Thèse n° 7834

Mechanics and co-evolution of allosteric materials and proteins

Présentée le 2 octobre 2020

à la Faculté des sciences de base
Laboratoire de physique des systèmes complexes
Programme doctoral en physique

pour l'obtention du grade de Docteur ès Sciences

par

Riccardo RAVASIO

Acceptée sur proposition du jury

Prof. F. Mila, président du jury
Prof. M. Wyart, directeur de thèse
Prof. R. Monasson, rapporteur
Prof. T. McLeish, rapporteur
Prof. A.-F. Bitbol, rapporteuse

École
polytechnique
fédérale
de Lausanne

2020

A mamma e papà

Acknowledgements

I express my gratitude to Matthieu Wyart for his supervision during my doctoral studies in the PCSL (Physics of Complex Systems Laboratory) group. I deeply appreciate his passion for science, his intuitive and precise way of thinking, and his constant support and guidance during the time that I spent here. I had the chance to find a thriving environment where I grew as a scientist and as a person.

I would like to acknowledge Anne-Florence Bitbol, Tom McLeish and Rémi Monasson who accepted to be part of the jury of my viva and provided deep insights that contributed to improve the final version of this manuscript along with my understanding. I also thank Frédéric Mila for his role as president in the doctoral commission.

Little of what is discussed in this manuscript would have been possible without the collaboration that Matthieu created with Carolina Brito and Le Yan, and also without the many other people I will acknowledge in the following. I would like to thank Carolina and Le for their patient and essential teaching, for the time spent in Lausanne discussing about our projects, and for the joyful environment created in and out of the lab, among which some intense, noteworthy debugging sessions that turned out to be fun with you. I also thank Barbara Bravi who joined this collaboration and with whom I had the chance to share two years of great work, enriching discussions, comforted by hot tea. I thank you for your door that you keep always open.

Duccio Malinverni offered me the first coffee when I arrived at EPFL and has also been patient in explaining the world of proteins and inference methods with his extreme competence. Stefano Zamuner has been a constant source of inspiration throughout these years and also a crucial reference point for learning how to analyse protein data. Without you two, it would have taken much longer to learn these new topics — and I am pretty sure it would have been boring!

I want to thank Solange Flatt for the time spent together during her master thesis working on the analysis of protein data, which would have been surely less efficient and less fun otherwise — not to mention our coffee breaks at Arcadie.

I thank Eric DeGiuli who guided me in the realm of elasticity — and not only — and has been a model to follow for my first experience as teaching assistant that we shared at EPFL.

I acknowledge fruitful and inspiring discussions with Paolo De Los Rios, who has been always available to talk about science and life, and visitors Christophe Eloy, Rama Ranganathan, Alberto Rosso, Boris Shraiman, Martin Weigt. In particular, I thank Rama for sharing empirical data that we used in our research.

I thank all members of PCSL for these past four years: Eric, Matteo, Wencheng, Yuanjian, Solange, Mario, Stefano, Barbara, Marko, Tom, Elisabeth, Levent, Jonas and Leonardo. It has been incredible to share our scientific and personal lives in such a supportive environment with dinners, *pauses gourmandes*, activities in the snow and on trees and just discussions on campus that will be unforgettable. A special thanks goes to Corinne that with her positive energy and patience participated and made all our group activities possible, among all the other help she provided.

During the PhD I participated to several meetings of the Simons collaboration on Cracking the glass problems where I could get inspired by new science and share time with colleagues and friends, in particular, I would like to thank Ada, Benjamin, Camille, Elisabeth, Giulio, Ludovic, Marylou, Pierfrancesco, Stefano and Valentina.

I thank Aubry Jacquier for his technical assistance and support with any IT issue I had.

I am also indebited to the people who gave feedback on shorter or longer parts of this manuscript: Matthieu Wyart, Carolina Brito, Barbara Bravi, Abigail Plummer, Gabriele Berruto.

To conclude, I warmly thank *all* my *friends*; from THEOS, LMSC, LMGN, PBL, LSU, LUMES, FIMAP, SUNMIL, PCSL and sionmaville, *Mellon*, Gianluca, Ghil and my family without whom it would have been just gray.

Lausanne, 11th June 2020 Riccardo Ravasio

Abstract

The regulation of several processes inside and outside the cell depends on the action of a particular class of enzymes, called allosteric. In allosteric macromolecules, binding a ligand at one site affects the binding activity at a distal functional site, providing a reliable tool to regulate the corresponding function. The physical mechanisms underpinning allostery and its long-range communication are not yet fully understood, despite a great number of advances were made possible by significant works spanning 60 years, in between biology, bioinformatics and physics.

In physics terms, proteins can be viewed as amorphous materials that however underwent billions of years of evolution to be functional as observed today. The framework introduced in this dissertation allows to explore how the structural organisation of an allosteric system is constrained by the function that it has evolved to perform. It allows a classification of allosteric architectures and suggests a physical explanation behind the emergence of such long-range allosteric coupling. Furthermore, it is also apt to build in silico a large amount of allosteric architectures that share the same evolutionary history. The constraints imprinted by evolution on sequences that share a common ancestor motivate the exploration of inference methods that try to predict the fitness of a protein solely from the knowledge of sequences.

The strategy used to build this framework is to resort to a coarse-grained model of a protein, on the line of elastic network models resulted successful in the description of the large-scale dynamics of proteins. Ideas on how to pursue these research directions further are discussed throughout the chapters.

Firstly, we introduce an in-silico model for the evolution of allosteric behaviour in discrete lattices of harmonic springs. The in-silico evolution is performed for two different allosteric tasks: one optimising for the transmission of strain between the allosteric and active site, while the other maximising the cooperative binding energy between the two. To optimise the transmission of strain, the network develops a lever that amplifies the response at the active site. In such a way, our model proposes a novel allosteric architecture, potentially in use in proteins as well. Cooperative architectures show, among others, hinge and shear motions and rationalise the observation of a low energy mode that describes conformational changes in proteins. Indeed, to achieve proper function, the mode is predicted to get softer as the size of the system increases. This prediction is tested by collecting a database of 34 high resolution structures of allosteric proteins and is proven valid even when elastic nonlinearities are introduced.

Secondly, the sequences generated with the in-silico model serve to benchmark existing methods that infer co-evolutionary couplings between amino acids, proven to be successful in predicting local structural constraints, but with unclear performance in the presence of global allosteric constraints. These models do predict local features reflecting structure, but fail in the prediction of long-range functional dependencies and are not able to generate synthetic sequences that function as native ones. Thus, the exploration of new directions is needed.

Keywords: Allostery, protein function, mechanical networks, amorphous materials, co-evolution, inference methods, normal mode analysis, population shift, induced fit.

Sinossi

La regolazione di numerosi processi all'interno e all'esterno delle nostre cellule dipende dall'azione di una particolare classe di enzimi, chiamati allosterici. Il legame di un ligando a un sito di una macromolecola allosterica influenza l'attività di un altro sito ad esso distante, fornendo uno strumento per controllare in maniera precisa la funzione corrispondente. I meccanismi fisici che sottostanno all'allosteria, tra cui la comunicazione a lungo raggio tra i due siti, non sono ad oggi ancora pienamenti compresi, nonostante i numerosi progressi degli ultimi sessanta anni provenienti da campi diversi come biologia, bioinformatica e fisica.

In fisica, una proteina può essere vista come un solido amorfo che ha subito miliardi di anni di evoluzione per esibire la funzione che si osserva oggigiorno. Le idee introdotte in questa dissertazione permettono di esplorare la relazione tra l'organizzazione strutturale di un sistema allosterico e i vincoli evolutivi dovuti alla sua funzione: lo sviluppo della comunicazione a distanza può essere spiegato in termini fisici e ciò rende possibile una classificazione delle diverse architetture allosteriche. L'approccio usato si basa su modelli di reti elastiche che, nonostante le forti approssimazioni, riescono a riprodurre la dinamica soggiacente a cambiamenti conformazionali in alcune proteine.

Il modello qui introdotto esegue evoluzioni in silico di reti elastiche che esibiscono un dato comportamento allosterico. Le reti elastiche sono descritte da un reticolo di nodi connessi da molle, cioè da oscillatori armonici. Per esplorare la relazione tra architettura e funzione, l'evoluzione in silico avviene considerando due diverse funzioni. La prima ottimizza per la propagazione di deformazione elastica tra il sito allosterico e il sito attivo, mentre la seconda seleziona architetture che mostrano grande energia cooperativa di legame tra i due siti. Si osserva che la rete, per ottimizzare la propagazione della deformazione elastica, sviluppa una leva che amplifica la risposta dal sito allosterico al sito attivo, proponendo una nuova architettura allosterica, possibilmente in uso anche nelle proteine. Le architetture evolute per ottimizzare i legami cooperativi tra i siti sono invece caratterizzate da movimenti simili a quelli di taglio e cerniera, già osservati in diverse proteine. Queste architetture giustificano la presenza di un modo vibrazionale a bassa energia osservato che spesso descrive il cambiamento conformazionale nelle proteine. Infatti, la frequenza di questo modo deve decrescere al crescere della taglia del sistema per far sì che funzioni. Questa previsione è studiata raccogliendo una banca dati contenente la struttura di 34 proteine allosteriche ed è confermata anche quando delle nonlinearità sono introdotte nel sistema.

Inoltre, il modello è anche in grado di generare un numero di sequenze che condividono la stessa storia evolutiva. Si può così studiare il comportamento di metodi di inferenza che usano l'informazione evolutiva per tentare una predizione dell'energia di una proteina a partire solamente dalle sequenze di amminoacidi. Metodi di inferenza esistenti hanno mostrato comprovato successo nel predirre quali amminoacidi sono in contatto nella struttura tridimensionale, sebbene falliscano sia nel catturare relazioni funzionali tra amminoacidi distanti, quali possono essere quelle allosteriche, che nel generare sequenze sintetiche che funzionano come quelle native. Questi fallimenti motivano la ricerca di nuovi metodi che siano più adatti a trattare la complessità della funzione allosterica.

Parole chiave: Allosteria, proteine, reti meccaniche, materiali amorfi, coevoluzione, metodi di inferenza, analisi dei modi normali, population shift, induced fit.

Contents

Acknowledgements **v**

Abstract (English/Italiano) **vii**

Introduction **1**

Preamble 1

Background 1

The role of crystallography in understanding protein allostery 3

The role of sequence data in understanding protein allostery 13

Scope of the thesis and outline 29

1 Mechanics of allosteric materials and proteins **33**

1.1 Minimal models connecting structure, motion and function in allosteric proteins 33

1.1.1 Rationale for the presence of a soft mode in allosteric proteins 34

1.1.2 Prediction of residues involved in the allosteric response 36

1.1.3 A model of action at a distance without structural change 36

1.1.4 Discussion 37

1.1.5 Outline 38

1.2 A model for evolution of action at a distance in proteins 39

1.2.1 Definition of the model 39

1.3 Geometric function 42

1.3.1 Evolution 42

1.3.2 Thermodynamics 43

1.3.3 Architecture: trumpet design 44

1.3.4 Isostatic materials act as levers 46

1.3.5 Thermal fluctuations 47

1.3.6 A tool to pick fruits 48

1.3.7 Conclusions 49

1.4 Cooperative function 49

1.4.1 Evolution 51

1.4.2 Architectures: shear, hinge, twist and many others 52

1.4.3 Specificity of geometric and cooperative designs 57

1.4.4 Classification 58

1.4.5 Optimal cooperative binding 61

- 1.4.6 Fluctuations-driven cooperativity 64
- 1.4.7 Conclusions 65
- 1.5 Effect of non-linearity: a mechanical view on induced fit and population shift . 65
  - 1.5.1 Induced fit scenario 66
  - 1.5.2 Population shift scenario 68
  - 1.5.3 Remark on the hypothesis of the theory 72
  - 1.5.4 Mechanical model of population shift allostery 72
  - 1.5.5 Conclusions 74

2 Empirical results **77**
- 2.1 Describing deformation: the strain tensor 78
- 2.2 The allosteric response 79
- 2.3 Amplification of allosteric response in a PDZ domain 80
- 2.4 Testing optimal cooperativity in 34 allosteric proteins 81
  - 2.4.1 Principle 1: an extended low-energy mode well describes the allosteric response 83
  - 2.4.2 Principle 2: the allosteric response consists in extended regions of large shear 84
  - 2.4.3 Principle 3: scaling of the stiffness of the allosteric response 85
  - 2.4.4 Summarising outcomes of the empirical testing 86
- 2.5 Conclusions 86

3 Benchmark of inference methods for the prediction of allosteric fitness **89**
- 3.1 Mutation costs and epistasis in the in-silico model 90
  - 3.1.1 Characterisation of single mutation costs 92
  - 3.1.2 Mechanical characterisation of epistasis 93
- 3.2 Overview of the performance of Direct Coupling Analysis 95
  - 3.2.1 Prediction of mutation costs 97
  - 3.2.2 Generative power 97
  - 3.2.3 Prediction of epistasis 100
- 3.3 A possible answer to why DCA fails in the prediction of long-range couplings . 102
  - 3.3.1 Empirical data on a PDZ domain support the failure of DCA in capturing epistasis 104
- 3.4 Performance of Statistical Coupling Analysis 106
  - 3.4.1 Geometric sequences: local structure 106
  - 3.4.2 Cooperative sequences: long-range epistasis 107
- 3.5 Conclusions 109
  - 3.5.1 Summary of results 109
  - 3.5.2 Further directions 110

Conclusions **113**
- Mechanics: classification of allosteric architectures 113
  - Towards a conclusive empirical validation 114

Inference: benchmark on allosteric dependencies 115
An alternative route for inferring fitness 116

List of figures **121**

Acronyms **123**

A Mechanical networks **125**
A.1 Construction of the embedding lattice 125
A.2 Linear response of elastic networks 126
A.2.1 Stiffness matrix 126
A.2.2 Linear response to an imposed displacement 127
A.2.3 Measure of thermally-induced motion 128
A.3 Trumpet architecture in networks with $z < 4$ 129
A.4 Cooperative energy of two dipoles in a continuous elastic medium 130

B Strain analysis on empirical data **135**
B.1 Computing the local strain tensor in a protein 135

C Co-evolutionary analysis **137**
C.1 Mechanical interpretation of mutation costs and epistasis 137
C.2 Prediction of epistasis 140
C.3 Direct Coupling Analysis: inference procedure 142
C.4 Mutation costs and generative performance in the inferred Ising model 144
C.4.1 Comparison with conservation 145
C.5 Simple model illustrating the failure of DCA 147
C.5.1 Feedforward neural network 148

Bibliography **171**

Curriculum Vitae **173**

Introduction

Preamble

This thesis deals with different fields: physics and biology. Like before the advent of thermodynamics there was no clarity about the functioning of engines, in biology there is currently no understanding of the general principles that rule evolution and function of biological systems. There is a tendency of studying the microscopic details without connecting to their defining principles. In this dissertation, we provide theoretical results on the mechanics of an ubiquitous kind of protein regulation, allostery, that involves communication between distant regions in the protein structure. Instead of focusing on the study of an individual protein where details of its architecture might be distracting, we introduce an in-silico evolution scheme of such allosteric behaviour in discrete elastic materials. The model is an extremely simplified description of what is a protein. Yet, it still shows rich features. It provides means to classify the mechanics of allosteric architectures, to assess the relationship between structure and function and to explore the sequence-to-function pardigm and its inference. Hence, despite its simplicity, it proves useful in proposing principles for allosteric communication that can be tested experimentally. Behold and beware.

Background

In the complex and crowded environment inside and outside cells a multitude of events takes place, leading to control the functioning of our organism. Inside the cell, *metabolic pathways* constitute the series of biochemical reactions resulting in the synthesis of molecules, from amino acids, the building blocks of proteins, to Adenosine TriPhosphate (ATP), the energy source of most processes in living cells. Outside, *signal-transduction pathways* create a communication between the surrounding environment and the cell, and ultimately result in cellular response, e.g. allowing the cell to keep growing and avoid imminent death. In healthy organisms, these processes do not happen by chance, but are regulated thanks to some specific macromolecules, enzymes, among which *allosteric* enzymes are responsible for the control of most. In allosteric macromolecules, binding a ligand at one site affects the binding activity at a distal functional site, allowing the regulation of the corresponding function. The world of allostery is largely diverse: from enzymes, Aspartate TransCarbamoylase (ATCase) responsible to start off the biosynthesis of pyramidines — fundamental units of our genetic code — to proteins like haemoglobin, efficiently transporting oxygen in our blood.

Pedestrian introduction to proteins

The focus in this dissertation will be on protein allostery. Proteins are the protagonists in cells [1]: they constitute most of the cell's dry mass and they are both the building blocks and the regulators of almost all cell's functions. Moreover, they display the most complex structures and sophisticated functions among molecules, which have been fine-tuned during billions of years of evolution. Yet, this evolutionary process not only contributed in a fine-tuning of their functions, but also with a surprising versatility of proteins to adapt to different environments.

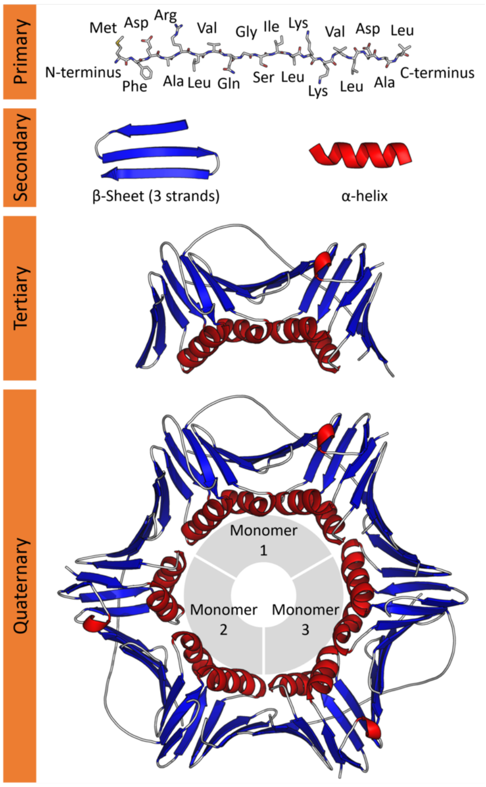

Figure 1 – Figure made by Thomas Shafee under the creative common license. This figure summarises the hierarchical structure of proteins into different level of characterisation, from its sequence to the way different proteins structurally form a protein complex. The protein is the proliferating cell nuclear antigen with Protein Data Bank (PDB) code 1AXC.

Each protein has its own sequence of amino acids — also called residues, which can be of 20 types with different chemical properties. The link between one amino acid in the chain to the next is guaranteed through a covalent bond, the peptide bond. In the core of the protein there is a repeating sequences of atoms, the backbone, where the portions of amino acids not involved in peptide bonds are attached, forming side chains that confer the special properties to proteins, as shown in the top panel of Fig. 1. Indeed, each side chain obeys to the chemical properties of the corresponding amino acid: either hydrophobic or positively charged, and so on. The typical length of the chain is 300 amino acids for bacteria and 400 for eukaryotes [2].

At this point the protein is just a polypeptide chain. How does it fold? The three-dimensional structure of a protein is determined by several interactions. First, steric constraints limit the three-dimensional arrangement of atoms. Nonetheless the protein chain being flexible, it has an extremely large number of ways to fold[1], constrained by different weak noncovalent bonds — hydrogen bonds,

---

[1] Interestingly, the Levinthal's paradox states that finding the native fold of a sequence composed of 100 amino

electrostatic and Van der Waals attractions — that form between two different parts of the chain.

Finally, the organisation of amino acids closely depends also on their degree of hydrophobicity: hydrophobic amino acids will tend to cluster in the interior of the protein, while the others at the exterior, in contact with water molecules of the cell. As a result of these interactions, the protein assumes a given conformation, typically found to be the one that minimises its free energy. Overall, the resulting fold is stable, even if upon changes in the environment it can modify its structure, which is fundamental to protein function.

From this discussion it emerges that a protein has a hierarchical structure organised in different levels of characterisation, as summarised in Fig. 1. The amino acid sequence is defined as the primary structure. The three-dimensional structures have commonalities in the way they fold locally. Two common patterns are found: α-helix and β-sheet, constituing the secondary structure of the protein. The global fold is the tertiary structure and the way proteins organise in space to form a protein complex is called quaternary structure.

Even if the folded conformation is stable, proteins are not rigid agglomerates of matter. They show a level of design and engineering as they present architectures that allow for mechanical motions under chemical events, like a ligand binding at a site. The coupling between mechanics and chemistry is responsible for the extraordinary spectrum of functionalities in proteins that underlie the dynamic processes happening in cells. Allosteric proteins fall in even a more special class given their ability to regulate function at a distal site from where the ligand has bound.

General questions arise from this pedestrian introduction to what is a protein. What can be learnt from the long evolution that proteins have undergone? Do sequences of the same kind of protein that has evolved under different conditions contain a trace of the constraints imposed by the mantaining of its structure and function? Is it possible to classify allosteric proteins into groups according to mechanical properties and allosteric tasks? Are the chosen architectures optimal to perform the corresponding allosteric function?

The role of crystallography in understanding protein allostery

In 1904, Christian Bohr already noticed that one molecule — carbon dioxide — affects the binding affinity of another molecule — oxygen — to a protein — haemoglobin [5]. This phenomenon was known as the Bohr effect and it was studied as cooperative binding of ligands to distinct protein sites leading to different equations describing such effects [6–9]. The term allostery was introduced almost 60 years later when Monod and Jacob [10] described several works where *the inhibitor is not a steric analog of the substrate*, as it is the

---

acids by exploring all of these conformations would take 10^{75} years at the speed of light [3]. However, as shown by Anfinsen, folding in nature is guided by the principle of free energy minimisation and thus takes only few milliseconds for a protein to fold in its native conformation in a funnel-like landscape [4].

case in the experiments by Changeux on the L-threonine deaminase [11, 12]. Indeed, its etimology combines *allos* for *other* and *steric* for *arrangement of atoms in space*. At the same time, a phenomenological model — the Monod-Wyman-Changeux (MWC) model — was published, predicting the thermodynamics and kinetics of the allosteric transition discovered in macromolecules like haemoglobin, where oxygen molecules bind cooperatively. This transition involves a conformational change between two well-defined structural states — the deoxygenated and the oxygenated state, which has higher affinity towards oxygen binding — that is so evident that Perutz noticed it decades before he could solve the X-ray structure [17]. Air could enter a sample of haemoglobin crystals which consequently started shattering from the point where air leaked in. This was a hint for a structural change upon oxygen binding that was large enough to break the bonds responsible for keeping the proteins in a crystal. The presence of an action at a distance was surely not assessed in the absence of structural information. Later on, the crystal structures of haemoglobin in the two states were solved, giving acess to the allosteric response upon oxygen binding [18], see Fig. 2. The clarification of heamoglobin structure made the problem of allosteric regulation even more challenging since it revealed that the oxygen binding sites are all distant one from the other in the molecule. How is it possible to couple energetically two distant sites? The answer to this question is still debated and it is one of the motivation of the work presented in this dissertation.

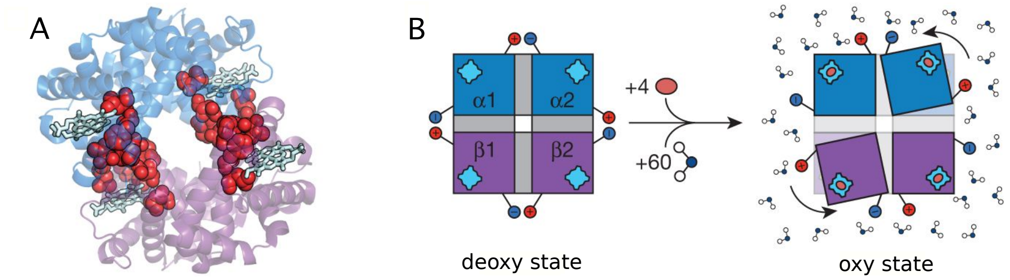

Figure 2 – Figure adapted from [13]. (A) Ribbon diagram representation of tetrameric haemoglobin rendered with the software Pymol [14]. A proposed pathway responsible for the cooperative transition from the deoxygenated and the oxygenated state is shown as red spheres. The pathway is identified by considering residues that are related in their evolutionary history, see the next section for a precise definition and discussion. The haem groups are represented as light blue sticks. (B) Allosteric transition of tetrameric haemoglobin, as proposed by Perutz model based on structural changes [15, 16]. Tetrameric haemoglobin in the deoxygenated state is depicted on the left with the two alpha-subunits (blue) and the two beta-subunits (purple) each with their own haem group (light blue). Salt bridges, depicted as the red positive and blue negative charges, hold the molecule in the deoxygenated conformation. These salt bridges are released upon binding of oxygen (orange oval) in the transition to the oxygenated conformation (on the right) accompanied by a 15° turn of the subunits relative to each another. In physiological conditions, also 60 additional water molecules contribute to the allosteric transition, preferentially binding the oxygenated state.

Phenomenological models of allostery

The availability of X-ray data of proteins in the two conformations, before and after the ligand binds at the allosteric site, opened the way to the study of allostery. Indeed, already the discovery of associated structural and functional changes in haemoglobin upong oxygen binding and in other systems as well [19] triggered the development of theoretical works implementing the observed features of the allosteric transition. This effort resulted in two phenomenologicals models, the Monod-Wyman-Changeux (MWC or later known as population shift) [20] and Koshland-Némethy-Filmer (KNF) or induced fit [21] model, both well describing the cooperative thermodynamics and kinetics of oxygen binding in haemoglobin [22] and other proteins [23].

The Monod-Wyman-Changeux model

For simplicity, two states of the macromolecule with different binding affinities, "Inactive" (In) and "Active" (Ac), are present — even in the absence of ligands — in the energy landscape of the MWC model, whose relative stability depends upon ligand binding. Supposing two binding sites, the macromolecule can be in four binding configurations, result of the binding combinations of the effector and the other ligand, see Fig. 3, where the effector is generally a small molecule that by binding preferentially to the allosteric site of the protein it regulates its activity.

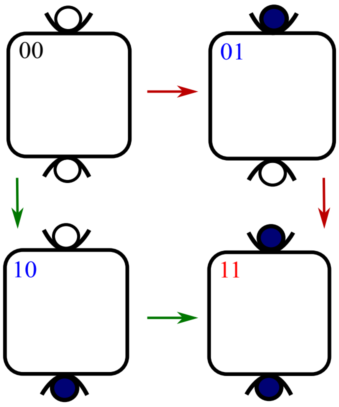

Figure 3 – Sketch of the four possible binding configurations of the network with two binding sites, xy, the allosteric site (x) and the active site (y). The values of $x, y = 0, 1$ specify the occupancy of the two binding sites. If no ligands are bound $x = 0$ and $y = 0$, if a ligand is bound at the allosteric site $x = 1$ and $y = 0$, and so on. Two paths are possible to get to the final configuration 11. A good allosteric function that favours cooperation between the two sites is $\mathscr{F}_{MWC} = (F_{01} - F_{00}) - (F_{11} - F_{10})$. Indeed, a macromolecule that obeys $\mathscr{F}_{MWC}$ the path $00 \rightarrow 10 \rightarrow 11$ is favoured since the energy difference $(F_{11} - F_{10})$ needs to be small, while the other path becomes very costly, since $(F_{01} - F_{00})$ gets at the same time larger and larger. Note that the two paths are not equivalent once the position of the allosteric site is chosen.

Let us consider the paradigmatic allosteric scenario where the effector binds at the allosteric site and elicits a change in binding activity at the active site, where the other ligand can now bind with a higher affinity. In the MWC model, the binding of the effector at the allosteric site shifts the equilibrium of the macromolecule from the inactive to the active state where the affinity of binding the other ligand is higher. Hence, it provides a simple mechanistic description of the cooperative allosteric transition.

	effector unbound	effector bound
inactive	$F_{00} = F_{In}$	$F_{10} = F_{In} - k_B T \ln\left(\frac{c}{K^{In}}\right)$
active	$F_{01} = F_{Ac}$	$F_{11} = F_{Ac} - k_B T \ln\left(\frac{c}{K^{Ac}}\right)$

Table 1 – Statistical mechanics of the MWC model for a macromolecule with two binding sites. The concentration of effectors is denoted with c and the dissociation constants of the inactive and active states with $K^{Ac} < K^{In}$, the inverse of the dissociation constant being proportional to the binding affinity. It is defined as $K_{In} = c \exp\left(-\beta F_{\text{bind,In}}\right)$, where $F_{\text{bind,In}}$ is the amount of energy gained after binding one ligand, $F_{\text{bind,In}} = F_{In} - F_{10}$.

The free energy of the macromolecule, F, can be written down for each configuration, resulting in a statistical mechanics description of the MWC model, see Tab. 1, which is useful since it allows to compute the probability for the macromolecule to be in the active state [17]. The cooperative behaviour between the allosteric and active site can be translated in a functional form describing cooperative fitness. Indeed, cooperative binding occurs when binding at one site lowers the energy cost of binding at the other site, and it is described by $\mathscr{F}_{MWC} = F_{10} + F_{01} - F_{00} - F_{11}$. Given the free energies in Tab. 1, $\mathscr{F}_{MWC}$ reads

$$\begin{aligned}
\mathscr{F}_{MWC} &= F_{10} + F_{01} - F_{11} - F_{00} && (1)\\
&= F_{In} - k_B T \ln\left(\frac{c}{K^{In}}\right) + F_{Ac} - \left[F_{Ac} - k_B T \ln\left(\frac{c}{K^{Ac}}\right)\right] - F_{In} && (2)\\
&= k_B T \ln\left(\frac{K_{In}}{K_{Ac}}\right) > 0 \quad K_{Ac} < K_{In} && (3)\\
&= F_{\text{bind,In}} - F_{\text{bind,Ac}}\,, && (4)
\end{aligned}$$

indeed describing cooperativity ($\mathscr{F}_{MWC} > 0$ for $K_{Ac} < K_{In}$) — note that if $K_{Ac} = K_{In}$ there is no cooperativity in the MWC model.

Generalisation of the MWC model

The models of allostery here discussed were specifically built thinking of the structural transitions like the one of haemoglobin upon oxygen binding as reported in Fig. 2. Hence, they depend on many details and in this context are not presented as in the original form. Thanks to the increase of experimental works exploring various allosteric systems, the view on allostery changed over time [13,24]. Indeed, allosteric behaviour has been identified in macromolecules composed of a sole subunit — monomers [25,26], without a significant structural change of the backbone in the case of nearby allosteric and active sites [27–29] and in proteins not classified to be allosteric [30–33], that display allostery when few point mutation are engineered or a ligand with a stronger affinity is introduced. These results contributed in a more general view of allostery as an ensemble of states [13]. The macromolecule is seen as having several possible states separated by energy barriers, changing the thermodynamic view of two pre-existing structures — inactive and active — to a dynamic free energy landscape that can be reshaped upon physiological conditions. The allosteric behaviour results from a weighted-average over all states that the macromolecule can be in and cannot thus be obtained from a single high-resolution structure. Indeed, functionally relevant conformational changes are often achieved through a shift in the equilibrium populations of the states [13, 34, 35] as can be measured by using NMR instead of X-ray data [36]. From this view comes the term *population shift*, used also in a more modern view of the MWC model.

The induced fit model

Few years after the works by Monod, Wyman and Changeux, the induced fit model was introduced [21], expanding the seminal ideas already present in [8]. This model still well describes cooperative thermodynamics and kinetics in haemoglobin even though differing in nature from the MWC model [22]. Moreover, the induced fit model also predicts that the binding of the first ligand can decrease the affinity to binding of subsequent ligands, in a negative cooperative manner observed in several proteins [37], feature that has been later

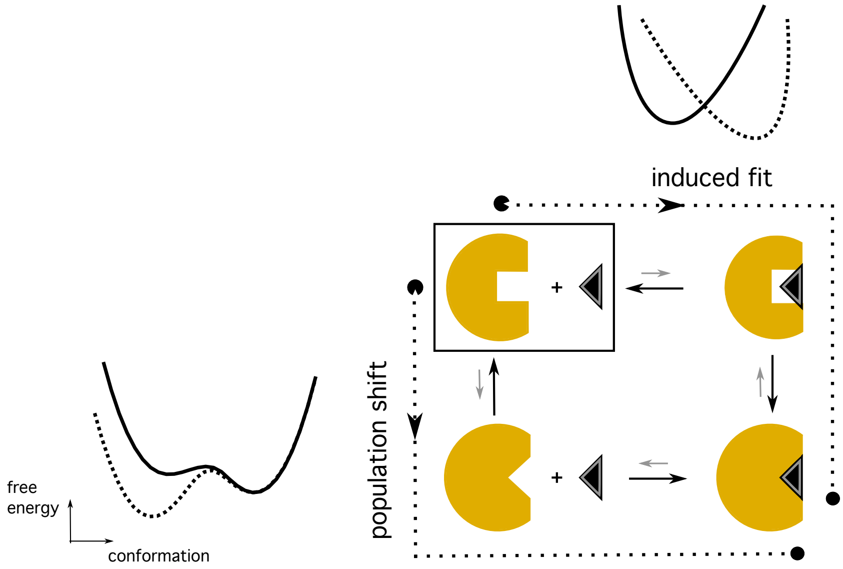

Figure 4 – The idea behind the population shift and induced fit scenario is sketched. The protein starts in the conformation with a binding pocket shape that does not match the ligand present in the solution — see illustration in the box. The direction taken in the two scenarios is indicated via dashed arrows. The population shift scenario predicts that the protein changes conformational state — from the inactive to the active state — so that the shape of the ligand is accomodated and can therein bind. The resulting energy landscape reproduces the two states. In the induced fit scenario, the protein also starts in the configuration with the wrong shape, but it is the ligand that by binding induces a change in the shape until it is matched, without another conformational state being populated. Hence, only one state is present in the energy landscape of the induced fit scenario.

introduced in an extended MWC model [38].

The MWC and induced fit models are based on two different scenarios for which a protein binding pocket changes its shape to let the ligand dock. As sketched in Fig. 4, the protein with the wrong configuration for binding the ligand has two ways of adapting its shape; either, driven by physiological conditions, it switches to a conformational state with the right geometry — population shift scenario — or the binding of the ligand makes the protein arrange via sequential events so that the shape is matched without switching from a state to another — induced fit scenario. Hence, the MWC model is characterised by a free energy landscape composed of two states separated by a barrier and the induced fit model just by one state, whose free energy changes during the allosteric transition, as sketched in Fig. 4 where free energy is viewed as a function of conformation. The two scenarios can be used by the same enzyme depending on ligand concentration and on the difference in free energy between the inactive and active state so that a continuous switching between the two is possible [39].

Hence, in the induced fit model, there is no concept of a shift between two states: ligand binding at the allosteric site of the macromolecule induces an adaptation of the — supposed flexible — active site so that the effector can bind. In an energy landscape picture, only one state is present and it changes in energy when the effector is bound.

Limitations of phenomenological models

Although the MWC and induced fit models presumably apply to various proteins, they do not specify which designs allow for efficient action at a distance — i.e. do not provide an explanation on how the protein is able to transmit allosterically the effect of ligand binding to the active site [13, 40–42] — and how robust these designs are to mutations, as discussed in this seminal article by Hopfield [43]. How is it possible to couple energetically two distant sites? The purely structural-based model of cooperative allostery by Perutz [15, 16], illustrated in Fig. 2, was the first to give insight on how the three-dimensional structure facilitates the allosteric communication between the sites in haemoglobin, suggesting that allostery can be understood via structural changes by inspecting the high-resolution structure. As discussed, this thermodynamic approach is successful, but is not complete without a mechanical picture of allostery.

Classification of allosteric motions

The increase in number of X-ray structures allowed to study the possible local motions that proteins exhibit when going from one conformation to the other upon ligand binding, defining the allosteric response as the displacement occurring between the two states. Small and large-scale protein motions could then be systematically observed. Allosteric proteins were found to be special for displaying on average more than double the percent of motion with respect to non allosteric proteins upon ligand binding [44], while noticing that in both cases the motion is more likely to occurr in weakly constrained regions, as loops or outer parts of the protein.

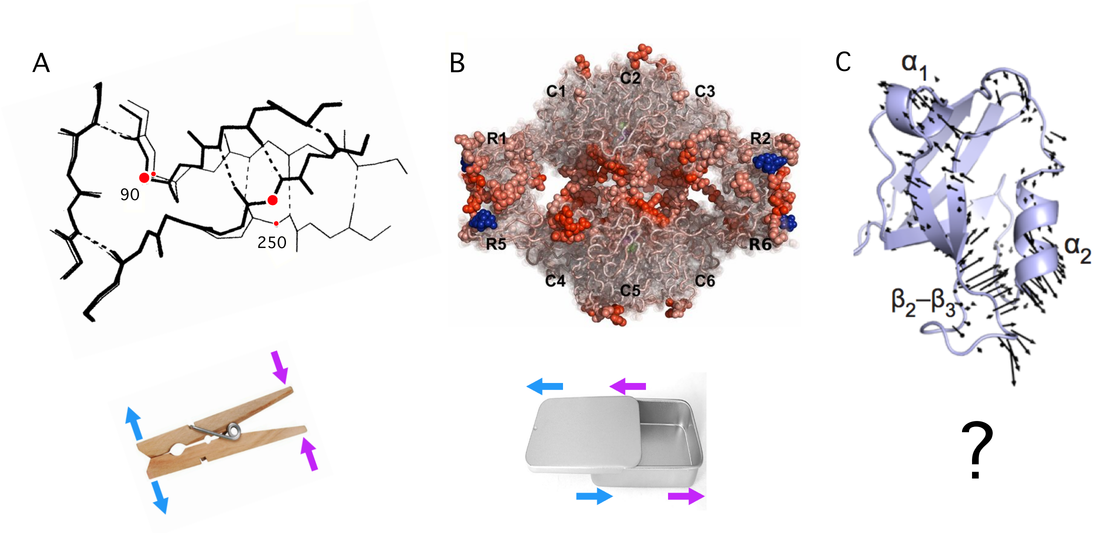

Figure 5 – Figure A is adapted from [45], figure B is adapted from [46] and figure C is adapted from [47]. (A) An example of the protein mechanics study of [45] from an analysis of high-resolution structures is shown. Lactoferrin displays a coupling of two simple hinges in a β-sheet. The rotation points (amino acids 90 and 250) are indicated by red circles. (B) The biological assembly of E. Coli ATCase is shown. The amount of shear per residue is displayed as color code (in log scale) increasing from grey to red — also only the most sheared residues are opaque. It seems that shear is transmitted from the allosteric sites (blue) to the buried active sites (green) via sliding of the unsheared regions of the catalytic domains. Indeed, a shear plane appears between the two halves of the protein complex. (C) Conserved motions of backbone atoms due to ligand binding in high-resolution structures of 11 diverse homologues of the PDZ family. The two active sites are located around the $\alpha 1$ helix and $\beta 2 - \beta 3$ loop, while the ligand binds in the pocket between the $\beta 2$ sheet and $\alpha 2$ helix.

Hinge design. An attempt of classification found two main types of motions, hinge and shear [45]. Fig. 5A illustrates an example of hinge motion, which is also the one used by haemoglobin as is shown in Fig. 2. These designs implement simple architectures that allow to propagate elastic information at long distance and are also common in daily-life objects as sketched in the bottom panels of Fig. 5. Indeed, a small displacement of the handles of a scissors makes the blades move at the other side. This is possible thanks to the presence of the hinge mechanism that can be easily excited. The very same mechanism is exploited by proteins as well.

Shear design. The shear design displays a weakly connected region between two rigid blocks, as exemplified by the mint box in Fig. 5B that opens by sliding. The allosteric response can be characterised further by looking at how much and which kind of deformation is associated to it. The strain tensor quantifies the amount of deformation occurring at each amino acid in the protein where shear and bulk components are disentangled. A significant amount of shear deformation can be measured residue by residue, showing that the allosteric response of a set of proteins studied in [46] is associated with large shear-like deformation. A shearing plane is also identified, see Fig. 5B.

Currently unclassified. Proteins whose allosteric architecture does not conform to these two groups are also found. An example is the PDZ domain, which is a structural domain composed of 80-90 amino acids that is commonly present in signaling proteins. Its response upon ligand binding is shown in Fig. 5C. Hence, new tools are needed to make a catalogue of allosteric responses and associated task.

Surely, X-ray structures have also their limitations. Given the importance of enviromental conditions in the allosteric transition, some proteins might not show any relevant conformational change when looked in their crystalline state even though they would undergo a conformational change when in solution with physiological conditions in a dynamical environment [48, 49]. Hence, when conformational change is not detected it does not mean that it is not present.

Relating motion to structure: elastic network models

As the classification presented in the previous section suggests, see Fig. 1, many allosteric proteins are composed of semi-rigid domains that can move relative to each others via softer flexible regions. Hence, this observation can justify substituting atomistic interactions for a more coarse-grained mechanical model of allosteric transition. An example is elastic networks where springs connect the amino acids in a protein with the assumption that its dynamics is fully determined by the topology of contacts. Such model relates structure to motion by giving direct access to the dynamics of the protein with collective motions at low-frequency. These motions are identified via normal modes that describe the intrinsic structural rearrangements of the considered protein as harmonic vibrational oscillations around its mean minimum energy structure, which should be close to its native state.

Definition of the model

An elastic network of springs can be directly built given the resolved positions of amino acids in the three-dimensional structure of the protein, which is supposed to be in the native minimum of the protein free energy landscape. This representation of a protein as masses and springs allows to access the mechanical features of the protein long-time and large-scale dynamics from structural data, as discussed in the following.

In the anisotropic network model [50], the harmonic potential between the residues i and j with stiffness constant k simply reads

$$V_{ij} = \frac{1}{2}k(l_{ij} - l_{ij}^0)^2 = \frac{1}{2}k\left\{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} - l_{ij}^0\right\}^2 \tag{5}$$

where l_{ij}^0 is the equilibrium distance between i and j, as illustrated in Fig. 6 where the beads represent protein residues with positions (x, y, z) in the three-dimensional crystalline structure. This model is distinguished from the previous gaussian network model [51] by taking into

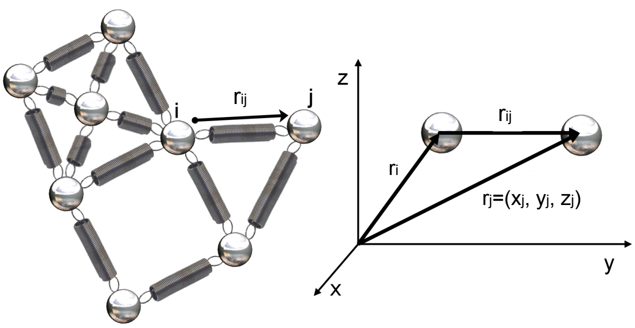

Figure 6 – Schema of an elastic network model where the metallic beads represent residues among which harmonic springs are inserted when their distance is below a chosen cutoff radius. The position vectors r_i and $r_j = (x_j, y_j, z_j)$ are highlighted, and $l_{ij} = ||r_{ij}|| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$. Public domain figure adapted from Wikipedia.

account the three-dimensionality of motion instead of just its radial component describing the expansion or compression of the spring.

Normal modes

The idea of supposing harmonic interactions between the residues of a protein to perform normal mode analysis was introduced by Tirion [52] and was consequently discussed in several works, notably by Hinsen [53, 54]. In Ref. [52] Tirion showed that the lowest energy modes obtained by supposing harmonic interactions where almost identical to the ones found with detailed force fields, thus supporting the use of elastic network models.

The potential in Eq. 5 acts for all residues at a distance within a chosen cutoff radius, R_c, whose value needs to be chosen to avoid a definition of neighborhood containing very few residues. Indeed, a poorly connected residue might create instabilities leading to more than the six zero modes associated to translations and rotations, i.e. the directions along which the system can move without any energy cost.

The Hessian matrix is obtained from Eq. 5 by computing its second derivatives and then it is evaluated in the equilibrium configuration, l_{ij}^0. The normal modes — and associated frequencies — are directly calculated via diagonalisation of the Hessian matrix, which reads

$$H_{ij} = \frac{-k}{l_{ij}} \begin{bmatrix} (x_j - x_i)^2 & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)^2 & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)^2 \end{bmatrix} . \tag{6}$$

Connection to allostery

Given the frequencies ω and the normal modes $\mathbf{R}_\omega$, the flexibility of the structure in this harmonic approximation can be directly obtained from the definition of B-factor [55] — also known as Debye-Waller factor in condensed-matter physics

$$B(i) = \sum_{\omega>0} \frac{1}{\omega^2} \mathbf{R}_\omega(i) \cdot \mathbf{R}_\omega(i)\,, \tag{7}$$

where, in this instance, the index i refers to a given residue. Elastic Network Models (ENM) predict accurately experimental temperature factors — of some proteins — with the B-factor computed in the harmonic approximation via Eq. 7. In a good quality model, the experimental factors indeed measure the flexibility of each component by the degree of uncertainty of observing that given atom position.

Moreover, the empirical allosteric response is found to be well described by few low energy, global normal modes — often mainly by just one of them [56–58]. Thus, the structural organisation of the protein in the unliganded state displays an intrinsic flexibility — in particular described by the low energy vibrational modes — that is exploited to perform the functional conformational change upon ligand binding. This result supports that in at least some proteins elasticity — possibly non-linear [59] — is an appropriate language to describe allostery[2]. However, ENMs do not provide any insight on the underlying protein structure that allows this soft mode to appear or on how the mode has to be designed for proper allosteric communication, thus other frameworks are needed, as argued in Chapter 1.

The choice of harmonic springs all of equal stiffness and defined by an arbitrary cutoff radius as proxy for native interactions is a major oversimplification [60]. Other works built refined versions of ENMs — yet still based on the same concept — aiming to describe some of the basic features of proteins not taken into account in anisotropic network models — e.g. side-chain interactions [61], and the presence of more rigid regions, as the backbone [62]. Such refined versions improved the prediction of motion and dynamics from structural data with respect to the usual ENM — the anisotropic network model, which still gives surprisingly good results.

Summary and open questions

Communication between distant sites in proteins that results in a change in activity has been posited since the works of Bohr who noticed that carbon dioxide affects the binding affinity of oxygen to haemoglobin [5]. Monod and Jacob [10] coined the word allostery and a phenomenological model, the Monod-Wyman-Changeux model, was later discussed, which was then followed by the induced fit model. Both models successfully describe allosteric cooperative effects of oxygen binding in haemoglobin. In the meanwhile, Perutz succeeded in crystallising haemoglobin in both the unliganded and liganded state, opening the way to understanding the role of structure in the allosteric communication. As discussed, the increasing availability of high-resolution structures made possible to attempt a classification of motions [44, 45], to access dynamical features via elasticity [50] and to characterise the

---

[2] There are however cases of intrinsically disordered proteins that may be considered more as liquids than solids, and for which elasticity is thus inappropriate.

mechanical properties of the allosteric response [46].

Some unanswered questions emerge from the (partial) description of the contributions to understand allosteric behaviour.

- **Nature of the information.** As discussed by Hopfield in [43], the phenomenological models introduced above do not clarify which is the nature of the information that propagates in the protein. What is the origin of the coupling between allosteric and active site? Which quantity is transmitted: strain or energy?
- **Designs.** Do the designs depend on the information that is transmitted? Are there other designs than hinge and shear that are apt to achieve such propagations of information at long distances?
- **Classification.** The classification presented in [44, 45] is an attempt. Is it possible to properly classify the designs and unravel which mechanical principles are responsible for making them work? Have these designs evolved to be close to optimality for the task they have to perform?
- **Elastic description.** The typical elastic network model involves linear elasticity, while proteins are intrinsically nonlinear, as discussed in the population shift scenario. How are the questions just raised affected by the presence of nonlinearities in the elastic description of the protein?

The role of sequence data in understanding protein allostery

The footprints left by the evolution that shaped proteins during billions of years are a rich source of information, putatively also for protein allostery. Mechanisms that occasionally duplicate genes allow one copy of the gene to evolve independently to enable it to perform a new function; these two divergent genes are called paralogs. Speciation is also responsible for a divergence, where genes usually share similarities in function and are called orthologs. Both processes have been abundant in the history of evolution [1], a sketch of an evolutionary tree summarising these effects is shown in Fig. 7.

As a consequence, many proteins that exist today possess similar sequence and structure so to be grouped in a protein family, even though they can express different biological activities depending on the required degree of sequence similarity. It happens also that the sequences of two proteins have diverged so much that their family relationship cannot be assessed without a structure determination. With the recent increase of the number of sequenced proteins in the same family, researchers started to ask how to extract information from the sequence variability acquired during evolution. In the following, the relevant ideas to progress in this directions are outlined, starting from the role of co-evolution [63], commonly defined as *reciprocal evolutionary change in interacting specie*s [64].

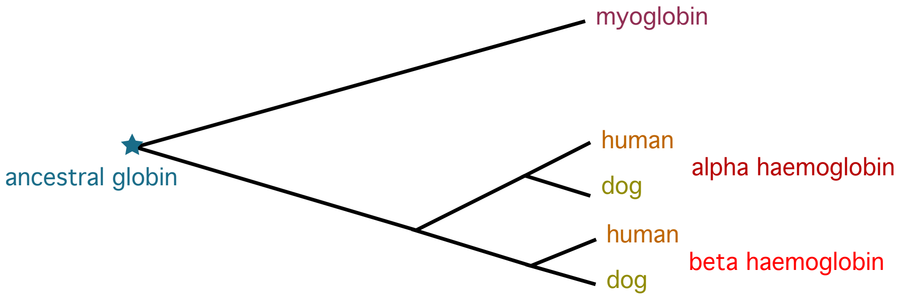

Figure 7 – A sketch of the phylogenetic tree of globins shows the divergence of the evolutionary path from the common ancestor, globin, to the appearance of myoglobin and heamoglobin chains as separation of branches in the tree — called paralogs. Each branching point indicates a common ancestor for the following lines. Smaller evolutionary differences are seen when the two haemoglobin chains form and differentiate in different organisms — called orthologs.

Proteins, big data and co-evolution

In the last 15 years (i.e. since 2005) massively parallel techniques for sequencing DNA known as next generation sequencing became commercially available [65], marking the steep growth of the number of protein sequences recorded on online databases, like UniProt. Indeed, DNA sequencing is the most efficient way to date to perform protein sequencing. This progress allowed to build *protein families* with a considerable amount of entries, frequently up to $10^3 - 10^5$ sequences per family, and made the use of big data in bioinformatics possible.

But what is the relevance of a protein family? A protein family is defined as a set of proteins that have common evolutionary history, known as homologs[3], conserved three-dimensional structure and function, but whose sequences have diverged in evolution — given the different organisms they are taken from or paths they have followed, see Fig. 7 — thus displaying only 20-30% of sequence identity. Indeed, mutations occurred in the course of evolution are responsible for the considerable amount of variability observed between sequences in the family. Nonetheless, this variability is not meaningless. Few mutations introduced at random in one protein are found to be enough to break its structure and function [67–69]. Hence, traces of what makes structure and function conserved in the protein family must be present in the observed variability. Indeed, two positions in the amino acid sequence can be found to have evolved together — *co-evolved* — to obey structural or functional constraints, even though singularly they would present large variability. This gives the hope to find information about the evolutionary constraints acting on a given protein that reflect structure and function from sequence data alone. Since a large amount of sequence entries in UniProt are not annotated — note that more than 5000 protein family do not have an example structure — the acquisition of information solely from sequences is of great interest.

---

[3] As an example, haemoglobin and myoglobin are homologous proteins since they share the same evolutionary history, as sketched in Fig. 7. Protein families can also be defined as being composed of solely orthologs. This definition would allow to explore the landscape that associates fitness to a sequence, see Fig. 13, around one native minimum, in contrast to a protein family between homologs that considers proteins better distributed over the space of all possible sequences [66].

Proteins in the same family may be composed of different numbers of amino acids, thus with sequences of different lengths that cannot be compared. An alignment of these sequences is then needed to be able to compute their statistical properties. Building a matrix composed of aligned sequences, the Multiple Sequence Alignment (MSA), constitutes a research field on its own [70] and the quality of the alignment deeply influences the outcome of any analysis that uses the MSA as starting point. Several methods exist to deal with the high complexity of the task, but they are overall based on adding gaps (–) until all sequences have the same length and both repetitive and conserved regions are matched without the possibility that a column in the matrix consists only of gaps.

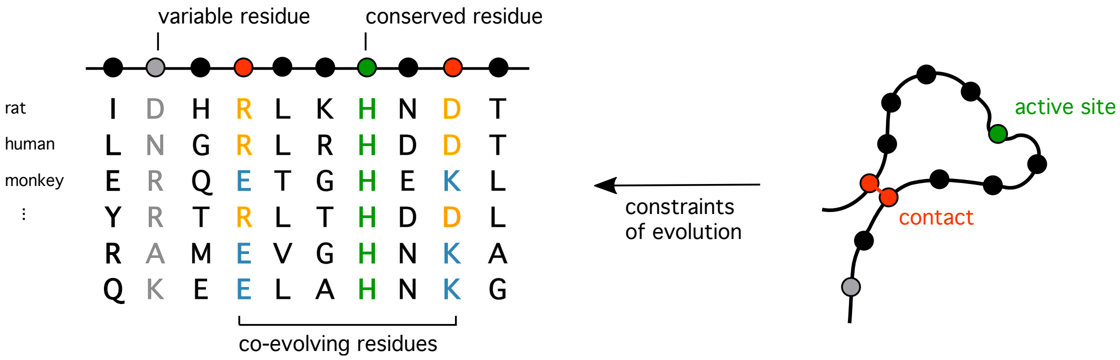

Figure 8 – Members of a protein family have evolved in different environments while still displaying overall the same structure and function. The interplay between these evolutionary constraints and the sequences in the MSA are summarised in this figure. On the left, few sequences in an example MSA are shown, where conserved, variable and co-evolved position are highlighted with colours. On the right, a sample sequence from the MSA is folded in the corresponding protein, where the connection between statistical properties of the MSA and evolutionary constraint is sketched.

Given an MSA, statistical observables can be directly measured. Conservation of amino acids among sequences is the most immediate observable to extract functional information from the MSA [71–75]. Indeed, from Fig. 8 it emerges that a position in the sequence diplays large conservation as a footprint of the crucial role it plays in the protein, as could be the case for a functional site (represented as a green circle in Fig. 8) or a position located in the core of the protein. Indeed, evolution will most likely keep the functional site of a protein in the family conserved, then imprinting this constraint in the MSA.

Further, more subtle information can be extracted if measuring correlations between amino acids in two different positions [76–78]. Fig. 8 shows that two positions co-evolve together when the interaction between the two is crucial, as it happens e.g. for two contacting residues shown as red disks in the figure. More generally, co-evolution appears when two positions display coordinated changes that occur to maintain or refine functional interactions between them, thus suggesting the presence of correlated mutations necessary to sustain overall protein stability, function or folding. Hence, the significant correlation displayed between the pair suggests the possibility to exploit it as a proxy for contact prediction [77].

Soon after the relevance of co-evolution in MSA was discovered, the biology, bioinformatics and physics community thrived to pin down the most efficient ways of extracting the structural and functional constraints imprinted in the MSA by evolution, as it is discussed in the following sections.

Inferring structure from sequence

The study of co-evolution in protein families started with measuring correlations between the amino acids of two different positions in the MSA. Positions displaying large correlation were found to correspond to neighbouring residues in the protein structure, which helped to guide structure prediction, but still provided a non negligible amount of false contacts. The accuracy of contact prediction improved once it was realised that the measure of correlations cannot disentangle between direct and indirect interactions: long-range correlations can arise from networks of short-range couplings, thus adding noise to the prediction, see Fig. 9. Indeed, Fig. 9B shows that direct pairwise couplings are needed to have a good prediction of contacts and not positions that display correlations and mutate in a coordinate fashion due to higher than primarly effects. But what are these direct couplings? How can they be extracted from the MSA?

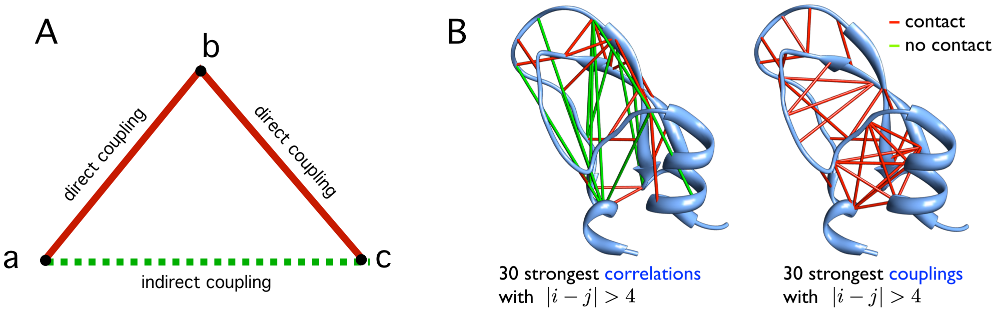

Figure 9 – Figure B is adapted from [69]. (A) This simplified interaction pattern aims to show the difference between direct and indirect interactions. If **a** interacts with **b**, e.g. through a contact, and **b** interacts with **c**, then a correlation is also measured between **a** and **c** originating from the indirect interaction of both **a** and **c** with **b**. This indirect correlation originates a false contact when correlation measures are used for contact prediction. (B) The largest pairwise correlations and couplings are shown for the PF00014 protein family (trypsin inhibitor) highlighting the amount of false positive contacts predicted when considering large pairwise correlations. The analysis excludes unsignificant predictions with separation $|i-j| \leq 4$ along the amino-acid sequence, where the value four is chosen since it corresponds to the distance between one turn in an α-helix.

