

**A computational method
for identifying allosteric binding sites in kinases**

A thesis submitted to the University of Strathclyde in fulfilment of the requirements
of the degree of Doctor of Philosophy

By

Nizar Ali Al-Shar'i

2013

University of Strathclyde

Strathclyde Institute of Pharmacy and Biomedical Sciences



Declaration of Authenticity and Author's Rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed:

Date:

Table of Contents

Topic	Page
Table of contents	iii
Table of figures	vi
Table of tables	xii
Abbreviations	xiii
Outline of the thesis	xvi
Acknowledgments	xviii
Abstract	xx
1 INTRODUCTION	1
1.1 Allosterism.....	2
1.1.1 Overview of allosteric regulation.....	2
1.1.2 Models of allosteric regulation: old and new views.....	4
1.1.3 Structural basis of allosteric regulation in proteins.....	8
1.1.4 Allosteric regulation in drug discovery.....	9
1.1.5 Methods for identifying allosteric sites.....	10
1.2 Molecular modelling.....	22
1.2.1 Overview.....	22
1.2.2 Quantum mechanical methods.....	24
1.2.3 Molecular mechanics.....	25
1.2.4 Treatment of solvent effect in MD simulations.....	43
1.2.5 Setting up and running a simulation.....	45
1.2.6 Use of grids in molecular modelling.....	47
1.2.7 Docking.....	49
1.2.8 Visualisation.....	51
1.3 Kinases.....	53
1.3.1 Overview.....	53
1.3.2 Structural biology and substrate binding of kinases.....	55
1.3.3 Regulation and conformational flexibility of kinases.....	57
1.3.4 Small molecule inhibitors of kinases.....	60
1.4 Rationale and objectives of the project.....	66
1.5 Future experimental plan.....	67
2 MATERIALS AND METHODS	68
2.1 Materials.....	69
2.1.1 Computational materials.....	69
2.1.2 Experimental materials.....	70
2.2 Computational methods.....	70
2.2.1 Molecular dynamics (MD) simulations.....	70
2.2.2 Simple Intrasequence Differences (SID) analysis.....	90
2.2.3 Correlations of energy fluctuations.....	91

2.2.4	The methods of virtual screening and docking	93
2.3	Experimental methods.....	97
2.3.1	Differential scanning fluorimetry (DSF).....	97
2.3.2	Inhibition and kinetics assays of DYRK2.....	97
3	PROOF OF CONCEPT (JNK1 AND CDK2)	99
3.1	c-Jun N-terminal Kinase (JNK-1).....	100
3.1.1	The role of JNKs.....	100
3.1.2	System stability and conformational flexibility	104
3.1.3	Analysis of correlated motion.....	121
3.1.4	Simple intrasequence difference analysis	129
3.1.5	Energy correlations	135
3.1.6	Summary of the overall results of JNK1.....	139
3.2	Cyclin-Dependent Kinase 2 (CDK2).....	141
3.2.1	The role of CDK2	141
3.2.2	System stability and conformational flexibility	145
3.2.3	Analysis of correlated motion.....	155
3.2.4	Simple intrasequence difference analysis	161
3.2.5	Energy correlations	167
3.2.6	Summary of the overall results of CDK2.....	170
4	CASE STUDY (DYRK2).....	172
4.1	The role of DYRK2	173
4.2	System stability and conformational flexibility	175
4.2.1	System stability.....	175
4.2.2	Conformational flexibility.....	178
4.3	Analysis of correlated motion.....	181
4.4	Simple intrasequence difference analysis	184
4.5	Energy correlations	185
4.6	Summary of the overall results of <i>DYRK2-apo</i>	188
4.7	Simulating the effect of ligand binding at the identified site.....	191
4.7.1	Analysis of the MD results of DYRK2-ligand complexes.....	195
4.8	Virtual screening and molecular docking	204
4.8.1	Structure-based pharmacophore generation.....	204
4.8.2	Virtual screening of commercial databases.....	207
4.8.3	Molecular Docking	208
4.9	Experimental evaluation of the binding affinities of the selected hits.....	222
4.9.1	Differential scanning fluorimetry (DSF).....	222
4.9.2	DYRK2 assays.....	231
5	CONCLUSION AND OUTLOOK.....	239
5.1	Concluding remarks	240
5.2	Outlook	244
6	APPENDICES.....	245
6.1	Appendix I: Relationship between force, potential energy, and positional changes as a function of time.....	246

6.2	Appendix II: minimisation, equilibration, and MD production input files	248
6.2.1	Minimisation input files	248
6.2.2	Equilibration input files	250
6.2.3	MD production input file	252
6.3	Appendix III: Input files used to analyse the generated trajectories	254
6.4	Appendix IV: conversion of SID scores or RMSF values into colour codes.....	257
6.5	Appendix V: Input files used to calculate the energy correlations	258
6.5.1	Perl scripts.....	258
6.5.2	Matlab input files	269
7	REFERENCES.....	270

Table of figures

Figure number	Page
Figure 1.1: Allosteric regulation of proteins	3
Figure 1.2: General model of allosteric regulation.	4
Figure 1.3: Schematic representation of MWC and KNF models of allostery	5
Figure 1.4: Schematic representation of the pre-existing equilibrium and post-binding rearrangement of proteins.	7
Figure 1.5: Allosteric regulation by phosphorylation	8
Figure 1.6: Determination of protein structure by x-ray crystallography.	11
Figure 1.7: The allosteric site in HIV-1 RT.	11
Figure 1.8: Determination of protein 3D structure using NMR.	13
Figure 1.9: A schematic representation of SID analysis	16
Figure 1.10: Graphical representation of SID score (HL).....	18
Figure 1.11: Comparative SID scoring	18
Figure 1.12: Illustration of the SCA through predicting the allosteric core of the G protein family.	20
Figure 1.13: Pictorial representation of the different terms incorporated in a MM force field.	27
Figure 1.14: Energy change with bond length.	29
Figure 1.15: Torsional profile for ethane as a function of torsional angle with an energy barrier of about 3 Kcal/mol, showing the periodicity of the potential energy on rotation around the carbon-carbon bond.....	30
Figure 1.16: Representation of the improper torsion and out of plane terms which are added to force fields to maintain planarity of unsaturated systems.....	31
Figure 1.17: The induced dipole and van der Waals attraction.....	31
Figure 1.18: A typical van der Waals curve, showing the attractive and repulsive terms.	32
Figure 1.19: The basic functional form of the force field most commonly used in AMBER 10 for macromolecular simulations.....	35
Figure 1.20: Schematic representation of the concept of energy minimisation.	36
Figure 1.21: The effect of energy minimisation on the overall geometry of the molecule. ...	37
Figure 1.22: A schematic representation of the potential energy landscape showing local and global energy minima.	39
Figure 1.23: This figure shows how MD can be used to explore the molecular energy landscape for stable conformations (ideally the global minimum).....	39
Figure 1.24: Treatment of solvent effects in MD.....	45
Figure 1.25: A schematic representation of the set up and the running of molecular dynamics simulations.	47
Figure 1.26: The use of grid to measure molecular properties.	48
Figure 1.27: The graphical interface of Discovery Studio 3.1 client	52
Figure 1.28: Phylogenetic tree of the human kinome highlighting the different subsets.....	54
Figure 1.29: The structure of the catalytic domain of kinases	55
Figure 1.30: The structure and nomenclature of a kinase ATP binding site.....	56

Figure 1.31: A 2D representation of the kinase ATP binding site	57
Figure 1.32: Schematic representation of the active and in active conformations of kinases.	59
Figure 1.33: Schematic representation of the different regions of the ATP binding site that are utilised by different kinase inhibitors.....	60
Figure 1.34: Chemical structure of sunitinib.	61
Figure 1.35: Chemical structure of Gleevec.	61
Figure 1.36: Chemical structure of compound AP23464.....	62
Figure 1.37: Chemical structure of BMS-006	62
Figure 1.38: Chemical structure of compound HKI-272	63
Figure 2.1: A simplified workflow in AMBER	71
Figure 2.2: A schematic representation of the workflow in ptraj.....	78
Figure 2.3: The starting crystal structures for the simulated JNK1 states.	82
Figure 2.4: The chemical structure of JNK1 allosteric inhibitor.	83
Figure 2.5: The starting crystal structures for the simulated CDK2 states.	84
Figure 2.6: The starting crystal structure of DYRK2-apo.....	86
Figure 2.7: The chemical structures of the two probes used to generate the starting structural models for the DYRK2-probe-1and DYRK2-com-6 respectively.....	86
Figure 2.8: The starting structural models for the DYRK2-complex states.....	88
Figure 2.9: A schematic description of the workflow of structure-based pharmacophore generation.....	94
Figure 2.10: The primary structure-based pharmacophore model generated based on the putative allosteric site in DYRK2.	95
Figure 3.1: Cartoon representation of the backbone structure of JNK1 complexed with pepJIP1.....	101
Figure 3.2: X-ray crystal structure of JNK-1 complexed with pepJIP1.	102
Figure 3.3: X-ray crystal structure of JNK-1 complexed with an allosteric inhibitor	103
Figure 3.4: Backbone atoms' RMSD versus time plot for the MD simulations of JNK1-apo and JNK1-allo using the starting structure of each trajectory as a reference.	104
Figure 3.5: Comparison between two trajectories of the JNK1-allo state where the pseudorandom seed generator (RSG) was activated in one of them but not in the other.	106
Figure 3.6: The effect of activating the random seed generator (RSG) and restarting the equilibration MD segments after each run on the quality of the generated trajectory.	107
Figure 3.7: Summary of the energy changes for the two JNK1 states versus time..	108
Figure 3.8: Summary of temperature changes for the two states of JNK1 plotted versus time	109
Figure 3.9: Comparison between the starting structures and the average structure of the two simulated states of JNK1.....	110
Figure 3.10: Comparison of the distance between the N and C-terminal domains in the starting and minimized average structures for JNK1-apo and JNK1-allo states.....	111
Figure 3.11: The elimination of rotational and translational interference with residual fluctuation by performing an RMS fitting to the starting crystal structure in each trajectory	112
Figure 3.12: Comparison of the RMSF values of the two simulated states of JNK1	113

Figure 3.13: Fluctuation difference between the two states of JNK1	114
Figure 3.14: Colour coded representation of the residual fluctuation values of the simulated states of JNK1.....	115
Figure 3.15: Comparison of the networks of hydrogen bonds in the minimised average structures of both states of JNK1.....	116
Figure 3.16: Comparison of the networks of hydrogen bonds in the starting structures of both states of JNK1	118
Figure 3.17: Comparison of the contact maps of the minimized average structures of JNK1-apo and JNK1-allo states.....	119
Figure 3.18: Structural mapping and comparison of the contact maps of the minimized average structures of the JNK1-apo and JNK1-allo states.	120
Figure 3.19: The residual cross-correlation matrices of the backbone atoms around their average position calculated from the simulated trajectories of the two JNK1 states and represented as heat maps.	123
Figure 3.20: Regions of correlated residues in the JNK1-apo state.....	125
Figure 3.21: The main anti-correlated region in JNK1-apo.....	127
Figure 3.22: The effect of binding the allosteric inhibitor on the correlated residues in JNK1-allo.....	128
Figure 3.23: Column chart of the individual SID scores of each residue position of the minimised average structures of the JNK1-apo and the JNK1-allo states	130
Figure 3.24: SID analysis of the minimized average structures of the two simulated states of JNK1.	131
Figure 3.25: Column chart of the RSD of SID scores of each residue position of all the extracted frames from the JNK1-apo and the JNK1-allo trajectories.	133
Figure 3.26: SID analysis of all frames extracted from the trajectories of the two simulated states of JNK1.....	134
Figure 3.27: The correlation peaks that define the interaction pathways in JNK1.....	137
Figure 3.28: Structural mapping of the residues that define the interaction pathways in JNK1-apo and JNK1-allo states.....	138
Figure 3.29: Summary of main computational results for JNK1.	139
Figure 3.30: Cartoon depiction of the backbone atoms of the catalytic domain of the CDK2.	141
Figure 3.31: The structural changes associated with cyclin-A binding.	142
Figure 3.32: The activation sequence of CDK2	143
Figure 3.33: The chemical structure of ANS that binds to the allosteric site of CDK2.....	143
Figure 3.34: The allosteric binding site in CDK2 and the conformational changes in the α C-helix induced by ANS binding.....	144
Figure 3.35: The RMS deviation of the backbone atoms of CDK2-apo and CDK2-ATP from their minimised starting structures.	145
Figure 3.36: Summary of the energy changes for the two CDK2 states versus time.	146
Figure 3.37: Summary of temperature changes for the two CDK2 states plotted versus time	147

Figure 3.38: Superimposition of the minimised average and the starting structure of the CDK2-apo system and the CDK2-ATP.....	148
Figure 3.39: Surface representation of the rotated N-domain in figure 3.38.	149
Figure 3.40: Residual fluctuations of CDK2.....	150
Figure 3.41: The effect of ATP binding on enzyme flexibility.	151
Figure 3.42: Colour coded representation of the residual fluctuation values of the two simulated states of CDK2.	152
Figure 3.43: The contact maps of the two minimised average structures of CDK2.	153
Figure 3.44: Comparison of the contact maps of the CDK2-apo and CDK2-ATP states.....	154
Figure 3.45: Residual correlation analysis of the two simulated states of CDK2.....	156
Figure 3.46: Analysis of the correlated residues in the CDK2-apo state.	157
Figure 3.47: The interactions between cluster I and other components of the phospho-CDK2-cyclin-A complex.	159
Figure 3.48: Analysis of the correlated residues in the CDK2-ATP state.	160
Figure 3.49: Column chart of the SID scores of each residue position of the minimised average structures of the CDK2-apo and the CDK2-ATP states.....	162
Figure 3.50: SID analysis of the minimised average structures of the two simulated states of CDK2.....	163
Figure 3.51: Column chart of the RSD of SID scores of each residue position of all the frames extracted from the CDK2-apo and CDK2-ATP trajectories	165
Figure 3.52: SID analysis of all extracted frames from the two CDK2 simulated trajectories.	166
Figure 3.53: The peaks of energy correlation that define the interaction pathways in CDK2.	168
Figure 3.54: Structural mapping of the residues that define the interaction pathways in CDK2-apo and CDK-ATP states.....	169
Figure 3.55: Summary of the main computational results of CDK2	171
Figure 4.1: The structure of DYRK2.....	174
Figure 4.2: Backbone atoms' RMSD versus time plot for the MD simulation of DYRK2-apo state using the starting structure of the trajectory as a reference.	176
Figure 4.3: Summary of the energy and temperature changes during the simulation of DYRK2-apo	177
Figure 4.4: Comparison between the starting structure and the average structure of the simulated DYRK2-apo state.....	178
Figure 4.5: The residual fluctuation of DYRK2-apo plotted as RMSF versus residue index.	179
Figure 4.6: Colour coded representation of the residual fluctuation values of DYRK2-apo.....	180
Figure 4.7: The matrix of the residual cross-correlation motion of the backbone atoms around their average position calculated from the simulation trajectory and represented as a heat map.	181
Figure 4.8: Structural mapping of the concentrated residue-residue coupling areas in DYRK2-apo heat map	183
Figure 4.9: Column chart of the SID scores of each residue position of the minimised average structure of the DYRK2-apo state	184

Figure 4.10: SID analysis of the minimized average structure of DYRK2-apo.....	185
Figure 4.11: The energy correlation peaks that define the interaction path in DYRK2.....	186
Figure 4.12: Structural mapping of the residues that define the interaction pathways in DYRK2-apo	187
Figure 4.13: Summary of the overall results obtained from applying different computational analysis in studying DYRK2-apo.	189
Figure 4.14: Design of the small molecule probe..	191
Figure 4.15: The docked pose used as a starting structure for the DYRK2-probe-1 complex MD simulations.	192
Figure 4.16: The detachment of probe-1 from the binding site after 7 ns.....	193
Figure 4.17: The chemical structures of the more flexible and complex probe, probe-2..	193
Figure 4.18: The binding mode of probe-2.....	194
Figure 4.19: The chemical structures of the derivatives that were obtained by substituting the pyridine ring in probe-2.....	194
Figure 4.20: The docked pose of compound 6 used as a starting model for the second artificial DYRK2-ligand complex (DYRK2-com-6) MD simulation	195
Figure 4.21: Comparison of the RMSDs values of the two simulated states of DYRK2.	196
Figure 4.22: Cartoon representation of the backbone atoms of the minimised average structures of the DYRK2-apo and DYRK2-com-6 showing the regions of major conformational changes resulting from ligand binding. Compound 6 is depicted in ball and stick.....	197
Figure 4.23: Comparison of the residual fluctuations of the simulated complex with that of the DYRK2-apo state.	198
Figure 4.24: Structural mapping of the regions that experienced a reduction in their RMSF values upon ligand binding.	199
Figure 4.25: The dynamic cross-correlation matrices represented as heat maps. DYRK2-apo state is shown in the left and DYRK2-com-6 in right.	200
Figure 4.26: Comparison of the energy correlations (represented as peaks in the plots) of the DYRK2-apo state and DYRK2-com-6 model.....	201
Figure 4.27: Structural mapping of the residues that define the interaction pathways in DYRK2-apo and DYRK2-com-6	202
Figure 4.28: DS definition of the binding site	204
Figure 4.29: Structure based pharmacophore generation..	205
Figure 4.30: The generated structure-based pharmacophore model from the putative allosteric site in DYRK2.....	205
Figure 4.31: The three sub-pharmacophores obtained by fragmenting the primary pharmacophore.	207
Figure 4.32: Overlay of the first retrieved hit from the Maybridge DB defined by each pharmacophore	207
Figure 4.33: A pipeline showing the filtration of the retrieved hits based on Lipinski's rule of five and a fit value of more than two using Pipeline Pilot.....	208
Figure 4.34: GOLD definition of the binding site used for docking the filtered hits that were retrieved from the Maybridge database.	209

Figure 4.35: Pruning of the hydrophobic fitting points to fit the site of interest.	209
Figure 4.36: The top ranked docked pose of the first ten retrieved hits.....	210
Figure 4.37: The virtual screening workflow.....	211
Figure 4.38: Schematic representation of the principle of the DSF method.....	223
Figure 4.39: Schematic representation of protein stabilisation upon the addition of a compound that binds the native protein.....	224
Figure 4.40: The DSF results for the 48 selected hits against DYRK2.	225
Figure 4.41: The chemical structures of the four compounds that showed good binding affinities in the DSF screening.....	225
Figure 4.42: Analysis of the binding modes of the four (A-D) differential scanning fluorimetry (DSF) hits in the putative binding site.....	227
Figure 4.43: Analysis of the binding modes of the four (A-D) differential scanning fluorimetry (DSF) hits in the ATP binding site.....	229
Figure 4.44: Comparison of the K_m values for ATP (recombinant DYRK2) obtained by using the purchased kit with that reported in the user manual of the manufacturer	231
Figure 4.45: Comparison of the inhibitory effect of staurosporine on the activity of recombinant DYRK2 obtained when conducting the inhibition assay with the results reported in the kit's user manual.	232
Figure 4.46: The concentration-response curves of RF05199 and RJC03509 at 125 μ M ATP.	232
Figure 4.47: Saturation curve for an enzyme showing the relation between the initial reaction rate (V_0) and substrate concentration [S], from which V_{max} and K_M can be determined.	233
Figure 4.48: The reaction schemes of the three classes of reversible enzyme inhibition...	236
Figure 4.49: Saturation curve plots for the six concentrations of RF05199 relative to the reference to determine the kinetic parameters and the type of inhibition.....	237
Figure 5.1: A scheme summarising the different steps that led to the identification of a non-competitive DYRK2 inhibitor.....	241
Figure 5.2: Interfacial connectivity in DYRK2. The major interfaces in DYRK2 are represented as solid surface around the backbone of the protein and within a mesh surface. It is clear that they are establishing a network of contacts within the 3D structure of the protein; hence disruption of the stability of any of these interconnected interfaces will propagate to distant loci in the protein.....	243

Table of tables

Table number	Page
<i>Table 1.1:</i> Docking programs commonly used in virtual screening.....	51
<i>Table 1.2:</i> Small molecule kinase inhibitors approved by the FDA for cancer treatment.....	64
<i>Table 2.1:</i> Summary of all simulated systems.....	89
<i>Table 3.1:</i> Statistical descriptors of the JNK1 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.....	130
<i>Table 3.2:</i> Statistical descriptors of the RSD of JNK1 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.....	133
<i>Table 3.3:</i> Statistical descriptors of the CDK2 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.....	162
<i>Table 3.4:</i> Statistical descriptors of the RSD of CDK2 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.....	165
<i>Table 4.1:</i> Statistical descriptors of the DYRK2 SID scores used to guide the selection of residues contributing to complex interfaces in the protein that are of possible allosteric potential.....	184
<i>Table 4.2:</i> The 48 compounds purchased for experimental testing of their binding affinities.....	212
<i>Table 4.3:</i> Summary of the number and types of interactions between the four hits and each of the putative and ATP binding sites.....	230
<i>Table 4.4:</i> The initial rates of RF05199SC at five different concentrations for eight different concentrations of ATP relative to the reference of no inhibitor.....	237
<i>Table 4.5:</i> The kinetic parameters for the DYRK2 kinase that were obtained from the saturation curve plots	238

Abbreviations

Abbreviation	Full name
3D-QSAR	3D quantitative structure activity relationships
ADP	Adenosine diphosphate
AM1	Austin model 1
AMBER	Assisted Model Building with Energy Refinement
ANS	8-anilino-1-naphthalene sulfonate
ATCM	Allosteric ternary complex model
ATP	Adenosine triphosphate
BO	Born-Oppenheimer
BPG	2,3-bisphosphoglycerate
CAK	CDK-activating kinase
CCDC	Cambridge crystallographic data centre
CDK	Cyclin-dependent kinase
CFE	Consistent force field
CG	Conjugate gradients
CMview	Contact map view
CPU	Central processing unit
CTP	Cytidine triphosphates
DB	Database
DFG motif	Aspartate, phenylalanine and glycine
DIFF	Differential
DMSO	Dimethyl sulfoxide
DS	Discovery studio
DSF	Differential scanning fluorimetry
DYRK2	Dual specificity tyrosine-phosphorylation-regulated kinase 2
ELISA	Enzyme-linked immunosorbent assay
ENM	Elastic network model
F16BPase	Fructose-1,6-bisphosphatase
FDA	Food and drug administration
GA	Genetic algorithm

Abbreviation	Full name
GAFF	General AMBER force field
GCT	CTP:glycerol-3-phosphate cytidyltransferase
GG	Greatest gap
GlyP	Glycogen phosphorylase
GNM	Gaussian network model
GOLD	Genetic Optimization for Ligand Docking
GPCR	G protein coupled receptors
GPU	Graphical processing unit
HL	Highest-lowest
HRP	Horseradish peroxidase
HSQC	Heteronuclear signal-quantum correlation
IQD	Inter-quartile distance
JIP1	JNK-interacting protein-1
JNK-1	C- Jun N-terminal protein kinase
K_M	Michaelis-Menten constant
KNF	Koshland, Nemethy and Filmer model
LBDD	Ligand based drug design
LQ	Lower quartile
MKK	Mitogen-activated protein kinase kinase
MD	Molecular dynamics
MM	Molecular mechanics
MMFF	Merck molecular force field
MQ	Median quartile
MSA	Multiple sequence alignment
MWC	Monod, Wyman and Changeux model
NGS	National grid service
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NR	Newton-raphson

Abbreviation	Full name
OS	Operating system
PDB	Protein databank
PDF	Probability density function
PES	Potential energy surface
PM3	Parametric method 3
PP	Pipeline pilot
QM	Quantum mechanics
RMS	Root mean squared
RMSF	Root mean square fluctuations
RSD	Relative standard deviation
RSG	Pseudorandom seed generator
RTP	Room temperature
SBDD	Structure-based drug design
SBP	Structure-based pharmacophore
SCA	Statistical coupling analysis
STD	Standard deviation
SD	Steepest descents
SGC	Structural genomic consortium
SID	Simple intrasequence differences
STP	Surface triplet propensities
T_m	Melting temperature
TMB	Tetra-methylbenzidine
TOCSY	Total correlation spectroscopy
UQ	Upper quartile
vdW	Van der Waals
V_{max}	Maximum rate
VS	Virtual screening
ΔG_u	Gibbs free energy of unfolding

Outline of the thesis

This thesis is organised in six chapters as follows:

Chapter 1: this chapter includes three main sections. The first section presents a general overview of the concept of allostery highlighting the old and new perception of this phenomenon, with coverage of the main experimental and computational methods that are used in the identification of new allosteric binding sites in proteins.

The second section discusses in more details the different tools and techniques used in the field of molecular modelling, particularly molecular dynamics (MD) simulations.

The third section focusses on the kinases, their structural biology, regulation and inhibition. These are the proteins that are the subject of the computational approach that is being applied to explore for allostery.

Chapter 2: this chapter lists the different materials and outlines all the methods used in this study. The procedures of the computational methods MD simulations, simple intrasequence difference (SID) analysis, energy correlations and virtual screening are presented in detail. Differential scanning fluorimetry and the enzyme assays used to measure binding *in vitro* are also detailed.

Chapter 3: the main aim of this study was to develop a computational method that can be used to search for allosteric binding sites in proteins. This chapter discusses in detail the different results which were obtained by applying our computational approach to studying JNK1 and CDK2. These two kinases have experimentally identified allosteric binding sites and were used for proof of concept purposes in order to evaluate the accuracy and the validity of our approach. MD simulations of different states of each kinase were performed to generate trajectories that describe the evolution of systems with time, from which functional dynamic information was extracted. Protein topology was also analysed by SID and the main interfaces in the two proteins were identified. The energy correlations between protein residues were

calculated and the energetic interaction pathways were identified. Results in this chapter showed very good agreement between the computational predictions of potential allosteric sites and those identified experimentally in both systems.

Chapter 4: this chapter represents a real case study, in which our computational approach was applied to study DYRK2 (which is a kinase with no previously identified allosteric sites). We present and discuss all computational and experimental data that led to the identification of a putative allosteric and a non-competitive inhibitor of DYRK2.

Chapter 5: in this final chapter, general conclusions are drawn along with proposals for future work.

Chapter 6: appendix

Acknowledgments

First of all I want to thank my supervisors Prof. Simon MacKay and Dr. Balir Johnston. I am very grateful for their support, encouragement, guidance and for giving me the freedom to pursue various ideas without objection during these past three years.

My deepest thanks also go to Dr. Nahoum Anthony for his continuous support and help throughout my PhD time; the same goes to Dr. Rachel Clark and Dr. Murray Robertson for their help in the start of my study.

Also I would like to thank Prof. Stefan Knapp and Dr. Jon Elkins at the SGC in Oxford University for conducting the DSF screening; Dr. Mark Dufton and Christopher Foley for performing the SID scoring; and Louise Young for conducting the enzyme assays.

I would like to thank all my colleagues in SIPBS: Bilal, Giacomo, Jude, Sabin, James, Jessica, Murad, Mohammad, Saud, Khalid, Osama, Ibrahim, Rabab, Ahmad, Sarah, Jordan, Adel, Tony and Chris for their friendship, support and lovely spirits.

I also thank my friends (Ali, Raid, Jehad, Muath, Omar, Hossam, Ahmad, Osama, Yousef, Hussain, Abdul rahman, Rami and Abdullah) for providing friendship and support that I needed.

Finally, my deepest thanks go to my family in Jordan for all their love, encouragement and support; especially to my Mom and Dad for their sincere prayers.

Thank you all for helping me to make this dream come true. May Almighty Allah bless you all.

*I dedicate this thesis to
my family for their unlimited support, unconditional love and sincere prayers.*

I love you all.

Abstract

There are 518 protein kinases encoded within the human genome [1] that control cellular signal transduction and play a major role in almost all cellular events. Aberrant kinase activity is linked to pathological conditions including cancer, inflammation, diabetes and many others, making them a tractable target for drug discovery research [2-5]. To date, most of the current medicinal chemistry efforts target the ATP binding site, which is highly conserved amongst the kinase family, and many compounds suffer from cross-activity leading to undesirable side effects and toxicity. The ability to target allosteric sites on the catalytic domain of kinases, which are less conserved compared to ATP binding sites, would therefore provide an avenue for greater selectivity.

Here we propose a computational approach to identify allosteric sites in target kinases. We use a combination of molecular dynamics (MD) simulations to explore the critical structural and dynamic conformational changes of the enzymes and simple intrasequence differences (SID) analysis which identifies the major interfaces in the enzyme that may be involved in allosteric modulation. This computational approach provides not only a new method of identifying allosteric sites but also a better understanding of the mechanisms of allosteric modulation of target kinases and the structural basis for the design and development of more selective and specific small molecules inhibitors as therapeutic agents.

Chapter one

1 INTRODUCTION

1.1 Allosterism

1.1.1 Overview of allosteric regulation

The interest in allosterism is increasingly attracting more attention because of its crucial role in living cells. It controls almost all metabolic processes and gene regulation; regulates catalytic activities; protein and ligand transport; and coordinates enzymatic and signaling pathways [6, 7]. The term *allosterism* was first introduced by Monod and Jacob in 1961 in reference to the biological process of regulating protein function through binding of effector molecules to sites in the protein that are topographically distinct (*allosteric sites*) from the active site leading to a functional change in the latter via conformational and (or) dynamical modifications. The allosteric site can be directly adjacent or remote from the active site [6, 8-13], and the conformational changes in the active site that are usually associated with binding an allosteric effector are referred to as *allosteric transitions* [10, 14]. The origin of the word allosterism came from the Greek *allos* (other or different) and *stereos* (solid or object); in reference to the fact that the allosteric site is topographically distinct from the active site, and because the concept was originally used to describe protein conformational “shape” changes associated with effector binding [8, 9].

The binding of an allosteric effector, which can be a small molecule or a macromolecule, to proteins (e.g. enzymes) may have either an activating or inhibitory effect depending on whether it increases or decreases the protein’s activity or affinity for its substrate. For example, the binding of oxygen to hemoglobin is *allosteric activation* since the binding of the first oxygen molecule enhances the affinity for other oxygen molecules [13, 15]. On the other hand, *allosteric inhibition* decreases the activity or affinity for the substrate; as when 2,3-bisphosphoglycerate (BPG) binds to hemoglobin which causes stabilization of the hemoglobin in its inactive state thereby decreasing its affinity to oxygen (its natural substrate). Another example of allosteric inhibition is the locking of caspase-7 in the inactive zymogen conformation by the allosteric inhibitor DICA [16]. A schematic representation of the allosteric regulation of proteins (e.g. enzymes) is shown in figure 1.1.

Allosteric regulation of proteins could be homotropic or heterotropic depending on the chemical nature of the allosteric effector. If the protein is regulated by two

chemically identical ligands, such as when the enzyme is regulated by its own substrate, it is *homotropic allostery*. Conversely, *heterotropic allostery* corresponds to regulation of protein activity by two chemically different ligands [8, 12, 15]. A good example that shows both types of allosteric regulation is haemoglobin, where homotropic allostery is demonstrated by the binding of the first oxygen molecule which enhances the affinity for other oxygen molecules (the ligand and the substrate have the same chemical structure); and heterotropic allostery is demonstrated by binding of BPG which decreases the affinity to oxygen (chemically different from oxygen).

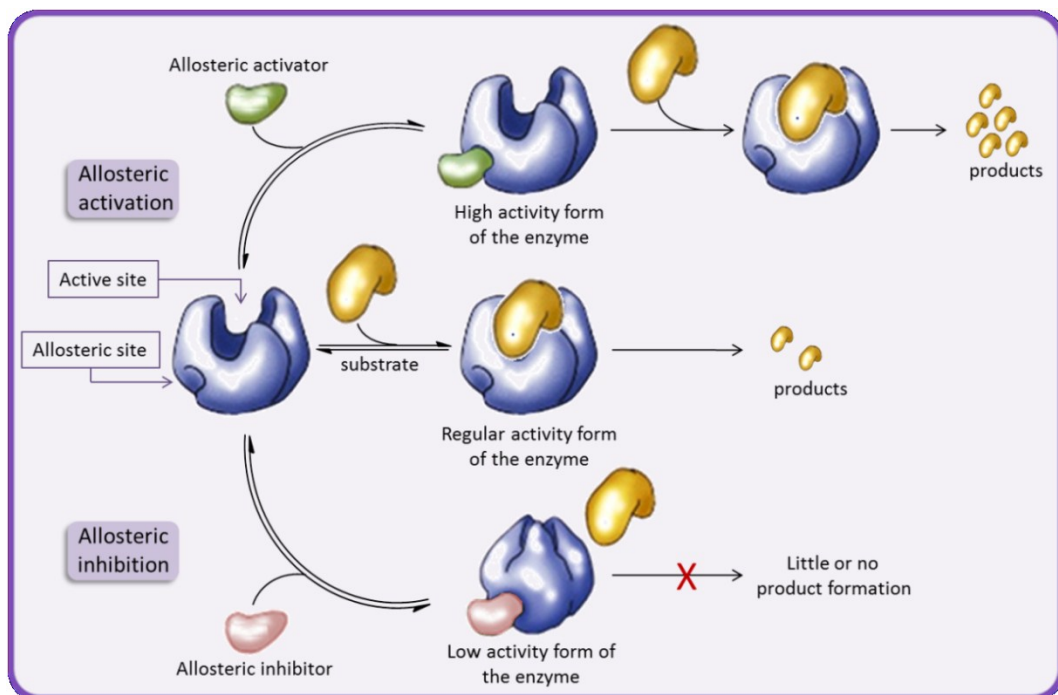


Figure 1.1: Allosteric regulation of proteins. An allosteric effector binds to a site in the protein that is distinct from the active site. This binding alters the conformation (and/or dynamics) of the active site turning it off (or on); therefore preventing (or enhancing) the binding of substrate molecule. The process of binding the allosteric effector and the substrate is under equilibrium control. Modified from [17] and [18].

Another term that is frequently associated with allostery is protein *cooperativity*. Cooperativity can be regarded as being a special case of allostery between binding sites, especially in *polymeric proteins* [19]. In general, allostery and cooperativity are

considered as being subclasses of the same phenomenon [8, 12]. Hill's coefficient, H , although not ideal [19, 20], is still used as a quantitative measure of cooperativity [10, 19], where a positive cooperativity ($H > 1$) is observed when ligand binding to one site of an enzyme enhances the affinity of other sites for another ligand. Conversely, a negative cooperativity is observed when the binding of a ligand at one site decreases the affinity of other sites [8, 19]. The binding of oxygen and BPG to hemoglobin are, respectively, the classical examples of positive and negative cooperativity in a polymeric protein. Another example of negative cooperativity in a dimeric protein is binding of the second cytidine triphosphates (CTP) ligand to the enzyme CTP:glycerol-3-phosphate cytidylyltransferase (GCT) [21]. To understand allostery and cooperativity and the underlying mechanisms by which they interact, it is important to identify the residues involved in the communication process between allosteric and active sites [22].

1.1.2 Models of allosteric regulation: old and new views

1.1.2.1 The old (classical) view of allostery

To account for allosteric mechanisms, two models were proposed in the 1960's, and both were applicable to oligomeric proteins of identical monomers or subunits. Both assumed that protein subunits exist in either a tensed (T) or relaxed (R) state, with the R state being of higher affinity to ligand binding than the T state [13]. The difference between the two models relates to how subunits interact in different conformations [19]. Figure 1.2 shows a general model of allosteric regulation. The enzyme (dynamic protein) exists in equilibrium between low-energy conformations, active and inactive, and ligand binding affects enzyme activity by shifting this equilibrium [11].

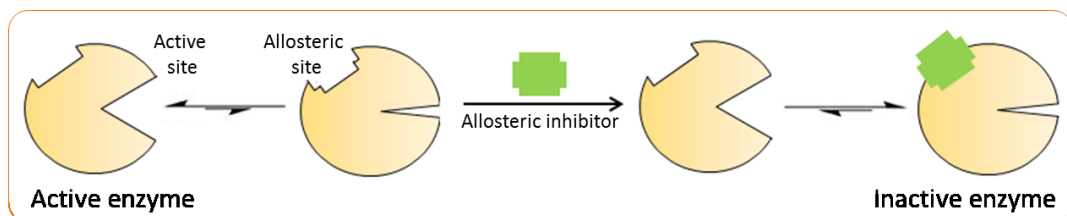


Figure 1.2: General model of allosteric regulation.

The first model is the *concerted or symmetry model* which was proposed by Monod, Wyman and Changeux (MWC). This model postulates that all subunits in a protein must be in the same conformational state (R or T) in such a way that a conformational change in one subunit is essentially causing an equivalent change in all other subunits, an all-or-non transition. The equilibrium between the two states can be shifted to the R or T state upon binding a ligand [13, 15, 23] (figure 1.3). The other model is the *sequential model* which was proposed by Koshland, Nemethy and Filmer (KNF). The sequential model differs from the MWC model in that the protein's subunits are not necessarily in the same state; binding occurs via induced fit, and conformational change in one subunit does not propagate to all other subunits, it just increases substrate affinity in the adjacent subunit. Ligand binding thereby causes the protein to undergo a transition of sequential structural rearrangements rather than an all-or-non transition [13, 15, 24] (figure 1.3).

The old view can be then described by two main characteristics: the first is there are only two states of the protein (T and R); and the second is there is a conformational change in the active site associated with ligand binding at the allosteric site.

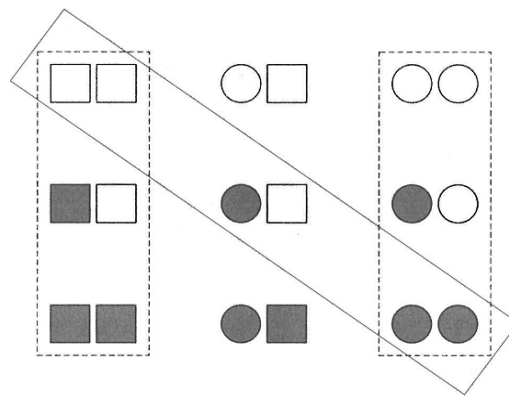


Figure 1.3: Schematic representation of MWC and KNF models of allostery for a homodimer. The left and right dashed columns represent the symmetric MWC model (only T or R). The diagonal box represents the sequential KNF model (hybrid of T and R). Squares and circles are T and R states respectively, and filled and empty symbols are ligated and non-ligated states [8, 19].

1.1.2.2 The new view of allostery

The accumulation of recent allosteric observations in different proteins has changed the old view of allostery; for example, allosteric behavior in single domain proteins has changed the old view that allostery is only observed in multidomain proteins [15].

Current evidence indicates that the flexibility of proteins makes them exist in the native state as a conformational ensemble; of particular interest are conformations that are accessible near the energy minimum which are called substates when separated by low energy barriers. Furthermore, within this conformational ensemble that the allosteric proteins sample there are conformations that are approximate to those of biological function, even in the absence of the allosteric effector. Binding of the allosteric effector causes a shift of the pre-existing conformations, known as a *population shift*, towards a different ensemble of conformations which entrap it in a certain state [7, 13, 15, 16, 25]. Subsequently, the binding may be optimized by induced fit [25]. This changes the old view of just two states R and T. Figure 1.4 shows a schematic representation of the pre-existing equilibrium of the ensemble of conformations that are accessible in the native state of a protein in its energy minimum, involving passage to substates as they pass the low energy barriers; the population shift that results from substrate binding as it selects the conformation that fits it the most (conformational selection); and the final optimization of the binding by induced fit. Note that the energy landscape has changed after binding the substrate shifting the population of the substates towards the conformer that fits it.

Recent experimental data have shown that in cases where the active and inactive forms of the protein show little or no conformational differences, allosteric mechanisms will not be identified by structural studies: the allosteric effect will not necessarily be mediated by a conformational change (at the backbone level) but rather by a change in protein's dynamics [6, 13], and this has changed the old view that allostery must involve a conformational change of the active site. Thus, all dynamic proteins can be considered as being potentially allosteric [15]. This can be fortified by the emerging thermodynamical definition of allostery in which proteins are divided into three groups based on whether the allosteric effect is driven by

entropy (type I), entropy and enthalpy (type II), or enthalpy (type III). According to this classification, the conformational changes associated with binding allosteric effectors are ranging from none or subtle backbone changes in type I; minor in type II; to large changes in type III [6, 26].

The new view of allostery can then be described by two characteristics: firstly, that proteins exist in ensembles rather than only two conformational states; secondly, that allostery is a thermodynamic process which can be controlled by enthalpy, entropy, or a combination of the two. There may not be conformational changes; therefore absence of conformational changes does not imply that allostery is not in play.

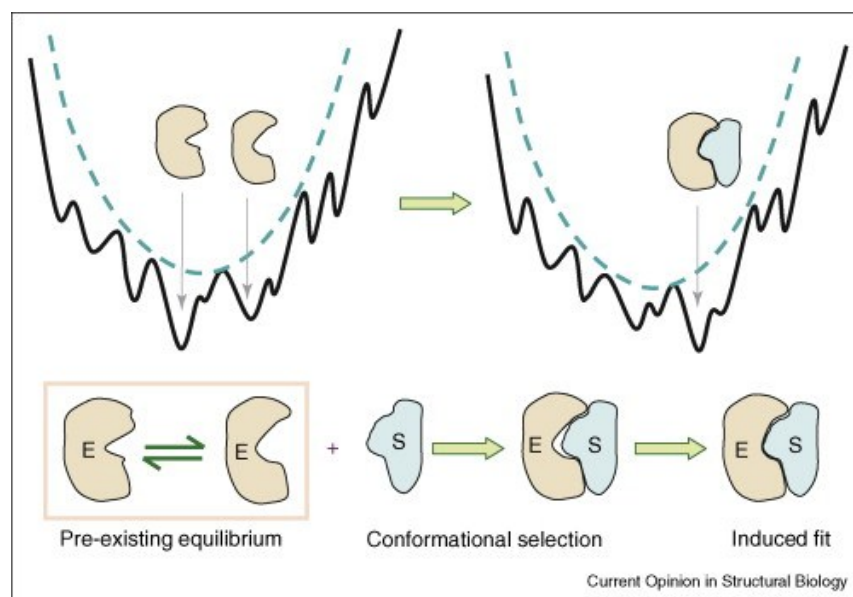


Figure 1.4: Schematic representation of the pre-existing equilibrium and post-binding rearrangement of proteins. Within the energy well (top panel), which is approximated by a harmonic potential (dashed line), the protein (E) samples an ensemble of conformations as it crosses small energy barriers. Two conformations are shown which are in equilibrium before substrate (S) binding; the substrate selects and binds to the conformer that allows for optimal interactions which in turn shifts the equilibrium [25].

1.1.3 Structural basis of allosteric regulation in proteins

Allosteric perturbation may result from binding of an effector, either a small molecule or a macromolecule; changes in the medium of the protein such as changes in pH, temperature, or ionic strength; also it may result from phosphorylation, glycosylation and other similar covalent modifications [26]. The following section discusses the structural basis of allosteric regulation:

Allosteric regulation by small-molecule binding

Small molecules are attracted to allosteric cavities because the latter are rich in functional groups suitable for binding [16]. This kind of binding is the most common form of allosteric regulation, and it modulates protein activity through different ways such as [13]: opening or closing of the active site; change of the active site conformation; and allosteric control of complex formation.

Allosteric regulation by phosphorylation

The most subtle trigger for allosteric regulation is possibly by phosphorylation [13]. For example phosphorylation of Ser14 in mammalian glycogen phosphorylase 45Å away from the active site causes a 10° rotation of the dimer's subunits, which activates the enzyme by opening up the catalytic site (figure 1.5) [16, 27].

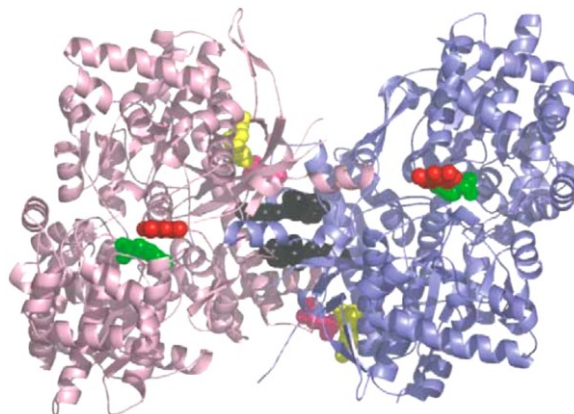


Figure 1.5: Allosteric regulation by phosphorylation. Enzyme monomers (pink and blue), the active site (red), substrate (green) and Ser14 phosphorylation site (bright pink), (black and yellow) are other regulatory sites [16].

Allosteric regulation by disulfide bonds

Protein conformation can be changed following breakage or formation of a disulfide bond in a site distinct from the active one [13, 28].

Allosteric regulation by protein binding

Proteins can also allosterically modulate the function of other proteins. For example, protein cyclin controls the function of cyclin-dependent kinase by displacing the CDK's activating segment making the active site accessible for binding ATP, and exposing Thr160 to phosphorylation by CDK-activating kinase (CAK) for full activation of the enzyme [13].

Allosteric ternary complex model (ATCM)

This is mainly seen in GPCRs where binding of the allosteric modulator alters ligand affinity for the active site and in the absence of the natural ligand the allosteric modulator alone will mediate no effect. Binding of either ligand is affected by concentration, equilibrium dissociation constant and cooperativity factor [29].

1.1.4 Allosteric regulation in drug discovery

Any misregulation of protein function, especially those involved in signalling pathways, will often result in disease. Generally drug design efforts focus on developing compounds that bind to the active site of the protein. Since the topology and homology of active sites are often conserved, these competitive inhibitors frequently lead to side effects. On the contrary, allosteric sites tend to be much more specific because they are often less conserved [30]. From the standpoint of drug design, allosteric modulation of protein activity has many advantages such as the potential for greater protein subtype selectivity and the ability to selectively tune responses in specific tissues [15]. Allosteric drugs can also target different conformational states of the protein [30]. The main problem in allosteric drug design is finding the allosteric site and then designing a ligand that binds to it. Historically, all clinically used drugs that act via an allosteric mechanism have been discovered initially by serendipity. The following sections introduce existing approaches that may be utilised when searching for allosteric sites.

1.1.5 Methods for identifying allosteric sites

1.1.5.1 Experimental approaches

1.1.5.1.1 X-ray crystallography

X-ray crystallography is one of the main methods for protein structure determination, and has greatly influenced structure based drug design. A 3D structure from a protein crystal is generally obtained via applying the following steps [31] (figure 1.6): protein purification, protein crystallization, crystal mounting, diffraction analysis and data collection, construction of electron density map, data processing and finally structure refinement and model building. Usually, new protein crystals are tested firstly on an in-house x-ray generator, then the best crystals are tested in a synchrotron source which produces a very intense monochromatic x-ray beam with a wave length in the range of 0.7-1.5 Å which is equivalent to the interatomic distances in molecules. These synchrotrons have high quality optics giving rise to high signal to noise ratio of the diffraction image, thereby producing highly resolved structures [31-33].

The use of x-ray crystallography to determine the binding site of an allosteric effector is usually preceded by high-throughput screening to identify a compound that binds followed by rigorous kinetic experiments to determine if the mechanism of inhibition is allosteric [16]. The major limitation of crystallography is its static nature (proteins are naturally dynamic) [28] and the crystal packing effect on protein structure that can influence backbone conformations, hinge-like motions and side-chain conformations [34]. However, it still has the major advantage of having virtually no size limitations compared to NMR techniques [35]. Many allosteric sites have been discovered by this method such as human liver glycogen phosphorylase (GlyP), fructose-1,6-bisphosphatase (F16BPase), glucokinase, and HIV-1 reverse transcriptase [16]. Figure 1.7 shows the allosteric and active sites of HIV-1 RT [16].

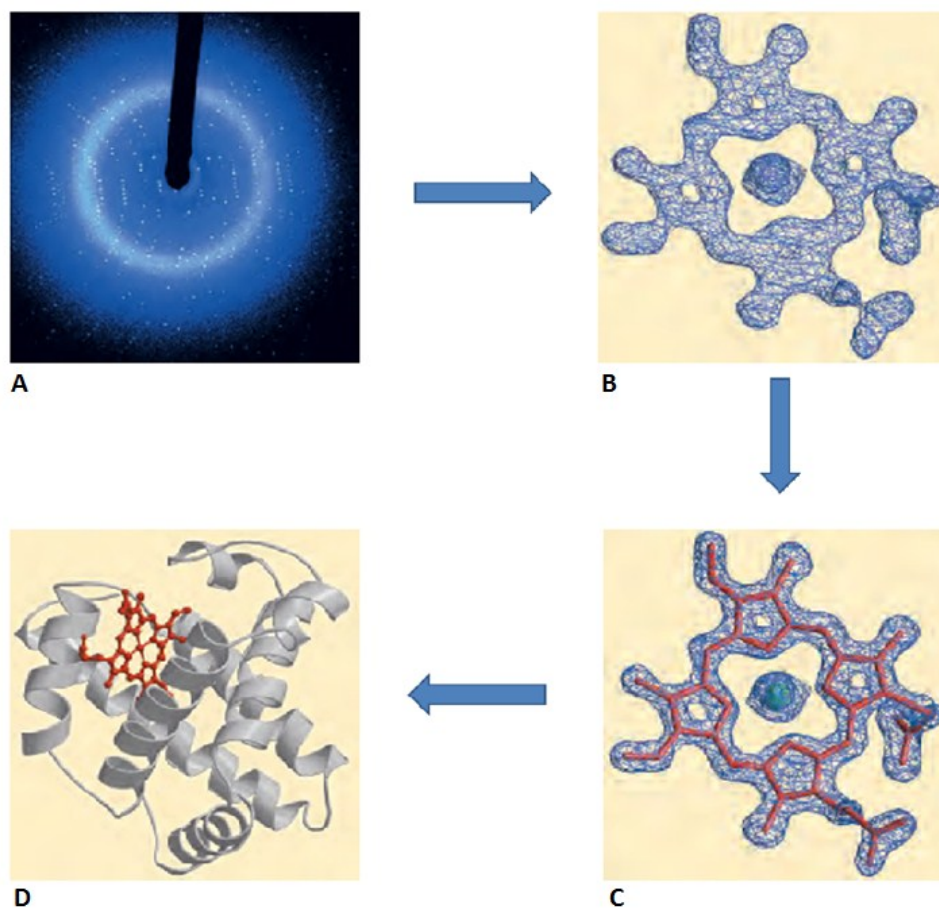


Figure 1.6: Determination of protein structure by x-ray crystallography; (A) the x-ray diffraction pattern which is converted to (B) the electron density map. (C) Analysis of the electron map will locate the mean position of atoms (greatest electron density) to generate (D) the completed 3D structure [36].

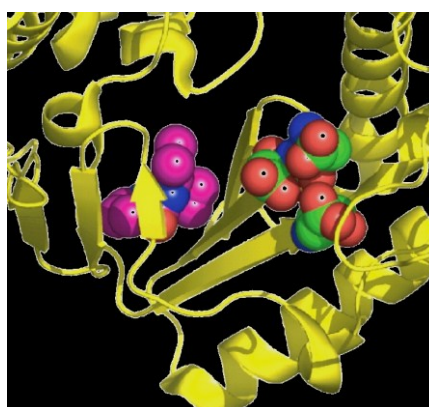


Figure 1.7: The allosteric site in HIV-1 RT (pink spheres) is distinct from the active site (green and red spheres) [16].

1.1.5.1.2 Nuclear Magnetic Resonance (NMR)

NMR is an important complementary method to x-ray crystallography for determining 3D protein structures [36]. It is also an excellent method for identifying those residues involved in protein-protein and protein-ligand interactions (allosteric sites) [37], such as the spin-labelled adenine analogue technique that is used to identify allosteric sites in kinases [38]. The major advantage of this method is the ability to conduct the analysis in solutions approximating physiological conditions, which in turn allows dynamic properties of proteins to be studied [35, 36], along with their thermodynamic and kinetic interactions with other macromolecules or low molecular weight ligands [36]. Unfortunately protein size, limited to around 30 KDa for complete atomic resolution, is still the major limitation for this method because of signal overlapping [39]; however, backbone assignment for proteins up to 100 KDa have been described [28].

Even for small proteins, the ^1H NMR spectrum can be complicated (figure 1.8a), and many techniques have been devised to ease the problem of interpreting complicated spectra especially for larger proteins, including 2D techniques, such as nuclear Overhauser effect spectroscopy (NOESY), which measures distance-dependant coupling, and total correlation spectroscopy (TOCSY) which measures coupling between covalently bonded atoms [36, 40] (figure 1.8b). One important problem with homonuclear NMR is peak overlap, which can be overcome by focusing on atoms other than ^1H that provide a heteronuclear signal-quantum correlation (HSQC) spectrum, and when combined with (NOESY) or (TOCSY), produce a 3D NMR spectra to facilitate assignment [41]. H/D solvent exchange studies can provide topological data about the protein, such as which part is solvent exposed or buried, since solvent accessibility is reflected by exchange rate [42]. Once the spectra have been assigned to elucidate the 3D structure, spatial restraints such as distance, angle, vdW, and bond lengths are then applied to generate a group of structural conformations consistent with the entered constraints [36] (figure 1.8c). For large proteins where the peaks are broad and weak due to fast relaxation times, a new technique has been introduced to attenuate relaxation, called transverse relaxation optimized spectroscopy (TROSY) [43]. Furthermore, extensive efforts have been

directed towards automating the NMR process to increase the throughput of protein structure determination [35, 44].

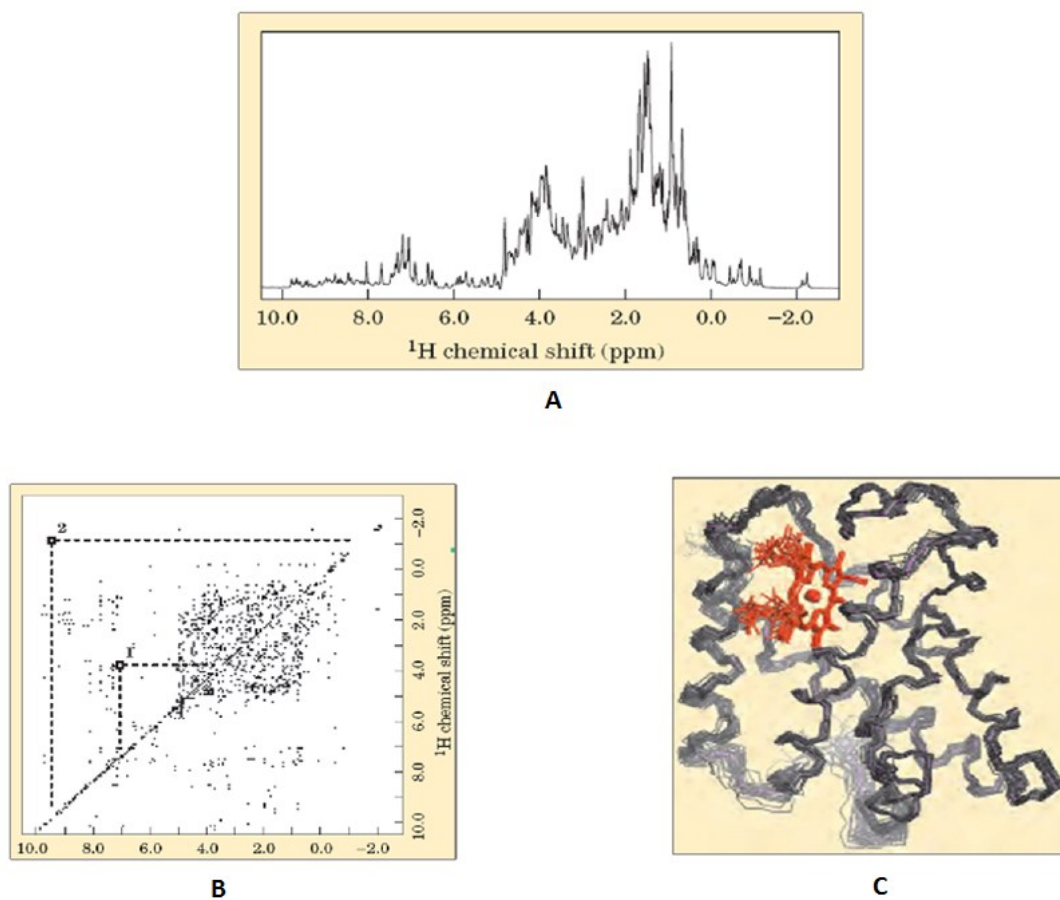


Figure 1.8: Determination of protein 3D structure using NMR. (A) One dimensional (1D) NMR spectrum of a protein molecule. (B) Protein 2D NMR spectrum, the off-diagonal dots are the NOE signals of coupled H atoms. (C) Protein 3D structure, the multiple lines represent the structural conformers [36].

1.1.5.2 Computational approaches

Computational methods to characterize and identify binding sites in proteins are needed to understand molecular interactions and to design compounds of pharmaceutical and biotechnological interest. However, the complexity of protein-ligand interactions, which involves kinetics, pharmacodynamics and physicochemical complementarity, makes the task of full binding site characterization a demanding computational problem [45]. Methods that are used to predict new binding sites can be categorized into:

- *Geometric approaches* such as POCKET [46], LIGSITE and its extensions [47-49], and SURFNET [50] which identify pockets or cavities on the protein surface by utilizing the geometric features of protein shape without need to any knowledge of the ligands.
- *Energetic approaches* such as GRID [51] and CS-Map [52] which utilise many physicochemical descriptors of ligand binding to calculate interaction energy between a probe molecule and the target protein to identify interaction sites rather than pockets or cavities.
- Methods using *structure and sequence comparison* such as FINDSITE [53] which utilises similarities between proteins of known function to identify binding sites.
- Methods that study the *dynamics* of protein structures such as molecular dynamics (MD) to simulate localized protein motion coupled with normal mode analysis (NMA) which identifies large scale motions to generate ensembles of representative protein conformations that accounts for protein flexibility.
- Methods that detect *cooperative coupling* between different protein regions such as the statistical coupling analysis (SCA) [54] and the COREX algorithm [55].
- Methods that *identify potential binding sites and predict their druggability* such as SiteMap, which characterise the binding site in a similar manner to the GRID algorithm by using a probe molecule to define binding sites by linking 'site points' that are most likely to be involved in protein-protein or protein-ligand binding, and these points are determined according to their sheltering

from solvent and closeness to the protein surface [56, 57]. A good account of these methods can be found in [45].

- Finally the *surface triplet propensities (STP) method*, which is based on a score table of triplets of protein surface atoms that can be simultaneously touched by a solvent probe molecule to give the propensities of atom types of surface atoms to be located in a binding site [58].

The following section discusses the methods that are most related to the identification of allosteric binding sites in proteins.

1.1.5.2.1 Simple Intrasequence Differences (SID) analysis

SID analysis is a bioinformatics method designed to search protein fold topologies and to identify and grade interfaces of potential contribution to molecular stability. An understanding of the internal arrangement of these interfaces helps predict conformational change in the fold resulting from site specific inductions (e.g. via mutations or ligand binding) [59]. To achieve this, every residue in a protein 3D structure is graded numerically according to its topological situation in the folded chain. This grading highlights the potential contribution of every position and its vicinity to the overall conformational stability of the molecule, and assesses where each position exerts its effect, whether within an element of local structure or at an interface between such elements. By merging the overlapping localities of high grades, it is possible to identify the extent and topology of significant sub-structural interfaces existing within the folds of domains and sub-domains [59]. The approach undertaken to grade residues is to firstly define a volume around the α -carbon of each residue position in the folded protein. Then, travelling sequentially from N to C-terminus, clusters are formed by grouping residues whose α -carbon is contained within the defined volume, and these clusters are then scored. Clusters are assigned a score value based on the maximum chain separation of residues within the cluster. The scoring methods include (figure 1.9):

(a) *Simple difference (highest residue number – lowest, HL)*; this score indicates how far in the primary sequence are the two extreme residues of protein segments that have encompassed within the same sphere. Therefore, the higher the number of

protein segments that are sequentially distant the higher the score. A cluster that is overlying an *interface* will have a high score, contrary to that found within an *element* which will have low score.

(b) *Simple difference (Greatest gap, GG)*; where all positions within a cluster are numerically ordered and the difference between consecutive residue numbers is calculated, the GG value is assigned the largest difference. A cluster that contains only a single protein segment will have a low score, while if two non-overlapping segments are found the score will be high. However, if more than two segments are found then the cluster will have a score that is lower than the HL score (intermediate).

(c) *Differential score (DIFF)*; which is the difference between the HL and GG scores ($\text{DIFF} = \text{HL} - \text{GG}$). In clusters that contain single segment of structure or two segments of chain-distant structures, the HL and GG scores will be small, hence the DIFF score. While in the case of having three or more segments, the DIFF score will be large as the GG is lower than the HL score.

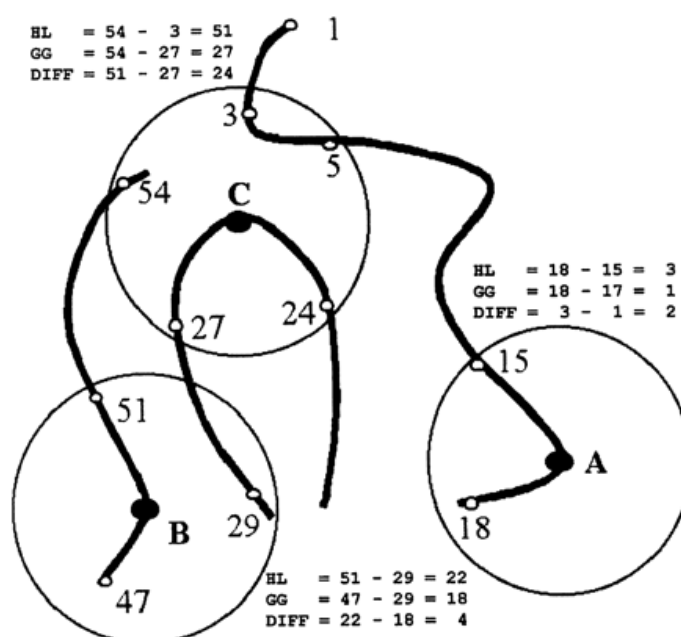


Figure 1.9: A schematic representation of SID analysis of three adjacent sections of protein backbone. The bold lines represent the backbone, and circles represent the spherical volumes defined around residues A, B and C. SID scores (HL, GG, and

DIFF) are shown for each cluster. Cluster A with low SID scores represents a locally independent element of structure. Cluster B with high SID scores represents the interface of two sections of structure. Cluster C with high HL, but intermediate GG scores represents the interface between three separate sections of structure. Thus elements are distinguishable from other interfaces between two, and three or more sections of chain [59].

In summary, when the number of protein segments (sections) in the scored cluster is one all scores (HL, GG, and DIF) will be low. When it is two, the HL and GG scores will be high and the DIFF will be low. When it is three or more, the HL will be high and the GG and DIFF scores will be intermediate. Therefore, consideration of these score values and the number of residues contained within that cluster enables absolute identification of the type of interface enclosed by that cluster [59].

Comparing SID scores for members of a protein family, or different states of the same fold, illustrates how SID scores change upon protein perturbation (e.g. ligand binding) and focuses attention on areas where such perturbation has the most topological impact. A standard deviation for every position can be obtained from this comparison, which can guide identification of regions of potential motion or adjustment in the protein. Values of standard deviation (STD) close to zero indicate little change in SID score, which implies that the region around the position in question is not vulnerable to conformational change, whilst large values can imply some degree of conformational or structural rearrangement [59].

In order to visualize the structural elements and interfaces in a protein fold, clusters obtained by SID analysis need to be merged. This merge proceeds by combining clusters having one or more residue in common. This process is repeated until all clusters have been examined [59]. Figure 1.10 and figure 1.11 illustrate interfaces highlighted by SID in the pancreatic trypsin inhibitor fold (PDB code 4pti) [59].

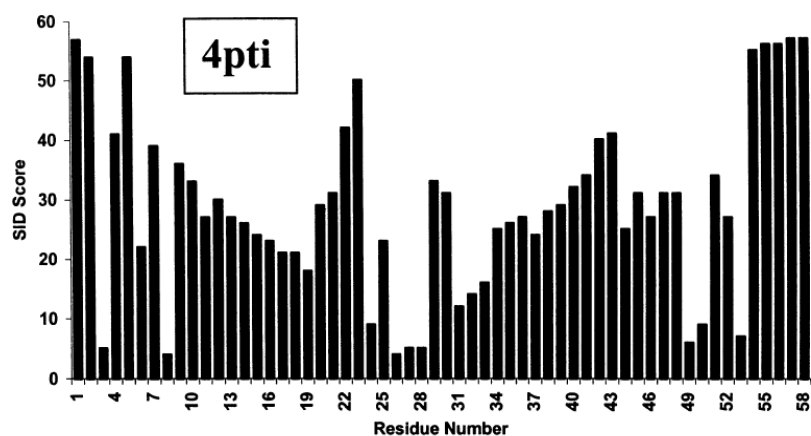


Figure 1.10: Graphical representation of SID score (HL) for each residue position in the analysed sequence fold [59].

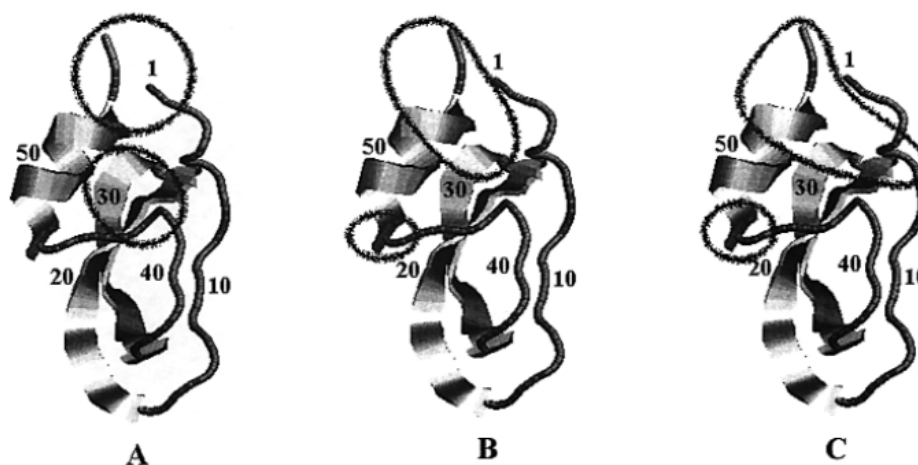


Figure 1.11: Comparative SID scoring showing cartoons of chain fold with the notably changed sites being circled. (A) Sites of high SID scores (HL calculation). (B) Sites of considerable variation in SID score amongst homologues. (C) Sites of considerable changes in SID score upon complexation with other enzyme [59].

SID analysis of protein fold topologies has shown that major interfaces tend to be externally accessible. These interfaces are often closely connected to the active site, which explains how alterations to the interfacial juxtapositions (allosteric binding) can affect the properties of the active site via interfacial realignment [59].

1.1.5.2.2 *Statistical Coupling Analysis (SCA)*

The protein-protein (or protein-ligand) interaction can be envisioned as an energetic perturbation to the binding surfaces of proteins that diffuses through their 3D structure causing specific changes in their function. In order to understand allosteric coupling in proteins (the long range energetic interactions between sites on the protein), the energetically coupled residues involved in the allosteric communication process need to be identified first to understand their physical mechanism of communication [22, 54]. A possible way to achieve this is through systematic mutagenesis, but this approach does not necessarily provide information about allosteric communication between sites, or is limited to small regions in the protein. For a full scale mapping of protein residues, Ranaganathan and co-workers introduced a method called statistical coupling analysis (SCA), which is a sequence-based technique used to identify coevolution of energetically coupled residues involved in allosteric communication networks. This method benefits from the fact that evolution is a vast experiment in mutagenesis selecting for function, and a functional energetic coupling of a pair of residues will drive their coevolution [54]. Two hypotheses of molecular evolution establish the basis of SCA. First, frequencies of amino acid with no evolutionary constraints (conservation) at any position will approach their mean in all protein families, and the conservation at any site j is the degree of its frequency deviation from that mean, which is measured by ΔG_j^{stat} (energy-like statistical parameter). Second, functional coupling of two positions will drive their mutual coevolution [54]. In SCA, the coevolution of two sites can be quantified by statistical analysis of large and diverse multiple sequence alignment (MSA) of a protein family and selecting a subset of sequences in which a fixed amino acid appears at a specified position, followed by evaluation of the statistical perturbation of the distribution of that amino acid on the distribution of amino acids at other sites [22, 54]. The effect of perturbation of an amino acid frequency at site i , on amino acid x at another site j is measured by another energy-like statistical parameter $\Delta\Delta G_{i,j}^{\text{stat},x}$. Calculating $\Delta\Delta G_{i,j}^{\text{stat},x}$ for all sites j will generate a map of how perturbation at site i is felt by all other sites and is an evolutionary prediction of global patterns of energetic coupling for test site i [54]. Using this method, Ranaganathan and coworkers have predicted the networks of residues responsible for

the conduction of allosteric signaling in the G-protein family [60] and the RXR nuclear receptor heterodimers [22]. Figure 1.12 illustrates the complete SCA for the G protein family as a matrix of $\Delta\Delta G^{\text{stat}}$ values.

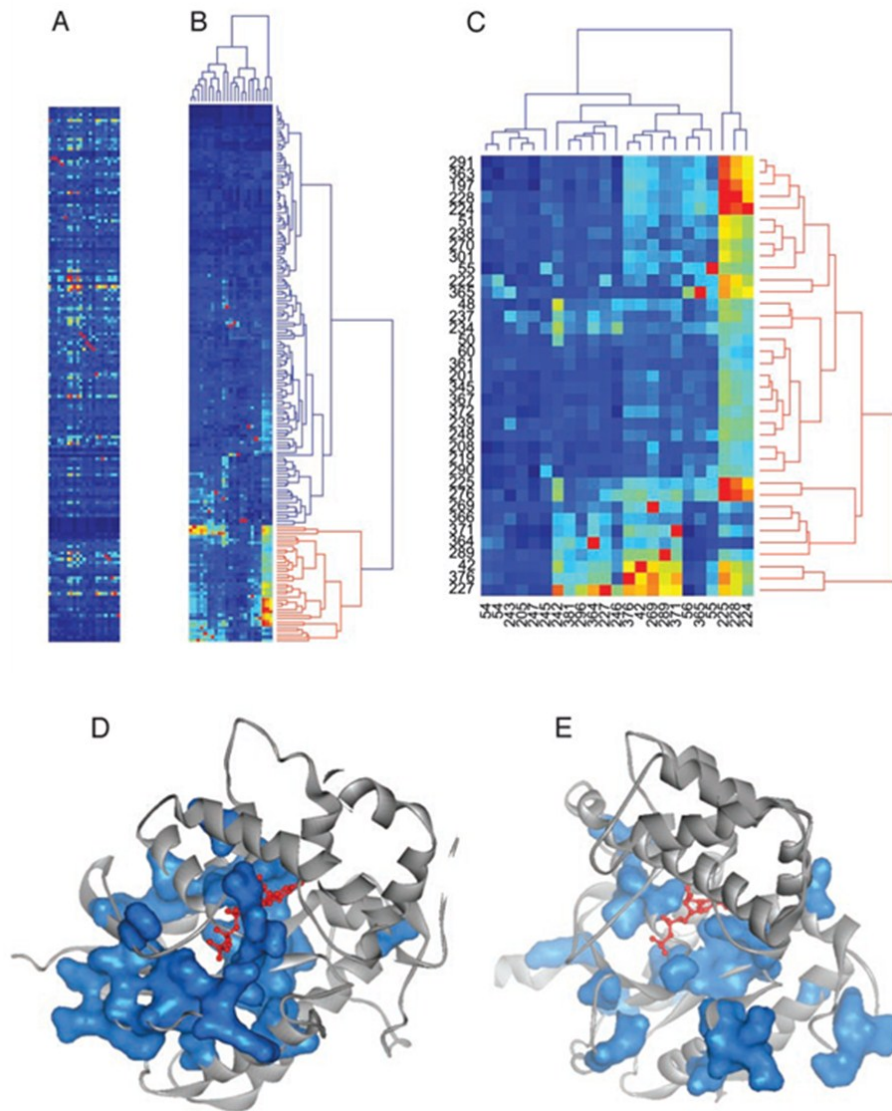


Figure 1.12: Illustration of the SCA through predicting the allosteric core of the G protein family [60]. A) Rows are positions (N to C terminus) on the MSA, and columns are the perturbation experiments. Thus, each column shows the statistical coupling between one perturbed test site (red pixels) and all other sites, the global evolutionary coupling between many pairs of residues is represented by the whole matrix. (B) 2D clustering of the SCA matrix identifies clusters of residues showing

similar evolutionary coupling profiles. (C) Extraction and reclustering of the primary positions (red cluster in B) reveals their mutual coevolution. (D-E) Mapping the primary clustered residues on the structure of a G protein generates a van der Waals surface (blue) around the cluster residues. In (D) the activated state (bound to GTP) the residues form a network surrounding the nucleotide-binding pocket and linking it with those at the effector-binding sites. But they are fragmented in (E) the inactive state (bound to GDP) [60].

1.2 Molecular modelling

1.2.1 Overview

Molecular modelling is centred on the application of theoretical and computational techniques, such as physical and mathematical laws, to create models and simulations that mimic the behavior of molecules or molecular systems (ranging from small molecules to large biomolecules such as proteins and nucleic acids), which in turn help in predicting and understanding the properties of these molecules [61].

In the field of drug discovery, the main aim of molecular modelling is to model and simulate molecules implicated in human diseases to help predict and estimate molecular properties to enable a more rational process of drug development. The prominent application of structure-based molecular modelling is the identification of the active sites for enzymes and receptors, and finding clefts implicated in protein-protein interactions [62].

The modelling of drugs, receptors and their interactions can be divided into two broad categories; ligand-based modelling, and structure-based drug design (SBDD) [62]. The ligand based drug design (LBDD) method utilises the information from a compound or a set of compounds that have known biological activity against the target of interest to search databases of small molecules for similar compounds with better biological activity. This can be done by pharmacophore matching, similarity and substructure search, or 3D shape search. In the case of the SBDD, the structure of the target is known and searching for possible ligands is performed by molecular docking of the ligands in the available databases into the binding site of the target to obtain a predicted binding mode. Poses are then scored to assess the binding affinity of the ligands to rank them and select the best ones for experimental testing [63]. Also in SBDD, the amino acids of the binding site of the target can be used to generate a pharmacophore to perform preliminary screening of databases.

The different operations that are conducted in the field of molecular modelling are carried out using programs or algorithms that calculate different aspects of the studied system ranging from calculating the structure to calculating its molecular properties such as conformational energy, energy minima, atom charges and many

other properties. This becomes feasible nowadays because of the huge advances in computer power, both the hardware and the software.

The computational methods that are used in molecular modelling to calculate structure and property data can be divided into two groups, quantum mechanics and molecular mechanics [64]. *Quantum mechanics (QM)* is the most accurate in calculating the energy of a given system since electrons are explicitly considered (although in some cases such as self-consistent field theory, each electron only experiences the smeared-out charge of all of the other electrons in the molecule). Unfortunately, because it is computationally very demanding, its application is limited to small systems of limited numbers of atoms. A more applicable alternative for computational simulation of biological systems is *Molecular Mechanics (MM)* which utilizes the Born-Oppenheimer (BO) approximation. BO approximation simplifies the Schrödinger equation for a molecule by assuming that nuclei masses are much heavier than that of electrons and that electrons are much faster than nuclei. Therefore, the electronic and nuclear motions in molecules can be separated; consequently atoms are treated as classical particles [65].

The solvation effect should also be taken into account in molecular modelling because ligand-receptor non-bonded interactions occur in an aqueous environment and ignoring water molecules will affect the molecular geometry and the energetics. Depending on the system representation and computational cost, solvent can be introduced either explicitly or implicitly [62].

A technique known as energy minimisation, based on MM, is used to find stable conformations of the system that corresponds to a local energy minimum. It is used to relax the geometry and optimize ligand-receptor contacts. *Molecular Dynamics (MD)* simulations which are based on Newton's laws of motion, describes the behaviour of the system as a function of time (spatial information for each atom in the molecule). All of these approaches will be discussed in more detail in the following sections.

1.2.2 Quantum mechanical methods

Quantum mechanics (QM) uses theories developed for quantum physics and provides the most accurate results when calculating the energy of a given system since electrons are explicitly considered. QM methods are often designated as *ab initio* because they are capable of reproducing experimental data without the need for empirical parameters by solving Schrödinger's equation from first principles (indeed, *ab initio* is the Latin translation for "from the beginning"). Because electrons are considered explicitly, *ab initio* methods are indispensable when investigating chemical reactions involving bond formation or breaking and in cases where little or no experimental data are available. Unfortunately, because it is computationally very demanding, its application is limited to small systems of limited numbers of atoms (hundreds of atoms), such as calculating ligand internal energy. It is noteworthy that the quality of the QM calculation is dependent on the basis set used for the calculation as well as the particular level of theory chosen for the calculation [62, 65, 66].

To reduce the high computational expense of the *ab initio* method and to fill the gap between it and the MM method, empirical parameters were introduced to the *ab initio* method to produce what is called the *semiempirical molecular orbital method*. Here, only the valence electrons are taken into account as the main molecular properties of an atom are influenced mainly by these electrons. Semiempirical methods differ only in the kind of approximation they make, and many of them such as Austin Model 1 (AM1) [67] and Parametric Method 3 (PM3) [68] have shown a good compromise between the accuracy of the results and the computational cost [66]. Semiempirical methods are faster than *ab initio* methods and are applicable to bigger systems up to hundreds of atoms, because of their approximations and the use of stored parameters. In general, QM methods are suitable for calculating properties such as molecular orbital energies and coefficients, electrostatic potential, partial atomic charges as well as reliable structural parameters [64]. Because biological macromolecules are relatively huge, to allow QM treatment of ligand-receptor complexes, a hybrid QM-MM method, which combines the advantages of relatively fast MM and accurate QM calculations, can be applied to tackle such systems. For example, in a protein ligand simulation the ligand and the active site (where the

process needs to be described electronically and accurately) is treated quantum mechanically, whereas the rest of the protein and the solvent are described using molecular mechanics [62, 66].

1.2.3 Molecular mechanics

Unlike QM, which deals with electrons and nuclei explicitly, *Molecular Mechanics (MM)* utilizes the BO approximation, where atoms (and therefore molecules) are treated as classical particles and their interactions and energies (force fields) are governed by bonded (bond distance, bond angles, dihedral angles) and non-bonded (van der Waals and electrostatic interactions) terms. The neglect of electrons within these calculations significantly reduces the number of bodies to simulate and makes MM a fast and practical method even for large biological systems that contains thousands of atoms such as proteins [62]. MM is a parametric method, in which a set of molecules are used to obtain parameters of general applicability. These parameters are derived from high level QM calculations and/or experimental data. In the context of molecular modelling, the set of parameters and the mathematical function used to calculate the potential energy of the system are collectively known as a force-field. The most common force-fields for biological macromolecules are AMBER and CHARMM [62]. In general, MM is suitable for energy minimisation of the system and in identification of stable conformations, generating different conformations, calculating the conformational energy and to study the motion of the system over time [64].

1.2.3.1 Force fields

Force fields are used to optimize molecular geometry and calculate energies, and operate on the assumption that the total potential energy of a molecule can be expressed as the sum of contributions from many interaction types between its atoms [69]. In MM, atom types reflect hybridization and the surrounding environment of each atom, and are treated as spheres of different sizes connected together by bonds represented by springs of different lengths, where Hooke's law can be applied to calculate the potential energy of the atomic ensemble. The motion of these atoms can

then be described by applying laws of classical physics. The total potential energy (steric energy) is a function of the atomic positions of all the atoms in the system (atomic coordinates), and its value is calculated as the sum of the bonded and the nonbonded terms. The energy of a simulated system will change as bond lengths, angles and torsions together with nonbonded interactions deviate from the ideal geometries specified in the force field [65, 66, 70, 71]. These reference values are empirical parameters which are derived from high level QM calculations and/or experimental data. In this context, a force field can be defined as the collection of reference unstrained values together with a set of parameters (force constants) used to calculate the total potential energy of a given system. A general functional form of a force field is given in equation 1.1 [66]. Different groups have developed many force fields for different molecular types to calculate the energy of a molecule, and they all follow the scheme shown in equation 1.1, and most have additional terms encapsulated within ($E_{\text{miscellaneous}}$) [65].

$$E_{\text{total}} = E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{vdw}} + E_{\text{elec}} + E_{\text{miscellaneous}} \quad (1.1)$$

E_{total} is the total energy of the molecule, E_{str} is the bond stretching energy term, E_{bend} is the angle bending energy term, E_{tors} is the torsional energy term, E_{vdw} is the van der Waals energy term, E_{elec} is the electrostatic energy term. The $E_{\text{miscellaneous}}$ term is a collection of energy terms used in different force fields such as: E_{improper} which is added to keep chirality and planarity of sp^2 atoms; $E_{\text{cross-terms}}$ which are added to account for the possible effects of bond stretching, angle bending and torsion on nearby atoms bond lengths, angles and torsion; and other terms are included that account for hydrogen bonding [65]. The force field equation is shown pictorially in figure 1.13.

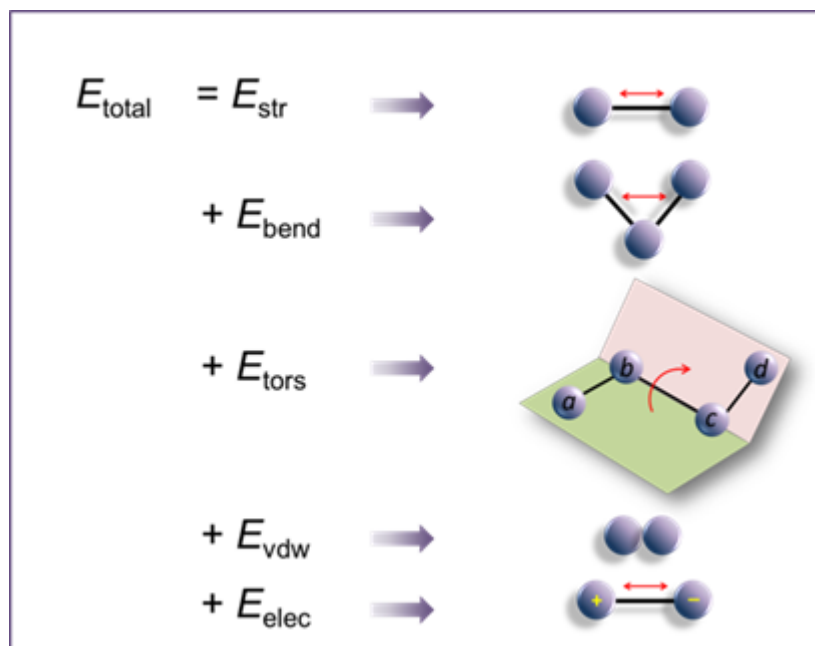


Figure 1.13: Pictorial representation of the different terms incorporated in a MM force field.

The fundamental idea of MM is that bonds have natural lengths and angles, and their equilibrium values and the corresponding force constants applied to maintain these equilibria are referred to as force field parameters. Each deviation from these values will increase the total molecular energy, which is a measure of the intermolecular strain relative to hypothetical unstrained reference structure; therefore, the total energy by itself has no physical meaning [66]. The main drawback of MM is its dependence on empirical parameters which are obtained from a small number of molecular model systems [69]. To understand how force fields are utilised in molecular modelling, each energetic term is described below:

i. Bond stretching energy (E_{str}): this is a harmonic potential that describes the energy changes associated with stretching or contracting a bond from its unstrained length. The simplest way to represent this is by treating bonds as ideal springs (figure 1.13 first term). The stretching energy is simply a constant multiplied by the square of the displacement from the unstrained bond length (equilibrium position) as described by Hooke's law (equation 1.2) [66], giving a harmonic energy curve (figure 1.14).

$$E_{str} = \frac{1}{2}k_b(b - b_0)^2 \quad (1.2)$$

where k_b is the bond stretching force constant, b_0 is the unstrained bond length and b is the actual bond length.

As can be seen from the curve, the approximation only fits well near the equilibrium bond length. Fortunately, this is not a problem because covalent bonds are stronger than forces affecting the molecule and rarely change very much from their equilibrium [65]. In more refined force fields a Morse function may be included at the expense of complexity [65, 66, 69]; or equation 1.2 may be modified with another term proportional to the cubic (as in the MM2-based force fields [72]) or quartic (as in MM3[73-75], CFF [76], and MMFF [69] force fields) change in bond length which makes the curve fit the more accurate Morse curve more precisely[65].

ii. Angle bending energy (E_{bend}): Again, for angle bending energy the harmonic, simple spring model is used (figure 1.13 second term). As shown in equation 1.3 the angle bending energy is proportional to the square of the deviation from the equilibrium position [66]:

$$E_{bend} = \frac{1}{2}k_\theta(\theta - \theta_0)^2 \quad (1.3)$$

where k_θ is the angle bending force constant, θ_0 is the equilibrium value for the bond angle and θ is the actual value.

