

## University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF NATURAL & ENVIRONMENTAL SCIENCES

School of Chemistry

**Prediction of the binding free energies of inhibitors of Epidermal Growth Factor  
Receptor kinase and the identification of the dynamics thereof**

by

**Christopher Bull**

Thesis for the degree of Doctor of Philosophy

September 2013

**Supervisor:** Prof. Jonathan Essex

**Industrial supervisor:** Dr. Richard Ward

**Advisor:** Dr. Syma Khalid

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL & ENVIRONMENTAL SCIENCES

SCHOOL OF CHEMISTRY

Thesis for the degree of Doctor of Philosophy

**PREDICTION OF THE BINDING FREE ENERGIES OF INHIBITORS OF EPIDERMAL  
GROWTH FACTOR RECEPTOR KINASE AND THE IDENTIFICATION OF THE  
DYNAMICS THEREOF**

By Christopher William Bull

Epidermal Growth Factor Receptor (EGFR) kinase is a signalling protein implicated in a number of cancers, including non-small cell lung cancer (NSCLC). As well as activating mutations of EGFR kinase being oncogenic, the prognosis of NSCLC correlates with the impact of EGFR mutations on inhibitor binding affinities. However, treatment with tyrosine kinase inhibitors is particularly vulnerable to resistance mutations. The exact mechanisms by which EGFR kinase mutations impart activation or resistance has not been clearly defined at an atomistic level, and attempts to elucidate these mechanisms *in silico* are hindered by the long time scales over which the conformational dynamics of EGFR kinase occur. In this thesis rigorous free energy calculations are employed to investigate the relative binding free energy of inhibitors of EGFR kinase, and elucidate the hydration of the binding pocket. Additionally, various enhanced molecular dynamics (MD) sampling methods are utilised alongside conventional MD to investigate their ability to overcome the challenge of the long time scales of conformational change in EGFR kinase. The complementary use of dimensionality reduction techniques such as principal components analysis and locally scaled diffusion map analysis is shown to be useful in characterising long time scale dynamics, as well as in validating the sampling of enhanced MD methods. Using these

techniques alongside traditional analyses, new insight into the role of three activating mutations was gained; however, the results suggest that accessible simulation times are still too short, implying a continuing role for enhanced MD methods in the future.

# Contents

Contents .....	i
DECLARATION OF AUTHORSHIP .....	vii
Acknowledgements .....	ix
Chapter 1: Introduction .....	1
Chapter 2: Epidermal Growth Factor Receptor.....	3
2.1 Molecular biology of EGFR.....	3
2.1.1 Physiological role of EGFR.....	3
2.1.2 Signalling pathways .....	4
2.1.3 EGFR in cancer .....	6
2.2 Structure of EGFR .....	7
2.2.1 Structural regulation of EGFR compared to other kinases.....	11
2.2.2 Overview of available structural data .....	12
2.3 EGFR kinase mutations .....	14
2.3.1 L858R .....	15
2.3.2 G719X.....	16
2.3.3 Exon 19 deletions.....	17
2.3.4 T790M.....	18
2.3.5 Other mutations.....	19
2.3.6 Mutations summary .....	19
2.4 EGFR kinase inhibitors.....	20
2.5 Computational studies of EGFR .....	22
2.5.1 Insight into the WT dynamics of EGFR .....	22
2.5.2 Insights into the effect of EGFR mutations .....	26
2.5.3 Prediction of EGFR inhibitor binding affinity.....	29
2.6 Conclusions.....	29
2.6.1 Implications on the experimental design.....	31
Chapter 3: Computational methods .....	35

3.1	Statistical mechanics .....	35
3.1.1	Statistical ensembles.....	39
3.1.2	Assumptions.....	40
3.1.3	The partition function.....	40
3.2	Molecular dynamics .....	43
3.2.1	Newton's second law .....	43
3.2.2	Timesteps .....	45
3.2.3	Thermodynamic conditions .....	46
3.3	Monte Carlo .....	48
3.3.1	The Metropolis test.....	48
3.3.2	Observing detailed balance.....	49
3.3.3	Generating configurations.....	50
3.3.4	Grand Canonical Monte Carlo .....	51
3.3.5	Just Add Water Molecules .....	52
3.4	Modelling interactions of a system.....	54
3.4.1	Force fields.....	54
3.4.2	Periodic boundary conditions.....	56
3.4.3	Water models.....	57
3.4.4	Energy minimisation.....	58
3.5	Calculation of binding free energies .....	59
3.5.1	Relative binding free energies.....	59
3.6	Hybrid Quantum Mechanics/Molecular Mechanics .....	61
3.6.1	The Schrödinger equation .....	61
3.6.2	Density functional theory .....	64
3.6.3	QM/MM rescoring of MM free energies .....	65
3.7	Analysis methods.....	67
3.7.1	Secondary structure prediction.....	67
3.7.2	Dimensionality reduction.....	68
3.7.3	Principal Component Analysis.....	68
3.7.4	Multidimensional scaling .....	69
3.7.5	Diffusion map analysis .....	70

3.8	Enhanced sampling techniques .....	71
3.8.1	Accelerated Molecular Dynamics .....	72
3.8.2	Diffusion Map Directed Molecular Dynamics .....	74
3.8.3	Reversible Digitally Filtered Molecular Dynamics .....	76
3.9	Summary .....	79
<b>Chapter 4: Prediction of relative binding free energies .....</b>		<b>80</b>
4.1	Introduction to azoquinazoline inhibitors.....	80
4.2	System setup and protocols .....	82
4.3	Initial results.....	85
4.4	Water site prediction.....	95
4.4.1	System setup.....	96
4.4.2	GCMC protocol .....	96
4.4.3	GCMC results .....	98
4.4.4	JAWS protocol.....	104
4.4.5	JAWS results.....	105
4.4.6	Context of identified water sites .....	107
4.5	Results with predicted waters .....	108
4.6	QM/MM corrections.....	114
4.6.1	System setup.....	115
4.6.2	Results.....	115
4.7	Conclusions.....	117
<b>Chapter 5: Conformational dynamics of EGFR.....</b>		<b>119</b>
5.1	System setup .....	120
5.2	Setup of enhanced sampling methods.....	124
5.2.1	AMD .....	125
5.2.2	DMDMD.....	126
5.2.3	RDFMD .....	127
5.3	RMSD results.....	129
5.3.1	Active A-loop RMSDs.....	130
5.3.2	Inactive A-loop RMSDs .....	135

5.3.3	Active C–helix RMSDs .....	140
5.3.4	Inactive C–helix RMSDs .....	145
5.3.5	RMSD summary.....	150
5.3.6	RMSF results.....	151
5.3.7	Active RMSFs.....	152
5.3.8	Inactive RMSFs.....	163
5.3.9	RMSF summary.....	174
5.4	Secondary structure analysis .....	174
5.4.1	Active conformation structure .....	175
5.4.2	Inactive conformation structure .....	180
5.5	Dimensionality reduction .....	187
5.5.1	PCA results .....	187
5.5.2	Diffusion map results .....	205
5.6	Visualisation and other analyses .....	213
5.6.1	Sampling of the active trajectories .....	214
5.6.2	Sampling of the inactive trajectories.....	218
5.7	Summary of the Molecular Dynamics results.....	221
5.8	Evaluation of sampling methods .....	224
5.9	Investigation of the relevance of supertrajectory PCA results.....	229
5.9.1	Comparison of sampling methods by subspace overlap.....	238
<b>Chapter 6:</b>	<b>Conclusions .....</b>	<b>241</b>
6.1	Future work .....	244
<b>Appendices .....</b>	<b>247</b>	
Appendix 1:	Secondary structure analysis.....	247
Appendix 2:	RMSF analysis of RDFMD trajectories.....	248
	Impact of mutations on RDFMD simulations.....	249
	Impact of different RDFMD targets on sampling .....	250
Appendix 3:	PCA of each mutant and WT .....	252
	PCA of active conformations of each mutant.....	252
	PCA of inactive conformations of each mutant.....	261

Appendix 4: Additional thermodynamic cycles .....	270
Appendix 5: Elucidation of water sites by GCMC .....	271
Appendix 6: Principal component dot products .....	277
<b>Bibliography.....</b>	<b>285</b>



# **DECLARATION OF AUTHORSHIP**

**I, Christopher Bull**

**declare that the thesis entitled**

**Prediction of the binding free energies of inhibitors of Epidermal Growth Factor Receptor kinase and the identification of the dynamics thereof**

**and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:**

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed: .....

Date:.....

# Acknowledgements

I could have gained much less from my studies, and it is my supervisor Jon, who I would most like to thank, for his insight, enthusiasm and patience, virtues of which he had plenty to share when I had run short.

This thesis owes a lot to the people at Astra Zeneca, who made this work possible. Among them, my industrial supervisor, Richard Ward, whose discussions and advice has been particularly influential. I would also like to thank the Cecilia Clementi, Wenwei Zheng and Mary Rohrdanz at Rice University, Texas, who went out of their way help us implement their analyses on our work, as well as Charlie Laughton of Nottingham University for providing perspective in the sometimes confusing topic of dimensionality reduction.

All of the above have contributed much, but none more so than the combined friendship of members of the Essex, Khalid and Skylaris groups. To list them each by name would be impossible, but I particularly want to thank Michael Bodnarchuk, without whom my simulations would have been much drier, as well as the other members of the 2009 quartet: Michael Carter, Barbara Sander, and Ioannis Haldoupis, for their unfaltering companionship that made me look forward to each morning, not to mention each Friday evening!

Finally, without the support and love of my family I may never have embarked on these studies at all; all that I am, and all that I ever hoped to be, I owe to you. And, most heartfelt of all, is my thanks to my beautiful wife Masako, for leaving the security and luxury of Japan to slum it with me, and for filling every moment of this journey with life.







## Chapter 1: Introduction

Epidermal Growth Factor Receptor (EGFR) is a signalling protein involved in the correct development of embryos, and has become an important target because of its involvement in a number of cancers, where EGFR may be found upregulated or mutated[1]. EGFR mutations are important due to their diagnostic value, but also because the presence of EGFR mutations has a significant impact on how well a cancer responds to certain drugs, such as gefitinib[2]. Moreover, it has also been found that treatment with such drugs often leads to resistance[3]. Thus, there is ample motive to study EGFR and, owing to this, EGFR has been studied extensively *in silico* in an attempt to characterise its role at an atomistic level; however, the precise atomistic role of mutations in EGFR remains unclear.

The *in silico* study of EGFR kinase generally concerns itself with one or both of two lines of inquiry: How do EGFR mutations impact the conformational dynamics of EGFR kinase, and how do EGFR kinase mutations impact the ability of drug molecules to bind to the kinase?

Investigating the impact of EGFR mutations on the conformational dynamics of the kinase is problematic due to the long time scales over which kinase activation and deactivation may occur[4]; indeed, the activation of EGFR kinase has never been simulated in full. The deactivation process is also challenging, having only been observed previously[5] by utilisation of a specialist supercomputer, ANTON, which was designed specifically to probe such problems. Such simulations require dozens of  $\mu$ s of simulation time, which is beyond the available computing resources of institutions such as the one the present study originates. Methods exist for accelerating the dynamics, but few of these have been applied to EGFR. In the present study three

enhanced sampling techniques (accelerated MD, reversibly digitally filtered MD, and diffusion map directed MD) are compared against conventional MD (cMD) utilising more commonly available software and computational resources (though still limited to supercomputing).

Previous investigations into the impact of mutations on inhibitor binding free energies has also been carried out on EGFR kinase, but only using approximate methods such as MM/GBSA and MM/PBSA[6],[7]. To perform a more rigorous analysis requires that the more rigorous methods are capable of identifying trends within the binding of inhibitors to the WT. The present study attempts to realise this first step by utilising Monte Carlo simulations to predict the binding affinities of azoquinazolines to EGFR kinase using the replica exchange thermodynamic integration method.

Our knowledge of EGFR kinase is improving constantly, and chapter 2 will summarise the current understanding of the protein, discuss the importance of EGFR kinase in greater detail, as well as introducing the computational studies carried out on EGFR to date. Chapter 3 is devoted to a description of the computational methods employed in the present study. Chapter 4 will discuss the results of the prediction of EGFR inhibitor binding free energies and chapter 5 will deal with the conformational dynamics of EGFR and performance of the utilised enhanced sampling methods.

## Chapter 2: Epidermal Growth Factor Receptor

### 2.1 Molecular biology of EGFR

Epidermal Growth Factor Receptor (EGFR) is a membrane bound signalling protein of the ErbB family. It is essential for the normal development of various tissues, including bone [8], mammary ducts [9], vascular system [10], and others[11]. Owing to its role in development, EGFR is normally found at low levels in most tissues [12][13], where it is usually highly regulated both transcriptionally and mechanically by existing as an inactive monomer, requiring dimerization that is facilitated by the binding of extracellular signals such as Epidermal Growth Factor (EGF)[11].

EGFR gained pharmaceutical significance with the discovery of its involvement in a number of cancers, most notably Non-Small Cell Lung Cancer and some head and neck cancers, although it is thought that EGFR is upregulated in a large proportion of cancers[1]. As well as being implicated in the development of cancer, the mutational status of cancer patients has been found to correlate with patient outcomes[14].

#### 2.1.1 Physiological role of EGFR

EGFR exists as a transmembrane protein in a variety of cell types, where it receives signals in the form of small proteins (including EGF) from the extracellular environment. Many of these signals are produced in times of cellular stress by the signalling molecule's cleavage from membranes in the signal's source cell [15]. EGFR forms inactive dimers on the cell surface which prime the protein for binding of the

protein signal, which stimulates kinase activation[16]. Different ligands have different binding affinities, and also impact the rate of EGFR degradation, leading to a certain degree of ligand-dependent EGFR signalling regulation[17]. EGFR's ability to heterodimerise with other members of the ErbB family broadens its possible range of responses[18].

### 2.1.2 Signalling pathways

Downstream pathways include the phospholipase C $\gamma$  (PLC $\gamma$ ) pathway, which regulates cell motility, the MAPK pathway, which regulates cell proliferation, the STAT pathway which also regulates cell proliferation, as well as differentiation and apoptosis[17][18], and the PI3K/AKT pathway which prevents apoptosis. Many of these pathways are viable due to the C-tail domain of EGFR, whose phosphorylation motifs enable binding to the SH2 regions of various signalling proteins[17][18].

In the MAPK pathway, EGFR binds to a Grb2:SOS protein complex, which then activates the GTPase Ras, which in turn activates Raf kinase. The activation cascade then proceeds via MEK and MAPK1, the latter of which is translocated into the nucleus to act as a transcription factor to promote cell growth[18].

EGFR activates PLC $\gamma$  directly, which can activate RAS (via SOS) to promote cell growth (as previously discussed). Additionally, PLC $\gamma$  can interact with protein kinase C, stimulating gene expression[19].

EGFR has been shown to bind STAT proteins, including STAT3, which it is able to achieve in the nucleus, acting as a nuclear transcription factor capable of inducing the expression of nitric oxide synthase, which is associated with tumour growth[20].

In the PI3K/AKT pathway, PI3K binds to EGFR. Its presence on the membrane allows PI3K to start phosphorylating phosphatidylinositol 4,5-bisphosphate (PIP<sub>2</sub>) to produce PIP<sub>3</sub>, which activates AKT[21]. AKT inhibits the apoptosis signalling protein BAD[18].

However, it is important to note that these downstream pathways are not only accessible to the ErbB family, and interactions from other proteins help the cell to elicit appropriate responses to the environment (see figure 2.1 for a simplified schematic). For example, it has been shown that by blocking the MAP kinase kinase (MEK), can prevent motility while retaining some other functions of EGFR[22]. Indeed, this is a relatively simple example in the complex network of signalling that regulates gene expression [23].

EGFR's role as a promoter of motility, proliferation and differentiation makes it an important protein for the natural development of animals. EGFR has been shown to be necessary for the correct development of tissues both during adult physiological changes (such as in gestating mothers) and foetus development [24], [25]. However, the prominent position of this protein in the proliferation pathway also makes it an important participant in a number of cancers, and EGFR is found to be particularly important in non-small cell lung cancer (NSCLC)[26].

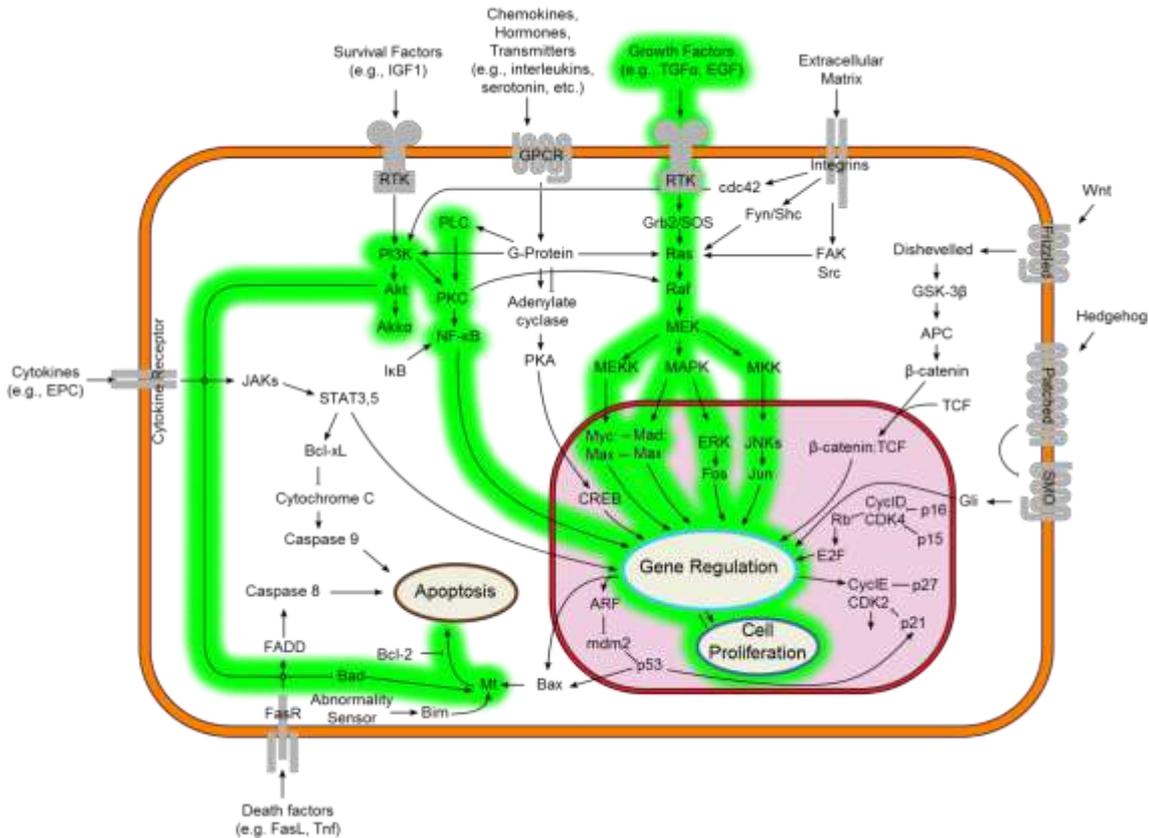


Figure 2.1: Simplified schematic of cell signalling pathways involved in cancer, with downstream pathways of EGFR highlighted in green. Adapted from ref. [27], itself based on Figure 2 of Hanahan & Weinberg (2000)[23].

### 2.1.3 EGFR in cancer

The current study deals primarily with mutations of the EGFR kinase region, however there are a number of mechanisms by which EGFR dysregulation can predispose an individual to cancer. Firstly, EGFR may be upregulated, leading to a larger proportion of EGFR propagating signals on the cell surface [28]. Secondly, EGFR may become mutated leading to increased EGFR activity [29][26]. Thirdly, EGFR segregation or breakdown processes may be interrupted[30]. Each of these lead to increased signalling along the proliferation pathways, increasing the likelihood of a cell becoming cancerous.

## 2.2 Structure of EGFR

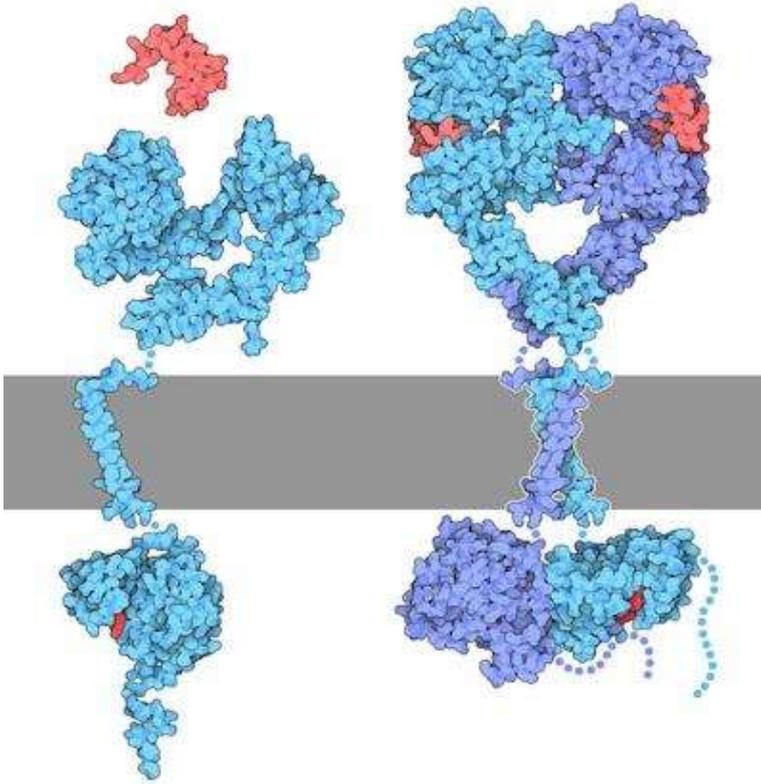


Figure 2.2: Illustration of EGFR (blue) in its monomeric (left) and dimeric (right) forms, with the extracellular region (top) shown binding EGF (red). The kinase is the intracellular (bottom) portion of the protein. Image from June 2010 Molecule of the Month by David Goodsell DOI: 10.2210/rcsb\_pdb/mom\_2010\_6

EGFR consists of a 622 amino acid extracellular region linked to a 542 amino acid intracellular region by the transmembrane helix. The extracellular region comprises of a ligand binding domain and a dimerisation domain, allowing the protein to sense growth factors in the environment, and dimerise in response to them [17]. Dimerisation of the extracellular region is accompanied by asymmetric intracellular dimerisation, which is thought to activate the kinase region of EGFR within the cell[29]. EGFR also

has a long carboxy-terminal tail which is recognised by a number of signalling proteins[17].

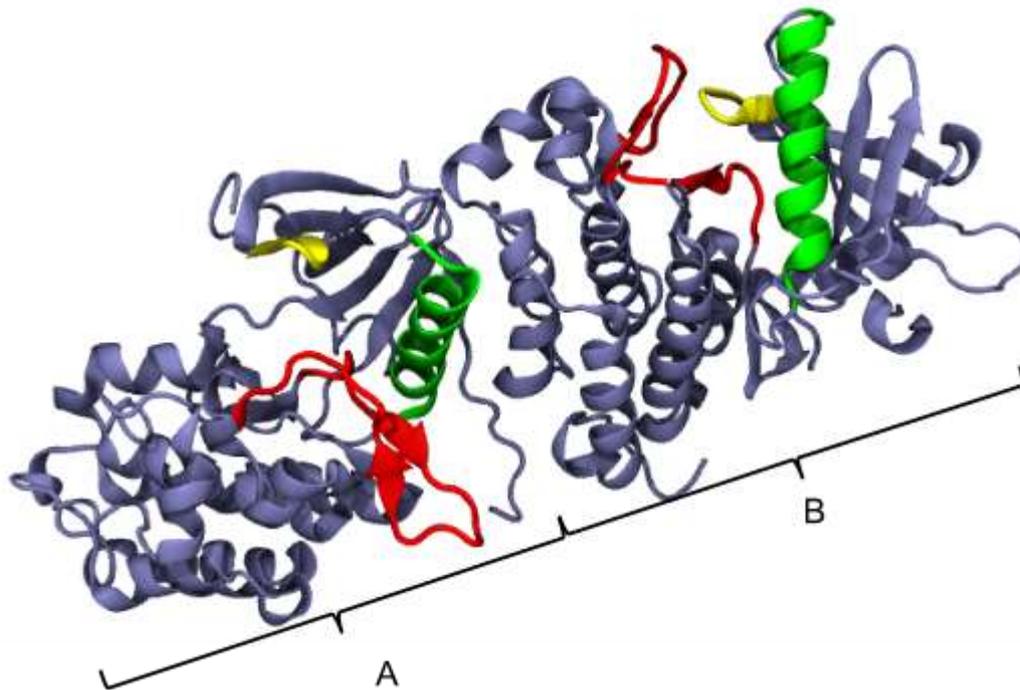


Figure 2.3: Crystal structure of the intracellular EGFR kinase dimer consisting of an active monomer (A) and activating monomer (B). PDB code 3IKA.

Crystallographic structures of the extracellular and intracellular region exist. The extracellular region has not been simulated in this study, but has been dealt with by other institutions[31][32][33]. An extensive array of crystal structures of the intracellular kinase have been reported to date, including those of a number of mutants both with and without inhibitors bound, providing abundant information for simulation.

The first successful crystallographic analysis of EGFR kinase was by Stamos et al. (2002)[34], and showed the active kinase both with and without an azoquinazoline bound to the ATP binding site. This active structure is shown in figure 2.4A.

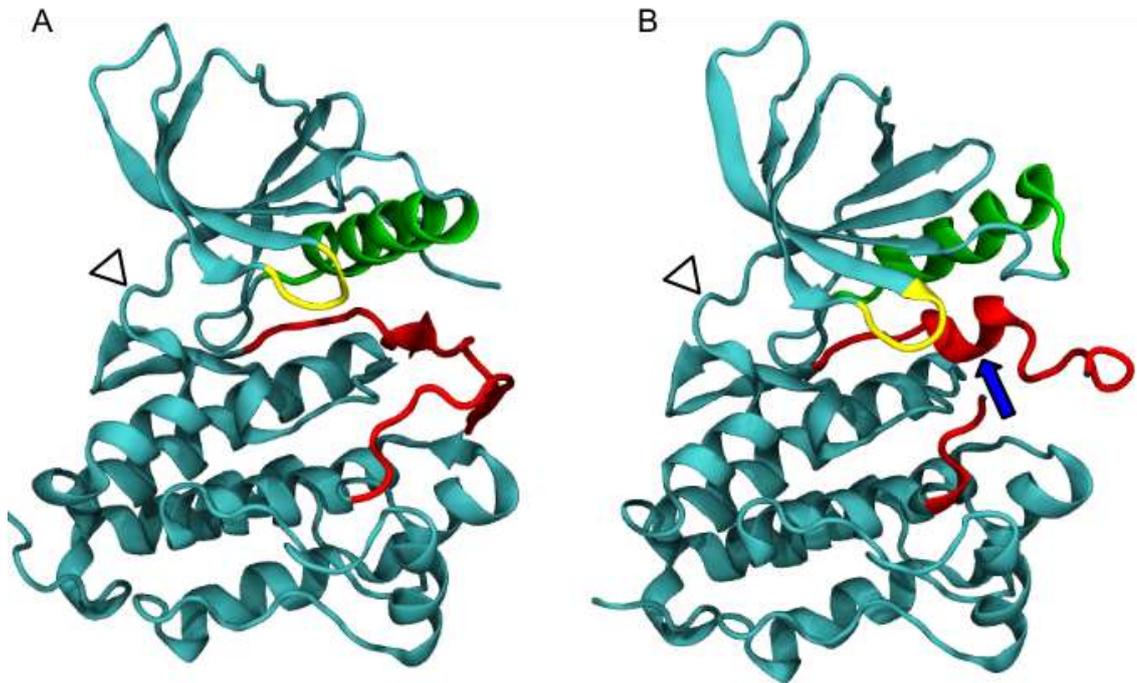


Figure 2.4: Crystal structures of the active (A; PDB: 1m17) and inactive (B; PDB: 2gs7) monomers of EGFR kinase. The activation loop is shown in red, C-helix in green and P-loop in yellow. The white arrowhead indicates the hinge region, and the blue arrow indicates the A-loop helical coil in the inactive structure.

Crystallographic structures of the inactive form of EGFR (see figure 2.4B) revealed large structural differences between the active and inactive conformations. Firstly, there is a 30 degree rotation of the C-helix (green), which exhibits an “in” conformation in the active structure, and an “out” conformation in the inactive structure. Also there is an extensive difference between the active and inactive structure of the activation loop (A-loop, red). The inactive A-loop displays large fluctuations, resulting in many crystallographic structures of inactive EGFR kinase missing a number of A-loop residues. Towards the N-terminal from this poorly resolved region exists a helical turn between the C-helix and the hinge region (see figure 2.4B; blue arrow), a feature only found in the inactive conformation, and has been suggested to act as a lock keeping the C-helix in its inactive “out” conformation[35]. Additionally, in the inactive structure, the P-loop (yellow) residue F723 packs against this helical turn, but tucks under the P-

loop in the active conformation. The active EGFR conformation is stabilised by a hydrophobic spine (residues L777, M766, D856 and H835), as has been found in many other kinases (Kornev 2006)[36], as well as the E762-K745 salt bridge, which helps maintain the C-helix “in” conformation.

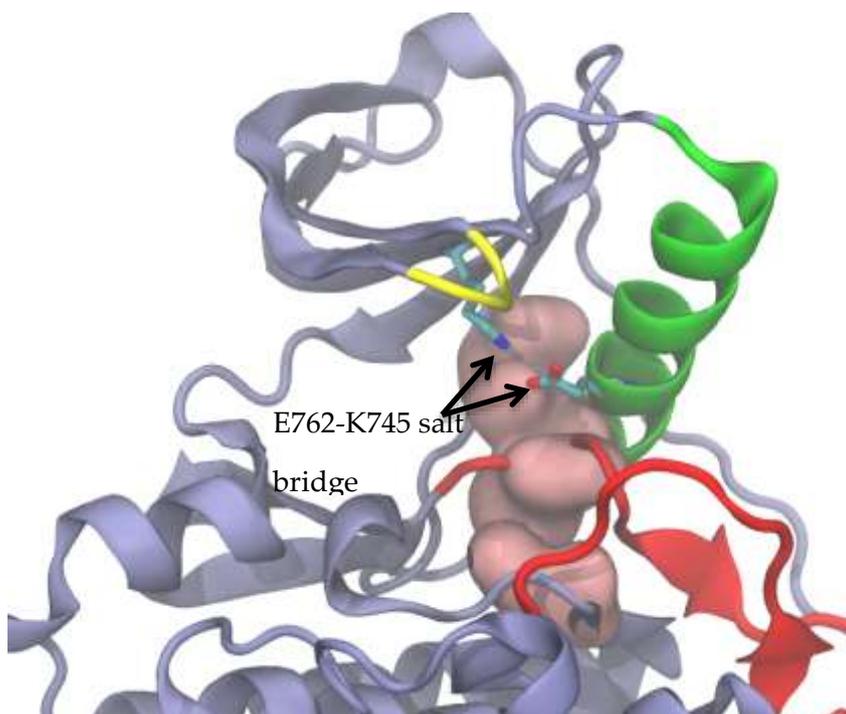


Figure 2.5: Interactions stabilizing the active conformation of EGFR kinase, including the E762-K745 salt bridge (labeled and shown as stick structures) and the hydrophobic spine (shown as a pink surface).

Zhang et al., drawing on biochemical and crystallographic evidence, proposed an asymmetric dimerisation mechanism whereby an activating monomer (activator) makes contact with a receiving monomer (receiver) using the activator’s C-lobe against the receiver’s N-lobe[29]. Later crystallographic structures revealed important interactions of the juxtamembrane (JM) region of EGFR that lies between the kinase and the transmembrane region [37]. It was found that the receiver’s JM region forming both close contacts with the activator’s C-lobe and a JM dimer with the activator’s JM

region, an interaction that has been coined the “juxtamembrane latch”. The JM latch has been implicated in the activation process of EGFR by helping to stabilise the asymmetric dimer[38]. Also, the presence of the juxtamembrane latch has been used to help explain why EGFR dimers that form in the absence of EGF remain inactive. These inactive dimers have an extracellular conformation that separates the JM regions of the EGFR dimer, preventing formation of the stabilising JM latch, and resulting instead in an inactive, symmetrical intracellular dimer[38].

### 2.2.1 Structural regulation of EGFR compared to other kinases

For most kinases phosphorylation of the A-loop is a necessity for activation; however, for EGFR kinase, this condition is not a requirement for kinase activity[17]. Nonetheless, the C-terminal tail of EGFR kinase contains 5 phosphorylation sites which, when phosphorylated, participate in the recruitment of proteins in the downstream signalling pathways[39].

The DFG loop is conserved across many kinase A-loops, with the DFG aspartate holding ATP in the correct position for catalysis through its interactions with Mg<sup>2+</sup>. The DFG is orientated differently in the active and inactive conformations, with the DFG aspartate often buried in the inactive conformation (“DFG-out”), and available for its stabilising interactions in the active conformation (“DFG-in”)[40]. In inactive EGFR, however, the A-loop adopts a src-like conformation where the aspartate of the DFG is not buried in the protein[41].

### 2.2.2 Overview of available structural data

Table 2.1 summarises the available crystal structures of EGFR kinase. The structures of the “closed” conformation correspond roughly to the “semi-closed” conformation observed in simulations by Sutto et al (2012)[42], as discussed later.

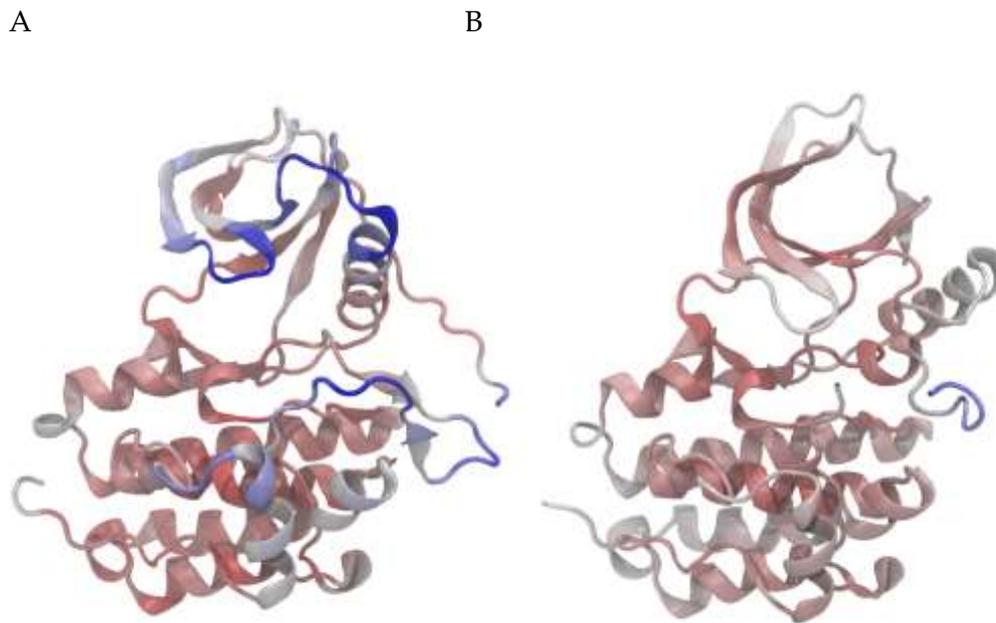


Figure 2.6: Crystal structures of the active (1m14) and inactive (2gs7) conformations of EGFR kinase coloured by B-factor (from 0 to 150, dark red to dark blue).

The crystal structures suggest that the inactive conformation, with the exception of the A-loop, is relatively rigid, with relatively low beta factors across the protein. On the other hand, the active conformation has high beta factors in the P-loop, A-loop, and the N-terminal side of the C-helix. It has been suggested that the high beta factors on the C-helix retards the ability of the unliganded WT to form the asymmetric dimer[43]

PDB ID	Ligand	Conformation	Mutation	Resolution (Å)
<a href="#">1M14</a>	None	Active	WT	2.60
<a href="#">1M17</a>	Erlotinib	Active	WT	2.60
<a href="#">1XKK</a>	Lapatinib	Inactive	WT	2.40
<a href="#">2EB2</a>	None	Active	G719S	2.50
<a href="#">2EB3</a>	ATP analogue	Active	L858R	2.84
<a href="#">2GS2</a>	None	Active	WT	2.80
<a href="#">2GS6</a>	ATP analogue	Active	WT	2.60
<a href="#">2GS7</a>	ATP analogue	Inactive	V924R	2.60
<a href="#">2ITN</a>	ATP analogue	Active	G719S	2.47
<a href="#">2ITO</a>	Iressa	Active	G719S	3.25
<a href="#">2ITP</a>	AEE788	Active	G719S	2.74
<a href="#">2ITQ</a>	AFN941	Active	G719S	2.68
<a href="#">2ITT</a>	AEE788	Active	L858R	2.73
<a href="#">2ITU</a>	AFN941	Active	L858R	2.80
<a href="#">2ITV</a>	ATP analogue	Active	L858R	2.47
<a href="#">2ITW</a>	AFN941	Active	WT	2.88
<a href="#">2ITX</a>	ATP analogue	Active	WT	2.98
<a href="#">2ITY</a>	Iressa	Active	WT	3.42
<a href="#">2ITZ</a>	Iressa	Active	L858R	2.80
<a href="#">2J5E</a>	13-JAB	Active	WT	3.10
<a href="#">2J5F</a>	34-JAB	Active	WT	3.00
<a href="#">2J6M</a>	AEE788	Active	WT	3.10
<a href="#">2JIT</a>	None	Active	T790M	3.10
<a href="#">2JIU</a>	AEE788	Active	T790M	3.05
<a href="#">2JIV</a>	HKI-272	Active	T790M	3.50
<a href="#">2RF9</a>	Mig6 peptide	Active	WT	3.50
<a href="#">2RFD</a>	Mig6 peptide	Inactive	K799E	3.60
<a href="#">2RFE</a>	Mig6 peptide	Inactive	K799E	2.90
<a href="#">2RGP</a>	Hydrazone	Inactive	WT	2.00
<a href="#">3BEL</a>	Oxime	Inactive	WT	2.30
<a href="#">3GOP</a>	None	Active	K721M	2.80
<a href="#">3GT8</a>	ATP analogue	Active	WT	2.95
<a href="#">3IKA</a>	WZ4002	Active	T790M	2.90
<a href="#">3LZB</a>	imidazo[2,1-b]thiazole	Inactive	V924R	2.70
<a href="#">3POZ</a>	tak-285	Inactive	WT	1.50
<a href="#">3VJN</a>	ATP analogue	Active	G719S, T790M	2.34
<a href="#">3VJO</a>	ATP analogue	Active	WT	2.64
<a href="#">3W2O</a>	tak-285	Active	L858R, T790M	2.35
<a href="#">3W2P</a>	Pyrimidine-based ligand	Active	L858R, T790M	2.05
<a href="#">3W2Q</a>	HKI-272	Active	L858R, T790M	2.20
<a href="#">3W2R</a>	Pyrimidine-based ligand	Active	L858R, T790M	2.05
<a href="#">3W2S</a>	Pyrimidine-based ligand	Inactive	WT	1.90
<a href="#">3W32</a>	Pyrimidine-based ligand	Inactive	WT	1.80
<a href="#">3W33</a>	Pyrimidine-based ligand	Inactive	WT	1.70
<a href="#">4G5J</a>	BIBW2992	Active	WT	2.80
<a href="#">4G5P</a>	BIBW2992	Active	T790M	3.17
<a href="#">4HJO</a>	Erlotinib	Inactive	V924R	2.75
<a href="#">4I1Z</a>	None	Closed	V948R, L858R, T790M	3.00
<a href="#">4I20</a>	None	Closed	V948R, L858R	3.34
<a href="#">4I21</a>	Mig6 peptide	Closed	V948R, L858R, T790M	3.37
<a href="#">4I22</a>	Gefitinib	Inactive	V948R, L858R, T790M	1.71
<a href="#">4I23</a>	Dacomitinib	Active	WT	2.80
<a href="#">4I24</a>	Dacomitinib	Inactive	T790M	1.80
<a href="#">4LI5</a>	Pyrimidine-based ligand	Active	WT	2.64
<a href="#">4LLO</a>	PD168393	Active	L858R, T790M	4.00

Table 2.1: list of PDBs of EGFR kinase, the type of ligand bound (if any), conformation, mutational state and resolution of the structures. Green entries represent those

structures used in the present study; blue entries represent those structures available when the simulations were being set up.

## 2.3 EGFR kinase mutations

As has been discussed previously, EGFR's position in the cell proliferation signalling network in conjunction with EGFR kinase mutations makes it a critical component in our understanding of a number of cancer types. Interestingly, despite the ability of activating mutations of EGFR kinase to transform normal cells into cancer cells [44], there is also a correlation between activating mutations and patient response to therapy, such that mutant EGFR kinase is more susceptible to inhibition by small drug molecules[45], [46], a behaviour that may be due to cells becoming dependent on anti-apoptotic signalling by mutant EGFR[46]. Additionally, despite the efficacy of inhibitors to the activating mutants, in most cases, patients develop resistance[47]. These characteristics of EGFR kinase mutants have made them a major focus for research and drug development. The following section will deal with a number of mutations, in relation to the structure of EGFR (discussed in the previous section).

## 2.3.1 L858R

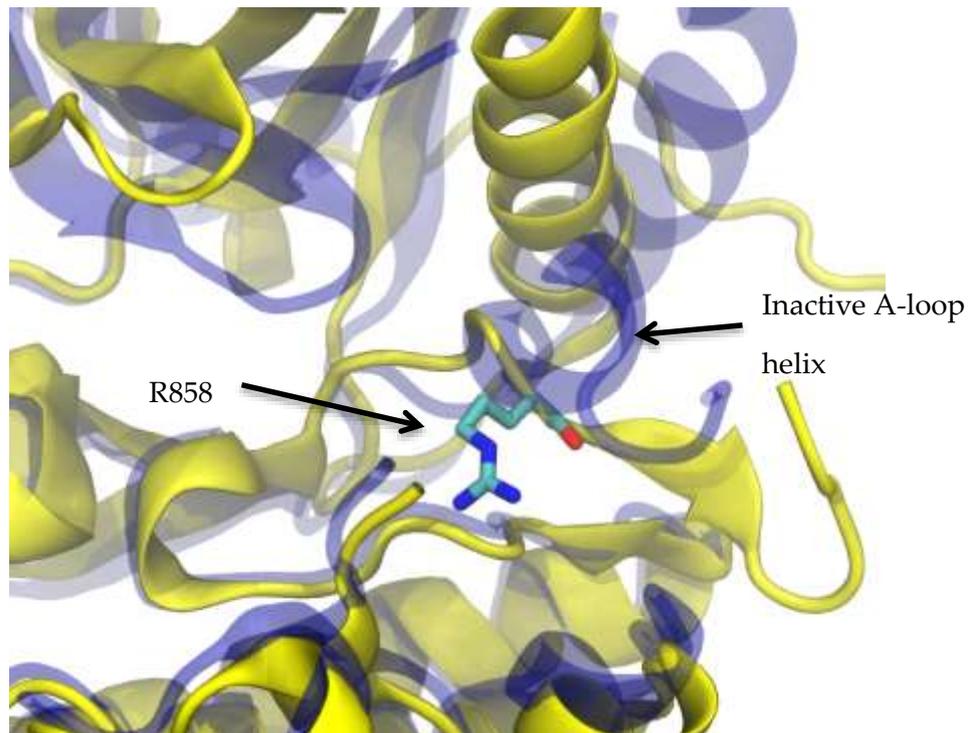


Figure 2.7: R858 (stick representation) in its structural context, with the L858R mutant structure (yellow; PDB code: 2ITV), and the WT inactive structure (blue; PDB code: 1XKK)

The L858R is the most prevalent single point activating mutation[47]. In inactive WT EGFR L858 is incorporated into the A-loop's hydrophobic helical turn and mutation of this leucine into a bulkier, charged arginine has been suggested to prevent the helical turn's formation, thus making the inactive conformation inaccessible, leading to activation[35]. Several studies have noted the increased catalytic activity of the L858R mutant [35],[29],[48], and results suggest that the mutation induces catalytic activity regardless of dimerisation, thus bypassing EGFR's normal signal-based regulation[29].

2.3.2 G719X

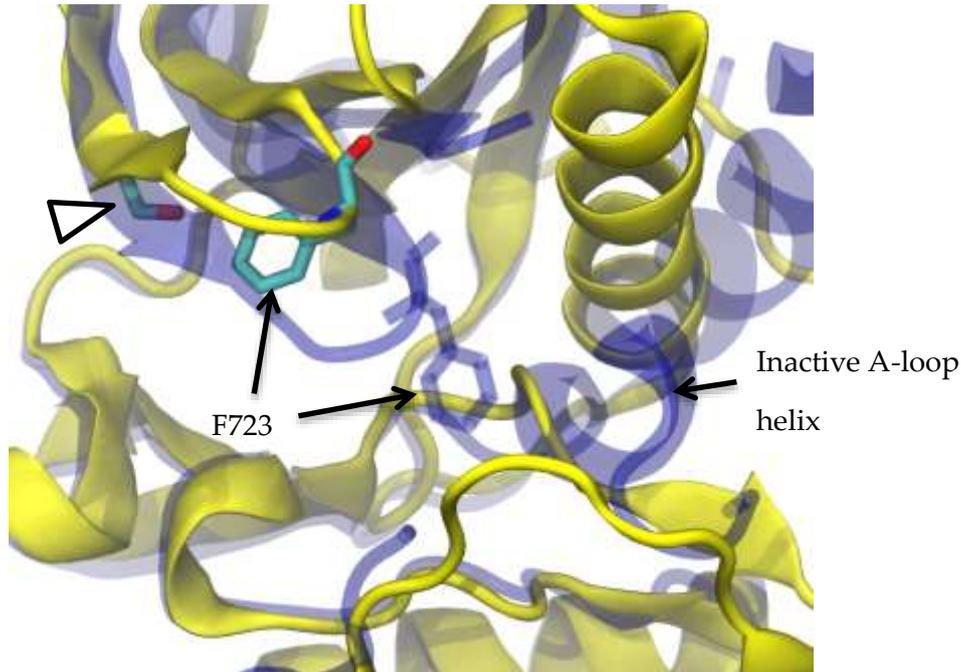


Figure 2.8: S719 (Arrowhead) in its structural context, with the G719S mutant structure (yellow; PDB code: 2ITP), and the WT inactive structure (blue; PDB code: 1XKK).

The G719X activating mutations occur on the P-loop of EGFR kinase, increasing its catalytic activity[35]. Its mechanism of activation has been suggested to be similar to that of L858R, despite its distance from the A-loop. The hypothesis set out by Yun et al. in 2007 states that mutations at this residue (which will introduce bulkier residues than the WT glycine) will reduce the flexibility of the P-loop, this in turn reduces the ability of the downstream F723 to pack against the hydrophobic helical turn on the A-loop (see figure 2.7), resulting in the reduction of stabilising interactions for the inactive A-loop, thus promoting conformational transition to the active conformation[35].

### 2.3.3 Exon 19 deletions

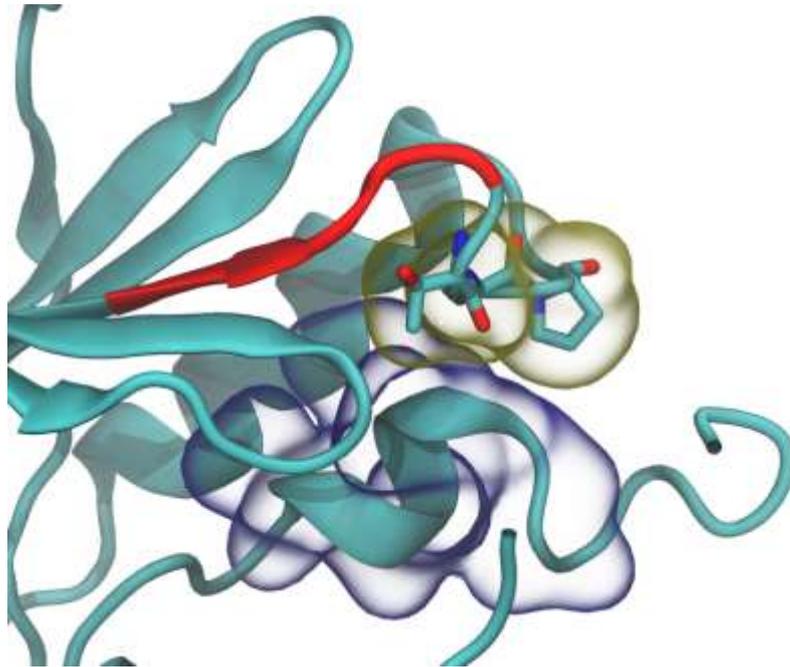


Figure 2.9: Inactive conformation of EGFR kinase with residues of the ELREA deletion highlighted in red, and the  $\alpha$ C- $\beta$ 4 loop represented as sticks. The surfaces of the  $\alpha$ C- $\beta$ 4 loop and A-loop helix are also shown in a glass surface representation (red and blue, respectively).

Exon 19 deletions account for almost half of EGFR kinase mutations[47], and these mutants respond significantly better to inhibitors than the L858R mutation (although the reason for this is unclear)[45]. There are a number of in-frame deletions that occur in exon 19 of the EGFR gene, which corresponds to a region close to the N-terminal side of the C-helix and is adjacent to a set of residues (the  $\alpha$ C- $\beta$ 4 loop) that pack against the A-loop's helical turn. Similarly to the G719S and L858R, this mutation appears to cause activation by disrupting the interactions of the A-loop helical turn[29][49]. Alternatively, it has been suggested that the deletion directly acts on the C-helix, effectively "pulling" it into the active conformation[50].

#### 2.3.4 T790M

The T790M mutation is a resistance mutation termed the “gatekeeper”, responsible for approximately 50% of cases of resistance to EGFR kinase inhibitors[47]. It was identified in 2005 by Kobayashi et al, after a patient exhibiting an exon 19 deletion stopped responding to gefitinib (known for being particularly potent against such mutants, see section 2.4), prompting the re-sequencing of the patient’s EGFR gene. Modelling the mutation into the crystal structure of EGFR kinase with erlotinib bound, Kobayashi et al. noted that the increased bulk of a methionine at position 790 clashed with the predicted surface of erlotinib. Additionally, it was noted that threonine at position 790 is able to make water mediated hydrogen bonds to erlotinib whereas methionine would not be able to[51].

Yun et al. produced crystal structures of the T790M mutant, and found the inhibitors could bind despite the T790M mutant, suggesting that steric hindrance was not the sole contributor to resistance. by analysing WT and mutant EGFR kinetics, the study identified a loss in ATP affinity associated with the L858R mutation, and found that the T790M mutation restored much of the loss in ATP affinity when introduced into an L858R mutant. The restoration of ATP affinity is not accompanied by a very large change in inhibitor affinity, and thus it appears that the T790M mutant confers resistance by enabling ATP to outcompete inhibitors at the binding site[52].

In 2009, Balias and Rizzo performed an in silico study to probe the effect of EGFR kinase mutations, and found that the L858R:T790M double mutant reduced the favourable electrostatic interactions with binding pocket waters along with disruption to the binding pocket water network[6].

### 2.3.5 Other mutations

The mutations described above account for the majority of known cases of EGFR kinase mutation, however there is a much wider spectrum of “uncommon” EGFR mutations comprised of both activating and resistance mutations. Many have been identified from patients, including the L747X mutations, V769M and A871E mutations[53], as well as a number of rare exon 20 mutants that have been shown to respond poorly to treatment with EGFR inhibitors[54]. Interestingly, some of these mutations were also found during in vitro resistance mutation screening, which highlighted a very broad spectrum of possible resistance mutation sites[3].

### 2.3.6 Mutations summary

EGFR kinase mutations have a significant impact on clinical outcomes and the efficacy of EGFR kinase inhibitors. However, it is worth noting that these mutations are not the only factors at work even in those cancers most commonly associated with them. Overexpression of EGFR is another common feature of a lung cancers [55], and EGFR interacts with other members of the ErbB family of kinases, in some cases mitigating the effect of mutation[3]. Additionally, mutations may occur in the extracellular region of the protein [56], such as have not been considered in this study. Nonetheless, the prevalence of EGFR kinase mutations and their interesting effects on inhibitor binding and activity makes the kinase mutations and important target for study. Additionally, it appears that the majority of activating EGFR kinase mutations have a conserved mechanism of action that pivots on destabilisation of the inactive A-loop helical turn, which provides a convenient starting point for investigation of EGFR kinase dynamics.

## 2.4 EGFR kinase inhibitors

EGFR signalling can be prevented artificially via a number of routes, including RNA interference[57], EGFR targeting antibodies and tyrosine kinase inhibitors (TKI). RNA interference, while promising, is still an emerging technology [58], whereas antibodies and TKIs have been used widely to treat cancer[59]. Of most importance to the current study are the 4-anilinoazoquinazoline TKIs, which are accommodated in the ATP-binding site of EGFR kinase[34],[41], [35], and found to be particularly effective against mutated forms of the kinase [60],[2],[45].

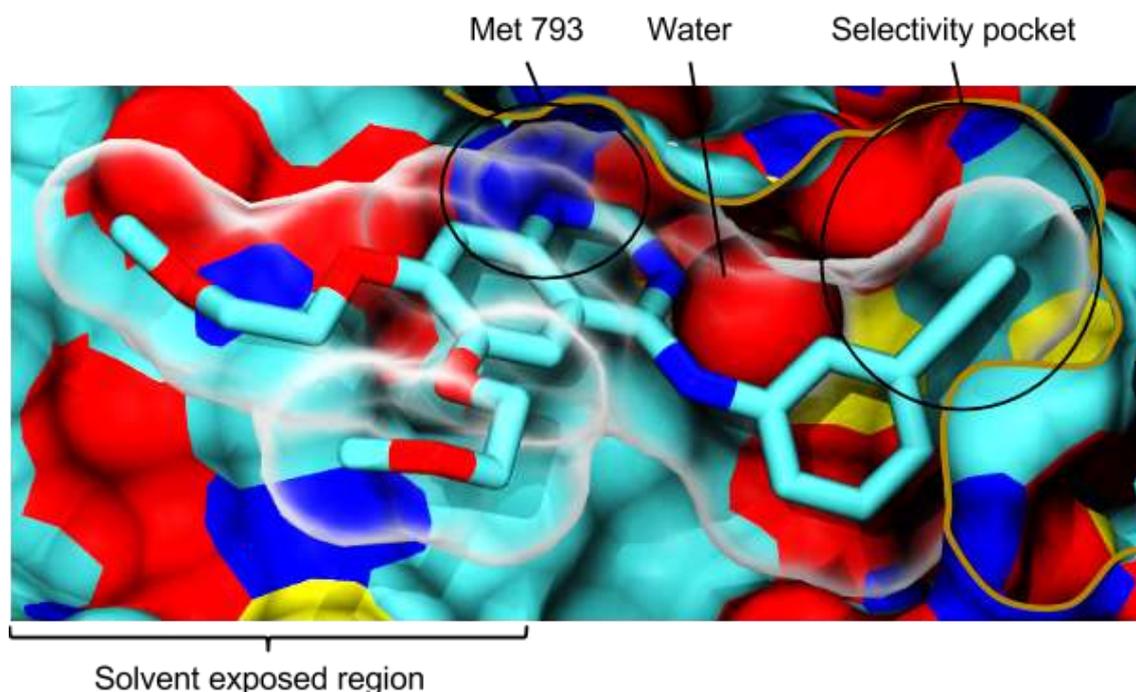


Figure 2.10: Erlotinib (sticks and translucent surface) bound to EGFR (opaque surface), with important regions of the binding pocket labelled. The ochre line highlights the point at which the protein surface intersects the page

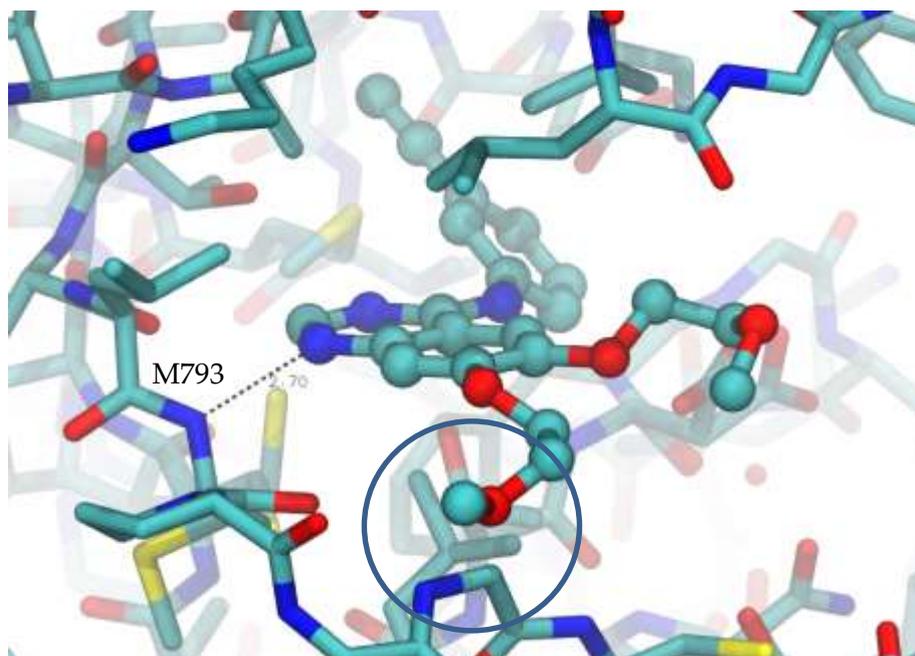


Figure 2.11: Coulombic interactions of Erlotinib (represented by sticks and balls) with EGFR kinase (represented by sticks): The hydrogen bond to the hinge region is marked as a dotted line, the circled region indicates where moieties in the solvent exposed region may interact with the solvated protein surface.

Analysis of the crystal structures shows that the 4-anilinoazoquinazolines make a number of important interactions with EGFR kinase[34]. Firstly, the quinazoline nitrogen opposite the aniline group is able to make a hydrogen bond to the backbone of the hinge region residue Met 793 (see figure 2.10). A second set of hydrogen interactions are made between the other quinazoline nitrogen and the binding pocket via water, the oxygen of which is within 4 Å of polar substituents of Gln 791, Thr 790 and Thr 854. There is also evidence for additional waters in the binding site[35][6], which would likely further stabilise the inhibitor in the pocket. Lastly, the anilino group extends into the back of the binding pocket, an area termed the “selectivity pocket”, which is also important for binding [3].

Initial crystal structures of tyrosine kinase inhibitors bound to EGFR kinase showed them occupying the ATP binding pocket of the active conformation[34]; however, tyrosine kinase inhibitors targeting the inactive conformation also exist (as in the case of lapatinib). Furthermore, recent studies suggest that inhibitors such as gefitinib, which were previously thought to bind exclusively to the active conformation, may bind inactive closed conformations EGFR kinase[61]. Given that these inhibitors can bind monomeric EGFR kinase domains, it seems likely that the binding of inhibitors is not restricted by extracellular or intracellular dimerization; however, there is still no data with regards to the difference in binding affinity of the active and inactive conformations.

## 2.5 Computational studies of EGFR

EGFR's importance as a cancer therapy target is reflected in the large body of computational research that has been carried out. This section will discuss the insights gained into both the conformational dynamics and inhibitor binding of EGFR kinase.

### 2.5.1 Insight into the WT dynamics of EGFR

Papakyriakou et al. (2009) carried out MD on the 5-20 ns time scale, for the WT, L858R and L861Q. An RMSD comparison between the active WT simulation and the inactive crystal structure suggested a "drift" toward the inactive structure, particularly prominent for the C-helix. Interestingly, simulation of the dimer did not exhibit this drift, and remained stable. Additionally, the K745-E762 salt bridge was found to be unstable in the WT monomer simulation, but stable in the dimer simulation[62]. More recent findings from a metadynamics study of EGFR kinase by Sutto et al. (2012)

revealed that the WT monomer naturally exhibits an inactive structure[42], further supporting the idea of the WT being autoinhibited.

The formation of the hydrophobic spine has been well documented both by Papakyriakou (2009) and Dixit and Verkhivker (2009) in their TMD studies into the active-inactive transition, and show that the hydrophobic spine is formed during the inactive to active transition, being formed alongside the K745-E762 salt bridge[62][63], which is perhaps unsurprising given that each of these structures have residues on the C-helix, which must rotate considerably to achieve the conformational transition. Papakyriakou et al. describe the movement of L858 away from the immediate area of the salt bridge to be a prerequisite for the formation of the above structures[62].

Mustafa et al. (2010) carried out MD 10 ns simulations on EGFR with the juxtamembrane (JM) and C-tail regions included, and found that motions of those structures correlate with an opening/closing motion in the protein, and that dimerisation reduced the closing of the activated member of the dimer[64]. The opening and closing of the N-lobe appears to be one of the main long time-scale motions of the protein, appearing in the 200 ns MD study by Wan et al. (2011) on both the WT and L858R mutant, even in the absence of the JM and C-tail regions[65], despite the assertion by Mustafa et al. (2010) that these regions were necessary for the opening/closing motion. Nonetheless, the opening/closing motion in Mustafa et al's (2010) study are interesting in that it appears the JM and C-tail segments allow the motion to occur over much shorter time scales[64].

From the above studies, it has been well established that dimerisation has a significant stabilising effect on the active conformation of the kinase, and while the position of the activating monomer in contact with the N-lobe of the activated monomer (see figure 2.3) obviously has implications for the dynamics of the N-lobe, Dixit and Verkhivker's

(2011) analysis of allosteric communication shows that dimerisation has a coupling effect on much more distant parts of the activated dimer, with the C-helix playing a central role in this coupling. Further, the T790M mutation was found to strengthen this coupling[66].

A number of studies have highlighted the hydrogen bonding patterns in EGFR[50], with Shih et al. (2011) providing perhaps the most comprehensive view of the hydrogen bonding patterns of the inactive and active conformations[50]. Seemingly contrary to the idea that EGFR kinase is autoinhibited in the WT, all computational evidence suggests that the active conformation of EGFR has the larger set of internal Coulombic interactions. This discrepancy appears to be due to the increased importance of hydrophobic interactions in the inactive monomer; an observation that, as Shih et al. (2011) [50] point out, is logical bearing in mind the hydrophobic nature of the allosteric dimer interface. However, it has been predicted that dimerisation of EGFR disrupts the hydrogen bonding network of inactive EGFR, and indeed, many of the inactive conformation's stabilising hydrogen bonds are in the region of the dimer interface. The importance of the EGFR's non-covalent bonding network is further highlighted by the tendency for some (but not all) mutations to reside at, or close to, residues involved in these hydrogen bonds.

The hydrophobic interactions within the HER family have also been studied[50], with the concept of inactive EGFR autoinhibition being supported by the observation that the more restrictively autoinhibited HER2 has a larger hydrophobic patch at the C-terminal side of the C-helix. Also, it has been predicted, using Targeted MD (TMD), that one of the first triggers of activation may be the formation of the hydrophobic spine[63].

Recent  $\mu\text{s}$  timescale MD studies into EGFR kinase by Shan et al. (2012) have shown that the C-helix exhibits disorder over long timescales, and that this disorder allows the C-helix “in” to “out” transition to occur on much shorter time scales than previously expected: in the  $\sim 100$  ns time scale. Evidence for this disorder was also found in crystal structures, which exhibit high B-factors in the N-terminal side of the C-helix[43].

Although the active conformation was found to be unstable, the inactive conformation appeared highly stable even after 100  $\mu\text{s}$ . Despite this, Shan et al. (2012) maintain that the disordered C-helix conformation is the dominant one, based on the slow binding kinetics of lapatanib, which they assume to require the inactive conformation due to its slow binding kinetics (with the conformational change to the inactive conformation being a possible rate-limiting step)[43].

Shan et al. (2012) noted that the disorder of the C-helix is suppressed by the dimer, as well as simulations in a mixture of water and isopropyl alcohol. The latter finding suggesting that the disorder is mediated by hydrophobicity, as the isopropyl alcohol clustered around the dimerization regions which includes C-helix[43].

A recent study from the same group rigorously reconstructed the entire EGFR protein in both its monomeric and dimeric forms (including an active, EGF-bound dimer and an inactive dimer). The simulations show that the intracellular region makes interactions with anionic species of the cell membrane, and that these interactions favour the inactive conformation. One implication of this is that overexpression of EGFR may lead to further activation by sequestering the excess of anionic lipids[32].

To summarise, the simulations of the dynamics of WT EGFR kinase are generally consistent, and complement the experimental observations: dimerization has a

stabilising effect on the active conformation of the kinase, and the active monomer appears to be relatively unstable, favouring an inactive C-helix “out” configuration; with the key K745-E762 salt bridge being disrupted easily in the WT. It appears that the kinase undergoes an opening/closing motion, which may be promoted by the JM segment and C-terminal tail, but is also present in longer time-scale metadynamics simulations.

### 2.5.2 Insights into the effect of EGFR mutations

The study Dixit and Verkhivker (2009) into EGFR mutation provides a multiple pieces of evidence for the theory of Yun et al. (2008) that activating mutations destabilise the inactive conformation. First, they demonstrated that the introduction of the L858R into the inactive conformation *in silico*, with subsequent temperature replica exchange dynamics results in a final conformation almost identical to that of the active L858R crystal structure. Secondly, in 10 ns simulations of the L858R and T790M mutations introduced into the inactive conformation, they observed increases in the RMSF of the C-helix and A-loop. Thirdly, a MM-GBSA study of protein stability suggested that the inactive conformation was less stable for the mutants. Additionally, TMD of the inactive conformation was used to pull the A-loop into the active conformation, with the L858R mutant completing the inactive to active transition much more rapidly[63].

The same study provides observations to support the complementary theory that activating mutations stabilise the active structure, both in the decrease in RMSF of the A-loop in the 10 ns active L858R simulation, and in the increased stability of the active L858R according to the MM-GBSA experiment[63].

An interesting complement to the above simulations is the TMD simulations in Papakyriakou et al. (2009), which show that the L858R mutant could not be forced into the inactive conformation using a  $2 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$  force constant. The L861Q could be forced into the inactive state, but did so over a slightly longer time scale than the WT. Papakyriakou et al. also noted the ability of R858 to interact with the glutamate that forms the K745-E762 salt bridge in the active structure. Since E762 is a C-helix structure, they reasoned that this interaction may promote transitions from inactive C-helix “out” to active C-helix “in” structures[62].

An interesting, but confusing aspect of the effect of the L858R on EGFR kinase dynamics was elucidated by the 200 ns simulations by Wan et al. (2011), which found that the active conformation was less stable for the L858R than the WT, and that the mutation biased towards a new, previously unrecognised conformation with an inactive C-helix orientation and an active A-loop conformation[65].

The results of the metadynamics study by Sutto et al. (2012) tie together the findings of Wan et al. (2011) and Papakyriakou (2009) in that while the L858R appears to have a deeper energy minima in the “semi-closed” conformation (with an inactive C-helix conformation, but an A-loop conformation unlike the inactive crystal structure) than the active conformation, the L858R is unable to visit the inactive conformation as exhibited by in the inactive crystal structures[42]. Additionally, the T790M was found to promote an active conformation, and when introduced into the L858R to produce a double mutant, the active conformation had a deeper minimum than with the L858R alone[42].

As stated previously, recent  $\mu\text{s}$  timescale MD studies into EGFR kinase by Shan et al. (2012) have shown that the C-helix exhibits disorder over long timescales. As well as elucidating the role of this disorder in the WT, they found that activating mutants act

to reduce the intrinsic disorder of the C-helix. Backed up with wet-lab experiments, they demonstrated that this stabilisation of the C-helix allowed for dimerisation to occur more readily, and proposed that this is one of the primary mechanisms of the activating mutants, based on their simulations of the deletion, L861Q, G719S, and S768I mutants[43].

Interestingly, Shan et al. (2012) does not cite Wan (2011), and these papers effectively contradict each other directly, with the former arguing for a stabilising effect of the L858R on the active conformation, and the latter noting a destabilisation. Nonetheless, Shan et al. (2012) have scrutinised the stability of the active conformation, demonstrating the L858R mutant to maintain its K745-E762 salt bridge for between 2-30  $\mu$ s.

Another interesting finding of Shan et al. (2012) was that the L858R mutant was well accommodated by the inactive conformation, and actually helps stabilise the inactive conformation, contrary to the orthodoxy, which predicted destabilisation.

To summarise the current literature on simulations of the EGFR kinase mutants, earlier studies appear to have favoured the simpler view of mutations stabilising the active conformation and destabilising the inactive conformation. More recent studies (utilising longer time scale simulations) propose a more complicated situation, with the identification of new minima corresponding to disordered C-helix configurations and a “semi-closed” conformation, with the mutants having a significant impact on the frequency with which these minima are visited. However, there is also inconsistency among these newer studies, with studies suggesting that either the L858R mutation destabilises or stabilises the active conformation.

### 2.5.3 Prediction of EGFR inhibitor binding affinity

There is a large body of literature for the prediction of EGFR inhibitor binding affinity, including QSAR[67][68], and docking [69],[70]–[72], often alongside wet-lab experiments. Liu et al. (2006) created a computational descriptor to predict the impact of mutation on binding affinity by computing the volume of a tetrahedron residing in the binding site: Each vertex of the tetrahedron lay on the midpoint of the line between two different residues[7]. In effect, the volume of the tetrahedron is proportional to the volume of the binding site, and so it seems that the degree to which mutants close the pocket impacts greatly on the binding affinity of inhibitors.

Particularly striking in the above study was the tendency for gefitinib to leave the pocket during 9 ns MD simulations of the T790M mutants[7]. Balias and Rizzo (2009) also investigated the impact of EGFR mutation on inhibitor binding affinity, and predicted an enhanced binding affinity due to the L858R activating mutation [6]. However, Balias and Rizzo went further, examining the effect of 4 mutants on 3 inhibitors with respect to the WT, using MM/GBSA on 5 ns of MD simulation, and were able to calculate the fold resistance (the factor by which inhibitor binding affinity is reduced by a resistance mutation) of most mutations to within approximately 1 kcal mol<sup>-1</sup> of the experimentally determined value. Intriguingly, Balias and Rizzo do not observe the movement of the ligand out of the pocket as in Liu et al's (2006) study, and note the impact of the mutation on the binding pocket waters and electrostatics rather than in terms of steric hindrance.

## 2.6 Conclusions

Despite the large body of computational work described briefly above, there is much still to learn of EGFR kinase dynamics. Is the sole effect of the mutations one of counteracting EGFR kinase's intrinsic disorder, or is there a noticeable impact of these mutations on the active and inactive conformations, as earlier studies suggest? Longer timescale simulations of the inactive conformation, combined with new analytical techniques may provide additional insight. The possibility of using new techniques for accelerating dynamics is another relatively unexplored option in the context of EGFR kinase.

A clear problem in simulating EGFR is accessing the long time scales of its dynamics, which have been largely out of the grasp of even the longest MD simulations of specialist supercomputers such as ANTON[43], until very recently[5]. To combat this, a number of studies have implemented enhanced sampling methods, however these have only been carried out at relatively short time scales (less than 100 ns) or by directing the system along reaction coordinates (as in the extensive TMD simulations of EGFR) that risk failure to sample important states, though metadynamics has been shown capable of sampling the disordered C-helix state[42].

Discussion of the role of each of the key structural features of EGFR kinase in activation, and the ability of mutations to perturb them, has yet to reach a consensus. Consequently, any MD techniques that can be used to target either the individual structures or simply to simulate a wider region of conformational space seem likely to provide additional insight. In fact, the above discussion shows that disagreement is still ongoing as to whether the activating mutants destabilise the inactive conformation[65],[42] or not[43], and whether the mutants stabilise the active conformation[43] or some conformation close by [42],[65].

The present study focuses on utilising new analysis techniques and enhanced sampling methods that do not require knowledge of the activation pathway, and in doing so investigate the ability of the current software, hardware and data available to most computational scientists to elucidate the impact of mutations on long timescale dynamics.

In addition, although EGFR inhibitor binding affinity has been widely investigated (as discussed above), no rigorous free energy calculations have been performed to date. Thus, there is still much scope for applying rigorous techniques to the system, not only to get a more detailed picture of the binding site of EGFR kinase, but also to investigate how binding is affected by mutation, and to test the plausibility of using such calculations in drug discovery.

### **2.6.1 Implications on the experimental design**

As previously discussed, approximate methods for the prediction of inhibitor binding affinities have been widely applied to EGFR kinase; however, as of yet rigorous comparison of EGFR kinase inhibitors by techniques such as thermodynamic integration and free energy perturbation are lacking. Thus the question of whether these more rigorous techniques are of use in the context of EGFR kinase remains open, and the present study aims to help fill this gap by carrying out replica exchange thermodynamic integration on EGFR kinase inhibitors. The primary aim of experiments outlined in chapter 4 was the robust prediction of EGFR kinase inhibitor binding affinities.

Additionally, while work has begun on the prediction of the effect of EGFR kinase mutations on binding affinity[6],[7], no rigorous approaches have been attempted. A secondary aim of the experiments was to reproduce and extend this work using rigorous methods; however, successful prediction of EGFR kinase inhibitors was a prerequisite for initiation of this work.

The previous discussion also outlines several gaps in our understanding of EGFR kinase dynamics. The aim of the experiments in chapter 5 was to attempt to fill these gaps using freely available or cheap computational resources, and to determine how useful a range of enhanced sampling methods are in building a picture of kinase dynamics. Specifically, the analyses presented in chapter 5 aim to address the issue of intrinsic disorder as a method of EGFR kinase regulation and its counteraction by mutation (as set out by Shan et al. (2012)[43]), and the issue of whether activating mutations have a stabilising/destabilising effect on the active/inactive conformations.

To achieve the above aim, molecular dynamics simulations were performed. To validate these simulations, as well as provide possible insight, RMSD and RMSF analyses were employed, which provide a quick means of comparing the characteristics of the sampling between simulations, and how differences in sampling is related to specific structures within the kinase domain. To complement these analyses dimensionality reduction techniques have been employed in an effort to map out the sampling of the protein with respect to A) the differences between the active and inactive conformations and B) the most important motions of the protein, if possible. These dimensionality reduction techniques also have the advantage of identifying which motions are most dominant in a subset of trajectories, depending on how they are implemented. Investigating the issue of disorder in the C-helix requires analysis of protein secondary structure, which was also carried out. In addition to these analyses, trajectory visualisation, particularly when guided by the other analyses, has

been employed to gain understanding into the conformational dynamics of EGFR kinase.

Another important aim of the experiments in chapter 5 was the evaluation of enhanced sampling methods, particularly the degree to which those methods could increase the amount of sampling while avoiding unrealistic sampling. To this end, a comparison of each of the MD analyses mentioned above in addition to clustering analysis was carried out, primarily for comparison with conventional MD.



## Chapter 3: Computational methods

As discussed in the previous chapter, computational methods can provide atomistic insight into protein systems. However, it is important to ensure that whatever models are produced are sufficiently grounded in reality to provide meaningful insights. Additionally, once a model is obtained, it is necessary to apply suitable analyses to them; otherwise no insight can be gained at all.

This chapter will discuss the theory of the methods applied in the present study, particularly Molecular Dynamics (MD) and Monte Carlo (MC), and the statistical mechanics from which both of these methods gain their utility. Extensions to both of these methods will also be discussed including Quantum Mechanics/Molecular Mechanics hybridisation (QMMM) of MC, and enhanced MD sampling methods.

The analytical methods will also be discussed, most prominently dimensionality reduction techniques, due to their utility in guiding the understanding of large datasets such as those produced in the present study.

### 3.1 Statistical mechanics

It goes without saying that drug binding and mutation can have dramatic macroscopic effects, but these effects are somehow propagated on the atomistic level. Statistical mechanics provides a route for bridging macroscopic observation with atomistic reality, and is the basis of all the simulations discussed in this thesis.

Essentially, statistical mechanics is an application of our understanding of probability to atomistic systems, much of the following portrayal and equations are derived from Chang's *Physical Chemistry for the Chemical and Biological Sciences*[73]. To illustrate, we can imagine a box containing a gas of  $N$  particles. There is no physical law to prevent a state where all  $N$  particles reside in the left side of the box, in fact if  $N=1$ , this would occur half the time, with the other microstate (with the particle in the right of the box) occurring the rest of the time. If we increase  $N$  to 2, then we create a new state where one particle is in the left side, and one in the right side. However, the new state can be arrived at by either the first particle being in the right side and second particle in the left side or visa versa, thus the new macrostate is formed of two microstates. This new macro state is more likely to occur than the other states, which each require having both particles in a particular side of the box, simply because it has more microstates.

As  $N$  increases, not only does the number of microstates increase, but the number of microstates within the most probable macrostates increase, due to the following principle, first derived by Boltzmann:

$$W = \frac{N!}{\prod_i n_i} \quad (3.1)$$

Where  $W$  is the number of microstates in a macrostate,  $N$  is the number of particles,  $n_i$  is the number of particles occupying a specific region ( $i$ ) of the box, and  $0! = 1$

Evidently,  $W$  reaches a minimum when all the particles occupy the same region of the box, and a maximum when the particles are spread evenly through all regions of the box. Of course, this simple view of the particles in a box is not sufficient for describing realistic problems as it does not account for interactions within the system, however it does indicate that there is some set of configurations that have a greater contribution to

the properties of the system, implying that simulations need to take this tendency into account.

From the above, we are interested in finding the most probable configurations of the system (that is, those which contribute the most to the properties of the system). The logarithm of equation 3.1 (being monotonic with that equation) can be used to explore this problem better than equation 3.1 itself:

$$\ln W = \ln N! - \ln \sum_i \ln n_i! \quad (3.1)$$

Which can be expressed using Stirling's approximation (since  $n_i$  is a large number) to give:

$$\ln W = N \ln N - N - \sum_i (n_i \ln n_i - n_i) \quad (3.2)$$

Assuming our toy system is comprised of a constant number of particles:

$$\sum_i n_i = N \quad (3.4)$$

Also, since we are interested in finding the maximum of  $\ln W$  its derivative is 0, and equation 3.3 can be expressed as:

$$d \ln W = - \sum_i \ln n_i dn_i = 0 \quad (3.5)$$

Equally, if each particle  $i$  has an energy  $\varepsilon_i$  such that the total energy  $E$  is the sum of the energy of each particle:

$$\sum_i n_i \varepsilon_i = E \quad (3.6)$$

Bearing in mind the derivative of a constant is 0, it can be subtracted from the other terms of equation 3.5, thus incorporating the energetics of the system. However, is necessary to use Lagrange's method of undetermined multipliers with the constraints (equations 3.6 and 3.4), which produces the following equation:

$$\sum_i -(\ln n_i dn_i + \alpha - \beta \varepsilon_i) dn_i = 0 \quad (3.7)$$

Or:

$$\ln n_i = \alpha - \beta \varepsilon_i \quad (3.8)$$

$$n_i = e^{\alpha - \beta \varepsilon_i} \quad (3.9)$$

We can substitute equation 3.9 into equation 3.4 to attain

$$\sum_i e^{\alpha - \beta \varepsilon_i} = N \quad (3.10)$$

Thus we arrive at the Boltzmann distribution:

$$\frac{n_i}{N} = \frac{e^{\beta \varepsilon_i}}{\sum_i e^{\beta \varepsilon_i}} \quad (3.11)$$

It should be noted that while  $\beta$  has been referred to only in terms of its role as a Lagrange multiplier, here  $\beta$  takes the form  $\frac{1}{k_B T}$  and is known as the *thermodynamic*  $\beta$ ,

and that the temperature ( $T$ ) has been considered constant throughout. Under these conditions, equation 3.11 expresses the probability of finding the system in state  $n_i$ .

### 3.1.1 Statistical ensembles

A representative subset (or all) of the possible configurations of a system is collectively called an *ensemble*. In the previous section, the Boltzmann distribution was outlined for a system under constant volume, temperature and number of particles, this corresponds to the *canonical ensemble*.

However, there are systems for which we might not wish to apply these specific restrictions, and thus we can look to the grand canonical, microcanonical or more generalised ensembles.

The grand canonical ensemble fixes the chemical potential of a system, rather than the number of particles, thus allowing the investigation of adding or removing particles; however this also necessitates modification of the partition function to take into account the relative ease of introducing new particles.

The microcanonical ensemble (also NVE ensemble) keeps the total energy of the system fixed along with the number of particles and volume. It is appropriate for looking at isolated systems, and is useful for determining a number of thermodynamic properties, or when a system's density requires equilibration (as is impossible in the canonical ensemble).

Additional ensembles can also be constructed: For example, parallel tempering (PT) utilises multiple NPT simulations at different temperatures while allowing periodic

(and conditional) exchange of the systems. This allows the higher temperature simulations to cover a large conformational space without being caught in the minima dominating the lower temperatures, while at the same time allowing the high-temperature configuration to exchange to lower temperatures which are likely to be of more interest.

### 3.1.2 Assumptions

The power of statistical mechanics relies on two main assumptions: The *postulate of a priori probabilities* and the *postulate of ergodicity*[74].

The *postulate of a priori probabilities* states that there is an equal probability of reaching any given microstate with an equal energy. In other words, if  $E_i = E_j$ , then probability  $P_i = P_j$ .

The *postulate of ergodicity* states that the time average of quantities of a system is equal to the ensemble average quantities of the system, because a system evolving over an infinite time scale will theoretically explore all possible microstates.

### 3.1.3 The partition function

From a probabilistic viewpoint, with the above assumptions, we are now in a good position to attempt to extract properties from a system. The partition function is a mathematical tool that can be used to do so. If the number of particles, volume and temperature are constant, the partition function ( $Z$ ) is:

$$Z = \sum_i e^{-\beta \varepsilon_i} \quad (3.12)$$

This is an integral part of the Boltzmann distribution (see equation 3.11), and can be used directly to obtain properties from a system. For example, the total energy of a system under constant number of particles, temperature ( $T$ ), and volume ( $V$ ) would be calculated like so:

$$E = \frac{k_B T}{Z} \left( \frac{\delta Z}{\delta T} \right)_V \quad (3.13)$$

However, while this relationship holds true for quantum systems, which can take discrete energy states, a classical treatment requires continuous properties for the particles, effectively producing an infinite set of microstates. The classical partition function overcomes this problem via integration:

$$Z = \frac{1}{N! h^{3N}} \iint d\mathbf{p}^N d\mathbf{r}^N e^{-\beta E(\mathbf{p}^N, \mathbf{r}^N)} \quad (3.14)$$

Where  $h$  is the Planck length,  $\mathbf{p}$  is momentum, and  $\mathbf{r}$  positions of the particles. The denominator originates in part from both the indistinguishability of the particles, and as a whole ensures comparability between classical and quantum calculations.

An experimental measurement ( $A_{ave}$ ) can be expressed as a function of the positions and momenta of all the particles over the time the measurement was made:

$$A_{ave} = \int_{t=0}^{t=1} A(\mathbf{p}^N(t), \mathbf{r}^N(t)) dt \quad (3.15)$$

According to the *postulate of ergodicity*, as  $t$  approaches infinity, the observable property approaches the ensemble average. This ensemble average can be calculated from the following:

$$A_{obs} = \iint d\mathbf{p}^N d\mathbf{r}^N A(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{p}^N, \mathbf{r}^N) \quad (3.16)$$

Where the probability density,  $\rho(\mathbf{p}^N, \mathbf{r}^N)$ , is the probability of finding a configuration of the system with the given momenta and positions, and is given by the Boltzmann distribution:

$$\rho(\mathbf{p}^N, \mathbf{r}^N) = \frac{1}{Z} e^{-\beta E(\mathbf{p}^N, \mathbf{r}^N)} \quad (3.17)$$

So long as the simulation samples with a probability density such as that above, it is possible to simply derive properties as the average of observations made during the simulation. In the case of the internal energy:

$$U = \langle E \rangle = \frac{1}{M} \sum_{i=1}^M E_i \quad (3.18)$$

Where  $M$  is the number of observations made during the simulation.

## 3.2 Molecular dynamics

Molecular dynamics (MD) generates an ensemble of configurations by allowing the system to evolve through time. Given sufficient time, the system will produce an ensemble from which meaningful thermodynamic properties can be extracted as discussed previously (equation 3.18). Additionally, because MD explores configurations of the system as a function of time, it is possible to elucidate dynamics such as conformational change due to activation events or inhibitor binding etc. As such MD has become a useful tool for investigating a variety of systems.

### 3.2.1 Newton's second law

Molecular dynamics can be thought of as solving Newton's laws of motion for a system. Newton's second law states:

$$\mathbf{F} = m\mathbf{a} \quad (3.19)$$

Where  $\mathbf{F}$  is the vector force acting on a particle,  $m$  is the particle's mass, and  $\mathbf{a}$  is the vector acceleration provided by the force. Given the atomic coordinates of a system, an approximation for the force acting on a particle can be found using the classical potential (discussed later), which can then give the corresponding acceleration of the particle. With this information, it is possible to calculate the positions of the atoms after a period of time  $\delta t$  has elapsed by integration of the acceleration. Specifically, the majority of the present study utilises the leapfrog algorithm (as implemented in the

sander program of the AMBER package). The *leapfrog* algorithm calculates the new positions of the particles like so:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \frac{1}{2}\delta t\mathbf{v}(t + \frac{1}{2}\delta t) \quad (3.20)$$

Where  $\mathbf{r}$  is the position of a particle,  $\mathbf{v}$  is the velocity of the particle, which is calculated at  $t + \frac{1}{2}\delta t$  as follows:

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t - \frac{1}{2}\delta t) + \delta t\mathbf{a}(t) \quad (3.21)$$

Where  $\mathbf{a}$  is the acceleration of the particle.

The above equations allow the system to evolve over time, by alternately calculating the velocities then positions (hence the name “*leapfrog*”). Evidently this scheme requires some initial velocities, which are generated randomly to correspond with the desired temperature of the system.

The current work also utilises the *velocity Verlet* algorithm (for those simulations using NAMD), which calculates the new positions of the particles like so:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) \quad (3.22)$$

Evidently, to implement the above equation, it is necessary to calculate the velocities for each step, which is done by first calculating the velocities after a half timestep:

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t\mathbf{a}(t) \quad (3.23)$$

Then the final velocity is calculated like so:

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t + \frac{1}{2}\delta t) + \frac{1}{2}\delta t\mathbf{a}(t) \quad (3.24)$$

The *velocity Verlet* approach is useful as it calculates the velocities at the same time step during which the positions are calculated, unlike the *leapfrog* algorithm, which requires extra calculation if the velocities at a particular time step are required.

### 3.2.2 Timesteps

Evidently, while the force acting on the particle will be dependent on the potential discussed earlier, it is necessary to choose an appropriate time step for the system in question. A low time step may be computationally wasteful if the time step is much less than the rate at which acceleration varies between time steps. A high time step may also lead to problems associated with particles travelling in areas where they may experience large discontinuous forces, leading to a loss of conservation of energy.

In the current work a time step of 2 fs is used unless otherwise stated, which is made possible by using the SHAKE algorithm to constrain bonds involving hydrogen, as these fluctuate at a very high frequency while contributing little to the longer time-scale dynamics of the system. Including the high frequency motions of hydrogen would require a smaller time step, leading to a reduction in the ability to sample the longer time scale dynamics of the system in which we are generally interested.

### 3.2.3 Thermodynamic conditions

Because the current work aims to gain insight into biological processes, it is necessary to adjust the simulation parameters such that they best represent the system in a natural environment. Temperatures of 300 K have been used with 1 atm of pressure, however the maintenance of these conditions deserves some explanation.

Various methods for controlling temperature exist. The simplest is velocity scaling, where particle velocities are scaled in to correct the deviation of the system temperature to the reference temperature. Other methods include those utilising heat baths such as the Berendsen thermostat, Andersen thermostat and Nosé-Hoover thermostat.

In the current work temperature has been regulated using a Langevin thermostat. This introduces additional acceleration randomly to particles in the system to simulate collisions with a heat bath. These accelerations are provided by forces randomly generated from a Gaussian distribution appropriate for the desired temperature. Additionally a damping constant is introduced to simulate a friction force ( $F_{friction}$ ):

$$F_{friction} = -\xi v \quad (3.25)$$

Where  $v$  is the velocity of a particle, and  $\xi$  is a frictional constant, which is proportional both to the collision frequency and the mass of the particle. This frictional force dampens the impact of higher velocity collisions over normal simulation time scales, and thus ensures the system is maintained at a particular temperature; however, the possibility of higher velocity particles provided by Langevin dynamics increases jostling of the solute, which may enhance barrier crossing[75].

Except for some short time scale fluctuations, the above scheme should maintain a constant temperature; however, it is particularly important to be able to monitor the temperature, especially for ensembles like the NVE ensemble, where temperature can fluctuate over time. The temperature is a function of the total kinetic energy of the system ( $E_{kin}$ ):

$$T = \frac{2 E_{kin}}{3 N k_B} \quad (3.26)$$

$$E_{kin} = \sum_{i=1}^N m_i v_i^2 / 2 \quad (3.27)$$

Where  $v_i$  is the magnitude of the velocity  $\mathbf{v}_i$  of particle  $i$ , and  $m_i$  is the mass of particle  $i$ .

When a fixed pressure is required a barostat is used whereby the volume of the system is scaled along with the positions of the particles. In the current work, this is achieved using a weak-coupling Berendsen barostat[76]. A scaling factor ( $\lambda$ ) is used to scale the positions of the molecules in the simulation at each time step:

$$\lambda = \left\{ 1 + \frac{\delta t}{\tau_p} (P(t) - P_{bath}) \right\}^{1/3} \quad (3.28)$$

Where  $\tau_p$  is a relaxation constant, and  $P(t)$  is the pressure at time  $t$ , and  $P_{bath}$  is the pressure of the pressure bath.

