the interaction energy of that bond or weak interaction. The construction step was initially developed by Delmotte[54] and is described briefly here.

The crystal structures typically do not contain hydrogen atoms and so the program Reduce[243] is used to add these. Following this, the software FIRST[101, 102] identifies: covalent bonds, which are weighted using standard bond energies[105]; hydrogen bonds, given a value according to the potential of Mayo[48] using a threshold of 0.01kcal mol^{-1}; and hydrophobic interactions (threshold of 8Å), weighted using the potential developed by Lin et al[137]. Finally, electrostatic interactions are accounted for using a standard Coulomb potential and atomic charges for the residues are assigned using the OPLS-AA force field[108].

4.5 Results and Discussion

We investigate here three different scenarios: the binding of the positive allosteric effector, ATP, to the active R state of ATCase and binding of inhibitory CTP to the inactive T state (the inverted case, that is ATP with

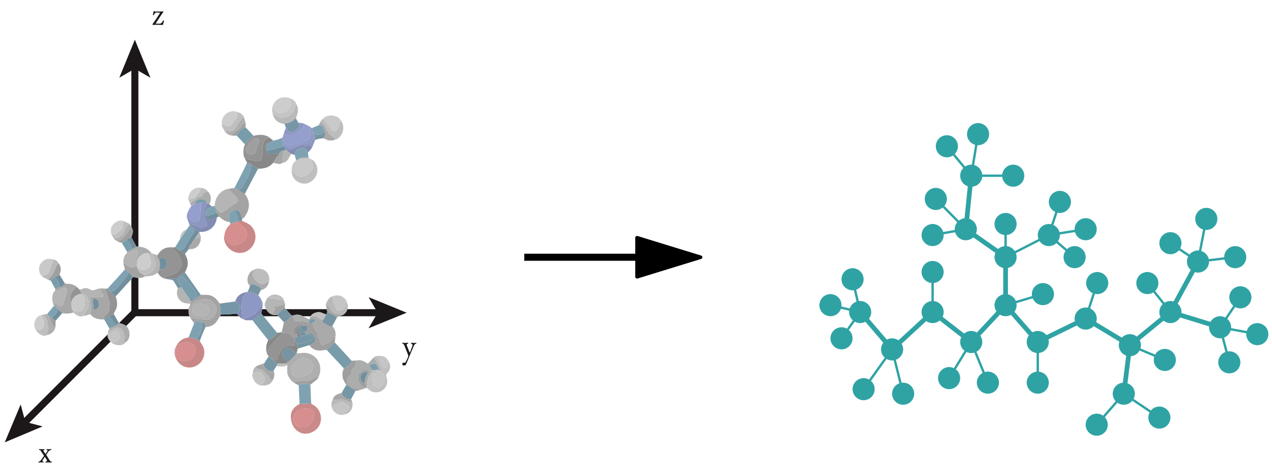


Figure 4.4: The three dimensional coordinates from the PDB file are converted into a *graph* of the protein with edges weighted by bond energies.

the T state and CTP with the R state show negligible effects experimentally so are not considered[111]), which is the heterotropic case. We then study the interaction of the bisubstrate analogue N-(phosphonacetyl)-l-aspartate (PALA) with the active site of the R state in order to elucidate the homotropic mechanism.

Active R State with ATP source

Identification of key residues involved in allostery

ATP is an allosteric activator of ATCase, able to increase the activity of the enzyme by 180% at a 2mM concentration[111]. ATP does not affect the maximal rate of the enzyme, but instead induces a shift from the inactive T state to the active R state. The MWC and EAM models would suggest this shift is caused by a preferential stabilisation of the active R state over the inactive T state, whilst the KNF model would attribute this to the binding of ATP to the inactive state driving it towards the active state. The two models are not necessarily orthogonal however, and we instead focus on the energy transfer within the individual states as a result of ligand binding.

The crystal structure of ATCase chosen (4KGV) has six binding sites for ATP. Bond-to-bond propensities relies on the selection of one or more source residues, such that the weak bonds associated with the atoms in those residues act as the source of the perturbation (Figure 4.3). In order to assess the effect of ATP binding to ATCase, ATP was chosen as the source residue, so that the source of the perturbation was the bonds between ATP and the protein. By representing the protein structure as a graph, bond- to-bond propensities uses a full atomistic description of the protein and so it was possible to select multiple source sites, allowing

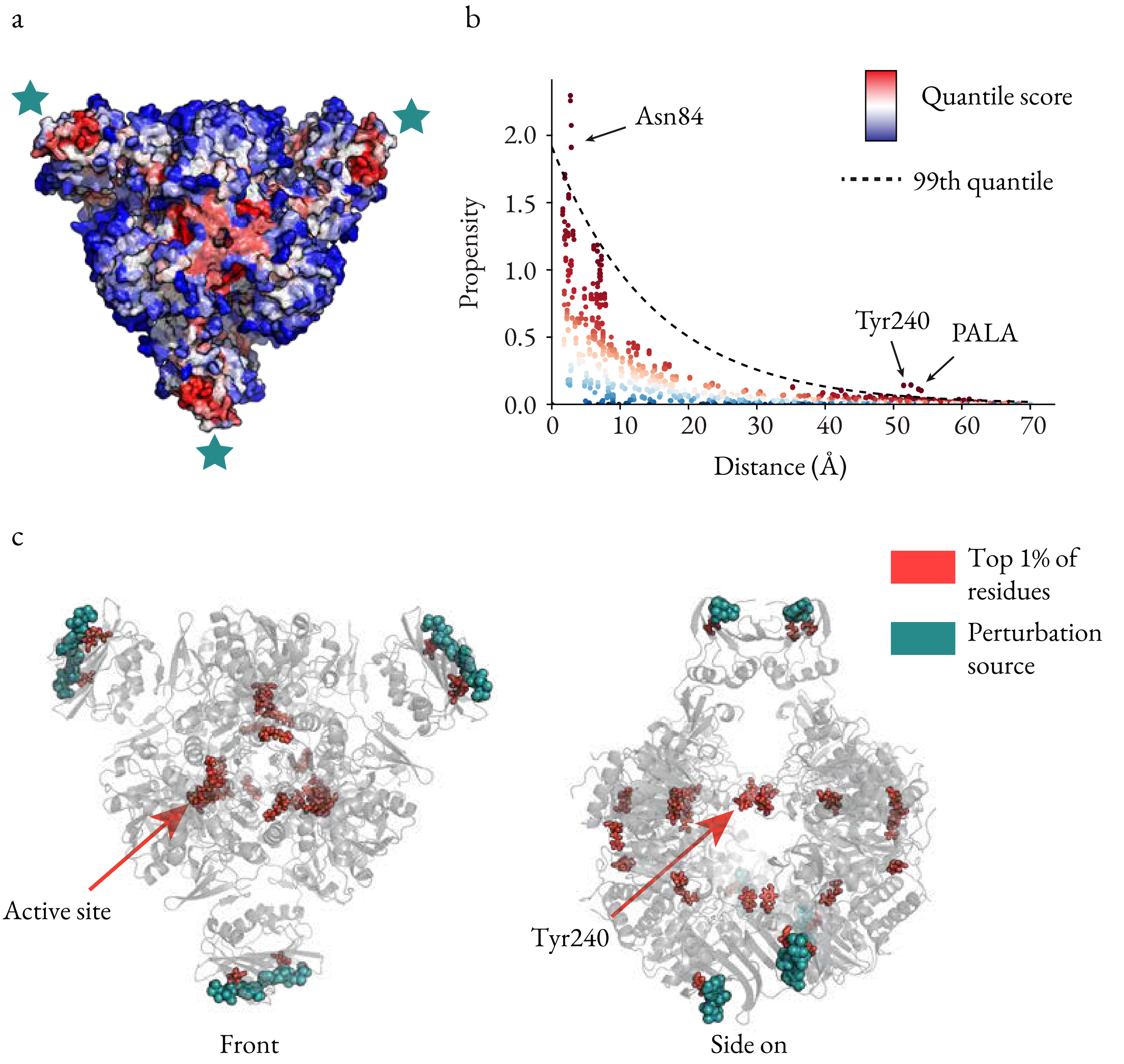


Figure 4.5: Residue ranking of the active R state of ATCase with 6 ATPs as the source by bond-to-bond propensities and conditional quantile regression. All residues are ranked (shown from a red to blue scale) and can be seen either directly on the structure (a) or plotted against distance from the source (b). Here, we further focus on the top 1% as the most significant and plot them on the protein structure (c). Thus (c) displays entirely equivalent results to (a) but the method allows us to highlight those residues that are *particularly* important to energy distribution without making any changes to the underlying data.

us to examine the effect of changing the number of ATP residues to the source and simulating the effect of altering ATP concentration. Initially, all six ATP molecules were included as the source residues in order to clearly identify those residues that score particularly highly and are thus in some way significant to energy transfer in the active R state.

Figure 4.5a demonstrates the output of the method. Each residue in the protein receives a propensity value, which is then ranked by conditional quantile regression, taking account of the distance of the residue from the source sites. Thus in Figure 4.5b, residues are coloured red if they rank highly and blue if they achieve a low rank. In order to investigate the effect of energy flow on allostery, we are interested in the highest scoring residues and so Figure 4.5c displays in red only those residues that obtained a score in the top 1% ($p_R \geq 0.99$). The highest scoring residue is Tyr240, with each of the six residues scoring $p_R = 1$ (see Table A.1). Tyr240 is known to play a role in the T ↔ R transition, with each pair of tyrosine residues forming bonds between their phenyl rings in the R state across the gap between the two catalytic trimers, as opposed to an hydrogen bond to Asp271 in the T state[36, 31]. Cherfils *et al* used site directed mutagenesis to substitute Tyr240 for phenylalanine[36], which is has the effect of removing the hydroxyl group that forms the hydrogen bond in the T state. The resulting mutated enzyme (in the presence of a subsaturating level of PALA such that the amount of T and R state ATCase was approximately equal) shifted strongly towards the R state upon addition of ATP, in contrast to the wild-type protein where no effect was observed when ATP was added.

The other residue which scores highly across all six instances is PALA (average $p_R = 0.996$) , the bisubstrate analogue that sits in the active site, indicating a very strong link between the allosteric and active sites. Indeed it can be seen starkly from Figure 4.5 that the highest scoring residues are concentrated at both the allosteric and active sites. Interestingly, there does not appear to be a clear path between the two sites, a result that is somewhat similar to the strain analysis carried out by Mitchell *et al*[162] suggesting a more complex form of communication than energy simply being shunted along residue "pathways".

Table 4.1: Active R state with six ATP sources, showing the top 20 residues by *quantile score*. All six active site substrate PALA residues and all six Tyr240 residues score above the 99.5% quantile.

Residue Name and Chain	Quantile Score	Residue Name and Chain	Quantile Score
Tyr240 A	1	Pala401 G	0.996
Tyr240 I	1	Pala401 A	0.996
Pala401 K	1	Pala401 K	0.996
Pala401 C	1	Arg65 C	0.996
Tyr240 E	1	Arg65 K	0.996
Tyr240 C	1	Pala401 E	0.996
Tyr240 K	1	Pala401 I	0.996
Tyr240 G	1	Asn84 J	0.996
Asn84 B	1	Asn84 L	0.996
Asn84 F	1	Asn84 D	0.996

The propensity score of a residue is simply the sum of its bond propensity scores. It is often important then to look at the bond scores directly, as key bonds within residues may be missed if other low scoring bonds in the residue "average out" the overall propensity score for the residue. Indeed, this approach highlights the vital importance of understanding proteins at the bond-level, as even course-graining to the residue level can remove crucial information.

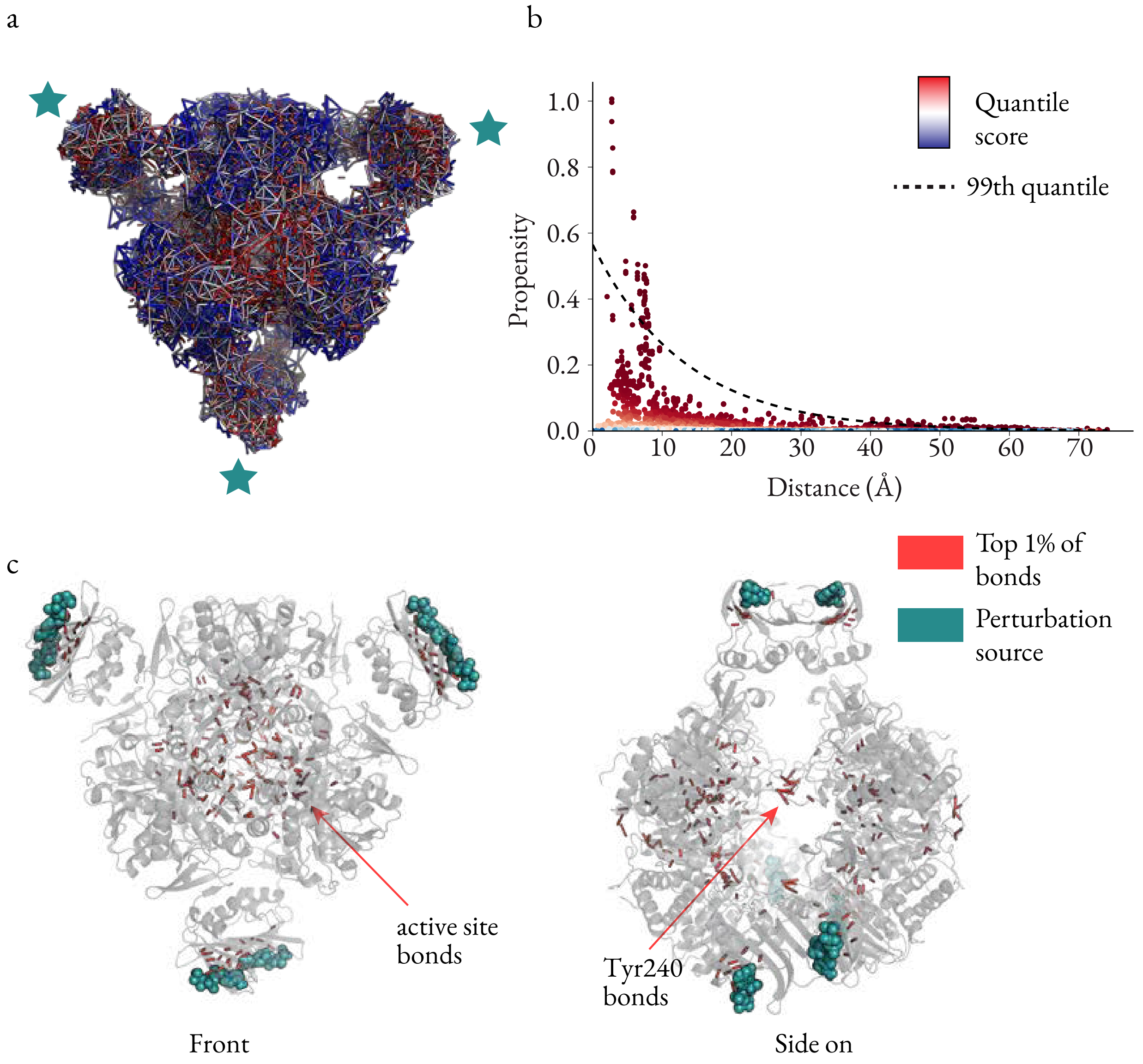


Figure 4.6: Bonds ranking in the active R state of ATCase with 6 ATPs as the source by bond-to-bond propensities and conditional quantile regression. Each bond receives a propensity score, which is then ranked by conditional quantile regression, (a) and (b). We can clearly highlight the highest scoring bonds by only selecting those that have scored above the 99^{th} percentile and display those bonds that are disproportionately affected by the perturbation at the six allosteric sites (c).

To this end, one of the key bonds that emerges is the in hydrogen bond between Lys164 and Glu239, which is a bond that forms in the R state but is not present in the T state[111] (a different Lys164 - Glu239 interaction exists in the T state) and has been highlighted as important in the T ↔ R transition. All six instances of the bond (from each of the six catalytic subunits) score very highly, with an average score of $p_b = 0.997$. When either of Lys164 or Glu239 is substituted with glutamine and lysine respectively, the mutant ATCase protein exists in the R state even in the absence of PALA and does not exhibit homotropic or heterotropic effects[171], highlighting the role of this interaction for cooperativity and allostery. Similarly, Asn111 in the regulatory chain forms a new bond in the R state with Glu109 in the catalytic chain and again this hydrogen bonds scores very highly across all six instance, though interestingly there is a slight asymmetry across the two trimers. In chains C, G and K (See Figure 4.1), the average bond score is $p_b = 0.997$, whilst it is slightly slower for chains A, E and I on the other catalytic trimer ($p_b = 0.985$). Mutation of Asn111 to alanine also leads to the absence of homotropic and heterotropic effects and a shift to the R state[62]. Another interdomain interaction identified as being highly important for stabilisation of the active R state is the Glu50 - Arg234[225] and indeed two different hydrogen bonds score very highly across all six catalytic chains: 0.995 for one set of six hydrogen bonds and 0.994 for the other, suggesting that the link between these two residues is particularly important for energy transfer.

Formation of allosteric pathway appears to require three ATP sources in cyclic formation

Whilst in the previous section, all six ATP molecules were used as the source of the perturbation, the method allows us to select any number of ligand sites as the source. We can then model the effect of progressively adding more ATP molecules to ATCase to investigate how energy flow occurs when different numbers of ligand sites are bound to. Starting with a single ATP source on chain B, further ligands are added to the perturbation source on chain F (i.e. two sources), then chain J (three sources) and the bonds and residues are scored in each case.

When just a single ATP source is used (in this case, arbitrarily, on the regulatory chain B), it is immediately apparent that the ranking of the residues in the protein (and thus their response to the energy perturbation originating at the allosteric ligand) is different to when all six ATP molecules are used as the source. For example, the active site residue PALA was one of the highest scoring residues when six ATP molecules were used (average quantile score = 0.996) but it scores lower here (average score = 0.941). Instead, most of the highest scoring residues appear to be located near to the allosteric site on chain L, which is situated across the multimer from the chain B source (see Fig 4.1) Asp19 (which binds to ATP) on chain L scores $p_b = 1$, whilst Lys56 scores $p_b = 0.996$. Mutation of Lys56 to alanine led to the disappearance of homotropic cooperativity in the presence of ATP, but not CTP[44], suggesting it is involved in the communication pathway between ATP and the active site. The other highly significant residue in the case of the six ATP sources was Tyr240 ($p_b = 1$) and the results here are interesting. The pair of Tyr240 residues in chains E and K still score highly ($p_b = 0.993$ and 1 respectively whilst the Tyr240 residues in the other four catalytic chains score lower (average of 0.932 across the four catalytic chains). As Fig 4.1 shows, catalytic chains E and K are in fact situated on the

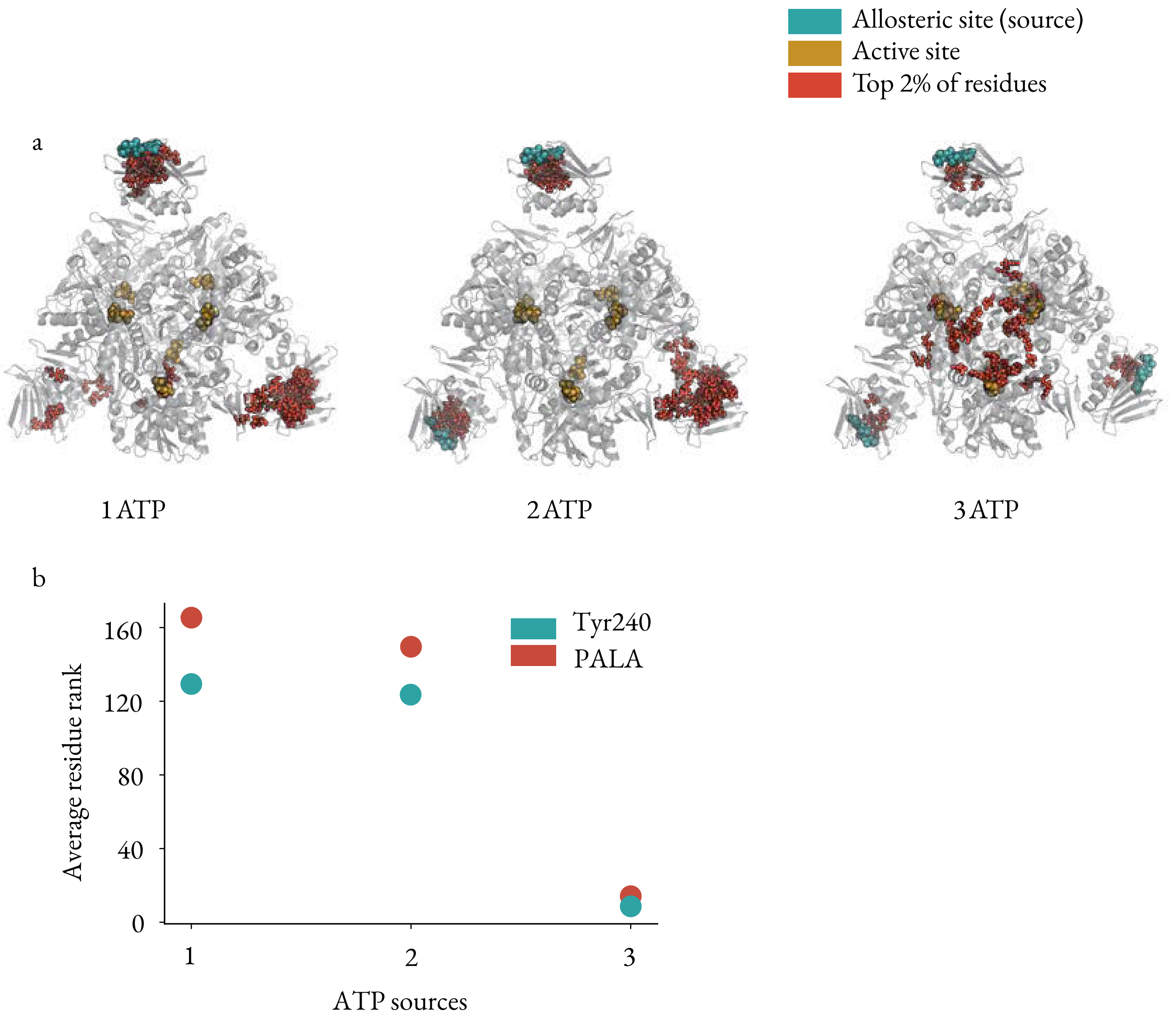


Figure 4.7: a) Binding of the first two ATP molecules does not appear to show communication between the allosteric source sites and the active site (identified by the gold PALA residues) in the active R state by bond-to-bond propensities. However, binding of a third ATP ligand leads to a switching effect, at which point all six active site PALA residues score within the top 20 residues out of 2790. b) Scatter plot showing the average rank of the two highest scoring residues (out of 2790) from the 6 ATP case.

other side of the protein to the chain B source, again suggesting that communication within ATCase is both long range and not driven by a contiguous pathway of individual residues.

As can be seen in Figure 4.7, the pattern of high scoring residues is similar when a second ATP molecule (on chain F) is included in the perturbation source, with significant residues appearing again in the region of the allosteric sites on chains J and L distal to both the source sites on chains B and F. Indeed PALA's score is very similar to the single ATP source case ($p_b = 0.946$, showing little change in the communication to the active site upon "binding" of a second ligand. Tyr240 scores slightly higher on average here ($p_b = 0.955$), though no single Tyr240 residues scores as highly as in the six ATP case. The overall effect is that there does not appear to be a significant change in the response to the perturbation between one or two ATP molecules binding.

In contrast, a significant change occurs upon addition of a third ATP molecule to the perturbation source. The average score for PALA now jumps to $p_b = 0.996$, the same as it is in the six ATP case, whilst Tyr240 receives a score of $p_b = 0.998$. If a third ATP source is added instead to chain D (see Fig 1), the increases are not as stark (PALA = 0.948 and Tyr240 = 0.962), suggesting that the "cyclical" distribution of the ATP sources around the protein may be important for facilitating communication with the active site.

Active unligated R state with PALA source

In order to study energy flow in homotropic case, PALA, which acts as a bisubstrate analogue in the active site, was used as the perturbation source (or more specifically, the bonds between PALA and the active site were the source). Again, all six PALA residues were included as source residues in order to clearly identify those residues in the structure that are particularly significant with respect to energy distribution and thus may be implicated in the cooperative mechanism.

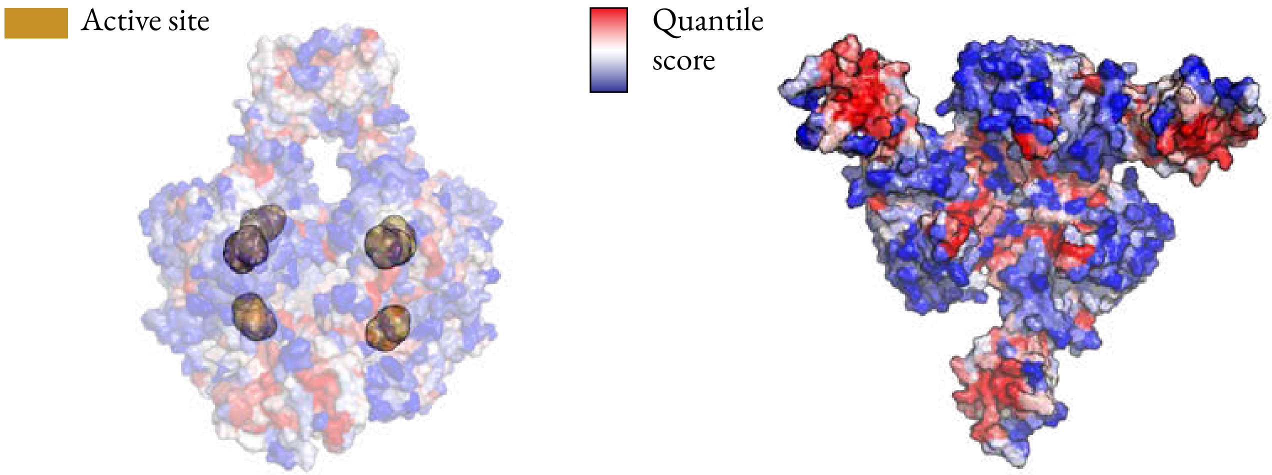


Figure 4.8: To investigate homotropic cooperativity, the six PALA substrates were selected as the source on the active unligated R state. The structure on the right shows just one half on the protein for clarity and here it is clear that the highest scoring regions (in red) of ATCase are located around the active and allosteric sites.

Figure 4.8 shows the overall effect of a perturbation at the six active sites. Similarly to when ATP is used as the source, the highest scoring regions are clustered around both the allosteric and active sites. The result reinforces the idea that there is a form of communication between these distal sites and once again, it is interesting to observe that there do not appear to be obvious, individual pathways between the two types of site.

Table 4.2: R state unligated (1D09), showing the top 20 residues by quantile score. Every Glu50 residue scores the maximum of 1.

Residue Name and Chain	Quantile Score	Residue Name and Chain	Quantile Score
Asp19 D	1	Ile44 L	0.996
Asp19 H	1	Ile44 H	0.996
Asp19 L	1	Ile44 D	0.996
Glu50 E	1	Asp90 E	0.996
Glu50 A	1	Asp90 A	0.996
Glu50 K	1	Asp90 C	0.996
Glu50 G	1	Arg105 G	0.996
Glu50 I	1	Arg105 C	0.996
Glu50 C	1	Arg105 K	0.996
Met1 H	0.996	Met1 D	0.996

As seen in Table 4.2, the highest scoring residue is Glu50, with all six instances of the residue scoring the maximum of $p_b = 1$. As mentioned earlier, Glu50 is a crucial residue for the stability of the R state and substitution of glutamic acid for alanine leads to dramatic changes in the enzyme. Activity is reduced by 15-fold and cooperativity is completely lost[174]. Significant communication to the allosteric sites is also seen, with Asp19 (one of the residues that interacts with ATP and CTP) scoring $p_b = 0.992$ over the six sites, whilst Lys60, another allosteric residue, scores highly ($p_b = 0.989$) over the regulatory chains on one side of the protein (chains D, H and L), again demonstrating asymmetry in the distribution of energy over the structure. Glu233 forms a salt link with Arg229 only in the R state, which orients Arg229 into the active site[112]. The removal of the salt link via mutation of glutamic acid to serine leads a significant decrease in both catalytic activity and cooperativity[18] and indeed Glu233 scores highly overall ($p_b = 0.985$), though once again a difference is seen between the two catalytic trimers (trimer AEI scores $p_b = 0.993$ vs $p_b = 0.977$ for the opposite trimer CGK).

Analysis of the *bond level* data reveals further information. As expected, the previously mentioned Glu233 - Arg229 salt bridge ranks very highly ($p_b = 0.996$) whilst the Glu50 interaction with Arg167 (which itself interacts with PALA in the active site, being positioned correctly by its association with Glu50) involving two types of bonds scores above $p_b = 0.995$ across all bonds. The Asp19 - Lys56 link scores an average of $p_b = 0.999$ over its six instances and it was found that substitution of lysine by alanine affected not only cooperativity but also removed the ability of ATP to activate the enzyme[44]. As Asp19 is one of the allosteric residues, it appears that this bond to Lys56 may be crucial in communicating with the active site.

Table 4.3: T state with CTP (5AT1). The top 20 residues by quantile score are listed and both Arg56 and Arg65 appear six times each. These residues sit at the C1-C2 interface within the catalytic subunits.

Residue Name and Chain	Quantile Score	Residue Name and Chain	Quantile Score
Arg65 G	1	Arg65 I	0.996
Arg56 G	1	Arg65 E	0.996
Arg65 C	1	Arg56 C	0.996
Arg65 A	1	Arg56 I	0.996
Arg56 A	1	Arg56 E	0.996
Arg65 K	1	Arg85 H	0.996
Asn84 H	1	Ile86 F	0.996
Asn84 D	1	Ile86 B	0.996
Ile86 J	1	Asn84 L	0.996
Arg56 K	0.996	Ile86 H	0.996

Stabilisation of the catalytic trimer: inactive T state with CTP

In contrast to both the case of ATP and PALA being used as a the perturbation source in the active R state, the highest scoring regions of the protein when CTP is the source in the inactive T state appear most strongly at the C1-C2 interface (See Figure 4.9) instead of the active site. The catalytic trimers move as essentially rigid units during the T ↔ R transition so there is little change between the inactive T state and the active R state in this region.

Two residues in particular stand out in Table 4.3: Arg65 (average $p_b = 0.999$) and Arg56 (average $p_b = 0.998$). It can be seen from Figure 4.9 that both these residues bridge the C1-C2 interface, though they do not form links to each other. Looking at the bond data, one of the key interactions made by Arg65 is with Asp100 (average $p_b = 0.999$). This specific interaction was identified as being important for the stability of the catalytic trimer[14] and replacement of Asp for either Asn or Ala reduces the half life of inactivation of the catalytic subunit. Arg65 additionally forms a hydrogen bond to His41, another residue implicated in catalytic subunit stability and this interaction also scores highly ($p_b = 0.983$), though once again there is a significant difference between the two catalytic subunits, with the interactions in the AEI trimer scoring $p_b = 0.999$, compared to $p_b = 0.968$ in trimer CGK. There is possibly a link here with experimental data showing that in the R state, only half (i.e. three) of the His41-Glu37 interactions are broken[83, 223] during the transition from the T state, demonstrating an intriguing asymmetry that appears to be captured by bond-to-bond propensities. In fact, the results for Glu37 are even more stark, with the average quantile score across chains A, E and I 0.990 versus 0.262 for chains C, G and K, a remarkable difference between essentially symmetrically equivalents sets of residues. Glu37 itself has been associated with stabilising the catalytic trimer[14].