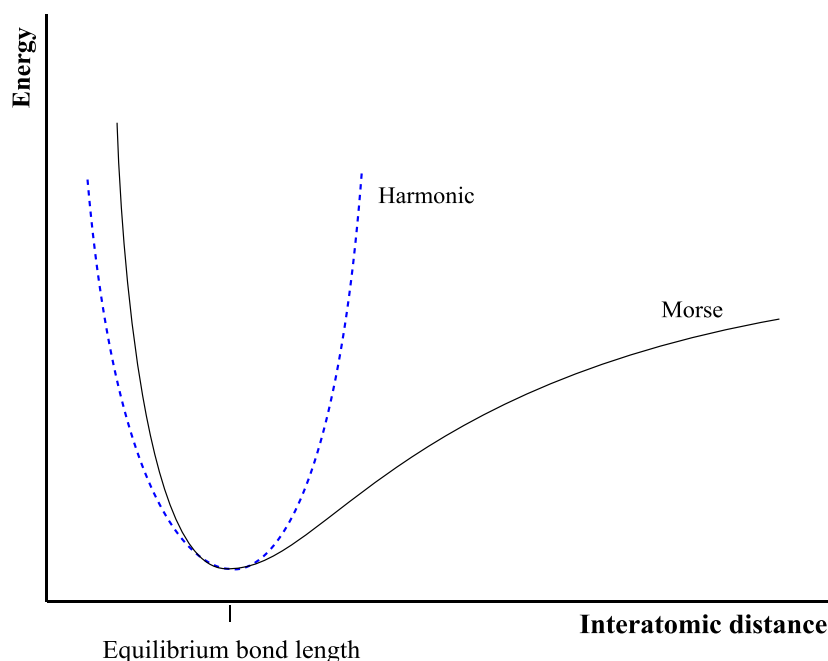


Figure 1.14: Energy change with bond length. The simple harmonic potential of Hook's approximation (dashed blue line) is compared with the Morse potential (solid black line).

iii. *Torsional energy (E_{tors}):* this models the rotational barriers attributed to the presence of steric interactions between atoms separated by 3 covalent bonds (4 atoms a , b , c and d). The dihedral angle, which is the source of most conformational flexibility in biomolecules, is the angle between the plane containing the first three atoms (abc) and the plane containing the last three atoms (bcd) [65] (figure 1.13 third term). The potential energy on rotation around a dihedral is periodic in nature. Thus, the torsion angle potential is often expressed as a cosine function as shown in equation 1.4 [66]:

$$E_{tors} = \frac{1}{2}k_{\varphi}[1 + \cos(n\varphi - \varphi_0)] \quad (1.4)$$

where k_{φ} is the torsional barrier, φ is the actual torsion angle, n is the periodicity (number of energy minima within one full cycle) and φ_0 the reference torsional angle. Figure 1.15 shows the torsional profile for ethane.

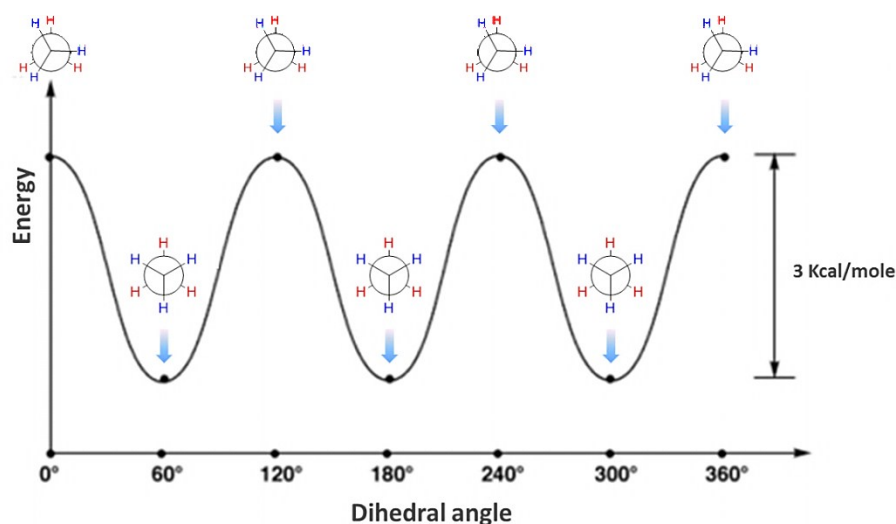


Figure 1.15: Torsional profile for ethane as a function of torsional angle with an energy barrier of about 3 Kcal/mol, showing the periodicity of the potential energy on rotation around the carbon-carbon bond.

Torsion terms are also used to maintain chirality and planarity of sp^2 atoms via introducing an additional energy term called *improper torsion* (equation 1.5). [65, 77]. Figure 1.16 A shows the improper torsion angle of the carbonyl oxygen atom in cyclobutanone relative to the other four atoms of the ring for which an improper torsional potential can be used to maintain it at 0° or 180° [78].

$$E_{improper} = k_{\omega}(\omega - \omega_0)^2 \quad (1.5)$$

where k_{ω} is the improper angle force constant, ω is the improper angle and ω_0 is the reference equilibrium value. There are other ways to account for plane bending. One way is to calculate the angle between a bond and the central atom relative to the plane formed by the central atom and the other two atoms. Another approach is to measure the height of the fourth atom above the plane formed by the other three. The out of plane torsion deviation (either in angle or distance) can be accounted for by a harmonic potential in equation 1.6 (figure 1.16B-C) [78, 79].

$$E_{opl} = \sum k_x X^2 \quad (1.6)$$

where, E_{opl} is the out of plane energy, k_x is the force constant and X is either the height or the angle of the out of plane deviation.

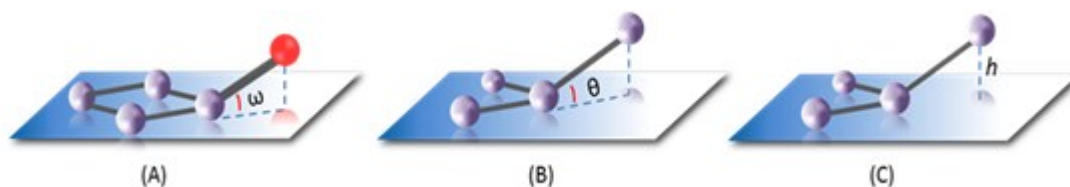


Figure 1.16: Representation of the improper torsion and out of plane terms which are added to force fields to maintain planarity of unsaturated systems. A) Improper torsion angle ω between the oxygen atom (red) and the plane identified by the other four atoms in cyclobutanone. B and C) out of plane deviation measured as an angle or distance [78].

iv. *van der Waals energy (E_{vdw}):* In spite of the fact that atoms have clouds of electrons surrounding the nucleus, they behave as if they have definite size, and a measure of this atomic size is the van der Waals radius. Therefore, if atoms get too close to each other without a bond being formed, then the interaction energy will go up rapidly because of the repulsive forces between electron clouds. Nevertheless, atoms like to be adjacent, and the reason for this mutual interaction is the induced dipole interaction. The induced dipole results from fluctuations in the charge distribution in the electron clouds as a result of movement around the nucleus, and the instantaneous dipole on one atom will induce a dipole on a second nearby atom giving rise to an attractive interaction [65] (figure 1.17).

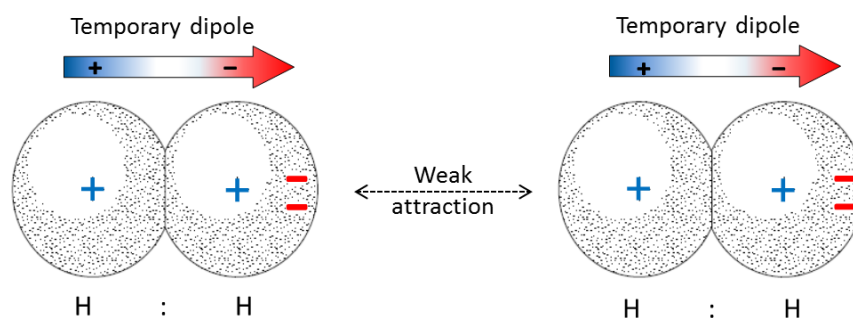


Figure 1.17: The induced dipole and van der Waals attraction.

This attractive force is also referred to as the dispersion force or London force, and the net effect is proportional to $1/r^6$, which is most often expressed using a Lennard-

Jones 6-12 potential shown in equation 1.7 and figure 1.18 [65]. There are several other forms of modified Lennard-Jones potential that are used in different force fields [66].

$$E_{vdw} = \sum \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \quad (1.7)$$

where, A_{ij} is the repulsive term coefficient which is accounted for by r^{-12} , B_{ij} is the attractive term coefficient which is mediated by r^{-6} and r_{ij} is the distance between the atoms i and j .

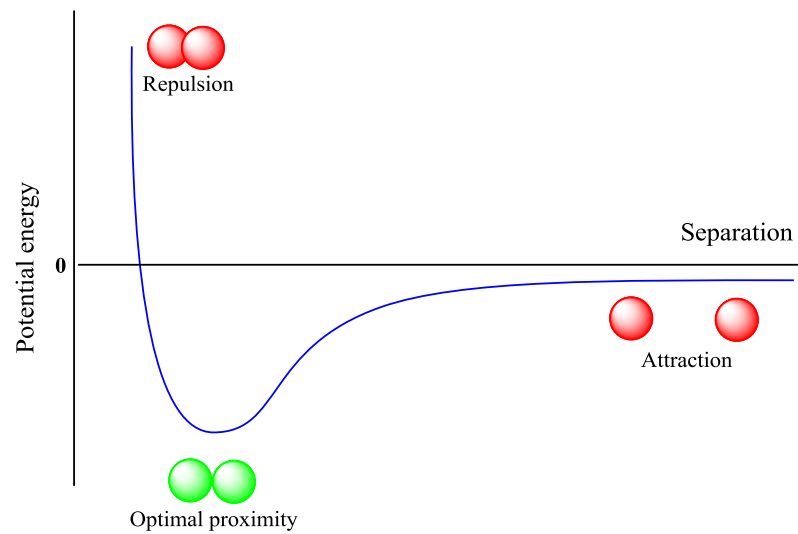


Figure 1.18: A typical van der Waals curve, showing the attractive and repulsive terms.

v. *Electrostatic energy (E_{elec}):* The electrostatic interaction between a pair of atoms is represented by Coulomb's Law (equation 1.8) [66], after assigning a partial charge to every atom in the system. The force of the attraction between charges is inversely proportional to the square of the distance ($1/r^2$), so the energy of interaction is proportional to $1/r$ [65]:

$$E_{elec} = \frac{1}{\epsilon} \left[\frac{(Q_1 \cdot Q_2)}{r} \right] \quad (1.8)$$

where ε is the dielectric constant, Q_1, Q_2 are the atomic charges of the interacting atoms and r is the interatomic distance. The value of E_{elec} is calculated for all pairs of atoms in the system, which is proportional to the square of the number of atoms (pairwise model): for a system with N atoms, there will be $\frac{n(n-1)}{2}$ interactions. Generally, in large systems, most of the computational time required to calculate the energy of the molecule is accounted for by the large number of nonbonded interactions. To reduce this computational expense, a cut-off distance is used beyond which separated atoms are assumed to have negligible interactions. This approximation works well for E_{vdw} , as the attractive interactions fall off rapidly with distance, but less well with E_{elec} because there can be significant interactions over large distances, [65]. Often a short cutoff is used for the VdW term and a longer one for electrostatic interactions.

vi. *Cross terms (Emiscellaneous)*: Most force fields add other terms such as cross terms, out of plane terms and hydrogen bonding terms [65, 66]. These terms account for the possible changes that might occur to some components in the force field as a result of their interdependence. For example, if a bond is stretched, then the associated bond angles will be easier to bend, and if a bond angle is opened, the rotational barrier may be reduced [65]. Other commonly used cross terms include bend-bend, stretch-bend and torsion-bend terms [79].

Stretch-bend term:

$$E_{b\theta} = \sum \sum k_{b\theta} (b - b_0)(\theta - \theta_0) \quad (1.9)$$

Bend-bend:

$$E_{\theta\hat{\theta}} = \sum \sum k_{\theta\hat{\theta}} (\theta - \hat{\theta})(\theta - \hat{\theta}_0) \quad (1.10)$$

Torsion-bend:

$$E_{\theta\hat{\theta}\omega} = \sum \sum k_{\theta\hat{\theta}\omega} (\theta - \hat{\theta})(\theta - \hat{\theta}_0)\cos\omega \quad (1.11)$$

The k terms are the force constants; b, b_0, θ, θ_0 and ω have been described previously [79].

1.2.3.2 Transferability and Parameterization of force fields

The objective of a force field is to calculate the potential energy of molecules with good accuracy. This depends on the quality of the potential energy function and the set of parameters incorporated into these functions [66]. These parameters are experimentally derived; for example, values for bond lengths and bond angles and their constants can be obtained from X-ray diffraction data and infra-red spectroscopy [65]. Spectroscopic experiments also provide information on rotational barriers and vibrational frequencies [79]. Another approach is to consider the electronegativity of the atoms, which naturally depends on the atomic charge [65]. An empirical charge scheme can also be applied to calculate charges. When there is no adequate experimental data, QM can be used in the parameterization of force fields [79]. Since force fields include long lists of parameters of the different components of the force field, they should be consistent. Furthermore, because force fields differ in terms of the presence or absence of cross terms, it makes it difficult to transfer parameters between them. There is no best force field for all problems; rather, the choice should be based on the system of interest [65]. Force fields cannot be applied for a particular problem unless all the necessary parameters are included. Several force fields have been developed to examine small molecules and many others are developed primarily for proteins and other biomolecules. Inadequate availability of experimental data has led to the development of so called class II force fields such as the consistent force field (CFF) and the Merck molecular force field (MMFF), where QM is used in both to calculate the energy surface. MMFF was developed mainly to handle all functional groups of pharmaceutical importance, including small molecules and macromolecular structures. The MMFF 94 version of this force field has been incorporated in commercial software packages such as SYBYL and MOE [66].

1.2.3.3 Force fields for protein modelling

Protein models are derived either from crystal structures or from homology modelling and in some cases these models require further refinement: in homology

derived models, the loops and side-chain conformations are chosen arbitrarily and they may not have an energetically reasonable structure; in a crystal structure, the internal strain which results from the crystal packing forces must be removed. The force fields used in protein modelling are based on equation 1, but they are slightly different from those for small molecules. Besides the special parameterization of proteins, some simplifications are introduced into protein force fields to reduce the computational expense because of the large number of atoms in such systems. Several force fields for protein modelling are incorporated in software programs such as AMBER, CHARMM and GROMOS [66]. Figure 1.19 shows the basic form of the AMBER force field that is implemented for macromolecular simulations [80-82].

$$V(r) = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta))$$

$$+ \sum_{\substack{non-bonded \\ pairs}} \left\{ \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{r_{ij}} \right\}$$

Figure 1.19: The basic functional form of the force field most commonly used in AMBER 10 for macromolecular simulations. See section 1.2.3.1 for definition of the terms.

1.2.3.4 Energy minimisation

The first step in a molecular modelling study is to generate a model by defining the relative positions of atoms using a set of coordinates, and because the starting geometry of the molecule determines the quality of the subsequent investigations, it is usually beneficial to find an energy minimum state representative of a realistic molecular geometry. [66].

A stable conformation of a molecular system is achieved when the potential energy of the system matches a global or local minimum on its potential energy surface. Finding the low energy conformation, which is the aim of molecular mechanics, is

achieved via a process called *energy minimisation*, in which atoms are moved from a starting non-equilibrium molecular geometry so as to reduce the net forces (the gradient of potential energy) on the atoms until an equilibrated conformation with negligible forces is obtained. The minimisation algorithm makes small changes in the position of every atom and calculates the energy after every move; a full cycle where each atom is moved once is called an iteration. The process continues until no further reduction in the energy is achieved, and is repeated many times until an overall energy minimum is reached (figure 1.20 and figure 1.21). It is important to note that minimisation algorithms do not necessarily find the global energy minimum; in fact, they often only find local minima. Many energy minimisation methods have been developed, and to calculate the change in the coordinates for each step, they use information from the derivatives of the potential energy function. The first derivative is called the gradient and the second derivative matrix is the Hessian. Minimisation methods can be divided into two classes: the first derivative method includes steepest descents and conjugate gradient; the second derivative method includes the Newton-Raphson technique and related algorithms [65, 66, 70].

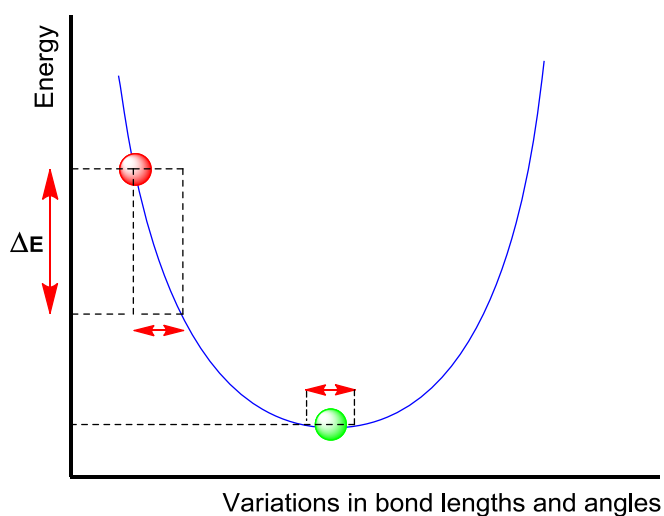


Figure 1.20: Schematic representation of the concept of energy minimisation. The red sphere ● represents an energetically unstable conformation. The minimisation algorithm makes small changes in the position of every atom and calculates the energy after every move until it reaches a stable conformation which corresponds to an energy minimum (green sphere ●).

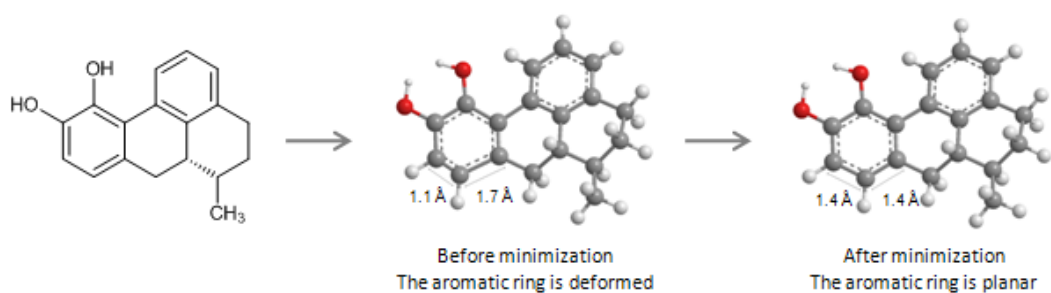


Figure 1.21: The effect of energy minimisation on the overall geometry of the molecule. Generated using Chem3D Pro 12.0® [83].

In steepest descents (SD), the minimisation proceeds rapidly downhill on the energy surface following the direction of the steepest slope, but performs more slowly near the energy minimum where the potential energy surface (PES) is flatter (this is referred to as having poor convergence properties). It is often used for structures far from the minimum, such as poorly refined crystal structures, and is used as a rough introductory run followed by a more advanced method such as conjugate gradients (CG). CG accumulates information from one iteration to the next, thus refining the direction towards the minimum, and has a more efficient convergence near the minimum compared to SD. The Newton-Raphson (NR) method uses both the gradient and the curvature of the energy surface. The second derivative is used to predict where the function will pass through a minimum, and is very effective especially when convergence is approached. However, it is time consuming to calculate the energy surface curvature, so approximations are often made to speed up the calculations. The application of this method is limited to preoptimised systems where a precise minimum is required, and should be avoided in cases involving poor structures as the minimisation process might be unstable and lead to catastrophic results (maximized structure). For both CG and NR, the major disadvantage is the computational expense and storage requirements for calculating large systems [65, 66, 70, 79].

The general approach to energy minimisation depends on two factors; the size of the system and the current state of the optimization. For example, if the structure is far from a minimum, a cascade of techniques is used. Generally SD is used for first 10-100 steps to remove close contacts, then minimisation can be completed to

convergence using either CG or NR. Different minimisation criteria can be used to determine when to stop the minimisation process. Defining the number of steps is not ideal, and stopping when the gradient is less than a selected value measured by root mean squared RMS is preferable. When choosing a method, there should be a balance between attaining a reasonable minimum and avoiding unnecessary calculations [65, 66]. Crystal coordinates sometimes have unfavourable atomic interactions which cause large initial forces and result in artificial movement of these atoms away from the original structure when minimized. To avoid this, it is recommended to relax the protein model gradually through the application of tethering or restraints. In the first minimisation stage, a restraining force is assigned to all heavy atoms to fix the atomic coordinates in predefined positions, and allows all hydrogens and solvent molecule to adjust their positions. In the second stage of minimisation, only the well-defined main chain atoms are restrained and the side chains are allowed to move to adjust their positions. Finally, all restraints are removed and the entire system is allowed to move so that the final minimum represents a totally relaxed conformation. The SD algorithm is suitable in the first two stages of minimisation, after which the CG method should be used to reach convergence efficiently [66].

1.2.3.5 Molecular Dynamics

The structural model that results from a minimisation process is not necessarily the most stable conformation because the minimisation algorithm stops once it reaches any stable conformation (local energy minimum) that might be separated from the most stable conformation (global minimum) by one or more energy barriers (figure 1.22). The minimisation program does not know if there is a more stable conformation beyond that barrier; therefore, getting to the most stable conformation necessitates the generation of different conformations of the molecule and comparing their steric energies [64]. One of the most widely used methods to achieve this is molecular dynamics (MD), which is used to explore and sample conformational space to find favoured 3D structures as well as the global energy minimum (figure 1.23).

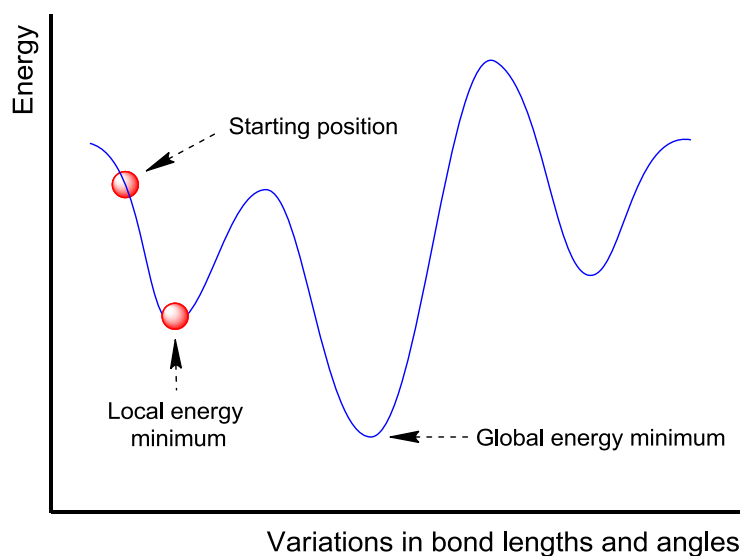


Figure 1.22: A schematic representation of the potential energy landscape showing local and global energy minima. From the starting position, energy minimisation will find the local energy minimum. In order to reach the global energy minimum an energy barrier must be crossed which can be achieved by simulated heating of the system.

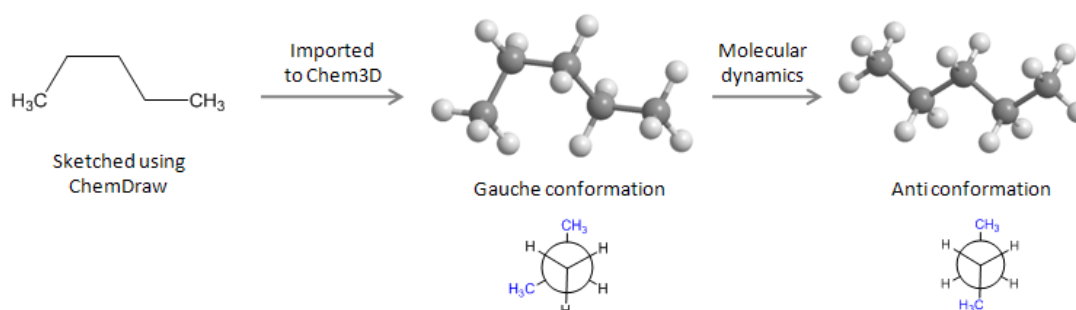


Figure 1.23: This figure shows how MD can be used to explore the molecular energy landscape for stable conformations (ideally the global minimum). From the starting sketched structure, energy minimisation was only capable of finding the stable “gauche” conformation while MD successfully found the most stable “anti” conformation. Generated using Chem3D Pro 12.0® [83].

MD can generate different conformations for the same molecule by heating it to high temperatures up to 900 K (only around 300 K for biomolecules) followed by gradual

cooling to 300K to explore the degrees of freedom (conformational space) in a process called simulated annealing. This process allows more flexible bond stretching and rotation which allows the molecule to overcome the energy barriers between different conformations and to reach a more stable one (preventing the system from being trapped in a local minimum). In all cases, a minimized structure is required as a starting structure for MD [66]. MD simulations are based on the application of Newton's laws of motion to describe the evolution of the system with time. MD requires a set of initial coordinates for the system in hand, and then the forces acting on each atom are calculated by applying a force field. From the force, acceleration can be derived and then integrated to determine the new velocity and position of each atom after a time increment (the kinetic energy). After the new coordinates have been determined, the same steps are repeated iteratively to generate a set of structures that represent the evolution of the system with time (trajectory) [62, 78]. The trajectory is obtained by solving the differential equation of Newton's second law ($F = ma$):

$$\frac{d^2 x_i}{dt^2} = \frac{F_{xi}}{m_i} \quad (1.12)$$

Equation 1.12 describes the motion of a particle of mass m_i having a force F_{xi} applied in the x_i direction. Appendix I shows the full mathematical description of the relationship between force and potential energy.

In MD simulations of macromolecules such as proteins, the forces on particles depend on their position relative to other particles, and under such conditions where the motions of all particles are coupled together, the equations of motion cannot be solved analytically and need to be integrated using a *finite difference method*. This method assumes that the potential model is a pairwise additive, and the integration is divided into many small stages that are separated by a fixed time δt . The total forces on each particle at time t is calculated as the vector sum of its interaction with other particles. From the total force, accelerations of the particles can be determined, which are then combined with positions and velocities at time (t) to calculate positions and velocities at time($t + \delta t$). Then the same process is repeated to calculate new positions and velocities at time ($t + 2\delta t$), and so on, assuming the

force to be constant during the time step [78]. Different algorithms for integrating the equation of motion can be used, and all assume that the dynamic properties of the system can be approximated as a Taylor series expression:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \frac{1}{6} \delta t^3 \mathbf{b}(t) + \frac{1}{24} \delta t^4 \mathbf{c}(t) + \dots \quad (1.13)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \delta t \mathbf{a}(t) + \frac{1}{2} \delta t^2 \mathbf{b}(t) + \frac{1}{6} \delta t^3 \mathbf{c}(t) + \dots \quad (1.14)$$

$$\mathbf{a}(t + \delta t) = \mathbf{a}(t) + \delta t \mathbf{b}(t) + \frac{1}{2} \delta t^2 \mathbf{c}(t) + \dots \quad (1.15)$$

$$\mathbf{b}(t + \delta t) = \mathbf{b}(t) + \delta t \mathbf{c}(t) + \dots \quad (1.16)$$

where \mathbf{v} is the velocity (first derivative of the position), \mathbf{a} is the acceleration (second derivative), \mathbf{b} is the third derivative and so on [78]. The most commonly used algorithm for integrating the equation of motion is the Verlet algorithm [84], which uses the positions and accelerations at time t , and the positions from the previous step $\mathbf{r}(t - \delta t)$, to calculate the positions at time $(t + \delta t)$, $\mathbf{r}(t + \delta t)$.

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \dots \quad (1.17)$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) + \dots \quad (1.18)$$

Adding the two equations gives equation 19 which is used to calculate the positions at time $(t + \delta t)$.

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2 \mathbf{a}(t) \quad (1.19)$$

Velocities are missing in this expression; one way to calculate them is by dividing the difference between the positions at $(t + \delta t)$ and $(t - \delta t)$ by $2\delta t$:

$$\mathbf{v}(t) = \frac{[\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)]}{2\delta t} \quad (1.20)$$

The advantage of the Verlet algorithm is that its implementation is straight forward with modest storage requirements. Unfortunately, it suffers some drawbacks such as its low precision because it adds a small term $\delta t^2 \mathbf{a}(t)$ to the difference of larger ones $2\mathbf{r}(t)$ and $\mathbf{r}(t - \delta t)$. Also, it is not self-starting since it uses positions at $(t - \delta t)$ to

calculate position at time(t). Consequently, at time $t = 0$ there is only one set of positions and positions at $(t - \delta t)$ need to be calculated by other means. Because of this, several variations and improvements of the Verlet algorithm have been developed, such as the *leap-frog* algorithm and the *velocity Verlet* method. The latter gives positions, velocities and accelerations at the same time without any loss of precision using the following equations [78]:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2} \delta t^2 \mathbf{a}(t) \quad (1.21)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2} \delta t [\mathbf{a}(t) + \mathbf{a}(t + \delta t)] \quad (1.22)$$

Another related algorithm called the *Beeman's algorithm* can also be used:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{2}{3} \delta t^2 \mathbf{a}(t) - \frac{1}{6} \delta t^2 \mathbf{a}(t - \delta t) \quad (1.23)$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{3} \delta t \mathbf{a}(t) + \frac{5}{6} \delta t \mathbf{a}(t) - \frac{1}{6} \delta t \mathbf{a}(t - \delta t) \quad (1.24)$$

It uses a more accurate expression for the velocity, which results in better energy conservation because kinetic energy is calculated directly from the velocity. However, because of the complexity of the expressions used, it is computationally more expensive [78]. The availability of many integration algorithms makes it difficult to decide which one to use. The basic criteria for a good integrator are: it should be fast; require little computational cost; permit a long time step, δt , which is related to computational cost since long time steps require fewer iterations to cover the conformational space; finally, one of the most important criteria is energy and momentum conservation, which can be calculated as the RMS and plotted against time. It is important to use the correct time step in MD simulations. If too small, only a small proportion of the conformational space will be covered; if too large, atoms might move too far and end up occupying the same area of space, resulting in huge, unrealistic potential energies which may lead to instabilities or total failure of the integration algorithm. The aim is therefore, to find the correct balance between the trajectory and covering the conformational space [78].

In flexible molecules such as proteins, it is important to find how short a time step is necessary (the characteristic time for the system) which can be determined by the

fastest mechanical movement of the system. A good guide is to choose a time step that is one-tenth the highest frequency movement, which is the stretching vibration of the C-H bonds, as inferred from IR spectroscopy. The C-H vibration is repeated in a frequency of about 10 fs, so the time step is generally set to 1 fs. Since the C-H vibrations are of little importance in MD simulations, freezing out this movement by constraining these bonds at their equilibrium values and allowing the rest of the molecule to move freely will allow longer time steps to be used. It also reduces the computational cost because the bond stretching energy will not need to be calculated for the frozen bonds during each iteration. The most commonly used method of applying constraints in MD is the SHAKE algorithm developed by Ryckaert [78, 85]. There are other methods that can be used to further reduce the computational demand in MD such as the use of the united-atom potential energy function in which a group of atoms such as *CH2* and *CH3* are treated as one unified interaction site (one particle) that approximately represent the molecular mechanical properties of the group; and the use of cut-off radii (mentioned previously) to neglect the nonbonded interactions beyond a certain distance [66, 78]. Regarding the length of the simulation, a MD simulation takes a while to equilibrate, so the run should be long enough to allow for this [65].

1.2.4 Treatment of solvent effect in MD simulations

Solvents have an enormous influence on the geometry and the energetics of biological molecules, especially on those containing charges and dipoles. Therefore, the effect of solvent should be taken into account in MD simulations [62, 78, 79]. Solvent molecules can be incorporated in the system explicitly, but the computational cost will be increased significantly as a result of the massive increase in atom numbers. Thus, depending on the system and the computational cost, solvent can be represented either explicitly or implicitly [66].

In the *implicit* approach, the simplest way to simulate solvent effects is based on the assumption that the prominent effect of the solvent is to screen the electrostatic interactions in the solute [79]. Solvent effect is mimicked by using the corresponding solvent dielectric constant (ϵ) (ϵ water = 80), and because the electrostatic effects

decrease with r^{-1} (equation 1.8), some force fields use the distance dependent dielectric constant. Other implicit models have been developed where the solvent is treated as a continuous medium surrounding the molecule. These models have formulas to account for solvent-mediated charge-charge interactions and surface area terms for hydrophobic and van der Waals interactions. The two major continuum models applied in MD account for charge-charge interactions by using the Poisson-Boltzmann equation or the generalised Born approximation [66].

In the *explicit* approach there are many ways to incorporate actual solvent molecules into the system and these can be grouped into two categories; those that use periodic boundary conditions or those that do not. *Periodic boundary conditions* enable the use of small numbers of particles in such a way that they experience forces as being in a bulk solution. In this technique, the molecule is placed in the centre of a cell (for example in a cubic box) and the space between the molecule and the boundaries is filled with solvent. The image of this box is then replicated in all directions to give a periodic array. The coordinates of the image particles are a simple translation of the central box, and if a particle moves out of one side of the central box, another will enter from the opposite side so that the number of particles remains the same. The box size and the cutoff distance should be chosen carefully so that a particle in the central box does not see its image in the surrounding boxes (figure 1.24a) [79]. There are different shaped cells that can be used in this technique such as the cube, parallelepiped, hexagonal prism, truncated octahedron, rhombic dodecahedron, and the elongated dodecahedron. Ideally, to reduce the number of atoms in a simulation, a periodic cell should be chosen that reflects the shape of the molecule [66] (figure 1.24b).

The simplest way to incorporate explicit water in the system in a *non-periodic boundary conditions* technique is by wrapping the molecule with a skin of solvent molecule which uses much fewer particles than the periodic boundary method. In some cases when interested only in a part of the molecule such as the active site of an enzyme, the molecule can be divided into two regions: the first is the reaction zone, which is the part of interest and undergoes the full simulation; the second region is

that outside the reaction zone called the reservoir region, which is held fixed or harmonically restrained [66].

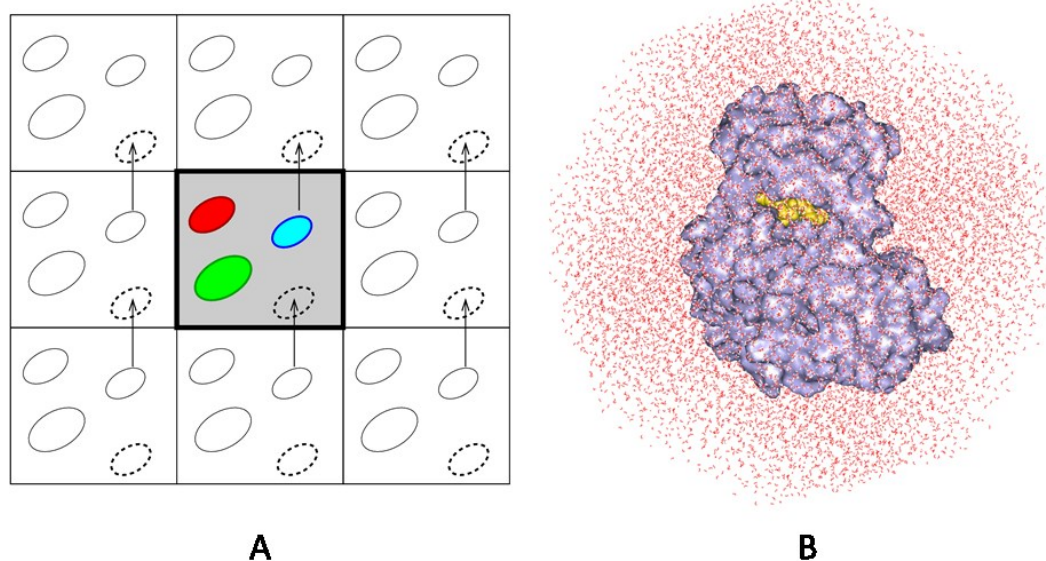


Figure 1.24: Treatment of solvent effects in MD. (A) The periodic boundary conditions in 2D. The central square is surrounded by eight image squares and whenever a particle leaves the central square another will enter from the opposite side. (B) Solvated protein (CDK2) using a truncated octahedron box. The protein is represented in its solvent accessible surface (violet), the ATP in CPK representation (yellow) and the water molecules are represented as lines (red and white) the image was generated using Discovery Studio Accelrys® [86].

1.2.5 Setting up and running a simulation

To set up and run MD simulations the following steps are applied (figure 1.25): *Choosing the initial configuration*; to start an MD simulation, an initial configuration must be chosen which represents the starting point for the simulation ($t = 0$). In general, the initial biomolecule structure for simulation is the x-ray derived or NMR structure that is obtained from the Brookhaven Protein Databank (<http://www.rcsb.org/pdb/>), although structures generated by homology modelling can also be used. The second step is *adding hydrogens*, because hydrogen atoms have only one electron they don't diffract the x-ray beam thus they don't appear in

the electron density of the molecule and consequently they are missing in the solved crystal structure. Therefore, they need to be added to the molecule in hand before starting the simulation. The third step is *solvation*, where water is added to the structure either implicitly or explicitly. The fourth step is *minimisation*; before starting an MD simulation, it is important to minimize the structure to remove all strong repulsive interactions which may lead to structural distortion and an unstable simulation. The fifth step is *assigning the initial velocities*; the initial velocities are assigned to each atom of the system by heating and integration of Newton's law of motion to propagate the system in time. During the heating process, temperature increases gradually and new velocities are assigned with each slight increase and the simulation is allowed to continue. This is repeated until the desired temperature is reached. The sixth step is *equilibration*; after reaching the desired temperature, the simulation of the entire system continues until thermodynamic and structural properties become equilibrated (stable with time). Then the *production phase* commences for the time length desired, and the thermodynamic parameters can be calculated. Finally, the simulation will be *analysed* to calculate the properties of interest and to check whether there were any problems with the simulation [78].

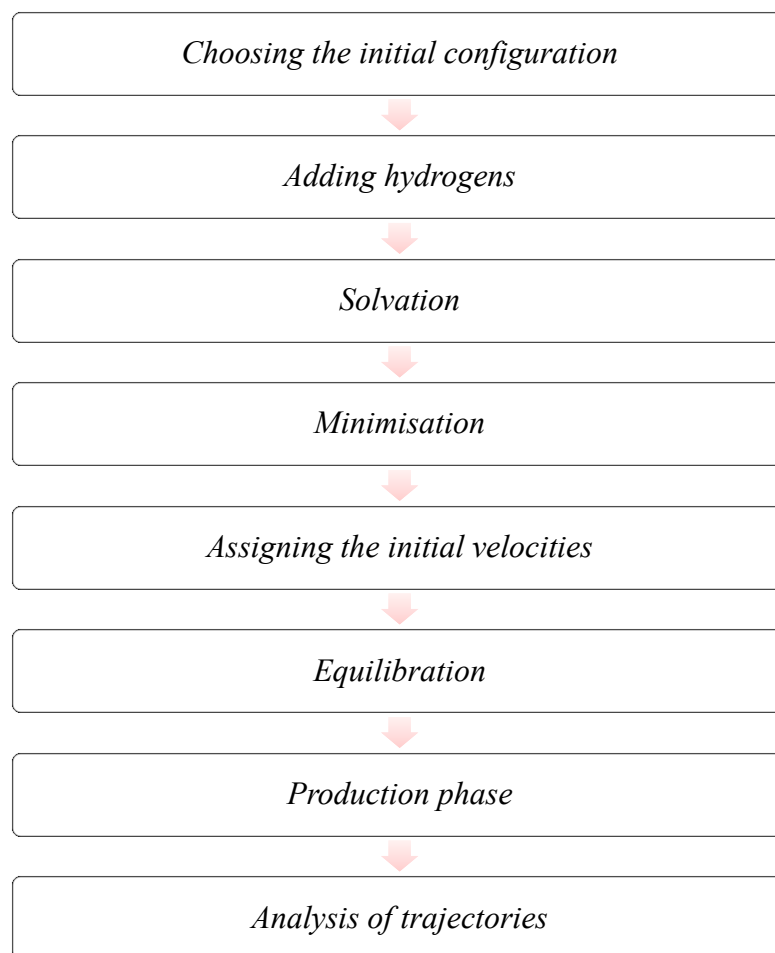


Figure 1.25: A schematic representation of the set up and the running of molecular dynamics simulations.

1.2.6 Use of grids in molecular modelling

The use of grids to measure some molecular properties has proved to be of high value in molecular modelling, which has resulted in them being intensively used and implemented in many software programs, especially for docking and 3D quantitative structure activity relationships (3D-QSAR). A grid is a preconstructed 3D lattice (figure 1.26) in which the molecule of interest is placed to measure its properties.

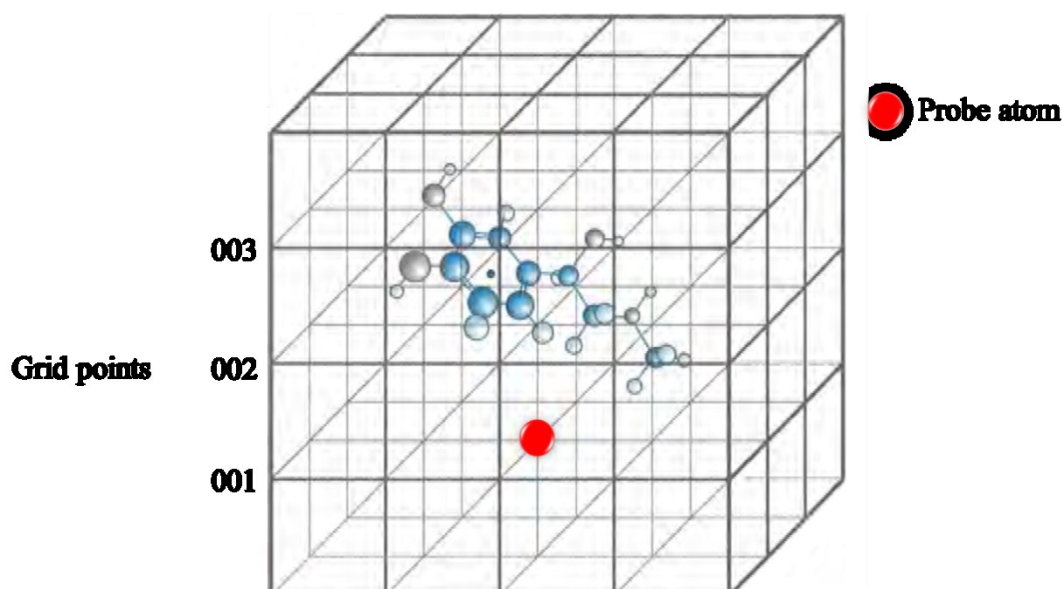


Figure 1.26: The use of grid to measure molecular properties by placing a probe atom at each grid point [64].

Some molecular properties can be measured as fields that influence the space surrounding the molecule; the most commonly measured fields are steric and electrostatic. To measure these fields, the molecule is placed inside the grid, a probe atom is placed at each grid point in turn, and the interaction between the probe atom and the molecule is measured. The probe atoms used to measure steric and electrostatic fields might be a proton or a sp^3 hybridized carbon atom. Hydrophobicity can be measured using water molecule as a probe. For steric fields, the interaction between the probe atom and the molecule will increase as the probe get closer to the molecule; and for electrostatic fields there will be an attraction when the positively charged probe nears an electron rich area in the molecule and a repulsion when approaching an electron poor area. Fields at each grid point are tabulated and a value for steric energy is chosen to define the shape of the molecule, where grid points having that value will be connected by contour lines to define the steric field of the molecule. A similar process applies for the electrostatic fields to define electron rich and electron poor regions on the molecule.

As mentioned above, the grid method has been intensively used in many docking software programs to measure the interaction between the docked molecule and the

active site of the protein. In this context, a grid is constructed within the binding site and different atoms such as H, N, C or O or fragments of molecules such as C=O or CO₂⁻ are used as probes to measure the interaction with the amino acids located within the binding site surface. The interactions here are typical binding interactions such as ionic, hydrogen bonding and van der Waals. It is possible to identify possible interactions with the binding site and to measure their strength, which in turn defines the binding characteristics of the active site. Once all these binding interactions have been calculated, they will be stored in a tabulated format for each atom or fragment probe (the calculations are intensive but they are performed only once). Therefore it is possible to calculate the binding energy of different docked molecules by identifying the atoms or fragments of the docked molecule that coincide with certain grid points by looking up the value of each interaction and summing them to give the total binding energy. In this way, it is possible to screen large libraries of small molecules within a reasonable time [64].

1.2.7 Docking

Molecular docking is a key computational tool in molecular modelling and is the most commonly used structure-based method in virtual screening. Docking software uses a search algorithm that generates a presumed mode of binding between the ligand and the receptor, along with a scoring function that ranks them by evaluating their interaction. There are many docking programs that can automatically dock ligands into a binding site and be used in virtual screening such as Autodock [87], LigandFit [88], DOCK [89], FLOG [90], FlexX [91], LIDAEUS [92], Glide [93] and GOLD [94]. These docking programs differ mainly in the sampling algorithm used to generate different conformations of the docked ligands, the scoring function, and the consideration of ligand and receptor flexibility [63].

The fact that ligands in complexes with macromolecules do not necessarily adopt a global minimum conformation, and that proteins show structural rearrangements upon binding of the ligand, necessitate the inclusion of flexibility in the docking algorithms. This will add to the complexity of the search algorithm because there is a need to account for fluctuations in bond lengths, angles and torsions besides the

rotational and translational degrees of freedom. In the case of macromolecules, their size and flexibility pose a significant challenge, and most current docking algorithms consider the receptor as a rigid body, while treating ligands as flexible by considering the degrees of freedom corresponding only to dihedral angles since the conformational space of a ligand is directly proportional to the number of dihedral angles in that ligand. Ligand conformations can be pre-generated as a library of conformations before the starting of the docking process (fixed docking); or can be generated as the docking process proceeds (flexible docking) [62]. Different methods have been employed in docking programs to explore the conformational space of ligands as the docking calculations proceed such as the use of a genetic algorithm [95], which mimics the processes of biological evolution such as mutations and crossover; Monte Carlo which is a stochastic simulation method that generates different conformations by random bond rotations [64]; the systematic search or stepwise bond rotation, which exhaustively enumerates all conformations resulting from rotating each rotatable bond [64]; and the incremental growth methods in which the molecule is divided into small fragments and then incrementally rebuilt in the binding site [63].

Once the docking program generates a pose (orientation or binding mode) for a ligand in the binding site, it needs to be scored to identify the most likely binding mode in terms of shape and physicochemical complementarity with the binding site. Based on this score, a ligand pose can be ranked with respect to other poses of the same ligand and with respect to other molecules in the database. There is a wide range of scoring functions that can be categorized into physical-based (force-field), empirical or knowledge-based. The physical-based scoring functions employ force fields to score ligand poses by estimating their binding free energy, and to reduce the time and complexity of calculations those force-fields are employed in a minimalistic manner on a grid with no explicit solvent to obtain a single point energy value as the score for the pose. Empirical scoring functions use physicochemical interactions such as H-bonding, ionic and hydrophobic interactions which are derived from fitting to known experimental binding energies for a variety of different protein–ligand complexes to estimate the binding energy of a ligand pose. Knowledge-based scoring functions are derived from protein–ligand atom pair interactions by extracting

structural information from a subset of protein-ligand complex structures from the PDB. Statistical mechanics is used to evaluate the binding free energy of a complex by summing the free energies (potentials of mean force) of all interatomic contacts based on their interatomic distance frequencies in the subset of experimental complexes [63, 96]. Table 1.1 lists some of the most commonly used docking programs along with the type of conformational sampling and scoring algorithm used.

Table 1.1: Docking programs commonly used in virtual screening.

Docking program	Ligand conformational sampling	Scoring function
AutoDock	Genetic algorithm	Force field
LigandFit	Monte Carlo	Empirical score
Dock	Incremental growth	Force field
FlexX	Incremental growth	Empirical score
Glide	Systematic search	Empirical score
Gold	Genetic algorithm	Empirical score

1.2.8 Visualisation

The 3D molecular structure is at the core of all molecular simulation studies and there is therefore a need for good visualisation software to provide a means of viewing structural data from different positions. There are a plethora of software programs that are designed to perform different modelling tasks ranging from visualisation to running MD simulations such as AMBER [80], CHARMM [77], VMD [97] and GROMACS[98], and all of them are capable of displaying molecules in relatively high resolution and in dealing with different file formats. Figure 1.27 shows the user-friendly interface of the Discovery Studio 3.1 client [86].

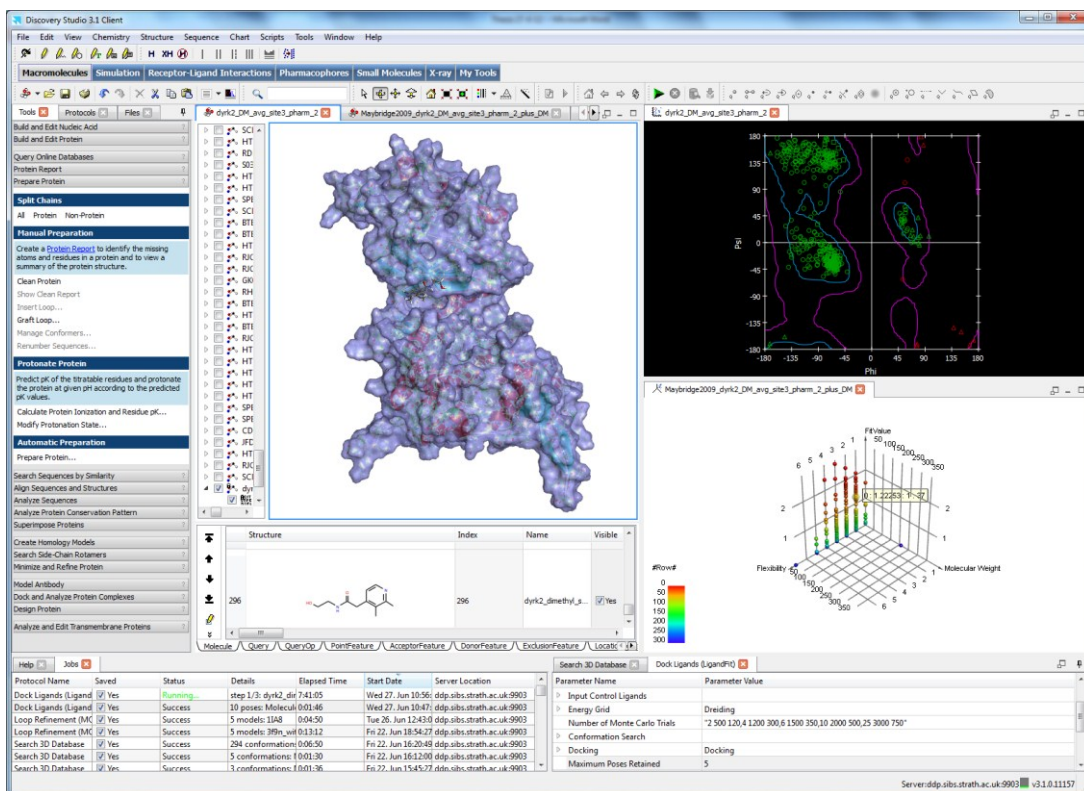
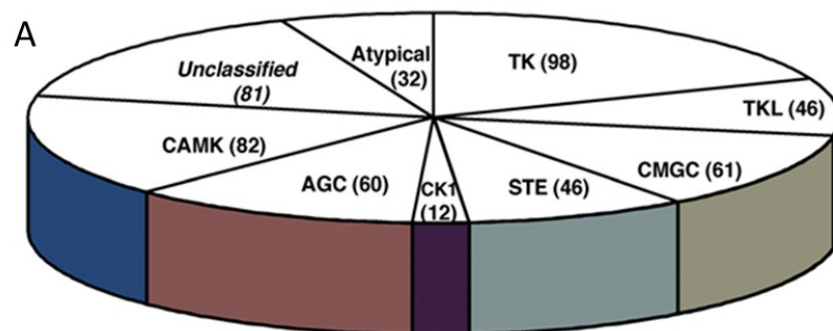


Figure 1.27: The graphical interface of Discovery Studio 3.1 client. In DS there are a plethora of functionalities that facilitate the visualization, manipulation and analysis of modelled systems such as advanced molecular visualizations; generation of 2D ligand receptor interaction diagrams; a variety of charts such as 3D point plots, heat maps and Ramachandran plots; sequence alignments and comparisons.

1.3 Kinases

1.3.1 Overview

There are 518 protein kinases encoded within the human genome, constituting 1.7% of the whole genome [1], which can be classified and categorized into subsets according to their sequence and structural similarities [1, 5] (figure 1.28). These kinases phosphorylate their substrate proteins through catalysing the transfer of the terminal γ phosphate group from an ATP molecule onto the hydroxyl side chain of a serine, threonine or tyrosine residues within the substrate protein, and because of these specific substrate recognition sites they are divided into serine/threonine kinases, tyrosine kinases, and dual-specificity kinases which have mixed serine/threonine and tyrosine kinase activity [2, 3]. There are also protein kinases that phosphorylate histidine. They are signal transduction enzymes mainly found in prokaryotes and lower eukaryotes, where they autophosphorylate on a conserved histidine within the kinase that is then transferred to a conserved aspartate in the so-called 'receiver domain' within the target protein [99-101]. In addition to protein kinases there is another group of kinases called lipid kinases which phosphorylate lipids at specific OH groups to generate a range of lipid second messengers [102-104]. Phosphorylation is part of the process that controls cellular signal transduction and kinases play a major role in many cellular events such as cell growth, differentiation, metabolism and apoptosis. Thus irregular kinase activity due to abnormal expression or mutation in the gene sequence is linked to many pathological conditions including cancer, inflammation, diabetes, autoimmune diseases, rheumatoid arthritis, psoriasis, atherosclerosis, neurological and metabolic disorders, and makes them a tractable target for drug discovery research [2-5].



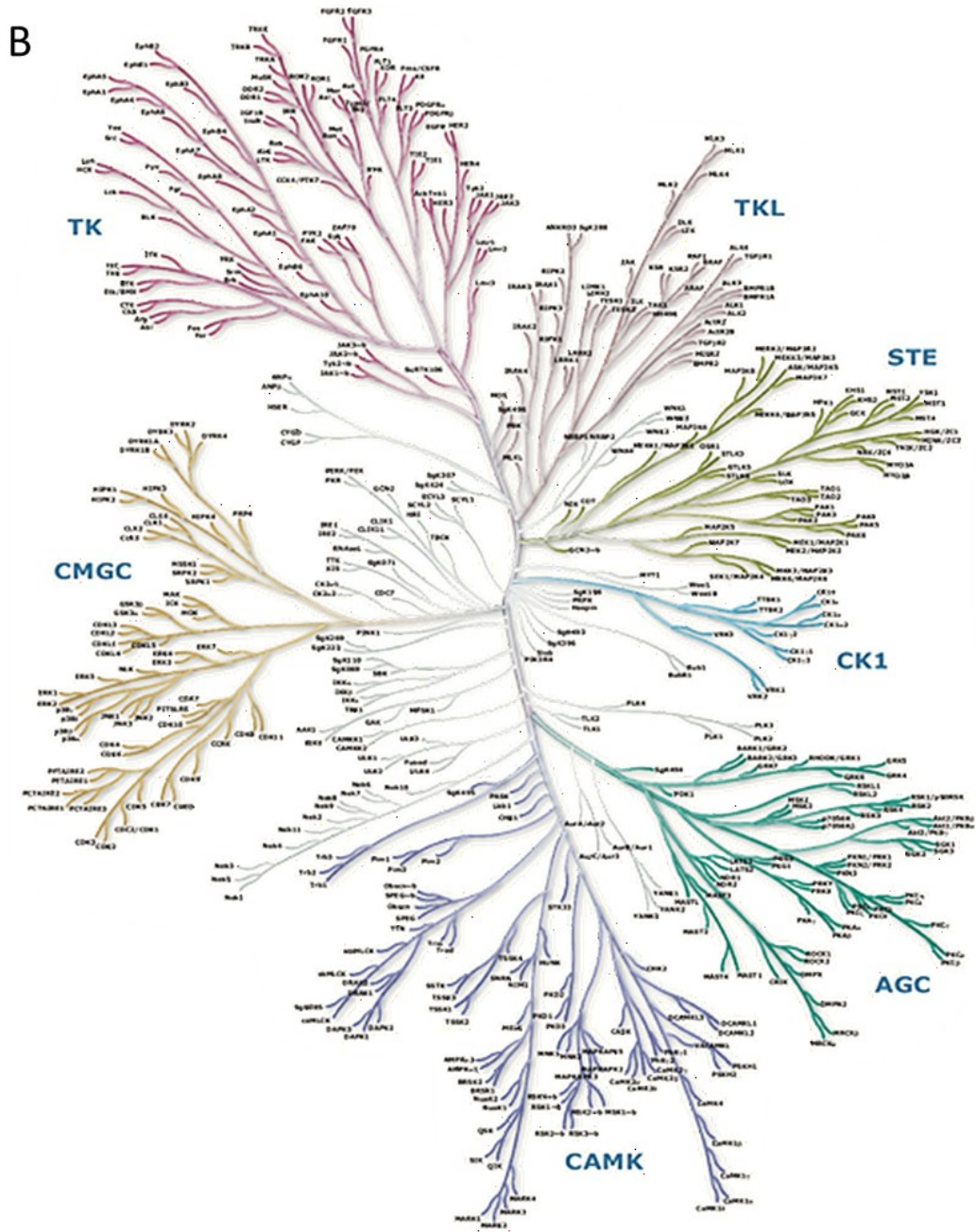


Figure 1.28: A) a pie chart representing the names of kinase subsets along with the number of kinases in each subset [5]. B) Phylogenetic tree of the human kinome highlighting the different subsets. The kinase subsets names are: TK: tyrosine kinase; TKL: tyrosine kinase like; CMGC: CDK, MAPK, GSK3 and CLK; STE: homologues of Sterile 7, Sterile 11 and Sterile 20 kinases; CK1: casein kinase 1; AGC: PKA, PKG and PKC; CAMK: calcium /calmodulin-dependent protein kinase. obtained from www.kinase.com/human/kinome according to reference [1].

1.3.2 Structural biology and substrate binding of kinases

In most cases kinases consist of at least two domains; the catalytic domain which contains the ATP binding site and is responsible for the phosphorylation of the target substrate, and a regulatory domain which allosterically modulates the activity of the catalytic one. The two domains could be in the same macromolecule or they might be separate molecules. The catalytic domain of the kinase is highly conserved among the entire kinase family due to the fact that all kinases recognize and bind ATP [4]. The catalytic domain has a bilobal structure, and consists of approximately 300 amino acids; the N-terminal domain is linked to the C-terminal domain via a hinge, and the ATP binding site is located between the two lobes. The N-terminal domain is the smaller of the two lobes and consists mainly of β -sheets and contains a conserved α -helix (α C-helix). In contrast, the C-terminal domain is comprised mainly of α -helices and contains the activation loop (figure 1.29).

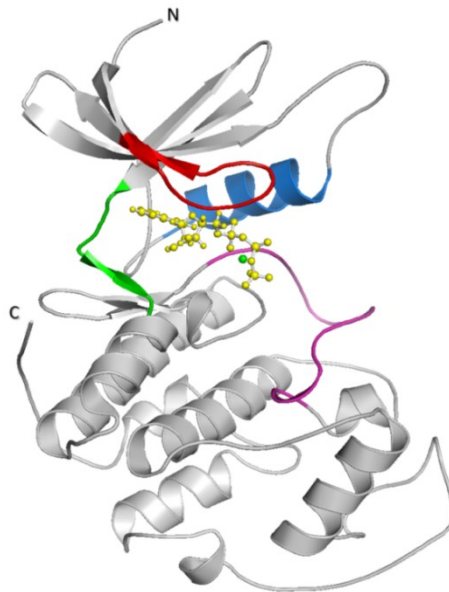


Figure 1.29: The structure of the catalytic domain of kinases (CDK2 kinase pdb code 2CCI) in complex with ATP showing the two domains (N and C) and the ATP binding site. The key features of the ATP binding site are coloured. ATP (shown in yellow ball and stick representation) is bound in the cleft between the N-terminus and C-terminus; the hinge is green; the glycine-rich loop is red; the activation loop (or the T-loop) is magenta; α C-helix is blue and the green sphere is Mg^{+2} ion.

The hinge region is crucial for forming the catalytic active site in which ATP is bound in a hydrophobic pocket; usually the hinge has one hydrogen donor flanked by two hydrogen acceptors that interact with the adenine ring of ATP (figure 1.30). The hydroxyl groups of the ribose moiety form hydrogen bonds with a polar residue located at the beginning of the C-lobe. The triphosphate group extends towards an opened hydrophilic region and interacts with a conserved lysine residue through hydrogen bonding. This lysine residue is held in position by forming a salt bridge with a conserved glutamic acid within the C-helix [105]. As a single residue, the gatekeeper is an important residue in the ATP binding site because the size of its side chain determines the access to the hydrophobic pocket located behind it (hence the name), thereby defining the selectivity of the ATP site for inhibitors (figure 1.30). Based on the gatekeeper's position, the two hydrogen bonding residues in the hinge are referred to as GK+1 and GK+3 (figure 1.31) [3].

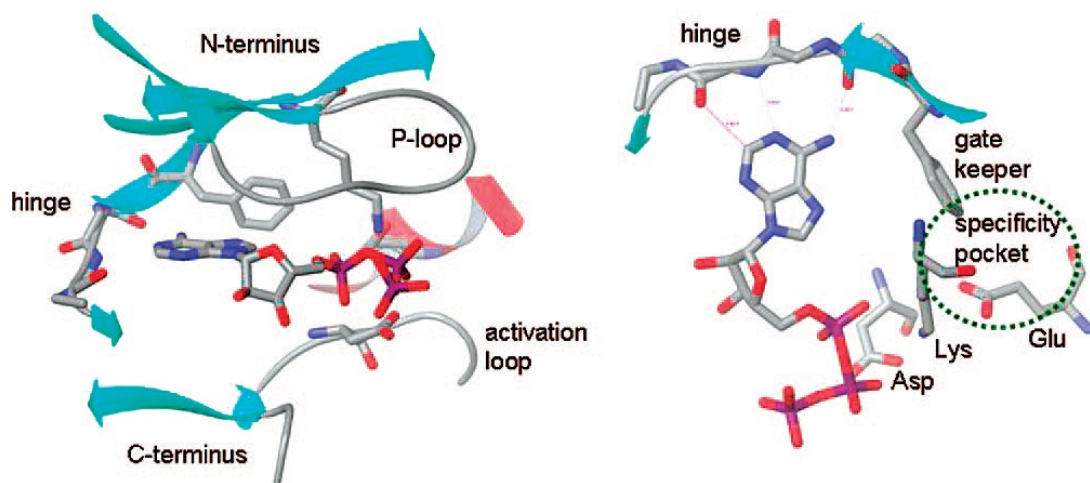


Figure 1.30: The structure and nomenclature of a kinase ATP binding site. Left: P-loop (glycine rich loop) forms the roof, and a C-terminus B sheet forms the floor. Right: the hinge region participates in hydrogen bonds with ATP or with competitive inhibitors. The gatekeeper and three other residues, the conserved lysine, glutamate (of helix C), and aspartate (of the DFG triplet, at the beginning of the activation loop) control access to hydrophobic specificity pocket (green dashed line) [3].

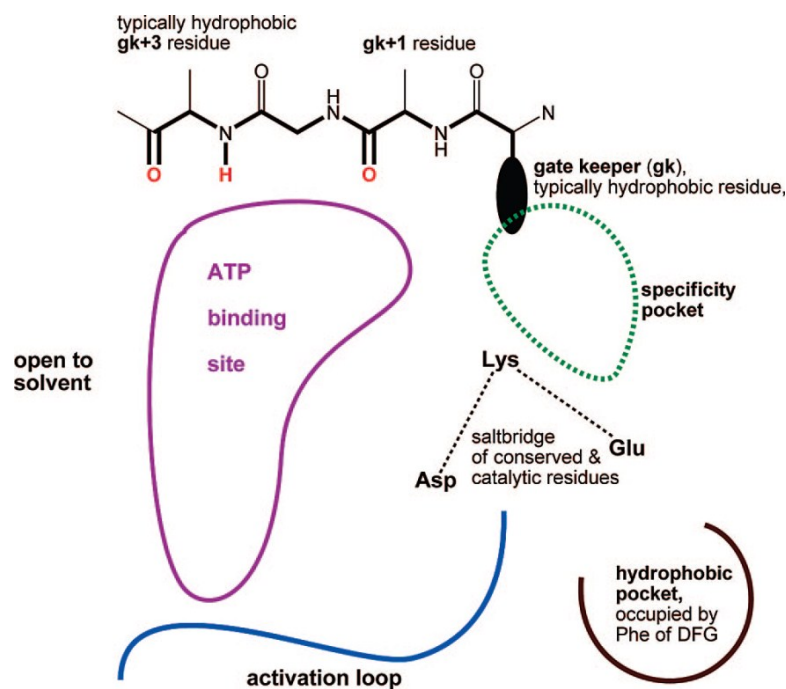


Figure 1.31: A 2D representation of the kinase ATP binding site with the key features being labelled. The kinase ATP site is generally hydrophobic except for the hinge region and the catalytic salt bridge regions [3].

1.3.3 Regulation and conformational flexibility of kinases

Kinases are highly regulated because of their crucial role in almost all cellular processes, and this regulation involves conformational changes in structure that affect the shape of the ATP binding site thereby determining the activity of the kinase. Of prominent importance in these regulatory mechanisms are the activation loop (T-loop), the glycine-rich loop and the C-helix. The activation loop contains serine, threonine or tyrosine residues which may be phosphorylated. When they are unphosphorylated, the loop partly occupies the ATP binding site, but when phosphorylated (by an upstream kinase), it becomes hydrophilic and moves toward the solvent exposing the site to allow ATP to bind [3, 106]. The N-terminal side of the activation loop begins with a highly conserved triad which is comprised of aspartate, phenylalanine and glycine (DFG motif), of which aspartate has a catalytic role in the transfer of the γ -phosphate group and also forms a salt bridge with a conserved lysine residue. The inactive closed state of the kinase is referred to as *DFG out*, and the active opened state is referred to as *DFG in*. In the inactive DFG

out state, the activation loop is compact and covers the substrate binding site [4], whilst the aspartate of the DFG motif is rotated out of the ATP binding site accompanied by rotation of the other two DFG residues. The movement of the phenylalanine side chain in turn reveals a hydrophobic pocket (allosteric site) that can be utilised to target more selective inhibitors [2, 3]. In the active state (DFG in), the activation loop opens up uncovering the ATP binding site of the kinase [4]. The activation loop is in equilibrium between these conformations [105].

The roof of the binding site is made of another loop with a conserved glycine rich motif GXGXXG, known as the glycine rich loop or the P-loop (figure 1.29) [3]. Which is less flexible than the T-loop but can wrap around the ligand for better binding [105]. Concerning the C-helix, its conformational changes can affect the activity of the kinase through modulating the orientation of the conserved glutamic acid which is involved in fixing the catalytic lysine residue in its proper position [105].

To summarize, kinases exist in two forms, the active and inactive form, which are a consequence of the conformation of the activation loop and the position of the DFG motif, glycine rich loop and the C-helix. Conformational changes occur as a result of substrate binding, the phosphorylation of the activation loop, or through induction by allosteric modulators [4] (figure 1.32).

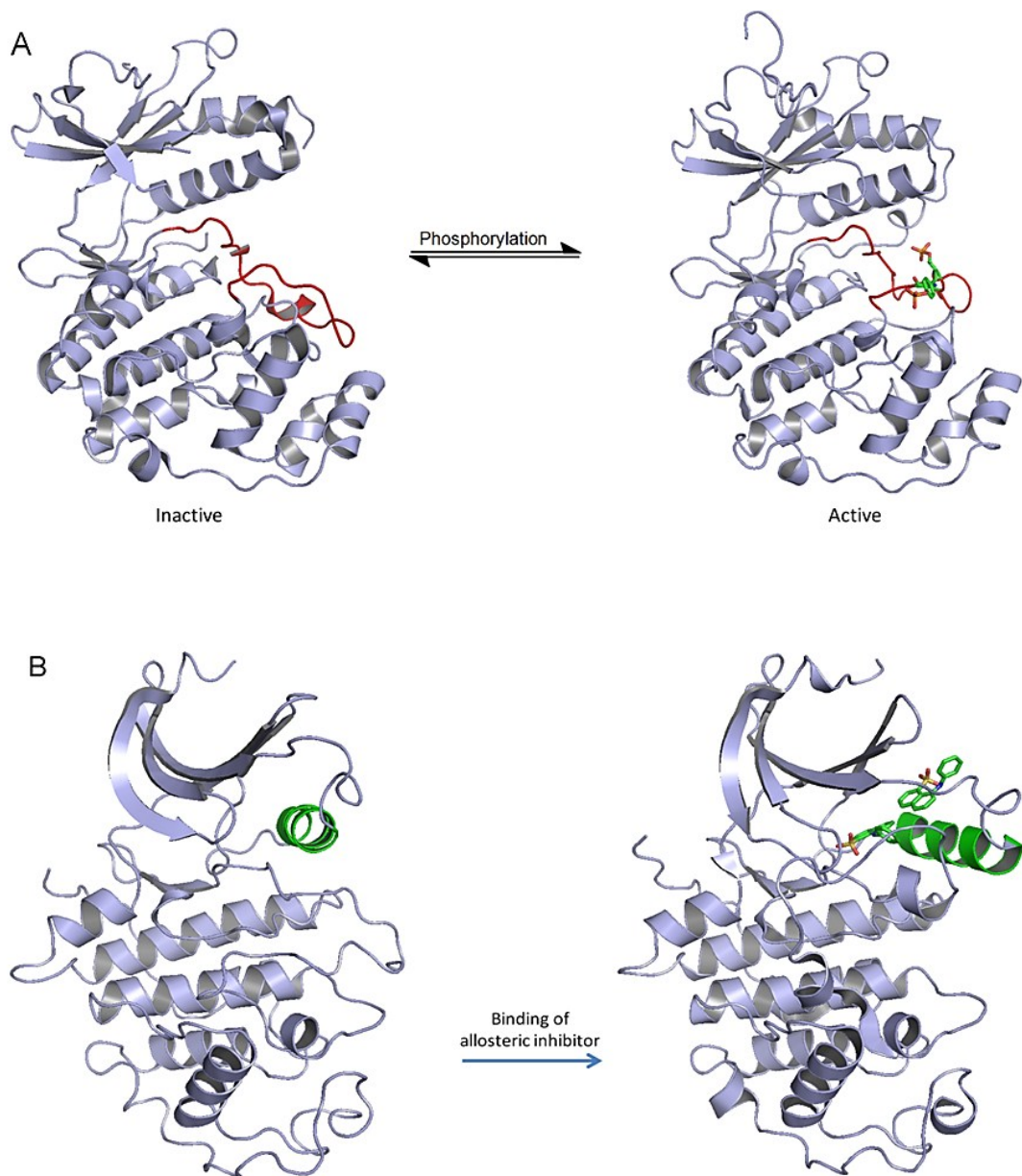


Figure 1.32: Schematic representation of the active and inactive conformations of kinases. A) Conformational changes in the extracellular signal-regulated protein kinase (ERK2) are induced by phosphorylation of the T-loop (red) [107] (PDB codes, 2Z7L left and 2ERK right). B) Conformational changes in cyclin-dependent kinase 2 (CDK2) are induced by binding of allosteric inhibitor [108] (PDB codes, 2CCI left and 3PXF right). Note the conformation of the α C-helix (green). (The ribbon representations were generated using PyMOL).

1.3.4 Small molecule inhibitors of kinases

1.3.4.1 Classification of small molecules inhibitors

Small molecule kinase inhibitors can utilize different regions of the ATP binding site (figure 1.33). Based on their mode of inhibition and the conformation of the target kinase (DFG in or out) they can be divided into:

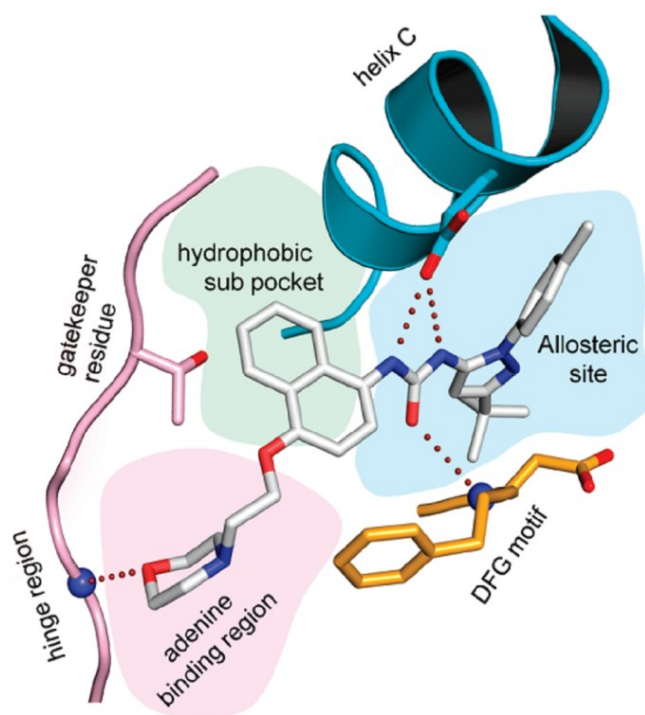


Figure 1.33: Schematic representation of the different regions of the ATP binding site that are utilised by different kinase inhibitors. This view shows BIRB-796, a type II inhibitor of p38 α (PDB code: 1kv2)[109].

a. *First generation or type I inhibitors:* these mimic the ATP binding to the kinase, and usually bind to the preformed ATP binding site in its active open form (DFG in) (figure 1.33). An example of this type is Sunitinib (Pfizer) (figure 1.34) which is a multikinase inhibitor approved by the FDA for treatment of gastrointestinal tumors. The main disadvantage of this type of inhibitors is the cross-activity within the target kinase family [2].

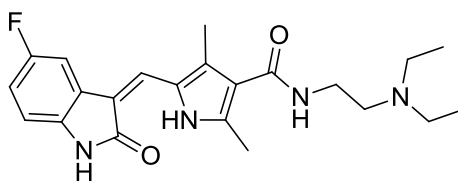


Figure 1.34: Chemical structure of sunitinib.

b. *Type II inhibitors:* these compounds bind and stabilize the target kinase in its closed inactive conformer (DFG out) preventing the binding of both ATP and the substrate protein. Compounds of this class bind to the same area as those of type I but also extend to the allosteric hydrophobic area present in the inactive conformation, and are more selective (figure 1.33). The main problem with type II compounds is their vulnerability to the emergence of resistance due to mutations in the catalytic domain. The first small molecule kinase inhibitor to reach the market and bind the inactive conformation of the enzyme was imatinib (Gleevec, Novartis) (figure 1.35) [2].

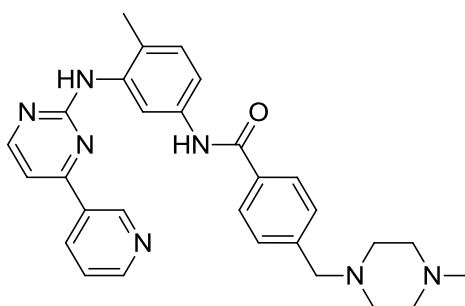


Figure 1.35: Chemical structure of Gleevec.

c. *Type I_{1/2} inhibitors:* these are a hybrid of both type I and type II inhibitors and they target the hydrophobic back cavity of the kinase in either the active (DFG in) or the inactive (DFG out) conformation. Like type I compounds, they bind the ATP binding site (hinge and adenine ring region), and they extend to the back cavity to interact with residues that are involved in type II inhibitors binding. The shape and size of the back pocket, (the hydrophobic sub pocket in

figure 1.33) is governed by the nature of the gatekeeper residue and a small gatekeeper residue results in a larger and more accessible cavity, and this class of inhibitors target kinases of known small gatekeeper residues. An example of this type of inhibitors is compound AP23464 (ARIAD) (figure 1.36), which targets Src tyrosine kinase with picomolar affinity [2].

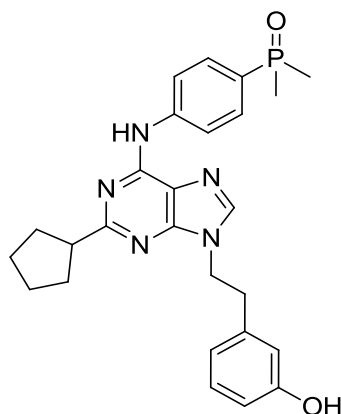


Figure 1.36: Chemical structure of compound AP23464.

d. *Type III inhibitors:* they bind to allosteric sites distinct from the ATP binding site. Binding to these sites can modulate the activity of the enzyme in either a negative or positive way by changing its conformation. Because these inhibitors do not compete with ATP, they have the advantage of greater selectivity and possibly higher affinity. An example is compound BMS-006 (Bristol-Myers Squibb) (figure 1.37) which inhibits IKK- β and has shown promising results in preclinical studies to treat rheumatoid arthritis [4].

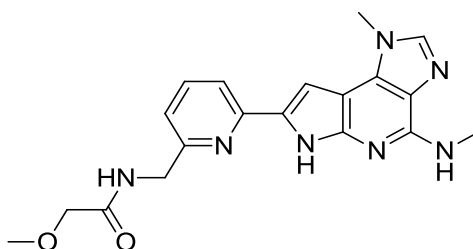


Figure 1.37: Chemical structure of BMS-006 [110].

e. Type IV inhibitors: they covalently inhibit kinases by forming a covalent bond with the ATP binding site residues, usually by targeting cysteine residues in the active site via a nucleophilic reaction. An example of this type is the potent HER-2 inhibitor HKI-272 (figure 1.38) which is in clinical development for the treatment of HER-2 induced breast cancer [110, 111].

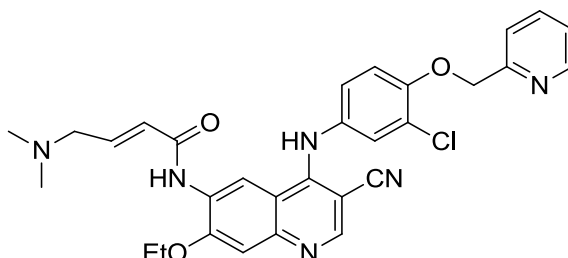


Figure 1.38: Chemical structure of compound HKI-272 [111].

1.3.4.2 Approved small molecule kinase inhibitors

The last few decades have witnessed intensive research targeting kinases as potential therapeutic targets. However, only eight compounds have found their way to the market as small molecule kinase inhibitors after being approved by the FDA for cancer treatment (table 1.2). All of these eight compounds compete with ATP either directly (type I): sunitinib, erlotinib, gefitinib, dasatinib and lapatinib; or indirectly (type II): imatinib, sorafenib and nilotinib[2].

Table 1.2: Small molecule kinase inhibitors approved by the FDA for cancer treatment.

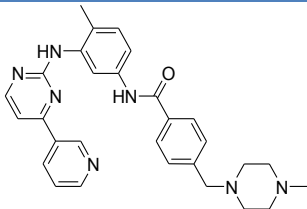
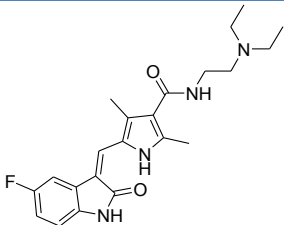
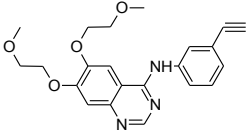
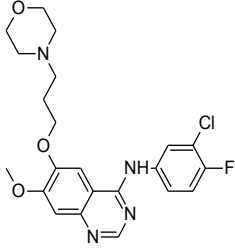
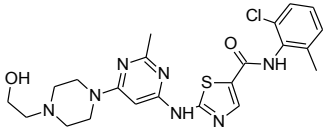
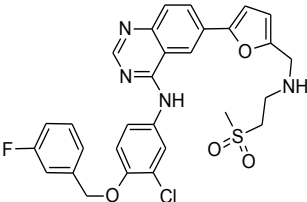
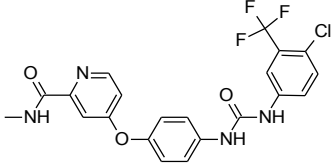
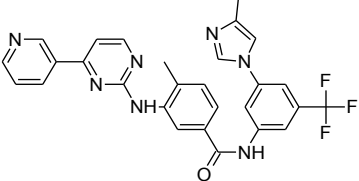
Chemical structure, name, and manufacturer	Indications
 <p><i>Imatinib</i> (Gleevec®, Novartis)</p>	<p>Approved on December 20, 2002 for the treatment of Philadelphia chromosome-positive chronic myelogenous leukemia (CML) [112].</p>
 <p><i>Sunatinib</i> (Sutent®, Pfizer)</p>	<p>Approved on January 26, 2006 for the treatment of patients with imatinib refractory or intolerant gastrointestinal stromal tumors (GIST) [113, 114].</p>
 <p><i>Elrotinib</i> (Tarceva®, OSI and Genentech)</p>	<p>Approved on November 18, 2004, for the treatment of patients with locally advanced or metastatic non-small cell lung cancer (NSCLC) [115].</p>
 <p><i>Gifitinib</i> (Iressa®, AstraZeneca)</p>	<p>Approved on May 5, 2003 as monotherapy for the treatment of (NSCLC) after failure of other treatments [116].</p>

Table 1.2: Continued...

Chemical structure, name, and manufacturer	Indications
 <p><i>Dasatinib</i> (Sprycel®, Bristol-Myers Squibb)</p>	<p>Approved on June 28, 2006 for the treatment of adults with chronic phase, accelerated phase, or myeloid or lymphoid blast phase chronic myeloid leukemia (CML) or philadelphia chromosome-positive acute lymphoblastic leukemia (Ph⁺ ALL) [117].</p>
 <p><i>Lapatinib</i> (Tykerb®, GSK)</p>	<p>Approved on March 13, 2007 for the treatment of breast cancer [118].</p>
 <p><i>Sorafinib</i> (Nexavar®, Bayer)</p>	<p>Approved on December 20, 2005 for the treatment of patients with advanced renal cell carcinoma (RCC) [119].</p>
 <p><i>Nilotinib</i> (Tasigna®, Novartis)</p>	<p>Approved On October 29, 2007 for the treatment of chronic-phase and accelerated-phase chronic myelogenous leukemia (CML) resistant to or intolerant of imatinib [120].</p>

1.4 Rationale and objectives of the project

The crucial role of kinases in controlling cellular processes makes them a tractable target for drug discovery projects. However most of the current medicinal chemistry efforts target the ATP binding site, which is highly conserved amongst the kinase family, and many compounds suffer from cross-activity leading to undesirable side effects and toxicity. Compounds that are capable of binding the catalytic domain of the kinase with greater selectivity to block ATP binding and inhibit activity are therefore required.

An appealing, yet challenging approach to achieve selective inhibition is through identifying new allosteric sites on the target enzyme and developing allosteric modulators that target these sites. This approach offers many advantages over competitive ATP inhibition, such as:

1. In contrast to the ATP-binding site, which is highly conserved among the protein kinase family, allosteric sites are more diverse because they are much less conserved, and represent greater opportunities as drug targets. Hence, targeting allosteric sites increases the possibility of selective modulation by small molecules within the same enzyme family with fewer side effects [11, 38]. Also as a result of the high conservation of the ATP binding site, the intellectual property scene for kinase inhibitors is highly congested, making it difficult to find new chemical entities[121] .
2. Allosteric modulators offer a higher degree of specificity, which is important especially within a highly homologous kinase subfamily of high sequence similarity [4].
3. Allosteric inhibitors may regulate excessive kinase activity without affecting the basal activity of the enzyme, and preserve the beneficial function while blocking the harmful effect of the overstimulated kinase [4].
4. Allostery provides the opportunity to develop allosteric activators compared to other compounds targeting kinases which are almost all inhibitors. Allosteric activators can be used in cases where the normal stimulatory pathway is blocked,

or when activating a targeted kinase can negate the deleterious effect or abnormalities of another kinase. [4].

5. Using allosteric modulators as therapeutic agents for producing a physiological response has the advantage of offering a much easier approach for producing orally active drug-like inhibitors. Competitive inhibitors often mimic an enzyme's transition state substrate and are often non-drug like, i.e. usually charged or very non polar, which makes orally active drug-like inhibitors a challenging process [16].