Regulation of pressure is requires a reliable method of calculating the pressure. In an ideal gas, this is can be obtained from the ideal gas law; however, the interactions between the particles cause a significant deviation from the ideal gas law. Instead, the

pressure must be calculated as a function of the positions of the particles and the forces acting on them:

$$P = \frac{1}{V} \left[ Nk_B T - \frac{1}{3} \sum_{i=1}^N \sum_{j=i+1}^N r_{ij} \mathbf{f}_{ij} \right] \quad (3.29)$$

### 3.3 Monte Carlo

Unlike MD, Monte Carlo (MC) simulations do not follow the evolution of a system through time, instead progressing in a stochastic manner. In brief, the system is allowed to undergo random changes in configuration along predetermined degrees of freedom to produce an ensemble from which thermodynamic properties can be determined.

#### 3.3.1 The Metropolis test

While it is possible to make random changes to a system to generate new conformations, then weight these conformations by the Boltzmann factor ( $\exp(E/k_B T)$ ), this will lead to a large number of unfavourable, heavily weighted configurations that do not contribute significantly to the dataset.

Metropolis et al (1953) developed a method whereby each configuration could be weighted evenly by preferentially sampling the more favourable regions of phase space. In a Metropolis MC simulation the energy of new conformations is tested; if the energy of the new conformation is lower than the current conformation, it is always accepted. Otherwise, the following condition is required for acceptance:

$$\xi_a < e^{-\Delta E/k_B T} \quad (3.30)$$

Where  $\xi_a$  is a random number between 0 and 1, and is regenerated for each configuration. If the above condition is not met, then the new configuration is discarded.

In this way, a Metropolis MC simulation samples those regions of phase space with the greatest contribution to the integral, and thus converged results can be obtained in much shorter time scales.

### 3.3.2 Observing detailed balance

According to the principle of detailed balance, at equilibrium, a system should be able to transition from the arbitrary equilibrium configuration  $m$  to the equilibrium configuration  $n$  and vice versa[74]:

$$\pi_{mn}P_m = \pi_{nm}P_n \quad (3.31)$$

Where  $\pi_{mn}$  is the probability of making a transition from configuration  $m$  to configuration  $n$ ,  $\pi_{nm}$  is the reverse transition,  $P_m$  is the probability of the system being in configuration  $m$ ,  $P_n$  is the probability of the system being in configuration  $n$ .

Despite the constraint applied to the acceptance of MC simulations by the Metropolis test, it is possible to demonstrate that Metropolis MC still obeys detailed balance by rearranging the above equation, and substituting  $\pi$  for the Metropolis acceptance probability:

$$\frac{p_n}{p_m} = \frac{\pi_{mn}}{\pi_{nm}} = \frac{e^{(-E_n/k_B T)}}{e^{(-E_m/k_B T)}} = e^{(-E_n - E_m/k_B T)} \quad (3.32)$$

If  $e^{(-E_n/k_B T)} > e^{(-E_m/k_B T)}$  then

$$\frac{\pi_{mn}}{\pi_{nm}} = \frac{e^{(-E_n - E_m/k_B T)}}{1} = e^{(-E_n - E_m/k_B T)} \quad (3.33)$$

If  $e^{(-E_n/k_B T)} < e^{(-E_m/k_B T)}$  then

$$\frac{\pi_{mn}}{\pi_{nm}} = \frac{1}{e^{(-E_m - E_n/k_B T)}} = e^{(-E_n - E_m/k_B T)} \quad (3.34)$$

Thus the ratio of the probabilities of sampling between two specific configurations is independent of how favourable either configuration is.

### 3.3.3 Generating configurations

To generate configurations as outlined above, it is necessary to specify degrees of freedom to alter during each iteration of MC. MC moves can include molecular rotations and translations as well as internal flexing: bond stretching, angle bending and torsion of dihedrals.

It is desirable to avoid very large changes each iteration, as this will lead to the majority of configurations being discarded (due to high-energy atomic overlaps), and thus less phase space will be explored. Equally, flexing by too little will lead to less efficient sampling of phase space. An additional complication arises for larger molecules, where a small change in dihedral may lead to downstream atoms moving a large distance. The amount by which each degree of freedom should be allowed to change each iteration can be probed by analysing move acceptance rates using short test runs.

### 3.3.4 Grand Canonical Monte Carlo

Grand Canonical Monte Carlo (GCMC), with its ability to insert or remove waters from a simulation, is useful in determining the hydration of protein binding pockets[77][78], which is the capacity in which it has been used in the present study.

GCMC samples the Grand Canonical ensemble, where the number of particles ( $N$ ) is allowed to fluctuate, while the chemical potential ( $\mu$ ), volume ( $V$ ), and temperature ( $T$ ) are kept constant. To enable changes in  $N$ , two new types of move are introduced: insertions and deletions. The chemical potential scales with the chemical activity  $z$  like so:

$$z = \frac{e^{\mu/k_B T}}{\Lambda^3} \quad (3.35)$$

Where  $\Lambda$  is the de Broglie wavelength. The acceptance tests for an insertion and deletion (respectively) are[79]:

$$\xi < \left(1 + \frac{N+1}{zV} e^{\Delta E/k_B T}\right)^{-1} \quad (3.36)$$

$$\xi < \left(1 + \frac{zV}{N} e^{\Delta E/k_B T}\right)^{-1} \quad (3.37)$$

Where  $\Delta E$  is the change in energy due to the insertion or deletion, and  $\xi$  is a random number between 0 and 1. In the present work, the Adams parameter ( $B$ ) is used instead of  $z$ . This parameter is related to the excess chemical potential ( $\mu_{ex}$ ) like so[80]:

$$B = \frac{\mu_{ex}}{k_B T} + \ln \bar{n} \quad (3.38)$$

Where  $\bar{n}$  is the expected number of particles in the simulated region:

$$\bar{n} = \bar{p}V \quad (3.39)$$

Here,  $\bar{p}$  is the number density of the particles. Because B scales with the excess chemical potential, a simulation of constant B is equivalent to a simulation with constant  $z$ [79].

By simulated annealing of B values, it is possible to calculate the binding free energy ( $\Delta G_{bind}$ ) of water (or other particles) using the following relationship[81]:

$$\Delta G_{bind} = \Delta G_{hyd} + k_B T (B - \ln \bar{n}) \quad (3.40)$$

Where  $\Delta G_{hyd}$  is the hydration free energy of the water. The value of B that provides an occupancy of 0.5 for a water will thus provide the binding free energy.

### 3.3.5 Just Add Water Molecules

Just Add Water Molecules (JAWS) is a technique first described by Michel et al (2009)[82]. A JAWS simulation, like a GCMC simulation can be used to investigate the free energy associated with the insertion/deletion of a particle; however, instead of inserting and deleting water molecules, water molecules are allowed to sample a new degree of freedom,  $\theta$ . This scaling factor linearly interpolates the water's intermolecular interactions, such that a  $\theta=0$  water does not interact with the system, and a  $\theta=1$  water interacts fully with the system.

Putative water sites can be identified by letting these  $\theta$ -waters sample around a grid that corresponds to the region of interest. This sampling consists of rotations, translations and changes in  $\theta$ . The probability density of  $\theta$ -waters sampling high values of  $\theta$  (above 0.95) across the grid provides a map of where waters can interact favourably, and where they cannot.

The binding affinity of waters at these putative water sites can be estimated by separate simulations whereby a  $\theta$ -water is inserted into the site, and restrained in a  $3 \text{ \AA}^3$  portion of the site. This  $\theta$ -water is then allowed to sample as previously, but only within the putative site. The final binding free energy of the water at this site is calculated by computation of the ratio of frames for which the water was “off” or “on”, as follows:

$$\Delta G_{bind} = -k_B T \ln \left( \frac{P(\theta > 0.95)}{P(\theta < 0.05)} \right) \quad (3.41)$$

Where  $P(\theta > 0.95)$  is the probability of the water having a  $\theta$  value of above 0.95, and  $P(\theta < 0.05)$  is the probability of the water having a  $\theta$  value of below 0.05. However, in the above implementation the water is introduced from the gas phase, and so a biasing potential  $V(\theta_i)$  is applied which not only accounts for the free energy of hydration (to get a true estimate binding affinity), but also accounts for the energy required to restrain the water in the volume  $V^{\text{constr}}$ :

$$V(\theta_i) = (-\Delta G_{hyd} + \Delta G_{constr})\theta_i \quad (3.42)$$

Where  $\theta_i$  is the  $\theta$  value of water  $i$ ,  $\Delta G_{hyd}$  is the free energy of hydration of a water molecule (+6.4 kcal/mol) and  $\Delta G_{constr}$  is as follows:

$$\Delta G_{constr} = -kT \ln \left( \frac{V^{\text{constr}}}{V^0} \right) \quad (3.43)$$

Where  $V^0$  is the volume which the water would be able to access in the bulk solvent, which is taken to be 55.55 mol/L.

## 3.4 Modelling interactions of a system

Molecular Dynamics and Monte Carlo methods both require accurate modelling of the interactions of the system. This section will introduce those aspects of this modelling which applies to both these methods.

### 3.4.1 Force fields

Force fields are central to the production of a system's ensemble, as they provide a route to estimating the energy of a given configuration, allowing us to weight a given configuration, prevent the simulation from producing impossible configurations, and/or follow the time-dependent evolution of the system.

In a classical potential, the energy of the system is taken to be the addition of the sum of the (bond, angle and dihedral) stretching and electrostatic terms. The bond stretching term is often a simple harmonic function dependent on the deviation between a reference bond length and the distance between atoms, as well as the bond force constant. Similarly, the bond angle is also harmonic, with its own force constants specific to the atoms involved. Dihedral angle stretching terms are modelled using a Fourier series (and may also include scaled non-bonded interactions). The electrostatic terms are divided into the Lennard Jones potential to describe the van der Waals forces (repulsion and dispersion) and a Coulombic potential that describes long ranged

electrostatics. The general form of such a potential is shown below:

$$\begin{aligned}
 V(r^N) = & \sum_{bonds} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 \\
 & + \sum_{torsions} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \\
 & + \sum_{i=1}^N \sum_{j=i+1}^N \left( 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)
 \end{aligned} \tag{3.44}$$

Where  $k_i$  is the force constant for the bond/angle  $i$ ,  $l_i$  is the bond length, and  $l_{i,0}$  is the reference bond length. The angle magnitude is represented by  $\theta_i$ , with  $\theta_{i,0}$  corresponding to the reference angle. The periodicity of the force constant of a dihedral is determined by  $n$ , with  $V_n$  being the force constant for the dihedral,  $\omega$  the magnitude of the torsion, and  $\gamma$  the phase angle for the torsion.

The final term expresses the non-bonded interactions, and itself is comprised of two terms, the first is a Lennard-Jones potential with a “well depth” parameter  $\varepsilon_{ij}$  (that functions like the force constants in the previous terms to determine the strength of the potential), and a collision diameter  $\sigma_{ij}$  that corresponds to the separation at which the van der Waals interaction energy is 0. The parameter  $r_{ij}$  corresponds to the separation of atoms  $i$  and  $j$ . The second term of the non-bonded interactions is the Columbic term, describing the interaction of the atomic partial charges ( $q_i$  and  $q_j$ ) given their separation ( $r_{ij}$ ) according to the permittivity of free space ( $\varepsilon_0$ ).

A number of other potentials exist with a wide range of applications. Non-atomistic potentials attempt to increase computational efficiency by unifying adjacent atoms into beads which may interact in a repulsive/attractive manner as in the Go model, or via elastic constraints as in the elastic network model. Since a bead may negate the

requirement to calculate the interactions of several atoms, such methods are much faster; however, atomistic insight into the dynamics of such models is lost, and their dynamics is biased towards the configuration they were parameterized for[83].

Polarisable force fields allow the charge distributions to change during a simulation.

The fluctuating charge method utilises partial point charges, but these charges are variable, and change according to their surroundings. The inducible point dipole calculates induced dipoles that are iteratively minimised (interacting with other induced dipoles and fixed charges) until self-consistency is attained[84]. Chemical reactions can be modelled using modified forcefields[85] or using QM/MM methods[86]. The current work, however, utilises classical potentials, and is restricted to atomic representations that treat atoms as having fixed partial point charges.

### 3.4.2 Periodic boundary conditions

One problem with the forcefield as set out above is that it is constructed only to take into account particles within the system. In reality a protein system, such as we have studied here, would be surrounded by a comparatively vast volume of water and other molecules. The existence of a wider environment is not such a problem for the Van der Waals forces, which decay very rapidly over distance (and for which cutoffs can be implemented), but for the Coulombic interactions, this is problematic[87].

To emulate this environment, periodic boundary conditions are used, whereby the system is replicated into images. These images then tessellate through the space, extending outside of the original system. Theoretically, the long range interactions of particles in a box separated by  $r_{box}$  can be accounted for like so:

$$V = \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0 |r_{ij} + r_{box}|} \quad (3.45)$$

Where  $V$  is the energy contribution of the interaction between particles  $i$  and  $j$  where  $j$  is an image in a box separated by  $r_{box}$  from the central simulation box. The calculation could be reproduced for a finite number of simulation images; however, this does not sufficiently converge as  $r$  increases. A better approximation of the long range interactions utilises Ewald summation that can account for an infinite number of images[88].

Unfortunately the computational expense of calculating the long range interactions increases with the square of the number of particles, even using Ewald summation. To compensate, the particle mesh Ewald (PME) method is employed[88], which simplifies the problem by distributing the charges in the simulation onto a grid. In doing so, not only is the number of interactions that need to be calculated reduced, but the problem also becomes more amenable to fast Fourier transforms, providing even greater computational efficiency, scaling  $N \log N$ , as opposed to  $N^2$ .

### 3.4.3 Water models

Since the classical potential does not incorporate polarisation effects explicitly, the modelling of water is problematic. The most extensively used water models simply utilise point charges, and in the TIP3P model of water, this provides a reasonable model of the structural properties of water, with little computational cost. However, for reproducing the energetics of water, a more complicated model is required. TIP4P uses an extra site displaced from the position of the oxygen away from the hydrogens, in the direction of the normal of the angle, which leads to modest improvements in its thermodynamic properties. In the current work, in cases where increased accuracy is desired, as in rigorous free energy calculations, TIP4P is used[89].

### 3.4.4 Energy minimisation

Configurations in energy minima (especially global energy minima) correspond to important states of the system, additionally, identification of minima and the lowest-energy pathway between them can be used to identify transition states[90]. In the present study, however, the systematic location of energy minima is most useful in providing a realistic configuration from which to commence simulation by alleviating steric clashes from the crystal structure as well as relaxing the structure in its new simulation environment (having previously been determined from its crystal environment). In the present study two techniques have been utilised: the steepest descents method, and the conjugate gradients method.

The steepest descents method perturbs the  $N$  degrees of freedom in the system along a vector  $\mathbf{s}$  corresponding to the net force acting on those degrees of freedom, which also happens to correspond to the gradient of the energy landscape. It is then possible to find the minimum along this vector by searching along this vector. This process can be repeated until a minimum is found; however, where the energy landscape has a narrow valley the steepest descents method may become inefficient, as each iteration is orthogonal to the last, leading to an inefficient zigzagging across the energy landscape.

The conjugate gradients method does a line search along a different vector  $\mathbf{v}_i$ , which is related to the previous vector  $\mathbf{v}_{i-1}$  like so:

$$\mathbf{v}_i = -\mathbf{g}_i + \gamma_i \mathbf{v}_{i-1} \quad (3.46)$$

Where  $\mathbf{g}_i$  is the gradient at the start of iteration  $i$ , and  $\gamma_i$  is a constant:

$$\gamma_i = \frac{\mathbf{g}_i \cdot \mathbf{g}_i}{\mathbf{g}_{i-1} \cdot \mathbf{g}_{i-1}} \quad (3.47)$$

The conjugate gradients method is more efficient at low gradients, but at steeper gradients the steepest descent method is more efficient. Thus, steepest descents minimisation is usually performed before conjugate gradient minimisation[74].

### 3.5 Calculation of binding free energies

Binding free energies are highly valuable data that allow us to differentiate poor drug candidates from promising ones, and as such are important in the field of molecular simulation. Theoretically a binding free energy can be obtained by running a simulation where a drug molecule is allowed to bind spontaneously to its target, and while this has been accomplished[43], it requires exceptionally long simulation times for each drug candidate.

#### 3.5.1 Relative binding free energies

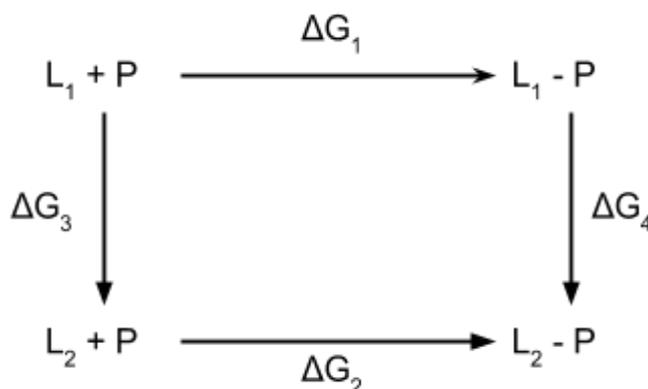


Figure 3.1: Thermodynamic cycle for the binding of ligands L1 and L2 to protein P

An alternative to evaluating ligands by comparing their absolute binding free energies ( $\Delta G_1$  and  $\Delta G_2$  in figure 3.1), is to calculate their relative binding free energy. Since  $\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$ , it is not necessary to calculate the time-consuming processes that provide  $\Delta G_1$  and  $\Delta G_2$ . Instead, the ligand can be mutated both in protein and in solvent to produce a rigorous and quick method of obtaining the relative binding free energy.

The free energy difference ( $\Delta G$ ) between two states (A and B) can be expressed as:

$$\Delta G = k_B T \ln \langle e^{-H(A,B)/k_B T} \rangle_A \quad (3.48)$$

Where  $H(A, B)$  is the difference in the free energy between state A and B.  $\langle \rangle_A$  denotes that the ensemble average is taken over state A. However, calculation of the free energy of state A and B separately may not lead to converged results, as the phase space between states A and B may lack significant overlap.

Alternatives exist, including the free energy perturbation method, however in the present study thermodynamic integration (TI) has been utilised. In TI the above equation is expressed as the integral of the rate of change in free energy with respect to  $\lambda$ [91]:

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\delta G}{\delta \lambda} \right\rangle_{\lambda} d\lambda \quad (3.49)$$

Where  $\lambda$  is a linear interpolation factor which is applied to the forcefield parameters, scaling them between those of state 1 and those of state 2. One convenience of this relationship is that it is possible to utilise numerous  $\lambda$  values, each consisting of a separate free energy calculation. The phase space overlap will be correspondingly improved as the number of  $\lambda$  windows is increased, leading to better convergence of the calculations.

In the present study, these  $\lambda$  windows have been allowed to undergo replica exchange, such that after a given number of iterations  $\lambda$  adjacent windows are allowed to exchange configurations. This allows for the individual  $\lambda$  windows to exchange atomic coordinates (subject to a Metropolis test), and thus increase their phase space overlap, and further increasing the accuracy.

By performing small perturbations from one ligand into consecutive ligands, the relative free energy of binding of a whole dataset of ligands can be compared. While this does not provide specific binding affinities for a given ligand, it does allow for the identification of promising drug-like molecules from the dataset.

## 3.6 Hybrid Quantum Mechanics/Molecular Mechanics

Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods arise from the dual requirement for speed and accuracy: MM, being computationally very efficient, allows for the very fast generation of an ensemble of structures, whereas QM, by incorporating the electronic structure of the system we are interested in, provides extra detail and allows for the explicit treatment of phenomena such as polarisation. MM has been described previously (see sections 3.1 to 3.4), and so this section will cover the basics of QM theory as applied in the present study, as well as how QM and MM can be meaningfully unified.

### 3.6.1 The Schrödinger equation

The following discussion on Quantum mechanics and equations are derived from Leach's *Molecular Modelling Principles and Applications*[74], and a more detailed

introduction can be found therein. The Schrödinger equation describes the motion of a particle through space over time:

$$\left\{ -\frac{\hbar^2}{2m} \left( \frac{\delta^2}{\delta x^2} + \frac{\delta^2}{\delta y^2} + \frac{\delta^2}{\delta z^2} \right) + \mathcal{V} \right\} \Psi(\mathbf{r}, t) = i\hbar \frac{\delta \Psi(\mathbf{r}, t)}{\delta t} \quad (3.50)$$

Where  $\hbar$  is Planck's constant divided by  $2\pi$ ,  $m$  is the mass of the particle in the external field  $\mathcal{V}$ , and  $\mathbf{r}$  its position. The square root of -1 is represented as  $i$ , and the wavefunction of the particle is  $\Psi$ .

By assuming the potential to be independent of time we arrive at:

$$\mathcal{H}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \quad (3.51)$$

Where  $\mathcal{H}$  is the Hamiltonian operator:

$$\mathcal{H} = -\frac{\hbar^2}{2m} \left( \frac{\delta^2}{\delta x^2} + \frac{\delta^2}{\delta y^2} + \frac{\delta^2}{\delta z^2} \right) + \mathcal{V} \quad (3.52)$$

It is possible, from the above to determine atomic orbitals for mono-electronic systems such as the hydrogen atom; however, the exact calculation of multi-electronic systems is not possible. To address this, the Born-Oppenheimer approximation was produced; which constrains the positions of the nuclei. This approximation is based on the assumption that the nuclei (being at least 1836 times heavier than the electrons) are not significantly perturbed by the motions of the electrons, and the electrons adapt almost instantaneously to changes in the nuclear positions. This provides the possibility of solving the Schrödinger equation for a molecule with only one electron, but not for multi-electronic systems, where the electrons may interact with each other.

The Hartree-Fock method attempts to solve this problem by treating each electron as existing in a fixed field (determined by the nuclei and other electrons), and then

examining the electrons individually. A solution to one electron will affect the solutions to the remaining electrons; however, continued iteration will ideally result in convergence.

Accurate representation of molecules require the construction of molecular orbitals. The Hartree-Fock equations are not suitable for application directly to molecules, and so molecular orbitals are constructed as the linear combination of atomic orbitals (LCAO). The atomic orbitals may be represented using Slater type orbitals (STO), or Gaussian orbitals, the latter being an approximation of the former (STO being more computationally demanding). The set of functions used to calculate these molecular orbitals are called *basis sets*.

A minimal basis set contains only those orbitals required for the number of filled orbitals in all the atoms; however, since the molecular orbital is constructed using the LCAO approximation the asymmetry introduced by the presence of other nuclei in the molecule is not accounted for. To address this, polarisable basis sets incorporate p-type orbitals with symmetrical orbitals such as s orbitals to approximate this asymmetry. However, the change in shape of an orbital due to containing more than 1 electron also needs to be accounted for. This can be achieved by expressing the orbital as a combination of a “contracted” orbital and a “diffuse” orbital, the diffuse orbital allowing a change in the overall shape of the orbital according to the number of electrons it contains. Another problem with minimal basis sets are their inability to appropriately describe deviations of orbitals from a spherical electron distribution. To address this extra functions can be added; however, this is often omitted for the inner orbitals of an atom, which contribute far less to the properties of the molecule. These are called *split valence* basis sets. In the present study the 6-31G\* basis set was used: this corresponds to utilising 6 Gaussian functions to represent the core orbitals, and 4 Gaussians (3 for the contracted and 1 for the diffuse functions) for the valence orbitals, The asterisk denotes that the basis set also includes polarisation functions.

### 3.6.2 Density functional theory

Density Functional Theory (DFT) relies on the existence of a relationship between the electronic energy and electron density, such that the energy  $E$  is:

$$E[\rho(\mathbf{r})] = \int V_{ext}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + F[\rho(\mathbf{r})] \quad (3.53)$$

Where  $V_{ext}$  is an external potential acting on the electrons,  $\rho(\mathbf{r})$  is the electron density, and  $F[\rho(\mathbf{r})]$  is the sum of the kinetic energy of the electrons and the inter-electron interaction energy[92]. Minimisation of this function gives the ground state electron density, under the condition:

$$N = \int \rho(\mathbf{r})d\mathbf{r} \quad (3.54)$$

Since there is a fixed number of electrons ( $N$ ). By applying a Lagrange multiplier ( $-\mu$ ) we get:

$$\frac{\delta}{\delta\rho(\mathbf{r})} \left[ E[\rho(\mathbf{r})] - \mu \int \rho(\mathbf{r})d\mathbf{r} \right] = 0 \quad (3.55)$$

However, it is still necessary to calculate  $F[\rho(\mathbf{r})]$ . To do this, the kinetic energy ( $E_{KE}[\rho(\mathbf{r})]$ ), electron-electron Coulombic energy ( $E_H[\rho(\mathbf{r})]$ ) and exchange and correlation energies ( $E_{XC}[\rho(\mathbf{r})]$ ) are summed[93]:

$$F[\rho(\mathbf{r})] = E_{KE}[\rho(\mathbf{r})] + E_H[\rho(\mathbf{r})] + E_{XC}[\rho(\mathbf{r})] \quad (3.56)$$

For brevity, the precise form of the terms in equation 3.56 will not be discussed here (see ref. [93]); however, it is worth noting that  $E_{XC}[\rho(\mathbf{r})]$  was originally defined by this equation. By making an additional guess at the electron density, and deriving a set of one-electron orbitals from this, an improved electron density can be obtained repeatedly until self consistent.

The local density approximation (LDA) and local spin density approximation (LSDA) assume that the electron density is constant around the point  $\mathbf{r}$ , the exchange-correlation energy is then calculated by integrating the exchange-correlation energy per electron over all space. However, to improve the accuracy of the DFT method, a number of techniques have been employed to effectively calculate  $E_{xc}$ , including gradient corrections, which take into account the gradient of the electron density at point  $\mathbf{r}$  as well as its magnitude. Additionally, the Hartree-Fock theory can calculate the exchange energy precisely, though the exchange-correlation term can be constructed in a variety of ways. The present work uses the Becke 3 parameter Lee-Yang-Parr (B3LYP) hybrid functional, that gives  $E_{xc}$  as the following approximation:

$$E_{XC} = E_{XC}^{LSDA} + a_0(E_X^{exact} - E_X^{LSDA}) + a_X\Delta E_X^{B88} + a_C\Delta E_C^{PW91} \quad (3.57)$$

Where  $E_X^{LSDA}$  and  $E_{XC}^{LSDA}$  are the exchange energy and exchange-correlation energy as calculated using the local spin density approximation, respectively.  $E_X^{exact}$  is the exchange energy,  $\Delta E_X^{B88}$  is Becke's 1988 gradient correction for the exchange energy, and  $\Delta E_C^{PW91}$  is Perdew and Wang's 1991 correction for the correlation energy. The coefficients  $a_0$ ,  $a_X$  and  $a_C$  are semiempirical and fitted to experimental data[94].

### 3.6.3 QM/MM rescoring of MM free energies

The QM techniques explained thus far have exceptionally broad applications; however, in this section only the application in rescoring of MM free energies will be considered. The advantage of using QM in the context of rescoring MM free energies is that the MM representation of the system can be produced very rapidly compared to QM, and polarisation effects that are not accounted for in MM simulations can be quantified using QM.

There are problems associated with combining the QM and MM systems: MM geometries may result in over/under polarization of the QM/MM system due to the simpler MM representation producing slightly incompatible configurations[95]. Additionally, if a cutoff is used for the QM region that cuts through an MM bond it will lead to half-filled orbitals, a problem that can be circumvented by dummy “link” atoms or using a hybrid  $sp^2$  orbital with one electron[74]. Additionally sampling of the QM region is still considerably expensive.

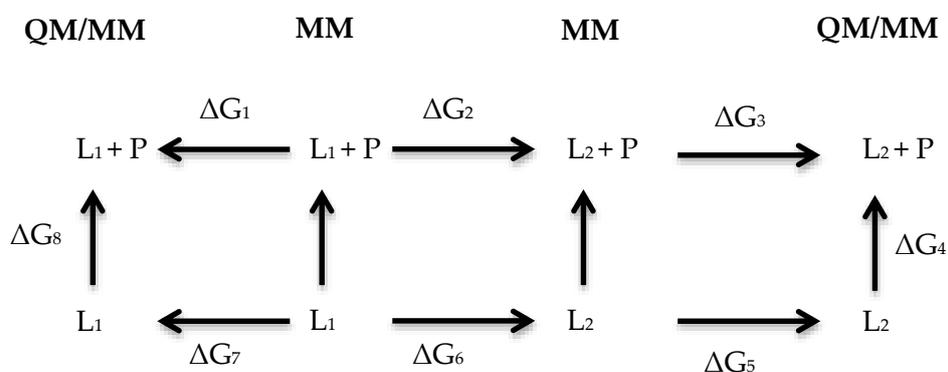


Figure 3.2: Thermodynamic cycle including the MM legs ( $\Delta G_2$  and  $\Delta G_6$ ) and the QM/MM legs ( $\Delta G_1$ ,  $\Delta G_3$ ,  $\Delta G_5$ , and  $\Delta G_7$ ) from which the relative QM/MM energy of binding ( $\Delta G_4 - \Delta G_8$ ) is calculated.

The calculation of MM relative binding affinities has already been discussed; similarly, QM/MM methods can be applied to this problem through the construction of thermodynamic cycles (see section 3.5.1). The present work draws on thermodynamic cycles designed only to incorporate the effect of polarization on the ligand. To achieve this, QM/MM simulations are constructed using the configurations generated during the perturbation from  $\lambda_0$  and  $\lambda_1$  (which results in  $\Delta G_2$  and  $\Delta G_6$  from figure 3.2).

Specifically, the MM configurations at  $\lambda_0$  and  $\lambda_1$  are used to construct a system with a QM ligand suspended in a sea of MM point charges.

The difference in free energy between the QM/MM system and MM system can be calculated using the Zwanzig equation (see equation 3.48) like so:

$$\Delta G_{MM \rightarrow QM/MM} = k_B T \ln \langle \exp(-(U_{QM/MM} - U_{QM}^{vac} - U_{MM}^{charges} - U_{solute-solvent,MM}^{Coul})/k_B T) \rangle \quad (3.58)$$

Where  $U_{QM/MM}$  is the energy calculated from a QM/MM calculation of the system with point charges,  $U_{QM}^{vac}$  is the single point energy of the solute in vacuum,  $U_{MM}^{charges}$  is the sum of all Coulombic interaction energies between the point charges, and  $U_{solute-solvent,MM}^{Coul}$  is the solute-solvent Columbic interaction energy.

The above calculation can be applied to obtain  $\Delta G_1$ ,  $\Delta G_3$ ,  $\Delta G_5$ , and  $\Delta G_7$  in figure 3.2, and thus obtain the corrected relative free energy of binding:

$$\Delta G_4 - \Delta G_8 = -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_5 - \Delta G_6 + \Delta G_7 \quad (3.59)$$

## 3.7 Analysis methods

### 3.7.1 Secondary structure prediction

The secondary structure of a protein has important implications for its functional role. In the present study the STRIDE algorithm has been used to analyse secondary structure. The STRIDE algorithm utilises both hydrogen bonding patterns and backbone torsions to assign a secondary structure to a given set of amino acids, this is

implemented in such a way that a poor score for torsion angles for a given type of structure can be offset if the hydrogen bonding score is appropriately high[96]. Other software exists, including the Dictionary of protein secondary structure (DSSP)[97], which is used extensively in secondary structure prediction, and DSSP and STRIDE were shown to agree well on the assignment of secondary structure[98].

### 3.7.2 Dimensionality reduction

Protein dynamics are often difficult to understand, since even small proteins will likely have upwards of 100 atoms, with each atom moving differently to its neighbours. Combining all this information and discarding noise from these movements is evidently impossible to do without computational effort.

Dimensionality reduction techniques are used increasingly to discern the main long-timescale motions involved in conformational change, to eliminate noise, and identify minima in terms of just a handful of collective variables.

### 3.7.3 Principal Component Analysis

At its simplest, Principal Component Analysis (PCA) reduces the dimensionality of a dataset by generating a matrix of the covariance between each datapoint (or, in the case of the present study, each trajectory snapshot). For a protein system the covariance matrix can be produced like so[99]:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \quad (3.60)$$

Where  $x$  comprises all the degrees of freedom,  $i$  and  $j$  are frames in a trajectory, and  $\langle x_i \rangle$  is the average positions of the coordinates. Diagonalisation of this matrix provides a set of eigenvectors which are comprised of the coefficients of the principal components[74].

The PCs are ranked by the amount of the variance in the data they account for, with PC1 being the eigenvector accounting for the most variance. As a consequence, the first PCs have the property of corresponding to large, slow motions in proteins, and subsequent PCs corresponding to less variance, and due more to noise than conformational change[99].

By projecting the data into PC space, and visualising its distribution on the first PCs it is possible to map the sampling of a trajectory based on just a few low frequency motions. In this way, it was hypothesised that the effects of the activating mutations of EGFR kinase would be discernable from their impact on these low frequency motions, as has been recognised in a previous study on the L858R mutation of EGFR[65].

#### **3.7.4 Multidimensional scaling**

In Multidimensional scaling (MDS) dimensionality is reduced such that the information is rearranged into  $N$  dimensions (where  $N$  is specified by the user). MDS attempts to arrange the data points across these  $N$  dimensions using a matrix of similarities/differences, and in such a way as to preserve the original data as much as possible[100].

In the present study, MDS has only been used to perform processing on data during

the diffusion map analysis[101].

### 3.7.5 Diffusion map analysis

Diffusion map analysis is similar to PCA, but rather than construction of a matrix of covariance, the analysis stems from a matrix  $K$  derived from pairwise RMSDs:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2\varepsilon_i\varepsilon_j}\right) \quad (3.61)$$

Where  $x$  are configurations of the system, and  $\|x_i - x_j\|$  represents the RMSD between configurations  $i$  and  $j$ . The  $\varepsilon$  parameters are the local scale parameters, which traditionally have been constants across all values of  $i$  and  $j$ . In a conventional diffusion map, the constant local scale would be chosen to be slightly larger than the proportion of the RMSD due to noise[101].

As in PCA, eigenvectors can be extracted from the matrix  $K$  by diagonalisation, and the eigenvector with the largest eigenvalue will be the one that correlates with whichever motion of the protein involves the greatest change in diffusion space. The theoretical advantage of the diffusion map over PCA is that diffusion maps approximate the Fokker-Planck equation (which describes how a probability distribution of a system evolves over time), and so the diffusion coordinate (DC) of a snapshot relates to how accessible a configuration with a similar/dissimilar DC is rather than simply how similar the snapshots are with respect to a given motion (as with a PC).

In other words, inclusion of the local scale parameters means that configurations are considered in their local environment; data points will cluster more tightly where there is more accessibility between similar configurations. This is in contrast to PCA, which

plots points proportionately to their interatomic distances.

A significant disadvantage with previous diffusion map implementations was the choice of the local scaling parameter, since the sampling of configurations of a molecular system involves motions occurring over multiple time scales. The current work utilises locally scaled diffusion maps, which employ multidimensional scaling to estimate  $\epsilon$  for each configuration, in this way the multiple time scales over which motions occur is automatically included in the analysis. In this way, the local scale is small enough that accessibility is not erroneously inferred between highly different structures, and yet large enough that accessibility is inferred to important, local conformations that do not differ only due to noisy motions.

### 3.8 Enhanced sampling techniques

The present study aims to investigate the impact of mutations on the conformational dynamics of EGFR kinase; however, it has been demonstrated both experimentally and computationally that the time-scales of kinase dynamics (in the millisecond to microsecond time scale[102]) are too long to elucidate the complete repertoire of EGFR sampling with conventional MD (cMD).

A number of techniques for enhancing the sampling of MD have been produced to date. Here, we will examine in detail the enhanced sampling methods employed in the present study.

### 3.8.1 Accelerated Molecular Dynamics

In the present study, Accelerated Molecular Dynamics (AMD) refers to the method developed by Hamelberg et al.(2004)[103]. AMD introduces two boost potentials, one to the dihedral energy, and another to the potential energy. The boost is applied while the energy is below a given level, and acts to reduce the conformational trapping in a similar manner to umbrella sampling[104].

The method introduces the following bias potential:

$$\Delta V(\mathbf{r}) = \frac{(E - V(\mathbf{r}))^2}{\alpha + (E - V(\mathbf{r}))} \quad (3.62)$$

Where  $\Delta V(\mathbf{r})$  is the boost potential,  $V(\mathbf{r})$  is the true potential,  $E$  is the boost threshold, and  $\alpha$  is a weighting factor. The bias potential is only applied when  $E > V(\mathbf{r})$ , where it is added to the true potential. Clearly, any increase in energy allows the system to sample over energy barriers that are particularly difficult for cMD to overcome; however, simply preventing  $V(\mathbf{r})$  from dropping below  $E$  is problematic for two main reasons. Firstly, if  $V(\mathbf{r})$  is not allowed to drop below  $E$  the system will undergo a random walk over the energy surface until it reaches a region of phase space where  $E < V(\mathbf{r})$ . Such a random walk will lead to a non-Boltzmann distribution of configurations, and while reweighting of results can assist in obtaining meaningful observables from the simulation, the isoenergetic surface must be well sampled to do so. Secondly, if the modified potential (ie.  $\Delta V(\mathbf{r}) + V(\mathbf{r})$ ) is constant when  $E > V(\mathbf{r})$ , then at points where the modified potential intersects the true potential ( $\Delta V(\mathbf{r}) + V(\mathbf{r}) = E$ ), the free energy landscape will be discontinuous, and integrating the equations of motion across them becomes problematic[103].

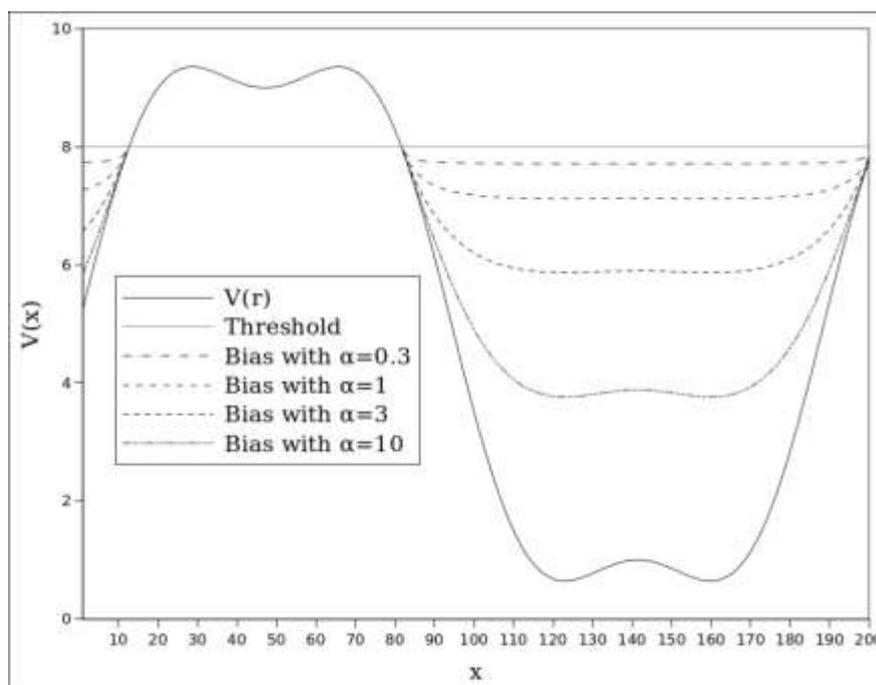


Figure 3.3: Effect of AMD bias potentials with a threshold of 8 and varying  $\alpha$  parameters on a hypothetical energy surface.

AMD avoids the above problems by reducing the depth of minima while maintaining their general shape. The intersection between the true potential and the bias potential is continuous, because the effect of the bias potential is increasingly less profound as the difference between the actual energy of a configuration and the threshold energy decreases (see equation 3.61 and figure 3.3). Meanwhile, by maintaining the shape of the minima, a larger proportion of relevant configurations are explored.

Nonetheless, high energy thresholds with low  $\alpha$  parameters still leads to isoenergetic regions of phase space (see figure 3.3), and so some care is required when assigning these parameters.

Experiments using AMD on test systems have shown that while applying such a boost to the torsional terms increases conformational change, in large protein systems the

relative immobility of the solvent becomes problematic. To account for this, Hamelberg et al (2007) developed the method to apply two bias potentials, one to the overall potential energy, and one to the torsional terms. This dual boosting technique was found to be effective not only in increasing the sampling of the system and convergence of results, but also produced distributions of configurations that would be expected from long time-scale cMD of the same system[104].

AMD has been shown to have a clear potential to increase the sampling of a system in a manner that mimics the sampling of long time-scale cMD[105]. The present study aims to utilise the AMD results directly in the elucidation of the effects of mutation on EGFR kinase sampling, as well as comparing it with other enhanced sampling methods.

### 3.8.2 Diffusion Map Directed Molecular Dynamics

Diffusion Map Directed Molecular Dynamics (DMDMD) is a method of increasing the range of sampling by favouring conformations that exist further away from energy minima than its neighbours, thus making the crossing of energy barriers more likely, and speeding up conformational change.

To identify those points furthest from the minima (frontier points), a diffusion map is constructed using snapshots from a short MD run. By the nature of diffusion maps (see section 3.7.5), the first diffusion coordinate (DC) will correspond to the slowest timescale motion occurring in the simulation. Since snapshots will tend to be spaced closer together towards the centre of a minimum, those snapshots with the greatest (negative or positive) DC1 value will be the snapshots furthest from a minimum. By restarting the system from such a snapshot, and repeating the process, the amount of phase space explored is increased.

If the length of the DMDMD iterations is too short, not only does DMDMD become inefficient because more time is spent calculating the diffusion map, but also the simulation may not sufficiently sample its current minimum to identify which snapshots are furthest from the centre of the minimum. This is manifested in the difference between each eigenvalue: The difference between the first and second eigenvalues should be much larger than the difference between the lower ranked eigenvalues. Simulations sampling too short a timescale will have very little difference between the first and second eigenvalues. Thus, the initial parameterisation of DMDMD can be performed by investigating the impact of the time covered by an iteration on the difference between the eigenvalues: ideally a diffusion map should produce a clear spectrum of eigenvalues (private correspondence with Wenwei Zheng, Rice University, 2012).

Since the snapshots contained within a diffusion map are taken from a period of time chosen such that no conformational change is likely to have occurred, instead of locally scaling the diffusion map, a constant local scale is applicable. This is useful because it reduces error from calculating the local scale (which is particularly high with a small number of snapshots), and decreases calculation time[106].

Although DMDMD increases the amount of conformational space explored, since it biases against sampling of minima, it does not produce a Boltzmann distribution of conformations, with less favourable conformations being more likely. Thus extra computational effort is required to generate a Boltzmann distribution from the frontier points collected during the DMDMD simulation before drawing quantitative conclusions from the data[107].

Being among the latest enhanced sampling techniques, DMDMD has only been applied to one protein system previously[107]. In the present study it has been utilised to probe

the conformational dynamics of EGFR kinase, to test whether it can explore similar conformations to long time-scale cMD and AMD, and to identify differences between DMDMD's unique method of sampling with that of the other methods employed.

### 3.8.3 Reversible Digitally Filtered Molecular Dynamics

Reversible Digitally Filtered Molecular Dynamics (RDFMD) is a method for accelerating low frequency motions (that is, those motions correlated with conformational change) of a protein. To achieve this, RDFMD must first filter out the “noisy” higher frequency motions using digital filters. The resulting isolated low frequency motion can then be amplified. In practice, this is realised using a digital filter that amplifies the low frequency motions while leaving the other frequency motions untouched. Additionally, the technique can be applied to specific regions of the protein, and has been used previously to promote the opening and closing of the catalytic loops of the Escherichia coli dihydrofolate reductase (DHFR)[108].

To ensure that the protein remains stable during the RDFMD simulation a delay is set between filter applications to allow energy to dissipate, the filter applications are additionally separated by 40 ps of NVT cMD, and a temperature cap is employed such that if possibly destabilising temperatures are reached, subsequent filter applications are abandoned until after the next NVT simulation. In particular, this temperature cap is employed to prevent *cis-trans* isomerisation of the protein backbone.

In RDFMD and digitally filtered molecular dynamics (DFMD), the filter itself is a collection of coefficients ( $c$ ) that has been designed specifically to boost low frequency signals in the input data. The final velocity ( $\mathbf{v}'_{n,t}$ ) given to an atom is:

$$\mathbf{v}'_{n,t} = \sum_{i=-m}^m c_i \mathbf{v}_{n,t-i} \quad (3.63)$$

Where  $2m+1$  is the total number of coefficients, and  $\mathbf{v}_{n,t-i}$  is the velocity of atom  $n$  at time step  $t$ .

Since the filter application process requires the same number of snapshots as coefficients in the filter, in DFMD, filters with a large number of coefficients lead to simulations with more snapshots. The longer simulation times used to generate these snapshots leads to the energy provided by the filter application to be lost during the next filter application.

RDFMD circumvents the problems of DFMD by running both a forward and reverse simulation, such that the user can specify a desired time delay (known as the filter delay) between the input structure and output structure or the digital filtering process. Thus, a filter delay can be specified that allows the energy of filter applications to build up over time, without dissipating, and without overheating[109].

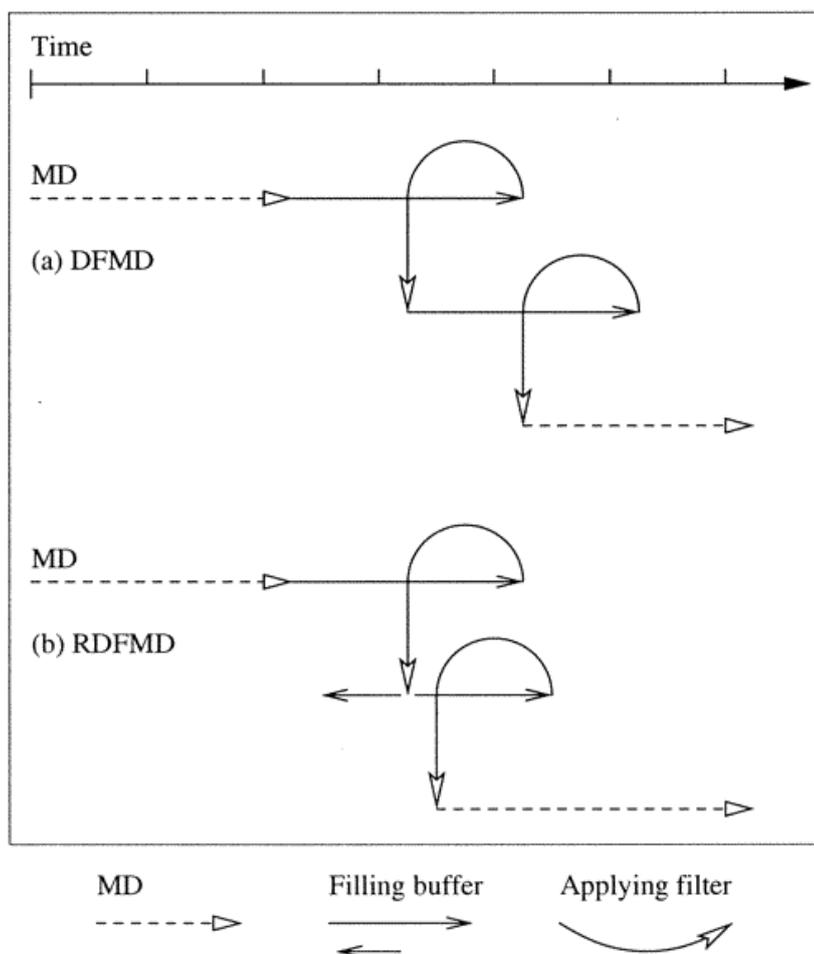


Figure 3.4: Comparison of DFMD and RDFMD methods, taken from ref. [109]

Evidently it is necessary to parameterise RDFMD for the specific system and amplification targets of interest. Inappropriate parameters may lead to inefficient sampling of the system, since several filter applications are required to effectively boost sampling, and exceeding the temperature cap results in a reduction in filter applications.

The ability of RDFMD to produce Boltzmann distributed configurations has not been thoroughly explored to date; however, it has been shown to be capable at inducing conformational change [108], [109], making it of particular interest in the present study, where the changes between the active and inactive conformations are predicted to occur over very long time-scales.

### 3.9 Summary

This chapter has described the basis behind the methods utilised in the present study. The entire work is essentially underpinned by statistical mechanics, and the assumptions that allow the extrapolation between experimental observations and computer models.

Monte Carlo provides a convenient method for the efficient sampling of systems where we are not interested in evolution over time, but still wish to accurately (and, where possible, quickly) determine the properties of the system. In the present study, MC was employed in the calculation of relative binding free energies. To supplement this, the water prediction methods GCMC and JAWS were also implemented, as the hydration of binding pockets has a significant impact on the binding free energy of inhibitors[110]. Polarisation effects in this investigation were also accounted for using QM/MM methodologies.

To attempt to elucidate the conformational dynamics of EGFR kinase, MD was utilised. Since MD simulates the time-dependent evolution of a system, it is a logical choice for this task; however, the long time scales over which conformational change occurs in kinases[4] renders simulation by MD less practical. To address this issue, several of the enhanced sampling methods mentioned in this chapter have been applied, both to gain more insight into the dynamics of EGFR kinase, and to compare those methods.

## Chapter 4: Prediction of relative binding free energies

### 4.1 Introduction to azoquinazoline inhibitors

The present work utilises experimental data summarised in Vema (2003)[67], drawn from two experimental studies[111], [112]. The ligands used in the present work are all azoquinazolines (see figure 4.1) of the “Family A” pyrido[3,2-d]pyrimidines described in Vema (2003), and were chosen for their pharmacological interest (the quinazolines include prominent EGFR inhibitors such as erlotinib and gefitinib), wide range of pIC50 values, and their well determined binding modes.

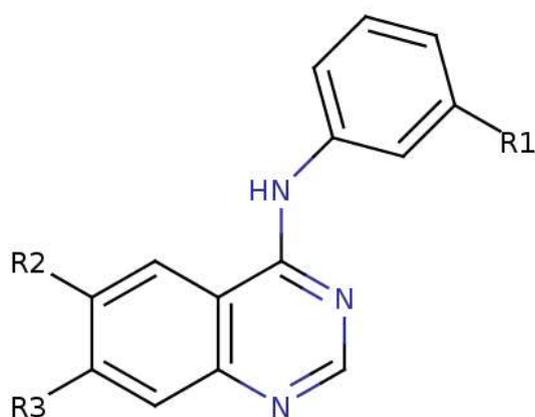


Figure 4.1: Representation of the azoquinazoline structure

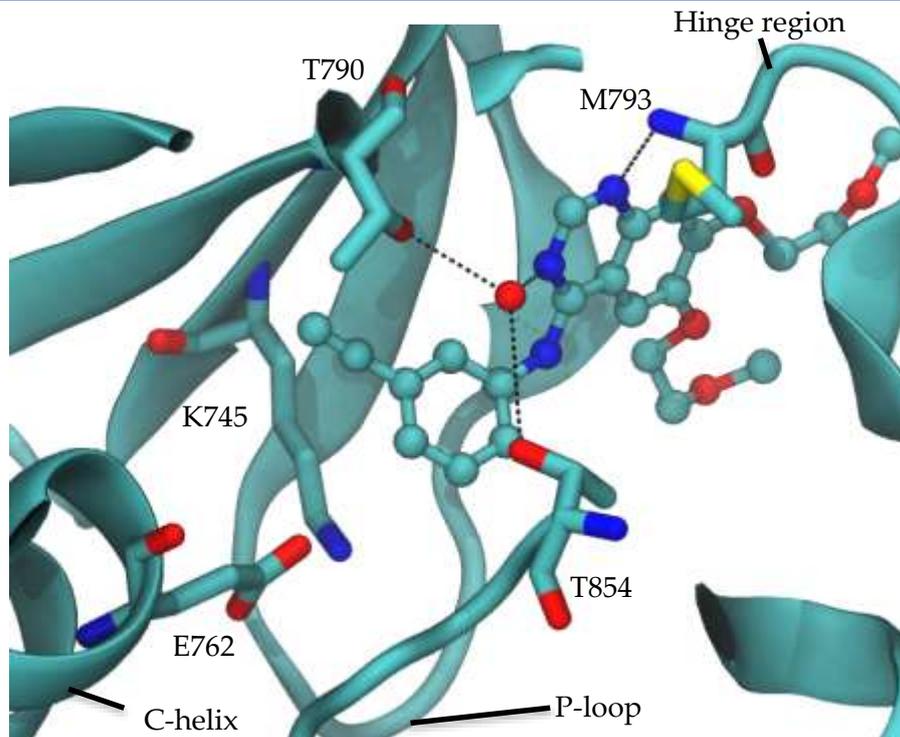


Figure 4.2: Interactions of Erlotinib (ball and stick representation) with EGFR residues, with important hydrogen bonds shown, as found in the crystal structure 1M17.

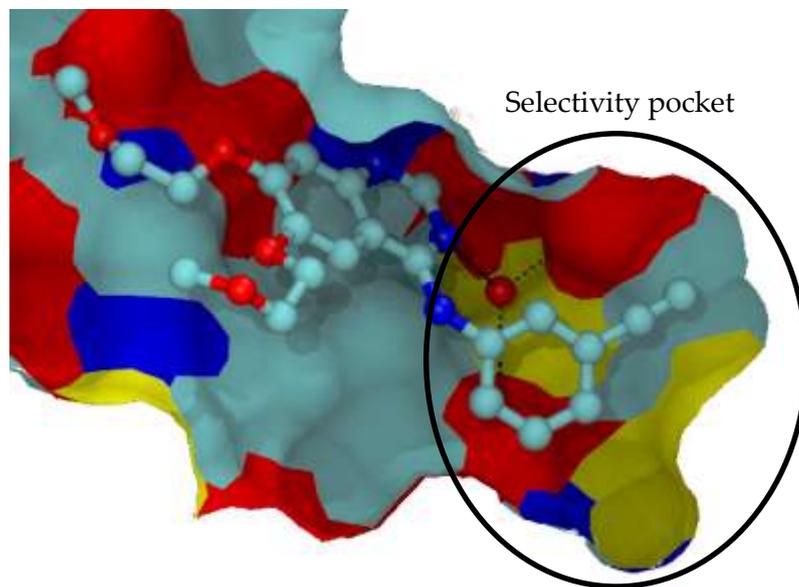


Figure 4.3: Erlotinib (ball and stick representation) in the EGFR kinase binding pocket (surface representation), with important hydrogen bonds shown, as found in the crystal structure 1M17.

Ligand	R1	R2	R3	pIC50
1	H	H	H	6.46
2	Me	H	H	6.04
3	Cl	H	H	7.63
4	Br	H	H	7.56
8	Br	OMe	H	6.45
10	Br	H	OMe	8
14	H	OMe	H	7.25
23	H	OMe	OMe	7.53
25	Cl	OMe	OMe	9.5
28	Br	NHMe	H	8.39
29	Br	N(Me)2	H	7.07

Table 4.1: Experimentally determined pIC50s of the compounds examined in the present study[67].

## 4.2 System setup and protocols

Starting coordinates for EGFR protein were taken from the crystallographic structure 1M17, since it already contains an azoquinazoline (erlotinib). This may bias the results to favour ligands that mimic the crystallographic ligand's interactions; however, by superimposing the ligands in table 4.1 onto the crystallographic positions, it provides a binding mode without using docking software or positioning manually, and thus hopefully removes an extra source of error. The crystallographic waters of this structure were kept throughout the following process.

The protein structure was passed into the WHAT\_IF program[113], using the HBONDS module to predict the positions of polar hydrogens throughout the protein. Asn, His and Gln residues were inspected by eye to check that the orientation of these residues were correct, and that flips of these residues performed by WHAT\_IF were reasonable (Gln849 and His988 were flipped by WHAT\_IF, and were retained). The partially protonated protein was then run through AMBER's tleap software[114] using the AMBER99FFSB forcefield[115], adding the remaining (non-polar) hydrogens. The protein was then subjected to 300 steps of steepest descent and 1200 steps of conjugate gradients minimisation.

The quinazoline structure from the ligand present in 1M17 was used as the base from which to construct the ligands presented in table 4.1. This was achieved using Accelrys Discovery Studio Visualizer[116]. The resulting pdb was passed to the antechamber program from the AmberTools software suite, which was used to assign charges (to a net charge of 0) using the AM1BCC method, and generate parameters from the General Amber Forcefield (GAFF)[117]. Protonation states were assigned according to advice from Richard Ward (Astra Zeneca, private correspondence, 2010). Missing parameters were checked using the parmchk program, also of the AmberTools suite. A Z-matrix for each molecule was then produced by hand, and flexibility of each relevant degree of freedom defined to ensure realistic flexibility of the ligand.

Ligand 1 was then placed into the protein structure, and a scoop was produced 15 Å around the ligand, such that residues that had no atoms closer than this threshold were discarded, and residues with no atoms closer than 10 Å around the ligand were made rigid. However, to maintain neutrality of the protein, peripheral residues E749, E829 and E906 were added and K757, K860 and K875 were removed.

Ligand 1 was then substituted within the protein for another ligand in table 4.1, and the protein scoop was placed in a sphere of TIP4P waters with a radius of 27 Å.

Crystallographic waters were converted into TIP4P waters, and everything was subjected to Monte Carlo equilibration using a residue-based cutoff of 10 Å, with a feather of 0.5 Å, and imposing a half-harmonic restraint force of 1.5 kcal mol<sup>-1</sup> on water molecules moving further than 27 Å from the simulation centre. The equilibration was run initially on only the water molecules, keeping the protein and ligand rigid for 40,000 steps to allow the water to relax. 160,000 steps were then performed on the entire system to equilibrate the protein, ligand and waters.

All equilibration and production runs were performed using protoMS[118]. Production simulations used single topology replica exchange thermodynamic integration (RETI), with 16 lambda windows (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, 1.00), and a replica exchange was attempted every 200000 moves. For every solute move, there were 9 protein moves, and 60 solvent moves. The solute and solvent were allowed to make rotations and translations, and the solute torsions and bond angles were allowed to flex, as were those protein sidechains not specified as rigid (as discussed above). A temperature of 300K was used. Bonds to dummy atoms were given a minimum length of 0.2 Å, and interpolation of the charge and other parameters between lambda windows was linear.

Simulations were run both bound (as described above) and free (in solvent), with the free simulations constructed similarly to the bound simulations, except with the protein excluded.

With 3 repeats for each perturbation listed in table 4.3, the protein, solvent and ligand of each lambda window was equilibrated for 20 M steps. Finally at least 80 M moves were performed at each value of lambda until the calculated free energy change converged, and the calculated average free energy change of the thermodynamic cycles (see figure 4.6) converged.

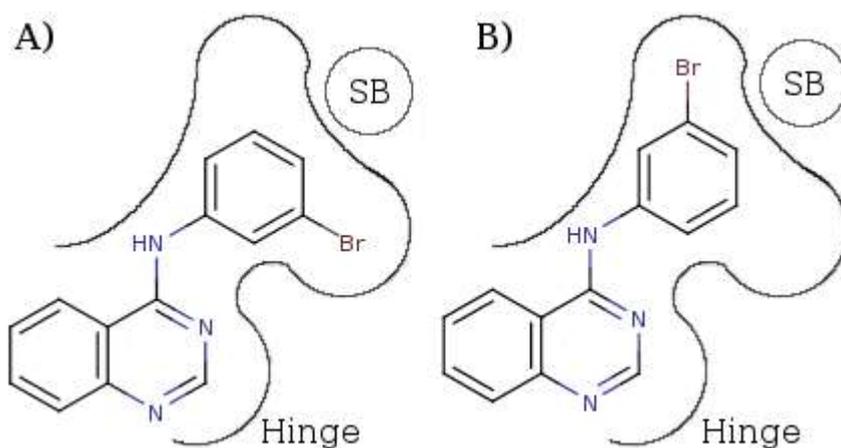


Figure 4.4: Schematic of ligand 4 in the binding pocket of EGFR. “SB” denotes the position of the K745-E762 salt bridge. A) The initial conformation corresponding to that found in the crystal structure 1M17 B) conformation of ligand 4 with the halophenyl group flipped approximately  $180^\circ$  (4f)

Owing to the ability of the phenyl group to rotate, it is possible that ligands 2, 3, 4, 8, 10, 25, 28, and 29 can bind the protein with the phenyl group in one of two orientations (for ligands with R1=H, the flipped orientation is indistinguishable). Though the crystal structures suggest the R1 group extends into the selectivity pocket, it is impossible to rule out the other alternative. Thus, to investigate the effect of this orientation change, an additional set of simulations were set up in an identical manner to those above, except that the halogen was removed from one side of the phenyl ring, and grown into the equivalent position on the other side (see figure 4.4), both in the bound and free legs.

### 4.3 Initial results

The perturbations and their resulting calculated relative binding free energies are presented alongside experimentally determined relative binding free energies in table

4.3, and plotted in figure 4.5. The method for estimating the relative binding free energies from the pIC50s provided in Vema 2003 proceeds from the Cheng-Prusoff equation[119]:

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}} \quad (4.1)$$

Where  $K_i$  is the inhibition constant, which should equal the binding constant of the inhibitor.  $IC_{50}$  is the concentration required to reduce enzymatic function by 50%,  $[S]$  is the concentration of the enzyme's substrate, and  $K_m$  is the value of  $[S]$  where the enzyme activity is half its maximum.

$K_i$  is related to free energy of binding by the following equation[120]:

$$\Delta G = -RT \ln K_i \quad (4.2)$$

Where  $R$  is the ideal gas constant,  $T$  is the temperature,  $K_i$  is the inhibition constant. We can use the above to relate relative binding free energies to  $IC_{50}$  like so:

$$\begin{aligned} \Delta\Delta G_{AB} = G_B - G_A &= (-RT \ln k_i^B) - (-RT \ln k_i^A) = -RT \ln \frac{k_i^B}{k_i^A} \quad (4.3) \\ &= -RT \ln \frac{IC_{50}^B}{IC_{50}^A} \end{aligned}$$

Because the data provided in Vema et al. was derived from the same assay method it was assumed that the same substrate and enzyme were used to measure the IC50 of each compound, and thus the ratio of the inhibition constants is equal to the ratio of IC50 values. The Cheng-Prusoff equation also assumes that the mechanism of inhibition is the same for all inhibitors, given that the predicted binding mode involves

the conserved hydrogen bond to the hinge region[34], [35], [41]. A more likely limitation is the fact the Cheng-Prusoff equation becomes decidedly less accurate at low agonist concentrations[121]. Given that the concentration of EGFR was  $0.05 \text{ ng mL}^{-1}$  and the concentration of EGF was  $2 \text{ } \mu\text{g mL}^{-1}$ , this limitation also seems unlikely to be a problem.

Errors were calculated as per the following equation:

$$SE_i = \frac{\sigma}{\sqrt{N}} \quad (4.4)$$

Where  $SE_i$  is the standard error for a simulation,  $i$ .  $\sigma$  is the standard deviation in results across the results from the simulations, and  $N$  is the number of simulations utilised in the calculation of the free energy (in the present study,  $N=3$ ). Because calculation of binding affinity requires the summation of the results of two simulations, the errors calculated for each simulation were then combined like so:

$$SE_{\Delta\Delta G} = \sqrt{\sum_{i=1}^{N=2} SE_i^2} \quad (4.5)$$

Additionally, predictive indices (PI) were calculated for the whole dataset using the following equation[122]:

$$PI = \frac{\sum_{j=i+1} \sum_{i=1} w_{ij} c_{ij}}{\sum_{j=i+1} \sum_{i=1} w_{ij}} \quad (4.6)$$

Where:

$$w_{ij} = |E(j) - E(i)| \quad (4.7)$$

and

$$\begin{aligned} \text{if } \frac{E(j) - E(i)}{P(j) - P(i)} < 0 \text{ then } C_{ij} &= -1 \\ \text{if } \frac{E(j) - E(i)}{P(j) - P(i)} > 0 \text{ then } C_{ij} &= 1 \\ \text{if } P(j) - P(i) = 0 \text{ then } C_{ij} &= 0 \end{aligned} \quad (4.8)$$

$E(x)$  is the experimental value for  $\Delta\Delta G_{\text{bind}}$  of the  $x^{\text{th}}$  perturbation, whereas  $P(x)$  is the  $x^{\text{th}}$  perturbation as calculated by the simulations detailed above. A PI of 0 corresponds to a random distribution of data, whereas values of 1 and -1 correspond to consistently correct predictions and consistently incorrect predictions, respectively. An additional point of note is that the weighting factor  $w_{ij}$  is proportional to the difference between the experimental binding free energies; thus the incorrect prediction of relative binding free energies for compounds with drastically different binding affinities will lead to greater penalty than for compounds with similar binding affinities.

Perturbation	$\Delta G_{\text{free}}$ (kcal mol <sup>-1</sup> )	$\Delta G_{\text{bound}}$ (kcal mol <sup>-1</sup> )	$\Delta\Delta G_{\text{bind}}$ (kcal mol <sup>-1</sup> )
4→4f	0.31 ± 0.32	0.85 ± 0.85	0.54 ± 0.91
8→8f	-2.82 ± 0.25	-0.75 ± 0.31	2.06 ± 0.40
28→28f	-0.97 ± 0.21	0.82 ± 0.37	1.79 ± 0.43
3→3f	4.32 ± 0.04	7.54 ± 0.06	3.22 ± 0.07
2→2f	7.64 ± 0.14	8.66 ± 0.09	1.02 ± 0.17

Table 4.2: Free energy change for flipping the halophenyl group bound to EGFR and in solution. The suffix “f” denotes the halophenyl group of the ligand is flipped 180 degrees (see figure 4.4).

Perturbation	$\Delta\Delta G_{\text{bind}}$ (kcal mol <sup>-1</sup> )	Exp <sup>a</sup> (kcal mol <sup>-1</sup> )	$\Delta G_{\text{free}}$ (kcal mol <sup>-1</sup> )	$\Delta G_{\text{bound}}$ (kcal mol <sup>-1</sup> )
4→1	-3.04 ± 0.07	1.50	0.60 ± 0.04	-2.44 ± 0.05
4→8	0.45 ± 0.17	1.51	4.59 ± 0.03	5.04 ± 0.17
1→2	1.74 ± 0.15	0.57	-4.51 ± 0.07	-2.77 ± 0.13
3→2	3.93 ± 0.09	2.17	5.77 ± 0.08	9.70 ± 0.02
3→8	4.57 ± 0.32	1.61	14.75 ± 0.10	19.31 ± 0.31
8→28	-1.60 ± 0.05	-2.64	-11.13 ± 0.03	-12.73 ± 0.04
28→29	-0.16 ± 0.32	1.80	8.23 ± 0.22	8.08 ± 0.23
8→25	-3.97 ± 0.21	-4.16	-8.59 ± 0.11	-12.56 ± 0.18
4→10	-0.16 ± 0.21	-0.60	-34.37 ± 0.06	-34.53 ± 0.20
1→14	-0.28 ± 0.21	-1.08	0.72 ± 0.09	0.44 ± 0.19
14→23	-0.74 ± 0.25	-0.38	-3.15 ± 0.18	-3.89 ± 0.17
4→2	-1.16 ± 0.32	2.07	-4.14 ± 0.05	-5.30 ± 0.32
1→3	-1.38 ± 0.23	-1.59	-10.41 ± 0.08	-11.79 ± 0.21

Table 4.3: Relative binding free energies compared with experimental values. a) errors were around ± 15% of original IC50 values[112], which corresponds to ± 0.21 kcal mol<sup>-1</sup>

To check consistency, thermodynamic cycles were constructed from the perturbations (see figure 4.6). The thermodynamic cycles would be expected to close to 0 kcal mol<sup>-1</sup> in a real system, with any deviation due to inconsistencies in the simulation.

Table 4.2 shows the relative binding free energy (calculated by subtracting the free energy change in the free leg from the same of the bound leg) for systems that undergo

a flip of the phenyl group. Flipping of the ring away from the orientation found in the crystal structure is universally unfavourable for the binding of the examined ligands, and thus only the results for the non-flipped simulations were used. The results also show a non-zero value for the free energy change in the free leg; which indicates that the ring does not rotate freely during the MC simulations.

The thermodynamic cycles for the calculations close appreciably well to zero (see figure 4.6), indicating consistency among the simulations, and a predictive index of 0.71 was obtained, suggesting that the calculations are more likely to rank perturbations in order of relative binding affinity than at random; however, the correlation coefficient was determined to be just 0.26, which is not statistically significant (to within 95% certainty).

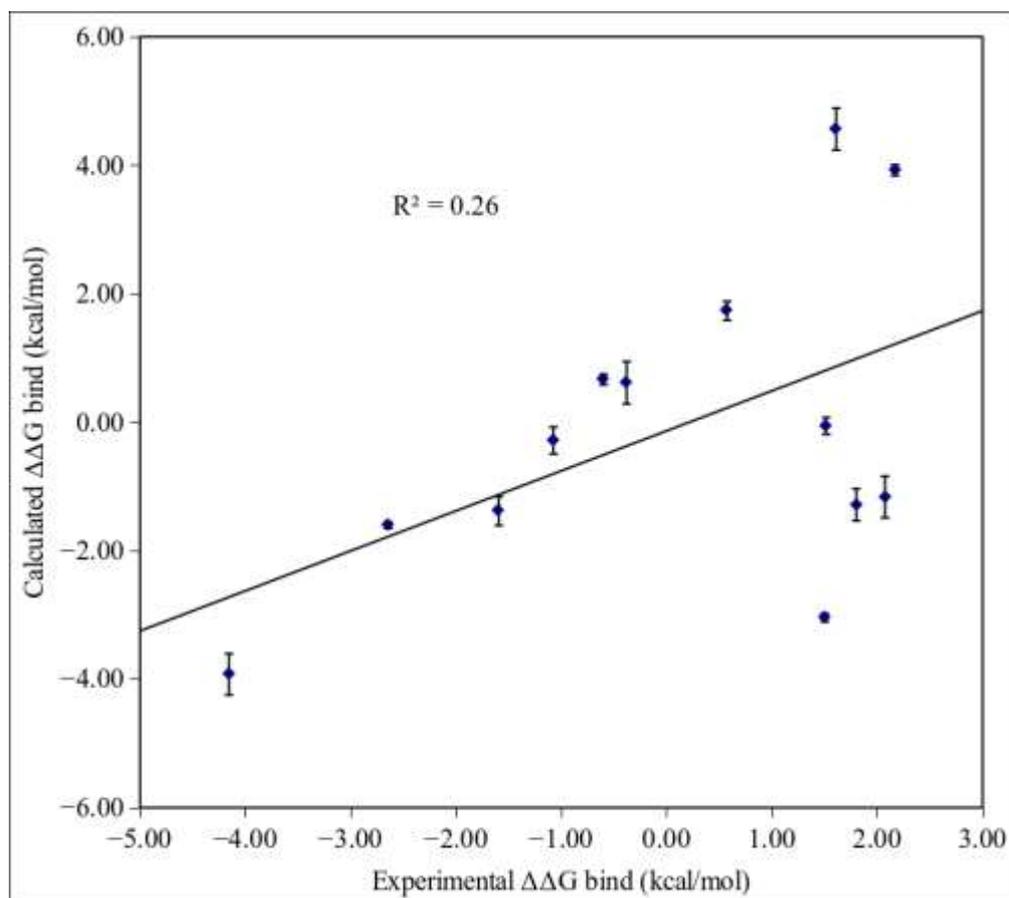


Figure 4.5: Experimental relative binding free energies versus the initial calculated binding free energies.

Thermodynamic cycles constructed from the free legs of the perturbations close to within 1 kcal mol<sup>-1</sup> of zero (see figure 4.6, and appendix 4), indicating that the simulations are relatively well behaved; however, it appears that simulations including the loss/gain of an OMe group concurrently with a halogen perturbation produce less reliable results, and this may account for the relatively large deviation of 0.79 kcal mol<sup>-1</sup> for one of the perturbation cycles. The bound legs are somewhat less well behaved. To some extent, this is expected due to the longer convergence times required for protein systems; however, considering the relatively small deviations from 0 for the other bound thermodynamic cycles, there may be some systematic problem. Additionally, from similar work, closure in the range of 0.4-0.9 kcal mol<sup>-1</sup> would seem more reasonable[123]. Causes for this problem were investigated. Lack of sampling was investigated by extending the calculation by 80 M moves. The free energy gradients appeared clear of inconsistency, no inconsistencies were found in the assigned flexes and point charges, and inconsistencies in sampling were investigated by visualisation; however, no causes for the bad closure of the thermodynamic cycles were found.

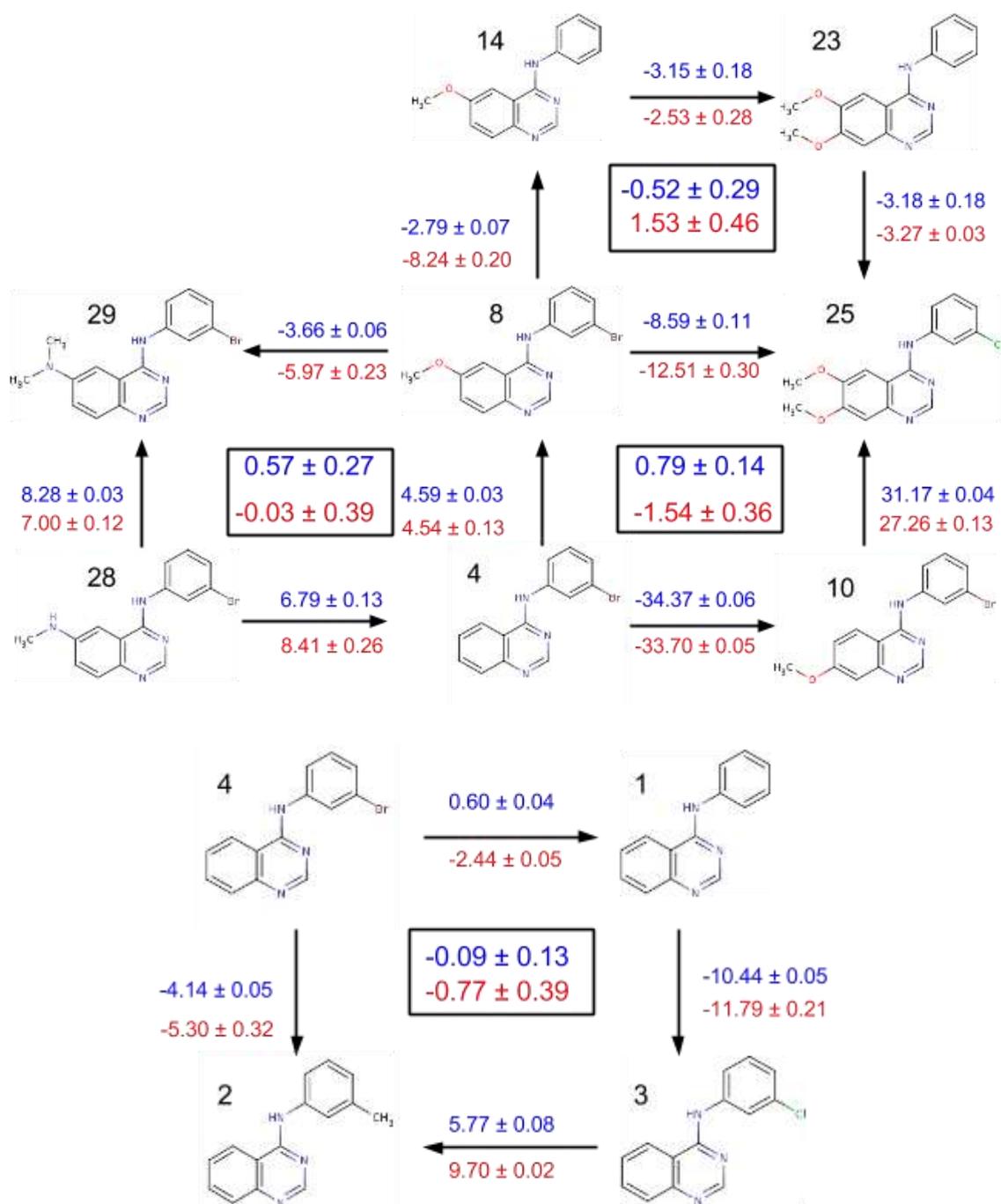


Figure 4.6: Closure of thermodynamic cycles: calculated relative free energy change for the free leg (blue) and calculated relative free energy change for the bound leg (red).

Totals in bold, in the centre of each cycle. All quantities are in kcal mol<sup>-1</sup>.

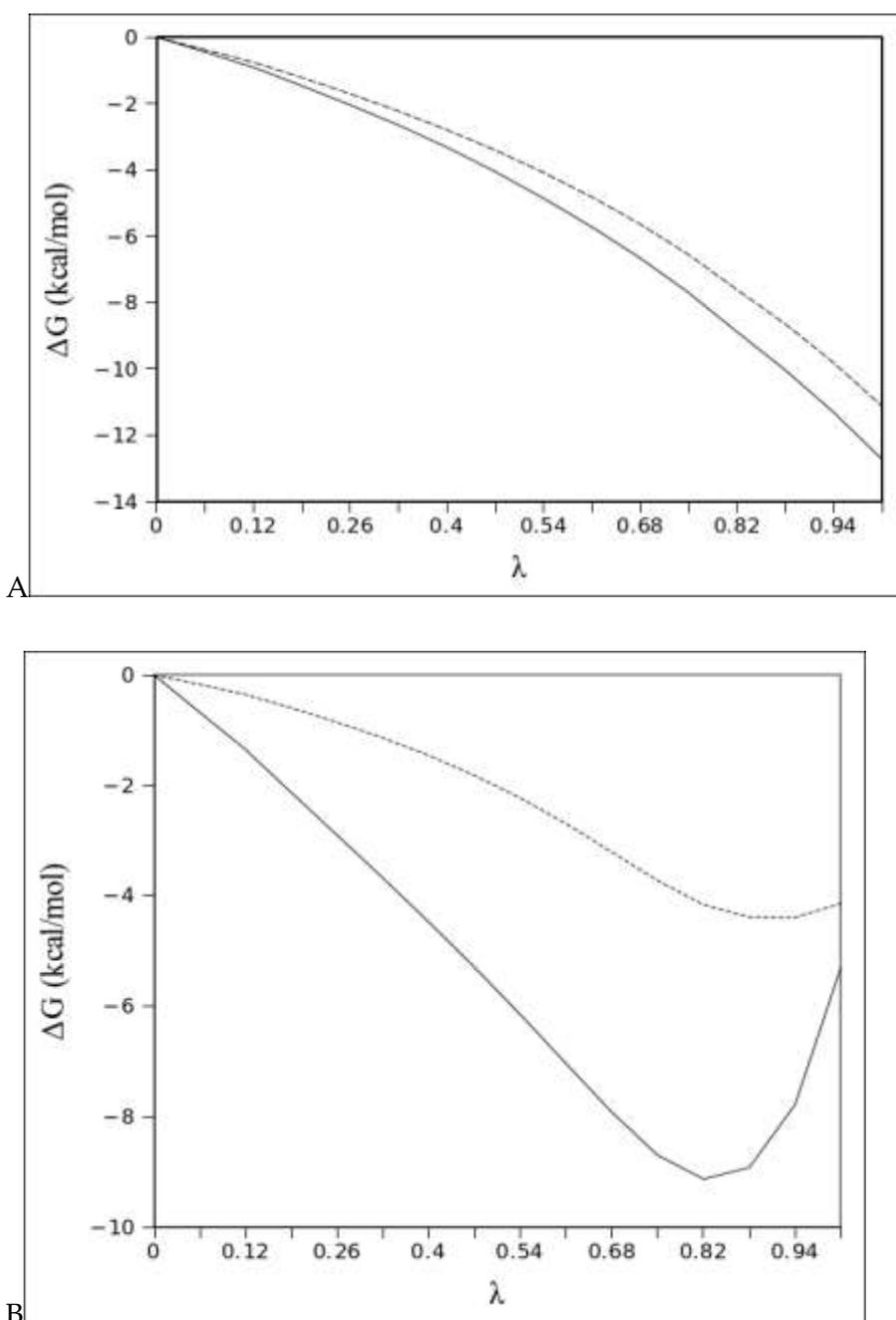


Figure 4.7: The change in free energy with the  $\lambda$  coupling parameter for (A) perturbation 8 to 28 (B) perturbation 4 to 2. Solid line represents the change when bound to EGFR kinase, the dashed line represents the change in water.

Figure 4.7 shows the free energy profiles of perturbations from 8 to 28 and 4 to 2. The profiles appear smooth, and thus do not indicate significant problems with the system. The profile for 8 to 28 suggests that the NHMe group interacts more favourably than

the OMe group both in the protein and in water, which may be expected since amines can both donate and accept protons. The amine group of ligand 28 projects into the solvent when bound to EGFR kinase, which appears to be responsible for the similarity of the profiles for the perturbation in protein and water. The profile for 4 to 2 shows a markedly more favourable free energy change in the protein, which is likely to be due to the reduced electronegativity of the Me group of ligand 2 versus the Br of ligand 4 (which will better accommodate solvation). However, from the experimental results, it was expected that ligand 2 would bind poorly compared to ligand 4, and the reason for the opposite being seen in the calculated results is not clear from the data presented thus far.

The primary conclusion of the initial results is that it has not been possible to predict the relative binding free energies of the inhibitors with certainty. The relatively well behaved free energy profiles and closed thermodynamic cycles suggest that the system has been adequately sampled, thus it appears that some systematic error is present. This could be a limitation of the force field, a problem with the system's configuration, or a sampling problem, such as the lack of rotation around the phenyl group in the free legs.

Particularly worrying were the incorrect sign of predictions for the perturbations 4→1 and 4→2. Visualisation of the trajectory of perturbation 4→1 revealed that the smaller ligand 1 was able to move into the selectivity pocket (see figure 4.8), at the expense of its hydrogen bond to the hinge region. The loss of this hydrogen bond is particularly problematic as the consensus binding mode requires a hydrogen bond to the hinge region[124]. Additionally, since ligands containing large R1 moieties cannot adopt the same conformation as ligand 1, this to some degree lessens the phase space overlap required to produce useful data.

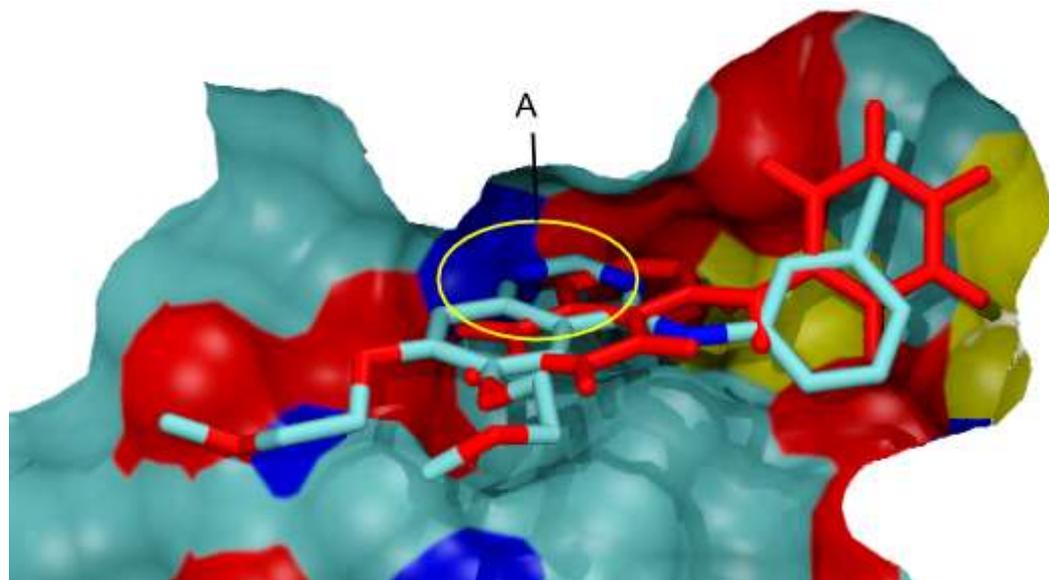


Figure 4.8: The binding pocket of EGFR with erlotinib and ligand 1 bound (PDB: 1M17). Final position of ligand 1 shown in red, the hydrogen bonding interaction with the hinge region is circled in yellow (A). Erlotinib is coloured by atom.

Analysis of all crystal structures revealed that a water molecule may exist in the selectivity pocket (PDB: 1M14, 2ITV, ITW). The absence or presence of this water with small inhibitors such as ligand 1 has not been determined in the literature, and would likely have a significant effect on the position of ligand 1 in the binding pocket. Thus, an evaluation of water sites within the binding pocket of EGFR was carried out.

#### 4.4 Water site prediction

In an effort to identify the cause of the poor correlation to experiment found in the previous section, the binding pocket waters were investigated. From studies by Balias and Rizzo (2009)[6], and Michel et al. (2009)[110], it seems that the binding pocket waters have a structure more complicated than suggested by the crystallographic structure 1M17. This can conceivably cause large errors in the above results due to the

lack of interactions of waters which should be present, and the large free energies that may be involved in adding or removing waters from the binding site[82]. Additionally, as mentioned previously (see section 4.3), it was found that without a halogen to fit into the selectivity pocket, ligand 1 was able to move its phenyl group into the selectivity pocket, resulting in the loss of ligand hydrogen bonding to the hinge region.

Since the crystal structures do not provide a consensus for the number of present water molecules, it has been necessary to utilise techniques that can systematically elucidate probable water sites. For this purpose, the Grand Canonical Monte Carlo (GCMC) and Just Add Water molecules (JAWS) techniques have been applied.

### 4.4.1 System setup

To maintain consistency, the same protein, water cap and ligands were used as those produced in section 4.2, with the exception that the crystallographic waters were removed to confirm that the methods were sensitive enough to identify those known waters. The prediction of water sites was carried out with ligand 1,2,3 and 4 to ascertain whether changes in the R1 group lead to a change in hydration pattern, ligands 8, 10 and 28 were included in the analysis to ascertain the same for changes to the R1 and R2 group.

### 4.4.2 GCMC protocol

For ligands 1-4, 8, 10 and 28, a box of dimensions 16 x 14 x 16 Å was defined to incorporate the binding site. These dimensions did not extend into the solvent exposed region, as waters are freely able to sample that region by diffusion during the MC

production runs. Apart from the normal moves available to the Monte Carlo simulation, the system was allowed to attempt to add or remove waters within the box such that for every solute move 3 attempts were made to insert and delete TIP4P waters into the box. Each system was run for 40 million moves with a B factor of -2, -4, -6, -8, -10, -12, and -14. The B factor is defined as follows:

$$B = \frac{\mu_{ex}}{k_B T} + \ln \bar{\rho} V \quad (4.9)$$

Where  $T$  is temperature,  $k_B$  is the Boltzmann constant,  $\mu_{ex}$  is the excess chemical potential.  $\bar{\rho}$  is the number density of water in the bulk (corresponding to a value of 0.0334), and  $V$  the simulation volume.

These simulations were used to search for water sites. Upon their identification, an additional set of simulations was run using a  $3 \times 3 \times 3 \text{ \AA}$  box around each water site, with other parameters the same as above. This allows for a more rigorous investigation of the binding free energy of waters at each site.

To investigate water sites in more detail, GCMC can be used to determine the binding free energy of a water molecule in the particular site using the following equation:

$$\Delta G_{bind} = -\Delta G_{hyd} + k_B T (B - \ln \bar{\rho} V) \quad (4.10)$$

Where  $G_{hyd}$  is the hydration free energy of water (taken to be  $-6.4 \text{ kcal mol}^{-1}$ , from previous experiments within the laboratory). In this instance  $V$  is taken to be a volume of  $16 \times 14 \times 16 \text{ \AA}$  for the water site search, and  $3 \times 3 \times 3 \text{ \AA}$  for more rigorous estimates at each high occupancy site.

Because the GCMC simulations produce very noisy results, 3 separate runs were performed for each ligand to produce better statistics. Initial analysis is performed using the results from the  $16 \times 14 \times 16 \text{ \AA}$  box: MC snapshots were examined using the Visual Molecular Dynamics software[125] to identify where waters were inserted during the simulation. The occupancy of these water sites over the course of the simulation was measured, and sites with an occupancy within standard error of 0.5 at a B factor corresponding to a negative free energy of binding were used for further investigation. This further investigation involved encasing the water sites in  $3 \times 3 \times 3 \text{ \AA}$  boxes and repeating the simulation at each water site to estimate the binding free energy of the waters at these sites. This estimation was achieved by extrapolating the B factor at which the occupancy at each site reaches 0.5, this B factor is then used to calculate the binding free energy (see equation 4.10).

### 4.4.3 GCMC results

Initial GCMC results are presented in figure 4.9 and appendix 5. As seen in figure 4.9, water W1 clearly has a favourable binding free energy, being occupied at all ranges of B value tested. W2 and W3 are more ambiguous, with consistently lower occupancy. From these results it seems that W2 and particularly W3 may have a lower binding affinity, however even W3, which appears to have a less favourable binding affinity than W2, has a tendency to remain within error of the critical 0.5 occupancy for a range of B-factors. Since the GCMC results over the whole pocket were noisy, it was decided to investigate sites W2 and W3 further, in addition to W1.