In statistical physics, direct pairwise interactions between degrees of freedom are the basis of the Ising model [79], whose energy function for a discrete variable $\sigma_i = \{-1, 1\}$ reads

$$\mathscr{E}(\{\sigma\}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i\,. \tag{8}$$

The rules of interaction between the binary variables $\{\sigma\}$ are set by the choice of the coupling matrix J. The term h_i acts like a field favouring the variable σ_i to align along the field, mimicking the effect of an external magnetic field in physics or of conservation in the context of proteins. If $J_{ij} > 0$ the two variables are favoured to have the same value when minimising Eq. 8, while the opposite when $J_{ij} < 0$. Moreover, J_{ij} can be set to be nonzero only when (i, j) satisfy specific conditions, e.g. they are first-neighbours when $\{\sigma\}$ are embedded in a lattice. Hence, the common use of the Ising model is to produce configurations $\{\sigma\}$ that minimise Eq. 8 with the chosen microscopic interactions and then study macroscopic observables, like magnetisation and correlation, $\langle\sigma_i\rangle$ and $\langle\sigma_i\sigma_j\rangle$, where the average is performed over the sequences $\{\sigma\}$ sampled at equilibrium by the Boltzmann distribution $P(\{\sigma\}) = e^{-\beta\mathscr{E}(\{\sigma\})}$. The inverse problem can also be studied, where the microscopic parameters are inferred from the knowledge of the configurations, giving a useful framework where direct couplings are well defined and can be inferred from sequence data [80]. In the following, this framework — first applied to proteins in [81] — is discussed in the context of proteins with its successes in predicting contacts and structure.

Direct coupling analysis

One successful inference method of couplings and fields from sequence data takes the name of Direct Coupling Analysis (DCA). Each entry of the MSA of a given protein family is seen as a sequence σ^a with $a = 1, \dots, M$ where each element σ_i^a represents one amino acid with $i = 1, \dots, N$ and can take $q = 21$ values. Thus, it defines an extension of the Ising model to variables with $q > 2$ degrees of freedom, and is known as the Potts model. Note that M denotes the number of proteins inside the family and N the number of amino acids in the sequences. The desired framework for performing inference is a global statistical model that assumes sequences in the MSA of the protein family to be generated from the same unknown energy function $\mathscr{E}(\{\sigma\})$ and hence to obey to the same fitness energy landscape. How to define the probability distribution $P(\{\sigma\})$ — of observing a sequence $\{\sigma\}$ — for whole amino acid sequences in the MSA under study?

The method

First, the model to be inferred needs to be consistent with the statistics of the empirical data. A numerous group of works [68,81,82] showed the sufficiency of pairwise interactions to capture the effective features of the sequence ensemble, hence supporting to fix the marginals of $P(\{\sigma\})$ to the empirical single- and double-site frequency observed in the MSA. This assumes that pairwise couplings are sufficient to capture the amino acid variability in the MSA and that correlations between different positions result collectively from a network of direct couplings. The conditions on the marginals read

$$\begin{aligned} P_i(\sigma_i) &= \sum_{\sigma_k | k \neq i} P(\{\sigma\}) = f_i(\sigma_i) \\ P_{ij}(\sigma_i, \sigma_j) &= \sum_{\sigma_k | k \neq i,j} P(\{\sigma\}) = f_{ij}(\sigma_i, \sigma_j), \end{aligned} \tag{9}$$

where $f_i(\sigma_i)$ counts the number of proteins in the MSA that display amino acid σ_i in position i, while $f_{ij}(\sigma_i,\sigma_j)$ counts the number of entries for which amino acids σ_i and σ_j are correlated. Since there is no additional knowledge on how $P(\{\sigma\})$ should look like other than obeying the constraints of Eq. 9, looking for the most general $P(\{\sigma\})$ would give the least biased form of this probability. One way of finding the least-biased model is to maximise the entropy of the probability distribution [83][4]

$$S = -\sum_{\sigma_i|i=1,\ldots,N} P(\{\sigma\}) \ln P(\{\sigma\}), \tag{10}$$

while imposing the constraints Eq. 9 as Lagrange multipliers. The maximisation gives the functional form for $P(\{\sigma\})$ which is the same as the equilibrium distribution of the Ising or Potts model, the Boltzmann weigth

$$P(\{\sigma\}) = \frac{1}{Z} \exp\left[\sum_{i<j} J_{ij}(\sigma_i,\sigma_j) + \sum_i h_i(\sigma_i)\right] \equiv \frac{1}{Z} \exp\left[-\mathscr{E}(\{\sigma\})\right] \tag{11}$$

$$Z = \sum_{\sigma_i|i=1,\ldots,N} \exp\left[\sum_{i<j} J_{ij}(\sigma_i,\sigma_j) + \sum_i h_i(\sigma_i)\right], \tag{12}$$

where $J_{ij}(\sigma_i,\sigma_j)$ and $h_i(\sigma_i)$ are the Lagrange multipliers resulting from imposing Eq. 9.

A limitation resides in computing the summations in Eq. 9 since they are performed over all amino acids apart i and j, thus implying a computational time that goes like q^N and becomes intractable when the length of the sequences N increases. The partition function itself, Z, also becomes intractable for large N for the same reason. Moreover, the marginals in Eq. 9 can be obtained via the partition function by derivation

$$\begin{aligned} \frac{\partial \ln Z}{\partial h_i(\sigma_i)} &= P_i(\sigma_i) \qquad (13)\\ \frac{\partial \ln Z}{\partial h_i(\sigma_i)\partial h_j(\sigma_j)} &= -P_{ij}(\sigma_i,\sigma_j) + P_i(\sigma_i)P_j(\sigma_j) \equiv -C_{ij}(\sigma_i,\sigma_j), \end{aligned}$$

where the connected correlation matrix $C_{ij}(\sigma_i,\sigma_j)$ is defined. Several ways to approximate Z have been laid out to perform DCA with a different accuracy in inferring couplings and fields — where accuracy can be either measured via the fractions of predicted contacts or via comparison with the exact solution for small N. The approximations consist in message-passing algorithms [80, 84], improved Monte Carlo sampling [85], perturbative expansions [86–90], pseudo-likelihoods [91, 92]. The simplest method to implement computationally is mean-field DCA (mfDCA) [81] based on the mean-field approximation for which the joint distribution of all amino acids factorises. Indeed, the mfDCA-inferred coupling simply reads

$$J_{ij}(\sigma_i,\sigma_j) = -(C^{-1})_{ij}(\sigma_i,\sigma_j). \tag{14}$$

---

[4] Note that the uniform distribution carrying no information maximises the entropy by construction.

Success in the prediction of contact in the three-dimensional structure

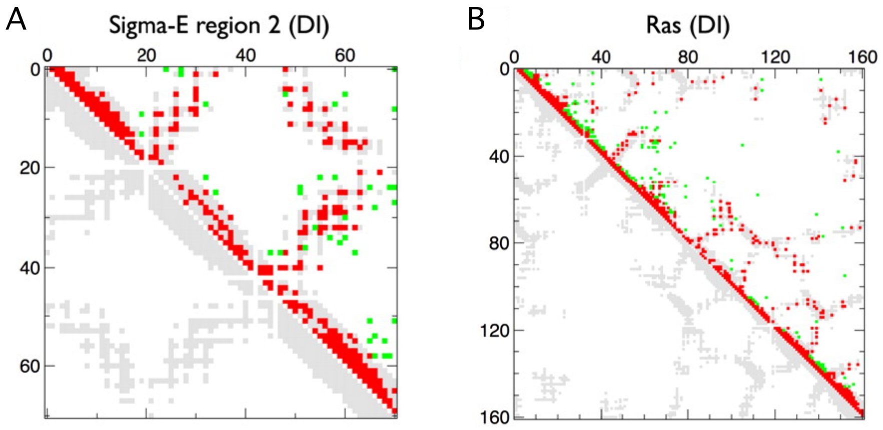

Figure 10 – The figure is adapted from [81]. Given a protein sequence of length N, the $2N$ highest-ranking couplings inferred via mfDCA are superposed to the contact map of the associated protein (in grey). The contact map is zero if residue i and j are at distance greater than a threshold, here set to 8 Å, and one if the residues are closer. Red squares identify true positive, while green squares spatially distant pairs. The minimum separation between two amino acids along the sequence is of five positions, $|i-j|>5$. Two examples of contact map prediction are shown, on the left for the SigmaE factor and on the right for the eukaryotic signaling protein Ras.

Albeit the simplicity of the approximation, the couplings inferred via mfDCA show to predict with a good accuracy couples of positions that are close in space, as shown in Fig. 10 with the results from [81]. Contacting amino acids are found among the couples (i,j) with largest $J_{ij}(\sigma_i,\sigma_j)$. The accuracy of the predictions increase with more refined methods as pseudolikelihood or Adaptive Cluster Expansion (ACE), where maximum likelihood is performed with an initial condition for the parameters given by constructing a sparse network of interactions sufficient to reproduce the observed correlation data, thus avoiding potential overfitting [89, 93].

Overall, DCA is able to predict with high accuracy the contact map of a protein family, i.e. the couples of positions that are close in space, serving as a valuable information for the prediction of structure from sequence, as shown in Ref. [94] where it guides the discovery of new folds. In 2019, a research group at DeepMind (Google) created a machine learning scheme, AlphaFold [95, 96], to predict structure from sequence data where contact predictions from DCA are used as one of the informative features to get distances between pairs of amino acids necessary for structure prediction. Indeed, the use of contacts maps from DCA was already shown to improve significantly the performance of structure prediction [97], but never with an improvement as impressive as the one of AlphaFold.

Extracting function from sequence

A pair of amino acids in contact is predicted by looking at the largest inferred couplings $J_{ij}(\sigma_i,\sigma_j)$. When the method is applied on a known structure it identifies a certain number of false positive contacts, among which some correspond to residues that are far away in space if looked at in the corresponding crystalline structure. Are these positions encoding for some long-range functional coupling or do they originate solely from X-ray artifacts? Is it possible to extract functionally related positions with methods like DCA or other approaches are needed?

In 2017, a research article [98] tackled the issue of the origin of coevolution between residues that are distant in the three-dimensional structure. A systematic study over 4000 protein families showed that the majority of pairs are found to be in contact in at least one structure in the family and thus little evidence is found to support the idea of large couplings between distant allosteric and functional sites. Hence, it seems that the long-range couplings inferred with DCA are not related to allosteric communication. In the following subsection, another method based on co-evolutionary information is introduced [99], which provides more insight on functional features by identifying larger groups of co-evolving residues with respect to pairwise couplings inferred by DCA.

Statistical coupling analysis

The correlation matrix C_{ij} defined in Eq. 13 can be directly used to study co-evolution [99] where — instead of looking at the most correlated *pairs* of residues — *groups* of co-evolving positions are identified via the largest entries of the modes of C_{ij} with highest variance, thus counting for most of the variability in the dataset. This approach takes inspiration from the sector analysis developed for financial markets, where relevant correlations of stock performance over a finite sized time window are extracted from the noise. Indeed, most correlations simply arise due to the limited period of time during which the prices are sampled and are then non significant. It emerges that the few relevant correlations are organised in collective modes, each describing a sector of the economy whose performance fluctuates together over time. This study inspired the search for sectors in proteins by studying the modes of C_{ij}, hence defining the Statistical Coupling Analysis (SCA).

In particular, the matrix used in SCA is the conservation weighted analog of C_{ij}, justified as to possibly look at the contribution of correlations to conservation

$$\tilde{C}_{ij}(\sigma_i,\sigma_j) = \phi(\Sigma_i^{\sigma_i})\phi(\Sigma_j^{\sigma_j})|C_{ij}(\sigma_i,\sigma_j)|, \tag{15}$$

where the weights are a function of the conservation of amino acid σ_i at site i, $\Sigma_i(\sigma_i)$[5]. The

---

[5] In the literature, they are usually taken to be gradient of conservation, i.e. $\phi = \partial\Sigma/\partial f$.

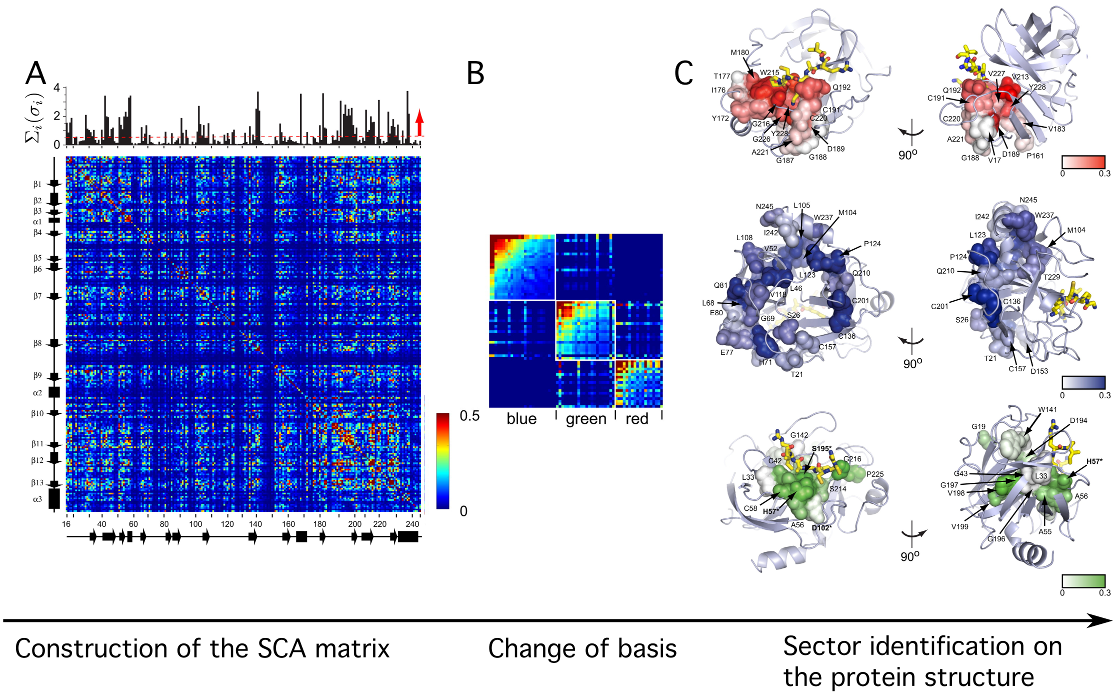

Figure 11 – All the figures are adapted from [100]. Statistical coupling analysis is performed on the S1A protease family. (A) The SCA matrix is shown. Secondary structures elements are specified with arrows and squares and the profile of amino acid conservation $\Sigma_j(\sigma_j)$ is shown above. (B) The matrix rewritten in the basis of the chosen modes as in Eq. 17 shows a structure with three independent diagonal blocks identifying three sectors, coloured as blue, green and red. (C) The three sectors are highlighted in the protein structure of a member of the S1A protease family, the rat trypsin (with PDB identifier 3TGI). The colour is according to the strength of contribution in each sector.

conservation, with $\overline{f}(\sigma_i)$ the frequency of σ_i in all proteins in the family, reads

$$\Sigma_i(\sigma_i) = f_i(\sigma_i)\ln\frac{f_i(\sigma_i)}{\overline{f}(\sigma_i)} + \left[1 - f_i(\sigma_i)\right]\ln\frac{1 - f_i(\sigma_i)}{1 - \overline{f}(\sigma_i)}\,. \tag{16}$$

An example of application of SCA to the S1A protease family is shown in Fig. 11 from [100].

The SCA analysis consist in a sophisticated principal component analysis of the SCA matrix, Eq. 15. The first N_Γ modes $\boldsymbol{\psi}^\gamma$ of $\tilde{C}_{ij}(\sigma_i,\sigma_j)$ are extracted, with N_Γ identifying the number of modes with largest eigenvalue that are clearly separated from the spectrum of random data, admitting that such a clear separation occurs. A position i is part of a sector when the corresponding component i of at least one of these N_Γ modes has an absolute value larger than a threshold[6] ε, $|\psi_i^\gamma| > \varepsilon$. Once the relevant modes are chosen, the significant correlations

---

[6] Other relationships involving components of different modes have also been used to define sectors, as discussed in [100], where relationships between two modes, as could be e.g. $\psi_i^2 > |\psi_i^4|$, translate an observed correlation between those components and define directions in modes space where sectors emerge. Moreover, a sector could also be defined by a linear combination of different eigenvectors. Hence, these observations underline the details in the method that do not make it robustly and systematically applicable.

can be represented by the following matrix reduced to sector position shown in Fig. 11B

$$\tilde{C}_{ij}(\sigma_i,\sigma_j) = \sum_{\gamma=1}^{N_\Gamma} \lambda_\gamma |\psi^\gamma\rangle\langle\psi^\gamma| . \tag{17}$$

A variant approach solely based on the covariance matrix C_{ij} has been introduced in [101] where is shown to be more robust than SCA and to perform better when the conservation bias in Eq. 15 is dropped. The method is based on the observation that sectors are identified from the low energy modes of the covariance matrix and considers the inverse of the off-diagonal terms of the covariance matrix C_{ij} as observable for predictions — it is indeed called ICOD, which stands for Inverse Covariance Off-Diagonal.

Success in the prediction of functional dependencies

The way of defining sectors is somewhat arbitrary — it depends on the two parameters N_Γ and ε and also on the weights ϕ in the definition of the SCA matrix — and it is not clear how it is applicable to proteins other than the few ones studied in the literature for which a thorough knowledge about their functioning is already at hand to guide sector identification. Anyhow, the available results show that meaningful sectors can be identified in some proteins.

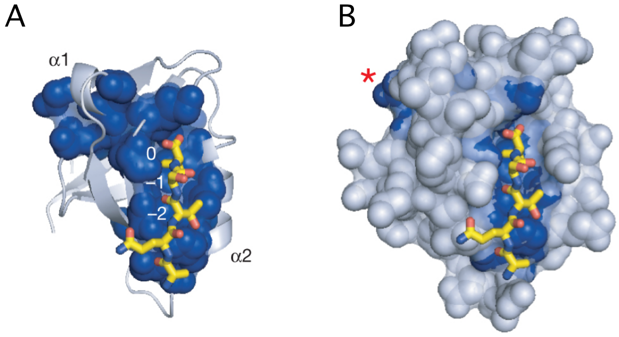

Figure 12 – The figure is adapted from [102]. The residues involved in the sector are shown as blue spheres, while a cartoon (A) or space filling (B) representation shows the structure of rat PSD95$^{\mathrm{pdz3}}$ with PDB identifier 1BE9. Yellow stick bonds represent the co-crystallized peptide ligand. The sector is composed of a sparse network of residues around the ligand-binding pocket and that connects to a distant active surface site [103], labelled with an asterisk, through a subset of amino acids within the protein core.

For the S1A protease family shown in Fig. 11, the positions belonging to the three sectors are located in areas responsible for different and independent functions according to the sector they belong to [100]. Moreover, in distinct protein families SCA identifies one sector [102, 104–107] whose amino acids form a chain connecting the allosteric and active site, thus supposedly identifying a pathway of residues mediating the allosteric communication solely from the co-evolutionary information in the MSA [103, 104, 108, 109], see Fig. 12 for the case of PDZ domain and Fig. 2A for haemoglobin. [110].

However, the idea of allosteric pathways should be taken with care. Indeed, the stress induced by binding in the conformation does not have necessary to generally follow residue by residue the allosteric pathway, but it could propagate non-locally at the distal functional site thanks to the presence of rigid parts — as could be β-sheets in the secondary structure, Fig. 1. Moreover, in recent molecular dynamics experiments of a photoswitchable PDZ domain[7] [111] it has been shown that different trajectories are highly heterogenous, suggesting for the existence of several possible pathways of the allosteric transition. A successful alternative is to define the pathway in Fourier space instead of real space as the dominant normal mode [112], usually global, or a sum of normal modes [113].

Moreover, Refs. [33, 114, 115] used the information obtained by SCA to locate[8] functional hotspots, then used it to engineer proteins with controllable functions. Other works [117, 118] use information on conserved residues to locate functional sites and engineer functionalities, which performs similarly than SCA given the strong component that conservation plays in the SCA matrix [119].

Overall, sectors obtained from SCA are found to be sparse, being composed of 20% of the total number of amino acids, and to be organised into structurally adjacent networks of co-evolving residues. These findings suggests that natural proteins are organised into groups of amino acids each related to conserved functional activities and that can be identified from the different evolutionary pressures they have been subjected to [100, 120–122].

A model for protein fitness

The success of co-evolutionary methods in predicting structural and functional constraints hints that models describing the fitness of proteins may also be built starting from sequence data of one protein family [66] — supposing that a fitness which is a global function of sequences in the protein family is biologically relevant. This is true since the proteins in a family are the result of evolutionary pathways, like the one sketched in Fig. 7, that are in turn determined by the shape of the fitness landscape. Thus, if co-evolutionary methods are able to capture some of the evolutionary constraints, they may also succed in modelling, to a given extent, the landscape responsible to generate those constraints.

Fitness can be defined in different ways, one definition that is straigthforward measures fitness as how well the protein performs its task. Hence, fitness can be experimentally measured by looking at binding affinity or at the activity upon external responses as for photosensible domains. The concept of fitness landscape was introduced in 1931 in a seminal article by Wright [124] as to understand the effect of mutations on evolutionary processes. A sketch of a fitness landscape is displayed in Fig. 13. There, fitness $\mathscr{F}$ is shown as a function of sequence $\{\sigma\} = \sigma_1, \ldots, \sigma_N$, representing the genotype-to-phenotype — i.e. sequence-to-fitness — map,

---

[7] PDZ2S with PDB identifiers 2M0Z — inactive — and 2M10 — active.

[8] See [116] for a scheme to locate functional hotspots via SCA and how to engineer new functions via their knowledge.

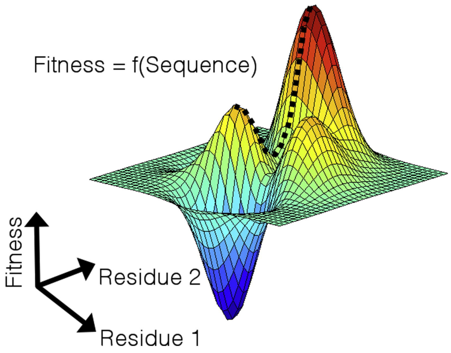

Figure 13 – This figure is adapted from [123]. Sketch of a genotype-to-phenotype mapping, $\mathscr{F}(\{\sigma\})$. The amino acid sequence specifies a location on the landscape and the height of the landscape its fitness. For visualisation purposes, only two residues are shown, but the landscape is generally high-dimensional.

$\mathscr{F}(\{\sigma\})$; N denotes the length of the sequence. For limits of visualisation, only two amino acids are shown in sequence space.

The definition of such mapping identifies evolution as the process of selecting sequences in the landscape according to their fitness. Its characterisation, however, is non trivial given both the impossibility to experimentally measure the fitness of all mutated sequence and the effect of non additivity of mutations, called *epistasis*, which may lead to a rugged landscape with an extensive number of local optima. The advance of experimental techniques determing fitness changes upon mutations, as deep mutational scanning[9] [126], allows to probe local patches of the fitness landscape centered around the native sequence.

The fitness landscape is crucial to determine the effect of amino acids mutations on function. Which are the positions in the sequence that affect function the most upon mutation? Which are the effects of simultaneous mutations at different sites? The group of Chakraborty with the work of Ref. [123] studied the fitness landscape of HIV to find which regions are the most vulnerable to be then targeted via an immune therapy. Moreover, the knowledge of the fitness landscape allows to perform directed evolution [127]: exploring the landscape in new ways that were not chosen in nature to discover possibly new proteins.

The experimental measure of the fitness cost of mutations is however involving and becomes intractable at higher orders purely for combinatorial reasons — even with the new advances in mutational scans, in the case of simultaneous mutations at several sites. Hence, the interest of building a model for protein fitness able to predict mutation costs from co-evolutionary information.

---

[9] On this line, a recent work [125] defines and evaluate different strategies to choose a minimal set of experimental mutational data and use it to guide predictive computational models so to reduce the number of mutations required for predictions.

The role of epistasis

Among the set of observables characterising the fitness landscape, one in particular has received considerable attention in recent research. Several works have highlighted the role of the non-additivity of mutations defined as epistasis [128]. The effect of epistasis to fitness is found to be either beneficial or deleterious, as discussed in [129]. Moreover, epistasis is responsible to influence evolutionary paths by reshaping the fitness landscape itself [130]. Indeed, epistatic interactions with deleterious effects may restrict the trajectories available to a protein during evolution or if beneficial, on the contrary, create new paths leading to sequences and functions that would otherwise have been inaccessible. Overall, epistasis increases the ruggedness of the fitness landscape since the physical effect of a mutation depends on where in the sequence it is introduced, i.e. on which directions it takes in the landscape.

Given the fitness $\mathscr{F}$ and the fitness cost upon mutation at position i, $\Delta\mathscr{F}_i = \mathscr{F}_i - \mathscr{F}$, second order epistasis reads

$$\Delta\Delta\mathscr{F}_{ij} = \Delta\mathscr{F}_{ij} - \Delta\mathscr{F}_i - \Delta\mathscr{F}_j \,. \tag{18}$$

Its empirical measurement, involving mutations at at least two pairs of positions, remains elusive given the combinatorial complexity, even though deep mutational scans are able to efficiently reduce this complexity [131–134]. The resulting epistatic interactions are found to involve positions that can be distant in the structure of the protein [132, 134, 135], usually related to distant functional dependencies, as allosteric interactions described in the previous section.

A recent work [136] measured epistatic interactions up to the 13th and showed that frequencies and correlations of a MSA of fluorescent proteins give a good prediction of the true second order epistasis, suggesting that the co-evolutionary methods discussed above should be able to capture it as well.

Fitness landscape as a tool to generate new sequences

Ultimately, a model accurately predicting the fitness landscape $\mathscr{F}(\{\sigma\})$ would be able to generate new artificial sequences that fold in the corresponding native structure, opening new possibilities for protein and drug design. This means that such a model would capture the overall statistics of the landscape, or the non additivity of many mutations, not just of two.

The seminal work by Ranganathan and collaborators [105, 137] goes in this direction. They showed that the information contained in the frequencies and correlations of the MSA of a small protein module, the WW domain, is sufficient to design artificial sequences in vitro whose a sizeable fraction folds correctly in the native state. The sequences were generated by recombining natural sequences in a way that preserves the measured single- and double-site frequencies. The fraction of correct fold vanishes when sequences are generated by using information only on single-site frequencies. Even though this result is valid only for a tiny protein it demonstrates the power of a free energy function inferred from statistical properties of the MSA, over an ad-hoc physical free energy function.

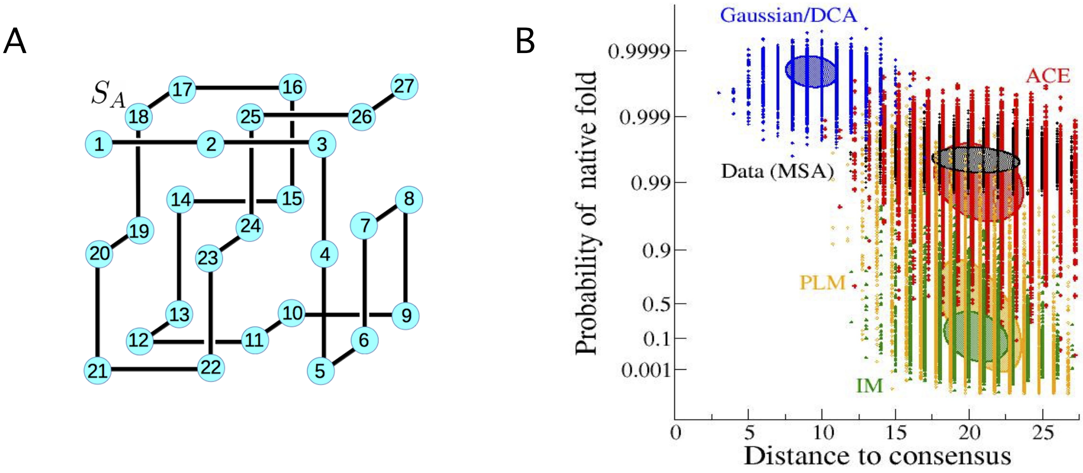

Figure 14 – The figure is adapted from [138]. (A) A structure S_A of a protein lattice is shown where amino acids (blue circles) are identified with their number, $n = 1, \dots, 27$, along the protein backbone (black line). There are 28 additional contacts between nearest-neighbour amino acids that are not along the backbone, e.g. between 1 and 18. (B) Probability of finding a native fold against their Hamming distances to the consensus sequence of the original MSA of the structure used to infer the four models. The results are shown for sequences randomly generated with different inference procedures: the independent-site model (IM, green) — fully determined by conservation, the Potts-ACE [89, 90] (red), the Potts-pseudo-likelihood method [92] (orange) and the Potts-Gaussian [81] (blue). Black circles encode the results for the original sequences. Filled ellipses represent domains corresponding to one standard deviation around the mean in both directions.

Given the experimental difficulty of testing the generative power of a model with protein data, simple protein models come at hand as a benchmark of existing inference methods. Lattice proteins have been introduced in [139–142] to provide a coarse-grained description of protein folding. See Fig. 14 for a sketch of a lattice proteins with 27 amino acids. Protein lattices are used in Ref. [138] to benchmark the performance of pairwise models like DCA in generating new sequences in silico. The probability distribution inferred via DCA is found to generate sequences that are close both in variability and fitness to the native ones, see Fig. 14B. A model based only on conservation is proven to fail in generating sequences with the proper fold and function. This suggests again that structural constraints are captured by the inference method, where pairwise interactions are needed to achieve generative power. However, protein lattices do not account for functional dependencies as allosteric interactions. Hence, in-silico sequences evolved with a model enforcing allostery are needed to see whether the remarkable generative power of DCA is kept also when functional constraints are imposed.

In the following, the contributions of the complementary co-evolutionary methods, SCA and DCA, to predict protein fitness are overviewed in more detail.

Contributions of Statistical Coupling Analysis

SCA makes predictions directly using the statistical properties of the MSA — conservation and covariance — and it does not learn a probability distribution, in contrast with DCA, see Eq. 11. How to then define a model of fitness from SCA from which is possible to predict mutation costs and generate sequences?

Prediction of binding affinity. Suppose an MSA where additional new sequences do not change significantly the distribution of amino acids at positional sites. The frequencies of mutations in such an MSA obey a Boltzmann distribution as function of the binding free energy ΔG_i^x

$$\frac{P_i^x}{P_{\mathrm{MSA}}^x} = \exp\left(\frac{\Delta G_i^x}{kT}\right), \tag{19}$$

where P_i^x is the probability of any amino acid x to be at position i and P_{MSA} the mean frequency of amino acid x in the MSA. Hence, the statistical binding free energy reads

$$\Delta G_i^x = kT\left(\ln P_i^x - \ln P_{\mathrm{MSA}}^x\right). \tag{20}$$

The prediction of the empirical effect of mutations to the binding energy can be successfully performed by using this formalism, as reported in [99, 102].

Generate new sequences. As mentioned above, in the work of Refs. [105, 137] the information from the SCA matrix is combined with the knowledge of the number of amino acids of each type to construct new sequences that are found to fold and function correctly, suggesting that SCA — and other models based on conservation and correlation as well — captures significant features of the fitness landscape for a small protein domain.

Prediction of epistasis. Another way of using the co-evolutionary information from SCA is to directly consider the covariance matrix reduced to the relevant modes (Eq. 17) to predict the effects of mutations. Ref. [134] showed that in the case of the PDZ domain the SCA matrix is able to capture the strength of the measured epistatic interaction, even between long-range functionally related pairs.

While SCA showed success in the predictions of mutation effects and in guiding synthetic sequence design, it does not provide a unique and consistent way of performing these tasks. On the contrary, as how DCA is built, it learns the probability distribution of a sequence with a given fitness, i.e. it learns the fitness landscape from co-evolution. Hence, it provides a well-defined mathematical framework to predict mutation costs and generate new sequences.

Contributions of Direct Coupling Analysis

The other co-evolutionary approach discussed in this dissertation is DCA and infers evolutionary couplings between amino acids in the sequence alignment that are found to predict successfully contacting residues, as discussed above. The inferred energy function of the Potts model $\mathscr{E}(\{\sigma\})$ can be identified with the fitness of a protein sequence via $\mathscr{F}(\{\sigma\}) = -\mathscr{E}(\{\sigma\})$. This identification is valid if the fitness landscape is explored ergodically so that fitness obeys a Boltzmann law as the one of the inferred energy Eq. 11. Are the parameters inferred by imposing frequencies and correlations observed in the MSA sufficient to capture properties of the protein fitness landscape through $\mathscr{E}(\{\sigma\})$?

Prediction of fitness landscapes. Several works [123, 143–147] focused on the description of evolutionary fitness landscapes with methods based on DCA showing that the effect of mutations are well predicted at different orders, while not focusing on mutations involving functional residues that are of interest for allostery and protein and drug design.

Generate new sequences. The knowledge of the functional form of the inferred energy function allows to generate new sequences distributed according to its Boltzmann distribution. The generative power of DCA is found to be good in lattice models as shown in [138] where new sequences are drawn from the inferred distribution Eq. 11 and are found to reproduce the fitness of native ones. This analysis is based on lattice models for protein folding, thus describing structural constraints and not including functional allosteric interactions. Moreover, the recent work in Ref. [148] shows that DCA is a good generative model for the enzyme chorismate mutase with 45% of the designed sequences found to be comparabe with natural-like sequences, which is again in strong support for DCA as a good generative model.

Prediction of epistasis. In the work of Ref. [134] the performance of SCA on predicting epistasis is compared to DCA. It is found that DCA is outperformed by SCA in the inference of epistatic interactions, which is largely do to a failure in the prediction of long-range epistatic terms.

Summary and open questions

Two complementary methods of extracting the information imprinted by evolution in sequence alignments have been discussed. Direct coupling analysis (DCA) [81] proved to be successful in systematically inferring which amino acids are in contact in the three dimensional structure, while statistical coupling analysis (SCA) [99] identifies groups of residues that co-evolved that are sparse, contiguous in space and correlate with functional features with evidence in some proteins. How do these two methods compare for the inference of couplings between residues involved in allostery?

Allosteric proteins display a coupling between distant sites that are functionally related via the use of different architectures. A systematic study of DCA on thousands of proteins [98] shows that there is no statistical evidence for the existence of long-range direct couplings that involve allosteric communication. This is in striking opposition with the presence of extended sectors connecting functional sites reported in [102, 107] and the observation of long-range epistasis [135]. An open question is then why a pairwise model like DCA should be successful at predicting protein structure, but not long-range functional dependencies. The limitations in this direction can be summarised in the following questions

- **Performance of co-evolutionary methods.** An open question is the benchmark on how current methods perform in the prediction of mutation costs and epistasis and in generating new sequences when an allosteric task functionally coupling distant residues is present. This would extend the benchmark of DCA on lattice proteins built to mimic protein folding [138] where the inferred pairwise model seem sufficient to generate new sequences with the true value of fitness and variability in the amino acid distribution.
- **Function prediction.** In the Introduction, an attempt of classification of allosteric architectures is discussed. Does the co-evolutionary information of the MSA contain also information about this classification? Is it possible to guess which architecture is in use in a given protein family by using solely sequence data without knowledge of the high resolution structure?

Sequences proved to be enough to predict the contact map of the protein. Being able to systematically discover whether such protein is allosteric or not from the same data would be a breakthrough that may answer fundamental questions on which are the requirements that made allosteric behaviour so common and yet crucial in proteins. Moreover, co-evolutionary information could also be used to engineer de novo allosteric proteins with novel functions that fold as naturally occurring proteins, with obvious applications in drug design.

Scope of the thesis and outline

The introduction to protein allostery and inference delineated the general context of the research addressed in this thesis.

Minimal models of in-silico evolution of allostery are a powerful tool to gain insight in the open questions formulated in this introduction. Quantities like fitness that are usually very hard to measure, can be readily accessed. In the line of the works on elastic network models previously discussed, in Chapter 1 we introduce an evolution scheme to evolve a harmonic spring network to display long-range communication. Indeed, an unevolved elastic material does not display action at a distance: the response upon a local perturbation decays rapidly as a powerlaw as function of distance [149]. The same holds for an amorphous material, as a protein can be related to [150], that exhibits long-range response only in critical conditions [151–154]. Hence, the elastic material needs to be evolved to perform an allosteric task. Which information is transmitted? Which architectures are good at transmitting what? The dependence of the allosteric architecture on the chosen task can be automatically probed by changing the fitness function in the evolution scheme. We explored two different tasks that resulted to give two dramatically different scenarios. The first optimises for a given strain upon perturbation at a distal site and produced a novel architecture functioning as a lever, possibly extending the observed categories of hinge and shear in proteins. The second optimises for cooperative binding energy and recovers the usual hinge and shear designs, as well as many others. With theoretical arguments and numerical verifications we are able to pinpoint which are the requirements that the network has to obey to display optimal cooperative binding. The appearance of an extended low energy mode is found to be related with the design of shear and hinge architectures and its stiffness is not random, but is shown to obey a precise scaling, it scales inversely with the size of the system to function properly. The results discussed in Chapter 1 are published in [155, 156].

Given the simplified assumption of harmonicity, we also characterise the behaviour of the network in the presence of non-linear springs. Chapter 1 discuss the work we did in this directions to study the mechanics of the induced fit and population shift models and that has been published in [157]. The energy landscape of the in-silico model in the harmonic framework always displays one energy minimum. Hence, it can be reconduced to a induced fit model. Oppositely, the network with additional elastic nonlinearity displays the two-state landscape of the population shift scenario, see Fig. 4. The cooperative fitness is found to be more robust towards mutations that increase the stiffness in the population shift scenario. Moreover, in the population shift model the onset of optimality happens dynamically at the same frequency scaling of the soft mode as in the induced fit scenario. It supports the prediction that that for proper function proteins must evolve a functional elastic mode that is softer as their size increases.

The classification resulting from the in-silico evolution can be tested with high-resolution structural data. First, architectures evolved to optimise the transmission of strain are found to display a lever mechanism to function. An analysis of the allosteric response of a PDZ domain shows that such lever design may also be used in proteins. Second, a database of allosteric proteins is built from the available literature to study the classification of cooperative architectures. A systematic analysis over the resulting dataset of 34 allosteric proteins confirms that in most of the cases one extended low energy mode describes a large fraction of the allosteric response. Moreover, the scaling of the frequency of the soft mode as function of the number of residues is consistent with the predicted scaling, indeed suggesting that larger proteins deveolp a softer functional elastic mode to function properly. An experimental work on a single protein is needed to confirm the validity of these findings. This empirical analysis are presented in Chapter 2.

The in-silico model is able to produce as many solutions as desired that can be translated in binary sequences reflecting the occupancy or not of springs. Hence, it serves as a benchmark to explore the performance of the existing inference methods to predict epistatic interactions — and more generally fitness landscapes — in a well controlled framework where an allosteric function, coupling distant regions of the protein, is considered. We find that usual inference methods are able to identify which links are relevant for the functioning of the architecture.

Furthermore, the inferred energy can be used to predict the mutational landscape of the system under study. We find that Direct Coupling Analysis (DCA) well predicts single mutations costs of the cooperative fitness governing the spring network. Still, it is not able to generate synthetic sequences that achieve to function. DCA is also found not to capture epistatic interactions involving pairs of distant residues, as could be two positions coupled by the allosteric function. We argue that the reason is that DCA fails when subparts of the system work in concert as it is the case in allostery. Methods usually more successful in extracting functional information like SCA do not provide much improvement in these tasks. Chapter 3 discuss the work we did in this directions which has been published in [155, 158].

The outlooks of this work are discussed in the Conclusions of the dissertation. To support the classification of allosteric architectures discussed in Chapter 1 and tested empirically in Chapter 2, further tests and possible experimental works are discussed. Furthermore, to seek for an improvement of the inference of fitness landscape we argue that different methods may be useful. We discuss a new direction that aims to develop a theory for identifying the hidden variables of allostery in the minimal in-silico model. Indeed, we show that the mechanical energy can be rewritten as function of few variables that are found to be crucial for a proper functioning. This idea goes in the direction of a fitness that is a nonlinear function of few variables that are linearly dependent [136]. Two roads open. Does the inference with an ad hoc non linear function of the energy provide better information than the usual DCA inference? Is the inference able to capture the linear variables instead? This connects with old ideas in deep learning where identifying the relevant hidden variables is deemed crucial for learning, as discussed recently in [159].

1 Mechanics of allosteric materials and proteins

This chapter is organized as follows. Section 1.1 serves as an introduction to the minimal models developed by other research groups aiming to understand the underlying physical relationship between motion, structure and function in allosteric proteins. The limitations of these works to answer the questions raised in the Introduction are discussed, leading to the research performed in this dissertation. The results presented in the following sections are published; Sec. 1.3 is based on [155], Sec. 1.4 on [156] and Sec. 1.5 on [157].

1.1 Minimal models connecting structure, motion and function in allosteric proteins

The approach used in this chapter studies the protein at a coarse-grained level, following the works on lattice proteins [139, 141] and elastic networks models [50–52]. Forces between amino acids depend on several features like charge, hydrophobicity, polarity and can be modeled precisely in molecular dynamics simulations. Minimal models, however, can also display rich features even though utterly approximating the interactions between amino acids. The advantage is that they usually are easier to handle than atomistic simulations. Indeed, they can be applied more systematically in situations where they are found to provide sensible predictions, thus compensating the naivety of the model construction.

The main example of these models is Elastic Network Models (ENMs). While they assume the force between two amino acids to simply be an harmonic interaction, they still reproduce protein conformational changes upon ligand binding, like allosteric transitions. A number of studies on diverse proteins that undergo ligand-induced conformational changes have shown that the resulting domain movements are dominated by a few — and often just one — normal modes [50, 51, 56, 57, 160–167]. Fig. 1.1 shows that one mode alone indeed reproduces very closely the conformational change of the LAO binding protein. The same results hold for other cases.

The success of ENM in the prediction of conformational changes provided a new perspective

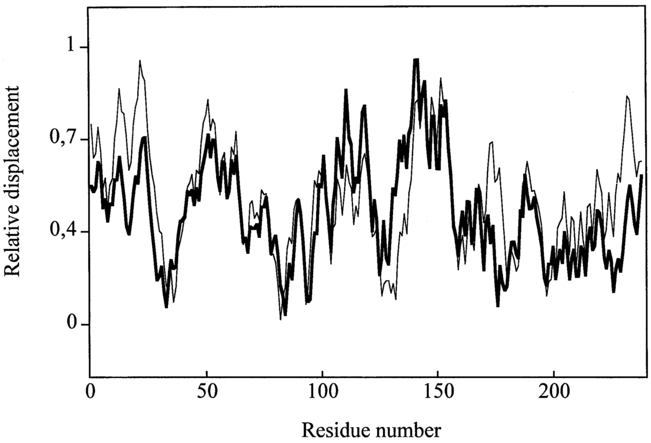

Figure 1.1 – Figure adapted from [56]. The conformational change of the LAO binding protein is shown together with the low energy normal mode that overlaps the most with it as a function of the residue number. The conformational change is computed among the open (PDB identifier 2LAO) and closed (PDB identifier 1LST) crystallographic structures and is shown with a thick line. The thin line shows the displacement of the atoms along the corresponding normal mode direction.

where to study the connection between structure and protein motion and function. Which is the underlying organisation of proteins that allows for this normal mode to appear? Given a high resolution conformation, is it possible to detect which are the amino acids that are involved in the conformational change? This view is purely structural and does not consider the contributions of conformational fluctuations of proteins. What is the role of entropy in allosteric transitions?

The following sections address these questions with the use of minimal models each characterised by a different level of coarse-graining, from networks built upon the high resolution structure to more abstracted toy models, as sketched in Fig. 1.2.

1.1.1 Rationale for the presence of a soft mode in allosteric proteins

The main feature connecting allostery and mechanics is the presence of a soft mode of deformation in several allosteric macromolecules that coarse-grained models like ENM are able to capture. Soft is used as a synonym for low energy, thus it is easy to couple to a soft mode upon a perturbation in the structure given the resulting small increase in energy. From the description of functional conformational changes by a single or few low energy modes, it can be argued that allosteric properties are strictly dependent on the pre-existence of global soft modes in the unliganded protein structure that are typically over-damped. Their revelance is also supported by the observation that these soft modes are evolutionary conserved in the same proteins evolved in different species [171]. But why is the mode recruited for the allosteric interaction low energy (long wavelength) and not high energy (short wavelength)? As discussed in Refs. [113, 172], many proteins are characterised by inhomogeneities in their structure that tend to localise normal modes within a few wavelengths, according to Anderson

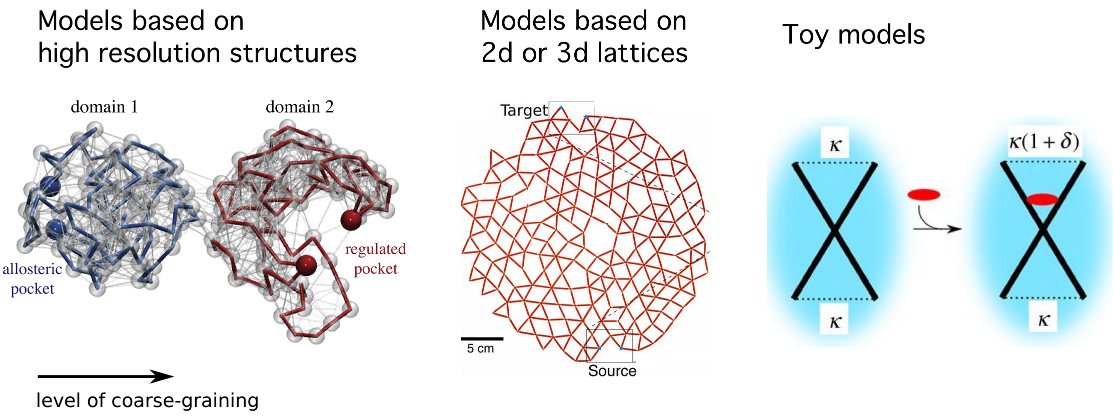

Figure 1.2 – Figures adapted from [168] (left) [169] (middle) and [170] (right). Different levels of coarse-graining in minimal models are summarised in this figure. On the *left*, an elastic network is built from two randomly folded three-dimensional chains and evolved to display long-range communication. The figure is taken from [168]. This method can be used on high-resolution structures to monitor the conformational dynamics. In the *middle*, a more coarse-grained model of evolution of long-range communication is shown where a two-dimensional triangular lattice is used as starting point of the design algorithm, taken from [169]. Finally, on the *right*, a toy model illustrate how it is possible to have allosteric communication driven by entropy without structual change by considering two rods joint in the middle and connected by springs of constant κ that increase by δ upon ligand binding (red circle), taken from [170].

localisation [173, 174]. Hence, a long wavelength mode involving a large amount of the structure is necessary to explain the non local allosteric effects, since a short wavelength mode is generally localised within a lengthscale smaller than the distance between the allosteric and active site, which is usually of order of the total diameter of the protein.

The idea of designing an elastic network with desired properties started with the work of Togashi and Mikhailov [175] who looked for modelling a macromolecule able to undergo a large-amplitude motion, like a hinge, upon ligand binding. The network is built from the three-dimensional structure of a protein in the inactive state where each node is an atom or residue, as in the left panel of Fig. 1.2. The evolution of the resulting elastic network is set by the requirement of displaying a well separated mode at low energy. Hence, this mode is by construction the dominant contribution to the long-time relaxation in the linear regime. Designed networks are found to display a precise structural organisation. They are composed of rigid parts linked by soft joints, as common when the mechanics is characterised by one soft mode responsible for the collective motion, usually called *mechanism*. The terminology comes from mechanical engineering where the design of a material that displays a mechanism is an active research topic and the appearance of less coordinated, softer regions in the evolved materials is found to be crucial for design [176]. However, in the work of Togashi and Mikhailov the soft mode is selected to be present. Thus, its presence does not stem from an evolution scheme imposing functional requirements, like allosteric communication.

1.1.2 Prediction of residues involved in the allosteric response

In an article by Amor, Barahona and collaborators [177] there is no concept of designing a network to be allosteric, the network resulting from high-resolution structures is taken as input. A graph representation of a protein is introduced, where nodes are atoms and weighted edges are both covalent and non covalent bonds. The goal is to reveal which pathways are strongly coupled to the active site, without any a-priori information on the location of functional sites.

The way fluctuations of edge weights propagate into the network is modeled with a diffusion equation on the graph where the weights are set by the interaction stiffnesses defined by detailed atomic potentials. The method is able to predict which are the sites relevant for allostery in a set of 103 proteins. However, it does not contribute to a physical understanding of the structural reasons for which allosteric residues can be detected. The study of a network with an elastic coupling between the nodes would help to advance in this direction.

1.1.3 A model of action at a distance without structural change

McLeish and collaborators studied in several works [113, 170, 178–180] the mechanisms behind communication at a distance when there is no evident structural change between the inactive and active state, following the seminal work of Cooper and Dryden where a cooperative allosteric free energy composed of only entropic terms is proposed [27]. Hence, in this context, allosteric communication emerges solely from the changes in entropy due to thermal fluctuations. Indeed, allosteric communication can emerge when ligand binding changes the amplitude of thermal fluctuations around the mean structure of the protein — described by a global soft mode, thus changing the entropy and building a free energy like the one of Cooper and Dryden.

Which are the mechanisms that can lead to long-range coupling and large entropic terms? A toy-model of fluctation allostery, the allosteron, helps the investigations of fluctuations allostery. It consists of two rods joint in one point — like in a scissor — with internal side chains that fluctuate depending on binding events, as shown in the right panel of Fig. 1.2. It is known that a slow global mode in the system is able to couple two distant sites, what about the role of fast local modes like the ones of the side chains? The allosteron model although it is one of the most coarse-grained, is able to make non-trivial predictions about fluctuations-based allostery. It shows that local modes can couple to global ones to amplify the entropic allosteric effect in agreement with the observation in the met repressor [179]. Furthermore, it is possible to obtain pathways by which fluctuation allostery affects self-assembly of protein complexes [170] and also identifies the physical origin of negative cooperativity in a particular protein [181].

1.1.4 Discussion

This section overviewed minimal models describing long-range communication in proteins, however several questions are left unanswered, notably the physical origin of the soft mode. In this dissertation, we propose a model where soft modes appear as a result of a specific functional task. This framework allows to

- study the relationship between structural organisation and function;
- classify architectures — indeed, the structural organisation results into different architectures depending on the evolved function;
- understand whether the systems are poised to functional optimality;
- investigate the consequences of elastic nonlinearities on allosteric behaviour.

The strategy we adopt is the one of an in-silico model for the evolution of allosteric interactions in elastic networks. A similar framework is used in mechanical engineering where topology optimisation designs materials that exhibit a specific function upon perturbation. There, the topological optimisation of the structure is usually performed via finite elements analysis according to a chosen objective function [176, 182, 183].

Related research

Other works tackling these issues have been published in parallel or after the research presented in this dissertation and they address related, but complementary questions to the ones just raised.

Reproduce the allosteric transition in proteins. The design of optimised long-range communication between two distant sites in elastic network models is also studied by Flechsig in [168]. A network composed of two domains is randomly folded and evolved by changing the position of one bead at a time and measuring the transmission of the response between the allosteric and active site. The dynamical equation obeyed by the beads is the Newton's equation in the overdamped limit, so that no assumption of linear elasticity is needed. The result is a model able to reproduce the allosteric transition of the protein myosin V — and possibly generalisable to other proteins as well — when the initial configuration of the network evolution is set to the high resolution structure of the inactive state.

Mechanical complexity of long-range functionality Rocks, Nagel, Liu and collaborators explored how difficult is to have a long-range communication in a two-dimensional elastic network [169]. The task required is to display a given strain at the active site when another strain is imposed at the opposite allosteric site. The evolution is set by pruning one bond at a time and measuring the strain at the active site. It is found that only 1% of the bonds need to be removed for the network to be designed. Two- and three-dimensional networks evolved with this scheme have been fabricated. The framework where multiple tasks are satisfied is discussed in another work [184].

Evolution of proteins in a percolation-like model. Tlusty and collaborators approached the problem of a mechanical model of protein evolution from a different perspective [185, 186]. The beads in the elastic network are amino acids that can fall into two species — either hydrophobic or polar — whose kind and interaction strength are encoded by binary codons in the gene. This structure, gene⇒amino acid network⇒evolution⇒function, is useful to build a physical model of the genotype-to-phenotype map. The evolution acts by changing the position of the beads and monitor the response in the network upon local perturbation until the network achieves the target response that is set to be a large-scale motion. The transition to designed network is defined by the emergence of weakly-connected regions and of a soft mode related to function. However, the soft mode and the flexible path are inserted ad hoc and do not result from underlying physical principles. Indeed, the beads of the network are not elastically coupled and the importance of weakly-connected regions in conducting strain and enabling such mode is not discussed.

1.1.5 Outline

The following section, Sec. 1.2, serves to define the in-silico model as a network of harmonic springs and its evolution scheme. To study the relationship between function and architecture, the evolution is performed by optimising two different allosteric tasks. One optimises for the transmission of strain from the allosteric to the active site and is discussed in Sec. 1.3. The other, optimises the cooperative binding energy between the two sites and is reviewed in Sec. 1.4. The results show that two scenarios arise and deeply different mechanism are used to implement the two allosteric functions. The optimisation of the geometric fitness leads to an exotic architecture where the response is amplified from the allosteric to the active site (Subsec. 1.3.3) thanks to the presence of special modes that act as a lever (Subsec. 1.3.4). On the other hand, cooperative architectures show a variety of designs among which shear and hinge (Subsec. 1.4.2), already discussed in the introduction to proteins. Subsec. 1.4.6 shows that fluctuations-driven cooperativity can happen in the in-silico model as well, once an extendend low-energy mode is responsible for function.

Despite the differences, all designs can be classified (Subsec. 1.4.4) and are found to work thanks to the presence of a soft mode, as the one identifed in proteins by ENMs. Furthermore, the scaling of the stiffness of this mode as function of the size of the network is tuned to display optimal cooperative binding (Subsec.1.4.5). Indeed, it is found that for proper function, the network must evolve a functional elastic mode that is softer as its size increases. Sec. 1.5 deals with the effect of nonlinear elasticity on cooperative allostery, connecting with the two-state population shift scenario of the MWC model, discussed in the Introduction. The resulting cooperative networks can switch from two equilibrium states and show a soft mode that connects the two states and obeys the same scaling of stifness than the linear case to properly function.

1.2 A model for evolution of action at a distance in proteins

In this section, we look at the evolution of allosteric communication in elastic materials, where we consider a network of springs instead of a protein — this lattice model has been previously used to characterise the behaviour of covalent glasses [187, 188]. Allosteric behaviour in an elastic material is already a surprising property. Indeed, the elastic response upon perturbation decays as a powerlaw over distance in a continuous elastic medium and a long-range effects cannot appear without a properly structured material. Which are the architectures that allow for action at a distance? Is evolution of these architecture hard or easy? How does the architecture depend on the chosen task? Is it possible to achieve a classification of optimal designs?

In the following, we address these questions in detail by studying in-silico evolution of allosteric communication in springs networks optimising for two different tasks. One optimises for the display of a given strain at the active site upon perturbation at the allosteric site and the other maximises for the cooperative energy between the allosteric and active site, like the cooperative binding of oxygen in haemoglobin discussed in the introduction. The two settings give rise to two drastically different scenarios that are discussed in details in the following sections. The insights provided by these results are tested in proteins as discussed in the next chapter, Chap. 2.

1.2.1 Definition of the model

As shown in Fig. 1.3, the considered network is composed of $N = L^d$ nodes organised in a d-dimensional lattice, where d can be either two or three. Harmonic springs[1] of unit stiffness k connect some of the adjacent nodes where the number N_s is fixed according to the the coordination $z = 2N_s/N$ counting the average number of links per node. These N_s springs are initially randomly distributed in the network and then swapped according to the evolution discussed below [2]. The allosteric and active sites are located at two opposite sides and each consists of $n = 4$ nodes. The geometry of the network can be chosen with both periodic and open boundary conditions on two of the four sides, corresponding to a cylinder and to a sheet.

---

[1] Each node of the triangular lattice is distorted by adding a random displacement in order to avoid straight lines in the network that could cause unphysical localised modes orthogonal to them. Here, harmonic springs of stiffness k are considered, non harmonic springs will be discussed later in Subsec. 1.5.

