Developing New Applications for Perturbation Response Scanning Method to Study Conformational Modulation of Globular Proteins

By

Farzaneh Jalalypour

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Sabanci University
December 2020

Developing New Applications for PRS Method to Study Conformational Modulation of Globular Proteins

APPROVED BY:

DATE OF APPROVAL: 25/12/2020

Developing New Applications for PRS Method to Study Conformational Modulation of Globular Proteins

Farzaneh Jalalypour
Ph.D. Dissertation, December 2020

Supervisor: Prof. Dr. Canan Atilgan

Abstract

Conformational transitions in proteins facilitate precise physiological functions. Therefore, it is crucial to understand the mechanisms underlying these processes to modulate protein function. Yet, studying structural and dynamical properties of proteins are notoriously challenging due to the complexity of the underlying potential energy surfaces (PES). Perturbation Response Scanning (PRS) method has previously been developed to identify key residues that participate in the communication network responsible for specific conformational transitions. PRS is based on a residue-by-residue scan of the protein to determine the subset of residues/forces which provide the closest conformational change leading to a target conformational state, inasmuch as linear response theory applies to these motions. In this thesis, two novel methods are developed to further explore the dynamics of proteins. Perturb-Scan-Pull (PSP) method evaluates if conformational transitions may be triggered on the PES. It aims to study functionally relevant conformational transitions in proteins using results obtained by PRS and feeding them as inputs to steered molecular dynamics simulations. The success and the transferability of the method are evaluated on three protein systems having different complexity of motions on the PES: calmodulin, adenylate kinase, and bacterial ferric binding protein. Results indicate that PSP method captures the target conformation, while providing key residues and the optimum paths with relatively low free energy profiles.
Unlike PSP method, which is developed to study conformational changes between two known states of a protein and considers the best force vector toward the target state, protein perturbation responses can be clustered with the hope of exploring the collective variables (CV) toward new conformations of a protein. The perturbation response clustering (PRC) technique is developed to study the alternative conformations available to proteins for which these have not yet been detected via experimental methods. Using collective variables predicted via clustering of the response vectors, new conformations are sampled, which capture low lying energy states that exist under specific circumstances in vivo. The methodologies developed in this thesis can be applied on a wide range of proteins having different functions and displaying various types of motions. More importantly, these methods can be extended to study nucleic acids (DNA, RNA) or membrane proteins by considering lipids molecules.

Keywords: molecular simulation, conformational change, potential of mean force, steered molecular dynamics, perturbation-response scanning

Globüler Proteinlerin Yapısal Modülasyonunun İncelenmesi Amacıyla Etki Tepki Taraması Yöntemine Geliştirilen Yeni Uygulamalar

Özet

Proteinlerin konformasyonları arasındaki yapısal geçişler belli fizyolojik işlevlerin gerçekleşmesine olanak sağlar. Bu sebeple protein işlevini yöneten bu mekanizmaları anlamak büyük önem arz eder. Proteinlerin yapısal ve dinamik özelliklerini çalışmak, temel potansiyel enerji yüzeylerinin (PEY) karmaşıklığı sebebiyle basit değildir. Daha önce geliştirilen, etki-tepki taraması (ETT) metodu seçilmiş yapısal değişimleri tetikleyen iletişim ağındaki anahtar amino-asitleri tespit eder. ETT, amino asitleri teker teker tarayarak, doğrusal tepki kuramının izin verdiği ölçüde, hedef yapısal duruma en yakın yapısal geçişe sebep olan amino asit/kuvvet altkümesini bulur. Bu tezde, proteinlerin dinamiklerini araştırmak için iki yeni metot geliştirilmiştir. Sars-Tara-Çek (STÇ) metodu, PEY üzerinde yapısal geçişlerin tetiklenip tetiklenmeyeceğini değerlendirir. Metodun amacı, EET'den alınan fonksiyonel olarak ilintili yapısal geçişleri, yönlendirilmiş moleküler dinamik benzetimlerine girdi olarak kullanmaktır. Metodun başarısı ve farklı sistemlere uygulanabilirliğini ölçmek için, PEY üzerindeki hareket örüntüsü farklı olan üç protein sistemi kullanıldı; kalmodülin, adenilat kinaz ve bakteriyel demir bağlayan protein. Sonuçlar gösterdi ki, STÇ metodu anahtar amino asitleri ve görece düşük enerjili en iyi yolakları bularak hedef üç boyutlu yapıya ulışlmasını sağlamaktadır. Proteinin iki bilinen durumunu arasındaki yapısal geçişi, hedef yapıya giden en iyi kuvvet vektörünü bularak çalışan ETÇ metodunun aksine; protein etki-tepkileri, proteinin yeni üç boyutlu yapılarına ulaşan kolektif değişkenler (KD) bulma amacıyla ile kümelenebilir. Etki-tepki kümelemesi (ETK) metodu, x-ışınımı kristalografisi, NMR ve diğer bilinen deneysel yöntemlerle tespit edilemeyen, protein dinamiğince mümkün alternatif üç boyutlu yapıları çalışmak için geliştirilmiştir. Tepki vektörlerinin kümelenmesiyle tahmin edilen KD'leri kullanarak sadece in vivo ortamda var olan ve düşük enerjili durumlara karşılık gelen yeni yapılar örneklenebilir. Bu tezde geliştirilen metotlar, farklı fonksiyonlara sahip ve farklı hareket düzenleri gösteren, geniş çeşitlilikte proteinlere uygulanmıtır. Daha da önemlisi, bu metotların kullanımı, nükleik asitler (DNA ve RNA) ya da lipit molekülleri de gözetilerek hücre zarı proteinlerine kadar genişletilebilir.

Anahtar kelimeler: moleküler benzetim, yapısal değişim, ortalama kuvvet potansiyeli, yönlendirilmiş moleküler dinamiği, etki-tepki taraması

Dedicated to my partner, Pouya Y. Louyeh, who has been a constant source of support and encouragement during the challenges of my PhD.

تقدیم به عشقم پویا که اگه نبود نمی شد

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my supervisor and academic mom, Professor Canan Atılgan, for her guidance, support, excellent advices, and patience, which made her a successful role model for all women. Canan Hocam, thank you for having faith on me and giving me the opportunity to be a member of MIDSTLAB. This dissertation would not have been possible without your persistent help.

I would like to thank Professor Ali Rana Atılgan, my academic dad, who motivated me with his amazing discussions and invaluable ideas. Ali Rana Hocam, thank you for your support and positive energy during hard times when I was away from my home and couldn't visit my family due to Covid-19 pandemic. In every MIDSTLAB Zoom meeting we had, you reminded that we are a family and you treated each of us like one of your children.

I admire the help and guidance of Asst. Prof. Özge Şensoy, who always has a smile on her face. Özge Hocam, working with a successful woman like you, gave me a new perspective on the future.

I would like to thank Prof. Aatto Laaksonen, Dr. Lilit Axner, and HPC-Europa3 Transnational Access programme for giving me the golden opportunity to visit KTH Royal Institute of Technology and perform the second part of this thesis.
Special thanks go to Prof. Erik Lindahl, who gave me his time and the opportunity to join Scilifelab meetings during my visit and for several informative discussions we had.

I also would like to thank our previous dean, Prof. Yusuf Z. Menceloğlu, for his support during difficult times, Prof. Ersin Göğüş who helped me to adapt to university campus life, Dr. Haleh Abdizadeh, Asst. Prof. Jamal S.M Zanjani, and Dr. Morteza Ghorbani for their admirable advises in every aspect of my Ph.D. life.

I am also thankful to MIDTSLAB members, Dr. Sofia Piepoli, Tandac Furkan Guclu, Kurt Aricanli, Gokşin Liu, Ebru Cetin, Isik Kantarcioglu, Erhan Ekmen and Nazli Kocatug, and in particular my brilliant research colleagues and friends Metehan Ilter and Mohamed Shehata who made PhD life much easier and so enjoyable.

I wish to thank all my dear friends, Assoc. Prof. Ali Zarrabi, Dr. shiva Taghizadeh, Babak Bahrami, Taha Behroozi kohlan, Dr. Ferdows Afghah, Ali Nadernezhad, Ema Silva, Francisco Pereira Brandão de Sá Rodrigues, Atefeh khorsand, Sepideh Shemshad, Sina Zarre, Dr. Ali Asgharpour, Dr. Araz Sheibani Aghdam, Saeede Nazari, Amin ghasemzadeh, Dr. Isa Emami Tabrizi, Kaveh Rahimzadeh Berenji, Ahmad Reza Motezakker, Abdul Rahman Dabbour, Ata Golparvar, Sara Atito Ali Ahmed, Saman Hosseini Ashtiani, Nooshin Masoudi, Rana Abedi, Soroush Afkhami, Amir Safari, Dr. Nur Kocaturk, Dr. Yunus Akkoc, Dr. Deniz Gulfem Ozturk, Dr. Ezgi Karakaş Schüller,

Biran Musul, Sinem Aydin, Banu Akinci, and all beloved friends who helped me to accomplish my PhD with their supports and positive vibes.

I am truly grateful to my parents, Farideh and Hojjat, Pouya's parents, Zeynab and Yagoub, my sister, Shahrzad, and my brothers in law, Vahid and Pedram, and also Dayi Rasul and Zandayi robab, for their immeasurable love and care.

I am particularly indebted to my partner, Pouya, who I met at Sabanci University. Pouya thank you for your good humour in thesis anxiety and stress, in darkest hours of my PhD life, for always supporting me and my dreams, believing in me when I lost belief in myself, and love me despite all absences and difficulties. I thank our cats Vahshi, Zardalu, Kuchulu, Ramtin, and Madam Coco who boost our energy.

Finally, I would like to thank Sabanci University for financial and academic support, Technical Research Council of Turkey (TUBITAK) with the project numbers of 116F229 for providing computational facilities, and finally TRUBA (in Turkey), and PDC (in Sweden) centers of High Performance Computing where I conducted this research in the past 3 years.

I would like to thank Turkey and Turkish people who made me feel like at home, and the leader, Mustafa Kemal Atatürk, for his life lessons: "Herkesin kendine göre bir zevki var. Kimi bahçe ile uğraşmak, güzel çiçekler yetiştirmek ister. Bazı insanlar da adam yetiştirmekten hoşlanır. Bahçesinde çiçek yetiştiren adam çiçekten bir şey bekler mi? Adam yetiştirebilen adam da, çiçek yetiştirendeki duygularla hareket edebilmelidir. Ancak bu biçimde düşünen ve çalışan adamlar, memleketlerine ve milletlerine ve bunların geleceğine faydalı olabilirler." (Mustafa Kemal Atatürk)

Table of Contents

List of Tables

List of Figures

Part I. General Introduction

Proteins are dynamical entities, having high degree of conformational flexibility and plasticity which are mediated by breaking and reforming noncovalent bonds in a fluctuating environment. Protein motions are not random and they are evolutionarily conserved to carry out biological functions [Marsh and Teichmann, 2014]. In most proteins, including enzymes and those which are involved in signaling or membrane transport, functionally relevant conformational transitions occur between a pair of states such as active/inactive, bound/unbound, open/closed, or apo/holo forms [Atilgan et al., 2010; Echols et al., 2003]. This conformational transition process is tightly regulated to maintain equilibrium between these end-states to control biological processes [Wu and Post, 2018].

Tracing conformational modulation of a protein is crucial for understanding the way it functions. Yet, analyzing dynamics and conformational modulation of a protein is a challenging procedure due to the lack of desired tools and the complexity of the information it contains [Jalalypour et al., 2020].

Experimental techniques such as X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) and cryo-electron microscopy provide probable conformations for the abovementioned end-states. X-ray crystallography and NMR have both been used to determine the conformational states and dynamics of proteins [Atilgan, 2018]. One limitation of X-ray crystallography is that not all proteins are amenable to crystallization, and even if they are, the structure of flexible parts such as loops might not be correct in a packed crystal [Srivastava et al., 2018]. NMR on the other hand is limited by protein size [Frueh et al., 2013]. In recent years, high-resolution cryo-electron microscopy has developed significantly, which has led to an exponential growth of available structures [Carroni and Saibil, 2016] and used to identify alternative conformations of a protein. However, they provide little, if any, information regarding intermediate states sampled during these transitions. Moreover, such structures may not represent physiologically relevant conformations as they may have been obtained under non-physiological conditions, e.g. non-cellular pH, low temperature, etc [Grant et al., 2010; Nussinov, 2016]. it has turned out to be surprisingly difficult to correctly assign such conformations to functional states, not to mention explain the dynamics of the complex motions itself. Several limitations of experimental methods highlight the need for fast and efficient computational approaches – but they must also have good predictive power. Subjecting these experimental structures to **molecular dynamics (MD) simulations** that mimic

physiological conditions may help achieve a biologically relevant conformational ensemble or relax a structure [Atilgan, 2018; Nussinov, 2016].

In classical MD, each protein is assumed to have a unique energy landscape that is captured under a particular set of conditions where individual distinguishable configurations represent a distinct energy level. Significantly populated conformations sampled under physiological conditions are generally associated with the "native state" and are postulated to correspond to the minimum on the free energy surface [Mallamace et al., 2016; Papaleo et al., 2016]. Elucidation of the underlying molecular mechanisms which govern conformational transitions is crucial to modulate protein function [Haspel et al., 2010]. While MD simulations may be used to investigate time-dependent structural and dynamical properties of proteins at the atomistic level, in general, classical MD cannot achieve time scales accessed by experiments [Karplus and Kuriyan, 2005] due to the rugged energy landscape of proteins that are decorated by various local minima. These states are usually separated by high energy barriers which further impede achieving complete sampling of the available conformational space [Gedeon et al., 2015].

The conformational transitions are typically too rare to be captured within the limited timescales of MD [Karplus and Kuriyan, 2005]; even if they are captured, the methods are too slow to systematically identify the roles of various residues e.g. for allosteric modulation and critically test computational predictions against site-directed mutagenesis results.

In silico methods Alternatively, enhanced sampling methods such as accelerated MD [Hamelberg et al., 2004], milestoning [Faradjian and Elber, 2004], adaptive biasing force [Darve et al., 2008], and metadynamics [Leone et al., 2010] may be used to surpass these high energy barriers [Pierce et al., 2012]. In particular, a number of conformational change specific enhanced-sampling methods that rely on variants of metadynamics simulations have been developed, and they have been successful in characterizing the free energy landscape of proteins that display conformational multiplicity[Brotzakis and Parrinello, 2018; Wang et al., 2016]. At the other end of the spectrum, non-equilibrium methods such as steered molecular dynamics (SMD) may be utilized to get free energy profiles for a given process [Isralewitz et al., 2001]. **Non-trajectory techniques** have also made a significant breakthrough in the field of protein dynamics. In particular, network-based methods such as anisotropic network model (ENM)[Atilgan et al., 2001], Gaussian network model (GNM) [Bahar et al., 1997], torsional network models (TNMs)[Mendez and Bastolla, 2010] as well as normal mode analyses (NMAs)[Bahar

and Rader, 2005; Case, 1994] have proved to be successful in analyzing conformational transitions, while each has its own limitations [Atilgan, 2018]. For instance, NMA is an effective tool for exploring functionally relevant protein motions if they are dominated by relatively simple movements such as hinge bending; however, problems arise while studying more complex conformational changes that are co-represented by several modes [Petrone and Pande, 2006; Yang et al., 2007]. This problem has also been identified in allosteric systems where the dominant motion may be localized to a limited number of key residues and it is not representable by a single dominant mode [Petrone and Pande, 2006].

Perturbation Response Scanning method (PRS) was developed to invoke the modes of motion relevant to the conformational change of interest and to identify key residues having significant role in modulation of the transitions between various states.

PRS is based on **protein perturbation** which uses the idea of applying linear response theory to study conformational changes undergone by proteins under selected external perturbations [Ikeguchi et al., 2005]. Using PRS, protein perturbations responses are applied to a sequential scanning of all residues to search for the perturbations that invoke the response closest to the targeted conformational change [Atilgan and Atilgan, 2009]. PRS method has been tested on a number of systems of varying size and complexity of motion. PRS has been shown to capture **key residues** contributing to the residue-residue interaction network in the protein structure [Abdizadeh et al., 2015; Atilgan et al., 2011; Gerek and Ozkan, 2011; Guven et al., 2014; Penkler et al., 2017; Seyler and Beckstein, 2014; Stetz et al., 2017]. One can combine PRS with other approaches to create new applications and methods. For example, using PRS-based methods, sensor and effector residues of proteins which are responsible for sensing and conveying the allosteric signal, respectively, have been identified [Dutta et al., 2015]. PRS has also been used in a pathogenicity prediction tool to study the significant impact of missense mutations on protein dynamics [Ponzoni and Bahar, 2018]. PRS output has also been utilized to quantify the resilience of individual residues to perturbation by calculating a metric called dynamic flexibility index (dfi) [Nevin Gerek et al., 2013]. dfi has later been used to study the differences in conformational dynamics of evolutionarily related proteins [Zou et al., 2014]. The method has also been coupled with protein-ligand docking in a method called Backbone Perturbation-Dock (BP-Dock) whereby PRS is used to perturb the structure so as to imitate the force that the ligand exerts on the binding site, resulting in generation of a wide range of different conformations as binding-induced states for ensemble docking

[Bolia et al., 2014]. Combination of evolutionary analysis with PRS has been used to characterize the functional and regulatory role of post-translational modification sites in allosteric mechanism of Hsp90 proteins [Stetz et al., 2018]. In another study, a systematic analysis regarding allosteric roles of mutational hotspots in tumor suppressor genes has been provided [Verkhivker, 2019]. Specifically, residue interaction network and PRS results have been compared with experimental studies to reveal cancer mutations responsible for not only local, but also global dynamic fluctuations. Recently, the allosteric regulation of ABL tyrosine kinase using a methodology combining MD simulations and PRS has led to a novel network-centric approach to identify allosteric hotspots and important interactions [Astl and Verkhivker, 2019].

Besides providing effective key residues in a conformational transition, protein perturbation also carries information on the **direction** along which those residues may be manipulated to achieve the target structure. It was postulated that pulling along the possible best direction given by PRS might help overcome the energetic barrier and allow the conformational change to achieve completion [Atilgan et al., 2011]. The most straightforward method to simulate forces acting on given residues is by using the SMD technique [Zhang and Lou, 2012]. SMD was designed for investigating unbinding or unfolding processes in which one has to define an optimal direction to obtain accurate prediction of work and free energy difference estimates. While different methods to find the optimal direction have been proposed [Baker et al., 2013; Sankar et al., 2015; Vuong et al., 2015], in most studies the pulling direction is chosen heuristically which may not lead to a favorable pathway, and increases the chances of inefficient SMD simulations [Liu et al., 2008]. A number of reaction coordinate protocols that are used to identify optimal directions to accelerate convergence of enhanced free energy surface sampling calculations rely on defining a reference path, usually calculated by a direct connection between the initial and target states (see e.g [Czerminski and Elber, 1990]). SGOOP [Tiwary and Berne, 2016] and VAC-MetaD [McCarty and Parrinello, 2017] methods both optimize the number of collective variables, the former based on the assumption that the best collective variables are those with maximum time scale separation between their slow and hidden fast processes, and the latter adopting a signal processing technique to identify slow order parameters that are used as collective variables. RAVE method [Ribeiro et al., 2018] on the other hand, cycles through MD and a deep learning approach to produce the reaction coordinate and its free energy simultaneously. While all these methods lead to the selection of reaction coordinates that greatly enhance the sampling,

the coordinates themselves do not necessarily display a direct biological relevance to the conformational change.

In this thesis, protein perturbation is extended to investigate conformational transitions in protein systems by feedings its findings into SMD as novel methodologies entitled **Perturb-Scan-Pull (PSP)** and **Perturbation Response Scanning (PRC).**

The main advantage of the PSP method is that it is computationally cheap, only needs a short MD simulation of about 100 ns as input, and direction is a natural output of PRS that is fed automatically into SMD. Moreover, we have previously shown that the residues and directions produced by PRS have direct biological relevance, e.g. by pointing to allosteric sites that manipulate the conformational change [Abdizadeh and Atilgan, 2016; Aykut et al., 2013; Penkler et al., 2017]. Once the collective variable is determined, PRS may be combined with, e.g. metadynamics or umbrella sampling, instead of SMD to determine the landscape more precisely. Since PRS provides a single collective variable, it proves to be advantageous over the other approaches, as the number of collective variables selected in reaction coordinate methods should be small in free energy surface construction methods. For example, in metadynamics, the cost of reconstructing the free energy grows exponentially with the number of collective variables used. Similarly, VAC-MetaD [McCarty and Parrinello, 2017] attempts to improve the initial, non-optimal collective variables using a variational principle approach that is based on the time-lagged independent component analysis. Consequently, one needs a long trajectory as opposed to PRS.

Up to now, different conformational states have been captured experimentally both for prokaryotic and eukaryotic proteins, but there are many cases with limited data e.g., where structure on one side of a transition is not available. Unlike PRS and PSP methods, which are developed to study conformational changes between two known states of a protein and only consider the best force vector toward a target state, protein perturbation responses can be clustered with the hope of exploring the collective variables toward new conformations of a protein. Protein Perturbation Response Clustering (PRC) takes all responses into account to indicate the total movements of a protein as a whole. Two different clustering approaches were utilized to classify displacement vectors using K-means algorithm [Lloyd, 1982]. The first approach aims to identify residues with a significant role in the protein dynamics and conformational modulations, by considering the fluctuation of the structure in response to a single residue perturbation. In the second approach, all possible displacement vectors are considered to predict a collective variable

which then feed to MD simulation in order to predict new conformations. The summary of the input necessary for these methodologies and the information obtained as a result is illustrated in Figure 1.

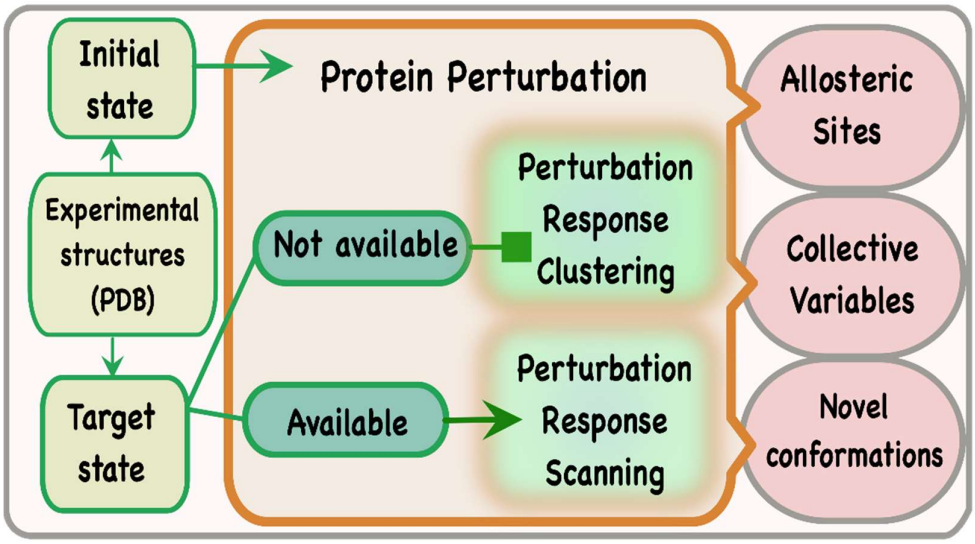

Figure 1. Summary of the perturbation response scanning and clustering

Part II. Methodology

Details of experimental structures

The Protein Data Bank (PDB) is a free open access data archive which provides 3D structure of wide range of biomolecules including proteins, DNA, and RNA [Berman et al., 2000]. Protein structures studied in this thesis were retrieved from the PDB and information regarding all proteins are summarized in Table 1.

Table 1. PDB structures utilized in this thesis

Protein	Motion	State	PDB code	Residue indices	Resolution (Å)	Ligand	reference
ADK	Hinge	Open	4AKE	214	2.2	-	[Müller et al., 1996]
		Closed	1AKE	214	2	substrate-mimicking inhibitor	[Müller and Schulz, 1992]
FBP	Hinge	Open	1D9V	309	1.75	Phosphate ion	[Bruns et al., 2001]
		Closed	1MRP	309	1.6	Phosphate ion/ Fe^{3+}	[Bruns et al., 1997]
Ras	Loop Motion	open	5P21	166	1.35	Phosphoaminophospho - nic acid-guanylate ester (GNP), Magnesium ion	[Pai et al., 1990]
CaM*	Complex motions	Extended	3CLN	5-147	2.2	Calcium ions	[Babu et al., 1988]
		Compact	1PRW	148	1.7	Calcium ions	[Fallon and Quiocho, 2003]
		Compact	1LIN	3-148	2	Calcium ions/ TFP	[Vandonselaar et al., 1994]
		Compact	1CDL	5-146	2	Calcium ions	[Meador et al., 1992]
		Extended	1RFJ	147	2	Calcium ions/ MPD	[Yun et al., 2004]
		Compact	1QIW	2-146	2.3	Calcium ions/DPD	[Harmat et al., 2000]
		Compact	2BBM9	148	NMR	Calcium ions	[Ikura et al., 1992]
		Extended	1MUX	148	NMR	Calcium ions/ WW7	[Osawa et al., 1998]

	Ensemble of 160 models	2K0E	148	NMR	Calcium ions	[Gsponer et al., 2008]
	Ensemble of 20 models	2KDU	148	NMR	Calcium ions	[Rodríguez-Castañeda et al., 2010]

* For all CaM systems, the C_α coordinates of residues 5-147 as well as four calcium ions are utilized. We exclude coordinates for 1−4 and 148 since they are not reported in the 3CLN x-ray data.

Principles of Molecular Dynamics (MD) simulation

In the molecular dynamic (MD) simulations, the physical motion of molecule/particle is calculated in silico [Alder and Wainwright, 1959; Carlo, 1995] via Newton's law of motion

$$\mathbf{F}_\mathrm{i} = m_i\, \mathbf{a}_\mathrm{i} \tag{1}$$

or

$$\ddot{\mathbf{x}}_\mathrm{i} = -\left(\frac{\partial\, \mathbf{x}_\mathrm{i}\, V_i(x)}{m_i}\right) \tag{2}$$

Where $\mathbf{X}_\mathrm{i}$, m_i and $\mathbf{F}_\mathrm{i}$ are position, mass and applied force on particle i. To simulate how the atoms, move according to time (trajectory), the initial configuration including initial position and initial velocity of each atom is defined for a system having N number of atoms. Then the displacement from initial position is calculated via the force acting on each atom. The force can be measured from the interatomic energy or simply the energy between atoms which mainly depends on their distance (**r**). In another word, force can be obtained via differentiating the potential energy function (U) with respect to the position of all atoms

$$\mathbf{F}_i = -\nabla_\mathrm{i}\, U(\mathbf{r}_1, \ldots, \mathbf{r}_\mathrm{N}) \tag{3}$$

The force field is a mathematical model of the potential energy function based on the chemical and physical characteristics of a system in which the potential energy (U), is divided into two parts for bonded and non-bonded interactions (Figure 2).

$$U(\mathbf{r}_1, \ldots, \mathbf{r}_\mathrm{N}) = U_\mathrm{bonded}\,(\mathbf{r}_1, \ldots, \mathbf{r}_\mathrm{N}) + U_\mathrm{non\text{-}bonded}\,(\mathbf{r}_1, \ldots, \mathbf{r}_\mathrm{N}) \tag{4}$$

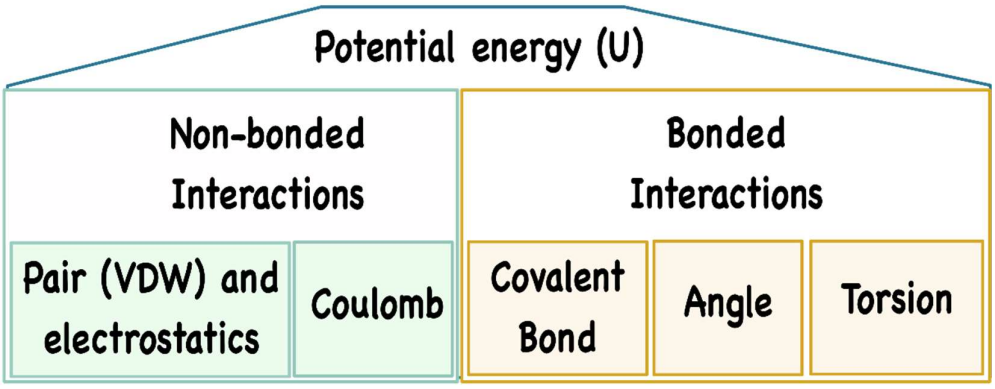

Figure 2. Potential energy (U) is defined based on two types of interactions; bonded and non-bonded.