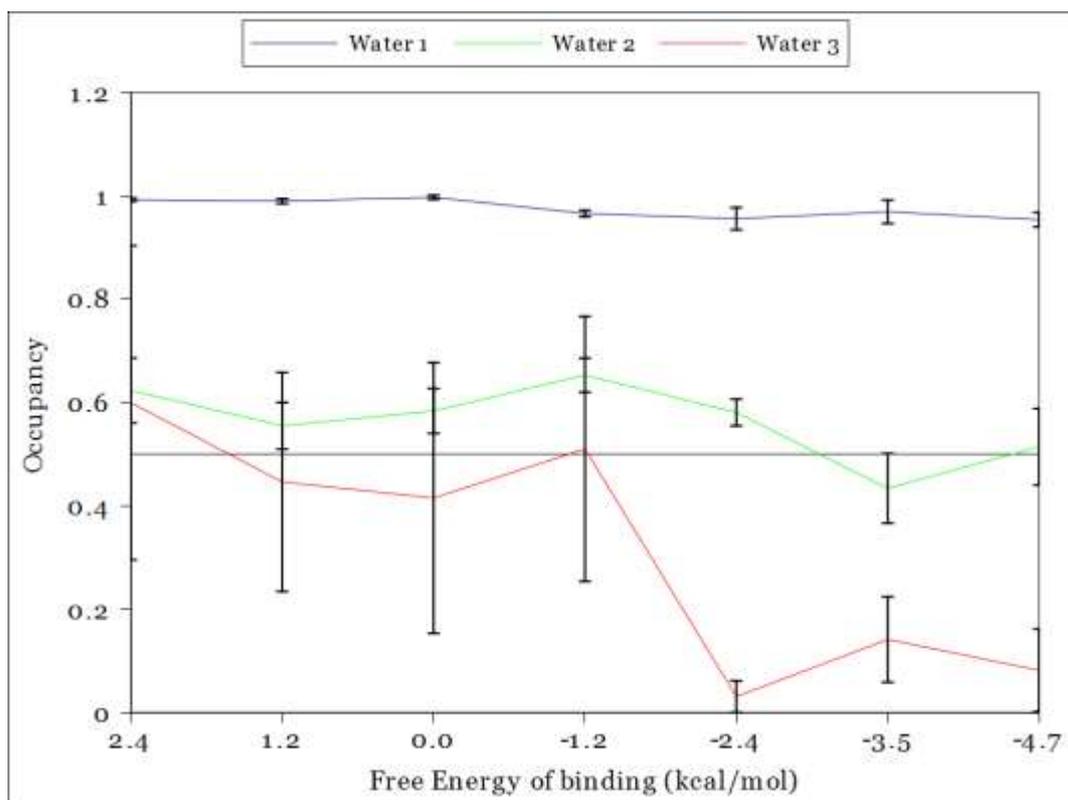


Figure 4.9: GCMC derived occupancy of waters in the binding site with ligand 10 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

W2 corresponds to the crystallographic water seen in 1M17 (see figures 4.10 and 4.11), W1 and W3 are seen in some crystal structures, including 2ITP, 2ITQ and 2ITV. The presence of W1 and W2 are also described in the literature, both in Balias & Rizzo (2009)[6] and the study of kinase binding site waters by Michel et al (2009)[110]. W3, however is not mentioned in either study. It is possible that the latter study defined a smaller binding pocket, indeed site W3 was only found on extension of the defined binding pocket to include the region of the K745-E762 salt bridge. Additionally values for the occupancy of W3 are very noisy (see figure 4.9 and appendix 5), but generally close enough to 0.5 to warrant further investigation, and thus was subjected to the second stage of GCMC (see figure 4.15).

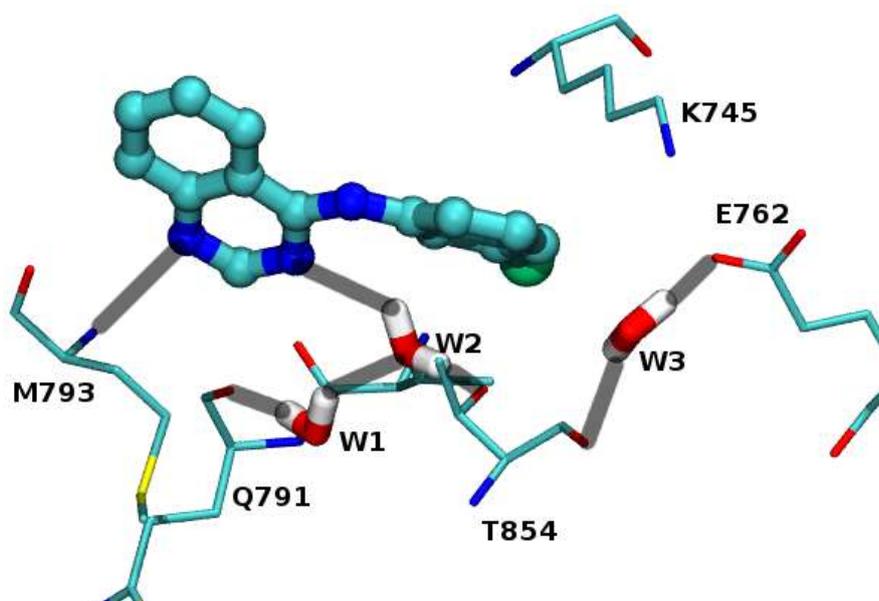


Figure 4.10: Ligand 4 bound to EGFR, GCMC waters with a binding free energy within error of 0 kcal mol<sup>-1</sup> are labelled W1-3, with possible hydrogen bonds highlighted in grey.

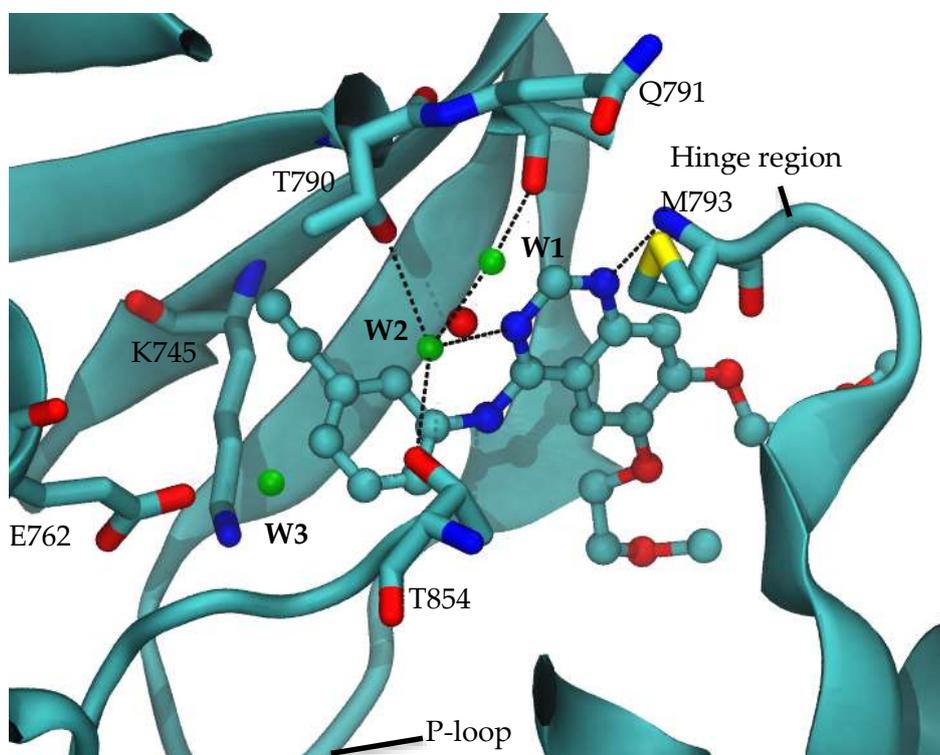


Figure 4.11: Erlotinib (ball and sticks) in the binding pocket. The positions of predicted waters (green) are shown alongside the crystal structure water of 1M17 (red)

Separate GCMC simulations on  $3 \times 3 \times 3 \text{ \AA}$  boxes around each water described above produced better statistics (see figures 4.12, 4.13, 4.14, and 4.15), with a clear titration curve in occupancy as the B factors (and thus free energies) are decreased. The reduction in noise is likely due to the insertions being concentrated in the vacant areas of the pocket defined by the smaller box, rather than the entire pocket.

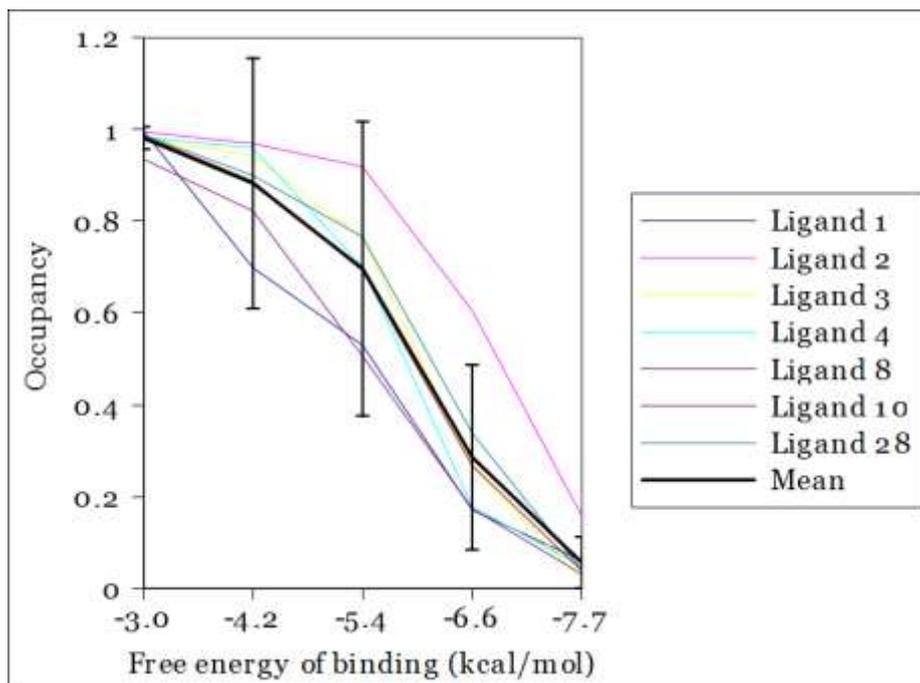


Figure 4.12: Occupancy of water 1 in GCMC simulations for a range of ligands and B factors corresponding to the free energy of binding (x-axis).

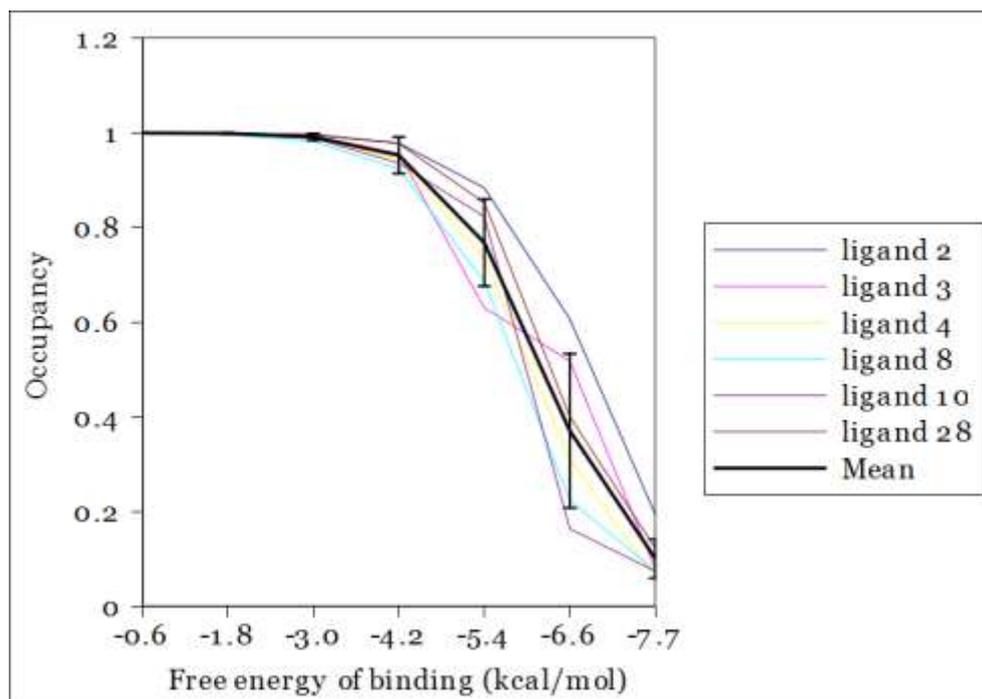


Figure 4.13: Occupancy of water 2 in GCMC simulations for a range of ligands and B factors corresponding to the free energy of binding (x-axis).

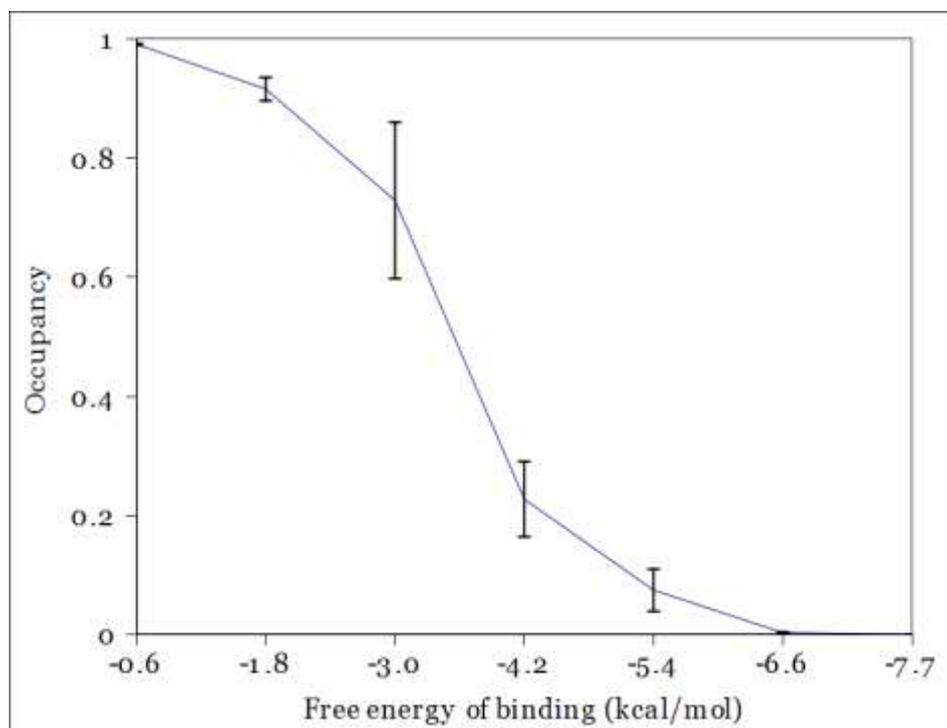


Figure 4.14: Occupancy of water 2 in GCMC simulations for ligand 1 and a range of B factors corresponding to the free energy of binding (x-axis).

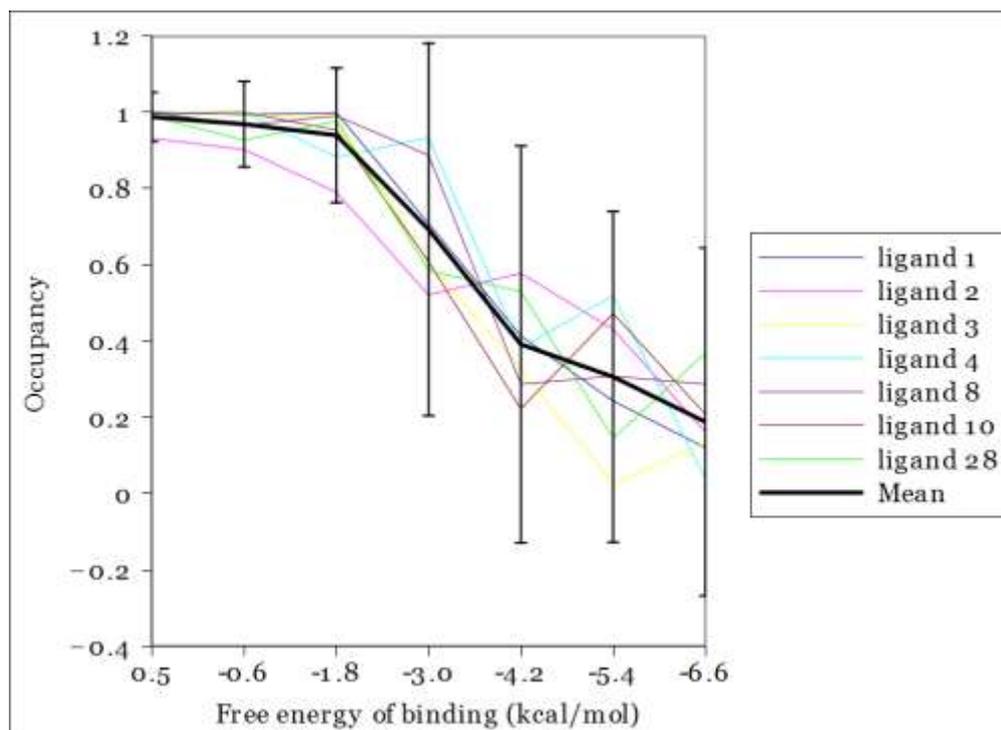


Figure 4.15: Occupancy of water 3 in GCMC simulations for a range of ligands and B factors corresponding to the free energy of binding (x-axis).

GCMC predicts waters at site 1 to have a binding free energy of between -5 and -7 kcal mol<sup>-1</sup>, and between -6 and -7 kcal mol<sup>-1</sup> for site 2, with waters at site 3 having a binding free energy of between -3 and -5 kcal mol<sup>-1</sup>. These figures are different to those obtained during the water site search, since the larger box area would have been partially occluded by ligand and protein, resulting in a reduction in insertion moves.

Despite some differences between the calculated binding free energies of the waters between ligands the results indicate that all of the identified waters have favourable binding free energies. It is interesting to note that water 2 has a markedly less favourable binding free energy (approximately -3.6 kcal mol<sup>-1</sup>) when ligand 1 is present, as indicated by a shift towards higher B factors (see figure 4.14). Since ligand 1 was found to sample configurations closer to the back of the selectivity pocket in the initial simulations, it may be that the increased mobility of ligand 1 is associated with

the reduced binding free energy of this water. This indicates that some significant proportion of the difference in binding free energies of ligand 1 compared to the other ligands is due to the ligand's interactions with binding pocket water 2.

#### 4.4.4 JAWS protocol

For ligands 1-4, 8, 10 and 28, a  $16 \times 14 \times 16 \text{ \AA}$  3D grid was defined in the binding site of EGFR, with a spacing of  $1 \text{ \AA}$ , as with the GCMC setup; these dimensions did not extend into the solvent exposed region. A biasing potential of  $6.4 \text{ kcal mol}^{-1}$  was used to ensure sufficient sampling of the on and off states. 40  $\theta$ -waters were injected into the binding site and allowed to sample without interacting for 1 million MC moves. Then, to discover water sites, 20 million MC moves were performed. During these moves, as well as sampling the solvent, protein and ligand, the waters were allowed to sample an interaction scaling factor,  $\theta$ , between 0 and 1 (inclusive), with 13  $\theta$  moves performed for each 6 physical (ie translation or rotation)  $\theta$ -water move. Each grid point is associated with an array element that is incremented when a nearby water samples a  $\theta$  value greater than 0.95. Cumulative data regarding the presence of interacting  $\theta$ -waters within the binding site was viewed as a density plot using the Open Astex Viewer[126]. Putative water sites were then examined by introducing a  $\theta$ -water into them and placing the grid centred on the water site. In each case a hardwall potential was introduced to prevent waters leaving the grid. The system was once again allowed to sample as previously for 40 million MC moves, and the binding affinity estimated (see section 3.3.5).

## 4.4.5 JAWS results

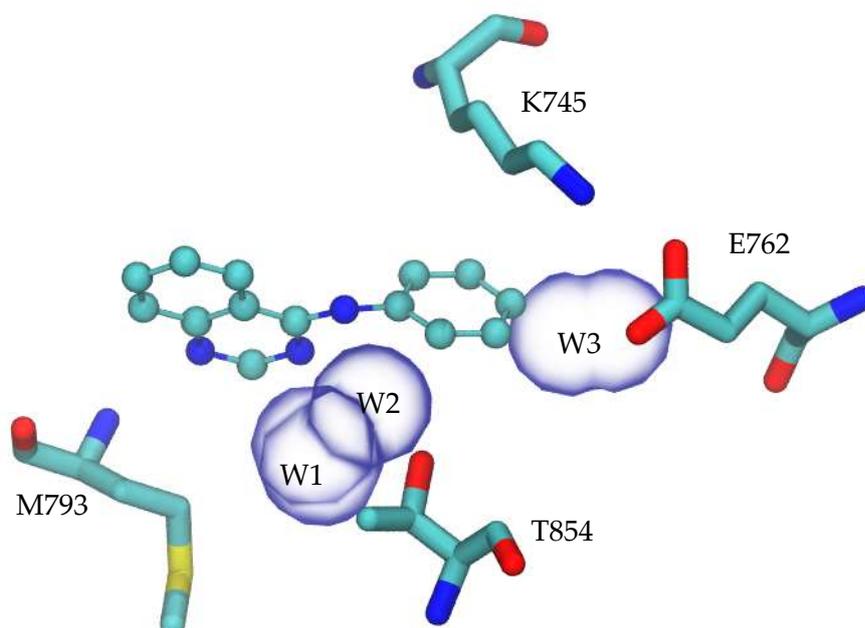


Figure 4.16: Representative JAWS results. Ligand 1 shown with important surrounding residues. Water sites predicted by JAWS shown as translucent blue spheres, labelled with the prefix W.

The initial JAWS simulations predict several water sites within the binding pocket (see figure 4.16). All the sites shown in figure 4.16 have an approximate probability of containing a water of greater than 0.5. Sites W1, W2 and W3 correspond to the GCMC determined sites (labelled identically in figure 4.10).

Water	Binding Free Energy (kcal mol <sup>-1</sup> )
W1	-6.00 ± 0.28
W2	-5.50 ± 0.26
W3	-4.76 ± 0.21

Table 4.4: Binding free energies of waters in sites elucidated by GCMC and JAWS stage 1, calculated using the JAWS stage 2 method. Results shown as an average of the values across 7 ligands.

The binding free energies were well conserved between ligands, as shown by the low error produced when averaging values over all the ligands simulated in JAWS (see table 4.4). Table 4.4 shows the predicted free energy of binding for waters W1, W2 and W3. Each of the waters has a favourable binding free energy. However, there is a possibility that the binding free energies have been overestimated, as the ratio of sampling between “on” and “off” waters was around 8000:1. Nonetheless, with the applied bias of  $6.4 \text{ kcal mol}^{-1}$ , these ratios indicate a highly favourable binding free energy, which is also in agreement with the GCMC results.

## 4.4.6 Context of identified water sites

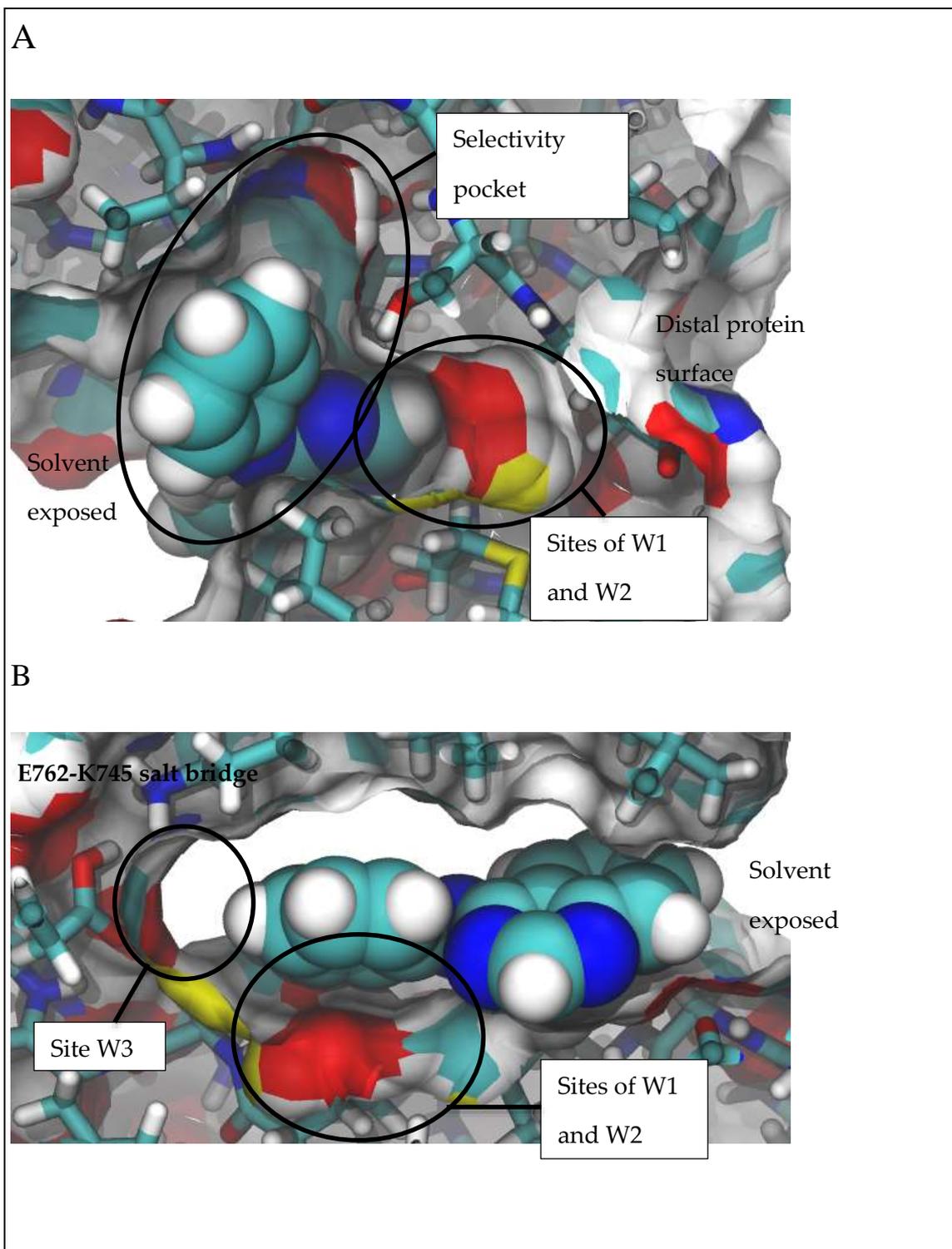


Figure 4.17: Cross sections of the binding pocket of EGFR kinase with Ligand 1 (the smallest of the ligands examined; represented by VDW spheres) bound.

Figure A shows cross section of the ATP binding pocket of EGFR with ligand 1 bound. Cross section A (figure 4.17(A)) clearly shows the selectivity pocket, as well as the space where W1 and W2 are found. This space is comprised of a channel that extends to the distal surface of the kinase; however no waters are found to diffuse into the binding pocket from the bulk in any of the simulations, and there are no reports in the literature where waters were found to utilise this channel. It seems likely that diffusion through the channel requires far more sampling than is obtained in the MC simulations presented here.

Bulk water is additionally obstructed from diffusing into the pocket by the positioning of the ligand; even the smallest ligand prevents water from entering sites W1 and W2 (see figure 4.17(A)). Site W3 appears more open to the solvent (see figure 4.17(B)), and there is some flexibility in the side chains of the salt-bridge, which suggests that sufficient sampling might allow bulk waters to enter site W3 if the salt-bridge moves enough; however, in the simulations of the present study no waters were seen to move to or from site W3.

## 4.5 Results with predicted waters

To investigate the effect of including the waters as predicted in the previous sections, the systems were set up as before (see section 4.2), except the crystallographic water in the binding site was replaced with W2 from the GCMC simulations, and W1 and W3 were included too.

Figure 4.18 shows that the simulations incorporating the waters predicted with GCMC and JAWS produces a moderately improved correlation with experiment compared to

the results obtained with 1M17's crystallographic water, although the correlation is still not statistically significant ( $p > 0.05$ ). Additionally, a PI of 0.78 was calculated for the dataset, when compared with the PI of 0.71 in the original dataset, this suggests that the addition of the predicted waters had a mild effect on the calculated binding free energies.

Perturbation	$\Delta\Delta G_{\text{bind}}$ (kcal mol <sup>-1</sup> )	Exp <sup>a</sup> (kcal mol <sup>-1</sup> )	$\Delta G_{\text{free}}$ (kcal mol <sup>-1</sup> )	$\Delta G_{\text{bound}}$ (kcal mol <sup>-1</sup> )
4→1	-2.23 ± 0.05	1.50	0.60 ± 0.04	-1.63 ± 0.02
4→8	0.36 ± 0.07	1.51	4.59 ± 0.03	4.95 ± 0.06
1→2	1.81 ± 0.14	0.57	-4.51 ± 0.07	-2.71 ± 0.12
3→2	3.74 ± 0.10	2.17	5.77 ± 0.08	9.50 ± 0.06
3→8	4.24 ± 0.53	1.61	14.75 ± 0.10	18.99 ± 0.52
8→28	-1.76 ± 0.14	-2.64	-11.13 ± 0.03	-12.89 ± 0.14
28→29	0.03 ± 0.57	1.80	8.28 ± 0.22	8.31 ± 0.53
8→25	-3.24 ± 0.24	-4.16	-8.59 ± 0.11	-11.83 ± 0.21
4→10	0.13 ± 0.14	-0.60	-34.37 ± 0.06	-34.24 ± 0.12
1→14	0.05 ± 0.09	-1.08	0.72 ± 0.09	0.77 ± 0.01
14→23	-0.12 ± 0.27	-0.38	-3.15 ± 0.18	-3.27 ± 0.20
4→2	-0.68 ± 0.18	2.07	-4.14 ± 0.05	-4.82 ± 0.17
1→3	-1.38 ± 0.10	-1.59	-10.41 ± 0.08	-11.80 ± 0.06

Table 4.5: Relative binding free energies compared with experimental values. a) errors were around ± 15% of original IC50 values[112], which corresponds to ± 0.21 kcal mol<sup>-1</sup>

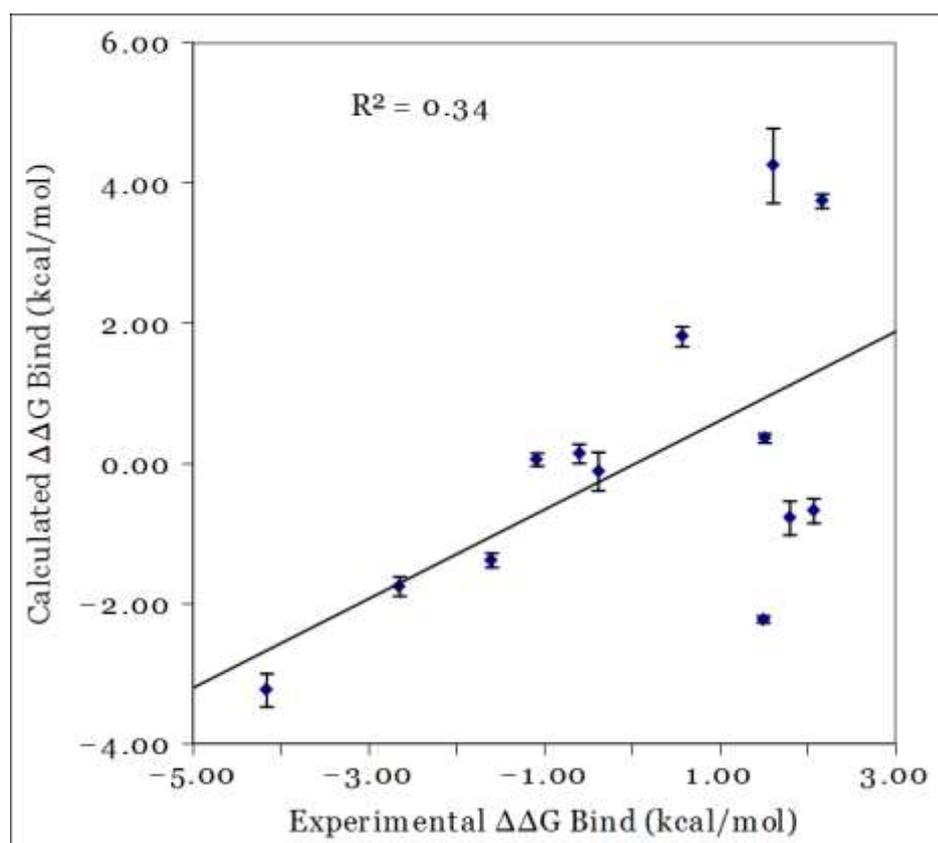


Figure 4.18: Experimental relative binding free energies versus the calculated binding free energies with predicted waters included.

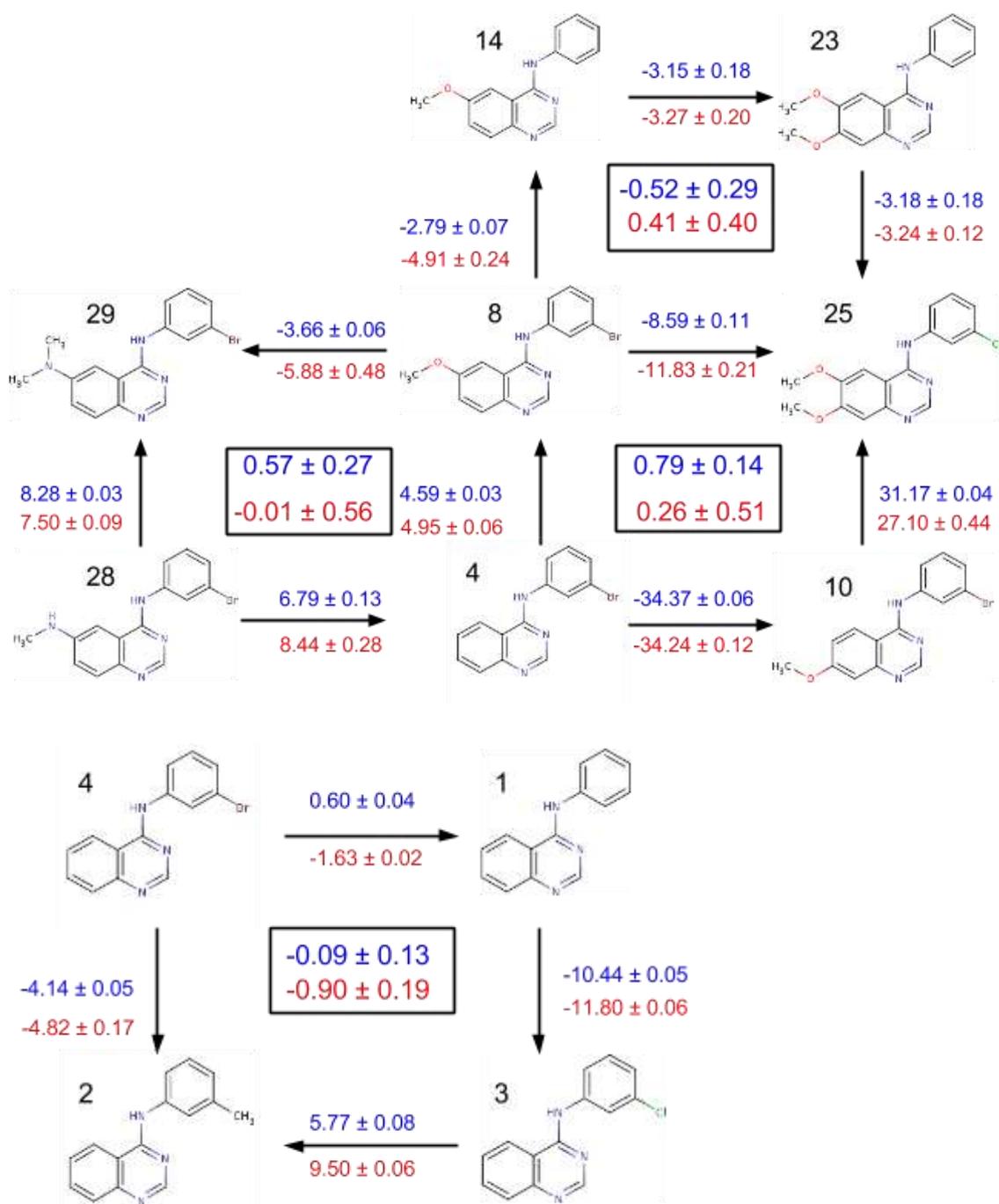


Figure 4.19: Thermodynamic cycles for perturbation simulations. Numbers accompanying arrows indicate the free energy change of a perturbation (blue for solvent leg, red for bound leg). Numbers in the middle of a cycle denote the sum of all steps in a cycle (blue for solvent leg, red for bound leg).

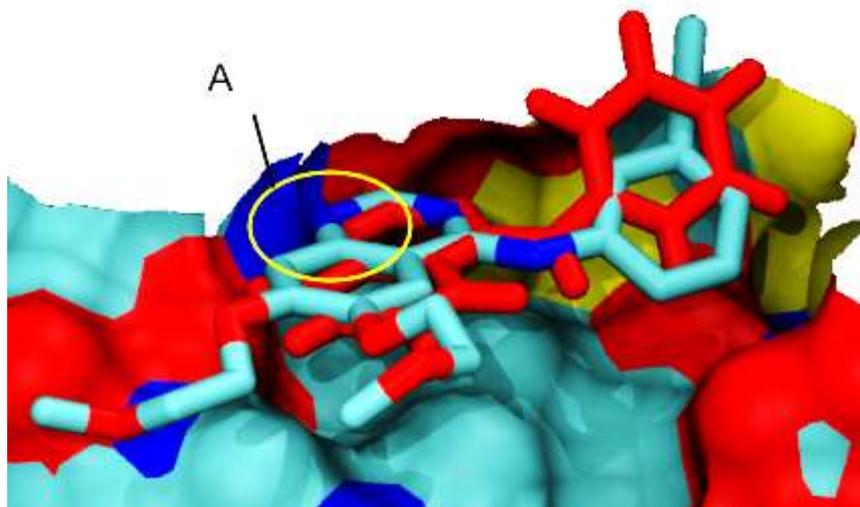


Figure 4.20: The binding pocket of EGFR with erlotinib bound (PDB: 1M17). Final position of ligand 1 when additional waters are present is shown in red, the hydrogen bonding interaction with the hinge region is circled in yellow (A).

The addition of the predicted waters has not prevented ligand 1 from moving into the back of the selectivity pocket (see figure 4.20); however, it appears that the ligand is rotating to fill the pocket in such a way as to maintain the conserved hydrogen bond to the hinge region. The bound leg thermodynamic cycles also converge well to 0 (see figure 4.21, and appendix 4. Also, refer to section 4.3 for a discussion on the free legs, which remain unchanged), suggesting that the protein-ligand system is relatively self-consistent. This closure is better than the results without the predicted waters, which may suggest that the reduced ability of the ligand to move in the pocket allows for better phase space overlap. If so, this may explain why no cause for the poor closure in the previous simulations was found.

Taken together, it appears that the inclusion of the predicted waters does make some difference to the quality of the simulations and the results; however, the correlation is still considerably weak, and it therefore seems likely that some of the remaining deviation in the calculated results from the experimentally determined values stems

from either some other aspect of the system setup, or from the forcefield itself. The possibility of the the phenyl group flipping is countered by the new hydration state: water W3 would need to be displaced, and this would incur a considerable penalty. Of particular interest is the fact that many of the outliers in the dataset appear to involve perturbations of the halogen group.

A possible cause for the poor scoring of perturbations involving halogens is how halogens are handled by the GAFF forcefield. The high electronegativity of halogens makes them particularly susceptible to polarisation effects, which also relates to the ability of halogens to make noncovalent halogen bonds[127]. Interestingly, QM/MM computational studies have identified halogen bonding of inhibitors with protein residues, including halogen bonding to backbone carbonyl groups[128][129]. In the case of EGFR kinase, the hydrogen bond donor groups of the selectivity pocket are all sequestered in the  $\beta$  sheet, so a halogen bond directly to the protein seems unlikely. Nonetheless, the halogen bonding may be of importance in the free simulations, and so further simulations were performed to attempt to probe this effect, as detailed in the following section.

### 4.6 QM/MM corrections

To investigate the possibility of the AMBER treatment of halogens being a weak link in the simulation protocol, hybrid QM/MM simulations were run for those perturbations involving halogens.

### 4.6.1 System setup

Two sets of systems (ligand in water, and ligand in protein) were produced as previously described (see section 4.2), except using the GCMC predicted waters for the bound leg. Notable exceptions include that only the lambda 0 and lambda 1 states were subjected to MC simulation, as the intermediate states are not required to produce the correction. Additionally, 100 million MC moves were performed for both these lambda windows of the perturbations listed in table 4.6, with snapshots produced every 20000 moves to produce 5000 snapshots.

The snapshots were then used to produce input for the *Gaussian 09 rev A.02* program. This input consisted of a conversion of the ligand into a Gaussian-compatible format, as well as the conversion of other atoms within 15 Å into point charges. Additionally, a ligand in vacuum system was produced from each snapshot.

*Gaussian 09 rev A.02* was then used to calculate the single point QM energies of each snapshot, both surrounded by point charges and in vacuum, using B3LYP density functional with the 6-31G\* basis set, with no symmetry constraints.

### 4.6.2 Results

The QM/MM corrected binding free energies are overall closer to the experimental values than the MM calculations (see figure 4.21); however, even when the differences between the experimental and QM/MM results are summed, the total is just 0.65 kcal mol<sup>-1</sup> closer to experiment than the MM calculations. Given that the average error for the QM/MM calculations are greater than this, it is not possible to say with any certainty that the QM/MM corrections actually have any utility.

Perturbation	QM/MM corrected $\Delta\Delta G_{\text{bind}}$ (kcal mol <sup>-1</sup> )	MM $\Delta\Delta G_{\text{bind}}$ (kcal mol <sup>-1</sup> )	Exp <sup>a</sup> (kcal mol <sup>-1</sup> )
4→1	-0.76 ± 0.45	-3.04 ± 0.07	1.50
3→2	3.53 ± 0.81	3.93 ± 0.09	2.17
3→8	3.02 ± 1.18	4.57 ± 0.32	1.61
8→25	-1.34 ± 1.18	-3.97 ± 0.21	-4.16
4→2	-0.50 ± 0.82	-1.16 ± 0.32	2.07
1→3	-2.33 ± 0.64	-1.38 ± 0.23	-1.59

Table 4.6: QM/MM corrections to the binding free energy. a) errors were around ± 15% of original IC50 values[112], which corresponds to ± 0.21 kcal mol<sup>-1</sup>

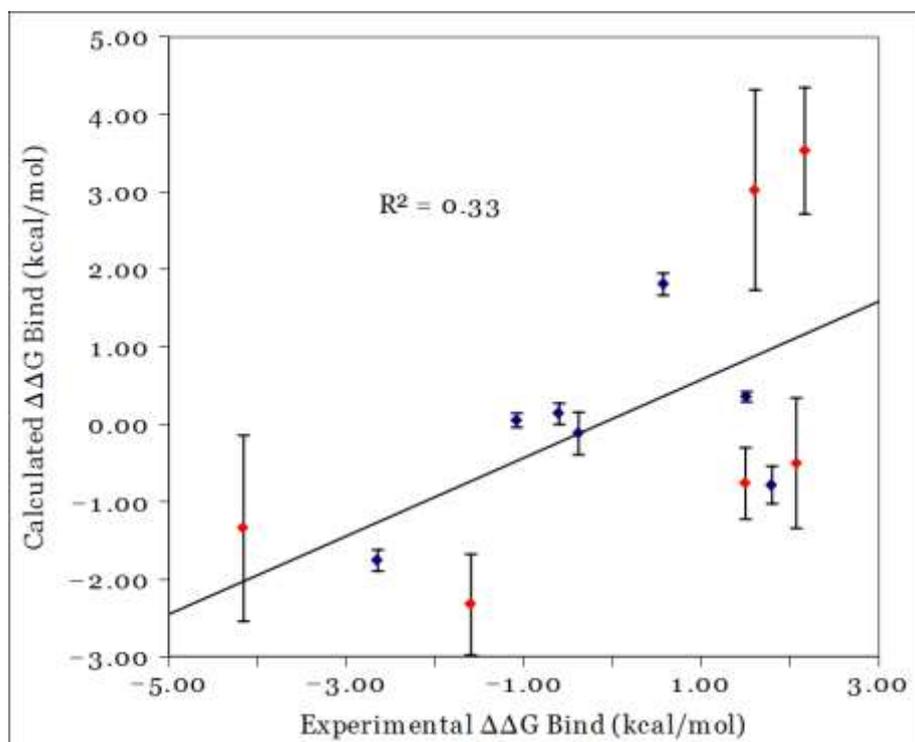


Figure 4.21: Experimental relative binding free energies versus the calculated binding free energies with predicted waters. MM (blue) and QM/MM corrected (red) results are included.

When plotting the QM/MM corrected energies against the experimentally determined values, the regression coefficient remains at 0.33 (compared to 0.34 for the MM calculations). The PI was calculated as 0.81, which represents a mild improvement (MM alone resulted in 0.78), though such an improvement is unlikely to be significant given the increase in error associated with QM/MM.

The lack of improvement due to the inclusion of polarization effects by QM/MM makes it impossible to rule out some unknown deficiency in the system setup, or experimental values, or to comment further on whether the treatment of halogens is a limiting factor in the prediction of binding affinities in the present study. Nonetheless, while the QM/MM corrections do not have a significant effect on the dataset, considering the corrections are relatively simple, it is encouraging that the method performs about as well as MM. It may be that future refinements of this QM/MM methodology will result in increasingly accurate free energy predictions.

## 4.7 Conclusions

The present study has investigated the potential of widely available, cheap, and rigorous free energy techniques in the prediction of the relative binding free energies of inhibitors of EGFR kinase. In the process, a number of potential pitfalls of these methods have been highlighted, including the hydration state of the pocket (as has also been highlighted by Balius & Rizzo (2009)[6] and Michel et al. (2009)[110]), and the possible error introduced by the lack of polarization in the MM forcefield.

The correlation between calculated and experimental relative binding free energies was shown to be poor for the techniques implemented, the cause of which remains unclear. Nonetheless, reasonable predictive indices were obtained, suggesting that the ranking

of the inhibitors is generally good. Additionally, given that the QM/MM corrections maintained the correlation and predictive indices, in the future, further refinement of these relatively naïve QM/MM techniques may help in systems where polarization is important

## Chapter 5: Conformational dynamics of EGFR

As discussed previously (see section 2.2), there are large conformational changes involved in the activation of EGFR kinase, and a number of theories regarding the mechanism by which mutations activate the protein have suggested the stabilisation of the active conformation and/or destabilisation of the inactive conformation[35], [62], [63]. Other theories suggest mutation perturbs the kinase into mutant-specific conformations[65], or limits the accessibility of important WT conformations (for example, the disordered C-helix conformation[43], and the inactive A-loop helix formation[62]).

Intuitively, one could test these theories by examining the mechanism by which mutations introduced *in silico* into the inactive structure promote activation, and by examining the relative stability of the active and inactive mutant structures relative to the WT.

Unfortunately, it is very difficult to sample the activation pathway due to the long time scales upon which activation occurs (estimated to be on the microsecond-millisecond time scale, based on studies on ABL kinase)[102]; until recently[5], even the longest MD simulations failed to sample the entire pathway[43], and such sampling is only feasible with exceptional computing resources. Nonetheless, even in much shorter time scale simulations, a number of studies have noted differences in the sampling of EGFR kinase between mutants[62], [63], [65], confirming that much can be learned even through comparatively short simulations.

In this chapter, the results of various MD techniques will be used to evaluate the effect of activating mutations on EGFR kinase dynamics, and attempt to relate these back to

the nature of the mutation. Additionally, since the enhanced sampling techniques trialed in the present study have not been compared previously, the relative performance of each MD technique will be discussed.

## 5.1 System setup

System	Conformation	template PDB structure	Residues utilised
WT	Active	1M17	672-962*
WT	Inactive	2GS7	672-958*
G719S	Active	2ITP, 2ITQ	671-981*
G719S	Inactive	2GS7	672-957*
L858R	Active	2ITV	701-987
L858R	Inactive	2GS7	678-958*
ELREA deletion	Active	1M17	672-958*
ELREA deletion	Inactive	2GS7	678-958*

Table 5.1: PDB structures and residues utilised for each system of the present study.

Note that the numbering system differs between PDB files. \*these structures use an alternate numbering system to the one used in the present study, which is 24 amino acids shorter[29], [35], [34]

The PDB structures in table 5.1 were used to create models of the WT and mutant EGFR kinase monomers. The decision to utilise only the monomers was made for a number of reasons: Firstly, the literature is abundant with studies on EGFR kinase monomers, making comparison of observations relatively straightforward. Secondly, at the time the present study was commenced, it was suggested that activating mutants of

EGFR were active in their monomeric form[29]; however, more recent studies cast doubt upon this assumption [32], [43]. Thirdly, simulation of the dimer is considerably more demanding, and would have provided less data regarding the ability of the mutations to perturb the WT dynamics.

The choice of using only the kinase domains of EGFR was made similarly: The kinase region alone was predicted to be activated in the mutants and inactive in the WT[29], and modelling of the entire EGFR protein would have required more computational resources than were available at the commencement of the study (however, such modelling has been recently accomplished by Arkhipov et al. (2013)[32]).

Since no inactive mutant structures exist, and there are missing residues in all but the 1M17 structure, modelling of the mutant structures and missing residues is necessary, the procedure was as follows:

The Ensembl database[130] was used to extract the WT amino acid sequence to use as a template for MODELLER[131], and mutant sequence templates were adapted from the Ensembl sequence by hand, comparing with the sequences in the PDB files where available (see table 5.1). These amino acid sequences and PDB structures were used as the inputs for the MODELLER program, which was used to produce 20 models per mutant in both the active and inactive conformation. Models were then validated using the program WHAT\_CHECK[132]. Four models were chosen for each system (with 3 mutants and the WT in both the inactive and active conformations, this totalled 32 models) on the basis of their Ramachandran plots, WHAT\_CHECK Ramachandran Z-score and chi-1/chi-2 rotamer normality scores. Structures were also visualised in VMD to ensure the structures conformed to the current structural knowledge of EGFR kinase. All these models were utilised in the cMD simulations, and half of the models

were used for the enhanced sampling methods (of the 4 models for each system, 2 were taken at random).

Crystallographic waters were retained, and all other molecules discarded. Only the main chain of the protein was used, and for 1M17 structures, a C-terminal extension (residue numbers 959+) was also discarded since its conformation (extended away from the protein) was not in the expected conformation (where it should wrap up beyond the hinge region to interact with the N-lobe[38]), and is therefore a possible crystallographic artefact.

These proteins were then protonated on polar hydrogen sites using the HBONDS module of the WHATIF program[113]. Due to the challenging nature of assigning histidine, glutamine and asparagine protonation[133], the WHATIF program calculates the hydrogen bonding network of the protein, and may flip side chain residues based on this network. To ensure these flips were indeed reasonable, the resultant structures were examined by eye. In each case, the flips were retained. The system was then run through the tleap module of the AMBER Tools package[114] to add the non-polar hydrogen atoms and solvate the system with a box of TIP3P waters extending 10Å from the surface of the protein. At this stage, Cl<sup>-</sup> ions were also added to neutralise the system (the number of Cl<sup>-</sup> ions was different between mutants and conformational states due to the differences in charge introduced by mutants, as well as differences in the number of residues in the models. 3 were required for the active L858R, 4 for the inactive L858R, 1 for the active G719S, 2 for the inactive G719S, 2 for active deletion, 3 for the inactive deletion, 1 for the active WT, 2 for the inactive WT).

A WT crystal structure (1XKK) exists for the inactive conformation, however it contains a number of missing residues (residues 750-753 and 867-876); in an effort to eliminate the introduction of errors by modelling in so many missing residues, the crystal

structure 2GS7 was used. Since the 2GS7 crystal structure contains a V924R mutation, reverse mutants were produced to which activating mutations were modelled later (with the exception of the WT system).

All molecular dynamics trajectories (both production and equilibration) were produced using the sander program from the AMBER simulation package[114], and the AMBER99 Stony Brook forcefield[115].

The solvated system was equilibrated first by two rounds of minimisation. First a minimisation was performed with a restraint of  $1000 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$  on all non-hydrogen atoms, with 5000 steepest descent and 5000 conjugate gradient steps. Next, a similar minimisation was performed with restraints only on the protein heavy atoms. The system was then heated to 300 K over 500 ps, maintaining the protein restraints, and using the NVT ensemble. In this way, the waters were able to become disordered. To equilibrate the pressure, the NPT ensemble was then used, allowing the volume to adapt to the new configuration for another 500 ps. The system was then cooled to 100 K using the NVT ensemble in preparation for 6 rounds of minimisation (1000 steepest descent, 1000 conjugate gradients), reducing the restraints by 80% each round, with the 6th round having no restraints. The system was then heated in 10 ps steps by 80 K until the temperature was at 300 K. Finally, the system was run using the NPT ensemble for 200ps until the total energy appeared equilibrated.

The deletion models were subjected to an additional, initial round of minimisation, to alleviate possible steric clashes arising from the modelling of the 5 amino acid deletion. This consisted of restraining all the residues with a force of  $1000 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ , except for 4 residues in the immediate vicinity of the ELREA deletion, and performing 1000 steepest descent, and 4000 conjugate gradient minimisation steps.

Thus 4 models of each system in each mutational state (WT, L858R, G719S, deletion) were run in both the inactive and active conformations (total of 32 models) for 1  $\mu$ s at 300 K at 1 atm using the PMEMD module of the AMBER software suite. A timestep of 2 fs was used with SHAKE constraints, a cutoff of 10 Å. A Langevin thermostat was employed (a collision frequency of 5 ps<sup>-1</sup>) with a Berendsen barostat to maintain the NPT ensemble. Periodic boundary conditions were maintained throughout using the particle mesh Ewald (PME) method.

## 5.2 Setup of enhanced sampling methods

All the enhanced sampling methods employed in the present study require some degree of parameterisation. The process of parameterising the sampling methods is relatively straightforward; however, reckless assignment of parameters can easily lead to an inefficient boost to sampling, or an excessive boost that causes the system to become unstable.

It is important to note that the parameterisation of all the enhanced sampling methods is dependent to some degree on the size of the system in question. For AMD and RDFMD, a boost in potential is given that increases the energy of the system in proportion to the number of protein residues receiving a boost. In DMDMD, the local scale parameter is roughly proportional to the system size. To overcome these obstacles a systematic parameterisation was carried out for each method, as described below.

### 5.2.1 AMD

The parameterisation process is outlined in the AMBER manual[134]. Briefly, an overall energy threshold ( $E_{\text{threshP}}$ ), a torsion term energy threshold ( $E_{\text{threshD}}$ ) and weighting factors for each of the boost potentials ( $\alpha_{\text{P}}$  and  $\alpha_{\text{D}}$ , respectively) must be chosen (see section 3.8.1).

The AMBER manual states that a value of  $E_{\text{threshD}}$  should correspond to the average dihedral energy of a test simulation of the system, plus  $3.5 \text{ kcal mol}^{-1} \text{ residue}^{-1}$ .  $\alpha_{\text{D}}$  should have a value of  $0.7 \text{ kcal mol}^{-1} \text{ residue}^{-1}$ .  $\alpha_{\text{P}}$  should have a value equal to a fifth of the number of atoms.  $E_{\text{threshP}}$  should correspond to the average potential energy of a test simulation of the system, plus a fifth of the value of  $\alpha_{\text{P}}$ . However, the manual also states that these value can be increased or decreased depending on how mild or severe a boost is desired, giving recommendations for the lower limits of values, but not the upper limits.

In the present study, a test simulation was run from the equilibrated inactive WT protein, using the average total energy (instead of the average potential energy) to investigate the effects of an increased boost potential. This represented an increase of approximately  $30000 \text{ kcal mol}^{-1}$  compared to the average potential energy.

Visualisation of the resulting trajectory showed immediate unwinding of helices in the C-lobe, which seems likely due to the excessive boost provided by the AMD potential, since no evidence of this kind of motion in the protein has been reported to date.

Given the propensity of the system to sample unreasonable configurations during AMD, a gentle approach was taken using the guide to parameters given in Hamelberg et al. (2007)[104]. Essentially, the overall boost parameters are the same as those

suggested in the AMBER manual (the calculation of EthreshP and AlphaP is described above); however, the dihedral boost (EthreshD) is calculated as the average dihedral energy of the system during a normal cMD run multiplied by 1.3, whereas AlphaD is simply the difference between EthreshD and the average dihedral energy. The effect of this should be (on average) to maintain a dihedral energy half way between the boost potential and the real energy surface. These parameters were not found to lead to the destabilisation of the C-lobe helices for the duration of the 1 ns test simulation.

Production runs were run using 2 models of each system (total 16 models) for 1  $\mu$ s with all other parameters as per the cMD simulations.

### 5.2.2 DMDMD

In DMDMD, there are effectively three parameters: the local scale parameter and the number and timing of snapshots. The local scale parameter in a diffusion map is usually calculated using multidimensional scaling, however, over the short time scales examined using the diffusion map method in DMDMD, it is assumed that the local scale does not significantly vary in each DMDMD iteration. The number of snapshots utilised in each iteration is important as the diffusion map is constructed using the  $\alpha$ -carbon positions of these snapshots, and DMDMD is more accurate as the number of snapshots increases. Additionally, the total time spanned by one DMDMD iteration must be enough to adequately sample the local minima such that DMDMD can recognise which snapshot is furthest away from that local minima in phase space[106].

To check how effectively the diffusion map is identifying these leading snapshots, the largest output eigenvectors (corresponding to the sampling along the slowest motions of the protein) can be used to rate a given combination of input parameters: a useful

combination of parameters should result in a sizeable difference between the first and second largest eigenvectors. However, assuming sufficient sampling is performed, the choice of local scale parameter does not appear to be critical, with a variety of ranges providing robust results (see section 3.8.2).

In the present study, it was found that where each iteration produced 100 snapshots spaced by 20 ps, with a local scale parameter of 0.9, a separation of around 0.2 was produced between the first and second eigenvectors (with each eigenvector having a range between 0 and 1). The production DMDMD simulations used these parameters for 500 iterations, corresponding to approximately 500 ns of simulation time (however, it should be noted that since each iteration does not necessarily start from the final snapshot of the previous iteration, so the trajectories do not represent 500 ns of accumulated simulation time. This also results in simulations having a different number of snapshots).

### 5.2.3 RDFMD

In RDFMD a target frequency and target atoms must be specified, as well as the scale of the applied boost. The target atoms were taken to be all the heavy atoms of the residues of the A-loop (residues 855-877), C-helix (residues 753-769) or P-loop (residues 720-725). These targets were chosen due to their roles in the activation process: The A-loop and C-helix undergo considerable conformational change during activation (see 2.2), and the flexibility of the P-loop has been suggested to be important in the conformational transition between the inactive and active states[35]. Additionally, each structure contains or is close to the site of an activating mutation.

Since the boost is applied to all the specified atoms of the targeted region, the different size of these regions requires that they be given boosts of a different scale. In particular, the A-loop (being the longest section to receive a boost) was particularly prone to exceeding the temperature cap (hereafter referred to as “overheating”). The temperature cap, itself a parameter, was set to 800K, as was found to prevent *cis-trans* isomerisation of the protein backbone in a previous RDFMD study[135]. Using the paper on parameterisation of RDFMD[108] as a guide, a filter of 201 coefficients was used with a 0-100  $\text{cm}^{-1}$  response range; however, the other parameters must be assigned by trial and error as described below.

To avoid structural instability, overheating automatically terminates the digital filtering, requiring the system to undergo the next iteration of NVT simulation to allow it to cool down. It has been shown that to obtain the best sampling, several digital filtering steps are necessary[108]; however, an excessive boost in an RDFMD simulation can lead to overheating after just one application of digital filtering, and thus inadequate sampling. Thus, in the present study, each target region was tested on the WT active structure using 100 RDFMD iterations (as per a production run). A range of filter scale and filter frequency parameters were tested and evaluated on how many rounds of digital filtering (out of a total of 5) were applied without overheating, and on how the average RMSD of the affected region was affected. Only those combinations of parameters that lead to a probability of greater than 0.2 of completing all 5 rounds of digital filtering per iteration were considered, and among the considered combinations, those resulting in the greatest RMSD in the target region were chosen for production runs. All these test runs were visualised to ensure the given parameters did not lead to structural instability.

Target (number of residues)	Filter scale	Filter delay (fs)
P-loop (6)	1.25	50
C-helix (16)	0.75	50
A-loop (21)	0.75	50

Table 5.2: Filter parameters used for each target in the RDFMD simulations.

The parameters used in the RDFMD simulation are shown in table 5.2. Additionally, a filter delay of 50 steps was used alongside a 201 coefficient filter (see section 3.8.3 for further explanation of these parameters). The AMBER99 Stony Brook forcefield[115] was used with NAMD[136] to perform the RDFMD simulations. To maintain consistency, the equilibrated structures from the other MD simulations were utilised, and to mitigate against possible inconsistencies between the NAMD and PMEMD calculations, an initial equilibration stage of 150 ps was run in NAMD. RDFMD was run for 100 filter applications each spaced by 4 ps steps of conventional NVT MD on all structures.

### 5.3 RMSD results

The RMSD of the C-helix and A-loop backbone C $\alpha$  atoms were calculated with respect to the active and inactive crystal structures (1m17 and 2gs7, respectively). These RMSDs were used to produce box and whisker plots (figures 5.1 to 5.19), which show the minimum and maximum RMSD in a given simulation as vertical bars terminating in a horizontal cap. The box represents the range of RMSDs spanned between the 1<sup>st</sup> and 3<sup>rd</sup> quartiles, with the median represented as a horizontal line intersecting this box.

For cMD, AMD, and DMDMD each simulation was treated separately; however, due to the problems with representing the hundreds of RDFMD simulations individually, a supertrajectory of the RDFMD simulations was constructed before performing the analysis. While this comes at the cost of losing the ability to visualise the RMSD behaviour of each individual simulation, the plots are included to provide insight into the combined sampling of the RDFMD trajectories compared to that of the other enhanced sampling methods.

### 5.3.1 Active A-loop RMSDs

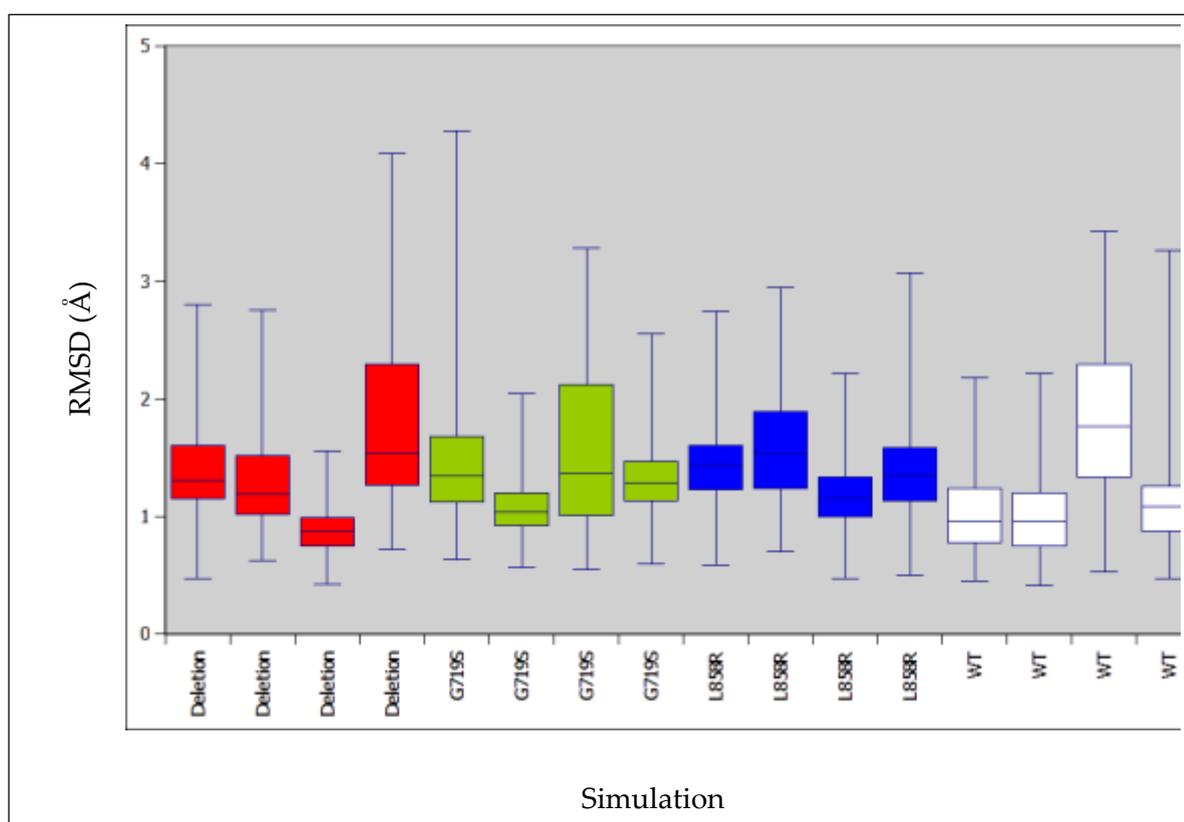


Figure 5.1: Box and whisker plot showing the RMSD of the A-loop backbone C $\alpha$  atoms with respect to the active crystal structure (1m17) for the active cMD simulations.

Figure 5.1 shows a trend for the A-loop of mutants to spend a greater proportion of their time further from the active conformation than does the WT; however, this difference appears modest in most of the systems, and exceptions also exist: one of the deletion mutants exhibits sampling unusually close to the active conformation, and one of the WT simulations exhibits the opposite.

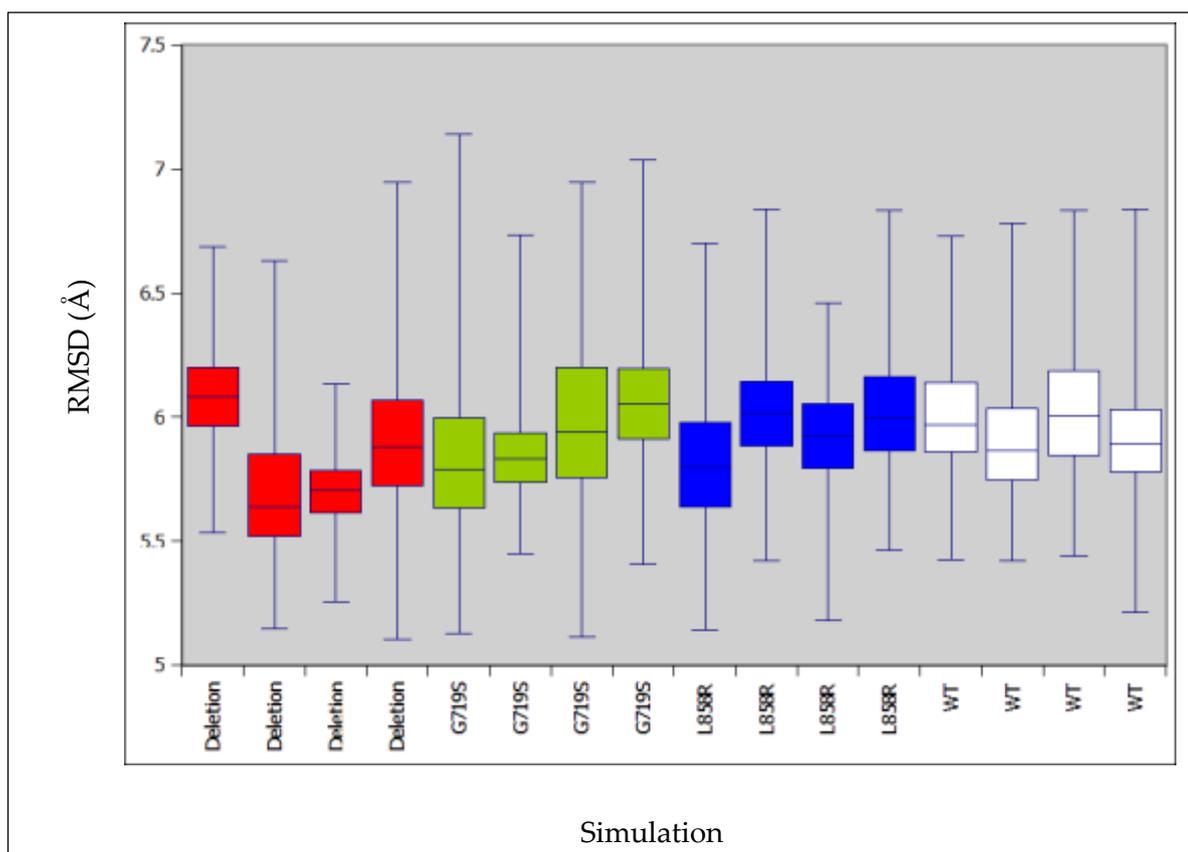


Figure 5.2: Box and whisker plot showing the RMSD of the A-loop backbone C $\alpha$  atoms with respect to the inactive crystal structure (2gs7) for the active cMD simulations.

Figure 5.2 shows a trend for the mutant A-loops to more closely resemble the A-loop of the inactive crystal structure (2gs7) than does the WT. While the trend is only present in the mutants, it is not entirely conserved between simulations of the same system, and is of only a very small magnitude (a difference in medians of approximately 0.3 Å).

at most). Nonetheless, the trend is also reflected in the minimum RMSD to the inactive A-loop

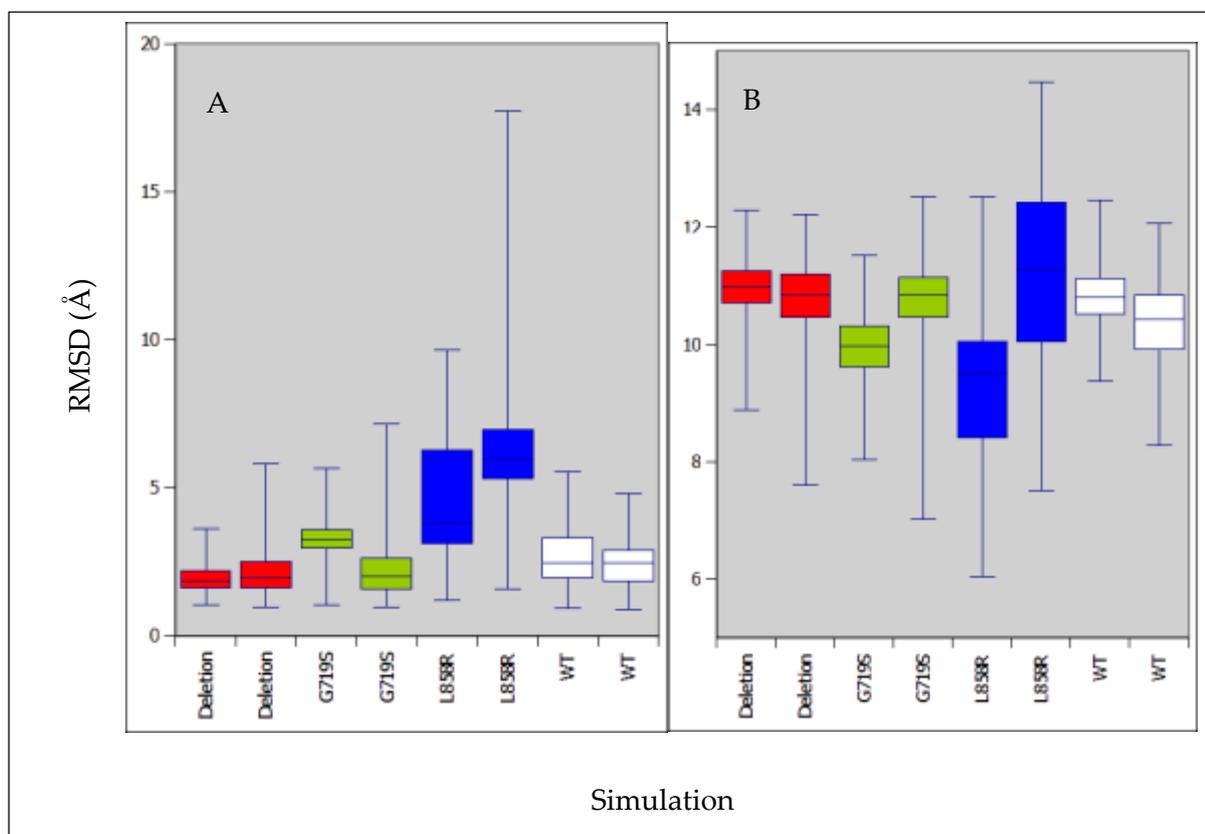


Figure 5.3: Box and whisker plot showing the RMSD of the A-loop backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the active AMD simulations.

In the AMD results (figure 5.3) the A-loop of the point mutants appears to be able to maintain conformations much further from the active crystal structure than the WT, particularly for the L858R mutant. Additionally, the L858R (and one of the G719S simulations) appears to more readily access A-loop configurations more similar to the inactive conformation than the WT. The deletion on the other hand appears to be the least inclined to follow this behaviour, even in comparison to the WT; albeit by a very small margin (with a difference in medians of approximately  $0.6 \text{ \AA}$  for figure 5.3(A)).

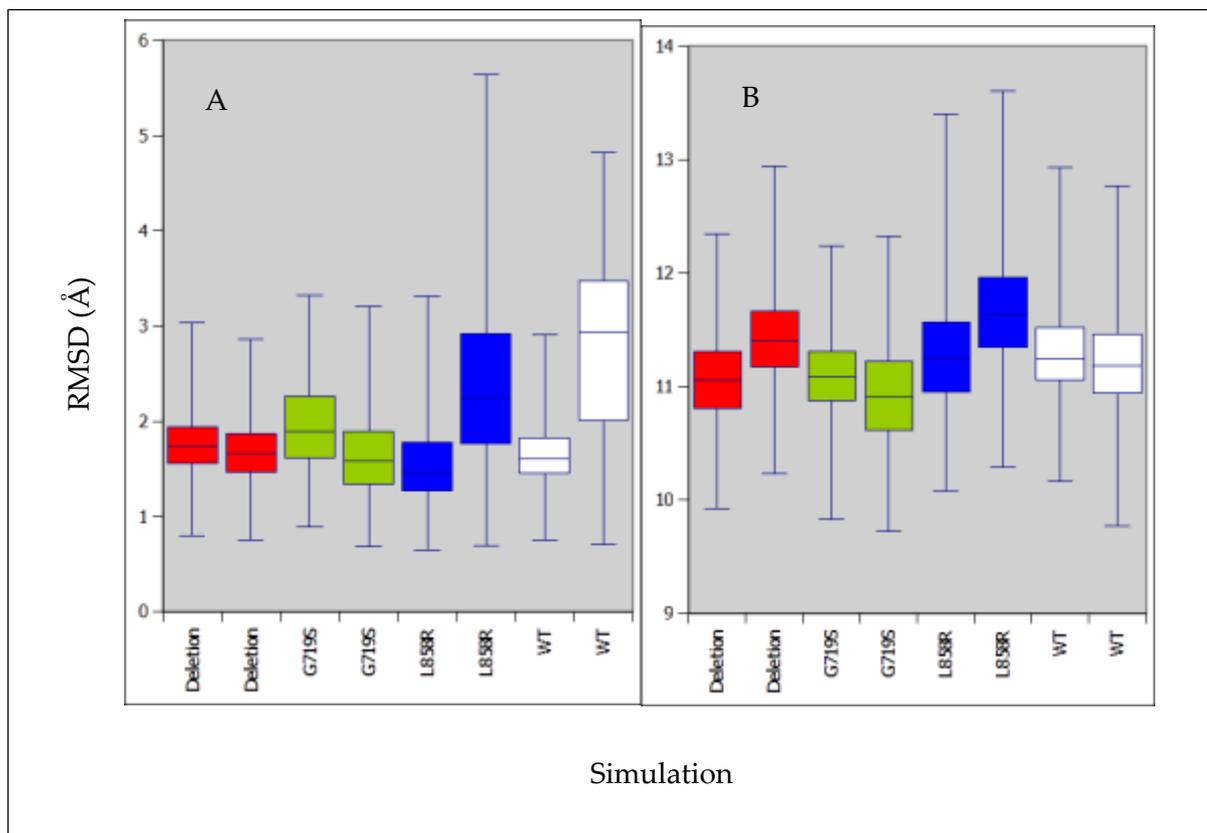


Figure 5.4: Box and whisker plot showing the RMSD of the A-loop backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the active DMDMD simulations.

The A-loop of the L858R also appears, in one simulation, to be able to move considerably away from the active crystal structure (figure 5.4(A)); however, one of the WT simulations also appears able to exhibit this behaviour. Additionally, the RMSD with respect to the inactive crystal structure does not appear to follow any previously identified trend, with the possible exception of the G719S mutant, which appears to have a marginally reduced RMSD (figure 5.4(B)).

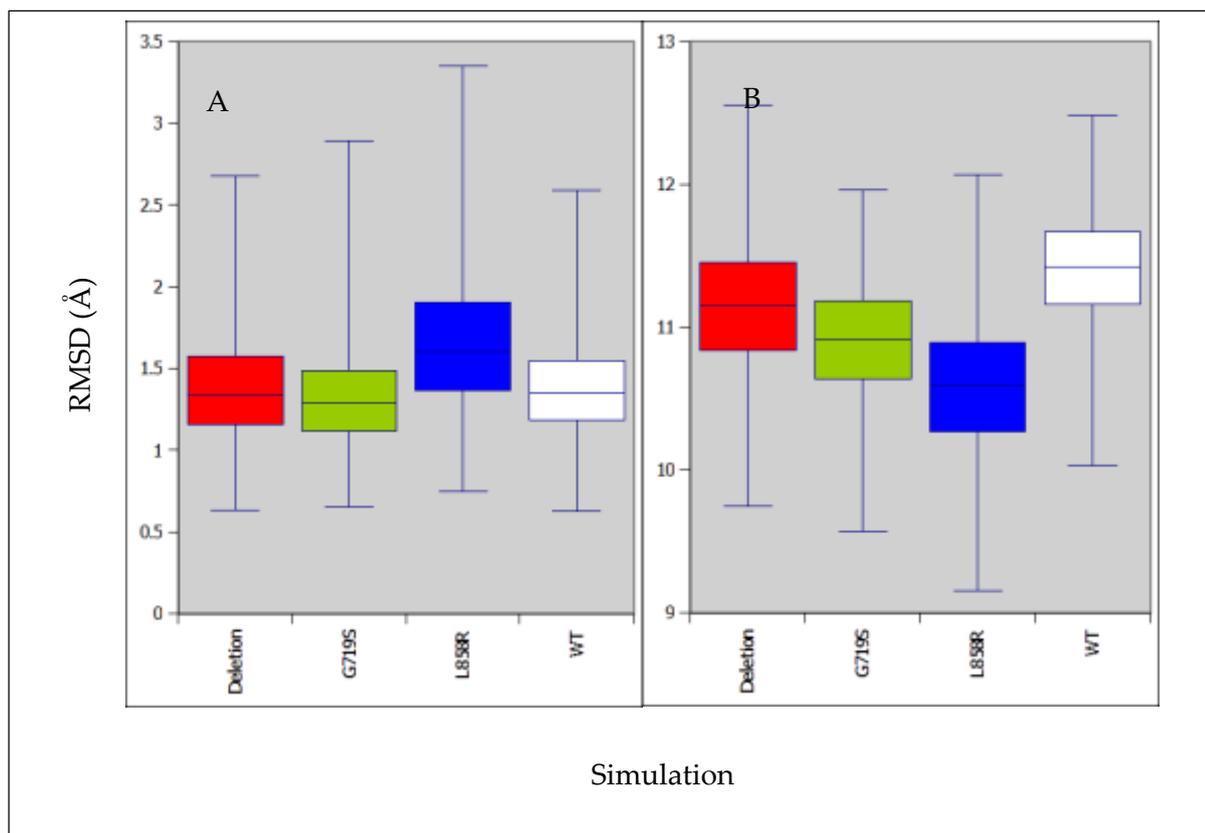


Figure 5.5: Box and whisker plot showing the RMSD of the A-loop backbone C $\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the active RDFMD simulations. Note that each box in the diagram represents the combined data of 30 RDFMD simulations.

In the RDFMD simulations, the A-loop exhibits some of the aforementioned trends, with the L858R exhibiting greater RMSDs with respect to the active structure (figure 5.5(A)), and lower RMSDs with respect to the inactive structure (as does the G719S; see figure 5.5(B)).

Taken together, with the active conformation as a starting point, it appears that the point mutants have a tendency towards higher A-loop RMSDs with respect to the active crystal structure, and (more tentatively) lower A-loop RMSDs with respect to the inactive crystal structure, when the mutants are compared to the WT. Nonetheless, this

difference is often quite small, and possibly not significant in most cases. Additionally, the analyses do not suggest that the WT is completely incapable of similar behaviour to the mutants (in figures A,B and D *at least* one WT simulation has a similar range of RMSDs in its 1<sup>st</sup> to 3<sup>rd</sup> quartiles, in comparison to the point mutants with the most extreme RMSD values). Nonetheless, the data suggests that the configurations of the A-loop away from the active crystal structure, and closer to the inactive crystal structure seem to be more accessible for the point mutants; possibly due to the point mutants transitioning out of the active conformation (this will be discussed in more detail in the following sections). The role of the deletion mutant, however, appears no clearer from the analysis of the A-loop RMSDs.

### 5.3.2 Inactive A-loop RMSDs

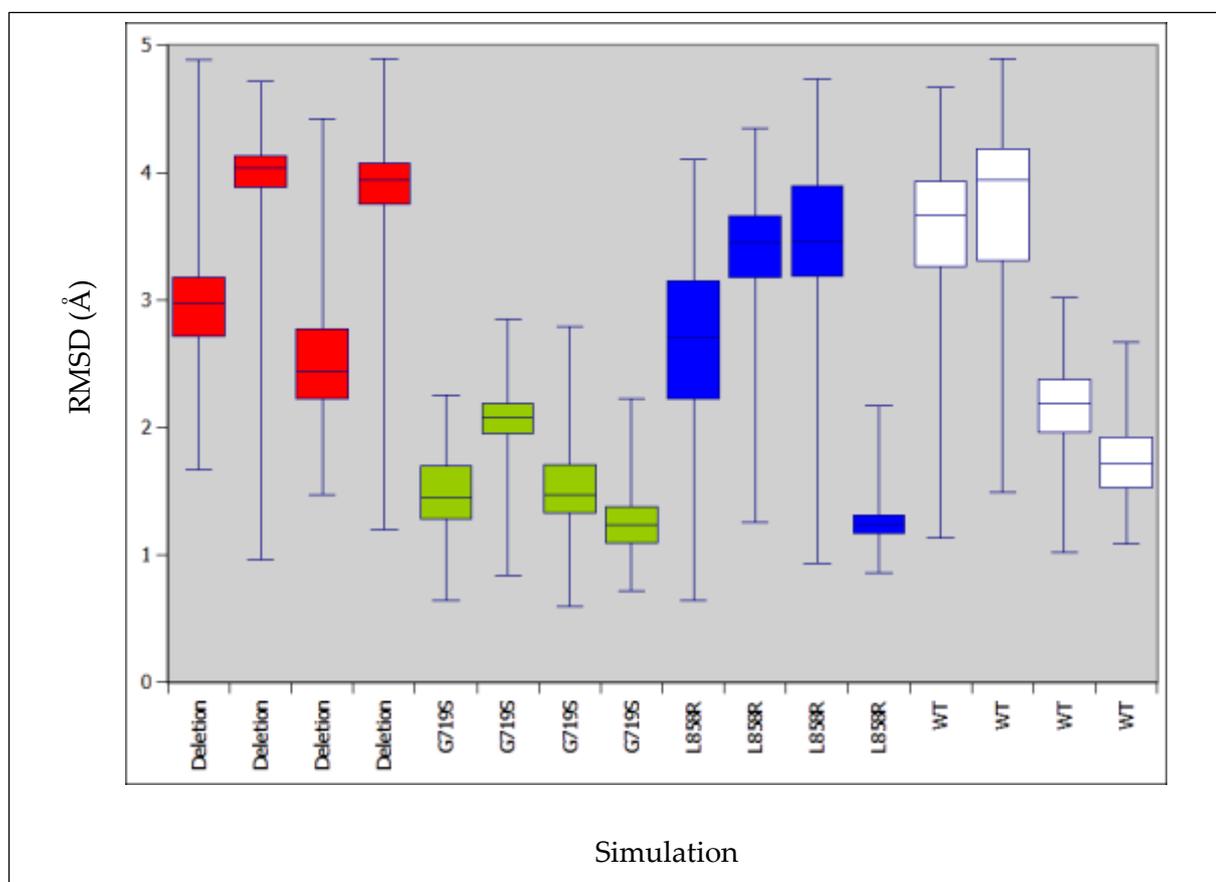


Figure 5.6: Box and whisker plot showing the RMSD of the A-loop backbone C $\alpha$  atoms with respect to the inactive crystal structure (2gs7) for the inactive cMD simulations.

Figure 5.6 shows a much larger spread in median RMSD amongst the simulations (in comparison to figure 5.1), which is probably due to the mobile nature of the inactive A-loop allowing for a much larger range configurations. Interestingly, the G719S mutant appears to remain the closest to the inactive crystal structure on average; however, no other trends amongst the simulations are apparent.

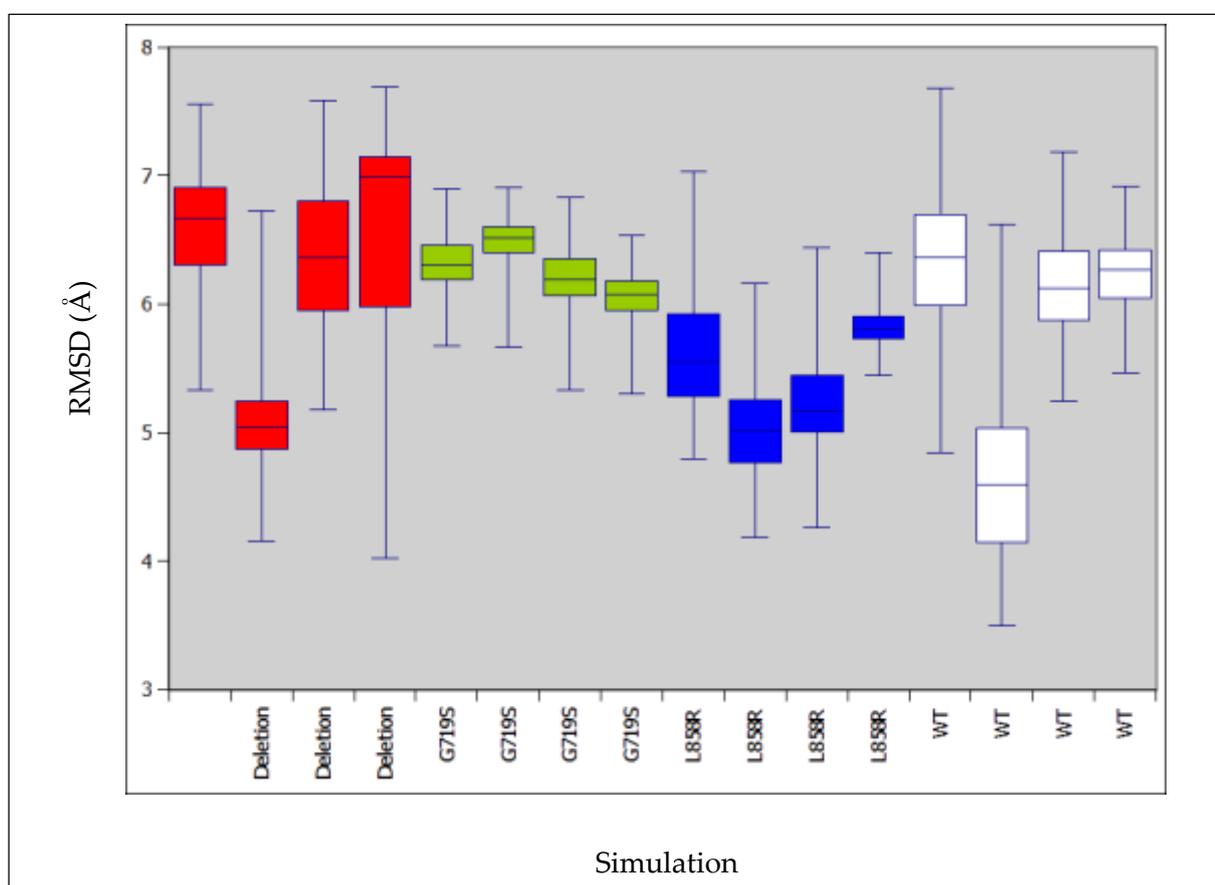


Figure 5.7: Box and whisker plot showing the RMSD of the A-loop backbone C $\alpha$  atoms with respect to the active crystal structure (1m17) for the inactive cMD simulations.

Figure 5.7 shows that the RMSD of the A-loop of the inactive cMD simulations with respect to the active crystal structures also has a large RMSD spread amongst the simulations. It appears that the L858R is the most likely to adopt A-loop conformations with a lower RMSD with respect to the active conformation; however both the deletion and the WT have one simulation each which shows a similar spread of RMSDs, so the lowering of RMSD with respect to the active crystal structure does not seem to be limited to any one mutant. It is perhaps unsurprising that the G719S mutant, which exhibited the least difference in RMSD from the inactive crystal structure, also shows fairly high RMSDs with respect to the active crystal structure (although not remarkably different from the RMSDs of a number of mutant and WT simulations).