[2] An extension of the model is briefly discussed in Subsec. 3.5.2, where the presence of the springs is dictated by the interactions acting between the two adjacent nodes, which can each have N_a degrees of freedom, building a toy model describing the variability and role of amino acids in proteins.

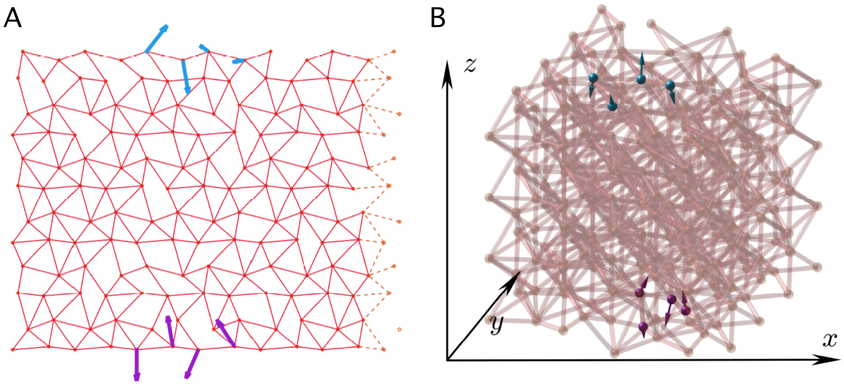

Figure 1.3 – (A) The triangular lattice in $d = 2$ is shown for a system of size $L = 20$ and $N = L^2$ nodes where N_s springs are occupied with coordination $z = 5$. Periodic boundary conditions are chosen along the horizontal axis, mimicking a cylindrical geometry. Springs crossing the periodic boundary are shown in dashed lines, and are not present when open boundary conditions are used. (B) The extension to $d = 3$ is a face-centered cubic lattice with open boundaries. Its linear size is $L = 12$ and $z = 8.4$. In both (A) and (B), occupied links that display a spring of stiffness unity are indicated by lines. The displacement imposed at the allosteric site — composed of four nodes — is denoted by violet arrows, while the target response that can asked to be displayed at the active site upon perturbation is shown in blue arrows.

The occupancy configuration of the network is described by a connection vector

$$|\sigma\rangle = \begin{cases} \sigma_\alpha = 1 & \text{if the link } \alpha \text{ is occupied} \\ \sigma_\alpha = 0 & \text{otherwise} \end{cases} \tag{1.1}$$

whose size is the number of all possible links. The notation is summarised in Tab. 1.1 and sketched in Fig. 3 in the Introduction.

Notation	Definition
10	one ligand bound at the allosteric site
01	one ligand bound at the active site
11	one ligand bound at both the allosteric and active site
Al	allosteric site
Ac	active site
r	indicated the $N-4$ nodes excluding the allosteric site

Table 1.1 – Definition of notation as introduced in Fig. 3.

The effect of ligand binding is modelled as imposing a displacement, $d\mathbf{R}^{10}$, on the allosteric site that is defined to be a subset of four nodes, as shown in purple arrows in Fig. 1.3. Indeed, when a ligand binds to a protein, the area around the binding site rearranges to accomodate the ligand, therein inducing a local strain[3]. The allosteric site is located on a side of the network with free boundaries, see purple arrows in Fig. 1.3. After relaxing the elastic energy of the

---

[3] This description of binding assumes the rigidity of ligands, which thus are not deformed in the process. The

entire network, the response of all other $N-4$ nodes to such perturbation $d\mathbf{R}_r^{10}(|\sigma\rangle)$ can be found. It depends on the connection vector $|\sigma\rangle$ and one example is denoted by black arrows in Fig. 1.3A. The energy cost associated with the binding event is

$$E^{10}(\sigma) = \frac{1}{2}\langle d\mathbf{R}_r^{10}|\mathcal{M}|d\mathbf{R}_r^{10}\rangle, \tag{1.2}$$

where $\mathcal{M}$ is the stiffness matrix of the network of dimension $Nd \times Nd$ and it depends on the considered network via the occupancy vector $|\sigma\rangle$. The same procedure is used to model the binding of another ligand at the active site, shown in blue in Fig. 1.3, defining $E^{01}(\sigma)$. If the ligands are bound simultaneously, the resulting joint binding energy is $E^{11}(\sigma)$. The definition of the procedure is detailed in Appendix A.2. The specificity of binding is discussed in Subsec. 1.4.3.

In the case of harmonic springs, $d\mathbf{R}_r^{10}(|\sigma\rangle)$ can be computed via linear response, see Appendix A.2 and via energy minimisation when non harmonic springs are involved, see Subsec. 1.5.

Initially, the N_s springs are randomly distributed in the network and the response decays as a power law at the opposite side, where the four nodes active site is located, showing no long-range function. Indeed, the network is made functional by the optimisation of a chosen fitness, $\mathcal{F}(|\sigma\rangle)$, reflecting the ability of the network to perform the desired function involving a specific outcome at a distal active site. Two different fitness functions are considered, one — called *geometric* — optimising for the display of a given strain at the active site upon perturbation at the allosteric site and the other — *cooperative* — maximising for the cooperative energy between the allosteric and active site. The two settings are discussed in details in the following sections.

The optimisation algorithm works following a Metropolis algorithm. The probability of a link — chosen at random — to change its state is

$$P(|\sigma\rangle \to |\sigma'\rangle) = \min\left\{1, \exp\left(\frac{\mathcal{F}(|\sigma'\rangle) - \mathcal{F}(|\sigma\rangle)}{T_e}\right)\right\}, \tag{1.3}$$

where the two configurations $|\sigma\rangle$ and $|\sigma'\rangle$ differ only by one spring and T_e, the evolution temperature, tunes the stochasticity with $T_e = \infty$ corresponding to undesigned, random networks. The evolution algorithm acts on all springs apart the ones connecting the allosteric and active site that are always present.

During evolution, the number of springs per node averaged over N — the average coordination z defined as $z = 2N_s/N$ — is kept fixed to a desired value defining the average rigidity of the network. This implies that each evolution step consist in a swap of the occupancy of two links, keeping the number of occupied links fixed. The frequency of springs f can be found as $f = \frac{zL^2}{N_{ps}}$, where N_{ps} is the number of all possible springs. Maxwell derived a bound to

---

results are expected to hold true qualitatively as long as the ligands are not significantly softer than the protein itself.

be satisfied by the coordination for the system to be rigid [189]. The condition $Nd - N_s > 0$ implies[4] a rigidity bound for the average coordination z, $z > 2d$. The value $z_c = 2d$ corresponds to the transition between a rigid or floppy network and defines the *isostatic* regime, which is responsible for non trivial features in the response that will be discussed in Subsec. 1.3.4. Since proteins fold in a stable structure, it is reasonable to fix the average coordination of the network in evolution to $z = 5$ for $d = 2$. Anyhow, the behaviour of the maximum fitness reached in the case of the geometric task is studied as function of coordination in Fig. 1.5C. The architecture found in the case of average coordination fixed to a lower value, $z = 3$, is discussed in Appendix A.3.

1.3 Geometric function

The first task to be discussed is a very simple example of allosteric communication. The network is optimised to display a given strain when a perturbation is applied at the opposite side, as shown in Fig. 1.4 where the perturbation at the allosteric site is represented with purple arrows whereas the target response at the active site with blue arrows. The *geometric* fitness function describing the matching of the target strain with the strain induced by the perturbation at the active site reads

$$\mathscr{F}(|\sigma\rangle) \equiv -E(|\sigma\rangle) \equiv -\min_{|\mathbf{U}\rangle} \sqrt{\sum_{i\in\{01\}} \left(d\mathbf{R}(|\sigma\rangle)_i^r - d\mathbf{R}_i^{01} - \mathbf{U}_i\right)^2} \tag{1.4}$$

$$= -\sqrt{\sum_{i\in\{01\}} (d\mathbf{R}_i^r - d\mathbf{R}_i^{01})^2 - \sum_{i\in\{01\}} (d\mathbf{R}_i^r - d\mathbf{R}_i^{01})(d\mathbf{R}_{i+1}^r - d\mathbf{R}_{i+1}^{01})}\,, \tag{1.5}$$

where the summation is done over the nodes in the active site, {01}, and $|\mathbf{U}\rangle$ represents the vector of global translations ($|\mathbf{U}_r\rangle = c$ with c constant) and rotations of the displacement at the active site, $d\mathbf{R}_i^{01}$. The minimisation over $|\mathbf{U}\rangle$ ensures that what is optimised at the active site is not the absolute displacement, but the strain, since it is the deformation of the structure at the active site that is physically responsible for the change in functionality, not the displacement itself. Indeed, when a ligand binds to a protein, the area around the binding sites rearranges to accomodate the ligand, therein inducing a strain. It is easy to see that if the displacement at the active site is a global translation, $d\mathbf{R}_i^r - d\mathbf{R}_i^{01} = c$, Eq. 1.5 consistently gives back zero energy. For the case of a global rotation, the calculation is more involved.

1.3.1 Evolution

For each chosen coordination number z and evolution temperature T_e, we consider 20 Monte Carlo sampling series with 10^5 Monte Carlo steps in each, and discard the first half of the

---

[4] In a mean-field-like argument where the network consists of disconnected parts, the number of zero modes is equal to the number of degrees of freedom, $N_{dof} = Nd$, substracted to the number of constraints $N_c = N_s$, where d is the space dimension. If the fraction of zero modes — once the trivial modes of translation and rotation are taken out — is greater than zero, the system is rigid. Thus, the bound $Nd - N_s > 0$ for rigidity.

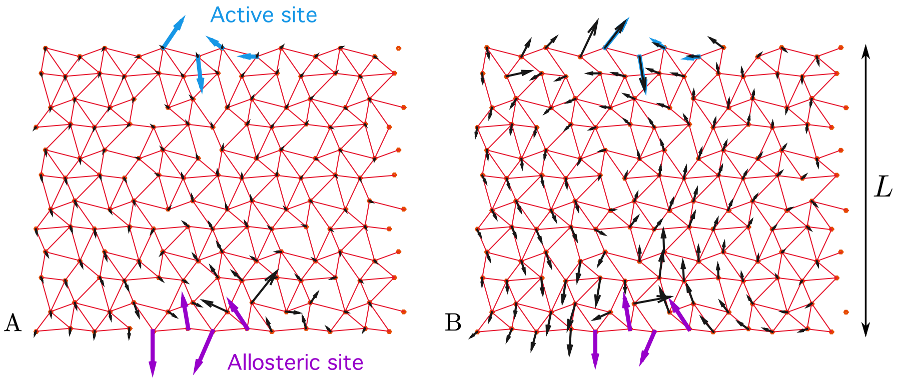

Figure 1.4 – An undesigned (A) and designed (B) network are shown for a system of size $L = 12$ and $N = L^2$ nodes where N_s springs are occupied. The displacement imposed at the allosteric site is denoted by violet arrows, while the linear response $d\mathbf{R}_r^{10}(|\sigma\rangle)$ of the nodes upon such perturbation is shown as black arrows. In the undesigned network, where the springs are randomly placed, the response decays as expected and does not match the required shape at the active site (A). After evolution, the linear response matches with the required response, represented with blue arrows, at the active site, showing the emergence of functionality in the designed network (B).

series to eliminate transient effects, resulting in a sample of 10^6 configurations. The evolution algorithm is able to easily find functional networks when the temperature, T_e, is sufficiently lowered, as shown in Fig. 1.5A where the average fitness function over different configurations, $\langle\mathscr{F}\rangle$, is shown as function of T_e. The undesigned network corresponding to a random configuration of springs and the designed functional network are shown in Fig. 1.4A and B, respectively.

1.3.2 Thermodynamics

The results in Fig. 1.5 involve four different system sizes ($L = 4, 6, 8, 10, 12$ and $L = 20$ for Fig. 1.5B) to study the thermodynamics of our system. Fig. 1.5A shows that the transition in the average fitness becomes sharper as the size of the network increases, hinting for a transition happening at the value of T_e corresponding to the appearance of the desired functionality in the configurations. To test this hypothesis, the specific heat $c(T_e) = dE(T_e)/dT_e$ is computed and shown in Fig. 1.5B where its peak becomes sharper as the size increases, indeed supporting a transition from random to functional networks. Drawing a parallel from statistical mechanics, the behaviour found for the specific heat suggests that the allosteric transition in these networks is a collective process where the springs cooperatively arrange to build the functional configuration, even if data from a much larger size are needed to test the validity in the thermodynamic limit. Moreover, within this setting, it is also possible to monitor how the fitness landscape is explored by looking at whether evolution gets stuck in a local maximum of bad fitness, which would be a typical situation for systems like amorphous solids and glasses that display a rugged energy landscape.

Fig. 1.5C shows that the fitness reached with an evolution at $T_e = 0$ — the maximum averaged

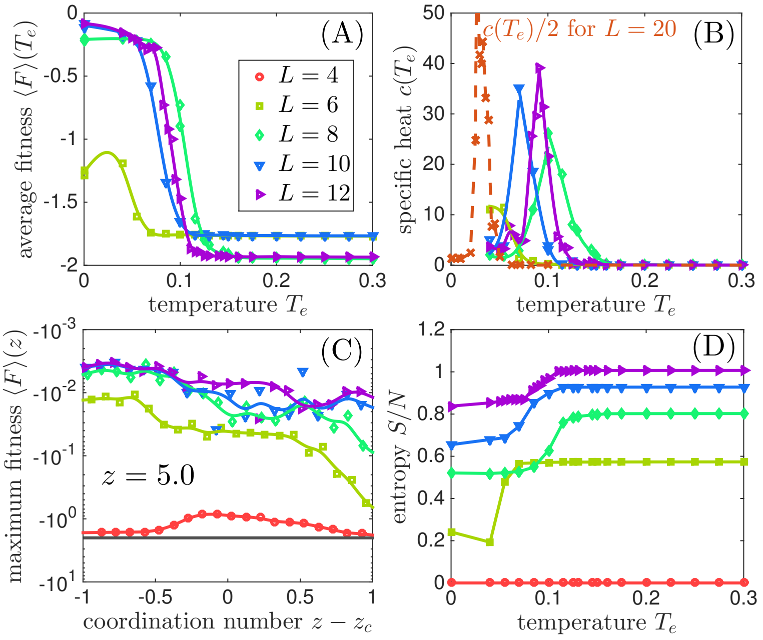

Figure 1.5 – (A) Fitness averaged over 10^6 configurations shown as function of the evolution temperature, T_e. A jump in fitness is found for large systems around $T_c \approx 0.09$. (B) The specific heat as function of T_e shows a peak that increases as the size L increases, supporting the existence of a transition at a certain temperature T_c. (C) Averaged fitness computed with a pure gradient ascent result in an exploration of local maxima in the fitness landscape and it is shown as function of coordination in a log-linear plot. The black line indicates the fitness for random networks when there is no mechanical response at the active site. (D). The entropy density as function temperature is shown.

fitness corresponding to a pure gradient ascent with zero stochasticity — is 200-fold higher than the one of random networks, which is of order one, proving that indeed the algorithm does not get stuck in a maximum with a low value of fitness. For the largest sizes, the algorithm shows consistent performance as function of coordination, since the maximum average fitness is probed in the range $z \in [3,5]$ and fluctuates around high fitness values. The number of these allosteric networks can be estimated as $e^{S(T_e)}$ via the entropy $dS = C(T_e)dT_e/T_e$. At the transition, some degrees of freedom have to be tuned to have a designed network, implying a decrease in entropy, as consistently shown in Fig. 1.5D. The number of $L = 12$ networks at temperature $T_e = 0.05$, well inside the allosteric phase, is large $e^{S(T_e)} \approx 10^{53}$, but the probability to find such a network by chance is low, $e^{S(T_e)-S(\infty)} \approx 10^{-10}$.

1.3.3 Architecture: trumpet design

The previous results on thermodynamics proved that it is possible to find networks with allosteric behaviour via evolution in the geometric fitness landscape. Now, we would like to see whether the 10^6 evolved allosteric networks display some similar features in their structure. The coordination number monitors the local rigidity of the network and its mean over the configurations can be used as a measure to see whether a common design is shared between the networks.

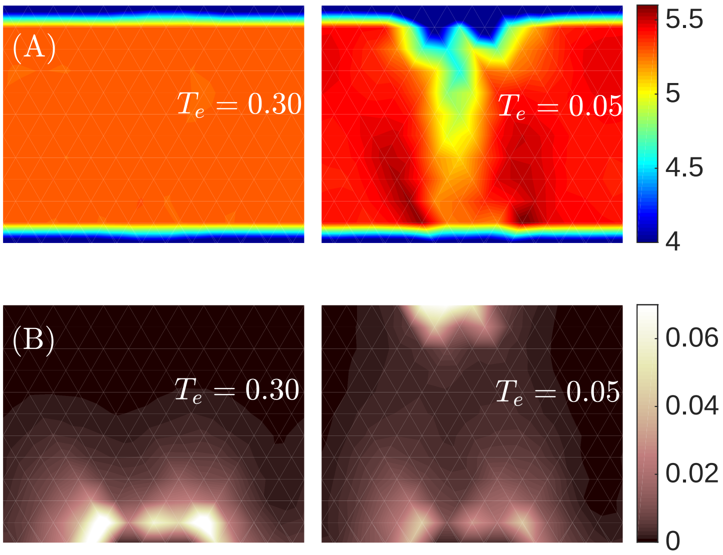

Figure 1.6 – (A) Mean coordination maps and (B) mean-squared magnitude of the normalised response at different nodes for $L = 12$ networks evolved at $T_e = 0.30$ (left) and $T_e = 0.05$ (right). For functional networks (right), a trumpet in the mean coordination connecting the allosteric and active site appears and the associated response is non-monotonic, being amplified at the active site.

Before the transition, the networks do not display any particular structure, in consistence with the absence of design, as shown by the the homogeneity in the mean coordination map of networks evolved at $T_e = 0.30$ in Fig. 1.6A-left for $L = 12$. For networks evolved well below the transition, at $T = 0.05$, a non trivial structure appears in the mean coordination: a softer trumpet-like region is carved into a more rigid matrix, with coordination lower than the fixed average $z = 5$ and monotonically decreasing from the allosteric to the active site where an isostatic area with mean coordination close to $z_c = 4$ appears, see Fig. 1.6A-right.

How does the response elicited in the network upon perturbation at the allosteric site look like in this *trumpet design*? The behaviour of the response can be monitored by looking at the mean-squared magnitude of the normalised response at different nodes i, $\langle |d\mathbf{R}_i^r|^2 / \sum_i |d\mathbf{R}_i^r|^2 \rangle$ averaged over different initial conditions. Fig. 1.6B-right shows that the magnitude of the response has a non-monotonic behaviour in the trumpet design: it is large at the allosteric site, it vanishes in the middle and appears again amplified at the active site, where it needed to be, contrarily to undesigned network where the magnitude is large only at the allosteric site, Fig. 1.6B-left. The amplification is striking since the response is amplified by a factor five, reminiscent of levers that amplify motion by large amounts. How is the trumpet design providing the right architecture for such an amplification? Is it using a mechanism similar to a lever? The answer lies in the special features emerging from isostaticity, which are discussed in the next paragraph.

1.3.4 Isostatic materials act as levers

The relevance of isostaticity needs to be discussed in order to understand which is the mechanism underlying the non monotonic behaviour of the response in the designed network, as shown in Fig. 1.6B-right.

Isostatic materials are known to have exotic features since they live on the verge of instability: if a link is broken, the system looses its mechanical stability and a zero energy mode of deformation appears. In Fig. 1.7A an isostatic network of linear size L is cut along the vertical axis in the middle, thus removing order of L^{d-1} constraints and generating a number of zero modes of the same order. Some of these zero modes will be localised on the nodes of one of the two boundaries, however others will propagate through the network [191–193]. The zero modes that have an amplitude that decreases or increases exponentially with distance as they penetrate in the bulk of the system — instead of displaying the usual power law decay — are called *edge modes*. In the case of exponential increase, they act as a powerful lever that amplifies motion exponentially towards free boundaries.

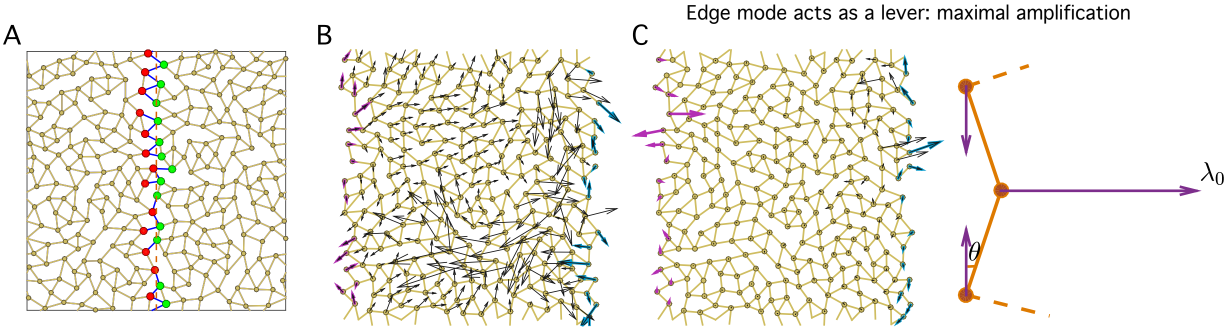

Figure 1.7 – Figures adapted from [190]. (A) An isostatic network in two dimensions ($z = 4$) has periodic boundary conditions in the horizontal direction and it is cut in the middle. The cut bonds are shown in blue and the nodes are coloured according to the boundary they belong to, either left (red) or right (green). (B-C) Response (cyan arrows on the right boundary and black arrows in the bulk) to an imposed displacement on the left boundary (magenta arrows) in an isostatic network with free boundary conditions in the horizontal axis and a periodic boundary condition in the vertical one. Two examples of zero modes are displayed in (B) and (C). The zero mode in (C) shows a significant amplification of a factor 20000, similar to the amplification of a lever where an applied displacement of unitary magnitude gets amplified by λ_0.

The appearance of edge modes was explained theoretically for some crystalline solids [194] and also for disordered elastic networks [190], allowing to use them in this framework, as well. In the linear regime, the effect of a displacement imposed at one of the boundaries[5] to the opposite one is seen as a transmission problem between slabs of width l_c. The spectrum of the transmission matrix contains a characterisation of displacement propagation. Through its manipulation, it can be proved that the maximal amplification of the response upon perturbation at one of the free boundaries is of order L^{L/l_c}, where the slab width l_c is empirically fixed to

---

[5] The boundary needs to be *squeezed*, i.e. all the boundary needs to move.

$l_c = 4$ [190]. Fig. 1.7B and Fig. 1.7C show two examples of zero modes where the response at the right boundary upon an imposed displacement on the left boundary is shown as black arrows. Fig. 1.7C displays an edge mode where the response is amplified by a factor ≈ 20000 [190].

In the light of these findings, the trumpet design of Fig. 1.6A-right can be identified to be responsible for the amplification of the response at the active site, where the designed network develops a patch of coordination close to $z_c = 4$, the isostatic value, as sketched in Fig. 1.8A.

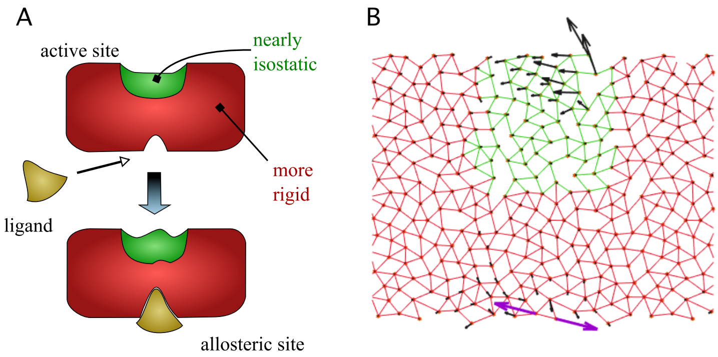

Figure 1.8 – (A) Sketch of the allosteric behaviour found when the geometric fitness function Eq. 1.5 is optimised. The system can be viewed as a nearly isostatic patch around the active site embedded in a rigid matrix. When the ligand imposes a given strain at the allosteric site, the desired strain is obtained at the active site. (B) Toy network with an isostatic patch around the active site embedded in a rigid matrix showing that, when a dipole perturbation is applied (purple arrows), the response (black arrows) decays in the bulk, but get indeed amplified at the active site.

The observation of a less constrained, floppy region connecting the allosteric to the active site is not surprising since it serves to enable the propagation of the response over a longer range. However, this is not what is observed in geometric networks where the response is amplified from the two sites. The coordination is asymmetric, with a more coordinated region around the allosteric site and an isostatic patch at the active site, which is crucial for amplification. To test this proposal, Fig. 1.8B shows a toy network built to have an isostatic patch around the active site ($z = 4$) embedded in a rigid matrix ($z = 5$). When a dipole pertubation is applied at the allosteric site, the response in the bulk decays, but a small displacement at the boundary of the isostatic patch is sufficient to excite an edge mode that amplifies the response at the other boundary where the active site is located, consistent with the theoretical framework [190].

1.3.5 Thermal fluctuations

Thermally-induced motion can be characterised by B-factors [55] that measure the flexibility of the structure as discussed in Appendix A.2.3. Fig. 1.9 shows the B-factor map for the configurations at high and low evolution temperature. Thermal motion appears to be larger in the trumpet region, particularly in the marginally-connected region of the active site. There, B-factors are about 20 times larger than in the other boundary nodes of the system,

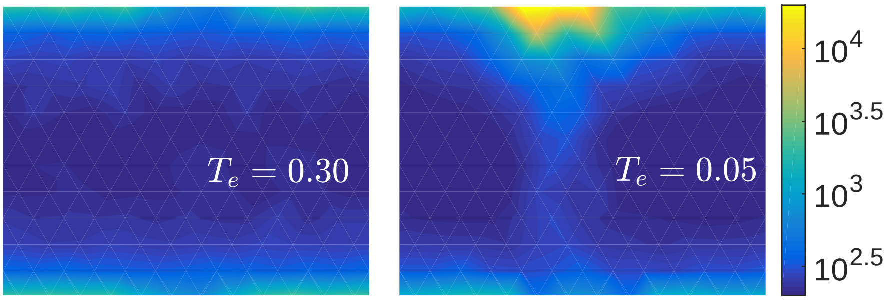

Figure 1.9 – Spatial distribution of B-factor for $T_e = 0.30$ (left) and $T_e = 0.05$ (right). The colorbar is in log scale.

corresponding to an amplitude of motion about four times larger. This high softness at the active site upon thermally-induced motion could be avoided by changing the fitness function and penalising thermal motion at the active site.

1.3.6 A tool to pick fruits

The novel edge mode lever architecture is studied for a network of linear size $L = 12$. Does the response can still get amplified when L increases? Fig. 1.10D shows that when $L = 20$ the same structure in the coordination map appears, with the only difference of a more loosely connected region nearby the allosteric site. The normalised magnitude of the allosteric response (Fig. 1.10B) also shows the same non monotonic behaviour as in Fig. 1.6B, but a larger

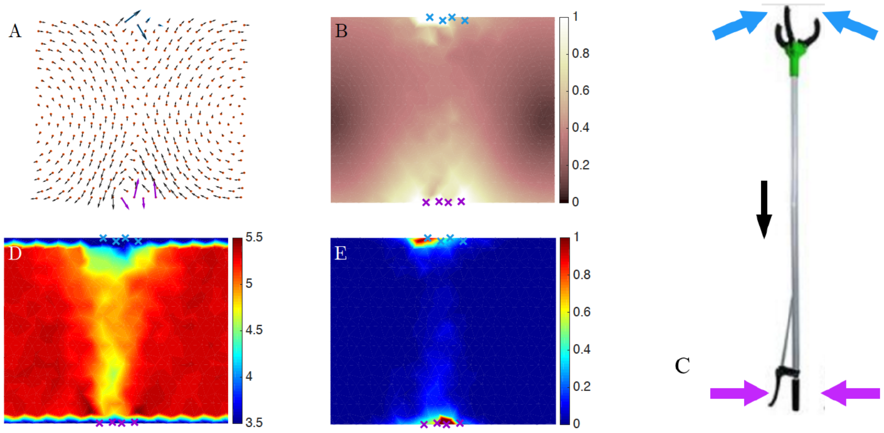

Figure 1.10 – (A) The average response induced by binding at the allosteric site is shown as black arrows. (B) Map of the average magnitude of response $\langle |d\mathbf{R}_i^r|^2 / \sum_i |d\mathbf{R}_i^r|^2 \rangle$ as shown in Fig. 1.6B. (C) A fruit picker illustrates the combination of edge mode lever and shear to perform the geometric task when the linear size increases. (D) Map of the average coordination number z. (E) Map of the average intensity of shear deformation, see Eq. 1.8 for its definition.

magnitude characterises the region around the allosteric site. An inspection of the allosteric response (Fig. 1.10A) and of the intensity of shear deformation of the response (Fig. 1.10E) proves that the region where the network displays a larger magnitude of response undergoes a shear motion. The combination between shear motion and edge mode lever assures that the response is transmitted at a distance such that the lever mechanism can get in action. The architecture resembles the design of a fruit-picker, where the shear exerted by pushing a button transmits to the other end of the stick regulating the opening of the claw, as sketched in Fig. 1.10C.

1.3.7 Conclusions

The evolution scheme easily finds networks that optimise the geometric function, where a given strain is elicited at the active site upon perturbation at the opposite allosteric site. A transition from undesigned to designed networks is found in the behaviour of the specific heat, suggesting that the emergence of allosteric behaviour is a collective process.

A shared architecture, the trumpet design, results from a statistical analysis of the designed networks. The mean coordination decreases from the allosteric to the active site where a nearly isostatic patch is present. This isostatic region is proposed to be responsible for the observed non-monotonic behaviour in the response — which is amplified at the active site — through the appearance of nearly zero energy modes that, once excited at one boundary of the isostatic matrix, grow exponentially towards the opposite boundary. Since the geometric function optimises strain only at the active site, the trumpet architecture is asymmetric: the elastic information cannot propagate from the active to the allosteric site, but it is greatly amplified in the other direction.

This architecture is novel. Indeed, the optimisation of strain revealed not to recover the usual designs found in proteins, like shear and hinge, reported in Fig. 5 in the Introduction. Fig. 5 also highlights that not all proteins conformational changes fall in the category of hinge and shear, as the PDZ domain shown in Fig. 5C whose functioning architecture is unclear. In Chapter 2, we argue that the PDZ domain may follow precisely this novel design showing an amplification of the response between its functional sites. Anyhow, the absence of shear and hinge designs motivates for looking at the in-silico evolution under the optimisation of another allosteric task, as discussed in the next section.

1.4 Cooperative function

Biology provides another allosteric task that is abundantly observed: cooperative binding. As discussed in the introduction, cooperative allosteric interactions consist in a coupling between two or more binding sites, as exemplified by haemoglobin where the binding of oxygen at one heam lowers the energy cost of binding at the others. The cooperation is encoded in the fitness function described for the MWC model. In the context of spring newtorks, the temperature is zero and the binding free energy reduces to the elastic energy contribution

$$\mathscr{F} \equiv \Delta\Delta E = E_{10} + E_{01} - E_{11}\,, \tag{1.6}$$

where E is the elastic energy associated to the displacement or force induced at the allosteric site (10), active site (01), or both simultaneously (11), see Fig. 3, where $E_{00} = 0$ since the network is at equilibrium and there is no pre-stress. Maximising Eq. 3.1 favours the binding of the substrate at the active site when the ligand has already been bound at the allosteric site. This is apparent by re-ordering the terms in the fitness as $\mathscr{F} = E_{10} - (E_{11} - E_{01})$, where in the optimisation of $\mathscr{F}$, E_{10} needs to be large and $(E_{11} - E_{01})$ small, implying that the amount of energy to bind at the active site when there is no ligand at the allosteric site is significantly larger than when it is present, favouring the binding of the substrate.

The form of the fitness function involves symmetrically both the allosteric and the active site, hence more symmetric architectures than the trumpet design discussed before are expected. To interpret the functional constraint imposed by the cooperative fitness, it is useful to consider the case of a weak coupling between the sites. The approximated fitness function reads[6]

$$\mathscr{F} \approx \langle \mathbf{F}^{01} | d\mathbf{R}^{10\to Ac} \rangle\,, \tag{1.7}$$

where $\mathbf{F}^{01}$ is the force exerted by the substrate when it binds at the active site and $d\mathbf{R}^{10\to Ac}$ is the displacement induced at the active site when the ligand binds at the allosteric site. Note that each field in Eq. 1.7 is of dimension nd, where $n = 4$ is the number of nodes in the active site and $d = 2$ the spatial dimension. Hence, maximising $\mathscr{F}$ leads to both a large and specific response at the active site, as for the geometric task, and a large force scale, requiring the

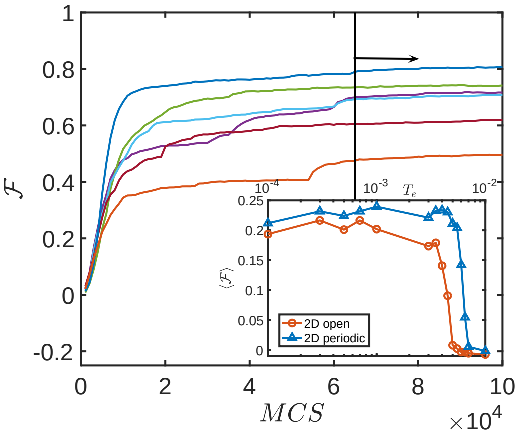

Figure 1.11 – Fitness function $\mathscr{F}$ as function of the number of steps. Some initial conditions display transient plateau in $\mathscr{F}$, suggesting an underlying rugged fitness landscape. Inset: the fitness averaged over the last 3.5×10^4 (see black vertical line) and 25 initial conditions $\langle\mathscr{F}\rangle$ is shown as function of the evolution temperature, T_e.

---

[6] See Sec. C.1 in the Appendix for the details of the calculation.

material to be rigid enough at the active site. This additional constraint is susceptible to make the evolution harder than for the geometric fitness function, as discussed in the following.

1.4.1 Evolution

The evolution scheme is run for 25 Monte Carlo sampling series with 10^5 steps each and at different values of T_e while keeping $z = 5$. To avoid transient effects, only the configurations from the last 3.5×10^4 steps in each series are kept for the analysis. Two different boundaries conditions are considered: open and periodic. The size considered is $L = 20$ unless stated otherwise.

The fitness function averaged over the number of steps and the 25 initial conditions, $\langle \mathscr{F} \rangle$, as function of T_e shows a transition from zero to finite fitness as the temperature is lowered, corresponding to the emergence of design, see the inset of Fig. 1.11 and Fig. 1.12 where the response in two networks equilibrated at low and large temperature is shown.

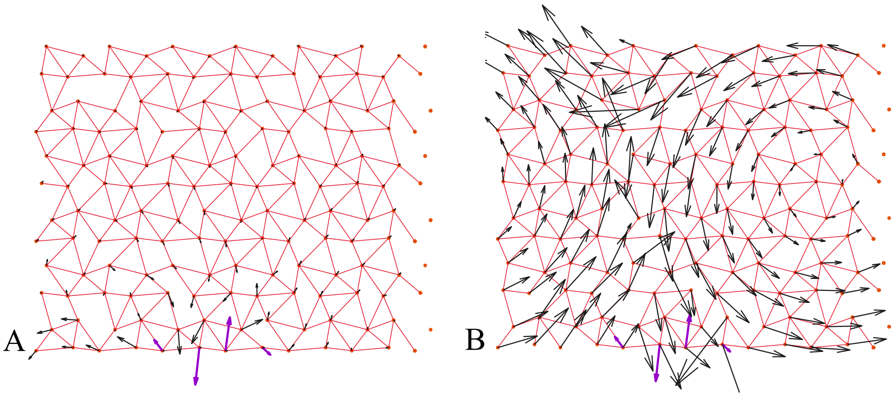

Figure 1.12 – An undesigned ($T_e = 10^{-1}$) (A) and designed ($T_e = 10^{-4}$) (B) network are shown for a system of size $L = 12$, with $N = L^2$ nodes and periodic boundary conditions. The displacement imposed at the allosteric site is denoted by violet arrows, while the linear response of the nodes upon such perturbation is shown as black arrows. In the undesigned network, where the springs are randomly placed, the response is not specific and decays in the bulk (A). After evolution, the designed network optimises the cooperative fitness by displaying a specific architecture, in this case a shear mechanism (B).

The fitness for different initial conditions plotted against the number of steps suggests that the landscape is more rugged than the geometric one, with the presence of large energy barriers. As Monte Carlo steps increase, the system spends more and more time in one state, see the plateau in $\mathscr{F}$ for the last 3.5×10^4 steps. The evolution in such a landscape is consistent to producing architectures that differ from one initial condition to another as it is shown by the Hamming distance between configurations $|\sigma_i\rangle$ and $|\sigma_j\rangle$ averaged over all pairwise combinations, $d \equiv \langle d(i,j) \rangle$. Given two configurations, the Hamming distance counts, link by link, the number of different spring occupancies normalised by the number of possible springs, N_{ps}. This distance is to be compared to

- the average distance between random networks $\overline{d} \equiv 1 - f$, where f is the frequency of springs fixed by the coordination and size of the system, $f = \frac{zL^2}{N_{ps}}$ — note that $\overline{d} \neq 0.5$ since the networks have a fixed frequency of springs that is larger than one half;
- the maximum distance between two sequences d_{max} which — for this case where the frequency of springs is very large — results to be $d_{max} = 2(1 - f)$.

Within the same initial condition, the distance between different configurations is found to be small, $d \approx 0.06 \ll \overline{d} \approx 0.12$, while the distance between configurations from two different initial conditions is large, $d \approx 0.19$ given that $d_{max} = 0.2353$ — using data from a network of size $L = 12$ with periodic boundaries, $z = 5$ and $N_{ps} = 408$. These results support the presence of a rugged fitness landscape which makes the evolution harder than for the previous geometric fitness function, as discussed intuitively above by looking at the quantities optimised for in cooperative binding, Eq. 1.7.

1.4.2 Architectures: shear, hinge, twist and many others

The properties described above about the exploration of the fitness landscape during evolution suggest that different architectures may be found according to different initial conditions and/or different system parameters, like boundary conditions. Due to this heterogeneity, averages should be performed solely within the same initial condition to capture which design is used, and not across all configurations as done for the geometric fitness. Hence, in the following, observables are averaged over the last 3.5×10^4 steps within the same initial condition, unless stated otherwise. It is interesting to remark that even within the same initial condition, there are different local realisation of the same architecture, reflecting the idea discussed in the Introduction that several allosteric pathways are possible to perform a given task.

Shear design

If *periodic* boundary conditions are chosen on the horizontal axis, the geometry of the network is a cylinder as previously discussed for the geometric fitness funtion. Among the 25 initial conditions, all of them display a shear design, as we show in the following by highlighting the features of the architectures.

- The allosteric response, i.e. the response induced by binding at the allosteric site $d\mathbf{R}^{10}$, of one sample is plotted in Fig. 1.13A, showing that two parts of the network move in opposite directions in a shear-like motion.
- Except from a path, linear in L, connecting the allosteric to the active site, the motion involved is the one of a rigid body, consistent with a shear design[7]. This can be seen by

---

[7] The presence of such continous path connecting the two sites is indeed the discriminator to identify a shear

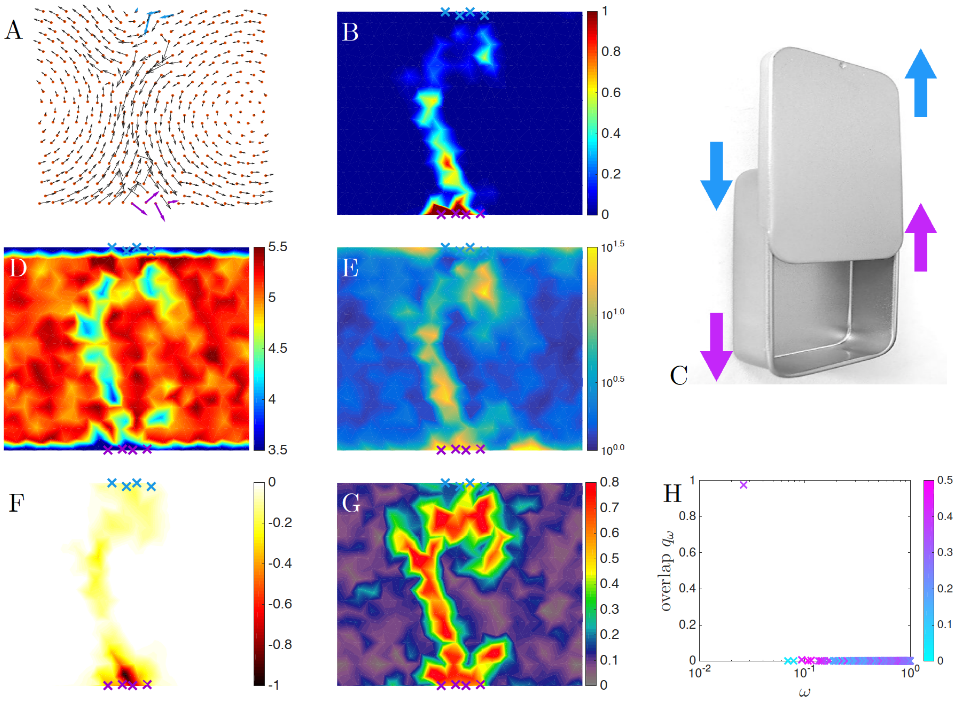

Figure 1.13 – Shear design. (A) The averaged response induced by binding at the allosteric site $d\mathbf{R}^{10}$ is shown as black arrows. (B) The averaged shear pseudo-energy map indicates that shear deformation happens along a linear path connecting the allosteric (violet) and active site (blue). (C) Pictorial sketch of the shear mechanism via a sliding mint box. (D) Map of the average coordination number z. (E) Map of the average strain B-factor. (F) Map of the fitness cost of performing a single mutation in the network normalised as $\Delta\mathscr{F}/\mathscr{F}(|\sigma\rangle)$, where $\Delta\mathscr{F} = \mathscr{F}(|\sigma'\rangle) - \mathscr{F}(|\sigma\rangle)$ and $|\sigma'\rangle$ is the mutated network. (G) Map of conservation of spring occupancy in the evolution scheme. (H) The overlap q_ω of the response with the vibrational modes is shown as function of the mode frequency ω and color-coded according to the participation ratio of the mode, describing how the response is distributed in the network, i.e. either extended or localised. A single extended mode of low frequency describes the response to binding.

looking at the shear pseudo-energy map E_{sh} Fig. 1.13B, quantifying the amount of shear deformation for each node of the network. E_{sh} and its counterpart E_{bulk} describing volume deformations as compressions or dilations are defined as follows

$$E_{sh}(i) = \frac{1}{2}\sum_{a,b=1}^{3}[\gamma_{ab}(i)]^2 \tag{1.8}$$

$$E_{bulk}(i) = \frac{1}{2}\sum_{c=1}^{3}[\epsilon_{cc}(i)]^2, \tag{1.9}$$

where $\boldsymbol{\epsilon}$ is the strain tensor that can be directly computed from any displacement and $\boldsymbol{\gamma}$ is the local shear tensor obtained by removing the trace of $\boldsymbol{\gamma}$; see Section 2.1 for a detailed discussion.

- Structurally, the strain path connecting the two sites correspond to a loosely connected

---

design.

region, with low coordination as shown in Fig. 1.13D.

- The strain path presents also large thermal fluctuations of strain, characterised by the strain B-factor, whose map is shown inFig. 1.13E. Strain B-factors extend the B-factors introduced in the introduction for the Elastic Network Models leaving out the effect of rigid motions. Given a vibrational mode $\mathbf{R}_\omega(i)$ at node i, the strain B-factor is defined as

$$SB(i) = \sum_{\omega>0} \frac{2}{\omega^2} \left[E_{sh,\omega}(i) + E_{bulk,\omega}(i)\right] , \tag{1.10}$$

  where the shear and bulk pseudo-energies are computed taking the mode $\mathbf{R}_\omega(i)$ as displacement. The map of strain B-factors identifies the most flexible regions in the network that correspond to the shearing links and shows a structure that is robust upon thermal flutuations.

A pictorial sketch of the shear design is depicted in Fig. 1.13C with the use of a mint box.

The softest and more strained region turns out to be the one where the fitness cost of performing a single mutation of the spring occupancy is the largest, as shown in Fig. 1.13F. This whole region is also the most conserved (Fig. 1.13G), where conservation is computed starting from the mean occupancy of a link α averaged over a period of observation τ

$$\langle\sigma_\alpha\rangle \equiv \frac{1}{\tau}\sum_{t=1}^{\tau}\sigma_\alpha(t) \tag{1.11}$$

and the deviation from this average — if there is no selection pressure on a link α, $\langle\sigma_\alpha\rangle = \overline{\sigma}$ — defines the conservation Σ_α that reads

$$\Sigma_\alpha = \langle\sigma_\alpha\rangle \ln\frac{\langle\sigma_\alpha\rangle}{\overline{\sigma}} + (1-\langle\sigma_\alpha\rangle)\ln\frac{1-\langle\sigma_\alpha\rangle}{1-\overline{\sigma}} . \tag{1.12}$$

Hinge design

If *open* boundary conditions are chosen on the horizontal axis, around 40 – 50 percent of the initial configurations lead to hinge designs, while the remaining to shear designs. In the following the features justifying the hinge design are reported.

- The response $d\mathbf{R}^{10}$ of one sample candidate to display a hinge design is plotted in Fig. 1.14A and shows that the network is undergoing two rotating, hinge-like motions.

- The shear pseudo-energy map, Fig. 1.14B, shows that the network is behaving like two rigid bodies connected by a hinge in the middle, since strain is located only around the allosteric and active site. A posteriori, a hinge design can then be identified in cooperative networks that do not display a path of large strain connecting the two sites.

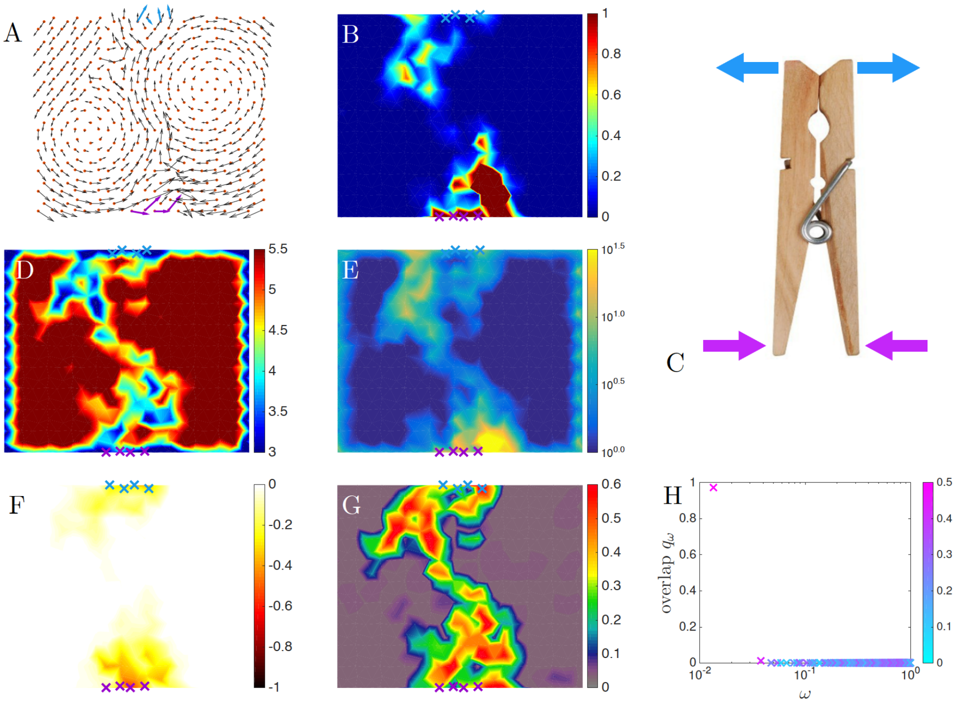

Figure 1.14 – Hinge design. (A) The averaged response induced by binding at the allosteric site $d\mathbf{R}^{10}$ is shown as black arrows. (B) The averaged shear pseudo-energy map indicates that shear deformation happens only in the surrounding of the allosteric (violet) and active site (blue). (C) Pictorial sketch of the hinge mechanism via the movement of a clothespin. (D) Map of the average coordination number z. (E) Map of the average strain B-factor. (F) Map of the fitness cost of performing a single mutation in the network normalised as $\Delta\mathscr{F}/\mathscr{F}$. (G) Map of conservation of spring occupancy in the evolution scheme. (H) The overlap q_ω of the response with the vibrational modes is shown as function of the mode frequency ω and color-coded according to the participation ratio of the mode. A single low energy extended mode is found to contribute to the allosteric response.

- Structurally, the network displays two disconnected regions with low coordination as shown by the coordination map in Fig. 1.14D, consistent with a hinge design.
- The strain B-factors in Fig. 1.14D show that the most thermal fluctations of strain are located around allosteric and active sites.

A pictorial sketch of the hinge design is depicted in Fig. 1.14C with the use of a clothespin.

As it is the case for the shear design, the regions that are less constrained and show more shear energy are also the ones where both the fitness cost of single mutations and conservation are the largest, see Fig. 1.14F.

Twist design

The network can be extended in $d = 3$ dimensions considering a cube with face-centered symmetry of linear size $L = 12$, as shown in Fig. 1.3B, see Appendix A.1 for the details of construction.

The optimisation gives rise to a plethora of solutions, whose architecture is in general difficult to relate to simple designs of daily objects. However, all the architectures are found to fall in the same classification as the ones found in two dimensions. In the following, one solution is described in detail and is proved to be a twist design, akin to the motions of a Rubik's cube.

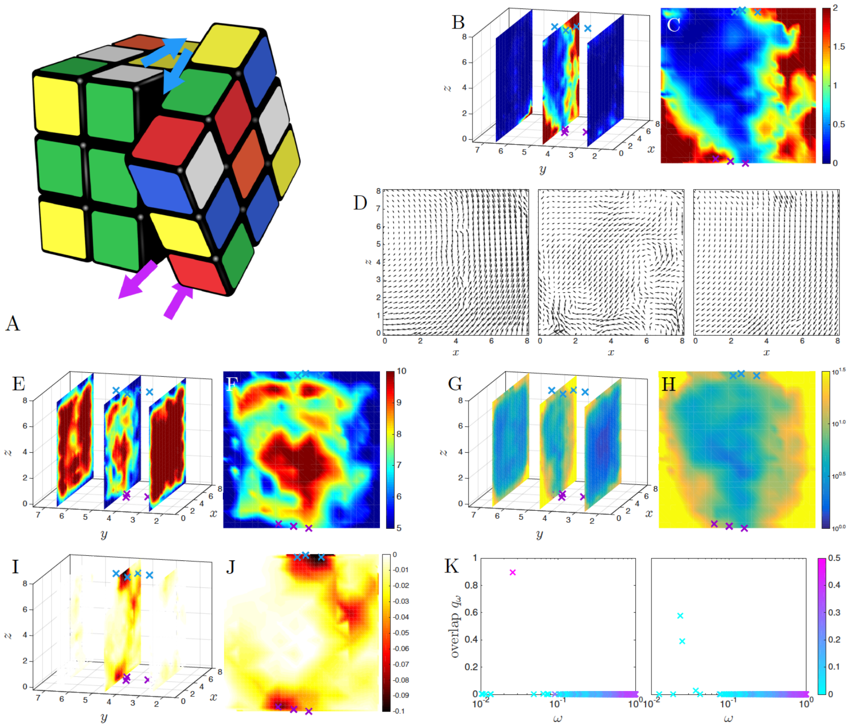

Figure 1.15 – Twist design. (A) Pictorial sketch of the twist mechanism via the movement of a Rubik's cube. (B) Two dimensional sections of the shear pseudo-energy are shown where the allosteric (violet) and active site (blue) are also reported. (C) The shear pseudo-energy in the middle section is shown where deformation is mostly located at the sides. (D) The response induced by binding at the allosteric site $d\mathbf{R}^{10}$ is shown as black arrows in the same sections as of panel B. (E) Map of the average coordination number z on the three sections, where the middle section is shown in panel (F). (G) Map of the strain B-factor on the three sections where the middle one is highlighted in panel (H). (I) Map of the fitness cost of performing a single mutation in the network normalised as $\Delta\mathscr{F}/\mathscr{F}$ is shown on the three sections, with the middle one shown separately in panel (J). (K) The overlap q_ω of the response with the vibrational modes is shown as function of the mode frequency ω and color-coded according to the participation ratio of the mode. It is shown for two different time points in the evolution run. For most time points, several vibrational modes projects into the allosteric response (right), but it is always possible to find a point where just one mode contributes to most of it (left) as it is the case in two dimensions.

The visualisation of the three-dimensional solutions is more challenging and for that three sections of the cube are probed, one at the middle and two at the outer parts. Combining measures of the allosteric response, intensity of shear deformation and rigidity shown in Fig. 1.15B-C-D-E-F, it emerges that the most outer sections behave like rigid-bodies moving in opposite directions, while the middle section display a hinge in the middle, supporting that the solution behaves like a Rubik's cube, sketched in Fig. 1.15A. The intensity of strain fluctuations is also consistent.

Single mutations affect differently the three sections of the cube. Mutations in the outer rigid-body sections are not costly, while in the middle section the nodes close to the boundaries that experience large shear deformation are the ones where single mutations affect the most the fitness, see Fig. 1.15I-J.

1.4.3 Specificity of geometric and cooperative designs

Proteins bind specifically to the ligand at the allosteric site, are the evolved networks also responding only to the originally imposed displacement? Both designs result to be specific with respect to the type of perturbation that is applied at the allosteric site. In the simpler case of an applied force **f** instead of a displacement, the cooperative fitness can be easily written as function of the modes $\boldsymbol{\omega}$ and in particular as function of the mechanism $\boldsymbol{\omega}^\star$

$$\Delta\Delta E = \frac{1}{2}\sum_{\omega}\frac{1}{\omega^2}\left(\langle \mathbf{f}^{11}|\boldsymbol{\omega}\rangle^2 - \langle \mathbf{f}^{10}|\boldsymbol{\omega}\rangle^2 - \langle \mathbf{f}^{01}|\boldsymbol{\omega}\rangle^2\right) \tag{1.13}$$

$$\approx \frac{1}{2\omega^{\star 2}}\left(\langle \mathbf{f}^{11}|\boldsymbol{\omega}^\star\rangle^2 - \langle \mathbf{f}^{10}|\boldsymbol{\omega}^\star\rangle^2 - \langle \mathbf{f}^{01}|\boldsymbol{\omega}^\star\rangle^2\right), \tag{1.14}$$

where it can be seen that if the applied force $\mathbf{f}^{10}$ is orthogonal to the mechanism, the fitness goes to zero. Hence, the fitness is large only if the force is along the mode. A similar argument holds in the case of an applied displacement.

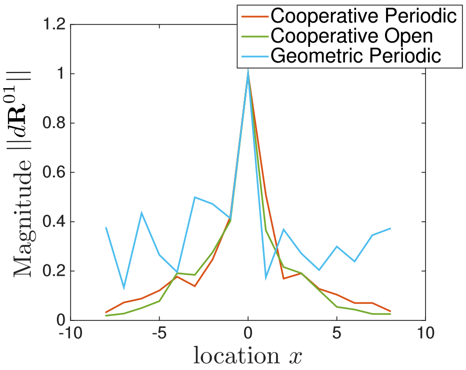

Figure 1.16 – Networks evolved to respond to a perturbation imposed at the allosteric site ($x = 0$) are perturbed at different sites located at the right ($x > 0$) and left ($x < 0$) of the original allosteric site to study how sensitive the designs are when the location of the imposed perturbation is changed. The sensitivity is measured by looking at the average magnitude of the allosteric response at the active site, $||d\mathbf{R}^{01}||$. The analysis is performed for the cooperative designs with both periodic and open boundaries and for the geometric designs with periodic boundaries.

In the case of geometric design, a numerical test shows that applying a different displacement from the original one at the allosteric site results in a considerable decrease in the value of fitness, supporting the specificity of the design upon a change in the imposed displacement.

Specificity on different locations of the imposed perturbation can also be tested. Fig. 1.16 shows that cooperative networks are more specific than geometric ones in responding to the same displacement applied in different locations than the allosteric site. The geometric designs show a less pronounced peak around $x = 0$ than the cooperative ones. Indeed, in the case of geometric designs, the same displacement applied at a different location on the bottom boundary of the network still results in a relativey large response at the active site given the presence of the lever mechanism that amplifies any displacement received on the marginally-connected boundary. On the contrary, an evolved cooperative network shows a smaller response at the active site when perturbed at different sites ($x < 0$ and $x > 0$) .

1.4.4 Classification

Common features of cooperative design appear from the results described in Fig. 1.13, Fig. 1.14 and Fig. 1.15, unifying the different architectures found to optimise cooperative fitness. These common features could be looked for by averaging over all initial conditions. Fig. 1.17 shows that the features observed through different architectures are indeed statistically significant, supporting the definition of principles of optimal cooperative binding. The solutions in Fig. 1.17 are from periodic boundary networks in two dimensions, similar results are obtained for open boundary and three dimensional networks.