1.5 Future experimental plan

Here we propose a computational approach to identify allosteric sites in target kinases. We use a combination of MD simulations to explore the critical structural and dynamic conformational changes of the enzymes and simple intrasequence differences (SID) analysis which identifies the major interfaces in the enzyme that may be involved in allosteric modulation. Potential allosteric binding sites will be identified through correlation analysis based on the simulated kinases. This computational approach provides not only a new method of identifying allosteric sites but also a better understanding of the mechanisms of allosteric modulation of target kinases and the structural basis for the design and development of more selective and specific small molecules inhibitors as therapeutic agents.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Computational materials

MD simulations were performed using the molecular dynamics software package AMBER (Assisted Model Building with Energy Refinement) versions 10 and 11 [80, 122, 123]. Simulations were run on a range of computational hardware to benchmark performance and ensure that the lengthy calculations involved could be carried out as quickly and reliably as possible. In the first instance, the UK National Grid Service (NGS) was used with 16 processors dedicated to each MD job. The surprisingly slow computational performance observed prompted upgrading of in-house facilities including: a 6-node Hewlett-Packard cluster, with each node consisting of 2x3.2 GHz quad core opteron CPUs and at least 16 Gb of RAM interconnected using a low latency Myrinet Myri10-G switch; and a Viglen GPU server containing 1x6-core Xeon 2.66 GHz CPU, 12 Gb RAM and 2 x Nvidia Tesla M2050 GPU cards. Both in-house machines were running the 64-bit Red Hat Enterprise Linux Server 5.5 Operating System (OS). The approximate relative performance in hours per nanosecond of simulation for the NGS 16 processors, the in-house cluster of 48 processors and the in-house 2 GPUs (using the same system) was 50 hours, 8 hours and 10 hours respectively; which shows that the in-house machines were about 5 times faster than the NGS. Moreover, increasing the number of NGS processors was resulting in a much slower performance, suggesting a messaging problem.

Preparation of the starting protein structures for MD simulations along with the virtual screening studies were performed using Discovery Studio (DS) 2.5 and 3.1, and Pipeline Pilot (PP) 7.5 and 8.5 from Accelrys® Software Inc. [86].

Docking of retrieved hits was performed using GOLD from the Cambridge Crystallographic Data Centre (CCDC) [94, 124]. The server hosting DS, PP and GOLD runs 64-bit Red Hat Enterprise Linux Server release 5.3 and has four quad core Xeon 3.4 GHz CPUs and 16 Gb RAM.

Presentation quality images were generated using the PyMOL Molecular Graphics System [125]. Contact maps were generated using the CMview plugin [126] for

PyMOL. Energy correlation analysis was performed using statistical analysis software MATLAB, version 7.12.0 [127].

2.1.2 Experimental materials

DYRK2 Kinase activity was determined using an enzyme-linked immunosorbent assay (ELISA) kit - CycLex DYRK2 Kinase Assay (Caltag Medsystems, Buckingham, UK). Materials supplied with the kit included a 96-well microplate coated with recombinant p53 N-terminus (1-99 amino acids) substrate, kinase buffer, wash buffer, ATP, horseradish peroxidase (HRP) conjugated anti-phospho-p53 S46 (TK-4D4) antibody, chromogenic substrate - tetra-methylbenzidine (TMB), and stop solution (1N H₂SO₄). Human recombinant DYRK2 (Caltag Medsystems, Buckingham, UK) was purchased in addition to the kit.

Optical density was measured at 540nm on a Spectromax 190 absorbance plate reader (Molecular Devices, Wokingham, UK). The IC₅₀ of the phosphorylated substrate was calculated for each compound using the nonlinear curve fitting program Graphpad Prism 4 (Sigma software, Ashburton, UK).

2.2 Computational methods

2.2.1 Molecular dynamics (MD) simulations

All methods described in this section are specific for AMBER 10. Running a simulation using AMBER (and most other MD software) requires the following [80, 122]:

1. Cartesian coordinates for each atom in the system. They are usually obtained from X-ray crystallography, NMR spectroscopy, or model-building and are stored in Protein Databank "pdb" or Tripos "mol2" file format.
2. Topology information which includes connectivity, atom names, atom types, residue names, and charges. This information is stored in a database within the program. It contains topology information only for standard residues such as standard amino acids, DNA, RNA, and common sugars. For nonstandard molecules topology information is user-generated.

3. A force field which is the functional (mathematical) form and the set of parameters for all of the bonds, angles, dihedrals, and atom types in the system used to calculate its potential energy.
4. Commands which are issued by the user in an input file to specify the procedural options and the desired parameters.

AMBER includes a set of functionalities and programs that are designed to read and build molecular topologies; carry out minimisation and MD simulations; and analyse MD results (figure 2.1). These programs can be categorized as follows: (i) *Preparatory programs*: such as LEaP which is used to create new systems or modify old ones and to generate coordinates and topology files; and *Antechamber* which is used to generate structural information (parameters) for nonstandard residues; (ii) *Simulation programs*: there are two simulation programs in AMBER, the first is SANDER which is the basic energy minimiser and MD program (it is the MD engine); the second is PMEMD (Particle Mesh Ewald Molecular Dynamics) which is a modified version of SANDER for speed and parallel scaling; and (iii) *Analysis programs*: the main analysis program in AMBER is *PTRAJ*, which is a general purpose facility for analysis and processing of MD trajectories [80, 122]. The following procedure was used for setting up structural models for systems of interest:

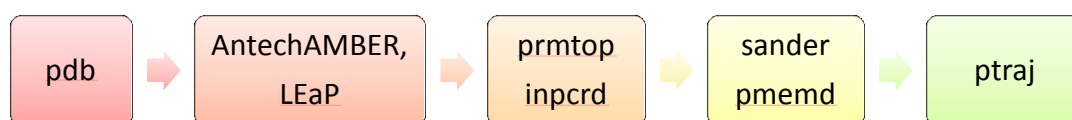


Figure 2.1: A simplified workflow in AMBER. A pdb file of standard residues is loaded into LEaP (or to Antechamber if it has nonstandard residues and then to Leap) to generate topology and coordinate files that are used by sander to generate a trajectory for the system, which can be later analysed by *ptraj*.

2.2.1.1 Preparation of the simulated models

The structural models of the studied systems were prepared using Discovery Studio 2.5.5 (DS) from Accelrys; specific details for each system will be provided later. The

initial coordinates for each structural model were retrieved directly from its corresponding crystal structure from the Protein Data Bank (PDB). In some of the crystal structures used there were missing loops that need to be inserted and treated before running the simulation.

2.2.1.2 Treatment of missing loops in the starting crystal structures

All missing loops or protein segments in the starting crystal structures were inserted using DS via applying the following procedure: (1) the loop was built using the *Build and Edit Protein* tools within DS which adopt the random tweak algorithm developed by Shenkin et al. [128] whereby the missing loop is built with conformations of no or minimal bumps with the rest of the protein by iteratively sampling closed loop conformations. (2) Cleaning and typing the protein using *Protein Report and Utilities* and the *Simulation* tools within DS. (3) Loop refinement using *Loop Refinement (MODELER)* protocol within DS. The default parameters of the protocol were used. The MODELLER energy function can be used to refine one or more loop regions in a protein while other parts of the protein structure are held rigid. The generated models are sorted by the total Probability Density Function (PDF) energy of each model. The PDF total energy is used to give an indication of the quality of the model compared to the other models and the lower the value the better the quality of the model.

2.2.1.3 Generating the structural coordinates and topology files

Once the initial pdb structures are prepared, the next step is to start building the necessary input files to be processed by *Sander* (the MD engine in AMBER). To run a MD simulation using AMBER all of the residues within the studied system must be pre-defined in the AMBER database (standard residues). For non-standard residues we need to provide structural information and force field parameters for them before creating the input files. Those parameters were generated using *Antechamber* program which utilise the General AMBER Force Field (GAFF) [129, 130].

Afterwards, *Sander's* input files were created using AMBER 10's *Xleap* module. Since the AMBER MD package has a range of different force fields dedicated for

different types of simulations, it is important to explicitly specify which force field is to be used. In AMBER 10, the force fields recommended for the simulation of proteins and nucleic acids in explicit solvent are either the FF99SB or FF03 force fields. In this study we used the FF99SB force field [80]. In order to create the input files, the following steps were carried out sequentially within *Xleap*. Firstly, hydrogens were added at pre-determined positions. Secondly, each protein was solvated by immersing it in a truncated octahedral box of pre-equilibrated TIP3P water [131]. Thirdly, counter ions were added at positions of high electric potential around the molecule in order to neutralize the system. Finally, the input files were created and saved. The input files we created using *Xleap* are the *prmtop file* which is the parameter/topology file that defines the connectivity and parameters for our model. The information in this file does not change during the simulation. The other input file is the *inpcrd file* which contains information about the coordinates and velocities. Data in this file are updated during the simulations to store new coordinates and velocities.

2.2.1.4 Simulation protocol

Once the initial structures and the input files have been created, the simulation process is started. AMBER allows the simulation of the energetics and dynamics of a molecular system; in this study it was used to perform energy minimisations and MD simulations.

2.2.1.4.1 Minimisation

Since the starting geometry of the molecule determines the quality of the subsequent analysis, it should always be optimised to find the energy minimum state which corresponds to a stable conformation of the molecular system. Finding the low energy conformation is achieved via *energy minimisation*, whereby the atoms are moved from a starting non-equilibrium molecular geometry so as to reduce the net forces (the gradient of potential energy) on the atoms until an equilibrated conformation with negligible forces is obtained. In this study minimisation was applied to relax the system and remove all potential collision contacts between the protein and the added solvent and ions. It is important to minimise the simulated

structure before running molecular dynamics because un-minimised structures will result in instabilities when running MD. For example, the initial structures in our simulations were a gas phase (isolated) representation with no protons; we used *Xleap* to add these protons and then solvated the systems, and later neutralized them by adding suitable counter ions. The added water has not felt the influence of the solute or the counter ions and more importantly there may be gaps between the solvent and the protein and solvent and box edges. These gaps in turn can lead to vacuum bubbles within the system and eventually to instability in the MD simulation. Therefore, a careful minimisation of the simulated system must be completed before running MD.

The minimisation procedure for the solvated systems consisted of three stages, each of 50,000 steps where the first 250 steps utilised the steepest descent algorithm and the remaining steps used the conjugate gradients algorithm. In the *first stage* only hydrogens, water and counter ions were allowed to move while all other atoms in the system were restrained by applying a restraining force of $100 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ on each heavy atom of the simulated system. The strong positional restraints on each of the heavy atoms keep them more-or-less fixed in the same position. Such restraints use a harmonic potential to restrain the selected atoms to a reference structure, in this case our starting structure. In the *second stage*, only backbone atoms were restrained by applying a force of $100 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$, while all other atoms were allowed to move. In the *last stage* no restraints were applied and the entire system was allowed to move.

2.2.1.4.2 Equilibration

The first stage of the MD simulations is the equilibration phase, which enables production dynamics at a constant temperature and pressure that closely resembles laboratory conditions. However, during the first few ps of simulation systems will still be at low temperatures, and the calculation of pressure at this stage is very inaccurate because the density of the system has not yet equilibrated, and using constant pressure periodic boundaries can lead to instabilities in the generated trajectory. Therefore, MD simulations were initially run at constant volume and once equilibrated switched to constant pressure.

In this study, the equilibration phase was subdivided into four stages. In all four stages a weak restraint of $10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ was applied to ensure that temperature increased from 0 to 300 K without any wild fluctuation in the solute (protein). In the *first stage*, the system was gradually heated to 100 K over the course of 20 ps under constant volume conditions (NVT, the canonical ensemble).

In classical thermodynamics of main importance is the bulk properties of the system such as the number of particles (N), the temperature of the system (T), and the volume of the container (V); rather than properties of the particles. Based on this an ensemble can be defined as a collection of systems that have different microstates but have the same thermodynamic or macroscopic state. In MD simulations trajectories are generated from which time average properties of the system are calculated, but experimental observables such as thermodynamic properties are ensemble averages rather than time averages. However, the Ergodic hypothesis [132] states that an ensemble average equals a time average, provided that the MD simulation is long enough to sample enough representative conformations. Four major types of ensembles are commonly used in MD simulations: the microcanonical ensemble NVE in which the thermodynamic state is characterized by the fixed number of atoms (N), fixed volume (V), and fixed energy (E); the canonical ensemble NVT in which number of atoms (N), volume (V) and temperature (T) are fixed; the isobaric-isothermal ensemble NPT where number of atoms, pressure (P), and temperature are fixed; and finally the grand canonical ensemble μ VVT where the number of atoms, temperature and chemical potential (μ) are fixed [70, 133].

In the *second stage* of equilibration the temperature was gradually raised to 200 K over the course of 40 ps under constant volume (NVT) conditions. In the *third stage* the system temperature was gradually raised to 300 K over the course of 60 ps under constant volume (NVT) conditions. In the *last stage* the temperature was kept at 300 K for another 60 ps but switched to constant pressure (NPT) rather than constant volume to equilibrate the entire system and to achieve density equilibration in order to obtain a system ready for running production dynamics.

In order to maintain good control over temperature and pressure in MD simulations and to run the simulations in as realistic conditions as possible, different algorithms

have been developed that couple the system to temperature and pressure baths. These algorithms are divided into velocity rescaling approaches such as the Berendsen [134] and Nose-Hoover [135] thermostat; and velocity modification approaches such as the Andersen [136] and the Langevin [137] thermostats. The Langevin dynamics scheme [137] was used in all the equilibration stages. In the Langevin scheme, the temperature of the system is maintained by modifying the velocities of the particles in the simulations with pseudorandom forces where they undergo stochastic collisions with imaginary particles whose momenta follow a Maxwell distribution at a given reference temperature.

2.2.1.4.3 MD production

After equilibrating the simulated systems the *MD production phase* commenced. Although the application of the Langevin dynamics scheme ensures even distribution of temperature over the entire system, it unfortunately alters the fast dynamics of the studied system. Since the interest here is in studying the correlation functions of the simulated systems, the first part of the production phase was run using Langevin dynamics and once the system is very well equilibrated we shifted to Berendsen dynamics [134] where the simulation was continued at constant temperature and pressure (NPT). In the Berendsen scheme, the system is weakly coupled to an external bath where the temperature is maintained by scaling the velocities in each step of the simulation using equation 2.1:

$$v' = \lambda v, \quad \lambda = \left[1 + \frac{\Delta t}{\tau_T} \left(\frac{T_0}{T} - 1 \right) \right]^{\frac{1}{2}} \quad (2.1)$$

where, T_0 is the reference temperature of the heat bath, τ_T the coupling constant, and Δt the integration time step. The same is applicable to pressure. Therefore, the simulation was run in isobaric-isothermal NPT ensemble conditions.

2.2.1.5 Treatment of solvent effects in MD simulations

In order to simulate the properties of a bulk solvent all of the simulations in this study were performed under periodic boundary conditions [79] using a truncated octahedral box of pre-equilibrated TIP3P water [131].

2.2.1.6 Reducing the computational complexity

All of the simulated systems in this study were solvated using explicit water molecules which significantly increases the number of atoms, adding to the computational cost. As a result, it is important to reduce the computational complexity as much as possible. Systems were solvated using the triangulated water TIP3P model, in which the angle between the hydrogens (H-O-H) is kept fixed. Using such a water model necessitates restraining the motion of the hydrogen atoms of the water to avoid inaccuracies in the calculation of the densities. Also the motions of hydrogen atoms in the protein itself were constrained since their motion is of little importance in MD. This was achieved using the SHAKE algorithm [85] to constrain all bonds involving hydrogen which allows the MD time step to be increased to 2 fs from 1 fs without introducing any numerical instability into the simulation.

2.2.1.7 Treatment of the non-bonded cut off values

In all steps of the MD simulation the long-range electrostatic interactions were treated by the particle mesh Ewald method [138] using a non-bonded cut off value of 15 Å [139, 140]. Snapshots of the simulations were saved every 1 picosecond for analysis. A full discussion of the input files that were used to run the MD simulations is in appendix II.

2.2.1.8 Analysis of the MD trajectories

The generated trajectories were analysed using the *ptraj* module of AMBER 10. To use *ptraj* it is necessary to (i) read in a parameter/topology file, (ii) set up a list of input coordinate files, (iii) optionally specify an output file, and (iv) specify a series of actions to be performed on each input set of coordinates (figure 2.2). *ptraj* as used to calculate the average structure, the root-mean-square-deviation of generated structures from a reference, the atomic positional fluctuations, and the residual dynamic correlation matrices. This was done as follows:

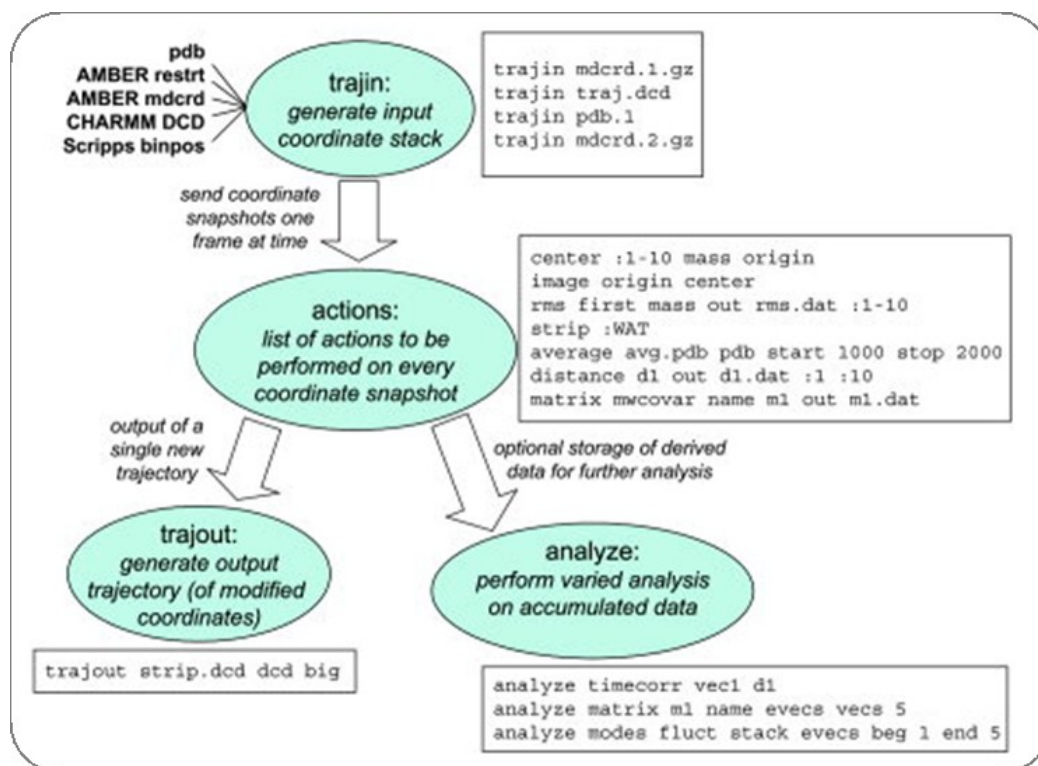


Figure 2.2: A schematic representation of the workflow in *ptraj* [122].

Average structure:

In MD simulations time averages such as average structure, average potential energy, etc. can be calculated. An average structure of a number of conformations (a trajectory) of the same protein is defined by the average value of the coordinates of the N atoms in the system, and is obtained by the following equation:

$$\langle X_i \rangle = \frac{1}{M} \sum_{j=1}^M X_{ij} \quad (2.2)$$

where X corresponds to the coordinates; i corresponds to the atom, j corresponds to the conformation; and M is the total number of conformations for which an average structure was calculated. Since the Ergodic hypothesis states that an ensemble average equals the time average, the average structure of a MD trajectory can be considered representative of the microstate of that system [133, 141].

Root mean square (RMS) deviation between structures:

The RMSD is the most commonly used measure of structural similarity (spatial difference) of macromolecules, defining the difference between two structures or two conformations of the same structure by a single value (equation 2.3) [142, 143]:

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i^X - r_i^Y)^2} \quad (2.3)$$

where N is the number of atoms, i the current atom, r^X the target structure, and r^Y the reference structure. The reference structure can be the starting structure of the simulated system, the average structure of the simulation or any selected frame (conformation) of the simulation. In this study the starting structure (the energy minimised crystal structure) was used as a reference. The target structure is usually all the succeeding frames in the trajectory. In order to generate accurate results the starting and the average structures were aligned before calculating the RMSD; only the backbone atoms (C, C α and N) were considered. The RMSD values along the trajectory were then represented as a line plot to examine the convergence and stability of the simulation.

Care needs to be taken when calculating RMSD values since loops and free termini can add to the overall value giving the impression the system is not well equilibrated. Therefore, it is advisable to perform RMSD calculations based on fit utilising the well-formed secondary structures of the protein.

Atomic position fluctuations:

The atomic positional fluctuations in the studied systems were calculated from their corresponding trajectories using the atomic fluctuation facility of *ptraj* (equation 2.4):

$$\text{rmsf} = \sqrt{\frac{1}{N} \sum_{i=1}^N \langle (r_i^t - r_i^{\text{ref}})^2 \rangle_t} \quad (2.4)$$

where, r_i corresponds to the x,y,z coordinates of atom i at time t of the simulation and atom i in the *reference* structure respectively, the $\langle \dots \rangle_t$ corresponds to time average [144]. So the RMSF can be viewed as the standard deviation of the atomic positional

movements throughout the simulation time. In this study the average fluctuations (RMSF) were calculated for the backbone atoms (C, C α and N) of the studied proteins.

Residual dynamic correlation matrices:

The residual correlation analysis of the generated trajectories was based on the calculation of the pairwise residual dynamic correlation matrices generated using the correlation and fluctuation tools within *ptraj*. The correlation matrix $Corr_{ij}$ is a diagonally symmetric N x N array, whose elements C_{ij} describe the cross correlation between atom i and atom j and is given by equation 2.5 [145-147]:

$$C_{ij} = \frac{\langle (\vec{r}_i - \vec{r}_{i,ave}) \cdot (\vec{r}_j - \vec{r}_{j,ave}) \rangle}{\sqrt{\langle \vec{r}_i - \vec{r}_{i,ave} \rangle^2 \langle \vec{r}_j - \vec{r}_{j,ave} \rangle^2}} \quad (2.5)$$

where \vec{r}_i and \vec{r}_j are the position vectors of two atoms i and j respectively and $C_{ij} = C_{ji}$. The correlation matrix describes the linear correlation between any pairs of residues or set of atoms as they move around their average position during dynamics; where a positive correlation (correlated residues) reflects a concerted motion along the same direction and a negative correlation (anti-correlated residues) indicates a motion in the opposite direction [145]. Cross-correlation coefficients range from a value of -1 (completely anti-correlated motions) to a value of +1 (completely correlated motions) [146]. The diagonal elements of the correlation matrix will be 1 since they are the correlation of a residue with itself (self-correlation); however, because *ptraj* was used to calculate the residual correlation as the average residual (rather than atomic) correlation coefficient for each residue, the values of self-correlated residues were less than 1. A full discussion of the input files that were used to analyse MD simulation trajectories is in appendix III.

2.2.1.9 Calculating contact maps

A contact map of a protein is a two-dimensional representation of its three dimensional structure. It is a binary symmetrical matrix that can be defined for any protein P by the following expression:

$$M_{ij}^P = \begin{cases} 1 & \text{if the distance between residues } i, j \text{ is } \leq \tau \text{ \AA} \\ 0 & \text{otherwise} \end{cases}$$

(2.6)

where i and j are two residues, and τ is the distance cut-off between the two residues. Usually the distance between the two residues is measured between the coordinates of their $C\alpha$. The cut-off value for the distance between the two residues in a contact map ranges from 6 to 16 Å [148]. In this study, contact maps were generated using the Contact Map View (CMview) plugin [126] for PyMOL [125]. The cut-off distance for the contacts between the α -carbons of any residue pair was set to 8 Å.

2.2.1.10 Preparation of the individual simulated systems

AMBER always renumbers the amino acid residues in simulated models to start from number one, therefore, to remove any subsequent confusion, residues in the original pdb files were initially renumbered in this way.

2.2.1.10.1 *c-Jun N-terminal protein kinase (JNK-1)*

a. Model preparation

Two structural models of the catalytic domain of JNK1-alpha1 isoform were prepared using Discovery Studio (DS). The first is the apo form (*JNK1-apo*); and the second is complexed with an allosteric inhibitor, a biaryl tetrazole, (*JNK1-allo*). The initial coordinates for both structures were retrieved directly from their crystal structures in the Protein Data Bank. The corresponding PDB code for *JNK1-apo* is 3O17 with a crystal resolution of 3 Å; and for *JNK1-allo* is 3O2M with a crystal resolution of 2.7 Å [121]. The crystal structure of 3O17 is composed of a homodimer of the catalytic domain of JNK-1 in complex with a peptide segment of JNK-interacting protein-1 (pepJIP1). JIP1 is a scaffolding protein responsible for the assembly of the components of JNK cascade [149]. The crystal structure of 3O2M is a homodimer of pepJIP1-JNK1 complexed with the allosteric inhibitor. The structural models for both states were prepared by deleting one of the monomers, the peptide segment (pepJIP1), and all of the SO_4^{-2} ions (figure 2.3).

b. Generating the topology and coordinates (prmtop and inpcrd) files

Having prepared the two initial pdb structures, it was then necessary to create sander input files: the *prmtop*; and the *inpcrd* files. As discussed, to run a MD simulation using AMBER all of the residues within the studied system must be pre-defined in the AMBER database. In the case of *JNK1-apo* all of the residues in the system are standard amino acids and by default are pre-defined. Sander's MD input files can therefore be generated without any further processing of the prepared pdb file. In the case of *JNK1-allo* the allosteric inhibitor (figure 2.4) is not a standard residue. Structural information and force field parameters for this non-standard residue needed to be calculated before creating the input files. These parameters were generated using the *Antechamber* program, which utilises the general AMBER force field (GAFF).

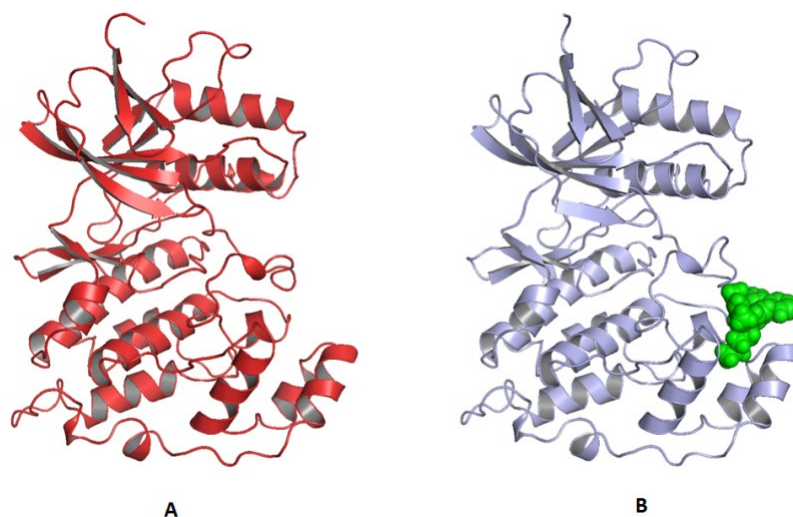


Figure 2.3: The starting crystal structures for the simulated JNK1 states. (A) *JNK1-apo* state (PDB code 3O17). (B) The *JNK1-allo* state (PDB code 3O2M) with allosteric inhibitor shown in green spheres.

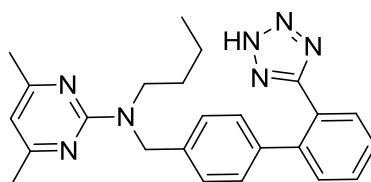


Figure 2.4: The chemical structure of JNK1 allosteric inhibitor.

Next, the two sander input files were created using the *Xleap* module within AMBER and the FF99SB force field as previously described. When solvating the prepared models the distance between the edge of the box and the protein was set to 12 Å for (*JNK1-apo*) and 10 Å for (*JNK1-allo*). After solvation and neutralisation the (*JNK1-apo*) system consisted of 357 amino acids, four Na⁺ ions and more than 14,000 water molecules; the (*JNK1-allo*) system consisted of 358 amino acids, the allosteric inhibitor, five Na⁺ ions and more than 12,000 water molecules. Finally, the input files were created and saved and the MD simulation protocol applied. The simulation time for *JNK-apo* and *JNK-allo* systems was 51 and 50 ns respectively. The first 7 ns of *JNK-apo* simulation and the first 6 ns of *JNK-allo* simulation utilised the Langevin dynamics scheme and were considered as an extended equilibration phase, and the other 44 ns of production dynamics were continued at constant temperature and pressure (NPT) utilising the Berendsen dynamics scheme.

2.2.1.10.2 Cyclin-dependent kinase 2 (CDK2)

a. Preparation of the simulated models

Two structural models of the catalytic domain of CDK2 were prepared using DS: the nucleotide-free state (*CDK2-apo*); and the unphosphorylated ATP-bound state (*CDK2-ATP*). The initial coordinates for the *CDK2-apo* and the *CDK2-ATP* states were retrieved directly from their crystal structures in the Protein Data Bank. The corresponding PDB codes are 1HCL for the *CDK2-apo* state (crystal resolution is 1.8 Å) [150] and 1HCK for the *CDK2-ATP* state (crystal resolution is 1.9 Å) [150]. Surprisingly, the two crystal structures were almost identical with an RMSD of their backbone atoms of 0.36 Å. The structural models of *CDK2-apo* and *CDK2-ATP* (figure 2.5) were prepared as follows: the *Protein Report* functionality in DS revealed that both systems had a missing loop, Leu37-Glu40; the missing LDTE

segment was inserted using the procedure described above with another CDK2 crystal structure (PDB code 3PY1, resolution 2.05 Å) [108] used as a reference. Based on the total energy score of the generated models, the top ranked models with PDF total energy of -10204.16 kcal/mol for 1HCL (model number 3) and -10145.23 Kcal/mol for 1HCK (model number 5) were selected. Both crystal structures (1HCL and 1HCK) have two conformations for Gln131; for the apo state (1HCL); conformation A was selected as it forms a hydrogen bond with the backbone of the protein, and for the ATP-bound state (1HCK) conformation B was selected, which forms a hydrogen bond with ATP.

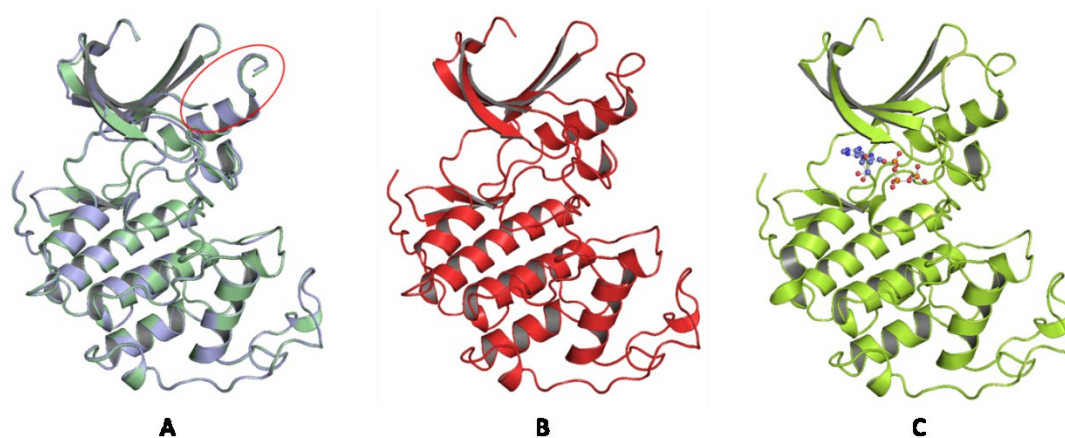


Figure 2.5: The starting crystal structures for the simulated CDK2 states. (A) Superimposition of the two crystal structures showing the high degree of conformational similarity between the two structures and highlight the location of the missing loops (red ellipsoid). 1HCL is pale green and 1HCK is light blue. (B and C) The refined structural models of *CDK2-apo* (PDB code 1HCL) and *CDK2-ATP* state (PDB code 1HCK) respectively.

b. Generating the topology and coordinates (prmtop and inpcrd) files

Having prepared the two initial pdb structures, Sander's input files were created. The ATP molecule in the *CDK2-ATP* model is a non-standard residue that needs to be parameterized. Therefore, the polyphosphate parameters developed by Meagher et al. [151] for ATP were adopted. When solvating the prepared models the distance between protein surface and the edge of the box was set to 12 Å for both systems.

After solvation and neutralisation the (*CDK2-apo*) system consisted of 298 amino acids and four Cl⁻ ions; the (*CDK2-ATP*) system consisted of 298 amino acids, two Cl⁻ ions, two Mg²⁺ ions and one ATP molecule. Each system contained more than 12000 water molecules. Finally, the *inpcrd* and *prmtop* files were created and saved; and the MD simulation protocol was applied. The simulation time for the *CDK2-apo* system was 50 ns, while that for *CDK2-ATP* was 60 ns. In both systems the first 6 ns of simulation was utilising the Langevin dynamics scheme and considered as an extended equilibration phase; afterwards the production dynamics commenced at constant temperature and pressure utilising the Berendsen dynamics scheme.

2.2.1.10.3 Dual specificity tyrosine-phosphorylation-regulated kinase 2 (*DYRK2*)

a. Preparation of the simulated models

Three structural models of the catalytic domain of *DYRK2* were prepared using DS which are the nucleotide-free state (*DYRK2-apo*); and two modelled complex states (*DYRK2-probe-1*) and (*DYRK2-com-6*). The initial coordinates for the (*DYRK2-apo*) state were retrieved directly from its crystal structure from the Protein Data Bank. The corresponding PDB code is 3K2L and crystal resolution is 2.36 Å. In this crystal structure there was three missing residues Gln70, Ser71 and Gly123; two residues with alternate conformations which are Arg253 and Arg325; five phosphorylated residues Ser159, Thr308, Tyr309, Ser369 and Ser385; and 5 ions (2 SO₄⁻², 2 Cl⁻ and 1 Na⁺).

The structural model of *DYRK2-apo* was prepared by deleting all ions; completing all missing side chains; keeping one set of the alternate conformations (conformation A); mutating all of the five phosphorylated residues back to their native state using the *Build and Edit Protein* tool in DS; and inserting and refining the missing residues following the procedure described above for inserting missing loops. Based on the total energy score of the generated models we selected and saved the top ranked model and the other models were deleted (figure 2.6).

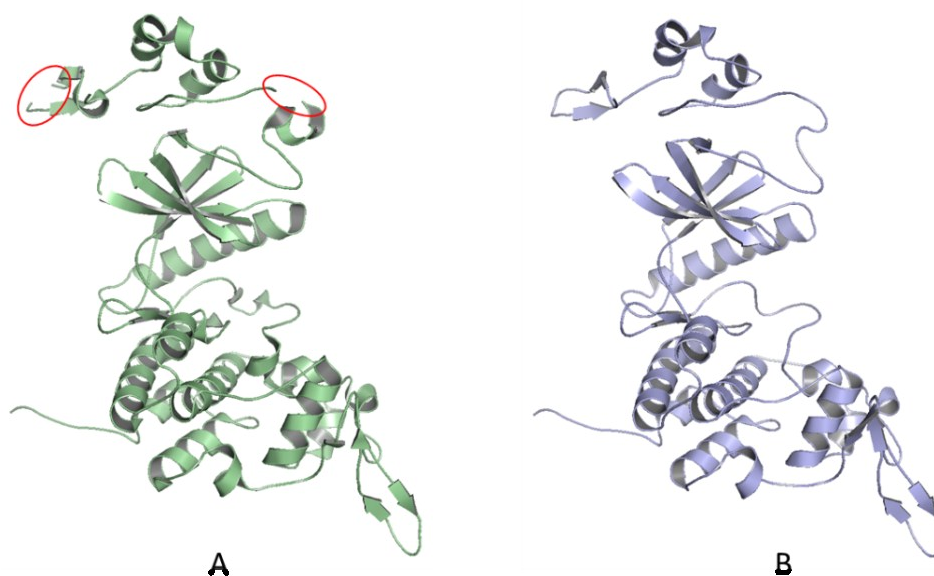


Figure 2.6: The starting crystal structure of *DYRK2-*apo**. (A) The missing residues are highlighted by red ellipsoids. (B) The starting structural model after inserting and refining the missing segments.

Initial coordinates for the *DYRK2-probe-1* and *DYRK2-com-6* models were obtained by docking a small molecule probe (figure 2.7), which were sketched and minimised using DS, into the identified putative allosteric binding site in the N-domain of the minimised average structure of the equilibrated part of the *DYRK2-*apo** trajectory.

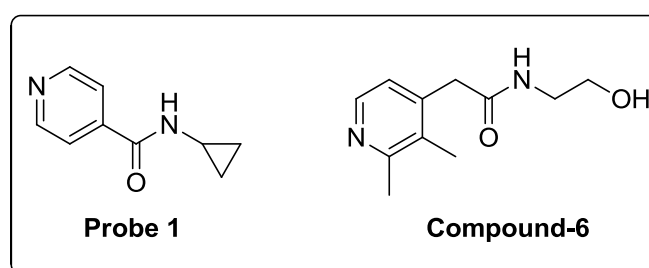


Figure 2.7: The chemical structures of the two probes used to generate the starting structural models for the *DYRK2-probe-1* and *DYRK2-com-6* respectively.

In order to run MD simulations of the *DYRK2-complex* states, the probe molecules were docked into the identified binding site to serve as starting structural models. To this end, the first probe molecule was manually docked into the identified binding

site of the minimised average structure of *DYRK2-apo*, so that DS can define the binding site based on the location of the docked ligand for subsequent automated docking. Having defined the binding site, the *Generate Conformations* protocol in DS was used to generate conformations for the probe molecules. The *Best* method for conformations generation was selected and all other parameters were left at their default values. The generated conformation were docked into the binding site using CDOCKER and the best docked pose for each of the two probes was selected and saved as the starting structural model for the corresponding *DYRK2-complex* state (figure 2.8). CDOCKER (CHARMm-based DOCKER) is a grid-based MD docking algorithm. It treats the protein as a rigid molecule but accounts for full ligand flexibility including bonds, angles, and dihedrals and to refine the docked poses it performs a final minimisation step [152].

b. Generating the Topology and Coordinates (prmtop and inpcrd) Files

Having generated the initial pdb structures, sander was then used to create the *prmtop* and *inpcrd* files. All residues in the *DYRK2-apo* state are standard amino acids; but for the *DYRK2-complex* models the small molecules (probes) were non-standard residues, so *Antechamber* was used to generate the necessary force field parameters. The distance between the protein and the edge of the solvation box was 12 Å for *DYRK2-apo* and 10 Å for the *DYRK2-complex* systems. After solvation and neutralisation, the *DYRK2-apo* system consisted of 410 amino acids, 14 Cl⁻ ions, and more than 28,000 water molecules; the (*DYRK2-complex*) systems consisted of 410 amino acids, 14 Cl⁻ ions, the corresponding probe molecule, and more than 29,000 water molecules. The *inpcrd* and *prmtop* files were created and the MD simulation protocol was applied. The simulation time for *DYRK2-apo* was 50 ns. The first 10 ns of simulation utilised the Langevin dynamics scheme and was considered an extended equilibration phase; the other 40 ns of production dynamics utilised the Berendsen dynamics scheme. The simulation time for the *DYRK2-complex* systems was 15 ns. The first 5 ns utilised the Langevin dynamics scheme and the remaining 10 ns of production dynamics utilised the Berendsen dynamics scheme. A comprehensive summary of all simulated systems is given in table 2.1.

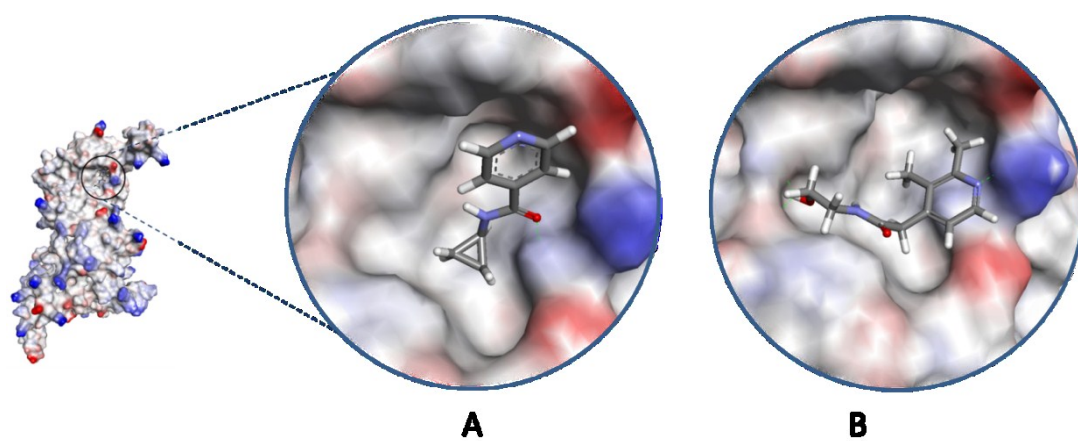


Figure 2.8: The starting structural models for the DYRK2-complex states. Complexes (A and B) correspond to *DYRK2-probe-1* and *DYRK2-com-6* models respectively.

Table 2.1: Summary of all simulated systems.

Kinase name	Simulated state	PDB code	Preparation of the structural models	Model components	Simulation time and type of dynamics*	
					Langevin	Berendsen
JNK1	<i>JNK1-apo</i>	3O17	Deletion of one of the monomers, the pepJIP1, and SO ₄ ⁻² ions.	Consisted of 357 amino acids, 4 Na ⁺ ions and more than 14000 water molecules.	7 ns	44 ns
	<i>JNK1-allo</i>	3O2M	Same as the apo state, in addition the allosteric inhibitor was parameterized.	Consisted of 358 amino acids, the allosteric inhibitor, 5 Na ⁺ ions and more than 12000 water molecules.	6 ns	44 ns
CDK2	<i>CDK2-apo</i>	1HCL	Gln131 has 2 conformations, conformation A was selected. The missing loop was inserted and refined.	Consisted of 298 amino acids, 4 Cl ⁻ ions, and more than 12000 water molecules.	6 ns	44 ns
	<i>CDK2-ATP</i>	1HCK	Conformation B of Gln131 was selected. The missing loop was inserted and refined, and ATP parameters by Meagher et al. were utilised.	Consisted of 298 amino acids, 2 Cl ⁻ ions, 2 Mg ⁺² ions, ATP molecule, and more than 12000 water molecules.	6 ns	54 ns
DYRK2	<i>DYRK2-apo</i>	3K2L	All ions were deleted; conformation A for residues with alternate conformations was selected; the five phosphorylated residues were mutated back to their native state; and the missing loops were inserted and refined.	Consisted of 410 amino acids, 14 Cl ⁻ ions; and more than 28000 water molecules.	10 ns	40 ns
	<i>DYRK2-probe-1</i> <i>DYRK2-com-6</i>		Each of the small molecule probes was docked into the identified putative allosteric site in the minimised average structure of the of the <i>DYRK2-apo</i> state.	Consisted of 410 amino acids, the probe, 14 Cl ⁻ ions, and more than 29000 water molecules.	5 ns	10 ns

* For all systems, energy minimisation was carried out using steepest descents for the first 250 steps, then conjugate gradient. They were then heated (0-300 K) over 120 ps under NVT conditions, then equilibrated at 300 K for 60 ps under NPT conditions. In the heating and equilibration stages, Langevin dynamics were applied. Langevin dynamics is suitable for equilibrating the system, and Berendsen dynamics is better suited for studying structural correlation.

2.2.2 Simple Intracequence Differences (SID) analysis

2.2.2.1 Individual and comparative SID analysis of simulated systems

SID analysis can be applied to single protein folds to provide insight on the susceptibility of topological and sub-structural elements to perturbations. Additionally, comparative SID analysis can be applied to related protein folds (related variants, different states of the same fold, or frames obtained from MD simulations) to draw attention to regions where ligand binding and activation events have had the most prominent topological effect [59]. Since MD simulations are being run for different proteins (some simulations are for different states of the same protein) SID analysis can be utilised to study the minimised average structures of the simulated systems and also comparatively with different frames (structure snapshots) from the same trajectory. This provides a means for identifying regions of high allosteric potential.

Individual SID analysis: an average structure was generated from the trajectory of each simulated system using the *ptraj* module within AMBER10. These were then processed with SID. SID analysis proceeds as follows:

- i) Sphere definition*, a 7 Å sphere is defined around C α -carbon of each amino acid residue in the structure.
- ii) Clustering*, starting from the first residue in the N-terminus and proceeding to the C-terminus sequentially, other residues whose C α -carbons are located within the sphere of the considered residue are clustered.
- iii) Scoring clusters*, clusters are assigned a score based on the maximum chain separation of residues within the cluster. Scoring methods include: a) simple difference (highest-lowest, *HL*), which is the residue with highest sequence number minus the one with lowest sequence number; b) simple difference (greatest gap *GG*), where all positions within a cluster are numerically ordered and the difference between consecutive residue numbers is calculated, the GG value is assigned the largest difference; c) differential score (*DIFF*), $DIFF = HL - GG$; and d) the count score (*count*), which is the total number of residues encompassed within the sphere.

All SID scores (HL, GG, DIFF, and Count) were calculated for each residue in the structure.

iv) Selection of important SID scores, the outliers in the SID data were of particular importance because they identify residues with allosteric potential. Those residues were identified by determining the median, the upper (UQ) and lower (LQ) quartiles of the SID score distribution (HL, GG or differential). The inter-quartile distance (IQD) is then defined as (UQ – LQ). Clusters that have HL and DIFF scores greater than the value of their median quartile, and a GG scores lying within their IQD value were clustered as outliers.

Comparative SID analysis: in order to analyse the trajectories using SID, all of the saved frames during the simulation were extracted and saved in PDB format using the *ptraj* module within AMBER10. Saved frames were then scored using all SID scores. In order to identify regions of potential motion or adjustment in the protein, the standard deviation (STD) and the average (Avg) for each of SID scores for every residue in the protein throughout the analysed frames were calculated, afterwards the relative standard deviation (RSD) was obtained by dividing the standard deviation by the average. The RSD is the absolute value of the coefficient of variation and can be used to differentiate between regions in the protein that are vulnerable to conformational change and those that are not. Small or close-to-zero values of RSD indicate little change in SID score, implying that the region around the position in question is not vulnerable to conformational change, whereas large values can imply some degree of conformational rearrangement. For a better visualisation of the SID analysis, the SID scores for all systems were converted into a colour code and structurally mapped onto the backbone of the corresponding protein. This conversion of SID scores into colours was achieved using a python script that runs within PyMol (see appendix IV).

2.2.3 Correlations of energy fluctuations

This method is based on the adaptation of two related approaches. The first was developed by Bahar and co-workers in which they relate signal propagation in

proteins with their residual equilibrium dynamical fluctuations (fluctuations in inter-residue distance). They showed that elements of secondary structures are more efficient in processing signals compared to coiled residues and that catalytic residues have enhanced communication propensities [153]. The other approach is that of Burak Erman where he related, based on a canonical model of protein systems, the fluctuations in the energy of the surroundings of the protein to the fluctuations in the residue positions within the protein [154]. Both approaches studied the signal propagation in proteins and fluctuations in energy and how they relate to fluctuations in residue positions within the 3D structure of proteins, thereby linking the energy fluctuations and signal propagation pathways in proteins with their 3D architecture.

This approach was applied in our study to identify hot spots on protein surface that are more responsive to energy fluctuations in the surroundings, “energy gates”; and to identify the pathways between residues through which energy diffuses, “interaction pathways” [154, 155]; which have an obvious connection to the concept of allosteric regulation of proteins.

In order to identify these energy gates and interaction pathways (the correlations of energy fluctuations) MATLAB [127] was employed. Firstly, the affinity matrix which represents the interaction strength between any two residues i and j in the protein was calculated and defined as follows:

$$a_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}} \quad (2.7)$$

where N_{ij} is the total number of contacts between all atoms in residue i and residue j based on a cutoff distance of 7 Å, and N_i , N_j are the total number of heavy atoms in each residue. Based on the affinity matrix the local interaction density d_j for each residue j was measured as:

$$d_j = \sum_{i=1}^n a_{ij} \quad (2.8)$$

Secondly, the stiffness matrix or the *Kirchhoff matrix* (Γ) was calculated, which is defined in terms of the affinity $A = \{a_{ij}\}$ and the degree matrices $D = \text{diag}\{d_j\}$ such that:

$$\Gamma = D - A \quad (2.9)$$

Then the correlations of the energy fluctuations were calculated using equation 2.10:

$$C_{T,i} = k \sum_j (\Gamma_{ii}^{-1} - 2\Gamma_{ij}^{-1} + \Gamma_{jj}^{-1}) \quad (2.10)$$

Finally, the energetic interaction, ΔU_i , of residue i with all other residues in the protein was calculated using equation 2.11:

$$\Delta U_i = \sum_k \langle \Delta \hat{U}_i \Delta \hat{U}_k \rangle \quad (2.11)$$

Following the calculation of the energy correlation matrix of all residues within the minimised average structure for each simulated system, the average value of the energetic interactions for each residue were calculated and plotted against residue index then mapped onto the backbone of the corresponding structure. Appendix v shows the perl scripts and the Matlab input files that were used to calculate the energy correlations.

2.2.4 The methods of virtual screening and docking

Having identified a putative allosteric binding site in DYRK2, virtual screening (VS) of databases (DB) was started and followed by docking of retrieved hits into the identified site in order to search for compounds that show good binding affinities. A fast way to search 3D databases is to generate a 3D pharmacophore.

2.2.4.1 Structure-based pharmacophore generation

The identified putative allosteric binding site in DYRK2 was used to generate a 3D structure-based pharmacophore (SBP) model for the virtual screening of small molecules databases. A SBP utilises known or suspected protein active site to select compounds most likely to bind within that site. In this study the minimised average structure of the *DYRK2-com-6* simulated trajectory was used to generate the pharmacophore by using the *Interaction Generation Protocol* available in DS following the sequence shown in figure 2.9.

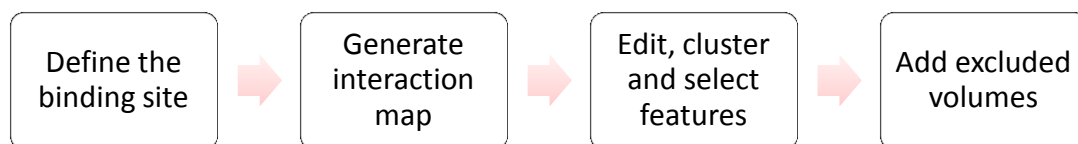


Figure 2.9: A schematic description of the workflow of structure-based pharmacophore generation.

This protocol applies the Ludi algorithm to generate the interaction map by numerating interaction points from a protein binding site that are important for ligand binding which are then converted to pharmacophoric features [156, 157]. In order to run the protocol the binding site needs to be defined with a sphere that covers all important amino acid residues. The sphere was created using the *Define and Edit Binding Site* tool within DS by defining the enzyme as a receptor then defining the binding sites from receptor cavities that are identified using a flood filling algorithm. The sphere is then created around the cavity that represents the site of interest. The sphere encompassed all residues in the binding site that might be of relevance to ligand binding. Having defined the binding site, the protocol was employed to identify all hydrophobic and hydrophilic interaction points within the sphere that can be complemented by a ligand. The identified hydrogen bond acceptors, hydrogen bond donors, and hydrophobic features were then clustered and edited using the *Edit and Cluster Pharmacophore Features* tool in DS and the most important features were selected and included in the construction of the 3D pharmacophore model (figure 2.10a). The generated 3D pharmacophore consisted of seven features (three hydrogen-bond donors, three hydrogen-bond acceptor and one hydrophobic region). Each of these features has a position constraint which consists of the ideal location of that feature in 3D-space surrounded by a 1.6Å radius tolerance sphere. In addition to the tolerance sphere, hydrogen-bond donor and acceptor features also have a 3Å long vector to indicate the direction of interaction and at the end of the vector there is another 2.2 Å tolerance sphere to accommodate the other (opposite) interacting group. Functional groups of candidate ligands must map all the pharmacophoric features and reside within the tolerance sphere to be retrieved as hits.

In order to account for the steric interactions with the target, which give rise to false positive hits in virtual screening (ligands that map the pharmacophore but do not show good docking scores because of steric clashes with target), excluded volumes were added to the generated pharmacophore to remove these false positives. All carbon atoms in residues within 6Å of the sphere defining the binding site were selected and used to place exclusion constraints around each of them. The exclusion volumes were then clustered in the same way as the pharmacophoric features (figure 2.10b), whilst further irrelevant excluded volumes were manually removed.

The generated pharmacophore comprises all features that are important for ligand binding, but having seven features in this pharmacophore proved over restrictive when it was used to screen Maybridge database with only two hits were returned. Therefore, it was fragmented into three smaller pharmacophores that complemented each other and allowed for thorough screening of databases. The excluded volumes were kept intact in all of the three small pharmacophores.

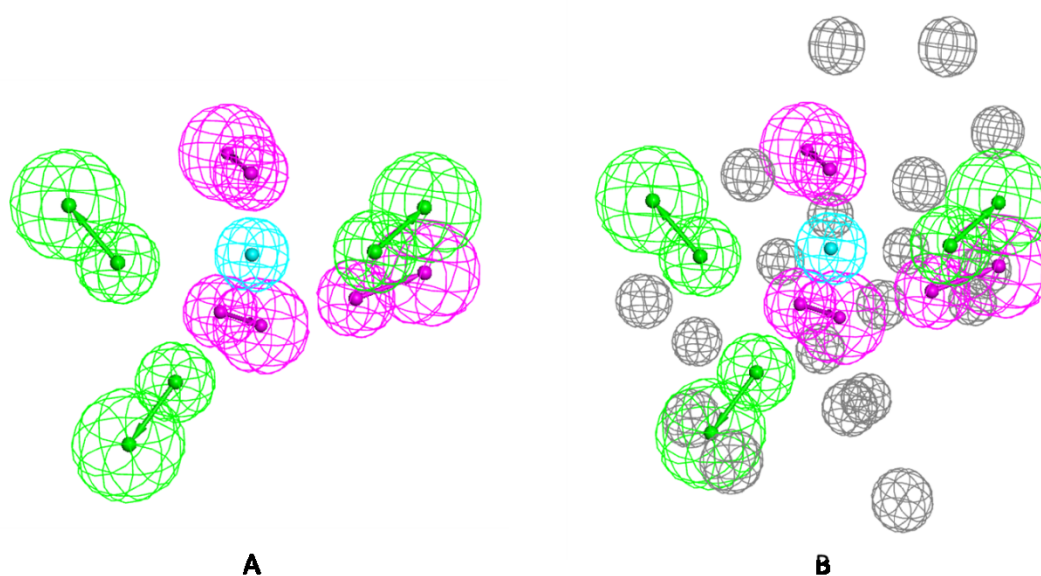


Figure 2.10: The primary structure-based pharmacophore model generated based on the putative allosteric site in DYRK2 before adding the exclusion volumes (A); and after the addition of the excluded volumes (B).

2.2.4.2 Virtual screening of commercial databases

The three generated pharmacophores were then used for virtual screening of databases of small molecules. DS offers two methods for database searching; the *Fast* rigid and the *Best* flexible methods. In general better results are obtained using the best flexible method. In our study the *Best flexible search* method in the *Search 3D Database* protocol in DS was used to screen the Maybridge 2009 Database [158]. 56,969 compounds were screened and all ligands that mapped any of the three pharmacophores were returned. Pipeline Pilot (PP) was then used to filter the retrieved hits based on Lipinski's rule of five of drug like properties and consideration of fit values. Hits that passed filtration criteria were selected for molecular docking.

2.2.4.3 Molecular Docking

Molecular docking of the filtered hits was performed using GOLD (Genetic Optimization for Ligand Docking) docking program version 5.1. GOLD uses a genetic algorithm (GA) to perform fully flexible ligand docking with partial flexibility of the binding site [94, 95]. GOLD has proved to be one of the most accurate ligand docking programs. Different studies have shown that GOLD was more than 70% (in some cases more than 80%) accurate in reproducing the crystal complex binding mode of the ligand [159, 160]. The minimised average structure of the *DYRK2-com-6* was used to define the binding site. The binding site in GOLD was defined by selecting the ligand and all atoms within 6Å radius of the ligand. The ligand was extracted from the complex and a cavity detection algorithm, LIGSITE, was used to restrict the region of interest to concave, solvent-accessible surfaces. LIGSITE is an automatic program that searches for and detect pockets on protein surfaces that may serve as a binding site for small ligands [47]. Only the top three scored docks out of ten per compound were stored, and early termination was activated. GOLD uses a fitness score to separate and rank all generated conformations of the docked compounds. The Gold fitness score accounts for protein-ligand H-bonding and van der Waal interactions; and ligand's internal van der Waals and torsional strain energy [94]. GoldScore was used as the fitness function in docking the Maybridge retrieved filtered hits. The ligands that showed

good binding modes and good molecular interactions with the binding site were selected for experimental evaluation of their binding affinities.

2.3 Experimental methods

2.3.1 Differential scanning fluorimetry (DSF)

The DSF screening was performed in prof. Stefan Knapp's laboratory at the Structural Genomics Consortium (SGC) at the Oxford University. Screened compounds were dissolved to a concentration of 50 mM in DMSO, afterwards the SGC protocol [161] was employed.

2.3.2 Inhibition and kinetics assays of DYRK2

The inhibition and kinetics assays were performed by Louise Young in SIPBS at the University of Strathclyde. DYRK2 Kinase activity was determined using an enzyme-linked immunosorbent assay (ELISA) kit - CycLex DYRK2 Kinase Assay (Caltag Medsystems, Buckingham, UK), and the recommended protocol was adhered to [162]. The materials below were made up as follows:

- *Lyophilized ATP* was reconstituted in 800µl distilled deionised water to give a final concentration of 2.5mM.
- *Wash buffer* was diluted 10x to give a working solution.
- *DYRK2 enzyme* was diluted 1:20 in Assay buffer.

All other reagents listed below were ready to use:

- *Microplate* coated with recombinant p53 N-terminus (1-99 aa) as a substrate for DYRK2.
- *Kinase buffer*
- *Horse radish peroxidase conjugated detection antibody*
- *Tetra-methylbenzidine (TMB)*, as a substrate reagent for the HRP.

Basically, human recombinant DYRK2, 20m units per well and ATP were incubated in a kinase reaction buffer on the plate coated with a recombinant p53 substrate containing the serine 46 residue which is phosphorylated by DYRK2 in the presence and absence of test compounds.

For inhibition assays the compounds or standard were added in the concentration range of 10nM to 30 μ M using an ATP concentration of 125 μ M as recommended by the manufacturer of the kit. For kinetic assays doubling dilutions of ATP were used starting at 125 μ M and ending at 0.5 μ M; compound concentrations were 0.1, 1, 2.5 and 5 μ M. The assay plate was incubated for 30 minutes at 30°C.

After five washes with wash buffer containing 2% Tween-20, the antibody against the phosphorylated substrate (HRP conjugated anti-phospho-p53 S46 (TK-4D4) antibody) was added and incubated at room temperature (RTP) for 30 minutes. After a further five washes, TMB was added as a chromogenic substrate for HRP and incubated for 15 minutes at RTP. Then optical density was measured at 540nm.

For the inhibition assays, an apparent K_i of the phosphorylated substrate was calculated for each compound using the Cheng-Prusoff equation by nonlinear curve fitting program Graphpad Prism 4 (Sigma software, Ashburton, UK).

For the kinetic assays Lineweaver -Burk analysis was employed to determine the V_{max} and K_m of the compounds (at four different concentrations) against the ATP substrate.

3 PROOF OF CONCEPT (JNK1 AND CDK2)

For the proof-of-concept purpose two protein kinases with known experimentally identified allosteric binding sites were selected and analysed in order to validate and examine the predictive potential of the computational method we are proposing in this study. These were c-Jun N-terminal protein kinase (JNK-1) and cyclin-dependent kinase 2 (CDK2). Both were selected because the allosteric site was very well defined and distinct from the ATP binding site, and binding of the allosteric inhibitor (the one co-crystallised with each kinase) has affected the function of the enzyme, which confirms that these sites are functionally important. Moreover, in both cases there was a solved crystal structure for the *apo* enzyme which enabled running MD simulations of the *apo* and complex forms of the same enzyme to compare their dynamics and energetics without the need to use a modelled *apo* structure that may not reflect the native state of the enzyme. The dual specificity tyrosine-phosphorylation-regulated kinase 2 (DYRK2), which has no known allosteric binding sites was used as a real case study. This kinase was selected based on a recommendation from prof. Knapp's research group at the SGC-Oxford University who solved all crystal structures of this kinase. Solving the crystal structure of a DYRK2 complex with any potential ligand from this study would therefore be feasible.

3.1 c-Jun N-terminal Kinase (JNK-1)

3.1.1 The role of JNKs

c-Jun N-terminal Kinases (JNKs) are serine/threonine protein kinases that are members of the mitogen-activated protein (MAP) kinases family (figure 3.1). There are ten isoforms of JNKs proteins derived from three encoding genes *jnk1*, *jnk2* and *jnk3* [149, 163]. JNKs are activated in response to cytokines or environmental stress [164]. Complete activation of the JNKs is achieved via dual phosphorylation of threonine and tyrosine in the threonine-proline-tyrosine motif, and this phosphorylation is carried out by mitogen-activated protein kinase kinase (MKK) four and/or seven. They can be deactivated by serine and tyrosine phosphatases [164].

Biochemical and genetic studies have shown that JNKs' signaling pathway regulates many physiological processes, such as cellular proliferation, apoptosis, and tissue morphogenesis [164]; they are also involved in many pathological conditions including atherosclerosis, diabetes, stroke, Parkinson's and Alzheimer's diseases. These numerous and seemingly contradictory cellular responses of JNKs' signaling pathway make it difficult to regulate or control their activity. For example deletion of the encoding gene of JNK-1 on one hand protects against obesity-induced insulin resistance, but on the other it initiates Alzheimer's disease [121, 149, 165].

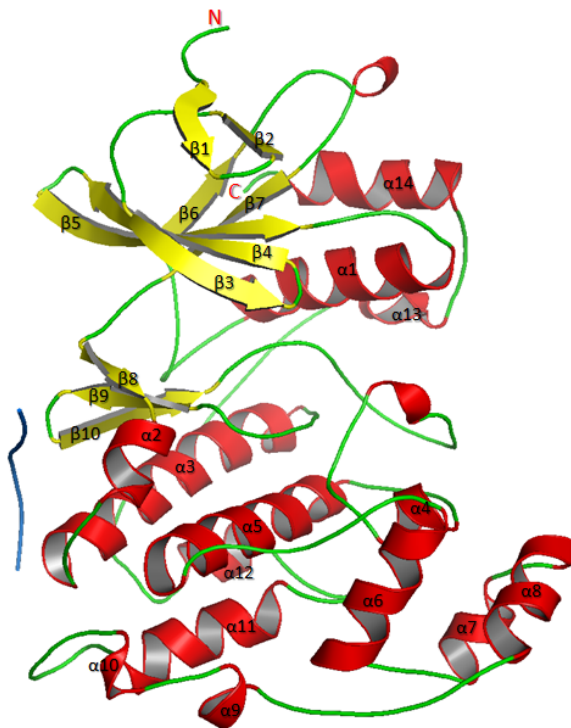


Figure 3.1: Cartoon representation of the backbone structure of JNK1 complexed with pepJIP1 (blue ribbon) pdb code 3O17. α -helices are in red, β -sheets are in yellow, and loops and turns are in green.

The signaling pathway of JNKs is regulated by a scaffolding protein, JNK-interacting protein-1 (JIP1) that assembles the components of the JNK cascade. Overexpression of JIP1 results in deactivation of JNKs' signaling pathway. The region of JIP1 that retains JNK-inhibition ability has been identified (pepJIP1),

which is a peptide segment derived from the docking site of JIP1. This peptide segment inhibits JNK through competing with the upstream kinases and other substrates for their binding grooves in JNK. Also, binding of pepJIP1 causes a distortion of the ATP binding site in JNK which significantly reduces ATP affinity which altogether explains the allosteric inhibitory effects of pepJIP1 [149] (figure 3.2).

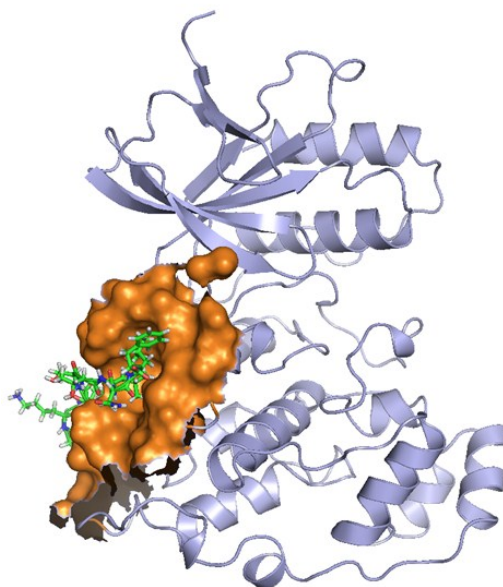


Figure 3.2: X-ray crystal structure of JNK-1 (light blue ribbons) complexed with pepJIP1 (coloured by atom type). The binding site of pepJIP is represented as a brown surface (PDB code 3O17).

Most of the deposited crystal structures of JNK in complex with small molecule inhibitors in the protein data bank show that these inhibitors are ATP competitive inhibitors, also there are some crystal structures in which JNK1 is complexed with peptide inhibitors. Recently an allosteric site distant from the ATP binding site has been identified in JNK1 and crystallised (3O2M). This allosteric inhibitor was identified via an affinity-based, high-throughput screening technique where exposed sites on the surface of a protein are examined against libraries of small molecules. Figure 3.3 shows the position of the allosteric site in JNK1 relative to the ATP binding site [121].

The main aim in this study is to computationally investigate the allosteric modulations in JNK1 to see whether the allosteric binding site is identified. Dynamical and structural changes in proteins are essential for allosteric transitions, and to explore these changes in JNK1, two molecular dynamics simulations were carried out; one on the apo state (*JNK1-apo*) and one on JNK1 complexed with an allosteric inhibitor (*JNK1-allo*). In many MD studies the *apo* form of the protein of interest is obtained by removing the ligand, usually ATP or ADP, which is then compared with the ligated state to extract structural and dynamical information [166, 167]. In this study, the apo structure was nucleotide free but complexed with pepJIP1, and to obtain the apo form of the enzyme the pepJIP1 was deleted. The results of these simulations will be presented as follows: firstly, the stability of the simulations will be inspected, followed by analysis of the conformational flexibility of the different parts in each system. Then residual dynamic correlation analysis, SID analysis and finally energy correlation analysis will be performed. In order to ensure that the sampling focuses on the equilibrated part of the trajectory, the first few nanoseconds of each trajectory that was considered as an extension of the equilibration phase were discarded.

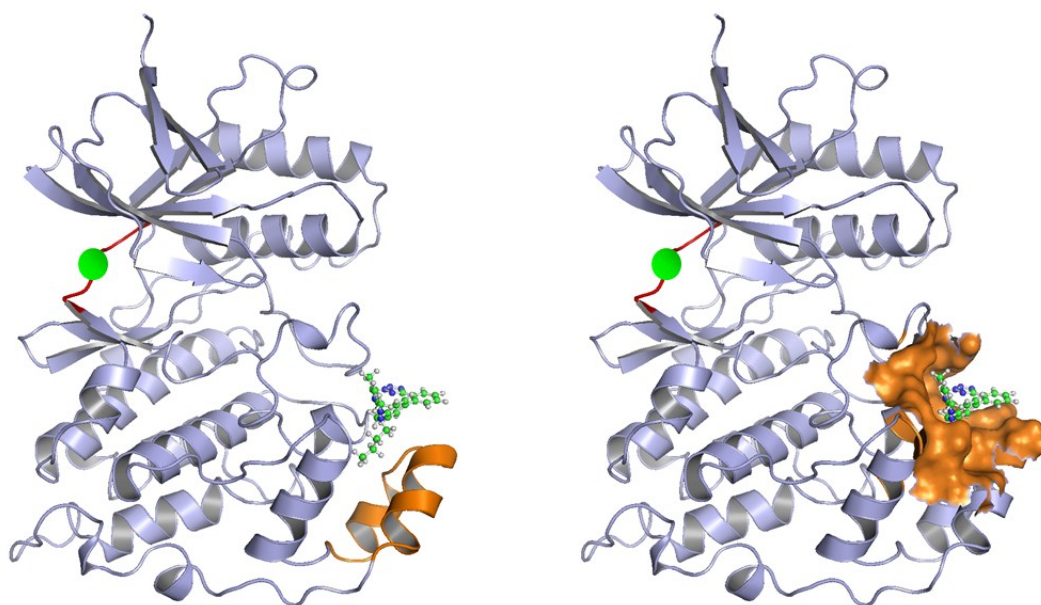


Figure 3.3: X-ray crystal structure of JNK-1 (light blue ribbons) complexed with an allosteric inhibitor (ball and stick). As shown in the left panel, the allosteric binding

site is distant from the ATP-binding site (hinge region; coloured red and denoted by the green sphere) and is located in the region between the MAPK insert (orange) and the body of the protein. The right panel shows a surface representation of the allosteric binding site (orange).

3.1.2 System stability and conformational flexibility

3.1.2.1 System stability

Root mean square deviation:

The stability of the simulated systems was explored by calculating the mass weighted root-mean-square deviation (RMSD) of their backbone atoms from the corresponding starting structures versus time (figure 3.4).

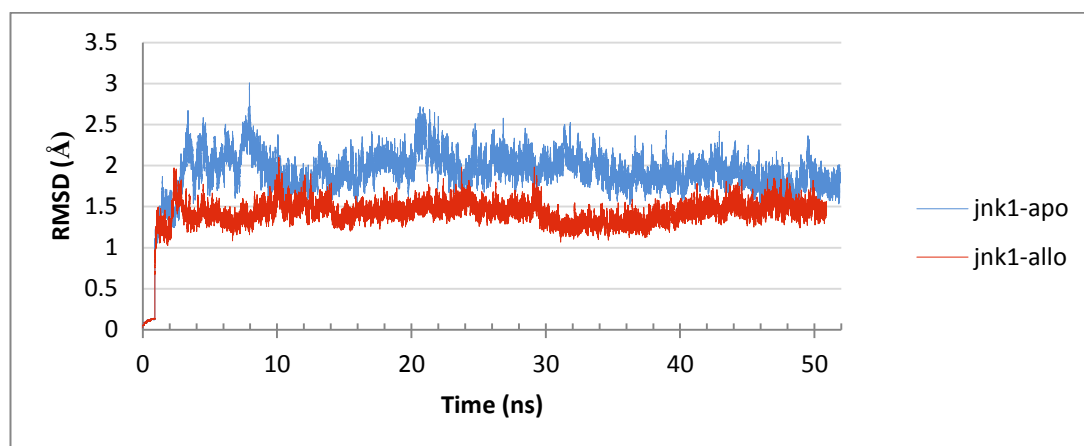


Figure 3.4: Backbone atoms' RMSD versus time plot for the MD simulations of *JNK1-apo* (blue) and *JNK1-allo* (red) using the starting structure of each trajectory as a reference.

Figure 3.4 shows that the RMSD of the backbone atoms of the two systems remained low for the first 180 ps (heating phase), which is due to the restraining force that was applied on the protein to prevent any wild fluctuations; once the restraints were removed the RMSD increased sharply as the protein relaxed within the solvent. The simulations then remained stable without any large oscillations. The two systems display moderate RMSD values of 1.94 Å and 1.43 Å for *JNK1-apo* and *JNK1-allo* respectively. The plots of both trajectories show that they have reached a stable and

equilibrated dynamical state relatively quickly, in around 1 ns. Afterwards, *JNK1-apo* (blue line) continued to drift from the initial structure for about 2-3 ns before a stable conformational ensemble was obtained; for *JNK1-allo* (red line) this drift lasted for about 2 ns. After these initial equilibrations, all conformational ensembles remain stable with average RMSD values of less than 2 Å with respect to their corresponding starting structures. The RMSD plots also show that the RMSD values for *JNK1-allo* were less than those of *JNK1-apo* throughout the simulation by nearly 0.5 Å, suggesting that *JNK1-allo* has not experienced the same degree of conformational change as *JNK1-apo*.

Effects of the pseudorandom seed generator:

It has been shown that running long MD simulations in many segments using either Andersen or Langevin dynamics without controlling the pseudorandom seed generator (RSG) can result in artefacts in the generated trajectory leading to inaccurate results. These artefacts arise from the residual forces that are applied in these algorithms to maintain the temperature at the desired value. However, these residual forces were shown to be fading out as the number of simulation steps in each MD segment increases. Generally speaking, MD simulations of 1,000,000 steps or more and a weak thermocoupling (collision frequency) of 3 ps⁻¹ or less are safe from artefacts [168]. Since the Langevin thermostat was used to control the temperature of the systems during the equilibration phase of the MD simulation without activating the RSG flag in the *md input files*, it was necessary to assess the health of the generated trajectories despite reasonably long segments (500,000 steps) and a collision frequency of 1 ps⁻¹ being used. Therefore, one of the trajectories was extended using the same *md input file* that was used during the equilibration phase (without the RSG being activated) for 20 ns and compared with a new trajectory with the RSG being activated. Figure 3.5 shows the RMSD versus time plots of the two trajectories and shows that there was no unusual behaviour in either trajectory confirming the quality of the simulations.

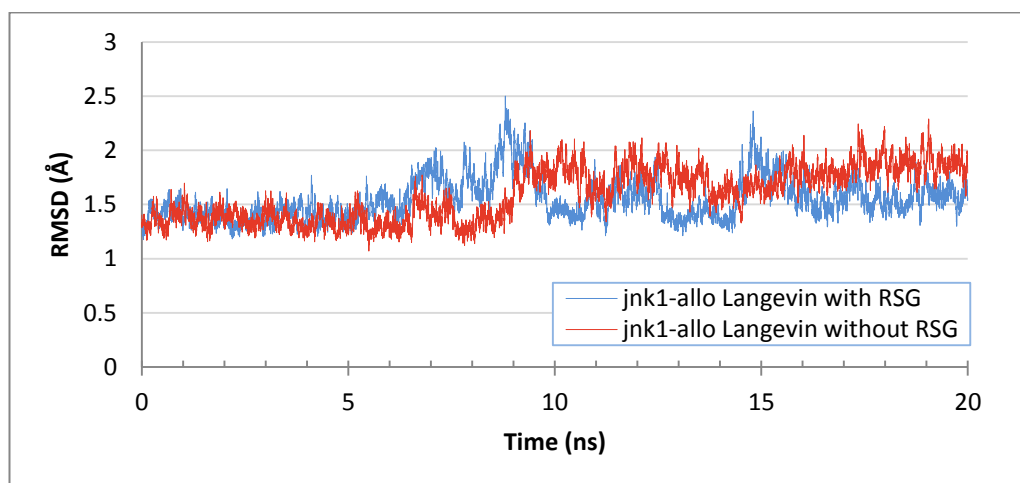
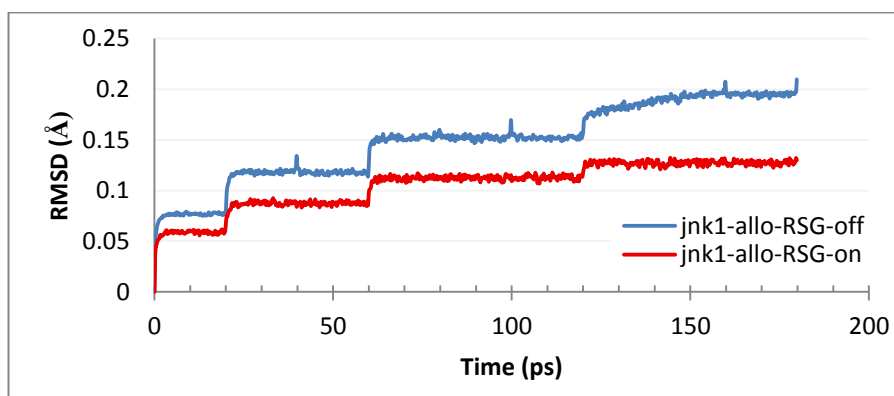


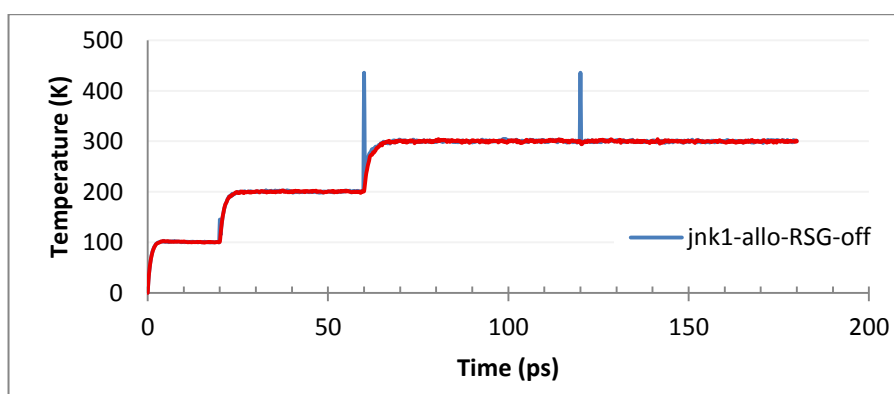
Figure 3.5: Comparison between two trajectories of the *JNK1-allo* state where the pseudorandom seed generator (RSG) was activated in one of them (blue line) but not in the other (red line).

The effect of these residual forces is more noticeable in short simulation segments such as in the heating phase (figure 3.6), yet the difference between the two RMSD values of the two trajectories was very small (typically less than 0.05\AA) and considered negligible.

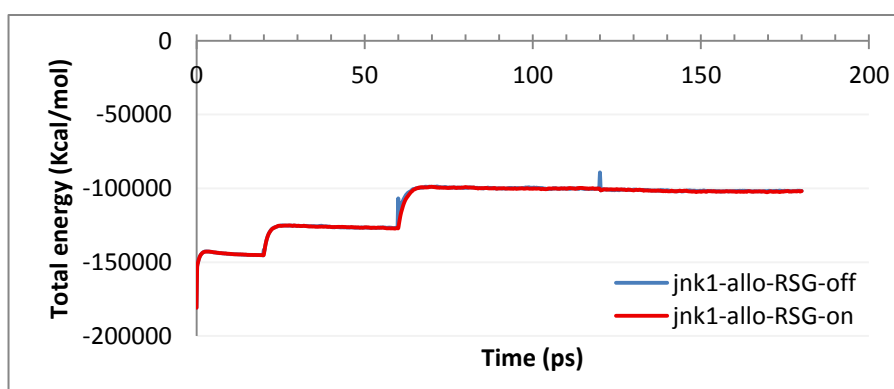
In figure 3.6, the spikes that exist in the temperature and total energy plots are attributed to not restarting each segment of the equilibration phase from the last saved frame of the previous one. In this case the simulated system does not have any record of the saved value of the temperature from the last step (the same applies for the kinetic energy since it is directly related to the velocity which in turn is related to the temperature of the system) and it will adapt to the new temperature value in the new input file (the input file specifies the values of all the parameters that control the behaviour of the system), and because Amber adopts the leap-frog algorithm (as the integrator of Newton's equations of motion) the velocities are written to the restart file at a different time point than the coordinates; hence there will be a very short time (less than a time step) between the last frame from the previous segment and the next one in the new segment at which the velocities (and temperature) will not be controlled which give rise to this spike.



A



B

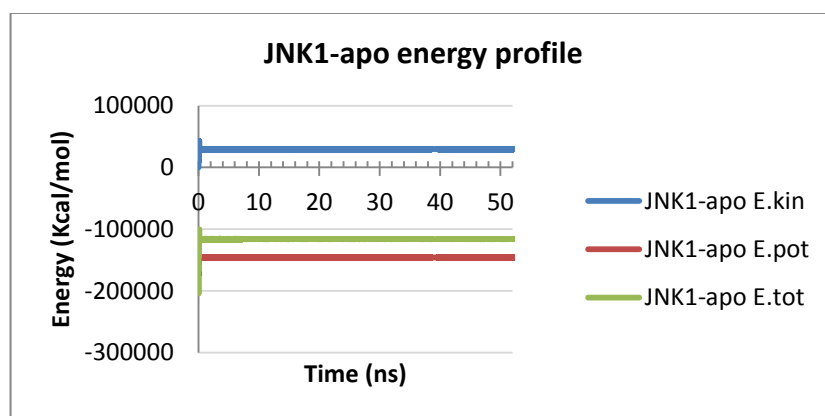


C

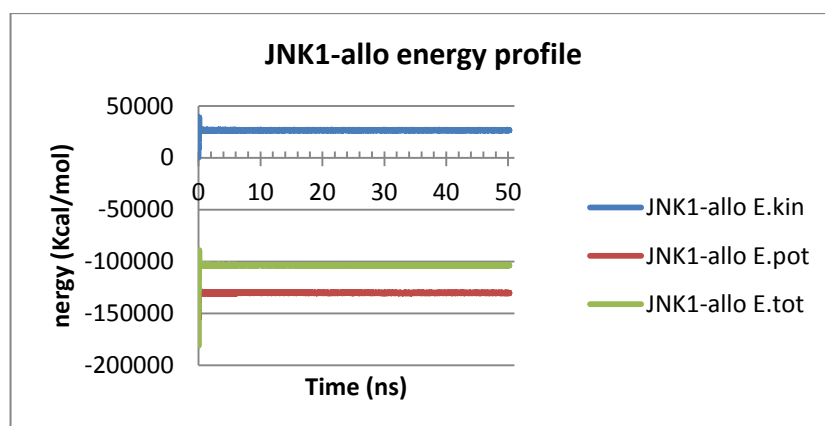
Figure 3.6: The effect of activating the random seed generator (RSG) and restarting the equilibration MD segments after each run on the quality of the generated trajectory: (A) the RMSD; (B) temperature; and (C) the total energy. In each of the three plots the blue line corresponds to simulations without using the RSG and without restarting the MD runs; and the red line corresponds to simulations where the RSG was used and MD segments were restarted. Those plots are for the *JNK1-allo* state.

Energy conservation and temperature stability:

The stability of the simulation can also be assessed by tracing the energy conservation of the system throughout the simulation time, which can be analysed by extracting kinetic, potential and total energy values from the generated trajectory (figure 3.7); along with examining the maintenance of a stable temperature (figure 3.8).



(A)

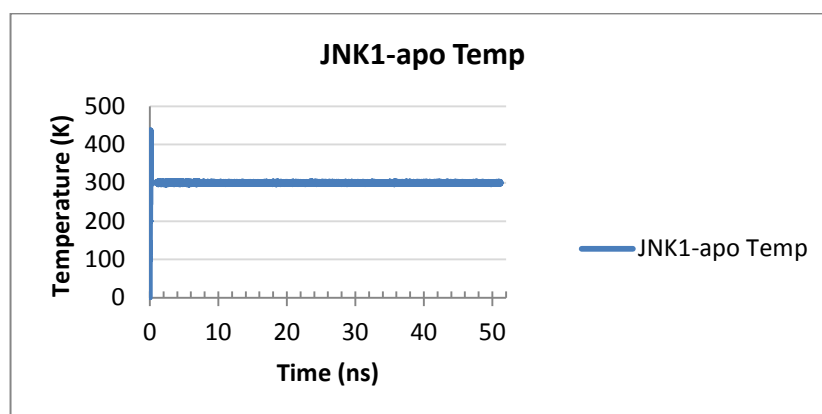


(B)

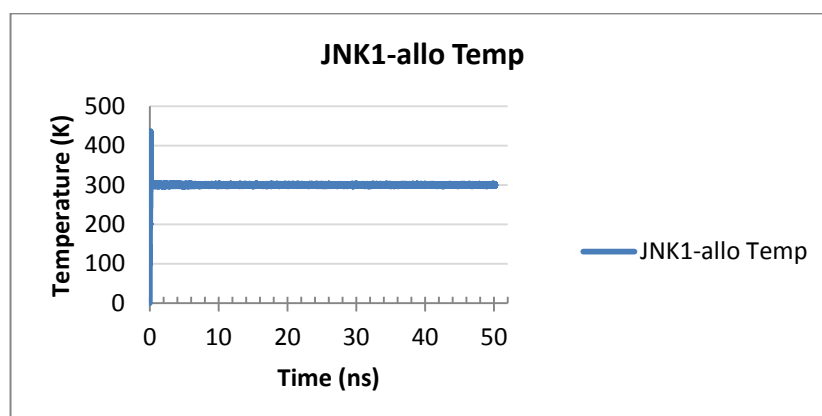
Figure 3.7: Summary of the energy changes for the two JNK1 states versus time. Plot (A) corresponds to the *JNK1-apo* state; and (B) to *JNK1-alo*. Blue lines correspond to the total kinetic energy; red lines correspond to the total potential energy; and the green lines correspond to the total energy.