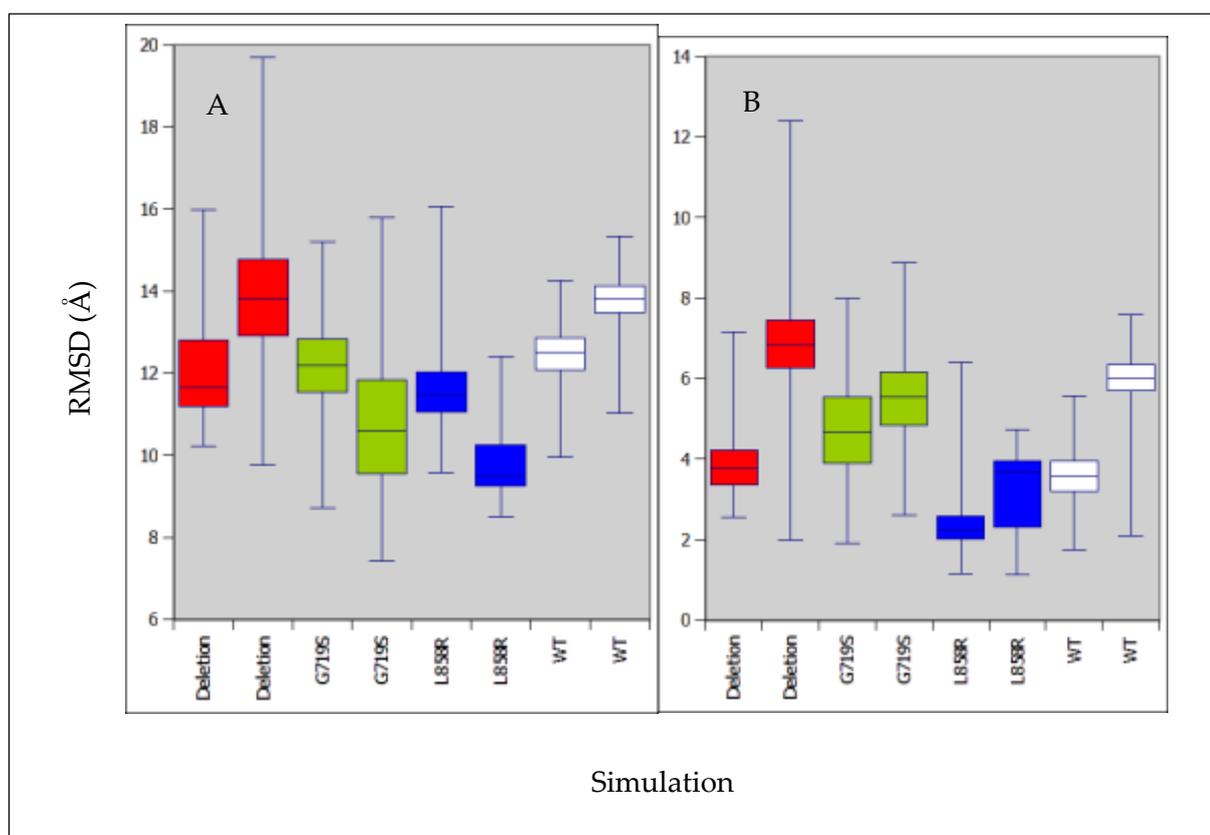


Figure 5.8: Box and whisker plot showing the RMSD of the A-loop backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the inactive AMD simulations.

As with the results from cMD (figure 5.6 and 5.7), the AMD elucidates very little in the way of trends amongst the mutants, the only exception being one of the deletion mutants, which appears to be able to sample a very wide range of RMSDs (figure 5.8). The DMDMD simulations produced even less difference between the simulations (figure 5.9).

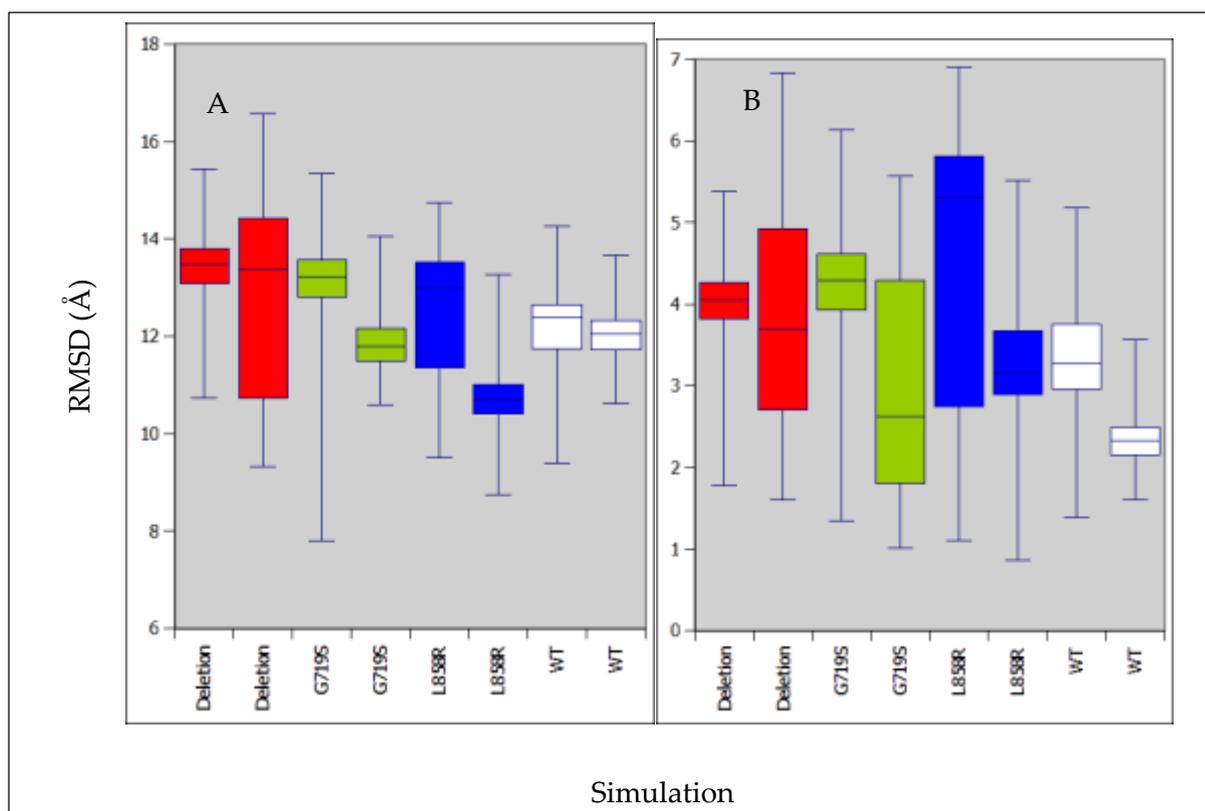


Figure 5.8: Box and whisker plot showing the RMSD of the A-loop backbone C $\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the inactive DMDMD simulations.

The RDFMD simulations (figure 5.9) show a unique trend in RMSD, where an elevated RMSD dominates in the deletion (the median being approximately 1 Å greater than the WT for the deletion).

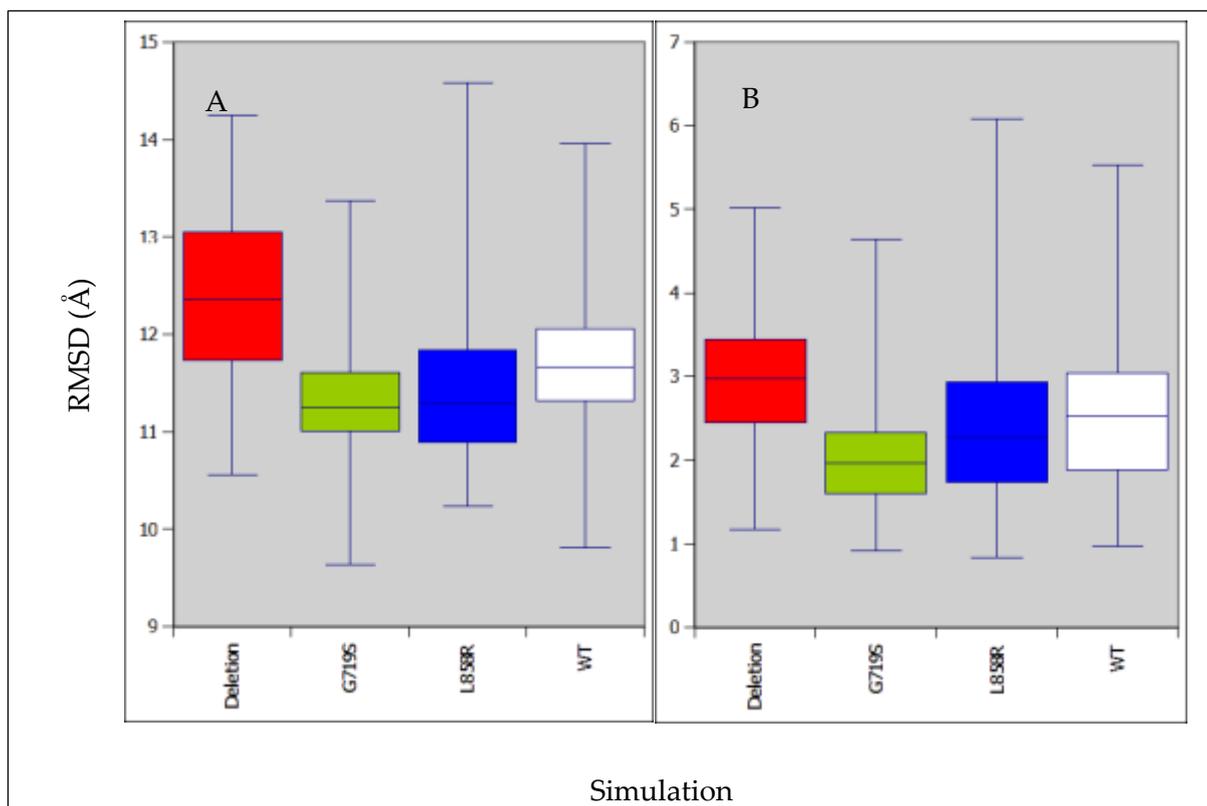


Figure 5.9: Box and whisker plot showing the RMSD of the A-loop backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the inactive RDFMD simulations. Note that each box in the diagram represents the combined data of 30 RDFMD simulations.

Overall, the AMD and RDFMD results suggest that the deletion has a tendency towards accessing configurations with a higher RMSD; however, this tendency is not clear in the DMDMD or cMD results. There also appears to be a tendency for the WT simulations to sample a smaller RMSD range than the mutants in the AMD and DMDMD simulations.

Taken together, it appears that, in the inactive conformation, the A-loop is more liable to produce conflicting RMSD results; none of the above trends are consistent through more than two sampling methods. This could be due to the more mobile nature of the A-loop, as discussed previously, or a lack of repeats.

## 5.3.3 Active C-helix RMSDs

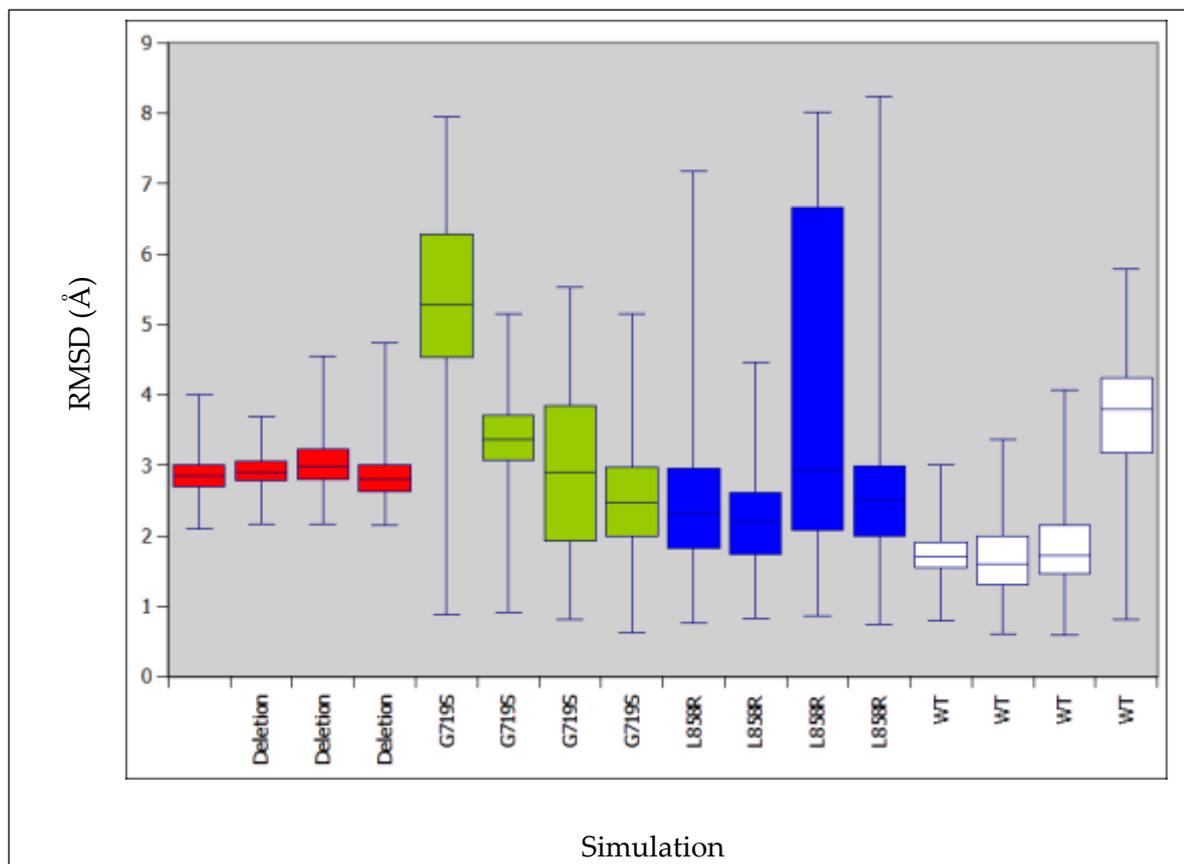


Figure 5.10: Box and whisker plot showing the RMSD of the C-helix backbone C $\alpha$  atoms with respect to the active crystal structure (1m17) for the active cMD simulations.

Figure 5.10 shows the RMSD of the C-helix backbone C $\alpha$  atoms with respect to the active crystal structure, the deletion appears to have the most tightly constrained RMSDs, with the point mutations exhibiting large fluctuations in RMSD, in comparison to the WT. The same simulations with their RMSD calculated with respect to the inactive crystal structure (figure 5.11) show almost the opposite, with the point

mutants again able to sample RMSDs over a much wider range than the WT, and critically, much closer to the inactive conformation.

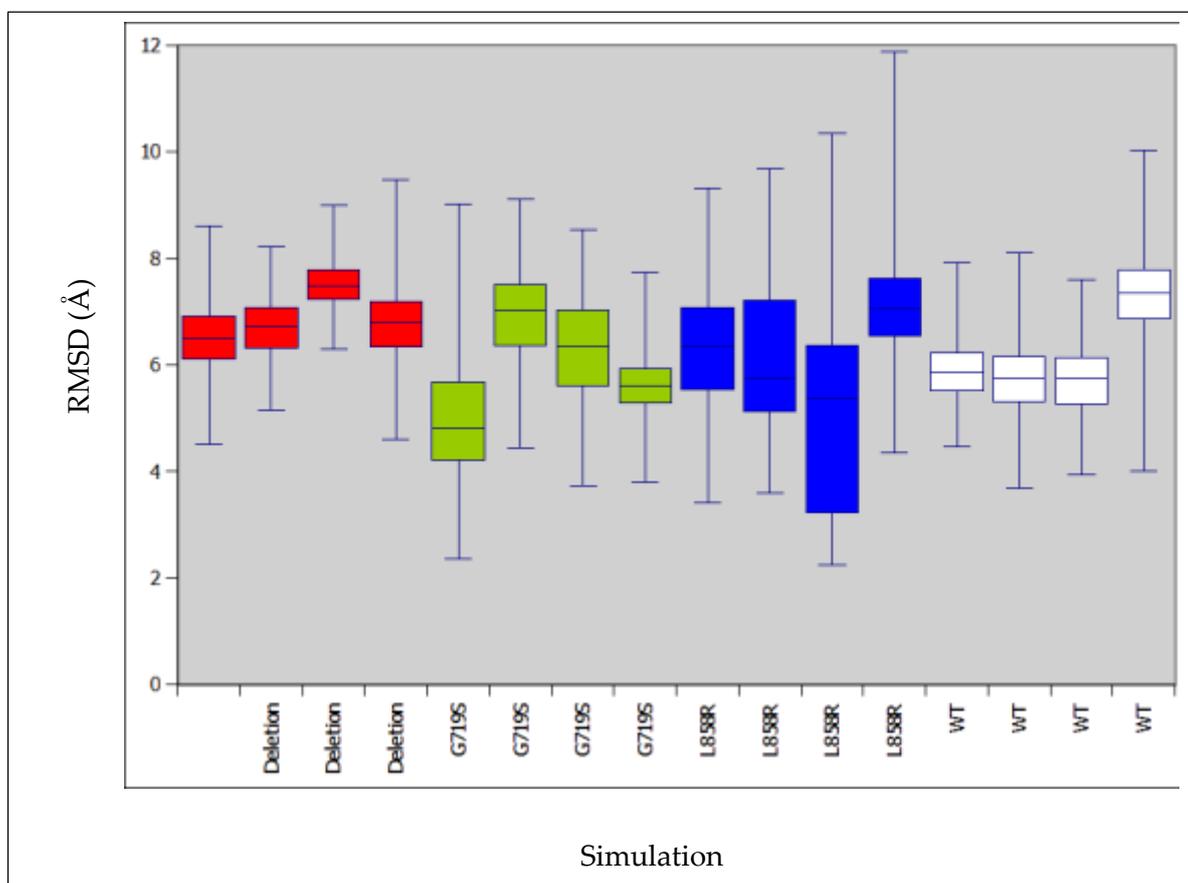


Figure 5.11: Box and whisker plot showing the RMSD of the C-helix backbone C $\alpha$  atoms with respect to the inactive crystal structure (2gs7) for the active cMD simulations.

The AMD results (figure 5.12) are in general agreement with the cMD results; however, it appears that for the AMD simulations, the L858R had a greater impact than did the G719S. Additionally, the WT appears much more mobile in the AMD simulations than in the cMD simulations.

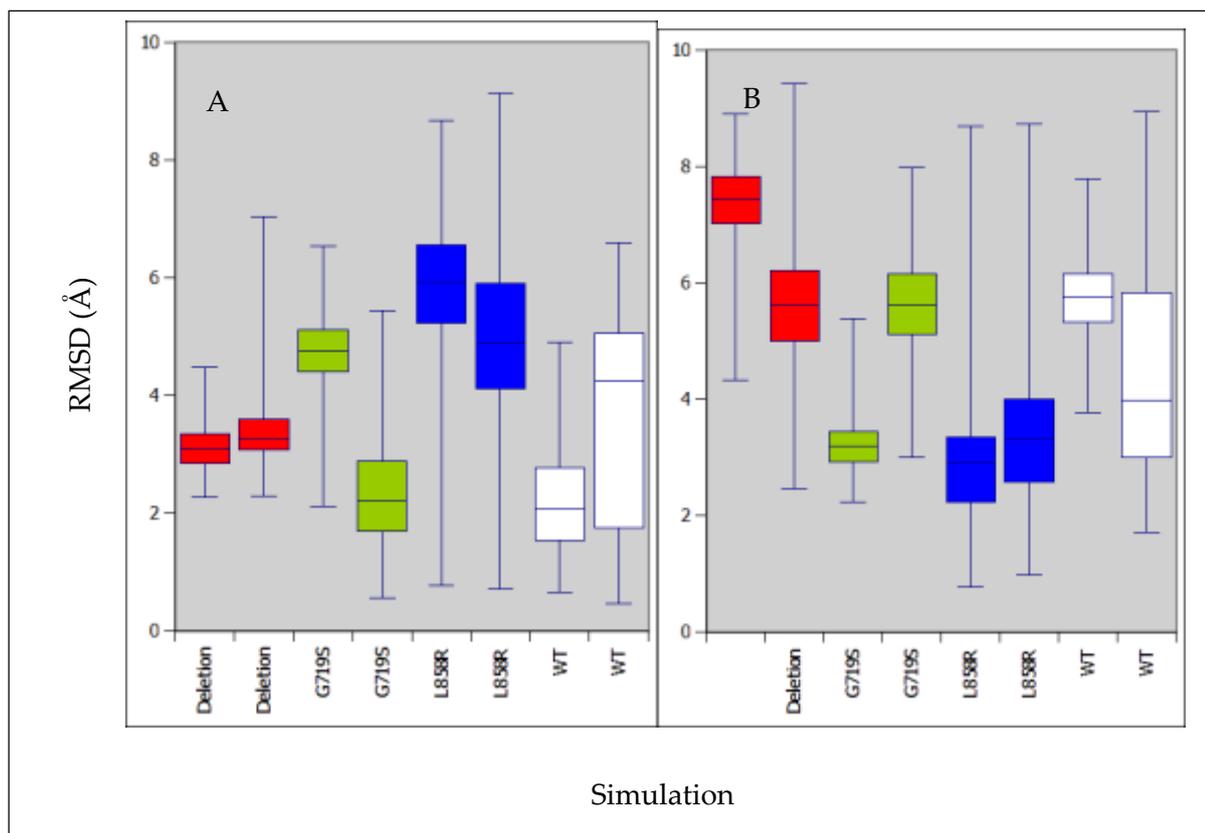


Figure 5.12: Box and whisker plot showing the RMSD of the C-helix backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the active AMD simulations.

The DMDMD simulations (figure 5.13) show a greater range of sampling by the point mutations between the 1<sup>st</sup> and 3<sup>rd</sup> quartiles; however, this sampling is not necessarily further from the active crystal structure or closer to the inactive structure, as was observed for the cMD and AMD simulations. Additionally, the deletion simulations do not appear to sample closer to the active crystal structure than the point mutants; however the range of RMSDs sampled by the deletion is reduced in the DMDMD simulations.

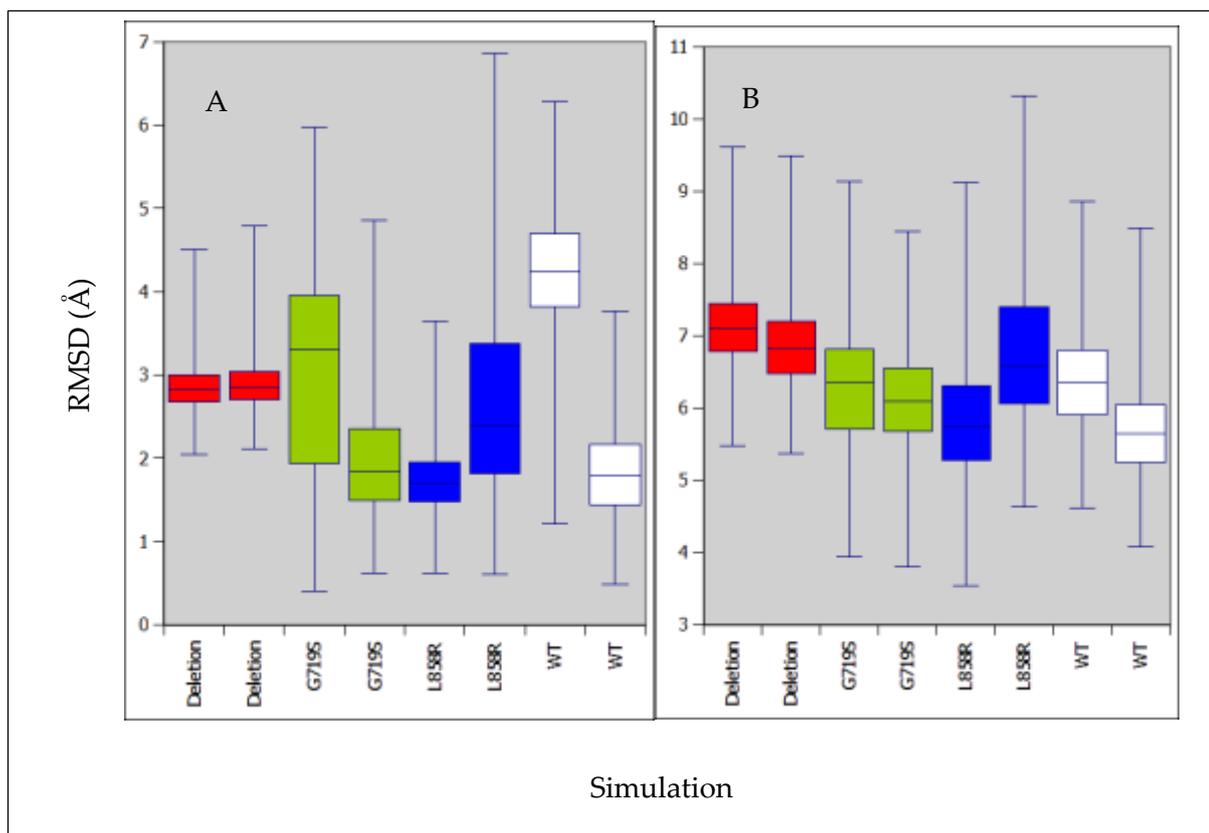


Figure 5.13: Box and whisker plot showing the RMSD of the C-helix backbone C $\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the active DMDMD simulations.

The RDFMD simulations (figure 5.14) show none of the features identified previously; the only characteristic consistent with the results from the other sampling methods is the tendency for the deletion to sample higher RMSDs with respect to the inactive conformation.

The lack of consistency between the RDFMD results and the other sampling methods may be due to the large timescales over which C-helix motions occur (unlike motions of the more mobile A-loop), which would suggest that RDFMD has not been successful in sufficiently boosting the sampling of the regions we are most interested in.

Nonetheless, it is also possible that the analysis of the RMSD of the RDFMD

simulations, in being constructed differently to the other sampling methods, is not conducive to comparison with the other results.

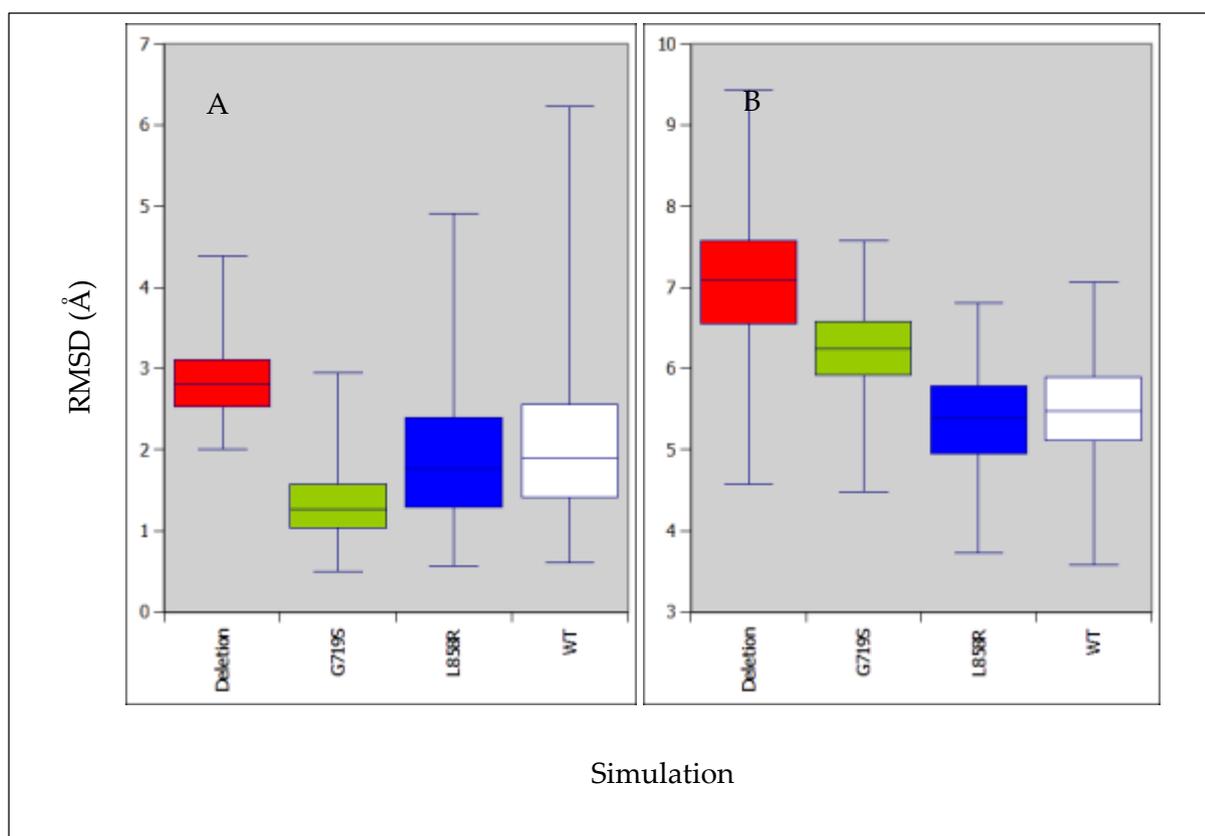


Figure 5.14: Box and whisker plot showing the RMSD of the C-helix backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the active RDFMD simulations. Note that each box in the diagram represents the combined data of 30 RDFMD simulations.

Taken together, the cMD, AMD and DMDMD results show that there is a remarkably large range of RMSDs accessible to the point mutants, and the deletion mutant appears to sample only a very narrow RMSD range, albeit further from the active crystal structure than the WT. The cMD and AMD results suggest the larger range of sampling of RMSDs for the C-helix in the active point mutants leads to decreased RMSDs with respect to the inactive crystal structure, suggesting that the C-helix has become more “inactive-like” in the point mutant simulations, although these findings are not clear in the DMDMD results.

The lack of consistency between the RDFMD results and the other sampling methods is not encouraging, as outlined previously, and further analysis is required to gain an understanding of how the RDFMD simulations differed from the other sampling methods.

### 5.3.4 Inactive C-helix RMSDs

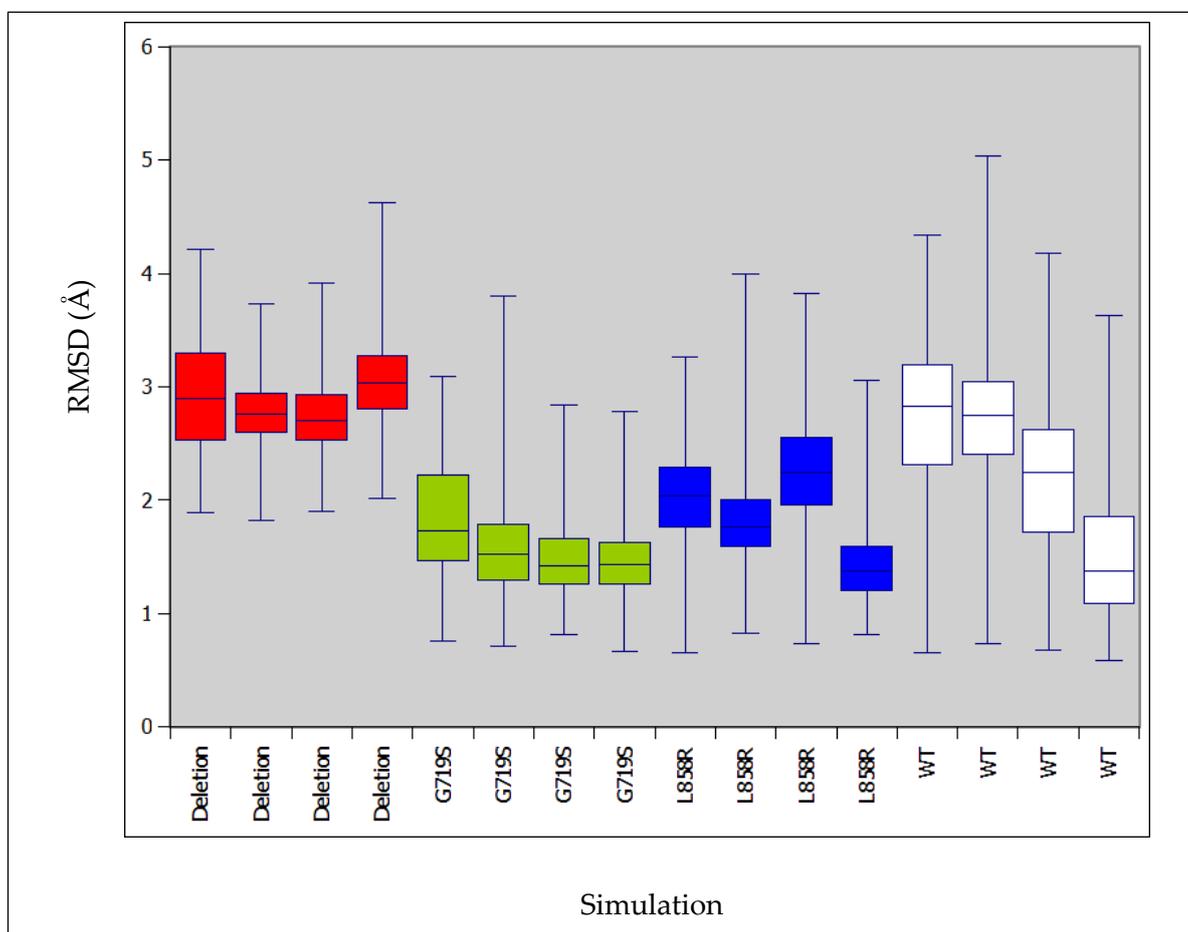


Figure 5.15: Box and whisker plot showing the RMSD of the C-helix backbone Ca atoms with respect to the inactive crystal structure (2gs7) for the inactive cMD simulations.

The trends of the C-helix in the active conformation simulations disappears in the inactive simulations, with the point mutations exhibiting low RMSDs in comparison to the WT and deletion. The deletion mutant C-helix appears to be more mobile in the inactive conformation (compare figure 5.15 with figure 5.10). The deletion mutant also appears to sample slightly closer to the active conformation than the other simulations on average (see figure 5.16), a trend which is not seen in the point mutants.

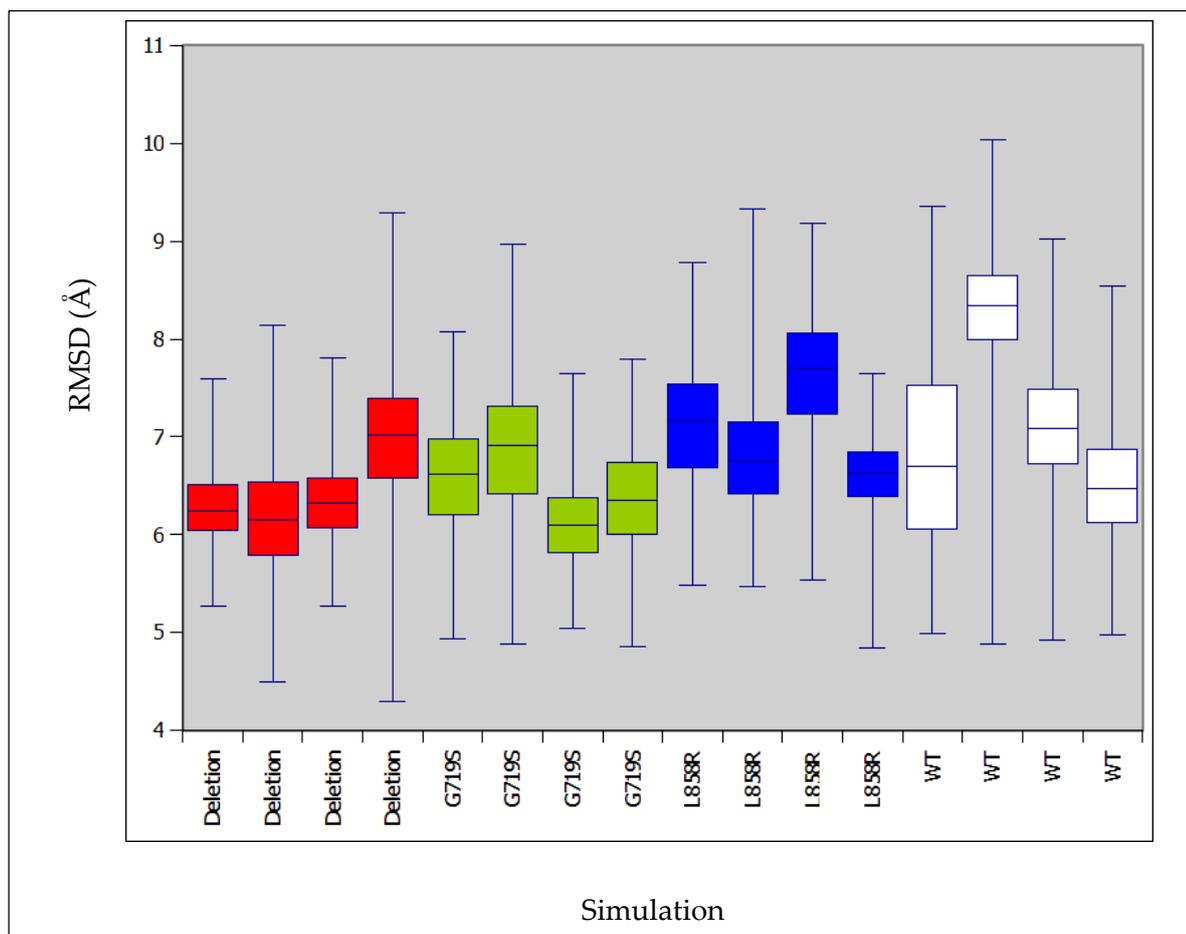


Figure 5.16: Box and whisker plot showing the RMSD of the C-helix backbone  $C\alpha$  atoms with respect to the active crystal structure (1m17) for the inactive cMD simulations.

In the AMD simulations, the trends with respect to the active crystal structure seen in the cMD simulations are lost; however, the tendency for the deletion to exhibit higher

RMSDs with respect to the inactive crystal structure is greatly increased. Additionally, the G719S appears also appears to access higher RMSDs relative to the inactive conformation (see figure 5.17).

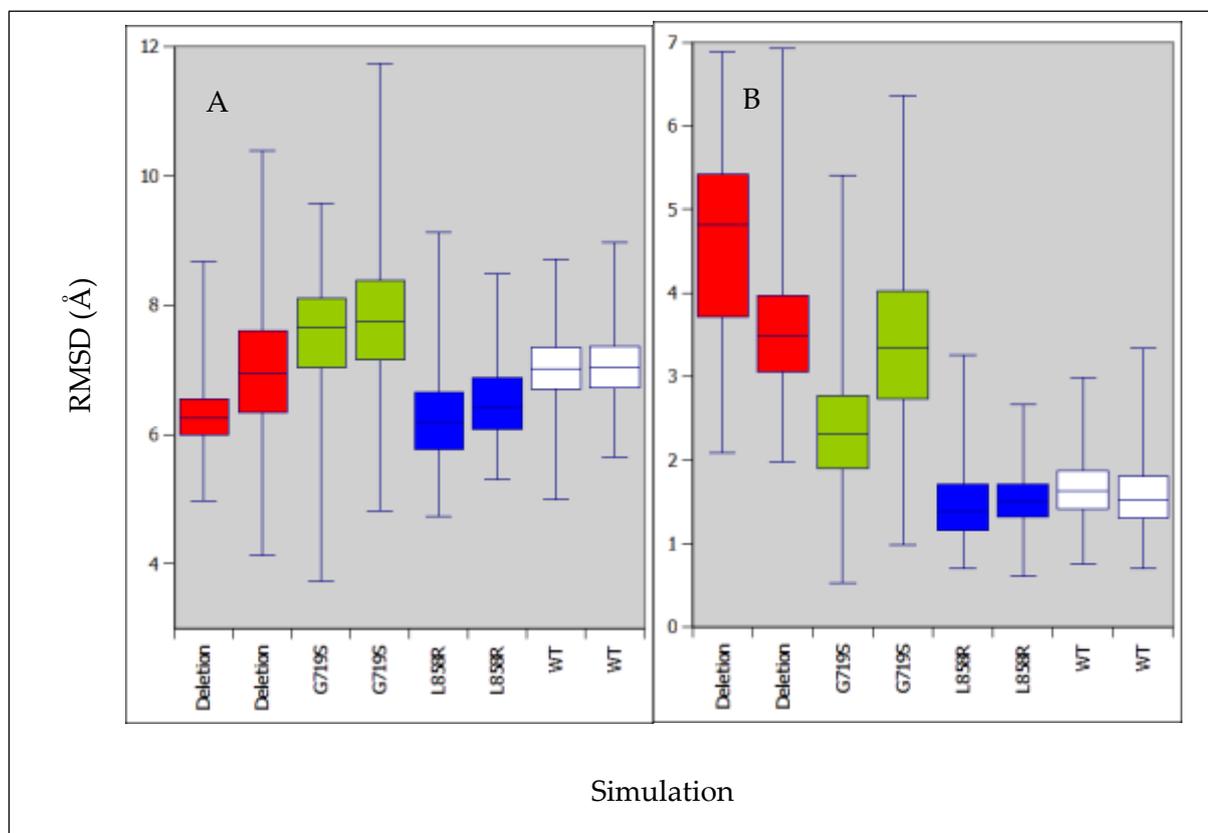


Figure 5.17: Box and whisker plot showing the RMSD of the C-helix backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the inactive AMD simulations.

As with the AMD simulations (figure 5.17), the DMDMD simulations (figure 5.18) do not exhibit the RMSD trends of the inactive simulations with respect to the active crystal structure; however, there appears to be little agreement between DMDMD and cMD with respect to the inactive crystal structure either (compare figure 5.18(B) with figure 5.15), although it could be argued that, on average, the point mutations have a lower RMSD than the deletion, this is not the case for all the simulations.

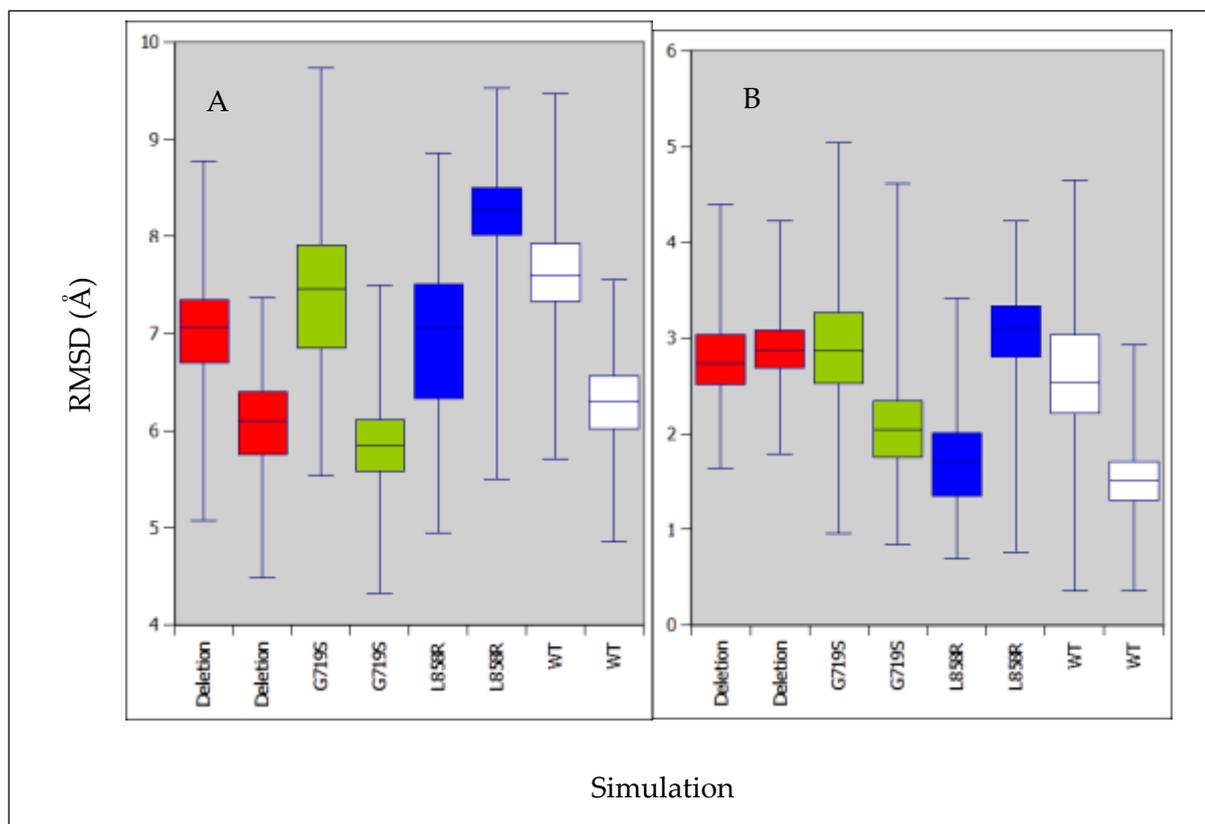


Figure 5.18: Box and whisker plot showing the RMSD of the C-helix backbone  $C\alpha$  atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the inactive DMDMD simulations.

The RDFMD results appear in agreement with the cMD results, with the deletion having the lowest RMSD with respect to the active crystal structure, and the highest RMSD with respect to the inactive crystal structure.

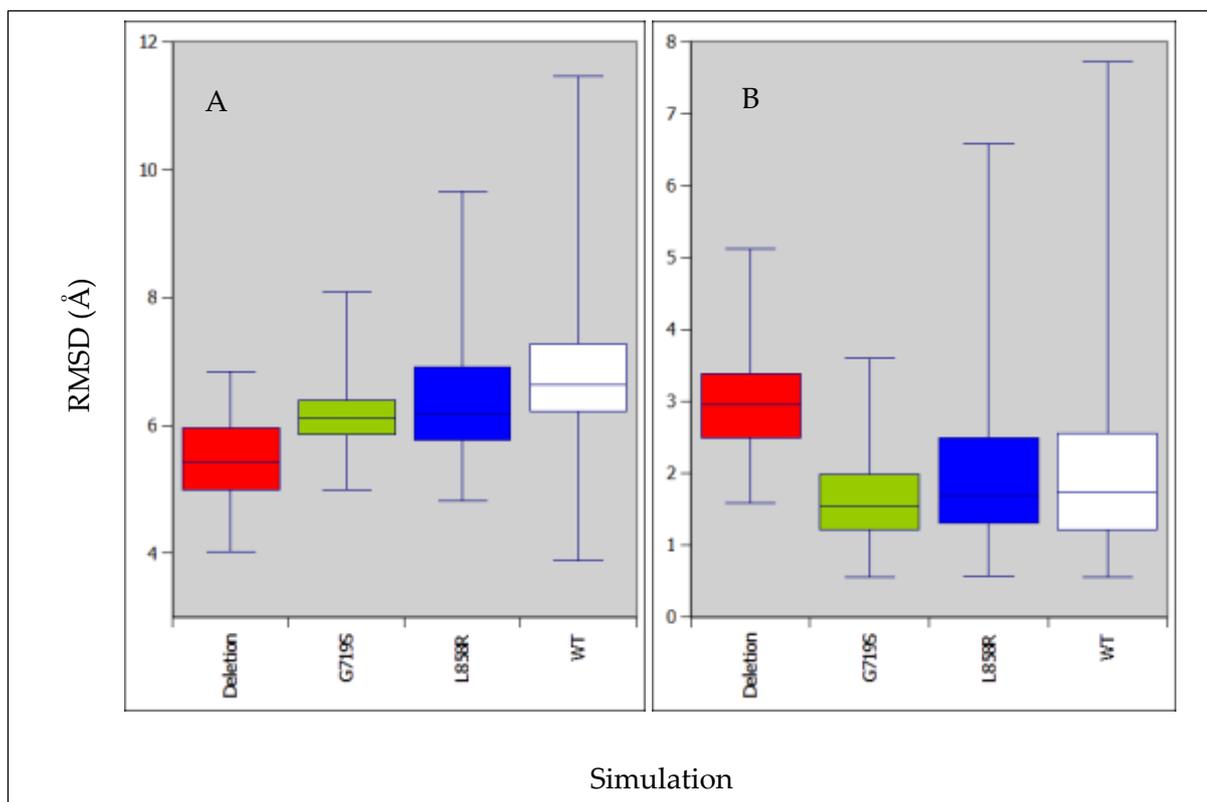


Figure 5.19: Box and whisker plot showing the RMSD of the C-helix backbone Ca atoms with respect to the active crystal structure (A; 1m17), and the inactive crystal structure (B; 2gs7) for the inactive RDFMD simulations. Note that each box in the diagram represents the combined data of 30 RDFMD simulations.

When all the sampling methods are taken together, it appears that the deletion promotes higher C-helix RMSDs with respect to the inactive crystal structures, in many cases concurrently reducing the RMSD of the C-helix with respect to the active crystal structure. For the AMD and DMDMD simulations, the G719S mutant also had a greater RMSD on average with respect to the inactive crystal structure; however, this observation was poorly conserved even between the AMD and DMDMD simulations.

Nonetheless, it appears fairly clear that the deletion has some impact on the C-helix, and is pulling the C-helix into a more active conformation in the majority of inactive simulations.

### 5.3.5 RMSD summary

In summary, the RMSD analysis, albeit a simple approach, successfully elucidates a number of interesting features of the sampling of EGFR. Firstly, with regards to the A-loop, the point mutants simulated from the active conformation sample configurations much further from the active crystal structure than is typical for the WT or deletion, a trend that is sometimes accompanied (seemingly dependent upon which sampling method is employed, although this could be due to the lack of repeats) with the sampling of configurations closer to the inactive crystal structure. Secondly, the active point mutants exhibit higher RMSDs for the C-helix with respect to the active crystal structure, once again, possibly sampling closer to the inactive crystal structure. Finally, the inactive deletion appears to result in elevated RMSDs with respect to the inactive structure relative to the other inactive simulations.

Overall, this hints at the point mutations *destabilising* the active conformation, a possibility that has been raised previously by Wan et al. (2011)[65] The other finding, which fits more easily into the orthodox, is that the deletion appears to destabilise the inactive conformation of the C-helix. These two points will gain further examination in the following sections.

Another interesting feature of the RMSD analyses is the ability to observe differences in sampling between the sampling methods. It is particularly interesting that, where a trend is identified, it is usually identifiable in the AMD and DMDMD simulations, with the AMD usually being the most distinct. In this respect, RDFMD performed relatively poorly, and in one instance produced very contradictory results to the other methods (compare figure 14 to figures 10, 11, 13, and 13).

It is unclear to what extent the inconsistency and insensitivity of the RDFMD results is due to the short time scales employed, the sampling method itself or the handling of data. Further discussion on this will appear through the more advanced analyses of later sections.

### 5.3.6 RMSF results

RMSFs were calculated for the backbone  $\alpha$  carbons over the course of the entire trajectory for all simulations. To investigate the impact of mutations on the RMSF profile of EGFR kinase, the WT RMSF profile and the difference between each mutant and the WT RMSF profile was calculated. Additionally, to quantify the significance of the calculated difference in RMSF values between the WT and each mutant, an unpaired t-test was performed.

## 5.3.7 Active RMSFs

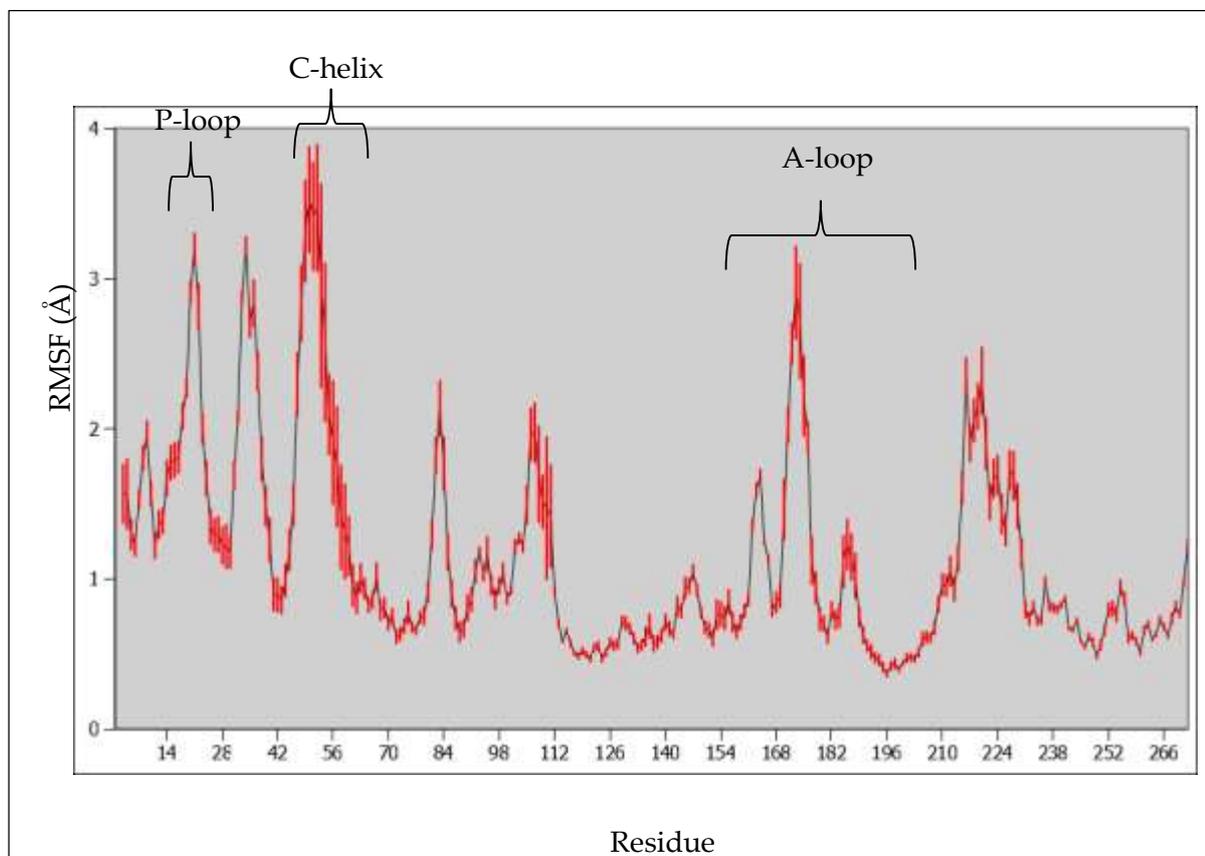


Figure 5.20: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT active cMD simulations, with the standard error shown as red error bars.

For the cMD simulations, the active WT RMSF profile is shown in figure 5.20.

Compared to the rest of the protein, the C-helix, P-loop and A-loop (also the loop downstream of the P-loop and upstream of the C-helix) have very high RMSF values, with the error bars greatest for the C-helix and A-loop, suggesting these structures differ more substantially between repeats.

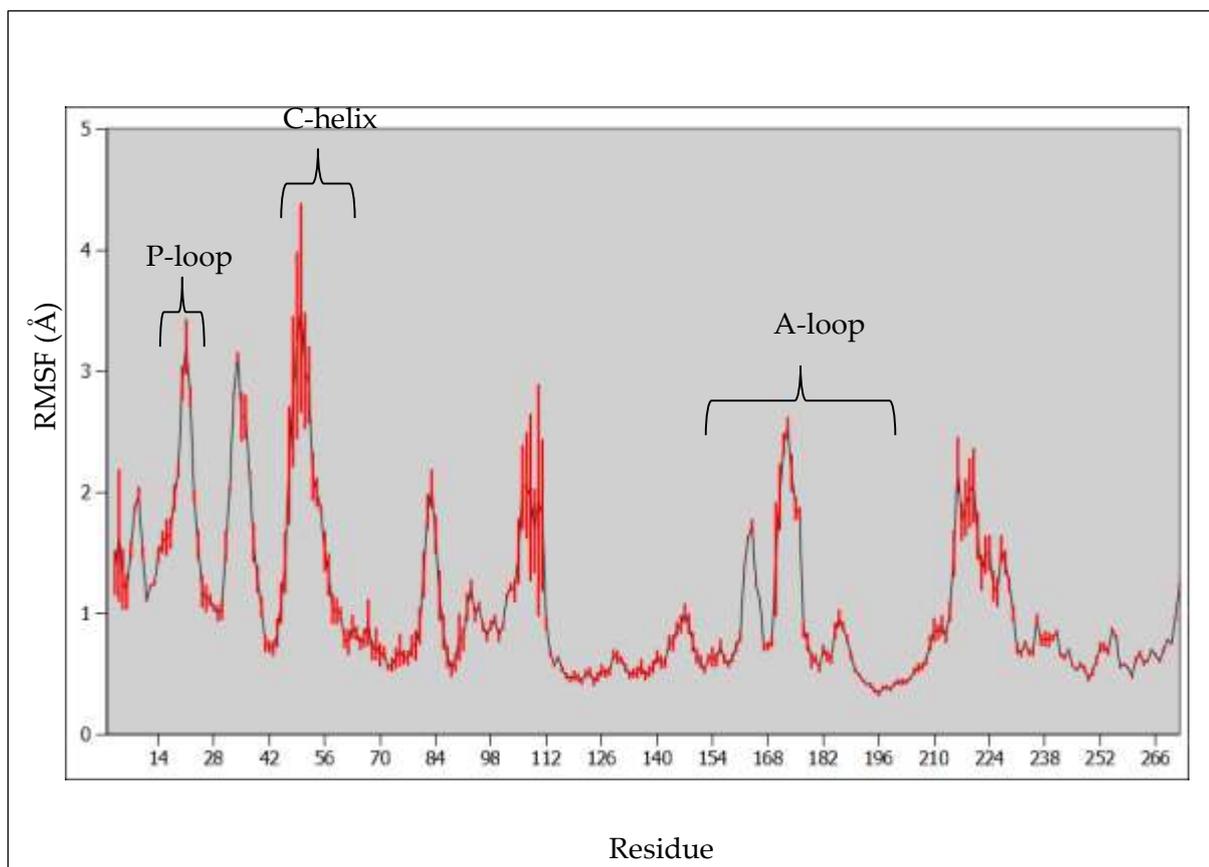


Figure 5.21: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT active AMD simulations, with the standard error shown as red error bars.

The AMD simulations show a similar RMSF profile (figure 5.21), but with much greater error in the region of the C-helix; this could be due to greater variation between repeats; however, there were only 2 AMD simulations for starting configuration of each system, compared to 4 for the cMD, and so the increase in error likely derives greatly from the lack of repeats.

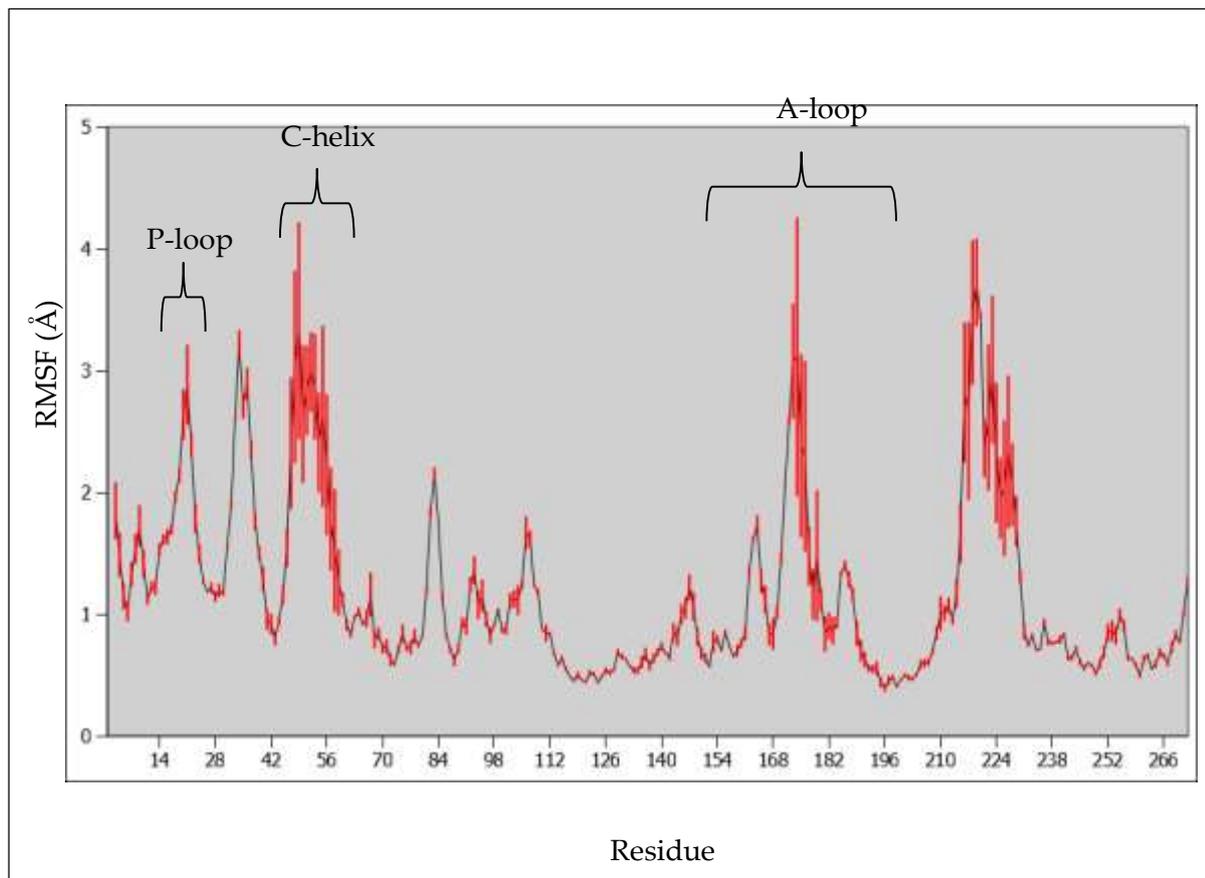


Figure 5.22: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT active DMDMD simulations, with the standard error shown as red error bars.

The DMDMD simulations of the active WT (figure 5.22) produce a similar profile as before, with the exception of a greater increase in the magnitude of the errors for the C-helix and A-loop. Since this increase in the errors not only occurs with respect to the cMD simulations, but the AMD simulations (which also had only 2 simulations per starting conformation), it indicates a greater degree of variation between the repeats, in comparison to the AMD simulations.

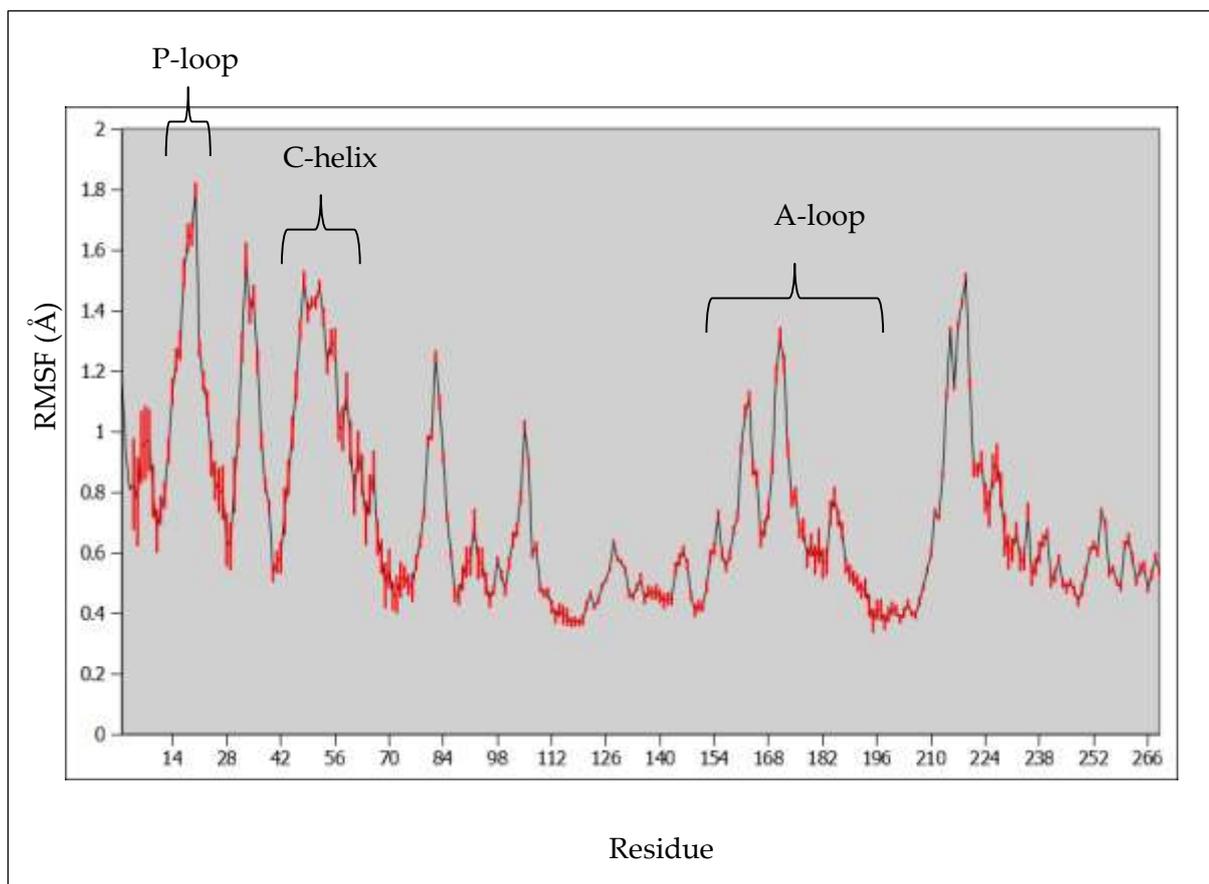


Figure 5.23: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT active RDFMD simulations, with the standard error shown as red error bars.

The RDFMD simulations show a slightly different pattern (figure 5.23) than before, with the peaks having generally a much lower magnitude than previously, presumably due to the short time scales of the RDFMD simulations, as highlighted in the discussion on the RDFMD RMSDs. There is also a reduction in standard error due to the large number of repeats employed for the RDFMD simulations.

Overall, the sampling methods all produce similar RMSF profiles, qualitatively, it appears that the sampling methods are all sampling similarly; however, from the RMSD results it is already apparent this is not entirely the case, and so the results should be considered with caution. Interestingly the RDFMD results, which were the

most inconsistent in terms of RMSDs with respect to the other sampling methods, also appear the least consistent in this RMSF analysis. However, as discussed above, this may be simply due to the short time-scales employed in this instance.

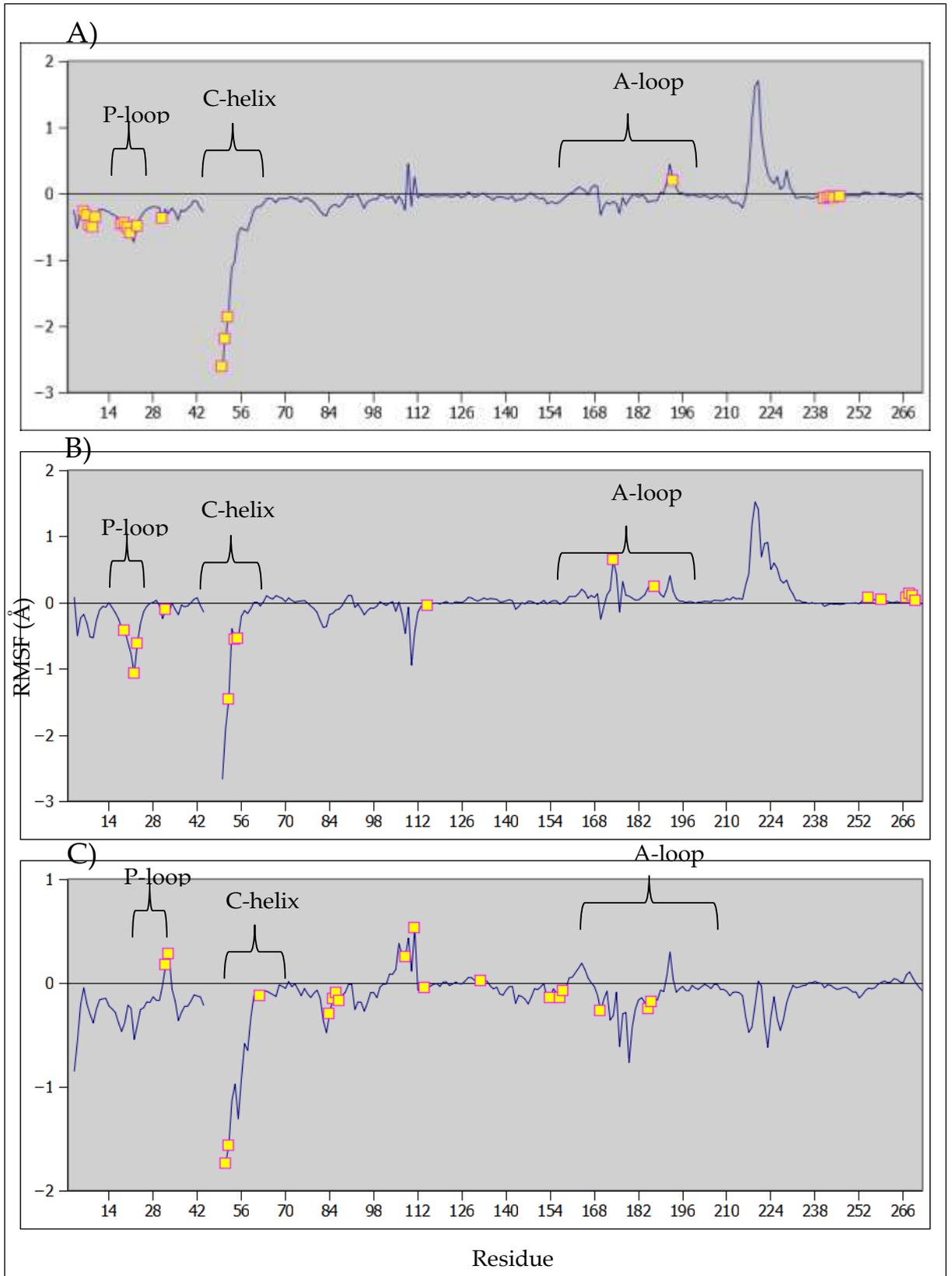


Figure 5.24: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the deletion and WT active cMD (A), AMD (B), and DMDMD (C) simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

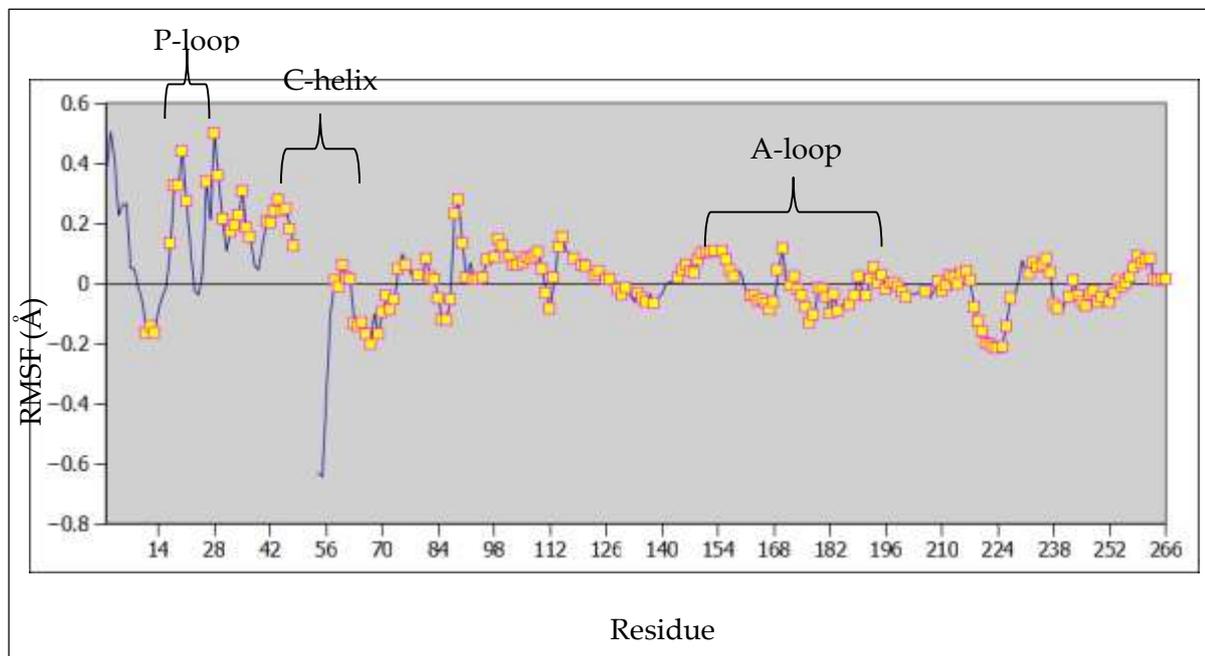


Figure 5.25: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the deletion and WT RDFMD simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

The difference between the RMSF of the deletion mutant and the RMSF of the WT deletion mutant is shown for each sampling method in figures 5.24 and 5.25. Very little of the profile is conserved between sampling methods, with the only real exception being the region of the C-helix just following the deletion, which is shown to be considerably less mobile for all sampling methods (though not statistically significantly for the RDFMD simulations). This observation correlates well with the previous observation of the deletion reducing the range of RMSDs accessible to the deletion mutant's C-helix, relative to the WT.

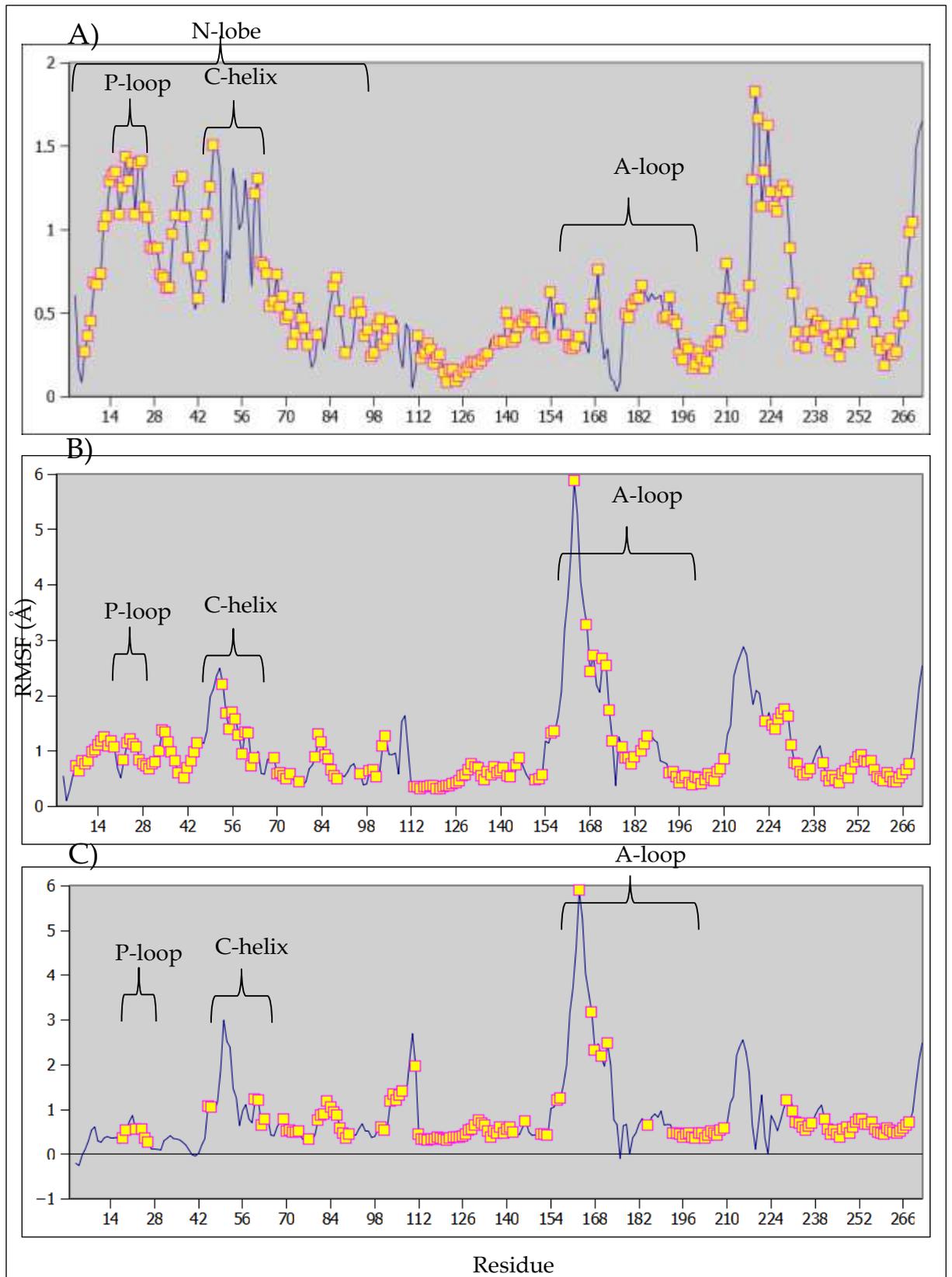


Figure 5.26: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the L858R and WT active cMD (A), AMD (B), and DMDMD (C) simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

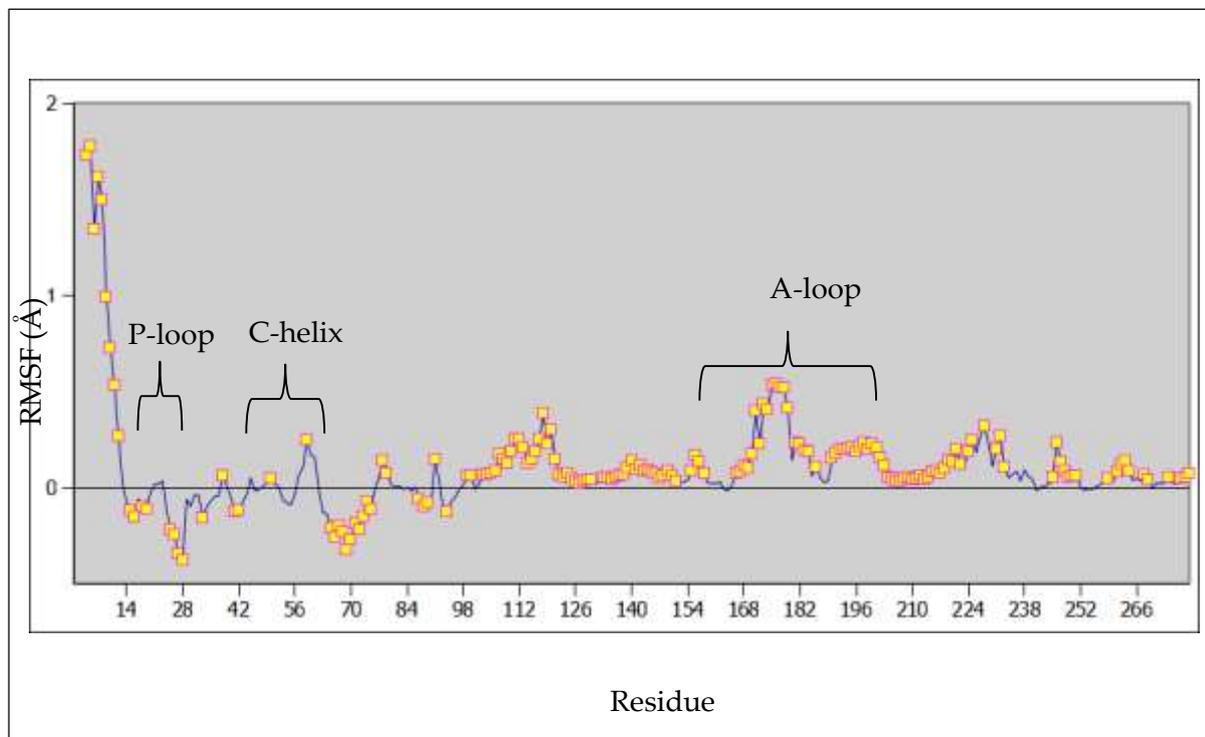


Figure 5.27: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the L858R and WT RDFMD simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

The active conformation of the L858R mutant has increased RMSFs for the C-helix across the cMD, AMD and DMDMD simulations and an increased A-loop RMSF for all the enhanced sampling methods (figure 5.26 and 5.27). Unlike the RMSF profile for the deletion, the profile for the L858R shows a much higher level of significance throughout. This observation correlates well with the RMSD results, which also showed the L858R to increase C-helix and A-loop RMSDs in a number of simulations.

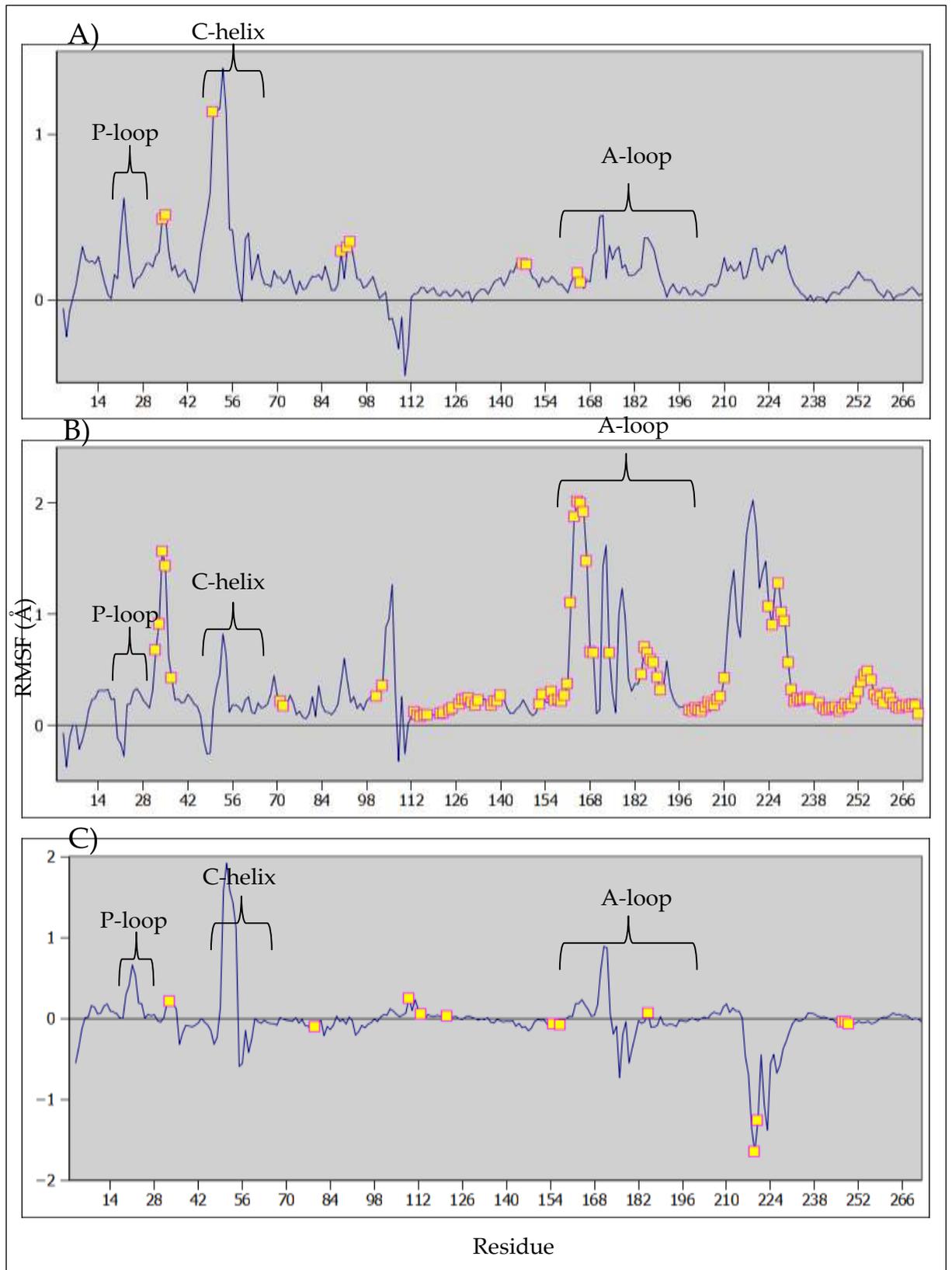


Figure 5.28: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the G719S and WT active cMD (A), AMD (B), and DMDMD (C) simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

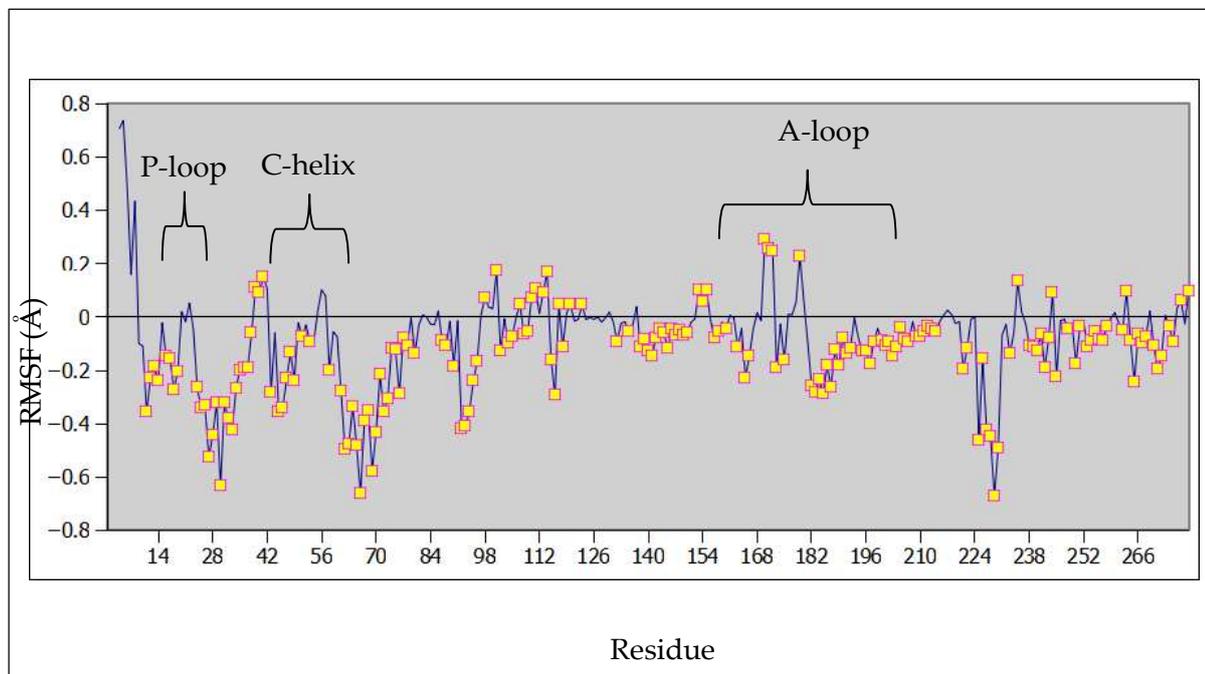


Figure 5.29: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the G719S and WT RDFMD simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

For the G719S, there is very little in the way of significant differences from the WT that are conserved through the different sampling methods (figure 5.28 and 5.29). There is a tendency for a peak to exist at the A-loop and C-helix; however, these peaks are not always statistically significant, particularly in the case of the C-helix. Nonetheless, this would correspond with the RMSD results which seem to suggest a similar role for G719S to the L858R mutant, and that the impact of the G719S was not as great as the L858R.

Taken together, the RMSF profiles of the active conformation mutants in comparison to the WT appear to show a similar picture to the RMSD analysis: the deletion appears to

reduce the mobility of the C-helix, while the point mutants appear to increase it, as well as increasing A-loop fluctuations.

### 5.3.8 Inactive RMSFs

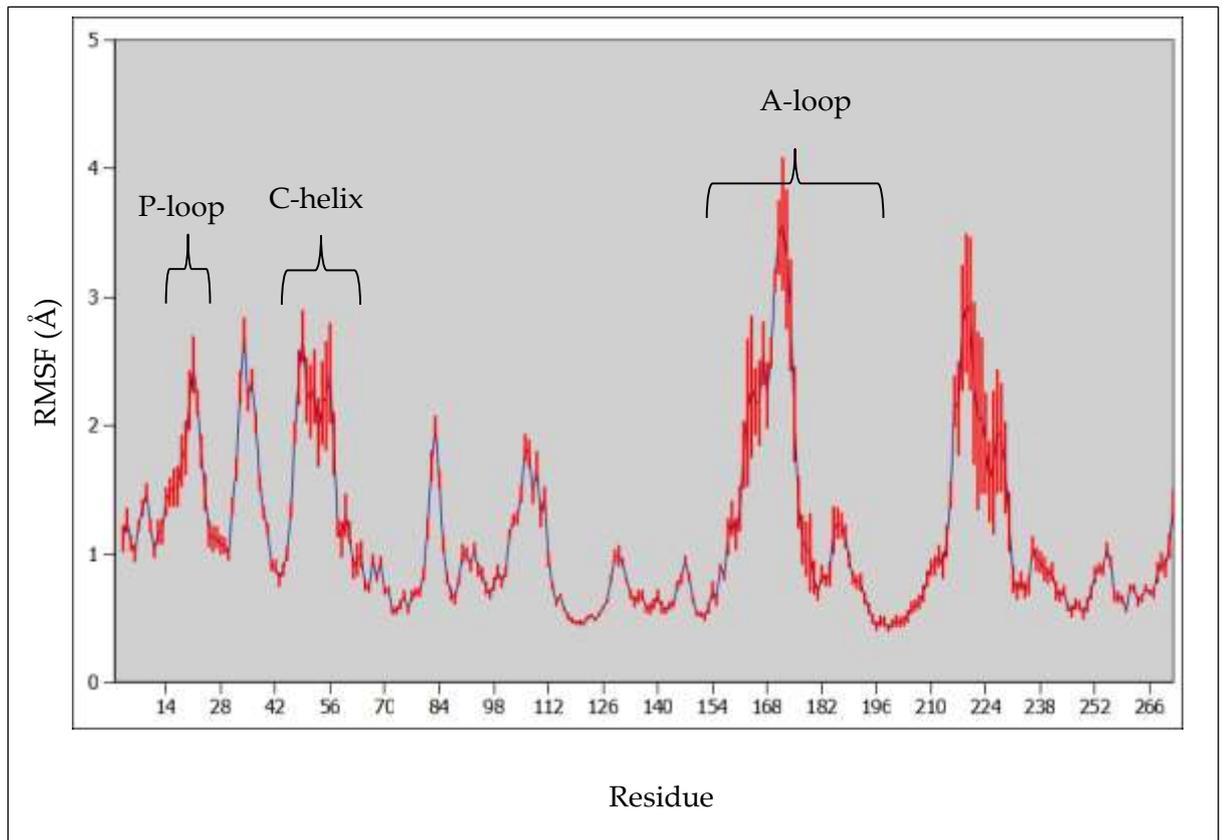


Figure 5.30: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT inactive cMD simulations, with the standard error shown as red error bars.