- Less cordinated regions are carved in a more rigid structure as it was already observed in a class of proteins [195] and in protein models [180, 196].
- The strain computed from the allosteric response is small in the rigid manifold, supporting rigid-body or long-wavelength motion, and large in the soft manifold. Indeed, the nodes in the network for which the shear deformation is large are the most flexible ones, see Fig. 1.17A. It turns out that these nodes are also the ones most affected by single-point mutations as shown inFig. 1.17B.

Normal mode analysis in Fig. 1.13H, Fig. 1.14H and Fig. 1.15K shows that the response induced by binding at the allosteric site is described by a single mode which is extended — as shown by the participation ratio encoded in the color bar — and has low frequency. This is valid also for the solutions in three dimensions where there are several different designs not easy to connect with a known one. In that case, in most of the time points during evolution several low energy modes contribute to the allosteric response, but there is always one point whereby just one mode describes most of the response.

The presence of a mode easy to excite is a common property of daily life objects designed to be functional — like scissors — and such a mode is called *mechanism.* It has been also observed in allosteric proteins, where it represents the allosteric response, or in usual proteins, where it describes the accessible collective mechanical movements, as discussed in the introduction.

- Low frequency extended modes appear systematically in the density of vibrational modes (Fig. 1.17C) and are the ones identified to be descriptive of the allosteric response (Fig. 1.17D). This is further proven by the cumulative overlap between the vibrational modes and the allosteric response, see Fig. 1.17D. The cumulative overlap is larger than 80 percent for the lowest frequency modes (rank equal to 1) and quickly saturates to 100 percent (rank equal to 3) for networks evolved to optimise cooperative binding both in two and three dimensions, manifestation of the presence of a mechanism. The same measure for networks evolved to perform the geometric task, optimising for Eq. 1.5, shows a slow saturation to large values of the overlap, consistent with the fact that function in those networks is not attained via mechanisms but edge modes.

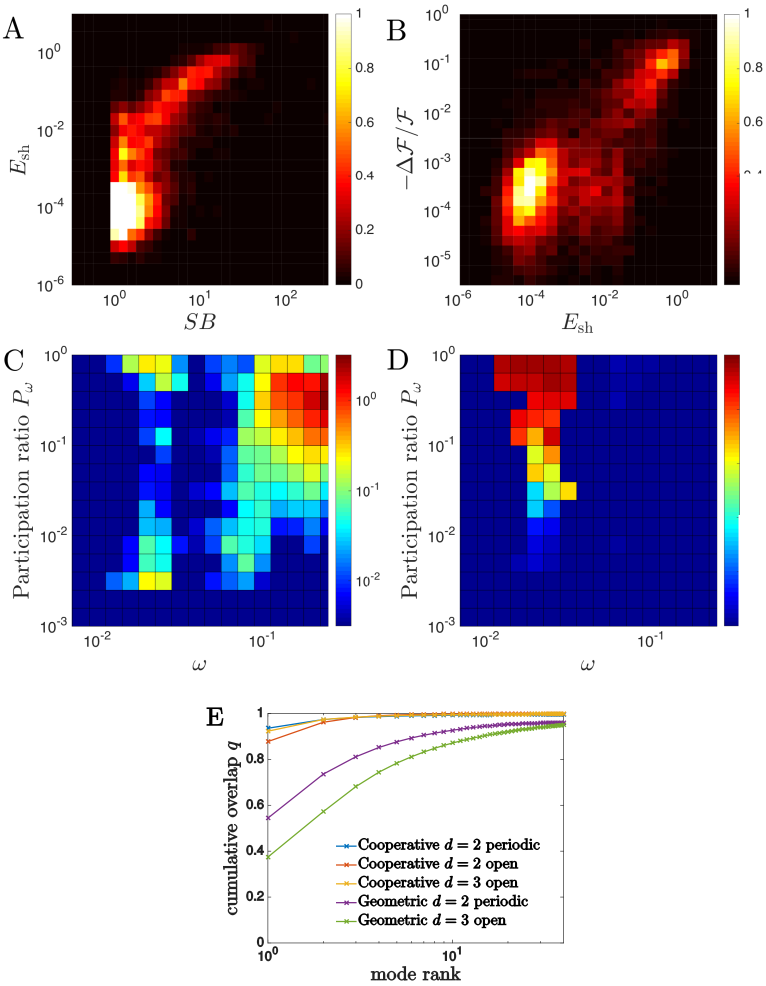

Figure 1.17 – Averages are performed on all 25 initial conditions to elucidate the common features shared by the designs. The bivariate histograms of nodes in the network displaying (A) a given strain B-factor (SB) and shear pseudo-energy (E_{sh}) and (B) a given shear pseudo-energy (E_{sh}) and normalised mutation cost ($-\Delta\mathscr{F}/\mathscr{F}$) are shown. The color bar codes for the relative abundance of data points in each bin. The positive correlation for large values found in both measures indicates that the most flexible regions (larger strain B-factors) and the ones most susceptible to mutations (larger $-\Delta\mathscr{F}/\mathscr{F}$) are also the ones with the most shear deformation (larger E_{sh}). (C) The density of vibrational modes plotted as function of both the frequency ω and the participation ratio P_ω, $D(\omega, P_\omega)$ shows systematically the presence of an extended mode with low frequency. (D) Moreover, the overlap $q(\omega, P_\omega)$ is close to unity exactly in the same region (ω, P_ω), proving that the extended soft mode describes most of the response induced by binding at the allosteric site. (E) The cumulative overlap is shown as function of its rank for the first modes both for the geometric and the cooperative fitness function, highlighting the difference in the physical mechanism used to be fit, respectively edge modes and mechanisms.

1.4.5 Optimal cooperative binding

The common features shared by the in-silico evolved networks can be rationalised by looking at cooperative binding in a continuum elastic medium in the absence and presence of a mechanism.

Absence of design

In the absence of design, the cooperative fitness decreases very rapidly as function of distance betwen the allosteric and active site, L, as discussed in the following in the case of a medium with elastic modulus G. The effect of a strain imposed by binding at one site can be decomposed into multipole moments that describe the displacement field at distant regions where the dominant contribution is dipolar, since higher multipoles decay faster. More precisely, the imposed force $\mathbf{f}$ acts on the n nodes of the site at positions $\mathbf{a}^s$ with $s = 1, \ldots, n$, and the resulting displacement is computed via the Green function $G_{ij}(\mathbf{r} - \mathbf{a}^s)$ (see Appendix A.4 for more details)

$$\begin{aligned} u_i(\mathbf{r}) &= \sum_{s=1}^{n} G_{ij}(\mathbf{r} - \mathbf{a}^s) f_j^s && (1.15) \\ &= G_{ij}(\mathbf{r}) \sum_{s=1}^{n} f_j^s - G_{ij,k}(\mathbf{r}) \sum_{s=1}^{n} f_j^s a_k^s + \mathcal{O}(||\mathbf{a}^s||^2) && (1.16) \\ &= -G_{ij,k}(\mathbf{r}) \sum_{s=1}^{n} f_j^s a_k^s + \mathcal{O}(||\mathbf{a}^s||^2) \quad \text{(no net force on the nodes: } \sum_{s=1}^{n} \mathbf{f}^s = 0) . && (1.17) \end{aligned}$$

The Green function in $d > 2$ dimensions being

$$\begin{aligned} G_{ij}(\mathbf{r}) &= g_{ij}\left(\frac{\mathbf{r}}{r}\right) \frac{1}{r^{d-2}} && (1.18) \\ G_{ij,k}(\mathbf{r}) &= h_{ijk}\left(\frac{\mathbf{r}}{r}\right) \frac{1}{r^{d-1}} , && (1.19) \end{aligned}$$

where g and h contain the angular dependence. Hence, the imposed strain will induce a dipolar response sufficiently far away from the binding site, thus allowing to approximate the effect of a binding event as a dipole applied at the allosteric and active site, as sketched in Fig. 1.18A. The magnitude of each dipole is set to fc where f is the norm of the applied force and c the distance between the two poles. The two terms in the approximated form of the fitness Eq. 1.7 can be explicited. The response $|d\mathbf{R}^{01\to 11}\rangle$ is dipolar with magnitude dependent on the distance r given by fc/Gr^{d-1}. The force field $|\mathbf{F}\rangle$ is also dipolar, hence its scalar product with $|d\mathbf{R}^{01\to 11}\rangle$ acts as a derivative at $r = L$ for $L \gg c$ giving

$$\mathcal{F} \approx \langle d\mathbf{R}^{01\to 11} | \mathbf{F} \rangle \approx |||\mathbf{F}\rangle|| \frac{d}{dr} \frac{fc}{Gr^{d-1}}\bigg|_{r=L} = (1-d)\frac{f^2 c^2}{GL^d} \sim L^{-d} \quad d > 2 , \tag{1.20}$$

which is indeed a very rapid decay of the cooperative fitness with the distance L. For a precise derivation of this result and the two-dimensional case see Appendix A.4.

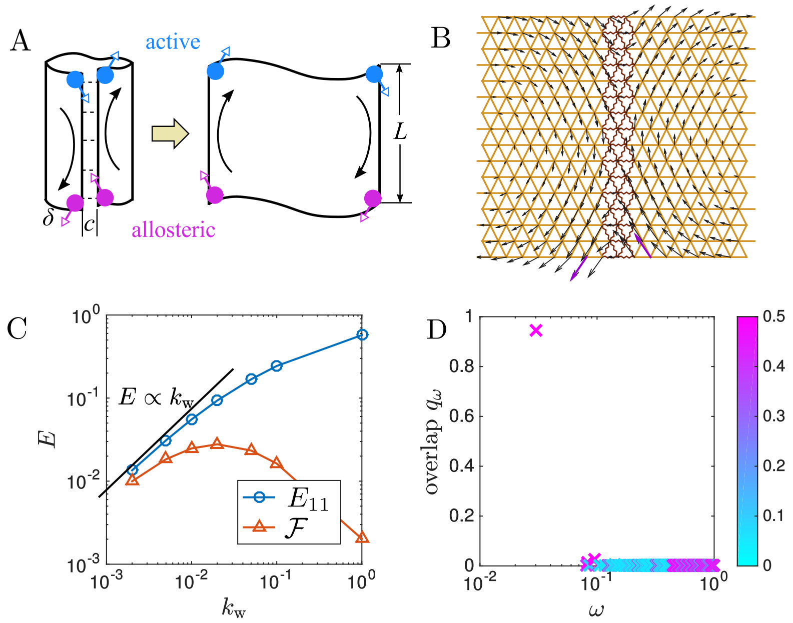

Figure 1.18 – (A) A shear mechanism is designed in a cylinder of height L by cutting a slice of width c. The resulting object has the topology of a square with a rotation as an additional mode with zero energy cost. If the cut is filled with a soft material, the frequency of the mode gets finite. The response induced by imposing displacements at the allosteric or active site will be dominated by the mechanism, as long as the material in the cut is soft enough. (B) A shear architecture is built in a triangular lattice of size $L = 16$ with the introduction of a soft band of width $c = L/10$ with springs of stiffness $k_w \ll k = 1$ such that $G_w/G = k_w/k$ ($k_w = 0.05$). The displacement imposed at the allosteric site is shown as purple arrows, while the resulting response as black arrows. (C) The energy of binding simultaneously at the allosteric and active site E_{11} and the cooperative fitness $\mathscr{F}$ are shown as function of k_w for a network of size $L = 32$. The dependence of $\mathscr{F}$ on k_w is indeed non-monotonic, with an optimum for a certain $k_w^\star$. (D) The overlap q_ω between the response $\mathbf{R}^{10}$ and the vibrational modes is shown as function of the frequency ω at optimal $k_w^\star = 0.036$ for $L = 32$ and $c = L/10$. The color code is according to the participation ratio.

Presence of design

How does the presence of a mechanism affect the cooperative fitness? The continuum elastic medium is designed in $d = 2$ to have a shear design with the considered geometry depicted in Fig. 1.18A, where a band of width c is cut in a cylinder of height L. The two dimensional equivalent is also shown. Let G_w be the elastic modulus of the band and δ the displacement imposed at the allosteric or active site. If the cut surface is empty ($G_w = 0$) the material will display a zero energy mode, corresponding to the sliding of the two ends of the cylinder or to the rotation of the two dimensional sheets. This is valid also for G_w sufficiently small. Since the mode costs no energy, an applied displacement will couple only to it and lead to the same response when applied at any of the sites, leading to $E_{10} \approx E_{01} \approx E_{11}$ and to a cooperative fitness set by the elastic energy stored in the band

$$\mathscr{F} = E_{10} + E_{01} - E_{11} \approx E_{10} \approx \frac{L}{c} G_w \delta^2 . \tag{1.21}$$

The dependence on G_w in Eq. 1.21 shows that $\mathscr{F} = 0$ when a mechanism is present ($G_w = 0$) and $\mathscr{F}$ increases linearly with G_w. However, cooperative fitness will not grow indefinitely. When the band becomes more rigid than the material, it is more favorable to deform the latter with respect to the band, thus coupling to modes other than the mechanism and breaking the validity of Eq. 1.21. This crossover happens when the energy of deforming a continuous medium of modulus G, E_c, is lower than the energy associated to the mechanism, which happens for a given $G_w^\star$

$$E_c = \frac{G\delta^2}{\log(L/c)} < \frac{L}{c} G_w \delta^2 \quad \Rightarrow \quad \frac{G_w^\star}{G} \approx \frac{c}{L\log(L/c)}. \tag{1.22}$$

For $G \gg G_w^\star$ the imposed displacement will excite modes typical of a continuous medium, resulting in a cooperative fitness rapidly decaying as in Eq. 1.20. Hence, the value $G_w \approx G_w^\star$ corresponds to an optimum in cooperative fitess of the order

$$\mathscr{F}^\star \approx \frac{G\delta^2}{\log(L/c)}. \tag{1.23}$$

These theoretical predictions can be tested in the discrete setting of a triangular lattice, as introduced in Fig. 1.18B, with $N = L^2$ nodes and a band of width $c = L/10$ where the spring stiffness is $k_w \ll k = 1$ chosen such that $G_w/G = k_w/k$. The measured fitness is shown to have a non-monotonic behaviour as function of k_w in Fig. 1.18B. In Fig. 1.18C the emergence of a mechanism appears when the stiffness of the soft springs $k_w^\star \sim L^{-1}$ ($k_w^\star = 0.036$ for $L = 32$ and $1/32 \simeq 0.031$) as predicted in Eq. 1.22.

Optimal cooperative binding is associated to the excitation of a mechanism with a very small value of the cooperative energy of the response to the imposed displacement δ, as it is seen from Eq. 1.23 when L is large. The mechanism can be found by looking at the vibrational spectrum of the stiffness matrix where, in a harmonic approximation, the inverse of each eigenvalue is proportional to the extent of thermal fluctuations along the associated eigenvector. The order of the eigenvalue $k_\psi^\star$ associated to the mechanism $d\boldsymbol{\psi}$ can be estimated via a simple argument

$$k_\psi^\star \approx \frac{\mathscr{F}^\star}{||d\boldsymbol{\psi}||^2} \approx \frac{1}{\delta^2 L^2} \frac{G\delta^2}{\log(L/c)} \sim \frac{1}{N\log\left(\sqrt{N}/c\right)}, \tag{1.24}$$

since the shear mechanism is extended, all the particles — whose number in the case of a discrete material is $N = L^2$ — move of a distance δ and its square norm is then $||d\boldsymbol{\psi}||^2 \approx \delta^2 N$. The frequency of the mode can be directly obtained by supposing that all particles have identical mass m

$$\omega_\psi^\star \approx \sqrt{\frac{k_\psi^\star}{m}} \sim \frac{1}{\sqrt{N\log\left(\sqrt{N}/c\right)}}. \tag{1.25}$$

For the range of values of N considered, the logarithmic correction is negligible and the scaling reduces to $k^\star_\psi \sim N^{-1}$. The result is easily extended to the three-dimensional shear mechanism where $E_c = G\delta^2$ and $\mathscr{F} \approx Lc^{-2}G_w\delta^2$, giving $\mathscr{F}^\star \approx Gc\delta^2$ and $k^\star_\psi \sim N^{-1}$ since $||d\boldsymbol{\psi}||^2 \approx \delta^2 L^3 = \delta^2 N$.

The value of $k^\star_\psi$ sets the scaling of the stiffness of the mechanism with respect to the number of particles N in order to have optimal cooperative binding. Equivalently, it sets the scaling of thermal fluctuations along the mechanism: the fluctuations increase as N for systems where cooperative binding is optimal. The presence of a mechanism not only is responsible for an increase of cooperative fitness with respect to the undesigned case, but it poises the system to its optimal value.

In this derivation, it is assumed that maximising the cooperative binding energy is the best strategy for the protein to achieve allosteric function. Another assumption to be explored is to maximise sensitivity, the derivative of fitness with respect to the stiffness, as suggested to be the strategy used by the transcription factor CAP in Ref. [197].

1.4.6 Fluctuations-driven cooperativity

As discussed in Subsec. 1.1.3 and in Refs. [27,180], the existence of a low energy extended mode of frequency $\omega^\star_\psi$ (the mechanism in the case of cooperative designs) leads to the possibility of a cooperative effect without any observed mean displacement, once thermal contributions are considered. Indeed, a binding event can affect how particles near the site fluctuate around their mean position, while leaving the mean itself unaltered. Fluctuations-driven cooperativity can then arise as follows. Binding at the active site freezes motion in that region and results in an increase of the soft mode frequency, leading to an entropic cost that can be diminished if binding already took place at the allosteric site.

Let us define ω_{10}, ω_{01} and ω_{11} the frequencies of the mechanism after binding at the allosteric site, active site and both respectively. We can estimate these quantities as

$$\omega_{10}^2 = \omega^{\star\,2}_\psi + e_{10} \tag{1.26}$$

$$\omega_{01}^2 = \omega^{\star\,2}_\psi + e_{01} \tag{1.27}$$

$$\omega_{11}^2 = \omega^{\star\,2}_\psi + e_{11}\,, \tag{1.28}$$

where e_{10} (e_{01}) characterizes the additional energy required for the mode to move when a ligand is bound at the allosteric (active) site. Assuming harmonic dynamics, the entropy of a normal mode of frequency ω reads

$$S = k_B \ln\left(\frac{k_B T}{\hbar\omega}\right). \tag{1.29}$$

Using this expression, the cooperative free energy is estimated as

$$\Delta\Delta F = -T\Delta\Delta S = k_B T \ln\left(\frac{\omega_{11}\omega_{\psi}^{\star}}{\omega_{10}\omega_{01}}\right) \tag{1.30}$$

$$= -k_B T \ln\left(1 - \frac{e_{11}}{({\omega_{\psi}^{\star}}^2 + e_{10})({\omega_{\psi}^{\star}}^2 + e_{01})}\right), \tag{1.31}$$

which can indeed be large if ${\omega_{\psi}^{\star}}^2$ is small compared to both e_{10} and e_{01}.

1.4.7 Conclusions

The choice of the cooperative fitness function in the evolution scheme produces a new class of architectures, like shear or hinge, consistent with known designs adopted by allosteric proteins shown in Fig. 5. These architectures use different physical principles to function with respect to the edge modes underlying the designs optimising the geometric fitness. Evolution is harder in the cooperative fitness landscape: different initial conditions correspond to different configurations, as can be seen by computing their Hamming distance. Although this heterogeneity, all the cooperative architectures work according to the same principles. They develop an extended mode with very low frequency describing most of the structural change upon perturbation. This mode is called mechanism and it is easy to couple to when a strain mimicking binding is imposed at one of the two sites, then inducing the allosteric response.

The presence of a mechanism rationalises the aforementioned empirical success of normal mode analysis in detecting few soft modes responsible for the observed allosteric response. Indeed, in this context, soft modes are found to appear as the result of evolution of architectures that display cooperative binding.

Moreover, cooperative architectures not only increase cooperative fitness with respect to the undesigned ones, but are poised to display an optimal value. Indeed, the stiffness of the mode is not random: it cannot be too low nor too large in order to have a finite value of cooperative fitness. An optimum is found for a given stiffness, whose scaling with the number of particles in the system $N = L^d$ is $k_{\psi}^{\star} \sim N^{-1}$ where d is the spatial dimension and L the linear system size. From this result, it follows the prediction that larger proteins may need softer modes to operate cooperatively than smaller ones. In Chapter 2 this prediction is tested by computing normal modes with ENM in a dataset of 34 allosteric proteins supporting the predicted trend.

1.5 Effect of non-linearity: a mechanical view on induced fit and population shift

The in-silico evolution of disordered mechanical networks leads to the emergence of allosteric behaviour and physical requirements to be satisfied by the networks in order to have optimal cooperative binding between the allosteric and active site. The considered networks are composed of harmonic springs, displaying a linear behaviour. Hence, the transition from non functional to functional networks happens in an energy landscape that is harmonic, listing the previously discussed results in the class of induced fit models (KNF model) presented in the introduction, see Fig. 4.

It comes natural to ask which is the effect of non-linearities in the evolution of allosteric behaviour with a landscape as the one of the population shift scenario, see Fig. 4. In particular, do such networks in the population shift scenario feature the same behaviour in cooperative energy as function of stiffness as the one in the induced fit, Fig. 1.20C? Is the stiffness of the soft mode still optimal, Eq. 1.24, in the population shift scenario?

1.5.1 Induced fit scenario

The extension of the mechanical framework to the induced fit scenario involves a geometric interpretation of the results discussed in the previous section.

Consider a protein that displays cooperative binding thanks to the presence of an elastic soft mode $\boldsymbol{\psi}$, which, when excited, will induce the allosteric response in the protein. Cooperate fitness is defined as in Eq. 3.1, but with the symmetric assumption between the two binding sites leading to $E_{01} = E_{10}$ and

$$\Delta\Delta E = 2E_{01} - E_{11} \,. \tag{1.32}$$

By introducing a variable x that denotes the motion along the mode $\boldsymbol{\psi}$ and varying from 0 to 1 as the protein undergoes its allosteric response, the energy profile of the unbound state can be expressed as function of x and it reads

$$E_{00}(x) = k_a ||d\mathbf{R}_a||^2 f(x) \,, \tag{1.33}$$

where k_a is the stiffness of the mode[8] and $f(x)$ a function of order unity and such that $f''(x = 0) = 1$. To be consistent with the induced fit scenario, $f(x)$ displays a unique minimum at $x = 0$ as sketched in Fig. 1.19. For a linear elastic material, like the one discussed previously, $f(x) = x^2/2$. The squared norm is defined as $||d\mathbf{R}_a||^2 = \sum_{i=1}^{N} ||d\mathbf{R}_a(i)||^2$ where $d\mathbf{R}_a(i)$ is the allosteric response field at position i. In the presence of a mechanism, $d\mathbf{R}_a$ and $\boldsymbol{\psi}$ have almost unitary overlap. The term $||d\mathbf{R}_a||^2$ takes into account the volume where the allosteric response takes place: if the protein has N residues, then $||d\mathbf{R}_a||^2 \approx Na^2$ where a is the interatomic distance. The dependence of $||d\mathbf{R}_a||^2$ with N in allosteric proteins is discussed in Chap. 2.

If after some motion x_0 along the mode $\boldsymbol{\psi}$ the protein can accomodate the ligands without deforming, the energy profiles of the bound states $E_{10}(x)$ and $E_{11}(x)$ satisfy $E_{10}(x_0) = E_{11}(x_0) =$

---

[8] Here, the stiffness of the mode is denoted as k_a instead of k_ψ since $k_\psi \approx k_a$ given the relationship between the mechanism and the allosteric response, where k_a is the stiffness of the allosteric response. Indeed, the allosteric response is a more direct observable than the mechanism in proteins.

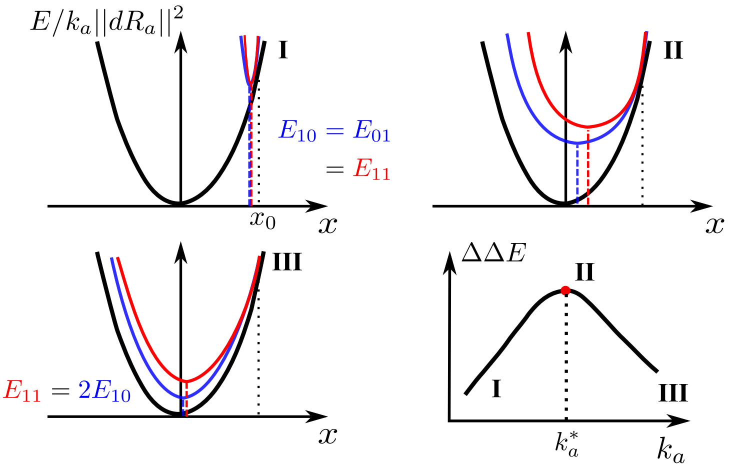

Figure 1.19 – Induced fit scenario. The elastic energy of the unbound configuration $E_{00}(x)/k_a$ (in black), the singly bound configuration $E_{01}(x)/k_a$ (in blue), and doubly bound configuration $E_{11}(x)/k_a$ (in red) as a function of the imposed motion x along its soft mode are shown, rescaled for visibility by the stiffness of the soft mode k_a. The values E_{00}, E_{10}, and E_{11} correspond to the minima of the black ($E_{00}(x)/k_a$), blue ($E_{01}(x)/k_a$), and red ($E_{11}(x)/k_a$) curve, respectively. We assume that the protein accomodates the ligands without deforming when the motion along the soft mode is $x = x_0$; thus, the three energy profiles are identical at that point. Three scenarios (I, II and III) appear depending on the value that the stiffness of the mode Nk_a takes with respect to the stiffness of interatomic interactions nk. Cooperative fitness $\Delta\Delta E$ as function of k_a is sketch, highlighting the resulting behaviour in the three scenarios, I, II and III.

$E_{00}(x_0)$. However as x departs from x_0, the shape is not matched anymore with the result of imposing an elastic strain at the binding sites that leads to an increase of elastic energy in the protein, triggering motion along other elastic modes than $\boldsymbol{\psi}$. The difference in energy profiles for one binding event with the assumption of rigid ligands is

$$E_{10}(x) - E_{00}(x) = nka^2 g(x - x_0)\,, \tag{1.34}$$

where the binding site is composed of order n atoms that move by a distance a as the protein undergoes its allosteric response, k is the stiffness of interatomic interactions, which is significantly larger than the one of the soft mode $k \gg k_a$. The function g is a dimensionless function that vanish quadratically in zero, but is possibly nonlinear at large arguments. Similarly,

$$E_{11}(x) - E_{10}(x) = nka^2 g(x - x_0)\,. \tag{1.35}$$

The energy values used to compute cooperative fitness, E_{10} and E_{11}, are directly obtained by minimising the energy profiles, $E_\bullet \equiv \min_x E_\bullet(x)$. Two extreme cases arise, leading to different results in this minimisation, as illustrated in Fig. 1.19, panel I and III.

- **Region I** (Fig. 1.19-I): if $k_a N \ll nk$, as x departs from x_0, the elastic energy induced by binding $E_{01}(x) - E_{00}(x)$ is significantly larger than the energy of the mode $E_{00}(x)$ meaning that $E_{10}(x) = nka^2 g(x - x_0) + E_{00}(x)$ has a sharp minimum in $x = x_0$ given the functional form of $g(x-x_0)$. Hence, the minimisation of $E_{10}(x)$ gives $E_{10} \approx k_a ||d\mathbf{R}_a||^2 f(x_0)$. Similarly for $E_{11}(x)$, $E_{11} \approx k_a ||d\mathbf{R}_a||^2 f(x_0)$. Cooperative fitness is then

$$\Delta\Delta E = 2E_{10} - E_{11} \approx E_{10} \approx k_a ||d\mathbf{R}_a||^2 f(x_0)\,, \tag{1.36}$$

which vanishes linearly as $k_a \to 0$.

- **Region III** (Fig. 1.19-III): if $k_a N \gg nk$, $E_{00}(x)$ is large with respect to $nka^2 g(x-x_0)$ so that $E_{00}(x)$, $E_{10}(x)$ and $E_{11}(x)$ are all similar and present a minimum at $x=0$. When $x=0$ the system does not move along the mode $\boldsymbol{\psi}$ upond binding: the allosteric and active site are not coupled via an extended mode and thus $E_{11} \approx E_{10} + E_{01}$. Hence, the resulting cooperative fitness vanishes $\Delta\Delta E \to 0$ as $k_a \to \infty$.

- **Region II** (Fig. 1.19-II): a sweet spot where cooperative binding is optimal is found for

$$k_a^\star \sim \frac{nk}{N}\,, \tag{1.37}$$

rederiving the result obtained in Sec. 1.4 (Eq. 1.24). The argument for the presence of the optimum does not require $f(x)$ to be a parabola, but a function which grows monotonically in the directions around its minimum, $x=0$.

1.5.2 Population shift scenario

The population shift scenario, exemplified by the MWC model as overviewed in the Introduction, assumes a protein with two states, inactive (In) and active (Ac) between which it switches even in the absence of ligands. Binding events have the effect of shifting the equilibrium towards one or the other.

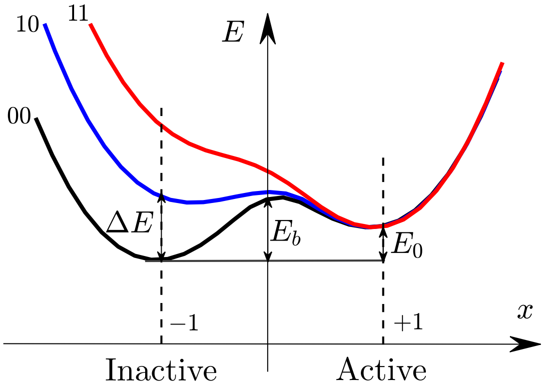

Figure 1.20 – The energy profiles for the different binding events $E_{00}(x)$ (black), $E_{10}(x)$ (blue) and $E_{11}(x)$ (red) are sketched as function of the amplitude of motion x along the path connecting the inactive (In) and active (Ac) state. The features of the potential are also highlighted, such as the height of the energy barrier between the two states E_b, the energy cost of binding one ligand ΔE and the difference in energy between the inactive and active state when no ligands are bound E_0.

In the absence of ligands, the energy of the inactive state can be set as reference $E_{00}^{In} = 0$ keeping the energy of the active state $E_0 \equiv E_{00}^{Ac}$ as variable. In the active state, the protein is assumed to have the right structure to bind the ligand without deforming, hence binding events cost no energy $E_{10}^{Ac} = E_{01}^{Ac} = E_{11}^{Ac} = E_0$. This is not the case in the inactive state where the binding energy cost is ΔE so that $E_{10}^{In} = E_{01}^{In} = \Delta E$ and $E_{11}^{In} = 2\Delta E$, since within a given configuration if the two binding sites are distant enough — i.e. the elastic coupling between them is very weak — the binding energies are additive. A sketch of the energy landscape can be found in Fig. 1.20.

For each binding configuration, the favorable state is the one that has minimal energy, e.g. $E_{01} = \min\left(E_{10}^{In}, E_{10}^{Ac}\right)$. Cooperative fitness can be then directly computed and three different cases arise according to the relationship between E_0 and ΔE, sketched in Fig. 1.21

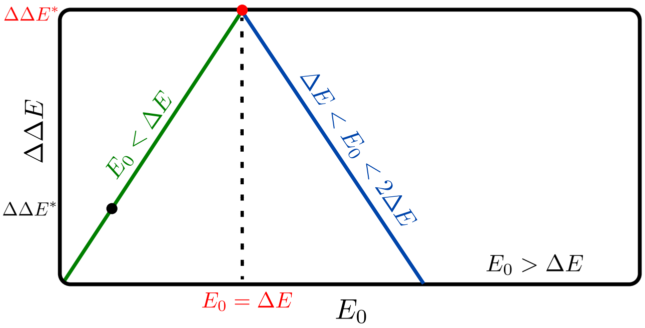

Figure 1.21 – Cooperative energy $\Delta\Delta E$ as function of the energy cost of binding a ligand in the active state, E_0 is shown in the context of the population shift scenario. The maximum value of cooperative fitness that the system can reach is denoted as $\Delta\Delta E^*$.

- if $E_0 < \Delta E$, the binding of one ligand is sufficient to shift the equilibrium of the system to the active state and according to what defined above it implies $E_{10} = E_{11} = E_0$ and $\Delta\Delta E = E_0$;
- if $\Delta E < E_0 < 2\Delta E$, the binding of one ligand is no more sufficient. Two ligands need to bind to shift the equilibrium towards the active state, leading to a cooperative fitness $\Delta\Delta E = 2\Delta E - E_0$;
- if $E_0 > 2\Delta E$, the system is always in the inactive state even when two ligands are bound and $\Delta\Delta E = 0$.

The resulting cooperative fitness displays a maximum for $E_0 = \Delta E$, as shown in Fig. 1.21.

In the presence of a soft mode responsible for the allosteric function, the inactive and active states are connected by a favoured path. Similarly to what discussed for the induced fit scenario, the energy in the absence of ligands can be expanded as function of the amount of motion x along that path, considering here x as varying from $+1$ to -1 given the two states nature of the problem, which also enforces a fourth degree polynomial. The energy $E_{00}(x)$ reads

$$E_{00}(x) = k_a ||d\mathbf{R}_a||^2 \left(\frac{1}{8} x^4 - \frac{1}{4} x^2 + xb \right), \tag{1.38}$$

where out of the five parameters of a polynomial of fourth degree, three are fixed by redefining the reference energy and changing the value of x both via multiplicative and additive constants. The parameter b describes the amount of tilt of the potential towards the inactive state, k_a the stiffness of the soft mode and $||d\mathbf{R}_a||^2$ the squared norm of the allosteric response. A typical profile that satisfies Eq. 1.38 is shown in Fig. 1.20 where the inactive state $E_{00}(x = -1)$ has lower energy than the active state $E_{00}(x = +1)$.

In the case of no ligands bound, when $b = 0$, $E_{00}(x)$ is symmetric and it presents two minima at $x = \pm 1$ with the same energy and separated by an energy barrier $E_b = \frac{1}{8} k_a ||d\mathbf{R}_a||^2$. When b is positive, $E_{00}^{Ac} > E_{00}^{In}$ up to $b = b_c = 1/(3\sqrt{3})$, where the active state become unstable and only one stable state is left. The energy difference between the inactive and active state E_0 is maximal at $b = b_c$ for fixed k_a, finding $E_0 = \frac{3}{8} k_a ||d\mathbf{R}_a||^2$. When b is small, $E_0 \simeq 2 k_a ||d\mathbf{R}_a||^2 b$. Hence, when $b < b_c$ Eq. 1.38 is consistent with the phenomenology of the population shift model and it constitutes the regime considered in the following.

Binding events that occur when the protein is in the active state cost no energy, as discussed before. However, by departing from the active state, the topology of the protein does not match the ligand and it needs to deform elastically near the binding site to accomodate the ligand. Similarly as before

$$E_{10}(x) - E_{00}(x) = nka^2 g(x - x_{Ac}), \tag{1.39}$$

where x_{Ac} is the location of the active state along the path and for b small $x_{Ac} \approx 1$. This binding energy can be found as the difference between the blue and black curves in Fig. 1.20. In the inactive state $x = x_{In}$ the binding energy

$$\Delta E = E_{10}(x_{In}) - E_{00}(x_{In}) = nka^2 g(x_{In} - x_{Ac}) \approx 4nka^2, \tag{1.40}$$

is independent both of the size of the protein N and the stiffness of the mode k_a, where the factor 4 comes from the choice $g(x - x_{Ac}) = (x - x_{Ac})^2$[9] and the limit of b small where $x_{In} \approx -1$ and $x_{Ac} \approx 1$. Supposing that the two binding sites are distant enough, the energy cost of the double binding event for a given value of x is $E_{11}(x) - E_{00}(x) = 2nka^2 g(x - x_{Ac})$.

The goal is to understand how the stiffness of the mode k_a constrains cooperative fitness in the case of a population shift scenario like the one just introduced. The maximal cooperative

---

[9] The choice $g(x - x_{Ac}) = (x - x_{Ac})^2$ serves as an illustration. Qualitatively, the results will not change for other choices of $g(x - x_{Ac})$ apart from the multiplying factor that will change from 4.

fitness over all possible values of tilt b given k_a is

$$\Delta\Delta E^* \equiv \max_b\{\Delta\Delta E(k_a, b)\}\,. \tag{1.41}$$

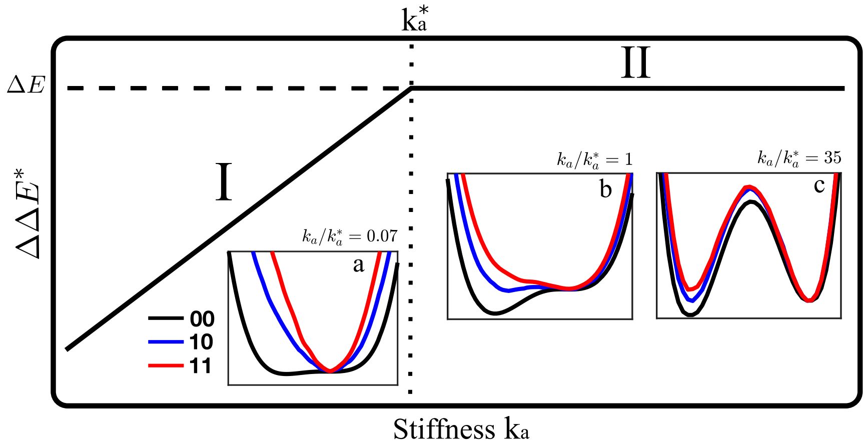

Figure 1.22 – The behaviour of the maximal cooperative energy $\Delta\Delta E^* \equiv \max_b\{\Delta\Delta E(k_a, b)\}$ as function of the stiffness of the mode is shown, where a linear growth (region I) and plateau (region II) appear for, respectively, small and large k_a. In the insets the energy profiles for $k_a \ll k_a^\star$ (a), $k_a \sim k_a^\star$ (b) and $k_a \gg k_a^\star$ (c) are plotted. The data are from a nonlinear elastic model of allostery whose implementation is discussed below.

Two cases appear as function of the relationship between Nk_a and nk, see regions I and II in Fig. 1.22

- **Region I**: if $Nk_a \ll nk$, the elastic cost associated to binding, ΔE, is much larger than E_0 and both $E_{10}(x)$ and $E_{11}(x)$ are peaked around x_{Ac}, as illustrated in Fig. 1.22-a. Hence, $E_0 < \Delta E$, which implies $\Delta\Delta E = E_0$, as summarised in Fig. 1.21 and it is maximal for $b = b_c$, giving

$$\Delta\Delta E^* = \frac{3}{8} k_a ||d\mathbf{R}_a||^2 \sim k_a N a^2\,. \tag{1.42}$$

  This proves that maximal cooperative fitness vanishes linearly for small k_a as skecthed in Fig. 1.22. The behaviour of cooperative fitness for small k_a in the population shift scenario is similar to the one of the induced fit scenario, also linearly vanishing.

- **Region II**: if $Nk_a \gg nk$, the opposite case arises: the elastic cost ΔE is much smaller than E_0 and all energy profiles $E_{00}(x)$, $E_{10}(x)$ and E_{11} are approximately identical, as illustrated in Fig. 1.22-c. Again referring to Fig. 1.21, cooperative fitness is maximised by choosing a value of tilt b that fixes E_0 to ΔE, implying

$$\Delta\Delta E^* = \Delta E \sim nka^2\,, \tag{1.43}$$

  which is independent of the stiffness of the mode, k_a. The plateau of constant cooperative fitness appears at a value of the stiffness of the mode whose scaling with N is

$$k_a^\star \sim \frac{nk}{N}. \tag{1.44}$$

In opposition to the induced fit scenario, maximal cooperative fitness in the population shift framework does not vanish as the stiffness of the mode increases, but it displays a plateau at a value of stiffness $k_a^\star$, which has the same scaling with the number of residues N as in the induced fit scenario. There, the vanishing of cooperative fitness at large k_a stems from the behaviour of $E_{00}(x)$, $E_{10}(x)$ and $E_{11}(x)$. These binding energies all have very similar profiles when k_a is large and all present a single minimum, whose position is also approximately identical and reduces to be $x \approx 0$ independently of binding. Introducing two minima in the energy landscape allows to add a degree of freedom that does not constrain x to be zero, hence allowing the coupling to the soft mode. As a consequence, non vanishing cooperative fitness is displayed also at large k_a: the maximal cooperative fitness $\Delta\Delta E^*$ saturates to a constant value for $k_a \gg k_a^\star$ in the population shift scenario. This result suggests that mutations that increase the stiffness of the low energy mode have a milder effect in the population shift scenario, than in the induced fit, where cooperativity reaches an optimum at $k_a^\star$.

Anyhow, the energy barrier between the inactive and active state increases linearly with k_a in that limit, meaning that the transition between the two states becomes less and less favourable. Hence, given the slow transition rates for $k_a \gg k_a^\star$, the value of k_a reasonably lies close to $k_a^\star$ and the same condition as in the induced fit scenario holds.

1.5.3 Remark on the hypothesis of the theory

Note that the main assumption of the theory is the presence of a mechanism, a low-energy extended mode that describes the behaviour of the system resulting from the allosteric transition. As far as the mechanism resists to the introduction of temperature in the system, the theory stays valid — where instead of an energy landscape there will be a free-energy landscape.

1.5.4 Mechanical model of population shift allostery

The theoretical results discussed in the previous section can be tested in a mechanical model with the same properties of population shift allostery. In-silico evolution of allosteric materials lead to the design of architectures with soft extended regions where most of the strain is concentrated, as also found in few proteins [46] and discussed in a larger dataset in Chap. 2. If such a protein presents two states and the allosteric transition between them involves a conformational change, then the associated deformation will happen in the soft region, where at least two ways of properly organising the amino acids must exist. These assumptions help defining the following mechanical model where springs can have two different rest lengths.

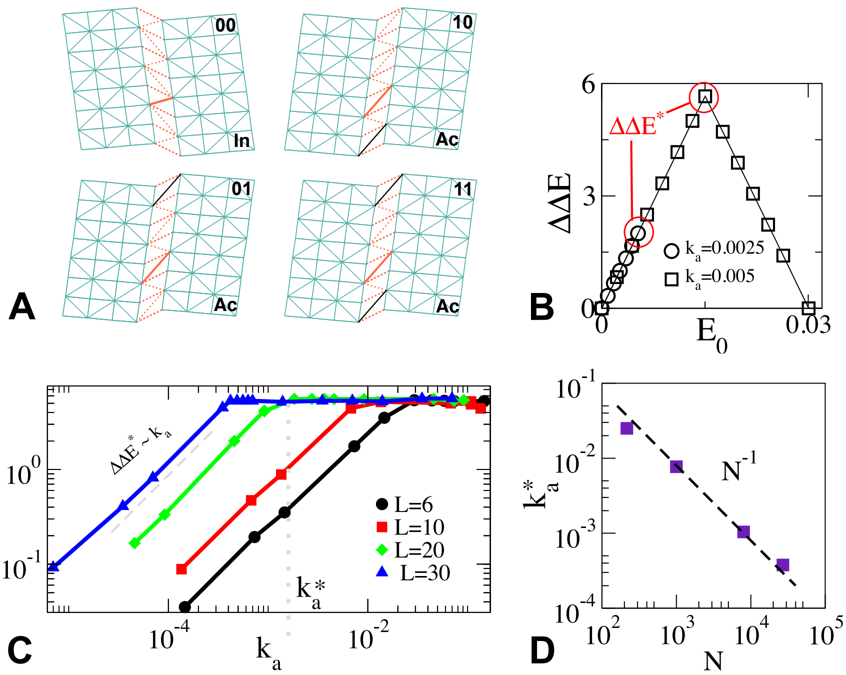

Figure 1.23 – (A) Sliced view of the three-dimensional mechanical model for population shift allostery. Rigid elastic regions made of harmonic springs (in blue) of stiffness k are connected through a weak nonlinear region made of nonlinear springs (in red) of stiffness k_w, with $k_w \ll k$. The inactive state (In) favors short springs, whereas the active state (Ac) favors long springs, as exemplified by a solid orange spring in the weak band. The location of the binding sites is represented by a solid black spring shown when the ligand is bound $(00, 10, 01, 11)$. (B) $\Delta\Delta E$ as function of E_0 for different values of k_a for a linear length $L = 20$ is shown, where $L^3 = N$. The same behaviour discussed in Fig. 1.21 is found. (C) Maximal cooperative energy $\Delta\Delta E^*$ as function of k_a for $L = 6, 10, 20$, and 30 is shown. The elbow in these curves defines the crossover stiffness $k_a^\star$, as in Fig. 1.22. (D) $k_a^\star$ versus number of nodes in the model N, supporting $k_a^\star \sim 1/N$.

Among the various architectures, the shear design can be picked as model where two three-dimensional rigid blocks of linear size $L^3 = N$ are connected through a soft planar region, which is easy to be deformed. The rigid regions are composed of harmonic springs of stiffness k, shown in blue in Fig. 1.23A, while the soft layer is made of anharmonic springs of stiffness $k_w \ll k$, represented as red in the same figure. Each of the anharmonic springs α follows a potential equivalent to Eq. 1.38 imposing that the springs have two different lengths when in the inactive and active states

$$E_w^\alpha = p_0^\alpha + p_1^\alpha x_\alpha + p_2^\alpha x_\alpha^2 + p_3^\alpha x_\alpha^3 + p_4^\alpha x_\alpha^4\,, \tag{1.45}$$

where $x_\alpha = l_\alpha / a$, with l_α the distance of two particles connected by α and a a parameter capturing the typical local deformation of adjacent particles between the inactive and active states. The prefactors can be chosen by imposing the following five conditions for all springs α

$$\left.\frac{dE_w^\alpha}{dx_\alpha}\right|_{x_{In}} = \left.\frac{dE_w^\alpha}{dx_\alpha}\right|_{x_{Ac}} = 0, \tag{1.46}$$

$$p_0^\alpha = 0, \tag{1.47}$$

$$p_4^\alpha = \frac{1}{2} k_w a^2, \tag{1.48}$$

$$p_1^\alpha (x_{In} - x_{Ac}) + p_2^\alpha (x_{In}^2 - x_{Ac}^2) + p_3^\alpha (x_{In}^3 - x_{Ac}^3) + p_4^\alpha (x_{In}^4 - x_{Ac}^4) = k_w a^2 b_w, \tag{1.49}$$

where b_w captures the energy difference between the inactive and active state. Indeed, the nonlinear springs present two stable extensions at which they exert no force and whose relative energy is controlled by the bias b_w. The material also presents two binding sites defined such that, upon binding, the distance between two nodes (represented with a black solid line in Fig. 1.23A) is the same as the one in the active state, thus favouring the material to be in the active configuration.

The energy profiles $E_{00}(x)$, $E_{10}(x)$ and $E_{11}(x)$ can be computed numerically as function of x which represents the motion along the shear mode. The energies are found by letting the elastic energy of the material relax after a shear displacement, i.e. a value of x, is imposed. The system relaxes along all possible directions apart the one of the shear mode and the ones corresponding to translations and rotations of the rigid blocks, by setting the force gradient in the minimisation procedure orthogonal to these modes. The energy profiles are shown in Fig. 1.22 for three different values of $k_a/k_a^\star$, denoting their differences as the stiffness of the shear mode changes. The relevant parameters discussed above ($k_a, E_0, \Delta E$ and $\Delta\Delta E$) can be directly computed from the profiles. The stiffness k_a can be found from fitting Eq. 1.38 to $E_{00}(x)$ while measuring $||d\mathbf{R}_a||^2$ and can be increased by varying the value of k_w. The energy difference E_0 can be also computed directly from $E_{00}(x)$ and can be changed by tuning the value of the bias b_w. The binding cost ΔE and cooperative fitness $\Delta\Delta E$ can be extracted by computing the minima of $E_{00}(x)$, $E_{10}(x)$ and $E_{11}(x)$. The results are shown in Fig. 1.23. Fig. 1.23B shows cooperative energy $\Delta\Delta E$ as function of E_0 for two values of k_a. As expected from Fig. 1.21, when k_a is large enough $\Delta\Delta E$ reaches the maximal $\Delta\Delta E^* = \Delta E$, while, when k_a is smaller, $\Delta\Delta E^*$ is fixed by the largest achievable energy difference E_0. Fig. 1.23C confirms the behaviour of the maximal cooperative fitness $\Delta\Delta E^*$ upon changing k_a for different system sizes N. Fig. 1.23D shows that the predicted scaling $k_a^\star \sim 1/N$ matches the scaling of the measured $k_a^\star$ from the data of Fig. 1.23C.

1.5.5 Conclusions

The insights from mechanical materials endowed with allosteric behaviour show that the cooperative energy depends on the stiffness of the low energy mode responsible for the allosteric transition. In an induced-fit-like scenario, cooperative fitness reaches an optimum for a given value of stiffness $k_a^\star$, whose scalings with the size of the system N is derived, $k_a^\star \sim 1/N$, see Fig. 1.19. The formalism is then extended to describe the two-states switch dynamics typical of the population shift scenario. Also in this context, cooperative fitness has a precise dependence on the stiffness, it displays a plateau after a value $k_a^\star$ that is found to have the same scaling with N as in the induced fit scenario. Even if the cooperative fitness displays a plateau at $k_a^\star$ and not an optimum, it does not mean that there is no optimality. Indeed, when $k_a \gg k_a^\star$ the transition between the inactive and active state becomes very slow given to very high energy barriers. Hence, the system is more likely to be at $k_a \sim k_a^\star$ also in the population shift scenario. Again, these results suggest that larger proteins display softer modes to accomplish allosteric cooperative binding, prediction tested in Chapter 2. Moreover, the presence of plateau of cooperative energy after $k_a^\star$ suggests that the effect of mutations that increase the stiffness of the low energy mode affect fitness by a small amount, in contrast to the induced fit scenario, where an optimum is displayed at the same scaling of $k_a^\star$. This prediction could also be tested in experiments, as briefly elaborated in Chapter 2.

Overall, these results identify the stiffness of the low energy mode as a significant observable to be monitored in order to study cooperative binding in proteins. Do protein actually maximise the cooperative energy by developing a soft mode with a stiffness whose magnitude scales inversely with the number of residues? Do extended soft modes systematically exist in proteins? Do they act in regions where strain deformation is the largest? In the next chapter, Chap. 2, structural data of 34 allosteric proteins are used to explore these questions and the necessary tools are introduced alongside.

2 Empirical results

High resolution structures can be easily found on online databases allowing to access elasticity via Elastic Network Models (ENM) and study their conformational changes. However, structures of allosteric proteins are more difficult to identify among the plethora of entries in the Protein Data Bank (PDB). We built a dataset of 34 allosteric proteins taken from previous studies on allosteric motions to test the results of the evolution of allosteric binding in in-silico models, discussed in Chapter 1.

An analysis over 34 allosteric proteins systematically shows that the allosteric response is mostly described by one low energy mode when it involves a large amount of amino acids, extending the results already present in the ENM literature e.g. [56], as reported in both the Introduction and Chapter 1.

The main goal is to perform a first test of the outcomes of the in-silico evolution of allosteric materials under both tasks. Do some proteins make use of the novel architecture for long-range communication found when optimising for the geometric task? Do allosteric protein work at optimality as cooperative architectures do?

In this chapter, first we provide preliminary evidence that some proteins present an amplification of the allosteric response — in particular the PDZ domain whose allosteric response is not well captured by a simple hinge or shear motion, as summarised in Fig. 5 in the Introduction. It suggests that levers may be present in proteins — a novel architecture that emerged from the evolution of the in-silico scheme presented in this thesis and discussed in Chapter 1.

Next, we build a database of 34 allosteric proteins and confirm that many indeed appear to display a mechanism that governs most of the allosteric response. We find empirical support that bigger protein needs a lower frequency mechanism to function, in consistence with the idea that they have evolved to present an architecture with strong cooperative binding.

The results presented in Sec. 2.4 are published in [157], while the preliminary discussion in Sec. 2.3 is unpublished.

2.1 Describing deformation: the strain tensor

The empirical understanding of allosteric communication from a mechanical standpoint involves the measure of the relative structural changes inside the material, whether this is a protein or an elastic network. In continous mechanics, strain is defined to describe such changes by removing rigid translations and rotations so that the structural deformation is well captured. Suppose that the position is changed from $\mathbf{x}$ to $\mathbf{y}$ after a motion Ω, $\mathbf{y} = \Omega(\mathbf{x})$. A tensor mapping the initial position $\mathbf{x}$ to the gradient of the motion, the deformation gradient, is defined as

$$\mathbf{F} = \frac{\partial \Omega}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}+\mathbf{u})}{\partial \mathbf{x}} = \mathbf{I} + \frac{\partial \mathbf{u}}{\partial \mathbf{x}}, \tag{2.1}$$

where $\mathbf{u}$ is the displacement field $\mathbf{u} = \mathbf{y} - \mathbf{x}$ and $\mathbf{I}$ the identity matrix. In this way, a rigid translation will not contribute to $\mathbf{F}$. What about rigid rotations? The deformation gradient can be decomposed in the product of two tensors [198], an orthogonal tensor $\mathbf{R}$ corresponding to a rotation and a positive definite symmetric one $\mathbf{V}$ (or $\mathbf{U}$)

$$\mathbf{F} = \mathbf{RU} = \mathbf{VR}. \tag{2.2}$$

Hence, the deformation gradient is seen as an application of a rigid rotation followed by a stretch along a set of orthogonal axes, or viceversa. It appears now clear that given the orthogonality of $\mathbf{R}$, the deformation tensor built as $\mathbf{F}^t\mathbf{F}$ ($\mathbf{F}\mathbf{F}^t$) does not contain rigid rotations. This deformation tensor naturally enters the definition of the strain tensor found by looking at the difference between the differential change between two points in the initial — $d\mathbf{X}$ — and deformed — $d\mathbf{x}$ — conformation. Choosing the material coordinates — the initial state — as frame of reference, this difference comes out to be

$$d\mathbf{x} = \mathbf{F}d\mathbf{X} \tag{2.3}$$

$$d\mathbf{x}^2 = d\mathbf{x}\cdot d\mathbf{x} = d\mathbf{X}\cdot \mathbf{F}^t\mathbf{F}\cdot d\mathbf{X} \tag{2.4}$$

$$d\mathbf{x}^2 - d\mathbf{X}^2 = d\mathbf{X}\cdot(\mathbf{F}^t\mathbf{F} - \mathbf{I})\cdot d\mathbf{X}, \tag{2.5}$$

defining the Cauchy-Green strain tensor as

$$\boldsymbol{\epsilon} \equiv \frac{1}{2}\left(\mathbf{F}^t\mathbf{F} - \mathbf{I}\right) \tag{2.6}$$

$$\epsilon_{ab} = \frac{1}{2}\left(\frac{\partial \mathbf{x}}{\partial X_a}\cdot\frac{\partial \mathbf{x}}{\partial X_b} - \delta_{ab}\right) \tag{2.7}$$

$$= \frac{1}{2}\left(\frac{\partial u_a}{\partial x_b} + \frac{\partial u_b}{\partial x_a} + \frac{\partial u_c}{\partial x_a}\frac{\partial u_c}{\partial x_b}\right) \approx \frac{1}{2}\left(\frac{\partial u_a}{\partial x_b} + \frac{\partial u_b}{\partial x_a}\right), \tag{2.8}$$

where a, b label the spatial dimension and the approximation in the last line supposes small deformations. The strain tensor can be then directly computed from any displacement field $\mathbf{u}$. However, the mechanics of proteins is not continous, but discrete, leading to a strain tensor that is defined for each component or particle i, $\boldsymbol{\epsilon}(i)$. The discrete version of the strain tensor is computed in [199] and discussed in Appendix B.1.

The strain tensor contains information about both components regarding volume expansions or contractions — bulk strain — and arising from relative motions of two volumes — shear strain. Proteins cannot be significantly compressed and expand only in the case of unfolding. Hence, the relevant deformation to be measured is linked to the relative motion of adjacent residues. First, the diagonal components of the strain tensor ϵ_{aa} are the ones associated to relative volume changes [149]. Hence, the strain tensor can be decomposed into a sum of pure shear and compression as

$$\epsilon_{ab} = \left(\epsilon_{ab} - \frac{1}{3}\delta_{ab}\epsilon_{cc}\right) + \frac{1}{3}\delta_{ab}\epsilon_{cc}, \tag{2.9}$$

where in the first term in brackets the volume component of $\boldsymbol{\epsilon}$ is substracted. From this last result follows that the shear tensor — that is not sensible to compression or dilation — is defined as

$$\boldsymbol{\gamma} = \boldsymbol{\epsilon} - \left(\frac{1}{3}\mathrm{Tr}(\boldsymbol{\epsilon})\right)\mathbf{I}. \tag{2.10}$$

A scalar quantity describing shear deformation can be defined for each component of the discrete system — residues in the case of a protein — called shear pseudo-energy, E_{sh}. For the component i, this linear elastic energy reads

$$E_{sh}(i) = \frac{1}{2}\sum_{a,b=1}^{3}[\gamma_{ab}(i)]^2. \tag{2.11}$$

Similarily, also a bulk pseudo-energy can be computed, describing volume changes

$$E_{bulk}(i) = \frac{1}{2}\sum_{c=1}^{3}[\epsilon_{cc}(i)]^2. \tag{2.12}$$

2.2 The allosteric response

The allosteric response is defined as the displacement between the inactive and active conformation of the considered protein, where in the inactive state no ligand is bound at the allosteric site, while in the active state the ligand is bound there. Given the two conformations, a structural alignment needs to be performed before computing the displacement. The rotations and translations that minimise the difference between the squared distances among the structures are found via a least-squared fit. The algorithm used to perform this conformation superposition is from the software Pymol [14] and performs also a sequence alignment whose effect is negligible in the case of structures of the same protein as it is the case here. Once the inactive and active structures are aligned, the allosteric response $d\mathbf{R}_a$ is directly computed as the resulting displacement.