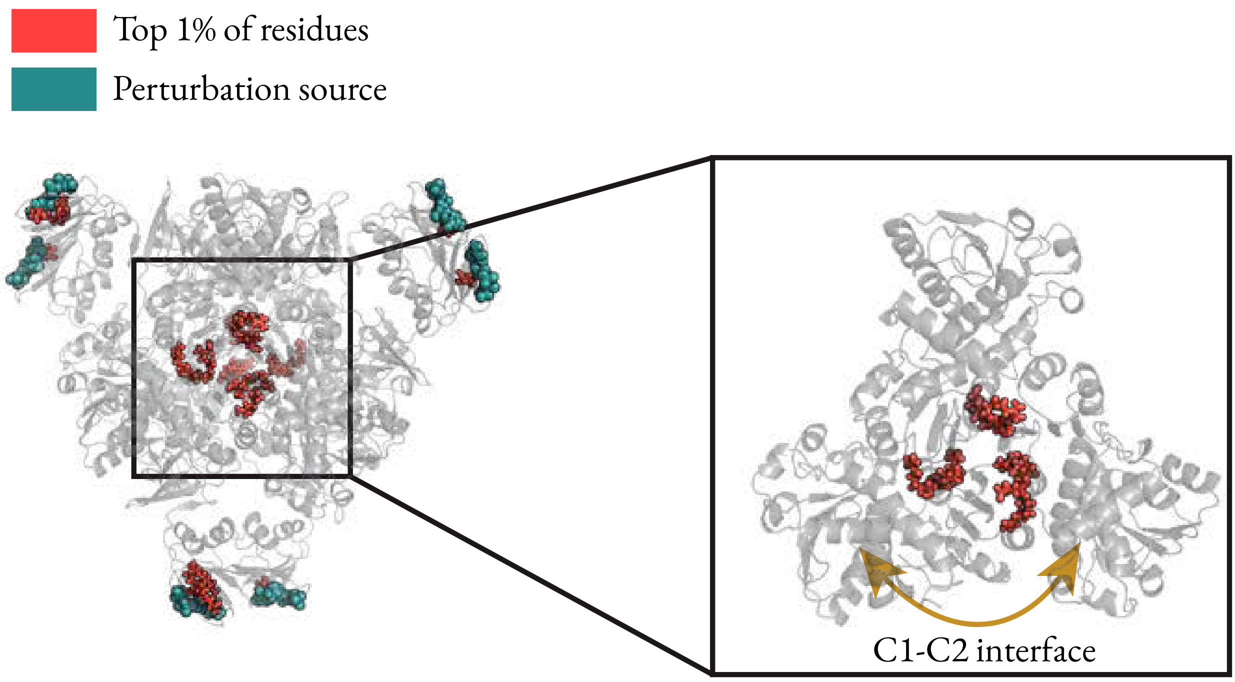


Figure 4.9: When six CTP molecules are used as the perturbation source on the inactive T state, the highest scoring residues appear at the C1-C2 interface, which is the boundary between catalytic subunits within the catalytic trimer. Arg56 and Arg65 are two of the highest scoring residues, shown on the right, forming a link across the C1-C2 interface.

Conversely, there appears to be little experimental data on Arg56, nor on the two highest scoring links it makes: to Gly72 ($p_b = 0.999$) and Gln60 ($p_b = 0.986$), though the Gly72 interaction occurs across the C1-C2 interface[107, 223] so it would seem possible that this interaction is also involved in stability of the trimer. Perhaps less surprisingly, a number of residues located close to the CTP site also rank highly: Ile86, which forms a non-polar interaction with the nucleotide[111], and Asn84, which interacts with the phosphate part of CTP[223] score $p_b = 0.993$ and 0.985 respectively. Val17 also forms a non-polar interaction with CTP, though scores slightly lower with an average quantile score of 0.978.

Sequential binding of CTP shows a similar pattern to that of ATP

Whilst the identity of the highest scoring residues when CTP is used as the perturbation source is different to that when ATP is used, there is a similar "switching effect" when a third CTP molecule is included as a source residue in a cyclic arrangement around the ATCase structure. As seen in Figure 4.10, inclusion of a third ligand leads to the clustering of high scoring residues in the region of C1-C2 interface, between the catalytic subunits within a trimer, in a similar fashion to the previously discussed six CTP source case. It appears to be the interaction between the CTP ligands located in such a way around the ATCase protein that leads to energy flow focusing on those residues identified as particularly significant.

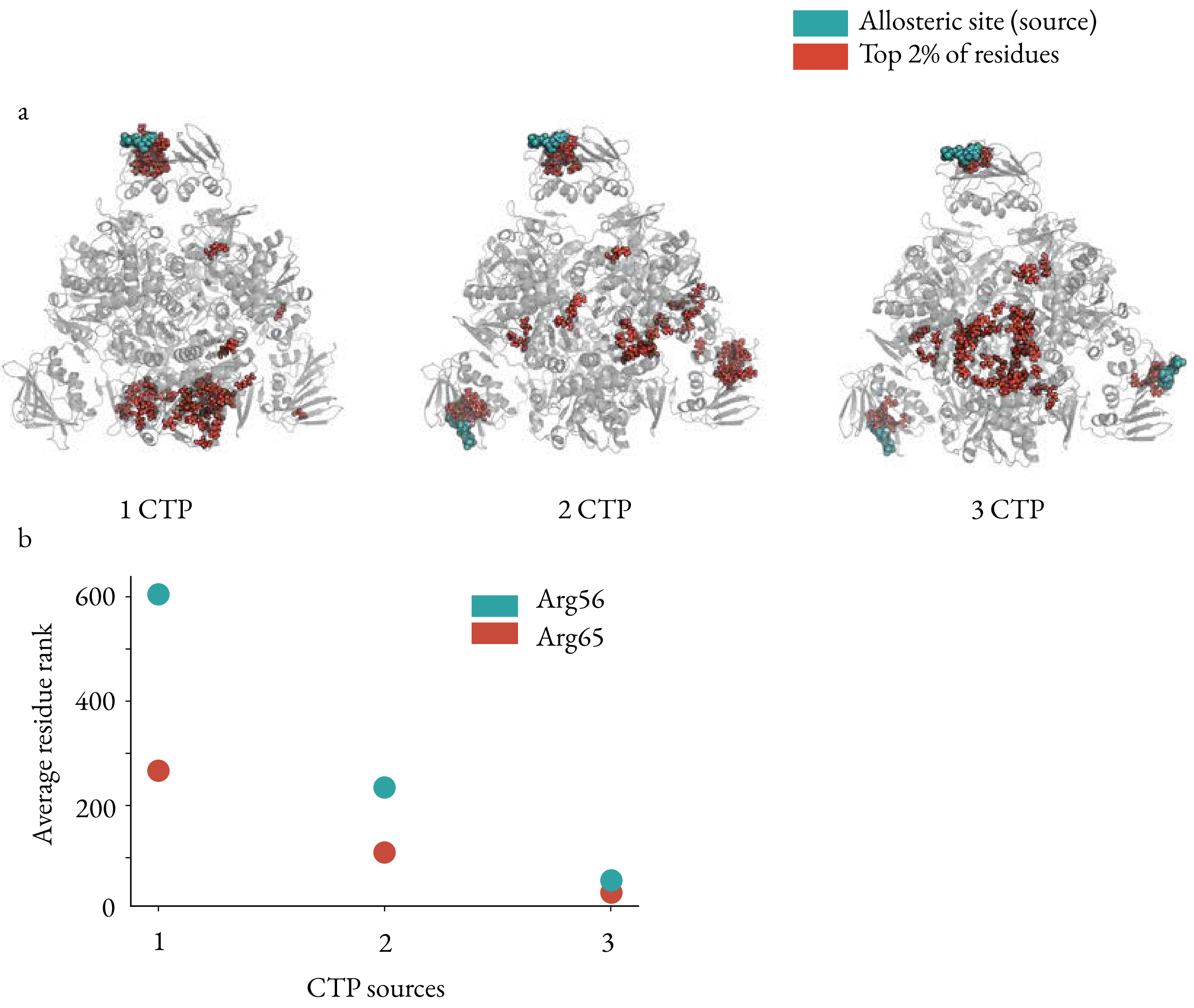


Figure 4.10: a) The top 2% of residues displayed when varying numbers of CTP molecules are included as source residues. In contrast to the ATP case, there does not appear to be as much communication with the distal allosteric sites for one or two source ligands but again inclusion of a third ligand on chain J leads to the results resembling the six CTP case described previously. b) Scatter plot showing the average rank of the two highest scoring residues (out of 2790) from the 6 CTP case.

This effect is illustrated by focusing on two of the highest scoring residues from the previous case where all six CTP residues were included as source residues: Arg56 and Arg65. Starting from a single CTP source, the scores for Arg65 progress from 0.904 to 0.961 and then to 0.989 when a third ligand is included whilst equivalently for Arg56, the scores are 0.779, 0.916 and 0.982 as each of the CTP ligands is added. The increases in scores here of the two highest scoring residues are actually more "linear" than in the case of ATP but it is still only when a third CTP ligand is included cyclically that the results from the six CTP case are replicated. When the third ligand is instead added to chain D, such that the three CTP ligands are now bound to chains B, D and F, the increase in score upon addition of the third ligand is smaller for both Arg65 ($p_b = 0.973$)

and Arg56 ($p_b = 0.922$) which again suggests that it is a particular feature of the geometric arrangement of the allosteric ligands that facilitates communication to the key residues within the protein.

4.6 Conclusions

In this chapter, we have demonstrated that bond-to-bond propensities, having previously been used to predict allosteric sites from knowledge only of the active site of a protein, can be used to investigate the energy flow process of the "reverse process" of a ligand binding to an allosteric site. In the active R state, using ATP as a source of a perturbation reveals a number of residues as being particularly significant, including Tyr240, which links the two sides of the ATCase protein, and PALA, which sits in the active site. There is thus a clear communication pathway between the allosteric and active sites in ATCase but in accordance with other computational studies of ATCase[162], this communication does not appear to occur through discrete pathways of residues but instead via a collective of lower scoring residues. A qualitatively similar effect is observed in the 'scissor model' of allostery in which the normal mode of the system leads to large changes at either end of the molecule, whilst leaving the centre of the molecule largely unperturbed[159]. In each case, the network structure of the protein appears to arranged in order to facilitate long range communication between sites.

Furthermore, it appears that the geometrical distribution of the ligands is important. Only when three ATP residues arranged cyclically around the ATCase structure are used as the perturbation source are the previously mentioned residues identified as high scoring; when a single or two ATP molecules are used then there does not appear to be a strong link the active site, though there does seem to be communication between the distal allosteric sites.

Homotropic allostery was investigated by using the six PALA substrates as the perturbation source. The regions that scored most highly in this case were the active and allosteric sites, reinforcing the idea that the two types of sites are highly coupled in the active state and also hinting that homotropic and heterotropic effects are not orthogonal phenomena and are instead closely intertwined.

Finally, allosteric inhibition of ATCase by CTP was studied by using the CTP molecules as the perturbation source. Interestingly, rather than the active site region being identified as significant, it was instead the C1-C2 interface of the catalytic trimers that was found to be particularly coupled to the allosteric sites. The boundary between the catalytic subunits has been found to be important for stability of the enzyme but not particularly vital for catalytic activity. It is possible that the allosteric ligands may play subtlety different roles when binding to the active and inactive states of the enzyme.

Chapter 5

Elastic networks

5.1 Introduction

Given the success of the bond-to-bond propensity method in both predicting allosteric sites using only the crystal structure of a protein[5] and its ability to elucidate key aspects of the allosteric mechanism as presented in Chapter 4, the question is raised as to the particularity of the physical process propensity describes. Consideration of how binding of a ligand to a protein propagates energy across the protein structure thus lead to the development of a novel computational method: *Elastic network response*. In this chapter, we describe the motivations for the method and provide a derivation based on the equations of equilibrium for a mechanical system. We then apply the method to a range of proteins and supplement the analysis using a technique called *infinitesimal rigidity* to provide a complete description of how mechanical perturbations are distributed over the network structure of the protein.

5.2 Motivation: A mechanical view of allostery

Using a combination of the *Elastic response* method and *Infinitesimal rigidity*, we are able to build up a complete description of how a structural perturbation at a chosen site (modelled as the compression of a set of bonds, for example those between a ligand and the protein allosteric site) is distributed over the entire protein structure. We would tend to expect that a highly flexible protein would be unlikely to transfer energy in this fashion so infinitesimal rigidity allows us to measure and visualise those parts of the protein that are *rigid* and thus more likely to be amenable to communication through structural perturbations.

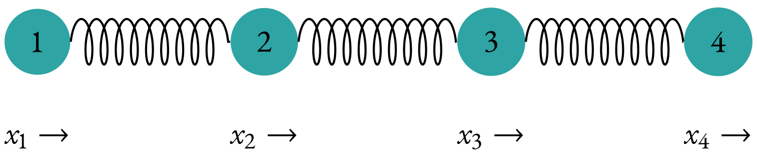
The potential mechanism we can attempt to elucidate using our method is the *extended conformational selection* model[7, 47, 28], invoked to explain how conformational selection can occur in proteins where either the active or allosteric sites are closed in the active state[228]. Here, we consider that the active state itself is composed of two (or perhaps more) substates so that binding of a ligand to the allosteric site of an "active state" protein causes a small change at the active site, perhaps leading to some structural rearrangement that further stabilises the active site, via the breaking of various weak interactions in a specific manner. We therefore hypothesise that any perturbation of the allosteric site by our ligand would cause a particularly high response at or around the active site in the active state of the protein. The methods described in this chapter strictly only hold in the regime of infinitesimally small changes and our assumption is that it will approximately hold for the small structural changes often seen upon allosteric binding.

5.3 Elastic network response

The 1-dimensional case

Two centre interactions

The mathematical underpinnings of bond-to-bond propensity were first introduced by Schaub *et al*[206] in the context of electrical networks. We make use of the well known "mechanical-electrical analogy" to interpret the presented equations instead in terms of a 1-dimensional ball-and-spring model.



We start by considering the displacement of the atoms (mechanical *displacement* therefore plays the role of *potential* in the electrical network) and how they relate to the edge variables, which here are *stretches* (or *compressions*) of the springs, analogous to *potential difference* using a small model as an example:

$$
\begin{aligned}
y_1 &= x_2 - x_1 \\
y_2 &= x_3 - x_2 \\
y_3 &= x_4 - x_3
\end{aligned} \tag{5.1}
$$

We can write this more succinctly in matrix form by utilising the *incidence matrix B*, that maps node variables to edge variables:

$$\mathrm{y} = B^T x \tag{5.2}$$

In order to find the force on each spring, we multiply by the spring constant, k, according to Hooke's Law:

$$f_1 = k_1 y_1 \tag{5.3}$$

$$f_2 = k_2 y_2 \tag{5.4}$$

$$f_3 = k_3 y_3 \tag{5.5}$$

or in vector form by compiling the spring constants into a diagonal matrix G:

$$f = G\mathrm{y} \tag{5.6}$$

We then consider the corresponding force on the balls by Newton's Third Law, taking into account any external forces f_{ext} on the balls:

$$f_{ext,1} = -f_1 \tag{5.7}$$

$$f_{ext,2} = f_1 - f_2 \tag{5.8}$$

$$f_{ext,3} = f_2 - f_3 \tag{5.9}$$

$$f_{ext,4} = f_3 \tag{5.10}$$

which is just:

$$f_{ext} = Bf \tag{5.11}$$

Overall therefore we can write:

$$\begin{aligned} f_{ext} &= BGB^T x \quad &(5.12)\\ &= Lx \quad &(5.13)\end{aligned}$$

L being our Laplacian from previously, which in this context therefore takes the form of a *stiffness matrix*. In order to solve this equation for a set of displacements given some input forces, we multiply both sides by the Moore-Penrose pseudoinverse of L:

$$x = L^{\dagger} f_{ext} \tag{5.14}$$

as L is singular and has a single zero eigenvalue, which here corresponds to the rigid translational degree of freedom along the 1-dimensional line. The presence of this rigid mode presents a problem: imagine we introduce a set of external forces on the nodes that lead to a displacement and obtain a numerical value. We are then able to add any multiple of $\begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}$ and still obtain a solution. In other words the solution we obtain is not *unique*. The situation is resolved by working in the *translationally* and *rotationally invariant* edge-space, which we know from Eq. (5.2) how to do, using the incidence matrix:

$$\mathrm{y} = B^T L^{\dagger} f_{ext} \tag{5.15}$$

Whilst we have amended the issue of infinite solutions in the node-space, we face a similar problem of *redundancy* when considering our input force vector on the nodes. If we choose an input force of $\begin{bmatrix}3 & -2 & 0 & 0\end{bmatrix}$, we will obviously compress the first spring, but we will also still have a net force to the right, leading to translational motion. What would be more convenient is if we restrict the choice of input forces to only those that lead to *pure* stretches or compressions, so that we do not accidentally choose force vectors that have large translational components, given our desire is to study internal energy transfer within the protein. We make use once again of the incidence matrix:

$$f_{ext} = Bf_0 \tag{5.16}$$

where f_0 now represents a set of external input forces on the bonds. Putting this all together:

$$\mathrm{y} = B^T L^\dagger Bf_0 \tag{5.17}$$

Giving us an expression for the change in length of the various springs in the system, given a particular input set of external forces applied to the springs. We can then convert this to the output force on each bond by simply multiplying through by each spring constant:

$$f = GB^T L^\dagger Bf_0 \tag{5.18}$$

and the potential energy increase of each bond is:

$$V = \frac{1}{2}\mathrm{y} \odot \mathrm{y} \tag{5.19}$$

where $\odot$ is the Hadamard, or elementwise, product. We now have an equation (5.18) that appears similar to our expression for bond-to-bond propensity. Can we therefore conclude that we are modelling our protein as a elastic model? Whilst clearly a protein is not a 1-dimensional object, might we propose that we can separate out the three directional components so each has a expression in the form of Equation (5.18)? The precedent for doing so comes from the formulation of the Gaussian Network Model (GNM) of Bahar *et al*[12, 90], who justified their model on the basis that motion of a residue within a protein is determined by the number of contacts around it and thus drew inspiration from models such as those of Flory[72] and Rouse[109] who studied polymer physics. The Hamiltonian for the GNM, defining a displacement of a node from equilibrium position as $u = x - x_0$, is:

$$V = \frac{1}{2} \sum_{\langle i,j \rangle,\alpha} (u_i^\alpha - u_j^\alpha) K_{ij}^\alpha (u_i^\alpha - u_j^\alpha) = V_x + V_y + V_z \tag{5.20}$$

Unfortunately we run into a serious problem, pointed out by Thorpe[232] that this model, whilst possessing the necessary translational invariance, lacks *rotational invariance*. The consequence of this is that if we substitute rotational motion, as defined by the cross product, into the Hamiltonian we calculate a non-zero contribution to the energy:

$$u_i = \Theta(R_i \times \hat{z}) \tag{5.21}$$

with Θ a small angle, $\hat{z}$ an arbitrary rotational axis and R_i a vector from any point on that axis to a point in the protein (for example an atom or residue). If we set $u = d$ for all atoms to represent a rigid translation, we can see that all terms in u simply cancel out in Equation (5.20), but in general u terms in the rotational case will not. As a rigid body rotation involves no change in the lengths of any of the springs in the model, it should not induce a change in the potential energy of the system. Thorpe further notes that Equation (5.20) corresponds to a Born model, that was initially used to study motions of lattices but was eventually realised to be inadequate because of its inability to account for rotations[114].

If we instead write the form of the Hamiltonian for what is known as the Anisotropic Network model (ANM) (in a slightly unusual form that will be convenient later):

$$V = \frac{1}{2} \sum_{\langle i,j \rangle} k_{ij} [(u_i - u_j) \cdot \hat{r}_{ij}]^2 \tag{5.22}$$

where $\hat{r}_{ij}$ is the unit vector between nodes i and j. The ANM is a 3-dimensional model that incorporates information about the orientation of the springs joining pairs of atoms. Again, it is obvious that setting all the displacement terms to the same value will result in a zero energy contribution from translational motion as required. Here, however we correctly capture the necessary rotational invariance. By plugging in the rotation term to the Hamiltonian again:

$$
\begin{aligned}
V &= \frac{\Theta}{2}\sum_{\langle i,j\rangle} k_{ij}\left[\left((r_i \times \hat{z}) - (r_j \times \hat{z})\right)\cdot \hat{r}_{ij}\right]^2 \\
&= \frac{\Theta}{2}\sum_{\langle i,j\rangle} k_{ij}\left[\left((r_i - r_j) \times \hat{z}\right)\cdot \hat{r}_{ij}\right]^2 \\
&= 0
\end{aligned}
$$

where we have used the fact the cross product is distributive. Using the definition of the cross product, the term inside the brackets will produce a vector that is orthogonal to the vector $r_i - r_j$ and thus orthogonal to $\hat{r}_{ij}$ which is just the unit vector in the same direction as $r_i - r_j$. Therefore we are taking the dot product of two orthogonal vectors, which gives us the required zero energy contribution.

It is clear then that if we wish to interpret Equation (5.17) as representing an elastic model representation of a protein, we must construct the model including the full 3-dimensional structure of the protein.

The 3-dimensional case

We build up the framework using similar arguments to those of Strang[226] who deals with the 2-dimensional case of static structures in the context of structural stability. However, whilst Strang uses angles to represent spatial information, we instead use vectors here as this turns out to be more transferable to later arguments. Throughout we use a right handed coordinate system, as this is used by the PDB files that will eventually provided the atom coordinates for the method.

Firstly we consider how the extension of an edge can be written in terms of the displacements of its associated nodes for a single spring:

$$
e = |r_{ij}| - |r_{ij,0}| \tag{5.23}
$$

or simply the extension is the length after displacement minus that before. In order to find how this expression is related to the components of the displacement (x_i for node i, x_j for node j and likewise in the y and z directions), we first expand $|r_{ij,0}|$:

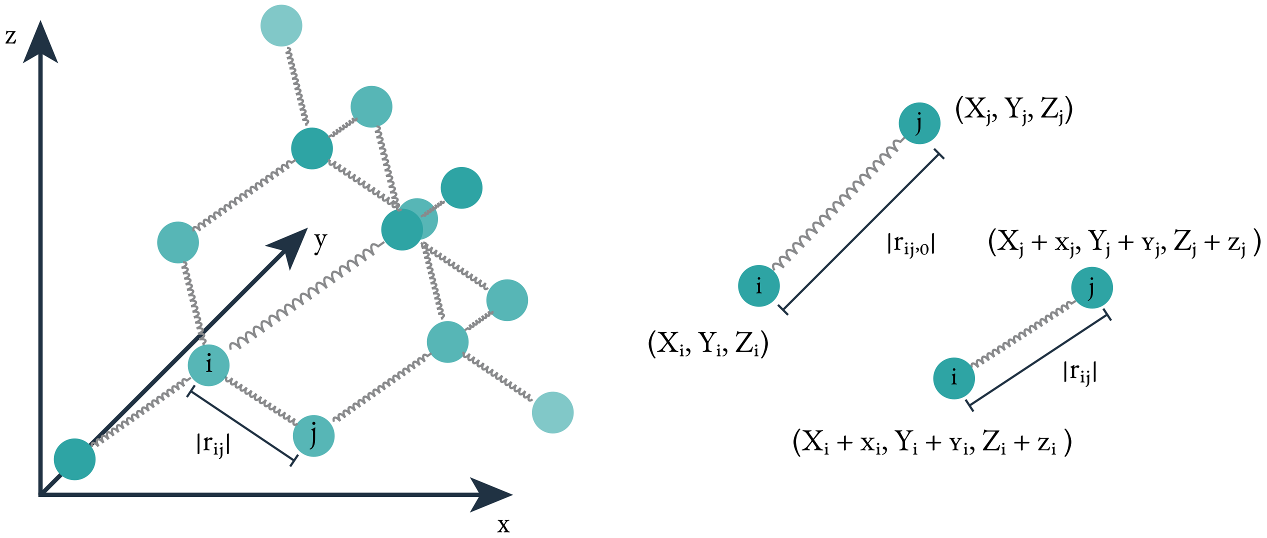

z
y
x
i
j
|r_{ij}|
j
(X$_j$, Y$_j$, Z$_j$)
|r$_{ij,0}$|
(X$_j$ + x$_j$, Y$_j$ + y$_j$, Z$_j$ + z$_j$)
i
j
(X$_i$, Y$_i$, Z$_i$)
i
|r$_{ij}$|
(X$_i$ + x$_i$, Y$_i$ + y$_i$, Z$_i$ + z$_i$)

Figure 5.1: Edge displacement in a three dimensional elastic network model in terms of node position changes.

$$\begin{aligned} |r_{ij,0}|^2 &= \left[(X_i - X_j)^2 \;\; (Y_i - Y_j)^2 \;\; (Z_i - Z_j)^2\right] \\ &= \left[(X_i^2 - 2X_iX_j + X_j^2) \;\; \cdots \;\; \cdots\right] \end{aligned}$$

where X_i represents the initial position of node i in the x-axis and similarly for the other directions and we have abbreviated the y and z terms in the third line. We can do the same for the stretched spring:

$$\begin{aligned} |r_{ij}|^2 &= \left[\big((X_i + x_i) - (X_j + x_j)\big)^2 \;\; \big((Y_i + y_i) - (Y_j + y_j)\big)^2 \;\; \big((Z_i + z_i) - (Z_j + z_j)\big)^2\right] \\ &= \left[X_i^2 + 2x_iX_i - 2(X_iX_j + x_iX_j + x_jX_i + x_ix_j) + X_j^2 + 2x_jX_j + x_j^2 \;\; \cdots \;\; \cdots\right] \\ &= |r_{ij,0}|^2 + 2x_iX_i - 2x_iX_j - 2x_jX_i + 2x_jX_j + \;\cdots + \;\cdots \;+\; \Theta(x^2) \end{aligned}$$

In the last line, we have collected nonlinear terms together and will drop them from the calculation. Physically, this means the model will only hold for *small displacements*. We then used the expansion of $|r_{ij,0}|^2$ from earlier to substitute the for the relevant terms. We now rewrite the second part of the equation such that we can complete the square:

$$|r_{ij}|^2 = |r_{ij,0}|^2 + 2|r_{ij,0}|\Big(\frac{x_iX_i - x_iX_j - x_jX_i + x_jX_j + \cdots + \cdots}{|r_{ij,0}|}\Big) + O(x^2)$$

$$= \big(|r_{ij,0}| + \frac{(x_iX_i - x_iX_j - x_jX_i + x_jX_j + \cdots + \cdots)}{|r_{ij,0}|}\big)^2 + O(x^2)$$

Square rooting both sides and ignoring nonlinear terms:

$$|r_{ij}| = |r_{ij,0}| + \frac{(x_iX_i - x_iX_j - x_jX_i + x_jX_j + \cdots + \cdots)}{|r_{ij,0}|}$$

$$= |r_{ij,0}| + \frac{\big((X_i - X_j)x_i - (X_i - X_j)x_j + (X_i - X_j)y_i + (Y_i - Y_j)y_j + (Y_i - Y_j)z_i + (Z_i - Z_j)z_j\big)}{|r_{ij,0}|}$$

We therefore have an expression for the extension e, which can be written more conveniently using vector notation:

$$e = \begin{pmatrix} \frac{(X_i-X_j)}{|r_{ij,0}|} & \frac{(Y_i-Y_j)}{|r_{ij,0}|} & \frac{(Z_i-Z_j)}{|r_{ij,0}|} & -\frac{(X_i-X_j)}{|r_{ij,0}|} & -\frac{(Y_i-Y_j)}{|r_{ij,0}|} & -\frac{(Z_i-Z_j)}{|r_{ij,0}|} \end{pmatrix} \begin{pmatrix} u_i \\ u_j \end{pmatrix}$$

or even more compactly:

$$e = \begin{pmatrix} \frac{r_{ij,0}}{|r_{ij,0}|} & -\frac{r_{ij,0}}{|r_{ij,0}|} \end{pmatrix} \begin{pmatrix} u_i \\ u_j \end{pmatrix} \tag{5.24}$$

It is apparent now that the row vector containing the terms in $r_{ij,0}$ is our incidence matrix for the 3-dimensional case, where instead of each row containing a 1 and -1 in the positions i and j for the nodes joined by the edge, we have $r_{ij,0}$ and $-r_{ij,0}$, each being a $1 \times 3N$ vector. We can then follow the argument through in a similar manner to the 1-dimensional case, using Hooke's law and Newton's 3^{rd} law to match the x,y and z components of the external forces to reach the equation:

$$f_{ext} = BGB^{T}x \tag{5.25}$$

$$= Ku \tag{5.26}$$

where K is the stiffness matrix and u is the $3N \times 1$ vector of node displacements. Equation (??) is one of the central equations in mechanical engineering (via the Finite Element Method) and takes the form of a discrete Green's function describing an impulse-response. We can use the same arguments as for the 1-dimensional case, where we solved for x and then crucially restricted the input force to being either a pure stretch or compression in the edge-space and converted the output node displacement to a series of edge variables so that they were invariant to rigid body motion of the entire protein structure:

$$e = B^{T}K^{\dagger}Bf_0 \tag{5.27}$$

$$f_{\text{out}} = GB^{T}K^{\dagger}Bf_0 \tag{5.28}$$

now using e to represent the edge length changes to avoid confusion with the y variable used previously for node displacements. Potential energy changes can also be determined as before.

Three centre interactions

In the preceding work, we showed how the equation for bond-to-bond propensity could be re-derived in a mechanical context so as to describe how an input force is propagated across a protein structure modelled as an elastic network. Each of the springs represented an interaction between two nodes, which may be chosen to be atoms or residues in elastic models. However, if we use the procedure described in Chapter 4 to construct our bond network, we are able to identify specific chemical interactions using the software FIRST, such as covalent bonds and hydrophobic interactions. We know that in addition to bond interactions, chemistry in many cases places further restrictions on the *angle* between two bonds, for example an sp^3 hybridized carbon atom has an preferred bond angle of 109.28°. As described in the previous section however, there are no such restrictions on the angle between springs and as such, relative to a real protein system, our elastic model is likely too "floppy". We therefore introduce, using the same linear framework, a set of angle interactions within the structure.

We start this time from the distance between the nodes i and k. As before we want to see how this distance changes in relation to the displacements of each of the nodes:

$$e = |r_{ik}| - |r_{ik,0}| \tag{5.29}$$

$$|r_{ik}|^2 = |r_{ik,0}|^2 + 2x_iX_i - 2x_iX_k - 2x_kX_i + 2x_kX_k + \cdots + \cdots + O(x^2) \tag{5.30}$$

However here we have just obtained the same expression as for the two centre case and thus is no different to simply placing another spring between nodes *i* and *k*. The key step then is to additionally *restrict* the lengths of bonds (*i, j*) and (*j, k*), so that we only allow motions that change the angle but not the lengths of the bonds.

$$\begin{aligned} |r_{ij}|^2 &= |r_{ij,0}|^2 \\ |r_{jk}|^2 &= |r_{jk,0}|^2 \end{aligned}$$

When we expand both of these equations, we obtain the relations:

$$\begin{aligned} 2x_jX_j - 2x_jX_i - 2x_iX_j + 2x_iX_i + \cdots &= 0 \\ 2x_jX_j - 2x_jX_k - 2x_kX_j + 2x_kX_k + \cdots &= 0 \end{aligned}$$

which can be substituted in Equation (5.30):

$$|r_{ik}|^2 = |r_{ik,0}|^2 + 2\big[(X_j - X_k)x_i - (2X_j - X_i - X_k)x_j + (X_j - X_i)x_k + \cdots\big] + O(x^2, y^2, z^2) \tag{5.31}$$

Completing the square and square rooting both sides as in the two centre case results in an expression for the extension of the three centre (*i, k*) interaction:

$$e = \frac{(X_j - X_k)x_i - \big[(X_j - X_i) + (X_j - X_k)\big]x_j + (X_j - X_i)x_k + \cdots}{|r_{ik,0}|} \tag{5.32}$$

and again in a more convenient vector notation:

$$e = \begin{pmatrix} \frac{r_{jk}}{r_{ik,0}} & -\frac{r_{ji}+r_{jk}}{r_{ik,0}} & \frac{r_{ji}}{r_{ik,0}} \end{pmatrix} \begin{pmatrix} u_i \\ u_j \\ u_k \end{pmatrix} \tag{5.33}$$

Now our incidence matrix has three entries per row as might be expected now that we are relating three nodes. Once again, we have assumed linearity and can construct our stiffness matrix for the set of angle interactions we choose.