The bonded interactions include the covalent bonds between two atoms (2-body), angle between three atoms (3-body), and torsion angle between four atoms (4-body). The covalent bond potential is calculated via

$$U_{\text{bond}} = k_{ij} (b_{ij}\text{-}b_{eq})^2 \qquad (5)$$

Where b_{ij} is the distance between pairs of atoms, i and j, (b_{ij}=|$\mathbf{b}_j$-$\mathbf{b}_i$|) and b_{eq} is the reference distance or the average bond length. k is the spring constant rely on type of the atoms. The angular bond potential of three covalently bonded atoms, i, j, and l, is calculated via

$$U_{\text{angle}} = k_{ijl} (\theta_{ijl} - \theta_{eq})^2 \qquad (6)$$

Where k_{ijl} is the constant, θ_{eq} and θ_{ijl} are equilibrium angle and angle between two vectors, ij and jl, respectively.

The torsion angle potential is calculated via

$$U_{\text{torsion}} = k_{ijlh} (1+ \cos (m\ \varphi_{ijlh} - \gamma)) \qquad (7)$$

In which φ, m, and γ are the angle between planes (ijl and jlh), phase shift angle, and integer constant, respectively.

The non-bonded interactions consist of Coulombic, van der Waals (VDW) and electrostatic interactions, which the latter can be calculated via Lennard-Jones potential

$$U_{LJ,t}(r) = \begin{cases} 4 \in [(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6} & r \le r_c \\ 0 & r > r_c \end{cases} \quad (8)$$

where r, σ, and ϵ are the distance between two atoms, size of atom, and the depth of the potential, respectively. And the former can be measured via

$$U_{Coulomb} = \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \quad (9)$$

After measuring the force between particles, as the following, the $\mathbf{r_i}(t + \Delta t)$ is calculated by substituting the force with second derivative of the position. The displacement from initial position after a short time interval (timestep) is given by a standard taylor expansion series

$$\mathbf{r_i}(t + \Delta t) = \mathbf{r_i}(t) + \frac{d\mathbf{r_i}(t)}{dt}\Delta t + \frac{d^2\mathbf{r_i}(t)}{dt^2}\frac{\Delta t^2}{2} + \cdots \quad (10)$$

The Verlet integration is a numerical method to integrate equation 1 to obtain the position $\mathbf{r_i}(t + \Delta t)$. The Verlet algorithm [Darden et al., 1999] truncate the series after third order by summing up equation 2.10 with $\mathbf{r_i}(t - \Delta t)$ one, leading to equation 11 by substituting the force with second derivative of the position with respect to time.

$$\mathbf{r_i}(t + \Delta t) = 2\mathbf{r_i}(t) + \mathbf{r_i}(t - \Delta t) + \frac{\mathbf{F_i}(t)}{m_i}\Delta t^2. \quad (11)$$

Principle of Steered Molecular Dynamics (SMD)

The rationale behind Steered molecular dynamic simulation is to study protein dynamics by exerting an external force on a selected atom (SMD atom), while a certain atom is fixed. SMD simulations can be performed in two types of either constant force or constant velocity. The constant force pulling is utilized when one knows the exact force value to apply, whereas constant velocity pulling is proper for unknown systems.

In constant velocity pulling, force is applied on a dummy atom which is attached to SMD atom with a virtual spring and moves with a constant velocity. The force between dummy atom and SMD atom is calculated via equation 12 and 13

$$\mathrm{F} = -\nabla U \tag{12}$$

and

$$U = \frac{1}{2} k[vt - (\mathrm{r} - \mathrm{r}_0) \,.\, \mathrm{n}]^2 \tag{13}$$

Where *U, v, k, t* is potential energy, pulling velocity, spring constant, and time, respectively. The higher the speed is, the stronger is the force. The SMD atom will be moved from its initial position r_0 to r along direction n. As shown in Fig. the external force is applied on the dummy atom depicted in green, which allows a linear motion with respect to time. The force applied on SMD atom relies on the distance between SMD and dummy atom. The SMD atom shown in red, as well as the rest of the structure attached to it via covalent bonds, follow the harmonic spring (Figure 3) [Célerse et al., 2019]

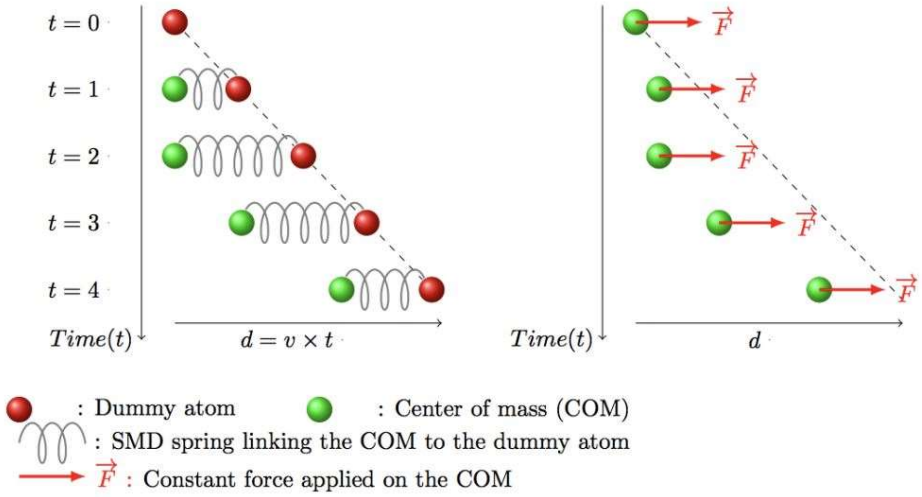

Figure 3. Principle of the constant velocity and constant force SMD simulations. Left: The constant velocity pulling; Green: SMD atom, red: dummy atom, Gray: harmonic spring. Dummy atom moves with constant velocity which enforces SMD atom to move via the force exerted by the spring. Right: The constant force pulling; the force is directly applied on SMD atom. The figure is taken from ref (Célerse, et al. 2019)

Classical MD simulations protocol

We have performed classical MD simulations starting with the listed PDB structures in Table 1. Each protein is solvated in a water box of at least 10 Å from all directions using

the VMD 1.8.7 program [Humphrey et al., 1996] (https://www.ks.uiuc.edu/Research/vmd/) with solvate plugin version 1.2 (https://www.ks.uiuc.edu/Research/vmd/plugins/solvate/). The NAMD package [Phillips et al., 2005] (https://www.ks.uiuc.edu/Research/namd/) is used to model the dynamics of the protein-water system. The CharmM36 [Best et al., 2012] force field parameters (http://mackerell.umaryland.edu/charmm_ff.shtml) are used with the TIP3P model for water. The protein-water system is neutralized by addition of ions to reach 150 mM KCl for calmodulin (CaM), Ras, and adenylate kinase (ADK) as intracellular proteins, and 150 mM NaCl for the periplasmic FBP protein (FBP). Particle mesh Ewald method with periodic boundary conditions was used for calculating electrostatic interactions, with a cutoff distance of 12 Å and a switching function at 10 Å [Darden et al., 1999]. RATTLE algorithm [Andersen, 1983] is applied to use a step size of 2 fs in the numerical integration with the Verlet algorithm [Darden et al., 1999]. Temperature control is carried out by Langevin dynamics with a damping coefficient of 5 ps^{-1}. Pressure control is attained by a Langevin piston. Volumetric fluctuations are preset to be isotropic. The system was first minimized for 5000 steps to remove bad contacts. Then, the MD simulation is run in the isothermal-isobaric (NPT) ensemble at 1 atm and 310 K until volumetric fluctuations are stable to maintain the desired average pressure. Production runs are made for the next 100 ns and the coordinate sets are saved at 2 ps intervals leading to 50000 snapshots for each trajectory. More details regarding MD simulation are listed in
Table 2.

Table 2. Details of the classical MD simulations.

Protein	**Starting structure**	**Equilibrated box size (Å)**	**Number of atoms**	**Ionic concentration**	**Simulated molecules**
CaM	3CLN	75 × 90 × 70	47778	150 mM KCl	Protein, calcium ions
CaM	1PRW	70 × 69 × 72	32521	150 mM KCl	Protein, calcium ions
ADK	4AKE	78 × 95 × 95	67335	150 mM KCl	Protein
FBP	1D9V	92 × 82 × 69	49580	150 mM NaCl	Protein, phosphate anion*
FBP	1MRP	78 × 86 × 80	49580	150 mM NaCl	Protein, phosphate

Ras	5P21	73× 72 × 72	35145	150 mM KCl	anion*, ferric iron GTP

*phosphate group is modeled as $H_2PO_4^-$ in all FBP simulations.

Additionally, in order to sample more conformations for clustering purposes, 400 ns simulations are performed for wild type, protonated, and mutated CaM system start from extended form. The protonation state of residues are determined using H++ [Anandakrishnan et al., 2012] (http://biophysics.cs.vt.edu) and PROPKA web servers [Rostkowski et al., 2011] (http://server.poissonboltzmann.org) for CaM system. The point mutation is performed using the Mutator plugin (https://www.ks.uiuc.edu/Research/vmd/plugins/mutator/) in the VMD software.

Steered MD simulations (SMD) protocol

SMD simulations were performed under the same conditions as those set in the classical MD runs, starting from the same initial snapshot as that used for the starting PRS structure, but with a timestep of 1 fs. SMD runs are continued to the extent that required approaching the target state. SMD simulations were carried out by fixing a residue along the pulling direction and applying external forces to the key residues, K*. All SMD (pulling) simulations are named with a job ID that is prefixed P, followed by a job index. These simulations are performed with various combinations of parameters for constant velocity, v (0.01, 0.02, and 0.03 Å ps^{-1}) and spring constant, k (80, 90, and 100 kcal mol^{-1} $Å^{-2}$). Relaxation simulations are carried out under the same conditions as the classical simulations; they are named with the job ID of the SMD simulation that led to the new point followed by R and a job index.

PRS calculations and analysis

The PRS methodology was introduced and described in ref [Atilgan and Atilgan, 2009]. Briefly, PRS requires two distinct forms of a protein, denoted by initial (I) and target structures (T) as inputs. The objective is to find a perturbed residue and perturbation direction in I that leads to displacements that are most similar to the conformational change between I and T. Based on linear response theory, PRS relates the external force

($\Delta\mathbf{F}$) to displacements ($\Delta\mathbf{R}$) via a covariance matrix, **C**, derived from MD simulations. The shift in the coordinates is calculated by

$$\Delta\mathbf{R}_1 = \langle\mathbf{R}\rangle_1 - \langle\mathbf{R}\rangle_0 \cong \beta\langle\Delta\mathbf{R}\Delta\mathbf{R}^T\rangle_0\Delta\mathbf{F} = \beta\mathbf{C}\Delta\mathbf{F} \quad (14)$$

where $\mathbf{R}_0$ and $\mathbf{R}_1$ represent initial conformation of protein (unperturbed state) and the predicted coordinates (perturbed state), respectively, and $\beta = 1/k_B\,T$. $\Delta\mathbf{F}$ vector contains the coordinates of the force vectors applied on the residues. **C** is the cross-correlation matrix of the fluctuations of the nodes in the initial state of the protein.

To this end, the coarse-grained representation of each state is constructed by taking the C_α atom of each residue as a node. Then, many random, fictitious, forces ($\Delta\mathbf{F}$) in different directions are sequentially introduced on each node to perturb the structure, leading to the predicted $\Delta\mathbf{R}_1$ values for each $\Delta\mathbf{F}$. PRS records the resulting predicted displacements to compare them with the conformational change determined from experimental coordinates, here difference between initial and target crystal structures ($\Delta\mathbf{S}$). As the final step, the overlap between the predicted and measured directions is evaluated by:

$$O^i = \frac{\Delta\mathbf{R}^i\cdot\Delta\mathbf{S}}{\left|(\Delta\mathbf{R}\cdot\Delta\mathbf{R})^i(\Delta\mathbf{S}\cdot\Delta\mathbf{S})\right|^{\frac{1}{2}}} \quad (15)$$

High PRS overlaps (O^i) indicate high similarity of the predicted and experimental displacement vectors implies a good choice for perturbing vectors (force vectors) termed as best PRS direction (D*) to achieve the desired conformational transition. Calculations of PRS as well as the analyses of results have been performed using MDToolbox [Matsunaga and Sugita, 2018] implemented in MATLAB (https://github.com/ymatsunaga/mdtoolbox). The PRS steps are schematically shown in Figure 4.

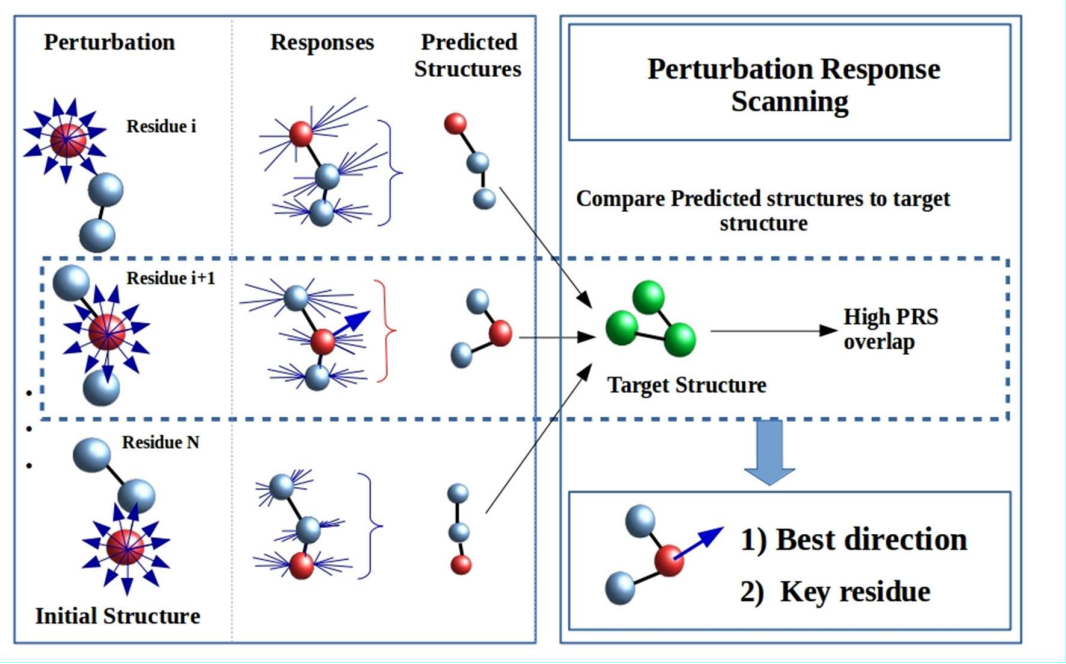

Figure 4. PRS steps illustrated schematically.

Potential of mean force calculations

SMD is used to accelerate the process of conformational transition by pulling a specific atom of the protein along a predefined direction. The energy profile, approximated by the PMF along the free energy pathway, can probe the underlying mechanisms of the conformational transition. Jarzynski's equality [Eq. (16)] is applied to SMD simulation results for PMF calculation (see ref [Park and Schulten, 2004] for details).

$$\mathbf{F}_{\lambda(t)} - \mathbf{F}_{\lambda(0)} = \frac{-1}{\beta}\log\langle\exp[-\beta W(t)]\rangle \qquad (16)$$

Here, $\mathbf{F}_\lambda$ is the Helmholtz free energy of the system, t is time, W is the work done, calculated from the integral of force as a function of the distance the SMD atom has moved during the conformational transition. This equality links non-equilibrium processes of SMD simulations with equilibrium properties manifested in the PMF [Park and Schulten, 2004].

In constant velocity SMD, a force is imposed on the center of mass of the SMD atom via a virtual spring. Having a stiff spring, its position changes along the pulling coordinate (λ) via

$$\lambda(t) = \lambda(0) + vt \qquad (17)$$

To calculate the PMF, each simulation is repeated 10 times to generate several SMD trajectories for each pathway. The spring constant (k) is chosen large enough to avoid fluctuation of the SMD atom. Thus, the reaction coordinate is calculated via equation 4. Stiff spring also minimizes the deviation between the reaction coordinate among the repeated trajectories [Park and Schulten, 2004] which are overlapped to ensure that SMD atom is moving through similar reaction coordinates.

PMF is calculated using the second order cumulant expansion formula [Eq. (18)], which has proved to work better than exact formula [Eq. (16)] for SMD method with limited sampling (see ref [Park and Schulten, 2004]).

$$\mathbf{F}_{\lambda(t)} - \mathbf{F}_{\lambda(0)} = \langle W(t) \rangle - \frac{1}{2k_B T} (\langle W(t)^2 \rangle - \langle W(t) \rangle^2) + \cdots \quad (18)$$

PMF calculation is performed with 0.02, 0.01, 0.02 and 0.02 Å ps^{-1} constant velocity for CaM, ADK, Ras and FBP, respectively. These values are slower than those for common SMD simulations, to improve PMF values. Spring constant is set 100 kcal mol^{-1} $Å^{-2}$, which is large enough to prevent the fluctuation of the reaction coordinate among different trajectories. All simulations are performed at 310 K. Free energy path is computed with respect to the distance between initial and final position of the SMD atom which is measurable by equation 17; it is 45, 11, 18 and 6 Å for CaM, ADK, Ras and FBP systems, respectively. The final position is considered as where the SMD trajectory reach the minimum RMSD with T.

K-means clustering

K-means clustering is one of the most commonly used unsupervised machine learning methods; it aims to find groups with certain similarities in a data set with no predefined labels [Lloyd, 1982]. K-means is an iterative algorithm which ies dataset having N data points into k groups or clusters. Briefly, the number of clusters is determined as k (e.g. k=2, cluster data into two groups), and data points are randomly assigned to one of the clusters accordingly. Then the centroid is computed for each cluster to represent the cluster center; this can be an imaginary or a real location. Subsequently, data points are reassigned to the closest centroid (nearest mean) to generate new clusters for which the centroids are recomputed. The final two steps are repeated in each iteration until

convergence is reached and no improvement is possible. Convergence is defined as the iteration at which data points are allocated to the same cluster and there is not a significant change in the centroid of newly generated clusters. In this thesis, the perturbation response vectors are clustered using K-means algorithm.

Scripting and programming language

Matrix laboratory (MATLAB) is a programming language developed by MathWorks which allows matrix modification and manipulation, as well as algorithm implementation. MDToolbox [Matsunaga and Sugita, 2018] implemented in MATLAB (https://github.com/ ymatsunaga/mdtoolbox) was developed to analyze data generated via molecular dynamics (MD) simulations and provides a collection of required functions. In this thesis, PRS calculations as well as the analyses of results have been performed using MDToolbox. The required scripts to perform perturbation clustering have been implemented using the Statistics and Machine Learning Toolbox of MATLAB.

Part III. Perturb-Scan-Pull methodology

Perturb-Scan-Pull (PSP) as a methodology to determine conformational switching pathways in proteins

PSP method is exemplified by three protein systems which are represented by different type of motions: (i) calmodulin (CaM; complex combination of rotation and bending)[Atilgan et al., 2011], (ii) adenylate kinase (ADK; open-to-closed loop motion)[Müller et al., 1996], and (iii) ferric binding protein (FBP; hinge bending)[Atilgan and Atilgan, 2009](Figure 5). Results show that PSP effectively identifies key residues and pulling directions for the three systems studied which are needed to accomplish the conformational change from the initial to the target structure. We also show that the energetically more favorable path is followed upon usage of the best possible direction and key residues provided by PRS.

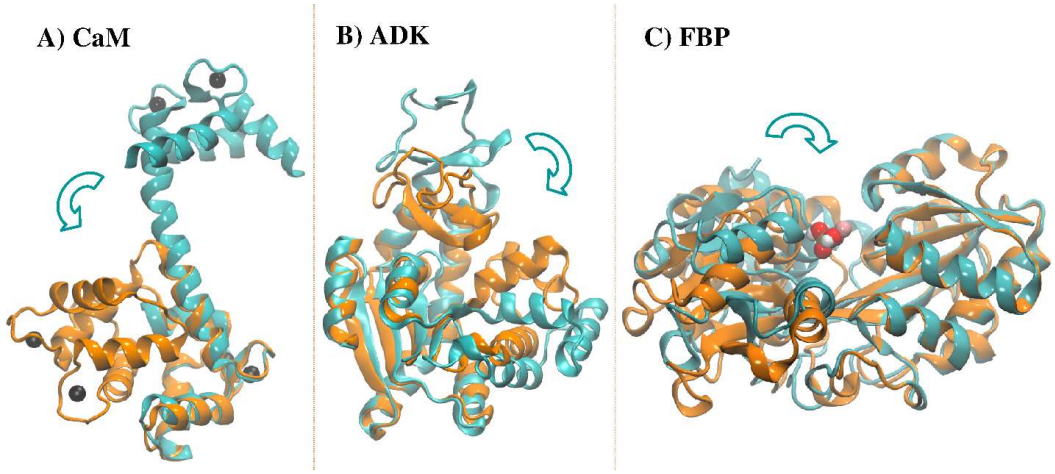

Figure 5. Motions of the three protein systems studied in this section. Extended form of proteins colored in cyan and compact structures colored in orange. Arrows indicate direction of motion. A) Calmodulin displays a complex transition which is represented by twisting and bending around the central helix. Cyan: 3CLN; orange: 1PRW. Calcium ions shown as black beads. Two structures are superimposed on the C-domain. B) Adenylate kinase (ADK); hinge motion of the flexible loop. Cyan: 4AKE; orange: 1AKE. Two structures are superimposed on the core domain. C) Ferric binding protein (FBP); hinge motion of the moving domain on the fixed domain. Cyan: 1D9V; orange: 1MRP. Ferric iron colored in pink and phosphate group shown in space filling. Two structures are superimposed on the so-called fixed-domain.

Development and parameter optimization of the PSP methodology

The PSP methodology consists of two main stages: 1) PRS is applied to the system (see section II-Methods) to select candidate residues and directions that have the propensity

to accomplish the pursued conformational change, 2) information regarding key residues and force directions are used in SMD simulations to trigger the conformational change. PSP procedure is summarized in Figure 6.

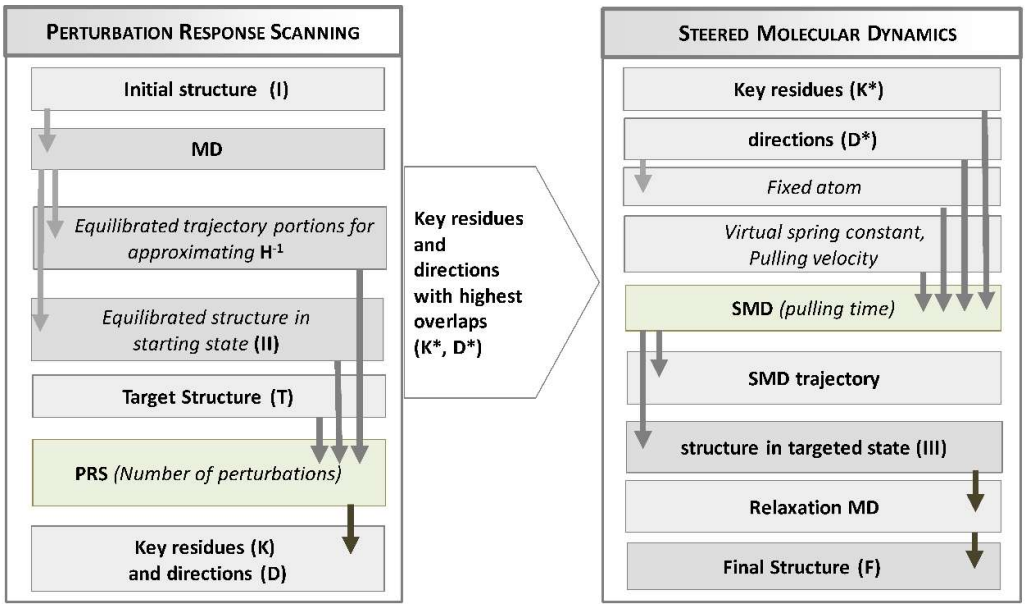

Figure 6. Summary of the PSP methodology; parameters to optimize are displayed in italics; arrows show the information fed from one box to another; the main component is in a green box. The first column displays the flow of PRS calculations. Structure I and T indicate the initial and target structures, respectively. Structure I is not directly used in PRS, but is fed to a classical MD simulation which yields a trajectory by which one can choose an equilibrated trajectory chunk for the approximation of the inverse Hessian ($\mathbf{H}^{-1}$) as well as a compatible well-equilibrated initial snapshot (structure II). A residue with high PRS overlap (K*) and its corresponding direction (D*) define the SMD atom and the best pulling direction, respectively. The second column displays the flow of SMD simulations. The fixed atom is defined according to pulling direction (see text). Pulling simulation starts with the same initial structure as used in PRS (structure II). A frame of pulling simulation having minimum RMSD with the target structure is recorded (Structure III) and subjected to relaxation simulation. Final structure (F) is obtained as the most similar frame of relaxation simulation to target structure.

Information regarding the PRS part of the PSP methodology (column 1 in Figure 6) for each system is listed in Table 4. For this phase of PSP, 100 ns simulations were performed for each system, starting from the initial (PDB) structure which we label I. For PRS calculations, it is imperative that a well-equilibrated chunk of an MD trajectory is selected to construct the variance-covariance matrix which is used to approximate the inverse Hessian, $\mathbf{H}^{-1}$ (equation 1). For this purpose, RMSDs of the proteins from the starting structures are measured and displayed in Figure 7 for all the systems studied in PSP

section, wherein the portions of the trajectories used in matrix construction are marked with red horizontal lines.

