

**Some pages of this thesis may have been removed for copyright restrictions.**

If you have discovered material in Aston Research Explorer which is unlawful e.g. breaches copyright, (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please read our [Takedown policy](#) and contact the service immediately (openaccess@aston.ac.uk)

# **Computational Investigation of Small-Molecule Human Tissue Transglutaminase Inhibitors**

Mahmood H M Jasim

Doctor of Philosophy

Aston University

December 2016

© Mahmood H M Jasim, 2016

Mahmood H M Jasim asserts his moral right to be identified as  
the author of this thesis

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without appropriate permission or acknowledgement.

**Aston University**  
**Computational investigation of small-molecule human tissue transglutaminase inhibitors**

**Mahmood H M Jasim**

**PhD**

**2016**

## **Summary**

Human tissue transglutaminase (TG2) catalyses transamidation and deamidation reactions through a nucleophilic cysteine residue (CYS277). TG2 activity was found to increase in celiac disease, cystic fibrosis, neurodegenerative disorders and cancer. For this, TG2 has received much focus as a target for drug discovery and many inhibitors have been designed and tested. The most important of these have an electrophilic warhead that reacts covalently with CYS277 resulting in an irreversible inhibition of TG2. The work presented in this thesis aimed at the development of computational methods that could aid in the design and testing of potential TG2 inhibitors. 3-D models of TG2 active site were developed starting from published X-ray crystal structures by means of docking experiments with known irreversible inhibitors followed by molecular dynamics (MD) simulations. The models were validated by additional docking runs and MD simulations involving a larger set of compounds with a range of activities against TG2. The models performed reasonably well in the validation process and were, therefore, chosen as active site models of TG2. No straightforward correlation could be found to rank the compounds based on their activities. This was the rationale for the next stage of the work, where the mechanism of inhibition of TG2 by two classes of inhibitors was studied. The covalent-bond-forming events for the inhibitors bearing acrylamide warheads were followed by applying quantum mechanics/molecular mechanics (QM/MM) umbrella sampling MD simulations to the reaction. The produced activation energies correlated well with the biological activities for the inhibitors and a mechanism with an oxyanion intermediate was proposed. The mechanism of inhibition by compounds having sulfonium ion warheads was investigated using reaction path experiments, where a transition state was first identified and verified and was used as a starting point for the reaction path. The activation energies again produced a reasonable correlation with biological activity and an S<sub>N</sub>2 mechanism was suggested for this inhibition.

On a different level, two allosteric inhibitors proposed in the literature were docked into an allosteric site in TG2 predicted by a collaborator from the University of Strathclyde, and docking complexes were subjected to accelerated MD (aMD) to inspect whether the binding would induce significant conformational changes in TG2. The binding of one inhibitor in the predicted site caused bending in TG2 structure that could be a starting event for complete TG2 inactivation. The other inhibitor seemed to produce a similar effect when bound to the original GDP binding site. An even more profound conformational change was reported due to the binding of GDP in its original binding site. aMD, for the simulation times used (400-1000 nanoseconds), was able to represent some large conformational changes in TG2 brought about by the binding of allosteric inhibitors. To sum up, the work presented in this thesis was successful in applying various computational approaches to the analysis of inhibition of TG2 with irreversible and allosteric inhibitors.

Key words: docking, molecular dynamics, umbrella sampling, allosteric inhibition.

## **Acknowledgements**

Firstly, I would like to express my sincere gratitude to my supervisor, Dr Dan Rathbone for his continuous support during my entire PhD study. His door was always open for my questions. He provided me with valuable guidance in all aspects of the research, scientific and technical. His advice helped me greatly during the writing of the thesis as well.

My profound gratitude goes to the Iraqi Ministry of Higher Education and Scientific Research, through the Iraqi Cultural Attaché in London for sponsoring me during my PhD study, in addition to my home institution, the College of Pharmacy at the University of Mosul for making life easier when applying to the sponsorship.

Special thanks is extended to Andrew Stanczak from Aston University IT Services for his tremendous efforts in setting up a suitable and convenient working environment for the AMBER program package.

I would also like to thank Dr Blair Johnston and his PhD student Antony Vassileiou from the Strathclyde Institute of Pharmacy and Biomedical Sciences at the University of Strathclyde for their help with the prediction of an allosteric binding site for TG2.

Finally, but by no means least, all the thanks go to my wife Atyaf, and my lovely boys Yousif and Mohammed for their support and patience for the entire period of my PhD.

# Table of Contents

List of Abbreviations .....	7
List of Figures .....	8
List of Tables.....	12
<b>1 General Introduction .....</b>	<b>14</b>
1.1 Transglutaminases .....	14
1.2 Tissue Transglutaminase.....	15
1.2.1 Biological functions of TG2 .....	17
1.2.2 Regulation of TG2 activity .....	18
1.2.3 Structure of TG2.....	19
1.3 Role of TG2 in Disease States .....	19
1.3.1 Celiac disease .....	20
1.3.2 Inflammatory disorders .....	21
1.3.3 Neurodegenerative diseases .....	22
1.3.4 Cancer.....	23
1.3.5 TG2 and metabolic disorders.....	24
1.4 Inhibitors of TG2 .....	24
1.4.1 Reversible inhibitors.....	25
1.4.2 Irreversible inhibitors .....	29
1.5 Computational Techniques.....	39
1.5.1 Energy minimisation.....	39
1.5.2 Molecular dynamics.....	43
1.5.3 Molecular docking.....	55
1.6 Computational Chemistry and TG2 Inhibitors .....	62
1.7 Aim and Objectives.....	63
<b>2 Materials and Methods .....</b>	<b>65</b>
2.1 Materials.....	65
2.1.1 Computers .....	65
2.1.2 Programs .....	65
2.1.3 Compounds .....	66
2.2 Methods .....	67
2.2.1 Chapter 3 methods (model development and validation).....	67
2.2.2 Chapter 4 methods (QM work) .....	75
2.2.3 Chapter 5 methods (allosteric inhibition) .....	79
<b>3 Development and Validation of TG2 Active Site Models.....</b>	<b>81</b>
3.1 Results of Initial Docking .....	81

3.1.1	TG2 active site.....	82
3.1.2	Initial docking.....	84
3.2	Analysis of MD Trajectories of Initial Docking.....	86
3.2.1	Trajectory of 1a-85ns-3.....	86
3.2.2	Trajectory of 1b-30ns-2.....	92
3.2.3	Trajectory of 1c-30ns-14.....	98
3.2.4	Trajectory of 1d-95ns-9.....	99
3.2.5	Trajectory of 1e-40ns-10.....	100
3.2.6	Trajectory of 1f-80ns-6.....	100
3.3	Docking into TG2 Models from MD Trajectories of Complexes.....	103
3.3.1	Docking into TG2 models from the 1a complex trajectory.....	104
3.3.2	Docking into TG2 models from the 1b complex trajectory.....	107
3.3.3	Docking into TG2 models from the 1c complex trajectory.....	111
3.3.4	Docking into TG2 models from the 1d complex trajectory.....	112
3.3.5	Docking into TG2 models from the 1e complex trajectory.....	112
3.3.6	Docking into TG2 models from the 1f complex trajectory.....	113
3.4	Validation of Active Site Models.....	116
3.4.1	Validation process 1 (3 inactive compounds).....	116
3.4.2	Validation process 2 (water docking).....	120
3.4.3	Validation process 3 (more test compounds).....	122
3.4.4	Validation process 4 (all compounds).....	128
3.4.5	Validation process 5 (5-ns MD simulations).....	131
3.4.6	Validation process 6 (more active compounds).....	146
3.5	Inflexible Docking.....	149
3.6	Additional Scoring Functions.....	150
3.6.1	ASP scoring function.....	151
3.6.2	ChemScore.....	151
3.6.3	CHEMPLP.....	151
3.6.4	Rescoring.....	152
3.7	Covalent Docking and MD.....	154
3.7.1	Docking.....	154
3.7.2	Molecular dynamics.....	155
3.8	Analysis of the Valid Active Site Models of TG2.....	160
3.9	Analysis of the MD Trajectory of Empty TG2.....	164
3.9.1	RMSD and PCA.....	165
3.9.2	Water behaviour.....	169
3.10	Conclusions.....	173
4	Quantum Mechanical Experiments.....	176
4.1	Umbrella Sampling.....	176
4.1.1	Introduction.....	176

4.1.2	Results and Discussion.....	177
4.2	CAChe Reaction Path Experiments .....	204
4.2.1	Introduction .....	204
4.2.2	Results and Discussion.....	206
4.3	Conclusions .....	214
5	Allosteric Inhibition of TG2 .....	216
5.1	Introduction .....	216
5.2	Methods .....	218
5.2.1	Prediction of the allosteric binding site .....	218
5.2.2	Docking.....	219
5.2.3	Molecular dynamics.....	219
5.2.4	Principal component analysis .....	221
5.3	Results and Discussion .....	221
5.3.1	Predicted allosteric binding site .....	221
5.3.2	Docking results .....	223
5.3.3	Results of accelerated MD .....	223
5.4	Conclusions .....	246
6	General Discussion and Conclusions.....	249
	References .....	255

# List of Abbreviations

Human tissue transglutaminase	TG2
Extracellular matrix	ECM
Guanosine triphosphate	GTP
Guanosine diphosphate	GDP
Celiac disease	CD
Human leukocyte antigen	HLA
Cystic fibrosis	CF
Rheumatoid arthritis	RA
Alzheimer's disease	AD
Huntington's disease	HD
Parkinson's disease	PD
Glucose-stimulated insulin secretion	GSIS
Phospholipase A <sub>2</sub>	PLA <sub>2</sub>
Benzyloxycarbonyl protective group	Cbz
tert-butyloxycarbonyl protective group	<i>t</i> -Boc
Diazo-oxo-L-norleucine	DON
Florenylmethyloxycarbonyl	Fmoc
Molecular mechanics	MM
Density functional theory	DFT
Parameterisation method 3	PM3
Molecular dynamics	MD
Assisted model building with energy refinement	AMBER
Umbrella sampling	US
Potential of mean force	PMF
Reaction coordinate	RC
Weighted histogram analysis method	WHAM
Steered molecular dynamics	SMD
Accelerated molecular dynamics	aMD
Self-consistent density functional tight-binding	SCC-DFTB
Principal component analysis	PCA
Genetic optimisation for ligand docking	GOLD
Computer-aided chemistry	CAChe
Visual molecular dynamics	VMD
General atomic and molecular electronic structure system	GAMESS
General AMBER force field	GAFF
Potential energy surface	PES
Intrinsic reaction coordinate	IRC
Astex statistical potential	ASP
Restrained electrostatic potential	RESP
Root-mean-squared-deviation	RMSD
Root-mean-squared-fluctuation	RMSF
Transition state	TS
Activation energy	AE
Reaction energy	RE
Radial distribution function	RDF
Molecular orbital package	MOPAC
Principal components	PCs

# List of Figures

Figure 1-1: Reactions catalysed by transglutaminases.....	15
Figure 1-2: TG2 in open (upper) and closed (lower) conformations.....	20
Figure 1-3: Chemical structures of the amine reversible inhibitors of TG2.....	26
Figure 1-4: Compound i1 from the work by Pardin, Pelletier, et al. (2008).....	27
Figure 1-5: Compound i2 from the work by Pardin, Roy, et al. (2008).....	28
Figure 1-6: TG2 allosteric inhibitor (i3) by Case and Stein (2007).....	29
Figure 1-7: 3-Bromo-4,5-dihydroisoxazole derivatives, left is compound i4 by Choi et al. (2005) and right is compound i5 by Watts et al. (2006).....	31
Figure 1-8: Compound i6 by Klock et al. (2014).....	32
Figure 1-9: Different warheads attached to the scaffold Cbz-Gln-Gly (R is the scaffold).....	33
Figure 1-10: Compound i7, the basic scaffold used by Pardin et al. (2006) with the different warheads.....	33
Figure 1-11: Peptide-based inhibitors derived from DON by Griffin et al. (2008) (i8).....	34
Figure 1-12: Examples of modifications to the imidazolium warhead by Badarau, Mongeot, et al. (2013) and their effect as TG2 inhibitors.....	35
Figure 1-13: Four compounds representing different series of TG2 inhibitors from the work by Prime, Andersen, et al. (2012).....	36
Figure 1-14: Compound i13, which was the basis for the compounds designed by Badarau et al. (2015).....	37
Figure 2-1: Restrained distances in umbrella sampling simulations.....	76
Figure 2-2: PES produced by map reaction.....	79
Figure 3-1: TG2 closed and open conformations.....	82
Figure 3-2: Active site tunnel and the hydrophobic loop in TG2.....	83
Figure 3-3: H-bonds between the inhibitor and TG2 in the original crystal structure.....	83
Figure 3-4: Hydrophobic interactions between the inhibitor and TG2.....	84
Figure 3-5: Docking complexes for the 6 compounds selected to be taken to molecular dynamics.....	87
Figure 3-6: Hydrophobic interactions between TG2 and 1a.....	88
Figure 3-7: The tunnel in the starting conformation of 1a trajectory. A: Surface view. B: space filling view of the bridging tryptophans.....	88
Figure 3-8: Graph of the distance between the acrylamide warhead and SG atom of CYS277 in the 1a trajectory along with RMSD for 1a alone.....	89
Figure 3-9: Distance of the H-bond between ASN333 and 1a (left) and the bond itself (right).....	89
Figure 3-10: Distance between the lipophilic part and SER309 during the simulation.....	90
Figure 3-11: Hydrophobic interactions stabilising the adamantyl group in the hydrophobic loop in 1a trajectory..	90
Figure 3-12: RMSD for TG2 in the presence and absence of compound 1a.....	91
Figure 3-13: Atomic fluctuation for TG2 residues in the trajectory of 1a.....	91
Figure 3-14: The high fluctuating helix from 1a trajectory.....	91
Figure 3-15: Changes in temperature and total energy during the simulation of 1a.....	92
Figure 3-16: A graph of the distance between the warhead carbon of 1b and SG of CYS277, with the RMSD for 1b during the MD simulation.....	92
Figure 3-17: Hydrogen bonds in 1b trajectory with ASN333 (A), PHE334 (B) and ARG317 (C) showing the acceptor atoms on 1b.....	93
Figure 3-18: Pocket formed when there is H-bonding with ARG317 (right) and absence of the pocket when the interaction is not present (left).....	93
Figure 3-19: A: distance between C42 of 1b and SER303 in the 1b trajectory. B: the same distance on a PDB structure.....	94
Figure 3-20: How the tertiary butyl group of 1b moved further into the hydrophobic loop from the starting conformation (A) to the conformation at 65 ns (B).....	95
Figure 3-21: Example of hydrophobic interactions involving the tertiary butyl part of 1b in the hydrophobic loop of TG2.....	95
Figure 3-22: Pi-cation interaction between 1b and PHE334.....	96
Figure 3-23: Plot of the distance between sulphur atom of 1b and benzene ring of PHE180.....	96
Figure 3-24: Example of pi-sulphur interactions involving TRP332 and PHE334 with the positively charged sulphur atom of 1b.....	97
Figure 3-25: RMSD of TG2 alone in 1b trajectory compared to that applied on an empty TG2.....	97
Figure 3-26: The loop between ASP152 and LEU157.....	98
Figure 3-27: Temperature and total energy graphs for the trajectory of 1b.....	98
Figure 3-28: Interactions between 1c and TG2 active site in the starting conformation.....	99
Figure 3-29: Time plot for the distance between warhead of 1c and sulphur atom of CYS123 showing the increased distance after 32 ns.....	99
Figure 3-30: Number of hydrogen bonds with 1c showing their disappearance after 32 ns.....	100
Figure 3-31: Interactions between 1e and TG2 active site in the starting conformation.....	101
Figure 3-32: Graph for the distance between warhead of 1e and sulphur of CYS123.....	101
Figure 3-33: Frequency of the H-bond between 1e and GLN276.....	101

Figure 3-34: RMSD graphs for TG2 in the simulations for compounds 1c, 1d, 1e and 1f compared to the simulation of empty TG2.....	102
Figure 3-35: Atomic fluctuations for TG2 residues in the 6 MD simulations.....	102
Figure 3-36: Example pose of 1b with the lipophilic part outside the hydrophobic loop (coloured red).....	104
Figure 3-37: Poses of 1b in the 95ns model showing interactions with TG2 residues.....	105
Figure 3-38: Poses of 1a, 1b, 1d and 1e in the 90ns model showing interactions with TG2 residues.....	107
Figure 3-39: Poses of 1a, 1b, 1e and 1f in the 100ns model of the 1a trajectory showing interactions.....	108
Figure 3-40: Poses of 1d and 1c in 65ns and 95ns models of the 1b trajectory after CAChe docking.....	108
Figure 3-41: Poses for 1b, 1c and 1e in the 65ns model showing interactions with TG2 residues.....	109
Figure 3-42: Poses for 1c, 1d, 1e and 1f in the 155ns model showing interactions with TG2 residues.....	110
Figure 3-43: Poses for 1b, 1c, 1d and 1e in the 245ns model showing interactions with TG2 residues.....	111
Figure 3-44: Pose of 1b in the 155ns model of 1e trajectory after CAChe docking.....	112
Figure 3-45: The pose of 1d in the 125ns model of 1e trajectory.....	113
Figure 3-46: Orientation of TRP332 in the 1e trajectory compared to the starting conformation (top left).....	114
Figure 3-47: The pose of 1d in the 245ns model of 1f trajectory.....	114
Figure 3-48: Poses of compounds 1a, 1b and 1f in model 65ns showing the role of ASN333 in blocking the hydrophobic region of TG2 active site. 1b is shown for comparison.....	119
Figure 3-49: Model 90ns showing the water molecules in the active site.....	120
Figure 3-50: The seven water molecules chosen by GOLD for docking of compound 1e in the 100ns model. ...	121
Figure 3-51: Hydrogen bonds between compound 1b and water in models 155ns (left) and 245ns (right).....	121
Figure 3-52: A: Pose of compound 3b in 95ns model showing the pi-sulphur interaction. B: Pose of compound 3o in 90ns model as an example.....	126
Figure 3-53: Poses of compounds 3e (left) and 3i (right) in the 95ns and 245ns models respectively, showing the pi interactions with the piperazine ring.....	128
Figure 3-54: A graphical presentation for information from Table 3-15.....	129
Figure 3-55: Correlation between scores for active compounds and biological activity in 95ns model, left is GoldScore and right is CHEMPLP.....	131
Figure 3-56: The location of residue ASN333 relative to the active site of TG2 and showing a docked inhibitor and CYS277 for reference.....	135
Figure 3-57: Percentages of active and inactive compounds that passed the 5-ns MD simulations without a significant change in their starting positions in each of the 6 active site models of TG2.....	140
Figure 3-58: Graphs for best correlation of a binding free energy value and biological activity as TG2 IC <sub>50</sub> for the 6 models.....	143
Figure 3-59: IC <sub>50</sub> versus GB (left) and PM3-GB (right) for 10 sulfonium ion compounds in Table 3-26.....	145
Figure 3-60: Correlations between IC <sub>50</sub> and GB, PM3-GB and PB for the sulfonium ion compounds from Table 3-26 after removing the least active compound (3i).....	145
Figure 3-61: A: pose of i10 showing the upright pose of the compound. B: example pose of i11 in 100ns and C is example pose of i12 in 95ns showing interactions.....	148
Figure 3-62: Performance of the 6 models during inflexible docking showing the percentages of good poses and passing the 4 criteria for active and inactive compounds.....	150
Figure 3-63: Minimised (grey carbons) and non-minimised (green carbons) conformations of compound 1e before rescoring with GOLD.....	153
Figure 3-64: Correlation between TG2 IC <sub>50</sub> and GoldScore minimised rescoring for model 65ns with the inversed trend.....	154
Figure 3-65: Correlations of minimised GoldScore and non-minimised ChemScore with IC <sub>50</sub> in models 100ns and 155ns respectively.....	154
Figure 3-66: Performance of the 6 models in covalent docking experiment.....	156
Figure 3-67: The pose of compound 1e in model 155ns as a representative for covalent docking.....	156
Figure 3-68: RMSD for the ligand only during the 15-ns covalent MD simulations for active (1e) and inactive (4e) compounds.....	157
Figure 3-69: Compound 1a from one of its simulations in which the lipophilic part of the compound straightened in the active site.....	157
Figure 3-70: Right: the angle measured between SG, EC and Cα and Left: graphs of this angle for all the simulations in which the ligands maintained their starting position.....	158
Figure 3-71: The radial distribution function of water molecules around the covalent bond during the simulations of active and inactive compounds.....	159
Figure 3-72: The same RDF graphs as in Figure 3-71 with the addition of RDF graphs for the initial simulations of compounds 1b and 1e.....	159
Figure 3-73: The conformation of the loop between residues 319 and 327 in the 6 valid models compared to the original.....	161
Figure 3-74: The orientation of ASN333 in the 6 models compared to the original.....	163
Figure 3-75: The orientation of the bridging tryptophan residues in the 6 models along with their orientation in the original crystal structure.....	164
Figure 3-76: Energy and temperature graphs for the 3 simulations applied to TG2 for 500 ns.....	164
Figure 3-77: RMSD curves for TG2 from the 3 runs.....	165
Figure 3-78: RMSD of the active site residues only for the 3, 500-ns runs on empty TG2.....	166

Figure 3-79: The first 20 principal components (PCs) generated by pyPcazip for the 3 MD runs with their eigenvalues. ....	166
Figure 3-80: The first versus second PCs for the 3 runs.....	167
Figure 3-81: Trajectories for the 1 <sup>st</sup> PC of the 3 runs.....	168
Figure 3-82: Atomic fluctuation for the individual residues of TG2 for the 3 runs.....	168
Figure 3-83: Water molecules in the 1 <sup>st</sup> (left) and 2 <sup>nd</sup> (right) solvation shells of the 1 <sup>st</sup> MD run as representative of distribution of water molecules in the 3 runs.....	169
Figure 3-84: RDF of water molecules relative to SG atom of CYS277 in the 3 MD runs .....	170
Figure 3-85: RDF of water molecules in the 275-ns trajectories of the active compounds.....	170
Figure 3-86: The location of ILE331 used as reference for the RDF of water molecules in the hydrophobic loop of TG2.....	172
Figure 3-87: RDF of water molecules in the 3 runs in the hydrophobic loop of TG2.....	172
Figure 3-88: RDF graphs for the hydrophobic loop in the trajectories of complexes.....	173
Figure 4-1: The reaction of inhibition of TG2 by acrylamide inhibitors. ....	177
Figure 4-2: Distances used during umbrella sampling for the 1-stage and 2-stage simulations. ....	178
Figure 4-3: PMF graphs of the 6 acrylamide compounds generated from WHAM program.....	181
Figure 4-4: A: Overlap graph for compound 1e using the default values for the force constant, B: overlap for compound 1e using a force constant of 300 after RC reached 4.1 Å. C: PMF graphs of the 2 simulations.....	182
Figure 4-5: The dihedral angle that was measured during umbrella sampling.....	183
Figure 4-6: Graphs of the dihedral angle in compounds 1a and 3j .....	184
Figure 4-7: RMSD graphs for the simulations of the 6 acrylamide compounds. ....	184
Figure 4-8: left: IC <sub>50</sub> versus AE, right: log of IC <sub>50</sub> versus RE.....	185
Figure 4-9: Changes in the charges (lines, left axis, Ch-O, Ch-SG and Ch-C2) of compounds 3j and 3l .....	186
Figure 4-10: Possible intermediates formed during single-stage simulation of compound 3l.....	187
Figure 4-11: The overlap for the 1 <sup>st</sup> step during the 2-step simulations for compound 1a .....	188
Figure 4-12: PMF graphs of the 6 acrylamide compounds during the 1 <sup>st</sup> step of the 2-stage reaction. ....	188
Figure 4-13: Correlation between TG2 IC <sub>50</sub> and the energy at the saddle point of the PMF graphs of the 1 <sup>st</sup> step of the 2-stage simulations for acrylamide compounds. ....	189
Figure 4-14: The overlap for the 2 <sup>nd</sup> step during the 2-step simulations for compound 1e.....	190
Figure 4-15: PMF graphs of the 6 compounds for the 2 <sup>nd</sup> step of the 2-step US simulations .....	190
Figure 4-16: Correlation between biological activity and AE for the 2 <sup>nd</sup> step of simulations on acrylamides.....	191
Figure 4-17: Charges of the 6 compounds during the 2-step simulations.....	192
Figure 4-18: The change in the charge of the carbonyl carbon of the acrylamide warhead of compound 1e during the 2-step US experiment with DFTB. ....	194
Figure 4-19: Distances around the reaction centre in the 2-step simulations for compound 3o as a representative for the remaining compounds. ....	194
Figure 4-20: RMSD for the simulations of the 6 acrylamide compounds with the 2-step US. ....	194
Figure 4-21: RDF graphs for the 6 compounds during the single-stage simulations.....	195
Figure 4-22: Graphs of the number of water molecules in the 1 <sup>st</sup> solvation shell, as a function of simulation time all for single-stage simulations.....	197
Figure 4-23: The 3 water molecules (coloured blue) that were trapped in the 1 <sup>st</sup> solvation shell during the single-stage simulation of compound 3l, in a frame after the SG-EC bond has been formed. ....	198
Figure 4-24: Distance between SG of CYS277 and the oxygen atom of the problematic water molecule during the single-stage simulation for compound 3o. ....	198
Figure 4-25: Hydrogen bonds with water molecule in compound 3o single-stage simulation. ....	198
Figure 4-26: Correlation between biological activity as IC <sub>50</sub> and AE for the acrylamide compounds, excluding compound 3o, during the single-stage PM3 US simulations.....	199
Figure 4-27: RDF graphs for the simulation of compound 3j, showing the probability calculated for the entire simulation and that calculated for the final part of the simulation after the barrier. ....	199
Figure 4-28: Graphs of the number of water molecules in the 1 <sup>st</sup> solvation shell for 2-step simulations .....	201
Figure 4-29: The closest arginine (ARG317) and lysine (LYS176) residues to C2 of the acrylamide within TG2 active site.....	203
Figure 4-30: The single-stage mechanism for the inhibition of TG2 with sulfonium ion inhibitors .....	203
Figure 4-31: A typical reaction path between reactants (R) and products (P) showing the activation energy. ....	205
Figure 4-32: PES produced from map experiment on 1b with the minimum and TS. ....	207
Figure 4-33: Left: TS structure before refinement. Right: TS structure after PM3 refinement. ....	208
Figure 4-34: IR spectrum of refined TS showing ONE negative vibration (red). ....	208
Figure 4-35: reaction path for 1b starting from reactants (left) to products (right). ....	208
Figure 4-36: Distance between SG and S+ and of the bond to be broken in the 8 TSs.....	210
Figure 4-37: Mechanism of the reaction including the change in charges and reaction coordinates between reactants, TS and products.....	211
Figure 4-38: TS structure for compound 1b in the stick form, showing EC and the three non-reacting atoms attached to it in the ball form to confirm the coplanar geometry they adopt in accordance with S <sub>N</sub> 2 mechanism. ....	212
Figure 4-39: IC <sub>50</sub> values for the 8 compounds versus their activation energies. ....	212
Figure 4-40: Correlation with biological activity for compounds with IC <sub>50</sub> of 1 μM or less. ....	212