As per the active simulations (figure 5.20), the inactive simulations (figure 5.30) have prominent peaks for the P-loop, C-helix and A-loop structures.

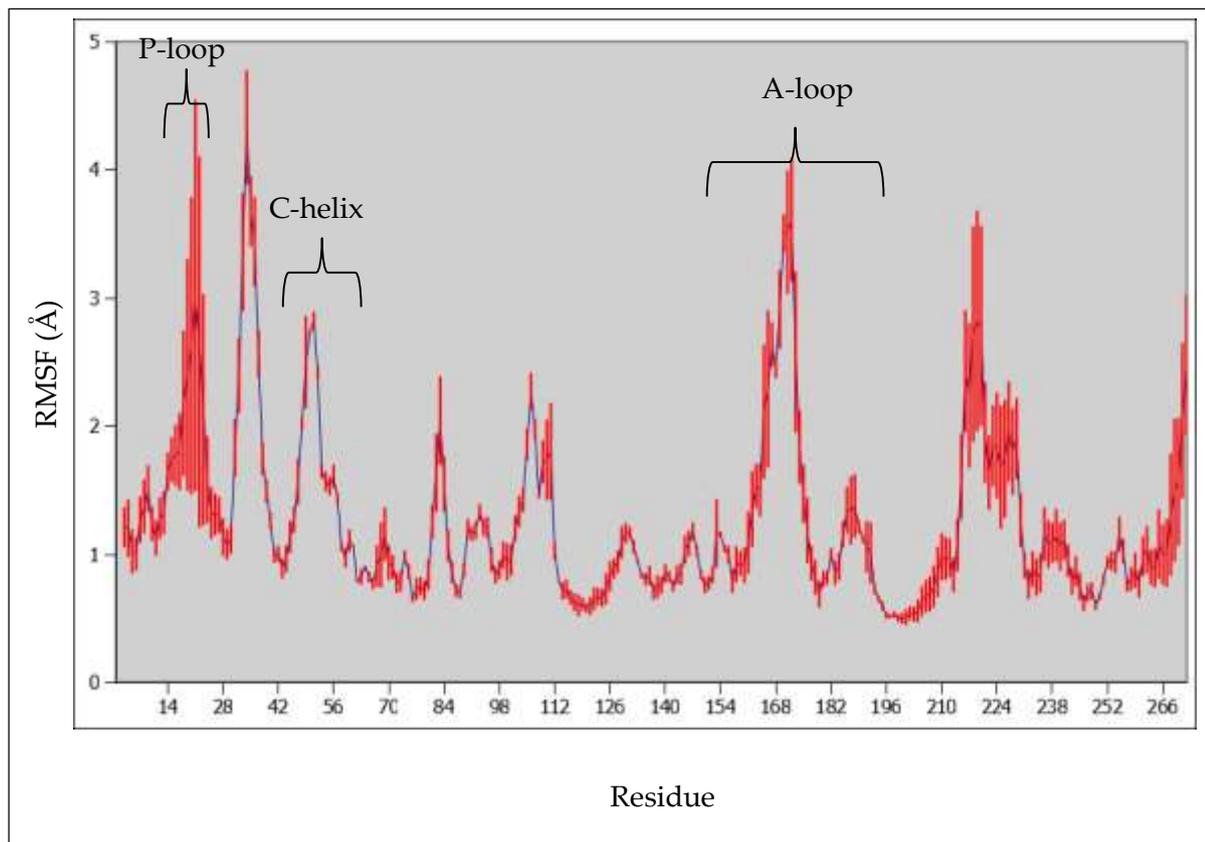


Figure 5.31: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT inactive AMD simulations, with the standard error shown as red error bars.

The RMSF profile for the WT in the AMD simulations (figure 5.31) has a similar arrangement of peaks, particularly for the A-loop, C-helix and P-loop; however the RMSF of the P-loop appears to vary considerably between the AMD repeats. Additionally, the peaks are somewhat higher compared to the cMD simulations, possibly indicating that more conformational space has been sampled.

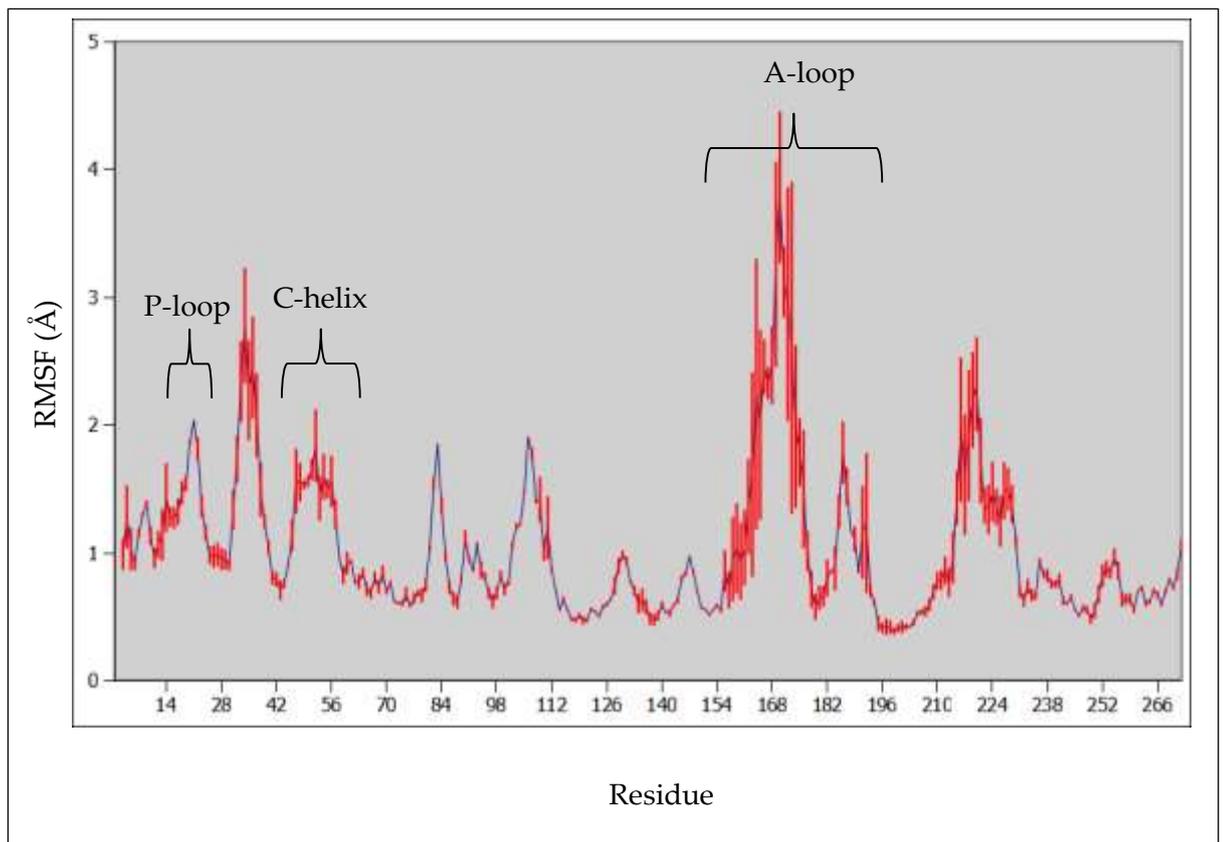


Figure 5.32: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT inactive DMDMD simulations, with the standard error shown as red error bars.

The DMDMD simulations (figure 5.32) show a slightly different WT RMSF profile to the other sampling methods, with a higher peak for the A-loop, and lower peaks for the P-loop and C-helix.

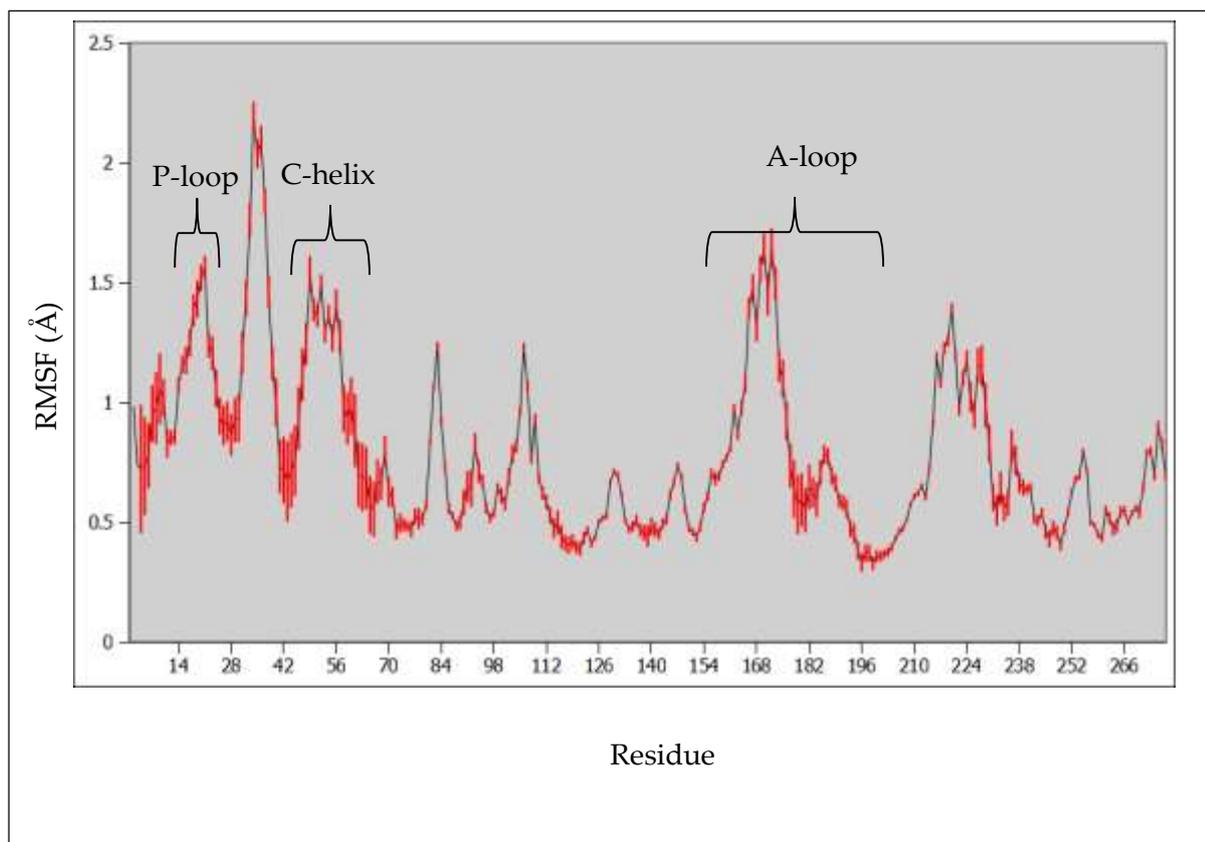


Figure 5.33: Average RMSF of the  $\alpha$ -carbons of the backbone for the WT inactive RDFMD simulations, with the standard error shown as red error bars.

The RDFMD simulations (figure 5.33) show a similar inactive WT RMSF profile to the cMD simulations, with the exception of the loop downstream of the P-loop and upstream of the C-helix, which appears to have a higher RMSF. The magnitudes of the RMSFs for the RDFMD simulations appear to be much lower than the other sampling methods.

As with the active simulations, all of the sampling methods appear to show peaks where expected for the A-loop, C-helix, and P-loop, and while the magnitude of the RMSF for these peaks is not always consistent between the methods, it seems that the sampling is being performed by the same regions for each sampling method; however, as discussed previously, the RMSDs show the sampling can be very different between

sampling methods, and so it is not necessarily the case that the simulations are sampling similar configuration space.

There is a consistent trend for the deletion to increase the RMSF of the A-loop with respect to the WT (figure 5.34), with this change being statistically significant for at least one point on the peak. There is also a significant increase in C-helix RMSFs for the AMD and DMDMD simulations, which appears to be accompanied by an RMSF increase across much of the N-lobe.

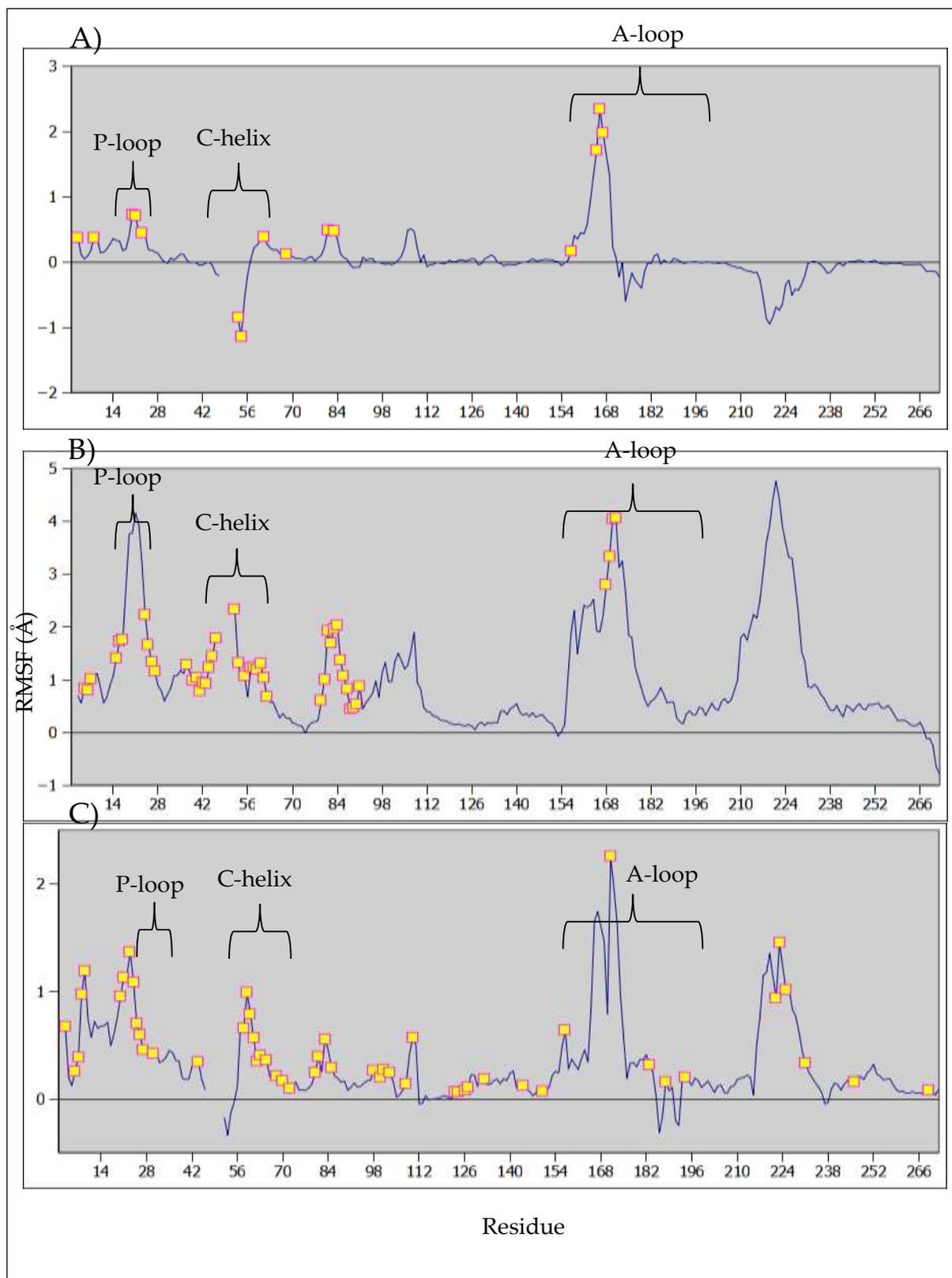


Figure 5.34: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the deletion and WT inactive cMD (A), AMD (B), and DMDMD (C) simulations.

Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

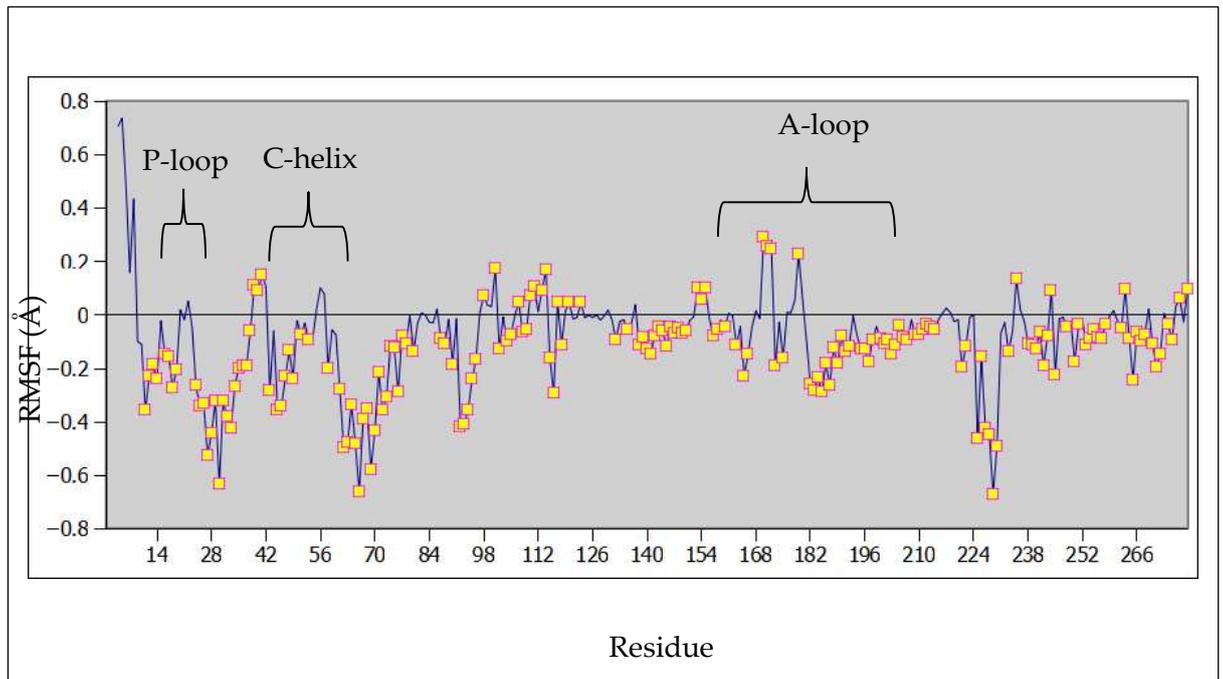


Figure 5.35: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the inactive deletion and WT RDFMD simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

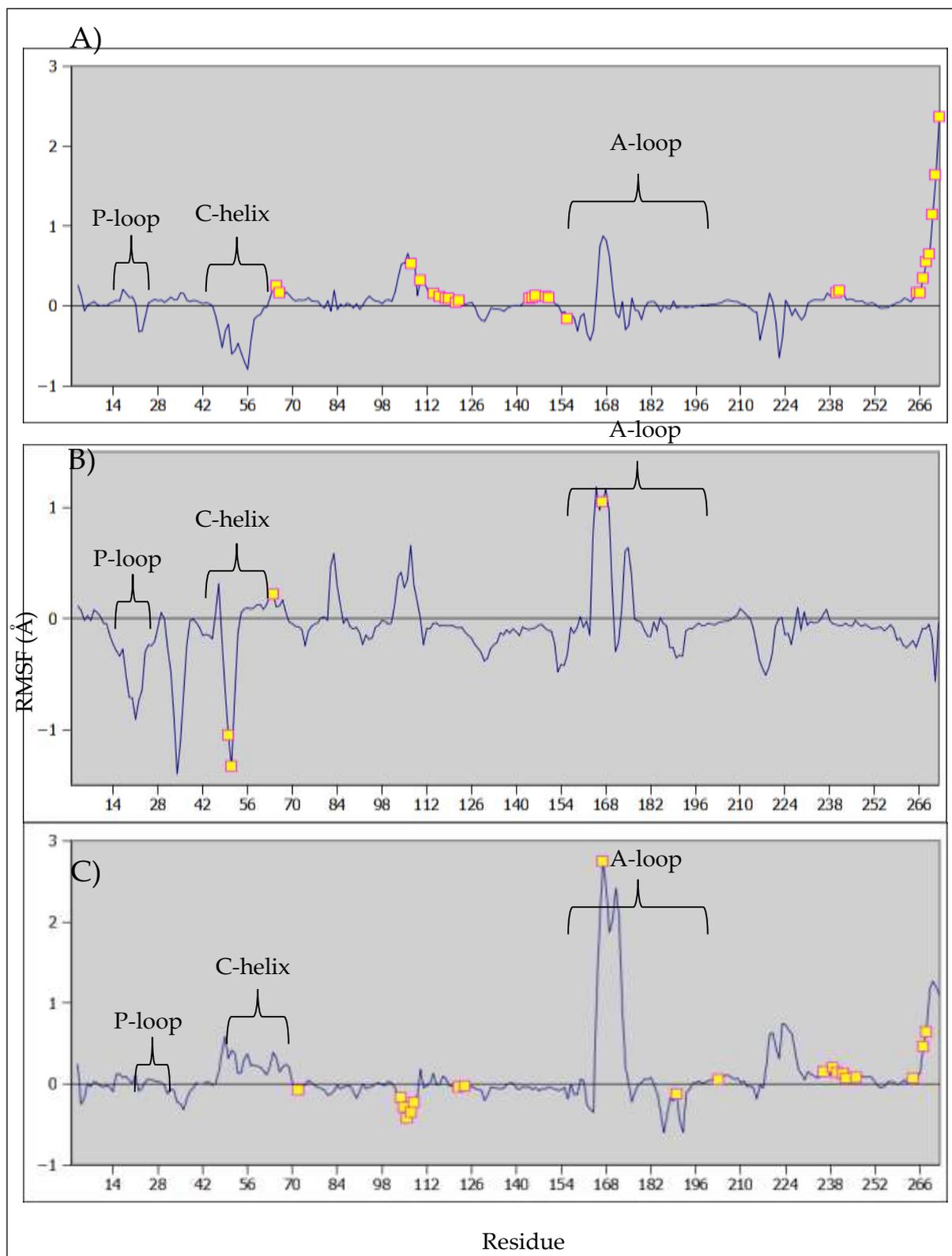


Figure 5.36: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the L858R and WT inactive cMD (A), AMD (B), and DMDMD (C) simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

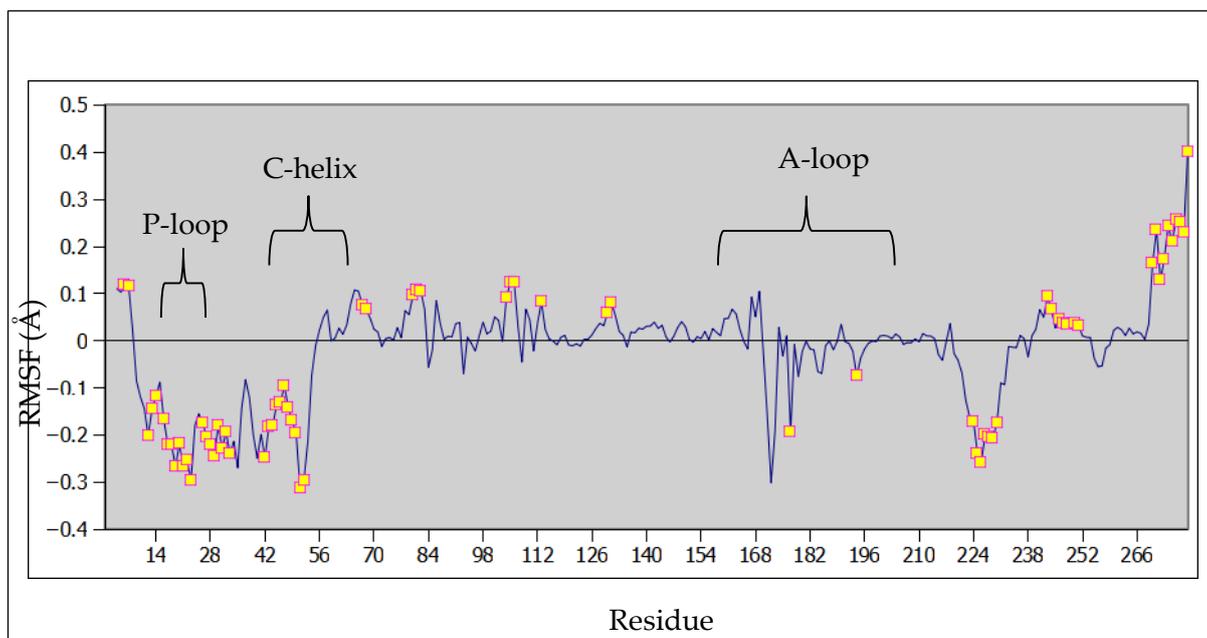


Figure 5.37: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the inactive L858R and WT RDFMD simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

With the exception of the DMDMD simulations, there appears to be a tendency for the L858R to reduce RMSF in the region of the C-helix (figures 5.36 and 5.37). Interestingly, most of the sampling methods show an RMSF peak at the beginning of the A-loop (where the L858R is situated); however this is only statistically significant in the AMD and DMDMD simulations. Additionally, the magnitude of the peaks is generally quite small, suggesting that for the most part, the L858R does not have a very significant impact upon RMSFs in comparison to the WT.

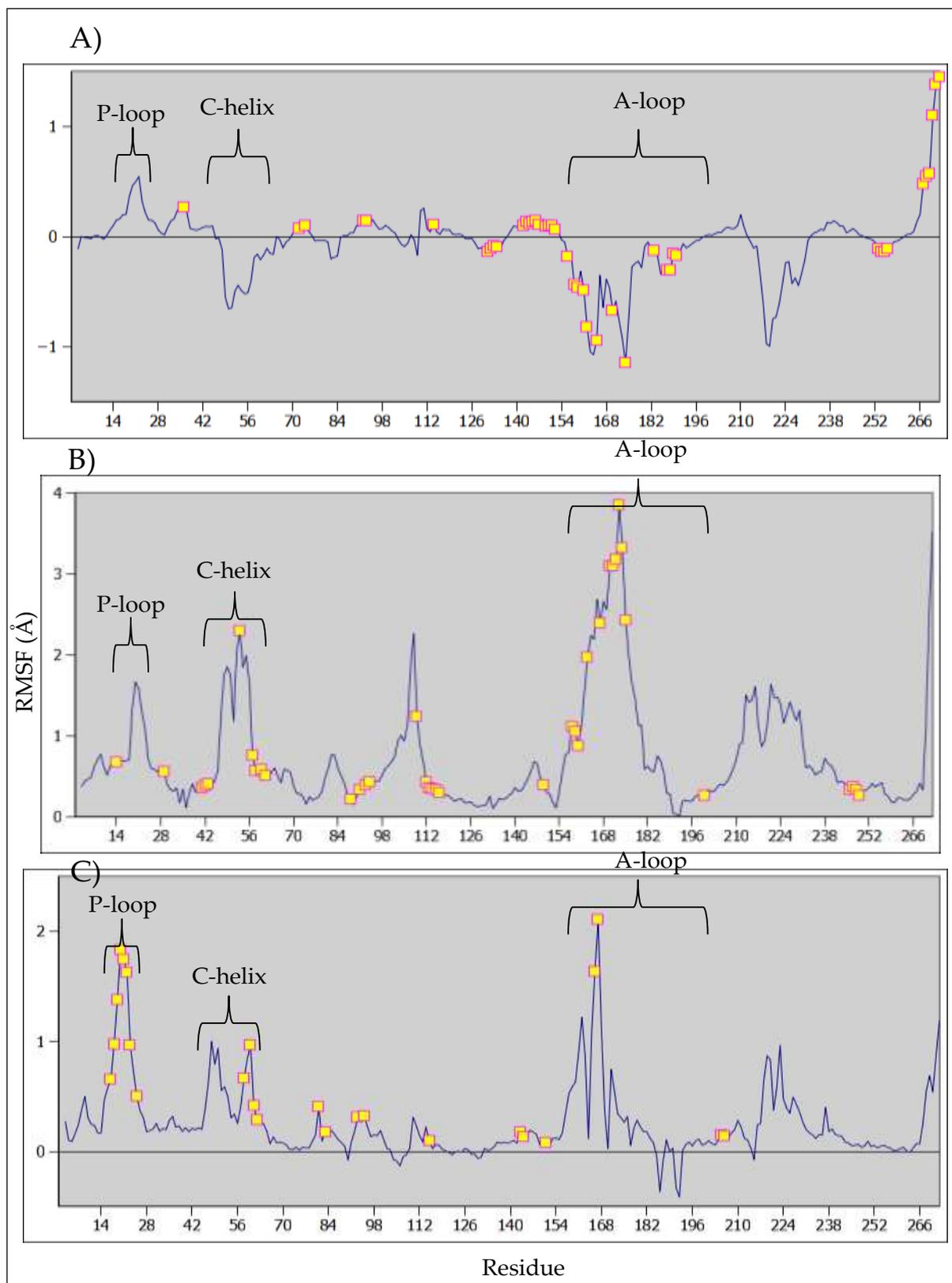


Figure 5.38: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the G719S and WT inactive cMD (A), AMD (B), and DMDMD (C) simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

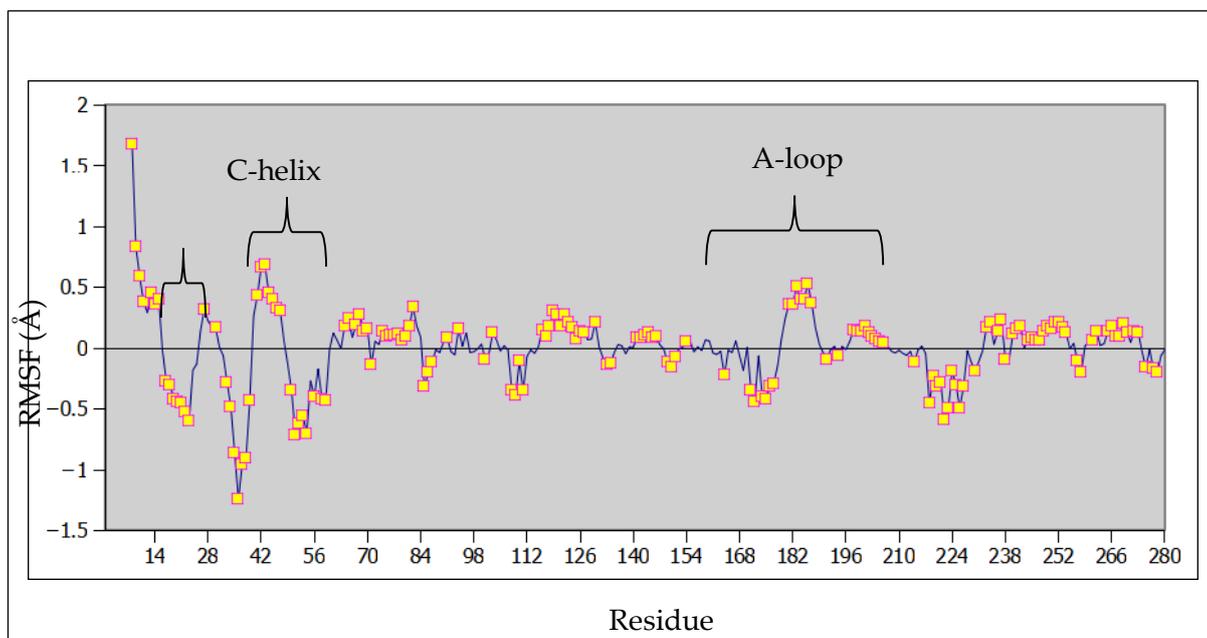


Figure 5.39: Difference in average RMSF of the  $\alpha$ -carbons of the backbone between the inactive G719S and WT RDFMD simulations. Statistically significant differences ( $p < 0.05$ ) are highlighted in yellow.

The G719S (figures 5.38 and 5.39) has a statistically significant increase to C-helix RMSD in comparison to the WT for all sampling methods except cMD (which exhibited a reduced RMSF, and was not statistically significant). There is also a tendency for an increase to the P-loop RMSF; however this was only statistically significant in the DMDMD simulations, and the opposite was observed for the RDFMD simulations. A statistically significant increase was also observed for the A-loop in the enhanced sampling methods; however, a significant reduction was seen in the cMD simulations, and the RDFMD simulations show a significant reduction in RMSFs for the first half of the A-loop (with the latter residues being significantly higher) with respect to the WT

The RMSF results for the inactive simulations show a picture complementary to the RMSD results: with the deletion increasing C-helix RMSFs in the AMD and DMDMD simulations. Similar results for the G719S helps corroborate the finding that some G719S simulations appeared to exhibit higher C-helix RMSDs (see section 5.3). The

relatively low L858R RMSFs also corroborate with the tendency for lower C-helix RMSDs for that mutant.

### 5.3.9 RMSF summary

Overall, the RMSF results corroborate well with the RMSD analysis, and help to highlight how the mutants impact the sampling of EGFR kinase. As with the RMSD analysis, some differences become apparent when comparing between sampling methods; however exploration into how the sampling differs between sampling methods will require more advanced analysis, as discussed in the later sections of this chapter.

## 5.4 Secondary structure analysis

The secondary structure of proteins is critical to their function; indeed discussion of EGFR in the present study has been made mostly in terms of the sampling of the A-loop and C-helix. An understanding of how these structures evolve over time has been shown to be important in a number of studies [5], [43], [62], and here the secondary structure assignment algorithm STRIDE[96] has been applied to the cMD, AMD, and DMDMD trajectories. As with the RMSDs, the RDFMD simulations have been excluded for expediency.

Each trajectory was used as input for version 1.9.1 of the VMD program[125], which utilises the STRIDE algorithm in the secondary structure prediction function of the *timeline* plugin (a default plugin included with VMD version 1.9.1).

Secondary structure analysis of EGFR kinase is only performed in one report of simulations on EGFR kinase in the literature. The report, by Shan et al. (2012)[43], limited this analysis to the C-helix, which they predicted to lapse into disorder in the active WT, and be stabilised by the activating mutants. In the present study, a secondary structure analysis was performed on the A-loop and C-helix of both the active and inactive conformations.

#### **5.4.1 Active conformation structure**

The results of the secondary structure analysis on the active C-helix of the cMD simulations is shown in figure 5.40. There are significant fluctuations in the L858R simulations, such that differentiation of the WT from the L858R is very difficult. Only the C-helix of the deletion appears to be significantly stabilised compared to the WT. A similar picture emerges from the AMD simulations (see figure 5.41); however, the DMDMD simulations appear to be in closer agreement with Shan et al. (2012)[43] in the degree of disorder (see figure 5.42).

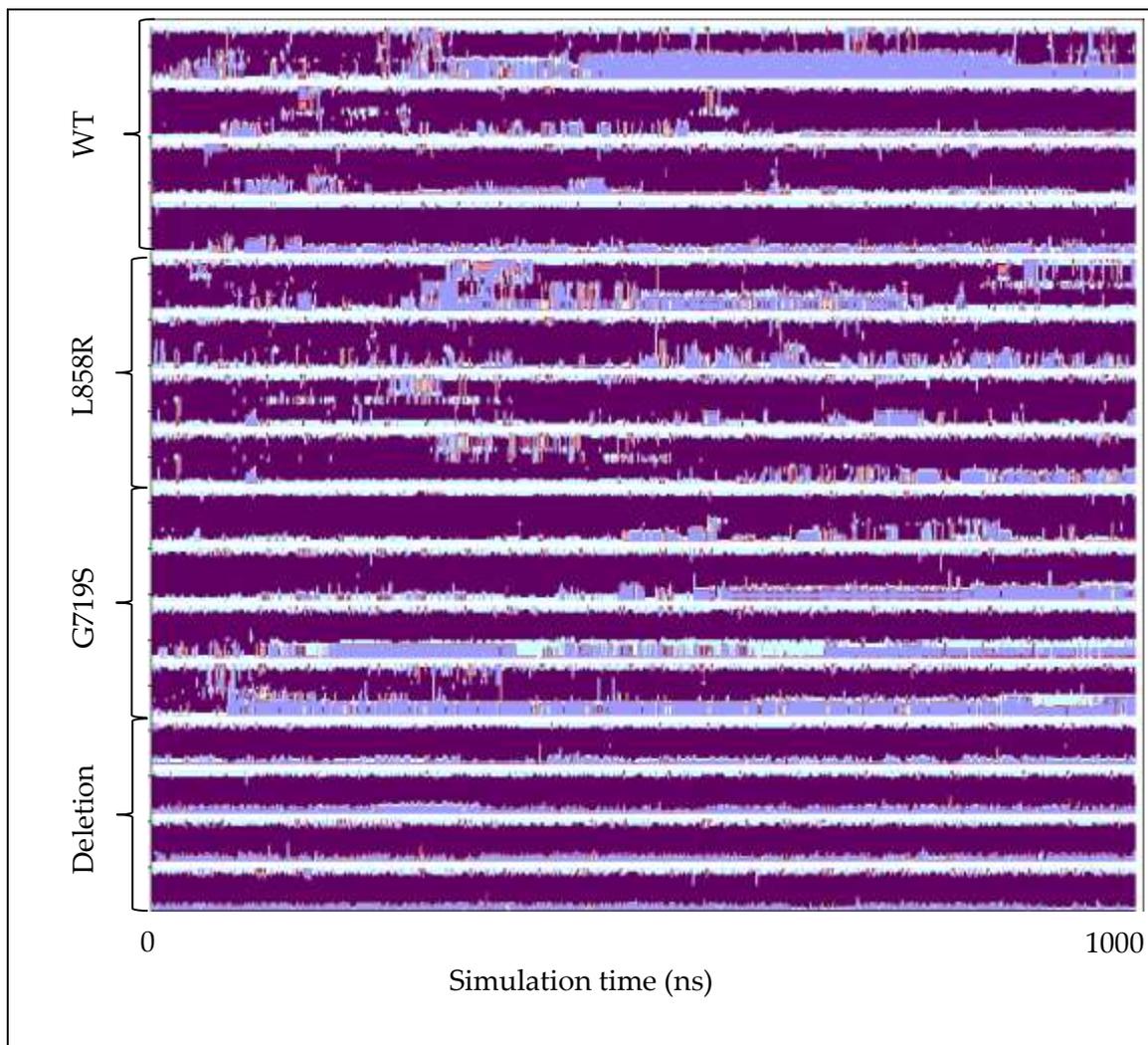


Figure 5.40: STRIDE secondary structure analysis of the C-helix over the course of the active cMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

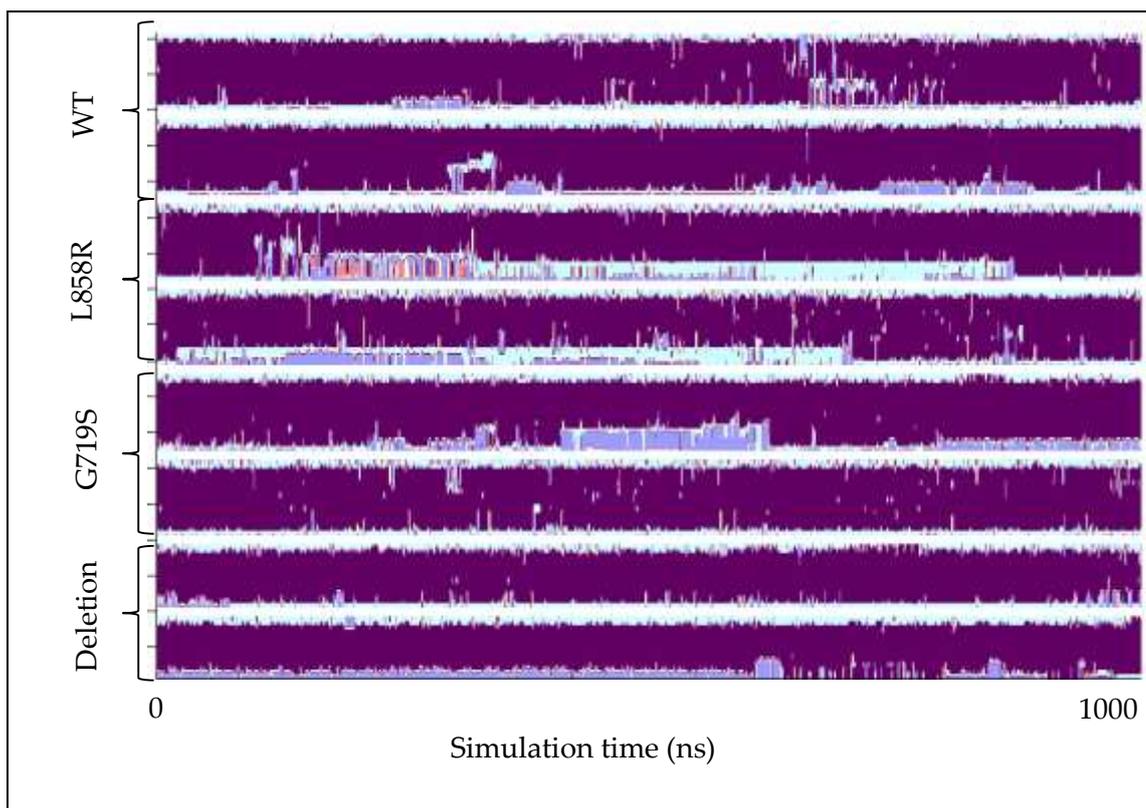


Figure 5.41: STRIDE secondary structure analysis of the C-helix over the course of the active AMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

In the active DMDMD simulations there appears to be a considerable level of disorder in the WT C-helix, with the mutants exhibiting more stability. It may be that the increased propensity for DMDMD to sample conformations away from local minima leads to an increased sampling of disordered conformations, and it is encouraging that the WT appears to have the most readily perturbed C-helix; however, the DMDMD results (see figure 5.42) only show significant disorder in one of the two simulations, and do not help explain the observations from the other sampling methods.

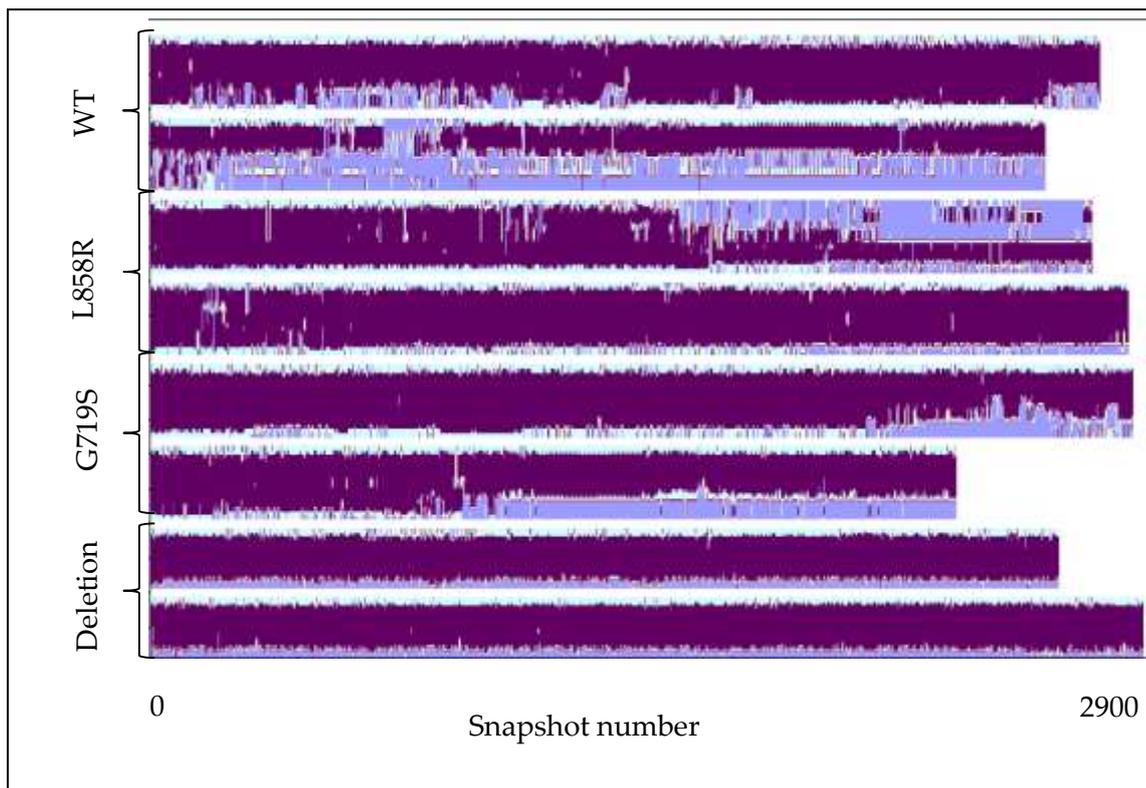


Figure 5.42: STRIDE secondary structure analysis of the C-helix over the course of the active DMDMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

Analysis of the A-loop over the course of the cMD (see figure 5.43) and DMDMD (see appendix 1) simulations found little change in the secondary structure over time; however, the AMD simulations show a significant change in the A-loop of the L858R mutant (see figure 5.44).

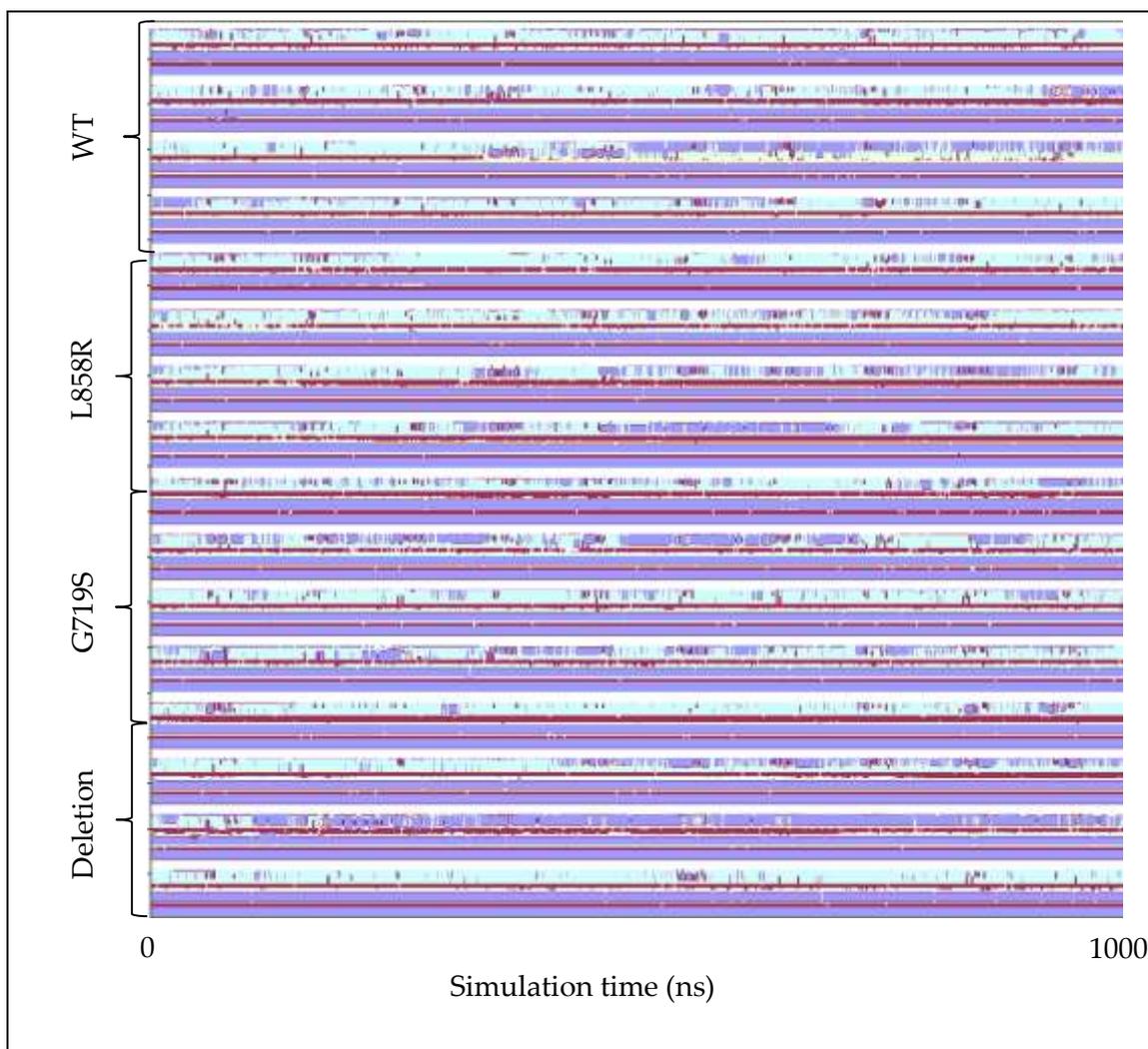


Figure 5.43: STRIDE secondary structure analysis of the A-loop over the course of the active cMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

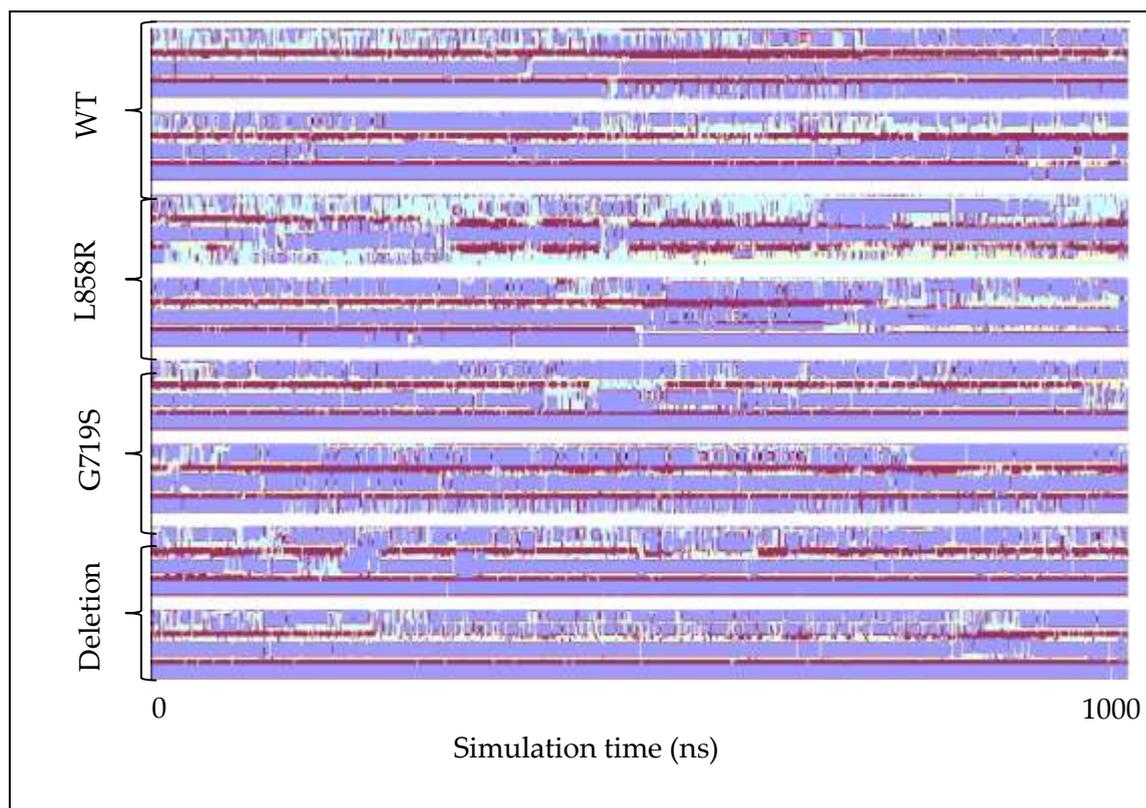


Figure 5.44: STRIDE secondary structure analysis of the A-loop over the course of the active AMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

#### 5.4.2 Inactive conformation structure

The secondary structure of the inactive C-helix is fairly noisy, however it appears that the L858R has a mild stabilising effect on the C-helix, and the deletion seems to have the opposite effect. In general the G719S has a similar C-helix structure to the WT, although one of the WT simulations exhibits considerable reorganisation of the C-helix (see figure 5.45). As might be expected from increasing the sampling using AMD, instabilities of the inactive A-loop helix occur more frequently. This is especially true for the deletion mutant and G719S, whereas the WT and L858R mutants appear as stable in this region as in the cMD (see figure 5.46), moreover, these observations are consistent with the DMDMD results (see figure 5.47).

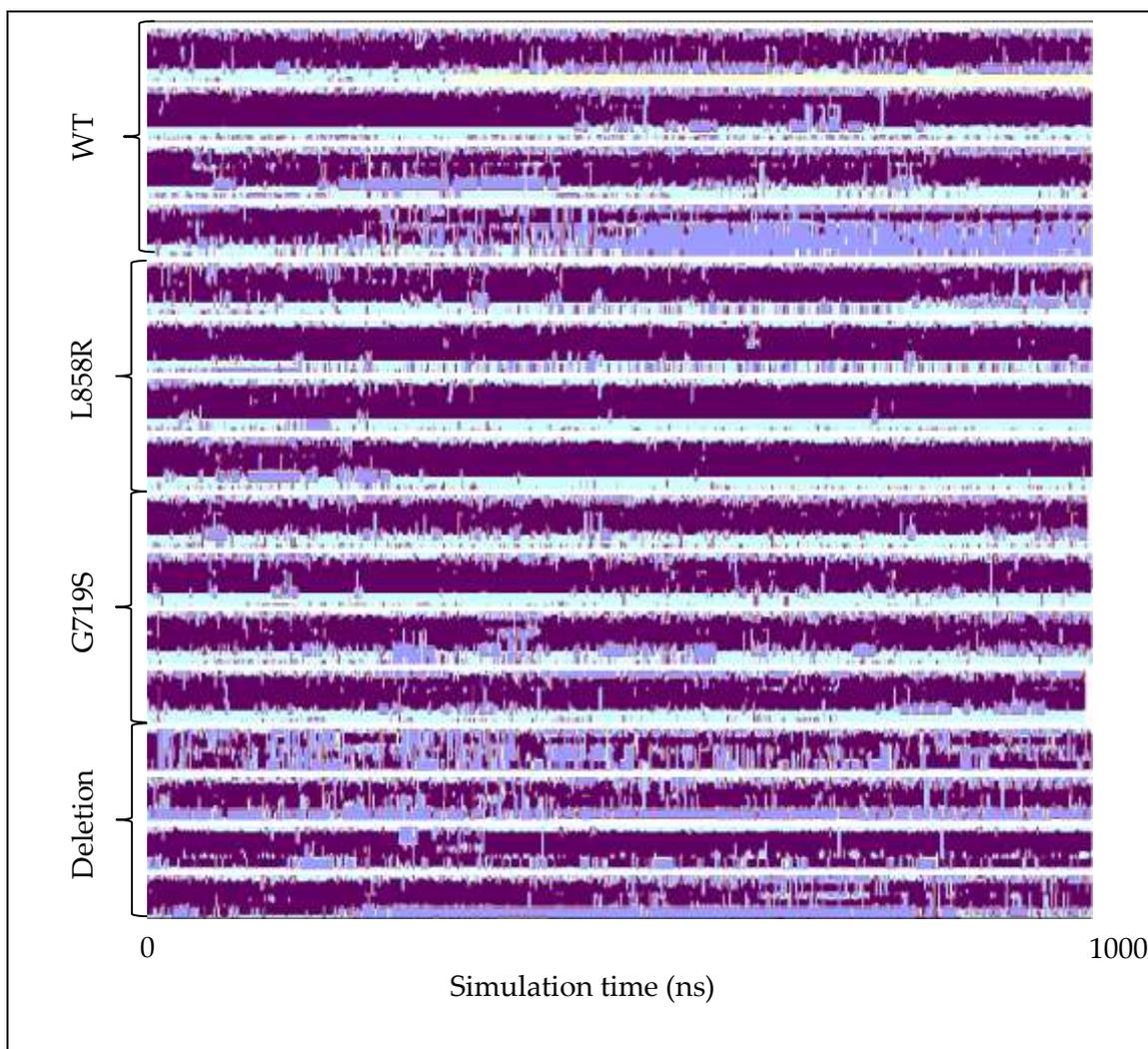


Figure 5.45: STRIDE secondary structure analysis of the C-helix over the course of the inactive cMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

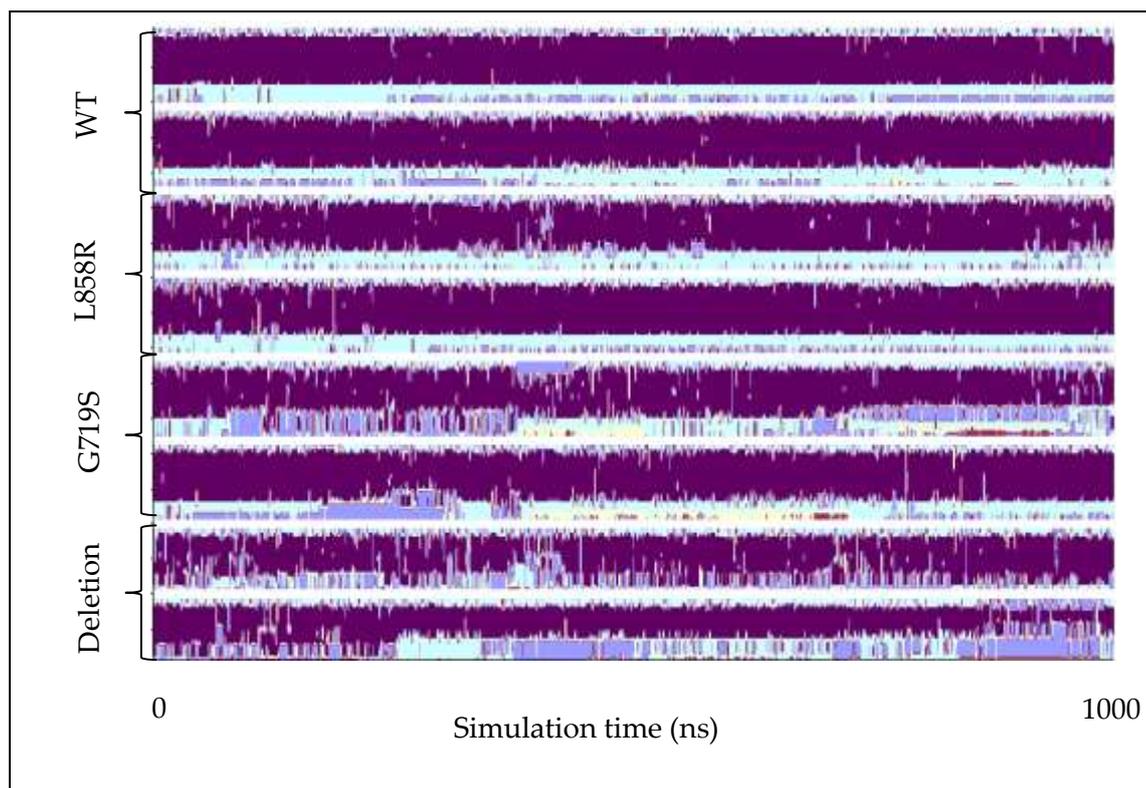


Figure 5.46: STRIDE secondary structure analysis of the C-helix over the course of the inactive AMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

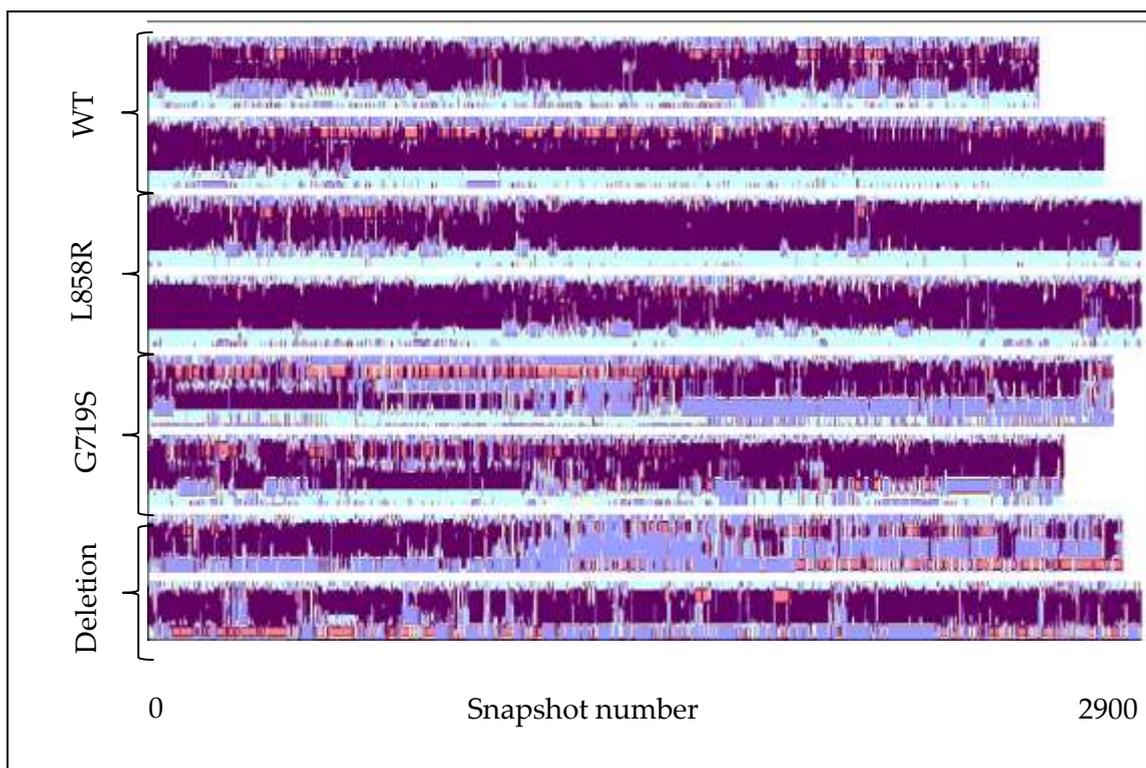


Figure 5.47: STRIDE secondary structure analysis of the C-helix over the course of the inactive DMDMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

To investigate the impact of the mutational status of EGFR on the A-loop helix, a STRIDE secondary structure analysis was performed on the region for all simulations. The results (see figure 5.48) show that all the deletion simulations have at least some destabilisation of the helix (though in the case of the cMD simulations this destabilisation only leads to unwinding in the above example). Interestingly, this destabilisation is not consistent throughout the mutant trajectories, suggesting that the mutants do not necessarily induce activation by destabilising the A-loop helix. Even more curious is the effect of the L858R to apparently incorporate more residues into the A-loop helix, and the fact that both the L858R and G719S appear to have a more stable A-loop helix than the WT.

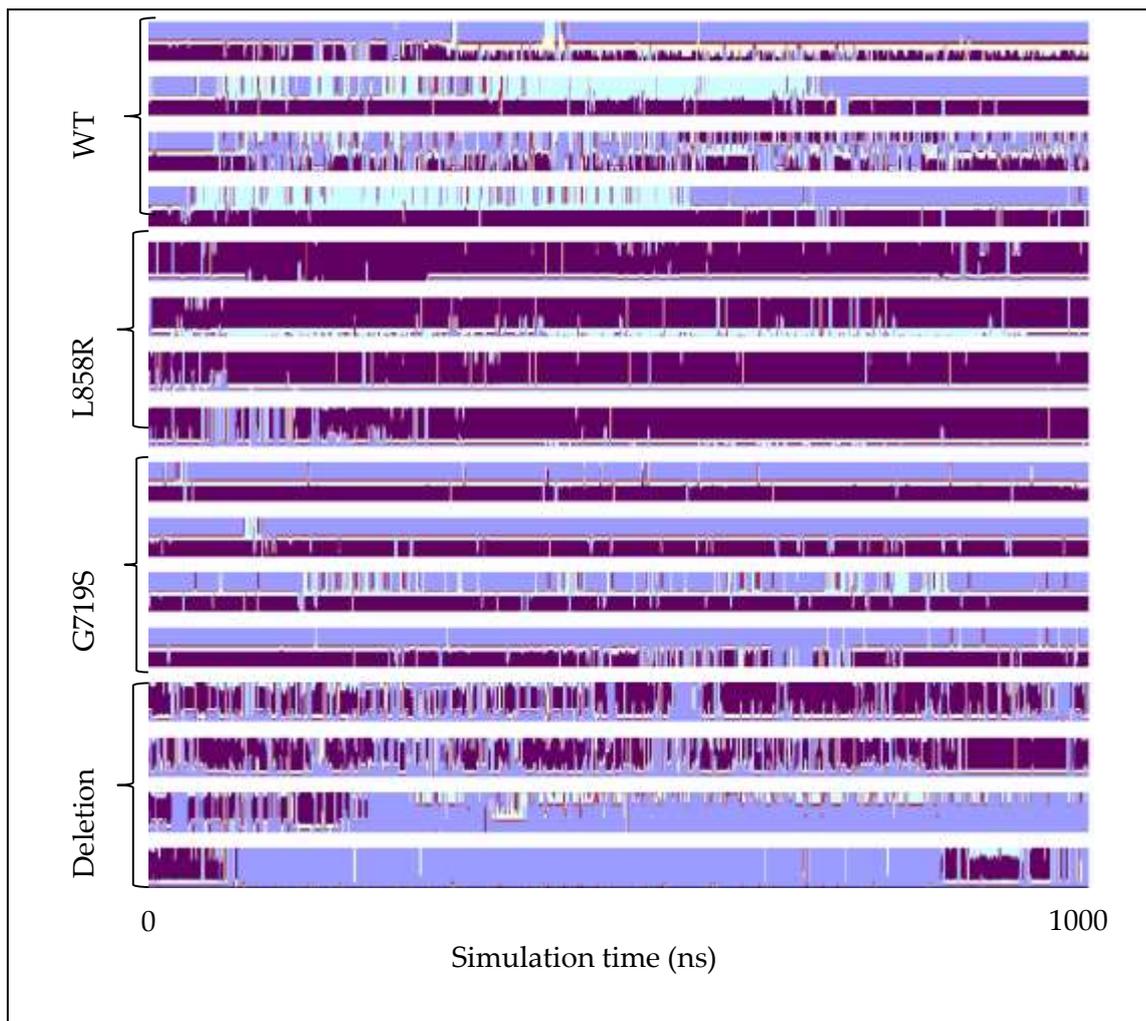


Figure 5.48: STRIDE secondary structure analysis of the A-loop helix over the course of the inactive cMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

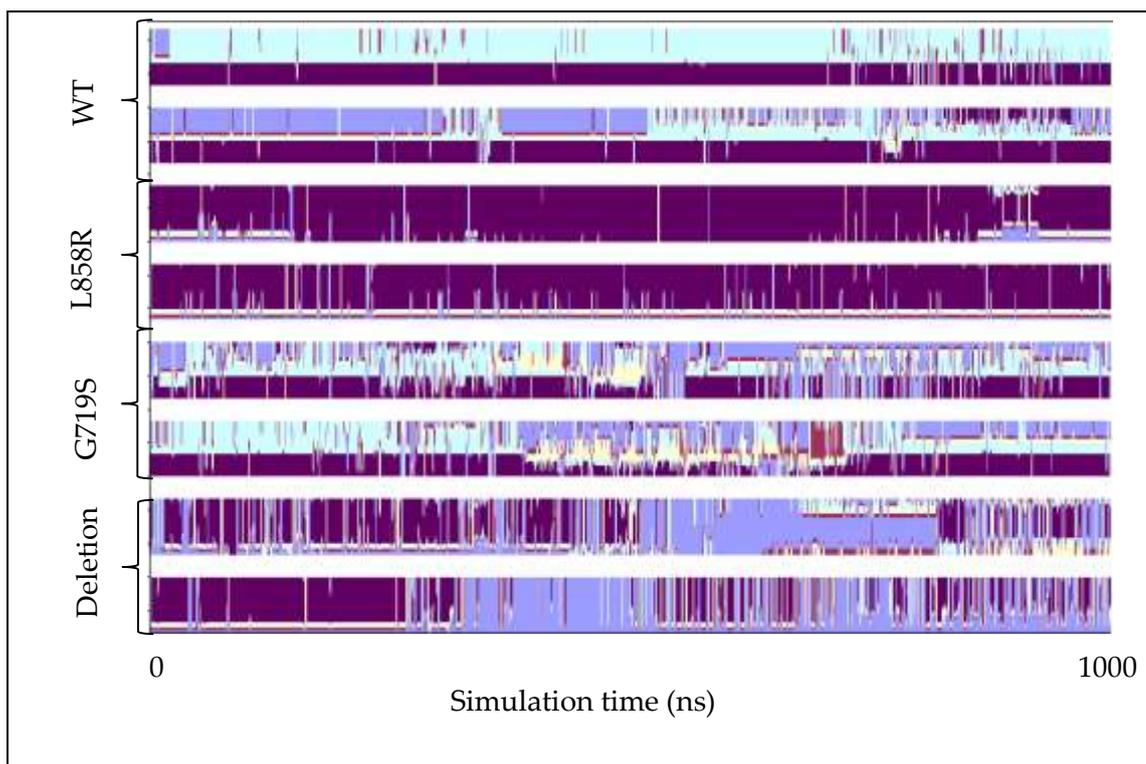


Figure 5.49: STRIDE secondary structure analysis of the A-loop helix over the course of the inactive AMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

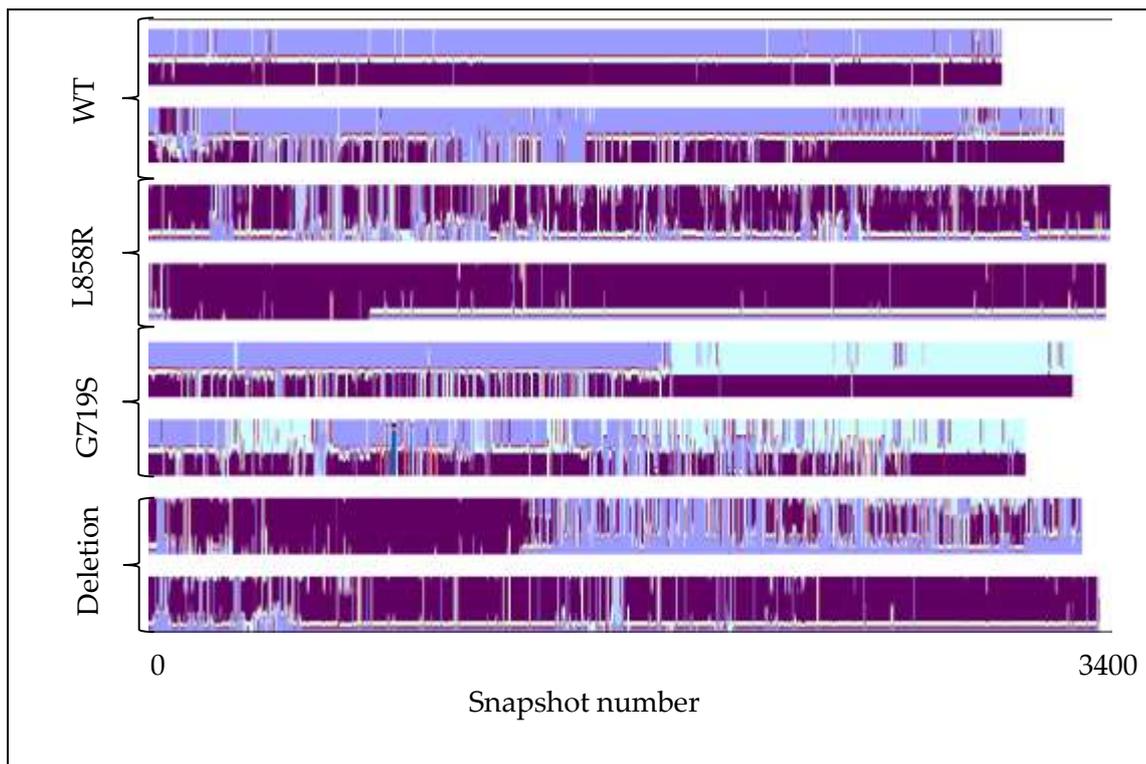


Figure 5.50: STRIDE secondary structure analysis of the A-loop helix over the course of the inactive DMDMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

The increased stability of the A-loop helix for the L858R is surprising, since theories to date have mostly rationalised the L858R as destabilising the inactive A-loop helix. It may be that this is not the case, but that the energy barriers for forming the inactive A-loop helix in the L858R mutant from the active conformation are prohibitively high: The TMD study by Papakyriakou et al. (2009) demonstrated that even using a  $2 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$  force constant, it was not possible to force the active A-loop into an inactive conformation[62].

## 5.5 Dimensionality reduction

### 5.5.1 PCA results

As described previously (see section 3.7.3), principal components analysis (PCA) is a useful tool for reducing the dimensionality of a dataset based on the covariance of data. In the present study, these aspects of PCA are doubly useful: the reduction in dimensionality makes rationalising the results simpler and the grouping of structures by similarity aids in the identification of the differential sampling between the mutants and WT, as we shall see later.

Supertrajectories were constructed by concatenating all trajectories of the same sampling method; additionally, the atoms of the deletion region were removed from the non-deletion snapshots to ensure that the number of  $\alpha$ -carbon atoms included from each simulation was consistent. These trajectories were aligned by RMSD fitting of backbone  $\alpha$ -carbon positions to those of the first frame of the supertrajectory (which corresponded to the deletion model) using the ptraj program from the AMBER tools software suite[114]. Since the main purpose of the analysis was to identify backbone motions, only the backbone  $\alpha$ -Carbons were kept for the PCA.

PCA was carried out using the bio3D module[137] (an extension to the R program[138]). It is important to note that only the first two principal components were analysed in detail, since initial investigation suggested that these would be the most useful in categorising the different sampling that occurs between the mutant simulations, with higher ranking PCs showing little ability to differentiate between the sampling of different mutations. Nonetheless, as the first two PCs account for between 40.7-68% of the variance in the datasets, this does represent a considerable approximation.

As can be seen from figure 5.51 the first principal component (PC) differentiates between the active and inactive trajectories. Indeed, the first PC corresponds roughly to the difference between the active and inactive conformations (see figure 5.52a), a trend that is consistent regardless of the sampling method used in the present study. Considering the large difference between the inactive and active conformations, this is unsurprising, and has been seen in a previous study on EGFR[65]. PC2 appears to correlate with a torsional motion accompanied by some closure of the N-lobe (particularly the P-loop) against the C-lobe (see figure 5.52b).

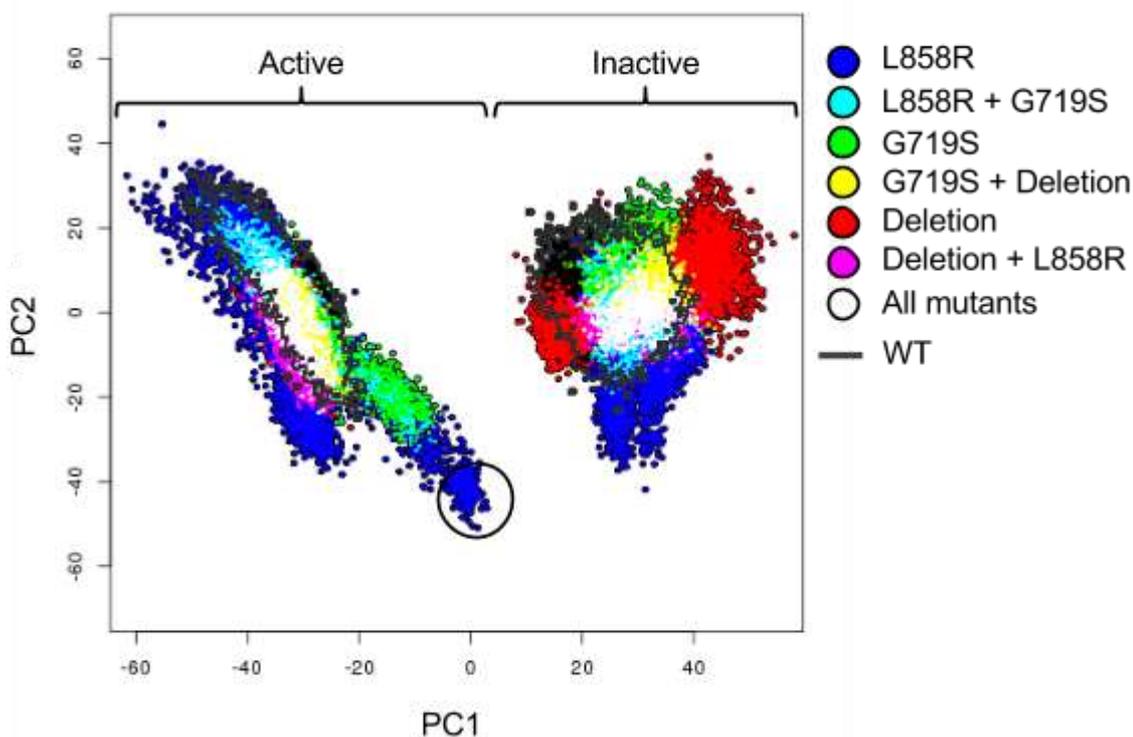


Figure 5.51: Snapshots from all cMD simulations projected onto PC1 and PC2 of the supertrajectory PCA with the mutant status of the snapshots differentiated by colour. Where two or more snapshots with different mutant status occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the WT, which is represented in black encased in a dark grey outline. The final frames of the active L858R model 3 simulation are circled.

Interestingly, the active deletion mutation appears highly stabilised with barely any sampling across the first two PCs, compared to the other mutants and WT (see figure 5.51), a feature which is present in all the sampling methods (see figures 5.53 5.54) except RDFMD (where the active G719S appeared to sample less; see figure 5.58). Conversely, the deletion appears to be the most capable mutant at sampling PC1 in the inactive conformation. Thus, when limiting discussion to the sampling across PC1, the deletion appears to conform entirely to the notion of activating mutants stabilising the active conformation, and destabilising the inactive conformation.

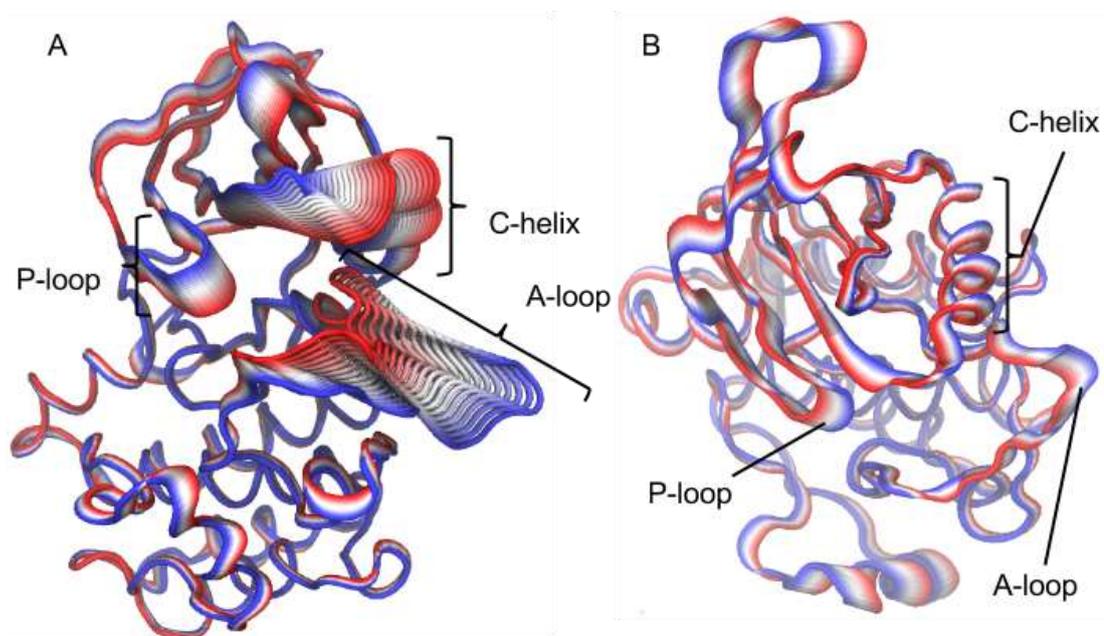


Figure 5.52: Representations of the backbone atomic displacements captured in PC1 (A) and PC2 (B) of the cMD supertrajectory. In (B), the protein has been rotated approximately 90 degrees along the horizontal axis of the page, relative to (A), in order to better show the torsional motion.

While the remaining simulations are each able to sample different PC1/PC2 space, it is difficult to draw any convincing conclusions from this PCA data alone. Nonetheless, on closer inspection, some of these differences are correlated with seemingly relevant

conformations. The most prominent difference is perhaps the exceptional sampling of the active L858R mutant compared to the other simulations, particularly where the PC score becomes close to that of the inactive trajectories (see figure 5.51, circled), which will be discussed later. Again, the greater propensity for the active L858R mutant to sample PC1 and PC2 is an observation that is consistent across the sampling methods, with the exception of RDFMD (see figures 5.54, 5.56, and 5.58).

Applying PCA to the AMD supertrajectory, it was found that PC1 is consistent with the cMD simulations, but that PC2 corresponds to an opening/closing of the N-lobe (see figure 5.53). On closer inspection, it seems that this opening/closing is also a feature of the cMD simulations, but is incorporated into PC1 (see figure 5.52a)

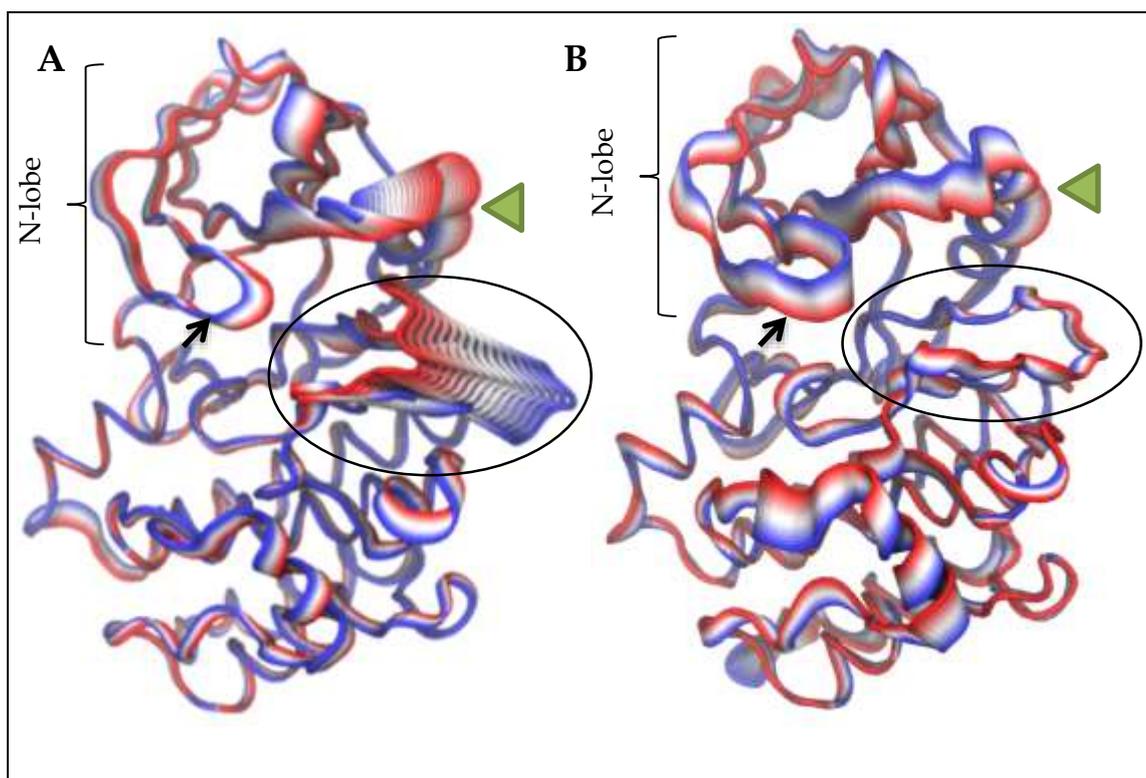


Figure 5.53: Representations of the backbone atomic displacements captured in PC1 (A) and PC2 (B) of the AMD supertrajectory. The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

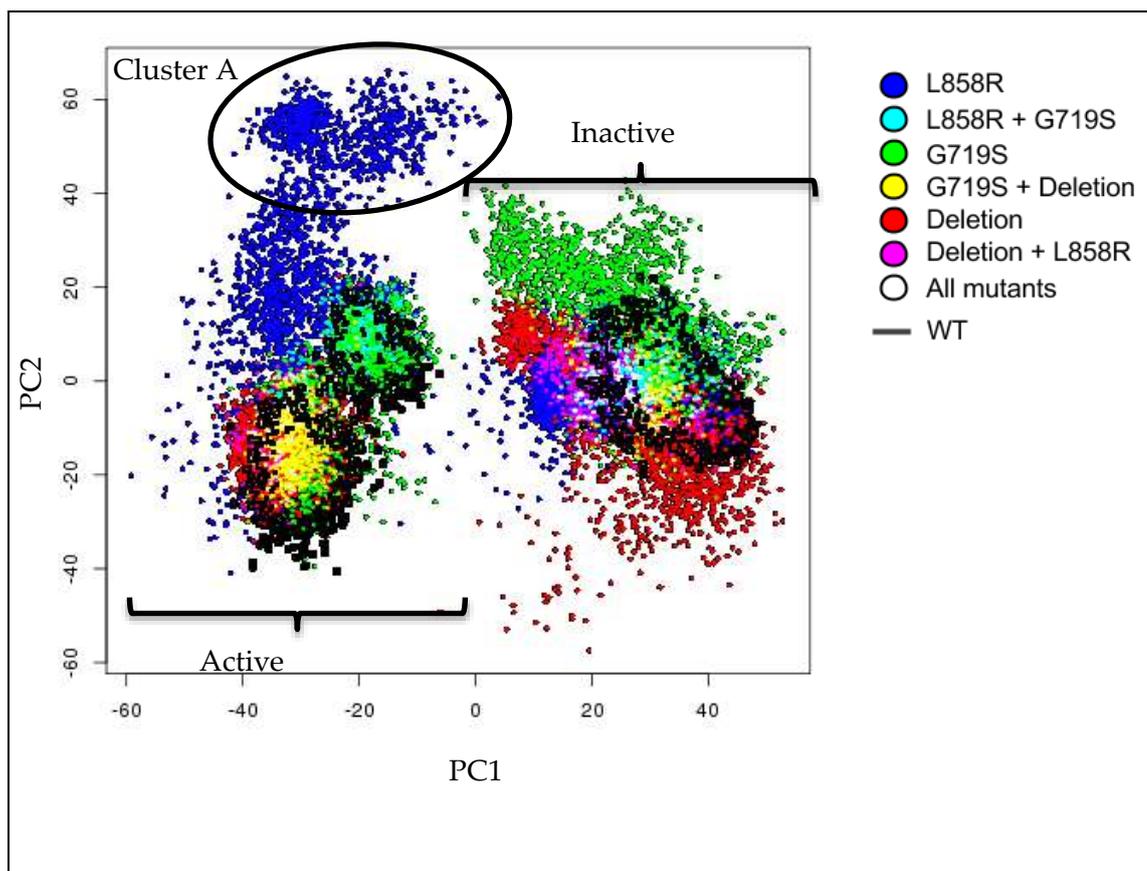


Figure 5.54: Snapshots from all AMD simulations projected onto PC1 and PC2 of the supertrajectory PCA with the mutant status of the snapshots differentiated by colour. Where two or more snapshots with different mutant status occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the WT, which is represented in black encased with a black outline. A cluster unique to one of the L858R models is circled (A).

Given that closing of the N-lobe (sampling on PC2) does not necessarily lead to better sampling along the first PC (and thus presumably less progress along the active-inactive transition) in the AMD simulations (see figure 5.54), it may be possible that the propensity of the L858R mutant to exhibit closing of the N-lobe may represent a “dead-end” (cluster A, figure 5.54), from which sampling to the inactive conformation is impossible, but return to the active conformation is possible. Nonetheless, given the short simulation times, it is impossible to rule out the possibility of an active-inactive transition occurring from this cluster.

The AMD supertrajectory PCA appears to have sampled the most out of the enhanced sampling methods, based on the diffuse distribution of the snapshots and a mild overlap in the sampling of the active and inactive trajectories along PC1 (see figure 5.54). By the same criteria, the other enhanced sampling methods appear to have performed worse (see figures 5.56 and 5.58).

The DMDMD supertrajectory has a similar PC1 to the AMD and cMD trajectories, however PC2 correlates to a motion in the C-terminal (see figure 5.55B), where the L858R appears to have sampled more extensively (see figure 5.56).