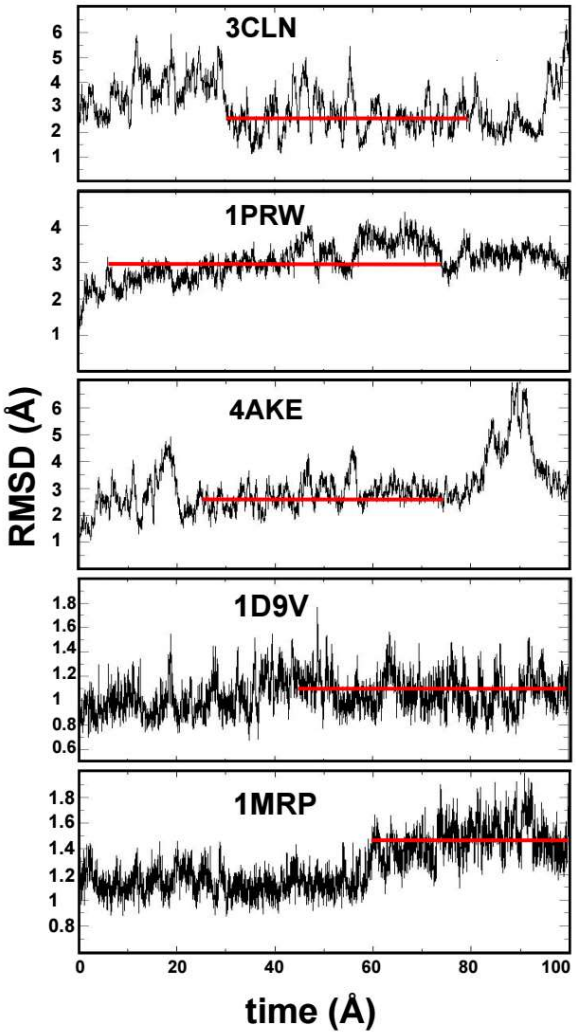

Figure 7. RMSD plots of the classical MD simulations performed starting from the PDB

Previously, crystal structures were used as initial structures in the standard PRS method [Atilgan et al., 2011]. However, the output of such a PRS is not applicable to the PSP procedure, because the SMD simulation cannot be initiated from a non-equilibrated crystal structure. To address this issue, we have used the initial snapshot of the equilibrated chunk (II) for PRS calculations; this selection has served the purpose of reaching high overlaps in PRS. Regarding selection of the number of perturbations introduced on each residue, we have found that 600 perturbation vectors which are randomly distributed within a sphere sample all directions that provide large PRS

overlaps (O^i). Note that having maximum O^i < 0.6 may not lead to the targeted conformational changes in the second stage of the PSP methodology.

Table 3. Summary of PRS results for the three protein systems including best overlap values and key residues

Protein	PRS type	Initial structure (II)*	Target structure (T)	Equilibrated trajectory chunk	Average RMSD of chunk (Å)	best overlap (O^i)	Key Residues (K*)
CaM	Extended to compact	3CLN (30ns)	1PRW	30-80 ns	2.5	0.75	106, 105, 26, 122, 118
	Compact to extended	1PRW (5 ns)	3CLN	5-75 ns	3	0.71	59, 108, 53, 58, 17
ADK	Open to closed	4AKE (25ns)	1AKE	25-75 ns	2.5	0.89	146, 153, 151, 147, 149
FBP	Open to closed	1D9V (45ns)	1MRP	45-100 ns	1	0.91	31, 6, 33, 34, 37
	Closed to open	1MRP (60ns)	1D9V	60-100 ns	1.5	0.91	71, 70, 46, 45, 231

* The time point of the trajectory used as the initial structure is displayed in parentheses and is selected as the initial structure of the trajectory chunk used.

At the next phase of PSP (column 2 in Figure 6), using PRS outputs as input to SMD simulations, several constant velocity pulling simulations have been performed; see Methods section for details of SMD simulations. Thus, key residues (K*) and the

corresponding best PRS direction (D*) which are suggested by PRS, define the pulling (SMD) atom and pulling direction, respectively. To apply the same direction obtained from PRS, structure II is used to initiate both PRS and SMD (green boxes in Figure 6). To avoid rotations during the pulling, another residue that resides on the opposite domain and is alongside the vector defining the pulling direction is selected as the fixed atom. To this end, the angle between the pulling direction and the vectors between the Cα atoms of the SMD residue and all other residues are calculated on structure II. The Cα atom with the smallest angle and is at least 20 Å from the SMD atom is selected for constraining; note that this distance is based on size of the molecule and might be modified depending on the protein being studied. An SMD simulation is considered successful if the protein is not distorted, the secondary structure of the protein is stable and the pulling leads to conformations resembling the target state. Subsequently, a relaxation simulation is performed to eliminate the artificial tensions built in the pulled structure. The relaxation step also allows the protein to complete the process; either by reaching the target state, or by returning to the initial one. For this purpose, the frame of trajectory with minimum RMSD from the target structure is saved as the desired snapshot (structure III) for each pulling simulation and subjected to a relaxation simulation. The obtained final structure (F) is saved for further analysis.

In what follows, we present the details of the PSP method on three different proteins. As the calmodulin (CaM) presents the hardest test case in terms of the complexity of the motion in the conformational change, we start out by outlining the optimization of PSP parameters on CaM. To test the generality of our approach, we follow by applying the optimized PSP methodology to ADK and FBP, two proteins which have long been utilized as test cases for studying conformational multiplicity in proteins.

PSP proof-of-concept in the complex motions of calcium bound calmodulin

CaM acts as an intracellular Ca^{2+} sensor, and plays an important role in calcium signaling pathways in eukaryotic cells [Dagher et al., 2009]. It regulates a variety of biological processes by its propensity to bind a wide range of targets. CaM can adopt a large variety of conformations which proves to be important for carrying out its function [Liu et al., 2017]. CaM is made up of the N-domain (residues 1−68) and the C-domain (residues 92−148) which are connected by a flexible helical linker (residues 69−91). The three dimensional structure of CaM includes seven α-helices which consist of mainly polar and

charged residues; the protein displays four EF-hand motifs, two on each domain, each of which can coordinate one Ca^{2+} ion [Atilgan et al., 2011].

Apo-CaM is the ion-free state of the protein which is activated as a result of an increase in the intracellular concentration of free Ca2+ due to transient opening of the trans-membrane Ca2+ channels. Calcium loaded CaM (holo-CaM) is the active form of the protein [Komeiji et al., 2002]. Ion binding causes large conformational changes in CaM, leading to significant exposure of non-polar regions on the protein surface that facilitate interaction between those non-polar regions of the protein and its targets [Vetter and Leclerc, 2003]. Besides the large structural change accompanying apo → holo transformation, holo CaM may also undergo large spatial changes to adopt a compact form after loading with a ligand. Moreover, even in the absence of the ligand, holo CaM is known to display conformational multiplicity, whereby the probability distribution between the available states depends strongly on the environmental conditions [Gsponer et al., 2008; Slaughter et al., 2005]. However, the structural details of the latter process remain elusive [Fallon and Quiocho, 2003]. As revealed by x-ray crystallography, extended dumbbell shape and compact form of the protein with a bent central linker are both representative of the active state of holo CaM [Babu et al., 1988; Fallon and Quiocho, 2003](Figure 5A).

In MIDSTLAB previous work utilizing PRS to study the conformational transitions of holo CaM [Atilgan et al., 2011], 3CLN was selected as the initial (extended) state and seven alternative forms with distinct conformations represented by the PDB codes 1RFJ, and 1MUX (extended states) and 1CDL, 1PRW, 1QIW, 2BBM, and 1LIN (compact states) were each considered as target structures. RMSD values were measured for each initial - target protein pair. 1PRW gave the highest RMSD value of 16 Å with respect to 3CLN among all of the target structures [Atilgan et al., 2011].

Despite other protein systems which easily lead to PRS overlap values of ~0.9, CaM has lower maximum values due to the complexity of its conformational change. Therefore, here we describe PSP optimization for CaM as an example of how to achieve the highest possible value using this method. The inputs to PRS that have been monitored, the resulting maximum overlaps, and key residues determined are listed in Table 4.

Table 4. PSP optimization part 1: PRS input optimization on the starting structure 3CLN

PRS input monitored	Initial structure time point	Target structure	P	Trajectory Chunk	Oi	Key Residues, K*
A. Perturbed structure and trajectory chunk	Crystal	1PRW	600	10-90ns	0.60	27, 106, 105, 110, 28
	100 ps	1PRW	600	10-90ns	0.65	106, 28,105,110, 27
	20 ns	1PRW	600	10-90ns	0.68	106, 27, 110, 28, 123
	30 ns	1PRW	600	10-90ns	0.69	27, 106, 124, 105,110
	20 ns	1PRW	600	20-80ns	0.70	106, 124, 122, 27,31
	30 ns	1PRW	600	30-80ns	0.75	106, 105, 26, 122, 118
B. Number of perturbations, P	30 ns	1PRW	400	30-80ns	0.71-0.74	106, 105, 26, 122, 118
	30 ns	1PRW	500	30-80ns	0.74-0.75	106, 105, 26, 122, 118
	30 ns	1PRW	600	30-80ns	0.75	106, 105, 26, 122, 118
	30 ns	1PRW	700	30-80ns	0.75	106, 105, 26, 122, 118
	30 ns	1PRW	1000	30-80ns	0.75	106, 105, 26, 122, 118
C. Target Structures, T	30 ns	1CDL	600	30-80ns	0.65	106, 105, 122, 27,26
	30 ns	1MUX	600	30-80ns	0.84	17, 41, 16, 39, 101
	30 ns	2BBM	600	30-80ns	0.70	105, 26, 122, 106, 27
	30 ns	1LIN	600	30-80ns	0.72	106, 105, 118, 28,110
	30 ns	1PRW	600	30-80ns	0.75	106, 105, 26, 122, 118
	30 ns	1QIW	600	30-80ns	0.68	26, 105, 106, 118,108
	30 ns	1RFJ	600	30-80ns	0.79	107, 108, 106, 105,30

Different lengths of trajectory chunks and initial snapshots (structure II) were utilized to achieve the highest PRS overlap value (O^i) (see row A in Table 4). Based on RMSD time-series of the 3CLN trajectory, a suitable chunk for PRS is selected from the plot (Figure 7). From the 100 ns trajectory, 10-90 ns, 20-80 ns and 30-80 ns parts have been selected for $\mathbf{H}^{-1}$ construction and subjected to PRS together with snapshots at the 100 ps time

point, as well as snapshots selected from within the chunk. Results show that the first frame of each chunk can be considered as a compatible initial state (II) for PRS calculations. Comparing the results, 30-80 ns of 3CLN trajectory together with the initial structure at the 30 ns time point give the highest overlap and are considered as a proper input for subsequent steps. Note that using more than P = 600 perturbations does not improve the overlaps (see row B in Table 4). Finally, the conformational change from 3CLN to each of the other target x-ray structures is examined in a series of PRS runs (row C in Table 4). These structures vary in conformation due to the type of ligands which bind the protein. 1MUX and 1RFJ, having extended structures similar to 3CLN, give higher PRS overlaps in comparison to compact conformations (0.84 and 0.79, respectively). On the other hand, the PRS overlaps corresponding to compact structures are ~0.7 at best. The highlighted residues in these analyses are 106, 105, 27 and 26. Amongst the compact states, 1PRW indicates the greatest PRS overlap of O^i = 0.75 due to key residues 106, 105, and 26 (Table 4). The 3CLN to 1PRW is selected as the test case to validate the PSP methodology, because the structural difference between 1PRW and 3CLN is the highest amongst all pairs studied (RMSD is 16 Å) and this transition the most complicated one. The rationale behind the PSP method is to use an educated guess for the collective variables to drive the protein from a starting conformation to another, targeted stable state. To test this hypothesis, key residues K* are pulled along the best PRS directions D* in SMD simulations. To optimize pulling velocity and spring constant, SMD experiments are repeated several times. Simulations along the direction which leads to the distortion of the protein and its secondary structure are discarded. Details of SMD simulations performed for parameter optimization are listed in
Table 5.

Table 5. PSP optimization part 2: SMD input optimization due to holo-CaM extended to compact conformational transition.

SMD Label	Selected trajectory chunk(ns)	Initial structure (III)	SMD atom	Fixed atom	k (kcal mol^{-1}Å^{-2})	v (Å ps^{-1})	Pulling Direction *	Time (ns)
P1	10-90	100 ps	106	7	80	0.05	-0.58, 0.27, 0.70	2.1

P2	10-90	100 ps	28	65	80	0.05	0.26, 0.78, -0.59	2.1
P3	10-90	20 ns	106	5	80	0.05	-0.57, 0.26, 0.53	2.1
P4	10-90	20 ns	27	113	80	0.05	-0.19, -0.69, 0.32	2.1
P5	10-90	20 ns	27	80	80	0.05	-0.19, -0.69, 0.32	2.1
P6	10-90	20 ns	106	5	80	0.05	-0.87, 0.40, 0.88	2.1
P7	10-90	20 ns	28	80	80	0.05	-0.26, -0.78, 0.49	2.1
P8	10-90	20 ns	110	80	80	0.05	-0.39, 0.84, 0.48	2.1
P9	10-90	20 ns	110	80	80	0.05	-0.49, 0.94, 0.51	2.1
P10	20-80	20 ns	106	6	80	0.05	-0.19, 0.53, 0.64	2.1
P11	20-80	20 ns	106	6	80	0.05	-0.14, 0.94, 0.94	2.1
P12	30-80	30 ns	106	7	80	0.05	0.09, -0.80, -0.35	2.1

P13	30-80	30 ns	106	7	80	0.05	-0.19, 0.88, 0.50	2.1
P14	30-80	30 ns	26	80	80	0.05	0.07, -0.18, 0.13	2.1
P15	30-80	30 ns	106	7	80	0.05	-0.09, 0.80, 0.35	2.1
P16	30-80	30 ns	105	7	80	0.05	-0.10, 0.64, 0.61	2.1
P17	30-80	30 ns	105	7	80	0.05	-0.13, 0.84, 0.76	2.1
P18	30-80	30 ns	26	80	80	0.05	0.07, -0.18, 0.13	2.1
P19	**30-80**	**30 ns**	**106**	**7**	**90**	**0.03**	**-0.09, 0.85, 0.35**	**2.2**
P20	30-80	30 ns	105	7	90	0.03	-0.10, 0.44, 0.61	2.2
P21	**30-80**	**30 ns**	**106**	**7**	**90**	**0.03**	**-0.19, 0.58, 0.53**	**2.3**
P22	30-80	30 ns	106	7	90	0.03	0.09, -0.80, -0.35	2.3
P23	30-80	30 ns	26	80	90	0.03	0.07, -0.18, 0.13	2.3

P24	30-80	30 ns	105	7	90	0.03	-0.13, 0.84, 0.76	2.3
P25	30-80	30 ns	26	80	90	0.03	0.07, -0.18, 0.13	2.3
P26	**30-80**	**30 ns**	**106**	**7**	**90**	**0.03**	**-0.02, 0.31, 0.51**	**2.3**
P27	**30-80**	**30 ns**	**105**	**7**	**90**	**0.03**	**-0.09, 0.58, 0.86**	**2.3**
P28	**30-80**	**30 ns**	**106**	**7**	**90**	**0.03**	**-0.34, 0.71, 0.86**	**2.3**

*Directions used as SMD inputs obtained from PRS runs are relative to the same coordinate frame and are listed to enable comparison between pulling directions

We find that a pulling velocity of 0.03 Å/ps, force constant of 90 kcal mol^{-1} $Å^{-2}$ are optimal for utilizing the SMD simulations for our current purposes; these constants are typical of values used in the SMD simulation literature [Martin et al., 2009]. To this aim, the RMSD between the reference 1PRW structure and each of the snapshots recorded in the SMD trajectories is monitored. Of the 28 pulling simulations performed, the five (shown in bold amongst the listed in Table 5) that display the lowest RMSD values along the pulling trajectory compared to 1PRW are studied in more detail. Note that, while we have performed PSP on the conformational change from 3CLN to 1PRW, we also compare the results with other target structures including all the crystal forms and an NMR ensemble of 160 structures (PDB code 2K0E). The RMSDs for selected structures are listed in Table 6. We find that there are many other compact structures that are attained in these simulations that have lower RMSD than 1PRW (lowest values are shaded in gray in Table 6), although the key residues and pulling direction is selected based on the latter.

Table 6.Minimum RMSD between states of top ranked SMD trajectories compared to selected PDB structures (Å).[1]

	PDB code of compared structure[2]	Simulation label					
		P19	**P21**	**P26**	**P27**	**P28**	**P19.R6**
extended forms[3]	3CLN	12.16	10.59	10.11	9.71	9.13	13.85
	1RFJ	11.71	9.44	9.13	8.55	8.19	13.69
	1MUX	14.32	11.64	11.56	10.41	10.15	16.38
compact forms	1CDL	5.56	8.48	7.01	11.47	8.69	1.94
	1LIN	5.58	8.92	7.58	11.76	9.11	2.89
	1QIW	5.34	8.60	7.17	11.51	8.86	2.11
	1PRW	**6.23**	**9.46**	**8.34**	**12.23**	**9.86**	**4.34**
	2BBM	5.55	8.55	6.85	11.57	8.73	2.43
	2K0E (36)	3.44	2.89	2.43	5.57	3.25	5.28
	2K0E (52)	2.90	4.38	3.23	7.53	4.52	3.55
	2K0E (114)	3.57	2.90	2.43	5.51	3.42	4.73
	2K0E (138)	4.76	7.03	5.34	10.15	7.06	2.95

[1] the targeted 1PRW comparisons are shown in bold; cells are shaded for the most similar structures in a given pulling / relaxation experiment. For the extended forms, the first 500 ps is not used in comparison, as these are similar structure I.

[2] For the NMR ensemble of 160 structures, the number in parentheses represent the index in the reported PDB structure 2K0E.

[3] for the extended forms, the first 500 ps is not used in comparison, as these are similar structure I.

The frame with the lowest value (III) for 1PRW (row shown in bold in Table 6) is saved for classical MD simulations to test if these transient structures are near the transition state that leads to the compact form; these relaxation simulations are labelled with the extension R; e.g. P19.R1, P19.R2, etc. The removal of the restraint many eventually lead to the return of the structures to the extended forms, finalize a conformational change to the compact form, or they remain in the transient state during the relaxation trajectory.

We exemplify further compaction in the P19 simulation which gives the lowest RMSD with respect to 1PRW. In this SMD, residue 106 located on C-lobe is pulled along the largest overlap PRS direction, and residue 7 residing on the opposite domain is fixed (Figure 8, thick black line). Six relaxation simulations are performed for durations of 5 (twice), 8 (once) and 10 ns (thrice) labelled P19.R1 – P19.R6 (Figure 8, gray lines). In

the relaxation simulation P19.R6 yields the lowest RMSD, not only with 1PRW, but also with other structures with bent central helix (Table 6, last column). By comparing the results of P19 and P19.R6 runs, we can follow how the structures tip over from the extended forms (3CLN, 1RFJ, 1MUX) to the compact ones upon relaxation. The final RMSDs between the best sampled states and 1CDL, 1LIN, 1QIW, 1PRW, 2BBM are 1.94, 2.89, 2.11, 4.34, 2.43 Å, respectively. Thus, 19R.6.CaM attains a conformation similar to the compact crystal form and some of the NMR ensemble structures while other relaxation simulations stay near closed transient molecular structure within the 10 ns window of observations.

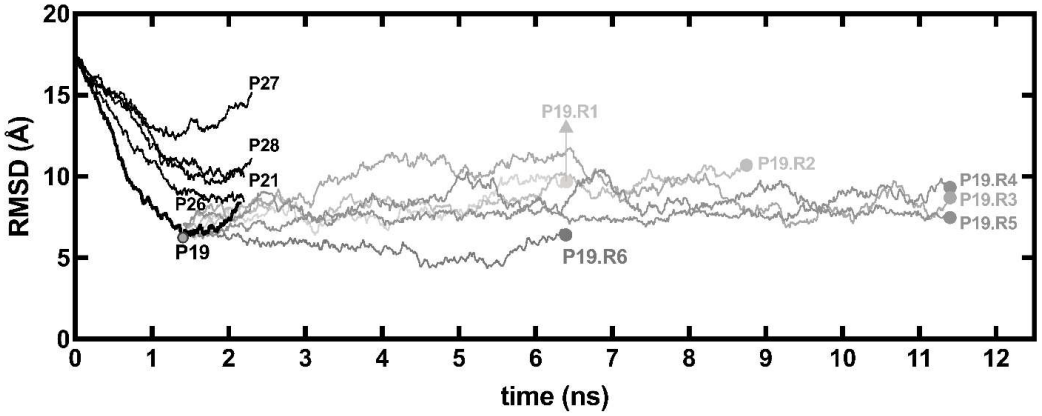

Figure 8. Progress of 3CLN (extended form) towards 1PRW (compact form) for selected five SMD trajectories (black), monitored by the RMSD between each point and the targeted 1PRW crystal structure. Swarms of relaxation runs is generated from the minimum point (III) of each trajectory; that for the P19 trajectory is displayed with six trajectories (shades of gray; termination points emphasized by filled circles) emanating from the gray encircled minimum point. Trajectory labels are illustrated on the figure.

Using PSP method, our assumption is to perturb a structure located in one minima of the free energy landscape and find the pathway that connects that minimum to another one. This process is reproduced using SMD simulations and completed by the relaxation simulation. Obviously, P19 that has the lowest RMSD with target structures likely passed the barriers that enabled it to stay at the transient state (P19.R1 – P19.R5) or even complete the process during the relaxation (P19.R6). While RMSD is an overall measure of the similarity between structures, the conformational change of CaM is quite complex and the steps to achieve a compact form are better described by a combination of motions. In MIDTSLAB previous work, we showed that displaying the CaM structure in a reduced conformational space described by two degrees of freedom conveniently

illustrates the main features of its motions [Aykut et al., 2013]. A dihedral angle that shows the motion of domains; and a distance which indicates the bending of linker were considered as reduced variables therein, which we also adopt here (schematically shown on the left of Figure 9). The former is described by the four consecutive points, center of mass (COM) of the N-domain, Cα atom of residue 69, Cα atom of residue 92 (residues at the two end points of linker) and COM of the C-domain. The latter is the distance between the Cα atoms of residues 69 and 92. Selected coordinates from the SMD and relaxation trajectories are projected on this space along with the crystal structures and 160 structures (PDB code 2K0E) which are listed in the NMR ensemble (Figure 9; individual NMR models are enumerated and plotted in Figure 10. The P19 simulation is colored in dark green and its relaxation simulation P19.R6 is colored in light green. Together they display the progress of 3CLN towards 1PRW in two steps. The relaxation trajectory covers the area of compact CaM structures and includes most of the compact target structures studied.

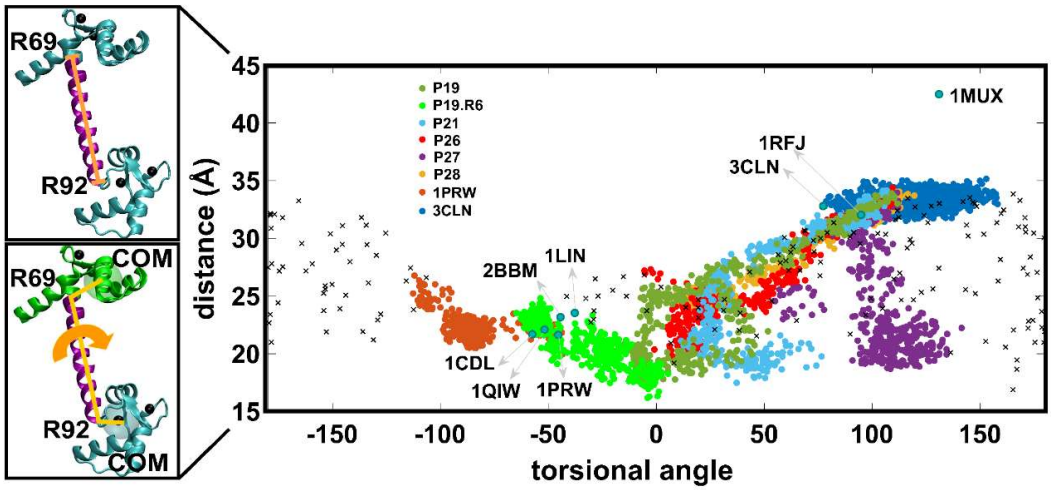

Figure 9. Conformations sampled by calmodulin, projected on the simplified two-degree-of-freedom model. Dihedral angle was measured between four points: center of mass (COM) of N-Domain, residues 69 and 92 and COM of C-Domain (lower left). Distance was measured between residues located on each side of central helix, 69 and 92, to trace its bending (upper left). Encircled dots: crystal structures; crosses: 2K0E NMR ensemble structures; colored dots: simulation trajectories as labelled in the inset. 3CLN and 1PRW represent the respective classical MD simulations.

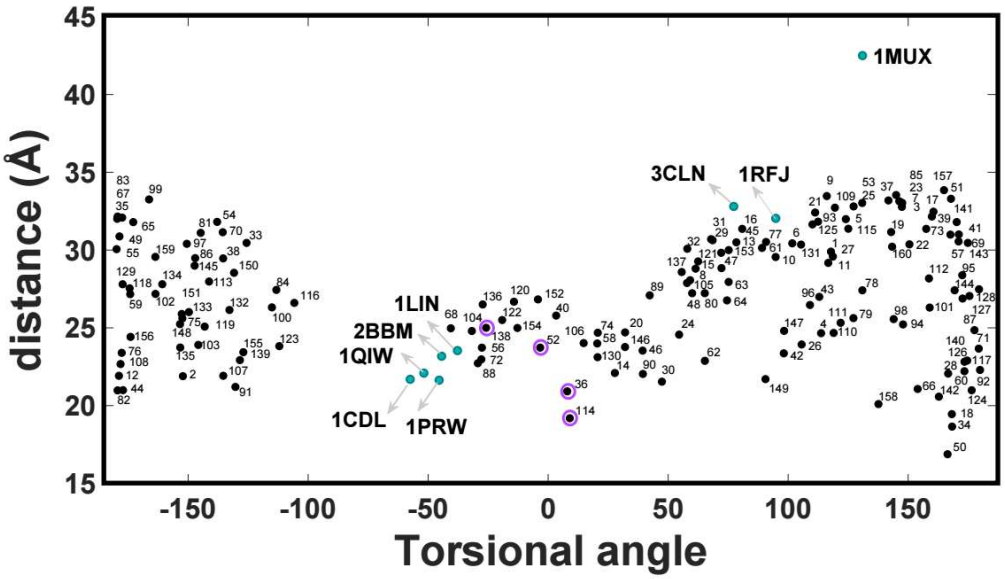

Figure 10. CaM structures projected on the two-dimensional reduced space of helix end-to-end distance vs. torsional angle representing the relative placement of the two lobes (see Figure 9, left). Crystal structures are displayed in cyan and labelled with their PDB codes; the NMR ensemble (PDB code 2K0E) containing 160 model structures are displayed by the numbered black dots or by encircled dots if they are similar to our PSP structures.

In MIDSTLAB previous work, we have shown that the extended linker of CaM is extremely stable. While it is relatively easy to achieve population shifts that sample the relative positioning of the N- and C-domains by changing the salt and pH conditions, it is a rare event to achieve a bent-linker conformation in classical MD simulations, even on the order of microseconds [Atilgan et al., 2011]. On the other hand, NMR and Förster resonance energy transfer (FRET) experiments show that both extended and compact conformations can be sampled in solution [Johnson, 2006]. It is evident that the landscape towards the bent-linker conformation is tightly controlled. Note that while in the SMD simulation set only P19 has a relatively low RMSD with respect to the compact crystal structures, they all achieve the bent linker conformation. Therefore, PRS can identify residues and directions along the free energy pathway connecting the two states.