Figure 4-41: PES (A) and reaction path (B) graphs for compound 1b when using 1 label as the reaction coordinate.....	213
Figure 5-1: Compounds i3 (left) and i2 (right) which were used as allosteric inhibitors.....	216
Figure 5-2: TG2 closed, inactive (top) and open, active (bottom) conformations.....	217
Figure 5-3: The predicted allosteric binding site by the Strathclyde protocol (top) versus the site used for docking purposes (bottom).....	219
Figure 5-4: The starting structures for the aMD simulations on TG2 surrounded by 24 copies of i3 (left) and i2 (right).....	221
Figure 5-5: TG2 in open conformation, showing the predicted allosteric binding site (yellow), the folding segment (green) and GDP binding site on the 1 <sup>st</sup> $\beta$ -barrel (red).....	222
Figure 5-6: Closed conformation of TG2 showing the binding site predicted from the trajectory of closed TG2 (red) versus the original prediction (yellow).....	222
Figure 5-7: Docking complexes for compounds i3 and i2 in the predicted allosteric site of TG2. ....	224
Figure 5-8: Atomic fluctuation values for TG2 residues during Run 1.....	225
Figure 5-9: PCs obtained by pyPcazip for the trajectory of Run 1. ....	225
Figure 5-10: Animation of PC1 for Run 1.....	226
Figure 5-11: RMSD for TG2 backbone atoms during Run 1. ....	226
Figure 5-12: Atomic fluctuation values for TG2 residues during Run 2.....	227
Figure 5-13: A comparison of TG2 structure at the end (blue and yellow) and start (red) of Run 2. ....	228
Figure 5-14: RMSD for TG2 in Run 2, compared to that of the 2 terminal $\beta$ -barrels during the same run. ....	228
Figure 5-15: The six GDP copies (in the stick form) that settled within distinct TG2 locations in Run 2.....	229
Figure 5-16: Interactions between GDP and TG2 residues, in 1KV3 crystal structure (left) and in the last frame in Run 2 (right).....	230
Figure 5-17: RMSD values for compound i2 during Run 3.....	230
Figure 5-18: A comparison between the pose of compound i2 at the start of Run 3 (left) and the pose after 395 ns of simulation time (right). ....	231
Figure 5-19: A comparison between the structure of TG2 at the start (blue) and the end (red) of Run 3.....	231
Figure 5-20: RMSD for TG2 during Run 3. ....	232
Figure 5-21: Atomic fluctuation scores for TG2 residues in Run 3.....	232
Figure 5-22: TG2 with the 24 copies of i2 at the end of Run 4.....	233
Figure 5-23: Another angle for TG2 from Figure 5-22 showing i2 copy (orange) on the GDP binding site within the 1 <sup>st</sup> barrel (coloured red). ....	233
Figure 5-24: RMSD graph for i2 copy that positioned itself in the GDP binding site during Run 4.....	233
Figure 5-25: Left: Residues within 5 Å of i2 copy at the end of Run 4. Right: Residues within 5 Å of GDP in 1KV3 crystal structure. ....	234
Figure 5-26: Atomic fluctuation values for TG2 residues during runs 3 and 4.....	235
Figure 5-27: TG2 structure at the end of Run 5 (red) compared to that at the end of Run 4 (blue). ....	236
Figure 5-28: RMSD for i2 during Run 5.....	236
Figure 5-29: Atomic fluctuation values for TG2 residues in Run 5 compared to those from Run 2. ....	237
Figure 5-30: Conformation at the end of Run 6 (red) compared to that at the end of Run 4 (green).....	237
Figure 5-31: RMSD for Run 6, compared to that of Run 5.....	238
Figure 5-32: Atomic fluctuations for TG2 residues in Run 6 compared to those in Run 5.....	238
Figure 5-33: A comparison of the pose of i3 within the predicted site (red) between the docking complex (left) and the conformation at the end of the Run 7 (right). ....	240
Figure 5-34: RMSD graphs, the Y-axis represents RMSD values in Å. ....	240
Figure 5-35: TG2 structure at the start of Run 7 (blue) and after 1 $\mu$ s (red), showing the bending of the terminal barrel.....	241
Figure 5-36: RMSD graph for TG2 residues in Run 7.....	241
Figure 5-37: Atomic fluctuations for TG2 residues during Run 7, compared to those of Run 4. ....	241
Figure 5-38: PCs produced for Run 7.....	242
Figure 5-39: The animation produced by pyPcazip for the first PC of Run 7.....	242
Figure 5-40: The animation produced by pyPcazip for the second PC of Run 7.....	243
Figure 5-41: The projections of the top 2 PCs from Run 7 into the general subspace as a function of simulation time.....	244
Figure 5-42: PC1 as a function of time for Run 1 and Run 7. ....	244
Figure 5-43: RMSD graph for TG2 during Run 8. ....	245
Figure 5-44: A comparison of TG2 structure at the start and the end of Run 8.....	245
Figure 5-45: i3 copies in the final frame in Run 8.....	246
Figure 5-46: Energy and temperature during Run 5 (top) and Run 7 (bottom). ....	247

# List of Tables

Table 1-1: Members of the transglutaminase family of enzymes (Odi & Coussons 2014).	16
Table 3-1: Structures and IC <sub>50</sub> for compounds 1a-1f.	85
Table 3-2: Good complexes from initial docking.	86
Table 3-3: Docking results for active compounds into the 95ns model.	105
Table 3-4: Docking results for active compounds into the 90ns model.	106
Table 3-5: Docking results for active compounds into the 100ns model.	106
Table 3-6: Docking results for active compounds into the 65ns model.	109
Table 3-7: Docking results for active compounds into the 155ns model.	110
Table 3-8: Docking results for active compounds into the 245ns model.	110
Table 3-9: Inactive compounds used in validation process 1.	117
Table 3-10: Docking results for validation process 1 into models from the 1a trajectory.	118
Table 3-11: Docking results for validation process 1 into models from the 1b trajectory.	119
Table 3-12: Active and inactive compounds used in validation process 3.	124
Table 3-13: Docking results of validation process 3 into models from the 1a trajectory.	125
Table 3-14: Docking results of validation process 3 into models from the 1b trajectory.	127
Table 3-15: Performance of the 6 active site models in validation process 4.	129
Table 3-16: Scores of the active and inactive compounds in model 95ns from validation process 4.	130
Table 3-17: Results of validation process 5 on the 90ns model of the trajectory of 1a.	133
Table 3-18: Results of validation process 5 on the 95ns model of the trajectory of 1a.	133
Table 3-19: Results of validation process 5 on the 100ns model of the trajectory of 1a.	134
Table 3-20: Results of validation process 5 on the 65ns model of the trajectory of 1b.	137
Table 3-21: Results of validation process 5 on the 155ns model of the trajectory of 1b.	138
Table 3-22: Results of validation process 5 on the 245ns model of the trajectory of 1b.	139
Table 3-23: Summary of the performance of the 6 models in validation process 5.	139
Table 3-24: Binding free energy values for the complexes in the 5-ns MD simulations in the models from the trajectory of 1a.	141
Table 3-25: Binding free energy values for the complexes in the 5-ns MD simulations in the models from the trajectory of 1b.	142
Table 3-26: Average binding free energy values for compounds with sulfonium ion warhead taken from all the 6 models.	144
Table 3-27: Structures of compounds from (Prime, Andersen, et al. 2012) used in validation process 6.	147
Table 3-28: Docking results of validation process 6 on all valid models.	147
Table 3-29: The results of inflexible docking on the individual models showing percentages of good poses and passing the 4 criteria in each model.	149
Table 3-30: Results of covalent docking using acrylamide compounds on the 6 models.	155
Table 3-31: RMSD values of the 6 models using the original crystal structure as a reference.	161
Table 3-32: Atomic fluctuation values of the added loop from Figure 3-73.	162
Table 3-33: Average values for the number of water molecules in the 1 <sup>st</sup> and 2 <sup>nd</sup> solvation shells in the 3 MD runs applied on empty TG2.	169
Table 4-1: Distances involved in the RC for the 1-stage simulations and the corresponding angles for the SG-EC-C2 bond.	180
Table 4-2: RC distances and SG-EC-C2 and dihedral angles for TS structures of the 6 compounds.	181
Table 4-3: AE and RE (both in kcal/mol) for the 6 acrylamide compounds resulting from US simulations using PM3 as the QM method.	185
Table 4-4: Parameters for the structures at the saddle point of the PMF graph of the simulation of the 1 <sup>st</sup> step of the 2-stage simulations.	189
Table 4-5: Parameters for the structures at the TS for the 2 <sup>nd</sup> stage of US on acrylamides using DFTB as the QM method.	191
Table 4-6: Source of the starting structures of sulfonium ion compounds. Their structures have been presented in Table 3-1 (p86) and Table 3-12 (p125).	206
Table 4-7: Reaction coordinates and energy values for the remaining 7 compounds.	209
Table 4-8: Charges and bond orders.	210

# **Chapter 1**

## **General Introduction**

# 1 General Introduction

This chapter of the thesis gives an introduction to the transglutaminase enzyme family, human tissue transglutaminase, its biological functions, regulation and role in disease states, the available classes of inhibitors and their mechanisms with example structures, and the different computational techniques applied throughout the course of this work including molecular dynamics and docking.

## 1.1 Transglutaminases

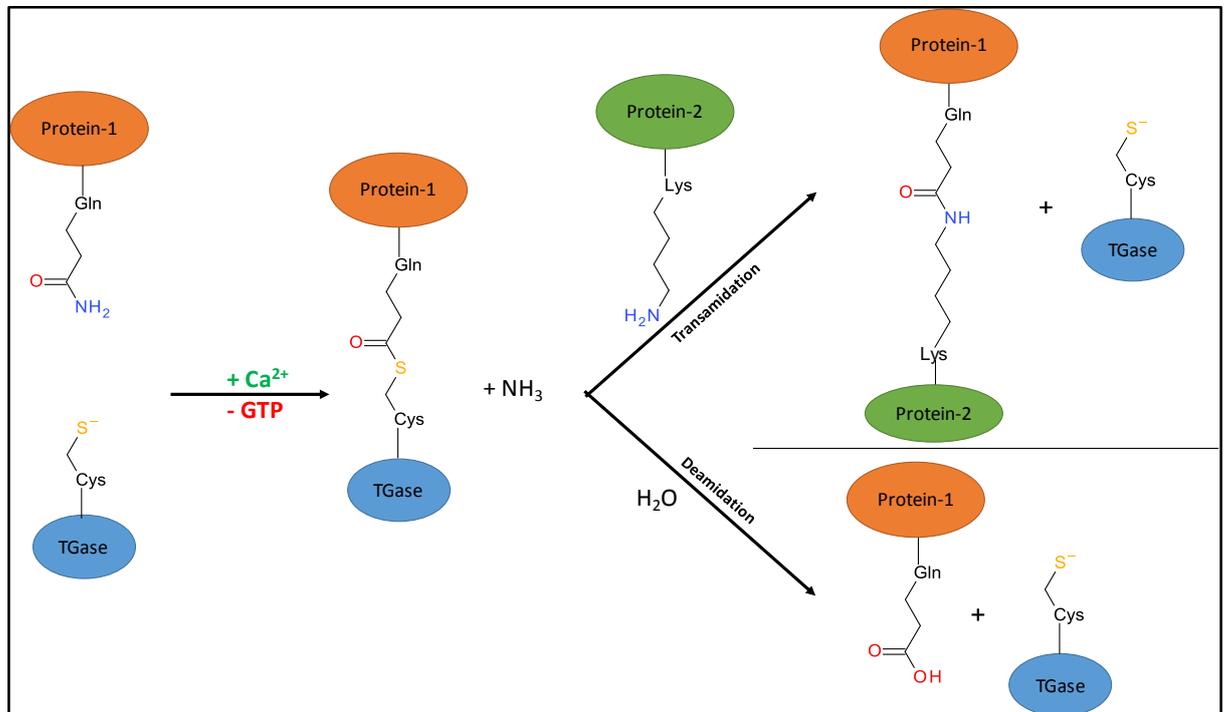
Human tissue transglutaminase (TG2) belongs to the family of transglutaminases (Martin et al. 2013). The family has an EC code of 2.3.2.13 and the other names given to the enzymes include protein-glutamine gamma-glutamyltransferase, fibrinolyase, glutamylpeptide gamma-glutamyltransferase, polyamine transglutaminase and TGase (Gasteiger et al. 2003) (<http://enzyme.expasy.org>). The enzymes, in general, are involved in the post-translational modifications of proteins (Martin et al. 2013).

Transglutaminases act as aminoacyltransferases, where they catalyse the calcium ion-dependent transamidation and deamidation reactions of the side chain of a peptidyl glutamine residue (Siegel & Khosla 2007). The amide group of a glutamine side chain acts as an acyl donor (electrophile) and a primary amine from another protein acts as an acyl acceptor (nucleophile) in transamidation reactions leading to the formation of an isopeptide bond between the side chains of the two residues on the two proteins. In deamidation reactions, water acts as the nucleophile and the result is the conversion of glutamine to glutamate. Both reactions proceed with the liberation of ammonia and the formation of an intermediate thio-ester bond between glutamine amide and an active site cysteine residue of the transglutaminase. The intermediate acts the initial electrophile (Figure 1-1). This catalytic activity is inhibited by GTP (Badarau, Collighan, et al. 2013; Siegel & Khosla 2007).

Cross-linked proteins resulting from transglutaminase catalysis strengthen the tissues against degradation by chemicals and by proteolytic enzymes, since the bond that forms during transamidation reaction is an iso-peptide bond, rather than a peptide bond that is sensitive to proteolysis. The structures are also insoluble (Pinkas et al. 2007; Candi et al. 1998). Therefore, the resultant proteins contribute to the stabilisation of hair, skin and blood clots (Badarau, Collighan, et al. 2013).

The members of the transglutaminase family are distributed in various tissues of the body and within these tissues, they may be located inside the nuclei, in the cytoplasm or in the

extracellular space. Their deficiency or increase may be associated with various pathological conditions. A summary of the locations, biological functions and the associated pathological conditions for the transglutaminases is presented in Table 1-1.



**Figure 1-1: Reactions catalysed by transglutaminases.**

All the enzymes mentioned in Table 1-1 are involved in catalysing the formation of complex protein structures through transamidation reaction shown in Figure 1-1, with the exception of band 4.2. The latter is located in the membrane of the erythrocytes and has an alanine residue instead of the active site cysteine, making it useless with this type of catalysis. Band 4.2 contributes to maintaining the shape of erythrocytes through mechanical support (Odi & Coussons 2014). Another point to mention regarding Factor XIIIa is that this is the active form of Factor XIII. The activation occurs in response to high calcium ion concentrations (Muszbek et al. 2011).

## 1.2 Tissue Transglutaminase

Tissue transglutaminase or TG2 is probably the most important member of the transglutaminase family due to its widespread distribution in the tissues of human body and its involvement in a variety of disease states (Badarau, Collighan, et al. 2013). This is the reason why TG2 was the focus of this thesis. The following sections will include more detailed information about TG2, its biological functions and substrates, its biochemical structure, its role in disease states and some of its established inhibitors.

<b>Transglutaminase</b>	<b>Other names</b>	<b>Location</b>	<b>Biological function</b>	<b>Associated pathology</b>
<b>TG1</b>	Keratinocyte TG, particulate TG, TGK	Keratinocytes and squamous epithelia. Cytosolic and membrane	Cell envelop in squamous epithelia and differentiating keratinocyte	Mutations cause Lamellar ichthyosis (Odi & Coussons 2014; Candi et al. 1998)
<b>TG2</b>	Liver TG, tissue TG, endothelial TG, erythrocyte TG	In many tissues, in the membrane, cytoplasm, nucleus, extracellular	Apoptosis, cell survival signalling, cell differentiation, matrix stabilization, endocytosis	Increased activity causes Neurodegenerative, autoimmune, malignant diseases (Facchiano et al. 2006; Odi & Coussons 2014)
<b>TG3</b>	Epidermal TG, hair follicle TG	Keratinocytes, corneocytes, hair follicles, cytoplasm	Differentiation of hair follicles	Loss may cause thinner hair (John et al. 2012; Odi & Coussons 2014)
<b>TG4</b>	Prostate TG, TGP	Prostate gland, extracellular	Reproduction and fertility in rodents (Cho et al. 2010)	Increased activity recorded in prostatic cancer (Jiang et al. 2013)
<b>TG5</b>	TGX, type 5 TG	Mainly in female reproductive system, keratinocytes, cytoplasm	Keratinocyte differentiation	Increased activity in hyperkeratotic ichthyosis and in psoriasis (Odi & Coussons 2014; Candi et al. 2002)
<b>TG6</b>	TGY, type 6 TG	Testis, lungs, brain, unknown cellular distribution	CNS development, motor function	Spinocerebellar ataxias, polyglutamine diseases (Odi & Coussons 2014; Wang et al. 2010; Guan et al. 2013)
<b>TG7</b>	TGZ, type 7 TG	Mainly testis and lungs	Unknown	Unknown (Kuramoto et al. 2013; Odi & Coussons 2014)
<b>FXIII</b>	Factor XIII A, fibrin stabilising factor, fibrinolygase, plasm TG	Platelets, placenta, astrocytes, macrophages, heart	Wound healing, bone growth, maintaining pregnancy, improving vascular permeability (Muszbek et al. 2011; Odi & Coussons 2014)	Deficiency causes impaired wound healing, abortion and severe bleeding tendency (Muszbek et al. 2011; Odi & Coussons 2014)
<b>Band 4.2</b>	Erythrocyte membrane protein band 4.2, B4.2, ATP-binding erythrocyte membrane protein band 4.2	Surface of erythrocyte membranes, bone marrow, foetal liver, spleen, found in membranes	Key component of erythrocyte skeletal network maintains erythrocyte shape and mechanical properties	Spherocytic elliptocytosis (Odi & Coussons 2014)

**Table 1-1: Members of the transglutaminase family of enzymes (Odi & Coussons 2014).**

## 1.2.1 Biological functions of TG2

Despite the widespread distribution of TG2 in tissues, it was found that mice, with genetically removed TG2, led normal lives in terms of anatomy, development and reproduction; indicating that TG2 is not a vital enzyme (Nanda et al. 2001; De Laurenzi & Melino 2001). Regarding the non-vital functions that TG2 is involved in, they can be broadly categorised into those that are calcium-dependent (transamidation and deamidation post-translational modifications), and functions that are independent of calcium and do not involve reactions at the active site cysteine residue (Odi & Coussons 2014).

### 1.2.1.1 Calcium-dependent functions of TG2

The  $\text{Ca}^{2+}$ -dependent actions of TG2 are the same as those for other members of the family, where TG2 catalyses transamidation and deamidation reactions through an active site cysteine residue (CYS277). Transamidation leads to the formation of insoluble protein-like structures, whereas deamidation leads to the conversion of glutamine to glutamate (Pinkas et al. 2007; Candi et al. 1998). The specific functions, or role in diseases, of transamidation-resultant structures depend on the location of their substrate and TG2.

TG2 located in the extracellular matrix (ECM) can catalyse the crosslinking of ECM proteins such as fibronectin, fibrin, collagen, vitronectin and osteopontin and the crosslinked proteins contribute to the stabilisation of ECM (Aeschlimann & Thomazy 2000; Belkin 2011). Within the cells, in the intracellular space whether in the nucleus or in the cytoplasm, TG2 crosslinking potential is activated in the event of necrosis. The result is the formation of the proteinaceous structures that prevent the release of inflammatory materials outside the apoptotic cells. A variety of intracellular proteins and enzymes could act as substrates for TG2 (Nicholas et al. 2003).

Besides transamidation crosslinking, TG2 also catalyses deamidation reactions in which water acts as the nucleophile and the result is the conversion of a glutamine residue in the protein to a glutamate (Figure 1-1). Of the glutamine-containing proteins susceptible to deamidation are lens  $\beta$ -crystallins (eye lens proteins responsible for lens integrity and reactivity), where age-related deamidation of glutamine residues in the N-terminal arms of the crystallins can be mediated by TG2, thereby facilitating cataract formation (Boros et al. 2008). Probably the most important deamidation reaction catalysed by TG2 is that affecting the wheat protein gliadin, resulting in epitopes that activate T cells and elicit immune response in celiac disease (Nurminskaya & Belkin 2013). This will be discussed in more detail in the section of diseases associated with TG2.

It has been reported that deamidation is favoured at low concentrations of TG2, at low pH and with poor substrates (Nurminskaya & Belkin 2013). Furthermore, deamidation and transamidation were found to be substrate specific; TG2 deamidates or transamidates specific and different glutamine residues (Boros et al. 2006).

### 1.2.1.2 Calcium-independent functions of TG2

Studies performed with the active site cysteine of TG2 (CYS277) mutated to serine have found that TG2 was still performing some biological functions, indicating that such functions are not related to TG2 calcium-dependent transamidating and deamidating activities (Siegel & Khosla 2007). One of those functions is the ability of TG2 to bind and hydrolyse GTP (guanosine triphosphate) converting it to GDP (guanosine diphosphate). This type of ability allows TG2 to function as a G-protein and transfer signals into the cells (Gundemir et al. 2012). In this fashion, TG2 was found to transport signals from  $\alpha_1$ -adrenoceptor to phospholipase C (Hwang et al. 1995). Other receptors to which TG2 was shown to act as a G-protein are oxytocin receptor (Baek et al. 1996) and thromboxane receptor (Veza et al. 1999). G-protein action of TG2 is not only  $\text{Ca}^{2+}$ -independent, but also inhibits the  $\text{Ca}^{2+}$ -dependent transamidating and deamidating activities (Gundemir et al. 2012).

The ability of acting as a protein kinase (ability to phosphorylate proteins) is another calcium-independent action of TG2. TG2 was found to be able to phosphorylate a variety of proteins, including insulin-like growth factor-binding protein-3 (IGFBP-3), histones, p53 and retinoblastoma protein (Mishra & Murphy 2004; Mishra et al. 2006; Mishra & Murphy 2006; Mishra et al. 2007). This activity of TG2 was *in vitro* only, and whether TG2 can act as a kinase *in vivo* is still unknown (Gundemir et al. 2012). TG2 can also act as protein disulphide isomerase (PDI), aiding in the formation and breakage of disulphide bonds between cysteine residues. This action was shown to occur both *in vivo* and *in vitro* (Gundemir et al. 2012).

In addition to crosslinking the proteins of ECM, TG2 was found to facilitate the adhesion of ECM to cell surfaces by non-covalent binding to fibronectin (Akimov et al. 2000). TG2 has high affinity to fibronectin and they both bind independent of  $\text{Ca}^{2+}$  or G-protein or kinase activities of TG2 (Odi & Coussons 2014). The result is improved cellular adhesion which can contribute to wound healing and embryogenesis (Akimov & Belkin 2001).

## 1.2.2 Regulation of TG2 activity

TG2 is normally latent with regard to its crosslinking and deamidating activities, which usually require the lack of GTP or GDP and high concentrations of  $\text{Ca}^{2+}$ . GTP and GDP normally bind to TG2 and induce a large conformational change in the enzyme that renders

TG2 inactive. In stressful conditions that involve cell injury, calcium homeostasis is disturbed, increasing  $\text{Ca}^{2+}$  concentration, and activating TG2 by releasing the bound GDP or GTP. This is true for intracellular TG2. Outside the cells, TG2 is also normally inactive, but not because of GTP binding. Instead, the majority of the enzyme is non-covalently bound to fibronectin making it inactive. When cells are injured, activated intracellular TG2 is released to the extracellular space causing detachment and activation of the extracellular TG2 (Wang & Griffin 2012; Siegel et al. 2008).

### 1.2.3 Structure of TG2

The PDB structure of TG2 that was used as the basis for this work had the code 2Q3Z (Pinkas et al. 2007). TG2 in this structure is in the form of active enzyme that is not bound to GTP or GDP. Active TG2 has an “open” conformation, in contrast to a “closed” conformation when the enzyme is bound to GTP or GDP (for example PDB code 1KV3 (Liu et al. 2002)). In both cases, TG2 is composed of 4 distinct regions; an N-terminal  $\beta$ -sandwich (green in Figure 1-2), a catalytic core (red in Figure 1-2) and two C-terminal  $\beta$ -barrels (yellow and cyan in Figure 1-2). The N-terminal  $\beta$ -sandwich is the part of TG2 that binds to fibronectin, and the catalytic core contains the active site cysteine residue and represents the site at which transamidation and deamidation reactions take place. In active TG2, the 4 major components are in the form of a line exposing the catalytic core and CYS277 to substrates. In the inactive closed conformation, the two C-terminal barrels fold on the catalytic core, blocking the access of substrates to CYS277 and inhibiting the transamidating and deamidating actions. A cleft between the catalytic core and the first  $\beta$ -barrel in the closed conformation holds the site at which GTP/GDP bind (Pinkas et al. 2007; Jang et al. 2014; Liu et al. 2002) (Figure 1-2).

## 1.3 Role of TG2 in Disease States

As mentioned earlier, TG2-knocked out mice could lead normal lives, indicating that TG2 is not vital to mammalian life, despite its various functions. However, it is the involvement of TG2 in a variety of disease states that makes the enzyme an attractive target in the process of drug discovery and development. The enzyme has been linked to a number of pathological conditions, including autoimmune and inflammatory diseases, neurodegenerative diseases, certain tumours and some metabolic disorders such as diabetes mellitus (Griffin et al. 2002; Facchiano et al. 2006). The exact role of TG2 in these conditions is still not entirely clear, but from the proteinaceous materials detected, it is believed that it is the  $\text{Ca}^{2+}$ -dependent catalysis of TG2 that is important (Badarau, Collighan,

et al. 2013). In the following sections, a brief description of the role of TG2 in each of the disease states will be presented.

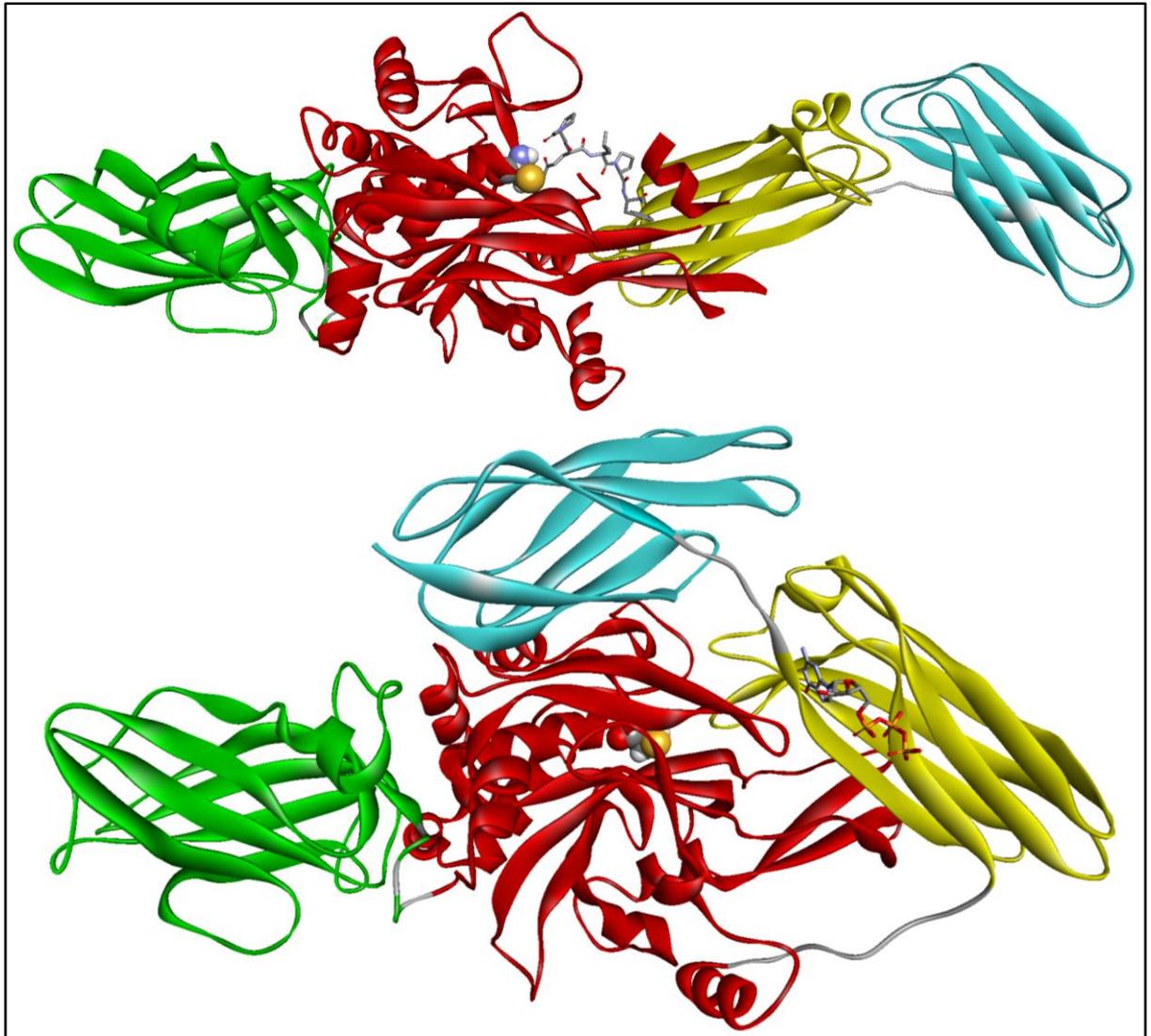


Figure 1-2: TG2 in open (upper) and closed (lower) conformations. CYS277 is shown in both in the space filling form.  $\beta$ -sandwich is coloured green, catalytic core is red, 1st  $\beta$ -barrel is yellow and the 2nd  $\beta$ -barrel is cyan. In the open conformation, there is a covalent inhibitor bound to CYS277 and shown in the stick form. In the closed conformation, GTP is shown also in the stick form. The open conformation is PDB 2Q3Z, while the closed conformation is PDB 4PYG (Jang et al. 2014).

