# INVESTIGATING DRUG-MEDIATED CONFORMATIONAL CHANGES IN KRAS

*A thesis submitted in partial fulfillment of the requirements for the award of the dual degree of*

**Bachelor of Science – Master of Science**

in

**Biological Sciences**

by


**SASWAT KUMAR MOHANTY**

**17MS204**


Under the supervision of

**DR. SUSMITA ROY**

&

co-supervision of

**DR. PURBA MUKHERJEE**

to the

**DEPARTMENT OF BIOLOGICAL SCIENCES**



**INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH**

**KOLKATA**


**MAY, 2022**

# INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH, KOLKATA

## Certificate

I hereby certify that the matter embodied in the thesis entitled, "**Investigating Drug-Mediated Conformational Changes in KRas**", is the result of investigations carried out by Saswat Kumar Mohanty at the Department of Biological Sciences and Department of Chemical Sciences, Indian Institute of Science Education and Research, Kolkata under my supervision and has not been submitted elsewhere for the award of any degree, diploma or other qualification.

Supervisor Signature

Date: *24 May, 2022*

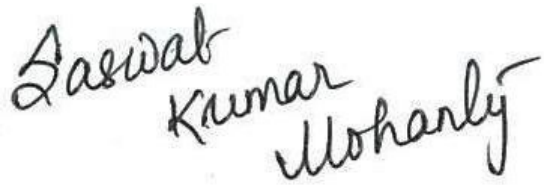
Place: *Mohanpur, Nadia, WB*

Co-Supervisor Signature

# Declaration

I hereby declare that this thesis entitled "**Investigating Drug-Mediated Conformational Changes in KRas**" is my own work, which has been carried out under the supervision of Dr. Susmita Roy and the co-supervision of Dr. Purba Mukherjee. To the best of my knowledge, it contains no materials which are previously published or written by any other person, or substantial proportions of material which have been accepted for the award of any other degree at IISER Kolkata or any other educational institution, except where due Acknowledgements are made in the thesis.

I certify that all copyrighted materials incorporated into this thesis complies with the Indian Copyright (Amendment) Act 2012 and I agree to indemnify and save IISER Kolkata from any and all claims that may be asserted, or that may arise from any copyright violation.



Date: *19 May, 2022* Saswat Kumar Mohanty

Place: *Mohanpur, Nadia, WB*

# Acknowledgement

First and foremost, I'd like to express my gratitude to my advisor, Dr. Susmita Roy, for piquing my interest in computational biophysics and her invaluable guidance, never-ending support, and encouragement throughout the research process. Being her Master's student has been a privilege. She introduced me to international collaborations so early in my career, which has been a lifetime experience. I have learned a lot from her over the last 2.5 years, including how to approach a problem from different perspectives, turn a problem into a new possibility, and, most importantly, look at failures positively. Therefore, I owe her a debt of gratitude for being an outstanding mentor.

I'd also like to express my heartfelt gratitude to all of the faculty members at IISER Kolkata, who have instilled in me the knowledge and enthusiasm to carry out this research work.

I want to take this opportunity to thank my parents, Anagendra Nath Mohanty and Binata Mohanty, as well as my sister, Swagatika Mohanty, for providing me with moral support and motivation throughout my time at IISER Kolkata. I also owe my deepest gratitude to my friends' group "Crazzzy Confused" for their constant support throughout this journey of ups and downs at IISER Kolkata.

I would also take the opportunity to thank my co-supervisor, Dr. Purba Mukherjee, who agreed to co-supervise my thesis. She has helped me with all the departmental logistics and has extended her constant support whenever I have needed it. Nevertheless, I'd also like to thank Raju Sarkar, Satyam Sangeet, Anushree Sinha, Avijit Mainan, and my other lab mates, who have been extremely helpful in providing me with their unwavering support and helping me grow. I am also grateful to the DIRAC supercomputing facility at IISER Kolkata and the supercomputing facility at IACS Kolkata, without which I would not have been able to carry out my research work.

# Abstract

Intrinsically Disorder Proteins (IDPs) are a significant part of the human proteome, and their involvement in numerous diseases is well documented. As IDPs have no single fixed structure, they represent an exception to the structured-protein concept, known as Anfinsen's dogma. On the other hand, there are Intrinsically Disorder Regions (IDRs) in some protein structures that can be fully or partially disordered, containing highly charged amino acid residues. Despite their unstructured regions, they are involved in critical roles in cellular functioning.

KRas, a member of the Ras GTPase family, is one of such proteins containing a number of IDRs known as switch regions. Mutations in wild-type KRas at the G12 position cause loss of GTPase activity and acquire oncogenic properties that result in tumour cell growth and cancer progression. Recently, AMG510 was one of the first KRas (G12C) inhibitors efficacious against KRas G12C tumors. However, a recent FDA-approved drug MRTX849 is more efficacious than AMG510 in tumour regression in KRas G12C mutant cell lines of multiple tumour types, especially patients with lung and colon cancer patients.

As acquired resistance to the mutant selective KRas G12C inhibitor like AMG510 is a major concern in lung cancer, to understand different drug-induced structural changes of KRas, this thesis work attempts to perform computational studies on the G12C mutated KRas, as well as the above two drug bound forms: AMG-510 and MRTX-849. This thesis contains four chapters as follows: Chapter-I contains introduction to IDPs/IDRs, the structural plasticity and the experimental and computational methods to characterise their properties. We refresh through the computational methods, in chapter-II, which we have used to run our molecular dynamics simulations and which we have used to extract meaningful data from our trajectories. Through our analysis of fluctuation, contact and correlation map, in chapter-III, we find that MRTX is potent in inhibiting the fluctuation of the IDR switches as compared to AMG. MRTX forms a large number of hydrogen and a few hydrophobic interactions with the Switch-II loop. In chapter-IV, our thorough free energy analyses and comparison of drug-bound and unbound forms of KRas explore all possible drug-induced conformational states. This exploration indicates that MRTX is likely to restrict the GDP-GTP exchange in its functional cycle, and hence, possible this is one of the reasons that may exert high efficacy. This study also predicts that switch-II inhibition can act as a potent target for any future drug to make KRas-G12C inhibition operational and devoid of acquired drug-resistance.

# Contents

**Acknowledgement** iv

**Abstract** v

**List of Abbreviations** xi

**List of Figures** xiii

**1 Introduction** 1

1.1 Order-Disorder Transition: Limit of Anfinsen's Dogma . . . . . 1

1.1.1 A Continuum rather than Binary . . . . . 1

1.1.2 Structural plasticity and implications . . . . . 2

1.1.3 Interactions of IDPs and mechanism . . . . . 3

1.2 Experimental characterization of IDPs . . . . . 4

1.3 Computational aspects of IDPs . . . . . 5

1.3.1 Molecular Dynamics Simulations . . . . . 5

1.3.2 Energy Landscape Visualization Method (ELViM) . . . . 6

1.3.3 Parallel Tempering . . . . . 6

1.4 IDP Examples . . . . . 8

1.4.1 Prostate-Associated Gene 4 (PAGE4) . . . . . 8

1.4.2 Ras-family proteins . . . . . 9

**2 Concepts of Computational Methods and Techniques** 11

2.1 Basics of Statistical Mechanics . . . . . 11

2.1.1 Phase Space Trajectory . . . . . 11

2.1.2 Ensembles . . . . . 12

2.1.3 The First Postulate: Time average is equal to Ensemble average . . . . . 14
2.1.4 The Second Postulate: Equal A Priori Probability . . . . 14
2.1.5 The Ergodic Hypothesis . . . . . 15
2.1.6 Canonical Partition Function . . . . . 15
2.1.7 Relating Thermodynamics and Partition Function . . . . 17
2.2 Basic Concepts of Molecular Dynamics Simulation . . . . . 19
2.2.1 General Features of Force Field . . . . . 19
2.2.1.1 Bonded Potential . . . . . 20
2.2.1.2 Angular Potential . . . . . 20
2.2.1.3 Torsional Potential . . . . . 21
2.2.1.4 Non-bonded Potential . . . . . 23
2.2.2 Energy Minimisation . . . . . 24
2.2.2.1 Steepest Descent . . . . . 25
2.2.2.2 Conjugate Gradient . . . . . 26
2.2.3 Basic Approach . . . . . 26
2.2.4 Numerical Integration Methods . . . . . 27
2.2.4.1 Verlet Algorithm . . . . . 27
2.2.4.2 Leap-Frog Algorithm . . . . . 28
2.2.4.3 Velocity-Verlet Algorithm . . . . . 28
2.3 Temperature and Pressure Control . . . . . 29
2.3.1 Temperature Coupling . . . . . 29
2.3.2 Pressure Coupling . . . . . 30
2.4 Tricks for Computational Efficiency . . . . . 31
2.4.1 Periodic Boundary Conditions . . . . . 31

2.4.2 Minimum Image Convention and Truncation of Intermolecular Interaction . . . . . 32
2.4.3 Long Range Forces: Ewald Summation and Particle Mesh Ewald . . . . . 33
2.4.4 Neighbour Lists and Cell Lists . . . . . 34
2.4.5 Free Energy Calculations: Umbrella Sampling . . . . . 34
2.4.6 Free Energy Calculations: Metadynamics . . . . . 37
2.4.6.1 Standard Metadynamics . . . . . 39
2.4.6.2 Well-tempered Metadynamics . . . . . 40
2.5 Computational Methods for Analysis . . . . . 41
2.5.1 Root-Mean Square Distance (RMSD) Analysis . . . . . 41
2.5.2 Root-Mean Square Fluctuation (RMSF) Analysis . . . . 41
2.5.3 Theory of Correlation Analysis . . . . . 42
2.5.3.1 Gaussian Network Model . . . . . 42
2.5.3.2 Covariance Matrix . . . . . 43
2.5.3.3 Correlation Matrix . . . . . 44

**3 Drug-induced conformational dynamics of oncogenic KRas: Comparing the effects of AMG-510 & MRTX-849 45**
3.1 Introduction . . . . . 45
3.1.1 KRas: A MAPK signalling protein & its cellular functioning 45
3.1.2 KRas in Lung Cancer . . . . . 47
3.1.3 Most fatal G12C mutation and its drug induced inhibition 48
3.2 Methodology: Details of atomistic simulation methods and IDP specific force fields . . . . . 49
3.2.1 System Preparation . . . . . 49
3.2.1.1 Mutated protein in no-drug state . . . . . 49

3.2.1.2 Mutated protein in drug-bound state: Specific interaction with AMG-510 & MRTX-849 . . . . . 49
3.2.2 Hybrid protein specific force field: CHARMM36IDPSFF 50
3.2.3 Atomistic simulation methods . . . . . 52
3.3 Results & Analysis: Comparison of conformational dynamics among the G12C variants, and the AMG and MRTX drug-bound forms . . . . . 53
3.3.1 Finding fluctuating motifs from RMSF analysis . . . . . 53
3.3.2 Quantifying & comparing fluctuation of different IDRs in KRas from RMSD . . . . . 54
3.3.3 Fluctuation-fluctuation correlation at residual level between Switch-I & Switch-II . . . . . 55
3.3.4 Structural investigation in the neighbourhood of switch regions–specifically focussing on $\alpha$-2 & $\alpha$-3 . . . . . 57
3.3.4.1 Temporal Helicity comparison . . . . . 57
3.3.4.2 Dihedral Analysis . . . . . 58
3.3.5 Exploration of drug-mediated interaction through contact map analysis . . . . . 59
3.4 Conclusion . . . . . 61

**4 Exploring conformational landscape of drug-bound and unbound forms of KRAS: Deducing switch-mediated kick-out mechanism 62**
4.1 Introduction . . . . . 62
4.2 Methodology: Well-tempered metadynamics simulation . . . . 64
4.3 Results . . . . . 65
4.3.1 Conformational states of Wild-type (WT) and G12C oncogenic variant of KRas . . . . . 65

4.3.2 Comparison of conformational states of the oncogenic variant & the drug-bound forms of KRas . . . . . 66
4.4 Conclusion . . . . . 68

**Future Aspects** **69**

**References** **70**

# List of Abbreviations

**MD** *Molecular Dynamics*

**IDP** *Intrinsically Disordered Protein*

**IDR** *Intrinsically Disordered Region*

**PIN** *Protein Interaction Network*

**MORF** *MOlecular Recognition Feature*

**SLiM** *Short-LInear Motif*

**EM** *Electron Microscopy*

**SAXS** *Small-Angle X-ray Scattering*

**DLS** *Dynamic Light Scattering*

**AFM** *Atomic Force Microscopy*

**CD** *Circular Dichorism*

**smFRET** *single molecular Förster Resonance Energy Transfer*

**2f-FCS** *Two-focus Fluorescence Correlation Spectroscopy*

**MS** *Mass Spectrometry*

**NMR** *Nuclear Magnetic Resonance*

**FTIR** *Fourier Transform Infrared Spectroscopy*

**PAGE4** *Prostate-Associated GEne 4*

**PCa** *Prostate Cancer*

**ELViM** *Energy Landscape Visualization Method*

**NVE** *Number of particles, Volume, Energy*

**NVT** *Number of particles, Volume, Temperature*

**NPT** *Number of particles, Pressure, Temperature*

**GROMACS** *GROningen MAchine for Chemical Simulations*

**MIC** *Minimum Image Convention*

**TIMI** *Truncation of InterMolecular Interaction*

**PBC** *Periodic Boundary Condition*

**CV** *Collective Variable*

**MAPK** *Mitogen Activated protein Kinase*

**ERK** *Extracellular Signal-Regulated Kinase*

**DNA** *Deoxyribose Nucleic Acid*

**KRas** *Kirsten Rat sarcoma*

**SOS** *Son of Sevenless*

**GEF** *Guanine Exchange Factor*

**GDP** *Guanosine DiPhosphate*

**GTP** *Guanosine TriPhosphate*

**GAP** *GTPase Activating Protein*

**P-loop** *Phosphate-binding loop*

**HVR** *Hyper-Variable Region*

**NSCLC** *Non-Small-Cell Lung Cancer*

**AMG** *AMGen*

**MRTX** *MiRati Therapeutics*

**FDA** *Food and Drug Administration*

**CHARMM36IDPSFF** *Chemistry at Harvard Macromolecular Mechanics-36 Intrinsically Disordered Protein Specific Force Field*

**CMAP** *Correction MAP*

**LINCS** *Linear Constraint Solver*

**GNM** *Gaussian Network Model*

**RMSD** *Root-Mean Square Distance*

**RMSF** *Root-Mean Square Fluctuation*

# List of Figures


1.1 Order-Disorder Transition; adapted from Ref. [10] . . . . . 2

1.2 ELViM representation of different phosphorylated states of PAGE4; adapted from Ref. [77] . . . . . 9

2.1 Schematic representation of the trajectory of a system of N particles 11

2.2 Schematic representation of microcanonical ensemble . . . . . 13

2.3 Schematic representation of macrocanonical ensemble . . . . . 13

2.4 Schematic representation of grandcanonical ensemble . . . . . 14

2.5 Atoms i and j connected by spring with force constant $k_{ij}$; adapted from Ref. [87] . . . . . 20

2.6 Atoms i, j and k making an angle $\theta_{ijk}$; adapted from Ref. [87] . 20

2.7 Atoms i, j, k and l making a proper dihedral $\theta_{ijkl}$; adapted from Ref. [87] . . . . . 21

2.8 Atoms i, j, k and l making a improper dihedral $\theta_{ijkl}$; adapted from Ref. [87] . . . . . 22

2.9 Charged atoms i and j separated by distance $r_{ij}$; adapted from Ref. [87] . . . . . 23

2.10 Atoms i and j separated by distance $r_{ij}$; adapted from Ref. [87] . 23

2.11 Energy Minimisation Scheme; adapted from Ref. [88] . . . . . 24

2.12 Periodic Boundary Conditions; adapted from Ref. [98] . . . . . 32

2.13 High Activation Energy Barrier separating State I and State II; adapted from Ref. [106] . . . . . 35

2.14 Reaction coordinates between two states divided into into distinct windows; adapted from Ref. [106] . . . . . 37

2.15 Well-tempered metadynamics simulation showing decreasing Gaussian height with time; adapted from Ref. [110] . . . . . 40

2.16 Nodes in the GNM model connected with springs; adapted from Ref. [111] . . . . . 42
3.1 Schematic of MAPK/ERK cellular signalling pathway; adapted from Ref. [112] . . . . . 45
3.2 KRas function under physiological and mutated states; adapted from Ref. [114] . . . . . 46
3.3 KRas Structure and schematic representation of helices and sheets; adapted from Ref. [115] . . . . . 47
3.4 Pie chart showing mutational distribution for KRas malignancy in NSCLC . . . . . 48
3.5 Mutated KRas at position 12 from Glycine to Cysteine . . . . . 50
3.6 Mutated KRas covalently attached to AMG-510 at position 12 . 51
3.7 Mutated KRas covalently attached to MRTX-849 at position 12 51
3.8 RMSF plot of GDP bound G12C variant, AMG and MRTX drug-bound . . . . . 54
3.9 RMSD plots of GDP bound G12C variant, AMG and MRTX drug-bound . . . . . 55
3.10 Correlation plots of GDP bound G12C variant, AMG and MRTX drug-bound . . . . . 56
3.11 $\alpha$-2 helix melting histogram comparison . . . . . 57
3.12 Frequency-dependent histograms of GDP bound G12C variant, AMG and MRTX drug-bound . . . . . 59
3.13 Contact map of Switch-II loop region with drugs . . . . . 60
3.14 Contact map of Switch-II's $\alpha$-2 region with drugs . . . . . 60
3.15 Contact map of $\alpha$-3 helix with drugs . . . . . 60

4.1 Schematic diagram of KRas-GEF Interaction scheme showing Kick-Out; adapted from Ref. [80] . . . . . 63
4.2 Binding and Unbinding mechanism of KRas-GEF interaction (PDB ID: 7KFZ) . . . . . 63
4.3 Well-tempered Metadynamics plot of WT KRas along with its most stable state . . . . . 65
4.4 Well-tempered Metadynamics plot of G12C-mutated KRas along with its most stable state . . . . . 65
4.5 Well-tempered Metadynamics plot of AMG-bound mutated KRas along with its most stable states . . . . . 66
4.6 Well-tempered Metadynamics plot of MRTX-bound mutated KRas along with its most stable states . . . . . 67

# 1 Introduction

## 1.1 Order-Disorder Transition: Limit of Anfinsen's Dogma

### 1.1.1 A Continuum rather than Binary

According to Christian B. Anfinsen, for a small globular protein in the normal physiological environment, the 3D native structure of it is solely determined by the protein's amino acid sequence [1, 2]. This hypothesis of Anfinsen, also know as the thermodynamic hypothesis, a postulate in molecular biology is infamously known as the Anfinsen's Dogma. He stated that a protein folds if these three conditions are satisfied: Uniqueness, Stability, and Kinetic Accessibility. It basically meant that proteins once folded must acquire the lowest accessible energy state in the free energy basin and must be unique. This implied that proteins mostly have very stable atomic positions that fluctuate slightly due to low-amplitude thermal perturbations.

However, it was discovered that almost all living biome on the earth consists of a proteome that is significantly constituted by Intrinsically disordered proteins (IDPs). These proteins, contrary to the Anfinsen's dogma, lack well-defined 3D structures in the normal physiological environment and have hence been a topic for active research recently. Though significant progress has been made, but there are still many potholes needed to be filled [3, 4, 5]. Therefore, IDPs are able to sample a huge number of dissimilar conformational states during their biological lifetime, as compared to the proteins having fixed structures corresponding to one global minima [6]. This is possible because, IDPs have global energy surface with multiple shallow minima having low energy barriers which enables the protein to rapidly visit multiple energy states in its lifetime [7, 8]. Therefore, in ordered proteins, the folded and unfolded forms are treated

as binary states, whereas in IDPs it essentially dictates a continuum of states [9].

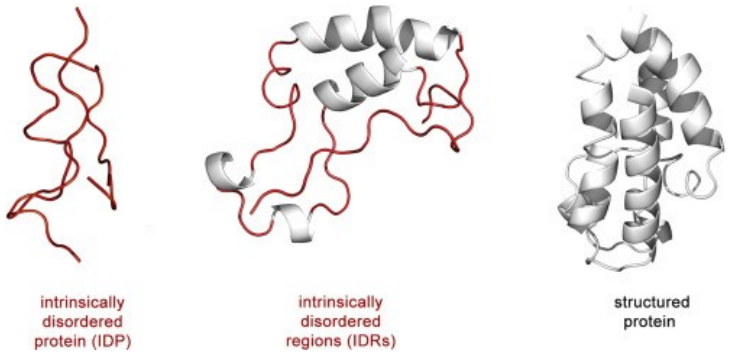



Figure 1.1: Order-Disorder Transition; adapted from Ref. [10]

A family of the IDPs constitutes of proteins where they show spatio-temporal heterogeneity in the form that, only certain parts of the proteins are disordered to a different degree. These fragments of disordered parts are known as foldons. They can be of many types such as, inducible foldons, morphing inducible foldons, semi-foldons and non-foldons [11, 12, 13, 14]. This spatio-temporal heterogeneity enables their multifunctional ability, making different parts of the protein to function differently under different conditions. Hence, instead of a classical "one gene–one protein–one structure–one function" model, the promiscuous IDPs/IDRs constitute a structure-function continuum [14, 15, 16].

### 1.1.2 Structural plasticity and implications

Because of the structural heterogeneity exhibited by the IDPs, they occupy the key nodal positions in the Protein Interaction Networks (PIN) [17, 18]. PINs being the channel system inside the cells, are essential for the coordinated functioning of the cell. However, due to the ability of IDPs to promiscuously interact, when overexpressed, IDPs can rewire and change the PINs adapting to the new environmental perturbations [19].

Being major hubs in the PINs, IDPs perform a multitude of functions such as

signaling via cellular protein networks, splicing, embryonic differentiation and development, and transcriptional regulation [6, 20]. The interactions exhibited by IDPs have high specificity with low affinity which lead to rapid and spontaneous dissociation and, hence, termination of the downstream signal, which allows high levels of cellular control [21, 22]. These let IDPs to function as sensitive rheostats and switches in the PIN regulatory circuits [22, 23].

Apart from these, IDPs are involved in major cellular events such as: regulation of cell cycle, phenotypic plasticity, stress response, and circadian rhythm [24, 25, 26, 27, 28, 29, 30]. Moreover, most IDPs can form protein-based memories that drive the development and inheritance of biological characteristics in a prion-like way [31]. IDPs are also involved in ensuring that other proteins fold properly. Therefore, several chaperones, heat-shock proteins (Hsp22 and $\alpha\beta$-e) and stress-response proteins are IDPs in nature [32].

Due to the vast repertoire of functions IDPs execute, any dysregulation can lead IDPs to cause pathological states [33]. Hence, in many diseases like cancer, neurodegenerative diseases, genetic diseases, diabetes, etc. IDPs are seen to be dysregulated [34, 35, 36].

#### 1.1.3 Interactions of IDPs and mechanism

It has been studied that several of IDPs undergo disorder to order transition upon binding. Once they bind to their cognate partners, they undergo the "coupled folding and binding" phenomenon [37]. For this to happen, two mechanisms must go simultaneously. The first one is the "induced fit" and the other one being "conformational selection" mechanism. The former one informs that IDPs fold after associating with the target, while the latter envisages all potential conformations of the ensemble pre-exist among which one is selected by the lig-

and [38]. However, both can co-exist suggesting that the binding mechanism of the IDPs are determined by their intrinsic secondary structure propensities [39]. Hence, the disorder to order transition in IDPs is referred to as "template folding", where the partner binding to the IDP dictates the route to the product formed, ensuring a cooperative binding [40]. However, for some of the IDPs it has been seen that they continue to stay disordered even when bound to their cognate partner. Such interactions have been described to be as "fuzzy complexes" [41]. IDPs mainly interact with their binding partners with the help of molecular recognition features (MORFs) and short-linear motifs (SLiMs), in addition to low complexity sequences [42, 43]. The interactions of IDPs through SLIMs and are particularly electrostatic in nature, either through a highly positively charges patch or a negatively charged ones [44, 45]. However, some hydrophobic regions are also found to be interacting [46].

