Alma Mater Studiorum – Università di Bologna
in cotutela con Università de Savoie Mont-Blanc

DOTTORATO DI RICERCA IN

CHIMICA

Ciclo XXXV

Settore Concorsuale: 03/A2

Settore Scientifico Disciplinare: CHIM/02

Un approccio multidisciplinare allo studio della dinamica delle proteine e della trasmissione del segnale

Presentata da: Aria Gheeraert

Coordinatore Dottorato

Prof. Luca Prodi

Supervisore

Prof. Ivan Rivalta

Supervisore

Prof. Laurent Vuillon

Esame finale anno 2022

UNIVERSITÉ SAVOIE MONT BLANC

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ SAVOIE MONT BLANC

Spécialité : **Mathématiques appliquées et applications des mathématiques**

Arrêté ministériel : 25 Mai 2016

Présentée par

Aria Gheeraert

Thèse dirigée par **Laurent Vuillon** et
codirigée par **Ivan Rivalta**

préparée au sein du **Laboratoire d'analyse et de mathématiques appliquées (LAMA - UMR CNRS 5127 & Université Savoie Mont Blanc)**
dans **l'École Doctorale Mathématiques, des Sciences et Technologies de l'Information et de l'Informatique (MSTII, ED 217)**

Une approche multidisciplinaire de l'étude de la dynamique des protéines et de la transmission de signaux

Thèse soutenue publiquement le **6 juillet 2022**,
devant le jury composé de :
M. Frédéric CAZALS
Directeur de recherche à Inria Sophia Antipolis, Rapporteur
M. Vittorio LIMONGELLI
Full professor à Universita della Svizerra Italiana, Rapporteur

UNIVERSITÉ SAVOIE MONT BLANC

M. Laurent VUILLON
Directeur de recherche à Université Savoie Mont-Blanc, Directeur de thèse
M. Ivan RIVALTA
Fonction et lieu de la fonction, rôle (Président, Rapporteur, Examinateur)
Mme Claire LESIEUR
Chargée de recherche à Laboratoire Ampère, Institut des Systèmes complexes de lyon (IXXI-ENS-Lyon), Examinatrice
M. Sergei GRUDININ
Chargé de recherche à Inria Grenoble, Examinateur
M. Victor S. BATISTA
Professor of Chemistry à Yale University, Invité

Summaries in French, English and Italian

Résumé en français

L'allostérie est un phénomène d'importance fondamentale en biologie qui permet la régulation de la fonction et l'adaptabilité dynamique des enzymes et protéines. Malgré sa découverte il y a plus d'un siècle, l'allostérie reste une énigme biophysique, parfois appelée « second secret de la vie ». La difficulté est principalement associée à la nature complexe des mécanismes allostériques qui se manifestent comme l'altération de la fonction biologique d'une protéine/enzyme (c.-à-d. la liaison d'un substrat/ligand au site active) par la liaison d'un « autre objet » ("allos stereos" en grec) a un site distant (plus d'un nanomètre) du site actif, le site effecteur. Ainsi, au cœur de l'allostérie, il y a une propagation d'un signal du site effecteur au site actif à travers une dense matrice protéique, où l'un des enjeux principaux est représenté par l'élucidation des interactions physico-chimiques entre résidus d'acides aminés qui permettent la communication entre les deux sites : les chemins allostériques. Ici, nous proposons une approche multidisciplinaire basée sur la combination de méthodes de chimie théorique, impliquant des simulations de dynamique moléculaire de mouvements de protéines, des analyses (bio)physiques des systèmes allostériques, incluant des alignements multiples de séquences de systèmes allostériques connus, et des outils mathématiques basés sur la théorie des graphes et d'apprentissage automatique qui peuvent grandement aider à la compréhension de la complexité des interactions dynamiques impliquées dans les différents systèmes allostériques. Le projet vise à développer des outils rapides et robustes pour identifier des chemins allostériques inconnus. La caractérisation et les prédictions de points allostériques peut élucider et exploiter pleinement la modulation allostérique dans les enzymes et dans les complexes ADN-protéine, avec de potentielles grandes applications dans l'ingénierie des enzymes et dans la découverte de médicaments.

Summary in english

Allostery is a phenomenon of fundamental importance in biology, allowing regulation of function and dynamic adaptability of enzymes and proteins. Despite the allosteric effect was first observed more than a century ago allostery remains a biophysical enigma, defined as the "second secret of life". The challenge is mainly associated to the rather complex nature of the allosteric mechanisms, which manifests itself as the alteration of the biological function of a protein/enzyme (e.g. ligand/substrate binding at the active site) by binding of "other object" ("allos stereos" in Greek) at a site distant (¿ 1 nanometer) from the active site, namely the effector site. Thus, at the heart of allostery there is signal propagation from the effector to the active site through a dense protein matrix, with a fundamental challenge being represented by the elucidation of the physico-chemical interactions between amino acid residues allowing communicatio n between the two binding sites, i.e. the "allosteric pathways". Here, we propose a multidisciplinary approach based on a combination of computational chemistry, involving molecular dynamics simulations of protein motions, (bio)physical analysis of allosteric systems, including multiple sequence alignments of known allosteric systems, and mathematical tools based on graph theory and machine learning that can greatly help understanding the complexity of dynamical interactions involved in the different allosteric systems. The project aims at developing robust and fast tools to identify unknown allosteric pathways. The characterization and predictions of such allosteric spots could elucidate and fully exploit the power of allosteric modulation in enzymes and DNA-protein complexes, with great potential applications in enzyme engineering and drug discovery.

Riassunto in italiano

L'allosteria è un fenomeno di fondamentale importanza in biologia che permette la regolazione della funzione e l'adattabilità dinamica di enzimi e proteine. Nonostante la sua scoperta più di un secolo fa, l'allosteria rimane un enigma biofisico, a volte chiamato "il secondo segreto della vita". La difficoltà è principalmente associata alla natura complessa dei meccanismi allosterici che si manifestano come l'alterazione della funzione biologica di una proteina/enzima (cioè il legame di un substrato/ligando al sito attivo) attraverso il legame di un "altro oggetto" (" allos stereos" in greco) ad un sito distante (più di un nanometro) dal sito attivo, il sito effettore. Pertanto, al centro dell'allosteria, c'è una propagazione di un segnale dal sito effettore al sito attivo attraverso una matrice proteica densa, dove una delle sfide principali è rappresentata dalla delucidazione delle interazioni fisico-chimiche tra i residui di amminoacidi che consentono la comunicazione tra i due siti: le vie allosteriche. In questo elaborato, viene proposto un approccio multidisciplinare basato sulla combinazione di metodi chimici teorici, che coinvolgono simulazioni di dinamica molecolare dei movimenti delle proteine, analisi (bio)fisiche di sistemi allosterici, inclusi allineamenti di sequenze multiple di sistemi allosterici noti e strumenti matematici basati sulla teoria dei grafi ed apprendimento automatico che può aiutare notevolmente a comprendere la complessità delle interazioni dinamiche coinvolte nei diversi sistemi allosterici. Il progetto mira a sviluppare strumenti veloci e robusti per identificare percorsi allosterici sconosciuti. Le caratterizzazioni e le previsioni dei punti allosterici possono chiarire e sfruttare appieno la modulazione allosterica negli enzimi e nei complessi DNA-proteina, con potenziali ampie applicazioni nell'ingegneria enzimatica e nella scoperta di farmaci.

Contents

Acknowledgments

First, I would like to offer my special thanks to Ivan Rivalta, Laurent Vuillon and Claire Lesieur who all have supervised some part of this PhD thesis. Coming from very different fields and with different expertise, they are also who allowed this thesis to be this unique and multidisciplinary. I would also like to thank Lorenza Pacini which was the PhD student of Claire that assisted me at the beginning of thesis and allowed me to understand much more easily preexisting tools. Finally, I would like to thank Andrea Piazzi and Sebastiano Cauzzi, two interns I helped Ivan to coach and that assisted me in my work. They have been a precious help and also gave me some experience in managing the work of other people.

Then, I would like to express my sincere gratitude to Carine Michel and Romain Réocreux. When I was only a L3 student in the ENS de Lyon, they introduced me to the ENS de Lyon Theoretical Chemistry Lab that paved my way to this thesis. Even if my favorite subjects were very different from what she was doing, Carine advised me for a long time, and it is notably her who advised me to contact Ivan to get this thesis.

During my PhD, I have had the chance to collaborate with many groups and I would like to extend my thanks to all these groups. Our first collaborators were the team of F. Javier Luque, Carolina Esterellas, Elnaz Aledavood and Alessia Forte at the Universitat de Barcelona. They were the first to contact us following our initial paper. In some ways, they were the first external members to show an interest in our work. For me, this was a major moral boost. This led to the publication of one paper on the AMPK and also to many improvements of my code and its availability. I also have had the chance to work with a team of collaborators based at Yale University thanks to Ivan's connections and notably with Victor S. Batista, Patrick Loria, Federica Maschietto, Apala Chaudhuri, Florentina Tofoleanu, Uriel Morzan, Brandon Allen, Gregory Kyro, Qu Zexing, Peter Nekrasov. These collaborations have resulted in the redaction of various articles, but also helped me to grow as a researcher thanks to their insightful comments and contributions. Finally, I have had the opportunity to myself establish connections during the first lockdown with a group aiming at studying the SARS-CoV-2 mechanism of action founded by Bernard Maigret. This group consists of a group in Nancy with Bernard Maigret, Marie-Dominique Devignes, Isaure Chauvot de Beauchêne, Vincent Leroux, Dominique Mias-Lucquin, a group in Montpellier with Laurent Chaloin and Olivier Moncorgé and finally Serge Perez in Grenoble. This collaboration was a long but very insightful project and many interesting works are still to come out of this.

Then I'd like to thank all the other students that have made my daily life at the laboratory more pleasing and that provided assistance when necessary. In Chambéry: Mickaël Nahon, Eloi Martinet, Tristan Humbert, Clément Lagisquet and in Bologna there's simply too many people, and I'm afraid to forget one.

Next, I would like to thank the most precious people in my life: Yann-Edwin Keta-Leroy, Camille Normand and Paul Mangold (I drew your names for the order!!!). They are not only my best friends, in some ways, they are also my soulmates. They are both life partners and also research partners, we positively influence each other in both areas. For now, in the scientific world, these collaborations have been stuck to the acknowledgments' section of some of our papers, but I truly hope that some day we will co-author together for real.

Among friends, I would like to offer my special thanks to all those that have hosted me during my academic trips. Briefly Miette, Nora, Lumi, Clair, Zach, Philippine, Marin and Claire.

Last but not least, I am deeply grateful to my parents and my little brother. They have always been there for me during this PhD thesis. I have always been able to count on them through my ups and downs. I certainly would not be where I am now if they had not always been there to push me to aim for the top.

Chapter 1

Introduction

1.1 Motivation

Proteins are some of the most compelling objects of terrestrial life. They fulfill the most diverse functions, and yet they are all constructed from the same elementary blocks: the 21 amino acids. This complexity and variety of functions contrast with their apparent simplicity. Alphabetical languages can also express countless ideas with a small set of letters, which is why it is tempting to consider proteins as the language of life. Understanding proteins is a crucial step in understanding many aspects of life, and much effort has been devoted to it. Proteins vary according to the sequence in which amino acids are ordered, and this sequence generally gives them a defined 3D structure that is related to their activity. In less than a millisecond and through a reproducible process, most proteins fold into a characterized shape in a given environment[1]. However, proteins are dynamical objects and adapt their shape to external conditions[2]. Even understanding how proteins fold has challenged more than two generations of scientists and has been referred to as the secret of life[3]. Only in very recent times precise theoretical predictions of protein folding are possible on the basis of a deep learning algorithm that does not explain the underlying process[4].

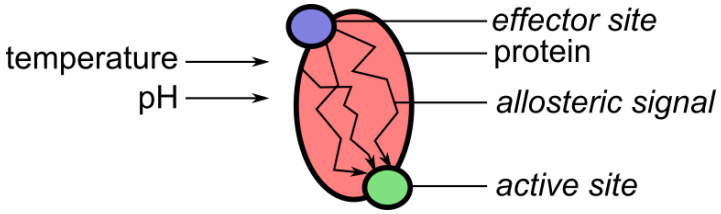


Figure 1.1: Principle of allosteric regulation

For life, it is crucial that proteins do not perform their activity aimlessly. Therefore, evolution has designed, with billions of years of evolution, regulatory processes that enable proteins to respond to their environment and modulate their activity according to external conditions. Allosteric regulation, or allostery, is an example of a regulatory process that is of fundamental importance in biology. It has even been hypothesized that allostery is an inherent property of all dynamic proteins[5]. In allostery, the binding of an effector (generally a small molecule, but it can be another protein) transmits a signal through the protein matrix, which disturbs another separate functional site, and then modulates its activity. The word allostery is a neologism formed from the ancient Greek allos stereos, meaning "other solid (object)" observing the distance between the sites in opposition to orthosteric : orthós (straight, correct). In practice, this process gives proteins interesting sensor behaviors, such as being active only in the presence of a molecule (positive allostery) or being able to reduce their activity in the presence of a molecule (negative allostery). The name allostery and its derivatives are also sometimes generalized to describe various environmental changes, such as changes in pH and temperature, or light activation.

Although the first allostery model was proposed more than 55 years ago[6, 7], the phenomenon remains poorly understood and has been nicknamed the "second secret of life"[8, 9]. A major obstacle in understanding allostery is that numerous studies have focused on a structural and static vision of the studied systems. Recent studies have shown that in some cases, allosteric mechanisms involve the less probable conformations of a protein, so restricting to the study of the most probable one is limited[10]. Dynamical motions of various scales have been shown to be involved in allostery such as rigid-body (i.e., ternary or quaternary structures) motion[11, 12, 13], conformational dynamics of folded structures[14, 15, 16, 17, 18, 19, 19, 20], local (un)folding[21, 22] and intrinsic disorder[23, 24, 25, 26]. Another aspect that makes the problem difficult to study is that in most proteins, the allosteric signal is carried by multiple redundant allosteric pathways[27] and residues that are critical for allosteric signaling are poorly conserved[28, 29]. Thus, many intuitive ideas to find allosteric path-

ways by experiment are simply wrong because blocking one pathway does not necessarily block all the pathways.

Designing experiments to study the allosteric pathways is hard because of the timescale and the precision required. Comparing X-ray crystal structures of the *reference* and *perturbed* proteins is tempting because it provides a detailed molecular information of the difference between two states, but allostery is an intrinsically dynamical process and this method only compares two snapshots of the systems, which is very limited. Mutagenesis experiments coupled with protein activity measurement (in the case of an enzyme, steady-state kinetics) can point to specific residues essential to the allosteric mechanism. However, these experiments usually cannot explain on their own why a specific residue is important to the mechanism, and predicting which amino acids are good candidates for mutagenesis is not obvious. NMR methods are another popular tool[30] as their time scale allows the capture of several states of each system, which makes for a more substantial comparison than single crystal structures. However, most NMR methods are blind to some amino acids or their signal are extremely congested and cannot provide molecular insights into the underlying phenomenon. Other experimental tools exist and are beyond the scope of this thesis, but this shows that allostery is a field where computational tools and particularly Molecular Dynamics (MD) simulations can provide a precious complementary vision of the phenomenon with information at the molecular level at the microsecond timescale.

The exponential improvement of technologies currently gives the opportunity to perform classical molecular dynamics (MD) simulations of systems with a relatively large size and timescale[31, 32, 33, 34, 35]. Analyzing such long and sizable simulations thus becomes increasingly challenging, and there is a need to develop tools to facilitate their analysis. Dynamical networks have emerged about 20 years ago to investigate allosteric signaling[36, 37, 38, 39, 40, 41] and very recently machine learning analysis of MD simulations to infer allosteric pathways has become an increasingly popular tool[42, 43, 44, 45]. In both dynamical networks and machine learning methods, one of the most important tasks is to select relevant features (sometimes called descriptors) for analysis.

Among the features studied, contacts have been described as a natural language for allostery[46]. Indeed, at their heart, the rearrangements that occur in a protein after perturbation are driven by the formation and disruption of contacts that involve effects similar to those of protein folding[47, 48]. Thus, contacts are some natural dimension in which we expect a protein to evolve. Contact conditions are sometimes used to infer allosteric pathways, but only as a filter to select relevant residue pairs. This happens principally in correlation analysis and, if a contact is considered present for a sufficient portion of the Molecular Dynamics simulation, the correlation is considered[36, 37, 38].

In static analysis of structures, contact networks, sometimes called Amino Acid Networks (AANs) have various different definitions. Some focus only on the C_α distance to infer a binary map of contacts[48, 49, 50, 51, 52, 53, 54], on the C_β[55] (with a cutoff between 7 Å and 8.5Å;) or residue centroid[56], while others infer the binary map from any heavy-atom contact[57, 51] (with a cut-off below 5 Å) or any sidechain contact[58]. Mapping with weights gives a richer description of the contact than unweighted networks and despite many approaches to map with energetic quantities[59, 60], it is possible to merely give a weight of the contact by the number of amino acid that satisfy the contact condition[61]. This type of weighted contact network analysis has been successfully used to infer protein dynamics and determine structural robustness to mutations in proteins, being powerful in understanding how a local change can produce global changes that are associated with retention or loss of protein functions[62, 63, 64]. A recent study has shown the potential of dynamical network of inter-residues contacts and was used to reveal the allosteric effects of mutations in the catalytic activity of the Cyclophilin A enzyme, proving to be potentially able to identify key residues in allosteric signal propagation[65].

Therefore, contact analysis to infer allosteric pathways is a promising emerging field. During this thesis, another group has also focused their effort into this, developing their own contact methodology to retrieve allosteric pathways[66, 67, 68]. Current methodologies however do not offer a complete description of the system dynamics, as contacts are usually time-averaged or reduced to their frequency. Dealing with contacts as a dynamical feature presents many challenges. First, on a technical level, computing a detailed view of the contacts is resource intensive, and storing all possible contacts can take a large amount of memory. Moreover, the total number of contacts is intrinsically fluctuating, and dealing with a variable number of features possesses its own challenges. This is why many contact-based approaches overly simplify the problem, and there is clearly a gap in methodology to adequately study the dynamical evolution of contacts during MD simulations.

In first, we developed a methodology to investigate the dynamical evolution of contacts, primarily to study allostery put in the scope of comparing sets of MD simulations. This approach can be generalized to compare all sorts of MD simulation with a *reference* simulation and a *perturbed* simulation, notably to compare MD simulations of mutants. Moreover, the latest developments of the methodology which manages takes into account the dynamics within a single simulation opens the door to studying contact changes happening in a single simulation. In particular, one application proved its strength to assess if an MD simulation has

converged (i.e., is near the equilibrium) and can show the degrees of relaxation from the non-equilibrated input structure to the equilibrated structure. In general, many more applications of this methodology are possible, despite being outside the scope of this thesis. Among potential applications of this methodology, are general conformational change analysis in other types of MD simulation (biased, targeted), analysis of protein-protein binding/unbinding events and protein folding analysis.

1.2 Imidazole Glycerol Phosphate Synthase

Our main proteic target is Imidazole Glycerol Phosphate synthase (IGPS) from the bacteria, *T. maritima* which is an archetypical allosteric enzyme. The history of the discovery of the IGPS is a very interesting example of how a scientific model has evolved over time. During the 1950s to 1960s, the majority of steps in histidine biosynthesis were gradually elucidated, and most metabolic intermediates and enzymes catalyzing the corresponding reactions were discovered in *S. typhimurium*[69, 70, 71, 72, 73, 74]. Originally, the steps encoded by the hisF and hisH genes were thought to be carried out consecutively by two separate enzymes, but the precise order of these reactions was still unknown[72]. Furthermore, this model did not provide an explanation for the link between histidine synthesis and *de novo* purine synthesis[75, 76]. In 1992, it was actually discovered in *E. coli* that the hisF and hisH genes encoded two proteins bound in a stable 1:1 dimeric complex: IGPS. This last piece of the histidine biosynthesis puzzle[77] revealed that IGPS is present at the fifth step of histidine biosynthesis. In fact, it was later found that in plants and fungi, the hisH and hisF genes actually fuse into a single gene called his7, leading to the formation of IGPS with a single chain[78].

IGPS is a GlutamineAmidoTransferase (GATase), that is, an enzyme that catalyzes the hydrolysis of glu-

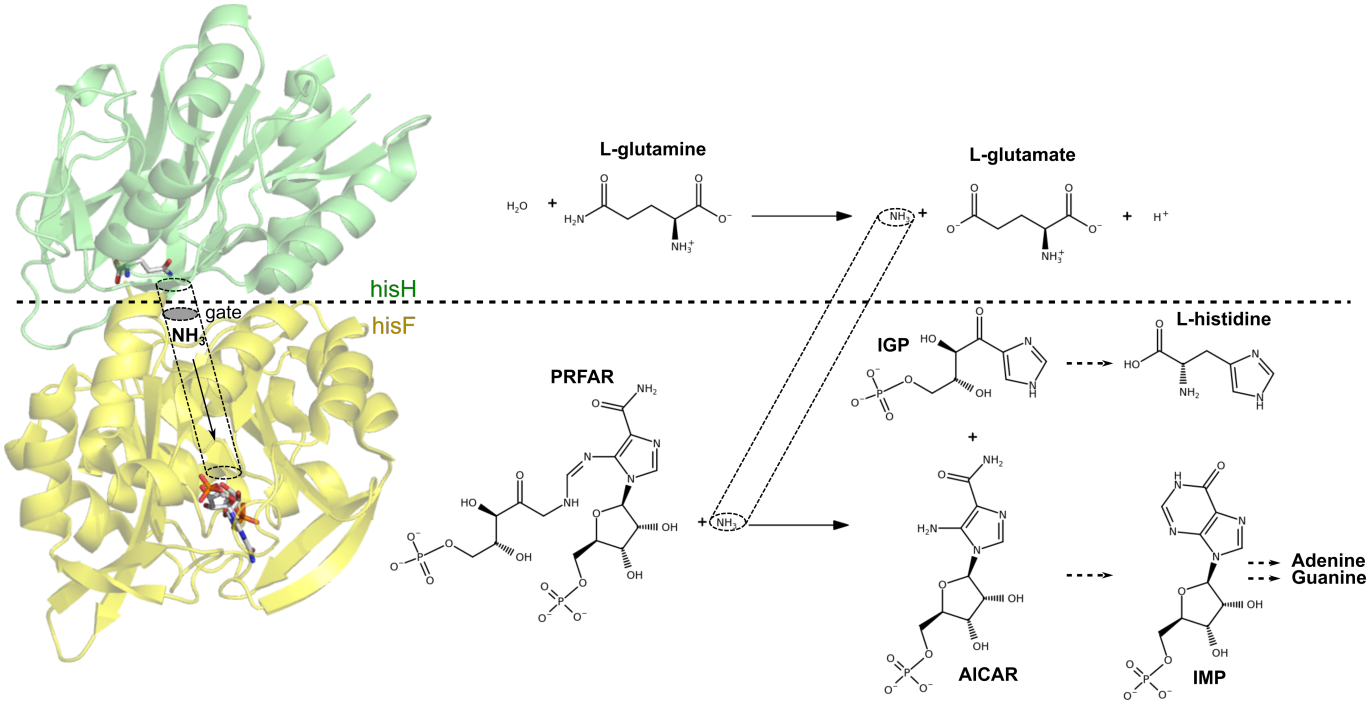


Figure 1.2: *T. maritima* IGPS 3D structure with glutamine bound in HisH and PRFAR bound in HisF (ternary complex) with reactions occurring at each active site. In this case, the effector site for reaction occurring in HisH is also the active site of HisF. The ammonia tunnel is represented by a dotted cylinder.

tamine to produce ammonia for another reaction. In *T. maritima* it is composed of two subunits: HisH which catalyzes glutamine hydrolysis into glutamate and ammonia and a cyclase HisF where the effector PRFAR, (N'-[(5'-phosphoribulosyl)formimino]-5-amino-imidazole-4-carboxamide-ribonucleotide) binds (see Figure 1.2. The nascent ammonia passes through a tunnel shielding it from water protonation and then approaches PRFAR near the effector site. Then, a cyclization occurs that releases ImidazoleGlycerol Phosphate (IGP, which gives its name to the enzyme), a precursor to histidine and 5'-(5-aminoimidazole-4-carboxamide) (AlCAR), later converted to Inosine MonoPhosphate (IMP), a precursor to adenine and guanine, the two purine nucleotides, finally explaining the link between histidine biosynthesis and the *de novo* purine synthesis pathways.

In their original work, Klem and Davisson note that in *E. coli*, the glutaminase efficiency by 39-fold[79] in presence of IGP. A similar effect is observed with a PRFAR precursor: N-[(5'-phosphoribosyl)formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (5'-ProFAR). This suggested that PRFAR and analogs are allosteric effectors of the glutaminase. The same effect is observed in IGPS from *S. cerevisiae* [80] and in this organism, and a comparison between many potential effectors showed that PRFAR is, in fact, the best effector

and increases the catalytic efficiency of glutaminase 4900 times[81]. This shows that IGPS is an allosteric enzyme that responds to PRFAR binding to trigger the ammonia transfer event and prevent the waste of glutamine. Furthermore, the increase in catalytic efficiency is primarily driven by an increase in catalytic activity and not an increase in substrate affinity, making IGPS a so-called V-type allosteric enzyme.

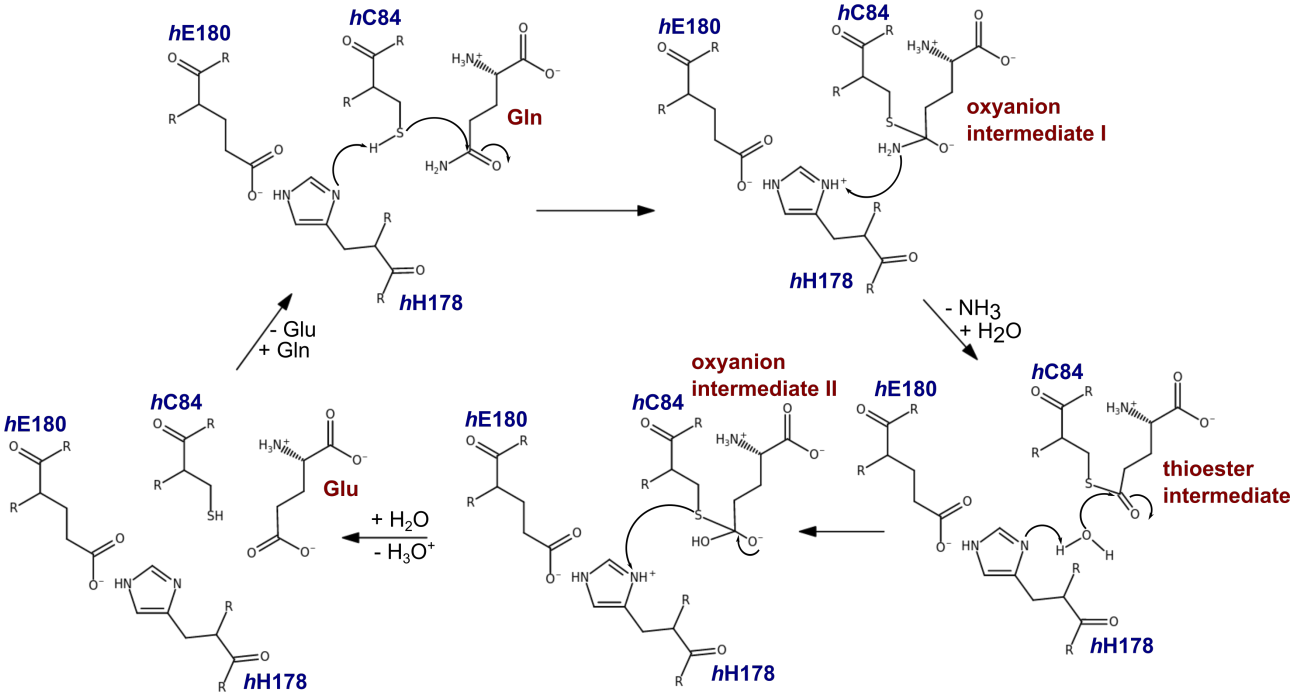


Figure 1.3: Glutaminase mechanism in IGPS from *T. maritima*. The mechanistic details of the reactions leading to the formation of the oxyanion intermediates are tentative, and the mechanism could be either concerted or sequential.

IGPS is a Class I GATase, which means that the glutaminase active site is composed of a conserved catalytic triad consisting of a cysteine (nucleophile), a histidine (base), and a glutamate (acid). In *T. maritima* these residues are *h*C84, *h*H178 and *h*E180 (the *h* or *f* prefix means that the residue is found in hisF or hisH, respectively). In Figure 1.3 the mechanism of the active site of glutaminase is represented. Although thoroughly studied, some details of the mechanism remain unclear, particularly in the two reactions leading to the formation of an oxyanion intermediate it is not clear if the protonation step of *h*H178 by the nucleophile is sequential or consecutive to the nucleophile attack. The two tetrahedral oxyanion intermediates are particularly unstable and are rate-limiting for the catalysis. Other Class-I GATase possess an oxyanion hole consisting of two backbone amide protons, which stabilizes the oxyanion intermediates[82, 83, 84, 85, 86, 87]. Crystal structures of *T. maritima* [88] and *S. cerevisiae* [89] show that the oxyanion hole is not completely formed in the apoenzyme (apo), as one amide proton is in place (*h*L85), but the other (*h*V51) is flipped away from the active site and shrouds the *h*L85 amide proton with its backbone carbonyl oxygen. It has been postulated that the allosteric mechanism in IGPS involves the formation of the oxyanion hole upon effector binding, explaining the strong different in catalytic efficiency between apo-IGPS and PRFAR-bound IGPS[88] and studies of PRFAR-bound *S. cerevisiae* IGPS have evidenced conformational variability in the oxyanion strand (49-PGVG segment)[90]. Despite numerous of experimental and theoretical[37] evidence, only very recently an experimental structural evidence definitely proved that in *T. maritima* the presence of PRFAR and glutamine results in the formation of an oxyanion hole[91].

In a series of computational studies of IGPS allostery that this thesis extends, allosteric propagation pathways were predicted to involve motions of various dynamical nature and are summarized in Figure 1.4 [37, 92, 93]. In first, a rigid-body motion between the HisF/HisH interface, named the *breathing motion*, is both faster and of lower amplitude upon effector binding. Higher in the dynamics spectrum, there are a series of sidechain rearrangements, involving in one instance the formation of a hydrophobic cluster and in another the alteration of a network of salt bridges. The mechanism also involves some backbone dynamics since a few key backbone hydrogen-bonding disruptions near the active site have been discovered. Finally, a refolding event occurs near the effector site with the folding of loop1 on the effector binding site and near the active site the 49-PGVG oxyanion segment is also shown to increase in flexibility. This latter part is the penultimate step of the allosteric mechanism which lead to the more probable formation of the oxyanion hole in holo-IGPS compared to apo-IGPS.

These theoretical predictions have been subsequently used to design a non-competitive allosteric drug[92] and IGPS mutants[94] that alter the IGPS allosteric pathways, resulting in inactive enzymes and proving the potential of elucidating allosteric pathways for drug design. Very recently, both short- and long-range predicted

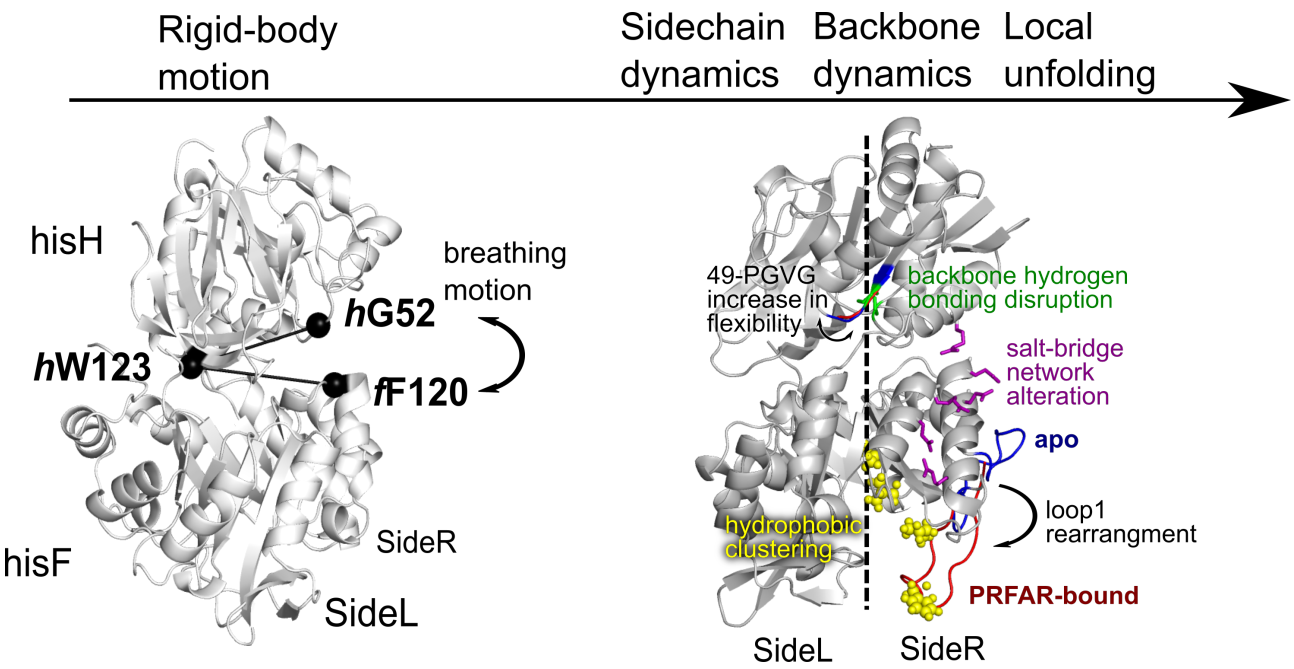


Figure 1.4: IGPS presents allosteric dynamics from most of the dynamic continuum of allostery.

effects have been demonstrated experimentally by X-ray structural characterization of active IGPS ternary complexes[91] and light-switching activation[95], respectively. IGPS is present in fungi, plants, bacteria, and archaea, but not in mammals effectively, making IGPS a target for the development of safe antipathogens[89, 96, 97]. Furthermore, it is a prototype allosteric system because its allosteric mechanism in *T. maritima* involves dynamics from all the dynamic continuum of allostery.

To evaluate our new tools, we use them for analysis of MD simulations of IGPS from *T. maritima* that have been previously studied and contain information on the allosteric mechanism. This set of MD simulations contains 4 simulations of 100 ns (1,000 frames each) for apo-IGPS (apo, simulations apo1-4) and 4 simulations of 100 ns for holo-IGPS (prfar, simulations prfar1-4).

References

[1] Jan Kubelka, James Hofrichter, and William A Eaton. "The protein folding 'speed limit'". In: *Current opinion structural biology* 14.1 (2004), pp. 76–88.

[2] Zimei Bu and David JE Callaway. "Proteins move! Protein dynamics and long-range allostery in cell signaling". In: *Adv. protein chemistry structural biology* 83 (2011), pp. 163–221.

[3] Ken A Dill and Justin L MacCallum. "The protein-folding problem, 50 years on". In: *science* 338.6110 (2012), pp. 1042–1046.

[4] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[5] K Gunasekaran, Buyong Ma, and Ruth Nussinov. "Is allostery an intrinsic property of all dynamic proteins?" In: *Proteins: Struct. Funct. Bioinform.* 57.3 (2004), pp. 433–443.

[6] Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux. "On the nature of allosteric transitions: a plausible model". In: *J Mol Biol* 12.1 (1965), pp. 88–118.

[7] Jean-Pierre Changeux. "Allostery and the Monod-Wyman-Changeux model after 50 years". In: *Annu. review biophysics* 41 (2012), pp. 103–133.

[8] Jacques Monod. *Chance and necessity: an essay on the natural philosophy of modern biology*. New York: Vintage, 1971.

[9] Aron W Fenton. "Allostery: an illustrated definition for the 'second secret of life'". In: *Trends biochemical sciences* 33.9 (2008), pp. 420–425.

[10] Fan Bai et al. "Conformational spread as a mechanism for cooperativity in the bacterial flagellar switch". In: *science* 327.5966 (2010), pp. 685–689.

[11] Joanna F Swain et al. "Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker". In: *Mol. cell* 26.1 (2007), pp. 27–39.

[12] Michael D Daily and Jeffrey J Gray. "Allosteric communication occurs via networks of tertiary and quaternary motions in proteins". In: *PLoS computational biology* 5.2 (2009), e1000293.

[13] Erik RP Zuiderweg et al. "Allostery in the Hsp70 chaperone proteins". In: *Mol. Chaperones* (2012), pp. 99–153.

[14] Hong Pan, J Ching Lee, and Vincent J Hilser. "Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble". In: *Proc. National Acad. Sci.* 97.22 (2000), pp. 12020–12025.

[15] Osamu Miyashita, Peter G Wolynes, and Jose N Onuchic. "Simple energy landscape model for the kinetics of functional transitions in proteins". In: *The journal physical chemistry B* 109.5 (2005), pp. 1959–1969.

[16] Nataliya Popovych et al. "Dynamically driven protein allostery". In: *Nat. structural & molecular biology* 13.9 (2006), pp. 831–838.

[17] Ernesto J Fuentes et al. "Evaluation of energetic and dynamic coupling networks in a PDZ domain protein". In: *J. molecular biology* 364.3 (2006), pp. 337–351.

[18] Robert G Smock and Lila M Gierasch. "Sending signals dynamically". In: *science* 324.5924 (2009), pp. 198–203.

[19] Chad M Petit et al. "Hidden dynamic allostery in a PDZ domain". In: *Proc. National Acad. Sci.* 106.43 (2009), pp. 18249–18254.

[20] Shiou-Ru Tzeng and Charalampos G Kalodimos. "Dynamic activation of an allosteric regulatory protein". In: *Nature* 462.7271 (2009), pp. 368–372.

[21] Travis P Schrank, D Wayne Bolen, and Vincent J Hilser. "Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins". In: *Proc. National Acad. Sci.* 106.40 (2009), pp. 16984–16989.

[22] Sean E Reichheld, Zhou Yu, and Alan R Davidson. "The induction of folding cooperativity by ligand binding drives the allosteric response of tetracycline repressor". In: *Proc. National Acad. Sci.* 106.52 (2009), pp. 22263–22268.

[23] Vincent J Hilser and E Brad Thompson. "Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins". In: *Proc. National Acad. Sci.* 104.20 (2007), pp. 8311–8315.

[24] Abel Garcia-Pino et al. "Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity". In: *Cell* 142.1 (2010), pp. 101–111.

[25] Eva Sevcsik et al. "Allostery in a disordered protein: oxidative modifications to α-synuclein act distally to regulate membrane binding". In: *J. Am. Chem. Soc.* 133.18 (2011), pp. 7152–7158.

[26] Allan Chris M Ferreon et al. "Modulation of allostery by protein intrinsic disorder". In: *Nature* 498.7454 (2013), pp. 390–394.

[27] Antonio Del Sol et al. "The origin of allosteric functional modulation: multiple pre-existing pathways". In: *Structure* 17.8 (2009), pp. 1042–1050.

[28] Dennis R Livesay, Kyle E Kreth, and Anthony A Fodor. "A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms". In: *Allostery* (2012), pp. 385–398.

[29] Megan Leander et al. "Functional plasticity and evolutionary adaptation of allosteric regulation". In: *Proc. National Acad. Sci.* 117.41 (2020), pp. 25445–25454.

[30] Sarina Grutsch, Sven Brüschweiler, and Martin Tollinger. "NMR methods to study dynamic allostery". In: *PLoS computational biology* 12.3 (2016), e1004620.

[31] Kresten Lindorff-Larsen et al. "How fast-folding proteins fold". In: *Science* 334.6055 (2011), pp. 517–520.

[32] David E Shaw et al. "Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer". In: *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. 2014, pp. 41–53.

[33] Andrea Saltalamacchia et al. "Decrypting the information exchange pathways across the spliceosome machinery". In: *J. Am. Chem. Soc.* 142.18 (2020), pp. 8403–8411.

[34] Jaewoon Jung et al. "Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations". In: *J. computational chemistry* 40.21 (2019), pp. 1919–1930.

[35] Jaewoon Jung et al. "New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems". In: *J. Comput. Chem.* 42.4 (2021), pp. 231–241.

[36] Anurag Sethi et al. "Dynamical networks in tRNA: protein complexes". In: *Proc. National Acad. Sci.* 106.16 (2009), pp. 6620–6625.

[37] Ivan Rivalta et al. "Allosteric pathways in imidazole glycerol phosphate synthase". In: *Proc. National Acad. Sci.* 109.22 (2012), E1428–E1436.

[38] Adam T VanWart et al. "Exploring residue component contributions to dynamical network models of allostery". In: *J. chemical theory computation* 8.8 (2012), pp. 2949–2961.

[39] Kyle W East et al. "NMR and computational methods for molecular resolution of allosteric pathways in enzyme complexes". In: *Biophys. reviews* 12.1 (2020), pp. 155–174.

[40] Onur Serçinoğlu and Pemra Ozbek. "gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations". In: *Nucleic Acids Research* 46.W1 (July 2018), W554–W562.

[41] Gennady M Verkhivker and Luisa Di Paola. "Dynamic Network Modeling of Allosteric Interactions and Communication Pathways in the SARS-CoV-2 Spike Trimer Mutants: Differential Modulation of Conformational Landscapes and Signal Transmission via Cascades of Regulatory Switches". In: *The J. Phys. Chem. B* 125.3 (2021), pp. 850–873.

[42] Mohsen Botlani, Ahnaf Siddiqui, and Sameer Varma. "Machine learning approaches to evaluate correlation patterns in allosteric signaling: A case study of the PDZ2 domain". In: *The J. Chem. Phys.* 148.24 (2018), p. 241726.

[43] Hongyu Zhou, Zheng Dong, and Peng Tao. "Recognition of protein allosteric states and residues: Machine learning approaches". In: *J. computational chemistry* 39.20 (2018), pp. 1481–1490.

[44] Hamed S Hayatshahi et al. "Probing protein allostery as a residue-specific concept via residue response maps". In: *J. Chem. Inf. Model.* 59.11 (2019), pp. 4691–4705.

[45] Filippo Marchetti et al. "Machine learning prediction of allosteric drug activity from molecular dynamics". In: *The journal physical chemistry letters* 12.15 (2021), pp. 3724–3732.

[46] Luisa Di Paola and Alessandro Giuliani. "Protein contact network topology: a natural language for allostery". In: *Current opinion structural biology* 31 (2015), pp. 43–48.

[47] Michele Vendruscolo et al. "Small-world view of the amino acids that play a key role in protein folding". In: *Phys. Review E* 65.6 (2002), p. 061910.

[48] Nikolay V Dokholyan et al. "Topological determinants of protein folding". In: *Proc. National Acad. Sci.* 99.13 (2002), pp. 8637–8641.

[49] Ganesh Bagler and Somdatta Sinha. "Network properties of protein structures". In: *Physica A: Stat. Mech. its Appl.* 346.1-2 (2005), pp. 27–33.

[50] Ganesh Bagler and Somdatta Sinha. "Assortative mixing in protein contact networks and protein folding kinetics". In: *Bioinformatics* 23.14 (2007), pp. 1760–1767.

[51] L Bartoli, P Fariselli, and R Casadio. "The effect of backbone on the small-world properties of protein contact maps". In: *Phys. biology* 4.4 (2008), p. L1.

[52] Omar Gaci and Stefan Balev. "Node degree distribution in amino acid interaction networks". In: *2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop*. IEEE. 2009, pp. 107–112.

[53] Saraswathi Vishveshwara, Amit Ghosh, and Priti Hansia. "Intra and inter-molecular communications through protein structure network". In: *Current Protein Pept. Sci.* 10.2 (2009), pp. 146–160.

[54] Susan Khor. "Towards an integrated understanding of the structural characteristics of protein residue networks". In: *Theory Biosci.* 131.2 (2012), pp. 61–75.

[55] Ali Rana Atilgan, Pelin Akan, and Canan Baysal. "Small-world communication of residues and significance for protein dynamics". In: *Biophys. journal* 86.1 (2004), pp. 85–91.

[56] Nelson Augusto Alves and Alexandre Souto Martinez. "Inferring topological features of proteins from amino acid residue networks". In: *Physica A: Stat. Mech. Its Appl.* 375.1 (2007), pp. 336–344.

[57] Lesley H Greene and Victoria A Higman. "Uncovering network systems within protein structures". In: *J. molecular biology* 334.4 (2003), pp. 781–791.

[58] Md Aftabuddin and S Kundu. "Hydrophobic, hydrophilic, and charged amino acid networks within protein". In: *Biophys. journal* 93.1 (2007), pp. 225–231.

[59] MS Vijayabaskar and Saraswathi Vishveshwara. "Interaction energy based protein structure networks". In: *Biophys. journal* 99.11 (2010), pp. 3704–3715.

[60] Xiong Jiao et al. "Construction and application of the weighted amino acid network based on energy". In: *Phys. Review E* 75.5 (2007), p. 051903.

[61] Md Aftabuddin and Sudip Kundu. "Weighted and unweighted network of amino acids within protein". In: *Physica A: Stat. Mech. its Appl.* 369.2 (2006), pp. 895–904.

[62] Laurent Vuillon and Claire Lesieur. "From local to global changes in proteins: a network view". In: *Current opinion structural biology* 31 (2015), pp. 1–8.

[63] Mounia Achoch et al. "Protein structural robustness to mutations: an in silico investigation". In: *Phys. Chem. Chem. Phys.* 18.20 (2016), pp. 13770–13780.

[64] Rodrigo Dorantes-Gilardi et al. "In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: not too many links, not too few". In: *Phys. Chem. Chem. Phys.* 20.39 (2018), pp. 25399–25410.

[65] Urmi Doshi et al. "Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation". In: *Proc. National Acad. Sci.* 113.17 (2016), pp. 4735–4740.

[66] Xin-Qiu Yao, Mohamed Momin, and Donald Hamelberg. "Elucidating allosteric communications in proteins with difference contact network analysis". In: *J. chemical information modeling* 58.7 (2018), pp. 1325–1330.

[67] Xin-Qiu Yao, Mohamed Momin, and Donald Hamelberg. "Establishing a Framework of Using Residue–Residue Interactions in Protein Difference Network Analysis". In: *J. chemical information modeling* 59.7 (2019), pp. 3222–3228.

[68] Xin-Qiu Yao and Donald Hamelberg. "Residue–Residue Contact Changes during Functional Processes Define Allosteric Communication Pathways". In: *J. Chem. Theory Comput.* (2022).

[69] Felix Haas et al. "A series of histidineless mutants of Neurospora crassa". In: *Genetics* 37.3 (1952), p. 217.

[70] Bruce N Ames, Robert G Martin, and Barbara J Garry. "The first step of histidine biosynthesis". In: *J. Biol. Chem.* 236.7 (1961), pp. 2019–2026.

[71] Bruce N Ames and PE Hartman. "Genes, enzymes, and control mechanisms in histidine biosynthesis". In: *The molecular basis neoplasia* (1962), pp. 322–345.

[72] David WE Smith and Bruce N Ames. "Intermediates in the early steps of histidine biosynthesis". In: *J. Biol. Chem.* 239.6 (1964), pp. 1848–1855.

[73] Michael Brenner and Bruce N Ames. "The histidine operon and its regulation". In: *Metabolic regulation.* Elsevier, 1971, pp. 349–387.

[74] Robert G Martin et al. "[147] Enzymes and intermediates of histidine biosynthesis in Salmonella typhimurium". In: *Methods in enzymology.* Vol. 17. Elsevier, 1971, pp. 3–44.

[75] Alexandra E Shedlovsky and Boris Magasanik. "A defect in histidine biosynthesis causing an adenine deficiency". In: *J. Biol. Chem.* 237.12 (1962), pp. 3725–3730.

[76] H Mark Johnston and John R Roth. "Histidine mutants requiring adenine: selection of mutants with reduced hisG expression in Salmonella typhimurium". In: *Genetics* 92.1 (1979), pp. 1–15.

[77] Pietro Alifano et al. "Histidine biosynthetic pathway and genes: structure, regulation, and evolution". In: *Microbiol. reviews* 60.1 (1996), pp. 44–69.

[78] M Kuenzler et al. "Cloning, primary structure, and regulation of the HIS7 gene encoding a bifunctional glutamine amidotransferase: cyclase from Saccharomyces cerevisiae". In: *J. bacteriology* 175.17 (1993), pp. 5548–5558.

[79] Thomas J Klem and V Jo Davisson. "Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis". In: *Biochemistry* 32.19 (1993), pp. 5177–5186.

[80] Sridar V Chittur, Yuan Chen, and V Jo Davisson. "Expression and purification of imidazole glycerol phosphate synthase from Saccharomyces cerevisiae". In: *Protein Expr. Purif.* 18.3 (2000), pp. 366–377.

[81] Rebecca S Myers et al. "Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase". In: *Biochemistry* 42.23 (2003), pp. 7013–7022.

[82] John JG Tesmer et al. "The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families". In: *Nat. structural biology* 3.1 (1996), pp. 74–86.

[83] James B Thoden et al. "The small subunit of carbamoyl phosphate synthetase: snapshots along the reaction pathway". In: *Biochemistry* 38.49 (1999), pp. 16158–16166.

[84] Thorsten Knöchel et al. "The crystal structure of anthranilate synthase from Sulfolobus solfataricus: functional implications". In: *Proc. National Acad. Sci.* 96.17 (1999), pp. 9479–9484.

[85] Hongmin Li et al. "Three-dimensional structure of human γ-glutamyl hydrolase: a class I glutamine amidotransferase adapted for a complex substrate". In: *J. Biol. Chem.* 277.27 (2002), pp. 24522–24529.

[86] Marco Strohmeier et al. "Structure of a bacterial pyridoxal 5'-phosphate synthase complex". In: *Proc. National Acad. Sci.* 103.51 (2006), pp. 19284–19289.

[87] Mariya Morar et al. "Formylglycinamide ribonucleotide amidotransferase from Thermotoga maritima: structural insights into complex formation". In: *Biochemistry* 47.30 (2008), pp. 7816–7830.

[88] Alice Douangamath et al. “Structural evidence for ammonia tunneling across the ($\beta\alpha$) 8 barrel of the imidazole glycerol phosphate synthase bienzyme complex”. In: *Structure* 10.2 (2002), pp. 185–193.

[89] Barnali N Chaudhuri et al. “Crystal structure of imidazole glycerol phosphate synthase: a tunnel through a (β/α) 8 barrel joins two active sites”. In: *Structure* 9.10 (2001), pp. 987–997.

[90] Barnali N Chaudhuri et al. “Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme”. In: *Biochemistry* 42.23 (2003), pp. 7003–7012.

[91] Jan Philip Wurm et al. “Molecular basis for the allosteric activation mechanism of the heterodimeric imidazole glycerol phosphate synthase complex”. In: *Nat. communications* 12.1 (2021), pp. 1–13.

[92] Ivan Rivalta et al. “Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein–protein interface”. In: *Biochemistry* 55.47 (2016), pp. 6484–6494.

[93] Christian FA Negre et al. “Eigenvector centrality for characterization of protein allosteric pathways”. In: *Proc. National Acad. Sci.* 115.52 (2018), E12201–E12208.

[94] George P Lisi et al. “Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity”. In: *Proc. National Acad. Sci.* 114.17 (2017), E3414–E3423.

[95] Andrea C Kneuttinger et al. “Significance of the protein interface configuration for allostery in imidazole glycerol phosphate synthase”. In: *Biochemistry* 59.29 (2020), pp. 2729–2742.

[96] Maria J Gomez and Alexander A Neyfakh. “Genes involved in intrinsic antibiotic resistance of Acinetobacter baylyi”. In: *Antimicrob. agents chemotherapy* 50.11 (2006), pp. 3562–3567.

[97] Katrin Breitbach, Jens Köhler, and Ivo Steinmetz. “Induction of protective immunity against Burkholderia pseudomallei using attenuated mutants with defects in the intracellular life cycle”. In: *Trans. Royal Soc. Trop. Med. Hyg.* 102.Supplement_1 (2008), S89–S94.

Chapter 2

Methodology development

2.1 Dynamical Perturbation Contact Network

2.1.1 Network theory and proteins

Protein folding and thus functionality is principally guided by the formation of non-covalent interactions between amino acid residues[1]. Therefore, proteins can be understood as a set of amino acids in a reciprocated interaction. In mathematics, a convenient way to model a set of objects in relation is by using a network (also called graph). Because contacts between residues are reciprocated (i.e. if residue A is in contact with B, then residue B is also in contact with A), the graph necessary to model contacts in a protein is undirected (i.e. there is no preferential direction A-to-B or vice versa). A graph G is a pair of sets $G = (V, E)$; V is the set of nodes (also called vertices), here are the protein amino acids, and E is the set of edges (also called links or lines), which are unordered pairs of nodes, here contacts between two amino acids. Edges can also be given attributes, such as weight or colors. The nodes u and v of an edge u, v are called the endpoints of the edge. A node can be isolated in a graph (i.e. not connected to any other node) and the process of removing these nodes from a graph to create a new graph is called pruning. Another convenient way to mathematically represent a graph is by its adjacency matrix. The values a_{ij} of the adjacency matrix are equal to one if i and j are in relation (here contact). Since the relations are reciprocal, $a_{ij} = a_{ji}$ and the matrix is symmetric. Graphs can be allowed to contain loops, that is a node is in relation with itself, here this would represent the self-interaction of atoms in an amino acid. Such a study might be valuable but is outside the scope of this work, thus we forbid loops. Then, the adjacency matrix diagonal is zero: $\forall i \in V, a_{ii} = 0$. To account for attributes (and especially weights), attribute adjacency matrices are defined, with the values corresponding to each edge being its attribute. In the case of weight, this is named the weighted adjacency matrix.

An example of graph is depicted in Figure 2.1. This graph G contains 4 nodes and 2 edges, such that:

$$
\begin{aligned}
&G = (V, E) \\
&V = \{u, v, w, x\} \\
&E = \{(u, v), (u, x)\}
\end{aligned}
$$

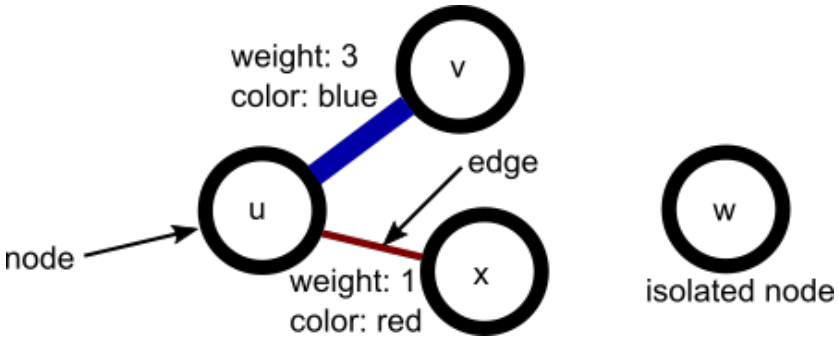


Figure 2.1: Example of weighted and colored graph.

Edges are both weighted and colored in this graph. The corresponding adjacency matrix A, the weighted adjacency matrix W and the color adjacency matrix C are thus:

$$A = \begin{array}{c} \begin{matrix} u & v & w & x \end{matrix} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} u \\ v \\ w \\ x \end{matrix} \qquad W = \begin{array}{c} \begin{matrix} u & v & w & x \end{matrix} \\ \begin{pmatrix} 0 & 3 & 0 & 1 \\ 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} u \\ v \\ w \\ x \end{matrix} \qquad C = \begin{array}{c} \begin{matrix} u & v & w & x \end{matrix} \\ \begin{pmatrix} 0 & \text{blue} & 0 & \text{red} \\ \text{blue} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \text{red} & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} u \\ v \\ w \\ x \end{matrix}$$

In most real graphs, and in the case of proteins where the number of interactions of a single node is limited, adjacency matrices are usually highly sparse (i.e. with a lot of zero elements). For operations that need using adjacency matrices, there is an incentive to represent these matrices with tools that are adapted to sparse matrices.

2.1.2 Amino Acid Networks

Proteins have thus been studied as networks of interacting residues for about 25 years[2]. In literature, such networks have been given many names: protein structure network[3], protein contact network[4], residue interaction graphs (RIG)[5], residue networks[2] or amino acid networks (AAN) used here. The original authors stated that using of the word "protein" in the name of such graphs may be misleading because it can imply that the proteins are the nodes of a graph which are connected by different types of interaction[2]. We recommend following their nomenclature and to not use the word "protein" in the network name if nodes are not proteins. Conceptually, AAN are closely related to protein contact maps. In fact, the protein contact map is simply a 2D image of the adjacency matrix of an AAN.

The different names given to AANs are also accompanied by a variety of definitions. The different types of

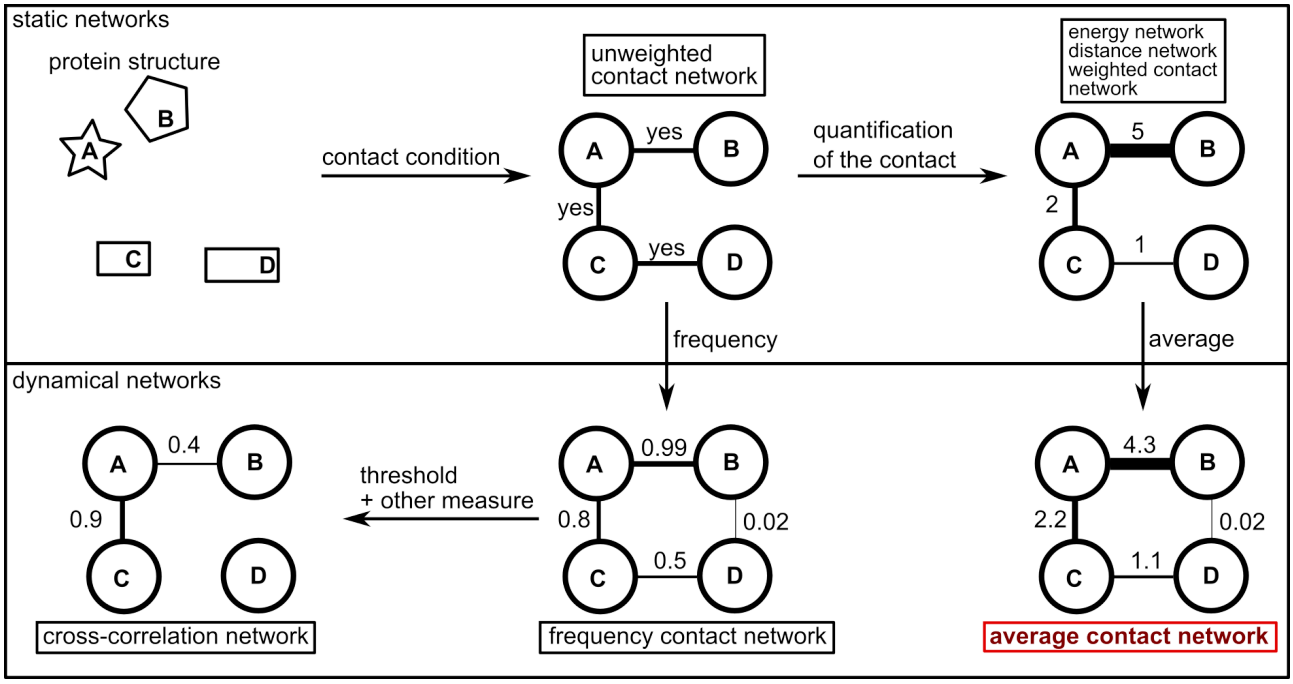


Figure 2.2: The different types of AANs. Our methodology uses average contact network.

AAN are summarized in Figure 2.2. One of the most important aspects of the definition of an AAN is the contact condition, i.e. the condition that two amino acids must satisfy in order to be considered in relation in the AAN. Generally, this condition uses both a *cutoff* distance and a *selection*, that is, residues are considered in contact if there exists a pair of atoms (one belonging to each residue) in *selection* that are at a distance below *cutoff*. There are two important groups of AAN: static networks, which describe a single structure (usually experimental), and dynamical networks, which aim at describing the dynamics of interactions (usually from an MD simulation). In dynamical networks, the contact condition is sometimes refined to residues that are in contact for a significant portion of the simulation (e.g. at least in 75% of the simulation). Static AAN predates dynamical AAN networks[2, 6] which are generally derived from static AANs.

There are two other important groups of AANs: unweighted and weighted networks. Unweighted networks are principally static networks and represent simply a binary relationship: the residues are either in interaction or not. This usually results from a simplification of the underlying problem where the contact condition uses

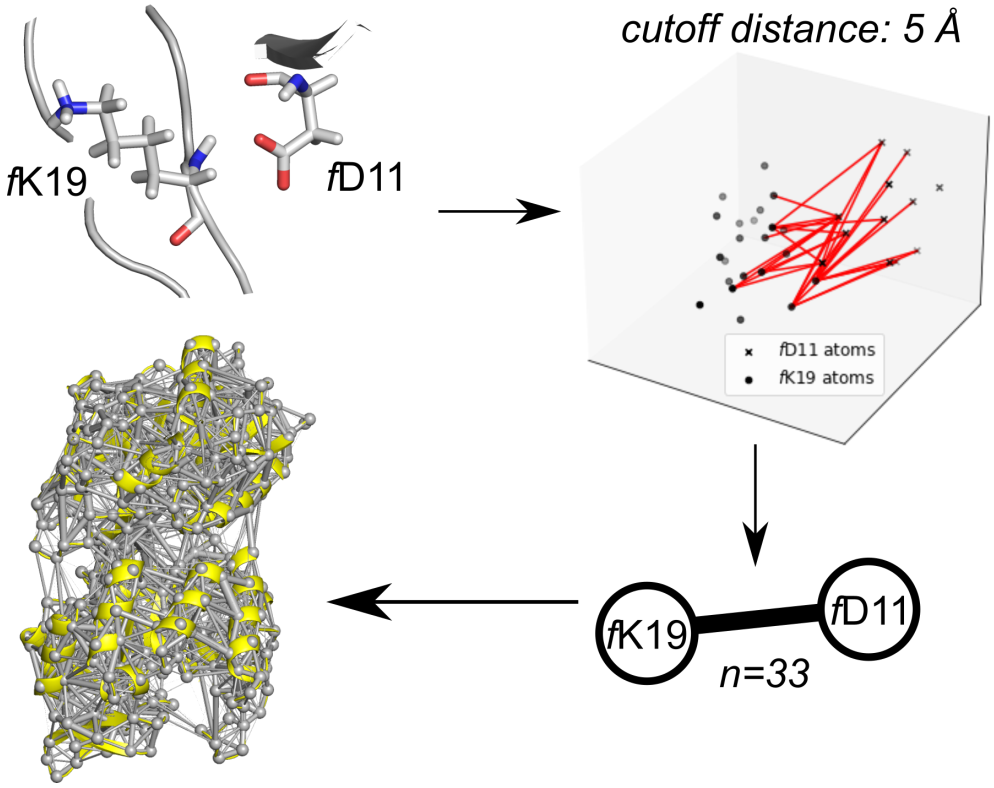


Figure 2.3: Our contact network condition and weighting.

only a single atom in the residue as a selection, such as C_α[7], C_β[8] or residue centroid[9]. This definition is very restrictive since residues have various shapes and sizes. Notably, sidechain contacts can easily be overlooked in these AANs. Because of this, the contact cutoff is usually much larger than a typical non-covalent distance and ranges from 7 to 8.5 Å. Other unweighted networks use heavy-atom[10] and sidechain heavy-atom as *selection* for contact[11]. This provides with a more complete description at the expense of more computation cost. This also allows us to lower the contact cutoff condition down to 5Å, which is the highest cutoff of the London-van der Waals forces[12]. Most static networks do not consider hydrogen atoms, since those are not resolved in experimental crystal structures, but in theory, networks built from MD simulations could consider hydrogen atoms, but this is rarely done because it vastly increases the system size and computation time. Still, in some contacts such as hydrogen bonds, taking into account hydrogen atoms may be important as the heavy-atom distance is not enough to define a hydrogen bond.

In weighted AANs, a weight is assigned to each contact to account for a more quantitative description of the contact. There are different types of weighted static AAN, and the first use of weights used the number of atomic couples that satisfy the cutoff condition as weights[13]. This introduces variability in the magnitude of contacts and discriminates residues with many close atoms from those with only a few. There is evidence that this weighting affects different types of contact differently in a protein[11]. Contacts can also be weighted by computing an interaction energy between the residues[14, 15] or by the shortest distance between two atoms in *selection*. There are many ways to creating dynamical networks from static networks and this produces generally weighted network. For instance, one can compute the frequency of presence of a contact probability along a MD simulation and assign this as the contact weight[16, 17]. These works seem to have an optimized cutoff at 4.5 Å[18] which is very close to the 5 Å suggested by London van der Waals forces. Another range of methods applies a threshold to the frequency contact network (usually around 0.75) and computes a different metric, such as the cross-correlation between the positions of C_α[6, 19, 20]. In this work, we introduced an average contact network based, instead, on the average number of interatomic contacts in the simulation (see Figure 2.3). This is a tradeoff that allows to capture more information than the frequency contact network and require less computation time than the cross-correlation networks.

AANs are representative of the general structure of a protein, but even on a relatively small protein such as IGPS, this network is really congested (see Figure 2.3) and is hard to interpret without additional analysis. Community analysis is a popular technique to facilitate AAN analysis[6, 19]. Community analysis is a distance-based method (here distance is employed as a general term, not meaning the distance between amino acids), but in AANs, weights usually grow with the magnitude of the contact. Therefore, the first step in community analysis is typically to convert the AAN to a distance network using functions such as a negative log. This weighted graph contains information on the critical nodes and paths that are important for communication within the protein.

Communities are then detected using the Girvan-Newman algorithm[21] and optimized using maximization of modularity[22]. Edges connecting communities are thus important for the communication in the protein and quantities can be averaged inside communities and between communities to facilitate analysis.

2.1.3 Perturbation Networks

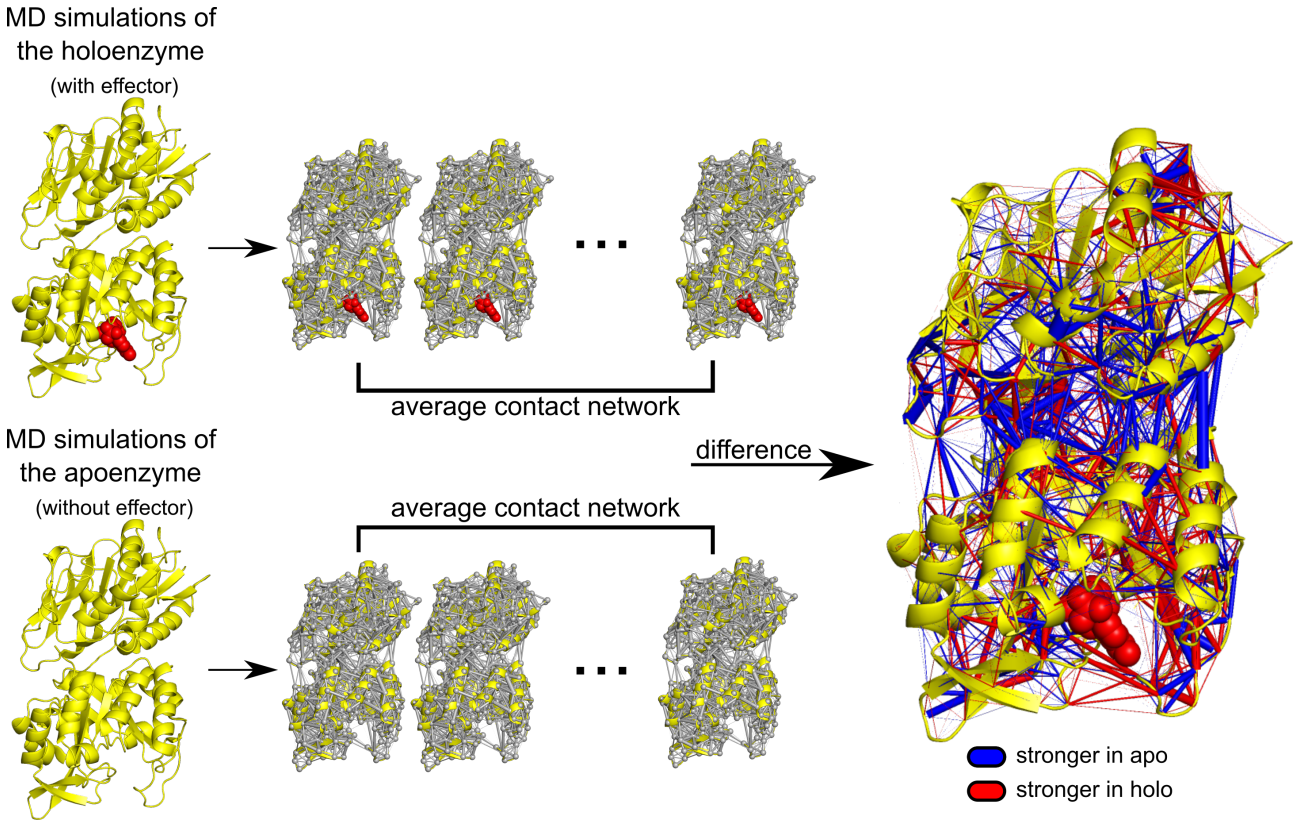


Figure 2.4: Methodology behind building a dynamical perturbation contact network

The concept of comparing different structures of proteins using a more simplified contact view dates back to contact maps[23]. Perturbation networks are generally defined as the difference between the AAN of a *reference* system and a *perturbed* system. Formally the difference is performed between the adjacency matrices of the two networks. This computation produces edges with negative and positive signs, indicating that a contact is more present in one of the systems. To visualize more easily the differences in contact, we can thus assign different colors to edges with negative and positive differences.
Differences of frequency contact networks between mutants have proven useful to study perturbations induced by mutations in the dynamical case[16] and also in the static case with differences of weighted contact networks[24]. Difference of frequency contact networks of different conformations in an allosteric system[17] managed to provide an explanation for the allosteric communication. In this work, we introduced the Dynamical Perturbation Contact Network (DPCN) (see Fig. 2.4) which is a difference of average contact networks between two proteins. To build the DPCN, we produce an AAN for the apoenzyme and an AAN for the holoenzyme using previously produced MD simulations from *T. maritima* IGPS[19]. These trajectories were previously successfully analyzed with cross-correlation networks to explain the allosteric mechanism.

The network topology of a DPCN is vastly different from the topology of an AAN. Most contacts do not change substantially and only some *outlier* contacts change weight: thus, most edges in the network have a value close to zero and only a few edges stand out while many are relatively small (see Figure 2.4). Still, a DPCN generally contains the union of edges from the two original AAN, since edges are rarely perfectly cancelling each other and the DPCN picture is quite congested which makes the relevant information is hard to interpret. There are two main ways to simplify the information. One is to create communities of a consensus network (i.e. a network which contains all edges with similar weights in the *reference* and *perturbed* system) and to sum the differences of probability between these communities[17]. The second, more simple, which we use, is to use a *threshold* value and to display only edges with a weight greater than *threshold* to emphasize areas where contact undergoes significant changes. In frequency contact networks, the threshold used for frequency change is 0.1[16] (i.e., 10% of change in the frequency). A very sensible choice of *threshold* for DPCN can be 1, as it means that on average less than 1 contact difference is established between the residues. However, this choice is not sufficient, and empirically, in this work, we found that for our systems and using a contact condition

of 5Å with heavy-atoms, a threshold of 5 or 6 produces a humanly readable graph (less than 150 edges) and accurately describes the most important contact changes in the allosteric mechanism. This methodology also pointed to contact changes that were overlooked in previous studies, and consistent with experimental findings.

The main issue with this method is that some contacts involved in certain types of contact, such as backbone hydrogen bonding and hydrophobic contacts, are significantly underweighted. We thus assumed that a heavy-atom based weighted contact network might be biased for long sidechains with polar heads. In particular, salt bridges and polar contacts stand out. This led to the idea of changing the *selection* and introducing backbone and all atoms contact networks. The use of backbone networks is very efficient in recovering backbone hydrogen bonding and proved also efficient in noticing a local unfolding event in the $h\alpha 4$ helix. However, the usage of an all-atom network showed that it was possible to highlight more hydrophobic contacts, particularly those involving residues I, L, and V (isoleucine, leucine and valine). For each of those new *selection* the *threshold* used had to be adapted: 3 for the backbone network and 25 for all atom networks.

2.1.4 Limitations of DPCNs

Despite being very successful at pointing key allosteric residues and contact changes, this methodology has a few limitations. First, two parameters are used to build the network: the contact cutoff and the weight threshold. While the first seems necessary and is easy to rationalize, the second is not. The use of a threshold was really arbitrary and produced disconnected graphs with a lot of unrelated edges. Moreover, having to optimize three different types of graph for a joint analysis seemed a bit tedious. Finally, all the contact information found in all the frames of the different simulations is collapsed into a single quantity: the average. The variability between and within simulations is completely overlooked. The DPCN strategy is intrinsically *supervised* (i.e. we have to label a system as *reference* and the other as *perturbed*). It is thus not clearly established if the variability detected between apo and effector bound IGPS is of strong magnitude, or if the variability within simulations is actually bigger than a difference that can be associated to the allosteric effect of effector binding.

References

[1] Charlotte W Pratt and Kathleen Cornely. *Essential biochemistry*. John Wiley & Sons, 2021.

[2] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential". In: *Fold. Design* 2.3 (1997), pp. 173–181.

[3] KV Brinda, Avadhesha Surolia, and Sarawathi Vishveshwara. "Insights into the quaternary association of proteins through structure graphs: a case study of lectins". In: *Biochem. journal* 391.1 (2005), pp. 1–15.

[4] Luisa Di Paola et al. "Protein contact networks: an emerging paradigm in chemistry". In: *Chem. reviews* 113.3 (2013), pp. 1598–1613.

[5] Adrian A Canutescu, Andrew A Shelenkov, and Roland L Dunbrack Jr. "A graph-theory algorithm for rapid protein side-chain prediction". In: *Protein science* 12.9 (2003), pp. 2001–2014.

[6] Anurag Sethi et al. "Dynamical networks in tRNA: protein complexes". In: *Proc. National Acad. Sci.* 106.16 (2009), pp. 6620–6625.

[7] Nikolay V Dokholyan et al. "Topological determinants of protein folding". In: *Proc. National Acad. Sci.* 99.13 (2002), pp. 8637–8641.

[8] Ali Rana Atilgan, Pelin Akan, and Canan Baysal. "Small-world communication of residues and significance for protein dynamics". In: *Biophys. journal* 86.1 (2004), pp. 85–91.

[9] Nelson Augusto Alves and Alexandre Souto Martinez. "Inferring topological features of proteins from amino acid residue networks". In: *Physica A: Stat. Mech. Its Appl.* 375.1 (2007), pp. 336–344.

[10] Lesley H Greene and Victoria A Higman. "Uncovering network systems within protein structures". In: *J. molecular biology* 334.4 (2003), pp. 781–791.

[11] Md Aftabuddin and S Kundu. "Hydrophobic, hydrophilic, and charged amino acid networks within protein". In: *Biophys. journal* 93.1 (2007), pp. 225–231.

[12] Ignacio Tinoco, Kenneth Sauer, and James C Wang. *Physical chemistry: principles and applications in biological sciences*. 544: 577 TIN. 1995.

[13] Md Aftabuddin and Sudip Kundu. "Weighted and unweighted network of amino acids within protein". In: *Physica A: Stat. Mech. its Appl.* 369.2 (2006), pp. 895–904.

[14] Xiong Jiao et al. "Construction and application of the weighted amino acid network based on energy". In: *Phys. Review E* 75.5 (2007), p. 051903.

[15] MS Vijayabaskar and Saraswathi Vishveshwara. "Interaction energy based protein structure networks". In: *Biophys. journal* 99.11 (2010), pp. 3704–3715.

[16] Urmi Doshi et al. “Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation”. In: *Proc. National Acad. Sci.* 113.17 (2016), pp. 4735–4740.

[17] Xin-Qiu Yao, Mohamed Momin, and Donald Hamelberg. “Elucidating allosteric communications in proteins with difference contact network analysis”. In: *J. chemical information modeling* 58.7 (2018), pp. 1325–1330.

[18] Xin-Qiu Yao, Mohamed Momin, and Donald Hamelberg. “Establishing a Framework of Using Residue–Residue Interactions in Protein Difference Network Analysis”. In: *J. chemical information modeling* 59.7 (2019), pp. 3222–3228.

[19] Ivan Rivalta et al. “Allosteric pathways in imidazole glycerol phosphate synthase”. In: *Proc. National Acad. Sci.* 109.22 (2012), E1428–E1436.

[20] Adam T VanWart et al. “Exploring residue component contributions to dynamical network models of allostery”. In: *J. chemical theory computation* 8.8 (2012), pp. 2949–2961.

[21] Michelle Girvan and Mark EJ Newman. “Community structure in social and biological networks”. In: *Proc. national academy sciences* 99.12 (2002), pp. 7821–7826.

[22] Mark EJ Newman and Michelle Girvan. “Finding and evaluating community structure in networks”. In: *Phys. review E* 69.2 (2004), p. 026113.

[23] Krzysztof Pawłowski, Andrzej Bierzyński, and Adam Godzik. “Structural diversity in a family of homologous proteins”. In: *J. molecular biology* 258.2 (1996), pp. 349–366.

[24] Rodrigo Dorantes-Gilardi et al. “In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: not too many links, not too few”. In: *Phys. Chem. Chem. Phys.* 20.39 (2018), pp. 25399–25410.

2.1.5 Published Article 1

The development of this methodology lead to the publication of an article in The Journal of Chemistry B in 2019. This work is the result of a collaboration with the team of Claire Lesieur at the Ampère Laboratory in Lyon and with the team of Victor S. Batista at Yale University.

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

Cite This: J. Phys. Chem. B 2019, 123, 3452−3461

pubs.acs.org/JPCB

Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks

Published as part of The Journal of Physical Chemistry virtual special issue "Young Scientists".

Aria Gheeraert,[†] Lorenza Pacini,[‡,§,∥] Victor S. Batista,[⊥] Laurent Vuillon,[§] Claire Lesieur,*[,‡,∥] and Ivan Rivalta*[,†,¶]

[†]Univ Lyon, Ens de Lyon, CNRS UMR 5182, Université Claude Bernard Lyon 1, Laboratoire de Chimie, F69342 Lyon, France

[‡]Institut Rhônalpin des systèmes complexes, IXXI-ENS-Lyon, 69007 Lyon, France

[§]LAMA, Univ. Savoie Mont Blanc, CNRS, LAMA, 73376 Le Bourget du Lac, France

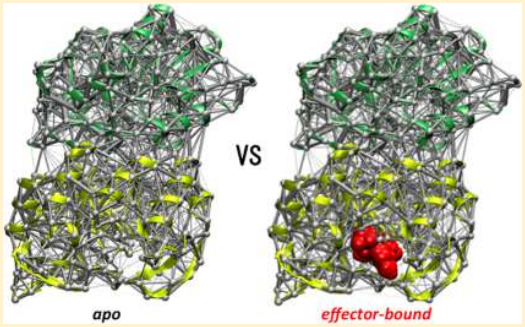
[∥]AMPERE, CNRS, Univ. Lyon, 69622 Lyon, France

[⊥]Department of Chemistry and Energy Sciences Institute, Yale University, P.O. Box 208107, New Haven, Connecticut 06520-8107, United States

[¶]Dipartimento di Chimica Industriale "Toso Montanari", Università degli Studi di Bologna, Viale del Risorgimento 4, I-40136 Bologna, Italy

Supporting Information

ABSTRACT: Elucidation of the allosteric pathways in proteins is a computational challenge that strongly benefits from combination of atomistic molecular dynamics (MD) simulations and coarse-grained analysis of the complex dynamical network of chemical interactions based on graph theory. Here, we introduce and assess the performances of the dynamical perturbation network analysis of allosteric pathways in a prototypical V-type allosteric enzyme. Dynamical atomic contacts obtained from MD simulations are used to weight the allosteric protein graph, which involves an extended network of contacts perturbed by the effector binding in the allosteric site. The outcome showed good agreement with previously reported theoretical and experimental extended studies and it provided recognition of new potential allosteric spots that can be exploited in future mutagenesis experiments. Overall, the dynamical perturbation network analysis proved to be a powerful computational tool, complementary to other network-based approaches that can assist the full exploitation of allosteric phenomena for advances in protein engineering and rational drug design.



INTRODUCTION

The characterization of allosteric mechanisms in proteic systems is a challenging task due to the intrinsically complex and elusive nature of protein allostery.[1,2] The allosteric phenomena, ubiquitous in biology and not exclusive of proteins, have been shown to feature both structural and energetic origins.[3,4] Statistical ensemble models rooted in the historical phenomenological models of allostery[5,6] have suggested a unifying view of the operational allosteric mechanisms.[7,8] Still, to fully exploit the potential of allosteric phenomena for protein engineering and rational drug design, where allosteric systems (and particularly enzymes) can be manipulated to inhibit/enhance their (catalytic) activity or new allosteric sites can be discovered,[9−15] system-specific information is required.

The fundamental process occurring in allosteric enzymes is the binding of an effector ligand at the allosteric site distant from the functional active site, enabling the regulation of the corresponding enzymatic function; see Figure 1. Modulation of functions in allosteric enzymes is linked to the communication from the active to the allosteric site,[4,13,16] with effector-induced changes of residues dynamics and protein disorder altering either the affinity of the substrate for the active site (K-type) or the reaction rate (V-type) of the enzymes. The allosteric signal has been found to propagate through conserved amino acid residues[17−19] and, in general, it is expected to involve physicochemical interactions between "allostery-related" residues that comprise various secondary structure elements, defining (multiple) "allosteric pathways" of the proteic systems.[20]

Classical molecular dynamics (MD) simulations provide invaluable information on protein dynamics at atomistic

Received: February 8, 2019
Revised: March 31, 2019
Published: April 3, 2019

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

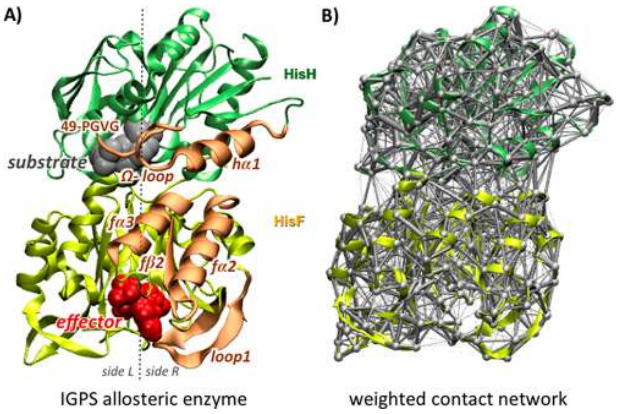


Figure 1. (A) IGPS allosteric (V-type) enzyme, with substrate (in gray) binding in the active site of the HisH glutaminase domain (in green) and the effector (PRAFR, in red) binding in the allosteric site at the bottom of the HisF cyclase domain (in yellow). Previously recognized secondary structure elements belonging to the IGPS allosteric pathways are shown (in orange), linking the allosteric and active sites at sideR of the enzyme. (B) Example of a 3D representation of the IGPS protein network, with nodes of the graph located at the α carbon atoms of the enzyme and edges connecting nodes weighted by the number of contacts between residues pairs.

resolution, representing a fundamental tool for the elucidation of such allosteric pathways,[21−23] whose experimental detailed characterization is certainly extremely challenging. While MD simulations enclose the dynamical information underpinning the allosteric effects,[24] analyzing complex networks of interactions between (a generally large number of) fluctuating amino acid residues and finding the allosteric signal paths within the wiring of such a network call for help from graph theory techniques. Network analysis of MD trajectories that incorporate allosteric motions has delivered, in fact, characterization of allosteric pathways and identification of allostery-related amino acid residues in various biological systems,[25−32] and helped rational discovery of allosteric modulators.[33,34] In particular, combining nuclear magnetic resonance (NMR) relaxation dispersion experiments with community analysis of dynamical networks,[35] based on mutual information on correlated protein motions obtained from MD simulations, we have revealed the allosteric pathways of the imidazole glycerol phosphate synthase (IGPS) enzyme from the thermophile *Thermotoga maritima*; see Figure 1.[25]

IGPS is a prototype allosteric enzyme absent in mammals but involved in essential biochemical pathways (histidine and purine synthesis) of pathogens, and thus, it is a potential target for antifungal, antibiotic, and herbicide development.[36−38] As shown in Figure 1, two tightly associated proteins constitute the IGPS V-type allosteric enzyme: (i) the HisH glutamine amidotransferase that catalyzes the hydrolysis of the substrate (glutamine) and (ii) the HisF cyclase where the effector PRFAR, i.e., *N*′-[(5′-phosphoribulosyl)formimino]-5-amino-imidazole-4-carboxamide-ribonucleotide, binds without affecting the glutamine binding affinity in HisH but accelerating its hydrolysis by ca. 5000-fold.[39] Our synergistic theoretical and experimental investigations suggested secondary structure elements and key residues involved in the allosteric signal propagation induced by the PRFAR binding to the apo IGPS protein. The IGPS allosteric mechanism involves a sequence of interactions that alter the dynamics of specific regions in one side of the IGPS complex (sideR; see Figure 1), including hydrogen bonds in the flexible loop1 and hydrophobic interactions in the fβ2 strand at the HisF allosteric site, ionic interactions at the HisF/HisH interface involving the fα2, fα3, and hα1 helices, as well as hydrogen bonding between the Ω-loop and a conserved (49-PGVG) sequence adjacent to the active site, namely, the oxyanion strand. These effector-induced interactions were shown to alter the overall HisF/HisH relative fluctuations (named *breathing motion*), promoting rotation of the conserved oxyanion strand associated with an inactive-to-active allosteric transition. The outcome of the community network analysis stimulated experimental mutagenesis studies focused on the suggested allostery-related amino acid residues,[40] as well as rational design of allosteric inhibitors able to knockout the IGPS allosteric signal propagation by interfering with the suggested allosteric pathways.[34] The proposed community network analysis employed the correlations of motion between residue pairs (in close contact) to weight the protein graph, resulting in a communication network where the betweenness centrality measure can decipher the most important nodes that transfer the allosteric signal. While proving to be an extremely powerful and transferable approach that has been employed to other allosteric systems,[26,27] this tool was revealed to be not very user-friendly and was particularly tedious to use when applied to large proteic systems. Very recently, we have proposed an alternative tool to the community network analysis that introduced the eigenvector centrality metric to analyze the correlated motions obtained from the MD simulations, providing a cost-effective approach that properly captures the IGPS allosteric pathways and allows the user to disentangle contributions to allostery due to short- or long-range correlations.[28] Nevertheless, both the betweenness and eigenvector centrality measures have been used to analyze protein graphs weighted by the correlated motions of α carbon atoms. These correlations certainly comprise only part of the network of interactions that are altered upon effector binding. Here, we explore the use of inter-residue physical contacts to build the weighted protein network, thus moving from a physical to a geometrical measure that tracks down and approximates the chemical interactions between residues. This type of weighted contact network analysis has been successfully used to infer protein dynamics and to determine structural robustness to mutations in proteins, it being powerful to understand how a local change can produce global changes that are associated with retention or loss of protein functions.[41−43] Here, we propose to use this weighted network approach to study allostery and to compute local perturbations of contacts induced by the effector binding, which are expected to propagate in the allosteric enzymes through protein dynamics. The use of unweighted networks based on a binary measure of dynamical contacts could be also envisioned to this aim, possibly providing a more coarse-grained picture of the effector-induced dynamical contacts with respect to networks weighted with the number of atomic contacts. In particular, taking advantage of the atomistic details contained in MD simulations one can account for dynamical contacts and their effector-induced modifications by averaging the number of contacts for each residue pair along a MD trajectory, using this information to weight the protein network. A similar approach, based on dynamical network of inter-residues contacts, has been used to reveal the allosteric effects of mutations in the

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

catalytic activity of the Cyclophilin A enzyme, proving to be potentially able to identify key residues in the allosteric signal propagation.[44] Here, we propose the use of the dynamical contact network approach to study allosteric perturbations induced by effector binding, instead of mutations, performing a dynamical perturbation network analysis of IGPS allostery. IGPS is, indeed, a prototypical allosteric enzyme whose allosteric pathways have been previously characterized in detail by means of MD simulations and network models and validated by NMR and biochemical and mutagenesis experiments, providing an ideal system to assess the performances of the perturbation network analysis for capturing allostery.

COMPUTATIONAL DETAILS

In this work, we used structural models of the apo and PRFAR-bound IGPS complexes and MD simulations that have been described elsewhere,[25] in order to fairly compare the results of the perturbation networks with those of the previously reported community network analysis. In our previous analysis we have showed that the time-averaged weighted networks, based on MD trajectories 100 ns (ns) long, adequately describe the dynamical networks, capturing the protein conformational changes induced by effector binding during the early dynamics of the IGPS complexes.[25] Therefore, previously obtained MD trajectories, including four independent simulations of 100 ns for the apo IGPS protein and four independent simulations of 100 ns for the PRFAR-bound IGPS complex, have been used.[25] MD simulations of the IGPS complexes were based on the AMBER-ff99SB[45] force field for the IGPS protein and on the generalized Amber force field[46] for the PRFAR ligand, using the NAMD2 software package.[47] Production run MD simulations succeeded a pre-equilibration procedure involving slow heating to 303 K, gradual release of atomic positions constraints, and subsequent unconstrained MD simulations of 4 ns in the canonical *NVT* ensemble using Langevin dynamics. Production runs were performed in the *NPT* ensemble at 303 K and 1 atm (using the Langevin piston) for 100 ns after reaching the equilibrium volume (i.e., after ca. 2−3 ns). Periodic boundary conditions and the particle mesh Ewald method[48] were employed, with van der Waals interactions calculated using a switching distance of 10 Å and a cutoff of 12 Å. A multiple time-stepping algorithm[49,50] was adopted, with bonded, short-range nonbonded, and long-range electrostatic interactions were evaluated at every one, two and four time steps, respectively, using a time step of integration set to 1 fs.

Protein Weighted Networks. In the protein network each node represents an amino acid residue (see Figure 1), with connections between nodes (namely, the graph edges) being defined according to atomic proximity: for each pair of residues, if there exists a couple of atoms, one in each residue, whose distance is below a given distance cutoff, then the two atoms satisfy the "contact condition" and the two corresponding nodes/residues are linked by an edge. In line with previously reported perturbation network analysis,[41,42] we used a 5 Å distance cutoff to define the contact condition, this choice allowing a fair comparison with previously reported community network analysis,[25] where the same distance cutoff has been adopted.[25] The effect of the distance cutoff parameter on the perturbation network analysis will deserve further investigation for application of the proposed network approach to other allosteric systems. The protein weighted network is then built by assigning to each edge (linking the *i*th and *j*th residues) a weight w_{ij}, which equals the number of contacts between two residues, i.e., the number of atom pairs that satisfy the contact condition between the *i*th and *j*th residues (see Figure S1 in the Supporting Information). To compute the number of contacts among the IGPS residues in the apo and PRFAR-bound complexes, and thus the corresponding contact weighted networks, we used the atomic coordinates extracted every 100 ps from the MD trajectories. The choice of the time interval to extract the atomic coordinates (and thus to compute the number of contacts) is bound to that one adopted in the community network analysis,[25] in order to provide a consistent comparison between the two different approaches. In particular, after concatenating the four independent simulations per each IGPS system (apo and effector bound) the number of atomic contacts are computed by averaging over the corresponding MD frames. If an edge is not present in a given frame, i.e., if two residues do not satisfy the contact condition in that very frame, its weight is set to zero and it will be still averaged with its weights at the remaining frames. As we will illustrate in the Results section, the computation of atomic contacts could include all protein atoms or it could exclude just the hydrogen atoms.

Dynamical Perturbation Network. The procedure described above generates two weighted contact networks, one for the apo protein and one for the PRFAR-bound complex, each one containing (in their average weights) information on the contacts dynamics of all residues pairs in the corresponding IGPS protein. As shown in Figure 2, a weighted network representing the perturbations of the contacts dynamics induced by the effector binding, i.e., the dynamical perturbation network, can be constructed by considering as edge weight for each residues pair the differences in weights (weight link) between the two IGPS proteins, i.e., using the perturbation weight ($w_p = w^{PRFAR} - w^{APO}$) to build the network. To simplify the visual inspection of such perturbation network, the edges are colored in red if

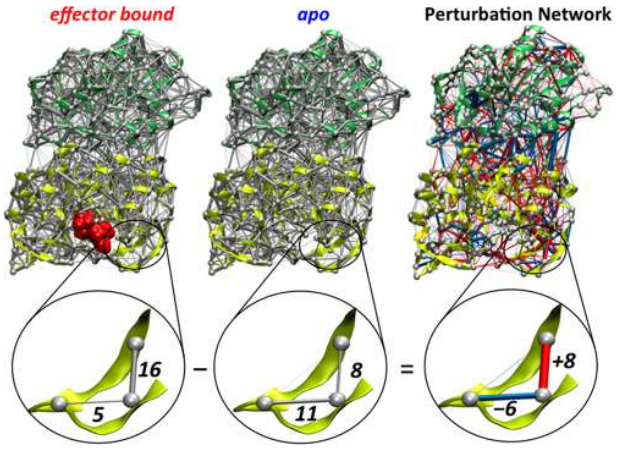


Figure 2. Construction of the IGPS perturbation network. The average contact weights of residue pairs of the apo IGPS are subtracted from that of the effector-bound binary complex. 3D representations of the average and perturbation networks and a corresponding close-up view are depicted. Reduction and increase of the number of contacts between residue pairs upon PRFAR binding are indicated with blue and red links, respectively. The widths of the links in each average and perturbation network are normalized to facilitate their visualizations.

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

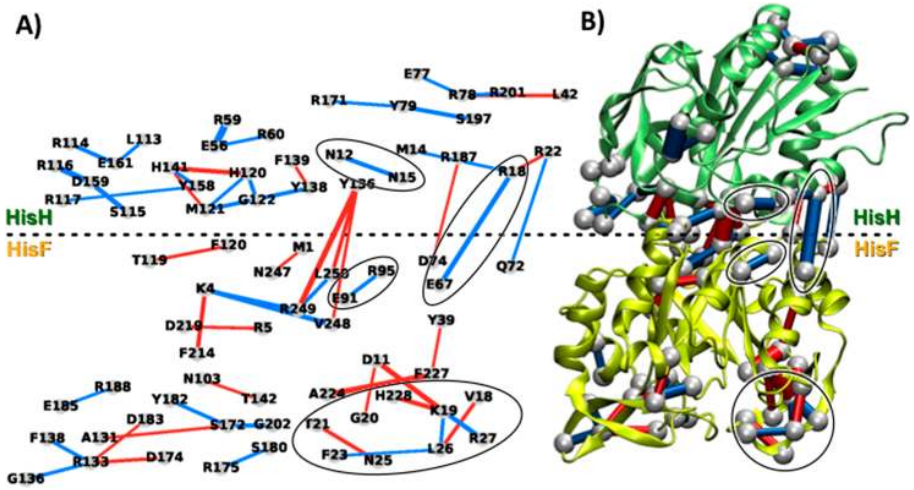


Figure 3. Perturbation network associated with PRFAR binding to IGPS, using a weight threshold $w_t = 6$ for the network visualizations. (A) 2D projection and (B) 3D representation of the perturbation network, showing reduction (blue lines) and increase (red lines) of the number of contacts between heavy atoms upon PRFAR binding. Perturbations associated with previously reported allosteric pathways[25] are highlighted with black circles.

PRFAR binding induces an increase in weight ($w_p > 0$), i.e., an increase in number of atomic contacts for a given residue pair, and in blue if PRFAR binding instead reduces the contact weight ($w_p < 0$); see Figure 2. To allow visualization of the 2D and/or 3D representations of the IGPS perturbation network, which contains around 10^4 edges, a weight link threshold (w_t) can be applied so that only the edges whose weight is greater than the chosen weight threshold, i.e., $|w_p| > w_t$, are kept for visualization. If a node loses all its edges during the subtraction process, it is also removed from the graph representation for simplicity. The impact of the weight threshold values on the graph visualization changes according to criterion used to compute the number of contacts. For instance, excluding hydrogen atoms from the count of atomic contacts reduces significantly the average weights values in each protein network and consequently also the weights in the dynamical perturbation network, allowing w_t values of 5 or 6 to be large enough to make the number of edges to visualize being less than one hundred. To obtain a similar number of edges while including all atoms in the counts of atomic contacts requires much larger weight thresholds ($w_t > 20$).

RESULTS AND DISCUSSION

Figure 3 shows the perturbation network associated with effector (PRFAR) binding to the IGPS protein, using a weight threshold $w_t = 6$ and considering only contacts between heavy atoms. Notably, the PRFAR perturbations are spread over different regions of the enzyme and reach also HisH residues located quite far from the effector site in HisF.

At the effector binding site, perturbations can be found at both sideL and sideR of the enzyme due to the hydrogen bonds created by the PRFAR molecule at these sides. In fact, it has been shown[25] that the hydroxyl groups of the PRFAR glycerol moiety create a hydrogen bonding network with the fG202 residue at the end of fβ7 (see Figure 4 in ref 25). The highly conserved fG202 residue is indeed detected by our network analysis, which further shows propagation of this perturbation across sideL. The fT142 and fR133 residues appear as central nodes for PRFAR signal propagation at HisF sideL. At sideR, the perturbation network analysis indicates that upon PRFAR binding contacts in the fβ8−fα8′ turn of HisF are significantly affected, with an increase of contact between the fA224 and fF227 hydrophobic residues. Indeed, the glycerol side phosphate group of PRFAR is known to be involved in hydrogen bonds with the backbone of fA224 and the fS225 side chain located in the fβ8−fα8′ turn.[18,25] Notably, near the fβ8−fα8′ turn is located the important loop1, for which the perturbation network analysis shows drastic modifications of contacts upon effector binding, in agreement with previous results (see Figure S2 in the Supporting Information for direct comparison).[25] In particular, the loss of contacts in the loop, associated with residues fK19, fF23, fL26, and fR27 (blue lines in Figure 3) is compensated by an increase of contacts between residues fD11, fK19, fG20, and fH228. Thus, the invariant fK19 plays a central role in the perturbation network being crucial for the signal transduction at sideR of HisF, as demonstrated by experimental biochemical data on the fK19A mutant.[40] In fact, our network analysis allows recognition of important interactions between the highly conserved fD11 (in fβ1) and fK19, occurring only upon PRFAR binding (see Figure S2 in the Supporting Information) and suggesting the participation of fβ1 in the allosteric pathways and fD11 as another possible allosteric spot in IGPS.

As shown in Figure 3, while in HisF the increase of contacts (red lines) induced by the effector binding is almost compensated by a few contact losses (total weight gain is ca. 19), in HisH most of the perturbations are characterized by contact losses (blue lines, with total weight loss ca. 111). Among the pairs that feature contact loss in HisH, it is worth highlighting the hN12−hN15 pair connecting the Ω-loop and the hα1 helix, two secondary structure elements that have been indicated among the allosteric pathways and a crucial connection already pointed up by the community network analysis.[25]

The HisF/HisH interface also features perturbation of relevant contacts upon effector binding, in agreement with the change in breathing motion between apo and PRFAR-bound IGPS previously reported.[25] As pointed out by previous analysis of MD simulations, the allosteric effect of PRFAR expresses at the protein−protein interface as rearrangement of

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

the ionic interactions among the fE67, fE71, and hR18 residues, with rupture of the hR18−fE67 salt bridge (connecting the hα1 and fα2 helices) upon effector binding (see Figure S3 in the Supporting Information).[25] Notably, the contact loss in the hR18−fE67 ion pair interaction appears as one of the largest perturbations in the network (see Figure 3a) and it is accompanied by other significant changes in sideR. In particular, the fE91−fR95 salt bridge within the fα3 helix is also detected by the perturbation network analysis, in agreement with the fact that both hR18−fE67 and fE91−fR95 salt bridges represent the most relevant changes in ionic interactions at sideR associated with the allosteric pathways (see Figure S3 in the Supporting Information).[25] Other interactions at the HisF/HisH interface are evidenced by the perturbation network analysis: (i) contacts between the hY136 residue[51] in hβ8 and residues fV248, fR249, and fL250 in the C-terminal domain of HisF and (ii) two contact pairs connecting the hα1 and hα4 helices with the fα2−fβ3 turn, i.e., hR22−fQ72 and hR187−fD74, respectively. The interactions involving the polar hY136 residue show a global increase of the number of contacts of this residue with HisF, upon effector binding. This is due to the change of H-bonding between hY136 and fN247, which brings hY136 closer to the flexible HisF C-terminus (see Figure S4 in the Supporting Information). These changes of contacts comprise fR249, a highly conserved residue involved in the π-cation hW123−fR249 molecular hinge,[18] but are not associated with formation/disruption of very strong interactions that might alter significantly the IGPS structure. Still, the observed rearrangement of the HisF C-terminus involving the molecular hinge is in line with modification of the relative HisF/HisH (breathing) motion, an indirect effect associated with the disruption of the hR18−fE67 interface salt bridge. However, a contact loss is observed for the hR22−fQ72 pair upon effector binding, which involves hα1 and the fα2−fβ3 turn, respectively, and it appears to be directly related to the breaking of the adjacent hR18−fE67 salt bridge also connecting hα1 with HisF. The hR22−fQ72 contact loss is somehow compensated by the formation of a nearby hR187−fD74 salt bridge, involving the hα4 helix. The contacts encompassing residues hR22, hR187, fE67, fQ72, and fD74 are all located at sideR of the HisF/HisH interface, which has been indicated as a crucial region for the IGPS allosteric communication and thus deserves a more detailed analysis.

Figure 4 shows the perturbation network representation using weight threshold $w_t = 5$ that allows a detailed view of the interactions involved in the important region around the hR18−fE67 salt bridge. In addition to detecting the hR18−fE67 salt-bridge breaking, a recognized effect of PRFAR binding inducing separation of the hα1−fα2 elements,[25] the perturbation network analysis also indicates that propagation of the allosteric signal through the HisF/HisH interface involves ionic interactions that were not previously detected. In particular, the formation of the hR187−fD74 salt bridge that connects hα4 helix with the fα2−fβ3 turn in the PRFAR-bound complex is concomitant with the breaking of the hR22−fD74 salt bridge between the hα1 and the fα2−fβ3 turn, which is thus involved in the modifications of ionic interactions promoted by the hR18−fE67 salt-bridge disruption (see Figure S3 in the Supporting Information).[25] Notably, these results are in agreement with NMR dispersion experiments indicating that residues in the fα2−fβ3 turn (e.g., fI73 and fI75) are among those that have the largest dynamical changes upon effector binding.[52] Therefore, we propose that the fα2−fβ3 turn and the hα4 helix are secondary structure elements that are involved in the allosteric communication in IGPS and that residues hR22, hR187, and fD74 are potentially good candidates for mutagenesis experiments.

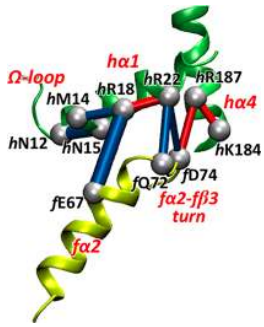


Figure 4. 3D representation of the perturbation network in the region close to the hR18−fE67 salt bridge. A weight threshold $w_t = 5$ is used for the network visualization. The perturbations associated with PRFAR binding show the relative reduction (blue lines) and increase (red lines) of the number of contacts between heavy atoms. The hR18−fE67 salt-bridge rupture upon effector binding is associated with modifications of polar and ionic interactions between the hα1 and hα4 helices and the fα2−fβ3 turn, along with contact losses and partial unfolding at the beginning of hα1 helix, where the Ω-loop is located.

In the proximity of the sideR interface region, the hN12 and hN15 residues belonging to hα1 and Ω-loop, respectively, have been suggested by the community network analysis to be important for the IGPS allostery,[25] allowing communication between the hα1 helix and the HisH active site via the Ω-loop. Figure 4 shows that the hN12−hN15 contact loss captured by the perturbation network is associated with other PRFAR-induced losses, i.e., the contacts in the hR18−hM14 and hR18−fE67 pairs. Overall, these modifications induced by PRFAR binding involve a partial unfolding of hα1 helix as a response to the hR18−fE67 salt-bridge rupture (see Figure S5 in the Supporting Information) and propagate toward the HisH active site via the Ω-loop.

The perturbation network analysis, thus, is quite useful for capturing the propagation of the PRFAR allosteric signals, providing direct visualization of allosteric effects as changes in the residue contacts. The above analysis based on the contacts between heavy atoms, indeed, detected most of the secondary structure elements in the known allosteric pathways,[25] including loop1, fα2, fα3, hα1, and Ω-loop, and indicated new secondary structures encompassing fβ1, fα2−fβ3 turn, and hα4 along with other key residues, like fK19, fD11, fD74, hR22, and hR187. Nevertheless, two important elements of the allosteric pathways, namely, the fβ2 strand in HisF and the 49-PGVG sequence in HisH active site, are not observed even among the nondescribed perturbations appearing in the computed network (see Figure 3). The missing secondary structure elements involve hydrophobic interactions (between fβ2 and loop1) and backbone hydrogen bonds (between 49-PGVG and Ω-loop), suggesting that the omission of hydrogen atoms (H's) in the count of residue contacts might be the reason for such a lack of detection of these important perturbations. However, H's are usually discarded in perturbation network analysis of mutated proteins because they are not resolved in X-ray structures and their presence

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

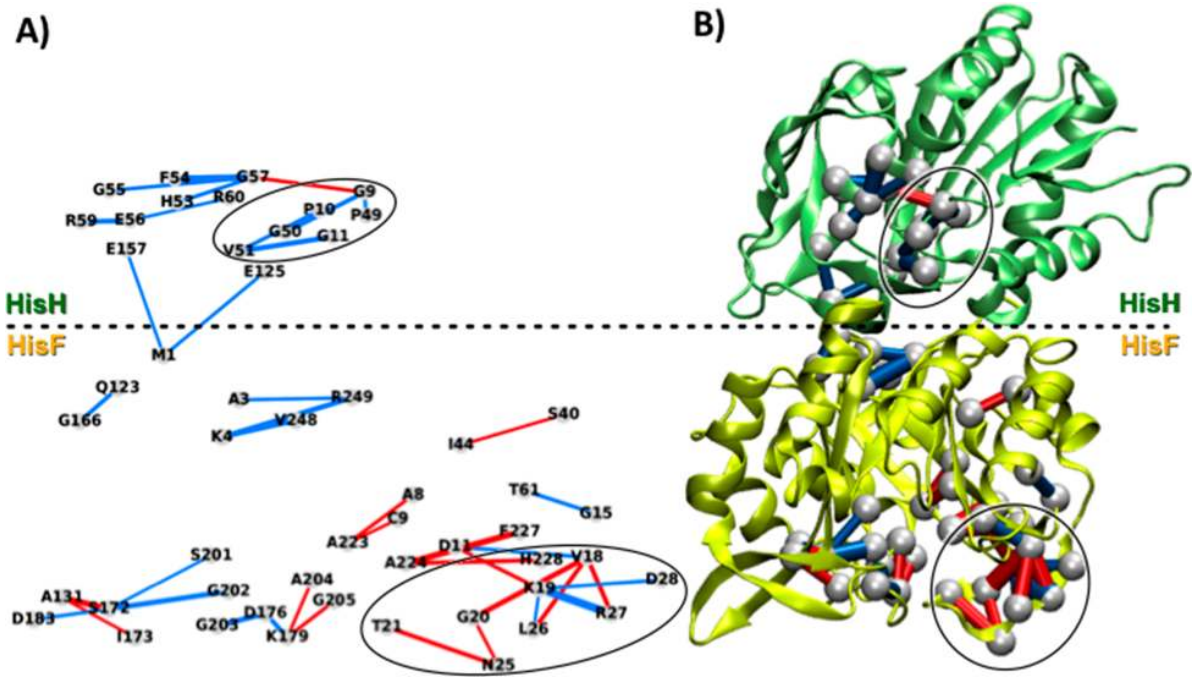


Figure 5. Perturbation network associated with PRFAR binding to IGPS, computed only for backbone atoms (including hydrogens) and using a weight threshold $w_t = 5$ for the network visualizations. 2D projection (A) and 3D representation (B) of the network, showing reduction (blue lines) and increase (red lines) of the number of contact atoms upon PRFAR binding.

significantly increases the number of contacts for each pair, adding sizable noise in the data analysis. To limit such a drawback, here we considered inclusion of the hydrogens in the perturbation networks while separating the analysis of backbone atoms (that do not contain many H's) from that of amino acid side chains.

Figure 5 shows the perturbation network analysis restricted to the backbone atoms while including hydrogens. This analysis allows focusing on the effector perturbations induced in the IGPS backbone. The backbone network shares some features with the perturbation network analysis of heavy atom contacts but it also highlights some perturbations previously overlooked. The backbone analysis, in fact, confirms the presence of strong perturbations in the PRFAR binding site, with detection of residues fG202 and fA224 and the H-bonds redistribution in loop1, as previously described. However, new perturbations stand out when the side chain contacts are removed from the network. In particular, the invariant fS201 and the highly conserved fG202, fG203, and fG205 residues of the SGGXG sequence at the fβ7−fα7 turn all feature perturbed backbone H-bonds. These perturbations can be viewed as a consequence of the hydrogen bonding network rearrangements induced by the PRFAR glycerol hydroxyls and phosphate groups at sideL of the effector binding site.[25] Moreover, the backbone analysis also catches the increase of contacts among the highly conserved residues fA224, fF227, and fH228, which is associated with a partial folding of the fβ8−fα8′ turn at sideR of the PRFAR binding site.