PSP distinguishes between the landscapes of the forward and reverse transitions of calmodulin

To test the transferability of the PSP methodology, the reverse transition on the PES should also be attainable using the outlined protocol. Therefore, we have next studied the conformational transition of compact to extended form in CaM. To this aim, a single 100

ns simulation is performed using 1PRW as the starting structure. The 5-75 ns trajectory chunk of the 1PRW simulation is depicted in Figure 7 and the PRS results are summarized in Table 3 Table 4. Best PRS direction (D*) is found to be on residue 59, while residue 115 is chosen as the fixed atom. Using the aforementioned SMD parameters, the extended form is obtained which remains stable upon relaxation. In Figure 11 we summarize the PSP findings for the forward and reverse transition of calcium loaded CaM. Key residues are mapped onto the CaM extended and compact structures as shown in Figure 11A; the best pulling residue and direction are displayed in Figure 11B and D, respectively. The resulting final structures are superimposed on the targeted crystal forms in Figure 11C and E. The superposition of the final structure and target state indicates a significant overlap.

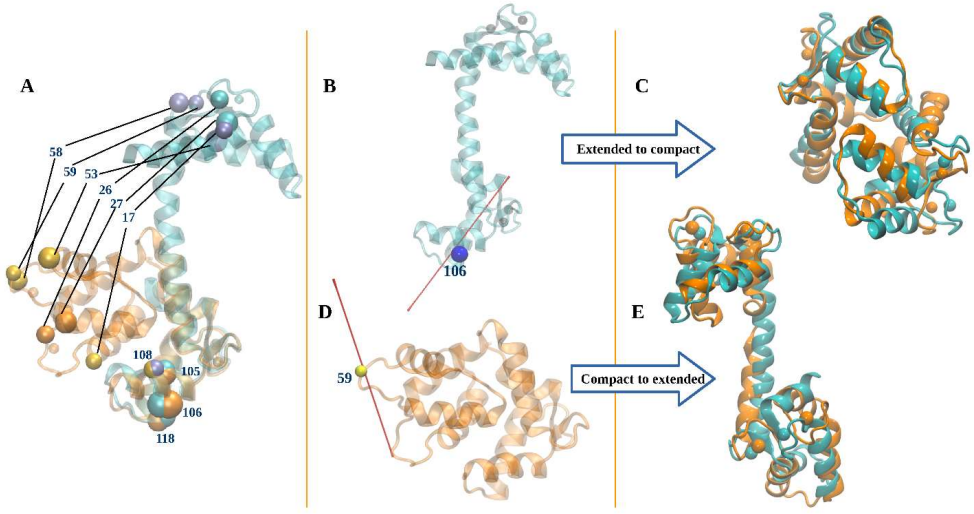

Figure 11. Applying PSP on CaM system to study extended to compact (A, B, C) and compact to extended (A, D, E) transition. A) Key residues which are effective in the transition identified by PRS. Cyan: 3CLN (extended); Orange: 1PRW (compact). Key residues effective in extended to compact transition colored in blue and orange on 3CLN and 1PRW, respectively. Key residues effective in compact to extended transition colored in purple and yellow on 3CLN and 1PRW, respectively. B) Residue 106 with the highest PRS overlap is depicted with the blue bead; red arrow shows the corresponding best PRS direction. C) Final structure (F) obtained from the extended → compact PSP scheme superposed on the target crystal structure (1PRW). Cyan: final structure (F); Orange: target structure (1PRW). D) Residue 59 with the highest PRS overlap depicted with the yellow bead; red arrow shows the corresponding best PRS direction. E) Final structure (F) which is obtained from the PSP methodology on CaM system superposed on the target crystal structure (3CLN). Orange: final structure (F); Cyan: target structure (3CLN).

Insofar as the PSP scheme allows for finding the ridges separating two desired conformations of a protein, it should be possible to quantify the relative free energy differences between the states through the identified transition state. Therefore the potential of mean force (PMF) profiles is reproduced using SMD (see ref [Park and Schulten, 2004] and section II-Methods). For the extended to compact PSP, we have selected P19 that has provided the best result. We have also chosen P28 that represents pulling of the same key residue (106), but along a slightly different direction (dot product between the two force directions is 0.86) as well as P27 representing pulling of the neighboring residue 105. PRS overlap is 0.75, 0.63 and 0.68 for P19, P27 and P28, respectively. To investigate the transition of the compact to extended transition, we selected the successful pulling of residue 59 along the best PRS direction (D*) and another one with a slightly lower PRS overlap (dot product between the two force directions is 0.81). A sample PMF of CaM computed with respect to the distance between initial and final position of the SMD atom is illustrated in Figure 12.

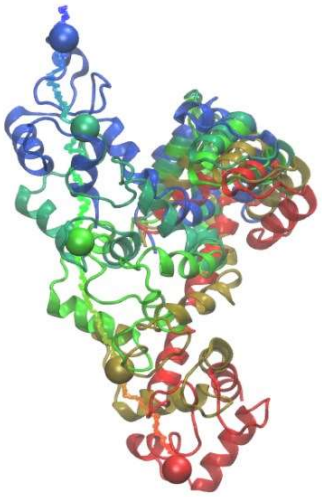

Figure 12. The distance between initial and final position of the SMD atom. Initial (red), midpoint (green) and last snapshots (blue) of the CaM conformational transition obtained in a sample SMD run. PMF profile is shown as a function of the distance this SMD atom has moved. Beads represent key residue 106 at different time steps; line indicates the pathway that the Cα atom of residue 106 travels from the initial to the final position.

Free energy paths of CaM from extended to compact form are plotted in Figure 13A, and the reverse cases are displayed in Figure 13B. We note that the second order cumulant expansion formula, which was used to obtain free energy difference values in this study, has been proved to perform better than exponential average under limited sampling. Still,

these values should not be used for quantitative interpretations [Patel and Kuyucak, 2017; You et al., 2019], but they should rather be utilized for ranking purposes as PMF values are overestimated. Both profiles in Figure 13 clearly display that the extended form of calcium loaded CaM is higher in energy than the compact conformation under the prevailing conditions. This is consistent with previous solution NMR studies reporting the linker's tendency to adopt the bent conformation in calcium-bound CaM [Barbato et al., 1992]. Additionally, FRET studies indicated that the compact form of holo CaM is more populated in solution except that the extended form is dominant in solutions with low calcium concentration [Johnson, 2006]. Moreover, the barrier to extended to compact transition is lower than that is required to get to the extended conformation as shown in Figure 13.

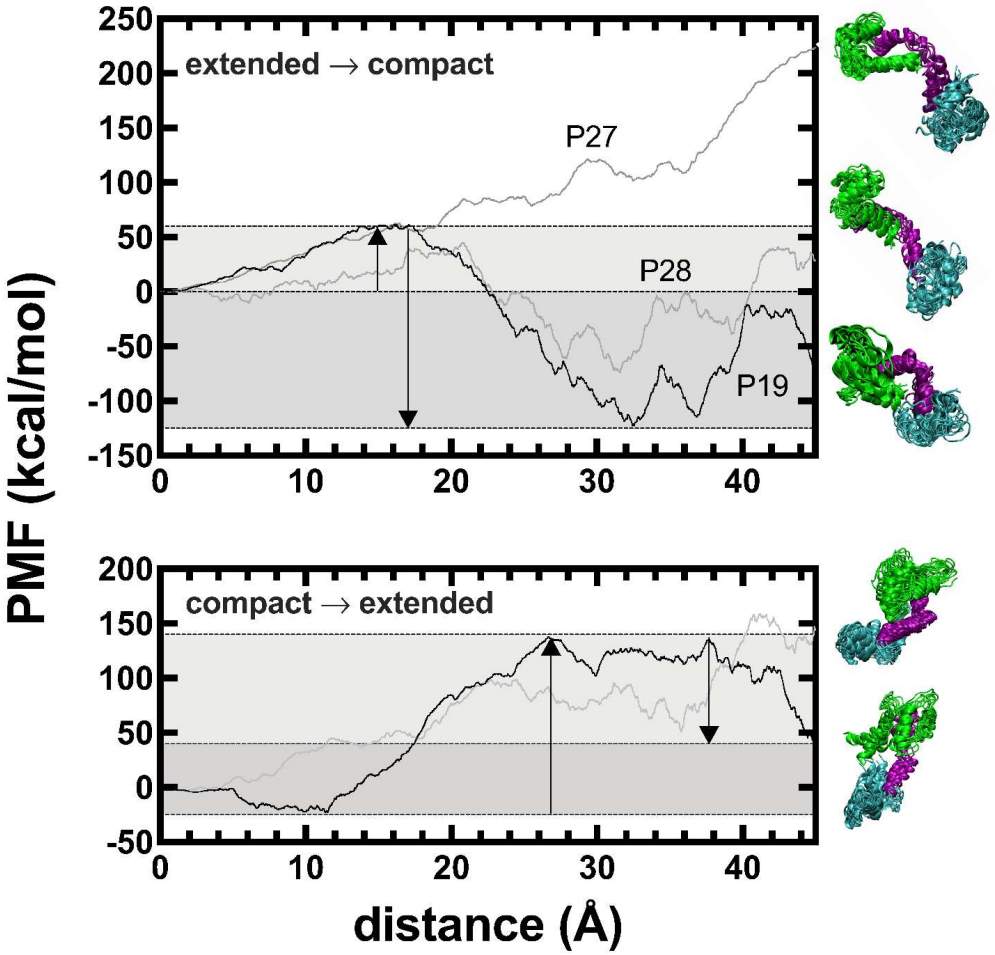

Figure 13. Reaction coordinates for conformational change of Ca-loaded CaM. Starting structures are depicted on the left; each simulation is repeated ten times and final structures are superposed on the right. In both cases, the compact form has lower energy. up) Extended to compact transition. 19P is pulled along the pathway, crosses a simple

barrier (up arrow) and reaches the minima corresponding to the compact conformation (down arrow); 27P and 28P selected as negative controls; 27P never reaches a low energy compact state, but explores a high, dead-end barrier; 28P enters a low energy barrier, but ends up in a semi-compact state that is less stable than that reached by P19. down) Compact to extended transition. 1PRW is pulled along the pathway, find a lower energy compact state before entering the on-pathway leading to extended form. The top of the barrier is more rugged than that for P19, but the pulling finally accomplishes reaching the minima corresponding to extended conformation (Black); Another pulling simulation along a direction with lower PRS overlap selected as negative control (gray); the final state is a compact, higher energy structure instead of the targeted extended form.

In fact, in the compact to extended transition, it is possible that the system attains an even lower energy compact state than that described by the x-ray structure. The barrier that is crossed is also flatter in the latter case. The results also indicate that the top ranked PRS results show a more favorable pathway for reaching the desired target structure and even a slight deviation might lead to a less favorable path with a higher energy final state despite a lower barrier (e.g. P28), or might enter a wrong path and not be able to attain a stable structure at all (e.g. P27). Subsequently, the error bars are presented on the PMF curves of successful forward and backward CaM pullings along the pathway results in lower energy in Figure 14.

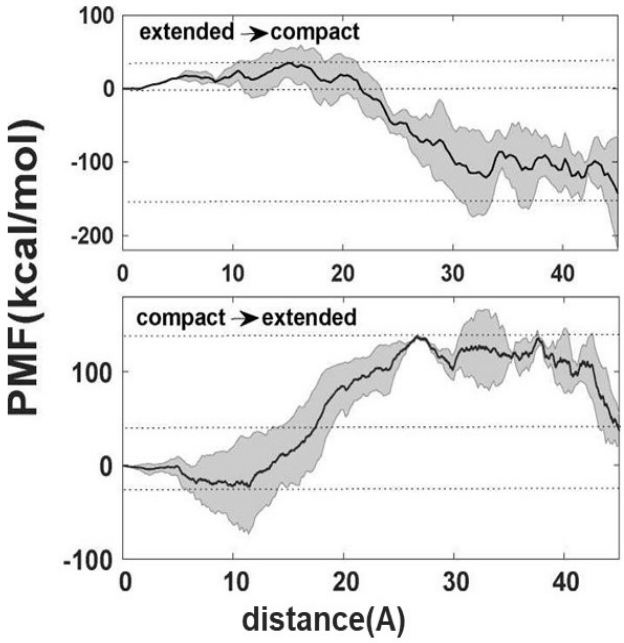

Figure 14. PMF calculation along the PSP determined the best-performing reaction coordinate based on results of 15 series of SMD simulations**,** which are depicted as bolded curves in Figure 13. PMF profile is shown as a function of the distance the SMD atom has moved. The error bars refer to the standard errors (gray areas). Up: The extended to compact calmodulin transition; down: compact to extended calmodulin transition.

PSP accomplishes the simple barrier crossing in adenylate kinase (ADK)

ADK can be considered as a simple model system for studying conformational changes in proteins, because it undergoes a large-scale, hinge-like loop motion which is relatively easy to conceptualize [Seyler and Beckstein, 2014]. ADK is a phosphotransferase enzyme that catalyzes the conversion of adenine nucleotides according to the chemical reaction, ATP+AMP ⇄ ADP+ADP, and plays an important role in cellular energy homeostasis by maintaining concentration of AMP, ADP, and ATP at the desired level. The three-dimensional structure of ADK consists of flexible regions and a central rigid part formed by parallel β sheets, surrounded by α-helices (Figure 5B). The central stable core is flanked by two highly dynamic domains, named LID-binding or LIDb (residues 114–164) and NMP-binding or NMPbd (residues 31–60)[Arora and Brooks, 2007]. LIDb and NMPbd are found close together when ADK is bound with a ligand or ligand-mimicking inhibitors (PDB ID: 1AKE)[Müller and Schulz, 1992] while they adopt a distant positioning in the absence of the ligand (PDB ID: 4AKE)[Müller et al., 1996]. Structural elasticity of the enzyme allows the interconversion between the closed and the open forms in the unbound state, while upon ligand binding the closed state is preferred. The closed conformation is the catalytically potent form since it provides the solvent-free environment needed for transferring the phosphoryl group [Pontiggia et al., 2008].

PSP is performed to study the ADK transition from the open state (4AKE), as the initial state, to the closed state (1AKE) as the target structure. As demonstrated in Figure 6 (flowchart), the first step provides a suitable chunk of the MD simulation which extensively samples the initial structure. Based on RMSD time-series of the simulation, the 25-75 ns chunk of the ADK trajectory, is selected for constructing $\mathbf{H}^{-1}$ (Figure 7). Subsequently, the 25 ns time point snapshot of the trajectory was subjected to PRS. Overlaps are calculated for all 214 residues present in the PDB file. Results summarized in Table 3Table 4 demonstrate that residues 146, 153, 151, 147, and 149, all of which are located on LIDb (Figure 15A), display the best overlaps (maximum $O^i = 0.89$).

The conformational transitions of ADK has been studied extensively via a plethora of biophysical and computational methods, and therefore it has become a testbed for evaluating path-sampling methods [Seyler and Beckstein, 2014]. We find that the top residue and direction plotted in Figure 15B closely mimics the first of the three collective variables used to bias the LID-core angle, the NMP-core angle, and the LID-NMP distances in, eg. bias-exchange metadynamics simulations[Li et al., 2015].[73] Moreover,

PRS generates a single "best" collective variable without resorting to the more complicated reaction coordinate optimization algorithms implemented in SCOOP [Tiwary and Berne, 2016], vac-MetaD [McCarty and Parrinello, 2017] or RAVE [Ribeiro et al., 2018].

Key residues are pulled along their best PRS direction (D*) in a series of SMD simulations. Here, C_α atom of residue 25 having the smallest deviation from the pulling direction and is at least 20 Å from the SMD atom is fixed. RMSD between 1AKE and each snapshot recorded during the SMD simulations is measured and the frame with minimum RMSD (structure III) is subjected to 5 ns relaxation simulations. The most similar frame to the target structure (1AKE) is saved as the final structure (F).

A sample RMSD value between the two end-point states, represented by crystal structures 4AKE and 1AKE, is 7.13 Å. RMSD between structure II and 1AKE is 6.80 Å which decreases to 3.11 Å after pulling residue 146 along the corresponding best PRS direction (D*). Relaxation simulation starting from final structure does not affect the RMSD value and validates the stability of the obtained structure. The whole process leads to the transition from the open to an intermediate state whereby the LID closes and contacts the NMP domain while the latter domain stays in the open state. The most probable transition pathway from the open to the closed form of apo ADK has been shown to go through this intermediate [Li et al., 2015; Shao, 2016; Wang et al., 2020] which was found to have 0.1 kcal/mol lower free energy than the open form [Li et al., 2015]. This intermediate is represented by several crystal structures in the PDB. The superposition of this final structure indicates low RMSD with 1DVR (1.1 Å), 1AK2 (1.2 Å), 2AK2 (1.2 Å), 2RH5 (1.3 Å), 2BBW (1.5 Å) (Figure 15C). Note that the final results are satisfactory despite the fact that for a small protein such as ADK, it is more difficult to find a fixed atom in the pulling scheme that will minimize rotational motions of the protein.

PMF results for this transition are displayed in Figure 15D (errorbars shown in Figure 16). Results indicate that PSP guides the protein through a relatively low-energy pathway, yielding the desirable target state. Note how the energy of the open and partially closed states are similar, separated by a relatively low energy barrier (~30 kcal/mol), consistent with the experimental observation that the interconversion between the two states are allowed in the unbound state of ADK [Henzler-Wildman et al., 2007]. A recent FRET study indicates that both open and closed conformations are sampled in the solution; however, the open form is dominant in the absence of ATP and the closed form is preferable in the presence of ligand [Aviram et al., 2018]. However, as we have noted

earlier, the PMFs obtained from SMD are not to be interpreted at a quantitative level; in fact, while our free energy pathway between these two conformers closely mimics that described by Li et al., it fails to clearly identify an energetically equivalent local minimum between these two end points. On the other hand, since the pulling direction and residue uniquely identified by PRS delineate a single collective variable for the favorable path, in future work, the SMD stage of the PSP may readily be replaced by more sophisticated sampling techniques to map the free energy surface of the protein.

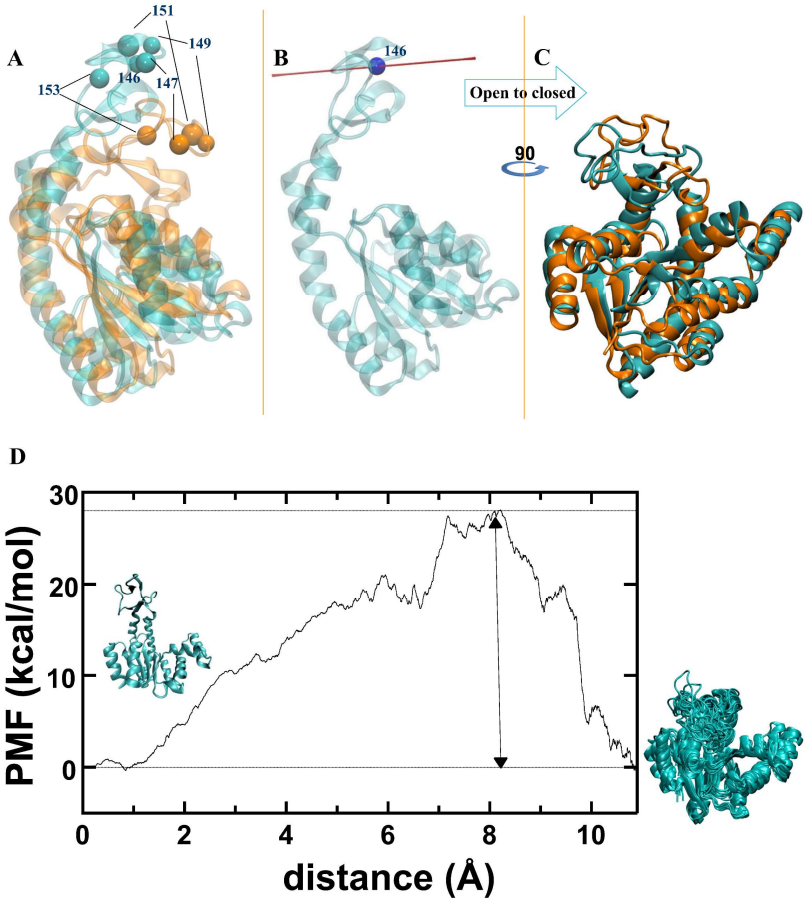

Figure 15. PSP on ADK, open to closed transition. **A)** Key residues effective in the transition identified by PRS. Cyan: 4AKE (initial structure); Orange: 1AKE (target structure). **B)** Residue 146 with the highest PRS overlap illustrated as blue bead and its corresponding pulling direction shown with red arrow. **C)** Final structure (F, cyan) obtained from the PSP methodology on ADK system superposed on the intermediate 2BBW (T, orange). Lateral view (up) and top view (bottom) indicate the proper overlap. **D)** Reaction coordinate for the conformational change of ADK obtained from SMD simulations. 4AKE is pulled along the PRS determined direction and reaches the minimum corresponding to closed conformation (Blue). The two conformations have similar energy; separated by a relatively low energy barrier (~30 kcal/mol in the PMF). Starting structure depicted on the left; SMD simulation repeated ten times and final structures superposed on the right.

We note that PSP readily lends itself to updates, so as to sample a conformational surface such as that of ADK which has several minima. Once a stable intermediate state is reached, one may update the pulled residue and direction by reapplying the PRS stage of the method followed by another SMD cycle. One may therefore navigate the conformational surface from minimum to minimum using sequential applications of PSP. There are other methods that couple enhanced sampling methods with continuous updating of collective variables obtained through slow modes calculated via ANM [Fuchigami et al., 2010; Fujisaki et al., 2013; Wang et al., 2019]. The advantage of PRS is to find a single collective variable that invokes several (collective) modes that are relevant to the sought-after conformational change. In contrast, using modes has the advantage of allowing continuous updating of the reaction coordinate throughout the sampling, but the inverse Hessian construction has the usual disadvantages within the framework of an elastic network model (e.g. cutoff distance selection).

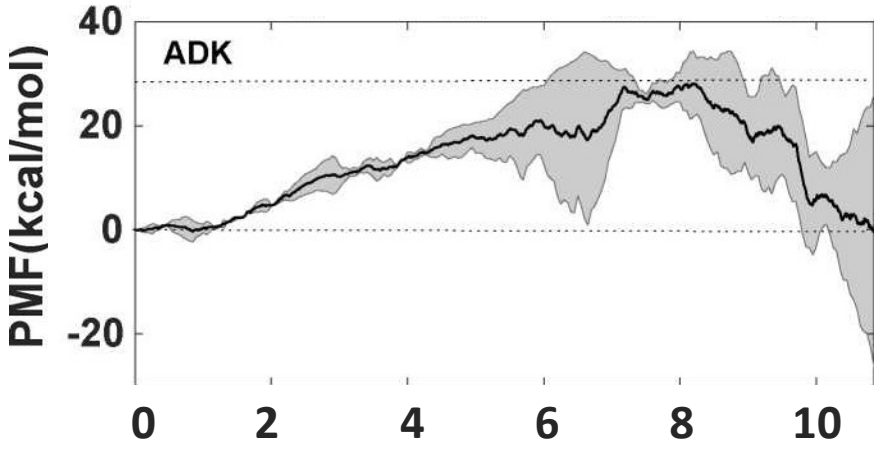

Figure 16. PMF calculation along the PSP determined the best-performing reaction coordinate of open to closed adenylate kinase transition, based on results of ten series of SMD simulations, which are depicted as bolded curves in Figure 15D. PMF profile is shown as a function of the distance the SMD atom has moved. The error bars refer to the standard errors (gray areas).

Iron binding dilemma observed in ferric binding protein (FBP) addressed by PSP scheme

Iron uptake pathway

Iron is a cofactor for many proteins engaged in fundamental biological processes and essential component in nearly all organisms. However, free iron can be toxic and harmful for the cell. Under physiological conditions, iron exists in its soluble form, as ferrous (Fe^{2+}), which is unstable and turns to poorly soluble ferric (Fe^{3+}) via the Fenton reaction ($Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^{\cdot} + OH^{-}$). The reactive oxygen species generated in this process are able to destroy proteins, lipids, and even DNA molecules [Krewulak and Vogel, 2008; Neumann et al., 2017]. Hence, iron uptake, transport and storage are tightly regulated extracellularly by eukaryotic binding proteins such as ferritin, transferrin, lactoferrin, and intracellularly by hemoglobin [Krewulak and Vogel, 2008; Sherman et al., 2018]. Transferrin scavenges free iron from the gastrointestinal tract and shuttles stored iron from liver to cells where it binds to its receptors and enter cells via endocytosis. The acidic environment of the endosome facilitates iron separation [Anderson and Frazer, 2017]. Lactoferrin is secreted from granules of the activated neutrophils to collect iron at inflammation sites. It is also found in mucus, tears, and milk [Tang et al., 2001; Weinberg, 2003]. Both gram-negative and gram-positive bacterial pathogens steal iron from their host to be able to survive at low iron concentrations. They either uptake "iron chelators," e.g. siderophores and heme, as intact molecule or they extract iron from host proteins [Krewulak and Vogel, 2008]. The iron uptake pathway in gram-negative bacteria such as *Neisseria gonorrheae* and *Haemophilus influenzae* occurs via a receptor-mediated mechanism. Two lipoproteins on outer membrane (OM), TbpA/TbpB and LbpA/LbpB, are involved in iron extraction from transferrin and lactoferrin, respectively. They form a trimeric complex which removes iron by forcing domain separation (Figure 17a). TbpB (LbpB in lactoferrin complex) is anchored to the OM by a long unstructured linker and facilitates the binding of transferrin, while TbpA (LbpA in lactoferrin complex) is a trans-membrane β-barrel protein which acts as a channel to import iron (Figure 17b) [Krewulak and Vogel, 2008; Neumann et al., 2017; Noinaj et al., 2012; Szewczyk and Collet, 2016]. The energy required for transport across the membrane is supplied by the proton gradient generated by the Ton system located in the inner membrane (IM) of bacteria. Ton system consists of three subunits: ExbB and ExbD proteins which together form the proton channel, and TonB which physically interacts with iron-loaded TbpA via its long

periplasmic domain. Ton box, a plug domain of TbpA, contains a conserved binding site for TonB protein which is exposed to the periplasmic side upon transferrin binding (Figure 17d)[Celia et al., 2019; Ciragan et al., 2020; Hickman et al., 2017].