Four centre interactions

The natural extension of this is to also consider four centre interactions, which in the language of chemistry, correspond to *dihedral* interactions. These are particularly important for double bonds, which restrict rotation around the bond. The procedure is identical to the three centre case, except this time we need to restrict the lengths of the three bonds (*i, j*), (*j, k*) and (*k, l*) and additionally the two angles (*i, k*) and (*j, l*). The final expression for the extension is:

$$e = \begin{pmatrix} -\frac{(r_{ik}+r_{ij})}{|r_{il,0}|} & -\frac{(r_{ji}+r_{jk}+r_{jl})}{|r_{il,0}|} & -\frac{(r_{kl}+r_{ki}+r_{kj})}{|r_{il,0}|} & -\frac{(r_{ik}+r_{ik})}{|r_{il,0}|} \end{pmatrix} \begin{pmatrix} u_i \\ u_j \\ u_k \\ u_l \end{pmatrix} \tag{5.34}$$

Note that each of the n-centre interactions corresponds to motion in (n-1) dimensions: the two centre interactions (or bonds) describe motions along a line, three centre (angle) interactions are movements of the nodes in the plane, whilst four centre (dihedral) interactions capture the motions in and out of the plane defined by the four atoms. We therefore do not need to go further and consider five centre interactions or above, as these would simply accord with combinations of the previous n-centre interactions.

5.4 Rigidity and Infinitesimal rigidity

We noted previously that it is important when constructing mechanical models to make sure they have both translational and rotational invariance. Mathematically, this equates to the stiffness matrix for the system having six 0 eigenvalues: three for translation and three for rotations, where the value of the eigenvalue represents the zero energy cost associated with that motion. Depending on the structure of the model, it is in fact possible to have more than the six 0 eigenvalues, which equate to additional motions of the model that lead to no change in the length of any of the springs. The study of such system is the basis of Rigidity Theory, which has a long history stretching back to Lagrange and Maxwell.

Rigidity

To introduce the basic ideas of rigidity theory, we work initially in 2-dimensions, where any mechanical system possesses three rigid motions: two translational modes in the x-y plane and a rotational mode about an imaginary z axis passing through the plane. When considering the rigidity of an object, we no longer imagine the edge variables as springs that can be deformed, but instead as *hard constraints* that we cannot violate. Equivalently, we only allow motions of the object that do not change the length of any of the edges.

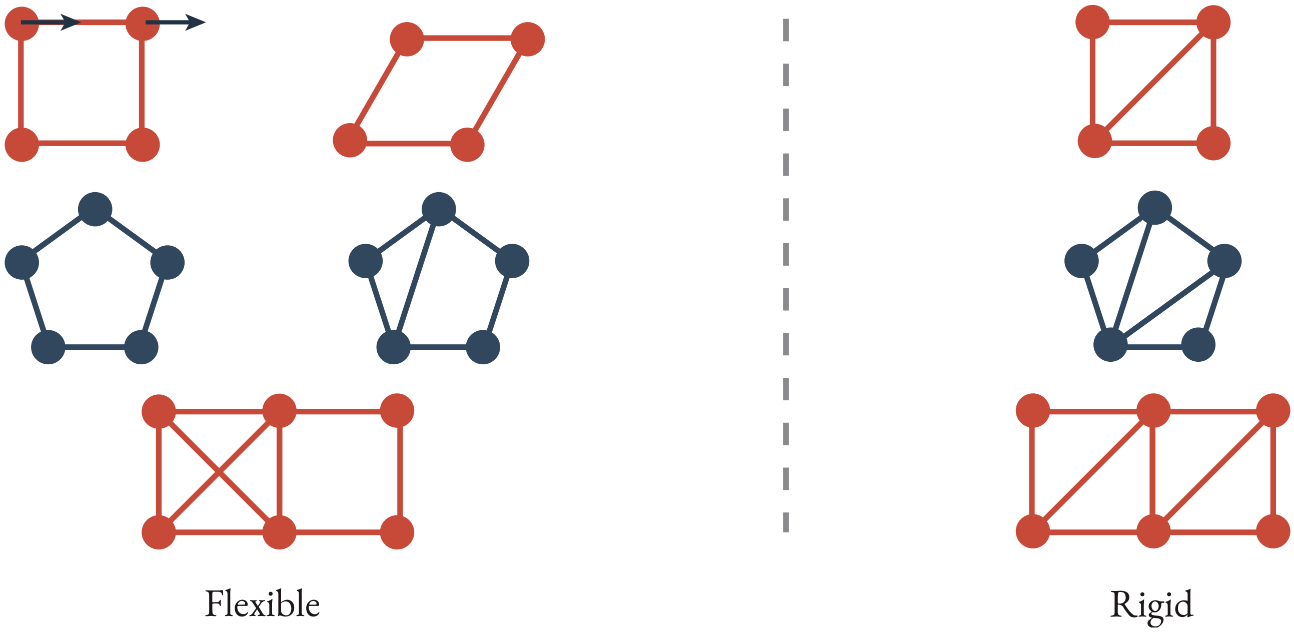


Figure 5.2: Mechanical networks can be classified as being either *rigid* or *flexible*. A flexible structure is one that can be *deformed* at zero energy cost. That is, without changing the length of any of the edges.

A triangle in the 2D plane cannot be deformed in any way, so its only rigid motions are the universal translations and rotation. However, a square can be deformed if we "push over" the top edge so that the structure resembles a rhombus, where all the sides have unchanged length compared to the square (Fig. 5.2). If we place an additional edge across one of the diagonals of the square, we form a structure that is now rigid. Meanwhile a pentagon requires two additional edges to make it into an undeformable shape. This pattern can be formalised as Laman's Theorem:

Theorem 5.1. *The edges of a graph G = (V, E) are independent in two dimensions if and only if no subgraph G' = (V, E) has more than 2N' - 3 edges, where N' is the number of vertices in G'.*

Corollary 5.1. *A graph G = (V, E) with 2N - 3 edges, where N is the number of vertices in G, is rigid if and only if no subgraph G' has more than 2N' - 3 edges.*

In other words, in order for a 2-dimensional structure to be rigid, we must have at least $2N - 3$ constraints but that alone is not a sufficient condition. Using a trivial example, if we had a square with 4 edges and wanted to satisfy Laman's Theorem, we would need $8 - 3 = 5$ edges. If we then added that edge "on top" of one of the existing edges, we would have satisfied the counting rule (which is known as Maxwell counting) but would still have a framework that was not rigid. The added edge must be *independent*. A more realistic example is shown in Fig. 5.2, which contains two squares joined by an edge. The Maxwell counting term $2N - 3$ is derived from the fact that a system of N nodes in 2-dimensions has 2N degrees of freedom. Each independent constraint added removes a degree of freedom and we have the three rigid body motions already so we require $2N - 3$ constraints to "freeze" the system.

Determining the rigidity of a structure in 2-dimensions is possible via the computationally efficient method known as the Pebble Game, which utilises Laman's Theorem to turn the determination of rigidity into a counting problem. Here of course we are interested in the 3-dimensional problem. The Maxwell counting rule for 3-dimensions is $3N - 6$, which accounts for the six rigid motions described earlier. Unfortunately, whilst Laman's Theorem holds as a *necessary* condition in 3-dimensions, it is no longer sufficient, with typical example of such a structure that satisfies Laman's Theorem in 3-dimensions but is nevertheless flexible is the "double banana"[98].

However, if we add the additional condition that every pair of two centre constraints is accompanied by an *angle constraint*, then Laman's Theorem appears to hold. These structures are known as *body-bar* frameworks and whilst Laman's Theorem has not been proven to hold generally, there have been no examples discovered in over 20 years that contravene the rule. The flexibility of these structures can then be determined by a 3-dimensional equivalent of the Pebble Game, which is the premise of the software FIRST[102]. In the case of proteins, the requirement to include angle constraints is not too arduous; as explained earlier, chemistry places restrictions on the angles of many pairs of bonds anyway. Again, due to the conversion of the rigidity problem into a combinatorial one, FIRST is highly efficient, running in $O(n \log n)$ time, which has allowed for its application to systems as large as viral capsids.

Infinitesimal rigidity

Whilst FIRST has been successfully applied to a range of proteins to examine their flexibility, if we wish to have more freedom over the set of constraints that we include in our structures, we must turn to other methods. We can write the rigidity problem for M constraints explicitly as a set of M equations:

$$|p_i - p_j|^2 = c_{ij},\ (i,j) \in \mathcal{M}, \tag{5.35}$$

where p_i is the 3×1 position vector of node i. Solving this set of M nonlinear equations directly is usually infeasible for anything other than very small systems. An alternative approach is *Infinitesimal Rigidity*.

We begin by taking the derivative of both sides of Equation (5.35) with respect to time t for all constrained pairs:

$$(p_i - p_j) \cdot (u_i - u_i) = 0,\ (i,j) \in \mathcal{M}, \tag{5.36}$$

with $u_i = \frac{dp_i}{dt}$. We expand out the brackets:

$$(p_i {-}_j)u_i - (p_i - p_j)u_j = 0,\ (i,j) \in \mathcal{M}, \tag{5.37}$$

and write in vector form:

$$Ru = 0 \tag{5.38}$$

R is called the Rigidity matrix and each row represents a single constraint. For example, a three node system with each pair of nodes joined by an edge would have the Rigidity matrix:

$$R = \begin{pmatrix} p_1 - p_2 & p_2 - p_1 & 0 \\ 0 & p_2 - p_3 & p_3 - p_2 \\ p_1 - p_3 & 0 & p_3 - p_1 \end{pmatrix} \tag{5.39}$$

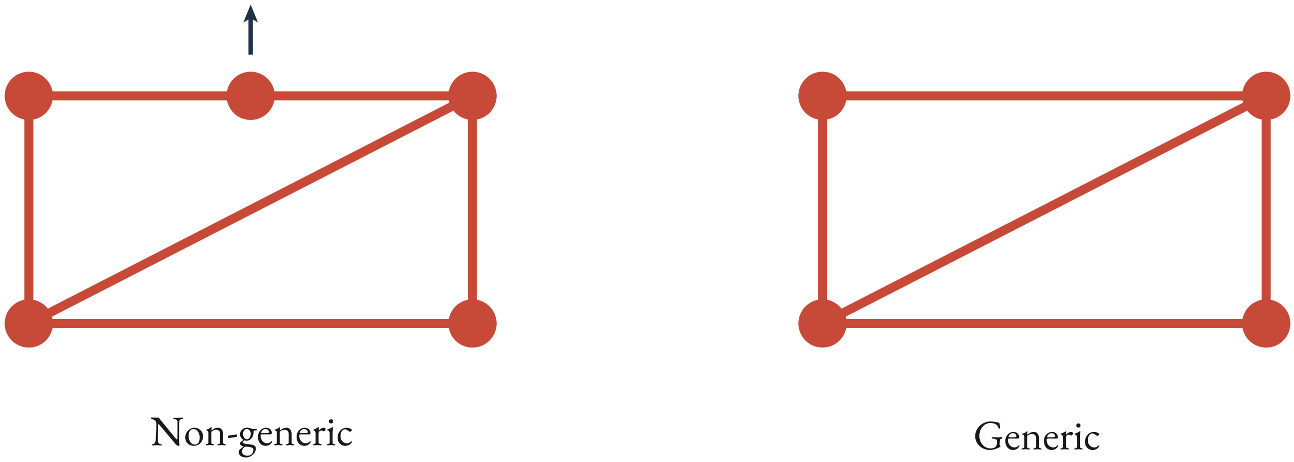


Figure 5.3: The network on the left is *rigid* but is not *infinitesimally rigid* as it has an infinitesimal flex. Structures such as this are described as *non-generic* and generally occur when colinear edges are present in the structure. The structure on the right by contrast is generic and is both rigid and infinitesimally rigid.

Our task then is to find those infinitesimal motions of the structure that have zero cost, or in other words, calculate the set of vectors that comprise the *nullspace* of the Rigidity matrix. It has been shown that infinitesimal rigidity and rigidity are equivalent[242] in the case of *generic* frameworks (see Fig. 5.3), which is the case for all of the structures examined in this thesis and therefore we are free from now on to refer to the rigidity of structures. There are many different ways to compute the nullspace of a matrix but here we use the somewhat standard approach of performing a singular value decomposition (SVD) of the Rigidity matrix:

$$R = U\Sigma V^T \tag{5.40}$$

In the case where we have more edges than nodes (M > N), the nullspace spans the rows of the right singular matrix V^T that correspond to the singular values in the diagonal matrix Σ whilst if M < N, we take the bottom (N - M) rows of V^T in addition to any rows associated with singular values.

Rigid cluster decomposition

Once we have our set of infinitesimal motions, or nullspace of R, we need a way to then ascertain those groups of atoms within the structure that act as rigid bodies. We use the algorithm described in Ref. [46], which has proven to be computationally efficient:

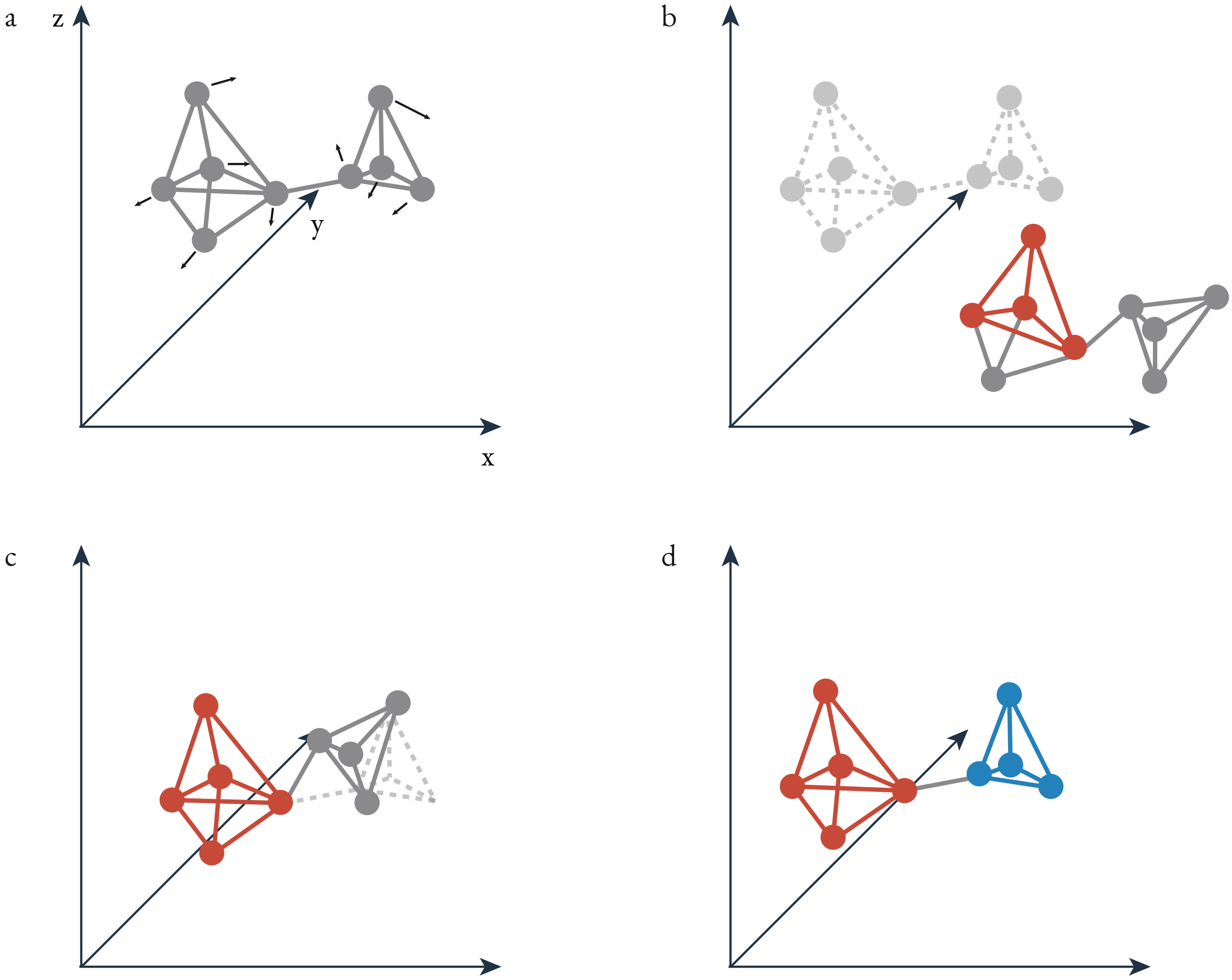


Figure 5.4: The steps for the rigid cluster decomposition algorithm. For each of the trivial infinitesimal motions, such as the one in a) the atoms are moved by a small distance along each $3N \times 1$ vector to a new position b). A rigid tetrahedron of atoms is selected in the new position then in c) this is moved back to its original position. Any atoms that also return to their original position at the same time (for all infinitesimal motions) are part of the same cluster. d) The process is repeated until all atoms are clustered into rigid regions or are assigned as floppy.

1. Identify a set of 4 atoms, T, that form a fully connected tetrahedron.

2. Translate the coordinate frame to the centre of the set T.

$$p_k := p_k - \frac{1}{4}\sum_{k\in T} p_k$$

3. Transform the three coordinate axes so that they correspond to the principle axes of the set T:

$$p_k := Kp_k$$

K is the rotation matrix whose rows are the eigenvectors of the matrix I:

$$I_{\alpha\beta} = \sum_{k\in T}(|p_k|\delta_{\alpha\beta} - p_{k\alpha}p_{k\beta}), \text{ where } (\alpha,\beta) = (x,y,z)$$

4. Generate the trivial motions in this new coordinate frame: three rotations (r_k^x, r_k^y, r_k^z) and three translations (t_k^x, t_k^y, t_k^z) for all of the atoms of the structure.

$$r_k^\alpha = p_k \times \hat{e^\alpha}; \quad t_k^\alpha = \hat{e^\alpha}$$

5. Transform the trivial motions back into the starting coordinate frame:

$$r_k^\alpha := K^T r_k^\alpha; \quad t_k^\alpha := K^T t_k^\alpha$$

6. Compile the trivial motions for each atom into column vectors so we have three 3N-dimensional translations t^α and three rotations r^α. Normalise each of these trivial motions:

$$r^\alpha := \frac{r^\alpha}{|\mathrm{r}_\mathrm{T}^\alpha|}; \quad t^\alpha := \frac{t^\alpha}{|\mathrm{t}_\mathrm{T}^\alpha|}$$

using the magnitude of the 12-dimensional vectors associated with the set T. Now the set of six 12-dimensional trivial motions of the set T are orthonormal.

7. The set of displacements of each of the atoms relative to the set T can then be calculated by returning the set T to its initial position:

$$\Delta p^\gamma = \mathrm{q}^\gamma - \sum_\alpha(\mathrm{q}_T^\gamma \cdot r_T^\alpha)r^\alpha - \sum_\alpha(\mathrm{q}_T^\gamma \cdot t_T^\alpha)t^\alpha$$

where we now use γ additionally index over the set of trivial motions.

8. For each atom, calculate its absolute displacement in space away from its initial position due to the infinitesimal motions. If the maximum displacement of the atom over the entire set of infinitesimal motions is below a chosen small threshold value then we say that atom is part of the same rigid cluster as the set T:

$$\max_{\gamma} |\Delta p_k^{\gamma}| < \delta$$

The value of δ used in this work is 10^{-4}

5.5 PDK1

3-phosphoinositide-dependent kinase 1, or PDK1, is a member of the AGC group of of Ser/Thr kinases and is itself an upstream kinase for other members of the AGC family, such as protein kinease A (PKA) and protein kinase B (PKB or Akt). As a consequence of its role in the PI3K pathway[167, 77], PDK1 has been implicated in a number of cancers[59, 192] and as such is considered a potential drug target[78]. PDK1 contains an allosteric site: the PDK1-interacting-fragment (PIF) pocket that binds to a hydrophobic motif on downstream substrates and activates PDK1[94]. Sadowksky *et al*[202] investigated the effect of binding at the allosteric site using small molecule sulphides that were able to increase the activity of PDK1. We use an active form of PDK1 as we are interested in the plausibility of propagation of strain as a mechanism for the *extended conformational selection* model.

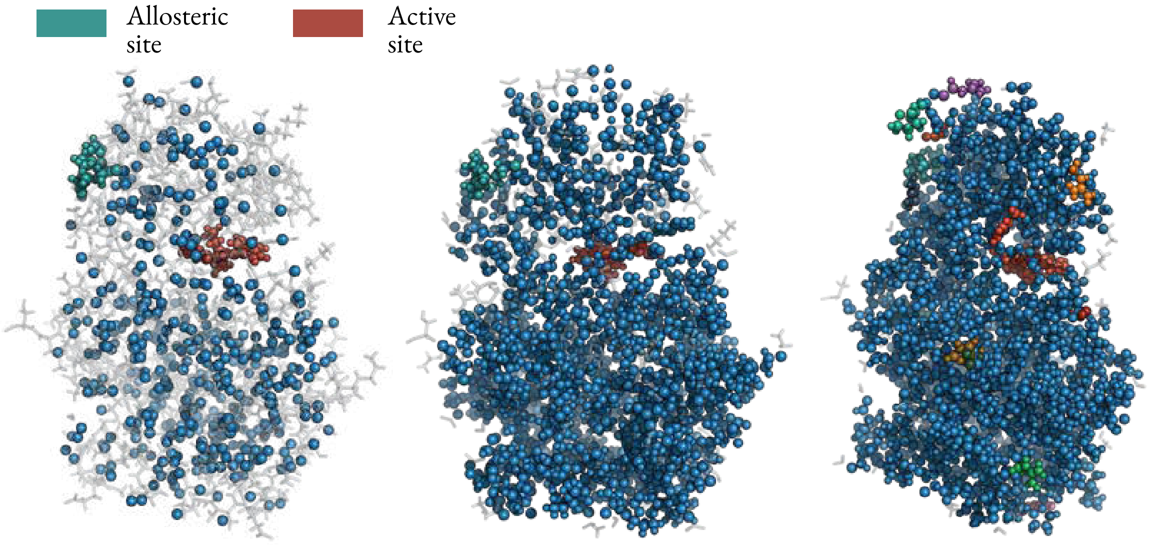
Here, we use the *elastic response* method to model the allosteric mechanism as a possible propagation of strain induced by the formation of bonds between the ligand and the protein. An atomistic network structure of the protein is constructed in a similar fashion to the *bond-to-bond propensities* method of Chapter 4, with hydrogen atoms added to the crystal structure obtained from the protein data bank using the program Reduce. The presence of the various bond types (covalent, hydrogen and hydrophobic interactions) is determined by the software FIRST. Now, instead of using bond energies to weight the network edges, we assign values to the edges according to their presumed spring constant. Using the Amber15fb force field[238], we assign spring constants to the (relative) correct order of magnitude:

The reason we do not use the exact values from the force field is that hydrophobic interactions are here modelled as two-centre interactions, whilst in molecular dynamics simulations they result from the presence of either implicit or explicit water that favours interactions to polar regions of the protein. As such, no such

Table 5.1: Springs constants for each of the elastic network interactions.

Interaction	Spring constant (relative)
Covalent	100
Hydrogen	10
Hydrophobic	1
Angle	1
Dihedral	0.1

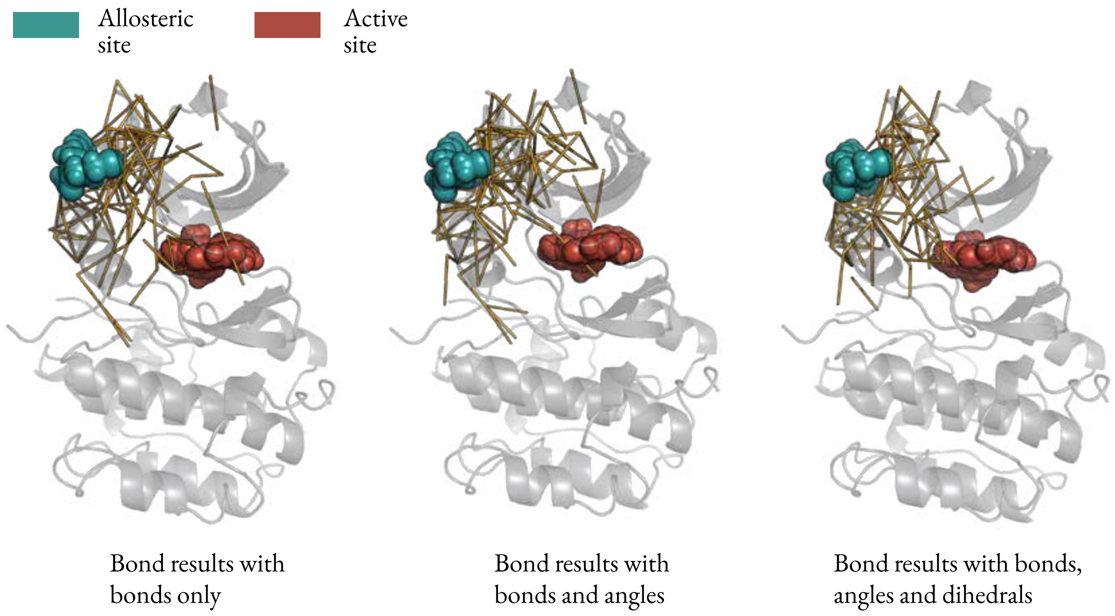
spring constant values exist for hydrophobic interactions. In fact in the Amber force field, hydrogen bonds are largely derived from electrostatic contributions and so again it is difficult to assign an exact value. In contrast to *bond-to-bond propensities*, we do not apply further statistical methods such as quantile regression here as we wish to model the actual, raw displacement felt by each of the interactions as would be the case for the real protein. We also apply *infinitesimal rigidity* to the network representation of the protein in order to decompose the structure into its rigid subparts.



Bonds only Bonds and angles Bonds, angles and dihedrals

Figure 5.5: Infinitesimal rigidity results for PDK1 where each cluster has a different colour and "floppy" atoms are shown in transparent grey. a) Only bonds included as constraints, leading to a single large cluster in blue with all other atoms floppy. b) Angle constraints included. c) Dihedral and angle constraints included.

We obtain the output displacement for all edges within the protein but remove the source edges from the results as these invariably score much higher. In Fig. 5.6 we show the top 2% of bonds by absolute length change (i.e. we do not discriminate between bond stretching or compression) in three scenarios: firstly where only the two-centre bond interactions are used to construct the elastic network as is traditionally the case with elastic network analysis of proteins. Then, we also construct networks where angle constraints between pairs of covalent bonds are included and finally ones in which dihedral angle constraints from double bonds are modelled. The choice is 2% here is arbitrary, however given the highest scoring interaction in the *bonds only* case (the hydrophobic interaction between Lys120 and Asn122) experiences a distance change of 0.766, and those interactions outside the top 2% experience a change of less than 0.01 it seems a reasonable assumption that we can neglect the output of any bonds scoring less than this. From the infinitesimal rigidity results presented in Fig. 5.5, even in the case where only two-centre constraints are included, the allosteric site and the region around the active site: Val96, Lys111, Tyr161, Ala162, Thr222, and Asp223 all form interactions with the active site and appear in the large cluster, with Leu88 the only active site residue that has no atoms within the rigid cluster. When 3-centre and 4-centre constraints are included, almost all of the protein becomes part of the rigid cluster, with some small groups of clustered atoms around the surface in the 4-centre case. It appears then at least plausible that propagation of strain may be emitted from binding at the allosteric site towards the active site, particularly through the rigid cluster formed by the 2-centre interactions that contains a smaller subset of the atoms in the protein.



Bond results with bonds only

Bond results with bonds and angles

Bond results with bonds, angles and dihedrals

Figure 5.6: The top 2% of 2-centre interactions (bonds) by absolute change in length in PDK1 (PDB code: 3ORZ[202]). a) Only bonds are included in the elastic network b) 3-centre (angle) interactions are included in the network but we calculate the *output displacement* of the bonds as we are interested in the potential change in bonding pattern driven by ligand binding. c) Dihedral angles are additionally included and again only output bond displacements are shown.

However the results from the *elastic response* results are less supportive of the possibility of long range mechanical transfer. As can be seen in Fig. 5.7, we obtain a linear decrease in the log edge displacement as distance increases (correlation coefficient = -0.603, standard error = 0.0022) even when angles and dihedrals are included, suggesting an exponential decrease decay in the mechanical propagation. Such a response is similar to random networks or continuum media[245] and is not suggestive of a structure optimised for directed perturbations. The two highest scoring interactions by some distance are the Lys120 - Asn122 (0.766) and the Val124-Pro125 (0.437) hydrophobic interactions with the next highest displaced bond (Val124-Val127) scoring just 0.189. Both interactions are within 5Å of the allosteric source site, with the active site around 17Å away. Indeed the highest scoring interactions involving active site residues are two Lys111-Phe157 hydrophobic interactions, which are displaced by 0.0122 and rank 130th and 131st. Of the top 2% (149 out of 7391) of interactions by output displacement, all but 4 are hydrophobic interactions which is unsurprising given they have the weakest spring constants but appears to lead to those weak interactions near the allosteric site effectively acting like a sponge for the inputted force and preventing long range transfer of displacement. We see this clearly when changing the force constant of the hydrophobic interactions to be 10 (the same as the hydrogen bonds) as in Fig. 5.8, the range of the propagation increases. However, as discussed, it is difficult to rationally assign spring constant values to the hydrophobic interactions without inadvertently biasing the analysis. Indeed, when all spring constants are simply set to be equal (which is unrealistic physically), the propagation extends even further, suggesting the method is highly sensitive to the set of spring constants chosen. At the very least, we can see that topology alone is not a dominant enough factor to determine if a mechanical explanation for allostery is plausible, the particular values of the edge variables are also crucial. Certainly there does not appear to be strong evidence here that the allosteric effect exhibited by PDK1 is mediated by traversal of strain energy. The results also appear to support of the use of a coarse-grained, residue level description of the protein when modelling using an elastic network as the residue-residue interactions result from an "averaging" of the total set of interactions between the atoms of the residues. Modelling at the atomistic level, as in molecular dynamics, requires a highly accurate parameterisation of the set of bonds.

5.6 h-Ras

The *elastic response* analysis was also performed on h-Ras, a GTPase involved in the regulation of growth factor mediated cell division[155] and mutations in h-Ras have been implicated in a range of cancers[38, 37]. Buhrman *et al*[29] discovered that h-Ras can be allosterically modulated by calcium acetate and postulated a network of hydrogen bonds that linked this allosteric site to the distal active site at catalytic residue Gln61 by comparing changes between the active and inactive states. Again, we decompose the protein network structure (PDB code: 3K8Y) into its rigid subunits via the infinitesimal rigidity algorithm and model the effect of a mechanical allosteric perturbation via elastic response.