It is worth noting that the fD11−fK19 ion-pair contact, strongly reinforced in the presence of PRFAR, unexpectedly appears in the backbone perturbation network. This result provides direct evidence of this interaction being not associated with the formation of a fD11−fK19 salt bridge (as could be expected for an ion pair) but to hydrogen bonding between the fD11 side chain and the fK19 backbone; see Figure 6.

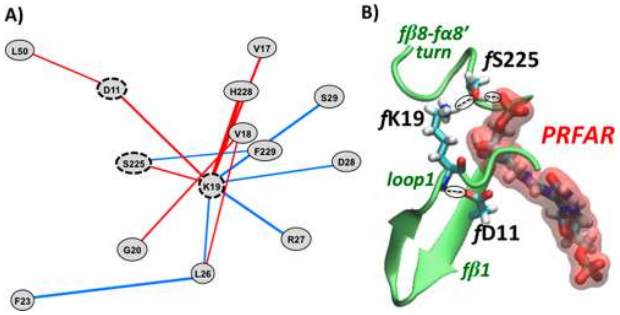


Figure 6. (A) Perturbation network around residue fK19 associated with PRFAR binding to IGPS, computed for all atoms (including hydrogens) and using a weight threshold of $w_t = 19$ for visualization (left panel) and (B) representative configuration of the H-bonding network in the PRFAR-bound complex associated with the fD11, fK19, and fS225 residues, also showing the partial folding of the fβ8−fα8′ turn.

The backbone interactions perturbed at the HisF/HisH interface are rather limited and are restricted to the highly flexible HisF N-terminus (fM1), getting in contact with the hβ7 and the hβ9 strands, i.e., with the backbone of residues hE125 and hE157, respectively. However, important backbone perturbations are found in a localized region of HisH, remarkably close to the active site. In fact, as shown in Figure 5a, the backbone network analysis clearly catches the allosteric effect associated with the 49-PGVG (oxyanion) strand that, as previously shown,[25] loses contacts with the Ω-loop due to the hydrogen bond breaking between hP10 and hV51 (see Figure 9 in ref 25). Notably, three residues of the conserved 49-

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

PGVG sequence (i.e., hP49, hG50, and hV51) are found to lose contacts with Ω-loop residues hG9, hP10, and hG11, in line with the fact that the separation of these two secondary structure elements is associated with rotation of the oxyanion strand near the substrate binding site. Beyond the remarkable ability of the perturbation network to retrieve the allosteric effects in the active site, this analysis also suggests effector-induced alterations that were overlooked in previous studies. In fact, the loss in contacts between 49-PGVG and the Ω-loop appears to be associated with a partial unfolding of the hα2 helix (which is next in sequence to the oxyanion strand), featuring the decrease of contacts between hH53, hF54, hE56, hG57, hR59, and hR60 residues, only slightly compensated by the strengthening of the hG9−hG57 interaction. The partial unfolding of helix hα2 is contiguous and it assists the mechanistically relevant oxyanion strand flip and it should be thus considered as part of the allosteric pathways.

The perturbation network analysis using heavy atom contacts did not show a significant number of interactions among hydrophobic residues being affected by the effector binding. By including hydrogens in the count of contacts, the percentage of hydrophobic interactions that participate in the perturbation network increases (see Figure S5 in the Supporting Information). However, as mentioned above, the whole network including H's represents a challenging graph to analyze since it contains a large number of contacts and a sizable amount of noise. As mentioned in the method section, in order to produce 2D (or 3D) representations that can be visually inspected (e.g., with number of perturbed pair <100) a large weight threshold (w_t around 20) has to be applied to such a network (see Figure S6 in the Supporting Information). More than looking at the whole network including H's, a more effective analysis can be performed by inspection of specific clusters of perturbations. For instance, Figure 6 shows the analysis of local perturbations around the key fK19 residue in loop1, indicating that rearrangements of contacts in loop1 are connected to residue fL50, previously reported as part of a hydrophobic cluster in the fβ2 strand,[25] via the fD11 residue. Moreover, the modifications of the fD11−fK19 contact upon effector binding are correlated with the partial folding of the fβ8−fα8′ turn, as detected by the backbone analysis but here involving residues fS225 and fH228. Notably, it has been shown that fS225 is H-bonded to the glycerol phosphate group of PRFAR,[25] and thus we performed a detailed investigation of the fS225-PRFAR H-bonds in relation to the fK19 residue along the MD simulations. We found that the fD11−fK19 contact modified upon effector binding promotes the formation of a H-bond network between fK19 (in loop1) and fS225 (in the fβ8−fα8′ turn) and the glycerol phosphate group of the PRFAR. All these observations explain the inhibition of allosteric signals in the K19A mutant[40] and confirm the importance of both fD11 and fK19 residues for the allosteric communication in HisF. At the same time, the outcome claims for inclusion of the folded fβ8−fα8′ turn as a secondary structure element of the IGPS allosteric pathways.

Finally, by limiting the perturbation network analysis to the contacts among side chains (while including hydrogen atoms), some interesting features stand out at the HisF/HisH interface. In particular, the hM121 residue stands out in the side chain network (see Figure S7 in the Supporting Information) as it features several contact perturbations with the invariant fR5, fK99, and fE167 residues that belong to the ammonia tunnel gate of the HisF barrel[51] and with the highly conserved fD98 of the structurally important fD98−hK181 salt-bridge anchor.[25,51] It has been previously shown that the PRFAR binding, indeed, alters the dynamics of these conserved residues that are associated with important structural features of the complex IGPS enzyme.[25,34] Thus, these results demonstrate that the perturbation network analysis of side chains can catch most of the structurally important conserved residues that are perturbed by the effector binding.

In summary, as shown in Figure 7, the perturbation network is a powerful tool for the characterization of the IGPS allosteric

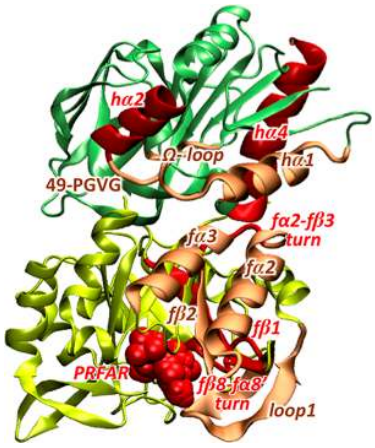


Figure 7. Representation of the secondary structure elements involved in the allosteric pathways as predicted by previously reported studies (in light orange) and by the perturbation network analysis in this work (in light orange and red).

pathways based on analysis of MD trajectories, allowing recognition of previously overlooked allosteric spots. In particular, the use of the perturbation network approach showed that with just the analysis of the heavy atom contacts most of the secondary structure elements involved in the allosteric pathways are already detected. In addition to that, the involvement of the fβ1, fα2−fβ3 turn, and hα4 secondary structures (and related key residues) in the allosteric signal propagation has been recognized by perturbation of heavy atom contacts. The addition of hydrogen atoms in the contact counting and the concomitant restriction of the analysis to the backbone atoms readily provided the detection of folding/unfolding events during the MD simulations that are strictly connected to the signal propagation, including partial folding of the fβ8−fα8′ turn in the effector binding site and the partial unfolding of the hα2 helix in the proximity of the substrate binding site.

■ CONCLUSIONS

The dynamical perturbation network analysis has been proposed and assessed for the investigation of allosteric pathways in the IGPS enzyme, a prototype allosteric system that involves known allostery-relevant amino acid residues and secondary structure elements. The network analysis of dynamical inter-residue atomic contacts, obtained from averaging several independent MD simulations of the apo and effector-bound IGPS complexes, is an effective tool, as shown by the good agreement with previously reported community network analysis based on mutual information on protein-correlated motions. In fact, limiting the count of

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

atomic contacts to heavy atoms already provided detection of strong effector-induced perturbations in the loop1, fα2, fα3, hα1, and Ω-loop secondary structure elements at the IGPS sideR, known to be involved in the allosteric signal propagation. Furthermore, the dynamical perturbation network analysis of heavy atom contacts also suggested previously overlooked residues fD11, fD74, hR22, and hR187 (located in the fβ1, fα2−fβ3 turn, and hα4 elements at sideR) as potential targets for future mutagenesis studies. Addition of hydrogen atoms in the computation of atomic contacts increases the complexity of the perturbation network, whose analysis has been separated in contributions from the backbone and the side chains atoms. The backbone network analysis, while sharing some features with the perturbation network analysis of heavy atoms contacts, highlighted some unknown allosteric perturbations, including the partial folding of the fβ8−fα8′ turn in the effector binding site and the partial unfolding of the hα2 helix in the proximity of the active site. Remarkably, restriction to the backbone atoms (including hydrogens) demonstrated how such network analysis provides rapid detection of folding/unfolding events induced by the effector binding that only time-consuming and tedious comparative analysis of MD trajectories can accomplish. However, the perturbation network analysis restricted to side chains contacts retrieved the structurally most important and highly conserved residues whose interactions are perturbed by the effector binding. Overall, by providing good agreement with previous theoretical and experimental studies and by recognition of new potential allosteric spots in the IGPS enzyme, the dynamical perturbation network analysis proved to be a powerful computational tool, complementary to other effective network-based methodologies for the characterization of allosteric pathways.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.9b01294.

Protein contact network modeling, Additional results including hydrogen bonds and ionic interactions modifications, analysis of pair contacts types and significant contact perturbations in the ammonia gate (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*Ivan Rivalta. E-mail: i.rivalta@unibo.it. Phone: +39 051 20 9 3617;.

*Claire Lesieur. E-mail: claire.lesieur@ens-lyon.fr. Phone: +33 (0) 4 26 23 38 06.

ORCID

Victor S. Batista: 0000-0002-3262-1237

Ivan Rivalta: 0000-0002-1208-602X

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Biography

Ivan Rivalta, born in Catanzaro, Italy, received his *Laurea* and Ph.D. in Chemistry at the Università della Calabria, Italy. After a visiting postdoctoral fellowship at the Department of Chemistry and Applied Biosciences of ETH Zurich (Lugano, Switzerland), he was Associate Research Scientist first at the Chemistry Department of Yale University (USA) and subsequently at the Dipartimento di Chimica "G. Ciamician" of the Alma Mater Studiorum, Università di Bologna, Italy. In 2014, he was appointed as CNRS permanent Researcher in the Laboratoire de Chimie UMR-5182 at the École Normale Supérieure de Lyon, France. Since 2018, he is Associate Professor of Physical Chemistry at the Dipartimento di Chimica Industriale "Toso Montanari" of the Alma Mater Studiorum, Università di Bologna, Italy. His research concerns the development and application of theoretical methods and computational techniques for the study of chemical and photochemical phenomena, with a focus on biological and biomimetic systems.

■ ACKNOWLEDGMENTS

The authors acknowledge the support of the Institut Rhônalpin des systèmes complexes, IXXI-ENS-Lyon, Lyon, France. V.S.B. acknowledges support from the NIH grant GM106121 and supercomputer resources from NERSC. I.R. acknowledges the use of HPC resources of the "Pôle Scientifique de Modélisation Numérique" (PSMN) at the École Normale Supérieure de Lyon, France.

■ REFERENCES

(1) Changeux, J. P. 50 Years of Allosteric Interactions: The Twists and Turns of the Models. *Nat. Rev. Mol. Cell Biol.* **2013**, *14*, 819−829.

(2) Fenton, A. W. Allostery: An Illustrated Definition for the 'Second Secret of Life'. *Trends Biochem. Sci.* **2008**, *33*, 420−425.

(3) Laskowski, R. A.; Gerick, F.; Thornton, J. M. The Structural Basis of Allosteric Regulation in Proteins. *FEBS Lett.* **2009**, *583*, 1692−1698.

(4) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The Ensemble Nature of Allostery. *Nature* **2014**, *508*, 331−339.

(5) Monod, J.; Wyman, J.; Changeux, J. P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* **1965**, *12*, 88−118.

(6) Koshland, D. E.; Nemethy, G.; Filmer, D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry* **1966**, *5*, 365−368.

(7) Hilser, V. J.; Wrabl, J. O.; Motlagh, H. N. Structural and Energetic Basis of Allostery. *Annu. Rev. Biophys.* **2012**, *41*, 585−609.

(8) Tsai, C. J.; Nussinov, R. A Unified View of "How Allostery Works". *PLoS Comput. Biol.* **2014**, *10*, e1003394.

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

(9) Christopoulos, A. Allosteric Binding Sites on Cell-Surface Receptors: Novel Targets for Drug Discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 198−210.
(10) Wootten, D.; Christopoulos, A.; Sexton, P. M. Emerging Paradigms in Gpcr Allostery: Implications for Drug Discovery. *Nat. Rev. Drug Discovery* **2013**, *12*, 630−644.
(11) Taly, A.; Corringer, P. J.; Guedin, D.; Lestage, P.; Changeux, J. P. Nicotinic Receptors: Allosteric Transitions and Therapeutic Targets in the Nervous System. *Nat. Rev. Drug Discovery* **2009**, *8*, 733−750.
(12) Gohara, D. W.; Di Cera, E. Allostery in Trypsin-Like Proteases Suggests New Therapeutic Strategies. *Trends Biotechnol.* **2011**, *29*, 577−585.
(13) Nussinov, R.; Tsai, C. J. Allostery in Disease and in Drug Discovery. *Cell* **2013**, *153*, 293−305.
(14) Makhlynets, O. V.; Raymond, E. A.; Korendovych, I. V. Design of Allosterically Regulated Protein Catalysts. *Biochemistry* **2015**, *54*, 1444−1456.
(15) Lisi, G. P.; Manley, G. A.; Hendrickson, H.; Rivalta, I.; Batista, V. S.; Loria, J. P. Dissecting Dynamic Allosteric Pathways Using Chemically Related Small-Molecule Activators. *Structure* **2016**, *24*, 1155−1166.
(16) Kar, G.; Keskin, O.; Gursoy, A.; Nussinov, R. Allostery and Population Shift in Drug Discovery. *Curr. Opin. Pharmacol.* **2010**, *10*, 715−722.
(17) Süel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins. *Nat. Struct. Biol.* **2003**, *10*, 59−69.
(18) Amaro, R. E.; Sethi, A.; Myers, R. S.; Davisson, V. J.; Luthey-Schulten, Z. A. A Network of Conserved Interactions Regulates the Allosteric Signal in a Glutamine Amidotransferase. *Biochemistry* **2007**, *46*, 2156−2173.
(19) Bruschweiler, S.; Schanda, P.; Kloiber, K.; Brutscher, B.; Kontaxis, G.; Konrat, R.; Tollinger, M. Direct Observation of the Dynamic Process Underlying Allosteric Signal Transmission. *J. Am. Chem. Soc.* **2009**, *131*, 3063−3068.
(20) del Sol, A.; Tsai, C. J.; Ma, B.; Nussinov, R. The Origin of Allosteric Functional Modulation: Multiple Pre-Existing Pathways. *Structure* **2009**, *17*, 1042−1050.
(21) Feher, V. A.; Durrant, J. D.; Van Wart, A. T.; Amaro, R. E. Computational Approaches to Mapping Allosteric Pathways. *Curr. Opin. Struct. Biol.* **2014**, *25*, 98−103.
(22) Martin, N. E.; Malik, S.; Calimet, N.; Changeux, J. P.; Cecchini, M. Un-Gating and Allosteric Modulation of a Pentameric Ligand-Gated Ion Channel Captured by Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005784.
(23) Markwick, P. R.; McCammon, J. A. Studying Functional Dynamics in Bio-Molecules Using Accelerated Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 20053−20065.
(24) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59*, 4035−4061.
(25) Rivalta, I.; Sultan, M. M.; Lee, N. S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1428−1436.
(26) Ricci, C. G.; Silveira, R. L.; Rivalta, I.; Batista, V. S.; Skaf, M. S. Allosteric Pathways in the Pparγ-Rxrα Nuclear Receptor Complex. *Sci. Rep.* **2016**, *6*, 19940.
(27) Palermo, G.; Ricci, C. G.; Fernando, A.; Basak, R.; Jinek, M.; Rivalta, I.; Batista, V. S.; McCammon, J. A. Protospacer Adjacent Motif-Induced Allostery Activates Crispr-Cas9. *J. Am. Chem. Soc.* **2017**, *139*, 16028−16031.
(28) Negre, C. F. A.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector Centrality for Characterization of Protein Allosteric Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E12201−E12208.
(29) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in Trna: Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6620−6625.
(30) Gasper, P. M.; Fuglestad, B.; Komives, E. A.; Markwick, P. R.; McCammon, J. A. Allosteric Networks in Thrombin Distinguish Procoagulant Vs. Anticoagulant Activities. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 21216−21222.
(31) Blacklock, K.; Verkhivker, G. M. Computational Modeling of Allosteric Regulation in the Hsp90 Chaperones: A Statistical Ensemble Analysis of Protein Structure Networks and Allosteric Communications. *PLoS Comput. Biol.* **2014**, *10*, e1003679.
(32) Stolzenberg, S.; Michino, M.; LeVine, M. V.; Weinstein, H.; Shi, L. Computational Approaches to Detect Allosteric Pathways in Transmembrane Molecular Machines. *Biochim. Biophys. Acta, Biomembr.* **2016**, *1858*, 1652−1662.
(33) Wagner, J. R.; Lee, C. T.; Durrant, J. D.; Malmstrom, R. D.; Feher, V. A.; Amaro, R. E. Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chem. Rev.* **2016**, *116*, 6370−6390.
(34) Rivalta, I.; Lisi, G. P.; Snoeberger, N. S.; Manley, G.; Loria, J. P.; Batista, V. S. Allosteric Communication Disrupted by a Small Molecule Binding to the Imidazole Glycerol Phosphate Synthase Protein-Protein Interface. *Biochemistry* **2016**, *55*, 6484−6494.
(35) Manley, G.; Rivalta, I.; Loria, J. P. Solution Nmr and Computational Methods for Understanding Protein Allostery. *J. Phys. Chem. B* **2013**, *117*, 3063−3073.
(36) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V. J.; Smith, J. L. Crystal Structure of Imidazole Glycerol Phosphate Synthase: A Tunnel through a (Beta/Alpha)(8) Barrel Joins Two Active Sites. *Structure* **2001**, *9*, 987−997.
(37) Breitbach, K.; Kohler, J.; Steinmetz, I. Induction of Protective Immunity against *Burkholderia Pseudomallei* Using Attenuated Mutants with Defects in the Intracellular Life Cycle. *Trans. R. Soc. Trop. Med. Hyg.* **2008**, *102* (Suppl 1), S89−94.
(38) Gomez, M. J.; Neyfakh, A. A. Genes Involved in Intrinsic Antibiotic Resistance of Acinetobacter Baylyi. *Antimicrob. Agents Chemother.* **2006**, *50*, 3562−3567.
(39) Myers, R. S.; Jensen, J. R.; Deras, I. L.; Smith, J. L.; Davisson, V. J. Substrate-Induced Changes in the Ammonia Channel for Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2003**, *42*, 7013−7022.
(40) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. Altering the Allosteric Pathway in Igps Suppresses Millisecond Motions and Catalytic Activity. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E3414−E3423.
(41) Vuillon, L.; Lesieur, C. From Local to Global Changes in Proteins: A Network View. *Curr. Opin. Struct. Biol.* **2015**, *31*, 1−8.
(42) Dorantes-Gilardi, R.; Bourgeat, L.; Pacini, L.; Vuillon, L.; Lesieur, C. In Proteins, the Structural Responses of a Position to Mutation Rely on the Goldilocks Principle: Not Too Many Links, Not Too Few. *Phys. Chem. Chem. Phys.* **2018**, *20*, 25399−25410.
(43) Achoch, M.; Dorantes-Gilardi, R.; Wymant, C.; Feverati, G.; Salamatian, K.; Vuillon, L.; Lesieur, C. Protein Structural Robustness to Mutations: An in Silico Investigation. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13770−13780.
(44) Doshi, U.; Holliday, M. J.; Eisenmesser, E. Z.; Hamelberg, D. Dynamical Network of Residue-Residue Contacts Reveals Coupled Allosteric Effects in Recognition, Catalysis, and Mutation. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 4735−4740.
(45) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(46) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(47) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(48) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N-Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

(49) Grubmuller, H.; Heller, H.; Windemuth, H.; Schulten, K. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-Range Interactions. *Mol. Simul.* **1991**, *6*, 121−142.
(50) Schlick, T.; Skeel, R. D.; Brunger, A. T.; Kale, L. V.; Board, J. A.; Hermans, J.; Schulten, K. Algorithmic Challenges in Computational Molecular Biophysics. *J. Comput. Phys.* **1999**, *151*, 9−48.
(51) Douangamath, A.; Walker, M.; Beismann-Driemeyer, S.; Vega-Fernandez, M. C.; Sterner, R.; Wilmanns, M. Structural Evidence for Ammonia Tunneling across the (Beta/Alpha)(8) Barrel of the Imidazole Glycerol Phosphate Synthase Bienzyme Complex. *Structure* **2002**, *10*, 185−193.
(52) Lipchock, J. M.; Loria, J. P. Nanometer Propagation of Millisecond Motions in V-Type Allostery. *Structure* **2010**, *18*, 1596−1607.

DOI: 10.1021/acs.jpcb.9b01294
J. Phys. Chem. B 2019, 123, 3452−3461

2.2 Connected Component Analysis of Dynamical Perturbation Contact Networks

2.2.1 Clustering edge weights

A key challenge in the study of DPCNs is to systematically extract relevant information from the network. In the DPCN, the network topology is vastly different from individual AANs; in particular, there are positive and negative edges, and most of the edges have a weight close to zero, while only a few *outlier* edges contain relevant information. In fact, here, the issue is a general feature selection problem. Our regular methodology uses a *threshold* based on the absolute value of the edge weights to remove low-value edges, but an appropriate threshold value is heavily dependent on the *selection* used in the contact condition (3 for the backbone, 5-6 for the heavy atom, 25 for the all atom), and we could not provide a systematic procedure to select this value. Moreover, the contact value is difficult to grasp intuitively by comparison with the 10% threshold used in frequency contact networks. The distribution of edges in the apo and holo AANs and DPCN using a 5 Å cutoff and heavy-atom selection is shown in Figure 2.5. The distribution of weights in the AAN of apo and holo is astonishingly similar, suggesting that this curve may be independent of the protein conformation. Interestingly, it has some irregular slope changes. Some studies have been devoted to the precise study of these slope changes, notably that they related with the nature of interactions captured in AANs, however they remain outside the scope of our problem, which focuses specifically in DPCN. In DPCN, most of the edge weights are clustered around zero, and we only want to extract the *outlier* edges with high values. An empirical way to select such a *threshold* is then to select the "the knee in the curve" on both sides of the DPCN curve. However, this approach is not systematical and subjective, since there is no precise definition of "knee in the curve".

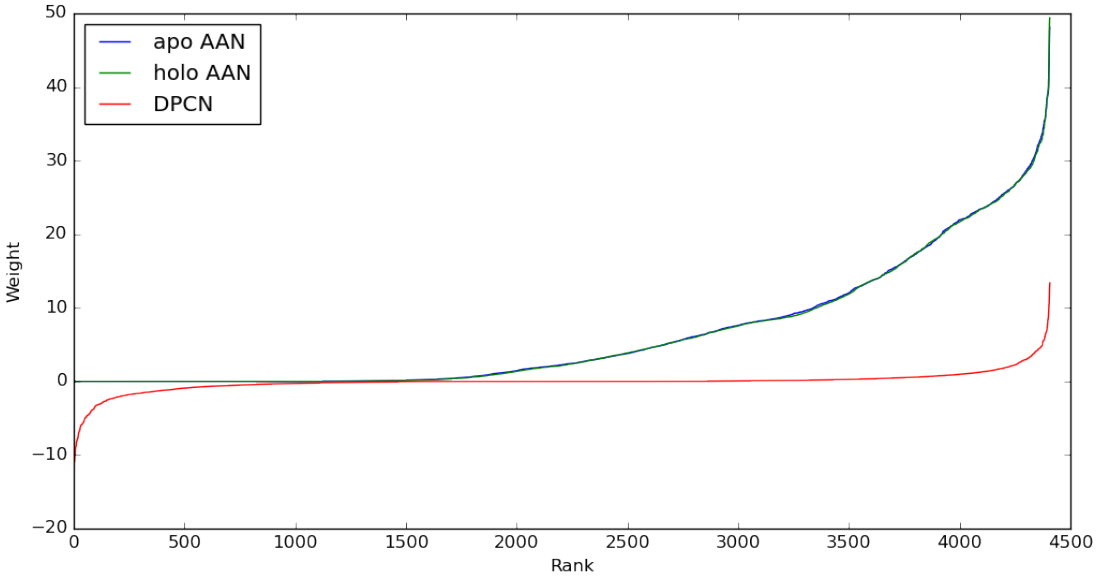


Figure 2.5: Ranked edge weights in increasing order for apo and holo AAN and the DPCN

One approach we then used is to cluster weights with machine learning techniques so that edges of similar importance are grouped together. There are many clustering techniques, but for practical purpose, we focused on clustering techniques implemented in sklearn[1]. To select the most appropriate clustering technique, we had two main criteria: first, all clustering techniques require some general parameters: the so-called *hyperparameters*; we wanted to find a technique which is the least dependent on hyperparameter choice so that clustering does not require more tuning than selecting a threshold. Second, the solution must be largely scalable with the number of samples, so that using it on large proteins would not substantially increase the computation time. For this, three algorithms were very promising: Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[2], Ordering Points To Identify the Clustering Structure (OPTICS)[3] and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) Clustering[4]. Notably, these three techniques do not ask for the number of cluster as hyperparameter and are good at detecting *outliers*[4, 5, 6].

In total, the DPCN built between simulations apo1-4 and prfar1-4 with a 5 Å *cutoff* and heavy-atom *selection* contains 4,088 edges. In Figure 2.6 we report the clustering of those weights using ten techniques implemented in sklearn with their standard hyperparameters. While most algorithms performs in a fraction of a second, Affinity Propagation[7] takes more than 100 seconds and actually fails to converge. Therefore, we firmly excluded this clustering technique. Spectral clustering and OPTICS also took more than a second. This suggests that issues may arise using these techniques with larger systems.

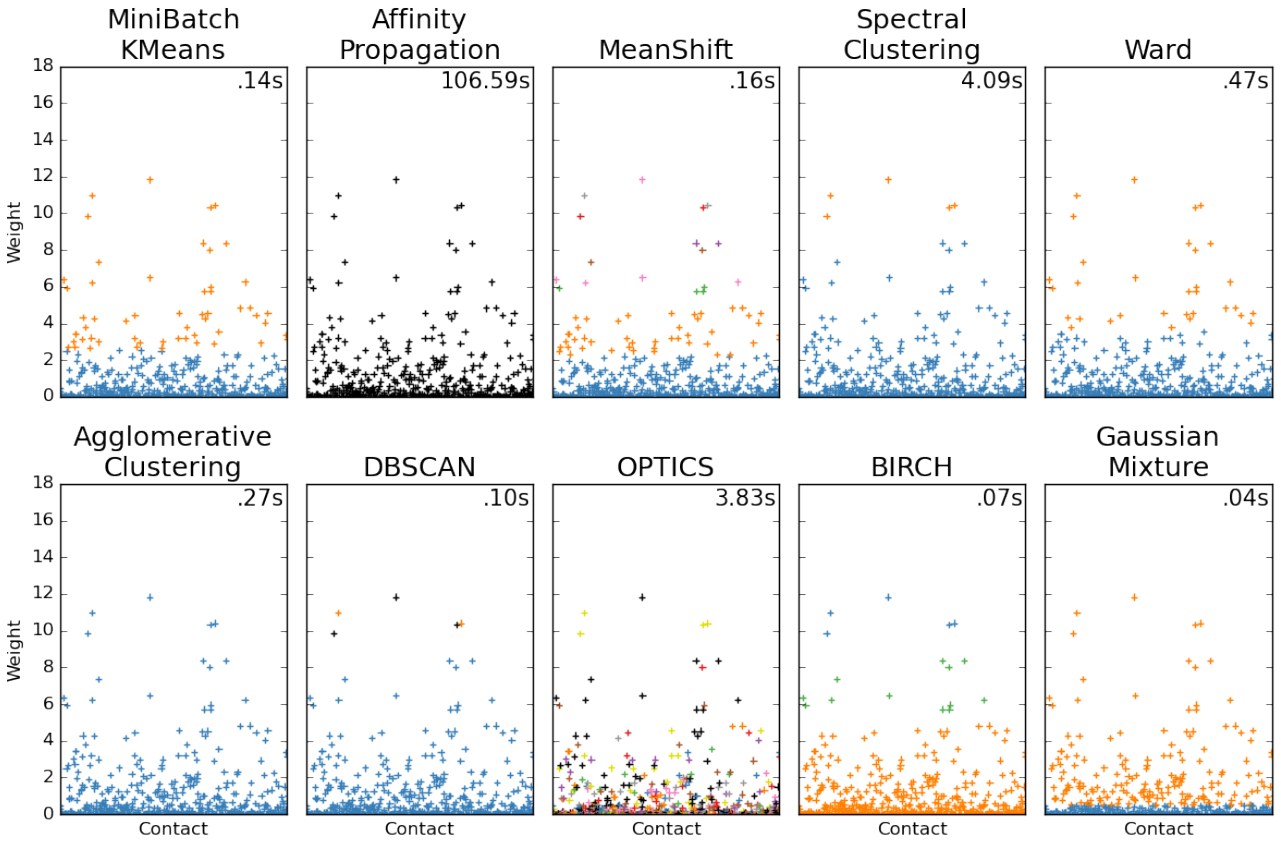


Figure 2.6: Clustering techniques with different colors indicating the clusters in which each data point belong. The time taken for each clustering is written in each plot. For algorithms asking specifically for a number of clusters, we chose 2 clusters in each case and for algorithms asking for a number of neighbors, we chose 10 neighbors in each case

In our previous works, we empirically found that a *threshold* of 5 produces a cleaned version of the graph. Therefore, a clustering algorithm that separates clusters close to this range is the most suitable. Kmeans[8], Spectral Clustering, Ward, Agglomerative Clustering, DBSCAN, OPTICS, and Gaussian Mixture all fail in doing so. This might be because the standard hyperparameters are not a good fit for our problem. Still, an important aspect of our approach is that we want as little influence of the hyperparameter as possible, thus not having to optimize the hyperparameters too much. Because of this, we discarded all these algorithms. On the contrary, the MeanShift algorithm[9] performs particularly well. One cluster separates weight between 0 and 2.5, the "very low" contact changes. Another cluster separates weights between 2.5 and around 5: the "low" contact changes. Finally, many clusters are created for the data above. This clustering is rather interesting because it fits nicely with our earlier analysis and is rather fast (0.16s). However, despite the fact that this algorithm is generally recognized as one of the best *unsupervised* clustering techniques, it scales terribly with the sample number[10] and we discarded it from the possibilities.

Of all the algorithms, BIRCH clustering really stands out. In theory, it was one of the best fit for an algorithm, since its main purpose is outlier removal and data reduction. Here, a single cluster removes all edges below a weight of 5, consistent with our previous analysis, and then two clusters separate the above data. Furthermore, it takes only 0.07 second and is very scalable with the number of samples. BIRCH clustering also does not ask for a specific number of clusters in hyperparameters. For all these reasons, we focus our analysis on BIRCH clustering.

2.2.2 Limitations of BIRCH clustering

While BIRCH clustering was particularly successful in detecting groups of importance in terms of weight. However, the precise meaning of these groups of importance is elusive. In fact, within a single group, there are contacts of different types (salt bridge, polar, hydrophobic, etc.) that are localized throughout the whole protein. Moreover, edges belonging to the same localized perturbation can be grouped in totally different groups. Although BIRCH clustering provides a way to select a threshold arbitrarily less, this study proved that the mere use of *threshold* to filter edges cannot provide a consistent network view (in terms of chemistry or geometry). In fact, clustering using only edge weight values does not take into account the network topology,

which is a central element in the study of DPCN.

2.2.3 Connected Component Analysis

In most cases, the AAN of a protein is *connected*, that is, we can construct a path between each pair of amino acids in the graph. A trivial example of a nonconnected AAN is a protein with two separate unbound chains. When successively higher threshold values are applied on the network, this connectivity can be lost because some connections are lost. In this case, we call the *connected components* (CCs) of the graph, the maximum size subgraphs that are *connected*. To account for all possible edge values, we can simply incrementally remove the edge with the lowest value. This process is explained in Figure 3 in Manuscript 1. A CC Analysis (CCA) investigates how quantities related to CCs evolve with successive removal of the edge with the lowest value. The CCs of a graph have an intrinsically local aspect because the path between a node and its neighbors is usually short and thus more resilient to edge remove. Some approaches focus on studying the number of nodes or edges of the graph's largest components, but here we focus on maximizing the number of CCs.

When removing an edge, two concurrent effects are at play: either an edge is the last remaining link between two subcomponents, thus removing it creates a new component, or it was the last edge of a given component, thus removing it destroys a component. Of course, the process of removing an edge can simply have no effect on the number of CCs. At the beginning of the process, there is generally only one CC, while at the end of the process there is always zero since the graph is *empty*. In practice, the component creation effect occurs before the component destruction effects, and then the number of CCs evolves in a bell curve during the edge removal process. Therefore, the weight which maximizes the CCs number is interesting because it shows precisely where the component destruction effect dominates the component creation one. We then hypothesized that this number can be selected as threshold (for a 5Å heavy-atom based DPCN, we found a maximum number of components attained at threshold 4.45, which is close to our empirical 5). In our exampled notably, several thresholds produced a maximum number of CCs as the bell curve had a plateau at its top, and we used the one that best represents where the component destruction effect dominates, that is, the one with the highest weight.

In the end, by selecting a threshold of 4.45 suggested by the CCA, this process produced a graph with 36 components, which remains much to analyze. Actually, almost half of the components consisted of a single edge, which is consistent with the fact that there is a significant decline in number of components after this point. In order to study perturbations that spread within the protein, in the first instance, we simply removed the components with a single edge. This still produced around 20 components, not entirely satisfactory. Using the same argument to study *local-to-global* perturbations, we used another metric to discriminate the CCs, which is their *diameter*. The diameter of a graph represents the maximum *eccentricity*, (i.e. the maximum length of all the shortest paths in the component). When only CCs with diameters > 3 were selected, this reduced the selection to nine CCs. Of the nine components, eight of them could be directly attributed to the allosteric pathways previously recognized for our system[11, 12, 13]. This analysis can, thus, very efficiently automatize the DPCN analysis and this, whether the contact conditions. In fact, our results also show that the procedure can also be applied to various types of Perturbation Networks such as perturbation frequency contact networks and perturbation cross-correlation networks.

References

[1] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *J. machine Learn. research* 12 (2011), pp. 2825–2830.

[2] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[3] Mihael Ankerst et al. "OPTICS: Ordering points to identify the clustering structure". In: *ACM Sigmod record* 28.2 (1999), pp. 49–60.

[4] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases". In: *ACM sigmod record* 25.2 (1996), pp. 103–114.

[5] Muhammad Fazal Ijaz et al. "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest". In: *Appl. Sci.* 8.8 (2018), p. 1325.

[6] Ching-Heng Lin et al. "Applying density-based outlier identifications using multiple datasets for validation of stroke clinical outcomes". In: *Int. journal medical informatics* 132 (2019), p. 103988.

[7] Brendan J Frey and Delbert Dueck. "Clustering by passing messages between data points". In: *science* 315.5814 (2007), pp. 972–976.

[8] David Sculley. "Web-scale k-means clustering". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 1177–1178.

[9] Yizong Cheng. “Mean shift, mode seeking, and clustering”. In: *IEEE transactions on pattern analysis machine intelligence* 17.8 (1995), pp. 790–799.

[10] Stefan Craciun et al. “A scalable RC architecture for mean-shift clustering”. In: *2013 IEEE 24th International Conference on Application-Specific Systems, Architectures and Processors*. IEEE. 2013, pp. 370–374.

[11] Ivan Rivalta et al. “Allosteric pathways in imidazole glycerol phosphate synthase”. In: *Proc. National Acad. Sci.* 109.22 (2012), E1428–E1436.

[12] Christian FA Negre et al. “Eigenvector centrality for characterization of protein allosteric pathways”. In: *Proc. National Acad. Sci.* 115.52 (2018), E12201–E12208.

[13] Aria Gheeraert et al. “Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks”. In: *The J. Phys. Chem. B* 123.16 (2019), pp. 3452–3461.

2.2.4 Manuscript 1

This work led to the redaction of a manuscript for the Journal of Chemical Theory and Computation.

Connected Component Analysis of Dynamical Perturbation Contact Network

Aria Gheeraert,[†,‡] Laurent Vuillon,[*,†] and Ivan Rivalta[*,‡,¶]

†*LAMA, Université Savoie Mont-Blanc, CNRS, Bourget-du-Lac, France*

‡*Dipartimento di Chimica Industriale "Toso Montanari", Università di Bologna, Viale Risorgimento 4, I-40136 Bologna, Italy*

¶*Univ Lyon, Ens de Lyon, CNRS UMR 5182, Université Claude Bernard Lyon 1 Laboratoire de Chimie, F69342, Lyon, France*

E-mail: laurent.vuillon@univ-savoie.fr; i.rivalta@unibo.it

Phone: +33 4 79 75 87 33; +39 051 209 3617

Abstract

Introduction

The exponential improvement of technologies currently grants the opportunity to perform classical molecular dynamics (MD) simulations of systems with quite large size and time scale.[1–5] Analyzing such long and sizable simulations thus becomes increasingly challenging with dynamical network approaches emerging about a dozen years ago as valuables tools.[6–13] This approach was used to understand the way atoms in protein arrange themselves[14] and to investigate allosteric signaling.[6,8,13,15,16] The usage of network theory on static protein systems (i.e. crystal structures) is more than 30 years old[17] and since then several types of

network have been defined.[18] The shift from static to dynamical network was first successfully implemented to cross-correlation network[8,19] Another widely used network approach is that considering atomic contacts between amino acid residues, namely contact networks, which is generally used in a "static" way, i.e. by analyzing crystal structures of multiple protein types/families[20–25] or, in a more strict comparative fashion, by monitoring contact perturbations induced by mutations,[26,27] namely perturbation contact networks. This latter methodology has proven to be an efficient tool to analyze dynamical protein networks, i.e. time-averaged graphs associated to MD trajectories, where perturbations could be due to mutations[28] or to effector binding, thus associated with allosteric signaling.[29] Dynamical perturbation contact networks of allosteric signals have proven to be particularly powerful at pointing out local differences in conformation during the allosteric regulation of the imidazoleglycerol phosphate synthase (IGPS) enzyme from *Thermotoga maritima*, using an *reference* set of MD simulations (modeling the apoenzyme) and a *perturbed* one (modeling the holoenzyme). Remarkably, this analysis captured allosteric pathways previously described in literature[8,30] and added even more information about local contact changes between two sets of MD simulations. IGPS is an archetype allosteric enzyme participating in fundamental biochemical pathways that is lacking in all mammals but present in fungi, plants and bacterias. Hence, IGPS is a target for safe antipathogens development.[31–33] Its place is during the fifth step of histidine synthesis and is composed of two subunits: a glutamineamidotransferase (GATase) HisH which catalyzes the hydrolysis of glutamine into glutamate and ammonia and a cyclase HisF where the effector PRFAR, (N'-[(5'-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide) binds. The ammonia released at the GATase active site then tunnels through the protein, crossing an ion gate, approaches PRFAR at the effector site and generate a cyclization releasing Imidazoleglycerol Phosphate (ImGP), a precursor to histidine and 5'-(5-aminoimidazole-4-carboxamide) (AlCAR) later used in the synthesis of purines. Upon effector binding, the affinity of glutamine slightly increase (5-fold) while the catalytic activity increases by 3 orders of magnitude (1,000-fold) thus making IGPS a V-type

allosteric enzyme.[34] Here, the aim of the allosteric regulation is to moderate the hydrolysis of glutamine to only make it happen when PRFAR is also there to react. One of the impressive feature of IGPS is that active and effector sites are at a distance of about 25 Å. In our series of computational studies of IGPS allostery,[8,30,35] we predicted that the allosteric propagation mechanism involves a collection of both short- and long-range displacements, i.e. a set of local (hydrophobic, salt-bridges and H-bond) interactions that increase inter-residue motion correlations on one side of the protein (sideR)[8,30] and a slow collective motion that alters the HisF/HisH interface, namely the breathing motion.[30] The theoretical predictions have been subsequently used to design allosteric drugs[35] and IGPS mutants[36] that alter the IGPS allosteric pathways, resulting in inactive enzymes. Very recently, both short- and long-range predicted effects have been demonstrated experimentally by X-ray structural characterization of active IGPS ternary complexes[37] and light-switching activation,[38] respectively. The whole allosteric mechanism is summarized in Figure S1.

In this context, our dynamical perturbation contact network analysis of IGPS allostery has been crucial to discover the role of specific secondary structure elements ($f\beta1$, $f\alpha8$-$f\beta8$ turn, $h\alpha2$ and $h\alpha4$) and key allosteric contacts (fD11, fD74, fR22, and fR187). However, this analysis required, first, a brute-force approach (i.e. the weight threshold) in order to reduce the number of perturbed contacts for an eye-friendly visualization and, then, a biased selection of the most relevant perturbations based on the previous knowledge of the IGPS allosteric pathways, e.g. focusing on contact changes at sideR.

Therefore, since dynamical perturbation contact network proved to be an adequate tool to enlarge the comprehension of allostery, it is of fundamental importance to develop unbiased approaches without parameters with completely arbitrary selection, in order to make this method of more general use. For instance, the choice of a weight threshold as well as that of the atomic contact types are not obvious tasks (usually requiring many attempts) and should be avoided. Here, we propose two alternative methodologies to overcome these limitations of the dynamical perturbation contact network analysis. The first uses clustering,

an unsupervised learning technique that groups data points together into sets, namely Birch clustering.[39] This method established itself as a reliable way to partition large datasets, with the advantage of avoiding arbitrary parameters (e.g. number of clusters) and of good scalability (that can be of help for studying very large protein complexes). The second one involves a connected component analysis[40–43] that can partition the contact network by grouping connected nodes, similar to what has been successfully done for energy-weighted protein graphs,[44] with the advantage of providing information on the local propagation of the perturbations. Here, we present how the usage of clustering and the connected component analysis of dynamical perturbation contact graphs could provide generalized network analysis of MD simulations, by showing their applications to the allosteric pathways of IGPS (from *T. maritima*), a excellent test-case with well-known allosteric features.

Materials and Methods

Aiming to achieve a comparison with ref. 29 which introduced dynamic perturbation contact network on IGPS, we used the same structural models of apo and PRFAR-bound IGPS complex that are described in ref[8] in order to adequately assess the impact of the extension presented here. In this previous analyzes it was shown that 100 ns of simulation were enough to acquire allosteric effects in all the simulations. Accordingly here, we used these previous MD trajectories, comprised of four replica simulations of 100 ns for IGPS apoenzyme and four replica simulations of 100 ns for the holoenzyme (PRFAR-bound).[8] MD simulations of the IGPS complexes used the AMBER-ff99SB[45] force field for the IGPS protein and the generalized Amber force field[46] for the PRFAR ligand. Computations were run using the NAMD2 software package.[47] A pre-equilibration procedure was performed on both systems including addition of hydrogen atoms and explicit TIP3[48] water solvent molecules approximately (22,500) optimization constraining the rest of the atoms at the crystal structure positions, a slow heating to 303K, gradual release of atomic positions constraints, and un-

constrained MD simulations of 4 ns in the canonical NVT ensemble using Langevin dynamics. Thereafter production runs were simulated in the NPT ensemble at 303 K and 1 atm (using the Langevin piston) for 100 ns after reaching the equilibrium volume (after ca. 2-3 ns). Periodic boundary conditions and the particle mesh Ewald method[49] were employed, with van der Waals interactions calculated using a switching distance of 10 Å and a cutoff of 12 Å. A multiple time-stepping algorithm[50,51] was adopted, with bonded, short-range nonbonded, and long-range electrostatic interactions were evaluated respectively at every one, two and four time steps, using a time step of integration of 1 fs.

In general most computations were performed using the NumPy package,[52] handling of MD trajectories and topologies was done with MDTraj[53] and network theory analyzes with NetworkX.[54]

Dynamical Perturbation Contact Network

At each frame, we use the Cython[55] implementation of the KD-tree algorithm[56–58] found in scipy[59] (scipy.spatial.cKDTree) to get all the list of all atomic pairs at a distance below a cutoff of 5 Å. The atomic contact matrix A_{ij} such that $a_{ij} = 1$ if atom i and j are in contact or 0 in the opposite case is built thanks to this list. Here the cutoff value of 5 Åwas used in consistence with the previous analyzes.[8,29] The average atomic contact matrix of a set of simulations is defined by averaging each element on all the individual matrices i.e. $a_{ij,avg} = \frac{\sum_t a_{ij,t}}{n_{frames}}$. Note that in contrary to each frame atomic contact matrix which is binary, the average will frequently produces decimal numbers thus requiring floating-point arithmetic. Finally this matrix can be converted to the residue contact matrix using transformation matrices T such that $t_{ij} = 1$ if atom i is in residue j or 0 elsewhere. Note that if we want to avoid counting an atom i in the transformation matrix simply by setting all the t_i row to be equal to 0. The average residual contact matrix R can be expressed as $R = T^t AT$. This definition allow to use different transformation matrices to describe asymmetric contacts (i.e. contacts between different selections). Here when not mentioned,

the default selection used is the protein stripped from hydrogen atoms. Looking at intra-residual contacts is beyond the scope of this study, thus we set all the diagonal elements of the average residue contact matrix to be equal to 0. The average residual contact matrix is then the adjacency matrix of the contact network. The average perturbation contact matrix between an *initial* set of simulations and a *perturbed* on is here defined as the subtraction of their two average residual contact matrices. For visualization purposes we add a coloring scheme to the edges: blue if the weight is bigger in the *initial* state and red if the weight is bigger in the *final* state. The dynamical perturbation contact network is the network created from the latter adjacency matrix. All this procedure is strictly equivalent to the one described in ref 29.

Birch Clustering

To cluster the weights, we represented the unsigned list of weights in the dynamical perturbation contact network as a one-dimensional vector. It is on this vector that we used the scikit-learn[60] implementation of Birch clustering[39] with a threshold of 0.5 and a branching factor of 50. In order to bypass the arbitrary choice of a number of clusters we usually did not performed the final clustering step except if specially mentionned.

Connected component analysis

Weighted networks can be cleared of their faintest connections by removing all edges present in the network that have a weight lower than a threshold value and pruning isolated nodes. A connected component of this new graph (constructed with a given threshold) is a component C in which each pair of nodes is connected with each other via a path in the component. Thus two connected components are distrinct if there is no link between these two connected components. Connected components are found using the Breadth-First search algorithm.[61–63] Here we assess properties of connected components, namely the number of components in the graph using different threshold values. Dynamical perturbation network have the specificity

to have edges weighted by decimal numbers. Therefore, scanning the number of connected components by increasing the threshold using a weight step may neglect some relevant data points (see Fig. S3). To avoid this problem, here we compute instead the number of connected components after successive removals of the edges with lowest weight and pruning isolated nodes. This gives the exact representation of the number of connected components in function of the threshold.

Results and discussion

The entire perturbation contact network between PRFAR-bound and apo-IGPS represented in Figure 1A is quite congested graph, from which one can still extract general clues about the overall contact changes due to effector binding. In fact, a majority of contact loss upon effector binding (blue edges) is detectable between the two subunits, consistently with the alteration of the breathing motion observed in previous studies [REF]. Furthermore, another striking pattern is present at sideR of HisF at sideR and involves the effector site, loop1, $f\alpha1$, $f\alpha2$. All these secondary structure elements have been shown to be essential to the propagation of allosteric motions. Other smaller patterns of interaction are also noticeable, but the complete graph contains more than 4,000 edges and is difficult for a human to exploit.

To ensure a reliable way to have a visualisation that is comprehensible for a human eye, one can remove all edges below a certain threshold weight, thus displaying only a specific number of the top biggest edges. For instance, if one selects the top 50 edges this will correspond to a threshold weight of 6.38, as represented in Figure 1B. With this crude selection criterion, one can visualize the HisF perturbations near the effector site at sideR that propagate to loop1, $f\alpha1$ and $f\beta1$ along with other displacements at sideL. Moreover, at the HisF/HisH interface the contact changes related to the salt bridge network between $f\alpha2$, $f\alpha3$ and $h\alpha1$ is observed along with their propagation to the Ω-loop and other shifts at sideL near the hM121 residue,

which are linked to the opening of the ion gate. Finally, in HisH is evident the strongest perturbation corresponding to the partial unfolding of the $h\alpha 2$ helix upon binding and a cluster of interactions at the top of HisH, with hR78 as hub.

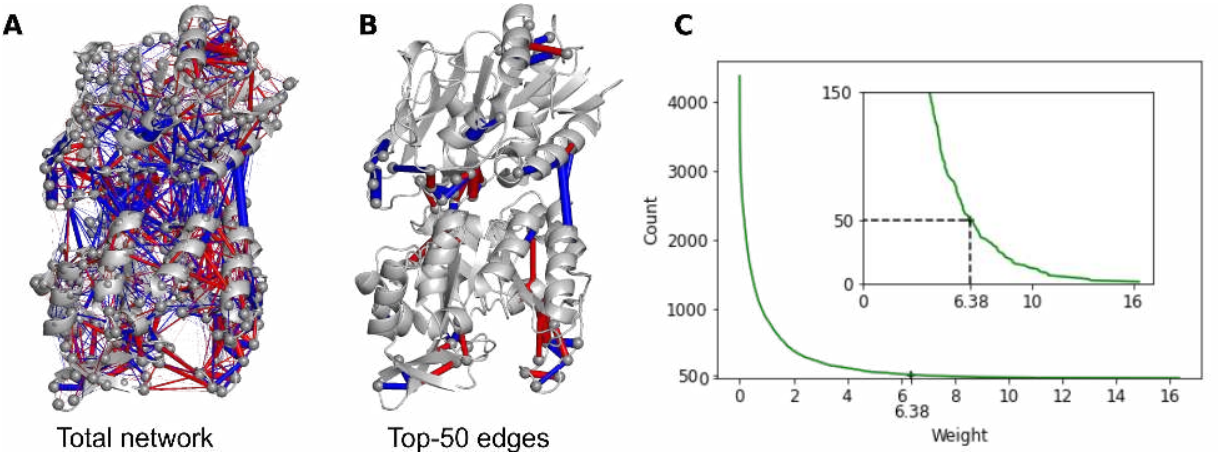


Figure 1: A. Complete Dynamical Perturbation between PRFAR-bound IGPS and apo IGPS. Blue edges represent a stronger contact in apo while red edges a stronger contact in PRFAR-bound B. With only the top-fifty biggest edges shown. C. Repartition of the edges in terms of weight showing the weight of the 50th biggest edge.

Clustering

The perturbation network edge counting decays as a function of weight as represented on Fig. 1C, featuring in the tail of the curve (i.e. at the largest weights) some sort of discretisation, with some edges grouped together (appearing as small plateaux) and separated from other groups of edges. An arbitrary choice of threshold weight is then arguable, because it will not necessarily preserve these groups. Therefore, selecting a threshold at the edge of these groups could be ideal because it would allow keeping edges with similar weights together in a separated group of relevance. The Birch clustering of edges by weight allowed us to distinguish these groups of relevance and to select those corresponding to the biggest weights. In Figure 2, the perturbation network associated with edges in the top-four group of relevance are depicted with a color for each group. The first group corresponds to the

single outlier pair hE56-hR59 involved in the unfolding of the $h\alpha 2$ helix. In the second group are found three edges: fA224-ff227, hR116-hD159, fR249-hY136. The first pair has already been attributed to a displacement of a hydrogen-bonding near the effector site,[29] marking the beginning of the allosteric pathways, while the two other interactions are related to the breathing motion (with fR249 being part of the hinge) and local contacts at sideL. The third group of relevance contains four edges: fD11-fK19 fE67-hR18, hS115-hD159, fK4-fV248. The first two are key elements of the allosteric mechanism, the third one forms a triad with the hR116-hD159 pair in group 2, while the latest involves the ion gate (i.e. fK4) and its connection to the breathing motion hinge (i.e. fV248 is adjacent to fR249) . The fourth group contains six edges (including a triad): fK4-fF214, hH120-hH141, fH228-fK19-fR27, hN12-hN15 and hY79-hS197. Among these six edges, the first one is correlated to the opening of the ion gate (via fK4) as well as the second one (as part of the cluster involving hM121), while the triad (comprising effector binding site and loop1) and the hN12-hN15 pair have been established as part of the allosteric signaling mechanism. The last perturbation of this group has not been previously highlighted since it belongs to a cluster of interaction in the top of HisH not directly linked with other relevant perturbations. Overall, the most interesting outcome of the clustering procedure providing groups of relevance is the appearance of a direct propagation from the effector site to loop1 and then to $f\beta 1$, involving residues fA224, fF227 fH228, fK19, fD11 and fR27 at the bottom of HisF (sideR). However, these connections are located in three different groups of relevance. Indeed, as expected, this analysis does not provide insights of local propagation of contact perturbations and, thus, direct information on the allosteric signaling mechanism. In fact, increasing the number of groups of relevance (e.g. more than four considered so far) will not necessarily provide better local information, while it complicates the analysis by quickly increasing the total number of pairs to be considered.

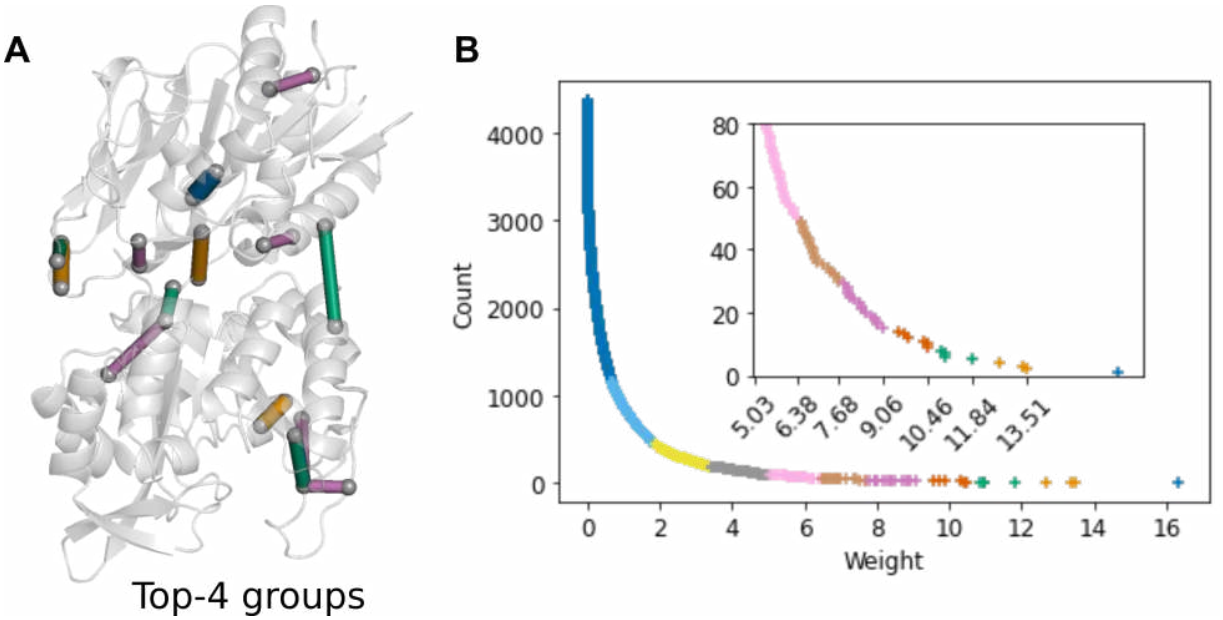


Figure 2: A. Representation of the top-four groups in term of edge weight B. Representation of the Birch clustering of edges by weight.

Connected component analysis

Considering this limitation of the cluster analysis, we performed a connected component analysis that provides a way to cluster edges by closeness, possibly granting information on the local propagation of the allosteric signals. As shown in Figure 3, when sequentially removing edges with the lowest weight, the number of connected components (after pruning isolated nodes) can either increase (if this edge is the last one connecting two components), decrease (if the edge is the last member of a component), or make a plateau (if the edge is inside a connected component with more than one edge). The complete graph of IGPS (see Fig. 1A) contains initially a single connected component, which is conserved until edges with weights <0.80 are removed. By increasing the weight threshold, the first split into two components occurs and then the number of connected components steadily increases until around 35 (at weight equal to 3.03), meaning that in this range of weights removing edges creates new components. In the range of weights between 3.03 and 4.45 , the number of connected components oscillates, initially decreasing (up to 30 until a weight of 3.39) and then increasing up to its maximum at 36 components. From weight 3.82 to 4.45 , the

components are created and destroyed approximately at the same rate, thus featuring fast and small oscillations for sequential edge cuts. The network with the maximum number of components at the largest weight (i.e. at 4.45) is considered as the graph containing its "final" components, since from this point the number of components is destined to quickly decrease after each edge cut (only occasionally it can slightly increase upon new edge cut, with total number of components obviously smaller than the maximum, i.e. <36, see Fig. 2B). At this point, in fact, edge removal will create just (pruned) isolated nodes, indicating that a graph structure where the components are interconnected by edges with large weights is reached. This structure resembles a community structure where the edges with smallest weights are removed and the corresponding nodes pruned. Indeed, as shown in Fig, 2B, there is a fast decrease in the number of connected components from 36 to 20 between weights 4.45 and 6. This corresponds to the removal of the smallest components, usually containing a single edge. At weights >6, the number of components undergoes a much slower decay (even plateauing in some ranges) until zeroing at the final weight around 16. This behavior refers to the strongest components slowly disappearing, where the strength of a given component is related to its edge with the largest weight, which then corresponds to the vanishing point of this component.

Figure 3C represents the distribution of vanishing points for the 36 final components, i.e. at threshold weight equal to 4.45 . Around 50% of these components have a vanishing point between 4.45 and 6, which belongs to the initial fast decrease in the number of components down to 20. The median value of the whole distribution of vanishing points of the final components is 6.01 that, notably, is quite close to the threshold weight corresponding for the top 50 edges (i.e. 6.38). Above 6, the distribution of vanishing points is really spread, with maximum one or two components sharing the same vanishing point, in line with the slow disappearance of components following edge removal with weights >6, as discussed above. Two interesting metrics about the components are the size, the diameter. The size

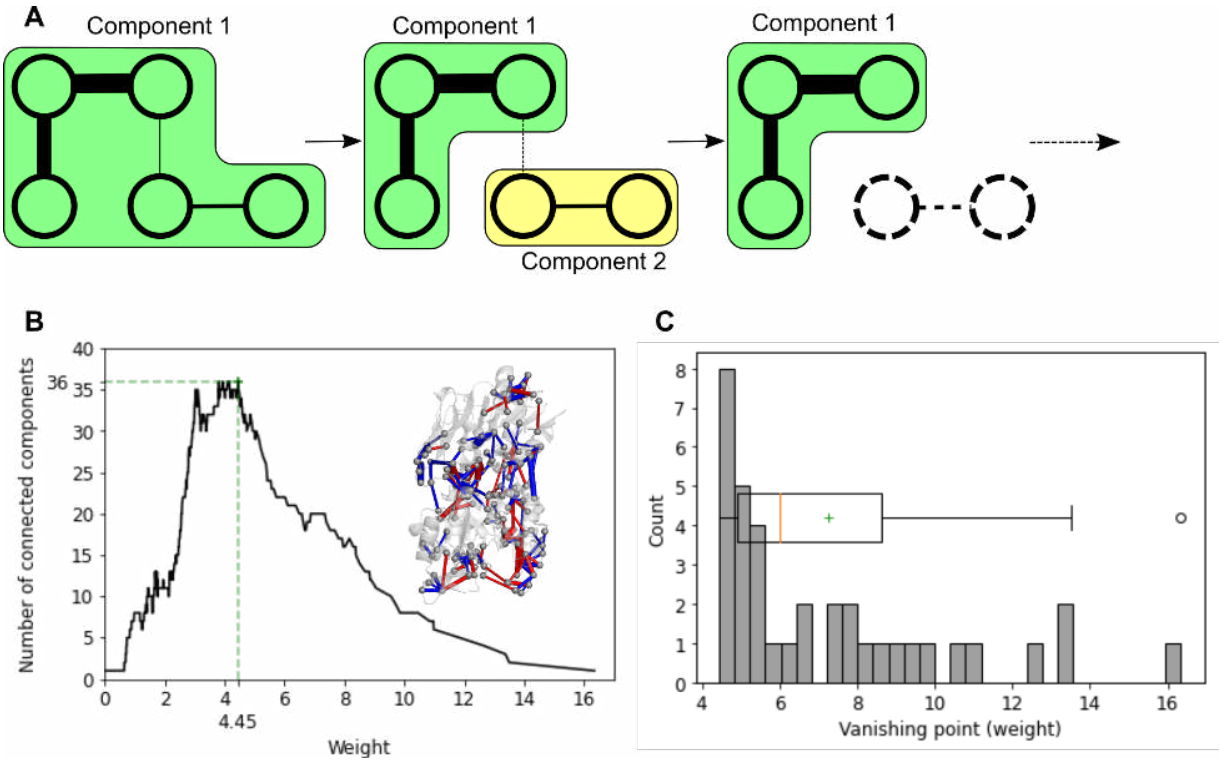


Figure 3: A. Diagram representing the count of components while successively removing edges with lowest weight B. Number of connected components in function of the lowest weight in the graph. C. Distribution of the vanishing points of each component for the network at threshold 4.45 corresponding to the final components.

is the number of edge, the diameter the is the greatest distance between any pair of nodes in the component. In the case of the perturbation contact network analyzed here, the sizes of the final components indicate how much a local perturbation has an influence on other amino acids whatever their position. On the other hand, the diameter evaluates how much a perturbation can spread to different parts of the protein. Notably, we observed an interesting trend of the diameters of the final components and their vanishing points, as depicted in Fig. 4A. Most of the final components are comprise of two vertices and one edge so they have a diameter of 1 but also a good portion of them have a diameter of 2 and represent the most trivial example of propagation. We thus chose to only look at the nine major components of diameter > 2 represented on Figure 4. They involve specific secondary structures that are reported in Table 1 alon with the metrics associated to each component. Among those components, two relates to the alteration of motion in loop1 associated with its closing on PRFAR (components attributed 1 and 9), three relates to the alteration of breathing motion observed at sideL (components attributed 2, 3 and 6) while two relates to the alteration of breathing motion at sideR coupled with rearrangements at the surface of HisF (components attributed 4 and 8). Component 5 with the biggest vanishing point reveals the formation of the oxyanion hole. Only component 7 cannot be attributed to known allosteric effects but it contains the N-terminus part of HisF which was not resolved in the crystal structure used to perform MD simulations and is thus prone to thermal fluctuations.

Table 1: Secondary structure elements of the seven major components

Attr.	**d**	**v.p.**	**Size**	**Secondary structure elements**
1	7	13.38	23	loop1, fα1, fβ1, fβ5-fα5, fβ8-fα8
2	6	10.46	15	fα3-fβ4, fα4-fβ5, fα6-fβ7, fβ4, fβ6, hβ6-hβ7, hβ7, hβ8', hβ9
3	4	13.51	6	hβ10, hβ6-hβ7, hβ8, hβ9-hβ10
4	4	10.97	9	fα2, fα2-fβ3, hα1, hα4
5	3	16.34	6	oxyanion strand, hα2, Ω-loop
6	3	12.68	9	fC-term, fN-term, fα7, fα8, hβ8'
7	3	9.72	6	hN-term, hα4, hβ10-hβ11, hβ11, hβ4
8	3	9.56	4	fα2, fα3, fα3-fβ4, hα1, Ω-loop
9	3	5.39	4	loop1, fβ6-fα6, fβ7-fα7

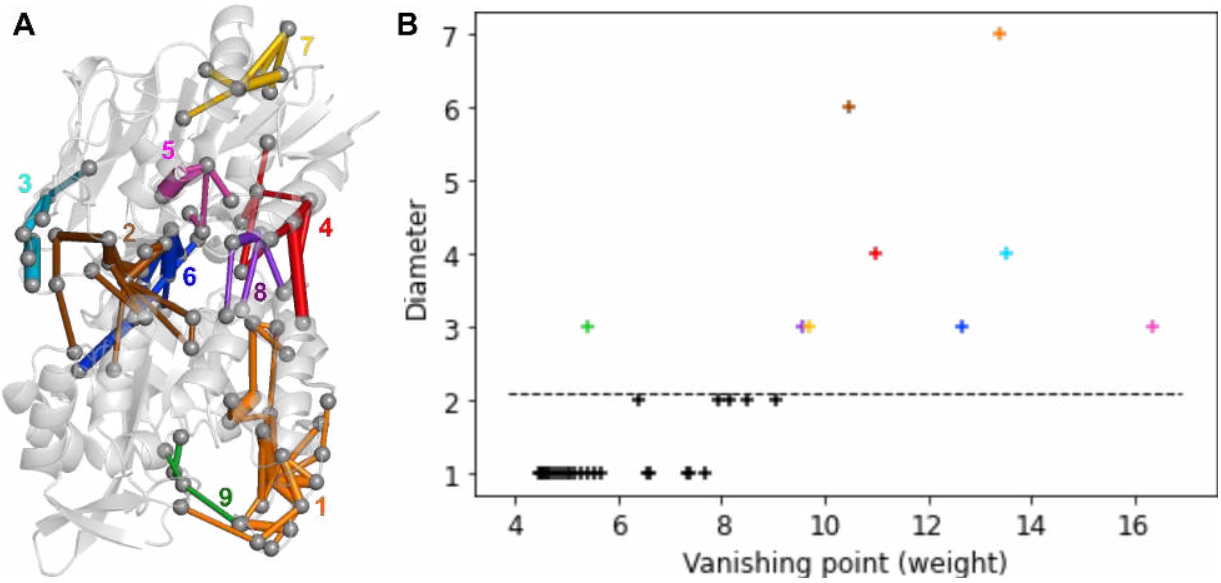


Figure 4: A. Dynamical Perturbation Network containing the 9 major component with diameter > 2. B. Scatter plot of the correlation between the vanishing point of a component and its diameter.

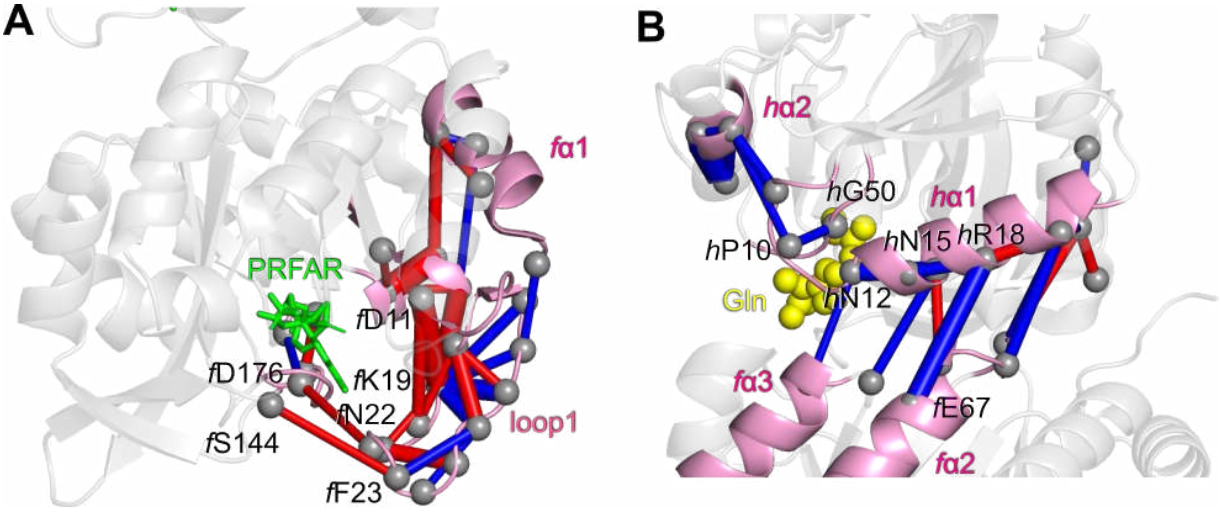


Figure 5: A. Representation of the Dynamical Perturbation Contact Network of component 1 and 9. B. Representation of components 4, 5 and 8.

The main asset of this approach to draw dynamical perturbation contact network is that the links from local to global perturbation is easier to grasp. The rudimentary threshold selection was producing many isolated edges and with this approach it was essential to mix the use of several threshold in order to be able to look at the entirety of a perturbation. Here knowing that the effector site is at the bottom of HisF makes it clear that the information about the origin of the allosteric mechanism is contained in components 1 and 9 show in Figure 5A. These components are neighbors and almost connected since component 1 contains residues fT21 and fF23 while component 9 contains residue fN22. We retrieve the contact network corresponding to the top-four groups of section between residues fA224, fF227 fH228, fK19, fD11 and fR27. But this time we are able to capture even more local perturbations. Almost all residues in loop1 from V18 to G30 undergoes rearrangements with another loop1 residue. But some interactions previously overlooked during perturbation network contact analysis stand out. Displacements from the effector site (fβ8-fα8 turn) and loop1 propagates to $f\alpha$1. This is in agreement with the latest results which shows that this secondary structure is also involved at sideR in HisF rearrangement. At the same time, two red edges connects sideR and sideL: fS144–fF23 and fD176–fN22. These increase in contact upon effector binding can be rationalized as loop1 folding up on PRFAR congruent to its binding at the effector site. Components ranked 4, 5 and 8 are represented on Figure 5B. Components 5 and 8 are also neighbors with hM14 belonging to component 5 and hN15 belonging to component 8. Remarkably component 4 is very similar to the zoomed contact perturbation network obtained in ref.[29] that we obtained using a lower threshold. Indeed, the connected component analysis present a generalization of this idea of zooming on local regions without having to choose a specific region. Component 5 is centered around the biggest edge of the graph hE56-hR59 that was highlighted as the single member of the top group of relevance. Thanks to the connected component analysis, we are now able to directly associate this perturbation with other displacements in the protein. hE56 hR59 are both linked with hR60 which itself linked with hV8 and hP10 (located in the Ω-loop). Finally

the hP10–hG50 blue edge highlights that the hydrogen bond between these two residues break, marking the penultimate step of the allosteric mechanism permitting the flipping of the oxyanion strand.

Conclusion

Birch clustering and connected component analysis have been proposed to further process dynamical perturbation contact network. Birch clustering tool established itself as a more reliable way to select a threshold than an arbitrary selection. Unfortunately this procedure remains tiresome to investigate propagation of signaling in proteins since the clustering is performed on the weight space rather than in the 3-dimensional distance space. On the other hand, the connected component analysis is a straightforward tool, removing the need to arbitrarily select a value (we can rationalize the only parameter which is the diameter of the component), directly displaying perturbations associated to the allosteric mechanism. Indeed of the nine components with a diameter >2 are staged four key allosteric motions: stiffening of loop1 and its folding on the effector binding pocket, rearrangements at sideR of HisF, alteration of the breathing motion and the flipping of the oxyanion strand. Here thanks to the local nature of the connected component analysis, we are able to recapture allosteric motions that previously required a fine tuning of parameters (hP10-hG50 hydrogen bond breaking, extension of the salt bridge alterations to $h\alpha 4$). Moreover some new information emerge, namely the role of residue fF23 in folding on the effector binding pocket upon effector binding and the involvement in $f\alpha 1$ in the rearrangements at HisF. Overall, we evaluated the connected component analysis to be a remarkable tool to facilitate the analysis of perturbation contact network. This procedure which is easy to implement and applicable to all weighted networks, has the potential to become a standard procedure to guide the investigation of other type of dense protein weighted networks whether static or dynamic.

Acknowledgement

LV and AG thank 80 prime CNRS programme and MITI

AG thanks FEBS Short-Term Scholarship

Supporting Information Available

Additionnal procedures including Birch clustering with fixed number of clusters, rough connected component analysis, distribution of size, order and diameter of the final components, and full representation of components.

References

(1) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

(2) Shaw, D. E.; Grossman, J.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H., et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2014; pp 41–53.

(3) Saltalamacchia, A.; Casalino, L.; Borisek, J.; Batista, V. S.; Rivalta, I.; Magistrato, A. Decrypting the information exchange pathways across the spliceosome machinery. *Journal of the American Chemical Society* **2020**, *142*, 8403–8411.

(4) Jung, J.; Nishima, W.; Daniels, M.; Bascom, G.; Kobayashi, C.; Adedoyin, A.; Wall, M.; Lappala, A.; Phillips, D.; Fischer, W., et al. Scaling molecular dynamics beyond 100,000

processor cores for large-scale biophysical simulations. *Journal of computational chemistry* **2019**, *40*, 1919–1930.

(5) Jung, J.; Kobayashi, C.; Kasahara, K.; Tan, C.; Kuroda, A.; Minami, K.; Ishiduki, S.; Nishiki, T.; Inoue, H.; Ishikawa, Y., et al. New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems. *Journal of Computational Chemistry* **2021**, *42*, 231–241.

(6) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical networks in tRNA: protein complexes. *Proceedings of the National Academy of Sciences* **2009**, *106*, 6620–6625.

(7) Alexander, R. W.; Eargle, J.; Luthey-Schulten, Z. Experimental and computational determination of tRNA dynamics. *FEBS letters* **2010**, *584*, 376–386.

(8) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric pathways in imidazole glycerol phosphate synthase. *Proceedings of the National Academy of Sciences* **2012**, *109*, E1428–E1436, Publisher: National Academy of Sciences Section: PNAS Plus.

(9) VanWart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring residue component contributions to dynamical network models of allostery. *Journal of chemical theory and computation* **2012**, *8*, 2949–2961.

(10) Gasper, P. M.; Fuglestad, B.; Komives, E. A.; Markwick, P. R. L.; McCammon, J. A. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *Proceedings of the National Academy of Sciences* **2012**, *109*, 21216–21222, Publisher: National Academy of Sciences Section: Biological Sciences.

(11) Miao, Y.; Nichols, S. E.; Gasper, P. M.; Metzger, V. T.; McCammon, J. A. Activation and dynamic network of the M2 muscarinic receptor. *Proceedings of the national academy of sciences* **2013**, *110*, 10982–10987.

(12) Stolzenberg, S.; Michino, M.; LeVine, M. V.; Weinstein, H.; Shi, L. Computational approaches to detect allosteric pathways in transmembrane molecular machines. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2016**, *1858*, 1652–1662.

(13) Serçinoğlu, O.; Ozbek, P. gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations. *Nucleic Acids Research* **2018**, *46*, W554–W562.

(14) Bowerman, S.; Wereszczynski, J. Detecting allosteric networks using molecular dynamics simulation. *Methods in enzymology* **2016**, *578*, 429–447.

(15) East, K. W.; Skeens, E.; Cui, J. Y.; Belato, H. B.; Mitchell, B.; Hsu, R.; Batista, V. S.; Palermo, G.; Lisi, G. P. NMR and computational methods for molecular resolution of allosteric pathways in enzyme complexes. *Biophysical reviews* **2020**, *12*, 155–174.

(16) Verkhivker, G. M.; Di Paola, L. Dynamic Network Modeling of Allosteric Interactions and Communication Pathways in the SARS-CoV-2 Spike Trimer Mutants: Differential Modulation of Conformational Landscapes and Signal Transmission via Cascades of Regulatory Switches. *The Journal of Physical Chemistry B* **2021**, *125*, 850–873.

(17) Artymiuk, P. J.; Rice, D. W.; Mitchell, E. M.; Willett, P. Structural resemblance between the families of bacterial signal-transduction proteins and of G proteins revealed by graph theoretical techniques. *Protein Engineering, Design and Selection* **1990**, *4*, 39–43.

(18) Böde, C.; Kovács, I. A.; Szalay, M. S.; Palotai, R.; Korcsmáros, T.; Csermely, P. Network analysis of protein dynamics. *Febs Letters* **2007**, *581*, 2776–2782.

(19) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proceedings of the National Academy of Sciences* **2009**, *106*, 6620–6625, Publisher: National Academy of Sciences Section: Physical Sciences.

(20) Aftabuddin, M.; Kundu, S. Weighted and unweighted network of amino acids within protein. *Physica A: Statistical Mechanics and its Applications* **2006**, *369*, 895–904.

(21) Barah, P.; Sinha, S. Analysis of protein folds using protein contact networks. *Pramana* **2009**, *71*, 369.

(22) Silveira, C. H. d.; Pires, D. E. V.; Minardi, R. C.; Ribeiro, C.; Veloso, C. J. M.; Lopes, J. C. D.; Meira, W.; Neshich, G.; Ramos, C. H. I.; Habesch, R.; Santoro, M. M. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics* **2009**, *74*, 727–743, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22187.

(23) Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Giuliani, A. Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chemical Reviews* **2013**, *113*, 1598–1613, Publisher: American Chemical Society.

(24) K. Grewal, R.; Roy, S. Modeling proteins as residue interaction networks. *Protein and Peptide Letters* **2015**, *22*, 923–933.

(25) Vuillon, L.; Lesieur, C. From local to global changes in proteins: a network view. *Current Opinion in Structural Biology* **2015**, *31*, 1–8.

(26) Achoch, M.; Dorantes-Gilardi, R.; Wymant, C.; Feverati, G.; Salamatian, K.; Vuillon, L.; Lesieur, C. Protein structural robustness to mutations: an in silico investigation. *Physical Chemistry Chemical Physics* **2016**, *18*, 13770–13780.

(27) Dorantes-Gilardi, R.; Bourgeat, L.; Pacini, L.; Vuillon, L.; Lesieur, C. In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: not too many links, not too few. *Physical Chemistry Chemical Physics* **2018**, *20*, 25399–25410.

(28) Doshi, U.; Holliday, M. J.; Eisenmesser, E. Z.; Hamelberg, D. Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proceedings of the National Academy of Sciences* **2016**, *113*, 4735–4740.

(29) Gheeraert, A.; Pacini, L.; Batista, V. S.; Vuillon, L.; Lesieur, C.; Rivalta, I. Exploring allosteric pathways of a v-type enzyme with dynamical perturbation networks. *The Journal of Physical Chemistry B* **2019**, *123*, 3452–3461.

(30) Negre, C. F.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences* **2018**, *115*, E12201–E12208, Publisher: National Acad Sciences.

(31) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Chittur, S. V.; Davisson, V. J.; Smith, J. L. Crystal structure of imidazole glycerol phosphate synthase: a tunnel through a (β/α) 8 barrel joins two active sites. *Structure* **2001**, *9*, 987–997.

(32) Gomez, M. J.; Neyfakh, A. A. Genes involved in intrinsic antibiotic resistance of Acinetobacter baylyi. *Antimicrobial agents and chemotherapy* **2006**, *50*, 3562–3567.

(33) Breitbach, K.; Köhler, J.; Steinmetz, I. Induction of protective immunity against Burkholderia pseudomallei using attenuated mutants with defects in the intracellular life cycle. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **2008**, *102*, S89–S94.

(34) Lisi, G. P.; Currier, A. A.; Loria, J. P. Glutamine hydrolysis by imidazole glycerol phosphate synthase displays temperature dependent allosteric activation. *Frontiers in molecular biosciences* **2018**, *5*, 4.

(35) Rivalta, I.; Lisi, G. P.; Snoeberger, N.-S.; Manley, G.; Loria, J. P.; Batista, V. S. Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein–protein interface. *Biochemistry* **2016**, *55*, 6484–6494.

(36) Lisi, G. P.; East, K. W.; Batista, V. S.; Loria, J. P. Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity. *Proceedings of the National Academy of Sciences* **2017**, *114*, E3414–E3423.

(37) Wurm, J. P.; Sung, S.; Kneuttinger, A. C.; Hupfeld, E.; Sterner, R.; Wilmanns, M.; Sprangers, R. Molecular basis for the allosteric activation mechanism of the heterodimeric imidazole glycerol phosphate synthase complex. *Nature communications* **2021**, *12*, 1–13.

(38) Kneuttinger, A. C.; Rajendran, C.; Simeth, N. A.; Bruckmann, A.; König, B.; Sterner, R. Significance of the Protein Interface Configuration for Allostery in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **2020**, *59*, 2729–2742.

(39) Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record* **1996**, *25*, 103–114.

(40) Gauvin, L.; Panisson, A.; Cattuto, C. Detecting the Community Structure and Activity Patterns of Temporal Networks: A Non-Negative Tensor Factorization Approach. *PLOS ONE* **2014**, *9*, e86028, Publisher: Public Library of Science.

(41) Latapy, M.; Viard, T.; Magnien, C. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining* **2018**, *8*, 61.

(42) Viard, T.; Magnien, C.; Latapy, M. Enumerating maximal cliques in link streams with durations. *Information Processing Letters* **2018**, *133*, 44–48.

(43) Karsai, M. Computational Human Dynamics. *arXiv:1907.07475 [physics]* **2019**, arXiv: 1907.07475.

(44) Vijayabaskar, M. S.; Vishveshwara, S. Interaction Energy Based Protein Structure Networks. *Biophysical Journal* **2010**, *99*, 3704–3715.

(45) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005**, *26*, 1668–1688.

(46) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **2004**, *25*, 1157–1174.

(47) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **2005**, *26*, 1781–1802.

(48) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.

(49) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N×log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993**, *98*, 10089–10092.

(50) Grubmüller, H.; Heller, H.; Windemuth, A.; Schulten, K. Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Molecular Simulation* **1991**, *6*, 121–142.

(51) Schlick, T.; Skeel, R. D.; Brunger, A. T.; Kalé, L. V.; Board Jr, J. A.; Hermans, J.; Schulten, K. Algorithmic challenges in computational molecular biophysics. *Journal of Computational Physics* **1999**, *151*, 9–48.

(52) Harris, C. R. et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362.

(53) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MD-

Traj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **2015**, *109*, 1528 – 1532.

(54) Hagberg, A.; Swart, P.; S Chult, D. *Exploring network structure, dynamics, and function using NetworkX*; 2008.

(55) Behnel, S.; Bradshaw, R.; Citro, C.; Dalcin, L.; Seljebotn, D. S.; Smith, K. Cython: The Best of Both Worlds. *Computing in Science Engineering* **2011**, *13*, 31–39.

(56) Bentley, J. L. Multidimensional binary search trees used for associative searching. *Communications of the ACM* **1975**, *18*, 509–517.

(57) Maneewongvatana, S.; Mount, D. M. It's okay to be skinny, if your friends are fat. Center for geometric computing 4th annual workshop on computational geometry. 1999; pp 1–8.

(58) Moore, A. W.; Connolly, A. J.; Genovese, C.; Gray, A.; Grone, L.; Kanidoris II, N.; Nichol, R. C.; Schneider, J.; Szalay, A. S.; Szapudi, I., et al. *Mining the Sky*; Springer, 2001; pp 71–82.

(59) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.

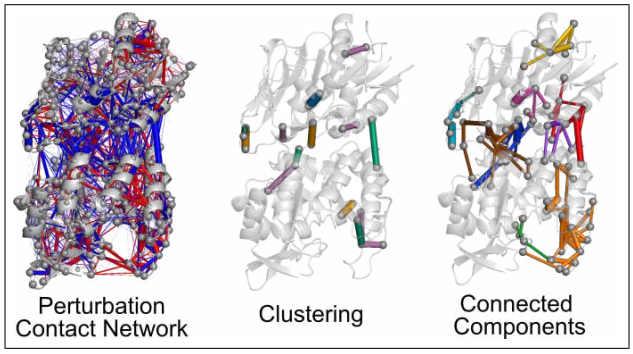
(60) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(61) Zuse, K., et al. Der Plankalkül. **1972**,

(62) Moore, E. F. The shortest path through a maze. Proc. Int. Symp. Switching Theory, 1959. 1959; pp 285–292.

(63) Lee, C. Y. An algorithm for path connections and its applications. *IRE transactions on electronic computers* **1961**, 346–365.

Graphical TOC Entry



2.3 Generalization of Perturbation Contact Analysis

2.3.1 From a *supervised* to an *unsupervised* procedure

The problems that DPCN can help to solve are usually quite narrow: they involve a set of MD simulations in a *reference* and *perturbed* system. All the contact information contained in each frame of the MD simulations is simply averaged in a *supervised* manner. When running MD simulations, it is good practice to run several simulations of the same system (also called *replicas*) to explore the most probable conformations of a protein. DPCN can also be used to investigate differences between replicas of the same system and between single replicas of different systems. Besides, the conformation of a protein can change in a single replica. Thus, averaging it can produce unexpected results. This work quickly becomes tedious because we do not know *a priori* which are replicas and time windows with interesting behaviors. To understand this issue in more depth, we can investigate DPCN between different time windows of a replica. Moreover, this approach does not provide with a real way of quantifying the importance of difference between replicas. In fact, DPCN do not guarantee that the difference displayed is significant (i.e. not noise) nor reproducible (i.e. found in every replica and within all frames of a replica). In practice, problems can also be much more complex than studying only two systems and having one *reference* system and multiple *perturbed* system. Running DPCN to solve such problems also becomes tedious.

Instead of averaging the contacts of frames with a *supervised* labeling, we introduce here a procedure where all contact data are stored in a *contact matrix*. In the *contact matrix*, the rows corresponds to the different samples (i.e. frames) and the columns corresponds to the different features (i.e. contacts). From this contact matrix, we extract the principal axes of variance using Principal Component Analysis (PCA). With standard linear PCA, the eigenvectors corresponding to a principal component are linear combinations of all contacts and can thus also be represented as contact networks. In our main system of interest, IGPS from *T. maritima*; we actually discovered that the first principal component (PC1) of variance using the concatenated contact matrices of all simulations of apo and PRFAR-bound IGPS completely separates frames of apo from frames of prfar. Moreover, the corresponding PC1 Network (PC1N) is highly correlated to the DPCN. This proves that the differences that we previously investigated are both significant and reproduced in all replicas and within replicas.

2.3.2 Other methodological development and applications

Contact Principal Component Analysis

We report the contacts weights of different frames in a matrix C of size $N_{frames} \times N_{contacts}$. If a contact is not present in a frame, its weight is simply put as zero for this frame. We use Principal Component Analysis (PCA, cPCA for contacts) to extract the k-first principal components (PCs). The PCs are each of size N_{frames} and represent the projection of the frames in this component. During the decomposition, we compute the (ordered) eigenvectors of the covariance matrix. Each of these eigenvectors corresponds to a principal component and is of size $N_{contacts}$, thus representing a linear combination of all contacts in the system. We define a new type of contact network: the ith PC Network (PCiN) in which nodes are amino acids of the protein, edges are all contacts, and weights are the value of the contact in the eigenvector. These eigenvectors also corresponds to an eigenvalue, which is representative of the importance of the principal component. In PCA, the eigenvalues and eigenvectors are ordered so that the PCs decrease importance with the component number.

As described in reference [1], we can restrict ourselves to one or two given eigenvectors i and j, we use them to get a projected 2D free energy-landscape of the system along the eigenvector dimensions:

$$\Delta G(PC_i, PC_j) = -k_{\mathrm{B}}T\ (P(PC_i, PC_j)) - G_0 \tag{2.1}$$

where $P(PC_i, PC_j)$ is a probability estimate obtained from the MD frames and G_0 is the free energy of the most probable state.

Other features and their challenges

Simple atomic contacts can be compared with other features of dynamical protein structures, such as: Cartesian coordinates of the C_α (xPCA), linearized ϕ and ψ dihedral angles (dPCA), binary contact (fPCA) and transformed distance of contact (tdistPCA). Features, thus, can be external measures of the system (i.e., dependent on a frame of reference outside the protein, such as Cartesian coordinates) or internal measures (all other features studied). Internal features have an intrinsic advantage over external features because they are independent of the translations or rotations of the system and thus do not need alignment of frames.

Another discrimination between features is how they are indexed. In cartesian coordinates, three value (x, y, z) corresponds to each atom of single amino-acid. By contrast, binary contact, distance contact or simply contact have a single value that corresponds to two amino acids. When multiple features values corresponds to a single contact, we have to slightly modify the PCA procedure. The PC vectors and 2D free-energy landscape are not affected but the eigenvector interpretation must change. To avoid compensatory effects, we compute the importance of an amino acid as the square root of the sum of squares of the influence of each of its individual features in the eigenvector. In cartesian coordinates, because this importance corresponds to a single amino acid, we cannot represent any PCiN. All features that are indexed in correspondence with two amino acids can be represented as PCiN. Other features, not presented here, such as volume elements, relate to more than two amino acids. In such cases, outside the scope of this thesis, hypergraphs can be defined.

Finally, the importance of some features does not necessarily grow linearly. This is notably the case for ϕ and ψ backbone dihedral angles which have to be linearized from the circular space using the transformations:[1]

$$q_{4n} = \sin\phi_n; \quad q_{4n+1} = \cos\phi_n; \quad q_{4n+2} = \sin\psi_n; \quad q_{4n+3} = \cos\psi_n \tag{2.2}$$

with $n = 1, \ldots, N$ corresponding to the N pairs of consecutive residues from which dihedral angles are considered (in practice $= N_{\text{residues}} - N_{\text{chains}}$). Similarly, the shortest distance between heavy atom contacts between two residues does not evolve linearly with the energy, but $\frac{1}{d^2}$ does. Thus we transform the closest heavy-atom distance matrix between residues in contact with the inverse squared function.

Single trajectory PCA

First, we restrain our analysis on a single trajectory: apo1 (100 ns, 1,000 frames). In total, 3,668 pairs of residue establish a contact during this trajectory, resulting in a $1,000 \times 3,668$ contact matrix. In order to assess the number of relevant components in PCA, we first report the PC1 using the first one hundredth PCs (in PCA, the truncation of the smaller components does not affect the bigger). In Figure 2.7 the evolution of the explained variance ratio of each component is reported with each component and its cumulative counterpart. The explained variance ratio indicates how much variance of an overall signal is explained by a component. It is directly related to the eigenvalue of a component and thus descent in order of magnitude, each component explains less and less variance. Many selection criterions exist. For instance, one of them is the Kaiser criterion, which says that we should keep only components explaining more variance than a single feature. For 3,668 contacts, this means selecting components explaining more than 0.027% of the overall variance. In our case, this criterion is impractical, since it require selecting more than the 100 original components for analysis. A more traditional criterion selection is to select components in order to select the "elbow" of the graph where the explained variance ratio seem to level off. This criterion is criticizable because of its subjectivity, but here gives a more reasonable choice, selection between 5 and 15 components. Another criterion we can use is to select only components that explain more than 1% of the system variance. Here this means selecting the 9th first PC, which matches nicely with the elbow selection and is our choice in the rest of the analysis. Despite the fact that the two first PC combined explains about 16% of the variance, even selecting components up to the 100th explains only 60% of the system variance, and the growth of the cumulative explained variance slows very quickly. This implies that most of the variance cannot be simply explained by global motions. This result is actually in good agreement with the fact that a nicely equilibrated MD simulations has many random thermal fluctuations, and that much of this variance in the simulation simply is not explicable.

In Figure 2.8A we represent the time-evolution of the values associated to the 9th first PC and PC20 (for comparison) during simulation apo1. PC1 display a particularly interesting behavior since at the beginning of the simulation, its value is high in the positives (around 100), but before 20 ns it quickly decreases and then stabilizes around a value of -50. This suggests that PC1 represents a relaxation signal. At the beginning of the simulation, the protein is folded in a position out of equilibrium but then during the simulation it quickly relaxes. Interestingly, in Figure 2.8B we show that in the corresponding PC1N the contacts associated with this relaxation are mainly found within the allosteric pathways and notably the $f\alpha1$ and $f\alpha2$ helices and at the interface between HisF and HisH. The pre-equilibrated structures were built using the 1GPW chain C and D PDB structure. This structure stands out particularly because it has loop1 in a folded position and phosphate groups are bound at the effector site at the precise spot where PRFAR binds. Therefore, this structure may possess some degree of activation which could explain why the relaxation signal is mainly found in the allosteric pathways.
By contrast PC2 and PC3 shows oscillating behaviors in Figure 2.8A. The period of PC2 is about the length of the simulation (100 ns), and its amplitude is around 100 (ranging from -50 to +50). The period of PC3 is twice as small (50ns) and has a similar amplitude. Periodic motions like these are more likely to represent the system intrinsic dynamics than PC1 which represents an artifact from the system preparation. The contacts involved in the signal described by PC2 are mainly found at the interface, and a 100-ns period perfectly matches previous descriptions of the breathing motion in simulations of apo[2] so this suggests that PC2 represents contacts involved in the breathing motion. In contrast, influences in PC3 are mainly localized to a few

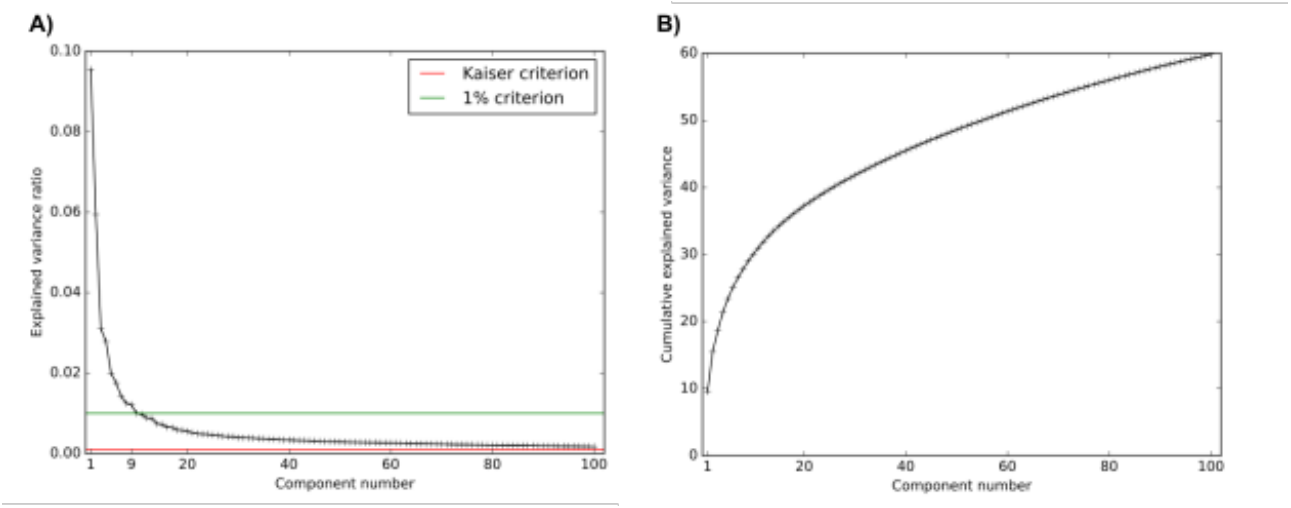


Figure 2.7: A. Explained variance ratio in function of the component number, with the Kaiser criterion displayed in red and the 1% criterion in green. B. Cumulative explained variance ratio in function of the component number.

contacts. Some of these contacts are directly located near the effector site and in secondary structures involved in allosteric pathways. The PC3 signal accurately represents the system dynamics, therefore its concentration in the allosteric pathways may indicate that it is intrinsically labile showing that the allosteric pathways are encoded in the protein structure.
PC4, similar to PC1, starts at a high positive value (75) and quickly decreases to negative values before 20 ns. Then it stabilizes around zero and oscillates between positive and negative values. By the same argument we can postulate that this signal represents a relaxation degree of the system. However the main difference is that in this case, an oscillation around zero starts after relaxation. Here this suggests that the component actually represents some of the system dynamics but that the initial position is a far from equilibrium.

Each smaller component shows an oscillating behaviors around zero. This suggests that there is no more big relaxation component and that, in fact, the system is nicely equilibrated after 20 ns. Comparison between molecular dynamics and random walks of protein systems has shown that the more PCA biplots resemble a cosine function, the more this trajectory is similar to a random walk, non-convergent, and therefore contains little dynamical information[3]. In Figure 2.9 we represent the biplots between each combination of the 9th first PC. None of the biplots represents a cosine for the full duration of the simulation. However, in the projections associated with PC1 and PC4, the early steps of the simulation (40 first ns for PC1 and 20 first ns for PC4) approximate a cosine shape in most of the associated biplots. This further indicates that PC1 and PC4 are not really representative of the system dynamics, and instead represent the shift from the initial structure towards a more equilibrated system. By contrast all other plots, including those displaying PC2 are much different from a cosine and this suggests they contain valuable information about the protein dynamics and the underlying free energy landscape.
Interestingly, both the period and the amplitude of the value of the components are shortening with the component number. This proves that the biggest components are representative of large scale slow collective motions while smaller components are fast and localized motions and the smallest components are virtually only noise such as shown in the PC20 graph whose value is oscillating so fast with such a small amplitude that it can hardly be explained. We proved that cPCA is a valuable tool for extracting different contact signals from a MD simulation. A first application is for convergence analysis and shows if we can extract some relaxation signals from the dynamics. Another application is to capture contacts involved in the main motions of the intrinsic dynamics of the protein. We suggest that, if they happen, events such as local (un)folding, intrinsic disorder, or large-scale conformational shifts should be extracted by cPCA. In particular, cPCA facilitates the analysis because it is an unsupervised technique which finds the time-window in which the corresponding events are happening by contrast to the DPCN technique in which frames have to be labeled to perform averages and differences. Similarly, by concatenating contact matrices of different systems and/or replicas (when they can be concatenated), the cPCA technique could provide a vision of the most important differences between the systems and whether they are reproducible.

Multi-trajectory analysis and allosteric signal decomposition

To compare contact signals in simulations of apo and prfar, we concatenate the contact matrices of those simulations (apo1 to apo4, prfar1 to prfar4) and perform PCA on this contact matrix. This matrix is of size $8,000 \times 4,408$ and in a first study, we select the first hundredth PC to select an appropriate number of components to study. In Figure 2.10, we show that the scree plot variance ratio shows again an elbow between

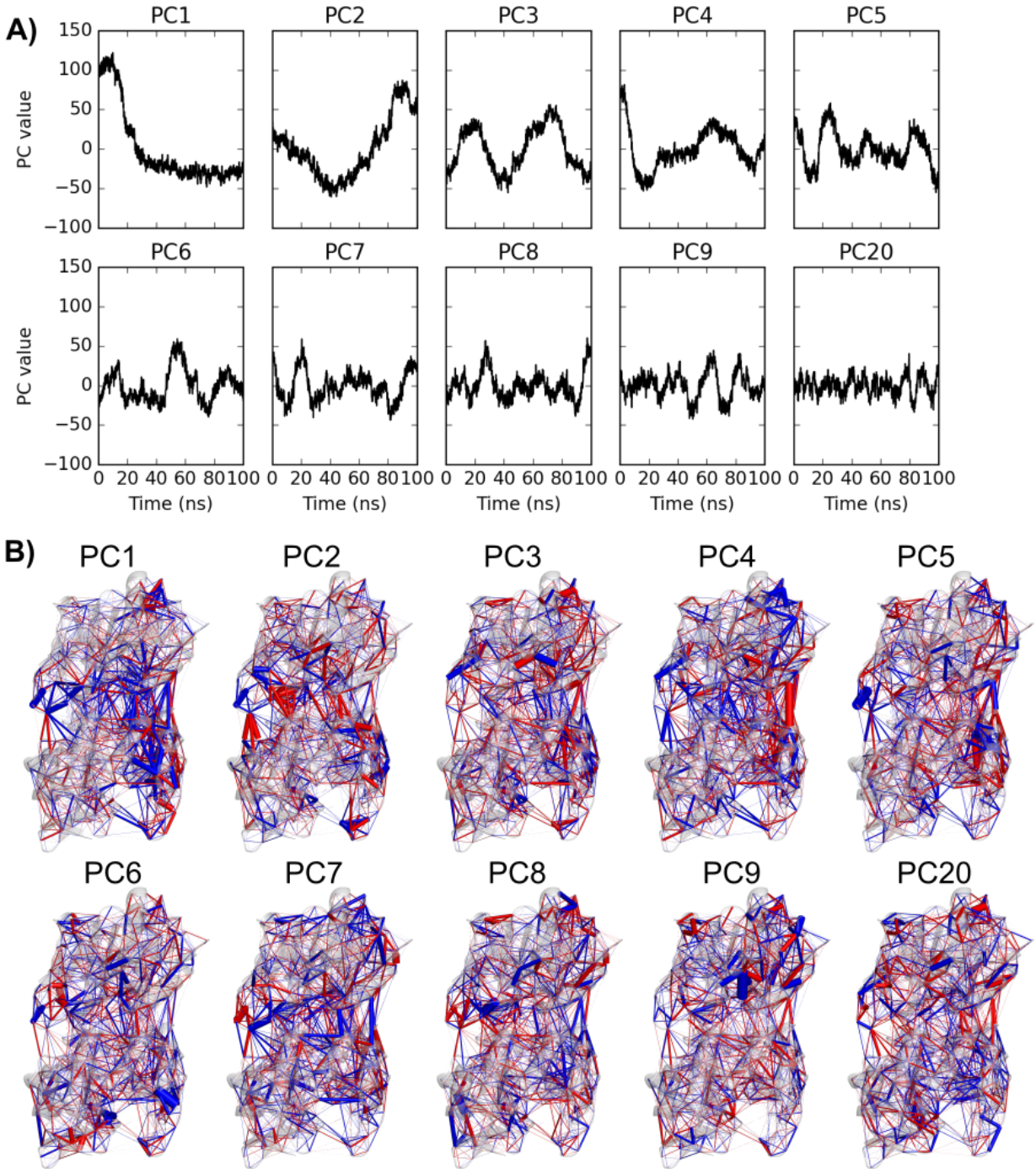


Figure 2.8: A. Time-evolution of the 9th first principal components and the 20th. B. Network representation of the eigenvector associated to the 9th first principal components and the 20th. Edge width is proportional to the contribution of each edge in its PCN. Edge color is blue if a contact is typically stronger in frames with negative PC value and red if this contact is stronger in frames with a positive PC value and vice-versa.

5 and 15 and that the 1% criterion rule shows that a good choice is to select the first nine PCs. This number is the same as the number of components we selected during the single-trajectory analysis. This is somehow surprising because we expect that adding more signal (produced in very different conditions) will produce more relevant components. Still, this effect can be moderated by the fact that the addition of more input data helps define more clearly if the signal carried by a component is reproducible along all simulations or if it is an artifact found in a single simulation.

In Figure 2.11A we show the evolution of the ninth first principal components values during time in the different simulations. Very interestingly, values of PC1 perfectly discriminate frames of apo (with a PC1 around 50) from frames belonging to prfar simulations (PC1 around -50). This suggests that the first principal component contains essential information about contacts that differentiate the frames of apo and the frames of prfar. Therefore, PC1 is a suitable candidate to represent the allosteric contact signal. Among the ninth first principal components, the first is the only one which is able to strictly discriminate the frames this way. This is expected

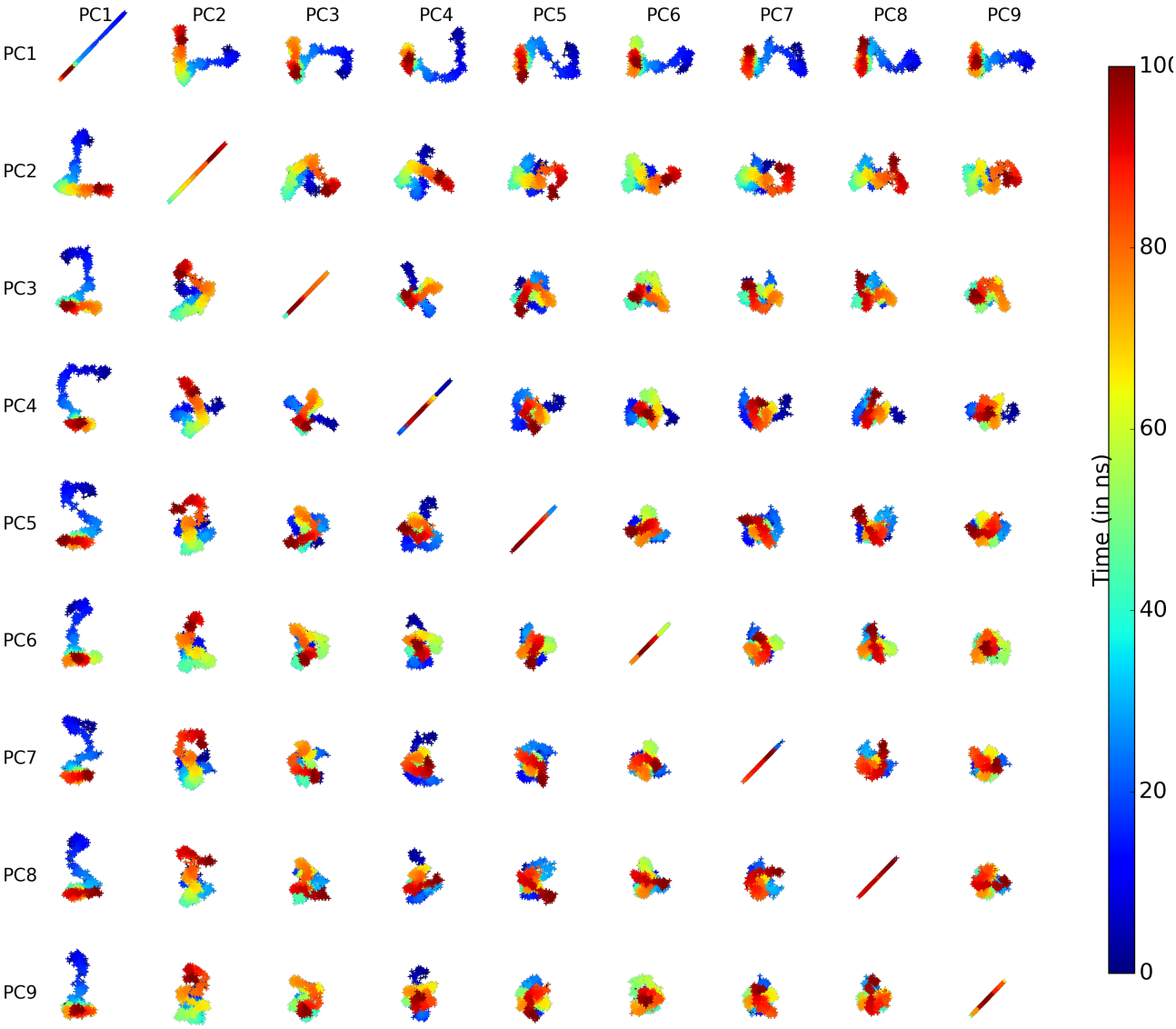


Figure 2.9: Biplot between the 9th first principal components

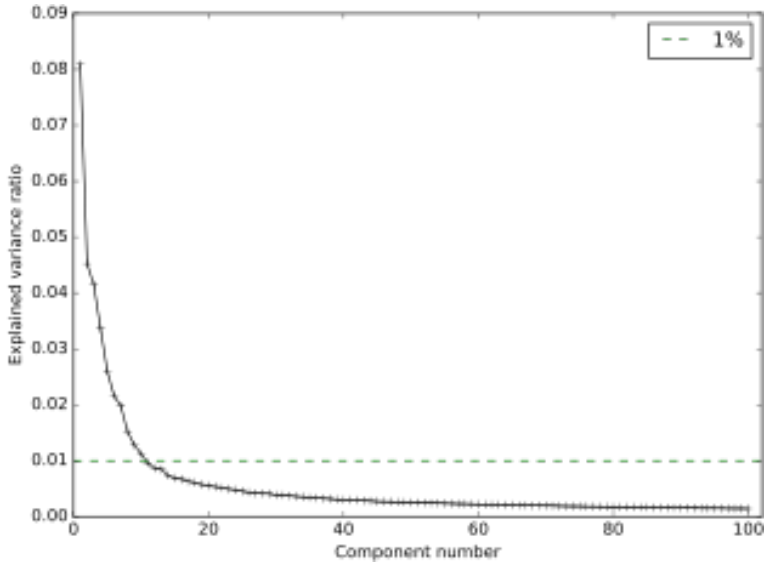


Figure 2.10: Scree plot for the 100 first principal components of the multiple trajectory contact signal. The 1% criterion is displayed in green dotted line.

from PCA properties and shows that PC1 extracts all the allosteric information found in our simulations.