### 1.2 Experimental characterization of IDPs

Experimentally characterizing the IDPs, especially the IDR regions in the large proteins and complexes still remains a major challenge. Since the well-known techniques like cryo-EM and X-Ray crystallography provide only static state images of proteins in the frozen and crystallized states, respectively, they are not adequate to study the vast ensemble of structural heterogeneity posed by the IDPs [47]. Therefore, the go-to techniques for the characterization of IDPs are as follows: small-angle X-ray scattering (SAXS), dynamic light scattering (DLS), atomic force microscopy (AFM), circular dichorism (CD), single molecular Förster resonance energy transfer (FRET), fluorescence, two-focus fluorescence correlation spectroscopy (2f-FCS), mass spectrometry (MS), nuclear magnetic resonance (NMR), Fourier transform infrared spectroscopy (FTIR)

and Raman spectroscopy [48, 49, 50, 51, 52, 53, 54, 55]. However, all these experiments are able to provide only limited resolution, structure, and dynamics. Here, the computational methods have come to the rescue by elucidating the conformational ensemble to a higher degree, along with validating the experimental results [56, 57]. Nevertheless, the visualisation of the whole complex energy landscape still remains a computational challenge.

## 1.3 Computational aspects of IDPs

Due to the challenges faced by the experimental techniques, computational methods such as explicit solvent, atomistic molecular dynamics simulation, coarse-grained molecular dynamics simulations, and enhanced sampling methods are extensively used to study IDPs. Recent advancements in molecular dynamics simulations and energy landscape visualization techniques have shed new light on conformational dynamics and their functional implications at the system level.

### 1.3.1 Molecular Dynamics Simulations

The aim of MD trajectory analysis is to capture the properties of a system as a function of few-dimensional reaction coordinates, such as dominant kinetics and structural features of transition state ensembles. Alternative strategies for inferring suitable reaction coordinates to describe the energy landscape exist beyond the straightforward structure-based coordinates, such as the fraction of native contacts and the root mean square distance (RMSD) from reference structures. Transition-path analysis, for example, can be used to determine the coordinates that best represent the underlying free-energy barrier[58]. Time-correlation analysis, on the other hand, allows for the classification of collec-

tive variables associated with the slowest motions[59]. These techniques have a common limitation in that they require a priori definition of coordinates, which can be computationally expensive.

Local minima can be addressed individually, and visualization of the distances between local minima in a hierarchical representation is also an appealing way to probe the energy landscape[60]. The methods described above are well suited to studying funnel-like landscapes with well-defined energy basins. IDPs, on the other hand, are far more difficult systems to study because of their high disorder, shallow energy minima, and lack of reference structures.

### 1.3.2 Energy Landscape Visualization Method (ELViM)

A multidimensional scaling (MDS) method, is used in a recent successful approach to investigating IDPs. The goal of MDS methods is to represent an ensemble of objects in a low-dimensional space for easier analysis, given an ensemble of objects in the original multidimensional phase space. The ELViM method is one of the most intimidating of them all[61, 62]. This method is based on pairwise distances between all structures in the ensemble and is reaction coordinate-free[63]. The energy landscape can be visually analysed using this method. Furthermore, multiple ensembles can be mapped into a single phase space, allowing comparison of ensembles studied under various physical and chemical conditions. This MDS strategy appears to give an accurate depiction of the IDP energy landscape.

### 1.3.3 Parallel Tempering

IDPs are found in shallow, rugged free energy landscapes with multiple conformational populations in dynamic equilibrium. As a result, using experimental techniques to structurally resolve them at high resolution is difficult. Molecular

simulation has recently been used in conjunction with low-resolution ensemble-averaged data to elucidate the structural and dynamical features of IDPs at higher resolution[64, 65, 66]. Despite many advances, extracting an IDP's experimentally consistent ensemble remains a difficult task. This is due in part to the presence of multiple conformational states in an ensemble, which makes experimental data noisy, sparse, and/or ambiguous. Molecular simulations, on the other hand, typically sample only a small portion of an IDP ensemble's phase space, despite the fact that the underlying free energy landscape is shallow. The presence of significant entropic barriers between different population clusters is an often overlooked aspect of IDPs sampling and the main reason for samples failing to replicate the ensemble and thermodynamic averages of experiments. Adequate sampling is required for the determination of experimentally consistent ensemble data from simulation, which is typically accomplished in advanced sampling approaches by either applying structural restraints using collective variables or re-weighting the obtained conformations to arrive at Boltzmann weighted populations[67, 68]. Parallel tempering (PT) sampling is appealing because it can be used effectively without any reweighing or restraining, and it does not require the use of a low-dimensional collective variable (CV) to define the ensemble states. Furthermore, in cases where sampling results do not match experimental data, PT can be seamlessly combined with other CV-based restraining methods or re-weighted to solve the problems of interest[69, 70, 71]. Several variants of PT have evolved in recent years like TREMD, REST/REST2, REHT and gREST.

## 1.4 IDP Examples

### 1.4.1 Prostate-Associated Gene 4 (PAGE4)

Prostate Cancer (PCa) is a leading cause of mortality and morbidity around the world. PAGE4, a protein that appears to act as both an oncogenic factor and a metastasis suppressor, has been identified as a novel therapeutic target for PCa. PAGE4 is a prostate-specific Cancer/Testis Antigen that is highly upregulated in the human foetal prostate and its diseased states, but not in the adult normal gland[72]. The PAGE4 protein is predicted to be highly disordered by bioinformatic algorithms[73]. PAGE4 is expected to have several regions (most notably residues 13–19 and 86–92, and to a lesser extent 49–61) with a slightly increased propensity to order, according to these analyses[74]. Also, it has been reported that, PAGE4 has metastable secondary structures, according to nuclear magnetic resonance (NMR) experiments.

PAGE4 acts as a stress-response protein by suppressing reactive oxygen species and preventing DNA damage. The kinase HIPK1 can phosphorylate PAGE4 at two residues (S9, T51); phosphorylation of PAGE4 allows it to interact with the AP-1 transcription factor complex [75]. Another kinase, CLK2, can phosphorylate PAGE4, and the two phosphorylated versions of PAGE4 (HIPK1-PAGE4 and CLK2-PAGE4) have opposing functions due to their different conformational dynamics[76, 72]. CLK2-PAGE4 has a reduced affinity for AP-1 due to its random coil-like structure, whereas HIPK1-PAGE4 has a compact conformational ensemble that can bind AP-1 and potentiate c-Jun[72]. Because c-Jun potentiation indirectly increases CLK2 levels via AR, a negative feedback loop is formed, resulting in oscillations in AR levels and those of different phosphorylated versions of PAGE4. These oscillations can cause non-genetic

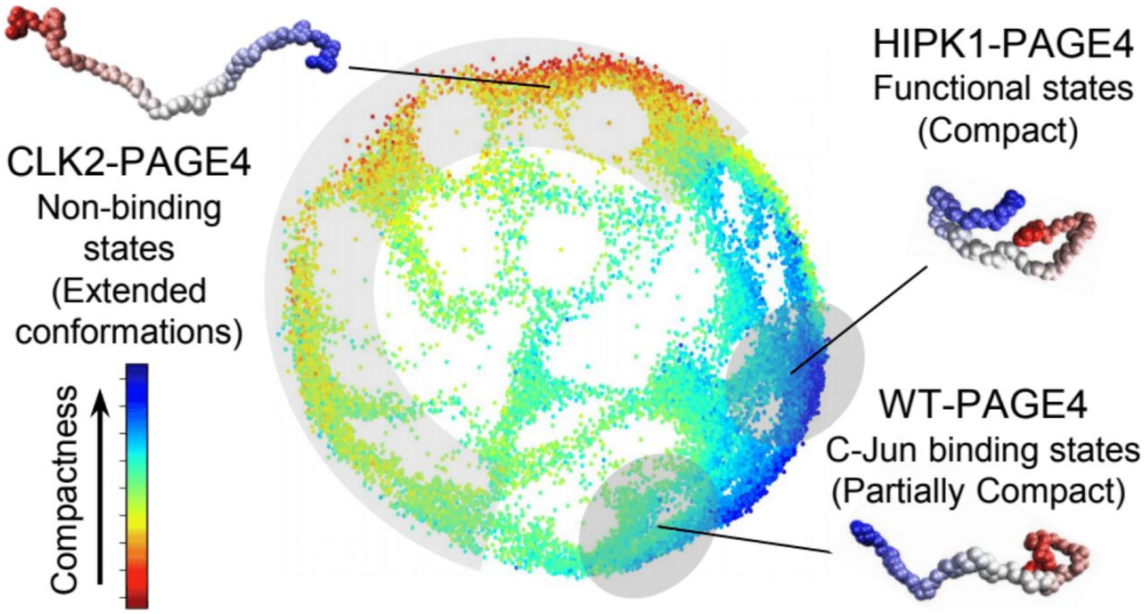



Figure 1.2: ELViM representation of different phosphorylated states of PAGE4; adapted from Ref. [77]

heterogeneity in a clonal prostate cancer cell population, as well as dynamic levels of AR in individual cells, which can affect their therapeutic sensitivity.

### 1.4.2 Ras-family proteins

Ras, which stands for "Rat Sarcoma Virus," is a group of proteins found in all animal cell types and organs. Ras proteins are members of the small GTPase protein family, which is involved in signal transmission within cells. When incoming signals "switch on" Ras, it activates other proteins, which in turn activate genes involved in cell growth, differentiation, and survival. Ras gene mutations can result in the production of permanently activated Ras proteins, which can cause unintended and overactive signalling within the cell even when no external signals are present. Overactive Ras signalling can lead to cancer because these signals cause cell growth and division[78]. HRAS, KRAS, and NRAS are the three most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20 to 25 percent of all human tumours, and up to 90 percent in certain types of cancer[79]. As a result, Ras inhibitors are being

investigated as a potential treatment for cancer and other diseases characterised by Ras over-expression.

Six beta strands and five alpha helices make up Ras[80]. It has two domains: a G domain that binds guanosine nucleotides and a C-terminal membrane targeting region (CAAX-COOH, also known as CAAX box) that is lipid-modified by farnesyl transferase, RCE1, and ICMT. The G domain has five G motifs that directly bind GDP/GTP. The P-loop, also known as the G1 motif, binds the beta phosphate of GDP and GTP. The threonine35 in the G2 motif, also known as Switch-I or SW1, binds the terminal phosphate (-phosphate) of GTP and the divalent magnesium ion bound in the active site. The DXXGQ motif is found in the G3 motif, also known as Switch-II or SW2. The D stands for aspartate57, which is specific for guanine versus adenine binding, and the Q stands for glutamine61, which activates a catalytic water molecule for GTP to GDP hydrolysis. The LVGNKXDL motif in the G4 motif provides specific interaction with guanine. A SAK consensus sequence can be found in the G5 motif. The A is alanine146, which provides guanine specificity rather than adenine specificity. When GTP is hydrolyzed into GDP, the two switch motifs, G2 and G3, are the main parts of the protein that move. The basic functionality of a molecular switch protein is mediated by this conformational change mediated by the two switch motifs. The “on” state of Ras is the GTP-bound state, while the “off” state is the GDP-bound state. Ras also binds a magnesium ion, which aids in nucleotide binding coordination.

# 2 Concepts of Computational Methods and Techniques

## 2.1 Basics of Statistical Mechanics

### 2.1.1 Phase Space Trajectory

Let a particle be in a one-dimensional space at the time $t = 0$. So to describe the particle's future trajectory at any time t, we need the time-dependent evolution of the position and momentum coordinates. Therefore phase space is the two-dimensional coordinate space constructed by these two coordinates, position and momentum, when plotted together. When this phase space is extended to three-dimensional space the dimensionality of the phase space becomes six and is referred to as the $\mu$-space. When extended to N number of particles it creates a 6N-dimensional phase space, also called the $\Gamma$-space. Thus, the particle's temporal position can be represented by the curve depicted by the representative point with the time (t). This representation of the time evolution of the position and moment of representative points in phase space known as the trajectory.

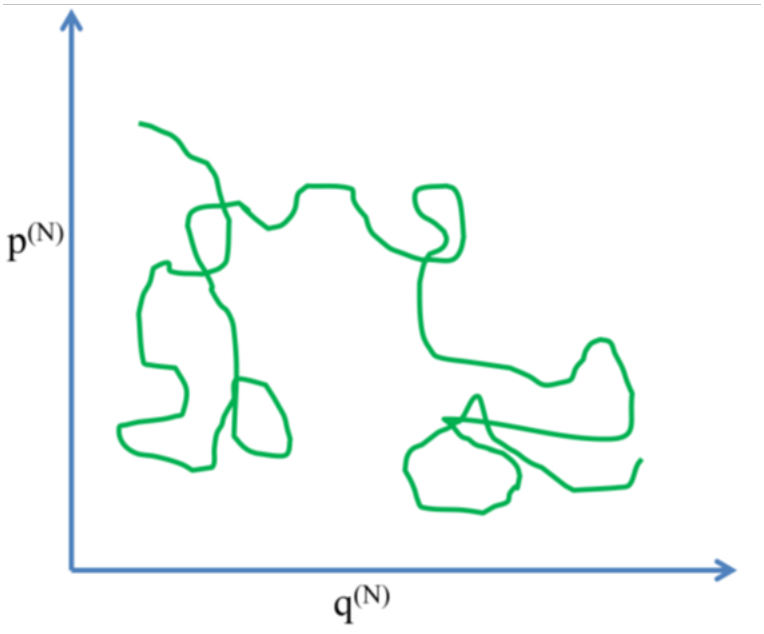



Figure 2.1: Schematic representation of the trajectory of a system of N particles

### 2.1.2 Ensembles

The concept of the ensemble is based on the fact that an equilibrium system is made up of a large number of microscopic states, also known as microstates. When the system's temperature is non zero, the system's natural motion takes it through some of these microstates in a timescale comparable to the macrostate's measurement timescale. If we assume that the system has constant energy at all times, the trajectory moves on a constant energy surface. As a result, the probability distribution of the thermodynamical and mechanical properties at these microstates is required to calculate the system's average equilibrium properties. Instead of chasing the time evolution of these microstates to obtain this distribution, we consider a mental picture of a large number of systems with similar macroscopic properties such as the number of particles (N), pressure (P), volume (V), and energy (E). Given the large number of microstates, it is highly likely that the microstates inhabited by each system are distinct. Hence, ensemble is a mentally constructed collection of thermodynamically identical systems. The ensembles can be classified based on the thermodynamic constraints.

- Microcanonical Ensemble (NVE): The number of particles, volume, and energy of the system all remain constant in this ensemble; thus, each microstate in the ensemble should have the same energy, volume, and number of particles. As a result, the system must be an isolated system that cannot exchange energy or particles with its surroundings in order to obtain this ensemble. The schematic is represented in the Figure 2.2.

- Macrocanonical Ensemble (NVT): Although the energy of the system can vary, the number of particles, volume, and temperature of the system remain constant in this ensemble, so each microstate in the ensemble should

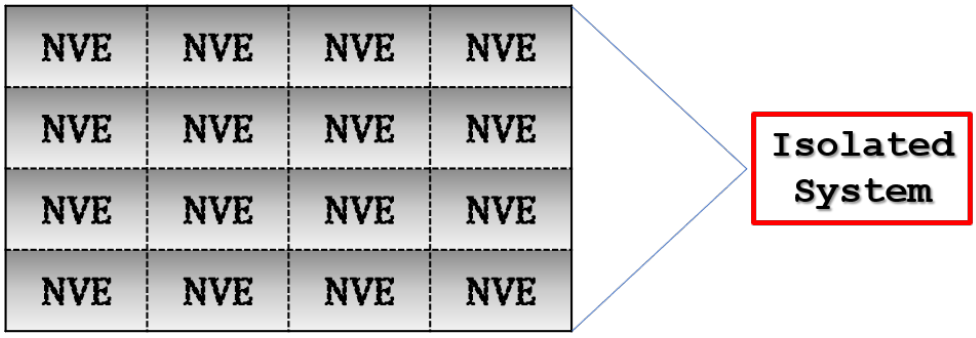



Figure 2.2: Schematic representation of microcanonical ensemble

correspond to the same temperature and have the same number of particles and volume. To achieve this ensemble, the system must be a closed system that cannot exchange particles with the environment but can exchange energy in order to maintain an equilibrium temperature. In molecular dynamics simulation, this is the most commonly used ensemble. As a result, we'll talk about it more in the second half of the chapter.The schematic is represented in the Figure 2.3.

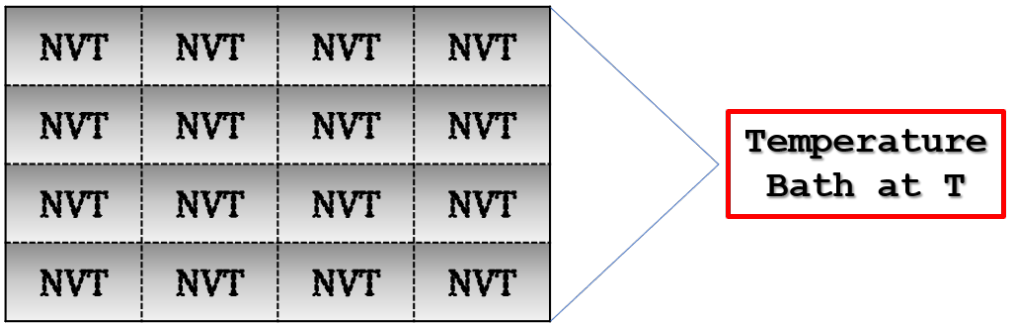



Figure 2.3: Schematic representation of macrocanonical ensemble

- Grandcanonical Ensemble ($\mu$VT): Neither the energy nor the number of particles are fixed in this ensemble. The chemical potential, volume, and temperature in this system remain constant. The system should be open for this ensemble. The schematic is represented in the Figure 2.4.

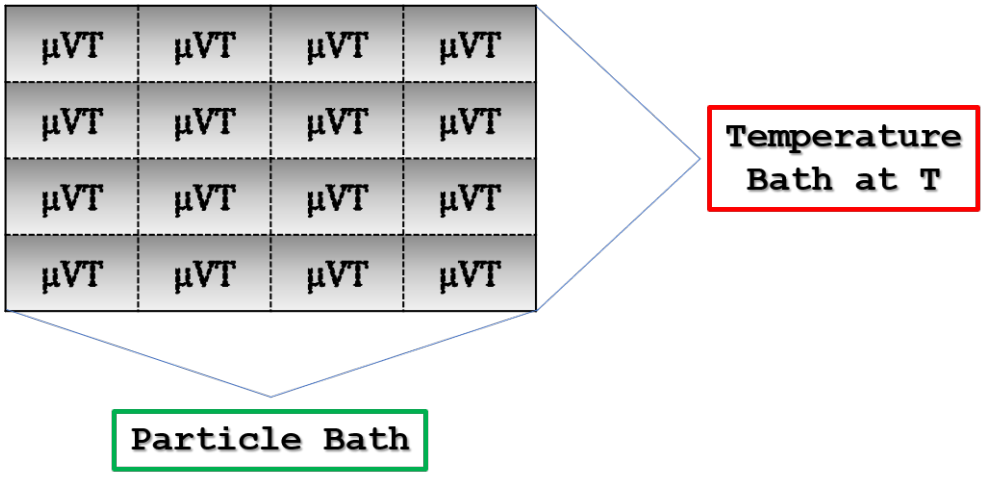



Figure 2.4: Schematic representation of grandcanonical ensemble

### 2.1.3 The First Postulate: Time average is equal to Ensemble average

The time average of a quantity ($\bar{O}$) can be defined as the average of the quantity over a long period of time. Mathematically,

$$\bar{O} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t_0}^{t_0+\tau} O(t)dt \tag{2.1}$$

On the other hand, the ensemble average can be defined as:

$$\langle O \rangle = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} O_i p_i \tag{2.2}$$

here $p_i$ is the probability of the particle in the ensemble to be in $i^{th}$ microstate. For large N and $\tau$, the first postulate of statistical mechanics states that "the time average is equal to the ensemble average", therefore,

$$\bar{O} = \langle O \rangle \tag{2.3}$$

### 2.1.4 The Second Postulate: Equal A Priori Probability

When given a very long time, this postulate states that a system has an equal chance of being in any microstate corresponding to the system's macrostate.

In other words, given enough time, the system visits each of the microstates an equal number of times. However, because the criteria for a long time have not been defined properly, it is assumed that this time is much longer than any relaxation time of the system, preventing us from capturing it in all possible states.

### 2.1.5 The Ergodic Hypothesis

The Ergodic hypothesis states that if a system is given enough time, it is free to explore all of the microstates associated with it, and the time spent in each microstate is proportional to its volume in phase space. However, if a system becomes trapped in a region of phase space, this hypothesis is broken, and the ensemble average is no longer equal to the time average, violating the statistical mechanics' first postulate. As a result, the Ergodic hypothesis serves as a link between the two statistical mechanics postulates. The second postulate is valid because of the Ergodic hypothesis, and the ensemble average is done because of the second postulate.

### 2.1.6 Canonical Partition Function

We would refer from here the macrocanonical ensemble as the canonical ensemble. We now consider a system consisting of $N_p$ number of replicas of the original NVT system, as we know in canonical ensemble N, V, and T are kept constant. All of the ensemble's systems are placed next to each other so that they can exchange heat through their heat-conducting walls, but not matter. This entire ensemble is placed in a heat bath, and after reaching equilibrium, a thermal insulation is placed around the entire ensemble, forming an isolated super-system. This is done to convert the canonical ensemble to a microcanonical ensemble, allowing statistical thermodynamics postulates that are only true

for microcanonical ensemble to be applied to canonical ensemble as well. The microcanonical super-ensemble of the system under consideration is made up of mental replicas of such a super-system.

Now we denote $n_i$ as the number of systems in our super-energy system's $E_i$, and $E_t$ as the total energy of the super-system. Hence,

$$\sum_i n_i = N_p \text{ and } \sum_i n_i E_i = E_t \tag{2.4}$$

The number of possibilities for distributing $n_i$ number of systems with energy $E_i$ in $N_p$ number of states are:

$$\Omega(n_i) = \frac{N_p!}{n_1! n_2! n_3! \dots} \tag{2.5}$$

Hence , probability of occurrence of a given system $n_i$ with energy $E_i$ is:

$$P_i = \frac{\sum_i n_i \Omega(n_i)}{\sum_i \Omega(n_i)} \tag{2.6}$$

Using Stirling's approximation,

$$\ln \Omega_j(n) = \left( \sum_j n_j \right) \ln \left( \sum_j n_j \right) - \left( \sum_j n_j \ln n_j \right) \tag{2.7}$$

We now use the lagrange undetermined multiplier method to incorporate the conditions:

$$\frac{\partial}{\partial n_i} \left[ \ln \Omega_j(n) - \alpha \sum_j n_j - \beta \sum_j n_j E_j \right] = 0 \tag{2.8}$$

The two undetermined multipliers here are $\alpha$ and $\beta$. Now we will differentiate using the maximum term method to obtain the following expression:

$$\ln \left( \sum_j n_j \right) - \ln n_i^* - \alpha - \beta E_i = 0 \tag{2.9}$$

$$\frac{n_i^*}{N} = e^{-\alpha-\beta E_i}, \quad j = 1, 2, \ldots \tag{2.10}$$

Here $n_i^*$ is the most probable distribution. Here sum over $n_i^*$ is equal to the total number of systems in the ensemble.

$$e^{\alpha} = \sum_i e^{-\beta E_i} \tag{2.11}$$

Hence,

$$\bar{P}_i = \frac{n_i^*}{N} = \frac{e^{-\beta E_i(N,V)}}{\sum_i e^{-\beta E_i(N,V)}} \tag{2.12}$$

The denominator becomes the canonical partition function (Z).

$$Z_N(V,T) = \sum_i e^{-\beta E_i(N,V)} \tag{2.13}$$

### 2.1.7 Relating Thermodynamics and Partition Function

The average energy, $\bar{E}$ can be defined mathematically as follow:

$$\bar{E} = \sum_i P_i E_i = \frac{\sum_i E_i e^{-\beta E_i(N,V)}}{\sum_i e^{-\beta E_i(N,V)}} \tag{2.14}$$

Hence,

$$\begin{aligned} d\bar{E} &= \sum_i (E_i dP_i + P_i dE_i) \\ &= \frac{1}{\beta} \sum_i \left[ (\ln P_i + \ln Z) dP_i + P_i \left( \frac{\partial E_i}{\partial V} \right)_N dV \right] \end{aligned} \tag{2.15}$$

Now , we know that,

$$\sum_i P = 1 \Rightarrow \sum_i dP_i = 0 \tag{2.16}$$

$$S = -k_B \sum_i P_i \ln P_i \tag{2.17}$$

We also know that,

$$dS = -k_B d\left(\sum_i P_i \ln P_i\right) = -k_B\left(\sum_i dP_i + \sum_i \ln P_i dP_i\right) \tag{2.18}$$

These two equations can be combined, to get,

$$-\frac{1}{\beta} d\left(\sum_i P_i \ln P_i\right) = d\bar{E} + pdV \tag{2.19}$$

Comparing with, $TdS = dE + pdV$, we get,

$$TdS = -\frac{1}{\beta} d\left(\sum_i P_i \ln P_i\right) \tag{2.20}$$

Hence,

$$\beta = \frac{1}{k_B T} \tag{2.21}$$

Using the thermodynamic relationship between entropy, internal energy and Helmholtz free energy, we get,

$$S = \frac{\bar{E}}{T} + k_B \ln Z = \frac{\bar{E}}{T} - \frac{A}{T} \tag{2.22}$$

Hence,

$$A = -k_B \ln Z(N, V, T) \tag{2.23}$$

Moreover, we also know that,

$$dA = -SdT - pdV \tag{2.24}$$

Hence,

$$S = -\left(\frac{\partial A}{\partial T}\right)_{V,N} = k_B T\left(\frac{\partial \ln Z}{\partial T}\right)_{V,N} + k_B \ln Z \tag{2.25}$$

and

$$p = -\left(\frac{\partial A}{\partial T}\right)_{T,N} = k_B T \left(\frac{\partial \ln Z}{\partial T}\right)_{T,N} \tag{2.26}$$

## 2.2 Basic Concepts of Molecular Dynamics Simulation

### 2.2.1 General Features of Force Field

Force Fields are the most important part of molecular dynamics simulations. Force fields contain terms that help to calculate the overall energy of the system at any time point. As according to the Born-Oppenheimer approximation, the electronic and nuclear motion of the system can be decoupled, in the force fields we only account for the nuclear part of the system[81]. This greatly reduces the computational cost because we do not have to look for the electronic motions[82, 83]. However, the drawback is that it is unable to predict bond formation and breakage.