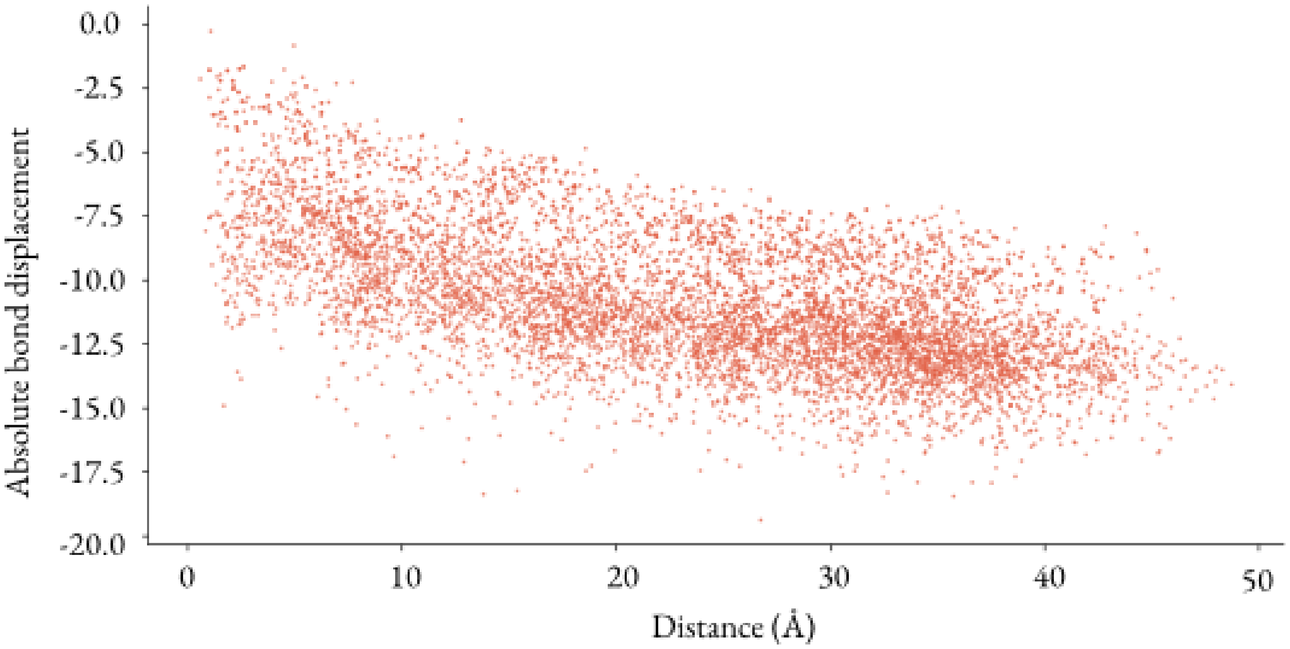
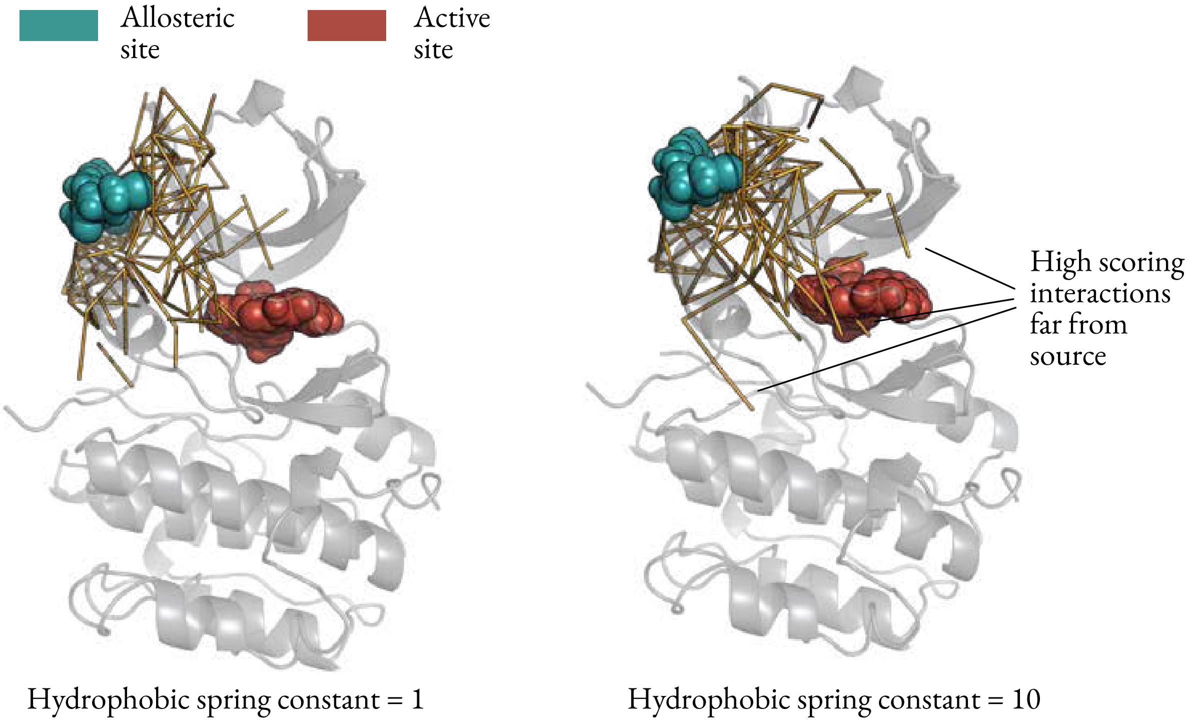


Figure 5.7: Plot of how log absolute displacement of interactions varies with increasing distance from the allosteric source site. A linear decrease is seen with slope -0.142 (correlation coefficient = -0.603, standard error = 0.0022), suggesting that the effect of the perturbation decays exponentially away from the allosteric site. Whilst we note the possible existence of two bands of points (top and bottom), when the data clustering algorithm DBSCAN[65] is run on the data set, no significant partition of the data is seen.



Hydrophobic spring constant = 1

Hydrophobic spring constant = 10

Figure 5.8: Changing the value of the spring constant for the hydrophobic interactions, we see a greater propagation of strain is possible but there is still an exponential fall off with distance.

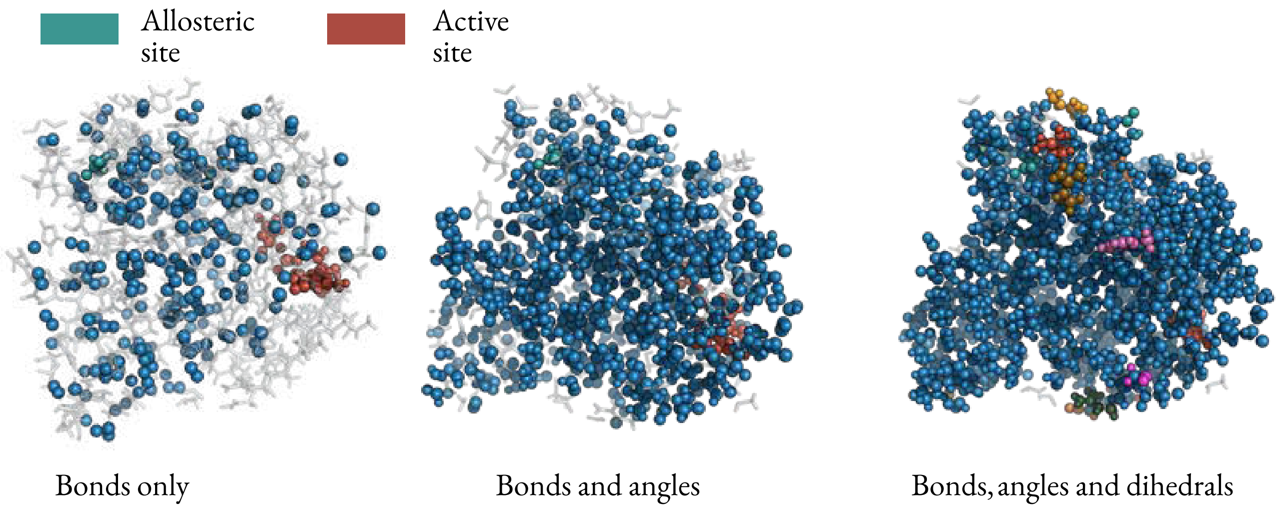
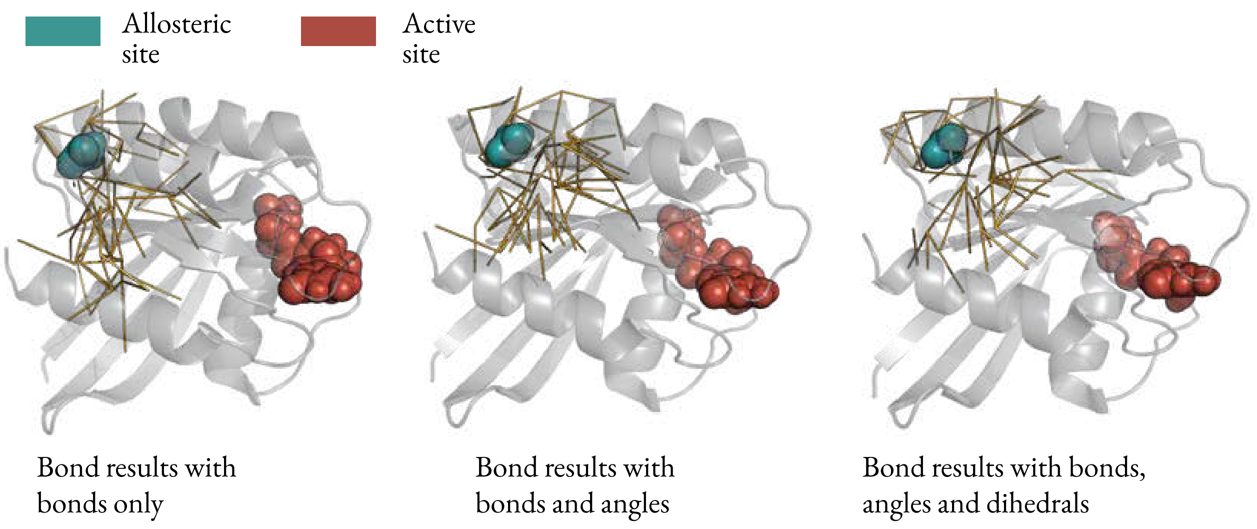


Figure 5.9: Infinitesimal rigidity results for h-Ras where each cluster has a different colour and "floppy" atoms are shown in transparent grey. a) Only bonds included as constraints, leading to a single large cluster in blue with all other atoms floppy. b) Angle constraints included. c) Dihedral and angle constraints included.

Whilst one of the atoms (the methyl carbon) of the allosteric ligand ACT forms part of the large cluster in the bonds only case in Fig. 5.9a, neither the active site ligand GNP, nor any of the active site residues Gln61, Thr35 or Tyr32 appear in this cluster. Even when both angles and dihedrals are included in the analysis (Fig. 5.9c), only the same atom from ACT forms part of the single large cluster, though now the entire GNP active site substrate is present in the cluster, as well as 10 out of the 18 atoms of the Gln61 active site residue. However, once dihedrals are included, 2063 out of the 2673 total atoms in the protein are part of the large cluster, which is at odds with the idea that fast pathways exist in the protein just below the percolation threshold that lead to specific, directed flow[132].

Similarly to the PDK1 case, the elastic response results (Fig. 5.10) do not suggest the presence of pre-existing pathways in the protein between allosteric and active sites, whether or not we include angle and dihedral constraints in the analysis or not. Again, the majority of the highest scoring bonds by absolute displacement are the weak hydrophobic interactions, rather than the pathway of hydrogen bonds suggested by Buhrman. Absolute displacement similarly falls away exponentially with distance, with values outside the top 2% of bonds by rank being displacement less than 2.5% of the value of the top scoring bonds, and there does not appear to be any particular directionality to the mechanical propagation.



Bond results with bonds only

Bond results with bonds and angles

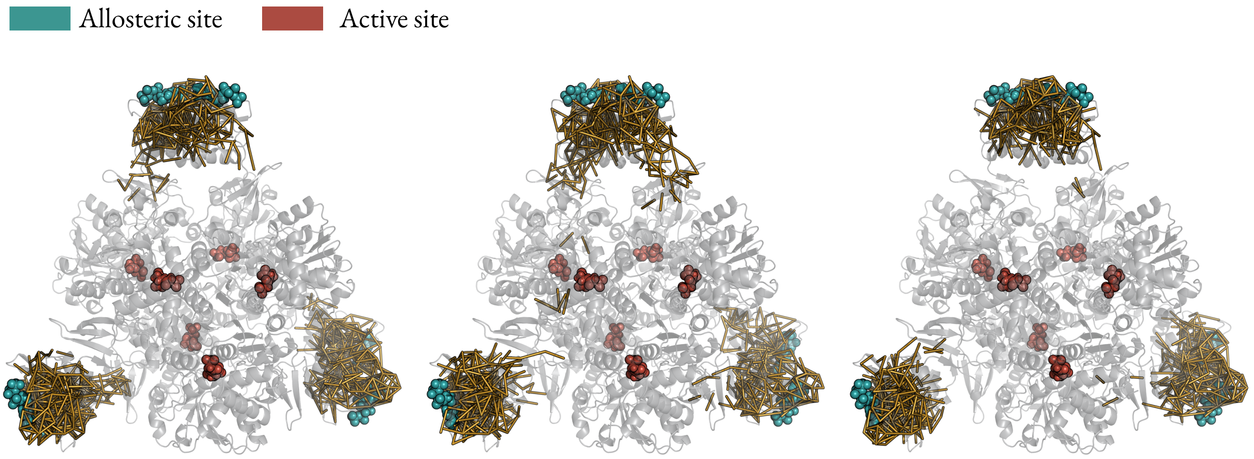
Bond results with bonds, angles and dihedrals

Figure 5.10: The top 2% of 2-centre interactions (bonds) by absolute change in length in h-Ras (PDB code: 3K8Y[29]). a) Only bonds are included in the elastic network b) 3-centre (angle) interactions are included in the network but we calculate the *output displacement* of the bonds as we are interested in the potential change in bonding pattern driven by ligand binding. c) Dihedral angles are additionally included and again only output bond displacements are shown.

5.7 ATCase

We also calculated the elastic response for the ATP-protein interactions in ATCase. However we were unable to perform rigidity analysis due to the large size of ATCase; with 80,000 atoms, the infinitesimal rigidity algorithm involves calculating the SVD of a matrix of size $240,000 \times 240,000$. Whilst software does exist for calculation of eigenvalues in large, sparse matrices via iterative Arnoldi methods[?], even these methods were unable to perform the decomposition in a reasonable time.

Much as with the previous two examples, we do not see significant propagation of strain from the allosteric sites occupied by ATP towards the active sites. Instead we observe a similar outcome to that of PDK (See Fig. 5.7) where the absolute bond displacement appears to fall off exponentially. Again, we perform the elastic response calculation with just two-center interactions included, and with additional three- and four-center interactions. Curiously, the longest range effects seem to occur when both two and three-center interactions are included, with the inclusion of the four-center dihedral interactions giving the shortest range results of the three cases. It can be seen in Fig. 5.11 that there are some high scoring edges nearer one of the active sites. One of these is an Arg234 - Asn113 hydrophobic interaction, and Arg234 has been shown to be important for R state stabilisation[225], though specifically through its hydrogen bond interactions with Glu50 rather than to Asn113. Additionally the nearby Arg167 - Asn132 hydrophobic interaction also scores highly, with Arg167 being one of the residues that interacts with the active site PALA residue. However, only a small



Bond results with bonds only

Bond results with bonds and angles

Bond results with bonds, angles and dihedrals

Figure 5.11: The top 2% of 2-centre interactions (bonds) by absolute change in length in ATCase (PDB code: 4KGV[42]). a) Only bonds are included in the elastic network b) 3-centre (angle) interactions are included in the network but we calculate the *output displacement* of the bonds as we are interested in the potential change in bonding pattern driven by ligand binding. c) Dihedral angles are additionally included and again only output bond displacements are shown.

number of other bonds that are more than 20Å from the allosteric sites also score highly and once again, it appears unlikely that long range mechanical pertubations are a plausible mechanism for communication between allosteric and active sites.

5.8 Regular lattice

We also performed the elastic response calculation in a regular hexagonal lattice, which acts as a null model, in order to see if the propagation may still in fact be significant in comparison to a structure not optimised for long range mechanical transfer. We used a 16 × 16 × 16 lattice with 4096 "atoms", which most closely corresponds in size to PDK1 and chose 18 atoms to act as the allosteric ligand and 28 edges to represent the bonds between the ligand and the "protein", as shown in Fig. 5.12.

Again, we display the top 2% of edges by absolute edge displacement as below this, the values quickly become negligable. Much as in the case of the real protein structures, the propagation of strain falls off in an exponential manner (Fig. 5.14) rather than exhibiting any long range effects. In contrast to the proteins however, the pattern of strain here is symmetrical, decaying in a spherical shape around the allosteric site and the values for the edge displacements are more tightly distributed due to the symmetry of the lattice (Fig 5.14). When we compare the results for PDK to the lattice, we can see that there is some heterogeneity

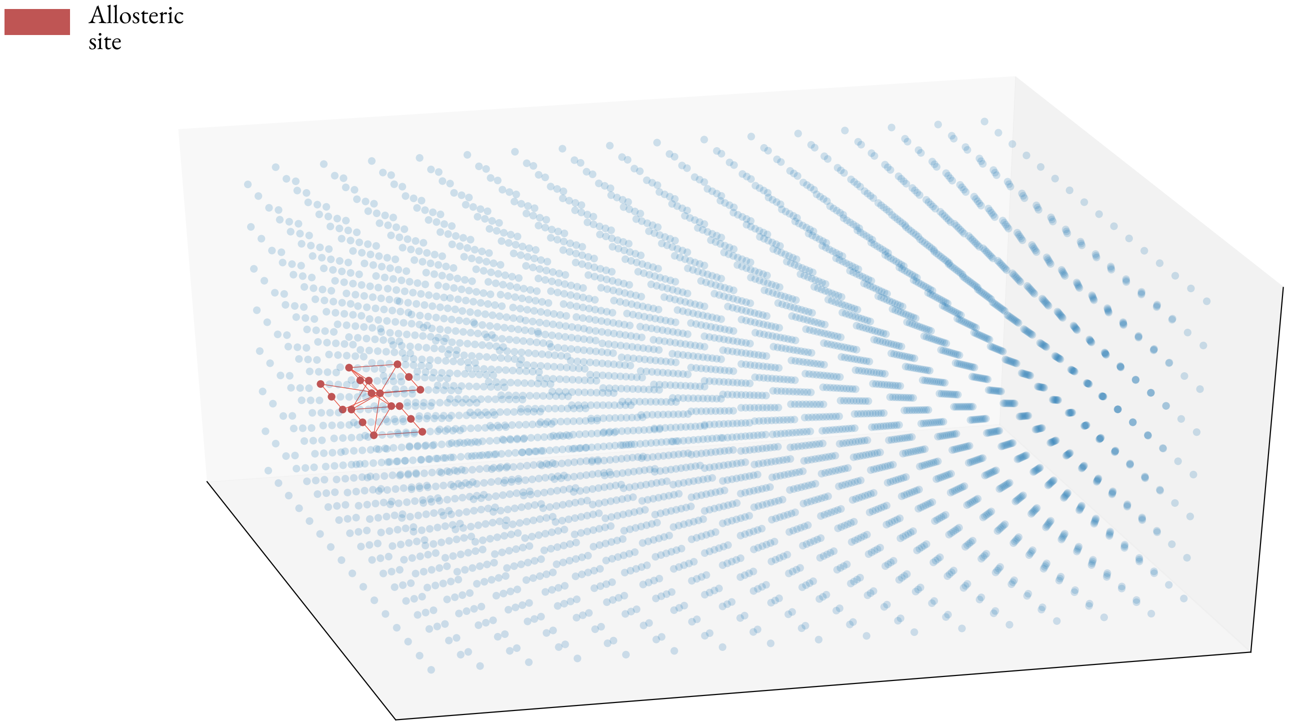


Figure 5.12: 28 atoms are chosen as the "allosteric ligand" along with 28 edges representing bonds between the ligand and the "protein", here portrayed as a 16 × 16 × 16 lattice.

in the propagation of strain away from the allosteric site, perhaps suggestive of some optimisation of longer range transfer, the values for the edge displacements appear to decay too quickly to make mechanical transfer between the allosteric and active sites plausible.

5.9 Elastic response conclusions

In this part of the chapter, we investigated whether a method, *elastic response*, that could model mechanical perturbations, could explain how binding of a ligand at an allosteric site may lead to perturbations at a distal active site. The idea of strain propagating from the allosteric site in a directed manner towards the active site is often invoked as a mechanism of allostery[235, 187]. By applying the method to the highly symmetrical hexagonal lattice, we see that long range transfer does not occur in that case and could not just result as an artefact of all elastic network structures, leading to the hypothesis that the protein structure may be optimised to facilitate mechanical communication between the allosteric and active sites.

However we do not find strong evidence this is possible, with perturbations in the proteins studied tending to die away exponentially and in an essentially isotropic manner. Traditionally elastic models include only 2-centre interactions, which in this case correspond to chemical bonds. In order to model the full set of con-

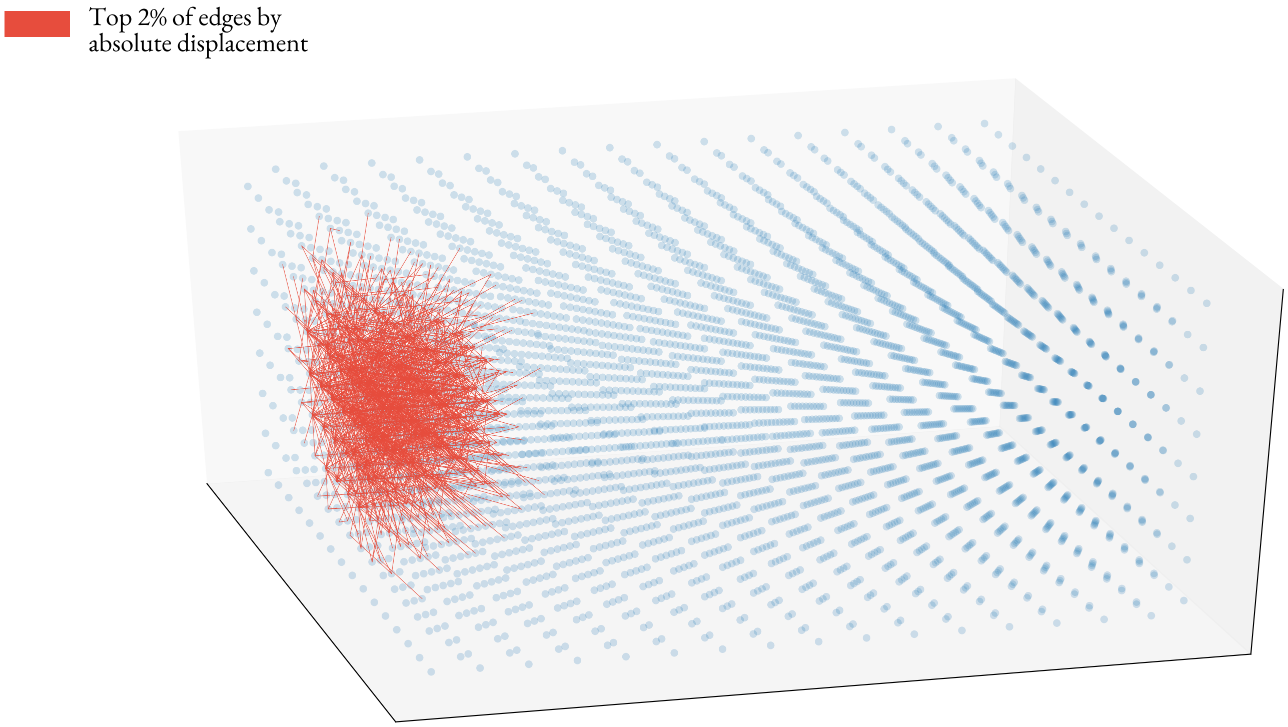


Figure 5.13: Shown are the top 2% of edges by absolute displacements. All of these edges are clustered in a spherical region around the "allosteric site", demonstrating an absence of any long range mechanical transfer.

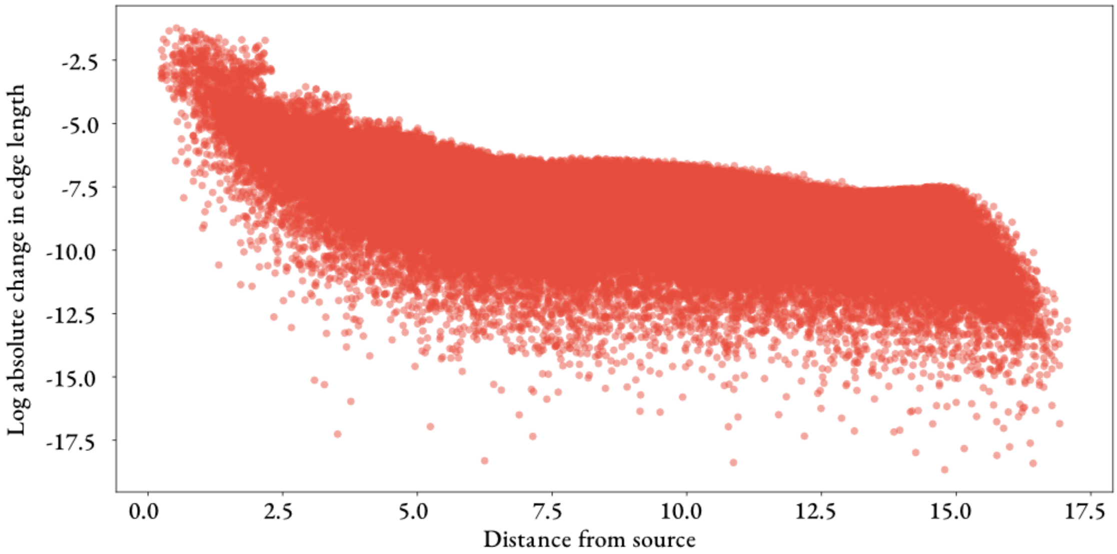


Figure 5.14: Similar results to those obtained in the case of PDK are found for the hexagonal lattice whereby absolute change in bond length falls away exponentially with distance. However, the results are more tightly distributed than in the PDK case, most likely due to the symmetry of the lattice so that points at a similar distance from the allosteric also have a similar local packing environment.

straints that chemistry places on the protein, we also devised expressions for angle and dihedral interactions, which were presumed to increase the stiffness of the overall structure and make the elucidation of long range effects more likely. Despite this, even the inclusion of additional constraints did not lead to a convincing case for long range mechanical structural transfer. Infinitesimal rigidity analysis did show cases where both allosteric and active sites were present within the same rigid cluster, however the clusters tended to be large and cover much of the protein, rather than resembling fractal structures just below the percolation threshold that may be conducive to directed paths[132].

We note also that whilst here we do not cast our methods in terms of the dynamics of the protein (*elastic response* as formulated below considers a static picture of the elastic model and considers how an input force on a set of bonds leads to an output change in length of the bonds in the rest of the system), the equations we derive are similar to those used in entropic studies of allostery. Capetelli *et al*[32] applied their dynamic flexibility index (DFI) to residue-residue elastic models of proteins. They apply a perturbation to each residue in turn and measure the global response on the set of other residues. The response to the perturbation is written as:

$$[\Delta R]_{3N\times 1} = [H]^{-1}_{3N\times 3N}\,[F]^{-1}\,3N \times 1 \qquad (5.41)$$

which has precisely the form of Eq. (5.28) except in the node space rather than the edge space. Usually, in elastic models the Hessian H is precisely the stiffness matrix K but Capetelli *et al* in fact use a different form of H by constructing it from the covariance matrix C of a molecular dynamics simulation. They call the resulting matrix G which has the form: $G = C^{-1}$. A promising future direction for this work therefore would be to apply our edge-centric framework to a *dynamical* model of perturbations, using the matrix G as this would allow for a more principled means of constructing our Hessian as it would ultimately be derived from well tested molecular dynamics forcefields.

5.10 Residue-Residue interaction embeddedness

Another method that we have developed within the elastic network framework is *interaction embeddedness*, based on the general edge centrality measure of embeddedness introduced by Schaub *et al*[206] in the context of random walks on networks. Importantly, embeddedness is not the same as *betweenness centrality*[75], a popular measure of edge centrality. The shortest path for every pair of edges is calculated and the betweenness of an edge is equal to the number of shortest paths that pass through that edge. Instead, embeddedness is related to the property in electrical networks known as *effective resistance*[81]. Similarly to propensity, we reinterpret these measures in the case of a *mechanical* system.

If we first define the *bond-to-bond force transfer matrix* by grouping together terms in Eq. (5.28):

$$f_{\text{out}} = GB^T K^\dagger B f_{\text{in}} \tag{5.42}$$

$$f_{\text{out}} = G\mathbb{R} f_{\text{in}} \tag{5.43}$$

$$f_{\text{out}} = \mathcal{M} f_{\text{in}} \tag{5.44}$$

Embeddedness is then defined as $1 - \mathcal{M}_{bb}$, in other words we are interested in the diagonal elements of the transfer matrix. We can extract this entry for a particular bond b by setting the b^{th} entry of f_{in} to 1 (and all others to 0) and looking at the b^{th} entry of f_{out}. Physically then, the diagonal elements of $\mathcal{M}$ tell us what proportion of the force applied to a bond is actually transmitted to that bond versus what is redistributed over the rest of the elastic network. Those edges that are highly *embedded* are therefore those that have a high value of $1 - \mathcal{M}_{bb}$, or equivalently small values for the diagonal elements of M.

Given an input force on a particular edge, the corresponding diagonal element of $\mathbb{R}$ tells us the change in length of that edge, which (depending on the location of the spring within the network topology) is not necessarily the same as if the spring was isolated (note: this is *not* the same R as the rigidity matrix). The situation is therefore the mechanical analogue of *effective resistance* in electrical networks[81], also known as the resistance distance[120]. In the electrical case, the effective resistance is defined as the drop in potential difference across a wire given injection of a unit current; here the equivalent situation is the change in length of an edge given the application of a unit force to it. However, in the mechanical case, it is both the connectivity *and* the geometry of the network of the network (for two and three dimensions) that determines edge responses.

Another (equivalent) interpretation that is perhaps more natural from the point of view of protein dynamics is that the diagonal values of $\mathbb{R}$ equate to the expectation value for the interaction length changes $\mathbb{E}\left[y^2\right]$ and the diagonal values of $\mathcal{M}$, the average potential energy of the interactions, given the elastic network sitting in a heat bath. This can be seen from the following derivation, where we begin from the Langevin equation in the *overdamped* regime, as discussed in Chapter 3 such that the inertia of the system is damped out and we have Brownian motion as in Eq. (3.4):

$$\frac{dx}{dt} = -K(x - x_{\text{eq}}) + \phi(t) \tag{5.45}$$

$$x_t - x_{\text{eq}} = \int_{-\infty}^{t} \exp\left[K(t-s)\right] \phi(s)\, ds$$

$$\begin{aligned} Y_t &= B^T \left(x_t - x_{\text{eq}}\right) \\ &= B^T \int_{-\infty}^{t} \exp\left[K(t-s)\right] \phi(s)\, ds \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[Y_t Y_t^T\right] &= \int_{-\infty}^{t} \int_{-\infty}^{t} B^T \exp\left[K(t-s)\right] \left[\phi(s)\phi(\xi)^T\right]\, \exp\left[K(t-\xi)\right]^T B\, ds\, d\xi \\ &= \int_{-\infty}^{t} \int_{-\infty}^{t} B^T \exp\left[K(t-s)\right] \left[\delta(s-\xi)I\right]\, \exp\left[K(t-\xi)\right]^T B\, ds\, d\xi \\ &= \int_{-\infty}^{t} B^T \exp\left[K(t-s)\right]\, \exp\left[K(t-\xi)\right]^T B\, d\xi \\ &= \int_{-\infty}^{t} B^T \exp\left[K(2t-2s)\right] B\, d\xi \end{aligned}$$

Decomposing K into its eigenvalues and eigenvectors:

$$\begin{aligned} \mathbb{E}\left[Y_t Y_t^T\right] &= \sum_{i=1}^{3} N \int_{-\infty}^{t} B^T \exp\left(-\lambda_i \left(2t - 2\xi\right)\right) v_i v_i^T B d\xi \\ &= \frac{1}{N} \int_{-\infty}^{t} B^T 11^T B\xi + \sum_{i=2}^{3} N \int_{-\infty}^{t} B^T \exp\left(-\lambda_i \left(2t - 2\xi\right)\right) v_i v_i^T\, B d\xi \\ &= B^T \left[\sum_{i=2}^{3} N \frac{e^{-\lambda_i(2t-2\xi)}}{2\lambda_i} \Bigg|_{-\infty}^{t} v_i v_i^T\right] B = \frac{1}{2} B^T \sum_{i=2}^{3} N \frac{1}{\lambda_i} v_i v_i^T B \\ &= \frac{1}{2} B^T K^{\dagger} B \end{aligned}$$

The expectation for the potential energy of each interaction $g_b\, \mathbb{E}\left[y_b^2\right]$ are then the diagonal entries of $\frac{1}{2} G B^T K^{\dagger} B = M$, where we have multiplied each entry by its associated force constant g_b. Here, rather than use atoms for the nodes, we use residues such that the edges in this case are the residue-residue interactions as rather than measure the (small) mechanical response to an applied perturbation as in the previous section, we instead study the long time fluctuations of the protein at equilibrium.