PC2, on the other hand, has mostly positive values for apo1, prfar1-2 and negative values for apo2-4 and

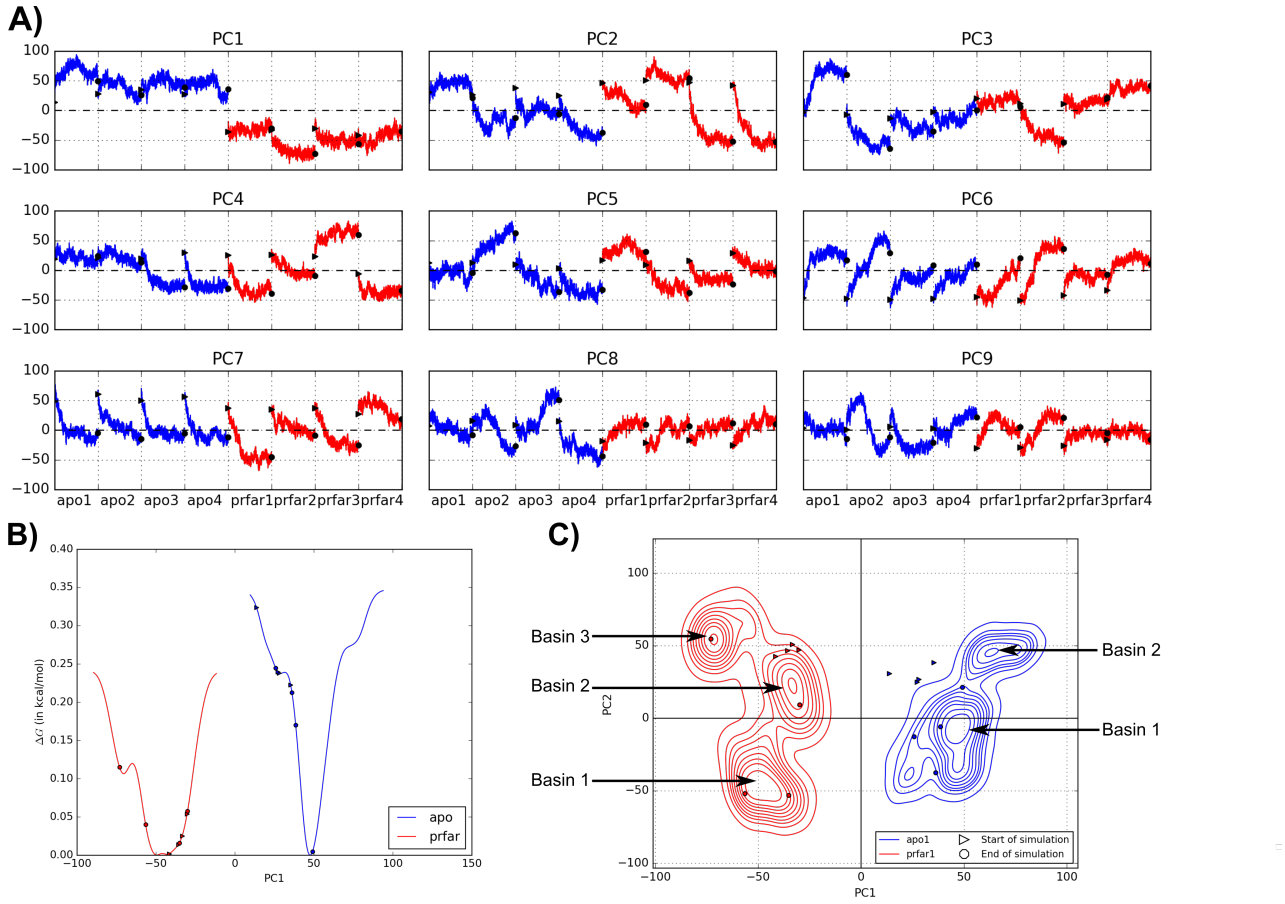


Figure 2.11: (A) Time-evolution along the different MD simulations of the PC values of the 9th principal components 1 to 9. Simulations of apo are colored in blue and simulations of prfar in red. (B) Free-energy computed (separately) along the PC1 axis for simulation apo1-4 and prfar1-4. Start of simulations are flagged with a right triangle and end of simulations flagged with an octagon. (C) 2D Free-energy landscape of simulations apo1-4 and prfar1-4 shown as contour lines computed separately by system. For each system, there are 9 contour lines separating the plot in 10 areas with iso-proportions of probability (hence a 10% probability). The path taken by a simulation starts from a right triangular flag and ends in an octagonal flag.

prfar3-4. This indicates a signal of a different nature which is not exclusive to either apo or prfar. This signal is probably representative of the exploration of different conformations by different replicas. In this case, the conformations obtained are not exclusive to either apo or prfar, which notes that this conformational change can be achieved in presence or absence of PRFAR. Still, whether PRFAR affects this conformational change remains open.

PC3, PC4, and PC5 also point out signals that are representative of the difference between replicas. PC3 shows positive values for apo1 and prfar4 and negative values for apo2 and prfar2 while the other replicas are fluctuating around zero. PC4 shows mostly positive values in simulation apo1-2 and prfar3 and negative values in apo3-4, prfar1 and prfar4 while prfar3 is oscillating around zero. In PC5, the simulations apo2 and prfar1 are strongly positive and apo3-4 and prfar2 are strongly negative. Other simulations are fluctuating near zero. These last three components are representative of most subtle behaviors, that are not consistently reproduced among simulations and not exclusive to either apo or prfar. While PC1 and PC2 both separates roughly 50% of the frames from each other, which maximizes variance, and it has to be noted that here, PC3, PC4 and PC5 separates much less neatly the frames, which accounts for component that are explaining less variance.

In PC6, the value in each simulation starts around -50, but then quickly increases to a positive value in the simulations apo1, apo2, and prfar2, but around zero in the simulations apo3, apo4, prfar1, prfar3, and prfar4. Similarly, in PC7, the value of the component in each simulation starts around 50 but then quickly drops around zero in apo1, apo2, apo3, apo4. In prfar simulations, the behaviors are much more nuanced; in prfar1 and prfar2 it quickly drops to a negative value, but in prfar2 it slowly relaxes towards zero, while in prfar4 it oscillates in the positive values. Both these signals are representative of a relaxation signal from the input structure to equilibrium. The difference between PC6 and PC7 is that in PC6, the relaxation is along a dimension where the system is already fluctuating (behaviors are rather different in all simulations after the initial relaxation), but in PC7 this is a pure relaxation for all apo simulations (converging fast towards zero), but is a fluctuating

dimension for frames of prfar (different behaviors in this case). Although these components contain an artificial component due to the input geometry, this shows that PCA is able to decompose this relaxation signal from the other components, which suggests that other components are not tarnished by relaxations occurring during the early dynamics of the simulation. Moreover, because these two components are now number 6 and number 7, this shows that the use of more frames in PCA grants a lower weight to these initial artifacts.

In PC8, the oscillations in the simulations apo1 and prfar1-4 are close to noise. The signal in the other simulations is very replica specific. Similarly, in PC9, the oscillations of simulations apo1, prfar3-4 are close to noise, with the other simulations showing a small but very specific behavior. These specific behaviors are achieved in particular simulations and are not consistently reproduced in different replicas of the same system or across systems. The ninth first principal components can be organized in groups which follow a relative order. First the difference between systems (here, PC1 the allosteric signal), second the differences between replicas (here PC2-5), then relaxation signals reproduced in different replicas (here PC6-7) and then system-specific behaviors with PC8-9. This goes with the PCA properties, which finds the axes of maximum variance. PC1 offers then a rather unique view of the allosteric signal, since it represents the differences between frames from apo and prfar that are consistently reproduced in all replicas.

In Figure 2.11B we represent the free-energy potential well calculated from the probabilities of the PC1 value in the simulations apo1-4 and prfar1-4. The two wells do not overlap because values taken in PC1 in apo and prfar also do not overlap. The apo simulations potential well is much more narrow than the one of prfar simulations. At first glance, this seems in contradiction to the fact that PRFAR binding is hypothesized to tense the protein in a more defined conformation. This phenomenon can be explained by two different factors. In first, PC1 is only one dimension of evolution of the system, which means that this situation may be different in PC2, PC3 and so on. In second, the starting point of all apo simulations is high in energy in the PC1 well, and they all start with a relatively low PC1 which is high in energy. This echoes with our previous result in the single-trajectory analysis, which found that the relaxation signal (which was also PC1) was found predominantly in the allosteric pathways. This suggests that the initial structure of the apo simulations still contains some elements of "activation" that quickly disappears. This matches with the ensemble view of allostery. This may cause exploration of the energy landscape to be biased by the initial position.
In Figure 2.11C, we show a 2D projection of free-energy landscape (or biplot) of simulations apo1-4 and prfar1-4 on the PC1 and PC2 axes. This view complements complementary the view which uses only the PC1 axis. This picture shows that all starting points for the simulations of apo are found outside the contour lines (i.e. in the 10% of frames that are found in the rest of the plane). This again shows that the initial positions for the apo simulations are a bit out of equilibrium. The situation for prfar simulations is more nuanced with this 2D projection. Indeed, the start of prfar simulations are found inside the contour lines but are in a high energetic position considering the PC2 axis.

In apo simulations, two main energetic basins form, one centered around the point (50, 0) and one around (60, 40). According to the PC1 values and the PC2 values in Figure 2.11B (or Figure 2.12), we attribute the first basin to the simulations apo2-4 and the second to the simulation apo1. Interestingly, the end point of simulation apo1 is in a position close to a transition towards the first basin. This suggests that the second basin is not stable (maybe metastable) and that only the first basin is the most representative of the system at equilibrium. In prfar simulations, we show three different energetic basins, the first around (-50, -50), the second around (-30, 25) and the third around (-60, 50). Similarly, we can attribute the first basin to simulations prfar3-4, the second to simulation prfar1 and the third to simulation prfar4. Despite all simulations starts in basin 2 they mostly end up in different basins which shows that the system is more relaxed in the PC2 axis. By contrast with apo simulations which finds stable basins only near zero PC2, prfar simulations have stable basins with various PC2 values, even in strongly negative or positive PC2. This is in fact in good accordance with the ensemble view of allostery, in which the binding of an effector affects not only the most stable conformation of a protein but also the least stable conformations. The depiction of the PC2N in Figure 2.13 shows that the contact changes associated to this component are very localized and associated to the (un)folding of the $h\alpha 4$ helix and to the breathing motion. Still, to analyze in depth this more subtle behavior, the length of the simulations may be an obstacle, because we probably did not explore enough of the energy landscape. This is contrasts with PC1, which extracts the allosteric signal, which here is shown to be perfectly reproducible in all simulations and at every time step.

Comparison between PC1N and DPCN

In Figure 2.14A is represented the PC1N and DPCN obtained from simulations apo1-4 and prfar1-4. At first glance, the resemblance between the networks is really striking. Both graphs are very dense and hard to read. In Figure 2.14 B, we represent a more digestible representation of those networks based on a Connected Component Analysis (CCA, note that the word component here has a different mathematical meaning than in the Principal Component). CCA shows that the networks are actually a bit different because applied to the PC1N

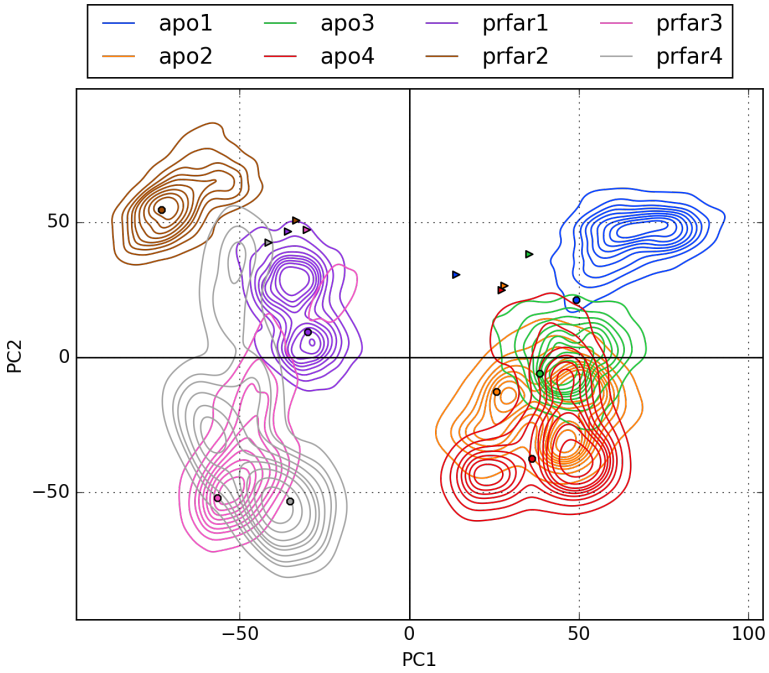


Figure 2.12: 2D Free-energy landscape of simulations apo1-4 and prfar1-4 shown as replica-disjoint contour lines. For each simulation, there are 9 contour lines separating the plot in 10 areas with iso-proportions of probability (hence a 10% probability). The path taken by a simulation starts from a right triangular flag and ends in an octagonal flag.

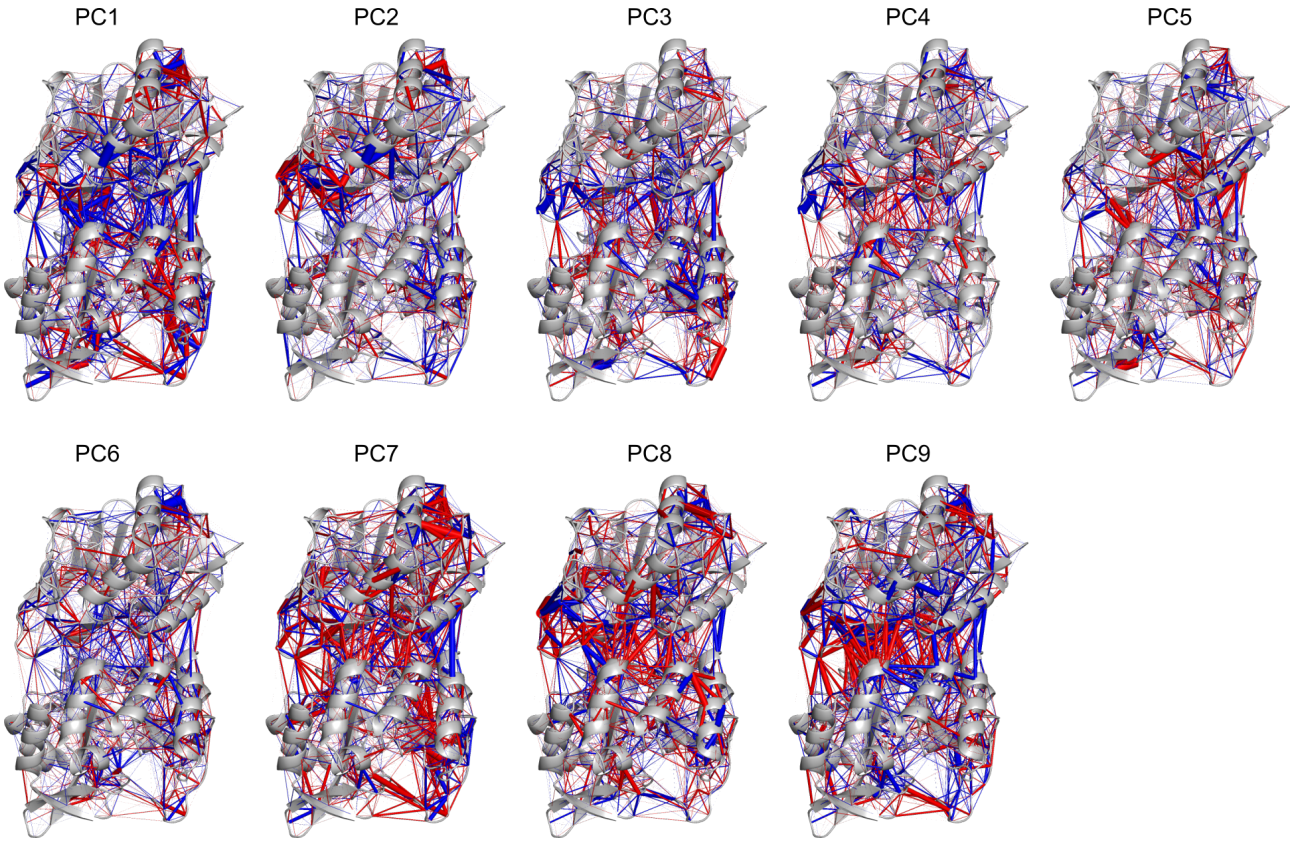


Figure 2.13: Ninth first principal components networks in the multi-trajectory analysis

we find 11 Connected Components (CC), with 95 edges and 98 nodes compared to the 9 CC, 82 edges and 82 nodes in the DPCN. Most of the components are identical between the two analyses. One slight difference is that one CC representing the propagation of perturbation near the effector site at sideR is broken in two in the PC1N. Then, two components appear with PC1N: one located near the effector site at sideL and one located between the $f\alpha1$ and $f\alpha2$ helices. Finally, after cleaning with CCA, one component disappear in the PC1N which is found at sideL near the interface in HisH and was attributed to side effects of the breathing motion.

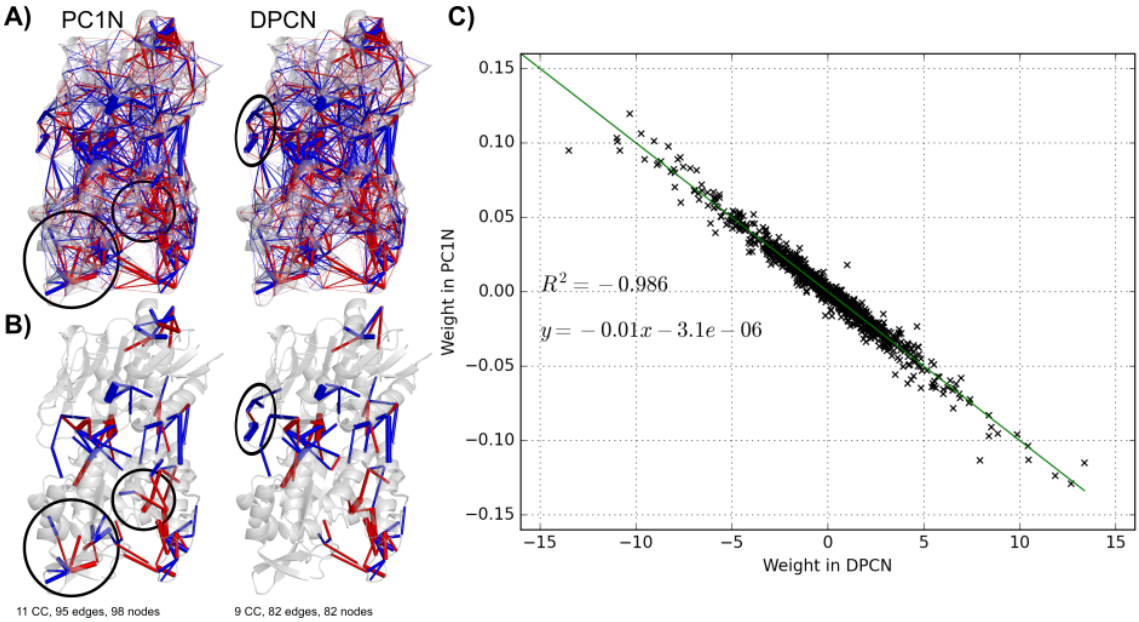


Figure 2.14: Representation on the protein structure of the complete (A) with connected component analysis (B) DPCN and PC1N between the 4 simulations of apo and PRFAR-bound. Edge colors are assigned so that a blue edge represents a contact stronger in apo while a red edge represents a contact stronger in the PRFAR-bound complex. Edge width are normalized by the value of the maximum edge and are proportional to the weight in each network. Black ellipses surrounds components that are present in the CCA for each network but not in the other. (C) Correlation plot between the weights in the DPCN and the PC1N.

Components which are appearing and disappearing in the CCA of PC1N have indeed respectively stronger and lower weights in the PC1N than in the DPCN. Interestingly, the two components which are appearing are desirable because one represents effects associated to the effector binding and the other a salt bridge network alteration between $f\alpha 1$ and $f\alpha 2$ previously reported in correlation analysis[2] and DPCN[4] but not in CCA of DPCN. Finally, the component that disappears in the PC1N is not much studied and was not really related to any relevant experimental data. This shows that despite the strong similarity between PC1N and DPCN, PC1N has a superior ability to extract the allosteric signal. To quantify this similarity, we report in Figure 2.14C the correlation plot between the weights in the DPCN and in the PC1N and the linear regression between the two data sets. A very strong anti-correlation is found here ($R^2 < -0.98$). The anticorrelation appears here because PCA suffers from sign indeterminacy, and here it found weights in the opposite direction as the DPCN (we reversed the color signs in the PC1N 2.14A-B for simplicity). This very strong correlation is in good accordance with the graph that look very similar. Still, the average relative error between the fit and the PC1N data is of 16% which shows that the graphs are not perfect matches. Interestingly, the slope of the fit is equal to 0.01 up to the fifth decimal. In fact, the norm of the DPCN is equal to 98 while the PC1N is by nature normalized, showing that the 1/100 factor here is mostly coincidental. The intercept of the fit is equal to 0 up to the 6th decimal, which shows that the fit is in fact linear.

Comparison between cPCA and other PCA techniques

In Figure 2.15A we report the 2D projection of the free-energy landscape in the PC1 and PC2 eigenvector dimensions computed separately for simulations of apo and prfar using atomic displacements of alpha carbons. A major difference between atomic displacements and backbone dihedral angles or contacts is that the atomic displacements are computed using Cartesian coordinates which are external to the system. Such measures are not invariant by rotation or translation of the complete protein or even wrapping within a box. This means that before computing atomic displacements, trajectories must be correctly wrapped and aligned. Using only atomic displacements, the free-energy landscape shows that simulations of apo and prfar overlap a lot. In fact, the landscapes computed separately for each replica show that while individual replicas can be assigned a mostly negative or positive PC1 and PC2, this does not translate to the systems and differences are not consistently replicated. This contrasts with studies of correlations of this quantity, which are in good agreement with contact networks.

In Figure 2.15B we report the free-energy landscape in the PC1 and PC2 axes using backbone dihedral angles as features. Here, the frames of apo and prfar are conveniently separated. However, the separation is not entirely perfect and some frames of prfar have a negative PC1, which is typical of frames of apo. The landscapes computed separately for each replica shows that the simulation prfar1 actually differs much from the

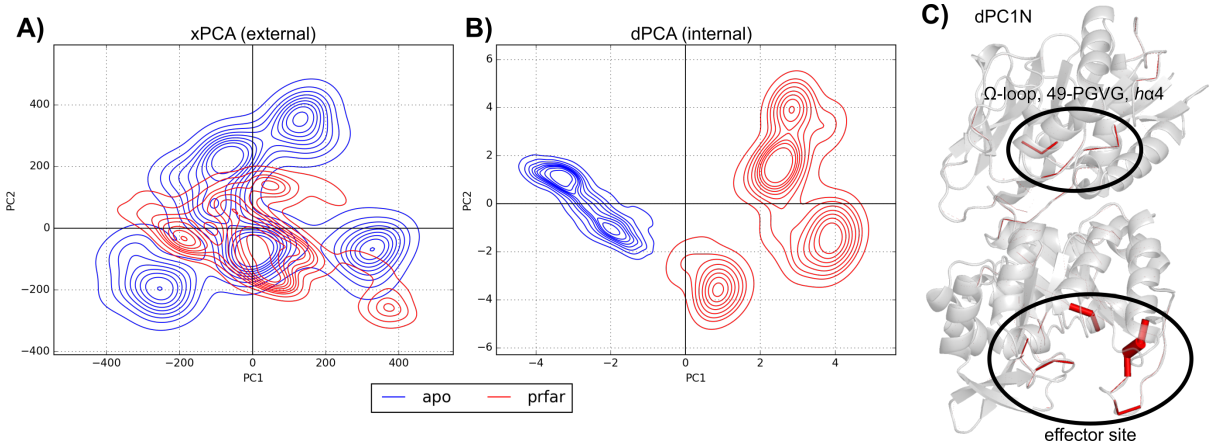


Figure 2.15: 2D projection of the free-energy landscape in the PC1 and PC2 axes computed separately for simulations of apo (blue) and prfar (red) using atomic displacements (A) and backbone dihedral angles (B) as features. (C) Backbone-dihedral angles PC1N. Edge width is proportional to the weight. No threshold is used, if edges are not visible it is because they are too small to be displayed.

other simulations of prfar and in some aspects is even closer to frames of apo. Overall, the backbone dihedral angles, despite providing an approximately satisfactory separation between frames of apo and frames of prfar, extract a less perfect allosteric signal. In Figure 2.15 C, we report the most important pairs of covalently bound residues that participate the most in the PC1 eigenvector. The technical details behind dPCA are a bit different from cPCA because from two circular features we produce four linearized features for each residue pair. The sign in the eigenvector thus makes less sense, and to avoid compensatory effects, the influence of the pair of residues is considered to be the square root of the sum of squares of the influence of the four features. Because of this, PC1N using dihedral angles has only positive values. The pairs of residue which shows the most variation are predominantly found in the effector site (notably in loop1) and near the active site in the Ω-loop, 49-PGVG and the $h\alpha 4$ helix. Very interestingly, this study shows elements of the allosteric pathways that change in flexibility upon binding of the effector (loop1, 49-PGVG, the Ω-loop) and local (un)folding in the $h\alpha 4$ helix. This picture is much cleaner than the PC1N or the DPCN because the number of features is fixed ($N_{\text{residue}} - N_{\text{chain}} = 451$) and much lower than the number of contacts. Moreover, the square of sums may have the effect to strongly diminish small influences. In fact, here we show that the dPCA can complement cPCA by focusing on allosteric events with highly disordered dynamics. However, the dPCA is blind to the propagation of perturbation between loop1 and the Ω-loop because they involve more ordered motions, such as the breathing motion (a rigid body motion) or sidechain dynamics (the salt-bridge network alteration).

In Figures 2.16 A-B, we show the free energy landscapes of PCA using the contact frequency and transformed contact distances. The two landscape perfectly separates the apo and prfar simulations on the PC1 axis, and thus corresponding PC1N are representative of the allosteric contact signal. The two landscape have very similar shapes with apo simulations extending to extremal negative and positive values of PC2 while prfar simulations are much narrower in PC2 evolution and are oscillating around zero. Interestingly, this shape is rather different from the energy landscape cPCA in which both the apo and prfar simulations evolved along the PC2 axis with large amplitude, and prfar has the largest. This suggests that fPCA and tdistPCA share more similarities together than cPCA.

In Figure 2.16C-D, we show the corresponding fPC1N and tdistPC1N. The two networks share many similarities together and with the cPC1N. A minor difference is that tdistPC1N seems a little more dense, with low-value edges adding noise. The two networks show a high density of loss of contact at the interface, which is characteristic of *breathing motion*. The correlation plots between the edge weights of cPC1N, fPC1N and tdistPC1N are reported in Figure 2.17 and show that, in fact, tdistPC1N and fPC1N are the most correlated together ($R^2 = 0.77$) and cPC1N and fPC1N are the least correlated ($R^2 = 0.51$). This suggests that the average interatomic contact number possess some peculiar information compared to the frequency of contact or the transformed distance.

In Figure 2.16C the CCA applied to the fPC1N shows 5 components for 97 edges and 95 nodes. Two key elements of the allosteric pathways are not found inside fPC1N: the vast reorganization of loop1 and the alteration of the salt bridge network between the $f\alpha 1$ and $f\alpha 2$ helices. However, two elements of the allosteric pathways are emphasized: the large motions at the interface due to the breathing motion and the propagation of perturbation from the $h\alpha 1$ helix to the Ω-loop and to the PGVG sheet. With the amplification, many contact redundancies are found, which explains the low number of components (5) by comparison to the number of edges (97). This suggests that fPC1N is particularly good at detecting large conformational changes that affect the distance between elements of the secondary and tertiary structure. However, it is less able to detect more subtle sidechain reorganization such as the salt bridge network between $f\alpha 1$ and $f\alpha 2$ or even large motions within a secondary structure element as in loop1.

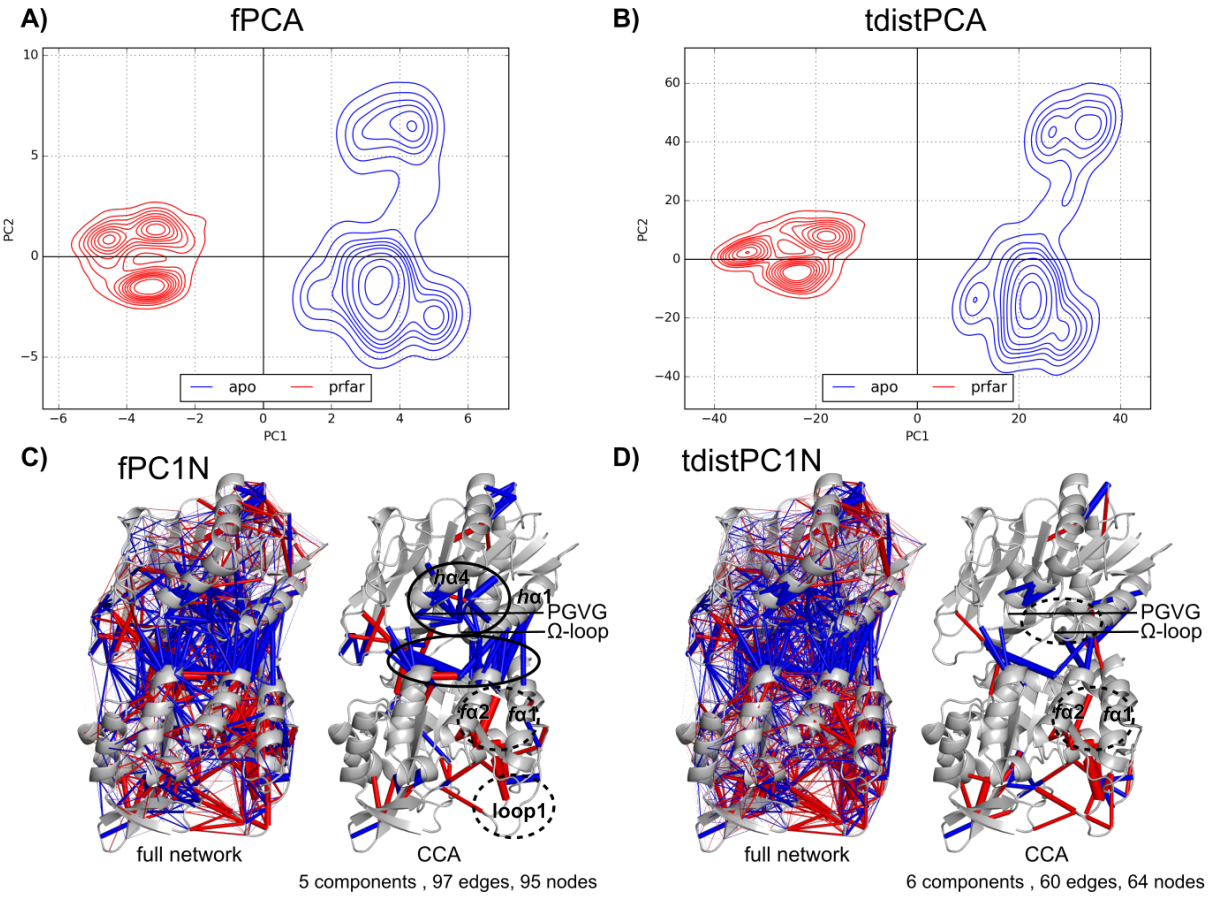


Figure 2.16: 2D projection of the free-energy landscape in the PC1 and PC2 axes computed separately for simulations of apo (blue) and prfar (red) using contact frequency (A) and transformed contact distances (B) as features. frequency-(C) and transformed distance-(D) PC1N with and without CCA. Edge width is proportional to the weight.

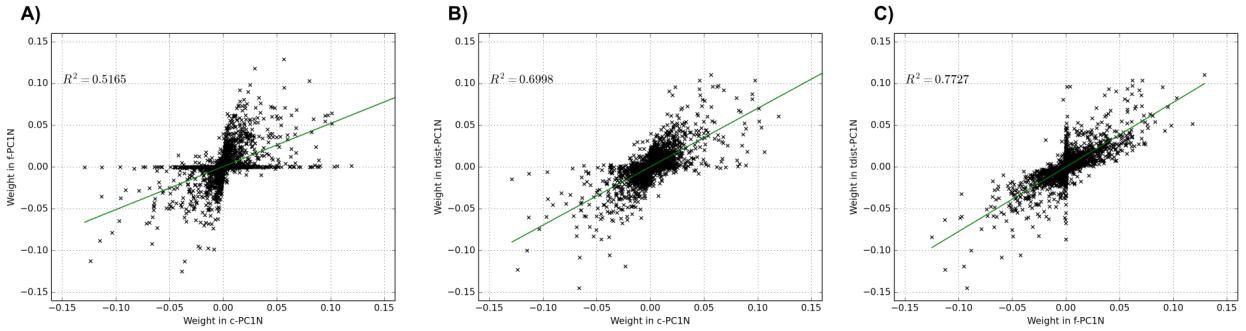


Figure 2.17: Correlation plot between weights in the cPC1N and fPC1N, (A) cPC1N and tdistPC1N (B), fPC1N and

In Figure 2.16C the CCA applied to the fPC1N shows 6 components for 60 edges and 64 nodes. With fewer edges, this network shows better the reorganization occurring in loop1 but is still unable to display the salt-bridge network alteration, does not emphasize the contact losses at the interface, and loses the propagation of perturbation from the $h\alpha 1$ helix to the Ω-loop and to the PGVG sheet. This network contains fewer edges, which are less redundant and more components than the fPC1N. This shows that the tdistPC1N is more precise in terms of contact importance, but this comes at the cost of producing components which are less able to show local-to-global contact perturbations.

The fPC1N can be seen as a cPC1N in binary and thus contains less information. In fact, in Figure 2.17 we see that some edges with almost no importance in the fPC1N have a strong importance in the tdistPC1N and the cPC1N. This suggests that the fPC1N is blind to some events. In fact, the number of interatomic contacts can fluctuate without the contact strictly breaking, and these events are captured by the cPC1N or the tdistPC1N but not by the fPC1N. Moreover, any contact breaking (even if loose) is weighted the same in the fPC1N, explaining why many edges are redundant. The tdistPC1N is able to capture more subtle information about a contact and its strength, but in fact, we only compute one aspect of the contact shape (the closest-heavy

distance) and a contact could still be fluctuating but keeping a similar closest-heavy distance. In the end, fPC1N and tdistPC1N show relevant information but are indifferent to some subtleties of the complex nature of contacts.

References

[1] Yuguang Mu, Phuong H Nguyen, and Gerhard Stock. "Energy landscape of a small peptide revealed by dihedral angle principal component analysis". In: *Proteins: Struct. Funct. Bioinform.* 58.1 (2005), pp. 45–52.

[2] Ivan Rivalta et al. "Allosteric pathways in imidazole glycerol phosphate synthase". In: *Proc. National Acad. Sci.* 109.22 (2012), E1428–E1436.

[3] Berk Hess. "Similarities between principal components of protein dynamics and random diffusion". In: *Phys. Review E* 62.6 (2000), p. 8438.

[4] Aria Gheeraert et al. "Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks". In: *The J. Phys. Chem. B* 123.16 (2019), pp. 3452–3461.

2.4 From Amino Acid Networks to Chemical Group Networks

2.4.1 The chemical nature of contacts

The formation of the contacts is the driving force behind the shapes of protein, and contact changes explain why protein change their shapes. Generally, we separate residues that are polar and electrically charged (arginine, lysine, aspartic acid, glutamic acid, sometimes a protonated histidine) from uncharged polar residues (generally serine, threonine, asparagine, glutamine, cysteine, and histidine) and hydrophobic residues (all the others). In water, because "birds of a feather flock together", hydrophobic residues tends to be buried within the protein while polar residues faces water[1]. Hydrophobic residues then cluster within the protein while mostly hydrogen bonds, salt bridges (i.e., contacts between oppositely charged residues) stabilizes the rest of the folding. Some interactions involve the π resonance of aromatics, usually classified as hydrophobic, such as π-polar, π-stacking, or π-cation interactions, which tempers the assertion that hydrophobic residues are buried within the protein. More generally, residues can possess both hydrophobic and polar moieties, such as lysine which has a long hydrophobic chain before its charged head, or tyrosine which possess an alcohol function bound to its aromatic ring. This further tempers a strict categorization of residues in strict categories. Thus, it is then more appropriate to dissect residues in chemical groups which are themselves either hydrophobic or polar. In fact, changes in contact between these chemical groups are the cornerstone of protein structuration and dynamics.

The regular AAN and DPCN methodologies compute only contacts between residues, and as such completely neglect the chemistry of contacts and the chemical groups. Moreover, when a contact change of nature between two residues between a *perturbed* and *reference* AANs, it is entirely possible that this is not displayed in the corresponding DPCN which suggests that a huge part of the reality of the contact network is hidden by the simplification behind AAN. This is a serious limitation of AANs which leads to our development of Chemical Group Networks (CGNs), that is, networks representing contacts in proteins using a smaller coarse-grain description of the proteins with its chemical groups.

2.4.2 Other methodological development and applications

Chemical Group Contact Networks

In AAN, a protein is represented as a collection of amino acids (nodes) linked by a certain quantity, such as the number of interatomic contacts. In CGNs, the protein is instead represented as a collection of chemical groups (nodes), and we analogously use a contact condition to link and weight those chemical groups. This kills two birds one stone because, at the same time, we gather information about the chemistry of a contact, and such networks can detect when a contact changes in nature between two residues.

Here, we present two ways of dividing the system that are incremental: the first simply separates each residue in its backbone and its sidechain. Except when specifically mentioned, hydrogen atoms are discarded from the analysis. In IGPS, backbone heavy atoms account for 1,816 atoms and sidechain heavy atoms for 1,762 atoms. We call this decomposition the Two Group Network (2GN). The second decomposition further decompose the sidechain in a hydrophobic and polar parts. We define the hydrophobic part as carbon atoms that are not covalently bonded to an oxygen or a nitrogen (that is, only bonded to carbon, hydrogen, or sulfur atoms). The polar part is defined as the carbon atoms bonded to oxygen or nitrogen, as well as oxygen, nitrogen, and sulfur atoms. This hydrophobic selection account for 1,086 atoms, while the polar selection accounts for 676 atoms

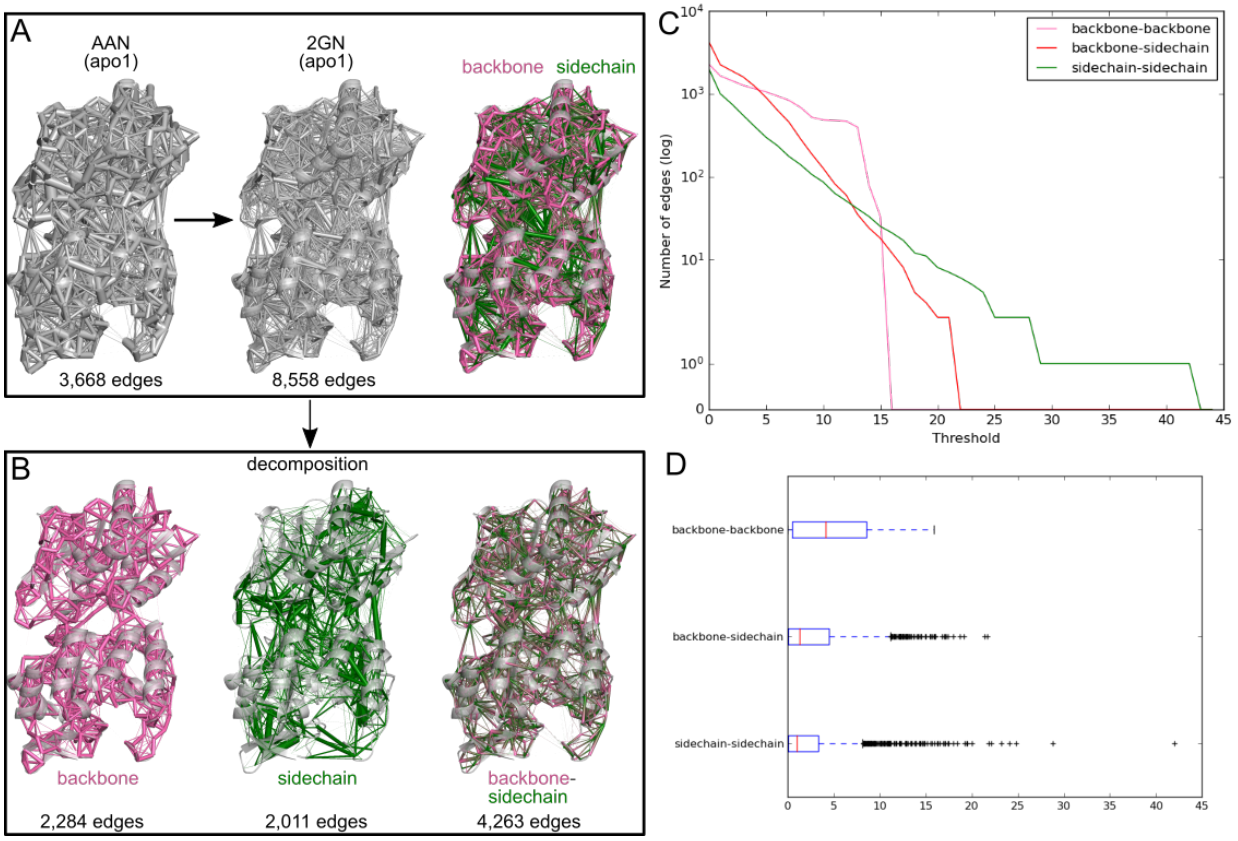


Figure 2.18: A) Comparison of the projections on the 3D structure of IGPS from *T. maritima* of the average AAN built with a 5 Å cutoff on heavy-atoms and the average 2GN built using the same contact condition. The AAN network endpoint for edges are centered on the C_α and for the 2GN, to split visually contacts, the center for backbone endpoints is located on the C_α and for sidechain endpoints, located on the backbone C. Edge width is proportional to the weight of the contact. A second coloring scheme is shown for the 2GN where backbone endpoints are represented in pink, sidechain in green with edges colored in plain if the contact is between two endpoints of the same chemical group and with a gradient between the two endpoints if they are of a different nature (backbone-sidechain). B) 2GN broken down into three components: backbone-backbone contacts, sidechain-sidechain contacts and backbone-sidechain contacts with the corresponding coloring scheme. C) Number of edges in function of the used threshold for each group of contact. D) Boxplot of the distribution of each type of contact in the 2GN. The median is shown with a read line and the blue box delimits the the first and third quantiles. Whiskers extends to 1.5 times the interquartile range (i.e. distance between first and third quantile). Outliers are marked with a black cross.

in IGPS. Theoretically, each contact in an AAN can be split into four and nine components, respectively, in the 2GN and 3GN and there are, respectively, three and six types of contact possible between chemical groups. Because intraresidual contacts are discarded in AANs, during network building, we also discard contacts inside the chemical groups of the same residue. After building CGNs, we can perform the same set of analysis developed for AANs, such as DPCN or PC*i*N.

Adaptation of connected component analysis

The process of regular connected component analysis to clean a congested network was developed for DPCN built with AANs but because chemical groups are disjointed in 2GN and 3GN, a contact change involving two different groups of the same amino acid are not connected and can belong to different Connected Component (CC). Therefore, we also propose a variation of the CC Analysis (CCA) strategy, in which all groups belonging to the same residue are artificially linked together. In theory, this can modify both the threshold that maximizes the number of CCs and the structure of the final core CCs. Because we add *artificial* edges, this generally reduces the number of *artificial* CC, which is different from the number of *true* CC.

Chemical Group Networks

In Figure 2.18A-B, we represent in the IGPS 3D structure a projection of the average AAN and 2GN during the simulation apo1. The 2GN contains 2.4 times more edges than the AAN (respectively 8,558 and 3,668 edges). Each contact can be decomposed into four new contacts in the 2GN; thus, the theoretical upper limit for an

increase is a factor of four. In this regard, a 2.4-time increase is actually impressive and suggests that most contacts in an AAN have more than one component in the 2GN description. Of these 8,558 edges, 2,284 are in the backbone network (27%), 2,011 in the sidechain network (23%), and 4,263 in the backbone-sidechain network (50%). Backbone-sidechain edges are twice as common, notably because between two residues there is two possible way to establish a backbone-sidechain contact whereas there is only one way to establish a backbone or a sidechain contact. Therefore, this suggests a good balance between the four possible decomposition of each AAN contact in the 2GN.

The backbone network is principally centered around covalent bonds. In fact, the network contains a continuous line of sizable edges in each chain, which is the protein backbone. Other contacts are principally located within the α helices, between β sheets on a β strand or inside loops, which is consistent with the fact that the backbone contacts are principally involved in the secondary structures of proteins. Interestingly, there are a few sizable backbone-backbone edges at the interface between HisF and HisH, suggesting that backbone contacts can be involved in interfaces. The sidechain network is completely different from the backbone one. Whereas backbone edges follow the structuration of the protein, sizable edges in the sidechain network are much more diverse, sometimes present inside α helices, between different secondary structures (mostly between α helices and loops), or at the interface, but with no precise pattern. This suggests that sidechain contacts have an influence in the arrangement of secondary structures and chains i.e. the ternary structure, as expected. Some of them are present at the extremities of α helices, thus we can hypothesize that they are also sometimes involved in the interruption of these helices. The backbone-sidechain network is also very diverse and does not follow regular patterns. Edges of this network are found in almost every possible context with decent magnitudes (covalent, within a secondary structure, between secondary structures, at the interface). The precise role of this network remains elusive and is probably hard to grasp simply because of the large number of edges.
In Figure 2.18C we represent the number of edges in each decomposed network in function of a *threshold* applied to the global 2GN. The evolution of the backbone network is quite interesting because the decrease in number of edges is very slow at first, but around threshold 15, the number of edges abruptly decreases and reaches zero. This suggests that a lot of contacts in the backbone network have a weight centered around this 15 value which may be a practical limit of the atoms that backbone can share. This huge drop starts around 700 edges, but since there are only 451 covalent bonds in IGPS, this significant decrease cannot be explained solely by covalent bonds. This suggests that other types of backbone contacts can reach similar weights. The sidechain curve has a much different trend. Starting with the lowest number of edges, it gradually decreases and surpasses the backbone-sidechain around a threshold of 13 and the backbone network after its significant drop around 15. Then it remains the network with the largest number of edges until the end and thus possesses edges with the largest weight. This is consistent with the fact that in the sidechain network, some edges truly stand out, while the majority of edges have a low weight. This shows a fundamental differences between backbone edges where large-value edges are centered around the same value and sidechain edges where they only a few outliers. The backbone-sidechain network also possess a steady decrease which is faster than the decrease of number of edges in the sidechain network. Interestingly, this shows that the backbone-sidechain network has a behavior in between the backbone and sidechain one. The vanishing point of this network (i.e. the threshold at which the network becomes empty) is lower than the vanishing point of the sidechain network despite containing twice as many edges. This suggests that contacts established by backbone groups are more limited than contacts established by the sidechain which can be much larger and are more flexible in contact creating. In Figure 2.18D we represent the distributions of each type of contact as a box plot. This further proves that the backbone network is the most uniform, as it contains no outlier edges. By contrast, the backbone-sidechain and sidechain are increasingly less uniform and the sidechain network contains mostly outliers.

In Table 2.1 we report the top ten contacts in each 2GN decomposition. In the top ten for the backbone network, nine are representative of covalent bonds, while one is a backbone hydrogen bond between two β sheets. All these contacts have weight between 15.5 and 15.9. This indicates a practical upper limit to backbone contacts: in fact, since all backbone nodes have the same shape and size (with the notable exception of proline), they cannot exceed this limit. Interestingly, the only link between two amino acids not in covalent interaction is between a glycine and an alanine, the two amino acids with the smallest sidechain: this suggests that it is the smallness of their respective sidechain that allows sharing so much contact between their backbone and that covalent backbone-backbone contact usually slightly dominates other types of contact, notably those within α-helices and between β-sheets. Their ranks in the global 2GN are between 37 and 50, which shows that they are the smallest of the top contacts, but they are still largely represented overall in the top contacts of the complete 2GN.

Interestingly, among the backbone-sidechain contacts, most are still between residues that are covalently bound (7 out of 10), one is within a α helix and two between β sheets in a β strand. Contrary to backbone-backbone contacts, these cannot simply be explained by a covalent bond between the residues. The sidechains involved are disproportionately aromatic compounds (9 out of 10). This suggests that these interactions are principally π-polar interactions between an aromatic and the backbone of a protein. The weights in this network are between 16.7 and 21.7, showing more spread than in the backbone network, and the ranks are between 8

backbone-backbone					backbone-sidechain*				
residue 1	residue 2	rank	weight	attribution	residue 1	residue 2	rank	weight	attribution
hL105	hI106	37	15.851	covalent	hH120*	hM121	8	21.662	covalent
hT142	hY143	38	15.844	covalent	hE77	hR78*	9	21.428	covalent
fD45	fE46	39	15.825	covalent	fK206	fH209*	14	19.155	α-helix
hK169	hG170	40	15.813	covalent	hG122	hY138*	15	18.697	β-strand
fD219	fA220	41	15.803	covalent	hG174	hF175*	17	17.99	covalent
hG154	hA166	42	15.718	β-strand	hY138*	hF139	20	17.435	covalent
fD98	fK99	43	15.663	covalent	hF128*	hF132	22	17.185	β-strand
hL90	hF91	46	15.607	covalent	hY158*	hD159	23	17.112	covalent
fE231	fI232	47	15.594	covalent	hY136	hY137*	26	16.98	covalent
fS180	fG181	50	15.522	covalent	hN26	hF27*	28	16.788	covalent

sidechain-sidechain				
residue 1	residue 2	rank	weight	attribution
fR249	hW123	1	42.015	hinge, π-cation
hY79	hR171	2	28.812	π-polar
hY137	hF177	3	24.825	π-π
fR133	fF138	4	24.112	π-cation
hE96	hY143	5	23.181	
fF138	fF189	6	22.111	π-π
hR114	hE161	7	21.836	salt-bridge
fF86	fQ115	10	20.028	π-polar
hR2	hY43	11	19.493	π-cation
fD45	fR249	12	19.432	salt-bridge

Table 2.1: Top 10 contacts for each group of contacts with the corresponding rank, weight and attribution. For backbone-sidechain contacts, an asterisk (*) indicates which endpoint is the sidechain.

and 28, which suggests that they are decently weighted in the 2GN.

Finally, in the sidechain-sidechain network are found the seven biggest edges of the whole 2GN and the 10th, 11th and 12th. This suggests that sidechain contacts are the ones which have the potential to be the biggest, but overall this phenomenon remains quite rare. In fact, there is a huge disparity in contact size, from 19.4 to 28.8 excluding the first, while the most important is at 42. Very interestingly, the biggest contact is fR249–hW123, a contact well identified in IGPS, named the *hinge*, which is located at the interface between hisF and hisH. It is known to be the principal contact of the interface and the junction of the *breathing motion*, but here we show that it is even shown to be the strongest contact in the whole protein by a large amount. The fact that it is such larger than the others raises two questions: a) can interface contacts be much larger than intradomain contacts because not being part of the same chain gives more lability to the sidechain? and b) does the contact value grow linearly? This requires additional studies on large datasets of MD simulations of proteins. Other contacts are all attributable to interactions between aromatic, polar, charged and salt bridges. None of the contacts are between two ILV (isoleucine, leucine and valine) residues, which are some essential hydrophobic contacts. This suggests that using 2GN, hydrophobic residues are significantly underweighted.

In Figures 2.19 A-B, we represent the 3GN projected on the 3D structure of the protein and the breakdown of each subcomponent. By design, compared to 3GN, the backbone network does not change at all, but the backbone-sidechain and sidechain-sidechain are broken down into five new possibilities. In the 2GN the sidechain network contained 2,011 edges which can theoretically be decomposed into four here, the hydrophobic network contains 1,430 (71%), the polar network only 628 (31%) and the hydrophobic-polar one 1,611 (80%). This shows that this decomposition is more specific than the previous one and not so many new components are created by this process. The effect holds for the decomposition of the backbone-sidechain network, since the 4,263 edges are divided into 3,588 edges (84%) in the backbone-hydrophobic network and 2,067 in the backbone-polar network (48%).

Although the backbone network remains the same in 3GN and 2GN, other networks are very different. In the hydrophobic network; most of the edges are located inside the protein structure and in very dense clusters, which is in good agreement with the fact that hydrophobic residues are usually buried in the protein and form clusters. Still, a few edges are present at the interface, and some sizable edges are even found on the exterior of IGPS. This could be explained in part by the fact that aromatic rings are considered hydrophobic in our model, despite having favorable interactions with water and other aromatic rings. Some of these contacts could simply be π-π interactions. Then, the polar network is much smaller and is complementary to the hydrophobic network, most edges are located external to the protein. The interface is the region where polar contacts are the densest. In the backbone-hydrophobic network, most of the contacts are also found buried in the protein

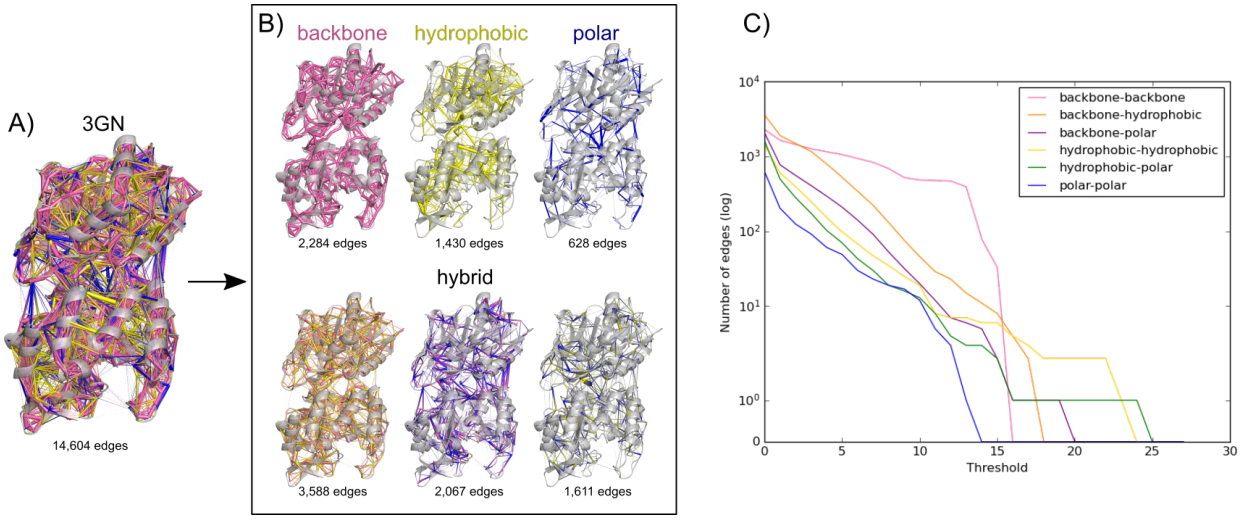


Figure 2.19: A) Projections on the 3D structure of IGPS from *T. maritima* of the average 3GN built with a 5 Å cutoff on heavy-atoms. To split visually contacts in the 3GN, the center for backbone, hydrophobic and polar endpoints are located respectively on the backbone C_α, C and N atoms. Edge width is proportional to the weight of the contact. The coloring scheme of edges represents backbone endpoints in pink, hydrophobic in yellow and polar in blue with edges colored in plain if the contact is between two endpoints of the same chemical group and with a gradient between the two endpoints if they are of a different nature (backbone-sidechain). B) 3GN broken down into six components: backbone-backbone contacts, hydrophobic-hydrophobic contacts and polar-polar contacts (*pure* contacts) and backbone-hydrophobic, backbone-polar and hydrophobic-polar contacts (*hybrid* contacts) with their respective coloring scheme. C) Number of edges in function of the used threshold for each group of contact.

and only a few are found at the interface or externally. By contrast, in the backbone-polar network, many links are found at the interface and external to the protein. Interestingly, at the interface, the vast majority of backbone endpoints are located in HisF, whereas most of the polar endpoints are located in HisH. It would be interesting to study this effect on a bigger variety of interfaces. Finally, the hydrophobic-backbone network is the smallest of all *hybrid* networks. Furthermore, the majority of edges are quite small, and only a handful are sizable. Of these sizable edges, most are located externally to the protein, and the largest are located at the interface. This principally suggest that these contacts are actually interactions between an aromatic ring and a polar or charged residue, which are only quite few.

In Figure 2.19C we represent the evolution of the number of edges in each subnetwork with the threshold. Again, the evolution of the backbone network is the same as in 2GN. In fact, interestingly, all the curves except the backbone network have a similar evolution and keep with a slight decrease until threshold 15 where the backbone network has a significant decrease. The polar network is a bit different from the other because it reaches zero around threshold 13. In contrast, in the 2GN, the salt bridges (which are polar-polar contacts) were decently weighted. This can be explained by the fact that here, only polar heads of the sidechain are considered in interaction. This suggests that these types of contact are underweighted in the 3GN. All the other types of contact possess elements which have a weight bigger than 16 (i.e., the biggest backbone edge) and thus are probably more accurately weighted.

In Table 2.2 we report the top ten contacts for each possible type of contact. The backbone network does not change except for contact ranks, now between the 12th position and the 26th. By breaking down sidechain elements into different parts, the backbone contacts have a relatively larger weight (i.e., compared to other contacts). Still, this type of contact never reaches the first position, indicating that the 3GN does not significantly overweight backbone contacts compared to the other types of contact. In the hydrophobic network, the top six contacts are among the overall strongest (six are between rank 2 and 20, weights from 15.6 to 23.1), but the four next are much lower (between rank 328 and 553, weights from 10.7 to 13.5). Five of the contacts are π-π interactions between aromatic rings, and five are hydrophobic contacts between an aromatic ring (four times a phenylalanine and one time a tyrosine) and an ILV residue (that is, isoleucine, leucine, or valine). Despite the fact that two types of contact we expect to detect with this weighting are detected, no hydrophobic contact is detected only between ILV residues. There are two hypotheses to explain this: a) contacts with aromatic rings are overweight, and b) ILV contacts are underestimated. The comparison with other types of contacts suggests that it is the second case. Contacts in the polar network have a similar problem. The top 10 contacts represents the many types of polar contact that can coexist: seven are salt-bridges, two are hydrogen bonds and one is the polar contact inside the catalytic triad (hH178–hE180, the stabilization of protonated histidine during the mechanism is critical for catalysis). However, their weights and ranks are the lowest of the top

backbone-backbone					hydrophobic-hydrophobic				
residue 1	residue 2	rank	weight	attribution	residue 1	residue 2	rank	weight	attribution
hL105	hI106	12	15.851	covalent	hY137	hF177	2	23.084	π-π
hT142	hY143	13	15.844	covalent	fF138	fF189	3	22.111	π-π
fD45	fE46	15	15.825	covalent	hF175	hF177	6	17.42	π-π
hK169	hG170	16	15.813	covalent	fI93	fF120	10	16.214	hydrophobic
fD219	fA220	17	15.803	covalent	hF139	hF177	11	15.947	π-π
hG154	hA166	18	15.718	H-bond	fF210	fL237	20	15.662	hydrophobic
fD98	fK99	19	15.663	covalent	hF128	hF132	328	13.504	π-π
hL90	hF91	22	15.607	covalent	hF47	hV81	535	11.15	hydrophobic
fE231	fI232	23	15.594	covalent	fF214	fV246	546	10.91	hydrophobic
fS180	fG181	26	15.522	covalent	hY17	hL34	553	10.758	hydrophobic
polar-polar					backbone-hydrophobic*				
residue 1	residue 2	rank	weight	attribution	residue 1	residue 2	rank	weight	attribution
hR114	hE161	416	13.007	salt-bridge	hG174	hF175*	5	17.99	covalent
hR144	hE161	490	12.706	salt-bridge	hF128*	hF132	7	17.185	π-polar (turn)
fR16	fD28	492	12.678	salt-bridge	hN26	hF27*	8	16.788	covalent
hR78	hR201	516	11.911	H-bond	hY136	hY137*	9	16.403	covalent
hE56	hR59	534	11.163	salt-bridge	hY158*	hD159	14	15.842	covalent
fR235	fE251	545	10.925	salt-bridge	hY138*	hF139	24	15.578	covalent
fN25	fR27	548	10.788	H-bond	hT131	hF132*	25	15.529	covalent
fD74	hR22	549	10.765	salt-bridge	hT142	hY143*	44	15.187	covalent
fD45	fR249	563	10.589	salt-bridge	hL42	hY43*	53	14.938	covalent
hH178	hE180	574	10.322	catalytic triad	hI127	hF128*	64	14.487	covalent
backbone-polar*					hydrophobic-polar*				
residue 1	residue 2	rank	weight	attribution	residue 1	residue 2	rank	weight	attribution
hH120*	hM121	4	19.071	covalent	fR249*	hW123	1	24.556	hinge, π-cation
fK206	fH209*	21	15.609	H-bond	fR133*	fF138	49	15.067	π-cation
hE77	hR78*	62	14.561	covalent	hL46	hH73*	75	14.272	
fA224	fH228*	80	14.133	H-bond	hF54	hS94*	507	12.185	π-polar
fQ72	hR22*	81	14.13	H-bond	hY79	hR171*	513	11.941	π-polar
fI42	fR235*	341	13.465	H-bond	fF141	fN148*	521	11.696	π-polar
hH73*	hR78	504	12.243	H-bond	hR2*	hY43	529	11.382	π-polar
fG181	fH209*	517	11.859	H-bond	fF86	fQ115*	531	11.251	π-polar
fM1	hN124*	522	11.684	H-bond	fR191*	fI198	542	10.974	
hH120*	hG122	525	11.523	H-bond	hF139	hH141*	543	10.968	

Table 2.2: Top 10 contacts for each group of contacts with the contacts with the corresponding rank, weight, and attribution. For hybrid contacts, an asterisk (*) indicates which endpoint corresponds to which group (also labeled in the group of contact).

of any subnetwork (ranks between 416 and 574, weights between 10.3 and 13). In fact, backbone groups and hydrophobic aromatic groups generally possess less hydrogen atoms per heavy atom (-CH, -NH at best and even no hydrogen atom carried) than some polar heads or hydrophobic aliphatic compounds (-NH2, -NH3+, -CH2, -CH3). Therefore, the slight underweighting could be explained by an absence of hydrogen counting.

In the *hybrid* networks, the backbone-hydrophobic shows the top ten edges that are well balanced with weights between 14.4 and 18 and ranks between 5 and 64. The top ten contacts of this subnetwork are between an aromatic ring and a backbone, which suggests π-polar interactions. Of the ten contacts, nine are between residues that are covalently bound. Since the sidechain and the backbone are not directly bound, the presence of strong contact cannot be simply explained by a covalent bond, but is probably favored by this fact. The other contact (which is actually top two of the subnetwork and top seven of the full network) is found between the extremities of a turn. The backbone-polar network is more disparate in the top ten contacts. The fifth strongest in this subnetwork have ranks between 4 and 81 in the full network and weights between 14.1 and 19.1 but the fifth next have ranks between 341 and 525 and weights between 11.5 and 13.5. Two of the contacts can be attributed to covalently bound residues, and all others are sidechain-backbone hydrogen bonds. Finally, in the hydrophobic-polar network, we find the top interaction of the network, which is the aforementioned π-cation *hinge*. In fact, of the top ten hydrophobic-polar contacts, the two biggest are π-cation interactions (ranks 1 and 49, weights 15.1 and 24.5) and the top four to eight are π-polar interactions (ranks between 507 and 531, weights between 11.2 and 12.2), so most hydrophobic-polar contacts can be explained by the ambivalent role of aromatic rings. The precise nature of the three other contacts remains elusive, and they cannot be explained simply.

Principal Component Analysis of Chemical Group Networks

In Figure 2.20 we report PC1N obtained with two groups (PC1-2GN) and three groups (PC1-3GN) on the whole set of simulations (apo1-4, prfar1-4). They are very similar in appearance, which is not surprising, since they are merely a decomposition of the complete AAN. The subdivision of each subnetwork clears the overall dense picture. This is especially true for the backbone-backbone network in Figure 2.21 which shows that most of the backbone-backbone perturbations are found near the effector site and especially in loop1, between the Ω-loop and PGVG and within the $h\alpha 4$ helix. This is consistent with previous studies of the allosteric mechanism; loop1 undergoes hydrogen bond rearrangements upon PRFAR binding, a backbone hydrogen bond breaks between the Ω loop and PGVG and gives more flexibility to the PGVG segment while the $h\alpha 4$ helix unfolds upon PRFAR binding. In fact, the overall clear picture of the backbone-backbone network suggests that backbone perturbations are fewer. Indeed, a good portion of the most important contacts are found inside secondary structures (which are difficult to perturb) and between covalent residues (which are even harder to perturb). By contrast, the sidechain and sidechain-backbone subnetworks remains quite congested, which indicates that the decomposition of the sidechain in two in the 3GN decompose more efficiently the information than in the 2GN. In the pure sidechain network, rearrangements are found that occur between $f\alpha 1$, $f\alpha 2$ and $h\alpha 1$, which is consistent with the fact that they are part of a salt bridge network alteration. Some other contact changes are detected, but more difficult to attribute. In the sidechain-backbone network, two main areas show a high density in terms of contact change: loop1 and the interface. However, the 2GN description remains quite limited and the subnets quite congested for a more detailed analysis.

The backbone-backbone network in the PC1-3GN decomposition is very similar to that obtained in the PC1-2GN. Only some edges have slightly larger weights in PC1-3GN. In fact, because sidechain groups are divided in smaller groups in the 3GN, the weight of these new types of contacts are smaller or equal than in the 2GN. This impacts the backbone-backbone contacts, which seem to be comparatively larger or equal. Interestingly, this effect is only small, which shows that this methodology is resilient in the groups we used. The hydrophobic subnetwork posses the majority of its edges being internal to the protein. This effect is more pronounced than in the average 3GN of simulation apo1 in Figure 2.19, suggesting that the few external contacts are resilient (and mediated by π-π interactions). Most of the edges in this decomposition are quite small and cannot be attributed to known allosteric pathways. The polar subnetwork shows much larger edges, and the majority of these edges are external to the protein. In fact, we know that the allosteric mechanism in IGPS from *T. maritima* involves principally external residues. This is consistent with the chemical groups in a protein that tend to be external to accommodate water solvation. The polar-polar network notably shows the salt bridge alteration between $f\alpha 1$, $f\alpha 2$ and $h\alpha 1$.

Next, the backbone-hydrophobic subnetwork contains perturbations in loop1 and near the breathing motion, while the backbone-polar contains perturbations at the interface and near the effector site. This picture is much clearer than the backbone-sidechain network in the PC1-2GN. This restrains the analysis to some definite areas. Finally, the hydrophobic-polar network is probably the one which shows the fewer alterations. Intriguingly, the few large contact changes are between distant parts of the protein: at the interface or between sideR and sideL near the effector site. From the average 3GN study, we know that the largest contacts in the hydrophobic-polar network are, in fact, π-polar or π-cation contacts, and this is consistent with such a long-range contact change.

CCA adaptation to CGN

In Figure 2.21B we report the CCA applied to the PC1N obtained with AAN, 2GN and 3GN. The network do not change substantially in shape, but the information gathered in the 2GN and in the 3GN is compelling, and the type of each contact change is clearly suggested. In particular, only very few edges are redundant between two residues. The threshold that maximizes the number of components in the AAN is of 0.043 (i.e., 4.3% of contribution to the eigenvector). This makes sense because, with more edges, their relative importance in the entire network is diminished. With 2GN and 3GN, this threshold decreases, respectively, to 3.8% and 2.8%. By increasing the detail of the decomposition, the number of components (respectively 11, 12 and 17), edges (respectively 95, 104, and 169), and nodes (98, 106, and 171) also increase. This is reasonable, since the corresponding full networks have also more edges and nodes. In both AANs and CGNs, the core CCs are describing fairly accurately the allosteric mechanism. Two key difference between the three networks however is shown in the connections between the Ω-loop and 49-PGVG and the hN12–hN15 contact loss. These are the last steps of the allosteric mechanism in which a polar contact is broken between two aspartic acids and a backbone hydrogen bond between residue hP10 and hV51 is broken, which allows the 49-PGVG segment to flip in a position where residues hV51 and hL85 forms an oxyanion hole stabilizing the oxyanion intermediates in glutamine hydrolysis. After CCA in the 2GN these two contact changes are absent from the network and in the 3GN only the contacts between 49-PGVG and the Ω-loop are detected and appear clearly as backbone contacts. However, in the complete PC1N, the hN12–hN15 is still much visible, in the sidechain contacts in the 2GN and in the polar contacts in the 3GN. In standard CCA, the chemical groups of the same residue defined in 2GN and 3GN are considered disconnected. This might produce undesirable results in CCA because effects in one

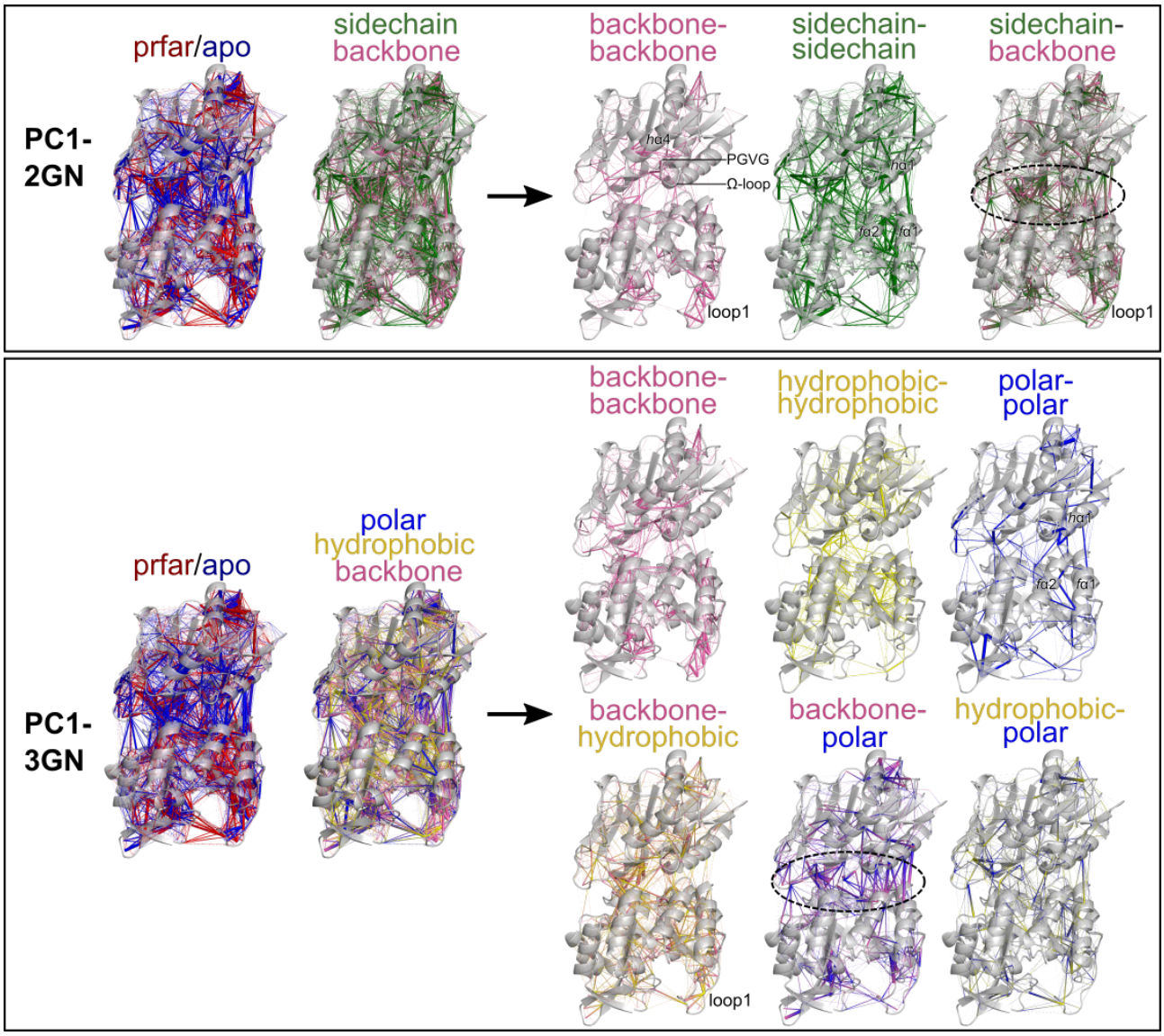


Figure 2.20: PC1N obtained decomposing the system in backbone/sidechain groups (2GN) and backbone/sidechain/polar groups (3GN). Edge width is proportional to the weight of the contact in the eigenvector, and are normalized by the same factor in every figure. To separate edges 2GN and 3GN from different groups located in the same residues, edges endpoint are artificially put at the C_α of the backbone, and the backbone C for the sidechain (2GN) or the hydrophobic group (3GN) and at the backbone N for the polar group (3GN). Each network is first represented with a blue/red color scheme, where blue represents a contact which is typically bigger in apo and red in PRFAR. Next, they are represented with edges that are either gradient or flat colors given the contact type. Backbone endpoints are represented in pink, sidechain in green, hydrophobic in yellow and polar in blue. For clarity, the networks are also broken down into every possible component.

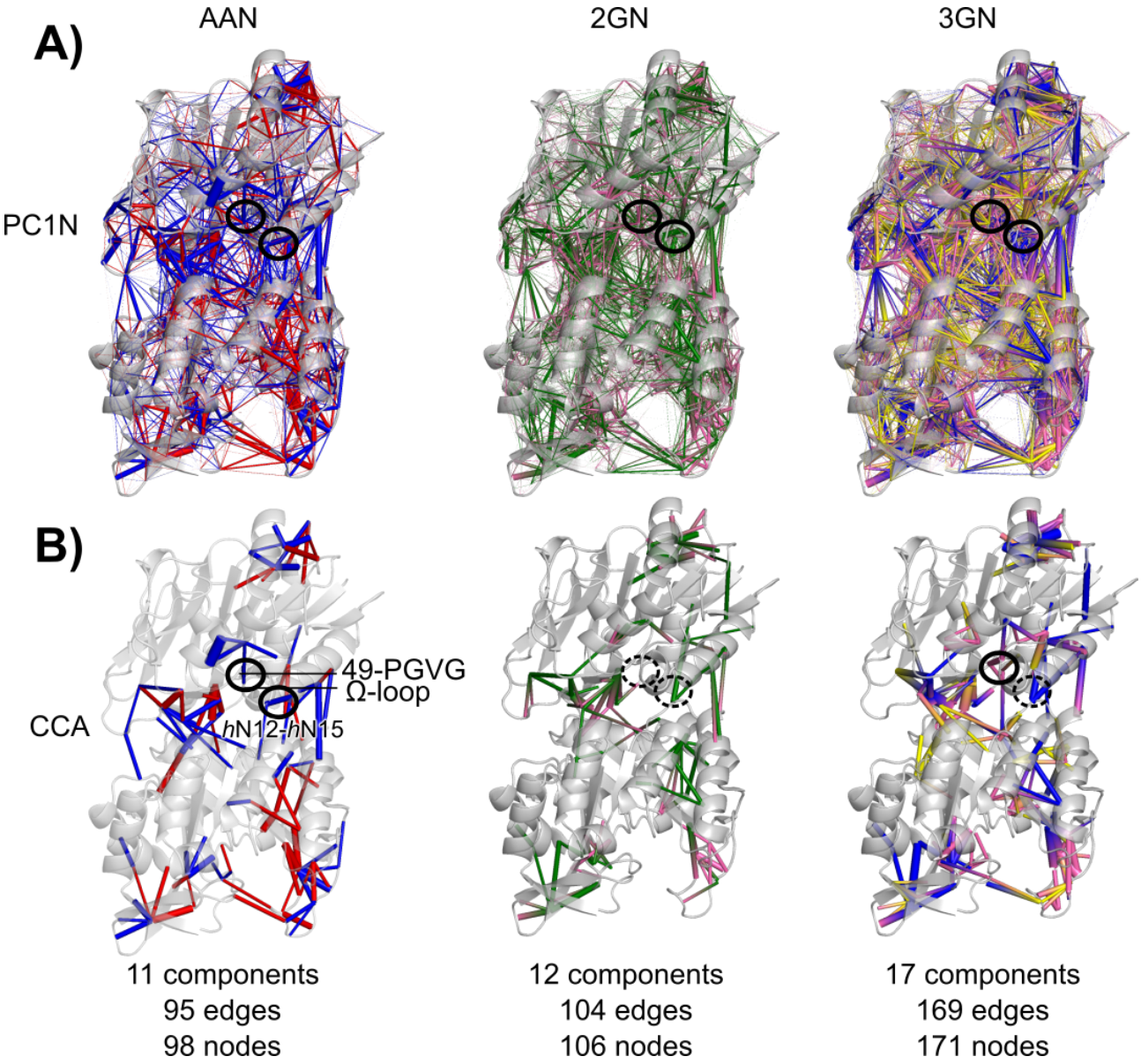


Figure 2.21: (A) PC1N obtained decomposing the system in amino acids (AAN), backbone/sidechain groups (2GN) backbone/sidechain/polar groups (3GN). Edge width is proportional to the weight of the contact in the eigenvector. All edge widths use the same normalization factor. For AAN, the endpoints of the edges are the C_α of each residue. For 2GN and 3GN for simplicity and to separate edges from different groups located in the same residues, the endpoints are artificially put at the C_α for the backbone, and the backbone C for the sidechain (2GN) or the hydrophobic group (3GN) and at the backbone N for the polar group (3GN). (B) Connected component analysis for each PC1N.

chemical group might propagate to a different chemical group. In fact, allosteric effects on the backbone of a residue can even translate into the neighbor residue[2] so this may affect even regular CCA with AAN. We can easily fix the CCA procedure for the "same residue, different group" issue by introducing artificial links between the different chemical groups. However, there is no easy fix for the neighbor residue issue because connecting neighbor residues connects the whole protein, and the number of CC remains one during the procedure.

In Figure 2.22 we show the adaptation of CCA to CGN. While the adapted procedure may in theory suggest using a different threshold value, here thresholds maximizing the number of *artificial* connected components are the same as the number of *true* connected components in both 2GN and 3GN. This suggests that the addition of artificial links to the network does not substantially change the topology of the network. This shows that the adapted procedure produces results that are consistent with regular CCA. For both 2GN and 3GN, the adaptation of cleaning with CCA increases most metrics in the network. In 2GN, the number of *true* CCs grows from 12 to 19 *true* CCs, the number of edges from 104 to 115 edges, and the number of nodes from 106 to 124 nodes. In the 3GN, the *true* number of CC grows from 17 to 42, the number of edges from 169 to 215 and the number of nodes from 171 to 241. Instead, the number of *artificial* CC (that is, CC where intraresidual connections are artificially added) decreases from 12 to 9 in the 2GN and from 17 to 12 in the 3GN. This can vastly simplify the analysis of individual CCs.

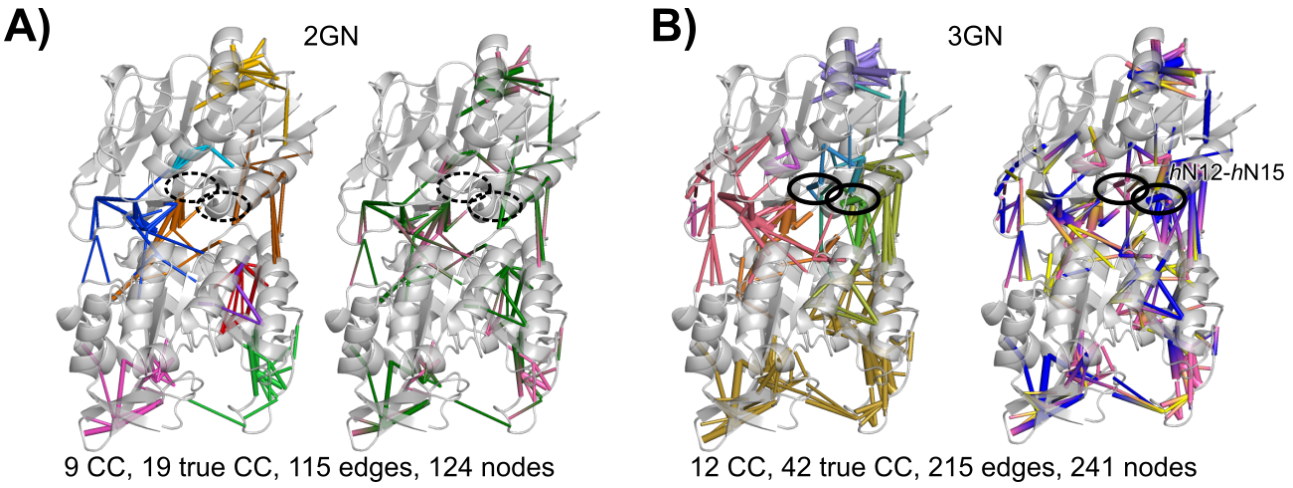


Figure 2.22: (A) PC1N after adapted CCA in the 2GN with a coloring scheme coloring differently each individual *artificial* CC and one coloring the endpoints of edges in green for sidechain groups and pink for backbone groups. (B) PC1N after adapted CCA in the 3GN with a coloring scheme coloring differently each individual *artificial* CC and one coloring the endpoints of edges in yellow for hydrophobic groups, blue for polar groups and pink for backbone groups.

The adaptation of the CCA to 2GN does not allow capture the Ω-loop–PGVG and *h*N12–*h*N15 contact losses. However, in the 3GN, the Ω-loop–PGVG contact is already present without CCA adaptation and with adaptation the *h*N12–*h*N15 contact loss reappear. This contact reappears notably because the polar network around the *h*N12–*h*N15 contact does not extend between a diameter of three, but a contact change involving the backbone of residue *h*N15 and the polar head of residue *h*E180. This suggests that overall, the 3GN decomposition is the best fit for investigating allosteric pathways and contact changes in proteins in general.

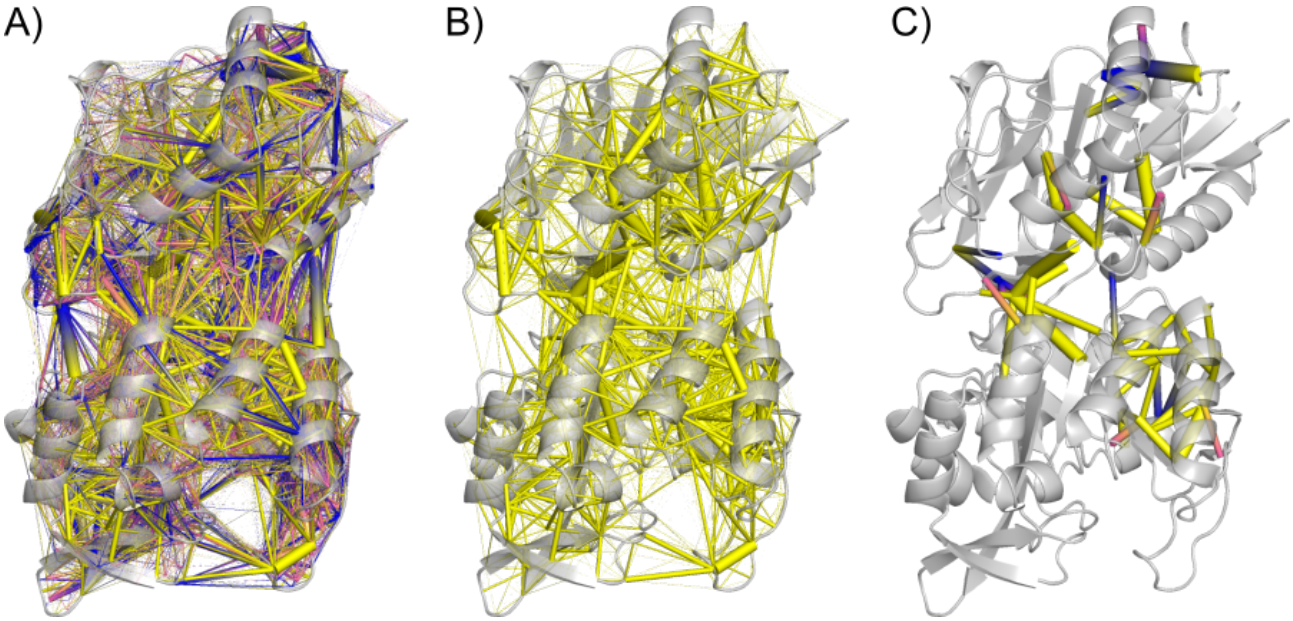


Figure 2.23: (A) PC1-3GN obtained including hydrogens in the *selection*. In practice, hydrogens are added to the selection of the heavy atom they are attached to. The endpoints of edges are colored in yellow for hydrophobic groups, blue for polar groups and pink for backbone groups. (B) hydrophobic component of this network. (C) CCA of this network

In 2GN, the adaptation of CCA produces a network with 11% backbone, 39% sidechain, and 49% backbone-sidechain links. The 3GN has 16% backbone, 6% hydrophobic, 13% polar, 30% backbone-polar, 20% backbone-hydrophobic, and 11% hydrophobic-polar. First, this suggests that the 2GN underweights the backbone contacts compared to the 3GN. This is particularly noticeable in Figure 2.22 where there are much more changes in backbone contact in loop1 and *h*α4 in the 3GN than in the 2GN. This better description of backbone contacts is due to a better balance between the sizes of the backbone, polar, and hydrophobic chemical groups compared to those of backbone and sidechain groups. Second, this suggests in the 3GN that some hydrophobic contacts

are underweighted. In fact, of the 6% hydrophobic contacts, 71% involve an aromatic ring, 51% an ILV residue, but only 4% involve two ILV residues. Nevertheless, a simple solution such as adding hydrogen atoms in each *selection* (i.e., the hydrogen atom is added to the *selection* of its heavy-atom) does not provide a satisfactory solution and such networks are completely biased for hydrophobic contacts (see Figure 2.23). Notably, using such a selection, the CCA erases some of the most important edges in the networks, which shows that the network topology is vastly different and actually less well-balanced between chemical groups. In fact, this comparatively proves that our definition of chemical groups that we suggest is a good-balanced, albeit not perfect, one.

The original motivation behind the CGNs is that interresidual contacts are principally guided by chemical groups. While a contact between two residues can change in nature without affecting much the PC1-AAN, but this should appear in the PC1-CGNs. We can detect this phenomenon when an edge between two residues is replicated, but the edges have different signs. Incidentally, this happens with low-value edges because they fluctuate around zero, but in the network cleaned with CCA, such changes have a profound meaning. In the PC1-2GN cleaned with CCA, seven residual contacts are duplicated, but only one shows a difference in sign: hY79–hR171. This shows that the majority of duplicated edges are synergistic (i.e. they influence positively each other), but contact nature change still occurred. Here, the contact change shows more sidechain-backbone (backbone of hR171, sidechain of hY79) component in prfar and more sidechain-sidechain contact in apo.

In PC1-3GN, 17 residual contacts are duplicated and 2 are tripled. Among duplicates, only 2 have an opposite sign: hY79–hR171 and hH120–hF139. The hY79–hR171 contact refine the PC1-2GN finding and shows that the contact has a bigger backbone-hydrophobic component in prfar (backbone moiety of hR171, hydrophobic part of hY79) and more hydrophobic-polar component in apo (polar moiety of hR171, hydrophobic part of hY79). This suggests that the hydrophobic moiety of hY79 exchanges contacts between the polar head of residue hR171 and the backbone head of residue hR171 upon PRFAR binding. Since hY79 being aromatic, this suggests that the contact exchanges from a π-cation interaction with the sidechain to a π-polar interaction with the backbone after PRFAR binding. Similarly, the other contact involves an exchange of contact from the polar part of residue hH120 between the hydrophobic and backbone moiety of residue hF139, suggesting that the contact is π-stacking in apo and changes to a polar interaction with the backbone (either strictly polar, π-polar or hydrogen bond) upon binding of PRFAR.

Interestingly, both triplicates have one sign different from the other two. This suggests that while synergistic effects are the norm in duplicate edges, they are an exception in triplicates. The two triplicates are hN15-hE180 and hD11-hV18. It has to be noted that hE180 is the acid from the catalytic triad and hN15 belongs to the allosteric pathways, thus by such we actually discovered a new interesting connection to the active site that may serve as an additional allosteric mechanism. In apo, the fD11–fV18 is typically both backbone-polar and hydrophobic-polar (between the backbone and hydrophobic parts of hV18 and the polar head of fD11), while in prfar it is typically a backbone-backbone contact. This suggests that there is a change in the hydrogen bonding between the two residues from a sidechain-backbone hydrogen bond in apo to a backbone-backbone hydrogen bond in prfar. In apo, the hN15–hE180 contact is typically between the polar head of residue hN15 and both the backbone and the hydrophobic part of residue hE180. In prfar, it is typically a backbone-polar contact between the backbone of hN15 and the polar head of hE180. Interestingly, this seems to suggest that upon effector binding, the backbone of residue hE180 is freed from its contact with residue hN15 but the polar head of residue hE180 is in contact with residue hN15.

In the catalytic triad, the role of residue hE180 is to stabilize the protonated base during catalysis (here, residue hH178) and in fact there is a stable hydrogen bond between the glutamine and histidine sidechains (see Figure 2.24E). Using our model, histidine has an NH group on the δ nitrogen and the hydrogen bond is between these hydrogen and glutamine sidechain oxygens (see Figure 2.24). Interestingly, the hH178-hE180 hydrogen bond is slightly more stable in apo. The effect is small, and that is probably why it does not appear in PC1N. To explain this, we must recall what has been documented in allosteric pathways[3, 4]. In apo there is a stable polar contact between the polar sidechains of residues hN12 and hN15 which breaks upon PRFAR binding (see Figure 2.24A-C). In apo, there is also a hydrophobic contact between the tiny hydrophobic moieties of hN15 and hE180. This contact is not really detected in the PC1-3GN because of the small size of the moieties, but it still recorded thanks to the backbone of residue hN15 which remains close to hE180. In this context, in a few snapshots, there is an unstable self-hydrogen bond between the sidechain and the backbone of residue hE180 (see Figure 2.24A) which breaks the hH178-hE180 hydrogen-bond. This also happens in prfar, but is less frequent (see Figure 2.24D). In prfar, the polar contact between hN12 and hN15 is broken and thus the polar head of hN15 is in an indirect, probably water-mediated, contact with the backbone of residue hE180. Although this effect discovered thanks to 3GN is only very tiny, there is reason to suspect that it may play a bigger role during catalysis. This may require additional studies and modeling of the precise catalytic mechanism in IGPS.

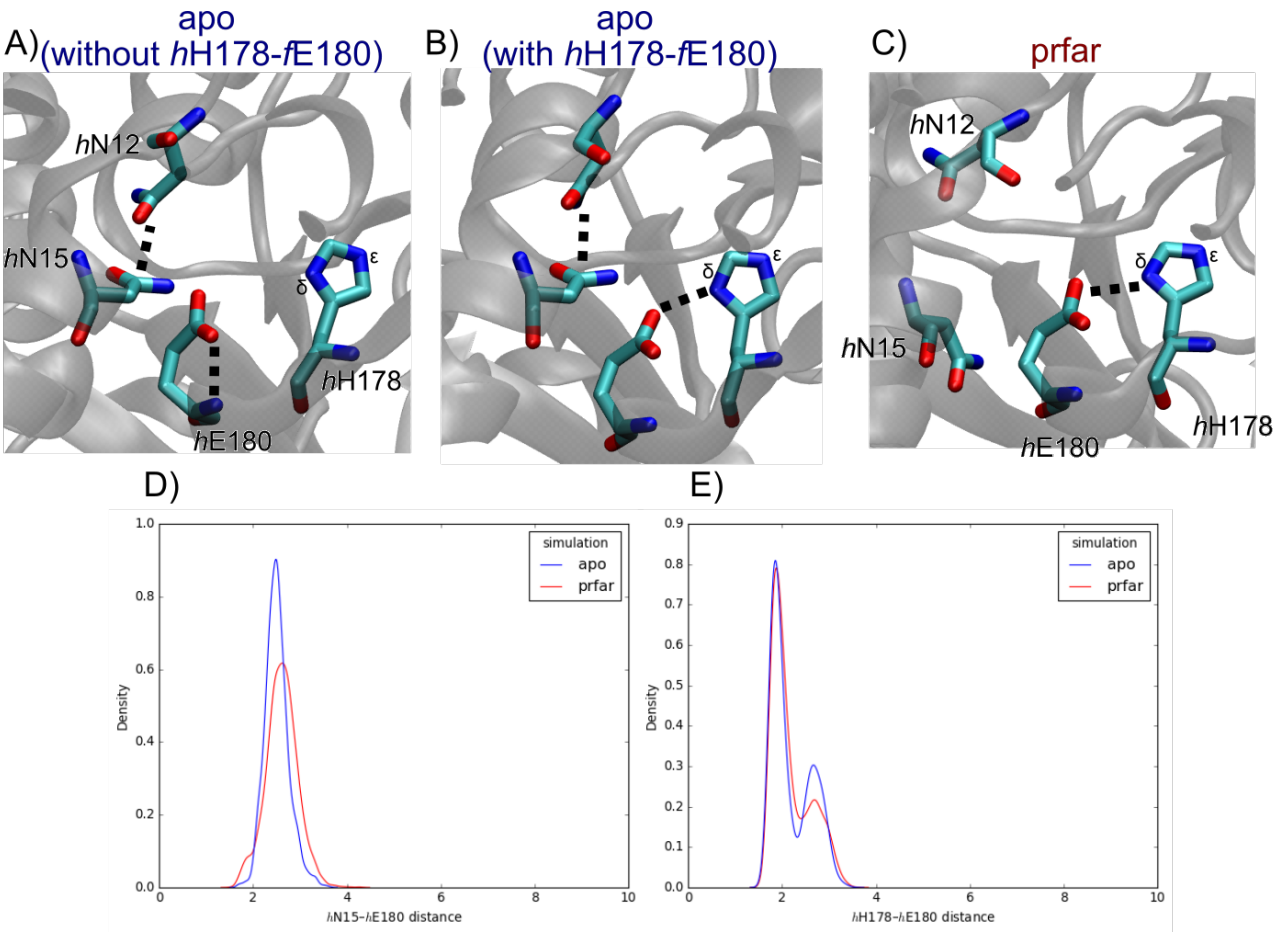


Figure 2.24: Snapshot of the hN15-hE180 contact with IGPS in cartoon representation and hN12, hN15, hH178 and hE180 shown in licorice in apo in a snapshot where the hH178–hE180 contact is broken in apo (A) or formed in apo (B) and formed in prfar (C). The polar contact between hN12 and h and the self hydrogen bond between residue hE180 is displayed in black dotted line in apo. The hydrogen bond between residue hH178 and hE180 in prfar is displayed in black dotted line. Kernel density estimate of the hN15-hE180 (C) and hH178-hE180 (D) distances in apo and prfar.

Conclusion on the use of Chemical Group Networks

The CGNs are a very powerful tool in combination with cPCA and despite they produce a more complex view of the network, it can easily be simplified thanks to an adapted version of the CCA. Of the two proposed variations of CGN: one with two groups (sidechain, backbone) and the one with three groups (backbone, hydrophobic and polar moieties), the version with three groups produces particularly better results, shows a better balance between groups, and shows all the important allosteric contacts found in AANs. Still, groups in the 3GN are not perfect, and some contacts seem to be slightly underweighted (notably hydrophobic contacts) and sometimes the network captures more easily an incidental contact between other groups instead of the true hydrophobic contact. Probably, different definitions of CGN may probably solve this problem. Interestingly, after the networks were cleaned with the adapted CCA, only a handful of edges are duplicated between the same pair of residues. This shows that CGN can provide approximate knowledge about the chemistry of a contact change. The only triplicate edges (in the 3GN) we found were even representative of a contact change between the residues. More generally, thanks to CGNs, we were able to capture phenomena in which a contact change of nature occurred between two residues. This feat is not possible with AAN, which proves the interest of this methodology, and we were notably able to capture a new impact of PRFAR binding on the active site. More generally, this methodology also showed that deeper studies of IGPS, with water molecules and modeling of the catalytic mechanism, are required to fully understand the allosteric mechanism in IGPS.

References

[1] Donald Voet, Judith G Voet, and Charlotte W Pratt. *Fundamentals of biochemistry: life at the molecular level*. John Wiley & Sons, 2016.

[2] Dengming Ming and Michael E Wall. "Allostery in a coarse-grained model of protein dynamics". In: *Phys. review letters* 95.19 (2005), p. 198103.

[3] Ivan Rivalta et al. "Allosteric pathways in imidazole glycerol phosphate synthase". In: *Proc. National Acad. Sci.* 109.22 (2012), E1428–E1436.

[4] Aria Gheeraert et al. "Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks". In: *The J. Phys. Chem. B* 123.16 (2019), pp. 3452–3461.

Chapter 3

Applications of the methodology

In this chapter, the DPCN and PC1N methods were applied in new investigations which were stimulated by other research groups that contacted us upon publication of our methodology.

3.1 Elucidating the Activation Mechanism of Adenosine MonoPhosphate-activated protein Kinase by Direct Pan-Activator PF-739

3.1.1 Adenosine MonoPhosphate-activated protein Kinase

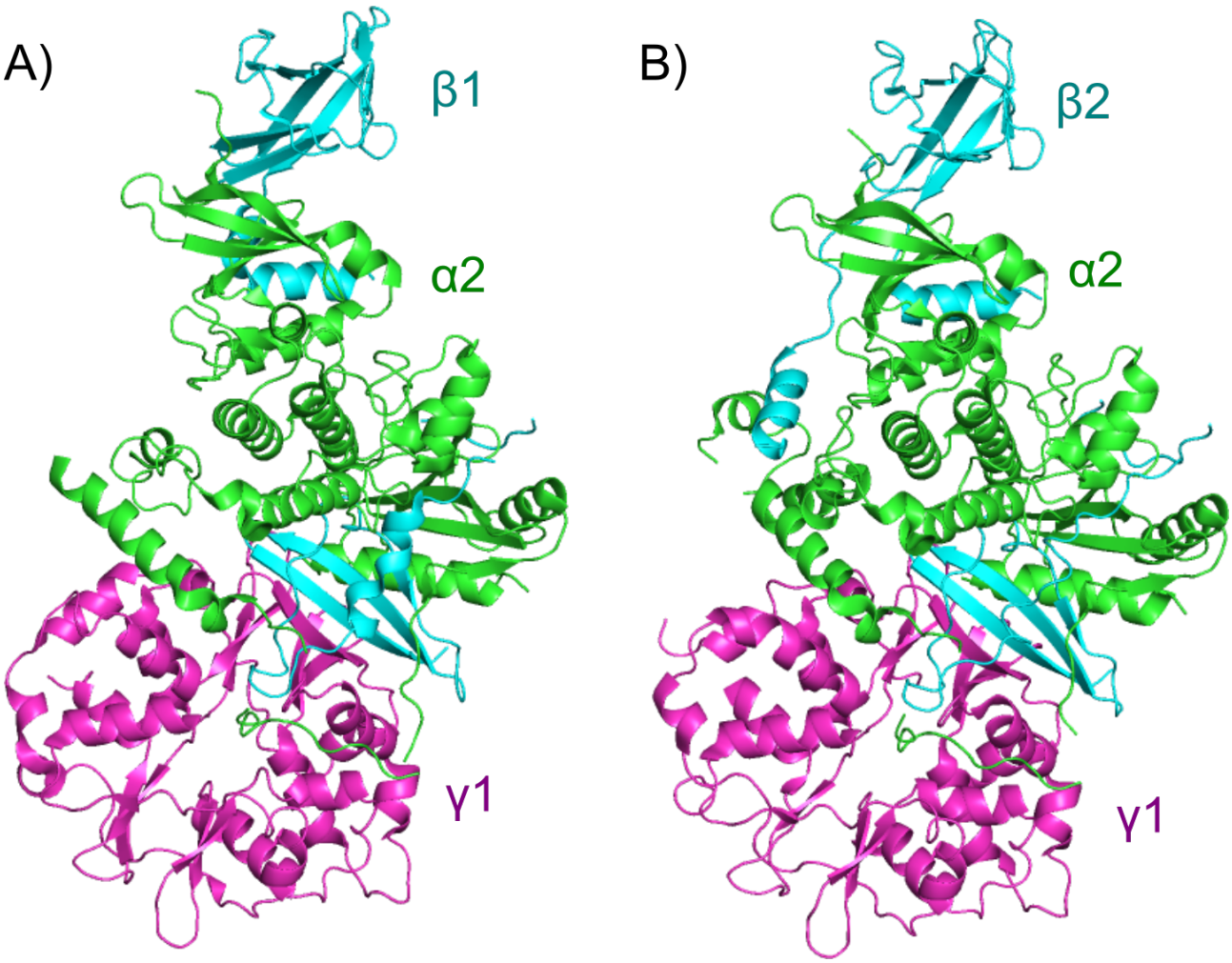


Figure 3.1: Crystal structure of two isoforms of human AMPK, $\alpha 2\beta 1\gamma 1$ (A, PDB entry 4CFF) and $\alpha 2\beta 2\gamma 1$ (B, PDB entry 6B2E)

Adenosine triphosphate (ATP) is a ubiquitous organic compound that provides energy to many molecular processes. When consumed it can convert either to a diphosphate (ADP) or a monophosphate AMP (AMP). Regulatory processes exists so that ATP is regenerated. The AMP-activated protein kinase (AMPK), is a highly

conserved protein that can sense low levels of ATP and responds by phosphorylating specific proteins to increase ATP generation and decrease ATP consumption. AMPK has been described as the "guardian of metabolism and mitochondrial homeostasis"[1]. Due to its role, AMPK is involved in various metabolic disorders such as type 2 diabetes, cardiovascular diseases, and obesity[2]. AMPK is a heterotrimeric complex composed of three different subunits designated α, β and γ. The catalytic site of the enzyme is located in the α subunit. In humans, each subunit is found in different isoforms, two for α, ($\alpha1$, $\alpha2$), two for β, ($\beta1$, $\beta2$) and three for γ ($\gamma1$, $\gamma2$, $\gamma3$) which makes a total of 12 variations of AMPK that have various distributions in different tissues[3]. Notably, while the $\alpha1$, $\beta1$ and $\gamma1$ have a low specificity, $\alpha2$ is found mainly in the heart and skeletal muscle, $\beta2$ in the skeletal muscle and $\gamma2$ in the heart muscle, and $\gamma3$ is found in the skeletal muscle[4, 5]. The different functionalities of these isoforms is an active research topic.

Several external factors can lead to AMPK activation[1], usually related to energy deprivation. Among them many small-molecule allosteric activators of AMPK have been discovered, and some of them shows a specificity towards a particular isoform[6, 7, 8, 9]. Among them PF-379 was discovered as a non-selective activator that binds in a pocket located at the interface between the α and β subunits, sometimes designated as the allosteric drug and metabolite (ADaM) site. This pan-activator is notably able to activate AMPK in the skeletal muscle which stimulates glucose uptake and glucose lowering which makes it a potential therapeutic approach to treat diabetic patients[8]. Understanding its mechanism of action could be valuable for designing selective activators to improve specificity in therapeutic treatments.

3.1.2 Molecular dynamics simulations and scalability of the code

The team of Carolina Esterellas and Elnaz Aledavood modeled AMPK neglecting the γ subunit following a "design and conquer" strategy. Indeed, the γ subunit does not contain the effector nor the active site, the experimental structures near α-γ and β-γ interface are not well resoluted which could add some uncertainty. Finally, increasing the system size would require drastically the computation time due to the need of a larger sampling. Thus they modelled the $\alpha2\beta1$ and $\alpha2\beta2$ proteins in three different configuration: the apoenzyme (free enzyme), the holoenzyme (enzyme bound to PF-379) and the ternary complex (enzyme bound to PF-379 and ATP) for a total of six systems. For each system was run three replicas of 1 µs for a total of 18 µs. Prior to this project, we ran the analysis focusing on two systems (apo and holo IGPS) and a maximum amount of 800 ns so this was a tremendous leap in time scalability. Still, the system is less large than IGPS containing only 2,955 heavy-atoms (against 3,578 in IGPS). This was an opportunity to test and adapt the scalability of the code. Notably, it is possible to highly parallelize the code because after MD the analysis of each individual frame are independent of one another. This is particularly helpful on a cluster where many processors are available and can drastically cut down the computation time. Thanks to the parallelization of the code, overall the practical computation time could be reduced by a factor equal to the number of processors used. At that point, computing the DPCN on a 2,955 heavy-atoms system for 10,000 frames (1 µs) was taking 1 hour on sixteen processors. The complete analysis of the system thus took 18 hours.

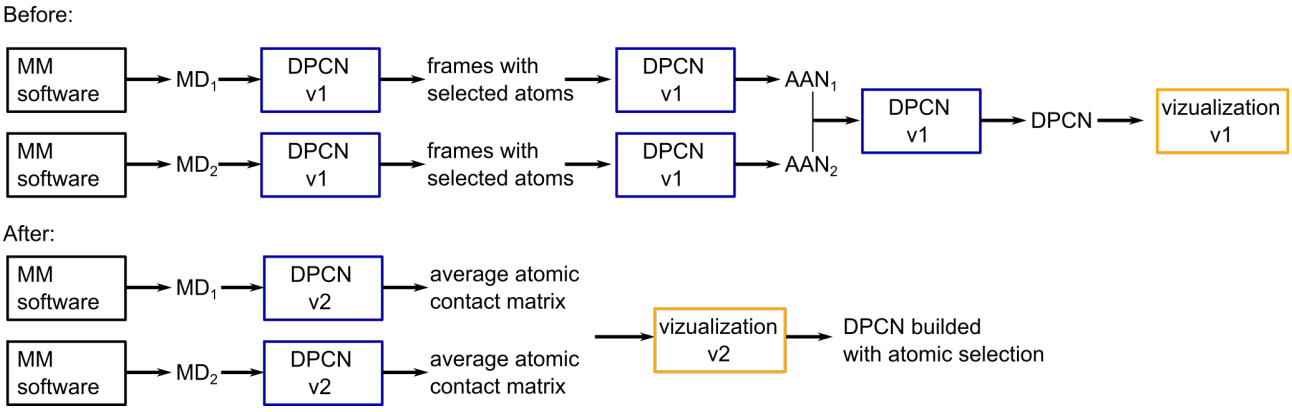


Figure 3.2: Evolution of the DPCN algorithm thanks to AMPK analysis

Six different systems makes for fifteen possible DPCN, considering all individual replicas, could total up to more than 153 DPCN to build. Looking also at both backbone contacts and all-atom contacts further increases the number of possibilities. We needed a different way to look at the procedure in order to only have to produce a single output file per system and one file per replica. Originally, one script was dedicated to building the DPCN and another to the visualization of those DPCN. Here, instead of computing the average amino acid network, we focused on computing the average atomic contact matrix. In the average atomic matrix, the frequency along the simulation of all atomic contacts are reported. The values taken in this matrix are thus binary, and this matrix is sparse. With an updated version of the visualization code, it could compute on the fly the DPCN between any two atomic matrices files by transforming this data into the amino acid contact matrix and then

doing the subtraction (which is almost instantaneous). Furthermore, we were able to tune the transformation applied to the atomic matrix to match a specific selection.

References

[1] Sébastien Herzig and Reuben J Shaw. "AMPK: guardian of metabolism and mitochondrial homeostasis". In: *Nat. reviews Mol. cell biology* 19.2 (2018), pp. 121–135.

[2] David Carling. "AMPK signalling in health and disease". In: *Current opinion cell biology* 45 (2017), pp. 31–37.

[3] Fiona A Ross, Carol MacKintosh, and D Grahame Hardie. "AMP-activated protein kinase: a cellular energy sensor that comes in 12 flavours". In: *The FEBS journal* 283.16 (2016), pp. 2987–3001.

[4] Mathias Uhlén et al. "Tissue-based map of the human proteome". In: *Science* 347.6220 (2015), p. 1260419.

[5] Human Protein Atlas. *Human protein atlas*. 2021.

[6] Barbara Cool et al. "Identification and characterization of a small molecule AMPK activator that treats key components of type 2 diabetes and the metabolic syndrome". In: *Cell metabolism* 3.6 (2006), pp. 403–416.

[7] Bing Xiao et al. "Structural basis of AMPK regulation by small molecule activators". In: *Nat. communications* 4.1 (2013), pp. 1–10.

[8] Emily C Cokorinos et al. "Activation of skeletal muscle AMPK promotes glucose disposal and glucose lowering in non-human primates and mice". In: *Cell metabolism* 25.5 (2017), pp. 1147–1159.

[9] Robert W Myers et al. "Systemic pan-AMPK activator MK-8722 improves glucose homeostasis but induces cardiac hypertrophy". In: *Science* 357.6350 (2017), pp. 507–511.

3.1.3 Published Article 2

This work led to the publication of an article in 2021 Frontiers in Molecular Biosciences in collaboration with the team of F Javier Luque and Carolina Esterellas.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 05 November 2021
doi: 10.3389/fmolb.2021.760026

Elucidating the Activation Mechanism of AMPK by Direct Pan-Activator PF-739

*Elnaz Aledavood[1], Aria Gheeraert[2,3], Alessia Forte[1], Laurent Vuillon[3], Ivan Rivalta[2,4], F. Javier Luque[1,5] and Carolina Estarellas[1]**

[1]Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Sciences, and Institute of Theoretical and Computational Chemistry (IQTCUB), University of Barcelona, Barcelona, Spain, [2]Dipartimento di Chimica Industriale "Toso Montanari" Università di Bologna, Bologna, Italy, [3]LAMA, University of Savoie Mont Blanc, CNRS, LAMA, Le Bourget du Lac, France, [4]Université de Lyon, École Normale Supérieure de Lyon, CNRS UMR 5182, Laboratoire de Chimie, Lyon, France, [5]Institute of Biomedicine (IBUB), University of Barcelona, Barcelona, Spain

Adenosine monophosphate-activated protein kinase (AMPK) is a key energy sensor regulating the cell metabolism in response to energy supply and demand. The evolutionary adaptation of AMPK to different tissues is accomplished through the expression of distinct isoforms that can form up to 12 heterotrimeric complexes, which exhibit notable differences in the sensitivity to direct activators. To comprehend the molecular factors of the activation mechanism of AMPK, we have assessed the changes in the structural and dynamical properties of β1- and β2-containing AMPK complexes formed upon binding to the pan-activator PF-739. The analysis revealed the molecular basis of the PF-739-mediated activation of AMPK and enabled us to identify distinctive features that may justify the slightly higher affinity towards the β1–isoform, such as the β1–Asn111 to β2–Asp111 substitution, which seems to be critical for modulating the dynamical sensitivity of β1- and β2 isoforms. The results are valuable in the design of selective activators to improve the tissue specificity of therapeutic treatment.