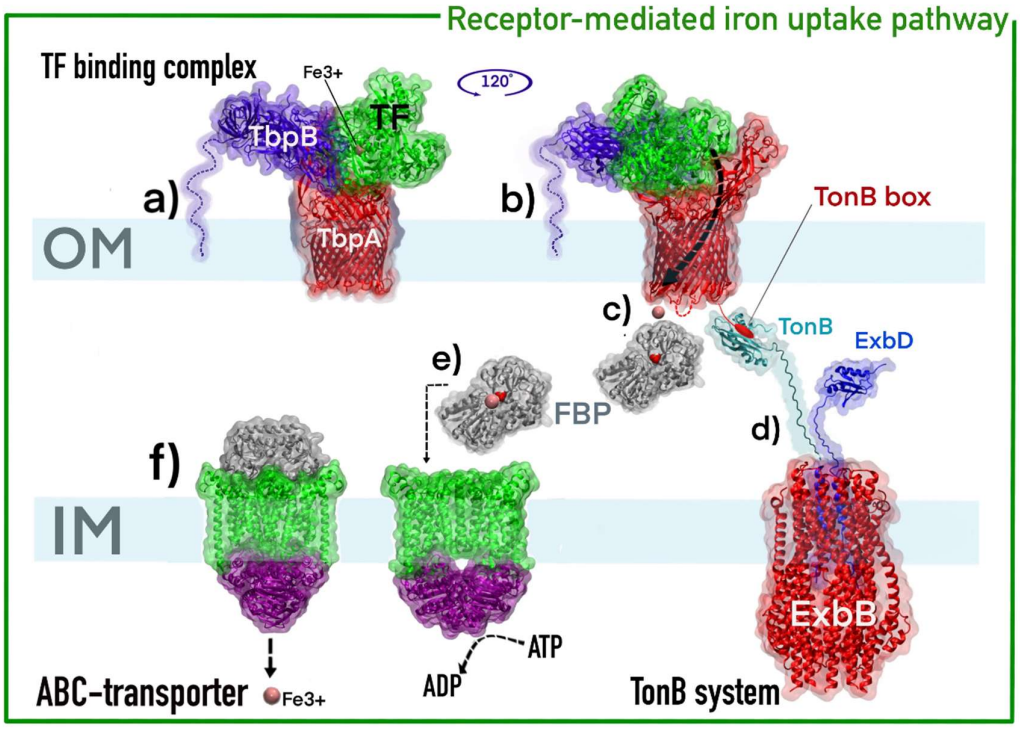

Figure 17. The iron uptake pathway by gram-negative bacteria. OM:Outer membrane; IM:Inner membrane; (a, b) Transferrin (TF) binding complex: Two lipoproteins, TbpA and TbpB, attract transferrin and form a trimeric complex to extract iron by forcing domain separation. The complex is modeled based on PDB codes 3V8X (Transferrin bound to TbpA) and 3VE1 (Transferrin bound to TbpB). (b, c) The extracted ferric ion is transferred to the periplasmic FBP (PDB code 1D9V). The black and red arrows indicate the pore formed to facilitate iron translocation and putative binding site for FBP (loop B), respectively. Ton box, a plug domain of TbpA, contains a conserved binding site for TonB protein which is exposed to the periplasmic side upon transferrin binding (shown in red). (d) Ton system is located in the inner membrane (IM) of bacteria and is comprised of ExbB, ExbD, and TonB proteins. It supplies the required energy for transporta across the membrane. TonB physically interacts with Ton Box via its long periplasmic domain. The 3D model is based on the recently reported Cryo-EM structure of bacterial Ton motor (PDB code 6TYI; ExbB shown in Red, ExbD shown in Blue) as well as 2PFU (the periplasmic domain of ExbD; shown in Blue), and 1XX3 (the periplasmic domain of TonB). (e) Iron loaded FBP binds to the ABC transporter (PDB code 1L7V: BtuCD). Upon binding, the ferric ion is released into a hydrophobic channel of transporter due to the distortion of the iron-binding pocket and the subsequent steric clash caused by a loop of the transmembrane domain (f) ATPase activity of the transporter leads to conformational change which allows ferric ion to translocate across inner membrane. Due to lack of experimental structure of FBP related ABC transporter, *Escherichia coli* vitamin B12 transporter is used as a similar model for this mechanism (PDB code 2QI9: BtuCD in complex with BtuF; Substrate binding protein).

Subsequently, the extracted ferric ion is transferred to the ferric binding protein which is responsible for shuttling iron in the periplasmic space (Figure 17c)[Guven et al., 2014; Jalalypour et al., 2020; Sensoy et al., 2017]. Finally, FBP interacts with ATP-binding cassette (ABC) transport system in the IM stimulating the conformational change of FBP. ABC transporter is comprised of two nucleotide-binding domains (NBD) having ATPase activity and two transmembrane domains (TMDs) so that the majority of the protein is located in the cytoplasmic side. The structure of FBP related ABC transporter is not available and iron release process is not fully clear. However, based on a previous hypothesis, the iron-loaded FBP binds to the periplasmic part of the transporter in its open state. Then, iron-binding pocket of FBP is distorted upon binding and iron is picked up by a steric clash caused by the loops of the TMD (Figure 17e). Subsequently, iron is released into the hydrophobic cavity of the transporter and the ATP-driven conformational change of the transporter allows iron to cross the membrane into the bacterial cytoplasm (Figure 17f)[Hollenstein et al., 2007; Locher, 2016; Rees et al., 2009].

Ferric binding protein structure

Typical structure of eukaryotic ferric binding proteins such as transferrin or lactoferrin consists of two lobes, each has the potential to bind one ferric ion, and connected by a short peptide linker. Each lobe is further divided into two similar sized domains [Abdizadeh et al., 2017]. This structural organization allows the protein to grab ligands between the two domains with a Venus flytrap mechanism [Berntsson et al., 2010]. Bacterial FBP which consists of two domains that are connected by a pair of antiparallel β-strands, exhibits a remarkable structural similarity to a single lobe of eukaryotic FBPs [Shouldice et al., 2004]. However, it has higher affinity to iron in acidic environment which helps bacteria obtain iron from the host transferrin [Khan et al., 2007a].

FBP is found in a variety of bacterial species such as N. gonorrhoeae [Bekker et al., 2004] and H. influenza [Bruns et al., 1997].This periplasmic transport protein hijacks the host iron and delivers it to the bacterial ABC transport system for releasing it into the cytoplasm[Wyckoff et al., 2006]. It consists of 46% α-helix, 17% β sheets and 37% disordered regions. Two characteristic domains of FBP, namely N- and C-domain, contain residues 1–82, 88–101, 226–276 and residues 83–87, 102–225, 277–309, respectively, in H. influenzae. Fe^{3+} is captured between these two domains along with engagement of a

phosphate anion in the binding site [Sensoy et al., 2017]. FBP has been crystallized in both iron-free, open (PDB ID: 1D9V) and iron-bound, closed (PDB ID: 1MRP) conformations. The open form dominates the population in the absence of the ligand, whereas the closed conformation is preferred when the protein forms a complex with Fe^{3+} and a phosphate anion [Guven et al., 2014]. FBP exhibits an octahedral coordination of iron by utilizing conserved residues from both domains including two tyrosines (Y195 and Y196), a histidine (H9), and a glutamic acid (Q57) residue as well as an exogenous phosphate ion and a water molecule [Khambati et al., 2010; Khan et al., 2007a]. FBP binds iron with remarkably high affinity in the open form on the order of $\sim 10^{18}$ M^{-1} [Bruns et al., 1997; Khan et al., 2007a; Sensoy et al., 2017], but is somehow able to readily release it at the inner membrane side of the periplasm, a phenomenon sometimes termed as "iron binding dilemma." We have applied the PSP scheme to address this problem, in particular to determine the relative positioning of the two conformations along the PES, as well as the free energy path connecting the two forms in both the forward (iron binding; open → closed) and the reverse (iron release; closed → open) transition.

For the forward PSP, the 45-100 ns chunk of the 1D9V classical MD simulation is selected based on the RMSD plot (Figure 7) and fed to PRS together with the 45 ns time point as the initial conformation (). Overlaps are calculated for all 309 residues and residues 31, 6, 33, 34, and 37 give the highest value of approximatively 0.91 (Table 4 and Figure 18A). SMD simulations are performed by pulling residue 31 and fixing residue 138. Cα atom of SMD residue is pulled along the best PRS direction D* (red arrow in Figure 18B). RMSD value between 1D9V (structure I) and 1MRP (T) crystal structures is 2.5 Å while that for structure II (45 ns time point) and T is 2.4 Å. The RMSD decreases to 0.98 Å after the SMD simulation which is further reduced to 0.80 Å during the relaxation run. Superposition of the final structure and target structure is illustrated in Figure 18C.

Regarding the reverse conformational transition, from closed to the open state with the Fe^{+3} bound to the protein, the 60-100 ns chunk of the 1MRP classical MD simulation was subjected to PRS along with the 60 ns time point of the trajectory as the initial structure (Figure 7 and Table 4). Residues 71, 70, 46, 45, and 231 give the highest PRS overlaps of approximatively 0.91 (Figure 18A, D). C_α atom of residues 71 and 143 are selected as SMD atom and fixed point, respectively. The RMSD value was measured as 3.1 Å between the initial conformation (60 ns time point of 1MRP) and the target crystal structure (PDB ID: 1D9V) which was further decreased to 2.1 Å after pulling (Figure 18E) and to 2.0 Å after the relaxation simulation. Previous studies reported that iron

uptake by FBP can occur in a thermally fluctuating environment, while its release is allosterically controlled [Atilgan and Atilgan, 2009]. A loop that spans residues 44–49, which contains conserved residues E45, G46 and T49, was reported as a significant region [Guven et al., 2014].

Our new results identify the same (E45, G46), and spatially nearby residues (L70, L71, and A231) as being operative in the closed to open transition which imitated iron release. These residues may be considered as allosteric residues involved in motions required for iron binding. In open to closed conformational transition, the identified residues are located on the N-domain near the binding site (Figure 18A).

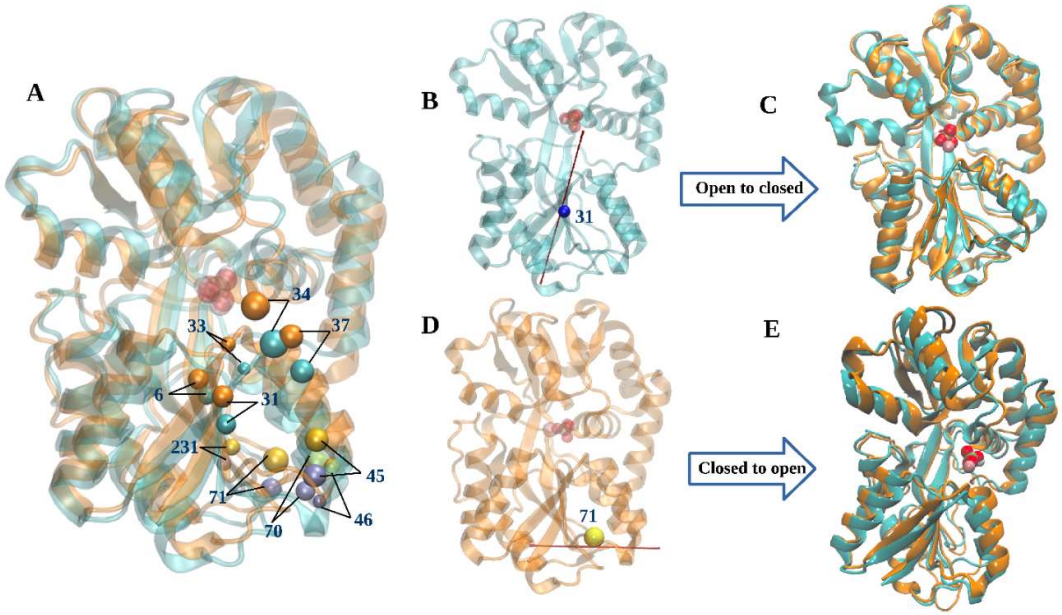

Figure 18. Applying PSP on FBP system to study open to closed (A, B, C) and closed to open (A, D, E) transition. **A)** Key residues in FBP conformational transition identified by PRS. Cyan: 1D9V (open); Orange: 1MRP (closed); RMSD is 2.5 Å. Key residues effective in the open to closed transition colored in blue and orange on initial structure and target structure, respectively. Key residues effective in the closed to open transition colored in purple and yellow on initial and target structure, respectively. **B)** Residue 31 with the highest PRS overlap illustrated as blue bead and its corresponding direction is shown as red arrow on the 1D9V crystal structure. **C)** Final structure (F) obtained from the PSP methodology on FBP system superposed on top of target crystal structure (1MRP). Cyan: final structure (F); Orange: target structure (1MRP); RMSD is 0.8 Å. **D)** Residue 71 with the highest PRS overlap and its corresponding direction is shown as yellow bead and red arrow, respectively, on the 1MRP crystal structure. **E)** Final structure (F) obtained from the PSP methodology on FBP; system superposed on the target crystal structure (1D9V). Orange: final structure (F); Cyan: target structure (1D9V).

As the conformational changes are predicted to be triggered by different phenomena, distinct PMF profiles are expected to accompany the two mechanisms. Also note that the open to closed simulations are devoid of the Fe^{3+} ion in the binding site, while the closed to open simulations include the bound cation. For the former transition, we find as a result of PSP that the closed form is located at a lower energy than the open form, and the transition is accompanied by a nearly flat barrier (Figure 19A); the error bars on the PMF (Figure 20) are large, pointing to the free sampling of the available space during the transition. This observation is consistent with the postulation that the apo FBP freely samples the open and closed forms and the transitions are triggered by thermal fluctuations in the absence of the iron and anion [Atilgan and Atilgan, 2009]. On the other hand, the opening process of FBP is not well understood and remains a controversial subject. Our PMF results indicate that in the prevailing conditions, i.e. 150 mM ionic strength, physiological pH and the presence of the phosphate anion coordinating the binding site, the iron bound open form is predicted to be more populated (Figure 19B); however, unlike the case for apo FBP, the barrier between the open and closed states is high. The error bars on the PMF (Figure 20) also point that unlike the open to closed conformational change which is controlled by thermal fluctuations, there is a specific pathway required to reach the open form from the iron-bound closed conformation.

The available direct experimental evidence indicates that the most populated state of the holo form is highly dependent on the environmental conditions. Thus, there is evidence to support our observations; e.g. several studies report holo form in the open conformation and mention that the closed conformation is obtained when specialized approaches such as crystallization in a proper buffer [Khambati et al., 2010] or in the presence of an anion such as phosphate were taken [Bekker et al., 2004]. However, phosphate ion is not essential for iron binding as confirmed by site-directed mutagenesis studies whereby crystal structures complex to iron without the synergistic anion [Bekker et al., 2004; Khan et al., 2007b]. It has been suggested that phosphate might be regulating FBP function by stabilizing the closed conformation [Bekker et al., 2004]. However, a small-angle X-ray scattering (SAXS) study indicated that the holo form tends to adopt an open conformation even in the presence of phosphate ion [Bulbul et al., 2018]. In fact, all mutated FBP proteins display open conformation while they still have iron or phosphate ion attached to them [Khambati et al., 2010; Khan et al., 2007a; Khan et al., 2007b; Shouldice et al., 2003].

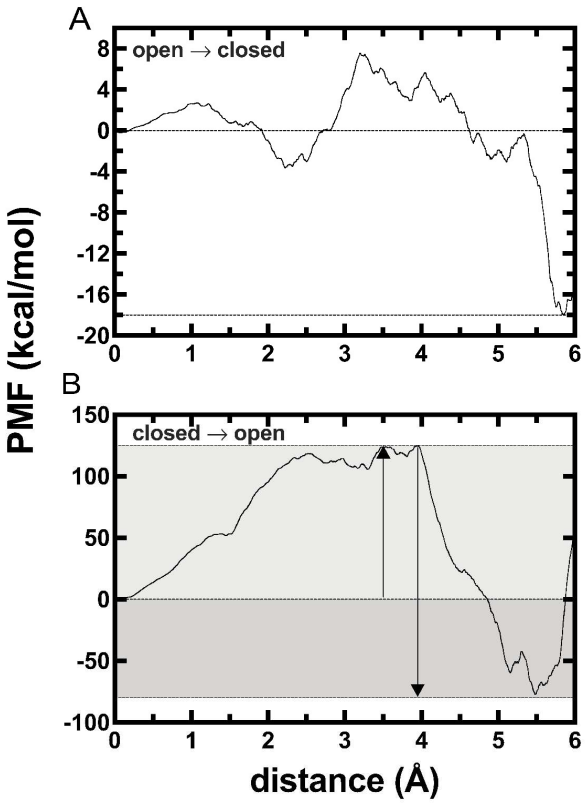

Figure 19. Reaction coordinate for the conformational change of FBP **A)** from open to closed, and **B)** from closed to open form. Starting structures are pulled along the favorable pathway and reach the minima corresponding to the targeted conformation; each simulation repeated ten times.

Site-directed mutagenesis of iron-coordinating residues as well as phosphate-coordinating residues indicate that, unlike transferrin, only one tyrosine is sufficient for iron binding and transfer process of bacterial FBP [Khambati et al., 2010; Khan et al., 2007a]. Mutation of coordinating residues only decrease the iron binding affinity which can be improved in an acidic environment, while mutants harboring double mutation of key tyrosine residues (Y195A/Y196A) is totally defective in iron binding and transfer [Khambati et al., 2010; Khan et al., 2007a]. Based on the crystal structures of reconstituted FBPs, wild type holo FBP can sample both open and closed form [Khambati et al., 2010; Shouldice et al., 2004]. Thus, even a slight change in coordination of the iron binding site would lead to open conformation. The details of structures are listed in Table 7.

Table 7. The details of experimental structures ()

Type	PDB code	Ligands	State		Organism
Wild type	1D9V	PO4	Apo	open	Haemophilus influenzae
Mutant (E57A)	2O6A	FE	Holo	open	Haemophilus influenzae
Mutant (Y195A)	3KN7	FE/PO4	Holo	open	Haemophilus influenzae
Mutant (Y196A)	3KN8	FE/PO4	Holo	open	Haemophilus influenzae
Mutant (H9Q)	1NNF	FE/EDT	Holo	open	Haemophilus influenzae
Mutant (Q58L)	2O68	FE/PO4	Holo	open	Haemophilus influenzae
Mutant (N193L)	2O69	FE	Holo	open	Haemophilus influenzae
Mutant (H9A)	1QVS	FE	Holo	open	Haemophilus influenzae
Mutant (N175L)	1QW0	FE	Holo	open	Haemophilus influenzae
Wild type	1MRP	FE/PO4	Holo	**closed**	Haemophilus influenzae
Wild type	1D9Y	FE/PO4	Holo	**closed**	Neisseria gonorrhoeae
Wild type	3ODB	FE	Holo	open	Haemophilus influenzae
Wild type	3OD7	FE/PO4	Holo	**closed**	Haemophilus influenzae
Wild type	1SI1	FE	Holo	open	Mannheimia haemolytica
Wild type	1SI0	FE, EDO, CO3	Holo	**closed**	Mannheimia haemolytica
Wild type	1O7T	Metal nanoclusters	-	open	Neisseria gonorrhoeae
Wild type	1R1N	Tri-nuclear oxoiron clusters	-	open	Neisseria gonorrhoeae

Nevertheless, FBP closing, opening, and iron release probably occur in several stages and are triggered by ionic concentration, pH modification or protein-protein interaction with ABC system subunits. Thus, the picture provided in Figure 19B reproduce only the first step in a multi-step process.

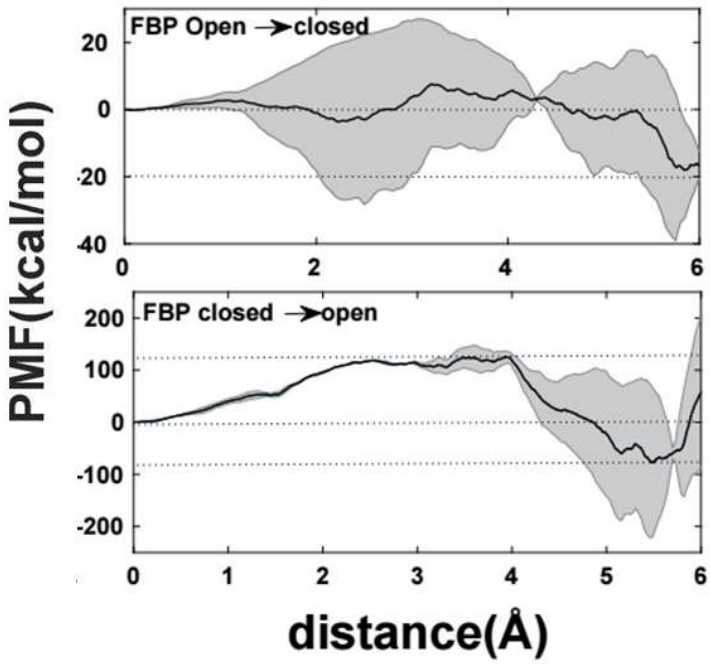

Figure 20**.** PMF calculation along the PSP determined the best-performing reaction coordinate based on results of ten series of SMD simulations. PMF profile is shown as a function of the distance the SMD atom has moved. The error bars refer to the standard errors (gray areas). up) open to closed ferric binding protein transition (apo form). down) closed to open ferric binding protein transition (holo form).

PSP method highlights the residues effective in conformational dynamics of Ras protein

The Ras protein family (H-RAS, N-RAS, K-RAS4A, and K-RAS4B) belongs to a small GTPases superfamily, having affinity for guanosine diphosphate (GDP) or guanosine triphosphate (GTP) nucleotides as well as intrinsic GTP hydrolysis ability, which is involved in cellular signal transduction[Azmi and Philip, 2017; McCormick, 1995] . Ras protein functions as an on/off switch and cycle between the active (GTP-bound) and inactive (GDP-bound) forms to mediate cell differentiation, proliferation, apoptosis, and survival. Hence, the Ras GDP/GTP switch process is tightly controlled in the cell by regulator proteins including guanine nucleotide exchange factors (GEFs)[Cook et al.,

2014] and GTPase activating proteins (GAPs)[ten Klooster and Hordijk, 2007]. GAP expedites the intrinsic GTPase activity of Ras and facilitates GTP hydrolysis leading to the formation of inactivate Ras, whereas GEFs stimulate the release of GDP which allows binding of GTP and activation of Ras protein [Ilter and Sensoy, 2019].

Ras mutations are responsible for relatively 30 % of several types of human cancers. Hotspot mutations occur frequently at either residues 12 (G12D, G12V), 13, or 61 and hinder binding of GAP, which subsequently increases the lifetime of activated Ras by reducing its intrinsic GTPase activity [Ilter and Sensoy, 2019; Parker and Mattos, 2015]. Targeting these proteins could be considered a novel approach for cancer treatment. However, Ras is thought to be undruggable due to the lack of deep binding pockets on its surface as well as the high binding affinities of the Ras ligands (picomolar level) which prevent competitive binding of the inhibitors and therapeutic agents [Cox et al., 2014].

The typical structure of Ras consists of two lobes; G/catalytic domain which is highly conserved among Ras family members (residues 1-166), and allosteric domain (residues 87-166). Ras family members anchor to membrane by hypervariable region (HVR) at their C-terminus (residues 173-189). The catalytic domain harbors the effector lobe (residues 1-86) which includes P-loop (residues 10-17), switch I (residues 30-40) and switch II (residues 60-76) loops which mediate GTP hydrolysis and interact with other regulator proteins such as RAF, GAP, GEF, and PI3K [Johnson et al., 2017; Parker and Mattos, 2015].

Switch I, in its open form, prevents the interaction of the Ras with its effector proteins, whereas binding of these proteins results in stabilization of Switch I while conferring flexibility to switch II. According to previous studies, phosphorylation of Y32 or Y64 residues has an impact on the conformation of switch I and II loops [Bunda et al., 2014]. The phosphorylation-driven conformational changes might potentially preclude protein interactions and inhibit the signaling pathway regardless of the activation state of Ras. Therefore, as a new approach, a flexible loop was designed by docking a drug to mimic the conformational change trigger by phosphorylation of the residue Y32 to increase flexibility of Switch I loop and alter protein stability with hopes of attenuating the Raf/Ras interaction. To follow the conformational modulation of Ras protein, PSP technique coupled with all-atom MD simulations is utilized. To do this, 2.5 μs MD simulations of wild-type and Y32-phosphorylated Ras, provided by the Sensoy lab [Ilter and Sensoy, 2019], is scanned to collect desired initial and target states based on the spatial positions of switch I and II loops. First, Ras structure is described with two degrees of freedom

which best indicate the protein motion. The distance between the C_α atoms of residue D12 and P34 shows the movement of switch I loop, and the distance between NH group of residue G60 and β-phosphate of GTP nucleotide illustrates the motion of switch II loop.

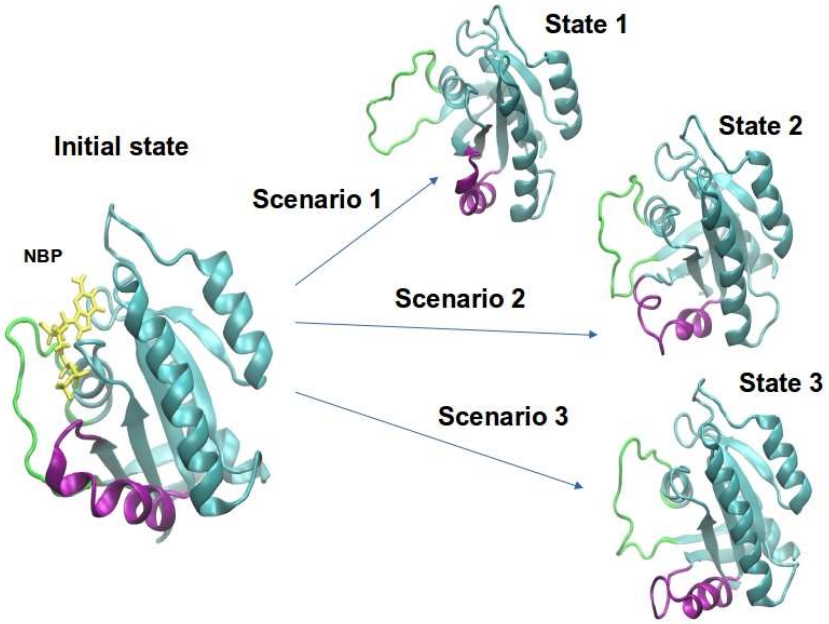

Figure 21. The initial state as well as three different conformations of Ras protein as target states determined based on the position of switch I and II loops. State 1) Switch I displaced from the NBP, while switch II remains in its initial state. State 2) Switch I is partially open and switch II is placed far from NBP. State 3) both loops are open and displaced from the NBP. Transition scenarios from initial state to each target state are numerated based on target state.

Accordingly, three conformers of the Ras protein are selected as target states; 1) switch I is displaced from the **nucleotide-binding pocket (NBP)**, while switch II remains in its initial state, 2) switch I is partially open and switch II is placed far from NBP, 3) both loops are open and displaced from the NBP. In total, four frames of trajectories including an initial state (fully closed) as well as target state 1 (switch I in open state), 2 (partially open), and 3 (fully open) are selected to represent the different structural minima of the Ras protein. Transition scenarios from initial state to each of target are numerated based on target states (Figure 21). The allosteric modulation between the initial and target states is investigated via separate PRS calculations. PRS calculations are optimized to obtain the highest overlap values. This method permits the identification of residues involved in Ras structural modification in describing various conformational changes from the initial

to each of the target states. Details of the structures as well as PRS results are summarized in Table 8.

Table 8. PRS calculation results of Ras protein system

structural states of Ras protein	Conformation states of switch I (D12-P34, Å)	switch II (G60-GTP, Å)	PRS selected residues (Transition scenario)	PRS overlap (O^i)
Initial state	closed (11.47)	closed (8.11)	-	-
Target state 1	open (25)	closed (9.72)	34,37,33,32,17 (scenario 1)	0.76-0.71
Target state 2	partially open (13.17)	open (12.01)	32,30,31 (scenario 2)	0.51-0.53
Target state 3	open (18.43)	open (22.33)	60,61,12,64,38 (scenario 3)	0.64-0.50

Results indicate that in the two first scenarios, residues located on switch I (34, 37, 32, and 30) are important in conformational modulation. However, in the conformational change from the structure with both loops closed to the fully open Ras protein, residues located on switch II (60, 61, 64) as well as residue 12 play a key role. Results are consistent with refereed studies which highlight the importance of these residues in Ras dynamics[Azmi and Philip, 2017; Ilter and Sensoy, 2019; Parker and Mattos, 2015].