The terms in the force fields are simple and come from several molecular motions of bonds, for example stretching and bending, described by Hooke's law[84, 85, 86]. Most of the force fields describe the motion with four components, the first two arising from bonds, the third one arising from the bond rotations, and the last one pertains to the non-bonded interactions. Therefore, the general form looks like this:

$$V(r^N) = \sum_{bond} V_{ij} + \sum_{angle} V_{ijk} + \sum_{dihedral} V_{ijkl} + \sum_{nb} V_{ij} \tag{2.27}$$

Here, the r signifies that V is the function of particle coordinate and the N is the number of particles in the system.

#### 2.2.1.1 Bonded Potential

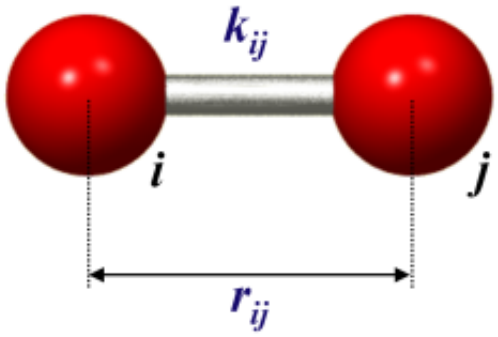



Figure 2.5: Atoms i and j connected by spring with force constant $k_{ij}$; adapted from Ref. [87]

The bonded potential corresponds to the energy term contributed by the covalent bonds present in the system and is derived from Hooke's law. The expression is as follows:

$$\sum_{bond} V_{ij} = \sum_{bond} \frac{1}{2} K_{ij}(r_{ij} - r_{ij}^{eq})^2 \tag{2.28}$$

Here, $K_{ij}$ refers to the force constant of the covalent bond, $r_{ij}$ is the instantaneous bond length, and $r_{ij}^{eq}$ is the equilibrium bond length between atoms i and j.

#### 2.2.1.2 Angular Potential

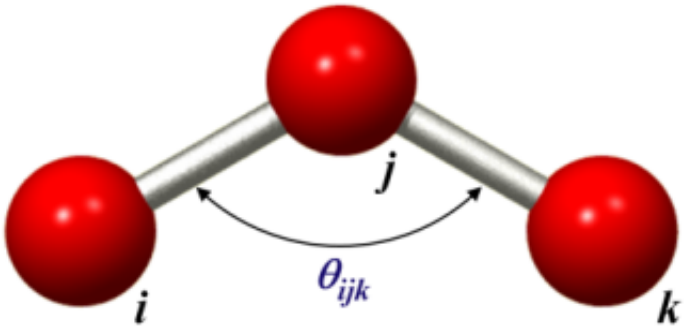



Figure 2.6: Atoms i, j and k making an angle $\theta_{ijk}$; adapted from Ref. [87]

The angular potential term corresponds to the energy term contributed by the vibrational angular motion present in the system corresponding to atoms i, j, and k. The expression is as follows:

$$\sum_{angle} V_{ijk} = \sum_{angle} \frac{1}{2}[K_{ijk}(\theta_{ijk} - \theta_{ijk}^{eq})^2 + K_{UB}(r_{ik} - r_{ik}^{eq})^2] \tag{2.29}$$

Here $K_{ijk}$ refers to the angle constant and $K_{UB}$ is the Urey-Bradley constant used to describe a non-covalent spring between the i*th* and k*th* atom. $\theta_{ijk}$ is the instantaneous angle term and $\theta_{ijk}^{eq}$ is the equilibrium angle between i, j, and k. $r_{ik}$ refers to the instantaneous distance between atoms i and k, and $r_{ik}^{eq}$ is the respective equilibrium term.

#### 2.2.1.3 Torsional Potential

The torsional potential term corresponds to the energy term contributed by the dihedral angular spring present between the two planes made by the first three atoms and by the last three atoms. There are two types of dihedral terms:

- *Proper Dihedral*: The proper dihedral consists of four atoms which are joined consecutively in a chain fashion.

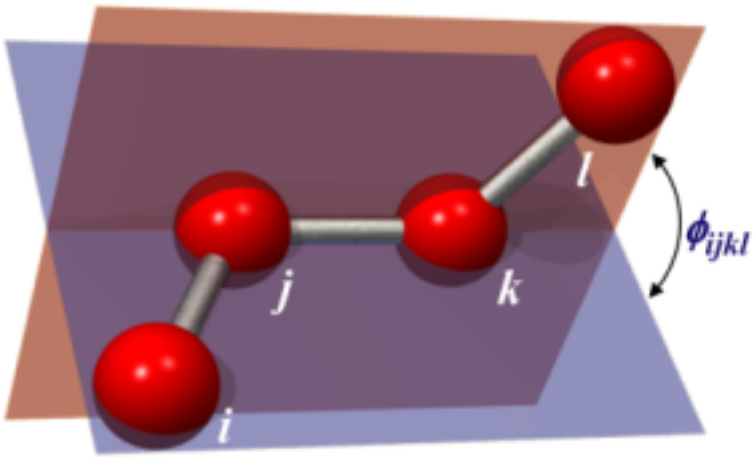



Figure 2.7: Atoms i, j, k and l making a proper dihedral $\theta_{ijkl}$; adapted from Ref. [87]

The potential can be expressed as follows, taking the first few terms from the Fourier transform:

$$\sum_{dihedral} V_{ijkl} = \sum_{dihedral} \frac{1}{2} K_{ijkl}(1 + cos(n\phi_{ijkl} - \phi_0)) \tag{2.30}$$

Here $K_{ijkl}$ is the torsional angle constant, the $\phi_{ijkl}$ is the instantaneous dihedral angle and $\phi_0$ is the minimum-potential angle. n is the *multiplicity*, which implies the number of minima present in a complete $360^o$ rotation of the dihedral.

- *Improper Dihedral*: The improper dihedral consists of four atoms, where the three atoms are connected to one central atom.

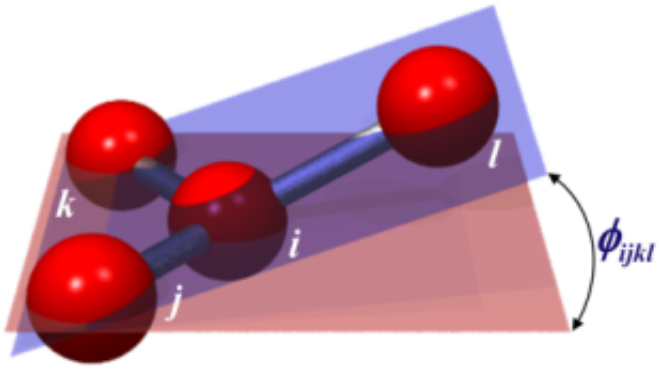



Figure 2.8: Atoms i, j, k and l making a improper dihedral $\theta_{ijkl}$; adapted from Ref. [87]

It helps to maintain the chirality of a heavy atom or to maintain the planarity. This is approximated as a harmonic potential. The expression is as follows:

$$\sum_{dihedral} V_{ijkl} = \sum_{dihedral} \frac{1}{2} K_{ijkl}(\phi_{ijkl} - \phi_0)^2 \tag{2.31}$$

Here $K_{ijkl}$ is the torsional angle constant, the $\phi_{ijkl}$ is the instantaneous dihedral angle and $\phi_0$ is the reference dihedral angle.

#### 2.2.1.4 Non-bonded Potential

The non-bonded potential can be of two types:

- *Electrostatic Potential*: The electrostatic potential accounts for the interaction of the charged species. It is repulsive for species with the same charge and is attractive otherwise.

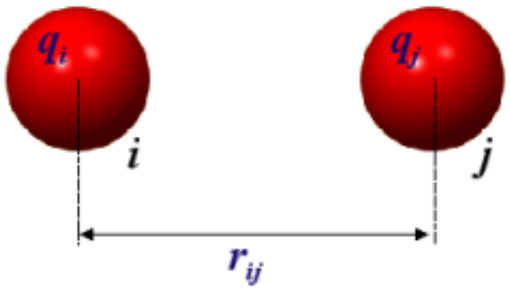



Figure 2.9: Charged atoms i and j separated by distance $r_{ij}$; adapted from Ref. [87]

The expression is given by the coloumbic potential expression as follows:

$$\sum_{elec} V_{ij} = \sum_{elec} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{2.32}$$

Here, $q_i$ and $q_j$ are the charges of species i and j, respectively. The $r_{ij}$ corresponds to the distance between the two species.

- *Lennard-Jones Potential*: The Lennard-Jones potential accounts for the weak dipole interaction energy between any two species i and j.

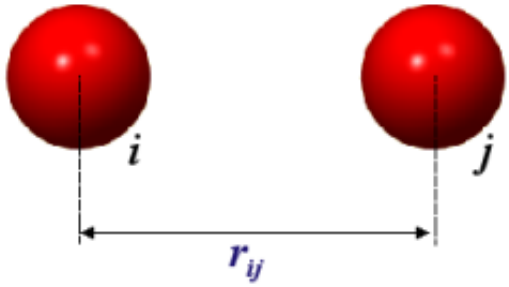



Figure 2.10: Atoms i and j separated by distance $r_{ij}$; adapted from Ref. [87]

The expression is as follows:

$$\sum_{LJ} V_{ij} = \sum_{LJ} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.33}$$

Here $\epsilon_{ij}$ is the depth of the potential well. The $r_{ij}$ corresponds to the distance between the two species and $\sigma_{ij}$ refers to the distance at which the species-species potential energy becomes zero.

### 2.2.2 Energy Minimisation

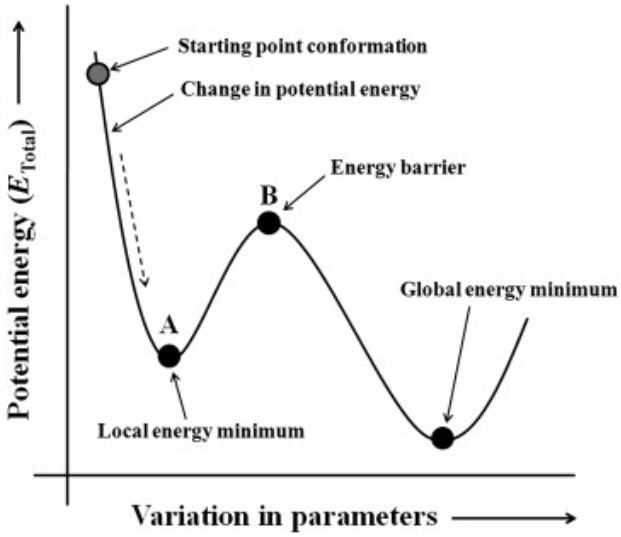



Figure 2.11: Energy Minimisation Scheme; adapted from Ref. [88]

Before starting of any simulation, the system must be at the minimum energy conformation. However, the structures reported in the databases are not in the energy-minimised state and the randomly added solvent configurations also can create strong steric clashes among themselves or with the solvent. Therefore, to obtain the local minima, the process of energy minimisation is performed. Under this algorithm, the slope is equated to zero and the differential gradient

is used for obtaining the local minima, as follows:

$$\frac{\partial V}{\partial R_i} = 0 \,\&\, \frac{\partial^2 V}{\partial R_i^2} > 0 \tag{2.34}$$

Here V is the potential associated with the system and $R_i$ is the coordinate in 3-dimension for the atoms in the system.

Most of the minimisation algorithms tend to locate the minima closest to the initial configuration. The most used ones are: Steepest Descent and Conjugate Gradient[81].

#### 2.2.2.1 Steepest Descent

This algorithm lets the system to take one step each time and moves in the direction opposite to the gradient of the potential energy, a coordinate function, as follows:

$$r_{n+1} = r_n - \alpha_n \nabla V(r_n) \tag{2.35}$$

where $\alpha_n$ is the step-size and $\nabla V(r_n)$ is the potential energy gradient function. However, *GROMACS* instills this in a slightly conditional way:

$$r_{n+1} = r_n - \frac{\nabla V(r_n)}{\max[\nabla V(r_n)]} \alpha_n \tag{2.36}$$

here $max[\nabla V(r_n)]$ is the largest scalar force on any atom in the system. Then,

- If $V_{n+1} < V_n$, the new position is added and $\alpha_{n+1}$ is rescaled to $1.2\alpha_n$
- If $V_{n+1} > V_n$, the new position is rejected and $\alpha_{n+1}$ is rescaled to $0.2\alpha_n$

This algorithm can be stopped manually by the user by giving a specific number of steps for iteration, or giving an accuracy term.

#### 2.2.2.2 Conjugate Gradient

This method can be used to choose successive search directions that avoid the constraint of repeated minimization in the same direction[81].A minimum range is first determined in each direction, then converged using either a golden section search or a quadratically convergent method. It considers the gradient history when determining the best next step direction.
In the early stages of the minimization, this approach is slower than steepest descent, but it becomes more efficient as you get closer to the energy minimum. The stop criterion and parameters used in conjugate gradient are the same as those used in steepest descent in GROMACS[89].

### 2.2.3 Basic Approach

Here now we have the interaction potential terms for the atomic particles and the corresponding energy minimised structure. Hence, now we can use the Newton equation to calculate the position and velocities of the atoms at different instances. Therefore, from the potential, we need to find the force as follows:

$$F_i = -\frac{\partial V}{\partial r_i} \tag{2.37}$$

After finding the force we now need to find the particle position from the momentum equation as follows:

$$F_i = \frac{dp_i}{dt} = m\frac{dv_i}{dt} = m\frac{d^2r_i}{dt^2} \tag{2.38}$$

The position and velocity can then be found by integrating the above equation 2.38.

### 2.2.4 Numerical Integration Methods

As described above, the position and velocities of the particles can be calculated using Newton's equations of motion. However, the initial position and velocity needs to be specified. The position is specified by the energy minimised structure and the velocity is randomly assigned by a Maxwell-Boltzmann distribution, as follows:

$$P(v_i) = \sqrt{\frac{m_i}{2\pi k_B T}} \exp\left(-\frac{m_i v_i^2}{2k_B T}\right) \tag{2.39}$$

Here $P(v_i)$ is the probability of particle i with mass $m_i$ to have the velocity $v_i$ at temperature T.

After this has been implemented different integration algorithms are used, such as (i) Verlet algorithm[90], (ii) Leap-frog algorithm[91], and (iii) Velocity Verlet algorithm[92] for numerically performing the integration and obtain the future coordinates and velocity.

#### 2.2.4.1 Verlet Algorithm

Using the Verlet algorithm[90], the velocity at $t$ and position at time $t+\delta t$ can be calculated using positions from time $t$ and $t-\delta t$. This is an iterative algorithm. The position at time $t+\delta t$ and $t-\delta t$ can be expressed as:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)(\delta t)^2 + ... \tag{2.40}$$

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)(\delta t)^2 + ... \tag{2.41}$$

Adding Equation 2.40 and 2.41 yields:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)(\delta t)^2 \tag{2.42}$$

And the velocities can be calculated as follows:

$$v(t) = \frac{[r(t+\delta t) - r(t-\delta t)]}{2\delta t} \tag{2.43}$$

#### 2.2.4.2 Leap-Frog Algorithm

Under the Leap-Frog Algorithm[91], the following terms are used:

$$v\left(t + \frac{1}{2}\delta t\right) = v\left(t - \frac{1}{2}\delta t\right) + a(t)\delta t \tag{2.44}$$

$$r\left(t + \delta t\right) = r(t) + v\left(t + \frac{1}{2}\delta t\right) \tag{2.45}$$

So as can be seen, the velocity is implemented first and then the position equation. The velocity leaps over the position to give the next midpoint values. The new positions $r(t+\delta t)$ depend on $v\left(t+\frac{1}{2}\delta t\right)$, which in turn depends on $v\left(t-\frac{1}{2}\delta t\right)$ and $a(t)$. Hence, the velocities at time t is as follows:

$$v(t) = \frac{1}{2}\left[v\left(t + \frac{1}{2}\delta t\right) + v\left(t - \frac{1}{2}\delta t\right)\right] \tag{2.46}$$

Therefore, this algorithm calculates the half-integer step velocities which are used to calculate the coordinates of the next steps.

#### 2.2.4.3 Velocity-Verlet Algorithm

Under the Velocity-Verlet Algorithm[92], the following terms are used:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)(\delta t)^2 \tag{2.47}$$

$$v(t+\delta t) = v(t) + \frac{1}{2}\left[a(t) + a(t+\delta t)\right]\delta t \tag{2.48}$$

The $v(t+\delta t)$ is determined substituting:

$$v\left(t+\frac{1}{2}\delta t\right) = v(t) + \frac{1}{2}a(t)\delta t \tag{2.49}$$

Hence, the equation becomes:

$$v(t+\delta t) = v\left(t+\frac{1}{2}\delta t\right) + \frac{1}{2}a(t+\delta t)\delta t \tag{2.50}$$

## 2.3 Temperature and Pressure Control

For performing a simulation of biomolecules a constant number, volume and temperature (NVT) needs to be maintained, which is essentially a canonical ensemble. In addition to that, the density needs to be maintained, and hence we use the constant temperature and pressure coupling, which is also an extended Hamiltonian method, known as the isothermal-isobaric (NPT) ensemble.

### 2.3.1 Temperature Coupling

During the molecular dynamics simulation, the velocity ($v_i$) are rescaled for each of the atoms in the system according to the equipartition theorem so as to maintain the temperature of the system. Hence, the temperature at each step of the trajectory is calculated and the velocities are rescaled to bring the instantaneous temperature (T) to the required temperature ($T_{req}$).

$$\frac{3}{2}k_B T = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}m_i v_i^2 \tag{2.51}$$

$$v_i \rightarrow v_i\sqrt{\frac{T_{req}}{T}} \tag{2.52}$$

Here $v_i$ is the velocity of the i*th* particle having mass $m_i$ in a system of N atoms. The equation implies an isokinetic thermostat.

However, their is an intrinsic problem with isokinetic thermostat. It strictly maintains a constant temperature, which is highly unrealistic in a biomolecular system. Rather, biomolecular systems experience a range of temperatures maintained on an average with the required temperature.

Therefore, small kinetic energy fluctuations, which is dependent on the temperature, are allowed. And as the number of particles (N) increases the fluctuations reduce. However, since simulation can't account for huge number of particles due to computational cost, appropriate thermostats are required which can allow thermal fluctuations along with maintenance of an average required temperature. Therefore, various thermostats have been developed such as Anderson[93], Berendsen[94], Nose-Hoover thermostat[95, 96] etc. These thermostats generate thermodynamics ensembles where the average temperature is maintained throughout the trajectory.

For example, in the Nose-Hoover thermostat the Boltzmann distribution is retained along with an extended ensemble approach. The system is strongly coupled with the required temperature ($T_{req}$), giving the Hamiltonian extra degrees of freedom.

### 2.3.2 Pressure Coupling

The major aim of the pressure coupling is to maintain a constant pressure throughout the simulation[84]. The equipartition theorem can be used and can be written as:

$$PV = Nk_BT + \langle W \rangle \tag{2.53}$$

Here $W = -\frac{1}{3}\sum_i \vec{r_i}\vec{F_i}$, where $\vec{F_i}$ is the force vector on each individual particle and $\vec{r_i}$ is the coordinate of it. Here is the pressure coupling the pressure is kept constant by rescaling the box vectors.

Similar to the thermostat described above, the Parinello-Rahman [95, 97] barostat also uses an extended ensemble approach, generating a isothermal-isobaric (NPT) ensemble in combination with Nose-Hoover thermostat. Under this, the Netwon's equation are modified to also incorporate the pressure, volume and temperature. In addition to that, it includes additional terms to account for the strength of the coupling between the thermostat and barostat to the system.

## 2.4 Tricks for Computational Efficiency

### 2.4.1 Periodic Boundary Conditions

The goal of molecular dynamics is to study bulk properties of a system. However, a real system consists of particles in the scale of Avogadro number of particles ($\approx 10^{23}$) or more. But with the current computational power mankind has a system in the order of thousand particles can be simulated in a reasonable amount of time. And small sized systems can give rise to inadvertent errors effecting the bulk properties. For example, surface effect is a bulk property. In a real system, the ratio between number of particles at the surface and total number of particles is insignificant and hence the surface effect can be neglected. However, in simulated cells, for a particle of $500$ particles, $500^{2/3} = 63$ particles stay at the surface and hence the influence of surface effect can't be neglected. This drawback is therefore overcome by replacing the boundaries of the system with periodic images. This technique is called periodic boundary condition.

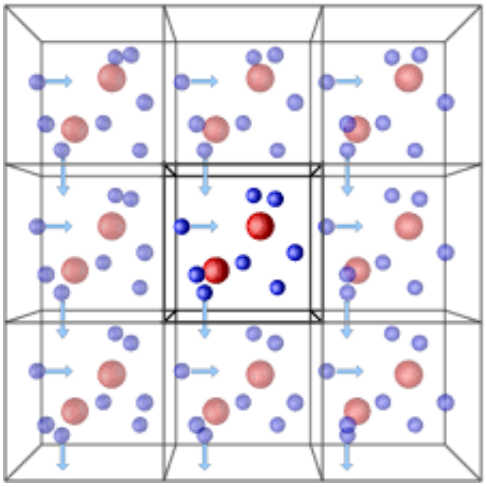


Figure 2.12: Periodic Boundary Conditions; adapted from Ref. [98]

Using this concept, the periodic images are arranged in all possible directions in a 3-D lattice. Therefore, the particle coordinates are calculated by adding integral multiples of the length of the box edges to the coordinates. Therefore, if a real particle goes out of the box during the simulation, then an image particle enters the box from the opposite side to mimic the real system. And for the calculation of particle interactions within the cutoff range, both the particle neighbours are included.

### 2.4.2 Minimum Image Convention and Truncation of Intermolecular Interaction

This concept of Minimum Image Convention (MIC) and Truncation of Intermolecular Interaction (TIMI)[84] are very closely coupled to the PBCs. The new problem introduced by PBC is of the force calculation in the system because technically periodic boundary condition technically implies infinite atoms. So, a cut-off zone must be introduced, which must be less than or equal to the box dimension. This is called minimum image convention or MIC. The force here is calculated both for the real and periodic images; however, a particle should not see its own image. The cutoff is fixed to the original box dimension for particle at the exact center.