Figure 3.7 shows that all of the energies increased during the first few ps corresponding to heating from 0 K to 300 K. Thereafter, the kinetic energy remained

constant throughout the simulation time which demonstrates that the temperature thermostat was working correctly. The potential energy also plateaued after the initial increase indicating that the system had relaxed and reached equilibrium. The total energy is the sum of the kinetic and potential energy, and is very well conserved for both systems highlighting their stability.



(A)



(B)

Figure 3.8: Summary of temperature changes for the two states of JNK1 plotted versus time. Plot (A) corresponds to *JNK1-apo* state and (B) to *JNK1-allo*.

The temperature of the two systems started at 0 K and then increased to 300 K over a period of 60 ps. The temperature then plateaued at around 300 K for the remainder of the simulation indicating that the Langevin and Berendsen thermostats were successfully controlling the temperature.

3.1.2.2 Conformational flexibility

Average structures:

To evaluate the overall structural and conformational changes in both simulated states, an average structure from the equilibrated part (last 44 ns) of each trajectory was calculated for each state and compared with its corresponding starting structure (figure 3.9). Structural changes in each system were characterised by calculating the RMSD of the backbone atoms of the entire protein of the average structure from its corresponding starting crystal structure. Figure 3.9 shows that in the apo state there were clear conformational changes affecting the N-terminal domain, the MAPK insert and the terminal α -helix of the C-terminus ($\alpha 14$); while in the allo state, changes between the starting and average structures were almost negligible.

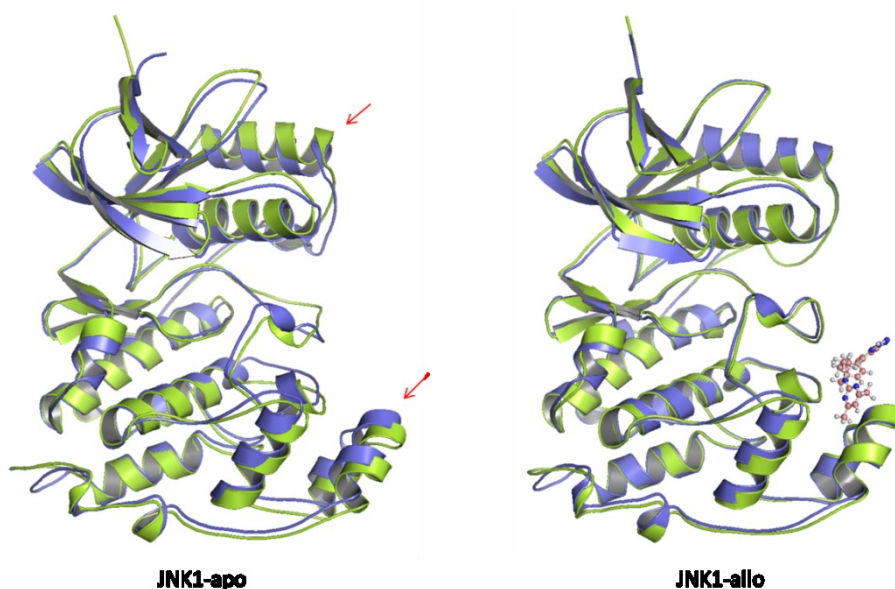


Figure 3.9: Comparison between the starting structures (blue) and the average structure (green) of the two simulated states of JNK1. The superimposition of the starting and average structures was based on alignment of the backbone atoms of the entire protein. Red arrows highlight the regions of the enzyme with substantial conformational changes.

The RMSD of the backbone atoms of the *JNK1-apo* and *JNK1-allo* average structures relative to their minimized starting structures were 1.28 Å and 0.77 Å

respectively. The negligible conformational changes in the complex state and its low RMSD value indicate that the allosteric inhibitor appears to restrain the flexibility of the protein. Figure 3.10 shows the difference in distance between the N and C-terminal domains in the average structures of the two states compared to their starting counterparts. The distance in the *JNK1-apo* state has increased by 2.12Å revealing that it is expanding with time. On the contrary, the distance in the *JNK1-allo* state has increased only by 0.07Å supporting the idea of the restraining effect of the allosteric inhibitor.

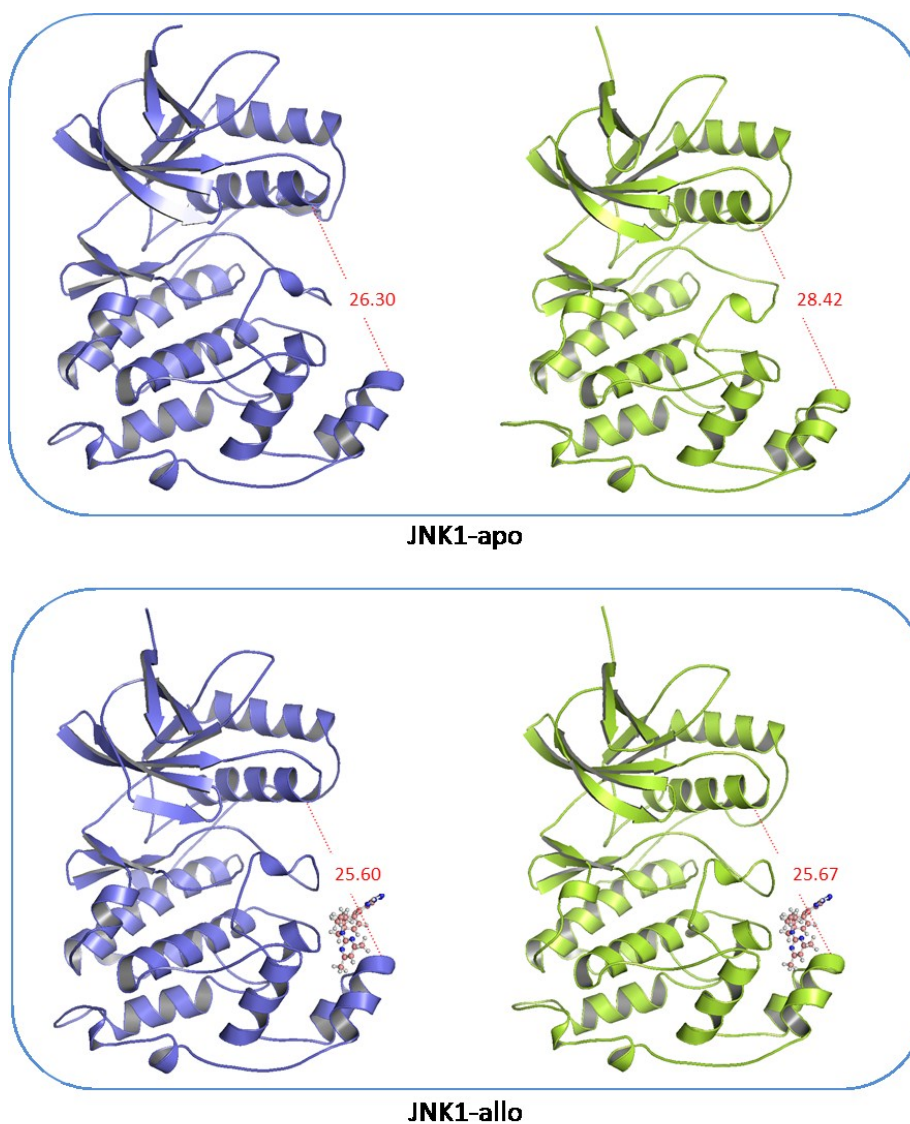


Figure 3.10: Comparison of the distance between the N and C-terminal domains in the starting and minimized average structures for *JNK1-apo* state (upper panel) and

JNK1-allo state (lower panel). Starting structures are represented in blue (left) and average structures are in green (right).

Residual fluctuations:

In order to investigate and explore the conformational variability of each trajectory, their residual fluctuation were calculated and compared, and to obtain the fluctuations without rotations or translations, a root mean square (RMS) fitting of the trajectory to the reference starting structure was performed prior to carrying out this calculation (figure 3.11).

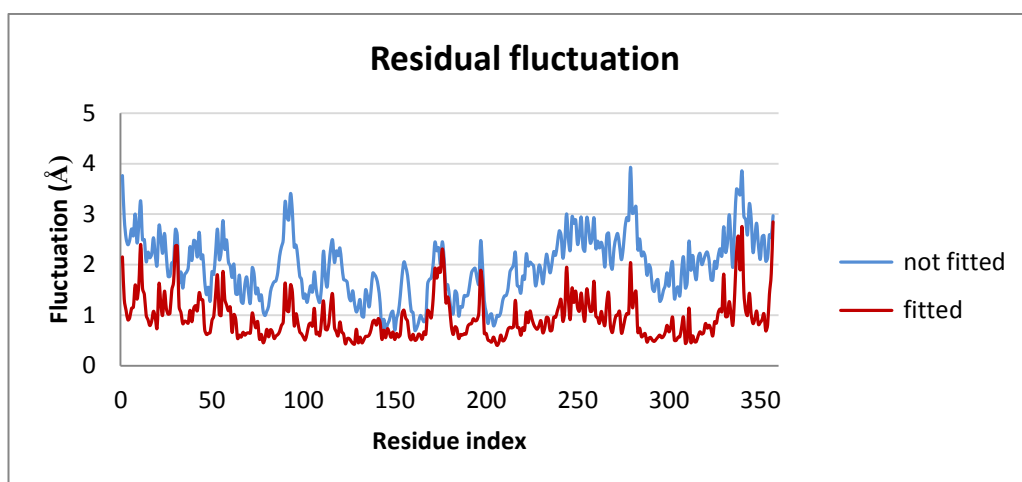


Figure 3.11: The elimination of rotational and translational interference with residual fluctuation by performing an RMS fitting to the starting crystal structure in each trajectory. The blue line corresponds to the residual fluctuation without RMS fitting and the red line corresponds to the RMS fitted plot (without rotations or translations). The plots in this figure correspond to *JNK1-apo* state.

As figure 3.11 shows, the RMS fitting maintained exactly the same pattern of fluctuations but it reduced their magnitude; thereby, giving a real picture of the residual fluctuation that reflects the local conformational variability of each trajectory rather than the overall movement of the protein.

The root mean square fluctuations (RMSF) of *JNK1-apo* residues compared to that of *JNK1-allo* are shown in figure 3.12A. The average residual fluctuation values of *JNK1-apo* and *JNK1-allo* were 1.213Å and 1.078Å respectively. Although the overall difference between the two RMSF values is small it still indicates that binding of the allosteric inhibitor has a restraining effect on the enzyme. Regions in the protein with relatively high fluctuations are highlighted by coloured bars underneath the plot and structurally mapped onto the backbone of the average structure of the *JNK1-allo* state in figure 3.12B using the same colour code.

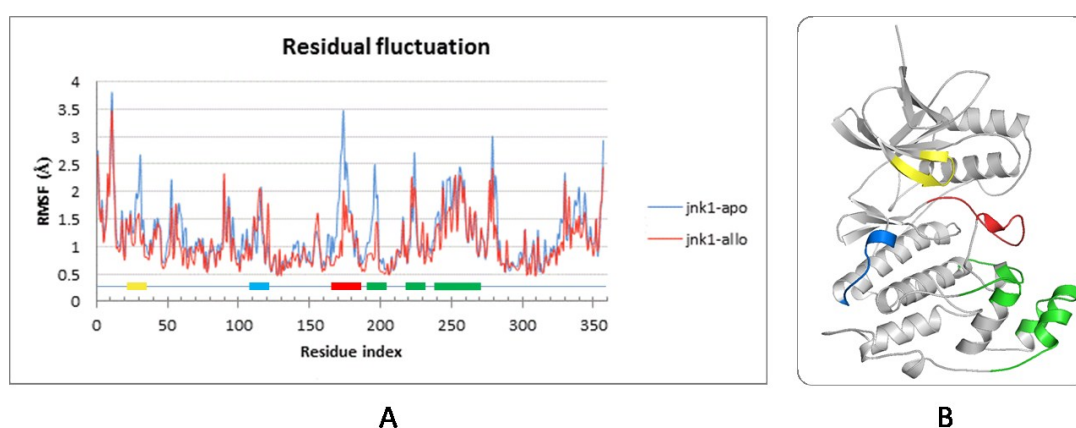


Figure 3.12: Comparison of the RMSF values of the two simulated states of JNK1. (A) RMSF versus residue plot of the *JNK1-apo* state (blue) and *JNK1-allo* state (red). (B) Structure of JNK1 with the most flexible regions highlighted using the same colour code as in (A). The yellow bar corresponds to residues 26-33 which includes the G-loop that forms the roof of the ATP binding site. The blue bar corresponds to residues 110-119 which are part of the docking groove of the JIP1 peptide. The red bar corresponds to residues 163-180 which form the T-loop. The green bar corresponds to residues 191-198, 220-227, and 242-260 that collectively form the allosteric binding site. All of these regions are among the most flexible regions in the protein and they are also functionally important.

To have a clearer perception of regions that experienced a notable change in the magnitude of their flexibility as a result of binding the allosteric inhibitor, the fluctuation values of the *JNK1-apo* trajectory were subtracted from that of the *JNK1-*

allo and the difference was plotted against residue index in figure 3.13A. The regions of highest flexibility difference were structurally mapped onto the backbone of the average structure of *JNK1-allo* state as shown in figure 3.13B using the same colour code in figure 3.12 for comparison.

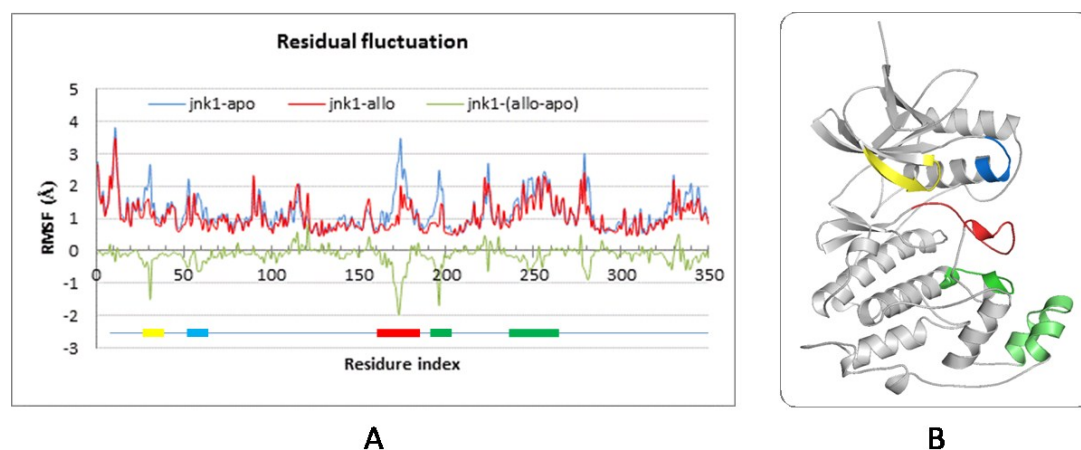


Figure 3.13: Fluctuation difference between the two states of JNK1. (A) Subtraction of the RMSF values of *JNK1-apo* (blue line) from that of *JNK1-allo* (red line) revealed the regions in the enzyme that encountered a reduction in their fluctuation values upon binding the allosteric inhibitor (green line). (B) Structural mapping of the regions that experienced the highest flexibility difference. The yellow bar corresponds to residues 25-32 (the G-loop). The blue bar corresponds to residues 55-60 which are part of the α C-helix. The red bar corresponds to residues 163-180 (the T-loop). The green bar corresponds to residues 190-200, and 240-255 (the allosteric binding site). Flexibility of residues 240-255 was slightly reduced compared to the other coloured regions.

As figure 3.13B shows, all of the regions whose flexibility was reduced because of the allosteric inhibitor match the most flexible regions in the protein as shown in figure 3.12, except the JIP1 docking groove. This is interesting in terms of how the allosteric inhibitor may affect the function of the enzyme by restraining the motion of functionally important segments.

In order to facilitate the visual inspection of the overall fluctuation profiles of the two simulated JNK1 states, the residual fluctuation values were converted into a colour code and structurally mapped onto the backbone of the corresponding average structure of each state (figure 3.14).

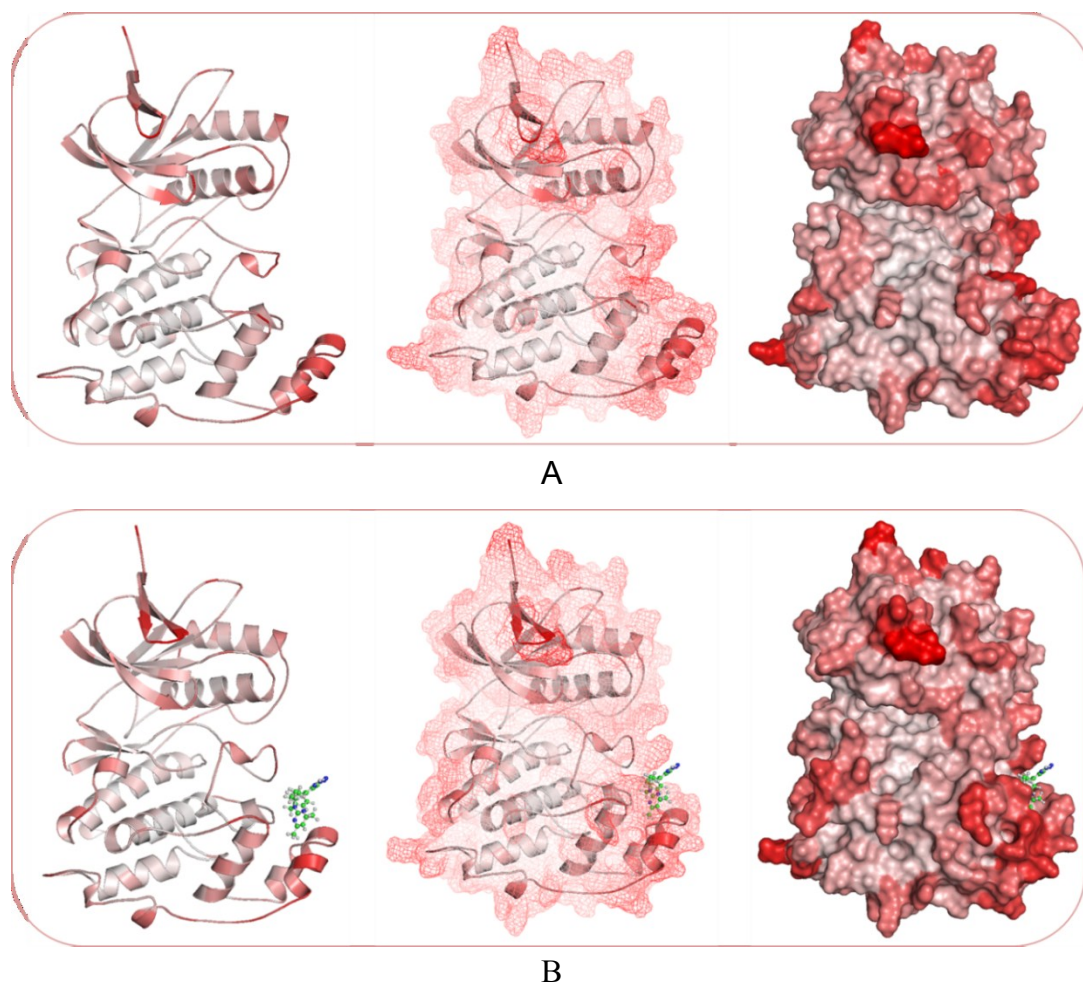


Figure 3.14: Colour coded representation of the residual fluctuation values of the simulated states of JNK1. (A) Represents the fluctuations of *JNK1-apo* state; and (B) the *JNK1-allo* state. The highest fluctuating regions are coloured red and the lowest are coloured white. For each state there is a cartoon, a mesh surface and a solid surface representation. The allosteric inhibitor is shown using a ball and stick representation.

Analysis of the residual fluctuation of the two systems in figures 3.12-14 revealed that the regions of the enzyme with the highest fluctuation values correspond to the G-loop, the docking groove of JIP1 ($\alpha 2$ helix), the T-loop, and the residues forming the allosteric binding site. It is clear from figure 3.12A that *JNK1-allo* has lower flexibility compared to *JNK1-apo*, especially in the region of the activation loop. A clearer picture of the interaction changes that seems to be responsible for this restraining effect of the allosteric inhibitor can be revealed by examining the network of hydrogen bonds in both states (figure 3.15).

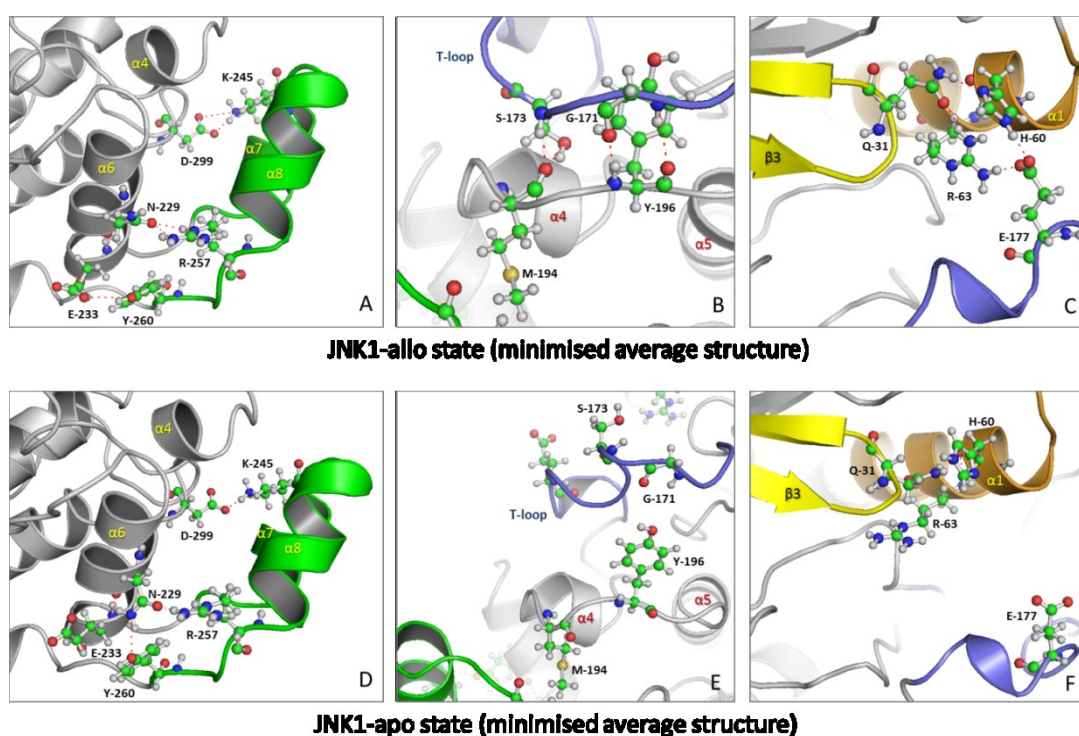


Figure 3.15: Comparison of the networks of hydrogen bonds in the minimised average structures of both states of JNK1. The upper panel corresponds to *JNK1-allo* and the lower one *JNK1-apo*. (A) and (D) show the differences in the network of hydrogen-bonds in the MAPK insert region of *JNK1-allo* and *JNK1-apo* states respectively. (B) and (E) compare those between the T-loop and the loop connecting helices $\alpha 4$ and $\alpha 5$. (C) and (F) compare those between the T-loop and the αC -helix. MAPK insert is in green, T-loop is in blue, the αC -helix is in orange, the G-loop is in yellow, and the rest of the protein is in grey. Residues involved in hydrogen bonds are coloured by element and represented as balls and sticks.

Most of the differences in the hydrogen-bond network involve the T-loop and the MAPK insert. In the MAPK insert region there are stronger interactions with the body of the protein in the *JNK1-allo* state compared to that in the *JNK1-apo* state. In the *JNK1-allo* state, Lys245 forms two hydrogen bonds with Asp299; Arg257 forms another two hydrogens bonds with Asn229; and Tyr260 forms one hydrogen bond with Glu233 (figure 3.15A). In the *JNK1-apo* state, each of Lys245 and Arg257 has only one hydrogen bond with Asp299 and Asn229 respectively, and Tyr260 forms a hydrogen bond with Asn229 instead of Glu233 (figure 3.15D). More important is the network of hydrogen bonds in the T-loop region.

The T-loop in the *JNK1-allo* state has adopted a different conformation from that of the *JNK1-apo* state which is accompanied by different networks of hydrogen bonds above and below the loop. For example there are four hydrogen bonds between the T-loop and the loop connecting helices $\alpha 4$ and $\alpha 5$ where Gly171 forms two hydrogen bonds with Tyr196, and Ser173 forms another two hydrogen bonds with Met194; there are none in the equivalent region of *JNK1-apo* (figure 3.15B and E). Furthermore, there is an interesting network of hydrogen bonds between the T-loop and the αC -helix where the orientation of the acidic residue Glu177 in the T-loop plays a critical role in the overall conformation of the region where it forms two hydrogen bonds with His60 and Arg63 of the αC -helix, and each of the latter two has another hydrogen bond with Gln31 of the G-loop (figure 3.15C). Again, none of these hydrogen bonds exists in the *JNK1-apo* state (figure 3.15F). This network of hydrogen bonds is the most likely explanation for the reduced flexibility of the enzyme in the *JNK1-allo* state.

In order to check whether this network of hydrogen bonds was established as the complex evolved with time or if it already existed in the starting structure from which the enzyme was unable to escape and become more flexible, the same analysis of the hydrogen bonds in the starting structures of the two states was conducted (figure 3.16).

Figure 3.16 shows that the starting crystal structure of the *JNK1-allo* state has a very similar network of hydrogen bonds as the average structure. It has exactly the same interactions in the MAPK insert region and has three identical hydrogen bonds

between the T-loop and the loop connecting helices α_4 and α_5 (there is only one hydrogen bond between S173 and M194). The main difference between the two is in the region between the T-loop and the α_C -helix where Glu177 forms two hydrogen bonds to His60 and there is no involvement of either Arg63 or Gln31 in the G-loop. While in the *JNK1-apo* state, the networks of hydrogen bonds and the orientation of the residues forming them in the two regions above and below the T-loop are quite different, although there is a similar pattern in the MAPK insert region with four hydrogen bonds between the MAPK insert and the body of the enzyme. These findings are very significant because they suggest that the allosteric inhibitor has restrained the flexibility of the enzyme from the outset, by establishing a strong network of hydrogen bonds that trap the enzyme in a non-functional state. The comparatively weak network of hydrogen bonds in the *JNK1-apo* state enabled the enzyme to evolve more freely and experience more conformational changes.

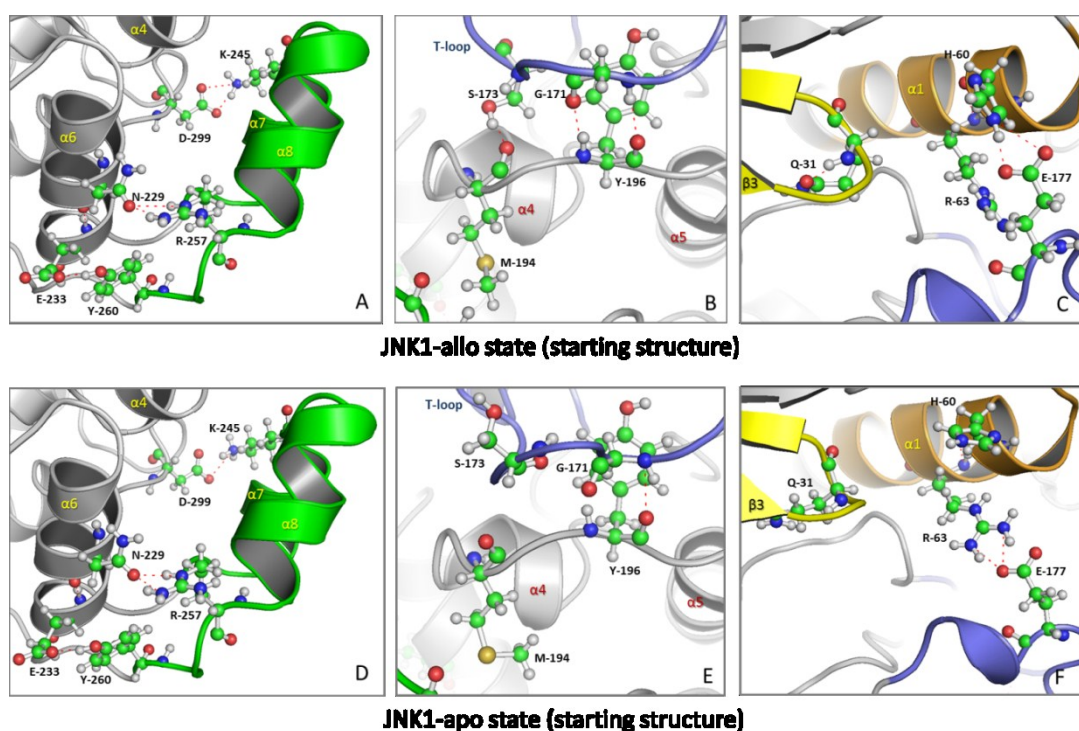


Figure 3.16: Comparison of the networks of hydrogen bonds in the starting structures of both states of JNK1. The upper panel corresponds to JNK1-allo and the lower one JNK1-apo. The same orientation as in figure 3.15 is used here.

3.1.2.3 Contact maps:

A contact map of a protein is a two dimensional representation of the distance between all amino acid residue pairs in its 3D structure represented in the form of a binary symmetrical matrix. Residue pairs that are located within a predetermined cut-off in the 3D structure of the protein will have a value of 1 in the matrix and those further than the cut-off will have a value of 0. The cut-off distance can take a value ranging from 6-16 Å and is usually measured between the C α of residue pairs [148]. In order to describe the structural (conformational) differences between the two simulated states of JNK1 in more detail, the contact maps of their average structures were calculated and compared to highlight the regions that have changed upon binding of the allosteric inhibitor (figure 3.17).

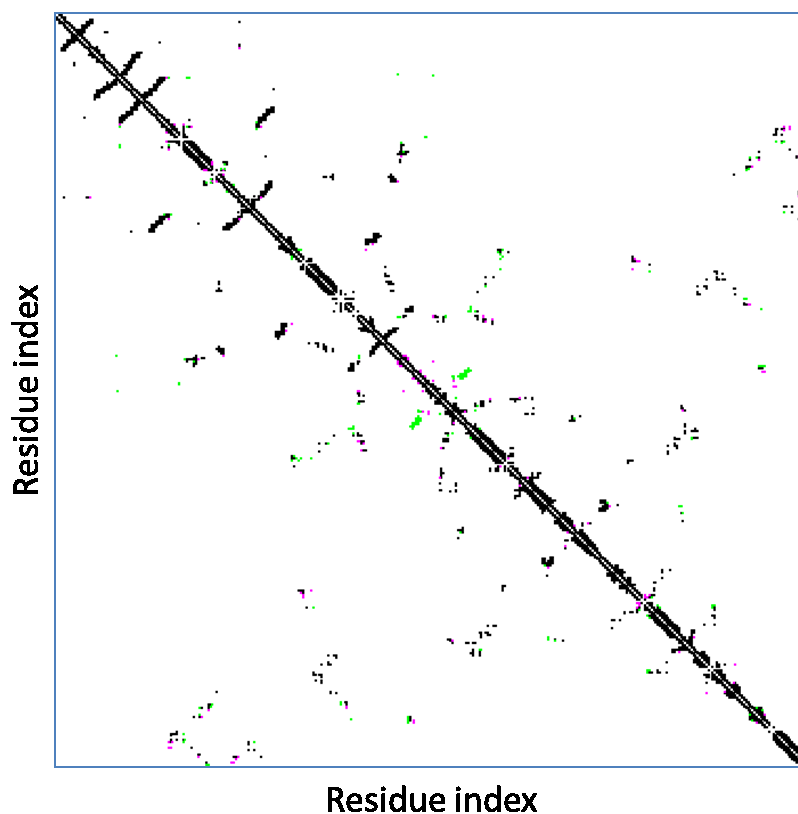


Figure 3.17: Comparison of the contact maps of the minimized average structures of *JNK1-apo* and *JNK1-allo* states. The black dots are common contacts present in both states; pink dots are unique contacts for the *JNK1-apo* state; and green dots are unique contacts for the *JNK1-allo* state.

In the *JNK1-apo* state based on a cut-off of 8 Å, there were 1,553 contacts of which 1,484 were common contacts (existing in both states) and 69 were unique contacts for this state. In the *JNK1-allo* state there were 1,579 contacts of which 1,483 were common and 95 were unique. In total there were 26 more unique contacts in the *JNK1-allo* state which seemingly have an effect on the overall configuration of the protein. Structural mapping of the unique contacts of each state onto the backbone of its corresponding average structure allowed the identification of those regions that experienced conformational changes resulting from binding the allosteric inhibitor (figure 3.18).

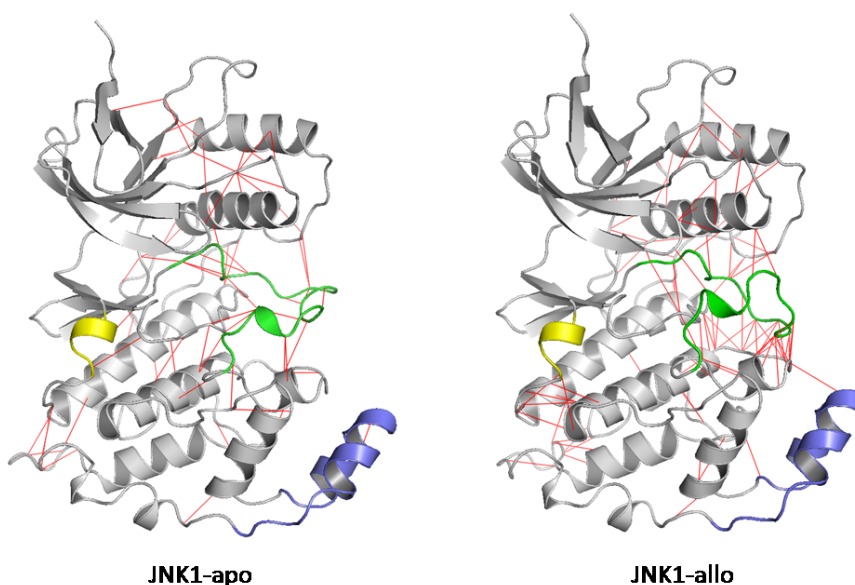


Figure 3.18: Structural mapping and comparison of the contact maps of the minimized average structures of the *JNK1-apo* (left) and *JNK1-allo* (right) states. The unique contacts in each state are represented as red lines, the T-loop is shown in green, the MAPK insert is in blue, and the $\alpha 2$ helix (part of the JIP1 docking groove) is in yellow.

As figures 3.17-18 show, there are more unique contacts in the *JNK1-allo* state which are more localized and concentrated compared to those in the *JNK1-apo* state. They are mainly concentrated around and involve the T-loop where it forms a bridge between the N and C-domains of the enzyme and appear to restrain the free motion of the two domains and that of the T-loop itself. It is clear from figure 3.18 that the

network of unique contacts in the two states is quite different. For example in the *JNK1-allo* state the T-loop has 23 contacts with the $\alpha 4$ helix and the loop beneath it that connects helices $\alpha 4$ and $\alpha 5$ compared to only 4 contacts in the *JNK1-apo* state. It also has 10 contacts with the αC -helix and the loop connecting helices $\alpha 13$ and $\alpha 14$ compared to 3 contacts in the *JNK1-apo* state. There are two unique contacts in the *JNK1-allo* state involving the MAPK insert region that have no equivalents in the *JNK1-apo* state. Conversely, there are two contacts in the *JNK1-apo* state involving the $\alpha 2$ helix (JIP1 docking groove) that are absent in the *JNK1-allo* state.

These networks of unique contacts and hydrogen bonds suggest that the binding of the allosteric inhibitor has changed the conformation of the MAPK insert along with that of helices $\alpha 4$ and $\alpha 5$ which in turn has affected the conformation of the T-loop which regulates the activity of the enzyme. It seems that the *JNK1-allo* state has been locked in a conformation that is functionally inactive, whereas the *JNK1-apo* state is more flexible and is capable of populating more conformational ensembles that enable productive binding with its substrates.

3.1.3 Analysis of correlated motion

In biomolecules there is a relationship between the correlated motion of structural elements and their function (i.e. the transduction of an allosteric signal). In enzymes for example, it has been proposed that the motion of residues both in and distant from the active site contribute to the catalytic activity of the enzyme. These correlated motions are difficult to identify and assess experimentally, which makes MD simulations a particularly useful for their analysis [169, 170].

The potential allosteric effects of JNK1 were investigated by studying the residual dynamic cross-correlation motion based on the simulation results as described in the Computational Methods section. Correlation analysis shows how residues communicate within the protein to propagate a structural or dynamical change in one site of the protein to another (such as the active site) thereby modulate its activity. By applying this approach the concerted, non-random fluctuations of residues can be identified as a function of the protein state [145]. Figure 3.19 shows the residual dynamic cross-correlation matrices calculated from the simulation trajectory of each

state and represented as heat maps for easier visual comparison. A large positive correlation (red pixels) corresponds to highly coordinated motion of the atom-set pair along the same direction, whereas a negative correlation (blue pixels) indicates motion in opposite directions. The scale of the colouring scheme is from -0.5 (negatively correlated residues) to 1.0 (positively correlated residues). The overall average of residual correlations in the *JNK1-apo* state was 0.007 and the average of the positively correlated residues was 0.200; for the *JNK1-allo* state the average of the overall correlations was 0.009 and that of the positively correlated residues was 0.187. Residues with correlation values of more than 0.4 (which is twice the average of positive correlations) were therefore considered as highly correlated with each other.

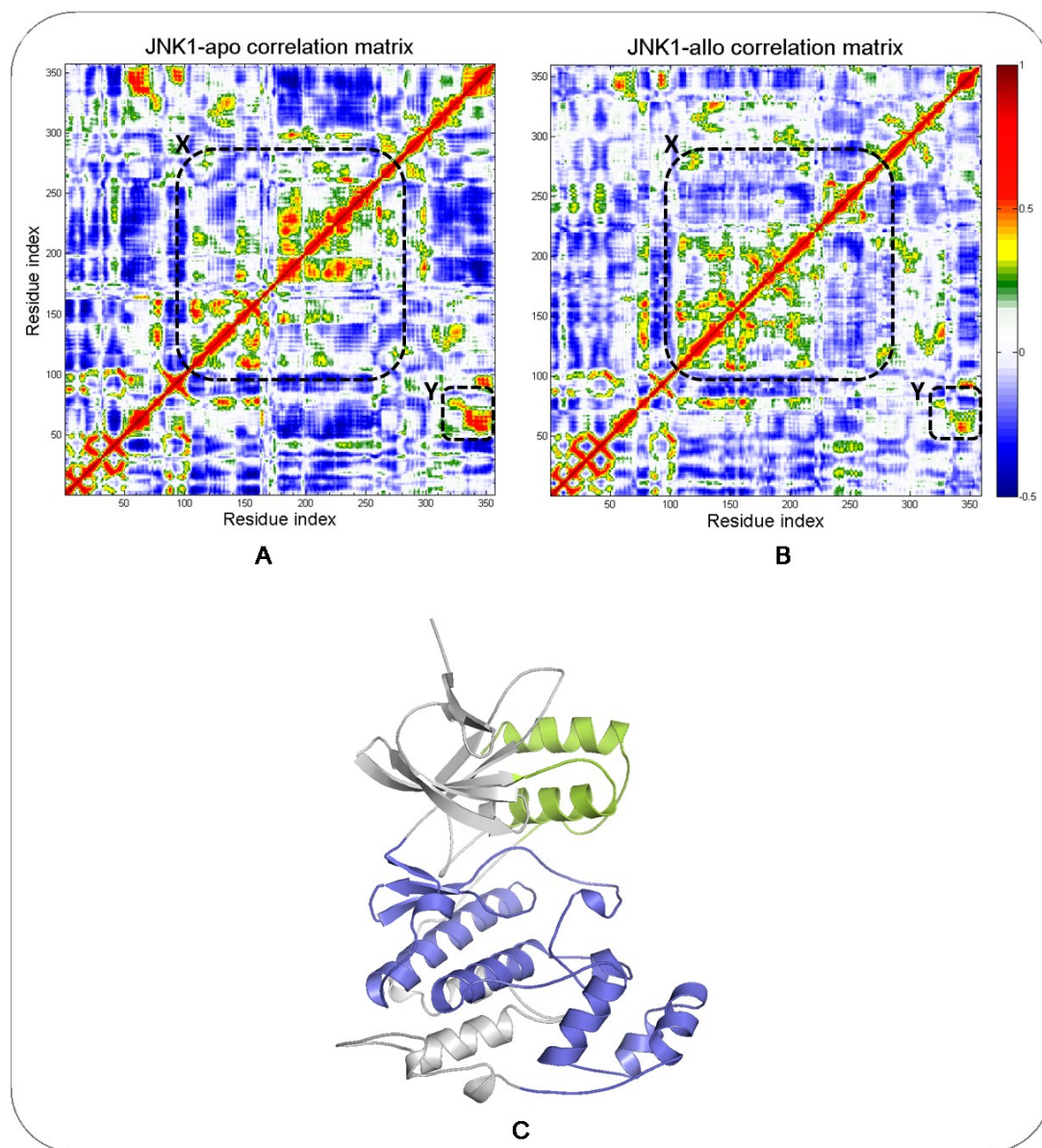


Figure 3.19: The residual cross-correlation matrices of the backbone atoms around their average position calculated from the simulated trajectories and represented as heat maps. (A) The heat map of the cross-correlation matrix of the *JNK1-apo* state, and (B) shows the *JNK1-allo* state. (C) Cartoon representation of the protein segments corresponding to the concentrated residue-residue coupling areas in (A) (dashed black lines) where blue segment corresponds to rectangle X and the pale green segment corresponds to rectangle Y.

As figure 3.19 shows, the number and nature of correlated residues varies for the two states with the total number of highly correlated residues in the *JNK1-apo* state being more than those in the *JNK1-allo* (more red pixels). Since the residual correlation matrix is a symmetrical (N×N) where N is equal to the number of residues in the protein of interest, the total number of correlations in the *JNK1-apo* state were 63546 [(357×356)/2] of which 3557 were highly correlated; while in the *JNK1-allo* state, the total number of correlations were 63903 [(358×357)/2] excluding the allosteric inhibitor from residue index] of which 2676 were highly correlated.

Inspection of the regions of major residual correlations (dashed rectangles) in figure 3.19 revealed that many (more than 24%) of the correlated residues in the *JNK1-apo* system have been erased or severely attenuated in the *JNK1-allo* state, suggesting that binding of the allosteric inhibitor has reduced many residue-residue couplings. If residues whose correlations have been attenuated are involved in conveying intramolecular residual communication important for proper coordination of protein function (maybe through an allosteric manner), disrupting their native states might affect protein function. The *JNK1-apo* state in this instance could provide a special conformational ensemble representative of intrinsic allosteric correlation. A closer examination of the correlations in this state revealed that the most concentrated residue-residue coupling area in the heat map was found in the region between residue 105 and residue 263, corresponding to a protein segment structurally located between the ATP binding site and the allosteric binding site (figure 3.19C). The existence of correlated residues in this region seems to be functionally important for the allosteric (MAPK insert) binding site to respond to conformational changes in the active site and vice versa. Another region of highly correlated residues is between residues 51-70 and 331-357, which corresponds to the α C-helix and the C-terminal helix (α 14). Generally, most of the highly correlated residues are either sequentially or structurally close to each other. Therefore, distant correlated residues are more likely to be involved in intramolecular interaction pathways that are implicated in controlling allosteric effects, since they can mediate signal propagation between distant parts of the protein. For a better visual inspection, the off-diagonally correlated residues in the *JNK1-apo* state heat map were identified and colour-mapped onto the backbone of its average structure (figure 3.20).

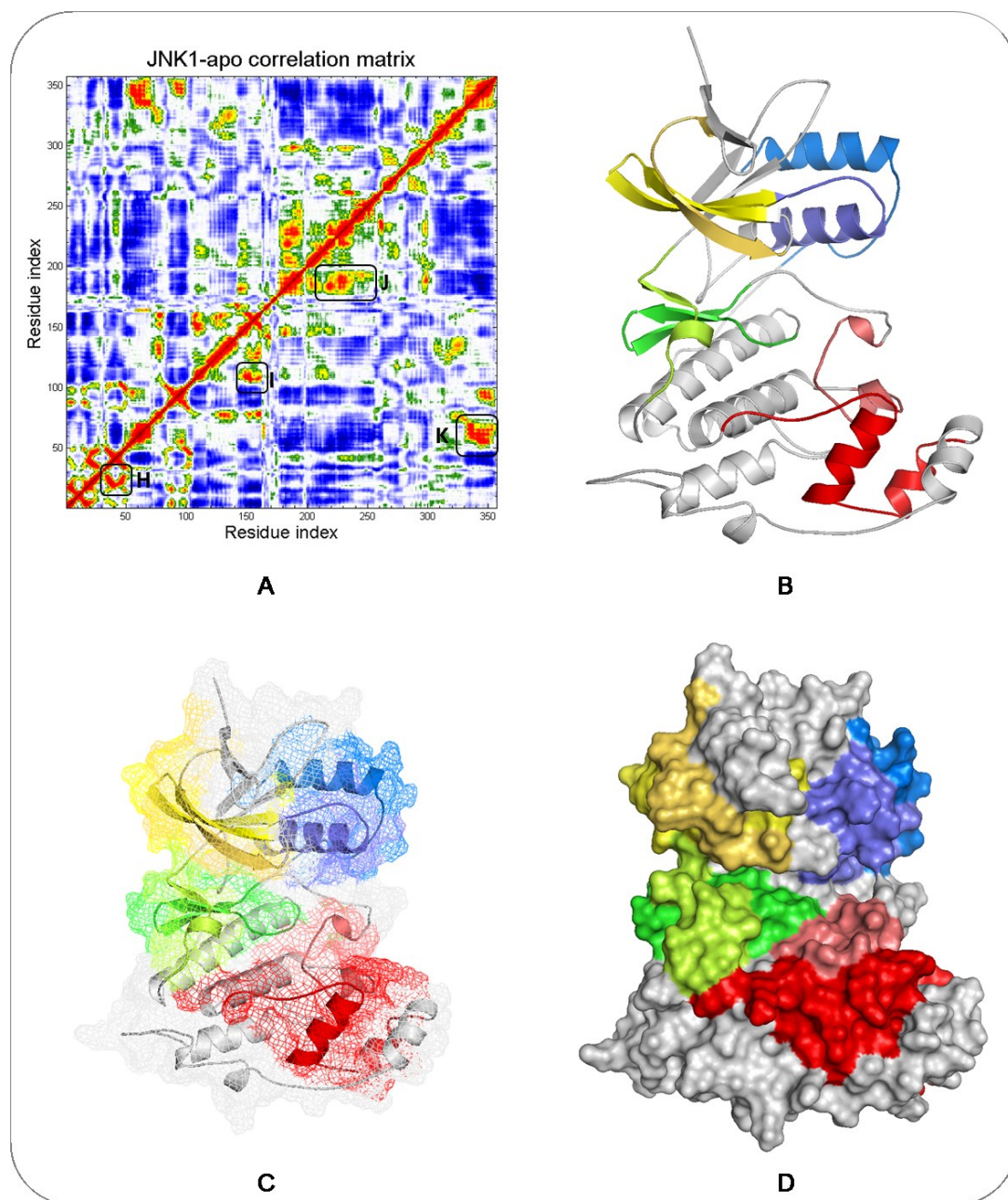


Figure 3.20: Regions of correlated residues in the *JNK1-apo* state. (A) The heat map of the cross correlation matrix of the *JNK1-apo* state with areas of high off-diagonal correlation highlighted by black rounded rectangles H-K. (B) Structural mapping of the highly correlated residues in (A); yellow segment corresponds to box H, green segments corresponds to box I, red segment corresponds to box J, and blue segment corresponds to box K. Residues of each coloured segment are positively correlated together. (C and D) are respectively the mesh and surface representation of the correlated residues as coloured in (B).

Figure 3.20 shows that areas of highly correlated residues can be defined into four distinct regions. The first region corresponds to rectangle H (coloured yellow in figure 3.20B-D) in which residues 16-29 are correlated with residues 33-50; this region includes the G-loop which is the roof of the ATP binding site. The second region corresponds to rectangle I (coloured green in figure 3.20B-D) in which residues 104-115 (part of the allosteric binding site of the JIP1 peptide) are correlated with residues 144-162 that forms the floor of the ATP binding site. The third region corresponds to rectangle J (coloured red in figure 3.20B-D) where residues 178-196 are correlated with residues 215-247 that form the allosteric binding site. The last region corresponds to rectangle K (coloured blue in figure 3.20B-D) in which residues 51-70 (the α C-helix) are correlated with residues 331-357 that are the C-terminal helix (α 14).

Besides those correlated regions in *JNK1-apo*, the heat map also shows an anti-correlated region which is highlighted in figure 3.21 by a rounded red rectangle. This region corresponds to residues 178-256 (the allosteric binding site) which are anti-correlated with residues 307-352 (helices α 12, α 13 and α 14). Binding of the allosteric inhibitor caused most of the anti-correlated residues in these regions to become non-correlated (neither correlated nor anti-correlated). Most importantly, it appears that the binding of the allosteric inhibitor has erased the correlations of residues in box J (the allosteric site) in the *JNK1-apo* state, and severely attenuated the correlations between residues in box K (helices α 1 and α 14) and to a lesser extent those in box I (the ATP binding site and the JIP1 docking groove) (figure 3.22).

The four identified regions of high residue-residue coupling can be viewed as being sub-domains within the enzyme and since protein dynamics are crucial for its function, perturbation of the dynamics of any of these sub-domains (such as ligand binding) will interfere with the overall intrinsic dynamics of the protein and ultimately affecting its function. In the case of JNK1, binding of the peptide segment (JIP1) to a region on the protein involving box I in figure 3.20 alters protein movement and has an allosteric inhibitory effect on protein activity. Likewise, the allosteric inhibitor which binds to a region on the protein corresponding to box J in

figure 3.20 alters protein movement and inhibits activity. If this is the case, it is possible that ligand binding to the other identified regions in boxes H and K (figure 3.20) may allosterically modulate protein activity.

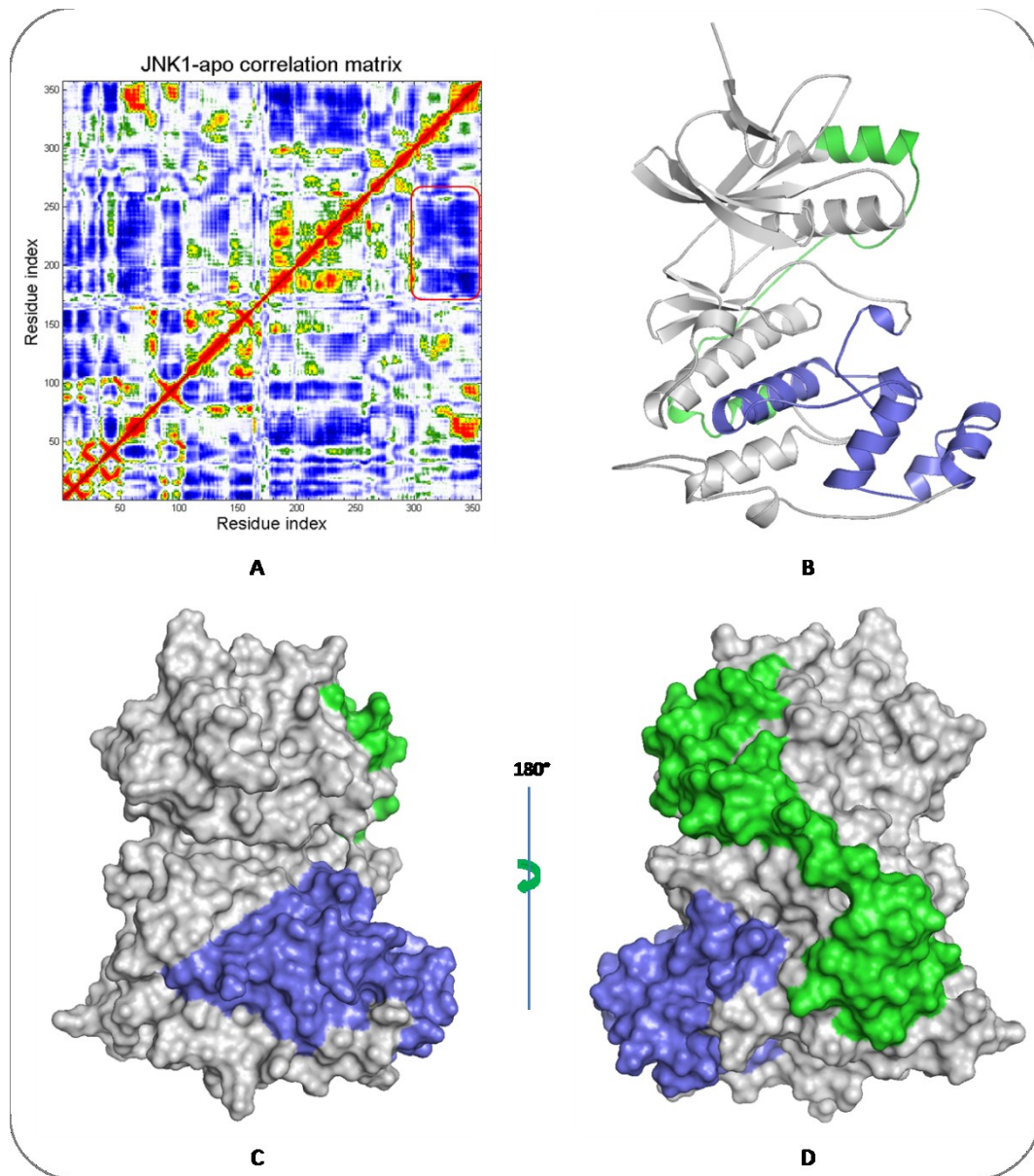


Figure 3.21: The main anti-correlated region in *JNK1-apo*. The anti-correlated region is highlighted by the rounded red rectangle in (A), and structurally mapped onto the backbone of the protein in (B). (C) Surface representation of the anti-correlated regions as coloured in (B). (D) Back view of the anti-correlated region by rotating (C) 180° about the y-axis.

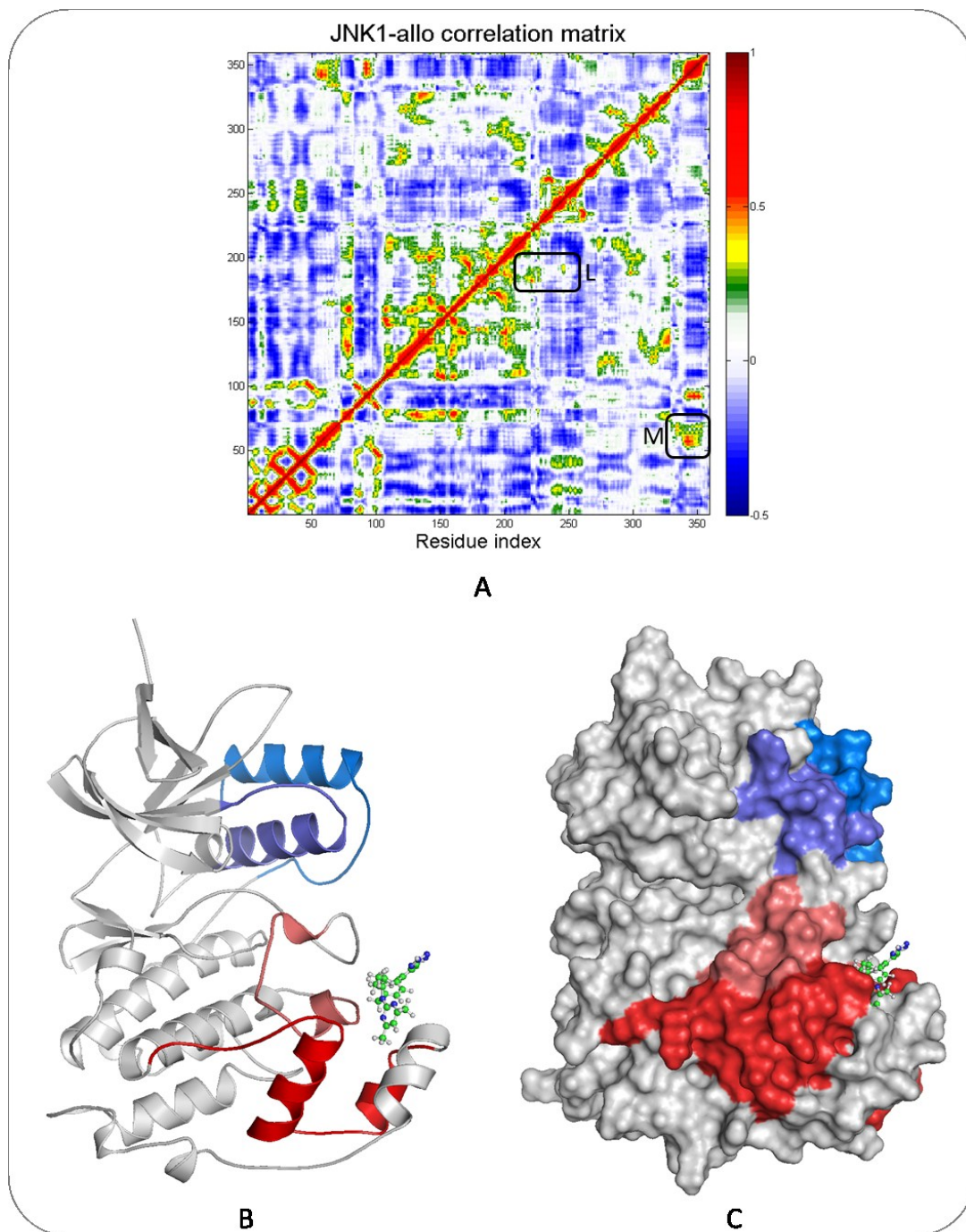


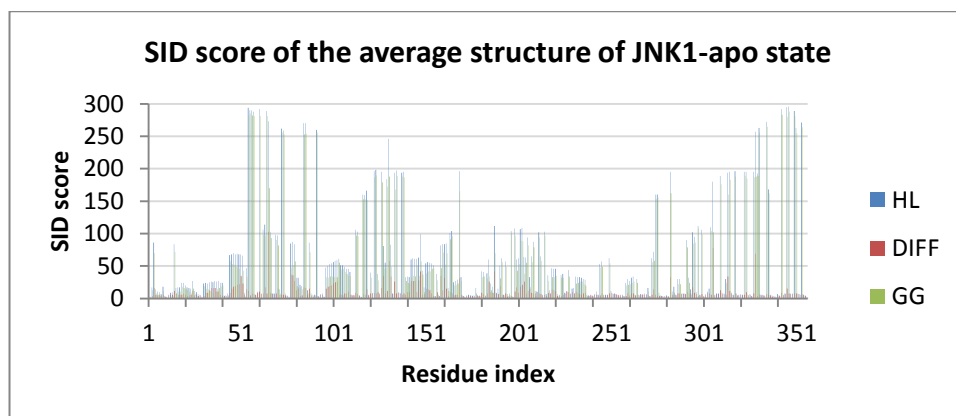
Figure 3.22: The effect of binding the allosteric inhibitor on the correlated residues in *JNK1-alo*. The black rounded rectangles in (A) shows the region where the correlations between residues have been erased or severely attenuated upon binding of the allosteric inhibitor (compared to the *JNK1-apo* heat map). Those regions are structurally mapped on the backbone of the protein in (B) and represented as surface in (C). The same colouring scheme as in figure 3.20 is used.

3.1.4 Simple intrasequence difference analysis

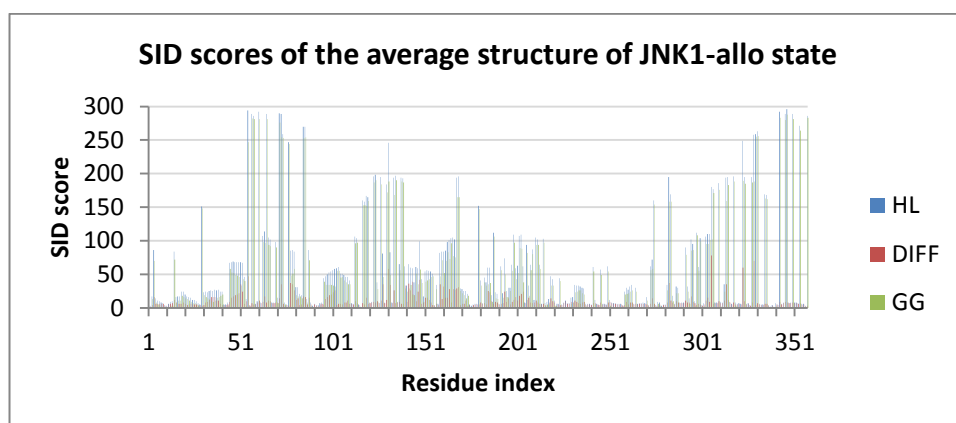
3.1.4.1 SID analysis of the minimised average structures of JNK1

SID analysis was applied to the single average structure of each simulated state of JNK1 to examine the basic potential of topological and sub-structural vulnerabilities to perturbations in order to identify the major interfaces in the enzyme that may be involved in allosteric modulation. In this context, the *JNK1-apo* state represents the native state of the protein and SID analysis of the average structure obtained from the simulation trajectory highlights the natural (intrinsic) interfaces within the protein that may be sensitive to structural perturbation. Because there are two simulated states of the same protein, their minimised average structures can be utilised to perform a comparative SID analysis in order to track interfacial changes resulting from ligand binding (the allosteric inhibitor). Ligand binding at sites covering an interface can destabilize the underlying interface because the residues of that interface will become involved in forming new interactions with the ligand at the expense of forming intramolecular interactions. As a consequence of this destabilization, there may be conformational changes in the protein that result in modulation of its function.

As mentioned in section 1.1.5.2.1, complex interfaces between protein segments tend to be more vulnerable to reorientation and motion induced by distant structural perturbation; and that the general trend in SID scores that identify such interfaces is to have a high HL and intermediate-to-high DIFF and intermediate GG scores. These three SID scores were collectively considered as described previously, to produce a single consensus score that guides the identification of protein interfaces of allosteric potential. Individual SID scores of the minimised average structures of the JNK1 states are presented as a column chart in figure 3.23. In order to highlight the residues that are implicated in multi-way interfaces, statistical functions were used to calculate the upper, median and lower quartiles of each of the SID scores (table 3.1), and then the logical functions within MS[®] Excel were utilised to select those residues which collectively have high HL and DIFF scores and intermediate GG scores. To be selected, residues must have HL and DIFF scores larger than their corresponding median value (MQ) and a GG score lying within the inter-quartile distance (IQD).



A



B

Figure 3.23: Column chart of the individual SID scores of each residue position of the minimised average structures of the *JNK1-apo* state (A) and the *JNK1-allo* state (B). HL scores are coloured blue, DIFF scores are coloured red and GG scores are coloured green.

Table 3.1: Statistical descriptors of the JNK1 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.

Statistical descriptors	SID scores of <i>JNK1-apo</i> state			SID scores of <i>JNK1-allo</i> state		
	HL	DIFF	GG	HL	DIFF	GG
Upper quartile (UQ)	68	10	52.5	84	11	57.25
Lower quartile (LQ)	7	5	1	8	5	1
Median (MQ)	25	7	16	27	7	19
Inter-quartile distance (IQD)	61	5	51.5	76	6	56.25
Average (AVG)	56.83	9.823	47.011	59.927	10.151	49.776
Minimum value (MIN)	2	1	1	2	1	1
Maximum value (MAX)	296	103	288	296	78	288

Since a graphical representation of SID scores is more informative through allowing the visualisation of the highlighted residues and clusters within the 3D structure of the protein, the numerical value of the consensus SID score of each residue in the two simulated states of JNK1 was converted into a colour code and structurally mapped onto the backbone of the corresponding average structure (figure 3.24).

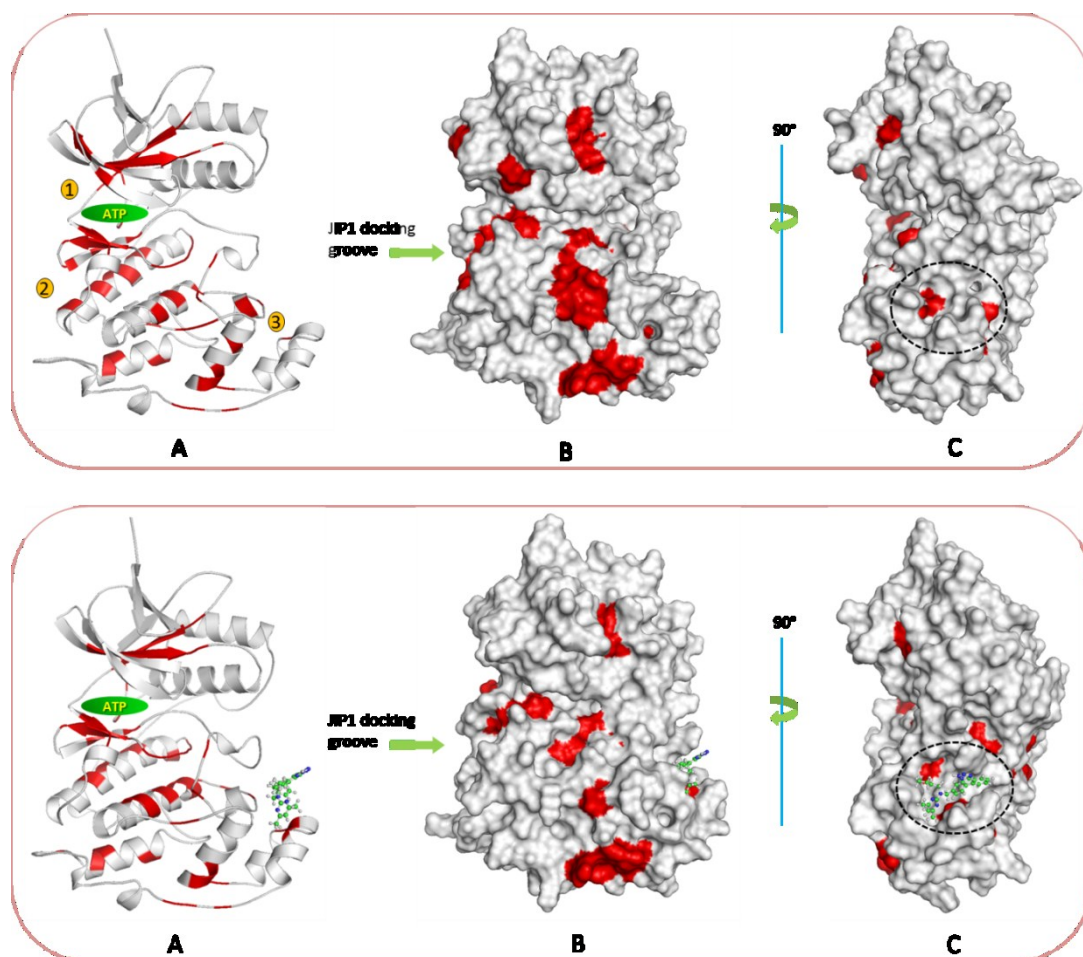


Figure 3.24: SID analysis of the minimized average structures of the two simulated states of JNK1. The upper panel is for *JNK1-apo* and the lower one is for *JNK1-allo*. The SID consensus scores are colour coded and structurally mapped onto the backbone of the corresponding average structure where red colour stands for high scores and white is for low scores. (A) Cartoon representation of the protein backbone. The ATP binding site is highlighted by a green ellipsoid. (B) Surface representation of the protein. The JIP1 binding groove is highlighted by a green

arrow. (C) Rotation about the y-axis of (B) to show the allosteric binding site which is highlighted by a dashed black ellipsoid. The allosteric inhibitor is shown in ball and stick. The numbered yellow circles highlight the three binding sites of JNK1 and at the same time highlight the regions of the protein that experienced a change in their SID score upon binding the allosteric inhibitor.

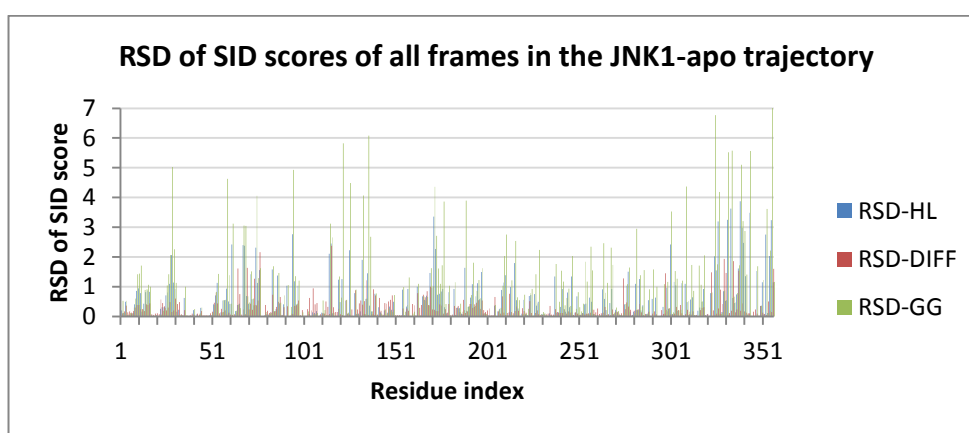
Figure 3.24 shows that SID analysis was not only successful in identifying the known binding sites of JNK1 namely the ATP binding site, the JIP1 docking groove and the allosteric binding site but also in tracking the changes in SID scores that result from changes in protein 3D structure associated with ligand binding. Although the allosteric binding site in JNK1 is distant from both the ATP and JIP1 binding sites, structural perturbation at that site (ligand binding) has propagated to the other two which is manifested as a change in SID scores of some residues in those sites and reflected in figure 3.24 as the disappearance of some red coloured residues in the *JNK1-allo* state compared to *JNK1-apo*.

3.1.4.2 Comparative SID analysis of the entire trajectory of each simulated state

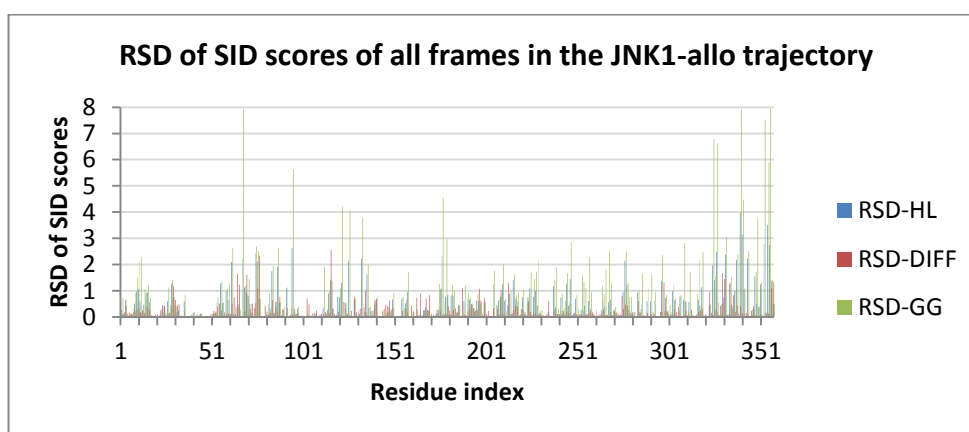
Since we are interested in studying the behaviour of the protein as it evolves with time so that regions in the protein that underwent prominent topological changes can be identified, all the frames from each simulated trajectory were extracted in PDB file format and subjected to comparative SID analysis. In order to investigate the effect of protein dynamics on SID scoring and to exclude the inherently high SID scores of residues that are located in structurally compacted regions of the protein, the relative standard deviation (RSD) of the SID scores for each residue throughout the simulation time was calculated. The RSD represents the absolute coefficient of variation where a high value represents large changes. This highlights residues that have variable SID scores as a result of conformational changes rather than those having high SID scores because of their location in rigid condensed regions with low conformational flexibility. This expression of variation in the score of each cluster will guide the identification of regions of potential motion and adjustment in the

protein corresponding to residues that experience large changes in their topological situation, such as widening or narrowing of the interface. For widening, this would be represented by some residues leaving the 7Å cluster thereby reducing the score, and for narrowing, by new residues entering the cluster and increase its score.

The RSD of SID scores (HL, DIFF and GG) for each simulated state was calculated and shown in figure 3.25 and table 3.2 as before. The consensus SID scores were then converted into a colour code and structurally mapped onto the backbone of their corresponding average structure (figure 3.26).



A



B

Figure 3.25: Column chart of the RSD of SID scores of each residue position of all the extracted frames from the *JNK1-apo* trajectory (A) and the *JNK1-allo* trajectory (B). HL scores are coloured blue, DIFF scores are coloured red and GG scores are coloured green.

Table 3.2: Statistical descriptors of the RSD of JNK1 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.

Statistical descriptors	SID scores of <i>JNK1-apo</i> state			SID scores of <i>JNK1-allo</i> state		
	HL	DIFF	GG	HL	DIFF	GG
Upper quartile (UQ)	0.776	0.3552	1.152	0.631	0.309	0.937
Lower quartile (LQ)	0.033	0.0713	0.032	0.034	0.074	0.029
Median (MQ)	0.219	0.1665	0.288	0.170	0.151	0.252
Inter-quartile distance (IQD)	0.743	0.2838	1.120	0.597	0.235	0.908
Average (AVG)	0.530	0.2749	0.840	0.457	0.267	0.754
Minimum value (MIN)	0	0	0	0	0	0
Maximum value (MAX)	3.877	2.3703	8.043	3.983	2.552	7.955

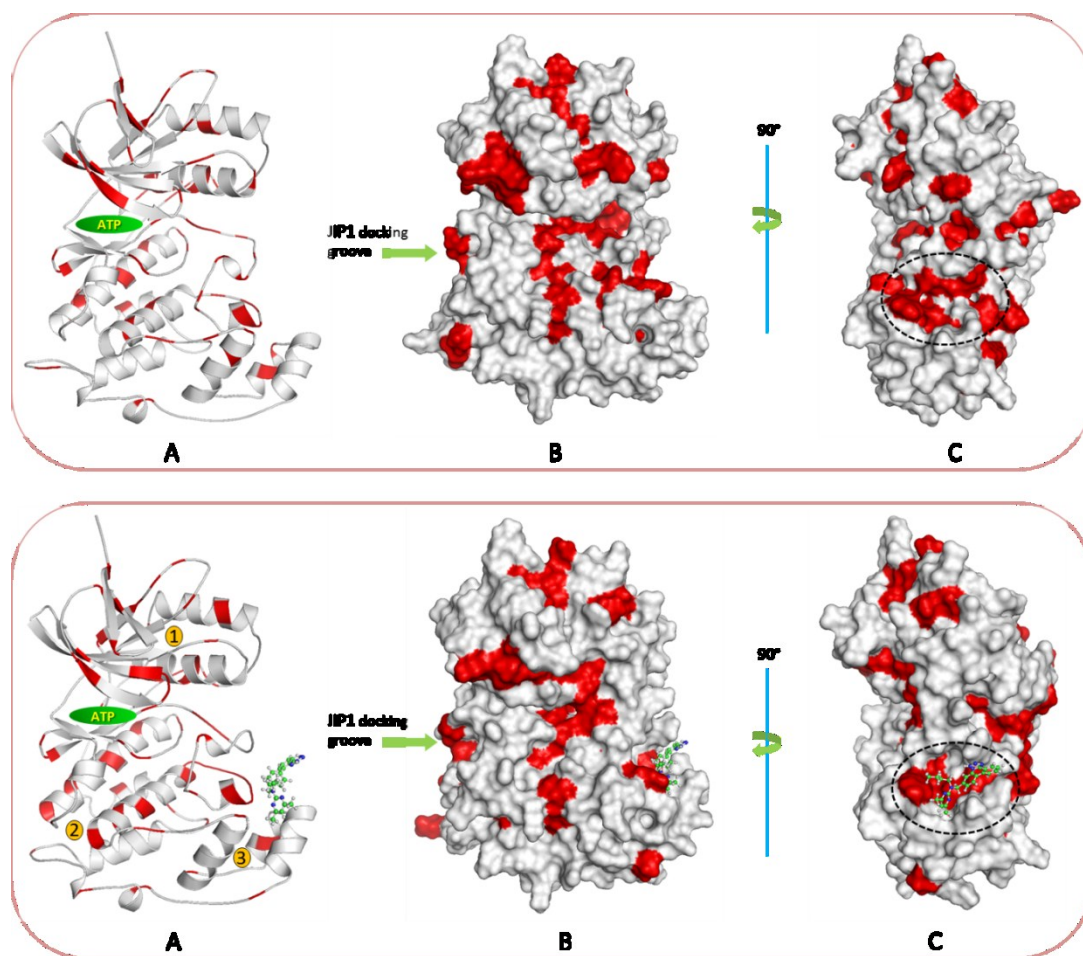


Figure 3.26: SID analysis of all frames extracted from the trajectories of the two simulated states of JNK1. The upper panel is for *JNK1-apo* and the lower one is for *JNK1-allo*. The overall consensus score of the RSD of SID scores (RSD-HL, RSD-DIFF, and RSD-GG) are colour coded and structurally mapped onto the backbone of

the corresponding average structure where red represents a high consensus score and white low scores. (A) Cartoon representation of the protein backbone. The ATP binding site is highlighted by a green ellipsoid. (B) Surface representation of the protein. JIP1 binding groove is highlighted by a green arrow. (C) Rotation about the y-axis of (B) to show the allosteric binding site which is highlighted by a dashed black ellipsoid. The allosteric inhibitor is shown in ball and stick. The numbered yellow circles highlight the regions of the protein that experienced a change in the variation of their SID score upon binding the allosteric inhibitor.

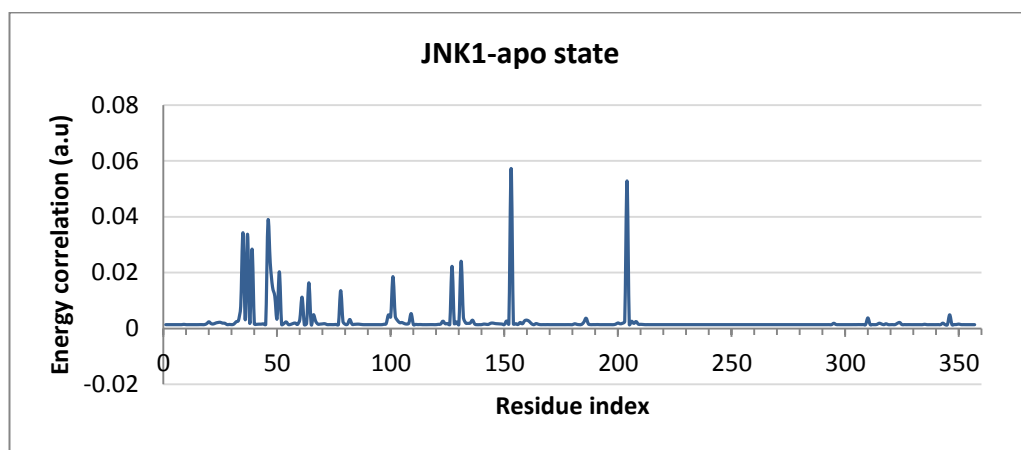
In figure 3.26 red coloured regions do not represent regions of high SID score as was the case in the single structure (figure 3.24), rather the magnitude of variation in the values of SID scores for residues involved in forming multi-way interfaces. Interestingly, among the residues that showed considerable variation in their SID scores as the system evolved with time are those implicated in forming the binding sites of the enzyme and where the allosteric inhibitor bound to the enzyme. Some of them (highlighted by the yellow numbered circles) no longer have large variation in their scores. The results of the comparative SID analysis taken with the single structure analysis indicate that both of the ATP binding site and the JIP1 docking groove are vulnerable to intermolecular recognition events and can be regulated by ligand binding at distant sites.

3.1.5 Energy correlations

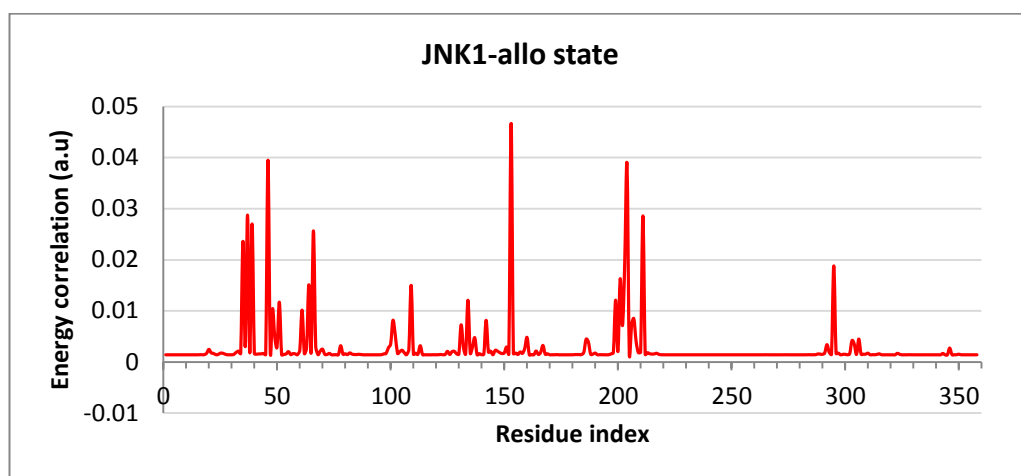
This method is based on the work of Bahar and Erman as described in the Materials and Methods section and is based on the elastic network model (ENM) and the Gaussian network model (GNM) which reflect the statistical thermodynamics of native proteins. Using this thermostistical formalism they related signal propagation in proteins and fluctuations in energy to the fluctuations in residue positions within the 3D structure of proteins, which in turn is related to the connectivity matrix of protein residues. Ultimately, this links the energy fluctuations and signal propagation pathways in proteins with their 3D architecture. The idea of this approach originates from the fact that native proteins within the cell are in a continuous exchange of energy with their surroundings which affects their functions,

and that protein 3D structures are engineered to handle this energy exchange in a productive way. The results of Erman's work showed that this exchange is neither spatially isotropic nor random: there are certain residues on a protein surface called "energy gates" that are more responsive to energy fluctuations than others and the energy that is taken from the surroundings by these residues does not diffuse randomly within the protein, but rather follows specific paths which are called "interaction pathways". These findings establish a connection between energy fluctuation (which can result from ligand binding) and protein topology (its 3D architecture), and this method was applied (see materials and methods) to analyse energy interactions of the minimised average structures of the two simulated states of JNK1. In all calculations, the cut-off distance was set to 7.0 Å. The eigenvalues that contribute most to the characterization of studied systems are those at the large end of the gamma matrix, therefore the largest seven eigenvalues of the gamma matrix were retained in calculating Γ^{-1} (the pseudo-inverse *Kirchhoff matrix*). Different scenarios have been tried to select the optimum number of eigenvalues to describe the system in hand, and using the largest seven eigenvalues seemed to show good balance between missing some features when using fewer eigenvalues; and adding to the redundancy of results when using more. Furthermore, different representations of the studied systems were tried ranging from using $C\alpha$, the backbone atoms, to the entire heavy atoms of the protein; the $C\alpha$ representation appeared to give a meaningful picture of the energetic coupling between protein residues.

A matrix of the energetic interaction of each residue in the protein with all other residues was calculated. To find those residues with high energetic correlation, the average value for the energetic interactions of each residue were calculated and plotted versus residue index (figure 3.27). The energy correlations were calculated based on the largest seven eigenvalues of the gamma matrix and using $C\alpha$ to represent protein residues. The appearance and disappearance of new peaks in the *JNK1-allo* state reflects the disruptive effects of the allosteric inhibitor on the interaction pathway as it binds the native state of the enzyme.



A



B

Figure 3.27: The correlation peaks that define the interaction pathways in JNK1. (A) The *JNK1-apo* state. (B) The *JNK1-allo* state. The correlations are given in arbitrary units.

To visualize the energetic interaction pathways defined by the residues corresponding to the peaks in figure 3.27, they were structurally mapped onto their corresponding average structure as shown in figure 3.28.