Designing a flexible loop as a new strategy to alter protein stability and interrupt RAS/RAF interaction

To prove that the flexible loop designed by drug docking can mimic phosphorylation of residue 32 and is able to alter the stability of structure, PSP was applied on drug free and drug-bound Ras proteins to get the free energy profile. Binding a drug molecule is postulated to induce opening of switch I loop as best described in the first transition scenario. To prove the effectiveness of the drug molecule used to destabilize the structure, PMF is calculated in the first transition scenario for both drug-free and drug-bound Ras systems. Y32 and its best direction with overlap values (O_i) of 0.76 were used to perform SMD simulations. The structure was perturbed by pulling the C_α atom of Y32 along the

best direction. Results indicate the significant difference between PMF values of two systems (Figure 22). Manifestly, drug binding increases the flexibility of the structure and decreases its stability, which in turn leads to high PMF values.

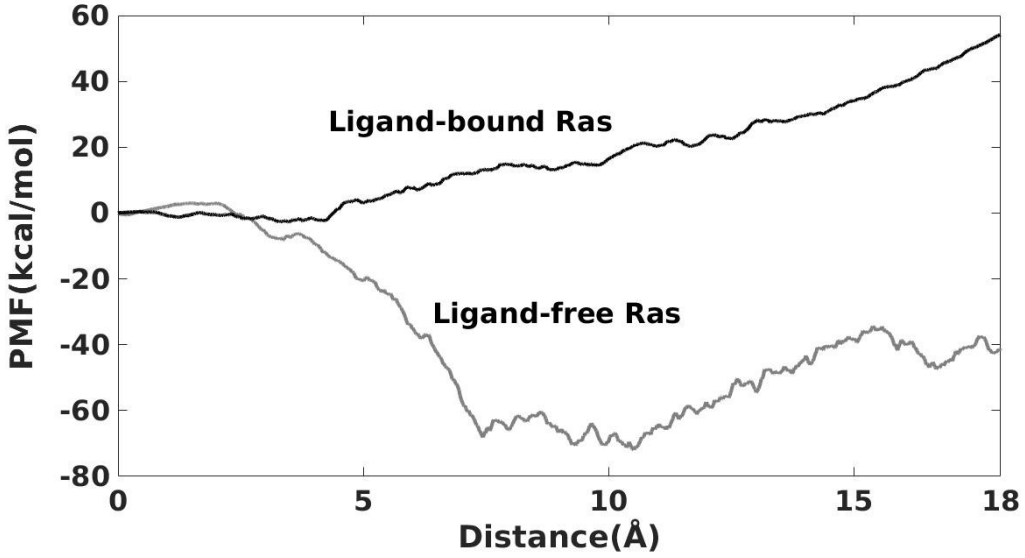

Figure 22. PMF calculated using the Jarzynski equality along PSP predicted coordinate with the highest overlap for the Ras protein transition scenario 1 (Switch I loop opening motion) as a function of distance.; each simulation was repeated 10 times.

Part IV. Study dynamics of a model protein via protein perturbation

Protein perturbation identifies residue 75 as an effective residue in dynamics of calmodulin

Introducing fictitious perturbations on proteins was first described as a systematic approach for conducting a survey of the conformations available to a protein in the perturbation response scanning (PRS) method [Atilgan and Atilgan, 2009]. Protein perturbation relates external forces applied on a structure, to displacements as responses via a covariance matrix derived from the classical MD simulations. Briefly, for a protein with N residues (i to N residues), the coarse-grained representation of structure is generated in which the C_α atoms are selected as N nodes. Then, to perturb the structure, random forces ($\Delta\mathbf{F}$) in various directions are applied on each node sequentially to generate displacement vectors ($\Delta\mathbf{R}$). The shift in the coordinates is calculated by equation 14.

Here, a similar approach to a method described by Nevin Gerek, *et al*. [Nevin Gerek et al., 2013] is proposed which utilizes the Hessian matrix ($\mathbf{C}$) to identify effective residues in the dynamics of the protein by considering the fluctuation of a structure in response to a single residue perturbation in M directions. To this end, displacement vectors generated by perturbation of each residue ($N \times M$ vectors) are sequentially recorded as the datasets (dataset i: displacement vectors of residue i perturbation, dataset i+1: displacement vectors of residue i+1 perturbation, …, dataset N: displacement vectors of residue N perturbation). For each dataset, displacement vectors of each residue are separately classified using K-means algorithm and cluster centroids are recorded. A centroid can be defined as a vector which represents the average displacement of a single residue. Lastly, obtained centroids are utilized to construct the final structures. To this end, coordinates of each centroid vector is added to the coordinates of the C_α atom of the residue on which

it is located. Eventually, each set is used to predict a single conformation. Figure 23 shows these steps schematically.

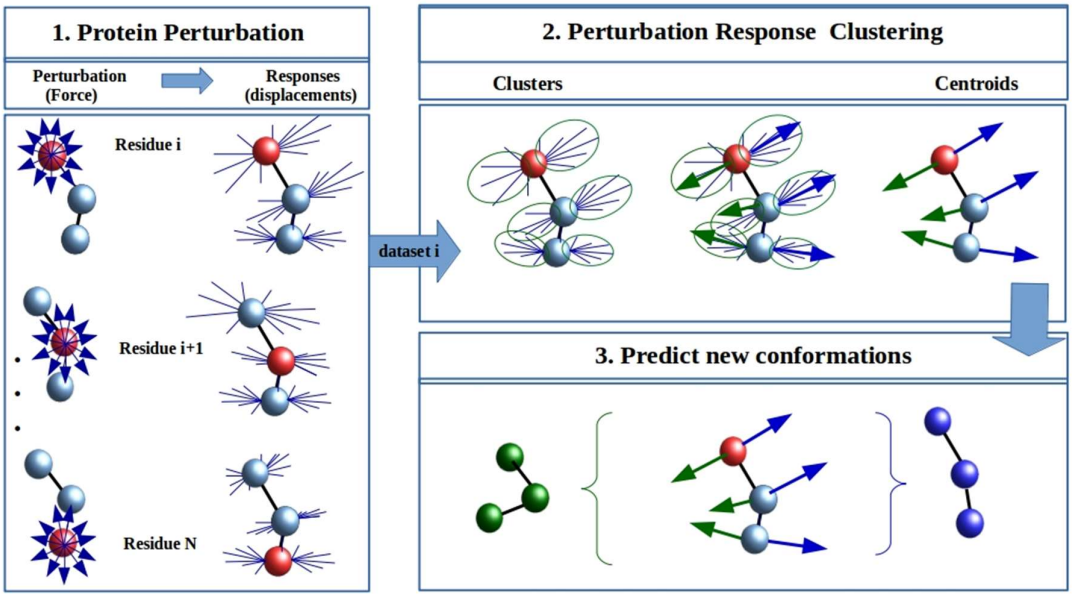

Figure 23. Perturbation response clustering to identify the key residues effective in dynamics of a protein is illustrated schematically.

Finally, the predicted structures are compared to the initial structure by measuring the RMSD to indicate the deviation from the initial state due to single residue perturbations and identify the residues with a significant role in the protein fluctuation and conformational modulations. The highlighted residue is mutated in silico and studied via a separate MD simulation to evaluate its effect on the protein dynamics.

Protein perturbation is applied on calmodulin protein based on the methodology previously described in section II. To do the perturbations, 400 ns MD simulation of calmodulin in its extended form (PDB code: 3CLN) is performed under physiological conditions. A 150 ns chunk of the trajectory (200-350 ns) is selected as the well-equilibrated part of the MD simulation (Figure 24; part shown in red), based on the RMSD of the protein compared to its initial state, and utilized to construct the cross-correlation matrix.

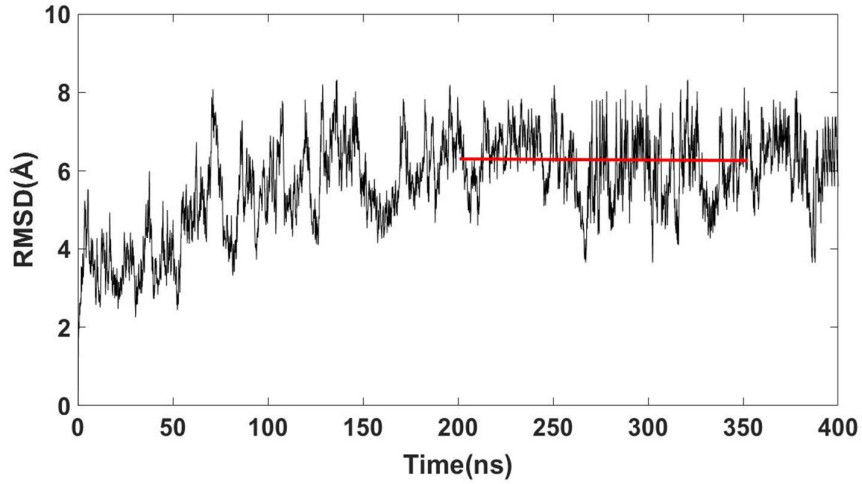

Figure 24. RMSD of CaM protein under physiological condition (pH=7.4)

Each and every one of the 143 residues of calmodulin (5-147) is perturbed 600 times to generate perturbation response vectors (M=600) and the obtained vectors ($\Delta\mathbf{R}$) are then accumulated in $N\times3N\times M$ matrix as the main input. Perturbing a single residue in 600 directions leads to 600 displacement vectors on each residue of the structure, including the perturbed residue itself. According to the methodology, 143 datasets are accumulated in separate matrices ($N\times3N\times M$). For each dataset, which only contains the response

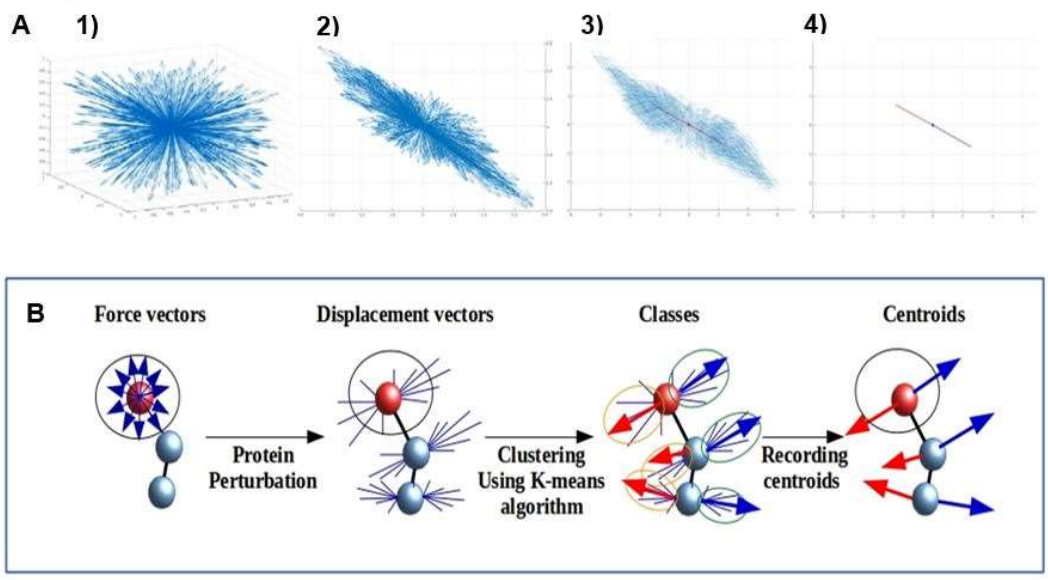

Figure 25. Protein perturbation clustering data presented on the first residue (i). In the actual implementation, the total movements of the protein as a whole, in response to single perturbations is clustered. A) data presented on residue 1; B) same data illustrated schematically. 1) external force vectors applied on selected residue to perturb the structure; 2) response (displacement) vectors obtained from PRS calculation; 3) clustering the data into k groups (here k=2); 4) cluster centroids indicate the average displacement of a single residue.

vectors of a single residue perturbation, the vectors on each residue are clustered into two groups (k=2) and the centroids are recorded in a $k \times 3N \times M$ matrix. For ease of visualization, the residue-by-residue prediction of the clustered displacements is presented in Figure 25. Subsequently, the Cartesian coordinates of each centroid vector is added to the coordinates of the residue of initial structure on which it is located, which results in constructing k new conformations from the perturbation of each single residue. Then, RMSD values are measured to compare the similarity between the predicted and the initial structures. The highest RMSD between the predicted structures and the initial structure (PDB code 3CLN) were 4.54 and 4.31 Å, yielded by perturbing residues 75 and 42, respectively (Figure 26).

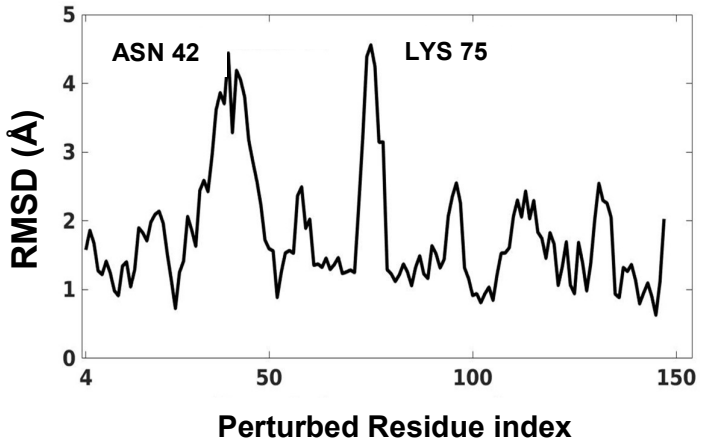

Figure 26. RMSD measured between clustering predicted structures and the initial state (open state, 3CLN). Perturbation of residue 42 and 75 lead to significant deviation of the structure from its initial state.

Additionally, clustering is applied on the closed form of calmodulin (1PRW) using previous simulations (Figure 7), and the highlighted residue was found to be 75 again. To evaluate the effect of residue K75 pinpointed via clustering, whose perturbation leads to significantly different displacement of the structure, mutation analysis was performed in silico. The K75A mutant was subjected to MD simulation. RMSD of the mutated protein was measured compared to the initial structure (Figure 27). Additionally, distance-torsional angle data, as described in Figure 9, were calculated for each frame of the simulation (Figure 27).

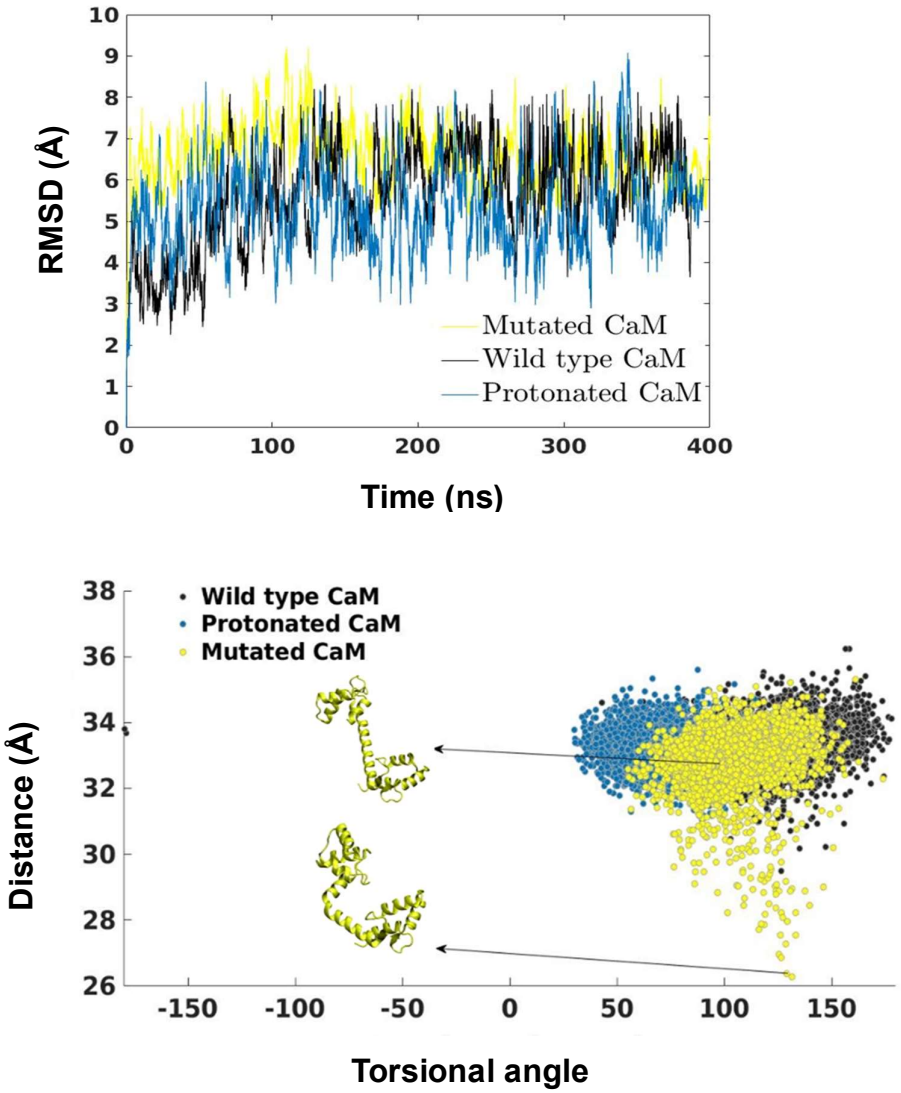

Figure 27. RMSD of the wild type, protonated, and mutated CaM systems. top) RMSD of wildtype CaM (black) protonated (blue) and mutated (yellow) compared to initial state; below) bending-torsional angles: wildtype CaM (black) protonated (blue) and mutated (yellow). Right: extended linker of mutated CaM. Right left: compact and bent linker of mutated CaM.

For comparison purposes, 400 ns simulations of open form calmodulin in acidic environment (pH = 5.5) is performed and the protonation state of the residues are determined using the H++ [Anandakrishnan et al., 2012] and PROPKA [Rostkowski et al., 2011] web servers. The acidic residues 11, 31, 67, 84, 93, 104, 122, 133, and 140 are upshifted in pK_a from the standard values of ~4–5.5, therefore they are protonated in the simulation to mimic pH 5.5 conditions. Results show that RMSD values are relatively similar for wild type and protonated CaM protein systems (Figure 27). However, protonation altered the position and orientation of the domains but does not have an impact on the linker which is also consistent with previous studies [Atilgan et al., 2011;

Aykut et al., 2013]. On the other hand, CaM with the K75A mutation tends to bend and form the structure similar to the closed form as shown in Figure 27.

To further study the role of residue 75 in bending of the central helix, the interactions of residues are examined in two trajectories, mutated CaM and P19 simulations (section V), both representing the linker bending. In P19 SMD simulation, the SMD atom is pulled along the best PSP direction with constant velocity and the applied force is recorded every 1 femtosecond (fs). Figure 28 shows the force values recorded in P19 SMD simulation plotted against the time. The interaction of residues including hydrogen bonds and salt bridges are monitored using the VMD timeline plugin.

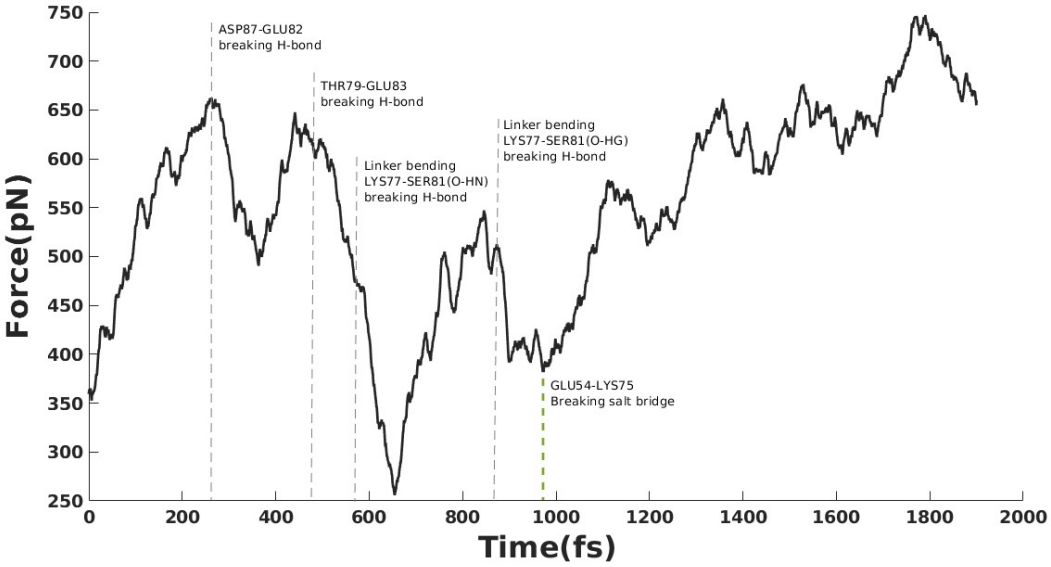

Figure 28. The force values applied on the SMD atom in P19 SMD simulation. The interactions between residues of the linker are determined using the timeline plugin and labeled on the figure.

Force plot indicates three main peaks which are associated with the breaking of hydrogen bonds between residues 87-82, 79-83, and 77-81 (Figure 28 and Figure 29). Finally, the salt bridge between residue 54 and 75 is destroyed which leads in linker bending.

As described in section V, the frame with minimum RMSD with 1PRW, the closed target state, is selected from P19 SMD simulation and subjected to relaxation simulation to complete the closing motion. Hence, the salt bridges are also monitored in two classical MD simulations, relaxation P19.R6 and mutated CaM K75A, which both indicate bending motion of CaM linker without applying external forces (Table 9).

Table 9. Salt bridges monitored in classical MD simulations for linker bending

Wild type simulation (P19.R6 relaxation simulation)	**Mutated CaM (K75A) Simulation**
GLU82-ARG86 GLU127-ARG126 ASP118-ARG106 GLU87-ARG90 GLU123-ARG126 ASP122-ARG126 **ASP80-LYS75** **GLU84-LYS75** **GLU54-LYS75** GLU45-LYS30 GLU114-LYS115 GLU83-ARG86	GLU127-ARG126 GLU83-ARG90 ASP78-ARG74 GLU123-ARG126 ASP22-LYS30 GLU87-ARG86 **ASP78-LYS77** ASP118-LYS115 **ASP80-LYS77** GLU87-ARG90 GLU45-LYS30 GLU83-ARG86 **GLU84-LYS77** GLU45-ARG37 ASP93-ARG90 GLU120-LYS115 GLU82-ARG86 GLU14-LYS13 GLU6-LYS13 ASP118-ARG106 ASP122-ARG126

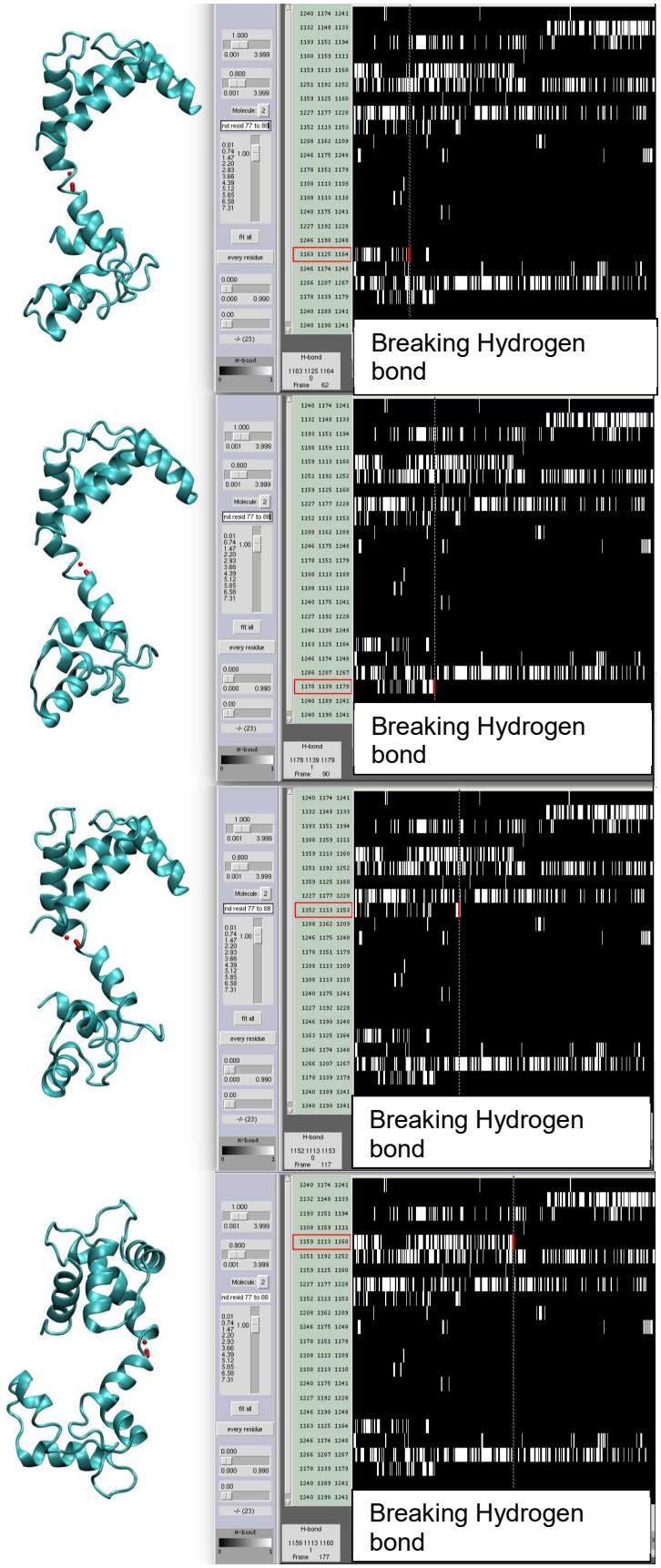

Figure 29. The hydrogen bonds monitored in P19 SMD simulation using timeline VMD plugin.

Lys75 is located in the middle of the linker, and forms a salt bridge with Asp78 in the extended state of CaM [Chattopadhyaya et al., 1992; Houdusse et al., 1997; Medvedeva et al., 2001]. Medvedeva *et al.* reported that the mutation in the position of residue 75 may have an impact on the dynamics and stability of the central linker [Medvedeva et al., 1999]. They claimed that the displacement of K75 with proline or charged residues e.g., glutamic acid may lead to an instability in the linker and its bending whereas, having hydrophobic residues (Ala, Val) at this position makes the linker more stable due to hydrophobic interactions with residues, 71 and 72 [Medvedeva et al., 2001; Medvedeva et al., 1999]. This assumption is in contradiction with our results which show that K75A mutation also destabilizes the linker.