### 1.3.1 Celiac disease

Celiac disease (coeliac disease or CD), also called celiac sprue and gluten-sensitive enteropathy, is an autoimmune disease in which TG2 plays a very distinct role. CD is characterised by sensitivity to gluten in wheat and other cereals. Gluten contains the glutamine-rich peptide, gliadin. TG2 can deamidate some of the glutamine residues on gliadin producing negatively charged glutamate residues. CD affected patients express human leukocyte antigen (HLA) molecules DQ2 and DQ8. Glutamate residues in deamidated gliadin can specifically bind to these antigenic molecules inducing an

inflammatory T-cell response that results in the flattening of the intestinal mucosa. Subsequently, nutrient absorption from the small intestine is impaired and if gluten ingestion is continued, symptoms would include chronic diarrhoea, malabsorption, anaemia, failure to thrive and weight loss (Facchiano et al. 2006; Sakly et al. 2006; Rauhavirta et al. 2013).

TG2-autoantibodies are found in the serum of CD patients, further confirming the role of TG2 in CD (Dieterich et al. 1997). These antibodies provide a diagnostic tool for CD and can be found in the intestinal mucosa even before their appearance in the serum, making their detection in the intestine a useful tool for confirming CD in potential patients (Tosco et al. 2013).

The only available standard treatment for CD is dietary gluten restriction, which is almost impossible to achieve successfully. This is due to the cost of the gluten free diet, the difficulty of obtaining gluten free diet in many regions of the world and the fact that many gluten-free diets still contain trace amounts of gluten that can induce a response in some patients (Schuppan et al. 2009). Therefore, TG2 inhibition has been suggested as a promising treatment for CD (Sollid & Khosla 2011). In a proof-of-concept study by Rauhavirta et al. (2013), the authors found that treating cells and CD-patients' intestinal biopsies with gliadin in the presence and absence of TG2 inhibitors resulted in reduction of gliadin toxic effects in the inhibitors-treated cells and biopsies. A reduction in the immune response has also been observed.

### **1.3.2 Inflammatory disorders**

Cystic fibrosis (CF) is a genetic disorder caused by mutations in the gene encoding for cystic fibrosis transmembrane regulator (CFTR) and is characterised by chronic lung inflammation and infection. It can affect other organs of the body causing pancreatic insufficiency, biliary disorders and male infertility (Ratjen & Döring 2003). Peroxisome proliferator-activated-receptor- $\gamma$  (PPAR- $\gamma$ ) is an inflammatory regulator whose ability to prevent or reduce CF pulmonary inflammation was proven through its regulation of the nuclear factor- $\kappa$ B (NF- $\kappa$ B), one of the inflammatory mediators involved in CF (Maiur et al. 2008; Perez et al. 2008). Defects in CFTR genes responsible for CF were found to be associated with increased levels and activity of TG2. It was also shown that increased TG2 activity was linked to reduced PPAR- $\gamma$  levels; where TG2 would crosslink PPAR- $\gamma$  into aggregates reducing its anti-inflammatory action. In addition, inhibiting TG2 irreversibly was found to restore the normal levels of PPAR- $\gamma$  in cells (Maiur et al. 2008).

NF- $\kappa$ B indirect activation by TG2 has also been linked to rheumatoid arthritis (RA), in which increased levels of TG2 in the synovial fluid have been reported. Increased TG2 activity was

linked to up-regulation of neutrophil gelatinase-associated lipocalin (NGAL) protein in RA patients (Katano et al. 2009). The role of TG2 in inflammation is not always harmful, where inflammation is a defence mechanism by which the body can fight back external injurious stimuli. In response to cutaneous injury, inflammation participates in the process of wound healing and TG2 activity is increased in response to some inflammatory mediators [for example, transforming growth factor (TGF)- $\beta$ ], and TG2 itself activates other mediators with the ultimate purpose of enhancing wound healing (Mehta et al. 2010).

### 1.3.3 Neurodegenerative diseases

TG2 is thought to be involved in the pathogenesis of some neurodegenerative disorders such as Alzheimer's, Huntington's and Parkinson's diseases. The mechanism is probably related to the formation of insoluble protein aggregates through calcium-dependent catalysis (Siegel & Khosla 2007; Facchiano et al. 2006). In Alzheimer's disease (AD), the most important pathological lesions are extracellular senile plaques (SP) and intracellular neurofibrillary tangles (NFT). SPs are mainly composed of amyloid beta (A $\beta$ ) peptides, whereas NFTs mainly contain hyper-phosphorylated tau. Both A $\beta$  and tau have been shown to act as substrates for TG2 (Wang et al. 2008). By its crosslinking activity, TG2 was shown to be involved in the formation of the AD lesions (SPs and NFTs), where it was found that TG2 catalyses the crosslinking between A $\beta$  peptides in SPs and between tau peptides in NFTs (Wilhelmus et al. 2009). In addition, many of the risk factors for AD, such as head trauma, ischemia and brain stress, have been shown to increase the activity of TG2 (Wang et al. 2008).

Huntington's disease (HD) is a neurodegenerative disorder characterised clinically by motor, psychological and cognitive anomalies. It is caused by a defect in the gene encoding for the huntingtin (HTT) protein. The defect includes an expansion of CAG repeats resulting in an extended polyglutamine tract at the N-terminal of HTT (Munoz-Sanjuan & Bates 2011). The polyglutamine tract gets separated from the defective HTT and forms aggregates with itself and with other proteins. The resultant macromolecular structures eventually lead to neuronal death. TG2 may catalyse the formation of these aggregates (Dedeoglu et al. 2002). Lesort et al. (1999) found that TG2 activity is increased in the brains of HD patients and this increase is directly proportional to the severity of the disease. Inhibition of TG2 with cystamine (reversible TG2 inhibitor) was shown to reduce polyglutamine aggregates and improve motor function and survival in HD mice (Dedeoglu et al. 2002). Similar outcomes were obtained by crossing HD mice with TG2 knock-out mice (Mastroberardino et al. 2002).

Parkinson's disease (PD) is another neurodegenerative disorder affecting dopaminergic neurons and is characterised by motor and non-motor dysfunctions (Jankovic 2008). The defective neurons in PD patients contain inclusions called Lewy bodies. Lewy bodies are insoluble aggregates containing a protein called  $\alpha$ -synuclein which is very prone to form aggregates in pathological conditions including PD and Lewy body dementia. It has been proven that TG2 is involved in the crosslinking of  $\alpha$ -synuclein with other proteins producing Lewy bodies *in vitro* and in cell lines (Nemes et al. 2009; Junn et al. 2003).

### 1.3.4 Cancer

Siegel & Khosla (2007) reviewed a number of studies that showed increased levels of TG2 in more than one type of cancer including glioblastomas, malignant melanomas and pancreatic ductal adenocarcinomas (PDAC). Verma et al. (2006) found that high TG2 levels in PDAC were associated with nodal metastasis and resistance to chemotherapy. The ability of TG2 to induce invasion was related to its potential to activate FAK (focal adhesion kinase), an enzyme involved in metastasis by promoting anti-apoptosis. The authors showed that this role of TG2 is not related to its transamidating or deamidating actions, by mutating the active site cysteine residue into serine residue and still obtaining comparable results (Verma et al. 2006). Similar results regarding the association between increased levels of TG2 and resistance to chemotherapy and metastasis were reported for breast cancer (Mehta et al. 2004). In addition to inducing metastasis and resistance to chemotherapy, increased TG2 levels in malignant cells was linked to their ability to survive and defy apoptosis (Mangala et al. 2007).

Park et al. (2011) proposed another mechanism by which TG2 can induce metastasis in pancreatic cancer cells. Their work examined sphingosylphosphorylcholine (SPC)-induced keratin 8 reorganisation in pancreatic cells as a possible pathway to metastasis by improving viscoelasticity and the migration properties of the malignant cells. The authors found that TG2 is activated by SPC, and the activated TG2 induces keratin 8 reorganisation by SPC. They also found that inhibiting TG2 by cystamine or genetically silencing TG2 would block the pathway resulting in reduced metastasis (Park et al. 2011). By a similar approach (gene silencing and cystamine inhibition), the research group, in another study, proved the inductive effect of TG2 on the metastasis in lung cancer cells, which is initiated through epithelial-mesenchymal-transition with the conversion of E-cadherin to N-cadherin (Park et al. 2013).

The relationship between TG2 and cancer is not simple. There are studies that showed an important role for TG2 in resisting cancer. For example, it was noted that TG2 inhibits

angiogenesis in an *in vitro* assay, independently of its transamidating activity, and that TG2 inhibits tumour growth and improves survival in mice melanoma models, compared to the same model from which TG2 was genetically removed (Jones et al. 2006). Therefore, inhibiting TG2 can be useful in certain types of cancers but the concept should not be generalised to all cancer cases.

### **1.3.5 TG2 and metabolic disorders**

Many of the enzymes that are involved in the production of energy within the body can be modified by TG2; they can act as substrates for isopeptide bond formation catalysed by TG2. Examples include glyceraldehyde 3-phosphate dehydrogenase, alpha-ketoglutarate dehydrogenase, phosphoglycerate dehydrogenase and fatty acid synthase. The same applies for hormones and receptors involved in metabolism (Facchiano et al. 2006). In diabetes mellitus, TG2 is believed to have a beneficial role where it was found that TG2-knocked out mice are intolerant to glucose and their glucose-stimulated insulin secretion (GSIS) is compromised. In addition, three TG2 point mutations were found in patients with type 2 diabetes mellitus (Bernassola et al. 2002). Salter et al. (2012) investigated the effects of these mutations on insulin secretion and found a positive correlation between the mutations and the impaired GSIS. They linked the defect in insulin secretion to a reduction in the transamidating action of TG2 initially and then to defective binding to GTP as a result of the mutations.

The association of TG2 in the above-mentioned wide range of pathologies provides the opportunity for a new therapeutic target, where inhibiting TG2 may offer possible pharmacological solutions to one or more of the pathological conditions linked to increased TG2 activity. An imperative characteristic of TG2 as a therapeutic drug target is the fact that TG2 has been shown to be non-vital to mice, reducing the possibilities of serious side effects arising from its inhibition. Because the role of TG2 in many pathologies has not yet been fully understood, TG2 potential inhibitors, in this stage of TG2-related research, can be used to provide a better appreciation of the mechanism by which TG2 can be involved (by inhibiting TG2 and monitoring the effects on the pathological condition). This concept has actually been employed in more than one of the studies discussed above. Different classes of tested TG2 inhibitors will be presented in the next section.

## **1.4 Inhibitors of TG2**

Two central themes should be considered when discussing TG2 inhibitors. The first is that the exact mechanism by which TG2 is involved in the various diseases is not fully

understood; it has been shown that sometimes it is the transamidating activity that is increased (as in celiac disease) and sometimes  $\text{Ca}^{2+}$ -independent actions are detected (some types of cancer). Therefore, it should not be expected from an inhibitor that interacts with CYS277 to be effective in suppressing all the pathologies associated with TG2. Secondly, the other members of the transglutaminase enzyme family share similar actions, especially those that are  $\text{Ca}^{2+}$ -dependent, which means that inhibiting TG2 may inhibit these enzymes as well. This is especially important for FXIIIa, whose inhibition may result in uncontrolled bleeding; and inhibition of TG1 which may lead to ichthyosis (Table 1-1). Thus, selectivity of potential inhibitors to TG2 should be verified before moving further in the process of the validation of the inhibitors.

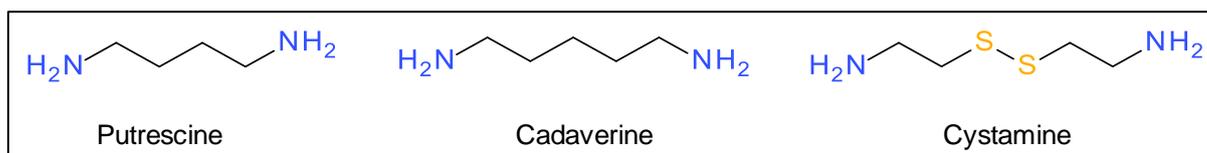
TG2 inhibitors can be classified into two broad categories, reversible inhibitors and irreversible inhibitors. Reversible inhibitors either compete with substrate lysine residues to bind to the thio-estered glutamine residue inhibiting the formation of the isopeptide bond, or they act in a similar fashion to GTP, where they bind somewhere in TG2, other than the transamidating active site, and induce a conformational change in the entire enzyme that blocks the access of substrates to CYS277. Irreversible inhibitors act by covalently binding to TG2 CYS277, and therefore inhibiting the enzyme in a suicidal manner (Badarau, Collighan, et al. 2013; Siegel & Khosla 2007).

### **1.4.1 Reversible inhibitors**

The net result from these inhibitors is either the formation of the isopeptide bond with the wrong amine (competitive inhibitors) or a temporary loss of TG2 action as a result of a large conformational change induced through an allosteric inhibition (non-competitive inhibitors).

#### **1.4.1.1 Competitive reversible inhibitors**

This class of TG2 inhibitors is characterised by a structure having a saturated aliphatic chain and a terminal primary amine. They act as substrates for the 2<sup>nd</sup> step in TG2-catalysed transamidation and deamidation reaction and the result is still an isopeptide bond but with the amine inhibitor instead of being with the natural amine substrate or with water. Examples of these inhibitors include putrescine, cadaverine derivatives (Lorand & Conrad 1984; Badarau, Collighan, et al. 2013) and cystamine (Jeitner et al. 2005) (Figure 1-3). Putrescine has been used by many for assays of TG2 activity and examples for putrescine use include the work by Schaertl et al. (2010) and Ohtake et al. (2006). Cadaverine derivatives were also used for the same purpose and examples include van den Akker et al. (2012), Olsen et al. (2011) and Sarang et al. (2007).



**Figure 1-3: Chemical structures of the amine reversible inhibitors of TG2.**

Cystamine is actually composed of 2 molecules of cysteamine [ $\beta$ -mercaptoethanolamine (MEA),  $\text{NH}_2\text{-(CH}_2\text{)}_2\text{-SH}$ ] connected by a disulphide bond. Two mechanisms have been proposed, by which cystamine can inhibit TG2. The first involves the intracellular reduction of cystamine to 2 molecules of MEA by glutathione (GSH). MEA then acts as a substrate for the 2<sup>nd</sup> step of transamidation, and an isopeptide bond forms between MEA and the 1<sup>st</sup> protein glutamine residue (Jeitner et al. 2005). This has been proven by the finding that it requires twice the molar equivalents of MEA to inhibit TG2 than what is required of cystamine (Cooper et al. 2002).

Jeitner et al. (2005) have shown that the inhibition of TG2 by MEA is concentration-dependent confirming that the inhibition is reversible. The authors suggested that the MEA thiolate group, by having an electron-withdrawing effect, would help in positioning MEA perfectly to be involved in the formation of the isopeptide bond through interacting with some positively charged moiety in the active site of TG2 due to its negative charge. The neutralisation of the negative charge by this interaction would also increase the nucleophilicity of MEA amine, thus increasing the chance for the formation of the bond.

The second mechanism by which cystamine may inhibit TG2 is the formation of disulphide link with CYS277 causing irreversible inhibition of TG2. This mechanism was supported by the evidence that cystamine inhibited TG2 in a time-dependent manner (Lorand & Conrad 1984; Lorand 1998; Siegel & Khosla 2007). Jeitner et al. (2005) reported that if cystamine was reduced intracellularly to MEA, then the disulphide bond formation will be an unlikely event, implying that cystamine will act as an irreversible inhibitor only in the absence of intracellular GSH.

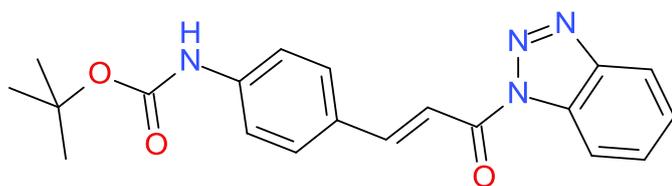
In addition to TG2, cystamine has been found to inhibit caspase 3 (an enzyme whose activation was linked with Huntington's disease progression). Inhibition of caspase 3 may contribute to the beneficial effects cystamine has in Huntington's disease besides inhibiting TG2 (Lesort et al. 2003). The authors indicated that caspase inhibition is independent of TG2 inhibition.

Problems associated with inhibition by cystamine include the formation of proteinaceous structures resulting from the use of MEA as the amine source in transamidation reactions. These structures may elicit an immune response for being unrecognisable by the body.

Another problem with cystamine is its lack of selectivity; cystamine was shown to inhibit caspase 3 and can inhibit other members of transglutaminase family, and this could result in undesirable effects (Badarau, Collighan, et al. 2013). Nevertheless, cystamine can still be used in research on cells overexpressing TG2 as an indicator of the beneficial effects arising from TG2 inhibition, and it has been used for this purpose as indicated earlier in HD and in cancer.

Other competitive TG2 inhibitors were proposed by Sohn et al. (2003) who synthesised recombinant peptides with dual inhibitory effect against phospholipase A<sub>2</sub> (PLA<sub>2</sub>) and TG2. The former is one of the enzymes involved in inflammatory processes and is activated by TG2 through the formation of intramolecular crosslinks between glutamine and lysine residues in PLA<sub>2</sub>. The inhibitory peptides were derived from TG2 substrates lipocortin-1 and pro-elafin and contained glutamine and/or lysine residues, making them suitable substrates for either step of the transamidation reaction, and hence the reversible action. The synthesised peptides were able to inhibit PLA<sub>2</sub> alone, TG2 alone and PLA<sub>2</sub> activated by TG2. Their results showed improved allergic conjunctivitis in a guinea pig model treated with these peptides (Sohn et al. 2003). The study did not examine the selectivity of these inhibitors against other members of the transglutaminase family.

Another group of reversible inhibitors was studied by Pardin, Pelletier, et al. (2008). The researchers synthesised a series of cinnamoyl derivatives and tested them against TG2. Their findings showed two classes of potent reversible inhibitors of TG2 that compete with the acyl donor substrate during transamidation reactions, the cinnamoyl benzotriazolyl amides and 3-(substituted cinnamoyl) pyridines. The most active of their compounds (**i1**) was a benzotriazolyl derivative with a TG2 IC<sub>50</sub> of 18 μM (Figure 1-4).



**Figure 1-4: Compound i1 from the work by Pardin, Pelletier, et al. (2008) with an IC<sub>50</sub> against TG2 of 18 μM.**

From structure activity relationship studies, the authors proposed the importance of hydrogen bond acceptors in the inhibitor structures for better activity. They also proposed that non-planar rings can reduce the inhibitory effect due to steric hindrance at the transamidation active site in TG2. The most active compounds were also tested against FXIIIa and caspase 3. There was very little or no activity against these enzymes inferring selectivity against TG2 (Pardin, Pelletier, et al. 2008). It is worth mentioning here that the

researchers used guinea pig liver TG2 for testing for its ease of attaining and its 80% structure similarity with human TG2.

The same research group followed on their work and synthesised a new series of compounds based on their most active compounds, keeping the cinnamoyl and the triazole groups and modifying the aromatic ring to enhance the inhibitor binding within TG2 active site (Pardin, Roy, et al. 2008). Their most active compound (**i2**) (Figure 1-5) in this group had an  $IC_{50}$  against guinea pig liver TG2 of 2.1  $\mu\text{M}$ , compared to 18  $\mu\text{M}$  from their previous research. The authors showed that a similar mechanism was followed by this group of inhibitors in being reversible and competitive TG2 inhibitors (Pardin, Roy, et al. 2008). No report of examining the selectivity of this inhibitor group was mentioned in the paper.

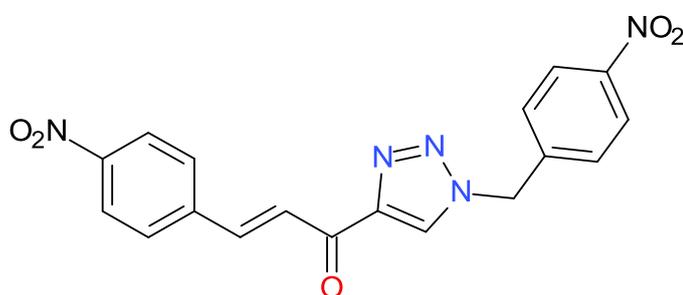


Figure 1-5: Compound **i2** from the work by Pardin, Roy, et al. (2008).

#### 1.4.1.2 Non-competitive reversible inhibitors

It has been mentioned earlier that TG2 activity is dependent on calcium ions. This activity can be inhibited by the binding of guanine nucleotides such as GTP and GDP. These natural inhibitors bind to an allosteric site in TG2 other than the transamidation active site and induce conformational changes that result in the loss of enzyme activity. This inhibition can be reversed by high concentrations of calcium ions (Liu et al. 2002). The conformational change has been introduced previously and the literature showed that the site at which GTP and GDP bind in TG2 is distinct from the transamidation active site. In theory, it is therefore possible to inhibit TG2 reversibly using compounds that would induce the conformational change triggered by the natural TG2 inactivators, GTP and GDP.

Case and Stein (2007), after screening a library of 110,000 drug-like molecules, were able to find one compound that inhibited TG2 reversibly in a manner similar to that of GTP. The  $IC_{50}$  of this compound (**i3**, Figure 1-6) was recorded to be 0.6  $\mu\text{M}$  and a selectivity study showed no effect of the compound on FXIIIa and caspase.

The authors proposed an allosteric inhibition mechanism through proving that the inhibitor blocks the activation of TG2 by  $\text{Ca}^{2+}$  by a steady state velocity analysis. They could not, however, prove that the inhibitor binds to GTP binding site, so they suggested that the

inhibitor could bind to the GTP binding site or to a different site in TG2 that regulates the binding of GTP, but not at the transamidating site. They have also shown that the inhibitory effect of their compound is dependent on the structure of the substrate used in the assay, suggesting a competitive mechanism. The authors explained this by the fact that the allosteric inhibitors can bind not only to the active free enzyme, but also to enzyme-substrate complexes and in the latter case, the structure of the substrate is relevant (Case & Stein 2007).

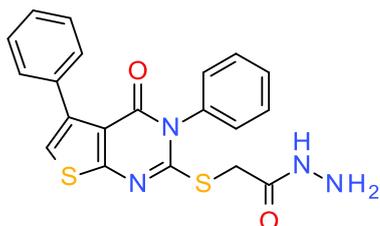


Figure 1-6: TG2 allosteric inhibitor (i3) by Case and Stein (2007).

Caron et al. (2012) used Förster resonance energy transfer (FRET) as an alternative method to X-ray crystallography for elucidating the changes in conformation of TG2 in response to activation. They used the peptide inhibitor that was bound to TG2 in the crystal structure with the PDB entry 2Q3Z by Pinkas et al. (2007) and proved that this compound inhibits TG2 by covalently binding to the active site cysteine residue locking TG2 in the open conformation. They also used the most potent inhibitor from the work by Pardin, Roy, et al. (2008) (compound **i2** Figure 1-5) and showed that **i2** inhibits TG2 by maintaining the closed conformation of the enzyme and blocking the transamidating activity. This may contradict the results reported by Pardin, Roy, et al. (2008), but the closing of TG2 conformation may occur after binding of the inhibitor.

Both types of reversible inhibitors have the potential to inhibit other members of the transglutaminase family, indicating a lack of selectivity. In addition, amine competitive inhibitors may induce an immune reaction resulting from the formation of non-native structures, which could limit their applications. These limitations, besides the moderate potencies of the reversible inhibitors and the availability of the active site nucleophilic cysteine residue, were probably the reasons for studying irreversible inhibitors.

### 1.4.2 Irreversible inhibitors

These TG2 inhibitors usually have an electrophilic warhead that can easily react with the nucleophilic sulphur atom of CYS277 by forming a covalent bond. Since the nucleophilic cysteine residue is essential for TG2 activity, the result of the reaction with these inhibitors is complete loss of TG2 activity. The design of such inhibitors is usually based on the

glutamine-bearing substrates of TG2, where small similar peptides are designed, to which an electrophilic warhead is incorporated to achieve the irreversible inhibition (Badarau, Collighan, et al. 2013).

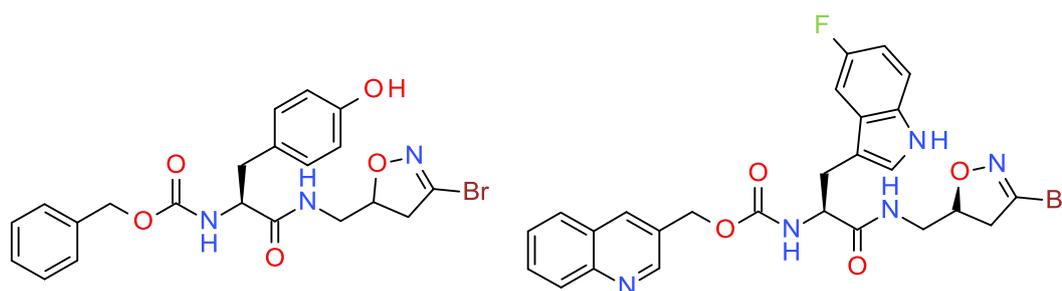
The simplest compound that has been used as an irreversible inhibitor is iodoacetamide [ $\text{NH}_2\text{-CO-CH}_2\text{-I}$ ] which is a known blocker of biochemical thiols, including the thiol group of cysteine residues (Reisz et al. 2013). Iodoacetamide has been used to inactivate TG2 during tests of the activity of the enzyme. Examples of such uses include the work performed by Dørum et al. (2010) and de Macédo et al. (2000). Iodine in the structure of iodoacetamide serves as a good leaving group increasing the electrophilicity of the terminal methyl group. However, the small size and high reactivity makes it possible for iodoacetamide to interact with other thiols within TG2 or within other thiol bearing enzymes. This probably limits the use of the compound to testing the activity of TG2 (Siegel & Khosla 2007).

#### 1.4.2.1 Peptide based inhibitors

In addition to the electrophilic warhead and the amino acids, peptide-based inhibitors should have in their structure a hydrophobic group on the opposite side from the warhead to improve the positioning of the inhibitor within the active site of TG2, which has a hydrophobic loop (Pinkas et al. 2007) (more details on this will follow in the relevant Results sections). Among the peptide-based inhibitors, the dihydroisoxazole derivatives received the most attention and were validated extensively. They are derived from acivicin which is an isoster of glutamine. These inhibitors are characterised by a 3-halo-4,5-dihydroisoxazole derivative warhead attached to an amino acid by a peptide bond, with an aromatic group to increase specificity to TG2 active site (Choi et al. 2005; Watts et al. 2006).

One of the best members of this group was discovered by Choi et al. (2005). The compound (**i4**) (Figure 1-7) has tyrosine as the amino acid and benzyloxycarbonyl (Cbz) group as the hydrophobic group. **i4** was found to be very active as a TG2 inhibitor and, when tested on human colonic cancer cell line, it sensitised the cells to epothilone C, a tubulin binding anticancer agent. Similar results were obtained when compound **i4** was injected into mice with glioblastoma. The compound also showed good oral bioavailability and low toxicity when administered to mice. The compound was not tested against other transglutaminases but it was incubated with GSH as a sulphur containing electrophile and GSH was not affected. GSH was chosen because it represents the most abundant physiological thiol. Structure activity relationship (SAR) studies showed that a bulky group such as the Cbz was important for activity as was the amide group connecting the dihydroisoxazole group to the remainder of the inhibitor molecule (Choi et al. 2005).