An elastic network model for ADK (4AKE[169]) was constructed (Fig. 5.15c) and the average displacement for each of the edge interactions was calculated. A range of distance cutoffs were tested (7Å, 10Å, 12Å and 15Å) and the Spearman's rank correlation coefficient (ϱ) between the set of scores for the coincident interactions was calculated. Between 7Å and 10Å, $\varrho = 0.216$, for 10Å and 12Å, $\varrho = 0.679$ and between 12Å and 15Å, $\varrho = 0.801$, demonstrating a greater robustness in the results for larger cutoff values (indeed below 7Å, zero energy modes appear in the network as revealed by singular value decomposition of the rigidity matrix R). As such we use a cutoff of 12Å for the following results, which is in line with other reports in the literature[10]. A relatively right skewed distribution of edge displacement values is observed (Pearson median skewness = 0.580); the average value for 4AKE was 0.236 with a small number of interactions scoring significantly highly.

The top 2% of interactions by rank are those scoring above 0.409 and the top 1% above 0.452. The most highly scoring interactions are clustered primarily in the lid and AMP$_{\text{bind}}$, corresponding closely to those regions of the protein that are structurally altered during the open to closed transition. Qualitatively similar results were obtained by Mitchell *et al*[162], who performed residue strain analysis of ADK by comparing residue displacements across an NMR ensemble of structures to calculate local strain. Here, however just a single structure is used and strain is predicted *a priori*, emphasizing that the intrinsic topology of the protein determines where strain is distributed to assist function. The highest scoring interaction with 0.701 is Gly56 - Lys57; Lys57 is one of the residues that shifts more than 10Å during the open - closed transition[87] whilst Gly56 has been shown to display particularly high fluctuations in coarse-grained MD simulations[239].

In order to judge whether these results are significant or simply an artifact, we additionally constructed an elastic model for a regular hexagonal lattice to act as a null model. Using a 6 × 6 × 6 hexagonal lattice with nearest neighbour interactions only, we calculated the average fluctutations of each of the interactions and found there was a clear difference between those edges in the centre of the lattice versus those at the surface. The distribution of average edge fluctuations in Fig. 5.17b shows two distinct regions above and below 0.55, with edges at the centre experiencing a lower average strain. It appears then that strain within the protein structure is harnessed to achieve function, with highly strained interactions within the protein being localised to specific regions rather than being randomly distributed over the surface of the protein.

We also carried out the edge displacement calculation for the allosteric protein ATCase, a multimer consisting of six catalytic subunits, each hosting an active site, and six regulatory subunits where the allosteric sites are

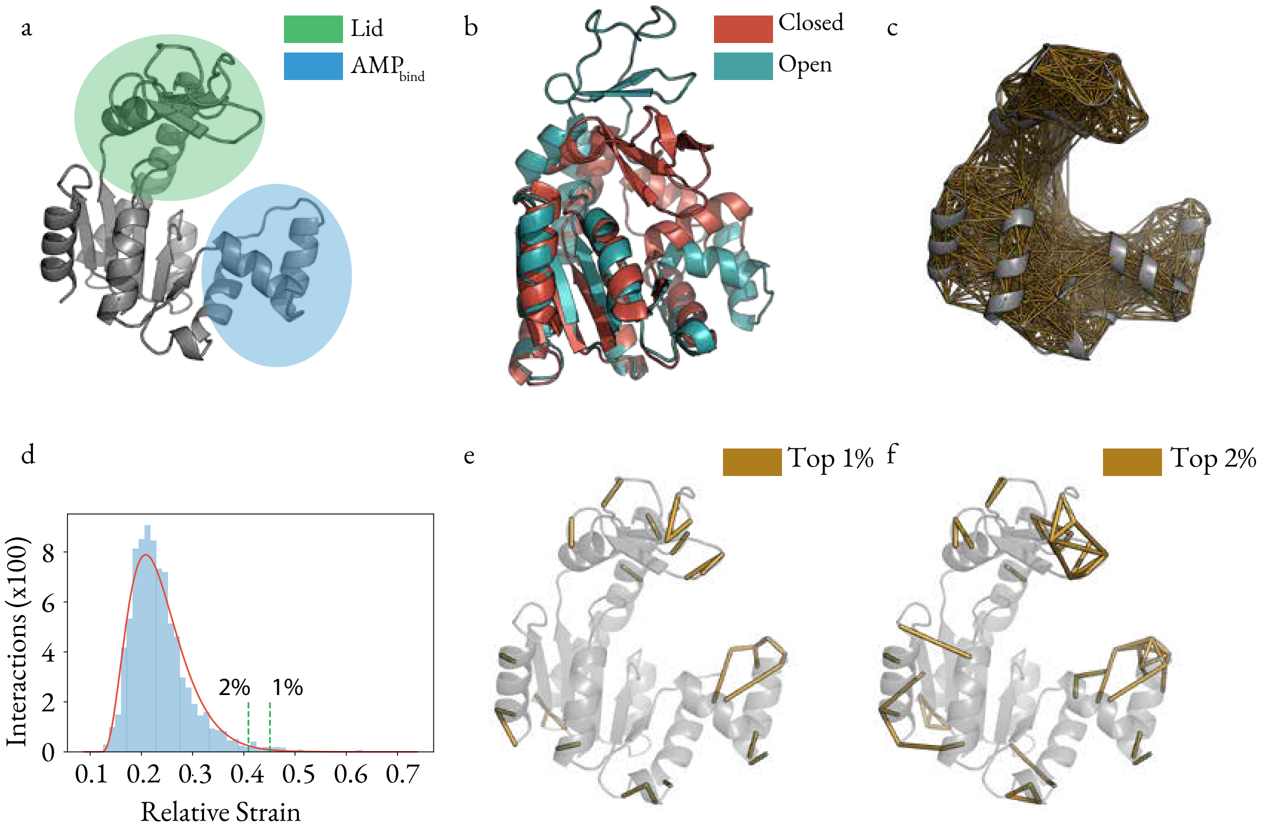


Figure 5.15: a) The structure of ADK from *Escherichia coli* (PDB: 4AKE). The lid and AMP$_{\text{bind}}$ domains are highlighted. b) Comparison of the closed (1AKE) and open (4AKE) forms. The main differences are in the lid and AMP$_{\text{bind}}$ domains. c) An elastic network model of ADK was constructed using a 12Å distance cutoff between residues. d) The distribution of average edge displacements in ADK, showing a relatively long tail of high scoring interactions such that the top 2% are significantly higher than the mean. e) Top 1% and f) Top 2% of highest scoring residue-residue interactions shown only. Significant clusters of high scoring interactions occur in the lid and AMP$_{\text{bind}}$ domains.

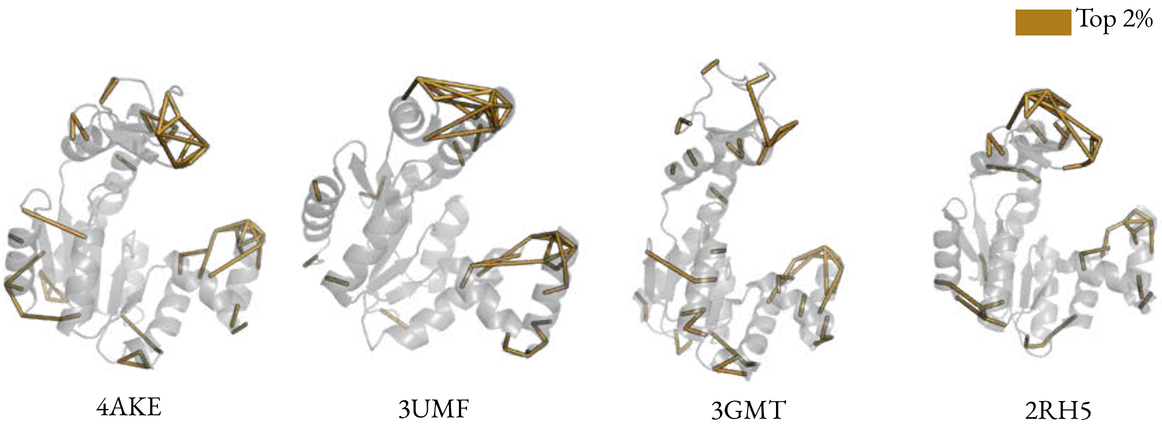


Figure 5.16: Top 2% of interactions by average strain across all 4 available open ADK structures. There is a strong consistency is the location of the highest scoring interactions, which appear mostly in the lid and AMP$_{\text{bind}}$ domains that undergo large transitions to the closed state.

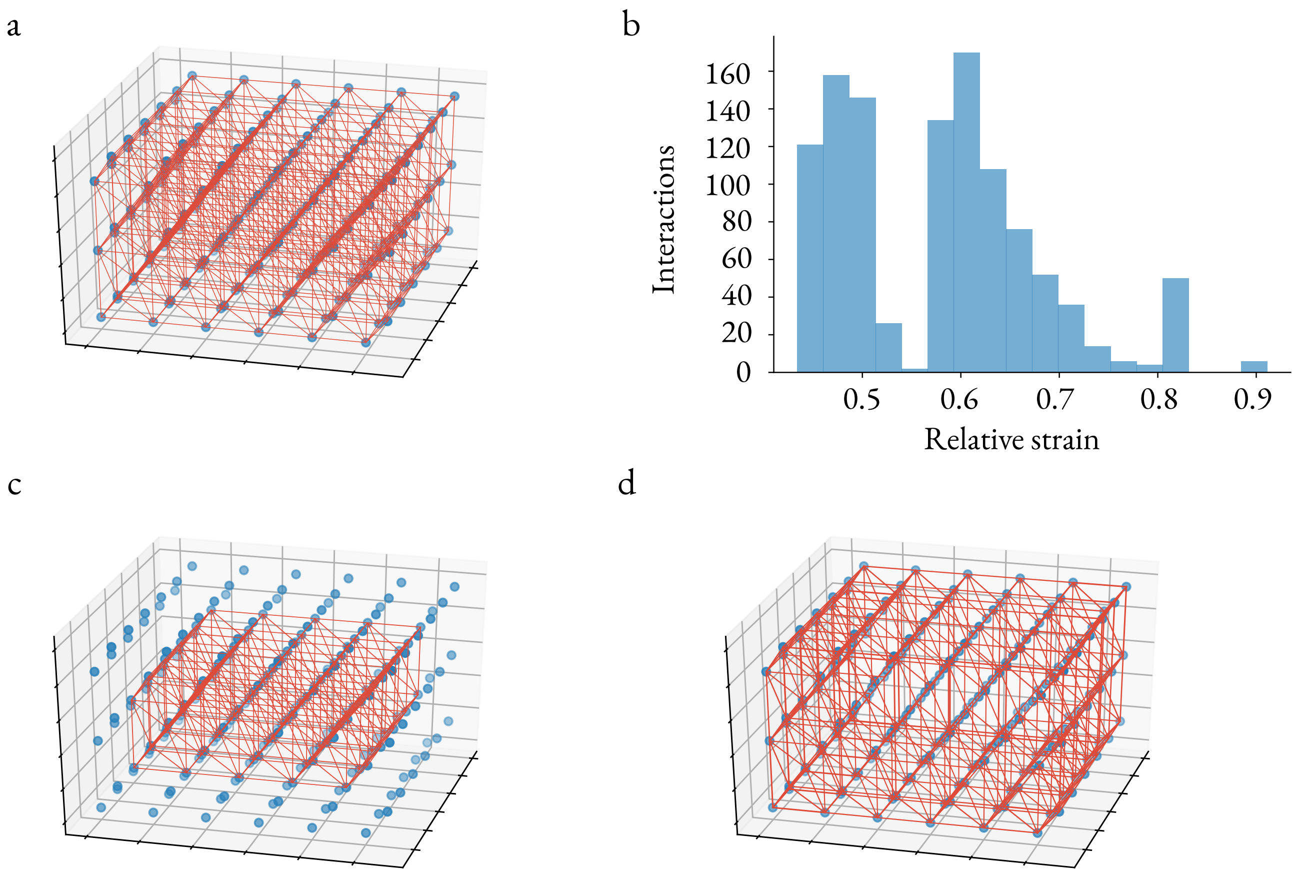


Figure 5.17: a) A hexagonal lattice with nearest neighbour interactions only. b) The distribution of average edge displacements shows two sub-distributions: below 0.55 are the interactions within the bulk, whereas those edges at the surface all score above 0.55. c) Bulk interactions. d) Surface interactions.

located. The distribution of the interaction scores is again right skewed (Pearson median skewness = 0.617). The highest scoring interactions in the active state bound to ATP (4KGV[42]) are strongly grouped at the six allosteric sites (with the top 1% shown in Fig.5.18). A number of residues that make up the allosteric site form interactions that score particularly highly: three of the Val9 - Glu10 interactions across the interface between the two regulatory chains rank 83rd, 84th and 85th out of 44979 interactions. 27 interactions involving ATP are present in the top 1%, including those to Ala11, Lys94 and Tyr89 that are part of the allosteric site itself.

However, similarly to ADK, it is apparent that the localisation of high average residue-residue fluctuations at functionally important sites is intrinsic to the protein structure; very similar results are obtained for the active state with no bound ligands (1D09[107]) such that it is not the ATP ligand itself that generates the strain at the allosteric site. For example three Asp19 - Asp4 interactions rank in the top 30 out of 45103, Asp19 forms part of the allosteric site, whilst Asp4 mutation to alanine is known to alter allosteric regulation of ATCase. Indeed, Asp4, Lys6 and Leu7 at the N-terminus of the regulatory chain all appear to be involved in allosteric signalling[56] and interactions involving one of these residues appear 75 times amongst the 423 interactions that rank in the top 1%.

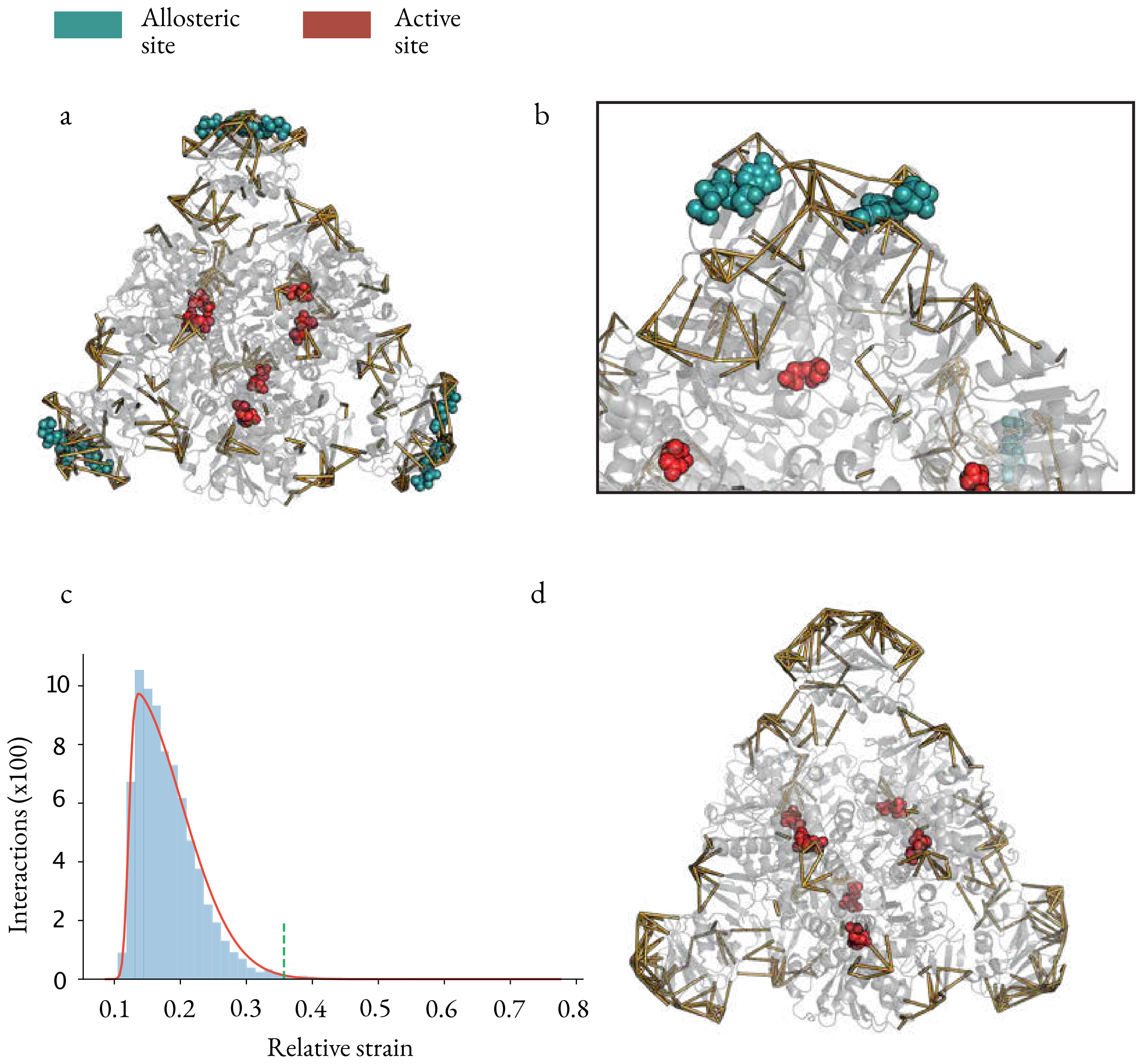


Figure 5.18: a) The top 1% of interactions by average edge displacement are shown in ATCase. The largest clusters of interactions appear strongly around the six allosteric sites, two of which are shown close up in b). c) The distribution of strain in each of the interactions. d) The allosteric sites are also seen to be the most strained regions in the unbound case.

5.11 Interaction embeddedness conclusions

Shortest path measures of residues are often cited as being an important indicator of communicability in proteins[9, 52], yet it is not immediately clear what process is being modelled in that instance. In contrast, in this part of the chapter, we demonstrate that the network betweenness measure *embeddedness*[206] has a explicit physical meaning in proteins (modelled as elastic networks), where a highly embedded edge (residue-residue interaction) is one that exhibits low strain under equilibrium conditions and conversely, highy strained interactions are those that have low embeddedness scores. Furthermore, we see that in two proteins, it is those regions of the protein that are functionally important that are most highly strained. We see that in ATCase, it is the allosteric sites that score highly and there are a number of potential explanations for this. One is that strain is localised at the allosteric site to aid crossing of the energy barrier upon binding (or release) of the allosteric ligands, such that the protein is "primed" for conformational changes. Another is that these sites may display large changes in entropy upon ligand binding or unbinding, which is what alters the energy landscape and thus the conformational populations of the various active and inactive states.

It should also be possible to experimentally verify the results of such analyses; Kolodziej *et al*[?] performed site directed mutagenesis to a single residue in the aspartate receptor of Salmonella typhimurium and were able to switch the protein from negative to positive cooperativity, indicating that single residue effects can be crucial. In Fructose-1,6-Bisphosphatase, Lu *et al*[145] mutated a number of residues, again discovering that changes in individual residues could have dramatic residues on allostery.

Chapter 6

Optimization of allosteric materials

6.1 Allosteric materials

There has been significant recent interest in the idea of developing "allosteric materials"; principled design of structures that exhibit a significant mechanical change at a target site, in response to an input force at distal site. These materials are thus inspired by the biological process of allostery, though specifically by the idea of long range structural propagation rather than by, say, thermodynamic explanations of allostery. A particular reason for this interest is that mechanical perturbations in randomly packed materials (a class believed to include proteins [136]) tend to be both nondirectional and quickly decaying, properties that are unconducive to the possibility of long range modulation of an active site.

Yan *et al*[245] then posed the question: is it possible to achieve a specific output displacement of a set of chosen target nodes, given a set of forces on some input nodes, by optimizing the topology of an elastic network? That is, given a set of nodes, how should we arrange a limited set of edges to achieve the desired displacement, which is expressed as a fitness function. The authors used a Monte Carlo based method for the arrangement of edges, in which the probability of moving an edge from one location to another was determined by:

$$P(\,|\sigma\rangle \rightarrow |\sigma'\rangle\,) = \min\left[1,\ \exp\left(\frac{F(|\sigma\rangle) - F(|\sigma'\rangle)}{T_e}\right)\right] \tag{6.1}$$

where σ and σ' refer to the configuration of edges before and after the movement of an edge, and T_e is an evolution temperature that determines the influence of the fitness function (so that at infinitely high temperatures, the algorithm simply produces random configurations). An average degree for the nodes of the network was chosen to be 5, roughly corresponding to the case in proteins and just above the *isostatic* value of $2d - d(d+1)/N$, whilst the lattice size was 12×12. Below the isostatic value, the network is very floppy and unable to transmit mechanical strain. It was then found that below a certain evolution temperature, it was always possible to achieve a perfect response in the network.

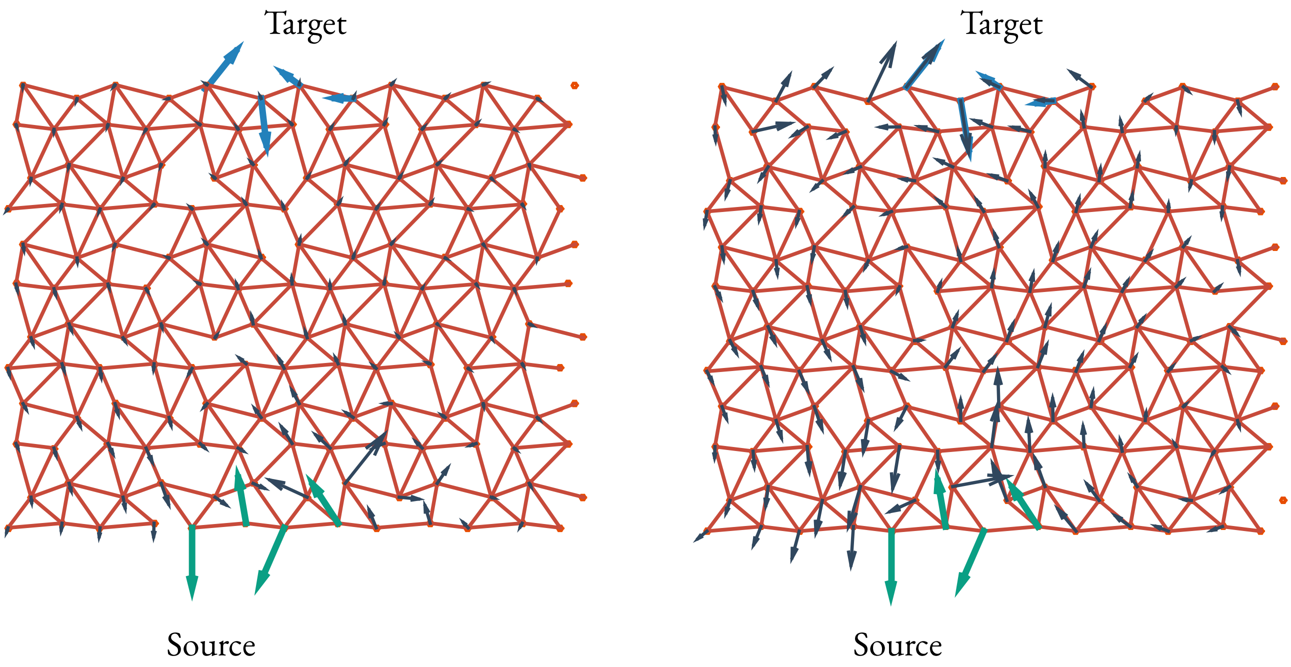


Figure 6.1: Yan *et al*[245] demonstrated using a Monte Carlo method that by optimising the topology of a network (given a fixed set of node positions and a predetermined number of edges) that a perfect output response is achievable (on the right) given a defined input force (green arrows) and output displacement (blue arrows). In contrast, a network with randomly placed edges displays a propagation of strain that decays quickly away from the source site (left diagram).

The authors noted a number of features of the high fitness networks: the part of the networks near the "active site" (target) was "soft", in the sense that the average node degree was very close to the isostatic point. Further, the average node degree decreased monotonically from the "allosteric site" (source) to the active site, inside a trumpet shaped region, that was rigid, but had a lower average degree than the bulk. The trumpet was then flanked on either side by more rigid regions. The evolved structures then lead to an intriguing outcome: whilst in random networks (at high evolution temperature), propagated displacements decay away quickly from the source site, in the high fitness networks, the displacement virtually disappears in the bulk, before reappearing strongly at the target site. The same authors subsequently[246] applied similar principles to three dimensional elastic networks and recovered a number of structures adopted by allosteric proteins: shear, hinge (synonymous with the 'scissor' mechanism[159]) and twist. Analogous results are seen as with the two-dimensional case, in which there is heterogeneity in the response to the input force at the allosteric site across the network, with large parts of the network effectively acting like a rigid block whilst other display a large response.

A similar approach was taken by Rocks *et al*[197], in a 2D network of 190 nodes and 400 bonds. Two pairs of nodes were chosen to be the source and target and the average node degree was chosen to be just above the isostatic value. By selectively removing certain edges, the authors attempted to maximise the ratio of the strain on the target edge to that on the source edge. Remarkably, only 5 bonds on average needed to be removed to achieve a strain ratio of 1, demonstrating the plausibility of protein networks evolving to allow for long range mechanical transfer. Similar results were recorded for 3D networks, though even fewer (4 on average out of 740) bonds needed to be removed. Impressively, the authors then used 3D printing to build physical elastic networks based on their theoretical results and were able to achieve the calculated strain ratios with 98% accuracy.

Both the above approaches focused on the *linear* response of strain in elastic networks, which holds in the regime of small displacements and is calculated using the standard stiffness matrix approach described in Chapter 5. Flechsig[71] instead studied the full, nonlinear dynamics of elastic networks by numerically solving the overdamped Langevin equation:

$$\frac{dr}{dt} = B(r)GB(r)^T \cdot (r - r_0)^T + f_{in} \tag{6.2}$$

using our notation from Chapter 5 but noticing that the incidence matrix is no longer a constant but is now a function of node position that is recalculated at each timestep. A random network was then constructed using a series of distance constraints and springs were added to the network according to a distance cutoff. Once again, the propagation of strain was optimized between two sites by progressively changing the location of springs and qualitatively similar results to the previous two analyses were achieved: random networks showed no directional flow of mechanical strain whilst optimization of the elastic networks lead to softer regions around the target. Furthermore, it was seen that in certain cases, a single mutation (here modelled by the removal of a spring) could completely disrupt allosteric communication - analogous to the situation in real proteins. By initially building the network as two district domains, Fleisig was also able to observe the development of a flexible region joining the two domains that appeared to function as a hinge, again recovering a common feature of real proteins.

In each of these cases, a number of assumptions are made when drawing a comparison between the evolved, artificial elastic networks and allostery in proteins: firstly that we can use a harmonic approximation for residue-residue interactions as discussed in Chapter 5 and secondly that thermal fluctuations can be ignored and the nodes only respond to the chosen input. It is nevertheless an important result that it is in fact possible to transmit mechanical strain across reasonably large distances within elastic networks such that this is a plausible mechanism of allostery. Additionally, these evolved networks recovered important features of many allosteric proteins, such as hinges, distribution of more rigid and more flexible regions and vulnerability to mutations at critical sites.

One drawback of the previous methods is that by optimizing a mechanical response based on the *location* of a set of springs in the network, the authors have formulated integer programs, which are NP-hard and are thus computationally demanding. In this chapter, we take a different approach: given a fixed set of springs, can we optimize the value of the *spring constants* in order to maximise the response at a target edge, given an input force on one (or more) input edges? We show that this can in fact be presented as a *convex optimization* problem, or more specifically, a subset of convex optimization called *semidefinite programming*. The advantage of this approach is two-fold: convex optimization problems have in general the property that any local minimum is also a global minimum; in practice this usually means there is just a single optimal solution. Secondly, this optimum can be quickly reached using efficient methods for which a range of software packages exist.

6.2 Convex Optimization

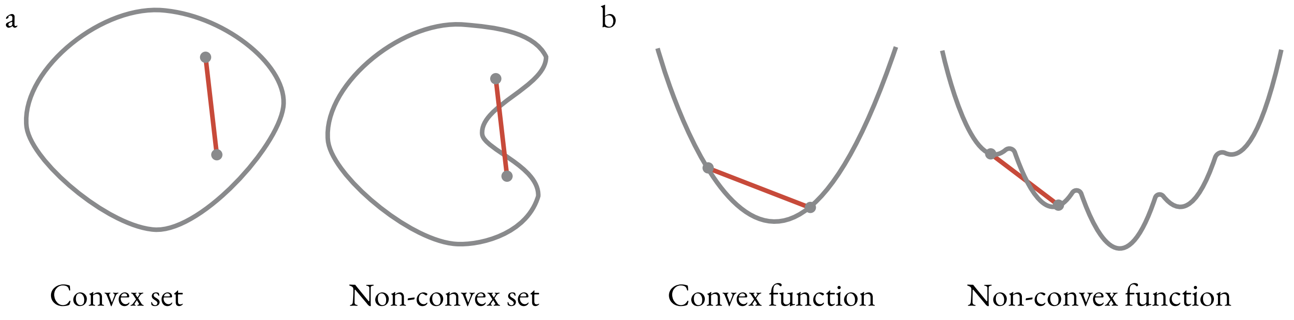


Figure 6.2: Convex optimisation is concerned with the minimisation of convex functions over convex domains. a) A convex set is one that satisfies the relation that for any two points x_1 and x_2 in the set, all points defined by $\vartheta x_1 + (1 - \vartheta)x_2$ are also present in the set, which can be seen visually by drawing a straight line between the two points as in the figure. b) Convex functions are those that satisfy *Jensen's inequality*: $f(\vartheta x_1 + (1-\vartheta)x_2) \leq f(\vartheta x_1) + f((1-\vartheta)x_2))$ for any two points x_1 and x_2 in the domain of f. Intuitively, this can be seen by drawing a straight line between any two parts of the function and observing that this line must always lie above said function.