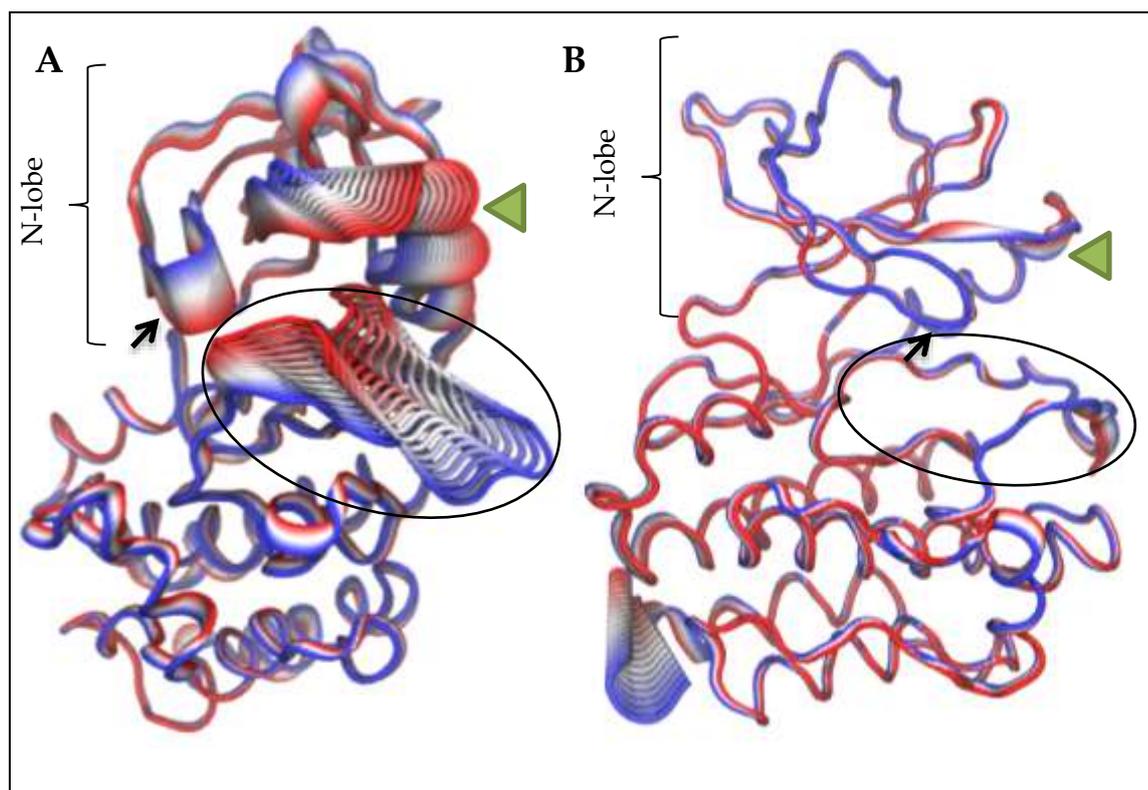


Figure 5.56: Representations of the backbone atomic displacements captured in PC1 (A) and PC2 (B) of the DMDMD supertrajectory. The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

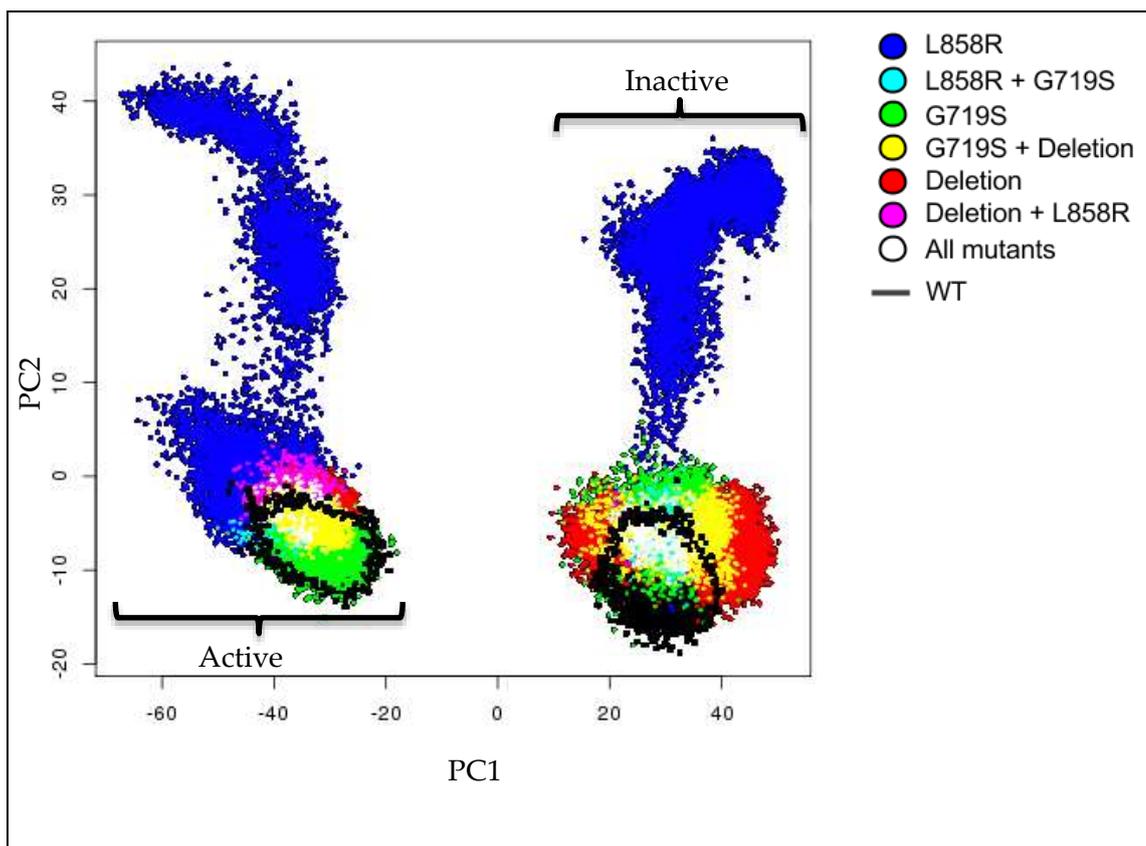


Figure 5.56: Snapshots from all DMDMD simulations projected onto PC1 and PC2 of the supertrajectory PCA with the mutant status of the snapshots differentiated by colour. Where two or more snapshots with different mutant status occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the WT, which is represented in black encased with a black outline.

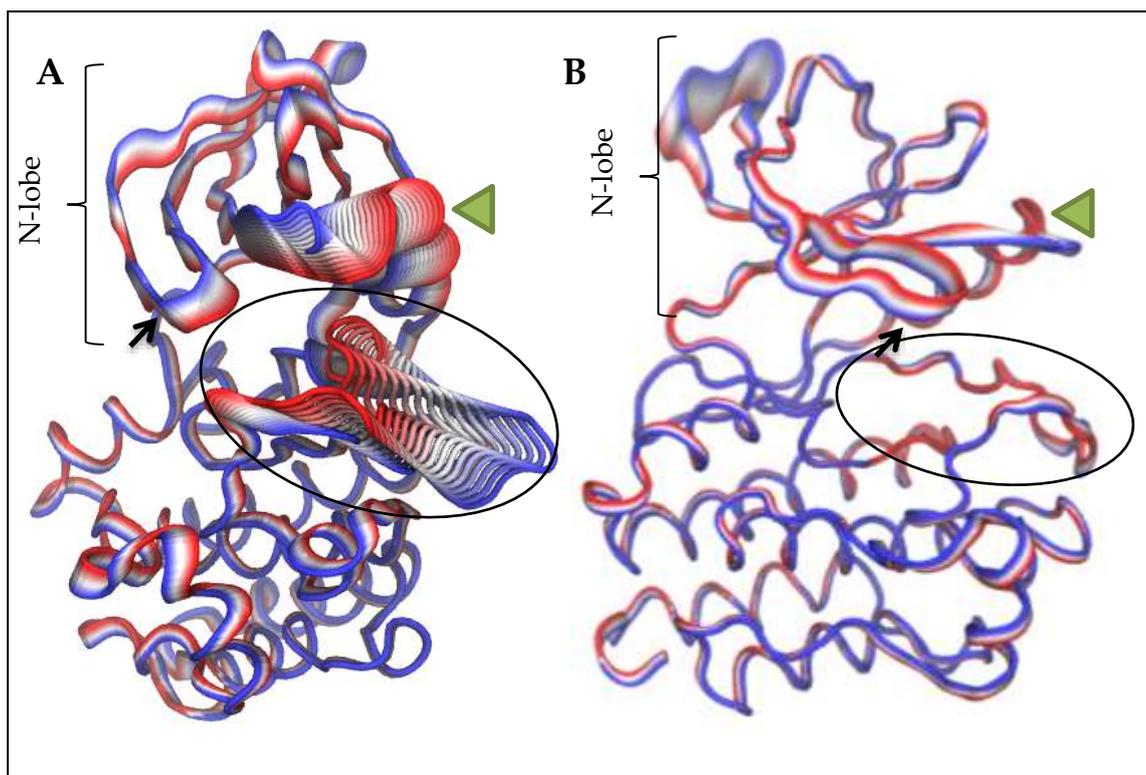


Figure 5.57: Representations of the backbone atomic displacements captured in PC1 (A) and PC2 (B) of the RDFMD supertrajectory. The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

The RDFMD supertrajectory appears to sample the least PC1 space out of the enhanced sampling methods (see figure 5.58), as discussed above, the RDFMD also has different sampling patterns for the mutants, with the deletion sampling more in the active conformation when compared to the WT, than in the other sampling methods.

Additionally, the sampling of the L858R mutant (particularly when comparing with the active conformation) appears less pronounced than AMD, despite the similarities in the motions captured by PC1 and PC2 (see figure 5.53 5.57). The markedly different sampling of PC2 in the RDFMD simulations (with the deletion mutant dominating, rather than the L858R mutant) may be due to the shorter timescales employed in the RDFMD simulations (400 ps per simulation): the shorter time scales appear to have led to a reduced proportion of variance captured by the higher ranking PCs (see appendix 3), thus it seems reasonable that the difference between the starting configurations of the deletion and the other simulations (since, in the deletion, residues surrounding the

deletion region effectively had to be pulled together) may be captured by lower ranking PCs than in the other sampling methods. This appears to be the case from visualisation of representations of the PCs, but translates poorly into a static image (see figure 5.57).

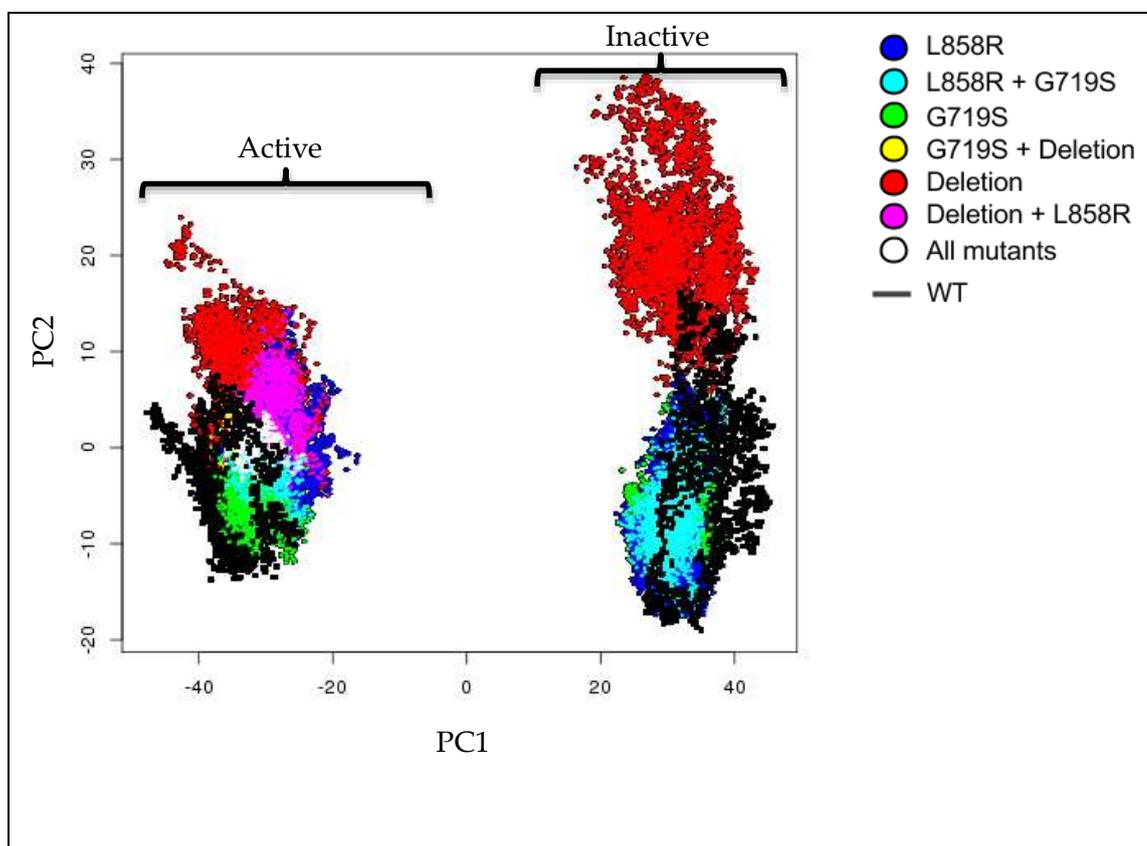


Figure 5.58: Snapshots from all RDFMD simulations projected onto PC1 and PC2 of the supertrajectory PCA with the mutant status of the snapshots differentiated by colour. Where two or more snapshots with different mutant status occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the WT, which is represented in black encased with a black outline.

To probe the different sampling of the deletion mutant along the second PC of the RDFMD simulations in more detail, the Pearson's correlation between PC1 and PC2 with inter-residue distances was calculated. This is achieved by calculating the Pearson's correlation coefficient between the inter-residue  $\alpha$ -carbon distances and the PC in question. A Pearson's R value of -1 or 1 corresponds to a perfect correlation or

anti-correlation, respectively, between the interatomic distance and the PC or DC; a Pearson's R value of 0 corresponds to the lack of a correlation.

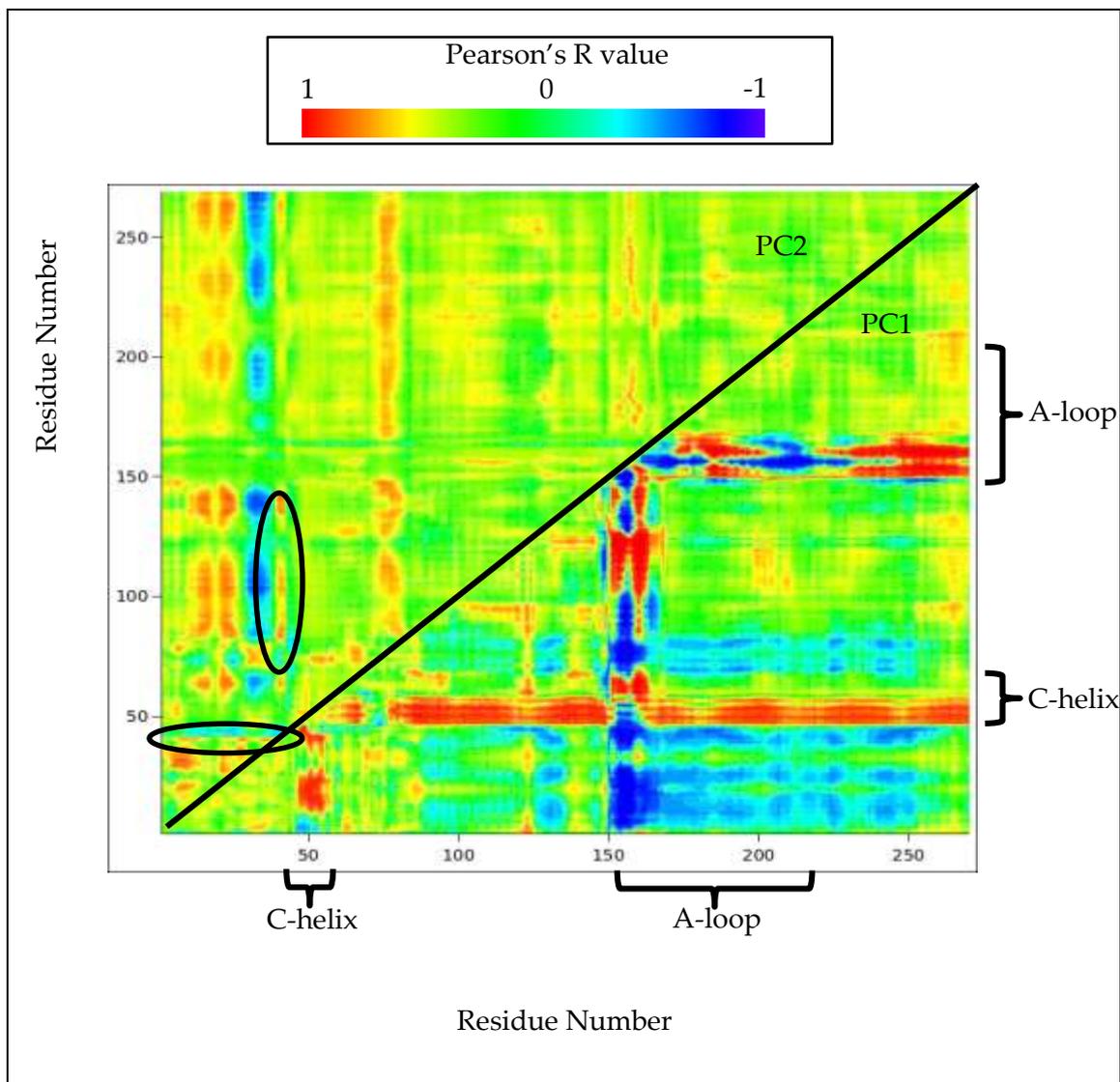


Figure 5.59: Pearson's correlation coefficient of the correlation between inter-residue  $\alpha$ -carbon distances and both PC1 and PC2 of the RDFMD simulation. Correlations between PC2 and the distance between the deletion region and the rest of the N-lobe are circled.

If the different sampling of the deletion on PC2 was due the differences between the starting structures, this may show up as a correlation between PC2 and the distance

between the deletion region and the rest of the N-lobe, which appears to be the case (see figure 5.59); however, the correlation in this region is mostly obscured by the other motions in the protein that have been incorporated into this PC.

Another feature of the PCA is the difference in the proportion of variance accounted for by each PC. The proportion of variance accounted for by PC1 tends to correlate loosely with how close the active and inactive trajectories sample on the first PC1. For example, PC1 of the AMD supertrajectory PCA accounts for 29.3% of the variance in the dataset, and there is a slight overlap between the sampling of the active and inactive trajectories on PC1; whereas the first PC of the RDFMD trajectory accounts for 61.5% of the variance of the dataset and has a large gap between the sampling of the active and inactive trajectories. Reductions in the proportion of variance captured by PC1 suggest that other motions are being more extensively sampled (although this does include motions corresponding to noise). The fact that the enhanced sampling methods (with the exception of RDFMD) incorporate a larger proportion of variance in their other PCs suggests that at least some of this extra sampling is occurring in important motions rather than noise, which is expected to occupy the smaller PCs with the smaller eigenvalue (compare figures 5.60, 5.61, 5.62, and 5.63).

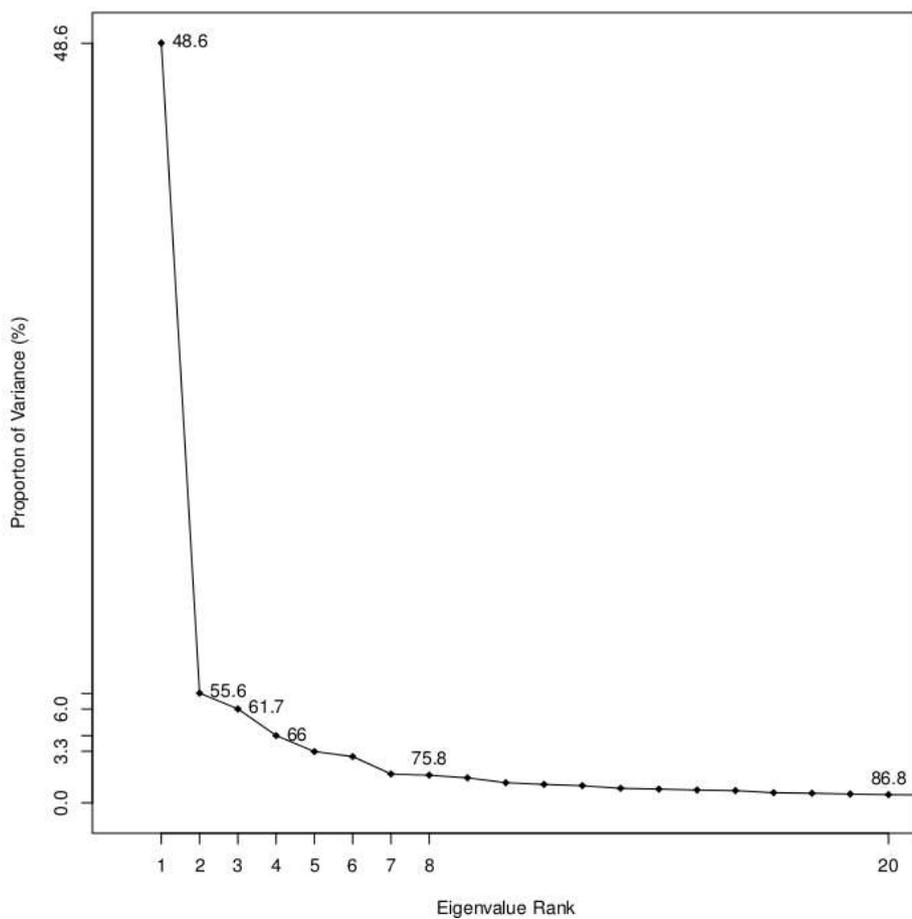


Figure 5.60: Scree plot of the proportion of variance accounted for by each principal component of the cMD supertrajectory

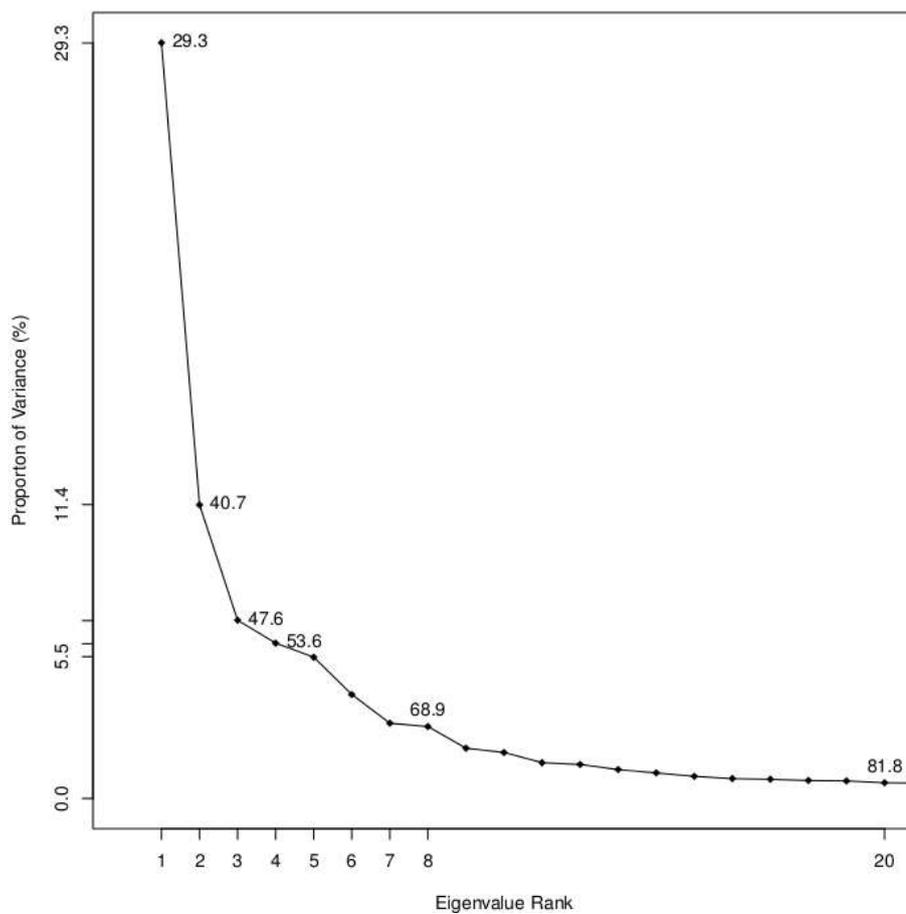


Figure 5.61: Scree plot of the proportion of variance accounted for by each principal component of the AMD supertrajectory

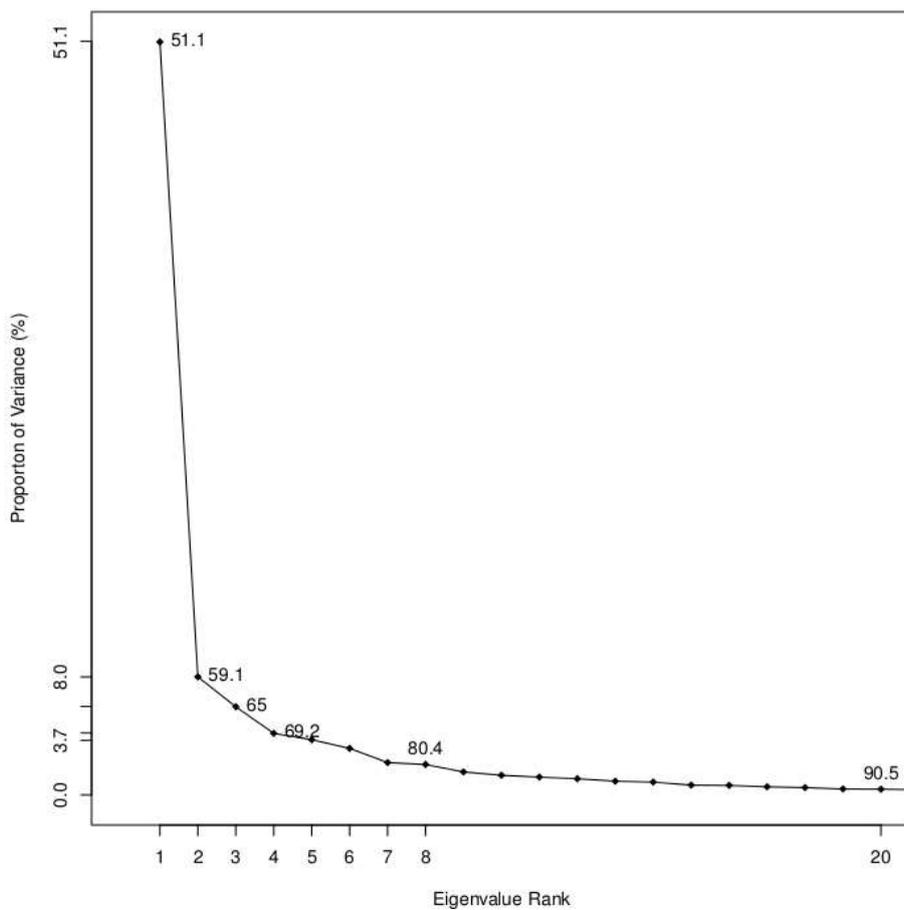


Figure 5.62: Scree plot of the proportion of variance accounted for by each principal component of the DMDMD supertrajectory

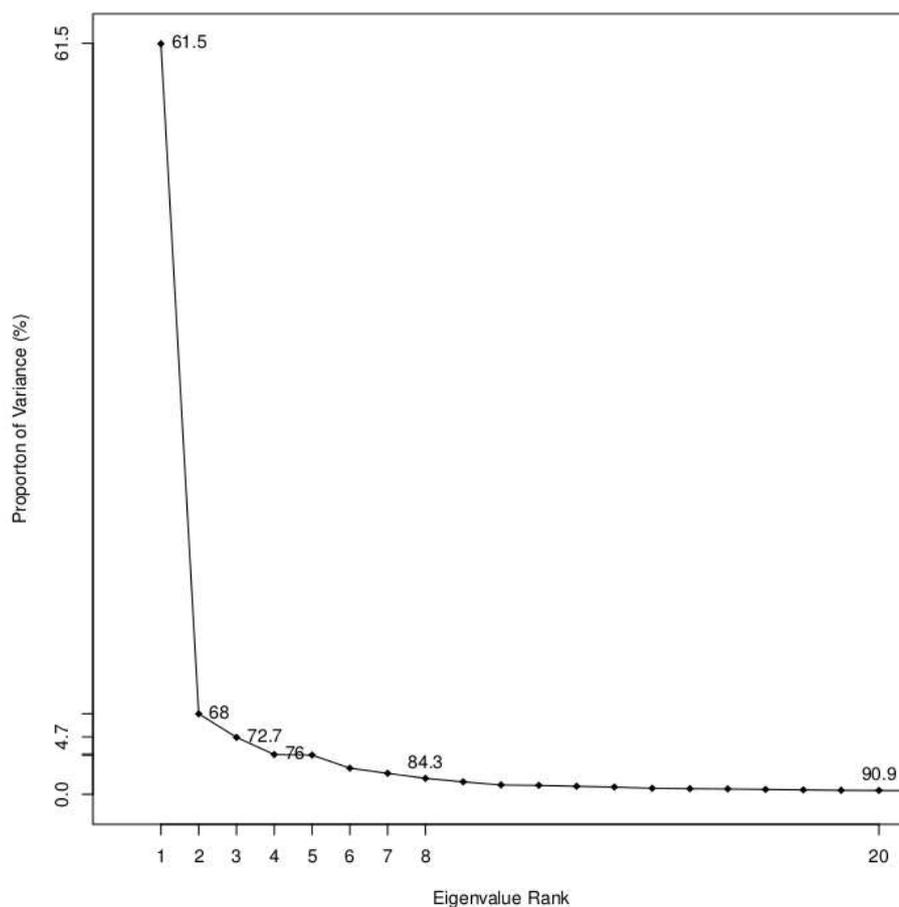


Figure 5.63: Scree plot of the proportion of variance accounted for by each principal component of the RDFMD supertrajectory

To attempt to gain a clearer picture of the sampling of the each conformation of the EGFR kinase monomer, an additional set of supertrajectories were produced for PCA. The new supertrajectories were comprised of all the simulations for a particular mutant in a particular conformation (see appendix 3). For example, one such supertrajectory consisted of all simulations (cMD, AMD, DMDMD, RDFMD) of the active L858R monomer (see appendix 3.1). The benefits of these supertrajectories are twofold: Firstly, it becomes possible to examine the general dynamics of a mutant without the biasing effect of the inactive structures on the first PC, secondly it allows us to examine the extent to which using different sampling methods leads to different conformations, a point that will be discussed later.

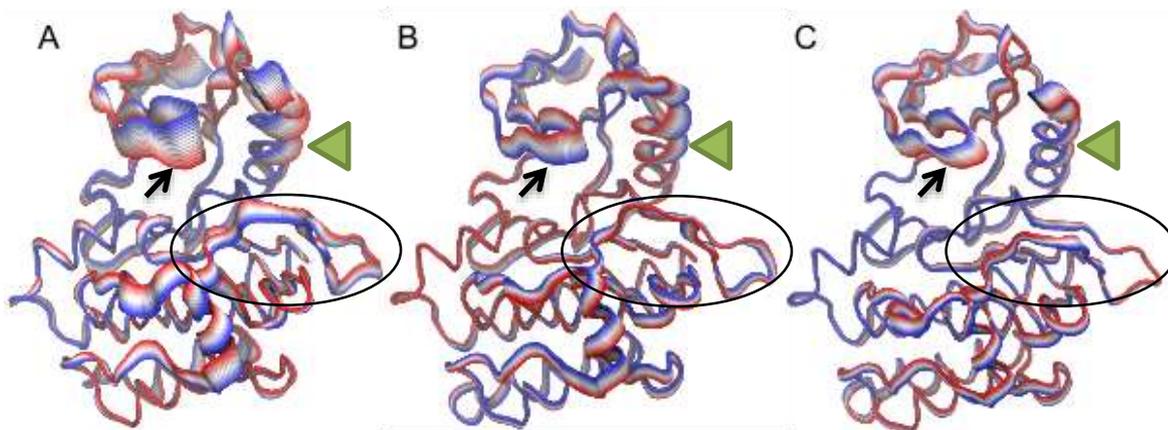


Figure 5.64: Representations of the backbone atomic displacements captured in PC1 of the active simulations of the L858R mutant (A), G719S mutant (B), and WT (C). The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

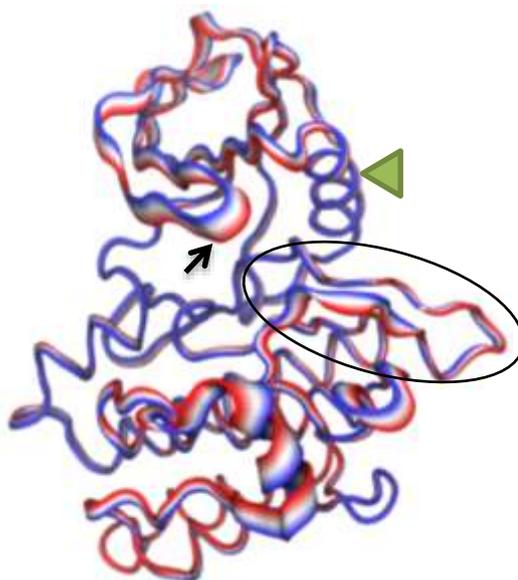


Figure 5.65: Representations of the backbone atomic displacements captured in PC1 of the active simulations of the Deletion mutant. The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

Interestingly the first PC of the active trajectories captures not only the motion of the C-helix to the inactive “out” conformation, but the rotation of the entire N-lobe such that the P-loop closes on the A-loop (see figure 5.64). This finding is consistent across all the active trajectories except the deletion (the C-helix movement does not occur; see figure 5.65). The closing motion is most pronounced in the L858R, with the G719S being somewhat less pronounced, and less still in the WT.

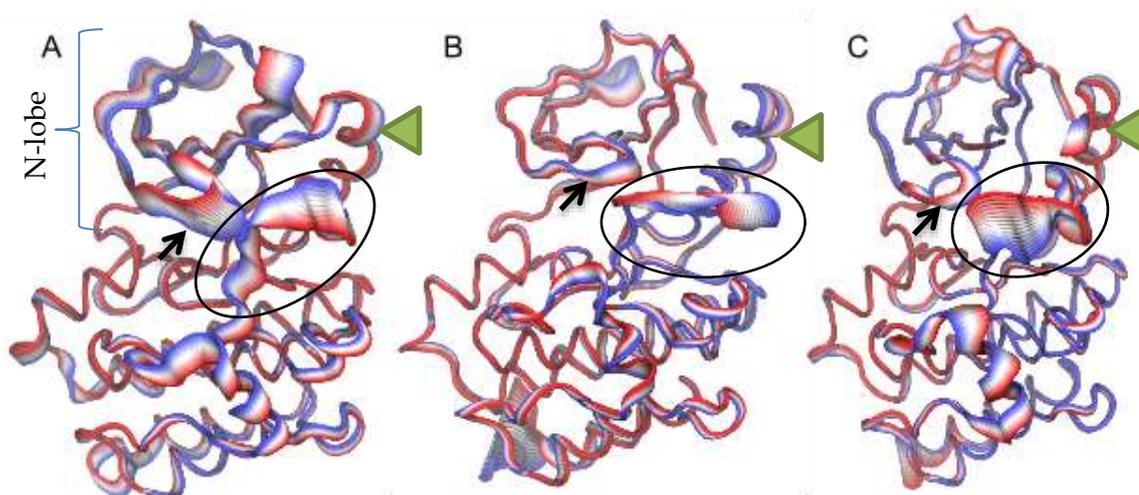


Figure 5.66: Representations of the backbone atomic displacements captured in PC1 of the supertrajectory for the inactive conformation of the Deletion (A), L858R (B), and G719S (C) mutants. The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

The PCAs of each supertrajectory of the inactive systems (again, each including all sampling methods) reveal a similar motion in the first PC, that resembles the opening/closing of the N-lobe coupled with rotation of the C-helix (see figure 5.66), as in the active trajectories (see figure 5.64). In the inactive trajectories, the motion of the C-helix is far less pronounced than in the active trajectories, while the flexibility of the rest of the N-lobe for the deletion is comparable to that of the active L858R mutant (compare figure 5.66 with figure 5.64). However, more of the variance in the first PC is

due to motions of the A-loop (see figure 5.66, which differ for all the systems, and are not consistent among different simulations of the same system), due to the high mobility of the inactive A-loop.

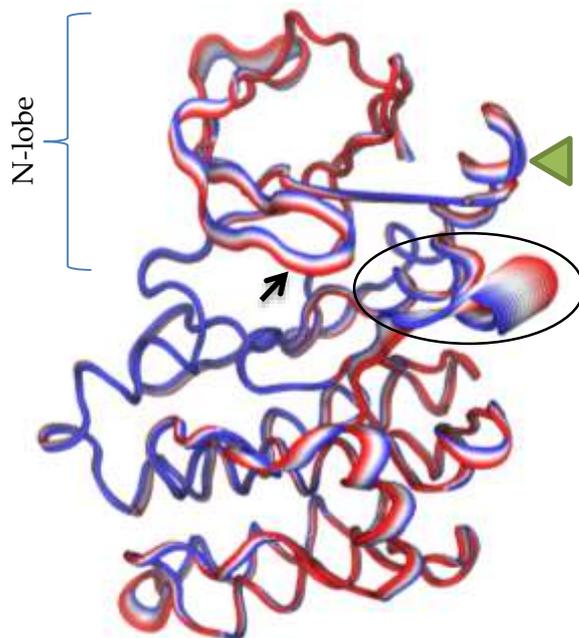


Figure 5.67: Representations of the backbone atomic displacements captured in PC1 for the inactive WT supertrajectory. The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

Figure 5.66 shows the motion of the N-lobe, which is particularly prominent in the region of the P-loop. The coupling of this motion to the C-helix does not appear to be on the same scale as was observed for the active trajectories (see figure 5.64), but C-helix conformations closer to the “in” state do appear to be correlated with a more open N-lobe (see figure 5.66).

### 5.5.2 Diffusion map results

Diffusion maps were prepared using the same supertrajectories as the PCA, which were used as input for the Locally Scaled Diffusion Map software[101]. The diffusion map results are similar to the PCA results, separating the dataset into active and inactive trajectories along the first diffusion coordinate (DC). The second diffusion coordinate also appears to correspond to the same motions as captured by the second PC of the PCA (compare figure 5.68 with figure 5.51). These similarities were further investigated by quantifying the correlation between the inter-residue distances and the DCs.

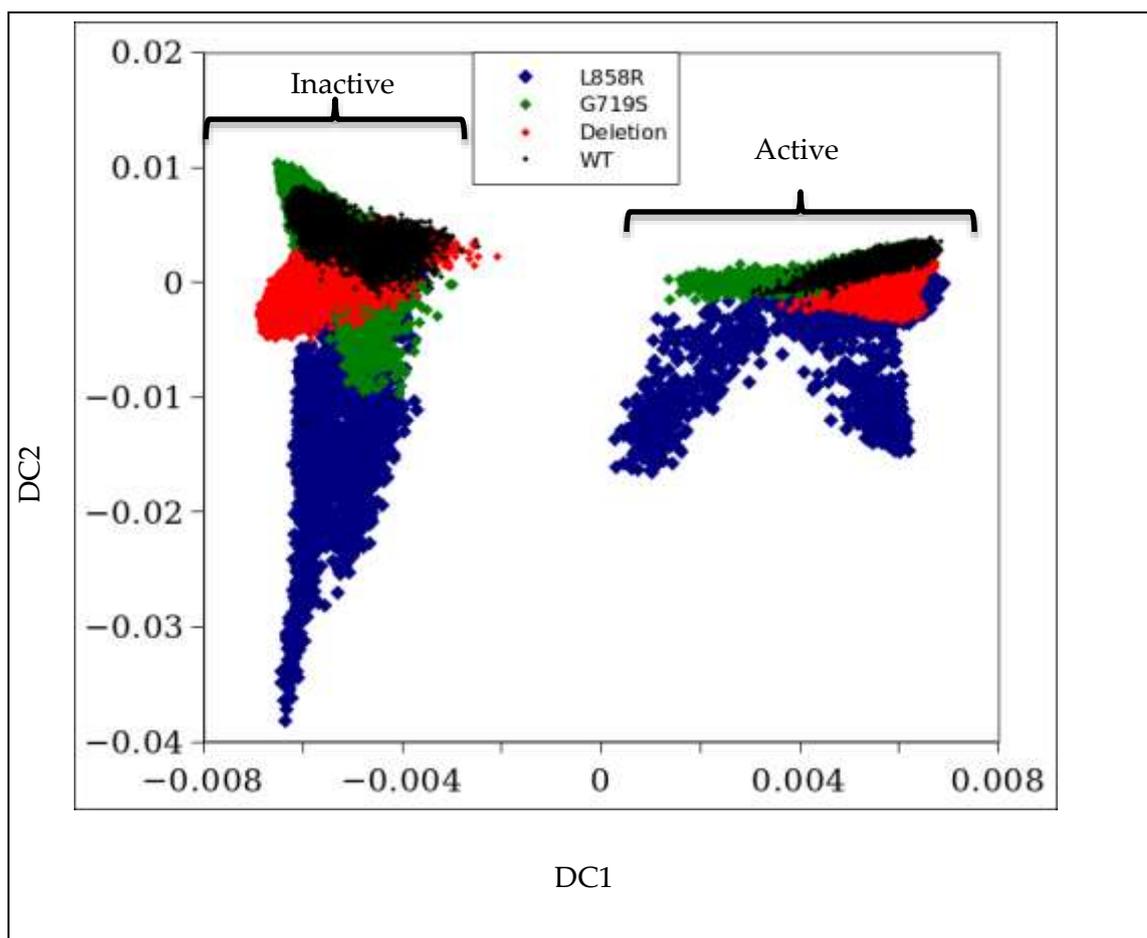


Figure 5.68: cMD supertrajectory projected onto the first two DCs of the diffusion map analysis, each mutant and WT is represented by a different colour (see legend).

Figure 5.69 shows a strong correlation between DC1 and the A-loop and C-helix; when taken together with the separation of the active and inactive trajectories on the first DC, it seems likely this DC is due to differences between the active and inactive conformations, similarly to PC1 of the PCA. Another prominent feature is the high correlation between DC1 and the distance between the N-lobe and C-lobe, which may correspond to the opening/closing motion also found in the first PC of the PCA (see figure 5.52). DC2 shows little correlation with inter-residue distances except for at the C-terminal, which seems likely to be due to the relatively mobile C-terminal region.

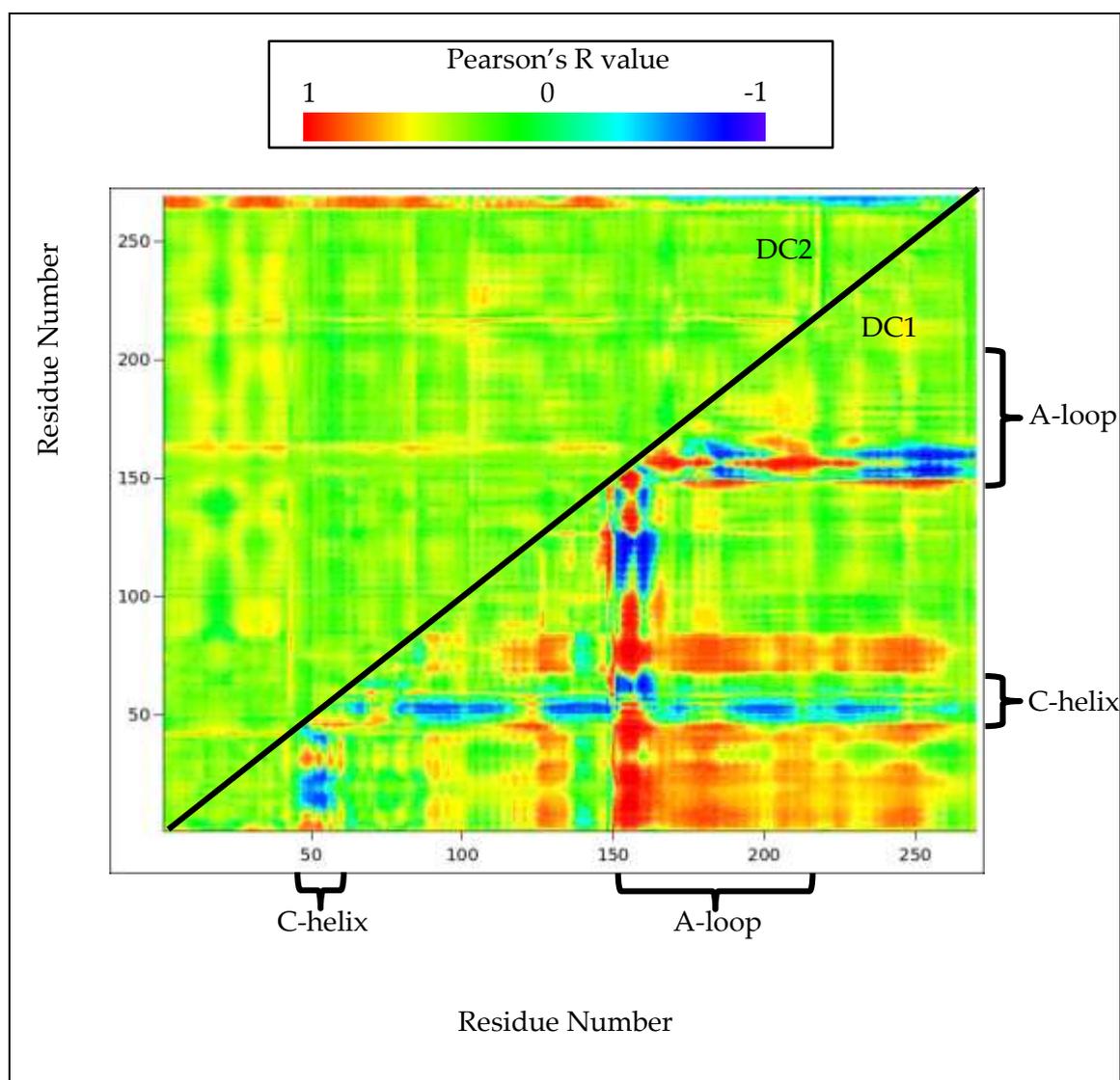


Figure 5.69: Pearson's correlation coefficient of the correlation between inter-residue  $\alpha$ -carbon distances and both DC1 and DC2 of the cMD simulation.

The AMD supertrajectory diffusion map (see figure 5.70) is interesting in that it plots the active trajectories much closer to the inactive trajectories, additionally, it was found that one cluster in the PCA (cluster A, see figure 5.54) disappears in the diffusion map. The identification of this cluster in the diffusion map was performed by cross-referencing the supertrajectory frames (since the supertrajectory used in the diffusion map analysis and PCA were the same).

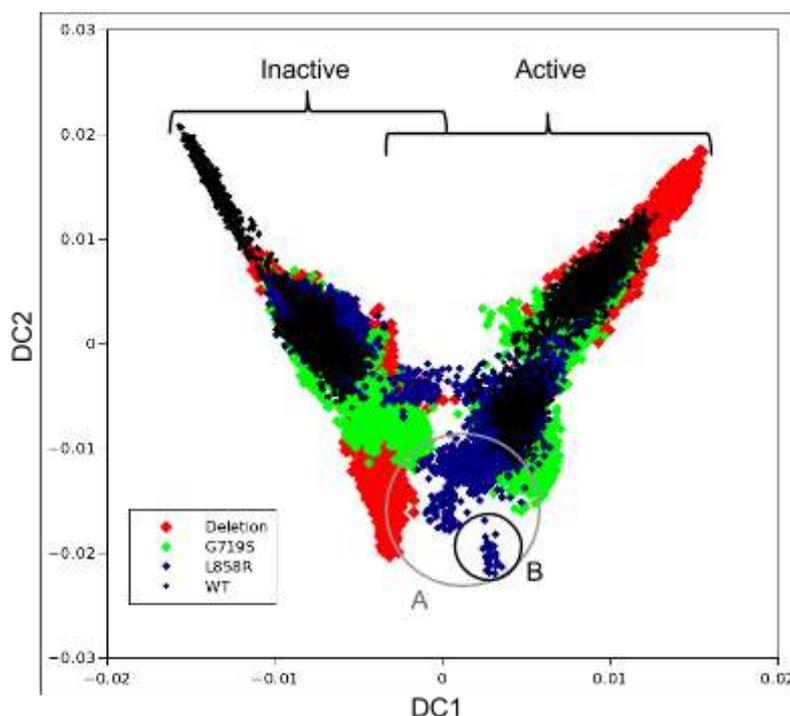


Figure 5.70: AMD supertrajectory projected onto the first two DCs of the diffusion map analysis, each mutant and WT is represented by a different colour (see legend). Cluster A from figure 5.54 consists of the blue markers (L858R snapshots) circled in grey, with a sub cluster (B) circled in black.

To investigate this further, additional Pearson's correlation coefficient plots were produced for PC1 and PC2 to determine whether the DCs correspond to the same motions as the PCs (see figure 5.71). Although there are some differences between each PC and the corresponding DC (which is perhaps unsurprising considering the different

way the analyses are performed), the DC correlation patterns are very similar to the PC ones.

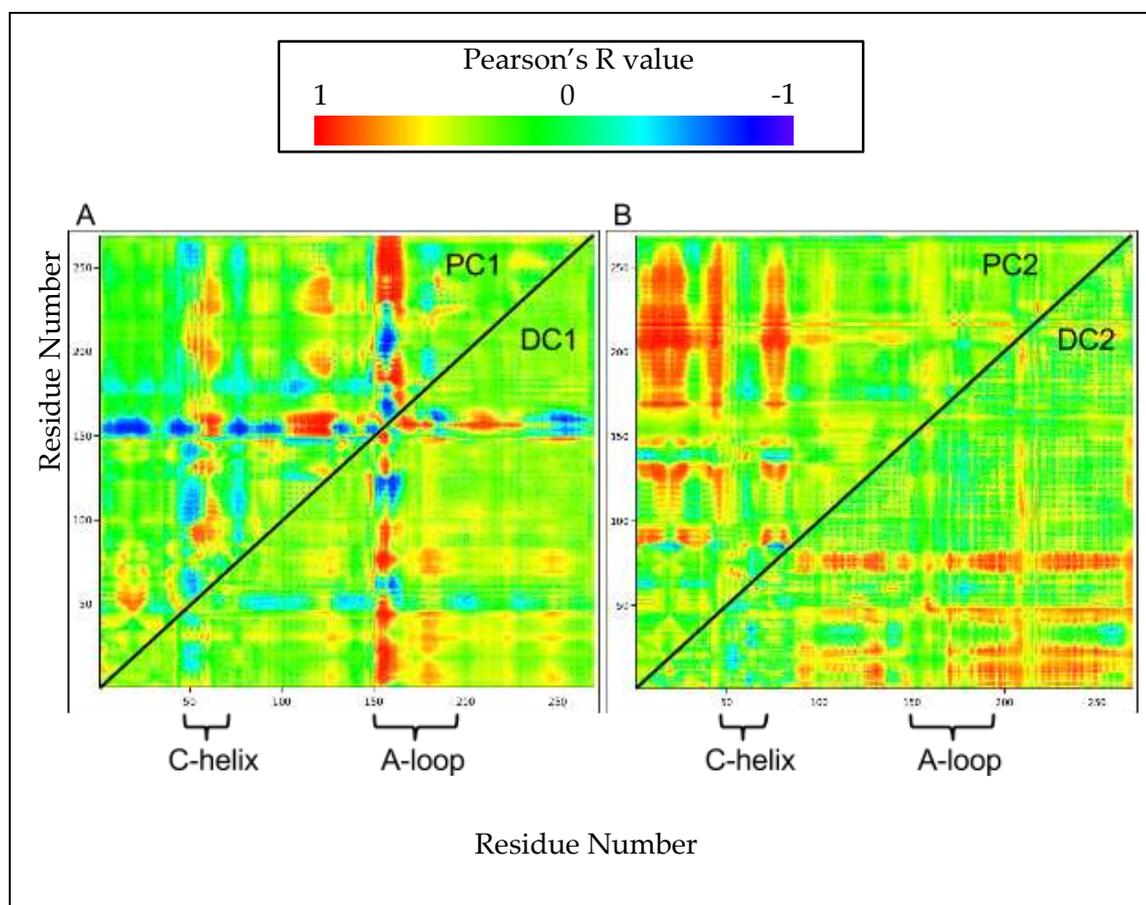


Figure 5.71: Pearson's Correlation coefficient of the correlation between inter-residue  $\alpha$ -carbon distances and both PC1 and DC1 (A), as well as PC2 and DC2 (B) of the AMD supertrajectory.

Since the PCs and DCs investigated here appear to correspond well with each other, the fact cluster A disappears in the diffusion map analysis is probably not due to the diffusion coordinates corresponding to different motions. It is therefore likely that cluster A, though significantly different from the other L858R conformations (as evidenced by the distance between cluster A and the other conformations on the PCA; see figure 5.54), it is readily accessible from the other L858R conformations. The fact that cluster A appears closer to the inactive conformations on the diffusion map would

also seem to suggest that rather than representing a possible “dead-end” (see section 5.5.1), cluster A may represent a possible transition pathway.

A sub-cluster of cluster A exists that is further from the inactive structures than the rest of cluster A (cluster B, figure 5.70), however sampling between cluster B and the rest of cluster A happens over such a short time scale, it is unlikely to be a significant impediment to transitioning between the active and inactive conformations (see figure 5.72). Additionally, the relatively low number of snapshots in and around cluster B means that the error associated with that cluster will be relatively high (Wenwei Zheng, private communication, Rice University, 2012).

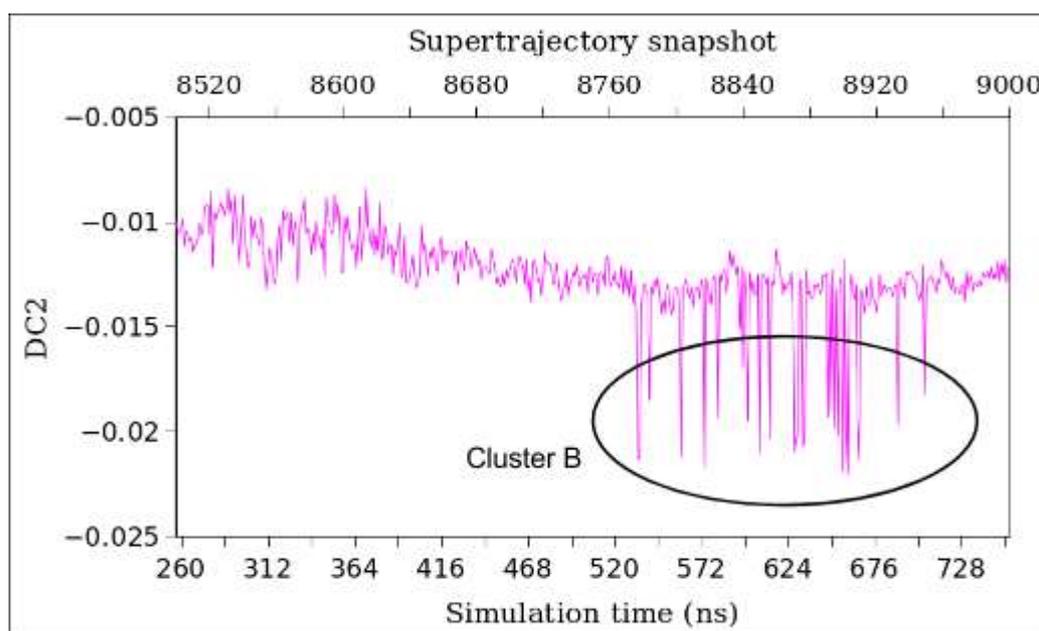


Figure 5.72: Sampling of a portion of model 2 of the active L858R mutation simulation on DC2, showing the short time-scale sampling between cluster B (circled) and the other regions of DC2 space.

The DMDMD trajectory diffusion map again appears similar to the PCA, with the L858R sampling extensively along DC2 (see figure 5.73) as it does along PC2 of the

PCA (see figure 5.56). Indeed, DC2 appears to correlate with the C-terminal (see figure 5.74), just as PC2 appears to correspond to C-terminal motions (see figure 5.55).

The extensive sampling of DC2 by the L858R, however, only appears to be a feature of the inactive L858R simulations, despite the PCA showing this extensive sampling for all the L858R simulations. This suggests that while considerable motions in the C-terminal do occur, the conformations produced by these motions more readily interconvert in the active trajectories. Unfortunately, it is not clear why this should be the case.

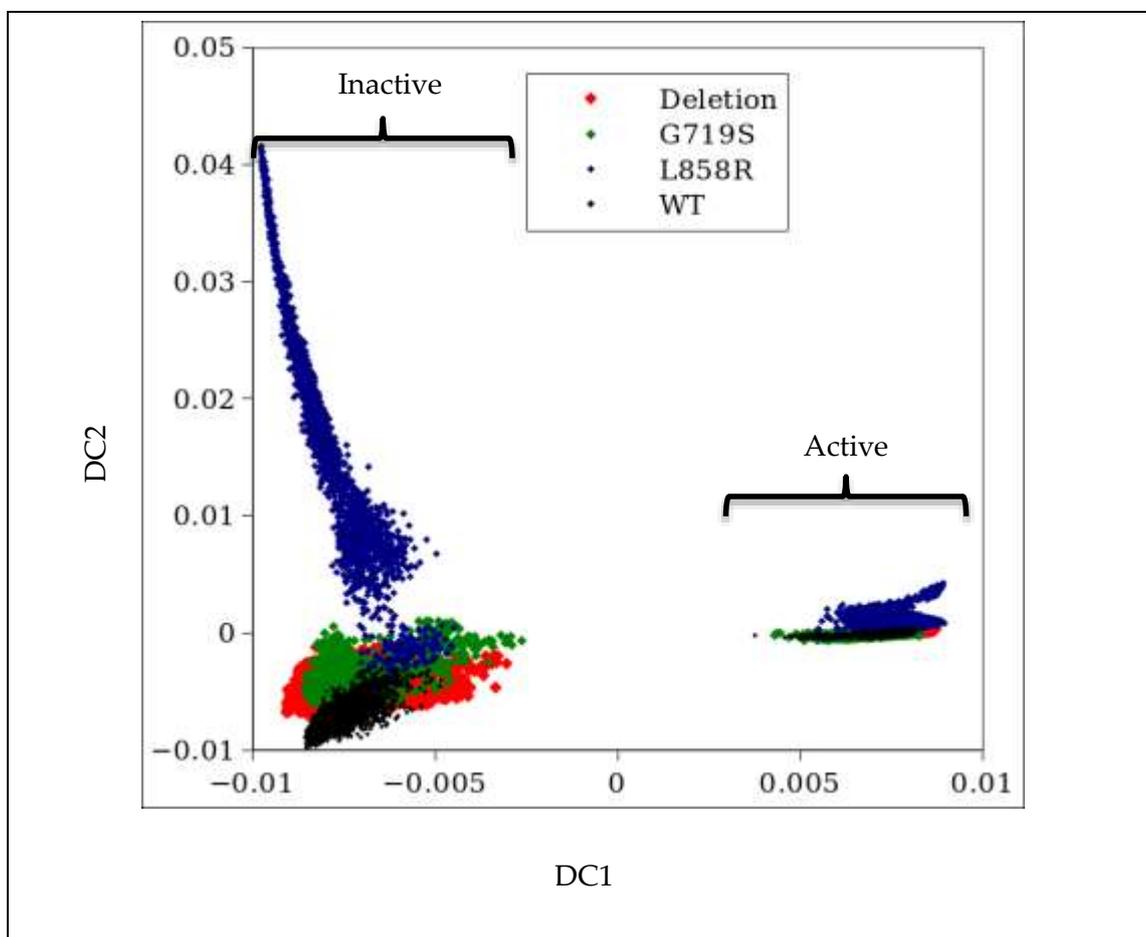


Figure 5.73: DMDMD supertrajectory projected onto the first two DCs of the diffusion map analysis, each mutant and WT is represented by a different colour (see legend).

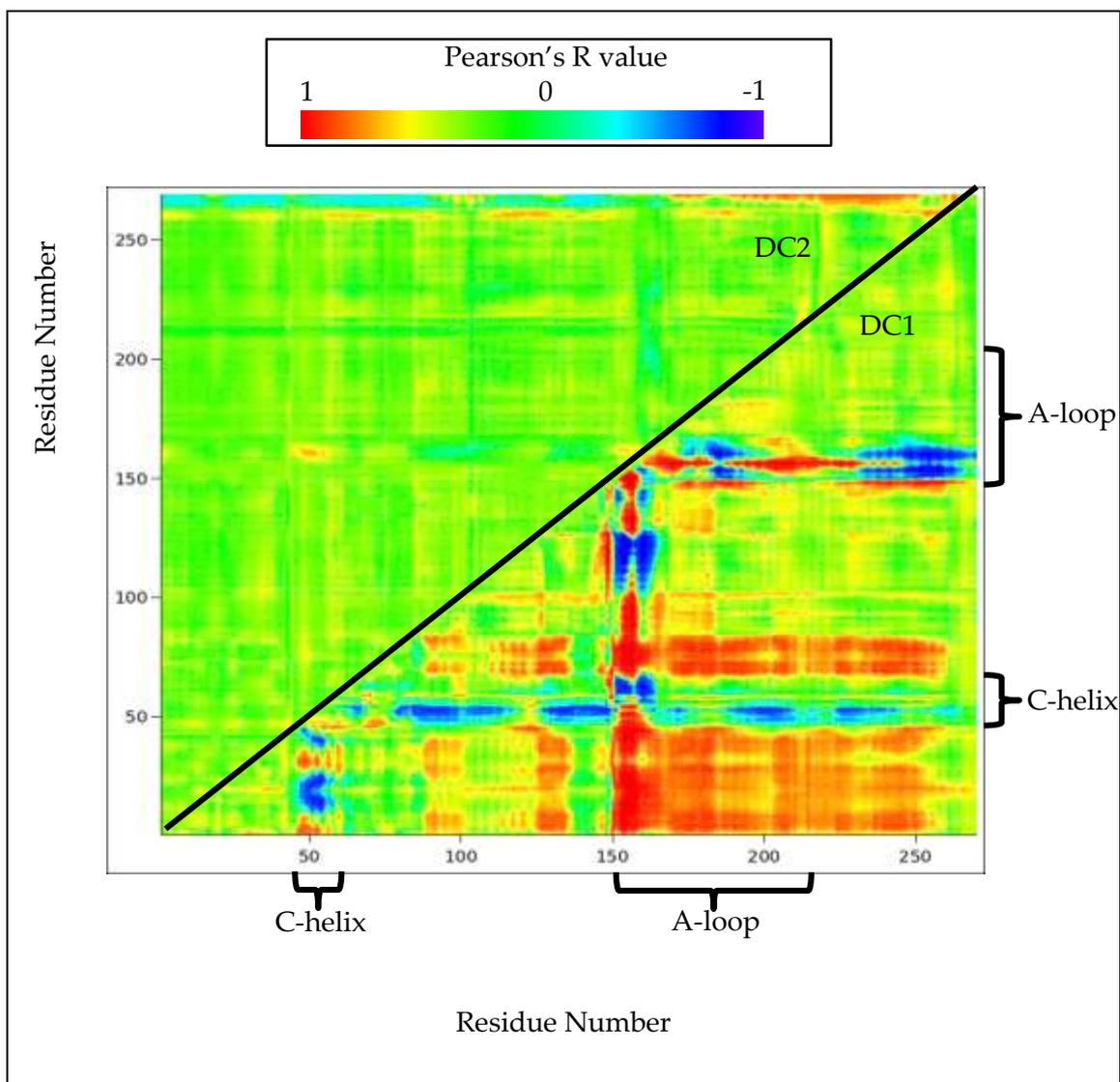


Figure 5.74: Pearson's correlation coefficient of the correlation between inter-residue  $\alpha$ -carbon distances and both DC1 and DC2 of the DMDMD simulation.

Similarly to the PCA (see figure 5.58), the diffusion map of the RDFMD simulations shows the deletion sampling more DC2 space, and this DC is strongly correlated with motion in the N-lobe (see figure 5.76). The pattern of correlation between DC2 and inter-residue distances (see figure 5.76) is very similar to that of PC2 (see figure 5.59), which suggests that DC2 and PC2 correspond to the same motion: an opening and closing of the N-lobe (see figure 5.57).

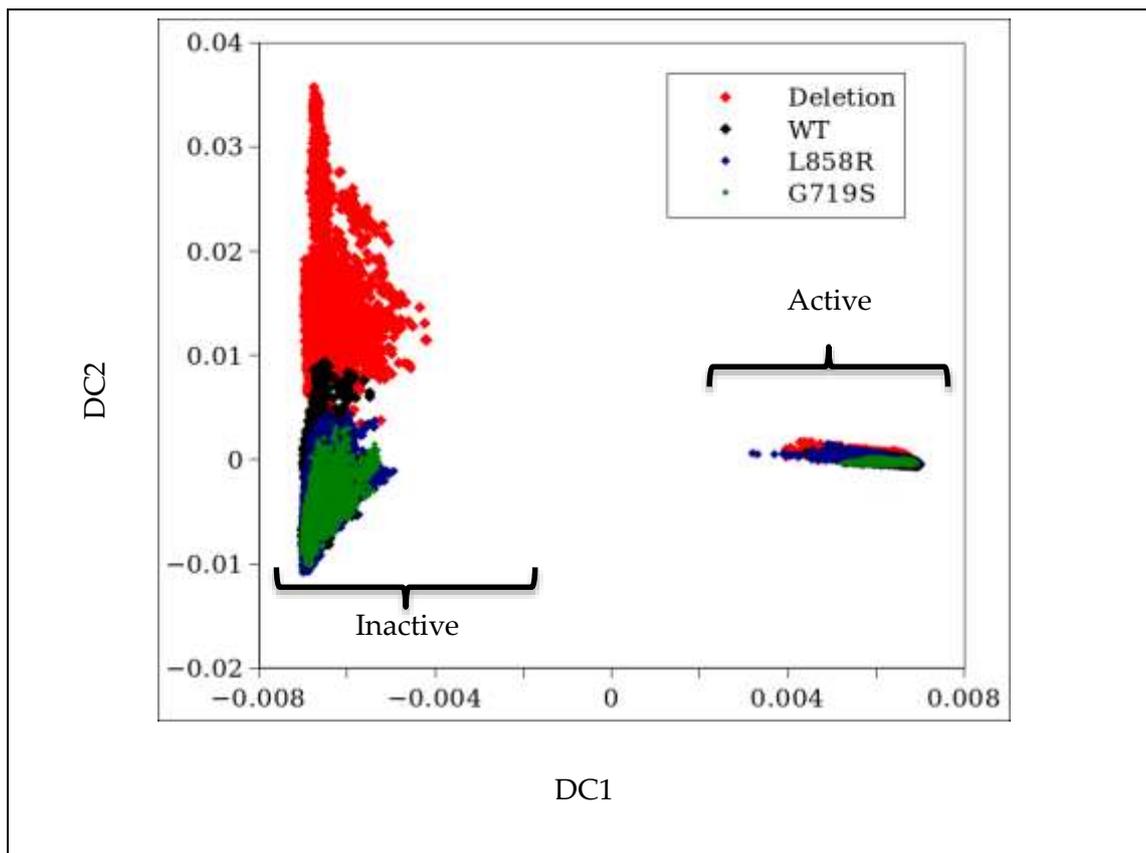


Figure 5.75: RDFMD supertrajectory projected onto the first two DCs of the diffusion map analysis, each mutant and WT is represented by a different colour (see legend).

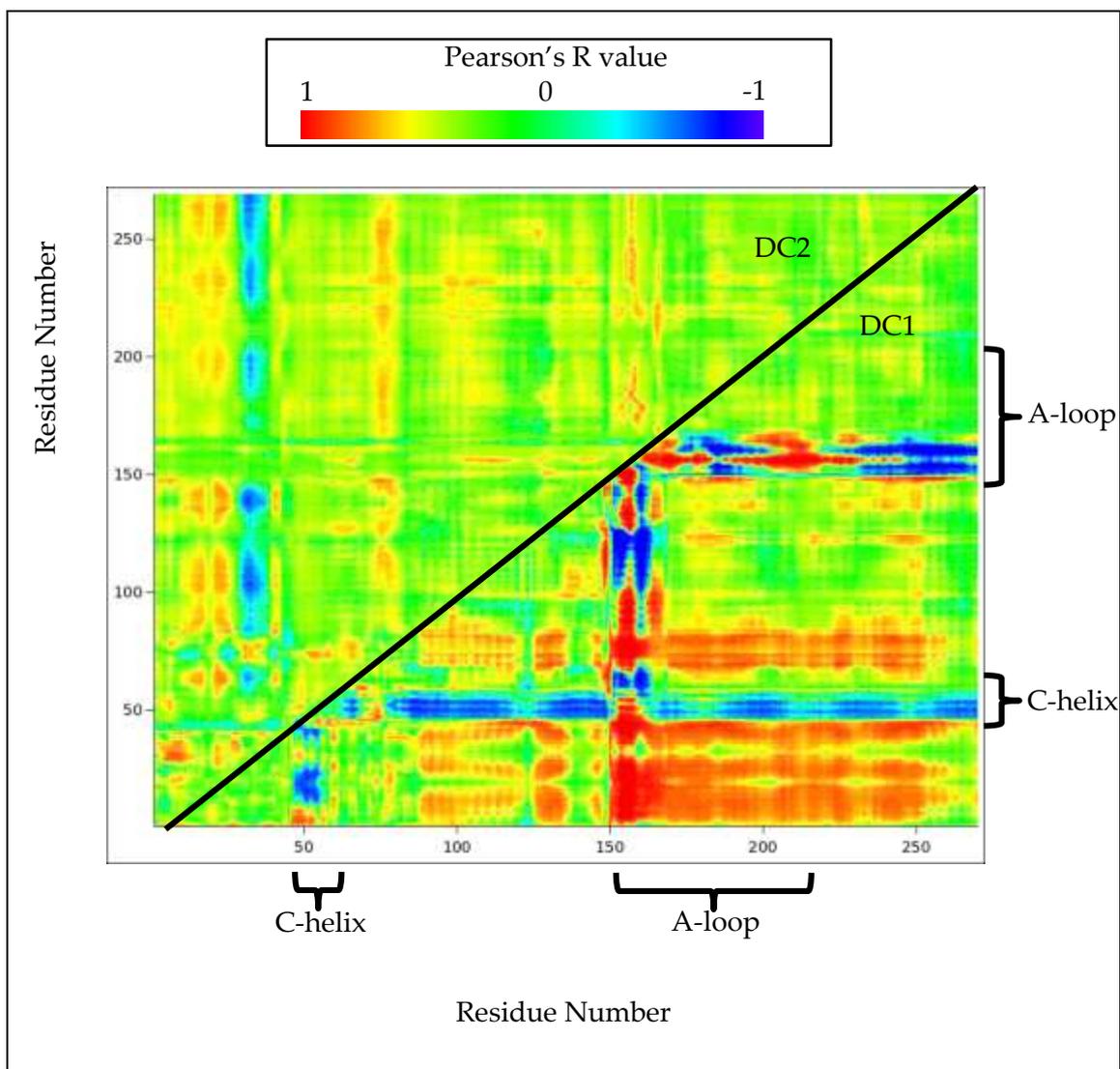


Figure 5.76: Pearson's correlation coefficient of the correlation between inter-residue  $\alpha$ -carbon distances and both DC1 and DC2 of the RDFMD simulation.

## 5.6 Visualisation and other analyses

The volume of data produced is not conducive to analysis by visualisation alone; however, the analyses above provide useful insights into which trajectory frames may be particularly important in gaining atomistic insight into the dynamics of the EGFR kinase. This section will deal with visualisation and other analyses that were guided by

the aforementioned analyses. Visualisations and bond distance analyses were performed using Visual Molecular Dynamics version 1.9.1[125].

### 5.6.1 Sampling of the active trajectories

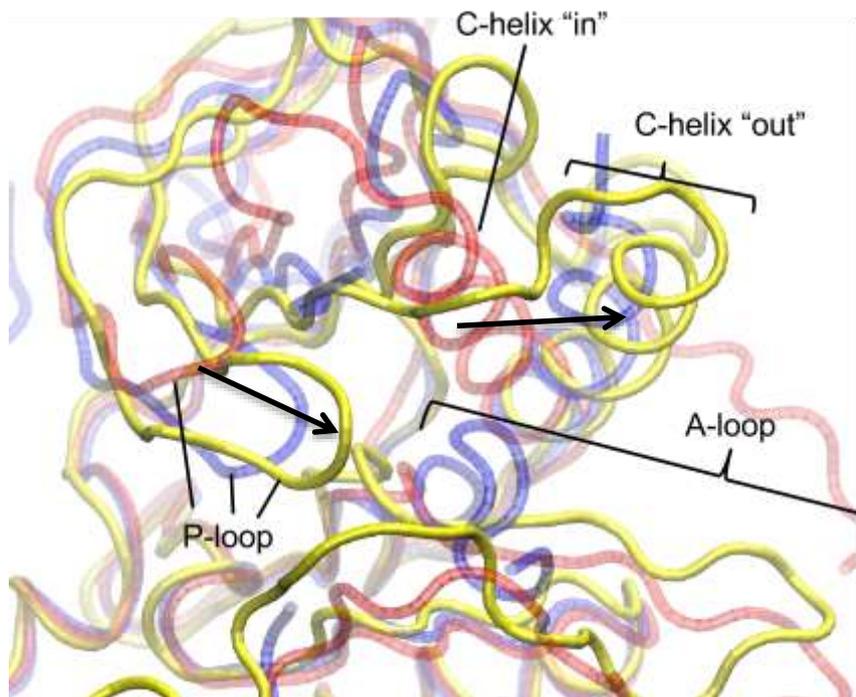


Figure 5.77: Conformation of a random snapshot from the circled region on figure 5.51, from the simulation of the L858R in the active conformation (yellow) superimposed on the WT inactive crystal structure (1xkk; blue), and the WT active crystal structure (1m17; red). The black arrows highlight the extent of the movement of the P-loop and C-helix during the simulation, as compared to the active crystal structure.

Visualisation of many of the active L858R trajectories reveals that the C-helix undergoes a rotation from the active “in” conformation to the inactive “out” conformation (see figure 5.77). This example of the L858R sampling the frontier of accessible PC1 space (see figure 5.51) towards the inactive trajectories is not limited to

the cMD simulations, but also appears in the AMD supertrajectory (see figure 5.54). Indeed, the sampling of C-helix “out” conformations by the L858R has been reported elsewhere [5], [42], [65].

It appears that the G719S mutant is also able to sample the C-helix “out” conformation; however, this is not as frequent as in the L858R mutant, suggesting the G719S has a more subtle impact on the C-helix.

The L858R may have a more potent impact on C-helix mobility due to the ability of R858 to interact with the glutamate residue from the K745-E762 salt bridge (see figure 5.78). While R858 interacts with E762, K745 is usually interacting with the DFG aspartate, D855. The interaction between R858 and E762 is seen in most of the instances where the C-helix of the L858R mutant rotates to the “out” conformation, but it is transient, not present in all the L858R simulations, and the K745-E762 salt bridge may be broken (and reformed) even when R858 is rotated well away from the salt bridge. Thus, while the R858 may play a role in “passing” the E762 away from K745, it also seems likely that a more subtle impact is imparted by the L858R, perhaps by the increased charge introduced into this region, that allows the C-helix to rotate.

It is also worth noting from the PCA (see figure 5.51) that simulations sampling more positive PC1 space also sample more negative PC2 space, which appears to correlate with a torsional motion of the N-lobe (see figure 5.52). Visualisation of the trajectories confirms that the movement of the C-helix into the “out” conformation is accompanied by the rest of the N-lobe rotating, which can be most prominently seen in the P-loop, which appears to follow the C-helix as it makes the transition (see figure 5.77).

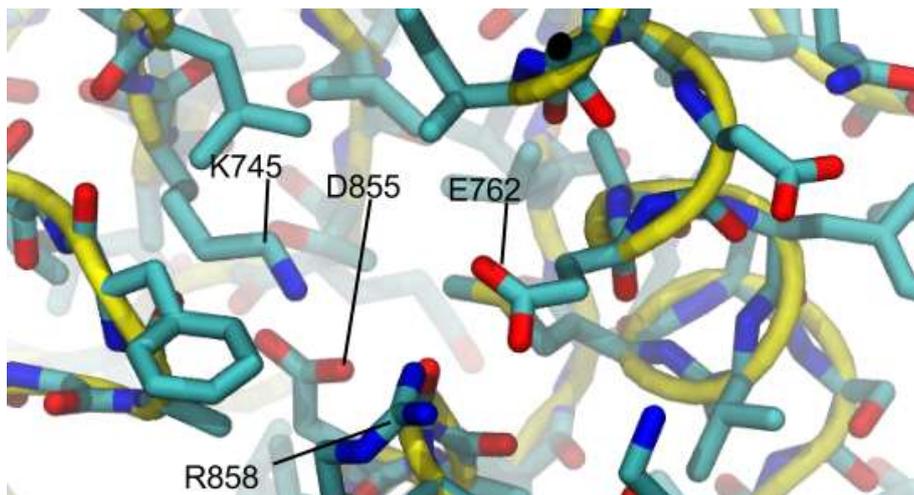


Figure 5.78: Snapshot from the L858R cMD simulation showing the interaction between R858 and E762 of the K745-E762 salt-bridge, while K745 interacts with the DFG aspartate D855.

Shan et al. (2012) propose that the C-helix of the L858R exhibits less disorder, and that the K745-E762 salt bridge is maintained for longer periods, compared to the wild type[43]. In the present study, however, the L858R mutant was found the most likely to lose this salt bridge and adopt a C-helix “out” conformation. Despite this, the L858R C-helix appears relatively stable (i.e. without disorder) (see section 5.4.1), possibly due to the motion being propagated by the rotation of the whole N-lobe, rather than the C-helix acting independently from the N-lobe, which might be expected to incur greater disorder.

It is not clear why the L858R shows such propensity to closure of the N-lobe, however, it is interesting to note that during such a closure the P-loop becomes very close to the mutant R858 residue (see figure 5.79). R858 makes various interactions, which are not consistent across the trajectories; however, no direct interactions appear to be made between R858 and the P-loop in those simulations where a closure occurs, which suggests that R858 has a more subtle effect, perhaps providing an environment more amenable to N-lobe closure.

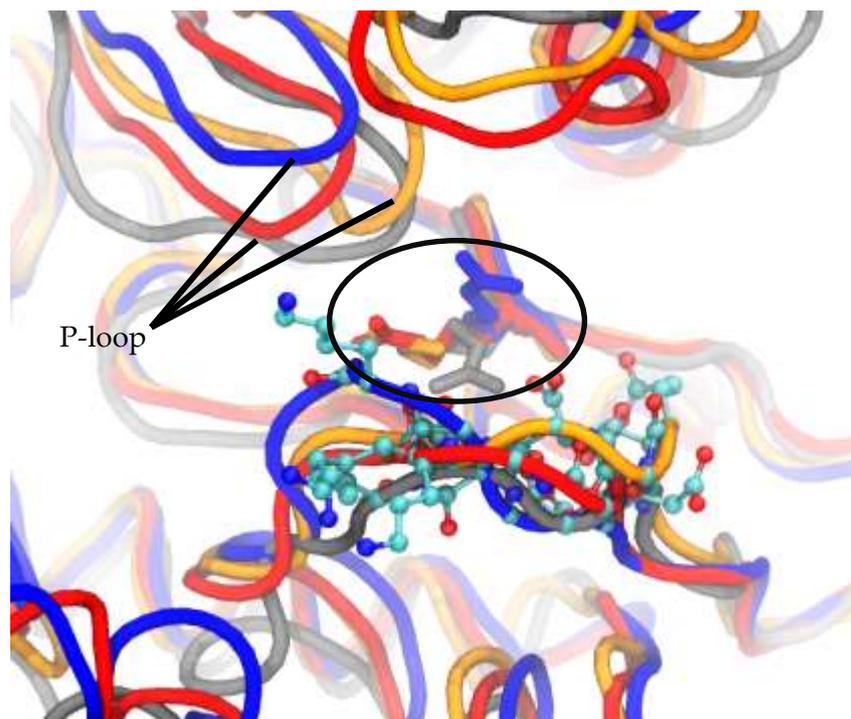


Figure 5.79: R858 (stick representation, circled) as found in a random snapshot. All models are represented, each in a different colour, residues of the A-loop that interact with R858 are shown in ball and stick representation.

The G719S mutant's propensity towards closure of the N-lobe is also difficult to explain, however direct interactions are made between S719 and any of R841, D823 and C797. On the other hand, these interactions appear to be somewhat transient (even in the simulation where it is most prevalent; see figure 5.80), and the real impact of this mutation may be a more subtle perturbation of the environment (as appears to be the case for the L858R). In this case, S719 introduces a polar group into the relatively hydrophobic binding-site side of the P-loop.

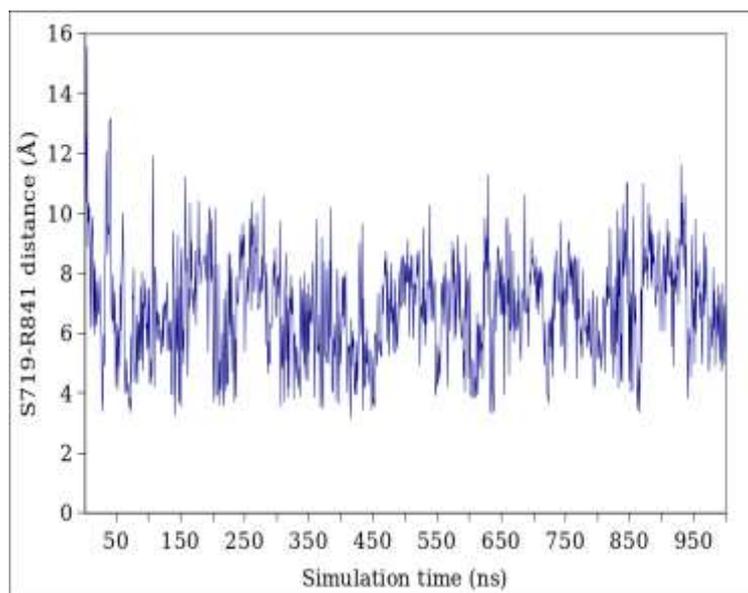


Figure 5.80: Distance between the polar oxygen of S719 and the guanadinium carbon of R841 during the AMD simulation of model 2 of the G719S active conformer.

To summarise, visualisation of the trajectories, guided by the PCA and diffusion maps, has helped to rationalise some of the behaviour of the kinase: Arginine 858 (of the L858R mutant) is in an ideal position to interrupt the important K745-E762 salt bridge, and is close to the P-loop following closure. Additionally, the Serine introduced in the G719S mutation appears to interact with some of the C-lobe residues, albeit transiently. In both instances, this may help to explain why the studied point mutations appear to promote the closure of the N-lobe. However, it is still unclear why these mutants are activating.

### 5.6.2 Sampling of the inactive trajectories

As with the analysis of the active trajectories, the PCA results can be utilised in the search for evidence of conformational change by the selection of inactive simulation snapshots on the frontier of PC1/PC2 sampling space towards the active conformation.

In the supertrajectory comprised of all the cMD simulations, this frontier is most widely explored by the deletion mutant (see figure 5.81). Examination of the trajectory sampling this frontier, shows that the A-loop helix has unwound (see figure 5.82). At the same time, it appears the C-helix has made a small rotation into the space created by the unwinding of this helix, although this rotation is on a slightly different axis to that found in the inactive-active transition.

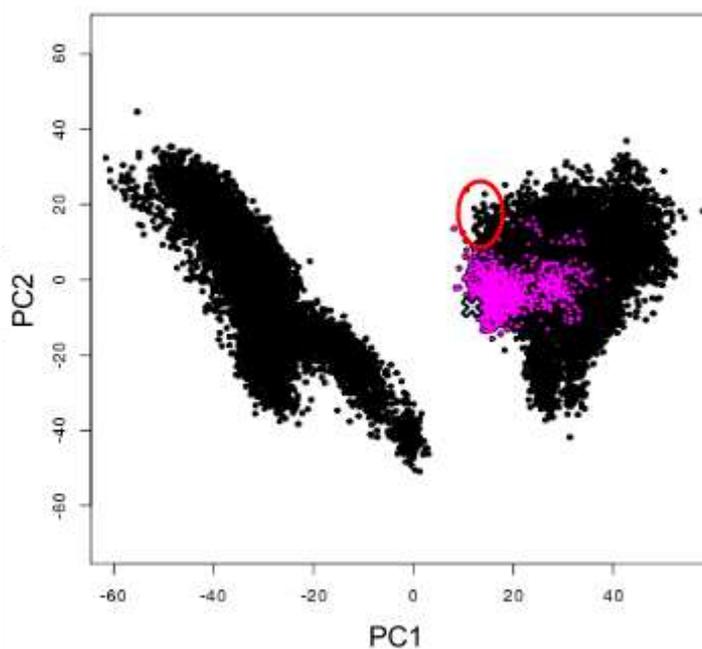


Figure 5.81: Trajectory of deletion model 2 (purple) projected onto PC1 and PC2 of the supertrajectory of all cMD simulations (black). The final frame of the trajectory is marked with a cross. A selection of WT snapshots undergoing skewing of the C-helix is circled in red.

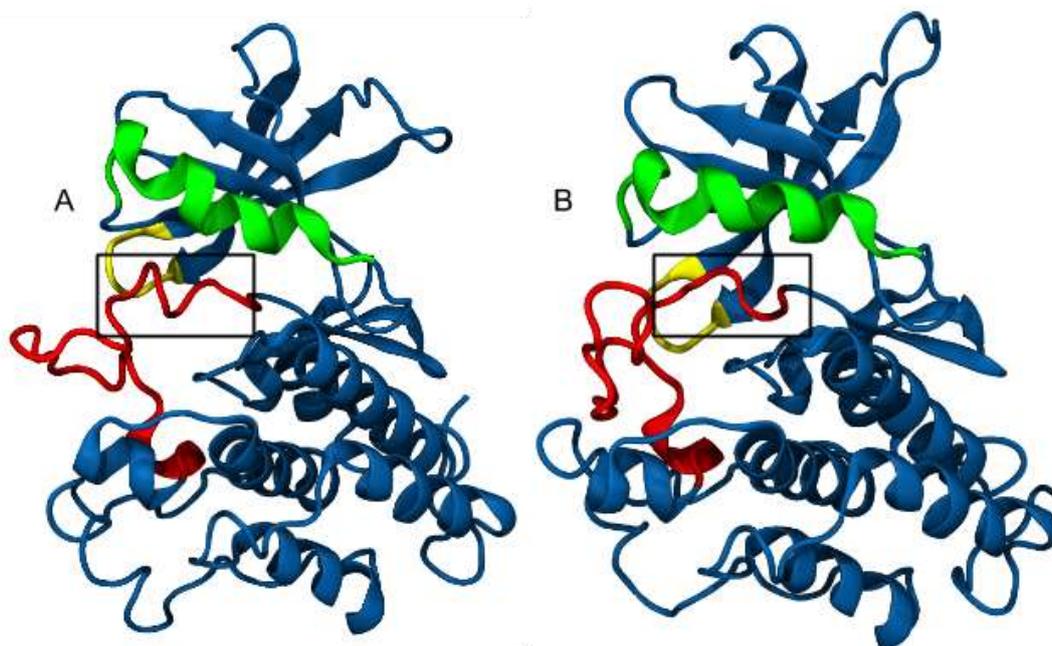


Figure 5.82: Snapshots of the cMD simulation of deletion model 2, including the first frame (A), and the last frame (B), each with the region of the A-loop helix highlighted. The C-helix is shown in green, A-loop in red and P-loop in yellow.

The relevance to PC1 of the unwinding of the A-loop in this manner is unclear, since the WT simulations sample a similar region of PC1 space without exhibiting this unwinding. Instead, the WT simulation's propensity to sampling lower values of PC1 appears to be due to torsion of the C-helix such that just a few residues at the N-terminal of the C-helix rotate towards the "in" conformation (see figure 5.83).

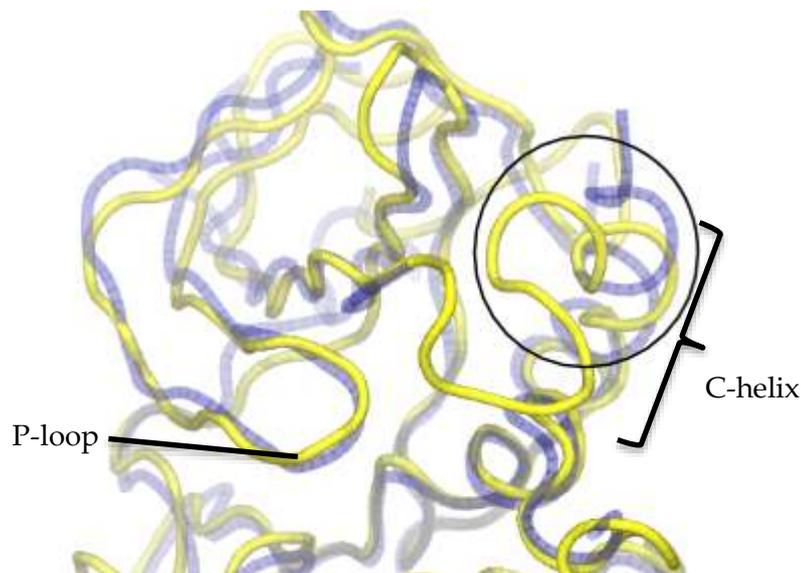


Figure 5.83: Conformation of a random snapshot from the circled region on figure 5.81, from the simulation of the WT inactive conformation (yellow) superimposed on the WT inactive crystal structure (1xkk; blue). The N-terminal side of the C-helix is shown twisted toward the P-loop (circled).

To summarise, although the inactive simulation dataset is punctuated by the intriguing unwinding of the A-loop helix, it appears that this is not necessarily driving the sampling of the inactive trajectories along PC1. Instead, as with the active trajectories, PC1 is dominated by C-helix movements.