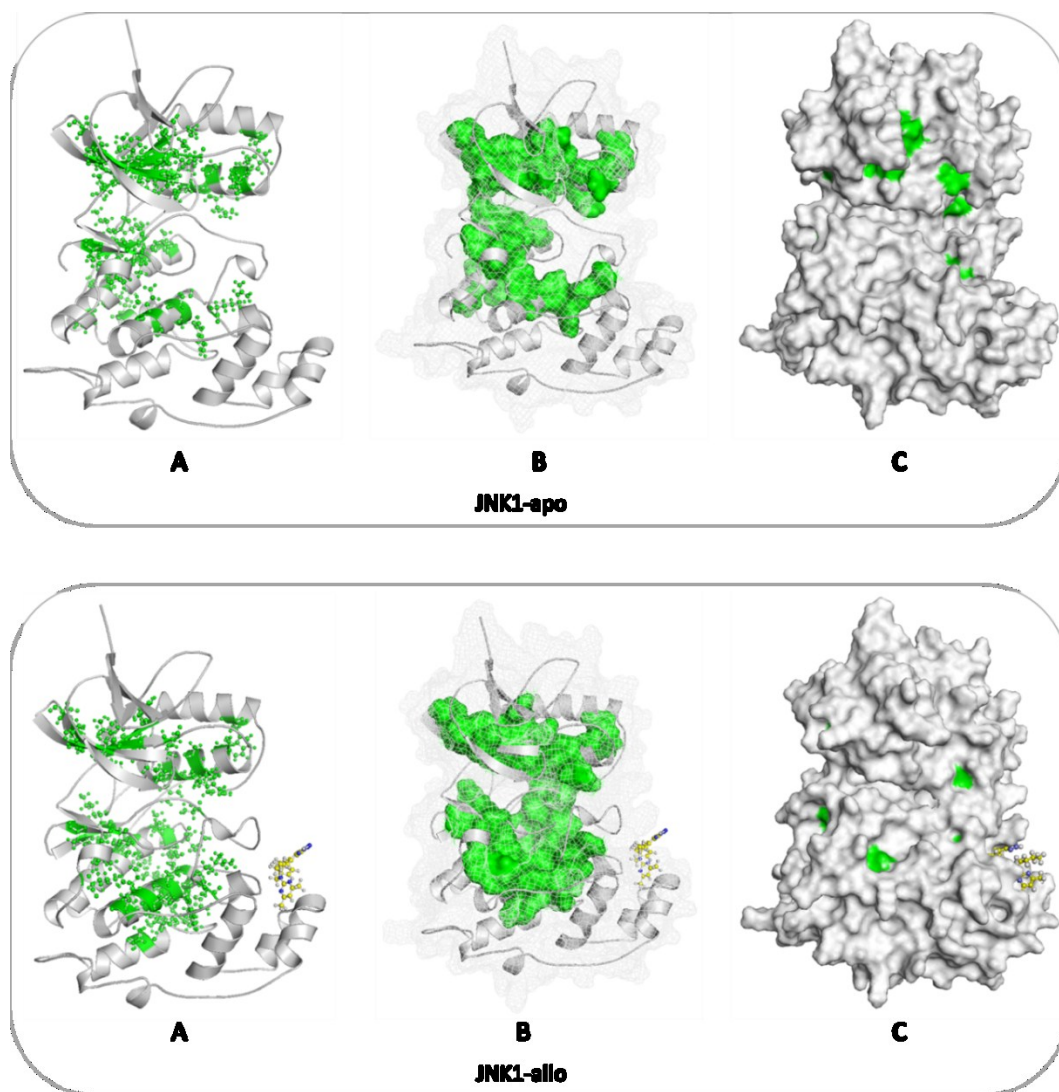


Figure 3.28: Structural mapping of the residues that define the interaction pathways in *JNK1-apo* (top) and *JNK1-allo* (bottom) states. (A) Residues forming the interaction pathway are represented by green balls and sticks, and other residues in the protein are coloured grey. (B) Surface representation of the residues forming the interaction pathway within a mesh surface of the entire protein. (C) Full surface representation of all residues in the protein showing the energy gates on the surface. The allosteric inhibitor is represented as balls and sticks and its carbon atoms are coloured yellow.

Figure 3.28 shows that residues of the interaction pathway of the *JNK1-apo* state actually form a continuous path that links the ATP-binding site (including the gate keeper Met102 and the catalytic Lys49) with the JIP1 docking groove and reaches

the boundaries of the allosteric binding site. This suggests that perturbing either site via ligand binding will be transmitted to the other. Such effects of ligand binding are obvious in the *JNK1-allo* state where the interaction pathway and the energy gates are different from that in the apo state. The binding of the allosteric inhibitor in this case can be thought of, from an energetic point of view, as an energetic perturbation of the native network of interactions in the protein that alters the energy gates and the interaction pathways crucial to functional substrate binding, which affects the function of the protein.

3.1.6 Summary of the overall results of JNK1

All of the computational analyses are assembled in figure 3.29.

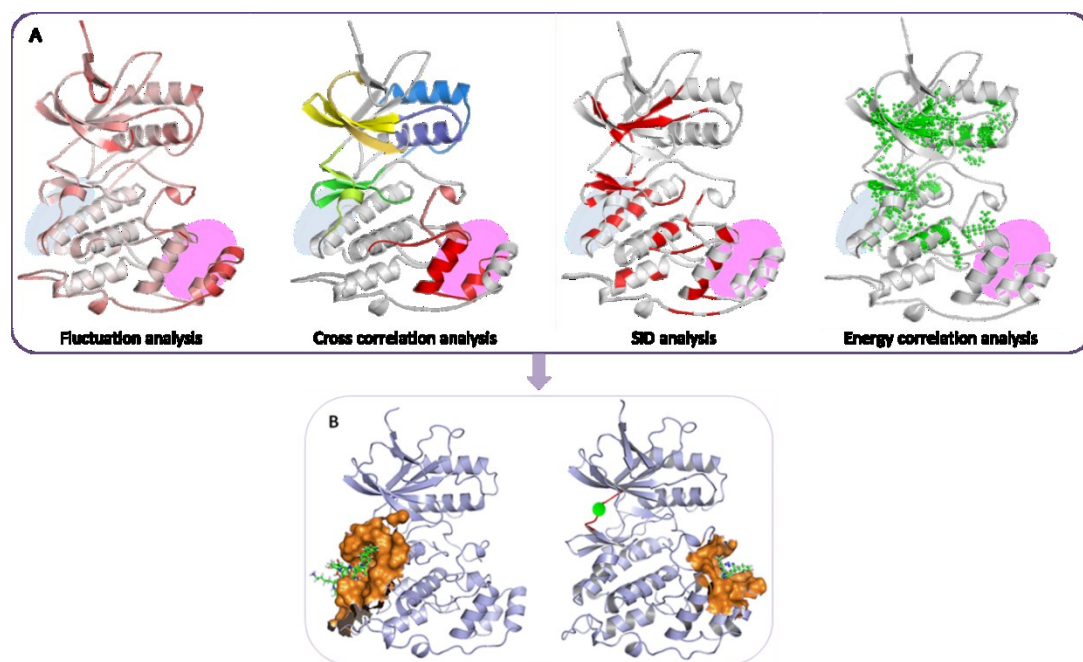


Figure 3.29: Summary of main computational results for JNK1. (A) Assembly of the results of different types of computational analysis implemented to study JNK1. The allosteric site is highlighted by pinkish shadows; and the JIP1 binding site is highlighted by bluish shadows. (B) The experimentally identified allosteric sites in JNK1. The allosteric inhibitors (ball and stick representations) occupy their corresponding binding sites (brown surfaces).

This comprehensive picture of the overall results shows that our computational predictions of putative allosteric sites were in excellent agreement with the experimentally (x-ray crystallography) identified sites. As figure 3.29 shows, in all of the computational methods the allosteric site (highlighted by pinkish shadow) and the JIP1 binding site (highlighted by bluish shadow) were shown to be of high allosteric potential.

3.2 Cyclin-Dependent Kinase 2 (CDK2)

3.2.1 The role of CDK2

Cyclin-dependent kinases (CDKs) are serine/threonine protein kinases that are critically dependent on their partner proteins, the cyclins, for their activity and substrate specificity (figure 3.30). CDKs have a key regulatory role in cell cycle progression; for example the CDK2-cyclin-A complex mainly controls the G1 to S-phase checkpoint [108, 171, 172].

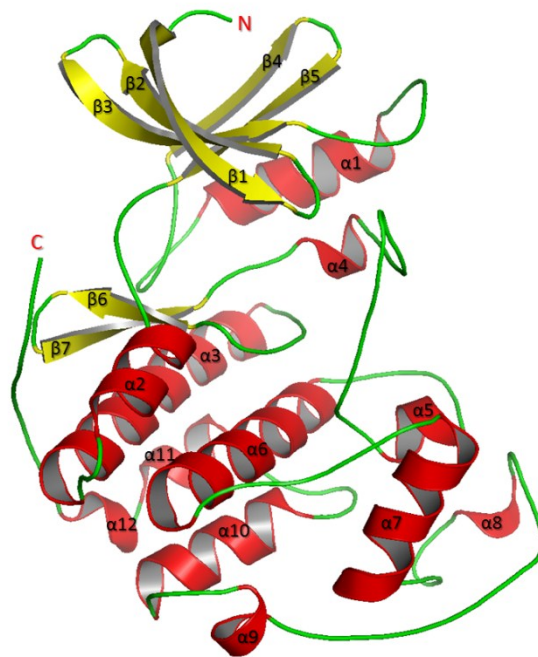


Figure 3.30: Cartoon depiction of the backbone atoms of the catalytic domain of the CDK2. The α -helices are in red; β -sheets are in yellow; and loops and turns are in green.

CDK2 is inactive as monomer because of two structural constraints. Firstly, the orientation of the side chains of some residues in the ATP binding (such as Lys33 and Asp145) inhibits the phosphotransfer. Secondly, the T-loop blocks the substrate-binding site (the cleft between the N and C-terminal domains) [172]. The first step in CDK2 activation is cyclin binding which is not only necessary for activity but also for selectivity. The main structural change in CDK2 associated with cyclin-A

binding is the migration of the PSTAIRE helix (the α C-helix, or α 1 in figure 3.30), which is highly conserved among CDKs (figure 3.31A). This rearrangement of the α C-helix promotes reconfiguration of the ATP binding site through reorientation of the residues that interact with the ATP phosphate. Also, cyclin-A binding causes α L12-helix (α 4 at the beginning of the T-loop) to change its conformation into a β -sheet which increases the flexibility of the T-loop allowing it to move away from the active site (figure 3.31B). Subsequently, phosphorylation of Thr160 in the T-loop by the cyclin-dependant kinase activating kinase 1 (CAK1) causes further reorientation and stabilization of the T-loop in its active conformation and generates the fully active complex [171-174] (figure 3.32).

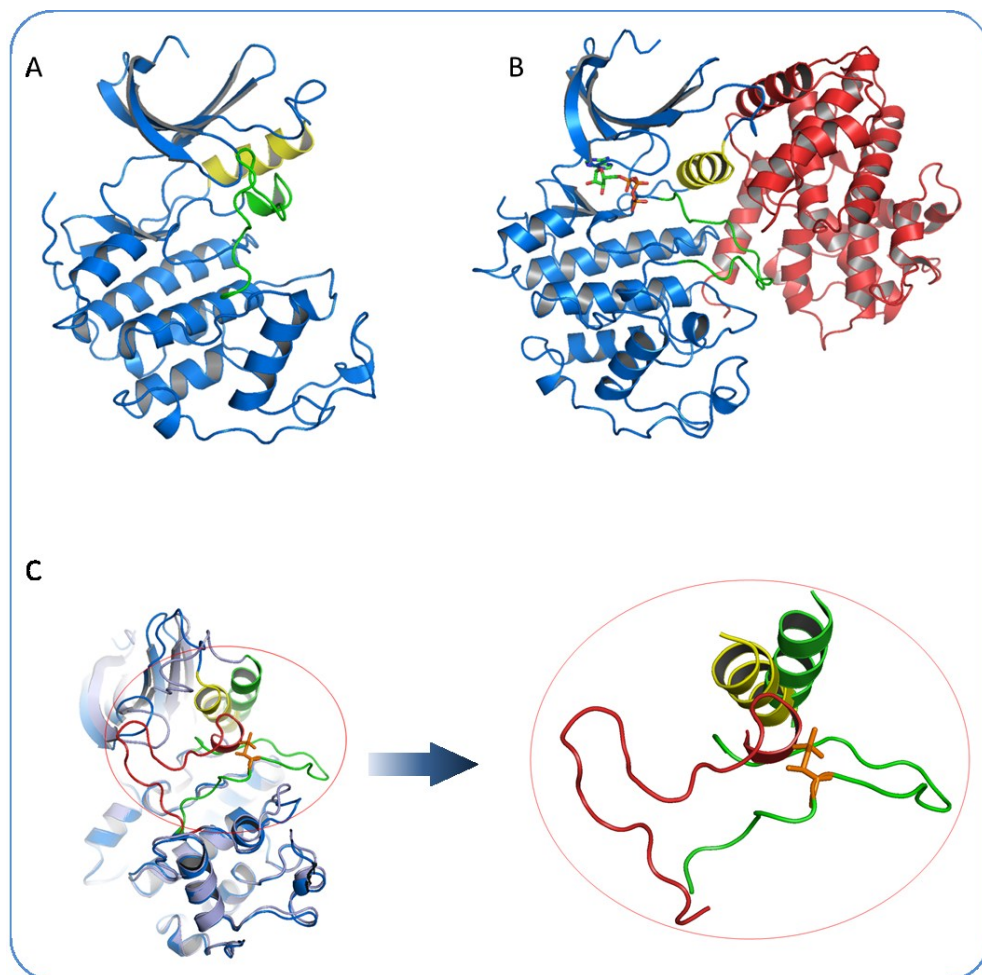


Figure 3.31: The structural changes associated with cyclin-A binding. (A) CDK2 without cyclin (inactive). α C-helix (yellow); the α L12-helix at the very beginning of the T-loop (green) (PDB code 1HCL [150]). (B) Phospho-CDK2 in complex with

cyclin-A (fully active), (PDB code 2CCI [171]). (C) Superimposition of the two crystal structures (1HCL, green α C-helix and red T-loop; and 2CCI yellow α C-helix and green T-loop with phospho-threonine in orange sticks) showing the structural changes in the α C-helix and the α L12-helix associated with cyclin-A binding. Note the migration of the α C-helix and the melting of the α L12-helix into a β -sheet.

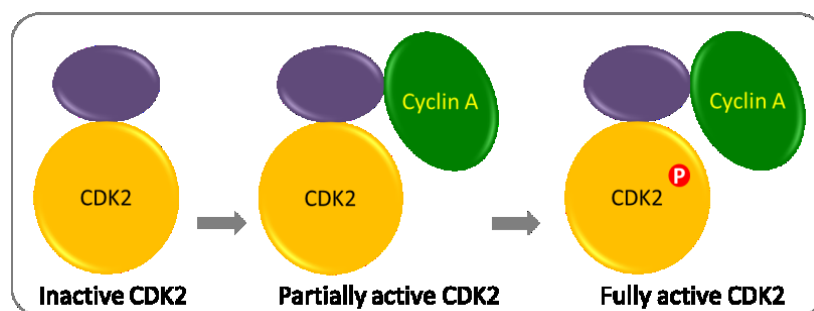


Figure 3.32: The activation sequence of CDK2. The initial step is cyclin-A binding, then full activation is achieved via phosphorylation of Thr160 in the T-loop.

There are more than 140 crystal structures of CDK2 in complex with small molecule inhibitors deposited in the protein data bank, and most of these inhibitors are competitive ATP inhibitors [108]. Recently, an allosteric site distant from the ATP binding site has been identified in CDK2 by x-ray crystallography: two 8-anilino-1-naphthalene sulfonate (ANS) molecules (figure 3.33) bound adjacent to one another between the α C-helix and β 3-5. ANS binding provoked large conformational changes in the α C-helix (figure 3.34) rendering it incompatible for cyclin-A association [108].

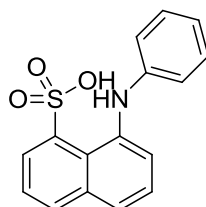


Figure 3.33: The chemical structure of ANS that binds to the allosteric site of CDK2 [108].

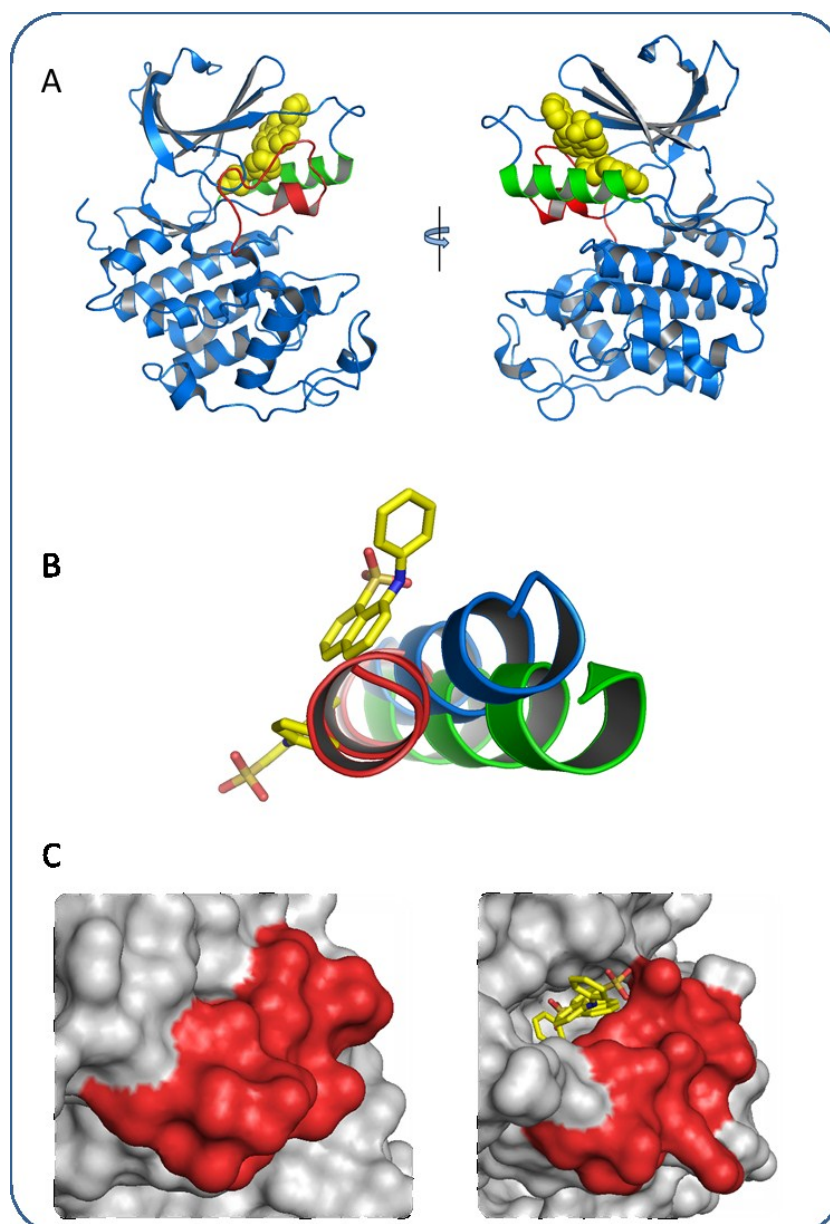


Figure 3.34: The allosteric binding site in CDK2 and the conformational changes in the α C-helix induced by ANS binding. (A) Front and back view of the allosteric site (α C-helix, β -sheets 3-5, and the DFG motif of the T-loop). ANS are in yellow; α C-helix is green; and T-loop is red (PDB code 3PXF [108]). (B) Superimposition reveals the conformational changes in α C-helix induced by ANS in the CDK2-ANS complex PDB code 3PXF (green) and in free CDK2 PDB code 1HCL (blue) as compared to the CDK2-cyclin A complex PDB code 2CCI (red). (C) Surface representations of the α C-helix (red) for free CDK2 (left) and the CDK2-ANS complex (right) shows partial opening of the ANS pocket toward solvent.

3.2.2 System stability and conformational flexibility

Two molecular dynamics simulations of CDK2 were performed to compare the dynamics of different states of this kinase. The two studied states were the nucleotide-free state (*CDK2-apo*); and the unphosphorylated ATP-bound state (*CDK2-ATP*).

3.2.2.1 System stability

Root mean square deviation:

Similar to JNK1, to explore the simulation stability of each of the studied states, mass weighted RMSD values of the backbone atoms from their starting structure were calculated and plotted versus time in figure 3.35. Both systems equilibrated fairly well with an average RMSD of 1.63Å and 1.87Å corresponding to *CDK2-apo* and *CDK2-ATP* respectively. Both of the trajectories reached a stable and equilibrated dynamical state relatively quickly in around 1 ns. As figure 3.35 shows, the RMSD values of *CDK2-ATP* was higher than those of *CDK2-apo* up to the 39th ns of simulation time. However, it is interesting to see that the RMSD values of *CDK2-apo* were increasing gradually up to the 39th ns where they coincide with those of *CDK2-ATP*, which suggests that the apo system has encountered more dynamical and conformational changes during this period.

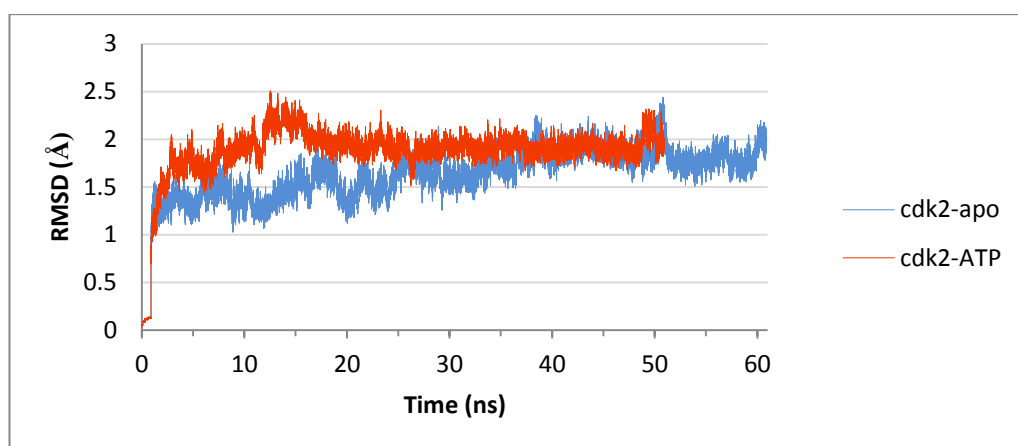
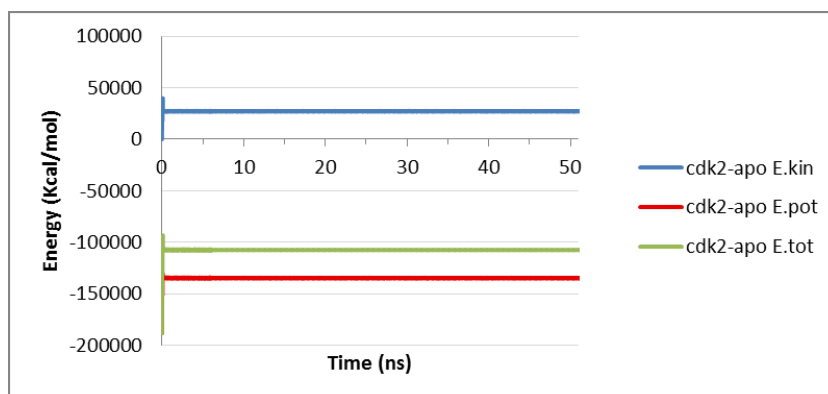


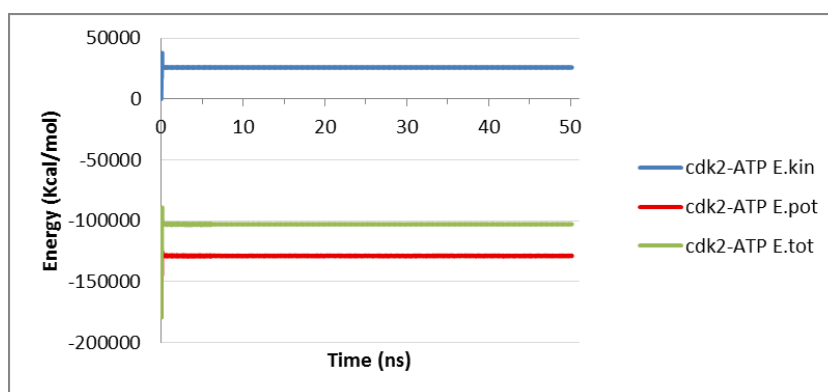
Figure 3.35: The RMS deviation of the backbone atoms of *CDK2-apo* (blue line) and *CDK2-ATP* (red line) from their minimised starting structures.

Energy conservation and temperature stability:

Figure 3.36 shows energy plots versus time and figure 3.37 shows the temperature plots versus time. The two sets of plots show that the energy components were very well conserved and the temperature was very stable, which means that the two systems were stable throughout the simulation time.

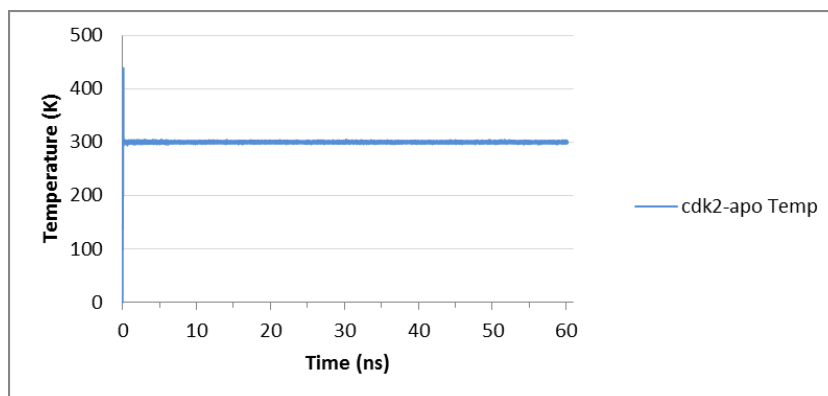


(A)

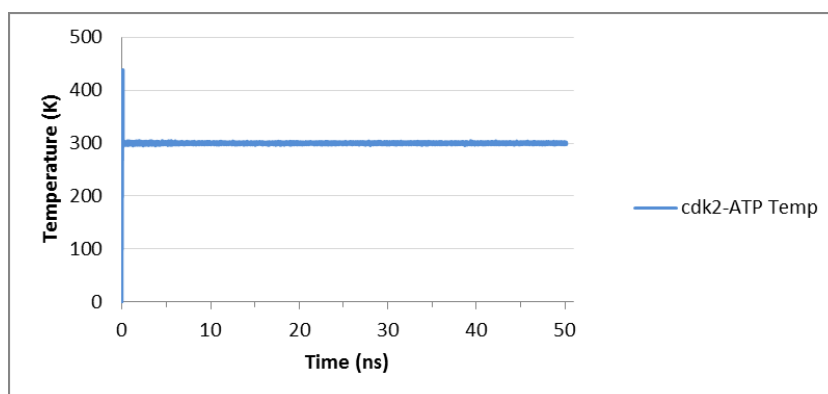


(B)

Figure 3.36: Summary of the energy changes for the two CDK2 states versus time. (A) Corresponds to the *CDK2-*apo** state; and (B) to *CDK2-ATP* state. Blue lines correspond to the total kinetic energy; red lines correspond to the total potential energy and the green lines correspond to the total energy.



(A)

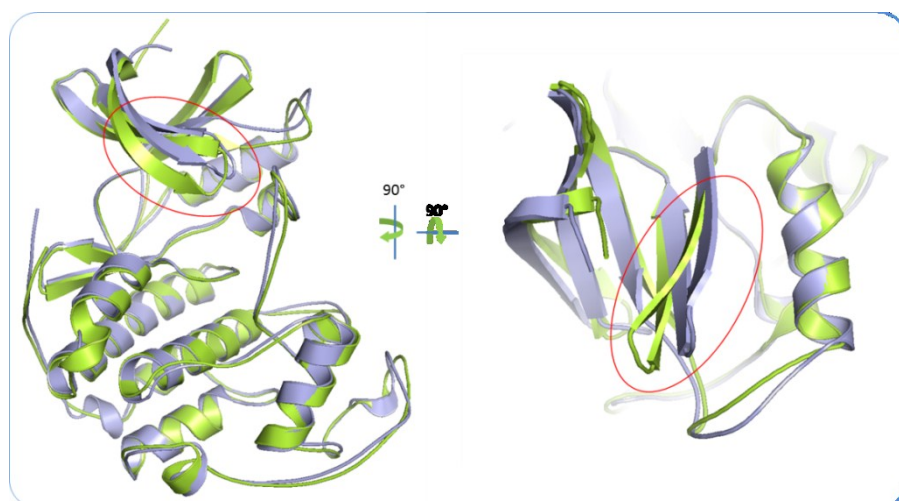


(B)

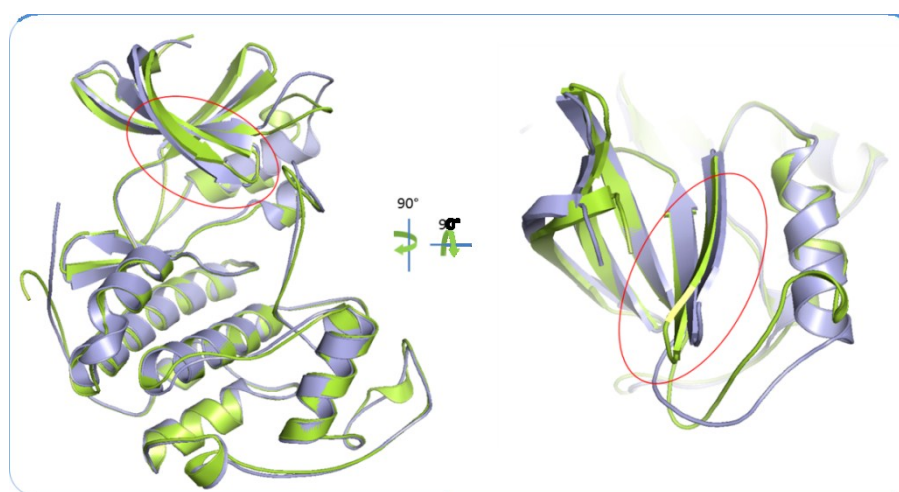
Figure 3.37: Summary of temperature changes for the two CDK2 states plotted versus time. (A) Corresponds to the *CDK2-apo* state and (B) to *CDK2-ATP*.

Average structures:

The average structures from the equilibrated parts (last 44 ns of *CDK2-apo* state and the last 54 ns of the *CDK2-ATP* state) of the two trajectories were calculated for each system and compared with the corresponding starting crystal structures (figure 3.38 and 3.39).



(A)



(B)

Figure 3.38: Superimposition of the minimised average and the starting structure of the *CDK2-apo* system (A) and the *CDK2-ATP* (B). Red ellipsoids highlight areas of major conformational changes. The starting structure is in lightblue and the average one is in green.

Figure 3.38 shows that in the apo structures there is a clear conformational change affecting the N-terminal domain of the enzyme in which the ATP binding site became tighter due to the movement of $\beta 1$ -(G-loop)- $\beta 2$ segment, whereas the C-terminal domain is almost unchanged. Rotation reveals that there is an increase in the distance between the αC -helix and the nearby $\beta 4$ and $\beta 5$ as a result of the movement of those two β -sheets. Surface representation of the enzyme in figure 3.39 clearly

shows the formation of a groove between those secondary structures (the α C-helix and the nearby β 4-5) which shows a conformational ensemble that could represent the initial formation of a binding groove for binding the small molecule allosteric effectors. Given the location of the allosteric inhibitor of CDK2, the movement and flexibility of these sheets could be used by small molecules to gain access to their final binding site. In the ATP bound structure none of these changes are evident and suggests that the bound nucleotide has prevented the movement of the β 1-(G-loop)- β 2 segment through a deformation of the last turn of the α C-helix as shown in figure 3.38. The RMSD of the backbone atoms of the minimised average structures of *CDK2-apo* and *CDK2-ATP* relative to their starting structures were 0.875 Å and 0.744 Å respectively. Although the overall RMSD values were not high, localised conformational changes were still noticeable.

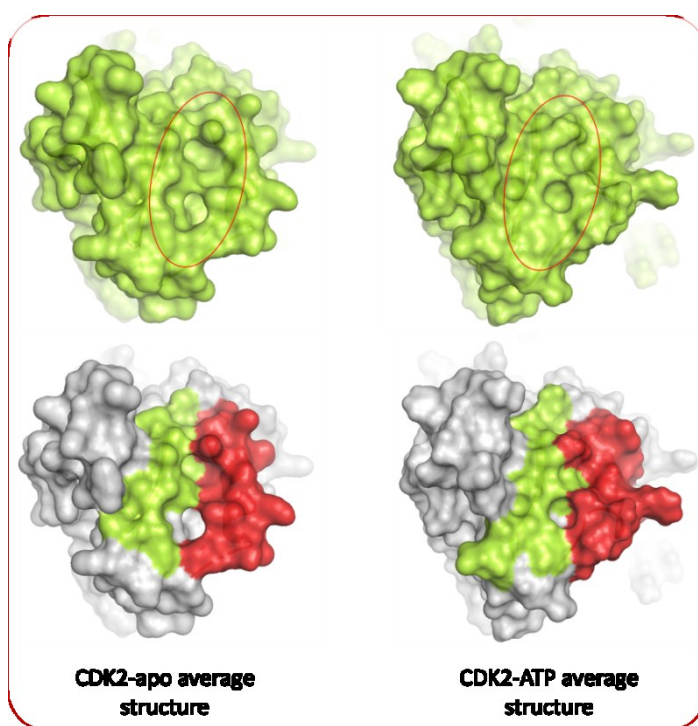


Figure 3.39: Surface representation of the rotated N-domain in figure 3.38. The red ellipsoid highlights the formation of the groove between the α C-helix and the nearby β 4 and β 5 in the *CDK2-apo* state which are coloured in the lower panel in red and green respectively. Left panel is for the *CDK2-apo* system and the right one is for *CDK2-ATP*.

3.2.2.2 Conformational flexibility

The conformational variability of the CDK2 was investigated by calculating and comparing the residual fluctuation for each of the two trajectories. Figure 3.40A shows the RMSF fitted plot of the residual fluctuation versus residue index for the two states of CDK2, where the most flexible regions in the protein have been highlighted by coloured bars underneath the plot and structurally mapped onto the protein backbone in B using the same colour code. The average residual fluctuation values of *CDK2-apo* and *CDK2-ATP* states were 1.17 Å and 1.05 Å respectively.

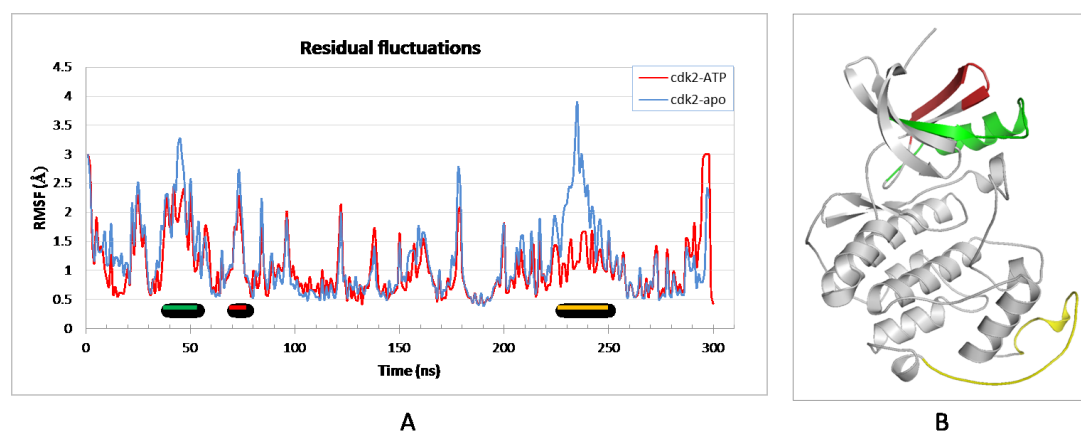


Figure 3.40: Residual fluctuations of CDK2. (A) RMSF versus residue index plot for *CDK2-apo* (blue) and *CDK2-ATP* (red). Regions of high flexibility are highlighted by coloured bars underneath the plot. The green bar corresponds to the α C-helix (residues 31-60), the red bar corresponds to β 4 and β 5 (residues 66-76), and the yellow bar correspond to the loops and α -helices between α 7- α 10 in the C-domain (residues 223-247). (B) Structure of CDK2 with the most flexible regions highlighted using the same colour code as in (A).

The effect of ATP binding on the residual flexibility of the enzyme is presented in figure 3.41. The RMSF values of the *CDK2-apo* state were subtracted from those of the *CDK2-ATP* state, and regions that have experienced the largest change are highlighted by a coloured bar underneath the plot and structurally mapped onto the backbone atoms of the protein in B.

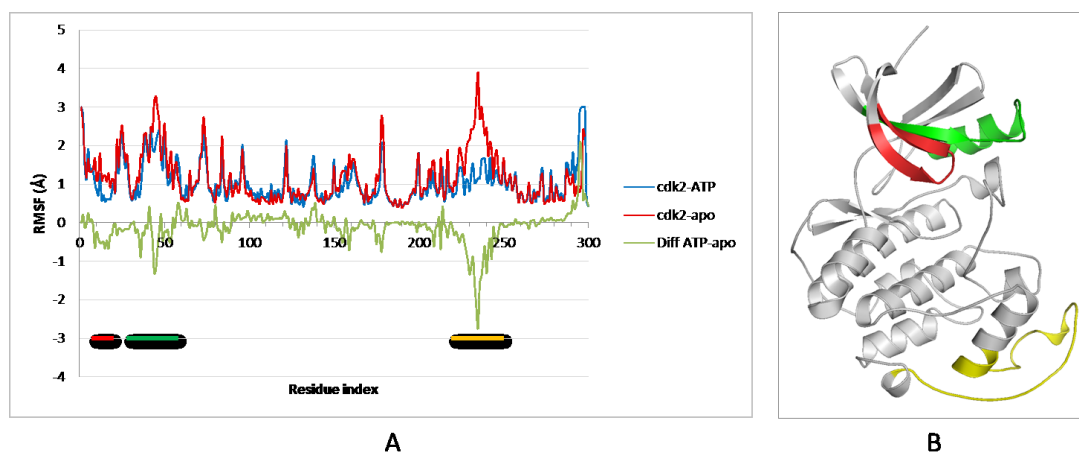
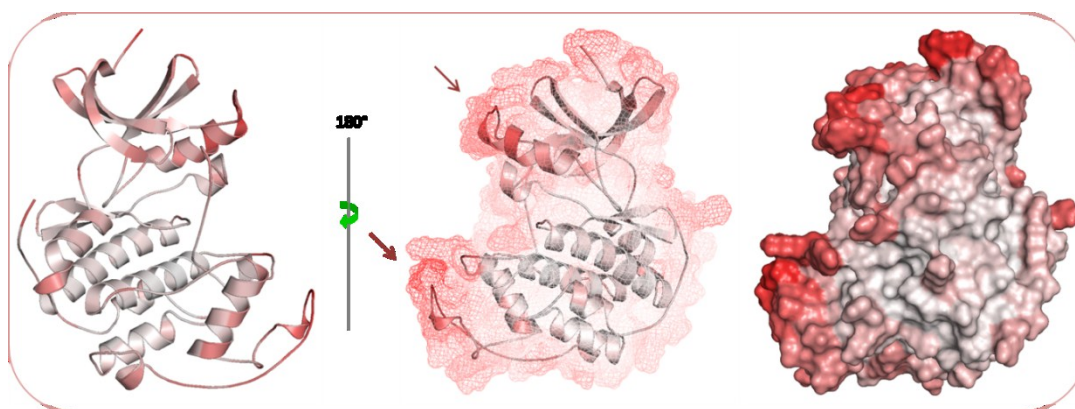


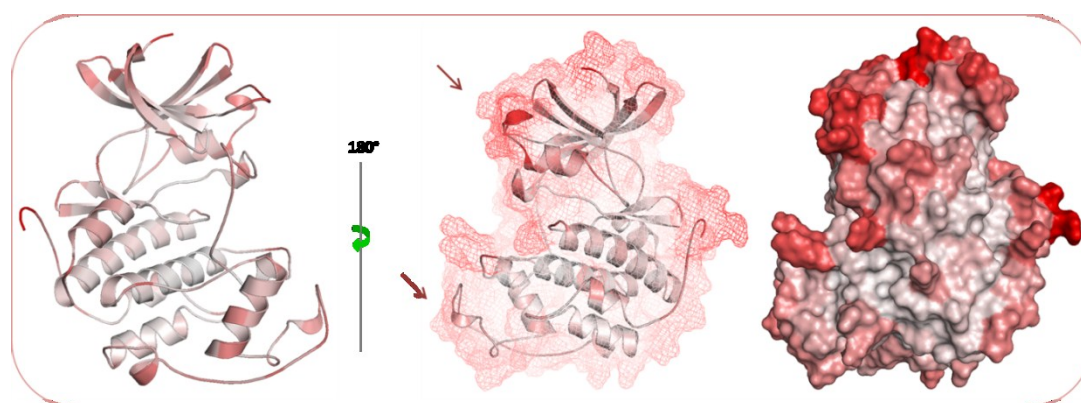
Figure 3.41: The effect of ATP binding on enzyme flexibility. (A) Subtraction of the RMSF values of *CDK2-apo* state from those of the *CDK2-ATP* state. The coloured bars highlight the regions of largest changes. (B) Structural mapping of the most affected regions upon ATP binding using the same colour code of the bars in part-A.

The largest change in protein flexibility upon ATP binding was in the loops and α -helices between $\alpha 7$ - $\alpha 10$ in the C-domain. Also, there were interesting changes in the N-domain that affected the αC -helix, $\beta 3$ (which includes Lys33) and the $\beta 1$ -(G-loop)- $\beta 2$ segment. The latter changes are expected because these segments are among the key features of the ATP binding site and are crucial for optimal binding and positioning of ATP via establishing a network of important hydrogen bonds so that the terminal phosphate group will be aligned for transfer to the substrate serine or threonine residue. For example, Lys33 and the αC -helix are responsible for the positioning of ATP to achieve productive binding; and the G-loop moulds down on ATP to accomplish this productive binding. As a result of these interactions with ATP, the flexibility of these protein segments is expected to be decreased.

To have a comprehensive visual picture of the residual fluctuation in both of the simulated states of CDK2, the RMSF values of both simulations were converted into colour codes and structurally mapped onto the backbone atoms of their corresponding minimised average structure (figure 3.42).



(A)



(B)

Figure 3.42: Colour coded representation of the residual fluctuation values of the two simulated states of CDK2. (A) Represents the fluctuations of the *CDK2-apo* state; and (B) is for the *CDK2-ATP* system. The highest fluctuating regions are coloured red and the lowest are coloured white. For each state there are cartoon, mesh surface and solid surface representations. The red arrows highlight regions with high flexibility.

Analysis of the residual fluctuation of the different states of CDK2 in figures 3.40-42 reveals that the regions of the enzyme with the highest fluctuation values correspond to the α C-helix, β 4 and β 5, the β 1-(G-loop)- β 2 segment in the N-domain; and the loops and α -helices between α 7- α 10 in the C-domain. It is well known that the G-loop and the α C-helix have a key role in ATP binding, but in the case of the CDK2,

the α C-helix has a special significance as it forms part of the binding pocket for the allosteric inhibitor shown in figure 3.34 [108].

3.2.2.3 Contact maps:

The contact maps of the minimised average structures of the two systems of CDK2 are presented in figure 3.43. The cutoff distance between α -carbons of residue pairs was set to 8Å. The number of contacts in the *CDK2-apo* system was 1282 of which 1228 were common and 54 were unique. The number of contacts in the *CDK2-ATP* system was 1311 of which 1228 were common leaving 83 as unique. All of the unique contacts in each system were structurally mapped in figure 3.44.

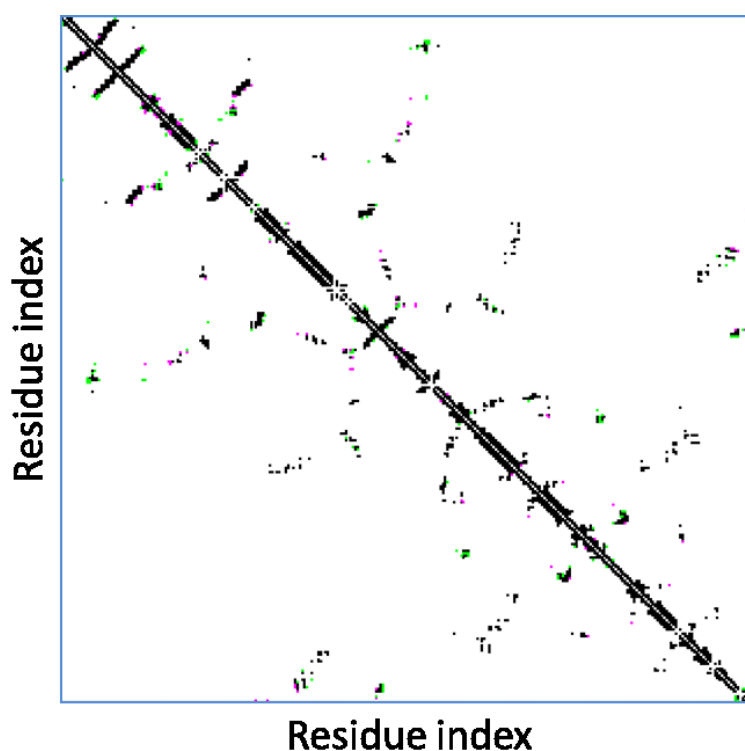


Figure 3.43: The contact maps of the two minimised average structures of CDK2. Black dots are common contacts present in both states; pink dots are unique contacts for the *CDK2-apo* state; and green dots are unique contacts for the *CDK2-ATP* state.

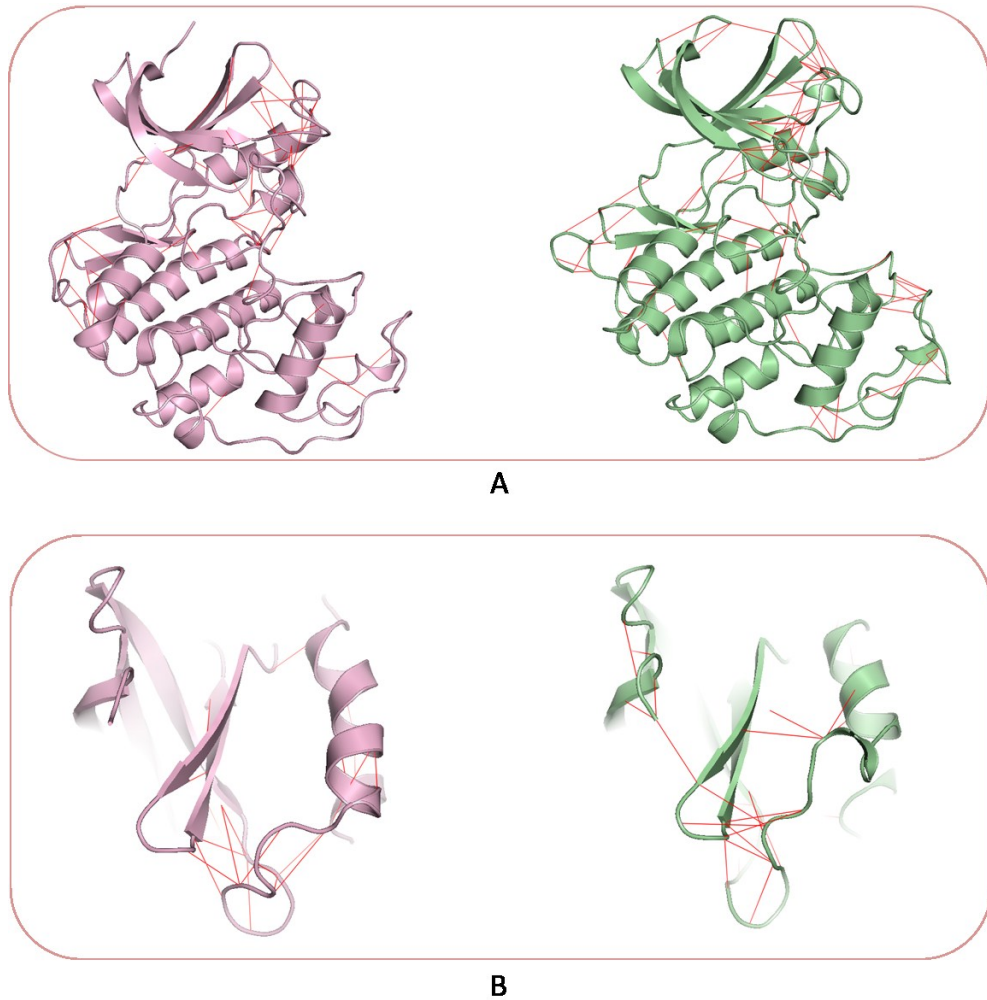


Figure 3.44: Comparison of the contact maps of the *CDK2-apo* (pink) and *CDK2-ATP* (pale green) states. (A) Shows the structural mapping of all of the unique contacts for both states (red lines). (B) A close up view of the N-domain.

Figure 3.44 shows that there are more unique contacts in the *CDK2-ATP* state. Most of those unique contacts are localized and concentrated in the N-domain particularly around and involving the α -C helix and the loops connecting β 3- α C helix and β 4- β 5 of the N-domain. The close-up view in figure 3.44B clearly shows the difference between the two states where there are no contacts between the α -C helix and the nearby β 4- β 5 in the *CDK2-apo* system, while there are four contacts in the *CDK2-ATP* state mediated by the deformed last turn of the α -C helix. This may explain the very small conformational change in this region in the *CDK2-ATP* state compared to the apo form.

3.2.3 Analysis of correlated motion

Figure 3.45 shows the heat maps of the dynamic cross-correlation matrices calculated from each simulation trajectory of the two CDK2 states. The scale of the colouring scheme is from -0.5 (negatively correlated residues) to 1.0 (positively correlated residues). The overall average of residual correlations in the *CDK2-apo* state was 0.013 and the average of the positively correlated residues was 0.150; for the *CDK2-ATP* state, the average of the overall correlations was 0.012 and 0.161 for the positively correlated residues. Residues with correlation values of more than 0.3 (which is twice the average of positive correlations) were considered as highly correlated with each other.

As figure 3.45 shows, the number and nature of correlated residues vary among the two heat maps. The total number of possible correlations in CDK2 is 44253; of which the number of highly correlated residues (having a correlation value of more than 0.3) in the *CDK2-apo* and *CDK2-ATP* states was 2281 and 2461 respectively, which represent around 10% increase in the total number of correlated residues in the *CDK2-ATP* state. This indicates that ATP binding has brought some order to the dynamical motion of the enzyme, and since ATP is the natural ligand for the enzyme, this seems to be a functional change in protein dynamics toward achieving its catalytic function.

By examining the areas of most concentrated residue-residue couplings in the heat maps of the two states of the enzyme and their structural mapping, we can see that in the apo state they are located in the region between residues 54 and 150, corresponding to a protein segment that includes the ATP binding site and major parts of the allosteric binding site in the N-domain. It seems that these residues are important to convey conformational responses between the two sites which could be related to preparing the enzyme for binding its regulatory protein (cyclin-A) in the N-domain. In the ATP bound state, correlated residues are located in the region between residues 105 and 213. These residues were more oriented towards the C-domain suggesting that the enzyme is preparing for proper binding of its substrate for catalytic function. A more focused and detailed analysis of the correlated residues in each state is presented in figures 3.46-3.48.

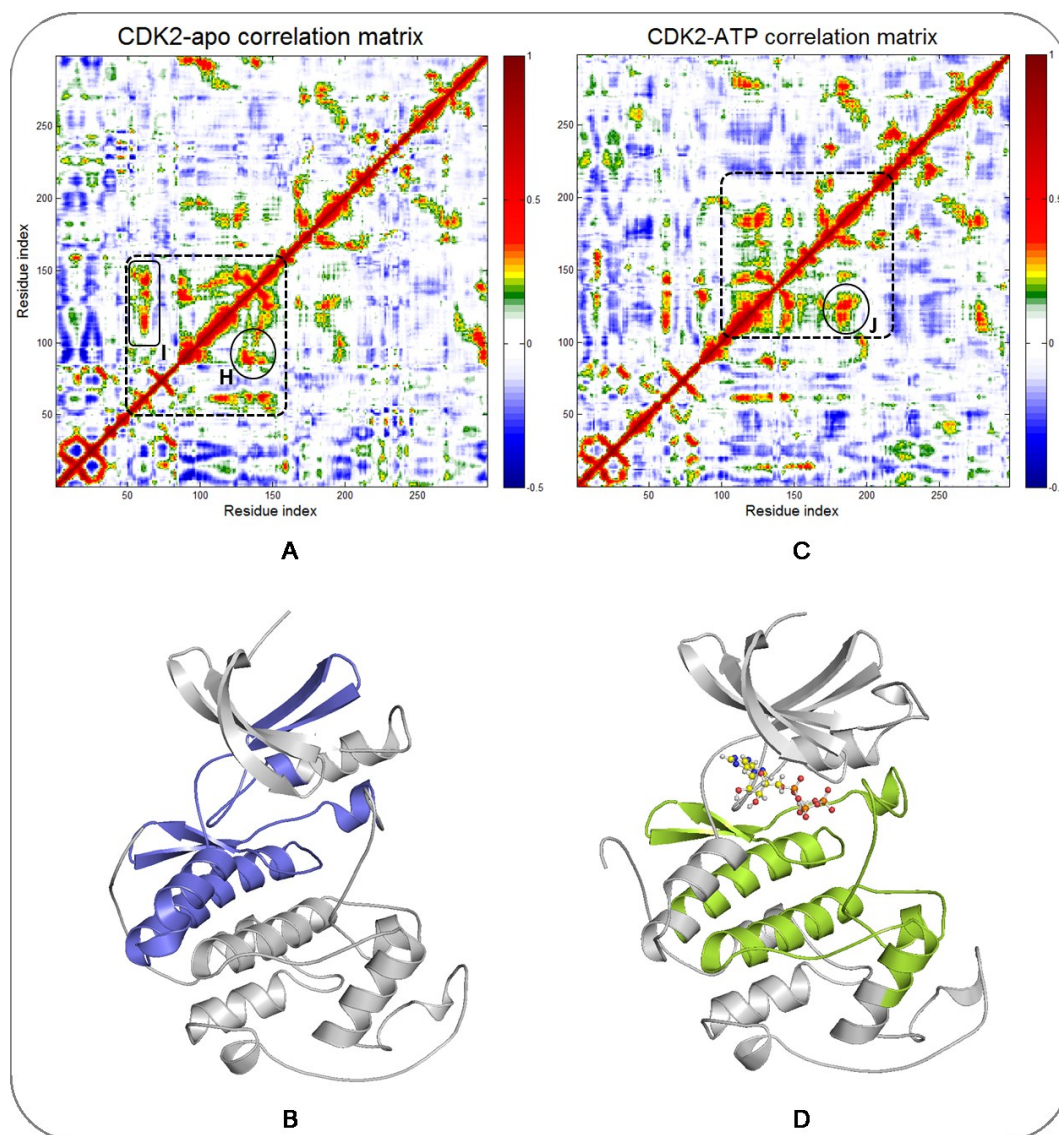


Figure 3.45: Residual correlation analysis of the two simulated states of CDK2. (A) The residual cross-correlation matrix of the backbone atoms around their average position calculated from the simulation trajectory of *CDK2-*apo** state and represented as a heat map. The most concentrated residue-residue coupling is highlighted by the dashed black rectangle and structurally mapped in (B); the off-diagonally correlated residues are highlighted by the solid black circle and rectangle H and I. (C) The heat map of the *CDK2-ATP* state. The most concentrated residue-residue coupling area is highlighted by the dashed black rectangle and structurally mapped in (D); the off-diagonally correlated residues are highlighted by the black solid circle J. In both heat maps, red pixels represent positive correlations and blue pixels are negative correlations.

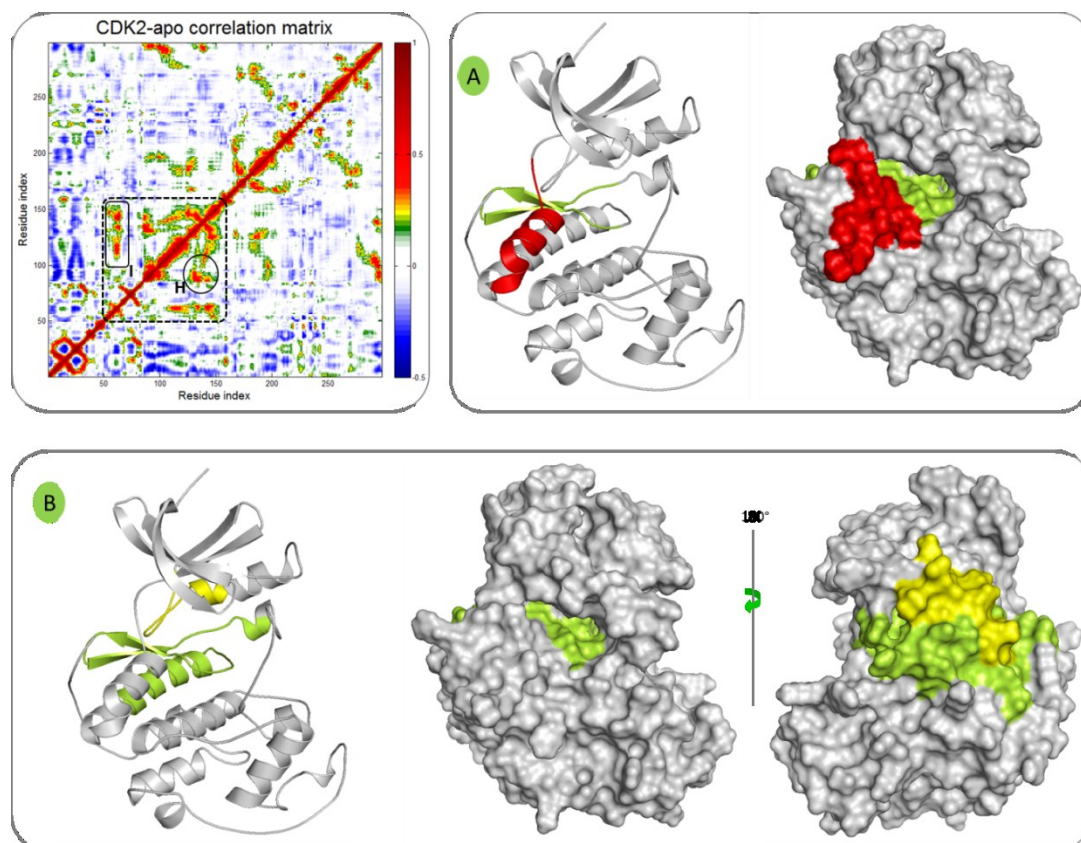


Figure 3.46: Analysis of the correlated residues in the *CDK2-apo* state. The heat map of the *CDK2-apo* state is shown in the upper left corner of the figure with the areas (clusters) of high off-diagonal residue-residue coupling are highlighted by black circle H and rectangle I. (A and B) are respectively the structural mapping of the correlated residues in cluster H (red and green) and I (green and yellow) in the heat map. Structural mapping of the clusters is depicted in cartoons and solid surfaces.

Figure 3.46 discusses the correlated residues in the *CDK2-apo* state. The heat map of this state shows that there are two main off-diagonal clusters of correlated residues, cluster H and I. Cluster H encompasses the red segment (residues 84-96) which is positively correlated with the green segment (residues 128-146); and cluster I is comprised of the yellow segment (residues 54-66) which is correlated with the green segment (residues 107-150). The two clusters shared large part of the green segment (residues 128-146) which implies that this segment might be coordinating the dynamical motion of the protein to guide and facilitate its association with cyclin-A.

It is interesting to note that cluster I contributes to a considerable portion of the protein surface that is implicated in binding cyclin-A (rotated view in figure 3.46B). Furthermore, cluster I includes three key residues, Lys56, Glu57 (in the N-terminal of α C-helix) and Arg122 (in the C-terminal of α 3) that are involved in forming a network of hydrogen bonds with cyclin-A residues Tyr185, Thr303, Asp305 and Ala307 as inferred from the fully active form of the enzyme (phospho-CDK2-cyclin-A; figure 3.47A-B). Moreover, another residue in the catalytic loop (the loop connecting β 6- α 3), Lys129, is also involved in hydrogen bonding with the phosphate-accepting Ser70 of the CDK2 substrate (figure 3.47C-D). These results suggest that ligand binding at, or near, this cluster might have the potential to allosterically affect the function of the enzyme by interfering with CDK2-cyclin-A association or the binding of the substrate.

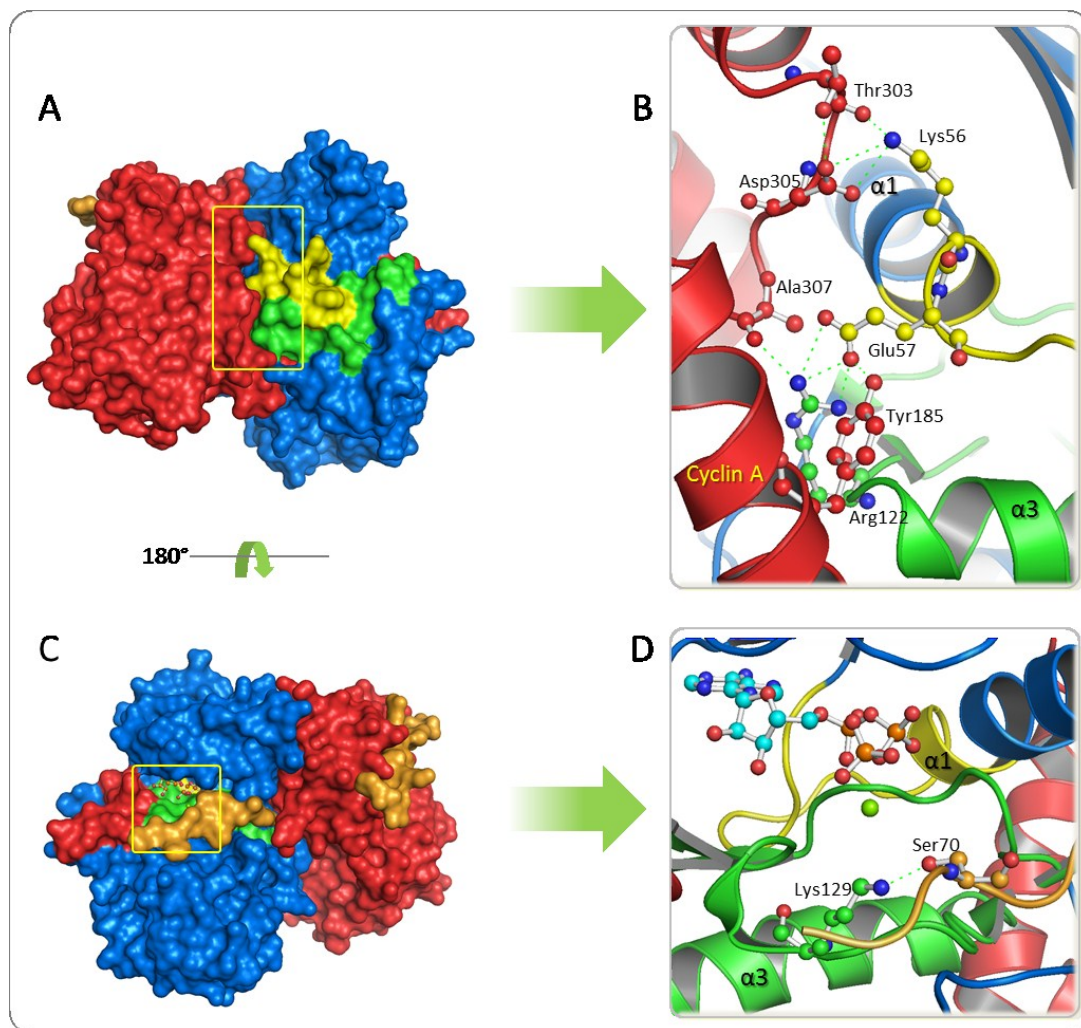


Figure 3.47: The interactions between cluster I and other components of the phospho-CDK2-cyclin-A complex. (A) Back view of the van der Waals surface of the phospho-CDK2-cyclin-A complex. Cyclin-A is in red, CDK2 is in blue, and cluster I is in yellow and green. (B) Close-up of the yellow rectangle in A showing the network of hydrogen-bonds between cluster I (yellow and green) and cyclin-A (red). (C) Front view of the complex. CDK2 substrate is shown in orange. (D) Close-up of the yellow square in C showing the interaction between Lys129 and Ser70. Residues that are involved in all of these interactions are depicted in ball and stick. Hydrogen bonds are depicted as dashed green lines.

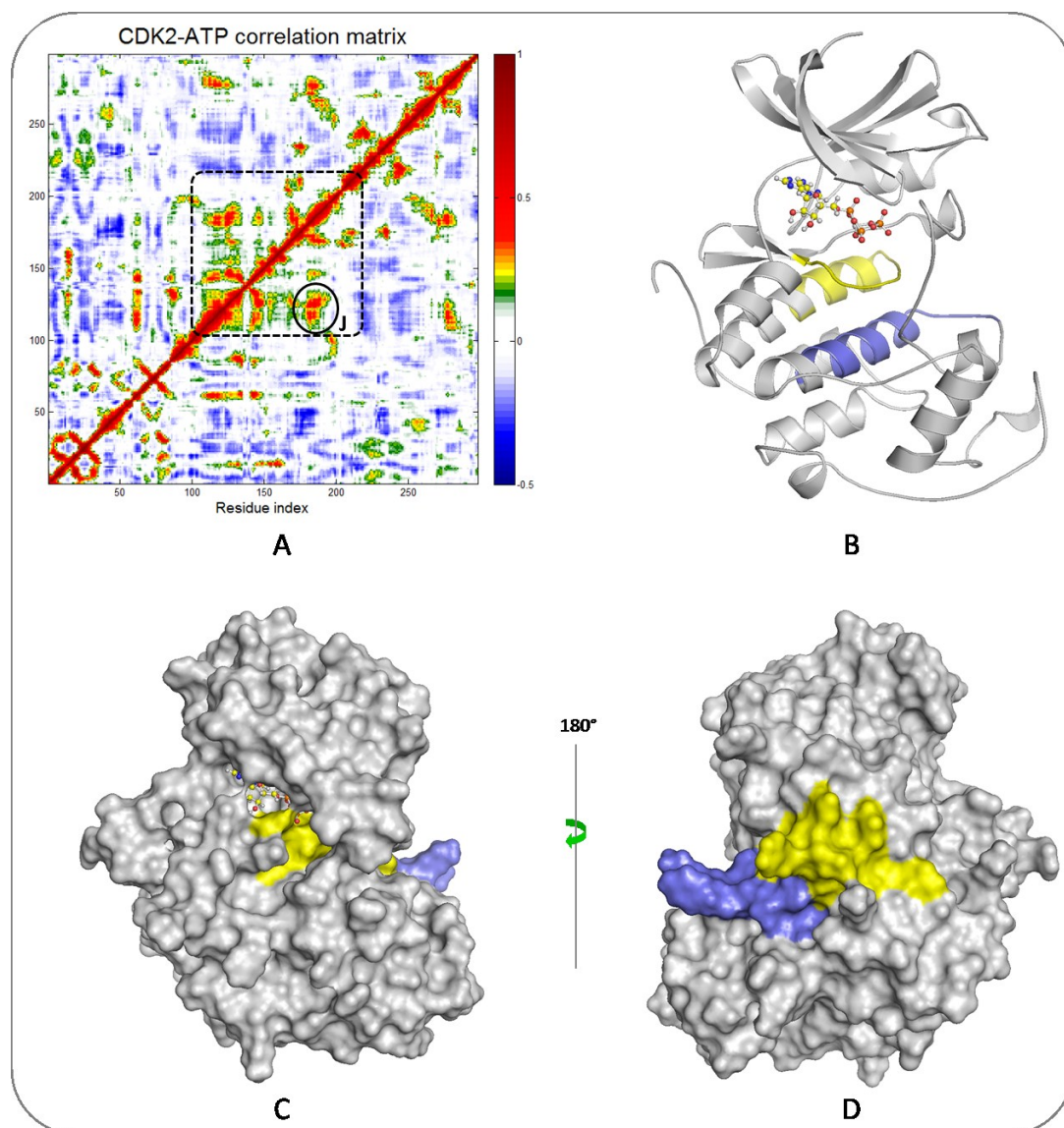


Figure 3.48: Analysis of the correlated residues in the *CDK2-ATP* state. (A) The heat map of the *CDK2-ATP* state with the region (cluster) of high off-diagonal residue-residue coupling is highlighted by black circle J. (B) Structural mapping of the correlated residues in cluster J where the yellow segment is correlated with the blue one. (C and D) are van der Waals surface depiction of the front and back view of the correlated residues as shown in B.

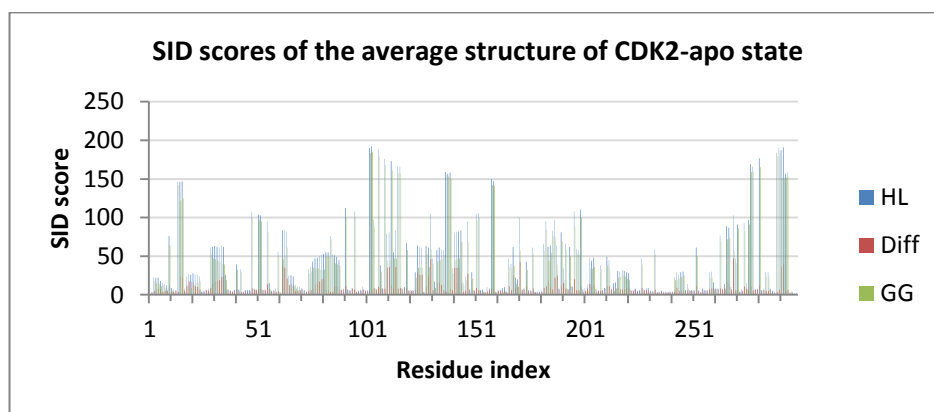
Figure 3.48 shows the correlated residues in the *CDK2-ATP* state. Comparing the heat maps of the two simulated states shows that the correlated residues in both clusters of the *CDK2-apo* state have experienced some attenuation upon binding

ATP. Nevertheless, new correlated residues have appeared (cluster J). Cluster J includes the yellow segment (residues 112-133) which is correlated with the blue segment (residues 178-193) which involve the C-domain. However, the yellow segment (which was implicated in important correlations in the apo state) is still relevant, which suggests that it is mediating the propagation of the dynamical changes in the protein from the ATP binding site and the N-domain towards the C-domain in order to achieve productive assembly of the CDK2 catalytic components.

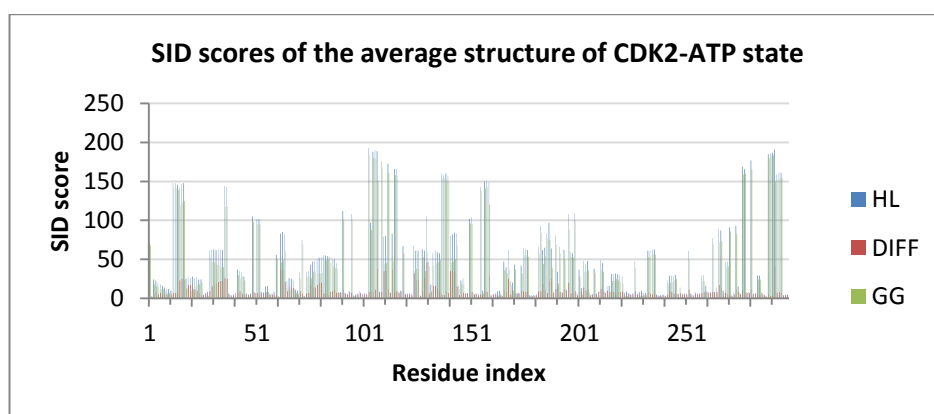
3.2.4 Simple insequence difference analysis

3.2.4.1 SID analysis of the minimised average structures of CDK2

In order to explore the topology of the enzyme and to identify major interfaces in the protein that might be of allosteric potential, SID analysis was applied to the minimised average structures of the two simulated states of CDK2. Following the same procedure described for JNK1, SID scores of the minimised average structures of CDK2 were exported into MS[®] Excel spread sheet and presented as a column chart in figure 3.49. The statistical functions in Excel were used to calculate the upper, median and lower quartiles of each of the SID scores (table 3.3) in order to highlight the residues that are implicated in multi-way interfaces, and then the logical functions within Excel were utilised to select those residues which collectively have high HL and DIFF scores and intermediate GG scores. Visual inspection of the positions of the residues that fulfil the above criteria and the clusters they formed within the 3D structure of the protein was achieved by converting the numerical value of the consensus SID score of each residue into a colour code and structurally mapping it onto the backbone of the corresponding average structure (figure 3.50).



A



B

Figure 3.49: Column chart of the SID scores of each residue position of the minimised average structures of the *CDK2-apo* state (A) and the *CDK2-ATP* state (B). HL scores are coloured blue, DIFF scores are coloured red and GG scores are coloured green.

Table 3.3: Statistical descriptors of the CDK2 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.

Statistical descriptors	SID scores of <i>CDK2-apo</i> state			SID scores of <i>CDK2-ATP</i> state		
	HL	DIFF	GG	HL	DIFF	GG
Upper quartile (UQ)	62	10	47	63	10	50.25
Lower quartile (LQ)	7	5	1	7	5	1
Median (MQ)	24	7	13.5	28.5	7	20
Inter-quartile distance (IQD)	55	5	46	56	5	49.25
Average (AVG)	42.57	9.25	33.31	47.38	9.42	37.96
Minimum value (MIN)	2	1	1	3	2	1
Maximum value (MAX)	192	47	185	193	46	185

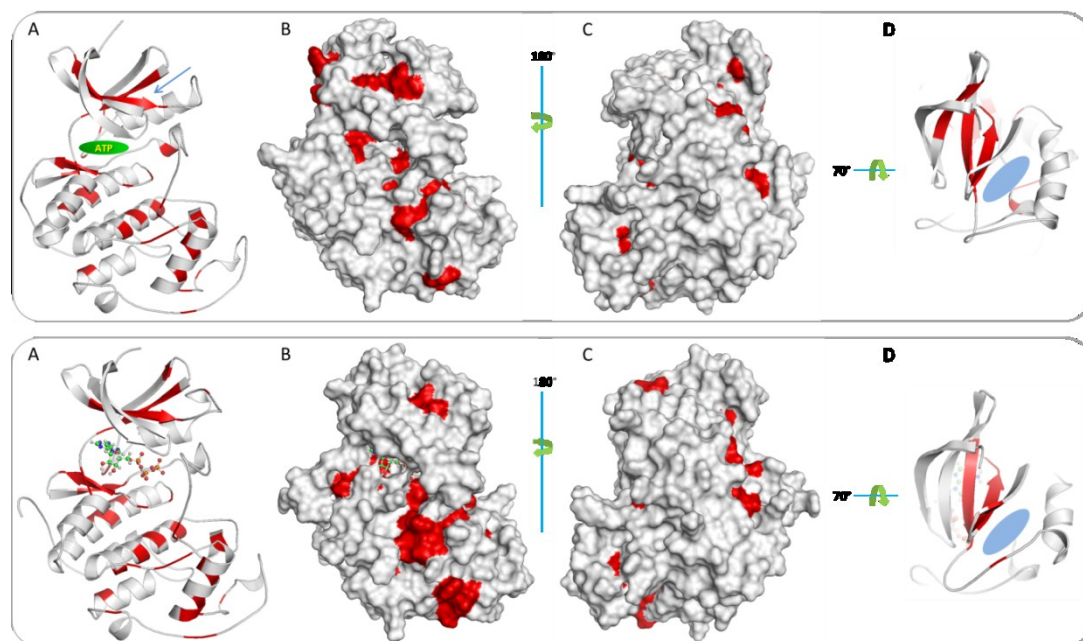


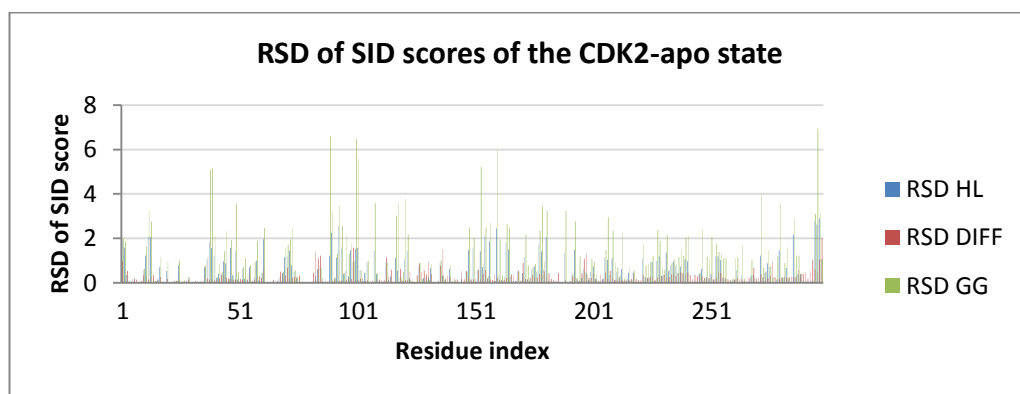
Figure 3.50: SID analysis of the minimised average structures of the two simulated states of CDK2. The upper panel is for *CDK2-*apo** and the lower one is for *CDK2-ATP*. The SID consensus scores (HL, DIFF, and GG) are colour coded and structurally mapped onto the backbone of the corresponding average structure where red colour stands for high scores and white is for low scores. (A) Cartoon representation of the protein backbone. In the apo state, the ATP binding site is highlighted by a green ellipsoid and the allosteric site is by a blue arrow. In the ATP bound state, the ATP molecule is shown in ball and stick representation. (B) Surface representation of the front view of the protein. (C) Back view of the protein by 180° rotation about the y-axis. (D) Top view of the protein backbone by rotating A 80° about its y-axis and 70° about its x-axis. The allosteric binding site is highlighted by the blue ellipsoid.

As figure 3.50 shows, SID analysis (especially of the *CDK2-*apo** state) has successfully identified both of the binding sites in the protein (the ATP binding site and the allosteric site). In the case of the ATP binding site, β -sheets 6 and 7 which form the floor of the ATP binding site showed high SID consensus scores. In the case of the allosteric site, β -sheets 3 and 5 and the DFG segment of the T-loop had a high SID score, which means that they are involved in a complex interface and might be of allosteric potential.

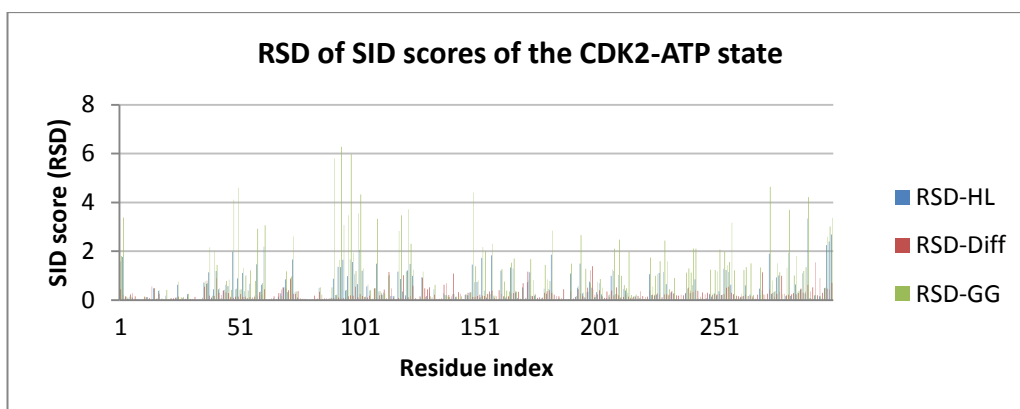
3.2.4.2 Comparative SID analysis of the entire trajectory of each simulated state

Comparative SID analysis of all extracted frames from the simulated trajectories of *CDK2-apo* and *CDK2-ATP* was performed in order to identify possible topological changes that might be associated with the evolution of the protein with time. The RSD for each of the SID scores (HL, Diff,a and GG) was calculated in order to highlight the regions in the protein that have experienced large variations in their SID scores which correspond to regions of potential motion and adjustment in the protein that could be of allosteric potential.

The RSD of SID scores (HL, DIFF and GG) for each simulated state was calculated and presented as column chart in figure 3.51. Those RSD values were then treated in the same way as in the minimised average structures using Excel to calculate the statistical descriptors that guide the selection of residues with prominent topological changes (table 3.4). The consensus SID scores were then converted into a colour code and structurally mapped onto the backbone of their corresponding average structure (figure 3.52).



A



B

Figure 3.51: Column chart of the RSD of SID scores of each residue position of all the frames extracted from the *CDK2-apo* (A) and *CDK2-ATP* (B) trajectories. HL scores are coloured blue, DIFF scores are coloured red and GG scores are coloured green.

Table 3.4: Statistical descriptors of the RSD of CDK2 SID scores used to guide the selection of residues contributing to potential interfaces in the protein.

Statistical descriptors	SID scores of <i>CDK2-apo</i> state			SID scores of <i>CDK2-ATP</i> state		
	HL	DIFF	GG	HL	DIFF	GG
Upper quartile (UQ)	0.763	0.354	1.225	0.614	0.323	1.214
Lower quartile (LQ)	0.053	0.123	0.053	0.037	0.116	0.051
Median (MQ)	0.250	0.201	0.426	0.194	0.197	0.301
Inter-quartile distance (IQD)	0.710	0.231	1.171	0.576	0.206	1.163
Average (AVG)	0.511	0.282	0.937	0.429	0.261	0.813
Minimum value (MIN)	0.001	0.014	0.001	0.001	0.008	0.001
Maximum value (MAX)	2.895	2.001	6.947	3.345	1.549	6.271

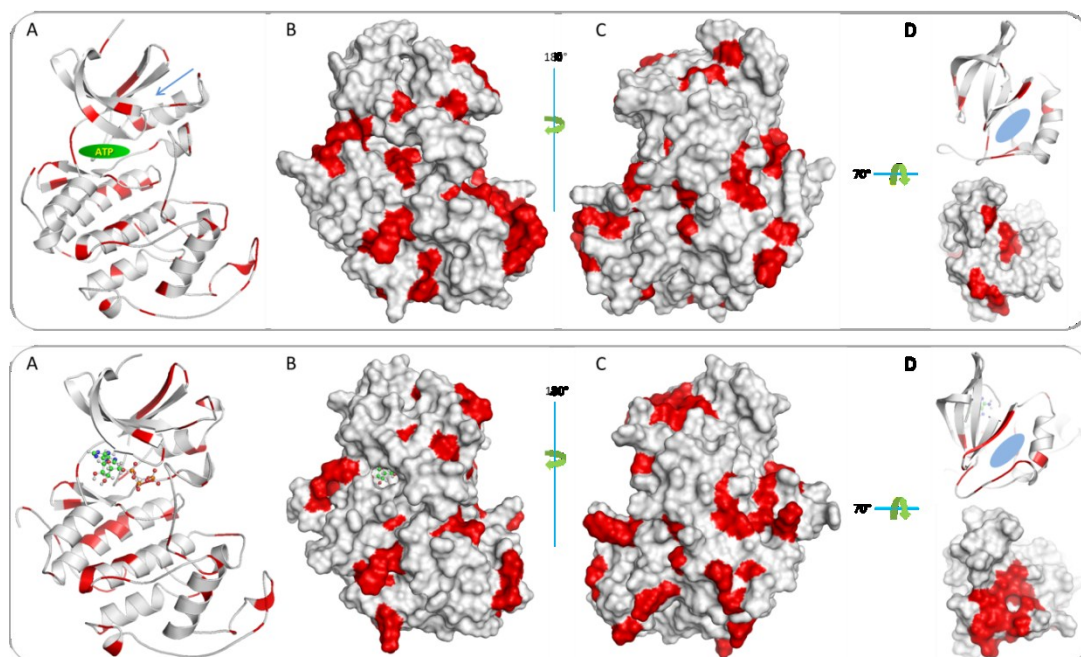


Figure 3.52: SID analysis of all extracted frames from *CDK2-apo* (upper panel) and *CDK2-ATP* (lower panel) simulated trajectories. The overall consensus score of the RSD of the SID scores (HL, DIFF, and GG) is colour coded and structurally mapped onto the backbone of the corresponding average structure where red colour stands for high score and white is for low scores. (A) Cartoon representation of the protein backbone. (B) Surface representation of the front view of the protein. (C) Back view of the protein by 180° rotation about the y-axis. (D) Top view of the protein depicted in cartoon and solid surface by rotating A 80° about its y-axis and 70° about its x-axis. The allosteric binding site is highlighted by the blue ellipsoid.