A detailed overview of convex optimization is not possible here but an excellent introduction to its theory and application is provided by Vandenberghe and Boyd[23]. The general form of a convex optimization problem is:

$$\begin{aligned} \text{minimise} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \qquad i = 1, \ldots, m \\ & h_i(x) = 0, \qquad i = 1, \ldots, p \end{aligned}$$

where the functions $f_i(x)$ are *convex* and $h_i(x)$ are *linear*. $f_0(x)$ is our *objective function* that we wish to minimise, whilst the addtional functions f_i are *inequality constraints*, that together with the *equality constraints* $h_i(x)$ define the *feasible region* over which we allow values of x. Thus, in order for a problem to be convex, we must be minimising a convex function over a domain that is a convex set. A generalisation of the above convex optimization problem is where the inequality constraints are vector valued. The inequalities thus become *generalised inequalities*:

$$\begin{aligned} \text{minimise} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \preceq_{K_i} 0, \qquad i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

The equality constraints have now been written in matrix form for convenience. Generalised inequalities extend the notion of ordering on the real number line $\mathbb{R}$ to vectors and matrices. The two most common generalised inequalities involve the *nonnegative orthant* and the set of *positive semidefinite* matrices. The nonnegative orthant in $\mathbb{R}^n$ is simply the set of vectors where each of the components is positive. For example, in $\mathbb{R}^2$ this would correspond to the vectors in the upper right quadrant. Thus $K = \mathbb{R}^n_+$ in the inequality above. Denoting by S_+ the set of positive semidefinite matrices (S being the set of symmetric matrices), setting $K = S_+$ leads to problems in *semidefinite programming*:

$$\begin{aligned} &\text{minimise} && c^T x \\ &\text{subject to} && x_1 F_1 + \ldots + x_n F_n + G \preceq 0 \\ & && Ax = b \end{aligned}$$

where $G, F_1, \ldots, F_n \in S^k$, and $A \in R^{p \times n}$. We have also dropped the K_i subscript as when the generalised inequality involves matrices, it refers to semidefinite matrices the vast majority of the time.

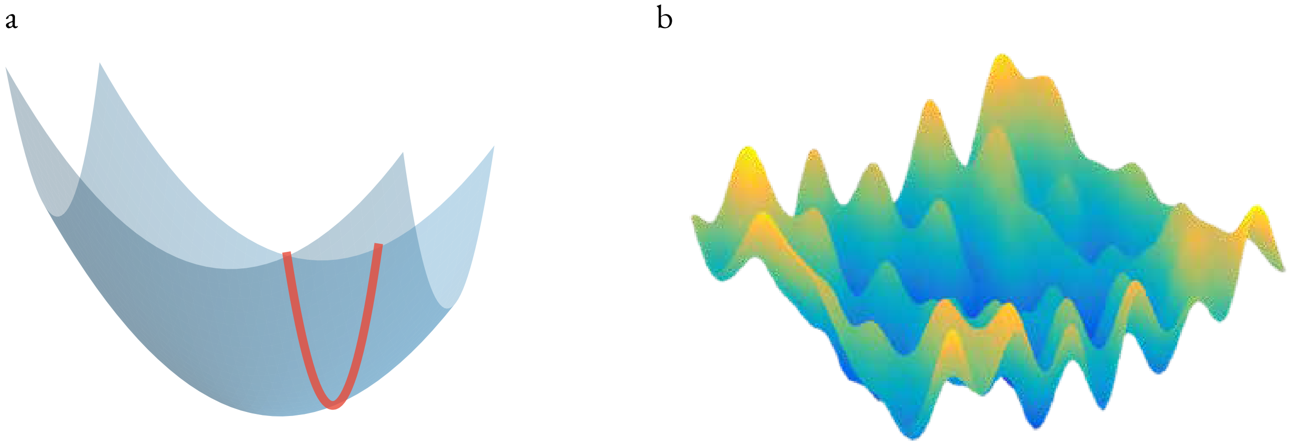


Figure 6.3: a) A convex function (in blue) has a single, global minimum, here just the bottom of the "bowl". A simple case of a *constrained* convex optimisation problem is also shown. If we imagine the quadratic function sitting in the $x - y$ plane, then we might wish to minimise the function *given a fixed value of y* and this is represented by the red line. In effect, we take a "slice" through the 2-dimensional well to obtain a new 1-dimensional quadratic that has a different minimum to the unconstrained problem. b) In contrast, non-convex problems often have a large number of local minima and typically obtaining the global minimum is very difficult.

Our aim then is to pose our problem, of optimizing the spring constants of an elastic network to achieve the greatest possible displacement response at a target edge, as a standard form convex optimization problem. More specifically as it turns out, the problem can be framed as a semidefinite program. In order to do so, and common to many semidefinite problems, we make use of a result involving the *Schur complement* (A.5.5 [23]) and here we introduce some necessary background. If we have a symmetric matrix X:

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{6.3}$$

where A is also symmetric, then (if A is nonsingular), the Schur complement of A in X is defined as:

$$S = C - B^T A^{-1} B \tag{6.4}$$

The Schur complement appears, amongst other places, in situations where we wish to minimise a quadratic form over a subset of the variables:

$$\min_{u} \quad u^T A u + 2 v^T B^T u + v^T C v \tag{6.5}$$

Minimising over u, we hold v constant, such that we wish to find the smallest value taken by the term:

$$f(u) = u^T A u + b^T u \tag{6.6}$$

where we have collected constant terms into b. Noting that:

$$\frac{1}{2}(u + A^{-1}b)^T A (u + A^{-1}b) = \frac{1}{2} u^T A u + u^T b + \frac{1}{2} b^T A^{-1} b \tag{6.7}$$

we can rewrite our function $f(u)$ as:

$$f(u) = \frac{1}{2}(u + A^{-1}b)^T A (u + A^{-1}b) - \frac{1}{2} b^T A^{-1} b \tag{6.8}$$

The second term is just a constant but we have free choice over the value of u in the first term. If A has a negative eigenvalue $-\lambda$ associated with an eigenvector w, such that $Aw = -\lambda w$, we can choose $u = \alpha w - A^{-1} b$ (as we can simply scale w however we wish):

$$\begin{aligned} f(u) &= \frac{1}{2} \alpha w^T A \alpha w - \frac{1}{2} b^T A^{-1} b \\ &= -\frac{1}{2} \alpha^2 \lambda \, |w|_2 - \frac{1}{2} b^T A^{-1} b \end{aligned} \tag{6.9}$$

so that our function can be made arbitrarily negative. To guarantee $f(u)$ is bounded below, we must have that $A \succeq 0$ (or here, really $A \succ 0$ as we assume A is invertible). Furthermore, we can see that the minimum of $f(u)$ is achieved when the first term in (6.8) is set to zero, which occurs when $u = -A^{-1}b$ with a minimum value of $f(u)$ of $-\frac{1}{2}b^T A^{-1}b$.

Substituting the solution for u back into the quadratic form:

$$\inf_u \quad \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \quad = \quad v^T S v \tag{6.10}$$

We can now see that if $A \geq 0$, then $X \geq 0$ if and only if $S \geq 0$. In our problem, however, we will be making use of the stiffness matrix K, which we know is singular (having at least 6 zero eigenvalues representing the rigid body motions) so possesses a pseudoinverse. We consequently require an additional restriction on the form of X, which we can uncover by extending the previous arguments for minimising a quadratic form.

Firstly, we write A as its eigendecomposition $W^T \Sigma W$ and substitute into the form of (6.6) where A is singular:

$$\begin{aligned} f(u) &= \frac{1}{2}u^T W^T \Sigma W u + u^T W^T W b \\ &\frac{1}{2}(Wu)^T \Sigma W u + (Wu)^T W b \end{aligned} \tag{6.11}$$

exploiting the fact that $W^T W = I$. Then, assuming the rank of A to be r, we set $Wu = \begin{pmatrix} x \\ y \end{pmatrix}$ and $Wb = \begin{pmatrix} c \\ d \end{pmatrix}$, where $x, c \in \mathbb{R}^r$ and $y, d \in \mathbb{R}^{n-r}$, we rewrite the expression for $f(u)$:

$$\begin{aligned} f(u) &= \frac{1}{2}\begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} \\ &= \frac{1}{2}y^T \Sigma_r y + y^T c + z^T d \end{aligned} \tag{6.12}$$

in which Σ is written in full showing the r nonzero eigenvalues contained in the invertible submatrix Σ_r and the $n - r$ zero eigenvalues in the bottom right. We know from the previous section that the sum of the first two terms has a minimum of zero, here if $\Sigma_r \geq 0$ which is just the same as $A \geq 0$. Our additional restriction then arises because in order for (6.12) to be bounded below, we must have $d = 0$ and therefore also $Wb = \begin{pmatrix} c \\ 0 \end{pmatrix}$. It can be shown that consequently b must lie in the range of A, which can be stated as $(I - AA^{\dagger})b = 0$. Comparing terms with the original problem in terms of X, we see that $b = 2Bv$, which implies we need to have $(I - AA^{\dagger})Bv = 0$. However, the expression must hold for all values of v if the problem is to have a minimum and therefore the additional restriction in the case of singular matrices A is $(I - AA^{\dagger})B = 0$. Finally, we then again also require the Schur complement, this time with singular A, to be positive semidefinite $S = C - B^T A^{\dagger} B \geq 0$. Putting this all together, the result that we will use in our semidefinite program is:

$$X \geq 0 \quad \Longleftrightarrow \quad A \geq 0, \quad (I - AA^{\dagger})B = 0, \quad C - B^T A^{\dagger} B \geq 0 \tag{6.13}$$

6.3 Spring constant optimization

We now have all the tools to set up our optimization problem. However, it turns out the presented solutions obtained for the chosen objective function *are not* the true solutions and unfortunately only arise as a result of the finite tolerance of the underlying convex optimization solver. Despite this, the results themselves are still interesting and valid as an example of long range mechanical transfer, and as such are presented here in full, before we discuss why the problem is in fact not solvable in the shown form.

If we firstly consider the case of applying a unit force to just a single source edge i as input to maximise the output displacement on an edge j, we know from Chapter 5:

$$\begin{aligned} e_{ji} &= B_j^T K^{\dagger} B_i \\ &= \operatorname{Tr} B_j^T K^{\dagger} B_i \\ &= \operatorname{Tr} (B_i B_j^T) K^{\dagger} \end{aligned} \tag{6.14}$$

where we have used the fact that the trace of a scalar is just the scalar itself and the cyclic property of the trace. The semidefinite program can now be written as such:

$$\begin{aligned}
\text{minimise} \quad & \mathrm{Tr} Y \\
\text{subject to} \quad & 1^T g = 1, \quad g \geq 0, \\
& \begin{bmatrix} K(B_i B_j^T)^\dagger & I \\ I & Y \end{bmatrix} \geq 0
\end{aligned} \tag{6.15}$$

where g is our vector of spring constants (such that $\mathrm{diag}(g) = G$) and we enforce the constraint that the springs constants must all be positive. Additionally, we normalise the spring constants so that they sum to 1, which prevents the solution for g from growing arbitrarily large. Finally, we have introduced the dummy matrix Y. If our generalised inequality constraint matrix is $X \geq 0$, then using our previous definition of the Schur complement: $S = Y - (B_i B_j^T)K^\dagger \geq 0$ and therefore:

$$Y \geq (B_i B_j^T)K^\dagger \tag{6.16}$$

and so we see that if we set $Y = (B_i B_j^T)K^\dagger$ by minimising $\mathrm{Tr} Y$, we are equivalently minimising $Tr(B_i B_j^T)K^\dagger$, allowing us to formulate our desired optimization of the output displacement as a standard form semidefinite program. Calculations were carried out using the python package for disciplined convex optimization CVXPY[57, 2], using the solver SCS[181].

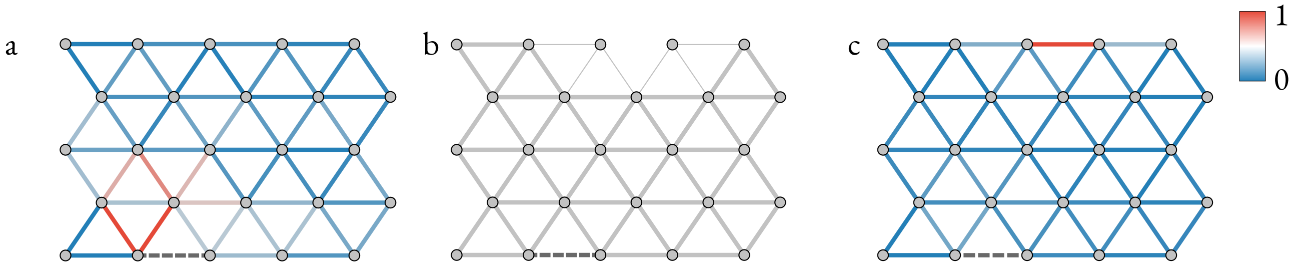


Figure 6.4: a) The edge displacement in a 5 × 5 lattice with the source edge shown as a dashed, dark grey line. Displacement quickly decreases radially and there is a negligible effect on the target edge (displacement values are normalised, not including the target edge). b) The spring constants are optimized as described in Eq. (6.15) and the strength is represented by the thickness of the edges. c) In the optimized network, edge displacement extends across the lattice, leading to the greatest output displacement at the target edge.

Using a simple example of a 5 × 5 2D hexagonal lattice, we see that compared to using a uniform spring constant for all weights, contraction of the target edge is increased in magnitude from 0.005 to 0.209. The spring constant of the target edge is unsurprisingly set to 0 during the optimization but more interestingly a "soft" region of weak springs is formed around the target spring, somewhat similar to previous simulations of allosteric materials[245], where the "softness" represented a lower node connectivity around the target site. When a uniform spring constant is used for all edges, the perturbation quickly decays radially away from the source edge and a negligible effect is felt at the target edge (Fig. 6.4a). However, after optimizing the spring constants, the target edge undergoes by far the largest displacement (other than the input edge itself) despite being located on the other side of the lattice to the perturbation site.

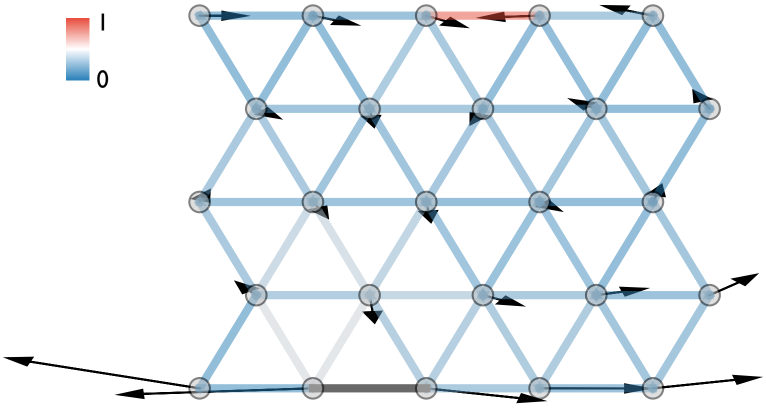


Figure 6.5: The results for the 5 x 5 lattice with the node displacements also shown.

A trivial solution

Unfortunately, whilst the networks above are certainly valid and *do* show long range mechanical transfer, the actual attainment of the networks via the semidefinite program is due to the small error tolerance present in all convex solvers. If we study Equation (6.13), we see that the condition our generalised inequality puts on our value for A is:

$$K(B_iB_j^T)^\dagger \geq 0 \tag{6.17}$$

that is $K(B_iB_j^T)^\dagger$ must be positive semidefinite. The (pseudo-) inverse of a positive semidefinite matrix must also be positive semidefinite and so we have necessarily set the condition:

$$(B_iB_j^T)K^\dagger \geq 0 \tag{6.18}$$

Our objective function is $\mathrm{Tr}(B_iB_j^T)K^\dagger$ and we are trying to make this as negative as possible. However, we know that the trace of a matrix is equal to the sum of its eigenvalues, but we have just said all the eigenvalues must be greater than or equal to zero. Thus the optimal solution to our convex problem is actually zero, negative values (corresponding to contraction of the output edge) only being obtained by an overshoot of the solver. The issue is not rectified by instead trying to maximise the output edge displacement (a stretch of the output edge) as we simply end up with the oppposite problem: our matrices then need to be negative semidefinite and once again the maximum of the trace of a negative semidefinite matrix is zero.

In fact we can go further and see that the problem itself is not well framed. By aiming to maximise (or minimise) the edge displacement of a particular edge, we can actually trivially obtain an infinite displacement by setting all of our edges to have zero spring constants (except, say one spring set to 1 so as to satisfy our constraint $1^T g = 1$). Then we will have a set of *floppy modes*, that are just the rigid motions of each of the nodes. Thus any solution to the equation $e_{ji} = B_j^T K^\dagger B_i$ must also include all contributions from the nullspace of $K^\dagger$ (which is just the same as the nullspace of K). One such contribution is the linear combination of the left node of edge e_{ji} moving to the right and the right node moving to the left (i.e. the edge contraction we are trying to maximise. See Fig.6.6). Further, any scalar multiple of this vector is also a solution and we can therefore scale this vector to be as large as we like. If we denote this node vector as u_{null} then the edge displacement $B_{out}u_{\text{null}}$ that we seek can be made infinitely large.

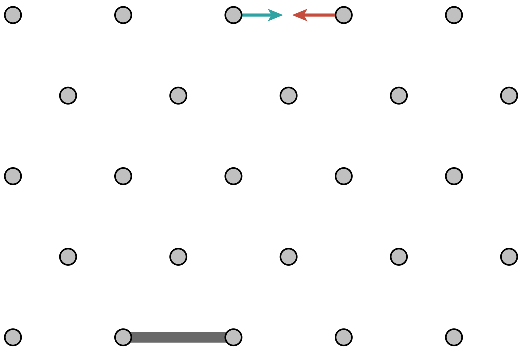


Figure 6.6: We can trivially make the target displacement as large as possible by setting the spring constant of one of the edges (say arbitrarily the source edge) to be 1 and all the rest to be 0. Then any solution to the stiffness equation $f = K^\dagger u$ must include the contributions from the nullspace, such as the linear combination of two vectors in red and green representing rigid motions of the two target nodes.

6.4 Optimization of correlated motions

An alternative approach is to instead try and optimize the set of spring constants so as to achieve the correlated motion of the active and allosteric sites. Such a coupling is one of the suggested explanations of *entropically* driven allostery[43, 159] and is mediated through the normal modes of the protein, as these global motions can couple distant parts of the structure[198]. We therefore consider the eigenvalue problem:

$$Kv_i = \lambda_i v_i \tag{6.19}$$

where K is our stiffness matrix, with λ_i and v_i the i^{th} eigenvalue and eigenvector of the matrix respectively. Note we would usually also have the mass matrix $\mathcal{M}$ on the right hand side of the equation but here we set $\mathcal{M} = I$, the identity for simplicity. Given a fixed set of nodes and edges, determined as before by B, the geometric incidence matrix, we can rewrite Eq. (6.20) as:

$$B^T GBv_i = \lambda_i v_i \tag{6.20}$$

The constraint we wish to place on the system is that for the motion of one of the eigenvectors, the displacement "allosteric site" edge should be coherent with that of the "active site" edge. More specifically, we impose that the allosteric site edge should have an equal and opposite displacement to the active site edge, which we can write as:

$$\begin{aligned} B_{ac}v &= -B_{al}v \\ (B_{ac} + B_{al})v &= 0 \end{aligned} \tag{6.21}$$

where B_{ac} refers to the row of B relating to the active site edge and likewise B_{al} for the allosteric site edge. Eq. (6.21) says then that we must choose a set of node motions v that lies in the nullspace of the vector $B_{ac} + B_{al}$. We can obtain the set of vectors that lie in the nullspace by calculating the singular value decomposition of $(B_{ac} + B_{al})$ in a similar manner to the rigidity matrix of Chapter 5. We can arbitrarily select any of the vectors v that lie in the nullspace and then find the spring constants g using the optimization:

$$\begin{aligned} &\text{minimise} \quad 1 \\ &\text{subject to} \quad (B^T GB - \lambda I)v = 0, \quad 1^T g = 1, \quad g \geq 0, \end{aligned} \tag{6.22}$$

where $G = \text{diag}(g)$. Here, we have used a common trick where we 'minimise' a number (chosen arbitrarily to be 1) whilst building our objective function (from Eq. (6.20)) into the constraint to find the set of spring constants g that give the chosen motion v, allowing free choice over the value of λ.

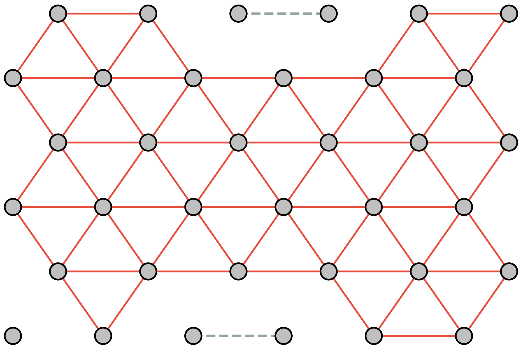


Figure 6.7: We optimize the set of springs so that the active and allosteric site edges (shown as grey dashed lines) display correlated dynamics for one of the normal modes, such that the active site compresses as the allosteric site extends and vice versa. The structure resembles the scissor molecule of Ref. ??, whereby the active and allosteric sites regions are "soft", whilst the rest of the molecular is stiff, allowing for the necessary hinge motion. In this case, either site could act as the active or allosteric site.

As seen in Fig. 6.7, the resulting structure resembles the "scissor molecule"[159, 158] that allows for correlated motions of the allosteric and active site. Furthermore, by performing the eigendecompostion of the resulting stiffness matrix $K = B^T GB$, we see that the normal mode associated with the eigenvalue λ that was allowed to freely vary in (6.22) is indeed one of the smallest eigenvalues (in this particular case the 4^{th} smallest) and thus corresponds to one of the slow, large amplitude, global modes. The result confirms (perhaps unsurprisingly) that interaction between two distant sites on the network structure must be mediated by global modes, rather than the shorter ranged, high frequency modes.

We can easily extend our approach to 3-dimensional structures and again use the example of a hexagonal lattice (here of dimensions $4 \times 4 \times 4$). We select two edges on either side of the lattice to act as the allosteric and active sites (it is arbitrary which site is which).

Shown in Figure 6.8 are the results for one of the motions that leads to correlated motion of the allosteric and active edges. Analogously to the 2-dimensional case, we see the edges around the two sites are set to nearly zero (the edges shown in dotted red have a spring constant of less than 10^{-15} as compared to those in green that all have values of 10^{-3}, again resulting in a "scissor-like" shape to the lattice. Again, the eigenvalue of the mode that leads to coherent motion is one of the small, low energy eigenvalues (in this case, the 10^{th} smallest) such that the motion is global.

As for the 2-dimensional case, we apply the constraint that the allosteric and active edges have an equal and opposite displacement for one of the normal modes of the lattice and then formulate the problem as a linear

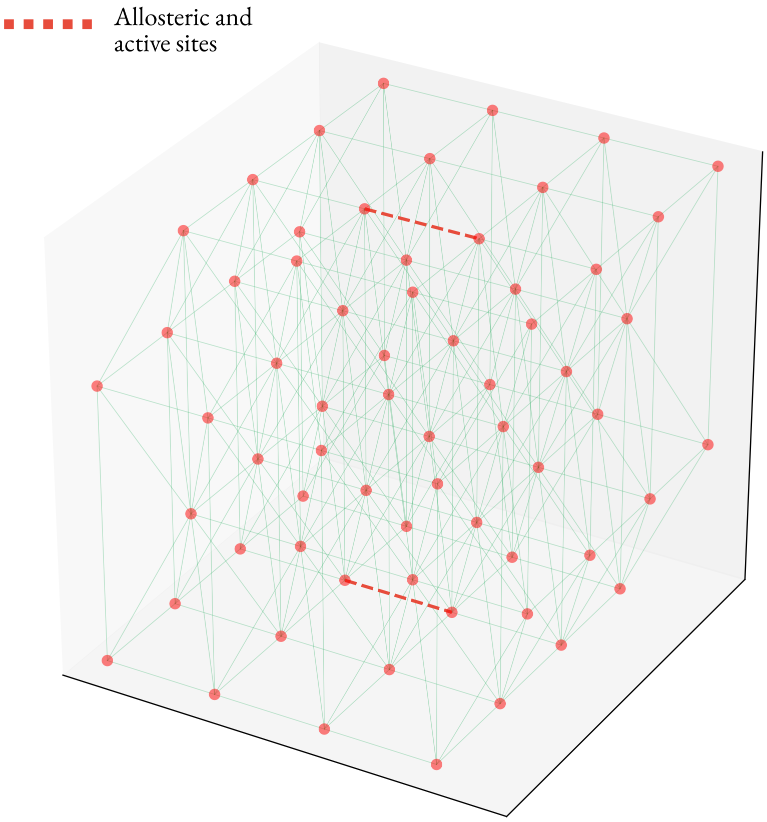


Figure 6.8: For the 3-dimensional case, a $4 \times 4/times4$ hexagonal lattice was chosen and again the allosteric and active sites (shown as dashed red lines) were selected to be on opposite sides of the lattice.

program to find the set of spring constants g that gives the correlated motion. The equations are thus have exactly the same form as Eq. (6.21) and (6.22) except with B now referring to a geometric incidence matrix of dimensions $M \times 3N$ with M edges and N nodes.

6.5 Conclusions

Despite the realisation that the results of the semidefinite program were not in fact optimal, we have serendipitously managed to design elastic networks that do exhibit allosteric-like properties, if we restrict ourselves to the case where allostery might be explained by long range mechanical perturbation. What is particularly interesting about the designed networks is that the displacement at the "active site" can be large, whilst the displacement of the rest of the network is very small and this is achievable without any sort of complex topology, but instead by simply setting the bonds around the active site to be weaker than the rest of the network. If all that is required is some mutation that causes a "softening" of the interactions around the active site, in addition to a pre-existing rigid pathway between the allosteric and active sites then the mechanism appears much more plausible than if strain had to propagate through a specific set of residues.

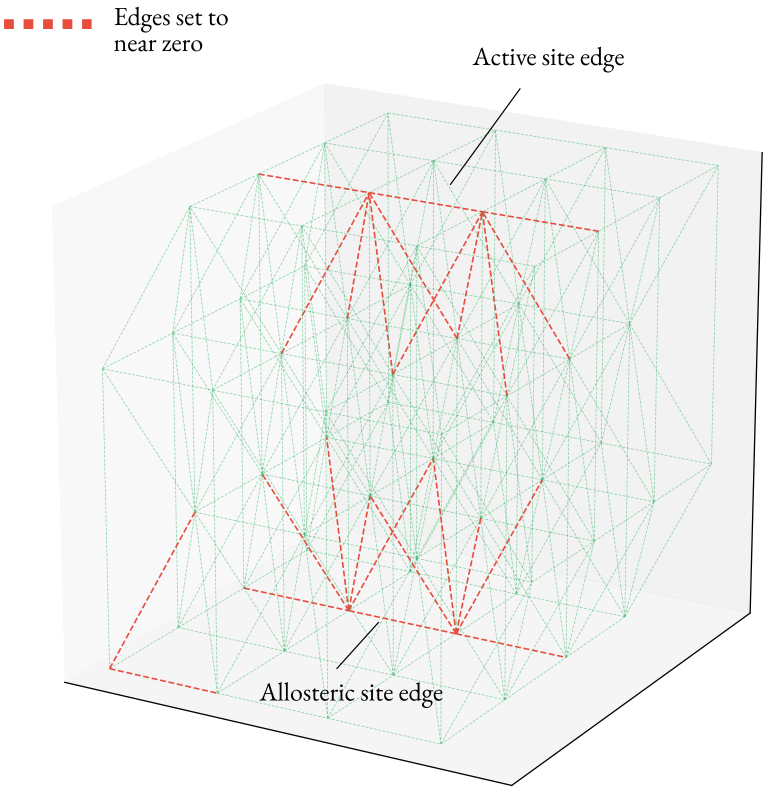


Figure 6.9: One of the vectors that satisfies the constraint in Eq. (6.21) is chosen and the spring constants are optimized. Those edges marked as dotted red are those that are set to be near zero ($< 10^{-15}$) whilst those edges in green have spring constants between 10^{-3}and10^{-2} (where the sum of the spring constants is set to 1).

On the other hand, by optimizing the set of spring constants such that the edge displacements for the allosteric and active sites were (negatively) correlated for one of the normal modes of the lattice, we were able to generate structures that resembled the "scissor molecule" of ?? and ??. The results support the idea that long range communication in proteins is mediated via the *global* modes or in other words, the "violin" model as opposed to the domino model of residue pathways[124].

Chapter 7

Conclusion

7.1 Summary of Thesis

In this thesis, we have primarily investigated the nature of the long range communication in proteins as a means to understand a potential mechanism of allostery. In particular we have focused on network methods in the edge space, which have particular relevance in the study of proteins, where it is ultimately changes in bonding or residue-residue interactions that modulate function. Initial work used a previously developed graph theoretical method *bond-to-bond propensities* to measure energetic coupling between functional sites in a large multimeric protein ATCase. We found strong and exclusive coupling between the active and allosteric sites, which displayed an intriguing nonlinear behaviour whereby the effect of the perturbation at the allosteric site had little effect on the bulk of the protein before reappearing strongly at the active site.

We then extended the ideas in *bond-to-bond propensity* to a three dimensional network description of proteins in order to understand the physical process at work. In this framework, the allosteric perturbation is modelled as the compression of the set of bonds between the ligand and the protein, and the mechanical response over the set of bonds can be efficiently calculated by solving a sparse linear equation. The generality of this approach also allowed us to design simple allosteric materials by using semidefinite programming to optimize the edge weights (here, spring constants) to maximise mechanical propagation in network. We found that long range effects can be achieved without large structural change in the bulk of the network by simply requiring that the "active site" of the network has much weaker springs relative to the rest of network, without any sort of complex topology being necessary. We were also able to use the elastic network framework to identify highly functional sites in proteins by considering where strain is localised at equilibrium.

Finally, we take a somewhat different approach in the final chapter where we apply a dynamics based community detection method called Markov stability to coarse grain a Markov state model of a model protein. Here our network is effectively a discrete approximation to the free energy landscape of the protein and we show that Markov stability is a principled way of linking the short timescales of protein motion as determined by Newton's Laws of Motion to the biologist's view of protein states.

7.2 Reflections

Can we say the results presented here provide strong evidence for the so called *structural view* of allostery? On the basis of the *elastic response* model of Chapter 5, the answer is: not particularly. In each case presented (and others not included here), the effect of the perturbation of the allosteric source site was consistently isotropic and decayed away exponentially, neither of which gives cause to believe long range perturbative effects are what drives allostery.

However, the results in Chapter 6, along with previous work by Le Yan[245] and Rocks[197], at least show that even within elastic networks, where only nearest neighbour interactions are present, long range structural changes can occur. Though in each case the networks are too small to resemble atomistic models, they may be a plausible description for coarse grained residue representations of proteins. Furthermore, the results from our convex optimization method seem to show (serendipitously) that all that is required for long range communication, that leaves the rest of the structure largely unchanged, is that the active site region is made "soft". That is, the interactions at the active site should be weaker than the rest of the network, and in fact this result is very similar to what was found by Le Yan in the context of optimizing the topology of the network, where in that case it was the average degree of the nodes near the active site that was lower, rather than the spring constants as it is here. Again, whilst we must be clear we are only dealing with toy models in this instance, it does appear to lend a more credible evolutionary explanation of how long range effects may have arisen. It is difficult to see how communication between a "future" allosteric site and the active site could have arisen as a series of mutations, progressively increasing increasing the communication between two distal sites. In contrast, if all that is required to form some sort of coupling to the active site is a mutation or small number of mutations near the active site, such that the region is made "soft", then the process appears far more plausible.