## 5.7 Summary of the Molecular Dynamics results

In the present study, it has been demonstrated (as many others have demonstrated) that mutations of EGFR kinase have a considerable impact on the protein's dynamics. However the fact that the literature is contradictory on the role of EGFR mutations makes it clear that analysis should be balanced not only by pre-existing knowledge but by the use of complementary analytical techniques. Indeed, in the present study the

possible existence of a conformational dead-end accessible only to the active L858R trajectories seems convincing from the PCA of the AMD simulations, yet mostly disappears in the corresponding diffusion map by shifting towards the inactive conformation, suggesting that the cluster could be important in a transition pathway.

Despite the above uncertainties, there are a number of observations that are universal across all the performed analyses. Firstly, the point mutations appear to promote the closure of the N-lobe against the C-lobe in the active conformation. Secondly, the deletion appears to act separately from the point mutations by locking the active conformation in place. Thirdly, the L858R appears stable in the inactive conformation. Finally, the deletion appears to destabilise the inactive conformation.

The finding that the point mutations promote the closure of the N-lobe against the C-lobe is consistent with some more recent observations of L858R in the literature[42], [65]. It seems likely that such observations were not made in earlier studies due to their short time scales, as the motions occur over the 10-500 ns time scale, with earlier simulations covering less than 50 ns[62], [63]. The implications of the N-lobe closure is impossible to ascertain from the results alone; however, the study by Sutto et al. (2013) seems to suggest that the closed conformation (at least for the L858R) is much more favourable than the inactive conformation[42]. Additionally, a recent study by Shan et al. (2013)[5] suggests that the conformational transition from the inactive conformation to the active conformation occurs over considerably longer time scales than the reverse, therefore mutations that stabilise the closed conformation may be able to more easily return to the active conformation. Additionally, Shan et al. (2013) suggest that transition to the inactive conformation involves pivoting of the N-lobe around the hinge region such that the C-helix separates from the C-lobe giving the A-loop more space to rearrange. Given the interactions between R858 and the C-helix noted by Sutto et al. (2013)[42], and the propensity for point mutants to form a closed N-lobe, it seems

unlikely that the mutants would perform this manoeuvre as favourably as the WT, and thus less likely to reach the inactive conformation.

There is little literature with which to compare the findings regarding the deletion mutant, but the evidence in the present study consistently points to the orthodox established by Yun et al. (2007)[35], that activation is achieved by destabilising the inactive conformation and stabilising the active conformation directly. It is perhaps worth noting in this context that the deletion exhibits a high magnitude of EGFR autophosphorylation even without EGF stimulation, when compared to the WT[139]; however, the study by Zhang et al. (2006) suggests that even the deletion mutant is inactive as a monomer[29]. Another possibility is that the deletion suppresses disorder of the C-helix, as suggested by Shan et al. (2012)[43]. Indeed, the active simulations of the deletion in the present study have a highly stable C-helix, which may promote formation of the dimer, perhaps even without the presence of EGF[32].

The present study finds little direct evidence in support of the reduction of disorder by mutants as proposed by Shan et al. (2012)[43]; this could be due to the relatively short time scales adopted; however, given that Shan et al. (2012) observe this disorder after approximately 1  $\mu$ s, it seems likely that it should have been observed in the AMD simulations, especially given that each of the WT simulations in Shan et al. (2012) exhibited this disorder. Shan et al. (2012) do not state how they determined secondary structures in their paper, and so another reason for this discrepancy may be the use of a different analysis technique.

To conclude, the exact mechanism by which mutants cause activation remains elusive, but the ability of the point mutations to promote a closed conformation may be a key property in hindering or even completely obstructing the kinase from reaching the inactive conformation. Additionally, although studies to date have tended to support

one theory over others (the notable exception being Sutto et al. (2013)[42]), the present study provides convincing evidence that some mutations have different mechanisms.

## 5.8 Evaluation of sampling methods

In the present study, cMD represents the most orthodox method, and so overlap between the sampling of the enhanced methods and cMD is desirable (in this section, overlap refers to the degree to which the simulations overlap each other on the PC1 v PC2 plots, not to subspace overlap, which is discussed in section 5.9); however, it is also desirable in an enhanced sampling method that extra conformational space is explored. Appendix 3 shows how sampling differs on the first two PCs between simulation methods for mutants and WT in both the active and inactive conformations.

Generally, there is a good overlap between RDFMD and cMD; however RDFMD tends to sample considerably less conformational space overall. This is likely due to the short length of the simulations, as shall be discussed later. Overlap in the sampling of DMDMD and cMD was generally good, with an overlap in sampling easily discernable. Overlap of the AMD and cMD appeared poor in a number of cases; however, the diffuse sampling across PC1 and PC2 space makes analysis difficult. Nonetheless, it remains a point of concern.

It is also interesting to note that the DMDMD and AMD succeed in sampling regions of space inaccessible to the cMD simulations in most cases. RDFMD also appears able to sample conformational space that is inaccessible to the cMD simulations, but this is not the case for all the mutants.

It is concerning that there is little overlap between the DMDMD and AMD sampling; however, this is not a consistent trend across all of the systems, and may be due to the lack of repeats for these methods (2 repeats each for DMDMD and AMD, versus 4 repeats for cMD). Nonetheless, to some extent this is expected due to the possibility that DMDMD is not expected to result in a Boltzmann distribution of conformations[106], unlike AMD, which mimics the Boltzmann distribution[103], though the minima are likely to be less well defined due to the way minima are “filled” (see section 3.8.2).

Clustering was utilised to probe how sampling differed between simulation methods, using the hypothesis that simulations that sample the same space should produce similar clusters. Clustering was performed using the backbone  $\alpha$ -carbons of the supertrajectories of each mutant and wild type in either the active or inactive conformation as input for the *cpptraj* program from the *AmberTools* program suite[114], an epsilon parameter (that determines the minimum distance between clusters) of 3 was used to generate the clusters.

Most of the active supertrajectories produced only 2 or 3 clusters, however the active L858R clustering analysis produced an extensive spectrum of clusters (see figure 5.84), which appears to be indicative of the different sampling previously discussed in terms of the first 2 PCs (see section 5.5.1), with a better overlap of cMD sampling with DMDMD rather than with the AMD. Nonetheless, there is evidently a significant overlap between all the sampling methods, as they all sample in cluster 1 (which accounts for over half the frames in the supertrajectory).

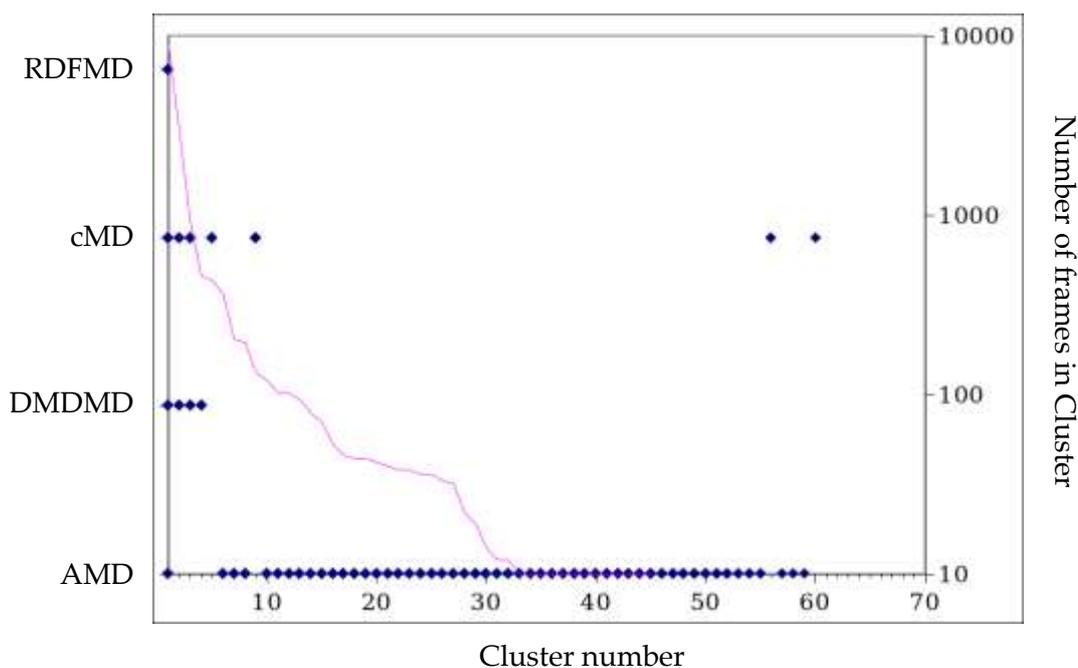


Figure 5.84: Clustering analysis of the active L858R supertrajectory. Blue diamonds indicate whether the sampling method (left Y-axis) populates a given cluster (X-axis). The total number of frames in each cluster is represented by the pink line (right Y-axis; logarithmic).

The inactive supertrajectories produced more clusters than the active, though less than the active L858R supertrajectory. Of the inactive supertrajectories, the deletion produced the greatest number of clusters (see figure 5.85). As with the previous example, there are significant differences between the sampling of the simulations, in this instance there is very little overlap between the cMD and DMDMD simulations. Nonetheless, as before, there is a universal overlap between sampling methods for cluster 1.

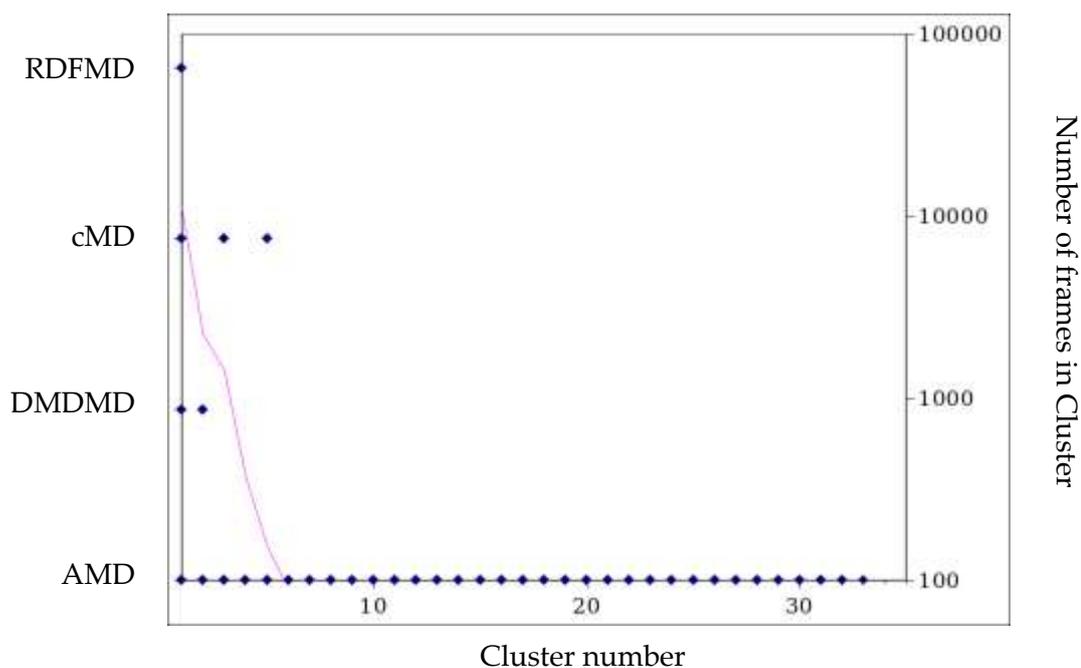


Figure 5.85: Clustering analysis of the inactive deletion supertrajectory. Blue diamonds indicate whether the sampling method (left Y-axis) populates a given cluster (X-axis). The total number of frames in each cluster is represented by the pink line (right Y-axis; logarithmic).

The minimal sampling on PC1 and PC2 by the RDFMD simulations is a feature in half of the active simulations (except the WT and deletion, see appendix 3.1), but not the inactive simulations (except the L858R). This could be due to the fact that RDFMD enhances all low-frequency motions, and that many such motions occur in the kinase besides those represented in the first 2 PCs; however, the amount of PC1/PC2 space sampled by the RDFMD simulations of each mutant seems to be independent of the proportion of variance accounted for by the first two principal components of the PCA of each mutant supertrajectory (see appendix 3). Additionally, the RDFMD simulations tend to occupy just one cluster in the clustering analysis.

It should be noted, however, that the RDFMD simulations each consisted of just 400 ps, giving the total RDFMD simulation time per combination of starting conformation and

mutational state of 12 ns. Despite this, the RDFMD simulations tend to sample an equivalent region of PC1/PC2 space to a 1  $\mu$ s cMD simulation, when combined. Considering this, the PCA and clustering analysis highlights an interesting feature of RDFMD, in that the RDFMD simulations can be shown to sample the important, low frequency motions relatively well considering the simulation times (as is evident from the PCA), but other motions in the protein that might lead to RDFMD simulations exploring multiple clusters are not very well explored. Nonetheless, the fact that cMD and AMD tend to sample a wider range of clusters (and those methods preserve, or at least mimic, the Boltzmann distribution) suggests that results from RDFMD, while much quicker to obtain, should be considered conservatively.

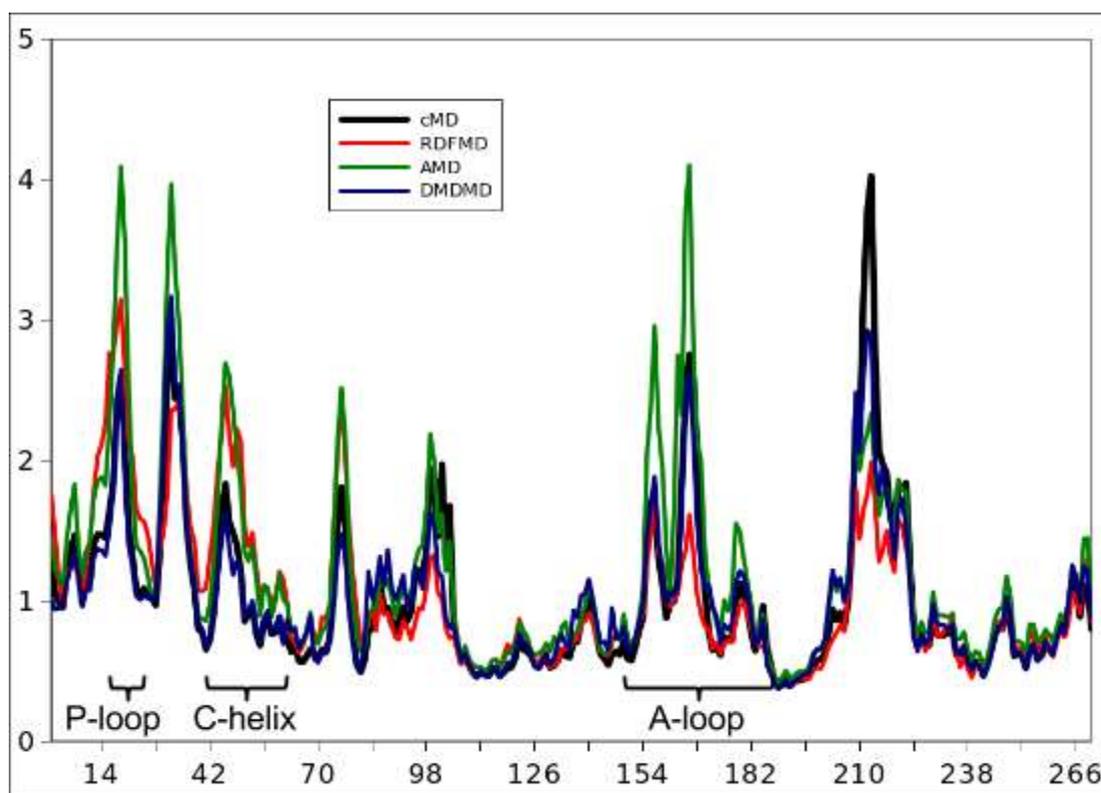


Figure 5.86: Average RMSF for the  $\alpha$ -carbon of each residue of the active deletion mutant for each simulation method.

The backbone  $\alpha$ -carbon RMSFs (see figure 5.86) show that those regions of the protein that are stable in the cMD simulations are also stable in the enhanced sampling methods. The DMDMD and cMD simulations have an almost identical RMSF profile, whereas AMD appears to have the greatest impact. Interestingly, RDFMD appears to have a comparable impact to AMD on the C-helix.

In summary, it appears that all the sampling methods examined in the present study are capable of broadly sampling the same space, but there are significant differences, especially when those motions relating to conformational change are considered (i.e. sampling along the low-ranking PCs). Unfortunately, it is difficult to determine whether these differences are inherent to the sampling methods, or due to the lack of repeats. Nonetheless, in terms of sampling beyond the time limits of cMD simulations, the enhanced sampling methods utilised in the present study all produce some increase in the sampling of the important motions of EGFR kinase, which appears to be most pronounced in the AMD simulations, and considering the short time scales utilised, the RDFMD simulations also show significant promise in their ability to sample important motions of the kinase.

## 5.9 Investigation of the relevance of supertrajectory PCA results

Having performed several PCAs, the question of whether the sampling on the first 2 principal components of the supertrajectories is actually representative of the natural sampling of the protein remains unanswered. This section aims to address this by examining the essential subspace overlap of the supertrajectories with the cMD WT simulations, under the assumption that the sampling of those simulations is

representative of the natural sampling of EGFR kinase. Additionally, the comparability of the supertrajectories of each sampling method will be analysed.

The “essential subspace” is defined as per Amadei et al. (1999) to be the root mean square inner product (RMSIP) of the first N eigenvectors of each principal components analysis:

$$RMSIP = \sqrt{\left(\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N (\eta_i \cdot v_j)^2\right)} \quad (5.1)$$

Where  $\eta_i$  and  $v_j$  are the eigenvectors of each principal component analysis. Amadei et al. (1999) demonstrated that subsequent 50 ps slices of a trajectory of protein L (and cytochrome c 551) have an RMSIP of approximately 0.6 (when N=10)[140]. Given that the analysis of the present study covers much larger time scales, and that data for the analysis was taken at a much larger interval (1 ns), this does not necessarily provide a useful guide for the characteristics of RMSIPs of simulations of the length described here. To investigate this further, the RMSIP of subsequent 200 ns slices of the cMD WT active trajectory were calculated for comparison, and found to average at 0.72.

Additionally, to investigate the expected subspace overlap between repeats of the same system, each WT cMD simulation was compared by RMSIP, as per figure 5.87 and 5.88:

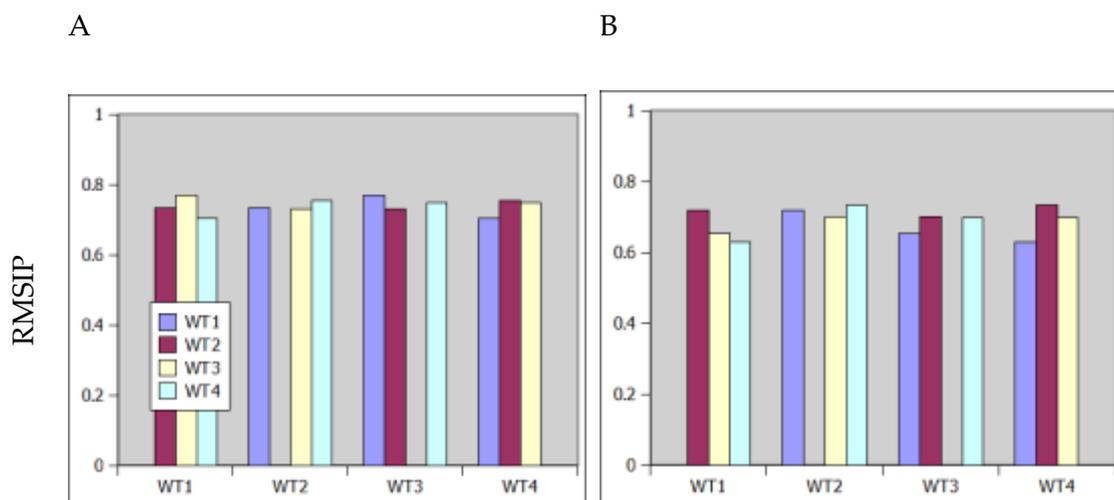


Figure 5.87: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each repeat (1-4) of the active WT cMD simulations.

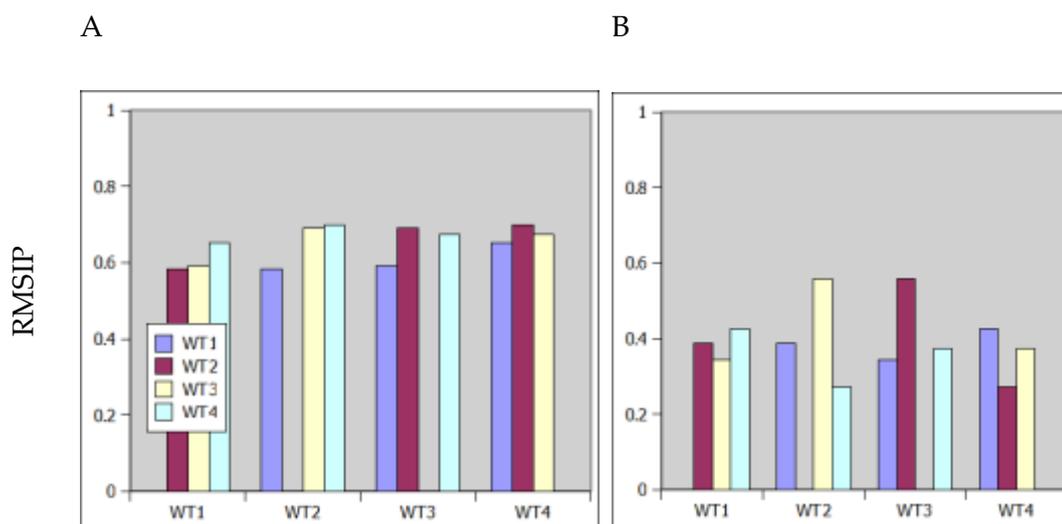


Figure 5.88: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each repeat (1-4) of the inactive WT cMD simulations.

The essential subspace overlap between the WT cMD simulations is between 0.6 and 0.7 for the active WT, with lower values when only 2 PCs are considered (Figure 5.87(B)). The inactive WT shows a poorer overlap, of between around 0.3 to 0.7, with low values particularly common when the RMSIP is composed of the first 2 PCs. The

higher overlap between the active simulations than the inactive simulations, particularly for the first 2 PCs, suggests that the active conformation is more likely to undergo the same motions between repeats. This could be due to the relative stability of the A-loop in the active conformation, with PC1 in particular exhibiting different motions in the inactive conformation between the supertrajectories of the mutants and WT (see figure 5.66 and 5.67).

Having established the subspace overlap to be expected from simulating systems with the same starting conformation, the essential subspace overlap of the principal components was calculated between each of the CMD WT trajectories and each of the WT and mutant supertrajectories. The results are shown in figure 5.89 and 5.90. The aim of this analysis was to evaluate the extent to which the supertrajectories capture natural motions of the protein.

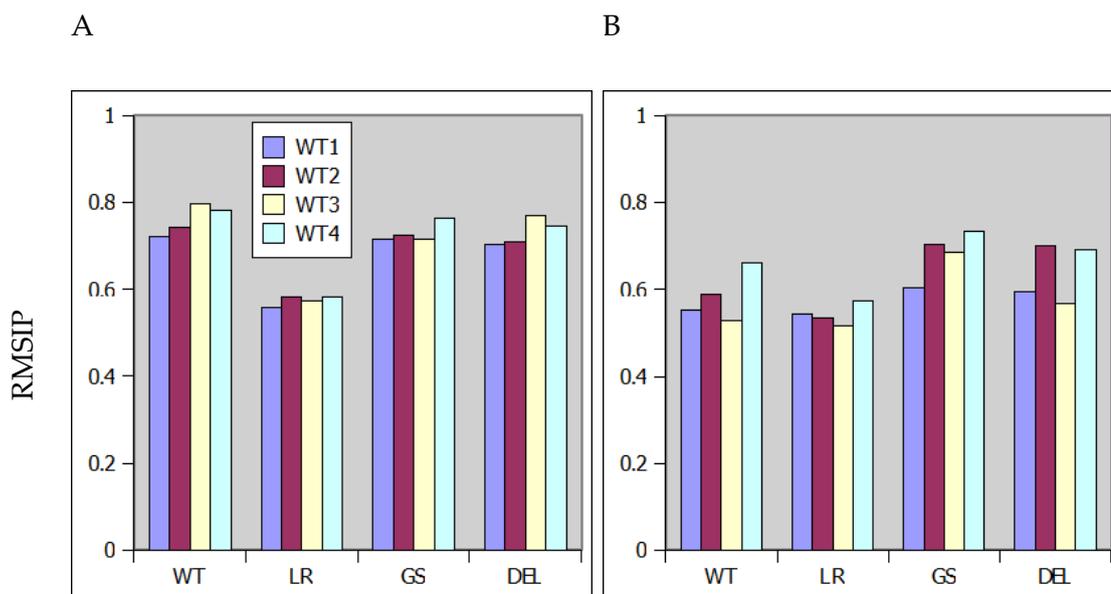


Figure 5.89: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each repeat (1-4) of the active WT cMD simulations and each active WT and mutant supertrajectory.

For the active trajectories, the WT trajectories have a high level of subspace overlap within the first 10 PCs (figure 5.89(A)), when compared against the WT, G719S, and deletion supertrajectories, however, the essential subspace of the L858R mutant has a considerably poorer overlap, possibly due to the L858R undergoing conformational change much more readily than the WT, as briefly discussed in section 5.5.1. However, the increased sampling of the L858R was only analysed with respect to the first 2 PCs, which only appear to have a marginal difference in overlap between the WT trajectories and the L858R supertrajectory and the overlap between the WT trajectories and the WT supertrajectory.

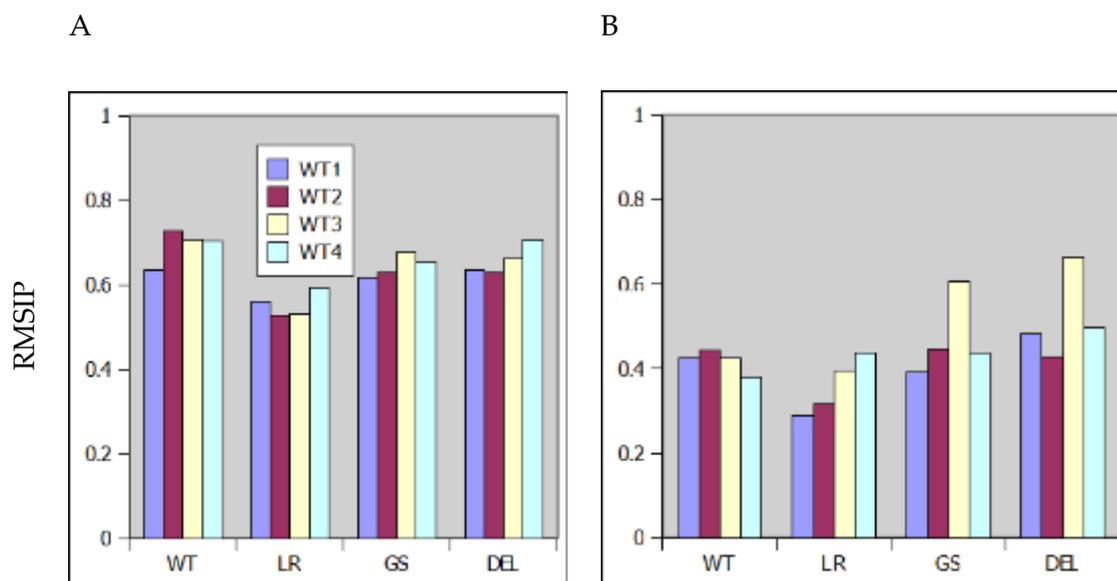


Figure 5.90: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each repeat (1-4) of the inactive WT cMD simulations and each inactive WT and mutant supertrajectory.

For the inactive WT trajectories and the inactive supertrajectories, much poorer subspace overlaps are obtained (see figure 5.90) overall, which may be due to the tendency for the inactive simulations to sample a wider range of A-loop configurations than the active simulations, as discussed previously. Interestingly, the inactive L858R supertrajectory appears to have a lower subspace overlap with the inactive WT trajectories, regardless of whether 10 or 2 PCs are included in the RMSIP calculation. The reason for this is unclear, but it is possible that this is due to the L858R also samples a much wider range of C-terminal tail conformations (see figure 5.66), which are not seen PC1 of the inactive WT (see figure 5.67; also see the appendix figure A3.18).

For the mutant and WT supertrajectories, it appears that a similar degree of subspace overlap is found with the individual WT trajectories (figures 5.89 and 5.90) as would be expected even between repeats of the same system (figures 5.87 and 5.88). Thus, it may be the case that the PCs of these supertrajectories are fairly representative of the

natural sampling of the system. This is especially the case for the active trajectories, where the dot product of the first PC of the WT trajectories and the first PC of the supertrajectories is often 0.6 or more (see appendix 6).

Next, the essential subspace overlap of the principal components was calculated between each of the cMD WT trajectories and the supertrajectories of each sampling method, the results are shown in figures 5.91 and 5.92:

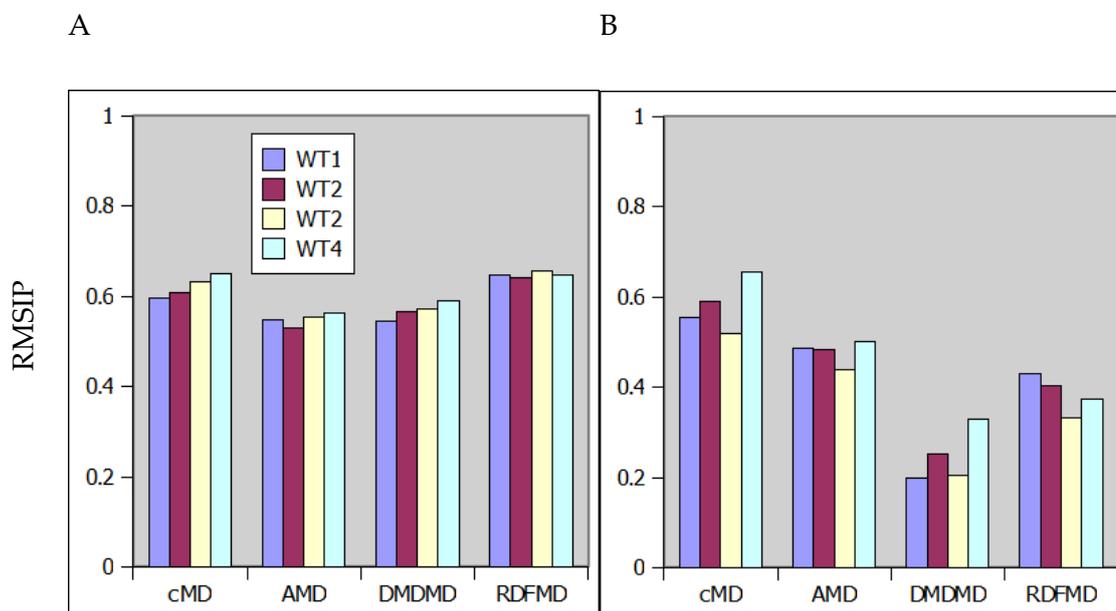


Figure 5.91: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each repeat (1-4) of the active WT cMD simulations and the supertrajectory for each sampling method.

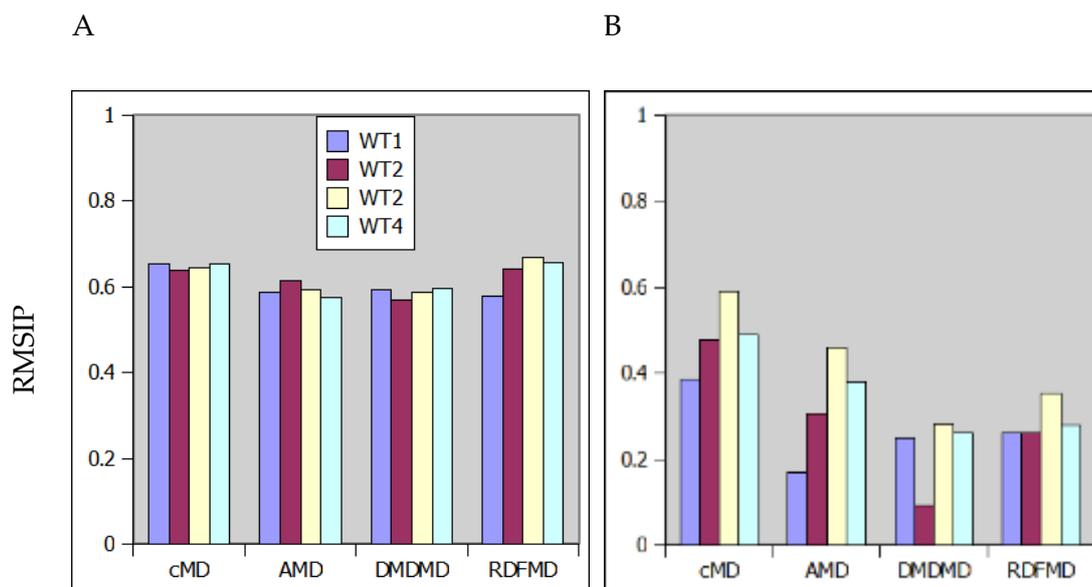


Figure 5.92: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each repeat (1-4) of the inactive WT cMD simulations and the supertrajectory for each sampling method.

A significant loss of overlap (of approximately 0.1, for the first 10 PCs) is seen when the subspace overlap of the WT trajectories with the supertrajectories (figures 5.91 and 5.92) is compared with the overlap between repeats of the WT trajectories (figures 5.87 and 5.88). When the first 2 PCs are considered, this loss is much greater, especially for the enhanced sampling methods. Given that the first PC of each of the sampling method's supertrajectories is the one that separates the component trajectories by conformation, it is expected that the first PC is due not to a natural motion of the protein, but due to the difference between the active and inactive conformations. Consequently, it is unsurprising that there be a significant loss in subspace overlap, particularly when only the first 2 PCs are considered (figures 5.91(B) and 5.92(B)).

Another interesting feature of figures 5.91 and 5.92 is the relatively poor essential subspace overlap of the DMDMD simulation, which may be due to the method's tendency to sample away from minima. It may be the case that this significantly

perturbs the dynamics of the simulation away from the natural dynamics. It may also be the case that the increase in conformational space explored by the method leads to the sampling of motions not captured in the cMD simulations. The slightly reduced overlap of the AMD simulations (especially with respect to the active WT trajectories; figure 5.91(A)) may also be evidence of this.

The investigation of the essential subspace overlap of the supertrajectories with the WT cMD trajectories has provided invaluable detail into the extent to which conclusions based solely on the distribution of PCA supertrajectory data can be made. Generally, supertrajectories constructed from simulations with similar starting configurations have a good subspace overlap in comparison to repeats of the same system, and even in comparison to subsequent 200 ns slices from a single trajectory. However, supertrajectories constructed from simulations with different starting configurations have a considerable loss in subspace overlap with the “natural” dynamics of the system, and most of this loss occurs within the first 2 PCs. Additionally, it is necessary to consider the limitations of the assumption that the WT cMD simulations are representative of the natural dynamics of EGFR kinase, with the relatively small simulation times of the present study being a considerable caveat.

Nonetheless, while an awareness of these issues is essential in considering the data presented, such supertrajectories remain useful in mapping the sampling of simulations of multiple configurations, as was demonstrated in section 5.6.

## 5.9.1 Comparison of sampling methods by subspace overlap

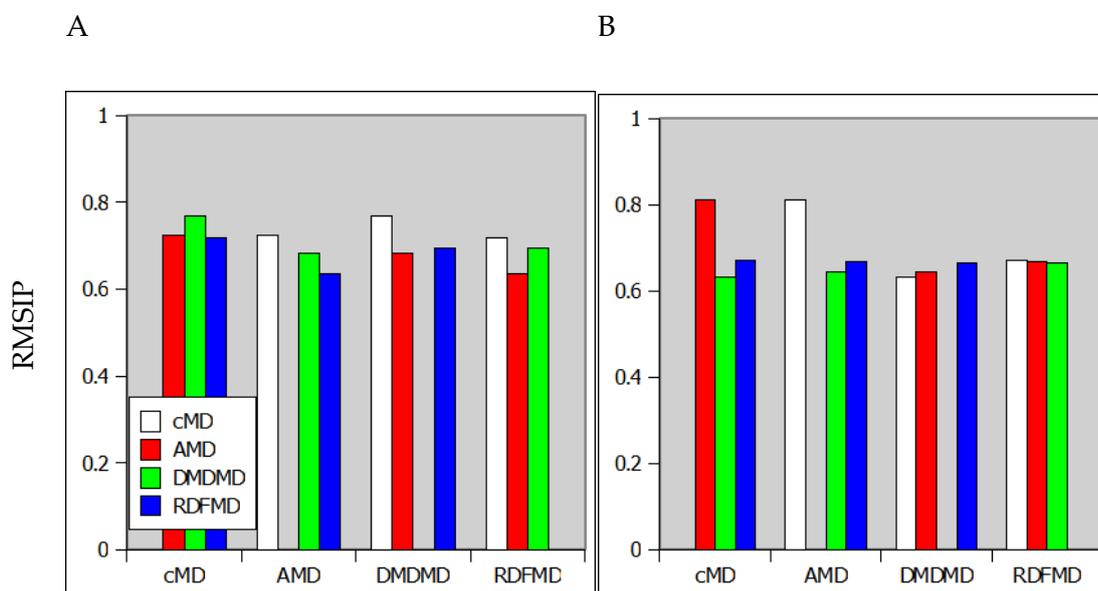


Figure 5.93: Essential subspace overlap of the first 10 (A) or 2 (B) PCA eigenvectors between each of the supertrajectories for each sampling method.

The essential subspace overlap between the sampling methods is generally of a similar magnitude to that expected between repeat simulations of a single system. This is particularly prominent for the cMD and AMD simulations when considering only the first two PCs (the dot products for the first and second PCs were 0.97 and -0.60, respectively), however the other sampling methods have poorer subspace overlaps. For the DMDMD simulations, the poor overlap with cMD may be due to the incorporation of C-terminal motions in PC2 of the DMDMD simulation shown in figure 5.55(A), which is not seen in PC2 of the cMD supertrajectory (the dot products of the first and second PCs with respect to the cMD supertrajectory were 0.88 and -0.17). For the RDFMD simulations the deletion mutant is separated along PC2, possibly due to differences in starting configuration caused by introduction of the deletion into the crystal structures discussed in section 5.5.1 (the dot products of the first and second PCs with respect to the cMD supertrajectory were 0.88 and 0.34). These results suggest

that the sampling of each method is roughly similar over the 10 most important motions of the protein, but over the 2 most important PCs AMD is in the most agreement with cMD, possibly indicating that AMD is a better candidate for realistic exploration of the long time-scale motions.



## Chapter 6: Conclusions

Knowledge of the role of EGFR kinase mutations in cancer has evolved considerably since their discovery, and has continued to present surprises since. The asymmetric dimerization process, the inhibitor binding kinetics of the T790M mutant, and more recently the possibility that activating mutations do not simply act by destabilising the inactive conformation, have all challenged previous understanding of EGFR. Today, the picture of EGFR is still incomplete, and the present study aimed to fill some of the gaps in the existing knowledge by elucidating the conformational dynamics of EGFR kinase mutants, and predict the binding affinity of EGFR inhibitors with a view to later investigating the impact of mutations on inhibitor binding free energies.

Chapter 2 framed present study in the context of its target, EGFR kinase, and relevant computational studies. EGFR kinase is an important target not only for its oncogenic mutations, but also its resistance mutations. Being a protein of considerable pharmacological interest, a wealth of knowledge has been accumulated to work with; this literature forms an evolving schematic of EGFR kinase, with the active and inactive crystal structures, and the impact of mutations and dimerisation in particular guiding the work in the later chapters. While there are some contradictions among the computational findings, these not only highlight possible areas of inquiry, but also the necessity for corroboration.

In chapter 3 the physical basis for the simulations of the present study were laid out. The Boltzmann distribution is particularly important in that it allows for the extraction of quantities of interest from simulations; however, since the Boltzmann distribution serves as the connection between simulated observations and real-world observations, the discussion also serves as a reminder that as computational methods such as the

enhanced sampling methods utilised in the present study, stray from the Boltzmann distribution care should be taken in how those methods are analysed. Indeed, in the original implementation of DMDMD, considerable computational effort was required to regain the Boltzmann distribution[106]. Chapter 3 also included discussion of a number of analysis methods, and as the time scales of computational studies increase, the methods used to analyse them must also adapt. Chapter 2 also bears testament to this, with earlier studies on EGFR kinase dominated by RMSD analyses and observations from visualisation[62], [63], and more recent studies employing PCA, secondary structure analysis, and the analysis of free energy landscapes [42], [43], [65].

The investigation into the prediction of relative binding free energies of inhibitors of EGFR was dealt with in chapter 4. This investigation turned out to be a far more complicated one than originally intended, and highlights a number of the pitfalls in carrying out these kinds of calculation rigorously: misrepresentation of the inhibitor's binding mode, the binding pocket hydration state, and possibly the lack of polarisation are all considerable problems in almost any attempt at binding free energy calculation[141]. It is encouraging to note that the use of a fast QM/MM method for incorporating polarisation effects does not lead to any detriment to the correlation between the calculated properties and the experimental observations; it seems possible that future refinement of the method might help improve the results of such calculations. Nonetheless, while experimental data exists[3] that might be useful in the rigorous characterisation of EGFR kinase resistance mutations, considerable improvement of the calculations in the present study may be necessary before such a goal is achieved.

In Chapter 5 the conformational dynamics of EGFR kinase and mutants thereof were probed with 4 sampling methods and several analytical techniques. The principal findings were that the point mutants are not stable in the active conformation, and that they appear to enhance an opening/closing motion of the N-lobe of the EGFR kinase

monomer. Additionally it was found that the deletion mutant appears to stabilise the active conformation, and destabilise the inactive conformation. These findings were primarily evident from the PCA; however, the diffusion map analysis also provides extra insight into the nature of clustering of the trajectories in the PCA. Furthermore, it is interesting to note that many of the observations of the PCA are evident in the simpler RMSD analyses, such as the increased stability of the active conformation of the deletion mutant relative to the WT; it seems these simple analyses will continue to be of use even as our repertoire of analysis methods expands.

The enhanced sampling methods employed were all capable, to some degree, of increasing the range of sampling compared to cMD. In this respect, AMD appears the most promising of the enhanced sampling methods, providing the greatest amount of sampling along those collective variables corresponding to the difference between the active and inactive conformations. However, it is important to note that the lack of consistency between the sampling of the trajectories across each of the mutants and the WT as determined by PCA. Whether this is due to some deficiency in the enhanced sampling methods, or the lack of repeats is not clear.

Despite the recent multi  $\mu$ s simulations of EGFR kinase[5], [43], the picture of EGFR's conformational dynamics still appears incomplete, and in places contradictory. Indeed, the present study finds little evidence to support the C-helix disorder observed by Shan et al. (2012)[43], and like Wan et al. (2011), contradicts the notion that the L858R is beneficial to the active conformation[65]. It would appear that further investigation is required before the role of EGFR mutation is sufficiently mapped out; however, pushing the boundaries of our understanding of EGFR kinase forward appears to be increasingly dependent on long time scale simulations. It seems likely then, that the majority of future work on EGFR kinase will involve enhanced sampling. Thus it is all the more important that suitable analytical techniques exist to validate such sampling methods.

The present study set out to probe the effect of mutations on the conformational dynamics of EGFR kinase, and on the binding affinity of its inhibitors. It could be argued that, while some interesting observations have been made, more has been learned about the challenges involved in the investigation process than the subject of the investigation itself. Nonetheless, by framing these problems overcoming them will hopefully become a simpler task.

## 6.1 Future work

A number of different approaches remain for future work into the rigorous free energy calculations. Further refinement of the QM/MM method is possible, and an evaluation of possible deficiencies of the QM/MM method as applied in the present study could be carried out. Over polarisation is a common problem, for example, and the use of smeared charges rather than point charges has been found to reduce over polarisation where MM point charges approach the QM system[142]. Additionally, with an ever-increasing repertoire of crystal structures available, there is much scope for investigating the impact of different starting configurations, which has not been thoroughly explored in the present study. The investigations of Balius & Rizzo (2009) [6] raise the interesting possibility of the T790M altering the hydration state of the pocket. The use of the more rigorous GCMC and/or JAWS methods to investigate possible changes in hydration patterns between the mutants could shed further light into this issue. Finally, of course, the secondary goal of rigorous prediction of the impact of mutations on binding affinity remains open, but still requires the foundation that inhibitor binding affinity prediction would provide.

With respect to understanding the conformational dynamics of EGFR kinase in greater detail, the inactive to active transition (or vice versa), and how mutations impact this transition, is still in greatest need of elucidation; however, it still appears that this transition is beyond the limits of accessible sampling of the techniques described in the present study.

Nonetheless, the investigation into the conformational dynamics of EGFR kinase is rich in opportunities for expansion. Firstly, the robustness of the results of the present study could be bolstered by an increase in the number of repeats. The necessity of a greater number of repeats was particularly clear in the AMD and DMDMD simulations, where only 2 simulations for each system were used, and often found to sample very differently. An increase in the number of repeats would allow for a much clearer evaluation of the methods, while at the same time increasing the amount of conformational space sampled: hopefully revealing even more details of the dynamics of the kinase. Additionally, while considerable effort was expended in parameterising the AMD and DMDMD enhanced sampling methods to ensure their validity, additional experiments into parameterisation to evaluate their impact on the degree to which different parameters increase the range of sampling of the most important motions, as well as how optimal parameterisation may differ between systems (which is also absent from the literature), would be useful in the wider application of these techniques. The combination of DMDMD with other enhanced sampling techniques (including AMD) is another interesting possibility; while the resulting distribution would bear the burden of the limitations of both techniques, it may then be possible regain the Boltzmann distribution using the original DMDMD protocol[106]. Of course a variety of other sampling methods exist, and coarse graining is a technique that has been used little in the context of EGFR.

The simulation of mutant dimers is still decidedly lacking from the literature, and represents another possible avenue of investigation, although one that requires

significantly more computational resources, and would benefit from the above investigations into enhanced sampling methods. The long time scale simulation of full-length EGFR including the extracellular region is beyond the resources available to most laboratories, but is proving to be enlightening where its implementation has been realised[32], and as shown in the present study, in some cases the sampling of mutants diverges from the WT after a relatively short amount of time, and rather than running exceptionally long time scale simulations, more effort might be applied to increasing the number of repeats, and the depth of the analysis while attempting to elucidate the conformational dynamics of EGFR kinase.

## Appendices

### Appendix 1: Secondary structure analysis

The secondary structure of the A-loop over the course of the active DMDMD trajectories was essentially the same as those of the cMD trajectories (for comparison, see section 5.4.1 figure 5.43).

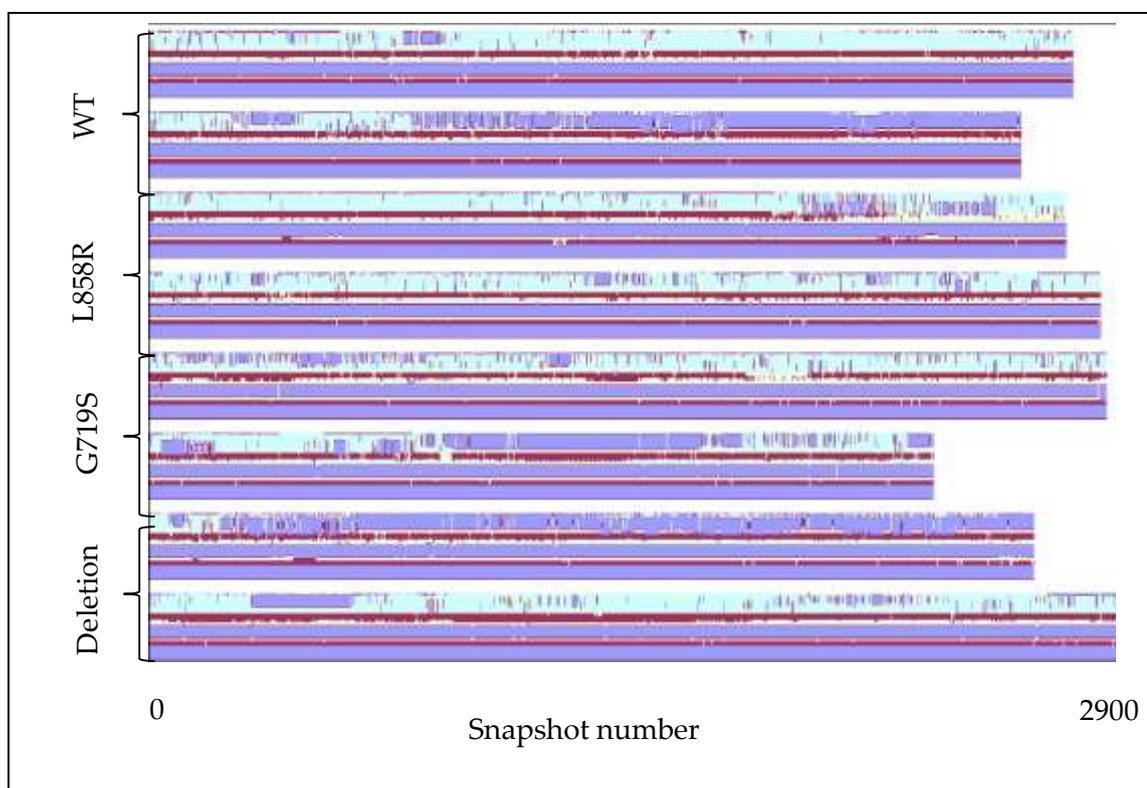


Figure A1.1: STRIDE secondary structure analysis of the A-loop over the course of the active DMDMD trajectories. Purple indicates the presence of a helix, blue represents a turn, and pale blue an isolated bridge structure.

## **Appendix 2: RMSF analysis of RDFMD trajectories**

The RMSF was not calculated for the RDFMD simulations in the same manner as the other sampling methods for brevity: the number of RDFMD simulations totals 240, and individual treatment of each simulation is difficult to analyse and visualise effectively by RMSF.

The RMSF taken over a supertrajectory compiled over all repeats was performed; however, it is not directly comparable to the RMSFs presented in section 5.3, which represent RMSFs derived from raw data. Nonetheless, for completeness, and to demonstrate that RDFMD simulations are subject to the subtle effects of mutation, this analysis is included here.

## Impact of mutations on RDFMD simulations

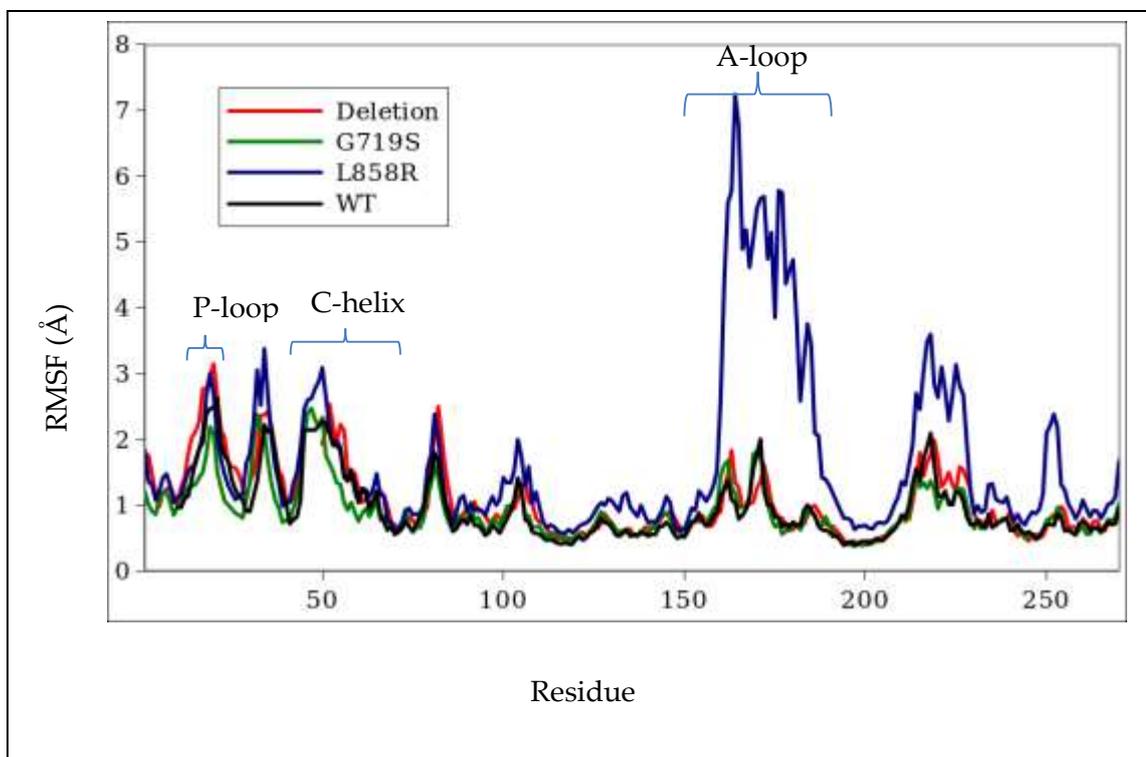


Figure A2.1: RMSF of each residue's  $\alpha$ -carbon over each active RDFMD simulation of the deletion (red), G719S (green), L858R (blue), and WT (black).

From figure A2.1, it appears that the RDFMD simulations do not capture the reduced motions of the active C-helix due to the deletion mutations, while there is some indication that the increased movements in the active C-helix due to the L858R mutation are captured, as are the increased RMSFs in stable regions of the protein. The most prominent feature is the increased L858R A-loop RMSF. With an RMSF of 7 Å, this is similar to the A-loop peak observed in the AMD simulations (see chapter 5.3, figure 5.26).

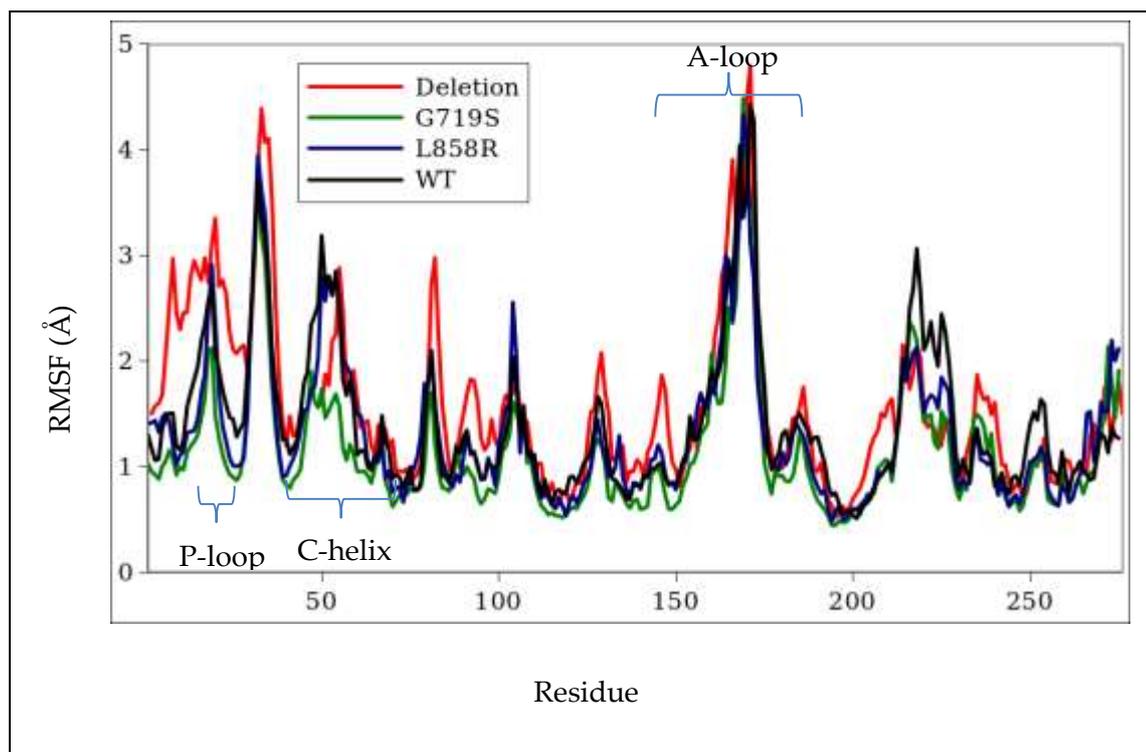


Figure A2.2: RMSF of each residue's  $\alpha$ -carbon over each inactive RDFMD simulation of the deletion (red), G719S (green), L858R (blue), and WT (black).

Figure A2.2 shows the RMSF of the combined inactive trajectories of each mutant and WT system. There is little deviation from the WT, as was also the case in the cMD simulations (compare with chapter 5.3, figure 5.30), however, the deletion does appear to be considerably destabilised, with stable regions surrounding the P-loop being disrupted to a greater extent than is seen in any of the other sampling methods (see chapter 5.3).

### Impact of different RDFMD targets on sampling

The RMSFs for the RDFMD experiments (see figure A2.3) suggest that the A-loop is somewhat restricted in the inactive conformation, as attempts to specifically boost the sampling of the A-loop were universally unimpressive compared to the boosts applied to the C-helix. This may be due in part to the relatively high flexibility of the A-loop, or

is possibly an indication that RDFMD is not suitable for longer stretches of amino acids.

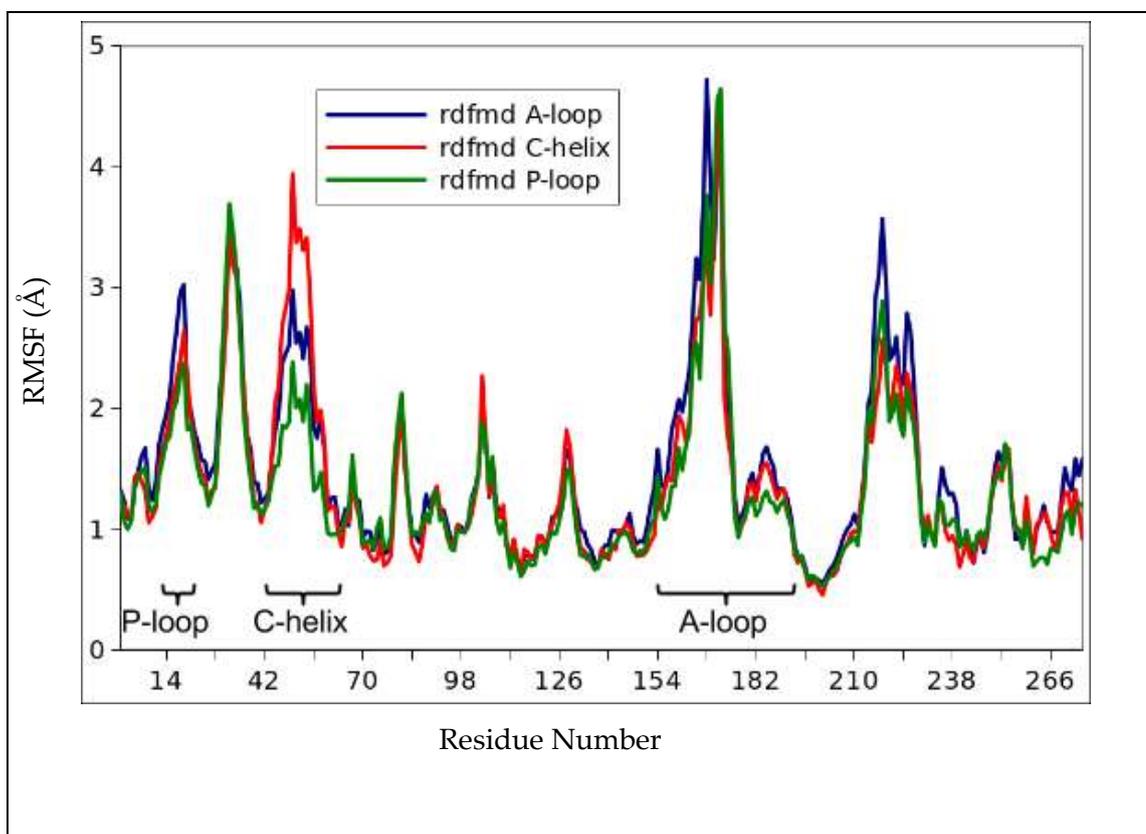


Figure A2.3: RMSF for the  $\alpha$ -carbon of each residue of the inactive WT for each RDFMD target region.

## Appendix 3: PCA of each mutant and WT

Here separate PCA plots for each mutant and WT in each of the active and inactive conformations are presented for all those plots not presented in the main text. The supertrajectories were constructed using simulations including all sampling methods. In each plot “pcax” and “pcay” correspond to PC1 and PC2, respectively. Additionally, the proportion of variance captured by each PC (ranked 1-20) is included, as well as representations of PC2 mapped onto the Cartesian coordinates.

### PCA of active conformations of each mutant

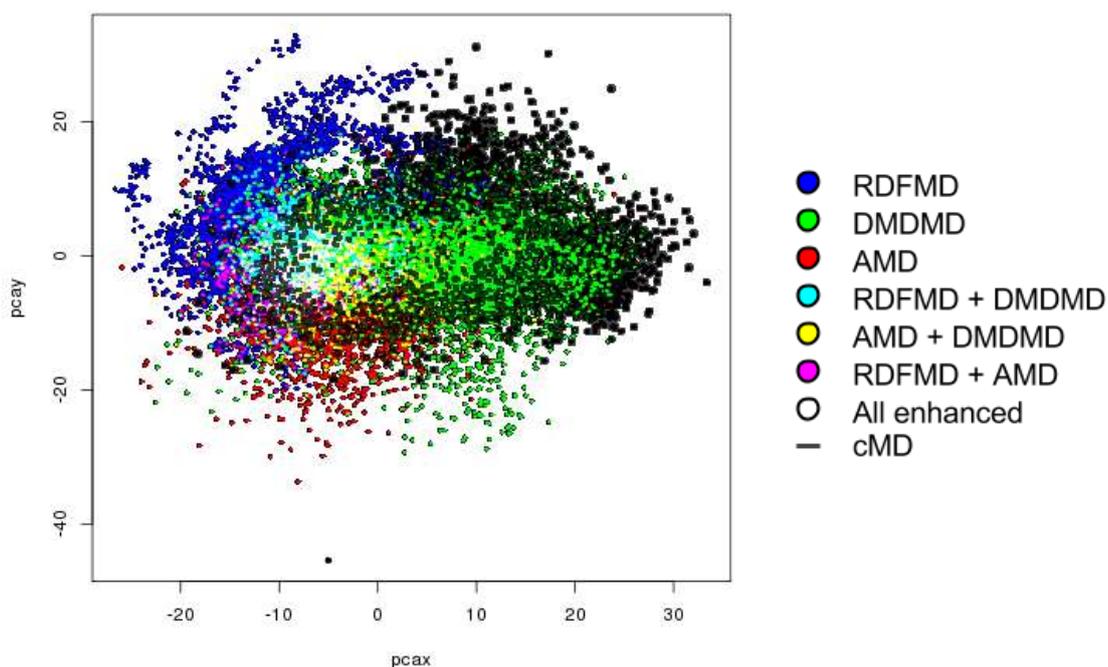


Figure A3.1: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all active deletion simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a dark grey outline.

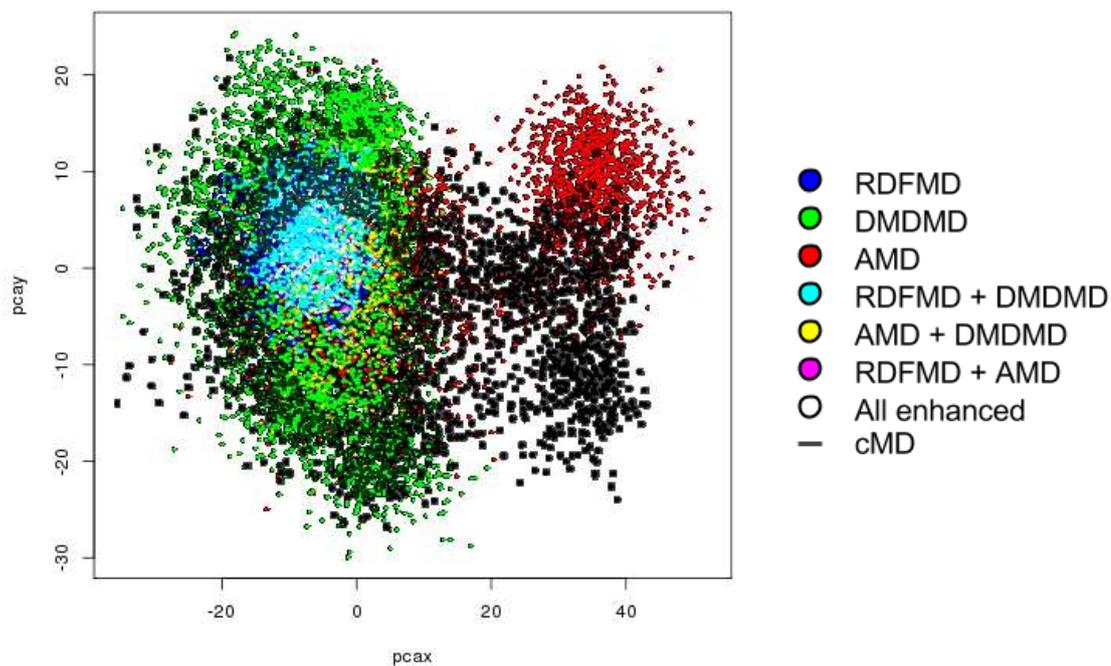


Figure A3.2: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all active G719S simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a dark grey outline.

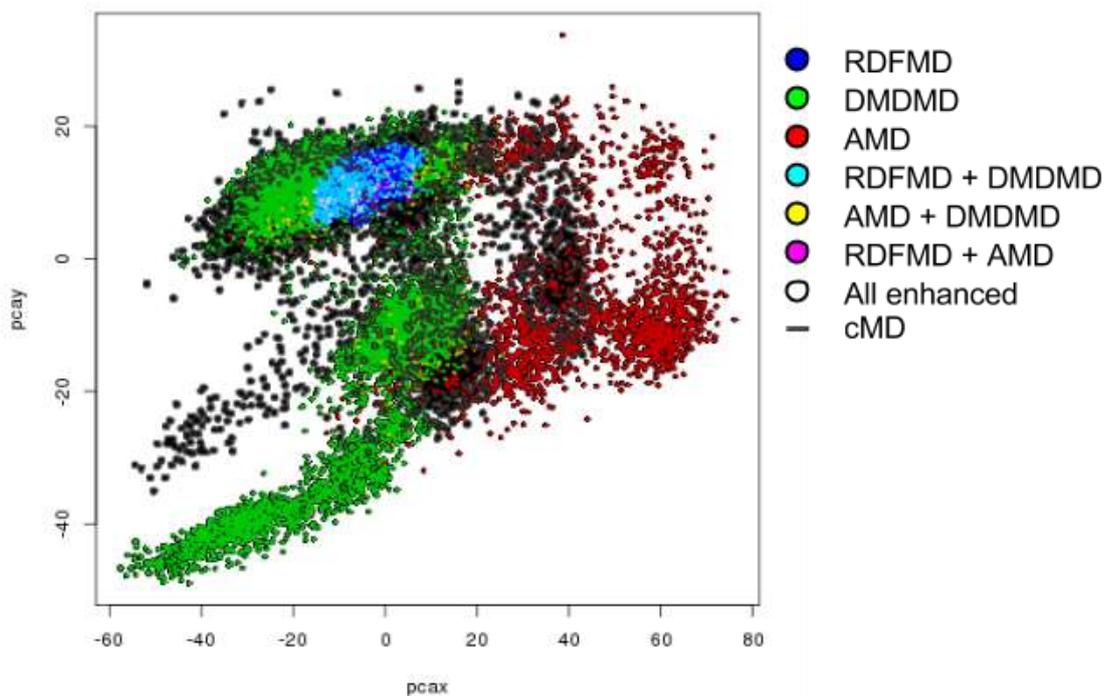


Figure A3.3: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all active L858R simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a dark grey outline.

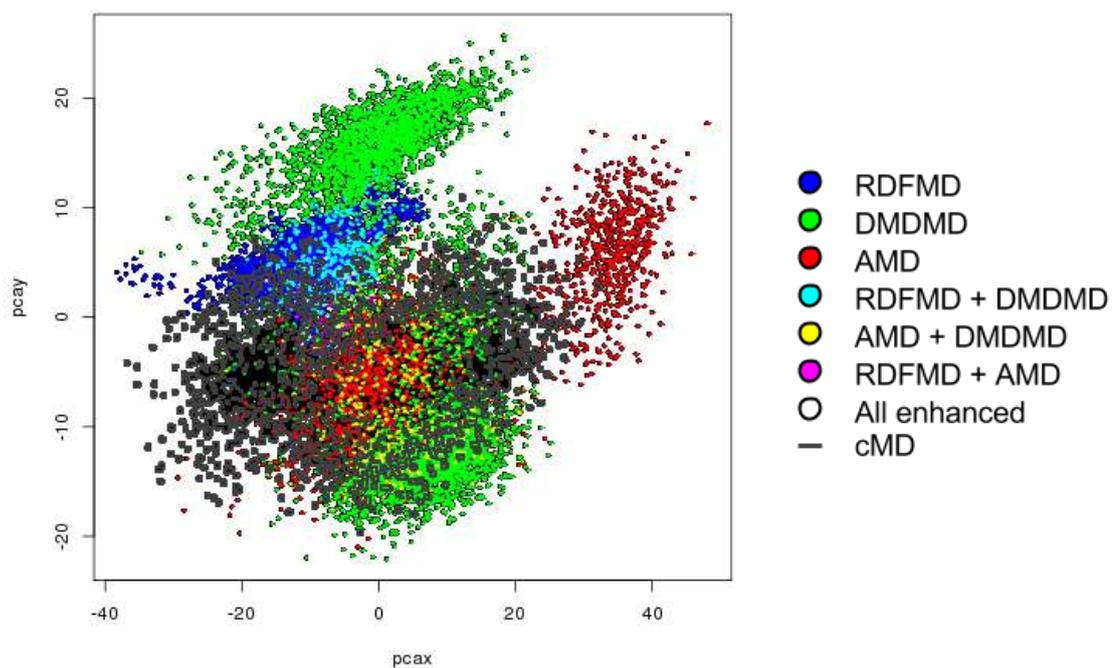


Figure A3.4: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all active WT simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a grey outline.

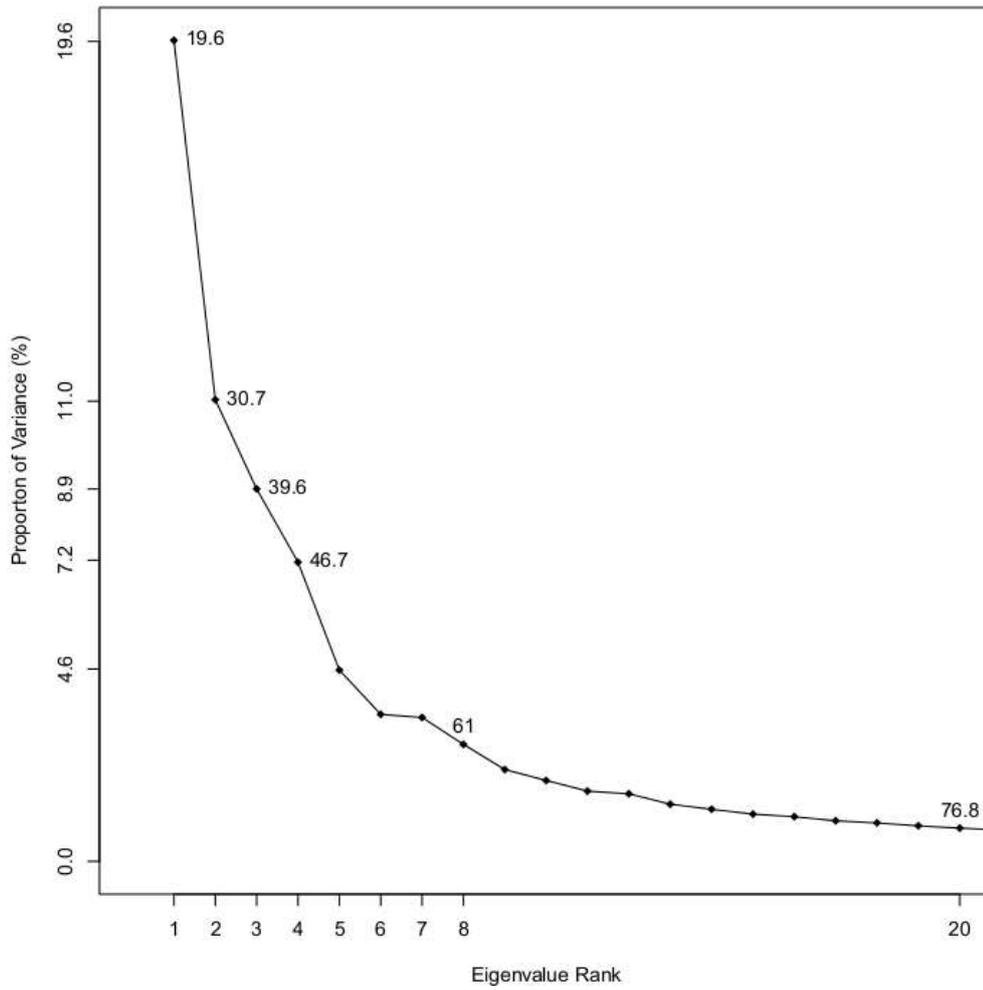


Figure A3.5: Scree plot of the proportion of variance accounted for by each principal component of the active deletion supertrajectory

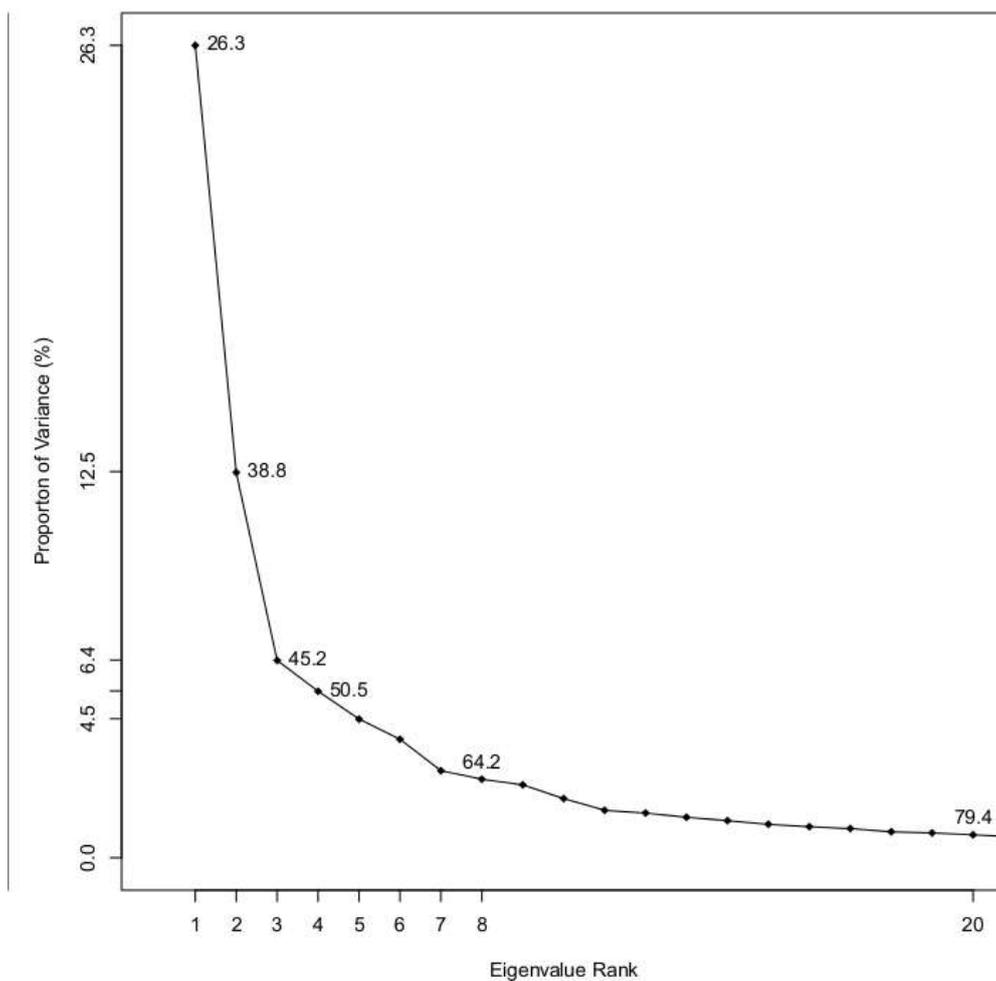


Figure A3.6: Scree plot of the proportion of variance accounted for by each principal component of the active G719S supertrajectory

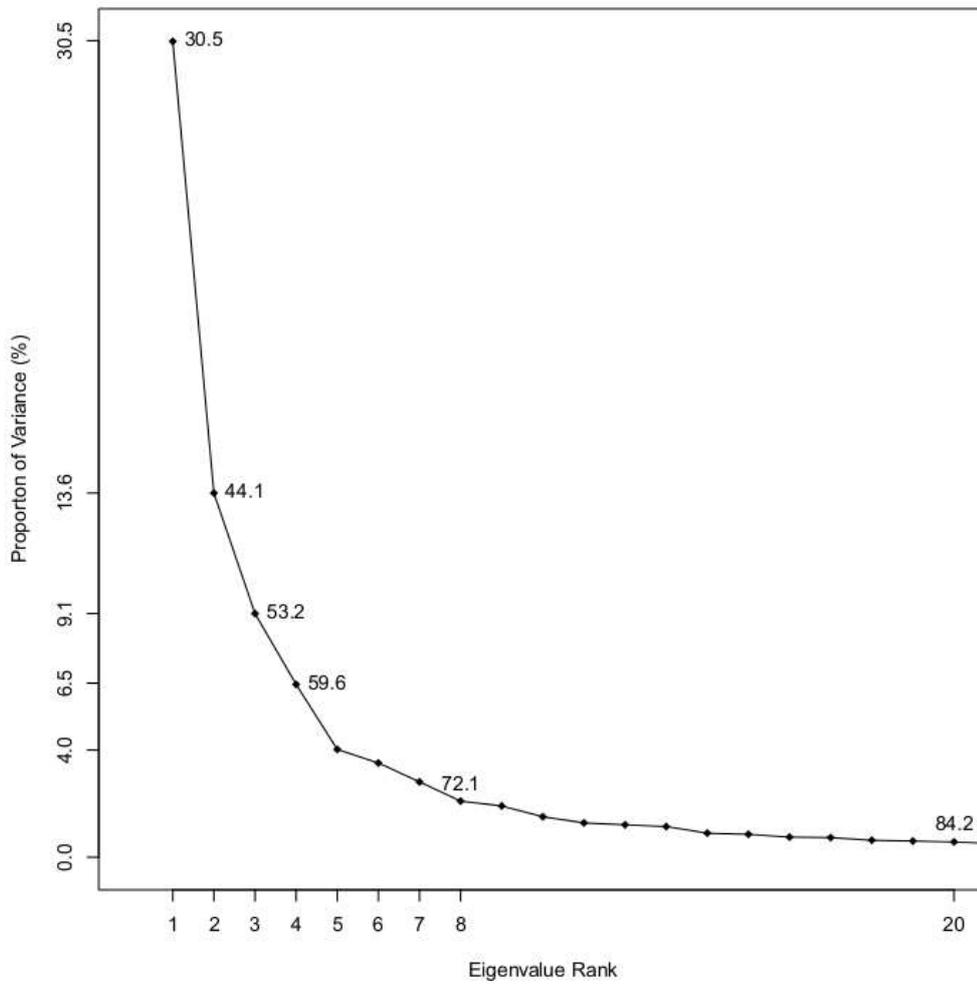


Figure A3.7: Scree plot of the proportion of variance accounted for by each principal component of the active L858R supertrajectory

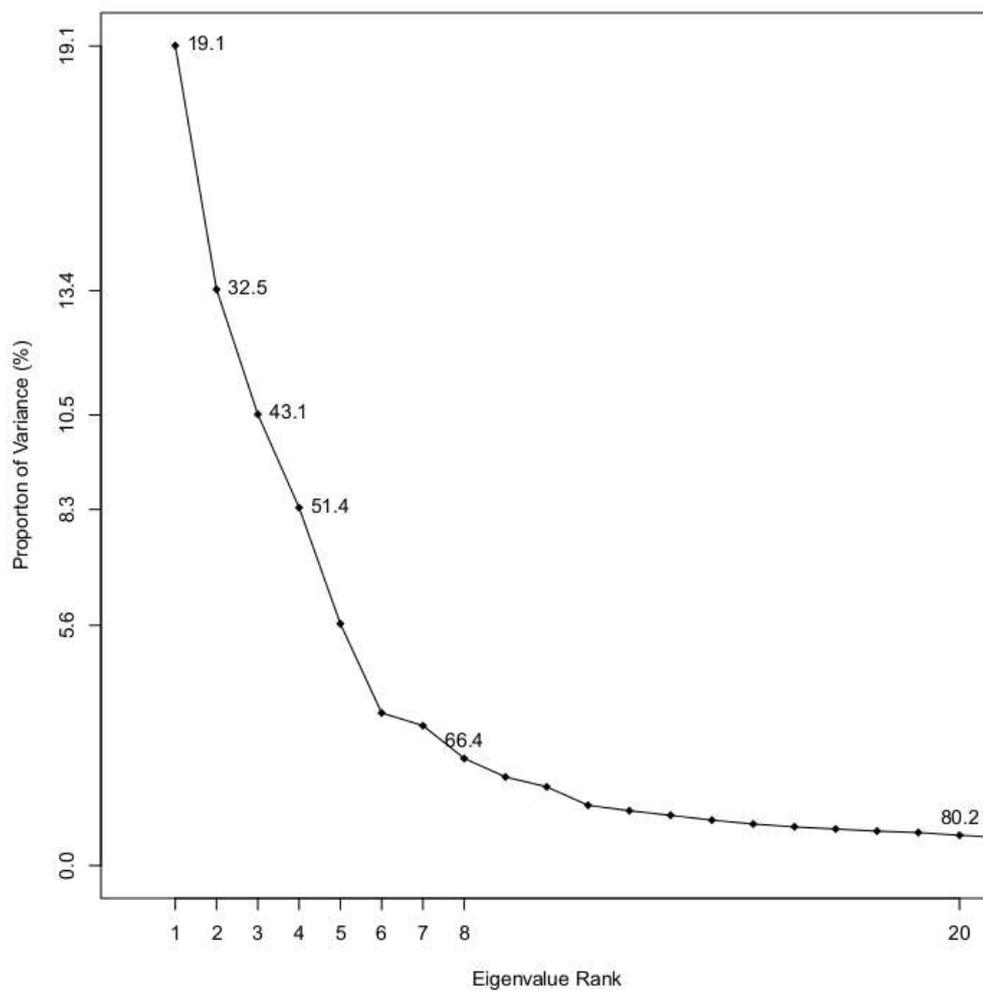


Figure A3.8: Scree plot of the proportion of variance accounted for by each principal component of the active WT supertrajectory

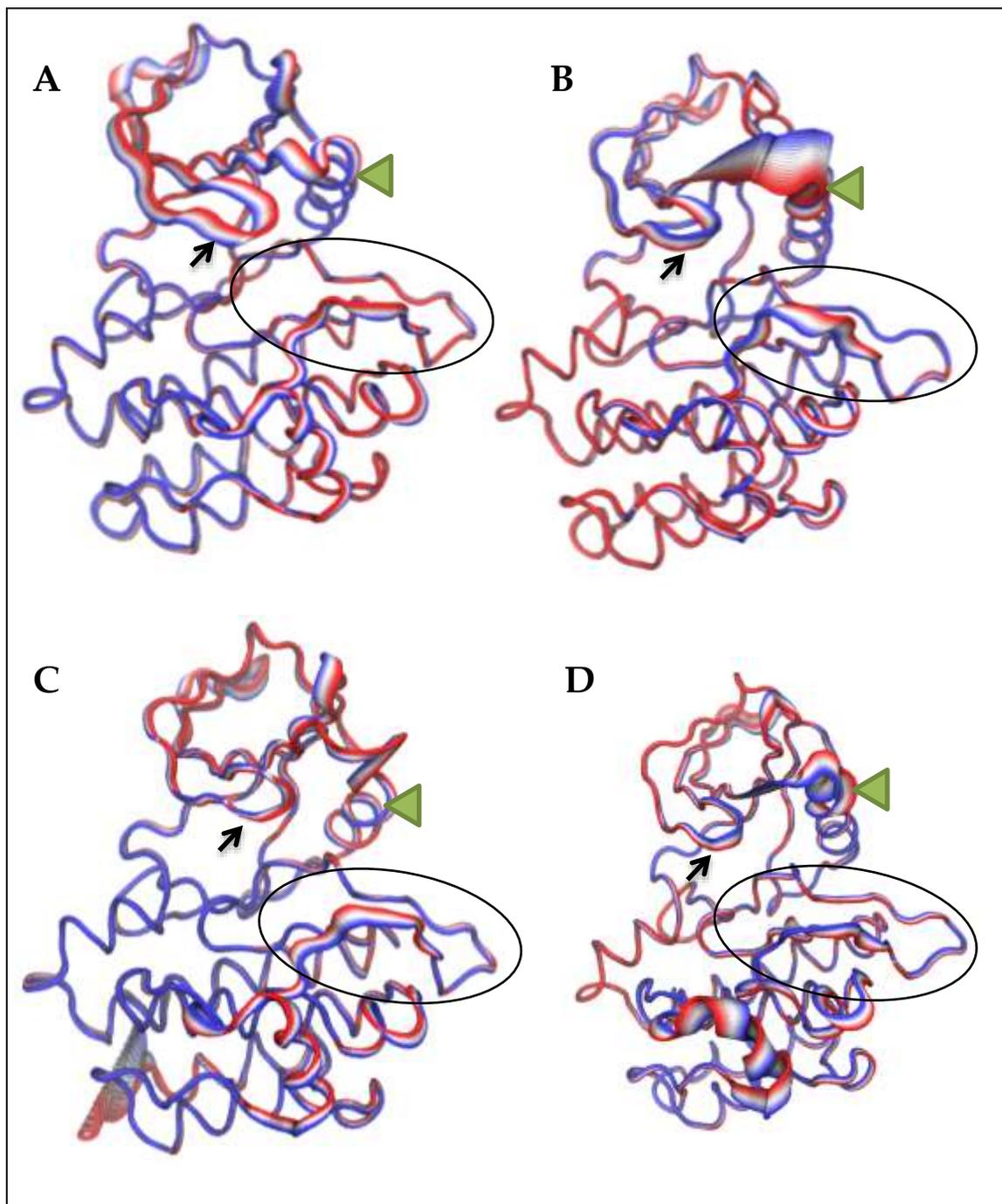


Figure A3.9: Representations of the backbone atomic displacements captured in PC2 of the active simulations of the deletion mutant (A), G719S mutant (B), L858R mutant (C), and WT (D). The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

## PCA of inactive conformations of each mutant

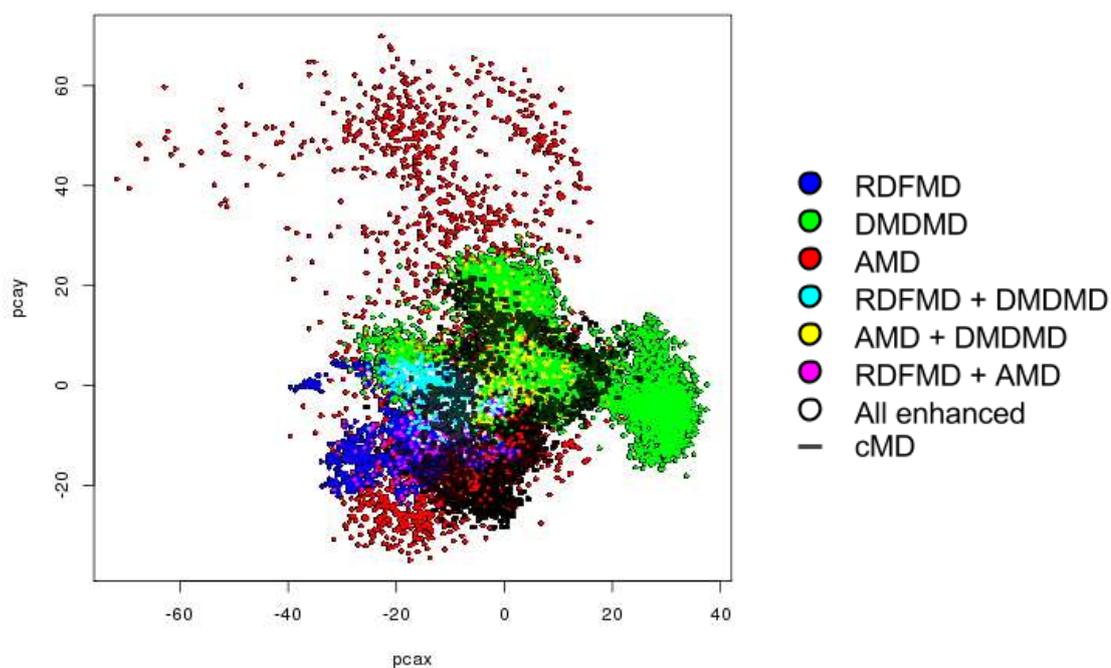


Figure A3.10: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all inactive deletion simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a grey outline.