The same group followed on the work on the same chemical class (Watts et al. 2006). They first proved that these inhibitors act by covalently binding to CYS277 of TG2. This was performed using mass spectrometry and it was found that the mass of the cysteine residue increases by a factor corresponding to the mass of the inhibitor minus that of bromine. Then, the authors synthesised a different series of compounds, changing the hydrophobic part and the amino acid from their previous set to enhance the activity. Their most active compound (**i5** Figure 1-7) was about 50-fold more potent than compound **i4**. **i5** had tryptophan as the amino acid and a quinoline group for the hydrophobic part.



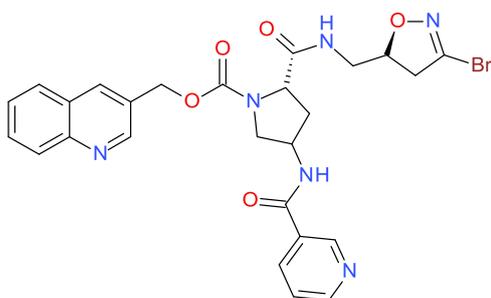
**Figure 1-7: 3-Bromo-4,5-dihydroisoxazole derivatives, left is compound **i4** by Choi et al. (2005) and right is compound **i5** by Watts et al. (2006).**

In addition, the authors have set the criteria for a dihydroisoxazole derivative to have increased activity against TG2. The criteria are (*S*)-stereochemistry at position 5 of the dihydroisoxazole moiety, the presence of a heteroatom in the side chain of the amino acid and the presence of hydrogen bond acceptor in the hydrophobic part of the inhibitor. Although the article stated that the inhibitors were selective, no measures were taken in the study to ascertain this (Watts et al. 2006).

Another study came out from the same laboratory (Klock et al. 2014) in which a library of published dihydroisoxazole TG2 inhibitors that includes around 60 compounds was tested for the selectivity of the inhibitors against TG1, TG3 and FXIIIa. It was found that many compounds were inactive against TG3 or FXIIIa, but were active in inhibiting TG1. The authors attributed TG1 activity to the presence of aromatic amino acids in the structures of the inhibitors. Consequently, they designed and tested a series of dihydroisoxazole derivatives containing substituted proline instead of tyrosine or tryptophan. Their most active compound was about 5 times more potent against TG2 than against TG1 (**i6** Figure 1-8). The compound has a quinoline group and a nicotinamide substituted proline residue (Klock et al. 2014).

For other peptide-based irreversible TG2 inhibitors, Chica et al. (2004) were able to define the possible structure of small peptide-based acyl donor substrates for TG2. Understanding how these substrates bind to TG2 would help in the design of related irreversible inhibitors.

They concluded that for best results, the peptide must contain in its structure a Cbz protective group with glutamine and glycine residues (Cbz-Gln-Gly) (carbobenzyloxy-L-glutaminyglycine). They also showed that tert-butyloxycarbonyl protective group (*t*-Boc) in the structure would make it unsuitable as a substrate. This is consistent with the finding from Choi et al. (2005) regarding the importance of the Cbz group and the requirement for the presence of glutamine residue in the structure.



**Figure 1-8: Compound i6 by Klock et al. (2014).**

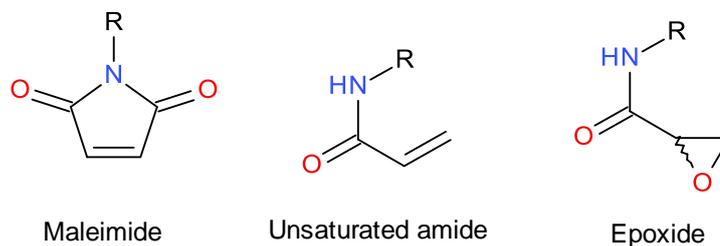
Even though the researchers used the crystal structure of red sea bream TG2, which has about 50% structure similarity in the active site with guinea pig TG2, their basic substrate was the basis for non-dihydroisoxazole inhibitors designed by others (Chica et al. 2004). Different functional groups have been attached to the glutamine residue in the above substrate by different researchers to produce various groups of irreversible TG2 inhibitors. Examples of these functional groups are maleimides, epoxides and  $\alpha,\beta$ -unsaturated amides.

Maleimide warhead inhibitors (Figure 1-9), based on the above scaffold, have been studied by Halim et al. (2007) who have shown that such inhibitors demonstrated time-dependent irreversible inhibition, suggesting reaction with the active site cysteine residue of TG2. Their compounds were not very effective when compared to other warheads, such as acrylamides. The authors attributed this to the physical size of the maleimides warhead group and the difficulty of its approach to CYS277 (Halim et al. 2007).

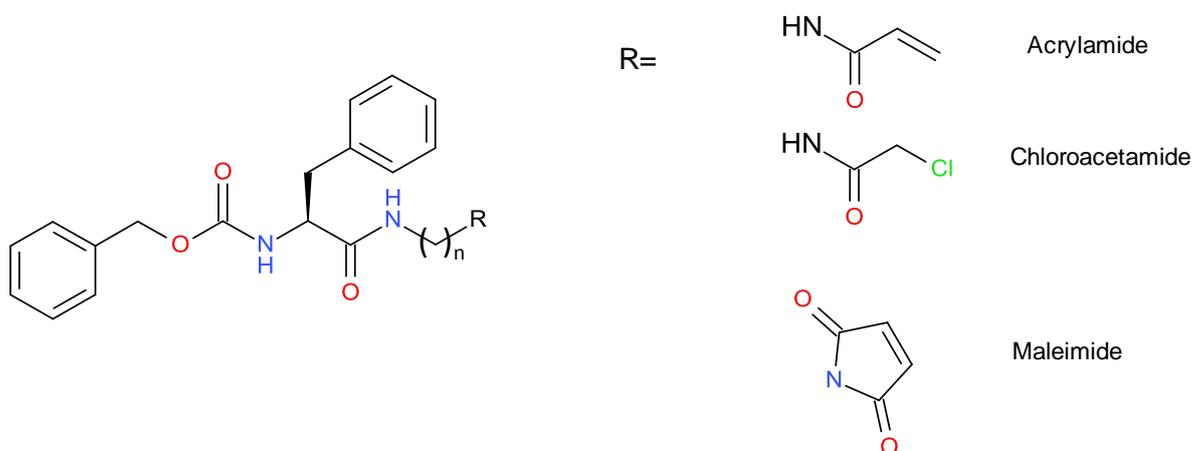
Based on the same scaffold, de Macédo et al. (2002) synthesised a series of compounds containing  $\alpha,\beta$ -unsaturated amides and their corresponding epoxides as the warhead (Figure 1-9). They did not test the compounds for their activity against TG2 but they proposed that they would be potential inhibitors for the enzyme.

Pardin et al. (2006) synthesised and tested a series of peptide based inhibitors. They used the scaffold by Choi et al. (2005) as the basis for their peptide but replaced tyrosine by phenylalanine. Their primary peptide structure composed of a Cbz protective group, phenylalanine, an aliphatic spacer and an electrophile. The electrophile was one of three

types, an acrylamide ( $\alpha,\beta$ -unsaturated amide), a chloroacetamide or a maleimide (compound **i7**, Figure 1-10).



**Figure 1-9: Different warheads attached to the scaffold Cbz-Gln-Gly (R is the scaffold).**



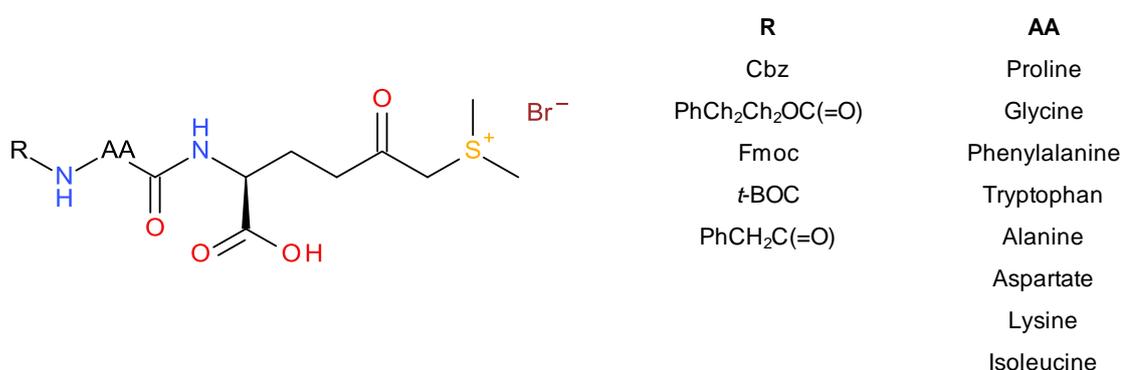
**Figure 1-10: Compound **i7**, the basic scaffold used by Pardin et al. (2006) with the different warheads.**

The acrylamide and chloroacetamide compounds showed good inhibitory activity against TG2, with inhibitory efficiencies that were 4-50 times higher than the dihydroisoxazole compound, **i4**, by Choi et al. (2005). It was also found that this activity is related to the length of the aliphatic spacer, where the activity increases with increasing the length of the spacer ( $n$  ranged between 2 and 8). The relation of the activity of the inhibitors with the length of the spacer was explained based on the shape of the active site in TG2. The authors suggested that a reasonable distance should separate the warhead from the hydrophobic Cbz group because the warhead is supposed to react with CYS277 and the hydrophobic group should accommodate itself in a hydrophobic region of TG2 active site; and CYS277 and the hydrophobic region of the active site are distant from each other, hence the requirement for the separation of the warhead from the Cbz group (Pardin et al. 2006).

The activity of the maleimide compounds was much lower than that of the other warheads, and that was attributed to the size of the warhead and the ease of access to the active site. All the inhibitors showed time-dependent inactivation of TG2. Irreversibility was further confirmed by removing the excess of the inhibitors from the reaction medium and noticing that TG2 activity was not restored. The authors also examined the specificity of the inhibitors

to TG2 by testing them against glutathione as a physiological thiol-containing nucleophile. The results showed that glutathione consumed only 10% of the inhibitor after 24 hours, compared to the fast reaction with TG2 (Pardin et al. 2006).

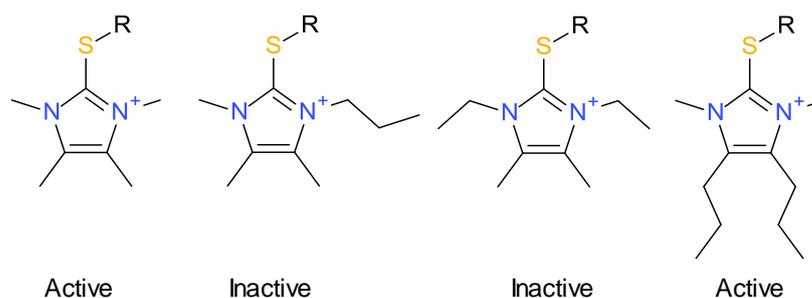
Dimethyl-sulfonium ketone is another warhead that has been used in a class of peptide-based inhibitors. Inhibitors containing the sulfonium warhead were derived from 6-diazo-5-oxo-L-norleucine (DON), using Cbz, Fmoc (fluorenylmethyloxycarbonyl) or *t*-BOC as the protective group with various amino acids by Griffin et al. (2008) (i8 Figure 1-11). An extra carboxyl group was added to the inhibitors to enhance water solubility and activity against TG2.



**Figure 1-11: Peptide-based inhibitors derived from DON by Griffin et al. (2008) (i8).**

The compounds showed good irreversible inhibitory effect with a TG2 IC<sub>50</sub> of as low as 5 μM. The Fmoc and *t*-Boc protective groups increased the IC<sub>50</sub> to more than 30 μM, again to signify the importance of the Cbz group for the activity against TG2. The enhanced water solubility would be useful in reducing toxicity as the compounds would be less likely to cross cell membranes. The study, however, did not include any attempt to examine the selectivity of the inhibitors for TG2. In addition to the sulfonium ion warhead, an imidazolium based warhead was also tried using the same peptide scaffold and was found to be more active than the sulfonium group with an IC<sub>50</sub> of 3 μM (Griffin et al. 2008). Consequently, the research group continued working by synthesising a new series with modifications of the imidazolium warhead (Badarau, Mongeot, et al. 2013).

The new study came up with some active TG2 inhibitors with IC<sub>50</sub> values of as low as 1 μM. Their main result was that the activity was lost by increasing the size of substitution at the imidazolium ring nitrogen atoms (Figure 1-12). The same observation was reported regarding the sulfonium warhead in the previous paper (Griffin et al. 2008). Another finding from this group was that the sulfonium warhead offered more specificity for TG2 over FXIIIA compared to imidazolium warhead (Badarau, Mongeot, et al. 2013).



**Figure 1-12: Examples of modifications to the imidazolium warhead by Badarau, Mongeot, et al. (2013) and their effect as TG2 inhibitors.**

The relatively large size of the imidazolium group can be compared to that of maleimide. Some of the imidazolium compounds were active and the authors have attributed this to the higher chemical reactivity of the imidazolium group, when compared to other chemical groups such as maleimide.

#### 1.4.2.2 Inhibitors discovered by screening of compound libraries

Several irreversible inhibitors were reported by Prime, Andersen, et al. (2012) who screened a library of 283,000 compounds for activity against TG2 and 4 chemotypic hits emerged. Three of these hits were excluded from further investigation because of high molecular weight or lack of selectivity. The 4<sup>th</sup> hit (4-bromophenylacrylamide, **i9**, Figure 1-13) showed good activity against TG2 with an  $IC_{50}$  of 3.3  $\mu$ M and promising selectivity characteristic when tested against TG1, TG3, FXIII A and caspase 3. Therefore this hit was used as a starting point for the synthesis of different groups of related compounds that were analysed *in vitro* and by a docking study (Prime, Andersen, et al. 2012).

The first modification involved the simple additions on the phenyl group. Only one compound of this group (sulphonamide modification, **i10**, Figure 1-13) was found to be active and showed a TG2  $IC_{50}$  of 1.9  $\mu$ M. A docking study on **i10** suggested that forming hydrogen bonds with residues TRP241 and GLN276 is important for a good pose. There was a possibility for another hydrogen bond between the sulphonamide group of the active compound and ASN333. Based on this assumption, the researchers prepared another series of compounds by extending the sulphonamide part to improve the chance of forming such hydrogen bond.

All the 16 compounds in this new series showed very good activity against TG2 with little or no effect on TG1, TG3 and FXIII A (**i11** is an example with a Cbz group and an  $IC_{50}$  of 0.12  $\mu$ M, Figure 1-13). The next step involved modifications in the acrylamide moiety of the latter active series. The alterations in the warhead lead to significant reductions in the activity with complete loss of selectivity; epoxides, esters, ketones and nitriles as substituents of the

acrylamide warhead all resulted in inactive inhibitors of TG2. The same applied to substitutions at the  $\alpha$ - and  $\beta$ -carbons of the acrylamide double bond with the exception of substituted chlorine. The same results (few active and none selective compounds) were reported for another series in which the modifications involved the phenyl group linking the acrylamide warhead to the sulphonamide part.

One final series of compounds was prepared by modifications to the Cbz protecting group, such as adding halo and alkyl substituents to the terminal phenyl group, or entirely replacing Cbz with urethanes and amides. Many of those modifications enhanced the activity and selectivity profiles for the inhibitors, and one compound in this series (**i12**, Figure 1-13), in which Cbz was replaced by an admantyl amide, showed the best activity and selectivity in this study with a TG2  $IC_{50}$  of 0.01  $\mu$ M and an 18-fold higher activity against TG2 than FXIIIa, and even higher selectivity against TG1 and TG3 (Prime, Andersen, et al. 2012). The results from the final series imply that the lipophilic parts of irreversible TG2 inhibitors represent a good area for development and modifications, in contrast to the warhead groups which can only be restricted to few classes of chemical groups.

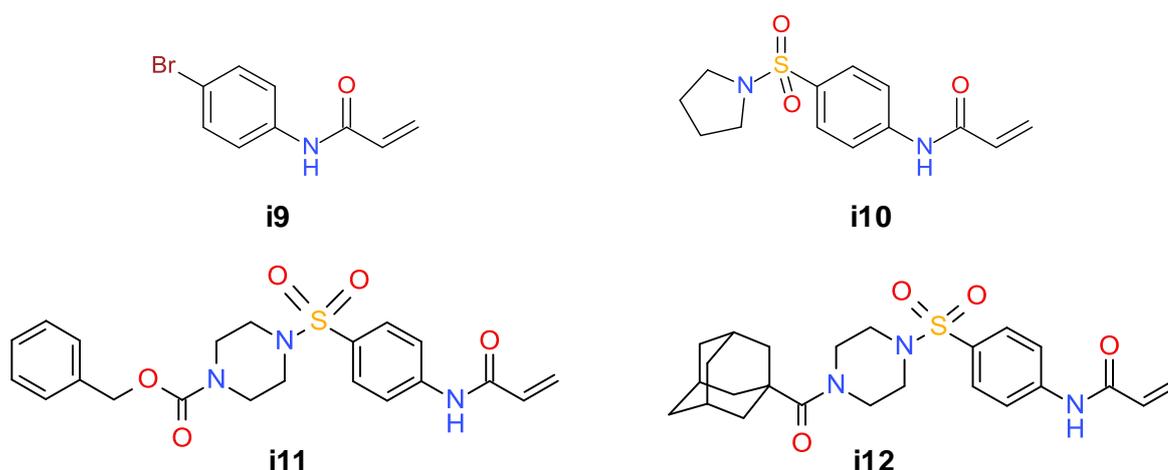


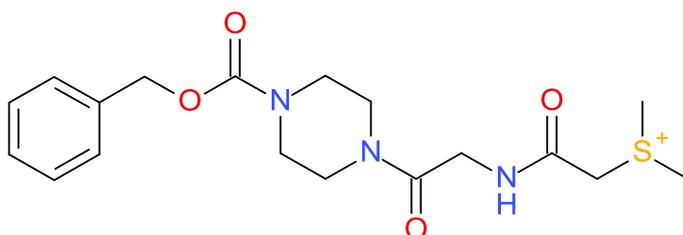
Figure 1-13: Four compounds representing different series of TG2 inhibitors from the work by Prime, Andersen, et al. (2012).

### 1.4.2.3 TG2 inhibitors used in this project

The structures of a total of 21 active TG2 inhibitors and 8 proven inactive compounds have been used during the course of this project in the various stages. These compounds were in-house compounds; designed, synthesised and tested for their activity and selectivity against TG2 in the laboratories of the School of Life and Health Sciences in Aston University, Birmingham with the participation of researchers from chemical and biological backgrounds. All the compounds have been patented (Griffin et al. 2014) and the most important of them were published (Badarau et al. 2015). The basis for the design of the new set of compounds

was molecular dynamic simulations applied on docking complexes for a sulfonium ion and an imidazolium inhibitors from previous work (compound **i7**, Figure 1-10) (Griffin et al. 2008; Badarau, Mongeot, et al. 2013). The simulations showed that the dihedral angle between the nitrogen from Cbz group, and the alpha carbon, the carboxylic carbon and the amine nitrogen of the central amino acid stabilised around 0°.

The value of the dihedral angle inferred stability around that region, therefore, that part was replaced by a more rigid piperazine ring. The resultant inhibitor (compound **i13**, Figure 1-14), which had glycine as the amino acid and a dimethyl sulfonium as the warhead, had a TG2 IC<sub>50</sub> value of 1.4 μM and showed an irreversible inhibition mechanism with minimal cellular toxicity. It was used as a prototype, from which several other compounds were derived. Changing the warhead to diethyl sulfonium retained the activity, while replacing it with an imidazolium group caused loss of activity. These same changes to diethyl sulfonium and imidazolium warheads produced the opposite effects on activity against TG2 in the previous set of compounds (Griffin et al. 2008), implying that the 2 sets, though related, cannot be compared to each other as they exhibited different activities.



**Figure 1-14: Compound i13, which was the basis for the compounds designed by Badarau et al. (2015).**

Lipophilic warheads were also tried; vinyl sulphonamide and acrylamide, with the aim of producing TG2 inhibitors with enhanced cellular permeability to specifically target intracellular TG2. Both warheads were active with the sulphonamide being slightly better (IC<sub>50</sub> for vinyl sulphonamide was 1.725 μM and for the acrylamide 5.925 μM). Next, the emphasis was turned on modifying the carbamate moiety in compound **i13** and combining this with various warheads. The rationale this time was to enhance the binding within the hydrophobic region of TG2 active site. Different sized aromatic and other hydrophobic groups were tried, with the best results obtained when the adamantyl hydrophobic part was combined with an acrylamide warhead (compound **1a**, Table 3-1, with a TG2 IC<sub>50</sub> of 0.125 μM), and when a *tert*-butyl part was used with a dimethyl sulfonium warhead (compound **1b**, Table 3-1, with and IC<sub>50</sub> of 0.273 μM). A naphthalene group for the lipophilic part with an acrylamide warhead also gave good activity against TG2 (compound **1c**, Table 3-1).

Amino acids other than glycine have also been tried. Changing to a bulky group such as phenylalanine resulted in complete loss of activity, while changing to alanine with the

dimethyl sulfonium warhead improved TG2 inhibition over that obtained with compound **i13** (compound **1d**, Table 3-1, with an  $IC_{50}$  of 0.7  $\mu$ M).

The final alteration involved applying a fluorescent tag on the lipophilic part of the inhibitors and the goal was to track the inhibition within the cells and to test that the inhibitor is physically bound to TG2. Two compounds having a dansyl group in their lipophilic end were synthesised and tested; the first had an acrylamide warhead (compound **1e**, Table 3-1) and produced the best inhibitory effect against TG2 with an  $IC_{50}$  of 0.0061  $\mu$ M. The second compound was designed with a dimethyl sulfonium warhead and showed descent TG2 inhibitory activity with an  $IC_{50}$  of 0.38  $\mu$ M (compound **1f**, Table 3-1).

An additional test was applied to confirm that these inhibitors act on the active “open” form of TG2. Two versions of TG2 were incubated for 30 minutes, once with TG2 activator  $Ca^{2+}$  and one with the TG2 inactivator GTP. Each TG2 version was then incubated with the inhibitors **1e** and **1f** for 30 minutes and the excess inhibitor removed. TG2 activity was measured and it was found that TG2 incubated with  $Ca^{2+}$  lost its activity upon the addition of the inhibitors while TG2 incubated with GTP retained its activity. This indicates that for these inhibitors to work,  $Ca^{2+}$  must be available to induce the open conformation of TG2 to give them access to CYS277.

To examine the selectivity of the inhibitors to TG2, the activity of compounds **1e** and **1f** against caspase 3 and 7 was measured and the compounds were found to be inactive against these enzymes. In addition, all the active compounds showed very good selectivity towards TG2 when tested against TG1, TG3 and FXIIIA and this was especially true for the dansyl compounds. Next, the toxicities of the active inhibitors were determined using XTT-base assay. The toxic concentrations for the compounds ranged between 25-100  $\mu$ M, with some halogenated inhibitors having the highest toxicities, whereas the water soluble compounds with sulfonium warheads were the best tolerated. The toxic concentrations were much higher than TG2  $IC_{50}$  values. It should be noted that the toxicity from the halogenated compounds was not unexpected, and they were never intended to be used as potential drugs; rather, they served as intermediates during the syntheses and were tested for their activity as TG2 inhibitors.

Compound **1e** with the dansyl fluorescent group was used to confirm the covalent inhibitory mechanism by incubating TG2 with the compound and  $Ca^{2+}$  for 3 hours. The same was applied for a mutant version of TG2 in which a serine residue replaced the active site cysteine residue. After incubation, the enzyme was separated from the remaining inhibitor and fluorescent bands were checked under the microscope. Wild type TG2 incubated with **1e** showed bands for the dansyl group while the mutant TG2 did not show any bands,

indicating the covalent binding of the inhibitor to the active site cysteine. When cellular permeability was tested on a subset of inhibitors, it was observed that the acrylamide-based inhibitors could penetrate cells and inhibit intracellular TG2 better than the sulfonium ion compounds. This is consistent with the water solubility profiles for both types of warheads (Badarau et al. 2015).

The version of TG2 that was used in the majority of studies for testing the inhibitors was the guinea pig liver TG2 (Sohn et al. 2003; Pardin et al. 2006; Pardin, Pelletier, et al. 2008; Griffin et al. 2008). Although this can be justified by the fact that this version of TG2 has 80% resemblance to the human version (Pardin, Pelletier, et al. 2008), testing the inhibitors against human TG2 will certainly produce more relevant results. The human version of the enzyme has been used in the studies by Choi et al. (2005), Watts et al. (2006), Prime, Andersen, et al. (2012), Badarau, Mongeot, et al. (2013) and Badarau et al. (2015). Griffin et al. (2008) used guinea pig TG2, but in addition they tested two of their compounds against human TG2 and found comparable results.

## 1.5 Computational Techniques

Computational chemistry is a general term which describes chemical events that can be represented by mathematical terms developed enough to be automated to make use of the rapidly increasing powers of computers (Young 2001). The term has been used interchangeably with molecular modelling and theoretical chemistry (Cramer 2013). Two main computational or modelling techniques have been used throughout this thesis; molecular dynamics (MD) and protein-ligand docking. Within each, several sub-techniques have been applied; with MD, for example, conventional MD was the major techniques used, in addition to enhanced sampling techniques with quantum mechanical treatment and covalent MD. With docking, flexible and rigid receptor approaches have been applied using more than one scoring function as well as rescoring a docked conformation and covalent docking. Both of these methods rely heavily on the principles of energy minimisation or geometry optimisation. In the following sections, brief descriptions for the various computational approaches will be presented.

### 1.5.1 Energy minimisation

The potential energy of any chemical system is a function of its atomic coordinates giving rise to the potential energy surface (PES), which is a multidimensional shape showing how the energy changes with the coordinates of the atoms. The surface is composed of hills, corresponding to high energy conformations, and valleys for the lower energy structures.

When building a computer model of a chemical system, the structure produced may not reflect the most stable conformation for that molecule. This results from the fact that the drawn structure will most probably be located outside the deepest valley in the PES (global minimum), which represents the most stable conformation for that system. The structure will have a higher chance of being within one of the local minima. The process of optimising the geometry of the chemical system to minimise its energy to the lowest possible value is called energy minimisation (Leach 2001).

Several minimisation methods are available. Examples include the Newton-Raphson method, which is the most computationally expensive because it calculates the first (slope) and second (curvature) derivatives of energy at each geometry point. Another example is the steepest descent method which calculates the first derivative only and assumes that the second derivative is a constant. Steepest descent is the fastest minimisation method and moves the system in the direction of the largest slope (steepest gradient). Conjugate gradient method is another minimisation method that is similar to the steepest descent, but the new search direction contains some information from the previous search direction (Höltje & Folkers 2008). All these are mathematical methods that optimise the geometry to minimise the energy. Energy is calculated at each step of the optimisation using one of two major methods. These are discussed in the next sections.

### **1.5.1.1 Molecular mechanics (MM)**

Molecular mechanics or force field methods make use of classical mechanics to handle chemical systems and treat molecules as being composed of atoms connected by bonds. Therefore, to model a new molecule, one can use the bond connecting parameters from known molecules (Hinchliffe 2003). Databases for those parameters can be obtained from experiments or from higher level calculations (Boeyens & Comba 2001). The force field refers to the energy terms investigated within a molecule, together with the parameters for those terms. The energy terms are generally divided into bonded terms (bond length, bond angle and dihedral angle) and non-bonded terms (van der Waals and electrostatic interactions). The total energy of the system is taken as the sum of these energy terms (Leach 2001).