Even this insight however does not give much credence to the idea of *pathways* within proteins. Instead long range perturbations appear to be mediated by bulk effects, rather than specific paths of residues linking sites together. Whilst a common approach in the literature has been to identify pathway residues via graph theoretical techniques such as measures of residue centrality based on shortest paths, it is not entirely clear what physical meaning this has in a protein. In contrast, the centrality measure *embeddedness*, introduced by Schaub *et al*[206] and applied here in Chapter 5 was shown to be equivalent in the case of elastic networks to the edges in the network that have the highest average strain, given the network sits in a heat bath at equilibrium. When applied to ATCase, it was the allosteric sites that showed the highest average strain (or alternatively, the lowest embeddedness). These results seem more supportive of the idea that allosteric sites are those that are sensitive to energy changes (and can thus significantly remodel the energy landscape), rather than necessarily having a particularly strong link to the active site. It has already been suggested the allostery is a property of *all* dynamic proteins[86] with "allosteric ligands" simply being a special case of molecules that lead to particularly large conformational changes. It would thus seem unlikely that a multitude of energetic pathways exist in proteins from various surface pockets to the active site; instead small changes at various regions on the protein surface that lead to favourable changes in the energy landscape appears more realistic from an evolutionary perspective.

7.3 Future Work

Markov state models

Markov state models (MSMs) offer a powerful means of characterising changes in the energy landscape as a result of ligand binding. MSMs are a discrete approximation to the energy landscape of a protein, such that the Markov matrix corresponding to the generated network of microstates contains all the information about the dynamics of the system, when considering a statistical *ensemble* of the protein at thermodynamic equilibrium[209]. Indeed, the fundamental approach is not especially new, being suggested by Zwanzig over 30 years ago[249] but a number of advances since have made MSMs a powerful modern technique for analysing protein function. As a result of the formulation of MSMs in terms of a variational approximation to the true propagator for the dynamical system[176], increases in computing power will allow for more and more accurate models for the energy landscape, whilst concomitantly producing better sampled MD simulations for larger systems. An obvious experiment then would be to simply simulate a particular allosteric protein both with and without allosteric ligands bound and assess the change in the populations of the various states. Traditionally, thermodynamic models of allostery have tended to be explained in terms of just two states: the active and inactive state and the relative adjustment between them. In the very high dimensional space of a protein, this mental model is clearly inadequate; for example studies of c-src tyrosine kinase showed a potential allostetic site in an intermediate state of the protein[217], which would not seen by only observing end states in the form of crystal structures.

Recently, more sophisticated methods for generating MSMs have been developed that leverage deep learning[93, 150]. Rather than employ a pipeline involving selection of features, dimensionality reduction and clustering to generate the MSM, the entire procedure is encoded in a single model that outputs the MSM directly from the MD trajectory. Furthermore, whilst methods such as tICA are easily interpretable, they are still ultimately linear decompositions of a highly dimensional, complex time series data and these newer, nonlinear methods have the potential to extract more meaningful information. Such approaches should simplify the use of MSMs even further.

Markov stability was introduced by Delvenne *et al*[55] as a general method for community detection on graphs, using the framework of a random walk on the network[130]. Qualitatively, Markov stability says that a *partition* of a network into a set of *communities* is "good", when a random walker placed on the network tends to stay within its starting community for a given *Markov time*. Thus we expect that for short Markov times, a large number of small communities will be found and as we move towards longer and longer times, larger communities will be favoured by the algorithm. Markov stability is a measure of the *clustered autocovariance* of the Markov process on the network: $\operatorname{cov}[X_t, X_{t+\tau}] = \mathbb{E}[X_t, X_{t+\tau}] - \mathbb{E}[X_t]\,\mathbb{E}[X_{t+\tau}]$, which can be written in matrix form as:

$$R(t) = H^T \left[\Pi T^\tau - \pi^T \pi\right] H \tag{7.1}$$

where $\Pi = \operatorname{diag}(\pi)$, π being the stationary distribution of the random walk. H is an $N \times c$ indicator matrix that possesses an entry 1 at entry H_{ij} if node i is in community j and 0 otherwise with c the number of communities. As Markov stability is defined in terms of the dynamics of probability flow on a network, it would make an ideal method for *clumping*, which is the term usually used in the MSM literature for grouping together microstates into larger macrostates[175], thus allowing a full mapping of the various resolutions of the free energy landscape.

Allostery through perturbation of charge

One aspect of allostery that we did not focus on in this report is the role of *charge*. Elastic models in particular are not well equipped to handle electrostatics but recent evidence suggests that charge redistribution as a result of ligand binding alters internal dynamics of proteins[141, 129]. Whilst here we extended the ideas of *bond-to-bond propensity* to a three dimensional mechanical model, the original motivation for the method came from electrical networks[206] and it is possible that in fact propensity captures some element of charge redistribution.

The equations are indeed similar to those of Chen and Martinez[34, 35] who studied fluctuating charge models. Here, a set of atoms each has an electronegativity value χ_i (playing the role of potential in the equilibrium equations for circuits) and each pair of atoms has a Coulomb interaction J_{ij} (analogous to the conductance in a circuit). When these atoms are brought together, such that there is a non-negligible Coulombic force between them, charge will transfer between the atoms according to the values of J_{ij} so as to make the sum of the pairwise differences in electronegativity 0 (which is simply the corresponding form of Kirchoff's Voltage Law for the system). The equation that is being solved is:

$$\chi = Jq \tag{7.2}$$

where J is analogous to the Laplacian L in *bond-to-bond propensity* or the stiffness matrix K in *elastic response* and as with those methods, the equation is actually formulated in the bond space. It seems possible that there is some relationship between *bond-to-bond propensity* and charge transfer and this seems worthy of future investigation. The Coulombic interactions falls off as $\frac{1}{r}$ and thus is very long ranged, in contrast to the two-centre Hooke springs on Chapter 5 and thus electrostatic perturbations may be more likely to induce changes distal to the allosteric site. Indeed ATP itself, the allosteric ligand of interest in Chapter 4 has been posited to mediate long range allosteric effects in myosin as a result of anistropic charge redistribution[204]. One of the difficulties of studying this approach thus far has been that MD simulations have tended to use non-polarizable force fields, where each atom acts as a fixed point charge throughout the simulation. With increasing computational power and specialised hardware[214], it is now becoming possible to implement polarizable force fields[216, 144] and this is offers a clear future direction for research into allostery.

Machine learning approaches

The popularity of machine learning has taken off in the last few years, aided by increases in computing power (particularly highly parallelisable GPU set ups) and the explosion in the quantity of available data. However, statistical approaches have a long history in the study of protein dynamics, particularly in the field of protein folding[199, 115, 6]. Machine learning offers an orthogonal, top down alternative to physical modelling methods such as MD or elastic network models. Recently, a number of deep learning approaches have appeared[106, 219], made possible a now reasonably large crystal structure dataset courtesy of the Protein Data Bank, which now has around 100,000 entries[152].

Another method that has gained popularity is the representation of molecules as *graphs*, where similarly to our approach in Chapter 4, 3-dimensional information is neglected in favour of a more simplistic, *connectivity focused* representation of the molecule that allows application of modern convolutional neural net methods[113, 60]. Such methods could be extended to the protein graphs we have used in this thesis, though labelling of allosteric sites would still likely need to be a time consuming manual task. It could in fact also be more powerful still to combine both the bottom up approach of MD with machine learning methods by providing both dynamical information and the far larger data sets that result from what is effectively a "video" of a protein. Currently, MD analysis is still largely confined to one protein at a time, but a robust statistical approach may allow more powerful generalisations to be made in the future.

Bibliography

[1] S. a. Adcock and J. A. Mccammon, *Molecular Dynamics : Survey of Methods for Simulating the Activity of Proteins Molecular Dynamics : Survey of Methods for Simulating the Activity of Proteins*, 106 (2006), pp. 1589–1615.

[2] S. D. Akshay Agrawal Robin Verschueren and S. Boyd, *A Rewriting System for Convex Optimization Problems*, Journal of Control and Decision, 5 (2018), pp. 42–60.

[3] B. J. Alder and T. Wainwright, *Phase transition for a hard sphere system*, The Journal of chemical physics, 27 (1957), pp. 1208–1209.

[4] B. Amor, S. N. Yaliraki, R. Woscholski, and M. Barahona, *Uncovering allosteric pathways in caspase-1 with Markov transient analysis and multiscale community detection*, Molecular bioSystems, 10 (2014), pp. 2247–2258.

[5] B. R. Amor, M. T. Schaub, S. N. Yaliraki, and M. Barahona, *Prediction of allosteric sites and mediating interactions through bond-to-bond propensities*, Nature Communications, 7 (2016), pp. 1–13.

[6] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling*, Bioinformatics, 22 (2006), pp. 195–201.

[7] K. Arora and C. L Brooks Iii, *Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism*, Proceedings of the National Academy of Sciences, 104 (2007).

[8] G. A. Arteca, *Scaling regimes of molecular size and self-entanglements in very compact proteins*, Physical Review E, 51 (1995), p. 2600.

[9] A. R. Atilgan, P. Akan, and C. Baysal, *Small-world communication of residues and significance for protein dynamics*, Biophysical journal, 86 (2004), pp. 85–91.

[10] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, *Anisotropy of fluctuation dynamics of proteins with an elastic network model*, Biophysical Journal, 80 (2001), pp. 505–515.

[11] C. Atilgan, Z. N. Gerek, S. B. Ozkan, and A. R. Atilgan, *Manipulation of conformational change in proteins by single-residue perturbations*, Biophysical Journal, 99 (2010), pp. 933–943.

[12] I. Bahar, A. R. Atilgan, and B. Erman, *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.*, Folding & design, 2 (1997), pp. 173–181.

[13] I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava, *Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins*, Chemical reviews, 110 (2009), pp. 1463–1497.

[14] D. P. Baker and E. R. Kantrowitz, *The conserved residues glutamate-37, aspartate-100, and arginine-269 are important for the structural stabilization of Escherichia coli aspartate transcarbamoylase.*, Biochemistry, 32 (1993), pp. 10150–10158.

[15] I. A. Balabin, W. Yang, and D. N. Beratan, *Coarse-grained modeling of allosteric regulation in protein receptors*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 14253–14258.

[16] A.-L. Barabási, *Scale-free networks: a decade and beyond*, science, 325 (2009), pp. 412–413.

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The Protein Data Bank.*, Nucleic acids research, 28 (2000), pp. 235–242.

[18] N. G. J. C. Biol, N. G. J. C. Biol, Y. J. Am, A. R. Liss, N. York, U. I. A. Biochem, S. A. Middleton, J. W. Stebbins, and E. R. Kantrowitz, *Forms of*, (1989), pp. 143–152.

[19] S. J. Blundell and K. M. Blundell, *Concepts in thermal physics*, OUP Oxford, 2009.

[20] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Transition path sampling: Throwing ropes over rough mountain passes, in the dark*, Annual review of physical chemistry, 53 (2002), pp. 291–318.

[21] P. Bonacich, *Factoring and weighting approaches to status scores and clique identification*, Journal of mathematical sociology, 2 (1972), pp. 113–120.

[22] G. R. Bowman and P. L. Geissler, *Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites*, Proceedings of the National Academy of Sciences, 109 (2012), pp. 11681–11686.

[23] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[24] S. R. Broadbent and J. M. Hammersley, *Percolation processes: I. Crystals and mazes*, in Mathematical Proceedings of the Cambridge Philosophical Society, vol. 53, Cambridge University Press, 1957, pp. 629–641.

[25] B. R. Brooks, C. L. Brooks, A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, and Others, *CHARMM: the biomolecular simulation program*, Journal of computational chemistry, 30 (2009), pp. 1545–1614.

[26] S. BruÌĹschweiler, P. Schanda, K. Kloiber, B. Brutscher, G. Kontaxis, R. Konrat, and M. Tollinger, *Direct observation of the dynamic process underlying allosteric signal transmission*, Journal of the American Chemical Society, 131 (2009), pp. 3063–3068.

[27] L. Bu and J. E. Straub, *Vibrational Energy Relaxation of âĂIJTailoredâĂİ Hemes in Myoglobin Following Ligand Photolysis Supports Energy Funneling Mechanism of Heme âĂIJCoolingâĂİ*, The Journal of Physical Chemistry B, 107 (2003), pp. 10634–10639.

[28] D. Bucher, B. J. Grant, and J. A. McCammon, *Induced fit or conformational selection? The role of the semi-closed state in the maltose binding protein*, Biochemistry, 50 (2011), pp. 10530–10539.

[29] G. Buhrman, G. Holzapfel, S. Fetics, and C. Mattos, *Allosteric modulation of Ras positions Q61 for a direct role in catalysis*, Proceedings of the National Academy of Sciences, 107 (2010), pp. 4931–4936.

[30] A. S. Burgen, *Conformational changes and drug action.*, in Federation proceedings, vol. 40, 1981, pp. 2723–2728.

[31] S. K. Burley and G. A. Petsko, *Aromatic-aromatic interaction: a mechanism of protein structure stabilization*, Science, 229 (1985), pp. 23–28.

[32] P. Campitelli, J. Guo, H.-X. Zhou, and S. B. Ozkan, *Hinge-shift mechanism modulates allosteric regulations in human Pin1*, The Journal of Physical Chemistry B, 122 (2018), pp. 5623–5629.

[33] J.-P. Changeux, *50 Years of Allosteric Interactions: the Twists and Turns of the Models*, Nature Reviews Molecular Cell Biology, 14 (2013), pp. 819–829.

[34] J. Chen, D. Hundertmark, and T. J. Martínez, *A unified theoretical framework for fluctuating-charge models in atom-space and in bond-space*, Journal of Chemical Physics, 129 (2008).

[35] J. Chen and T. J. Martínez, *Charge conservation in electronegativity equalization and its implications for the electrostatic properties of fluctuating-charge models*, Journal of Chemical Physics, 131 (2009), pp. 129–132.

[36] J. Cherfils, P. Vachette, P. Tauc, and J. Janin, *The pAR5 mutation and the allosteric mechanism of Escherichia coli aspartate carbamoyltransferase*, Embo J, 6 (1987), pp. 2843–2847.

[37] S. I. Chiosea, M. Miller, and R. R. Seethala, *HRAS mutations in epithelial–myoepithelial carcinoma*, Head and neck pathology, 8 (2014), pp. 146–150.

[38] S. I. Chiosea, L. Williams, C. C. Griffith, L. D. R. Thompson, I. Weinreb, J. E. Bauman, A. Luvison, S. Roy, R. R. Seethala, and M. N. Nikiforova, *Molecular characterization of apocrine salivary duct carcinoma*, The American journal of surgical pathology, 39 (2015), pp. 744–752.

[39] J. D. Chodera and F. Noé, *Markov state models of biomolecular conformational dynamics*, Current opinion in structural biology, 25 (2014), pp. 135–144.

[40] J. H. Choi, A. H. Laurent, V. J. Hilser, and M. Ostermeier, *Design of protein switches based on an ensemble model of allostery.*, Nature communications, 6 (2015), p. 6968.

[41] D. O. Clegg and D. E. Koshland, *The role of a signaling protein in bacterial sensing: behavioral effects of increased gene expression*, Proceedings of the National Academy of Sciences, 81 (1984), pp. 5056–5060.

[42] G. M. Cockrell, Y. Zheng, W. Guo, A. W. Peterson, J. K. Truong, and E. R. Kantrowitz, *New paradigm for allosteric regulation of escherichia coli aspartate transcarbamoylase*, Biochemistry, 52 (2013), pp. 8036–8047.

[43] A. Cooper and D. T. Dryden, *Allostery without conformational change - A plausible model*, European Biophysics Journal, 11 (1984), pp. 103–109.

[44] T. S. Corder and J. R. Wild, *Discrimination between nucleotide effector responses of aspartate transcarbamoylase due to a single site substitution in the allosteric binding site*, Journal of Biological Chemistry, 264 (1989), pp. 7425–7430.

[45] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*, Journal of the American Chemical Society, 117 (1995), pp. 5179–5197.

[46] J. R. Costa, *Infinitesimal and combinatorial rigidity approaches to coarse grain proteins*, PhD thesis, Imperial College London, 2008.

[47] P. Csermely, R. Palotai, and R. Nussinov, *Induced fit, conformational selection and independent dynamic segments: an extended view of binding events*, Trends in biochemical sciences, 35 (2010), pp. 539–546.

[48] B. I. Dahiyat, D. B. Gordon, and S. L. Mayo, *Automated design of the surface positions of protein helices.*, Protein science : a publication of the Protein Society, 6 (1997), pp. 1333–7.

[49] M. D. Daily, T. J. Upadhyaya, and J. J. Gray, *Contact rearrangements form coupled networks from local motions in allosteric proteins*, Proteins: Structure, Function and Genetics, 71 (2008), pp. 455–466.

[50] P.-G. De Gennes, *On a relation between percolation theory and the elasticity of gels*, Journal de Physique Lettres, 37 (1976), pp. 1–2.

[51] P. Dean, *A new Monte Carlo method for percolation problems on a lattice*, in Mathematical Proceedings of the Cambridge Philosophical Society, vol. 59, Cambridge University Press, 1963, pp. 397–410.

[52] A. Del Sol, H. Fujihashi, D. Amoros, and R. Nussinov, *Residues crucial for maintaining short paths in network communication mediate signaling in proteins*, Molecular Systems Biology, 2 (2006), pp. 1–12.

[53] A. del Sol, C. J. Tsai, B. Ma, and R. Nussinov, *The Origin of Allosteric Functional Modulation: Multiple Pre-existing Pathways*, Structure, 17 (2009), pp. 1042–1050.

[54] a. Delmotte, E. W. Tate, S. N. Yaliraki, and M. Barahona, *Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction.*, Physical biology, 8 (2011), p. 055010.

[55] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, *Stability of graph communities across time scales*, Proceedings of the National Academy of Sciences, 107 (2010), pp. 12755–12760.

[56] N. J. Dembowski and E. R. Kantrowitz, *The use of alanine scanning mutagenesis to determine the role of the n-terminus of the regulatory chain in the heterotropic mechanism of Escherichia coli aspartate transcarbamoylase*, Protein Engineering, Design and Selection, 7 (1994), pp. 673–679.

[57] S. Diamond and S. Boyd, *{CVXPY}: A {P}ython-Embedded Modeling Language for Convex Optimization*, Journal of Machine Learning Research, 17 (2016), pp. 1–5.

[58] T. Dobzhansky, *Nothing in biology makes sense except in the light of evolution*, The american biology teacher, 75 (2013), pp. 87–91.

[59] J. Du, M. Yang, S. Chen, D. Li, Z. Chang, and Z. Dong, *PDK1 promotes tumor growth and metastasis in a spontaneous breast cancer model*, Oncogene, 35 (2016), p. 3314.

[60] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, *Convolutional networks on graphs for learning molecular fingerprints*, in Advances in neural information processing systems, 2015, pp. 2224–2232.

[61] E. C. Dykeman and O. F. Sankey, *Normal mode analysis and applications in biological physics*, Journal of Physics: Condensed Matter, 22 (2010), p. 423202.

[62] E. Eisenstein, D. W. Markby, and H. K. Schachman, *Changes in stability and allosteric properties of aspartate transcarbamoylase resulting from amino acid substitutions in the zinc-binding domain of the regulatory chains.*, Proceedings of the National Academy of Sciences of the United States of America, 86 (1989), pp. 3094–8.

[63] M. B. Enright and D. M. Leitner, *Mass fractal dimension and the compactness of proteins*, Physical Review E, 71 (2005), p. 11912.

[64] D. L. Ermak and J. A. McCammon, *Brownian dynamics with hydrodynamic interactions*, The Journal of chemical physics, 69 (1978), pp. 1352–1360.

[65] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, and Others, *A density-based algorithm for discovering clusters in large spatial databases with noise.*, in Kdd, vol. 96, 1996, pp. 226–231.

[66] B. T. Falk, P. J. Sapienza, A. L. Lee, B. VanSchouwen, and G. Melacini, *Chemical shift imprint of intersubunit communication in a symmetric homodimer*, Proceedings of the National Academy of Sciences, 113 (2016), pp. 9533–9538.

[67] S. Feng and P. N. Sen, *Percolation on elastic networks: new exponent and threshold*, Physical review letters, 52 (1984), p. 216.

[68] X. Feng, Y. Deng, and H. W. J. Blöte, *Percolation transitions in two dimensions*, Physical Review E, 78 (2008), p. 31136.

[69] R. B. Fenwick, H. van den Bedem, J. S. Fraser, and P. E. Wright, *Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR*, Proceedings of the National Academy of Sciences, 111 (2014), pp. E445–E454.

[70] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, vol. 1, Basic books, 2011.

[71] H. Flechsig, *Article Design of Elastic Networks with Evolutionary Optimized Long-Range Communication as Mechanical Models of Allosteric Proteins*, Biophysical Journal, 113 (2017), pp. 558–571.

[72] P. Flory, *Statistical thermodynamics of random networks*, Proceedings of the Royal Society of London, 351 (1976), pp. 351–380.

[73] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, *The energy landscapes and motions of proteins*, Science, 254 (1991).

[74] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, *Conformational entropy in molecular recognition by proteins*, Nature, 448 (2007), p. 325.

[75] L. C. Freeman, *A Set of Measures of Centrality Based on Betweenness Author (s): Linton C . Freeman Published by : American Sociological Association Stable URL : http://www.jstor.org/stable/3033543 Accessed : 18-04-2016 12 : 00 UTC Your use of the JSTOR archive indicat*, Sociometry, 40 (1977), pp. 35–41.

[76] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, vol. 1, Elsevier, 2001.

[77] M. Frödin, T. L. Antal, B. A. Dümmler, C. J. Jensen, M. Deak, S. Gammeltoft, and R. M. Biondi, *A phosphoserine/threonine-binding pocket in AGC kinases and PDK1 mediates activation by hydrophobic motif phosphorylation*, The EMBO journal, 21 (2002), pp. 5396–5407.

[78] P. A. Gagliardi, A. Puliafito, and L. Primo, *PDK1: At the crossroad of cancer signaling pathways*, in Seminars in cancer biology, Elsevier, 2017.

[79] A. Garcia-pino, S. Balasubramanian, L. Wyns, E. Gazit, H. D. Greve, R. D. Magnuson, D. Charlier, N. A. J. V. Nuland, and R. Loris, *Allostery and Intrinsic Disorder Mediate Transcription Regulation by Conditional Cooperativity*, Cell, 142 (2010), pp. 101–111.

[80] J. Gerhart, *From feedback inhibition to allostery: The enduring example of aspartate transcarbamoylase*, FEBS Journal, 281 (2014), pp. 612–620.

[81] A. Ghosh, S. Boyd, and A. Saberi, *Minimizing Effective Resistance of a Graph*, SIAM Review, 50 (2008), pp. 37–66.

[82] A. Goncearenco, S. Mitternacht, T. Yong, B. Eisenhaber, F. Eisenhaber, and I. N. Berezovsky, *SPACER: server for predicting allosteric communication and effects of regulation*, Nucleic acids research, 41 (2013), pp. W266—-W272.

[83] J. E. Gouaux and W. N. Lipscomb, *Crystal structures of phosphonoacetamide ligated T and phosphonoacetamide and malonate ligated R states of aspartate carbamoyltransferase at 2.8-.ANG. resolution and neutral pH*, Biochemistry, 29 (1990), pp. 389–402.

[84] J. G. Greener and M. J. E. Sternberg, *AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis*, BMC bioinformatics, 16 (2015), p. 335.

[85] N. Greives and H.-x. Zhou, *Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit*, Proceedings of the National Academy of Sciences, 111 (2014), pp. 10197–10202.

[86] K. Gunasekaran, B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?*, Proteins: Structure, Function, and Bioinformatics, 57 (2004), pp. 433–443.

[87] J. Günther, A. Bergner, M. Hendlich, and G. Klebe, *Utilising structural knowledge in drug design strategies: Applications using relibase*, Journal of Molecular Biology, 326 (2003), pp. 621–636.

[88] J. Guo and H.-X. Zhou, *Protein Allostery and Conformational Dynamics*, Chemical Reviews, (2016), p. acs.chemrev.5b00590.

[89] M. Gur, J. D. Madura, and I. Bahar, *Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase*, Biophysical journal, 105 (2013), pp. 1643–1652.

[90] T. Haliloglu, I. Bahar, and B. Erman, *Gaussian dynamics of folded proteins*, Physical Review Letters, 79 (1997), pp. 3090–3093.

[91] P. Hamm, M. Lim, and R. M. Hochstrasser, *Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy*, The Journal of Physical Chemistry B, 102 (1998), pp. 6123–6138.

[92] K. Henzler-Wildman and D. Kern, *Dynamic personalities of proteins.*, Nature, 450 (2007), pp. 964–972.

[93] C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, *Variational encoding of complex dynamics*, Physical Review E, 97 (2018), p. 62412.

[94] V. Hindie, A. Stroba, H. Zhang, L. A. Lopez-Garcia, L. Idrissova, S. Zeuzem, D. Hirschberg, F. Schaeffer, T. J. D. Jørgensen, M. Engel, and Others, *Structure and allosteric effects of low-molecular-weight activators on the protein kinase PDK1*, Nature chemical biology, 5 (2009), p. 758.

[95] W. Huang, S. Lu, Z. Huang, X. Liu, L. Mou, Y. Luo, Y. Zhao, Y. Liu, Z. Chen, T. Hou, and Others, *Allosite: a method for predicting allosteric sites*, Bioinformatics, 29 (2013), pp. 2357–2359.

[96] T. I. Igumenova, K. K. Frederick, and A. J. Wand, *Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution*, Chemical reviews, 106 (2006), pp. 1672–1699.

[97] B. Isralewitz, M. Gao, and K. Schulten, *Steered molecular dynamics and mechanical functions of proteins*, Current Opinion in Structural Biology, 11 (2001), pp. 224–230.

[98] D. J. Jacobs, *Generic rigidity in three-dimensional bond-bending networks*, Journal of Physics A: Mathematical and General, 31 (1998), p. 6653.

[99] D. J. Jacobs and B. Hendrickson, *An algorithm for two-dimensional rigidity percolation: the pebble game*, Journal of Computational Physics, 137 (1997), pp. 346–365.

[100] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, *Protein flexibility predictions using graph theory.*, Proteins, 44 (2001), pp. 150–65.

[101] D. J. Jacobs and M. F. Thorpe, *Generic Rigidity Percolation: The Pebble Game*, Phys. Rev. Lett., 75 (1995), pp. 4051–4054.

[102] ——, *Computer-implemented system for analyzing rigidity of substructures within a macromolecule*, 2000.

[103] J. Janin and M. J. E. Sternberg, *Protein flexibility, not disorder, is intrinsic to molecular recognition*, F1000 biology reports, 5 (2013).

[104] V. A. Jarymowycz and M. J. Stone, *Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences*, Chemical reviews, 106 (2006), pp. 1624–1671.

[105] M. J.Huheey, E. Keiter, R.Keiter, *Inorganic chemistry: principles of structure and reactivity*, Pearson Education India, 2006.

[106] J. Jiménez, S. Doerr, G. Mart\'\inez-Rosell, A. S. Rose, and G. De Fabritiis, *DeepSite: protein-binding site predictor using 3D-convolutional neural networks*, Bioinformatics, 33 (2017), pp. 3036–3042.

[107] L. Jin, B. Stec, W. N. Lipscomb, and E. R. Kantrowitz, *Insights into the mechanisms of catalysis and heterotropic regulation of Escherichia coli aspartate transcarbamoylase based upon a structure of the enzyme complexed with the bisubstrate analogue N-phosphonacetyl-L-aspartate at 2.1 ??*, Proteins: Structure, Function and Genetics, 37 (1999), pp. 729–742.

[108] W. L. Jorgensen and J. Tirado-Rives, *The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin*, Journal of the American Chemical Society, 110 (1988), pp. 1657–1666.

[109] P. E. R. Jr, *A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers*, J. Chem. Phys., 21 (1953), p. 1272.

[110] R. Kalescky, H. Zhou, J. Liu, and P. Tao, *Rigid residue scan simulations systematically reveal residue entropic roles in protein allostery*, PLoS computational biology, 12 (2016), p. e1004893.

[111] E. R. Kantrowitz, *Allostery and cooperativity in Escherichia coli aspartate transcarbamoylase.*, Archives of biochemistry and biophysics, 519 (2012), pp. 81–90.

[112] H. Ke, W. N. Lipscomb, Y. Cho, and R. B. Honzatko, *Complex of N-phosphonacetyl-l-aspartate with aspartate carbamoyltransferase. X-ray refinement, analysis of conformational changes and catalytic and allosteric mechanisms*, Journal of Molecular Biology, 204 (1988), pp. 725–747.

[113] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, *Molecular graph convolutions: moving beyond fingerprints*, Journal of computer-aided molecular design, 30 (2016), pp. 595–608.

[114] P. N. Keating, *Effect of invariance requirements on the elastic strain energy of crystals with application to the diamond structure*, Physical Review, 145 (1966), pp. 637–645.

[115] L. A. Kelley and M. J. E. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server*, Nature protocols, 4 (2009), p. 363.

[116] J. A. Kelner and L. Orecchia, *A Simple , Combinatorial Algorithm for Solving SDD Systems in Nearly-Linear Time*, pp. 911–920.

[117] H. Kesten, *The critical probability of bond percolation on the square lattice equals 1/2*, Communications in mathematical physics, 74 (1980), pp. 41–59.

[118] Y. Kholodenko, M. Volk, E. Gooding, and R. M. Hochstrasser, *Energy dissipation and relaxation processes in deoxy myoglobin after photoexcitation in the Soret region*, Chemical Physics, 259 (2000), pp. 71–87.

[119] S. Kirkpatrick, *Percolation ancl Conduction*, Re, 45 (1973).

[120] D. J. Klein and M. Randić, *Resistance distance*, Journal of mathematical chemistry, 12 (1993), pp. 81–95.

[121] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, Journal of the ACM (JACM), 46 (1999), pp. 604–632.

[122] R. Koenker, *quantreg: Quantile Regression. R package version 5.19*, R Foundation for Statistical Computing: Vienna) Available at: http://CRAN. R-project. org/package= quantreg, (2015).

[123] R. Koenker and G. Bassett, *Quantile Regression*, Econometrica, 46 (1978), pp. 33–50.

[124] A. P. Kornev and S. S. Taylor, *Dynamics-Driven Allostery in Protein Kinases*, Trends in Biochemical Sciences, 40 (2015), pp. 628–647.

[125] D. E. Koshland, *Application of a theory of enzyme specificity to protein synthesis*, Proceedings of the National Academy of Sciences, 44 (1958), pp. 98–104.

[126] D. E. Koshland Jr, G. Némethy, and D. Filmer, *Comparison of experimental binding data and theoretical models in proteins containing subunits*, Biochemistry, 5 (1966), pp. 365–385.

[127] M. Koyama, S. Neya, and Y. Mizutani, *Role of heme propionates of myoglobin in vibrational energy relaxation*, Chemical Physics Letters, 430 (2006), pp. 404–408.

[128] P. Kukura, D. W. McCamant, S. Yoon, D. B. Wandschneider, and R. A. Mathies, *Structural observation of the primary isomerization in vision with femtosecond-stimulated Raman*, Science, 310 (2005), pp. 1006–1009.

[129] A. Kumawat and S. Chakrabarty, *Hidden electrostatic basis of dynamic allostery in a PDZ domain*, Proceedings of the National Academy of Sciences, 114 (2017), pp. E5825—-E5834.

[130] R. Lambiotte, J.-c. Delvenne, and M. Barahona, *Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks*, IEEE Transactions on Network Science and Engineering, 1 (2014), pp. 76–90.

[131] S. Lampa-Pastirk and W. F. Beck, *Intramolecular vibrational preparation of the unfolding transition state of ZnII-substituted cytochrome c*, The Journal of Physical Chemistry B, 110 (2006), pp. 22971–22974.

[132] D. M. Leitner, *Energy flow in proteins.*, Annual review of physical chemistry, 59 (2008), pp. 233–59.