On the other hand, Truncation of Intermolecular Interaction (TIMI) takes into account that two atoms separated by large distance negligibly interact or don't interact at all. This is true for short range interaction potentials where $V(r) \propto \frac{1}{r^n}$ and $n > 3$. Therefore, not all particles needs to be considered greatly reducing computing cost.

So, for a system of $N$ particles, there are a total of ${}^{N}C_2 \sim (N-1)N$ interacting pairs. Therefore, if $N >> 1$, then $(N-1)N \approx N^2$, which also implies that computational power is squared. Therefore, MIC and TIMI come to the rescue and save a magnitude of computational power by defining a spherical cutoff range.

### 2.4.3 Long Range Forces: Ewald Summation and Particle Mesh Ewald

Methods described above such as the TIMI only take account short-ranged interactions neglecting the Coloumbic and ion-dipole interactions. Therefore, in such interactions implementing a cutoff radius will result in a very high error in the interaction force calculations. To rescue these, the Ewald summation[99] and Particle Mesh Ewald[100, 101] were introduced.

For the long-range forces, all the periodic images of the box are used to calculate the electrostatic potential. Hence, the total electrostatic potential on an atom 'i' is derived as the infinite pair potential sum of all the charged particles in the box and their respective images. The sum that is derived is divided into two parts: long-ranged and short-ranged. The short-ranged part is calculated using the cutoff scheme, whereas the long-ranged part is usually taken care by the Ewald summation methods[99] by decomposition.

In addition to this, under the Particle Mesh Ewald[100, 101], each atom is represented on a mesh grid and the potential function is represented as the interaction

between the mesh points. Each mesh point is a separate particle that interacts with all other mesh particles in a convolution. The potential function in the mesh space is then evaluated via a Fast Fourier transformation. The mesh size and interpolation strategy determine the accuracy and computational efficiency.

### 2.4.4 Neighbour Lists and Cell Lists

However, the use of cutoff distance drastically decreases the efficiency of the interaction calculation. Because to implement the cutoff scheme all the atoms need to be mapped to look into which atoms come and do not come under the cutoff radius, which is then used to calculate the final interaction energy.

Therefore, the neighbour list scheme[90, 91, 102] is implemented to increase the computational efficiency. A list of nearby atoms to be included in the non-bonded interaction calculation is stored in an array and updated periodically in this method. The distance used to calculate each atom's neighbouring list must be greater than the non-bonded cutoff distance, so that no atom outside the neighbour cutoff gets closer than the non-bonded cutoff distance before the neighbour list is updated. A correction term can be added to the energy estimate at each step of updating the neighbour list.

Calculating the order on $N^2$ for an 'N' particle system is required to prepare and update such a neighbouring list. To reduce the number of calculations or neighbour searches, the entire simulation space can be divided into cells, with the search limited to particles that are present within the cells[102, 103, 104].

### 2.4.5 Free Energy Calculations: Umbrella Sampling

Under steered molecular dynamics, there can be some instances when the biomolecule of interest may get trapped in a local free energy minima due to the presence of a high energy barrier, violating the Ergodic hypothesis. Hence,

under these conditions the umbrella sampling technique is used which a type of enhanced sampling method. The technique of umbrella sampling was developed by two scientists, Torrie and Valleau in 1977[105]. In this method, a bias potential is used to increase the probability of the molecule to visit the unexplored minima at the other end of the high energy barrier. After sampling the whole phase space, the effect of this bias potential is then finally removed.

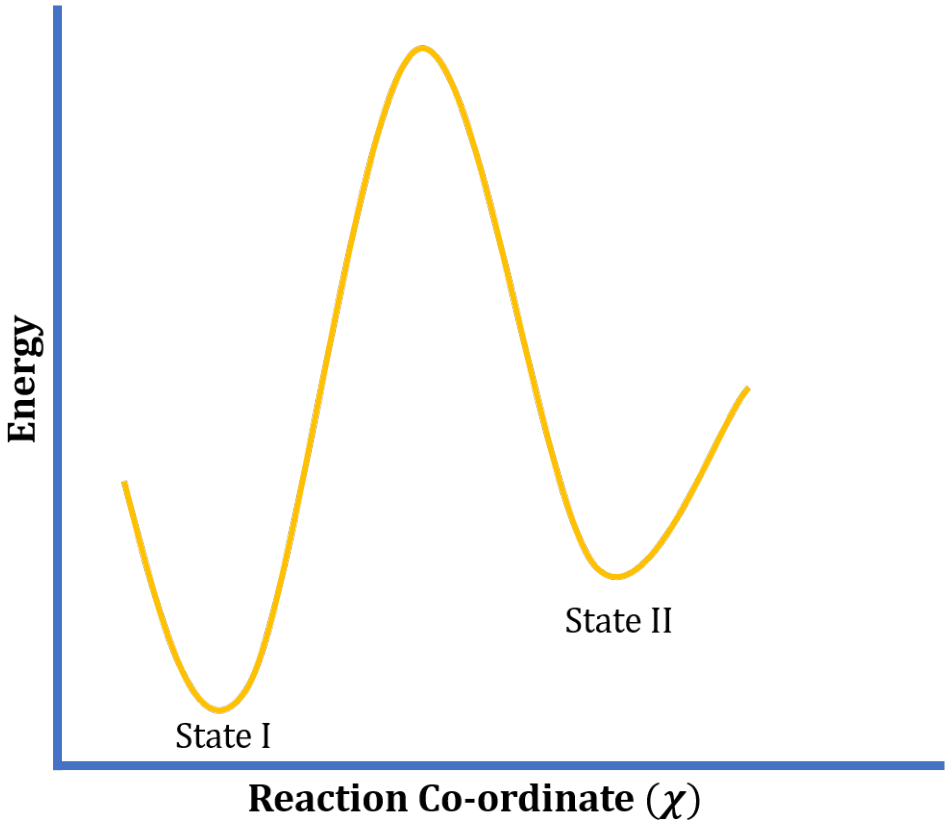



Figure 2.13: High Activation Energy Barrier separating State I and State II; adapted from Ref. [106]

As shown in the Figure 2.13, the states I and II are separated by a very high energy barrier. Therefore, a biomolecule trapped at state I has a very low chance of visiting state II in a real time frame and vise-versa. Therefore, a bias potential is as follows:

$$W(\chi) = K(\chi - \chi_0)^2 \tag{2.54}$$

Here K is the spring constant of the bias potential and $\xi$ is the reaction coordinate

of the system.

Therefore, the effective potential is:

$$V_0(\chi) = V(\chi) - W(\chi) \tag{2.55}$$

We must now recover the probability distribution for the unbiased trajectories after sampling the barrier top. This is done using a simple mathematical trick as written below.

$$\begin{aligned}
\langle O \rangle_{V_0} &= \frac{\int O(\chi) \exp(-\beta V_0(\chi)) d\chi}{\int \exp(-\beta V_0(\chi)) d\chi} \\
&= \frac{\int O(\chi) \exp(-\beta (V(\chi) + W)) d\chi}{\int \exp(-\beta (V(\chi) + W)) d\chi} \\
&= \frac{\int O(\chi) \exp(-\beta (V(\chi)) \exp(\beta W)) d\chi}{\int \exp(-\beta (V(\chi)) \exp(\beta W)) d\chi} \\
&= \frac{\langle O(\chi) \exp(\beta W) \rangle_V}{\langle \exp(\beta W) \rangle_V}
\end{aligned} \tag{2.56}$$

Using the above mathematical equations, we can adjust the bias potential (W) to achieve sufficient sampling in the desired region of the phase space, and then remove the bias potential to recover the unbiased ensemble average. Because we are using a harmonic potential that resembles an umbrella to constrain the system to a specific region of phase space in this case, this method is known as the umbrella sampling method [84, 102, 106].

However, umbrella sampling is computationally expensive. This is because the method involves creating multiple windows which also should have significant overlap to minimise errors, as errors from each window adds quadratically.

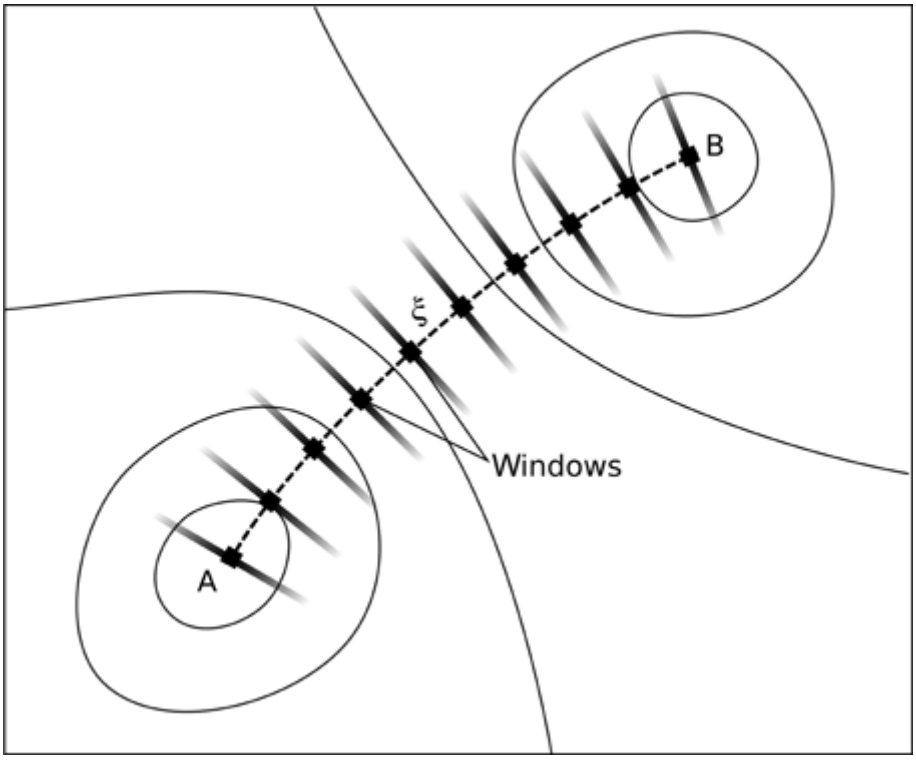



Figure 2.14: Reaction coordinates between two states divided into into distinct windows; adapted from Ref. [106]

However, the data from the windows is difficult to analyse. Therefore, a method called the Weighted Histogram Analysis Method [84, 106] is used to combine different simulations with different biasing potentials to generate the combined Potential Mean Force (PMF). The PMF provides with important insights as it gives the gives the free energy barrier separating different states. Hence, the information about the relative stability of different states present, can be inferred. The method along with other reaction coordinates can also be extended to multiple reaction coordinates as well.

### 2.4.6 Free Energy Calculations: Metadynamics

Metadynamics is another enhanced sampling method where the rare events beyond high energy barriers can be explored, implying ergodicity, and hence the free energy of the system can be estimated[107]. The process is well-known

as “filling the free energy with computational sand”. Under this algorithm, the assumption is made that the free energy of the system can be described by collective variables (CVs)[107]. Hence, during the simulation with metadynamics, more and more Gaussian hills are added as time progresses and the system is prevented to go back until all the system explores the complete energy landscape and starts making random walks.

Let the Hamiltonian of the system with the bias potential, $V_{bias}$, be: $H = K + V + V_{bias}$ where $V_{bias}$ is a function of CVs. Now, we will start updating the bias potential with the bias rate $\omega$ and $s_t$ is an instantaneous collective variable value at t. Thus,

$$\frac{\partial V_{bias}(s)}{\partial t} = \omega\delta(|s - s_t|) \tag{2.57}$$

which implies,

$$V_{bias} = \int_0^{t_{sim}} \omega\delta(|s - s_t|)dt \tag{2.58}$$

For computer simulations, the time t is discretized into $\tau$ intervals and the $\delta$ is replaced by a multidimensional positive Gaussian kernel function. This makes the equation as:

$$\begin{aligned} V_{bias} &\approx \tau \sum_{j=0}^{\frac{t_{sim}}{\tau}} \omega K(|s - s_j|) \\ &\approx \tau \sum_{j=0}^{\frac{t_{sim}}{\tau}} \omega \exp\left(-\frac{1}{2}\left|\frac{s - s_j}{\sigma}\right|^2\right) \end{aligned} \tag{2.59}$$

Metadynamics can be classified into two different categories based on deposited Gaussian heights: Standard Metadynamics and Well-tempered metadynamics[108].

#### 2.4.6.1 Standard Metadynamics

In standard metadynamics, the height of the deposited Gaussian kernels stay fixed through the simulation. The height of the Gaussian kernels ($W$) can be given by the product of the bias potential rate ($\omega$) and the Gaussian deposition stride ($\tau$) i.e., the time intervals. Therefore,

$$V_{bias} \approx \sum W(\omega\tau) \exp\left(-\sum_{i=1}^{d} \frac{1}{2} \left| \frac{s_i - s_i(q(\omega\tau))}{\sigma_i} \right|^2 \right) \tag{2.60}$$

As per the assumption of metadynamics, in the long-time limit, the bias potential converges to minus the free energy as a function of the CVs. Hence,

$$V_{bias}(s, t \to \infty) = -F(s) \tag{2.61}$$

where the free energy is defined as:

$$F(s) = -\frac{1}{\beta} \ln\left(\int dq\, \delta(s - s(q)) e^{-\beta U(q)}\right) \tag{2.62}$$

here $\beta = 1/k_B T$ and U(q) is the potential energy function.
However, standard metadynamics has its own sets of limitations, which include:

- The bias potential overfills the underlying Free Energy Surface and pushes the system toward high-energy regions of the CVs space, which makes it non-trivial to decide when to stop a simulation.
- Identifying a set of CVs appropriate for describing complex processes is far from trivial.

#### 2.4.6.2 Well-tempered Metadynamics

Therefore, to address the first limitation of the standard metadynamics, the concept of well-tempered metadynamics comes into the picture[109].

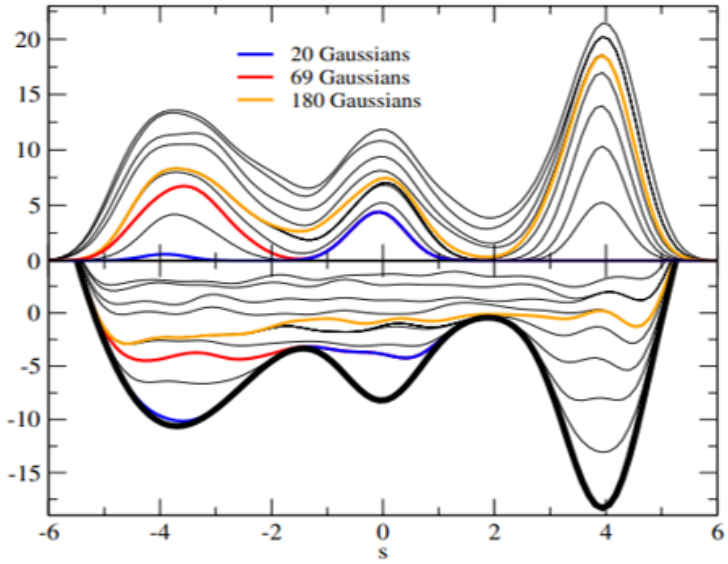



Figure 2.15: Well-tempered metadynamics simulation showing decreasing Gaussian height with time; adapted from Ref. [110]

Here, the bias deposition decreases with time. The new W therefore becomes,

$$W(\omega\tau) = W_0 \exp\left(-\frac{V_{bias}(s(q(\omega\tau)),\omega\tau)}{k_B\Delta T}\right) \tag{2.63}$$

This makes the bias potential to be

$$V_{bias} = \sum W_0 e^{-\frac{V_{bias}(s(q(\omega\tau)),\omega\tau)}{k_B\Delta T}} \exp\left(-\sum_{i=1}^{d}\frac{1}{2}\left|\frac{s_i - s_i(q(\omega\tau))}{\sigma_i}\right|^2\right) \tag{2.64}$$

However, this leads to non-compensation in the underlying free energy which is now given as,

$$V_{bias}(s, t\to\infty) = -\frac{\Delta T}{T+\Delta T}F(s) \tag{2.65}$$

From the Eqn. 2.65, it is evident that as the value of $\Delta T$ approches

$\Delta T \to 0 \quad \Rightarrow$ Standard Molecular Dynamics

$\Delta T \to \infty \Rightarrow$ Standard Metadynamics

This factor is defined as the bias factor in well-tempered metadynamics and is denoted by the symbol $\gamma$. Therefore,

$$\gamma = 1 + \frac{\Delta T}{T} \Rightarrow \frac{1-\gamma}{\gamma} = \frac{\Delta T}{T + \Delta T} \tag{2.66}$$

This leads to the bias potential, at long time, becoming,

$$V_{bias}(s, t \to \infty) = -\frac{1-\gamma}{\gamma} F(s) \tag{2.67}$$

## 2.5 Computational Methods for Analysis

### 2.5.1 Root-Mean Square Distance (RMSD) Analysis

RMSD is mainly useful in the quantifying how much a configuration of protein or segment has changed from its native state. This helps to analyse if a protein is in a different configurational space by undergoing a transition. It is calculated by averaging over particle coordinate giving time-specific values. For biomolecular systems, the RMSD is normalised over masses. The expression is given as:

$$RMSD(t) = \sqrt{\frac{1}{M} \sum_{i=1}^{N} m_i |x_i(t) - x_i(0)|^2} \tag{2.68}$$

### 2.5.2 Root-Mean Square Fluctuation (RMSF) Analysis

RMSF is mainly useful in the quantifying how much a protein segment is fluctuating through time. This particularly helps in differentiating between structured and unstructured regions in a protein of interest. It is calculated by averaging over time coordinate giving particle-specific values. The expression is given

as:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{\tau=1}^{T}|x_i(\tau) - \bar{x}_i|^2} \tag{2.69}$$

### 2.5.3 Theory of Correlation Analysis

#### 2.5.3.1 Gaussian Network Model

Gaussian Network Model is used to study the fluctuation and correlated motions of atoms. In this model, the $\alpha$-carbons of the amino acids of the proteins are identified as nodes, and all nodes are connected by springs within an interaction range generally with a cutoff distance ($r_c$) of 0.7Å.

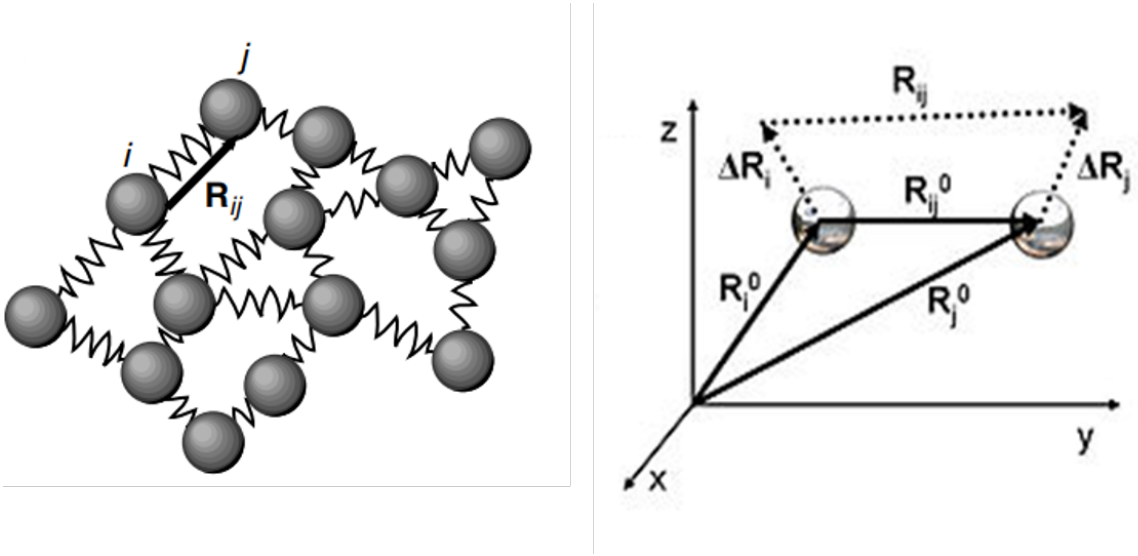



Figure 2.16: Nodes in the GNM model connected with springs; adapted from Ref. [111]

Here, we define the:

$$\Delta R_i = R_i - R_i^0 \text{ and } \Delta R_j = R_j - R_j^0 \tag{2.70}$$

Therefore,

$$\Delta R_{ij} = \Delta R_j - \Delta R_i \tag{2.71}$$

The potential energy of the model can be written as:

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Gamma_{ij}(\Delta R_i - \Delta R_j)^2\right] = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Gamma_{ij}\left[(\Delta X_i - \Delta X_j)^2 + (\Delta Y_i - \Delta Y_j)^2 + (\Delta Z_i - \Delta Z_j)^2\right]\right] \tag{2.72}$$

Here $\gamma$ is the spring constant and $\Gamma_{ij}$ is the ij*th* element of Kirchhoff's matrix of inter-residue contact, $\Gamma$, defined by:

$$\Gamma_{ij} = \begin{cases} -1, & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0, & \text{if } i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{i \neq j}\Gamma_{ij}, & \text{if } i = j \end{cases} \tag{2.73}$$

#### 2.5.3.2 Covariance Matrix

The general assumption of the GNM is that all fluctuations are isotropic and Gaussian in nature. After further derivations, it can be shown that the covariance matrix is a combination of expectation values of residue fluctuations and cross-correlations in the diagonal and off-diagonal elements, respectively. The covariance matrix (for X) is related to Kirchhoff's matrix as follows:

$$\Xi = \frac{k_B T}{\gamma}\Gamma^{-1} \tag{2.74}$$

Similarly, it can be written for Y and Z. Therefore, the residue fluctuations and cross-correlations can be expressed as follows:

$$\langle \Delta R_i^2 \rangle = \frac{3k_B T}{\gamma}(\Gamma^{-1})_{ii} \tag{2.75}$$

$$\langle \Delta R_i . \Delta R_j \rangle = \frac{3k_B T}{\gamma} (\Gamma^{-1})_{ij} \tag{2.76}$$

Therefore, the covariance matrix (for X) is as follows:

$$\Xi = \begin{bmatrix} \langle \Delta x_1^2 \rangle & \langle (\Delta x_1)(\Delta x_2) \rangle & \dots & \langle (\Delta x_1)(\Delta x_n) \rangle \\ \langle (\Delta x_2)(\Delta x_1) \rangle & \langle \Delta x_2^2 \rangle & \dots & \langle (\Delta x_2)(\Delta x_n) \rangle \\ \vdots & & \ddots & \\ \langle (\Delta x_n)(\Delta x_1) \rangle & \langle (\Delta x_n)(\Delta x_2) \rangle & \dots & \langle \Delta x_n^2 \rangle \end{bmatrix} \tag{2.77}$$

Here the cross-correlation terms are given by the following expression:

$$\langle (\Delta x_i)(\Delta x_j) \rangle = \langle (\Delta x_i^0 - (\langle x_i \rangle))(\Delta x_j^0 - (\langle x_j \rangle)) \rangle \tag{2.78}$$

#### 2.5.3.3 Correlation Matrix

Therefore, it can be summarised as:

$$\langle \Delta R_i . \Delta R_j \rangle = \langle \Delta x_i . \Delta x_j \rangle + \langle \Delta y_i . \Delta y_j \rangle + \langle \Delta z_i . \Delta z_j \rangle \tag{2.79}$$

$$\langle \Delta R_i . \Delta R_i \rangle = \langle \Delta x_i^2 \rangle + \langle \Delta y_i^2 \rangle + \langle \Delta z_i^2 \rangle \tag{2.80}$$

Now, the fluctuational cross-correlation matrix then was calculated as follows:

$$C(i,j) = \frac{\langle \Delta R_i . \Delta R_j \rangle}{\sqrt{\langle \Delta R_i . \Delta R_i \rangle \langle \Delta R_j . \Delta R_j \rangle}} \tag{2.81}$$

# 3 Drug-induced conformational dynamics of oncogenic KRas: Comparing the effects of AMG-510 & MRTX-849

## 3.1 Introduction

### 3.1.1 KRas: A MAPK signalling protein & its cellular functioning

The MAPK (Mitogen Activated protein Kinase) pathway or the ERK (Extracellular Signal-Regulated Kinase) pathway are a relay of proteins in the cell that communicate a signal to the cellular DNA from the extracellular receptor. It mainly consists of proteins that are involved in the phosphorylation of downstream proteins to make them "active" or "inactive" by acting as molecular switches.