#### **1.5.1.1.1 Bonded terms**

The energy associated with bond length is calculated based on Hook's law, in which the energy is related to the difference between the actual bond length and some value that represents the ideal or equilibrium length for that bond (Leach 2001). Bond angle term is also handled according to Hook's law as in bond length. However, its contribution to the total

energy of the system is less than that from bond length (Leach 2001). The angle between the 2 planes defining 4 successively bound atoms (the dihedral angle) is another bonded term in the force field. The energy of dihedral angle is calculated depending on the number of possible rotations around the angle. Its contribution to total energy is the least, but this should not make the angle less important than the previous term, because it means that it can be distorted very easily resulting in non-realistic conformations (Vinter & Gardner 1994).

#### **1.5.1.1.2 Non-bonded terms**

The terms describe the interactions between each atom pair that are not connected by a covalent bond and are of two types. Attraction or repulsion between non-bonded atoms, that are not caused by charges are represented by van der Waals forces. The contribution from this interaction is calculated by the Lennard-Jones 6-12 potential, where the energy will have a minimum at certain distance separating the atom pair. Electrostatic interactions arise from attraction and repulsion resulting from the charges on the various atoms in the system. The energy of the interaction between 2 atoms depends on the distance between them and the partial charges they are carrying and is calculated by Coulomb law (Vinter & Gardner 1994).

Non-bonded interactions tend to fade at higher distances, but in theory they extend infinitely. To save computer time, cut-off distances are introduced, which are distance limits beyond which the non-bonded energy will not be calculated for atom pairs. Force field methods perform the minimisation process by iteratively changing the parameters and calculating the energy at each step, until reaching the lowest possible energy for the system or the maximum number of iterations (Vinter & Gardner 1994).

Molecular mechanics methods are computationally very efficient, making them suitable to handle large systems including biological macromolecules. The force fields are transferrable, so they can be applied to a variety of systems (Leach 2001). On the other hand, their accuracy is not the best among the energy calculation techniques, and the lack of representation of electrons in the force fields means that they cannot be used to simulate chemical reactions (Vinter & Gardner 1994).

#### **1.5.1.2 Quantum mechanics (QM)**

These methods use the principles of quantum chemistry to describe a chemical system and calculate its energy. They handle molecules as being composed nuclei and electrons moving around them (Vinter & Gardner 1994). Quantum mechanical methods estimate the energy of a system by finding the appropriate wave function to describe the system. The wave function is a probability function for the movement of the electrons around the nuclei (orbitals). An

exact wave function can only be found for systems containing one electron only by solving the Schrödinger equation, and this is only applicable to hydrogen atom (Leach 2001). For poly-atomic systems, the energy of the electrons has more than one component; a core component (interaction between electron and nucleus), an exchange component (electrons on different orbitals) and a Coulomb component (electrons on the same orbital). Therefore, the Schrödinger equation cannot be solved exactly and the result is multiple wave functions to describe the system. Based on the variation principle, the best wave function is the one that gives the lowest energy to the system (Leach 2001; Vinter & Gardner 1994). Depending on the level of approximation, QM methods are of 2 major types.

#### 1.5.1.2.1 *Ab initio* methods

*Ab initio* is the Latin term for “from scratch”, which means that these methods do not use approximations or parameters from experiments, and perform optimisation by handling every single electron in the system (Dorsett & White 2000). Some approximations are, however, inevitable to make the calculations possible; one example is the use of the linear combination of atomic orbitals (LCAO) to represent the molecular orbital of the system in which the wave functions corresponding to the individual atomic orbitals are added together to give a wave function for the molecule, and the wave function with the lowest energy is the minimum for that molecule (Vinter & Gardner 1994).

*Ab initio* methods are defined by a choice of a method and a basis set. The method is a mathematical representation for the calculation of the wave function of the molecular orbital from the individual atomic orbitals. Hartree-Fock (HF) is an example of *ab initio* methods that deals with single-electron wave functions (Dorsett & White 2000) and density functional theory (DFT) which performs the calculations based on electron density, rather than single electrons (Koskinen & Mäkinen 2009). The basis set is an analytical term used to represent the atomic orbitals (Vinter & Gardner 1994).

*Ab initio* methods are highly accurate when compared to other means of geometry optimisation, and because they do not rely on parameters, they can be used for newly designed systems for which no experimental data exist. They can be used to derive parameters for other methods, such as force fields and semi-empirical methods (Dorsett & White 2000; Leach 2001). Their major disadvantage is that they are very computationally expensive, especially with the use of more complicated methods and basis sets, making them applicable only to small systems (Vinter & Gardner 1994).

### 1.5.1.2.2 Semi-empirical methods

They are methods that also deal with electrons and nuclei but they involve approximations that allow them to be much faster than *ab initio* methods (Hinchliffe 2003). The approximations include neglecting the integrals of overlaps between atomic orbitals (Leach 2001), substituting some integrals by parameters from experiments and performing the calculations on valence shell electrons only (Höltje & Folkers 2008). Depending on the level of neglect involved in the method, there are several types of semi-empirical methods, such as CNDO (complete neglect of differential overlap), MNDO (modified neglect of diatomic overlap) and PM3 (parameterisation method 3) (Leach 2001). Semi-empirical methods are faster than *ab initio* methods and their results are more reproducible, being based in part on parameters. They are less accurate, however, especially when applied on systems larger than the systems from which the parameters were derived (Vinter & Gardner 1994; Seabra et al. 2009).

### 1.5.2 Molecular dynamics

X-ray crystallography is probably the best technique that can be used to elucidate the 3-dimensional (3D) structure of biological macromolecules. The main drawback of this technique, among others, is that the structures produced are static and do not represent the entire conformational space the macromolecule can adopt. NMR spectroscopy is another technique that can give more insight on the flexibility of a biological system, through uncovering the structure of the macromolecule as an ensemble of conformations. Nonetheless, the technique remains extremely difficult to apply and time consuming (Mortier et al. 2015).

The static structures from X-ray crystallography and the ensembles from NMR spectroscopy have succeeded in providing essential information about biological macromolecules and their spatial arrangements. However, the inflexibility of the structures renders them less useful in recounting events that involve dynamics of the biological system. Example events include the continuous motion of proteins in biological solutions, the conformational changes in proteins brought about by binding of ligands, whether natural substrates or exogenous molecules, and the allosteric regulation of enzymes that signifies an important regulatory mechanism for many enzymes and usually involves a large conformational modification of the enzyme that alters its activity. Similar examples can be found within nucleic acids, where the binding of DNA to transcription factors is regulated by base sequence as well as the potential of the DNA molecule to accommodate the transcription factor (Hospital et al. 2015). Molecular dynamics (MD) is a computational method that could be used to predict the

possible conformational space a biological macromolecule can explore, as well as the paths taken by the biological system during this exploration. When applied to the static structures, MD can extend the level of flexibility of the system (Mortier et al. 2015).

Advances in computational powers witnessed in the last few years have enabled the application of MD on larger and larger systems. At the same time, these advances have allowed for extended simulation times; simulations on large proteins, receptors, membrane proteins, phospholipids, nucleic acids and other macromolecules for microseconds are now routinely conducted (Mortier et al. 2015).

### 1.5.2.1 Basic principle

MD simulations operate on the basis of solving Newton's equations of motion:

$$f_i = m_i a_i$$

where  $m_i$  is the mass of atom  $i$  and  $a_i$  is the acceleration of atom  $i$  and  $f_i$  is the force acting on the atom. The force is a function of the potential energy of the system and the latter is dependent on the positions of all atoms in the system. The result from MD is a time series of conformations that represent the dynamic spatial configuration of the biological system, and from which, thermodynamic properties can be calculated. The application of MD requires 4 basic fundamentals; the resolution of the system, description of the interactions within the system, production of the time series conformations and solvation (Mortier et al. 2015).

#### 1.5.2.1.1 System resolution

The resolution describes how the elementary particles of the system are given. The particles can be presented as nuclei and electrons, as atoms or as coarse-grained (CG) particles, each composed of several atoms. Nuclei and electrons description is utilised when there is QM treatment and the description gives the most accurate results, but is very slow that it would be impractical to apply to large systems. QM treatment is, therefore, reserved for the study of enzymatic reactions and reaction pathways where only a small fraction of the system is treated by QM while the remainder is handled conventionally. CG description is the other extreme where the number of particles in the system is reduced significantly allowing more space to be explored more rapidly, at the expense of the accuracy of the information obtained. Particles treated as atoms represents the best compromise between speed and accuracy and is the most popular approach in conventional MD (Mortier et al. 2015). Within the system resolution also lies the solvation of the system. This will be discussed in a separate section.

### 1.5.2.1.2 Force field

The potential energy function used within MD simulations describes the interactions within the system particles and is generally termed 'force field'. Force fields, as mentioned previously, commonly use two main terms to define the interactions between atoms; bonded interactions term and non-bonded interactions term (Mortier et al. 2015). Force fields, with their parameters, allow the calculation of the forces acting on each atom in the system, which in turn permit the prediction of the velocities and accelerations of these atoms and consequently the new atomic positions according to Newton's equations of motion. Due to the method of integration used in the calculation of the forces, a short time step must be used. The term 'short' means that the time step should be shorter than the fastest atomic motion and therefore, time steps are generally in the order of 1-2 femtoseconds. The use of larger time steps can significantly improve computation time and this is possible with CG representation of the system but on the expense of the accuracy of the information obtained (Hospital et al. 2015).

The parameters embedded in the force fields are derived from higher level computational methods or from experiments. Because biological macromolecules are composed of a finite set of building units (for example, amino acids for proteins and nucleotides for nucleic acids), the use of force fields would sound reasonable as those contain information on this finite set of building blocks. For organic, drug-like molecules or ligands, no building blocks can be used for the representations because of the diversity in the structures of these ligands. Separate force fields are used to describe these atoms which have the information for bonded and non-bonded terms taken from the general force fields while the partial atomic charges are derived from higher level QM calculations (Mortier et al. 2015).

Multiple reliable force fields exist today, and examples include the AMBER (assisted model building with energy refinement) force field (Hornak et al. 2006) used with the AMBER MD package (Case et al. 2012), CHARMM (Chemistry at HARvard Macromolecular Mechanics) (MacKerell et al. 1998) with the CHARMM program (Brooks et al. 2009) and GROMOS (Groningen Molecular Simulation) (Oostenbrink et al. 2004) with the GROMACS (GROningen MACHine for Chemical Simulations) program (Hess et al. 2008). Those force fields differ mainly in the methods used for their parameterisation and the sources of the parameters; the force fields cannot be used interchangeably, but the trajectories resulting from them are comparable (Hospital et al. 2015).

### 1.5.2.1.3 Production of the time series

After the forces acting on atoms are calculated and the new atomic positions are predicted, the new conformation is recorded and is used as the starting point for the next time step force calculation. The ultimate result of the MD simulation is a time series of the new conformations that would represent a trajectory for the dynamics of the biological macromolecule as predicted by Newton's equations of motion (Mortier et al. 2015). As stated earlier, the calculation of the new atomic positions requires integrating the equation of motion. An analytical integration is not possible due to the complexity of the equation. A numerical integration is used instead. Several numerical algorithms are available such as Verlet algorithm and the leap frog algorithm (Paquet & Viktor 2015).

### 1.5.2.1.4 Solvation and periodic boundary conditions

Biological macromolecules are found in solutions in biological system, and this should be considered when applying MD simulations. Two main solvation models exist; implicit solvation and explicit solvation. In the implicit solvation model, solvent, usually water, molecules are not added physically to the simulated system. Instead, they are replaced by a potential that is used to simulate the effect of the presence of the actual solvent molecules. This model is computationally fast because the number of the atoms of the system is not increased. In addition, the model allows for infinite representation of the solvent, which is the case in bulk liquid. On the other hand, there is the fundamental compromise in computation between speed and accuracy due to the absence of water molecules, and thereby the absence of the effects that could result from their presence such as hydrogen bonds and ligand solvation effects (Paquet & Viktor 2015; Onufriev 2008).

In the explicit solvation model, water molecules are added physically to solvate the biological system. Although water molecules in bulk solution are in the form of an infinite solvation box, only a finite number of water molecules can be added during MD due to the limited capacity of the computer memory. Solvation boxes are used for this purpose and these can take the shape of cubes, dodecahedrons or truncated octahedrons. The larger the solvation box, the more accurate the simulation would be, but this will have a great effect on the speed of the simulation. In fact, much of the computation time is spent on dealing with the solvent molecules as they usually constitute the majority of the components in the test system (Tuckerman & Martyna 2000; Paquet & Viktor 2015).

Because solvation boxes are finite, periodic boundary conditions are employed to make the representation more realistic. Periodic boundaries work by generating replicate translational images for the system in all the dimensions and the true system is in the centre of those

replicates. When a water molecule reaches the boundary of the box and eventually leaves it, its copy will replace it from the corresponding image. The representation is more realistic than the implicit solvent but explicit solvent with periodic boundary is computationally expensive due to the requirement to handle long distance electrostatic interactions between the atoms from various images. One approach around this is the use of the particle mesh Ewald method. This approach works by dividing the Coulombic potential into a short-range component (calculated by conventional means) and a long-range component whose contribution is calculated through a Fourier transform (the potential is calculated in the reciprocal space) (Tuckerman & Martyna 2000; Paquet & Viktor 2015).

### 1.5.2.2 Stages of MD

According to the authors of the CHARMM MD program (Stote et al. 1999), conventional MD is ideally performed in 5 stages. The first stage involves preparing the initial coordinates of the system to be simulated. These can be from a crystal structure from the Protein Data Bank, a homology model for a protein whose 3D structure has not been yet elucidated or a docking complex generated by one of the docking programs. The type of solvation should be chosen at this stage, whether implicit or explicit, as well as the force field that will be applied to define the atoms and the interactions. The second stage is performing energy minimisation on the system. The objective is to remove any bad clashes from the original crystal structure, and therefore, local optimisation would be sufficient where the goal is not to find a global minimum. Another objective for minimisation is to allow water molecules added during solvation to accommodate themselves to the biological system without any overlap.

The third step comprises heating the system. This is done because the system starts at a temperature of 0 K. During this phase, velocities are allocated to the atoms of the system at low temperature and MD is applied and the system is propagated over time by solving Newton's equations. Periodically, new velocities are allocated to the atoms while slightly increasing the temperature. Heating is continued in the same way (allocating velocities while raising the temperature) until the desired temperature is reached. Heating is followed by equilibrating the system (4<sup>th</sup> step). The purpose of this stage is to run the MD on the solvated system while monitoring properties such as pressure, temperature and energy. This phase should be continued until these properties become stable over time. When the properties are stable, the system is said to have equilibrated and can proceed to the final phase which is the production of the actual MD trajectory. Production phase can be extended for any desired period of time, and it is the trajectory produced during the production that could be inspected to gather information and perform analyses about the biological system.

The most popular MD packages such as AMBER, CHARMM and CROMACS have made a great use of the computational advances. For example, they all use the messaging passing interface (MPI), which allows the computation to be spread over multiple cores within the computer, and hence greatly speeding the process. Each core will handle a section of the system and will perform the calculations on that section. The communication between the cores is reduced, since it will be reserved for those handling adjacent sections of the system, which will also speed the process. Graphical processing units (GPUs) are components of computers that were originally intended to handle different aspects of graphics. A major enhancement of computer abilities to perform MD simulations was made when the GPUs were allowed to be used as processors to run MD. With GPUs, microsecond-long MD simulations are now conducted routinely (Hospital et al. 2015).

### **1.5.2.3 Enhanced sampling techniques**

Even with the recent advances in computational powers and the codes running MD, the complexity of the potential energy surface of biological macromolecules remains a major hurdle in the face of MD. The simulation will be trapped in an energy minimum near the starting conformation, without the possibility of crossing the high energy barriers separating it from the global minimum. Enhanced sampling techniques are methods developed to overcome the high energy barriers in the potential energy surface allowing more accessible exploration of conformational space. In addition, these techniques enable the sampling of events that accompany such high energy barriers, such as protein folding and bond breaking or formation (Hospital et al. 2015), keeping in mind that bond breaking and formation requires in addition QM treatment. Examples of those techniques are presented in the following sections, with a more detailed description of those used in this work.

#### **1.5.2.3.1 Replica exchange molecular dynamics**

This method of enhanced sampling involves simulating multiple copies or replicas of the biological system at the same time but at different temperatures. A swap is performed between the replicas every predefined number of steps, so that a replica starting at low temperature will be allowed to sample at a higher temperature and vice versa. A weighting analysis is performed to give the conformational space that was sampled during the method, which would be much wider than that sampled by conventional MD, since the simulations of replicas at higher temperature will overcome some of the high hills present in the potential energy surface of the molecule (Mitsutake et al. 2001). A major disadvantage of this approach, arising from the fact that multiple simulations are run at the same time, is its

computational cost (Michel & Essex 2010). The method has, however, proven superior to classical MD in predicting events such as protein folding (Bernardi et al. 2015).

### **1.5.2.3.2 Umbrella sampling**

The basic idea behind umbrella sampling is that a MD simulation is performed while adding a biasing potential to push the system to sample a predefined coordinate over a certain range of values. These values are defined by a restraint added to the coordinate. A record is kept for the manner in which the coordinate has changed over the simulation. The process is repeated by moving the minimum of the restraint governing the predefined coordinate to some higher or lower value. Each of the resultant trajectories is called a window. The individual windows must have some degree of overlap. This means that each recorded value of the coordinate must have been sampled in more than one window. The change in the coordinate between the windows must be small enough to ensure a quasi-static process. At the end, the biasing potential is removed and the potential of mean force (PMF) can be calculated, which represents the free energy change as a function of the coordinate (Lonsdale et al. 2012; Kästner 2011). Umbrella sampling can be used to calculate the free energy difference between two states (defined by the coordinate) through calculating PMF. That is why it was used in this work to simulate the reaction between some irreversible inhibitors and TG2. The elements of umbrella sampling are presented in the next sections.

#### **1.5.2.3.2.1 The coordinate**

Several system parameters can be used as the coordinate, or reaction coordinate (RC) for umbrella sampling simulations. The chosen RC should have the ability to represent the change between the initial and final stages of the system during the sampling simulations. It should also be possible to calculate the RC as a function of the atomic coordinates of the system being simulated. Accordingly, parameters such as distances, bond lengths, angles, dihedral angles, hydrogen bond lengths and root-mean-squared deviation (RMSD) from some reference structure, can all be used as reaction coordinates (Spiwok et al. 2015).

#### **1.5.2.3.2.2 The biasing potential**

During umbrella sampling, the RC is set to change from one value to another, where each of the 2 values characterises a definitive state of the system. As mentioned earlier, the change in RC between these 2 values is supposed to proceed over small steps, or windows and during each window a biasing potential is added to bring the desirable change. Usually, the biasing potential is applied in the form of a harmonic potential. The potential produced from applying the harmonic restraint is added to the energy of the system at that value of reaction

coordinate. This can then be removed to calculate PMF. The strength of the restraint is a criterion that requires attention; too low restraints will not be enough to push the system to cross the high energy barriers in the potential energy surface, and too high restraints will result in a very narrow distribution of RC with the subsequent lack of overlapping between the windows. Trying several values of the restraint, starting from the lowest possible and moving up until the change is observed with reasonable overlap indicates that the chosen restraint is suitable. This can be performed beforehand using smaller windows, but it will still be expensive in terms of computational time. The use of experimental values, if available, would produce more promising results (Kästner 2011; Mills & Andricioaei 2008).

#### 1.5.2.3.2.3 Potential of mean force

The result from umbrella sampling is usually expressed as a graph of the free energy change in the system along the sampled reaction coordinate. This is the definition of PMF (potential of mean force). The free energy or the PMF is obtained from the time series of the changes in the reaction coordinate in each window under the influence of the biasing potential. The weighted histogram analysis method (WHAM) (Souaille & Roux 2001) is used to remove the effect of the biasing potential and calculate PMF based on the probability of the distribution of the reaction coordinate over the windows. For the best PMF results, WHAM requires that umbrella sampling windows are overlapping (Kästner 2011; Baştuğ et al. 2008).

#### 1.5.2.3.3 Steered molecular dynamics

The concept of steered molecular dynamics (SMD) also involves the application of an external force or perturbation to drive a particular process in the desired direction, and thus would reduce the time needed to observe the final outcome. The work involved in the process could then be converted to binding free energy using the Jarzynski's equality (Jarzynski 1997b; Isralewitz et al. 2001; Park et al. 2003; Suan Li & Khanh Mai 2012). Jarzynski's equality (Jarzynski 1997a; Jarzynski 1997b) is a mathematical expression that is used to convert the work performed in non-equilibrium processes to a free energy. The basic equation is:

$$e^{\left(\frac{-\Delta G}{k_B T}\right)} = \left\langle e^{\left(\frac{-W}{k_B T}\right)} \right\rangle$$

where  $\Delta G$  is the free energy of the system,  $W$  is the work performed in the process,  $k_B$  is Boltzmann constant and  $T$  is the temperature of the system. The angled brackets indicate an average of the work over a number of the non-equilibrium processes.

The main drawback in Jarzynski's relationship lies in its very definition, where it applies only when the work is averaged over an infinite number of the non-equilibrium processes and when a finite number is used, correction must be applied to account for this bias (Dodson et al. 2012). Gore et al. (2003) have devised equations for the bias correction as well as the calculation of mean square errors associated with the bias correction. The equations involve calculating and correcting the dissipated work to ultimately calculate the bias-corrected Jarzynski's average.

#### **1.5.2.3.4 Accelerated molecular dynamics**

This is another enhanced sampling method that is aimed at improving the exploration of the potential energy surface of the studied biological systems by allowing the biological systems to break free from the local energy minima in which they are trapped. A major difference between accelerated MD (aMD) and umbrella sampling is that the latter requires a knowledge of the system to define the coordinate along which to apply the biasing potential. In aMD, a biasing potential is added such that it would be used every time the system is trapped in a local minimum, and if the system is next to a barrier or a hill on the energy surface then the biasing potential will not be used. In this way, prior knowledge of the system is not an absolute requirement. As in umbrella sampling, the effect of the biasing potential can be removed at the end of the simulation. aMD has enabled access to milliseconds of simulation information by performing nanoseconds of simulations (Pierce et al. 2012; Hamelberg et al. 2004).

The bias is added when the true potential of the system falls below a predefined value ( $E$ ), which is an indication that the system is in a local minimum. The bias addition is done in such a way to keep the shape of the minimum; in other words, the bias will help in raising the minimum and smoothing the associated well rather than making it flat. The predefined value of the true potential ( $E$ ), below which the bias will start acting, must be larger than the minimum potential of the system in its starting conformation; otherwise, the result would be conventional MD. The recommendation is to apply conventional MD for a short period and to calculate the average potential energy of the system at the end of this period. The average potential energy can then be used to define  $E$  (Pierce et al. 2012; Hamelberg et al. 2004). It is possible to obtain the original potential energy surface of the simulated system by reweighting the aMD simulation using exponential averaging and Boltzmann factor (Miao et al. 2014).

### 1.5.2.4 Quantum mechanical treatment

As mentioned in the sections covering system resolution and force fields, conventional force fields mostly used in MD handle the systems as being composed of atoms connected by strings and have special terms to account for non-bonded interactions. In some situations, the system involves a chemical reaction between 2 or more molecules and there is bond formation and/or breakage during the reaction. Classical force fields cannot simulate such types of system, regardless of the simulation time allowed.

QM methods deal with electronic structures; they account for electrons when handling the systems and hence can simulate chemical reactions. The main problem with these methods is their computational cost; because they deal with electrons, their simulations can take very long times. One way around the problem is to apply a combined quantum mechanics/molecular mechanics (QM/MM) method. In a QM/MM treatment, the reaction centre (for example and inhibitor and an active site residue) is treated by a QM method, while the rest of the system is treated by a classical force field using a MM method. The relatively small QM region will allow the simulation to be performed within an accessible time scale. The MM treatment enables the inclusion of the effect of the environment (remainder of protein structure, for example) on the reaction as well as any conformational changes associated with or resulting from the reaction (Ranaghan & Mulholland 2010).

Two approaches for calculating the total energy of the system in QM/MM have been proposed. The first is the additive approach, in which the QM region is only treated by a QM method and the total energy of the system would be simply the sum of that of the QM region, the MM region and the energy of the interaction between the 2 regions (van der Kamp & Mulholland 2013). The 2<sup>nd</sup> approach is a subtractive method, in which the energy of the total system (including the QM region) is calculated by MM, and the energy of the QM region is calculated additionally by the QM method. The energy of the system is then taken as the sum of the total energy calculated by MM for the whole system and the energy of the QM region calculated by QM, and from this sum the energy of the QM region calculated by MM method is taken away (Vreven et al. 2006).

If the QM region is covalently bound to the rest of the system (MM region), then additional atoms are added to the QM region at the positions of the covalent bonds linking the 2 regions. They are called link atoms and are usually hydrogen atom. They are treated as being a part of the QM region and their main purpose is to satisfy the valence of the atoms at the interface between the 2 regions (Walker et al. 2008). Many QM methods are available to be used for QM/MM simulations. Semi-empirical methods such as AM1, PM3, PM6 and

SCC-DFTB are reasonable choices, considering that higher level methods such as *ab initio* methods are much more computationally expensive. It should be noted that it is the QM method and the treatment of the QM region that will determine the speed at which the simulation is performed (van der Kamp & Mulholland 2013).

Two semi-empirical QM methods were used in QM/MM simulations in this work; PM3 and SCC-DFTB. Semi-empirical methods deal with electrons and molecular orbitals in their attempts to optimise or minimise molecular systems. They differ from the higher level *ab initio* methods in that certain aspects of the calculations have been simplified by parametrisation. PM3 (parameterisation method 3) belongs to a class of the semi-empirical methods called Modified Neglect of Diatomic Differential Overlap (MNDO) that is based on parametrisation of all two-electron integrals to avoid involving them in the calculations and therefore to save computational time. PM3 has been parametrised to reproduce large number of molecular properties. The method includes terms to handle hydrogen bonds, and many important elements were incorporated in the parameter set (Stewart 1989; Stewart 2007).

SCC-DFTB (self-consistent charge density functional tight-binding) is an approximation for the density functional theory (DFT) method, which is one of the *ab initio* methods (Elstner 2006). The Hartree-Fock method handles single electrons, and when an electron is being treated, the other electrons in the system are fixed. DFT moves around this approximation by dealing with the electron density. With the latter, it is possible to represent the density of all the electrons by 3 variables, rather than 3 variables for each electron in the system. The energy is then taken as a functional (function of a function) of the electron density, and by minimising the density, the geometry is optimised (Orio et al. 2009).

Tight binding is a semi-empirical approximation to DFT, which assumes that the electrons are tightly bound to the atoms and have limited interactions with the surrounding environment (Goringe et al. 1997). This allows introducing an approximation to the electron density by splitting it into a tightly bound reference region and an external region on which the calculation is performed (Foulkes & Haydock 1989). SCC-DFTB is a variation of DFTB in which the charge fluctuation is taken into account by decomposing the electron density into atomic-like contributions which fade as the distance from the atoms is increased (Elfturi 2014).

### 1.5.2.5 Principal component analysis

Principal component analysis (PCA) is a statistical method used to reduce the complexity of large data sets, such as MD trajectories. PCA extracts from the trajectories principal

components (PCs), each of which corresponds to a distinct molecular motion mode within the trajectory. Each PC is an eigenvector that has an eigenvalue, where the latter can numerically describe the contribution of that particular PC to the total motion within the MD trajectory. By comparing the PCs and their eigenvalues, it is possible to discriminate between important modes of motion from those that are not very influential. This will greatly facilitate the analysis of large MD trajectories, by giving the opportunity to focus the analysis on the major modes of motion observed during a given MD simulation (Ng et al. 2013; Skjaerven et al. 2011; David & Jacobs 2014).

### **1.5.2.6 AMBER overview**

AMBER 12 (Case et al. 2012) was the version of AMBER that was used to run various MD simulations in this work. AMBER refers to a collection of force fields designed to handle various biological molecules including proteins, nucleic acids and lipids in addition to small organic molecules. The term AMBER also applies to a package of programs and codes designed to perform various functions related to MD. The most important of these programs are presented below (Salomon-Ferrer, Case, et al. 2013).

#### **1.5.2.6.1 sander**

Sander (simulated annealing with NMR-derived energy restraints) is the main program within AMBER involved in running MD. It is also responsible for running replica exchange and QM/MM simulations.