Keywords: AMPK, protein dynamic, protein activation mechanism, pan-activator, isoform selectivity, molecular dynamics simulation

OPEN ACCESS

Edited by:
Yong Wang,
Zhejiang University, China

Reviewed by:
Jon Oakhill,
University of Melbourne, Australia
Zhaoxi Sun,
East China Normal University, China

****Correspondence:***
Carolina Estarellas
cestarellas@ub.edu

Specialty section:
This article was submitted to Biophysics, a section of the journal Frontiers in Molecular Biosciences

Received: *17 August 2021*
Accepted: *08 October 2021*
Published: *05 November 2021*

Citation:
Aledavood E, Gheeraert A, Forte A, Vuillon L, Rivalta I, Luque FJ and Estarellas C (2021) Elucidating the Activation Mechanism of AMPK by Direct Pan-Activator PF-739. Front. Mol. Biosci. 8:760026. doi: 10.3389/fmolb.2021.760026

INTRODUCTION

AMP-activated protein kinase (AMPK) is a Ser/Thr protein kinase with a key role as a sensor in cellular energy homeostasis (Xiao et al., 2011). Upon activation, AMPK increases the levels of ATP, favoring the reduction of anabolic pathways and up-regulation of catabolic pathways. Due to its critical role in cell metabolism, AMPK is implicated in numerous metabolic disorders such as type 2 diabetes, cardiovascular diseases, and obesity (Carling, 2017). However, one of the most interesting aspects of this enzyme comes from the different tissue distribution that is directly related to its structural complexity. AMPK is a heterotrimeric complex consisting of a catalytic α-subunit and two regulatory subunits, namely β and γ. Each subunit can be found in different isoforms, involving two for α (α1, α2), two for β (β1, β2), and three for γ (γ1, γ2, γ3) (Calabrese et al., 2014). The N-terminus of the α catalytic subunit contains a kinase domain, while its C-terminus is needed for the formation of the complex with the other subunits. The β-subunit has a central carbohydrate-binding module (CBM) that mediates AMPK interaction with glycogen, and the C-terminal region acts as a scaffold for the heterotrimeric assembly. Finally, the γ-subunit has four tandem repeats of the cystathionine

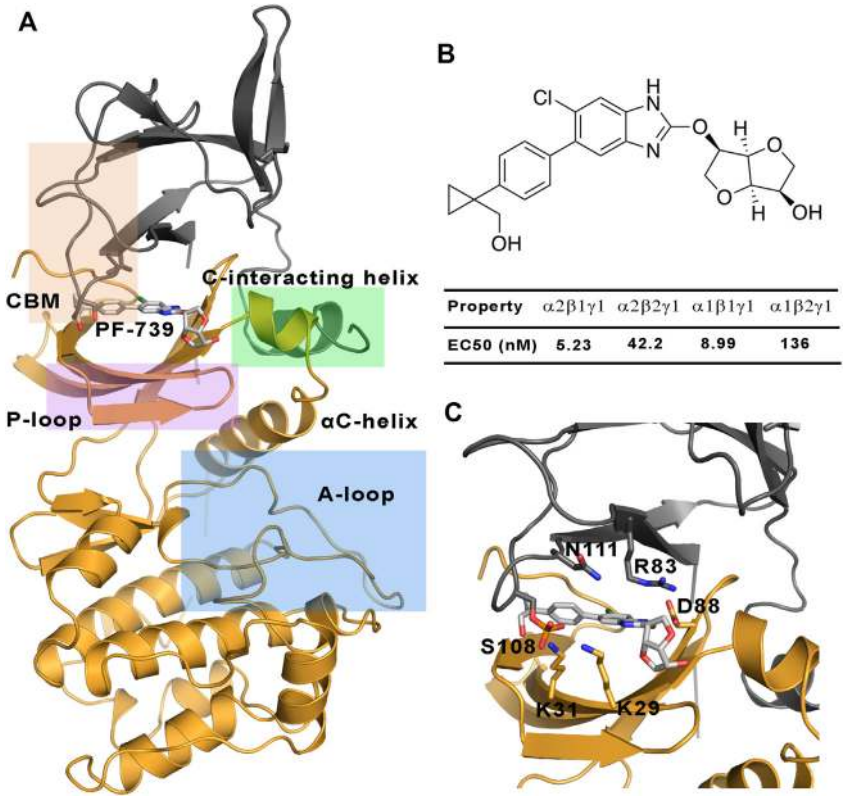


FIGURE 1 | (A) Representation of the system selected for the study of the direct activation mechanism, which is composed of α2 **(yellow cartoon)** and β1-/β2-subunits **(grey cartoons)**. The most important regions of these subunits are highlighted: P-loop **(purple)**, activation loop (A-loop, **cyan**), CBM domain **(orange)** and C-interacting helix **(green)**. **(B)** Chemical structure of PF-739, together with the experimental data obtained in Cokorinos et al. (2017) for the half maximal effective concentration (EC_{50}). **(C)** The activator bound to the ADaM site, with selected residues of the ADaM site shown as sticks.

β-synthase (CBS) domain, forming up to four potential nucleotide binding sites although only sites 1, 3 and 4 can really bind them (Scott et al., 2004; Scott et al., 2008; Carling et al., 2012).

AMPK is finely regulated by different mechanisms (Mahlapuu et al., 2004). An allosteric activation involves the phosphorylation of α2-Thr172 in the activation loop of the kinase domain by upstream kinases such as LKB1 and CaMKKb, together with the binding of AMP to the CBS domain in the γ–subunit. The active AMPK complex can thus respond to subtle fluctuations in the AMP/ATP ratio, it being several thousand-fold more active (Carling et al., 2012; Chen et al., 2012; Willows et al., 2017). On the other side, AMPK can also be indirectly activated by compounds such as metformin, phenformin and oligomycin (Vazquez-Martin et al., 2012), which are able to increase the intracellular levels of AMP. However, much interest is focused on the understanding of the direct activation mechanism of AMPK by small organic molecules. The first reported direct activator was the thienopyridone drug A-769662 (Cool et al., 2006), which is bound to a cavity located at the interface between the CBM domain of the β-subunit and the kinase domain of the α-subunit, namely the allosteric drug and metabolite (ADaM) site (Langendorf and Kemp, 2015). One of the main features of the direct activation is that this kind of activation is independent of the Thr172 phosphorylation, while it is enhanced by phosphorylation of Ser108 in the CBM domain of the β-subunit, increasing the AMPK activity by >90-fold (Hardie, 2014). Since then, a lot of efforts have been invested in obtaining direct AMPK activators, which in some cases exhibit a marked isoform selectivity (Olivier et al., 2018), while in other cases no significant selectivity is observed towards specific subunit isoforms. The isoform selectivity is relevant for the tissue distribution of the AMPK complexes. While α1, β1 and γ1 have low tissue specificity, α2 is basically found in the heart and skeletal muscle, β2 in the skeletal muscle and γ2 is mainly found in the heart muscle, and γ3 is found in the skeletal muscle (Uhlén et al., 2015; Human Protein Atlas (2021, 2021). The tissue specificity is related to the specific function of AMPK in these tissues, and therefore all the isoforms in the skeletal muscle have an important role in the glucose uptake, making AMPK a promising target for diabetes type 2 disease. In the last years an increasing effort has been devoted to design tissue-specific direct AMPK activators. As an example, the SC4 small-molecule, which was designed to increase the selectivity towards the α-subunit (being more selective for the α2-isoform) (Ngoei et al., 2018), can activate both β1- and β2-containing AMPK complexes, although a slightly higher activation is observed for the β1-isoform. Other interesting examples are the pan-activators

PF-739, which is able to activate both α2β1γ1 and α2β2γ1 (**Figure 1**), and MK-8722 which can activate the 12 heterotrimeric AMPK complexes (Myers et al., 2017). Regarding the selectivity of β-isoform, although the half maximal effective concentration (EC_{50}) determined for PF-739 and the binding affinity measurements for MK-8722 shows that they still exhibit a larger affinity for the β1-containing isoforms, they are the most potent activators of β2 complexes reported up to date (Cokorinos et al., 2017). However, it is still necessary to achieve a higher specificity to avoid off-tissue target effects. Accordingly, understanding of the molecular factors that favor the binding to specific isoforms is an outstanding issue.

In our previous works (Aledavood et al., 2019; Aledavood et al., 2021), we have studied the molecular factors that determine the selective activation of β1- and β2-containing AMPK complexes formed with A-769662 and SC4. We have hypothesized that the change of β1-Asn111 by β2-Asp111 could be a key factor in mediating the distinctive "mechanical" sensitivity of AMPK complexes to these activators. Here, we extent this analysis to the pan-activator PF-739 with the aim to examine how the binding of this compound affects the dynamical response of AMPK considering the trends disclosed for A-769662 and SC4. At this point, it is worth noting that while A-769662 is selective for β1-containing complexes, SC4 exhibits a mild preference for this isoform, a trend which was attributed to the presence of the carboxylate group present in the chemical structure of this activator. In contrast, PF-739 is a neutral compound, which suggests that other chemical features might also regulate the mild preference for binding to β1-containing AMPK complexes. Understanding the role of the factors that regulate the mechanical response of AMPK could thus be valuable for the tailored design of isoform-adapted pharmacophores useful in the search of selective direct activators. With this aim in mind, we have carried out extensive molecular dynamic simulations (MD) and network analysis to examine the differential trends in structural, dynamical and interaction patters emerging for AMPK complexes with PF-739.

RESULTS AND DISCUSSION

MD simulations were run to assess the structural and dynamics properties of the AMPK complexes formed by the α2-isoform bound to either β1- or β2- subunits. The neglect of the γ subunit in the simulated systems obeys two main motivations. First, following a divide-and-conquer strategy, this permits to focus the conformational sampling of the activator-induced changes on the ADaM site, which is shaped by residues in α and β subunits. Second, the adoption of these systems permits a direct comparison with the results obtained previously for the complexes formed with A-769662 and SC4 (Human Protein Atlas (2021, 2021; Ngoei et al., 2018). Accordingly, this study is focused on the conformational ensemble collected for the apo species of α2β1 and α2β2 systems, the corresponding complexes formed with PF-739 (holo species), and finally the complexes formed with both PF-739 and ATP molecule (holo+ATP), the latter being located in the ATP-binding site within the kinase domain of the α−subunit. For each system (apo, holo, and holo+ATP), the analysis involves the conformational ensemble explored in three independent replicas (1 μs/replica), leading to a total simulation time of 6 μs for the apo species and 12 μs for the ligand-bound complexes.

Structural Analysis of AMPK Complexes

We have examined the effect of PF-739 binding to the ADaM site (holo structures), and the simultaneous presence of PF-739 and ATP in both ADaM and ATP-binding sites (holo+ATP structures) on the global structural conformation of apo α2β1 and α2β2 by means of the root mean square deviation (RMSD) of the protein backbone along the corresponding 1 μs simulations (**Figure 2**). The RMSD was determined using the average structure of the holo+ATP species sampled in the last 200 ns of the three independent replicas run for either α2β1 or α2β2 species as reference. For the holo+ATP systems there is a high structural resemblance for all the replicas, as noted in the small fluctuations of the RMSD profiles (**Figure 2C**), which agrees with the preservation of the overall protein fold upon binding of both PF-739 and ATP. In particular, the RMSD values for the holo+ATP species range from 2.0 to 2.5 Å for α2β1 and from 2.7 to 3.0 Å for α2β2 (**Table 1**). These values are lower than the RMSD values obtained for the apo species (α2β1: 2.5–2.7 Å; α2β2: 2.9–3.4 Å).

Binding of PF-739 to the α2β1 species has no significant effect on the RMSD of the holo species (from 2.5 to 2.9 Å), which is close to the values obtained for the apo form. Only the presence of both the ligand and ATP (holo+ATP) gives rise to a reduction in the RMSD. This effect is even more remarkable in the α2β2 species, as the RMSD of the protein backbone is generally larger than the RMSD value determined for the α2β1 complex in all the states (apo, holo and holo+ATP; see **Figure 2** and **Table 2**). These findings suggest that PF-739 exerts a weak structural stabilization upon binding to both α2β1 and α2β2 species.

Regarding the per-residue mean square fluctuation (RMSF) profile, similar results are observed for both α2β1 and α2β2 species, as noted in the resemblance of the fluctuation patterns obtained by averaging the RMSF of the three replicas run for every system (**Figure 3**). The highest fluctuations in the α-subunit correspond to residues in the activation loop (residues 165–185, highlighted in blue in **Figure 3**) and the α-helix formed by residues 210–230. It is worth noting the higher fluctuation of the P-loop (residues 15–35; purple in **Figure 3**) in the holo state in comparison to both apo and holo+ATP systems. Thus, binding of PF-739 significantly affects the flexibility of the P-loop in both α2β1 and α2β2 species, which may have functional relevance since the P-loop contributes to shape both the ADaM and ATP-binding sites. Regarding the β-subunit, the largest fluctuations are in the CBM domain, which contains Ser108 (highlighted in orange in **Figure 3**; phosphorylated in both holo and holo+ATP states), and the regions near the C-interacting helix (residues 162–172, highlighted in green in **Figure 3**). It is worth noting that the binding of PF-739 (holo) and ATP (holo+ATP) increases the fluctuations of the α-subunit elements mentioned above, while reduces the fluctuations in the β-subunit, independently of the β-isoform. These findings are in

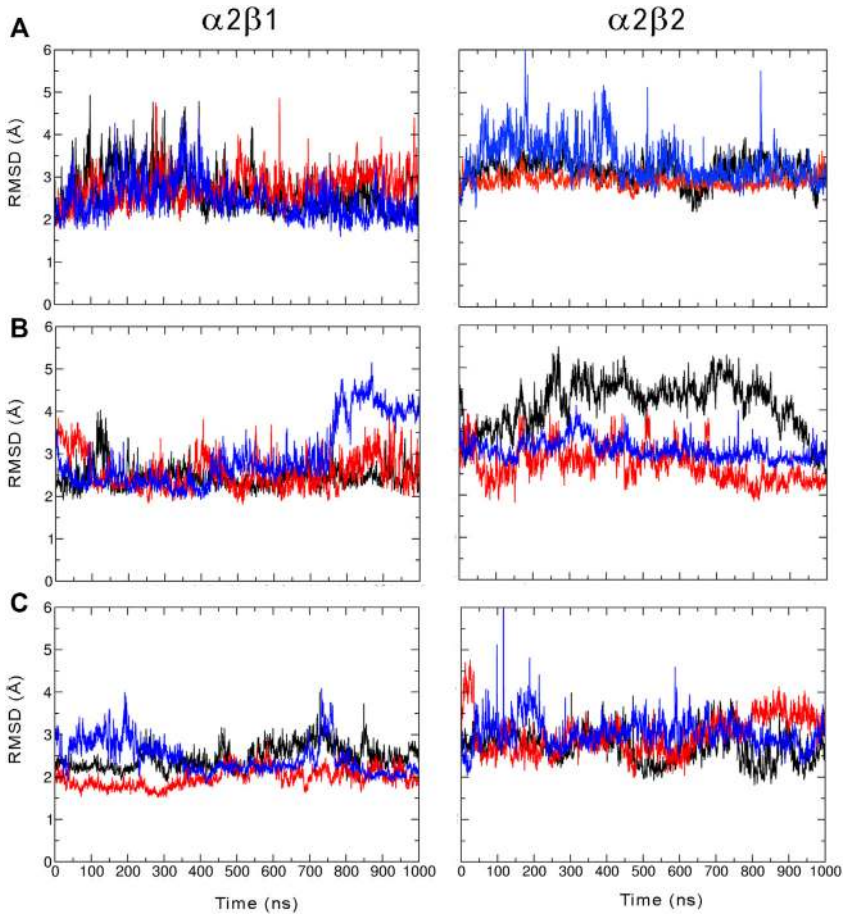


FIGURE 2 | Root mean squared deviation (RMSD, Å) determined for the protein backbone along the three 1 μs MD simulations run for the **(A)** apo, **(B)** holo and **(C)** holo+ATP species of AMPK isoforms α2β1 and α2β2 bound to PF-739 (each replica is shown in black, blue and red, respectively). For each analysis the reference structure used corresponds to the energy-minimized average structure of the holo+ATP sampled in the last 200 ns of the three independent MD simulations.

TABLE 1 | RMSD and standard deviation (Å) determined for the protein backbone of the snapshots sampled along the last 500 ns of MD simulations performed for all systems (apo, holo and holo+ATP states) of AMPK isoforms α2β1 and α2β2. Values were determined using the energy-minimized holo+ATP species averaged for the last 200 ns of each simulation system as reference structure.

System		Replica 1	Replica 2	Replica 3	Average
α2β1	**Apo**	2.6 ± 0.6	2.7 ± 0.4	2.5 ± 0.5	2.6
	Holo	2.5 ± 0.3	2.6 ± 0.5	2.9 ± 0.8	2.6
	holo+ATP	2.5 ± 0.3	2.0 ± 0.2	2.4 ± 0.4	2.3
α2β2	**Apo**	3.2 ± 0.3	2.9 ± 0.2	3.4 ± 0.5	3.2
	Holo	4.1 ± 0.6	2.7 ± 0.4	3.1 ± 0.3	3.3
	holo+ATP	2.7 ± 0.4	3.0 ± 0.5	3.0 ± 0.4	2.9

TABLE 2 | Contribution of the essential motion (%) to the structural variance of different AMPK systems and the total contribution of the first four projections.

Systems		Proj. 1	Proj. 2	Proj. 3	Proj.4	$Total_{(P1-P4)}$
α2β1	**apo**	41.2	12.0	8.1	4.6	66.0
	holo	38.6	12.1	7.6	4.0	62.3
	holo+ATP	30.7	12.6	7.0	5.1	55.4
α2β2	**apo**	30.9	12.6	8.8	4.8	57.1
	holo	33.1	12.7	8.6	5.3	59.7
	holo+ATP	29.0	13.2	7.2	5.3	54.7

agreement with the higher RMSD fluctuations observed in some replicas of the holo states for both α2β1 and α2β2 species.

Dynamic Properties of AMPK Complexes

In order to examine the effect of the activator on the conformational behavior of AMPK complexes, we have analyzed both the essential dynamics (ED) of the protein backbone and the dynamic correlation between residues.

The ED provides information about the essential motions of the protein and can be used to examine the effect of activator on the major motions of the protein skeleton. The results for the first essential motion for the apo (α2β1 and α2β2) states show a concerted bending that brings α- and β-subunits closer and then moves them apart (**Figure 4**). The most interesting feature is that the P-loop seems to act as a hinge, assisting the concerted bending

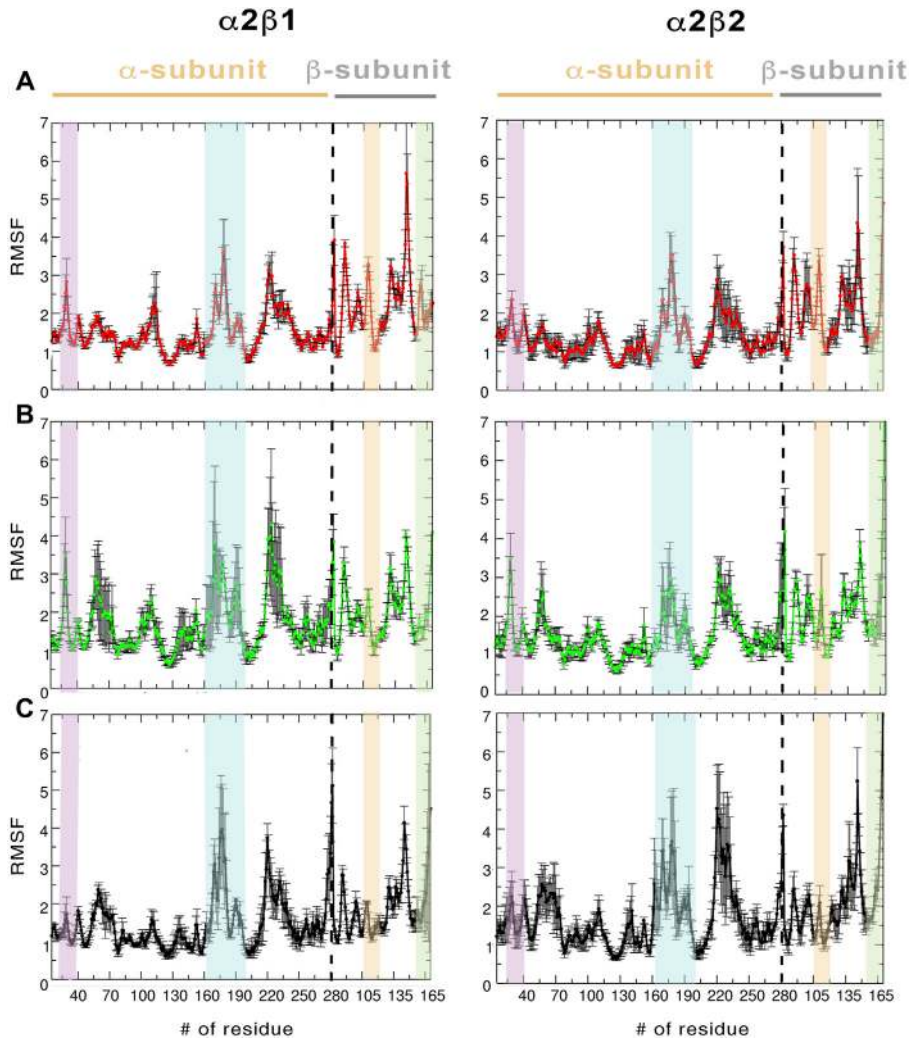


FIGURE 3 | RMSF (Å) average of the residues determined for the protein backbone along the last 500 ns of the three independent replicas runs for **(A)** apo, **(B)** holo and **(C)** holo+ATP. The standard deviation for each residue is shown as an error bar. The highlighted regions denote the moieties corresponding to P-loop **(purple)**, activation loop **(cyan)**, CBM domain **(orange)**, and C-interacting helix **(green)**.

between the subunits. Indeed, the first motion accounts on average for 41/31% of the structural variance in α2β1/α2β2 species, and the contribution of the first four motions accounts for 66/57% of the total structural variance (**Table 2**). This emphasizes the importance of the first essential motion to the conformational flexibility of the AMPK complexes.

Comparison of the ED results obtained for apo, holo and holo+ATP states reveals that binding of the activator has a mild effect on the conformational variance, which is reduced from 66% (apo) to 62% (holo) and 55% (holo+ATP) for the α2β1 species (**Table 2**). However, for the α2β2 species the activator triggers a slight increase in the conformational variance relative to the apo species, while subsequent binding of ATP results in a reduction of the structural variance (apo: 57.1%; holo: 59.7%; holo+ATP: 54.7%). These results are also reflected in the contribution of the first essential motion (**Figure 4** and **Table 2**). In the holo (α2β1 and α2β2) states, this motion reflects a synchronous motion of the P-loop and the CBM domain, which is in contrast with the increased stiffness observed in the holo+ATP state, especially regarding the P-loop, the helical domain in the α-subunit, as well as the region of the CBM domain nearest to the ADaM site. However, although the movements of the CBM domain are very similar between α2β1 and α2β2 species, the P-loop and the helices at the C-terminal region of the α-subunit exhibit higher fluctuations in α2β2 with respect to α2β1 (**Supplementary Figure S1**). Finally, it is worth noting that the enhanced stiffness achieved upon ATP binding to holo is again more remarkable in the case of the α2β1 complex (**Figure 4** and **Table 2**).

Besides the qualitative inspection of the overall dynamics of the systems shown in **Figure 4**, we have determined the similarity indices for the first essential motions (**Supplementary Table S1**). The similarity index for the apo species (i.e., the most flexible one) is close to 0.70 and 0.60 for α2β1 and α2β2, respectively, reflecting

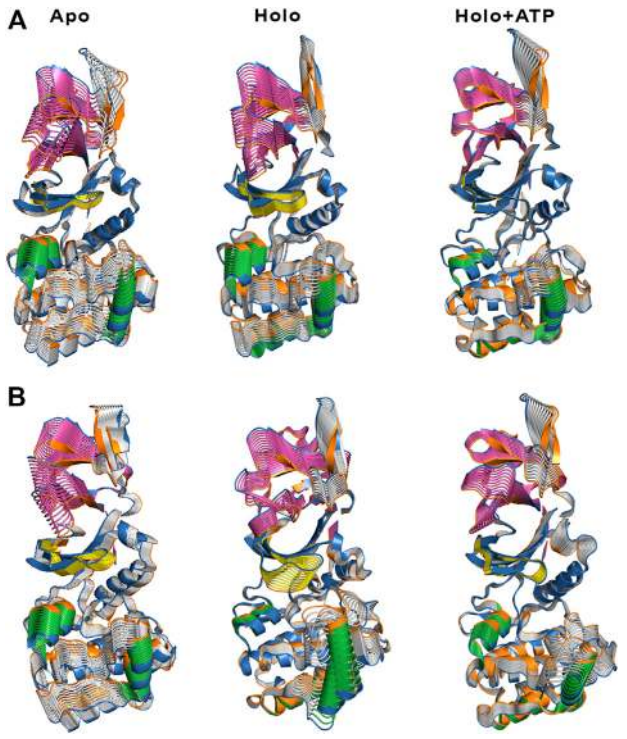


FIGURE 4 | Representation of the first essential motion derived from the ED analysis of the protein backbone for the **(A)** α2β1 and **(B)** α2β2 species, determined from the snapshots sampled along the last 500 ns of MD simulations. The P-loop is shown in yellow, and the helices formed by residues 100–110 and 220–229 in the α-subunit are shown in **green**. The CBM domain is colored in **magenta**.

the preservation of the major deformation of the protein skeleton in the three replicas. These results also agree with the higher conformational flexibility observed for α2β2 systems. In the holo species, the similarity indices are 0.75 for α2β1 and 0.47 for α2β2 systems. These results agree with the essential motion observed for the holo state of α2β1 (**Figure 4A**, middle panel) and α2β2 (**Figure 4B**, middle panel). In the former, the variance of the system is more balanced between certain regions, i.e., CBM domain, P-loop, A-loop and helices P220-G229 and E100-R110 (colored in green, **Figure 4**). However, higher fluctuations account for the structural elements in the α-subunit in α2β2. These findings are in agreement with the previous RMSD and RMSF results. Finally, for the holo+ATP systems the similarity index is close to 0.35 for α2β1 and α2β2, respectively. However, this simply means that binding of both activator and ATP rigidifies the protein skeleton, annihilating the large-scale deformations observed in the apo species as observed in **Figure 4**. The ED, shown in **Figure 4**, as well as the similarity indexes calculated, in **Supplementary Table S1**, have been obtained considering the last 500 ns of the simulation time of the three replicas. However, in order to check the statistical value of our simulations, we have also calculated the similarity indexes for the first three essential motions of the apo α2β1 and α2β2 derived from the ED analysis in time windows 200–600 and 600–1,000 ns for the three replicas (**Supplementary Table S2**). The similarity index amounts in general to 0.8. For the first replica of α2β1 system a lower similarity is observed, suggesting a slower structural relaxation, as noted in the similarity obtained for more advanced time windows (**Supplementary Table S3**). Overall, these results suggest that selection of the last 500 ns to perform the statistical analysis of the simulations is well suited for the comparison between replicas, although these results also suggest that shorter time periods might be also usable. For this reason this 500 ns time window has been used in further analysis.

To complement the results of ED analysis, we have performed two additional analyses with the aim to assess the dynamic correlation between residues and disclose specific relationships between the α- and β-subunits: a dynamical perturbation network (DPN, **Figure 5**) and a dynamic cross-correlation (DCC, **Figure 6**; see *Methods and Materials* for technical details) analysis.

The dynamical perturbation network (DPN) was calculated for apo and holo species as an average of the three independent replicas. **Figure 5** shows the changes in the correlation of residues between apo and holo states, where blue/red edges stand for contacts weakened/strengthened in holo relative to apo state. Thus, these networks provide information of how the interaction of the activator with the enzyme affects the contact network between residues. For the sake of comparison, this analysis was performed not only for PF-739, but also for A-769662, which exhibits a marked selectivity for β1-containing AMPK complexes. Our previous studies (Human Protein Atlas (2021, 2021; Ngoei et al., 2018) revealed that A-769662 acts as molecular glue between the α2- and β1-subunits, while this effect is lost in the α2β2 species due to the higher dynamical resilience of this specie towards the activator. The dynamical contact network for A-769662 (**Figure 5A**) perfectly agrees with these findings. In fact, the changes between apo and holo in α2β1 mainly reveal a higher number of contacts between the P-loop of the α2-subunit and the CBM of the β1-subunit as well as between the αC-helix of the α2-subunit and the C-interacting helix of the β1-subunit. Conversely, the contact network that emerges for the α2β2 complex is more complex, involving regions located far from the ADaM site. This result agrees with the higher flexibility of the α2β2 species, and the lower impact of A-769662 on the dynamical response of this complex.

For the pan-activator PF-739, the α2β1 complex exhibits fewer and more specific contacts, which primarily affect the CBM/P-loop and the αC-helix/C-interacting helix/A-loop, than the α2β2 species, thus resembling the results discussed for A-769662. However, the number of contacts weakened or even lost between the A-loop and the αC-helix in the α2β1 holo state is remarkably higher for PF-739-bound complexes compared to A-769662-bound ones (**Figures 5A,B**, left side). For α2β2, the number and weights of the edges are larger in this species, and the distribution of contacts involves wider regions from the CBM domain to the A-loop (see highlighted region in cyan, **Figure 5B**, right panel). Noteworthy, DPN analysis reveals that binding of A-769662 gives rise to a much larger difference in the dynamical network of α2β1/A-769662 and α2β2/A-769662 complexes than for α2β1/PF-739 and α2β2/PF-739 complexes, as the pattern

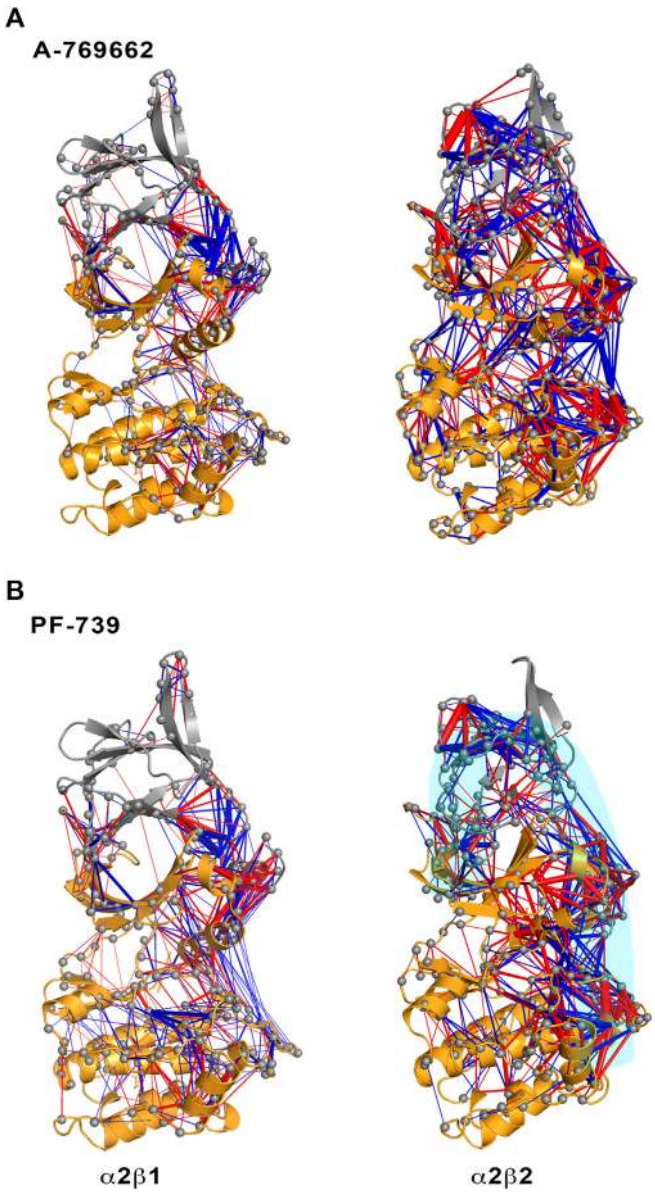


FIGURE 5 | Representation of the contact changes observed in the dynamical perturbation network between apo and holo states for **(left)** α2β1 and **(right)** α2β2 species. The holo states were calculated in presence of **(A)** A-769662 and **(B)** PF-739 activators. Red/blue edges show interresidue atomic contact increase/decrease in the holo state relative to the apo form. The magnitude of these changes is indicated by the width of the edges.

observed for the last two AMPK complexes exhibit a similar pattern (**Figure 5**). This is in agreement with the experimental results that indicate that the PF-739 is active against both β1- and β2-containing isoforms, in contrast with the selective activation of β1-containing AMPK complexes reported for A-769662.

Finally, the dynamic cross-correlation (DCC) analysis was performed to examine the correlated motions of residues in α2β1 and α2β2 AMPK complexes. For the apo systems (**Figure 6A**) one may notice a significant correlation between residues in the P-loop and the αC-helix, both in the α-subunit, and between the αC-helix from the α-subunit and the C-interacting helix from the β-subunit (as noted by the yellow marks). It is worth noting that there is a slight correlation between the P-loop and the CBM domain (β-subunit), more remarkable in α2β1 than in α2β2, as noted by the similarity indexes of 0.82 for α2β1, which is reduced to 0.75 in α2β2 (**Supplementary Table S4**). The holo+ATP systems show lower dynamical correlation between residues, as observed by the progressive reduction in the number and intensity of the areas that exhibit a pronounced correlation (shown in yellow and blue for highly correlated and anticorrelated fluctuations between residues, respectively). On the contrary, the correlation between the motion of the P-loop and the CBM domain is reinforced in the holo and holo+ATP states (black square in **Figure 6**). These effects are more noticeable for the comparison of holo in α2β1 (similarity indices of 0.63 in α2β1 vs 0.55 in

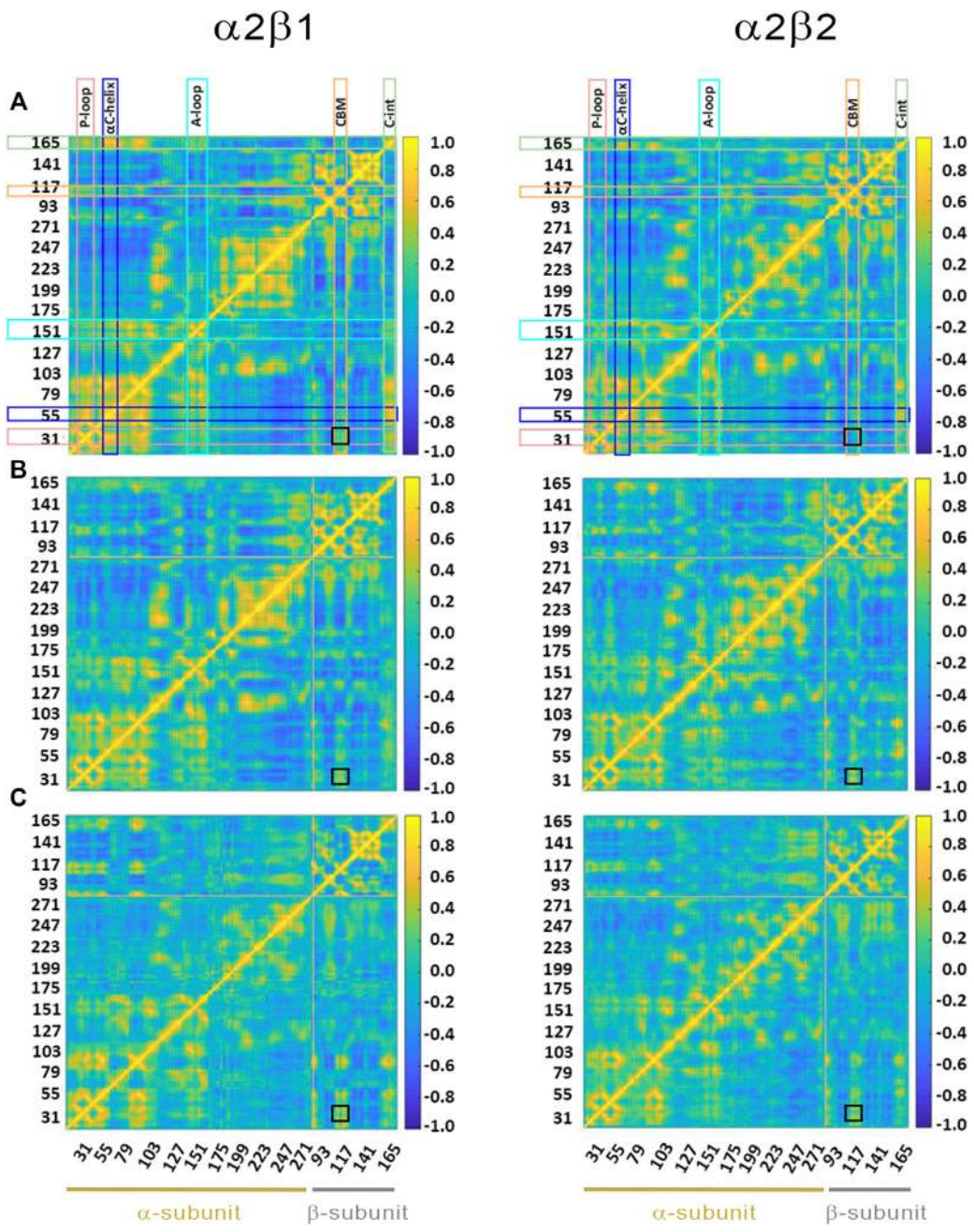


FIGURE 6 | Dynamic cross-correlation (DCC) matrices for **(A)** apo, **(B)** holo and **(C)** holo+ATP complexes of **(left)** α2β1 and **(right)** α2β2 with PF-739. The *x*- and *y*-axis denote the numbering of residue in the α- and β-subunit, indicated at the bottom of the plot (yellow and gray lines for α- and β- subunits, respectively). Specific regions of AMPK are highlighted with squares at the top of the plot: P-loop **(pink)**, αC-helix **(blue)**, activation loop (A-loop; **cyan**), CBM unit **(orange)** and C-interacting helix (C-int; **green**). Regions colored in yellow/blue show high correlated/anticorrelated fluctuations. The black square in the plots highlights the motion between the CBM domain and the P-loop.

α2β2, **Supplementary Table S4**), while lower differences exist for holo+ATP systems in α2β1 and α2β2, in agreement with previous analyses.

Although the preceding results show a high similarity in the dynamical behavior of both α2β1 and α2β2 species bound to PF-739 activator, which agrees with the definition of PF-739 as a pan-activator, these analyses still reveal subtle differences between β1- and β2-containing AMPK complexes. In particular, the results suggest that the α2β2 species have a larger resilience to the structural modulation exerted by the activator, whereas the α2β1 isoform is more sensitive to the conformational adaptation induced upon activator binding to the ADaM site, enhancing the stiffness of protein backbone for the β1-containing complex (**Figures 4**, **5**). These results agree with the fact that PF-

739, which can activate both α2β1γ1 and α2β2γ1 complexes, still exhibits a larger affinity for the β1-isoform (**Figure 1**) (Cokorinos et al., 2017).

Pre-Organization of ATP-Binding Site

To explore how PF-739 could influence the activation of AMPK, we have evaluated the dynamical response of the ATP-binding site due to the binding of the activator in the ADaM site. Specifically, we have assessed the pre-organization of the ATP-binding site in the apo, holo and holo+ATP states, using as a reference the average structure of the holo+ATP complex.

For the holo+ATP states, the residues of the ATP-binding site sample a conformational space with a high peak centered at a positional RMSD of 1.2 A° and a shoulder at 1.9 A° for α2β1, while a wider distribution is observed with a peak centered at 1.8 A° for α2β2 (**Figure 7**, Gaussian distributions colored in yellow). Unexpectedly, the apo state shows a narrower distribution with a unique peak centered at 2.0 A° for both α2β1 and α2β2 species. In fact, the conformations sampled by the apo state have a notable overlap with the distribution of holo+ATP, this resemblance being more significant for the α2β2 species. In contrast, the holo state exhibits a wider distribution, showing a bimodal RMSD profile, with peak values at 1.7 and 3.2 A° for α2β1, and at 1.8 and 2.5 A° for α2β2. These results suggest that the binding of PF-739 enhances the fluctuations of P-loop residues that shape the ATP-binding. Due to this higher conformational flexibility, the ATP-binding site can adopt conformations close to those populated in the holo+ATP state, but also visit more dissimilar conformational regions even in comparison with the apo state (**Figure 7**).

Structural Basis of the AMPK Activation by Pan-Activator PF-739 and Its Comparison With Other Direct Activators

To complement the previous analyses, we have examined the interaction network formed by PF-739 and the residues in both α- and β-subunits. To this end, we have clustered the snapshots sampled along the last 500 ns simulation of each replica for both holo α2β1 and α2β2 species, summing a total of 1.5 μs. The results for holo-α2β1 system display up to 4 different clusters, which account for 67.5, 11.8, 10.5 and 10.2% of the conformational ensemble, where the main difference is the conformation adopted by the sugar-like mannitol ring appendage of PF-739 (**Figure 8A**). In all cases two regions can be identified in the interaction network. The first one corresponds to the salt bridge formed between β1-Arg83 and α2-Asp88 (3.0 ± 0.3 Å), which at the same time is hydrogen-bonded to PF-739 (3.5 ± 0.6 Å). For the second cluster (11.8%), an additional interaction between β1-Arg83 and the sugar-like mannitol ring is observed (3.7 ± 0.6 Å; **Figure 8A**). The second region involves salt bridges between pSer108 located at the β-subunit CBM domain and α2-Lys29 (3.7 ± 0.9 Å) and α2-Lys31 (4.4 ± 1.3 Å), both from the P-loop of the α-subunit. Moreover, α2-Lys29 and α2-Lys31 establish contacts with PF-739, such as a hydrogen bond between the Lys31 and the hydroxymethyl-cyclopropyl group (3.2 ± 0.7 Å), which is found in all clusters, and an additional interaction between Lys29 and the N of the benzimidazole ring (3.9 ± 0.9 Å, **Figure 8A**) present in clusters 2 and 3. These interactions networks are very similar to those found in our previous study of SC4 (Aledavood et al., 2021), suggesting that the structural differences between these two compounds, mainly regarding the *o*-toluic substitution of SC4 by mannitol-like ring appendange in PF-739, and the 4′-nitrogen of imidazopyridine in SC4 by a carbon atom in PF-739, do not have a dramatic effect over the interaction at the ADaM site (see also **Supplementary Table S5**). Indeed, these findings remark the key role of the β1-Arg83/β2-Arg82 in the organization of these interactions networks as we explain below.

The cluster analysis performed for the holo-α2β2 system yields four clusters that differ in the orientation of the sugar-like mannitol ring of PF-739, accounting for 76.7, 13.3, 6.0 and 4.0% of the structural ensemble (**Figure 8B**). However, these clusters show higher structural diversity than those determined for the holo-α2β1 system. Thus, two distinct orientations of β2-Arg82 are found in all clusters (**Figure 9**). In one case (**Figure 9A**), β2-Arg82 interacts with α2-Asp88 (3.9 ± 1.3 Å), which forms a hydrogen bond with PF-739 (2.9 ± 0.2 Å). This arrangement represents 54.4% of all the conformations sampled for the α2β2 holo species. In the second orientation β2-Arg82 interacts with β2-Asp111 (3.8 ± 0.9 Å), accounting for 45.6% of the conformational ensemble (**Figure 9B**). Notably, in the α2β1 holo species this latter interaction is not observed, which can be attributed to the substitution of β2-Asp111 by β1-Asn111. The second orientation found for β2-Arg82 reinforces the interaction network observed through β-pSer108, which maintains its interactions with both αLys29 (3.3 ± 0.6) and αLys31 (3.8 ± 1.0) from the P-loop. Additionally, the interaction between αLys31 and the hydroxymethyl-cyclopropyl group (3.2 ± 0.6 Å) of PF-739 is maintained in all clusters, while the interaction between Lys29 and the N of the benzimidazole ring is less stable and only slightly observed in cluster #3 (4.4 ± 0.7 Å, **Figure 8B**).

These results suggest that the arrangement of the sugar-like mannitol unit structural, which exhibit notable differences between clusters, does not have a significant impact on the interaction network observed along the simulations, since the main interactions are preserved in all cases. Indeed, the arrangement of the sugar-like mannitol ring gives rise to new interactions between β1-Arg83 (β2-Arg82) and PF-739 only in cluster #2 (11.8%) for α2β1 and cluster #4 (4.0%) for α2β2. Furthermore, the conformation of the β1-Arg83/β2-Arg82 residue emerges as a key structural feature. While in the α2β1 holo specie, β1-Arg83 forms a salt bridge with α2-Asp88 in all sampled conformations, two orientations are found for β2-Arg82 in the α2β2 holo species (**Figure 9**). This distinctive trait can be attributed to the substitution β1-Asn111 → β2-Asp111, since the presence of β2-Asp111 in α2β2 promotes an electrostatic competition with α2-Asp88 for the interaction with β2-Arg82.

To confirm these results, we have calculated the major interaction pathways identified from WISP analysis for the holo species formed with PF-739. **Figure 10** show the WISP

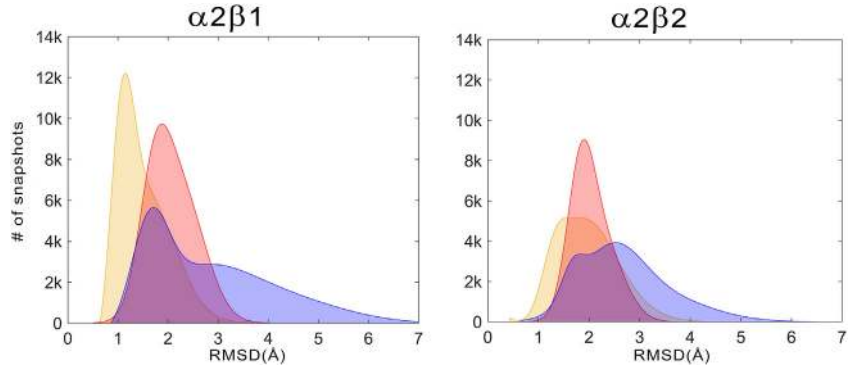


FIGURE 7 | Distribution of the positional deviation (RMSD; Å) of the structures sampled along the trajectories run for apo **(red)**, holo **(blue)**, and holo+ATP **(yellow)** for the residues that shape the ATP-binding site (residues α22–α32, α42–α46, α75–α79, α142–α147, and α153–α157). A total of 60,000 snapshots taken from the last 500 ns of MD simulations were considered for each system in the analysis.

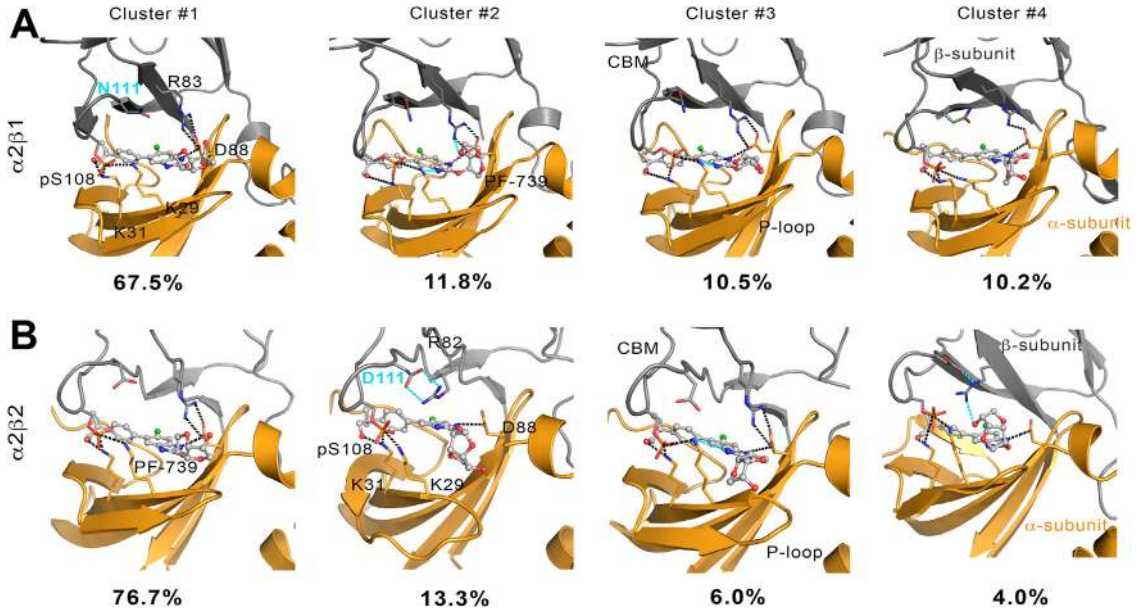


FIGURE 8 | Representation of main interactions between the CBM, P-loop and PF-739 for holo states of **(A)** α2β1 and **(B)** α2β2 species for the four clusters obtained along the last 500 ns of simulation of each replica. The α-subunit is shown in orange cartoon, while the β-subunit is shown in grey cartoons. PF-739 is shown in grey ball and sticks in the ADaM site. Selected polar interactions maintained through all the MD simulations and clusters are highlighted in black dashed lines, while those formed in specific clusters are shown in cyan.

results obtained in our previous work (Ngoei et al., 2018) for A-769662 (**Figure 10A**) and SC4 (**Figure 10B**), as well as the results obtained for PF-739 (**Figure 10C**). For the α2β1/A-769662 complex three major paths are found between the CBM domain and the P-loop, which involve i) pSer108, ii) the hydrophobic core of the ADaM site, and iii) the interaction β1-Arg83-α2-Asp88. All of them are directly connected with the activator through the residues participating in the path, supporting the role of A-769662 as a molecular glue between α2- and β1-subunits. However, only the pSer108 path is observed for the α2β2/A-769662 complex. This can be attributed to the β1-Asn111 → β2-Asp111 substitution, weakens the interaction between β2-Arg82 and α2-Asp88, and strengthens the path through pSer108. In turn, this agrees with the selective activation observed for AMPK complexes containing the β1-isoform. In contrast, two representative paths are found in the holo states formed with SC4 (**Figure 10B**), corresponding to the networks through pSer108 and through the pair β1/2-Arg83–α2-Asp88. Furthermore, SC4 exhibit a similar pattern in both α2β1 and α2β2, which is in agreement with the ability to activate both kinds of AMPK complexes (Hardie, 2014). Interestingly, the β1-Asn111 → β2-Asp111 substitution seems to be less sensitive to the presence of SC4, an effect that can be attributed to the negative charge of the activator that can

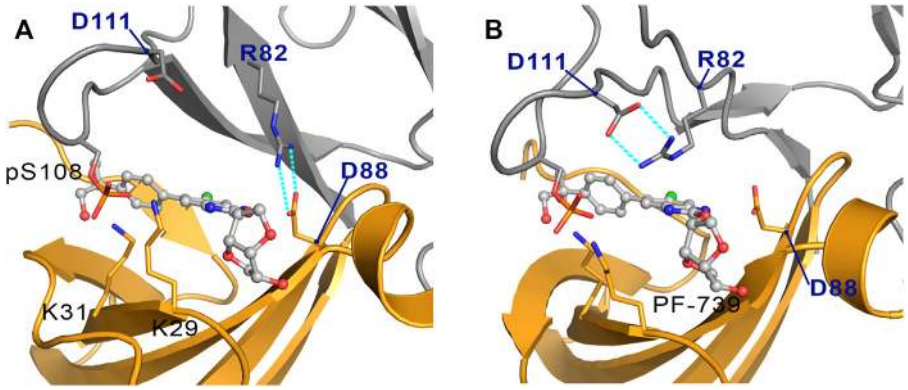


FIGURE 9 | Representation of the two orientations of the β2-Arg82 in α2β2 species, where the interaction with **(A)** the α2-Asp88 and **(B)** the β2-Asp111 are highlighted in cyan dashed lines. The α-subunit is shown in orange cartoon, while the β-subunit is shown in grey cartoons. PF-739 is shown in grey ball and sticks in the ADaM site.

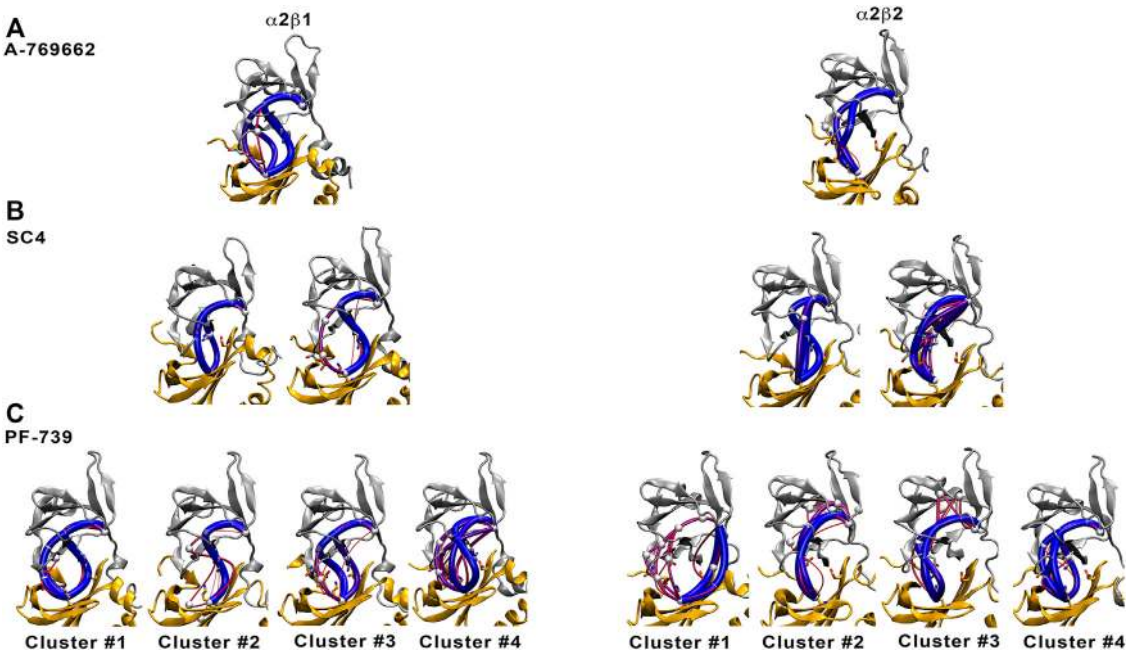


FIGURE 10 | Comparison of major interaction networks obtained from WISP analysis for α2β1 **(left panel)** and α2β2 **(right panel)** species of the holo states for **(A)** A-769662, **(B)** SC4 and **(C)** PF-739 direct activators.

modulate the linking role of β2-Arg82 towards a preferential interaction with either β2-Asp111 and α2-Asp88.

In light of these findings, we have performed the WISP analysis separately for the four main clusters obtained for PF-739 (**Figure 10C**). In the case of the holo-α2β1 state, the three pathways described above for A-769662 can be identified in the whole set of clusters. Although one may notice distinct traits for each cluster, at least two main paths can be observed for clusters #1, #3 and #4. In particular, for the most populated cluster (#1; 67.5%) they correspond to the paths mediated by pSer108 and the pair β1-Arg83–α2-Asp88, respectively. However, the analysis of the holo-α2β2 state reveals a weaker connectivity, since a single path dominates the interaction network in all clusters, For the most populated cluster #1 (76.7%), the path involves the β2-Arg82–α2-Asp88 pair, with a minor contribution of the pSer108-mediated path. In the other clusters, nevertheless, the pSer108 path is predominant, resembling the behavior found for A-769662 (**Figure 10A**, right panel).

These results suggest that the β1-Asn111 → β2-Asp111 substitution plays a critical role in defining the mechanical sensitivity of AMPK to the direct activator. Besides the pSer108-mediated path, the presence of β1-Asn111 in α2β1

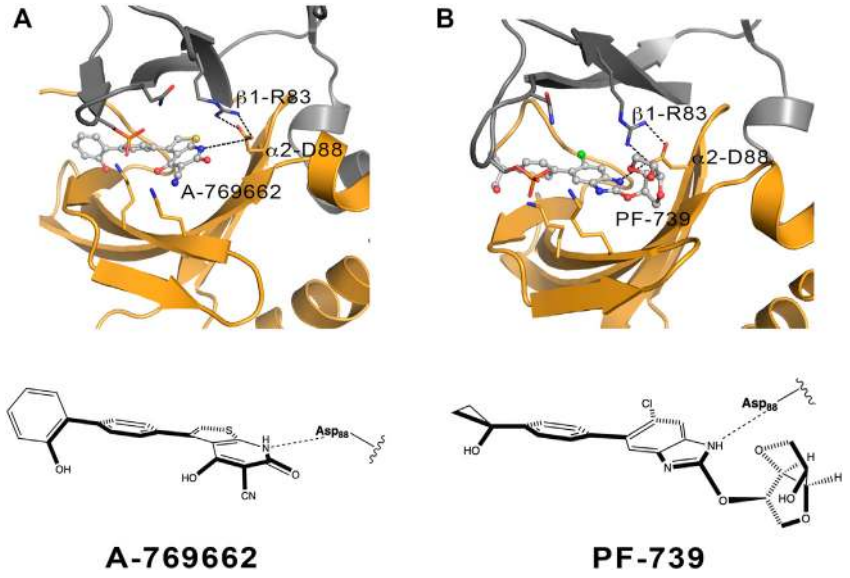


FIGURE 11 | Schematic representation of the α2β1 holo state for **(A)** A-769662 and **(B)** PF-739. The 2D chemical structure of the direct activators is presented with the orientation displayed in the ADaM site. The polar interactions between β1-Arg83, α2-Asp88 and the respective ligands are shown in dashed lines.

favors the formation of an additional path that involves the concerted interaction between β1-Arg83, α2-Asp88, activator and α2-Lys29/α2-Lys31. Nevertheless, the substitution β1-Asn111 → β2-Asp111 favors the breaking of the β2-Arg82–α2-Asp88 interaction and the formation of the salt bridge with β2-Asp111, which reinforces the contribution of the pSer108 path, making the α2β2 complex less sensitive to the modulation by the activator.

The chemical features of the activator also exerts role in assisting the conformational activation of both α2β1 and α2β2 species. The main difference between A-769662 and PF-739 is the replacement of the thienopyridone ring by a benzimidazole derivative with a sugar-like mannitol appendage in PF-739 (**Figure 11**). The β1-Arg83–α2-Asp88–A-769662–α2-Lys29/α2-Lys31 network of interactions acts as a transmission band that connect the dynamical motion of the CBM domain with the P-loop, assisting the effective transition toward conformations that resemble the ATP-binding site in the holo+ATP state for the α2β1 species (**Supplementary Figure S2**, left). However, breakage of this interaction path in the α2β2 holo species prevents the activator to mediate the transmission of the dynamical fluctuations of the CBM domain and the P-loop, which is reflected in a wider conformational distribution of the ATP-binding site (peak centered at 3.0 Å; see **Supplementary Figure S2**, top). This reflects the inability of A-769662 to pre-organize the ATP-binding site in β2-containing AMPK complexes.

The conformational response caused by PF-739 is more complex, reflecting the structural variability of the clusters regarding the orientation of the sugar-like mannitol appendage for both α2β1 and α2β2 species, and the two arrangements of β2-Arg82 in α2β2 compared to the single conformation of β1-Arg83 in α2β1. The analysis of the pre-organization of ATP-binding site (**Figure 7**) reveals that the activator is unable to reduce the conformational sampling to structures well suited for the binding of ATP, which would diminish the activation effect of PF-739. At this point let us remark the bimodal behavior shown in **Figure 7A**, with only 33.3/45.0% of the sampled structures of ATP-binding site resembling the holo+ATP in α2β1/α2β2, whereas A-769662 triggers a marked shift in the population distribution in the holo complex of α2β1 (**Supplementary Figure S2**). On the one side, this agrees with the ability of PF-739 to exert a mild activation in both α2β1 and α2β2. The distribution of holo+ATP–*like* conformations in α2β2 is wider than in α2β1, which reflects the higher structural plasticity observed in α2β2 species. On the other side, these findings are also in agreement with the WISP results, which show how PF-739 activator has higher gluing effect than A-769662 in α2β2, allowing the transmission of the information between α- and β-subunit through the pSer108 and β2-Arg82–α2-Asp88 pathways, explaining in this way why PF-739 acts as a pan-activator.

CONCLUSION

Discerning the molecular factors that regulates the structure-function relationships of AMPK isoforms is of utmost importance to rationalize the tissue-dependent expression of AMPK complexes, and thus enabling the design of specific compounds active against specific metabolic disorders. However, the recognition of the differences between isoforms that allow a different ligand behavior (i.e., selective activator, pan-activator or even inhibitor) is very challenging due to the high structural complexity of the enzyme and the highly correlated dynamics observed for both α2β1 and α2β2 species.

Our results confirmed that the subtle difference of β1-Asn111 to β2-Asp111 has great implications in the dynamical response of

AMPK to the binding of activators. This single substitution can change the interaction networks formed surrounded the activator, thus inducing a better mechanical response of the α2β1 specie towards the interaction of PF-739, than in the case of the α2β2 species. So, even in case of a pan-activator like the PF-739, able to activate both β-isoforms, still subtle residue substitutions in the ADaM site are responsible of difference in affinity towards the isoform. Additionally, we hypothesized that the bulkier substitutions in the chemical structure of the ligands located nearest to the α2-Asp88 residue could involve a higher variability in the conformational space, thus preventing to discern between β-isoforms.

In summary, we were able to characterize the key molecular features that mediate the activation of pan-activator towards α2β1and α2β2 species. All these findings shed light in the comprehension of the role of specific residues in the ADaM site that can modulate or completely change the direct activation mechanism of β1- and β2-containing AMPK complexes. Future studies will be appreciated to distinguish the structural basis of the different sensitivity of AMPK complexes formed by distinct α-subunits, and which is more important, the study of the full complex to disentangle the full allosteric network connection. This understanding will really enable us the design of tissue-selective modulators of this cellular energy sensor.

MATERIALS AND METHODS

Molecular Dynamics Simulations

Extended molecular dynamics (MD) simulations were utilized to analyze the structural and dynamical characteristics of the simulated system. For this purpose, the α2β1γ1 systems were built up using the complexes with A-769662 (PDB entry 4CFF) (Xiao et al., 2013). On the other hand, the system related to the complex of α2β2γ1 bound to SC4 (PDB entry 6B2E) (Ngoei et al., 2018) was also used as a template to model the complexes with PF-739. Following our previous studies, (Aledavood et al., 2019; Aledavood et al., 2021), the γ-subunit was not considered in MD simulations for several reasons. First, the ADaM site is shaped only by α- and β- isoforms. Furthermore, the lack of precise structural information about stretches of both α- and β-subunits, particularly regarding the C-terminal regions, which are located close to the γ-subunit, would introduce an additional level of uncertainty, opening the way to potential artefacts in the simulations. Finally, inclusion of the γ-subunit would have required a larger computational cost to guarantee a proper sampling of the dynamical motions of the three isoforms. Accordingly, following the "divide-and-conquer" strategy outlined above, the simulated systems comprise only α- and β-subunits. Specifically, simulations were performed for residues 8–278 of the α2 isoform, and residues 78–173 and 77–171 of the β1- and β2-isoforms, which were solved without disruptions in the X-ray structures. Finally, these structures were used to model the apo protein, the complexes of the activators bound to the phosphorylated Ser108 (pSer108)-containing isoforms (holo), and the corresponding holo+ATP complexes with both activator in the ADaM site and ATP in the ATP-binding site.

The Molecular dynamic (MD) simulations were performed using the AMBER18 package (Case et al., 2018) and the Amber ff99SBILDN force field (Lindorff-Larsen et al., 2010) for the protein, whereas the ligand (PF-739) were parameterized using the GAFF force field (Wang et al., 2004) in conjunction with restrained electrostatic potential-fitted (RESP) partial atomic charges derived from B3LYP/6-31G(d) calculations (Bayly et al., 1993). The parameters used for the ATP molecule were obtained from the Amber parameters database from Bryce group at the University of Manchester (AMBER parameter database, 2021; Meagher et al., 2003). The standard protonation state at physiological pH was assigned to ionisable residues, and a capping group (N-methyl) was added to the C-terminus of the α-subunit. The simulated systems were immersed in an octahedral box of TIP3P water molecules considering a solute-edge distance of 12 Å (Jorgensen et al., 1983), and counterions atoms were added to maintain the neutrality of the simulated systems (Joung and Cheatham, 2008). The final systems included the AMPK protein (368 residues for α2β1 and 367 residues for α2β2), around 25,000–26,700 water molecules, and a variable number of Na^+ and Cl^- ions, leading to simulated systems containing between 81,000 and 86,000 atoms (specific values are gathered in **Supplementary Table S6**).

Simulations were performed in the NPT ensemble for equilibration and NVT for MD productions using periodic boundary conditions and Ewald sums (grid spacing 1 Å) for treating long-range electrostatic interactions. Apo, holo and holo+ATP systems were simulated in triplicate. The minimization of the systems was performed refining the position of hydrogen atoms in the protein (2,000 cycles of steepest descent algorithm followed by 8,000 cycles of conjugate gradient), subsequently minimizing the position of water molecules (using again the previous scheme), and finally minimization of the whole system (4,000 cycles for steepest descent and 1,000 cycles of conjugate gradient). Later, the temperature of the system was gradually raised from 100 to 300 K in five steps, 50 ps each using the NVT ensemble and Langevin dynamics for the temperature regulation. In this process, suitable restraints (5 kcal mol^{-1} $Å^{-2}$) were imposed to keep the ligand (activator, ATP) in the binding pocket and prevent artefactual rearrangements along the equilibration stage. In order to equilibrate the density of the system an additional 5 ns step performed in the NPT ensemble using the Berendsen barostat. In addition, the restraints were progressively eliminated in this later step. Production MD simulations were run for 1 μs per replica, leading to a total simulation time of 12 μs for the ligand-bound AMPK complexes, and 6 μs for the two apo species of AMPK.

Essential Dynamics

This method was utilized to specify the most important motions from the structural variance sampled in MD simulations. In essential dynamics (ED) (Amadei et al., 1993), the dynamics along the individual modes can be studied and visualized separately, so we can filter the main collective motions during our simulations. Therefore, the positional covariance matrix is created and diagonalized in order to achieve the collective deformation modes, i.e., the eigenvectors, while the eigenvalues

account for the contribution of each motion to the structural variance of the protein. ED analysis was done for 25,000 snapshots from the last 500 ns of each simulation, taking into account only the backbone atoms and the calculations were performed with PCAsuite program (available at http://www.mmb.irbbarcelona.org/software/pcasuite/pcasuite.html), which is integrated in the pyPCcazip program, a suite of tools for compression and analysis of molecular simulations (Shkurti et al., 2016).

Dynamical Perturbation Network

Contact networks represent a protein as a collection of nodes, i.e., the residues that are connected by edges if those residues satisfy a contact condition. Here, in line with previous works (Vuillon and Lesieur, 2015; Dorantes-Gilardi et al., 2018; Gheeraert et al., 2019), the contact condition is satisfied if at least one heavy atom from a residue is at a distance below 5 Å from a heavy atom of another residue. Edges are then weighted by the total number of atomic couples that satisfy this contact condition. Individual contact networks from the frames of one MD simulation are built and averaged (considering the average total number of atomic contacts from various replicas) in order to create a dynamical weighted contact network, which represents a time-averaged contact network associated to the corresponding MD simulations.

To compare MD simulations of a protein in various states (i.e., apo, holo and holo+ATP complexes), we computed perturbation contact networks (Gheeraert et al., 2019) by subtracting two dynamical weighted contact networks associated to each pair of states. To differentiate increases and decreases in contact we assign colors to the edges of the dynamical perturbation network according to the sign of its edges. Finally, for visualization purposes a weight threshold can be applied so that only edges wit a weight greater than the threshold are kept for visualization, here set to 5 as in previous work (Gheeraert et al., 2019). Nodes isolated after this process are also pruned to simplify the visualization.

Dynamic Cross-Correlation Analysis

To complement the information gained from the ED analysis, dynamic cross-correlation (DCC) was used to examine the correlation motion of residues along a given trajectory. To this end, all the snapshots were aligned by means of least-square fitting of Cα atoms of the whole protein to the equilibrated starting configuration. Then, the DCC matrix was determined as noted in **Eq. 1**.

$$C_{ij} = \frac{c_{ij}}{c_{ii}^{1/2} c_{jj}^{1/2}} = \frac{\langle r_i r_j \rangle - \langle r_i r_j \rangle}{\left[\left(\langle r_i^2 \rangle \langle r_i^2 \rangle \right) \left(\langle r_j^2 \rangle - \langle r_j^2 \rangle \right) \right]^{1/2}} \quad (1)$$

where the position vectors of two Cα atoms i and j fitted in the structure at time t are denoted as $r_i(t)$ and $r_j(t)$, respectively.

The cross-correlation coefficients range from −1 to +1, which represent anticorrelated and correlated motions, respectively, whereas values close to zero indicate the absence of correlated motions (Hünenberger et al., 1995). This analysis was performed using the module available in AMBER package. The similarity between the DCC matrices computed for the three replicas run for apo, holo and holo+ATP systems was estimated using the Tanimoto similarity index. This parameter is a distance metrics used to quantify the degree of similarity between two sets of data. While this index is widely adopted to compare the descriptors that characterize the chemical structure of molecules, in this study it is used to compare the correlated motions determined for pairs of residues in the AMPK complexes.

Cluster Analysis

Cluster analysis is a way of determining structure populations from MD simulations. Clustering results in a partitioning data so that data inside a cluster are more similar to each other than they are outside a cluster. In MD, this is a mean of grouping similar conformations together. Similarity is defined by a distance metric, the smaller the distance, the more similar the structures. We used coordinate RMSD as the distance metric parameter. Additionally, we used K-means algorithm as implemented in cpptraj software (Shao et al., 2007), to perform cluster analysis. The K-means identifies k number of centroids, and then allocates every data point to the nearest cluster, while maintaining the centroids as small as possible (Shao et al., 2007). We set the sieve parameter to 10 to reduce the expense of generating the pair-wise distance matrix by using "total/10" frames for initial clustering. The sieved frames are then added to the initial clusters. This analysis was done for 100,000 snapshots from the last 500 ns of each simulation, considering only the backbone atoms.

Interaction Energy Network

Networks of local interactions are intrinsically linked to the structural response of proteins to external factors (O'Rourke et al., 2016). For our purposes, Weighted Implementation of Suboptimal Path (WISP) (Van Wart et al., 2014) was utilized to analyze the allosteric network. This method enabled us to perform a dynamic network analysis to understand how the binding of a ligand in an allosteric cavity could affect another binding site. In particular, WISP relies on the dynamical interdependence among the protein residues. To this end, each amino acid is treated as a node, which was located at the residue center-of-mass, and the interdependence among nodes is represented as a connecting edge with an associated numeric value that reflects its strength. The interdependence is determined from an *NxN* matrix *C* (N is the number of nodes) with values corresponding to the weights of each edge, reflecting the correlated motion among node-node pairs. Finally, the weight between the edge that connects nodes i and j is expressed as $wij = -\log(|Cij|)$, so that highly correlated or anticorrelated motions are characterized by small values of w_{ij}. This analysis was performed for the last 500 ns of the MD simulations.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: http://www.wwpdb.org/, 5UFU; http://www.wwpdb.org/, 6B2E; http://www.wwpdb.org/, 6B1U.

AUTHOR CONTRIBUTIONS

EA: Formal analysis, Investigation, Visualization, Writing–original draft. AG: Formal analysis, Investigation, Visualization. AF: Formal analysis, Investigation, Visualization. LV: Methodology, Investigation. IR: Methodology, Investigation. CE: Conceptualization, Methodology, Investigation, Supervision, Writing–review and editing. FL: Conceptualization, Methodology, Investigation, Supervision, Writing–review and editing, Funding acquisition.

ACKNOWLEDGMENTS

We thank the Spanish Ministerio de Economía y Competitividad (SAF2017-88107-R, and Maria de Maetzu MDM-2017-0767, AEI/FEDER), and the Generalitat de Catalunya (2017SGR1746) for financial support and the Barcelona Supercomputing Center (BCV-2019-2-0017 and BCV-2019-1-0009) and the Consorci de Serveis Universitaris de Catalunya (CSUC) for computational resources. EA thanks AGAUR (Generalitat of Catalunya; 2018FI-B1-00001) for a fellowship. IR and AG acknowledge the support of the Institut Rhônalpin des systèmes complexes, IXXI-ENS-Lyon, Lyon, France, and the use of HPC resources of the "Pôle Scientifique de Modélisation Numérique" (PSMN) at the École Normale Supérieure de Lyon, France.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.760026/full#supplementary-material

REFERENCES

Aledavood, E., Forte, A., Estarellas, C., and Javier Luque, F. (2021). Structural Basis of the Selective Activation of Enzyme Isoforms: Allosteric Response to Activators of β1- and β2-containing AMPK Complexes. *Comput. Struct. Biotechnol. J.* 19, 3394–3406. doi:10.1016/j.csbj.2021.05.056

Aledavood, E., Moraes, G., Lameira, J., Castro, A., Luque, F. J., and Estarellas, C. (2019). Understanding the Mechanism of Direct Activation of AMP-Kinase: Toward a Fine Allosteric Tuning of the Kinase Activity. *J. Chem. Inf. Model.* 59, 2859–2870. doi:10.1021/acs.jcim.8b00890

Amadei, A., Linssen, A. B. M., and Berendsen, H. J. C. (1993). Essential Dynamics of Proteins. *Proteins* 17, 412–425. doi:10.1002/prot.340170408

AMBER parameter database (2021). Bryce Group: Computational Biophysics and Drug Design. Available at: http://amber.manchester.ac.uk.

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: the RESP Model. *J. Phys. Chem.* 97, 10269–10280. doi:10.1021/j100142a004

Calabrese, M. F., Rajamohan, F., Harris, M. S., Caspers, N. L., Magyar, R., Withka, J. M., et al. (2014). Structural Basis for AMPK Activation: Natural and Synthetic Ligands Regulate Kinase Activity from Opposite Poles by Different Molecular Mechanisms. *Structure* 22, 1161–1172. doi:10.1016/j.str.2014.06.009

Carling, D. (2017). AMPK Signalling in Health and Disease. *Curr. Opin. Cel Biol.* 45, 31–37. doi:10.1016/j.ceb.2017.01.005

Carling, D., Thornton, C., Woods, A., and Sanders, M. J. (2012). AMP-activated Protein Kinase: New Regulation, New Roles. *Biochem. J.* 445, 11–27. doi:10.1042/bj20120546

Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., and Cruzeiro, V. W. D. (2018). *AMBER 2018*. San Francisco: University of California.

Chen, L., Wang, J., Zhang, Y.-Y., Yan, S. F., Neumann, D., Schlattner, U., et al. (2012). AMP-activated Protein Kinase Undergoes Nucleotide-dependent Conformational Changes. *Nat. Struct. Mol. Biol.* 19, 716–718. doi:10.1038/nsmb.2319

Cokorinos, E. C., Delmore, J., Reyes, A. R., Albuquerque, B., Kjøbsted, R., Jørgensen, N. O., et al. (2017). Activation of Skeletal Muscle AMPK Promotes Glucose Disposal and Glucose Lowering in Non-human Primates and Mice. *Cel Metab.* 25, 1147–1159. doi:10.1016/j.cmet.2017.04.010

Cool, B., Zinker, B., Chiou, W., Kifle, L., Cao, N., Perham, M., et al. (2006). Identification and Characterization of a Small Molecule AMPK Activator that Treats Key Components of Type 2 Diabetes and the Metabolic Syndrome. *Cel Metab.* 3, 403–416. doi:10.1016/j.cmet.2006.05.005

Dorantes-Gilardi, R., Bourgeat, L., Pacini, L., Vuillon, L., and Lesieur, C. (2018). In Proteins, the Structural Responses of a Position to Mutation Rely on the Goldilocks Principle: Not Too many Links, Not Too Few. *Phys. Chem. Chem. Phys.* 20, 25399–25410. doi:10.1039/c8cp04530e

Gheeraert, A., Pacini, L., Batista, V. S., Vuillon, L., Lesieur, C., and Rivalta, I. (2019). Exploring Allosteric Pathways of a V-type Enzyme with Dynamical Perturbation Networks. *J. Phys. Chem. B* 123, 3452–3461. doi:10.1021/acs.jpcb.9b01294

Hardie, D. G. (2014). AMPK-sensing Energy while Talking to Other Signaling Pathways. *Cel Metab.* 20, 939–952. doi:10.1016/j.cmet.2014.09.013

Human Protein Atlas (2021). Human Protein Atlas. available at: http://www.proteinatlas.org.

Hünenberger, P. H., Mark, A., and Mark, W. F. (1995). Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol.* 252, 492–503. doi:10.1006/jmbi.1995.0514

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79, 926–935. doi:10.1063/1.445869

Joung, I. S., and Cheatham, T. E. (2008). Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* 112, 9020–9041. doi:10.1021/jp8001614

Langendorf, C. G., and Kemp, B. E. (2015). Choreography of AMPK Activation. *Cell Res* 25, 5–6. doi:10.1038/cr.2014.163

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field. *Proteins* 78, 1950–1958. doi:10.1002/prot.22711

Mahlapuu, M., Johansson, C., Lindgren, K., Hjälm, G., Barnes, B. R., Krook, A., et al. (2004). Expression Profiling of the γ-subunit Isoforms of AMP-Activated Protein Kinase Suggests a Major Role for γ3 in white Skeletal Muscle. *Am. J. Physiology-Endocrinology Metab.* 286, E194–E200. doi:10.1152/ajpendo.00147.2003

Meagher, K. L., Redman, L. T., and Carlson, H. A. (2003). Development of Polyphosphate Parameters for Use with the AMBER Force Field. *J. Comput. Chem.* 24, 1016–1025. doi:10.1002/jcc.10262

Myers, R. W., Guan, H.-P., Ehrhart, J., Petrov, A., Prahalada, S., Tozzo, E., et al. (2017). Systemic Pan-AMPK Activator MK-8722 Improves Glucose Homeostasis but Induces Cardiac Hypertrophy. *Science* 357, 507–511. doi:10.1126/science.aah5582

Ngoei, K. R. W., Langendorf, C. G., Ling, N. X. Y., Hoque, A., Varghese, S., Camerino, M. A., et al. (2018). Structural Determinants for Small-Molecule Activation of Skeletal Muscle AMPK α2β2γ1 by the Glucose Importagog SC4. *Cel Chem. Biol.* 25, 728–737. doi:10.1016/j.chembiol.2018.03.008

O'Rourke, K. F., Gorman, S. D., and Boehr, D. D. (2016). Biophysical and Computational Methods to Analyze Amino Acid Interaction Networks in Proteins. *Comput. Struct. Biotechnol. J.* 14, 245–251. doi:10.1016/j.csbj.2016.06.002

Olivier, S., Foretz, M., and Viollet, B. (2018). Promise and Challenges for Direct Small Molecule AMPK Activators. *Biochem. Pharmacol.* 153, 147–158. doi:10.1016/j.bcp.2018.01.049

Scott, J. W., Hawley, S. A., Green, K. A., Anis, M., Stewart, G., Scullion, G. A., et al. (2004). CBS Domains Form Energy-Sensing Modules Whose Binding of Adenosine Ligands Is Disrupted by Disease Mutations. *J. Clin. Invest.* 113, 274–284. doi:10.1172/jci19874

Scott, J. W., van Denderen, B. J. W., Jorgensen, S. B., Honeyman, J. E., Steinberg, G. R., Oakhill, J. S., et al. (2008). Thienopyridone Drugs Are Selective Activators of AMP-Activated Protein Kinase β1-Containing Complexes. *Chem. Biol.* 15, 1220–1230. doi:10.1016/j.chembiol.2008.10.005

Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E. (2007). Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theor. Comput.* 3 (6), 2312–2334. doi:10.1021/ct700119m

Shkurti, A., Goni, R., Andrio, P., Breitmoser, E., Bethune, I., Orozco, M., et al. (2016). pyPcazip: A PCA-Based Toolkit for Compression and Analysis of Molecular Simulation Data. *SoftwareX* 5, 44–50. doi:10.1016/j.softx.2016.04.002

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based Map of the Human Proteome. *Science* 347, 1260419. doi:10.1126/science.1260419

Van Wart, A. T., Durrant, J., Votapka, L., and Amaro, R. E. (2014). Weighted Implementation of Suboptimal Paths (WISP): an Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theor. Comput.* 10, 511–517. doi:10.1021/ct4008603

Vazquez-Martin, A., Vellon, L., Quirós, P. M., Cufí, S., Ruiz de Galarreta, E., Oliveras-Ferraros, C., et al. (2012). Activation of AMP-Activated Protein Kinase (AMPK) Provides a Metabolic Barrier to Reprogramming Somatic Cells into Stem Cells. *Cell Cycle* 11, 974–989. doi:10.4161/cc.11.5.19450

Vuillon, L., and Lesieur, C. (2015). From Local to Global Changes in Proteins: a Network View. *Curr. Opin. Struct. Biol.* 31, 1–8. doi:10.1016/j.sbi.2015.02.015

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and Testing of a General AMBER Force Field. *J. Comput. Chem.* 25, 1157–1174. doi:10.1002/jcc.20035

Willows, R., Sanders, M. J., Xiao, B., Patel, B. R., Martin, S. R., Read, J., et al. (2017). Phosphorylation of AMPK by Upstream Kinases Is Required for Activity in Mammalian Cells. *Biochem. J.* 474, 3059–3073. doi:10.1042/bcj20170458

Xiao, B., Sanders, M. J., Carmena, D., Bright, N. J., Haire, L. F., Underwood, E., et al. (2013). Structural Basis of AMPK Regulation by Small Molecule Activators. *Nat. Commun.* 4, 3017. doi:10.1038/ncomms4017

Xiao, B., Sanders, M. J., Underwood, E., Heath, R., Mayer, F. V., Carmena, D., et al. (2011). Structure of Mammalian AMPK and its Regulation by ADP. *Nature* 472, 230–233. doi:10.1038/nature09932

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

3.2 Distinct allosteric pathways in Imidazole Glycerol Phosphate Synthase from *T. maritima* and *S. cerevisiae*

3.2.1 Structural and functional comparison between the enzymes

Our collaboration with the group of Prof. Victor S. Batista at Yale University initially focused on computational studies of allosteric pathways of IGPS from *T. maritima* . Numerous experimental studies on the allostery of IGPS from *S. cerevisiae* exists but to prior to this thesis, its allosteric pathways remained unknown. IGPS from *T. maritima* and *S. cerevisiae* is then a good test case to assess how evolution shape allosteric pathways, and especially to understand if they can be conserved in such distanced species (a bacteria and a fungi).

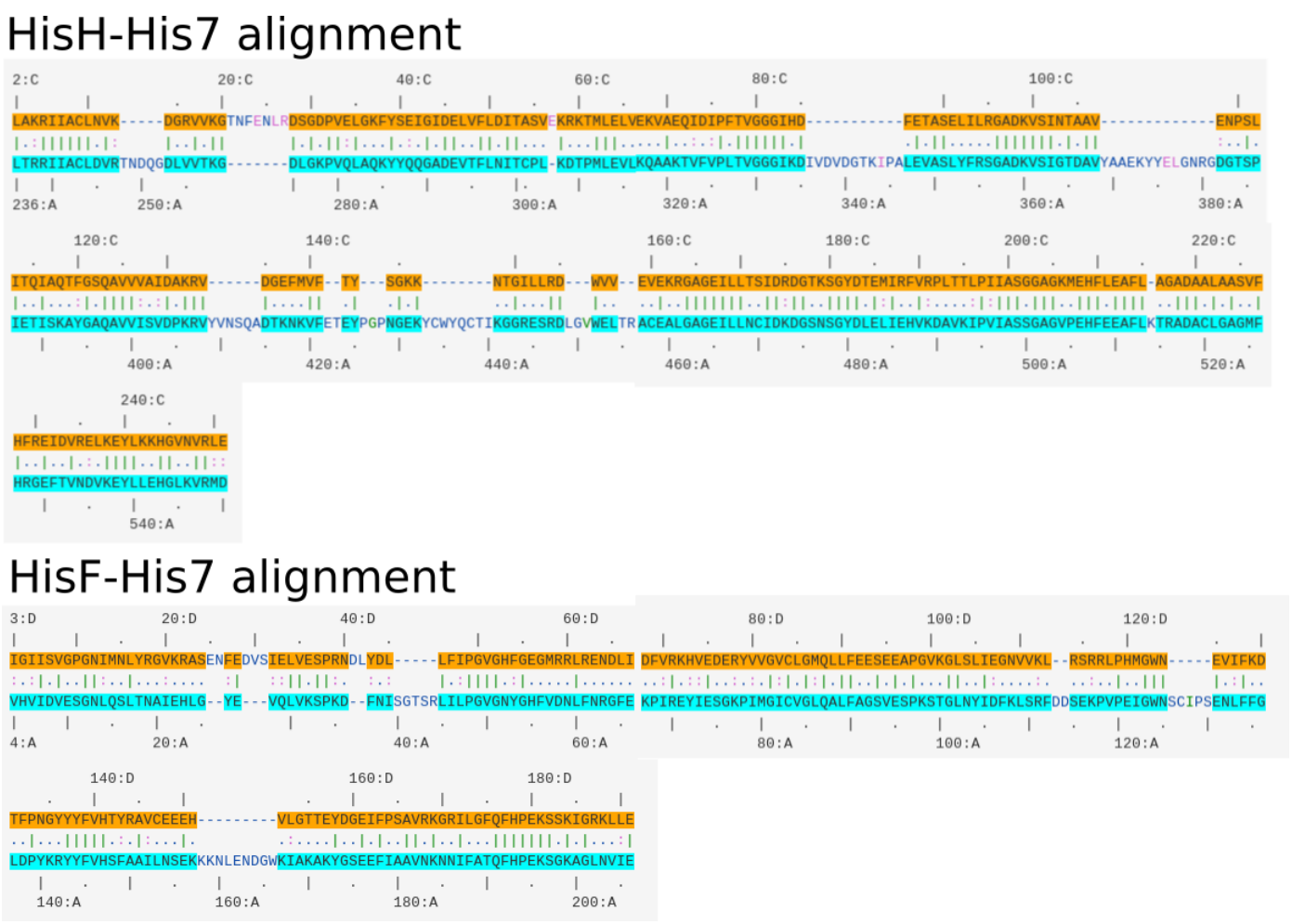


Figure 3.3: Amino acid alignment after using the Smith-Waterman structural alignment algorithm

Alignment	RMSD (in Å)	Gaps (%)	Identity	Similarity
HisH-His7	5.77	62 (20.39%)	40.46%	53.95%
HisF-His7	3.28	28 (13.08%)	28.97%	51.40%

Table 3.1: RMSD, identity and similarity of the alignments between the two chains of IGPS from *T. maritima* and IGPS from *S. cerevisiae*

The most importance difference between IGPS from *T. maritima* and *S. cerevisiae* is the fact that the first is a heterodimer composed by two chains, while the latter is a monomer. HisH and HisF from *T. maritima* respectively have 201 and 253 amino acids, and His7 from *S. cerevisiae* has 534. Using the Smith-Waterman algorithm[1] implemented in the RCSB PDB Comparison Tool Reference[2], we aligned structurally the amino acid sequence of HisH (PDB entry 1GPW.C), HisF (PDB entry 1GPW.D) and His7 (PDB entry 1OX4.A) (see Fig. 3.3). This shows that the structures of HisF and His7 are aligned between residues hI3–hE191 and V4–E206 and the sequence of HisF and His7 are aligned between residues fL2–fE251 and L236–D550. Despite only having a similarity of about 50% (see Table 3.1), these alignments shows a good structural identity with a RMSD between 3.28 and 5.77. The alignments of the 3D structures is also reported on Fig. 3.4. Looking more specifically at the allosteric pathways, we know that in *T. maritima* one key residue to begin the propagation

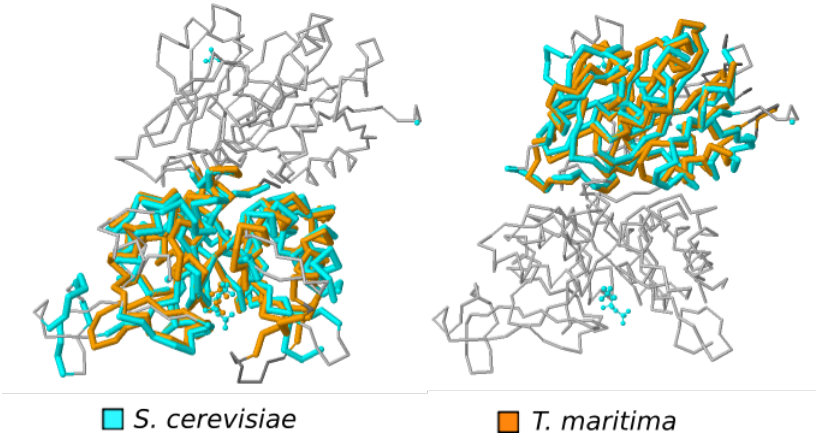


Figure 3.4: 3D structural alignment between IGPS from *T. maritima* (orange) and *S. cerevisiae* (cyan)

of perturbation is fK19 and its pairing with fD11. In *S. cerevisiae* they are both conserved with respectively residue K258 and D245. To get the crystal structure of 1GPW, the fD11N mutation was engineered. So considering the wild type IGPS from *T. maritima*, fD11 residue is conserved by D245. Upon PRFAR-binding, a hydrophobic cluster formation is reported in IGPS from *T. maritima* involving residues fF23, fV48, fL50 and fI52. In *S. cerevisiae* fF23 is absent, fV48 is substituted into T295 while L297 and I299 are conserved. This suggests that a different mechanism could be at play here. Notably, residue fF23 in loop1 establishes the connection with residues in the $f\beta$1 sheet. In *S. cerevisiae* the segment between fT21 and fR28 is deleted.

Another important part of the propagation mechanism is the alteration of the salt-bridge network involving residues fE91 fR95, fE67 and hR18. None of these residues are conserved in the structural alignment. This suggests that there again a different mechanism may be at play. Notably in IGPS from *T. maritima* these residues establishes the connection between HisF and HisH. It may be that the perturbation spread differently when the two chains are fused together. Upon PRFAR bonding, after the propagation from HisF to HisH, in *T. maritima*, the allosteric mechanisms channels through the disruption of the contact between hN12 and hN15. Only the first of these two residues is conserved with N13, the latter being replaced by S15, which is also versatile in hydrogen bonding. In *T. maritima* the last crucial bit of the mechanism is the disruption of hydrogen bond between residue hP10 and hV51 upon effector binding. The 49-PGVG strand is conserved in *S. cerevisiae* (and actually in every IGPS) but residue hP10 structurally overlaps with a dissimilar S11. After the 49-PGVG strand, the allosteric mechanism ends up in the catalytic triad hC84–hH178–hE180 completely conserved in *S. cerevisiae* which is of course conserved with residues C83, H193 and E195.

While key elements of the allosteric pathways are conserved, others are not. Intriguingly, the conserved

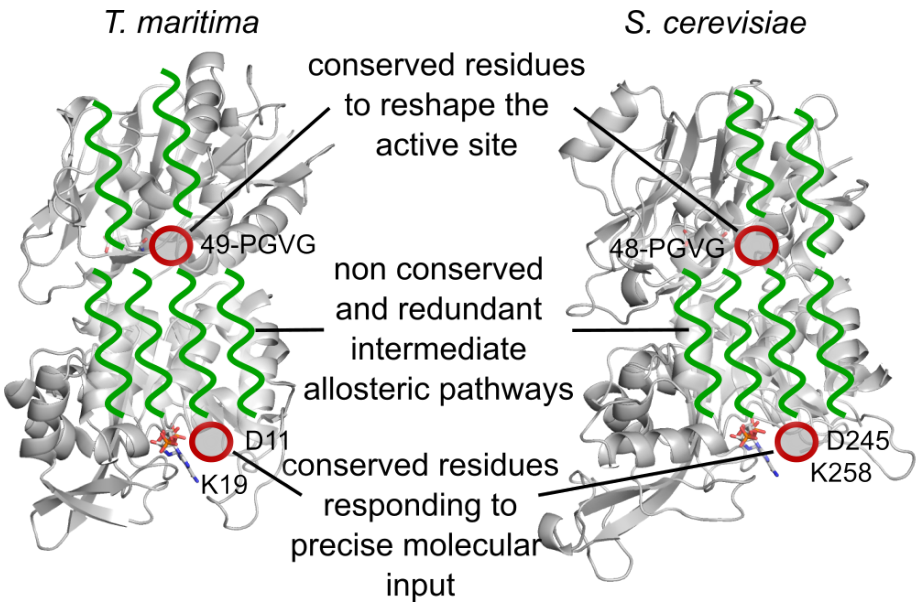


Figure 3.5: The endpoints' specificity hypothesis in allostery.

elements are either close to the effector site (fK19, fD11) or to the active site (PGVG strand) as reported in

Figure 3.5. One hypothesis is that at both ends (effector binding and substrate binding) the allosteric mechanism has to be very precise and residues involved in this particular mechanism should be conserved. Some degree of precision is required near the effector site so that allosteric effects are only triggered in response to a specific molecular input. In a V-type allosteric enzyme, the endpoint of the allosteric pathways is to reshape the active site cavity. By contrast with this view, the allosteric pathways are generally considered redundant, however it seems here that the redundancy affects mainly the intermediate pathways. This redundancy allows for the evolution of intermediate pathways, thus, key allosteric residues can differ from protein homolog to protein homolog.

3.2.2 Kinetics comparison between the enzymes

Michaelis–Menten steady state kinetics[3] of IGPS from *T. maritima*[4] and *S. cerevisiae*[5] have previously been experimentally determined, and the results are reported in Table 3.2. In both enzymes, the kinetics constants were determined in the absence of ligand (apo) and in saturation of PRFAR (holo). The catalytic rate constant k_{cat} is compared with the Michaelis constant K_M which corresponds to the concentration of substrate at which the reaction rate is at half-maximum. This Michaelis constant is inversely proportional to the affinity of the substrate for the enzyme. The constant k_{cat}/K_M (catalytic efficiency) is a measure of the efficiency at which an enzyme converts the substrate into a product.

	apo			holo			
Organism	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M	holo/apo
T. maritima[4]	3.72×10^{-3}	4.91	0.76	4.09	1.30	3150	4161
S. cerevisiae[5]	5.50×10^{-3}	4.70	1.18	6.80	1.20	5800	4900

Table 3.2: Michaelis–Menten steady stade kinetics of IGPS from *T. maritima* and *S. cerevisiae*

Interestingly, *T. maritima* and *S. cerevisiae* feature very similar Michael–Mentis parameters which always remains in the same order of magnitude in the apo and holo form. IGPS from *S. cerevisiae* has a slightly faster catalytic rate in apo and holo combined with a slightly bigger affinity, which in turn makes for a higher catalytic efficiency. Upon PRFAR-binding, it is the catalytic activity increase (3 orders of magnitude) that is principally responsible for the dramatic increase in catalytic efficiency (also 3 orders of magnitude) which shows that IGPS is a V-type allosteric enzyme in the two organisms.

Interestingly, both systems, despite featuring some structural differences in allosteric pathways, have almost identical kinetics. This can be probably attributed to the conservation of the catalytic site, the PGVG oxyanion strand and residue L85. It would appear here as if that the non-conservation of intermediate allosteric pathways has little to no effect on allosteric kinetics. If effectively demonstrated this effect could have huge repercussion in designing non-competitive allosteric inhibitors.

3.2.3 Molecular dynamics simulations

To provide computational structural biology elements, we ran MD simulations of the IGPS *S. cerevisiae* . IGPS from *S. cerevisiae* possess different crystal structures, including one with PRFAR-bound (PDB entry 1OX5). However, these structures include some missing segments. We built six different models in total, one using homology modeling for the missing loops and five using general purpose homology modeling. We then constructed for each model, the apo and holo models and ran MD simulations for 1 µs. For analysis, we focused extensively on the first 100 ns of simulation for a better comparison of allosteric pathways from *T. maritima* with previous references[6, 7, 8].

3.2.4 Allosteric pathways comparison

In general, we found that the allosteric pathways in the two enzymes are very different at every scale. Most of the key amino acids in *S. cerevisiae* IGPS allosteric pathways are different from those of *T. maritima* IGPS and moreover they are not located in the same secondary structure elements. In terms of global motions, *S. cerevisiae* is also vastly different from *T. maritima* featuring no breathing motion and overall different alterations of motion upon effector binding. Finally, the endpoint of *S. cerevisiae* IGPS allosteric mechanism is still the PGVG oxyanion strand. Still, in opposition with *T. maritima* IGPS, where the reason for PGVG flipping is the breaking of a hydrogen bond with the Ω-loop, in *S. cerevisiae* IGPS, these two secondary structures are not linked by a hydrogen bond, and it is the Ω-loop increase in flexibility that transmits to the PGVG segment. In consistence with the structural analysis here, we found that the allosteric pathways in *S. cerevisiae* are not evolutionary conserved, which validates our endpoint hypothesis.

References

[1] Temple F Smith, Michael S Waterman, et al. "Identification of common molecular subsequences". In: *J. molecular biology* 147.1 (1981), pp. 195–197.

[2] Andreas Prlić et al. "Pre-calculated protein structure alignments at the RCSB PDB website". In: *Bioinformatics* 26.23 (2010), pp. 2983–2985.

[3] Leonor Michaelis and Maud Leonora Menten. "Die Kinetik der Invertinwirkung". In: *Biochem Z* 49 (1913), pp. 333–369.

[4] George P Lisi, Allen A Currier, and J Patrick Loria. "Glutamine hydrolysis by imidazole glycerol phosphate synthase displays temperature dependent allosteric activation". In: *Front. molecular biosciences* 5 (2018), p. 4.

[5] Rebecca S Myers et al. "Reaction coupling through interdomain contacts in imidazole glycerol phosphate synthase". In: *Biochemistry* 44.36 (2005), pp. 11974–11985.

[6] Ivan Rivalta et al. "Allosteric pathways in imidazole glycerol phosphate synthase". In: *Proc. National Acad. Sci.* 109.22 (2012), E1428–E1436.

[7] Christian FA Negre et al. "Eigenvector centrality for characterization of protein allosteric pathways". In: *Proc. National Acad. Sci.* 115.52 (2018), E12201–E12208.

[8] Aria Gheeraert et al. "Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks". In: *The J. Phys. Chem. B* 123.16 (2019), pp. 3452–3461.

3.2.5 Published Article 3

This work started in our team in Bologna, and we offered our collaborators in Yale University to participate. This led to the publication of an article in the Biophysical Journal.

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

Biophysical *Journal*
Article

Biophysical Society

Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria

Federica Maschietto,[1] Aria Gheeraert,[2] Andrea Piazzi,[3] Victor S. Batista,[1,*] and Ivan Rivalta[2,3,*]
[1]Department of Chemistry, Yale University, New Haven, Connecticut; [2]Université de Lyon, CNRS, Institut de Chimie de Lyon, École Normale Supérieure de Lyon, Lyon Cedex 07, France; and [3]Dipartimento di Chimica Industriale "Toso Montanari", Alma Mater Studiorum, Università di Bologna, Bologna, Italia

ABSTRACT Understanding the relationship between protein structures and their function is still an open question that becomes very challenging when allostery plays an important functional role. Allosteric proteins, in fact, exploit different ranges of motions (from sidechain local fluctuations to long-range collective motions) to effectively couple distant binding sites, and of particular interest is whether allosteric proteins of the same families with similar functions and structures also necessarily share the same allosteric mechanisms. Here, we compared the early dynamics initiating the allosteric communication of a prototypical allosteric enzyme from two different organisms, i.e., the imidazole glycerol phosphate synthase (IGPS) enzymes from the thermophilic bacteria and the yeast, working at high and room temperatures, respectively. By combining molecular dynamics simulations and network models derived from graph theory, we found rather distinct early allosteric dynamics in the IGPS from the two organisms, involving significatively different allosteric pathways in terms of both local and collective motions. Given the successful prediction of key allosteric residues in the bacterial IGPS, whose mutation disrupts its allosteric communication, the outcome of this study paves the way for future experimental studies on the yeast IGPS that could foster therapeutic applications by exploiting the control of IGPS enzyme allostery.

SIGNIFICANCE Allosteric regulation is widely present in macromolecules and is essential to coordinate biochemical information transfer between spatially distant sites. Despite the growing interest dedicated to uncovering the mechanism of allosteric processes, the question of how allosteric enzymes from different evolutionary paths achieve the same catalytic function remains elusive. We examine the allosteric pathways of the imidazole glycerol phosphate synthase (IGPS) enzymes from yeast and thermophilic bacteria through the lens of molecular dynamics simulations and graph-theory-based network models. We find that protein-specific cooperative interactions between local and collective modes accomplish the same function of activating the catalytic site upon effector binding to the allosteric site, allowing the two enzymes to optimally function in their (different) natural environments.

INTRODUCTION

Allostery is an essential regulatory process of biological macromolecules of great interest for a wide range of applications, including drug discovery and gene-editing technologies (1–4). Allosteric mechanisms typically transmit the effect of binding of a ligand effector to a distant site, often responsible for catalytic activity (5). Targeting the signal transduction mechanism between the allosteric and catalytic sites can lead to suppression of substrate turnover at the active site, opening an opportunity for protein engineering or development of non-competitive small molecule inhibitors. An advantage of allosteric drugs is that they selectively tune responses in tissues where the endogenous agonists exert their physiological effects and only when the endogenous agonists are present (6). Such spatial and temporal selectivity cannot be achieved with traditional orthosteric agonists since those modify the receptor function continuously as long as they are present. Another important advantage is the intrinsic safety in overdosage since, once the allosteric sites are occupied, no further allosteric effect can be produced even with excessive doses (5,7). An outstanding challenge, however, is the development of fundamental understanding of allosteric pathways in

Submitted May 10, 2021, and accepted for publication November 29, 2021.
*Correspondence: victor.batista@yale.edu or i.rivalta@unibo.it
Editor: Alexandr Kornev
https://doi.org/10.1016/j.bpj.2021.11.2888

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

proteins (1,8–11). In fact, an allosteric mechanism encompasses all steps that are involved in the signal transduction extending from the effector to the active site. These steps include effector binding, allosteric communication (via local contacts and collective motions; i.e., the allosteric pathways) triggering alterations (usually associated to conformational changes) of the active site. It has been proposed that allostery may be an intrinsic property of virtually all proteins (12); however, the extent of conservation of allosteric mechanisms or absence of it across a protein family remains an open question (13). In fact, the similarity of protein structures does not necessarily imply a common function (proteins with different functions can share a common structural framework while the same function can be performed by proteins with different folds), suggesting that the structure/function relationship can be quite complex in terms of allostery (12).

On one side, there are examples of proteins with similar functions and structures retaining similar allosteric pathways that have been reported (14–17), pointing out the role of conserved network of residues in allostery. On the other hand, various studies have reported differences in the allosteric communication between protein homologs. For instance, the structural study of three bacterial chemotaxis protein Y orthologs showed divergent allosteric responses across the protein family, with allosteric signals found to be globally propagated in different, system-dependent, ways (18). Moreover, the characterization of three homologous of the HIV-1 envelope spike allostery has suggested that, despite the common modular structure of the allosteric network that remains highly conserved, the shortest path for communication between distal regions is sensitive to differences in the primary sequences of the individual proteins (19).

Therefore, the assumption that proteins with similar structures would have similar allosteric pathways is not always true, since allosteric communication in protein orthologs is often system specific (18,20,21). So, the extent to which allosteric pathways are conserved among protein orthologs remains an open question (22).

The intrinsic complexity of the question of conservation of allosteric pathways is due to the fact that differences in the allosteric communication between protein homologs can occur at different levels of the allosteric signaling pathways, i.e., involving both changes in local contacts and/or collective motions, suggesting that a detailed knowledge of these communication pathways is required. Here, in response to reviewers, we address this question for the allosteric pathways of imidazole glycerol phosphate synthase (IGPS) enzymes from two different organisms, bacteria (*Thermotoga maritima* [*Tm*]) and yeast (*Saccharomyces cerevisiae* [*Sc*]). IGPS enzymes are ideal for our analysis since they are prototypical systems for the study of allostery and have already attracted significant interest as targets for therapeutic applications (23–30). Our study is focused on understanding how these two allosteric enzymes with different evolutionary paths achieve the same allosteric function despite the significant differences in their primary sequences, and secondary structures. As a consequence of their structural analogy, IGPS enzymes from yeast and bacteria feature the same effector-binding site (31,32) and glutaminase active site (with analogous inactive/active conformations) (30). We focus on the characterization of their allosteric pathways (those of *Sc*-IGPS being unknown), exploring both local contacts and collective motion contributions to analyze whether or not the two enzymes have the same allosteric mechanism. We find that the early dynamics that initiate allosteric communication are rather different for the two enzymes, resulting in distinct allosteric pathways tailored for activity in the different natural environments of the two enzymes. Thermophiles exhibit robust functionality at high temperatures, while *Saccharomyces* function at room temperature. Their early allosteric dynamics involve differences in both collective motions and inter-residue interactions, which are likely due to the different adaptations of the enzymes to their native conditions.

Structural features of IGPS enzymes from *Thermophiles* and *Saccharomyces*

We begin by summarizing the similarities and structural differences between the two IGPS enzymes. In bacteria, IGPS is a tightly associated heterodimer complex formed by the glutaminase subunit HisH and the cyclase HisF (red and salmon, respectively, in Fig. 1 *A*) (33). In yeast *Sc*-IGPS, the two subunits are fused into a single polypeptide chain, His7 (green, in Fig. 1 *A*) with the two functional domains linked by a short polypeptide (i.e., the connector, circled in Fig. 1 *B*) (31). The aligned complexes share the same fold, as shown in Fig. 1 *A*, with a sequence similarity of 52% and 63% for HisH and HisF, and an RMSD of C-alpha carbon atoms of 1.93 and 2.03 Å, respectively (see sequence alignment in Table S2). Throughout this paper, we refer to secondary structural elements by increasing numbering and labeling the residues corresponding to HisH and HisF with prefixes *h* and *f*, respectively, following the standard *Tm*-IGPS nomenclature (11). The full topography of secondary structural elements of yeast IGPS is reported in the supporting material (Table S1), and for *Tm*-IGPS is reported in reference (33).

The same two reactions are catalyzed by the two domains of both IGPS enzymes from thermophiles and yeast. In the glutaminase domain, glutamine (Gln) is hydrolyzed to glutamate, releasing ammonia that migrates (29,31–34) to the cyclase domain, where it reacts with the effector 5-[(5-phospho-1-deoxy-D-ribulos-1-ylimino)methylamino]-1-(5-phospho-beta-D-ribosyl)imidazole-4-carboxamide (PRFAR) to form imidazoleglycerol phosphate (ImGP), a precursor to histidine and 5'-(5-aminoimidazole-4-carboxamide) (AlCAR),

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

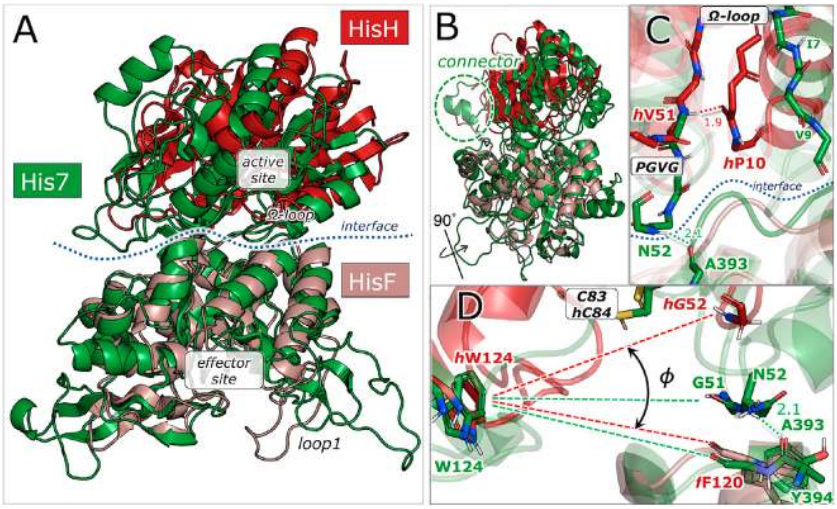


FIGURE 1 Molecular representation of IGPS from thermophile and yeast. (*A*) Front view of *Tm*-IGPS (red and salmon) as compared with *Sc*-IGPS from yeast *S. cerevisiae* (green). The structures are derived from the PDB models 1GPW and 1OX6 where the missing residues have been reconstructed (as described in the supporting material). The active site in the glutaminase domain and the effector site in the cyclase domain are more than 25 Å apart in both *Tm*-IGPS and *Sc*-IGPS. (*B*) Side view of aligned *Tm*-IGPS and *Sc*-IGPS structures, highlighting the position of the connector between the cyclase and glutaminase domains of His7. (*C*) Close-up of the glutaminase active site in *Tm*- and *Sc*-IGPS (3ZR4 and 1OX5, respectively), showing structural differences next to the active site (Gln substrate not shown). Loop β3α2 with hV51 is tightly bound to P10 in *Tm*-IGPS, but shifted toward the cyclase domain in His7. The β3α2 loop (also known as the PGVG strand) is highly conserved in all IGPS enzymes and is thought to stabilize the oxyanion intermediate formed during the catalytic reaction. (*D*) Close-up view of the interface between the cyclase and glutaminase domains in yeast and bacterial IGPS. In His7, the interface is closed, with the angle $\varphi = 15°$ between G51:Cα↔W124:Cγ↔Y394:Cγ, spanned by the green dotted lines, in the crystal structure (PDB: 1OX6) (34). In apo-*Tm* IGPS, however, the HisF:HisH interface is wide open. In the crystal structure (PDB: 1GPW) (33), the corresponding angle (*h*G52:Cα↔hW123:Cγ↔fF120:Cγ) $\varphi = 29°$ (between red dotted lines).

used in the synthesis of purines. While Gln hydrolysis could occur in the absence of the effector, the reaction is accelerated 5000-fold upon PRFAR binding, classifying IGPS as a V-type allosteric enzyme. Recent studies of *Tm*-IGPS (28) have shown that Gln has a different affinity for the enzyme with or without effector, although the major increase in turnover (K_{cat}) is predominant over the change in substrate dissociation constant, K_m^{Gln}.