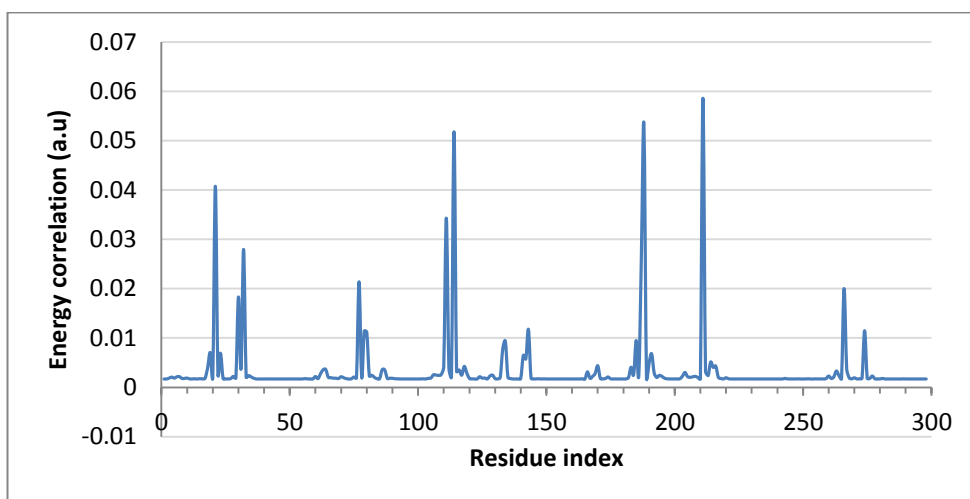
Figure 3.52 represents the comparative SID scoring between the different frames of each of the simulated states of CDK2 which are structurally mapped onto their corresponding minimised average structure; it also gives an overall comparison of regions that experienced large variations in their SID scores between the two states. The figure shows that in both simulated states the region corresponding to the allosteric site showed some variations in the SID scoring of its constituent residues, especially $\beta 4$. This segment ($\beta 4$) was not highlighted in the SID scoring of the average structures, which implies that its conformational flexibility along with that of the surrounding region might affect the underlying interface which was identified in the minimised average structures; and since it is on the surface of the protein, it could

have a role in the initial binding of the allosteric inhibitors and in propagating their perturbing effects to other regions or interfaces in the protein.

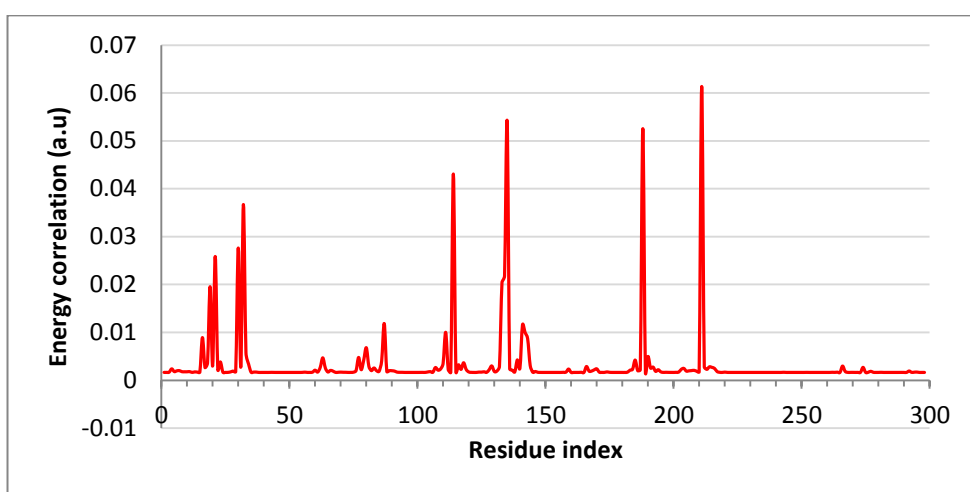
3.2.5 Energy correlations

The energy correlations in the minimised average structures of the two simulated states of CDK2 were calculated using the largest seven eigenvalues of the gamma matrix using C α to represent protein residues. The energy correlation matrix of each average structure was obtained by calculating the energetic interactions of each residue in the enzyme with all other residues. In order to highlight the residues that have high correlations with other residues in the protein, the average value of all correlations for each residue was calculated and plotted against residue index in figure 3.53.

The observed changes in the number of peaks in the *CDK2-ATP* state reflect the effects of ATP binding on the overall energetic interactions in the protein. Those changes can be better investigated by visualizing the energetic interaction pathways that are made of the residues corresponding to the peaks in figure 3.53. They were structurally mapped onto their corresponding average structure as shown in figure 3.54.



A



B

Figure 3.53: The peaks of energy correlation that define the interaction pathways in CDK2. (A) The *CDK2-*apo** state. (B) The *CDK2-ATP* state. The correlations are given in arbitrary units.

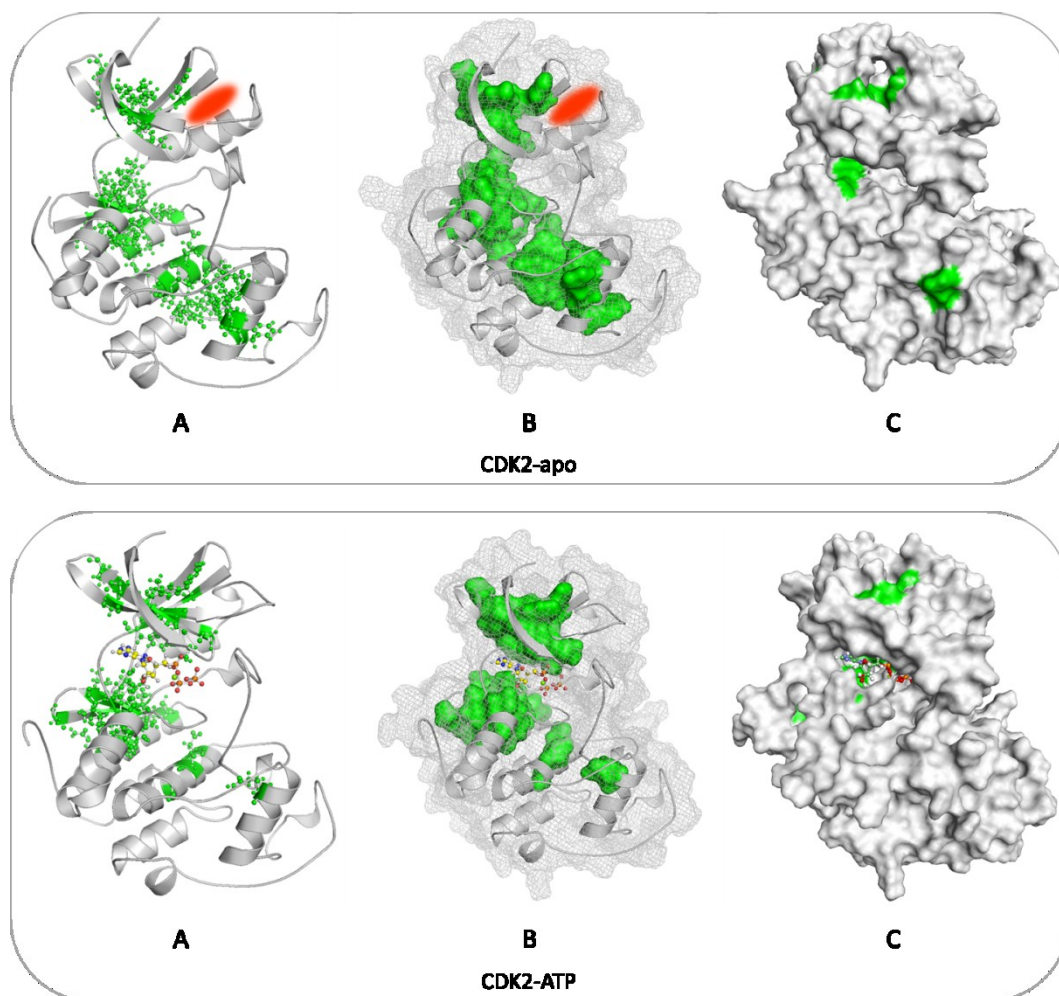


Figure 3.54: Structural mapping of the residues that define the interaction pathways in *CDK2-apo* (top) and *CDK-ATP* (bottom) states. (A) Residues forming the interaction pathway are represented by green balls and sticks, and other residues in the protein are coloured grey. (B) Surface representation of the residues forming the interaction pathway within a mesh surface of the entire protein. (C) Full surface representation of all residues in the protein showing the energy gates on the surface. The ATP molecule is represented by balls and sticks and coloured by element, and the allosteric binding site is highlighted by the red ellipsoid.

Figure 3.54 shows that in the *CDK2-apo* state, the residues of high energetic correlations formed a well-defined continuous interaction pathway extending from the bottom of the C-domain all the way through the ATP binding site up to the top of the N-domain where it involves parts of $\beta 3$ and $\beta 5$ that are located within the

allosteric binding site. This interaction pathway seems to be capable of mediating efficient communication between distant parts of the protein; such that a perturbation in remote sites in the protein will be transmitted to another. Such effects are evident in the *CDK2-ATP* state where ATP binding has disrupted the continuity of the native interaction pathway in the C-domain. However, the energetic correlations are now more localized around the ATP binding site, and the part of the interaction pathway in the N-domain has expanded a little to include the G-loop and the catalytic Lys33. It seems that ATP binding has modified the energetic correlations between protein residues and only maintained the functional ones.

3.2.6 Summary of the overall results of CDK2

In order to compose a comprehensive picture of the main findings in studying CDK2, all the results of the computational methods are assembled in figure 3.55.

Figure 3.55 shows that the results from different computations complement each other in identifying the allosteric binding site in the N-domain of the protein. For example, the fluctuation analysis has showed that all the components of the allosteric binding site are fairly flexible; the cross correlation analysis has revealed the dynamical correlations between the first part of the α C-helix and the DFG region of the T-loop; the SID and energy correlation analyses have revealed the potential and importance of β 3-5. Collectively, the overall findings have indicated that the α C-helix, the DFG segment of the T-loop, and β -sheets 3-5 could be of high allosteric potential, which agrees with the experimental findings.

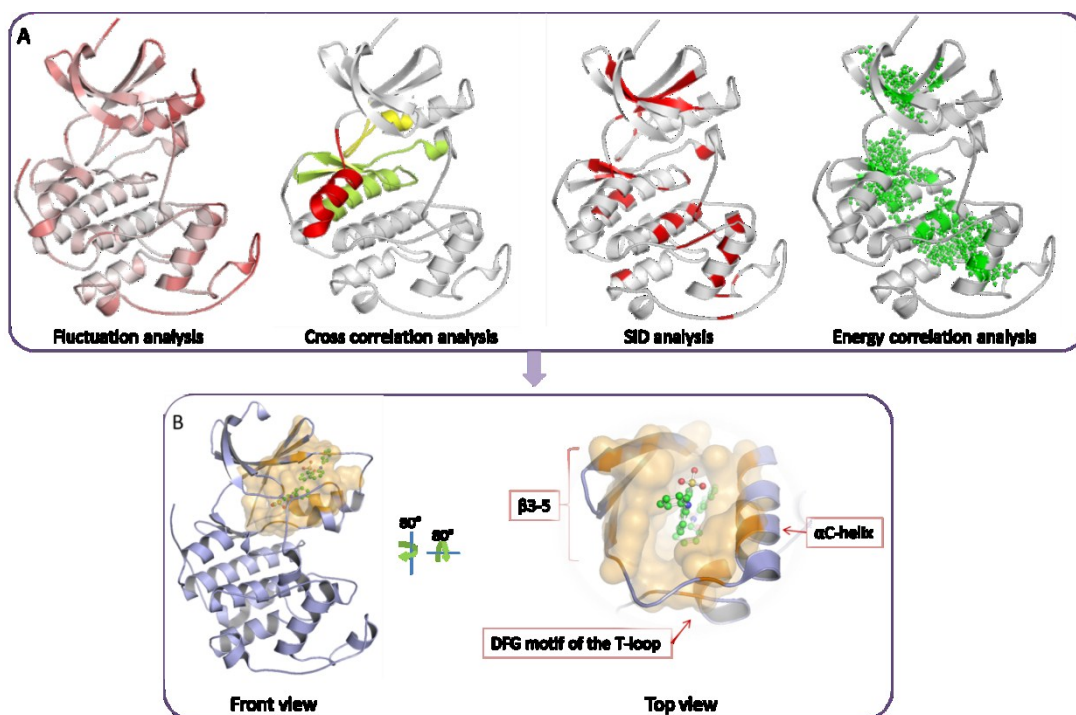


Figure 3.55: Summary of the main computational results of CDK2. (A) Assembly of the results of the different types of computational analysis implemented in studying CDK2. (B) The experimentally identified allosteric site. The allosteric inhibitors (ball and stick) are occupying their corresponding binding site (orange surface).

Chapter Four

4 CASE STUDY (DYRK2)

In the previous sections computational approaches were applied to kinases with known allosteric binding sites. Since the methodologies convincingly identified the experimentally determined allosteric sites, we applied the same approach to the kinase DYRK2 to search for and identify a putative binding site with allosteric potential. Allosteric small molecule regulators of DYRK2 have not previously been identified.

4.1 The role of DYRK2

DYRK2 is a member of a conserved dual-specificity tyrosine-regulated kinases (DYRKs) family that belongs to the CMGC group of protein kinases. The mammalian subfamily of DYRK comprises 5 members: DYRK1A, DYRK1B (also named Mirk), DYRK2, DYRK3 (also named REDK), and DYRK4 [175]. All of the DYRK family members share a conserved kinase domain and an adjacent upstream N-terminal DYRK homology (DH) box, but are divergent in their N- and C-terminal extensions [176, 177] (figure 4.1). Members of this family are defined as dual-specificity protein kinases because they phosphorylate tyrosine, serine, and threonine residues. Their activation mechanism involves autophosphorylation of a tyrosine residue at a conserved YXY sequence in the activation loop and distinguishes them from other kinases. This autophosphorylation is intramolecular and happens during protein translation. Their tyrosine kinase activity is limited to autophosphorylation at the conserved tyrosine residue, and once phosphorylated they lose their tyrosine kinase activity and functions only as a serine/threonine kinase [176]. The activity of DYRKs might be regulated by different mechanisms such as phosphorylation outside the activation loop; interaction with regulatory proteins; and control of subcellular localisation or protein stability [175].

DYRK2 is mainly expressed during development and in adult testes [175]. It has the ability to act as priming kinase, meaning that DYRK2 phosphorylation of a given residue in the target protein will facilitate the phosphorylation of another residue by a subsequent kinase. It is a priming kinase for phosphorylation of c-Jun/c-Myc, which coordinates their regulation [178]. DYRK2 also has a role in the DNA damage signalling pathway by regulating p53 via phosphorylating its serine 46 which

ultimately induces cellular apoptosis [179]. It is involved in many cellular processes such as cell proliferation, cytokinesis and cellular differentiation [180], and recently has been shown to have two roles in the E3 ubiquitin ligase complex; as a scaffolding protein for the formation of the complex and as a kinase to phosphorylate the ligase substrates [181]. It is overexpressed in adenocarcinomas of the oesophagus and lung, yet its exact mechanism of involvement in tumorigenesis is still to be clarified [182].

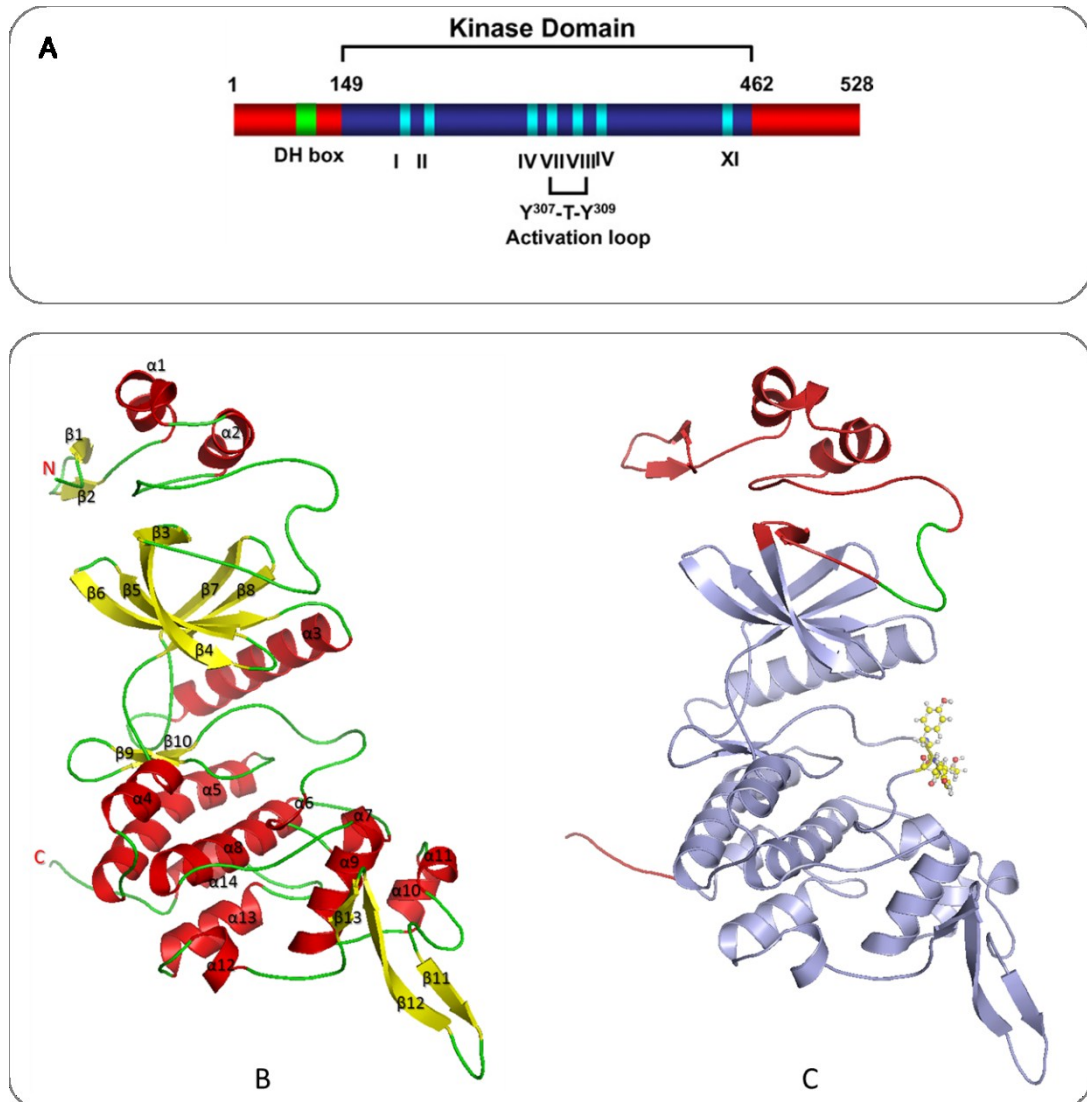


Figure 4.1: The structure of DYRK2. (A) Schematic depiction of DYRK2 structure. The kinase subdomains are numbered using roman numerals. The activation loop is highlighted along with the conserved YXY sequence [180]. (B) Cartoon representation of the backbone atoms of the catalytic domain coloured according to

their secondary structure; α -helices are in red, β -sheets are in yellow, loops and turns are in green. (C) Cartoon depiction of DYRK2 highlighting the substructures shown in the schematic representation in (A). The N and C-terminal extensions are coloured red; the DH-box which is located within the N-terminal extension is coloured green; the kinase domain is coloured light blue; and the residues of the conserved YXY sequence (YTY in DYRK2) within the T-loop are depicted as balls and sticks and their carbon atoms are coloured yellow.

A putative allosteric binding site has the following characteristics: it needs to be surface accessible within a pocket or a groove on the protein in order to establish interactions with the incoming ligand for proper binding and good affinity; it should be linked to the active site via interaction pathway(s) in order to transmit the dynamic and/or conformational effects associated with ligand binding to the active site of the protein, thereby modulating its activity; it is likely to be at a multi-way interface in the protein so that ligand binding will affect the overall stability of the protein (positively or negatively) through disturbing the stability of the interface itself; it needs to have flexibility to allow for optimal fitting of the ligand after initial binding.

4.2 System stability and conformational flexibility

The computational approach was implemented by running a 50 ns MD simulation of the *apo* state of the enzyme. The *apo* state of a protein represents its native state in which all the necessary information for performing its natural function is stored. The *DYRK2-apo* state model was prepared as described in the Materials and Methods chapter. The results of DYRK2 study will be presented in a similar way to that of JNK1 and CDK2.

4.2.1 System stability

Root mean square deviation:

The stability of the simulation was examined by calculating the mass weighted RMSD of the backbone atoms of the simulated state from its starting structure and

plotted versus time. In some cases where a simulated protein has very flexible segments (such as the loop extensions in DYRK2) the fluctuations in the RMSD maybe a misrepresentation of the overall conformational change in the backbone of the protein. Large changes in the RMSD can result from high fluctuations in these localized flexible segments rather than a real conformational change in the well-structured secondary structures. It is therefore advisable to check for such “un-realistic” fluctuations by recalculating the RMSD of the well-structured part of the enzyme and excluding the very flexible loop segments (figure 4.2).

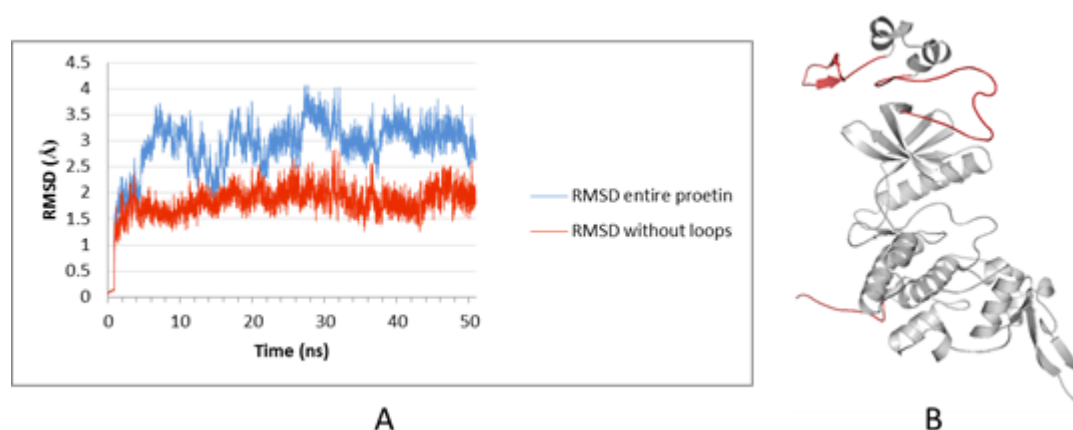


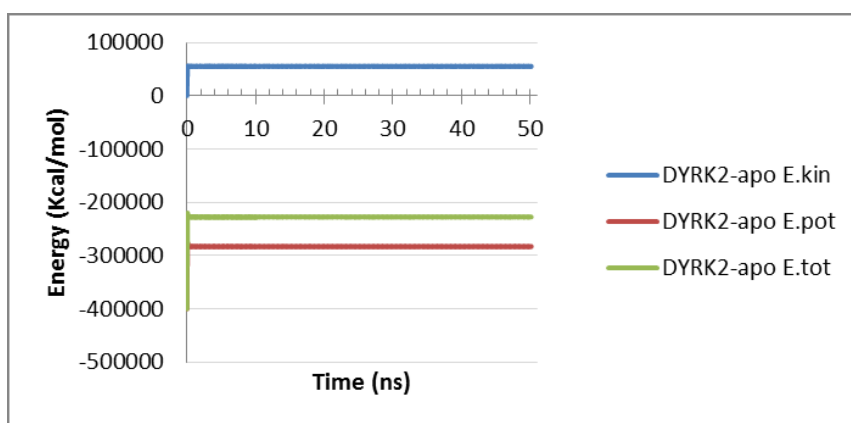
Figure 4.2: (A) Backbone atoms’ RMSD versus time plot for the MD simulation of *DYRK2-apo* state using the starting structure of the trajectory as a reference. The blue line corresponds to the RMSD calculated for the backbone atoms of the entire protein, and the red line represents the RMSD of the backbone atoms excluding the flexible segments of the protein (residues 1-22, 45-81 and 398-410). (B) The excluded flexible (loop) segment is highlighted in red.

Figure 4.2A shows that there is a clear difference between the two plots before and after excluding the flexible segment of the protein from the RMSD calculation. The average RMSD value of the backbone atom of the entire protein (blue line) is 2.9\AA , which is reduced to 1.8\AA when the flexible segment (red line) is neglected. Moreover, the shape of the plot is very different; there are fewer fluctuations which reflect a very well equilibrated and stable simulation. The system reached an initial

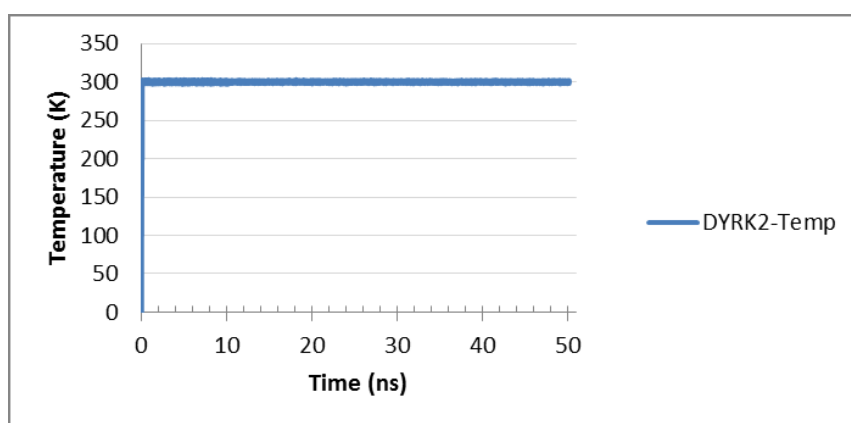
equilibrated state in about 4ns, and stabilised around an RMSD average value of 1.8Å. All analyses focused on the last stable 40 ns of the trajectory.

Energy conservation and temperature stability:

Figure 4.3 shows that all the energies increased during the heating phase of the simulation then they levelled off indicating that the total energy is very well conserved and the system is stable. There were no fluctuations in temperature after the heating stage indicating that the thermostats used were successful in controlling the temperature.



A



B

Figure 4.3: Summary of the energy and temperature changes during the simulation of *DYRK2-apo*. (A) Summary of energy changes. (B) Summary of temperature changes.

4.2.2 Conformational flexibility

Average structures:

The conformational flexibility of DYRK2 was assessed by calculating an average structure from the equilibrated phase (last 40 ns) of the simulated trajectory and comparing it with the corresponding starting structure (figure 4.4). Afterwards, the structural changes in the system were characterised by calculating the RMSD of the entire backbone atoms of the average structure from its starting crystal structure. Figure 4.4 shows that the major conformational changes are related to the two flexible termini, (β_1 , β_2 , α_1 , and α_2) in the N-terminus and (β_{11-13}) in the C-terminus. In the classical kinase domain, the main changes involved α_4 and α_{13} and to a lesser extent β_4 and β_5 .

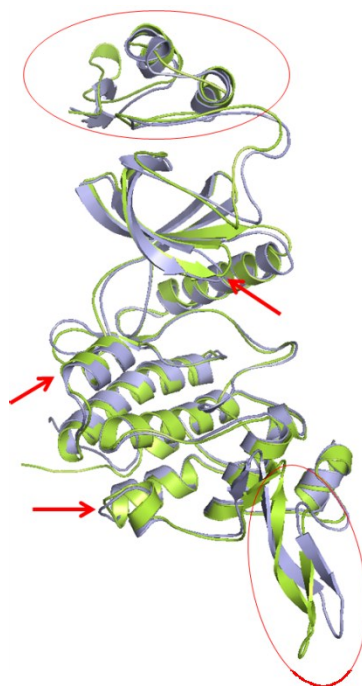


Figure 4.4: Comparison between the starting structure (light-blue) and the average structure (green) of the simulated *DYRK2-apo* state. The superimposition of the starting and average structures was based on alignment of the backbone atoms of the entire protein. Red arrows and ellipsoids highlight the regions of the enzyme with major conformational changes.

Residual fluctuations:

The conformational variability of the simulated system was studied by calculating the residual fluctuation based on the simulation trajectory (figure 4.5).

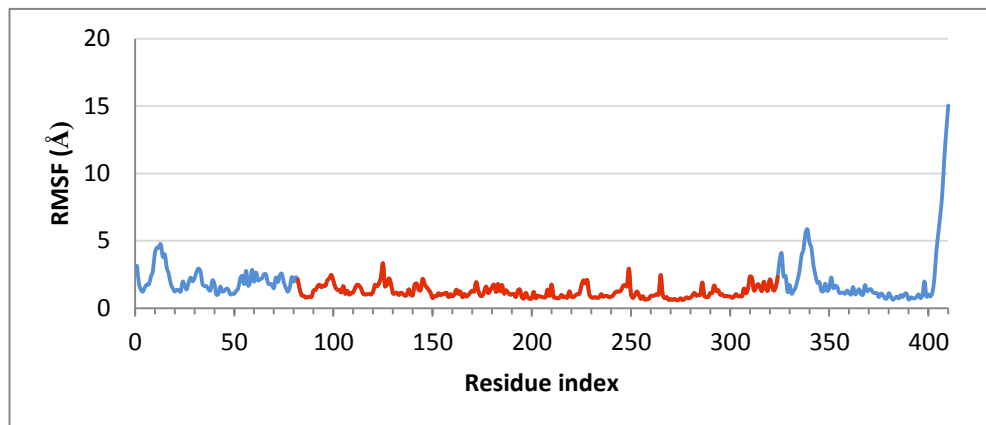


Figure 4.5: The residual fluctuation of *DYRK2-apo* plotted as RMSF versus residue index.

It is obvious from figure 4.5 that the C-terminal segment has a very high flexibility as reflected by the extremely high RMSF value (around 15Å). To focus on the fluctuations of the well-structured part of the enzyme, the flexible part (blue regions in the plot line) was neglected by setting the fluctuation values of its residues to 1, and the fluctuation values converted into a colour code and structurally mapped onto the backbone of the average structure to obtain an overall picture of protein flexibility (figure 4.6).

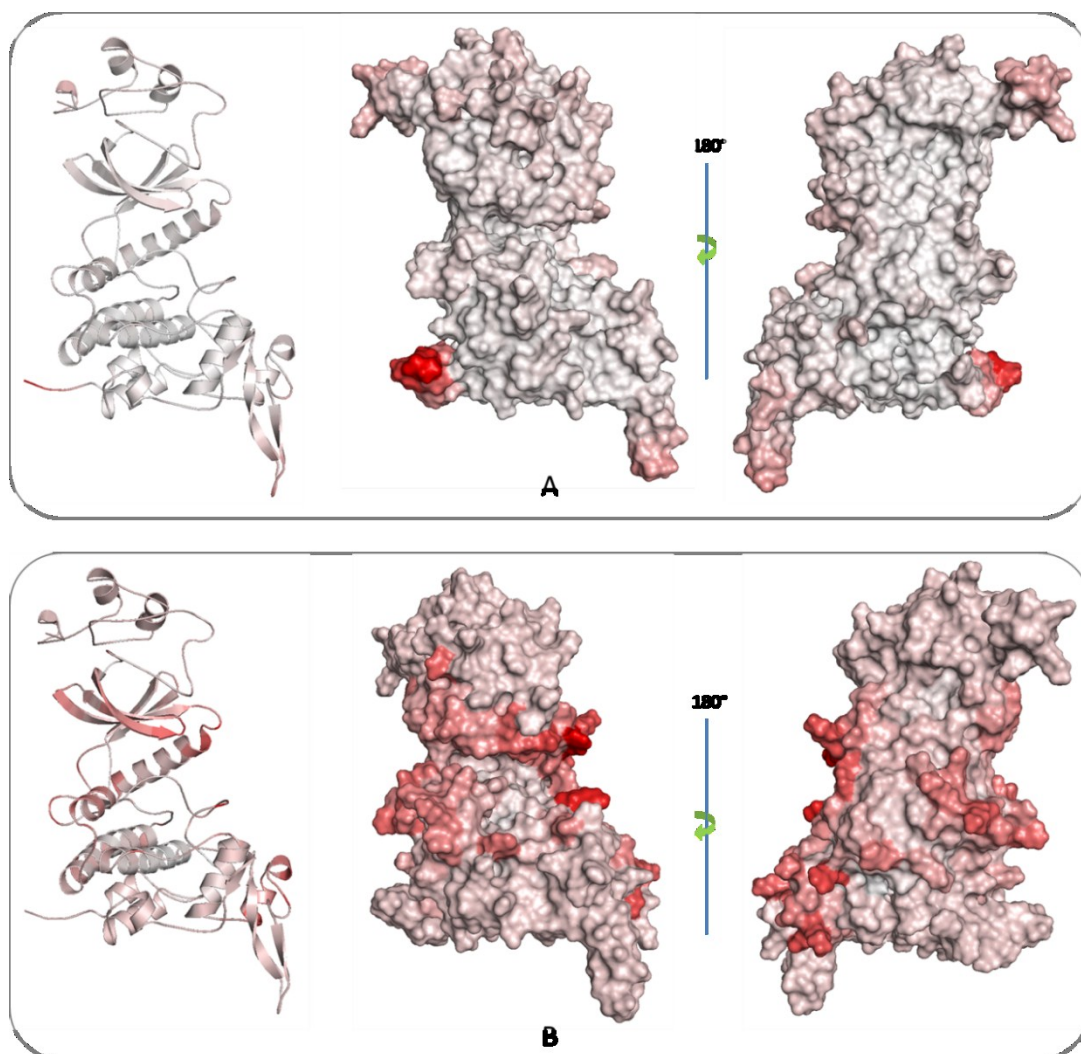


Figure 4.6: Colour coded representation of the residual fluctuation values of *DYRK2-apo*. (A) The protein is coloured using the RMSF values without manipulation. (B) Colouring after neglecting the RMSF values of the very flexible segments (residues 1-81 and 325-410). Red colour corresponds to high fluctuations and white is for low fluctuations. The protein is represented in backbone cartoon and in solid surfaces.

Figure 4.6A shows that because of the very high fluctuation value of the C-terminal loop (deep red), the rest of the protein appeared rigid (almost white). After neglecting the very flexible segments the fluctuation profile showed that the most flexible regions in the protein corresponds to β 4-6 and α 3 in the N-domain, and α -helices 4,10 and 11 in the C-domain.

4.3 Analysis of correlated motion

The residual cross-correlation motion of the *DYRK2-apo* trajectory was calculated and shown as the heat map of the calculated residual dynamic cross-correlation matrix of the backbone atoms in figure 4.7.

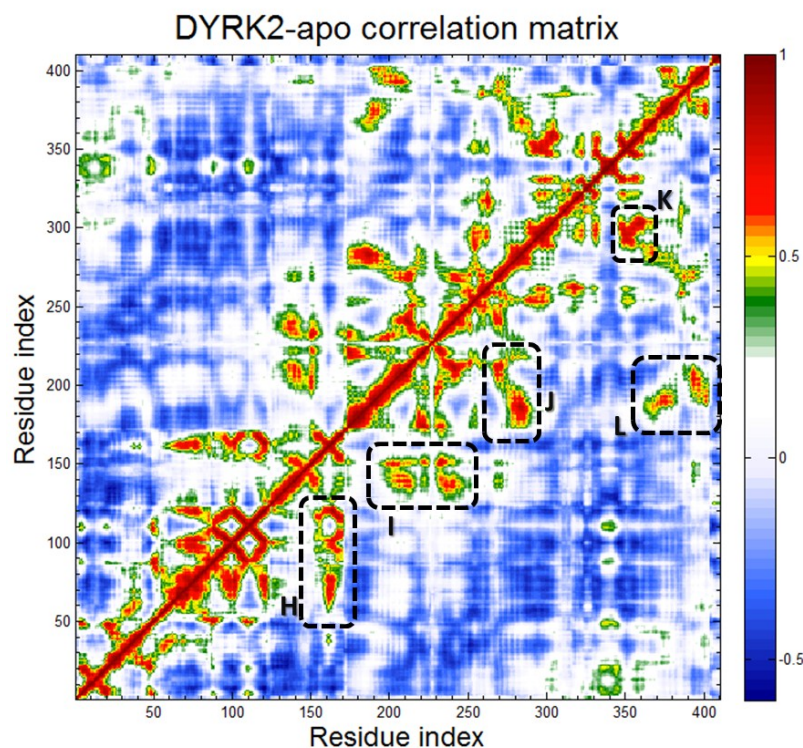


Figure 4.7: The matrix of the residual cross-correlation motion of the backbone atoms around their average position calculated from the simulation trajectory and represented as a heat map. Red pixels stand for high positive correlation, whereas blue pixels correspond to negative correlation. Black dashed boxes H-L highlight the off-diagonal clusters of concentrated residue-residue coupling in the protein.

In figure 4.7, the scale of the colouring scheme is from -0.6 (negatively correlated residues) to 1.0 (positively correlated residues). The overall average of residual correlations in the *DYRK2-apo* state was 0.028 and the average of the positively correlated residues was 0.277. Residues that have correlation values of more than 0.554 (which is twice the average of positive correlations) were considered as highly correlated with each other.

Regions of concentrated residue-residue coupling can be defined into five clusters, H-L. Cluster H comprises residues 57-124 that are correlated with residues 152-171. Cluster I involves residues 128-151 which are correlated with residues 195-249. Cluster J includes residues 174-222 which are correlated with residues 266-291. Cluster K encompasses residues 287-310 that are correlated with residues 346-369. Finally, cluster L, in which residues 180-212 are correlated with residues 363-404. In order to visually inspect these clusters or sub-domains that comprise those correlated residues within the 3D structure of the protein, they were structurally mapped onto the backbone of the average structure (figure 4.8).

Figure 4.8 shows the topological location of the five defined clusters in the heat map. Within this 3D distribution of the clusters we can see how the key features of the ATP binding site interplay to ensure optimal binding and alignment of the ATP molecule. For example in cluster I, the α C-helix is correlated with residues 195-249 which include the floor of the ATP binding site (β 9 and β 10), the catalytic loop and a large part of the T-loop, especially the DFG motif. All of these regions are functionally important not only for DYRK2 but for all kinases. Also in cluster H, β -sheets 4-6 that incorporate the G-loop and Lys118 which are crucial for productive ATP binding, are correlated with residues 152-171 that include the gate-keeper Phe168 and G+1 (Glu169) and G+3 (Lys171) in the hinge. Disruption of the natural dynamics of these regions may have a negative effect on the ATP binding site and the activity of the protein as a whole, which suggests that this subdomain may be allosterically relevant.

There is also some overlap between these clusters that may act to convey dynamic changes from one cluster into the other. For example, helix α 5 is shared by cluster I, J and L; and helix α 12 is shared by cluster K and L. This implies that perturbation (such as through ligand binding) at or near these clusters or sub-domains will alter their local dynamics and will dissipate to other regions of the protein leading to a global disruption of protein dynamics, conformation and function.

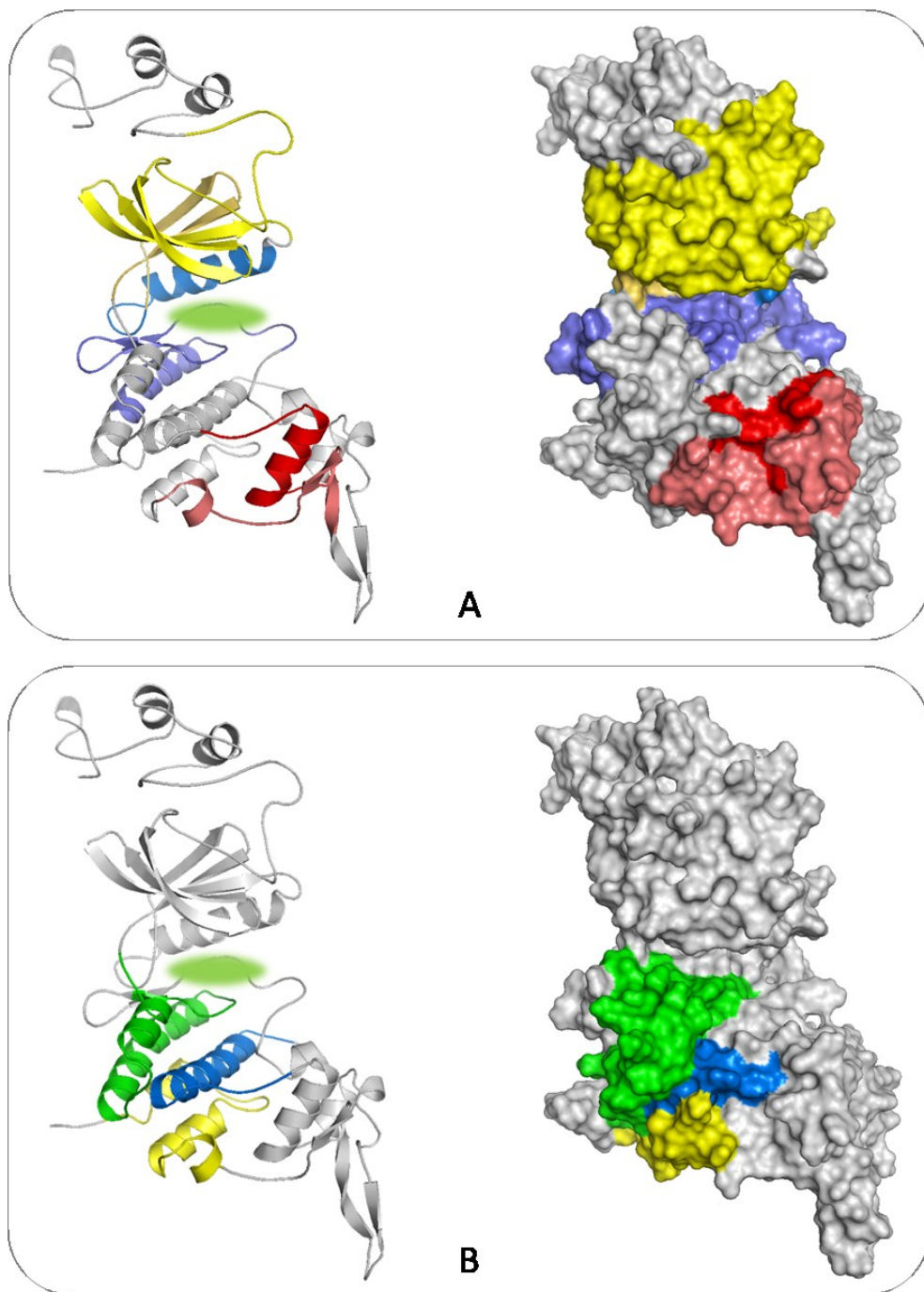


Figure 4.8: Structural mapping of the concentrated residue-residue coupling areas in *DYRK2-apo* heat map (boxes H-L in figure 4.7). (A) Cartoon and surface representation of clusters H (yellow), I (blue) and K (red). (B) Cartoon and surface representation of cluster J (green and blue) and L (yellow and green). Residues of each coloured segment are positively correlated together. The ATP binding site is highlighted by the green ellipsoid between the N and C-domains.

4.4 Simple intrasequence difference analysis

SID analysis was applied to the minimised average structure of DYRK2 as described before. SID scores of each residue (HL, DIFF and GG) were collectively considered to generate a consensus score that guides the identification of major interfaces in the protein. The values of different SID scores against residue index were plotted (figure 4.9) after which statistical and logical functions were used to select residues with consensus SID scores (table 4.1). Consensus SID score of each residue was converted into a colour code and structurally mapped onto the backbone of the average structure (figure 4.10).

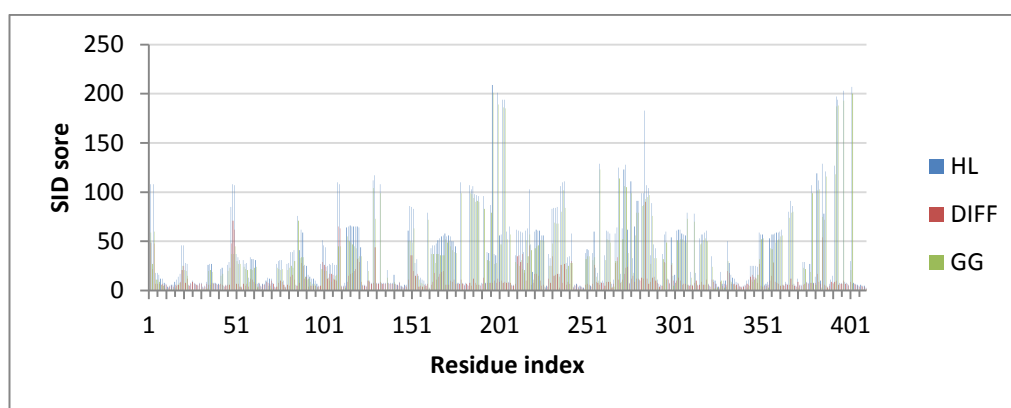


Figure 4.9: Column chart of the SID scores of each residue position of the minimised average structure of the *DYRK2-*apo** state. HL scores are coloured blue, DIFF scores are coloured red and GG scores are coloured green.

Table 3.3: Statistical descriptors of the DYRK2 SID scores used to guide the selection of residues contributing to complex interfaces in the protein that are of possible allosteric potential.

Statistical descriptors	SID scores of <i>DYRK2-<i>apo</i></i> state		
	HL	DIFF	GG
Upper quartile (UQ)	59	11	45
Lower quartile (LQ)	8	6	1
Median (MQ)	27	7	17.5
Inter-quartile distance (IQD)	51	5	44
Average (AVG)	39.609	10.729	28.880
Minimum value (MIN)	2	1	1
Maximum value (MAX)	209	90	201

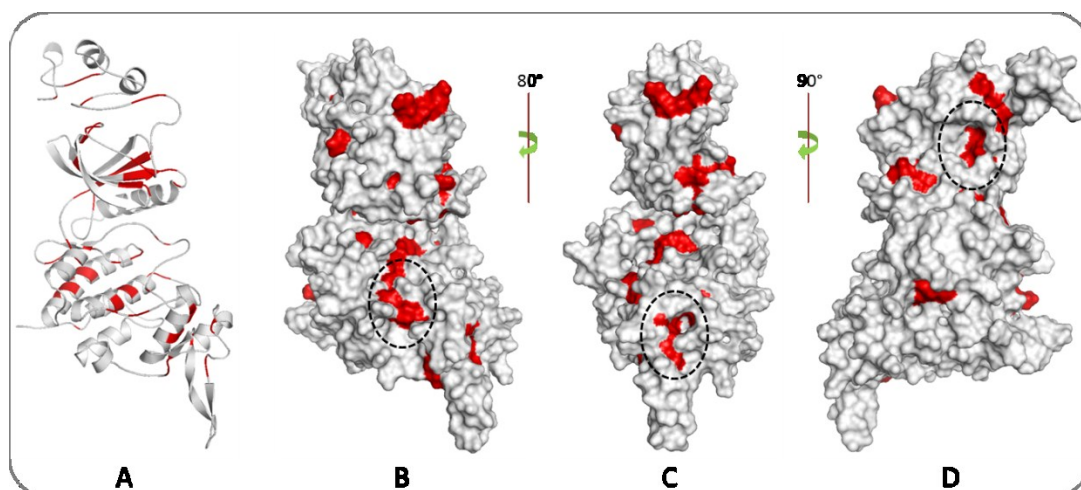


Figure 4.10: SID analysis of the minimized average structure of *DYRK2-apo*. The SID consensus scores (HL, DIFF, and GG) are colour coded and structurally mapped onto the backbone of the average structure where the red colour stands for high scores and white is for low scores. (A) Cartoon representation of the protein backbone. (B) Surface representation of the protein. (C) 80° rotation about the y-axis of (B). (D) 90° rotation about the y-axis of (C). The dashed black circles highlight regions with possible allosteric potential.

Figure 4.10 shows that SID analysis has highlighted few interfaces in both domains of the enzyme that are surface accessible (as shown from the solid surface depiction) and associated with a pocket or groove to enable ligand binding. The three such regions are highlighted by dashed black circles.

4.5 Energy correlations

The energy correlation matrix of *DYRK2-apo* was generated by calculating the energetic interactions of each residue in the enzyme with all other residues, followed by plotting the average correlation value for each residue against residue index (figure 4.11). Residues with high energetic correlations are important for effective communication between different parts of the protein during signal propagation. The energy correlations were calculated based on the largest seven eigenvalues of the gamma matrix using C α to represent protein residues.

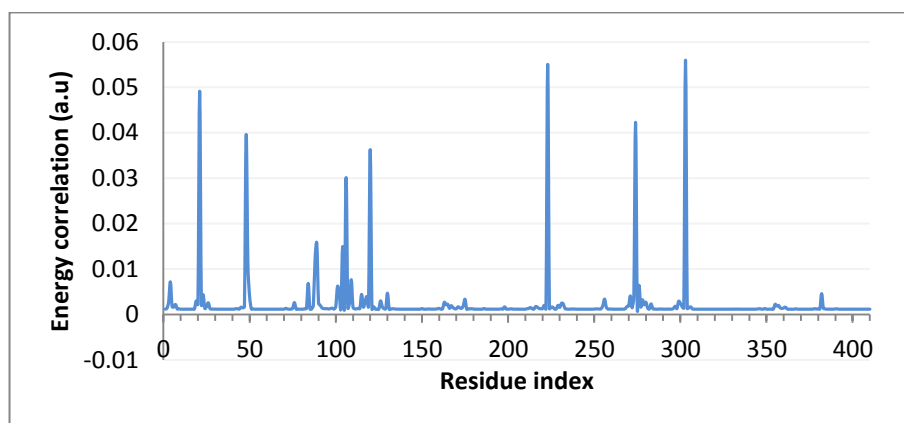


Figure 4.11: The energy correlation peaks that define the interaction path in DYRK2. The correlations are given in arbitrary units.

The peaks in figure 4.11 correspond to residues that have high average energy correlation values with other residues in the protein as calculated from the energy correlation matrix. These residues constitute the energy gates and the interaction pathways in *DYRK2-apo*, and have been structurally highlighted within the average structure in figure 4.12.

Figure 4.12 shows that the residues of the interaction pathway in *DYRK2-apo* are separated into two well defined pathways, one is in the N-domain of the protein and the other is in the C-domain. Both of the two pathways connect the ATP binding site with other regions in the corresponding domain. The gap between the two pathways is located exactly at the ATP binding site. In this context, the ATP molecule can be thought of as being the missing link that completes the chain, which when bound, orchestrates the necessary dynamic and conformational changes in both domains of the enzyme through these interaction pathways to achieve optimal activity. Energetic perturbation through ligand binding at suitable sites on either domain of the enzyme may disrupt the relevant interaction pathway by altering the optimal alignment of its residues thereby affecting the activity of the enzyme. Perturbation could also dissipate to the ATP binding site and alter the orientation of key residues in the active site. The solid surface rendering of the protein shows few green regions that correspond to the “energy gates”, and only one is located within a pocket which is highlighted by a dashed black circle.

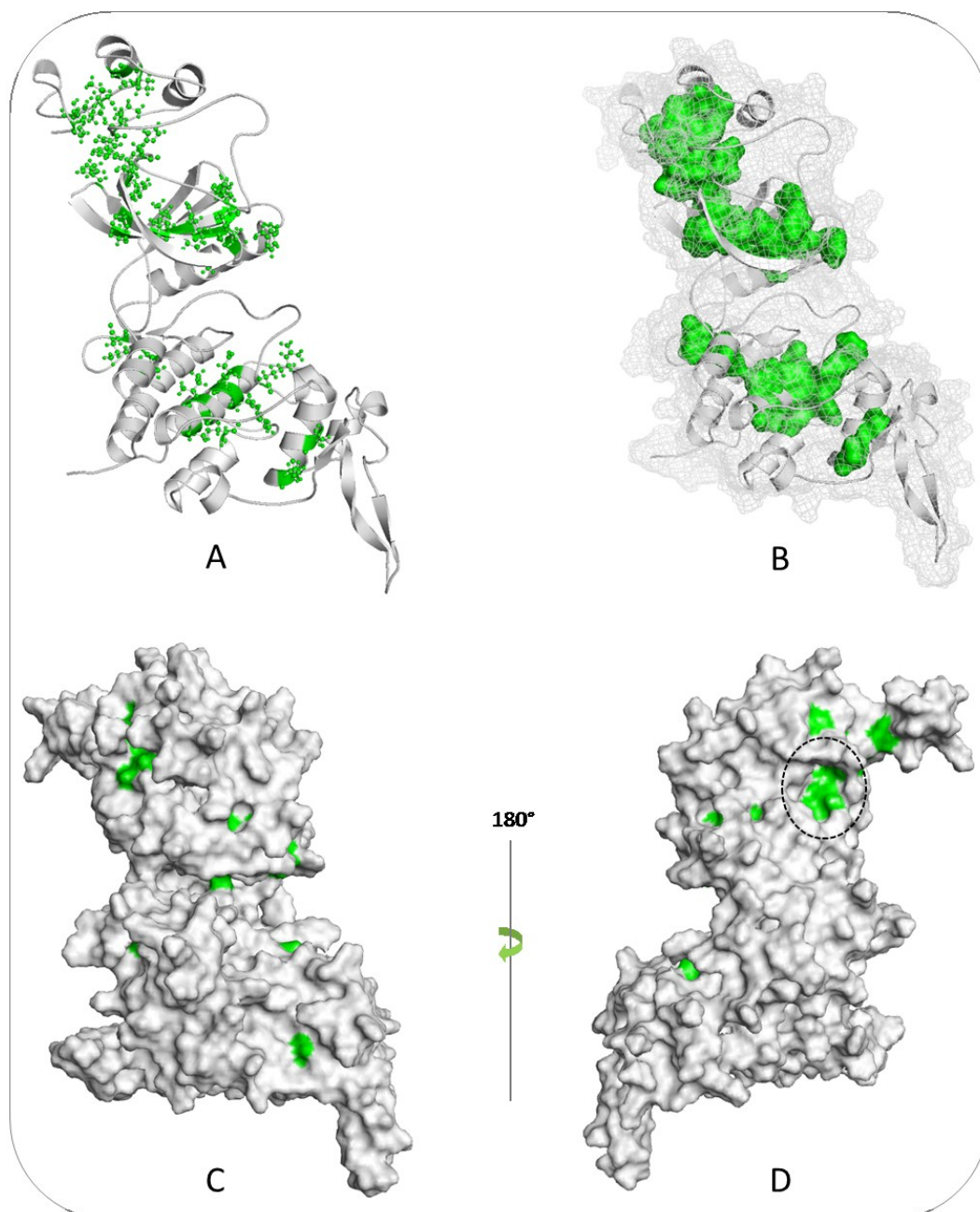


Figure 4.12: Structural mapping of the residues that define the interaction pathways in *DYRK2-apo*. (A) Residues forming the interaction pathway are represented by green balls and sticks and the rest of residues are coloured grey. (B) Surface representation of the residues forming the interaction pathway within a mesh surface of the entire protein. (C) Solid surface representation of all residues in the protein showing the energy gates on the surface of the protein. (D) Back view of protein surface in (C) by 180° rotation about the y-axis. The dashed black circle highlights the only surface accessible region that is located within a pocket.

4.6 Summary of the overall results of *DYRK2-apo*

The overall picture of the allosteric map in *DYRK2-apo* can be obtained by assembling the results of the different types of computational analysis that were employed in studying this kinase. By assessing these results in association with the characteristics of a good allosteric binding site that were listed at the beginning of this section, we identified the putative binding site in the N-domain of DYRK2 as having high allosteric potential (figure 4.13).

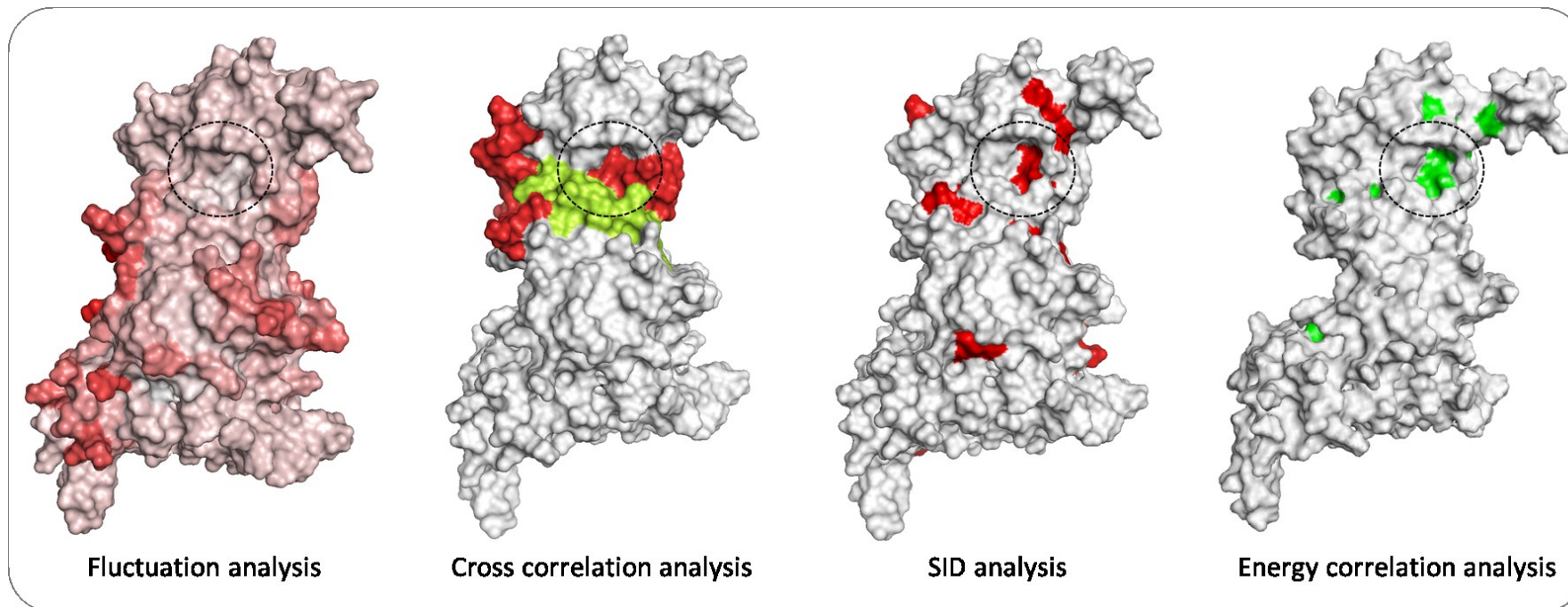


Figure 4.13: Summary of the overall results obtained from applying different computational analysis in studying *DYRK2-apo*. In each depiction, the residues that reflect the result of the pertinent analysis are in a different colour to those considered irrelevant (coloured grey). The putative allosteric binding site in the N-domain of the protein is highlighted by the dashed black circle. See text for more details.

Figure 4.13 shows the overall results of applying the computational approach in studying *DYRK2-apo*. The first depiction corresponds to the fluctuation analysis where the colouring is based on the RMSF values after neglecting the very flexible termini, the more red the colour the higher the flexibility. The second depiction is showing cluster A of the dynamic cross-correlation motion analysis in which the two segments that are correlated together (residues 57-124 and 152-171) are coloured red and green respectively. The third is the SID analysis of the average structure where the regions of the protein that corresponds to complex multi-way interfaces are coloured red. Finally, the energy correlation where the energy gates to the interaction pathways are coloured green. In all of these depictions, residues that are irrelevant to the analysis are coloured grey. The putative allosteric binding site is highlighted by a dashed black circle.

Figure 4.13 shows that the identified binding site fulfils the characteristics that define a plausible binding site. Firstly, it is located in a well-formed pocket between $\beta 4$, $\beta 7$, the loop that connects $\beta 3$ and $\beta 4$, and the loop connecting $\alpha 2$ and $\beta 3$; ligands that bind here could establish good interactions with the protein. Secondly, this site includes some loops that are flexible along with $\beta 4$ which has a considerable flexibility as shown by the residual fluctuation analysis which would allow for optimal fitting of the ligand after the initial binding. Thirdly, the dynamic cross-correlation motion analysis shows that it is located in cluster H at the interface between two correlated segments that incorporate the G-loop and Lys118 that are important for ATP binding, along with the gate-keeper, the G+1 and G+3 residues of the hinge region. The energy correlation analysis highlighted this site as being an energy gate to the interaction pathway in the N-domain that is capable of transmitting an energetic perturbation to the active site of the enzyme via the G-loop and Lys118. Fourthly, SID analysis shows that it is at a complex multi-way interface that is important to the overall stability of the protein. Finally, the solid surface rendering shows it is surface accessible by an approaching ligand without the need for the enzyme to undergo large conformational changes to enable access.

4.7 Simulating the effect of ligand binding at the identified site

In order to study the conformational effects of ligand binding at this site on the enzyme, an MD simulation of a modelled DYRK2-ligand complex was performed. As there is no existing information about ligands that bind to this site, a small molecule probe was designed (figure 4.14A) for the identified binding site within the minimized average structure of the *DYRK2-apo* state. The probe was designed such that the pyridine ring was selected to hydrogen bond with the backbone nitrogen atom of Thr159 in β 7, the amide to interact via a hydrogen bond with Arg88 in the loop connecting β 3- β 4 and the cyclopropyl ring to occupy the hydrophobic cleft formed by residues Tyr45, Pro46, Ile48, Tyr49 and Tyr87 (figure 4.14B).

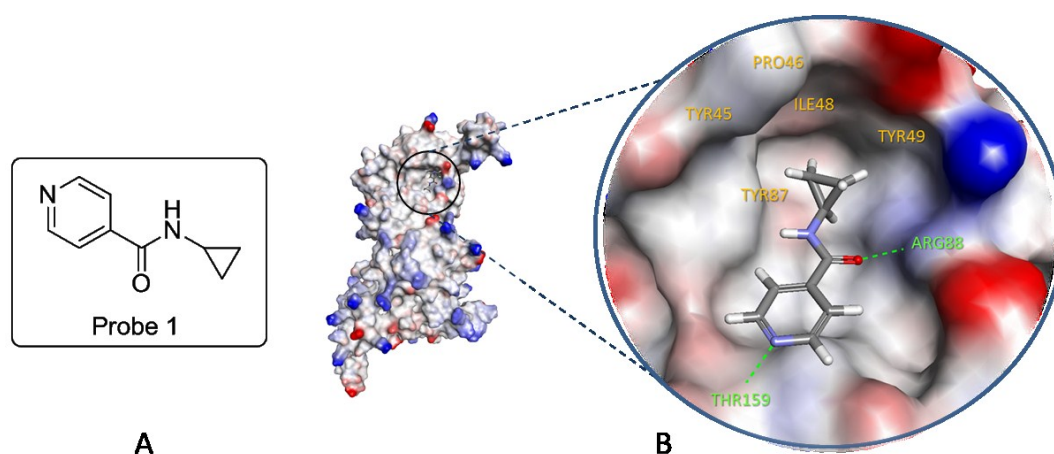


Figure 4.14: Design of the small molecule probe. (A) Chemical structure of probe-1. (B) The manually docked probe showing the interactions that it form with the putative allosteric site.

By establishing these interactions, the probe was intended to restrain the flexibility of the putative binding site, especially β 3, β 4 and β 7 (β 4 forms the roof of the ATP binding site and includes the G-loop at its end), thereby affecting the functional dynamics of the ATP binding site residues, which can ultimately affect the activity of the enzyme. Using the different tools and protocols available in DS, different conformations of the sketched probe were generated (only three were generated due to rigidity of the molecule) and docked into the site using CDOCKER as described in the materials and methods chapter. All conformations of probe-1 were docked to

ensure the optimum interactions between the probe and the enzyme were identified and to provide a reliable binding pose to serve as a starting model for the DYRK2-ligand complex (*DYRK2-probe-1*) for MD simulations (figure 4.15).

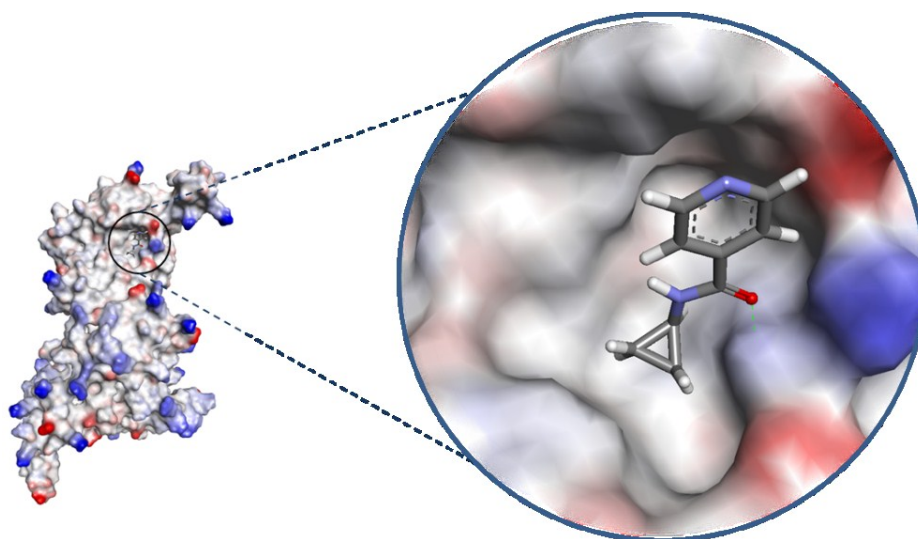


Figure 4.15: The docked pose used as a starting structure for the *DYRK2-probe-1* complex MD simulations.

After 7 ns the ligand (probe-1) detached from the binding site and traversed the surface of the protein until it finally bound underneath the T-loop (figure 4.16). This shows how weakly bound compounds can detach from their binding sites, particularly when they have very low flexibility, and unable to adapt to the conformational flexibility of the binding site.

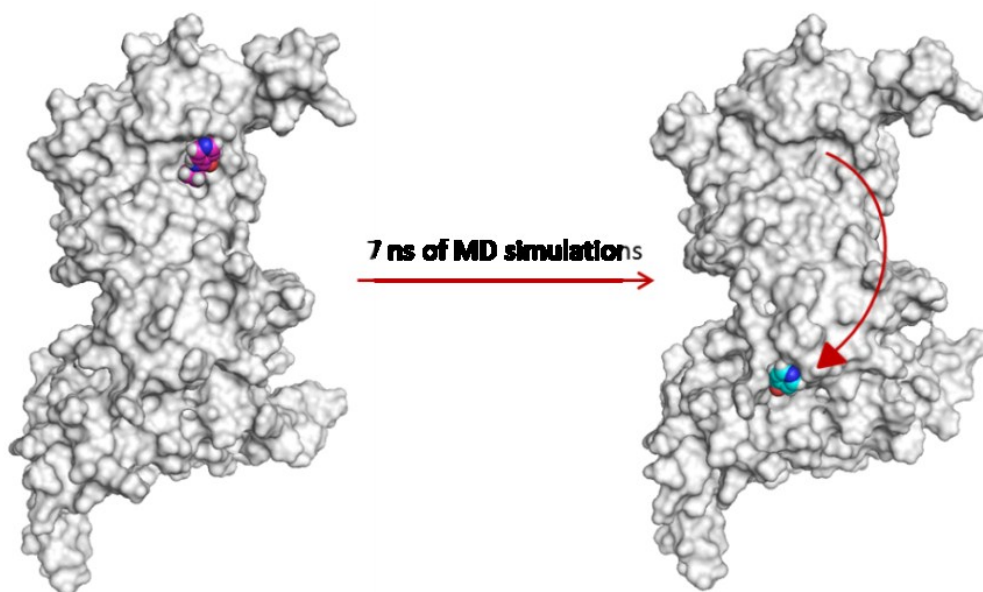


Figure 4.16: The detachment of probe-1 from the binding site after 7 ns.

To prevent this behaviour, another probe (probe-2) with greater flexibility and an extra hydrogen bonding functional group was docked in a similar pose (figure 4.17). Conformations for this probe were generated using the *Generate Conformations* protocol in DS using default values for the protocol parameters (generating 139 conformations); these were then docked into the binding site using *CDOCKER* with the default parameters without simulated annealing.

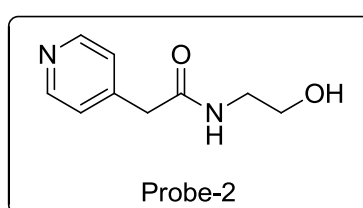


Figure 4.17: The chemical structures of the more flexible and complex probe, probe-2.

This probe showed good binding modes (figure 4.18). Because probe-2 showed good binding modes when docked in the identified binding site, it was structurally

modified in order to explore the hydrophobic region of the binding site by modifying the substituents on the pyridine ring (figure 4.19).

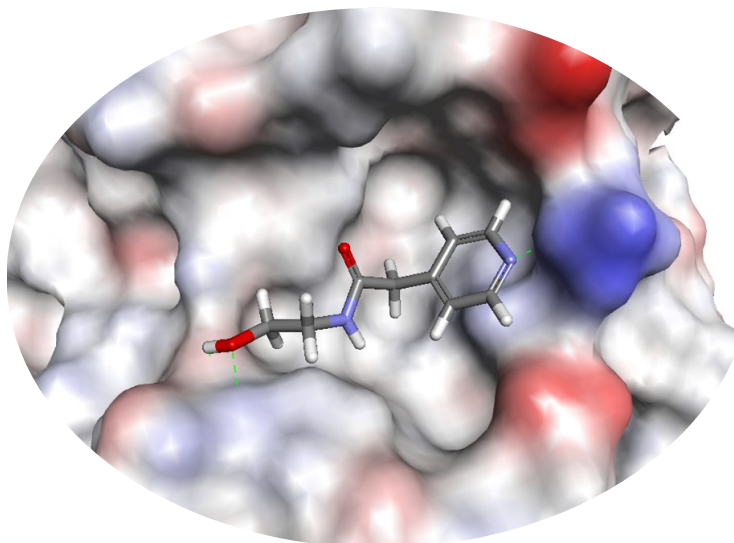


Figure 4.18: The binding mode of probe-2. The dashed green lines are the hydrogen bonds between the probe molecule and the protein.

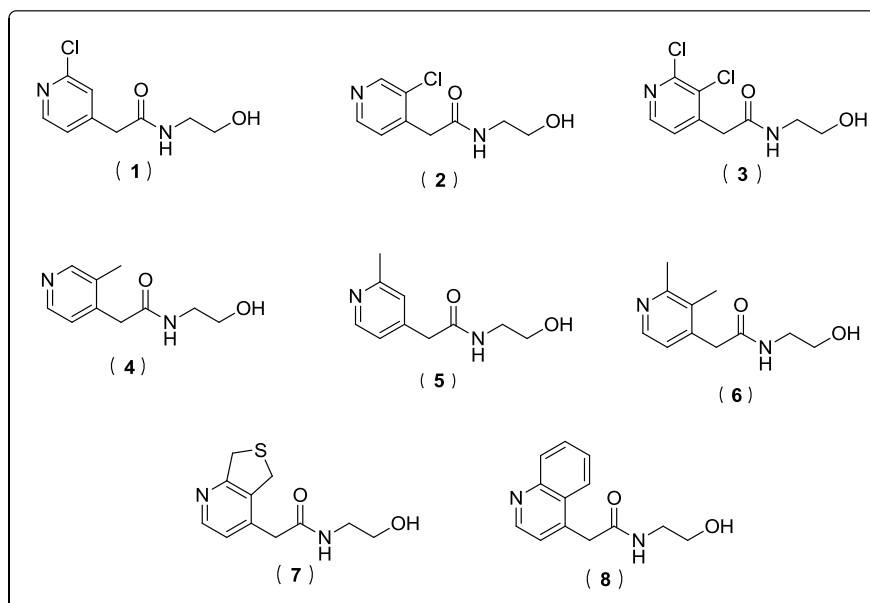


Figure 4.19: The chemical structures of the derivatives that were obtained by substituting the pyridine ring in probe-2.

All the derivatives of probe-2 were docked into the binding site using CDOCKER with default values for all parameters. Compound 6 (dimethyl substitution) was among the compounds that showed good binding modes, one of which was selected as a starting model for the MD simulation of the second artificial DYRK2-ligand complex (*DYRK2-com-6*) (figure 4.20).

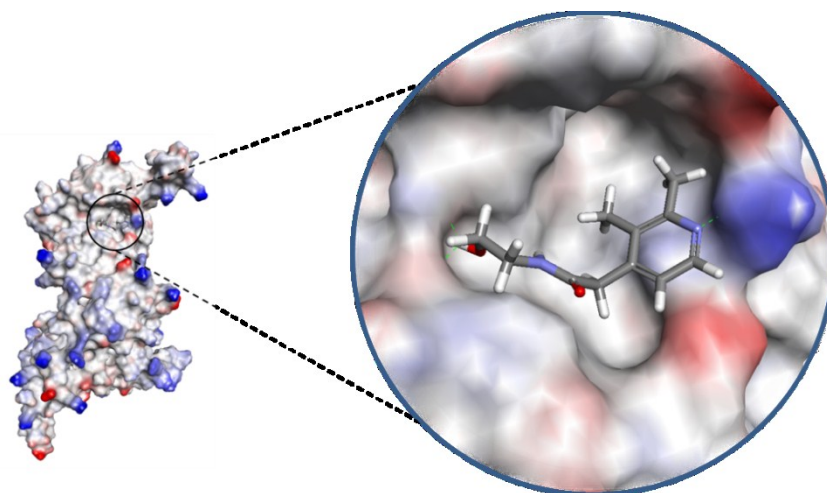


Figure 4.20: The docked pose of compound 6 used as a starting model for the second artificial DYRK2-ligand complex (*DYRK2-com-6*) MD simulation. The dashed green lines are the hydrogen bonds between the ligand and the protein.

4.7.1 Analysis of the MD results of DYRK2-ligand complexes

Considering the *DYRK2-apo* model as the native state of the enzyme, the effects of ligand binding at the identified site were assessed by comparing the results of the simulated *DYRK2-com-6* results with that of the simulated *DYRK2-apo* state. Firstly, the RMSD of the two systems was compared. This was followed by comparison of the minimised average structures, residual fluctuations, dynamic cross-correlation and the energy correlations.

Comparison of the RMSD values:

The RMSD values of the generated trajectory of the complex model from its minimised average structure was compared with that of the apo state to establish whether any notable changes in the behaviour of the simulated system upon ligand binding become evident over time. The RMSD values of the two trajectories were

calculated with reference to the corresponding minimised average structures (figure 4.21).

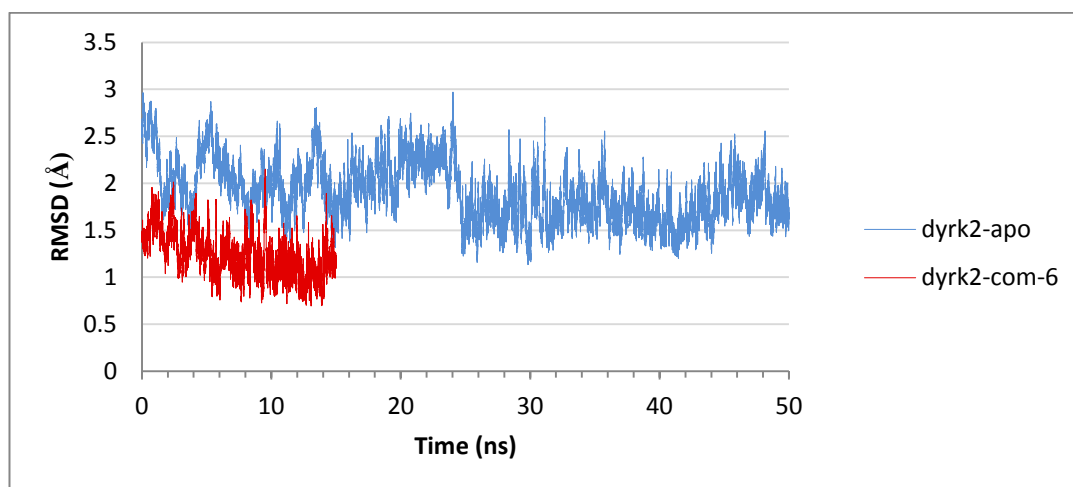


Figure 4.21: Comparison of the RMSDs values of the two simulated states of DYRK2. The RMSD values of each trajectory were compared with reference to their corresponding average structure. The blue trace corresponds to the *DYRK2-apo* state and the red one to the *DYRK2-com-6* model.

It is clear that the complex state of the enzyme has experienced a notable change in its RMSD values. The average values for *DYRK2-com-6* was 1.24Å, compared to 2.06Å for the *DYRK2-apo* state. These results imply that ligand binding in the identified site has had a restraining effect on the protein.

Comparison of the minimized average structures:

In order to check for any conformational changes resulting from ligand binding at the identified site average structures for the complex model was calculated and compared with that of the apo state (figure 4.22).

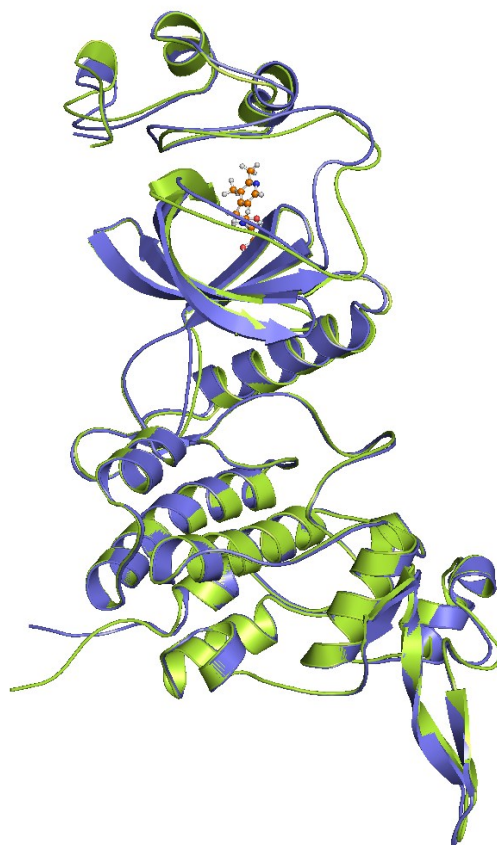
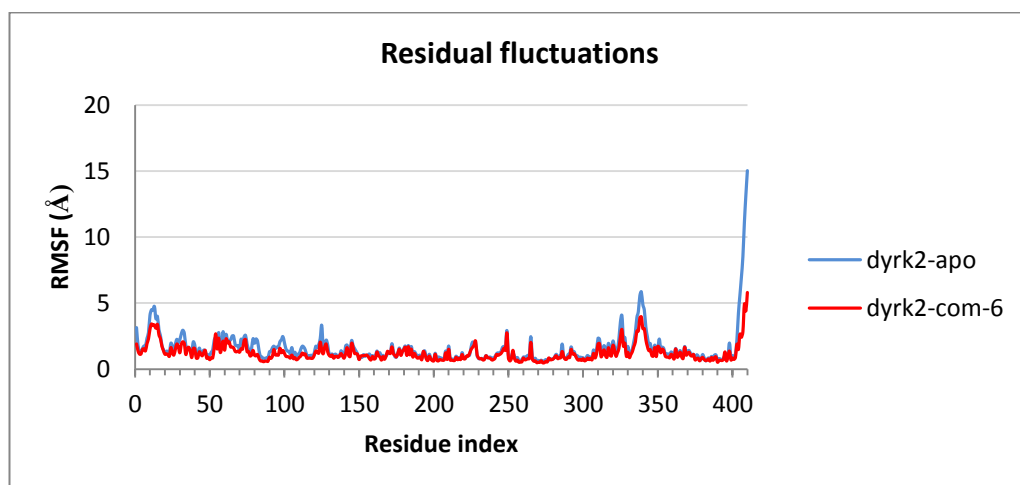


Figure 4.22: Cartoon representation of the backbone atoms of the minimised average structures of the *DYRK2-apo* (blue) and *DYRK2-com-6* (green) showing the regions of major conformational changes resulting from ligand binding. Compound 6 is depicted in ball and stick.

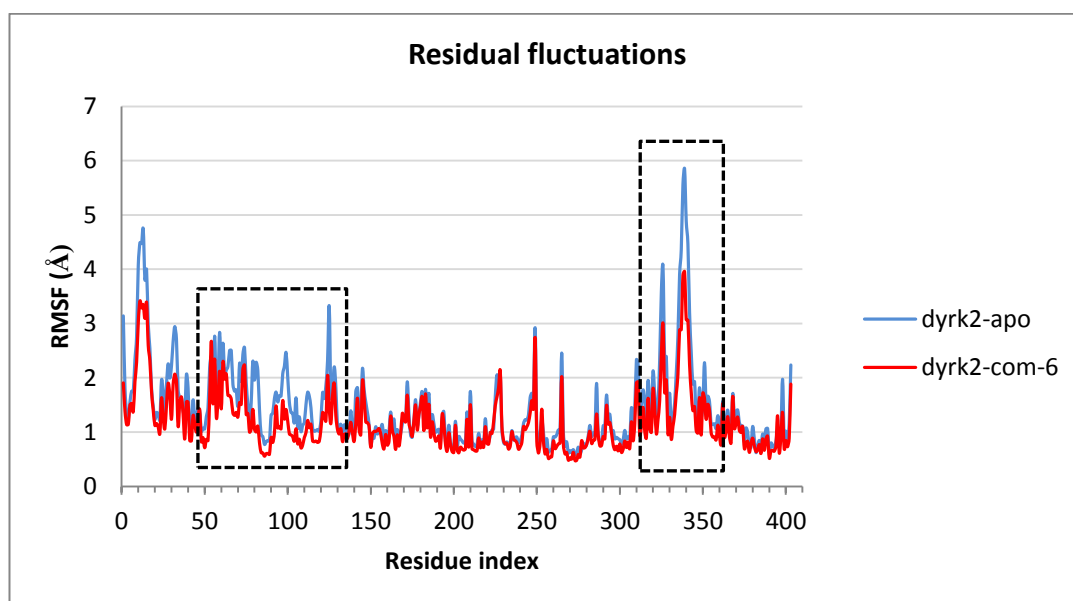
Figure 4.22 shows that the major conformational changes in the backbone atoms of the enzyme that resulted from ligand binding are located in the N-terminal extension of the protein, and to a lesser extent in the C-terminal extension. This change may not be solely driven by ligand binding since these regions are inherently flexible; yet, ligand binding must have contributed to the overall conformational change.

Residual fluctuations:

In order to assess whether there was any restraining effects on the enzyme attributed to ligand binding, the residual fluctuations of the new simulated complex model was calculated and compared with that of the apo state (figure 4.23).



A



B

Figure 4.23: Comparison of the residual fluctuations of the simulated complex with that of the *DYRK2-apo* state. (A) Comparison of the original RMSF values of *DYRK2-com-6* (red) with that of the *DYRK2-apo* (blue) states. (B) Comparison of the same RMSF values after excluding the last seven, extremely flexible residues. The regions of the protein that have had the highest changes in RMSF values are highlighted by dashed black rectangles.

Figure 4.23A shows that the C-terminus of *DYRK2* is very flexible. Therefore, the seven terminal residues (which have extremely high fluctuation values) were excluded from the plot to make the comparison simpler (part B). Part B shows the

regions of the protein that have experienced the largest reduction in fluctuation values upon ligand binding (highlighted by dashed black rectangles). These correspond to protein segments comprising residues 63-126, 325-327, and 337-344. These regions were highlighted on the average structure of the *DYRK2-com-6* model and are presented in figure 4.24.