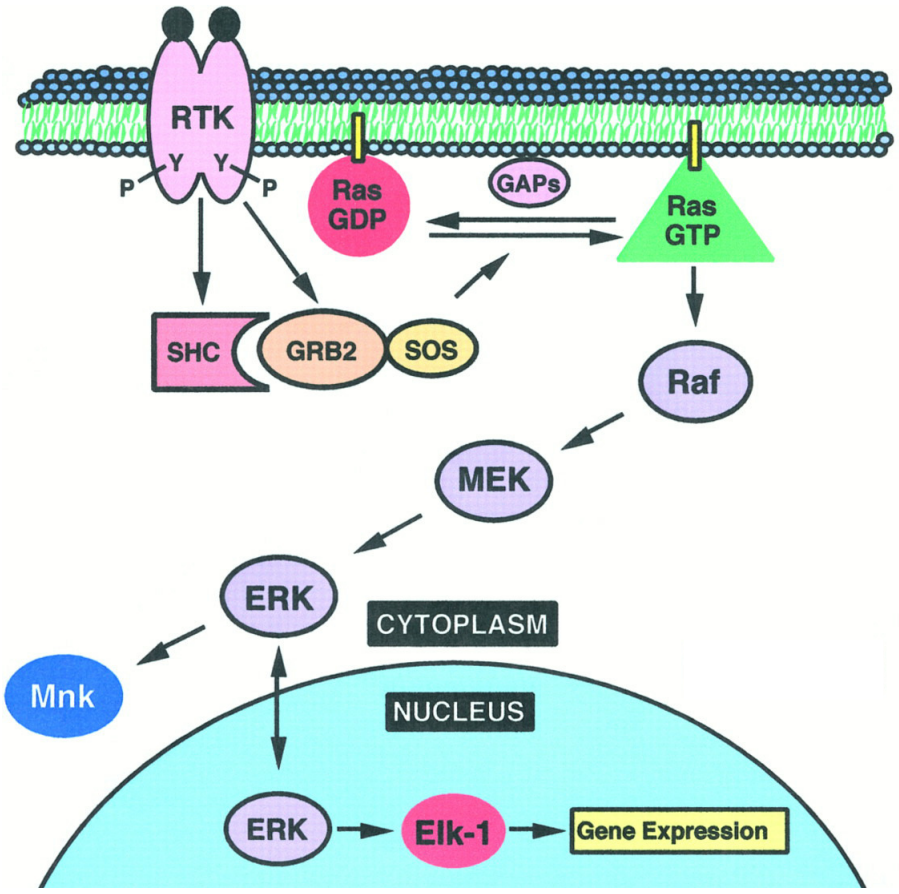



Figure 3.1: Schematic of MAPK/ERK cellular signalling pathway; adapted from Ref. [112]

The Ras-family proteins acts as a crucial relay in this chain. And KRas (Kirsten Rat sarcoma) is one of them [113]. It generally stays in an "Inactive" conformation when bound to GDP. When an upstream signal is intercepted, the SOS-family (Son of Sevenless) of proteins, which includes GEFs (Guanine Exchange Factors) catalyses the GDP-to-GTP exchange in the KRas, switching the latter from an inactive state to an active state. In this state, it is capable of activating downstream targets by phosphorylating them. Once it does the phosphorylation, under normal physiological conditions, it being a GTPase, hydrolyses the GTP to GDP with the help of a protein called GAP (GTPase Activating Protein), which catalyses the process. The schematic of the pathway has been shown in the Figure 3.1.

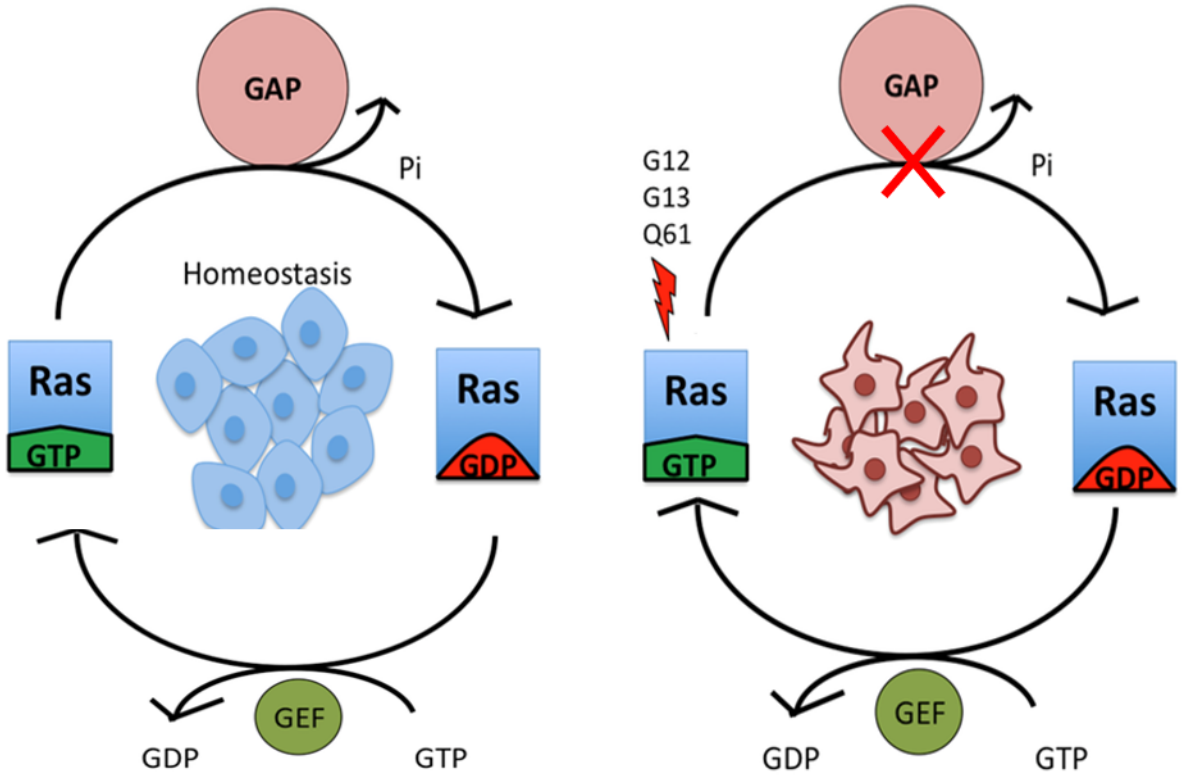



Figure 3.2: KRas function under physiological and mutated states; adapted from Ref. [114]

The schematic structure of KRas is as given below in the Figure 3.3. It has three important regions which act as the active site of the protein: the P-loop,

Switch-I and Switch-II, among which the last two act as Intrinsically Disordered Regions. There are a total of 5 $\alpha$-helices, 5 $\beta$-sheets and multiple loops forming a globular kind of protein structure. The HVR (Hyper-variable Region) region beyond 169 amino acid has been deleted for illustrative purposes.

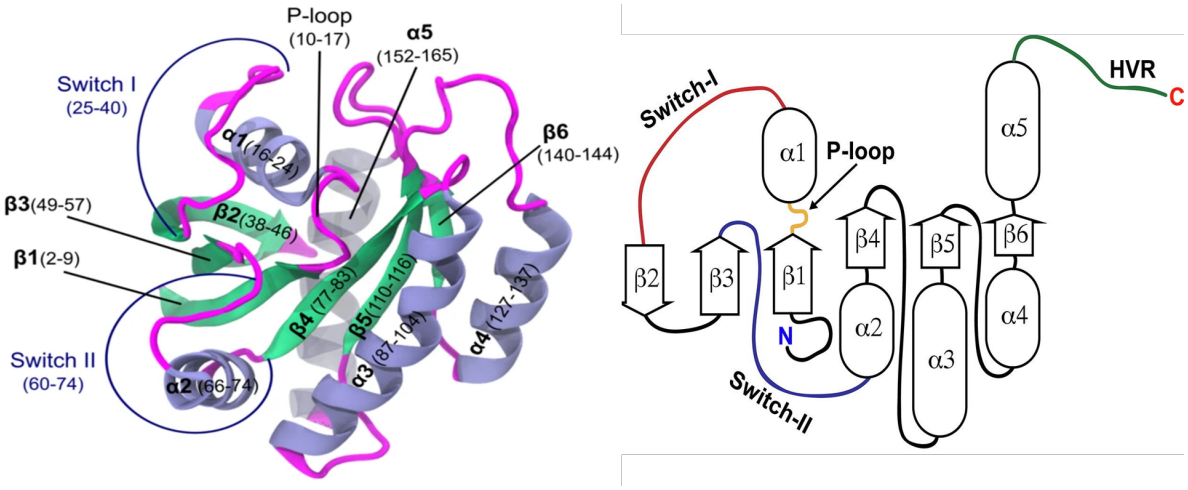



Figure 3.3: KRas Structure and schematic representation of helices and sheets; adapted from Ref. [115]

However, mutations can occur in the KRas at some of the potential sites, which can cause the protein to be oncogenic. Potential mutation sites are G12, G13, and Q61. These sites are essential in the functioning of KRas, and hence their mutation leads to abnormalities in the MAPK pathway. These mutations essentially lead to the loss in the GTPase activity of the KRas leading to cancer, predominantly seen in the lung cells. Therefore, the KRas always stays in the active state and promotes cell growth and proliferation. The schematic of the abnormal mechanism is shown in Figure 3.2.

### 3.1.2 KRas in Lung Cancer

Lung cancer is the most frequently diagnosed cancer and a leading cause of cancer-related death worldwide making up almost 25% of all cancer deaths [116]. Non-small-cell lung cancer (NSCLC) is the most commonly diagnosed

form of the disease, accounting for >85% of the total cases [117]. Squamous cell carcinoma and adenocarcinoma are examples of non-small-cell lung tumours that act similarly. Mutation in the protein involved in the MAPK signalling pathway, i.e. KRas, is the leading cause of NSCLC.

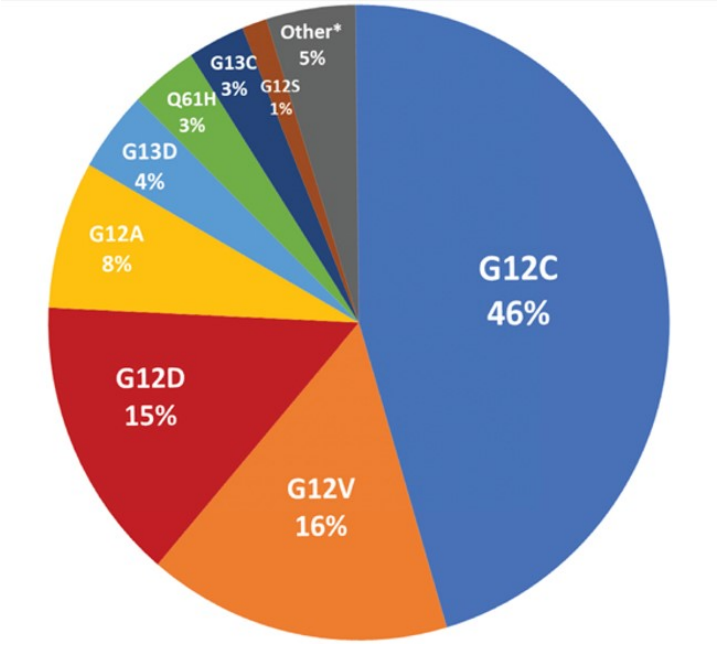



Figure 3.4: Pie chart showing mutational distribution for KRas malignancy in NSCLC

### 3.1.3 Most fatal G12C mutation and its drug induced inhibition

The major codon affected in these NSCLC cells is codon 12. The remaining being 13 and 61 [118]. And among the codon 12, G12C is the most common mutation accounting to around 46% as shown in Figure 3.4 [119]. The other major ones include G12V, G12D, and G12A. Recently, for the G12C-mutated KRas two drugs named AMG-510 and MRTX-849, were designed by Amgen and Mirati Therapeutics, respectively. The latter was recently approved by the FDA [120]. One aspect of this project mainly involves investigating the G12C-mutated KRas and its interaction in the presence of those two drugs: Adagrasib (MRTX-849) and Sotorasib (AMG-510). These two drugs covalently bind to the mutated site by forming a C–S bond and inhibit the binding of GTP to KRas,

leading the KRas to remain inactive for an indefinite period of time. The two drugs are very similar in structure. However, it has been shown in the literature that the former has an overall potency ($K_{inact}/K_I$) of $35\,mM^{-1}s^{-1}$, whereas the later has a potency of $9.9\,mM^{-1}s^{-1}$ [121, 122]. Here, $K_{inact}$ represents the rate of inactivation and $K_I$ represents the reversible affinity. Therefore, it was of interest to look into the mechanism of the drug binding and the differential inhibition caused by the two drugs at the molecular scale.

## 3.2 Methodology: Details of atomistic simulation methods and IDP specific force fields

### 3.2.1 System Preparation

#### 3.2.1.1 Mutated protein in no-drug state

The original KRas protein configuration was taken from the Protein Data Bank (PDB ID: 4OBE) from an X-Ray Diffraction experiment by Hunter et al.[123]. Subsequently, the glycine at position 12 was altered. Following that for the simulation of the G12C variant with GDP attached, PyMol's mutagenesis tool was used to convert glycine to cysteine, in-silico. The structure was homology modelled with the SWISS-MODEL server to account for missing regions[124]. The structure of the G12C-mutated protein is as shown in Figure 3.5.

#### 3.2.1.2 Mutated protein in drug-bound state: Specific interaction with AMG-510 & MRTX-849

X-Ray Diffraction structures from the protein data bank PDB ID: 6OIM[121] and PDB ID: 6UT0[122] were utilised to investigate the mutated KRas structure with the bound drugs Sotorasib and Adagrasib. Since, chemically both the drugs are covalently bound to the mutated site and the force field doesn't account for

this new bond, the simulation parameters were missing for this residue. Therefore, the bond was hypothetically created using a distance-constrained spring of force constant $k = 10,000\, kJ\, mol^{-1}\, nm^{-2}$. Also, because drug molecules were not contained in the initial force field of CHARMM36IDPSFF, the SWISS-PARAM module was used to determine their parameters. Both structures were homology modelled with the SWISS-MODEL server to account for missing regions[124]. The structures of the G12C-mutated drug-bound proteins are as shown in Figure 3.6 and Figure 3.7.

### 3.2.2 Hybrid protein specific force field: CHARMM36IDPSFF

The force field used for explicit solvent simulation of KRas with its ligands and drugs was CHARMM36IDPSFF, as this force field was specifically designed to simulate intracellular disordered proteins (IDP) or their regions (IDRs) [125].

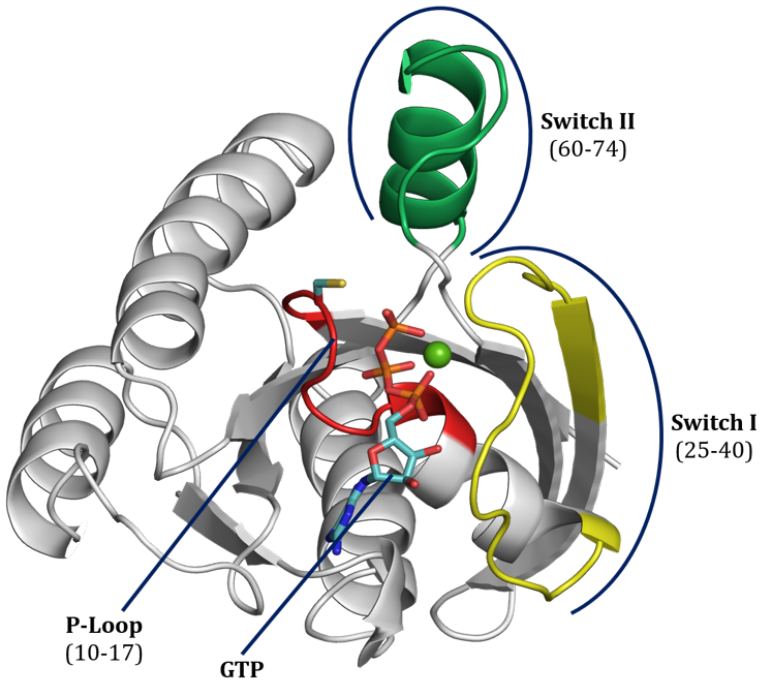



Figure 3.5: Mutated KRas at position 12 from Glycine to Cysteine

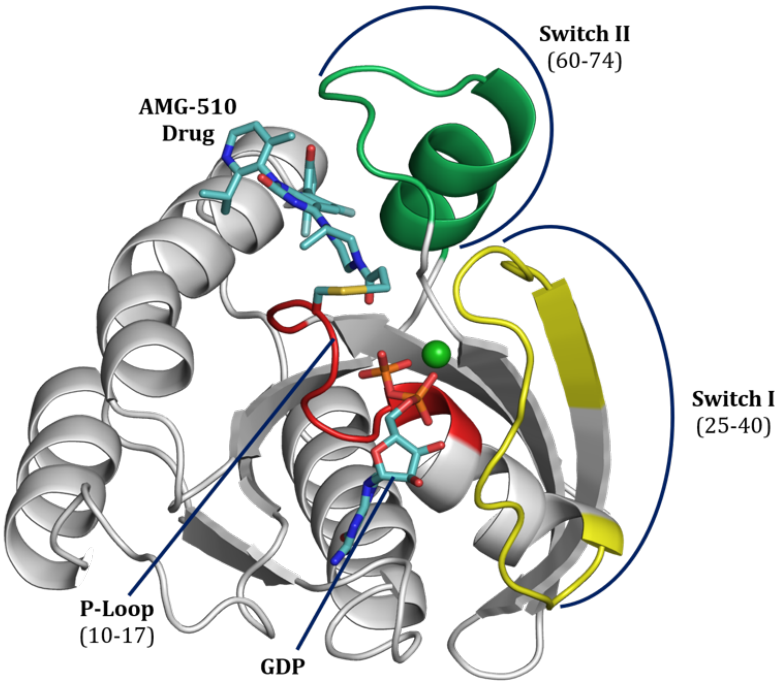



Figure 3.6: Mutated KRas covalently attached to AMG-510 at position 12

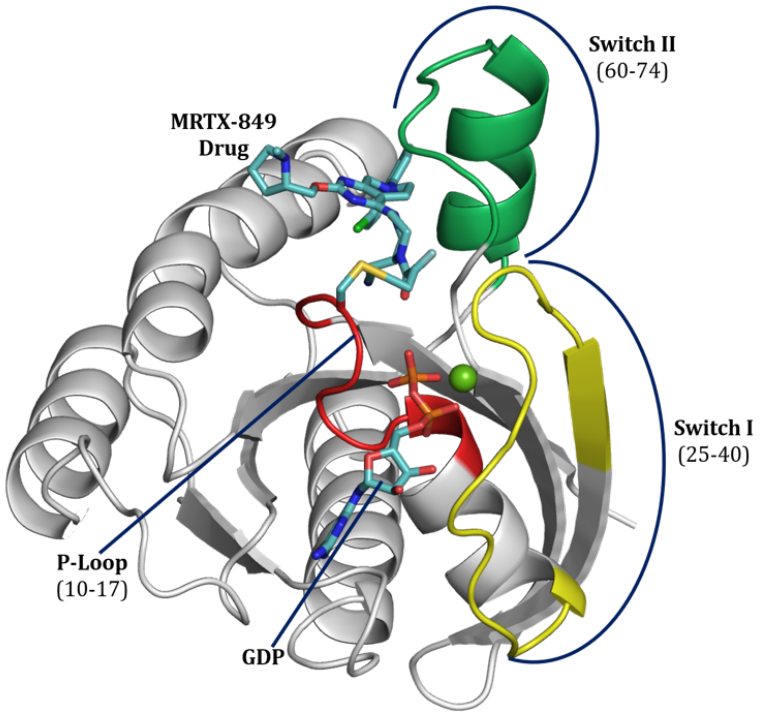



Figure 3.7: Mutated KRas covalently attached to MRTX-849 at position 12

This force field was improvised from the previously established force field CHARMM36m (C36m), which was itself an improved version of CHARMM36 (C36) [126]. The C36m force field yielded a high-energy barrier in the back-

bone dihedrals between the poly-proline II region and the helix region in the Ramachandran plot. Therefore, with modified Grid-based energy correction map (CMAP) parameters for all the 20 naturally occurring amino acids, the CHARMM36IDPSFF force field was developed. It is to be noted that, CMAP method was first used in CHARMM22 to account for improved sampling of backbone dihedrals [127]. Since, CHARMM36IDPSFF accounted for the backbone dihedrals better than the C36m, it was the go to force field for the hybrid protein system.

### 3.2.3 Atomistic simulation methods

GROMACS software was used to run the three simulations, and the topologies were created using the CHARMM36IDPSFF force field, which is specifically built for proteins having Intrinsically Disordered Regions[125]. After that, each of them were centered in a dodecahedral box and solvated using the TIP3P water model[128]. To mimic the normal physiological environment, the systems were then neutralised with sodium and chloride ions.

After the systems were prepared, they were subjected to energy minimisation using the steepest descent algorithm to remove steric clashes. The protein and bound ligands were first position constrained with a force constant of 1000 $kcal\,mol^{-2}\,nm^{-2}$ and the solvent was equilibrated. At this stage, the systems were allowed to go through an NVT equilibration at 300 K using the modified Berendsen thermostat[94] for 6 ns. The Parrinello-Rahman barostat[97] was then used to maintain an average pressure of 1 bar on all systems for 7ns. Throughout the simulations, a time step of 1fs was maintained and the leap-frog integrator was utilised. The position restrictions were eliminated in the final simulation and the simulations were run for 1$\mu$s each using the NPT pa-

rameters. Particle Mesh Ewald was used for electrostatic calculations, with a cubic interpolation of order 4 and a grid spacing of 0.16 for the Fast Fourier Transform. Periodic boundary conditions were used throughout all the simulations in all directions. After every 10 steps, the neighbour list was updated using a grid method, with a short-range neighbour list cut-off of 1 nm. LINCS constraints were applied to all of the bonds.

## 3.3 Results & Analysis: Comparison of conformational dynamics among the G12C variants, and the AMG and MRTX drug-bound forms

### 3.3.1 Finding fluctuating motifs from RMSF analysis

To better capture the fluctuative behaviour as proposed in the literature about Switch-I and Switch-II, the Root Mean Square Fluctuation of the three systems was calculated. Also, it proved as a validation of the force field we had chosen, if it was able to capture the disorderdness of the IDRs. The RMSF for all the systems were analysed as shown below. As can be seen, the RMSF shows that the region of Switch-I and Switch-II are having very high fluctuations as compared to all other segments in the protein for all the systems. The other conclusions that can be drawn from this Figure 3.8 are:

- For the GDP G12C variant, both the switch regions show very high fluctuation in the regime of 0.4-0.6nm.
- For GDP and AMG bound protein, the fluctuation is reduced only in the Switch-I region.
- The GDP and MRTX bound protein shows reduced fluctuations in the both the switch regions.