#### **1.5.2.6.2 pmemd**

This program is similar to sander but is applicable to a narrower range of MD simulations. It is faster than sander and performs better on parallel cores. It also includes the GPU acceleration recently introduced to AMBER (Salomon-Ferrer, Götz, et al. 2013).

#### **1.5.2.6.3 LEaP**

This is the program used to generate the parameter-topology and coordinate files that can be read by sander and pmemd to run MD. It can read PDB files and it can add hydrogen atoms, water molecules and neutralising ions.

#### **1.5.2.6.4 antechamber**

This program is similar to LEaP but is used to generate the input files for small organic molecules that are not included in the standard AMBER force fields. Its output files are then

loaded into LEaP to be used with the files for the biological molecule (Wang et al. 2006; Wang et al. 2004).

#### **1.5.2.6.5 CPPTRAJ**

This program is used for the analysis of the trajectories generated from various MD runs. It can calculate a wide range of information including distances between specific atoms, hydrogen bonds, root-mean-squared deviations from a reference structure, atomic fluctuations of the individual residues during a simulation, principal component analysis and many more (Roe & Cheatham 2013).

#### **1.5.2.6.6 MMPBSA.py**

This is another program that is used for post-processing purposes where it can be applied following a simulation of a complex between two molecules (ligand and protein for example) to calculate an estimate of the binding energy between the 2 molecules (Srinivasan et al. 1998; Miller et al. 2012).

AMBER programs and force fields are very popular MD tools. AMBER as a software has been cited more than 4,000 times until 2013 whereas AMBER force fields have been cited more than 9,000 times to that year (Salomon-Ferrer, Case, et al. 2013).

### **1.5.3 Molecular docking**

Molecular docking, or docking for short, is another computational method used to predict the interaction between two molecules, usually a ligand and a receptor. This prediction involves posing the ligand within a binding site in the receptor. An accurate guess of the binding mode is an important goal for docking, and together with correct estimation of the binding affinity that reflects biological activity of the ligand, represents the major sites of development of docking softwares. For posing the ligand in the binding site, the software uses a docking method to find the most plausible mode for ligand binding. This is not a straightforward task due to the large number of conformations a ligand molecule can adopt. For the prediction of the binding affinity, a scoring function is applied on the poses generated by the docking method to score the poses based on the interactions between the ligand and the receptor (Kitchen et al. 2004).

#### **1.5.3.1 The docking method**

The docking method or the search algorithm attempts to reproduce the experimental binding mode for a ligand in the binding site of a receptor, or to predict the binding mode if no

experimental one is identified. The algorithm accomplishes that by exploring the degrees of freedom of the ligand, and sometimes, of the receptor active site. Ideally, all the degrees of freedom of the involved residues should be tested and the resultant binding modes examined, but this approach is not applicable as it would be computationally very expensive. Therefore, approximations are applied to make the process more computationally accessible; restraints and constraints are applied to the ligand to limit its degrees of freedom, and the binding site of the receptor is treated as rigid body or partially flexible (Meng et al. 2011; Taylor et al. 2002). How the docking methods or the searching algorithms differ in the handling of the flexibility of the ligand and the receptor represents the major determinant of the method category.

### **1.5.3.1.1 Docking methods handling ligand flexibility**

These docking methods were designed mainly to find a way of posing the ligand within the binding site of the receptor. They are more concerned about handling the flexibility of the ligand than that of the receptor.

#### **1.5.3.1.1.1 Shape-matching search algorithms**

The basis of these docking algorithms is that they generate binding modes by fitting the ligand within the binding site of the receptor, such that the molecular surface of the ligand complements the molecular surface of the receptor binding site. They are the simplest of the docking methods and the most efficient in terms of computational time required to perform a docking task. Their accuracy, however, is limited by the rigid treatment of the ligand. The shape-matching methods are usually applied during virtual screening of libraries of thousands of compounds to identify hits, that could then be tested using more rigorous search methods (Meng et al. 2011; Huang & Zou 2010).

#### **1.5.3.1.1.2 Systematic search algorithms**

Two type of search algorithms fall under this category, the exhaustive search and fragmentation methods. In exhaustive search algorithms, ideally, all of the ligand's degrees of freedom are explored and since this is not computationally practical, some constraints are added to the ligand to reduce the number of rotatable bonds and hasten the searching process (Huang & Zou 2010; Friesner et al. 2004). The fragmentation method is also called incremental construction and involves breaking the ligand into small pieces at its rotatable bonds, and then rebuilding the ligand within the binding site piece by piece. Usually, a nucleus from the ligand is selected to be docked first (anchor). This, typically, is the largest core in the ligand and/or is responsible for important interactions with the binding site. The

remainder of the ligand is added incrementally to the anchor. The final outcome is the generation of several possible orientations of the ligand with better handling of its flexibility (Meng et al. 2011; Cross 2005).

#### **1.5.3.1.1.3 Stochastic or random searching algorithms**

These methods produce random changes to the ligand within the binding site to generate a binding mode. The changes are accepted or rejected after being judged by a predefined probability function. Different random algorithms have been proposed based on the criteria for accepting or rejecting the random changes (Sousa et al. 2006). Monte Carlo is an example of random searching method that utilises a Boltzmann-based probability function to judge the quality of the random alterations on the ligand. The function takes the decision based on the energy scores of the ligand before and after the alteration. Monte Carlo methods are computationally efficient when compared to MD methods, as the energy function is simpler and due to the ability of Monte Carlo methods to cross high energy barriers (Sousa et al. 2006; Meng et al. 2011). Tabu searching is another random algorithm, which performs docking by randomly changing the pose of the ligand within the binding site. The decision to accept or reject a new pose is based on comparing the RMSD of the new pose to all the previously generated poses. The algorithm also prevents the generation of previously explored poses by keeping a record of them (Kitchen et al. 2004; Sousa et al. 2006).

Another type of the random searching methods is the genetic algorithm. As the name suggests, these algorithms use the principles of genetics and evolution to perform docking. The ligand's degrees of freedom; its rotation, orientation and translation, are encoded into chromosomes as genes. Therefore, the chromosome would be composed of genes that define the states of the ligand (genotype) while the atomic coordinates resulting from the genotype represent the actual pose of the ligand within the binding site (phenotype). The chromosomes are altered randomly by applying genetic operations, namely mutations and crossovers to produce new chromosomes, and thereby, new poses for the ligand. Mutations involve making random changes to the genes constituting the chromosomes, and crossover operations exchange genes between chromosomes. Any new pose generated by the genetic operations is scored by a fitness or a scoring function; if the score is better than the that of the previous genotype, the resultant pose is retained and subjected to further alterations, and if the score was worse, the pose is rejected and the previous pose is altered differently. The process is repeated until a predefined number of operations is reached (Meng et al. 2011; Sousa et al. 2006; Jones et al. 1995).

### **1.5.3.1.2 Docking methods handling receptor flexibility**

Receptors, usually proteins, are dynamic molecules that are in continuous motion within the biological system. This is complicated by the conformational changes induced in the receptor by the binding of the ligand (ligand-induced fit). In order to simulate the docking process more realistically, the protein potential to move should be taken into consideration. However, because of the huge magnitude of the degrees of freedom of the protein, this is not usually a straightforward task (Huang & Zou 2010). The following methods have been used as approaches to tackle the issue of receptor flexibility in docking.

#### **1.5.3.1.2.1 Soft docking**

Using soft potentials in docking is the easiest and fastest method to handle protein flexibility. The method involves softening the van der Waals component of the scoring functions. The result is allowing or tolerating some clashes or overlaps between the ligand atoms and those from the receptor. Being fast and computationally efficient is the major advantage of the method. However, it can only handle minor side chain movements in the protein and cannot represent backbone dynamics (B-Rao et al. 2009; CCDC Software Limited 2013).

#### **1.5.3.1.2.2 Rotamer libraries**

Rotamer libraries represent a database for the most commonly observed conformations or rotamers for the side chains of the amino acids. Implementing such databases in the docking software allows the program to perform a survey of the possible conformations the side chain of a particular amino acid can adopt during docking to produce the most favourable complex for the ligand and the receptor. The computational cost is highly dependent on the size of the library. The method only handles side chains and cannot produce conformations that are not included in the library (B-Rao et al. 2009; Alberts et al. 2005). An example of rotamer libraries is the Penultimate library (Lovell et al. 2000) that is used by the docking program GOLD (Genetic Optimisation for Ligand Docking) (CCDC Software Limited 2013).

#### **1.5.3.1.2.3 Docking into multiple protein structures**

This approach is similar in concept to applying rotamer libraries and involves performing the docking task of the same ligand more than once into different conformations of the same receptor while treating each receptor conformation as a rigid body. Depending on the number of receptor structures used, the method can be computationally expensive but it will be able to handle side chain and backbone motions in the receptor. It should be noted that the chosen structures may not represent the entire conformational space of the receptor (Huang & Zou 2007; B-Rao et al. 2009). The multiple structures of the protein can be

obtained from a variety of sources including NMR structures, multiple X-ray structures of the same protein or conformations of the protein generated from a MD simulation on an initial structure (Carlson 2002).

Ensemble docking is a variation of docking into multiple protein structures, in which an average grid structure for the receptor is generated from the multiple conformations before docking, and a rigid docking is performed on the single average structure. The approach is more efficient in terms of the computational time required, but the drawbacks include that not all the features in the structures can be averaged, and the resultant structure may produce non-realistic 'artificial' complexes with the ligands (Cozzini et al. 2008).

#### **1.5.3.1.2.4 Molecular relaxation**

This approach involves firstly docking the ligand into the binding site of the receptor without any account for receptor flexibility. Then, the generated complex is relaxed as a whole or just the part that involves the ligand and the neighbouring receptor residues. The relaxation is performed through MD simulations. These methods allow for the exploration of a wider space of the conformational universe of the protein, including its backbone. They also give the possibility of testing the stability of the docking generated poses of the ligands; in other words, the methods enable the monitoring of the ability of the ligand to maintain its pose for any given period of simulations time. The main drawback of relaxation methods, however, is that they are computationally very expensive. MD can also be used on the receptor on its own to generate multiple conformations that could be used separately (Hospital et al. 2015; B-Rao et al. 2009; Huang & Zou 2010).

#### **1.5.3.2 The scoring function**

A scoring function is the second half of any docking protocol besides the searching algorithm. Scoring functions are responsible for estimating the binding free energy of the complexes generated by the searching algorithm. They are included in the docking process for several purposes; to distinguish between a correct binding mode of the ligand within the binding site from an incorrect one, to rank the generated poses of the ligand based on their binding affinities, to discriminate between active compounds and non-binders, and to rank several active compounds according to their biological activities toward the receptor. An ideal scoring function would be able to do all that, but this will be computationally impractical. Therefore, scoring functions are developed with approximations to achieve reliable results within feasible computation time (Meng et al. 2011; Sousa et al. 2006). The types of most widely used scoring functions are presented below.

### 1.5.3.2.1 Force field-based scoring functions

Force field functions attempt to estimate the binding energy between the ligand and the receptor by calculating two terms, the ligand internal energy and the energy of the non-bonded interactions between the ligand and the receptor. The non-bonded interactions involved are of two types, van der Waals interactions, which are calculated by a Lennard-Jones potential, and electrostatic interactions estimated through a Columbic formula (Sousa et al. 2006). Classical functions of this kind do not account for hydrogen bonds and solvation and entropic effects and include a cut-off distance for handling long range non-bonded interactions. Some newer functions based on force fields do include terms for hydrogen bonds or solvation (Meng et al. 2011). An example is the GoldScore function used within the GOLD docking program (CCDC Software Limited 2013; Verdonk et al. 2003).

### 1.5.3.2.2 Empirical scoring functions

The binding free energy between the ligand and the receptor in this type of functions is decomposed into several terms that describe hydrogen bonding, van der Waals interactions, metal ligation, hydrophobic effects, entropy, solvation and electrostatic interactions. Parameters for those terms are included in the scoring function for the calculation of the binding energy. The parameters are derived from a defined set of protein-ligand complexes, usually from the Protein Data Bank. Empirical scoring functions are more efficient than the force field functions because of their simpler energy terms. Nonetheless, their performance is highly dependent on the training set used to derive the parameters and may not function with the same efficiency between different protein-ligand systems (Meng et al. 2011; Huang & Zou 2010).

### 1.5.3.2.3 Knowledge-based scoring functions

Knowledge-based scoring functions estimate the score of a particular pose for the ligand within the binding site by calculating the sum of interaction energies between each atom pair from the ligand and the protein, within a predefined cut-off sphere. The interaction energies for the atom pairs are obtained from a statistical analysis of different ligand-receptor atom pairs in crystal structures, after converting them into distance dependent pairwise potential (potential of mean force or PMF). This type of scoring functions provides a compromise between speed and accuracy and is a middle ground between the previously mentioned scoring functions (Moitessier et al. 2008; Huang & Zou 2010).

Consensus and clustering scorings are scoring techniques designed to improve the scores obtained from the various scoring functions. In the former, the scoring information from

multiple scoring functions is merged to obtain better, consensus score. In clustering scoring, the generated binding modes for the ligand are clustered and the entropic contribution to scoring can be estimated by examining the clusters, thereby, adding entropic impact to the docking score (Huang & Zou 2010).

### 1.5.3.3 GOLD overview

GOLD (Genetic Optimisation for Ligand Docking) program version 5.2.2 (CCDC Software Limited 2013; Jones et al. 1997; Jones et al. 1995; Nissink et al. 2002; Verdonk et al. 2003; Cole et al. 2005; Verdonk et al. 2005; Hartshorn et al. 2007) was the main docking software used during the course of this thesis. The docking method used in the program is genetic algorithm that provides flexible treatment of the ligand and optional flexible handling of the side chains of up to 10 residues of the receptor active site. The flexible treatment of the receptor is achieved through one of 3 options; either allowing the side chain to rotate freely, choosing rotamers for the side chain from the Penultimate rotamer library (Lovell et al. 2000) or choosing the rotamers for the side chain that are specified in the PDB file for the protein. In addition to rotamers, the program offers the possibility of docking with soft potentials for any number of active site residues. GOLD is equipped with 4 different scoring functions; the Astex Statistical Potential (ASP) scoring function (Mooij & Verdonk 2005) (knowledge-based), ChemScore (Eldridge et al. 1997; Baxter et al. 1998) (empirical), CHEMPLP (Korb et al. 2009) (empirical) and GoldScore (Verdonk et al. 2003) (force field). Detailed description and comparisons between the scoring functions is presented in section 3.6.

Besides standard docking, GOLD provides some non-standard docking functions. Examples are the possibility of including up to 25 water molecules within the active site to test the effects of the presence of water on the docking in terms of water displacement and using water to mediate interactions with the protein. Another function is the use of hydrogen bonding constraints during docking. This utility forces the formation of a hydrogen bond between predefined atoms from the ligand and the protein and is used to reproduce certain structural hydrogen bonds during the docking. The third example involves docking with a distance constraint, which is similar to the hydrogen bond constraint in specifying a ligand atom and a protein atom and forcing them during the docking process to lie within a predefined distance range. The final example is the possibility of covalent docking, in which GOLD forces the formation of a covalent bond between predefined atoms from the ligand and the protein, and poses the rest of the ligand accordingly. In the biased docking methods (hydrogen bond and distance constraint), a penalty is added to the score if the specified conditions are not met.

## 1.6 Computational Chemistry and TG2 Inhibitors

Very few studies used computational chemistry techniques to aid in the design and testing of TG2 inhibitors. One of the earliest studies was by Iwata et al. (2000) and involved docking a series of *in vitro* tested compounds with a cyclopropanone ring against FXIIIa. Their results showed that the cyclopropanone ring should come close to the catalytic cysteine for activity. They also pointed the importance of a neighbouring tryptophan residue through forming a hydrogen bond with the inhibitor. The tryptophan residue (TRP279) is located at the top of the tunnel in which the catalytic cysteine is located and corresponds to residue TRP241 in TG2 (more details in section 3.1). Additionally, they showed that the terminal phenyl ring of the inhibitor would position in the hydrophobic region of the enzyme active site.

Chica et al. (2004) performed molecular modelling experiments involving their proposed peptide substrates of TG2 (section 1.4.2.1) to aid in the understanding of the mode of binding of the substrate. The researchers performed docking and molecular dynamics (MD) studies using the coordinates for red sea bream TG2. Their most significant findings were the role of TRP329 (TRP332 in human version) in acting as a gate to the active site, where its side chain moves to open and close the active site, possibly depending on calcium ion concentration. They also recorded a hydrophobic interaction with the substrate involving TRP236 (TRP241 in human TG2) that is important for positioning the glutamine residue to the catalytic cysteine.

Another study was performed by Tagami et al. (2009) to investigate the binding mode of a peptide substrate (Cbz-Gln-Gly) to microbial transglutaminase using a docking study and MD. They were able to conclude that a distance of around 4 Å between the side chain of glutamine residue in the substrate and the catalytic cysteine sulphur atom is important for activity along with some pi-interactions between the substrate and some active site residues such as leucine, isoleucine, phenylalanine and tyrosine.

Prime, Andersen, et al. (2012) conducted GOLD covalent docking experiments on compound **i10** (the initial sulphonamide modification, Figure 1-13). Their results identified the key residues of TG2 which are involved in forming interactions with the inhibitors. Examples of these interactions included hydrogen bonds between the warhead acrylamide CONH groups and TRP241 and GLN276. The generated poses also suggested the possibility of hydrogen bonding involving the central sulphonamide group and ASN333. This potential hydrogen bond would facilitate the interactions of the lipophilic part of the inhibitor within the hydrophobic loop of TG2 active site. That was the basis for their next series of modifications (section 1.4.2.2).

The study published by Badarau, Mongeot, et al. (2013) also included docking and MD. Two inhibitors with sulfonium ion and imidazolium warheads were docked into TG2 active site and MD simulation was applied on the docked complexes to examine the change in pose with time. The docking and MD results showed that the inhibitors adopt a bent conformation to facilitate the access of the warhead to the catalytic cysteine residue. The bent conformation involves posing the warhead deep in the catalytic tunnel. A hydrogen bond with residue TRP241 was seen to stabilise the bent conformation during MD simulation, in addition to interactions with TRP332 (the other bridging tryptophan residue). These interactions were mostly evident with the imidazolium warhead, and the authors suggested that they are responsible for the lower selectivity towards FXIIIa because these residues are conserved in FXIIIa (Badarau, Mongeot, et al. 2013).

No studies are available that include docking of larger sets of compounds or examination of the interactions within TG2 active site in more details. Furthermore, MD simulations, which are very powerful tools in computational chemistry, have not been used extensively with regard to TG2.

## 1.7 Aim and Objectives

The project involved computational analysis of TG2 enzyme and its inhibitors with the ultimate aim of developing modelling methodologies that could aid in the design and testing of new potential inhibitors for the enzyme. Towards this aim, a series of computational objectives have been sought. These included:

- Docking a small set of known TG2 inhibitors into the enzyme active site.
- Application of MD to the docked complexes.
- Development of TG2 active site models from the MD trajectories.
- Validation of the TG2 models by docking of additional known TG2 inhibitors, MD simulations, covalent docking and binding free energy calculations, with a view to establishing a correlation between the experimentally-determined biological activities and calculated molecular properties such as docking scores and binding free energies.
- Investigation of the mechanism of covalent inhibition of TG2 by the use of QM methods with the aid of techniques such as umbrella sampling and reaction path experiments, with a view to establishing a correlation between experimentally-determined biological activities and activation energies and reaction energies.
- Attempting to validate predicted allosteric sites of inhibition in TG2 for known allosteric inhibitors by means of docking and MD simulations.

# **Chapter 2**

## **Materials and Methods**

## 2 Materials and Methods

A general description of the methods followed during the course of this thesis is presented in this chapter, preceded by a listing of the computers and the programs used to carry out the simulations. Some specifics for certain calculations are mentioned in the relevant results chapters.

### 2.1 Materials

#### 2.1.1 Computers

For Microsoft Windows XP® Professional Version 2002 Service Pack 3, a desktop computer with the following specifications was used: Intel(R) Core(TM) i7-2600, CPU 3.4 GHz, RAM 2.94 GB. A Windows 7 Enterprise® 64-bit computer with the following specifications was also used; Intel(R) Core(TM) i5, CPU 3.4 GHz, RAM of 6.00 GB. Ubuntu 12.04 operating system was used to run the AMBER Package on an Intel(R) Core(TM) i7-3820 with a CPU of 3.6 GHz and a RAM 6 GB desktop computer. The latter computer is also equipped with two GeForce® GTX TITAN GPU cards from NVIDIA®, each with 6 GB of memory and 2688 CUDA cores. This was the main PC used for AMBER. In addition, AMBER was run on a number of other PCs and Aston University servers. The PCs were similar to the main Ubuntu PC but with GeForce® GTX 780 GPU cards, also from NVIDIA®. The servers had GeForce® GTX 780 and GeForce® GTX 780 Ti and their CPUs were Intel(R) Xeon(R) and varying memories. All the PCs and servers are the property of Aston University.

#### 2.1.2 Programs

1. AMBER (Assisted Model Building and Energy Refinement) Package, Version 12 (Case et al. 2012), was used for the minimisation of the protein and later on for the MD simulation. This version of AMBER supports GPU acceleration through CUDA (CUDART driver version 5.5 and CUDA Runtime version 5.0) (Salomon-Ferrer, Götz, et al. 2013).
2. AmberTools 13 (Salomon-Ferrer, Case, et al. 2013) is a package of supporting programs developed by AMBER group for the preparation of input and analysis of AMBER MD trajectories. Example programs used in this work include *LEaP* which was used for the generation of input files for MD, *antechamber* which was used for the generation of parameter and topology files of non-standard residues, *CPPTRAJ* was used to perform the post-MD analysis on the trajectories including calculating root-mean-square deviation (RMSD), hydrogen bond profiles and others, and

*MMPBSA.py* used for the calculation of the binding free energy between ligands and protein after a MD simulation (section 1.5.2.6).

3. GOLD Suite (Genetic Optimisation for Ligand Docking) Version 5.2.2 (CCDC Software Limited 2013; Jones et al. 1997; Jones et al. 1995; Nissink et al. 2002; Verdonk et al. 2003; Cole et al. 2005; Verdonk et al. 2005; Hartshorn et al. 2007) was used for various types of docking applied in the course of this work, for instance, flexible active site, covalent and water dockings.
4. CAChe (Computer-Aided Chemistry) WorkSystem Pro Version 7.5 (Fujitsu Limited 2006) was used for the drawing and minimisation of test compounds, docking and preparing the enzyme for MD. CAChe was also used to perform reaction path experiments on the sulfonium ion TG2 inhibitors.
5. Accelrys Discovery Studio (DS) Visualizer Version 3.5 and 4.0 (Accelrys Software Incorporation 2012; Accelrys Software Incorporation 2013) was used for visualisation of the docked complexes and analysis of the interactions between the compounds and the active site. It was also used to generate the 3-dimensional pictures presented in this work.
6. VMD (Visual Molecular Dynamics) Version 1.9 (Humphrey et al. 1996) was used for visual inspection of MD trajectories as well as for performing analysis on the trajectories including hydrogen bonding and RMSD.
7. GAMESS (General Atomic and Molecular Electronic Structure System) version 13-64 (Schmidt et al. 1993) was used to calculate the electrostatic potential for ligands during covalent docking and MD.
8. pyPcazip (Shkurti et al. 2016) was the program used to perform the principal component analysis (PCA) for the MD simulations.

AMBER, CPPTRAJ and pyPcazip were run under Ubuntu. The others were run on Windows (7 and XP). VMD was run on both.

### 2.1.3 Compounds

A range of active and inactive TG2 inhibitors were used in this work for the testing and validation of the various active site models and for the subsequent experiments. The details of the structures of the used compounds are in the relevant results chapters. The compounds were mainly adopted from the work by Badarau et al. (2015) and the associated patent (Griffin et al. 2014).

## 2.2 Methods

### 2.2.1 Chapter 3 methods (model development and validation)

#### 2.2.1.1 MD simulation on empty TG2

##### 2.2.1.1.1 Preparation of TG2 for molecular dynamics

The crystal structure of TG2 was downloaded from the Protein Data Bank ([www.rcsb.org](http://www.rcsb.org)) (Berman et al. 2000) as the PDB entry 2Q3Z (Pinkas et al. 2007). Human TG2 contains 687 amino acid residues; in the crystal structure 2Q3Z, 32 of these residues were missing. In addition, the structure contains a covalently bound inhibitor, Ac-P(DON)LPF-NH<sub>2</sub>, in the active site that is connected to CYS277 by a covalent linkage. The structure also contains 5 sulphate ions and 261 water molecules. The names and numbers of the missing residues are (Pinkas et al. 2007):

GLN307, ASN308, GLU319, PHE320, GLY321, GLU322, ILE323, GLN324, GLY325, ASP326, LYS327, GLN407, ASP408, ASP409, GLY410, SER411, VAL412, HIS413, LEU462, ASN463, LYS464, LEU465, ALA466, GLU467, LYS468, GLU469, GLU470, THR471, ILE684, GLY685, PRO686, ALA687

All the missing residues were added using CAChe Workspace's 'Insert Residue' function. Hydrogen atoms were added to the structure and water molecules and sulphate ions deleted. The whole protein was then locked (atoms' X, Y and Z coordinates frozen) with the exception of the added residues and an MM2 energy minimisation (Allinger 1977) was applied to relax these added residues. Residues 1-154 and 586-687 were deleted being far from the active site (Badarau, Mongeot, et al. 2013) to reduce computational cost. The remaining residues represent the catalytic core in TG2, while the deleted residues represent the N-terminal  $\beta$ -sandwich and the C-terminal  $\beta$ -barrels (section 1.2.3).

At this point the active site was defined. It was taken as a sphere of radius 8 Å working outwards from residues TRP241 and TRP332 (bridging residues of the active site tunnel, more detail in Chapter 3). The inhibitor was deleted and the residues forming the active site were recorded. CAChe Workspace was used for this. This is a list of these active site residues:

GLY170, SER171, ALA172, ASN229, LEU237, LEU238, GLY239, ARG240, TRP241, ASP242, ASN243, ASN244, TYR245, SER250, PRO251, MET252, TYR274, GLY275, GLN276, CYS277, TRP278, VAL279, PHE280, THR299, ASN300, TYR301, ASN302,

SER303, ALA304, PHE316, ARG317, ASP326, LYS327, SER328, GLU329, MET330, ILE331, TRP332, ASN333, PHE334, HIS335, CYS336, TRP337, ASP358, PRO359, THR360, PRO361, GLN362, ALA395, GLU396, LEU420, ILE421, VAL422, GLY423, LEU424.

Hydrogen atoms were deleted and a PDB file was saved as the starting point for the MD run. CYS277 was converted to CYM for AMBER to handle it as deprotonated and HIS335 to HIP to be treated as protonated.

The file was loaded into the LEaP program of the AMBER package to prepare the parameter-topology (PRMTOP) and coordinate (INPCRD) files using the AMBER force field ff99SB (Hornak et al. 2006). Before that 10 sodium ions were added to neutralise the charge of the protein. The latter was then solvated in a truncated octahedron of TIP3P (transferable intermolecular potential 3P) (Jorgensen et al. 1983) water molecules that extended for 8 Å from the protein surface. The total number of water molecules that were added was 7486. Finally, the PRMTOP and INPCRD files were generated and saved.

#### **2.2.1.1.2 Applying energy minimisation**

The PMEMD (Particle Mesh Ewald Molecular Dynamics) program of the AMBER package was used to run the minimisation on TG2. The minimisation was run for 2,000 cycles (maxcyc = 2000) using steepest descent for the first 1,000 cycles and conjugate gradient for the second 1,000 cycles (ncyc = 1000). The cut-off distance for treating non-bonded interactions was taken as 12 Å. Also, the minimisation was run at a constant volume (ntb = 1) with no pressure scaling, and using an explicit water model (igb = 0).