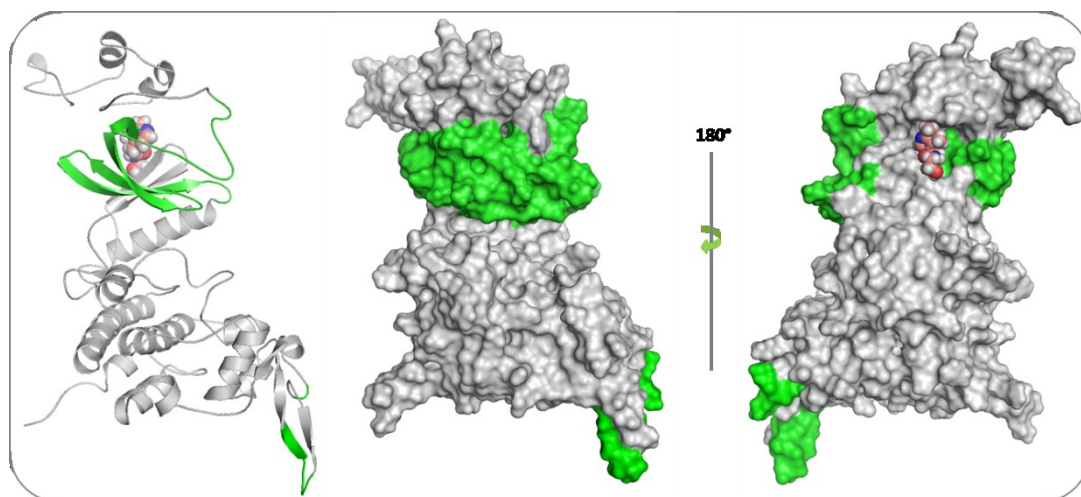


Figure 4.24: Structural mapping of the regions that experienced a reduction in their RMSF values upon ligand binding. (A) Cartoon representation of the backbone atoms. Regions with reduced flexibility are shown in green and the rest of the protein is in grey. Compound 6 is depicted in CPK pink spheres. (B) Solid surface representation of A. (C) rear view of the protein obtained by 180° rotation of B about the axis shown.

Figure 4.24 shows that the regions with restrained flexibility include the nearby binding site which involves β -sheets 4-6 that encompass the G-loop. The flexibility of the G-loop is important since it moves to achieve optimal binding with the ATP molecule. There is also another region in the C-domain that includes parts of β 12 and the loop connecting it with β 11, along with a small part of the loop connecting α 11 with β 11.

Dynamic cross-correlation:

Comparison of the cross-correlated motion matrices for each of the two trajectories revealed that ligand binding had resulted in a clear attenuation in the magnitude of the correlation among the residues in all clusters (figure 4.25).

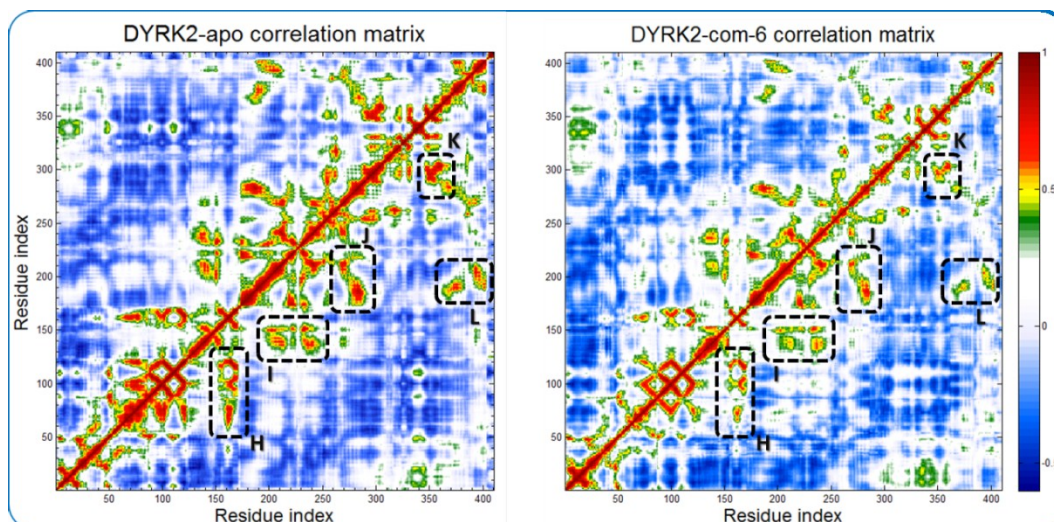


Figure 4.25: The dynamic cross-correlation matrices represented as heat maps. *DYRK2-apo* state is shown in the left and *DYRK2-com-6* in right. Clusters of main off-diagonal residue-residue correlations are highlighted by dashed black boxes.

The heat maps of the two DYRK2 models in figure 4.25 show that the cross-correlation pattern in the *DYRK2-apo* state is generally conserved in the complex model. However, the number of correlated residues and the magnitude of cross-correlation in the off-diagonal clusters of the complex model have been attenuated upon ligand binding. The total number of correlations in the *DYRK2-apo* state was 83845 of which 4718 were high correlations (above 0.554); while in the *DYRK2-com-6* state, there were 3007 high correlations (more than a 36% reduction in the number of correlated residues). This means that ligand binding at the putative allosteric site which is located at the interface between two correlated segments of the protein (cluster H) has affected the dynamics of not only the sub-domain where it binds but also the global dynamics of the enzyme; and since the well-orchestrated dynamical motions of proteins are important for their proper function, these findings strongly suggests that ligand binding at this site is likely to affect the function of the enzyme.

Energy correlation:

The final piece of analysis is the energy correlation. In order to explore the effect of ligand binding at the identified site on the overall energy correlation patterns in the enzyme, the energy correlation matrix for the complex model was calculated based on the largest seven eigenvalues. The average correlation value for each residue in the protein was calculated from the matrix and plotted versus residue index and compared with that of the apo state. Figure 4.26 compares the average energy correlations of the residues in the *DYRK2-apo*, and *DYRK2-com-6* states.

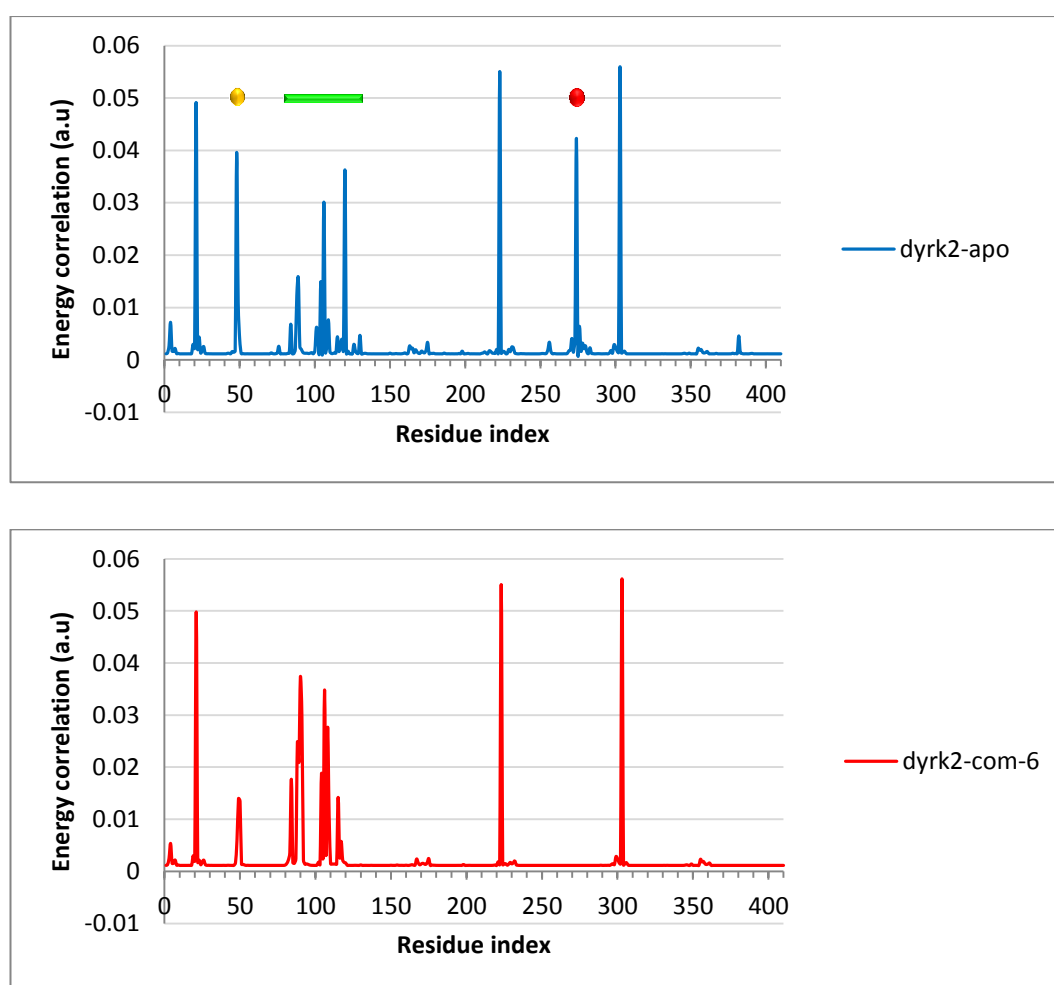


Figure 4.26: Comparison of the energy correlations (represented as peaks in the plots) of the *DYRK2-apo* state (blue), and *DYRK2-com-6* model (red). The correlations are given in arbitrary units. The coloured spheres and bar highlight the largest differences between the apo state and the complex model.

The peaks in figure 4.26 correspond to residues that have high energy correlation average values with other residues in the protein, and constitute the energy gates and the interaction pathways in the protein that are highly responsive to perturbations. Comparison of the *DYRK2-com-6* plot with that of the *DYRK2-apo* shows that ligand binding has introduced obvious changes in the plot. For example, the peak that corresponds to residue 48 (highlighted by a yellow sphere) has been suppressed dramatically; the peaks in the region extending from residue 110-120 (highlighted by the green bar) have also changed significantly, with some peaks disappearing and new ones appearing. The peak corresponding to residue 274 (highlighted by a red sphere) has completely disappeared. A better perception of how ligand binding has induced these dramatic changes in the energetic networks of the enzyme can be achieved by having a 3D representation of the residues with high correlation averages in all models structurally mapped onto the backbone of their corresponding average structure (figure 4.27).

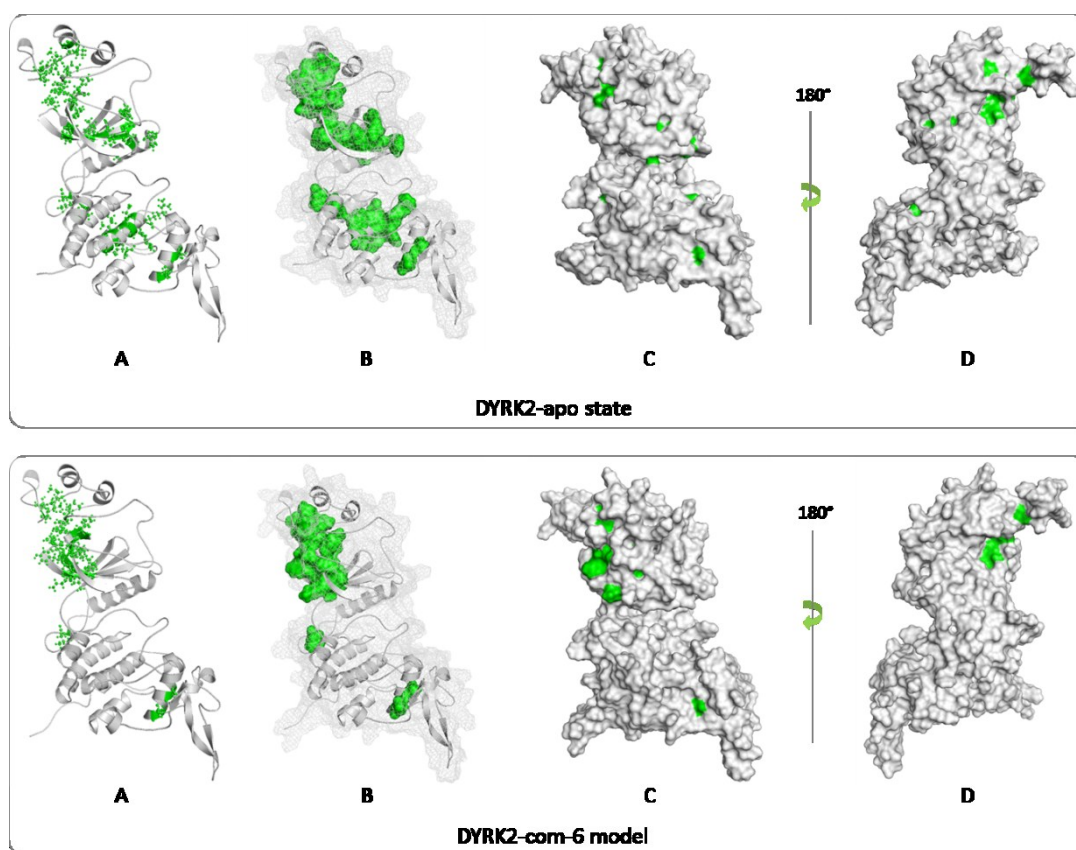


Figure 4.27: Structural mapping of the residues that define the interaction pathways in *DYRK2-apo* (upper panel), and *DYRK2-com-6* (lower panel). (A) Residues

forming the interaction pathways are represented by green balls and sticks and the rest of residues are coloured grey. (B) Surface representation of the residues forming the interaction pathway within a mesh surface of the entire protein. (C) Solid surface representation of all residues in the protein showing the energy gates on the surface of the protein. (D) Back view of protein surface in (C) by 180° rotation about the axis shown.

Figure 4.27 clearly shows that ligand binding has significantly affected the energy correlation profile in the protein. Comparison of the complex model with that of the apo state shows that the interaction pathway in the C-domain of the enzyme has almost disappeared. Regarding the other interaction pathway in the N-domain, ligand binding has severely affected the pathway in such a way that it is no longer involves the G-loop, Lys118 or any of the α C-helix residues. Altogether, this suggests that ligand binding has resulted in a massive disruption of the energetic interaction network of the enzyme and is expected to have an effect on its function, particularly on ATP binding.

Based on these promising results, a search of commercial chemical libraries was carried out to identify small molecules which could be screened against DYRK2. The minimised average structure of the *DYRK2-com-6* state was used as the structural model to perform virtual screening and docking studies. This model was used rather than the DYRK2-apo state because running the MD simulation with the probe had achieved the intended effect expanding the allosteric site to a state more likely to accept small molecule ligands for binding, thus expanding the possibility of identifying new hits.

4.8 Virtual screening and molecular docking

4.8.1 Structure-based pharmacophore generation

An efficient way of searching small molecules databases is to create a pharmacophore that contains all pharmacophoric features that correspond to chemical functionality deemed important for effective ligand binding. Therefore, the identified putative allosteric binding site in DYRK2 was used to generate a 3D structure-based pharmacophore model to be employed in the virtual screening (VS) of small molecules databases (DB). The minimised average structure of the *DYRK2-com-6* system was used as a structural model to generate the 3D pharmacophore using the *Interaction Generation Protocol* available in DS as described in the methods section. The binding site was defined with a sphere that encompasses all amino acid residues that could contribute to effective ligand binding (figure 4.28).

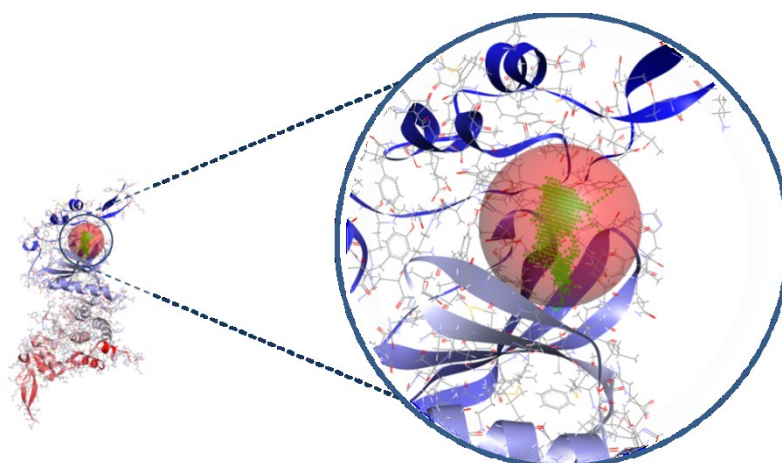


Figure 4.28: DS definition of the binding site. The binding site is defined by the transparent red sphere. The green dots correspond to the cavity of the site of interest that has been identified by the flood filling algorithm.

The binding site was analysed by applying the *Interaction Generation* protocol, which uses the Ludi algorithm to generate an interaction map of the binding site to include a list of all features (hydrogen-bond donors, hydrogen-bond acceptors, and hydrophobic points) that can be complemented by a ligand in order to have a reasonable interaction with the protein. A set of 3D pharmacophoric queries was then

derived from the interaction map (figure 4.29). The features were then clustered and edited where the most important features were selected and included in the construction of the primary pharmacophore model (figure 4.30A).

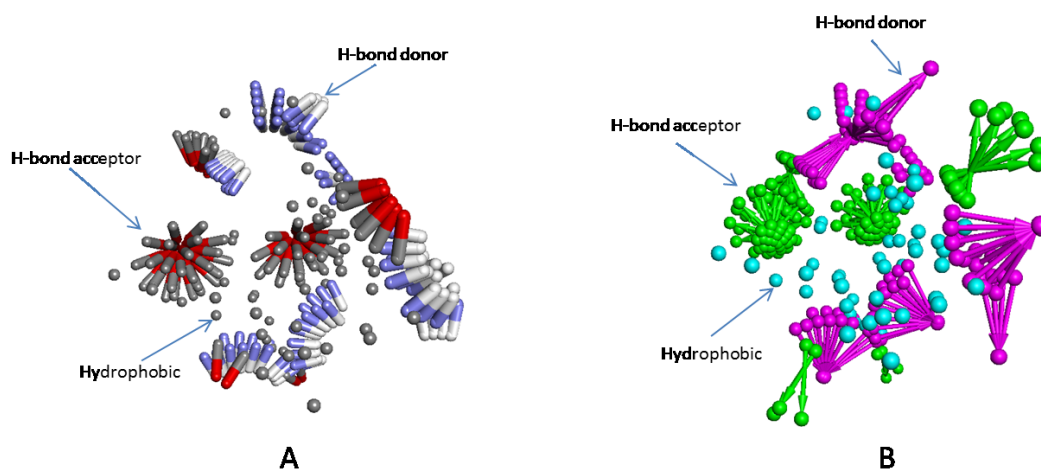


Figure 4.29: Structure based pharmacophore generation. (A) Ludi interaction map. (B) Interaction features are converted into pharmacophoric features. In the interaction map hydrogen-bond donors are in blue, hydrogen-bond acceptors are in red, and the hydrophobic regions are in grey; in the pharmacophoric features they are magenta, green, and cyan respectively.

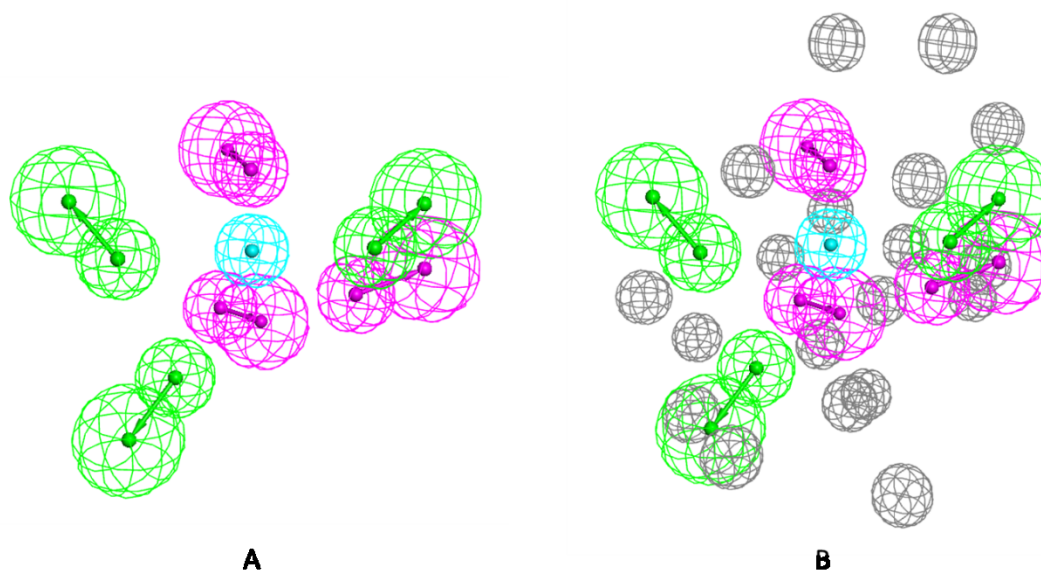


Figure 4.30: The generated structure-based pharmacophore model from the putative allosteric site in DYRK2. (A) The primary pharmacophore. (B) The pharmacophore

with the exclusion constraints (excluded volumes). The feature types in this pharmacophore are: hydrogen-bond donor (magenta), hydrogen-bond acceptor (green), hydrophobic (cyan), and excluded volumes (grey).

The generated 3D pharmacophore consisted of seven features: three hydrogen-bond donors, three hydrogen-bond acceptors, and one hydrophobic region. Functional groups of candidate ligands must map all the pharmacophoric features and reside within the spatial tolerance spheres in order to be retrieved as hits.

To account for the steric interactions with the target protein, excluded volumes were added to the generated pharmacophore as described in the methods section; these define regions within the binding site that a ligand may not overlap. This makes the search query more specific and excludes ligands that would clash with protein atoms in the binding site (figure 4.30B).

The generated pharmacophore comprises all features that are important for ligand binding, but having seven features made it so selective that when it was used to screen the Maybridge database, it returned only two hits. It was therefore fragmented into three smaller pharmacophores which complement each other and allow for thorough screening of the database. Pharmacophore number one is composed of four features: two hydrogen-bond donors and two acceptors; pharmacophore number two is made of four features: two hydrogen-bond donors, one acceptor and a hydrophobic feature; and the third pharmacophore comprises five features: two hydrogen-bond donors, two acceptors and a hydrophobic feature. The excluded volumes were kept intact in all of the three sub-pharmacophores (figure 4.31). The three pharmacophores were then used in the virtual screening of Maybrige database for possible hits.

Usually it is good practice to validate pharmacophores before using them to screen databases. The validation is performed by using a set of ligands with known binding modes or binding affinities. In this instance, there is no experimental information about ligands that bind to this site, and so they were used without standard validation.



Figure 4.31: The three sub-pharmacophores obtained by fragmenting the primary pharmacophore. The features are: hydrogen-bond donor (magenta), hydrogen-bond acceptor (green), hydrophobic (cyan), and excluded volumes (grey).

4.8.2 Virtual screening of commercial databases

The Maybridge 2009 Database [158] was searched for hits that fit to the identified binding pocket. The *best flexible* search method in the *Search 3D Database* protocol within DS was used. More than 56,000 compounds were screened and all ligands that mapped any of the three pharmacophores were returned. Pharmacophore one returned 725 hits, pharmacophore two returned 1,543 hits, and pharmacophore three returned 719 hits. The first retrieved hit in each of these pharmacophores is mapped upon its corresponding pharmacophore in figure 4.32. The total number of returned hits was 2,987 of which 869 were duplicates that were filtered out.

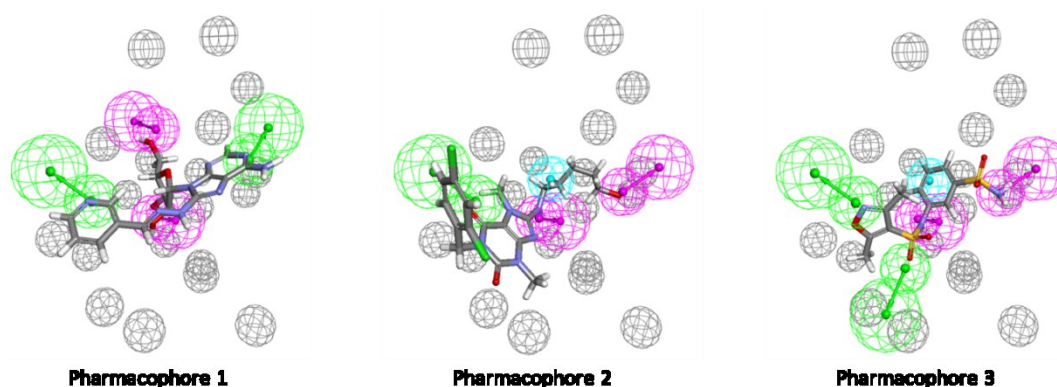


Figure 4.32: Overlay of the first retrieved hit from the Maybridge DB defined by each pharmacophore. The features are hydrogen-bond donor (magenta), hydrogen-bond acceptor (green), hydrophobic (cyan), and excluded volumes (grey).

Pipeline Pilot (PP) was used to filter the retrieved hits based on Lipinski's rule of five for drug like properties and consideration of the fit values. For a compound to obey Lipinski's rule it has to have a molecular weight of less than 500; less than 5 hydrogen-bond donor groups; less than 10 hydrogen-bond acceptor groups; and a partition coefficient (LogP) of less than 5. The threshold for the fit value used to filter the retrieved hits was set to be more than 2.0. The fit values of the retrieved hits are based on how well the hits map the pharmacophoric features and whether they deviate from the centre of the feature or not; the better the fit, the higher the fit value. Hits that passed all filtration criteria were 116 and they were selected for molecular docking (figure 4.33).

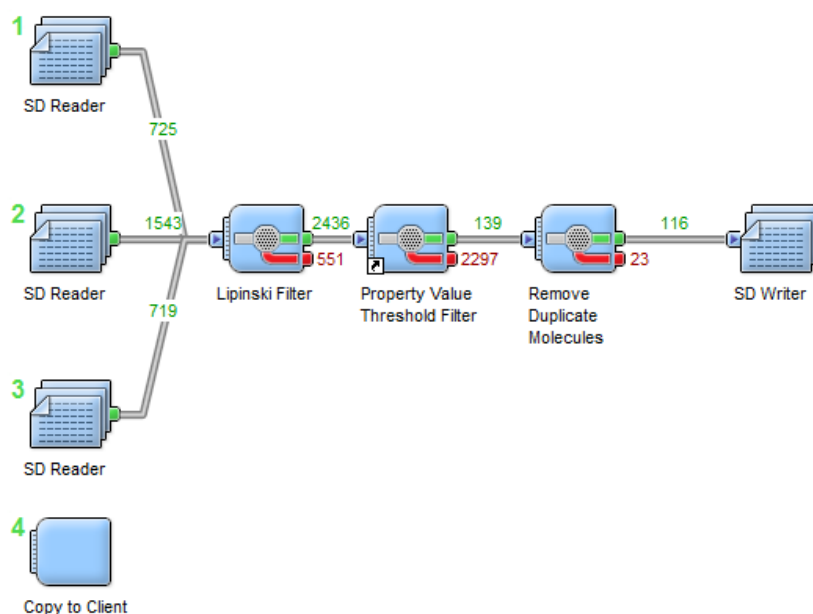


Figure 4.33: A pipeline showing the filtration of the retrieved hits based on Lipinski's rule of five and a fit value of more than two using Pipeline Pilot.

4.8.3 Molecular Docking

Molecular docking of the filtered hits was performed using GOLD version 5. The minimised average structure of the *DYRK2-com-6* system was used to define the binding site for docking purposes (figure 4.34). The binding site in GOLD was defined as described previously.

There are five different scoring functions within Gold; GoldScore, ChemScore, CHEMPLP, ASP, and User defined score; in this study GoldScore was used. Gold fitness scores account for protein-ligand hydrogen-bonding and van der Waal interactions as well as intramolecular van der Waals and torsional strain energy for the ligand [94]. The minimised average structure of the *DYRK2-com-6* complex was used and the target site was defined. The binding site was initially too large, and the hydrophobic fitting points were pruned by deleting the undesirable points distant from the site. The edited hydrophobic fitting points were then used in the GOLD configuration file to dock the 116 filtered hits (figure 4.35). Fitting points in Gold are generated automatically by placing a fine grid over the binding site. A bare carbon atom is placed at each grid point to measure the van der Waals interactions in order to construct a map of positions on which a hydrophobic ligand atom will have a favourable interaction [183].

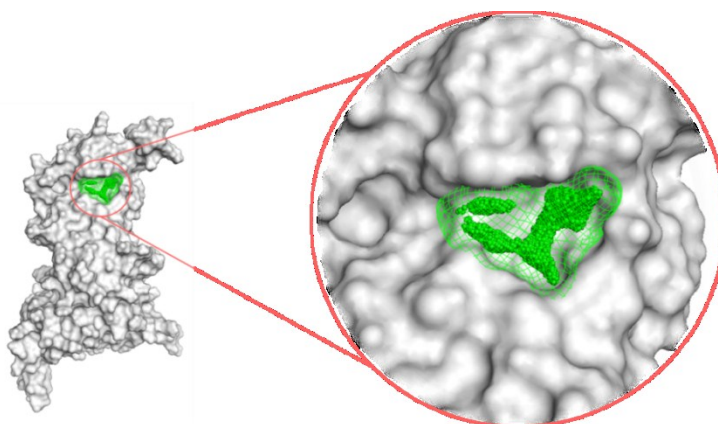


Figure 4.34: GOLD definition of the binding site used for docking the filtered hits that were retrieved from the Maybridge database.

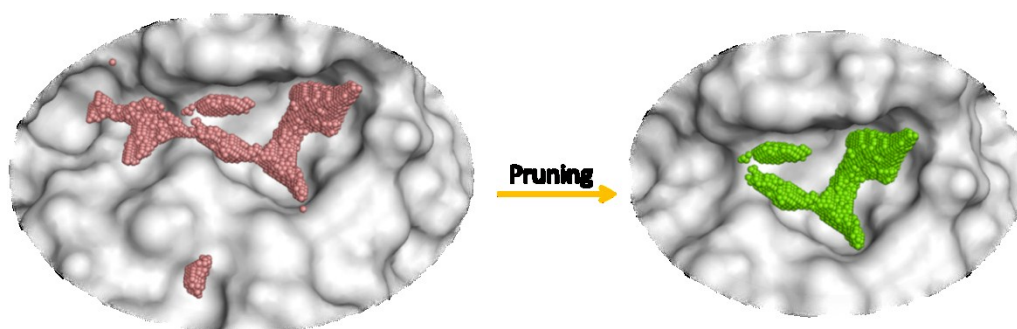


Figure 4.35: Pruning of the hydrophobic fitting points to fit the site of interest.

The resulting docked poses were ordered according to their fitness scores, so the best solution is placed first (figure 4.36). Ligands with strong interactions with the binding site and reasonable chemical structure (for example all non-aromatic hits were discarded) were selected for experimental evaluation of their binding affinities. Out of the 116 hits 48 compounds were selected and purchased from the vendor (table 4.2). The binding affinities of the 48 compounds were evaluated using the differential scanning fluorimetry (DSF) at the Structural Genomics Consortium (SGC), Oxford. Figure 4.37 summarizes the entire virtual screening workflow from the generation of the pharmacophore to the experimental validation of the selected hits.

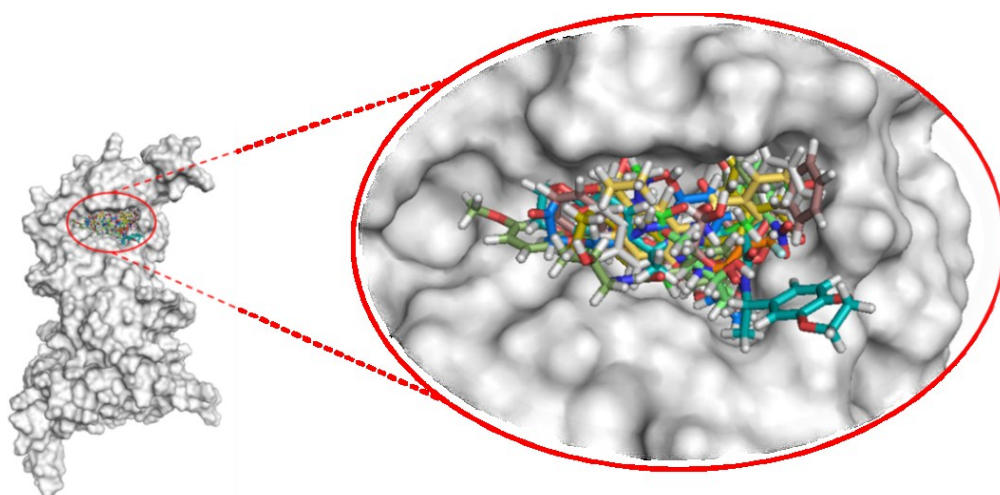


Figure 4.36: The top ranked docked pose of the first ten retrieved hits. Ligands are shown in sticks and protein surface is in grey.

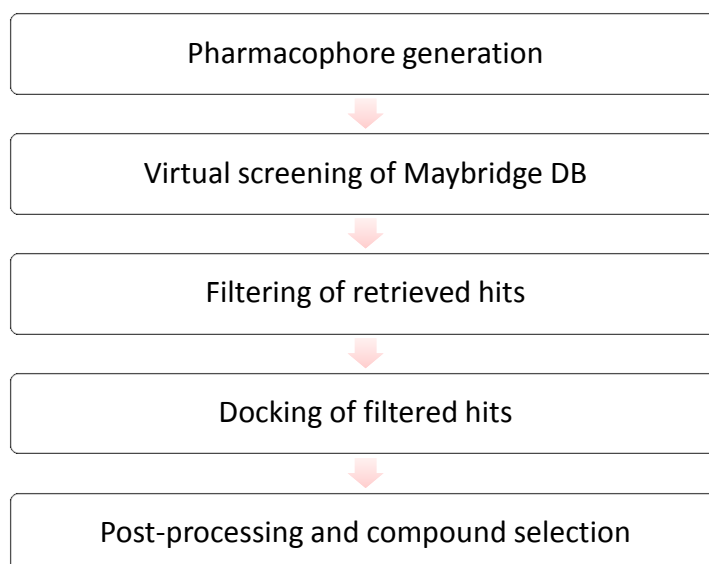


Figure 4.37: The virtual screening workflow. Pharmacophore generation allows large numbers of compounds to be rapidly screened for hits. Hits are then filtered for desirable chemical properties and then docked to provide a more reliable assessment of their potential affinity for the target. Finally, post-processing and manual inspection removes any erroneous results and ensures that sensible compounds are selected for real testing.

Table 4.2: The 48 compounds purchased for experimental testing of their binding affinities.

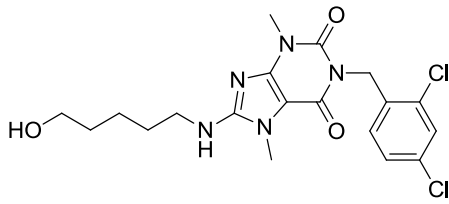
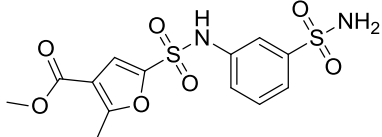
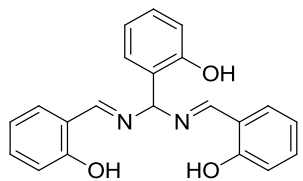
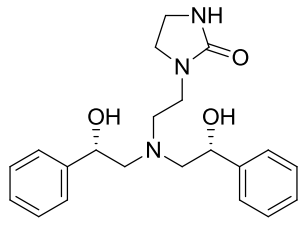
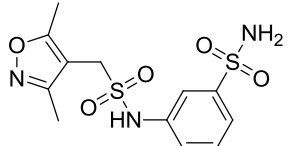
Index	Structures	Name	Fit value	Molecular weight
1		HTS11186	3.175	440.324
2		HTS06652	3.117	374.389
3		S04968	3.088	346.379
4		HTS02812	2.994	370.465
5		HTS05510	2.959	331.368

Table 4.2: Continued...

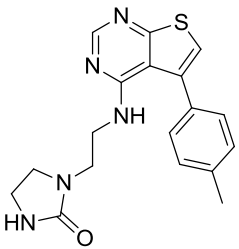
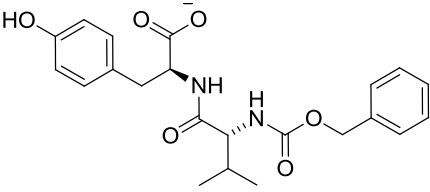
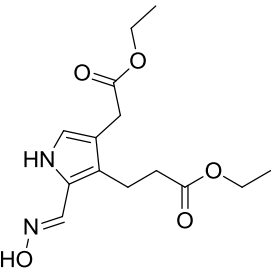
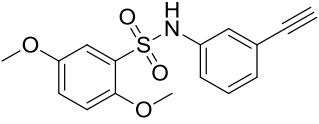
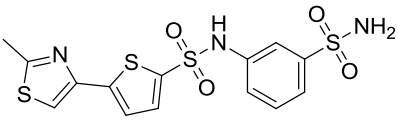
Index	Structures	Name	Fit value	Molecular weight
6		HTS03371	2.885	353.441
7		BTB15187	2.866	413.444
8		NRB02765	2.806	296.319
9		CD07857	2.802	317.36
10		HTS06647	2.72	415.531

Table 4.2: Continued...

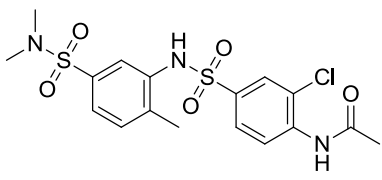
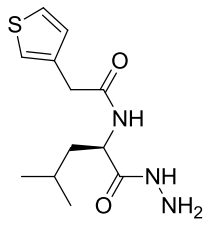
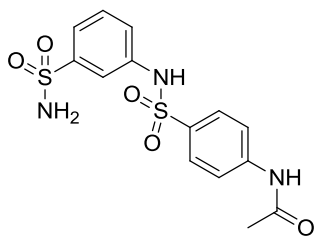
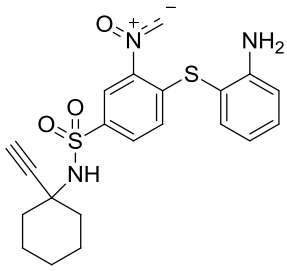
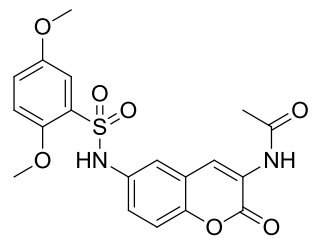
Index	Structures	Name	Fit value	Molecular weight
11		HTS05509	2.641	445.941
12		KM02891	2.566	269.363
13		HTS06639	2.513	369.416
14		DSHS00855	2.505	431.528
15		BTB04790	2.503	418.42

Table 4.2: Continued...

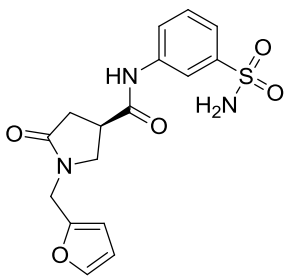
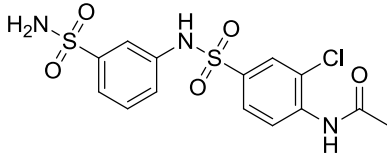
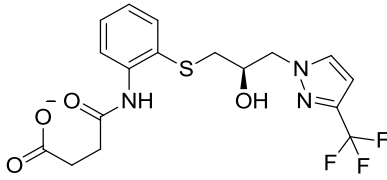
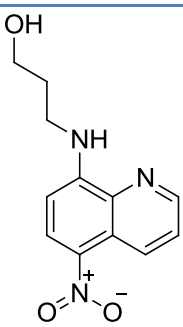
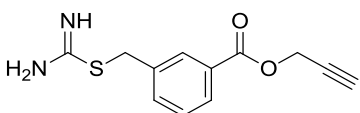
Index	Structures	Name	Fit value	Molecular weight
16		HTS09688	2.477	363.388
17		HTS05511	2.43	403.861
18		HAN00305	2.412	414.379
19		RJC03509	2.315	247.25
20		S07331	2.301	248.301

Table 4.2: Continued...

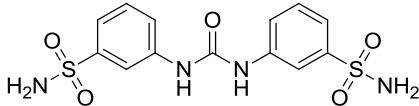
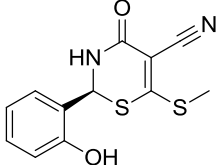
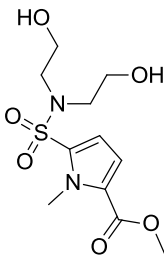
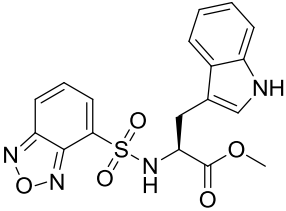
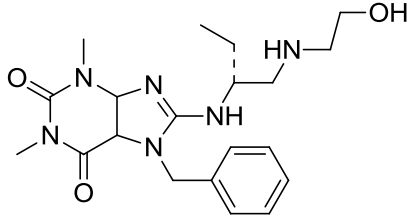
Index	Structures	Name	Fit value	Molecular weight
21		RJC00584	2.275	370.404
22		CD04067	2.271	278.35
23		HTS01746	2.268	306.335
24		RH01614	2.268	400.408
25		RJC03502	2.233	401.483

Table 4.2: Continued...

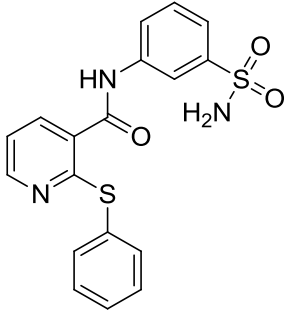
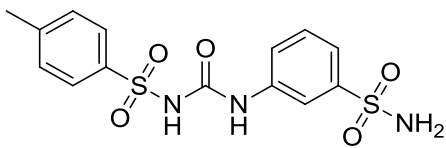
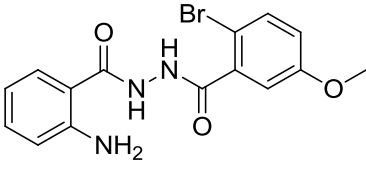
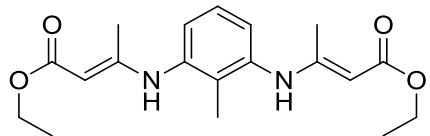
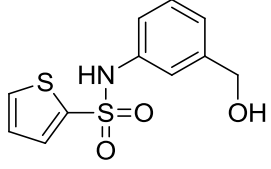
Index	Structures	Name	Fit value	Molecular weight
26		HTS06222	2.22	385.46
27		HTS05233	2.184	369.416
28		BTB06870	2.175	364.194
29		RH01034	2.175	346.421
30		HTS08047	2.169	269.34

Table 4.2: Continued...

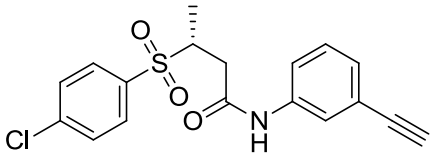
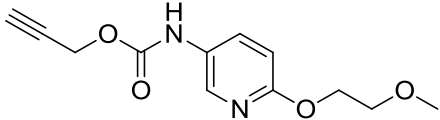
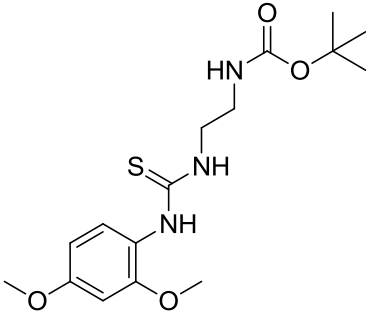
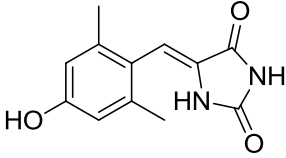
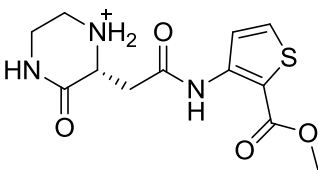
Index	Structures	Name	Fit value	Molecular weight
31		SEW00552	2.158	361.843
32		RF05238	2.144	250.251
33		AW00378	2.131	355.452
34		BTBS00039	2.122	232.235
35		HTS00383	2.114	298.338

Table 4.2: Continued...

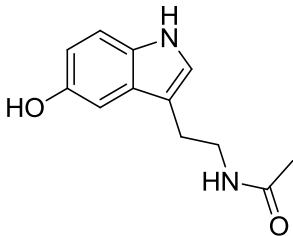
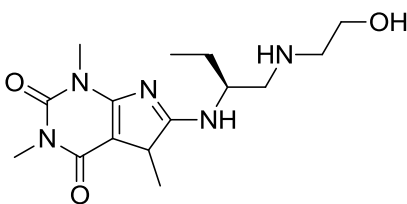
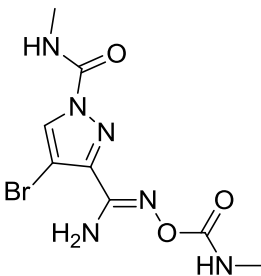
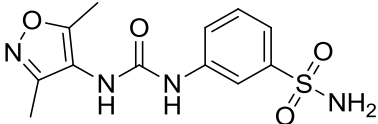
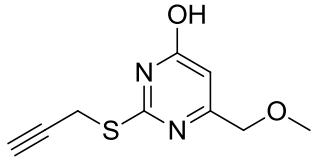
Index	Structures	Name	Fit value	Molecular weight
36		AC22693	2.114	218.252
37		RJC03501	2.102	325.387
38		RF05199	2.102	319.115
39		SCR01491	2.086	310.329
40		RJF00493	2.056	210.253

Table 4.2: Continued...

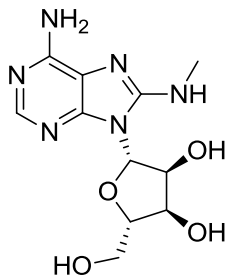
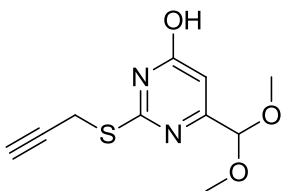
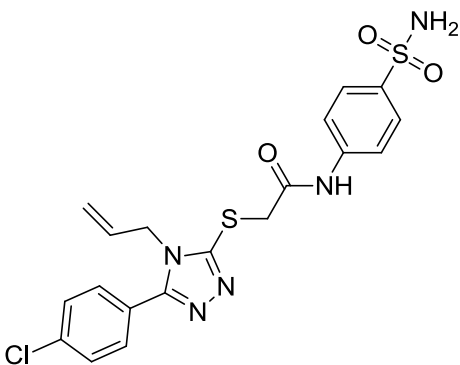
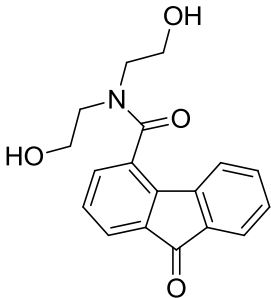
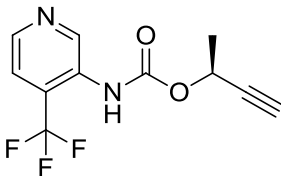
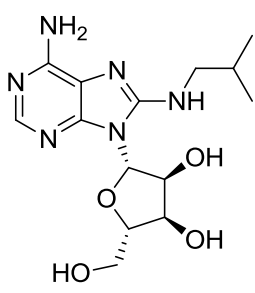
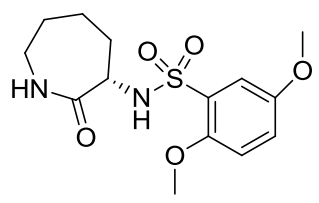
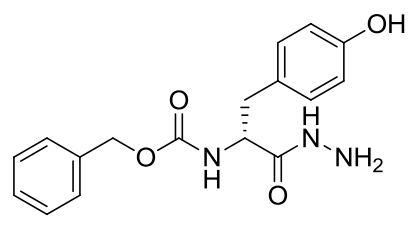
Index	Structures	Name	Fit value	Molecular weight
41		HTS11461	2.056	296.283
42		SPB05720	2.049	240.279
43		HTS13506	2.044	463.961
44		JFD03216	2.036	311.332
45		RF04724	2.024	258.197

Table 4.2: Continued...

Index	Structures	Name	Fit value	Molecular weight
46		HTS11216	2.02	338.362
47		SCR00973	2.01	328.384
48		BTB13337	2.029	329.35

4.9 Experimental evaluation of the binding affinities of the selected hits

4.9.1 Differential scanning fluorimetry (DSF)

DSF (also referred to as fluorescence thermal shift) is a fast and inexpensive method for the screening and identification of small ligands (can be small molecules, peptides or nucleic acids) that bind and stabilize purified proteins. Protein stability is related to its Gibbs free energy of unfolding, ΔG_u , which is temperature dependent, such that an increase in temperature will decrease protein stability and decrease ΔG_u until it reaches zero at the equilibrium point where the concentrations of folded and unfolded protein are equal. This equilibrium temperature is known as the *melting temperature* (T_m). Ligand binding in most cases increases protein stability via increasing its ΔG_u , which may increase T_m . The stabilizing effect of ligand binding is proportional to its concentration and binding affinity [161].

In the DSF method, the thermal unfolding of proteins is measured and monitored by incubating the target protein and ligand with an environmentally sensitive fluorescent dye that is highly fluorescent in a non-polar medium such as the hydrophobic regions of an unfolded protein, and quenched in aqueous environments. Therefore, the temperature at which a protein unfolds is measured by an increase in the fluorescence of the dye which has high affinity for the hydrophobic parts of the protein that become exposed upon unfolding. The fluorescence intensity is plotted as a function of temperature which gives a sigmoidal curve that can be described by a two-state transition (figure 4.38). The inflection point of the transition of the sigmoidal curve of fluorescence plotted against temperature defines the T_m . A high T_m indicates high protein stability. The transition midpoint in the presence and absence of a ligand is calculated and the difference in temperature (ΔT_m) between the two midpoints is a reflection of binding and is related to the binding affinity to which the ligand binds to the protein (figure 4.39). The most frequently used dye in DSF is SYPRO orange because of its favourable properties due to the high signal-to-noise ratio [161].

This method has been used in many studies and proved to be of high applicability. It has been used to identify stabilizing conditions that aid in protein crystallisation [184]; and also to rapidly screen libraries of various ligands against the proteins of interest for their binding affinities [185, 186]; and recently it has been utilised to

study enzyme inhibitor mode of action [187]. In general, DSF is used as a primary (yes/no) binding assay and then other methods are used to determine affinity values [188].

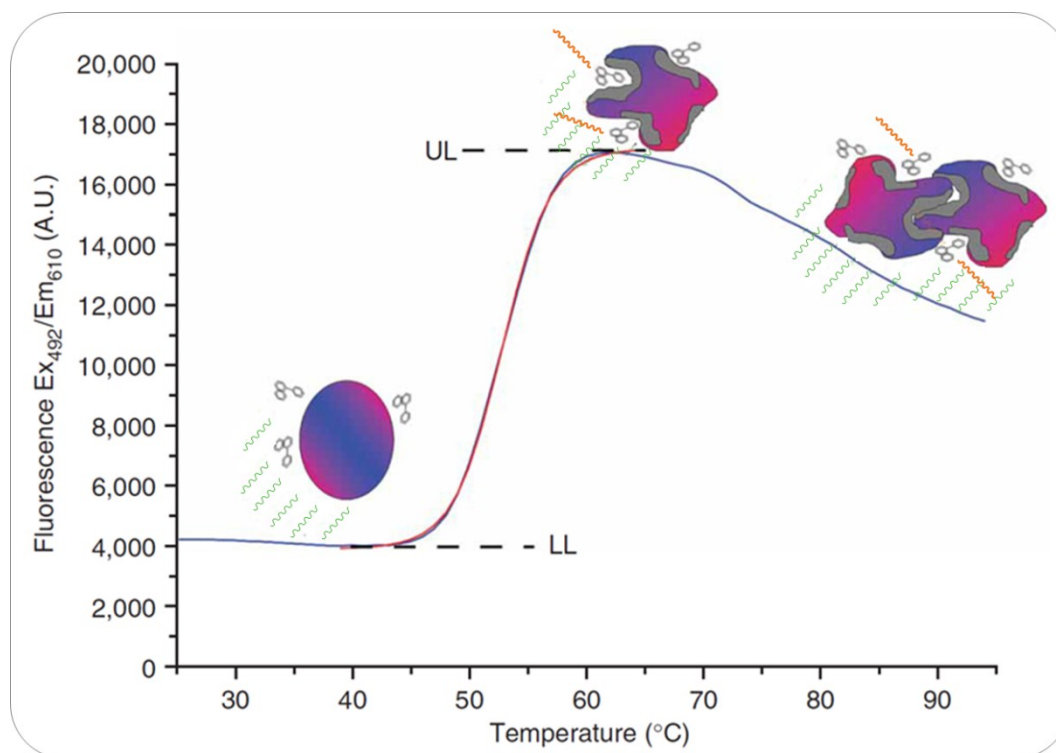


Figure 4.38: Schematic representation of the principle of the DSF method showing the recording of fluorescence intensity versus temperature for the unfolding of protein in the presence of a dye (depicted as three-ring aromatic molecule). In the case where the protein is intact (spherical shape at the baseline of the curve), a basic fluorescence intensity is excited by light of 492 nm (depicted as green curved lines). During the course of protein unfolding, the hydrophobic regions (in grey) become exposed, and strong fluorescent light of 610 nm (depicted as orange curved lines) is emitted by the bound dye molecules. The peak in fluorescence intensity is followed by a decline which is attributed to the removal of the protein from the solution because of precipitation and aggregation. LL and UL are the lower and upper level in fluorescence intensity respectively.

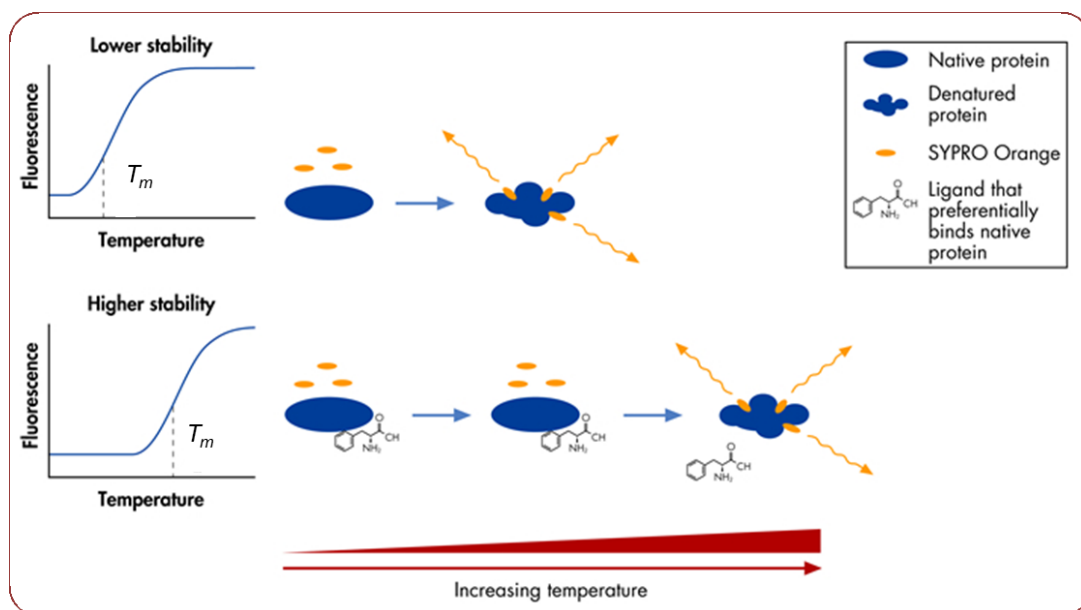


Figure 4.39: Schematic representation of protein stabilisation upon the addition of a compound that binds the native protein. The stabilising effect of the bound compound is manifested as an increase (shift) in the melting temperature (T_m) of the complex which allows the binding to be detected. Note the intense increase of the fluorescence of the dye (SYPRO Orange) upon unfolding of the protein as it binds the exposed hydrophobic patches of the protein. Adapted from [189].

Of the 48 compounds, four showed very good binding affinities based on their ΔT_m values (figure 4.40). Those compounds were HTS11216, HTS11461, RJC03509, and RF05199 and their chemical structures are shown in figure 4.41.

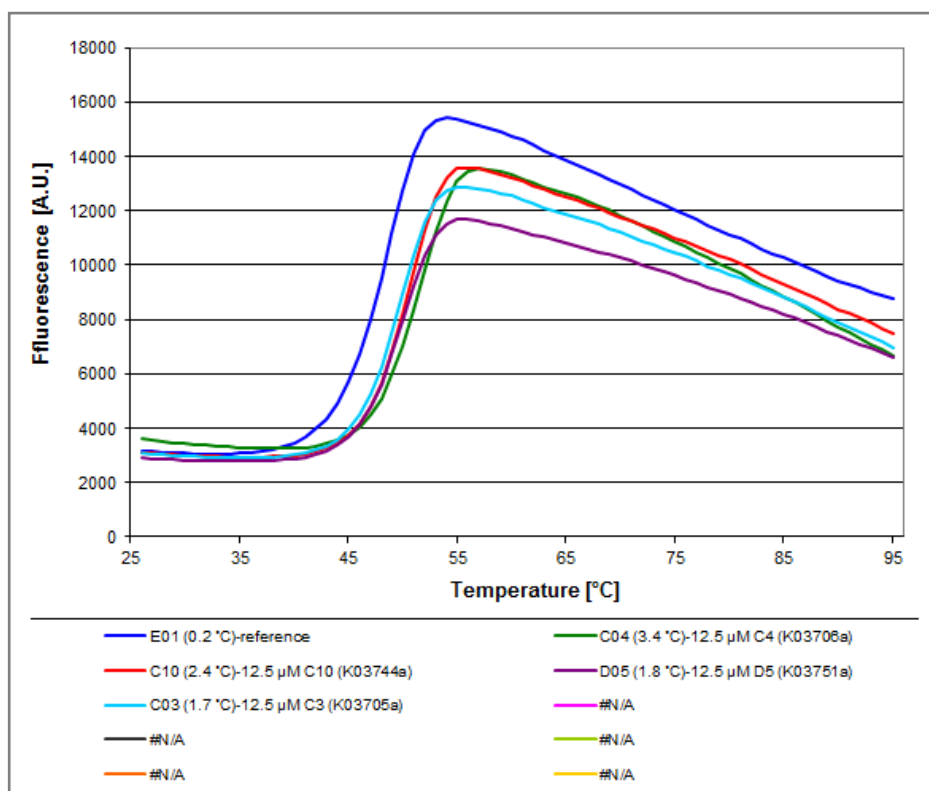


Figure 4.40: The DSF results for the 48 selected hits against DYRK2. The graph shows the curves that correspond to the four active hits. The blue line is the reference (no compound), red line corresponds to compound C10 (RF05199), cyan line corresponds to compound C03 (HTS11216), green line corresponds to compound C04 (HTS11461), and purple line corresponds to compound D05 (RJC03509).

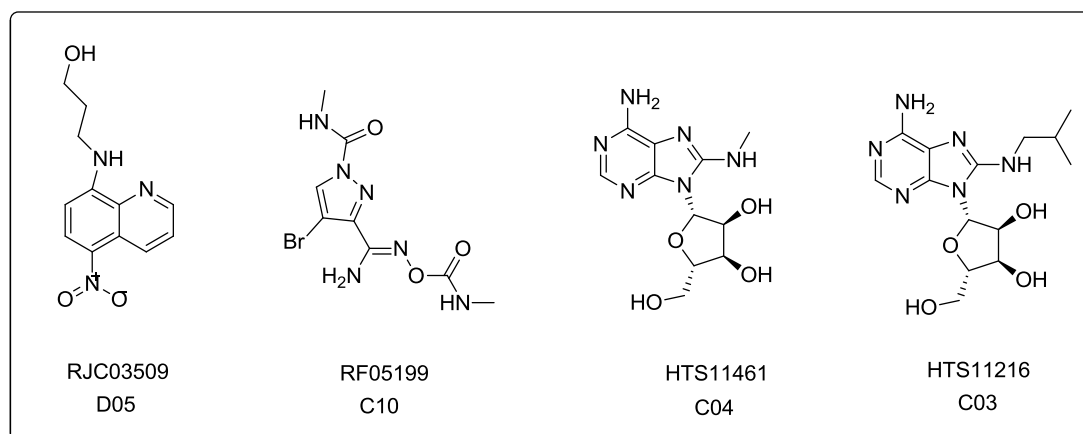


Figure 4.41: The chemical structures of the four compounds that showed good binding affinities in the DSF screening.

Figure 4.40 shows that the ΔT_m values of the four compounds ranged from 1.7-3.4 °C compared to the reference compound which has a ΔT_m value of 0.2 °C. In general, a ΔT_m value of ≥ 2 °C is considered promising, and values ≥ 4 °C have been found to correspond to IC_{50} 's of less than 1 μ M [161, 184].

Given the chemical structure of compounds C03 and C04 resembled potential ATP-site binders further investigation of the binding modes of these four compounds was carried out before any kinetic testing in order not to waste time and resources on compounds that could be false positives. Figure 4.42 shows a docked pose for each of the four compounds that lies within the top three scored poses along with a 2D ligand-receptor interaction diagram.

Figure 4.42 shows that all of the compounds had reasonable binding modes apart from compound HTS11461. Careful inspection of the 2D interaction diagram reveals noticeable differences in the type and total number of interactions between the ligands and the enzyme. Compound HTS11216 has one hydrogen bond with Glu156 and two Pi-Pi interactions with Tyr45; compound HTS11461 has two hydrogen bonds with Glu156 and Arg88 and a Sigma-Pi interaction with Phe158; compound RJC03509 has two hydrogen bonds with Ser44 and Arg88 and three Pi-Pi interactions with Tyr45; and compound RF05199 has three hydrogen bonds with Ser44, Tyr45 and Arg88 and a Pi-Pi interaction with Tyr45. RJC03509 and RF05199 seem to establish better binding modes and could potentially have better binding affinities compared to the other two. In addition, the four compounds were docked into the ATP binding site to compare the docked poses with their equivalents in the putative allosteric binding site. Appreciation of the nature of interactions and binding modes of the compounds in the two binding sites would help guide the prioritization by studying their kinetics (figure 4.43).

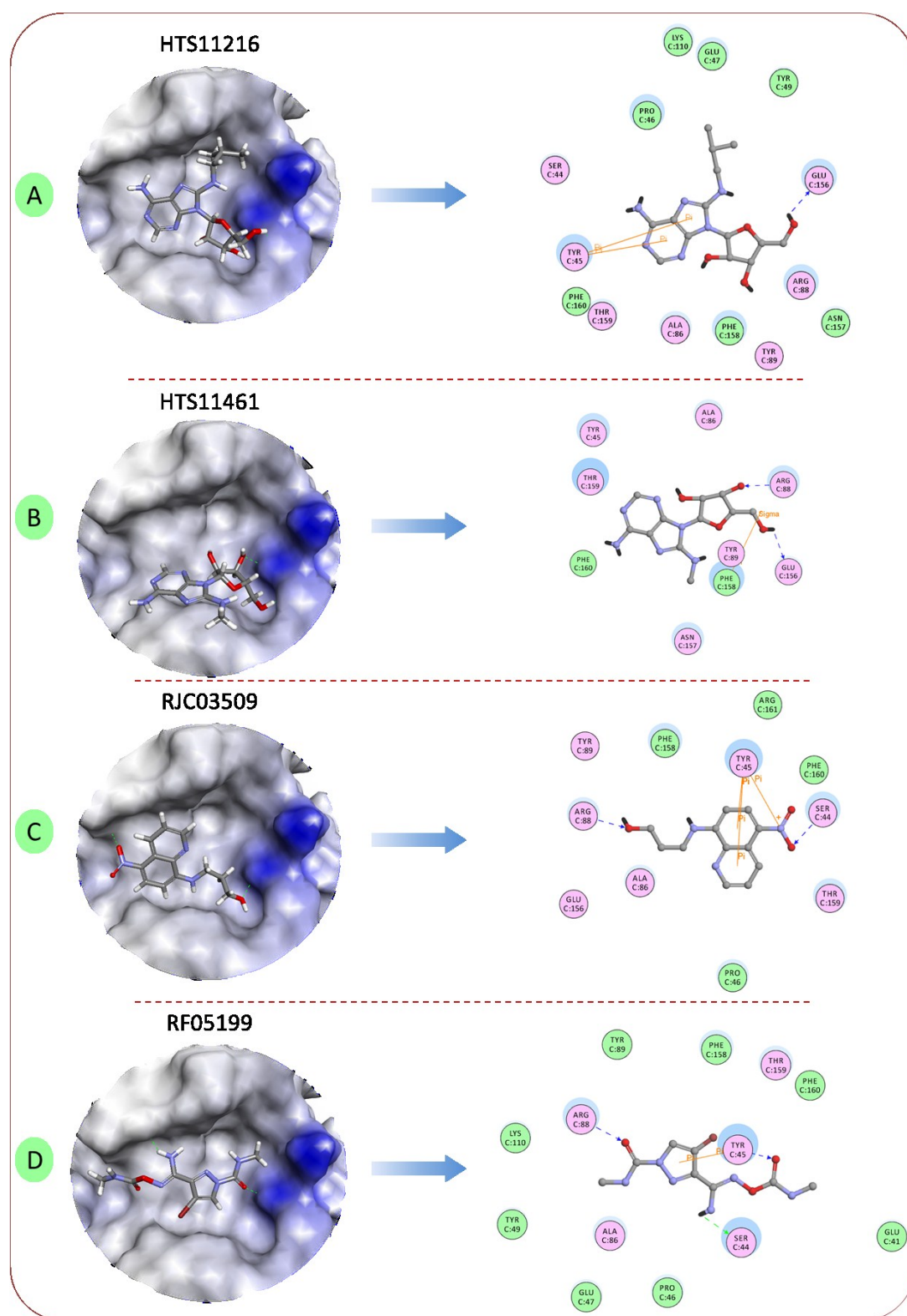


Figure 4.42: Analysis of the binding modes of the four (A-D) differential scanning fluorimetry (DSF) hits in the putative binding site. The left panel shows a docked pose for each of the compounds; (A) corresponds to compound HTS11216 docked

pose number 3; (B) corresponds to compound HTS11461 docked pose number 5; (C) corresponds to compound RJC03509 docked pose number 2; and (D) corresponds to compound RF05199 docked pose number 1. The enzyme is represented by a solid van der Waals surface, and the ligands are depicted in sticks where their atoms are coloured by element (C is in grey, N is in blue, O is in red, Br is in reddish brown and H is in white). Hydrogen bonds are shown as dashed green lines. The right panel shows the corresponding 2D interaction diagram between the four compounds and the enzyme. The ligands are represented by balls and sticks and their atoms are coloured by element in the same way as in the left panel apart from hydrogens which are coloured black. The residues are depicted as coloured discs including the name and number of the residue. The meaning of residue colouring is as follows: magenta coloured discs are residues involved in hydrogen-bond or charge or polar interactions; green discs are residues involved in van der Waals interactions; the solvent accessible surface of an interacting residue is represented by a blue halo around the residue and its diameter is proportional to the solvent accessible surface. The protein-ligand interactions are depicted as follows: hydrogen-bond interactions with amino acid main chains are represented by a green dashed line with an arrow head directed towards the electron donor; hydrogen-bond interactions with amino acid side-chains are represented by a blue dashed line with an arrow head directed towards the electron donor; pi interactions are represented by an orange line with symbols indicating the type of interaction (Pi-Pi or sigma-Pi).

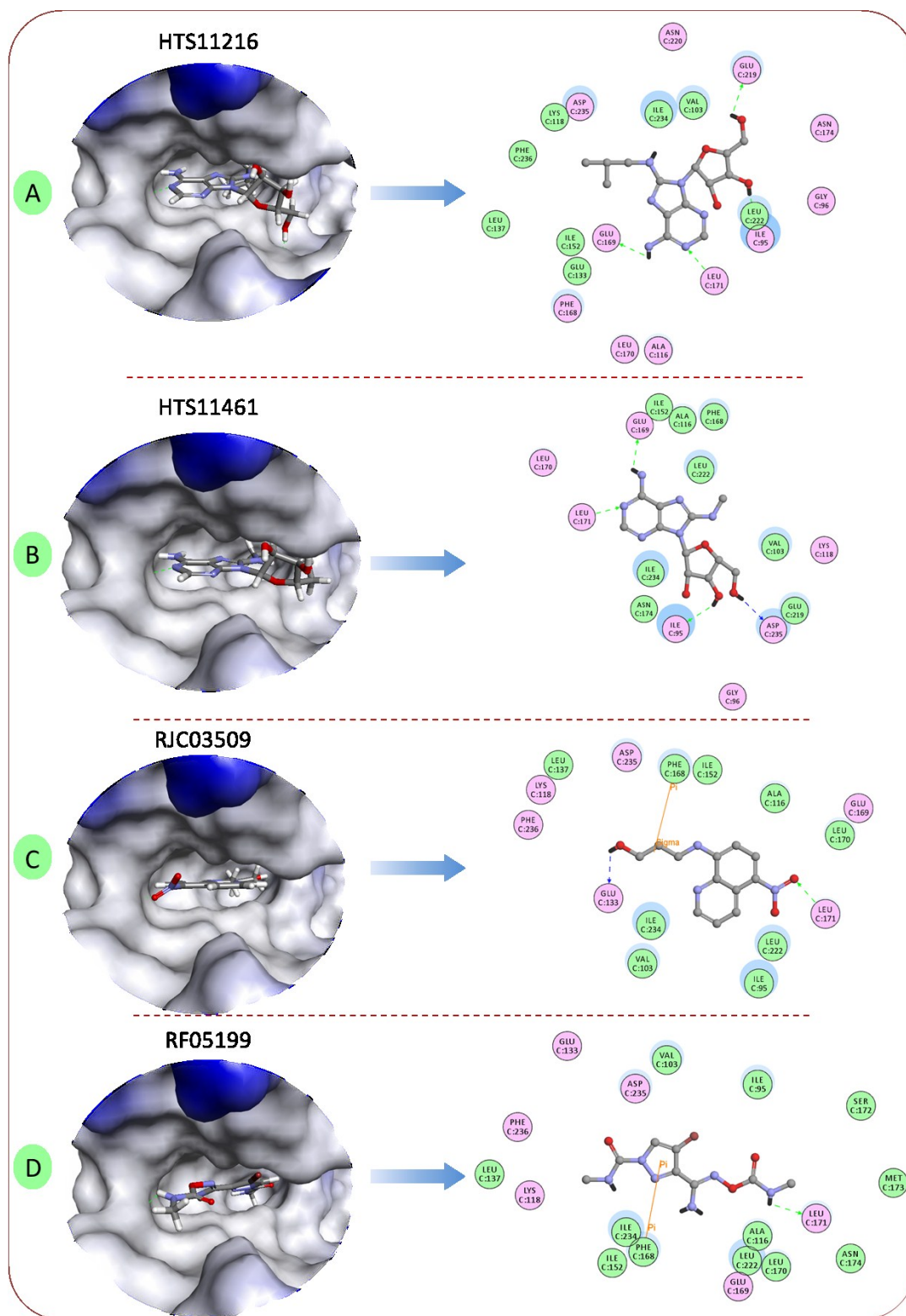


Figure 4.43: Analysis of the binding modes of the four (A-D) differential scanning fluorimetry (DSF) hits in the ATP binding site. The left panel shows a docked pose for each of the compounds; (A) corresponds to compound HTS11216 docked pose

number 5; (B) corresponds to compound HTS11461 docked pose number 9; (C) corresponds to compound RJC03509 docked pose number 3; and (D) corresponds to compound RF05199 docked pose number 2. The right panel shows the corresponding 2D interaction diagram between the compound and the enzyme. Representation of the enzyme, residues, ligands, and interactions are the same as in figure 4.42.

Careful analysis of the docked poses and more importantly the number and types of interactions between the compounds and the enzyme that are shown in figures 4.45 and 4.46 revealed that compounds HTS11216 and HTS11461 have better binding modes and interaction patterns in the ATP binding site, while compounds RJC03509 and RF05199 showed the opposite. Table 4.3 summarises the number and types of interactions for each compound in each of the two binding sites, and it shows that compounds RJC03509 and RF05199 are establishing better interactions with the putative binding site with a total number of interactions of 5 and 4 respectively, compared to that in the ATP binding site of 3 and 2 total interactions. Based on these results, compounds RJC03509 and RF05199 that are seemingly favouring the putative allosteric binding site were selected for conducting more informative experimental evaluation of their binding affinities and mode of inhibition to DYRK2 by studying their kinetics.

Table 4.3: Summary of the number and types of interactions between the four hits and each of the putative and ATP binding sites.

Compound name	Number of H-bonds		Number of Pi interactions		Total number of interactions	
	Putative site	ATP site	Putative site	ATP site	Putative site	ATP site
A. HTS11216	1	3	2	0	3	3
B. HTS11461	2	4	1	0	3	4
C. RJC03509	2	2	3	1	5	3
D. RF05199	3	1	1	1	4	2

4.9.2 DYRK2 assays

4.9.2.1 Determination of the IC_{50} values of the selected hits

Having identified two possible DYRK2 inhibitors (RJC03509 and RF05199) via DSF screening and detailed analysis of their binding modes from the docked poses, an experimental evaluation of their inhibitory concentration was conducted by performing an enzyme inhibition assay.

Concentration-response plots are commonly used to determine the effects of an inhibitor on a target enzyme. Experiments that are conducted to quantify the inhibitor concentration-response are performed at varying inhibitor concentrations while fixing the concentrations of the enzyme and substrate. To generate the plot, fractional enzyme activity (Y axis) is plotted as a function of inhibitor concentration (X axis), and the inhibitor concentration that causes 50% inhibition of maximal enzymatic activity is termed the IC_{50} . The range of inhibitor concentrations should be wide enough to provide well-defined top and bottom plateau values.

The DYRK2 screening kit documentation cited a K_m (Mechalis-Menten constant – see Section 4.9.2.2) of $4.6\mu\text{M}$ for ATP; therefore before screening the two compounds a K_m was calculated for ATP using the kit in order to validate its accuracy and reproducibility. A K_m of $3.8\mu\text{M}$ was observed which compares favourably with the $4.6\mu\text{M}$ cited (figure 4.44).

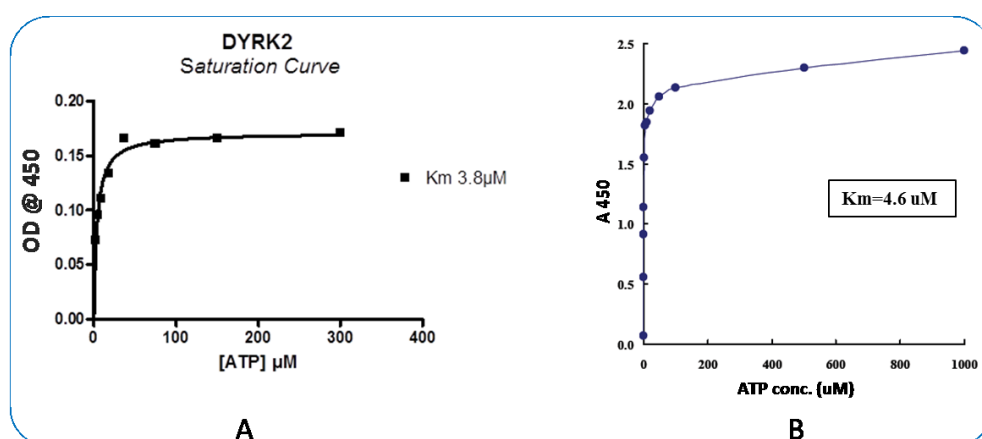


Figure 4.44: Comparison of the K_m values for ATP (recombinant DYRK2) obtained by using the purchased kit (A) with that reported in the user manual of the manufacturer (B).

As a control, the universal kinase inhibitor staurosporine was tested to determine inhibitory effects against DYRK2 in the concentration range of 10nM-10 μ M (figure 4.45).

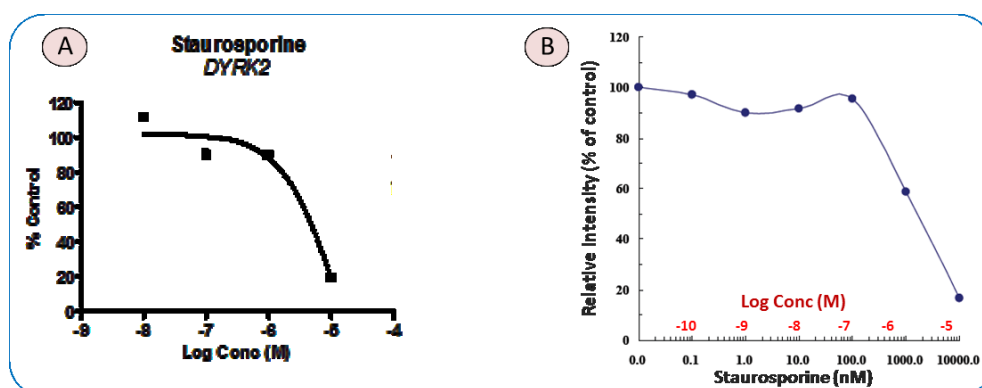


Figure 4.45: Comparison of the inhibitory effect of staurosporine on the activity of recombinant DYRK2 obtained when conducting the inhibition assay (A) with the results reported in the kit's user manual (B).

In order to calculate the IC₅₀'s of the two selected hits, their inhibitory effects were quantified for a concentration range of 10nM to 30 μ M against DYRK2 at 125 μ M ATP. Their concentration-response plots are presented in figure 4.46.

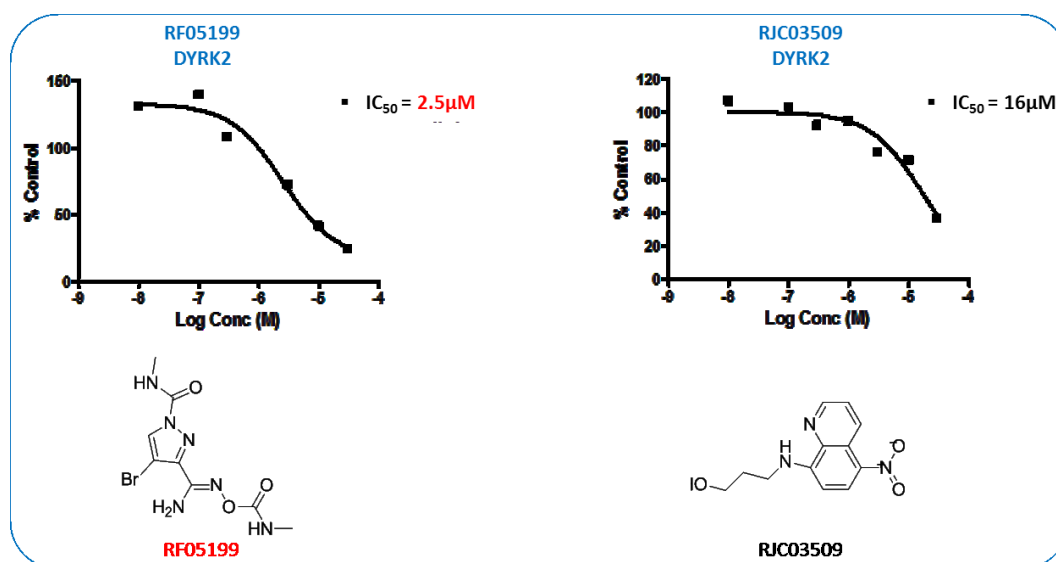


Figure 4.46: The concentration-response curves of RF05199 (left) and RJC03509 (right) at 125 μ M ATP.

As can be seen in figure 4.46, an IC_{50} value of $2.5\mu\text{M}$ and $16\mu\text{M}$ were determined for RF05199 and RJC03509 respectively; which are very promising values for hits obtained by virtual screening, especially for RF05199. In order to determine the mode of inhibition of RF05199 (whether it is a competitive or a non-competitive inhibitor), an investigation of its binding kinetics was conducted.

4.9.2.2 Determination of the mode of inhibition of the active hit

Enzyme kinetics refers to the study of enzyme-catalysed chemical reactions: measurement of the reaction rate and investigating the effects of varying reaction conditions can establish the catalytic mechanism, how activity is controlled and how drugs or small molecules might inhibit the enzyme.

Usually, the initial rate (v_0) of the enzyme-catalysed reaction is measured, which corresponds to a known substrate concentration $[S]$, with $[S]$ decreasing with time; figure 4.47 shows the variation of the initial reaction rate with $[S]$. The figure shows that the enzyme-catalysed reaction shows saturation kinetics, meaning that at low substrate concentrations the rate increases linearly, but as the concentration of the substrate increases the rate gradually levels off toward its maximum (V_{max}), when all enzyme active sites become saturated [190].

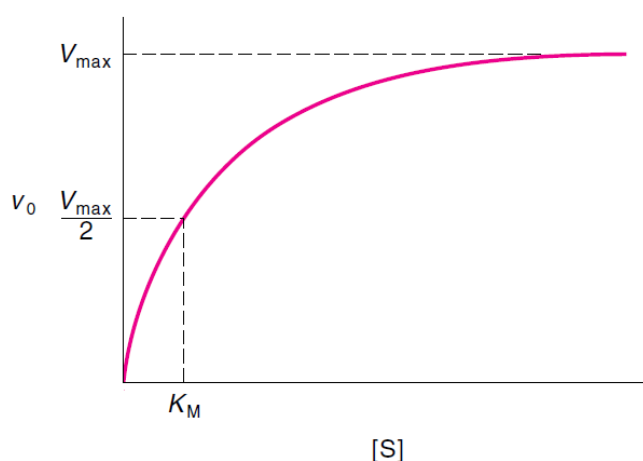
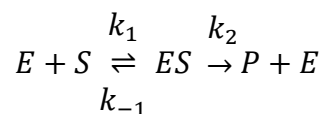


Figure 4.47: Saturation curve for an enzyme showing the relation between the initial reaction rate (V_0) and substrate concentration $[S]$, from which V_{max} and K_M can be determined [191].

One of the most important and fundamental equations that relate the initial rate of enzyme-catalysed reaction with substrate concentration is the Michaelis-Menten equation, which is derived by considering the following scheme:



The scheme above describes a single-substrate mechanism for an enzyme-catalysed reaction, where k_1 , k_{-1} and k_2 are the rate constants for the individual steps. The mathematical equation they derived that describes the dependence of the initial rate on substrate concentration is:

$$v_0 = \frac{V_{max} [S]}{K_M + [S]} \quad (4.1)$$

with

$$V_{max} = k_2[E]_{total} \quad (4.2)$$

and

$$K_M = \frac{k_{-1} + k_2}{k_1} \quad (4.3)$$

where v_0 is the initial reaction rate, V_{max} is the maximum rate, $[S]$ is the substrate concentration, $[E]_{total}$ is the total concentration of the enzyme, and K_M is the Michaelis-Menten constant, which is the substrate concentration midway between the initial rate and the maximum rate.

Therefore, measuring and plotting the initial reaction rate at increasing substrate concentrations will generate the saturation curve of the enzyme of interest which shows the relation between the reaction rate and substrate concentration (figure 4.47). In principle, both V_{max} and K_M can be determined from such a plot. However, this is not practical owing to the difficulty in locating the asymptotic value of V_{max} at very high substrate concentration [190]. Nevertheless, computer software that use nonlinear regression methods make it feasible.

The mode of inhibition of an enzyme by an inhibitor can be classified into reversible and irreversible inhibition. In the former case, there will be an equilibrium between the inhibitor and the enzyme, while in the latter, there will be a progressive inhibition with time until complete inhibition is reached when the concentration of the inhibitor exceeds that of the enzyme. Furthermore, reversible inhibitors are further sub-categorized into three mechanisms: competitive inhibitors, which compete with the substrate for the enzyme active site; non-competitive inhibitors, which bind to a distinct site from that of the substrate; and the uncompetitive inhibitors, which bind to the enzyme-substrate complex but not to the free enzyme (figure 4.48). In the case of competitive inhibition, V_{max} remains constant while K_M increases. In the case of non-competitive inhibition, the V_{max} decreases while K_M is constant. Finally, in the case of uncompetitive inhibition, both V_{max} and K_M decreases at equivalent ratios. In these three mechanistic groups, whenever the inhibitor binds the free enzyme or the enzyme-substrate complex, no products will be formed [190].