In the case were allostery is accompanied by a non trivial conformational change the allosteric response gives useful information about the underlying mechanisms. The following observables can be directly computed from $d\mathbf{R}_a$: (i) its magnitude $||d\mathbf{R}_a||^2$ or its normalised counterpart; (ii) the fraction of the residues involved in the allosteric response and (iii) the shear pseudo-energy E_{sh} as defined in Eq. 2.11. The participation ratio estimates how much the allosteric response is extended or localised — point (ii) — and it reads [200]

$$P = \frac{||d\mathbf{R}_a||^2}{N\sum_{i=1}^{N}||d\mathbf{R}_a(i)||^4}. \tag{2.13}$$

2.3 Amplification of allosteric response in a PDZ domain

The optimisation of the geometric fitness function Eq. 1.5 lead to an unusual principle for allosteric design where the magnitude of the allosteric response is non monotonic in the directions perpendicular to the axis connecting the allosteric and active sites. Hence, the question of looking for such a mechanism in allosteric proteins arises naturally. A putative candidate is the third PDZ domain of the synaptic protein PSD95 from the rat (Rattus norvegicus) brain studied in the Ranganathan lab (UChicago). Indeed, Fig. 5 in the Introduction argues that the PDZ domain displays allosteric behaviour, while its architecture is unclassified.

Fig. 2.1A shows the tertiary structure of PSD95$^{\text{pdz3}}$ where some positions are highlighted. The $\beta1-\beta2$ loop is where the carboxylate terminal (t-COOH) of the peptide ligand binds, creating a direct contact between the peptide and the protein. The active site is located in the pocket formed between the $\alpha2$ helix and the $\beta2$ sheet, where the peptide ligand binds. Allosteric effects are known to emerge from the $\beta2-\beta3$ loop [201] and $\alpha1-\beta4$ loop [103] controlling the function of the binding site, more specifically which class of ligands can bind there. Upon mutation of the unligated wild-type at a position near the $\beta2-\beta3$ loop, G330, the protein results to be able to bind one additional class of peptide ligands. The surprising result from the Ranganathan lab is that this mutation elicits a response at the $\beta1-\beta2$ loop in the absence of the peptide ligand, showing a precise long-range effect of the allosteric behaviour modulated by the $\beta2-\beta3$ loop.

In this case, the allosteric response can be computed as the displacement between the wild-type structure and the mutated one, both in the absence of ligands. The normalised magnitude of the allosteric response $||d\mathbf{R}_a||^2/\sum_i||d\mathbf{R}_a(i)||^2$ shows the presence of non-monotonic behaviour when looked in the plan orthogonal to the axis connecting the allosteric and active site chosen to be parallel to the y axis — see Fig. 2.1A-B. This result shows that amplification from the allosteric to the active site in the magnitude of the allosteric response occurs in PSD95$^{\text{pdz3}}$ upon the introduction of a mutation in the area nearby the allosteric site. Furthermore, this suggests that a mechanism like the one found in geometric mechanical networks, Fig. 1.6, could be at work in this PDZ domain. Information on the coordination in the protein would strengthen this claim.

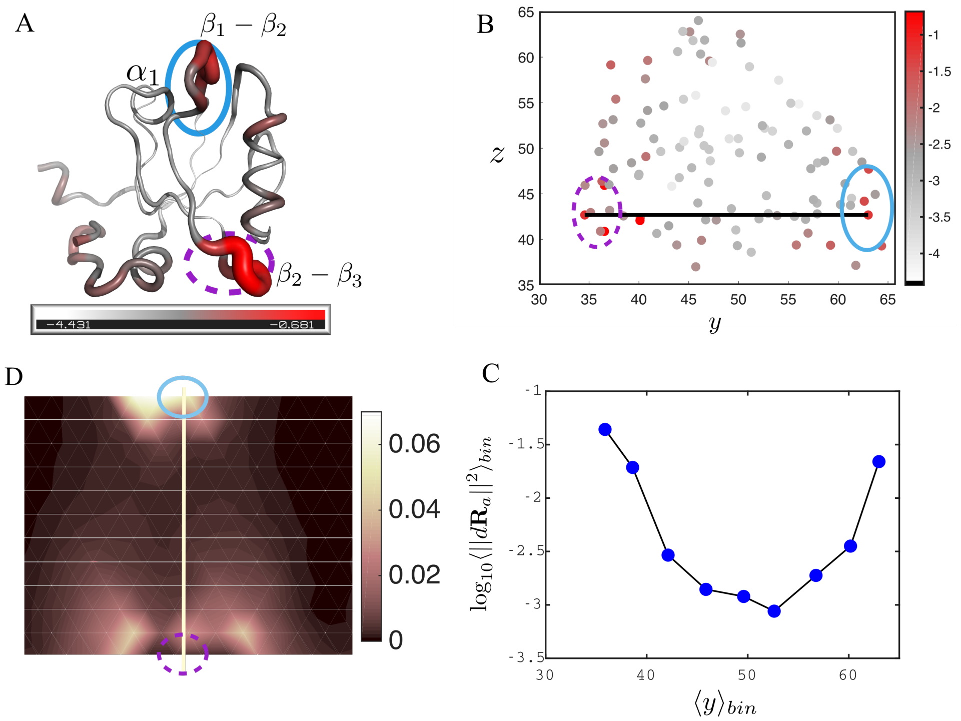

Figure 2.1 – (A) The tertiary structure of PSD95$^{\text{pdz3}}$ — a PDZ domain from the rat brain — is shown as a cartoon and coloured according to the magnitude of the allosteric response $||d\mathbf{R}_a||^2/\sum_i ||d\mathbf{R}_a(i)||^2$ in log scale. (B) A projection of the PDZ domain in the (z, y) plane, the same of figure A, is shown. The points represent the residues and are distinguished only by the colour coding for $||d\mathbf{R}_a||^2/\sum_i ||d\mathbf{R}_a(i)||^2$, in log scale. A line connecting the allosteric and active sites is shown in black, which is parallel to the y axis in the chosen projection. (C) The average magnitude of the displacement in the plane orthogonal to the y axis is shown as function of the average value of y, where the average is done over values in each of the chosen bins. This measure suggests that there is a non monotonic behaviour of the magnitude of the response from the allosteric to the active site, as found in geometric mechanical networks, see Fig. 1.6. (D) The magnitude of the allosteric response in the trumpet design Fig. 1.6 is reported for comparison.

2.4 Testing optimal cooperativity in 34 allosteric proteins

The optimisation of the cooperative fitness function Eq. 3.1 lead to architectures well known in protein mechanics, like the ones implementing hinge or shear motions. These architectures provide the allosteric functionality via an extended normal mode with low energy whose stiffness is not random, but has a precise scaling with the size of the system, let it be either the number of nodes in the network or the number of residues in the protein, which is reported in Eq. 1.24. This scaling ensures that the system is working at the point where cooperative energy is maximal, see Fig. 1.19 and Fig. 1.22. ENMs give a simple framework where to compute normal modes in proteins and investigate these findings.

To this scope, allosteric proteins for which the structure both in the inactive and active configuration is known need to be identified. A database of 34 allosteric proteins is constructed and reported in Tab. 2.1 with the respective PDB identifiers taken from [44,46] complementing the datasets in [56,58]. The set is diverse in functionality and includes enzymes (13), G-proteins (10), kinases (3), response regulators (3), DNA-binding proteins (4), and the human serum albumin. Moreover, 12 protein complexes are also present.

Protein type	Inactive (PDB)	Active (PDB)	Common residues	Chains	Overlap	Participation ratio
1. arf6	1E0S	2J5X	160	A	0.27	0.16
2. cdc42	1AN0	1NF3	183	AB	0.33	0.09
3. rab11	1OIV	1OIW	162	A	0.32	0.07
4. rac1	1HH4	1MH1	175	A	0.27	0.07
5. ras	4Q21	6Q21	164	A	0.44	0.08
6. rheb	1XTQ	1XTS	165	A	0.35	0.1
7. rhoA	1FTN	1A2B	173	A	0.24	0.08
8. sec4	1G16	1G17	152	A	0.23	0.08
9. IGF-1R	1P40	1K3A	283	A	0.23	0.03
10. met repressor	1CMB	1CMA	204	AB	0.24	0.08
11. tet repressor	2TRT	1QPI	190	A	0.47	0.32
12. glcN-6-P deaminase	1CD5	1HOT	262	A	0.45	0.26
13. EF-Tu	1TUI	1EFT	393	A	0.61	0.30
14. $G_{t\alpha}$	1TAG	1TND	310	A	0.36	0.07
15. ERK2	1ERK	2ERK	347	A	0.19	0.06
16. IRK	1IRK	1IR3	296	A	0.31	0.05
17. lac repressor	1TLF	1EFA	536	AB	0.71	0.48
18. PurR	1DBQ	1WET	271	A	0.71	0.34
19. anthranilate synthase	1I7S	1I7Q	1402	ABCD	0.61	0.45
20. chorismate mutase	2CSM	1CSM	241	A	0.54	0.25
21. FBPase-1	1EYJ	1EYI	323	A	0.46	0.04
22. phosphofructokinase	6PFK	4PFK	315	A	0.32	0.05
23. PTB1B	1T48	1PTY	287	A	0.48	0.02
24. ATCase*	6AT1	8AT1	2732	ABCDEFGHIJKL	0.64	0.55
25. hemoglobin	4HHB	1HHO	570	ACBD	0.66	0.52
26. NAD-malic enzyme	1QR6	1PJ2	2144	ABCD	0.72	0.57
27. phosphoglycerate DH	1PSD	1YBA	1580	ABCD	0.74	0.53
28. human serum albinum*	1E78	2BXB	574	A	0.36	0.09
29. fixJ	1DBW	1D5W	238	AB	0.69	0.65
30. DAHP synthase	1KFL	1N8F	1340	ABCD	0.64	0.55
31. SpoIIAA	1H4Y	1H4X	106	A	0.40	0.22
32. CheY	3CHY	1FQW	124	A	0.34	0.18
33. glycogen phosphorylase	1GPB	7GPB	1626	AB	0.53	0.24
34. ATCase	1RAC	1D09	2774	ABCDEFGHIJKL	0.65	0.55

Table 2.1 – Table with the X-ray structures (PDB identifier) used in the analysis. The number of the common residues and of the considered chains are also reported, along with the participation ratio of the allosteric response — defined as in Eq. 2.13 — and the maximal overlap found between the allosteric response and the normal modes — defined as in Eq. 2.14. The PDB identifiers are taken from [44] and [46] (asterisk).

The analysis can be performed considering residues instead of atoms without loss of generality. A sequence alignment is performed among the two structures so that only the residues in common are kept, whose number is denoted as N and reported in Tab. 2.1. Moreover, to avoid uninteresting fluctuations, the first and last two residues are excluded from the analysis. The resulting structures are then aligned by conformation superposition — as discussed before — and the allosteric response can be readily obtained for all entries in Tab. 2.1.

The allosteric response is shown in Fig. 2.2A for the elongation factor Tu (1TUI-1EFT), where a hinge-like motion seems to occurr upon binding. Indeed, the protein is colour-coded according to the shear pseudo-energy E_{sh} — in log scale — showing that the external regions are undergoing a rigid counter-rotating motion, while the deformation is centered in the middle, like what would be expected in a hinge design, see Fig. 1.14 for comparison with the result in two-dimensional mechanical networks.

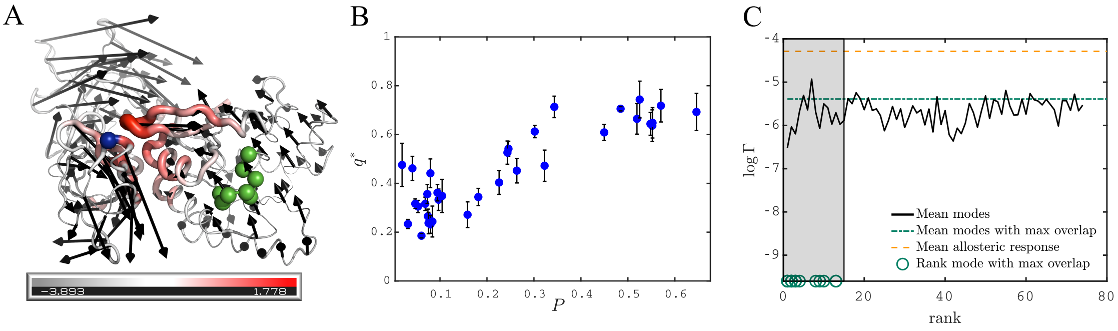

Figure 2.2 – (A) Allosteric response (black arrows) of elongation factor Tu corresponding to the displacement between the inactive state (where GDP is bound at the allosteric site) and active state (where GTP is bound at the allosteric site and the aminoacyl-tRNA can bind at the active site). The phosphate binding loop (allosteric site) is highlighted in green, while the active site is at the interface between the GDP binding domain and the two other domains, one residue of which is highlighted in blue [202]. The shear is encoded in both the color and the thickness of the structure in a logarithmic scale, red corresponds to large shear. The allosteric response is similar to that of a hinge. (B) Maximal overlap q^* as function of the participation ratio P. (C) The observable $\log\Gamma$ quantifying how modes are both extended and present large shear energy (see main text for definition) averaged over the proteins with overlap larger than 45%, as function of the mode rank for (i) the allosteric response , (ii) the modes with largest overlap and (iii) the first 75 modes (having subtracted the one with largest overlap) as indicated in legend. The green circles correspond to the rank of the mode with largest overlap. The shaded region highlights the range where these modes fall.

To study whether normal modes capture — at least partially — the allosteric behaviour, Hessians H are computed with the ENM formalism starting from the inactive structure at residue level. Results do not change if the computations are instead performed either with the active configurations or at atom level. As dicussed in the Introduction, the cutoff radius R_c used to define neighbouring residues is chosen consistently to have no more than six zero modes in the vibrational spectrum for each entry. To study the dependence of the results on the choice of R_c, the Hessians are built for 9 values of R_c equidistributed in the interval $R_c \in [8-12]$ Å. Mean and standard deviation of the observed quantities are then defined from the distribution of the observables defined from the different instances of R_c.

2.4.1 Principle 1: an extended low-energy mode well describes the allosteric response

Given a protein whose allosteric motion can be summarised as rigid blocks undergoing collective motion, the success of ENM is then proven by how much of the allosteric response is described by few normal modes, $\mathbf{v}_i$, by monitoring the maximal overlap between the allosteric response and normal modes. The overlap reads

$$q_i = \frac{|\langle d\mathbf{R}_a|\mathbf{v}_i\rangle|}{\sqrt{\langle d\mathbf{R}_a|d\mathbf{R}_a\rangle\langle\mathbf{v}_i|\mathbf{v}_i\rangle}}\,. \tag{2.14}$$

The measure of the maximal overlap $q^* \equiv \max_i q_i$ for the 34 allosteric proteins of Tab. 2.1 is shown in Fig. 2.2B as function of the participation ratio of the allosteric response. The results show that allostery indeed occurs mainly along one mode $\mathbf{v}^*$ for most of the proteins in the dataset — in half of the entries $q^* > 0.45$. Moreover, the allosteric response is more likely to be described by one mode if extended, i.e. if it involves a large number of residues, which is consistent with the assumption that ENM is more likely to work well when the allosteric behaviour is a collective motion of rigid blocks.

2.4.2 Principle 2: the allosteric response consists in extended regions of large shear

The evolution of cooperative mechanical networks shows that the allosteric response is characterised by extended regions of large shear deformation. Moreover, in [46], Leibler and coworkers show that shear planes naturally appear when looking at the strain associated with the allosteric response from X-ray structural data of some allosteric proteins. When looking at the vibrational spectrum, some modes are expected to be extended — as plane-wave-like modes — and others to bring a considerable amount of shear — as localised modes. Do normal modes capture the special feature of the allosteric response of being extended and associated to a large shear at the same time?

An observable combining information on both the participation ratio and the shear pseudo-energy can be introduced to systematically assess the unicity of the allosteric response over the normal modes to present both aspects. The following quantity can be defined from any displacement field

$$\log\Gamma \equiv \left[\gamma \log_{10}(\mathrm{P}) + \log_{10}(||E_{sh}||)\right]\,, \tag{2.15}$$

where $||E_{sh}||$ is the total magnitude of the shear energy, i.e. $||E_{sh}|| = (\sum_i E_{sh}(i)^2)^{1/2}$. $\log\Gamma$ is large if the displacement is extended and if the shear energy is large. The factor γ characterizes the trade-off between these two features. Here we choose $\gamma = 3.5$ reflecting the fact that for the considered vibrational modes, we find that P varies about 3.5 times less in relative terms than $||E_{sh}||$ as shown in Fig. 2.3. Thus for $\gamma = 3.5$, the spatial extension and the amount of shear equally affect $\log\Gamma$. Note that the value of γ found for the vibrational modes is equally good for compensating the values of the allosteric response. Fig. 2.2C shows $\log\Gamma$ averaged over the 17 proteins with $q^* > 45\%$ for the allosteric response (yellow line), the mode with maximum overlap (blue line) and the first 75 vibrational modes (having subtracted the one with largest overlap) as function of the mode rank. We find that Γ is typically 160 times larger for the allosteric response than for vibrational modes, a very significant difference underlying

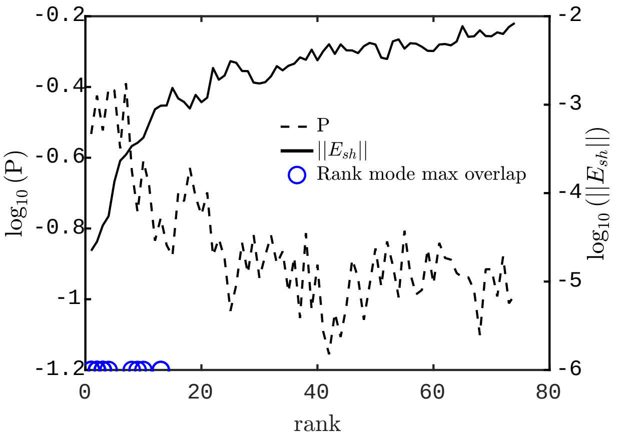

Figure 2.3 – The logarithm of the participation ratio P and of the norm of the shear pseudo-energy $||E_{sh}||$ is shown as function of the rank. The measured relative variation from the two curves is $\gamma^* \simeq 3.46$ and $\gamma = 3.5$ is chosen in Eq.2.15.

the specific geometry of the allosteric response. Generally, the modes identified to have a large overlap with the allosteric response result to be more localised than the response itself, hence the difference between the mean curves of the maximal modes $\mathbf{v}^*$ and the allosteric response — respectively green and yellow dashed lines in Fig. 2.2C.

2.4.3 Principle 3: scaling of the stiffness of the allosteric response

The precise scaling of the stiffness of the allosteric response poises the mechanical networks to optimally display cooperativity. Its scaling, $k_a^\star \sim 1/N$, can be directly tested with the dataset of Tab. 2.1. Indeed, given the Hessian matrix H, the stiffness k_a of the allosteric response $d\mathbf{R}_a$

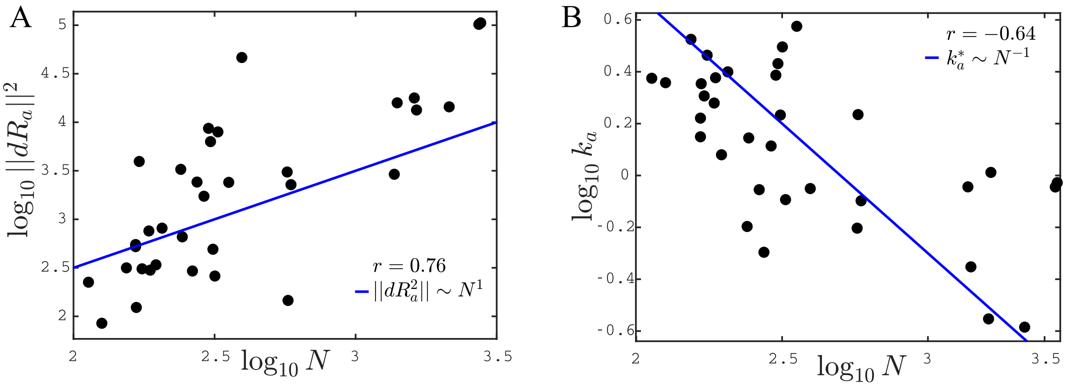

Figure 2.4 – (A) Logarithm of the square of the norm of the allosteric response, $||d\mathbf{R}_a||^2$, is shown as function of the logarithm of the number of residues, N. The solid line corresponds to $||d\mathbf{R}_a^2|| \sim N$. (B) The logarithm of the stiffness k_a as function of the logarithm of the number of residues N. The solid line represents the theoretical prediction $k_a^* \sim 1/N$. In both plots r indicates the Pearson coefficient.

can be estimated as the curvature of the elastic energy in that direction

$$k_a = \frac{\langle d\mathbf{R}_a | H | d\mathbf{R}_a \rangle}{||d\mathbf{R}_a||^2}\,. \tag{2.16}$$

In the analytical derivation of $k_a^\star \sim 1/N$, the scaling of the squared norm of the allosteric response as function of the number of residues was assumed to be linear, $||d\mathbf{R}_a||^2 \sim N$. Indeed, a larger protein is more likely to display an allosteric response involving more residues than a smaller one. The relationship can be directly tested with the dataset of Tab. 2.1: the 34 resulting values of $||d\mathbf{R}_a||^2$ are shown as function of N in Fig. 2.4A along with the line $||d\mathbf{R}_a||^2 = N$. There is a strong correlation between the logarithms of $||d\mathbf{R}_a||^2$ and N, as proven by the value of the Pearson coefficient $r = 0.76$ with p-value $p = 1.64 \times 10^{-7}$. Pearson coefficients are computed on the logarithmic values via the Matlab R2017b function *corr* and their value is consistent with the Spearman coefficients, for which a linear relationship between y and x is not supposed.

2.4.4 Summarising outcomes of the empirical testing

Fig. 2.2B provides systematic evidence that the allosteric response occurs along one soft elastic mode, while in Fig. 2.2C a novel observable Γ is introduced to establish that the allosteric response generally displays extended regions of high shear strain clearly recognisable from the normal modes. These observations support that for many proteins, elasticity is a useful starting point to describe allostery.

Furthermore, in Chapter 1 we have revisited the two classical thermodynamic models of allostery from this perspective, and provided a detailed study of how the energy profile along the soft mode evolves with binding. The results of Chapter 1 show that proper function is achieved if proteins evolve an elastic mode whose softness has to rapidly decrease with size and this prediction is supported by the anti-correlation observed between stiffness of the allosteric response and number of residues in Fig. 2.4B.

2.5 Conclusions

The numerical and theoretical study of allostery in mechanical networks provides principles for the transmission of the mechanical signal from the allosteric to the active site. The availability of several resolved structures of allosteric proteins in the inactive and active state allows a systematic study of these findings.

- **Geometric fitness: a novel architecture for propagating allostery**. The PDZ domain shown in the Introduction (Fig. 5C) not to conform to the classification of hinge and shear motions is found to display a surprising amplification of the response upon a mutation at its allosteric site. This amplification ressembles the trumpet architecture produced by in-silico evolution of geometric allosteric interactions optimising for strain between the allosteric and active site, as discussed in Chapter 1 (Sec. 1.3). Hence, the proposed edge mode mechanism could explain mechanically how such amplification is possible in the first place in the PDZ domain.

- **Cooperative fitness: cooperative binding is optimal**. One of the main result of Chapter 1 predicts a characteristic scaling of the stiffness of the allosteric response as function of the number of residues (Eq. 1.24) and can be directly tested from the X-ray sructures by computing the Hessian matrix with ENM. The result obtained in Fig. 2.4B is encouraging given the fair anti-correlation found between the measured stiffness and the number of residues, yet not definitive. Indeed, ENMs constitute a more than crude approximation for the energetics of protein interactions.

The results of these empirical analysis are encouraging, yet not conclusive.

- **Other features supporting amplification?**. The edge mode mechanism granting for amplification in the trumpet design is based on the presence of isostatic regions that are marginally connected, as shown by the coordination map in Fig. 1.6. Is it the case also in the PDZ domain? Moreover, are there other proteins, also in the database introduced in this chapter, that display such an amplification?

- **Testing optimal cooperativity on one protein**. The systematic study on a large dataset is crucial to test the predictions of the in-silico model, however it is based on a very simplifying and crude framework, namely ENM. Instead of looking at several protein, a precise study on *one* allosteric protein would clarify the validity of the findings. *Molecular dynamics* [111, 203, 204] would allow to accurately compute the vibrational modes and the stiffness of the allosteric response, while also monitoring how the energy profile along the allosteric response evolves with binding. Is the behaviour similar to the one of Fig. 1.20 or Fig. 1.22?

  Another way is through *mutation scans* [205], where cooperative fitness is measured via binding assays [206, 207], combined with *single molecule experiments* [208, 209] or *ultrafast laser pulses* [110,210,211] able to estimate the stiffness of the allosteric response to test the relationship between cooperative energy and stiffness. Is there an optimum? Or is there a small plateau? Indeed, in the mechanics of the population shift scenario, the cooperative binding energy is more robust to mutations that increase the stiffness of the soft mode, with respect to the induced fit scenario, compare Fig. 1.22 (population shift) with Fig. 1.18C (induced fit).

3 Benchmark of inference methods for the prediction of allosteric fitness

Fitness landscapes measure the mapping between fitness and sequence and are of utmost relevance in biology since they allow to access which are the effects of mutations on evolutionary paths. Thus, a model capable of sampling amino acid sequences in a way that they reproduce the fitness of the true landscape is of clear interest.

Direct Coupling Analysis (DCA) [81] has been proven successful in inferring evolutionary couplings that reflects contacts in the three-dimensional structure of proteins by using the information encoded by co-evolution in the Multiple Sequence Alignment (MSA) of protein families, see Fig. 10 in the Introduction. The success in contact prediction means that the DCA-inferred co-evolutionary couplings can be used to predict three-dimensional protein structures [94–96, 212] and to assemble protein complexes [213–215]. Moreover, DCA is proven successful also in the prediction of experimental results of deep mutational scans [66, 144, 146]. DCA is also found to generate sequences with the good value of fitness in the case of in-silico models of protein folding [138] — as shown in Fig. 14 in the Introduction. Note that a model depending only on conservation would fail in achieving the correct fold, underlying the relevance of pairwise interactions. The recent work in Ref. [148] shows that DCA is a good generative model for chorismate mutase enzymes reproducing the overall statistics of the MSA and finding that 45% of the designed sequences are comparabe with natural-like sequences.

Another successful inference method is Statistical Coupling Analysis (SCA) [99]. This method is able to identify groups of functionally related residues from the use of the covariance matrix weighted by conservation. The combination of conservation and correlations has been used to build in vitro synthetic sequences of a small protein domain — of around 40 amino acids — that fold as native ones, in contrast with the information contained solely in conservation that is not sufficient to achieve folding [105].

Hence, the success of DCA and SCA in capturing the evolutionary constraint in the MSA supports their use as models of the fitness landscape since it is this landscape itself that shaped the evolutionary pathways of proteins in the MSA of the family used in the DCA inference. However, DCA seems to fail in predicting long-range epistatic couplings [98, 134] that are experimentally observed in allosteric proteins where distant sites are functionally coupled [132, 134, 135]. Why does DCA seem to fail in predicting functional interactions? Moreover, its generative power has been thoroghly benchmarked only via an in-silico model built to mimick folding, thus with sequences displaying local constraints. Can DCA be a good generative model also for sequences that display more global, functional constraints?

In this chapter, these questions are addressed by using the in-silico sequences evolved to display cooperative allosteric binding introduced in Chapter 1. In particular, the shear design architecture of Fig. 1.13 is used.

In this context, epistasis, which reflects the coupling between different mutations, can be mechanically characterised, as achieved in Sec. 3.1. The characterisation shows that strong long-range epistatic interactions underlie the allosteric function proper of the in-silico sequences. Sec. 3.2 investigates the performance of DCA in (1) capturing single mutation cost, (2) being generative and (3) predicting the presence of long-range epistasis.

DCA well predicts single mutation costs, whose pattern of high cost mutation is located in the links associated to the functioning of the shear mechanism. However, the sequences generated according to the probability distribution inferred by DCA do not function as native ones, see Fig. 3.7. Hence, DCA is a poor generative model. Moreover, the DCA prediction of epistasis is found to be accurate for pairs of links close in the network, but fails dramatically for distant pairs, like the one involving the links in the functional regions of the allosteric and active site.

Sec. 3.3 proposes an explanation to this failure by introducing a toy model of allosteric function where several parts have to work in concert in order for the overall system to function, highlighting the hierarchical structure that underlies protein function, see Fig. 1 in the Introduction.

Sec. 3.4 focuses on the performance of the other co-evolutionary inference method discussed in the Introduction, Statistical Coupling Analysis (SCA) [100,101], in predicting such long-range epistasis. The observed absence of strong long-range correlations in the in-silico sequences, on which SCA is based, does not result in improved predictions.

In Sec. 3.5, concluding remarks suggest the possibility of using other models as fully-connected neural networks to improve the performance of DCA in capturing the more complex allosteric task.

The results presented in this chapter are published in [155, 158].

3.1 Mutation costs and epistasis in the in-silico model

The in-silico evolution model provides architectures that are designed to display long-range communication. In this chapter, the focus is on two-dimensional networks of size $L = 12$

evolved to optimise the cooperative fitness , which reads

$$\mathscr{F} = E_{10} + E_{01} - E_{11}\,, \tag{3.1}$$

and depends on the binding energy defined from the energy cost of imposing a displacement at the allosteric (E_{10}), active site (E_{01}) and both simultaneously (E_{11}), shown in colour in Fig. 3.1A. To better understand which functional constraints are involved in the cooperative fitness it is useful to consider the approximated form supposing a weak coupling between the two sites Eq. 1.7[1]

$$\mathscr{F} \approx \langle \mathbf{F}^{01} | d\mathbf{R}^{10\to Ac} \rangle\,. \tag{3.2}$$

This expression is an estimate of the change of mechanical work required for binding the substrate at the active site caused by binding the ligand at the allosteric site. Recall that $\mathbf{F}^{01}$ is the force exerted by the substrate when it binds at the active site and $d\mathbf{R}^{10\to Ac}$ is the displacement induced at the active site when the ligand binds at the allosteric site.

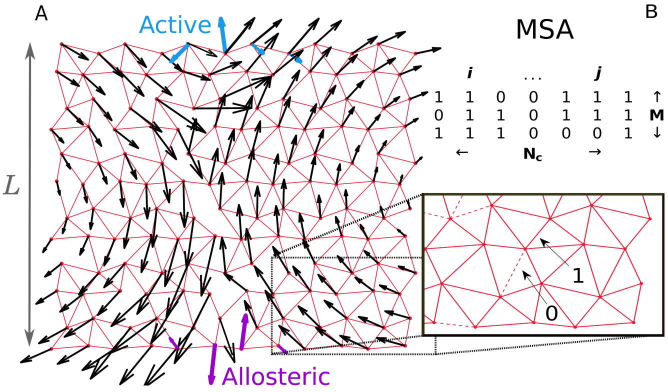

Figure 3.1 – (A) A solution of the in-silico evolution model optimising for cooperativity between the allosteric site (purple) and the active site (blue) is shown. The response to binding at the allosteric site is indicated via black arrows and it corresponds to a shear motion. (B) Each network can be mapped to a sequence of 0 and 1 coding for the occupancy of the springs. The in-silico scheme allows to generate a large number M of solutions, each corresponding to a slightly different shear architecture.

The boundary conditions of the elastic networks are set to be periodic so that the resulting architectures correspond to shear designs. The shear architecture is discussed in Chapter 1 (Fig. 1.13) where the emergence of a weakly-connected path between the allosteric and active site surrounded by rigid blocks is crucial to function. Indeed, this region corresponds to large shear deformation, which is also organised in a path connecting the two sites.

How to build a MSA comprising of sequences of different realisations of the shear design, as discussed for proteins in the Introduction? The resulting architectures of the thousands of evolved networks can be mapped into binary sequences encoding the presence or the absence

---

[1] The details are reported in the Appendix C.1.

of a spring at a given link, as illustrated in Fig. 3.1B. It is important to remark that, in this analogy, the role of an amino-acid in a protein MSA is played by a link, which can be occupied ($\sigma_i = 1$) or not ($\sigma_i = 0$).

The solutions, used in the following for the mechanical characterisation of epistasis, are obtained by sampling every 1000 time steps after an equilibration time of 10^5 steps at a high inverse temperature, $\beta = 10^4$, resulting in high fitness, typically of order $\mathscr{F} \sim 0.2$, see Fig. 1.11.

3.1.1 Characterisation of single mutation costs

Single mutations at a given link in the elastic network consist in changing its occupancy by adding or removing a spring. The resulting effect on fitness can be in principle either beneficial or detrimental. However, if single mutations are performed on evolved networks, their effect is expected to be positive given the high fitness of the configurations. The knowledge of single mutation is then of utmost importance to identify which positions along the sequence affect function the most.

The cost of a single mutation at link i is defined as

$$\Delta\mathscr{F}_i = \mathscr{F} - \mathscr{F}_i\,, \tag{3.3}$$

where $\mathscr{F}$ is the fitness of the original network and $\mathscr{F}_i$ the one of the network after changing the occupancy of link i, as discussed in the Introduction.

The result of single mutations on the shear architecture is discussed in Chapter 1 and reported here in Fig. 3.2, along with the measure of conservation. The path where the shear deformation is the strongest (Fig. 3.2A) turns out to be the one where the fitness cost of performing a single mutation of the spring occupancy is the largest, as shown in Fig. 3.2B. This region is also the most conserved (Fig. 3.2C), where conservation is computed according to Eq. 1.12. These results highlight that the functional region in the shear design identified from conservation and single mutations does not solely comprise of the neighbourhood of the binding sites, but extends to the whole path connecting them, in the view of its role in coupling the two.

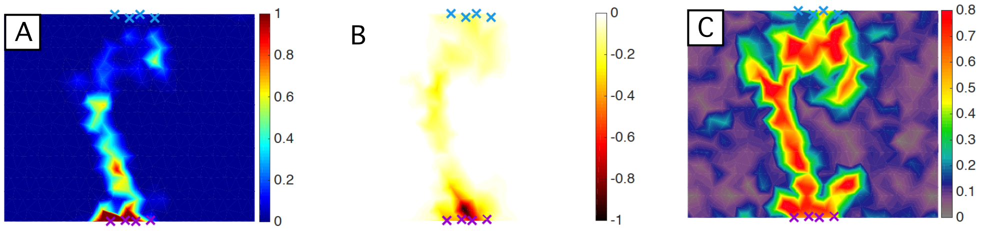

Figure 3.2 – (A) Map representing the average shear deformation, indeed happening along a linear path connecting the allosteric (violet) and active site (blue). (B) Map of the fitness cost of performing a single mutation in the network normalised as $\Delta\mathscr{F}_\rangle/\mathscr{F}$. (C) Map of conservation of spring occupancy.

3.1.2 Mechanical characterisation of epistasis

Epistasis measures the non-additivity of mutations and higher order mutation costs are needed to access it. In this analysis, the focus is on second order epistasis. Hence, only the fitness cost of a double mutation is needed. The cost of a simultaneous mutation at links i and j reads

$$\Delta\mathscr{F}_{ij} = \mathscr{F} - \mathscr{F}_{ij}, \tag{3.4}$$

and epistasis directly follows

$$\Delta\Delta\mathscr{F}_{ij} \equiv \Delta\mathscr{F}_{ij} - \Delta\mathscr{F}_{i} - \Delta\mathscr{F}_{j}. \tag{3.5}$$

The characterisation of how mutations affect function — a very hard task for proteins as reviewed in the Introduction — becomes accessible with the knowledge of the fitness function. In the approximation of weak couplings, where the fitness follows Eq. 3.2, epistasis can be explicitely written as function of mechanical observables: the force exerted by the substrate when it binds at the active site ($\mathbf{F}^{01}$) and the displacement induced at the active site when the ligand binds at the allosteric site ($d\mathbf{R}^{10\to Ac}$). Indeed, mutations in the networks have a significant effect when links responsible for the propagation of the signal $d\mathbf{R}^{10\to Ac}$ are involved.

It is found that the effect of single mutations is dominated by the changes of displacement at the active site, see Appendix C.4 for more details. Using this observation and starting from the approximated form of fitness in Eq. 3.2, epistasis is approximated as

$$\Delta\Delta\mathscr{F}_{ij} \approx -\langle \mathbf{F}^{10} | \left(\delta\boldsymbol{R}_{ij}^{10\to Ac} - \delta\boldsymbol{R}_{i}^{10\to Ac} - \delta\boldsymbol{R}_{j}^{10\to Ac} \right) \rangle, \tag{3.6}$$

where $\delta\mathbf{R}_i^{10\to Ac} = d\mathbf{R}_i^{10\to Ac} - d\mathbf{R}^{10\to Ac}$ with $d\mathbf{R}_i^{10\to Ac}$ the response at the active site when a ligand is bound at the allosteric site and link i has been mutated. $d\mathbf{R}_j^{10\to Ac}$ and $d\mathbf{R}_{ij}^{10\to Ac}$ are defined accordingly.

To monitor the effect of single mutations at different links i and j to the response at the active site it is useful to define the angle θ as the angle between $\delta\mathbf{R}_i^{10\to Ac}$ and $\delta\mathbf{R}_j^{10\to Ac}$. Indeed, if $\cos\theta = 1$, then only one mutation, the first, has an effect on the change of the response, while if $\cos\theta = -1$ the second mutation may counterbalance the negative effect on fitness of the first one.

In situations where the cost of a double mutation is dominated by the strongest point mutation, $\Delta\mathscr{F}_{ij} \approx \max(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j)$, epistasis can be simply rewritten as

$$\Delta\Delta\mathscr{F}_{ij} \approx -\min(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j). \tag{3.7}$$

The numerical exploration of epistasis clarifies the validity of these approximations in the sequences evolved by the in-silico model. The approximation resulting in Eq. 3.7 is found to

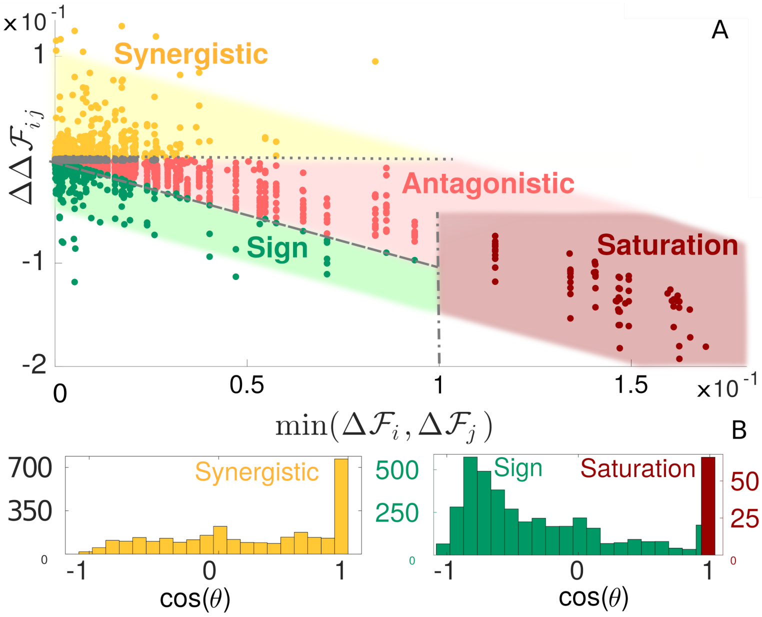

Figure 3.3 – (A) Phase diagram of epistasis in allosteric materials generated via the in-silico model. All quantities are averages over 50 configurations obtained in a single run. The shaded area is taken with arbitrary width and a -1 slope for clarity. Informative lines are shown. $\Delta\Delta\mathscr{F}_{ij} = 0$ (dotted style) corresponds to no epistasis and divides synergistic from antagonistic/sign epistasis. $\Delta\mathscr{F}_{ij} = \max(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j)$ (dashed style) separates sign and antagonistic epistasis and $\min(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j) = 0.1$ (dash-dotted style) is the threshold set to distinguish deleterious mutations, corresponding to the saturation region. Points in grey correspond to epistasis $< 5 \times 10^{-4}$ and are excluded from the analysis. (B) Histograms of $cos(\theta)$ for synergistic, sign and saturation epistasis.

capture the trend of measured epistasis for large values, as shown in Fig. 3.3A, with the dashed line indicating the assumption $\Delta\mathscr{F}_{ij} = \max(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j)$. No epistasis is found to correspond to purely additive mutations, as shown by the dotted line $\Delta\Delta\mathscr{F}_{ij} = 0$ in Fig. 3.3A. Moreover, epistasis between pairs of links can be classified in these different regimes

- **Saturation.** Mutations with $\Delta\mathscr{F} > 0.1$ and that correspond to a fitness loss of 50% are defined as deleterious. They represent $\sim 0.1\%$ of all pairs of mutations, in agreement with the sparsity observed in experimental measures as reported in [136]. Pairs of links with such deleterious mutations display the strongest epistasis in absolute value, and follow closely Eq. 3.7, as it appears from Fig. 3.3A. What do these epistatic interactions correspond to mechanically? The strength of epistasis can be reconducted to mutations that destroy signal propagation by themselves with $d\mathbf{R}_i^{10\to Ac} \approx d\mathbf{R}_j^{10\to Ac} \approx 0$, in such a way that the double mutation has the effect of a single one with $d\mathbf{R}_{ij}^{10\to Ac} \approx 0$. Hence, this kind of epistasis is classified as saturation. The measure of $\cos(\theta) \approx 1$ confirms this view as shown in Fig. 3.3B and naturally follows from $\delta\mathbf{R}_i^{10\to Ac} \approx \delta\mathbf{R}_j^{10\to Ac} \approx -d\mathbf{R}^{10\to Ac}$. This saturation regime can be seen as very high *diminishing-returns* epistasis[2] for which

---

[2] The use of the term diminishing-returns is borrowed from economics where adding an additional factor of

evidence from data and support from theoretical models are discussed in [216, 217].

- **Antagonistic.** Going up along the diagonal of Eq. 3.7 in Fig. 3.3A, the saturation effect becomes milder and changes to *antagonistic* epistasis [129, 218]. After a first mutation, making a second one results only in a weak additional change. Note that the double mutant is fitter than what would be expected from the additive case.

- **Sign.** In the intermediate range of mutation costs with $\min(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j) < 0.1$, the fitness cost of a deleterious mutation is diminished by the second mutation, i.e. $\Delta\mathscr{F}_{ij} < \max(\Delta\mathscr{F}_i, \Delta\mathscr{F}_j)$ leading to more compensatory epistatic interactions. Thus some mutations can increase the fitness and become beneficial in presence of another mutation, resembling the *sign* epistasis empirically detected [129, 219]. Mechanically, this compensatory effect orginates from situations where the two mutations deform the signal in opposite directions, so that the second one can partially re-establish fitness. Indeed, Fig. 3.3B shows that for sign epistasis $\cos(\theta)$ has a distribution peaked towards negative values.

- **Synergistic.** Positive-sign values of $\Delta\Delta\mathscr{F}_{ij}$ occur if two mutations perturb the elastic signal in the same direction, causing more damage than expected if they were purely additive and indicate *synergistic* epistasis. Indeed, Fig. 3.3B, shows that the distribution of $\cos(\theta)$ tends to be positive.

3.2 Overview of the performance of Direct Coupling Analysis

The role of mutations in the spring network has been discussed in details in the previous section, showing the presence of rich epistatic interactions. Are these mutation patterns captured by inference methods? The in-silico evolution scheme can be used to produce a large number of sequences needed to build the MSA necessary for any co-evolutionary analysis.

Spring configurations of the network can be mapped to binary sequences as summarised in Fig. 3.1B, where the binary variables encode for the occupancy of the corresponding link. To properly construct the MSA, $M = 135000$ configurations are evolved, each sampled from a different initial condition. This to avoid any bias that may originate from a high similarity in the sequences of the same run, differently than what done for the other analysis in this thesis. Correlations due to a common evolutionary history, known as phylogenetic effects, are thus avoided and no further ad hoc corrections in the inference procedure are needed to counterbalance these effects [69]. The similarity between sequences in the MSA can be measured by the average pairwise Hamming distance, which turns out to be $\sim 20\%$ of the length N_c. This similarity is consistent with what found in protein sequences, with the difference that the amino acid are the links between the nodes of the network and can then take two values, either zero or one according to their occupancy.

---

production results in smaller increases in output.

The inference method used for a statistical analysis of the MSA is DCA, which is based on fitting the measured single-site $\langle \sigma_i \rangle = 1/M \sum_m \sigma_i^m$ and pairwise $\langle \sigma_i \sigma_j \rangle = 1/M \sum_m \sigma_i^m \sigma_j^m$ frequencies of links by the probability distribution $P(\boldsymbol{\sigma})$ with maximal entropy, as discussed in the Introduction. This leads to

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\{(-\mathscr{E}(\boldsymbol{\sigma}))\} \tag{3.8}$$

$$\mathscr{E}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij} \sigma_i \sigma_j - \sum_i h_i \sigma_i , \tag{3.9}$$

where $\mathscr{E}$ is the inferred energy and is an estimate of $\beta \mathscr{F}$, with β the inverse evolution temperature. Since a measure of the proportionality between $\mathscr{E}$ and $\mathscr{F}$ is a sufficient requirement for the goodness of the inference procedure, the factor β is omitted when presenting the results.

As discussed in the Introduction, the fields h_i and couplings J_{ij} are inferred to match $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$. The scheme chosen to perform the inference is a combination of ACE (Adaptive Cluster Expansion) [89,90], an approximate technique developed from statistical physics ideas, combined with maximum likelihood, an exact technique, as reported in [93]. The resulting algorithm gives very accurate results and its performance is compared in the following with the more approximate mfDCA, as also discussed more in detail in Appendix C.3. Comparisons with other approximations are shown in a benchmark of DCA on protein lattices [138].

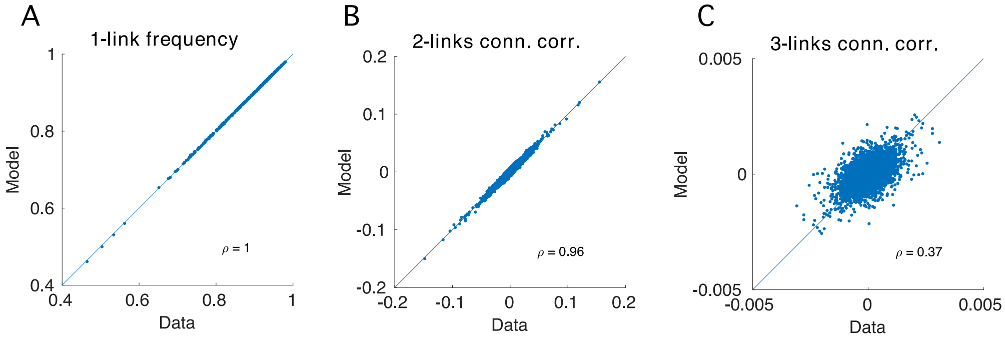

Figure 3.4 – Statistics of the model inferred by combining ACE and maximum likelihood. (A) single-link frequency and (B) pairwise-links connected correlations are very accurately reproduced, as they should by construction of the model. The relative errors, defined as in [90], are respectively $\epsilon_m = 2.45 \times 10^{-1}$ and $\epsilon_C = 1.30 \times 10^{-1}$). The third order connected correlations (C) are not constrained in the inference and are not well captured (Pearson correlation coefficient $\rho = 0.37$). This suggests that a pairwise probabilistic model of sequences is an approximation which becomes poor for estimating higher order moments.

The convergence of maximum entropy algorithms can be tested by looking at whether the single-site frequency and pairwise connected correrlations of links are reproduced as required by the construction of the model. Fig. 3.4A-B show the large correlation between true and inferred quantities in the case of ACE with maximum likelihood. Fig. 3.4C shows that the even this accurate DCA inference scheme is not able to capture third order correlations that are not constrained in the procedure. This result will be commented in the following when the generative power of the inference model is investigated. The details of the techniques are discussed in Appendix C.3.

All the tools are ready fo the benchmark of DCA in allosteric materials. The benchmark will focus on the ability of the inference method to (1) reproduce accurately the cost of single mutations, (2) generate new sequences with high fitness and (3) predict epistasis.

3.2.1 Prediction of mutation costs

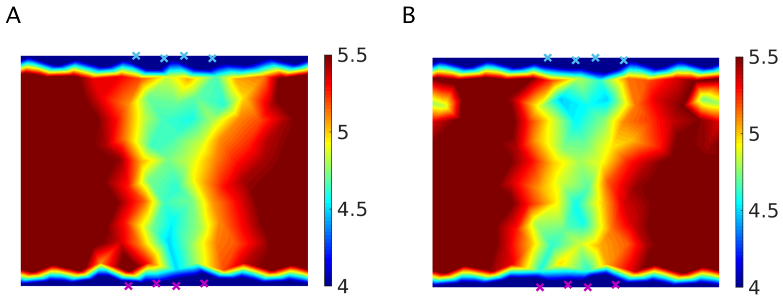

Figure 3.5 – Coordination map of original sequences (A) and generated ones (B) averaged over 50 solutions. They both exhibit a softer central path with coordination $z < 5$ joining active and allosteric sites, indicated respectively by blue and purple crosses, along which the shear-like sliding takes place. This path is embedded in a more connected, rigid region where the coordination is $z > 5$.

The cost of single mutations in the elastic networks displaying a shear design is reported in Fig. 1.14F and it is shown in Fig. 3.6A by colour-coding the links. The strongest mutation cost is located in the weakly-connected region which is indeed functionally important for the shear design, as elaborated in Sec. 1.4. The coordination map of the network that monitors the number of springs for each node in the network is shown in Fig. 3.5A, where the soft region is clearly visible. Fig. 3.5B shows that sequences generated with the inferred model Eq. 3.8 also reproduce the continous weakly-connected path between the allosteric and active site. Is the inference model also able to reproduce single mutation costs?

The inferred energy $\mathscr{E}$ allows to compute a prediction for the mutation cost at a link i $\Delta\mathscr{E}_i = \mathscr{E}_i - \mathscr{E}$. The result with the parameters inferred via DCA is shown in the map of Fig. 3.6B. A more quantitative comparison between the true and predicted mutation cost with a scatter plot finds a very high correlation between the two, see Fig. 3.6C. Hence, DCA is able to accurately predict single point mutation costs, improving the prediction via conservation alone that neglects the contribution of pairwise couplings. The poor performance of conservation is shown in the inset of Fig. 3.6C.

3.2.2 Generative power

The work discussed in [138] investigates the generative power of DCA in inferring the fitness of protein lattices that mimic folding. How does DCA performs in the case of the cooperative

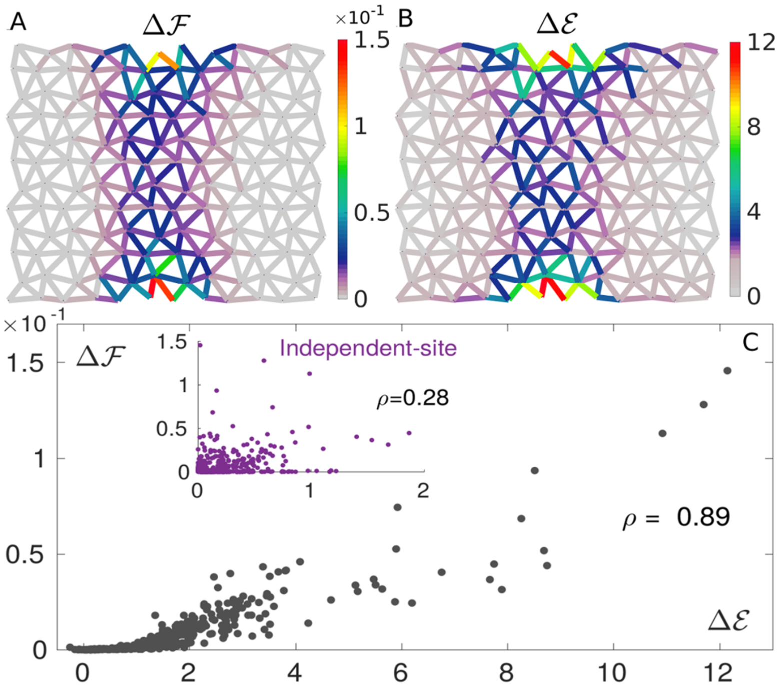

Figure 3.6 – Maps of true $\Delta\mathscr{F}$ (A) and DCA-inferred $\Delta\mathscr{E}$ (B) single mutation costs, averaged over 1.5×10^3 configurations randomly chosen from the MSA. Their patterns are very similar with high costs near the allosteric and active sites and in the shear path connecting them. (C) Scatter plot showing the strong correlation between $\Delta\mathscr{F}$ and $\Delta\mathscr{E}$ for all links, averaged over 1.5×10^3 configurations. The estimation of mutation costs based on an independent-site model — i.e. on conservation — correlates poorly with the true cost (inset), proving the need for incorporating correlations for proper prediction of mutation costs. The correlation between the two mutation costs is quantified via the Pearson correlation coefficient, ρ measured with the Matlab function *corrcoeff.*

architectures evolved to have an allosteric coupling? As discussed in the Introduction, eqs. 3.8 and 3.9 provide the model to generate sequences obeying the Boltzmann distribution which has as energy the inferred fitness $\mathscr{E}$. Jacquin and collaborators in [138] represented the fitness of sequences as function of their distance to the most representative sequence of the MSA, the consensus, where springs occupy the positions with largest mean occupancy, as overviewed in Fig. 14 in the Introduction. This representation allows to check at the same time for the prediction of both the distribution of fitness values and the variability of the MSA, quantified by the distance to consensus. Indeed, not only the prediction of fitness value is important for a good generative model, but also the ability to reproduce the variability of natural sequences. Fig. 3.7 shows the fitness of the original data, random data and sequences generated via ACE with maximum likelihood and mfDCA as function of the distance to consensus. The analysis shows that

- The sequences generated with the model inferred with ACE well reproduce the variability of the MSA, which is not the case for mfDCA whose variability is more close to the one of random sequences.

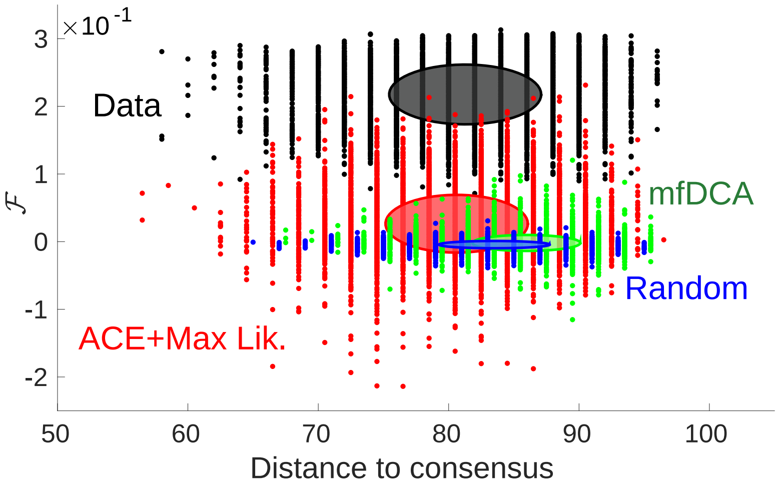

Figure 3.7 – The fitness as a function of distance to consensus of configurations generated by the inferred model is shown, following the representation of [138]. The sampling is done from the Boltzmann-Gibbs probability distribution $P(\boldsymbol{\sigma})$ of Eq. 3.8, whose parameters have been inferred via ACE with an additional maximum likelihood (red cloud) or mfDCA (green cloud). Original high fitness configurations (black cloud) and random ones (blue) are also shown as comparison. Each cloud consists of 10^4 sequences and the drawn ellipse gives one standard deviation around the mean in both horizontal and vertical directions. Distances to consensus of ACE + maximum likelihood, mfDCA and random sequences are shifted by respectively $+0.7$, -0.7 and -1.3 for better visibility.