Experimental and computational studies on PRFAR-bound and PRFAR-free forms of IGPS enzymes have identified flexible parts of the protein with potential allosteric roles in the communication between the effector and catalytic sites (23,26–28,31,34,35). These previous studies have provided evidence of an unformed oxyanion hole as the basis for low glutaminase activity in the effector-free form of the enzyme (31,34,36,37). The term "oxyanion hole" derives from the presence of a negatively charged oxygen on the Gln, generated by the reaction of the cysteine sulfur in the active site and the Gln substrate. The hole generated by the amino acid residues surrounding the anion stabilizes the negative charge before a neutral environment is restored.

The highly conserved sequence in the β3α2 loop of all IGPS enzymes, known as the PGVG (oxyanion) strand next to the glutaminase active site, hosts the charged intermediate. However, the crystal structures of IGPS from both yeast (31,34) and bacteria (33) suggest that the PGVG β strand has an improper conformation in the apo enzymes, with the NH group of *h*V51/V50 pointing out from the Gln-binding site. Therefore, a 180° turn of the whole oxyanion strand is necessary to stabilize the tetrahedral intermediate and to make the glutaminase enzyme catalytically active. Earlier studies are consistent with the formation of the oxyanion hole as the endpoint of the allosteric mechanism in IGPS enzymes (23,28,30,31,36).

Unlike allostery in bacterial IGPS, the allosteric pathway in IGPS from yeast remains uncertain. The comparative structural analysis of the two enzymes suggests that different allosteric mechanisms might operate in the two systems. For example, the PGVG strand in *Tm*-IGPS is more distant from the cyclase domain compared with *Sc*-IGPS. Further, the *h*V51-*h*P10 hydrogen bond (H-bond) that connects the PGVG oxyanion strand with the neighboring Ω-loop has been shown to be crucial in the allosteric mechanism of *Tm*-IGPS (23,38), although it is absent in *Sc*-IGPS (see Fig. 1 *C*) (31,34).

The cyclase:glutaminase interface in the single-chain *Sc*-IGPS is tighter than in the *Tm*-IGPS heterodimer (see Fig. 1 *D*) and the only H-bond near the PGVG strand is the N52-A393 interaction (weaker than *h*V51-*h*P10 in HisH) that connects the PGVG(N) strand to the *fα*4″ helix in the cyclase domain (see Fig. 1 *C*). Thus, the oxyanion strand is H-bonded to the HisH glutaminase in bacteria, while it remains at the interface between the two domains in *Sc*-IGPS. Therefore, the communication pathways along the two IGPS domains prior to the reaction at the glutaminase active site are presumably different in the two organisms. The flux of conformational changes associated with the allosteric mechanism of *Tm*-IGPS has been identified by computational studies and verified experimentally (23,25,28,30,38), but a comparative analysis of IGPS from different organisms was missing.

Here, we perform a comparative study of allosteric pathways in *Tm*- and *Sc*-IGPS adopting the same successful methodology used for the studies of bacterial IGPS. In particular, we used graph-theory-derived network models to analyze the correlations of nuclear fluctuations observed in molecular dynamics (MD) simulations of *Sc*-IGPS. This approach involves a set of computational tools that have previously been used to describe different aspects of the

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

protein dynamics in a variety of systems (39–42), including *Tm*-IGPS (23,38). Notably, in our earlier work on bacterial IGPS, the role of a HisF hydrophobic cluster in transmitting the effector binding signal has been confirmed in NMR titration experiments (23). Besides, mutation experiments coupled with kinetic essays have followed after our predicted allosteric pathways, targeting a few key residues where mutations induced the disruption of the allosteric effects (25). Moreover, our previous computational studies predicted that the allosteric pathways in *Tm*-IGPS involve an opening/closing (breathing) motion of the HisH domain relative to the HisF unit, supported by hinge-like interactions at the HisF:HisH interface (23,38). The crucial role of this interdomain collective motion was recently validated experimentally using an IGPS mutant involving a photo-responsive unnatural amino acid, which could lock the motion at the interface, resulting in modulation of the enzymatic activity (28). Finally, our previous MD simulations, which captured the early dynamics (100 ns) of bacterial IGPS, revealed how, for this time scale, the collective hinge motion is associated with local interresidue interactions that synergistically, and only in presence of the effector, initiate a conformational change in the HisH active site promoting the stabilization of an oxyanion hole. The hypothesis that the allosterically driven formation of such an oxyanion hole is essential for the IGPS catalytic activity, consistent with the active site conformational change seen in our MD simulations, was recently confirmed through experimental studies that finally characterized the pro-active configuration of *Tm*-IGPS (30). Altogether, the various experimental validations of our studies on *Tm*-IGPS allostery strongly support the robustness of our methodology in sampling the early allosteric dynamics of the IGPS enzyme and in characterizing the allosteric pathways (in terms of both local inter-residue interactions and collective protein motions), substantiating its application to the IGPS enzyme in another organism, such as the yeast *Sc*-IGPS.

In the present contribution, we thus compare the early allosteric dynamics and the well-established allosteric pathways of bacterial IGPS (23–26,28,30,38,43) with those of its yeast homolog, here obtained with the same methodology employed for *Tm*-IGPS, in conjunction with new complementary analysis of both *Tm*- and *Sc*-IGPS enzymes.

RESULTS AND DISCUSSION

Changes in correlations induced by PRFAR binding to IGPS from yeast and bacteria

Fig. 2 *A* shows the effect of PRFAR binding on the structure of correlations in IGPS from bacteria (left panel) and yeast (right panel), respectively. Specifically, Fig. 2 *A* shows maps of differences of generalized correlation coefficients, $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ (39) in PRFAR-bound and apo IGPS of *Tm*-IGPS (left panel) and *Sc*-IGPS (right panel), respectively. The generalized correlation coefficients $r_{MI}[\mathbf{x}_i,\mathbf{x}_j] = [1\text{-}\exp(-2/3\ I[\mathbf{x}_i,\mathbf{x}_j])]^{-1/2}$ provide a quantitative measure of correlations in the positions $\mathbf{x}_i$ and $\mathbf{x}_j$ of Cα atoms in residues *i* and *j*, based on the mutual information $H[\mathbf{x}_i,\mathbf{x}_j] = H[\mathbf{x}_j] + H[\mathbf{x}_i] - H[\mathbf{x}_i,\mathbf{x}_j]$. Here, $H[\mathbf{x}_i]$ and $H[\mathbf{x}_i,\mathbf{x}_j]$ are the marginal and joint (Shannon) entropies, respectively, for atomic vector displacements ($\mathbf{x}_i$ and $\mathbf{x}_j$) computed as ensemble averages over MD simulations of apo IGPS and PRFAR-bound states. The resulting correlation patterns reflect the early dynamics of *Sc*-IGPS (and *Tm*-IGPS), obtained by averaging the generalized correlation coefficients computed on six independent replicas of 100 ns (four replicas for *Tm*-IGPS), thus allowing for direct comparisons with earlier studies of *Tm*-IGPS (23,38) (further details provided in the supporting material). In addition, we performed a similar comparative analysis of correlations obtained instead using a gaussian network model (44) and based on the crystallographic structures of the IGPS enzyme from the two organisms (see Fig. S1). Notably, the resulting correlation matrices show evident differences, indicating that only part of the changes in correlations sampled with MD simulations are encoded in the structural differences between the two systems.

The distinct patterns of correlations, shown in Fig. 2 *A* for yeast and bacterial IGPS, suggest distinct allosteric motions triggered by PRFAR binding in the two enzymes. In particular, *Tm*-IGPS (Fig. 2 *A*, left) shows various domains within HisH and HisF where the residues are more correlated among themselves than with residues in other parts of the protein. This indicates a sort of internal division within HisH and HisF domains that clearly appears as blocks of reduced correlations (magenta features in Fig. 2 *A*, left panel) in one side of the *Tm*-IGPS (namely, sideL) and increased correlation (green features in Fig. 2 *A*, left panel) on the opposite side of the protein (namely, sideR). In *Tm*-IGPS, weaker correlations in the PRFAR-bound complex correspond to weaker interfacial HisH-HisF interactions upon effector binding. Reduced correlations affect the interdomain hinge-like breathing motion, as observed in MD simulations of apo and PRFAR-bound enzymes (23,38).

The hinge-like breathing motion plays a central role in the allosteric regulation of *Tm*-IGPS, as recently confirmed by experiments (28). The effector-induced internal division within HisH and HisF domains of *Tm*-IGPS is essentially absent in the *Sc*-IGPS cyclase and glutaminase domains of His7 (see Fig. 2 *A*, right panel), with a sizable increase of correlations observed only between cyclase residues 345–400 (belonging to *f*α3, *f*β4, and *f*α4) and the rest of the enzyme. Moreover, in contrast with *Tm*-IGPS, binding of PRFAR in *Sc*-IGPS induces milder effects on the correlations of motions in the whole enzyme. Therefore, it is clear that PRFAR binding to *Sc*-IGPS does not affect a hinge-like breathing motion as in *Tm*-IGPS, consistent with the

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

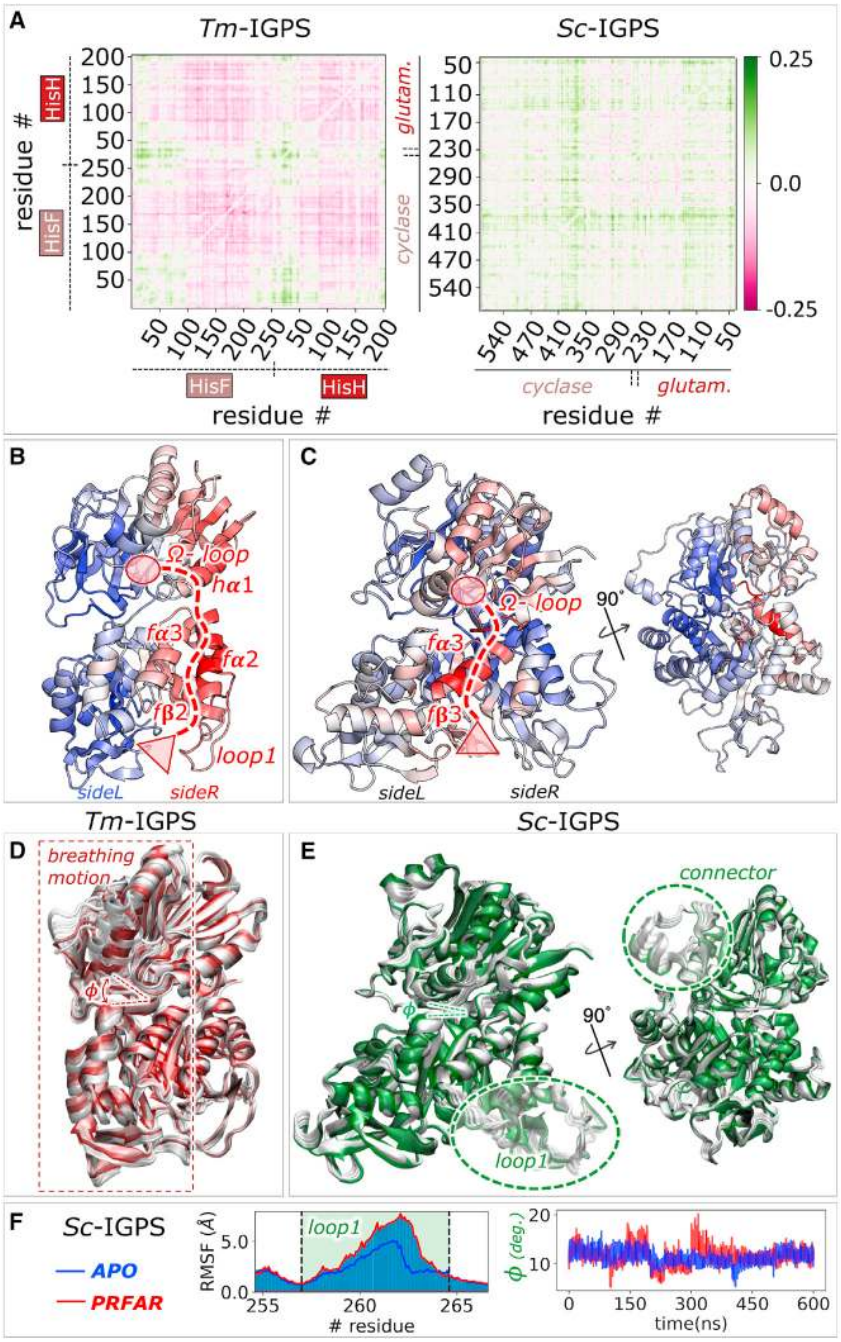


FIGURE 2 Analysis of correlated motions in IGPS from yeast and bacteria. (*A*) Comparison of generalized correlation coefficients $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ for PRFAR-*minus*-apo *Tm*-IGPS (left) and *Sc*-IGPS (right). In *Tm*-IGPS, PRFAR induces changes in both HisF and HisH, leading to innerly correlated domains (green features) as well as uncorrelated residues (magenta features), with amino acid residues 100–220 in HisF (sideL) featuring a decrease in correlations with the rest of the enzyme (black dotted lines). The PRFAR-*minus*-apo correlation matrix in *Sc*-IGPS does not exhibit similar features to those found in *Tm*-IGPS but rather milder changes of correlations due to effector binding, except for a sizable increase in correlations observed between cyclase residues 345 and 400 (belonging to *f*α3, *f*β4, and *f*α4) and the rest of *Sc*-IGPS (black dotted lines). The abbreviation glutam. refers to the glutaminase domain. (*B* and *C*) EC differences (PRFAR-*minus*-apo) projected onto the apo structure of *Tm*- (B) and *Sc*-IGPS (C), computed for local correlation values (damping-distance parameter $\lambda = 5$), featuring gains (in red) and loss (in blue) of centrality upon effector binding. The allosteric pathways from the effector site (red triangle) to the active site (red circle) in both enzymes are marked with red dotted lines. The main secondary structure elements along the pathways are labeled. (*D* and *E*) Differential (PRFAR-*minus*-apo) essential dynamics from the first PC of *Tm*- and *Sc*-IGPS MD trajectories. A rotated view of yeast IGPS is reported to visualize the motion of the connector. (*F*) RMSF of loop1 (left) and time evolution of the hinge breathing motion (right) in apo (blue lines) and PRFAR-bound (red lines) *Sc*-IGPS. The breathing motion is monitored by the G51(Cα)–W124(Cγ)–Y394(Cγ) angle (φ) over the concatenated (600 ns) MD simulations.

hypothesis of different allosteric mechanisms in the two organisms.

Long- and short-range allosteric communication in IGPS from yeast and bacteria

Our analysis of correlations in *Tm*-IGPS and *Sc*-IGPS shows distinct changes in correlated motions induced by PRFAR binding that result from changes in both long- and short-range interactions and enable allosteric activation of yeast and bacterial IGPS. Fig. 2 shows the principal component analysis (PCA) (40,45,46) and eigenvector centrality (EC) network analysis (38) of correlated motions. PCA selects the principal collective motions sampled from MD simulations by diagonalization of the covariance matrix of atomic displacements (see details in the supporting material), although it is limited to linear correlations. Thus, we employ the EC analysis to include non-linear correlations in an effort to disentangle long- and short-range contributions.

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

The EC methodology represents a cost-effective approach that yields fundamental understanding of allosteric mechanisms at the molecular level (38,47,48). Our implementation is based on a weighted graph with nodes corresponding to Cα atoms and weights between pairs of Cα atoms *i* and *j* determined by the corresponding generalized correlation coefficient $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$, as discussed above (see Fig. 2 *A* and *B*).

The centrality c_i of residue *i* is a real positive number defined by the *i-th* entry of the leading eigenvector of the weighted adjacency matrix $A_{ij} = (1-\delta_{ij})\ r_{MI}[\mathbf{x}_i,\mathbf{x}_j]\ \exp(-d_{ij}/\lambda))$. The damping parameter allows for the analysis of local correlations by simply dumping out the contributions from pairs of residues beyond a given range (see Fig. S5).

We initially focus on local centrality changes $\Delta c_i = c_i^{PRFAR} - c_i^{APO}$, induced by PRFAR, analyzed by defining A_{ij} with $\lambda = 5$ Å. Panels B and C in Fig. 2 show the normalized centrality differences Δc_i induced by PRFAR binding to *Tm*-IGPS and *Sc*-IGPS, respectively, with a color scale from minimal (blue) to maximal (red) values of Δc_i (details in the supporting material). The computed centrality differences reveal significant differences in the two organisms. For *Tm*-IGPS (38), only sideR transfers the allosteric signal through a pathway that involves multiple secondary structural elements: loop1, *f*β2, *f*α2, and *f*α3 in HisF and *h*α1, Ω-loop in HisH. The signal reaches the active site at the *h*C84 residue via alteration of H-bonding interactions with the highly conserved PGVG (oxyanion) strand, adjacent to the Ω-loop (23,38). In *Sc*-IGPS, however, the increased centralities induced by PRFAR binding are not localized on the sideR of the protein and involve a smaller number of secondary structure elements than in *Tm*-IGPS.

In fact, the PRFAR allosteric signal in His7 involves mainly *f*β3 and *f*α3 in the cyclase domain (where most of the increased values are found) with a direct link to the Ω-loop in the glutaminase domain that allows the signal to reach the active site (PGVG and C83) more directly than in *Tm*-IGPS.

Short-range correlations are affected by local contacts while long-range correlations involve collective modes that relate to slow protein motions. Here, we combine PCA and EC analysis to characterize the main collective modes and long-range correlations involved in the allosteric mechanisms. For *Tm*-IGPS, we have shown that the comparison of centrality differences obtained with $\lambda = \infty$ and $\lambda = 5$ allows for the characterization of long-range correlations in allosteric mechanisms that directly relate to the breathing motion of bacterial IGPS (38). Notably, we observed that the results agree with the essential motions induced by the effector as obtained by PCA (see Fig. S7). The essential motions are obtained by projecting the MD trajectories onto the main PRFAR-*minus*-apo difference principal components (PCs; ΔPC_1 and ΔPC_2 for first and second components, respectively). Fig. 2 *D* and *E* show the effector-induced essential motions described by the ΔPC_1 in both bacteria and yeast IGPS, indicating that there are significant differences in the two organisms. Indeed, the alteration of the breathing hinge motion in *Tm*-IGPS (see Fig. 2 *D*), upon effector binding, is replaced by a large motion of the loop1 (residues 250–275) and the connector site (residues 206–236) in *Sc*-IGPS. Analogously, the PRFAR-*minus*-apo difference for the second PC (ΔPC_2) reveals additional differences in the effector-induced essential dynamics of the two systems (see Figs. S8 and S9), with a mild movement of loop1 accompanying the *Tm*-IGPS hinge motion. In contrast, for *Sc*-IGPS, spring-like motion of the surface secondary structural elements of His7 was detected (see Videos S1–S3).

Overall, these results indicate that loop1 is involved in short-range interactions in *Tm*-IGPS allostery. However, in *Sc*-IGPS the loop1 is part of the long-range communication, becoming freer to fluctuate upon effector binding (see the root-mean-square fluctuations [RMSFs], reported in Fig. 2 *F*, left panel and in Fig. S8 compared with those in *Tm*-IGPS).

We note that loop1 is much shorter in *Tm*-IGPS than in Sc-IGPS so it might play different functional roles in the two systems. In fact, inspection of our MD trajectories suggests that loop1 of *Tm*-IGPS might play a 2-fold role in the *Tm*-IGPS by being involved in short-range allosteric communication and at the same time functioning as a gatekeeper to keep the effector in the binding pocket under high-temperature conditions. In *Sc*-IGPS, however, changes in the motion of loop1 induced upon effector binding are accompanied by the motion of the cyclase-glutaminase interdomain connector (see Fig. 2 *E*), alternatively to the breathing motion observed in *Tm*-IGPS (see Fig. 2 *F*, right panel), which is not present in *Sc*-IGPS (28). Moreover, the long and highly mobile loop1 of *Sc*-IGPS might facilitate PRFAR binding under room temperature conditions. In the absence of a prominent hinge-like motion as observed in Tm-IGPS, the role of the connector in *Sc*-IGPS is more related to the propagation of low-vibrational motions across the two domains. In this sense, while it was possible to successfully suggest point mutagenesis experiments targeting specific local contacts for loop1 in *Tm*-IGPS, the same is hard to do for loop1 and the connector site in *Sc*-IGPS, as their role is not associated with allosteric local contact changes but rather with the collective motions initiating allosteric communication.

Clearly, the combination of EC and PCA is a powerful methodology for identifying protein domains that are significantly affected upon binding of an allosteric effector and for characterization of essential motions, providing evidence of collective modes and inter-residue interactions that control the underlying allosteric mechanisms.

Besides, the residues showing the largest centrality values include those in fβ3 and fα3 in the cyclase domain and those in the Ω-loop and vicinity (Fig. 2 *C*, highlighted in red), which represent promising targets for site-directed mutagenesis studies since they exhibit the highest increase in centrality upon PRFAR binding. The impact of mutants on *Tm*-IGPS has been evaluated experimentally through mutagenesis studies coupled to kinetic experiments (25,32), confirming

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

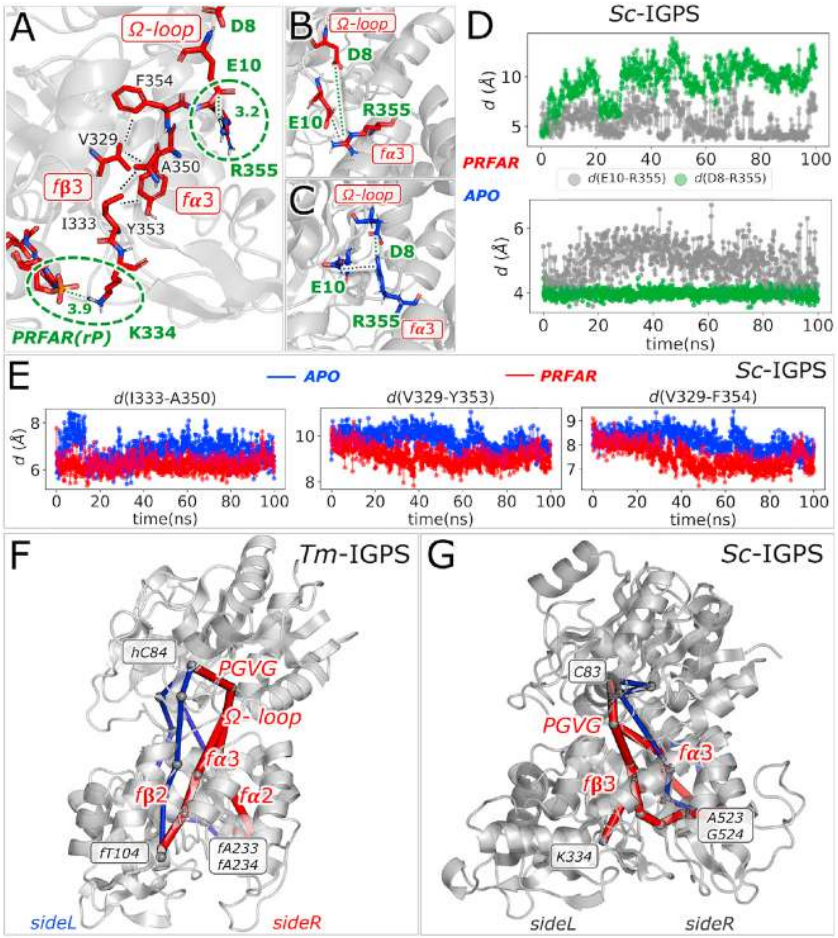


FIGURE 3 Allosteric communication between the effector and the glutaminase active site. (*A*) Local contacts spanning from the PRFAR binding site to the cyclase:glutaminase interface, involving at the extremes two salt bridges (green circles) between residue K334 in *f*β3 and the ribose-side phosphate of PRFAR (rP) and between R355 in *f*α3 and D8 and E10 in the glutaminase Ω-loop, bridged by a cluster of hydrophobic interactions between *f*β3 and *f*α3 residues (i.e., I333-A350-Y353-V329-F354). (*B* and *C*) Representative MD snapshots of the average R355-D8 and R355-E10 salt-bridge picture in the effector-bound (red sticks residues, *B*) and apo (blue sticks residues, *C*) complexes. (*D*) The effect of PRFAR binding on the time evolution of the R355-D8 and R355-E10 salt-bridge distances, along a representative 100 ns MD trajectory. (*E*) Tightening of the interactions in the *f*β3-*f*α3 hydrophobic cluster upon PRFAR binding, along a representative 100-ns MD trajectory. (*F* and *G*) Shortest communication pathways connecting the *f*T104, *f*A223, and *f*A224 residues and the K334, A523, and G524 residues in the PRFAR binding sites of *Tm*-IGPS (*F*) and *Sc*-IGPS (*G*), respectively, and the Gln substrate binding site, i.e., *h*C84 and C83, respectively.

that mutants that directly target the allosteric pathway have a strong impact on the allosteric communication. We anticipate that similar site-directed mutagenesis studies on *Sc*-IGPS targeting the residues along the highest centrality pathway could shed light on the adaptation of the allosteric pathways in these protein homologs. We emphasize that mutants that lie outside of predicted allosteric pathways have been found to be less disruptive of the allosteric function in other systems (49), suggesting that future mutagenesis studies targeting random mutations of both *Tm*- and *Sc*-IGPS would be very informative for further insights that foster therapeutic applications aimed at altering the functionality of IGPS enzymes by targeting residues that control the enzyme's dynamics.

While we performed 12 independent 100-ns runs, one for each model of the apo and holo systems *Sc*-IGPS, the results discussed above are obtained by averaging the calculated properties over all model replicas (see materials and methods section). Hence, the average picture discussed above (involving differences between apo and holo dynamics) is representative of the allosteric process, although the individual simulations would present different EC (and PCA) profiles (as reported in Fig. S5).

Remarkably, the average correlation and EC profiles over the different replicas resemble one of them (labeled as sim_1 in Figs. S2, S4, and S5), which seems to capture more clearly the allosteric effect (see additional comments in the supporting material documentation), so it has been selected as the most representative model replica in the following analysis.

In the next section, we analyze the allosteric pathways by inspecting those residues that are involved in short-range interactions responsible for information transfer across the catalytic units of IGPS. We do so by focusing on 100-ns snapshots that encompass most of the allosteric traits as identified by EC and PCA. The analysis provides understanding at the molecular level of the differences of the allosteric mechanism in the two organisms.

Allosteric pathways in IGPS from yeast and bacteria

Fig. 3 shows the analysis of allosteric pathways in *Tm*-IGPS and *Sc*-IGPS as determined by the influence of PRFAR on the correlations of thermal nuclear fluctuations. We find that

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

optimal communication pathways from the effector to the active sites are distinct in the two systems since PRFAR affects specific interactions in the two systems. In *Sc*-IGPS, the phosphate group at the ribose side (rP) of the effector forms a tight salt bridge with K334 in the *f*β3 sheet (see Fig. 3 *A*) that is favored over the D335-K334 H-bond present in the apo state (see Fig. S10). In *Tm*-IGPS, such ionic interaction with PRFAR is absent (23) as there is no residue capable of establishing a salt bridge with the effector in the bacterial enzyme. In *Sc*-IGPS, K334 is adjacent to I333, which belongs to a network of hydrophobic contacts (I333-A350-Y353-V329-F354) spanning over the whole *f*β3-*f*α3 region (see Fig. 3 *A*). Notably, these hydrophobic interactions are significantly strengthened upon PRFAR binding (see Fig. 3 *E*), and thus a hydrophobic cluster is most responsible for transmitting the effector signal through the cyclase domain (i.e., HisF in *Tm*-IGPS), similarly to the process in bacterial IGPS (23). However, the activation of the hydrophobic cluster in *Tm*-IGPS (comprising the *f*V48-*f*L50-*f*I52-*f*F23 residues) involves the *f*β2 sheet (not the *f*β3-*f*α3 region as in *Sc*-IGPS). More importantly, the activation mechanism involves the loop1, which is engaged in short-range allosteric interactions in *Tm*-IGPS. Furthermore, we note that changes in hydrophobic contacts due to PRFAR binding are primarily driven by interactions with the π-system of the imidazolecarboxamide group of PRFAR (Fig. S13). In *Sc*-IGPS, however, the allosteric signal is initiated upon formation of the K334-PRFAR(rP) salt bridge.

Changes in the hydrophobic contacts in *Tm*-IGPS induced by PRFAR binding affect a network of salt bridges on the surface of the IGPS sideR, involving ionic interactions between the charged residues *f*R59, *f*E67, *f*E71, *f*E91, and *f*R95 in the *f*α2 and *f*α3 helices (at HisF) and the *h*R18 residue in *h*α1 (at HisH) (23). In *Sc*-IGPS, however, there are no corresponding charged surface residues that can create a salt-bridge network and, thus, the signal travels from PRFAR through the *f*β3-*f*α3 hydrophobic cluster until it reaches the charged residue R355 (at the end of *f*α3), which interfaces the glutaminase domain (Fig. 3 *A*). As shown in Fig. 3 *B* and *C*, indeed, the R355 charged sidechain could engage in interface ionic interactions with either D8 or E10 sidechains, belonging to the Ω-loop of the glutaminase subunit. Notably, as shown in Fig. 3 *D*, the R355-D8 salt bridge is stably formed throughout the MD trajectories of apo *Sc*-IGPS. However, PRFAR binding induces a change in the R355 partner, favoring formation of the R355-E10 salt bridge, which is weaker than the apo R355-D8 bond. These results indicate that the effector alters the *f*α3/Ω-loop ionic interactions at the cyclase:glutaminase interface in *Sc*-IGPS, while in *Tm*-IGPS the affected salt bridges at the HisF/HisH interface involve the *f*α2/*h*α1 helices of sideR. We suggest that future mutagenesis studies of *Sc*-IGPS can target the important residues highlighted in our analysis; i.e. those along the allosteric pathway (I333-A350-Y353-V329-F354, R355, E10 and D8).

The comparison of *Tm*- and *Sc*-IGPS active sites in the crystallographic structures highlights how the effector-induced *h*V51-*h*P10 H-bond breaking (23) (a crucial allosteric step observed for the bacterial enzyme; see Fig. S15) is not plausible *Sc*-IGPS where H-bonding interactions near the PGVG oxyanion strand, stable throughout the dynamics, are limited to the A393-N52 H-bond at the interface (see Fig. 1 *C*). The interface H-bond in apo *Sc*-IGPS is weaker than the (buried) *h*V51-*h*P10 bond in apo *Tm*-IGPS and, despite weakening of the A393-N52 interaction upon effector binding (see Fig. S16), dynamical fluctuations are more related to the (quite narrow) breathing motion in His7 (see Fig. 2 *F*) than to allosteric signal propagation through local contacts. Therefore, the observation that PRFAR binding in the yeast affects the *f*α3/Ω-loop ionic interactions is not sufficient to explain how the effector signal is transferred from the interdomain interface to the active site of *Sc*-IGPS (i.e., there is no direct, allosterically modulated connection between the PGVG oxyanion strand and the Ω-loop in *Tm*-IGPS).

We analyze the communication pathways that link the effector site in the cyclase domain with the glutaminase active site and the activation mechanism toward the catalytically active state in both yeast and bacteria. The enzymatic communication pathways are computed as the optimal paths (i.e., paths with stronger correlation) connecting specific pairs of physically distant residues. Amino acid residues correspond to the nodes of a graph with edges defined by the strength of correlations between pairs of residues (11) (i.e., higher correlated pairs correspond to shorter bonds and are more likely to belong to the optimal communication path).

The communication pathways start at the PRFAR binding site with residues *f*T104, *f*A223, and *f*A224 of *Tm*-IGPS, and K334, A523, and G524 of *Sc*-IGPS. The target final node is the Gln substrate binding site (i.e., *h*C84 and C83 in *Tm*-IGPS and *Sc*-IGPS, respectively). As shown in Fig. 3 *F* and *G*, the resulting communication channels are affected by the effector binding (apo pathways in blue and PRFAR-bound in red), featuring significant differences between the two organisms. In accordance with our EC analysis (Fig. 2 *B* and *C*), the signal from the effector to the active site is triggered by PRFAR binding and is preferentially transferred through sideR in *Tm*-IGPS, involving PGVG and the Ω-loop. In *Sc*-IGPS, however, the allosteric pathway is more internal, allowing direct communication between the PGVG oxyanion strand and the cyclase domain, enhancing a spring-like PC motion of protein expansion and contraction.

The final step of allosteric activation at the IGPS active site is the rearrangement of the PGVG strand associated with the flipping of the amide N-H group of residue hV50/V51 toward the Gln-binding site (in *Tm/Sc*, respectively), which allows formation of an oxyanion hole (Fig. 4). In *Tm*-IGPS, we demonstrated that the initiation of the PGVG flipping requires breaking of the hV51-hP10

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

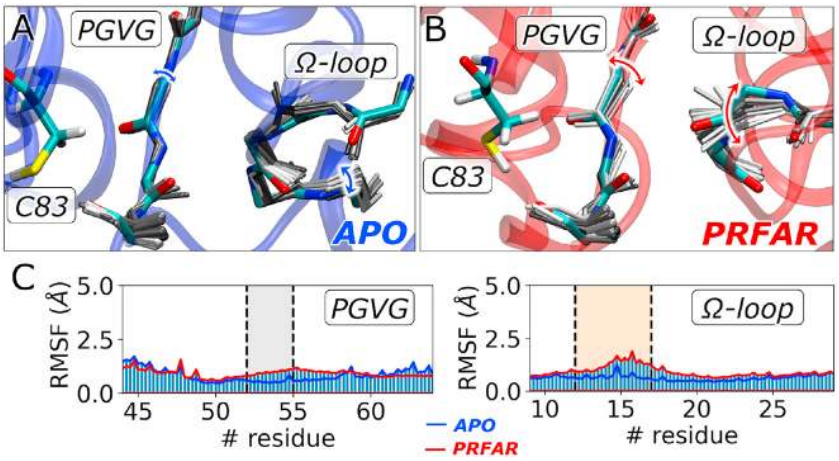


FIGURE 4 (*A* and *B*) Enhanced thermal fluctuations of the PGVG oxyanion strand and Ω-loop triggered by PRFAR binding in the glutaminase active site of *Sc*-IGPS. Average secondary structure in apo (blue), PRFAR-bound (red), and Gln-binding site (C83, colored sticks) are also depicted. (*C*) The RMSF profile of the PGVG oxyanion strand and Ω-loop in a representative (100 ns) trajectory in the apo (blue lines) and PRFAR-bound (red lines) complexes.

H-bond to separate the strand from the nearby Ω-loop (supporting material, Fig. S15). Notably, the breaking of this H-bond interaction has been resolved in the X-ray structure of the *Tm*-IGPS pro-active conformation (28), along with rearrangement of the PGVG strand and formation of the oxyanion hole. However, it remains to be established how the final allosteric step is initiated in *Sc*-IGPS, where PGVG and the Ω-loop are not linked by an H-bond.

Fig. 4 shows the early dynamics of PGVG and Ω-loop in *Sc*-IGPS and the differences observed (within 100 ns of a representative MD trajectory) between the apo and the PRFAR-bound complexes. The secondary structure elements PGVG and Ω-loop are not directly connected (e.g., by H-bonding) and are found to be more separated in *Sc*-IGPS than in *Tm*-IGPS (see Fig. 1 *C*). Nevertheless, both structural elements exhibit enhanced motion upon effector binding (Fig. 4 *A–D*), showing that changes in ionic interactions at the cyclase:glutaminase interface (e.g., R355-D8/E10 salt-bridge exchange; Fig. 3 *B–D*) correlates directly with motions in both secondary structural elements as the effector binds and promotes the interdomain signal transduction toward the active site.

In *Tm*-IGPS, changes in ionic contacts promote the HisF-HisH breathing motion that breaks the PGVG/Ω-loop H-bond and facilitates the PGVG flipping. In contrast, allostery in *Sc*-IGPS involves directly the Ω-loop, a structural element that affects the interface and enables the PGVG rearrangement in the absence of a hinge-like breathing motion. The limited interdomain motion in *Sc*-IGPS suggests that effector binding does not affect water accessibility to the glutaminase active site. Nevertheless, it is important to note that the reduced interdomain motion in *Sc*-IGPS is accompanied by enhanced collective motions of both the loop1 and the interdomain covalent connector, not present in *Tm*-IGPS.

Conclusions

We have characterized the early dynamics that involve the allosteric pathways of the IGPS enzyme in yeast and thermophilic bacteria by combining MD simulations and graph network analysis of correlated motions influenced by effector binding. We have found rather distinct allosteric pathways in the two enzymes, with specific inter-residues interactions and collective protein motions associated with conformational changes that initiate the communication between the allosteric and catalytic sites.

We speculate that the structural differences between yeast and bacterial IGPS are tailored to allow the proteins to function in their respective natural environments, leading to different allosteric mechanisms communicating distant sites in the IGPS enzymes of the two organisms. The heterodimer *Tm*-IGPS adapts the allosteric pathways to exploit a larger flexibility at high temperatures by allowing ample hinge-like motions of the two protein subunits. In contrast, the single-chain enzyme *Sc*-IGPS, which functions at room temperature, establishes more internal allosteric pathways in terms of inter-residues interactions, allowing for more direct communication between the PGVG oxyanion strand and the cyclase domain, enhanced by an overall spring-like motion of protein expansion and contraction, driven by flexible portions of the protein (loop1 and connector site). These predictions pave the way for future experimental validation (by mutagenesis, NMR, and kinetic essays) of the proposed differences between the allostery in the two organisms.

Our study contributes to understanding how proteins absolving for the same function, but from different evolutionary pathways, preserve their functionality in different environments by adapting their signaling pathway.

MATERIALS AND METHODS

Correlation matrices for *Tm*-IGPS are obtained from the same trajectories and following the same protocol as in reference (23), while yeast models are built *ex novo*.

The computational structural models for apo and PRFAR-bound yeast IGPS complexes are based on the crystal structure of the bienzyme complex from *Sc*-IGPS at 2.4 Å resolution (PDB: 1OX6-B) (31). The HisH-HisF apo-complex having several missing residues (261–275, 301–304, and

551–552) and three extra residues at the beginning of the chain required modeling prior to simulation. To complete the structure, first, we stripped the first three residues, then we aligned and added residues 256–260 and 299–310 from 1OX4-B (removing overlapping residues from 1OX6 due to poor alignment). Finally, we added residues 550–552 from 1JVN-A (removing residue 550 from 1OX6-B). We constructed the remaining residues (256–275) using different tools available online, using which we produced six different structural models. One model was generated using Modeller (50), a second one using Swiss-Model (51), and four suitable homology models were found on modbase. PRFAR was bound to each model by aligning each structure to the effector-bound crystal structure of yeast IGPS (34) (PDB: 1OX5).

The 12 generated structures (six in the apo state, six bound to the effector) align with RMSD <5 Å. To allow for a direct comparison between the dynamics of IGPS enzymes from *Tm*- and *Sc*-IGPS, we kept the simulation conditions analogous to the one used for bacterial IGPS in reference (23). Our choice of keeping the simulation conditions identical was motivated by recent studies demonstrating that PRFAR is a weaker allosteric activation at growth temperature than it is at room temperature (52). For the sake of clarity, we report some essential details below. MD simulations of the apo and PRFAR-bound structures of yeast IGPS are based on the AMBER-ff99SB (53) force field for the protein and generalized amber force field (54) for the PRFAR ligand (see supporting material), as implemented in the Amber20 software package (55). We performed 12 independent MD simulations, one for each complex (apo and PRFAR bound) for a total simulation time of 1.2 μs. Further details of the pre-equilibration procedure and MD production runs are described in the supporting material. Details on the computation of generalized correlation coefficients and covariances between pairs of residues and their analysis through the EC metrics and PCA as well as the description of how to compute allosteric pathways across yeast and bacterial IGPS are provided in the supporting material. Protein representations are obtained using the Pymol (56) software, with the exception of time-evolution representations, which are produced using VMD (56,57).

Determination of the allosteric pathways

The allosteric pathway for information transfer has been investigated by employing mutual information-based correlation analysis and network models from graph theory (39,40). Generalized correlations $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ capture noncollinear correlations between pairs of residues i and j, and are helpful in pointing out the residues that are most affected by the binding of an effector, and with it the information channels that govern the allosteric control. $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ alone can be hard to decipher and require some post-processing to interpret protein behavior. Network analysis tools (11,58), including different centrality metrics (59) , can be applied for the interpretation of correlated protein motions and their allosteric behavior. Here, the Cα-atoms of the proteins' amino acid residues constitute the nodes of a dynamical network graph, connected by edges (residue pair connection in terms of $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$). An adjacency matrix is then constructed such that it can be used to identify the key amino acid residues of IGPS with high susceptibility to effector binding. A simple, yet effective metric that extracts central nodes in the adjacency matrix is the EC (38). The basic idea behind this measure is the assumption that the centrality index of a node is not only determined by its position in the network but also by the neighboring nodes, hence it measures how well connected a node is to other well-connected nodes in the network. The protein network can be used to determine the optimal pathways for the information transfer between two nodes, defined as the shortest paths connecting a specific pair of nodes. In this context, edge lengths (i.e., the internode distances in the graph) are defined using the coefficients according to $-\log(r_{MI}[\mathbf{x}_i,\mathbf{x}_j])$, implying that highly correlated pairs (featuring good communication) are close in distance in the graph. In particular, we applied the Dijkstra algorithm to calculate the shortest pathways between residues *f*A233-*f*A234-A523/G524-R528 and hC84-C83, where each set of residues belongs to a different domain of bacterial and yeast IGPS, respectively. Hence, the computed pathways are composed of residue-to-residue steps that optimize the overall correlation (i.e., the momentum transport) between residues *f*A223-*f*A224 (at the effector site) and *h*C84 (in the glutaminase active site) in *Tm*-IGPS, and similarly residues K334, A523, G524, and C83 in His7. Additional details on the methods are included in the supporting material. As mentioned above, all analyses are performed on six different models (sim_0, sim_1, … sim_5) for yeast and four for bacterial IGPS, retrieved from reference (23), for which we examine both the apo and PRFAR-bound dynamics. Generalized correlation coefficients and covariances of atomic displacements are computed independently on each apo and PRFAR-bound 100-ns simulations. We compute the average PRFAR-bound-minus-apo correlation and covariance over each different model (four for bacteria and six for yeast). Remarkably, the average pictures depicted in Fig. 2, obtained as the average apo-minus-holo correlation (or covariance) computed across the different models, are representative of the allosteric process, although the individual simulations present different correlations matrices, EC, and PCA profiles (as shown in Figs. S2–S5). Among the six apo and PRFAR-bound replicas, the dynamics of sim_1 clearly resembles that of the average pictures, as illustrated in Fig. S5. Therefore, the characterization of shortest pathways and specific effector-induced contact changes has been reported in Figs. 3 and 4 using data from the representative model (i.e., sim_1).

PCs of protein dynamics

PCA (40) is a recognized approach to capture the essential motions of the simulated systems. In PCA, the covariance matrix of the protein Cα atoms is calculated and diagonalized to obtain a new set of coordinates (eigenvectors) to describe the system motions. Each eigenvector—or PC —is associated with an eigenvalue, which denotes how much each eigenvector is representative of the system dynamics.

To avoid translational artifacts, we set the center of mass of each frame at the origin and rotate each frame to its optimally aligned orientation relative to the average structure—computed over all apo trajectories—which also has its center of mass at the origin. Next, we evaluate the covariances of the positional fluctuations of each system over the apo and PRFAR-bound trajectories obtained by concatenation of the independent apo and effector-bound replicas. Because the motion of sidechains is mostly independent of the essential dynamics of IGPS, we restrict the covariance to the backbone atoms only. Projecting the original (centered) data onto the eigenvectors results in the PCs, whose associated eigenvalue (variance) is indicative of the portion of motion that the eigenvector describes. Together, the first two PCs relative to *Tm*-IGPS incorporate 44% and 33% of the total motion of the bacterial apo and PRFAR-bound trajectories, respectively (Fig. S3 *A*), while the percentages become 42% and 44% for His7 (Fig. S3 *B*). The contribution added by the third PC is much smaller, hence we limited our analysis to the first two.

By projecting the trajectory coordinates onto the PCs, one can visualize the essential motions induced by effector binding in yeast and bacterial IGPS on the protein structure, along the trajectory. The corresponding motions are shown in Figs. 2 *D*, *E*, S8 *A*, and *B*.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2021.11.2888.

AUTHOR CONTRIBUTIONS

F.M., A.G., and A.P. performed the research. I.R. and V.S.B. designed the research. F.M., V.S.B., and I.R. wrote the paper.

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

ACKNOWLEDGMENTS

I.R. and A.G. gratefully acknowledge the use of HPC resources of the Pôle Scientifique de Modélisation Numérique (PSMN) of the ENS-Lyon, France.
This work was supported by the NIH grant GM106121 (V.S.B.) and a generous allocation of high-performance computing time from NERSC.

REFERENCES

1. Wodak, S. J., E. Paci, …, T. McLeish. 2019. Allostery in its many disguises: from theory to applications. *Structure.* 27:566–578.
2. Greener, J. G., and M. J. Sternberg. 2018. Structure-based prediction of protein allostery. *Curr. Opin. Struct. Biol.* 50:1–8.
3. Loutchko, D., and H. Flechsig. 2020. Allosteric communication in molecular machines via information exchange: what can be learned from dynamical modeling. *Biophys. Rev.* 12:443–452.
4. East, K. W., E. Skeens, …, G. P. Lisi. 2020. NMR and computational methods for molecular resolution of allosteric pathways in enzyme complexes. *Biophys. Rev.* 12:155–174.
5. Birdsall, N. J., T. Farries, …, M. Sugimoto. 1999. Subtype-selective positive cooperative interactions between brucine analogs and acetylcholine at muscarinic receptors: functional studies. *Mol. Pharmacol.* 55:778–786.
6. Klein, J., and K. Löffelholz. 1996. Cholinergic Mechanisms: From Molecular Biology to Clinical Significance. Elsevier.
7. Christopoulos, A. 2002. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat. Rev. Drug Discov.* 1:198–210.
8. Guo, J., and H.-X. Zhou. 2016. Protein allostery and conformational dynamics. *Chem. Rev.* 116:6503–6515.
9. Bozovic, O., J. Ruf, …, P. Hamm. 2021. The speed of allosteric signaling within a single-domain protein. *J. Phys. Chem. Lett.* 12:4262–4267.
10. Stock, G., and P. Hamm. 2018. A non-equilibrium approach to allosteric communication. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 373:20170187.
11. Rivalta, I., and V. S. Batista. 2021. Community network analysis of allosteric proteins. *Methods Mol. Biol.* 2253:137–151.
12. Suplatov, D., and V. Švedas. 2015. Study of functional and allosteric sites in protein superfamilies. *Acta Naturae.* 7:34–45.
13. Hwang, P. K., and R. J. Fletterick. 1986. Convergent and divergent evolution of regulatory sites in eukaryotic phosphorylases. *Nature.* 324:80–84.
14. Micheletti, C. 2013. Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. *Phys. Life Rev.* 10:1–26.
15. Shulman, A. I., C. Larson, …, R. Ranganathan. 2004. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell.* 116:417–429.
16. Süel, G. M., S. W. Lockless, …, R. Ranganathan. 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10:59–69.
17. Chen, E., K. Reiss, …, G. P. Lisi. 2021. A structurally preserved allosteric site in the MIF superfamily affects enzymatic activity and CD74 activation in D-dopachrome tautomerase. *J. Biol. Chem.* 297:101061.
18. Mottonen, J. M., D. J. Jacobs, and D. R. Livesay. 2010. Allosteric response is both conserved and variable across three CheY orthologs. *Biophys. J.* 99:2245–2254.
19. Sethi, A., J. Tian, …, S. Gnanakaran. 2013. A mechanistic understanding of allosteric immune escape pathways in the HIV-1 envelope glycoprotein. *PLoS Comput. Biol.* 9:e1003046.
20. Gruber, R., and A. Horovitz. 2016. Allosteric mechanisms in chaperonin machines. *Chem. Rev.* 116:6588–6606.
21. Royer, W. E., Jr., H. Zhu, …, J. E. Knapp. 2005. Allosteric hemoglobin assembly: diversity and similarity. *J. Biol. Chem.* 280:27477–27480.
22. Livesay, D. R., K. E. Kreth, and A. A. Fodor. 2012. A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms. *Methods Mol. Biol.* 796:385–398.
23. Rivalta, I., M. M. Sultan, …, V. S. Batista. 2012. Allosteric pathways in imidazole glycerol phosphate synthase. *Proc. Natl. Acad. Sci. U S A.* 109:E1428–E1436.
24. Rivalta, I., G. P. Lisi, …, V. S. Batista. 2016. Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein–protein interface. *Biochemistry.* 55:6484–6494.
25. Lisi, G. P., K. W. East, …, J. P. Loria. 2017. Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity. *Proc. Natl. Acad. Sci. U S A.* 114:E3414–E3423.
26. Botello-Smith, W. M., and Y. Luo. 2019. Robust determination of protein allosteric signaling pathways. *J. Chem. Theor. Comput.* 15:2116–2126.
27. Lake, P. T., R. B. Davidson, …, M. McCullagh. 2020. Residue-level allostery propagates through the effective coarse-grained hessian. *J. Chem. Theor. Comput.* 16:3385–3395.
28. Kneuttinger, A. C., C. Rajendran, …, R. Sterner. 2020. Significance of the protein interface configuration for allostery in imidazole glycerol phosphate synthase. *Biochemistry.* 59:2729–2742.
29. Chittur, S. V., T. J. Klem, …, V. J. Davisson. 2001. Mechanism for acivicin inactivation of triad glutamine amidotransferases. *Biochemistry.* 40:876–887.
30. Wurm, J. P., S. Sung, …, R. Sprangers. 2021. Molecular basis for the allosteric activation mechanism of the heterodimeric imidazole glycerol phosphate synthase complex. *Nat. Commun.* 12:2748.
31. Chaudhuri, B. N., S. C. Lange, …, J. L. Smith. 2003. Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme. *Biochemistry.* 42:7003–7012.
32. List, F., M. C. Vega, …, M. Wilmanns. 2012. Catalysis uncoupling in a glutamine amidotransferase bienzyme by unblocking the glutaminase active site. *Chem. Biol.* 19:1589–1599.
33. Douangamath, A., M. Walker, …, M. Wilmanns. 2002. Structural evidence for ammonia tunneling across the (βα)8 barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure.* 10:185–193.
34. Chaudhuri, B. N., S. C. Lange, …, J. L. Smith. 2001. Crystal structure of imidazole glycerol phosphate synthase: a tunnel through a (β/α)8 barrel joins two active sites. *Structure.* 9:987–997.
35. Kneuttinger, A. C., K. Straub, …, R. Sterner. 2019. Light regulation of enzyme allostery through photo-responsive unnatural amino acids. *Cell Chem. Biol.* 26:1501–1514.e9.
36. Amaro, R. E., A. Sethi, …, Z. A. Luthey-Schulten. 2007. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry.* 46:2156–2173.
37. Amaro, R. E., R. S. Myers, …, Z. A. Luthey-Schulten. 2005. Structural elements in IGP synthase exclude water to optimize ammonia transfer. *Biophys. J.* 89:475–487.
38. Negre, C. F. A., U. N. Morzan, …, V. S. Batista. 2018. Eigenvector centrality for characterization of protein allosteric pathways. *Proc. Natl. Acad. Sci. U S A.* 115:E12201–E12208.
39. Lange, O. F., and H. Grubmüller. 2006. Generalized correlation for biomolecular dynamics. *Proteins.* 62:1053–1061.
40. Lange, O. F., and H. Grubmüller. 2008. Full correlation analysis of conformational protein dynamics. *Proteins.* 70:1294–1312.
41. Palermo, G. 2019. Structure and dynamics of the CRISPR-cas9 catalytic complex. *J. Chem. Inf. Model.* 59:2394–2406.
42. Melo, M. C. R., R. C. Bernardi, …, Z. Luthey-Schulten. 2020. Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *bioRxiv* https://doi.org/10.1101/2020.06.18.160572.

Please cite this article in press as: Maschietto et al., Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2021.11.2888

43. Gheeraert, A., L. Pacini, …, I. Rivalta. 2019. Exploring allosteric pathways of a V-type enzyme with dynamical perturbation networks. *J. Phys. Chem. B.* 123:3452–3461.

44. Atilgan, A. R., S. R. Durell, …, I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–515.

45. David, C. C., and D. J. Jacobs. 2014. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol. Biol.* 1084:193–226.

46. Amadei, A., A. B. Linssen, and H. J. Berendsen. 1993. Essential dynamics of proteins. *Proteins.* 17:412–425.

47. Jalili, M., A. Salehzadeh-Yazdi, …, K. Alimoghaddam. 2016. Evolution of centrality measurements for the detection of essential proteins in biological networks. *Front. Physiol.* 7:375.

48. Ashtiani, M., A. Salehzadeh-Yazdi, …, M. Jafari. 2018. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Syst. Biol.* 12:80.

49. Wang, J., A. Jain, …, N. V. Dokholyan. 2020. Mapping allosteric communications within individual proteins. *Nat. Commun.* 11:3862.

50. Webb, B., and A. Sali. 2016. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* 86:2.9.1–2.9.37.

51. Waterhouse, A., M. Bertoni, …, T. Schwede. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46:W296–W303.

52. Lisi, G. P., A. A. Currier, and J. P. Loria. 2018. Glutamine hydrolysis by imidazole glycerol phosphate synthase displays temperature dependent allosteric activation. *Front. Mol. Biosci.* 5:4.

53. Wang, J., P. Cieplak, and P. A. Kollman. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* 21:1049–1074.

54. Wang, J., R. M. Wolf, …, D. A. Case. 2004. Development and testing of a general amber force field. *J. Comput. Chem.* 25:1157–1174.

55. Case, D. A., H. M. Aktulga, …, P. A. Kollman. 2021. Amber 2020. University of California.

56. Schrödinger, L., and W. DeLano. 2020. PyMOL, Available at: http://www.pymol.org/pymol.

57. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38.

58. Yang, Z., R. Algesheimer, and C. J. Tessone. 2016. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* 6:30750.

59. Oldham, S., B. Fulcher, …, A. Fornito. 2019. Consistency and differences between centrality measures across distinct classes of networks. *PLoS One.* 14:e0220061.

3.3 Temperature increase mimics allosteric signaling in imidazole-glycerol phosphate synthase

3.3.1 Previous experimental findings

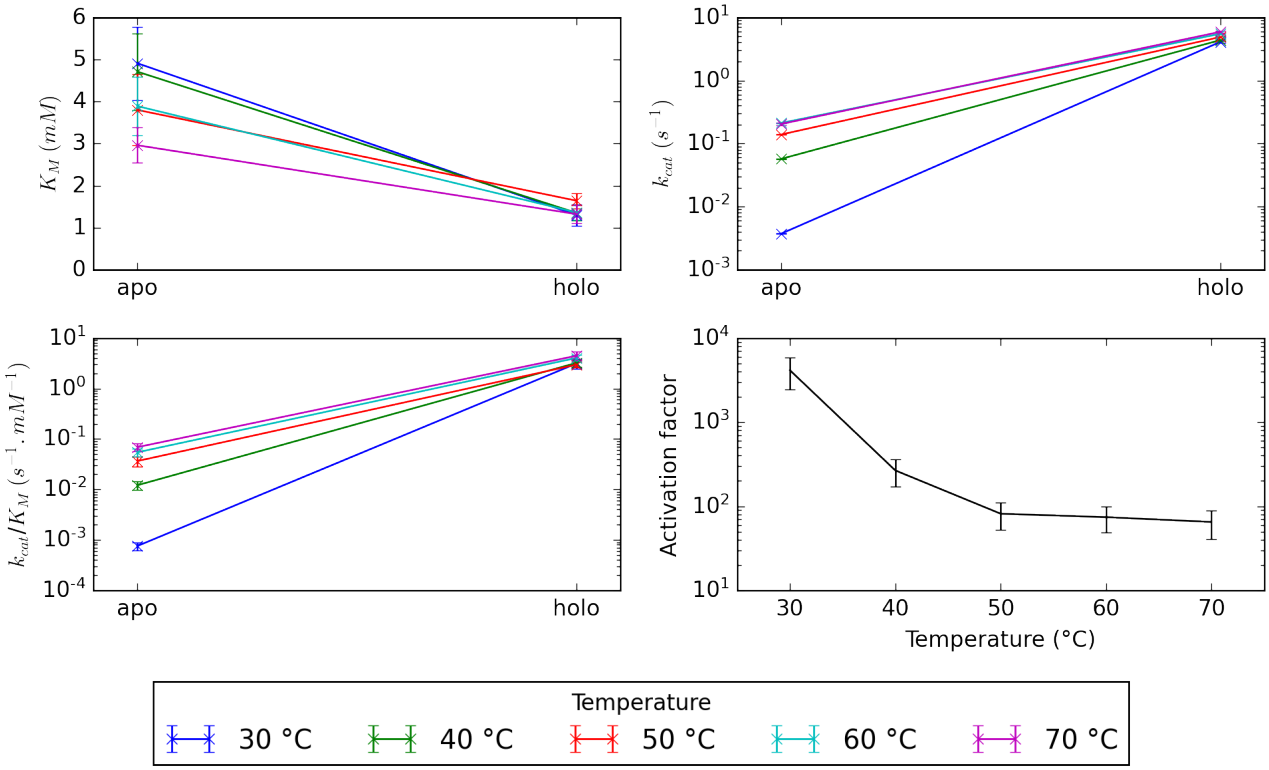


Figure 3.6: Evolution of the K_M dissociation constant (A), k_{cat} catalytic activity (B), k_{cat}/K_M (C) between apo and holo at different temperatures. (D) Evolution of the activation factor between apo and holo at the different temperatures. Data taken from ref [1]. A log-scale on the y-axis is used for catalytic activities, efficiencies and for the activation factor.

For many years, experimental data concerning IGPS form *T. maritima* was obtained at room temperature, but this bacteria is a thermophile growing only at temperatures between 55 °C and 90 °C with an optimum growth temperature of 80 °C (the highest in any bacteria)[2]. Still, the bacteria can survive for at least a year at −20 °C. Previous studies performed on IGPS from *T. maritima* are thus representative of a hibernating bacteria. To understand if the allosteric mechanism was temperature-dependent, the team of Patrick Loria and George Lisi reported Michaelis-Menten kinetics parameters of the apo-IGPS and holo-IGPS at temperatures ranging from 30 °C to 70 °C[1] (see Fig. 3.6A,B,C). The dissociation constant of glutamine is only slightly altered by temperature increase or effector binding and always remains in the same order of magnitude (mM).

The catalytic activity in apo increase by 1 order of magnitude between 30 °C and 40 °C and another order of magnitude between 40 °C and 50 °C and then only slightly increases between 50 °C and 70 °C. In holo, the catalytic activity slightly increases with temperature, but similarly always remains in the same order of magnitude. The evolution of the catalytic efficiency is mainly dominated to the evolution of the catalytic activity with IGPS being a V-type allosteric enzyme. Subsequently, the activation factor between the apo and holo enzymes vastly diminishes between 30 °C and 50 °C (from 4,161 to 82) and then remains mostly stable between 50 °C and 70 °C (from 82 to 65) (see Fig 3.6D). These experimental results prove that PRFAR is a weaker allosteric activator at high temperatures and suggests that temperature increase produces a similar effect as effector binding. Still, whether the same allosteric pathways are used in the temperature-dependent mechanism and the effector-dependent remains open.

3.3.2 Molecular Dynamics simulations analysis

To answer whether the temperature-dependent mechanism and the effector-dependent mechanism used the same allosteric pathways, we produced and analyzed Molecular Dynamics simulations of the apo30, holo30 and apo50 systems.

First, the eigenvector centrality analysis shows that upon temperature increase, residue displacements at sideR are becoming more correlated while at sideL they become less correlated. The same effect was identified

in previous studies and is here recovered for effector-binding. Moreover, the secondary structures involved in correlations increase at sideR due to temperature-increase and effector-binding are the same: loop1, $h\alpha 1$ and $h\alpha 4$. Finally, both effects are found to turn the optimal signaling pathways between the effector and active site from internal into external pathways.

Besides, similar effects are found again in terms of secondary structures changes in HisF in loop1 and $f\alpha 2$. The only notable exception is the $f\beta 6$-$f\alpha 6$ turn which folds into a helix upon PRFAR-binding but not upon temperature activation. This turn is located at the effector site and point at a small difference in overall activation mechanism that can be attributed to local effects of PRFAR-binding.

The DPCN analysis also show remarkable similarities between temperature and effector-binding effects. This analysis shows that not only the same secondary structures elements experience contact changes, but the exact same amino acids are involved. Consistently with secondary structure analysis, the main difference between temperature and effector-binding effect is found for the $f\beta 6$-$f\alpha 6$ turn whereupon PRFAR binding contacts increase substantially.

3.3.3 New experimental results and challenges

Our experimental collaborators at Yale University measured experimentally temperature coefficients of amide proton chemical shifts. They enabled to identify key amino acids that are impacted by temperature increase that changes amide proton environments. Among the residues mostly impacted by temperature changes, we identified fL63, which is involved in a backbone hydrogen bond with fR59. Consistent with perturbation networks and secondary structure analysis, this residue is located at the beginning of $f\alpha 2$ which unfolds at higher temperatures or when PRFAR binds.

To compare these results about amide proton environment with our MD simulations, we developed an "asymmetric" definition for the AAN. Instead of computing contacts within a selection (usually within heavy atoms), we compute contact between two selections. Here for instance we can compute the contacts between the backbone amide NH and the rest of the protein. This new way of computing contacts changes fundamentally how the contacts are computed and asked for an update of the algorithm. Remarkably, experimental temperature coefficient difference match nicely with "asymmetric" NH perturbation network and show a majority of perturbation initiated at sideR.

Our combined usage of experimental and theoretical tools on this prototype allosteric enzyme converge to the conclusion that temperature increase triggers communication pathways in IGPS from *T. maritima* which are akin to this enzyme allosteric pathways. Indeed, upon temperature increase, a series of dynamical and structural changes occur which mimics PRFAR binding at the notable exception of local perturbations near PRFAR binding site.

References

[1] George P Lisi, Allen A Currier, and J Patrick Loria. "Glutamine hydrolysis by imidazole glycerol phosphate synthase displays temperature dependent allosteric activation". In: *Front. molecular biosciences* 5 (2018), p. 4.

[2] Robert Huber et al. "Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 C". In: *Arch. Microbiol.* 144.4 (1986), pp. 324–333.

3.3.4 Manuscript 2

This work started in 2018 at Yale University in the team of Victor S. Batista and after publication of the DPCN methodology they proposed us to collaborate. The manuscript still need some adjustments before submission.

Temperature increase mimics allosteric signaling in imidazole-glycerol phosphate synthase

Florentina Tofoleanu,[*,1,2,a,†] Uriel Morzan,[*,1,3,a] Aria Gheeraert,[*,5,6] Apala Chaudhuri,[1] Zexing Qu,[1] Peter Nekrasov,[1] Bernard R. Brooks,[2] J. Patrick Loria,[1] Ivan Rivalta,[4,5,a] Victor S. Batista[1,a]

1. Department of Chemistry, Yale University, P.O. Box 208107, New Haven, CT, 06520-8107, USA
2. Laboratory of Computational Biology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892
3. The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy
4. Université de Lyon, École Normale Supérieure de Lyon, CNRS UMR 5182, Laboratoire de Chimie, 46 allée d'Italie, F69364 Lyon, France
5. Università degli Studi di Bologna, Viale del Risorgimento, 4I-40136, Bologna, Italy
6. Laboratory of Mathematics (LAMA), CNRS, University of Savoie Mont Blanc, France

[*]authors contributed equally to this work
[a] corresponding author
[†] current address: Novartis Institutes of BioMedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139

ABSTRACT

The enzyme imidazole glycerol phosphate synthase (IGPS) is a non-covalent complex of two subunit proteins (HisF and HisH) that catalyzes the hydrolysis of glutamine at the HisH active site, upon binding of the effector PRFAR to HisF at the allosteric site. IGPS is a potential target for antifungal, antibiotic, and herbicide development since it is not present in mammals and is involved in essential biosynthetic pathways of microorganisms. Here, we employ a combination of molecular dynamics simulations, network analyses, and NMR to demonstrate that temperature increase can induce a dynamics in IGPS that

resembles the allosteric activation by PRFAR binding at 25 °C. This justifies our previous enzyme kinetic and NMR dynamics studies indicating that at the growth temperature of *T. maritima* –a hyperthermophile organism– the enzymatic activity increases, while PRFAR is a weaker allosteric activator than it is at room temperature, evidencing a temperature-dependent allosteric mechanism. Our results pave the way for a more precise control of enzyme function and the expansion of drug discovery beyond the catalytic site.

Introduction

Allostery, the mechanism by which chemical signals are transmitted between spatially separated binding sites has been extensively investigated[1–14], due to significant interest in drug discovery applications.[5,15–20] Drug-like molecules that bind to allosteric sites offer advantages over traditional orthosteric modulators, including enhanced selectively in tuning responses[21] and intrinsic safeguards against overdosage.[21–23] Although concepts of allosteric drugs show tremendous promise in biomedicine, the lack of molecular level understanding of allosteric mechanisms that represent viable targets for drug discovery remains a major impediment.[15,16,24–26] Molecular level insight into the driving forces of allosteric mechanisms are necessary to elucidate and control enzyme function, expand the scope of enzyme engineering, and open new avenues for drug discovery.[18,27–30] Thus, it is critical to develop methods to establish paradigms for regulatory processes in prototypical enzymes.[2,31,3,4,6,8,12,13,32–42]

In particular, very little is known about the effect of temperature on allosteric mechanisms,[43–46] especially in thermophilic human pathogens that remain active at the elevated temperatures of their native environment. An understanding of the physico-chemical features that underlie this phenomenon could have profound implications for allosteric drug design against pathogenic organisms that survive in extreme environments.[47,48]

Here, we explore fundamental aspects of temperature-dependent allostery in the imidazole glycerol phosphate synthase (IGPS) enzyme from the thermophile *Thermotoga maritima (T. maritima)*.[5,9,49,50] IGPS is a potential therapeutic target,[51,52] since it is not present in mammals,

but rather, in opportunistic human pathogens that contain homologs of the *T. maritima* IGPS. Deletion of the IGPS gene in bacteria results in increased sensitivity to antibiotics,[53] and a decrease in infectivity.[49]

We analyze the effect of temperature on the allosteric mechanism,[46] with emphasis on the allosteric communication in the heterodimeric IGPS. The allosteric ligand PRFAR binds to the HisF subunit and enhances glutamine hydrolysis 5000-fold over its basal catalytic level at room temperature in the HisH subunit, over 30 Å away.[52] We have recently discovered that increasing the temperature of the native *T. maritima* environment drastically enhances millisecond dynamics in both PRFAR-free (apo) and PRFAR-bound (holo) IGPS. The catalytic enhancement in the holo IGPS is nearly independent of temperature in the 303-350 K range.[43] In contrast, basal levels of Gln hydrolysis increase sharply from 303 to 350 K resulting in PRFAR being a weaker activator at the physiological temperature for *T. maritima*.[43] In particular, it has been suggested that at 50 $^{\circ}$C the dynamics of the apo enzyme becomes comparable to the PRFAR-bound form, whereas at 30 $^{\circ}$C the difference between these two states is substantial.[43] Here, we show that both higher temperatures and PRFAR binding increase flexibility in some regions outside of the effector site in IGPS, enabling conformational sampling of an active enzyme form, and that PRFAR-induced motions propagate through well-defined secondary structure elements that are analogous of those involved by temperature increase.

Materials and Methods

We combined computational methods based on molecular dynamics (MD) simulations and network theory correlation analysis techniques, and nuclear magnetic resonance (NMR)[6,9,50] to study temperature-dependent allosteric communication in *T. maritima* IGPS. Computational methods have been previously used to investigate communication pathways and allostery in proteins and protein-tRNA/DNA molecular systems.[54–65]

Molecular dynamics simulations

The structural models for apo and holo structures for IGPS were based on the crystal structure of *T. Maritima* IGPS (PDB ID 1GPW, 2.4-Å resolution).[66,67] To build the apo structure, we extracted chains C and D of the HisH-HisF complex. We kept all water molecules associated with the two chains, and we further solvated the structures by using the explicit TIP3P model[66] to obtain a cubic box. Details on each system are found in the Supplementary Material.

The PRFAR-bound structure was built as previously described in Ref.[5] The protein-ligand complex was parameterized with the CHARMM36[68,69] and the generalized CHARMM force fields[70] by using the CHARMM-GUI.[71] We used AmberTools2017,[72] to convert the CHARMM file format to Amber, and the AmberGPU[73,74] package with the CHARMM36 force field for subsequent minimizations, heating, and production runs (we will make all simulation scripts available upon request). To compare the effect of PRFAR and of the increase in temperature, we simulated the apo structure at 30 ˚C and at 50 ˚C, and the IGPS-PRFAR (holo) structure at 30 ˚C. For an easier reading, the simulations of the apo system performed at 30 ˚C, and at 50 ˚C, will be referred to as apo30, and apo50, respectively. The simulation of the holo system performed at 30 ˚C and 50 ˚C will be referred to as holo30 and holo50, respectively. We simulated each system for 1 μs, and we extracted the last 0.5 μs of trajectories for analysis.

We postprocessed and analyzed the trajectories by using MDTraj,[75] CPPTRAJ[76] and pytraj.[77] The secondary structure analysis was performed by using a dictionary for the secondary structure of proteins (DSSP)[78] as implemented in MDtraj. Throughout the trajectories, a residue was assigned to the following secondary structure elements: helix (either α-helix, 3-helix, or 5-helix), sheet (extended strand, isolated β-bridge), or coil (turn, bend or loop and irregular elements). The secondary structure elements were assigned according to the information based on the crystal structure.

Eigenvector centrality analysis

In order to elucidate the allosteric pathways and pinpoint the changes in IGPS dynamics upon PRFAR binding and temperature increase, we employed the eigenvector centrality (EC) analysis recently developed within our group [1]. The method relies on mapping the MD trajectory into a graph composed by nodes separated by edges. Each node in the graph represents a α -carbon of a given amino acid. Edges between nodes are defined through an adjacency matrix A , where A_{ij} is the generalized correlation coefficient r_{MI} between nodes i and j given by

$$r_{MI}\left[x_i, x_j\right] = \left[1 - e^{-\frac{2}{3}\left(I[x_i, x_j]\right)}\right]^{1/2}, \quad (1)$$

where $I[r_i, r_j]$represents the mutual information between these amino acids[65,79]

$$I[x_i, x_j] = S[x_i] + S[x_j] - S[x_i, x_j] \quad (2)$$

$$S[x_i] = -\int p[x_i] \ln (p[x_i])\, dx_i \quad (3)$$

$$S[x_i, x_j] = -\int p([x_i, x_j]) \ln (p([x_i, x_j]))\, dx_i dx_j \quad (4)$$

Where $S[r_i]$ and $S[r_i, r_j]$ are the marginal and joint Shannon entropies respectively,[80] while $p[x_i]$ and $p([x_i, x_j])$ are probabilities of atomic displacement (x_i, x_j) computed over thermal fluctuations sampled by MD simulations at equilibrium. The generalized correlation coefficient r_{MI} ranges from zero for uncorrelated variables to 1 for fully correlated variables.

Once the adjacency matrix is obtained, diagonalizing the matrix provides an eigenvector whose values are related to each residue. The EC of an amino acid, c_i, can be defined as the weighted sum of the EC's of all the residues connected to it by an edge, A_{ij}

$$c_i = \frac{1}{\lambda} \sum_{j=1}^{N} A_{ij} c_j \,, \quad (5)$$

where λ is the leading eigenvalue of A. Hence, the EC coefficients c_i are the elements of the eigenvector associated to λ. Eigenvector centrality provides a measure of how well-connected each node is to other well-connected nodes in the network. This notion of eigenvector centrality allowed for recognition of patterns of dynamical changes associated to PRFAR binding and is here used to define and compare those associated with temperature increase.

Optimal pathways for motion transmission

In addition, to understand how the cross-talk between the active site (C84, H178, E180) and the effector binding site is altered by a temperature increase and by the PRFAR binding process, we studied the optimal pathways for motion transfer between them. This analysis was based on the

Dijkstra algorithm,[81] designed to find the roads that minimize the total distance traveled. In this study, the inter-node distance was defined as

$$w_{ij} = \; - log\left[r_{MI}(x_i, x_j)\right] \;, \tag{6}$$

therefore, the minimization of the total w travelled is equivalent to a maximal correlation between the initial and the final nodes of the path. The algorithm begins defining starting and destination nodes, which in our case were the residues hydrogen bonded to the PRFAR phosphates in the holo form, and hC84 in the active site (where the glutamine substrate binds). The pathway from the former to the latter is optimized iteratively, in each iteration the closest unvisited node is designated as the current node. From this current node, the distances to the remaining unvisited nodes are updated by determining the sum of the distance between the unvisited node and the value of the current node, if this value is less than the unvisited intersection's current value, the distance is updated. This process continues until the destination node is visited.

It is important to note that the cross-talk between two amino acids does not necessarily occur exclusively through the optimal path. Many *sub-optimal* paths with similar influence might contribute to the communication between distant residues. In this study, we built pathways merging the 50 *suboptimal* paths, representing the most likely pathways of motion transmission between the active and the effector sites.

Perturbation contact networks analysis

In order to figure out how much effector binding differs from effects of temperature increases, we applied the dynamical perturbation contact network (DPCN) analysis method recently proposed by our group.[82] Indeed, we have previously performed the DPCN to monitor the PRFAR binding effects on amino acid residues contacts and here we compare it with the temperature effect on the apo30 system by determining the contact changes with respect to the apo50. Each protein weighted contact network is built by assigning to each edge (linking the *i*-th and *j*-th residues) a weight w_{ij} that is the number of contacts between the residues. The contact condition is here defined for each pair of residues when it exists a couple of atoms (at least one

per residue) whose distance is below a given distance cutoff (here set to 5 Å) for each snapshot extracted from the MD trajectories (i.e., 10000 snapshots for each system). Further computational details can be found in our reference work on the effector binding DPCN in ref [82]. To allow easy visual inspection of DPCN results, the edges are colored in red if PRFAR binding or a temperature increase induce an increase in weight ($w_{ij} > 0$), and in blue if instead the contact number is reduced ($w_{ij} < 0$) and a weight threshold (w_t) is applied so that only the edges with $|w_{ij}| > wt$ are visualized. Here in the first part, atomic contacts are computed including only heavy atoms (i.e. excluding hydrogens) and in another part they are computed `asymmetrically" between atoms from the backbone NH and the rest of the protein.

Hydrogen bond analysis

The hydrogen bond (HB) analysis was performed by using PyHVis3D, a python-based package to calculate pairwise HBs between all donors and acceptors of all frames of the simulation trajectory.[83] The distance cutoff between acceptor and donor is 3.5 Å and the angle cutoff hydrogen–donor–acceptor is 30°. The algorithm calculates an NxN matrix (N = the number of donor/acceptor atoms in the protein), and each matrix element represents the average presence of a HB between two atoms over the simulation time.

Predicted NMR chemical shift

The structures sampled in the MD simulations were employed to predict the backbone ^{1}H and ^{15}N NMR chemical shifts using the SHIFTX2 method[84]. This program combines ensemble machine learning techniques with sequence alignment-based methods, its algorithm has been tested with high-resolution X-ray structures with verified chemical shift assignments. The SHIFTX2 analysis was performed on 20,000 configurations from each 1μs MD simulation to extract the backbone ^{1}H and ^{15}N chemical shifts at 30 $^{\circ}C$ and at 50 $^{\circ}C$; results were compared to experimental results. By combining ensemble machine learning methods to sequence alignment-based methods, SHIFTX2 data resulted in good correlation with experiment.

Experimental NMR chemical shift

Amide chemical shift data was collected for residues in the HisF subunit of IGPS. HisF was perdeuterated and ^{15}N labeled, whereas the HisH subunit was perdeuterated as described previously.[81] All data were acquired at a static magnetic field strength of 14.1 T on a Varian

Inova instrument. $^{1}H–^{15}N$ TROSY two-dimensional spectra were acquired with 32 scans and 64 increments in the t_1 dimension with corresponding spectral widths of 12000 Hz and 2800 Hz and a 1.3-second recycle delay. The temperature was calibrated using methanol as a calibration standard, and chemical shifts were recorded at 8 temperatures ranging from 20 to 55°C at 5 degree intervals. Chemical shifts were referenced using DSS as an internal standard with the ^{1}H resonance frequency of DSS set to 0 ppm.

Of the 253 residues in hisF (of which 239 amide resonances are assigned), we selected 164 residues in the HisF subunit, which were non-overlapping and for which the temperature shift was unambiguous across the temperature range. We determined the temperature coefficient in a neighborhood of 30 °C (between T_1=292.92 K and T_2=302.73 K) and in a neighborhood of 50 °C (between T_1=307.62 K and 322.41 K) as $\delta(\delta_{HN})/\delta T = (\delta_{HN,T2} - \delta_{HN,T1})/(T_2-T_1)$ and compared those temperature coefficients to investigate temperature dependent modification of the environnement for those amide protons.

Results

1. Effector binding vs. temperature increase: dynamical aspects

It has been previously shown that PRFAR induced allosteric activation is weaker at higher temperatures, and that the allosteric mechanism of IGPS is temperature dependent[43] i.e. the temperature dependence of the catalytic activity is steeper in the basal state than in the PRFAR-activated IGPS. Here, we explored the underlying molecular basis of this temperature dependence by performing molecular dynamics (MD) simulations at 30 °C and 50 °C in both the *apo* and *holo* states. We have recently shown that the EC provides a score indicating how correlated each residue is to the major motion modes, pinpointing key amino acids for IGPS dynamics. Furthermore, we have shown that the difference on the EC distribution of two equilibrium states enables the recognition of the main features associated to the transition between these two states, and hence providing a unique insight on the allosteric signalling process.[1]

Figure 1 shows the eigenvector centrality (EC) difference associated to PRFAR binding at 30 °C (left panel), 50 °C (right panel) and to the 30 °C → 50 °C temperature increase in the apo-IGPS (middle panel). The change in the EC indicates how the increase in temperature or the binding of PRFAR influences the relative contribution of each residue on the dynamics of IGPS. Remarkably, the EC trends upon temperature increase in apo-IGPS show a strong similarity to those observed on PRFAR binding process at 30 $^{\circ}$C. For the sake of clarity, IGPS can be divided in two sides, as illustrated in **Figure 1**, i.e. sideR and sideL, since a signature for the allosteric activation is the large increase in EC on loop 1 (HisF), hα1 (HisH) and hα4 (HisH) at sideR along with a depletion of EC at sideL, as previously observed.[1,5]

In contrast, the changes in EC associated with the PRFAR binding process at 50 °C are much more homogeneous amongst the residues, and qualitatively different from the behavior at 30 °C . This clearly shows that the presence of the effector has a substantially different effect on the protein dynamics at 50 °C, which can be connected to the much weaker PRFAR-induced activation at higher temperatures.

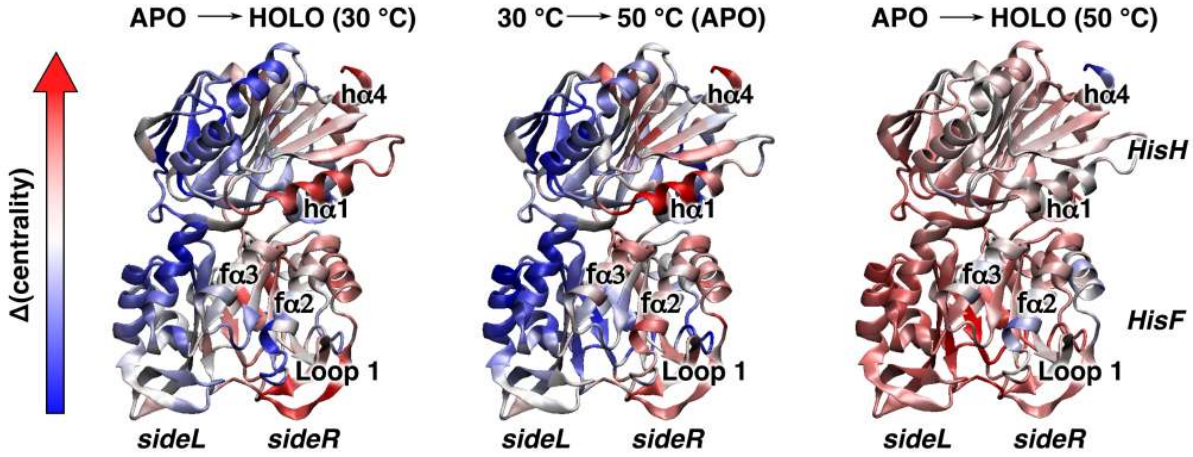


Figure 1: Eigenvector centrality difference associated with the binding of PRFAR at 30 °C (left panel), 50 °C (right panel) or temperature increase in the APO IGPS (middle panel). Residues shown in red have increased connectivity—in particular, loop 1 and hα1 become more central upon temperature increase and PRFAR binding.

In this context, it is interesting to analyze how the cross-talk between the effector binding site and the active site of IGPS, as modulated by local interactions, is modified by temperature. In order to shed light on this point, we studied the optimal residue-to-residue communication channels connecting the PRFAR phosphate binding sites with the catalytic site in HisH. All the amino acids belonging to these channels are depicted in solid color (see Figure 2), and they are distinguished between external (solvent exposed, depicted in red) and internal (surrounded by the protein matrix, depicted in blue). Noteworthy, while the communication pathways are almost purely internal for the *apo30*, the proportion of external residues is considerably increased both in *holo30* and in *apo50*. This internal-to-external transition, not only establishes another parallel between temperature increase and effector binding on the signaling pathway, but also suggests that the participation of these external residues is a key factor for the allosteric activation in IGPS from *T. maritima*.

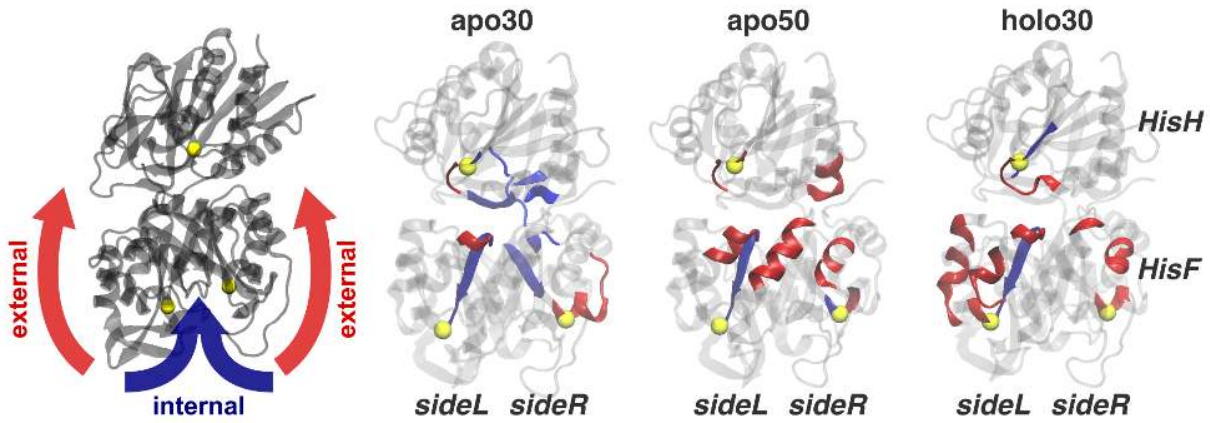


Figure 2. Optimal signaling pathways from the PRFAR phosphate binding sites (residues T104 fA224, lower yellow spheres) to the active site (hG50, upper yellow sphere). The amino acids highlighted with a solid color belong to this optimal pathway. The results for the *apo30*, *apo50* and *holo30*. As indicated in the left figure, the red fragments are composed of external amino acids (i.e. exposed to the solvent) and the blue fragments are internal residues (mainly exposed to other amino acids in the protein).

The effect of temperature increase can be regarded as an alternative route to activate the fluctuation of external amino acids, increasing the influence on the helices hα1 (HisH), hα4 (HisH) and the omega loop involved in the allosteric activation. While this activation shows clear differences between the effector binding effect and the temperature increase (**Figure 2**), in both cases we observe a strong internal-to-external transition in the communication pathway, which might be a key factor determining the catalytic activation and IGPS thermostability.

The interdomain hinge-like (breathing) motion has been recognized as one of the important elements of IGPS allosteric signalling mechanism at room temperature, representing a collective motion influenced by the effector binding.[1,5] PRFAR binding, in fact, slightly reduces the breathing motion angle (as defined by the Cα of residues fF120-hW123-hG52, see **Fig S8.** in the SI), while significantly shrinking the distribution of angle amplitudes explored by the IGPS complex. Moreover, as previously shown for 100 ns MD simulations,[1,82,83] these larger angle fluctuations in the *apo30* simulation are slower in time with respect to those in *holo30*. The breathing motion in *apo50*, instead, features a relatively small angle average (smaller than both *apo30* and *holo30*) and much more broad and asymmetric angle distribution, see **Fig. S8** in the SI.

These results overall suggest that the temperature increase has a similar impact to effector binding on the dynamics associated to local interactions during the allosteric propagation, but

somehow, less control can be achieved by temperature increase on the allosteric collective motions.

Allosteric activation vs. temperature increase: a structural perspective

In addition to the fundamental similarities between the dynamical patterns of allosteric activation and temperature increase, in this section we analyze and compare the structural changes associated with these processes. **Figure 3** shows the secondary structure changes that take place during the *apo50→apo30* and *holo30→apo30* transitions. In agreement with the dynamical changes discussed in the previous section, there are important similarities between the structural rearrangements associated to the 30 °C →50 °C temperature increase and to PRFAR binding. One of the main aspects of this resemblance is the conformation of loop1 (residues V17-D31), which adopts a combined β-sheet/helix structure in apo30, but is mostly devoid of regular secondary structure both in apo50 and holo30. This difference in secondary structure can be associated with the increase in flexibility of loop1, which has been suggested to play an important role in the activation process [2]. Furthermore, helix fα2 (R59-E71), which has previously been identified for being involved in the allosteric pathway of IGPS,[1,82,83] also shows an almost identical structural response to temperature increase and to PRFAR binding.

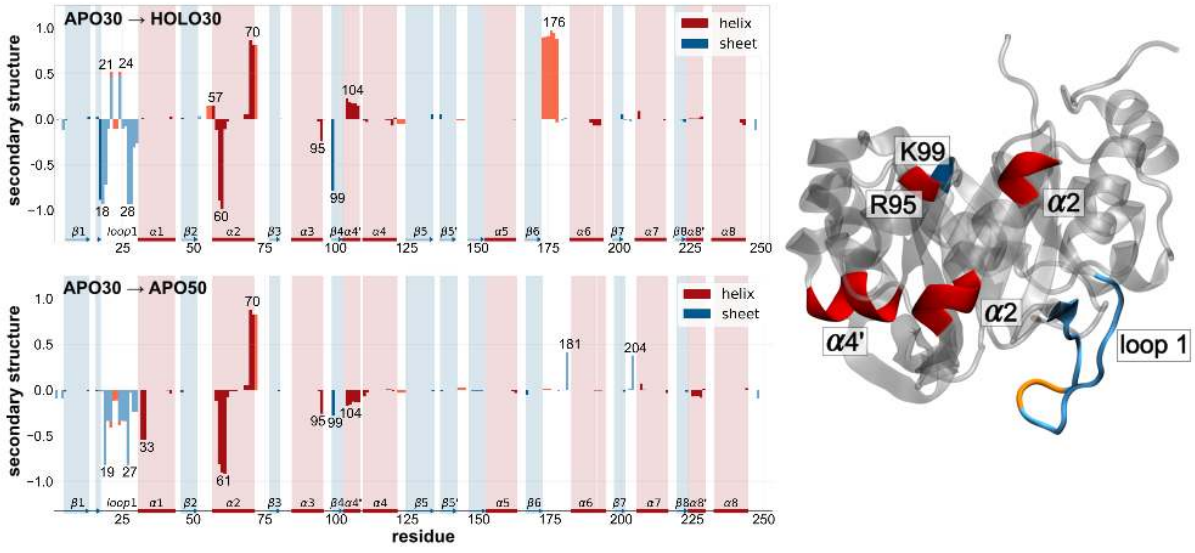


Figure 3. Secondary structure changes in HisF associated with the effector binding (upper panel), and with the temperature increase from 30 °C to 50 °C (lower panel). The right panel

represents the protein regions whose secondary structure is similarly affected by the temperature increase and effector binding.

Despite appearing as a small structural rearrangement, the reduction in secondary structure displayed by residues fR95 and fK99, which is observed in both apo30→apo50 and apo30→ holo30 transitions, is located in a critical spot for the allosteric transmission. These two residues situated on the interface of HisH and HisF are present in all the optimal signaling pathways presented in Figure 2, they also belong to the group of amino acids with higher EC increase upon PRFAR binding or temperature increase. Moreover, residue fR95 has been previously identified as one of the key step-stones in the allosteric transmission from HisF to HisH domain[6].

Conversely, there are some important structural differences observed in **Figure 3** in which the temperature increase and PRFAR binding lead to clearly different arrangements. An increased helicity in fβ6-fα6 and fα4 region is observed in holo30 , but absent in apo50. The formation of these helices is triggered by the interaction with the PRFAR phosphate groups at the ribose and the glycerol sides.

To further characterize the parallel between temperature increase and effector binding, we performed the DPCN analysis and we monitored the Hydrogen Bonds (HBs) at the HisF/HisH interface. **Figure 4** compares the changes in contacts upon PRFAR binding (left panel) with those found in the apo protein when the temperature increases from 30°C to 50°C. Notably, in both cases, the majority of contact alterations are located at the *sideR* of the protein, in analogy with the eigenvector centrality analysis (see Fig. 1). The DPCN results are also consistent with the signalling pathways analysis (see Fig. 2), since most of the contact perturbations due to both PRFAR binding and temperature increase involve solvent-exposed residues at the protein surface. Moreover, the detected contacts involve essentially the same set of nodes and edges (including perturbation signs), the differences being mostly about the absolute numbers of contact changes (i.e., the “perturbation intensities"). Interestingly, the alterations of the salt bridge network between *f*α2, *f*α3 and *h*α1 that have been recognized for the allosteric pathway of holo IGPS,[1,82,83] also appear upon temperature increase. In particular, the number of contacts between *f*α2 and *f*α3 helices increases in an almost identical way in the two cases, while *f*α2 and *f*α3 residues that lose contacts are more affected by PRFAR binding than by the rising temperature. In contrast, all changes in contacts between *f*α2 and *h*α1 are larger with temperature increase than upon effector binding. Overall, the correlation plot between the

contact perturbations induced by PRFAR binding and those due to temperature increase (from 30°C to 50°C) showed a Pearson correlation coefficient of 0.52 (see **Fig. S3**), clearly indicating the presence of similarities between the two activations of apo IGPS.

Still, some differences between the two effects are sizable, particularly near the effector site (see **Fig. S2)**, where multiple alterations are present at *sideL* upon PRFAR binding but not when increasing the temperature. We thus looked more closely at the perturbations induced around specific nodes (namely the "induced perturbation network", IPN) belonging to the fα7-fβ7 and the fβ6-fα6 turns near the effector binding site. For instance, the IPN of residue *f*D176 in the fβ6-fα6 turn showed that the presence of PRFAR increases contacts between residues fG202 and *f*G203 (directly in contact with PRFAR) and residues fR175, fS172 and fK179. In the recent PDB structure 7AC8[85], an IGPS mutant was crystallized in its active conformation. In this conformation, residue *f*D176 was found to form a salt bridge with *f*K19, which in turn forms a salt bridge with the PRFAR glycerol phosphate group. Notably, these changes indicate a propagation of contact perturbations that is consistent with our analysis of the secondary structure changes involving the fβ6-fα6 turn (see Fig. 3) and the RMSF difference plots, indicating that this element moves towards a helix structure upon PRFAR binding. When increasing the temperature, a gain of contacts in the fβ6-fα6 turn can also be noticed but not to the same extent as for PRFAR binding, in agreement with the lack of helicity increase in this region with rising temperature (see Fig. 3). Overall, these results highlight the main differences in contacts at the PRFAR binding site induced by effector binding vs temperature increase.

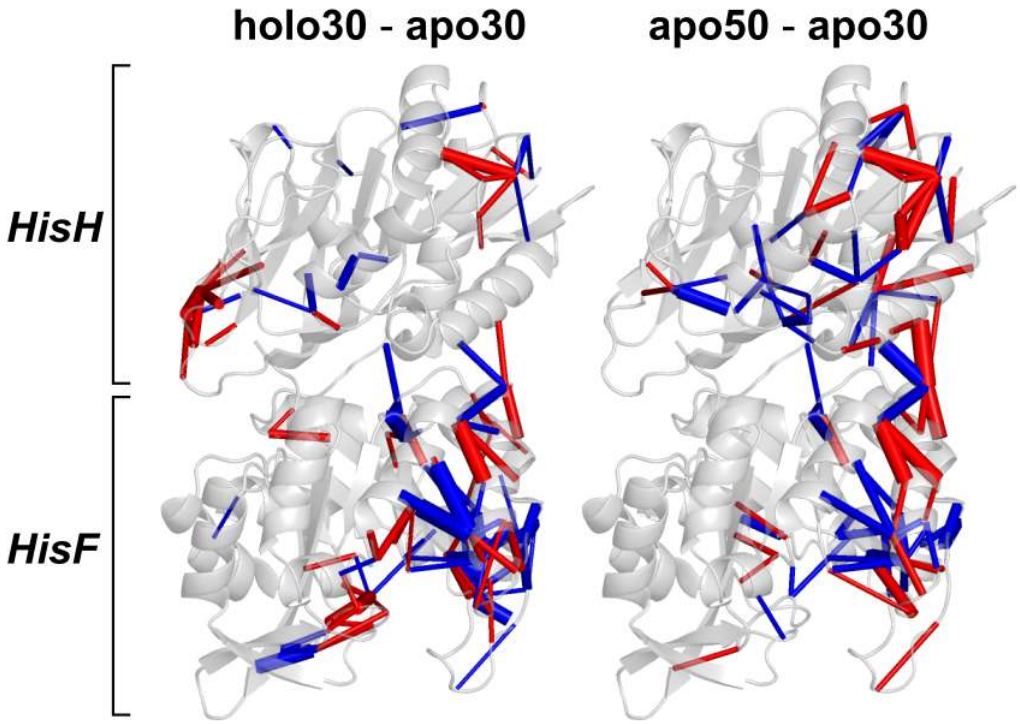


Figure 4. Perturbation contact networks between apo30 and holo30 (left panel) and between apo30 and apo50 (right panel), showing most relevant contact perturbations (i.e. a weight threshold of 5 contacts). Blue and red edges represent decrease and increase, respectively, of contacts upon effector binding (left panel) or temperature rising (right panel). Edge widths are proportional to the differences in number of contacts.

NMR chemical shifts and temperature-dependent dynamics

^{1}H and ^{15}N NMR measurements were performed in order to determine the temperature-induced chemical shifts in HisF. These measurements were compared with those computed from our simulations employing the SHIFTX2 package, and the dynamics of specific residues involved in chemical shift changes was analyzed. The SHIFTX2 results were averaged on 20000 configurations extracted from 1μs MD simulations at both 50 °C and 30 °C and compared with the experimental data (see **Fig. S5** in the SI), showing a good correlation (~0.81 for the ^{1}H chemical shifts at both temperatures). Considering the intrinsic limitations of SHIFTX2 simulated chemical shifts, these results suggest that our MD simulations can be fairly compared with our experimental NMR data.

Among the residues isotopically labelled in the HisF subunit (253 amino acid residues), 164 peaks were considered after ruling out unassigned as well as ambiguous/overlapped peaks. By careful analysis of all collected NMR data at various temperatures (from 20 to 40 °C with steps of 5 °C and at 50 °C), we observed that the chemical shift trends are quite distinct. A strong non-linear behavior (i.e. correlation coefficient of the linear fit < 0.93) is displayed by 22 residues while 32 display some curvature (i.e., p-value between a linear fit and a quadratic fit <0.05). The rest of the residues display the characteristic linear shift with increase in temperature typically observed in proteins.[86] Among these different trends, the most interesting one is that associated with residues featuring significant dynamical changes at temperatures around 30 °C and around 50 °C.

Figure 5 shows the temperature-dependent evolution of the five residues, with the most prominent change in temperature coefficient around 50° C (between 307.62K and 322.41K) and around 30 °C (between 292.92K and 302.73K). Among those top 5 residues, residue *f*L63 is the only one displaying a positive temperature coefficient. Upon temperature increase, the positive slope diminishes significantly around 307K and thereafter is near constant. Residues *f*G252 and *f*K60, featuring negative temperature coefficients, also present a change in slope at increasing temperature that tends to alleviate the temperature dependence (i.e. the slope becomes less negative around 50 °C than around 30 °C). On the contrary, residues *f*D14 and *f*D28, while featuring negative temperature coefficients, showed more negative slope around 50 °C than around 30 °C. As shown in Figure 5c, almost all of these top five residues are located at sideR near the allosteric pathways (with the notable exception of residue *f*G252, located at the C-term loop of HisF). Moreover, the asymmetric dynamical perturbation contact network between all backbone NH and the rest of the protein (i.e. representing changes in the environment around the amide proton monitored in NMR experiments) showed an interesting correlation with the temperature coefficient changes obtained experimentally. This outcome again suggests that our MD trajectories involve protein dynamics that are consistent with the available NMR data (i.e. those of the residues isotopically labelled in the HisF subunit). Thus, we looked more closely at the dynamical behavior of the residues with the greatest changes in temperature coefficient: *f*L63 (see Figure 5c). Residue *f*L63 is located in *f*α2, which undergoes rearrangement upon effector binding and is part of an altered salt bridge interaction (allosteric) network with two other helices, i.e. *f*α3 and *h*α1. Here, *f*L63, a hydrophobic residue, cannot be directly involved in the salt bridge network alteration, but is impacted through its neighbors. Upon temperature increase, the salt bridge between residues *f*K60 and *f*E90 breaks, while that between *f*K60 and *f*E64 forms. Overall, this change produces a partial refolding in the lower end of the *f*α2 helix and the backbone H-bond between residue *f*L63-NH and *f*R59-CO becomes less stable upon temperature increase (see Fig. 5d and **Fig. S10** in the SI). The presence of the intramolecular backbone H-bond is consistent with the positive temperature coefficient of this residue and its

weakening with temperature increase explains the reduction of its slope around 50 °C with respect to that around 30 °C. In contrast, residue *f*D14 is located after the end of *f*β1 and its amide proton is not involved in secondary structure formation and largely exposed to solvent, in line with its negative temperature coefficient. However, we found that the *f*D14-NH can make an H-bond with the solvent or with the sidechain of residue *f*T53, located at the end of the *f*β2 sheet, both in apo30 and apo50. The time evolution and the distribution of distances of this *f*D14-*f*T53 H-bond (see **Fig. S10** in the SI), such intramolecular interaction is occurring more often in apo50 than in apo30, suggesting a larger exchange of H-bond acceptor type (water molecule or *f*T53 sidechain) with temperature increase. Such dynamics is in line with the experimental observation of temperature coefficient decrease at around 50 °C with respect to that around 30 °C. In summary, the majority of notable changes in temperature coefficients between 30 °C and 50 °C are located near the effector site or at *sideR*, along the allosteric pathways in HisF. This reveals that temperature coefficients are suitable probes for investigating allosteric key spots, and additionally corroborates the similarities between temperature increase and effector binding effects on allosteric dynamics in IGPS.

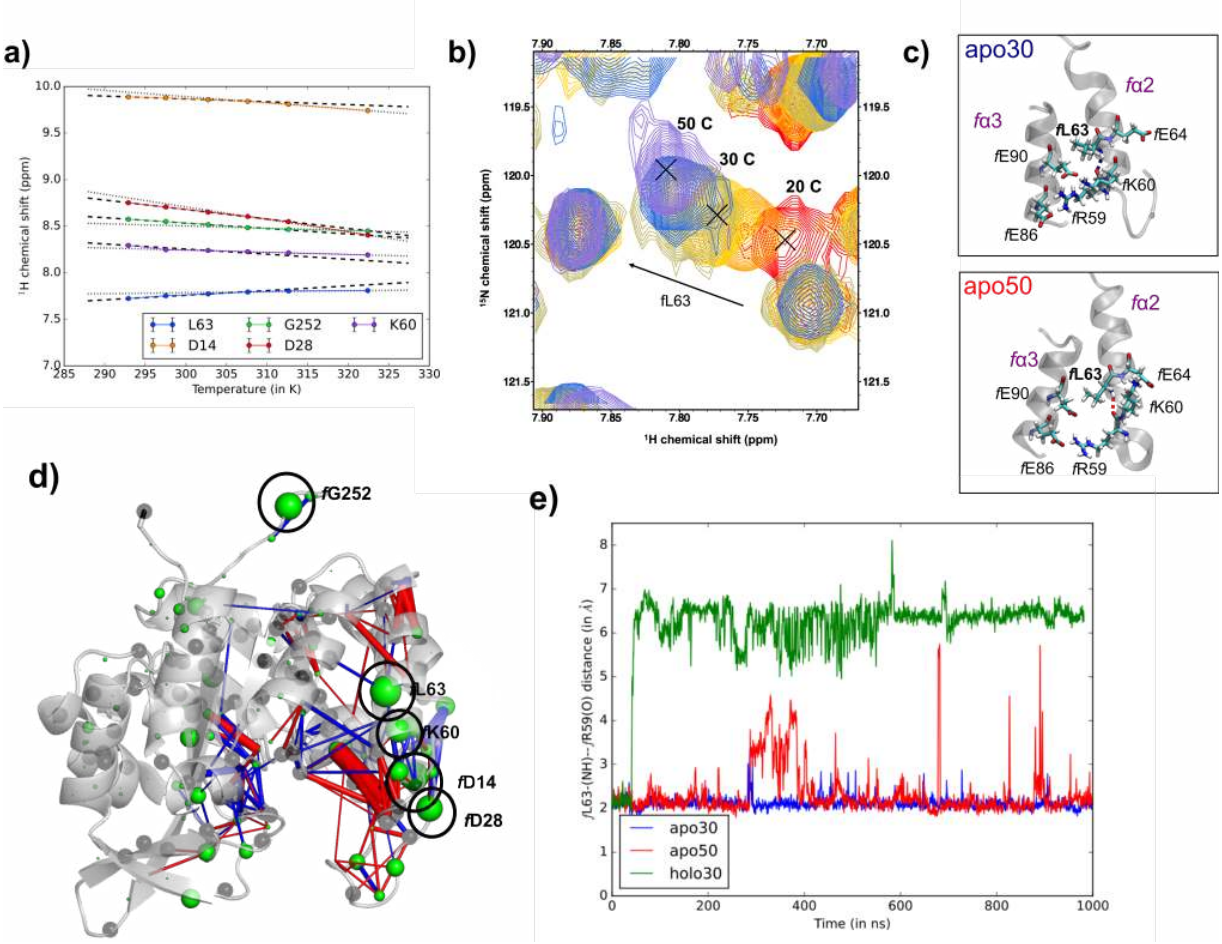


Figure 5. a) Experimental chemical shifts of the five residues with the biggest change in temperature coefficient around 30 °C (between 292.92K and 302.73K, with slope displayed by dashed line) and around 50 °C (between 307.62K and 322.41K, with slope displayed by dotted line). Typical experimental error bars for the temperature coefficients are too small to be visualized (<1ppb). b) Experimental spectral overlays for the most prominent change in NMR chemical shifts (i.e, NH of *f*L63) at all temperatures under consideration. c) H-bonding of the NH groups of *f*L63 in a representative snapshot of the apo30 and apo50 MD trajectories.

d) Superimposition of the experimental temperature coefficient changes between 30 °C and 50 °C and the asymmetric dynamical perturbation contact network between all backbone NH and the rest of the protein. The absolute variation between temperature coefficient around 30 °C and 50 °C are displayed in green spheres centered on the nitrogen atoms of N-H groups, with sphere sizes being proportional to the slope variation. Gray spheres refer to unlabeled residues in HisF, thus missing temperature coefficient values. In the perturbation network, blue and red edges represent decrease and increase, respectively, of contacts upon temperature rising. Edge widths are proportional to the differences in number of contacts. e) Evolution of the *f*L63-*f*R59 backbone hydrogen over time in the apo30, apo50 and holo30.

Conclusions

In the present work, we have demonstrated that a temperature increase from 30 °C to 50 °C in the apo state of IGPS can activate a structural and dynamical pattern that remarkably resembles the PRFAR-induced allosteric activation. We have identified the residues that belong to the signalling pathway, showing that both by binding PRFAR or increasing temperature there is an activation of an external communication channel composed by solvent-exposed residues. In agreement with this, the perturbation of the residue contacts due to both temperature increase and PRFAR binding involves mainly solvent-exposed residues at the protein surface, furthermore in both cases the majority of contact alterations belong to the *sideR* of the protein, as illustrated by our NMR temperature coefficient results and our eigenvector centrality analysis. On the other hand, the main structural and dynamical differences between the thermal and PRFAR activation, are located in the proximity of the effector binding pocket, where the thermal fluctuations cannot mimic the specific directional interactions caused by the presence of PRFAR.

The results presented here explain the origin of the weaker PRFAR-induced allosteric activation at elevated temperatures, since (i) the allosteric activation pattern is completely disrupted at 50 °C, and (ii) the intrinsic enzymatic activity of IGPS increases with temperature. In this context, the endothermic nature of PRFAR binding to IGPS[87] can be understood as an evolutionary adaptation strategy to high temperatures by compensating the loss of PRFAR-induced activation with an increased PRFAR binding affinity.

Overall, this study opens the doors for the development of novel tools to control IGPS activity, such as rationally designed allosteric drugs, pesticides or herbicides, as well as new engineered variants.

Bibliography

1. Negre, C. F. A. Eigenvector centrality for characterization of protein allosteric pathways. *Proc. Natl. Acad. Sci. U. S. A* **115**, 12201–12208 (2018).
2. Lisi, G. P., East, K. W., Batista, V. S. & Loria, J. P. Altering the allosteric pathway in IGPS suppresses millisecond motions and catalytic activity. *Proc. Natl. Acad. Sci.* **114**, E3414–E3423 (2017).

3. Lisi, G. P. & Loria, J. P. Allostery in enzyme catalysis. *Curr. Opin. Struct. Biol* **47**, 123–130 (2017).
4. Lisi, G. P. & Loria, J. P. Solution NMR spectroscopy for the study of enzyme allostery. *Chem Rev* **116**, 6323–6369 (2016).
5. Rivalta, I. Allosteric pathways in imidazole glycerol phosphate synthase. *Proc. Natl. Acad. Sci. U. S. A* **109**, 1428–36 (2012).
6. Rivalta, I. Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein-protein interface. *Biochemistry* **55**, 6484–6494 (2016).
7. Arora, K. & Brooks, C. L. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. U. S. A* **104**, 18496–18501 (3rd).
8. Malmstrom, R. D., Kornev, A. P., Taylor, S. S. & Amaro, R. E. Allostery through the computational microscope: cAMP activation of a canonical signalling domain. *Nat Commun* **6**, 7588 (2015).
9. VanWart, A. T., Eargle, J., Luthey-Schulten, Z. & Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* **8**, 2949–2961 (2012).
10. Myers, R. S., Amaro, R. E., Luthey-Schulten, Z. A. & Davisson, V. J. Reaction Coupling through Interdomain Contacts in Imidazole Glycerol Phosphate Synthase. *Biochemistry* **44**, 11974–11985 (2005).
11. Amaro, R. E., Sethi, A., Myers, R. S., Davisson, V. J. & Luthey-Schulten, Z. A. A Network of Conserved Interactions Regulates the Allosteric Signal in a Glutamine Amidotransferase †. *Biochemistry* **46**, 2156–2173 (2007).
12. Vu, P. J., Yao, X.-Q., Momin, M. & Hamelberg, D. Unraveling allosteric mechanisms of

enzymatic catalysis with an evolutionary analysis of Residue–Residue contact dynamical changes. *ACS Catal.* **8**, 2375–2384 (2018).