#### **2.2.1.1.3 Initial MD simulation (MD1), heating up the system**

The minimisation output file was used as the starting point for this stage of the MD simulation (ntx = 1). The heating stage of MD was also run under constant volume (ntb = 1) using a cut-off distance of 12 Å without any restraints (ntr = 0). An explicit water model was used (igb = 0) and the temperature of the system was taken from 0-300 K using Langevin dynamics (ntt = 3) at a collision frequency of 1 ps<sup>-1</sup> (gamma\_in = 1). The simulation was run for 20,000 steps (nstlim = 20000) at 0.001 ps time step for a total of 20 ps. The random number generator (ig = -1) was used during this stage.

#### **2.2.1.1.4 Second MD simulation (MD2), equilibrating the system**

The restart file obtained at the end of MD1 was used as the starting structure for this stage of the MD simulation. MD2 was run at constant temperature of 300 K for 5 nanoseconds

(5,000,000 steps of 0.001 picoseconds). The simulation was run under constant pressure periodic boundaries ( $ntb = 2$ ) using isotropic position scaling ( $ntp = 1$ ). Updating output files was done every 10,000 steps. The remaining criteria were the same as those of MD1. The random number generator was also applied during the equilibration phase.

#### **2.2.1.1.5 Third MD simulation (MD2\_fast), production phase**

Again, the restart file of MD2 was used to start this stage of MD. The simulation was run for 30 ns (15,000,000 steps of 0.002 ps). In this stage, constant volume periodic boundaries were applied with no pressure scaling ( $ntb = 1$ ,  $ntp = 0$ ). The run was also performed at constant temperature (300 K) but this time using the weak coupling algorithm with zero collision frequency ( $\gamma_{in} = 0$ ). The SHAKE algorithm was applied to restrain bonds involving hydrogen atoms ( $ntc = 2$ ,  $ntf = 2$ ). The wrapping function ( $iwrap = 1$ ) was used in this stage so that the coordinates written to the output files were to be wrapped into the primary box to make the trajectory more realistic. The output files were updated every 12,500 steps.

During the heating phase (MD1), using constant pressure will result in an unstable system as the system is cold at the start. However, constant pressure must be used during the equilibration phase (MD2) to allow for the density of the system, and water molecules in particular, to equilibrate. The weak coupling algorithm ( $ntt=1$ ) was used in the production phase (MD2\_fast) to hasten the calculation. This algorithm does not ensure even distribution of the temperature and therefore should only be used after equilibrating the system (Walker & Cheatham 2010).

The minimisation and the three stages of MD simulation were applied using the CUDA version of commands for AMBER 12 that involves the use of GPUs rather than CPUs to run AMBER calculations (Götz et al. 2012; Le Grand et al. 2013; Salomon-Ferrer, Götz, et al. 2013).

The production run was continued by applying new simulations at the end of every 30-ns simulation using the restart file from the last simulation as the starting structure for the new one. To take the system to a simulation time of 515 nanoseconds, MD2\_fast was repeated 17 times (5 ns from equilibration plus  $30\text{ns} \times 17$  for each production run). The CPPTRAJ program of the AmberTools was used with each resulting trajectory from MD2\_fast to centre the protein in the solvent box. CPPTRAJ was also used to generate a single trajectory combining all the individual trajectories to be used later by the same program for the generation of RMSD and atomic fluctuation charts. A PERL script was used to extract temperature and energy data from the OUT files for the analysis of the resulting trajectories.

### 2.2.1.1.6 Extracting models from MD trajectory

The VMD program was used for the extraction of conformations from the production phase trajectories. The conformations used for the initial docking were taken at every 5 ns starting from the tenth to the 100<sup>th</sup> nanosecond. The models were named by the nanosecond at which they were taken. These conformations were edited in a text editor to remove water molecules and sodium ions and to convert CYM and HIP to CYS and HIS respectively. Each of the conformations was checked for structural plausibility by means of a Ramachandran plot (Ramachandran & Sasisekharan 1968) using DS Visualizer.

### 2.2.1.2 Preparation of the compounds

All the test compounds used in this work were drawn and corrected for bond lengths, bond angles and atom valence and hybridisation states using CAChe Workspace. They were then energy minimised using a molecular mechanics force field (MM2) (Allinger 1977) and a semi-empirical quantum mechanics with PM3 parameter set (Stewart 1989; Stewart 2002) methods. CONFLEX conformational search (Gotō & Ōsawa 1993) was also applied to the rotatable bonds within the individual compounds prior to the PM3 energy minimisation. The energy minimisation and the CONFLEX search were performed using the CAChe Workspace.

### 2.2.1.3 Docking

At this stage of docking, the objective was to obtain some valid complexes to run MD on. For this reason, only six active test compounds were used.

#### 2.2.1.3.1 CAChe docking

The FastDock program within CAChe Project Leader was used to perform the docking study. The following specifications were used for the docking:

Type of docking: Flexible active site side chains and flexible ligands.  
Docking method: Lamarckian genetic algorithm  
Scoring Function: PMF (potential of mean force)  
Calculation Type: Dock (Use Grids)  
Use Amber van der Waals: Grid Spacing (Å) 0.30000  
Pop Size: 50  
Crossover Rate: 0.80000  
Elitism: 5  
Maximum Generations: 3000  
Mutation Rate: 0.20000  
Convergence: 1.0000

Four dockings were performed for each compound to increase the probability of finding the correct binding mode. All the complexes of the docking results were converted to PDB files to be viewed by DS Visualizer for analysis of the docking pose. Typically, the run time for a single docking was in the region of 4 hours.

#### **2.2.1.3.2 GOLD docking**

Hermes 1.6 of the GOLD Suite 5.2 was used to set up and run the dockings. The protein was loaded into Hermes as a PDB file and the active site was defined using the “List of atoms or residues” method of defining the binding site in Hermes. The compounds were then added to Hermes as MOL2 files. The number of GA (genetic algorithm) Run (corresponding to the number of docking solutions) for each compound was set to 20. GoldScore (Verdonk et al. 2003) was used as the primary scoring function for docking. Each docking was also rescored using the CHEMPLP (Korb et al. 2009) scoring function.

Early termination (a function in GOLD that terminates the docking if the top ranked solutions are very similar) was turned off to allow for complete exploration of the binding possibilities. “Genetic algorithm search options” was set to ‘Automatic’ where search efficiency was set to 200% (Very Flexible) with 10,000 operations to be performed as a minimum. Ten residues were selected from the active site residues to be treated as flexible during docking. The flexibility was applied using the “Library” option in “GOLD”. The “Library” option produces rotamers for the side chain that are consistent with the Penultimate Rotamer Library (Lovell et al. 2000) which includes the most commonly observed side chain conformations for the natural amino acids. Molecule Explorer in Hermes and DS Visualizer were both used for the analysis of the docked complexes.

#### **2.2.1.4 AMBER MD simulations on docked complexes**

A good docking complex for each of the six compounds was chosen to be taken into further MD simulation. The criteria chosen to define a “good docking complex” are presented in section 3.1 of the results. Each complex was first edited by removing the hydrogen atoms from the protein but not from the ligand and changing CYS277 to CYM and HIS335 to HIP. A separate file containing the ligand’s coordinates only was also saved.

The antechamber programme in the AMBER package was used to generate the parameter and topology files (PREPIN and FRCMOD files) for the ligand using the general AMBER force field (GAFF) (Wang et al. 2006; Wang et al. 2004). The generated files were loaded in the LEaP program of the AMBER package together with the complex PDB file to prepare the

parameter-topology (PRMTOP) and coordinate (INPCRD) files after neutralising the complex with sodium ions and adding an 8-Å truncated octahedron of TIP3P water molecules.

Minimisation and 3-stage MD simulation were then applied to the protein using the same settings as those used for empty TG2. The simulation in this case was allowed to continue for 275 nanoseconds for each ligand-protein complex. CPPTRAJ was also used here as for the empty TG2 to combine all the trajectories in one trajectory and to generate RMSD and atomic fluctuation charts. In addition, CPPTRAJ was used to generate, from the original combined trajectory, a new trajectory in which all the water molecules were deleted except 25 molecules that were closest to CYS277.

The trajectory obtained from each MD run was analysed thoroughly using VMD especially looking for hydrogen bonds between the ligand and active site residues. Also the distance between the warhead and the sulphur atom of the catalytic cysteine residue was measured over the whole trajectory. RMSD for the individual ligands and that of TG2 and the atomic fluctuation data were all computed for all the simulations.

### **2.2.1.5 Docking into models from MD trajectories of complexes**

Models of the active site were extracted from these trajectories at 30-nanosecond intervals to perform docking experiments and to select valid models. GOLD was the main docking program used at this stage with settings similar to those used previously with the exception of turning on 'Early Termination' and using default settings for docking speed. Valid models were TG2 active site models which gave good docking results (section 3.1 for a definition of "good"). These valid models were further validated by the docking of additional active and inactive compounds, docking with changed GOLD settings (docking with water molecules in the active site and covalent docking) and applying 5-ns MD simulations on different docked complexes.

#### **2.2.1.5.1 Binding free energy calculations**

Binding free energy calculations were applied to the trajectories obtained from the 5-ns MD simulations (section 3.4.5), where the trajectories were first subjected to post-MD treatment with CPPTRAJ (Roe & Cheatham 2013) that involved auto-imaging. This was followed by the generation a specific set of parameter and topology files for the ligand, the protein and the complex with LEaP program to be used in the calculation of the binding free energy. The complex was prepared by combining the PDB files of the ligand and the protein. Parameter-topology and coordinate files were additionally prepared for the complex after neutralising and solvating it with an 8 Å octahedron of TIP3P water molecules.

The binding free energy was calculated with Poisson Boltzmann (PB) and Generalized Born (GB) models using the MMPBSA.py program of the AMBER package (Srinivasan et al. 1998; Miller et al. 2012). In addition, QM treatment was also applied to the ligand and the active site cysteine residue during separate calculations with the GB model. The QM methods used were PM3 (Stewart 1989) and DFTB (Seabra et al. 2007; Elstner et al. 1998). For the GB calculations, *igb=5* [GB-OBC model (Onufriev et al. 2004)] was used for the normal and QM calculations. A value of 0.1 M was used for the salt concentration. For the PB model (Srinivasan et al. 1998), the default settings were used except for the ionic strength which was set to 0.1 M. The calculations were performed on all the 200 frames in each simulation (5 ns and saving every 0.025 ns) without omitting any frame.

### **2.2.1.6 Covalent docking and MD**

#### **2.2.1.6.1 Docking**

The GOLD software offers the possibility of performing docking experiments that involve the formation of a covalent bond between predefined atoms from the ligand and the receptor. The function requires that the same linking atom be present on both molecules, the protein and the ligand (CCDC Software Limited 2013). The protein atom selected was the SG atom of CYS277. A sulphur atom was added to the ligand as a linking atom. It was added to the electrophilic carbon atom of the acrylamide warhead. Other GOLD settings were similar to those used in the first validation experiment (20 solutions per ligand, GoldScore for primary scoring and CHEMPLP for rescoring and default docking speed settings).

#### **2.2.1.6.2 Molecular dynamics**

Since there is a covalent bond in the complex to be used as the starting structure in MD, a new residue should be defined for AMBER to recognise the ligand. This was performed over multiple steps.

##### **2.2.1.6.2.1 Capping the ligand**

The ligand was extracted from the complex file obtained after GOLD covalent docking and the connecting sulphur atom was deleted. A cap was introduced to account for the missing valence in the electrophilic carbon of the ligand. The cap chosen was -SCH<sub>3</sub>. The cap was added using CAChe Workspace. The whole molecule was then minimised by MM2. A CONFLEX (Gotō & Ōsawa 1993) search with MM2 was applied followed by a PM3 minimisation on the lowest energy conformation produced by CONFLEX. The ligand was then saved as a PDB file and taken to GAMESS program to calculate its charge.

#### 2.2.1.6.2.2 GAMESS calculation

The GAMESS program (Schmidt et al. 1993) was used to perform an energy minimisation with a high level method and to calculate the electrostatic potential of the molecule. The latter was then used to assign charges to the individual atoms of the ligand. Molby program (Nagata 2014) was used to prepare the input files of the various ligands for GAMESS calculations. The settings involved running an optimisation (RUNTYP=OPTIMIZE) for 100 cycles (NSTEP=100) using 6-31G\* method (GBASIS=N31 NGAUSS=6 NDFUNC=1). The electrostatic potential calculation was requested using IEPOP=1.

#### 2.2.1.6.2.3 RESP fitting

RESP (restrained electrostatic potential) (Bayly et al. 1993) is a free program incorporated into AMBER and can be used to assign charges to the atoms of a molecule whose electrostatic potential has been calculated. A script written by Hans de Winter (Rega Institute for Medical Research, Belgium) was used to convert the electrostatic potential grid produced by GAMESS into a format that is readable by RESP. Charge fitting was performed over 2 stages. In the first stage, the charges of the cap atoms were fixed to the values of the charges of atoms SD, CE, HE1, HE2 and HE3 of the side chain of methionine residue in the AMBER force field.

The second stage involved restraining all non-carbon, non-hydrogen atoms to the charges fitted in the first stage. Carbon and hydrogen atoms that are not parts of methyl or methylene ( $\text{CH}_2^-$  or  $\text{CH}_3^-$ ) were also restrained to the charges of the first stage. Methyl and methylene carbon atoms and the first hydrogen atom attached to them were fitted freely while the 2<sup>nd</sup> and/or 3<sup>rd</sup> hydrogen atoms were assigned the same charge as the first hydrogen of that group.

#### 2.2.1.6.2.4 Library building

The antechamber program of the AMBER package was used to assign atom types for the atoms of the capped ligands based on the GAFF force field (Wang et al. 2006; Wang et al. 2004). The Parmchk2 program of the AMBER package was used on the file generated by antechamber to check the parameters produced. Antechamber and Parmchk2 were used on the capped ligand to account for the types and parameters of the ligand atoms involved in forming the covalent bond. A PDB file of the original ligand without the cap and after removing the connecting sulphur atom was loaded into LEaP program of AMBER package to generate a library file of the ligand. Atom types and charges were assigned manually for

individual atoms from antechamber and RESP respectively using LEaP, and a library file was saved.

#### **2.2.1.6.2.5 Parameter-topology and coordinate files**

All the above stages were applied on individual ligands once and the library file produced from LEaP and frcmod file from Parmchk2 were used to build parameter and coordinate files for all the complexes involving that specific ligand. The complex PDB file was loaded into LEaP program of AMBER package together with the library and frcmod files. CYS277 was converted to CYX. A covalent bond was added explicitly between SG of CYS277 and the electrophilic carbon of the ligand using LEaP. The complex was then neutralised with sodium ions and solvated with an 8 Å octahedron of TIP3P water molecules and parameter and coordinate files were saved.

#### **2.2.1.6.2.6 Molecular dynamics**

After the generation of parameter and coordinate files, conventional MD was applied as before. The process started with a 2,000-cycle minimisation, followed by 20 ps of heating MD under constant volume, equilibration under constant pressure for 5 ns and production under constant volume for 15 ns. Only the production phase was analysed. The analysis included measuring the RMSD of the ligand using the “RMSD Trajectory tool” of VMD program as an indicator for the ligand behaviour during the simulation. Hydrogen bond analysis was also performed using the CPPTRAJ program of the AmberTools package.

### **2.2.2 Chapter 4 methods (QM work)**

#### **2.2.2.1 AMBER umbrella sampling simulations**

These experiments were applied to 6 active compounds, all having acrylamide warheads. Their structures as well as biological activities are presented in Chapter 3. The starting structures were either docked complexes taken from the various validation processes or snapshots from the 5-ns MD simulations. For each compound, the structure selected was the one that had the lowest distance between SG from CYS277 and the compound's electrophilic carbon (the beta carbon of the acrylamide double bond C $\beta$  or EC). Good docking rank and favourable interactions with active site residues were also considered in the selection process.

The starting structures were loaded in the LEaP program of the AMBER package to generate the parameter and coordinate files after generating those files for the compounds with antechamber with ff99sb and GAFF force fields respectively, neutralising and solvating

the complexes with in an 8-Å truncated octahedron of TIP3P water molecules. The systems were then minimised for 2,000 cycles and heated under constant volume from 0 K to 300 K over 20 picoseconds (ps) using Langevin dynamics. The heating was followed by equilibrating the system for 5 ns under constant pressure also using the Langevin dynamics. NMR restraints were used during MD to restrain the distances involved in the reaction at their starting values. The force constant used was 500 kcal/mol.Å<sup>2</sup>. The distances fixed were the one between SG of CYS277 and EC of the compound (D1 in Figure 2-1) and the distance between HD1 from HIS335 and the alpha carbon of the double bond of the acrylamide (C $\alpha$  or C2) (D2 in Figure 2-1).

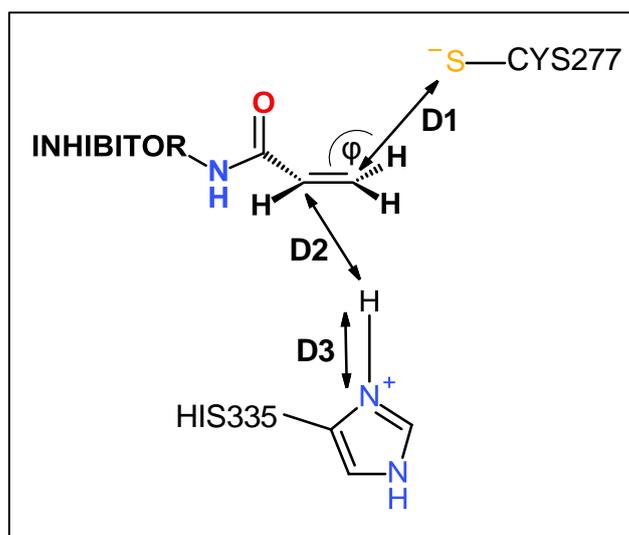


Figure 2-1: Restrained distances in umbrella sampling simulations.

### 2.2.2.1.1 QM/MM

Umbrella sampling (US) was performed using quantum mechanical treatment (Walker et al. 2008) of the region involved in the chemical reaction. The rest of the system was treated molecular mechanically using the ff99SB force field (Hornak et al. 2006). The QM region was composed of the acrylamide compound and the side chain of CYS277 (the negatively charged sulphur, SG and the connecting carbon with its 2 hydrogen atoms) and the imidazole ring of HIS335. PM3 (Stewart 1989) and SCC-DFTB (Elstner et al. 1998; Seabra et al. 2007) were the methods used to calculate the energy of the QM region in this work. PM3 was applied with a 12-Å cut-off distance. The same cut-off was used with SCC-DFTB along with an electronic temperature of 100 K.

### 2.2.2.1.2. Umbrella sampling (US)

For each compound and for each QM method, 2 US experiments were performed. The first simulation assumed that the reaction occurs over a single stage and the second simulation

involved two stages to drive the reaction to completion. All the simulations were performed under constant pressure. Before starting US, a 100-ps MD under constant pressure was applied on the structure resulting from the equilibration step. This MD was run with QM treatment of the reaction centre while fixing the distances between SG and EC and between HD1 and C2 to their starting values with a force constant of 250 kcal/mol.Å<sup>2</sup>. Each US simulation used the restart file from the 100-ps QM relaxation as the starting structure. The simulations were divided into windows along the reaction coordinate (RC); the number of these windows depended upon the starting value of the RC and the method used (1 or 2 stages). Each window was run over 5 ps. RC was controlled by a flat well restraint potential in all the US experiments.

#### 2.2.2.1.1.1 Single-stage US simulations

The reaction coordinate in this case was taken as a “generalised distance coordinate” involving the summation of 2 distances; the SG-EC distance (D1 in Figure 2-1) and the HD1-C2 distance (D2 in Figure 2-1). The RC for the 1-stage simulations was set to change over the windows from its starting value to 4.5 Å by 0.2-Å reductions and 50 kcal/mol.Å<sup>2</sup> force constant. After that, the reductions were 0.1 Å per window and the force constant was increased to 100 kcal/mol.Å<sup>2</sup>. The end value for the RC was set to 2.8 Å (≈1.8 Å for SG-EC bond and ≈1 Å for HD1-C2 bond). The same changes in the RC and the same number of windows were used for each compound with PM3 and SCC-DFTB. The number of windows ranged from 38 to 62 depending on the value of RC in the starting structure.

#### 2.2.2.1.1.2 Two-stage US simulations

The RC for the first stage of the 2-stage US simulation was a simple distance coordinate involving the distance between SG and EC. This was set to change from its original value to 1.6 Å over 0.1-Å reductions using a force constant of 250 kcal/mol.Å<sup>2</sup>; between RC values of 2.1 and 1.9 Å the reductions were set to 0.05 Å. These settings were used with PM3 and SCC-DFTB. The number of windows ranged from 19 to 43, also depending on the starting value of RC in the individual compounds.

For the 2<sup>nd</sup> stage of the 2-stage simulations, RC was defined as a “generalised distance coordinate” sampling the difference between the HD1-C2 distance (D2 in Figure 2-1) and HD1-ND1 distance (D3 in Figure 2-1). RC was set to change over 0.4-Å reductions using a force constant of 250 kcal/mol.Å<sup>2</sup> up to a value of 2.0 Å when the change was 0.2 Å per window with the same value for the force constant. The end point was set at -2.0 Å. The restart file generated from the last window of the 1<sup>st</sup> stage was used as the starting structure for this stage.

### 2.2.2.1.2 PMF calculations

The weighted histogram analysis method (Souaille & Roux 2001) was used to calculate the PMF profiles for the US simulations. The WHAM program (Grossfield 2013) was employed to construct the PMF graph along the reaction coordinate from the time series files generated during the US simulations. The number of bins for the 1-stage and the 2<sup>nd</sup> stage of the 2-stage simulations corresponded to how many 0.1-Å reductions were there during each simulation. There was 1 bin for each 0.1-Å reduction in the 1-stage experiments, while there were 4 bins for each 0.1-Å reduction for the 1<sup>st</sup> stage of the 2-stage simulations and there were 2.5 bins for each 0.1-Å reduction for the 2<sup>nd</sup> stage of the 2-stage simulations. The convergence tolerance used was 0.00001 in all the simulations.

### 2.2.2.2 CAChe reaction path experiments

QM reaction path experiments were applied on a set of 8 active compounds all having sulfonium ion warheads. The method used for the generation of the reaction path and the calculation of the activation energy was the same for all the compounds. The differences were in the choice of the starting conformation and whether it was energy minimised or not. The sources of the starting structures were MD frames or docking complexes (section 4.2.1.2, Table 4-6).

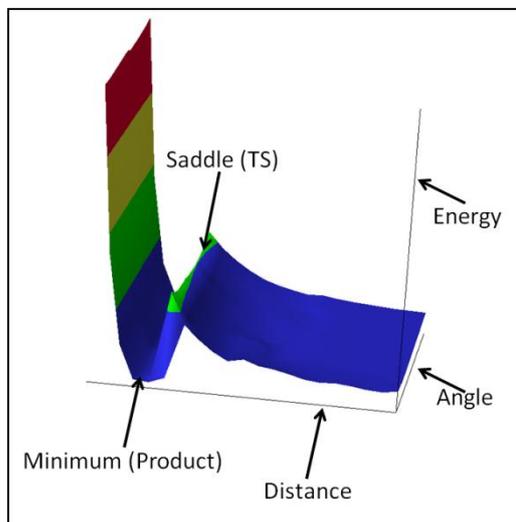
#### 2.2.2.2.1 Map reaction experiment

Minimisation within CAChe was applied to most of the starting structures using MM2. The process was applied on the compound and the cysteine residue while everything else in the system was locked. This was followed by deleting the entire system with the exception of the ligand and CYS277. Hydrogen atoms were added to CYS277 to complete its valence. Two reaction coordinates were chosen to map the reaction between the warheads containing the sulfonium ion and the negatively charged sulphur (SG) of CYS277. These were the distance between the electrophilic carbon (EC) of the ligand and SG of CYS277, and the angle between SG of CYS277, EC and the positively charged sulphur (S+) of the sulfonium ion in the ligand. In all the cases, the distance was allowed to change from its original value to 1 Å while the angle was changed from its original value to 180°. Both changes were applied over 20 steps. The QM method used with all the compounds was PM3.

#### 2.2.2.2.2 Transition state (TS) choice, refinement and verification

A map reaction experiment typically produces a 3-dimensional potential energy surface (PES) such as the one in Figure 2-2, when applied with 2 reaction coordinates. The axes

represent the reaction coordinates (distance and angle) and the energy of the system as a whole calculated by PM3. TS structures were chosen from the saddle points immediately next to the minimum having the same value for the angle as the minimum but with a different value for the distance (one step backwards).



**Figure 2-2: PES produced by map reaction.**

Refinement of the chosen TS was performed using a PM3 gradient minimisation. The refined TS was then verified again using a PM3 calculation of vibrational frequencies through the FORCE program of the CAChe package. Only TS structures that gave one, and only one, negative vibration in the verification process were considered.

#### **2.2.2.2.3 Reaction path experiment**

The refined and verified TS structure from each system was used as the starting point for the final experiment, the reaction path. The QM method used was also PM3 that was run for 200 cycles using the intrinsic reaction coordinate (IRC) method.

### **2.2.3 Chapter 5 methods (allosteric inhibition)**

Methods for this chapter are presented in the chapter itself (section 5.2) for their involvement of a technique that was detailed within Chapter 5.

# **Chapter 3**

## **Development and Validation of TG2 Active Site Models**

## 3 Development and Validation of TG2 Active Site Models

This chapter of the thesis covers the work that involved preparing models for the active site of TG2 which could be used for testing potential inhibitors for the enzyme by docking. The active site models were generated by an approach that combines MD simulations with docking. MD was applied initially on an empty version of TG2, and selected docking complexes were subjected to MD, and from the resultant trajectories, conformations were selected and tested for their ability to dock already tested inhibitors of TG2. The produced models were then validated by a series of docking experiments in which additional test compounds were introduced, some of them were inactive. Other docking experiments were also applied, that involved docking with water molecules in the active site, inflexible docking and covalent docking. Short MD simulations were also employed to judge the quality of the docking complexes produced in the various experiments. Binding free energy was calculated and correlated to biological activity. The behaviour of TG2 during various MD simulations was also discussed in this chapter.

### 3.1 Results of Initial Docking

The main purpose of the MD simulation performed initially on empty TG2 was to supply some conformations that could be used as models for the active site. Two reasons lie behind this justification for requiring more than one conformation of the enzyme for docking purposes. The first is that it has been around 40 years since the introduction of MD to biological macromolecules (McCammon et al. 1977) and the beginning of an era that started to consider proteins as molecules that are in continuous motion and cannot be represented by a single physical conformation (Karplus & McCammon 2002). The second reason is that a number of residues were missing from the original crystal structure and were added before the start of MD. These added residues needed to relax and accommodate themselves to their locations to improve the process of their integration into the protein, and molecular dynamics is an approach that could be used in this respect (Hospital et al. 2015).

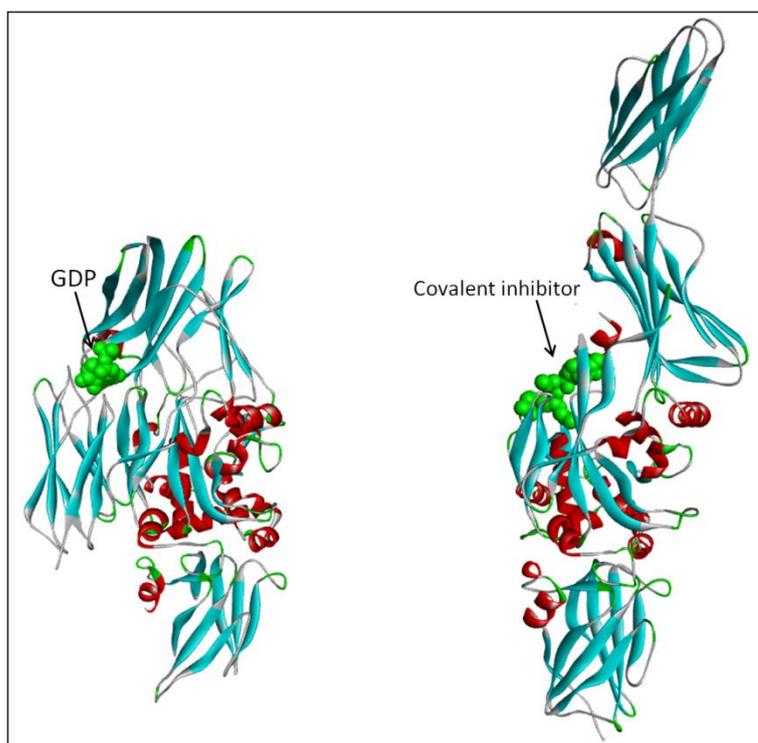
The behaviour of TG2 during this simulation is discussed in detail in a separate section of this chapter (3.9). For the purposes of this section of the work, it will be mentioned that the simulation was stable as indicated by the changes in the temperature and energy of the system with time as well as the change in the root-mean-square-deviation (RMSD) which are all presented in section 3.9.