- The distribution of the fitness of the sequences generated with ACE displays more variability than the random case, with some extreme values overlapping with the fitness of the data. This never occurs either for random data nor even for sequences generated with the more approximate mfDCA.
- However, the mean fitness is significantly lower than the one of the data, although larger than the mean fitness of random data, which is zero, and the one of mfDCA sequences that is slighly above zero.

These results suggest that the model inferred via DCA is not a good generative model when allostery is at play, contrasting its good performance in the case of protein lattices that are built to mimic protein folding [138]. Thus, a quadratic model accounting for conservation and correlations in the MSA, although it can capture some features of the shear design — as single mutation costs and the inhomogeneous distribution of coordination, as shown in Fig. 3.5 — is not enough for reproducing the allosteric fitness. Indeed, solutions sampled from the inferred energy landscape have the expected design but are not maximally fit. This suggests that structural components, as the distribution of links, are captured, but additional information is needed to reproduce a complex mechanical function as the cooperative fitness.

Moreover, higher order statistics as the third moment are not well reproduced, see Fig. 3.4, as instead it happens for protein structure predictions where local folding constraint are crucial as shown in [68]. This also suggests that the longer-range correlations induced by allostery might not be well captured by a pairwise model. The argument could be verified by checking whether in analysis like the one performed in [68] three-body correlations are captured even for amino acids that are far away in the structure. A study on a MSA made of solely allosteric proteins would also better study the performance of maximum entropy pairwise models in predicting allosteric features.

3.2.3 Prediction of epistasis

Another way to check whether the failure of DCA in generating fit sequences is due to the inability to capture allosteric interactions is through its performance in predicting epistasis. Indeed, as discussed in [134], epistasis is presumably large for functionally related sites, which

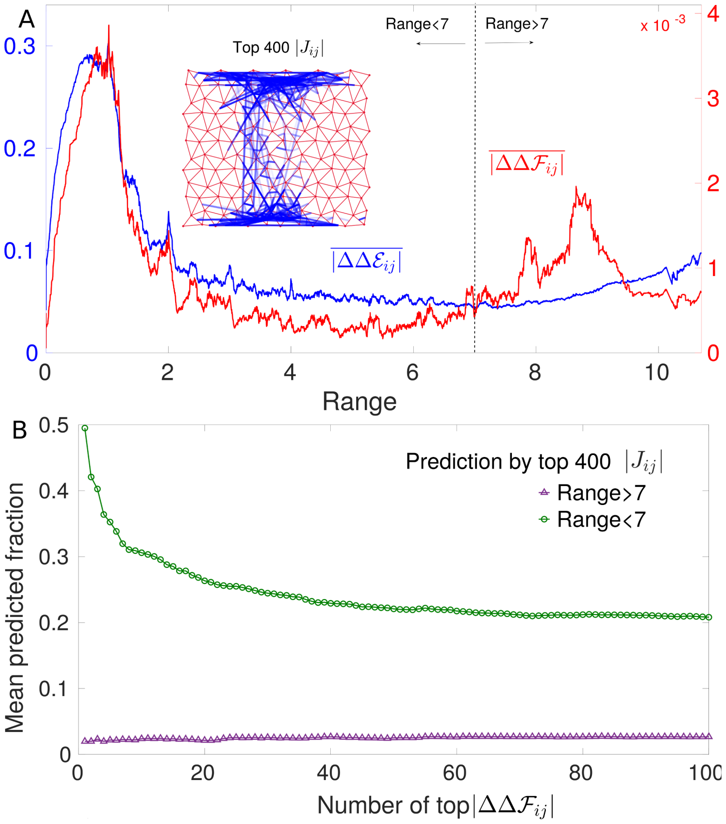

Figure 3.8 – (A) The running average of the absolute value of epistasis $\Delta\Delta\mathscr{F}_{ij}$ and of DCA prediction $\Delta\Delta\mathscr{E}_{ij}$ is shown for 1.5×10^3 configurations as a function of the distance between link i and j. The trends are well reproduced at short distances, while at long distance DCA underestimates epistasis. Inset: Top 400 inferred couplings. The couplings are mostly short range, with only a few long-range ones connecting the allosteric and the active site. (B) Assesment of the prediction of epistasis in single configurations by the top 400 couplings where long-range (> 7) and short-range (< 7) pairs of links are represented separately and ranked in terms of the epistasis magnitude $|\Delta\Delta\mathscr{F}_{ij}|$. The figure shows which fraction of these pairs — averaged over 100 randomly chosen configurations — belongs to the 400 largest couplings, as a function of the number of pairs with maximal epistasis considered. The random expectations for these mean predicted fractions are 0.0041 for short-range pairs and 0.0009 for long-range ones, which are both significantly lower than the values reported here. This result is robust also if the number of top couplings used in the prediction is increased, e.g. up to 1000.

in this case are also distant.

Firstly, the existence of strong long-range epistatic interactions in the MSA needs to be proven. Fig. 3.3A shows that pairs of links that both have a large single mutation cost systematically display strong epistasis, while Fig. 3.6A shows that links with a large mutation cost can be distant, as it happens for the neighbouring regions of allosteric and active site. These observations automatically imply that long-range and strong epistasis exists in cooperative allosteric sequences.

Can DCA identify such long-range epistatic interaction? From the functional form of the DCA-inferred energy, Eq. 3.9, the prediction for epistasis can be explicitely computed

$$\Delta\Delta\mathscr{E}_{ij} = -J_{ij}(2\sigma_i - 1)(2\sigma_j - 1) \tag{3.10}$$

$$|\Delta\Delta\mathscr{E}_{ij}| = |J_{ij}|, \tag{3.11}$$

showing that in a model inferred via DCA epistatic interactions simply reduce to the inferred co-evolutionary couplings.

The inset of Fig. 3.8A shows the top 400 couplings with highest magnitude as lines between each couple of links i and j. The couplings involve mostly links that are close in the network structure, with rare long-range couplings. This observation combined with Eq. 3.11 suggests that the predicted epistasis is very likely to fail to capture long-range interactions. The failure is evident when comparing the mean absolute epistasis $|\Delta\Delta\mathscr{F}_{ij}|$ and the mean absolutae DCA prediction $|\Delta\Delta\mathscr{E}_{ij}|$ as function of the distance between the pairs of links. The agreement between the predicted epistasis and the original one is accurate for short-distance links, while it becomes poor for long-range ones, where the strength of epistasis is strongly understimated by the predicted one, as shown in Fig. 3.8A.

This claim can be formalised by measuring the average fraction of pairs with the strongest values of epistasis that also fall in the list of the 400 pairs with largest couplings. A clear difference appears when this average fraction is looked at separately for long-range pairs — defined to have a range larger than 7 — and short-range ones, as shown in Fig. 3.8B. Long-range pairs have a much smaller average fraction than short-range ones, supporting the claim that the inferred model is not able to capture long-range epistasis. Even at short distance the prediction of epistasis by $|J_{ij}|$ is not excellent, the average fraction being ~ 0.5. The performance is significantly improved by considering epistasis averaged over several configurations, as done in Refs. [134, 136] and shown in Appendix C.2.

The poor performance of DCA for predicting long-range epistasis is in contrast with its success in contact prediction, where accurate algorithms like ACE are able to predict almost all true contacts [92,93]. How is it that DCA well captures local but not global constraints? As discussed in the Introduction, this failure is also documented in the empirical study of [98] showing that the majority of DCA-inferred long-range couplings are not related to functional dependencies.

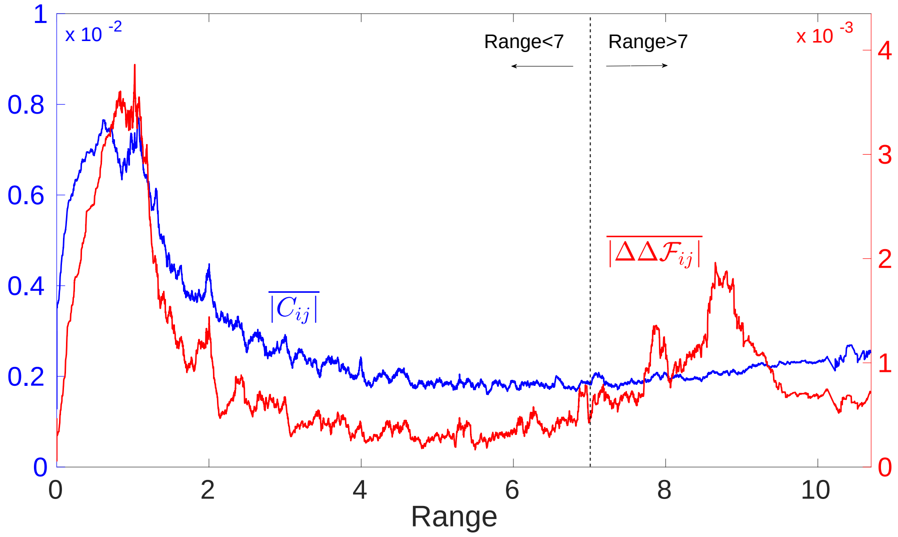

Figure 3.9 – Running average of the absolute value of connected correlations $C_{ij} = \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle$ and of epistasis $\Delta\Delta\mathscr{F}_{ij}$ for the same 1.5×10^3 configurations of Fig. 3.8A as a function of the distance between link i and j.

However, it is crucial to note that the sequences evolved with the in-silico model do not even display large long-range correlations — see Fig. 3.9 — that could still result from indirect interactions between networks of short-range couplings[3], see Fig. 9 in the Introduction. Thus, it is a more surprising result than the absence of strong long-range inferred couplings. The absence of long-range correlations suggests that other co-evolutionary approaches that use the principal components of the covariance matrix of the MSA, like SCA, are not likely to improve significantly the performance of DCA, as shown in Fig. 3.13. Finally, long-range couplings, as long-range correlations, are plausibly smaller than the ones at short-range, thus are more sensible to errors in the inference procedure. However, the discussion in the following paragraph suggests that there is something more conceptual beneath the poor performance of DCA in capturing long-range epistasis.

3.3 A possible answer to why DCA fails in the prediction of long-range couplings

Allosteric systems are characterised by a long-range communication that involves two binding sites or more and usually a specific architecture enabling such communication. As discussed in the Introduction, the local structure of some allosteric proteins is found to exhibit soft regions needed for the propagation of shear [46] implying the necessity of local constraints. On larger

---

[3] Another reason could explain the observation of long-range correlations. Indeed, it is interesting to remark that the long-range statistical couplings that are observed in the PDZ domain [99] may arise given that the covariance matrix used in SCA is conservation weighted and functional residues in the PDZ domain are found to be strongly conserved [119]

length scales, these regions are properly placed in the structure to enable an extended soft elastic mode [57, 58] that generates global constraints.

The shear design features a continous soft path in the structure connecting the allosteric and active site, which position can fluctuate from protein to protein in the same family. This is also the case for the sequences evolved with the in-silico model, where the position of the shear path varies from case to case. In the following, it is argued that the failure of DCA in predicting long-range functional dependencies originates from its inability to reproduce a function that requires many subparts of the system to work in concert — requiring global constraints — when each subpart can declinate in different types — implying local constraints. The hierarchical structure of this function reminds of the hierarchical three-dimensional organisation of proteins, shown in Fig. 1, where indeed several parts have to work in concert to make a protein complex function.

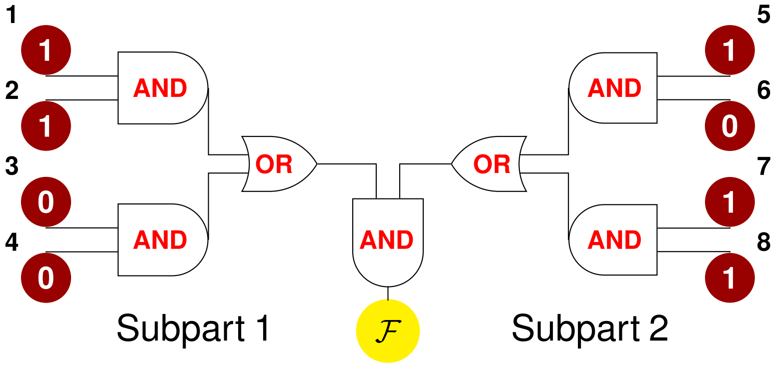

Figure 3.10 – A system is arranged into 2 subparts which must work together to accomplish a given function (AND gate). Each subpart is composed of 2 groups, i.e. can be of 2 types, (OR gate) and each type must satisfy some constraints to properly function (AND gate between single units).

A Boolean model, shown in Fig. 3.10, can be introduced as a toy model for a hyerarchical system displaying both global and local constraints to check the performance of DCA in predicting long-range epistatic interactions. An abstract task is performed by two subparts that must work in concert (AND gate) and that can be of two different types (OR gate), but such that each must be functional (AND gate). The boolean variables of the model are set to be eight each taking the value 0 or 1. The model is structured into four groups: two groups constist of the possible types of subpart 1 (left in Fig. 3.10) and the other two the possible types of subpart 2 (right). A configuration is functional if two variables of the same group are simultaneously in state 1 for each subpart.

The functional configurations can be enumerated and result to be 49, whose fitness is fixed to be $\mathscr{F}$, all other configurations are given a zero fitness. The value of $\mathscr{F}$ is assumed to be large so that the sequences in the MSA are only the 49 functional ones, with a uniform distribution. It is straightforward to calculate epistasis in this model, as well as single-site and pairwise frequencies from which couplings J_{ij} and fields h_i can be inferred. Hence, $\Delta\Delta\mathscr{F}_{ij}$ and $\Delta\Delta\mathscr{E}_{ij}$ can be compared for units i and j either (1) in the same group — or in the same subpart —

	$\lvert\Delta\Delta\mathscr{F}_{ij}\rvert$	C_{ij}	$\lvert\Delta\Delta\mathscr{E}_{ij}\rvert$	J_{ij}
Same group	1	0.061	0.51	1.18
Same subpart	0.33	−0.08	0.14	-1.01
Different subpart	0.43	0.00	0.07	0.40

Table 3.1 – Table summarizing true and predicted epistasis magnitude, $|\Delta\Delta\mathscr{F}_{ij}|$ and $|\Delta\Delta\mathscr{E}_{ij}|$, connected correlations C_{ij} and inferred couplings J_{ij} in the simple model for sites i and j in the same group, in the same subpart and in different subparts. For i and j in different subparts (third row) the non trivial magnitude of epistasis is not reflected in the values of correlations, and thus of the inferred couplings, so that it is underestimated by the DCA model. In Appendix C.5, an analytical derivation shows that $|\Delta\Delta\mathscr{F}_{ij}| = 21/49\mathscr{F}$ for i and j in the same group: since the prefactor $\mathscr{F}$ is not predicted, the choice $21/49\mathscr{F} = 1$ is without loss of generality. The numbers in the first column follow this definition.

which are then locally constrained by the defined function of the model, e.g. $i = 1$ and $j = 2$, or (2) in the two different subparts, thus globally constrained, e.g. $i = 1$ and $j = 5$.

By analytical manipulation discussed in detail in Appendix C.1 it can be shown that $|\Delta\Delta\mathscr{F}_{12}|/|\Delta\Delta\mathscr{F}_{15}| \approx 2.3$: global and local constraints lead to short range and long-range epistasis that is relatively similar in magnitude. Yet, long-range epistasis between subparts is significantly underestimated by DCA in contrast to short-range epistasis within subparts. This can be quantified by looking at the DCA prediction for the ratio of epistasis between two pairs of sites divided by the true ratio of epistasis. The epistasis of pairs of sites belonging to the same subpart is well predicted. As an example, the pair of sites $(1,2)$ and the pair $(1,3)$ display $|\Delta\Delta\mathscr{E}_{13}|/|\Delta\Delta\mathscr{E}_{12}| \times |\Delta\Delta\mathscr{F}_{12}|/|\Delta\Delta\mathscr{F}_{13}| \approx 0.86$. However, if the pair of sites belongs to different subparts, DCA strongly underestimates epistasis with $|\Delta\Delta\mathscr{E}_{15}|/|\Delta\Delta\mathscr{E}_{12}| \times |\Delta\Delta\mathscr{F}_{12}|/|\Delta\Delta\mathscr{F}_{15}| \approx 0.33$, i.e. by a factor 3.

Also in this context, long-range correlations are essentially absent being smaller than 1%, despite long-range epistasis is present. Hence, a functional constraint on the cooperation between subparts potentially far away in the structure, as allosteric and active site, implies strong long-range epistasis, but does not imply strong long-range correlations, which then implies small couplings. Numerical values for correlation, epistasis and inferred couplings are reported in Tab. 3.1. The prediction of epistasis via DCA in the toy model has the same features than in the MSA of the in-silico model, see Fig. 3.8 and Fig. 3.9. Since the toy model is a minimal model for a system where global and local features work together to perform a given task, the results suggest that DCA is not able to capture the properties of a system that requires several, variable parts to work in concert.

3.3.1 Empirical data on a PDZ domain support the failure of DCA in capturing epistasis

Recently epistasis was measured in an empirical setting by Salinas and Ranganathan [134] with the aid of deep mutational scan techniques applied to the PDZ domain α2-helix composed of only nine residues, which is part of an allosteric regulatory mechanism controlling ligand binding. Five homologs of PDZ domain were considered in the study. Epistasis is defined as

$$\Delta\Delta\mathscr{G}_{ij}^{xy} = \left(\Delta\mathscr{G}_i^x + \Delta\mathscr{G}_j^y\right) - \Delta\mathscr{G}_{ij}^{xy} \tag{3.12}$$

where $\mathscr{G}$ is the binding free energy and x, y correspond to mutations happening at positions i, j, respectively. DCA inference in [134] is performed on an alignment of 1656 eukaryotic PDZ

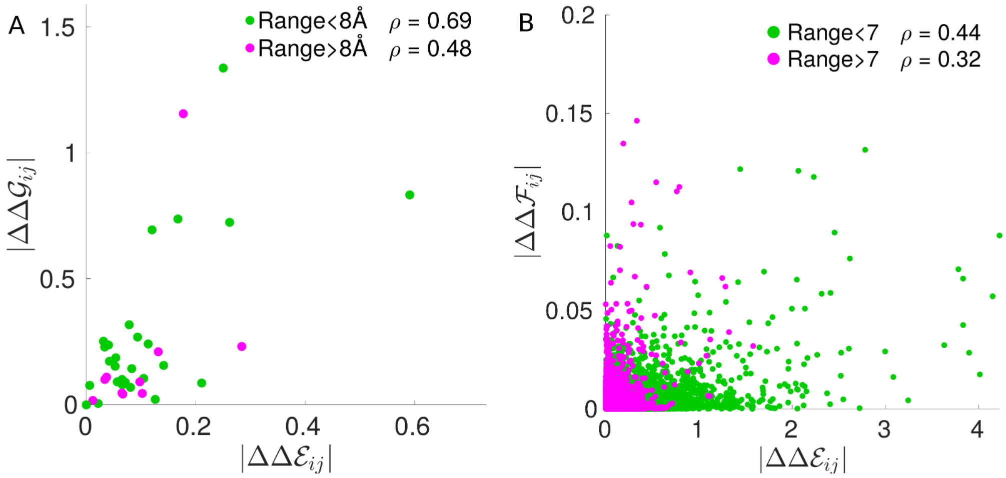

Figure 3.11 – (A) Scatter plot of average epistasis magnitude $|\Delta\Delta\mathscr{G}|$ plotted versus DCA-inferred energetic couplings $|\Delta\Delta\mathscr{E}|$, where the color code distinguishes short and long distance pairs of residues on the PDZ α2-helix three-dimensional structure. The data are from [134]. ρ, the Pearson correlation coefficient, indicates a better performance at short range. As a comparison, in (B) the scatter plot of average epistasis magnitude $|\Delta\Delta\mathscr{F}|$ as function DCA-inferred energetic couplings $|\Delta\Delta\mathscr{E}|$ is shown in the in-silico evolved networks. Similarly to (A), the prediction at long distance is poorer than at short distance.

domains, from where the DCA epistasis prediction $|\Delta\Delta\mathscr{E}_{ij}^{xy}|$ can be directly estimated. Averages over mutations x, y and over the five homologs are considered to improve the prediction — denoted here for simplicity as $\Delta\Delta\mathscr{E}_{ij}$ and $\Delta\Delta\mathscr{G}_{ij}$. Fig. 3.11A shows how well $|\Delta\Delta\mathscr{E}_{ij}|$ predicts the experimental energetic couplings $|\Delta\Delta\mathscr{G}_{ij}|$ for pairs of residues (i, j) at distance > 8Å and < 8Å, where distances are measured on the known three-dimensional crystal structure of the PDZ α2-helix and averaged over the five homologs. A stronger correlation between $|\Delta\Delta\mathscr{G}|$ and $|\Delta\Delta\mathscr{E}|$ for short range pairs (Pearson correlation $\rho = 0.69$), than for long range pairs ($\rho = 0.48$), as the long-range strong epistatic interaction between residues 1 and 8 is not captured by the DCA-inferred energetic couplings, as reported in the original article [134]. $|\Delta\Delta\mathscr{G}_{18}|$ in Fig. 3.11A is the point at largest $|\Delta\Delta\mathscr{G}|$ in the set of long-range pairs. This observation is consistent with the prediction on the limits of DCA in capturing strong long-range epistasis in the sequences evolved to display cooperative allosteric behaviour, as shown in Fig. 3.8 and Fig. 3.11B. Surely, this results is not enough to corroborate such hypothesis, but the advancement in deep mutational scan techniques may provide a less expensive way to measure epistatic interactions in several proteins, increasing the empirical data available, as discussed in the Introduction.

3.4 Performance of Statistical Coupling Analysis

Statistical Coupling Analysis (SCA) uses the co-evolutionary information in the MSA by extracting the principal components of its covariance matrix, thus defining groups of residues that have co-evolved together, as overviewed in the Introduction, see e.g. Fig. 11. SCA was originally based on a covariance matrix weighted by conservation [99] as in Eq. 15. Other variants use a non-weighted covariance matrix [134] and the inverse off-diagonal covariance [101]. In the following, these different definitions are used to test the performance on inferring functional information.

Subsec. 3.4.1 asks whether SCA is able to predict which region of the network is relevant for function in the case of sequences evolved in the in-silico model to optimise the geometric function. Subsec. 3.4.2, instead, focuses on the prediction of the observed long-range epistasis which has been seen to fail when DCA is used as inference model.

3.4.1 Geometric sequences: local structure

Key aspects of the design are more likely to stay conserved in evolution. The conservation map of allosteric networks evolved to perform the geometric task is shown in Fig. 3.12A. The presence of the trumpet is crucial for the amplification of the response that characterises function in geometric networks, as discussed in Chapter 1 (Sec. 1.3).

A trumpet pattern is found to emerge in the same location where coordination decrease monotonically from the allosteric to the active site, see Fig. 1.6. However, conservation is considerably higher around the active site, supporting the claim that the specificity of the response is essentially controlled by the geometry of the network near the active site, see Subsec. 1.3.4 where the physical principles underlying the geometric fitness are discussed.

The non-weighted covariance matrix C is used as a statistical observable to extract co-evolutionary information. The matrix C between the links α, β reads

$$C_{\alpha\beta}(\sigma_\alpha, \sigma_\beta) = \langle \sigma_\alpha \sigma_\beta \rangle - \langle \sigma_\alpha \rangle \langle \sigma_\beta \rangle, \tag{3.13}$$

where the $\langle \bullet \rangle$ denotes the ensemble average over different solutions. The eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_{N_s}$ of the matrix C can be then computed. Fig. 3.12B compares the spectrum of eigenvalues of a high temperature, essentially random, network with that of allosteric networks obtained at small evolution temperature, where the fitness of solution is high. In the latter case, few eigenvalues much larger than the continuum spectrum are found, which is itself more spread than in the random case.

As discussed in the Introduction, the sector of links that co-evolve is extracted by applying the following procedure [100]: (1) pick up the $N_\Gamma = 10$ eigenvectors $|\psi^\gamma\rangle$ with highest eigenvalues separated from the random spectrum; (2) include a given link α in the sector if, for at least one

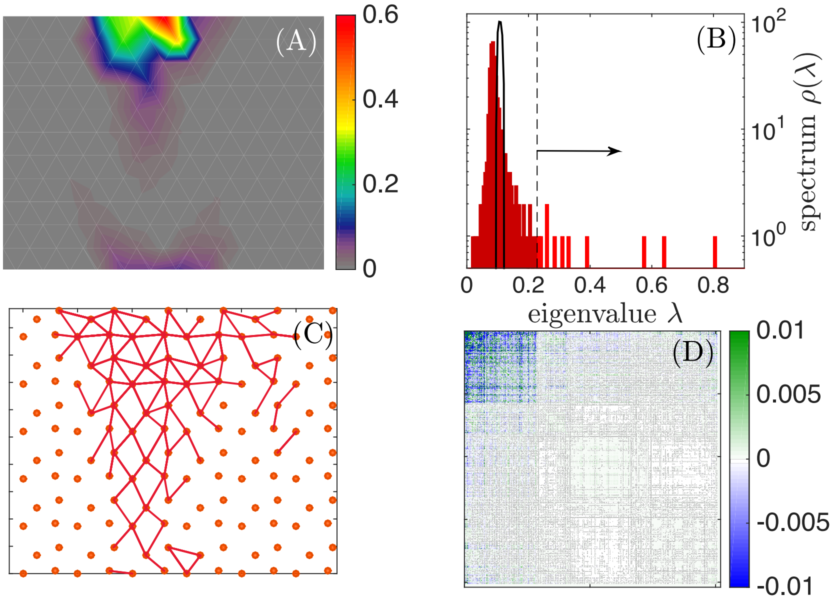

Figure 3.12 – (A) Spatial distribution of conservation, as defined in the text, for $T_e = 0.05$ and $z = 5.0$. (B) Spectrum of eigenvalues $\rho(\lambda)$ of C for the high temperature case ($T_e = 0.30$) in black and low temperature ($T_e = 0.05$) in red. The arrow indicates which eigenvalues are used to identify the springs shown in (C). (C) Springs selected using the procedure explained in the text. (D) $\tilde{C}$ is built using the same parameters as in (C). $\tilde{C}$ presents a clear separation in a region where the correlations are stronger, which corresponds to the trumpet shown in (C). All these figures are made using $L = 12$ and $z = 5$.

of this ten modes, $|\psi^\gamma_\alpha| > \epsilon = 0.05$. The links that are selected are highlighted in Fig. 3.12C and they belong to the region of the trumpet, supporting the idea that co-evolution has the power to uncover key functional aspects [100, 102]. The correlation matrix reconstructed from the 10 top eigenvectors of the covariance matrix can be also built as

$$\tilde{C}_{\alpha\beta} = \sum_{\gamma=1}^{N_\Gamma} \lambda_\gamma |\psi^\gamma\rangle\langle\psi^\gamma|. \tag{3.14}$$

$\tilde{C}$ is shown in Fig. 3.12D after re-ordering links in terms of the strength of their components in the top ten modes. One sector of links where correlations are strong in amplitude, but vary in sign, clearly appears and its links identify the springs involved in the trumpet structure.

3.4.2 Cooperative sequences: long-range epistasis

The recent results of Salinas and collaborators [134] show that SCA outperforms DCA in predicting long-range epistasis, thus supporting to test its performance on the cooperative sequences. However, as shown in Fig. 3.9, the cooperative sequences do not display strong

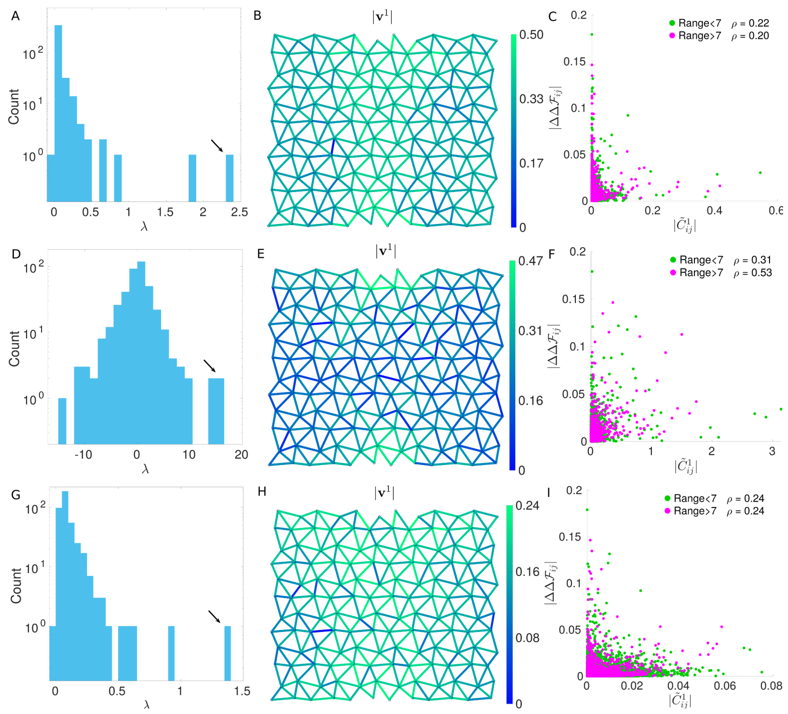

Figure 3.13 – Measure of epistasis by SCA with conservation weight (top row A, B, C), by Inverse Covariance Off-Diagonal (ICOD) [101] (central row D, E, F) and by SCA without conservation weight (bottom row G, H, I). (A, D, G) Spectrum of eigenvalues $\boldsymbol{\lambda}$ of the conservation-weighted covariance (A), of the inverse off-diagonal covariance (D) and of the covariance itself (G), where the highest value λ^1 — corresponding to the first principal component $\boldsymbol{\psi}^1$ — is highlighted by an arrow. (B, E, H) Absolute values of the first principal component $\boldsymbol{v}^1$ visualized on the network. The first principal component of the conservation-weighted covariance in (B), of the inverse off-diagonal covariance in (E) and of the covariance itself in (H). (C, F, I) Scatter plot of $\tilde{C}^1_{ij}$ shown against epistasis with a different color code for long and short distance pairs, where ρ is the Pearson correlation coefficient. $\tilde{C}^1_{ij}$ is constructed from the first top eigenvalue λ^1 and its corresponding principal component $\boldsymbol{\psi}^1$ of the conservation-weighted covariance in (C), of the inverse off-diagonal covariance in (F) and of the covariance itself in (I).

correlations at long range, hence it is very unlikely that a method precisely based on the covariance matrix can significantly improve the performance of DCA. The results with three different definitions of covariance matrices are discussed in the following.

Fig. 3.13A shows the spectrum of eigenvalues $\boldsymbol{\lambda}$ of the conservation-weigthed covariance matrix $\tilde{C}_{ij}$ as in Eq. 15. The top eigenvalue λ^1 is clearly separated from the bulk, thus it is likely to contain some information on the system. As in [134], the covariance from the top eigenmode only, $\tilde{C}^1_{ij} = \lambda^1 \psi^1_i \psi^1_j$, is reconstructed, where $\boldsymbol{\psi}^1$ is the eigenvector corresponding to λ^1 and its structure is visualised on the network in Fig. 3.13B. In Fig. 3.13C the absolute value of $\tilde{C}^1_{ij}$ against epistasis magnitude is shown. The prediction of epistasis compared to the inferred $\Delta\Delta\mathcal{E}_{ij}$ or J_{ij} — see Fig. 3.11 — is not improved, neither at short range nor at long range.

By including the conservation weight ϕ_i the result is slightly improved with respect to the principal components of the uncorrected covariance, see Fig. 3.13G-H-I. On the other hand, conservation only gives a particularly poor estimation regardless of the range.

The recent proposal by Wang et al. [101] of identifying groups of co-evolving amino acids by looking at the largest eigenvector of the inverse covariance matrix is also tested[4]. Given the nature of the chosen matrix, this method is called Inverse Covariance Off-Diagonal (ICOD). The top eigenvalue of the ICOD, see Fig. 3.13D, corresponds to a non-local mode mainly generated by links close to the active and allosteric site as shown in Fig. 3.13E and correlates to long-range epistasis to a larger extent than previous methods, Fig. 3.13F.

Summary. The application of a variant of SCA — without conservational weight in the definition of the covariance matrix — on sequences evolved to optimise the transmission of strain shows to recover the position of links in the region responsible for function, see Fig. 3.12. Moreover, SCA — with and without conservation weight — and ICOD are tested for the prediction of epistasis as shown in Fig. 3.13. SCA performs poorly in the prediction of long-range epistasis, while ICOD is found to provide a larger correlation with true epistasis. Hence, the performance of ICOD needs to be investigated in more detail.

3.5 Conclusions

3.5.1 Summary of results

This chapter main focus is on the performance of Direct Coupling Analysis (DCA) in predicting features of the fitness landscape, which relates the amino acid sequence of a protein to its fitness, as discussed in the Introduction, see Fig. 13. The in-silico sequences evolved to display cooperative binding in Chapter 1 (Sec. 1.4) are used to benchmark inference methods in this task. The design chosen for the analysis is the one of shear motion, mechanically characterised in Fig. 1.14. The results, however, do not depend on the specific design.

DCA shows good performance in the prediction of single mutations costs, which reflect functional features of the networks, as the weakly-connected region connecting allosteric and active site related to the shear motion, see Fig. 3.2, Fig. 3.5 and Fig. 3.6.

However, DCA performs poorly in both generating synthetic sequences with good fitness and in capturing the observed strong long-range epistatic interactions. This failure is supported by recent empirical data of deep mutational scan of a functional region of the PDZ domain [134], but it is still puzzling given the success of DCA in comparing with natural-like sequences [148].

A toy model introduced in Sec. 3.3 proposes an explanation of the failure of DCA in predicting

---

[4] Note that the top eigenvector of ICOD corresponds to eigenvectors with low eigenvalue of the usually used covariance matrix. The choice of low eigenvalues of the covariance matrix is considered crucial to avoid a contribution of phylogenetic effects as discussed in [220]. Moreover, low energy modes usually discarded by PCA are also found to be needed for a good prediction of contacts, further motivating their relevance [221].

such properties by considering a simple model with a hierarchical structure that need several parts to work in concert to function, see Fig. 3.10.

The performance of SCA in the prediction of long-range epistatic interactions is also revealed to be poor, while a more recent method, based on the inverse of the covariance matrix (ICOD) shows a larger correlations with the true epistasis, suggesting to explore this method further.

This chapter also studies the performance of Statistical Coupling Analysis (SCA) in predicting which are the links of the evolved network relevant for the allosteric functioning. The benchmark is done with the sequences optimised to transmit a strain from the allosteric to the active site, introduced in Sec. 1.3. The links identified with the co-evolutionary information contained in the SCA matrix reflect the trumpet structure which is crucial to achieve allosteric function in geometric architectures, see Sec. 1.3.4. The same results are obtained for cooperative architectures, but are not reported in this document.

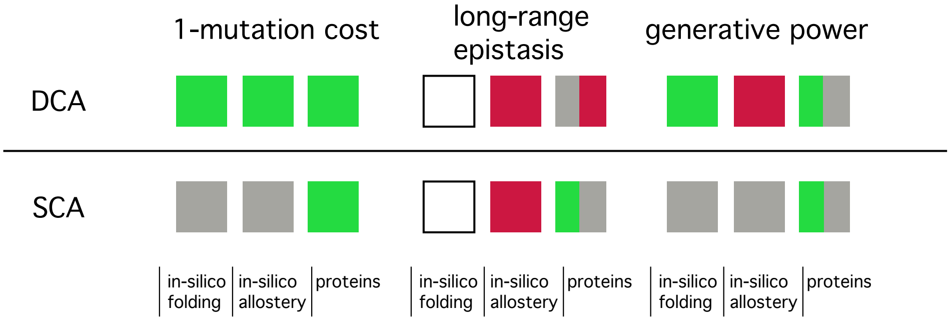

Figure 3.14 – This figure overviews schematically the performance of DCA and SCA in predicting the effect of single mutations on fitness, capturing the long-range epistatic interactions and achieving generative power. Green encode for success, red for failure and gray for unknown/not applicable. If a square is middle green, it means that the evidence supporting success is not strong enough. The inference is performed on MSA from three systems: in-silico models mimicking folding constraints [138, 139, 141] (left), the in-silico model for the evolution of allostery developed in this dissertation (middle) and proteins [105, 134, 144, 148] (right).

Fig. 3.14 shows an overview of the performance of DCA and SCA for the prediction of single mutations, long-range epistasis and for generating functional sequences in in-silico models of folding [138], in the in-silico model of allostery here discussed and in proteins [99, 105, 134, 144] with the goal of clarifying future directions.

3.5.2 Further directions

The following open questions can be delineated.

- **Incorporating folding constraints in the in-silico model**. The in-silico elastic model could be extended by considering the constraint that the protein must fold to operate, in addition to the considered allosteric constraint. The works on lattice proteins [138–140] suggest to consider nodes as amino acids. Hence, the stiffness of the spring between two

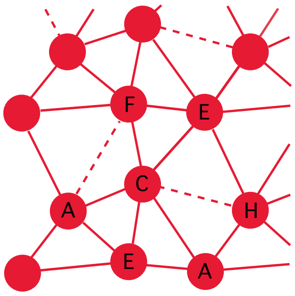

Figure 3.15 – The in-silico model introduced in Sec. 1.2 can be extended to include folding constraints. Indeed, the link occupancy can be dictated by the nature of the amino acids of the corresponding nodes pair, here denoted with letters in the alphabet. This could be easily introduced via a matrix Q_{ab} with binary values defining the interactions between amino acids, i.e. the occupancy of the corresponding link. In the sketch network it would enforce $Q_{AF} = 0$, $Q_{AC} = 1$, et cetera.

adjacent amino acids as well as their contribution to the total folding energy depend on the relationship between that pair, by e.g. defining a matrix encoding which pairs of amino acid can be connected via a spring, see Fig. 3.15 for a sketch of the resulting elastic model[5]. Although such a model is likely to lead to similar results as presented in this chapter for long-range evolutionary couplings, it will presumably differ significantly in the statistics of short-range ones. In particular, it may capture why three-body correlations are well described by two-body correlations in real proteins [68] which is not the case with the current in-silico sequences (Fig. 3.4C), and may lead to stronger conservation overall [222].

- **Neural networks as possible generative models**. Can one find better generative models than DCA for inferring long-range functional couplings? Several ways have been proposed to go beyond pairwise models by including nonlinearities, which implicitly take into account correlations at all orders, as nonlinear potentials in Restricted Boltzmann Machines [223], maximum-entropy probability measures with a nonlinear function of the energy [224], maximum-likelihood inference procedures based on nonlinear functions [225] and, finally, also deeper architectures [222,226].

  As a first test, a 3-layers feed-forward neural network with sigmoid activation functions can be trained to learn the values of fitness in the toy model of Fig. 3.10. Mutation costs and epistasis, even at long range, are correctly captured by this method. This observation suggests that neural networks may lead to better generative models in proteins, an hypothesis that could also be benchmarked in silico. Already, deep learning showed to increase the performance of protein structure prediction with the work of [95,96,222]. Further discussion on alternative inference models is layed down in the next chapter highlighting the conclusions of the thesis.

- **Classification of allosteric designs from co-evolution**. The co-evolved links identified

---

[5] Simulations of such models for the cooperative task lead to similar designs, with a lower average coordination given that with the new evolution algorithm the number of springs cannot be fixed as previously. This can be solved by adding a cost in the fitness functional that depends on the coordination, like a chemical potential.

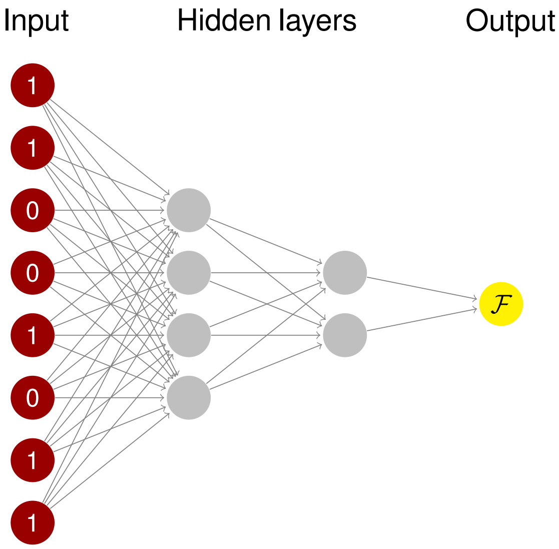

Figure 3.16 – The size of the input layer is eight, as is the size of the system. We add two hidden layers of four and two units and the final one-unit output is 1 if the input sequence has fitness $\mathscr{F}$ and 0 otherwise. The activation function from one layer to the successive one is a sigmoid and the network is fully connected, i.e. all units in one layer are connected to all units of the successive one.

by SCA reflect functional positions: the trumpet architecture can be identified by looking at the spatial organisation of the sector, Fig. 3.12. The same holds for the hinge and shear design, where however the distinction between the two architectures becomes unclear. Hence, the question of how to perform this classification without having to use as discriminant the spatial organisation of the links in the sector is still open.

Supervised learning via simple fully-connected networks is able to distinguish functional from random sequences, is it able also to distinguish allosteric designs? Anyhow, ultimately, the goal is to have a unsupervised scheme that can be applied also for proteins. Hence, other methods need to be explored, like e.g. clustering methods — that can be used for both unsupervised and supervised dimensionality reduction [227] — or generative deep learning techniques [228, 229].

Conclusions

In this dissertation allostery is approached from two different perspectives concerning mechanics and co-evolution.

Mechanics: classification of allosteric architectures

The physical mechanisms underlying allosteric regulation in proteins are not yet fully understood. To provide insights on the physical requirements for allostery, we introduced an in-silico model of evolution of allosteric behaviour in elastic materials. In particular, we consider discrete elastic networks where the nodes of the two- or three-dimensional lattice are connected by harmonic springs, supported by the success of elasticity in capturing the conformational dynamics in some proteins [56].

- **Classification**. The evolution is performed on two allosteric functions. One optimises the transmission of strain between the allosteric and active site, while the other seeks for optimal cooperative binding energy between the two sites. The evolution of the two tasks is found to implement drastically different architectures to achieve function. The first exploits a lever mechanism, rooted in the physical properties of the structure, to amplify the allosteric response from the allosteric to the active site. This architecture is novel and evidence supporting its use also in proteins is discussed in the case of a PDZ domain. Indeed, preliminary data suggest that a PDZ domain, whose allosteric response does not fit into any classification, uses an amplification mechanism similar to the lever found in the evolution of the in-silico model.

  The second, optimising for cooperative binding, recovers usual architectures found in proteins, like hinge and shear motions, and twist designs, while many others that are not reconducible to easily interpretable movements. This heterogeneity in structural organisations is unified by the same underlying physical principles. Indeed, all architectures exhibit a mechanism, an extended low-energy mode that describes almost perfectly the corresponding conformational change. The stiffness of this mode is not random: indeed, to properly function, the mode gest softer as the size of the system increases. This result provides a prediction that is tested in the case of allosteric proteins and proposes a physical explanation on the presence of such low-energy modes in proteins.

- **Effect of elastic non linearities**. The elastic network with harmonic springs discussed above belongs to the framework of the induced fit scenario of allostery, where only one state is present in the energy landscape. Conversely, the population shift scenario poses a two-state landscape where the system can switch from one state to the other upon a change in physiological conditions. A mechanical version of the population shift scenario for cooperative binding is discussed by introducing anharmonic springs to study the effects of elastic nonlinearities. The population shift scenario is found to be more robust to mutations that increase the stiffness of the soft mode than the induced fit scenario. However, to properly function the system is poised at the same condition: the larger the system, the softer the mode needs to be.
- **Database of allosteric proteins to test the predictions.** A database of 34 allosteric proteins is built from the literature to study systematically the prediction that a larger protein displays a softer mode. First, the results confirm that in most of the proteins the conformational change is indeed described by a low energy mode. Then, the allosteric response is proven to be special by the considerable amount of shear deformation that it carries, being larger than any of the modes. Finally, the stiffness of the allosteric response is shown as function of the number of residues and supports the predicted scaling, yet not providing a definite conclusion.

Towards a conclusive empirical validation

The in-silico model allows to capture the physical requirements for the optimisation of a given allosteric task. In the case of cooperative binding, a larger protein is expected to display a softer mode to properly function. The validation of this prediction with data from 34 allosteric proteins is encouraging, but not conclusive, given the large scatter between different proteins.

In the near future, the plan is to investigate in more detail the allosteric nature of all the proteins in the dataset. Are all proteins displaying allostery through cooperative binding? Which are the architectures used by the proteins where ENM fails to identify a functional soft mode? These questions also connect with the novel lever design discovered when the transmission of strain is optimised in the in-silico model. Do proteins in the dataset display such amplification?

Furthermore, an experiment focusing on one protein would be able to probe the validity of the predicted framework. On the one hand, molecular dynamics experiments allow to accurately compute the vibrational modes and the stiffness of the allosteric response, while also monitoring how the energy profile along the allosteric response evolves with binding. This could allow to test the dependence of fitness on the stiffness of the low energy mode and the idea that this mode enables the transition from the inactive to the active state. On the other hand, mutation scans where cooperative fitness is measured via binding assays, combined with single molecule experiments or ultrafast laser pulses able to estimate the stiffness of the allosteric response, can be used to test experimentally the relationship between cooperative energy and stiffness. In particular, it would be possible to investigate whether the cooperative binding energy is more robust to mutations that increase the stiffness of the soft mode, as suggested by the mechanics of the population shift scenario, in contrast with the induced fit framework.

Inference: benchmark on allosteric dependencies

Among the several methods to infer fitness, structure and function from the co-evolutionary information encoded in alignments of sequence of a protein family the focus in this dissertation is on Direct Coupling Analysis (DCA) [81] and Statistical Coupling Analysis (SCA) [99]. We use sequences evolved with the in-silico model for the display of allosteric communication to benchmark the performance of inference methods in dealing with a minimal sequence alignment where structurally distant and functionally related sites are crucial to be captured to predict fitness.

- **Classification from co-evolution (SCA)**. The exploration of functional networks revealed two classes of physical principles underpinning allosteric functioning. Is it possible to perform such classification solely from sequence data? We show that SCA is able to identify which links belong to functional regions, hence delineating the different links distribution that are observed for the lever design and for other cooperative designs, as hinge or shear motions. In the near future, other approaches as clustering methods and neural networks will be benchmarked to investigate whether an unsupervised method is able to directly distinguish the design, without having to look at the distribution of co-evolved links as happens for SCA.
- **Prediction of mutation costs (DCA)**. The knowledge of the mechanical fitness function allows to precisely characterise the physical effects of mutations on the fitness. Mutations at single sites in the network are well reproduced by the DCA-inferred energy function, consistently with the fact that positions of single mutations correlate with the local structural constraints that are well reproduced by DCA.
- **Prediction of epistasis (DCA and SCA)**. Epistasis quantifies the non additivity of mutations and is crucial in reshaping the fitness landscape, thus changing the resulting evolutionary paths. As in some allosteric proteins, the in-silico sequences display strong long-range epistatic interactions, in addition to short-range ones. However, a mutational scan study on a PDZ domain [134] argues that DCA does not to capture such long-range epistasis, while SCA does. We show that both DCA and SCA fail to capture the global epistatic interactions that involve pair of distant links in the in-silico sequences. However, ICOD based on the inverse of the covariance off-diagonal matrix needs to be investigated further as a predictive method for epistasis.
- **Importance of higher-order effects (DCA)**. The contribution of higher order terms to the fitness needs to be studied to check whether they actually improve the prediction

of long-range epistasis. The fitness of the toy model (Eq. C.25) can be expanded in powers of the spin variables to estimate effects of order $n \geq 3$ and check which is the contribution of such terms. Are they relevant for the fitness?

- **Generative power (DCA)**. DCA infers the probability distribution of the energy of a sequence in the alignment, given by the Boltzmann distribution. Thus, it allows to sample the inferred energy landscape and generate synthetic sequences. Does the inferred landscape reproduce the original fitness landscape? We find that DCA fails in generating sequences that reproduce the original value of fitness, even granted the generative power of the used algorithm on in-silico sequences that involve local structural constraints [138] and the good reproduction of natural-like sequences for chorismate mutase enzymes [148].

An alternative route for inferring fitness

Methods going beyond the inference of a pairwise model, like DCA — based on frequencies and correlations of the sequence alignment — may be needed to have sensible predictions about functional dependencies, as allosteric interactions, extending the success on predicting structural constraints, as contacts. In this favour, in Chapter 3 (Subsec. 3.5) we have shown that a fully-connected feed-forward neural network is able to learn epistatic interactions with a better accuracy than Direct Coupling Analysis, in a toy model of protein where several and variable parts work in concert to function. This observation suggests the possibility that neural networks constitute a better place where to look for generative models, possibility that can be benchmarked in silico.

In the deep learning community, it is argued that neural networks improve the performance of classification during training by lowering the dimensionality of the manifolds defined by the parameters of the network, as highlighted in a recent article [159]. In other words, the success in classification is deemed to be related to the identification of hidden variables that are relevant for learning the task, among the plethora of other uninformative directions in parameters space. If this is true, learning reduces to the application of a non-linear function, the activation function, to a linear combination of special variables. However, these hidden variables are not accessible a priori.

The richness of the simple in-silico model for evolution of allostery in elastic materials discussed in this thesis comes at hand another time. Indeed, the insight on the mechanical principles responsible for allostery in sequences evolved in silico can guide the definition of such variables for allosteric communication. Cooperative architectures are found to work via a mechanism, an extended soft mode of stiffness $k_a^\star$ which has a precise scaling with respect to the size of the system for proper function to occur. If $k_a^\star$ is too small or too large there is no cooperativity. Hence, the stiffness of the mechanism can be identified as a hidden variable for allosteric behaviour. Other variables may be related to coupling constants describing how the mode couples to the two sites.

Let us admit that the mechanical fitness of Eq. 3.1 can be approximately rewritten as a non-linear function f of some hidden variables η_i

$$\mathscr{F} = f\left(\{\eta_i\}\right) .$$

Then, inference methods can be used to check whether they actually perform better in the prediction of the hidden variables η_i, instead of the more complex function $\mathscr{F}$. How do these variables vary from sequence to sequence? Is it possible to identify combinations that are more conserved? How well the hidden variables are described by linear regression? Can a fully-connected neural network learn the fitness $\mathscr{F}$ of a sequence by knowing only the hidden variables? Hence, would it result in a model that correctly predicts epistasis and is generative?

This idea connects with recent works that consider inference models with an energy defined as a non-linear function of an independent or pairwise energy [224, 225]

$$\begin{aligned}
&\mathscr{F}(\{\sigma\}) = f\left(\sum_i h_i\sigma_i\right) \quad \text{independent [224, 225]}\\
&\mathscr{F}(\{\sigma\}) = f\left(\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i\right) \quad \text{pairwise [224]} .
\end{aligned}$$

The presence of the non linearity automatically allows to take into account correlations at all orders, as nonlinear potentials in Restricted Boltzmann Machines [223, 229] that are found to capture global features argued to be relevant for allostery. Moreover, the additional non linearity is found to provide good results on the prediction of epsitasis [225] where, however allostery is not systematically considered.

Interestingly, the independent model $\mathscr{F}(\{\sigma\}) = f(\sum_i h_i\sigma_i)$, with a non linearity applied on an additive trait, corresponds to a special kind of epistasis called global. Global epistasis has been recently argued to appear when the response to transient perturbations of proteins is described by few soft modes [230], suggesting that is the case also in the framework of allostery, at least for some proteins.

In practice, non-linear inference models of [223–225] can be tested against in-silico sequences that exhibit allosteric behaviour to study the generative power and the prediction of long-range epistatic interactions, both failed to be predicted with pairwise models. The complete analysis discussed in this dissertation on the mechanics of those sequences allows to go further and see whether the cooperative fitness can be written as a non-linear function of few variables relevant for function. Once the hidden variables are identified, fully-connected neural networks can be used to study whether the fitness, in particular the non-linear function f, can be learnt solely by the knwoledge of the relevant variables η_i.

List of Figures

1 Protein structure 2

2 Allosteric transition in haemoglobin. 4

3 Sketch of cooperativity 5

4 Pacman's diagram of allostery. 7

5 Classification 9

6 Elastic network model. 11

7 Phylogenetic tree of globin 14

8 Co-evolution 15

9 Direct vs indirect couplings 16

10 Contact prediction via mfDCA 19

11 Sector prediction via SCA 21

12 Allosteric pathway in a PDZ domain via SCA 22

13 Fitness landscape as genotype-phenotype mapping 24

14 Classification 26

1.1 One normal mode carries significant information about motion 34

1.2 Different level of coarse-graining of minimal models connecting structure and function 35

1.3 Definition of the network 40

1.4 Undesigned and designed (geometric) networks 43

1.5 Thermodynamics of geometric networks 44

1.6 Architecture of geometric networks 45
1.7 Role of isostaticity I 46
1.8 Role of isostaticity II 47
1.9 Thermal fluctuations in geometric networks 48
1.10 Increasing size: the fruitpicker 48
1.11 Evolution of cooperative networks 50
1.12 Undesigned and designed (cooperative) networks 51
1.13 Architecture of cooperative networks (shear) 53
1.14 Architecture of cooperative networks (hinge) 55
1.15 Architecture of cooperative 3d networks (twist) 56
1.16 Sensitivity of geometric and cooperative designs to a change in the location of the allosteric site 57
1.17 Common features of cooperative architectures 60
1.18 Shear in a cylinder 62
1.19 Extension to the induced fit scenario 67
1.20 Extension to the population shift scenario 68
1.21 Cooperative binding in the population shift scenario 69
1.22 Maximal cooperative fitness in the population shift scenario 71
1.23 Mecahanical model of population shift 73

2.1 Amplification of the allosteric response in a PDZ domain 81
2.2 Analysis of 34 allosteric proteins: geometry of the allosteric response 83
2.3 Competition between participation ratio and shear deformation 85
2.4 Analysis of 34 allosteric proteins: the scaling of stiffness 85

3.1 Study of co-evolution in artificial allosteric networks 91
3.2 Single mutation cost and conservation in the shear architecture 92

3.3 Classification and mechanical characterization of epistasis of cooperative sequences . . . 94
3.4 Performance of the inference procedure . . . 96
3.5 Properties of generated allosteric sequences . . . 97
3.6 Prediction of mutation costs by DCA . . . 98
3.7 Generative performance of DCA . . . 99
3.8 Prediction of epistasis by DCA . . . 100
3.9 Absence of long-range correlations . . . 102
3.10 Sketch of a simple model for protein function . . . 103
3.11 Prediction of experimentally measured epistasis by DCA from [134] . . . 105
3.12 SCA on geometric sequences . . . 107
3.13 Measure of epistasis by SCA . . . 108
3.14 Overview of performance of DCA and SCA. . . . 110
3.15 In-silico model with amino acid degrees of freedom . . . 111
3.16 Graphical representation of the feed-forward neural network for regression . . 112

A.1 Distortion of the lattice . . . 125
A.2 Inverse trumpet in weakly-connected networks . . . 130
A.3 Verification of decay of cooperative energy in discrete elastic medium . . . 133

C.1 Mechanics of mutations . . . 139
C.2 Prediction of epistasis by the DCA-inferred model . . . 141
C.3 Comparison with mfDCA . . . 144
C.4 Conservation taking into account the background frequency . . . 146

Acronyms

ACE Adaptive Cluster Expansion.

ATCase Aspartate TransCarbamoylase.

ATP Adenosine TriPhosphate.

DCA Direct Coupling Analysis.

DNA DeoxyriboNucleic Acid.

ENM Elastic Network Model.

ICOD Inverse Covariance Off-Diagonal.

KNF Koshland-Némethy-Filmer.

mfDCA mean-field DCA.

MSA Multiple Sequence Alignment.

MWC Monod-Wyman-Changeux.

NMR Nuclear Magnetic Resonance.

PCA Principal Component Analysis.

PDB Protein Data Bank.

SCA Statistical Coupling Analysis.

A Mechanical networks

A.1 Construction of the embedding lattice

2D triangular lattice. As discussed in Subsec. 1.2.1 in Chapter 1, we introduce a slight distortion of the lattice to remove straight lines that occur in a regular triangular lattice. Such straight lines are singular since they lead to unphysical localised floppy modes orthogonal to them. Ref. [231] suggests that the localised modes can be removed by imposing a random displacement on the nodes. However, avoiding the introduction of frozen disorder we opt for a different strategy. We group the nodes in the lattice by four, labeled as A B C D in Fig. A.1. One group forms a cell of the new distorted lattice. In each cell, node A stays in place, while nodes B, C, and D move by some distance δ: B moves along the direction perpendicular to BC, C along the direction perpendicular to CD, and D along the direction perpendicular to DB, as illustrated. We set δ to 0.2, since the straight lines are efficiently reduced with this distortion.