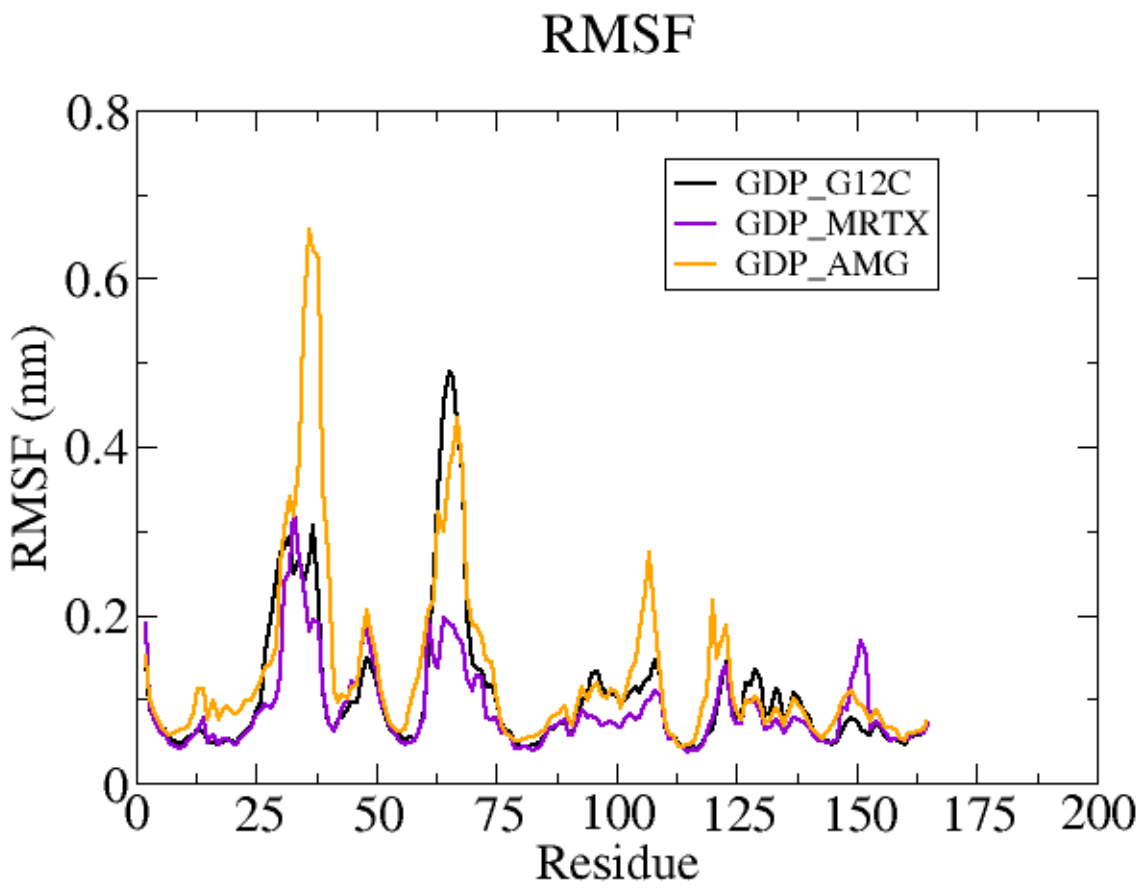



Figure 3.8: RMSF plot of GDP bound G12C variant, AMG and MRTX drug-bound

Hence, it conclusively shows that MRTX drug is significantly effective in reducing the fluctuation of the switch regions as compared to the GDP bound G12C variant of KRas.

### 3.3.2 Quantifying & comparing fluctuation of different IDRs in KRas from RMSD

Once the fluctuation was obtained, to understand their flexibility we have analysed and compared P-loop, Switch-I and Switch-II's Root Mean Square Deviation. The blue part indicated in the figure 3.9, represents the non-equilibrium part of the simulation and they have been ignored for all further analysis.

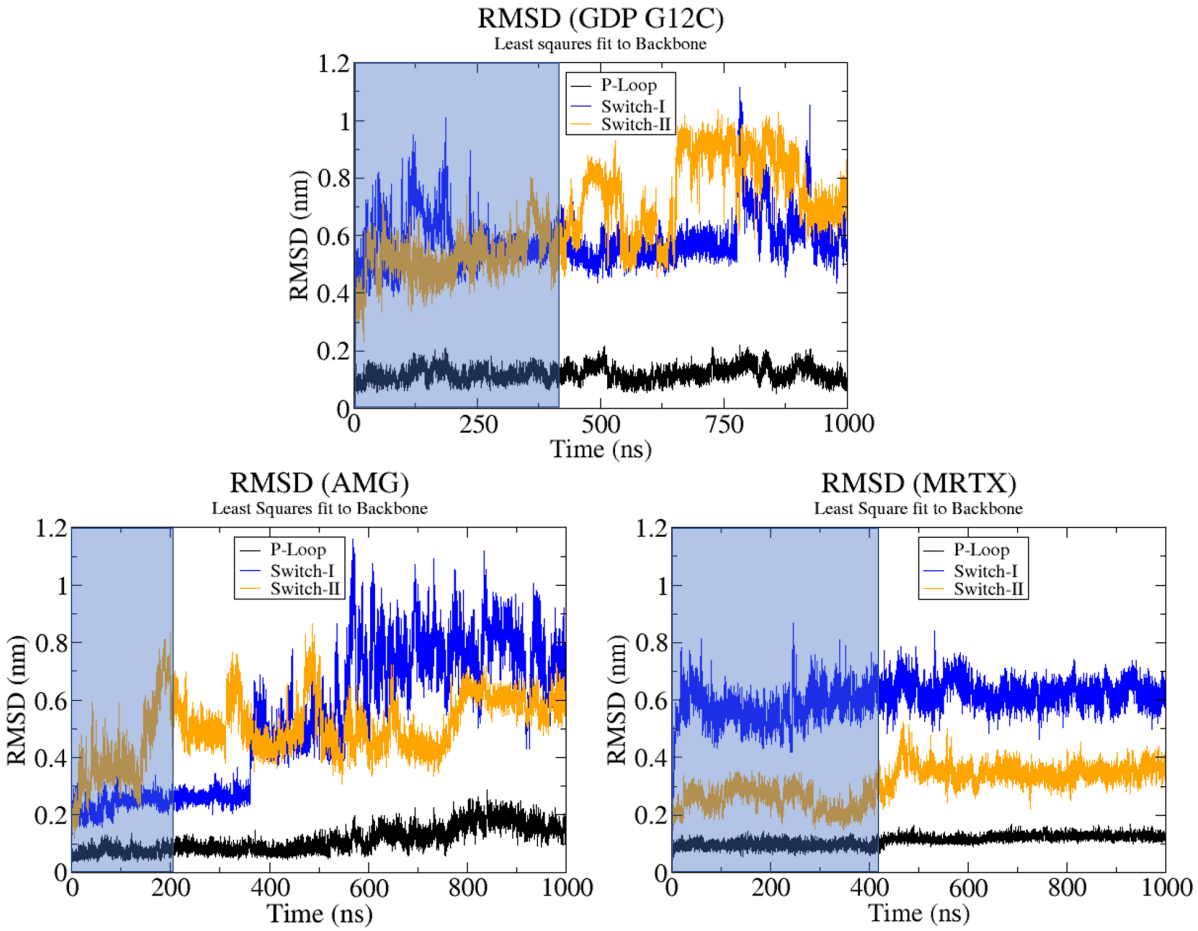



Figure 3.9: RMSD plots of GDP bound G12C variant, AMG and MRTX drug-bound

The RMSD for the P-loop, Switch-I and Switch-II of all the IDRs of the three systems were analysed as shown in the Figure 3.9. As is evident from the figure 3.9, the RMSD fluctuations for the Switch-II region of the MRTX-bound species is significantly lower than the RMSD fluctuation of the AMG-bound and no drug-bound G12C-mutated species. The Switch-I fluctuation is also restricted in one bound to MRTX as compared to the other two species.

### 3.3.3 Fluctuation-fluctuation correlation at residual level between Switch-I & Switch-II

Inferring from the upper sections that the Switch regions are highly flexible as compared to the other regions of the protein, it was of interest to see if the motions of the switch regions are correlated in some way. Therefore, the gmx

covar module of *GROMACS* was used to calculate the covariance matrix, which was indeed used to further calculate the correlation matrix using the theory from Section 2.5.3.

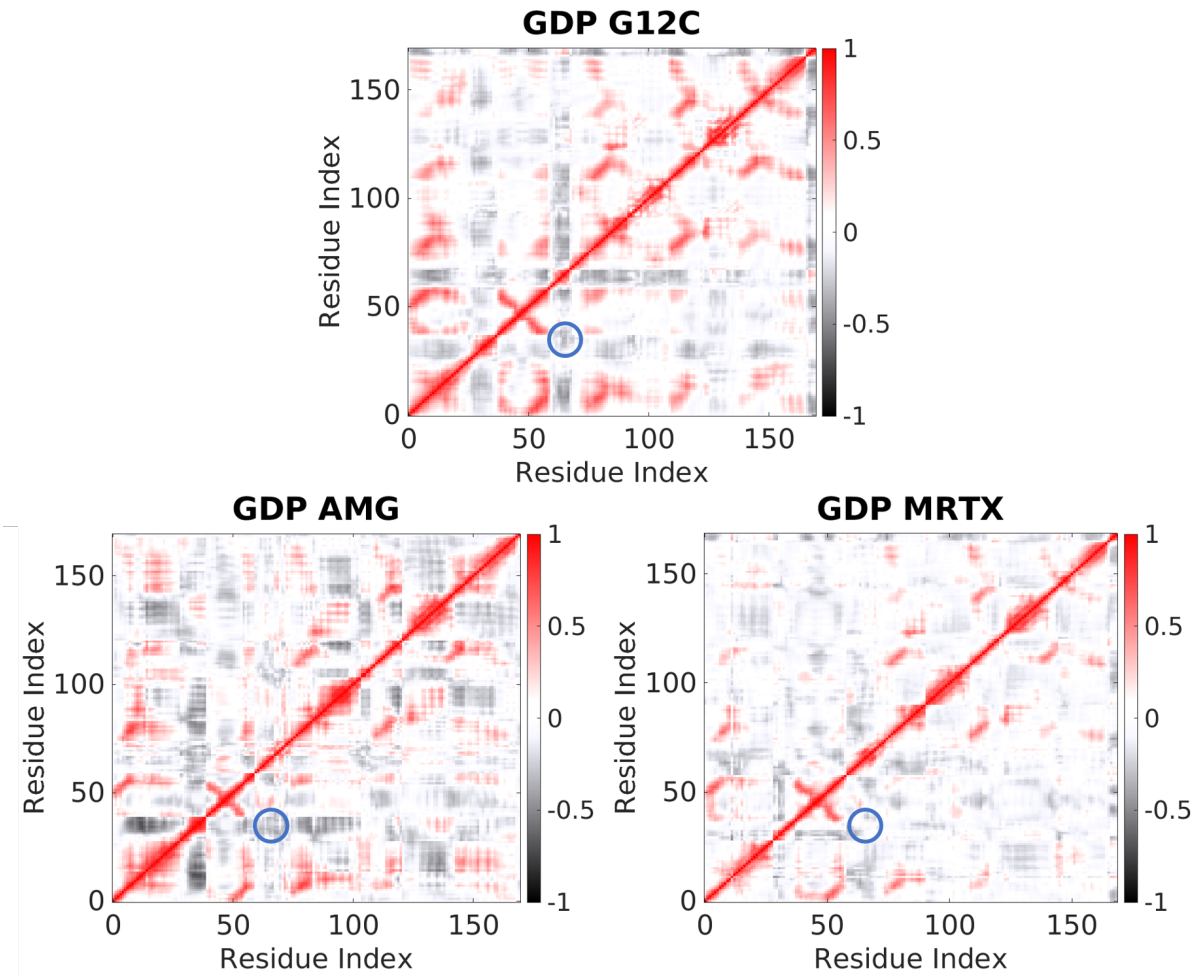



Figure 3.10: Correlation plots of GDP bound G12C variant, AMG and MRTX drug-bound

Therefore, the covariance matrix of the C$\alpha$ atoms was constructed and then the correlation matrix was calculated. The matrix were plotted as shown in the Figure 3.10.

Here in the Figure 3.10, the blue circles indicate the regions of Switch-I and Switch-II correlation. Though a prominent anti-correlation can be seen in the case of both GDP bound G12C variant and AMG-drug bound one, the correlation is seen to be completely diminished in the case of the MRTX drug bound one. Therefore, it can be concluded that MRTX not only reduces the fluctuation

of the Switch regions but also diminishes the correlated motion between the two Switch regions.

### 3.3.4 Structural investigation in the neighbourhood of switch regions–specifically focussing on $\alpha$-2 & $\alpha$-3

#### 3.3.4.1 Temporal Helicity comparison

The $\alpha$-2 helix in the KRas protein is a part of the Switch-II region as mentioned before. This is because, the helix has a very high propensity to fluctuate under normal conditions and continuously shifts its configuration between an alpha helix and loop. Therefore, a study was conducted to look upon if different systems we are studying have different helicity.

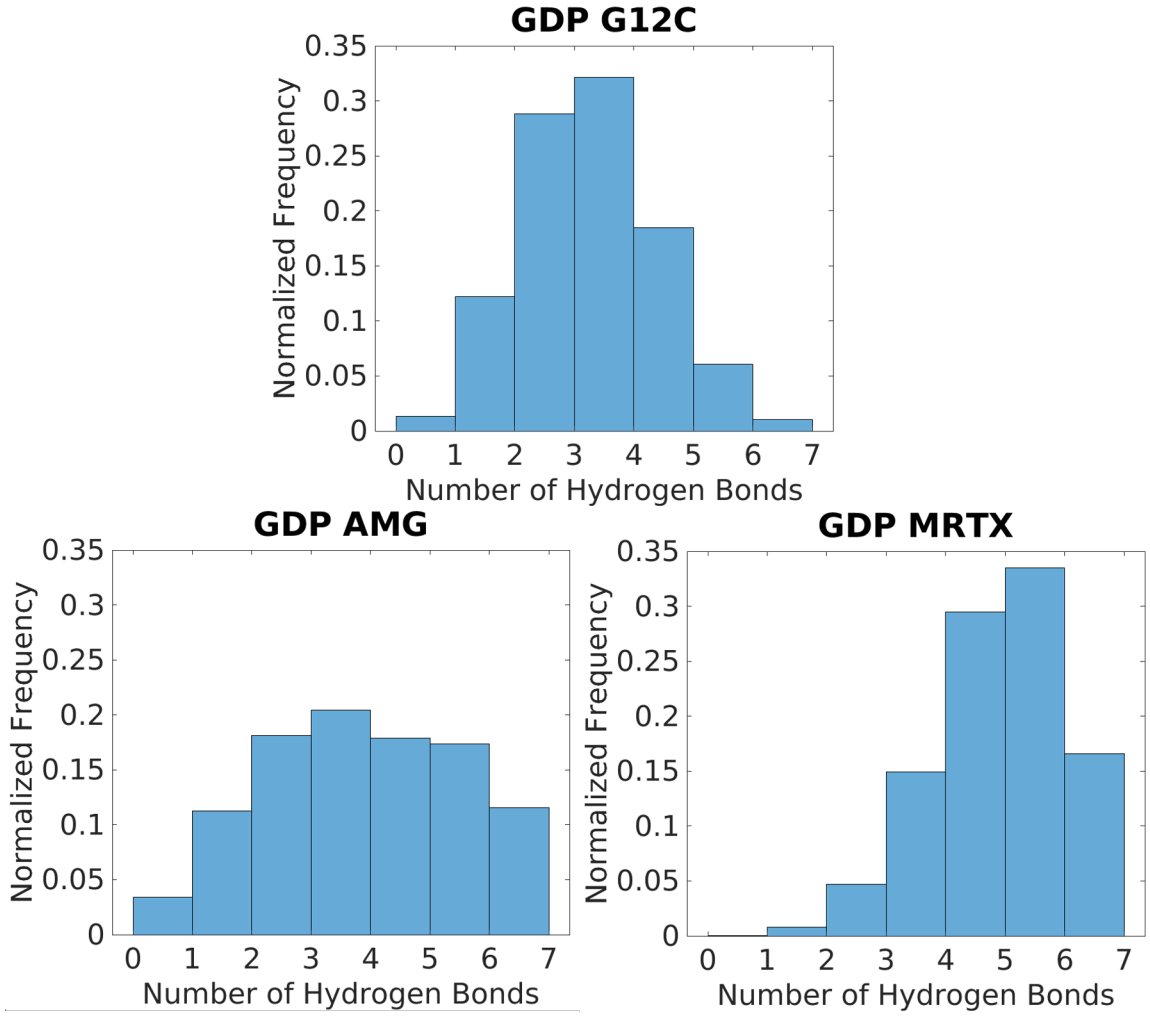



Figure 3.11: $\alpha$-2 helix melting histogram comparison

For this study, the gmx helix module of GROMACS was used to collect the time-dependent data and then a frequency-dependent histogram was plotted by normalizing all the plots with the global maxima of the frequency obtained form the three data sets. The data were obtained as shown in Figure 3.11.

The more the hydrogen bonds present better stabilized is the helix. As can seen in the Figure 3.11, the frequency-dependent normalised histogram for $\alpha$-2 helix for GDP-bound G12C variant and the AMG-drug bound protein mostly stays around the 3 to 4 hydrogen bonds during the simulation. However, for the MRTX drug bound one, the peak lies around 5 to 6 hydrogen bonds. Therefore, it can be concluded that Adagrasib successfully inhibits the melting of the helix as compared to the other two. And this might be one of the reason for which MRTX inhibits the fluctuation of the Switch-II region as shown in the RMSF Figure 3.8.

#### 3.3.4.2 Dihedral Analysis

The $\alpha$-2-$\alpha$-3 pocket is the main binding pocket of the drugs AMG and MRTX. Therefore, the dihedral angle formed these two loops were analysed so as to see if the drug-binding event has any effect on the bending of the two loops towards each other.

Therefore, the frequency-dependent histogram was plotted for all the three species as shown in the Figure 3.12. As can be seen in the Figure, both the G12C variant and AMG-bound KRas show bimodal distribution which ranges from $50^o$ to $148^o$, having peaks at around $80^o$ and $110^o$. Whereas for the MRTX-bound KRas, only a unimodal distribution is evident which is quite restricted between $70^o$ to $118^o$ peaking at around $96^o$. This also indicates that MRTX heavily restricts the fluctuation of the Switch-II's $\alpha$-2 helix and keeps it close

to the $\alpha$-3 helix, unlike the heavy fluctuation seen in case of G12C variant and AMG-bound KRas.

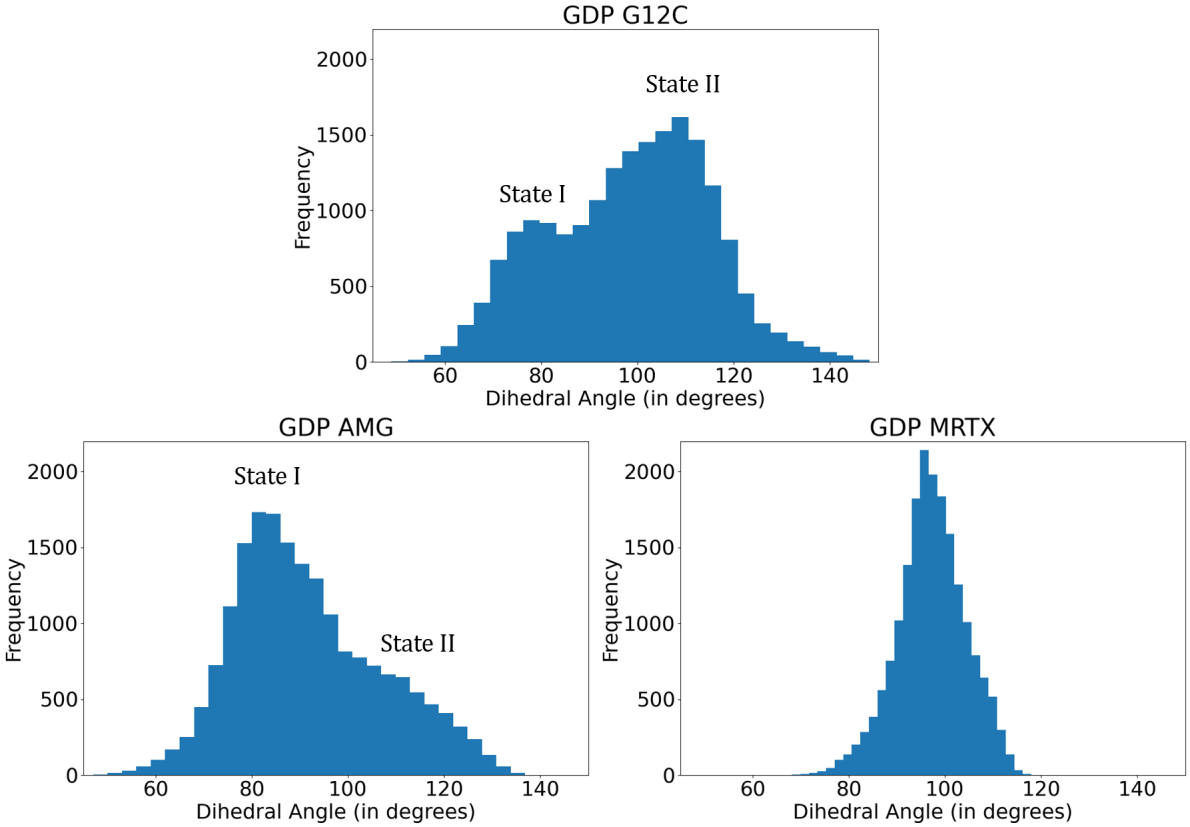



Figure 3.12: Frequency-dependent histograms of GDP bound G12C variant, AMG and MRTX drug-bound

### 3.3.5 Exploration of drug-mediated interaction through contact map analysis

Contact maps are a very good visualization tool for visualizing domain-domain interactions that are present in a biomolecular system. For this study, frequency dependent contact maps were used so as to analyse the major long-timescale interactions present between a protein segment and the associated ligand.

For our two drug systems, the drugs lie very close to the Switch-II and $\alpha$-3 regions. Therefore, the analysis was performed independently to all the combinations possible. As can be seen, in Figure 3.13 and Figure 3.14, the MRTX drug makes a significant number of more sustained contacts compared to AMG. The contacts formed are mainly forming through the hydrogen bonds and hy-

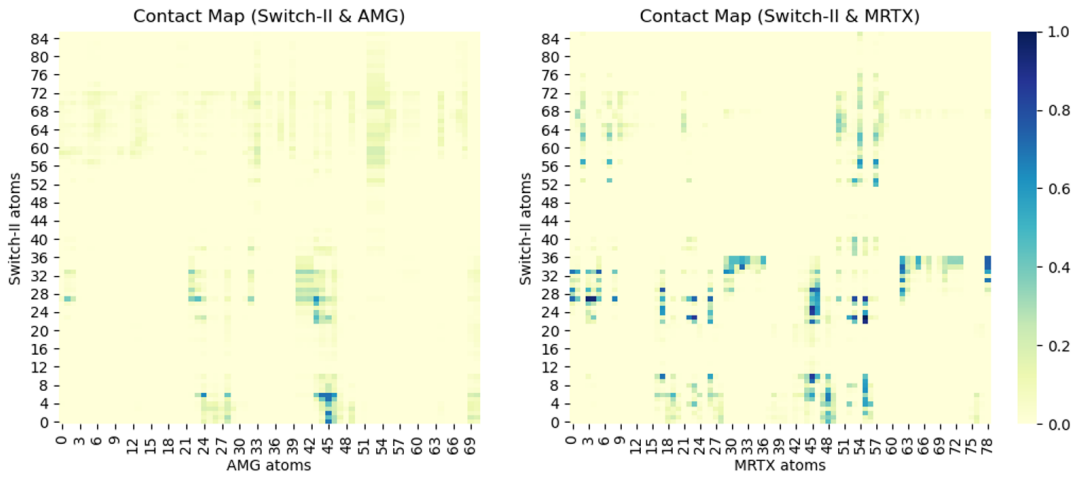



Figure 3.13: Contact map of Switch-II loop region with drugs

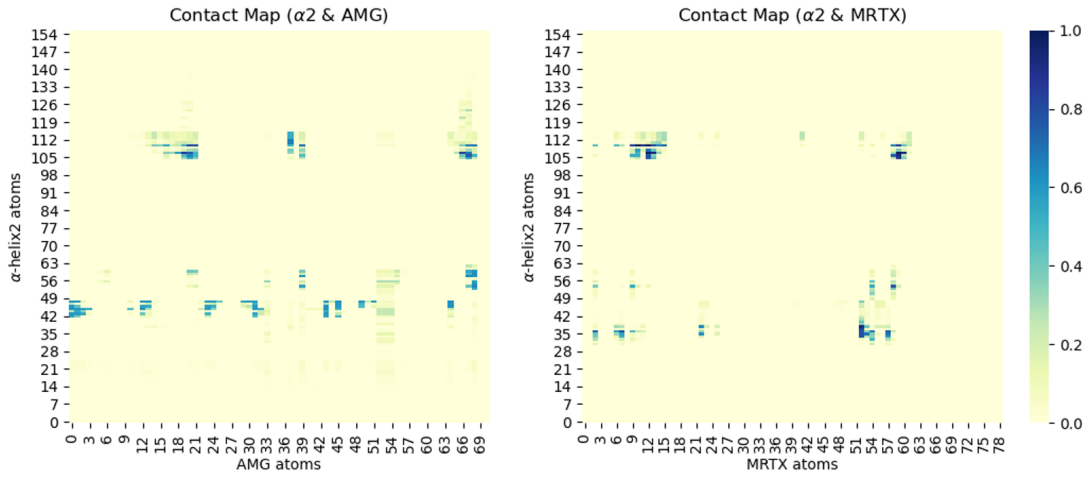



Figure 3.14: Contact map of Switch-II's $\alpha$-2 region with drugs

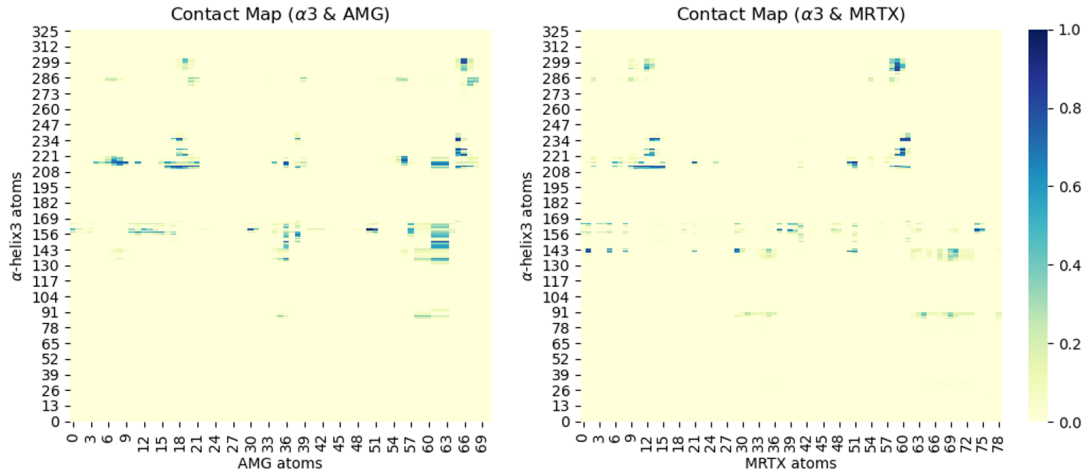



Figure 3.15: Contact map of $\alpha$-3 helix with drugs

drophobic contacts. Furthermore, it should be noted that there is no significant difference between the contacts formed between $\alpha$-3 and the drugs, as shown in Figure 3.15. This gives the conclusion that MRTX is able to form more contacts with the Switch-II region, as opposed to AMG, and hence able to highly reduce its fluctuation as confirmed through the RMSF and RMSD plots.