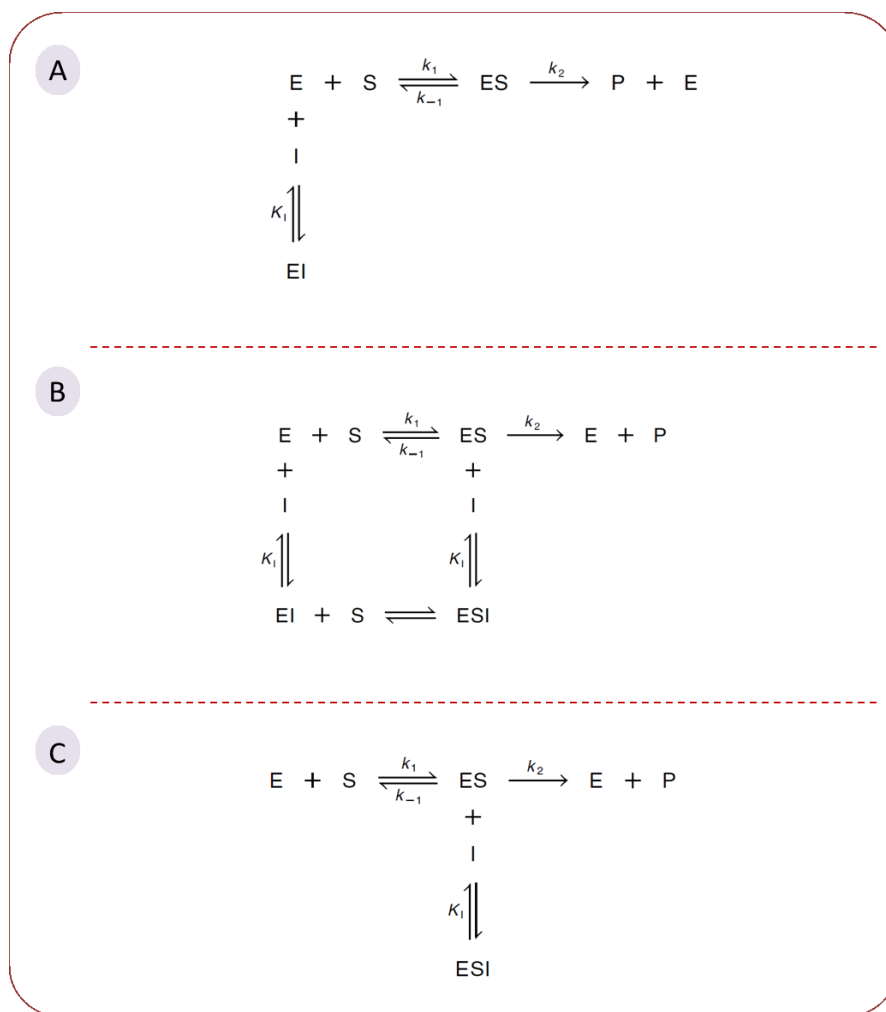


Figure 4.48: The reaction schemes of the three classes of reversible enzyme inhibition. (A) Competitive inhibition. (B) Non-competitive inhibition. (C) Uncompetitive inhibition [191].

The kinetics of DYRK2 inhibition by RF05199 were analysed by determining the initial rates at five concentrations of the inhibitor for eight different ATP concentrations (table 4.4). The initial rate values at zero inhibitor concentration along with the eight different ATP concentrations used to calculate it are in red. The kinetic parameters of DYRK2 in this study were determined by using the saturation curve plots of v_0 versus $[ATP]$ (figure 4.49), and the obtained values of V_{max} and K_M are summarised in table 4.5.

Table 4.4: The initial rates of RF05199SC at four different concentrations for eight different concentrations of ATP relative to the reference of no inhibitor.

DYRK2 initial rates ($\mu M \cdot s^{-1}$) at different [S] and [I] concentrations							
ATP conc (μM)	10 μM	5 μM	2.5 μM	1 μM	0.1 μM	ATP conc (μM)	0 μM
125	0.0393	0.0753	0.0849	0.1094	0.151	300.000	0.1704
75	0.0407	0.0365	0.0868	0.1352	0.1349	150.000	0.1657
25	0.0216	0.0256	0.0308	0.0683	0.0892	75.000	0.1610
12.5	0.0094	0.0176	0.0297	0.0457	0.1299	37.500	0.1659
6.25	0.0033	0.0099	0.0105	0.0396	0.055	18.750	0.1338
3.125	0.0033	0.0071	0.0067	0.0146	0.0392	9.400	0.1110
1	0.0036	0.0016	0.0054	0.0074	0.0177	4.700	0.0954
0.5	0.0005	0.0013	0.0024	0.0059	0.0105	2.350	0.0730

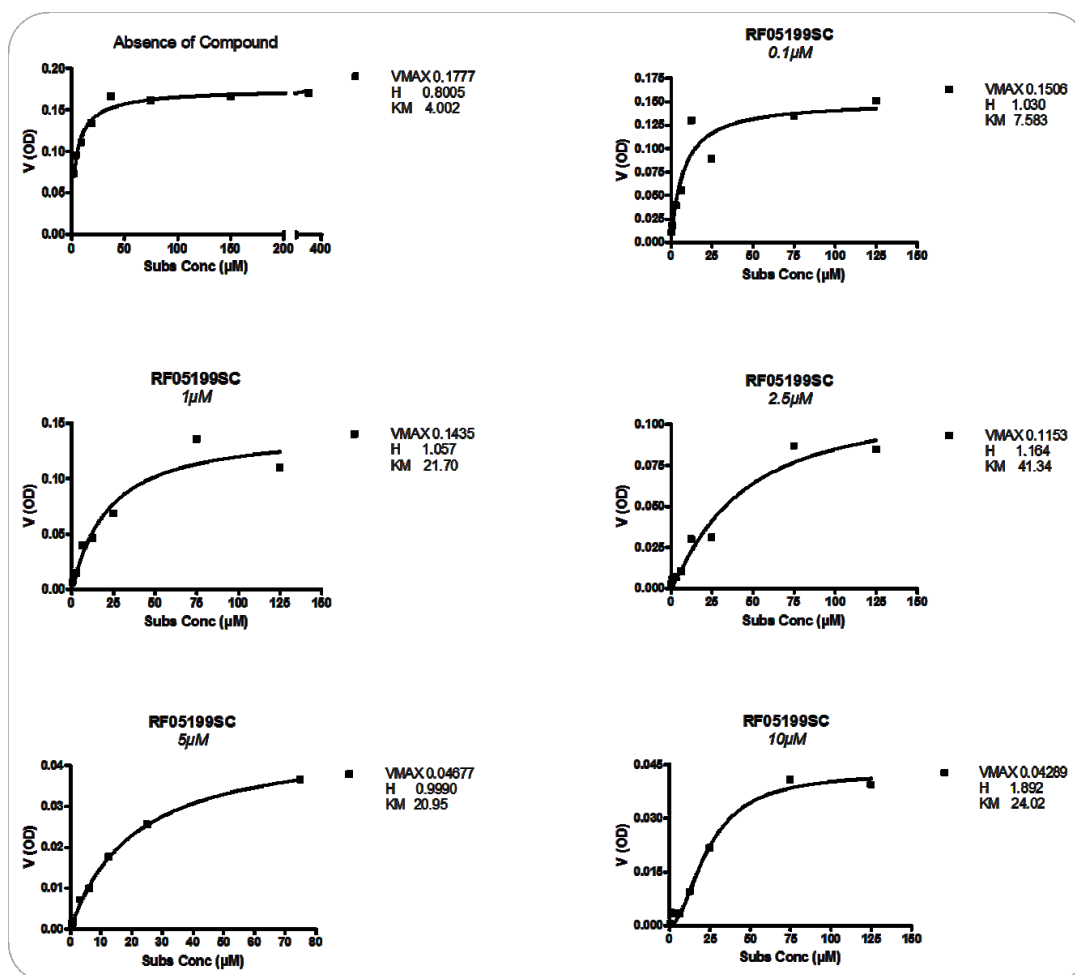


Figure 4.49: Saturation curve plots for the six concentrations of RF05199 relative to the reference to determine the kinetic parameters and the type of inhibition along with Hill's coefficient values.

Table 4.5: The kinetic parameters for the DYRK2 kinase that were obtained from the saturation curve plots.

RF05199 Conc (μM)	$V_{max}(\mu\text{M}\cdot\text{s}^{-1})$	$K_M (\mu\text{M})$
0	0.1778	4.002
0.1	0.1506	7.583
1	0.1435	21.70
2.5	0.1153	41.34
5	0.0468	20.95
10	0.0429	24.02

The initial rate data obtained from the saturation curve plots (table 4.5) show that the K_M increases from $4.002\mu\text{M}$ to $41.34\mu\text{M}$ with increasing concentration of the inhibitor from $0\mu\text{M}$ to $2.5\mu\text{M}$, while the V_{max} slightly decreases from $0.178\mu\text{M}\cdot\text{s}^{-1}$ to $0.115\mu\text{M}\cdot\text{s}^{-1}$. Afterwards K_M started to decrease at higher concentrations of the inhibitor while V_{max} remained the same. These findings show that the mode of RF05199 inhibition of DYRK2 appears to be mixed and could potentially involve an allosteric mechanism.

5 CONCLUSION AND OUTLOOK

5.1 Concluding remarks

The discovery of new allosteric sites in biomolecules in general and in proteins in particular can provide new avenues for the identification of new drugs for a wide range of diseases. Furthermore, it could offer innovative opportunities for better and deeper understanding of fundamental cellular processes. Ultimately, this could expand the chemical space of potential leads and aid the development of new drug chemotypes. Since protein-protein interactions in most cases are of an allosteric nature, the identification of allosteric sites that can interfere with the productive coupling of such co-dependent protein systems (as in the case of CDK2) may also lead to the development of inhibitors capable of disrupting such associations.

The results that have been presented in this study have clearly shown that protein dynamics and conformational flexibility are the major driving force in allosteric coupling events. These observations support the current view of allostery in which a protein is perceived as populating different dynamic conformational ensembles at equilibrium and perturbation of this equilibrium, for example through ligand binding, would induce a shift in the pre-existing populations (population shift) towards a different ensemble of conformations which trap the protein in a certain state.

In this study, our aim was to develop a new computational method that can be used to identify putative allosteric binding sites that can be utilised in drug discovery. Full characterisation of the residual dynamic cross-correlations, complete mapping of the complex multi-way interfaces and identification of long-range energetic interaction pathways in all of the examined proteins were performed to achieve this goal. Three protein kinases were studied; two for proof of concept purposes, and real study case to identify a new allosteric site and a ligand to bind.

In studying DYRK2 attention was focused on the apo form of the enzyme which represents its native state. Analysis of the correlated residues that could act as signal mediators, identification of the energetic interaction pathways and energy gates, and analysis of the major multi-way interfaces in DYRK2 enabled us to establish its allosteric profile and to identify a putative allosteric binding site in the N-domain. Protein 3D architecture appeared to be essential for effective propagation and transmission of signals between distant binding sites within the protein via pre-

existing interaction pathways that are linked to the dynamic motion of the protein (as inferred from SID and energy correlation analyses).

Once the putative allosteric site was identified, it allowed the construction of structure-based pharmacophore models based on the putative allosteric site using the minimised average structure of the *DYRK2-com-6* model. This was used to search the Maybridge small molecule database for ligands capable of interfering with the function of the protein by disrupting the network of interaction pathways related to the enzyme functional activity. The retrieved hits were then filtered and docked into the allosteric site. Experimental screening of the selected hits using DSF showed that four of the selected molecules bound to the enzyme. When evaluated in a quantitative way by examining binding kinetics, RF05199 had an IC_{50} of $2.5\mu\text{M}$ in a non-competitive nature. Figure 5.1 summarises the entire process of applying our approach in studying DYRK2, from the identification of the putative allosteric site through to the assessment of the mode of inhibition of the most active hit.

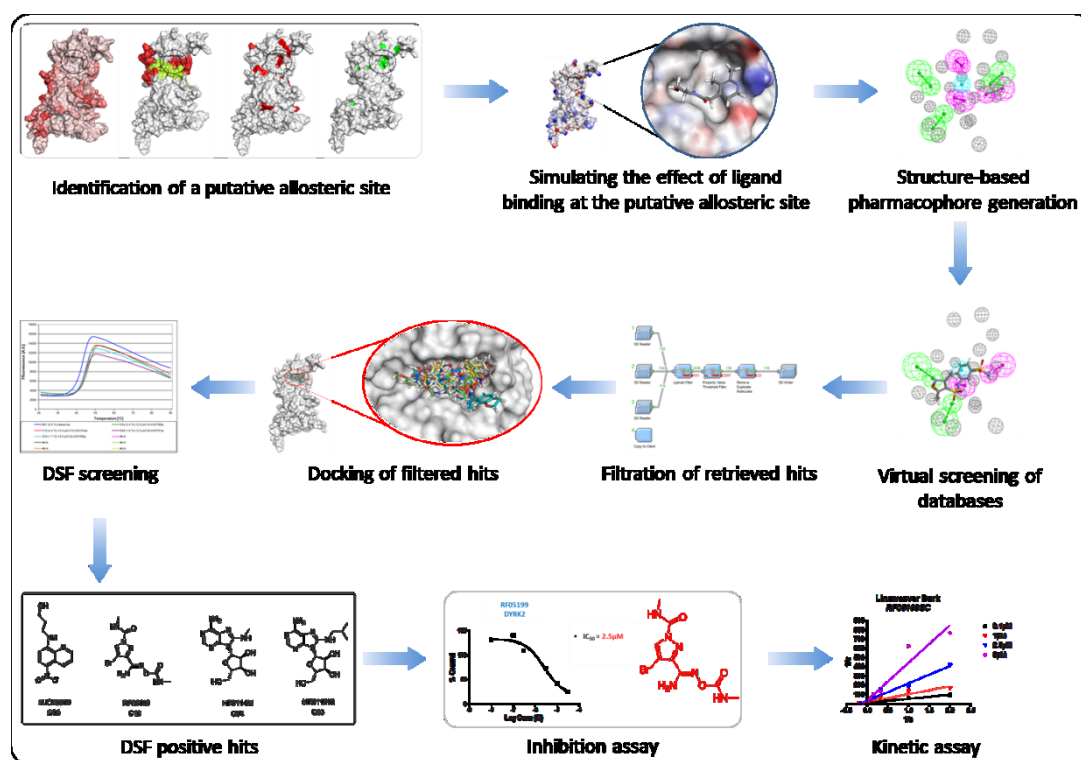


Figure 5.1: A scheme summarising the different steps that led to the identification of a non-competitive DYRK2 inhibitor. Application of our computational approach in

studying the apo form of DYRK2 resulted in the identification of a putative allosteric binding site in the N-domain of the enzyme, followed by the generation of a pharmacophore to screen commercial databases. Experimental evaluation of the binding affinities of the selected hits using DSF identified four positive hits which were further evaluated by enzyme kinetic analysis.

It is evident from this study that the explicit consideration of protein dynamics, complex interfaces and energetically coupled residues could be exploited to identify hot spots in a protein where small molecules can bind and cause perturbation in protein stability and its dynamics that may modulate its function. Analysis of protein flexibility and functional dynamics in association with topology and 3D architecture can reveal other states of the protein with possible binding sites that may not be available in a static representation of protein structure.

In summary, the combination of biophysical information and computational models with pharmacophore generation and screening represents a useful tool that can be utilised in identifying new binding sites to find hit compounds in drug discovery and chemical biology.

MD simulations

MD simulations has broadened and deepened our understanding of how biomolecules behave. They have been applied in many drug discovery projects to provide information complementary to that of experimental methods, especially in generating conformational ensembles that may not be accessible experimentally.

In this study, MD simulations were utilised to investigate the dynamic motion of kinases pertaining to their allosteric regulation; in particular via residual fluctuations and cross-correlation motions. The former was valuable in identifying regions of the proteins that are endowed with considerable flexibility thereby aiding the profiling of the allosteric map of that protein. The latter identified the coordinated and collective motion of the examined proteins. Those two dynamic properties when used to compare the different states of the same protein (free and complexed states) provided

valuable information about the possible mechanism of ligand intervention (e.g. allosteric) with protein activity.

SID analysis

SID analysis proved to be successful in identifying complex interfaces in the examined proteins that coincide with experimentally identified allosteric sites. The results suggest that the binding sites are at the main interfaces between the domains and sub-domains that establish a connection within the 3D structure of the protein (figure 5.2) (same applies for JNK1 and CDK2).

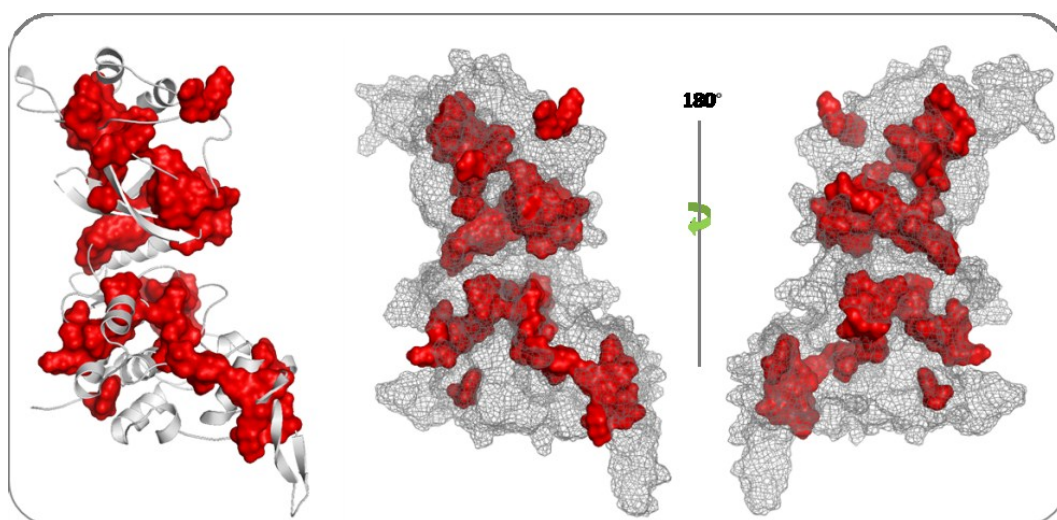


Figure 5.2: Interfacial connectivity in DYRK2. The major interfaces in DYRK2 are represented as solid surface around the backbone of the protein and within a mesh surface. It is clear that they are establishing a network of contacts within the 3D structure of the protein; hence disruption of the stability of any of these interconnected interfaces will propagate to distant loci in the protein.

This seems to provide residues remote from the functional site when perturbed to tweak the active site and subsequently the function of the protein. These perturbations through ligand binding could have caused adjustments of their underlying interfaces or in their vicinity which are consequently transferred to the active site via the interfacial connectivity. This can be seen in JNK1 upon binding of

the allosteric inhibitor and how it changed the interfacial picture at the ATP and JIP1 binding sites along with modulation of the activity of the enzyme. Most of the perturbation effects have been received at the active site through slight positional changes among the residues that border it. This in turn alters the directionality of certain components in the active site, thereby modulating protein function.

5.2 Outlook

Co-crystallisation of RF05199 with DYRK2 to provide decisive evidence of whether this compound is acting through binding to the identified putative allosteric binding site is underway. If the x-ray crystallography results confirm that this hit is binding to the identified putative allosteric site, then this will be the first allosteric site in proteins to be completely identified and characterised via computational approaches.

6 APPENDICES

6.1 Appendix I: Relationship between force, potential energy, and positional changes as a function of time

As pointed earlier, MD simulation is based on Newton's second law of motion. So, if we knew the force acting on each atom of the system, then it is possible to determine its acceleration. Newton's equation of motion states that:

$$F_i = m_i a_i \quad (6.1)$$

where F_i is the force exerted on particle i , m_i is the mass of particle i and a_i is the acceleration of particle i . Recall that the change in a particle's position with respect to time is its velocity:

$$v_i(t) = \frac{dr_i(t)}{dt} \quad (6.2)$$

where v_i is the velocity of particle i (atom i), r_i is the position of particle i . Also the change in a particle's velocity with respect to time is its acceleration:

$$a_i(t) = \frac{dv_i(t)}{dt} \quad (6.3)$$

where a_i is the acceleration of particle i , and v_i is the velocity of particle i . The force can also be expressed as the gradient of the potential energy [70]:

$$F_i = -\nabla V_i \quad (6.4)$$

where V is the potential energy of the system. Combining equations (6.1) and (6.4) with rearrangement yields:

$$-\frac{dV_i}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \quad (6.5)$$

Newton's equation of motion can then relate the derivative of the potential energy to the changes in position as a function of time.

$$F_i = m_i a_i = m_i \frac{dv_i(t)}{dt} = m_i \frac{d^2 r_i}{dt^2} \quad (6.6)$$

After integration of equations (6.2) and (6.3) we obtain expressions for the position and the velocity respectively as follows:

$$x_i = v_i t + x_{i0} \quad (6.7)$$

$$v_i = a_i t + v_{i0} \quad (6.8)$$

Substituting for the velocity in equation (6.7) with that in equation (6.8) gives the relation which gives the value of x at time t as a function of the acceleration, a , the initial position, x_0 , and the initial velocity, v_0 .

$$x_i = a_i t_i^2 + v_{i0} t_i + x_{i0} \quad (6.9)$$

From equation (6.4) the acceleration is given as the derivative of the potential energy with respect to the position, r :

$$a_i = -\frac{1}{m_i} \frac{dV_i}{dr_i} \quad (6.10)$$

Therefore, to calculate a trajectory, we only need the initial positions of the atoms which is obtained from x-ray crystal structure or NMR structures; an initial distribution of velocities obtained from Maxwell-Boltzmann distribution; and the acceleration, which is determined by the gradient of the potential energy function [70, 141].

6.2 Appendix II: minimisation, equilibration, and MD production input files

6.2.1 Minimisation input files

Minimisation of water and hydrogen

```
minimisation CUT15A restraint on heavy atoms except water
&cntrl
  imin      = 1,
  maxcyc    = 50000,  ncytc = 250,
  ntb       = 1,
  drms      = 0.1,
  ntpc      = 500,
  ntr       = 1,
  restraintmask='!(@H=) & !(:WAT | @Na+=)',
  restraint_wt=100.0,
  igb       = 0,
  cut       = 15,
&end
END
END
#restraintmask="!(@H=) & !(:WAT)",
```

The meaning of each of the terms is as follows:

- IMIN = 1: Minimisation is turned on (no MD).
- MAXCYC = 50000: Conduct a total of 50,000 steps of minimisation.
- NCYC = 250: Initially do 250 steps of steepest descent minimisation followed by (MAXCYC - NCYC) steps of conjugate gradient minimisation.
- NTB = 1: Use constant volume periodic boundaries (PME is always "on" when NTB>0).
- DRMS =0.1: The convergence criterion for the energy gradient. minimisation will stop when the root-mean-square of the Cartesian elements of the gradient is less than DRMS.
- NTPC= 500: Every 500 steps energy information will be printed in human-readable form to files "mdout" and "mdinfo".
- NTR = 1: We want to restrain specified atoms in Cartesian space using a harmonic potential. Restrained atoms are specified in the `restraintmask`. The force constant is given by `restraint_wt`.
- IGB=0: No generalized Born term is used.
- CUT = 15.0: Use a cutoff of 15 angstroms.

Minimisation of side chains and water

```
minimisation CUT15A restraint on backbone
&cntrl
  imin    = 1,
  maxcyc  = 50000,  ncyc    = 250,
  ntb     = 1,
  ntp     = 500,
  drms    = 0.1,
  ntr     = 1,
  restraintmask='@CA,C,O,N,OXT & !(:WAT | @Na+=) ',
  restraint_wt=100.0,
  igb     = 0,
  cut     = 15,
&end
END
END
```

Minimisation of the entire system, the entire system is allowed to move

```
minimisation CUT15A no restraint
&cntrl
  imin    = 1,
  maxcyc  = 50000,  ncyc    = 250,
  ntb     = 1,
  ntp     = 500,
  drms    = 0.1,
  ntr     = 0,
  igb     = 0,
  cut     = 15,
&end
END
END
```

Note how the restrainmask is changing from one minimisation input file into the other.

6.2.2 Equilibration input files

When using Langevin or Andersen dynamics

Stage one: equilibration 100 (heating the system to 100K)

```
MD with weak position restraint on solute
&cntrl
  imin=0,
  irect=0, ntx=1,
  ntb=1,
  cut=15,
  ntr=1,
  ntc=2, ntf=2,
  tempi=0.0, temp0=100,
  ntt=3,
  gamma_ln=1.0,
  ig=-1,
  iwrap=1,
  nstlim=10000, dt=0.002,
  ntp=100, ntwx=100,ntwr=1000
/
Keep Solute fixed with weak restraints
10.0
RES 1 n (n is the number of residues in the simulated system)
END
END
```

Stage two: equilibration 200 (heating the system to 200K)

```
MD with weak position restraint on solute
&cntrl
  imin=0,
  irect=1, ntx=7,
  ntb=1,
  cut=15,
  ntr=1,
  ntc=2, ntf=2,
  tempi=100.0, temp0=200,
  ntt =3,
  gamma_ln=1.0,
  ig=-1,
  iwrap=1,
  nstlim=20000, dt=0.002,
  ntp=100, ntwx=100,ntwr=1000
/
Keep Solute fixed with weak restraints
10.0
RES 1 n
END
END
```

Stage three: equilibration 300V1 (heating the system to 300K)

```
MD with weak position restraint on solute
&cntrl
imin=0,
irest=1, ntx=7,
ntb=1,
cut=15,
ntr=1,
ntc=2, ntf=2,
tempi=300.0, temp0=300,
ntt=3,
gamma_ln=1.0,
ig=-1,
iwrap=1,
nstlim=30000, dt=0.002,
ntpr=100, ntwx=100,ntwr=1000
/
Keep Solute fixed with weak restraints
10.0
RES 1 n
END
END
```

Stage four: equilibration 300P2

```
MD with weak position restraint on solute
&cntrl
imin=0,
irest=1, ntx=7,
ntb=2,
cut=15,
ntr=1,
ntc=2, ntf=2,
ntp=1,
tempi=300.0, temp0=300,
ntt =3,
gamma_ln=1.0,
ig=-1,
iwrap=1,
nstlim=30000, dt=0.002,
ntpr=100, ntwx=100,ntwr=1000
/
Keep Solute fixed with weak restraints
10.0
RES 1 n
END
END
```

6.2.3 MD production input file

Stage one: Langevin dynamics

```
MD
&cntrl
imin=0,
irest=1, ntx=7,
ntb=2,
pres0=1.0,
ntp=1,
taup=2.0,
cut=15,
ntr=0,
ntc=2, ntf=2,
tempi =300.0, temp0=300.0,
ntt=3,gamma_ln=1.0,
ig=-1,
iwrap=1,
nstlim=500000,dt=0.002,ntpr=500,ntwx=500,ntwr=1000
/
```

The meaning of each of the terms is as follows:

- IMIN = 0: Minimisation is turned off (run molecular dynamics)
- IREST = 1, NTX = 7: We want to restart our MD simulation where we left off. IREST tells sander that we want to restart a simulation, so the time is not reset to zero but will start from the last run. NTX = 7 means that we want to continue from where we finished so the coordinates, velocities and box information will be read from a formatted (ASCII) restrt file.
- NTB = 2, PRES0 = 1.0, NTP = 1, TAUP = 2.0: Use constant pressure periodic boundary with an average pressure of 1 atm (PRES0). Isotropic position scaling should be used to maintain the pressure (NTP=1) and a relaxation time of 2ps should be used (TAUP=2.0).
- CUT = 15.0: Use a cut off of 15 angstroms.
- NTR = 0: We are not using positional restraints.
- NTC = 2, NTF = 2: SHAKE should be turned on and used to constrain bonds involving hydrogen.
- TEMPI = 300.0, TEMP0 = 300.0: Our system was already heated to 300 K so here it will start at 300 K and should be maintained at 300 K.

- `NTT = 3, GAMMA_LN = 1.0`: The Langevin dynamics should be used to control the temperature using a collision frequency of 1.0 ps^{-1} .
- `NSTLIM = 500000, DT = 0.002`: We are going to run a total of 500,000 molecular dynamics steps with a time step of 2 fs per step.
- `Iwrap 1`: the coordinates written to the restart and trajectory files will be "wrapped" into a primary box. For better visual look.
- `NTPR = 500, NTWX = 500, NTWR = 1000`: Write to the output file (`NTPR`) every 500 steps (1 ps), to the trajectory file (`NTWX`) every 500 steps and write a restart file (`NTWR`), in case our job crashes and we want to restart it, every 1,000 steps.

Stage two: Berendsen dynamics

```

MD using Berendsen dynamics (NTT=1)
&cntrl
  imin=0,  irest=1,  ntx=7,
  ntt=1,  tempi=300.0,  temp0=300.0,  tautp=1.0,
  ntb=2,  ntp=1,  pres0=1.0,  taup=2.0,
  ntc=2,  ntf=2,
  cut=15,
  ntr=0,
  nstlim=500000,  dt=0.002,
  iwrap=1,  ntp=500,  ntwx=500,  ntwr=1000
/

```

`NTT=1` : Means we are using Berendsen thermostat.

`TAUTP=1` : Time constant, in ps, for heat bath coupling for the system

6.3 Appendix III: Input files used to analyse the generated trajectories

Input file to calculate the root mean square deviation (RMSD) from a reference

```
trajin name_md(i).mdcrd.gz
.
.
trajin name_md(n).mdcrd.gz
strip :WAT,Na+ (or Cl-)
center origin :1-n (n is the number of residues in the system)
image origin center
reference (name of reference file.pdb)
rms reference mass out name_of_output_file.rms @C,CA,N time
0.1
```

Input file to calculate the average structure of a trajectory

```
trajin name_md(i).mdcrd.gz 1 50 1
.
.
trajin name_md(n).mdcrd.gz 1 50 1
strip :WAT,Na+
center origin : 1-n
image origin center
rms first mass
average name_of_the_output_file.pdb pdb nobox
```

Input file to calculate the atomic positional fluctuations (apf) or the (RMSF)

```
trajin name_md(i).mdcrd.gz
.
.
trajin name_md(n).mdcrd.gz
strip :WAT,Na+
center origin :1-n
image origin center
reference (name of the reference file.pdb)
rms reference mass
atomicfluct out name_of_the_output_file.apf :1-n byres
```

Input file to calculate the correlation matrix

```
trajin name_md(i).mdcrd.gz
.
.
.
trajin name_md(n).mdcrd.gz
strip :WAT,Na+
center origin :1-n
image origin center
reference (name of the reference file.pdb)
rms first mass
matrix correl @C,CA,N out name_of_the_output_file.dat byres
mass
```

Input file to strip water from a solvated structure

```
trajin name_of_the_file_of_interest.pdb
strip :WAT,Na+ (or Cl-)
trajout name_of_the_output_file.pdb pdb nobox
```

Input file to extract all frames from the generated trajectories

```
trajin name_md(i).mdcrd.gz
.
.
.
trajin name_md(n).mdcrd.gz
strip :WAT,Na+
trajout name_of_the_output_file.pdb pdb nobox
```

A perl script to rename the pdb file extracted from generated trajectories

(Written by Dr. Nahoum Anthony at SIPBS-University of Strathclyde).

```
#This perl script should rename in batch files contained in a
directory

use Cwd;
my ($dir)=cwd();
print "\nwhat's the root name of the pdb files you want
renamed (ex: test.pdb, knowing your files are named test.pdb.1,
test.pdb.2...) ?\n";
$pdbfile = <stdin>;
chomp($pdbfile);
```

```
opendir (PDBFILES, $dir) or die "can't open $dir: $!";

while ($file = readdir PDBFILES)
{
    if ($file =~ m/$pdbfile.[0-9]+/)
    {
#       open(PDB, "<", $file);
#       @name = split(/\./,$file);
#       for (my $i=0; $i<$#name+1; $i++)
#       {
#           print "\n$name[$i]\n $i";
#       }
        my($number) = sprintf("%08d",$name[2]);
        my($newname) = "$name[0]".$number.".pdb";
        rename ($file,$newname);
#       print "$newname";
#       print "\n$file\n";
    }
#   close (PDB);
}
```

6.4 Appendix IV: conversion of SID scores or RMSF values into colour codes

The SID scores and the RMSF values of all simulated systems were edited using MS excel and saved in a comma delimited (*.csv) format which was then converted to a colour code in Pymol using the following Python script. (Written by Dr. Nahoum Anthony at SIPBS-University of Strathclyde).

```
from pymol import cmd
import linecache

cmd.viewport( 1024, 768 )      #Set the viewport (image)
dimensions

util.performance(0) #Set the visual quality to its highest
value (0)

cmd.load( 'file name.pdb', 'jnk1')    # load a pdb file

cmd.hide("lines","jnk1")
cmd.show("cartoon","jnk1")
cmd.bg_color('white')

for i in range(n, number of residues in the pdb file): #
change the value in the range to the number of residues in
your pdb file

    j=i+1          #change this j value if your residue number
in the      pdb file starts above 1
                #ex: j=i+5 if your residue number starts at
4

    sid = linecache.getline('vlues_file_name.csv', i+1) #
gets a numerical value from the i-th line in a file called
file_name.csv in running directory
    r= 1          # setting a red value in RGB based on sid
score
    g= 1-float(sid)# setting a green value of 0 in RGB
    b= 1-float(sid)# setting a blue value in RGB based on
sid score
    cmd.set_color('mycolour%d' %i,[r,g,b]) # sets your rgb
colour
    cmd.color('mycolour%d' %i,'resi %d' %j)
```

6.5 Appendix V: Input files used to calculate the energy correlations

6.5.1 Perl scripts

A perl script to calculate the A and D matrices for heavy atoms

(Written by Dr. Nahoum Anthony at SIPBS-University of Strathclyde).

```
#this script writes the A and D matrices
(PloSComputationalBiology2007_1716) from a pdb file containing
heavy atoms only.
#The PDB file has to start with the first heavy atom on the
first line and the last heavy atom on the last line.
#The script outputs a file called: "dist_your_pdb_file.txt"
#
#The script assumes the first residue as number 1, otherwise
need to modify the script !!
use Cwd;
my ($dir)=cwd();
print "\nWhat's you pdb file name? \n";
$pdb_file = <stdin>;
chomp($pdb_file);

#print "\nHow many residues does your protein contain? \n";
#$res_num = <stdin>;
#chomp($res_num);

open(SITE, "< $pdb_file") or die"input file doesn't exist";

open(A_MATRIX, ">> A_$pdb_file.txt") or die"output file not
created";
open(D_MATRIX, ">> D_$pdb_file.txt") or die"output file not
created";

my (@mat);
my ($line)=0;
my (@site_num);

my (@X);
my (@Y);
my (@Z);
my (@resnum); #will store which residue number the atom
belongs ie atom 1, belongs to residue 1, atom 7 to residue
2...
#my (@CHECK_TER);

my (@resatomnum); # will store how many atoms the residue has
is res 1 has 6 atoms, residue 2 has 7...

my(@dist);
my(@A_MAT);
my(@D_MAT);
```

```

my ($atomcount)=1;
my ($rescheck)=1;

my(@CONTACT_MAT); #(atomxatom) size matrix which will hold a
1 for residues in contact (dist<=7 A) and a 0 otherwise
my(@NUMCONTACT); #(resxres) size matrix which will hold the
total number of atom-atom contacts

while(<SITE>)
{
    @split_line = split(/\s+/, $_);
    push(@X, $split_line[5]);
    push(@Y, $split_line[6]);
    push(@Z, $split_line[7]);
    push(@resnum, $split_line[4]);
    push(@CHECK_TER, $split_line[0]);
    $line++;
}

for (my($i)=1; $i<=$#X; $i++)
{
    for (my($j)=1; $j<=$#X; $j++)
    {
        $dist[$i][$j]=sqrt(($X[$i-1]-$X[$j-1])**2 + ($Y[$i-
1]-$Y[$j-1])**2 + ($Z[$i-1]-$Z[$j-1])**2);
#        print "$dist[$i][$j]\t";
        if (0 < $dist[$i][$j] && $dist[$i][$j] <= 7 )
#check if I shouldn't add a condition here for distance=0
        {
            $CONTACT_MAT[$i][$j]=1;
        }
        else
        {
            $CONTACT_MAT[$i][$j]=0;
        }
#        print "$CONTACT_MAT[$i][$j]\t";
    }
#    print "\n";
}

for (my($i)=1; $i<=$#resnum; $i++)
{
    if($resnum[$i-1] == $resnum[$i])
    {
        $atomcount++;
    }
    else
    {
        push(@resatomnum, $atomcount);
        $atomcount=1;
    }
}

```

```

#print "\n";

my ($startcol)=1;
my ($startlin)=1;
my ($endcol)=$resatomnum[0];
my ($endlin)=$resatomnum[0];
my ($numcontact)=0;
#print "\n";
for (my ($col)=1; $col<=$#resatomnum+1; $col++)
{
    for (my ($lin)=1; $lin<=$#resatomnum+1; $lin++)
    {
        #      print"\ncol $col lin $lin startcol $startcol endcol
        $endcol startlin $startlin endlin $endlin resatomcol
        $resatomnum[$col-1] resatomlin $resatomnum[$lin-1]";
        if ($lin==$col)      #diagonal terms of the contact
        matrix (intra residue contacts)
        {
            for (my ($i)=$startcol+1; $i<=$endcol; $i++)
            {
                for (my ($j)=$startcol; $j<=$i-1; $j++)
                #shouldn't it be startlin instead of startcol ?
                {

                    $numcontact=$numcontact+$CONTACT_MAT[$i][$j];
                #      print "diag i=$i j=$j
                numcontact=$numcontact
                contact_mat(i)(j)=$CONTACT_MAT[$i][$j]\n";
                }
            }
            $NUMCONTACT[$col][$lin]=0;#$numcontact;
            #+$resatomnum[$col-1]; digonal term removed, no "self-contact"
            counted after discussion with author of paper

            $A_MAT[$col][$lin]=($NUMCONTACT[$col][$lin])/(sqrt($resatomnum
            [$col-1]*$resatomnum[$lin-1]));
            $numcontact=0;
        }
        elsif ($lin!=$col)      #off diagonal terms of the
        contact matrix (inter residue contacts)
        {
            for (my ($i)=$startcol; $i<=$endcol;
            $i++)
            {
                for (my ($j)=$startlin;
                $j<=$endlin; $j++)
                {

                    $numcontact=$numcontact+$CONTACT_MAT[$i][$j];
                #      print "off diag i=$i
                j=$j numcontact=$numcontact
                contact_mat(i)(j)=$CONTACT_MAT[$i][$j]\n";
                }
            }
        }
    }
}

```



```

        }
        $NUMCONTACT[$col][$lin]=$numcontact;

        $A_MAT[$col][$lin]=($NUMCONTACT[$col][$lin])/(sqrt($resatomnum[$col-1]*$resatomnum[$lin-1]));
        $numcontact=0;
    }
    $startlin=$startlin+$resatomnum[$lin-1];
    $endlin=$endlin+$resatomnum[$lin];
}
# $startcol=$startcol+$resatomnum[$col-1];
# $endcol=$endcol+$resatomnum[$col-1];
# $startlin=1;
# $endlin=$resatomnum[0];
  $startcol=$startcol+$resatomnum[$col-1];
  $endcol=$endcol+$resatomnum[$col];
  $startlin=1;
  $endlin=$resatomnum[0];
}
#print "\nNUM_CONTACT\n";
#print "\nA";
for (my($i)=1; $i<=$#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=$#resatomnum+1; $j++)
    {
        print A_MATRIX "$A_MAT[$i][$j] ";
    }
    print A_MATRIX "\n";
}

my($d_value)=0;
my(@DTEMP_MAT);
for (my($i)=1; $i<=$#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=$#resatomnum+1; $j++)
    {
        if ($i != $j)
        {
            $d_value=$d_value+$A_MAT[$i][$j];
        }
    }
    $DTEMP_MAT[$i]=$d_value;
    $d_value=0;
}

for (my($i)=1; $i<=$#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=$#resatomnum+1; $j++)
    {
        if ($i==$j)
        {
            $D_MAT[$i][$j]=$DTEMP_MAT[$i];
        }
    }
}

```

```

        }
        elsif($i!=$j)
        {
            $D_MAT[$i][$j]=0;
        }
    }
}
#print "\nD";
for (my($i)=1; $i<=$#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=$#resatomnum+1; $j++)
    {
        print D_MATRIX "$D_MAT[$i][$j] ";
    }
    print D_MATRIX "\n";
}

#print "\nD_MAT\n";

#     id_value=0;
#     push(@site_num,$separate_site[0]);
#     @residues = split(/,/, $separate_site[1]);

#     for (my($i)=1; $i<=$res_num; $i++)
#     {
#         if (grep(/^$i$/, @residues))
#         {
#             print "\nyes $i";
#             push(@$mat[$line],1);
#         }

#         else
#         {
#             push(@$mat[$line],0);
#         }
#     }
#     for( my($i)=0; $i<=$#residues; $i++)
#     {
#         print "\n$i:$residues[$i]";
#     }
#     $line++;
# }
#for (my($i)=0; $i<=$line-1; $i++)
#{
#     print MATRIX "$site_num[$i] ";
#     for (my($j)=0; $j<=$res_num-1; $j++)
#     {
#         print MATRIX "$mat[$i][$j] ";
#     }
#     print MATRIX "\n";
# }

```

A perl script to calculate the A and D matrices for Ca atoms

(Written by Dr. Nahoum Anthony at SIPBS-University of Strathclyde).

```
#this script writes the A and D matrices
(PloSComputationalBiology2007_1716) from a pdb file containing
heavy atoms only.
#The PDB file has to start with the first heavy atom on the
first line and the last heavy atom on the last line.
#The script outputs a file called: "dist_your_pdb_file.txt"
#
#The script assumes the first residue as number 1, otherwise
need to modify the script !!
use Cwd;
my ($dir)=cwd();
print "\nWhat's you pdb file name? \n";
$pdb_file = <stdin>;
chomp($pdb_file);

#print "\nHow many residues does your protein contain? \n";
#$res_num = <stdin>;
#chomp($res_num);

open(SITE, "< $pdb_file") or die"input file doesn't exist";

open(A_MATRIX, ">> A_$pdb_file.txt") or die"output file not
created";
open(D_MATRIX, ">> D_$pdb_file.txt") or die"output file not
created";

my (@mat);
my ($line)=0;
my (@site_num);

my (@X);
my (@Y);
my (@Z);
my (@resnum); #will store which residue number the atom
belongs ie atom 1, belongs to residue 1, atom 7 to residue
2...
#my (@CHECK_TER);

my (@resatomnum); # will store how many atoms the residue has
is res 1 has 6 atoms, residue 2 has 7...

my(@dist);
my(@A_MAT);
my(@D_MAT);
my ($atomcount)=1;
my ($rescheck)=1;
```

```

my(@CONTACT_MAT);    #(atomxatom) size matrix which will hold a
1 for residues in contact (dist<=4 A) and a 0 otherwise
my(@NUMCONTACT);    #(resxres) size matrix which will hold the
total number of atom-atom contacts

while(<SITE>)
{
    @split_line = split(/\s+/, $_);
    push(@X, $split_line[5]);
    push(@Y, $split_line[6]);
    push(@Z, $split_line[7]);
    push(@resnum, $split_line[4]);
    push(@CHECK_TER, $split_line[0]);
    $line++;
}
#print "\n";
#for (my($i)=0; $i<=$#X; $i++)
#{
#    print "$X[$i]\t";
#}
#print "\n";
#for (my($i)=0; $i<=$#Y; $i++)
#{
#    print "$Y[$i]\t";
#}
#print "\n";
#for (my($i)=0; $i<=$#Z; $i++)
#{
#    print "$Z[$i]\t";
#}
#print "\n";
#for (my($i)=0; $i<=$#resnum; $i++)
#{
#    print "$resnum[$i]\t";
#}
#print "\n";

for (my($i)=1; $i<=$#X; $i++)
{
    for (my($j)=1; $j<=$#X; $j++)
    {
        $dist[$i][$j]=sqrt(($X[$i-1]-$X[$j-1])**2 + ($Y[$i-1]-
$Y[$j-1])**2 + ($Z[$i-1]-$Z[$j-1])**2);
#        print "$dist[$i][$j]\t";
        if (0 < $dist[$i][$j] && $dist[$i][$j] <= 7 )
#check if I shouldn't add a condition here for distance=0
        {
            $CONTACT_MAT[$i][$j]=1;
        }
        else
        {
            $CONTACT_MAT[$i][$j]=0;
        }
    }
}

```

```

#         print "$CONTACT_MAT[$i][$j]\t";
#     }
#     print "\n";
# }

for (my($i)=1; $i<=$#resnum; $i++)
{
    if($resnum[$i-1] == $resnum[$i])
    {
        $atomcount++;
    }
    else
    {
        push(@resatomnum, $atomcount);
        $atomcount=1;
    }
}

#print "\n";
#for (my($i)=0; $i<=$#resatomnum; $i++)
#{
#    print "$resatomnum[$i]\t";
#}

my ($startcol)=1;
my ($startlin)=1;
my ($endcol)=$resatomnum[0];
my ($endlin)=$resatomnum[0];
my ($numcontact)=0;
#print "\n";
for (my($col)=1; $col<=$#resatomnum+1; $col++)
{
    for (my($lin)=1; $lin<=$#resatomnum+1; $lin++)
    {
#         print"\ncol $col lin $lin startcol $startcol encol
$endcol startlin $startlin endlin $endlin resatomcol
$resatomnum[$col-1] resatomlin $resatomnum[$lin-1]";
        if ($lin==$col) #diagonal terms of the contact
matrix (intra residue contacts)
        {
            for (my($i)=$startcol+1; $i<=$endcol; $i++)
            {
                for (my ($j)=$startcol; $j<=$i-1; $j++)
                {
#shouldn't it be startlin instead of startcol ?
                    $numcontact=$numcontact+$CONTACT_MAT[$i][$j];
#                     print "diag i=$i j=$j
numcontact=$numcontact
contact_mat (i) (j)=$CONTACT_MAT[$i][$j]\n";
                }
            }
        }
    }
}

```

```

        $NUMCONTACT[$col][$lin]=0;#$numcontact;
#+$resatomnum[$col-1]; digonal term removed, no "self-contact"
counted after discussion with author of paper

$A_MAT[$col][$lin]=($NUMCONTACT[$col][$lin])/(sqrt($resatomnum
[$col-1]*$resatomnum[$lin-1]));
        $numcontact=0;
    }
    elsif ($lin!=$col)    #off diagonal terms of the
contact matrix (inter residue contacts)
    {
        for (my($i)=$startcol; $i<=$endcol;
$i++)
            {
                for (my ($j)=$startlin;
$j<=$endlin; $j++)
                    {
                        $numcontact=$numcontact+$CONTACT_MAT[$i][$j];
#                                print "off diag i=$i
j=$j numcontact=$numcontact
contact_mat (i) (j)=$CONTACT_MAT[$i][$j]\n";
                    }
                }
        $NUMCONTACT[$col][$lin]=$numcontact;

        $A_MAT[$col][$lin]=($NUMCONTACT[$col][$lin])/(sqrt($resa
tomnum[$col-1]*$resatomnum[$lin-1]));
        $numcontact=0;
    }
    $startlin=$startlin+$resatomnum[$lin-1];
    $endlin=$endlin+$resatomnum[$lin];
}
# $startcol=$startcol+$resatomnum[$col-1];
# $endcol=$endcol+$resatomnum[$col-1];
# $startlin=1;
# $endlin=$resatomnum[0];
    $startcol=$startcol+$resatomnum[$col-1];
    $endcol=$endcol+$resatomnum[$col];
    $startlin=1;
    $endlin=$resatomnum[0];
}
#print "\nNUM_CONTACT\n";
#for (my($i)=1; $i<=$#resatomnum+1; $i++)
#{
#    for (my($j)=1; $j<=$#resatomnum+1; $j++)
#    {
#        print "$NUMCONTACT[$i][$j]\t";
#    }
#    print "\n";
#}
#print "\nA_MAT\n";

```

```

#for (my($i)=1; $i<=#resatomnum+1; $i++)
#{
#    for (my($j)=1; $j<=#resatomnum+1; $j++)
#    {
#        print "$A_MAT[$i][$j]\t";
#    }
#    print "\n";
#}

#print "\nA";
for (my($i)=1; $i<=#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=#resatomnum+1; $j++)
    {
        print A_MATRIX "$A_MAT[$i][$j] ";
    }
    print A_MATRIX "\n";
}

my($d_value)=0;
my(@DTEMP_MAT);
for (my($i)=1; $i<=#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=#resatomnum+1; $j++)
    {
        if ($i != $j)
        {
            $d_value=$d_value+$A_MAT[$i][$j];
        }
    }
    $DTEMP_MAT[$i]=$d_value;
    $d_value=0;
}

for (my($i)=1; $i<=#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=#resatomnum+1; $j++)
    {
        if($i==$j)
        {
            $D_MAT[$i][$j]=$DTEMP_MAT[$i];
        }
        elsif($i!=$j)
        {
            $D_MAT[$i][$j]=0;
        }
    }
}

#print "\nD";
for (my($i)=1; $i<=#resatomnum+1; $i++)
{
    for (my($j)=1; $j<=#resatomnum+1; $j++)

```

```

        {
            print D_MATRIX "$D_MAT[$i][$j] ";
        }
    print D_MATRIX "\n";
}

#print "\nD_MAT\n";
#for (my($i)=1; $i<=$#resatomnum+1; $i++)
#{
#    for (my($j)=1; $j<=$#resatomnum+1; $j++)
#    {
#        print "$D_MAT[$i][$j]\t";
#    }
#    print "\n";
#}

#    id_value=0;
#    push(@site_num,$separate_site[0]);
#    @residues = split(/,/, $separate_site[1]);

#    for (my($i)=1; $i<=$res_num; $i++)
#    {
#        if (grep(/^$i$/, @residues))
#        {
#            print "\nyes $i";
#            push(@$mat[$line],1);
#        }
#        else
#        {
#            push(@$mat[$line],0);
#        }
#    }
#    for( my($i)=0; $i<=$#residues; $i++)
#    {
#        print "\n$i:$residues[$i]";
#    }
#    $line++;
#}

#for (my($i)=0; $i<=$line-1; $i++)
#{
#    print MATRIX "$site_num[$i] ";
#    for (my($j)=0; $j<=$res_num-1; $j++)
#    {
#        print MATRIX "$mat[$i][$j] ";
#    }
#    print MATRIX "\n";
#}

```


6.5.2 Matlab input files

Matlab input file to create a function that calculates the inverted hessian

(Provided by Anindita Dutta at Bahar lab, University of Pittsburg - Pennsylvania, United States)

```
function [invhessian] = createInvHessian(eigenVal,eigenVector)
%function [invhessian,cov_mode] =
createInvHessianFast(eigenVal,eigenVector,noOfmodes)
eigenVal = diag(eigenVal);
noOfRes = length(eigenVector);
invhessian = zeros(noOfRes,noOfRes);
noOfmodes = length(eigenVal);
for i=1:1:noOfmodes
    invhessian = invhessian + (eigenVal(i)^(-
1))*eigenVector(:,i)*eigenVector(:,i)';
end
```

Matlab input file to create a function that calculates the energy correlation matrix

(Written by Dr. Nahoum Anthony at SIPBS-University of Strathclyde).

```
function [EcorrMatrix] = Ecorrel(invKirchoff,DMatrix)
EcorrMatrix = zeros(size(invKirchoff));
DMatrix = diag(DMatrix);
tempval = 0;
for j=1:1:length(diag(DMatrix))
    for i=1:1:length(diag(DMatrix))
        for k=1:1:length(diag(DMatrix))
            tempval = tempval + DMatrix(k);
        end
        EcorrMatrix(i,j) =
(invKirchoff(i,i)+invKirchoff(j,j)-
2*invKirchoff(i,j));%*tempval;
        tempval = 0;
    end
end
```

7 REFERENCES

1. Manning, G., et al., *The protein kinase complement of the human genome*. Science, 2002. **298**(5600): p. 1912-1934.
2. Zuccotto, F., et al., *Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation*. Journal of Medicinal Chemistry, 2010. **53**(7): p. 2681-2694.
3. Ghose, A.K., et al., *Knowledge based prediction of ligand binding modes and rational inhibitor design for kinase drug discovery*. Journal of Medicinal Chemistry, 2008. **51**(17): p. 5149-5171.
4. Eglen, R.M. and T. Reisine, *Human kinome drug discovery and the emerging importance of atypical allosteric inhibitors*. Expert Opinion on Drug Discovery, 2010. **5**(3): p. 277-290.
5. Bogoyevitch, M.A. and D.P. Fairlie, *A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding*. Drug Discovery Today, 2007. **12**(15-16): p. 622-633.
6. Tsai, C.-J., A. del Sol, and R. Nussinov, *Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play*. Journal of Molecular Biology, 2008. **378**(1): p. 1-11.
7. Hilser, V.J., *An Ensemble View of Allostery*. Science, 2010. **327**(5966): p. 653-654.
8. Qiang Cui and Martin Karplus, *Allostery and cooperativity revisited*. Protein Science, 2008. **17**: p. 1295-1307.
9. Omar N. A. Demerdash, Michael D. Daily, and Julie C. Mitchell, *Structure-Based Predictive Models for Allosteric Hot Spots*. PLoS Computational Biology, 2009. **5**(10).
10. Arthur Christopoulos and Terry Kenakin, *G Protein-Coupled Receptor Allostery and Complexing*. Pharmacological Reviews, 2002. **54**(2): p. 323-374.
11. Patrick Hauske, et al., *Allosteric Regulation of Proteases*. ChemBioChem, 2008. **9**: p. 2920-2928.
12. Aron W. Fenton, *Allostery: an illustrated definition for the 'second secret of life'*. Trends in Biochemical Sciences, 2008. **33**(9): p. 420-425.

13. Roman A. Laskowski, Fabian Gerick, and Janet M. Thornton, *The structural basis of allosteric regulation in proteins*. FEBS Letters, 2009. **583**: p. 1692-1698.
14. Brian A. Kidd, David Baker, and Wendy E. Thomas, *Computation of Conformational Coupling in Allosteric Proteins*. PLoS Computational Biology, 2009 **5**(8).
15. K. Gunasekaran, Buyong Ma, and Ruth Nussinov, *Is Allostery an Intrinsic Property of All Dynamic Proteins?* PROTEINS: Structure, Function, and Bioinformatics, 2004. **57**: p. 433-443.
16. Jeanne A. Hardy and James A. Wells, *Searching for new allosteric sites in enzymes*. Current Opinion in Structural Biology, 2004. **14**: p. 1-10.
17. *Metabolism and the Regulation of Enzymes*. [cited 2012 October 28th]; Available from: <http://course1.winona.edu/kbates/Bio241/images/figure-06-21b.jpg>.
18. Aquinox. *Allosteric modulation*. [cited 2012 October 28th]; Available from: <http://www.aqxpharma.com/content/allosteric-modulation>.
19. J. Krusek, *Allostery and Cooperativity in the Interaction of Drugs with Ionic Channel Receptors*. Physiological Research, 2004. **53**: p. 569-579.
20. Abeliovich, H., *An Empirical Extremum Principle for the Hill Coefficient in Ligand-Protein Interactions Showing Negative Cooperativity*. Biophysical journal, 2005. **89**(1): p. 76-79.
21. Stevens, S.Y., et al., *Delineation of the allosteric mechanism of a cytidyltransferase exhibiting negative cooperativity*. Nat Struct Mol Biol, 2001. **8**(11): p. 947-952.
22. Andrew I. Shulman, et al., *Structural Determinants of Allosteric Ligand Activation in RXR Heterodimers*. Cell, 2004. **116**(6): p. 417-429.
23. Monod, J., J. Wyman, and J.P. Changeux, *on the nature of allosteric transitions: a plausible model*. Journal of Molecular Biology, 1965. **12**: p. 88-118.
24. Koshland, D.E., G. Némethy, and D. Filmer, *Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits*. Biochemistry, 1966. **5**(1): p. 365-385.

25. Bahar, I., C. Chennubhotla, and D. Tobi, *Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation*. Current Opinion in Structural Biology, 2007. **17**(6): p. 633-640.
26. Tsai, C.-J., A. del Sol, and R. Nussinov, *Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms*. Molecular BioSystems, 2009. **5**(3): p. 207-216.
27. Oikonomakos, N.G., et al., *A new allosteric site in glycogen phosphorylase b as a target for drug interactions*. Structure (London, England : 1993), 2000. **8**(6): p. 575-584.
28. Bryan Schmidt and Philip J Hogg, *Search for allosteric disulfide bonds in NMR structures*. BMC Structural Biology, 2007. **7**(49).
29. A. Christopoulos, et al., *G-protein-coupled receptor allostery: the promise and the problem(s)*. Biochemical Society Transactions, 2004. **32**: p. 873–877.
30. Kar, G., et al., *Allostery and population shift in drug discovery*. Current Opinion in Pharmacology, 2010. **10**(6): p. 715-722.
31. M S Smyth and J H J Martin, *x Ray crystallography*. Journal of Clinical Pathology:Molecular Pathology, 2000. **53**(1): p. 8-14.
32. Mark J. Fogg and Anthony J. Wilkinson, *Higher-throughput approaches to crystallization and crystal structure determination*. Biochemical Society Transactions 2008. **36**: p. 771-775.
33. Wlodawer, A., et al., *Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures*. FEBS Journal, 2008. **275**(1): p. 1-21.
34. Eyal, E., et al., *The Limit of Accuracy of Protein Modeling: Influence of Crystal Packing on Protein Structure*. Journal of Molecular Biology, 2005. **351**(2): p. 431-442.
35. Wolfram Gronwald and Hans Robert Kalbitzer, *Automated structure determination of proteins by NMR spectroscopy*. Progress in Nuclear Magnetic Resonance Spectroscopy, 2004. **44** p. 33-96.
36. David L. Nelson (David Lee) and Michael M. Cox, eds. *Lehninger principles of biochemistry*. 4th ed. Vol. 1. 2005, W.H. Freeman New York 1119.

37. Suzanne B. Shuker, et al., *Discovering High-Affinity Ligands for Proteins: SAR by NMR*. Science 1996. **274** p. 1531-1534.
38. Wolfgang Jahnke, et al., *Strategies for the NMR-Based Identification and Optimization of Allosteric Protein Kinase Inhibitors*. ChemBioChem, 2005. **6**: p. 1607–1610.
39. Iwai, H. and S. ger, *Protein Ligation: Applications in NMR Studies of Proteins*. Biotechnology and Genetic Engineering Reviews, 2007. **24**(1): p. 129-146.
40. Wuthrich, K., *Protein Structure Determination in Solution by NMR Spectroscopy*. The Journal of Biological Chemistry, 1990. **265**(December 25): p. 22059-22062.
41. Ad Bax and Mitsuhiro Ikura, *An efficient 3D NMR technique for correlating the proton and ¹⁵N backbone amide resonances with the α -carbon of the preceding residue in uniformly ¹⁵N/¹³C enriched proteins*. Journal of Biomolecular NMR, 1991. **1**: p. 99-104.
42. Stanley J. Opella and Francesca M. Marassi, *Structure Determination of Membrane Proteins by NMR Spectroscopy*. Chemical Reviews, 2004. **104**(No. 8): p. 3587-3606.
43. Konstantin Pervushin, et al., *Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution*. Proceedings of the National Academy of Science of the United States of America, 1997. **94**: p. 12366-12371.
44. Gaohua Liu, et al., *NMR data collection and analysis protocol for high-throughput protein structure determination*. PNAS, 2005 **102** (no. 30): p. 10487-10492.
45. Henrich, S., et al., *Computational approaches to identifying and characterizing protein binding sites for ligand design*. Journal of Molecular Recognition, 2010. **23**(2): p. 209-219.
46. Levitt, D.G. and L.J. Banaszak, *POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids*. Journal of Molecular Graphics, 1992. **10**(4): p. 229-234.

47. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins*. Journal of Molecular Graphics and Modelling, 1997. **15**(6): p. 359-363.
48. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation*. BMC Structural Biology, 2006. **6**(1): p. 19.
49. Weisel, M., E. Proschak, and G. Schneider, *PocketPicker: analysis of ligand binding-sites with shape descriptors*. Chemistry Central Journal, 2007. **1**(1): p. 7.
50. Laskowski, R.A., *SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions*. Journal of Molecular Graphics, 1995. **13**(5): p. 323-330.
51. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*. Journal of Medicinal Chemistry, 1985. **28**(7): p. 849-857.
52. Landon, M.R., et al., *Identification of Hot Spots within Druggable Binding Regions by Computational Solvent Mapping of Proteins*. Journal of Medicinal Chemistry, 2007. **50**(6): p. 1231-1240.
53. Brylinski, M. and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation*. Proceedings of the National Academy of Sciences, 2008. **105**(1): p. 129-134.
54. Lockless, S.W. and R. Ranganathan, *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families*. Science, 1999. **286**(5438): p. 295-299.
55. Hilser, V.J. and E. Freire, *Structure-based Calculation of the Equilibrium Folding Pathway of Proteins. Correlation with Hydrogen Exchange Protection Factors*. Journal of Molecular Biology, 1996. **262**(5): p. 756-772.
56. Thomas A. Halgren, *Identifying and Characterizing Binding Sites and Assessing Druggability*. Journal of Chemical Information and Modeling, 2009. **49**: p. 377-389.
57. Tom Halgren, *New Method for Fast and Accurate Binding-site Identification and Analysis*. Chemical Biology and Drug Design, 2007. **69**: p. 146-148.

58. Mehio, W., et al., *Identification of protein binding surfaces using surface triplet propensities*. Bioinformatics, 2010. **26**(20): p. 2549-2555.
59. Leighton Pritchard, et al., *Simple intrasequence difference (SID) analysis: an original method to highlight and rank sub-structural interfaces in protein folds. Application to the folds of bovine pancreatic trypsin inhibitor, phospholipase A(2), chymotrypsin and carboxypeptidase A*. Protein Engineering, 2003. **16**(2): p. 87-101.
60. Mark E. Hatley, et al., *Allosteric determinants in guanine nucleotide-binding proteins*. PNAS 2003 **100**: p. 14445-14450.
61. Andrew R. Leach, *Molecular Modelling Principles and Applications*. first ed. 1996, Essex-England: Addison Wesley Longman Limited. 1.
62. Roderick E. Hubbard, ed. *Structure-Based Drug Discovery an overview*. first ed. Biophysical and Structural Aspects of Bioenergetics, ed. M. Wikstrom. 2006, The Royal Society of Chemistry: Cambridge, UK. 54-84.
63. Lyne, P.D., *Structure-based virtual screening: an overview*. Drug Discovery Today, 2002. **7**(20): p. 1047-1055.
64. Graham L. Patrick, *An Introduction to Medicinal Chemistry*. Fourth ed. 2009, New York: Oxford University Press. 752.
65. Jonathan M. Goodman, *Chemical Applications of molecular Modelling*. 1998, Cambridge: The Royal Society of Chemistry.
66. Hans-Dieter Holtje, et al., eds. *Molecular Modeling Basic Principles and Applications*. Third ed. 2008, Willy-VCH Verlag GmbH and Co.: Weinheim.
67. Dewar, M.J.S., et al., *Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model*. Journal of the American Chemical Society, 1985. **107**(13): p. 3902-3909.
68. Stewart, J.P., *Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements*. Journal of Molecular Modeling, 2004. **10**(2): p. 155-164.
69. Patrick Bultinck, et al., eds. *Computational Medicinal Chemistry for Drug Discovery*. 2004, Marcel Dekker: New York.

70. Hinchliffe, A., *Molecular Modelling for Beginners*. Second ed. 2008, Sussex: John Wiley and Sons Ltd.
71. Christopher J Cramer, *Essentials of Computational Chemistry Theories and Models*. Second ed. 2004, Sussex: John Wiley and Sons Ltd.
72. Allinger, N.L., *Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing VI and V2 Torsional Terms*. Journal of the American Chemical Society, 1977. **99**(25): p. 8127-8134.
73. Allinger, N.L., Y.H. Yuh, and J.H. Lii, *Molecular mechanics. The MM3 force field for hydrocarbons. 1*. Journal of the American Chemical Society, 1989. **111**(23): p. 8551-8566.
74. Lii Jenn Huei and Allinger Norman L., *Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics*. Journal of the American Chemical Society, 1989. **111**(23): p. 8566-8575.
75. Lii Jenn Huei and A.N. L., *Molecular mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons*. Journal of the American Chemical Society, 1989. **111**(23): p. 8576-8582.
76. Hwang, M.J., T.P. Stockfisch, and A.T. Hagler, *Derivation of Class II Force Fields. 2. Derivation and Characterization of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules*. Journal of the American Chemical Society, 1994. **116**(6): p. 2515-2525.
77. Brooks, B.R., et al., *CHARMM: The Biomolecular Simulation Program*. Journal of Computational Chemistry, 2009. **30**(10): p. 1545-1614.
78. Andrew R. Leach, *Molecular modelling Principles and Applications*. Second ed. 2001: Pearson Education limited.
79. Guy H. Grant and W. Graham Richards, *Computational chemistry*. 1998, Oxford Science Publications: Oxford.
80. Case, D.A., et al., *AMBER 10*. 2008, University of California: San Francisco.
81. Duan, Y., et al., *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations*. Journal of Computational Chemistry, 2003. **24**(16): p. 1999-2012.

82. Cornell, W.D., et al., *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*. Journal of the American Chemical Society, 1995. **117**(19): p. 5179-5197.
83. <http://www.cambridgesoft.com>.
84. Verlet, L., *Computer "Experiments" on Classical Fluids. I. thermodynamical Properties of Lennard-Jones molecules*. Physical reviews, 1967. **159**(1): p. 6.
85. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*. Journal of Computational Physics, 1977. **23**(3): p. 327-341.
86. *Discovery Studio*. 2009, Accelrys Inc.: San Diego, CA, USA.
87. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function*. Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
88. Venkatachalam, C.M., et al., *LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites*. Journal of Molecular Graphics and Modelling, 2003. **21**(4): p. 289-307.
89. P. Therese Lang, et al., *DOCK 6.4*. 2010, University of California: San Francisco.
90. Miller, M.D., et al., *FLOG - a system to select quasi-flexible ligands complementary to a receptor of known 3-dimensional structure*. Journal of Computer-Aided Molecular Design, 1994. **8**(2): p. 153-174.
91. Kramer, B., M. Rarey, and T. Lengauer, *Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking*. Proteins-Structure Function and Genetics, 1999. **37**(2): p. 228-241.
92. Taylor, P., et al., *Ligand discovery and virtual screening using the program LIDAEUS*. British Journal of Pharmacology, 2008. **153**(S1): p. S55-S67.
93. Friesner, R.A., et al., *Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy*. Journal of Medicinal Chemistry, 2004. **47**(7): p. 1739-1749.
94. Verdonk, M.L., et al., *Improved protein-ligand docking using GOLD*. Proteins-Structure Function and Genetics, 2003. **52**(4): p. 609-623.

95. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking*. Journal of Molecular Biology, 1997. **267**(3): p. 727-748.
96. Muegge, I., *PMF Scoring Revisited*. Journal of Medicinal Chemistry, 2005. **49**(20): p. 5895-5902.
97. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics*. Journal of Molecular Graphics, 1996. **14**(1): p. 33-&.
98. D. van der Spoel, et al., *Gromacs User Manual version 4.5.4*. 2010, www.gromacs.org.
99. Besant, P.G., E. Tan, and P.V. Attwood, *Mammalian protein histidine kinases*. The International Journal of Biochemistry & Cell Biology, 2003. **35**(3): p. 297-309.
100. Wolanin, P., P. Thomason, and J. Stock, *Histidine protein kinases: key signal transducers outside the animal kingdom*. Genome Biology, 2002. **3**(10): p. reviews3013.1 - reviews3013.8.
101. Marina, A., C. Waldburger, and W. Hendrickson, *Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein*. The EMBO Journal, 2005. **24**(24): p. 4247-59.
102. Knight, Z.A., et al., *A membrane capture assay for lipid kinase activity*. Nat. Protocols, 2007. **2**(10): p. 2459-2466.
103. Prescott, S.M., *A Thematic Series on Kinases and Phosphatases That Regulate Lipid Signaling*. Journal of Biological Chemistry, 1999. **274**(13): p. 8345-8346.
104. Katso, R., et al., *Cellular function of phosphoinositide 3-kinases: Implications for Development, Immunity, Homeostasis, and Cancer*. Annual Review of Cell and Developmental Biology, 2001. **17**(1): p. 615-675.
105. R.E. Babine and S.S. Abdel-Meguid, eds. *Protein Crystallography in Drug Discovery*. Methods and principles in medicinal chemistry, ed. R. Mannhold, H. Kubinyi, and G. Folkers. Vol. 20. 2004, WILEY-VCH Verlag GmbH and Co. KGaA: Weinheim. 262.
106. Martin E. M. Noble, Jane A. Endicott, and Louise N. Johnson, *Protein Kinase Inhibitors: Insights into Drug Design from Structure*. SCIENCE, 2004. **303**: p. 1800--1805.

107. Canagarajah, B.J., et al., *Activation Mechanism of the MAP Kinase ERK2 by Dual Phosphorylation*. *Cell*, 1997. **90**(5): p. 859-869.
108. Betzi, S., et al., *Discovery of a Potential Allosteric Ligand Binding Site in CDK2*. *ACS Chemical Biology*, 2011. **6**(5): p. 492-501.
109. Simard, J.R., et al., *Development of a Fluorescent-Tagged Kinase Assay System for the Detection and Characterization of Allosteric Kinase Inhibitors*. *Journal of the American Chemical Society*, 2009. **131**(37): p. 13286-13296.
110. Gillooly, K.M., et al., *Periodic, Partial Inhibition of I kappa B Kinase beta-Mediated Signaling Yields Therapeutic Benefit in Preclinical Models of Rheumatoid Arthritis*. *Journal of Pharmacology and Experimental Therapeutics*, 2009. **331**(2): p. 349-360.
111. Rabindran, S.K., et al., *Antitumor activity of HKI-272, an orally active, irreversible inhibitor of the HER-2 tyrosine kinase*. *Cancer Research*, 2004. **64**(11): p. 3958-3965.
112. John R. Johnson, et al., *Approval Summary : Imatinib Mesylate Capsules for Treatment of Adult Patients with Newly Diagnosed Philadelphia Chromosome-positive Chronic Myelogenous Leukemia in Chronic Phase*. *Clinical Cancer Research*, 2003. **9**: p. 1972-1979.
113. Goodman, V.L., et al., *Approval summary: Sunitinib for the treatment of imatinib refractory or intolerant gastrointestinal stromal tumors and advanced renal cell carcinoma*. *Clinical Cancer Research*, 2007. **13**(5): p. 1367-1373.
114. Rock, E.P., et al., *Food and drug administration drug approval summary: Sunitinib malate for the treatment of gastrointestinal stromal tumor and advanced renal cell carcinoma*. *Oncologist*, 2007. **12**(1): p. 107-113.
115. Cohen, M.H., et al., *FDA drug approval summary: Erlotinib (Tarceva (R)) tablets*. *Oncologist*, 2005. **10**(7): p. 461-466.
116. Cohen, M.H., et al., *United States Food and Drug Administration drug approval summary: Gefitinib (ZD1839; Iressa) tablets*. *Clinical Cancer Research*, 2004. **10**(4): p. 1212-1218.

117. Michael Brave, et al., *Sprycel for Chronic Myeloid Leukemia and Philadelphia Chromosome -Positive Acute Lymphoblastic Leukemia Resistant to or Intolerant of Imatinib Mesylate* Clinical Cancer Research, 2008. **14**(2): p. 352-359.
118. Ryan, Q., et al., *FDA Drug Approval Summary: Lapatinib in Combination with Capecitabine for Previously Treated Metastatic Breast Cancer That Overexpresses HER-2*. Oncologist, 2008. **13**(10): p. 1114-1119.
119. Robert C. Kane, et al., *Sorafenib for the Treatment of Advanced Renal Cell Carcinoma*. Clinical Cancer Research, 2006. **12**: p. 7271-7278.
120. Hazarika, M., et al., *Tasigna for chronic and accelerated phase Philadelphia chromosome-positive chronic myelogenous leukemia resistant to or intolerant of imatinib*. Clinical Cancer Research, 2008. **14**(17): p. 5325-5331.
121. Comess, K.M., et al., *Discovery and Characterization of Non-ATP Site Inhibitors of the Mitogen Activated Protein (MAP) Kinases*. ACS Chemical Biology, 2010.
122. Case, D.A., et al., *The Amber biomolecular simulation programs*. Journal of Computational Chemistry, 2005. **26**(16): p. 1668-1688.
123. Pearlman, D.A., et al., *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules*. Computer Physics Communications, 1995. **91**(1-3): p. 1-41.
124. *GOLD, version 5.0.1*, The Cambridge Crystallographic Data Centre (CCDC): Cambridge, England.
125. *The PyMOL Molecular Graphics System, Version 0.99rc6*, , Schrödinger, LLC.
126. Vehlow, C., et al., *CMView: Interactive contact map visualization and analysis*. Bioinformatics, 2011.
127. *MATLAB, version 7.12.0*. 2011, The MathWorks: Massachusetts, U.S.A.
128. Shenkin, P.S., et al., *Predicting antibody hypervariable loop conformation .I. Ensembles of random conformations for ring-like structures*. Biopolymers, 1987. **26**(12): p. 2053-2085.

129. Wang, J., et al., *Development and testing of a general amber force field*. Journal of Computational Chemistry, 2004. **25**(9): p. 1157-1174.
130. Wang, J., et al., *Automatic atom type and bond type perception in molecular mechanical calculations*. Journal of Molecular Graphics and Modelling, 2006. **25**(2): p. 247-260.
131. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water*. Journal of Chemical Physics, 1983. **79**(2): p. 926-935.
132. Oliveira, C.R.d. and T. Werlang, *Ergodic hypothesis in classical statistical mechanics*. Revista Brasileira de Ensino de Física, 2007. **29**: p. 189-201.
133. Berendsen, H., *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. 2007: Cambridge University Press.
134. Berendsen, H.J.C., et al., *Molecular dynamics with coupling to an external bath*. The Journal of Chemical Physics, 1984. **81**(8): p. 3684-3690.
135. Nose, S., *A unified formulation of the constant temperature molecular dynamics methods*. The Journal of Chemical Physics, 1984. **81**(1): p. 511-519.
136. Andersen, H.C., *Molecular dynamics simulations at constant pressure and/or temperature*. The Journal of Chemical Physics, 1980. **72**(4): p. 2384-2393.
137. Izaguirre, J.A., et al., *Langevin stabilization of molecular dynamics*. The Journal of Chemical Physics, 2001. **114**(5): p. 2090-2098.
138. Crowley, M., et al., *Adventures in improving the scaling and accuracy of a parallel molecular dynamics program*. The Journal of Supercomputing, 1997. **11**(3): p. 255-278.
139. Norberg, J. and L. Nilsson, *On the Truncation of Long-Range Electrostatic Interactions in DNA*. Biophysical journal, 2000. **79**(3): p. 1537-1553.
140. Fadna, E., K. Hladeckova, and J. Koca, *Long-range electrostatic interactions in molecular dynamics: An endothelin-1 case study*. Journal of Biomolecular Structure & Dynamics, 2005. **23**(2): p. 151-162.
141. *Theory of Molecular Dynamics Simulations*. [cited 2011 13 of November]; Available from: http://www.ch.embnet.org/MD_tutorial/.

142. Kuzmanic, A. and B. Zagrovic, *Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors*. Biophysical journal, 2010. **98**(5): p. 861-871.
143. B. KNAPP, et al., *Is an Intuitive Convergence Definition of Molecular Dynamics Simulations Solely Based on the Root Mean Square Deviation Possible?* Journal of Computational Biology, 2011. **18**(8): p. 997-1005.
144. Elofsson, A. and L. Nilsson, *How Consistent are Molecular Dynamics Simulations?: Comparing Structure and Dynamics in Reduced and Oxidized Escherichia coli Thioredoxin*. Journal of Molecular Biology, 1993. **233**(4): p. 766-780.
145. Morra, G., G. Verkhivker, and G. Colombo, *Modeling Signal Propagation Mechanisms and Ligand-Based Conformational Dynamics of the Hsp90 Molecular Chaperone Full-Length Dimer*. PLoS Comput Biol, 2009. **5**(3): p. e1000323.
146. Hünenberger, P.H., A.E. Mark, and W.F. van Gunsteren, *Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations*. Journal of Molecular Biology, 1995. **252**(4): p. 492-503.
147. Luo, J. and T.C. Bruice, *Ten-nanosecond molecular dynamics simulation of the motions of the horse liver alcohol dehydrogenase-PhCH₂O⁻ complex*. Proceedings of the National Academy of Sciences, 2002. **99**(26): p. 16597-16600.
148. Di Lena, P., et al., *Fast overlapping of protein contact maps by alignment of eigenvectors*. Bioinformatics, 2010. **26**(18): p. 2250-2258.
149. Heo, Y.-S., et al., *Structural basis for the selective inhibition of JNK1 by the scaffolding protein JIP1 and SP600125*. EMBO J, 2004. **23**(11): p. 2185-2195.
150. Schulze-Gahmen, U., H.L. De Bondt, and S.-H. Kim, *High-Resolution Crystal Structures of Human Cyclin-Dependent Kinase 2 with and without ATP: Bound Waters and Natural Ligand as Guides for Inhibitor Design*. Journal of Medicinal Chemistry, 1996. **39**(23): p. 4540-4546.

151. Meagher, K.L., L.T. Redman, and H.A. Carlson, *Development of polyphosphate parameters for use with the AMBER force field*. Journal of Computational Chemistry, 2003. **24**(9): p. 1016-1025.
152. Wu, G., et al., *Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMM-based MD docking algorithm*. Journal of Computational Chemistry, 2003. **24**(13): p. 1549-1562.
153. Chennubhotla, C. and I. Bahar, *Signal Propagation in Proteins and Relation to Equilibrium Fluctuations*. PLoS Comput Biol, 2007. **3**(9): p. e172.
154. Erman, B., *Relationships between ligand binding sites, protein architecture and correlated paths of energy and conformational fluctuations*. Physical Biology, 2011. **8**(5): p. 056003.
155. Tuzmen, C. and B. Erman, *Identification of Ligand Binding Sites of Proteins Using the Gaussian Network Model*. PLoS ONE, 2011. **6**(1): p. e16474.
156. Böhm, H.-J., *LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads*. Journal of Computer-Aided Molecular Design, 1992. **6**(6): p. 593-606.
157. Böhm, H.-J., *The computer program LUDI: A new method for the de novo design of enzyme inhibitors*. Journal of Computer-Aided Molecular Design, 1992. **6**(1): p. 61-78.
158. <http://www.maybridge.com>.
159. Hartshorn, M.J., et al., *Diverse, High-Quality Test Set for the Validation of Protein–Ligand Docking Performance*. Journal of Medicinal Chemistry, 2007. **50**(4): p. 726-741.
160. Nissink, J.W.M., et al., *A new test set for validating predictions of protein–ligand interaction*. Proteins: Structure, Function, and Bioinformatics, 2002. **49**(4): p. 457-471.
161. Niesen, F.H., H. Berglund, and M. Vedadi, *The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability*. nature protocols, 2007. **2**(9).
162. MBL international corporation. *CycLex DYRK2 Kinase Assay/Inhibitor Screening Kit*. 2012; Available from: <http://www.mblintl.com/product/cy-1181>.

163. Bode, A.M. and Z. Dong, *The functional contrariety of JNK*. Molecular Carcinogenesis, 2007. **46**(8): p. 591-598.
164. Ip, Y.T. and R.J. Davis, *Signal transduction by the c-Jun N-terminal kinase (JNK) — from inflammation to development*. Current Opinion in Cell Biology, 1998. **10**(2): p. 205-219.
165. Waetzig, V. and T. Herdegen, *Context-specific inhibition of JNKs: overcoming the dilemma of protection and damage*. Trends in Pharmacological Sciences, 2005. **26**(9): p. 455-461.
166. Zhang, W., *Exploring the Intermediate States of ADP–ATP Exchange: A Simulation Study on Eg5*. The Journal of Physical Chemistry B, 2010. **115**(5): p. 784-795.
167. Rodriguez Limardo, R.G., et al., *p38 γ Activation Triggers Dynamical Changes in Allosteric Docking Sites*. Biochemistry, 2011. **50**(8): p. 1384-1395.
168. Cerutti, D.S., et al., *A Vulnerability in Popular Molecular Dynamics Packages Concerning Langevin and Andersen Dynamics*. Journal of Chemical Theory and Computation, 2008. **4**(10): p. 1669-1680.
169. Agarwal, P.K., et al., *Network of coupled promoting motions in enzyme catalysis*. Proceedings of the National Academy of Sciences, 2002. **99**(5): p. 2794-2799.
170. Lange, O.F. and H. Grubmüller, *Generalized correlation for biomolecular dynamics*. Proteins: Structure, Function, and Bioinformatics, 2006. **62**(4): p. 1053-1061.
171. Cheng, K.Y., et al., *The role of the phospho-CDK2/cyclin A recruitment site in substrate recognition*. Journal of Biological Chemistry, 2006. **281**(32): p. 23167-23179.
172. De Bondt, H.L., et al., *Crystal structure of cyclin-dependent kinase 2*. Nature, 1993. **363**(6430): p. 595-602.
173. Chiariello, M., E. Gomez, and J.S. Gutkind, *Regulation of cyclin-dependent kinase (Cdk) 2 Thr-160 phosphorylation and activity by mitogen-activated protein kinase in late G(1) phase*. Biochemical Journal, 2000. **349**: p. 869-876.

174. Gondeau, C., et al., *Design of a novel class of peptide inhibitors of cyclin-dependent kinase/cyclin activation*. Journal of Biological Chemistry, 2005. **280**(14): p. 13793-13800.
175. Aranda, S., A. Laguna, and S. de la Luna, *DYRK family of protein kinases: evolutionary relationships, biochemical properties, and functional roles*. The FASEB Journal, 2011. **25**(2): p. 449-462.
176. Lochhead, P.A., et al., *Activation-Loop Autophosphorylation Is Mediated by a Novel Transitional Intermediate Form of DYRKs*. Cell, 2005. **121**(6): p. 925-936.
177. Becker, W., et al., *Sequence Characteristics, Subcellular Localization, and Substrate Specificity of DYRK-related Kinases, a Novel Family of Dual Specificity Protein Kinases*. Journal of Biological Chemistry, 1998. **273**(40): p. 25893-25902.
178. Taira, N., et al., *DYRK2 priming phosphorylation of c-Jun and c-Myc modulates cell cycle progression in human cancer cells*. The Journal of Clinical Investigation, 2012. **122**(3): p. 859-872.
179. Taira, N., et al., *DYRK2 Is Targeted to the Nucleus and Controls p53 via Ser46 Phosphorylation in the Apoptotic Response to DNA Damage*. Molecular cell, 2007. **25**(5): p. 725-738.
180. Yoshida, K., *Role for DYRK family kinases on regulation of apoptosis*. Biochemical Pharmacology, 2008. **76**(11): p. 1389-1394.
181. Maddika, S. and J. Chen, *Protein kinase DYRK2 is a scaffold that facilitates assembly of an E3 ligase*. Nat Cell Biol, 2009. **11**(4): p. 409-419.
182. Miller, C.T., et al., *Amplification and Overexpression of the Dual-Specificity Tyrosine-(Y)-Phosphorylation Regulated Kinase 2 (DYRK2) Gene in Esophageal and Lung Adenocarcinomas*. Cancer Research, 2003. **63**(14): p. 4136-4143.
183. *Gold user guide and tutorials version 5.1*, The Cambridge Crystallographic Data Centre (CCDC).
184. Vedadi, M., et al., *Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure*

- determination*. Proceedings of the National Academy of Sciences, 2006. **103**(43): p. 15835-15840.
185. Pantoliano, M.W., et al., *High-Density Miniaturized Thermal Shift Assays as a General Strategy for Drug Discovery*. Journal of Biomolecular Screening, 2001. **6**(6): p. 429-440.
186. Lo, M.-C., et al., *Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery*. Analytical Biochemistry, 2004. **332**(1): p. 153-159.
187. Lea, W.A. and A. Simeonov, *Differential Scanning Fluorometry Signatures as Indicators of Enzyme Inhibitor Mode of Action: Case Study of Glutathione S-Transferase*. PLoS ONE, 2012. **7**(4): p. e36219.
188. Kemp, M.M., M. Weïwer, and A.N. Koehler, *Unbiased binding assays for discovering small-molecule probes and drugs*. Bioorganic & Medicinal Chemistry, 2012. **20**(6): p. 1979-1989.
189. QIAGEN. *Rapid, high-throughput assessment of protein stability on the Rotor-Gene Q cyclers*. 2010 [cited 2012 29th of October]; Available from: http://www.qiagen.com/literature/qiagennews/weeklyarticle/10_07/e07/default.aspx#fig1.
190. Chang, R., *Physical chemistry for the chemical and biological sciences*. 2000, Sausalito, California: University Science Books
191. Chang, R., *Enzyme Kinetics*, in *Physical Chemistry for the Biosciences*. 2005, University Science Books. p. 363-400.

يُؤْتِي الْحِكْمَةَ مَنْ يَشَاءُ ۚ وَمَنْ يُؤْتَ الْحِكْمَةَ فَقَدْ أُوتِيَ خَيْرًا كَثِيرًا ۗ وَمَا يَذَّكَّرُ إِلَّا أُولُو الْأَلْبَابِ

﴿البقرة: ٢٦٩﴾

“He gives wisdom to whom He wills, and whoever has been given wisdom has certainly been given much good. And none will remember except those of understanding.”

(Qura'n, 2:269)