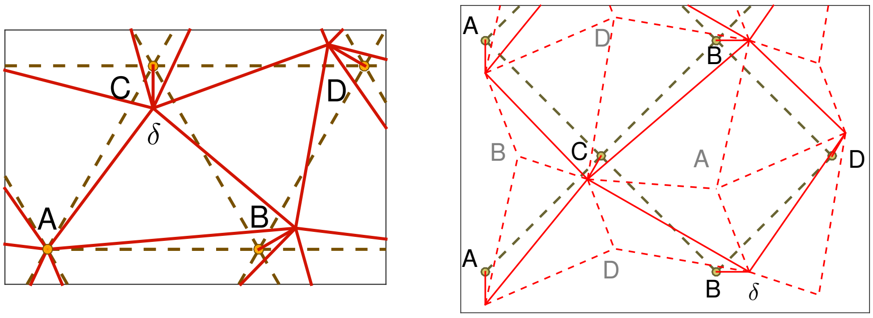

Figure A.1 – Illustration of the distorted triangular lattice (left) and distorted FCC lattice (right).

3D face-centered cubic (FCC) lattice. We introduce a similar distortion to the face-centered cubic lattice. Again, we label the lattice nodes into four different types A B C D, as shown in one layer in the z direction in Fig. A.1. The nodes are labeled in such a way that all 12 nearest neighbors of a node are different from it. For example, the center node in the bottom panel of Fig. A.1, labeled as C, is connecting to two As and two Bs (in solid lines) in the layer and four Ds with two other As and Bs (in dashed lines) out of the layer. To each layer in the z direction, there are two other layers, and in those two layers, D and A, B are located at the same x and y. Thus, we only see half of them (two D, one A and one B) connecting to C in a two dimensional projection along the z direction in Fig. A.1. We move all As along negative y direction, all Bs along positive x direction, all Cs along $(-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}})$, and Ds along $(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{3}})$ by $\delta = 0.2$. As shown in Fig. A.1, all straight lines composed of different links are thus avoided.

A.2 Linear response of elastic networks

A.2.1 Stiffness matrix

The minimum of the energy of a spring configuration can be calculated via gradient descent methods. However, via linear algebra arguments, discussed in this section, it is possible to show that minimisation resorts in simple equations that can be solved via matrix multiplication and inversion [232].

Let us consider a network composed of N nodes in d dimension where neighbouring nodes are connected by N_{psp} links.

Consider a displacement field $\delta\vec{R}_i \equiv \vec{R}_i - \vec{R}_{i0}$, where $\vec{R}_{i0}$ is the position of the node i in the initial mechanical equilibrium. The equilibrium distance $r_{\langle ij\rangle}$ can be defined as $r_{\langle ij\rangle} \equiv ||\vec{R}_{i0} - \vec{R}_{j0}||$ between node i and j. The distance among neighboring nodes at first order in $\delta\vec{R}_i$ changes by

$$||\vec{R}_i - \vec{R}_j|| = r_{\langle ij\rangle} + \sum_l S_{\langle ij\rangle,l}\delta\vec{R}_l + o(\delta\vec{R}^2), \tag{A.1}$$

where $\langle ij\rangle$ indicates the link between node i and j. The matrix $S_{\langle ij\rangle,\bullet} = \hat{n}_{\langle ij\rangle}(\langle i| - \langle j|)$ — where $\hat{n}_{\langle ij\rangle}$ is the unit vector along link $\langle ij\rangle$ from j to i — is the structure matrix, which defines the linear relation between displacements and changes of distances via Eq. A.1. Its size is given by the number of links N_{psp} times the number of nodes multiplied by the space dimension, Nd.

The force on a node can be written as composition of tensions via the structure matrix

$$\vec{F}_i = \sum_j \hat{n}_{\langle ij\rangle} f_{\langle ij\rangle} = \sum_{\langle lm\rangle} S_{\langle lm\rangle,i} f_{\langle lm\rangle}. \tag{A.2}$$

For linear springs between neighbouring nodes i and j, $f_{\langle ij\rangle} = k_{\langle ij\rangle}\delta r_{\langle ij\rangle}$, the force resulting from an applied displacement reads

$$|\mathbf{F}\rangle = \mathcal{M}|\delta\mathbf{R}\rangle, \tag{A.3}$$

where the stiffness matrix (also known as dynamical matrix) $\mathcal{M}_{i,j} = \sum_{\langle lm\rangle} k_{\langle lm\rangle} S_{\langle lm\rangle,i} S_{\langle lm\rangle,j} = kS^t S$ depends only on the occupancy vector $|\sigma\rangle$ and the link directions. It is a symmetric matrix for pairwise interactions such as springs.

The elastic energy can be then written as function of the displacement field $|\delta\mathbf{R}\rangle$ of dimension

Nd is

$$E = \frac{1}{2}\langle \mathbf{F}|\delta\mathbf{R}\rangle = \frac{1}{2}\langle \delta\mathbf{R}|\mathcal{M}|\delta\mathbf{R}\rangle. \tag{A.4}$$

A.2.2 Linear response to an imposed displacement

When a displacement is imposed on the subset ϵ of n nodes, $\delta\mathbf{R}^\epsilon$, forces are applied locally on the nodes mimicking the induced mismatch in the structure. However, all other $N-n$ nodes adapt to a new mechanical equilibrium with no net forces on them and follow a displacement $\delta\mathbf{R}_r$. Thus for this choice of basis Eq.(A.3) becomes

$$\begin{pmatrix} \vec{F} \\ \vec{0} \end{pmatrix} = \mathcal{M} \begin{pmatrix} \delta\mathbf{R}^\epsilon \\ \delta\mathbf{R}_r \end{pmatrix}. \tag{A.5}$$

which leads to:

$$\begin{pmatrix} \vec{F} \\ \delta\mathbf{R}_r \end{pmatrix} = Q^{-1}\mathcal{M} \begin{pmatrix} d\mathbf{R}^\epsilon \\ \vec{0} \end{pmatrix} \tag{A.6}$$

with

$$Q_{ij} = \begin{cases} \delta_{ij} & \text{if } j \in \epsilon \\ -\mathcal{M}_{ij} & \text{if } j \notin \epsilon \end{cases}. \tag{A.7}$$

When nontrivial zero modes are present in the network — like a dangling of one node, the linear equation (A.6) may not be solvable due to the matrix Q that becomes non invertible. Q^{-1} should be then understood as the pseudo-inverse, ensuring that the network does not respond along the directions of the zero modes. Indeed, the corresponding singular values are zero in Q. Another possibility to avoid zero modes is by imposing that each node also interacts with all its next-nearest neighbors via weak springs of stiffness $k_\mathrm{w} \ll 1$ that lift the eigenvalue of the nontrivial zero modes. Both methods lead to qualitatively identical results. For numerical costs, our results were computed using the second approach with $k_\mathrm{w} = 10^{-4}$. So our stiffness matrix reads $\mathcal{M} = S_\sigma^t S_\sigma + k_\mathrm{w} S_\mathrm{w}^t S_\mathrm{w}$.

Dealing with translations and rotations. When binding a ligand, applying the procedure discussed above, the translational and rotational degrees of freedom (TR) of the nodes are not determined and need to be taken care of.

Let us define Ψ^ϵ as $dN_\epsilon \times d_{TR}$-dimensional matrix, composed of a set of vectors with $d_{TR} = 6$ in $d = 3$ and $d_{TR} = 3$ in $d = 2$. If we represente the TR degrees of freedom at the nodes with this matrix Ψ^ϵ, any imposed displacement giving the same shape change results to follow

$$d\mathbf{R}^\epsilon = d\mathbf{R}_0^\epsilon + \Psi^\epsilon \cdot \vec{c}. \tag{A.8}$$

where $d\mathbf{R}_0^\epsilon$ is purely determined by the shape change, $d\mathbf{R}_0^\epsilon \cdot \Psi^\epsilon = \vec{0}$, and $\vec{c}$ is a parameter vector of dimension d_{TR} to count the translational and rotational contribution additional to the shape change. We can thus consider a new basis with a dN_ϵ by dN_ϵ transform matrix U to the original space on imposed nodes so that

$$d\mathbf{R}^\epsilon = U \begin{pmatrix} \vec{\delta}_0 \\ \vec{c} \end{pmatrix}, \tag{A.9}$$

translations and rotations are isolated from the shape change defined by $\vec{\delta}_0$. In this new basis, the forces $\vec{F}$ imposed on ϵ obey total force and torque balance

$$U^t \vec{F} = \begin{pmatrix} \vec{f} \\ \vec{0} \end{pmatrix}. \tag{A.10}$$

The linear response problem thus is rewritten as

$$\tilde{Q} \begin{pmatrix} \vec{f} \\ \vec{c} \\ d\mathbf{R}_r \end{pmatrix} = \begin{pmatrix} U^t & \mathbb{0} \\ \mathbb{0} & \mathbb{I} \end{pmatrix} \mathscr{M} \begin{pmatrix} d\mathbf{R}_0^\epsilon \\ \vec{0} \end{pmatrix} \tag{A.11}$$

where

$$\tilde{Q}_{ij} = \begin{cases} \delta_{ij} & \text{if } j \in \epsilon \setminus TR \\ -\tilde{\mathscr{M}}_{ij} & \text{otherwise} \end{cases} \tag{A.12}$$

with

$$\tilde{\mathscr{M}} = \begin{pmatrix} U^t & \mathbb{0} \\ \mathbb{0} & \mathbb{I} \end{pmatrix} \mathscr{M} \begin{pmatrix} U & \mathbb{0} \\ \mathbb{0} & \mathbb{I} \end{pmatrix}. \tag{A.13}$$

Note that given the separation of the two subspaces the matrix U is of dimension $dN_\epsilon \times dN_\epsilon$ and the identity matrix $\mathbb{I}$ is $d(N - N_\epsilon) \times d(N - N_\epsilon)$, consistent with $\mathscr{M}$ being $dN \times dN$. This formalism allow for efficient minimisation of the elastic energy in the linear approximation even for three-dimensional networks.

A.2.3 Measure of thermally-induced motion

Thermally-induced motion in a structure can be quantified via B-factors [55]

$$B = 8\pi^2 \langle u^2 \rangle \tag{A.14}$$

where u is the magnitude of the displacement around the mean position of nodes, and $\langle \bullet \rangle$ denotes thermal averaging.

In an elastic network, the energy due to displacement field $\vec{u}_i$ is,

$$H = \frac{1}{2}\sum_{i,j} \vec{u}_i \cdot \mathscr{M}_{ij} \cdot \vec{u}_j\,, \tag{A.15}$$

where $\mathscr{M}$ is the stiffness matrix defined in Eq.(A.3). The B-factor is directly related to the stiffness matrix $\mathscr{M}$. By definition, the thermal average of the correlation of the displacement at temperature T reads

$$\langle \vec{u}_k \cdot \vec{u}_l \rangle(T) = \frac{1}{Z(T)} \int \vec{u}_k \cdot \vec{u}_l e^{-\frac{1}{2T}\sum_{i,j} \vec{u}_i \cdot \mathscr{M}_{ij} \cdot \vec{u}_j} \prod_i \mathrm{d}\vec{u}_i\,, \tag{A.16}$$

and results to be a Gaussian integral. It can thus be carried out

$$\langle \vec{u}_i \cdot \vec{u}_j \rangle = T\left(\mathscr{M}^{-1}\right)_{ij}\,, \tag{A.17}$$

and

$$B_i = 8\pi^2 \langle u_i^2 \rangle = 8\pi^2 T (\mathscr{M}^{-1})_{ii}. \tag{A.18}$$

B-factors are computed at temperature $T = 1$.

A.3 Trumpet architecture in networks with $z < 4$

In Chapter 1, we have focused on the allosteric task in well-coordinated networks with $z = 5$. Whereas, thermodynamics results show that the task can be accomplished for very floppy networks, with coordination $z < 4$, which are interesting analogies to allostery of some unfolded floppy proteins [233].

In general, the displacement signal can only propagate a finite distance in rather homogeneous floppy networks [**?**]. This can be seen in the left panel of Fig. A.2B. Below the transition temperature ($T_c = 0.12$ for $z = 3.0$), we find that the network resolves this issue by generating a *trumpet* shape structure similar to the rigid case, connecting the allosteric site and the active site, as appears in the average coordination map shown in Fig. A.2A.

The coordination number in this trumpet structure is larger than in the rest of the material, allowing the signal to propagate. Similarly to the $\delta z > 0$ case, the mean coordination number decreases monotonically from the allosteric to the active site. However, the trumpet has the wider nearly-isostatic patch near the allosteric side. We also find that the amplitude of the displacement decreases monotonically from the allosteric to the active side in the bottom panel of Fig. A.2B.

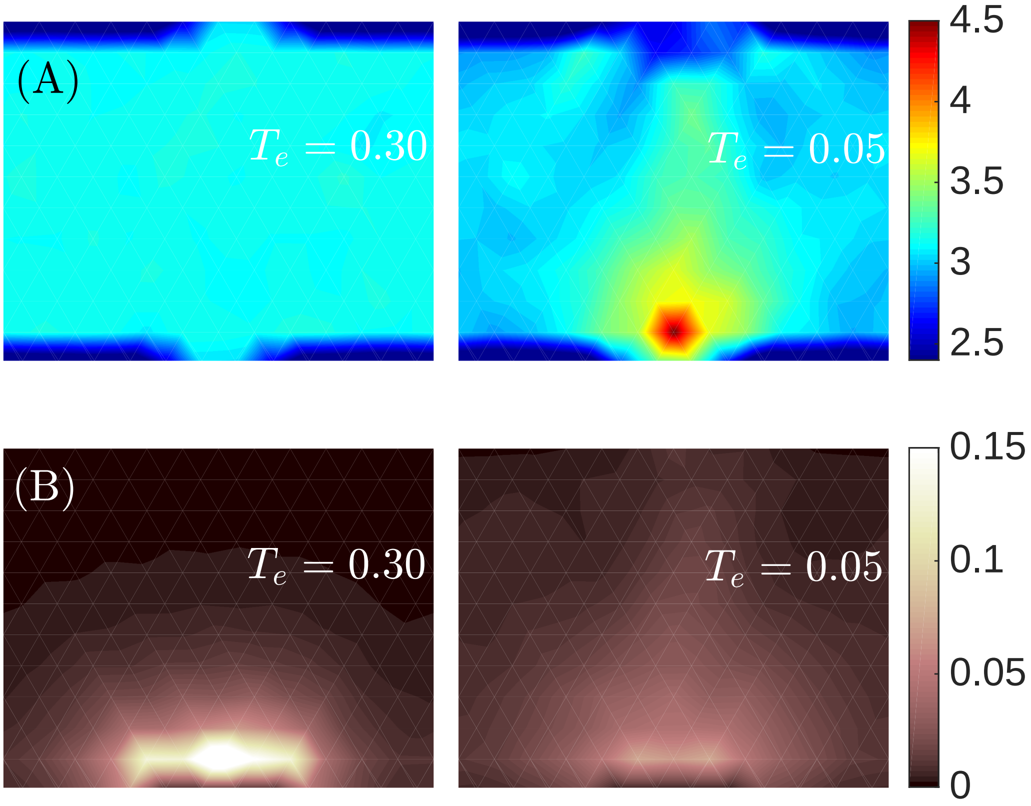

Figure A.2 – (A) Spatial distribution of coordination number. (B) Spatial distribution of the response magnitude to the excitation at allosteric sites. Left: $T_e = 0.3$. Right: $T_e = 0.10$. $z = 3.0$.

A.4 Cooperative energy of two dipoles in a continuous elastic medium

Elastic energy of a force monopole. In the following, we address the formalism to compute the elastic energy when a force is applied on a pathc in an elastic medium. We make the simplifying assumption that the velocity field is divergence free, without loss of generality for the predicted scaling behaviors. Thus, the equation determining the displacement field when a monopole of constant force $\vec{f}$ is applied over a spherical patch of radius c follows [149]

$$\Delta\vec{u} = \nabla\cdot\nabla\vec{u} = -\frac{d}{G\Omega_d c^d}\vec{f}, \tag{A.19}$$

where G is the shear modulus, Ω_d is the solid angle underlying the d-dimensional sphere. By defining $\vec{f} = f\hat{e}_y$, both force and displacement component are enforced to be along y direction. We then solve for the divergence of the displacement field using Gauss theorem

$$\nabla u_y = \begin{cases} -\frac{f}{G\Omega_d c^d} r\hat{e}_r, & r < c \\ -\frac{f}{G\Omega_d}\frac{1}{r^{d-1}}\hat{e}_r. & r \geq c \end{cases} \tag{A.20}$$

The total energy of the monopole reads approximately — for large system sizes

$$E_m = G\int \mathrm{d}^d\vec{r}(\nabla u_y)^2 = \frac{f^2}{G\Omega_d}\left(\int_0^c \frac{r^{d+1}}{c^{2d}} + \int_c^R \frac{1}{r^{d-1}}\right)\mathrm{d}r, \tag{A.21}$$

where R defines the system size. In two dimensions, the integral is dominated by the second term,

$$E_m = \frac{f^2}{2\pi G}\ln\frac{R}{c}. \tag{A.22}$$

In three dimensions and above, the integral of the second term converges in the large size limit $R \to \infty$, and it has the same scaling as the first term

$$E_m = \frac{2df^2}{(d^2-4)\Omega_d G}c^{2-d}. \tag{A.23}$$

Hence, the displacement δ can be achieved by an external force satisfying $\delta = \partial E_m / \partial f$,

$$\delta = f\frac{1}{\pi G}\ln\frac{R}{c}; \qquad d = 2 \tag{A.24}$$

$$\delta = f\frac{4d}{(d^2-4)\Omega_d G}c^{2-d}. \qquad d > 2 \tag{A.25}$$

Elastic energy of a force dipole. In Chapter 1, the cooperative energy between two force dipoles in a undesigned network was argued to vanish as function of the distance between the two dipoles. Here, based on the derivation of the energy for an elastic dipole, we compute the cooperative energy. We consider two dipoles of size c separated by L in a homogeneous medium, as illustrated in Fig. 1.18 in Chapter 1. The distance of the two monopoles forming one dipole is a.

The dipole energy can be defined similarly to the monopole energy computed above as

$$E_d = G\int \mathrm{d}^d\vec{r}\frac{f^2}{\Omega_d^2 G^2}\sum_{i=1}^{d}(x_{+,i} - x_{-,i})^2 = 2E_m + \tilde{E}_d, \tag{A.26}$$

where x_+, x_-, y_+, y_- are the components in x and y directions relative to the + monopole and − monopole, respectively, in the dipole. So the dipole self-energy is

$$\tilde{E}_d = -\frac{2f^2}{\Omega_d^2 G}\int \mathrm{d}^d\vec{r}\sum_{i=1}^{d} x_{+,i}x_{-,i}, \tag{A.27}$$

where the integral is over three regions, within c to the + monopole, within c to the − monopole, and the remaining space.

One can show that the contributions of the first two regions inside monopoles scale as c^{d+2}, while the contribution of the remaining region scales as c^{d+2}, comparable. Outside of the monopoles, $x_{\bullet,i} \approx \frac{1}{r^d} x_i$, so

$$x_+ = \frac{1}{((x-a/2)^2+y^2)^{d/2}}(x-a/2)\,, \tag{A.28}$$

$$x_- = \frac{1}{((x+a/2)^2+y^2)^{d/2}}(x+a/2)\,, \tag{A.29}$$

$$y_+ = \frac{1}{((x-a/2)^2+y^2)^{d/2}}y\,, \tag{A.30}$$

$$y_- = \frac{1}{((x+a/2)^2+y^2)^{d/2}}y\,, \tag{A.31}$$

and

$$\tilde{E}_d \sim -\frac{f^2}{G}\int_c^R r^{d-1}\mathrm{d}r\frac{1}{r^{2d-2}} \sim \begin{cases} -\frac{f^2}{G}\ln\frac{R}{a} & d=2 \\ -\frac{f^2}{G}c^{2-d} & d>2 \end{cases}\,. \tag{A.32}$$

Cooperative energy of two force dipoles. Again, similarly to the computation the dipole self-energy Eq.(A.27), the cooperative energy, defined as the extra energy from the interaction of two dipoles, can be computed as

$$E_{\text{coop}} = 2E_d - E_{\text{tot}} \tag{A.33}$$

$$= -\frac{2f^2}{\Omega_d^2 G}\int \mathrm{d}^d\vec{r}\sum_{i=1}^{d}(x^0_{+,i}-x^0_{-,i})(x^L_{+,i}-x^L_{-,i}). \tag{A.34}$$

where $x^L_\bullet$ are the contributions of the monopole at L. When $c/L \ll 1$, the contribution outside of both the monopoles and the dipoles dominates the energy

$$E_{\text{coop}} \sim \frac{f^2}{G}\int_0^\infty \mathrm{d}\rho\int_c^L \mathrm{d}z\rho^{d-2}\frac{c^2}{[\rho^2+z^2]^{d/2}[\rho^2+(L-z)^2]^{d/2}} \tag{A.35}$$

$$\sim \frac{f^2c^2}{GL^d}\ln\frac{L}{c}. \tag{A.36}$$

For given displacement δ applied at the dipoles we have

$$E_{\text{coop}} \sim \begin{cases} G\frac{c^2\delta^2}{L^2\ln\frac{L}{c}} & d=2 \\ G\frac{c^{2d-2}\delta^2}{L^d}\ln\frac{L}{c} & d>2 \end{cases} \tag{A.37}$$

showing that E_{coop} decays as fast as L^{-d} for two dipoles at a distance L from each other, with weak logarithmic corrections in $d=2$, as discussed in Chapter 1

Numerical verification. This section aims to justify that the results obtained for a continuous elastic medium are valid also for a discrete medium, as the spring network considered in our in-silico model. In Fig. A.3(A) we test our prediction for the cooperativity of a homogenous medium without any design (a distorted crystal with $\delta = 0.2$), and confirm Eq. A.37 for $d = 2$. Instead, in Fig. A.3(B) we test our prediction for an optimal shear design, and confirm the very weak logarithmic decay of the cooperative energy in two dimensions, as described in Eq. 1.22 in Chapter 1.

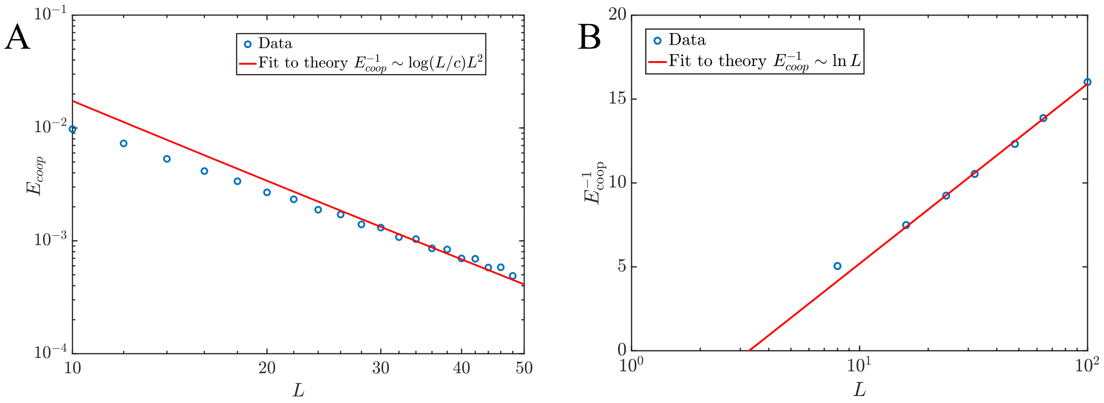

Figure A.3 – (A) Cooperative energy computed for a distorted crystal ($\delta = 0.2$) of varying size L with no mechanism. (B) Inverse of the cooperative energy for a crystal with a soft shear band presenting a mechanism, the softness of the band being chosen as the value of k_w where the cooperative energy is optimal, see Fig. 1.18. The two different scalings predicted from continuous elastic media are fitted and shown as solid lines.

B Strain analysis on empirical data

B.1 Computing the local strain tensor in a protein

In a continuous medium, a motion maps a point $\vec{X}$ in the reference configuration to a new point $\vec{x}$ in the current configuration, the strain tensor of the motion can thus be computed as

$$\epsilon_{ab}(\vec{X}) = \frac{1}{2}\left(\frac{\partial\vec{x}}{\partial X_a}\cdot\frac{\partial\vec{x}}{\partial X_b} - \delta_{ab}\right), \tag{B.1}$$

where a, b labels the spatial dimension.

In a discrete medium like proteins which are a collection of atoms (or residues as we consider), computing the partial derivative $\overleftrightarrow{\Lambda} = \partial\vec{x}/\partial\vec{X}$ at residue i is not straightforward. Ideally, for any neighboring residue j close enough in space

$$\Delta\vec{x}_{ij} = \overleftrightarrow{\Lambda}_i \cdot \Delta\vec{X}_{ij}, \tag{B.2}$$

where $\Delta\vec{X}_{ij} = \vec{R}_{i0} - \vec{R}_{j0}$ and $\Delta\vec{x}_{ij} = \vec{R}_i - \vec{R}_j$ in our setting, where $\vec{R}_i$ is the position of residue i taken from the X-ray structure. We have n_b number of such equations for $\overleftrightarrow{\Lambda}_i$ when n_b neighbors are considered. So $\overleftrightarrow{\Lambda}_i$ are usually over-determined when we consider all neighbors below a certain cutoff distance R_c (we choose R_{c1} = 8.5 Å for first nearest neighbors and R_{c2} = 10.5 Å for second nearest neighbors). Instead of solving Eq. (B.2), we define a mean squared error function [199]

$$MSE(i) = \sum_j (\Delta\vec{x}_{ij} - \overleftrightarrow{\Lambda}_i \cdot \Delta\vec{X}_{ij})^2 w_j(i), \tag{B.3}$$

where we have kept a weight function $w_j(i)$ of node j contribution to i in general. Specifically, we set as in [46] $w_j(i) = 1$ for all nearest neighbors to i ($R_{ij} < R_{c1}$), $w_j(i) = 1 - \frac{R_{ij} - R_{c1}}{R_{c2} - R_{c1}}$ for $R_{c1} < R_{ij} < R_{c2}$ and $w_j = 0$ otherwise. By minimizing the mean squared error with respect to

$\overleftrightarrow{\Lambda}_i$, we have

$$\overleftrightarrow{\Lambda}_i = \sum_j \Delta\vec{x}_{ij}\Delta\vec{X}_{ij}w_j(i)\cdot\left(\sum_j \Delta\vec{X}_{ij}\Delta\vec{X}_{ij}w_j(i)\right)^{-1}, \tag{B.4}$$

and

$$\overleftrightarrow{\epsilon}(i) = \frac{1}{2}\left(\overleftrightarrow{\Lambda}_i^t\cdot\overleftrightarrow{\Lambda}_i - \overleftrightarrow{\delta}\right), \tag{B.5}$$

where $\overleftrightarrow{\delta}$ is the identity tensor.

The shear pseudo-energy [46], a vector field whose components contain a measure of the relative motion of each residue, can be defined from the strain tensor $\overleftrightarrow{\epsilon}(i)$ computed above

$$E_{sh}(i) = \frac{1}{2}\sum_{l,m=1}^{3}[\gamma_{lm}(i)]^2,$$

where the local shear tensor $\overleftrightarrow{\gamma}(i) = \overleftrightarrow{\epsilon}(i) - (1/3)\mathrm{Tr}[\overleftrightarrow{\epsilon}(i)]\overleftrightarrow{\delta}$ depends only on the displacement between the two conformations via the strain tensor $\overleftrightarrow{\epsilon}(i)$ and $\overleftrightarrow{\delta}$.

C Co-evolutionary analysis

C.1 Mechanical interpretation of mutation costs and epistasis

Let us denote by ϵ the set of nodes where ligand binding takes place, e.g. for ligand binding at the allosteric site $\epsilon = (10)$ with size $\dim(\epsilon) = n_0 = 4$. Such event imposes a displacement $\mathbf{R}^{\epsilon}$ on the nodes ϵ which imparts locally a force $\boldsymbol{F}^{\epsilon}$ and induces a response $\mathbf{R}^{\epsilon\to r}$ on all the other nodes r. Clearly $\dim(\epsilon) + \dim(r) = L^d$ where L^d is the total number of nodes for a network of size L in d dimensions; for the example of binding to the allosteric site $r = (01, b)$, where b stands for the bulk of nodes belonging neither to the allosteric nor to the active site. Considering the deformation as a linear response to the external force, as in Eq. A.6, the relation between force and overall response field is written in terms of the dynamical matrix $\mathcal{M}$

$$\begin{pmatrix} \boldsymbol{F}^{\epsilon} \\ \mathbb{0} \end{pmatrix} = \mathcal{M} \begin{pmatrix} \mathbf{R}^{\epsilon} \\ \mathbf{R}^{\epsilon\to r} \end{pmatrix} \tag{C.1}$$

hence $\mathcal{M}$ is endowed with a block structure as follows

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}^{\epsilon,\epsilon} & \mathcal{M}^{\epsilon,r} \\ (\mathcal{M}^{\epsilon,r})^T & \mathcal{M}^{r,r} \end{pmatrix}$$

For pairwise interactions such as springs, the dynamical matrix $\mathcal{M}$ is symmetric. Forces, as well as resulting responses, can be calculated solely from the imposed displacement by introducing a matrix $\boldsymbol{Q}$

$$\boldsymbol{Q} = \begin{pmatrix} \mathbb{1}^{\epsilon} & -\mathcal{M}^{\epsilon,r} \\ \mathbb{0} & -\mathcal{M}^{r,r} \end{pmatrix}$$

such that

$$\begin{pmatrix} \boldsymbol{F}^{\epsilon} \\ \mathbf{R}^{\epsilon\to r} \end{pmatrix} = \boldsymbol{Q}^{-1} \mathcal{M} \begin{pmatrix} \mathbf{R}^{r} \\ \mathbb{0} \end{pmatrix} \tag{C.2}$$

Binding at ϵ costs an elastic energy E^{ϵ}

$$E^{\epsilon} = \frac{1}{2}\boldsymbol{F}^{\epsilon} \cdot \boldsymbol{R}^{\epsilon}, \tag{C.3}$$

thus defining the cooperative fitness as a combination of such elastic energies

$$\mathscr{F} = E^{01} - (E^{11} - E^{10}), \tag{C.4}$$

where E^{01}, E^{11} and E^{10} are given by Eq. C.3 with, respectively, $\epsilon = (01)$, $\epsilon = (11)$ and $\epsilon = (10)$.

Maximal cooperativity corresponds to making the binding of a substrate at the active site energetically favoured when a ligand is already bound to the allosteric site. Indeed, this situation reduces the energy of binding at the active site from E^{01} to $(E^{11} - E^{10})$. The energy of joint binding at the allosteric and active site $E^{11} = \frac{1}{2}\boldsymbol{F}^{11} \cdot \boldsymbol{R}^{11}$ can be espressed as

$$\frac{1}{2}\boldsymbol{F}^{11} \cdot \boldsymbol{R}^{11} = \frac{1}{2}\boldsymbol{F}^{10} \cdot \boldsymbol{R}^{10} + \frac{1}{2}\boldsymbol{F}^{Ac}_{|10} \cdot (\boldsymbol{R}^{01} - \boldsymbol{R}^{10\to Ac}). \tag{C.5}$$

After binding at the allosteric site with an energy cost $\frac{1}{2}\boldsymbol{F}^{10} \cdot \boldsymbol{R}^{10}$, the elastic energy of binding at the active site is determined by (1) the force at the active site when a ligand is already bound at the allosteric site ($\boldsymbol{F}^{Ac}_{|10}$ with subindex $|10$); (2) the displacement imposed at the active site $\boldsymbol{R}^{01}$ to which we subtract the response already caused by ligand binding at the allosteric site $\boldsymbol{R}^{10\to Ac}$. Eq. C.5 allows us to rewrite Eq. C.4 as

$$\mathscr{F} = \frac{1}{2}\boldsymbol{F}^{Ac}_{|10} \cdot \boldsymbol{R}^{10\to Ac} + \frac{1}{2}\delta\boldsymbol{F}^{10\to Ac} \cdot \boldsymbol{R}^{01}, \tag{C.6}$$

where one has the identity $\boldsymbol{F}^{01} - \boldsymbol{F}^{Ac}_{|10} = \delta\boldsymbol{F}^{10\to Ac}$.

Now, we consider the weak elastic coupling limit between the allosteric and active sites. Physically, we assume that the response induced at the active site by binding at the allosteric site is small compared to the one induced by binding at the active site. Mathematically, it corresponds to the assumptions that the elements of the dynamical matrix involving the communication between the two sites $\mathscr{M}^{01,b}(\mathscr{M}^{b,b})^{-1}\mathscr{M}^{b,10}$ are small. In this limit, expressing $\delta\boldsymbol{F}^{10\to Ac}$ and $\boldsymbol{R}^{10\to Ac}$ in terms of the imposed displacements by using Eq. C.2, we find that each term in Eq. C.6 follows, at first order in $\mathscr{M}^{01,b}(\mathscr{M}^{b,b})^{-1}\mathscr{M}^{b,10}$

$$\frac{1}{2}\boldsymbol{F}^{Ac}_{|10} \cdot \boldsymbol{R}^{10\to Ac} \approx \frac{1}{2}\delta\boldsymbol{F}^{10\to Ac} \cdot \boldsymbol{R}^{01} \approx \frac{1}{2}(\boldsymbol{R}^{01})^{T} \cdot (\mathscr{M}^{01,b})(\mathscr{M}^{b,b})^{-1}(\mathscr{M}^{b,10}) \cdot \boldsymbol{R}^{10}, \tag{C.7}$$

where a sum over b, the ensemble of bulk nodes, is taken. Hence, by using that $\frac{1}{2}\delta\boldsymbol{F}^{10\to Ac} \cdot \boldsymbol{R}^{01} \approx \frac{1}{2}\boldsymbol{F}^{Ac}_{|10} \cdot \boldsymbol{R}^{10\to Ac}$, we obtain from Eq. C.6

$$\mathscr{F} \approx \mathbf{F}^{01} \cdot \boldsymbol{R}^{10\to Ac}, \tag{C.8}$$

since $\boldsymbol{F}^{Ac}_{|10}$ can be approximated by $\boldsymbol{F}^{01}$ in the weak coupling limit.

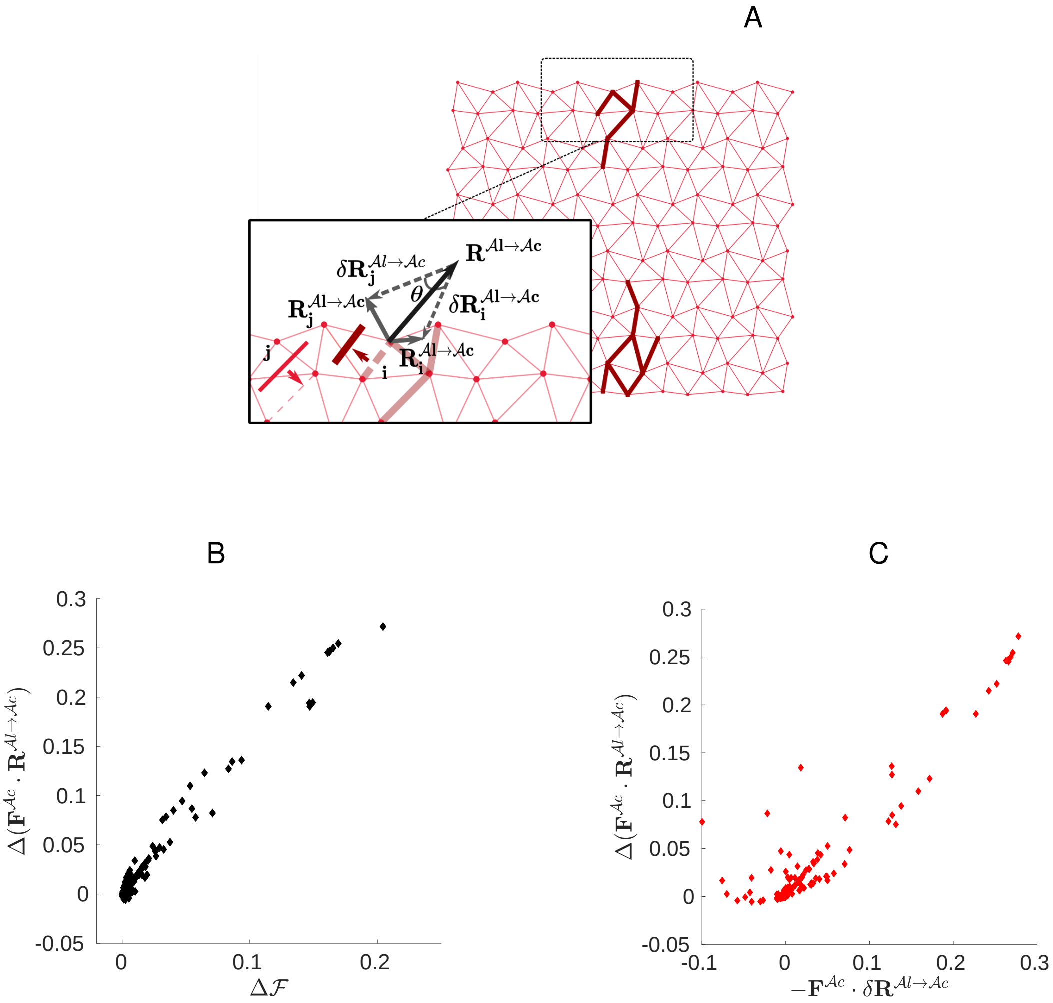

Figure C.1 – (A) The geometry of mutation costs is illustrated in the neighbourhood of the active site where only one of the $n_0 = 4$ nodes is considered. Thick dark red lines highlight links whose disruption would be deleterious for the allosteric fitness. (A,B) Numerical test of the approximation $\Delta\mathscr{F}_i \approx \Delta(\mathbf{F}^{Ac}\cdot\boldsymbol{R}^{10\to Ac})_i$ are shown in (B) and of $\Delta(\mathbf{F}^{Ac}\cdot\boldsymbol{R}^{10\to Ac})_i \approx -\mathbf{F}^{Ac}\cdot\delta\boldsymbol{R}_i^{10\to Ac}$ in (C). The latter is valid only for medium-high mutation costs.

Changes in displacement at the active site dominate the cost of mutation. If we denote by $\mathbf{F}_i^{01}$ and $\boldsymbol{R}_i^{10\to Ac}$ forces and displacements after a mutation at link i, the cost of one mutation can be expressed in this approximation as $\Delta\mathscr{F}_i \approx \Delta(\mathbf{F}^{01}\cdot\boldsymbol{R}^{10\to Ac})_i$, where $\Delta(\mathbf{F}^{01}\cdot\boldsymbol{R}^{10\to Ac})_i = \mathbf{F}^{01}\cdot\boldsymbol{R}^{10\to Ac} - \mathbf{F}_i^{01}\cdot\boldsymbol{R}_i^{10\to Ac}$. See Fig. C.1B for a numerical validation of our approximation. The expression can be further rewritten as

$$\Delta(\mathbf{F}^{01}\cdot\boldsymbol{R}^{10\to Ac})_i \approx -\left(\mathbf{F}^{01}\cdot\delta\boldsymbol{R}_i^{10\to Ac} + \delta\mathbf{F}_i^{01}\cdot\boldsymbol{R}^{10\to Ac} + \delta\mathbf{F}_i^{01}\cdot\delta\boldsymbol{R}_i^{10\to Ac}\right), \tag{C.9}$$

having defined changes in force as $\delta\mathbf{F}_i^{01} = \mathbf{F}_i^{01} - \mathbf{F}^{01}$, in analogy to changes in displacement $\delta\boldsymbol{R}_i^{10\to Ac}$ introduced in Chapter 3. We find numerically that the cost of single mutations, when it is not too small, is dominated by the changes in displacement at the active site

$$\Delta\mathscr{F}_i \approx -\mathbf{F}^{01}\cdot\delta\boldsymbol{R}_i^{10\to Ac} \tag{C.10}$$

as implied jointly by Fig. C.1B-C. As a consequence, epistasis between mutations at i and j with significant magnitude can be written $\Delta\Delta\mathscr{F}_{ij} \approx -\mathbf{F}^{01}\cdot\left(\delta\boldsymbol{R}_{ij}^{10\to Ac} - \delta\boldsymbol{R}_i^{10\to Ac} - \delta\boldsymbol{R}_j^{10\to Ac}\right)$, as presented in the main text.

Displacement vectors and their changes upon high-cost mutations at the active site are schematically depicted in Fig. C.1A. The few links located in the neighbourhood of both the allosteric and active site are crucial to the long-distance propagation of the allosteric response. They exhibit maximal epistasis along with maximal single mutation costs. Hence, they populate the saturation region of Fig. 3.3A shown in Chapter 3.

After a deleterious mutation consisting in removing a spring at link i, the displacement at the active site $\mathbf{R}_i^{10\to Ac}$ is significantly reduced with respect to the original optimal displacement $\mathbf{R}^{10\to Ac}$ and their difference is given by $\delta\mathbf{R}_i^{10\to Ac}$ shown as dashed arrow in C.1A. Given a second lethal mutation at j, the angle between the two mutations can be computed. We denote by θ the angle between $\delta\mathbf{R}_i^{10\to Ac}$ and $\delta\mathbf{R}_j^{10\to Ac}$. For lethal mutations $\cos(\theta) \approx 1$ as shown in Fig. 3.3B in Chapter 3, meaning that all mutations tend to have a homogeneous direction of action which is precisely the one opposite to the displacement at the active site.

C.2 Prediction of epistasis

The scaling of epistasis (Eq. 3.7 in Chapter 3) suggests that a measure simply based on the inferred single mutation cost is valid: $|\Delta\Delta\mathscr{F}_{ij}| \propto \min(\Delta\mathscr{E}_i, \Delta\mathscr{E}_j)$ with $\Delta\mathscr{E}$ inferred by DCA. We have verified that this approximation improves significantly the prediction of long-range epistasis in our in-silico model for the evolution of allostery in comparison to direct evolutionary couplings J_{ij}. This holds true both for single configurations and for the average epistatic pattern, as shown in respectively in Fig. C.2B-C.

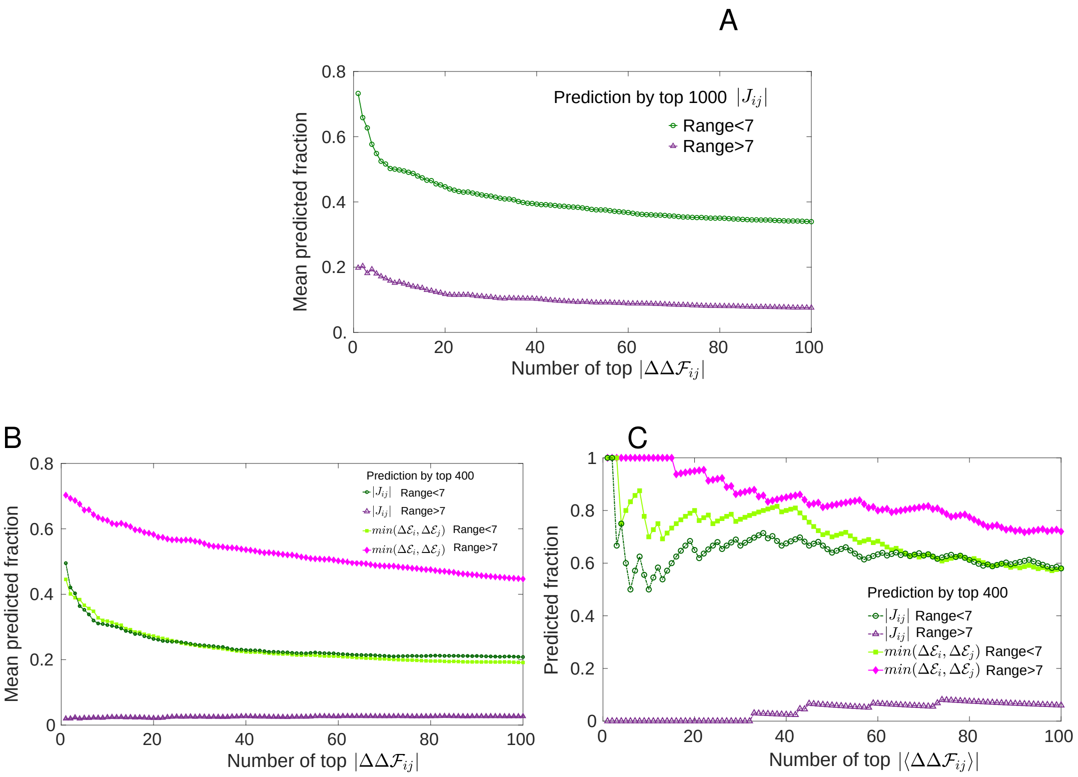

Figure C.2 – (A) Same plot as in Fig. 3.8B (Chapter 3) where we show the fraction of top rank epistasis $|\Delta\Delta\mathscr{F}_{ij}|$ predicted by top 1000 $|J_{ij}|$, averaged over 100 configurations. In comparison to Fig. 3.8B, here we consider a higher number of the couplings with largest magnitude to predict epistasis: the mean predicted fraction increases both for short range and long range epistasis, yet a clear difference between their values remains. (B) Same plot as Fig. 3.8B (Chapter 3) where we added curves for the prediction by $\min(\Delta\mathscr{E}_i, \Delta\mathscr{E}_j)$ — the minimum between average single mutation costs at i and j — corresponding to the scaling 3.7 in Chapter 3, which describes well the trend of epistasis (see Fig. 3.3A). As in Fig. 3.8B (Chapter 3), we rank separately long-range (> 7) and short-range (< 7) pairs of links i and j in terms of $|\Delta\Delta\mathscr{F}_{ij}|$ and we plot the fraction of these pairs — averaged over 100 configurations randomly chosen — falling either into the top 400 $|J_{ij}|$ (empty symbols) or into the top 400 values of $\min(\Delta\mathscr{E}_i, \Delta\mathscr{E}_j)$ (filled symbols). This second measure improves only slightly the estimation of strong short-range epistasis but it does so dramatically for long-range one. (C) Same plot as (B) where we show the fraction of the average epistasis $\langle\Delta\Delta\mathscr{F}_{ij}\rangle$ (estimated from 1.5×10^3 randomly chosen configurations of the MSA) that one would predict either via $|J_{ij}|$ or $\min(\Delta\mathscr{E}_i, \Delta\mathscr{E}_j)$. The prediction at short distance is rather accurate, with the predicted fraction reaching 1 for the maximally epistatic pairs; at long distance, signal on long-range epistasis captured by $|J_{ij}|$ is almost absent while the prediction by $\min(\Delta\mathscr{E}_i, \Delta\mathscr{E}_j)$ stands out for its precision.

C.3 Direct Coupling Analysis: inference procedure

In a maximum-entropy approach, extracting information from MSAs is cast as an inverse problem, i.e. inferring the set of parameters which enable the model, an Ising model in our framework, to reproduce certain observed statistical properties [79, 234]. The exact solution of this problem is found by maximum likelihood algorithms, which search for the set of couplings J_{ij} and fields h_i by maximising the likelihood of the model, specified by such parameters, to have produced data with the given statistics of single-site and pairwise frequencies, in our case. This exact maximization is often infeasible, therefore to tackle the inverse problem several approximate techniques have been developed.

Inference algorithm: ACE with a maximum likelihood refinement [93]. We resort to the Adaptive Cluster Expansion (ACE), an expansion of the entropy into contributions from clusters of spins [89, 90, 93]. We use the package made available by Barton https://github.com/johnbarton/ACE. The implementation consists of first a run of ACE followed by a proper maximum likelihood refinement via a Monte Carlo sampling (QLS routine), which takes as starting set of fields and couplings the ACE-inferred ones. Different parameters for the ACE and QLS routines can be set by the user, e.g. γ_2, the L_2-norm regularization strength for couplings which penalizes spurious large absolute values induced by undersampling. A natural value is $\gamma_2 = 1/M$, M being the size of the MSA.

To help convergence, we have chosen for ACE a higher value $\gamma_2 = 10^{-2}$ and $\theta = 10^{-5}$, this is the threshold at which the algorithm will run then exit, see [90]. In the further refinement by QLS, we have set mcb, the number of Monte Carlo steps used to estimate the inference error, to 200000 and $\gamma_2 = 1/M$. Having full control of the numerical evolution, we have tried to avoid undersampling issues by generating a large number of configurations $M = 135000$, which leads to $\gamma_2 \approx 0.7 \times 10^{-5}$. For the inference, we remove from sequences the 6 links at the active and allosteric sites as they are always associated to the symbol 1 (always occupied by a spring), so the number of parameters to infer is $N'_c + N'_c(N'_c - 1)/2 \sim 81000$ with $N'_c = N_c - 6 = 402$. Given the high frequency of 1s the sequences are given to the algorithm are flipped so to get a sparse matrix and improve numerical efficiency. We have verified that low values of the L_2-regularization allow us to obtain the maximal generative performance compatible with the model (in comparison to higher regularization). By default the L_2 regularization of fields is $0.01 \times \gamma_2$. In Fig. 3.4A, it is shown that the result of the inference is a model perfectly able to reproduce the first and second order statistics — as it should by construction, but that fails at reproducing higher order statistics.

Choice of the gauge. The chosen gauge is the consensus gauge, in the following the transformation between Ising and consensus gauge is discussed, with the use of the Ising Hamiltonian H, here rewritten by expliciting the parameters — field h and coupling J_{ij} — as function of the degrees of freedom — links occupancy $\sigma = 0, 1$

$$H = \frac{1}{2}\sum_{ij}\left[J_{ij}(1,1)\sigma_i\sigma_j + J_{ij}(1,0)\sigma_i(1-\sigma_j) + J_{ij}(0,1)(1-\sigma_i)\sigma_j + J_{ij}(0,0)(1-\sigma_i)(1-\sigma_j)\right] + \\ + \sum_i [h_i(0)(1-\sigma_i) + h_i(1)\sigma_i] \tag{C.11}$$

The Ising gauge imposes

$$J_{ij}(1,1) = J_{ij}(0,0) = J_{ij} \tag{C.12}$$

$$J_{ij}(0,1) = J_{ij}(1,0) = -J_{ij}, \tag{C.13}$$

and the resulting Hamiltonian H_I reads

$$\begin{aligned} H_I &= \frac{1}{2}\sum_{ij}\left[J_{ij}\sigma_i\sigma_j - J_{ij}\sigma_i(1-\sigma_j) - J_{ij}(1-\sigma_i)\sigma_j + J_{ij}(1-\sigma_i)(1-\sigma_j)\right] + \sum_i [h_i(0)(1-\sigma_i) + h_i(1)\sigma_i] \\ &= \frac{1}{2}\sum_{ij} 4J_{ij}\sigma_i\sigma_j + \sum_i \left[h_i(1) - h_i(0) - 2\sum_j J_{ij}\right]\sigma_i + \frac{1}{2}\sum_{ij} J_{ij} + \sum_i h_i(0) \qquad \text{(C.14)} \\ &\equiv H_C = \frac{1}{2}\sum_{ij} J^c_{ij}(1,1)\sigma_i\sigma_j + \sum_i h^c_i(1)\sigma_i + c. \qquad \text{(C.15)} \end{aligned}$$

Hence the transformations from the Ising (H_I) to the consensus gauge (H_C) are

$$J^c_{ij}(1,1) = 4J_{ij} \tag{C.16}$$

$$h^c_i(1) = h_i(1) - h_i(0) - 2\sum_j J_{ij} \tag{C.17}$$

$$c = \frac{1}{2}\sum_{ij} J_{ij} + \sum_i h_i(0). \tag{C.18}$$

Performance of mfDCA. For a comparison, we have considered also mean-field DCA (mfDCA) [81], derived from a mean-field factorized ansatz for the Boltzmann-Gibbs distribution Eq. 3.8. As discussed in the Introduction, couplings in mfDCA are given by $J_{ij} = -(\boldsymbol{C}^{-1})_{ij}$, where $\boldsymbol{C}_{ij} = \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle$ is the covariance of the MSA. Typically $\boldsymbol{C}$ is not invertible due to undersampling, making it necessary to add a pseudocount λ, see [69]. Ref. [235] shows that a pseudocount also helps correct for the systematic biases introduced by the mean field approximation. For this reason, we have used a pseudocount λ and chosen its value as $\lambda = 0.5$, which allows the best comparison to the ACE results, see Fig. C.3A.

It is noteworthy that in this way a computationally cheap technique as mfDCA yields a pattern of top J_{ij} strikingly similar to the one of a very accurate inference achieved by the combination of ACE and maximum likelihood. Therefore mfDCA, while extremely poor as a generative model, exhibits a good performance at reconstructing the distribution of relevant couplings,

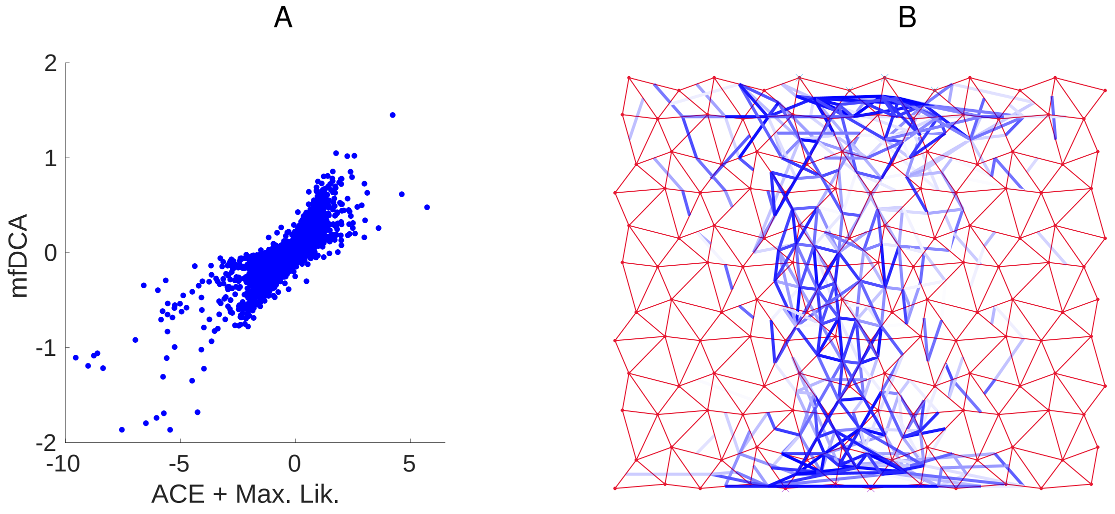

Figure C.3 – (A) Scatter plot comparing J_{ij} inferred via mfDCA to the direct couplings of ACE with refinement.: the pseudocount in mfDCA has been set to $\lambda = 0.5$ in such a way as to obtain the highest correlation between the two. (B) Spatial distribution of top 400 mfDCA-inferred couplings on the network. The reconstruction of the topology of relevant couplings is rather robust with respect to the choice of more approximate inference methods as mfDCA. As in Fig. 3.8A (inset) of Chapter 3, they are concentrated at short range, i.e. they connect links lying close either to the active site or the allosteric site and in the central high-shear path. Long-range mfDCA couplings, connecting links around respectively allosteric and active site, are weaker and appear among the top 600-1000 ones, implying an even worse performance at predicting long range epistasis than ACE + refinement.

as shown in Fig. C.3B.

C.4 Mutation costs and generative performance in the inferred Ising model

Swap mutations are enforced by the in-silico evolution. Costs of double mutations, i.e. joint mutations affecting links i and j, can be computed in the original model via fitness changes $\Delta\mathscr{F}_{ij} = \mathscr{F} - \mathscr{F}_{ij}$, where $\mathscr{F}_{ij}$ is the fitness after springs in i and j have been mutated. A double mutation can correspond either to (i) adding two springs at links i and j (i.e. $\sigma_i = \sigma_j = 1$) or removing them (i.e. $\sigma_i = \sigma_j = 0$) or to (ii) moving a spring from link i to link j or viceversa (i.e. $\sigma_i = 0, \sigma_j = 1$ or $\sigma_i = 1, \sigma_j = 0$). Let us call the former *non-swap* mutations and the latter *swap* mutations.

Swap mutations conserve the total amount of springs which is fixed to 360 given the fixed value of coordination during the in-silico evolution $\langle z \rangle = 5$. Hence, these swap mutations are the ones performed by the evolution algorithm. As optimal allosteric configurations maximize fitness with respect to this type of mutations, we opt to consider solely this kind when we compare mutation costs in terms of fitness and inferred energy (see Fig. 3.6C). We define *effective* single mutation costs $\Delta\mathscr{F}_i$ and $\Delta\mathscr{E}_i$ by taking, for each link, the swap with a link in the external region which is more rigid, as visible in e.g. Fig. 3.5, and thus mutations are completely neutral, whose cost is rougly zero.

Generative power. For the generative step, we implement a Monte Carlo sampling which which follows a swap-like dynamics as for the original numerical evolution. This ensures to select sequences from the inferred model that are structurally as close as possible to the initial data, i.e. with the same average coordination $\langle z \rangle = 5$, that is necessary for a consistent comparison.