## 3.4 Conclusion

From all these studies on the oncogenic variant of the KRas and its drug-bound states, it can concluded that:

- The two switch regions are highly disordered as compared to the other parts of the protein.
- MRTX-bound variant restricts the motion of both the Switch-I and Switch-II as compared to the G12C oncogenic variant and the AMG-bound one.
- MRTX-bound variant completely leads to the loss of the anti-correlation present in both the G12C-mutated and AMG-bound structures.
- MRTX-drug heavily restricts the fluctuation of the Switch-II's $\alpha$-2 helix, by significantly restricting the $\alpha$-2-$\alpha$-3 dihedral angle.
- The maximal contacts formed are in the MRTX-bound one are mainly through the Hydrogen bonds and Hydrophobic interactions.

# 4 Exploring conformational landscape of drug-bound and unbound forms of KRAS: Deducing switch-mediated kick-out mechanism

## 4.1 Introduction

Under the general working mechanism of KRas, the GDP-GTP exchange process is catalysed by the protein called Guanine Exchange Factor (GEF). The GEF belongs to the SOS (Son of Sevenless) family of proteins.

Under physiological conditions when KRas is in its inactive state, the high-affinity binding of the GDP to the protein is due to the interaction of the phosphates of GDP with the P-loop's K16 (Lysine-16) and the $Mg^{2+}$ ion. But GEF binding, leads to the pushing in of the Switch-II towards the P-loop which indeed leads to the pushing out of the $Mg^{2+}$ ion from its initial position and hence looses its interaction with the GDP. The residue A59 (Alanine-59) plays an important role in the above process. Therefore, since the interactions of the GDP is weakened and Switch-II position is pushed inside, the P-loop reorients and the K16's amino groups start forming interaction with the carboxylates of switch-II residues, i.e. D57 (Aspartic Acid-57) / E62 (Glutamic Acid-62). This mechanism of Pull-in of the Switch-II lets the P-loop to loose interaction with the GDP and and let the GDP to float in the cavity formed by Switch-I. Following this, the Switch-I is pulled apart by the GEF, leading to the flying out of the GDP from that cavity and letting the GTP in. Therefore, this mechanism is known as the Kick-Out mechanism or the Push-pull mechanism since the pushing of Switch-II and pulling of Switch-I is observed [80]. This mechanism has been schematically depicted in the Figure 4.1.

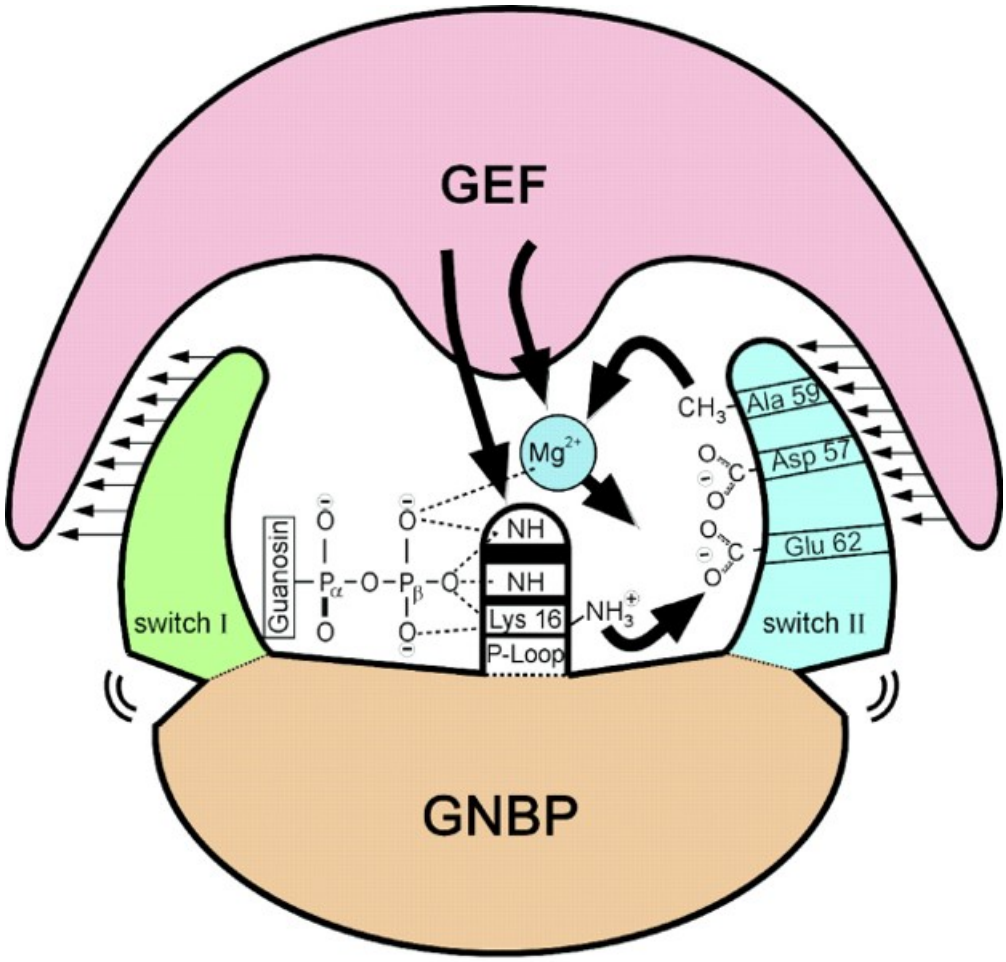



Figure 4.1: Schematic diagram of KRas-GEF Interaction scheme showing Kick-Out; adapted from Ref. [80]

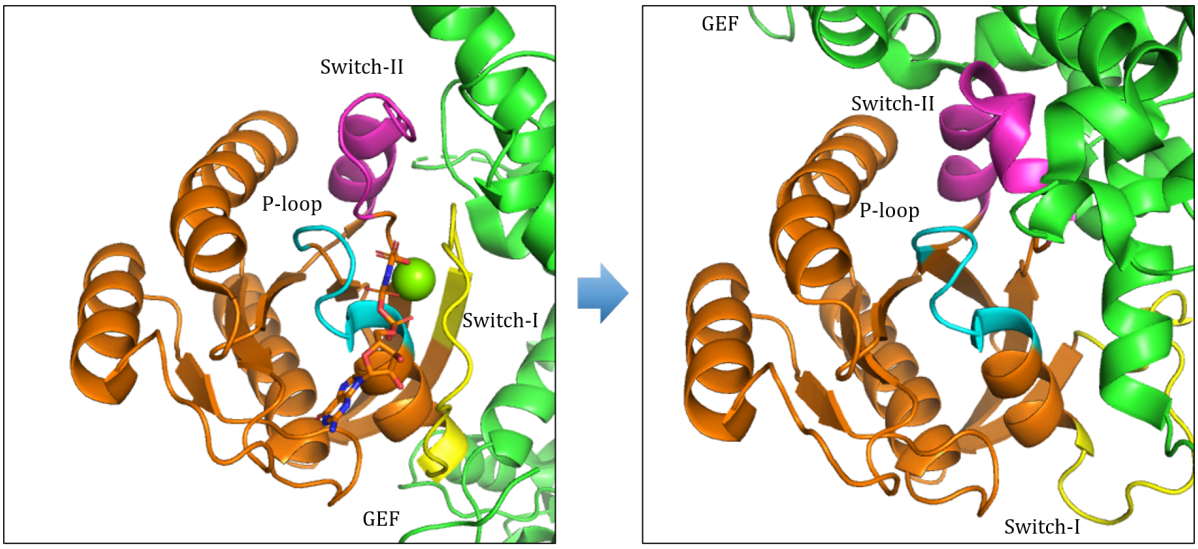



Figure 4.2: Binding and Unbinding mechanism of KRas-GEF interaction (PDB ID: 7KFZ)

Hence, our interest was to explore that if KRas can intrinsically reach this push-pull configuration in its energy landscape and how accessible it is in its drug-bound form. This could indicate an alternative pathway through which drugs might inhibit the GTP binding of mutated KRas.

## 4.2 Methodology: Well-tempered metadynamics simulation

Along with the systems described in Chapter 3, another extra system was incorporated here, which is the Wild-type KRas, so as to explore the conformation states visited by this native KRas system. Similar to the general MD protocol for all the three species described in the Section 3.2, all the steps were done exactly similar for all the four systems here, till the NPT equilibration. Once the systems were equilibrated, the simulation were patched with *PLUMED*[129, 130, 131], for the performance of the well-tempered 2D-metadynamics through Langevin dynamics. Also, the previously chosen distance order parameter was an 1D-order parameter and was unable to calculate the multidimensional free energy surface properly. Therefore, a collective variable such as RMSD was chosen to take care of it. Hence, the order parameters for this run were chosen as "RMSD of Switch-I" and "RMSD of Switch-II". For the Well-tempered metadynamics, the rate of hill deposition was set at 500, height and width of the Gaussian hills were set at 0.1 and 0.001 respectively, and order parameter bias factor was given a value of 15. The simulations were run through 200ns at the temperature 310K using grids for computational optimization. And the data were analysed to generate the free energy surface using the `sum_hills` module of *PLUMED*.

## 4.3 Results

### 4.3.1 Conformational states of Wild-type (WT) and G12C oncogenic variant of KRas

For the Wild-type (WT) variant of KRas, as can be seen from the Figure 4.3 there exists a single stable state. Also there is very less heterogeneity present in the structural landscape, which ranges from 0.12nm to 0.24nm for Switch-I RMSD, and from 0.23nm to 0.33nm for Switch-II RMSD. This region we propose to be the GDP-GTP exchange state. For the Figure 4.3 to 4.6, the red,

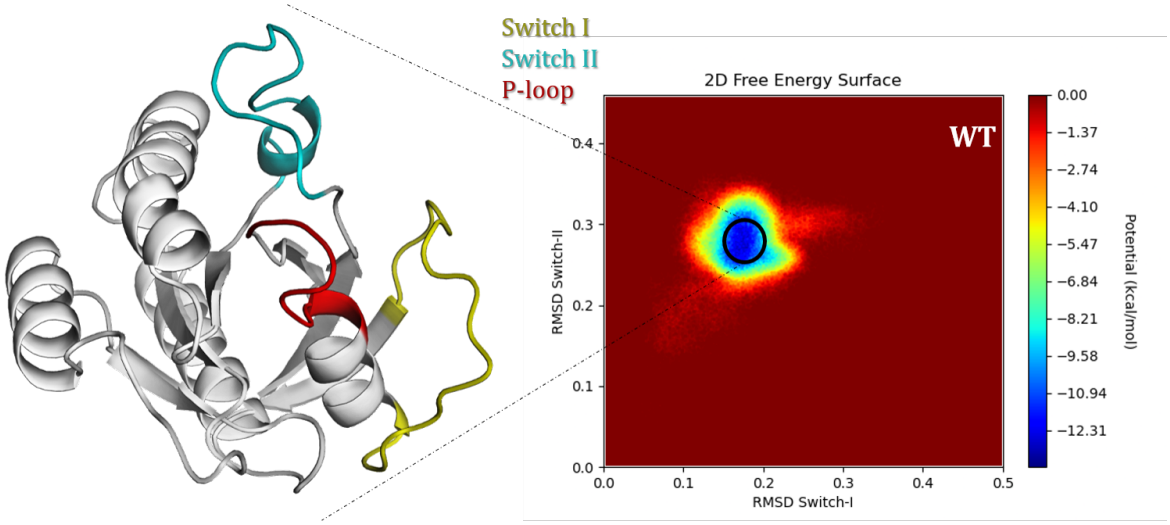



Figure 4.3: Well-tempered Metadynamics plot of WT KRas along with its most stable state

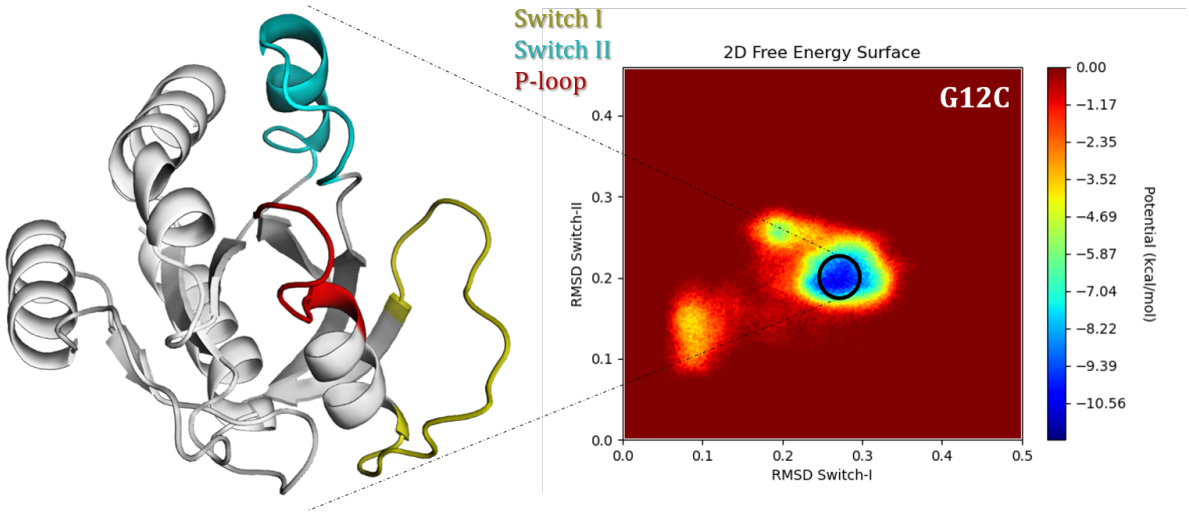



Figure 4.4: Well-tempered Metadynamics plot of G12C-mutated KRas along with its most stable state

yellow and cyan represent the P-loop, Switch-I and Switch-II, respectively. However, for the G12C-mutated oncogenic variant, there exists a different global minima as compared to the WT. This minima shows more deviation in the Switch-I and less in the Switch-II as compared to the WT. Nevertheless, it has a small population in the single state as WT as well. Also, it shows more heterogeneity than WT. Hence, we can propose that this variant has a small propensity for the GDP-GTP exchange to happen, but has a different major stable state, which might be leading to the oncogenicity.

### 4.3.2 Comparison of conformational states of the oncogenic variant & the drug-bound forms of KRas

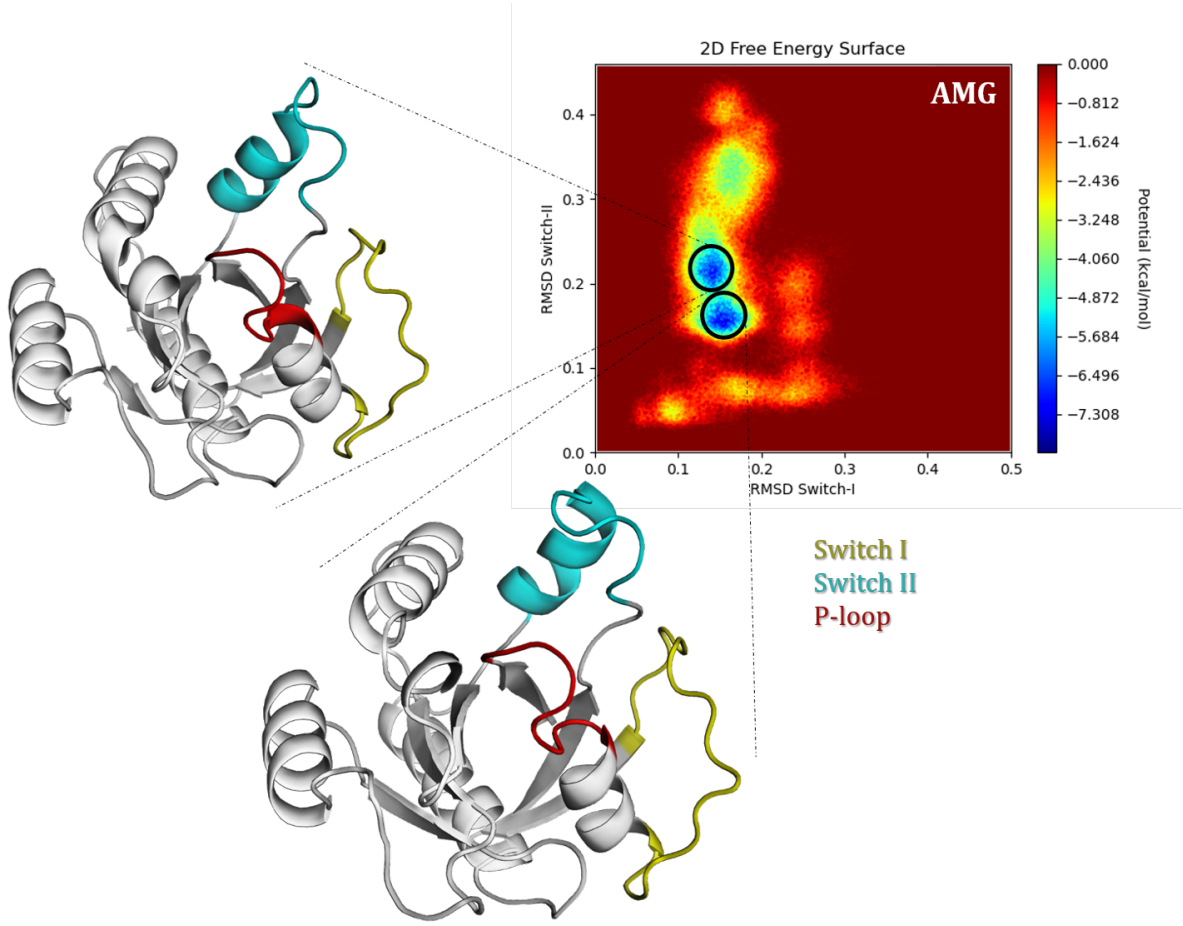



Figure 4.5: Well-tempered Metadynamics plot of AMG-bound mutated KRas along with its most stable states

As can be seen in the Figure 4.5, here for the AMG-bound mutated KRas, there exists very large heterogeneity in the energy landscape. There are two stable states as can be inferred from the landscape. However, the WT state still exists with a small population which indicates that GDP-GTP exchange pathway is still open even in the AMG bound state. This might be a reason why AMG is not a potent drug in inhibiting the mutated KRas's oncogenicity.

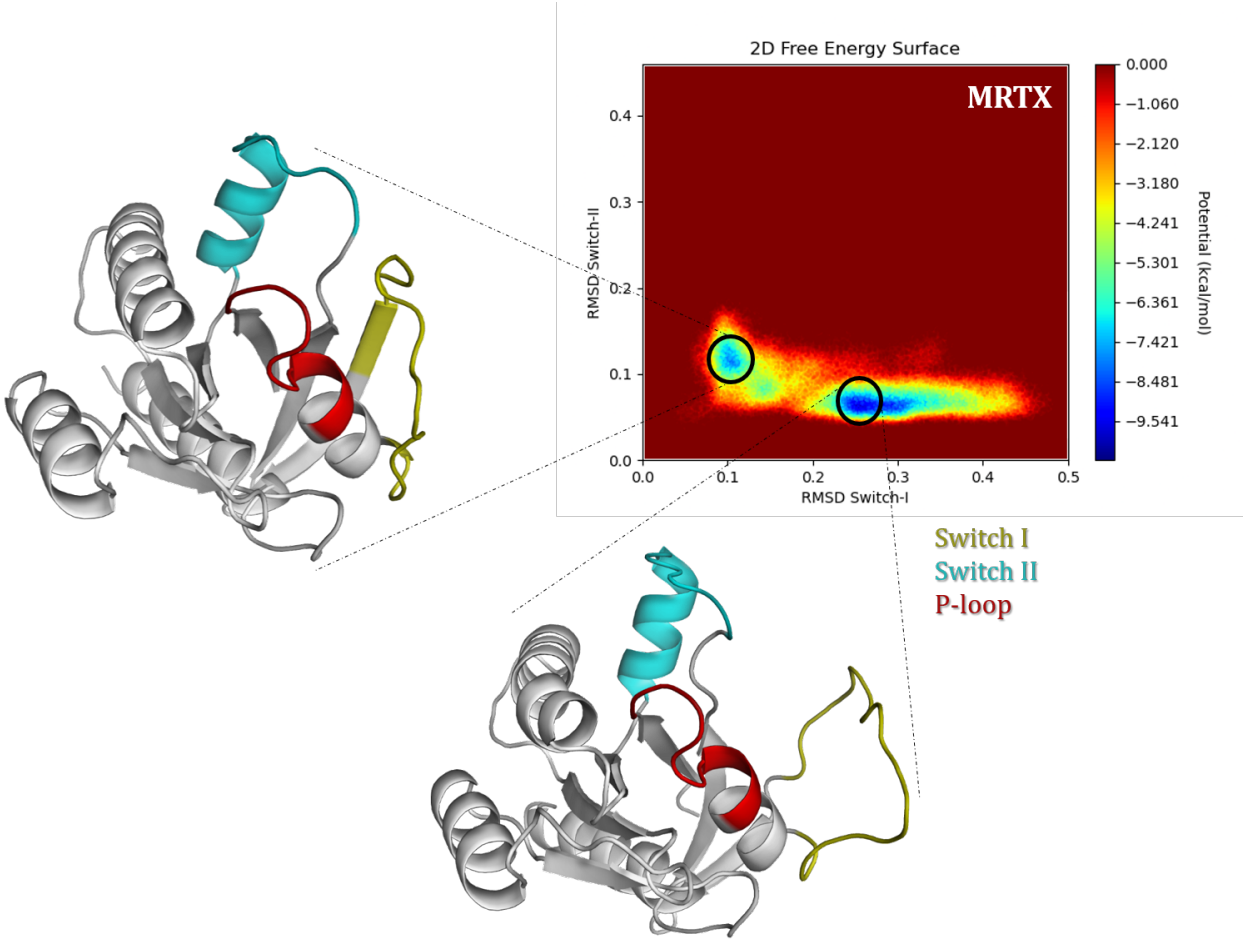



Figure 4.6: Well-tempered Metadynamics plot of MRTX-bound mutated KRas along with its most stable states

However, in the Figure 4.6, MRTX completely restricts the fluctuations in the Switch-II region of KRas in a controlled manner and has two stable states in its energy landscape. This also validates the conclusion made in the previous molecular dynamics data proposed in the previous chapter. Moreover, the WT state population doesn't exist. Therefore, it can be inferred that MRTX com-

pletely restricts the GDP-GTP exchange pathway.