13. Doshi, U., Holliday, M. J., Eisenmesser, E. Z. & Hamelberg, D. Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proc. Natl. Acad. Sci.* **113**, 4735–4740 (2016).
14. Law, S. M., Gagnon, J. K., Mapp, A. K. & Brooks, C. L. Prepaying the entropic cost for allosteric regulation in KIX. *Proc. Natl. Acad. Sci. U. S. A* **111**, 12067–12072 (3rd).
15. Amaro, R. E. Toward understanding 'the ways' of allosteric drugs. *ACS Cent. Sci.* **3**, 925–926 (2017).
16. Wagner, J. R. Emerging computational methods for the rational discovery of allosteric drugs. *Chem Rev* **116**, 6370–6390 (2016).
17. Appadurai, R. & Senapati, S. Dynamical network of HIV-1 protease mutants reveals the mechanism of drug resistance and unhindered activity. *Biochemistry* **55**, 1529–1540 (2016).
18. Bogoyevitch, M. A. & Fairlie, D. P. A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. *Drug Discov. Today* **12**, 622–633 (2007).
19. Cui, D. S., Beaumont, V., Ginther, P. S., Lipchock, J. M. & Loria, J. P. Leveraging reciprocity to identify and characterize unknown allosteric sites in protein tyrosine phosphatases. *J. Mol. Biol* **429**, 2360–2372 (2017).
20. Nussinov, R. & Tsai, C.-J. Allostery in disease and in drug discovery. *Cell* **153**, 293–305 (2013).
21. Birdsall, N. J., Lazareno, S. & Matsui, H. Allosteric regulation of muscarinic receptors. *Prog Brain Res* **109**, 147–151 (1996).
22. Birdsall, N. J. Subtype-selective positive cooperative interactions between brucine analogs

and acetylcholine at muscarinic receptors: functional studies. *Mol Pharmacol* **55**, 778–786 (1999).

23. Christopoulos, A. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat Rev Drug Discov* **1**, 198–210 (2002).
24. Ruscio, J. Z., Kohn, J. E., Ball, K. A. & Head-Gordon, T. The influence of protein dynamics on the success of computational enzyme design. *J. Am. Chem. Soc* **131**, 14111–14115 (2009).
25. Toledo, L., Masgrau, L., Maréchal, J.-D., Lluch, J. M. & González-Lafont, A. Insights into the mechanism of binding of arachidonic acid to mammalian 15-lipoxygenases. *J. Phys. Chem. B* **114**, 7037–7046 (2010).
26. Kar, G., Keskin, O., Gursoy, A. & Nussinov, R. Allostery and population shift in drug discovery. *Curr. Opin. Pharmacol* **10**, 715–722 (2010).
27. Nguyen, L. *et al.* Sialic acid-containing glycolipids mediate binding and viral entry of SARS-CoV-2. *Nat. Chem. Biol.* **18**, 81–90 (2022).
28. Amici, M., Dallanoce, C., Holzgrabe, U., Trankle, C. & Mohr, K. Allosteric ligands for G protein-coupled receptors: A novel strategy with attractive therapeutic opportunities. *Med. Res. Rev* **30**, 463–549 (2010).
29. Kanuma, K., Aoki, T. & Shimazaki, Y. Recent patents on positive allosteric modulators of the metabotropic glutamate 5 receptor as a potential treatment for schizophrenia. *Recent Pat CNS Drug Discov* **5**, 23–34 (2010).
30. Verkhivker, G. M., Dixit, A., Morra, G. & Colombo, G. Structural and computational biology of the molecular chaperone hsp90: From understanding molecular mechanisms to computer-based inhibitor design. *Curr. Top. Med. Chem* **9**, 1369–1385 (2009).
31. Lisi, G. P. Dissecting dynamic allosteric pathways using chemically related small-molecule

activators. *Struct. Lond. Engl. 1993* **24**, 1155–1166 (2016).

32. Lisi, G. P. & Loria, J. P. Using NMR spectroscopy to elucidate the role of molecular motions in enzyme function. *Prog. Nucl. Magn. Reson. Spectrosc* **92–93**, 1–17 (2016).
33. Lipchock, J. M. Characterization of protein tyrosine phosphatase 1B inhibition by chlorogenic acid and cichoric acid. *Biochemistry* **56**, 96–106 (2017).
34. Wang, J., Videla, P. E. & Batista, V. S. Effects of aligned α-helix peptide dipoles on experimental electrostatic potentials. *Protein Sci. Publ. Protein Soc.* **26**, 1692–1697 (2017).
35. Palermo, G. *et al.* Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *J. Am. Chem. Soc.* **139**, 16028–16031 (2017).
36. Ho, J. Triplet-triplet energy transfer in artificial and natural photosynthetic antennas. *Proc. Natl. Acad. Sci. U. S. A* **114**, 5513–5521 (2017).
37. Block, E., Batista, V. S., Matsunami, H., Zhuang, H. & Ahmed, L. The role of metals in mammalian olfaction of low molecular weight organosulfur compounds. *Nat. Prod. Rep* **34**, 529–557 (2017).
38. Askerka, M., Ho, J., Batista, E. R., Gascón, J. A. & Batista, V. S. The MOD-QM/MM Method. in *Methods in enzymology* 443–481 (2016).
39. Ricci, C. G., Silveira, R. L., Rivalta, I., Batista, V. S. & Skaf, M. S. Allosteric Pathways in the PPARγ-RXRα nuclear receptor complex. *Sci. Rep.* **6**, 19940 (2016).
40. Adeniran, C. & Hamelberg, D. Redox-specific allosteric modulation of the conformational dynamics of κB DNA by pirin in the NF-κB supramolecular complex. *Biochemistry* **56**, 5002–5010 (2017).
41. Feher, V. A., Durrant, J. D., Van Wart, A. T. & Amaro, R. E. Computational approaches to mapping allosteric pathways. *Curr. Opin. Struct. Biol.* **25**, 98–103 (2014).
42. Schloss, A. C. Fabrication of modularly functionalizable microcapsules using protein-based

technologies. *ACS Biomater. Sci. Eng.* **2**, 1856–1861 (2016).

43. Lisi, G. P., Currier, A. A. & Loria, J. P. Glutamine Hydrolysis by Imidazole Glycerol Phosphate Synthase Displays Temperature Dependent Allosteric Activation. *Front. Mol. Biosci.* **5**, 4 (2018).
44. Saavedra, H. G., Wrabl, J. O., Anderson, J. A., Li, J. & Hilser, V. J. Dynamic allostery can drive cold adaptation in enzymes. *Nature* **558**, 324–328 (2018).
45. McGresham, M. S., Lovingshimer, M. & Reinhart, G. D. Allosteric Regulation in Phosphofructokinase from the Extreme Thermophile Thermus thermophilus. *Biochemistry* **53**, 270–278 (2014).
46. Braxton, B. L., Tlapak-Simmons, V. L. & Reinhart, G. D. Temperature-induced inversion of allosteric phenomena. *J. Biol. Chem.* **269**, 47–50 (1994).
47. Rabkin, C. S. Thermophilic bacteria: a new cause of human disease. *J. Clin. Microbiol* **21**, 553–557 (1985).
48. Alvarez-Ordóñez, A., Broussolle, V., Colin, P., Nguyen-The, C. & Prieto, M. The adaptive response of bacterial food-borne pathogens in the environment, host and food: Implications for food safety. *Int J Food Microbiol* **213**, 99–109 (2015).
49. Breitbach, K., Köhler, J. & Steinmetz, I. Induction of protective immunity against Burkholderia pseudomallei using attenuated mutants with defects in the intracellular life cycle. *Trans. R. Soc. Trop. Med. Hyg.* **102**, S89–S94 (2008).
50. Manley, G., Rivalta, I. & Loria, J. P. Solution NMR and Computational Methods for Understanding Protein Allostery. *J. Phys. Chem. B* **117**, 3063–3073 (2013).
51. Chaudhuri, B. N. Crystal structure of imidazole glycerol phosphate synthase. *Struct. Lond. Engl. 1993* **9**, 987–997 (2001).
52. Sinha, S. C. Crystal structure of imidazole glycerol-phosphate dehydratase: duplication of

an unusual fold. *J. Biol. Chem* **279**, 15491–15498 (2004).

53. Gomez, M. J. & Neyfakh, A. A. Genes involved in intrinsic antibiotic resistance of Acinetobacter baylyi. *Antimicrob. Agents Chemother.* **50**, 3562–3567 (2006).
54. Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci.* **106**, 6620–6625 (2009).
55. Black Pyrkosz, A., Pyrkosz, A. B., Eargle, J., Sethi, A. & Luthey-Schulten, Z. Exit strategies for charged tRNA from GluRS. *J. Mol. Biol* **397**, 1350–1371 (2010).
56. Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol* **2**, (2006).
57. Bhattacharyya, M., Ghosh, A., Hansia, P. & Vishveshwara, S. Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. *Proteins-Struct. Funct. Genet.* **78**, 506–517 (2010).
58. Ghosh, A., Sakaguchi, R., Liu, C., Vishveshwara, S. & Hou, Y.-M. Allosteric Communication in Cysteinyl tRNA Synthetase: A NETWORK OF DIRECT AND INDIRECT READOUT. *J. Biol. Chem.* **286**, 37721–37731 (2011).
59. Ghosh, A. & Vishveshwara, S. A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc. Natl. Acad. Sci. U. S. A* **104**, 15711–15716 (2007).
60. Ghosh, A. & Vishveshwara, S. Variations in clique and community patterns in protein structures during allosteric communication: investigation of dynamically equilibrated structures of methionyl tRNA synthetase complexes. *Biochemistry* **47**, 11398–11407 (2008).
61. Bahar, I., Chennubhotla, C. & Tobi, D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr. Opin. Struct. Biol* **17**, 633–640 (2007).
62. Chennubhotla, C. & Bahar, I. Markov propagation of allosteric effects in biomolecular

systems: application to GroEL–GroES. *Mol. Syst. Biol.* **2**, 36 (2006).

63. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).
64. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
65. Lange, O. F. & Grubmüller, H. Generalized correlation for biomolecular dynamics. *Proteins Struct. Funct. Bioinforma.* **62**, 1053–1061 (2006).
66. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
67. Douangamath, A. Structural evidence for ammonia tunneling across the (beta alpha)(8) barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Struct. Lond. Engl. 1993* **10**, 185–193 (2002).
68. Huang, M., Giese, T. J., Lee, T.-S. & York, D. M. Improvement of DNA and RNA sugar pucker profiles from semiempirical quantum methods. *J Chem Theory Comput* **10**, 1538–1545 (2014).
69. Huang, J. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
70. Vanommeslaeghe, K. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem* **31**, 671–690 (2010).
71. Lee, J. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J Chem Theory Comput* **12**, 405–413 (2016).

72. Case, D. A. *et al. Amber 2017*. (University of California, 2017).

73. Götz, A. W. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. *Gen. Born J Chem Theory Comput* **8**, 1542–1555 (2012).

74. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J Chem Theory Comput* **9**, 3878–3888 (2013).

75. McGibbon, R. T. MDTraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys J* **109**, 1528–1532 (2015).

76. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (3rd).

77. Nguyen, H., Roe, D. R., Swails, J. & Case, D. A. PYTRAJ v1.0.0.dev1: Interactive data analysis for molecular dynamics simulations. (2016).

78. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

79. Kraskov, A., Stoegbauer, H. & Grassberger, P. Estimating Mutual Information. *Phys. Rev. E* **69**, 066138 (2004).

80. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

81. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer Math* **1**, 269–271 (1959).

82. Gheeraert, A. *et al.* Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks. *J. Phys. Chem. B* **123**, 3452–3461 (2019).

83. Knapp, B. pyHVis3D: visualising molecular simulation deduced H-bond networks in 3D: application to T-cell receptor interactions. *Bioinforma. Oxf. Engl.* **34**, 1941–1943 (2018).

84. Han, B., Liu, Y. F., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. Nmr* **50**, 43–57 (2011).

85. Wurm, J. P. *et al.* Molecular basis for the allosteric activation mechanism of the heterodimeric imidazole glycerol phosphate synthase complex. *Nat. Commun.* **12**, 2748 (2021).

86. Baxter, N. J. & Williamson, M. P. Temperature dependence of 1H chemical shifts in proteins. *J. Biomol. Nmr* **9**, 359–369 (1997).

87. Lipchock, J. M. & Loria, J. P. Nanometer Propagation of Millisecond Motions in V-Type Allostery. *Structure* **18**, 1596–1607 (2010).

3.4 Singular interface dynamics of the SARS-CoV-2 Delta variant uncovered with Perturbation Contact Analysis

3.4.1 The different models

When the Covid-19 pandemic hit the world, it quickly became one of the most intensively researched topic worldwide, and we have been involved in a collaboration with the Lorraine Research Laboratory in Computer Science and its Applications" in Nancy and the "Institut de Recherche en Infectiologie" in Montpellier. Our studies have focused on the first step of viral replication: the attachment of the virus to the cell. The SARS-CoV-2 virus primary target is the human ACE2 receptor and uses the so-called Spike protein for recognition and attachment. Both the spike protein and the ACE2 receptor are glycoproteins, i.e. oligosaccharide chains (glycans) are covalently attached to some residues sidechain. Furthermore, the Spike protein associates into a homotrimer and binds with three ACE2 receptors. The key part of the Spike protein that binds with the ACE2 receptor is called the Receptor Binding Domain (RBD) and the part of the ACE2 receptor that is located outside the cell is called the ectodomain. Knowing all this, our collaboration built three models of increasing

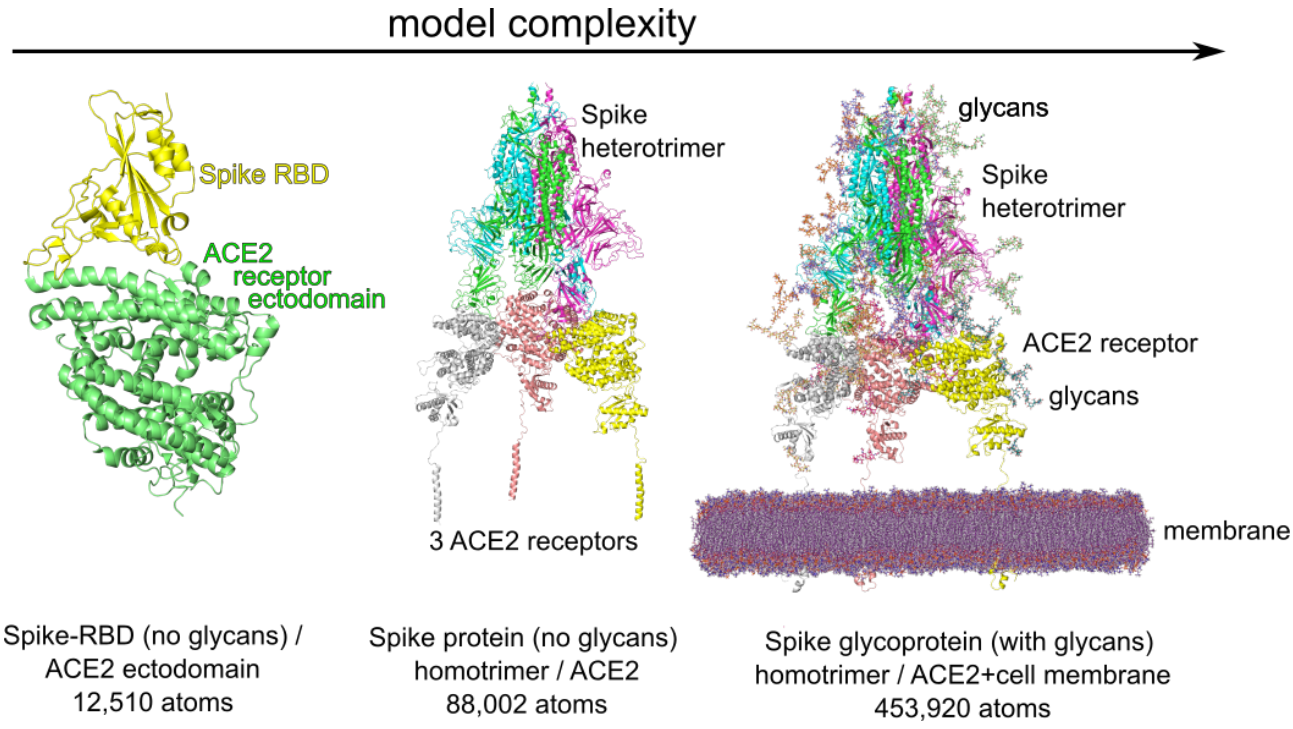


Figure 3.7: Three models

complexity (see Fig 3.7). First, a simple model with only the Spike RBD attached to the ACE2 ectodomain, without glycans, to really focus on the RBD/ACE2 interface. Then a model of the full spike protein homotrimer bound to three ACE2 receptors (without glycans). Finally, a model of the full spike heterotrimer with glycans bound to three ACE2 receptors buried in a membrane modelling the cell membrane. While the most simple model allows to multiply the simulations and to really focus on the RBD/ACE2 interface, more complex models may grant the possibility to investigate more complex mechanisms. The increase in model complexity is also rather interesting in the technical sense because it will grant us the possibility to assess how the DPCN analysis scale with the number of atoms in the system. In theory, the bottleneck is the query of distances, which is in should be in $\mathcal{O}(n \log n)$

3.4.2 Comparing SARS-CoV-2 variants

The emergence SARS-CoV-2 variants has refrained our ability to fight this pandemic through vaccines and natural immunity. Understanding key differences in the mechanism of action between the SARS-CoV-2 Wild-Type (WT) and its variants is thus crucial. Using the first model, we produced Molecular Dynamics simulations of RBD/ACE2 complexes of the WT and five variants that emerged in the year 2020 (RBD mutations in parenthesis): *Alpha* (N501Y), *Beta* (N501, K417N, E484K), *Gamma* (N501, K417T, E484K), *Delta* (L452R, T478K), *Epsilon* (L452R). While the first four were clearly more transmissible than the WT, the Epsilon had a similar transmissibility and could serve as a control.

With six different systems to compare, a total of thirty DPCNs would have been necessary to have a complete overview of the differences between the system. Even restricting ourselves to a comparative study of the WT,

looking at five different DPCNs is a tedious task. It actually is this study that motivated the development of the cPCA technique which allows extracting key contact differences and to visualize them without using prior knowledge on the system. The cPCA of the six concatenated trajectories shows that the trajectories clustered in four main groups: one with the WT and the *Epsilon* variant, one with the *Delta* variant, one with the *Alpha* variant and one with the *Beta* and *Gamma* variants. This confirmed that the cPCA can quantitatively assess when two trajectories are close in the contact space, a feature that the DPCN is unable to do. Furthermore, the PC1N and PC2N by contrast with individual DPCN contains information about all trajectories and thus have a higher degree of reproducibility in their results.

3.4.3 Scalability

Despite not being presented here since results are still too preliminary, the increasing complexity of the model provided the opportunity to test the scalability of the algorithm. In practice, DPCN and cPCA are built quasi instantaneously from the contact matrix. Therefore, the major bottleneck in computation time is the construction of the contact matrix. Contact analysis of individual frames can be conducted independently of each other. In practice, the time to compute the analysis grows then completely linearly with the number of frames. This time can also be reduced using parallelization on multiple processors. The interesting aspect of the scalability is not in the time dimensions, but rather in the system size dimension. As the number of atoms of a system increases, so does the contact analysis.

System	n_{atoms}	time (1 frame, s)	time (1,000 frames, min)
RBD/ACE2 (heavy atoms)	6,408	0.06	0.94
RBD/ACE2	12,510	0.22	3.69
Trimer (heavy atoms)	44,589	0.44	7.28
Trimer	88,002	1.84	31.7
Trimer+glycans+membrane (heavy atoms)	189,166	4.91	81.8
Trimer+glycans+membrane	453,920	32.96	549

Table 3.3: Time elapsed while building contact matrices of different systems.

In Table 3.3 are reported computation times for the different models using a single core of a Intel(R) Core(TM) i5-8350U CPU @ 1.70GHz unit. For each model we tried to compute all contacts and only heavy-atom contacts giving two different counts for the number of atoms. We report the time elapsed on 1,000 frames and averaged this time to produce a numerical value for a single frame. The systems span almost 3 orders of magnitude with the biggest system possessing almost half a million of atoms. The system with and without hydrogens are roughly separated by a factor of 2 in terms of number of atoms (respectively 1.95, 1.97 and 2.39) but the increase in computation time is much larger in each case (respectively 3.9, 4.2 and 6). This contrasts with the increase in system size between the first and the second model (multiplied by 6.95 using heavy atoms and 7.03 with all atoms) that is less translated in computation time (multiplied by 7.72 and 8.31 respectively). This suggests that adding hydrogen to the computation has another effect that drastically increases the computation time. We suggest that this may have to do with a higher density of atoms and thus a bigger number of contacts for the same amount of atoms. Another intriguing effect is the increase in system size between the second and the third model (multiplied by 4.24 for heavy-atoms and 5.15 for all atoms). while the computation time respectively increases by 11.24 and 17.91. There is a strong reason to suspect that the different density properties of the membrane there also drastically increases computation time. There may be an incentive to have a different approach in modelling contacts with a membrane. For instance, one can look at the membrane as a single unit and to study only interactions between protein and the membrane without considering internal interactions in the membrane. This approach can also be envisioned to take into account interactions between a protein and the solvent.

3.4.4 Submitted Article 1

This work led to the submission of an article in the Journal of Chemical Information and Modeling in collaboration with the team in Nancy and Montpellier in late March 2022. The reviews we received were encouraging and we present here an updated version of the manuscript based on our preliminary revisions.

Singular Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact Perturbation Analysis

Aria Gheeraert,[†,‡] Laurent Vuillon,[†] Laurent Chaloin,[¶] Olivier Moncorgé,[¶] Thibaut Very,[§] Serge Perez,[‖] Vincent Leroux,[⊥] Isaure Chauvot de Beauchêne,[⊥] Dominique Mias-Lucquin,[⊥] Marie-Dominique Devignes,[⊥] Ivan Rivalta,[*,‡,#] and Bernard Maigret[*,⊥]

†*LAMA, Univ. Savoie Mont Blanc, CNRS, LAMA, 73376 Le Bourget du Lac, France*

‡*Dipartimento di Chimica Industriale "Toso Montanari", Università degli Studi di Bologna, Viale del Risorgimento 4, I-40136 Bologna, Italy*

¶*Institut de Recherche en Infectiologie de Montpellier (IRIM), Univ. Montpellier, CNRS, 34293 Montpellier, France*

§*CNRS - IDRIS, rue John von Neumann BP 167 91403 Orsay cedex - France*

‖*University Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France*

⊥*University of Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

#*ENSL, CNRS, Laboratoire de Chimie UMR 5182, 46 allée d'Italie, 69364 Lyon, France*

E-mail: i.rivalta@unibo.it; bernard.maigret@loria.fr

Abstract

Emerging SARS-CoV-2 variants raise concerns about our ability to withstand the Covid-19 pandemic and, therefore, understanding mechanistic differences of those vari-

ants is crucial. In this study, we investigate disparities between the SARS-CoV-2 wild-type and five variants that emerged in late 2020, focusing on the structure and dynamics of the Spike protein interface with the human angiotensin-converting enzyme-2 (ACE2) receptor, by using crystallographic structures and extended analysis of microseconds molecular dynamics simulations. Dihedral angle principal component analysis (PCA) showed the strong similarities in the Spike RBD dynamics of the *Alpha*, *Beta*, *Gamma* and *Delta* variants, in contrast with those of WT and *Epsilon*. Dynamical perturbation networks and contact PCA identified the peculiar interface dynamics of *Delta* variant, which cannot be directly imputable to its specific L452R and T478K mutations since those residues are not in direct contact with the human ACE2 receptor. Our outcome shows that in the *Delta* variant the L452R and T478K mutations act synergistically on neighboring residues to provoke drastic changes in the Spike/ACE2 interface, thus a singular mechanism of action eventually explaining why it dominated over preceding variants.

Introduction

The SARS-CoV-2 virus, associated to the Covid-19 pandemic, has spread all over the world by first infecting human pulmonary cells. This critical step is achieved through specific interactions between the homotrimeric transmembrane Spike glycoprotein (S protein, with 1,273 residues in each monomer) and human angiotensin-converting enzyme-2 receptors (ACE2).[1,2] This attachment to cells is specifically mediated by the "receptor binding domain" (RBD, residues 319-541) of the Spike that binds with high affinity the N-terminal helix of ACE2,[3,4] allowing subsequent conformational changes and fusion between cell and viral membranes. As in many other viral infectious diseases, the emergence of mutant strains (or variants) ineluctably has arisen due to its zoonotic origin, interspecies transmission and human host adaptation. As the main important step in cell infection is the recognition of the specific ACE2 receptor, mutations occurring in Spike protein may confer increased or decreased in-

fectivity potential, contributing to changes in transmission rates. With the rapid emergence of variants of concern (VOC) that quickly spread worldwide, the characteristics of viral transmission, disease severity and neutralization susceptibility have been compromised. The first VOC was identified in the UK in late December 2020 (*Alpha* variant / B.1.1.7 lineage). While another variant (*Beta*, B.1.351) emerged independently in South Africa, new variants arose in Brazil (*Gamma*, P.1), in California (*Epsilon*, B.1.427/B.1.429) and finally in India (*Delta*/*Kappa*, B.1.617.1/2/3). The Alpha and Epsilon variant have been de-escalated as threat in summer 2021. In November 2021, the latest VOC (*Omicron*, B.1.1.529) was first detected in South Africa and has already spread to multiple countries and is now the current dominant form. Prior to this, the Delta variant was dominant for almost a year. The mechanisms by which these mutations modulate the infectivity or the severity of the disease are not fully understood, and only predictions can be drawn from phylogenetic studies[5] or binding free-energy calculations.[6] Focusing on the first step of viral infection or cell entry, several mutations encountered in the spike RBD are commonly shared by most variants, like N501Y or L452R. On the other hand, some mutations are more distinct, like T478K, which was exclusive to *Delta* prior to the discovery of the *Omicron* variant. The physicochemical interactions between hydrophobic and charged residues might greatly alter the recognition phase or the binding affinity between RBD and ACE2 receptors. For instance, the mutation N501T has been already shown to reduce the affinity of host ACE2 protein and S protein *in vitro*.[7] Here, we report an extensive investigation of the interaction of the Spike RBD domain with its human ACE2 receptor at the atomistic level, for the original SARS-CoV-2 virus as well as its five variants that emerged in late 2020, as detailed in Table 1. To this aim, we focus on the analysis of the primary molecular interactions between Spike and ACE2 based on experimental structural data available from the Protein data bank (PDB). First, we investigate contact changes between the available X-ray structures[3,18] Wild-Type (WT) and the *Alpha*, *Beta*, *Gamma* variants, at 2.85, 2.63 and 2.80 Å, respectively. Here, we did not compare those results with available Cryo-EM structures of the *Delta* and *Epsilon* variants[19]

Table 1: SARS-CoV-2 variants investigated in the present work. The epidemiological status is as reported by the European Center for Disease Prevention and Control (ECDC) as of 15 December 2021 (https://www.ecdc.europa.eu/en/covid-19/variants-concern). Mutations of interest found in Spike RBD compared to the WT SARS-CoV-2 strain are depicted in bold. *DE: De-escalated

WHO label (Lineage, PDB)	Status	First detected	Spike mutations	Impact on transmissibility	Impact on immunity	Impact on severity	Transmission in EU
Alpha (B.1.1.7, 7EKF)	DE*	UK (September 2020)	**N501Y**, D614G, P681H	Yes[8]	No	Yes[9,10]	Low
Beta (B.1.351, 7EKG)	VOC	SA (September 2020)	**K417N**, **E484K**, **N501Y**, D614G, A701V	Yes[11]	Yes[12,13]	Yes[9]	Medium
Gamma (P.1, 7EKC)	VOC	Brazil (December 2020)	**K417T**, **E484K**, **N501Y**, D614G, H655Y	Yes[14]	Yes[15]	Yes[9]	Medium
Delta (B.1.617.2, None)	VOC	India (December 2020)	**L452R**, **T478K**, D614G, P681R	Yes[16]	Yes[16]	Yes[16]	High
Epsilon (B.1.427/ B.1.429, None)	DE*	USA (September 2020)	**L452R**, D614G	Unclear[17]	Yes[17]	No	Very low

because the involved structures are resolved at a lower atomic resolution that does not allow appropriate computations of atomic contacts (i.e. > 3 Å). In the fight against the Covid-19 pandemic, Molecular Dynamics (MD) simulations were particularly successful at guiding vaccine development,[20,21] design RNA polymerase inhibitors,[22] investigate the binding of small molecules of the RBD,[23] designing main protease inhibitors[24] and elucidating the role of glycans in SARS-CoV-2 viral entry.[25] Notably, previous studies on the increased infectivity of variants have investigated the role of mutations on antibody-binding[26] and uncovered an allosteric signaling between mutations in the *Beta* variant.[27] In this line of works, we model all variants using a common modeling procedure starting from the WT structure with the highest resolution available at the time (PDB: 6M0J), introducing *in silico* mutations and equilibrating structures. Then we perform MD simulations of the monomeric form (1 unit of each protein) of various Spike-ACE2 systems. Thus, we performed the analysis of the primary molecular interactions between Spike and ACE2 focusing on the effect of the different mutations on the atomic contacts at the interface and the corresponding binding dynamics. This information is indeed not directly accessible from the crystallographic structural models available in the PDB ($>$ 200 X-ray or CryoEM-derived structures) and requires atomistic simulations. We adopted several tools to analyze the MD trajectories and to cross compare them, including dihedral angle principal component analysis (dPCA),[28] static and dynamical perturbation contact networks (PCN and DPCN, respectively) and contact principal component analysis (cPCA).[29] The dPCA shows that the different mutations trigger similar rearrangements inside the spike RBD in the *Alpha*, *Beta*, *Gamma* and *Delta* variant that are not fully reproduced in the *Epsilon* variant. Dynamical perturbation contact networks show that drastic differences in the interface dynamics arise between the *Delta* variant and the *Alpha*-to-*Gamma* group, despite the fact that these changes relate to mutations (L452R and T478K) that involve residues far from the interface. Finally, using cPCA, we show how synergistic effects of L452R and T478K mutations in *Delta* trigger a pattern of specific contact rearrangements that strongly affect the RBD/ACE2 interface. This knowledge on the

initial molecular mechanisms triggered by the Spike-ACE2 association provides a fundamental understanding of this critical aspect of viral infection, and may be very valuable for the rational design of antiviral therapies.

Materials and Methods

3D Models building and MD simulations

RBD/ACE2 wild type and mutants complexes

Several similar structures of the RBD/ACE2 wild type human monomer-monomer complex are available in the PDB database[1,3,7] (see Figure S1 in the SI) and we used the one with the highest resolution (2.45 Å): 6M0J.[3] The Visual Molecular Dynamics program (VMD)[30] was used to prepare the structural models starting from the WT PDB structure and to introduce *in silico* mutations. Molecular dynamics (MD) simulations were performed using the NAMD package[31] in conjunction with the recent CHARMM36 force field.[32] Six RBD/ACE2 complexes were considered in the present work: the WT and five variants among the most infectious strains (*Alpha* B.1.1.7, *Beta* B.1.351, *Gamma* P.1, *Delta* B.1.617.2 and *Epsilon* B.1.427 variant). Each protein-protein complex was placed in a TIP3P[33] water explicit solvent box of 150 $Å^3$ with periodic boundary conditions to simulate the biological environment realistically. Next, Na+ ions were added to ensure neutrality of the periodic box. Each system was firstly energy minimized performing 64,000 steps of conjugate gradient, next equilibrated (10 ns MD simulation) and a trajectory of 1 μs was then produced. The simulations were carried out in the isobaric-isothermal ensemble, maintaining constant pressure and temperature at 1 atm and 300K, respectively, by means of Langevin dynamics and Langevin piston approaches as implemented in NAMD. The equation of motion was integrated every fs, using the r-RESPA algorithm[34] to update short and long-range contributions at different frequencies. Long-range electrostatic interactions were treated using the particle-mesh

Ewald approach.[35] Every ps, one frame was saved from the trajectory file, leading to a total of 1,000,000 frames for further analysis.

MD Analysis tools

Root-mean-square deviation

The root-mean-square deviation of atomic positions is a first rough indicator of simulation convergence. First, we align trajectories with respect to their initial conformation by minimizing the RMSD of backbone atomic positions. Then we report minimal RMSD fluctuations over time. Since, in our models, the spike RBD contains 229 residues and the ACE2 protein 603, it is possible that averaging the RMSD on the global ACE2/RBD complex hides destabilization due specifically to mutations in the RBD. To assess more directly possible effects of mutations, we also compute the RMSD of backbone atomic positions restricted either to the RBD (excluding terminal segments, residues S325-N540) or to the Receptor Binding Motif (RBM, residues S438-Q506) where most mutations are located.

Dihedral angle principal component analysis

Principal component analysis (PCA)[36–44] of MD simulations is a general method to extract essential motions of a system and to reduce the high-dimensional evolution of a proteic system in a low-dimension landscape. In PCA, the feature choice is crucial and there has been an incentive to use internal coordinates like dihedral angles[28] over external coordinates (e.g. Cartesian coordinates).[45–47] In this formulation, for each frame we compute $2N$ dihedral angles and linearize them from the circular space using the transformations:

$$q_{4n} = \sin\phi_n; \quad q_{4n+1} = \cos\phi_n; \quad q_{4n+2} = \sin\psi_n; \quad q_{4n+3} = \cos\psi_n \tag{1}$$

with $n = 1, \ldots, N$ corresponding to the N pairs of consecutive residues from which dihedral angles are considered (in practice $= N_{\text{residues}} - N_{\text{chains}}$). In this study, we accounted for all

ϕ and ψ backbone dihedral angles. Since RBD variants only show single point mutations, the considered models have all the same number of backbone dihedral angles and can be compared straightforwardly. An observation matrix $Q_{i,j}$ of size $(N_{\text{frames}} \times 2N)$ is constructed, where the columns are all linearization of ϕ and ψ dihedral angles and the rows all possible observation states (10,000 frames for the WT and each variant so 60,000 frames in total). The scikit-learn[48] implementation of PCA decomposition to get the principal components (PCs) was used. Restricting to the two first eigenvectors, they can be used to obtain the free energy-landscape of the system:

$$G(\text{PC1}, \text{PC2}) = -k_B T \left[\ln P(v_1, v_2) - \ln P_{\max}\right] \tag{2}$$

Here $P(v1, v2)$ is the probability distribution obtained from a bivariate kernel density estimate,[49,50] which is subtracted to ensure that $\Delta\text{G} = 0$ for the lowest free energy minimum. Then the influence of the nth consecutive pair of residues in a component i is expressed as the sum of the squares of the influence of its features:

$$I_{i,n} = \sum_{j=n}^{n+3} v_{i,j}^2 \tag{3}$$

where v_i is the eigenvector corresponding to component i and $v_{i,j}$ the coefficient corresponding to feature $q_{i,j}$.

Ward's minimum variance method

Considering that the dPCA is built on maximization of variance property, in order to find clusters of frames in the highest density regions of the projection, it is meaningful to group together minimum variance regions. Thus, Ward's minimum variance method[51] has been used to build a hierarchical clustering of the frames in the projected space. We then measure the discrete acceleration of the height of each consecutive cluster, and we set the optimal number of clusters as the one that maximizes this acceleration. The acceleration on the

x-axis is shifted so that the initial acceleration value is for a number of clusters equal to two. The ensuing clustering of frames allows to differentiate regions with the highest density in the system energy-landscape. Ward's minimum variance method also provides a good way to detect key moments in a given simulation where the system undergoes large dynamical changes.

Perturbation contact network analysis

Contact networks represent a protein as a collection of nodes, i.e. residues, that are connected by edges if those residues satisfy a contact condition. Here, in line with our previous works,[52–54] the contact condition is achieved if at least one heavy atom from a residue is at a distance below 5 Å from another heavy atom in another residue. Edges between residues are then weighted by the total number of atomic contact pairs that satisfy this contact condition. Individual contact networks can be obtained from experimental PDB structures or from frames of MD simulations. "Static" contact networks are derived from a single experimental structure, while time-averaged networks of MD simulations correspond to dynamical contact networks. Then, in order to compare two contact networks (whether static or dynamical) and highlight contact differences between these structures, we subtract one from the other (formally, we subtract their weighted adjacency matrices). The differences between the two contact networks are visualized on the 3D model of the protein by assigning colors to the edges of the dynamical perturbation network according to the sign of the edges. Here, when we subtract the WT network from the mutant network, we assign the color red to a positive sign (i.e. stronger contacts in the mutant) and blue to a negative sign (i.e. stronger contacts in the WT). Finally, for visualization purposes, a weight threshold can be applied to select edges kept for display. Here, in line with previous works,[54] using a heavy-atom network, we used an absolute threshold of 5 when explicitly mentioned. Isolated nodes after this process are also pruned to simplify the visualization. The main advantage of such a method is to get a direct and global view of all interactions resulting from chain motions and to allow the

detection of subtle movements, including those occurring in loops.

Contact principal component analysis

We report the weights of the contact networks of every frame in a matrix C of size $N_{frames} \times N_{contacts}$. If a contact is not present in one frame, its weight is simply put as zero. We use Principal Component Analysis (cPCA for contacts) to extract the principal components. The PCs are each of size N_{frames} and represent the projection of the frames in this component. During the decomposition, we compute the (ordered) eigenvectors of the covariance matrix. Each of these eigenvectors correspond to a principal component and is of size $N_{contacts}$, thus representing a linear combination of all contacts in the system. We define a new type of contact network: the ith PC Network (PCiN) in which nodes are the amino acids of the protein, edges are all contacts and weights are the value of the contact in the eigenvector. These eigenvectors also corresponds to an eigenvalue, which is representative of the importance of the principal component. By design, the eigenvalues in PCA and eigenvectors are ordered, thus the PCs decrease in importance with the component number. Similarly to dPCA frames can be cluster using Ward's minimum variance method in the first principal components.

Results

Static perturbation contact analysis

Recently available structures of the *Alpha*, *Beta* and *Gamma* variant RBD in complex with the ACE2 protein[18] give precious molecular basis for the understanding of altered binding in emerging variants. In Figure 1B-D, we report the static perturbation contact network (PCN) between the RBD/ACE2 complex from the *Alpha Beta*, and *Gamma* variants (respectively PDB: 7EKF, 7EKG, 7EKC) and the WT (PDB: 6M0J, 1A), showing the main difference

in atomic contacts deducible from X-ray experiments. Focusing on the WT, the interface between the spike RBD and the ACE2 involves various secondary structure elements in the spike RBD. First, in the α3 helix, residue K417 is in contact with residue D30 located in the α1 helix of the ACE2 receptor. Then, the α4-β5 loop (residues D442-Y451) has a few contacts with the α1 helix of ACE2 (i.e. G446-Q42 and Y449-D38). In the β5 sheet (residues L452-R454), residue Y453 is in contact with H34 of the ACE2 α1 helix. The β5-β6 loop (residues L455-F490) is also mainly in contact with the α1 helix (L455-H34, F456-T27, N487-Q24, Y489-F28, Y489-T27, Y489-K31) but some residues are also interacting with the α2 helix of the ACE2 receptor (N487-Y83, F486-L79, F486-Y83). In the β6 sheet (residues P491-Q493), residue Q493 is in contact with H34 and E35 of ACE2 α1 helix. The nearby β6-α5 loop of RBD (residue S494-Y505) is also interacting with ACE2 α1 and with the β-turn (G352-D355), the most relevant contacts being: Q498-Y41, Q498-Q42, N501-Y41, N501-K353, Y505-K353. The largest number of atomic contacts (i.e. 43 atomic pairs) in the WT is found for the interaction Y505-K353 while the N501-K353 and Q498-Y41 contacts are tied second (with 25 pairs). Among all mutated residues involved in the variants studied here, only K417 and N501 have a significant contact across the interface ($<$ 5 atomic contacts) in the WT. It has to be noted that RBD residue E484 also possesses a minimal contact (1 atomic pair) with K31 in the ACE2 α1 helix.

In the *Alpha* variant, which contains only the N501Y mutation, the main contact changes are directly associated with this residue, featuring an increase in contact between the Y501-Y41 and Y501-K353 pairs (+14 and +11 atomic pairs, respectively). These increases in contacts are partially compensated by some contact losses, including those of the Q498-Q42 (-7) and Y505-E37 (-5) interactions. Interestingly, far from the mutation spot, there is also an increase in contact with H34 in the ACE2 α1 helix associated to the Q493-H34 and the Y453-H34 interactions (+13 and +10 atomic pairs respectively, see Table 2) and some decrease of contact with the ACE2 α2 helix, involving F486-L79 (-5) and F486-Y83 (-4).

Table 2: Contact values at critical residues in the Spike/ACE2 interface in X-Ray PDB structures of the WT (PDB: 6M0J) and *Alpha* (PDB: 7EKF), *Beta* (PDB: 7EKG) and *Gamma* variants (PDB: 7EKC).

Spike	K417	Y453	L455	F456	E484	F486	F486	Q493	Q493	Q493
ACE2	D30	H34	H34	K31	K31	L79	Y83	K31	H34	E35
WT	7	11	17	9	1	7	19	3	14	14
Alpha	6	21	16	10	1	2	15	11	27	0
Beta	0	14	11	12	0	9	18	10	7	2
Gamma	0	11	16	15	0	13	18	4	12	7

Spike	Q498	N501	N501	N501	Y505
ACE2	Q42	Y41	K353	G354	E37
WT	20	15	25	2	11
Alpha	13	26	39	2	6
Beta	14	28	35	2	4
Gamma	17	33	35	2	5

Overall, the increase in number of atomic contacts of *Alpha* variant with respect to WT is about 2%.

In the *Beta* variant, the same direct influence of the N501Y mutation is observed around residue N501. As expected, the K417N mutation breaks the K417-D30 salt-bridge and contact losses are observed for K417-D30 (-7) and for nearby contacts: Q493-H34 (-7), Q493-E35(-12) and L455-H34(-6) pairs. A slight increase in contacts for the Q493-K31 pair (+8) partially compensates this effect. The other mutation, i.e. E484K, breaks the weak E484-K31 contact (-1). Overall, the *Beta* variant features a loss of about a 5% of contacts with respect to WT.

Finally, the *Gamma* variant is very similar to *Beta* but there the intensification of the Y501-Y41 contact is further magnified (15 atomic contacts in the WT, 26 in *Alpha*, 28 in *Beta* and 33 in *Gamma*) while contacts losses due to the loss of the K417-D30 salt bridge (T417-D30 also loses the 7 atomic contacts) are mitigated: only the Q493-E35 pair (-5) undergoes contact loss. Two other inter-residues interactions show some indirect effects of those mutations at the interface, i.e. F456-K31 (+6) and F486-L79 (+6). Similar to the *Alpha*, the *Gamma*

variant features a ca. 1% of contact increase with respect to WT.

More general trends of intra-domain contact perturbations can be observed in the static PCN analysis, indicating that ACE2 contacts are more affected by mutations than RBD ones for the *Alpha*-to-*Gamma* variants and, overall, the *Gamma* variant features larger perturbations than the two others. The valuable information available from this static PCN analysis is however lacking dynamical effects that are going to be characterized in the following sections, where we also consider the comparison with the *Delta* and *Epsilon* variants that lack crystallographic structures with a resolution below 3 Å.

Dihedral angle principal component analysis

We performed microsecond MD simulations on the WT and its five variants *Alpha*-to-*Epsilon* to characterize the effects of mutations on the RBD and ACE2 dynamics. RMSD analysis of these MD trajectories (Figures S2-S4) indicated that all systems equilibrated within 200 ns after the pre-equilibration steps, including the domains where most mutations are present, i.e. RBD and RBM. The dPCA has been initially performed on the whole (1 microsecond) MD simulation of each system (i.e. the concatenated values of backbone dihedral angles in all the frames for each system, see Figure S5 in the SI). Because the ACE2 receptor is much more flexible than the RBD and to focus on dynamical changes in the RBD we restrain the dPCA analysis to dihedral angles of the RBD. For each simulation, the PC1 and PC2 values undergo drastic adjustments between 200 and 600 ns. This indicates that some major rearrangements occur in the system, some of which can be attributed to the incorporation of *in silico* mutations. The latest of these important shifts occurs at 600 ns in the *Gamma* variant. Since this variant contains three different mutations (the most in any studied variant, tied with *Beta*) this is not surprising that it is the last to converge. Ward's minimum variance method shows an optimal number of four, and each simulation remains in the same cluster during the last 400 ns. This indicates that our simulations

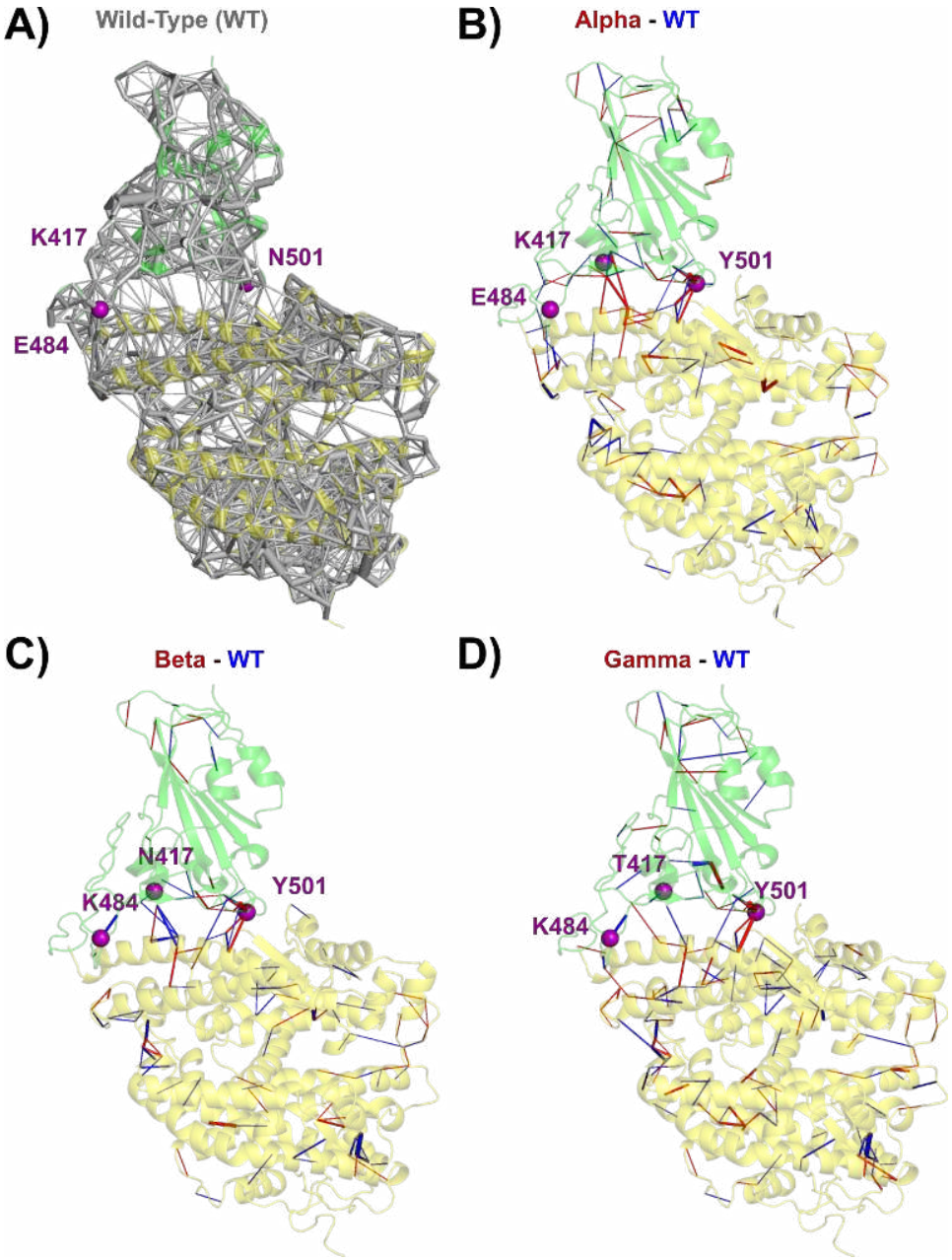


Figure 1: (A) Static amino acid network of the WT. (B-D) Static perturbation network at threshold 5 between *Alpha* (PDB: 7EKF, B), *Beta* (PDB: 7EKG, C), *Gamma* (PDB: 7EKC, D) and the WT (PDB: 6M0J)

have appropriately converged, and we can proceed with dPCA. Thus, here and in all the remaining analysis of this work, we focus on the frames of the last 400 nanoseconds for all MD simulations (employing then $N_{\text{features}} = 722$ and $N_{\text{frames}} = 24,000$).

When representing MD frames in a PC1 vs PC2 plane, as depicted in Figure 2, the WT

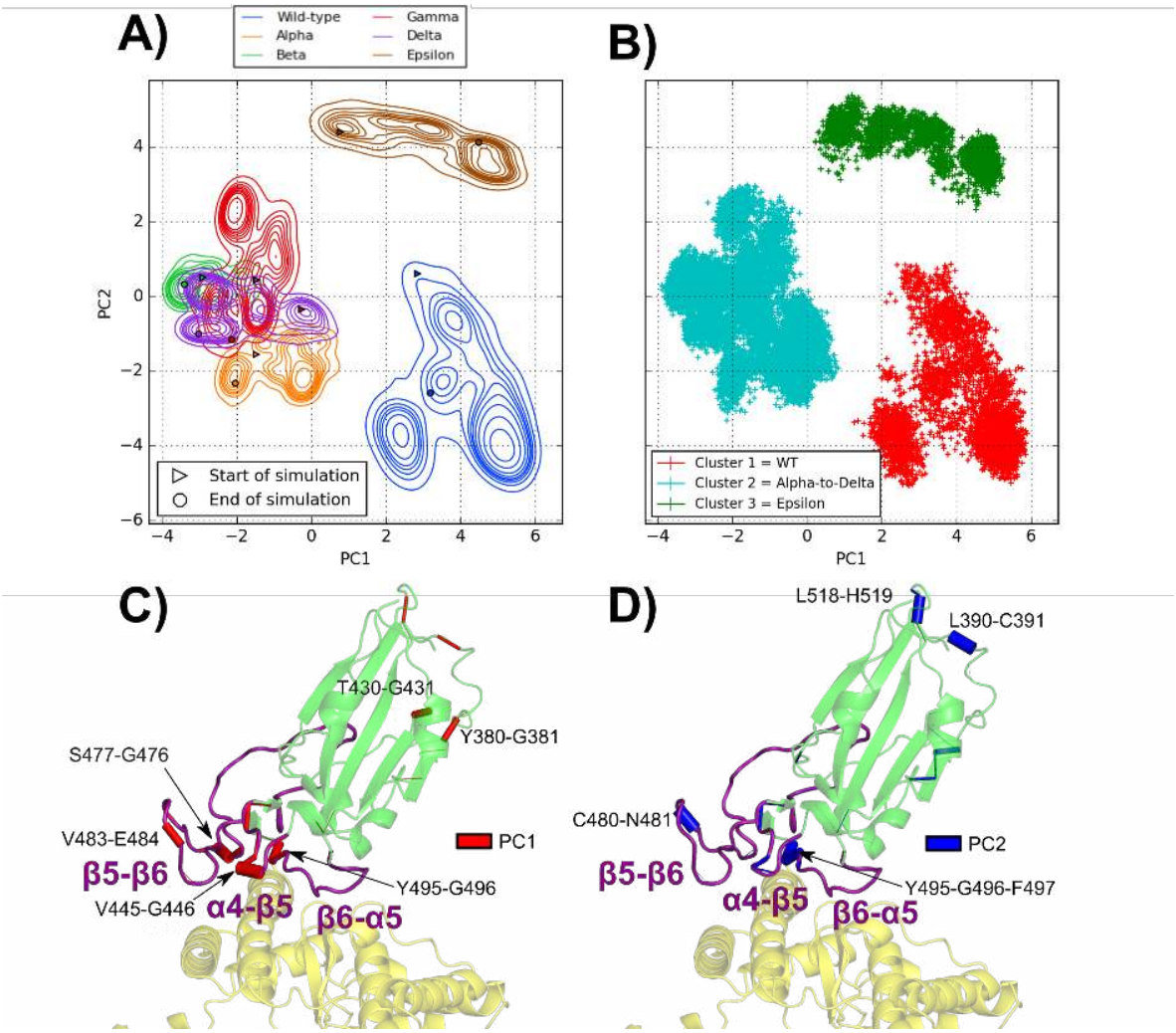


Figure 2: Projection of the frames corresponding to the final 400ns of simulation for the six studied complexes in the two dPCA eigenvector dimensions with (A) contour plots representing a kernel density estimate of the population of each complex, (B) scatter plot representing the three main clusters obtained through Ward's minimum variance method. Representation of the influence (as cylinders with a width proportional to the influence) of each dihedral angle in the PC1 (C) and PC2 (D) eigenvectors on the spike-RBD (green)/ACE2 (yellow) complex. The α4-β5, β5-β6 and β6-α5 loops are highlighted in purple.

and *Epsilon* systems are both isolated (PC1>0 and PC2<0 for WT; PC1>0 and PC2>0 for *Epsilon*) from *Alpha*-to-*Delta* variants that are grouped together (PC1<0, PC2≈0). This grouping of *Alpha*-to-*Delta* variants as function of the first two dPCA components suggests that different mutations might have similar effects on the RBD motion with respect to that of WT (see time evolutions of PC1 and PC2 in Figure S6D-E in the SI). In fact, the *Delta*

variant does not share mutations with *Alpha*, *Beta* and *Gamma* that, instead, all have in common the N501Y mutation. Notably, the *Epsilon* variant, despite sharing the L452R mutation with *Delta*, is separated from it (see also Fig. S6D-E). The dPCA results indicate that the PC1 (i.e. the largest variance axis) discriminates the *Alpha*-to-*Delta* group from both the WT and the *Epsilon* variant. Looking at the main conformational changes in the MD simulations, one can realize that the motion relating to WT and *Epsilon* (along PC1) refers to a large displacement of the α4-β5 loop (see Fig. S15 in the SI). On the other hand, the second principal component separates *Epsilon* from all the other systems, mainly because they feature different fluctuations of the β5-β6 loop (see Fig. S16 in the SI). Ward's minimum variance method quantitatively confirms this behavior, showing an optimal number of clusters (see Fig. S5 in the SI) equal to three, corresponding to the WT, *Epsilon* and *Alpha*-to-*Delta* groups. Interestingly, a previous study comparing the dynamics of the SARS-CoV-2 and SARS-CoV (responsible for the SARS 2003 outbreak) evaluated that the increased rigidity in the β5-β6 loop of the SARS-CoV-2 was linked to its higher infectivity because it enabled the formation of more stable bonds across the interface.[55] This is in line with our results and suggests that the higher rigidity in the *Alpha*, *Beta*, *Gamma* and *Delta* variant α4-β5 loop increases their transmissibility.

In Figure 2, the residue pairs with the most influence on the RBD dynamics are reported. The vast majority of these residues are located in three loops belonging to the RBM (438-506): α4-β5 (residues L455-F490), β5-β6 (residues L455-F490), β6-α5 (residue S494-Y505). It's interesting to note that the α4-β5 and β6-α5 loops are in contact and contain respectively mutations L452R (*Delta* and *Epsilon* variants) and N501Y (*Alpha*, *Beta* and *Gamma* variants). The time-evolution of the V483-E484 dihedral angles (see Fig. S8 in the SI) actually shows that their fluctuations are analogous in variants with (*Beta* and *Gamma*) or without (*Alpha*, *Delta*, *Epsilon*) the E484K mutation. On the other hand, in the WT, these dihedrals have a different behavior, i.e. featuring larger fluctuations and significant shifts in the microsecond simulations. This suggests that, while the V483-E484 dihedral

angle is involved in the main conformation motions of the RBD, the E484K mutation is not alone responsible for alterations of the RBM structure and motion. In fact, a previous study has uncovered an allosteric cross-talk between mutated residues K484 and Y501 mediated notably by N417[27] in the *Beta* variant. Other sources of this cross-talk are found near the mutation spots in the β5-β6 and β6-α5 loops, precisely where are located our main dihedral changes in PC1 and PC2. The present results suggest that in the different variants, there are cross-talks between β5-β6 and β6-α5 which affects the loop flexibility. The above analysis of critical dihedral angles is therefore useful to understand the dynamics of the RBD upon mutations and to characterize some similarities and differences among various variants. However, dPCA does not provide an atomistic picture of the ACE2 and Spike RBD proteins responses to mutations. In order to recover this important information, an analysis of atomic contacts is reported in the next section, with a focus on the ACE2/RBD interface.

Dynamical perturbation contact network analysis

The dynamical contact network of the WT simulation and dynamical perturbation contact network (DPCN) between variants and the WT are reported in Figure 3 (the individual amino acid networks are reported in Figure S9 in the SI). At first glance, the resemblance between DPCN from *Alpha*-to-*Delta* simulations is striking. Inside the Spike RBD, there is one main patch of contact changes located between the α4-β5 and β6-α5 loops that is present in the *Alpha*-to-*Delta* variant, while a similar (but not identical) patch exists in the *Epsilon* variant. Interestingly, parts of the RBD located farther from the interface with ACE2 appear significantly less affected by mutations. The interface between the two proteins displays some contact changes but, with the notable exception of *Delta*, these are of a lesser magnitude (i.e. smaller number of total atomic contacts for each residue pair) than internal contacts perturbations in the RBD and in the ACE2 receptor.

Surprisingly, the ACE2 receptor is subject to much more contact changes than the RBD upon mutations, and some resemblance between contact perturbations can be observed be-

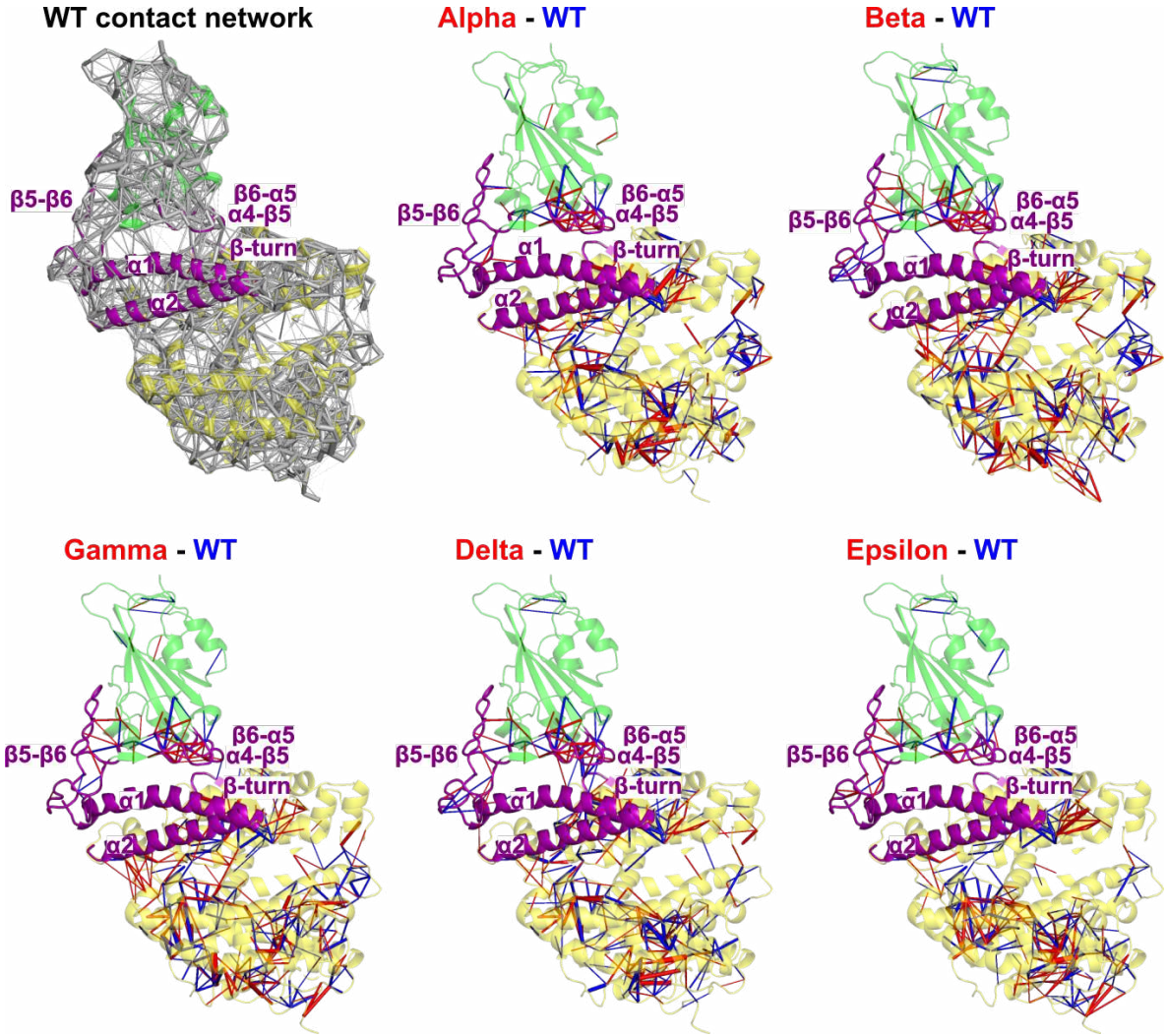


Figure 3: Complete Perturbation Network between each variant and the WT. The spike-RBD(green) / ACE2(yellow) complex is represented in cartoon representation. Stronger contacts in the WT are represented by a blue edge and in the variant in red. Edge width is proportional to their weight.

tween the five variants. This is consistent with studies showing that the ACE2 receptor is significantly flexible, in contrast with a high stability of the RBD/ACE2 interface.[56,57] In particular, simulations of a ACE2 homodimer bound to the RBD shows some conformations which may accommodate for the binding of a single SARS-CoV-2 RBD to multiple ACE2 units. Looking at the propagation of perturbations within the ACE2 receptor, from the RBD interface to the opposite side of the ectodomain, one could speculate that, upon mutations in RBD, the binding of these five Spike variants might eventually trigger a response of the

ACE2 receptor that significantly differs from that of the WT; shifting the conformational ensemble of the RBD/ACE2 interaction towards RBD units binding to multiple ACE2 receptors.

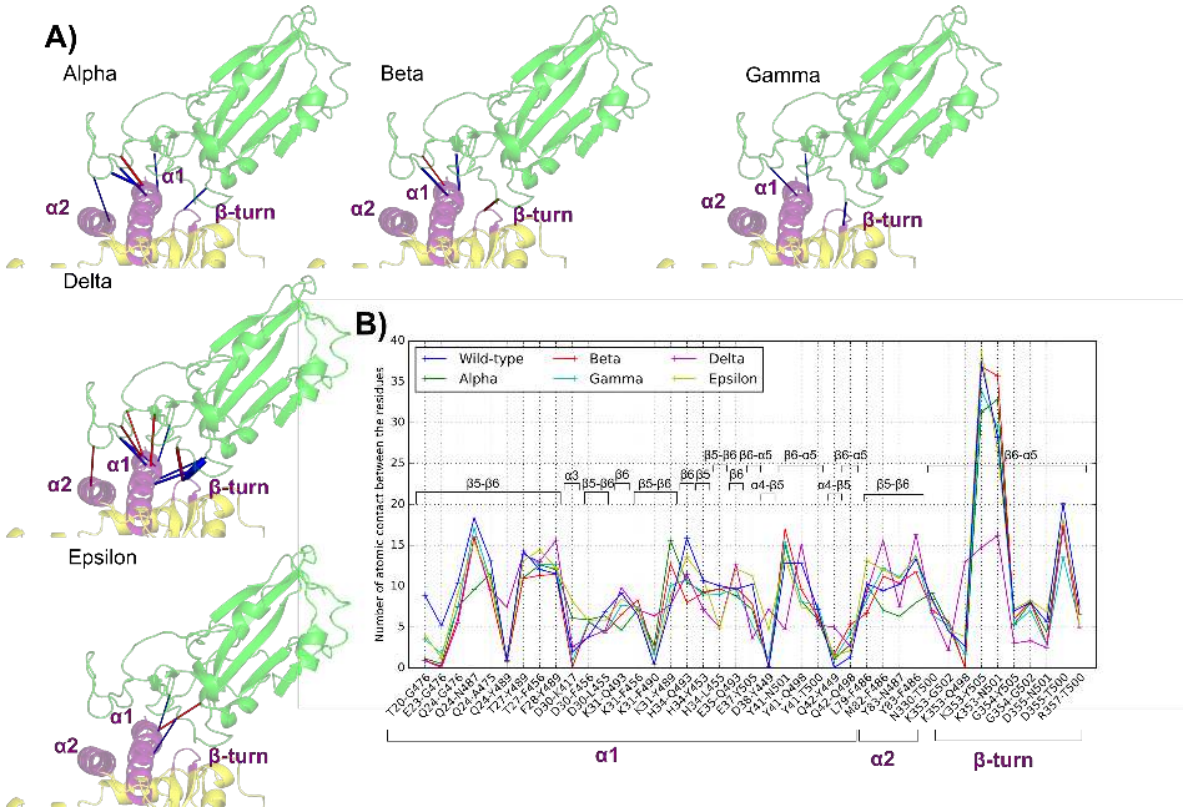


Figure 4: (A) Perturbation networks using a threshold value of 5 between the WT RDB (green)/ACE2 (yellow) complex and its mutants (*Alpha*, *Beta*, *Gamma*, *Delta* and *Epsilon*). Stronger contacts in the WT are represented by a blue edge and in the variant in red. Edges width is proportional to their weight and visualization factor the same for each variant. (B) Average number of interresidual atomic contacts in all pairs at the interface (labeled in the WT residue name) with more than 5 contacts in at least one simulation.

Notably, when considering the total number of average contacts at the interface in the last 400 ns of MD simulations (see Figure S11 in the SI), all variants feature less atomic contacts at the interface than the WT. In particular, the interface between the ACE2 receptor and the *Alpha* and *Beta* variant shows a decrease of 12% in atomic contacts, the *Gamma* interface a decrease of 11%, while the *Delta* and *Epsilon* interfaces decrease by 4%. This is

counterintuitive since we expect variants to show a higher RBD/ACE2 affinity, leading to an increase in contact count. In fact, experimentally, there is not a strict correlation between infective and transmissible variants and a higher affinity of the RBD/ACE2 complex.[58] This suggests that variants uses more complex mechanisms for cell entry and, in particular, a mechanism in which the RBD binds to more than one ACE2 unit is not predictable using our modeling. Therefore, the simplified mechanism described here at the RBD/ACE2 interface may be only the first step of a more complex mechanism in which the different variants facilitate the binding of the Spike trimer to more than one ACE2 receptor (e.g. PDB: 7V89 in the *Delta* variant). In fact, within this context, a slight destabilization of the monomeric RBD/ACE2 interface can be favorable to trigger RBD binding to multiple ACE2 receptors.

In Figure 4, a close view of the DPCN near the ACE2/RBD interface is reported along with the list of contact pairs involved (Figure 4B). The Spike RBD binds to three main areas of the ACE2 receptor, two helices, i.e. α1 (residues T20 to Y41) and α2 (mainly residues L79, M82 and Y83), and a β-turn (residues G352-D355). Among the WT residues mutated in the five variants, which are all located close to the RBD/ACE2 interface, only residues K417 and N501 are involved in the interface contacts during the MD simulation of WT, i.e. possessing (in average) > 5 atomic contacts with ACE2. Despite only two mutated residues are directly involved in the interface contacts, other atomic contacts at the interface are indirectly affected by mutations.

Here, we describe the direct and indirect contact perturbations upon mutations in the five variants. As shown in Figure 4A, the *Delta* is certainly the variant that features the largest number of interface contact perturbations despite the fact that, as described below, its mutations are not directly involved in interface contacts.

In the WT, K353 residue in a β-turn of the ACE2 belongs to a dense interface contact network with the β6-α5 loop of RBD (see main peak in Figure 4B), involving the K353-N501 and K353-Y505 interactions. Notably, in the *Delta* variant, while N501 is conserved, these

two contacts are disrupted and a new interface interaction is established between K353 and Q498. In the other variants, the K353-Y505 contact remains stable, but the K353-N501 interaction (stable in *Epsilon*) becomes slightly stronger in all N501Y variants, as a consequence of the π-cation formation mentioned above. Indeed, as discussed in the static PCN analysis, the K353-Y501 π-cation formation in the *Alpha*, *Beta* and *Gamma* variants is accompanied by that of a T-shaped π-stacking interaction between Y501 and Y41, located at the α1 of ACE2. In contrast, in the *Delta* variant, the Y41-N501 contact is substituted by a stronger Y41-Q498 interaction.

In all models, the α2 helix of ACE2 is in contact with two residues of the RBD β5-β6 loop: F486 and N487. With respect to the WT, the *Alpha* variant features a slight decrease of all contacts in this region, while *Beta* and *Gamma* remains relatively untouched. The *Delta* variant shows again the most disparities: an increase in the M82-F486 and Y83-F486 contacts and a decrease in the Y83-N487 contact are detected. The proximity of these residues with mutation T478K suggests an indirect effect of this mutation (specific of *Delta* variant) on the ACE2/RBD interface. In the *Epsilon*, just a slight increase in the L79-F486 contact is detected, and the rest of contacts remains similar to those of the WT.

The α1 helix of ACE2 is in contact with many secondary structures of the Spike. In particular, contacts with the β5-β6 loop of RBD involves the Y489 residue that features interesting contact perturbations upon mutations at the interface with ACE2 α1. In fact, Y489 strengthens the contact with residue F28 in WT while it establishes a new contact with residue Q24 in the *Delta* variant. In other variants, on the contrary, the Y489-K31 is strengthened as a consequence of the loss of the weak E484-K31 electrostatic interaction found in the WT. Still, at the α1 (nearby K31), D30 establishes a salt bridge with residue K417, another mutation spot. This K417-D30 salt-bridge has been found as a transient contact in MD simulations of *Epsilon* and *Alpha* variants, but this interaction is never observed in the *Beta* and *Gamma* trajectories, featuring the K417N and K417T mutations,

respectively. Surprisingly, in *Delta* and WT, without K417 mutation, this salt-bridge is also broken during the dynamics. While in the available X-ray structures (WT and *Alpha*) the K417-D30 salt-bridge results to be present, our MD simulations suggests that this interaction might be actually weak and prone to rupture.

The *Alpha*-to-*Gamma* dynamics reproduce the main interface perturbation found in all corresponding crystal structures, which is the enhanced interactions between Y501, K353 and Y41. In *Beta* and *Gamma*, the contact loss associated to the K417(N/T)-D30 salt-bridge breaking is also consistent with crystal structures. Interestingly, the WT dynamics shed light about the statistical significance of the K417-D30 interaction, since this salt-bridge features a breaking-formation dynamic even in absence of mutations.

Importantly, Delta mutation spots do not belong to the interface contacts, but they evidently have a significant impact at the interface. More generally, studying systematically the indirect effects of mutations is challenging, especially for comparative studies of mutants, and a more general type of analysis pointing at the most significant contact changes in various systems is required to understand why, for instance, the Delta variant features the largest interface perturbations despite the absence of interface mutations.

Contact principal component analysis

The cPCA is used to characterize the overall information on dynamical contacts resulting from MD simulations of WT and RBD variants into its PCs. In particular, we found 9,432 different contacts in the concatenated trajectories of the WT and five variants (considering the last 400 ns for each system). In Figure S7 in the SI we show that during the last 400 ns of each simulation, PC1 and PC2 values are stable, which shows that our simulations have appropriately converged. As shown in Figure 5A, the scatter plot of the first two PCs shows how cPCA can cluster frames featuring similar dynamical contacts and thus characterize different systems according to that. In contrast to dPCA, here frames are separated in four main clusters: one with the WT and the *Epsilon* variant (negative PC1 and PC2), one

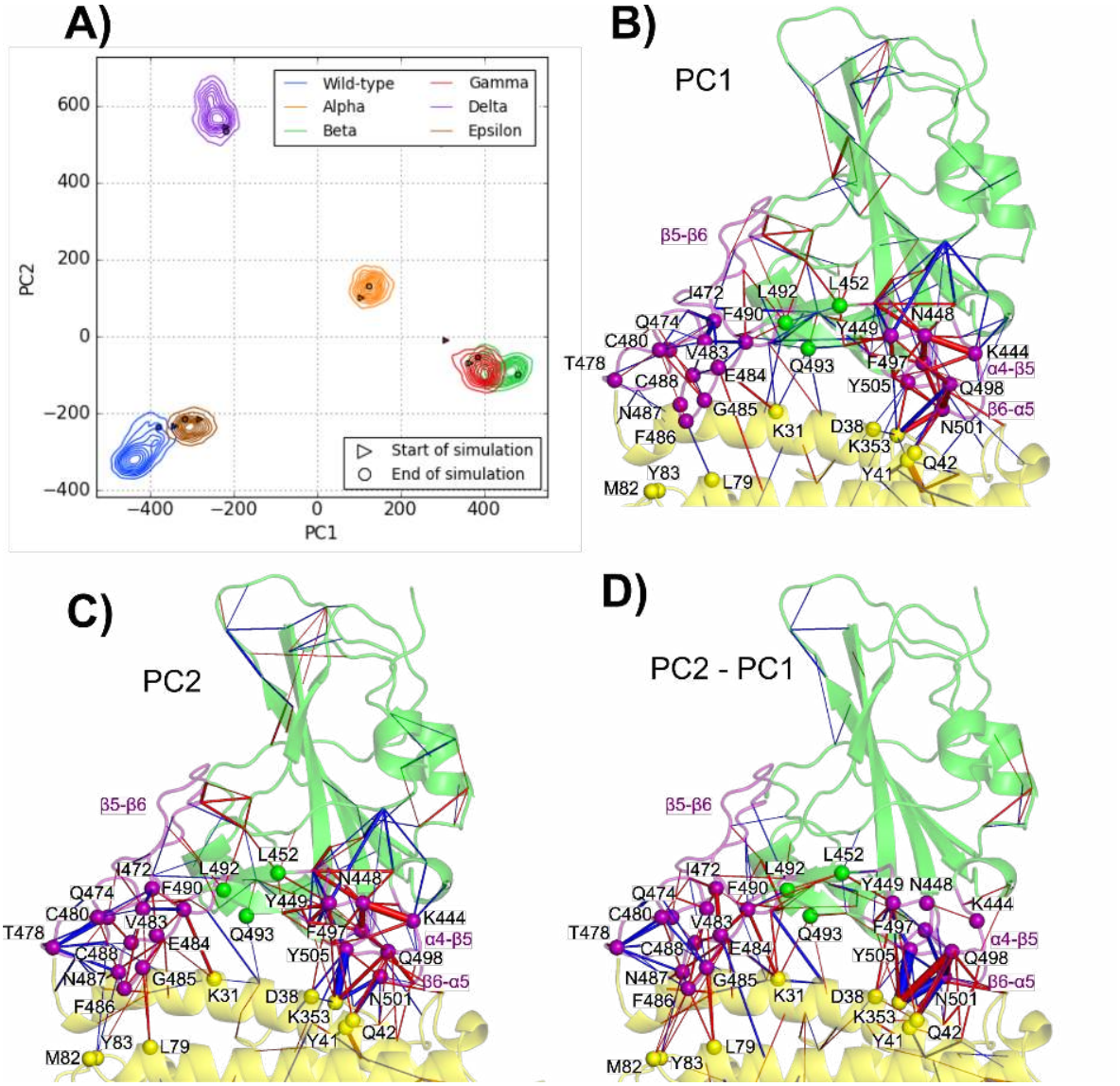


Figure 5: Projection of the frames corresponding to the final 400ns of simulation for the six studied complexes in the two cPCA eigenvector dimensions, with (A) terrain lines representing a kernel density estimate of the population of each complex. Network representation of the influence (as cylinders with a width proportional to the influence) of each contact in the PC1 (B) and PC2 (C) and PC2-PC1 (D) eigenvectors projected on the spike-RBD(green) / ACE2(yellow) WT complex. Blue edges show a negative contribution to the principal component, while red edges show a positive contribution to the principal component. Contacts with a contribution of less than 1% to the eigenvector were discarded.

with the *Delta* variant (positive PC2 and negative PC1), one with the *Alpha* variant (positive PC1 and PC2) and one with the *Beta* and *Gamma* variant (positive PC1 and negative PC2). In this representation, positive values of the PC1 separate *Alpha*, *Beta* and *Gamma* from

WT, *Delta* and *Epsilon*. Positive values of the PC2, instead, discriminate *Alpha* and *Delta* from *Beta*, *Gamma*, *Epsilon* and WT. The following PCs (see Fig. S11-14 in the SI), i.e. those referring to smaller eigenvalues than the two largest ones, are associated with specific separations between systems: the third component separates the WT (negative PC3), the *Epsilon* (positive PC3) from the rest while the fourth one separates *Alpha* (positive PC4) and *Gamma* (negative PC4) from the rest and, finally, the fifth component discriminates between *Alpha* and *Gamma* (negative PC5) from *Beta* (positive PC5) and the rest. Smaller components than PC5 are associated with dynamical contact changes within simulations of each system, e.g. PC6 relates to dynamic contacts occurring in the *Delta* variant. In the dPCA, instead, this kind of clustering associated with each specific system starts with the third principal component. Thus, cPCA provides finer distinctions between the systems under investigation, in terms of dynamical contact changes, with respect to dPCA, especially showing some characteristics of the *Delta* variant.

The representation of PC1 (with positive values for *Alpha*, *Beta* and *Gamma* and negative values for the rest) and PC2 (with positive values for *Delta* and *Alpha* and negative values for the rest) in terms of contact networks near the interface is depicted in Figure 5B-C. Therefore, in order to better differentiate the *Delta* network from the others, also the PC2-PC1 difference is represented in terms of contact network (see Figure 5D), with PC2-PC1 positive values being associated to number of contacts that are large in *Delta* and small in *Beta* and *Gamma* variants, while vice versa for negative values of PC2-PC1 difference (*Alpha*, *Epsilon* and WT contribute only minimally to this network since PC2-PC1 differences are small in these cases). Here, the analysis of the PC2-PC1 differences provides insights into the link between the two *Delta* mutations (T478K and L452R) and their indirect effects at the interface.

Starting from the T478K mutation (exclusive to *Delta*), located in the RBD β5-β6 loop, we found that this residue is a central hub of negative edges in both the PC2 and the PC2-PC1 networks (see Figure 5C-D), involving contacts with residues Q474, C480, F486, N487

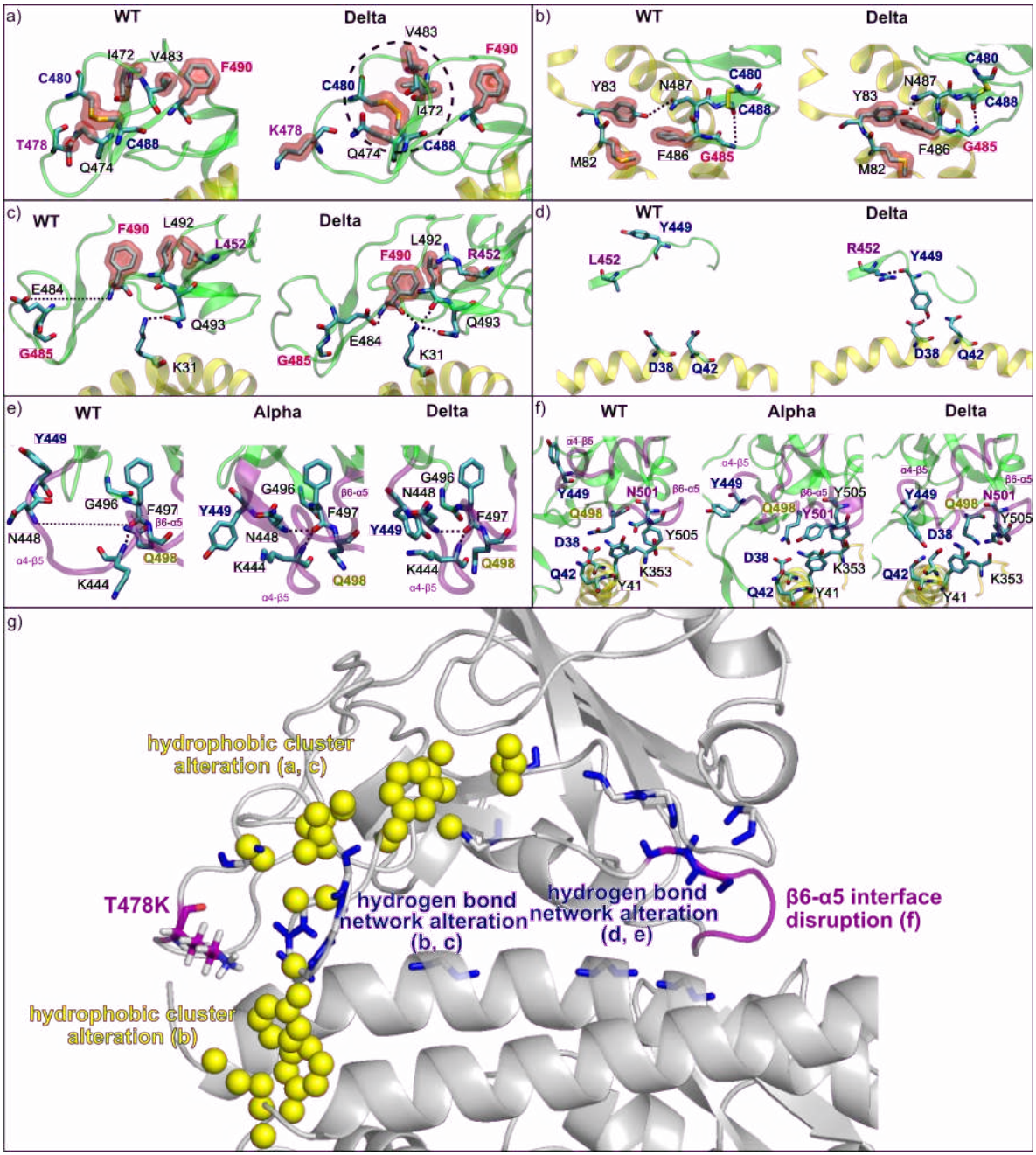


Figure 6: (a to f) Representative MD snapshots of some contacts in the different models emphasized by contact analysis. (g) Summary of the cross-talk between mutated residue T478 and the β6-α5 loop.

and C488. This indicates that few contacts between those residues are characteristic of the *Delta* variant. In fact, as shown in Figure 6A, a hydrophobic cluster is observed nearby the C480-C488 disulfide bridge in the β5-β6 loop of the WT (and also in all other variants but

Delta), involving the hydrophobic moieties of Q474 and T478 and residues I472, V483 and F490. Upon T478K mutation, in the *Delta* variant, the insertion of the lysine side chain does not allow for such arrangement and consequently residue K478 is repelled out of the cluster. This loss of interaction in *Delta* is associated with flipping of the C480-C488 bridge that in turns push residue F490 far from the cluster. As a consequence of this rearrangement of the hydrophobic cluster, a backbone G485(NH)-C488(O) hydrogen bond is stabilized in *Delta*, determining a better folding of the β5-β6 loop, as depicted in Fig. 6b. This differently folded structure also inevitably affects the dynamics of residues F486 and N487, which were previously highlighted in the DPCN of *Delta* at the RBD/ACE2 interface. These residues, indeed, show more contacts with M82 and Y83 (located in the ACE2 α2 helix) in the *Delta* variant than in the WT. This proves that the T478K mutation is indirectly responsible for the contact increase between the Spike-RBD and the α2 helix of ACE2.

The change in folding of the β5-β6 loop induced by the T478K mutation *Delta* has, moreover, other indirect effects on the RBD/ACE2 interface that are synergetic with the effects of the L452R mutation. In fact, as shown in the PC2-PC1 network in Fig. 6D, the negative edges around the T478K mutation in the RBD β5-β6 loop are somehow compensated by the positive edges around residue F490, i.e. the residue repelled out of the hydrophobic cluster in the *Delta* variant. This set of predominantly positive edges involves residues E484 (neighbor of G485), L492, L452 (mutated to R in *Delta*) and K31 across the interface. Indeed, the perturbations from the T478K appear to be connected to those induced by the L452R mutation through residue F490 (in the β5-β6-loop) and L492 (in the β6-sheet). Figure 6c shows that the contemporary T478K and L452R mutations in *Delta* have a significant effect on the hydrogen bonding network around the interface residue K31. In particular, the dynamics of residue F490 is synergistically affected by the two mutations from two different sides: on one side the change in folding of the β5-β6 loop upon T478K mutation stabilizes the E484-F490 hydrogen bond while, on the other side, since F490 is also in hydrophobic contact with L492 and L452, upon L452R mutation, the arginine sidechain promotes hydrogen bonding inter-

actions of the L492 and F490 backbones with the K31 sidechain. This finally results into three hydrogen bonds between the NH3+ head of K31 and the sidechain oxygen of Q493 and the backbone oxygens of L492 and F490, which is a characteristic interface arrangement of the *Delta* variant (i.e. in the WT only Q493 is hydrogen-bonded with K31) and it results from the combination of the two T478K and L452R mutations (far from the interface). Here, we should note that the Omicron variant also possess the T478K mutations (same as *Delta*) but in conjunction with the E484A mutation. This latter mutation is somehow surprising since the E484K mutation is very common in Spike's mutants (e.g. it is found in *Beta*, *Gamma*, *Mu*, *Lambda*, *Eta*, *Theta*) while E484A is exclusive to Omicron. This open to the question of how much the E484A mutation in *Omicron* could influence the effects of the *Delta* T478K mutations, which should be addressed in further studies. Notably, in the *Beta* variant, a cross-talk between mutated residue K484 and Y501 has been discovered.[27] The present results suggest that in the *Delta* variant, there is also an allosteric cross-talk between mutated residues K478 and the β6-α5 region (in which N501 is found). There is a possibility that the two cross-talks are incompatible with each other. It is worth mentioning that the E484K mutation, present in *Beta* and *Gamma* but not in *Alpha*, differentiates these variants in terms of the interface contacts between the ACE2 receptor and the RBD β5-β6 loop, involving the network of contacts highlighted in this region by the PC1 and PC2 components.

In the PC2-PC1 contact network (see Fig. 6D), residue L452(R) is a bridging node that connects the contact perturbations in the β5-β6 loop (described above and involving the T478K mutation) with those of the α4-β5 and β5-β6 loops. Residue L452(R) has a positive edge with residue Y449 in this network, meaning that a close L452(R)-Y449 contact is typical of the *Delta* variant. In turn, Y449 displays direct connections with interface residue, featuring positive edges with D38 and Q42 in the ACE2 receptor. Figure 6d shows that, indeed, upon L452R mutation, the arginine sidechain is able to make a hydrogen bond with Y449(O), which promotes a flipping of Y449 sidechain, allowing for the formation of a Y449-D38 in-

terface hydrogen bond that alters the surrounding H-bonding network, involving also Q42. Notably, the perturbations around residue Y449 in the PC2-PC1 network are minimal, as a consequence of the fact that perturbations inside the spike RBD (i.e. in the α4-β5 and β6-α5 loops) are rather similar in the PC1 and PC2 networks (see Fig. 5B,C). In particular, the K444-N448, N448-F497 pairs features numerous contacts in PC1 and PC2, but they virtually vanish in the PC2-PC1 network, indicating that rearrangements of contacts in this region are significant in all variants but somehow differ from *Epsilon* that is more similar to WT, in line with the DPCN results depicted in Figure 3. At the same time, the largest PC2-PC1 differences are found at the interface between these spike RBD loops and the ACE2 receptor. Here, in the DPCN interface analysis we highlighted the role of residues Q498, N501, Y505 in the β6-α5 loop in contact with Y41 in the α1 helix and K353 in the β-turn. Figure 6e shows how, in the *Delta* variant, upon the Y449 flipping mentioned above, the backbone N448(NH)-F497(O) hydrogen bond adds up to the preexisting K444(NH)-F497(O) one. Very interestingly, the very same two hydrogen bonds are also formed in the *Alpha* variant, featuring the sole N501Y mutation. This indicated that such a single mutation in the *Alpha* RBD creates a H-bonding network in the α4-β5 and β6-α5 loops of RBD similar to that produced by the indirect effects of the L452R mutation in the *Delta* variant (via residue Y449). Notably, these contact changes at the RBD common to both L452R and N501Y mutations in *Delta* and *Alpha*, respectively, have not an effect on the interface contacts. In fact, as shown in Figure 6f, the WT and *Alpha* interfaces involve interactions between the same residues (i.e. D38, Y41, Q42, K353, Q498, N501(Y), Y505) despite the presence of the N501Y mutation, which only changes the type of some interactions (most notably the Y41-N501 π-polar interaction is promoted to a Y41-Y501 π-π interaction). On the other hand, the interface in the *Delta* variant largely differs from those of the WT and *Alpha* since the involved residues now include Y449 instead of N501 and Y505. Interestingly, this shows how the indirect effect of *Delta* L452R mutation on the interface contacts, via the Y449 residue and the Y449-D38 interaction (see Figure 6d), has a large impact on the

RBD/ACE2 interface as previously mentioned in the DPCN analysis, see Fig. 4A. As a result of the L452R mutation in *Delta* variant, thus, the formation of the R452-Y449 interaction is associated to structural rearrangements of the α4-β5 and β6-α5 loops that modify the interface contacts by including the Y449-D38 hydrogen bond and substituting the N501-Y41 interaction with the Q498-Y41 hydrogen bond, pushing away residues N501 and Y505 from the interface (breaking their contacts with residue K353). As evident from Figure 5D, in fact, these interface changes are the most prominent in the PC2-PC1 network and represent the long distance effects of the L452R mutation on the RBD/ACE2 interface.

Conclusions

In this study, we first analyzed the (static) networks of atomic contacts between the Spike RBD protein and the ACE2 human receptor based on the available crystallographic structures of the *Alpha*-to-*Gamma* variants of SARS-CoV-2, capturing the contact changes with respect to the WT and thus perturbations due to RBD mutations. Then, in order to account for dynamical effects of RBD mutations on Spike/ACE2 interface contacts, microsecond MD simulations have been performed on the WT and the *Alpha*-to-*Epsilon* variants. Various tools for MD trajectories analysis have been used to recover the main similarities and differences between various Spike RBD variants interacting with the human ACE2 receptor.

First, the analysis of protein essential motions based on backbone dihedral angles, namely dPCA, allowed recognizing mobile RBD regions whose dynamics is altered by mutations. The first principal components of backbone dihedral angles is associated with motions in the α4-β5 loop, while the second principal components is associated with motions in the β5-β6 loop. Considering these essential motions, three distinct behavior have been observed for the various MD simulations: on one side, a cluster involving the *Alpha*, *Beta*, *Gamma* and *Delta* variants features a tight α4-β5 loop and a flexible β5-β6 loop; on the other hand, the WT features a flexible α4-β5 loop and a tight β5-β6 loop; while for the *Epsilon* variant, the

tightest β5-β6 loop was observed along with a partially flexible α4-β5 loop. Interestingly, this clustering correlates with the impact in transmissibility and severity of the SARS-CoV-2 disease in the studied variants. These results suggest that the L452R and N501Y mutation have closely related effects on RBD motions near the interface. However, as evidenced by the dPCA of the *Epsilon* variant, these motions are not fully reproduced in the absence of the T478K mutation, which indicates an interdependence between these mutations. In fact, this change in flexibility of the RBD near the interface may be a first step facilitating the Spike trimer binding to than one ACE2 receptor. Still, the dPCA analysis did not allow differentiating the *Delta* variant, the dominating one in most of the 2021 year, from the others.

Then, we were able to recover some specificity of the *Delta* variant by studying the dynamical perturbation contact network, with a focus on the RBD/ACE2 interface. The comparisons between WT atomic contact network with those of *Alpha*-to-*Epsilon* variants showed many similarities among the *Alpha*-to-*Gamma* variants that share the N501(Y) mutation, which promotes specific perturbations for the interface contacts of Y501 with K353 and Y41 residues, while the rest of interface contacts remains essentially preserved. By contrast, in the *Delta* variant, significant contact changes at the interface have been found despite the absence of interface mutations. Indeed, all interface contact changes in *Delta* cannot be directly attributed to the T478K and L452R mutations that must have indirect (but large) effects on the interface.

The subsequent cPCA analysis shed, finally, light on the propagation of contact perturbations induced by the T478K and L452R mutations in the *Delta* variant. This analysis showed that the T478K mutation alters the contacts of a hydrophobic cluster (involving residues Q474, T478, I472, V483 and F490) around the C480-C488 disulfide bridge inside the β5-β6 loop of the RBD and promotes the formation of a G485-C488 backbone hydrogen bond. In turn, this rearrangement affects the position of residue F486 and N487 that increase their interface contacts with the α2 helix of ACE2. At the same time, in the WT residue F490, L492

and L452 are involved in another hydrophobic cluster that upon L452R mutations adjusted because of both the presence of residue R452 and the alteration of F490 contacts due to the T478K mutation. In turn, residue F490 and L492 create a triple hydrogen bond at the interface of *Delta* with residue K31, which was H-bonded just to residue Q493 in the WT. Since it belongs to both the aforementioned T478- and L452-related hydrophobic clusters, residue F490 resulted to be central for the propagation of contacts changes due to the simultaneous T478K and L452R mutations that result to cooperate in inducing the interface perturbations found in *Delta*.

Our results highlight the singular mechanism of action of the mutations in the *Delta* variant that could eventually explain why it dominated over preceding variants. Moreover, since the recent *Omicron* variant possess the same T478K mutation but in conjunction with the E484A one, it remains to elucidate if a synergistic long-range effect of multiple mutations like that found here for the *Delta* variant is also operating for the currently dominating *Omicron*.

Data and Software Availability

Software to compute DPCN, dPCA and cPCA are available at the following github repository:

`https://github.com/agheeraert/pmdlearn`.

Molecular Dynamics simulations of all RBD/ACE2 complexes are available upon request.

Acknowledgement

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AP010712548 made by GENCI.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

AG and LV thank the support of CNRS the 80prime and MITI programs and acknowledge the support of the Institut Rhônalpin des systèmes complexes, IXXI-ENS-Lyon, Lyon, France.
AG and acknowledge the support of the Federation of European Biochemical Societies for its Short-Term Fellowship and the use of the "Pôle Scientifique de Modélisation Numérique" (PSMN) at the École Normale Supérieure de Lyon, France.
AG thanks Federica Maschietto for fruitful discussions about Ward's minimum variance method.

Supporting Information Available

RMSD fluctuations, dPCA and cPCA on the complete trajectory, time evolution of key distances and dihedral angles, cartesian coordinates PCA, cPCA components up to 8, time evolution of interface loops flexibility.

References

(1) Wang, Q.; Zhang, Y.; Wu, L.; Niu, S.; Song, C.; Zhang, Z.; Lu, G.; Qiao, C.; Hu, Y.; Yuen, K.-Y.; Wang, Q.; Zhou, H.; Yan, J.; Qi, J. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **2020**, *181*, 894–904.

(2) Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **2020**, *181*, 281–292.

(3) Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; Wang, X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. **2020**, *581*, 215–220.

(4) Yan, R.; Zhang, Y.; Li, Y.; Xia, L.; Guo, Y.; Zhou, Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **2020**, *367*, 1444–1448.

(5) Castillo, A. E.; Parra, B.; Tapia, P.; Acevedo, A.; Lagos, J.; Andrade, W.; Arata, L.; Leal, G.; Barra, G.; Tambley, C.; Tognarelli, J.; Bustos, P.; Ulloa, S.; Fasce, R.; Fernández, J. Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J. Med. Virol.* **2020**, *92*, 1562–1566.

(6) Chen, J.; Wang, R.; Wang, M.; Wei, G.-W. Mutations strengthened SARS-CoV-2 infectivity. *J. Mol. Biol.* **2020**, *432*, 5212–5226.

(7) Shang, J.; Ye, G.; Shi, K.; Wan, Y.; Luo, C.; Aihara, H.; Geng, Q.; Auerbach, A.; Li, F. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **2020**, *581*, 221–224.

(8) Davies, N. G.; Abbott, S.; Barnard, R. C.; Jarvis, C. I.; Kucharski, A. J.; Munday, J. D.; Pearson, C. A. B.; Russell, T. W.; Tully, D. C.; Washburne, A. D.; Wenseleers, T.; Gimma, A.; Waites, W.; Wong, K. L. M.; van Zandvoort, K.; Silverman, J. D.; CMMID COVID-19 Working Group,; COVID-19 Genomics UK (COG-UK) Consortium,; Diaz-Ordaz, K.; Keogh, R.; Eggo, R. M.; Funk, S.; Jit, M.; Atkins, K. E.; Edmunds, W. J. Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. *Science* **2021**, *372*.

(9) Funk, T.; Pharris, A.; Spiteri, G.; Bundle, N.; Melidou, A.; Carr, M.; Gonzalez, G.; Garcia-Leon, A.; Crispie, F.; O'Connor, L.; Murphy, N.; Mossong, J.; Vergison, A.; Wienecke-Baldacchino, A. K.; Abdelrahman, T.; Riccardo, F.; Stefanelli, P.; Martino, A. D.; Bella, A.; Presti, A. L.; Casaca, P.; Moreno, J.; Borges, V.; Isidro, J.; Ferreira, R.; Gomes, J. P.; Dotsenko, L.; Suija, H.; Epstein, J.; Sadikova, O.; Sepp, H.; Ikonen, N.; Savolainen-Kopra, C.; Blomqvist, S.; Möttönen, T.; Helve, O.; Gomes-Dias, J.; Adlhoch, C.; Groups, o. b. o. C. s. Characteristics of SARS-CoV-2 variants of concern B. 1.1. 7, B. 1.351 or P. 1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. **2021**, *26*, 2100348.

(10) Davies, N. G.; Jarvis, C. I.; Edmunds, W. J.; Jewell, N. P.; Diaz-Ordaz, K.; Keogh, R. H.

Increased mortality in community-tested cases of SARS-CoV-2 lineage B. 1.1. 7. *Nature* **2021**, *593*, 270–274.

(11) Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E. J.; Msomi, N.; Mlisana, K.; von Gottberg, A.; Walaza, S.; Allam, M.; Ismail, A.; Mohale, T.; Glass, A. J.; Engelbrecht, S.; Van Zyl, G.; Preiser, W.; Petruccione, F.; Sigal, A.; Hardie, D.; Marais, G.; Hsiao, N.-y.; Korsman, S.; Davies, M.-A.; Tyers, L.; Mudau, I.; York, D.; Maslo, C.; Goedhals, D.; Abrahams, S.; Laguda-Akingba, O.; Alisoltani-Dehkordi, A.; Godzik, A.; Wibmer, C. K.; Sewell, B. T.; Lourenço, J.; Alcantara, L. C. J.; Kosakovsky Pond, S. L.; Weaver, S.; Martin, D.; Lessells, R. J.; Bhiman, J. N.; Williamson, C.; de Oliveira, T. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **2021**, *592*, 438–443.

(12) Cele, S.; Gazy, I.; Jackson, L.; Hwa, S.-H.; Tegally, H.; Lustig, G.; Giandhari, J.; Pillay, S.; Wilkinson, E.; Naidoo, Y.; Karim, F.; Ganga, Y.; Khan, K.; Bernstein, M.; Balazs, A. B.; Gosnell, B. I.; Hanekom, W.; Moosa, M.-Y. S.; Lessells, R. J.; de Oliveira, T.; Sigal, A. Escape of SARS-CoV-2 501Y. V2 from neutralization by convalescent plasma. *Nature* **2021**, *593*, 142–146.

(13) Madhi, S. A.; Baillie, V.; Cutland, C. L.; Voysey, M.; Koen, A. L.; Fairlie, L.; Padayachee, S. D.; Dheda, K.; Barnabas, S. L.; Bhorat, Q. E.; Briner, C.; Kwatra, G.; Ahmed, K.; Aley, P.; Bhikha, S.; Bhiman, J. N.; Bhorat, A. E.; du Plessis, J.; Esmail, A.; Groenewald, M.; Horne, E.; Hwa, S.-H.; Jose, A.; Lambe, T.; Laubscher, M.; Malahleha, M.; Masenya, M.; Masilela, M.; McKenzie, S.; Molapo, K.; Moultrie, A.; Oelofse, S.; Patel, F.; Pillay, S.; Rhead, S.; Rodel, H.; Rossouw, L.; Taoushanis, C.; Tegally, H.; Thombrayil, A.; van Eck, S.; Wibmer, C. K.; Durham, N. M.; Kelly, E. J.; Villafana, T. L.; Gilbert, S.; Pollard, A. J.; de Oliveira, T.; Moore, P. L.; Sigal, A.; Izu, A. Efficacy of the ChAdOx1 nCoV-19 Covid-19 vaccine against the B. 1.351 variant. *N. Engl. J. Med.* **2021**, *384*, 1885–1898.

(14) Faria, N. R.; Mellan, T. A.; Whittaker, C.; Claro, I. M.; Candido, D. d. S.; Mishra, S.; Crispim, M. A. E.; Sales, F. C. S.; Hawryluk, I.; McCrone, J. T.; Hulswit, R. J. G.; Franco, L. A. M.; Ramundo, M. S.; de Jesus, J. G.; Andrade, P. S.; Coletti, T. M.; Ferreira, G. M.; Silva, C. A. M.; Manuli, E. R.; Pereira, R. H. M.; Peixoto, P. S.; Kraemer, M. U. G.; Gaburo, N.; Camilo, C. d. C.; Hoeltgebaum, H.; Souza, W. M.; Rocha, E. C.; de Souza, L. M.; de Pinho, M. C.; Araujo, L. J. T.; Malta, F. S. V.; de Lima, A. B.; Silva, J. d. P.; Zauli, D. A. G.; Ferreira, A. C. d. S.; Schnekenberg, R. P.; Laydon, D. J.; Walker, P. G. T.; Schlüter, H. M.; dos Santos, A. L. P.; Vidal, M. S.; Del Caro, V. S.; Filho, R. M. F.; dos Santos, H. M.; Aguiar, R. S.; Proença-Modena, J. L.; Nelson, B.; Hay, J. A.; Monod, M.; Miscouridou, X.; Coupland, H.; Sonabend, R.; Vollmer, M.; Gandy, A.; Prete, C. A.; Nascimento, V. H.; Suchard, M. A.; Bowden, T. A.; Pond, S. L. K.; Wu, C.-H.; Ratmann, O.; Ferguson, N. M.; Dye, C.; Loman, N. J.; Lemey, P.; Rambaut, A.; Fraiji, N. A.; Carvalho, M. d. P. S. S.; Pybus, O. G.; Flaxman, S.; Bhatt, S.; Sabino, E. C. Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **2021**, *372*, 815–821.

(15) Dejnirattisai, W.; Zhou, D.; Supasa, P.; Liu, C.; Mentzer, A. J.; Ginn, H. M.; Zhao, Y.; Duyvesteyn, H. M. E.; Tuekprakhon, A.; Nutalai, R.; Wang, B.; López-Camacho, C.; Slon-Campos, J.; Walter, T. S.; Skelly, D.; Costa Clemens, S. A.; Naveca, F. G.; Nascimento, V.; Nascimento, F.; Fernandes da Costa, C.; Resende, P. C.; Pauvolid-Correa, A.; Siqueira, M. M.; Dold, C.; Levin, R.; Dong, T.; Pollard, A. J.; Knight, J. C.; Crook, D.; Lambe, T.; Clutterbuck, E.; Bibi, S.; Flaxman, A.; Bittaye, M.; Belij-Rammerstorfer, S.; Gilbert, S. C.; Carroll, M. W.; Klenerman, P.; Barnes, E.; Dunachie, S. J.; Paterson, N. G.; Williams, M. A.; Hall, D. R.; Hulswit, R. J. G.; Bowden, T. A.; Fry, E. E.; Mongkolsapaya, J.; Ren, J.; Stuart, D. I.; Screaton, G. R. Antibody evasion by the P. 1 strain of SARS-CoV-2. *Cell* **2021**, *184*, 2939–2954.

(16) Sheikh, A.; McMenamin, J.; Taylor, B.; Robertson, C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **2021**,

(17) Deng, X.; Garcia-Knight, M. A.; Khalid, M. M.; Servellita, V.; Wang, C.; Morris, M. K.; Sotomayor-González, A.; Glasner, D. R.; Reyes, K. R.; Gliwa, A. S.; Reddy, N. P.; Martin, C. S. S.; Federman, S.; Cheng, J.; Balcerek, J.; Taylor, J.; Streithorst, J. A.; Miller, S.; Kumar, G. R.; Sreekumar, B.; Chen, P.-Y.; Schulze-Gahmen, U.; Taha, T. Y.; Hayashi, J.; Simoneau, C. R.; McMahon, S.; Lidsky, P. V.; Xiao, Y.; Hemarajata, P.; Green, N. M.; Espinosa, A.; Kath, C.; Haw, M.; Bell, J.; Hacker, J. K.; Hanson, C.; Wadford, D. A.; Anaya, C.; Ferguson, D.; Lareau, L. F.; Frankino, P. A.; Shivram, H.; Wyman, S. K.; Ott, M.; Andino, R.; Chiu, C. Y. Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *MedRxiv* **2021**,

(18) Han, P.; Su, C.; Zhang, Y.; Bai, C.; Zheng, A.; Qiao, C.; Wang, Q.; Niu, S.; Chen, Q.; Zhang, Y.; Li, W.; Liao, H.; Li, J.; Zhang, Z.; Cho, H.; Yang, M.; Rong, X.; Hu, Y.; Huang, N.; Yan, J.; Wang, Q.; Zhao, X.; Gao, G. F.; Qi, J. Molecular insights into receptor binding of recent emerging SARS-CoV-2 variants. **2021**, *12*, 1–9.

(19) Mannar, D.; Saville, J. W.; Zhu, X.; Srivastava, S. S.; Berezuk, A. M.; Zhou, S.; Tuttle, K. S.; Kim, A.; Li, W.; Dimitrov, D. S.; Subramaniam, S. Structural analysis of receptor binding domain mutations in SARS-CoV-2 variants of concern that modulate ACE2 and antibody binding. *Cell Rep.* **2021**, *37*, 110156.

(20) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Cent. Sci.* **2020**, *6*, 1722–1734.

(21) Arantes, P. R.; Saha, A.; Palermo, G. Fighting COVID-19 using molecular dynamics simulations. 2020.

(22) Munafò, F.; Donati, E.; Brindani, N.; Ottonello, G.; Armirotti, A.; De Vivo, M. Quercetin and Luteolin Are Single-digit Micromolar Inhibitors of the SARS-CoV-2 RNA-dependent RNA Polymerase. **2021**,

(23) Deganutti, G.; Prischi, F.; Reynolds, C. A. Supervised molecular dynamics for exploring the druggability of the SARS-CoV-2 spike protein. *Journal of computer-aided molecular design* **2021**, *35*, 195–207.

(24) Khoury, L. E.; Jing, Z.; Cuzzolin, A.; Deplano, A.; Loco, D.; Sattarov, B.; Hédin, F.; Wendeborn, S.; Ho, C.; Ahdab, D. E.; Inizan, T. J.; Sturlese, M.; Sosic, A.; Volpiana, M.; Lugato, A.; Barone, M.; Gatto, B.; Ludovica Macchia, M.; Bellanda, M.; Battistutta, R.; Salata, C.; Kondratov, I.; Iminov, R.; Khairulin, A.; Mykhalonok, Y.; Pochepko, A.; Chashka-Ratushnyi, V.; Kos, I.; Moro, S.; Montes, M.; Ren, P.; W. Ponder, J.; Lagardère, L.; Piquemal, J.-P.; Sabbadin, D. Computationally driven discovery of SARS-CoV-2 Mpro inhibitors: from design to experimental validation. *Chemical Science* **2022**,

(25) Sztain, T.; Ahn, S.-H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; Seitz, E.; McCool, R. S.; Kearns, F. L.; Acosta-Reyes, F.; Maji, S.; Mashayekhi, G.; McCammon, J. A.; Ourmazd, A.; Frank, J.; McLellan, J. S.; Chong, L. T.; Amaro, R. E. A Glycan Gate Controls Opening of the SARS-CoV-2 Spike Protein. *Nat. Chem.* **2021**, *13*, 963–968.

(26) Triveri, A.; Serapian, S. A.; Marchetti, F.; Doria, F.; Pavoni, S.; Cinquini, F.; Moroni, E.; Rasola, A.; Frigerio, F.; Colombo, G. SARS-CoV-2 Spike Protein Mutations and Escape from Antibodies: A Computational Model of Epitope Loss in Variants of Concern. *J. Chem. Inf. Model.* **2021**, *61*, 4687–4700.

(27) Spinello, A.; Saltalamacchia, A.; Borišek, J.; Magistrato, A. Allosteric Cross-Talk among Spike's Receptor-Binding Domain Mutations of the SARS-CoV-2 South African Variant Triggers an Effective Hijacking of Human Cell Receptor. *J. Phys. Chem. Lett.* **2021**, *12*, 5987–5993.

(28) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 45–52.

(29) Ernst, M.; Sittel, F.; Stock, G. Contact-and distance-based principal component analysis of protein dynamics. *The Journal of chemical physics* **2015**, *143*, 12B640_1.

(30) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(31) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130.

(32) Huang, J.; MacKerell Jr, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **2013**, *34*, 2135–2145.

(33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(34) Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* **1992**, *97*, 1990–2001.

(35) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(36) Kitao, A.; Hirata, F.; Gō, N. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* **1991**, *158*, 447–472.

(37) Hayward, S.; Go, N. Collective variable description of native protein dynamics. *Annu. Rev. Phys. Chem.* **1995**, *46*, 223–250.