Relaxing this constraint in the sampling leads to sequences endowed with higher internal variability, yet lying in the same range on fitness. Hence, the inferred model incorporates rather well the information on the fixed amount of springs. The parameters of the Ising model are inferred in such a way as to match single-site occupancy, which reflects the spatial pattern of coordination in the allosteric networks. In Fig. 3.5 we show that generated sequences, despite having lower fitness, reproduce successfully this property.

C.4.1 Comparison with conservation

Single-site frequency in protein alignments is informative about local conservation and is used as a standard measure of mutation costs at a certain position [236]. It can be fit by an independent-site Ising model where the energy (Eq. 3.9) contains only field terms and the inference from link occupancies $\langle \sigma_i \rangle$ translates to fitting the fields h_i to satisfy

$$\langle \sigma_i \rangle = \frac{e^{h_i \sigma_i}}{Z}, \tag{C.19}$$

where $Z = \sum_{\{\sigma\}} e^{h_i \sigma_i}$. Energy changes $\Delta \mathscr{E}_i$ upon point mutations can directly be computed. The energy cost of a mutation in an independent-site model is $\Delta \mathscr{E}_i = (2\sigma_i - 1) h_i$, where the field h_i can be determined up to a constant α by inverting Eq. C.19

$$h_i = \log(\langle \sigma_i \rangle) + \alpha . \tag{C.20}$$

The sequences in the MSA have a fixed average coordination $z = 5$ resulting in an average occupancy of $\bar{\sigma} = 360/408 = 0.88$ in the case $L = 12$ with periodic boundaries, see Chapter 1, Sec. 1.4. Hence, the field in the independent model should be non zero when the occupancy is away from $\bar{\sigma}$ to consistently describe conservation of both occupied and vacant links $\sigma = 0$ and $\sigma = 1$. In other words, the condition of zero field $h = 0$ should result in $\langle \sigma_i \rangle = \bar{\sigma}$. To satisfy this constraint the independent model of Eq. C.19 has to be modified as follows

$$\langle \sigma_i \rangle \equiv \langle \sigma_i \rangle_b = \frac{e^{h_i \sigma_i} f(\sigma_i)}{Z_b}, \tag{C.21}$$

where $f(\sigma_i = 1) = \bar{\sigma}$ and $f(\sigma_i = 0) = 1 - \bar{\sigma}$. The partition function Z_b is written as $Z_b = e^h \bar{\sigma} + (1 - \bar{\sigma})$.

The consensus gauge $h_i(\sigma = 0) = 0$ has been chosen, so that only $h_i(\sigma = 1)$ is non zero. Hence, the resulting field is

$$h_i = \log\left(\frac{\langle\sigma_i\rangle(1-\bar{\sigma})}{\bar{\sigma}(1-\langle\sigma_i\rangle)}\right), \tag{C.22}$$

describing how the observed occupancy of a link i, $\langle\sigma_i\rangle$, is biased away from the average occupancy $\bar{\sigma} = 360/408 = 0.88$. Indeed,$\langle\sigma_i\rangle = \bar{\sigma}$ is recovered when $h = 0$.

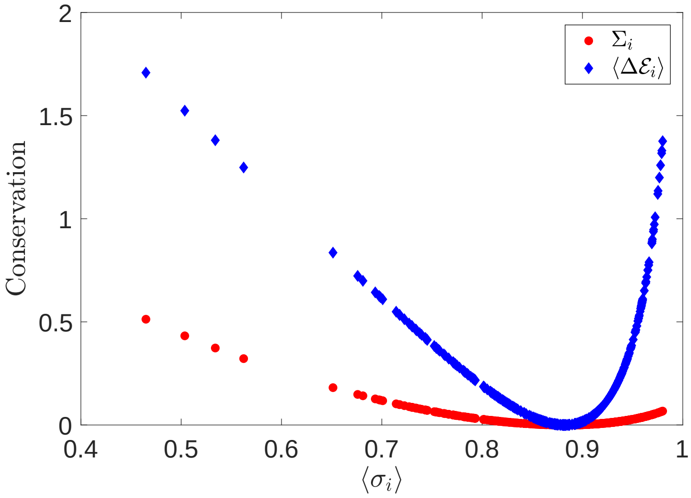

Figure C.4 – Conservation as function of the observed frequency $\langle\sigma_i\rangle$. The blue diamonds are obtained via Eq. C.23, while the red circles from the definition in Eq. 1.12 [100, 155]. The definition of Eq. C.23 better represents the structure of our data where a proper measure of conservation needs to take into account the background frequency $\bar{\sigma} = 0.88$.

The average energy cost of single mutations can be computed directly as

$$\langle\Delta\mathscr{E}_i\rangle = (2\langle\sigma_i\rangle_b - 1)h_i \tag{C.23}$$

$$= \left[\frac{2\langle\sigma_i\rangle(1-\bar{\sigma})}{\bar{\sigma}(1-\langle\sigma_i\rangle) + \langle\sigma_i\rangle(1-\bar{\sigma})} - 1\right] h_i\,. \tag{C.24}$$

On average the cost of single mutations, $\langle\Delta\mathscr{E}_i\rangle$, gives a measure of the conservation of link i, as it is 0 when $\langle\sigma_i\rangle = \bar{\sigma}$ and it increases the more link i tends to be either occupied or vacant. The performance of this measure of conservation is compared in Fig. C.4 with conservation defined as in Eq. 1.12 [100, 155].

The improvement achieved by the pairwise model over this conservation-based measure of mutation costs is extremely significant, see inset of Fig. 3.6C.

On the one hand, conservation is a purely local measure — it takes into account how a particular position is crucial to the propagation of the allosteric response. Including pairwise couplings proves to be necessary to capture the context-dependence of mutation costs, and thus must be included for their quantitative prediction.

On the other hand, the degree itself of structural conservation is rather low due to the heterogeneity of the shear-design MSA: the conformation, precise location and size of the shear path, hence the role of each link, can vary from architecture to architecture, leading to low structural conservation, with peaks only around the active and allosteric site. Conservation is found much higher *within* one set of dynamically related solutions (as for Fig. 3.3A), corresponding to one realization of the shear design among the many included in the MSA — see in particular Fig. Fig. 1.13.

C.5 Simple model illustrating the failure of DCA

To explain the discrepancy between short-range and long-range DCA-predictions of epistasis, we resort to the simple model of Fig. 3.10 as discussed in Chapter 3. We assign to all the 49 functional configurations the same fitness $\mathscr{F}$, all the other $2^8 - 49$ configurations would not belong to the sample of optimal configurations and are taken with zero fitness, thus $\Delta\mathscr{F} = 0$ if a mutation (single or double) results in a configuration still belonging to the optimal sample and $\Delta\mathscr{F} = \mathscr{F}$ otherwise. If we model each unit as a spin $\sigma = 0, 1$, this fitness function can be mathematically written as

$$\mathscr{F} = \mathscr{F}(\sigma_1\sigma_2 + \sigma_3\sigma_4 - \sigma_1\sigma_2\sigma_3\sigma_4) \cdot (\sigma_5\sigma_6 + \sigma_7\sigma_8 - \sigma_5\sigma_6\sigma_7\sigma_8) \tag{C.25}$$

i.e. it introduces high order couplings both at short (within groups and subparts) and long range (across subparts).
We can estimate average mutation costs by counting how frequently mutations would lead to a configuration outside of the optimal sample, yielding

$$\Delta\Delta\mathscr{F}_{12} = \Delta\mathscr{F}_{12} - \Delta\mathscr{F}_1 - \Delta\mathscr{F}_2 = 21/49\mathscr{F} - 21/49\mathscr{F} - 21/49\mathscr{F} = -21/49\mathscr{F} \tag{C.26}$$

$$\Delta\Delta\mathscr{F}_{15} = 33/49\mathscr{F} - 21/49\mathscr{F} - 21/49\mathscr{F} = -9/49\mathscr{F} \tag{C.27}$$

$$\frac{|\Delta\Delta\mathscr{F}_{12}|}{|\Delta\Delta\mathscr{F}_{15}|} = 21/9 \approx 2.3 \tag{C.28}$$

Next, by a simple likelihood maximization we infer the set of J_{ij} and h_i compatible with $\langle\sigma_i\rangle$ and $\langle\sigma_i\sigma_j\rangle$, single-site and pairwise frequencies of the optimal sample. We estimate $J_{12} = 1.18$ and $J_{15} = 0.40$, thus the prediction by DCA

$$\frac{|\Delta\Delta\mathscr{E}_{12}|}{|\Delta\Delta\mathscr{E}_{15}|} = \frac{|J_{12}(2\langle\sigma_1\rangle + 2\langle\sigma_2\rangle - 4\langle\sigma_1\sigma_2\rangle - 1)|}{|J_{15}(2\langle\sigma_1\rangle + 2\langle\sigma_5\rangle - 4\langle\sigma_1\sigma_5\rangle - 1)|} = \frac{|J_{12}(-21/49)|}{|J_{15}(-9/49)|} \approx 6.9 \tag{C.29}$$

i.e. the DCA prediction is significantly biased towards short-range epistasis. Due to symmetry of our model, epistasis and the DCA-prediction for any combination of units in the two subparts is the same as for units 1 and 5; similarly, the result for 2 units within the same group is given by the values for units 1 and 2. For the remaining combinations of units, i.e. the ones belonging the same subpart but to different groups (e.g. $i = 1$ and $j = 3$) we obtain that epistasis is weaker compared to units within the same group

$$\frac{|\Delta\Delta\mathscr{F}_{12}|}{|\Delta\Delta\mathscr{F}_{13}|} = \frac{|-21/49\mathscr{F}|}{|-7/49\mathscr{F}|} = 3 \tag{C.30}$$

Since each subpart can be of different type (OR gate), units from different groups (i.e. types) are less tightly constrained by function. The DCA-prediction does not underestimate epistasis as for units of different subparts (i.e. at long distance) with

$$\frac{|\Delta\Delta\mathscr{E}_{12}|}{|\Delta\Delta\mathscr{E}_{13}|} = \frac{|J_{12}(-21/49)|}{|J_{13}(7/49)|} \approx 3.5 \tag{C.31}$$

where $J_{13} = -1.01$. From Eq. C.28, Eq. C.29, Eq. C.30 and Eq. C.31 it is straightforward to calculate $|\Delta\Delta\mathscr{E}_{13}|/|\Delta\Delta\mathscr{E}_{12}| \times |\Delta\Delta\mathscr{F}_{12}|/|\Delta\Delta\mathscr{F}_{13}| \approx 0.86$ and $|\Delta\Delta\mathscr{E}_{15}|/|\Delta\Delta\mathscr{E}_{12}| \times |\Delta\Delta\mathscr{F}_{12}|/|\Delta\Delta\mathscr{F}_{15}| \approx 0.33$.

C.5.1 Feedforward neural network

To understand which machine learning tools could improve the prediction of epistasis in the simple model, we have built a feed-forward neural network performing least-squares regression of sequence data based on their fitness, see Fig. 3.16 in Chapter 3.

For data in the training set, we provide the network with both the input sequence and the target answer, i.e. a label 1 (standing for fitness $\mathscr{F}$) or 0. We vary the size of the training set from 10% to 80% of the $2^8 = 256$ total sequences and we keep the remaining sequences of the sample for validation of the accuracy of prediction. We learn the weights, i.e. the connections between layers, which minimize the mean squared error between the output of the network and the target answers by stochastic gradient descent from a random orthogonal initialization.

The 10% of learning runs with the best performance on the training dataset reach an average training error ranging between $\sim 4 \times 10^{-8}$ for a training set with 10% of the sample (25 configurations) to $\sim 3 \times 10^{-10}$ with 80%; the average validation error for the same runs is between $\sim 3 \times 10^{-1}$ and $\sim 2 \times 10^{-2}$ respectively.

We repeated the learning with an architecture where the width of the first hidden layer is bigger than the length of input data, for instance 16 and 32. For a width of 16 hidden units, the top 10% of trainings maintains an average accuracy on the training set of order 10^{-8} for the smaller training set (10% of the sample) and of order 10^{-10} for the largest one (80% of the sample); the corresponding average validation errors are $\sim 3 \times 10^{-1}$ and $\sim 4 \times 10^{-2}$.

When increasing further the first layer to a width of 32, we also added a dropout (here equal to 0.3) to balance the increase of parameters to learn with the same amount of data and avoid overfitting. In this way we obtained that the training error, averaged over the 10% best runs, was higher (from $\sim 8 \times 10^{-5}$ for a training set with 10% of the sample to $\sim 6 \times 10^{-6}$ with 80%) but the performance on the validation set was better (respective average errors of $\sim 2 \times 10^{-1}$ and 10^{-4}). Provided that the training set is not too small, these numerical tests confirm that a trained neural network, when presented with an optimal sequence mutated at some position, can predict the value of its fitness with good accuracy in such a way as to predict $\Delta\mathscr{F} \sim 0$ when it still belongs to the optimal sample or $\Delta\mathscr{F} \sim 1$ if it does not. This ensures that also epistasis would be accurately predicted at any range.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell.* Garland Science, 4th edition, 2002.

[2] Ron Milo, Rob Phillips, and Rob Phillips. *Cell Biology by the Numbers.* Garland Science, December 2015.

[3] Cyrus Levinthal. How to fold Graciously. In *Mossbauer Spectroscopy in Biological Systems Proceedings*, volume 67, pages 22–24, 1969.

[4] Christian B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, July 1973. Publisher: American Association for the Advancement of Science Section: Articles.

[5] Chr Bohr, K. Hasselbalch, and August Krogh. Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt1. *Skandinavisches Archiv Für Physiologie*, 16(2):402–412, 1904.

[6] Archibald Vivian Hill. The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. *The Journal of Physiology*, 40:i–vii, January 1910.

[7] G. S. Adair and Jr With the collaboration of A. V. Bock and H. Field. The Hemoglobin System Vi. the Oxygen Dissociation Curve of Hemoglobin. *Journal of Biological Chemistry*, 63(2):529–545, March 1925. Publisher: American Society for Biochemistry and Molecular Biology.

[8] Linus Pauling. The Structure and Entropy of Ice and of Other Crystals with Some Randomness of Atomic Arrangement. *Journal of the American Chemical Society*, 57(12):2680–2684, December 1935. Publisher: American Chemical Society.

[9] I. M. Klotz. The application of the law of mass action to binding by proteins; interactions with calcium. *Archives of Biochemistry*, 9:109–117, January 1946.

[10] Jacques Monod and François Jacob. General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:389–401, January 1961. Publisher: Cold Spring Harbor Laboratory Press.

[11] Jean-Pierre Changeux. The Feedback Control Mechanism of Biosynthetic L-Threonine Deaminase by L-Isoleucine. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:313–318, January 1961.

[12] Jean-Pierre Changeux. Allosteric Interactions on Biosynthetic L-threonine Deaminase from E. coli K12. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:497–504, January 1963.

[13] Hesam N. Motlagh, James O. Wrabl, Jing Li, and Vincent J. Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, April 2014. Number: 7496 Publisher: Nature Publishing Group.

[14] LLC Schrödinger. The PyMOL Molecular Graphics System, Version 1.8. 2015.

[15] M. F. Perutz. Stereochemistry of Cooperative Effects in Haemoglobin: Haem–Haem Interaction and the Problem of Allostery. *Nature*, 228(5273):726–734, November 1970.

[16] M. F. Perutz, A. J. Wilkinson, M. Paoli, and G. G. Dodson. The Stereochemical Mechanism of the Cooperative Effects in Hemoglobin Revisited. *Annual Review of Biophysics and Biomolecular Structure*, 27(1):1–34, 1998.

[17] William S. Bialek. *Biophysics: searching for principles*. Princeton University Press, Princeton, NJ, 2012.

[18] M. F. Perutz, M. G. Rossmann, Ann F. Cullis, Hilary Muirhead, Georg Will, and A. C. T. North. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature*, 185(4711):416–422, February 1960.

[19] Jacques Monod, Jean-Pierre Changeux, and François Jacob. Allosteric proteins and cellular control systems. *Journal of Molecular Biology*, 6(4):306–329, April 1963.

[20] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12(1):88–118, May 1965.

[21] D. E. Koshland, G. Némethy, and D. Filmer. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits*. *Biochemistry*, 5(1):365–385, January 1966.

[22] William A. Eaton, Eric R. Henry, James Hofrichter, and Andrea Mozzarelli. Is cooperative oxygen binding by hemoglobin really understood? *Nature Structural Biology*, 6(4):351–358, April 1999. Number: 4 Publisher: Nature Publishing Group.

[23] Jean-Pierre Changeux. Allostery and the Monod-Wyman-Changeux Model After 50 Years. *Annual Review of Biophysics*, 41(1):103–133, 2012.

[24] Chung-Jung Tsai and Ruth Nussinov. A Unified View of "How Allostery Works". *PLoS Computational Biology*, 10(2):e1003394, February 2014.

[25] Hans Frauenfelder, Benjamin H. McMahon, Robert H. Austin, Kelvin Chu, and John T. Groves. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proceedings of the National Academy of Sciences*, 98(5):2370–2374, February 2001.

[26] Brian F. Volkman, Doron Lipson, David E. Wemmer, and Dorothee Kern. Two-State Allosteric Behavior in a Single-Domain Signaling Protein. *Science*, 291(5512):2429–2433, March 2001.

[27] A. Cooper and D. T. Dryden. Allostery without conformational change. A plausible model. *European biophysics journal: EBJ*, 11(2):103–109, 1984.

[28] Nataliya Popovych, Shangjin Sun, Richard H. Ebright, and Charalampos G. Kalodimos. Dynamically driven protein allostery. *Nature Structural & Molecular Biology*, 13(9):831–838, September 2006.

[29] Chung-Jung Tsai, Antonio del Sol, and Ruth Nussinov. Allostery: absence of a change in shape does not imply that allostery is not at play. *Journal of Molecular Biology*, 378(1):1–11, April 2008.

[30] Catherine M. Falcon and Kathleen S. Matthews. Engineered Disulfide Linking the Hinge Regions within Lactose Repressor Dimer Increases Operator Affinity, Decreases Sequence Selectivity, and Alters Allostery. *Biochemistry*, 40(51):15650–15659, December 2001.

[31] Wendell A. Lim. The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Current Opinion in Structural Biology*, 12(1):61–68, February 2002.

[32] Srivatsan Raman, Noah Taylor, Naomi Genuth, Stanley Fields, and George M. Church. Engineering allostery. *Trends in Genetics*, 30(12):521–528, December 2014.

[33] David Pincus, Jai P. Pandey, Zoë A. Feder, Pau Creixell, Orna Resnekov, and Kimberly A. Reynolds. Engineering allosteric regulation in protein kinases. *Science Signaling*, 11(555):eaar3250, November 2018.

[34] Michele Vendruscolo and Christopher M. Dobson. Dynamic Visions of Enzymatic Reactions. *Science*, 313(5793):1586–1587, September 2006. Publisher: American Association for the Advancement of Science Section: Perspective.

[35] Michele Vendruscolo. Protein dynamics under light control. *Nature Chemical Biology*, 4(8):449–450, August 2008. Number: 8 Publisher: Nature Publishing Group.

[36] Xiaolan Yao, Michael K. Rosen, and Kevin H. Gardner. Estimation of the available free energy in a LOV2-J alpha photoswitch. *Nature Chemical Biology*, 4(8):491–497, August 2008.

[37] Hagai Abeliovich. An Empirical Extremum Principle for the Hill Coefficient in Ligand-Protein Interactions Showing Negative Cooperativity. *Biophysical Journal*, 89(1):76–79, July 2005.

[38] J. Kister, C. Poyart, and S. J. Edelstein. Oxygen-organophosphate linkage in hemoglobin A. The double hump effect. *Biophysical Journal*, 52(4):527–535, October 1987.

[39] Gordon G. Hammes, Yu-Chu Chang, and Terrence G. Oas. Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences*, 106(33):13737–13741, August 2009. Publisher: National Academy of Sciences Section: Biological Sciences.

[40] Qiang Cui and Martin Karplus. Allostery and cooperativity revisited. *Protein Science: A Publication of the Protein Society*, 17(8):1295–1307, August 2008.

[41] Gozde Kar, Ozlem Keskin, Attila Gursoy, and Ruth Nussinov. Allostery and population shift in drug discovery. *Current Opinion in Pharmacology*, 10(6):715–722, December 2010.

[42] D. Thirumalai, Changbong Hyeon, Pavel I. Zhuravlev, and George H. Lorimer. Symmetry, Rigidity, and Allosteric Signaling: From Monomeric Proteins to Molecular Machines. *Chemical Reviews*, 119(12):6788–6821, June 2019.

[43] J.J. Hopfield. Relation between structure, co-operativity and spectra in a model of hemoglobin action. *Journal of Molecular Biology*, 77(2):207–222, June 1973.

[44] Michael D. Daily and Jeffrey J. Gray. Local motions in a benchmark of allosteric proteins. *Proteins*, 67(2):385–399, May 2007.

[45] Mark Gerstein, Arthur M. Lesk, and Cyrus Chothia. Structural Mechanisms for Domain Movements in Proteins. *Biochemistry*, 33(22):6739–6749, June 1994. Publisher: American Chemical Society.

[46] Michael R. Mitchell, Tsvi Tlusty, and Stanislas Leibler. Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings. *Proceedings of the National Academy of Sciences*, 113(40):E5847–E5855, October 2016.

[47] Doeke R. Hekstra, K. Ian White, Michael A. Socolich, Robert W. Henning, Vukica Šrajer, and Rama Ranganathan. Electric-field-stimulated protein mechanics. *Nature*, 540(7633):400–405, December 2016. Number: 7633 Publisher: Nature Publishing Group.

[48] Gonzalo Jiménez-Osés, Sílvia Osuna, Xue Gao, Michael R. Sawaya, Lynne Gilson, Steven J. Collier, Gjalt W. Huisman, Todd O. Yeates, Yi Tang, and K. N. Houk. The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nature Chemical Biology*, 10(6):431–436, June 2014.

[49] Ruth Nussinov and Chung-Jung Tsai. Allostery without a conformational change? Revisiting the paradigm. *Current Opinion in Structural Biology*, 30:17–24, February 2015.

[50] 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(1):505–515, January 2001.

[51] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, 79(16):3090–3093, October 1997. Publisher: American Physical Society.

[52] Monique M. Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, 77(9):1905–1908, August 1996.

[53] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 33(3):417–429, 1998.

[54] Konrad Hinsen. Normal Mode Theory and Harmonic Potential Approximations. In *Normal Mode Analysis : Theory and Applications to Biologicaland Chemical Systems*, pages 1–16. 2006.

[55] K. N. Trueblood, H.-B. Bürgi, H. Burzlaff, J. D. Dunitz, C. M. Gramaccioli, H. H. Schulz, U. Shmueli, and S. C. Abrahams. Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica Section A: Foundations of Crystallography*, 52(5):770–781, September 1996. Number: 5 Publisher: International Union of Crystallography.

[56] F. Tama and Y. H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, 14(1):1–6, January 2001.

[57] Paolo De Los Rios, Fabio Cecconi, Anna Pretre, Giovanni Dietler, Olivier Michielin, Francesco Piazza, and Brice Juanico. Functional Dynamics of PDZ Binding Domains: A Normal-Mode Analysis. *Biophysical Journal*, 89(1):14–21, July 2005.

[58] W. Zheng, B. R. Brooks, and D. Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences*, 103(20):7664–7669, May 2006.

[59] O. Miyashita, J. N. Onuchic, and P. G. Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences*, 100(22):12570–12575, October 2003. ISBN: 9782135471108 Publisher: National Academy of Sciences Section: Physical Sciences.

[60] Jianpeng Ma. Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes. *Structure*, 13(3):373–380, March 2005.

[61] Cristian Micheletti, Paolo Carloni, and Amos Maritan. Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models. *Proteins: Structure, Function, and Bioinformatics*, 55(3):635–645, March 2004.

[62] Guang Song and Robert L. Jernigan. vGNM: A Better Model for Understanding the Dynamics of Proteins in Crystals. *Journal of Molecular Biology*, 369(3):880–893, June 2007.

[63] Paul R. Ehrlich and Peter H. Raven. Butterflies and Plants: A Study in Coevolution. *Evolution*, 18(4):586–608, 1964. Publisher: [Society for the Study of Evolution, Wiley].

[64] John N. Thompson. *The Coevolutionary Process*. University of Chicago Press, November 1994.

[65] Stephan C. Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, January 2008. Number: 1 Publisher: Nature Publishing Group.

[66] Giancarlo Croce. *Towards a genome-scale coevolutionary analysis*. PhD thesis, Université Sorbonne, 2019.

[67] Shimon Bershtein, Michal Segal, Roy Bekerman, Nobuhiko Tokuriki, and Dan S. Tawfik. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121):929–932, December 2006. Number: 7121 Publisher: Nature Publishing Group.

[68] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Molecular Biology and Evolution*, 35(4):1018–1027, April 2018. Publisher: Oxford Academic.

[69] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, January 2018. Publisher: IOP Publishing.

[70] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, November 2016. Publisher: Oxford Academic.

[71] Georg Casari, Chris Sander, and Alfonso Valencia. A method to predict functional residues in proteins. *Nature Structural Biology*, 2(2):171–178, February 1995. Number: 2 Publisher: Nature Publishing Group.

[72] Olivier Lichtarge, Henry R. Bourne, and Fred E. Cohen. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *Journal of Molecular Biology*, 257(2):342–358, March 1996.

[73] S. Karlin and L. Brocchieri. Evolutionary conservation of RecA genes in relation to protein structure and function. *Journal of Bacteriology*, 178(7):1881–1894, April 1996. Publisher: American Society for Microbiology Journals.

[74] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* April 1998.

[75] John A. Capra and Mona Singh. Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882, August 2007. Publisher: Oxford Academic.

[76] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.

[77] I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering, Design and Selection*, 7(3):349–358, March 1994. Publisher: Oxford Academic.

[78] E. Neher. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, 91(1):98–102, January 1994. Publisher: National Academy of Sciences Section: Research Article.

[79] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66(3):197–261, July 2017.

[80] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, January 2009. Publisher: National Academy of Sciences Section: Physical Sciences.

[81] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, December 2011. Publisher: National Academy of Sciences Section: PNAS Plus.

[82] William Bialek and Rama Ranganathan. Rediscovering the power of pairwise interactions. *arXiv:0712.4397 [q-bio]*, December 2007. arXiv: 0712.4397.

[83] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957. Publisher: American Physical Society.

[84] Marc Mézard and Thierry Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1):107–113, January 2009.

[85] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, April 2006. Number: 7087 Publisher: Nature Publishing Group.

[86] Simona Cocco, Stanislas Leibler, and Rémi Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, August 2009. Publisher: National Academy of Sciences Section: Physical Sciences.

[87] Vitor Sessak and Rémi Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42(5):055001, January 2009. Publisher: IOP Publishing.

[88] H. J. Kappen and F. B. Rodríguez. Efficient Learning in Boltzmann Machines Using Linear Response Theory. *Neural Computation*, 10(5):1137–1156, July 1998. Publisher: MIT Press.

[89] S. Cocco and R. Monasson. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. *Physical Review Letters*, 106(9):090601, March 2011. Publisher: American Physical Society.

[90] S. Cocco and R. Monasson. Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests. *Journal of Statistical Physics*, 147(2):252–314, April 2012.

[91] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, June 2010. Publisher: Institute of Mathematical Statistics.

[92] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1):012707, January 2013. Publisher: American Physical Society.

[93] J. P. Barton, E. De Leonardis, A. Coucke, and S. Cocco. ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, 32(20):3089–3097, October 2016. Publisher: Oxford Academic.

[94] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298, January 2017. Publisher: American Association for the Advancement of Science Section: Report.

[95] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones,

David Silver, Koray Kavukcuoglu, and Demis Hassabis. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics*, 87(12):1141–1148, 2019.

[96] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020. Number: 7792 Publisher: Nature Publishing Group.

[97] Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. DESTINI: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports*, 9(1):1–13, March 2019. Number: 1 Publisher: Nature Publishing Group.

[98] Ivan Anishchenko, Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences*, 114(34):9122–9127, August 2017. Publisher: National Academy of Sciences Section: Biological Sciences.

[99] Steve W. Lockless and Rama Ranganathan. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286(5438):295–299, October 1999. Publisher: American Association for the Advancement of Science Section: Report.

[100] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4):774–786, August 2009.

[101] Shou-Wen Wang, Anne-Florence Bitbol, and Ned S. Wingreen. Revealing evolutionary constraints on proteins through sequence analysis. *PLOS Computational Biology*, 15(4):e1007010, April 2019.

[102] Richard N. McLaughlin Jr, Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, November 2012. Number: 7422 Publisher: Nature Publishing Group.

[103] Francis C. Peterson, Rhiannon R. Penkert, Brian F. Volkman, and Kenneth E. Prehoda. Cdc42 Regulates the Par-6 PDZ Domain through an Allosteric CRIB-PDZ Transition. *Molecular Cell*, 13(5):665–676, March 2004.

[104] Gürol M. Süel, Steve W. Lockless, Mark A. Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology*, 10(1):59–69, January 2003. Number: 1 Publisher: Nature Publishing Group.

[105] William P. Russ, Drew M. Lowery, Prashant Mishra, Michael B. Yaffe, and Rama Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, September 2005. Number: 7058 Publisher: Nature Publishing Group.

[106] Robert G Smock, Olivier Rivoire, William P Russ, Joanna F Swain, Stanislas Leibler, Rama Ranganathan, and Lila M Gierasch. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Molecular Systems Biology*, 6(1):414, January 2010. Publisher: John Wiley & Sons, Ltd.

[107] Kimberly A. Reynolds, Richard N. McLaughlin, and Rama Ranganathan. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell*, 147(7):1564–1575, December 2011.

[108] Mark E. Hatley, Steve W. Lockless, Scott K. Gibson, Alfred G. Gilman, and Rama Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proceedings of the National Academy of Sciences*, 100(24):14445–14450, November 2003. Publisher: National Academy of Sciences Section: Biological Sciences.

[109] Ernesto J. Fuentes, Channing J. Der, and Andrew L. Lee. Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain. *Journal of Molecular Biology*, 335(4):1105–1115, January 2004.

[110] Gerhard Stock and Peter Hamm. A non-equilibrium approach to allosteric communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1749):20170187, June 2018. Publisher: Royal Society.

[111] Sebastian Buchenberg, Florian Sittel, and Gerhard Stock. Time-resolved observation of protein allosteric communication. *Proceedings of the National Academy of Sciences*, 114(33):E6804–E6811, August 2017. Publisher: National Academy of Sciences Section: PNAS Plus.

[112] Ilya A. Balabin, Weitao Yang, and David N. Beratan. Coarse-grained modeling of allosteric regulation in protein receptors. *Proceedings of the National Academy of Sciences*, 106(34):14253–14258, August 2009.

[113] Tom C.B. McLeish, Martin J. Cann, and Thomas L. Rodgers. Dynamic Transmission of Protein Allostery without Structural Change: Spatial Pathways or Global Modes? *Biophysical Journal*, 109(6):1240–1250, September 2015.

[114] Jeeyeon Lee, Madhusudan Natarajan, Vishal C. Nashine, Michael Socolich, Tina Vo, William P. Russ, Stephen J. Benkovic, and Rama Ranganathan. Surface Sites for Engineering Allosteric Control in Proteins. *Science*, 322(5900):438–442, October 2008. Publisher: American Association for the Advancement of Science Section: Report.

[115] Marc Ostermeier. Designing switchable enzymes. *Current Opinion in Structural Biology*, 19(4):442–448, August 2009.

[116] David Pincus, Orna Resnekov, and Kimberly A. Reynolds. An evolution-based strategy for engineering allosteric regulation. *Physical Biology*, 14(2):025002, April 2017. Publisher: IOP Publishing.

[117] Nikolay V. Dokholyan. Controlling Allosteric Networks in Proteins. *Chemical Reviews*, 116(11):6463–6487, June 2016.

[118] Onur Dagliyan, Nikolay V. Dokholyan, and Klaus M. Hahn. Engineering proteins for allosteric control by light or ligands. *Nature Protocols*, 14(6):1863–1883, June 2019. Number: 6 Publisher: Nature Publishing Group.

[119] Tiberiu Teşileanu, Lucy J. Colwell, and Stanislas Leibler. Protein Sectors: Statistical Coupling Analysis versus Conservation. *PLOS Computational Biology*, 11(2):e1004091, February 2015. Publisher: Public Library of Science.

[120] Olivier Rivoire. Elements of Coevolution in Biological Sequences. *Physical Review Letters*, 110(17):178102, April 2013. Publisher: American Physical Society.

[121] Mathieu Hemery and Olivier Rivoire. Evolution of sparsity and modularity in a model of protein allostery. *Physical Review E*, 91(4):042704, April 2015.

[122] Olivier Rivoire. Parsimonious evolutionary scenario for the origin of allostery and coevolution patterns in proteins. *Physical Review E*, 100(3):032411, September 2019. Publisher: American Physical Society.

[123] Andrew L. Ferguson, Jaclyn K. Mann, Saleha Omarjee, Thumbi Ndung'u, Bruce D. Walker, and Arup K. Chakraborty. Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity*, 38(3):606–617, March 2013.

[124] Sewall Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, March 1931.

[125] C. K. Sruthi and Meher Prakash. Deep2Full: Evaluating strategies for selecting the minimal mutational experiments for optimal computational predictions of deep mutational scan outcomes. *PLOS ONE*, 15(1):e0227621, January 2020. Publisher: Public Library of Science.

[126] Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, August 2014. Number: 8 Publisher: Nature Publishing Group.

[127] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, December 2009. Number: 12 Publisher: Nature Publishing Group.

[128] Patrick C. Phillips. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, November 2008. Number: 11 Publisher: Nature Publishing Group.

[129] J. Arjan G. M. de Visser, Tim F. Cooper, and Santiago F. Elena. The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences*, 278(1725):3617–3624, December 2011. Publisher: Royal Society.

[130] Tyler N. Starr and Joseph W. Thornton. Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218, 2016.

[131] Eric A. Ortlund, Jamie T. Bridgham, Matthew R. Redinbo, and Joseph W. Thornton. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science*, 317(5844):1544–1548, 2007. Publisher: American Association for the Advancement of Science.

[132] Chandrasekhar Natarajan, Noriko Inoguchi, Roy E. Weber, Angela Fago, Hideaki Moriyama, and Jay F. Storz. Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin. *Science*, 340(6138):1324–1327, June 2013. Publisher: American Association for the Advancement of Science Section: Report.

[133] Karthik Shekhar, Claire F. Ruberman, Andrew L. Ferguson, John P. Barton, Mehran Kardar, and Arup K. Chakraborty. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Physical Review E*, 88(6):062705, December 2013.

[134] Victor H Salinas and Rama Ranganathan. Coevolution-based inference of amino acid interactions underlying protein function. *eLife*, 7:e34300, July 2018. Publisher: eLife Sciences Publications, Ltd.

[135] C. Anders Olson, Nicholas C. Wu, and Ren Sun. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, 24(22):2643–2651, November 2014.

[136] Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10(1):1–11, September 2019. Number: 1 Publisher: Nature Publishing Group.

[137] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, September 2005. Number: 7058 Publisher: Nature Publishing Group.

[138] Hugo Jacquin, Amy Gilson, Eugene Shakhnovich, Simona Cocco, and Rémi Monasson. Benchmarking Inverse Statistical Approaches for Protein Structure and Design with Exactly Solvable Models. *PLOS Computational Biology*, 12(5):e1004889, May 2016. Publisher: Public Library of Science.

[139] Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, October 1989. Publisher: American Chemical Society.

[140] Eugene Shakhnovich and Alexander Gutin. Enumeration of all compact conformations of copolymers with random sequence of links. *The Journal of Chemical Physics*, 93(8):5967–5971, October 1990. Publisher: American Institute of Physics.

[141] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus. Protein folding bottlenecks: A lattice Monte Carlo simulation. *Physical Review Letters*, 67(12):1665–1668, September 1991. Publisher: American Physical Society.

[142] Leonid Mirny and Eugene Shakhnovich. Protein Folding Theory: From Lattice to All-Atom Models. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):361–396, 2001.

[143] Jaclyn K. Mann, John P. Barton, Andrew L. Ferguson, Saleha Omarjee, Bruce D. Walker, Arup Chakraborty, and Thumbi Ndung'u. The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing. *PLOS Computational Biology*, 10(8):e1003776, August 2014. Publisher: Public Library of Science.

[144] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, 33(1):268–280, January 2016.

[145] Pierre Barrat-Charlaix, Matteo Figliuzzi, and Martin Weigt. Improving landscape inference by integrating heterogeneous data in the inverse Ising problem. *Scientific Reports*, 6(1):1–9, November 2016. Number: 1 Publisher: Nature Publishing Group.

[146] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017. Number: 2 Publisher: Nature Publishing Group.

[147] Christoph Feinauer and Martin Weigt. Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease. *arXiv:1701.07246 [q-bio]*, January 2017. arXiv: 1701.07246.

[148] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. Evolution-based design of chorismate mutase enzymes. *bioRxiv*, page 2020.04.01.020487, April 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[149] L. D. Landau, E. M. Lifshitz, A. M. Kosevich, and L. P. Pitaevskii. *Theory of Elasticity*. Elsevier, January 1986.

[150] Jie Liang and Ken A. Dill. Are Proteins Well-Packed? *Biophysical Journal*, 81(2):751–766, August 2001.

[151] J.-P. Bouchaud. Granular media: some ideas from statistical physics. *arXiv:cond-mat/0211196*, December 2002. arXiv: cond-mat/0211196.

[152] Andrea J. Liu, Sidney R. Nagel, Wim van Saarloos, and Matthieu Wyart. *The jamming scenario—an introduction and outlook.* Oxford University Press, July 2011.

[153] Gustavo Düring, Edan Lerner, and Matthieu Wyart. Phonon gap and localization lengths in floppy materials. *Soft Matter*, 9(1):146–154, November 2012. Publisher: The Royal Society of Chemistry.

[154] Edan Lerner, Eric DeGiuli, Gustavo Düring, and Matthieu Wyart. Breakdown of continuum elasticity in amorphous solids. *Soft Matter*, 10(28):5085–5092, June 2014.

[155] Le Yan, Riccardo Ravasio, Carolina Brito, and Matthieu Wyart. Architecture and co-evolution of allosteric materials. *Proceedings of the National Academy of Sciences*, 114(10):2526–2531, March 2017.

[156] Le Yan, Riccardo Ravasio, Carolina Brito, and Matthieu Wyart. Principles for Optimal Cooperativity in Allosteric Materials. *Biophysical Journal*, 114(12):2787–2798, June 2018.

[157] Riccardo Ravasio, Solange Marie Flatt, Le Yan, Stefano Zamuner, Carolina Brito, and Matthieu Wyart. Mechanics of Allostery: Contrasting the Induced Fit and Population Shift Scenarios. *Biophysical Journal*, 117(10):1954–1962, November 2019.

[158] Barbara Bravi, Riccardo Ravasio, Carolina Brito, and Matthieu Wyart. Direct coupling analysis of epistasis in allosteric materials. *PLOS Computational Biology*, 16(3):e1007630, March 2020. Publisher: Public Library of Science.

[159] Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):1–13, February 2020. Number: 1 Publisher: Nature Publishing Group.

[160] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, June 1997.

[161] Ivet Bahar and Robert L. Jernigan. Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms11Edited by I. Tinoco. *Journal of Molecular Biology*, 281(5):871–884, September 1998.

[162] Melik C. Demirel, Ali Rana Atilgan, Ivet Bahar, Robert L. Jernigan, and Burak Erman. Identification of kinetically hot residues in proteins. *Protein Science*, 7(12):2522–2532, 1998.

[163] Moon K Kim, Gregory S Chirikjian, and Robert L Jernigan. Elastic models of conformational transitions in macromolecules. *Journal of Molecular Graphics and Modelling*, 21(2):151–160, October 2002.

[164] M Delarue and Y. H Sanejouand. Simplified Normal Mode Analysis of Conformational Transitions in DNA-dependent Polymerases: the Elastic Network Model. *Journal of Molecular Biology*, 320(5):1011–1024, July 2002.

[165] Wenjun Zheng and Sebastian Doniach. A comparative study of motor-protein motions by using a simple elastic-network model. *Proceedings of the National Academy of Sciences*, 100(23):13253–13258, November 2003. Publisher: National Academy of Sciences Section: Biological Sciences.

[166] A. J. Rader, Daniel H. Vlad, and Ivet Bahar. Maturation Dynamics of Bacteriophage HK97 Capsid. *Structure*, 13(3):413–421, March 2005.

[167] Chakra Chennubhotla, A. J. Rader, Lee-Wei Yang, and Ivet Bahar. Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical Biology*, 2(4):S173–S180, November 2005. Publisher: IOP Publishing.

[168] Holger Flechsig. Design of Elastic Networks with Evolutionary Optimized Long-Range Communication as Mechanical Models of Allosteric Proteins. *Biophysical Journal*, 113(3):558–571, August 2017.

[169] Jason W. Rocks, Nidhi Pashine, Irmgard Bischofberger, Carl P. Goodrich, Andrea J. Liu, and Sidney R. Nagel. Designing allostery-inspired response in mechanical networks. *Proceedings of the National Academy of Sciences*, 114(10):2520–2525, March 2017.

[170] Tom McLeish, C. Schaefer, and A. C. von der Heydt. The ‘allosteron’ model for entropic allostery of self-assembly. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1749):20170186, June 2018. Publisher: Royal Society.

[171] Sandhya P Tiwari and Nathalie Reuter. Conservation of intrinsic dynamics in proteins—what have computational models taught us? *Current Opinion in Structural Biology*, 50:75–81, June 2018.

[172] Ivet Bahar, Ali Rana Atilgan, Melik C. Demirel, and Burak Erman. Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability. *Physical Review Letters*, 80(12):2733–2736, March 1998. Publisher: American Physical Society.

[173] P. W. Anderson. Absence of Diffusion in Certain Random Lattices. *Physical Review*, 109(5):1492–1505, March 1958. Publisher: American Physical Society.

[174] P. W. Anderson. Local moments and localized states. *Reviews of Modern Physics*, 50(2):191–201, April 1978. Publisher: American Physical Society.

[175] Yuichi Togashi and Alexander S. Mikhailov. Nonlinear relaxation dynamics in elastic networks and design principles of molecular machines. *Proceedings of the National Academy of Sciences*, 104(21):8697–8702, May 2007. Publisher: National Academy of Sciences Section: Physical Sciences.

[176] Ole Sigmund and Kurt Maute. Topology optimization approaches: A comparative review. *Structural and Multidisciplinary Optimization*, 48(6):1031–1055, December 2013.

[177] B. R. C. 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(1):1–13, August 2016. Number: 1 Publisher: Nature Publishing Group.

[178] Rhoda J. Hawkins and Tom C. B. McLeish. Coarse-Grained Model Of Entropic Allostery. *Physical Review Letters*, 93(9):098104, August 2004. Publisher: American Physical Society.

[179] Rhoda J. Hawkins and Tom C. B. McLeish. Coupling of Global and Local Vibrational Modes in Dynamic Allostery of Proteins. *Biophysical Journal*, 91(6):2055–2062, September 2006.

[180] T C B McLeish, T L Rodgers, and M R Wilson. Allostery without conformation change: modelling protein dynamics at multiple scales. *Physical Biology*, 10(5):056004, September 2013.

[181] Hedvika Toncrova and Tom C. B. McLeish. Substrate-Modulated Thermal Fluctuations Affect Long-Range Allosteric Signaling in Protein Homodimers: Exemplified in CAP. *Biophysical Journal*, 98(10):2317–2326, May 2010. Publisher: Elsevier.

[182] Ole Sigmund. On the Design of Compliant Mechanisms Using Topology Optimization. *Mechanics of Structures and Machines*, 25(4):493–524, January 1997.

[183] Shinji Nishiwaki, Mary I. Frecker, Seungjae Min, and Noboru Kikuchi. Topology optimization of compliant mechanisms using the homogenization method. *International Journal for Numerical Methods in Engineering*, 42(3):535–559, 1998.

[184] Jason W. Rocks, Henrik Ronellenfitsch, Andrea J. Liu, Sidney R. Nagel, and Eleni Katifori. Limits of multifunctionality in tunable networks. *Proceedings of the National Academy of Sciences*, 116(7):2506–2511, February 2019. Publisher: National Academy of Sciences Section: Physical Sciences.

[185] Tsvi Tlusty, Albert Libchaber, and Jean-Pierre Eckmann. Physical Model of the Genotype-to-Phenotype Map of Proteins. *Physical Review X*, 7(2):021037, June 2017. Publisher: American Physical Society.

[186] Sandipan Dutta, Jean-Pierre Eckmann, Albert Libchaber, and Tsvi Tlusty. Green function of correlated genes in a minimal mechanical model of protein evolution. *Proceedings of the National Academy of Sciences*, 115(20):E4559–E4568, May 2018. Publisher: National Academy of Sciences Section: PNAS Plus.

[187] Le Yan and Matthieu Wyart. Evolution of Covalent Networks under Cooling: Contrasting the Rigidity Window and Jamming Scenarios. *Physical Review Letters*, 113(21):215504, November 2014. Publisher: American Physical Society.

[188] Le Yan and Matthieu Wyart. Adaptive elastic networks as models of supercooled liquids. *Physical Review E*, 92(2):022310, August 2015. Publisher: American Physical Society.

[189] J. Clerk Maxwell. On the calculation of the equilibrium and stiffness of frames. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 27(182):294–299, April 1864.

[190] Le Yan, Jean-Philippe Bouchaud, and Matthieu Wyart. Edge mode amplification in disordered elastic networks. *Soft Matter*, 13(34):5795–5801, August 2017. Publisher: The Royal Society of Chemistry.

[191] Matthieu Wyart, Leonardo E. Silbert, Sidney R. Nagel, and Thomas A. Witten. Effects of compression on the vibrational modes of marginally jammed solids. *Physical Review E*, 72(5):051306, November 2005. Publisher: American Physical Society.

[192] Matthew B. Pinson and Thomas A. Witten. Signal transmissibility in marginal granular materials. *Journal of Physics: Condensed Matter*, 28(49):495102, October 2016. Publisher: IOP Publishing.

[193] Daniel M. Sussman, Olaf Stenull, and T. C. Lubensky. Topological boundary modes in jammed matter. *Soft Matter*, 12(28):6079–6087, July 2016.

[194] C. L. Kane and T. C. Lubensky. Topological boundary modes in isostatic lattices. *Nature Physics*, 10(1):39–45, January 2014.

[195] Florence Tama, Florent Xavier Gadea, Osni Marques, and Yves-Henri Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Structure, Function, and Bioinformatics*, 41(1):1–7, 2000.

[196] Jean-Pierre Eckmann, Jacques Rougemont, and Tsvi Tlusty. Colloquium: Proteins: The physics of amorphous evolving matter. *Reviews of Modern Physics*, 91(3):031001, July 2019. Publisher: American Physical Society.

[197] Thomas L. Rodgers, Philip D. Townsend, David Burnell, Matthew L. Jones, Shane A. Richards, Tom C. B. McLeish, Ehmke Pohl, Mark R. Wilson, and Martin J. Cann. Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. *PLoS Biology*, 11(9):e1001651, September 2013.

[198] Nicholas J. Higham. Computing the Polar Decomposition—with Applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, October 1986.

[199] P. M. Gullett, M. F. Horstemeyer, M. I. Baskes, and H. Fang. A deformation gradient tensor and strain tensors for atomistic simulations. *Modelling and Simulation in Materials Science and Engineering*, 16(1):015001, December 2007.

[200] R. J. Bell and P. Dean. Atomic vibrations in vitreous silica. *Discussions of the Faraday Society*, 50(0):55–61, January 1970.

[201] Arjun S. Raman, K. Ian White, and Rama Ranganathan. Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell*, 166(2):468–480, July 2016.

[202] Morten Kjeldgaard, Poul Nissen, Søren Thirup, and Jens Nyborg. The crystal structure of elongation factor EF-Tu from Thermus aquaticus in the GTP conformation. *Structure*, 1(1):35–50, September 1993.

[203] Joanna F Swain and Lila M Gierasch. The changing landscape of protein allostery. *Current Opinion in Structural Biology*, 16(1):102–108, February 2006.

[204] Samuel Hertig, Naomi R. Latorraca, and Ron O. Dror. Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations. *PLOS Computational Biology*, 12(6):e1004746, June 2016. Publisher: Public Library of Science.

[205] Qingling Tang and Aron W. Fenton. Whole-protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism. *Human Mutation*, 38(9):1132–1143, 2017.

[206] Thomas D. Pollard. A Guide to Simple and Informative Binding Assays. *Molecular Biology of the Cell*, 21(23):4061–4067, December 2010. Publisher: American Society for Cell Biology (mboc).

[207] Pekka Poutiainen, Kun-Eek Kil, Zhaoda Zhang, Darshini Kuruppu, Bakhos Tannous, and Anna-Liisa Brownell. Co-operative binding assay for the characterization of mGlu4 allosteric modulators. *Neuropharmacology*, 97:142–148, October 2015.

[208] Shimon Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Structural Biology*, 7(9):724–729, September 2000. Number: 9 Publisher: Nature Publishing Group.

[209] Hubert M. Piwonski, Mila Goomanovsky, David Bensimon, Amnon Horovitz, and Gilad Haran. Allosteric inhibition of individual enzyme molecules trapped in lipid vesicles. *Proceedings of the National Academy of Sciences*, 109(22):E1437–E1443, May 2012. Publisher: National Academy of Sciences Section: PNAS Plus.

[210] Brigitte Buchli, Steven A. Waldauer, Reto Walser, Mateusz L. Donten, Rolf Pfister, Nicolas Blöchliger, Sandra Steiner, Amedeo Caflisch, Oliver Zerbe, and Peter Hamm. Kinetic response of a photoperturbed allosteric protein. *Proceedings of the National Academy of Sciences*, 110(29):11725–11730, July 2013. Publisher: National Academy of Sciences Section: Physical Sciences.

[211] Guifeng Li, Donny Magana, and R. Brian Dyer. Anisotropic energy flow and allosteric ligand binding in albumin. *Nature Communications*, 5(1):1–7, January 2014. Number: 1 Publisher: Nature Publishing Group.

[212] Debora S. Marks, Thomas A. Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, November 2012. Number: 11 Publisher: Nature Publishing Group.

[213] Alexander Schug, Martin Weigt, José N. Onuchic, Terence Hwa, and Hendrik Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, December 2009. Publisher: National Academy of Sciences Section: Physical Sciences.

[214] Thomas A Hopf, Charlotta P I Schärfe, João P G L M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M J J Bonvin, and Debora S Marks. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3:e03430, September 2014. Publisher: eLife Sciences Publications, Ltd.

[215] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, 3:e02030, May 2014. Publisher: eLife Sciences Publications, Ltd.

[216] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F. Delaney, Daniel Segrè, and Christopher J. Marx. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science*, 332(6034):1190–1192, June 2011. Publisher: American Association for the Advancement of Science Section: Report.

[217] Sijmen Schoustra, Sungmin Hwang, Joachim Krug, and J. Arjan G. M. de Visser. Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. *Proceedings of the Royal Society B: Biological Sciences*, 283(1837):20161376, August 2016. Publisher: Royal Society.

[218] Michael M. Desai, Daniel Weissman, and Marcus W. Feldman. Evolution Can Favor Antagonistic Epistasis. *Genetics*, 177(2):1001–1010, October 2007. Publisher: Genetics Section: Investigations.

[219] J. Lalić and S. F. Elena. Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. *Heredity*, 109(2):71–77, August 2012. Number: 2 Publisher: Nature Publishing Group.

[220] Chongli Qin and Lucy J. Colwell. Power law tails in phylogenetic systems. *Proceedings of the National Academy of Sciences*, 115(4):690–695, January 2018. Publisher: National Academy of Sciences Section: Physical Sciences.

[221] Simona Cocco, Remi Monasson, and Martin Weigt. From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes are Needed for Structure Prediction. *PLOS Computational Biology*, 9(8):e1003176, August 2013. Publisher: Public Library of Science.

[222] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, October 2018. Number: 10 Publisher: Nature Publishing Group.

[223] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning protein constitutive motifs from sequence data. *eLife*, 8:e39397, March 2019. Publisher: eLife Sciences Publications, Ltd.

[224] Jan Humplik and Gašper Tkačik. Probabilistic models for neural populations that naturally capture global coupling and criticality. *PLOS Computational Biology*, 13(9):e1005763, September 2017.

[225] Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, August 2018.

[226] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, December 2019. Number: 12 Publisher: Nature Publishing Group.

[227] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, December 2018. arXiv: 1802.03426.

[228] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[229] Jérôme Tubiana. *Restricted Boltzmann machines : from compositional representations to protein sequence analysis.* thesis, Paris Sciences et Lettres, November 2018.

[230] Kabir Husain and Arvind Murugan. Physical constraints on epistasis. *arXiv:1910.09491 [physics, q-bio]*, October 2019. arXiv: 1910.09491.

[231] D. J. Jacobs and M. F. Thorpe. Generic Rigidity Percolation: The Pebble Game. *Physical Review Letters*, 75(22):4051–4054, November 1995. Publisher: American Physical Society.

[232] Matthieu Wyart. On the Rigidity of Amorphous Solids. *Annales de Physique*, 30(3):1–96, 2005. arXiv: cond-mat/0512155.

[233] A. Keith Dunker, Christopher J. Oldfield, Jingwei Meng, Pedro Romero, Jack Y. Yang, Jessica Walton Chen, Vladimir Vacic, Zoran Obradovic, and Vladimir N. Uversky. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 9(2):S1, September 2008.

[234] Ludovica Bachschmid-Romano and Manfred Opper. A statistical physics approach to learning curves for the inverse Ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(6):063406, June 2017. Publisher: IOP Publishing.

[235] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson. Large pseudocounts and L2-norm penalties are necessary for the mean-field inference of Ising and Potts models. *Physical Review E*, 90(1):012132, July 2014. Publisher: American Physical Society.

[236] Prateek Kumar, Steven Henikoff, and Pauline C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, July 2009. Number: 7 Publisher: Nature Publishing Group.

Riccardo RAVASIO

Avenue de Longemalle 7, 1020 Renens VD, Suisse
Mobile: +41(0)786258483
E-mail: riccardo.ravasio@epfl.ch
Languages: Italian (mother tongue), French (fluent), English (fluent)

Education

April 2016 – present	PhD student with Matthieu Wyart École Polytechnique Fédérale de Lausanne *Mechanics and co-evolution of allosteric materials and proteins*
October 2013 – March 2016	Master Degree in Theoretical Physics Università degli Studi Milano-Bicocca 110/110 cum laude
September 2015 – March 2016	Research Internship – Master thesis IPhT/CEA Saclay, France *Activated Dynamics of disordered systems in the mean-field approximation,* advisor Giulio Biroli
September 2014 – March 2015	Exchange programme École Polytechnique, Palaiseau, France
September 2010 – September 2013	Bachelor Degree in Physics Università degli Studi Milano-Bicocca *Point-interactions in two- and three-dimensional Quantum Mechanics*, advisor Diego Noja, 107/110

Contributed talks, workshops and summer schools

- American Physical Society March Meeting 2018, Los Angeles – *Architecture of allosteric materials and principles for optimal cooperativity*
- Deutschen Physikalischen Gesellschaft Spring Meeting 2018, Berlin – *Architecture of allosteric materials*
- American Physical Society March Meeting 2019, Boston – *Optimality of cooperativity in allosteric materials and proteins*
- Workshop – *Coevolution in proteins and RNA, theory and experiments*, CSS-Cargèse 2016
- Boulder Condensed Matter Summer School 2017 – *Frustrated and disordered systems*
- Boulder Condensed Matter Summer School 2019 – *Theoretical biophysics*

Teaching assistant

- EPFL physics bachelor, *statistical physics 2*: phase transitions
- EPFL physics master, *statistical physics 3*: complex systems (random walks, polymers, glassy systems, information theory)

Publications

- L. Yan, R. Ravasio, C. Brito and M. Wyart, PNAS **114** (10) 2526-2531 (2017)
- L. Yan*, R. Ravasio*, C. Brito and M. Wyart, Biophysical Journal **114** 2787-2798 (2018)
- B. Bravi, R. Ravasio, C. Brito, M. Wyart, PLOS CB **3** e1007630 (2020)
- R. Ravasio, S. Flatt, L. Yan, S. Zamuner, C. Brito, M. Wyart, Biophysical Journal **117** (2019)
- I. Hartarsky, M. Baity-Jesi, R. Ravasio, A. Billoire, G. Biroli, J. Stat. Mech. 093302 (2019)
- R. Ravasio, A. Billoire, G. Biroli, C. Cammarota, in preparation (2019)
- S. J. Wodak et al., Structure **27** (4) 566-578 (2019)

* equal contribution

Other

I won a scholarship from ERASMUS+ for the exchange at École Polytechnique and one from Assolombarda for my internship at IPhT. In 2017, I supervised Solange Flatt for her master thesis, which contributed to a publication. In 2019, I have been officially acknowledged by the EPFL Physics Doctoral School as “outstanding teaching assistant”. As hobby, I regularly practice mountaineering, photography, cooking and bouldering.