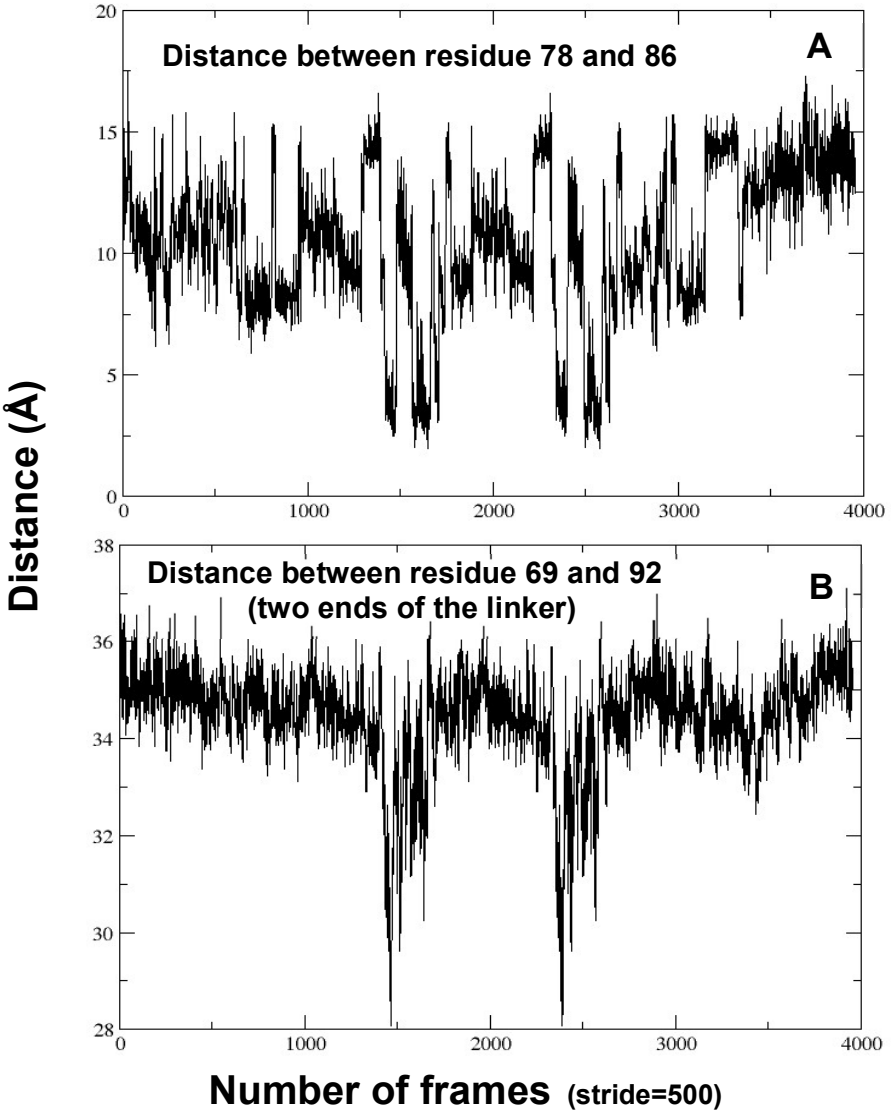

Figure 30. A. distance between residues 78 and 86. B. The distance between residues 69 and 92 which represents the bending of linker.

In P19.R6 relaxation simulation, residue 75 forms salt bridges with residues 80 and 84 to complete the closing movement (Table 9-bolded). In the simulation of mutated CaM, residue 77 mimics the situation where K75 is mutated to the neutral amino acid, Alanine. To understand the effective interactions in the bending of the central linker, the distances between residues located on the linker are investigated. Interaction between residues 78 and 86 is found to be associated with the linker bending. As shown in the Figure 30A, the distance between Asp78 and Arg86 is at its minimum at the time points where the linker is bent which occurs twice during the 400ns trajectory (Figure 30B).

The dynamics of calmodulin protein is altered in acidic environment

400 ns simulations of the CaM system under different conditions are used to evaluate the reproducibility of PRS and compare the results to previous findings. To this end, pdb codes 3CLN and 1PRW are selected as the open (initial) and closed (target) states, respectively. 400 ns simulations of the open form in the physiological (pH=7.4) and acidic (pH=5.5) environments are used to run PRS calculation from open→closed transition. PRS highlights residues 106, 105, 26, 118, and 115 as effective residues in the transition from open to closed form of CaM under physiological conditions. These results are consistent with our previous study (section III) [Jalalypour et al., 2020] and confirm the reproducibility of the PRS method. Applying PRS on CaM under acidic conditions highlights residues 30, 31, 101, and 118 as being effective in the open→closed transition which indicates that protonation modifies the dynamics of the protein.

Part V. Perturbation Response Clustering

Perturbation Response Clustering as a methodology to determine unknown conformational neighbors of a selected state

Previously developed methods, PRS and PSP, generate a set of N displacement vectors ($\Delta\mathbf{R}$) in response to each force, which can be utilized to predict a new conformation. Assuming the number of perturbations as M, in total, $N\times M$ structures will be predicted, which then a similarity parameter or overlap (O_i) is measured between each of them and the target structure. Accordingly, the best direction and its corresponding residue with highest O_i is selected to achieve the target state. Unlike PRS and PSP methods, which only consider the best force vector towards a known state and its corresponding set of displacements, clustering approach will take all sets of displacements, totally $N\times N\times M$ vectors, into account to explore new nearby conformations available to a protein. Here, a clustering approach was utilized to classify displacement vectors ($\Delta\mathbf{R}$) using K-means algorithm [Lloyd, 1982]. In this approach, all possible displacement vectors ($N\times N\times M$) are considered to indicate the total movements of a protein as a whole. To this end, centroids obtained from all datasets using the suggested method in section IV, are clustered based on the origin and direction of the vectors. Using this approach, the protein structure is divided into different domains based on the number of clusters (K), each having a representative vector which can be considered as collective variables (CV). K value is determined based on the size of a protein. shows these steps schematically.

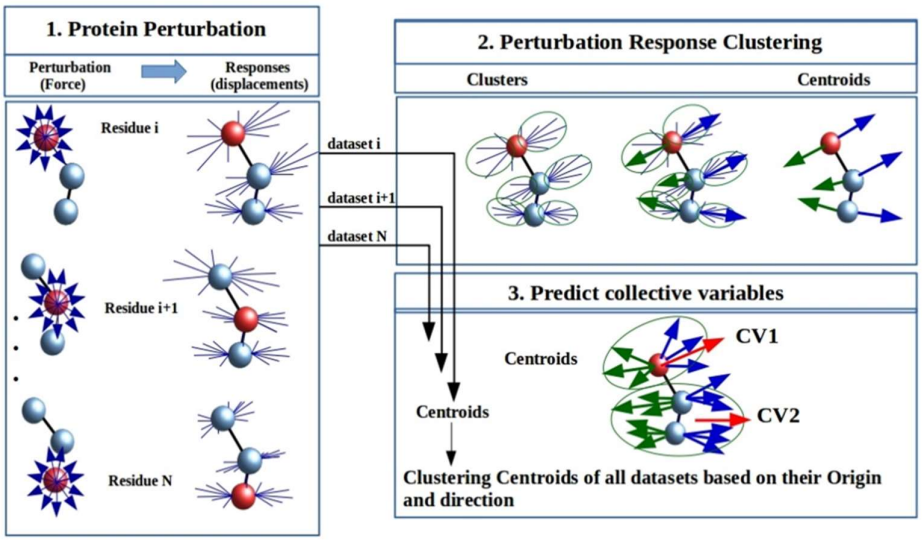

Figure 31. Perturbation response clustering to predict collective variables is illustrated schematically.

Perturbation response clustering reveals new conformational states of calmodulin

In the sections IV, we have discussed the $k \times 3N \times M$ matrix of centroids obtained from the perturbation of each residue, with k = 2. Here, all centroid matrices obtained from the perturbation of whole structure are accumulated and used as the main dataset (Figure 31). Eventually, the origin of each centroid vector is accumulated in the main dataset. Then, centroid vectors are clustered based on their origin and direction and classified into K subsets using the K-means algorithm. The number of clusters (K) is optimized based on the size of the protein. Calmodulin is divided to 6 regions (K=6) which covers all its moving parts. The final step generates the locations (origin) and directions of the average vectors of K clusters, called here as collective variables. The details of clustering via the clustering approach (run 1) is summarized in Figure 32 and Table 10. Besides, K-means starts with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. One can define initial values using k-means++ algorithm [David, 2007] to generate reproducible results. However, we take advantage of this issue and predict different CVs by repeating the runs. To this end, we produced 2 replicas titled as clustering run 2 and 3 (Figure 32).

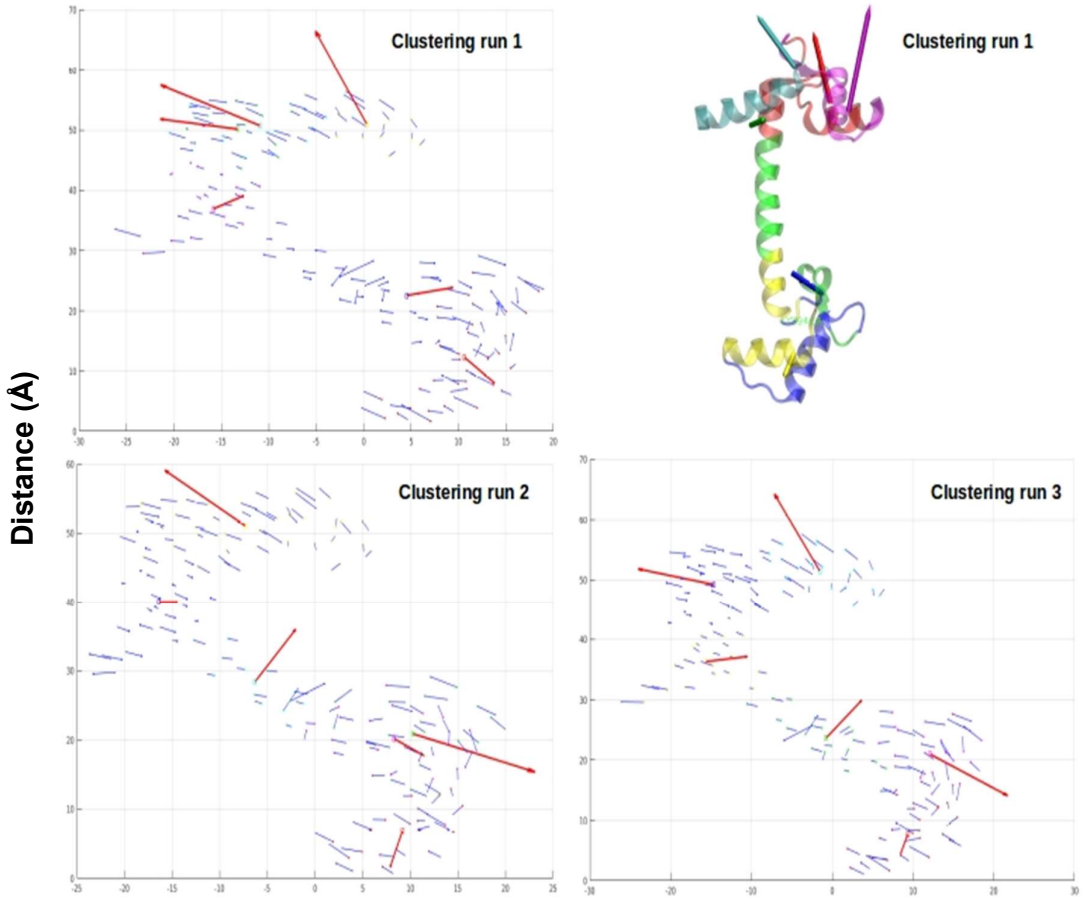

Figure 32. Perturbation response clustering to predict new collective variables. Clustering repeated 3 times to obtain different CVs. Top: left) The response vectors on each residue of CaM protein are clustered to 6 groups and the average of each group depicted as a red arrow. (For ease of visualization, average red arrow is magnified by the factor 1000) Right) clustering results represents in 3CLN pdb structure. Clusters 1 to 6 are shown in yellow, cyan, red, blue, purple, and green respectively. Bottom: clustering run is repeated twice using the same dataset (run 2 and 3).

Table 10. Residues resulting from the clustering shown in Figure 32, top, left (run 1).

Cluster (K)	Residues
1	90, 91, 92, 93, 94, 95, 96, 97, 98, 100, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116,117
2	5 , 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 65
3	26, 27, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 64, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75
4	99, 101,102, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137,140
5	28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51
6	76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 138, 139, 141, 142, 143, 144, 145, 146, 147

Steered molecular dynamics (SMD) simulations along perturbation response clustering predicted collective variables

The magnitude of the predicted CVs via each run is calculated and the top three CVs are fed into SMD simulations in order to predict new conformations. 100 ps frame of the trajectory (Figure 24) is used as the initial structure and the Cα atom of the closest residue to the origin of the CV is selected as the SMD stom and pulled along the CV (pulling direction). Furthermore, a Cα atom of the residue along the CV is selected as fixed residue based on dot product calculations as described in PSP methodology in section III. The details of the top three collective variables obtained from the different clustering runs which are used to perform SMD simulations are listed in Table 11.

Table 11. SMD simulation details performed along clustering predicted CVs; CVs are magnified by the factor 1000 to have consistence input with section III, However, SMD uses normalized direction. The hyphen indicates the distorted simulations.

CV origin (XYZ)			closest residue to origin	SMD Simulation	CV Direction (XYZ)			Fixed atom
Clustering run 1								
0.32	51.03	28.30	36	1.1	0.06	-0.17	-0.07	105
-10.94	50.76	18.82	52	1.2	0.12	-0.08	-0.02	133
-13.29	50.18	31.92	20	1.3	0.09	-0.02	-0.08	131
4.54	22.57	16.87	89	-	0.05	0.01	-0.05	136
10.62	12.31	8.82	121	-	0.04	-0.05	0.03	-
-15.86	36.93	28.31	69	-	0.04	0.02	0.03	-
Clustering run 2								
10.28	20.90	5.70	129	2.1	-0.14	0.06	0.01	57
-7.32	51.10	24.75	29	2.2	0.09	-0.09	-0.05	132
-6.30	28.41	20.13	78	2.3	0.05	0.08	-0.08	50
9.20	6.91	10.97	106	-	-0.01	-0.06	0.05	-
8.30	20.19	17.70	89	-	0.03	-0.03	-0.06	-
-16.39	40.04	30.43	12	-	0.02	0.00	0.04	-
Clustering run 3								
-1.55	51.52	25.62	33	3.1	0.06	-0.14	-0.05	136
12.13	20.97	11.72	136	3.2	-0.11	0.08	0.02	57
-14.65	49.24	25.51	63	3.3	0.10	-0.03	-0.05	131
-0.78	23.62	17.46	85	-	0.05	0.07	-0.07	-
-15.61	36.33	29.51	12	-	0.06	0.01	0.03	-
9.36	7.72	9.96	106	-	-0.01	-0.04	0.04	-

The RMSD between each frame of SMD simulations is measured compared to the target crystal and NMR structures listed in Table 1. The minimum values are reported in Table 12 for the crystal structures and Table 13 for the ensemble of NMR structures.

Table 12. The minimum RMSD between each frame of SMD simulations compared to target crystal structures

	1.1	**1.2**	**1.3**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**
3CLN	3.32	4.52	4.54	4.55	4.58	4.58	4.08	4.39	4.40
1RFJ	2.89	2.81	2.97	2.99	2.99	2.96	2.99	2.86	2.91
1MUX	6.17	6.80	6.80	6.78	6.80	6.80	6.80	6.80	6.80
1CDL	11.71	14.62	11.67	14.05	12.95	15.20	12.12	13.94	14.99
1LIN	12.28	14.66	11.62	14.46	13.40	15.34	12.58	14.42	15.06
1QIW	12.20	14.75	11.73	14.43	13.44	15.39	12.64	14.35	15.13
1PRW	13.33	15.64	12.17	15.56	14.22	16.31	12.87	15.60	16.04
2BBM	12.07	14.79	11.88	14.20	13.39	15.37	12.65	14.26	15.15

Table 13. The minimum RMSD between each frame of SMD simulations compared to target NMR structures.

	SMD Simulations (model/RMSD)								
	1.1	**1.2**	**1.3**	**2.1**	**2.2**	**2.3**	**3.1**	**3.2**	**3.3**
2K0E	**19** 3.85	**80** 3.62	**86** 4.22	**64** 3.61	**93** 1.93	**160** 6.04	**115** 2.64	**160** 4.30	**48** 3.40
	3 3.86	**48** 3.78	**22** 4.37	**80** 3.70	**21** 2.05	**9** 6.32	**19** 2.71	**115** 4.75	**96** 3.53
	79 3.90	**64** 3.85	**118** 4.42	**48** 3.96	**61** 2.24	**115** 6.41	**21** 2.96	**77** 4.80	**80** 3.56
2KDU	**19** 8.30	**18** 6.11	**19** 8.08	**3** 3.63	**8** 8.12	**18** 8.61	**19** 8.62	**18** 6.74	**18** 5.53
	13 8.36	**8** 6.45	**13** 8.28	**18** 4.25	**12** 8.61	**19** 8.62	**13** 8.67	**8** 7.55	**8** 6.08
	12 8.51	**1** 6.84	**7** 8.40	**1** 4.53	**19** 8.62	**13** 8.67	**12** 8.74	**1** 7.76	**1** 6.65

Finally, distance – torsion plots of CaM are prepared according to previously described protocol (Figure 9) to explore the conformational change of CaM. Unlike PSP which sampled structures along the transition path to achieve the target structure, PRC sampled new states which have not been captured earlier (Figure 33). All SMD simulations cover areas where experimental structures are located. In addition, simulation SMD 3.3 samples an area with no experimental structure. According to literature, new conformations of CaM (PDB code. 2KDU) are found in complex with the Calmodulin-binding domain of

Munc13-1 protein which is docked to a distinct binding site on calmodulin. The NMR structures contained in (2KDU) are located in the area where SMD 3.3 is sampled. The collective variables obtained from clustering method resulted in SMD simulations without distortion and they all sample different structural states of CaM.

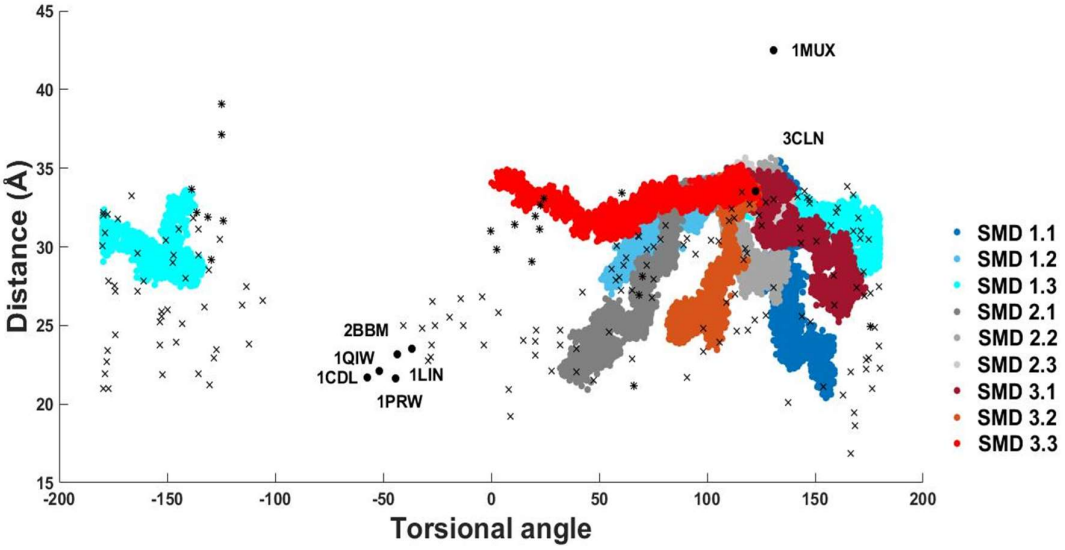

Figure 33. Conformations sampled by calmodulin, projected on the simplified two-degree-of-freedom model. Dihedral angle was measured between four points: center of mass (COM) of N-Domain, residues 69 and 92 and COM of C-Domain. Distance was measured between residues located on each side of central helix, 69 and 92, to trace its bending. Encircled dots: crystal structures; crosses: 2K0E NMR ensemble structures; stars: 2KDU NMR ensemble structures; colored dots: simulation trajectories as labelled in the inset: SMD 1.1 (dark blue), SMD 1.2 (blue), SMD 1.3 (light blue), SMD 2.1 (dark gray), SMD 2.2 (gray), SMD 2.3 (light gray), SMD 3.1 (dark red), SMD 3.2 (orange), SMD 3.3 (red).

VI. Conclusions

Proteins respond to various stimuli by adopting a variety of conformational states. Experimentally determined structures provide valuable information regarding conformations of the reactant and product, but they fall short of reporting intermediate states sampled along the pathway that links these states. On the other hand, achieving a holistic understanding of the mechanism of the reaction, hence the modulation of the function of the target protein, may be possible by investigating these intermediates which are located along the energy barrier and cannot be captured by means of classical MD simulations. Various computational schemes that direct the conformational change by imposing external forces or biasing the potentials have been devised, but the direction of change which is required to get to the final conformational state has usually been selected heuristically. In the first part of this thesis, PSP methodology was proposed which provides the key points and directions to apply forces to trigger a conformational change between the two preselected endpoints. The method readily handles the conformational landscape shifts accompanying changes that occur under selected environmental conditions. We have shown on four test systems that they could reach the target state from an initial conformation via the pre-designated pathway. The intermediate states sampled along the free energy pathways which are directed by high PRS overlaps may provide insight to the mechanism of the transition. Consequently, this knowledge can be used to guide development of therapeutics that can modulate the conformational transition processes of interest. PSP also has the potential to sample the high-energy transient conformations residing along the free energy pathway which are hard, if not impossible, to capture via common experimental methods. Thus, the PSP scheme outlines a clear approach to shed light on the underlying free energy landscape that governs the transition. Finally, we caution towards heuristic selection of force directions since we have also shown that even slight deviations in the selection of pulling directions may lead to false minima and a skewed view of the PES.

The PSP methodology gathers information by sampling the system in one conformational energy well and assumes that it provides information on how to reach a nearby free energy minimum. Moreover, it applies linear response theory to estimate this information. While, as shown in this study, it applies well to a range of proteins displaying different conformational landscapes, the main assumption of the methodology might fail in cases where minima have very different curvatures and are separated by high barriers. The limits of applicability of the protocol remains to be tested on a wider variety of systems in future work.

In the section IV of this thesis, calmodulin protein system is exploited as the model system to indicate the potentials of protein perturbation technique in studying proteins under different conditions and more importantly regardless of a target state. Additionally, calmodulin transition between two known states is studied in physiological and acidic conditions using PRS. In this section, a similar approach to the method developed by Nevin Gerek *et al.* [Nevin Gerek et al., 2013] is utilized to identify the critical residues in the dynamics of a protein. In this approach, the centroids obtained via clustering the response (displacement) vectors on each residue can be considered as predicted B-factors which describe the fluctuation of residues. Residue 75 of calmodulin, which is highlighted via this method, has been reported in several experimental studies [Medvedeva et al., 2001; Medvedeva et al., 1999] as a key residue having a significant role in CaM dynamics. However, its mechanism of function is not fully understood. To further investigate its role and to identify the effective interactions in bending of the central linker, hydrogen bonds and salt bridges are monitored via SMD and classical MD simulations which represent the linker bending event. The results show that the interaction between residues 78 and 86 is associated with the linker bending. On the other hand, the K75A mutant lacks formation of the salt bridge between residue Asp78 and Lys75, which is present in the extended form of wild type CaM [Chattopadhyaya et al., 1992; Medvedeva et al., 1999]. Lys75 may hinder interaction of residues 78 and 86.

In the section V, perturbation response clustering is proposed as a more generalized method to study the landscape of proteins. In particular, it may designate regions on the landscape that cannot be captured via crystallography or have not yet been detected via NMR or other methods. Using clustering method, new conformations of calmodulin are captured which are rarely available under normal experimental conditions and are only obtained when a protein interacts with a distinct binding site of calmodulin.

Biology has relied on trials and errors as a fundamental method for millennia. Nowadays, the amount of biological data is growing rapidly, and biologists prefer advanced methods to manage and analyze the big data rather than conventional ways. It is hard to suggest a treatment without knowing the exact molecular mechanism of function. Hence, tracing the structural change of a protein is important to understand the way it functions. However, studying protein motions and getting a full picture of its energy landscape is only possible via educated guesses and indirect approaches. Computational biology, molecular modeling and simulations provide a broad perspective in the hope of

understanding the function, structure, and the interactions of molecules at the atomic scale. *In silico* approaches such as MD simulations serve as a complement to *in vitro* and *in vivo* experiments and is able to reveal the details of molecular basis of biological functions, diseases as well as mechanism of action of conditions such as addiction or anesthesia. Besides, computer simulations provide effective ways to cut costs by reducing the number of time consuming, expensive wet-lab experiments, and more importantly, animal studies. For instance, computer simulations can be performed to test a theory in mutagenesis studies or structure-based drug design and narrow down a large number of candidates to be tested in wet-lab. Developing computational tools which are able to give us a full picture of protein dynamics in the least amount of time is of importance, particularly during times of crisis such as a pandemic. This thesis is focused on understanding the links between structure and function of proteins, in particular by identifying how individual residues in proteins influence the overall flexibility and function, say when a receptor binds a target or how proteins can be allosterically modulated by a small compound. It aims to combine simplified computational approaches with experimental methods on quite large scale by developing a practical toolbox to extract useful knowledge, facilitate the analysis, and be used by scientists all over the world. PSP and PRC generate new data regarding conformational transition and allosteric sites of proteins so as to assist in finding novel medications or medical applications.

This thesis has been written during the COVID-19 pandemic in 2020, and the above-mentioned methods are already used to study dynamic of SARS-CoV-2 spike glycoproteins to assist finding novel therapeutic agents and vaccines [Verkhivker, 2020], which indicates the importance of developing fast, precise, and user-friendly tools for the scientific community.

VII. Future work

The methodologies developed in this thesis can be applied to a wide range of proteins having different functions and displaying various types of motion. They can also be used to contribute new structures to the conformational ensembles provided by NMR. More importantly, these methods can be extended to study nucleic acids (DNA, RNA) or membrane proteins. For example, they can be utilized to study conformational changes which are allosterically triggered by lipid−protein interactions [Patrick et al., 2018].

In future work, the SMD stage of the PSP may readily be replaced by more sophisticated sampling techniques such as metadynamics [Leone et al., 2010] or enhanced sampling methods[Yang et al., 2019] to precisely map the free energy surface of a protein. It is also possible to combine PSP method with Markov state models (MSMs)[Chodera and Noé, 2014].

A method to investigate the effect of protein perturbation on the free energy of ligand binding may also be developed in future work. The protein perturbation predicted displacements could be linked to the thermodynamic cycle of system to identify functionally critical regions of proteins. Previously, a network-based method was reported to identify residues which thermodynamically coupled with binding of a ligand to GLIC (Li, X.Y, et al., 2014, Eur Biophys J). However, this approach could be extended to an advanced MD-based method such as protein perturbation which uses the trajectory to generate the fluctuation matrix and have the advantage over the structure-based approaches.

Secondly, the protein perturbation clustering can be improved in several ways: 1) The results of several runs can be collected and compared to determine the best CVs with the largest magnitude, 2) the results can be compared to the other methods such as normal mode analysis, ENM-based methods or methods which predict CV, and 3) K-means clustering method can be replaced by using more sophisticated machine learning approaches.

One major drawback of the methodologies developed in this dissertation is the limitation on performing the PRS calculations on complex protein systems (e.g. viruses) with an exceedingly large number of residues. To overcome this limitation, a proper coarse graining method such as that by Ross et al [Ross et al., 2018] can be utilized to reduce the number of nodes.