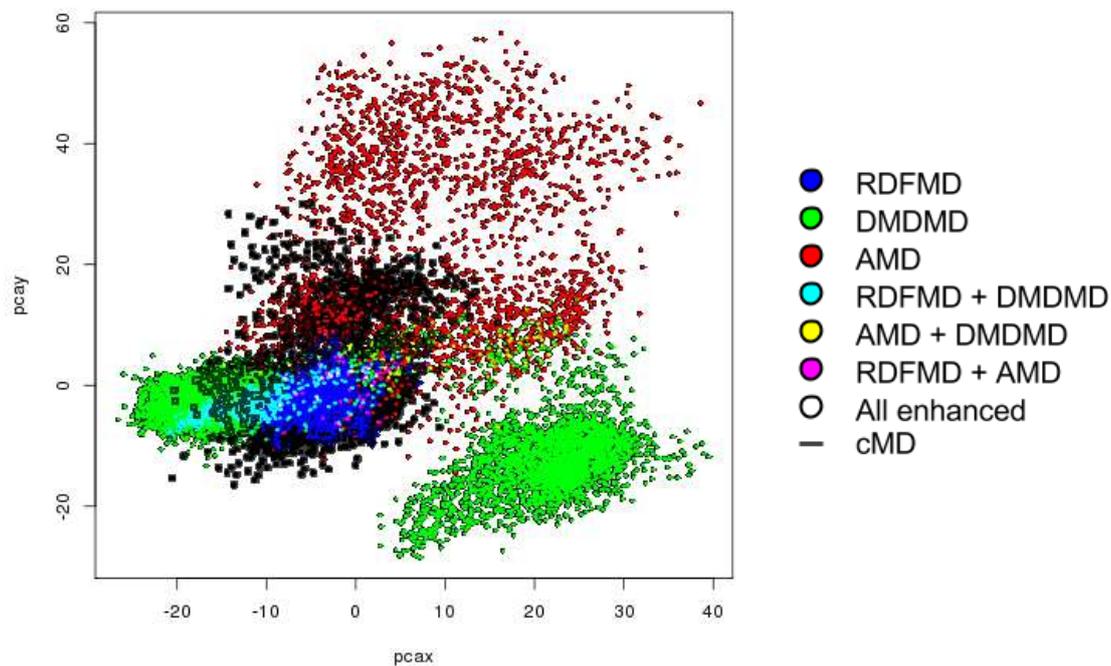


Figure A3.11: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all inactive G719S simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a grey outline.

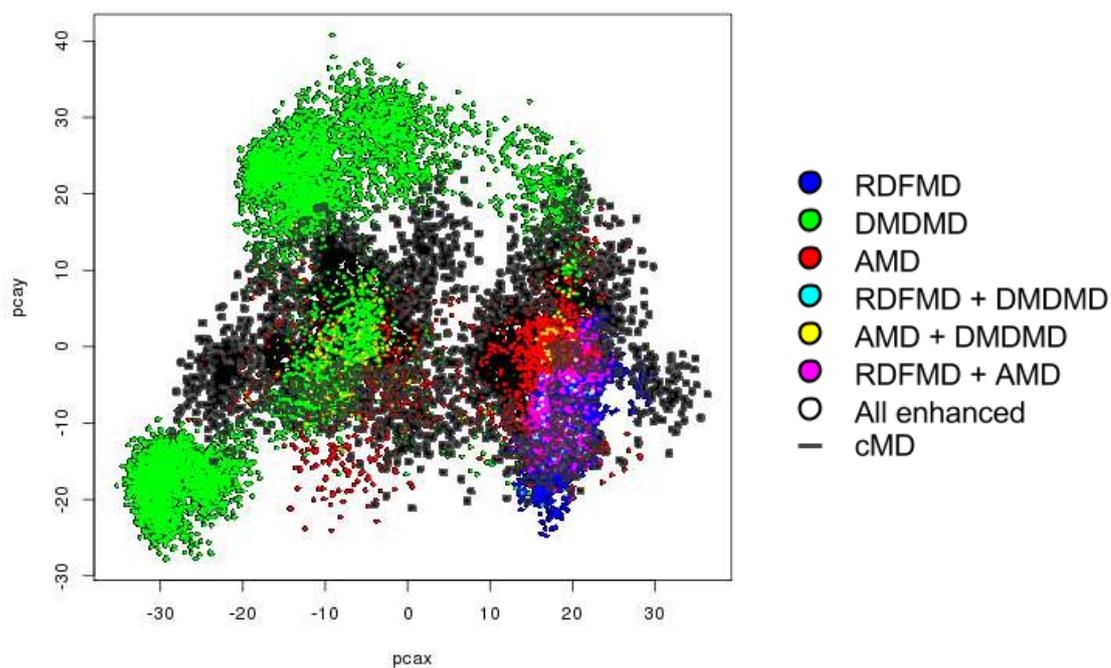


Figure A3.12: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all inactive L858R simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a grey outline.

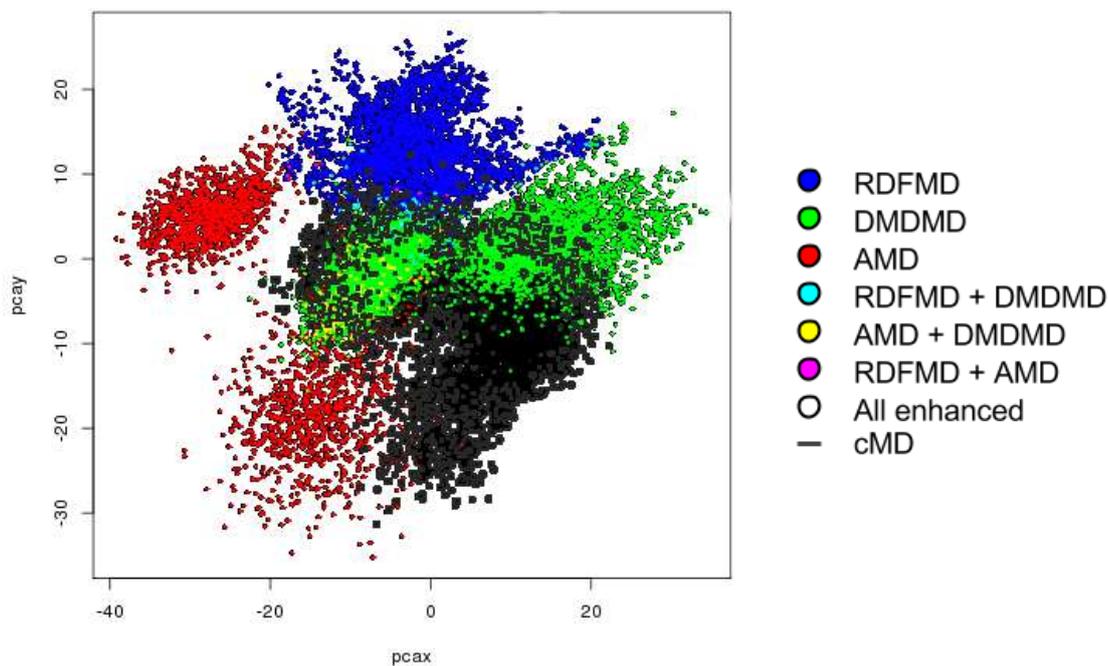


Figure A3.13: Sampling obtained using different sampling methods. Trajectories for each sampling method is projected onto the first two principal components of a supertrajectory composed from all inactive WT simulations. Where two or more snapshots with different sampling method occupy the same PC1/2 space the colours are merged additively (see legend), with the exception of the cMD, which is represented in black encased in a grey outline.

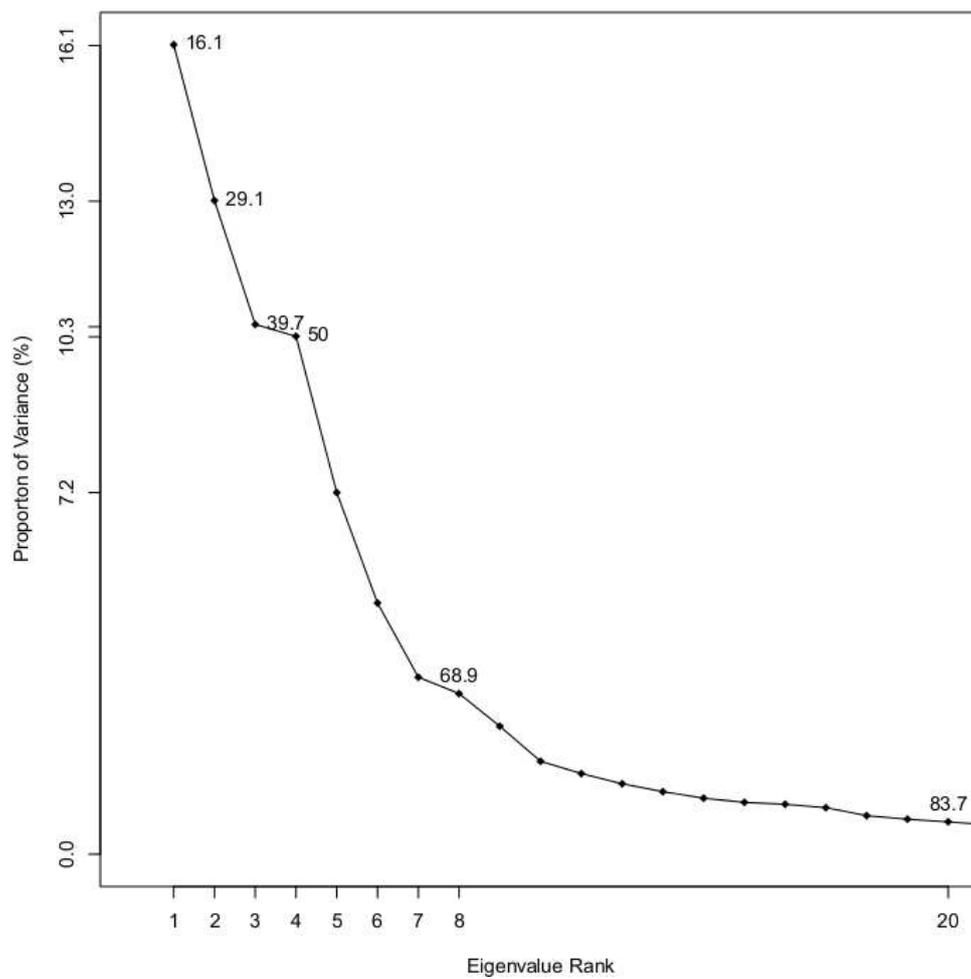


Figure A3.14: Scree plot of the proportion of variance accounted for by each principal component of the inactive deletion supertrajectory

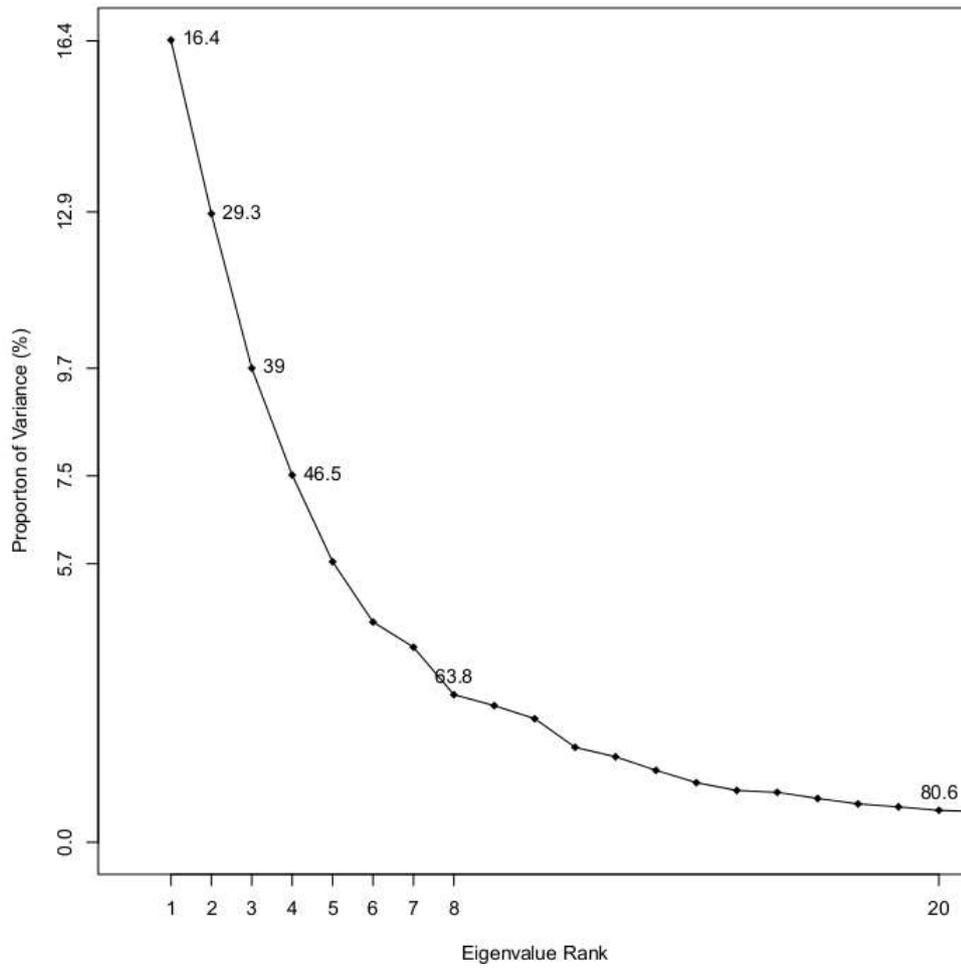


Figure A3.15: Scree plot of the proportion of variance accounted for by each principal component of the inactive G719S supertrajectory

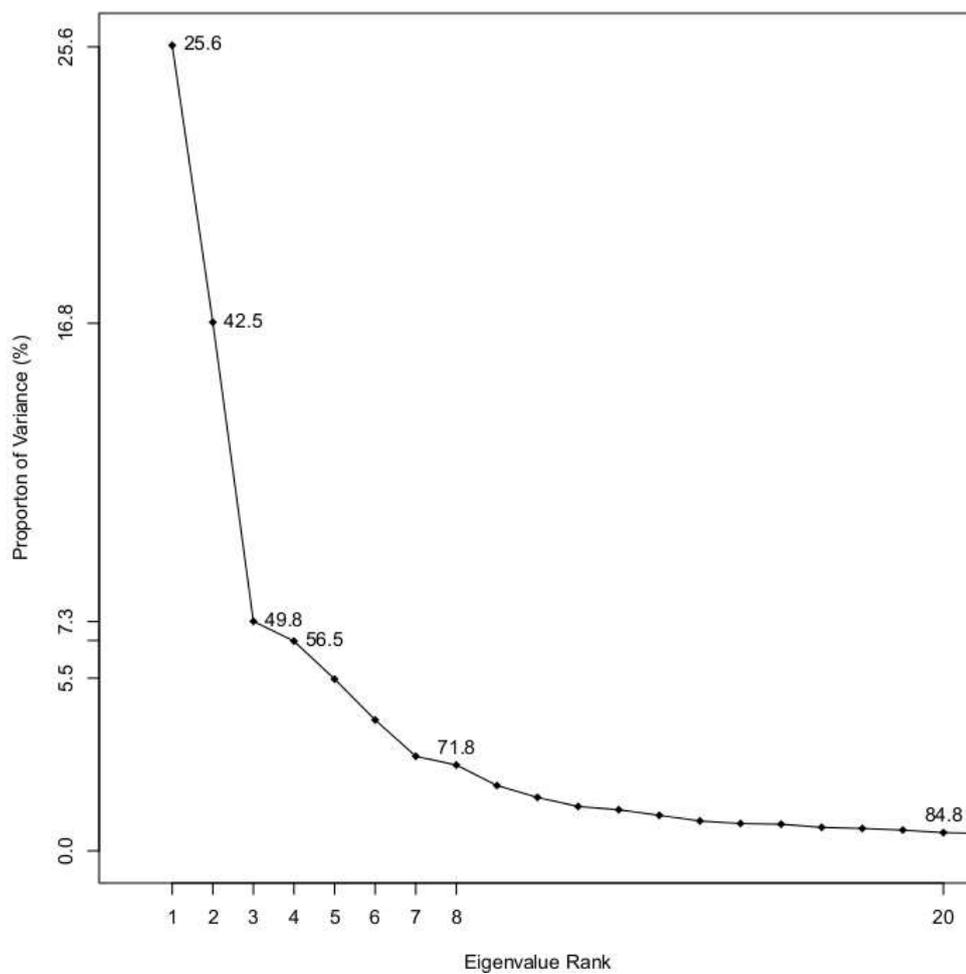


Figure A3.16: Scree plot of the proportion of variance accounted for by each principal component of the inactive L858R supertrajectory

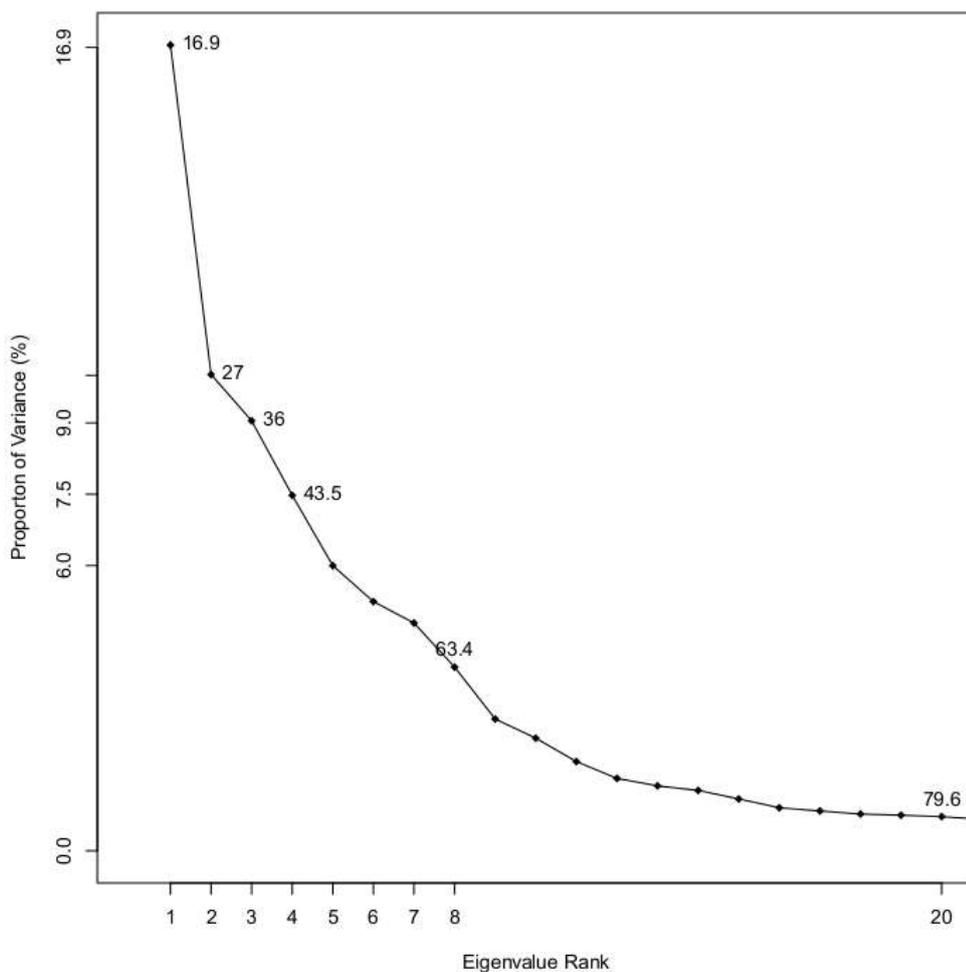


Figure A3.17: Scree plot of the proportion of variance accounted for by each principal component of the inactive WT supertrajectory

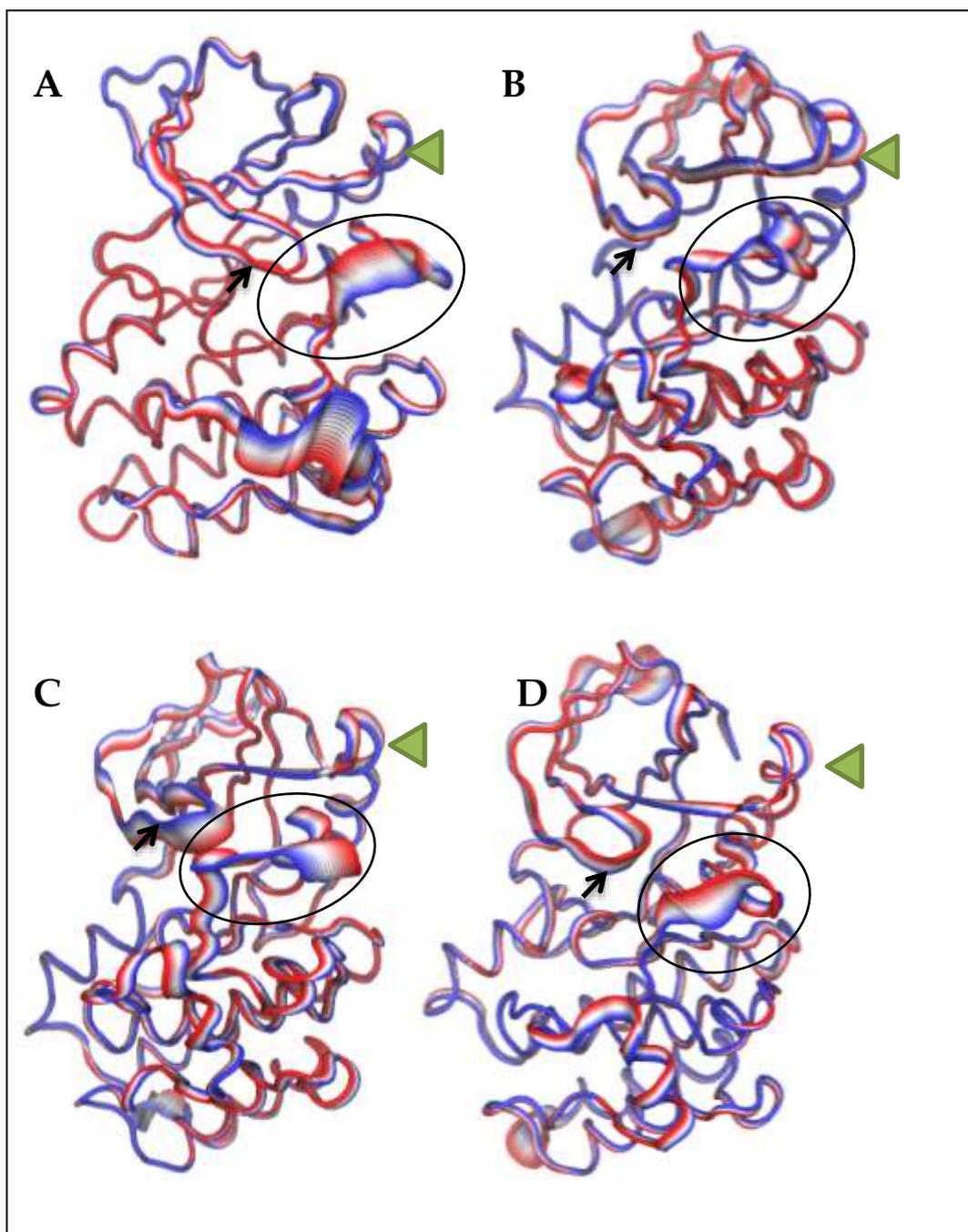


Figure A3.18: Representations of the backbone atomic displacements captured in PC2 of the active simulations of the deletion mutant (A), G719S mutant (B), L858R mutant (C), and WT (D). The P-loop is indicated by a black arrow, the C-helix by a green arrowhead, and the A-loop is circled in black.

## Appendix 4: Additional thermodynamic cycles

While the majority of thermodynamic cycles were presented in the main text, the remainder are tabulated here for expediency.

Cycle	Free leg closure (kcal mol <sup>-1</sup> )	Bound leg closure (kcal mol <sup>-1</sup> )
4-2-1-4	-0.23 ± 0.09	-0.09 ± 0.35
1-3-8-4-1	0.35 ± 0.14	0.55 ± 0.40
1-4-8-14-1	0.48 ± 0.12	-1.70 ± 0.31
4-8-28-4	0.25 ± 0.14	0.21 ± 0.29

Table A4.1: Thermodynamic cycles for simulations utilising the crystallographic waters.

Cycle	Free leg closure (kcal mol <sup>-1</sup> )	Bound leg closure (kcal mol <sup>-1</sup> )
4-2-1-4	-0.22 ± 0.10	-0.48 ± 0.21
1-3-8-4-1	0.34 ± 0.14	0.62 ± 0.53
1-4-8-14-1	0.49 ± 0.13	0.89 ± 0.25
4-8-28-4	0.25 ± 0.13	0.50 ± 0.32

Table A4.2: Thermodynamic cycles for simulations utilising the GCMC determined waters.

## Appendix 5: Elucidation of water sites by GCMC

The water site search using GCMC produces very noisy data, and justification of inclusion of water sites 2 and 3 was made on the basis of the results from all GCMC simulations on all the ligands for which GCMC was applied; however, to maintain readability of the main text, the remaining data is provided here.

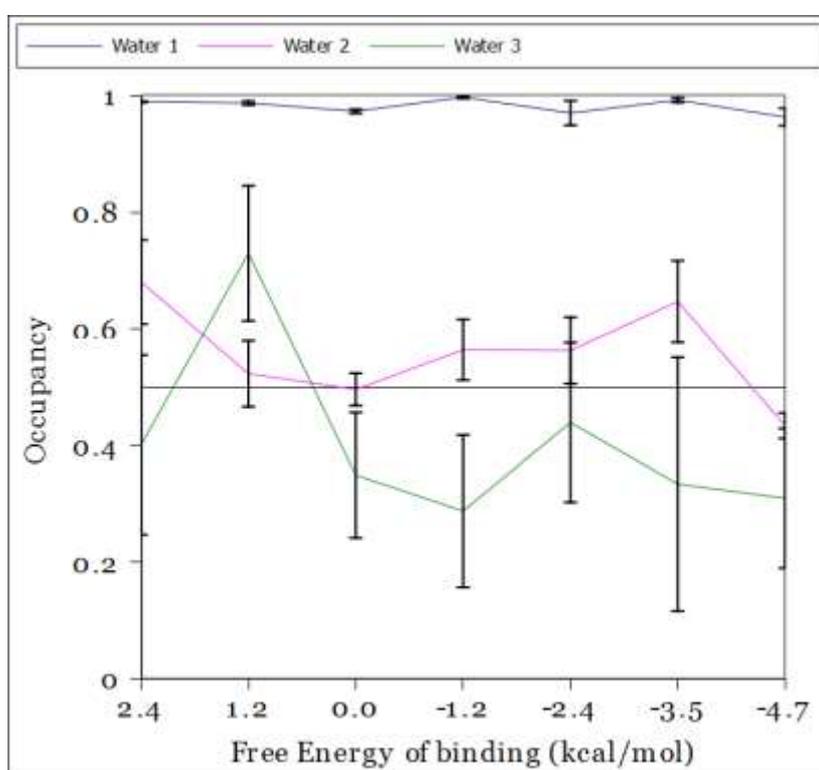


Figure A5.1: GCMC derived occupancy of waters in the binding site with ligand 1 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

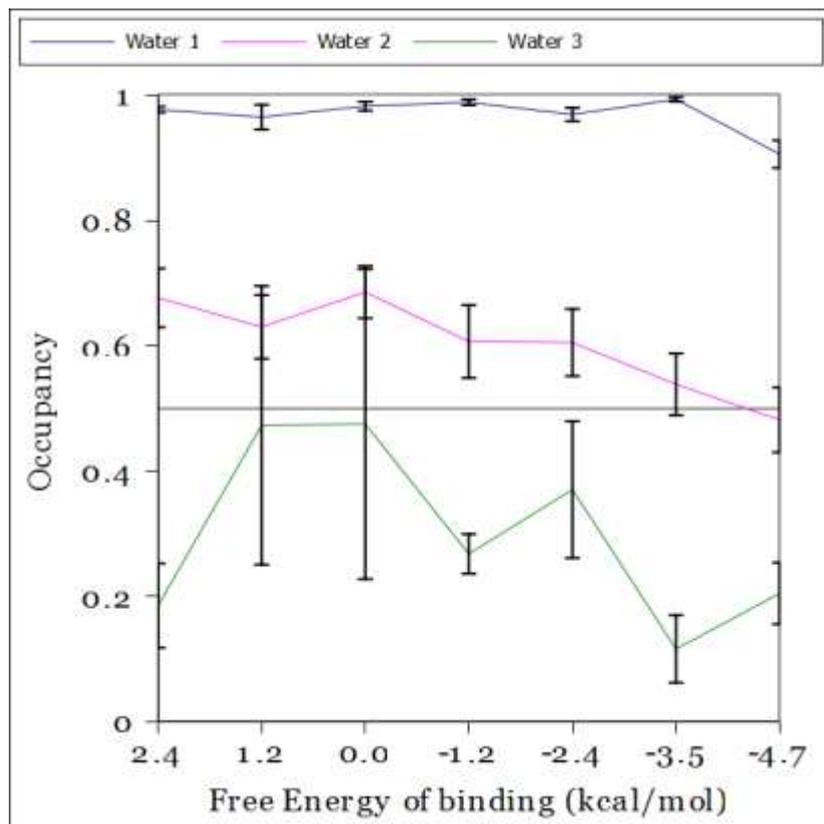


Figure A5.2: GCMC derived occupancy of waters in the binding site with ligand 2 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

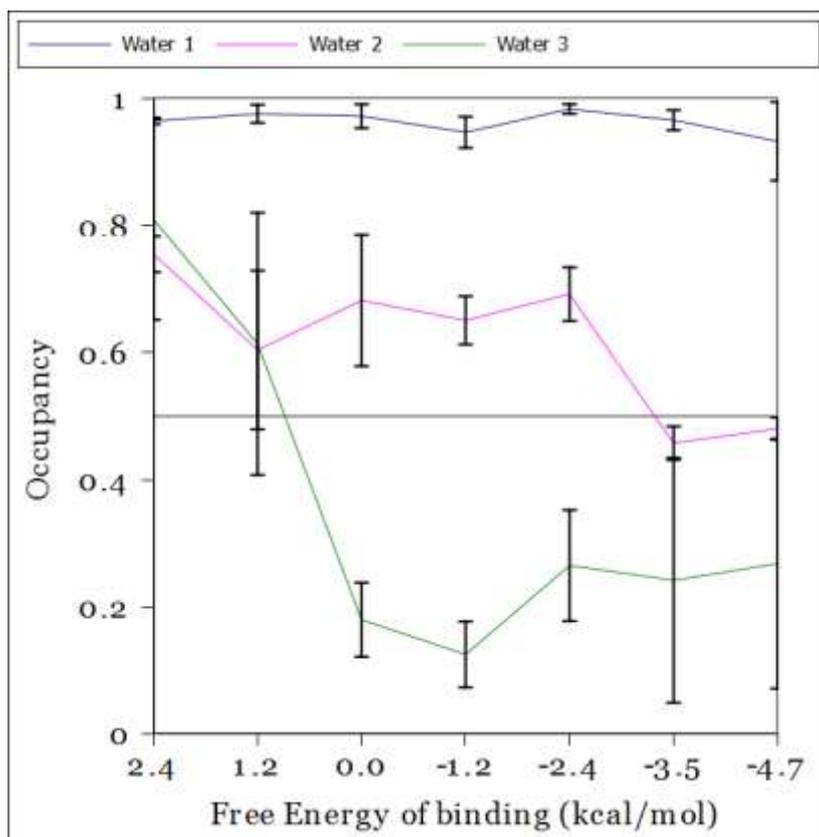


Figure A5.3: GCMC derived occupancy of waters in the binding site with ligand 3 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

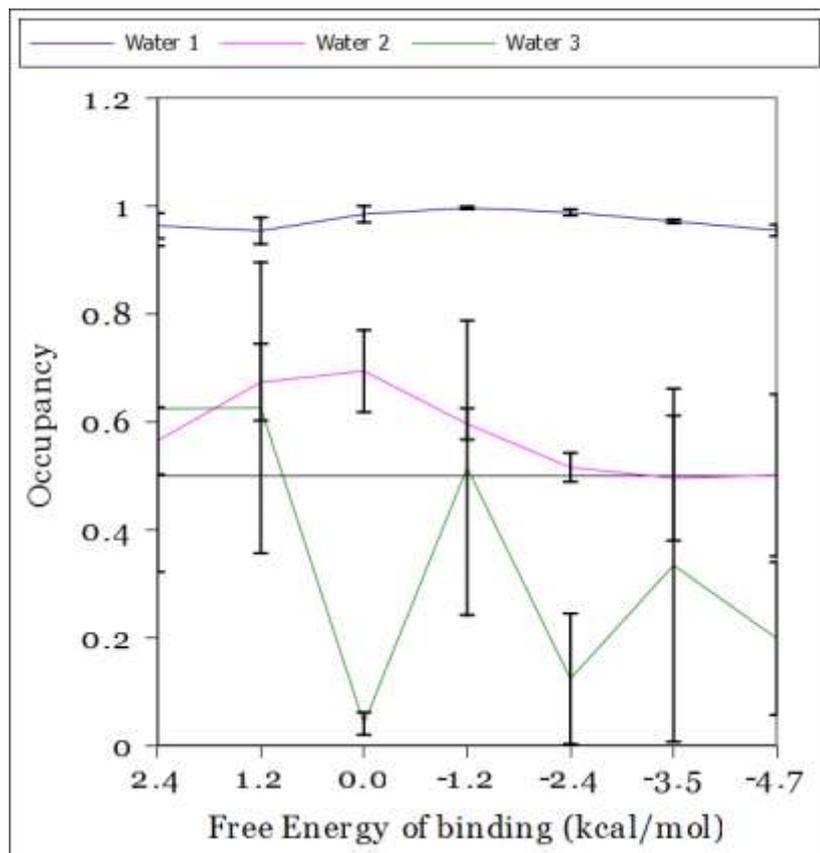


Figure A5.4: GCMC derived occupancy of waters in the binding site with ligand 4 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

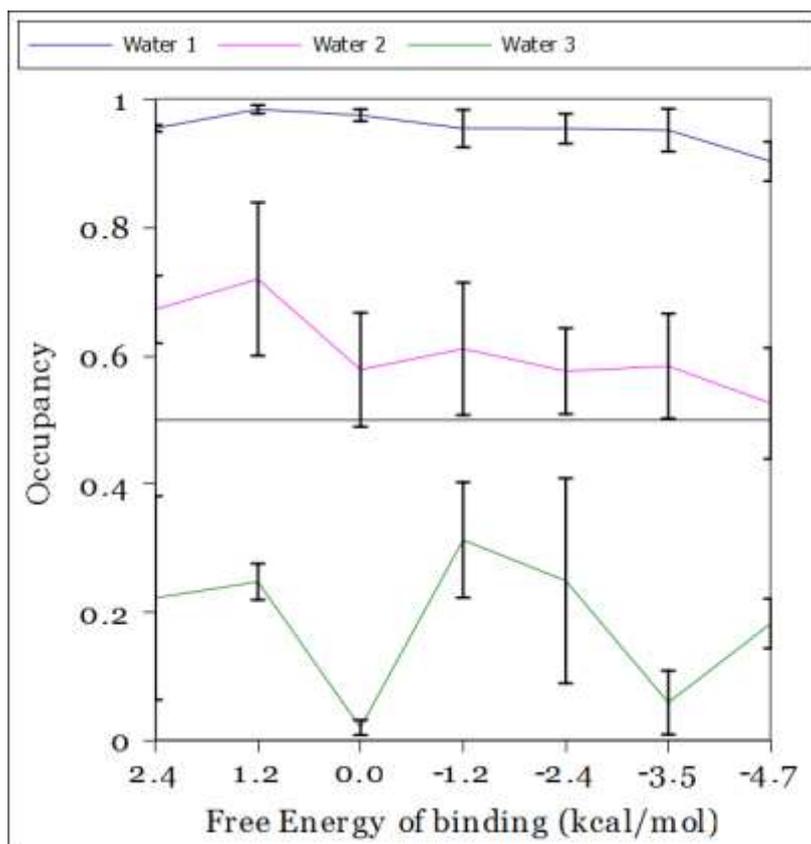


Figure A5.5: GCMC derived occupancy of waters in the binding site with ligand 8 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

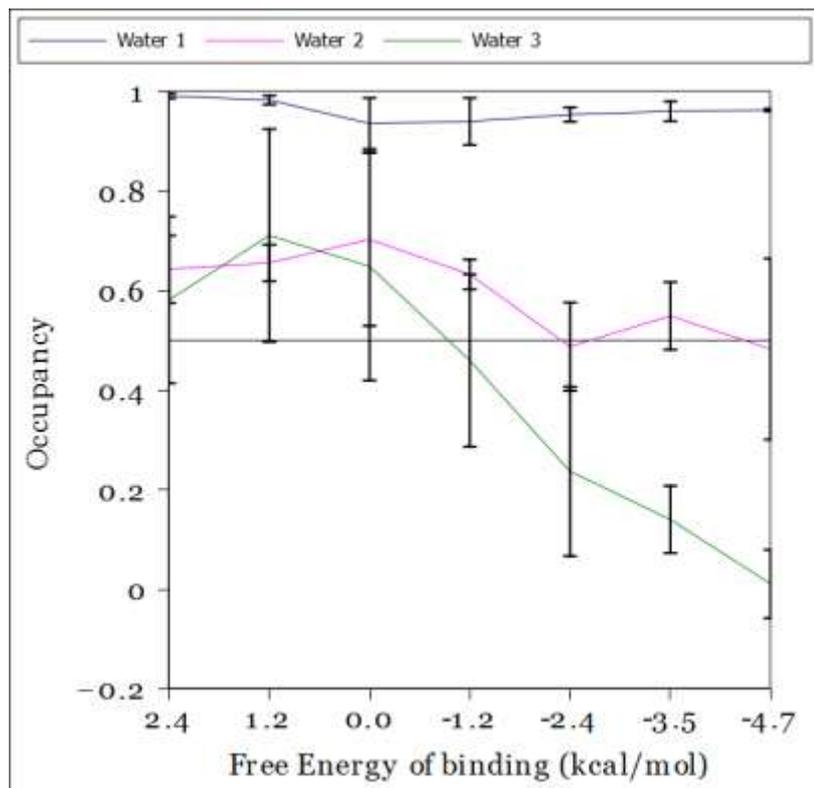


Figure A5.6: GCMC derived occupancy of waters in the binding site with ligand 28 bound, calculated at different B-factors corresponding to the binding free energies on the X-axis

## Appendix 6: Principal component dot products

The calculation of RMSIP performed in section 5.9 necessitates the calculation of the dot products of the principal components of each PCA to be compared; however, these dot products themselves can be useful in comparing separate PCAs, and being briefly discussed in the main text, they are included here for completeness.

For the comparison of the supertrajectories to the WT cMD trajectories the RMSIP was calculated for each WT cMD trajectory; however, it would require 16 tables to show the underlying dot products of that analysis, and so for brevity the following tables use the root mean square dot product taken across all of the WT cMD trajectories of a given starting conformation. Consequently, rather than scaling between -1 and 1, the RMS dot products scale between 0 and 1.

		WT Active supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT active trajectory PC	1	0.71	0.20	0.26	0.20	0.19	0.15	0.32	0.15	0.07	0.14
	2	0.33	0.19	0.65	0.16	0.34	0.21	0.18	0.13	0.14	0.06
	3	0.21	0.33	0.26	0.25	0.30	0.11	0.20	0.34	0.26	0.17
	4	0.16	0.18	0.28	0.20	0.13	0.37	0.34	0.33	0.20	0.28
	5	0.26	0.25	0.18	0.22	0.37	0.14	0.35	0.26	0.26	0.17
	6	0.09	0.21	0.19	0.13	0.35	0.24	0.24	0.19	0.29	0.09
	7	0.16	0.33	0.30	0.19	0.15	0.18	0.32	0.24	0.18	0.10
	8	0.13	0.25	0.14	0.13	0.10	0.27	0.20	0.07	0.20	0.27
	9	0.12	0.14	0.05	0.19	0.15	0.10	0.12	0.15	0.26	0.33
	10	0.15	0.17	0.10	0.24	0.28	0.24	0.32	0.15	0.20	0.28

Figure A6.1: Table of dot products between the first 10 PCs of the WT active cMD simulations and the WT active supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		L858R Active supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT active trajectory PC	1	0.62	0.14	0.32	0.13	0.22	0.12	0.18	0.11	0.20	0.06
	2	0.41	0.15	0.21	0.43	0.17	0.20	0.07	0.33	0.16	0.14
	3	0.14	0.12	0.13	0.33	0.18	0.08	0.10	0.23	0.18	0.17
	4	0.22	0.10	0.15	0.19	0.20	0.13	0.12	0.20	0.21	0.10
	5	0.19	0.11	0.09	0.32	0.16	0.04	0.13	0.12	0.26	0.22
	6	0.11	0.08	0.05	0.18	0.27	0.06	0.07	0.09	0.23	0.15
	7	0.12	0.13	0.09	0.15	0.31	0.10	0.09	0.16	0.23	0.09
	8	0.17	0.04	0.10	0.22	0.30	0.09	0.12	0.18	0.17	0.10
	9	0.09	0.04	0.15	0.14	0.21	0.07	0.05	0.20	0.12	0.11
	10	0.09	0.12	0.10	0.10	0.08	0.03	0.08	0.14	0.16	0.12

Figure A6.2: Table of dot products between the first 10 PCs of the WT active cMD simulations and the L858R active supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		G719S Active supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT active trajectory PC	1	0.65	0.32	0.07	0.16	0.40	0.13	0.04	0.07	0.12	0.14
	2	0.43	0.47	0.19	0.46	0.21	0.18	0.10	0.10	0.04	0.09
	3	0.13	0.37	0.32	0.26	0.24	0.37	0.19	0.14	0.14	0.07
	4	0.19	0.23	0.18	0.29	0.34	0.14	0.29	0.21	0.22	0.20
	5	0.12	0.27	0.35	0.34	0.24	0.21	0.15	0.16	0.15	0.12
	6	0.14	0.15	0.22	0.21	0.08	0.32	0.17	0.25	0.37	0.12
	7	0.12	0.21	0.27	0.15	0.13	0.32	0.15	0.18	0.33	0.20
	8	0.11	0.14	0.20	0.09	0.26	0.14	0.19	0.27	0.19	0.24
	9	0.06	0.11	0.24	0.20	0.11	0.17	0.24	0.19	0.26	0.31
	10	0.13	0.21	0.24	0.26	0.21	0.26	0.04	0.18	0.11	0.22

Figure A6.3: Table of dot products between the first 10 PCs of the WT active cMD simulations and the G719S active supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		Deletion Active supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT active trajectory PC	1	0.55	0.45	0.41	0.16	0.06	0.12	0.10	0.07	0.17	0.05
	2	0.27	0.49	0.51	0.15	0.21	0.18	0.24	0.10	0.09	0.06
	3	0.05	0.43	0.11	0.28	0.21	0.17	0.12	0.21	0.20	0.19
	4	0.23	0.26	0.25	0.14	0.30	0.26	0.21	0.09	0.08	0.27
	5	0.23	0.27	0.16	0.24	0.16	0.32	0.29	0.25	0.09	0.09
	6	0.27	0.13	0.24	0.31	0.12	0.37	0.24	0.20	0.11	0.24
	7	0.21	0.12	0.19	0.27	0.24	0.34	0.37	0.18	0.16	0.22
	8	0.25	0.14	0.24	0.16	0.20	0.20	0.09	0.03	0.21	0.19
	9	0.14	0.17	0.17	0.19	0.27	0.17	0.12	0.04	0.12	0.48
	10	0.09	0.04	0.04	0.32	0.29	0.30	0.09	0.19	0.28	0.13

Figure A6.4: Table of dot products between the first 10 PCs of the WT active cMD simulations and the deletion active supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		WT Inactive supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT Inactive trajectory PC	1	0.32	0.26	0.30	0.27	0.36	0.26	0.11	0.29	0.12	0.07
	2	0.28	0.32	0.26	0.13	0.16	0.21	0.36	0.26	0.23	0.08
	3	0.32	0.14	0.36	0.20	0.16	0.18	0.36	0.30	0.14	0.24
	4	0.17	0.15	0.16	0.19	0.17	0.24	0.21	0.16	0.26	0.29
	5	0.29	0.32	0.14	0.19	0.12	0.24	0.37	0.12	0.10	0.20
	6	0.13	0.11	0.10	0.20	0.19	0.24	0.20	0.21	0.40	0.32
	7	0.30	0.26	0.08	0.26	0.16	0.13	0.26	0.17	0.24	0.10
	8	0.15	0.06	0.20	0.35	0.24	0.11	0.14	0.16	0.20	0.27
	9	0.14	0.27	0.12	0.21	0.15	0.21	0.14	0.24	0.12	0.18
	10	0.23	0.21	0.16	0.16	0.15	0.22	0.12	0.09	0.13	0.15

Figure A6.5: Table of dot products between the first 10 PCs of the WT inactive cMD simulations and the WT inactive supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		L858R Inactive supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT Inactive trajectory PC	1	0.37	0.10	0.20	0.10	0.23	0.20	0.22	0.12	0.08	0.12
	2	0.32	0.13	0.20	0.39	0.29	0.11	0.16	0.16	0.06	0.11
	3	0.17	0.04	0.17	0.15	0.24	0.17	0.23	0.11	0.19	0.08
	4	0.13	0.11	0.12	0.12	0.35	0.13	0.14	0.23	0.20	0.18
	5	0.26	0.15	0.08	0.22	0.07	0.05	0.16	0.24	0.22	0.14
	6	0.28	0.11	0.05	0.29	0.12	0.08	0.14	0.17	0.26	0.15
	7	0.21	0.14	0.09	0.34	0.11	0.09	0.14	0.17	0.24	0.06
	8	0.17	0.14	0.11	0.12	0.19	0.05	0.09	0.26	0.12	0.04
	9	0.17	0.14	0.20	0.20	0.08	0.12	0.10	0.22	0.20	0.13
	10	0.08	0.18	0.14	0.13	0.12	0.09	0.09	0.12	0.18	0.08

Figure A6.6: Table of dot products between the first 10 PCs of the WT inactive cMD simulations and the L858R inactive supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		G719S Inactive supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT Inactive trajectory PC	1	0.43	0.15	0.13	0.14	0.26	0.12	0.13	0.04	0.15	0.21
	2	0.37	0.34	0.19	0.25	0.27	0.20	0.05	0.18	0.17	0.15
	3	0.27	0.22	0.21	0.25	0.31	0.15	0.12	0.10	0.21	0.21
	4	0.10	0.18	0.25	0.21	0.23	0.19	0.26	0.11	0.27	0.13
	5	0.31	0.17	0.29	0.33	0.14	0.14	0.21	0.12	0.15	0.11
	6	0.29	0.17	0.27	0.29	0.22	0.22	0.08	0.30	0.13	0.10
	7	0.17	0.20	0.16	0.32	0.18	0.25	0.25	0.18	0.09	0.13
	8	0.17	0.14	0.21	0.10	0.09	0.29	0.16	0.24	0.29	0.14
	9	0.16	0.16	0.15	0.15	0.26	0.30	0.17	0.16	0.12	0.15
	10	0.07	0.14	0.15	0.10	0.21	0.14	0.19	0.07	0.16	0.27

Figure A6.7: Table of dot products between the first 10 PCs of the WT inactive cMD simulations and the G719S inactive supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

		Deletion Inactive supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
WT Inactive trajectory PC	1	0.34	0.22	0.24	0.14	0.09	0.16	0.14	0.20	0.14	0.14
	2	0.36	0.51	0.24	0.12	0.17	0.10	0.18	0.16	0.13	0.15
	3	0.23	0.33	0.24	0.10	0.12	0.20	0.23	0.10	0.10	0.09
	4	0.23	0.11	0.41	0.13	0.15	0.36	0.21	0.13	0.14	0.13
	5	0.25	0.25	0.31	0.02	0.17	0.29	0.17	0.21	0.09	0.13
	6	0.30	0.20	0.29	0.21	0.15	0.40	0.16	0.13	0.14	0.07
	7	0.28	0.24	0.25	0.33	0.19	0.20	0.15	0.22	0.12	0.16
	8	0.17	0.08	0.20	0.39	0.24	0.29	0.13	0.19	0.17	0.10
	9	0.15	0.29	0.11	0.21	0.15	0.11	0.25	0.19	0.15	0.18
	10	0.13	0.21	0.07	0.02	0.09	0.17	0.13	0.17	0.22	0.25

Figure A6.8: Table of dot products between the first 10 PCs of the WT inactive cMD simulations and the deletion inactive supertrajectory. Each cell consists of the root mean square value of the dot product taken across each of the WT cMD simulations.

In the comparison of the supertrajectories of each sampling method the dot products were compared; however, to maintain readability in the main text, those tables of dot products are provided here.

		AMD supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
cMD trajectory PC	1	0.94	0.21	0.07	-0.04	-0.09	0.03	-0.07	0.01	0.02	0.05
	2	0.16	-0.60	-0.57	-0.01	-0.13	-0.29	0.11	0.09	-0.02	0.19
	3	-0.10	-0.03	0.34	-0.20	-0.61	-0.15	0.35	-0.04	0.24	0.07
	4	-0.10	0.35	-0.40	-0.02	-0.54	0.03	-0.32	0.10	-0.03	-0.02
	5	0.08	0.16	-0.24	-0.17	0.17	0.11	0.34	-0.26	0.20	0.02
	6	-0.11	0.25	-0.30	-0.28	-0.01	0.38	0.25	0.03	-0.33	0.09
	7	-0.02	0.08	0.03	-0.33	-0.01	-0.31	-0.13	0.01	-0.18	0.26
	8	-0.09	0.07	0.08	0.05	0.08	0.17	-0.15	0.00	0.33	0.57
	9	-0.04	-0.02	-0.13	0.12	0.05	0.16	0.15	-0.04	-0.05	0.01
	10	0.01	-0.18	-0.11	0.07	-0.02	-0.02	-0.09	-0.12	0.04	0.02

Figure A6.9: Table of dot products between the first 10 PCs of the cMD supertrajectory and the AMD supertrajectory.

		DMDMD supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
c MD trajectory PC	1	0.88	-0.07	-0.01	-0.04	0.04	0.21	0.10	-0.01	-0.04	0.01
	2	-0.05	-0.17	0.80	0.15	-0.31	0.16	-0.06	-0.06	0.06	0.06
	3	-0.05	-0.76	0.00	0.25	0.30	-0.15	0.16	-0.19	-0.02	0.01
	4	-0.04	-0.08	0.13	-0.70	-0.05	0.01	0.25	-0.28	0.06	-0.15
	5	0.02	0.26	0.08	0.18	-0.08	-0.33	0.33	-0.34	-0.17	0.10
	6	-0.10	0.34	0.14	0.30	0.19	0.17	0.45	-0.02	0.16	-0.16
	7	-0.05	0.01	0.17	-0.16	0.41	0.01	-0.11	-0.13	-0.34	-0.22
	8	-0.16	-0.27	-0.28	-0.01	-0.42	0.29	0.23	-0.06	-0.18	-0.15
	9	-0.15	0.04	-0.04	-0.07	0.15	0.24	0.15	-0.05	-0.07	0.61
	10	-0.05	0.04	0.06	-0.08	-0.10	-0.32	0.07	0.10	-0.19	-0.21

Figure A6.10: Table of dot products between the first 10 PCs of the cMD supertrajectory and the DMDMD supertrajectory.

		RDFMD supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
c MD trajectory PC	1	0.88	-0.07	-0.02	-0.16	0.06	0.12	0.09	-0.04	0.15	0.11
	2	0.02	0.34	-0.14	-0.34	-0.65	-0.04	0.28	0.00	-0.05	-0.03
	3	-0.04	-0.37	0.17	-0.13	0.13	-0.18	0.08	-0.04	-0.04	0.02
	4	0.06	0.06	0.67	0.20	-0.19	0.15	-0.12	0.26	0.26	0.04
	5	0.07	-0.19	0.06	-0.12	-0.28	-0.35	-0.20	0.17	-0.17	0.27
	6	-0.17	-0.25	0.30	-0.33	-0.15	-0.05	0.19	-0.06	0.33	-0.10
	7	-0.08	0.08	0.28	-0.43	0.15	0.33	-0.16	-0.07	-0.39	0.08
	8	-0.17	-0.15	-0.26	-0.13	0.16	0.11	-0.12	-0.01	0.39	0.40
	9	-0.08	0.13	0.13	0.20	-0.06	-0.23	0.23	-0.06	0.01	0.39
	10	-0.09	-0.02	-0.15	-0.04	-0.23	0.24	-0.27	-0.06	0.10	-0.12

Figure A6.11: Table of dot products between the first 10 PCs of the cMD supertrajectory and the RDFMD supertrajectory.

		RDFMD supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
AMD trajectory PC	1	0.88	0.02	-0.16	-0.19	-0.07	0.06	0.08	-0.04	0.03	0.05
	2	0.15	-0.31	0.44	0.16	0.28	0.06	-0.11	0.01	0.17	0.14
	3	0.05	-0.22	-0.26	0.02	0.72	-0.05	-0.12	-0.15	-0.05	-0.12
	4	-0.06	0.03	-0.23	0.34	-0.02	0.15	0.18	-0.05	0.07	0.03
	5	-0.08	0.16	-0.53	0.05	0.06	-0.02	-0.10	0.02	-0.16	0.11
	6	-0.09	-0.21	0.05	0.15	0.07	-0.08	0.04	0.08	0.32	0.15
	7	-0.07	-0.25	-0.06	-0.15	-0.12	-0.39	0.24	-0.05	-0.21	0.05
	8	-0.06	-0.03	0.02	0.01	-0.01	0.26	0.11	0.02	0.10	-0.02
	9	0.08	-0.19	-0.19	0.31	-0.05	-0.02	-0.07	0.03	-0.08	0.21
	10	-0.14	0.27	-0.03	-0.41	0.16	0.07	-0.04	-0.07	0.25	0.33

Figure A6.12: Table of dot products between the first 10 PCs of the AMD supertrajectory and the RDFMD supertrajectory.

		RDFMD supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
DMDMD trajectory PC	1	0.93	-0.07	-0.08	-0.11	0.08	0.03	-0.05	0.01	0.04	-0.07
	2	-0.03	0.14	0.03	0.10	-0.17	0.04	-0.06	0.09	-0.06	-0.08
	3	0.04	0.25	0.15	-0.38	-0.67	0.03	0.25	-0.08	-0.20	-0.18
	4	-0.05	-0.19	-0.46	-0.26	-0.05	-0.35	0.14	-0.19	-0.08	-0.09
	5	-0.02	-0.25	0.44	-0.02	0.19	0.11	0.15	-0.22	-0.33	-0.02
	6	0.14	0.14	0.02	0.12	0.04	0.17	0.37	-0.08	0.12	0.25
	7	0.03	-0.47	0.31	-0.10	-0.25	-0.26	-0.04	0.06	0.36	0.18
	8	-0.04	0.12	-0.20	0.02	0.12	-0.08	0.08	-0.33	0.08	-0.35
	9	-0.05	-0.22	-0.01	0.02	-0.05	0.03	0.38	0.26	0.10	-0.03
	10	0.04	0.28	-0.04	0.24	0.00	-0.44	0.28	0.05	-0.04	0.27

Figure A6.13: Table of dot products between the first 10 PCs of the DMDMD supertrajectory and the RDFMD supertrajectory.

		AMD supertrajectory PC									
		1	2	3	4	5	6	7	8	9	10
DMDMD trajectory PC	1	0.88	0.13	0.17	-0.06	-0.06	-0.07	-0.03	-0.07	-0.04	-0.14
	2	0.00	0.19	-0.37	-0.03	0.59	0.30	-0.04	-0.07	-0.33	-0.21
	3	0.14	-0.37	-0.56	-0.10	-0.24	-0.33	0.20	0.06	-0.23	0.06
	4	0.03	-0.26	0.13	-0.29	0.31	0.05	0.49	-0.12	0.18	-0.03
	5	-0.07	0.29	0.23	-0.22	-0.21	-0.13	0.12	0.10	-0.20	-0.26
	6	0.09	0.10	-0.13	0.02	0.14	0.11	-0.09	0.22	0.11	0.23
	7	0.00	0.28	-0.18	-0.21	-0.18	0.30	0.09	-0.20	0.07	0.08
	8	0.00	-0.16	0.31	0.28	0.12	-0.06	-0.10	-0.06	-0.24	-0.02
	9	-0.04	-0.03	-0.06	0.23	-0.11	0.07	0.15	0.34	-0.20	-0.27
	10	0.06	-0.15	-0.01	0.19	0.02	0.13	0.21	-0.29	-0.10	-0.04

Figure A6.14: Table of dot products between the first 10 PCs of the DMDMD supertrajectory and the AMD supertrajectory.

---

## Bibliography

- [1] C. L. Arteaga, "The Epidermal Growth Factor Receptor: From Mutant Oncogene in Nonhuman Cancers to Therapeutic Target in Human Neoplasia," *J. Clin. Oncol.*, vol. 19, no. suppl\_1, p. 32s–40, Sep. 2001.
- [2] J. G. Paez, P. a Jänne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel, P. Herman, F. J. Kaye, N. Lindeman, T. J. Boggon, K. Naoki, H. Sasaki, Y. Fujii, M. J. Eck, W. R. Sellers, B. E. Johnson, and M. Meyerson, "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.," *Science*, vol. 304, no. 5676, pp. 1497–500, Jun. 2004.
- [3] E. Avizienyte, R. a Ward, and A. P. Garner, "Comparison of the EGFR resistance mutation profiles generated by EGFR-targeted tyrosine kinase inhibitors and the impact of drug combinations.," *Biochem. J.*, vol. 415, no. 2, pp. 197–206, Oct. 2008.
- [4] N. Vajpai, A. Strauss, G. Fendrich, S. W. Cowan-Jacob, P. W. Manley, S. Grzesiek, and W. Jahnke, "Solution conformations and dynamics of ABL kinase-inhibitor complexes determined by NMR substantiate the different binding modes of imatinib/nilotinib and dasatinib.," *J. Biol. Chem.*, vol. 283, no. 26, pp. 18292–302, 2008.
- [5] Y. Shan, A. Arkhipov, E. T. Kim, A. C. Pan, and D. E. Shaw, "Transitions to catalytically inactive conformations in EGFR kinase.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 18, pp. 7270–5, Apr. 2013.
- [6] T. E. Balius and R. C. Rizzo, "Quantitative prediction of fold resistance for inhibitors of EGFR.," *Biochemistry*, vol. 48, no. 35, pp. 8435–48, Sep. 2009.
- [7] B. Liu, B. Bernard, and J. H. Wu, "Impact of EGFR Point Mutations on the Sensitivity to Gefitinib : Insights From Comparative Structural Analyses and Molecular Dynamics Simulations," vol. 346, no. April, pp. 331–346, 2006.
- [8] K. Wang, H. Yamamoto, J. R. Chin, Z. Werb, and T. H. Vu, "Epidermal growth factor receptor-deficient mice have delayed primary endochondral ossification because of defective osteoclast recruitment.," *J. Biol. Chem.*, vol. 279, no. 51, pp. 53848–56, Dec. 2004.
- [9] J. F. Wiesen, P. Young, Z. Werb, and G. R. Cunha, "Signaling through the stromal epidermal growth factor receptor is necessary for mammary ductal development.," *Development*, vol. 126, no. 2, pp. 335–44, Jan. 1999.
- [10] B. Chen, R. T. Bronson, L. D. Klamann, T. G. Hampton, J. F. Wang, P. J. Green, T. Magnuson, P. S. Douglas, J. P. Morgan, and B. G. Neel, "Mice mutant for Egfr and Shp2 have defective cardiac semilunar valvulogenesis.," *Nat. Genet.*, vol. 24, no. 3, pp. 296–9, Mar. 2000.

## Bibliography

---

- [11] E. D. Adamson and A. R. Rees, "Epidermal growth factor receptors," *Mol. Cell. Biochem.*, vol. 34, no. 3, pp. 129–152, Feb. 1981.
- [12] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian, "TiGER: A database for tissue-specific gene expression and regulation," *BMC Bioinformatics*, vol. 9, p. 271, 2008.
- [13] "TiGER," 2013. [Online]. Available: [http://bioinfo.wilmer.jhu.edu/tiger/db\\_gene/NM\\_005228-index.html](http://bioinfo.wilmer.jhu.edu/tiger/db_gene/NM_005228-index.html).
- [14] R. . Nicholson, J. M. . Gee, and M. . Harper, "EGFR and cancer prognosis," *Eur. J. Cancer*, vol. 37, pp. 9–15, Sep. 2001.
- [15] A. Herrlich, E. Klinman, J. Fu, C. Sadegh, and H. Lodish, "Ectodomain cleavage of the EGF ligands HB-EGF, neuregulin1-beta, and TGF-alpha is specifically triggered by different stimuli and involves different PKC isoenzymes.," *FASEB J.*, vol. 22, no. 12, pp. 4281–95, Dec. 2008.
- [16] I. Chung, R. Akita, R. Vandlen, D. Toomre, J. Schlessinger, and I. Mellman, "Spatial control of EGF receptor activation by reversible dimerization on living cells.," *Nature*, vol. 464, no. 7289, pp. 783–7, Apr. 2010.
- [17] A. Wells, "EGF receptor," *Int. J. Biochem. Cell Biol.*, vol. 31, no. 6, pp. 637–643, Jun. 1999.
- [18] P. Sanchez-Soria and T. D. Camenisch, "ErbB signaling in cardiac development and disease.," *Semin. Cell Dev. Biol.*, vol. 21, no. 9, pp. 929–35, Dec. 2010.
- [19] J. B. Park, C. S. Lee, J.-H. Jang, J. Ghim, Y.-J. Kim, S. You, D. Hwang, P.-G. Suh, and S. H. Ryu, "Phospholipase signalling networks in cancer.," *Nat. Rev. Cancer*, vol. 12, no. 11, pp. 782–92, Nov. 2012.
- [20] H.-W. Lo, S.-C. Hsu, M. Ali-Seyed, M. Gunduz, W. Xia, Y. Wei, G. Bartholomeusz, J.-Y. Shih, and M.-C. Hung, "Nuclear interaction of EGFR and STAT3 in the activation of the iNOS/NO pathway.," *Cancer Cell*, vol. 7, no. 6, pp. 575–89, Jun. 2005.
- [21] T. F. Franke, "Direct Regulation of the Akt Proto-Oncogene Product by Phosphatidylinositol-3,4-bisphosphate," *Science*, vol. 275, no. 5300, pp. 665–668, Jan. 1997.
- [22] A. Wells, K. Gupta, P. Chang, S. Swindle, A. Glading, and H. Shiraha, "Epidermal growth factor receptor-mediated motility in fibroblasts.," *Microsc. Res. Tech.*, vol. 43, no. 5, pp. 395–411, Dec. 1998.
- [23] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, 2000.

- [24] H. M. Miettinen, H. Huuskonen, A.-M. Partanen, P. Miettinen, J. T. Tuomisto, R. Pohjanvirta, and J. Tuomisto, "Effects of epidermal growth factor receptor deficiency and 2,3,7,8-tetrachlorodibenzo-p-dioxin on fetal development in mice.," *Toxicol. Lett.*, vol. 150, no. 3, pp. 285–91, May 2004.
- [25] S. Okamoto and T. Oka, "Evidence for physiological function of epidermal growth factor: pregestational sialoadenectomy of mice decreases milk production and increases offspring mortality during lactation period.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 81, pp. 6059–6063, 1984.
- [26] H. Shigematsu and A. F. Gazdar, "Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers.," *Int. J. cancer*, vol. 118, no. 2, pp. 257–62, Jan. 2006.
- [27] Anon, "[http://en.wikipedia.org/wiki/File:Signal\\_transduction\\_pathways.svg](http://en.wikipedia.org/wiki/File:Signal_transduction_pathways.svg)," 2007. .
- [28] C. Shang, Y. Guo, S. Fu, W. Fu, and K. Sun, "SH3GL2 gene participates in MEK-ERK signal pathway partly by regulating EGFR in the laryngeal carcinoma cell line Hep2.," *Med. Sci. Monit.*, vol. 16, no. 6, pp. BR168–73, Jun. 2010.
- [29] X. Zhang, J. Gureasko, K. Shen, P. A. Cole, and J. Kuriyan, "An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor.," *Cell*, vol. 125, no. 6, pp. 1137–49, Jun. 2006.
- [30] E. E. Er, M. C. Mendoza, A. M. Mackey, L. E. Rameh, and J. Blenis, "AKT facilitates EGFR trafficking and degradation by phosphorylating and activating PIKfyve.," *Sci. Signal.*, vol. 6, no. 279, p. ra45, Jun. 2013.
- [31] J. Kästner, H. H. Loeffler, S. K. Roberts, M. L. Martin-Fernandez, and M. D. Winn, "Ectodomain orientation, conformational plasticity and oligomerization of ErbB1 receptors investigated by molecular dynamics.," *J. Struct. Biol.*, vol. 167, no. 2, pp. 117–28, Aug. 2009.
- [32] A. Arkhipov, Y. Shan, R. Das, N. F. Endres, M. P. Eastwood, D. E. Wemmer, J. Kuriyan, and D. E. Shaw, "Architecture and membrane interactions of the EGF receptor.," *Cell*, vol. 152, no. 3, pp. 557–69, Jan. 2013.
- [33] C. J. Tynan, S. K. Roberts, D. J. Rolfe, D. T. Clarke, H. H. Loeffler, J. Kästner, M. D. Winn, P. J. Parker, and M. L. Martin-Fernandez, "Human epidermal growth factor receptor (EGFR) aligned on the plasma membrane adopts key features of Drosophila EGFR asymmetry.," *Mol. Cell. Biol.*, vol. 31, no. 11, pp. 2241–52, Jun. 2011.
- [34] J. Stamos, M. X. Sliwkowski, and C. Eigenbrot, "Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-

- anilinoquinazoline inhibitor.," *J. Biol. Chem.*, vol. 277, no. 48, pp. 46265–72, Nov. 2002.
- [35] C.-H. Yun, T. J. Boggon, Y. Li, M. S. Woo, H. Greulich, M. Meyerson, and M. J. Eck, "Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity.," *Cancer Cell*, vol. 11, no. 3, pp. 217–27, Mar. 2007.
- [36] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. F. Ten Eyck, "Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 47, pp. 17783–8, Nov. 2006.
- [37] M. Red Brewer, S. H. Choi, D. Alvarado, K. Moravcevic, A. Pozzi, M. a Lemmon, and G. Carpenter, "The juxtamembrane region of the EGF receptor functions as an activation domain.," *Mol. Cell*, vol. 34, no. 6, pp. 641–51, Jun. 2009.
- [38] N. Jura, N. F. Endres, K. Engel, S. Deindl, R. Das, M. H. Lamers, D. E. Wemmer, X. Zhang, and J. Kuriyan, "Mechanism for Activation of the EGF Receptor Catalytic Domain by the Juxtamembrane Segment," *Cell*, vol. 137, pp. 1293–1307, 2009.
- [39] M. A. Lemmon and J. Schlessinger, "Cell signaling by receptor tyrosine kinases.," *Cell*, vol. 141, no. 7, pp. 1117–34, Jun. 2010.
- [40] J. A. Endicott, M. E. M. Noble, and L. N. Johnson, "The structural basis for control of eukaryotic protein kinases.," *Annu. Rev. Biochem.*, vol. 81, pp. 587–613, Jan. 2012.
- [41] E. R. Wood, A. T. Truesdale, O. B. McDonald, D. Yuan, A. Hassell, S. H. Dickerson, B. Ellis, C. Pennisi, E. Horne, K. Lackey, K. J. Alligood, D. W. Rusnak, T. M. Gilmer, and L. Shewchuk, "A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells.," *Cancer Res.*, vol. 64, no. 18, pp. 6652–9, Sep. 2004.
- [42] L. Sutto and F. L. Gervasio, "Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 26, pp. 10616–21, Jun. 2013.
- [43] Y. Shan, M. P. Eastwood, X. Zhang, E. T. Kim, A. Arkhipov, R. O. Dror, J. Jumper, J. Kuriyan, and D. E. Shaw, "Oncogenic mutations counteract intrinsic disorder in the EGFR kinase and promote receptor dimerization.," *Cell*, vol. 149, no. 4, pp. 860–70, May 2012.
- [44] H. Greulich, T.-H. Chen, W. Feng, P. A. Jänne, J. V Alvarez, M. Zappaterra, S. E. Bulmer, D. A. Frank, W. C. Hahn, W. R. Sellers, and M. Meyerson, "Oncogenic

- Transformation by Inhibitor-Sensitive and -Resistant EGFR Mutants," *PLoS Med.*, vol. 2, p. e313, 2005.
- [45] D. M. Jackman, B. Y. Yeap, L. V Sequist, N. Lindeman, A. J. Holmes, V. a Joshi, D. W. Bell, M. S. Huberman, B. Halmos, M. S. Rabin, D. a Haber, T. J. Lynch, M. Meyerson, B. E. Johnson, and P. a Jänne, "Exon 19 deletion mutations of epidermal growth factor receptor are associated with prolonged survival in non-small cell lung cancer patients treated with gefitinib or erlotinib.," *Clin. Cancer Res.*, vol. 12, no. 13, pp. 3908–14, Jul. 2006.
- [46] R. Sordella, D. W. Bell, D. a Haber, and J. Settleman, "Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways.," *Science (80-. )*, vol. 305, no. 5687, pp. 1163–7, Aug. 2004.
- [47] A. F. Gazdar, "Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors.," *Oncogene*, vol. 28 Suppl 1, no. S1, pp. S24–31, Aug. 2009.
- [48] K. D. Carey, A. J. Garton, M. S. Romero, J. Kahler, S. Thomson, S. Ross, F. Park, J. D. Haley, N. Gibson, and M. X. Sliwkowski, "Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib.," *Cancer Res.*, vol. 66, pp. 8163–71, 2006.
- [49] R. Bose and X. Zhang, "The ErbB kinase domain: structural perspectives into kinase activation and inhibition.," *Exp. Cell Res.*, vol. 315, no. 4, pp. 649–58, Feb. 2009.
- [50] A. J. Shih, S. E. Telesco, S.-H. Choi, M. a Lemmon, and R. Radhakrishnan, "Molecular dynamics analysis of conserved hydrophobic and hydrophilic bond-interaction networks in ErbB family kinases.," *Biochem. J.*, vol. 436, no. 2, pp. 241–51, Jun. 2011.
- [51] S. Kobayashi, T. J. Boggon, T. Dayaram, P. a Jänne, O. Kocher, M. Meyerson, B. E. Johnson, M. J. Eck, D. G. Tenen, and B. Halmos, "EGFR mutation and resistance of non-small-cell lung cancer to gefitinib.," *N. Engl. J. Med.*, vol. 352, no. 8, pp. 786–92, Feb. 2005.
- [52] C.-H. Yun, K. E. Mengwasser, A. V Toms, M. S. Woo, H. Greulich, K.-K. Wong, M. Meyerson, and M. J. Eck, "The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 6, pp. 2070–5, Feb. 2008.
- [53] J.-Y. Wu, C.-J. Yu, Y.-C. Chang, C.-H. Yang, J.-Y. Shih, and P.-C. Yang, "Effectiveness of tyrosine kinase inhibitors on 'uncommon' epidermal growth factor receptor mutations of unknown clinical significance in non-small cell lung cancer.," *Clin. Cancer Res.*, vol. 17, no. 11, pp. 3812–21, Jun. 2011.

## Bibliography

---

- [54] J.-Y. Wu, S.-G. Wu, C.-H. Yang, C.-H. Gow, Y.-L. Chang, C.-J. Yu, J.-Y. Shih, and P.-C. Yang, "Lung cancer with epidermal growth factor receptor exon 20 mutations is associated with poor gefitinib treatment response.," *Clin. Cancer Res.*, vol. 14, no. 15, pp. 4877–82, Aug. 2008.
- [55] N. Fujimoto, M. Wislez, J. Zhang, K. Iwanaga, J. Dackor, A. E. Hanna, S. Kalyankrishna, D. D. Cody, R. E. Price, M. Sato, J. W. Shay, J. D. Minna, M. Peyton, X. Tang, E. Massarelli, R. Herbst, D. W. Threadgill, I. I. Wistuba, and J. M. Kurie, "High expression of ErbB family members and their ligands in lung adenocarcinomas that are sensitive to inhibition of epidermal growth factor receptor.," *Cancer Res.*, vol. 65, no. 24, pp. 11478–85, Dec. 2005.
- [56] C. T. Kuan, C. J. Wikstrand, and D. D. Bigner, "EGF mutant receptor vIII as a molecular target in cancer therapy.," *Endocr. Relat. Cancer*, vol. 8, no. 2, pp. 83–96, Jun. 2001.
- [57] G. Chen, P. Kronenberger, E. Teugels, I. A. Umelo, and J. De Grève, "Effect of siRNAs targeting the EGFR T790M mutation in a non-small cell lung cancer cell line resistant to EGFR tyrosine kinase inhibitors and combination with various agents.," *Biochem. Biophys. Res. Commun.*, vol. 431, no. 3, pp. 623–9, Feb. 2013.
- [58] S.-H. Chen and G. Zhaori, "Potential clinical applications of siRNA technique: benefits and limitations.," *Eur. J. Clin. Invest.*, vol. 41, no. 2, pp. 221–32, Feb. 2011.
- [59] M. Sattler, O. Abidoeye, and R. Salgia, "EGFR-targeted therapeutics: focus on SCCHN and NSCLC.," *ScientificWorldJournal.*, vol. 8, pp. 909–19, Jan. 2008.
- [60] T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, D. N. Louis, D. C. Christiani, J. Settleman, and D. A. Haber, "Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.," *N. Engl. J. Med.*, vol. 350, no. 21, pp. 2129–39, May 2004.
- [61] K. S. Gajiwala, J. Feng, R. Ferre, K. Ryan, O. Brodsky, S. Weinrich, J. C. Kath, and A. Stewart, "Insights into the aberrant activity of mutant EGFR kinase domain and drug recognition.," *Structure*, vol. 21, no. 2, pp. 209–19, Feb. 2013.
- [62] A. Papakyriakou, D. Vourloumis, F. Tzortzatou-Stathopoulou, and M. Karpusas, "Conformational dynamics of the EGFR kinase domain reveals structural features involved in activation.," *Proteins*, vol. 76, no. 2, pp. 375–86, Aug. 2009.
- [63] A. Dixit and G. M. Verkhivker, "Hierarchical modeling of activation mechanisms in the ABL and EGFR kinase domains: thermodynamic and mechanistic catalysts of kinase activation by cancer mutations.," *PLoS Comput. Biol.*, vol. 5, no. 8, p. e1000487, Aug. 2009.

- [64] M. Mustafa, A. Mirza, and N. Kannan, "Conformational regulation of the EGFR kinase core by the juxtamembrane and C-terminal tail: a molecular dynamics study.," *Proteins*, vol. 79, no. 1, pp. 99–114, Jan. 2011.
- [65] S. Wan and P. V Coveney, "Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor.," *J. Comput. Chem.*, vol. 32, no. 13, pp. 2843–52, Oct. 2011.
- [66] A. Dixit and G. M. Verkhivker, "Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in ABL and EGFR kinases.," *PLoS Comput. Biol.*, vol. 7, no. 10, p. e1002179, Oct. 2011.
- [67] A. Vema, "Design of EGFR kinase inhibitors: A ligand-based approach and its confirmation with structure-based studies.," *Bioorg. Med. Chem.*, vol. 11, no. 21, pp. 4643–4653, Oct. 2003.
- [68] C. La Motta, S. Sartini, T. Tuccinardi, E. Nerini, F. Da Settimo, and A. Martinelli, "Computational studies of epidermal growth factor receptor: docking reliability, three-dimensional quantitative structure-activity relationship analysis, and virtual screening studies.," *J. Med. Chem.*, vol. 52, no. 4, pp. 964–75, Feb. 2009.
- [69] E. B. Yang, Y. J. Guo, K. Zhang, Y. Z. Chen, and P. Mack, "Inhibition of epidermal growth factor receptor tyrosine kinase by chalcone derivatives.," *Biochim. Biophys. Acta*, vol. 1550, no. 2, pp. 144–52, Dec. 2001.
- [70] C. N. Cavasotto, M. a Ortiz, R. a Abagyan, and F. J. Piedrafita, "In silico identification of novel EGFR inhibitors with antiproliferative activity against cancer cells.," *Bioorg. Med. Chem. Lett.*, vol. 16, no. 7, pp. 1969–74, Apr. 2006.
- [71] T. Usui, H. S. Ban, J. Kawada, T. Hirokawa, and H. Nakamura, "Discovery of indenopyrazoles as EGFR and VEGFR-2 tyrosine kinase inhibitors by in silico high-throughput screening.," *Bioorg. Med. Chem. Lett.*, vol. 18, no. 1, pp. 285–8, Jan. 2008.
- [72] S. Kotra, K. K. Madala, and K. Jamil, "Homology models of the mutated EGFR and their response towards quinazoline analogues.," *J. Mol. Graph. Model.*, vol. 27, no. 3, pp. 244–54, Oct. 2008.
- [73] R. Chang, *Physical chemistry for the chemical and biological sciences*. University Science Books, 2000.
- [74] A. R. Leach, *Molecular Modelling Principles and Applications Second Edition*. 2001.
- [75] R. Loncharich, "Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide.," *Biopolymers*, 1992.

## Bibliography

---

- [76] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, p. 3684, Oct. 1984.
- [77] H. Resat and M. Mezei, "Grand Canonical Monte Carlo Simulation of Water Positions in Crystal Hydrates," *J. Am. Chem. Soc.*, vol. 116, no. 16, pp. 7451–7452, Aug. 1994.
- [78] H.-J. Woo, A. R. Dinner, and B. Roux, "Grand canonical Monte Carlo simulations of water in protein environments.," *J. Chem. Phys.*, vol. 121, no. 13, pp. 6392–400, Oct. 2004.
- [79] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, vol. p. 1987, p. 385.
- [80] Hyung-June Woo, A. R. Dinner, and B. Roux, "Grand canonical Monte Carlo simulations of water in protein environments," *J Chem Phys*, vol. 121, pp. 6392–6400, 2004.
- [81] F. Guarnieri and M. Mezei, "Simulated Annealing of Chemical Potential: A General Procedure for Locating Bound Waters. Application to the Study of the Differential Hydration Propensities of the Major and Minor Grooves of DNA," *J. Am. Chem. Soc.*, vol. 118, no. 35, pp. 8493–8494, Jan. 1996.
- [82] J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Prediction of the water content in protein binding sites.," *J. Phys. Chem. B*, vol. 113, no. 40, pp. 13337–46, Oct. 2009.
- [83] V. Tozzini, "Coarse-grained models for proteins," *Curr. Opin. Struct. Biol.*, vol. 15, no. 2, pp. 144–150, 2005.
- [84] T. A. Halgren and W. Damm, "Polarizable force fields," *Curr. Opin. Struct. Biol.*, vol. 11, no. 2, pp. 236–242, 2001.
- [85] A. Bernardi, A. M. Capelli, A. Comotti, C. Gennari, M. Gardner, J. M. Goodman, and I. Paterson, "Origins of stereoselectivity in chiral boron enolate aldol reactions: A computational study using transition state modellings," *Tetrahedron*, vol. 47, no. 20–21, pp. 3471–3484, 1991.
- [86] A. Warshel, "Computer simulations of enzyme catalysis: methods, progress, and insights," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 32, pp. 425–43, Jan. 2003.
- [87] D. M. York, "Atomic-Level Accuracy in Simulations of Large Protein Crystals," *Proc. Natl. Acad. Sci.*, vol. 91, no. 18, pp. 8715–8718, Aug. 1994.
- [88] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An  $N \cdot \log(N)$  method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, p. 10089, Jun. 1993.

- [89] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, p. 926, Jul. 1983.
- [90] S. Fischer and M. Karplus, "Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom," *Chem. Phys. Lett.*, vol. 194, no. 3, pp. 252–261, 1992.
- [91] P. A. Bash, U. C. Singh, R. Langridge, and P. A. Kollman, "Free energy calculations by computer simulation," *Science*, vol. 236, no. 4801, pp. 564–8, May 1987.
- [92] P. Hohenberg, "Inhomogeneous Electron Gas," *Phys. Rev.*, vol. 136, no. 3B, pp. B864–B871, Nov. 1964.
- [93] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Phys. Rev.*, vol. 140, no. 4A, pp. A1133–A1138, Nov. 1965.
- [94] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *J. Chem. Phys.*, vol. 98, no. 7, p. 5648, Apr. 1993.
- [95] E. Rosta, M. Klähn, and A. Warshel, "Towards accurate ab initio QM/MM calculations of free-energy profiles of enzymatic reactions.," *J. Phys. Chem. B*, vol. 110, no. 6, pp. 2934–41, Feb. 2006.
- [96] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment.," *Proteins*, vol. 23, no. 4, pp. 566–79, Dec. 1995.
- [97] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.," *Biopolymers*, vol. 22, no. 12, pp. 2577–637, Dec. 1983.
- [98] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A. G. de Brevern, and J.-F. Gibrat, "Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.," *BMC Struct. Biol.*, vol. 5, p. 17, Jan. 2005.
- [99] D. M. van Aalten, D. A. Conn, B. L. de Groot, H. J. Berendsen, J. B. Findlay, and A. Amadei, "Protein dynamics derived from clusters of crystal structures.," *Biophys. J.*, vol. 73, no. 6, pp. 2891–6, Dec. 1997.
- [100] G. E. Sims, I.-G. Choi, and S.-H. Kim, "Protein conformational space in higher order phi-Psi maps.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 3, pp. 618–21, Jan. 2005.
- [101] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map.," *J. Chem. Phys.*, vol. 134, no. 12, p. 124116, Mar. 2011.

## Bibliography

---

- [102] N. Vajpai, A. Strauss, G. Fendrich, S. W. Cowan-Jacob, P. W. Manley, S. Grzesiek, and W. Jahnke, "Solution conformations and dynamics of ABL kinase-inhibitor complexes determined by NMR substantiate the different binding modes of imatinib/nilotinib and dasatinib.," *J. Biol. Chem.*, vol. 283, no. 26, pp. 18292–302, Jun. 2008.
- [103] D. Hamelberg, J. Mongan, and J. A. McCammon, "Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules.," *J. Chem. Phys.*, vol. 120, no. 24, pp. 11919–29, Jun. 2004.
- [104] D. Hamelberg, C. A. F. de Oliveira, and J. A. McCammon, "Sampling of slow diffusive conformational transitions with accelerated molecular dynamics.," *J. Chem. Phys.*, vol. 127, no. 15, p. 155102, Oct. 2007.
- [105] L. C. T. Pierce, R. Salomon-Ferrer, C. Augusto F de Oliveira, J. A. McCammon, and R. C. Walker, "Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics.," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 2997–3002, Sep. 2012.
- [106] W. Zheng, M. A. Rohrdanz, and C. Clementi, "Rapid Exploration of Configuration Space with Diffusion-Map-Directed Molecular Dynamics.," *J. Phys. Chem. B*, Aug. 2013.
- [107] M. A. Rohrdanz, W. Zheng, B. Lambeth, and C. Clementi, "Multiscale characterization of macromolecular dynamics," in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment Gateway to Discovery - XSEDE '13*, 2013, p. 1.
- [108] A. P. Wiley, M. T. Swain, S. C. Phillips, J. W. Essex, and C. M. Edge, "Parametrization of reversible digitally filtered molecular dynamics simulations," *Journal of Chemical Theory and Computation*. 28-Feb-2005.
- [109] S. C. Phillips, M. T. Swain, A. P. Wiley, J. W. Essex, and C. M. Edge, "Reversible Digitally Filtered Molecular Dynamics," *J. Phys. Chem. B*, vol. 107, no. 9, pp. 2098–2110, Mar. 2003.
- [110] J. Michel, J. Tirado-Rives, and W. L. Jorgensen, "Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization.," *J. Am. Chem. Soc.*, vol. 131, no. 42, pp. 15403–11, Oct. 2009.
- [111] G. W. Rewcastle, W. A. Denny, A. J. Bridges, H. Zhou, D. R. Cody, A. McMichael, and D. W. Fry, "Tyrosine kinase inhibitors. 5. Synthesis and structure-activity relationships for 4-[(phenylmethyl)amino]- and 4-(phenylamino)quinazolines as potent adenosine 5'-triphosphate binding site inhibitors of the tyrosine kinase domain of the epidermal growth fa," *J. Med. Chem.*, vol. 38, no. 18, pp. 3482–3487, Sep. 1995.

- [112] A. J. Bridges, H. Zhou, D. R. Cody, G. W. Rewcastle, A. McMichael, H. D. Showalter, D. W. Fry, A. J. Kraker, and W. A. Denny, "Tyrosine kinase inhibitors. 8. An unusually steep structure-activity relationship for analogues of 4-(3-bromoanilino)-6,7-dimethoxyquinazoline (PD 153035), a potent inhibitor of the epidermal growth factor receptor.," *J. Med. Chem.*, vol. 39, no. 1, pp. 267–76, Jan. 1996.
- [113] G. Vriend, "WHAT IF: A molecular modeling and drug design program," *J. Mol. Graph.*, vol. 8, pp. 52–56, 1990.
- [114] and P. A. K. D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J., "AMBER 12." University of California, San Francisco, 2012.
- [115] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters.," *Proteins*, vol. 65, no. 3, pp. 712–25, Nov. 2006.
- [116] Accelrys Software Inc., "Discovery Studio Modeling Environment, Release 3.5." 2012.
- [117] J. Wang and R. Wolf, "Development and testing of a general amber force field," *J. ...*, 2004.
- [118] University of Southampton, "protoMS." 2012.
- [119] C. Yung-Chi and W. H. Prusoff, "Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction," *Biochem. Pharmacol.*, vol. 22, no. 23, pp. 3099–3108, 1973.
- [120] D. W. John Kuriyan, Boyana Konforti, *The Molecules of Life*, 1st ed. 2012, p. Ch 12.
- [121] S. Lazareno and N. J. Birdsall, "Estimation of competitive antagonist affinity from functional inhibition curves using the Gaddum, Schild and Cheng-Prusoff equations.," *Br. J. Pharmacol.*, vol. 109, no. 4, pp. 1110–9, Aug. 1993.
- [122] D. A. Pearlman and P. S. Charifson, "Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System †," *J. Med. Chem.*, vol. 44, no. 21, pp. 3417–3423, Oct. 2001.
- [123] J. Michel, "The Use of Free Energy Simulations as Scoring Functions," University of Southampton, 2006.

## Bibliography

---

- [124] T. Hou, L. Zhu, L. Chen, and X. Xu, "Mapping the binding site of a large set of quinazoline type EGF-R inhibitors using molecular field analyses and molecular docking studies.," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 1, pp. 273–87, 2003.
- [125] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics.," *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–8, 27–8, Feb. 1996.
- [126] M. Hartshorn, "OpenAstexViewer 3.0." 2013.
- [127] P. Politzer, P. Lane, M. C. Concha, Y. Ma, and J. S. Murray, "An overview of halogen bonding.," *J. Mol. Model.*, vol. 13, no. 2, pp. 305–11, Feb. 2007.
- [128] Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang, and W. Zhu, "Halogen bonding--a novel interaction for rational drug design?," *J. Med. Chem.*, vol. 52, no. 9, pp. 2854–62, May 2009.
- [129] J. J.-L. Liao, "Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors.," *J. Med. Chem.*, vol. 50, no. 3, pp. 409–24, Feb. 2007.
- [130] T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek, "Ensembl 2009.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D690–7, Jan. 2009.
- [131] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali, "Comparative protein structure modeling using Modeller.," *Curr. Protoc. Bioinforma.*, vol. Chapter 5, p. Unit 5.6, Oct. 2006.
- [132] R. W. Hooft, G. Vriend, C. Sander, and E. E. Abola, "Errors in protein structures.," *Nature*, vol. 381, no. 6580, p. 272, May 1996.
- [133] R. W. Hooft, C. Sander, and G. Vriend, "Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures.," *Proteins*, vol. 26, no. 4, pp. 363–76, Dec. 1996.
- [134] D. A. Case, T. Darden, T. E. C. Iii, C. Simmerling, S. Brook, J. Wang, U. T. Southwestern, R. E. Duke, U. Hill, R. Luo, U. C. Irvine, M. Crowley, R. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, U. C. S. Diego, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin,

- C. Sagui, N. C. State, V. Babin, P. A. Kollman, U. C. S. Francisco, D. A. Pearlman, R. V Stanton, J. Pitera, I. Massova, A. Cheng, P. State, J. J. Vincent, P. Beroza, V. Tsui, C. Schafmeister, W. S. Ross, R. Radmer, G. L. Seibel, J. W. Caldwell, C. Singh, P. Weiner, and P. Cieplak, "AMBER Users' Manual," pp. 1–304.
- [135] S. L. Williams, "The study of Conformational Motions using Enhanced Sampling Techniques," University of Southampton, 2007.
- [136] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD.," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–802, Dec. 2005.
- [137] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. D. Caves, "Bio3d: an R package for the comparative analysis of protein structures.," *Bioinformatics*, vol. 22, no. 21, pp. 2695–6, Nov. 2006.
- [138] R. C. Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing.
- [139] K. Sakai, H. Yokote, K. Murakami-Murofushi, T. Tamura, N. Saijo, and K. Nishio, "In-frame deletion in the EGF receptor alters kinase inhibition by gefitinib.," *Biochem. J.*, vol. 397, no. 3, pp. 537–43, Aug. 2006.
- [140] A. Amadei, M. A. Ceruso, and A. Di Nola, "On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations.," *Proteins*, vol. 36, no. 4, pp. 419–24, Sep. 1999.
- [141] J. Michel and J. W. Essex, "Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations.," *J. Comput. Mol. Des.*, vol. 24, pp. 639–658, 2010.
- [142] H. M. Senn and W. Thiel, "QM/MM methods for biomolecular systems.," *Angew. Chem. Int. Ed. Engl.*, vol. 48, no. 7, pp. 1198–229, Jan. 2009.