## 4.4 Conclusion

From all these Metadynamics studies, it can be concluded that:

- For the WT KRas, there exists a single stable state, that we propose is the state that favours the GDP-GTP exchange.
- G12C mutated KRas, shows heterogeneity and also exhibits a small population that favours GDP-GTP exchange.
- AMG-bound one is unable to protect KRas from the GDP-GTP exchange state and also exhibits high heterogeneity in its energy landscape. However, MRTX is successful in restricting the protein from visiting the GDP-GTP exchange favourable state.
- Restriction of the Switch-II fluctuation is necessary for inhibiting the oncogenicity exhibited by KRas, as in the case for MRTX-bound one. This can be noted for any future drug that needs to be designed in case the current drug is no more potent, and the oncogenic protein becomes drug resistant.

## Future Aspects

- More free energy enhanced sampling simulations needs to be performed for quantifying the energy basins and better inferring the thermodynamic data.
- Protein docking experiments can be performed to study the downstream signalling interaction of KRas.
- The allosteric affect of other G12 mutations (G12V, G12D, G12A) to the switch regions needs to be chalked out.

# References


[1] Christian B Anfinsen et al. “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain”. In: *Proceedings of the National Academy of Sciences of the United States of America* 47.9 (1961), p. 1309.

[2] Christian B Anfinsen. “Principles that govern the folding of protein chains”. In: *Science* 181.4096 (1973), pp. 223–230.

[3] Jonathan J Ward et al. “Prediction and functional analysis of native disorder in proteins from the three kingdoms of life”. In: *Journal of molecular biology* 337.3 (2004), pp. 635–645.

[4] Vladimir N Uversky. “The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome”. In: *Journal of Biomedicine and Biotechnology* 2010 (2010).

[5] H Jane Dyson. “Expanding the proteome: disordered and alternatively folded proteins”. In: *Quarterly reviews of biophysics* 44.4 (2011), pp. 467–518.

[6] Peter E Wright and H Jane Dyson. “Intrinsically unstructured proteins: reassessing the protein structure-function paradigm”. In: *Journal of molecular biology* 293.2 (1999), pp. 321–331.

[7] Peter Csermely, Robin Palotai, and Ruth Nussinov. “Induced fit, conformational selection and independent dynamic segments: an extended view of binding events”. In: *Nature Precedings* (2010), pp. 1–1.

[8] Malene Ringkjøbing Jensen et al. “Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy”. In: *Chemical reviews* 114.13 (2014), pp. 6632–6660.

[9] Shelly DeForte and Vladimir N Uversky. “Order, disorder, and everything in between”. In: *Molecules* 21.8 (2016), p. 1090.

[10] Catherine A. Musselman and Tatiana G. Kutateladze. “Characterization of functional disordered regions within chromatin-associated proteins”. In: *iScience* 24.2 (2021), p. 102070. ISSN: 2589-0042. DOI: https://doi.org/10.1016/j.isci.2021.102070. URL: https://www.sciencedirect.com/science/article/pii/S2589004221000389.

[11] Vladimir N Uversky. “Unusual biophysics of intrinsically disordered proteins”. In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1834.5 (2013), pp. 932–951.

[12] Vladimir N Uversky. “Dancing protein clouds: the strange biology and chaotic physics of intrinsically disordered proteins”. In: *Journal of Biological Chemistry* 291.13 (2016), pp. 6681–6688.

[13] Vladimir N Uversky. “p53 proteoforms and intrinsic disorder: an illustration of the protein structure–function continuum concept”. In: *International journal of molecular sciences* 17.11 (2016), p. 1874.

[14] Vladimir N Uversky. “Protein intrinsic disorder and structure-function continuum”. In: *Progress in molecular biology and translational science* 166 (2019), pp. 1–17.

[15] Alexander V Fonin et al. “Multi-functionality of proteins involved in GPCR and G protein signaling: making sense of structure–function

continuum with intrinsic disorder-based proteoforms". In: *Cellular and Molecular Life Sciences* 76.22 (2019), pp. 4461–4492.

[16] Prakash Kulkarni et al. "Structural metamorphism and polymorphism in proteins on the brink of thermodynamic stability". In: *Protein Science* 27.9 (2018), pp. 1557–1567.

[17] Chad Haynes et al. "Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes". In: *PLoS computational biology* 2.8 (2006), e100.

[18] A Keith Dunker et al. "Flexible nets: the roles of intrinsic disorder in protein interaction networks". In: *The FEBS journal* 272.20 (2005), pp. 5129–5148.

[19] Marija Buljan et al. "Alternative splicing of intrinsically disordered regions and rewiring of protein interactions". In: *Current opinion in structural biology* 23.3 (2013), pp. 443–450.

[20] A Keith Dunker et al. "Intrinsic disorder and protein function". In: *Biochemistry* 41.21 (2002), pp. 6573–6582.

[21] Brian W Pontius. "Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association". In: *Trends in biochemical sciences* 18.5 (1993), pp. 181–186.

[22] H Jane Dyson and Peter E Wright. "Intrinsically unstructured proteins and their functions". In: *Nature reviews Molecular cell biology* 6.3 (2005), pp. 197–208.

[23] Jörg Gsponer and M Madan Babu. "The rules of disorder or why disorder rules". In: *Progress in biophysics and molecular biology* 99.2-3 (2009), pp. 94–103.

[24] Charles A Galea et al. “Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits”. In: *Biochemistry* 47.29 (2008), pp. 7598–7609.

[25] Mi-Kyung Yoon et al. “Cell cycle regulation by the intrinsically disordered proteins p21 and p27”. In: *Biochemical Society Transactions* 40.5 (2012), pp. 981–988.

[26] Jennifer M Hurley et al. “Conserved RNA helicase FRH acts nonenzymatically to support the intrinsically disordered Neurospora clock protein FRQ”. In: *Molecular cell* 52.6 (2013), pp. 832–843.

[27] Pei Dong et al. “A dynamic interaction process between KaiA and KaiC is critical to the cyanobacterial circadian oscillator”. In: *Scientific reports* 6.1 (2016), pp. 1–11.

[28] Thomas C Boothby et al. “Tardigrades use intrinsically disordered proteins to survive desiccation”. In: *Molecular cell* 65.6 (2017), pp. 975–984.

[29] Steven M Mooney et al. “Phenotypic plasticity in prostate cancer: role of intrinsically disordered proteins”. In: *Asian journal of andrology* 18.5 (2016), p. 704.

[30] Prakash Kulkarni. *Intrinsically disordered proteins: insights from Poincare, Waddington, and Lamarck*. 2020.

[31] Sohini Chakrabortee et al. “Intrinsically disordered proteins drive emergence and inheritance of biological traits”. In: *Cell* 167.2 (2016), pp. 369–381.

[32] Peter Tompa and Denes Kovacs. “Intrinsically disordered chaperones in plants and animals”. In: *Biochemistry and Cell Biology* 88.2 (2010), pp. 167–174.

[33] Lilia M Iakoucheva et al. “Intrinsic disorder in cell-signaling and cancer-associated proteins”. In: *Journal of molecular biology* 323.3 (2002), pp. 573–584.

[34] Prakash Kulkarni and Vladimir N Uversky. *Intrinsically disordered proteins in chronic diseases*. 2019.

[35] Patricia Santofimia-Castaño et al. “Targeting intrinsically disordered proteins involved in cancer”. In: *Cellular and Molecular Life Sciences* 77.9 (2020), pp. 1695–1707.

[36] M Madan Babu et al. “Intrinsically disordered proteins: regulation and disease”. In: *Current opinion in structural biology* 21.3 (2011), pp. 432–440.

[37] H Jane Dyson and Peter E Wright. “Coupling of folding and binding for unstructured proteins”. In: *Current opinion in structural biology* 12.1 (2002), pp. 54–60.

[38] David D Boehr, Ruth Nussinov, and Peter E Wright. “The role of dynamic conformational ensembles in biomolecular recognition”. In: *Nature chemical biology* 5.11 (2009), pp. 789–796.

[39] Peter E Wright and H Jane Dyson. “Intrinsically disordered proteins in cellular signalling and regulation”. In: *Nature reviews Molecular cell biology* 16.1 (2015), pp. 18–29.

[40] Angelo Toto et al. “Molecular recognition by templated folding of an intrinsically disordered protein”. In: *Scientific reports* 6.1 (2016), pp. 1–9.

[41] Peter Tompa and Monika Fuxreiter. “Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions”. In: *Trends in biochemical sciences* 33.1 (2008), pp. 2–8.

[42] Amrita Mohan et al. “Analysis of Molecular Recognition Features (MoRFs)”. In: *Journal of Molecular Biology* 362.5 (2006), pp. 1043–1059. ISSN: 0022-2836.

[43] Norman E Davey et al. “Attributes of short linear motifs”. In: *Molecular BioSystems* 8.1 (2012), pp. 268–281.

[44] Nicolás Palopoli et al. “Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity”. In: *Protein Engineering, Design and Selection* 31.3 (2018), pp. 69–77.

[45] Andreas Prestel et al. “The PCNA interaction motifs revisited: thinking outside the PIP-box”. In: *Cellular and Molecular Life Sciences* 76.24 (2019), pp. 4923–4943.

[46] Heli I Alanen et al. “Beyond KDEL: the role of positions 5 and 6 in determining ER localization”. In: *Journal of molecular biology* 409.3 (2011), pp. 291–297.

[47] Rob Kaptein and Gerhard Wagner. *Integrative methods in structural biology*. 2019.

[48] Magnus Kjaergaard, Kaare Teilum, and Flemming M Poulsen. “Conformational selection in the molten globule state of the nuclear coactivator

binding domain of CBP". In: *Proceedings of the National Academy of Sciences* 107.28 (2010), pp. 12535–12540.

[49] Magnus Kjaergaard et al. "Temperature-dependent structural changes in intrinsically disordered proteins: Formation of $\alpha$–helices or loss of polyproline II?" In: *Protein Science* 19.8 (2010), pp. 1555–1564.

[50] Ewa Jurneczko et al. "Intrinsic disorder in proteins: a challenge for (un) structural biology met by ion mobility–mass spectrometry". In: *Biochemical Society Transactions* 40.5 (2012), pp. 1021–1026.

[51] Carlo Camilloni et al. "Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts". In: *Biochemistry* 51.11 (2012), pp. 2224–2231.

[52] Pau Bernado and Dmitri I Svergun. "Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering". In: *Molecular biosystems* 8.1 (2012), pp. 151–167.

[53] Malene Ringkjøbing Jensen, Rob WH Ruigrok, and Martin Blackledge. "Describing intrinsically disordered proteins at atomic resolution by NMR". In: *Current opinion in structural biology* 23.3 (2013), pp. 426–435.

[54] Yann GJ Sterckx et al. "Small-angle X-ray scattering-and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2". In: *Structure* 22.6 (2014), pp. 854–865.

[55] Alessandro Borgia et al. "Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods". In: *Journal of the American Chemical Society* 138.36 (2016), pp. 11714–11726.

[56] Supriyo Bhattacharya and Xingcheng Lin. “Recent advances in computational protocols addressing intrinsically disordered proteins”. In: *Biomolecules* 9.4 (2019), p. 146.

[57] Claire C Hsu, Markus J Buehler, and Anna Tarakanova. “The order-disorder continuum: linking predictions of protein structure and disorder through molecular simulation”. In: *Scientific reports* 10.1 (2020), pp. 1–14.

[58] Robert B Best and Gerhard Hummer. “Microscopic interpretation of folding $\phi$-values using the transition path ensemble”. In: *Proceedings of the National Academy of Sciences* 113.12 (2016), pp. 3263–3268.

[59] Frank Noé and Cecilia Clementi. “Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods”. In: *Current opinion in structural biology* 43 (2017), pp. 141–147.

[60] David J Wales. “Energy landscapes: some new horizons”. In: *Current opinion in structural biology* 20.1 (2010), pp. 3–10.

[61] Antonio B Oliveira Jr et al. “Visualization of protein folding funnels in lattice models”. In: *PloS one* 9.7 (2014), e100861.

[62] Antonio B Oliveira Jr et al. “Distinguishing biomolecular pathways and metastable states”. In: *Journal of chemical theory and computation* 15.11 (2019), pp. 6482–6490.

[63] Manon Ragonnet-Cronin et al. “Automated analysis of phylogenetic clusters”. In: *BMC bioinformatics* 14.1 (2013), pp. 1–10.

[64] Robert B Best. “Computational and theoretical advances in studies of intrinsically disordered proteins”. In: *Current opinion in structural biology* 42 (2017), pp. 147–154.

[65] Massimiliano Bonomi et al. “Principles of protein structural ensemble determination”. In: *Current opinion in structural biology* 42 (2017), pp. 106–116.

[66] Kresten Lindorff-Larsen et al. “Structure and dynamics of an unfolded protein examined by molecular dynamics simulation”. In: *Journal of the American Chemical Society* 134.8 (2012), pp. 3787–3791.

[67] Andrea Cavalli, Carlo Camilloni, and Michele Vendruscolo. “Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle”. In: *The Journal of chemical physics* 138.9 (2013), 03B603.

[68] Ramya Rangan et al. “Determination of structural ensembles of proteins: restraining vs reweighting”. In: *Journal of chemical theory and computation* 14.12 (2018), pp. 6632–6641.

[69] Gül H Zerze et al. “Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics”. In: *Journal of chemical theory and computation* 11.6 (2015), pp. 2776–2782.

[70] Shalini Awasthi and Nisanth N Nair. “Exploring high dimensional free energy landscapes: Temperature accelerated sliced sampling”. In: *The Journal of Chemical Physics* 146.9 (2017), p. 094108.

[71] Trang Nhu Do, Wing-Yiu Choy, and Mikko Karttunen. “Accelerating the conformational sampling of intrinsically disordered proteins”. In: *Journal of Chemical Theory and Computation* 10.11 (2014), pp. 5081–5094.

[72] Xingcheng Lin et al. “Structural and dynamical order of a disordered protein: molecular insights into conformational switching of PAGE4 at the systems level”. In: *Biomolecules* 9.2 (2019), p. 77.

[73] Renxiang Yan et al. “A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction”. In: *Scientific reports* 3.1 (2013), pp. 1–9.

[74] Yanan He et al. “Phosphorylation-induced conformational ensemble switching in an intrinsically disordered cancer/testis antigen”. In: *Journal of Biological Chemistry* 290.41 (2015), pp. 25090–25102.

[75] Steven M Mooney et al. “Cancer/testis antigen PAGE4, a regulator of c-Jun transactivation, is phosphorylated by homeodomain-interacting protein kinase 1, a component of the stress-response pathway”. In: *Biochemistry* 53.10 (2014), pp. 1670–1679.

[76] Prakash Kulkarni et al. “Phosphorylation-induced conformational dynamics in an intrinsically disordered protein and potential role in phenotypic heterogeneity”. In: *Proceedings of the National Academy of Sciences* 114.13 (2017), E2644–E2653.

[77] Antonio B Oliveira Junior et al. “Exploring Energy Landscapes of Intrinsically Disordered Proteins: Insights into Functional Mechanisms”. In: *Journal of Chemical Theory and Computation* 17.5 (2021), pp. 3178–3187.

[78] David S Goodsell. “The molecular perspective: the ras oncogene”. In: *Stem cells* 17.4 (1999), pp. 235–236.

[79] Julian Downward. “Targeting RAS signalling pathways in cancer therapy”. In: *Nature reviews cancer* 3.1 (2003), pp. 11–22.

[80] Ingrid R Vetter and Alfred Wittinghofer. “The guanine nucleotide-binding switch in three dimensions”. In: *Science* 294.5545 (2001), pp. 1299–1304.

[81] AR Leach. “Four challenges in molecular modelling: free energies, solvation, reactions and solid-state defects”. In: *Molecular Modelling: Principles and Applications, 2nd Ed. Prentice Hall, New York* (2001), pp. 563–639.

[82] Hans Martin Senn and Walter Thiel. “QM/MM methods for biomolecular systems”. In: *Angewandte Chemie International Edition* 48.7 (2009), pp. 1198–1229.

[83] Rodrigo Galindo-Murillo et al. “Assessing the current state of amber force field modifications for DNA”. In: *Journal of chemical theory and computation* 12.8 (2016), pp. 4114–4127.

[84] Biman Bagchi. *Statistical Mechanics for Chemistry and Materials Science*. CRC Press, 2018.

[85] Ralf Schneider, Amit Raj Sharma, and Abha Rai. “Introduction to molecular dynamics”. In: *Computational Many-Particle Physics*. Springer, 2008, pp. 3–40.

[86] M Sprik. “Effective pair potentials and beyond”. In: *Computer simulation in chemical physics* (1993), pp. 211–259.

[87] *Potentials*. URL: http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials.

[88] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press, 2015.

[89] Mark James Abraham et al. “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1 (2015), pp. 19–25.

[90] Loup Verlet. “Computer” experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules”. In: *Physical review* 159.1 (1967), p. 98.

[91] Roger W. Hockney and James W. Eastwood. *Computer simulation using particles*. Hilger, 1989.

[92] William C Swope et al. “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters”. In: *The Journal of chemical physics* 76.1 (1982), pp. 637–649.

[93] HC Andersen. “Molecular dynamics at constant temperature and/or pressure”. In: *J. Chem. Phys* 72 (1980), p. 2384.

[94] Herman JC Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[95] Shuichi Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. In: *The Journal of chemical physics* 81.1 (1984), pp. 511–519.

[96] William G Hoover. “Canonical dynamics: Equilibrium phase-space distributions”. In: *Physical review A* 31.3 (1985), p. 1695.

[97] Michele Parrinello and Aneesur Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *Journal of Applied physics* 52.12 (1981), pp. 7182–7190.

[98] perm_identity Posted by LAMMPS Tube. *Periodic boundary conditions*. Jan. 2022. URL: https://lammpstube.com/2019/10/30/periodic-boundary-conditions/.

[99] Paul Peter Ewald. "Ewald summation". In: *Ann. Phys* 369.253 (1921), pp. 1–2.

[100] Essmann U Perera L Berkowitz ML and Darden T Lee H Pedersen LG. "A smooth particle mesh Ewald method". In: *J. Chem. Phys* 103.19 (1995), pp. 8577–8593.

[101] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N log (N) method for Ewald sums in large systems". In: *The Journal of chemical physics* 98.12 (1993), pp. 10089–10092.

[102] Daan Frenkel and Berend Smit. "Chapter 4-molecular dynamics simulations". In: *Understanding Molecular Simulation* 2 (2002), pp. 63–107.

[103] Tim N Heinz and Philippe H Hünenberger. "A fast pairlist-construction algorithm for molecular simulations under periodic boundary conditions". In: *Journal of computational chemistry* 25.12 (2004), pp. 1474–1486.

[104] William Mattson and Betsy M Rice. "Near-neighbor calculations using a modified cell-linked list method". In: *Computer Physics Communications* 119.2-3 (1999), pp. 135–148.

[105] Glenn M Torrie and John P Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *Journal of Computational Physics* 23.2 (1977), pp. 187–199.

[106] Johannes Kästner. "Umbrella sampling". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.6 (2011), pp. 932–942.

[107] Alessandro Laio and Michele Parrinello. “Escaping free-energy minima”. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[108] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. “Well-tempered metadynamics: a smoothly converging and tunable free-energy method”. In: *Physical review letters* 100.2 (2008), p. 020603.

[109] Alessandro Laio and Francesco L Gervasio. “Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science”. In: *Reports on Progress in Physics* 71.12 (2008), p. 126601.

[110] Giovanni Bussi, Alessandro Laio, and Pratyush Tiwary. “Metadynamics: A unified framework for accelerating rare events and sampling thermodynamics and kinetics”. In: *Handbook of materials modeling: Methods: theory and modeling* (2020), pp. 565–595.

[111] Chakra Chennubhotla et al. “Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies”. In: *Physical biology* 2.4 (2005), S173.

[112] Anne B Vojtek and Channing J Der. “Increasing complexity of the Ras signaling pathway”. In: *Journal of Biological Chemistry* 273.32 (1998), pp. 19925–19928.

[113] Nobuo Tsuchida, Tom Ryder, and Eiichi Ohtsubo. “Nucleotide sequence of the oncogene encoding the p21 transforming protein of Kirsten murine sarcoma virus”. In: *Science* 217.4563 (1982), pp. 937–939.

[114] Daniel Zeitouni et al. “KRAS mutant pancreatic cancer: no lone path to an effective treatment”. In: *Cancers* 8.4 (2016), p. 45.

[115] Sezen Vatansever, Zeynep H Gümüş, and Burak Erman. “Intrinsic K-Ras dynamics: A novel molecular dynamics data analysis method shows causality between residue pair motions”. In: *Scientific reports* 6.1 (2016), pp. 1–12.

[116] Rebecca L Siegel et al. “Cancer statistics, 2021”. In: *CA: a cancer journal for clinicians* 71.1 (2021), pp. 7–33.

[117] Ravi Salgia. “Mutation testing for directing upfront targeted therapy and post-progression combination therapy strategies in lung adenocarcinoma”. In: *Expert Review of Molecular Diagnostics* 16.7 (2016), pp. 737–749.

[118] S Dearden et al. “Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap)”. In: *Annals of oncology* 24.9 (2013), pp. 2371–2376.

[119] Kathryn C Arbour et al. “Treatment Outcomes and Clinical Characteristics of Patients with KRAS-G12C–Mutant Non–Small Cell Lung Cancer”. In: *Clinical Cancer Research* 27.8 (2021), pp. 2209–2215.

[120] Hayley Virgil. *New Drug Application for ADAGRASIB accepted by FDA for Kras G12c+ NSCLC*. 2022. URL: https://www.cancernetwork.com/view/new-drug-application-for-adagrasib-accepted-by-fda-for-kras-g12c-nsclc.

[121] Jude Canon et al. “The clinical KRAS (G12C) inhibitor AMG 510 drives anti-tumour immunity”. In: *Nature* 575.7781 (2019), pp. 217–223.

[122] Jay B Fell et al. “Identification of the clinical development candidate MRTX849, a covalent KRASG12C inhibitor for the treatment of cancer”. In: *Journal of Medicinal Chemistry* 63.13 (2020), pp. 6679–6693.

[123] John C Hunter et al. “In situ selectivity profiling and crystal structure of SML-8-73-1, an active site inhibitor of oncogenic K-Ras G12C”. In: *Proceedings of the National Academy of Sciences* 111.24 (2014), pp. 8895–8900.

[124] Andrew Waterhouse et al. “SWISS-MODEL: homology modelling of protein structures and complexes”. In: *Nucleic acids research* 46.W1 (2018), W296–W303.

[125] Hao Liu et al. *Intrinsically disordered protein-specific force field CHARMM 36 IDPSFF*. 2018.

[126] Robert B Best et al. “Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone $\phi$, $\psi$ and side-chain $\chi 1$ and $\chi 2$ dihedral angles”. In: *Journal of chemical theory and computation* 8.9 (2012), pp. 3257–3273.

[127] Alexander D MacKerell Jr, Michael Feig, and Charles L Brooks. “Improved treatment of the protein backbone in empirical force fields”. In: *Journal of the American Chemical Society* 126.3 (2004), pp. 698–699.

[128] Pekka Mark and Lennart Nilsson. “Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K”. In: *The Journal of Physical Chemistry A* 105.43 (2001), pp. 9954–9960.

[129] Massimiliano Bonomi et al. “PLUMED: A portable plugin for free-energy calculations with molecular dynamics”. In: *Computer Physics Communications* 180.10 (2009), pp. 1961–1972.

[130] Massimiliano Bonomi et al. “Promoting transparency and reproducibility in enhanced molecular simulations”. In: *Nature methods* 16.8 (2019), pp. 670–673.

[131] Gareth A Tribello et al. “PLUMED 2: New feathers for an old bird”. In: *Computer Physics Communications* 185.2 (2014), pp. 604–613.