ABBREVIATIONS

PDB: protein databank

MD: Molecular dynamics

PRS: Perturbation response scanning

PSP: Perturb-Scan-Pull;

PRC: Perturbation Response Clustering

ADK: adenylate kinase;

CaM: calmodulin;

FBP: ferric binding protein;

PMF: potential of mean force;

PES: potential energy surfaces;

SMD: Steered molecular dynamics;

NBD: nucleotide-binding domain;

TMD: transmembrane domain;

RMSD: Root-mean-square deviation

VDW: Van der Waals;

NMR: Nuclear Magnetic Resonance spectroscopy

COM: center of mass

References

Abdizadeh H, Atilgan AR, Atilgan C. 2017. Mechanisms by which salt concentration moderates the dynamics of human serum transferrin. The Journal of Physical Chemistry B 121:4778-4789 %@ 1520-6106.

Abdizadeh H, Atilgan C. 2016. Predicting long term cooperativity and specific modulators of receptor interactions in human transferrin from dynamics within a single microstate. Physical Chemistry Chemical Physics 18:7916-7926.

Abdizadeh H, Guven G, Atilgan AR, Atilgan C. 2015. Perturbation response scanning specifies key regions in subtilisin serine protease for both function and stability. Journal of enzyme inhibition and medicinal chemistry 30:867-873.

Alder BJ, Wainwright TE. 1959. Studies in molecular dynamics. I. General method. The Journal of Chemical Physics 31:459-466.

Anandakrishnan R, Aguilar B, Onufriev AV. 2012. H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic acids research 40:W537-W541.

Andersen HC. 1983. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. Journal of Computational Physics 52:24-34.

Anderson GJ, Frazer DM. 2017. Current understanding of iron homeostasis. The American journal of clinical nutrition 106:1559S-1566S.

Arora K, Brooks CL. 2007. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. Proceedings of the National Academy of Sciences 104:18496-18501.

Astl L, Verkhivker GM. 2019. Atomistic modeling of the ABL kinase regulation by allosteric modulators using structural perturbation analysis and community-based network reconstruction of allosteric communications. Journal of chemical theory and computation.

Atilgan AR, Aykut AO, Atilgan C. 2011. Subtle p H differences trigger single residue motions for moderating conformations of calmodulin. The Journal of chemical physics 135:10B613.

Atilgan AR, Durell S, Jernigan RL, Demirel M, Keskin O, Bahar I. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysical journal 80:505-515.

Atilgan C. 2018. Computational Methods for Efficient Sampling of Protein Landscapes and Disclosing Allosteric Regions. Advances in protein chemistry and structural biology 113:33-64.

Atilgan C, Atilgan AR. 2009. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. PLoS computational biology 5:e1000544.

Atilgan C, Gerek Z, Ozkan S, Atilgan A. 2010. Manipulation of conformational change in proteins by single-residue perturbations. Biophysical Journal 99:933-943.

Aviram HY, Pirchi M, Mazal H, Barak Y, Riven I, Haran G. 2018. Direct observation of ultrafast large-scale dynamics of an enzyme under turnover conditions. Proceedings of the National Academy of Sciences 115:3243-3248.

Aykut AO, Atilgan AR, Atilgan C. 2013. Designing molecular dynamics simulations to shift populations of the conformational states of calmodulin. PLoS computational biology 9:e1003366.

Azmi A, Philip P. 2017. Targeting Rho, Rac, CDC42 GTPase Effector p21 Activated Kinases in Mutant K-Ras-Driven Cancereditor^editors. Conquering RAS. Elsevier, p 251-270.

Babu YS, Bugg CE, Cook WJ. 1988. Structure of calmodulin refined at 2.2 Å resolution. Journal of molecular biology 204:191-204.

Bahar I, Atilgan AR, Erman B. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design 2:173-181.
Bahar I, Rader A. 2005. Coarse-grained normal mode analysis in structural biology. Current opinion in structural biology 15:586-592.
Baker JL, Biais N, Tama F. 2013. Steered molecular dynamics simulations of a type IV pilus probe initial stages of a force-induced conformational transition. PLoS computational biology 9:e1003032.
Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. 1992. Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. Biochemistry 31:5269-5278.
Bekker EG, Creagh AL, Sanaie N, Yumoto F, Lau GH, Tanokura M, Haynes CA, Murphy ME. 2004. Specificity of the synergistic anion for iron binding by ferric binding protein from Neisseria gonorrhoeae. Biochemistry 43:9195-9203.
Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucleic acids research 28:235-242.
Berntsson RP-A, Smits SH, Schmitt L, Slotboom D-J, Poolman B. 2010. A structural classification of substrate-binding proteins. FEBS letters 584:2606-2617.
Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, MacKerell Jr AD. 2012. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. Journal of chemical theory and computation 8:3257-3273.
Bolia A, Woodrum BW, Cereda A, Ruben MA, Wang X, Ozkan SB, Ghirlanda G. 2014. A flexible docking scheme efficiently captures the energetics of glycan-cyanovirin binding. Biophysical journal 106:1142-1151.
Brotzakis ZF, Parrinello M. 2018. Enhanced Sampling of Protein Conformational Transitions via Dynamically Optimized Collective Variables. Journal of chemical theory and computation 15:1393-1398.
Bruns CM, Anderson DS, Vaughan KG, Williams PA, Nowalk AJ, McRee DE, Mietzner TA. 2001. Crystallographic and biochemical analyses of the metal-free Haemophilus influenzae Fe3+-binding protein. Biochemistry 40:15631-15637.
Bruns CM, Nowalk AJ, Arvai AS, McTigue MA, Vaughan KG, Mietzner TA, McRee DE. 1997. Structure of Haemophilus influenzae Fe+ 3-binding protein reveals convergent evolution within a superfamily. Nature structural biology 4:919.
Bulbul G, Liu G, Vithalapur NR, Atilgan C, Sayers Z, Pourmand N. 2018. Employment of Iron-Binding Protein from Haemophilus influenzae in Functional Nanopipettes for Iron Monitoring. ACS chemical neuroscience 10:1970-1977.
Bunda S, Heir P, Srikumar T, Cook JD, Burrell K, Kano Y, Lee JE, Zadeh G, Raught B, Ohh M. 2014. Src promotes GTPase activity of Ras via tyrosine 32 phosphorylation. Proceedings of the National Academy of Sciences 111:E3785-E3794.
Carlo M. 1995. Molecular Dynamics Simulations in Polymer Science, edited by K. Bindereditor^editors: Oxford University Press, New York.
Carroni M, Saibil HR. 2016. Cryo electron microscopy to determine the structure of macromolecular complexes. Methods 95:78-85.
Case DA. 1994. Normal mode analysis of protein dynamics. Current Opinion in Structural Biology 4:285-290.
Célerse F, Lagardère L, Derat E, Piquemal J-P. 2019. Massively parallel implementation of Steered Molecular Dynamics in Tinker-HP: comparisons of polarizable and non-polarizable simulations of realistic systems. Journal of chemical theory and computation 15:3694-3709.

Celia H, Botos I, Ni X, Fox T, De Val N, Lloubes R, Jiang J, Buchanan SK. 2019. Cryo-EM structure of the bacterial Ton motor subcomplex ExbB–ExbD provides information on structure and stoichiometry. Communications biology 2:1-6.
Chattopadhyaya R, Meador WE, Means AR, Quiocho FA. 1992. Calmodulin structure refined at 1.7 Å resolution. Journal of Molecular Biology 228:1177-1192.
Chodera JD, Noé F. 2014. Markov state models of biomolecular conformational dynamics. Current opinion in structural biology 25:135-144.
Ciragan A, Backlund SM, Mikula KM, Beyer HM, Samuli Ollila O, Iwaï H. 2020. NMR structure and dynamics of TonB investigated by scar-less segmental isotopic labeling using a salt-inducible split intein. Frontiers in chemistry 8:136.
Cook DR, Rossman KL, Der CJ. 2014. Rho guanine nucleotide exchange factors: regulators of Rho GTPase activity in development and disease. Oncogene 33:4021-4035.
Cox AD, Fesik SW, Kimmelman AC, Luo J, Der CJ. 2014. Drugging the undruggable RAS: mission possible? Nature reviews Drug discovery 13:828-851.
Czerminski R, Elber R. 1990. Self-avoiding walk between two fixed points as a tool to calculate reaction paths in large molecular systems. International Journal of Quantum Chemistry 38:167-185.
Dagher R, Brière C, Fève M, Zeniou M, Pigault C, Mazars C, Chneiweiss H, Ranjeva R, Kilhoffer M-C, Haiech J. 2009. Calcium fingerprints induced by Calmodulin interactors in eukaryotic cells. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research 1793:1068-1077.
Darden T, Perera L, Li L, Pedersen L. 1999. New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. Structure 7:R55-R60.
Darve E, Rodríguez-Gómez D, Pohorille A. 2008. Adaptive biasing force method for scalar and vector free energy calculations. The Journal of chemical physics 128:144120.
David A. 2007. Vassilvitskii S.: K-means++: The Advantages of Careful Seedingeditor^editors. 18th annual ACM-SIAM symposium on Discrete algorithms (SODA), New Orleans, Louisiana. p 1027-1035.
Dutta A, Krieger J, Lee JY, Garcia-Nafria J, Greger IH, Bahar I. 2015. Cooperative dynamics of intact AMPA and NMDA glutamate receptors: similarities and subfamily-specific differences. Structure 23:1692-1704.
Echols N, Milburn D, Gerstein M. 2003. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Research 31:478-482.
Fallon JL, Quiocho FA. 2003. A closed compact structure of native Ca2+-calmodulin. Structure 11:1303-1307.
Faradjian AK, Elber R. 2004. Computing time scales from reaction coordinates by milestoning. The Journal of chemical physics 120:10880-10889.
Frueh DP, Goodrich AC, Mishra SH, Nichols SR. 2013. NMR methods for structural studies of large monomeric and multimeric proteins. Current opinion in structural biology 23:734-739.
Fuchigami S, Omori S, Ikeguchi M, Kidera A. 2010. Normal mode analysis of protein dynamics in a non-Eckart frame. The Journal of chemical physics 132:104109.
Fujisaki H, Shiga M, Moritsugu K, Kidera A. 2013. Multiscale enhanced path sampling based on the Onsager-Machlup action: Application to a model polymer. The Journal of chemical physics 139:08B607_1.
Gedeon PC, Thomas JR, Madura JD. 2015. Accelerated molecular dynamics and protein conformational change: a theoretical and practical guide using a membrane embedded model neurotransmitter transportereditor^editors. Molecular Modeling of Proteins. Springer, p 253-287.

Gerek ZN, Ozkan SB. 2011. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. PLoS computational biology 7:e1002154.
Grant BJ, Gorfe AA, McCammon JA. 2010. Large conformational changes in proteins: signaling and other functions. Current opinion in structural biology 20:142-147.
Gsponer J, Christodoulou J, Cavalli A, Bui JM, Richter B, Dobson CM, Vendruscolo M. 2008. A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction. Structure 16:736-746.
Guven G, Atilgan AR, Atilgan C. 2014. Protonation States of Remote Residues Affect Binding–Release Dynamics of the Ligand but Not the Conformation of Apo Ferric Binding Protein. The Journal of Physical Chemistry B 118:11677-11687.
Hamelberg D, Mongan J, McCammon JA. 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. The Journal of chemical physics 120:11919-11929.
Harmat V, Böcskei Z, Náray-Szabó G, Bata I, Csutor AS, Hermecz I, Arányi P, Szabó B, Liliom K, Vértessy BG. 2000. A new potent calmodulin antagonist with arylalkylamine structure: crystallographic, spectroscopic and functional studies. Journal of molecular biology 297:747-755.
Haspel N, Moll M, Baker ML, Chiu W, Kavraki LE. 2010. Tracing conformational changes in proteins. BMC structural biology 10:S1.
Henzler-Wildman KA, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson MA, Petsko GA, Karplus M. 2007. Intrinsic motions along an enzymatic reaction trajectory. Nature 450:838.
Hickman SJ, Cooper RE, Bellucci L, Paci E, Brockwell DJ. 2017. Gating of TonB-dependent transporters by substrate-specific forced remodelling. Nature communications 8:1-12.
Hollenstein K, Frei DC, Locher KP. 2007. Structure of an ABC transporter in complex with its binding protein. Nature 446:213-216.
Houdusse A, Love ML, Dominguez R, Grabarek Z, Cohen C. 1997. Structures of four Ca2+-bound troponin C at 2.0 Å resolution: further insights into the Ca2+-switch in the calmodulin superfamily. Structure 5:1695-1711.
Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. Journal of molecular graphics 14:33-38.
Ikeguchi M, Ueno J, Sato M, Kidera A. 2005. Protein structural change upon ligand binding: linear response theory. Physical review letters 94:078102.
Ikura M, Clore GM, Gronenborn AM, Zhu G, Klee CB, Bax A. 1992. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. Science 256:632-638.
Ilter M, Sensoy O. 2019. Catalytically Competent Non-transforming H-RAS G12P Mutant Provides Insight into Molecular Switch Function and GAP-independent GTPase Activity of RAS. Scientific reports 9:1-10.
Isralewitz B, Gao M, Schulten K. 2001. Steered molecular dynamics and mechanical functions of proteins. Current opinion in structural biology 11:224-230.
Jalalypour F, Sensoy O, Atilgan C. 2020. Perturb–Scan–Pull: A Novel Method Facilitating Conformational Transitions in Proteins. Journal of Chemical Theory and Computation 16:3825-3841.
Johnson CK. 2006. Calmodulin, conformational states, and calcium signaling. A single-molecule perspective. Biochemistry 45:14233-14246.

Johnson CW, Reid D, Parker JA, Salter S, Knihtila R, Kuzmic P, Mattos C. 2017. The small GTPases K-Ras, N-Ras, and H-Ras have distinct biochemical properties determined by allosteric effects. Journal of Biological Chemistry 292:12981-12993.
Karplus M, Kuriyan J. 2005. Molecular dynamics and protein function. Proceedings of the National Academy of Sciences 102:6679-6685.
Khambati HK, Moraes TF, Singh J, Shouldice SR, Yu R-h, Schryvers AB. 2010. The role of vicinal tyrosine residues in the function of Haemophilus influenzae ferric-binding protein A. Biochemical Journal 432:57-67.
Khan AG, Shouldice SR, Kirby SD, Yu R-h, Tari LW, Schryvers AB. 2007a. High-affinity binding by the periplasmic iron-binding protein from Haemophilus influenzae is required for acquiring iron from transferrin. Biochemical Journal 404:217-225 %@ 0264-6021.
Khan AG, Shouldice SR, Tari LW, Schryvers AB. 2007b. The role of the synergistic phosphate anion in iron transport by the periplasmic iron-binding protein from Haemophilus influenzae. Biochemical Journal 403:43-48.
Komeiji Y, Ueno Y, Uebayasi M. 2002. Molecular dynamics simulations revealed Ca2+-dependent conformational change of Calmodulin. FEBS letters 521:133-139.
Krewulak KD, Vogel HJ. 2008. Structural biology of bacterial iron uptake. Biochimica et Biophysica Acta (BBA)-Biomembranes 1778:1781-1804.
Leone V, Marinelli F, Carloni P, Parrinello M. 2010. Targeting biomolecular flexibility with metadynamics. Current opinion in structural biology 20:148-154.
Li D, Liu MS, Ji B. 2015. Mapping the dynamics landscape of conformational transitions in enzyme: the adenylate kinase case. Biophysical journal 109:647-660.
Liu F, Chu X, Lu HP, Wang J. 2017. Molecular mechanism of multispecific recognition of Calmodulin through conformational changes. Proceedings of the National Academy of Sciences 114:E3927-E3934.
Liu X, Wang X, Jiang H. 2008. A steered molecular dynamics method with direction optimization and its applications on ligand molecule dissociation. Journal of biochemical and biophysical methods 70:857-864.
Lloyd S. 1982. Least squares quantization in PCM. IEEE transactions on information theory 28:129-137.
Locher KP. 2016. Mechanistic diversity in ATP-binding cassette (ABC) transporters. Nature structural & molecular biology 23:487.
Mallamace F, Corsaro C, Mallamace D, Vasi S, Vasi C, Baglioni P, Buldyrev SV, Chen S-H, Stanley HE. 2016. Energy landscape in protein folding and unfolding. Proceedings of the National Academy of Sciences 113:3159-3163.
Marsh JA, Teichmann SA. 2014. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. BioEssays 36:209-218.
Martin HS, Jha S, Howorka S, Coveney PV. 2009. Determination of free energy profiles for the translocation of polynucleotides through α-hemolysin nanopores using non-equilibrium molecular dynamics simulations. Journal of chemical theory and computation 5:2135-2148.
Matsunaga Y, Sugita Y. 2018. Refining Markov state models for conformational dynamics using ensemble-averaged data and time-series trajectories. The Journal of chemical physics 148:241731.
McCarty J, Parrinello M. 2017. A variational conformational dynamics approach to the selection of collective variables in metadynamics. The Journal of chemical physics 147:204109.
McCormick F. 1995. Ras-related proteins in signal transduction and growth control. Molecular reproduction and development 42:500-506.

Meador WE, Means AR, Quiocho FA. 1992. Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. Science 257:1251-1255.
Medvedeva MV, Djemuchadze DR, Watterson DM, Marston SB, Gusev NB. 2001. Replacement of Lys-75 of calmodulin affects its interaction with smooth muscle caldesmon. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology 1544:143-150.
Medvedeva MV, Polyakova OV, Watterson DM, Gusev NB. 1999. Mutation of Lys-75 affects calmodulin conformation. FEBS letters 450:139-143.
Mendez R, Bastolla U. 2010. Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. Physical review letters 104:228103.
Müller C, Schlauderer G, Reinstein J, Schulz GE. 1996. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 4:147-156.
Müller CW, Schulz GE. 1992. Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 Å resolution: A model for a catalytic transition state. Journal of molecular biology 224:159-177.
Neumann W, Hadley RC, Nolan EM. 2017. Transition metals at the host–pathogen interface: How Neisseria exploit human metalloproteins for acquiring iron and zinc. Essays in biochemistry 61:211-223.
Nevin Gerek Z, Kumar S, Banu Ozkan S. 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. Evolutionary applications 6:423-433.
Noinaj N, Buchanan SK, Cornelissen CN. 2012. The transferrin–iron import system from pathogenic N eisseria species. Molecular microbiology 86:246-257.
Nussinov R. 2016. Introduction to protein ensembles and allosteryeditor^editors: ACS Publications.
Osawa M, Swindells MB, Tanikawa J, Tanaka T, Mase T, Furuya T, Ikura M. 1998. Solution structure of calmodulin-W-7 complex: the basis of diversity in molecular recognition. Journal of molecular biology 276:165-176.
Pai EF, Krengel U, Petsko GA, Goody RS, Kabsch W, Wittinghofer A. 1990. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 A resolution: implications for the mechanism of GTP hydrolysis. The EMBO journal 9:2351-2359.
Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R. 2016. The role of protein loops and linkers in conformational dynamics and allostery. Chemical reviews 116:6391-6423.
Park S, Schulten K. 2004. Calculating potentials of mean force from steered molecular dynamics simulations. The Journal of chemical physics 120:5946-5961.
Parker JA, Mattos C. 2015. The Ras–membrane interface: isoform-specific differences in the catalytic domain. Molecular cancer research 13:595-603.
Patel D, Kuyucak S. 2017. Computational study of aggregation mechanism in human lysozyme [D67H]. PloS one 12:e0176886.
Patrick JW, Boone CD, Liu W, Conover GM, Liu Y, Cong X, Laganowsky A. 2018. Allostery revealed within lipid binding events to membrane proteins. Proceedings of the National Academy of Sciences 115:2976-2981.
Penkler D, Sensoy Oz, Atilgan C, Tastan Bishop Oz. 2017. Perturbation–Response Scanning Reveals Key Residues for Allosteric Control in Hsp70. Journal of chemical information and modeling 57:1359-1374.
Petrone P, Pande VS. 2006. Can conformational change be described by only a few normal modes? Biophysical journal 90:1583-1593.

Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. 2005. Scalable molecular dynamics with NAMD. Journal of computational chemistry 26:1781-1802.
Pierce LC, Salomon-Ferrer R, Augusto F. de Oliveira C, McCammon JA, Walker RC. 2012. Routine access to millisecond time scale events with accelerated molecular dynamics. Journal of chemical theory and computation 8:2997-3002.
Pontiggia F, Zen A, Micheletti C. 2008. Small-and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics. Biophysical journal 95:5901-5912.
Ponzoni L, Bahar I. 2018. Structural dynamics is a determinant of the functional significance of missense variants. Proceedings of the National Academy of Sciences 115:4164-4169.
Rees DC, Johnson E, Lewinson O. 2009. ABC transporters: the power to change. Nature reviews Molecular cell biology 10:218-227.
Ribeiro JML, Bravo P, Wang Y, Tiwary P. 2018. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). The Journal of chemical physics 149:072301.
Rodríguez-Castañeda F, Maestre-Martínez M, Coudevylle N, Dimova K, Junge H, Lipstein N, Lee D, Becker S, Brose N, Jahn O. 2010. Modular architecture of Munc13/calmodulin complexes: dual regulation by Ca2+ and possible function in short-term synaptic plasticity. The EMBO Journal 29:680-691.
Ross C, Nizami B, Glenister M, Sheik Amamuddy O, Atilgan AR, Atilgan C, Tastan Bishop Ö. 2018. MODE-TASK: large-scale protein motion tools. Bioinformatics 34:3759-3763.
Rostkowski M, Olsson MH, Søndergaard CR, Jensen JH. 2011. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. BMC structural biology 11:1-6.
Sankar K, Liu J, Wang Y, Jernigan RL. 2015. Distributions of experimental protein structures on coarse-grained free energy landscapes. The Journal of chemical physics 143:12B635_1.
Sensoy O, Atilgan AR, Atilgan C. 2017. FbpA iron storage and release are governed by periplasmic microenvironments. Physical Chemistry Chemical Physics 19:6064-6075.
Seyler SL, Beckstein O. 2014. Sampling large conformational transitions: adenylate kinase as a testing ground. Molecular Simulation 40:855-877.
Shao Q. 2016. Enhanced conformational sampling technique provides an energy landscape view of large-scale protein conformational transitions. Physical Chemistry Chemical Physics 18:29170-29182.
Sherman HG, Jovanovic C, Stolnik S, Baronian K, Downard AJ, Rawson FJ. 2018. New perspectives on iron uptake in eukaryotes. Frontiers in Molecular Biosciences 5:97.
Shouldice SR, Skene RJ, Dougan DR, McRee DE, Tari LW, Schryvers AB. 2003. Presence of ferric hydroxide clusters in mutants of Haemophilus influenzae ferric ion-binding protein A. Biochemistry 42:11908-11914.
Shouldice SR, Skene RJ, Dougan DR, Snell G, McRee DE, Schryvers AB, Tari LW. 2004. Structural basis for iron binding and release by a novel class of periplasmic iron-binding proteins found in gram-negative pathogens. Journal of bacteriology 186:3903-3910 %@ 0021-9193.
Slaughter BD, Unruh JR, Allen MW, Bieber Urbauer RJ, Johnson CK. 2005. Conformational substates of calmodulin revealed by single-pair fluorescence resonance energy transfer: influence of solution conditions and oxidative modification. Biochemistry 44:3694-3707.

Srivastava A, Nagai T, Srivastava A, Miyashita O, Tama F. 2018. Role of computational methods in going beyond X-ray crystallography to explore protein structure and dynamics. International journal of molecular sciences 19:3401.
Stetz G, Tse A, Verkhivker GM. 2017. Ensemble-based modeling and rigidity decomposition of allosteric interaction networks and communication pathways in cyclin-dependent kinases: Differentiating kinase clients of the Hsp90-Cdc37 chaperone. PloS one 12:e0186089.
Stetz G, Tse A, Verkhivker GM. 2018. Dissecting structure-encoded determinants of allosteric cross-talk between post-translational modification sites in the Hsp90 chaperones. Scientific reports 8:6899.
Szewczyk J, Collet J-F. 2016. The journey of lipoproteins through the cell: one birthplace, multiple destinationseditor^editors. Advances in Microbial Physiology. Elsevier, p 1-50.
Tang CM, Hood DW, Moxon ER. 2001. CHAPTER 14 - Pathogenesis of Haemophilus influenzae Infections. In Groisman EA, editor^editors. Principles of Bacterial Pathogenesis. San Diego: Academic Press, p 675-716.
ten Klooster JP, Hordijk PL. 2007. Targeting and localized signalling by small GTPases. Biology of the Cell 99:1-12.
Tiwary P, Berne B. 2016. Spectral gap optimization of order parameters for sampling complex molecular systems. Proceedings of the National Academy of Sciences 113:2839-2844.
Vandonselaar M, Hickie RA, Quail W, Delbaere LT. 1994. Trifluoperazine-induced conformational change in Ca 2+-calmodulin. Nature structural biology 1:795.
Verkhivker GM. 2019. Biophysical simulations and structure-based modeling of residue interaction networks in the tumor suppressor proteins reveal functional role of cancer mutation hotspots in molecular communication. Biochimica et Biophysica Acta (BBA)-General Subjects 1863:210-225.
Verkhivker GM. 2020. Molecular Simulations and Network Modeling Reveal an Allosteric Signaling in the SARS-CoV-2 Spike Proteins. Journal of proteome research 19:4587-4608.
Vetter SW, Leclerc E. 2003. Novel aspects of calmodulin target recognition and activation. European Journal of Biochemistry 270:404-414.
Vuong QV, Nguyen TT, Li MS. 2015. A new method for navigating optimal direction for pulling ligand from binding pocket: application to ranking binding affinity by steered molecular dynamics. Journal of chemical information and modeling 55:2731-2738.
Wang A, Zhang D, Li Y, Zhang Z, Li G. 2019. Large-scaled Biomolecular Conformational Transitions Explored by Combined Elastic Network Model and Enhanced Sampling Molecular Dynamics. The Journal of Physical Chemistry Letters.
Wang J, Peng C, Yu Y, Chen Z, Xu Z, Cai T, Shao Q, Shi J, Zhu W. 2020. Exploring Conformational Change of Adenylate Kinase by Replica Exchange Molecular Dynamic Simulation. Biophysical Journal.
Wang Y, Papaleo E, Lindorff-Larsen K. 2016. Mapping transiently formed and sparsely populated conformations on a complex energy landscape. Elife 5:e17505.
Weinberg ED. 2003. The therapeutic potential of lactoferrin. Expert opinion on investigational drugs 12:841-851.
Wu H, Post CB. 2018. Protein Conformational Transitions from All-Atom Adaptively Biased Path Optimization. Journal of chemical theory and computation 14:5372-5382.
Wyckoff EE, Mey AR, Leimbach A, Fisher CF, Payne SM. 2006. Characterization of ferric and ferrous iron transport systems in Vibrio cholerae. Journal of Bacteriology 188:6515-6523.

Yang L, Song G, Jernigan RL. 2007. How well can we understand large-scale protein motions using normal modes of elastic network models? Biophysical Journal 93:920-929.
Yang YI, Shao Q, Zhang J, Yang L, Gao YQ. 2019. Enhanced sampling in molecular dynamics. The Journal of chemical physics 151:070902.
You W, Tang Z, Chang C-eA. 2019. Potential mean force from umbrella sampling simulations: what can we learn and what is missed? Journal of chemical theory and computation 15:2433-2443.
Yun C-H, Bai J, Sun D-Y, Cui D-F, Chang W-R, Liang D-C. 2004. Structure of potato calmodulin PCM6: the first report of the three-dimensional structure of a plant calmodulin. Acta Crystallographica Section D: Biological Crystallography 60:1214-1219.
Zhang Y, Lou J. 2012. The Ca2+ influence on calmodulin unfolding pathway: a steered molecular dynamics simulation study. PloS one 7:e49013.
Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. 2014. Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme. Molecular biology and evolution 32:132-143.