[133] E. Lerner, E. DeGiuli, G. Düring, and M. Wyart, *Breakdown of continuum elasticity in amorphous solids*, Soft Matter, 10 (2014), pp. 5085–5092.

[134] C. Levinthal, *How to fold graciously*, Mossbauer spectroscopy in biological systems, 67 (1969), pp. 22–24.

[135] G. Li, D. Magana, and R. B. Dyer, *Anisotropic energy flow and allosteric ligand binding in albumin.*, Nature communications, 5 (2014), p. 3100.

[136] J. Liang and K. A. Dill, *Are Proteins Well-Packed ?*, Biophysical Journal, 81 (2001), pp. 751–766.

[137] M. S. Lin, N. L. Fawzi, and T. Head-Gordon, *Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction*, Structure, 15 (2007), pp. 727–740.

[138] A. B. Lindner, Z. Eshhar, and D. S. Tawfik, *Conformational changes affect binding and catalysis by ester-hydrolyzing antibodies*, J.Mol.Biol., 285 (1999), pp. 421–430.

[139] W. N. Lipscomb and E. R. Kantrowitz, *Structure and Mechanisms of Escherichia coli Aspartate Transcarbamoylase.*, Accounts of chemical research, 45 (2012), pp. 444–53.

[140] C.-h. Liu, S. R. Nagel, D. A. Schecter, S. N. Coppersmith, S. Majumdar, O. Narayan, and T. A. Witten, *Force fluctuations in bead packs*, Science, 269 (1995), pp. 513–515.

[141] J. Liu and R. Nussinov, *Energetic redistribution in allostery to execute protein function*, Proceedings of the National Academy of Sciences, 114 (2017), pp. 7480–7482.

[142] S. W. Lockless, R. Ranganathan, P. Kukic, C. Mirabello, G. Tradigo, I. Walsh, P. Veltri, G. Pollastri, M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, R. Ranganathan, K. E. Kreth, A. a. Fodor, S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmead, J. M. Thornton, C. a. Orengo, a. E. Todd, and F. M. Pearl, *Evolutionarily conserved pathways of energetic connectivity in protein families*, BMC Bioinformatics, 15 (1999), pp. 295–299.

[143] S. W. Lockless, M. A. Wall, and R. Ranganathan, *Evolutionarily conserved networks of residues mediate allosteric communication in proteins*, Nature structural biology, 10 (2003).

[144] P. E. M. Lopes, J. Huang, J. Shim, Y. Luo, H. Li, B. Roux, and A. D. MacKerell Jr, *Polarizable force field for peptides and proteins based on the classical drude oscillator*, Journal of chemical theory and computation, 9 (2013), pp. 5430–5449.

[145] G. Lu, E. L. Giroux, and E. R. Kantrowitz, *Importance of the Dimer-Dimer Interface for Allosteric Signal Transduction and AMP Cooperativity of Pig Kidney Fructose-1, 6-Bisphosphatase Site-Specific Mutagenesis Studies of Glu-192 and Asp-187 Residues on the 190's Loop*, Journal of Biological Chemistry, 272 (1997), pp. 5076–5081.

[146] B. Ma, S. Kumar, and C.-j. Tsai, *Folding funnels and binding mechanisms*, Protein Engineering, 12 (1999), pp. 713–720.

[147] L. Ma and Q. Cui, *Activation mechanism of a signaling protein at atomic resolution from advanced computations*, Journal of the American Chemical Society, 129 (2007), pp. 10261–10268.

[148] T. S. Majmudar, *TS Majmudar and RP Behringer, Nature (London) 435, 1079 (2005).*, Nature (London), 435 (2005), p. 1079.

[149] R. D. Malmstrom, A. P. Kornev, S. S. Taylor, and R. E. Amaro, *cAMP activation of a canonical signalling domain*, Nature Communications, 6 (2015), pp. 1–11.

[150] A. Mardt, L. Pasquali, H. Wu, and F. Noé, *VAMPnets for deep learning of molecular kinetics*, Nature communications, 9 (2018), p. 5.

[151] N. Masuda, M. A. Porter, and R. Lambiotte, *Random walks and diffusion on networks*, Physics Reports, 716-717 (2017), pp. 1–58.

[152] B. W. Matthews, *Which of the 100,000 structures in the protein data bank are reliable?*, Protein Science, 24 (2015), pp. 589–591.

[153] J. C. Maxwell, *JC Maxwell, Philos. Mag. 27, 294 (1864).*, Philos. Mag., 27 (1864), p. 294.

[154] J. A. McCammon, C. Y. Lee, and S. H. Northrup, *Side-Chain Rotational Isomerization in Proteins: A Mechanism Involving Gating and Transient Packing Defects*, Journal of the American Chemical Society, 105 (1983), pp. 2232–2237.

[155] F. McCormick, *Ras-related proteins in signal transduction and growth control*, Molecular reproduction and development, 42 (1995), pp. 500–506.

[156] L. R. McDonald, J. A. Boyer, and A. L. Lee, *Segmental motions, not a two-state concerted switch, underlie allostery in CheY*, Structure, 20 (2012), pp. 1363–1373.

[157] L. R. McDonald, M. J. Whitley, J. A. Boyer, and A. L. Lee, *Colocalization of fast and slow timescale dynamics in the allosteric signaling protein CheY*, Journal of molecular biology, 425 (2013), pp. 2372–2381.

[158] T. C. B. McLeish, M. J. Cann, and T. L. Rodgers, *Dynamic transmission of protein allostery without structural change: spatial pathways or global modes?*, Biophysical journal, 109 (2015), pp. 1240–1250.

[159] T. C. B. McLeish, T. L. Rodgers, and M. R. Wilson, *Allostery without conformation change: modelling protein dynamics at multiple scales*, Physical biology, 10 (2013), p. 56004.

[160] N. MEJ., *The structure and function of complex networks*, SIAM Rev, 45 (2003), p. 167.

[161] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines*, Journal of Chemical Physics, 1087 (1953).

[162] M. R. Mitchell, T. Tlusty, and S. Leibler, *Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings*, Proceedings of the National Academy of Sciences, (2016), p. 201609462.

[163] S. Mitternacht and I. N. Berezovsky, *Binding leverage as a molecular basis for allosteric regulation*, PLoS computational biology, 7 (2011), p. e1002148.

[164] L. Molgedey and H. G. Schuster, *Separation of a mixture of independent signals using time delayed correlations*, Physical review letters, 72 (1994), p. 3634.

[165] J. Monod, J.-P. Changeux, and F. Jacob, *Allosteric proteins and cellular control systems*, Journal of Molecular Biology, 6 (1963), pp. 306–329.

[166] J. Monod, J. Wyman, and J.-P. Changeux, *On the nature of allosteric transitions: a plausible model*, J Mol Biol, 12 (1965), pp. 88–118.

[167] A. Mora, D. Komander, D. M. F. van Aalten, and D. R. Alessi, *PDK1, the master regulator of AGC kinase signal transduction*, in Seminars in cell & developmental biology, vol. 15, Elsevier, 2004, pp. 161–170.

[168] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, *The ensemble nature of allostery*, Nature, 508 (2014), pp. 331–339.

[169] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, *Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding*, Structure, 4 (1996), pp. 147–156.

[170] T. Nakayama, K. Yakubo, and R. L. Orbach, *Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations*, Reviews of modern physics, 66 (1994), p. 381.

[171] J. Newell and H. K. Schachman, *Amino acid substitutions which stabilize aspartate transcarbamoylase in the R state disrupt both homotropic and heterotropic effects*, 37 (1990), pp. 183–196.

[172] M. Newman, *Networks: An Introduction*, Networks: An Introduction, (2010), pp. 1–784.

[173] M. E. J. Newman and R. M. Ziff, *Efficient Monte Carlo algorithm and high-precision results for percolation*, Physical Review Letters, 85 (2000), p. 4104.

[174] C. J. Newton and E. R. Kantrowitz, *Importance of domain closure for homotropic cooperativity in Escherichia coli aspartate transcarbamylase.*, Biochemistry, 29 (1990), pp. 1444–1451.

[175] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, *Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states*, The Journal of chemical physics, 126 (2007), p. 04B617.

[176] F. Noé and F. Nuske, *A variational approach to modeling slow processes in stochastic dynamical systems*, Multiscale Modeling & Simulation, 11 (2013), pp. 635–655.

[177] J. R. Norris, *Markov chains*, no. 2, Cambridge university press, 1998.

[178] S. H. Northrup and J. A. Mccammon, *Simulation Methods for Protein Structure Fluctuations*, Biopolymers, 19 (1980), pp. 1001–1016.

[179] R. Nussinov, *Introduction to Protein Ensembles and Allostery*, Chemical Reviews, 116 (2016), pp. 6263–6266.

[180] R. Nussinov and C.-J. Tsai, *Allostery in disease and in drug discovery*, Cell, 153 (2013), pp. 293–305.

[181] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd, *Conic optimization via operator splitting and homogeneous self-dual embedding*, Journal of Optimization Theory and Applications, 169 (2016), pp. 1042–1068.

[182] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, *Toward an outline of the topography of a realistic protein-folding funnel.*, Proceedings of the National Academy of Sciences, 92 (1995), pp. 3626–3630.

[183] N. Ota and D. A. Agard, *Intramolecular Signaling Pathways Revealed by Modeling Anisotropic Thermal Diffusion*, Journal of Molecular Biology, 351 (2005), pp. 345–354.

[184] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank citation ranking: Bringing order to the web.*, tech. rep., Stanford InfoLab, 1999.

[185] F. Palazzesi, A. Barducci, M. Tollinger, and M. Parrinello, *The allosteric communication pathways in KIX domain of CBP*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 14237–14242.

[186] V. S. Pande, K. Beauchamp, and G. R. Bowman, *Everything you wanted to know about Markov State Models but were afraid to ask*, Methods, 52 (2010), pp. 99–105.

[187] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio, and R. Nussinov, *The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery*, Chemical Reviews, 116 (2016), pp. 6391–6423.

[188] M. F. Perutz, *Stereochemistry of Cooperative Effects in Haemoglobin1*, in From theoretical physics to biology, Karger Publishers, 1973, pp. 247–285.

[189] J. C. Phillips and M. F. Thorpe, *Constraint theory, vector percolation and glass formation*, Solid State Communications, 53 (1985), pp. 699–702.

[190] H. Qin, L. Lim, and J. Song, *Protein dynamics at Eph receptor-ligand interfaces as revealed by crystallography, NMR and MD simulations*, BMC Biophysics, 5 (2012), pp. 1–11.

[191] A. J. Rader, B. M. Hespenheide, L. A. Kuhn, and M. F. Thorpe, *Protein unfolding: rigidity lost*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 3540–3545.

[192] C. Raimondi and M. Falasca, *Targeting PDK1 in cancer*, Current medicinal chemistry, 18 (2011), pp. 2763–2769.

[193] R. Rammal and G. Toulouse, *Random walks on fractal structures and percolation clusters*, Journal de Physique Lettres, 44 (1983), pp. 13–22.

[194] D. C. Rapaport, *The art of molecular dynamics simulation*, Cambridge university press, 2004.

[195] A. A. S. T. Ribeiro and V. Ortiz, *Energy propagation and network energetic coupling in proteins*, Journal of Physical Chemistry A, 119 (2015), pp. 1835–1846.

[196] A. A. S. T. Ribeiro and V. Ortiz, *A Chemical Perspective on Allostery*, Chemical Reviews, (2016), p. acs.chemrev.5b00543.

[197] J. W. Rocks, N. Pashine, I. Bischofberger, C. P. Goodrich, A. J. Liu, and S. R. Nagel, *Designing allostery-inspired response in mechanical networks*, Proceedings of the National Academy of Sciences, 114 (2017), pp. 2520–2525.

[198] T. L. Rodgers, P. D. Townsend, D. Burnell, M. L. Jones, S. A. Richards, T. C. B. McLeish, E. Pohl, M. R. Wilson, and M. J. Cann, *Modulation of global low-frequency motions underlies*

allosteric regulation: demonstration in CRP/FNR family transcription factors, PLoS biology, 11 (2013), p. e1001651.

[199] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker, *Protein structure prediction using Rosetta*, in Methods in enzymology, vol. 383, Elsevier, 2004, pp. 66–93.

[200] M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte, *Memory in network flows and its effects on spreading dynamics and community detection*, Nature Communications, 5 (2014), pp. 1–13.

[201] G. S. Rule and T. K. Hitchens, *Fundamentals of protein NMR spectroscopy*, vol. 5, Springer Science & Business Media, 2006.

[202] J. D. Sadowsky, M. A. Burlingame, D. W. Wolan, C. L. McClendon, M. P. Jacobson, and J. A. Wells, *Turning a protein kinase on or off from a single allosteric site via disulfide trapping*, Proceedings of the National Academy of Sciences, (2011).

[203] A. Sato and Y. Mizutani, *Picosecond structural dynamics of myoglobin following photodissociation of carbon monoxide as revealed by ultraviolet time-resolved resonance Raman spectroscopy*, Biochemistry, 44 (2005), pp. 14709–14714.

[204] T. Sato, J. Ohnuki, and M. Takano, *Dielectric allostery of protein: Response of myosin to ATP binding*, The Journal of Physical Chemistry B, 120 (2016), pp. 13047–13055.

[205] M. Sayar, M. C. Demirel, and A. R. Atilgan, *Dynamics of disordered structures: effect of non-linearity on the localization*, Journal of Sound Vibration, 205 (1997), pp. 372–379.

[206] M. Schaub, J. Lehmann, S. Yaliraki, and M. Barahona, *Structure of complex networks: Quantifying edge-to-edge relations by failure-induced flow redistribution*, Network Science, 2 (2014), pp. 1–24.

[207] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J. Tessone, and F. Schweitzer, *Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks*, Nature Communications, 5 (2014).

[208] M. Schuster, R. E. Silversmith, and R. B. Bourret, *Conformational coupling in the chemotaxis response regulator CheY*, Proceedings of the National Academy of Sciences, 98 (2001), pp. 6003–6008.

[209] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, Journal of Computational Physics, 151 (1999), pp. 146–168.

[210] G. M. Schütz and S. Trimper, *Elephants can always remember: Exact long-range memory effects in a non-Markovian random walk*, Phys. Rev. E, 70 (2004), p. 45101.

[211] E. Shakhnovich, *Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet*, Chemical Reviews, 106 (2006), pp. 1559–1588.

[212] K. Sharp and J. J. Skinner, *Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling*, Proteins: Structure, Function, and Bioinformatics, 65 (2006), pp. 347–361.

[213] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, and Others, *Millisecond-scale molecular dynamics simulations on Anton*, in Proceedings of the conference on high performance computing networking, storage and analysis, ACM, 2009, p. 39.

[214] D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, and Others, *Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer*, in Proceedings of the international conference for high performance computing, networking, storage and analysis, IEEE Press, 2014, pp. 41–53.

[215] D. E. Shaw, P. Maragakis, K. Lindorff-larsen, S. Piana, Y. Shan, and W. Wriggers, *Atomic-Level Characterization of the Structural Dynamics of Proteins*, Science, 330 (2010), pp. 341–347.

[216] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, *Polarizable atomic multipole-based AMOEBA force field for proteins*, Journal of chemical theory and computation, 9 (2013), pp. 4046–4063.

[217] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, *Activation pathway of Src kinase reveals intermediate states as targets for drug design*, Nature communications, 5 (2014), p. 3397.

[218] A. I. Shulman, C. Larson, D. J. Mangelsdorf, and R. Ranganathan, *Structural determinants of allosteric ligand activation in RXR heterodimers*, Cell, 116 (2004), pp. 417–429.

[219] M. Skalic, A. Varela-Rial, J. Jiménez, G. Mart\'\inez-Rosell, and G. De Fabritiis, *LigVoxel: Inpainting binding pockets using 3D-convolutional neural networks.*, Bioinformatics (Oxford, England), (2018).

[220] D. A. Spielman, *Nearly-Linear Time Algorithms for Graph Partitioning , Graph Sparsification , and Solving Linear Systems [Extended Abstract]*, (2004), pp. 81–90.

[221] D. Stauffer and A. Aharony, *Introduction to percolation theory: revised second edition*, CRC press, 2014.

[222] A. Stein, M. Rueda, A. Panjkovich, M. Orozco, and P. Aloy, *A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks*, Structure, 19 (2011), pp. 881–889.

[223] R. C. Stevens, Y. M. Chook, C. Y. Cho, W. N. Lipscomb, and E. R. Kantrowitz, *Escherichia coli aspartate carbamoylatransferase: the probing of crystal structure analysis via site-specific mutagenesis*, Protein Engineering, 4 (1991), pp. 391–409.

[224] R. C. Stevens, J. E. Gouaux, and W. N. Lipscomb, *Structural consequences of effector binding to the T state of aspartate carbamoyltransferase: crystal structures of the unligated and ATP- and CTP-complexed enzymes at 2.6-A resolution.*, Biochemistry, 29 (1990), pp. 7691–7701.

[225] K. Stieglitz, B. Stec, D. P. Baker, and E. R. Kantrowitz, *Monitoring the transition from the T to the R state in E. coli aspartate transcarbamoylase by X-ray crystallography: Crystal structures of the E50A mutant enzyme in four distinct allosteric states*, Journal of Molecular Biology, 341 (2004), pp. 853–868.

[226] G. Strang and K. Aarikka, *Introduction to applied mathematics*, vol. 16, Wellesley-Cambridge Press Wellesley, MA, 1986.

[227] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan, *Evolutionarily conserved networks of residues mediate allosteric communication in proteins*, Nature Structural and Molecular Biology, 10 (2003), p. 59.

[228] S. M. Sullivan and T. Holyoak, *Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection*, Proceedings of the National Academy of Sciences, 105 (2008).

[229] A. Szabo and M. Karplus, *A mathematical model for structure-function relations in hemoglobin*, Journal of molecular biology, 72 (1972), pp. 163–197.

[230] H. M. Taylor and S. Karlin, *An introduction to stochastic modeling*, Academic press, 2014.

[231] M. F. Thorpe, *Continuous deformations in random networks*, Journal of Non-Crystalline Solids, 57 (1983), pp. 355–370.

[232] ——, *Comment on elastic network models and proteins.*, Physical biology, 4 (2007), pp. 60–63; discussion 64–65.

[233] M. M. Tirion, *Large Amplitude Elastic Motions in Proteins from a Single-Parameter , Atomic Analysis*, Physical Review Letters, 77 (1996), pp. 1905–1908.

[234] M. M. Tirion and D. Ben-Avraham, *Normal mode analysis of G-actin*, Journal of molecular biology, 230 (1993), pp. 186–195.

[235] C. J. Tsai and R. Nussinov, *A Unified View of "How Allostery Works"*, PLoS Computational Biology, 10 (2014).

[236] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, *GROMACS: Fast, flexible, and free*, Journal of Computational Chemistry, 26 (2005), pp. 1701–1718.

[237] B. F. Volkman, D. Lipson, D. E. Wemmer, and D. Kern, *Two-state allosteric behavior in a single-domain signaling protein.*, Science (New York, N.Y.), 291 (2001), pp. 2429–2433.

[238] L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Mart\'\inez, and V. S. Pande, *Building a more predictive protein force field: a systematic and reproducible route to AMBER-FB15*, The Journal of Physical Chemistry B, 121 (2017), pp. 4023–4039.

[239] Y. Wang and L. Makowski, *Fine structure of conformational ensembles in adenylate kinase*, Proteins: Structure, Function and Bioinformatics, 86 (2018), pp. 332–343.

[240] D. J. Watts and S. H. Strogatz, *Collective dynamics of âĂŸsmall-world'networks*, nature, 393 (1998), p. 440.

[241] G. Wei, W. Xi, R. Nussinov, and B. Ma, *Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? the Diverse Functional Roles of Conformational Ensembles in the Cell*, Chemical Reviews, 116 (2016), pp. 6516–6551.

[242] W. Whiteley, *Counting out to the flexibility of molecules.*, Physical biology, 2 (2005), pp. S116–S126.

[243] J. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1*, Journal of Molecular Biology, 285 (1999), pp. 1735–1747.

[244] P. E. Wright and H. J. Dyson, *Intrinsically Unstructured Proteins : Re-assessing the Protein Structure-Function Paradigm*, Journal of Molecular Biology, 293 (1999).

[245] L. Yan, R. Ravasio, C. Brito, and M. Wyart, *Architecture and Co-Evolution of Allosteric Materials*, Proceedings of the National Academy of Sciences, (2016), pp. 1–6.

[246] ——, *Principles for optimal cooperativity in allosteric materials*, Biophysical Journal, 114 (2018), pp. 2787–2798.

[247] E. W. Yu and D. E. Koshland, *Propagating conformational changes over long (and short) distances in proteins*, Proceedings of the National Academy of Sciences, 27 (2001), pp. 2–5.

[248] L. Yu, S. M. Reutzel-Edens, and C. A. Mitchell, *Crystallization and polymorphism of conformationally flexible molecules: Problems, patterns, and strategies*, Organic Process Research and Development, 4 (2000), pp. 396–402.

[249] R. Zwanzig, *From classical dynamics to continuous time random walks*, Journal of Statistical Physics, 30 (1983), pp. 255–262.

chemfig

Appendix A

Further details of graph construction

The weights of the edges in the protein graphs of Chapter 4 are determined by the interaction energy of the bond. We include four different types of interaction: covalent bonds, hydrogen bonds, electrostatic interactions and hydrophobic interactions.

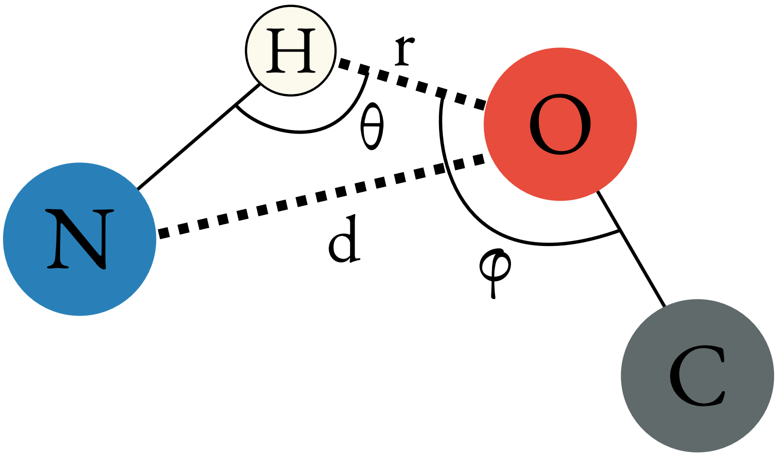
A.1 Covalent bonds

The actual presence of covalent bonds is determined by the software FIRST, based on simple distance cutoffs. Then standard chemical bond energies are used to weight the edges:

Bond	Energy (kJ/mol)	Bond	Energy (kJ/mol)	Bond	Energy (kJ/mol)
C-C	346	C=O	799	O-H	459
C=C	602	C-S	272	S-H	363
C-N	305	H-H	432	Se-H	276
C=N	615	C-H	411	P-O	335
C-P	264	N-H	386	P=O	544
C-O	358	P-H	322		

A.2 Hydrogen bonds

In the case of hydrogen bonds, both the position and the bond energies are calculated by FIRST, using the formula presented by Mayo *et al*[48]:



$$E_{\text{hydrogen bond}} = V_0 \left\{ 5 \left(\frac{R_0}{R} \right)^{10} - 6 \left(\frac{R_0}{R} \right)^{12} \right\} F(\vartheta, \phi, \psi) \tag{A.1}$$

with $V_0 = 8$ kcal/mol, $R_0 = 2.80$ Å as the equilibrium donor-acceptor distance, and R the actual distance.

Angles ϑ, ϕ and ψ are shown in Fig. A.2 and $F(\cdot)$ is a function of the three angles that depends on the hybridization of the donor - acceptor atoms.

A.3 Electrostatic interactions

Electrostatic interactions are included in the graph on the basis of the LINK entries in the protein structure's PDB file and the bond energies are calculated according to Coulomb's Law:

$$E_{\text{electrostatic}} = \frac{332}{\varepsilon} \frac{q_1 q_2}{r} \tag{A.2}$$

where $\varepsilon = 4$ is the dielectric constant, q_1 and q_2 are the charges on the atoms and r the distance between them. Atom charges are obtained from the OPLS-AA force field[108].

A.4 Hydrophobic interactions

C-C and C-S bonds may also have hydrophobic interactions between them: again FIRST identifies those pairs of atoms that are within a distance cutoff (2Å) but does not assign a specific energy. Instead the double well potential developed by Head-Gordon *et al*[137] is used.

Appendix B

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN)[65] is a clustering algorithm that groups together data points in space according to how closely spaced they are. Regions of densely packed points will be grouped together and in contrast to many clustering methods, DBSCAN is able to find non-linear groupings of the points, rather than simply drawing separating hyperplanes.

DBSCAN was performed on the data points in Fig. 5.7 to test whether there were two distinct clusters of points (here the absolute bond displacements) which may suggestive of certain bonds having a greater connection to the allosteric site. However, the two "bands" of points that are somewhat discernible by eye are not recovered by DBSCAN, instead an uninformative linear separation of the data points is seen.

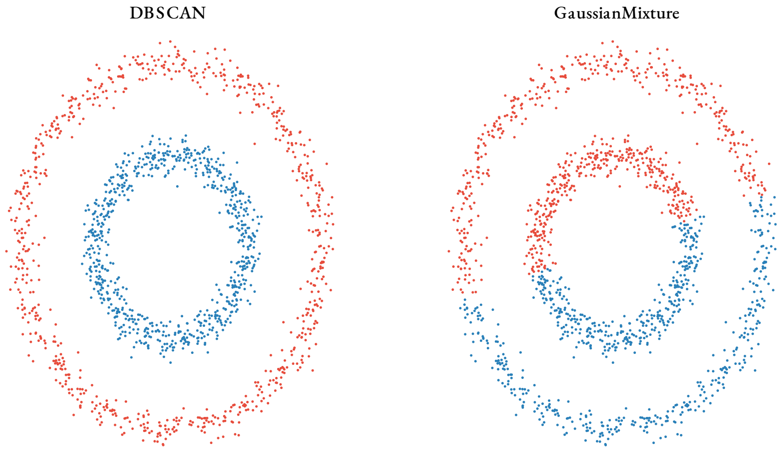
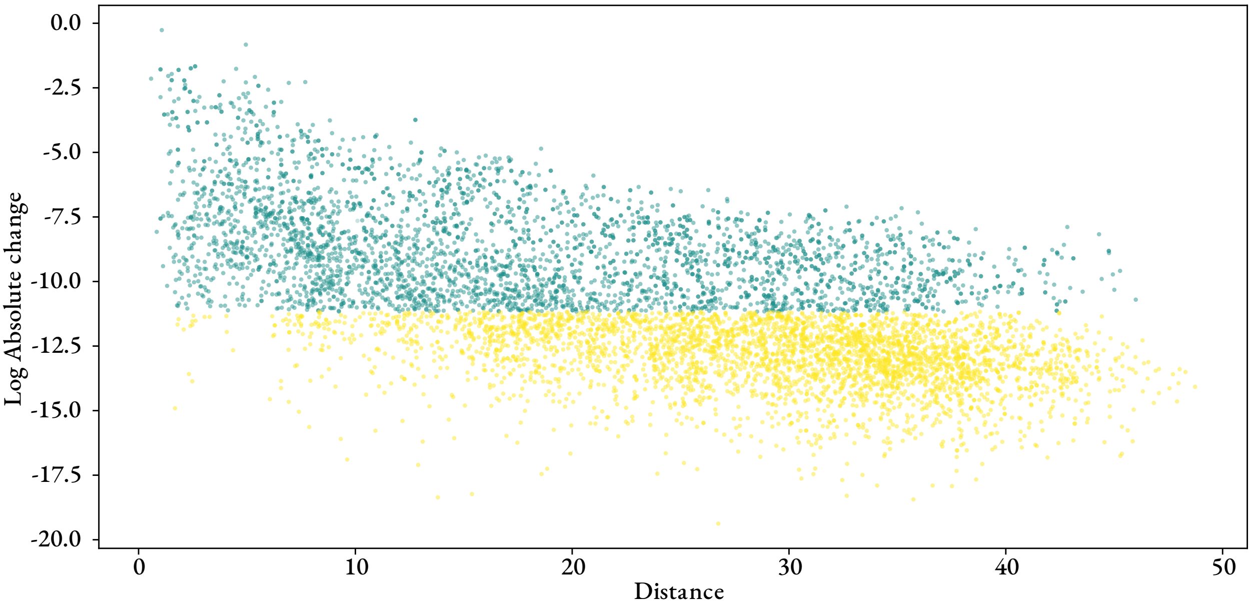


Figure B.1: On the left, DBSCAN is capable of clustering the points nonlinearly whilst many other standard methods, such as Gaussian Mixture Models on the right, are only able to linearly separate groups of points.


0.0
-2.5
-5.0
-7.5
-10.0
-12.5
-15.0
-17.5
-20.0
Log Absolute change
0
10
20
30
40
50
Distance

Appendix C

Quantile Regression

C.1 Generalising the median

Quantile regression is the generalisation of the idea of ranking 1-dimensional data. The most commonly used special case of quantile regression is the *median*, which is defined as the data point greater than (or equivalent less than) half of the total set of data points. The median for even numbers of data points is therefore not unique, though is usually assigned to be halfway between the point above and below the "midpoint". The value of the median (and quantile regression more generally) is therefore robust to extreme values, in contrast to the mean.

We can extend the idea of median using the following optimization problem for a set of points x:

$$\min_{\beta} \sum_{i=1}^{n} \varrho_p(x_i - \beta), \qquad p \in [0, 1]$$

$$\varrho_p(\xi) = \left|\xi(p - \mathbb{1}_{\xi<0})\right| \tag{C.1}$$

where we can now find any *quantile* for the data, with the median being the special case where $p = 0.5$. The indicator function $\mathbb{1}_{\xi<0}$ is equal to 1 when the term inside the function $\varrho_p(\cdot)$ is less than 0 and is 0 otherwise. The effect is to weight the penalty of having data points above and below the optimum point we wish to find β. So if we wish to find the 70$^{\text{th}}$ quantile, we would penalise points below by a weighting of 0.7 and those above by 0.3. Note that because the objective function (C.1) is non-differentiable, there is no analytical solution and instead linear programming is used to find an optimal solution. As an example:

Let Y be a discrete random variable that takes the values 1, 2, ..., 9 with equal probabilities. Find the value of Y so that p = 0.7.

$$L(\beta) = (p-1)\sum_{x_i<\beta}(x_i-\beta) \; + \; p\sum_{x_i>\beta}(x_i-\beta) \tag{C.2}$$

Trying different values:

$$L(6) = -0.3 \times (-5\ -4\ -3\ -2\ -1) + 0.7 \times (0\ +1\ +2\ +3) = 8.7$$

$$L(7) = -0.3 \times (-6\ -5\ -4\ -3\ -2\ -1) + 0.7 \times (0\ +1\ +2) = 8.4$$

$$L(8) = -0.3 \times (-7\ -6\ -5\ -4\ -3\ -2\ -1) + 0.7 \times (0\ +1) = 9.1$$

so we can see the value of our data that lies at the 70th quantile can be anywhere between 7 and 8.

C.2 Conditional quantile regression

In the same way the mean can be extended to the conditional mean, we can define *condition quantile regression* as:

$$\min_{\beta}\sum_{i=1}^{n}\varrho_p(y_i - Q(x_i,\beta)), \qquad p \in [0,1]$$

where $Q(x_i, \beta)$ is some assumed distribution, parameterised by β. In the special case where we suppose a linear relationship between the variables, we have:

$$\min_{\beta} \sum_{i=1}^{n} \varrho_p(y_i - (\beta_0 + \beta_1^T x_i)), \qquad p \in [0, 1]$$

Again, the objective function has no analytical solution but can be efficiently minimised via linear programming. We can now find a specific quantile for the data given the values of the variables x_i.