(38) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. Principal component analysis and long time protein dynamics. *J. Phys. Chem.* **1996**, *100*, 2567–2572.

(39) Kitao, A.; Hayward, S.; Go, N. Energy landscape of a native protein: Jumping-among-minima model. *Proteins: Struct., Funct., Bioinf.* **1998**, *33*, 496–517.

(40) Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **2000**, *62*, 8438.

(41) Hess, B. Convergence of sampling in protein simulations. *Phys. Rev. E* **2002**, *65*, 031910.

(42) Tournier, A. L.; Smith, J. C. Principal components of the protein dynamical transition. *Phys. Rev. Lett.* **2003**, *91*, 208106.

(43) Lange, O. F.; Grubmüller, H. Full correlation analysis of conformational protein dynamics. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1294–1312.

(44) David, C. C.; Jacobs, D. J. *Protein dynamics*; Springer, 2014; pp 193–226.

(45) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **2008**, *128*, 06B620.

(46) Sittel, F.; Jain, A.; Stock, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *J. Chem. Phys.* **2014**, *141*, 07B605_1.

(47) Jain, A.; Stock, G. Hierarchical folding free energy landscape of HP35 revealed by most probable path clustering. *J. Phys. Chem. B* **2014**, *118*, 7750–7760.

(48) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(49) Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.

(50) Davis, R. A.; Lii, K.-S.; Politis, D. N. *Selected Works of Murray Rosenblatt*; Springer, 2011; pp 95–100.

(51) Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

(52) Vuillon, L.; Lesieur, C. From local to global changes in proteins: a network view. *Curr. Opin. Struct. Biol.* **2015**, *31*, 1–8.

(53) Dorantes-Gilardi, R.; Bourgeat, L.; Pacini, L.; Vuillon, L.; Lesieur, C. In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: not too many links, not too few. *Phys. Chem. Chem. Phys.* **2018**, *20*, 25399–25410.

(54) Gheeraert, A.; Pacini, L.; Batista, V. S.; Vuillon, L.; Lesieur, C.; Rivalta, I. Exploring allosteric pathways of a v-type enzyme with dynamical perturbation networks. *J. Phys. Chem. B* **2019**, *123*, 3452–3461.

(55) Spinello, A.; Saltalamacchia, A.; Magistrato, A. Is the rigidity of SARS-CoV-2 spike

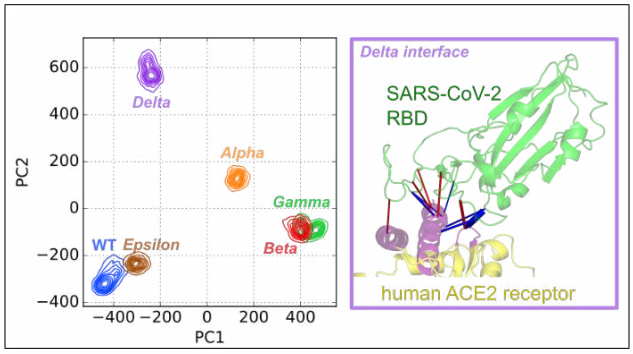
receptor-binding motif the hallmark for its enhanced infectivity? Insights from all-atom simulations. *The journal of physical chemistry letters* **2020**, *11*, 4785–4790.

(56) Barros, E. P.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Wang, Y.; Fallon, L.; Raguette, L.; Belfon, K.; Simmerling, C.; Amaro, R. E. The Flexibility of ACE2 in the Context of SARS-CoV-2 Infection. *Biophysical Journal* **2021**, *120*, 1072–1084.

(57) Fiorillo, B.; Marchianò, S.; Moraca, F.; Sepe, V.; Carino, A.; Rapacciuolo, P.; Biagioli, M.; Limongelli, V.; Zampella, A.; Catalanotti, B.; Fiorucci, S. Discovery of Bile Acid Derivatives as Potent ACE2 Activators by Virtual Screening and Essential Dynamics. *J. Chem. Inf. Model.* **2022**, *62*, 196–209.

(58) Han, P.; Li, L.; Liu, S.; Wang, Q.; Zhang, D.; Xu, Z.; Han, P.; Li, X.; Peng, Q.; Su, C.; Huang, B.; Li, D.; Zhang, R.; Tian, M.; Fu, L.; Gao, Y.; Zhao, X.; Liu, K.; Qi, J.; Gao, G. F.; Wang, P. Receptor Binding and Complex Structures of Human ACE2 to Spike RBD from Omicron and Delta SARS-CoV-2. *Cell* **2022**, *185*, 630–640.e10.

TOC Graphic


Delta
Alpha
Gamma
Beta
WT
Epsilon
PC2
PC1
600
400
200
0
−200
−400
−400 −200 0 200 400
Delta interface
SARS-CoV-2
RBD
human ACE2 receptor

Chapter 4

Conclusions

In this thesis, we developed various analytical tools aimed at the study of MD simulations. First, we developed a novel kind of amino acid network: the average contact network. This network, built from a purely geometrical perspective from the dynamics of a simulation can capture the fundamental contacts in a proteic system. When subtracted, the average contact networks produces a dynamical perturbation contact Network that emphasizes the biggest contact changes between two different systems. Then, to facilitate the analysis of DPCNs, we developed a connected component analysis to facilitate that discriminates among relevant edges. This approach was particularly useful at emphasizing local patches of perturbations that spread within a protein. To overcome intrinsic limitations of DPCNs, we developed a contact Principal Component Analysis, which can directly point at the principal axes of variation in a set containing numerous systems with numerous replicas, stressing which structures are the most different from others and why they differ. This analysis has a new ability to detect contact changes within a simulation and also can detect if the DPCN built between two trajectories is relevant or not. Finally, we developed a way to investigate proteic systems using smaller coarse-grains: the CGNs, which allows capturing when contact change of type between residue and overall, the chemistry of a contact.

The development of the average contact network and DPCN proved very successful in studying the allosteric pathways in IGPS. Later we successfully applied this tool to the study of another allosteric system (AMPK) and the temperature dependence effect of the allostery in IGPS. We finally used this tool, originally designed to study the effects of single point mutations in crystal structure and then generalized to MD simulations, to study the effects of mutations in the SARS-CoV-2 variants. In general, thus, the DPCN is a tool that can greatly facilitate the analysis of differences between a set of *reference* simulations and of *perturbed* simulations. The cPCA established itself as a complementary tool to the DPCN facilitate the investigations in the case of studying differences between the five SARS-CoV-2 variants and the WT. Furthermore, this tool also proved that it could point at differences within a simulation which additionally generalize it. The development of connected component analysis and machine learning contact analysis remains preliminary, but results based on the IGPS protein are very promising.

This average contact network is built only using a geometrical analysis of MD simulations but is able to discriminate contacts of different magnitudes in comparison with the frequency contact network[1, 2]. Interestingly, the same difference is found in contact principal component analysis of MD simulations[3]. Our work showed that taking into account the magnitude of contacts is an important aspect in AANs and can refine analysis of AAN. Some groups have already shown interest in the use of this new kind of AAN[4] and we hope that many groups will follow. While the initial goal was restricted to the study of allostery, in the end, this tool is completely general and can even be used basically to control the convergence of a simulation and to extract the relaxation signal from a simulation.

By introducing a variability in contact magnitude, different contacts are weighted differently. While this aspect possess numerous advantages and notably emphasize the biggest contact losses or gain, there is a reason to suspect that different interactions are weighted differently. Our methods to build contact networks, contact matrices, or in general to featurize contacts are purely geometric. Of course, the geometry of a contact is highly connected to the chemistry but the precise link between those two aspects remains elusive. Some scaling techniques have been proposed and introduced but the fix they provide does not bridge the gap in knowledge between geometry and chemistry. Another important limitation is the fact that our way of building AAN still depends on many parameter. One argument to take a cutoff parameter of 5 Å is that it represents the limit of Van der Waals interactions. While not all contacts are considered equal, all interatomic contacts are considered equal even if they are at the limit of the sphere of Van der Waals interactions. Some of our tests have shown that in some contexts, using a cut-off of 3.5 Å sometimes provides a better way to show contact losses and breaking because this way, the strongest contacts break and loss are particularly emphasized. In fact, contacts at the edge of being described are the most feeble and their description may provide unnecessary data.

There are countless projects which can be started from this work. One aspect we explored was building different kinds of AAN using different parameterless techniques, such a using Delaunay or Voronoi tessellations or a k-nearest neighbor approach. This early stage projects are promising, but a complete and concise implementation is still missing. Another very interesting aspect is to generalize the contact study to different interacting bodies. During these works, the idea to investigate molecule-protein, water-protein, sugar-protein and membrane-protein interactions have been proposed. In practice, our tools do not even need to be extensively modified for such analysis. The study of solvent-protein interaction notably is probably essential to fully understand the link between geometry and chemistry as some contacts are known to be water-mediated.

References

[1] Urmi Doshi et al. "Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation". In: *Proc. National Acad. Sci.* 113.17 (2016), pp. 4735–4740.

[2] Xin-Qiu Yao, Mohamed Momin, and Donald Hamelberg. "Elucidating allosteric communications in proteins with difference contact network analysis". In: *J. chemical information modeling* 58.7 (2018), pp. 1325–1330.

[3] Matthias Ernst, Florian Sittel, and Gerhard Stock. "Contact-and distance-based principal component analysis of protein dynamics". In: *The J. chemical physics* 143.24 (2015), 12B640_1.

[4] Elnaz Aledavood et al. "Elucidating the activation mechanism of AMPK by direct pan-activator PF-739". In: *Front. molecular biosciences* (2021), p. 1026.

Appendices

Supporting information to articles and manuscripts

.1 Supporting information to Published Article 1: Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks

Supporting Information

Exploring Allosteric Pathways of a V-Type Enzyme with Dynamical Perturbation Networks

Aria Gheeraert[♮], *Lorenza Pacini*[†§#], *Victor S. Batista*[‡], *Laurent Vuillon*[§], *Claire Lesieur*[†#]* *and Ivan Rivalta*[♮¶]*

[♮]Université de Lyon, CNRS, Institut de Chimie de Lyon, École Normale Supérieure de Lyon, 46 Allée d'Italie, F-69364 Lyon Cedex 07, France. [†]Institut Rhônalpin des systèmes complexes, IXXI-ENS-Lyon, 69007, Lyon, France. [§] LAMA, Univ. Savoie Mont Blanc, CNRS, LAMA, 73376 Le Bourget du Lac, France. [#]AMPERE, CNRS, Univ. Lyon, 69622, Lyon, France. [‡]Department of Chemistry, Yale University, P.O. Box 208107, New Haven, CT 06520-8107, and Energy Sciences Institute. [¶]Dipartimento di Chimica Industriale "Toso Montanari", Università degli Studi di Bologna, Viale del Risorgimento 4, I-40136 Bologna

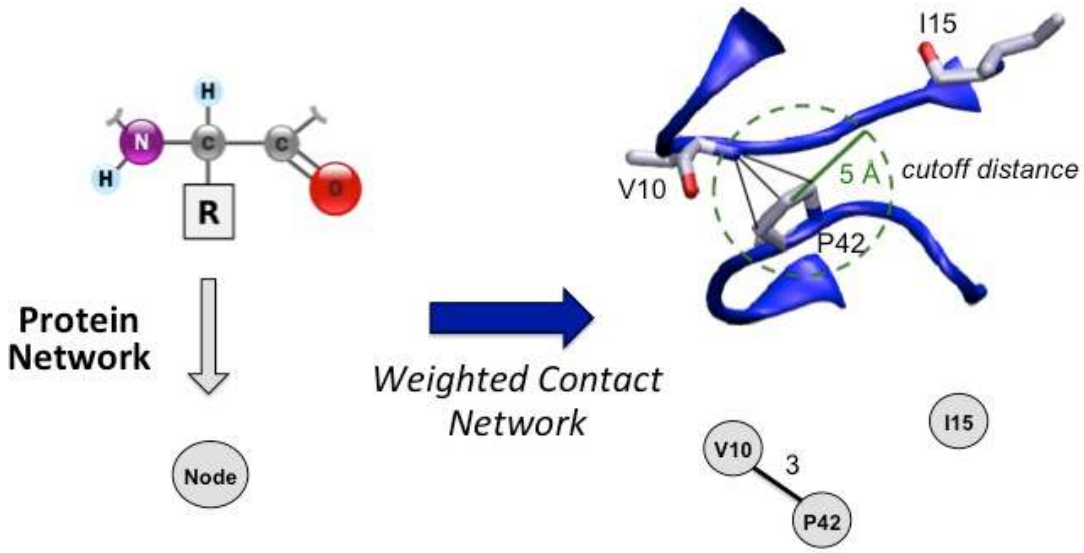


Figure S1. Each amino acid residue represents a node in the protein network. The presence of atomic contacts within the cutoff distance (5 Å) ensures the link between two nodes (i.e. the edge) in the protein network. The edges are weighted according to the number of atomic contacts for each residue pair. The picture shows a general example (not directly related to IGPS) for the construction of connections between three residues and assignment of weights to existing edges.

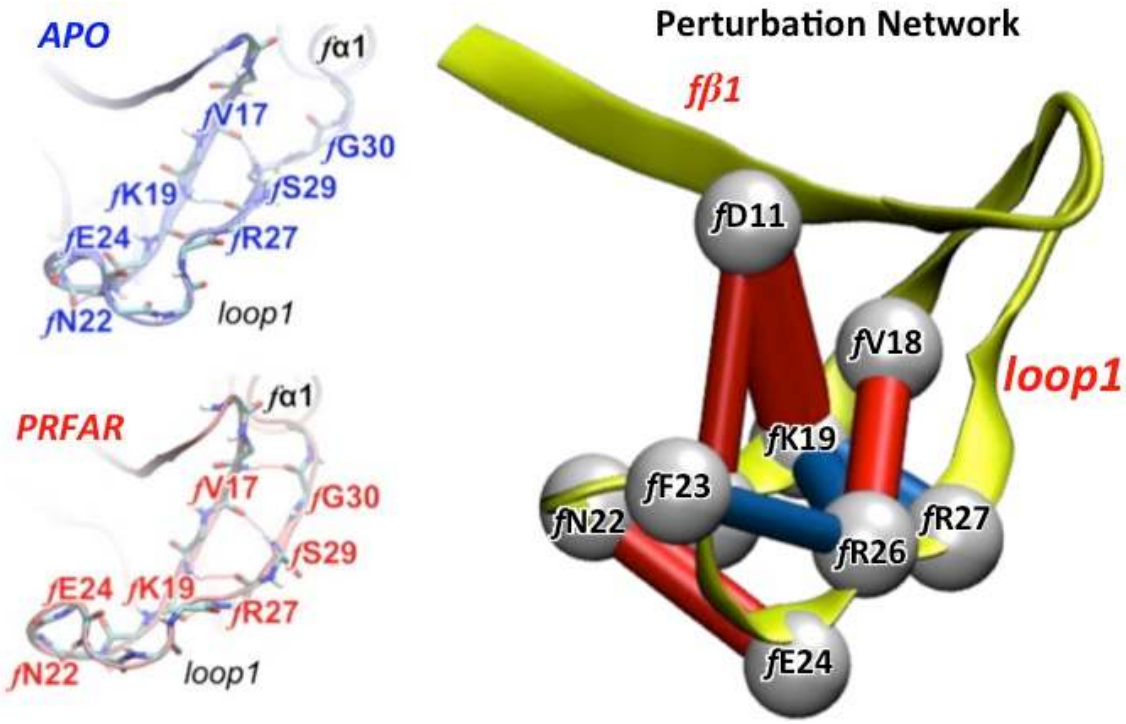


Figure S2. Comparison between hydrogen bonds modifications observed for *loop1* in the MD simulations of apo and PRFAR-bound complexes (left panels) and perturbations of heavy atoms contacts detected by means of the perturbation network analysis (right panel). A weight threshold $w_t = 6$ is used for the 3D representation of the network.

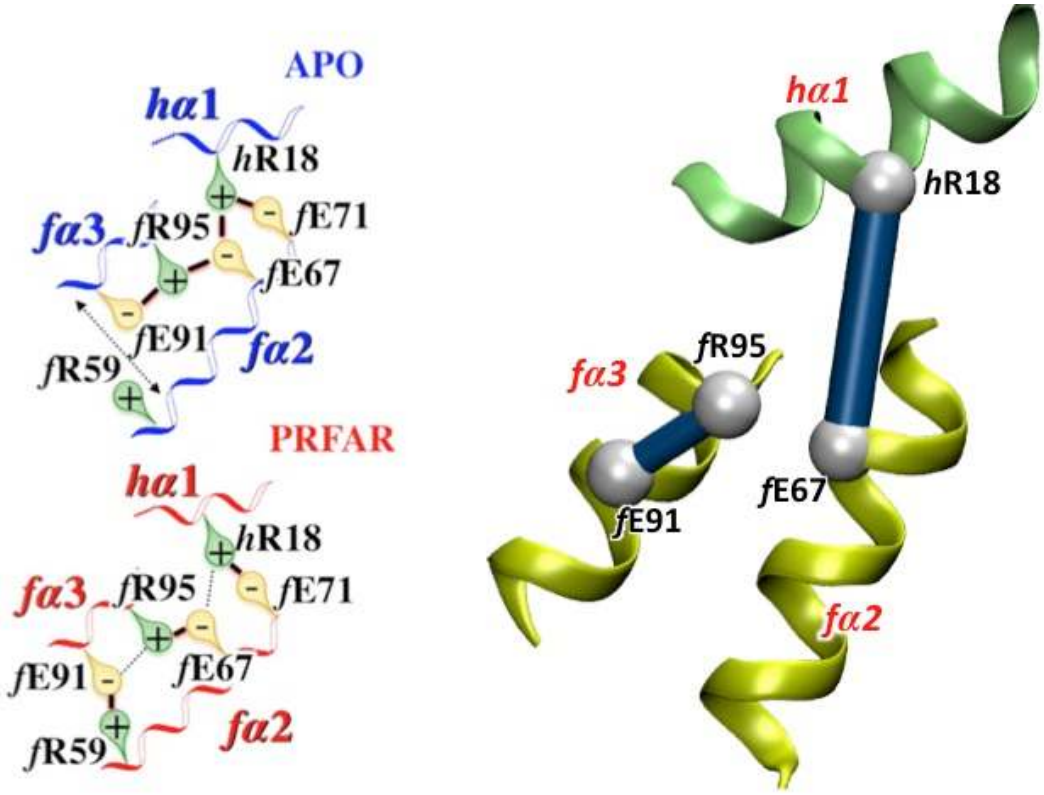


Figure S3. Comparison between ionic interactions modifications observed for $h\alpha 1$, $f\alpha 2$ and $f\alpha 3$ in the MD simulations of apo and PRFAR-bound complexes (left panels) and perturbations of heavy atoms contacts detected by means of the perturbation network analysis (right panel). A weight threshold $w_t = 6$ is used for the 3D representation of the network.

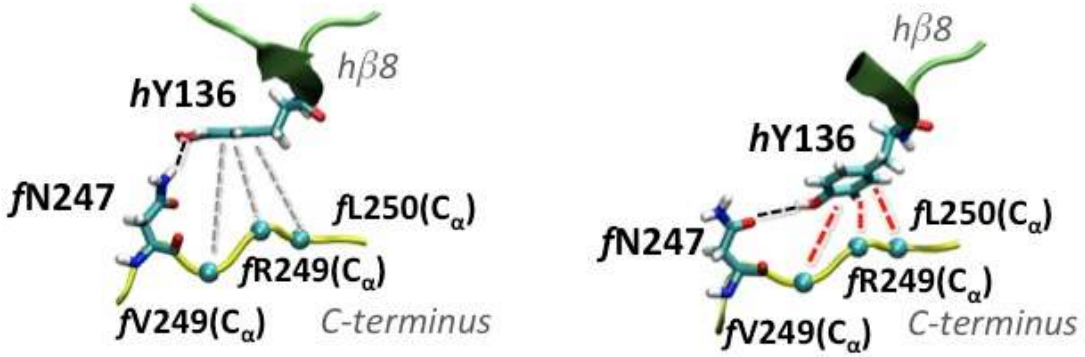


Figure S4. Schematic representation of contacts between the invariant *h*Y136 residue in *hβ*8 and residues *f*V248, *f*R249 and *f*L250 in the C-terminal domain of HisF, showing the change of H-bonding between *h*Y136 and *f*N247 that brings *h*Y136 closer to the flexible C-terminus.

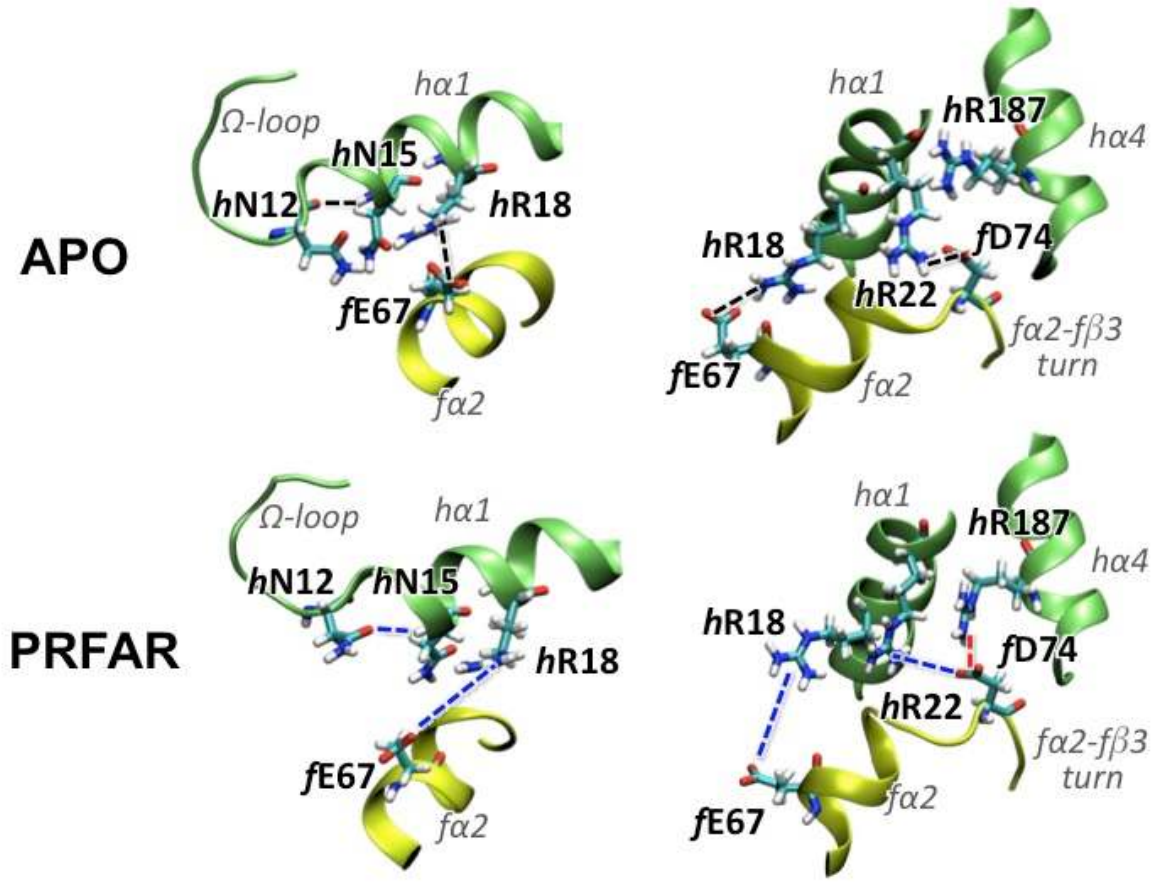


Figure S5. Representative configurations extracted from the MD simulations of the apo (top panels) and PRFAR bound (bottom panels) IGPS complexes, showing the *h*R18-*f*E67 salt–bridge disruption and the resulting partial unfolding of *hα*1 helix (propagating towards the active site via the *Ω-loop*) and rearrangement of interactions between polar/charged residues in *hα*1 and *hα*4 helices and the *fα*2–*fβ*3 turn.

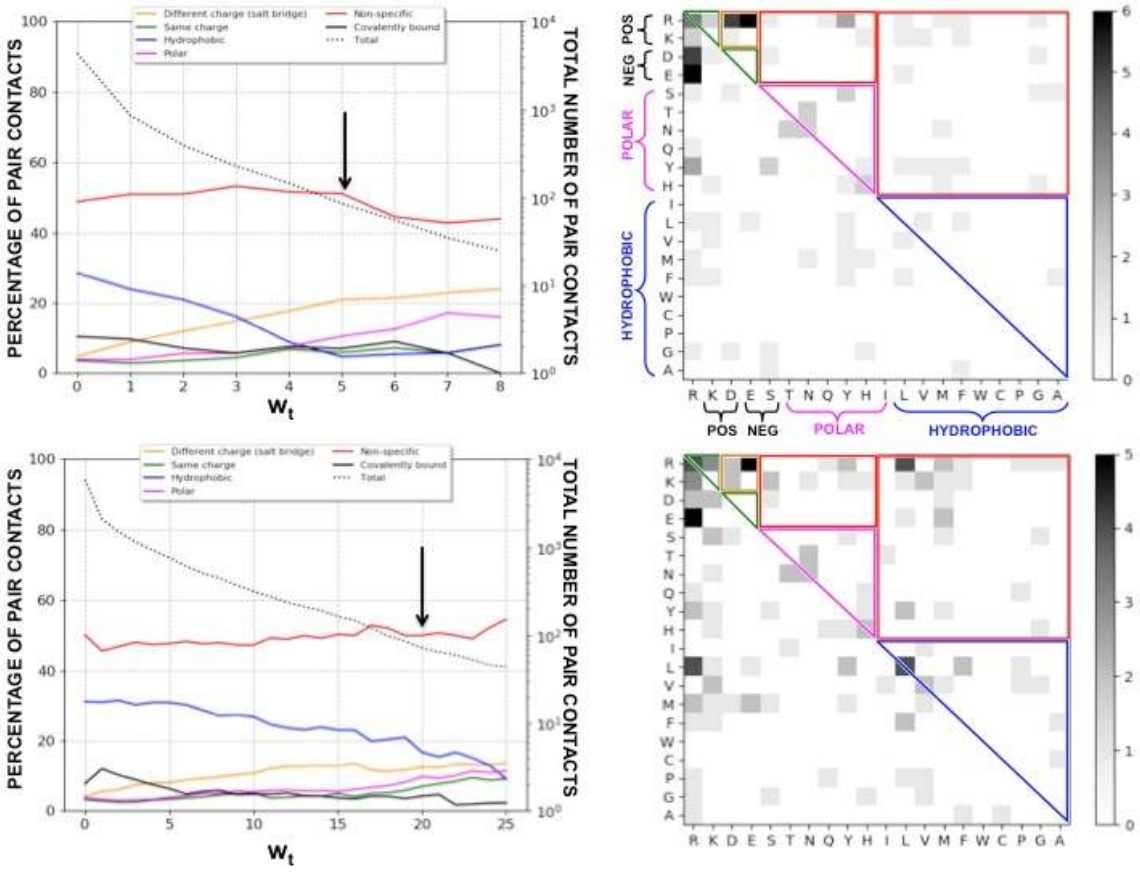


Figure S6. Analysis of the types of pair contacts detected by the perturbation networks using contacts between heavy atoms (top panels) and between all atoms including hydrogens (bottom panels). Left panels plots show the total number of pair contacts (dotted lines) and the percentage of pair contacts in the perturbation networks according to the type of interactions, which are defined as following: different charge (yellow lines) = R or K with D or E; same charge (green lines) = R with K or D with E; hydrophobic (blue lines) = I, L, V, M, F, W, C, P, G, A with themselves; polar (magenta lines) = S, T, N, O, Y, H with themselves. Note that since all histidine residues are not protonated (according to standard protonation at pH=7 for this enzyme), H is considered as a polar residue. Right panels maps show the contributions of specific amino acids to the pair contacts for a given perturbation weight threshold (w_t = 5 for heavy atoms and w_t = 20 for all atoms), with boxes highlighting the type of interactions involved.

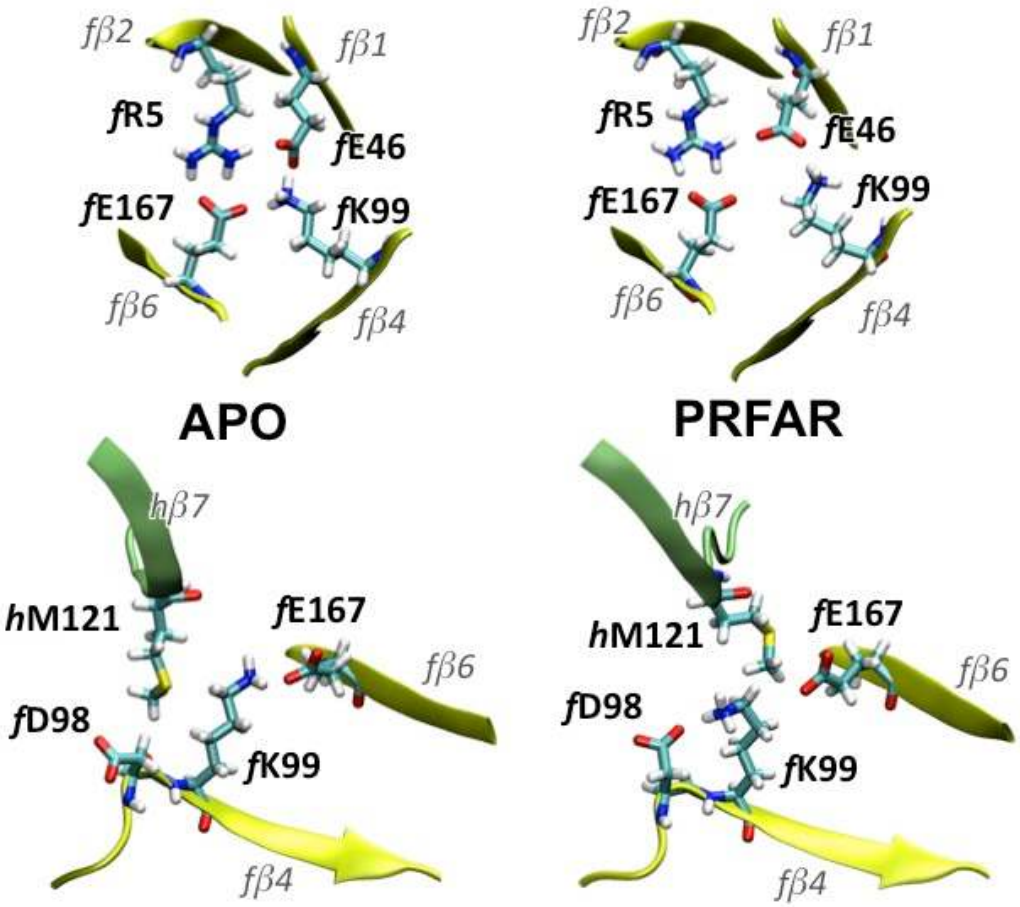


Figure S7. Representative configurations extracted from the MD simulations of the apo (left panels) and PRFAR bound (right panels) IGPS complexes, showing the effects of PRFAR binding to the interactions between the *h*M121 residue (features several contacts perturbations in the network analysis) and the invariant *f*R5, *f*K99 and *f*E167 residues that belong to the ammonia tunnel gate of the HisF barrel (top panels) and with the highly conserved *f*D98 (bottom panels) of the structurally important *f*D98–*h*K181 salt–bridge anchor.

.2 Supporting information to Manuscript 1: Connected Component Analysis of Dynamical Perturbation Contact Network

Supplementary information to Connected Component Analysis of Dynamical Perturbation Contact Network

Aria Gheeraert,[†,‡] Laurent Vuillon,[*,†] and Ivan Rivalta[*,‡,¶]

†*LAMA, Université Savoie Mont-Blanc, CNRS, Bourget-du-Lac, France*

‡*Dipartimento di Chimica Industriale "Toso Montanari", Università di Bologna, Viale Risorgimento 4, I-40136 Bologna, Italy*

¶*Univ Lyon, Ens de Lyon, CNRS UMR 5182, Université Claude Bernard Lyon 1 Laboratoire de Chimie, F69342, Lyon, France*

E-mail: laurent.vuillon@univ-savoie.fr; i.rivalta@unibo.it

Phone: +33 4 79 75 87 33; +39 051 209 3617

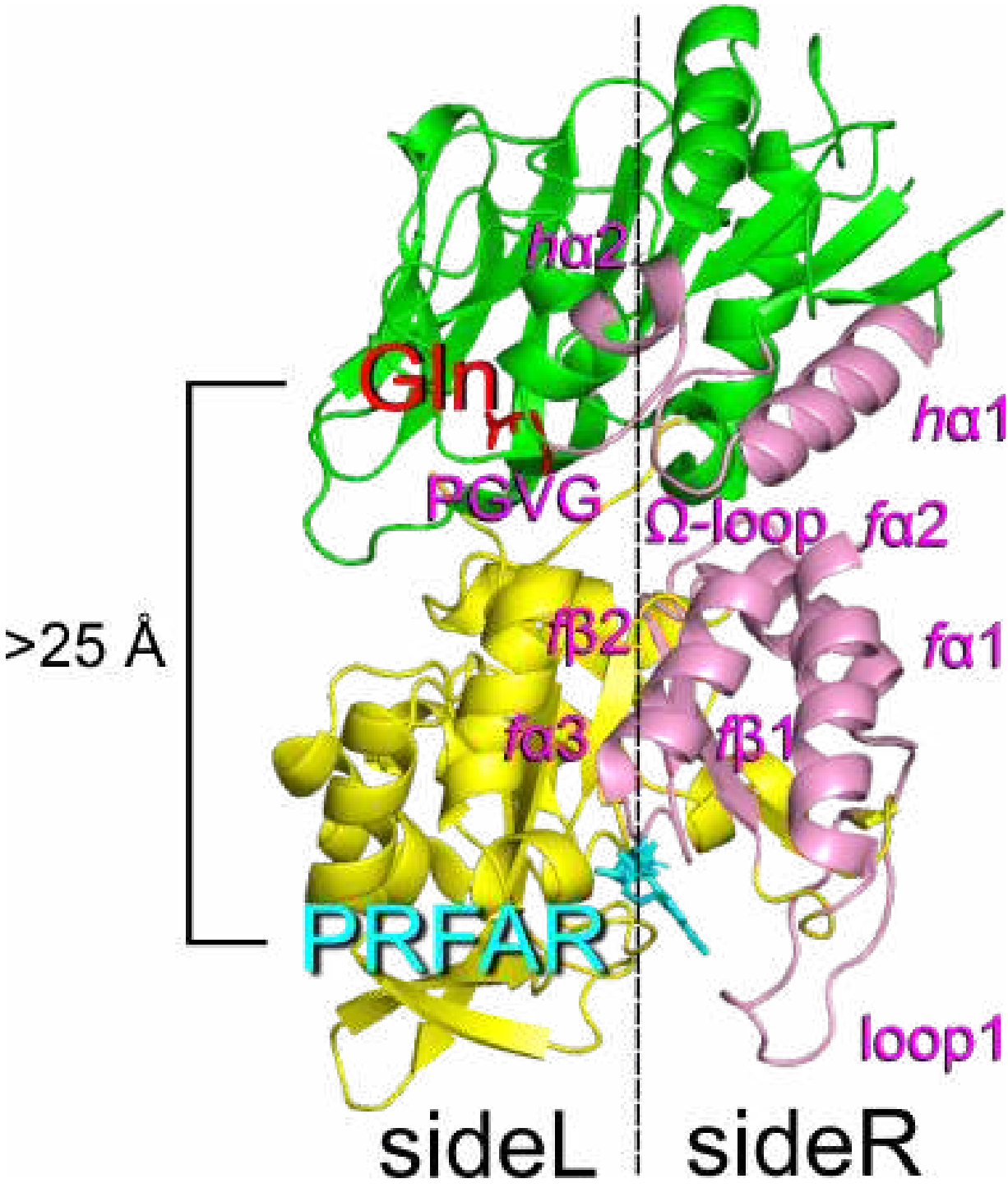


Figure 1: Allosteric mechanism of IGPS from *Thermotoga maritima*. The substrate (glutamine) is positioned in the active site and represented in red. The effector (PRFAR) is positioned in the effector site and represented in cyan. HisF is in yellow and HisH in green. Key secondary structure elements are represented and labeled in pink.

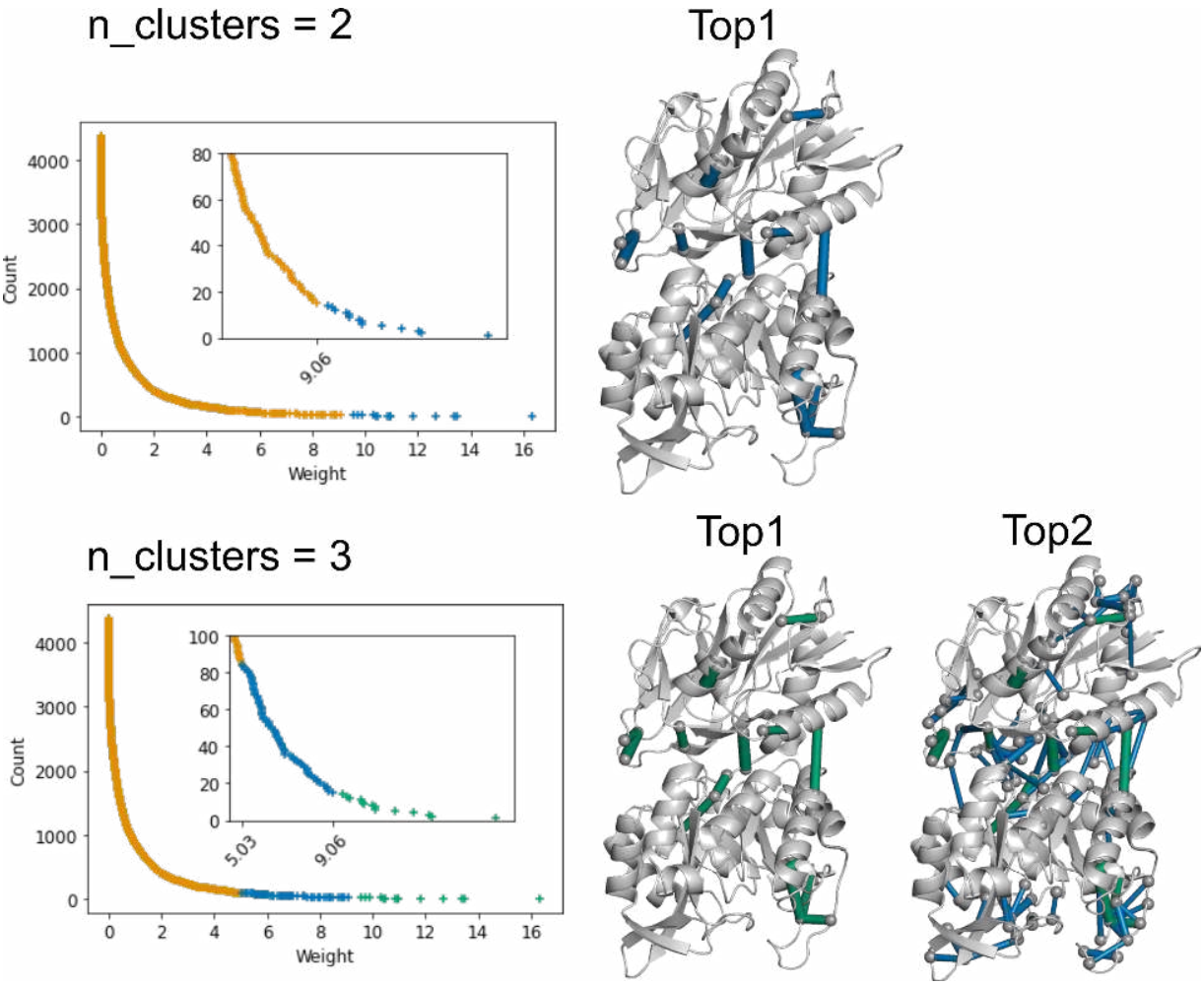


Figure 2: (*top*) Birch clustering with 2 clusters displaying the top cluster on the protein. (*bottom*). Birch clustering with 3 clusters displaying the top-one and top-two cluster on the protein.

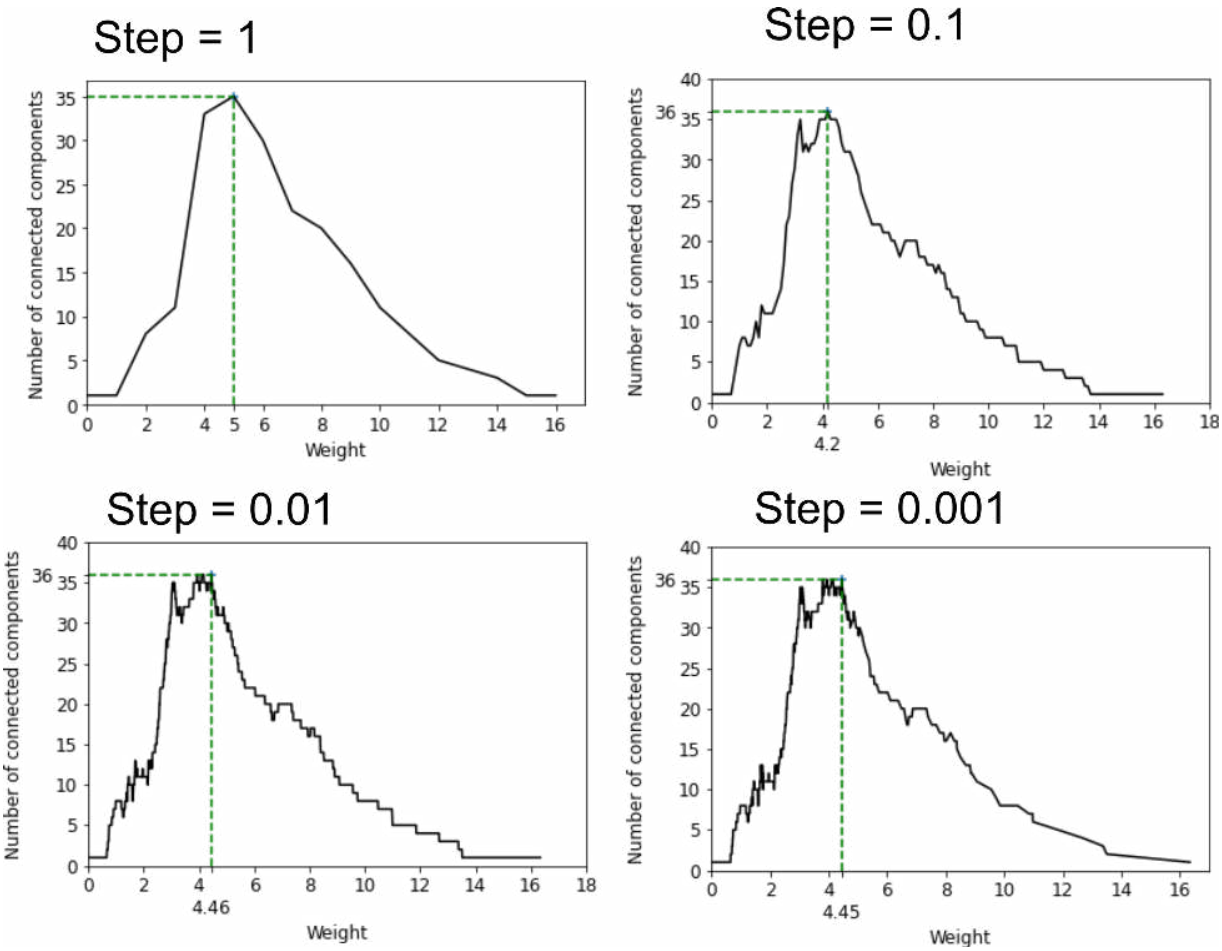


Figure 3: Connected component analysis procedure using steps instead of removing successively edges in the graph. This rough version gives a max component threshold of 5 with a step of 1, 4.2 with a step of 0.1, 4.46 with a step of 0.01 and 4.45 with a step of 0.001

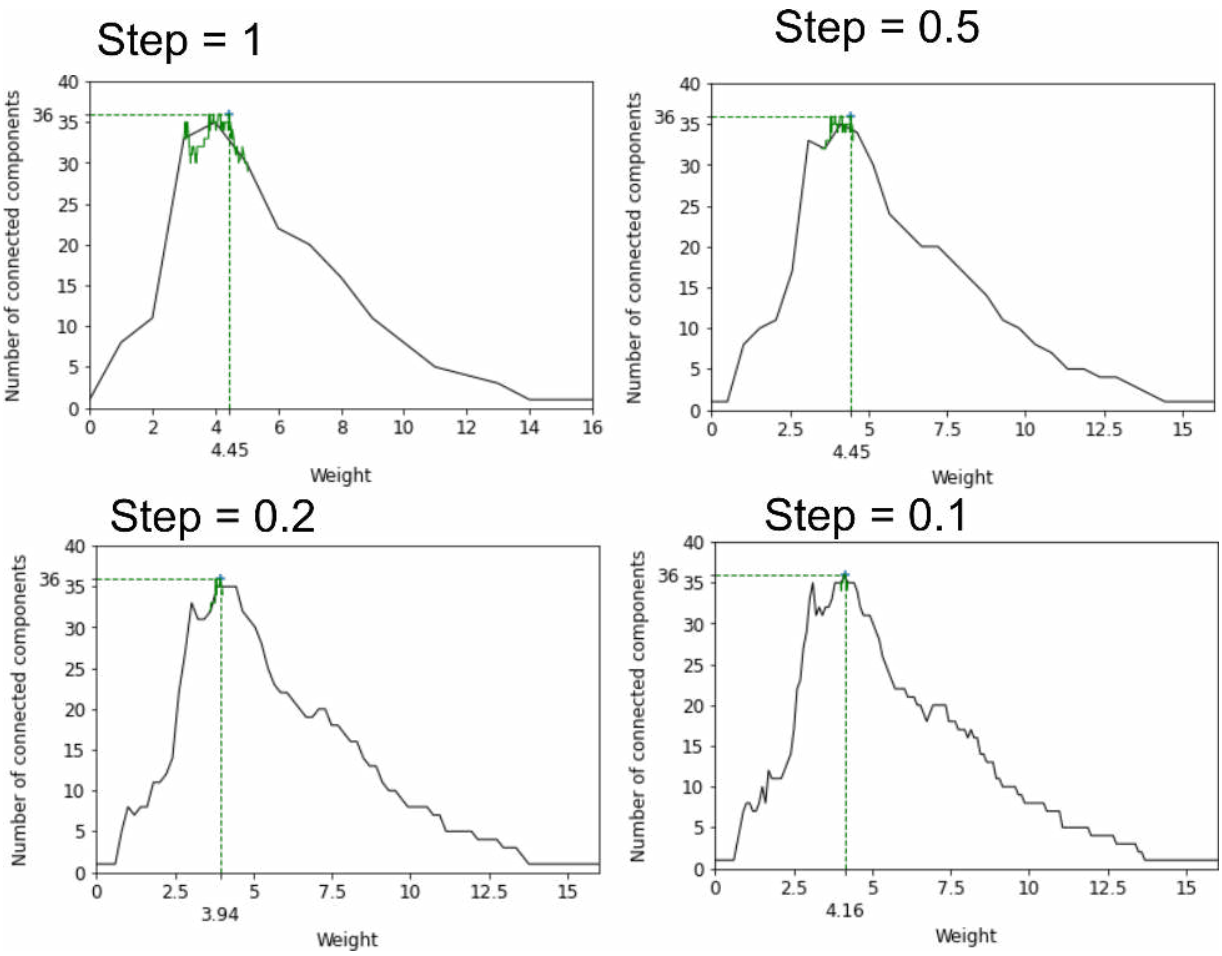


Figure 4: Combination of the step procedure and the exact procedure in the max interval found with different time steps. Only step of 1 and 0.5 are able to catch the true maximum but not 0.2 or 0.1.

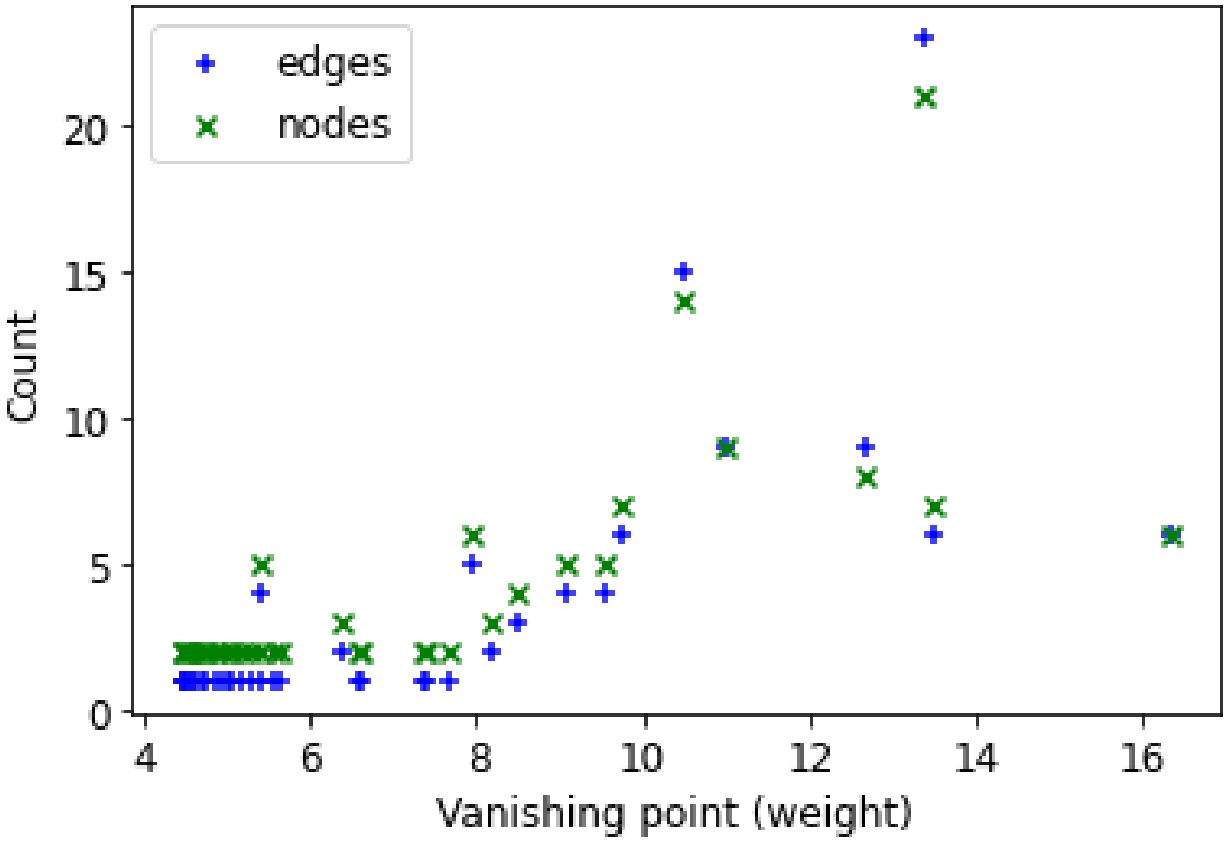


Figure 5: Scatter plot of the size (number of edges) and order (number of nodes) of each component against their vanishing point. There is a tendency of big vanishing points to create big components albeit not a complete correlation.

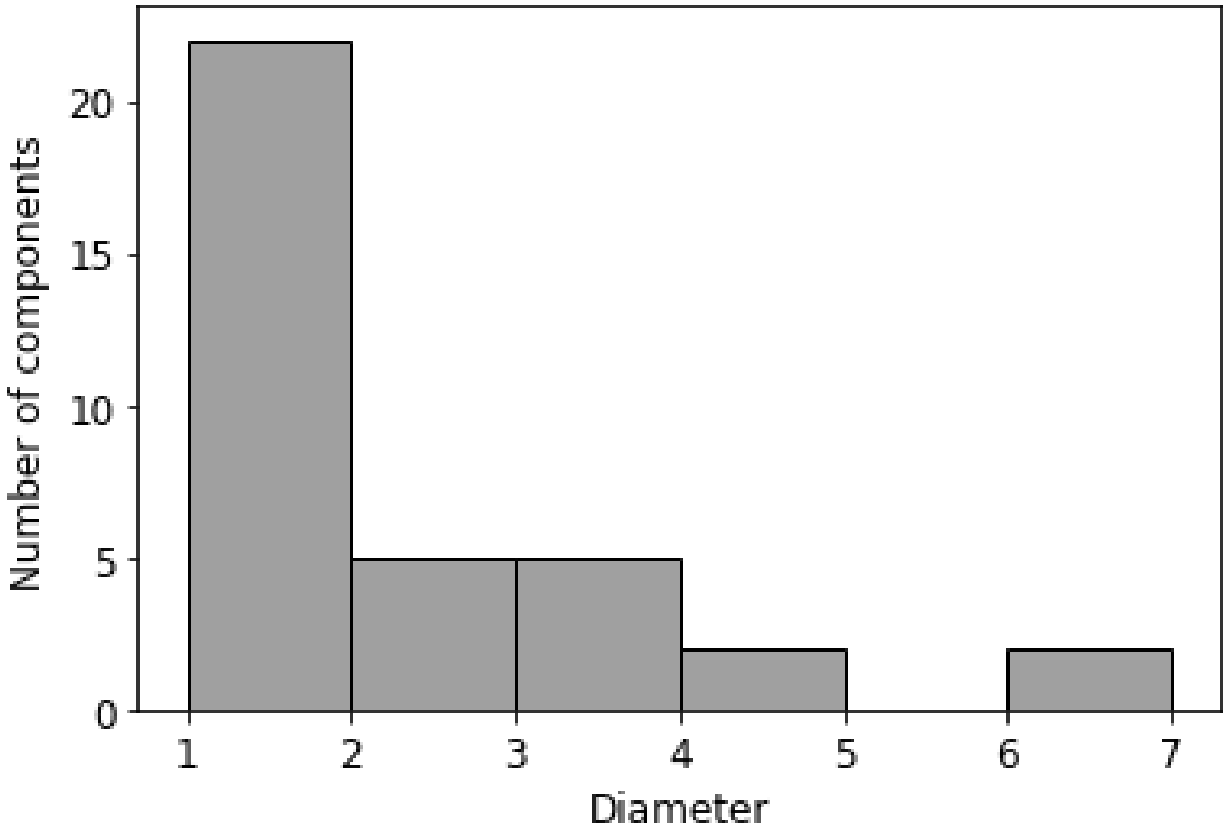


Figure 6: Distribution of the diameters in the final components. 22 components have a diameter of 1 (thus consisting of a single edge) while 5 have a diameter of 2 (trivial examples of propagations). The ninth major component have a diameter bigger or equal than 3.

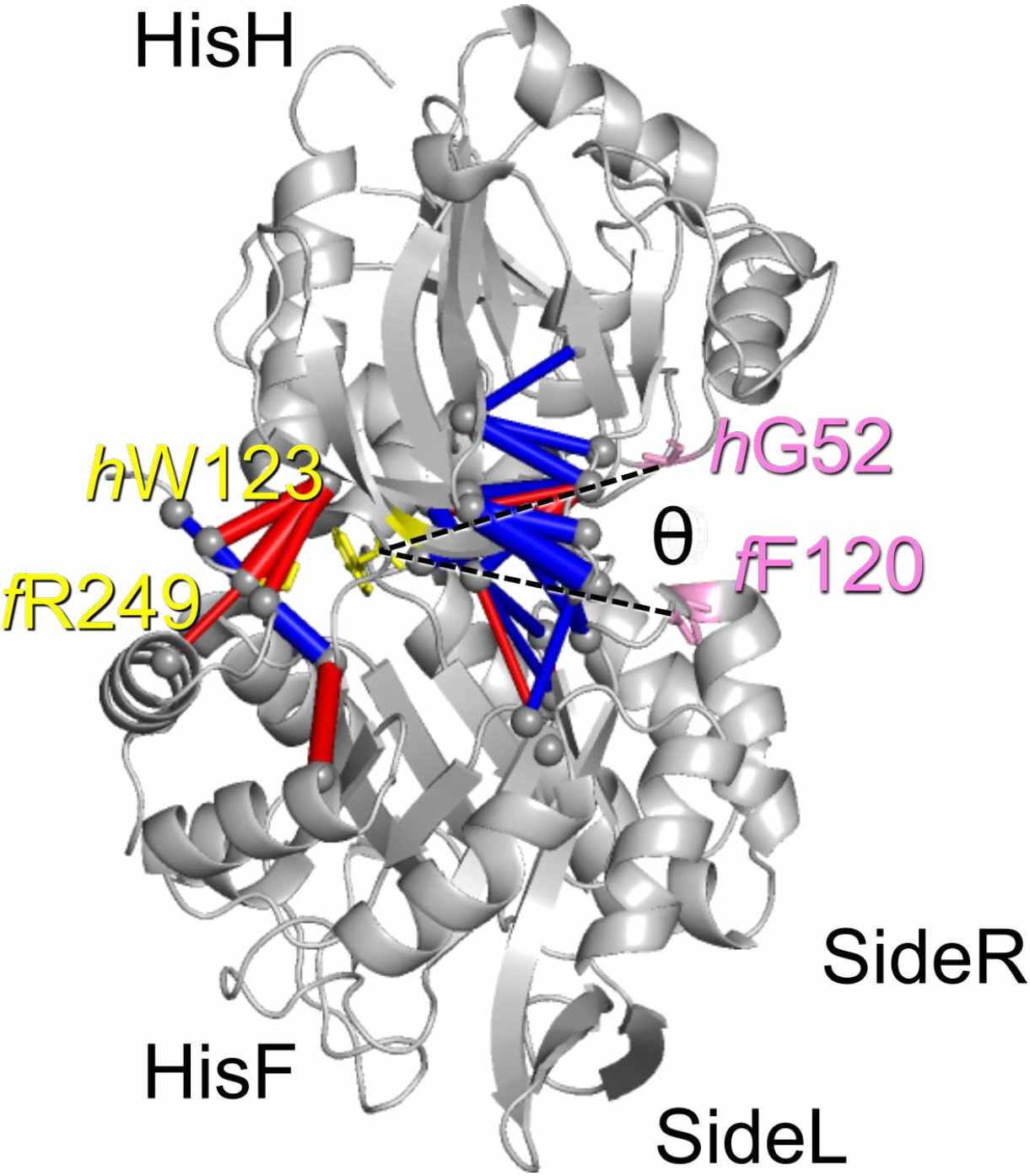


Figure 7: Components 2, 3 and 6 showing a tighter interactions near the hinge (red edges), while the overall alteration in breathing motion (angle between fF120, hW123 and hG52) .

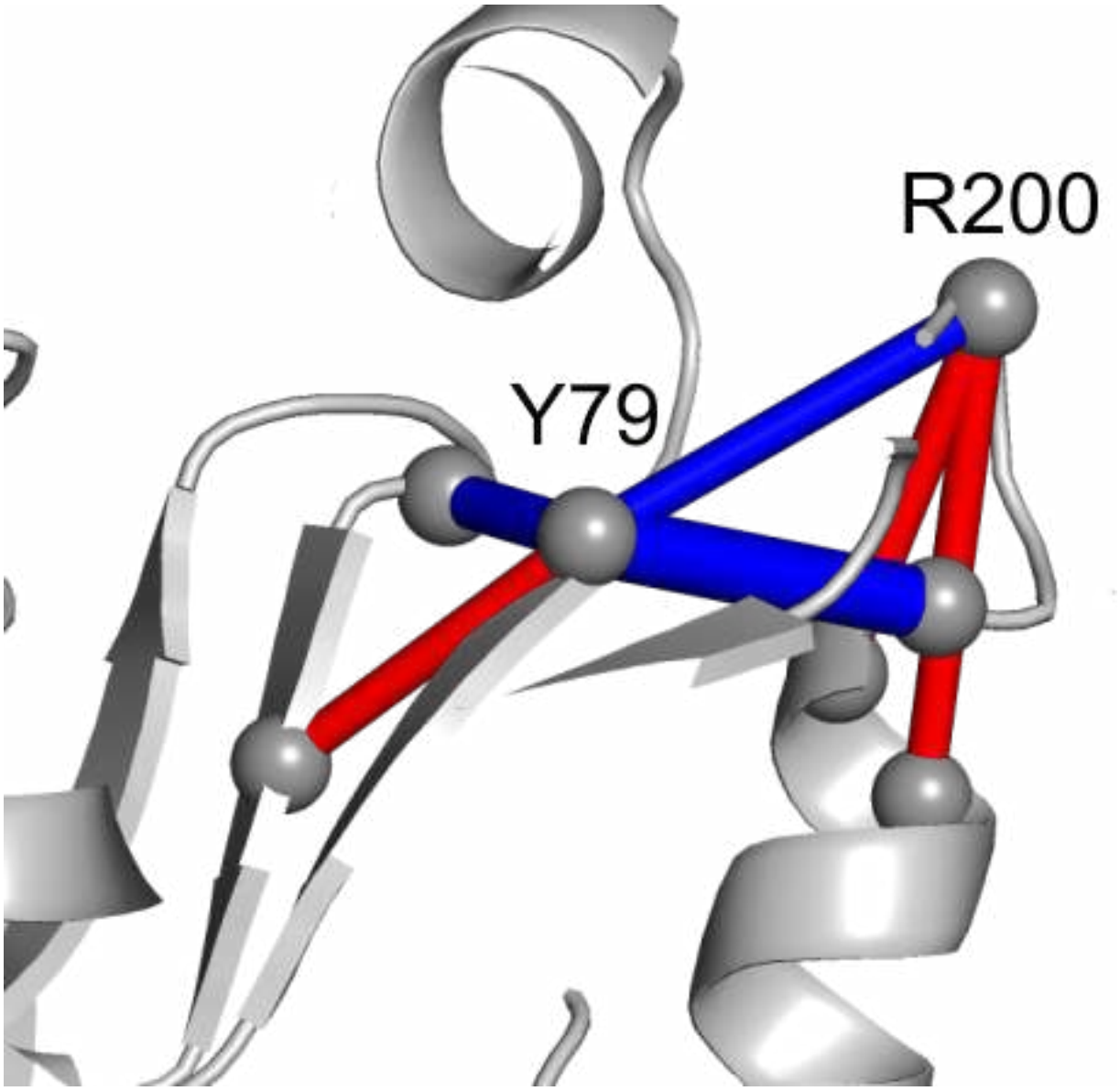


Figure 8: Component 7 including its two main hubs: hY79 and hR200. hR200 is one of the few unresolved residue from 1GPW crystal structure meaning that these displacements could be attributed to thermal fluctuations.

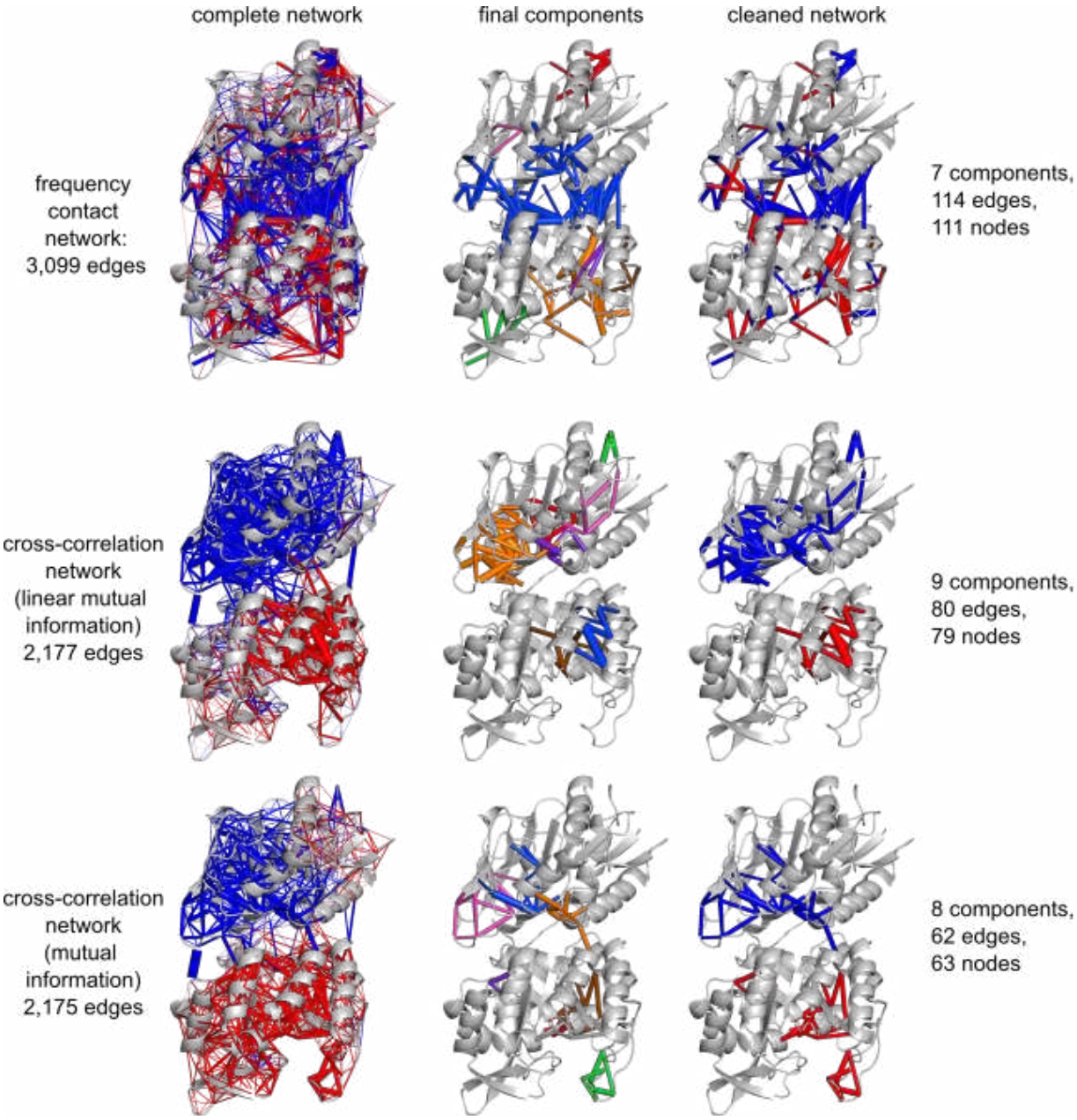


Figure 9: Connected component analysis applied to other perturbation networks (a frequency contact network, and two cross-correlation computed with linear and non-linear mutual information).

.3 Supporting information to Published Article 2: Distinct Allosteric Pathways in Imidazole Glycerol Phosphate Synthase from Yeast and Bacteria

Biophysical Journal, Volume *121*

Supplemental information

Distinct allosteric pathways in imidazole glycerol phosphate synthase from yeast and bacteria

Federica Maschietto, Aria Gheeraert, Andrea Piazzi, Victor S. Batista, and Ivan Rivalta

Supporting Information

Distinct Allosteric Pathways in Imidazole Glycerol Phosphate Synthase from Yeast and Bacteria

Federica Maschietto[1], Aria Gheeraert[2], Andrea Piazzi[3], Victor S. Batista[1]* and Ivan Rivalta[2,3]*

[1]Department of Chemistry, Yale University, New Haven, CT 06511, U.S.A.
[2]Université de Lyon, CNRS, Institut de Chimie de Lyon, École Normale Supérieure de Lyon, 46 Allée d'Italie, F-69364 Lyon Cedex 07, France
[3]Dipartimento di Chimica Industriale "Toso Montanari", ALMA MATER STUDIORUM, Università di Bologna, Viale del Risorgimento 4, 40126 Bologna, Italia

Section S1. Materials and Methods

Correlation matrices for IGPS from *Thermotoga maritima* (*Tm*-IGPS) are obtained from the same trajectories and following the same protocol as in reference [1], while yeast models (*Sc*-IGPS from *Saccharomyces cerevisiae*) are built ex-novo.

The computational structural models for apo and PRFAR bound yeast IGPS complexes are based on the crystal structure of the bienzyme complex from *S.cerevisiae* at 2.4 Å resolution (Protein Data Bank code 1OX6-B).[2] The HisH-HisF apo-complex having several missing residues (261-275, 301-304, and 551-552) and three extra residues at the beginning of the chain required modeling prior to simulation. To complete the structure, first, we stripped the first three residues, then we aligned and added residues 256-260 and 299-310 from 1OX4-B (removing overlapping residues from 1OX6 due to poor alignment). Finally, we added residues 550-552 from 1JVN-A, (removing residue 550 from 1OX6-B). We constructed the remaining residues (256-275) using different tools, using which we produced six different structures. One structure was generated using Modeller,[3] a second one using Swiss-Model,[4] and four suitable homology models were found on modbase. PRFAR was bound to each model by aligning each structure to the effector-bound crystal structure of yeast IGPS (PDB code 1OX5).

The twelve generated structures (six in the apo state, six bound to the effector) align with RMSD < 5 Å. To allow for a direct comparison between the dynamics of IGPS enzymes from *Tm*- and *Sc*-IGPS we kept the simulation conditions analogous to the one used for bacterial IGPS in reference.[1] For the sake of clarity, we report some essential details below. MD simulations of the apo and PRFAR-bound structures of yeast IGPS are based on the AMBER-ff99SB[5] force field for the protein and Generalized Amber Force Field[6] for the PRFAR ligand (see SI Text), as implemented in the Amber20 software package.[7] We performed twelve independent MD simulations, one for each complex (apo and PRFAR-bound) for a total simulation time of 1.2 μs.

Structure refinements such as addition of hydrogen and explicit TIP3 water solvent molecules (reaching density values >= 0.9 mol•Å^{-3}) are performed using AmberTools (2020). A constrained optimization with all atoms but solvent fixed at the crystal structure positions yields optimized solvated structures which are then slowly heated to 303 K, performing MD simulations (100 ps) in the canonical NVT ensemble using Langevin dynamics. We apply harmonic constraints to protein and PRFAR heavy atoms, with force constants set to 1 kcal•mol^{-1}. During the heating procedure all positional constraints are gradually lifted until all atoms are set freed.

Unconstrained MD simulations are run for more than 9 ns, for total pre-equilibration simulation time of at least 10 ns. The pre-equilibrated systems are simulated in the NPT ensemble at 300 K and 1 atm using the Langevin dynamics for 100 ns. All simulations are performed using periodic boundary conditions. Van der Walls interactions are calculated using a switching distance of 10 Å and a cutoff of 12 Å and electrostatic interactions are treated using the Particle Mesh Ewald method.[8] We employ the multiple time-stepping algorithm,[9] where bonded, shortrange nonbonded, and long-range electrostatic interactions are evaluated at every one, two, and four time steps, respectively, using a timestep of integration set to 1 fs.

Section S2. Details on the computation of correlation values and their analysis through the eigenvector centrality metrics, principal component analysis and allosteric pathways across yeast and bacterial IGPS.

Generalized correlation coefficients, eigenvector centrality and community network analysis

We quantify the extent of the dynamical correlation of fluctuations in the positions of Cα-atoms by computing the generalized correlation coefficient between each pair of residues,[10]

$$r_{\mathrm{MI}}[x_i, x_j] = (1 - exp(-\frac{2}{3} I[x_i, x_j]))^{1/2} \quad (1)$$

computed in terms of mutual information (MI), [11]

$$I[x_i; x_j] = [Hx_i] + H[x_j] - H[x_i; x_j]. \quad (2)$$

Here, $[Hx_i]$, $H[x_j]$, $H[x_i; x_j]$ are the marginal and joint (Shannon) entropies for atomic vector displacements $(x_i; x_j)$, computed along twelve independent 100 ns MD simulations for both apo and PRFAR-bound yeast IGPS complexes. The resulting generalized correlation coefficient values r_{MI} values fall in between 0 and +1, representing respectively uncorrelated and fully correlated variables. r_{MI} alone can be hard to decipher and require some post-processing to interpret protein behavior. Network analysis tools,[12,13] including different centrality metrics[14] can be applied for the interpretation of correlated protein motions and their allosteric behavior. Here, the Cα-atoms of the proteins' amino-acid residues constitute the nodes of a dynamical network graph, connected by edges (residue pair connection in terms of $r_{\mathrm{MI}}[x_i, x_j]$). An adjacency matrix A is then constructed such that it can be used to identify the key amino acid residues of IGPS with high susceptibility to effector binding. A simple yet effective metric extract "central" nodes in A is the eigenvector centrality EC. The basic idea behind this measure is the assumption that the centrality index of a node is not only determined by its position in the network but also by the neighboring nodes, hence it measures how well connected a node is to other well-connected nodes in the network. The EC of a node is defined as the weighted sum of the centralities of all nodes that are connected to it by an edge, A_{ij}:

$$c_i = \epsilon^{-1} \sum_{j=1}^{n} A_{ij} c_j, \quad (3)$$

where c is an eigenvector associated to the largest eigenvalue of A. Being any eigenvector defined only minus a multiplicative constant we orient the eigenvector in the positive quadrant (whatever the sign obtained from the diagonalization). Additionally, an exponential damping factor with a length parameter λ can be introduced to Eq. 3, by defining A as:

$$A_{ij} = \begin{cases} 0 & if\ i = j \\ r_{\mathrm{MI}}[x_i, x_j] exp(-d_{ij}/\lambda) & if\ i \neq j \end{cases}. \quad (4)$$

λ controls the locality of the correlations under consideration based on the average distance between residues d_{ij}.

Hence, using short enough values of λ will result in neglecting the correlation between residues that are far away from one another, revealing the effect of the locality in the allosteric pathway. On the other hand, by setting λ to a very large value, all correlations, including those between residues separated by long distances, will be retained and $A_{ij} = r_{\mathrm{MI}}[x_i, x_j] \quad \forall\, i \neq j$. In the main text the results presented correspond to a value of $\lambda = 5$.

Because we are interested in analyzing how the information transmission is affected by the allosteric stimulator, we focus on the difference centrality values computed by as $\Delta c = c_{\mathrm{PRFAR}} - c_{\mathrm{APO}}$. The nodes with higher eigenvector difference centrality are those acting as the principal "channels" for momentum transmission across the protein.

We visualize the c_i coefficient relative to each amino-acid in the protein structure, coloring each node from blue (zero centrality) to red (maximum centrality). In all of the cases, we apply a renormalization of the centrality values such that each falls in the -1, +1 range, as:

$$c_i' = 2\frac{c_i - min(c)}{max(c) - min(c)} - 1 \qquad i = 1, \text{number of nodes} \tag{5}$$

In the present study, we calculated generalized correlation coefficients based on mutual information and EC values independently on 100 ns apo and PRFAR-bound trajectories of yeast and bacterial IGPS and averaged over six and four replicas, respectively. As mentioned before, the trajectories used for bacterial IGPS are the same as in reference,[1] hence the EC values reported both in the main text (Figure 3A) and below (Figure S6) are the same as in reference [15] whereas those relative to yeast IGPS are computed ex novo, following the same procedure, as described Section S1.

The protein-network can be used to determine the optimal pathways for the information transfer between two nodes, defined as the shortest paths connecting a specific pair of nodes. In this context, edge lengths, i.e. the internode distances in the graph, are defined using the $r_{\mathrm{MI}}[x_i, x_j]$ coefficients according to *-log(*$r_{\mathrm{MI}}[x_i, x_j]$*)*, implying that highly correlated pairs (featuring good communication) are close in distance in the graph.

In particular we applied the Dijkstra algorithm to calculate the shortest pathways between residues *f*A233-*f*A234-A523/G524-R528 and hC84-C83, where each set of residues belongs to a different domain of bacterial and yeast IGPS, respectively. Hence, the computed pathways are composed of residue-to-residue steps that optimize the overall correlation (i.e., the momentum transport) between residues *f*A223-*f*A224 (at the effector site) and *h*C84 (in the glutaminase active site) in *Tm*-IGPS, and similarly residues K334, A523, G524 and C83 in His7.

Principal Component Analysis

Principal Component Analysis (PCA)[16,17] has been employed to capture the essential motions of the simulated systems. In PCA, the covariance matrix of the protein Cα atoms is calculated and diagonalized to obtain a new set of coordinates (eigenvectors) to describe the system motions. Each eigenvector – also called Principal Component (PC) – is associated with an eigenvalue, which denotes how much each eigenvector is representative of the system dynamics.

To avoid translational artifacts, we set the center of mass of each frame at the origin, and rotate each frame to its optimally aligned orientation relative to the average structure - computed over all apo trajectories - which also has its center of mass at the origin. Next, we evaluate the covariances of the positional fluctuations of each system over the apo and PRFAR-bound trajectories obtained by concatenation of the independent apo and effector-bound replicas. Because the motion of side-chains is mostly independent of the essential dynamics of IGPS, we restrict the covariance to the backbone atoms only. Projecting the original (centered) data onto the eigenvectors results in the PCs, whose associated eigenvalue (variance) is indicative of the portion of motion that the eigenvector describes.

Together, the first two principal components relative to *Tm*-IGPS incorporate 44% and 33% of the total motion of the bacterial apo and PRFAR-bound trajectories, respectively (Figure S3-A), while the percentages become 42% and 44% for His7 (Figure S3-B) . The contribution added by the third PC is much smaller hence we limited our analysis to the first two.
The interest in projecting the trajectory coordinates onto the PCs is that we can visualize the essential motions induced by effector-binding in yeast and bacterial IGPS on the protein structure, along the trajectory. The procedure is described below.
First, we project the original trajectory onto the first two PRFAR-minus-apo difference principal components (ΔPC) and visualize their motion (details in SI). The weights over the i^{th} principal component relative to a given trajectory are given as

$$w_i(t) = r(t) - \bar{r} \cdot PC_i, \tag{6}$$

where $r(t)$ is a vector containing the stacked cartesian coordinates of the selected group of atoms at time (t) and $\bar{r}$ are the mean (stacked x,y,z) coordinates along a selected (apo) trajectory. PC_i is the i^{th} principal component, having dimension ($3n$), with $n = \text{number of atoms selected}$. The resulting weight vectors $w(t)$ are ($3n$) dimensional and the dimension of n is equivalent to that of each row/column of the covariance matrix, and will coincide with the length of the PCs. Then, the projected coordinates on PRFAR-minus-apo difference principal components (ΔPC) are

$$r_i(t) = w_i(t) \times \Delta PC_i + \bar{r}, \tag{7}$$

Here, the product of the weights $w_i(t)$ - computed at each timestep of the apo trajectory - with the $i-th$ difference eigenvector ΔPC_i accounts for the fluctuations around the mean on that axis (i.e., the fluctuations induced by PRFAR binding), so the projected trajectory $\mathbf{r}_i(t)$ simply describes the effector-induced fluctuations added onto the mean positions $\bar{\mathbf{r}}$.

Additional comments on generalized correlation coefficients, EC and PCA

With regard to the analysis reported in the main text reported for yeast IGPS, it is worth discussing more in depth the outcomes of the single replicas as compared to the average. This analysis supports the finding in the text and shows the relevance of the simulations.
MD simulations are inherently chaotic, hence two simulations started from similar inputs may end up in significantly different configurations, making it hard to verify whether the process under interest is actually captured within the dynamic trajectory. This is why running a single trajectory may not mean much and replicates are almost always required. Indeed, allowing for high variance in the simulations - as we do, for instance, using different homology models to construct representative initial states from which to start the dynamics, is paramount to ensure that the simulations capture the process of interest (in our case the allosteric events in the enzyme's dynamics). We calculated generalized correlation coefficients based on mutual information and covariances of atomic displacements independently on each 100 ns apo and PRFAR-bound trajectories of yeast and bacterial IGPS and averaged over six and four replicas, respectively. A standard way to verify that a set of simulations contains a statistically relevant ensemble is to check that different simulations show similar ensemble average properties. The more unconstrained is the motion of a system of interest the more likely it will be that different dynamics sample different states of the system. The trade-off between considering a "large enough" number of independent simulations that will reliably capture a process of interest,

without averaging out important fluctuations, is system dependent and requires careful case-by-case examination. These observations apply to the simulations described in this work. Correlations, covariances (and therefore all the metrics derived from these) are subject to changes depending on the dynamics. For *Tm*-IGPS the four 100 ns apo/PRFAR-bound replicas, based on which we calculated the average properties discussed in the main text, showed similar features (as discussed in the original publication[1]. We find rather larger deviations in the yeast as compared to *Tm*-IGPS. However, the average picture -obtained as the average apo-minus-holo correlation profile computed across the different models (shown in Figure 2) - is representative of the allosteric process although the individual simulations present different EC and PCA profiles (as shown in Figure S5). Among the six apo and PRFAR-bound replicas the dynamics that encompasses most of the allosteric traits is labelled as $\underline{sim_1}$ in the figures reported below. Figures 3 and 4 in the main text are associated to the representative dynamics of sim_1.

Section S3. Supplementary figures and tables

Breakdown of the secondary structural elements of His7

Secondary Structural Element	Residue numbers	Label	length	Secondary Structural Element	Residue numbers	Label	length
Beta strand	3 – 7	*h*β1	5	Helix	277 – 288	*f*α1	12
Helix	15 – 23	*h*α1	9	Beta strand	292 – 299	*f*β2	8
Beta strand	27 – 33	*h*β2	7	Helix	307 – 309		3
Helix	34 – 36		3	Helix	311 – 319	*f*α2	9
Helix	39 – 41		3	Turn	320 – 322		3
				Beta strand	327 – 332	*f*β3	6
Beta strand	45 – 49	*h*β3	5				
Helix	53 – 62	*h*α2	10	Helix	346 – 356	*f*α3	11
Helix	66 – 74	*h*α2'	9	Beta strand	359 – 363	*f*β4	5
Beta strand	79 – 83	*h*β4	5	Helix	365 – 376	*f*α4	12
Helix	84 – 87	*h*α3	4	Helix	386 – 394	*f*α4'	9
				Helix	396 – 398	*f*α4"	3
Beta strand	90 – 93	*h*β5	4				
				Beta strand	399 – 403	*f*β5	5
Beta strand	104 – 111	*h*β6	8	Beta strand	405 – 412	*f*β5'	8
Turn	114 – 116		3	Helix	413 – 415	*f*α5	3
Beta strand	119 – 125	*h*β7	7	Beta strand	433 – 440	*f*β6X	8
Beta strand	143 – 150	*h*β8	8	Turn	441 – 444		4
Helix	155 – 163	*h*α4X	9	Beta strand	445 – 450	*f*β6	6
Beta strand	167 – 173	*h*β9	7	Helix	451 – 460	*f*α6	10
Beta strand	176 – 184	*h*β10	9	Beta strand	465 – 468		4
Beta strand	187 – 193	*h*β11	7	Helix	471 – 473		3
Helix	194 – 196		3	Turn	474 – 476		3
Helix	198 – 209	*h*α4	12	Helix	482 – 491	*f*α6	10
Helix	221 – 227	*h*α4'	7	Beta strand	496 – 498		3
Helix	232 – 235	*h*α4"	4	Helix	505 – 514	*f*α7	10
Beta strand	240 – 248	*f*β1	9	Beta strand	518 – 523	*f*β8	6
Beta strand	250 – 252		3	Helix	524 – 527	*f*α8'	4
Beta strand	254 – 257		4				

Table S1. Full topography of secondary structural elements of yeast IGPS from https://www.uniprot.org/uniprot/P33734

Sequence alignment

Using the jFATCAT rigid algorithm implemented in the RCSB PDB Comparison Tool Reference (https://www.rcsb.org/alignment), we aligned structurally the amino acid sequence of HisH (PDB entry: 3ZR4.C), HisF (3ZR4.D) and His7 (1OX5.A). The sequence alignments are reported in Table S2. The structures of HisF and His7 are aligned for residues hM1–hS197 and P5–Q215 and the sequences of HisF and His7 are aligned for residues fM1–fE251 and G238–D553. Despite a similarity of ~ 50-60% (see Alignment Summary Table), the alignments show good structural similarity with a RMSD of the C-alpha backbone atoms ~ 2 Å. The alignments of the 3D structures are also reported in Fig. 1 in the main text.

Alignment Summary	**RMSD**	**Sequence Identity%**	**Sequence Similarity%**	**Length**
HisF-His7	2.03	46	63	241
HisH-His7	1.93	30	52	192

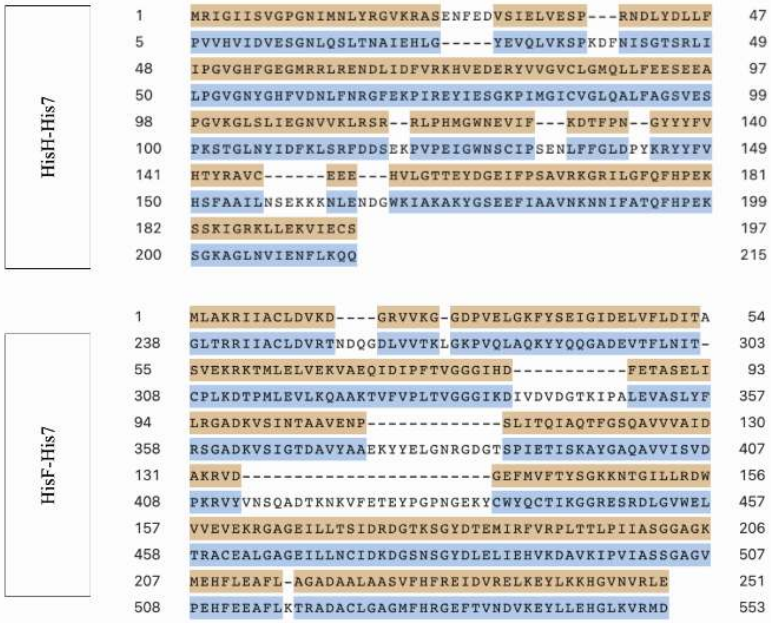


Table S2. Sequence alignment of His7 from *Saccharomyces cerevisiae* and HisH, HisF from *Thermotoga maritima*.

Correlation matrices from Elastic Network Model

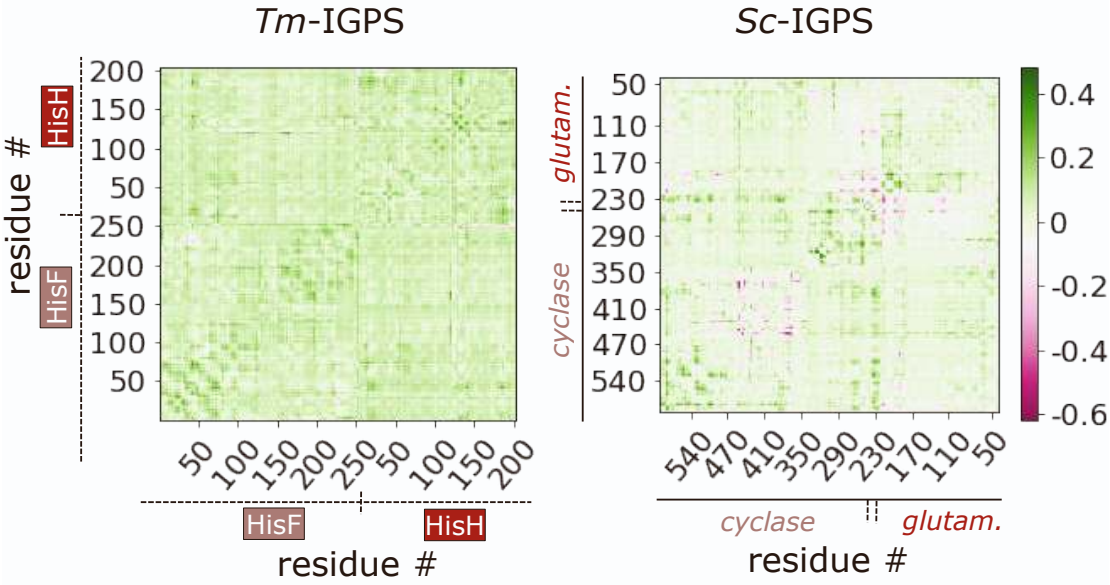


Figure S1. Difference of theoretical cross-correlation matrix between the holoenzyme and the apoenzyme in T. maritima (right) and S. cerevisiae (left). Cross-correlation matrices were computed with a Gaussian Network Model[18] using the pre-equilibrated structures of model 1 for apo and PRFAR-bound of T. maritima and S. cerevisiae. Kirchoff matrices build with a cutoff of 10 Å and a spring constant of 1. Only the first 20 modes were taken into account in the computation.

Correlation matrices

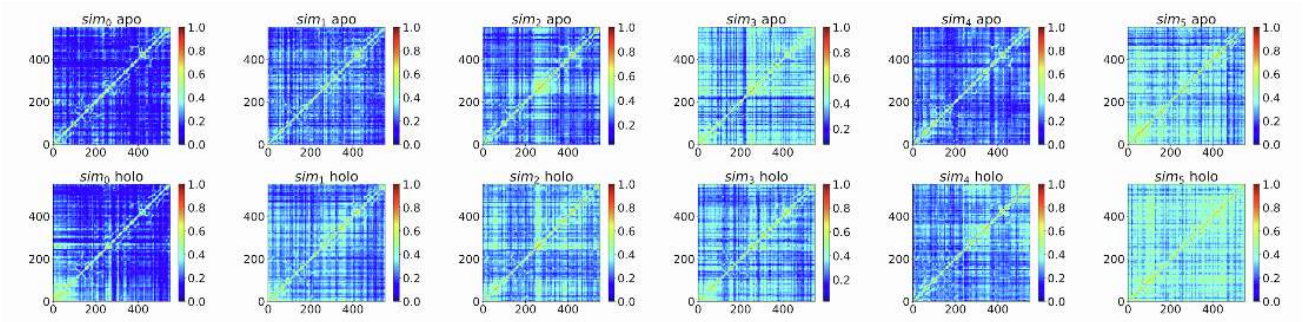


Figure S2. Generalized correlation coefficient matrices computed over six 100 ns replicas of simulated dynamics of apo and PRFAR-bound His7.

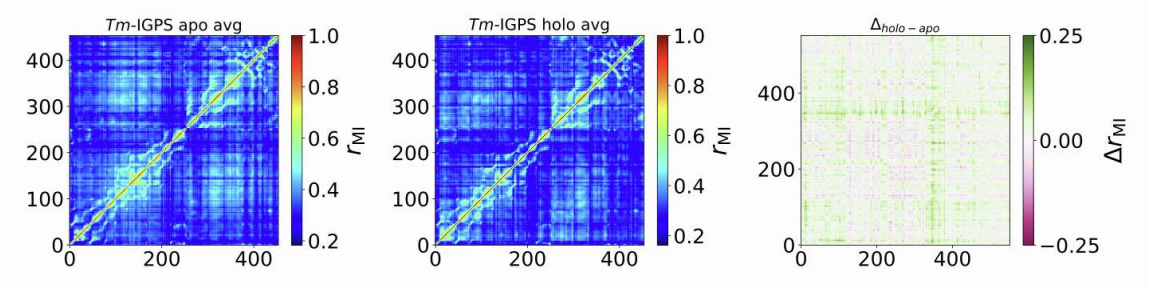


Figure S3. Left to right: average generalized correlation coefficient matrices over over apo, PRFAR-bound trajectories and difference (PRFAR-bound-minus-apo) computed over the six 100 ns replicas of simulated dynamics of *Tm*-IGPS.

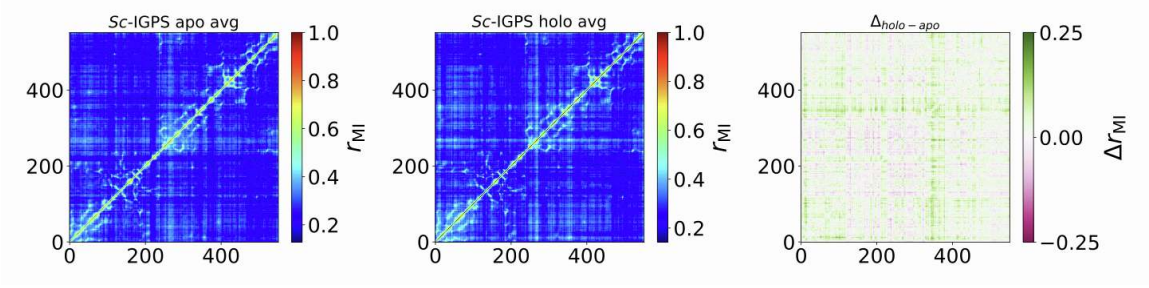


Figure S4. Left to right: average generalized correlation coefficient matrices over over apo, PRFAR-bound trajectories and difference (PRFAR-bound-minus-apo) computed over the six 100 ns replicas of simulated dynamics of His7.

Eigenvector centrality

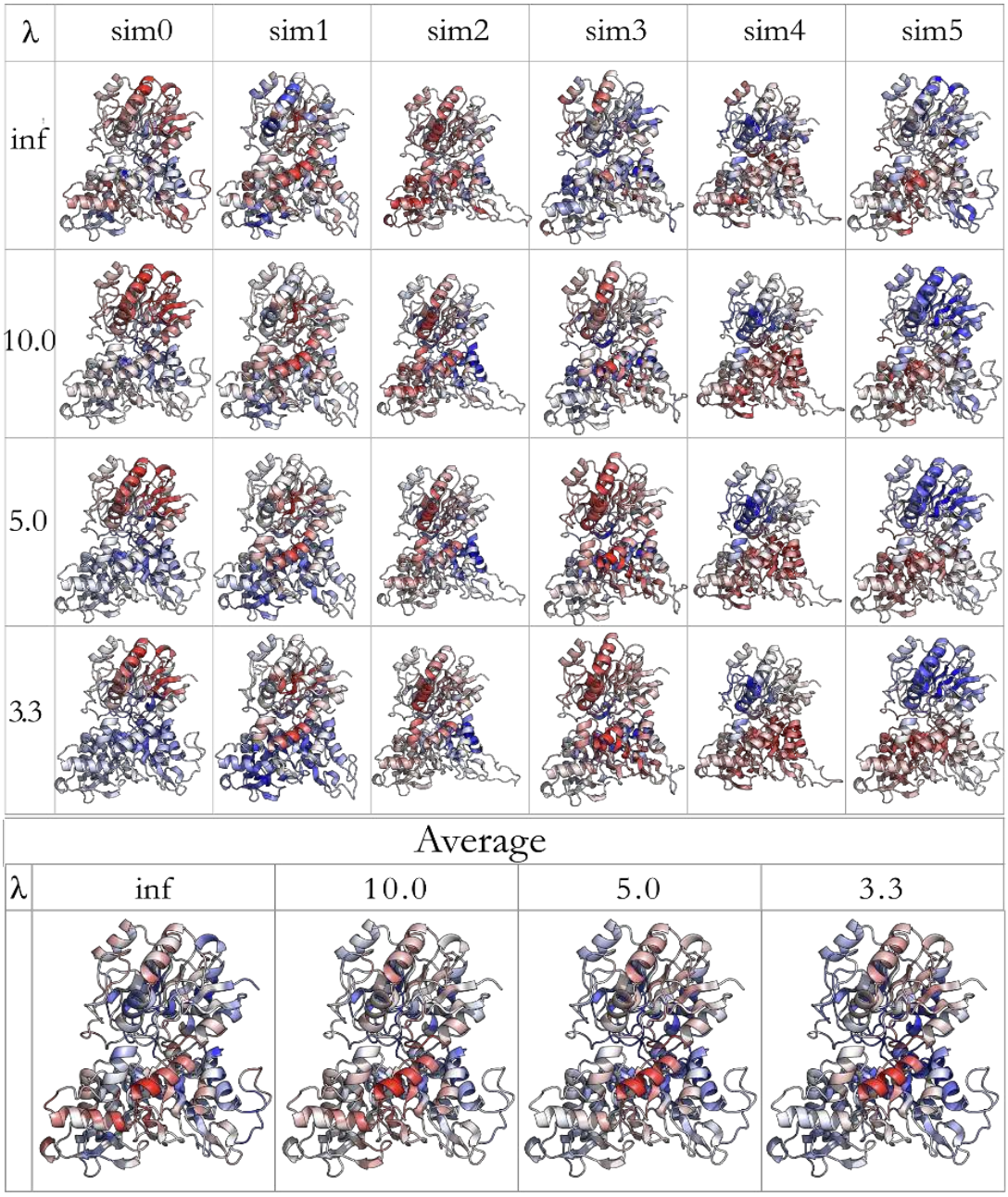


Figure S5. Centrality differences (PRFAR-bound-minus-apo) projected onto the apo structure of *Sc*-IGPS, computed different values of λ. Regions in red and blue correspond to gains and loss of centrality. To note, the EC values relative to sim_1 recover most of the allosteric traits as it can be inferred by the similarity of the centrality pictures showing the averaged values over the six independent replicas (last row).

Analysis of first and second principal components in apo and PRFAR-bound yeast and bacterial IGPS

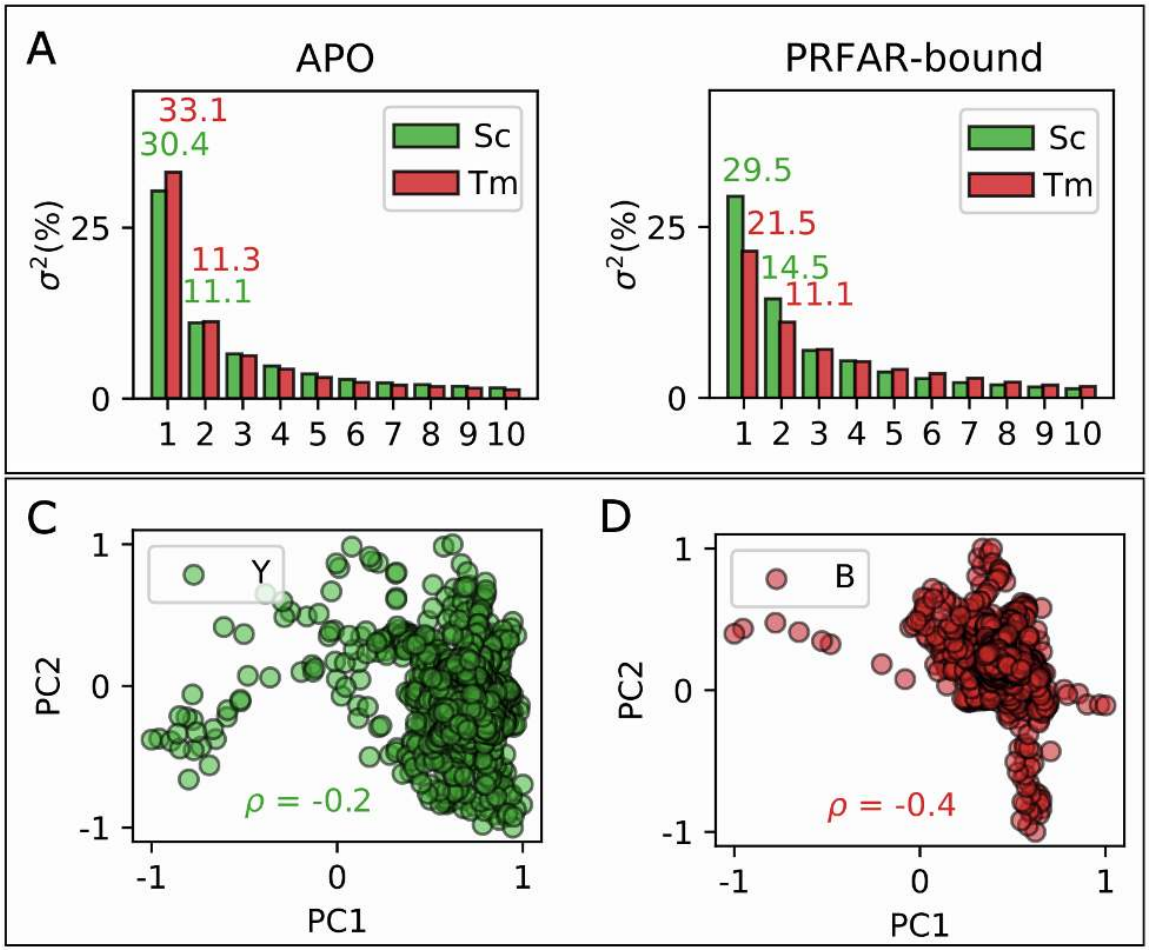


Figure S6. (A) and (B) Cumulative variance of the first and second PCs computed for the apo (A) and PRFAR-bound (B) trajectories, showing the comparison between *Tm*-IGPS (red) and *Sc*-IGPS (green). (C) and (D) show the correlation between first and second principal components computed along the trajectories of yeast and bacterial IGPS. (C) In *Sc*-IGPS PC1 and PC2 are poorly correlated, confirming that they account for distinct motions, while the higher correlation shown in (D) suggests that PC1 and PC2 in *Tm*-IGPS have some degree of overlap.

Essential motions of the trajectory through principal component analysis

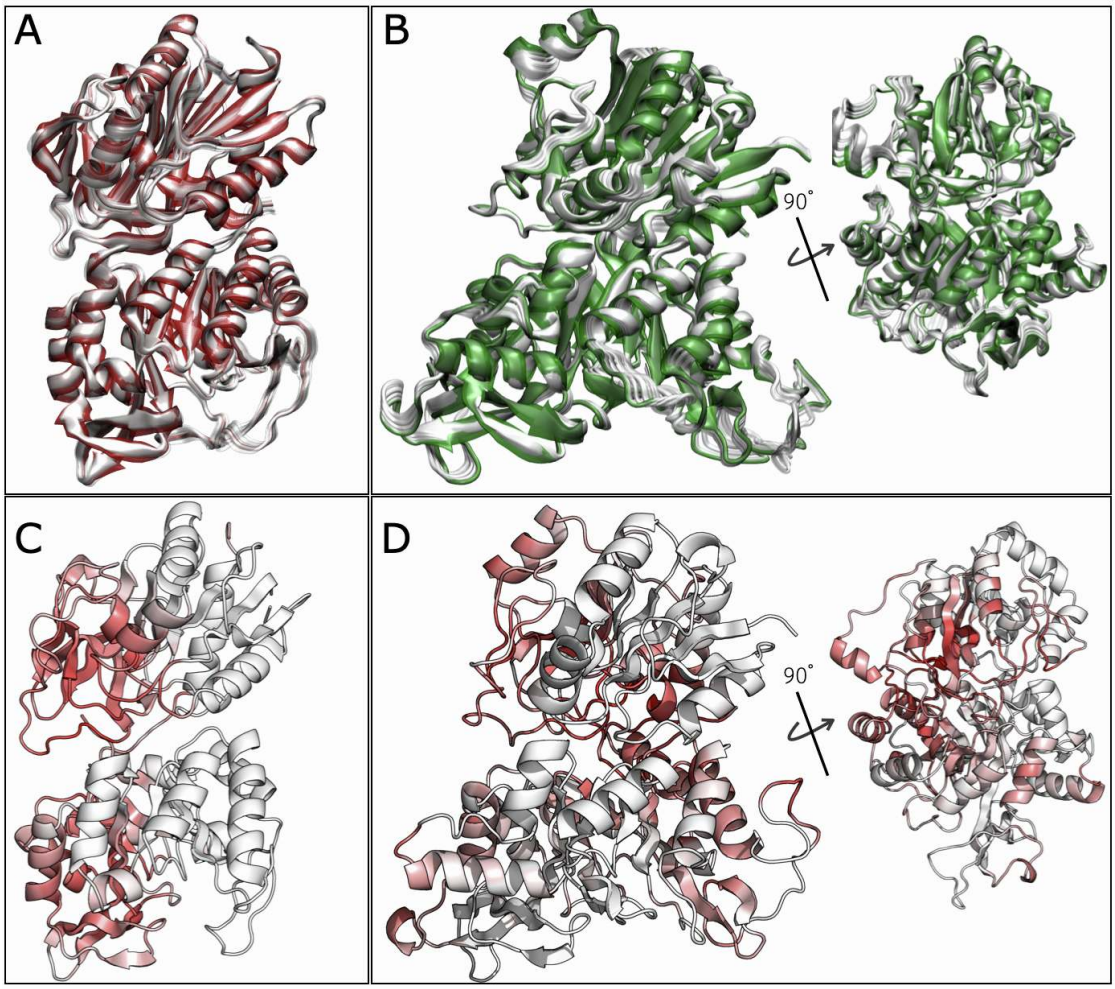


Figure S7. Projection of the original apo trajectory of *Tm*-IGPS (A) and *Sc*-IGPS onto the difference (PRFAR-minus-apo) second principal components computed along the yeast (ΔPC_1^Y) and bacterial (ΔPC_1^B) IGPS trajectories, as discussed in Section S2. This figure provides a zoom in of Figure 3E and 3F in the main text for better visualization of the dynamic low-vibrational motions of the two enzymes. Panels C and D show positive variations in the EC coefficients due to the long-range component of correlations in *Tm*-IGPS and His7 respectively. The largest increase in the long-range centrality coefficients upon PRFAR binding interests different regions in *Tm* and *Sc*. The values in *Tm* are consistent with the presence of an interdomain "breathing" motion shown with black dashed black lines and forming an angle φ). In *Sc,* the largest structural (long-range) rearrangements are associated with the motion of the connector and of the secondary structure elements *f*α8, *f*α1, *h*α4, *h*β9, marked in the figure. Long-range EC centralities match description of low vibrational motions provided by the analysis of first principal components.

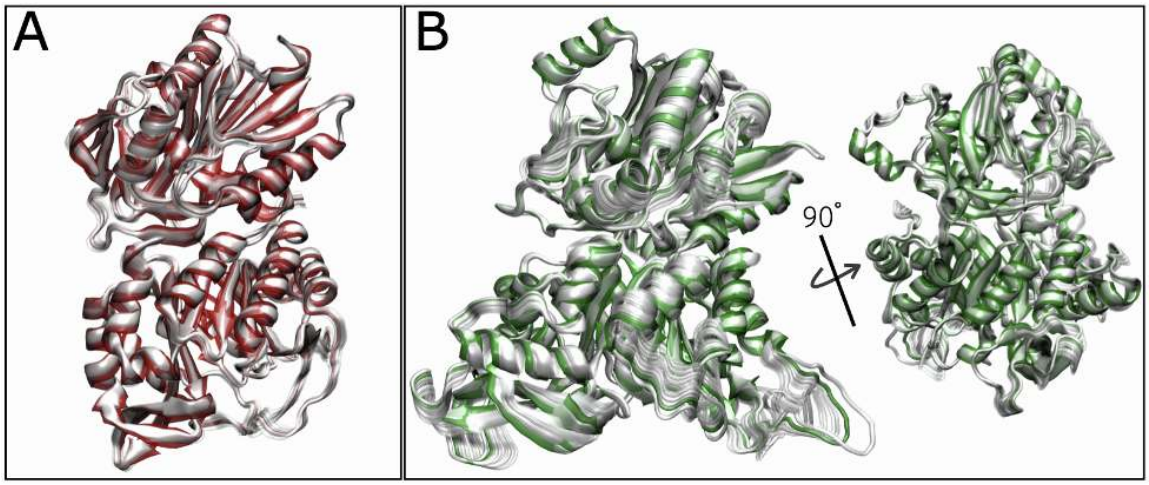


Figure 8. Projection of the apo trajectory of *Tm*-IGPS (A) and *Sc*-IGPS onto the difference (PRFAR-minus-apo) second principal components computed along the yeast (ΔPC_2^Y) and bacterial (ΔPC_2^B) IGPS trajectories, as discussed in Section S2.

Role of loop1 in Tm-IGPS and Sc-IGPS

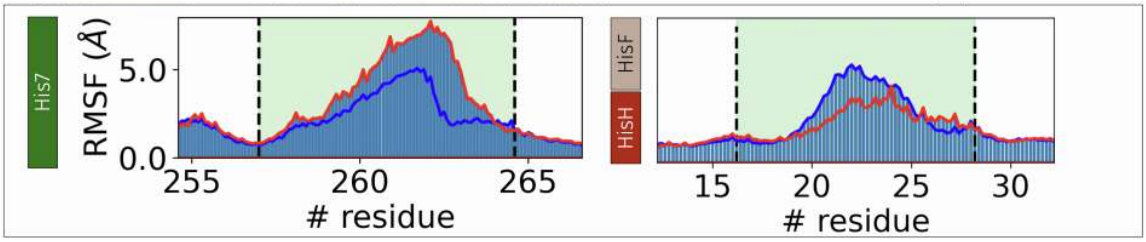


Figure S9. The motion of loop1 has a behavior in His7 and *Tm*-IGPS, upon binding of the effector. While in His7 the binding induces an increased mobility of loop1, in *Tm*-IGPS, binding of PRFAR constrains the motion of loop1. This behavior is consistent with the different role of the loop in the two systems, as suggested in the main text.

Distance profiles K334-D335 profile across six 100 ns replicas of simulated dynamics of His7.

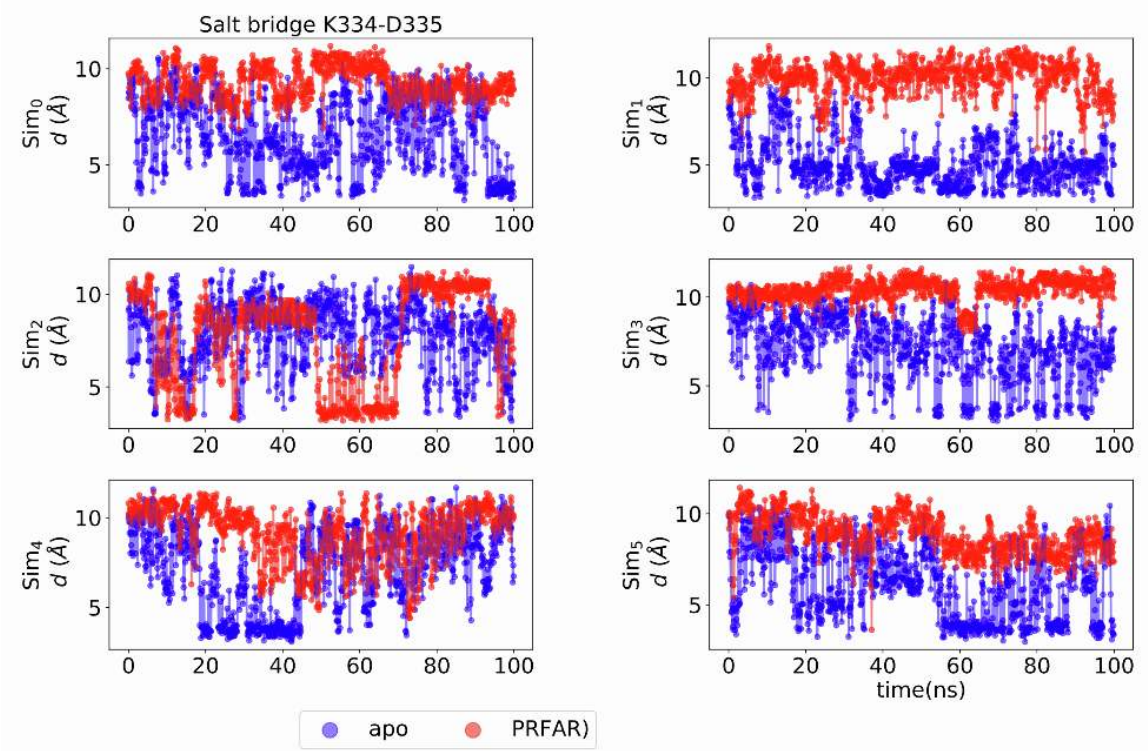


Figure S10. The K334-D335 salt bridge is mostly present in the APO simulation and breaks upon PRFAR binding as the effector interacts with residue K334. The dissolution of the K334-D335 is particularly evident in Sim_1, in accordance with our observation of Sim_1 best capturing the allosteric process.

Distance profiles of the hydrophobic cluster across six 100 ns replicas of simulated dynamics of His7.

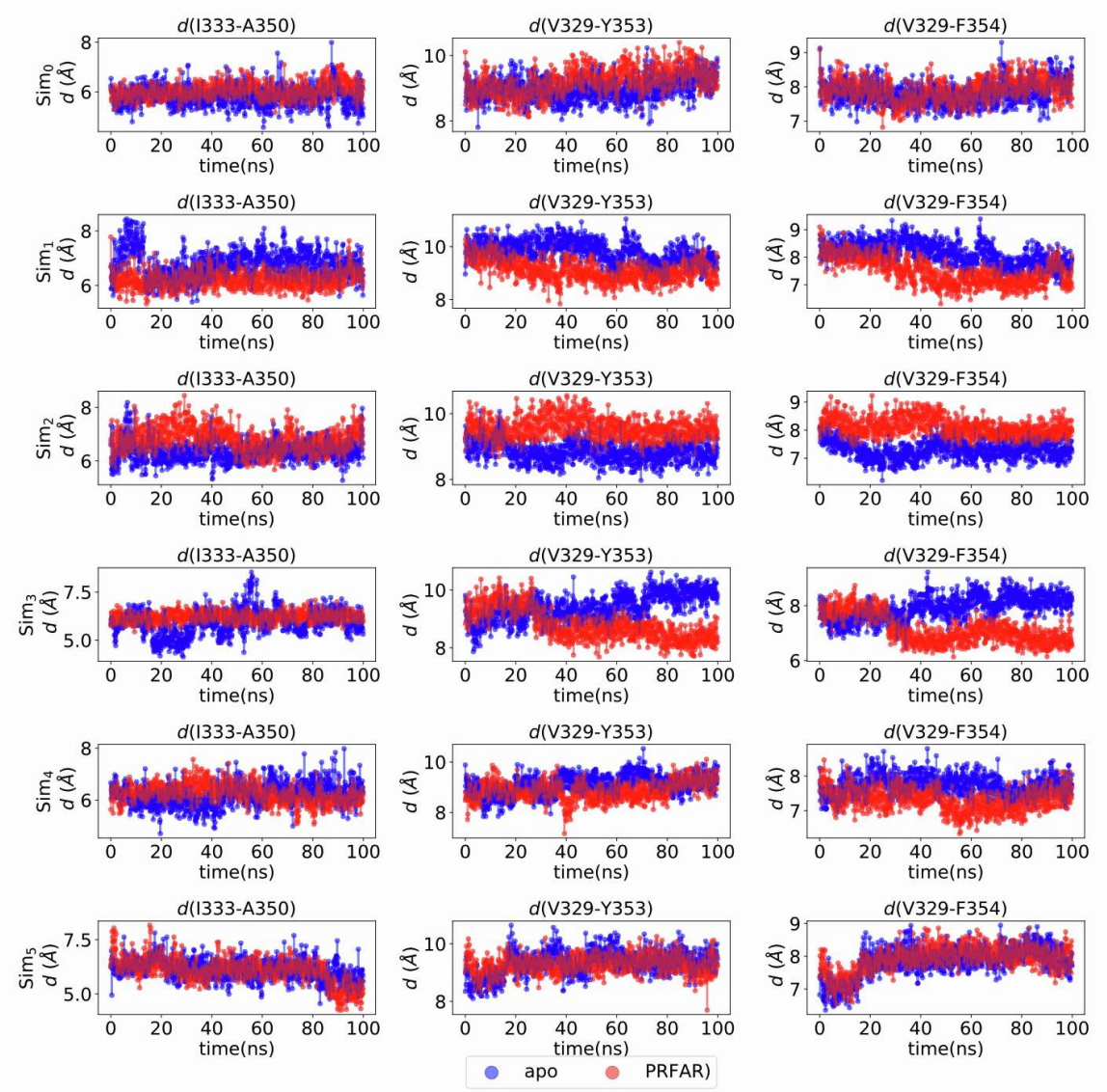


Figure S11. Distance profiles of I333-A350, V329-Y353, V329-F354 computed across six apo (blue) and PRFAR-bound (red) 100 ns replicas of simulated dynamics of His7.

Hinge Motion

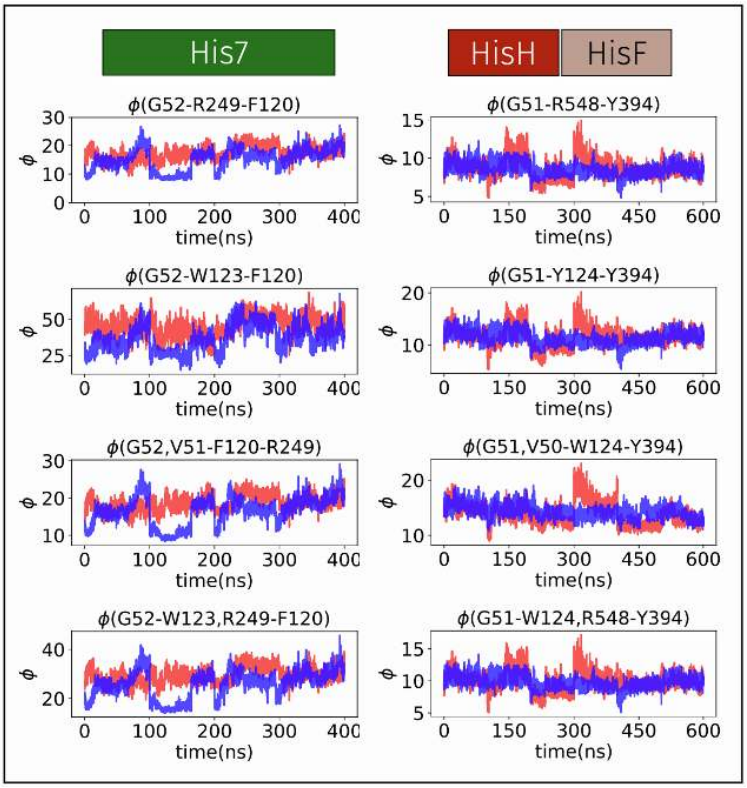


Figure S12. Hinge motion profile six apo (blue) and PRFAR-bound (red) over the concatenated dynamics of *Tm*-IGPS(left) and Hisy (right), measured through the angle ϕ defined using different residues. The standard definition of ϕ G51-W124-Y394) used in other publications is included. The oscillation mostly ranges between 10 and 20 degrees with no significant changes to the PRFAR bound profiles as compared to the apo.

The distribution of ϕ supports our hypothesis of a reduced importance of the hinge motion in the allosteric process of His7, as compared to that of *Tm*-IGPS.

Relevant salt-bridge interactions in Tm- and Sc-IGPS

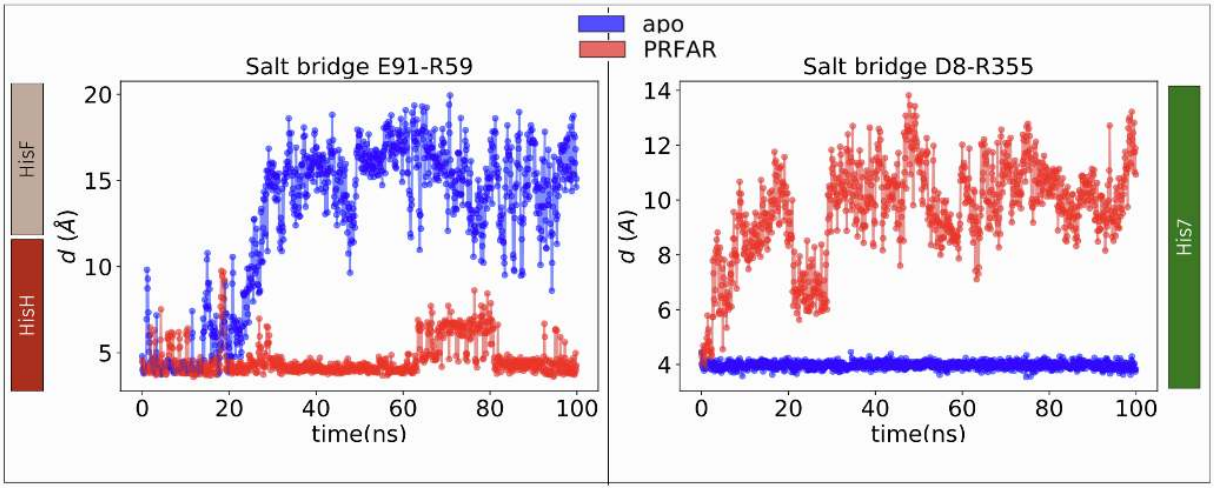


Figure S13. Profile of the E91-R59 (*Tm*-IGPS) and D8-R355 (His7) salt-bridge interactions along 100 ns of apo and PRFAR bound states of dynamics. As suggested by the large modification of the profiles upon effector binding these interactions are crucial in the signaling mechanism of bacterial and yeast IGPS, respectively.

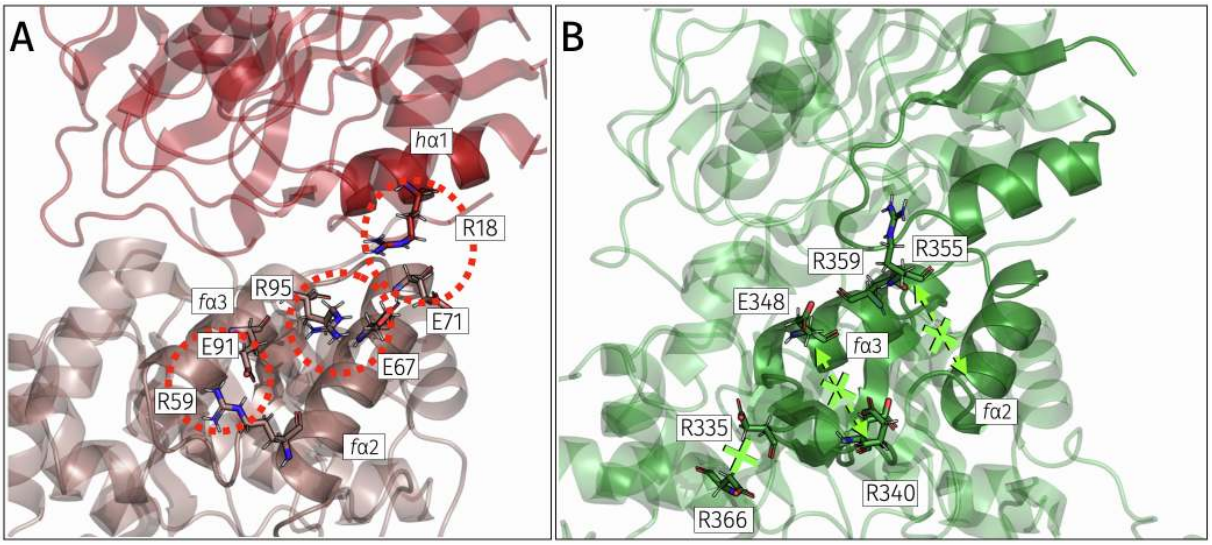


Figure S14. (A) Residues that participate in the network of salt-bridges at sideR of *Tm*-IGPS, induced by PRFAR binding. In *Sc*-IGPS (B), there are no corresponding surface-charged residues that can allow communication between the two active sites through the coupling of *f*α3-*f*α2 similar to that of *Tm*-IGPS. Instead, in *Sc*-IGPS the signal travels across *f*α3 and *f*β3.

Closeup view of the glutaminase active site and PGVG sequences in yeast and bacterial IGPS

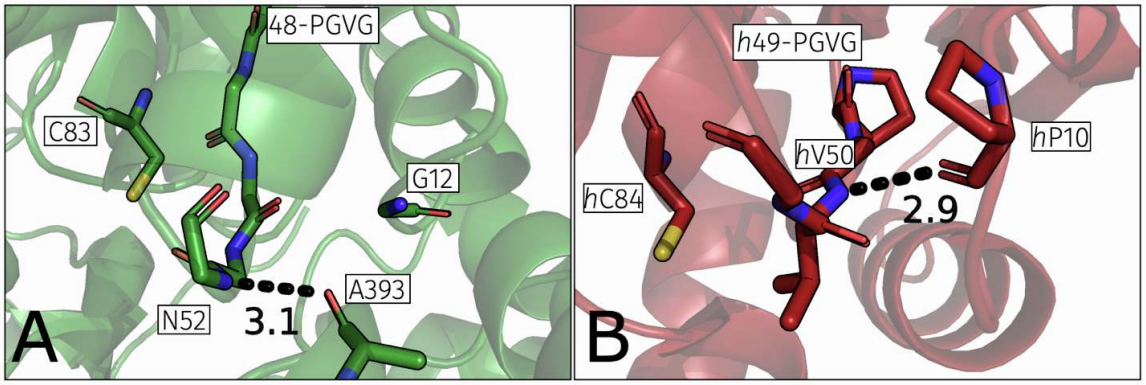


Figure S15. (A) The hydrogen bond between A393-N52 is mostly present in the apo structure and loosens upon PRFAR binding. Unlike in *Sc* apo, the *h*48-PGVG sequence *Tm*-IGPS is not within hydrogen-bonding distance to the glutaminase domain. (B) In *Tm* apo, residue *h*V50 is tightly bound in a hydrogen bond with *h*P10, while the corresponding distance varies significantly across the dynamics of *Sc*-apo, suggesting a different cross communication between the Ω-loop and PGVG in the two enzymes. The hydrogen bond between *h*V50 and *h*P10 in *Tm*-IGPS dissolves in the presence of PRFAR. This bond rupture marks the transition between the inactive state (apo) and the pro-active state.

Time-evolution of hydrogen bond at the interface of His7

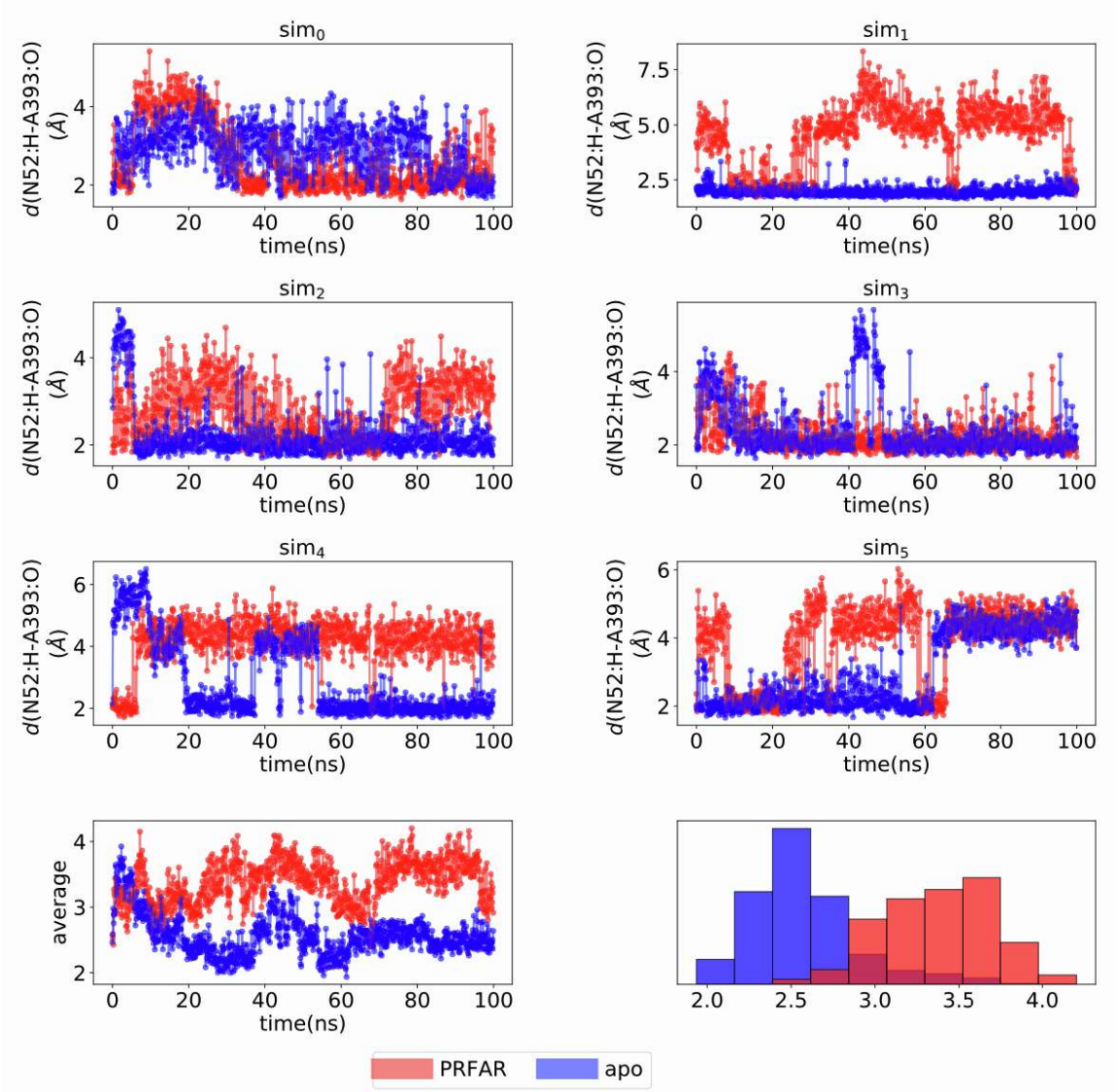


Figure S16. Distance profile of the N52:H-A393:O bond computed along the six replicas of the apo (blue) and PRFAR-bound (red) simulated dynamics of His7. The last row shows the mean values averaged on the different replicas, as well as a histogram representation of the same distribution. At the interface, the hydrogen bond between the backbone atoms of A393 and N52 elongates in the presence of the effector.

Movie legends

A way to investigate the essential motions of the trajectory is to project the original trajectory onto each of the principal components, to visualize the motion of the principal component. The resulting trajectories computed by projecting the original coordinates onto the first difference (ΔPC_1) and second (ΔPC_2) principal components are shown in the enclosed jupyter-notebook:

A way to investigate the essential motions of the trajectory is to project the original trajectory onto each of the principal components, to visualize the motion of the principal component. Instead of including all atoms of the trajectories one can focus on selected atom groups, for instance the backbone atoms.

The principal component analysis presented in this work is performed by selecting the backbone atoms of the apo and PRFAR-bound trajectories of either *Tm*-IGPS or *Sc*-IGPS.
The product of the weights $w_i(t)$ for the i^{th} principal component relative to the apo trajectory with the difference eigenvector $\Delta PC_i = PC_i^{PRFAR-bound} - PC_i^{APO}$ describes the fluctuations around the mean on that axis, induced by PRFAR binding,

$$r_i(t) = PC_i(t) \times u_i + r$$

The projected trajectory $r_i(t)$ is simply the fluctuations added onto the mean positions. (See description in section *Principal Component Analysis* at page 2).
The resulting trajectories computed by projecting the original coordinates onto the difference first (ΔPC_1) and second (ΔPC_2) principal components are shown in the three videos enclosed to the Supplementary material.

1- Video named DELTA_PC1.mov shows projected trajectories of *Tm*-IGPS and *Sc*-IGPS along the **first difference principal component ΔPC_1** (PRFAR-bound-minus-apo).

2- Video named DELTA_PC1rot.mov shows projected trajectories of *Tm*-IGPS and *Sc*-IGPS along the **first difference principal component ΔPC_1** (PRFAR-bound-minus-apo), where the Sc-IGPS is shown in a rotated view with respect to the Video 1, to highlight the motion of the connector.

3- Video named DELTA_PC2.mov shows projected trajectories of *Tm*-IGPS and *Sc*-IGPS along the **second difference principal component ΔPC_1** (PRFAR-bound-minus-apo).

The videos altogether show the difference in the dynamics of the low-vibrational motions of the two enzymes. While *Tm*-IGPS adopts a hinge-like breathing motion that modifies the opening of the interface between the two subunits, *Sc*-IGPS displays a rather different spring-like motion located at the core of the enzyme, coupled with large variations at the connector site and loop1.

Supporting References

(1) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric Pathways in Imidazole Glycerol Phosphate Synthase. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (22), E1428–E1436.

(2) Chaudhuri, B. N.; Lange, S. C.; Myers, R. S.; Davisson, V. J.; Smith, J. L. Toward Understanding the Mechanism of the Complex Cyclization Reaction Catalyzed by Imidazole Glycerolphosphate Synthase: Crystal Structures of a Ternary Complex and the Free Enzyme. *Biochemistry* **2003**, *42* (23), 7003–7012.

(3) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Protein Sci.* **2016**, *86*, 2.9.1–2.9.37.

(4) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296–W303.

(5) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049–1074.

(6) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *Journal of Computational Chemistry*. 2004, pp 1157–1174.

(7) Citations for Amber https://ambermd.org/CiteAmber.php (accessed Jan 7, 2021).

(8) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(9) Grubmüller, H.; Heller, H.; Windemuth, A.; Schulten, K. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-Range Interactions. *Mol. Simul.* **1991**, *6* (1-3), 121–142.

(10) Lange, O. F.; Grubmüller, H. Generalized Correlation for Biomolecular Dynamics. *Proteins* **2006**, *62* (4), 1053–1061.

(11) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; John Wiley & Sons, 2012.

(12) Yang, Z.; Algesheimer, R.; Tessone, C. J. A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Sci. Rep.* **2016**, *6*, 30750.

(13) Rivalta, I.; Batista, V. S. Community Network Analysis of Allosteric Proteins. *Methods Mol. Biol.* **2021**, *2253*, 137–151.

(14) Oldham, S.; Fulcher, B.; Parkes, L.; Arnatkevic Iūtė, A.; Suo, C.; Fornito, A. Consistency and Differences between Centrality Measures across Distinct Classes of Networks. *PLoS One* **2019**, *14* (7), e0220061.

(15) Negre, C. F. A.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector Centrality for Characterization of Protein Allosteric Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (52), E12201–E12208.

(16) Lange, O. F.; Grubmüller, H. Full Correlation Analysis of Conformational Protein Dynamics. *Proteins* **2008**, *70* (4), 1294–1312.

(17) David, C. C.; Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Methods Mol. Biol.* **2014**, *1084*, 193–226.

(18) Atilgan, A. R. *et al.* Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys J* **80**, 505–515 (2001).

.4 Supporting information to Manuscript 2: Temperature Increase Mimics Allosteric Signaling in Imidazole Glycerol Phosphate Synthase

Supplementary Information for

Temperature increase mimics allosteric signaling by PRFAR in imidazole-glycerol phosphate synthase

Florentina Tofoleanu,[*,1,2,a,†] Uriel Morzan,[*,1,3,a] Aria Gheeraert,[*,5,6] Apala Chaudhuri,[1] Zexing Qu,[1] Peter Nekrasov,[1] Bernard R. Brooks,[2] J. Patrick Loria,[1] Ivan Rivalta,[4,5,a] Victor S. Batista[1,a]

1. Department of Chemistry, Yale University, P.O. Box 208107, New Haven, CT, 06520-8107, USA
2. Laboratory of Computational Biology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892
3. The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy
4. Université de Lyon, École Normale Supérieure de Lyon, CNRS UMR 5182, Laboratoire de Chimie, 46 allée d'Italie, F69364 Lyon, France
5. Università degli Studi di Bologna, Viale del Risorgimento, 4I-40136, Bologna, Italy
6. Laboratory of Mathematics (LAMA), CNRS, University of Savoie Mont Blanc, France

[*]authors contributed equally to this work

[a] corresponding author

[†] current address: Novartis Institutes of BioMedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139

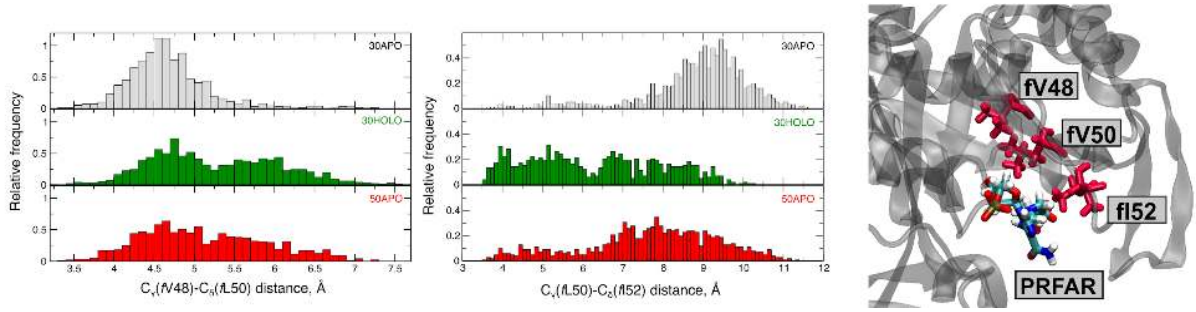


Figure S1. Histogram showing the relative frequencies of the distances between the γ-carbon of residue *f*V48 and the δ-carbon of residue *f*L50 (left) and the distances between the γ-carbon of residue *f*L50 and the δ-carbon of residue *f*I52 in apo30, holo30 and apo50 (middle). (right) Hydrophobic cluster composed of residues *f*V48, *f*L50, *f*I52 in red licorice close to the effector site and PRFAR.

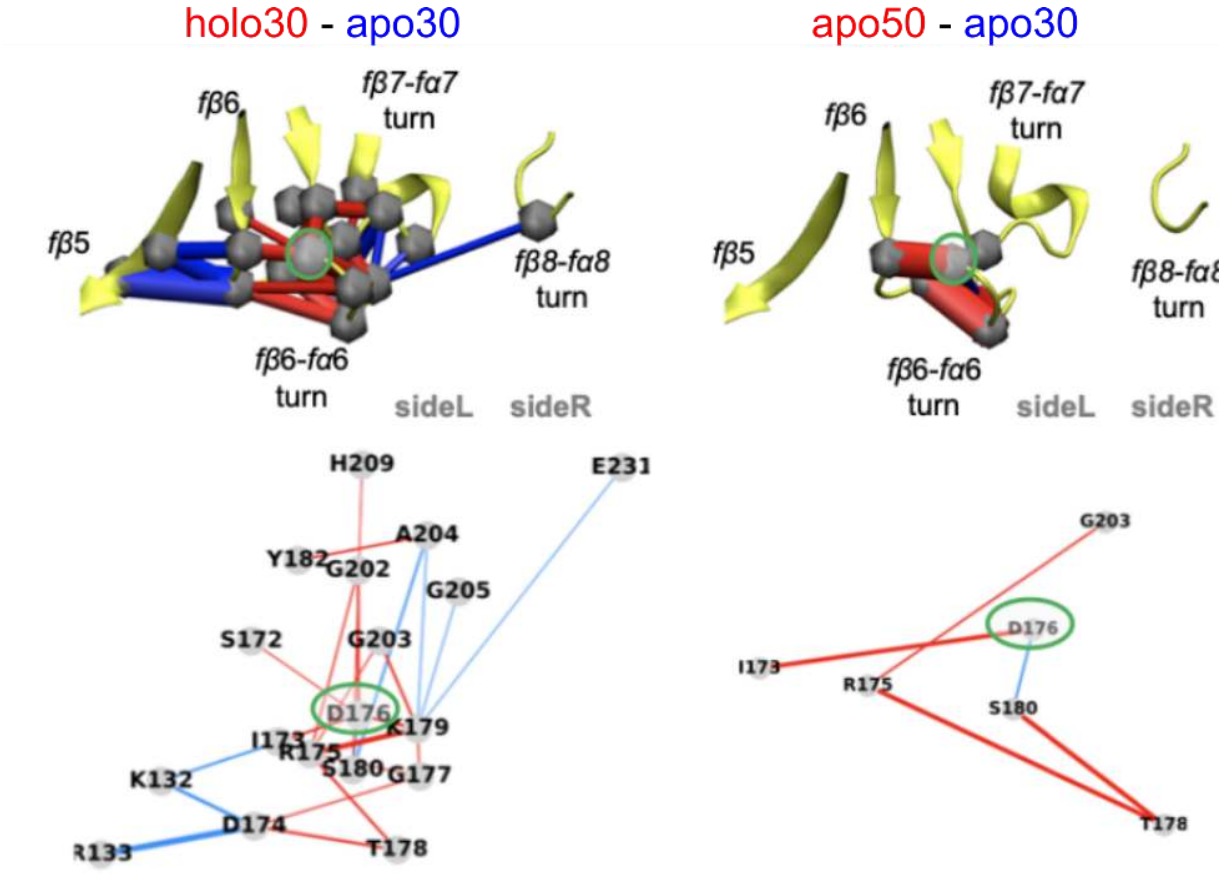


Figure S2. Induced perturbations for fD176 for the apo30/holo30 (left panel) and apo30/apo50 (right panel) perturbation networks. Blue and red edges represent a bigger number of contacts in the systems labeled with blue and red text, respectively. Edge widths are proportional to the number of contacts changes.

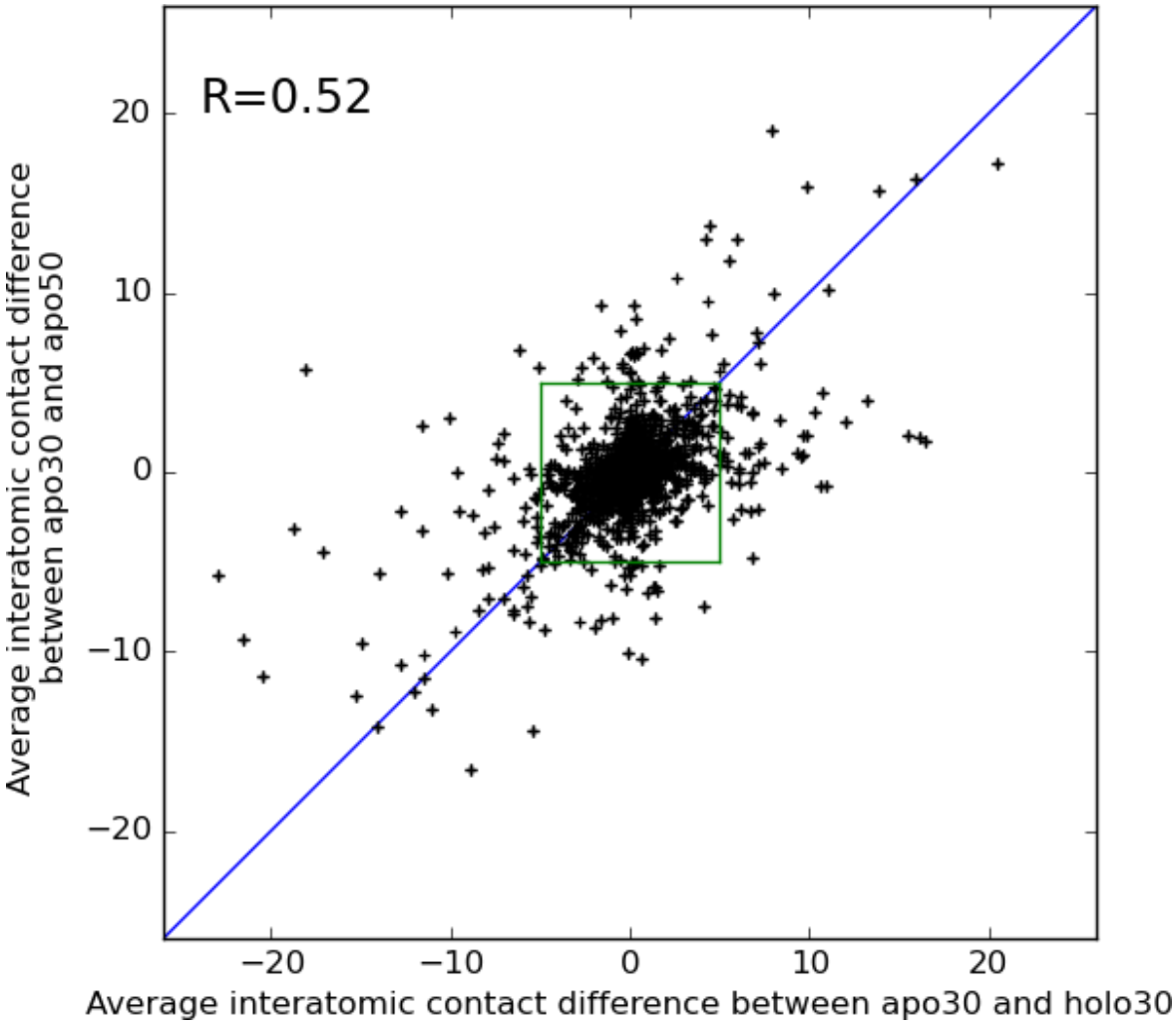


Figure S3 Correlation plot between the weight of edges in the DCPN between apo30 and holo30 and in the DPCN between apo30 and apo50. A blue line highlights the first bisector, (x=y) and a green box displays the threshold 5 limit.

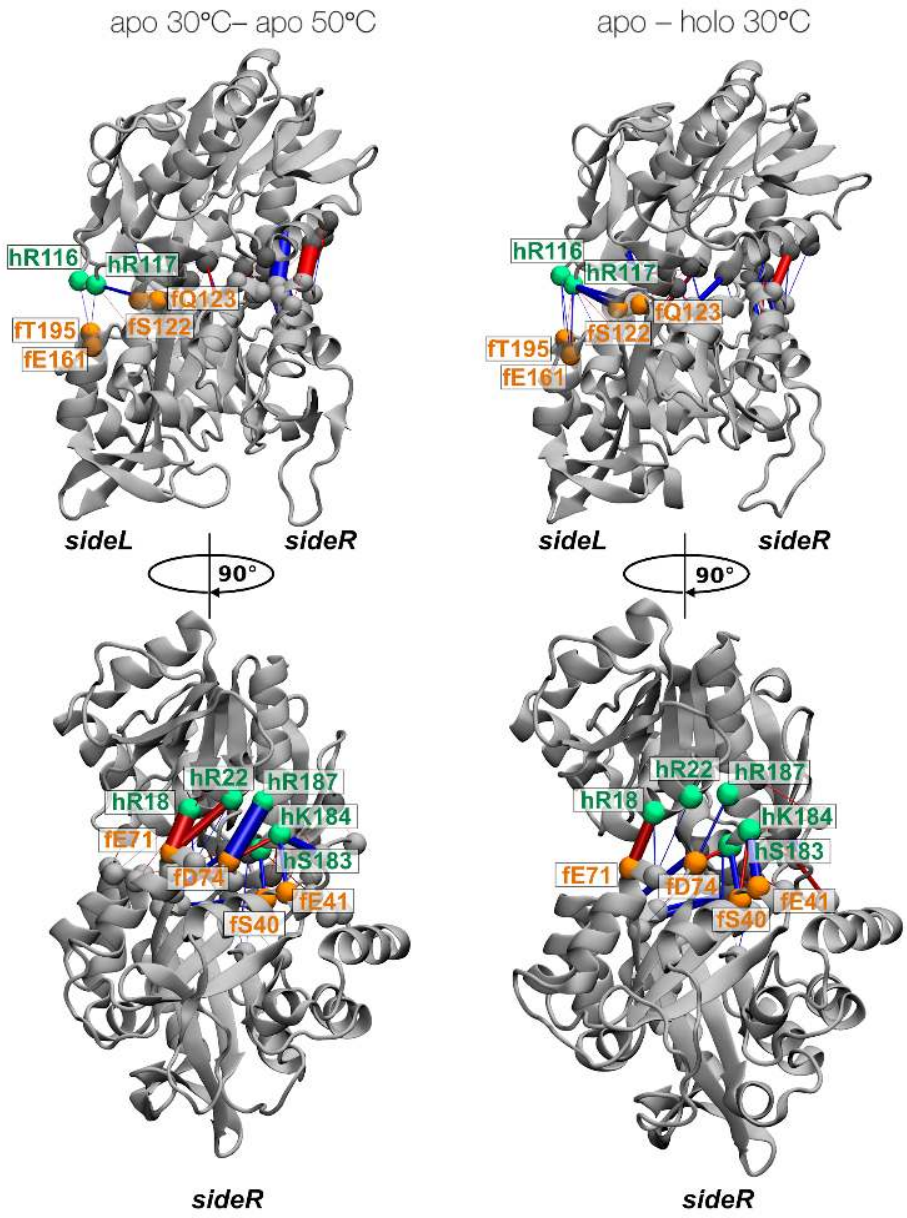


Figure S4. Hydrogen bonding at the hisF-hisH interface. The upper panel shows the HB network on sideL, bottom panel represents the HB networks at sideR. Residues that form strong HB in hisF and in hisH are represented as orange spheres, and green spheres, respectively. A blue cylinder indicates that the HBs were more persistent in the apo30 simulation, whereas a red cylinder indicates more HBs for the apo50 simulation, or the holo30 simulation.

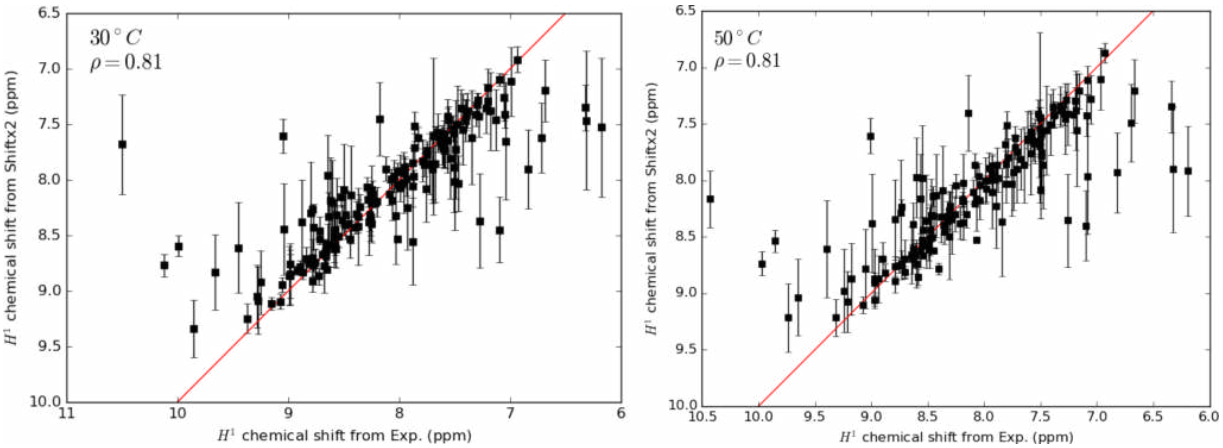


Figure S5. The correlation between SHIFTX2 and experimental for ^{1}H chemical shifts. SHIFTX2 chemical shifts computed at 30 °C (left) and 50 °C (right) on the corresponding trajectory versus experimental values obtained at the same temperature. Theoretical error bars are computed as the standard deviation of each chemical shift and experimental error bars are computed on four different calibrations of the experiment. Typical experimental error bars are too small to be visible (<1ppb).

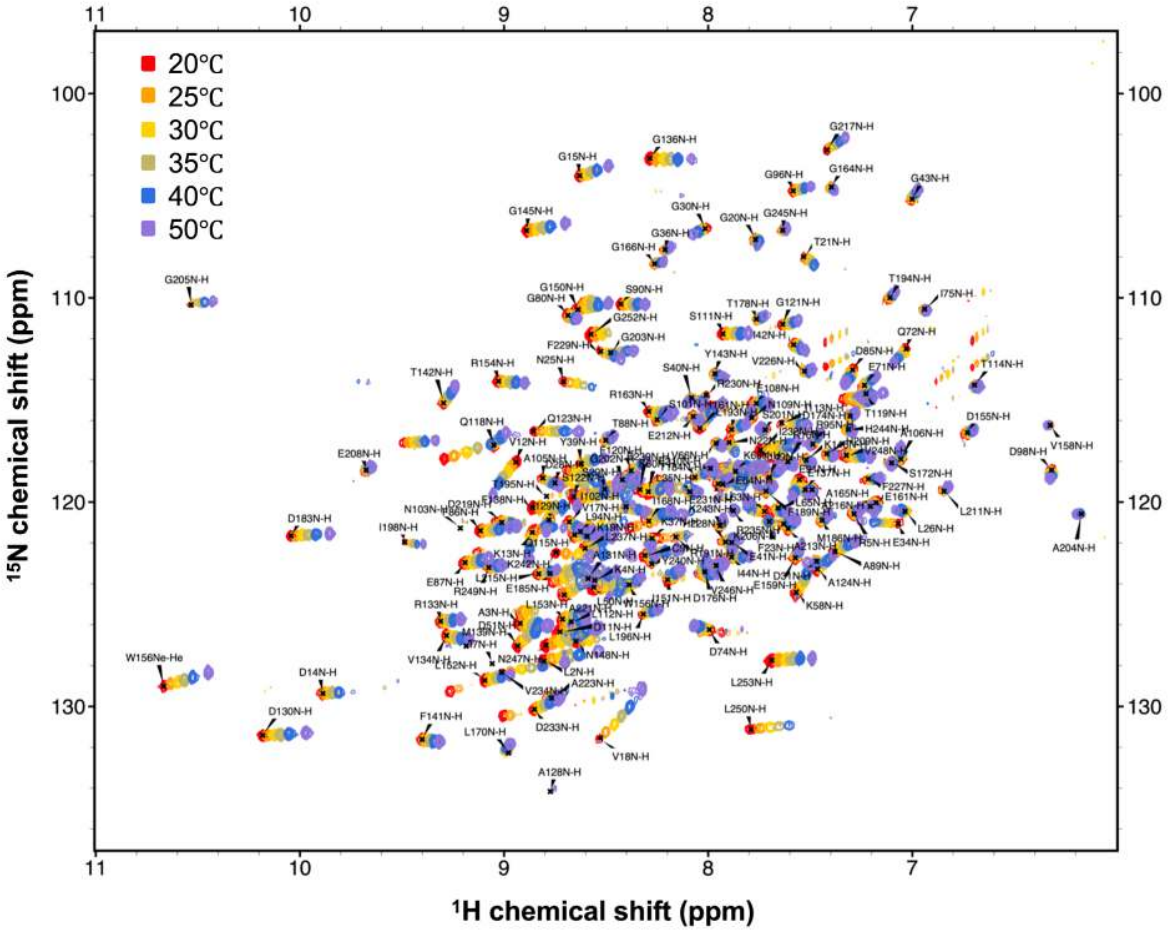


Figure S6. ^{1}H -^{15}N HSQC spectral overlay for the isotopically labelled hisF subunit in IGPS. Data was collected over a temperature range of 293-323 K (at 293, 298, 303, 308, 313 and 323 K), using a 600 MHz Varian spectrometer. A 500 μM sample of IGPS at pH 7.3 was used with DSS as an internal standard.

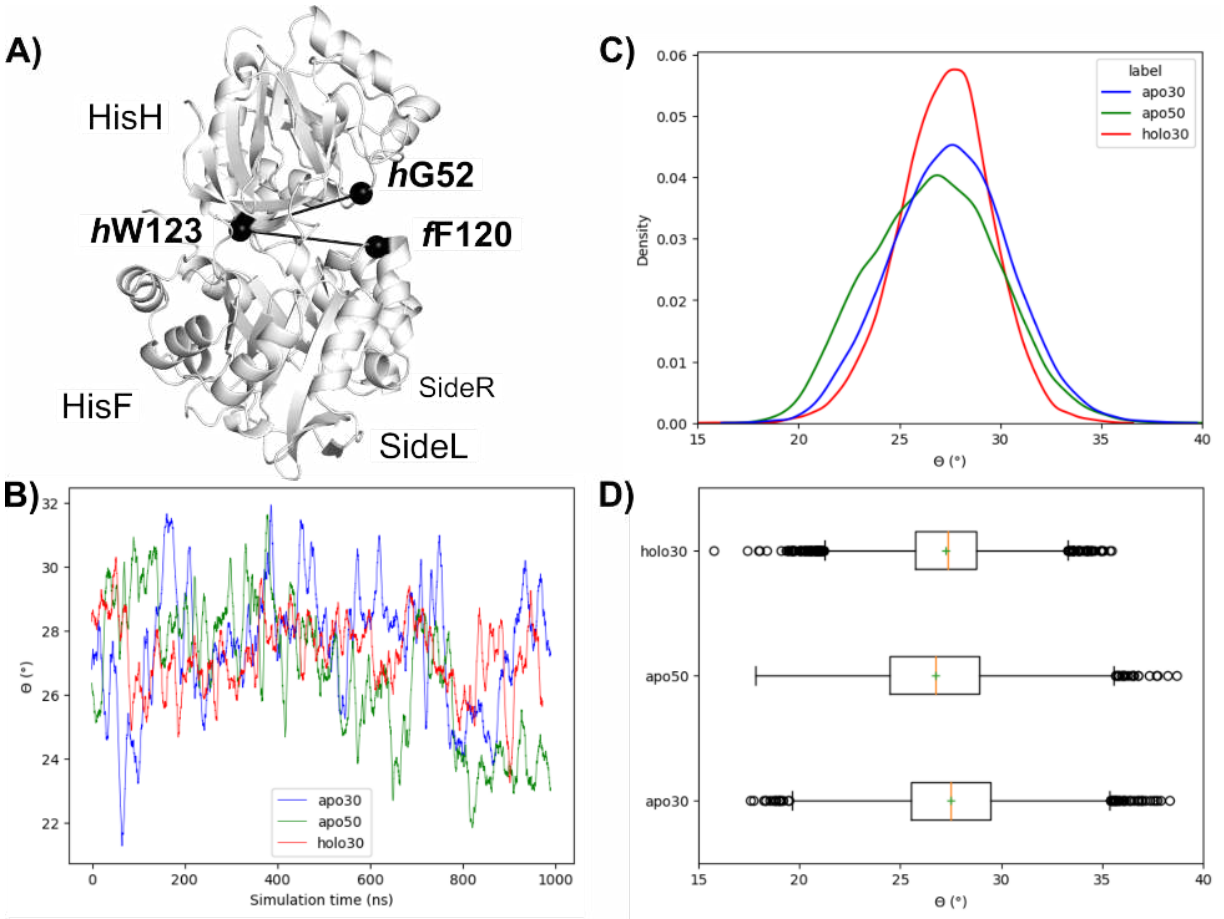


Figure S8. (A) Instantaneous representation of the breathing motion angle (between Cα atoms of *f*F120, *h*W123, *h*G52) during the first frame of the apo30 simulation. (B) Breathing motion angle evolution during the 1μs MD simulation in apo30 (blue), apo50 (green) and holo30 (red) with moving average with a time window of 100 frames=10ns. (C) Kernel density estimate of the distribution of breathing motion angle in each trajectory. (D) Boxplot representing the distribution of breathing motion angle in each trajectory. The mean is represented with a green cross and the median in orange line.

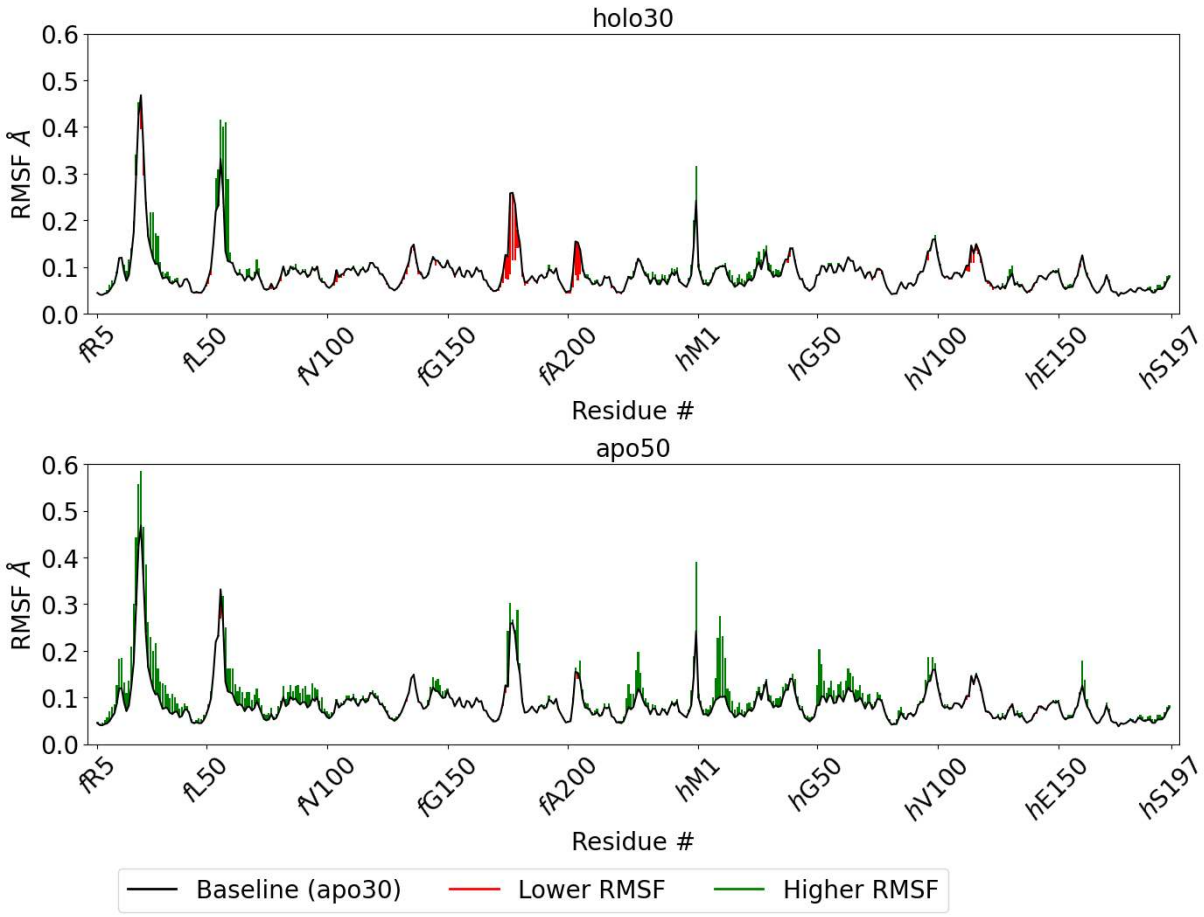


Figure S9. RMSF of apo30 (black lines, top and bottom) compared to holo30 (top) and apo50 (bottom) with green upward bars if the RMSF is bigger and red downward bars if the RMSF is lower. Terminal segments are excluded.

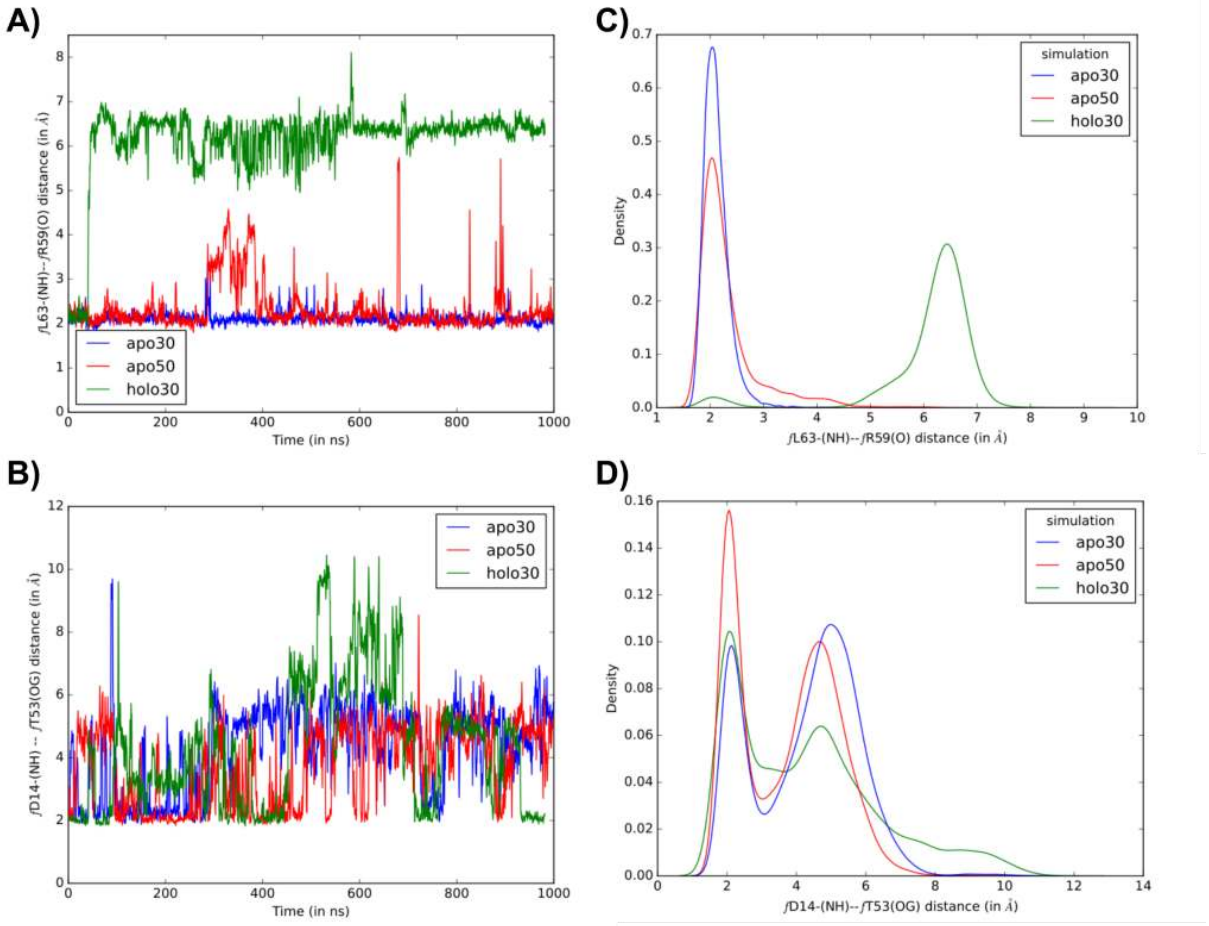


Figure S10. Hydrogen bond distance (A) along the 1μs simulation of apo30 (blue) and apo50 (red) between residues *ƒ*L63-*ƒ*R59 (A) and *ƒ*D14-*ƒ*T53 (B). Kernel density estimate of the length of the respective same hydrogen bonds (C and D).

.5 Supporting information to Submitted Article 1: Singular Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact Perturbation Analysis

Supporting Information to Singular Interface Dynamics of the SARS-CoV-2 Delta Variant Explained with Contact Perturbation Analysis

Aria Gheeraert,[†,‡] Laurent Vuillon,[†] Laurent Chaloin,[¶] Olivier Moncorgé,[¶] Thibaut Very,[§] Serge Perez,[‖] Vincent Leroux,[⊥] Isaure Chauvot de Beauchêne,[⊥] Dominique Mias-Lucquin,[⊥] Marie-Dominique Devignes,[⊥] Ivan Rivalta,[*,‡,#] and Bernard Maigret[*,⊥]

†*LAMA, Univ. Savoie Mont Blanc, CNRS, LAMA, 73376 Le Bourget du Lac, France*

‡*Dipartimento di Chimica Industriale "Toso Montanari", Università degli Studi di Bologna, Viale del Risorgimento 4, I-40136 Bologna, Italy*

¶*Institut de Recherche en Infectiologie de Montpellier (IRIM), Univ. Montpellier, CNRS, 34293 Montpellier, France*

§*CNRS - IDRIS, rue John von Neumann BP 167 91403 Orsay cedex - France*

‖*University Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France*

⊥*University of Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

#*ENSL, CNRS, Laboratoire de Chimie UMR 5182, 46 allée d'Italie, 69364 Lyon, France*

E-mail: i.rivalta@unibo.it; bernard.maigret@loria.fr

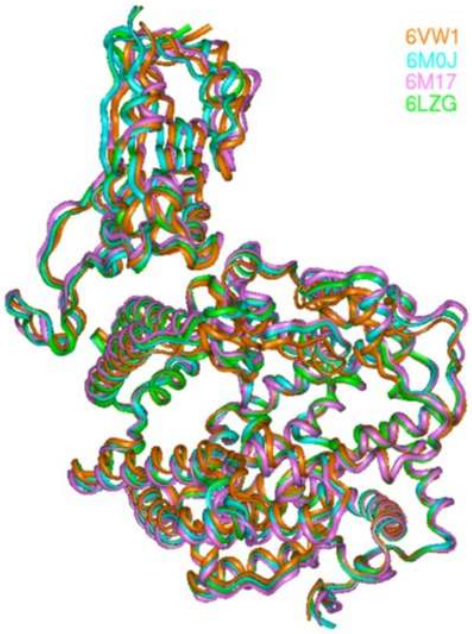


Figure S1. Superposition of four RBD/ACE2 complexes found in the PDB (PDB: GVW1, 6M0J 6M17 and 6LZG)

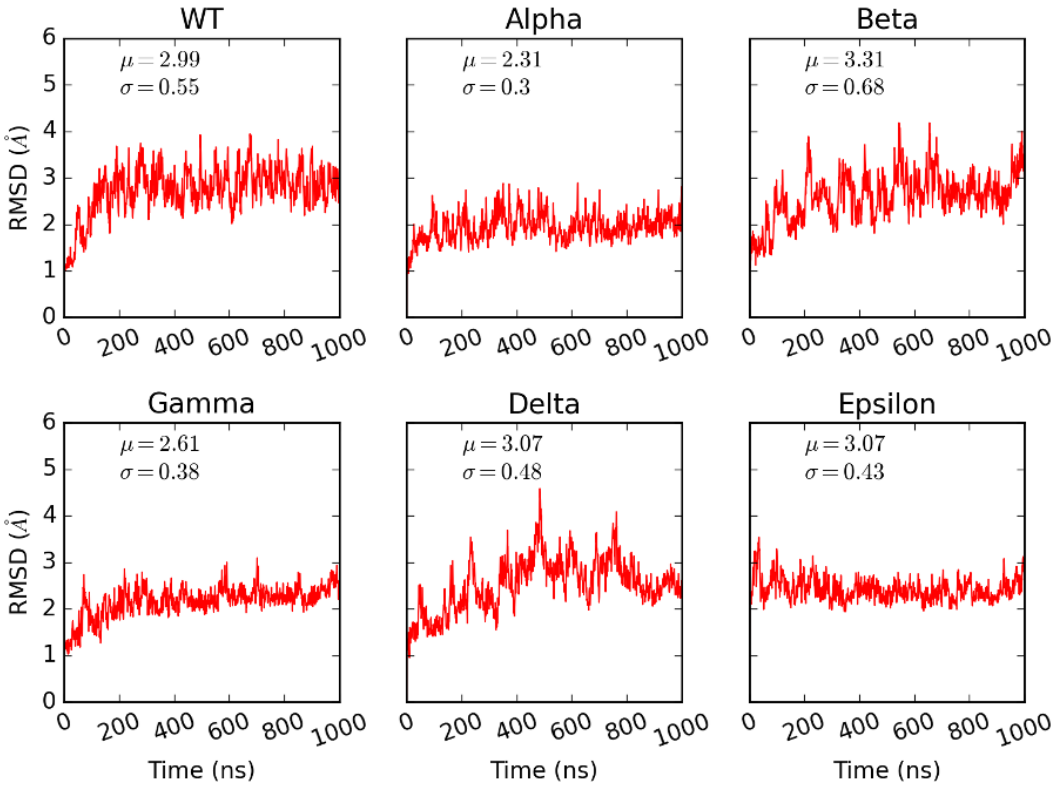


Figure S2. RMSD fluctuations during the microsecond MD simulations obtained for the wild type and five different variants of RBD/ACE2 complexes. For each complex, the RMSD were calculated only for the backbone atoms and excluding terminal loops (T27-D597 for ACE2 and S325-N540 for spike-RBD) and using the first frame of the simulation as reference.

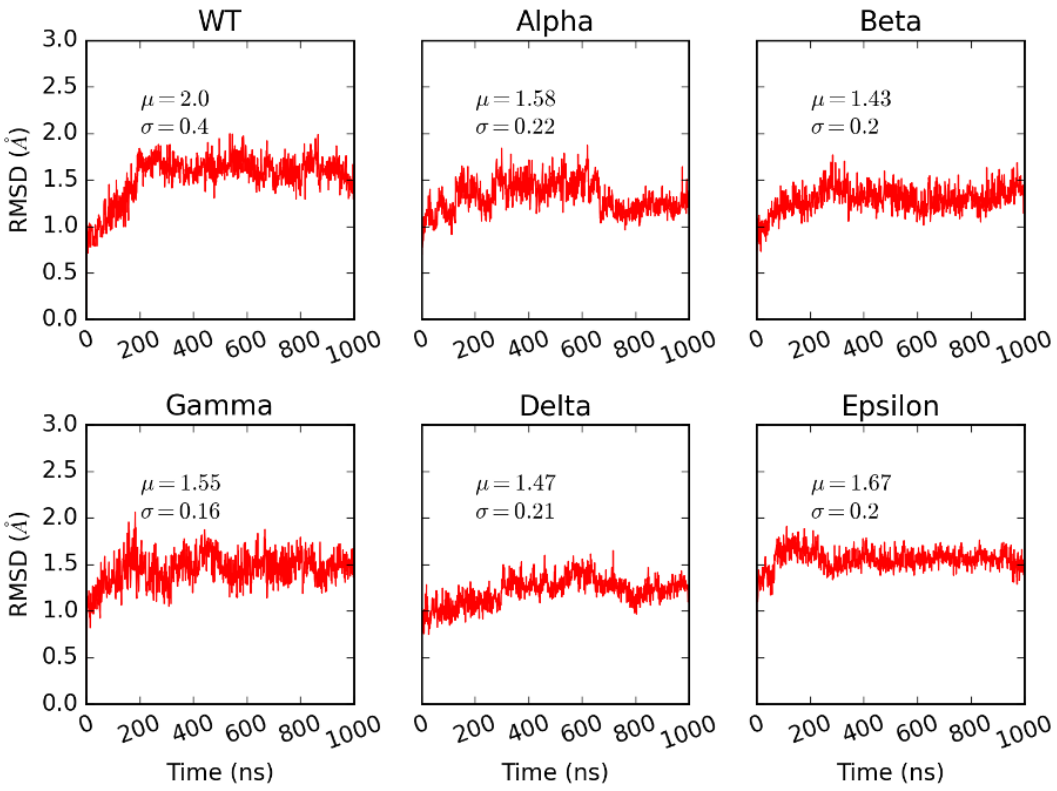


Figure S3. RMSD fluctuations during the microsecond simulations obtained for the six complexes. In each complex, the RMSD were calculated only for the backbone atoms of the RBD excluding terminal loops (residues S325-N540) and the initial frame as reference.

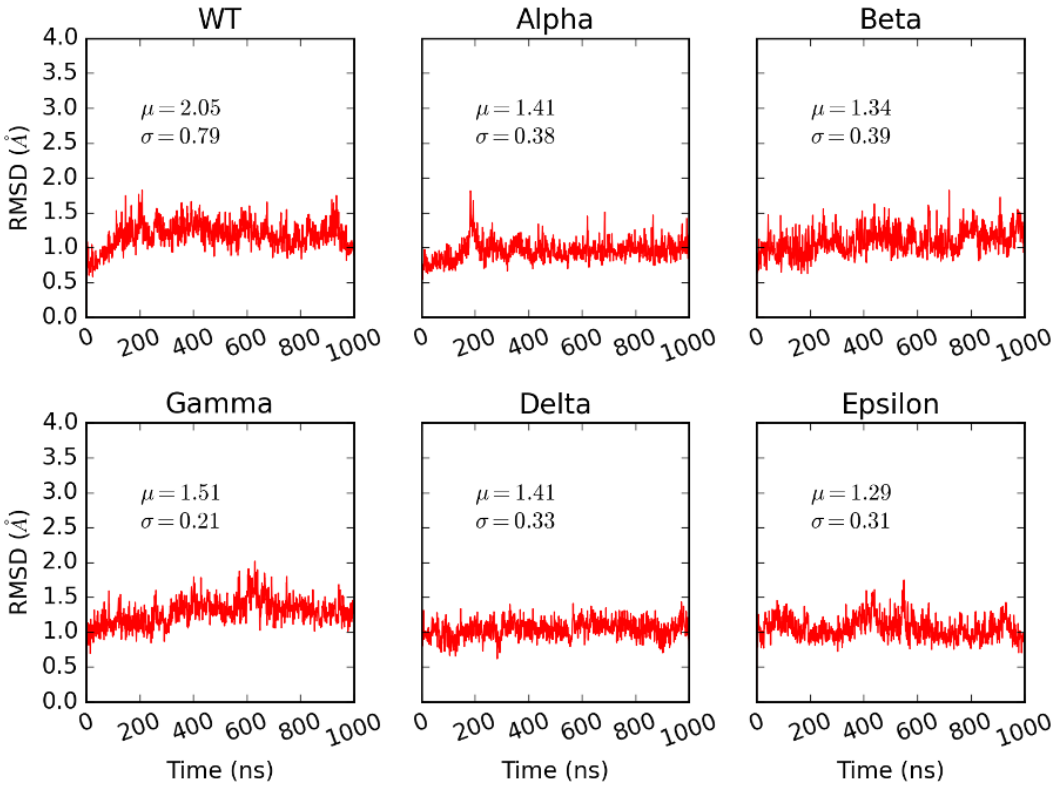


Figure S4. RMSD fluctuations during the microsecond simulations obtained for the six complexes. In each complex, the RMSD were calculated only for the backbone atoms of the RBM (residues S438-Q506) and using the initial frame as reference.

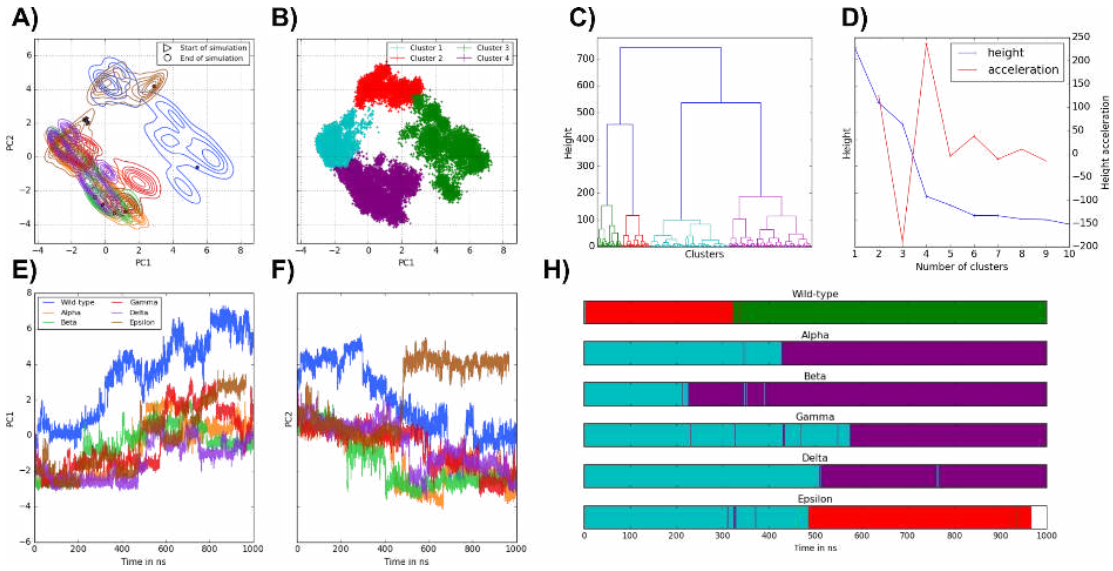


Figure S5. Projection of the frames corresponding to the microsecond simulations for the six studied complexes in the two dPCA eigenvector dimensions with (A) terrain lines representing a kernel density estimate of the population of each complex, (B) scatter plot representing the three main clusters obtained through Ward's minimum variance method. (C) Hierarchy obtained through Ward's minimum variance method and (D) acceleration plot displaying an optimal number of clusters equal to four. Time-plot of the (E) PC1 and (F) PC2 during each simulation. (H) Time evolution of each simulation in the different clusters.

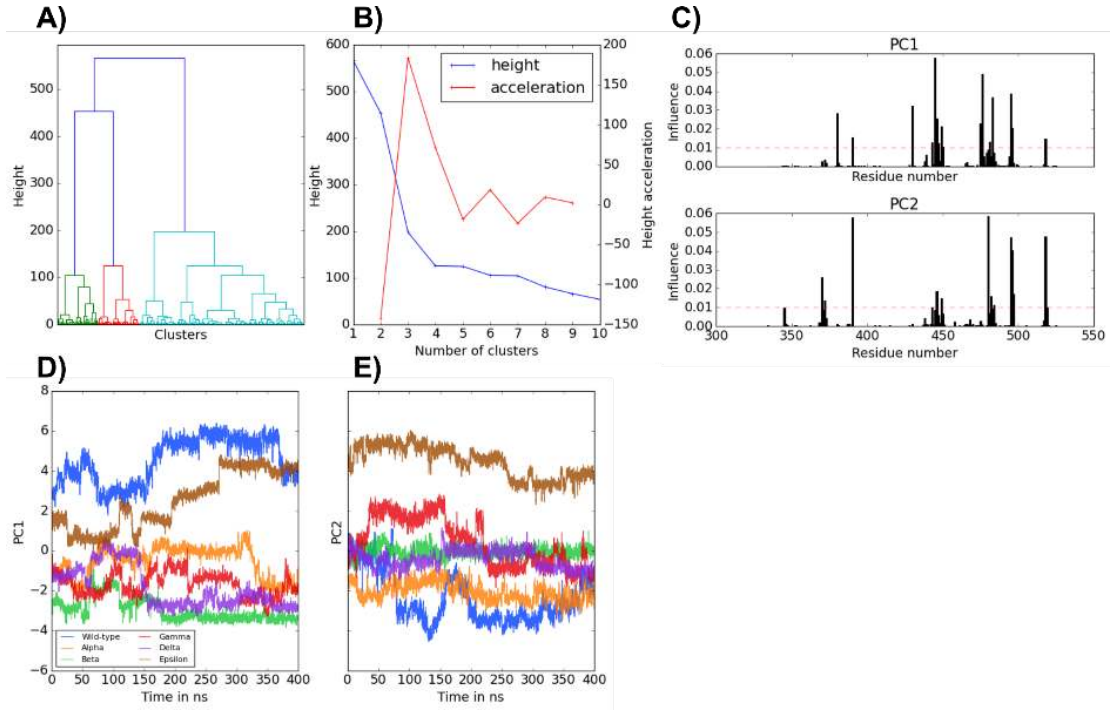


Figure S6. (A) Hierarchy obtained through Ward's minimum variance method and (B) acceleration plot displaying an optimal number of clusters equal to three. (C) Influence of each pair of consecutive residues in the PC1 and PC2. Time-plot of the (D) PC1 and (E) PC2 during each simulation.

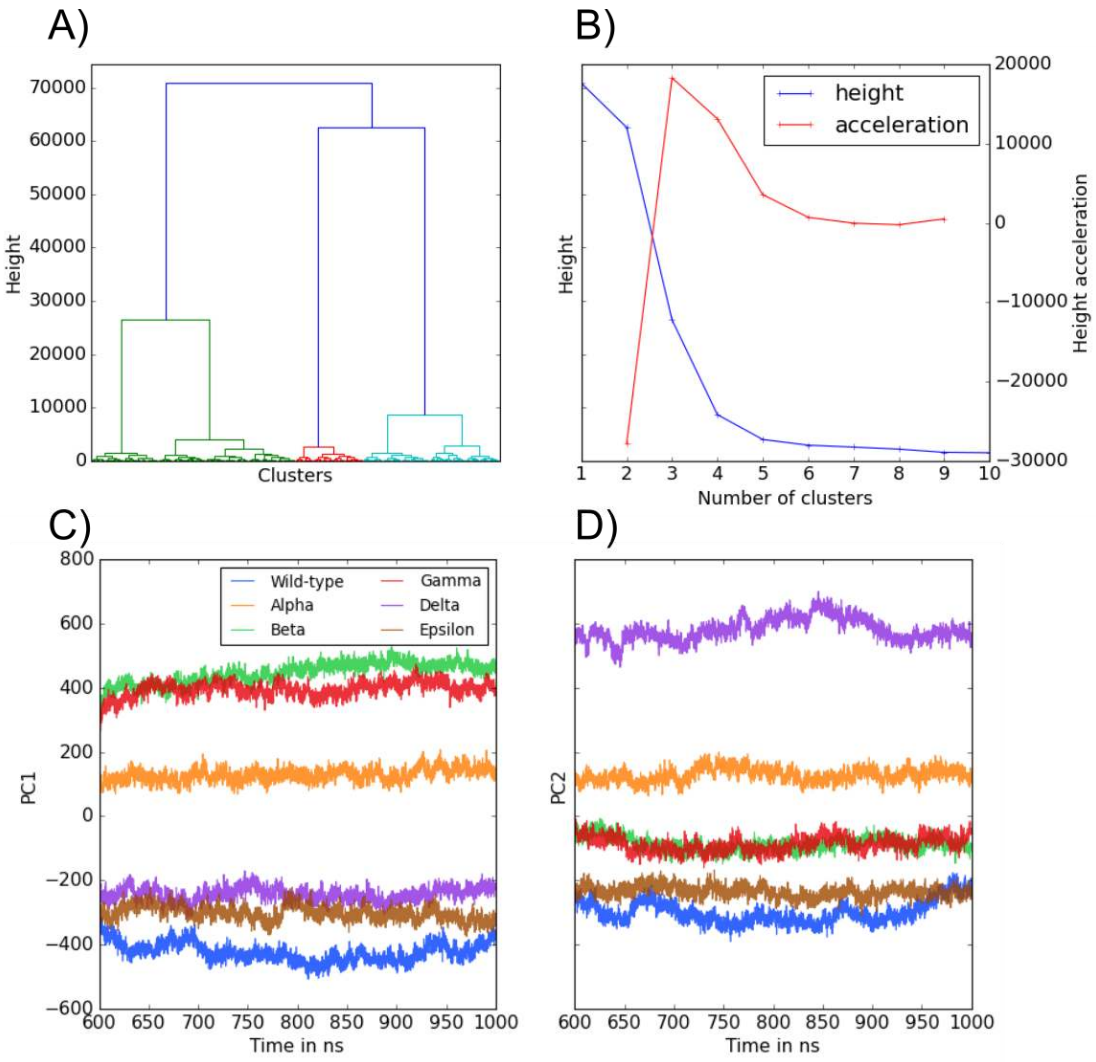


Figure S7. (A) Hierarchy obtained through Ward's minimum variance method and (B) acceleration plot displaying an optimal number of clusters equal to three. Time-plot of the PC1 (C) and PC2 (D) during the last 400 ns in each simulation.

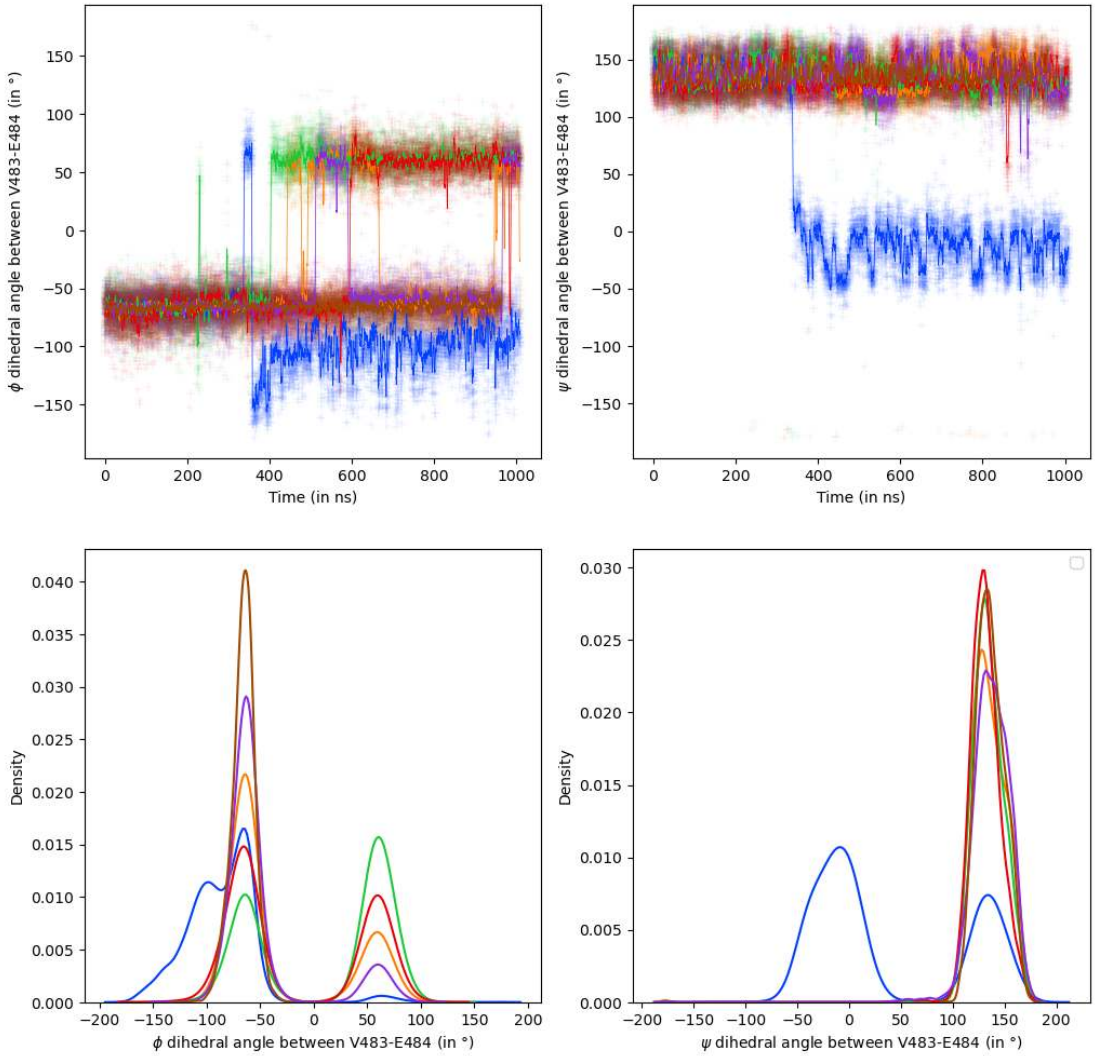


Figure S8. Time-evolution (left) and density (bottom) of the φ (left) and ψ (right) dihedral angles between V483 and E484.

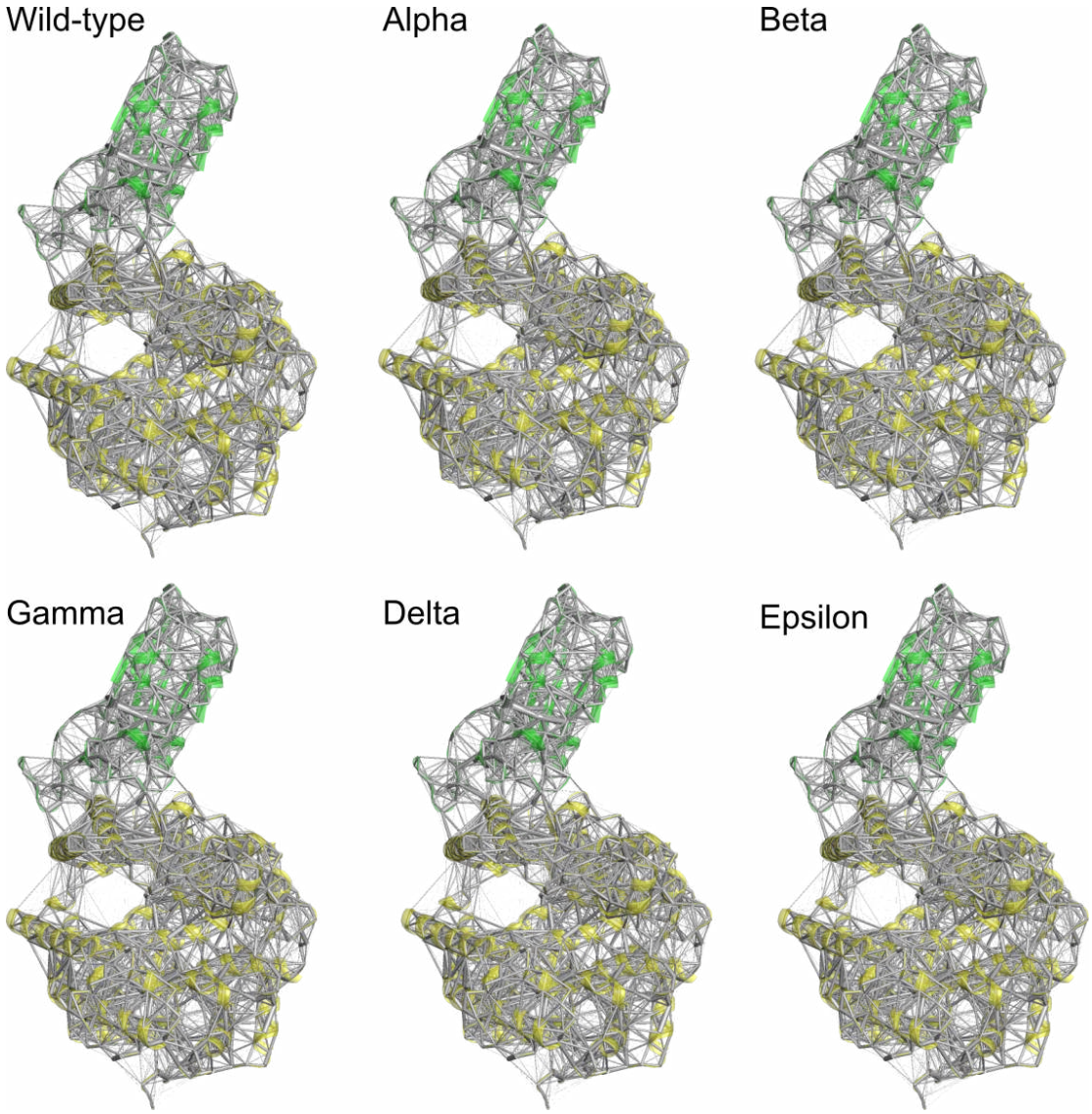


Figure S9. Complete contact network between the wild-type and each studied variant. The spike-RBD(green)/ACE2(yellow) complex is represented in cartoon representation. Contacts are represented with an edge width proportional to the number of interresidual atomic contacts.

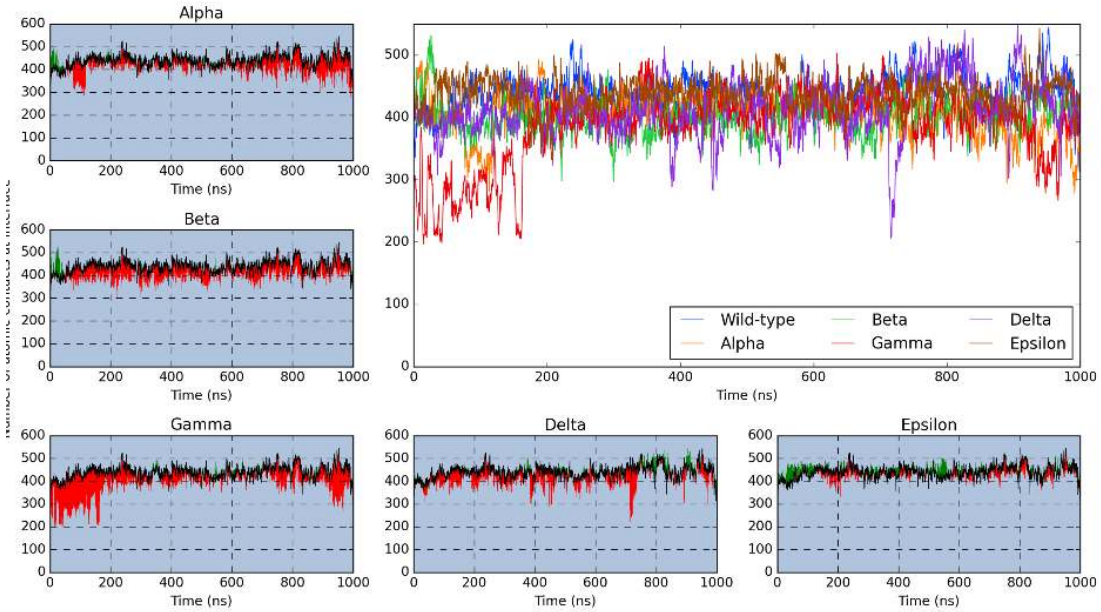
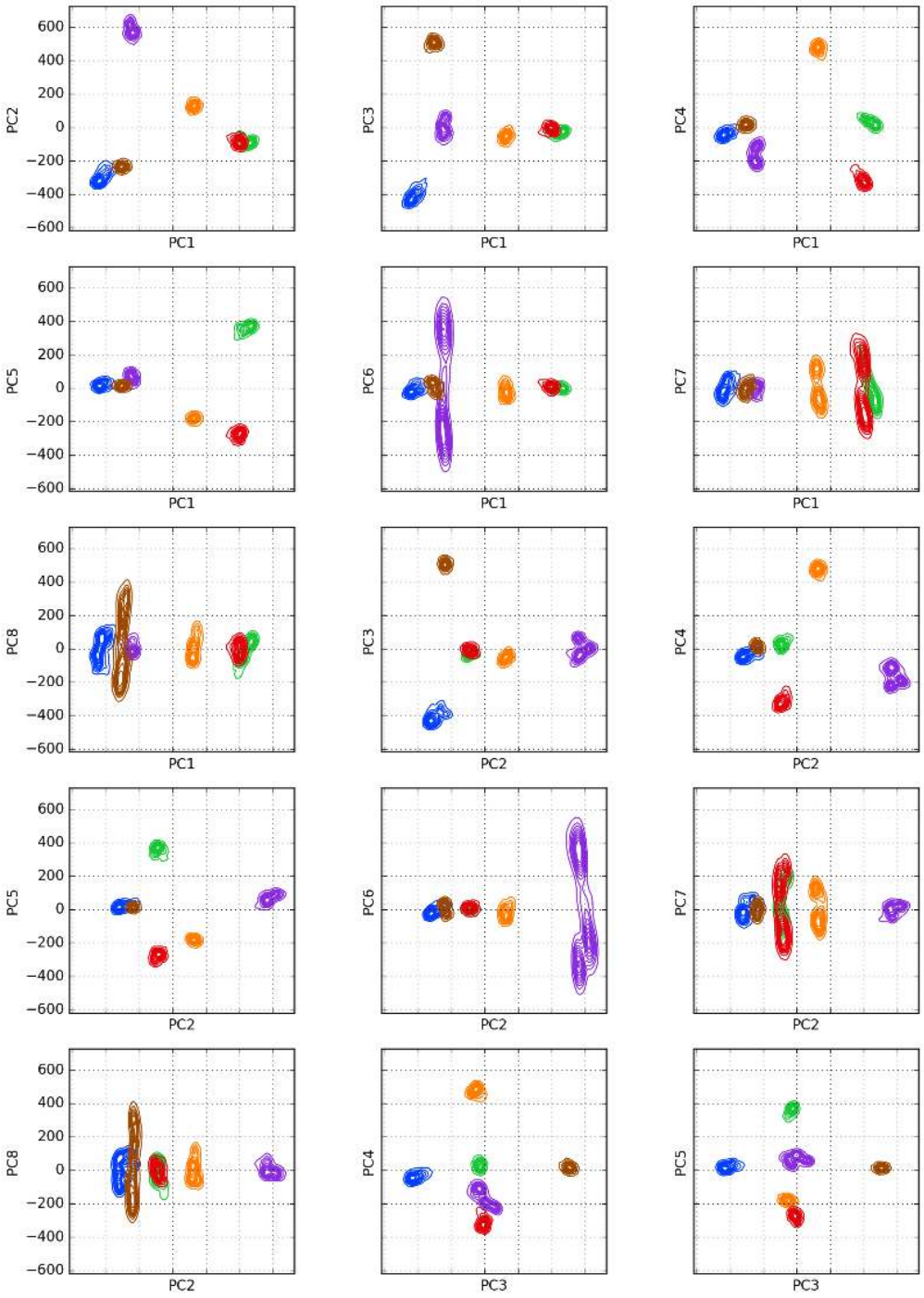


Figure S10. Number of heavy-atom contacts at the interface in function of the time for each simulation (top right panel). Individual comparison between each variant (in green if the number of contacts is bigger in the variant, red otherwise) and the wild-type (in black).


PC2
PC1
PC3
PC1
PC4
PC1
PC5
PC1
PC6
PC1
PC7
PC1
PC8
PC1
PC3
PC2
PC4
PC2
PC5
PC2
PC6
PC2
PC7
PC2
PC8
PC2
PC4
PC3
PC5
PC3
600
400
200
0
−200
−400
−600

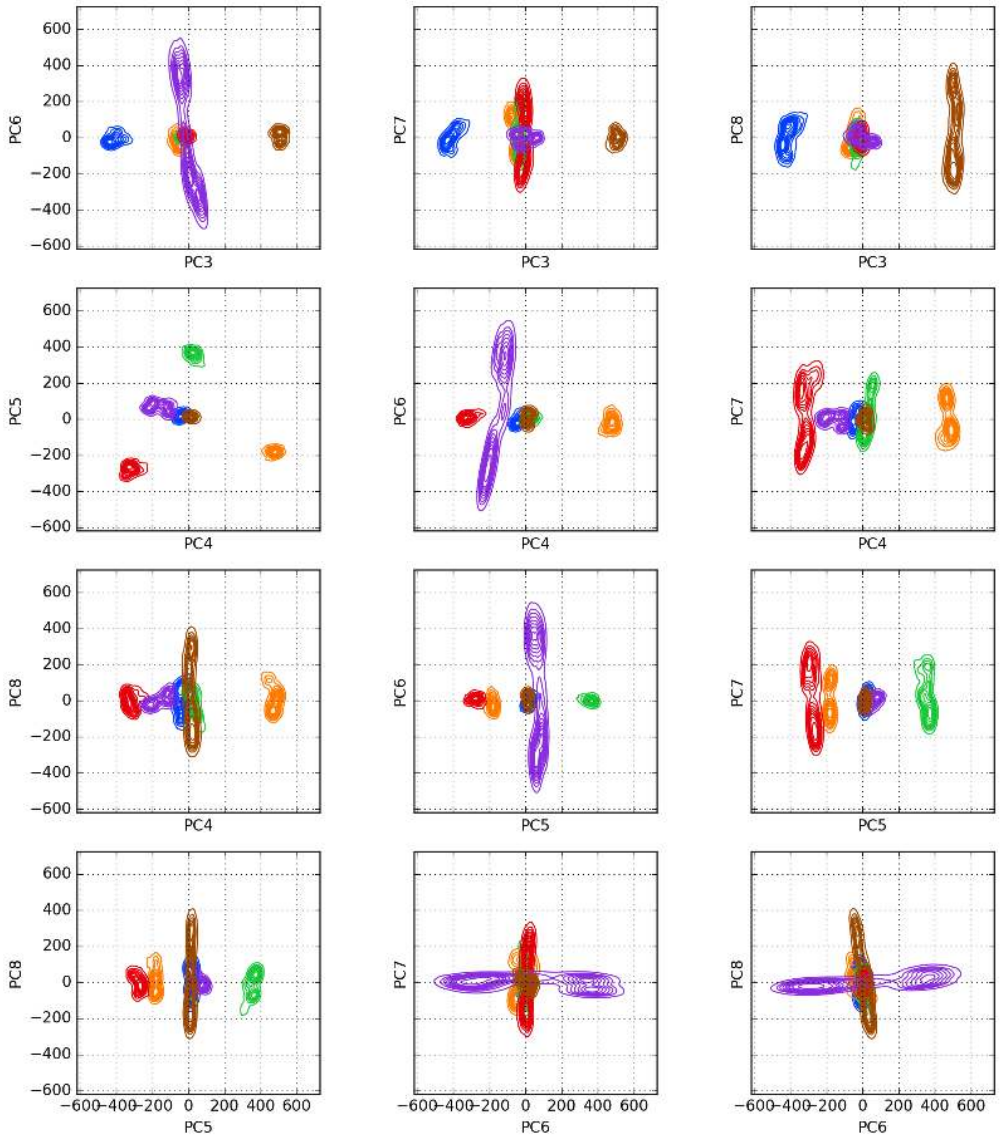


Figure S11. Free energy landscape associated with the different RBD variants in combination of the eight first cPCA eigenvector dimensions. Kernel density estimation shows regions with the highest population with terrain lines.

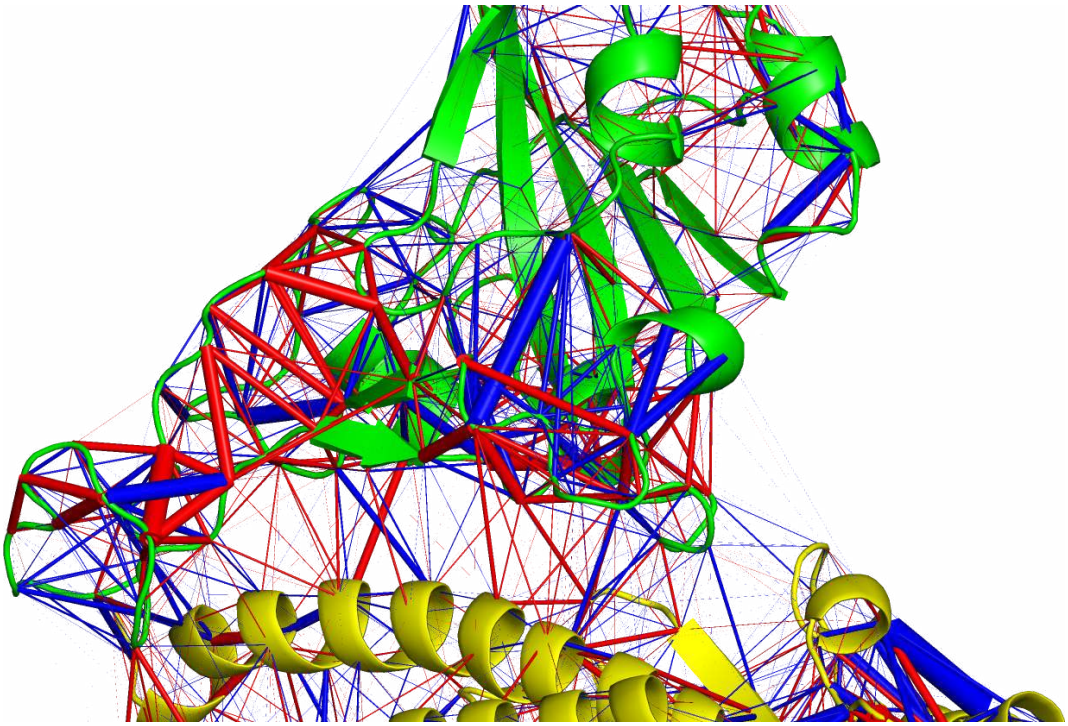


Figure S12. Eigenvector representation of the PC3 (a red edge means an increase in contact leads to positive values in PC3 and a blue edge means a decrease in this contact leads to negative values on PC3)

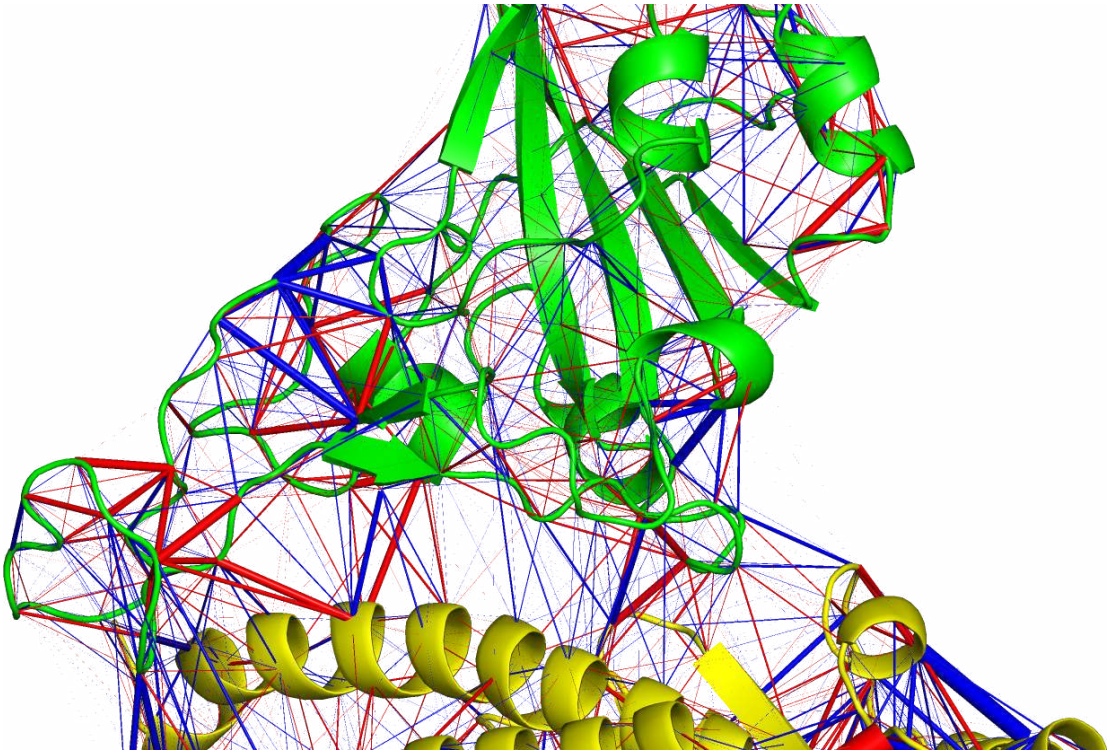


Figure S13. Eigenvector representation of the PC4 (a red edge means an increase in contact leads to positive values in PC4 and a blue edge means a decrease in this contact leads to negative values on PC4)

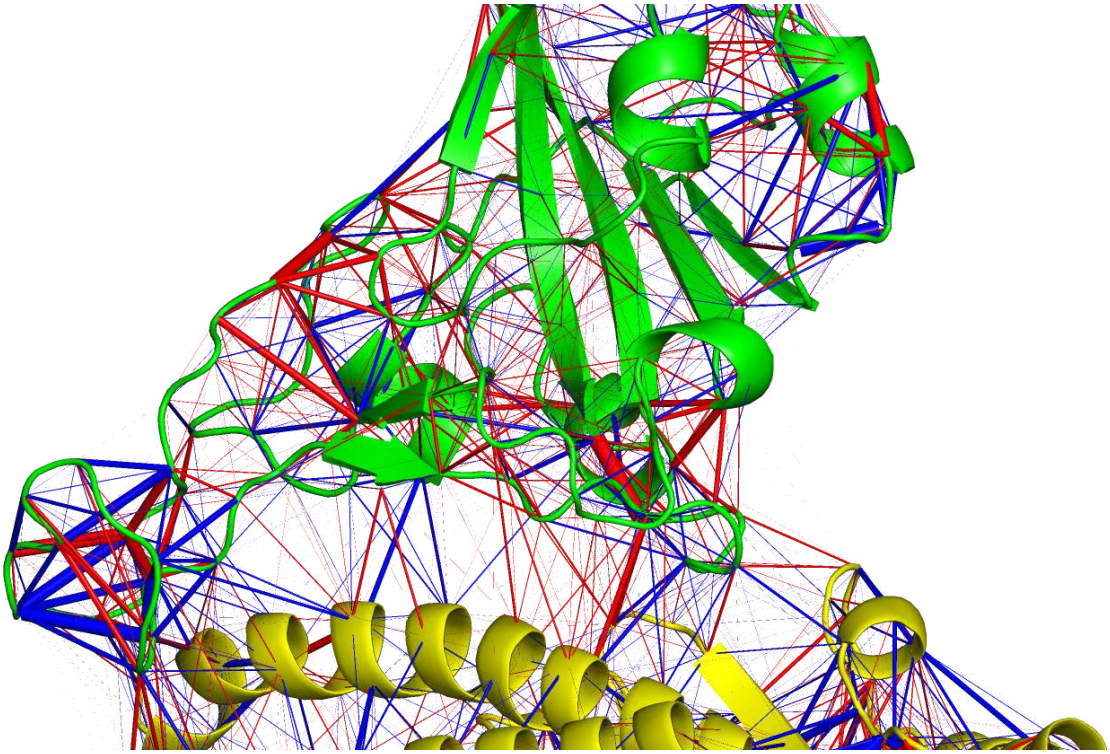


Figure S14. Eigenvector representation of the PC5 (a red edge means an increase in contact leads to positive values in PC5 and a blue edge means a decrease in this contact leads to negative values on PC5)

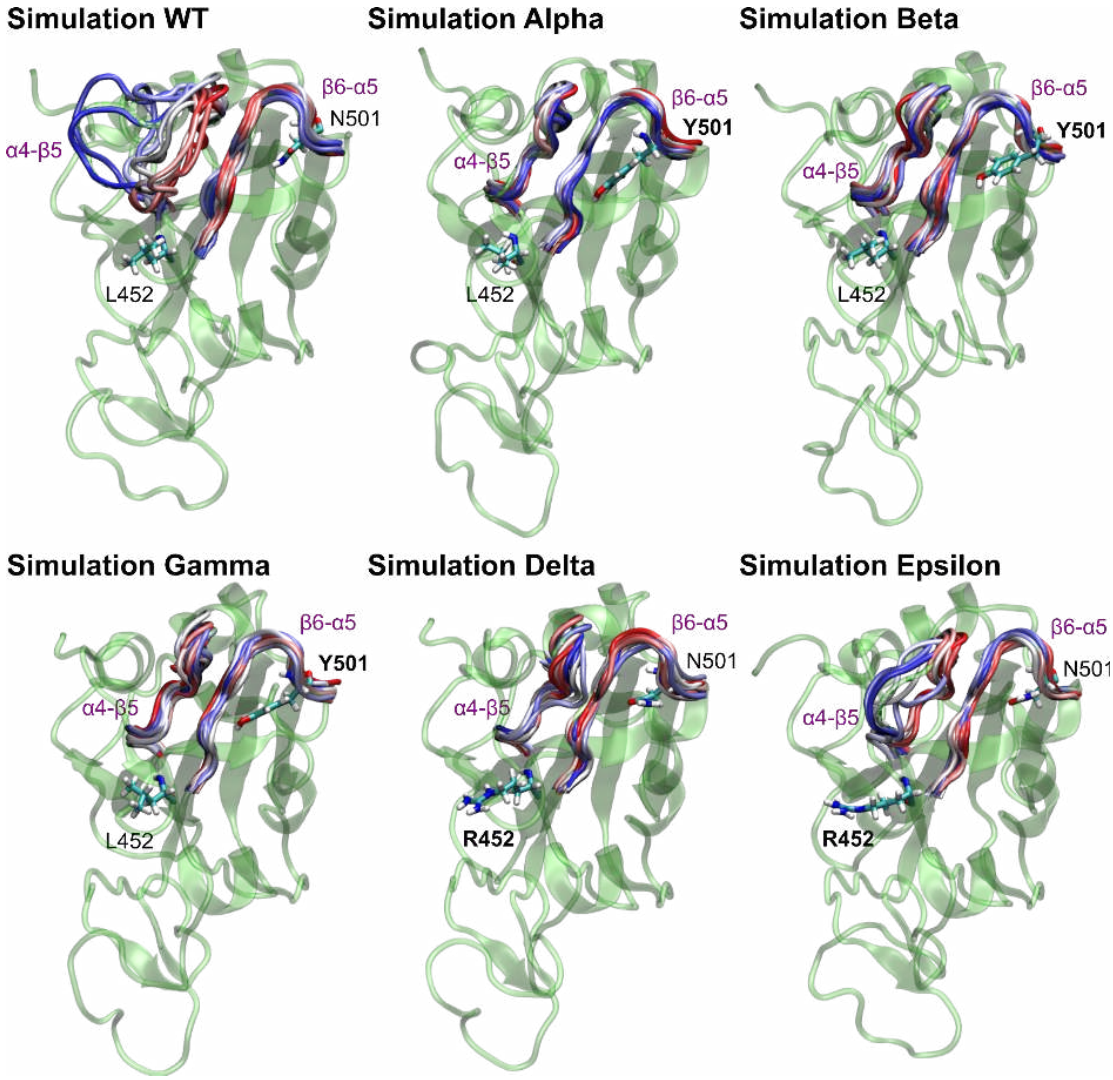


Figure S15. Spike RBD structure (in transparent lime) with the time-evolution of the α4-β5 and β6-α5 turn in contact (from the beginning of the simulation to the end from red to blue with structure printed each 100ns). The position of the mutated N501 and L452 residues are also shown in licorice.

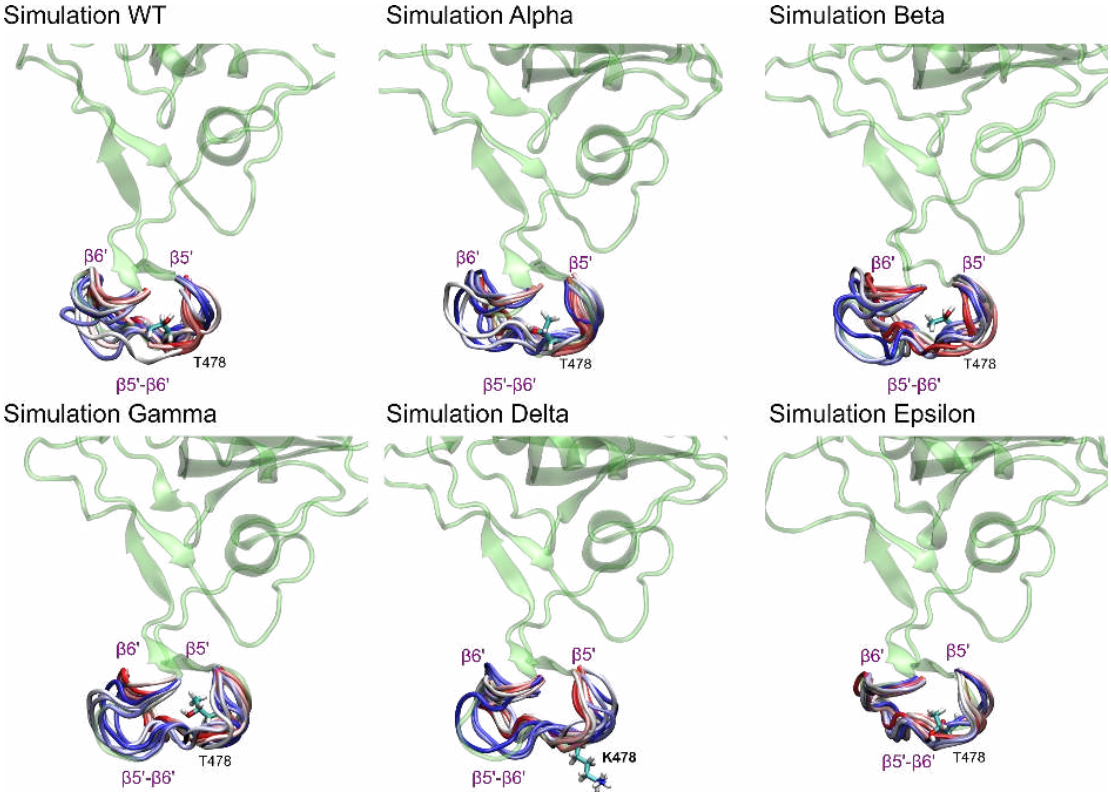


Figure S16. Spike RBD structure (in transparent lime) with the time-evolution of the β5'-β6' loop (from the beginning of the simulation to the end from red to blue with structure printed each 100ns). The position of the mutated T478 residue is also shown in licorice.