### 3.1.1 TG2 active site

The crystal structure that was used as the basis for this work (PDB 2Q3Z) had TG2 in an “open conformation”, in contrast to a GDP-bound “closed conformation” (PDB code 1KV3 (Liu et al. 2002)). In 1KV3, the two C-terminal barrels are bent towards the active site giving the “closed” shape (Figure 3-1, (Pinkas et al. 2007)). The crystal structure, 2Q3Z has a gluten-like pentapeptide inhibitor, Ac-P(DON)LPF-NH<sub>2</sub> covalently bound to CYS277.

The catalytic cysteine residue (CYS277) is located in a tunnel (Figure 3-2) bridged by two tryptophan residues (TRP241 and TRP332) and there is a threonine residue (THR360) at the edge of the tunnel. The inhibitor in the crystal structure exhibited a network of interactions with active site residues including hydrogen bonds and hydrophobic interactions. The hydrogen bonds were between the inhibitor and residues TRP241, GLN276, CYS277 and ASN333. In addition, there was the covalent bond with CYS277 (Figure 3-3).

As it can be seen from Figure 3-3, the inhibitor assumes a bent conformation in the active site with the warhead in the catalytic tunnel and the other end, which is mainly aromatic, embedded in a hydrophobic region or loop (Figure 3-2) (Badarau, Mongeot, et al. 2013; Pinkas et al. 2007). This hydrophobic region is mainly composed of the following residues; ALA304, LEU312, ILE313, PHE316, ILE331 and LEU420. Four of these residues, ALA304, ILE313, ILE331 and LEU420, form pi-alkyl interactions with the terminal phenyl ring of the inhibitor stabilising this part in the hydrophobic region (Figure 3-4).



**Figure 3-1: TG2 closed and open conformations. The closed conformation bound to GDP (left), and the open conformation bound to a covalent inhibitor (right). Adopted from Pinkas et al. (2007).**

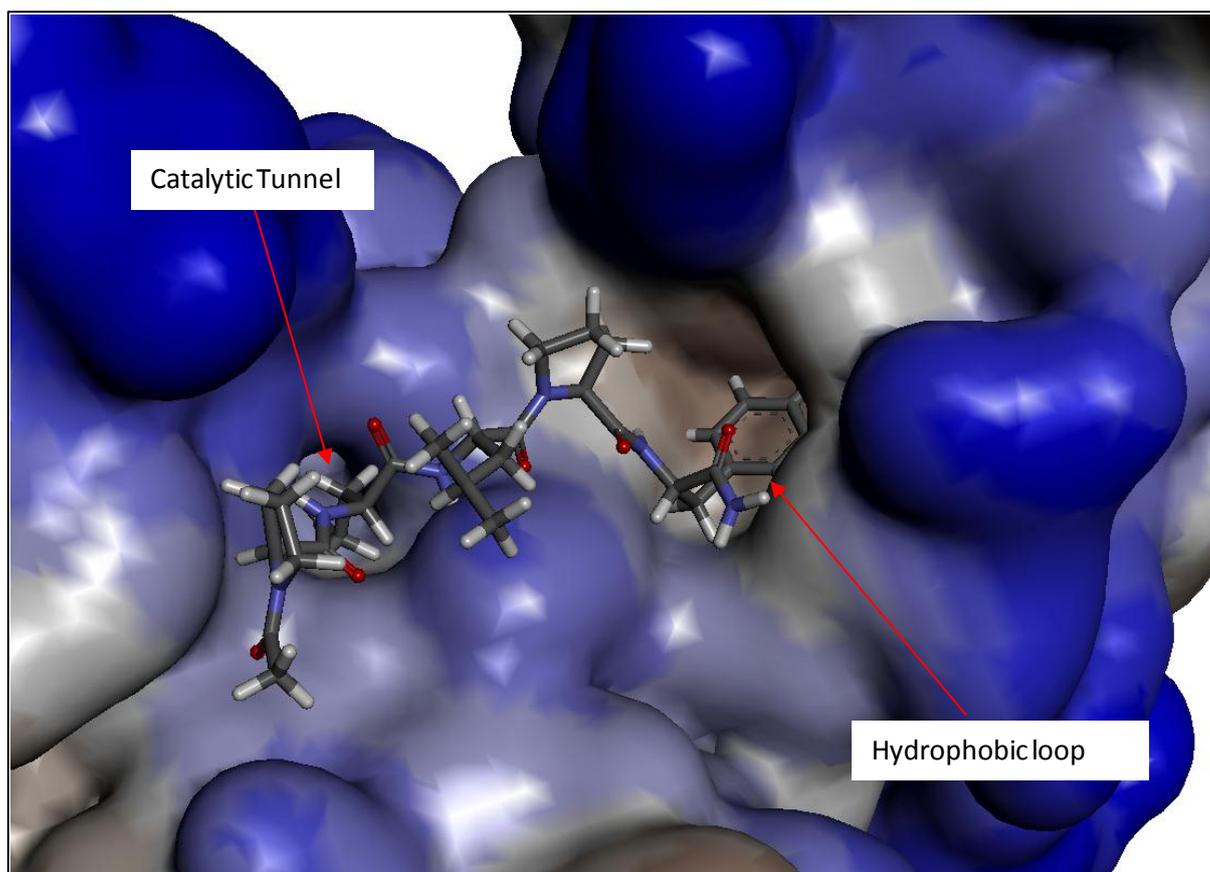


Figure 3-2: Active site tunnel and the hydrophobic loop in TG2. The surface is coloured by hydrophobicity and the inhibitor by element.

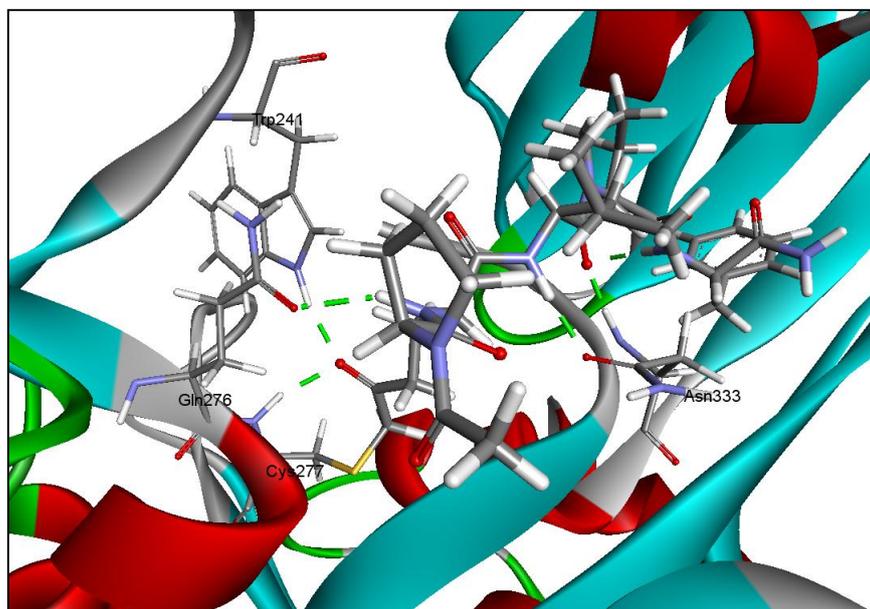
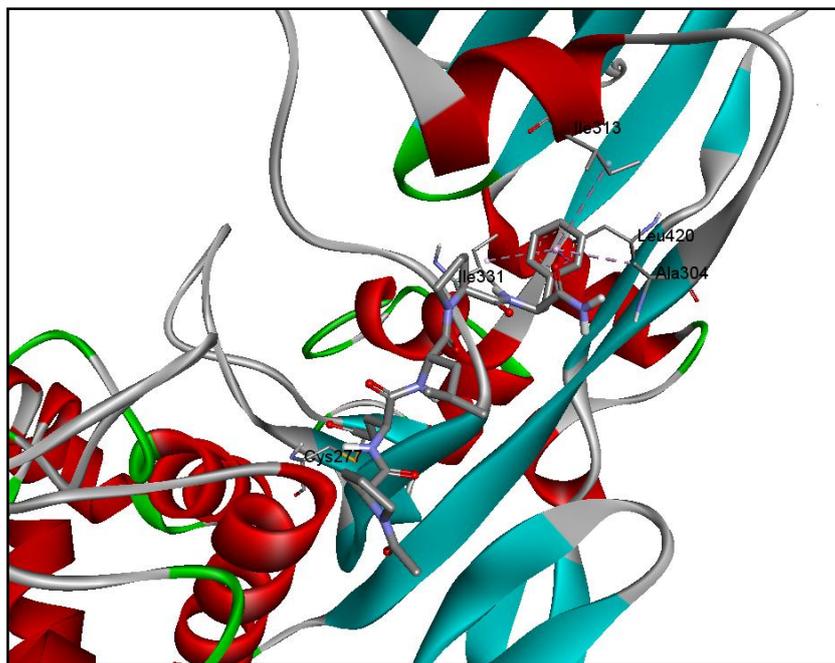


Figure 3-3: H-bonds between the inhibitor and TG2 in the original crystal structure.

The residues selected to be treated as flexible when using the GOLD docking program were based on the above information (residues involved in hydrogen bonds and lipophilic interactions with the inhibitor). These were TRP241, GLN276, CYS277, TRP278, TRP332, ASN333, PHE334, HIS335, THR360 and LEU420. Hydrogen bonds will be shown as green dotted lines and pi interactions as dotted light pink (pi-alkyl), dotted dark pink (pi-pi) or dotted

orange (pi-cation) lines. This summary of the interactions was introduced to be used as a guidance during the analysis of the subsequent docking and MD trajectories.



**Figure 3-4: Hydrophobic interactions between the inhibitor and TG2.**

### 3.1.2 Initial docking

Six active TG2 inhibitors (Badarau et al. 2015) were used for initial docking. Compounds **1a-1f** are listed in Table 3-1 along with their structures and  $IC_{50}$  values against human TG2. The goal at this stage was to obtain good poses for the active inhibitors within the active site of TG2 for the complexes to be taken to MD simulation. Therefore, a distance of less than 4 Å between the electrophilic carbon of the compound's warhead and the sulphur atom of CYS277 was used as the sole criterion for selecting a good complex at this stage.

Table 3-2 shows the six complexes selected with the distance between the warhead and the sulphur atom of CYS277, and the docking software used. The complexes were named in the form [inhibitor name-model name (ns at which the model was taken)-docking attempt].

Figure 3-5 shows the poses of the selected complexes. It has been mentioned earlier that a good pose for an inhibitor in the TG2 active site is in the form of a bent conformation with the lipophilic part of the inhibitor embedded in the hydrophobic loop of TG2 (Pinkas et al. 2007; Badarau, Mongeot, et al. 2013). However, this could not be achieved with all the compounds at this stage. Consequently, the selected complexes for compounds **1c**, **1d** and **1f** had the lipophilic part of the inhibitor outside the hydrophobic loop.

Neither docking programs nor the ligands could be blamed for the docking results obtained at this stage. The settings used during the dockings ensured that a most extensive search

was performed combined with flexible treatment of (the most important) active site residues. The ligands used were proven to be active and their  $IC_{50}$  values show that they are amongst the most active TG2 inhibitors (Badarau et al. 2015), and their structures, although sharing a similar core, have different lipophilic parts and two different warheads. This leaves two possible causes for the failure; either the active site itself is not fit to accommodate the ligands, or the fact that the inhibitors are irreversible and act through forming a covalent bond with CYS277, and the formation of the bond, although unlikely, may need to be the first event for the compound to position itself correctly within the active site of TG2. Molecular dynamics applied to these 6 docking complexes was a possible solution to the first cause of failure in docking at this stage by allowing the active site residues to move freely with the ligand within to accommodate it better by changing the conformations of the side chains or even the backbone structure to allow for better positioning of the ligands. Dealing with the bond is more complicated and requires quantum mechanical treatment or covalent docking. These techniques have been applied during the course of this work and are presented in the relevant Results sections.

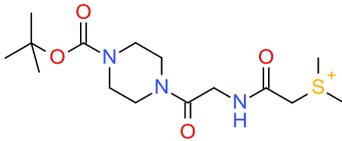
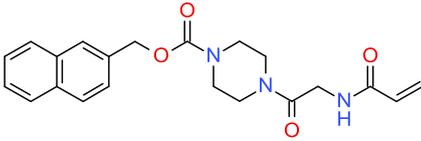
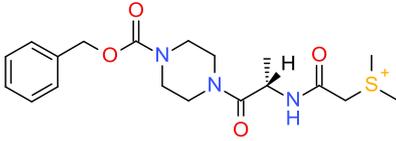
Compound name	Structure	$IC_{50}$ ( $\mu M$ )
1a		0.125
1b		0.273
1c		0.440
1d		0.700
1e		0.0061
1f		0.380

Table 3-1: Structures and  $IC_{50}$  for compounds 1a-1f.

Complex Name	Distance (Å)	Docking Program
<b>1a</b> -85ns-3	3.4	CACHe
<b>1b</b> -30ns-2	3.3	CACHe
<b>1c</b> -30ns-14	3.7	GOLD
<b>1d</b> -95ns-9	3.7	GOLD
<b>1e</b> -40ns-10	3.0	GOLD
<b>1f</b> -80ns-6	5.6	GOLD

**Table 3-2: Good complexes from initial docking.**

When the active site of TG2 to be used for docking was investigated, different centres for the sphere used in the definition were tried. These included the sulphur atom of CYS277, the whole CYS277 residue, each of the bridging tryptophan residues (TRP241 and TRP332) on its own and finally both the bridging residues at the same time. With each one of these centres, an 8-Å sphere of residues was used as the active site and a test docking was performed. The best results were obtained with the last definition, signifying the importance of these bridging residues in correctly posing TG2 inhibitors in the active site. That is why this active site definition was used throughout this work.

## 3.2 Analysis of MD Trajectories of Initial Docking

The main reason for performing MD simulations on docked complexes of TG2 inhibitors is to obtain better TG2 active site models. Taking 19 conformations of TG2 from a 100-ns simulation to be used for docking of 6 proven active inhibitors produced only 3 valid complexes with the correct proposed conformation of the inhibitor and 3 more complexes were assumed to be correct. Therefore, MD was applied to those complexes in an attempt to refine the active site by allowing for the inhibitors to adjust themselves correctly within the active site and at the same time permitting the active site residues to accommodate the inhibitors. Only the trajectories of compounds **1a** and **1b** will be discussed in details as they gave valid models in the subsequent docking. The other 4 trajectories will be discussed briefly.

### 3.2.1 Trajectory of 1a-85ns-3

In the starting conformation for this run, there were no hydrogen bonds between **1a** and the protein. However, there were pi-interactions between the residues in the hydrophobic loop and the adamantyl part of the ligand and a pi-interaction involving TRP332 and the piperazine ring of **1a** (Figure 3-6). Also in the starting conformation, the bridging tryptophan residues were very close to each other that the tunnel was almost closed (Figure 3-7).

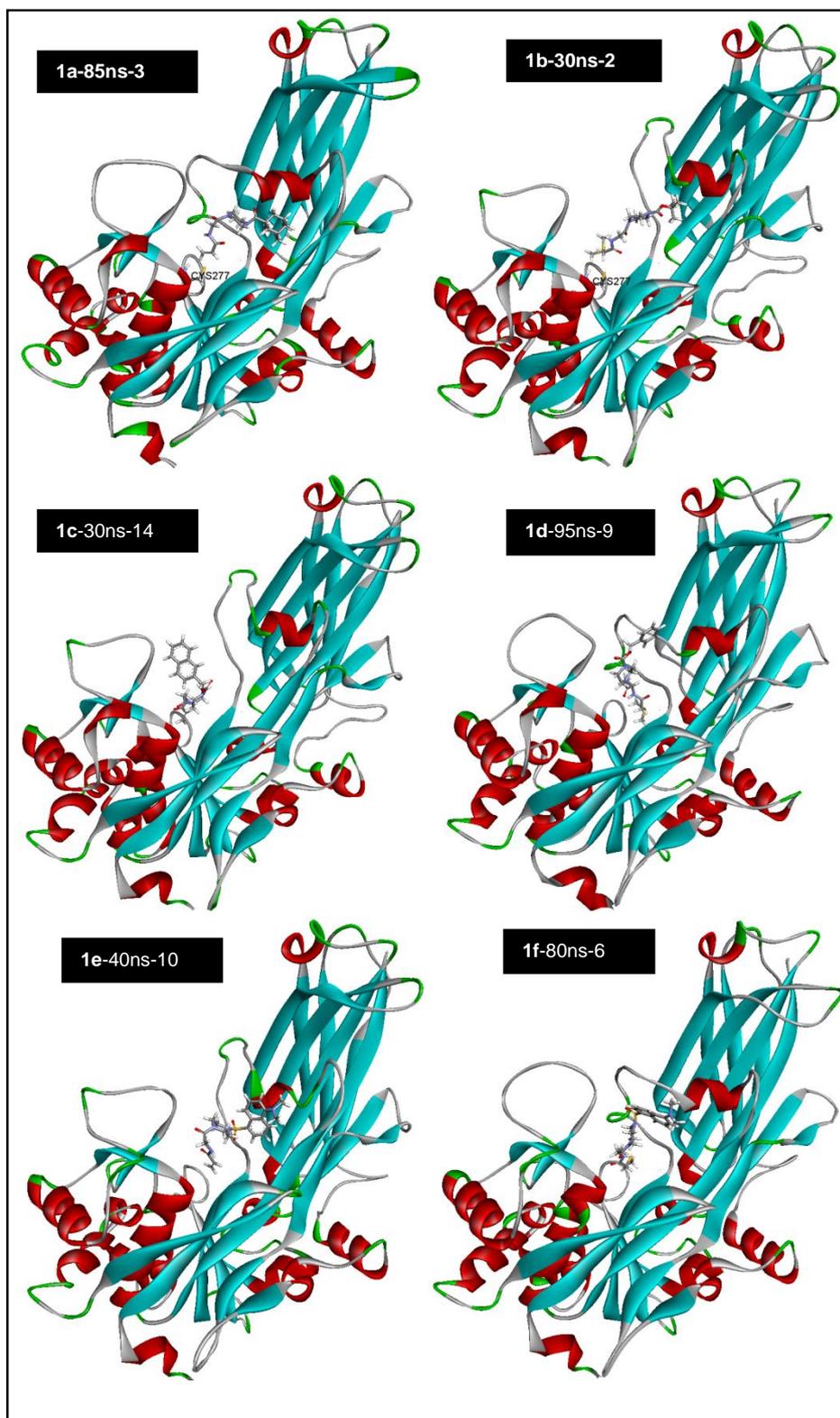


Figure 3-5: Docking complexes for the 6 compounds selected to be taken to molecular dynamics.

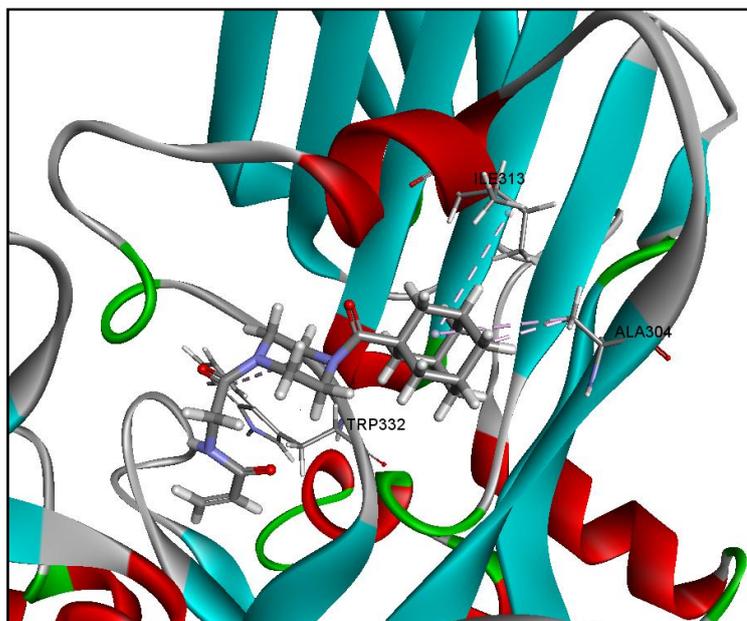


Figure 3-6: Hydrophobic interactions between TG2 and 1a.

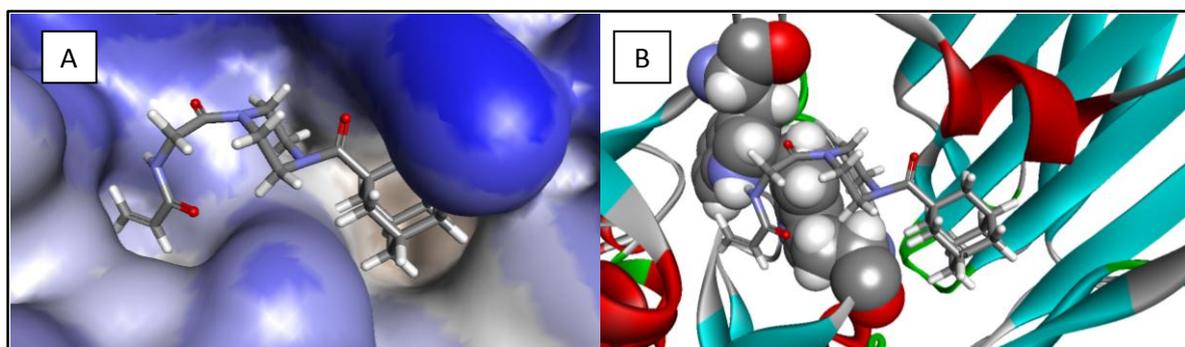


Figure 3-7: The tunnel in the starting conformation of 1a trajectory. A: Surface view. B: space filling view of the bridging tryptophans.

### 3.2.1.1 Warhead position

The acrylamide warhead of compound **1a** left the tunnel and moved away from CYS277 early in the simulation. This happened at 12.475 nanoseconds. In the frame immediately before this time point, the warhead was about 7.6 Å away from CYS277 sulphur atom but was essentially within the tunnel. The distance then started to increase to more than 20 Å (Figure 3-8). It has also been noted that at the same frame at which the warhead left, CYS277 has moved towards the bridging tryptophan residues causing the tunnel to close. The movement of the warhead at this time was also evident from the RMSD measured for **1a** alone during the trajectory and is shown along with the distance in Figure 3-8. By comparing the 2 traces, it is obvious that both have started to increase at the same time.

### 3.2.1.2 Hydrogen bonds

VMD program was used for hydrogen bond analysis using the default settings; 3.0 Å donor-acceptor distance and 20° angle cut-off. Very few hydrogen bonds were found during this

simulation, but there was a bond with ASN333 at the beginning that lasted to approximately the same time as the presence of the warhead within the catalytic tunnel. Figure 3-9 shows a graph for the distance between O38 of **1a** and H of ASN333 along with the bond itself. Both the closure of the tunnel and the disappearance of the H-bond with ASN333 may have contributed to the movement of the warhead of **1a** away from the catalytic tunnel.

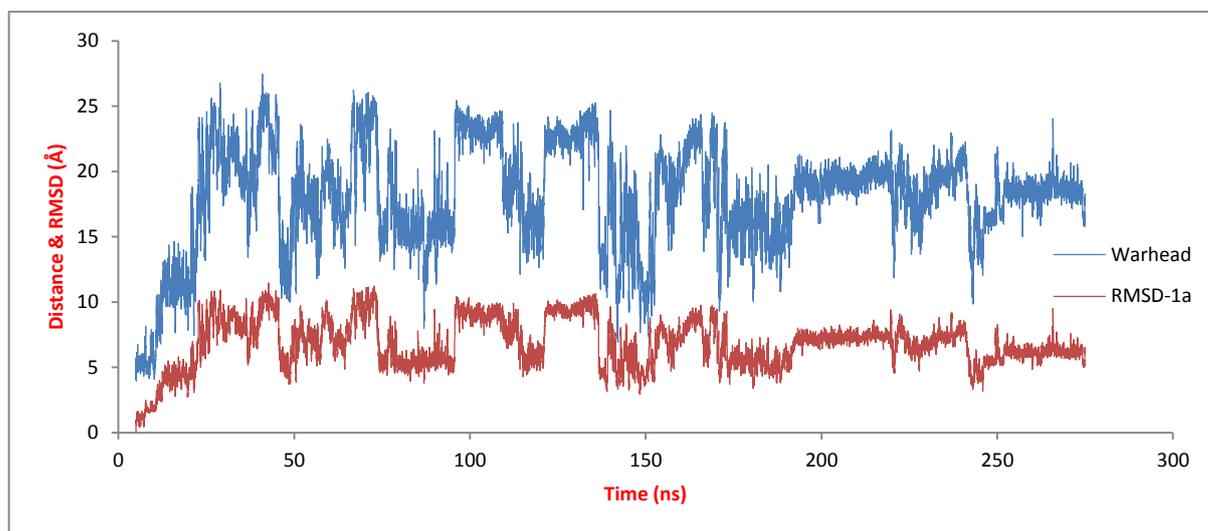


Figure 3-8: Graph of the distance between the acrylamide warhead and SG atom of CYS277 in the **1a** trajectory along with RMSD for **1a** alone.

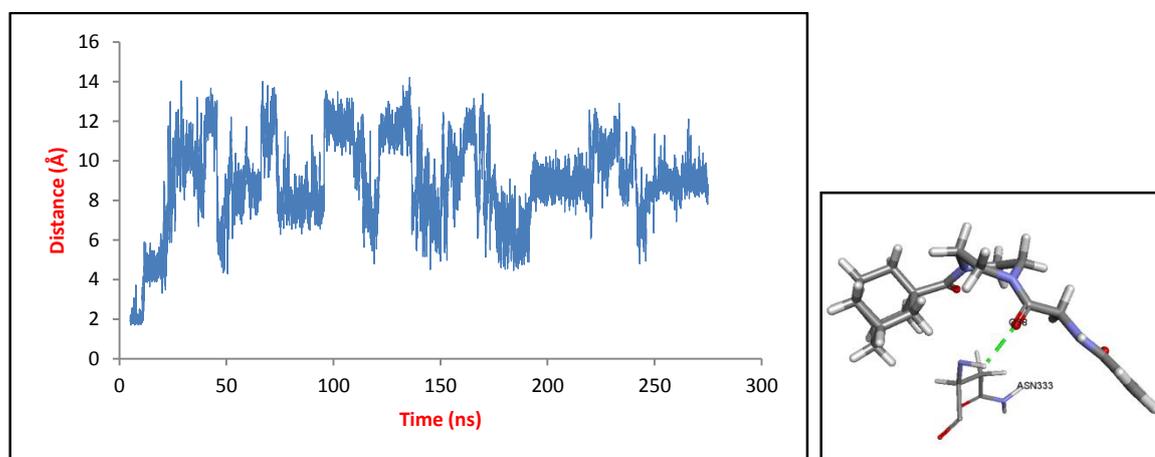


Figure 3-9: Distance of the H-bond between ASN333 and **1a** (left) and the bond itself (right).

### 3.2.1.3 The lipophilic part

The distance between C43 of **1a** and alpha carbon of SER309 (Figure 3-10 right) was used as a reference for the position of the lipophilic part. SER309 was chosen as a reference because of its position at the top of the hydrophobic loop, where it oversees the lipophilic part of **1a**. Figure 3-10 shows a graph of this distance with the position of the lipophilic part within the hydrophobic region. In the starting conformation, the distance was 13.6 Å and from the graph it is clear that the adamantyl group did not move that much during the simulation. The reason behind this is the persistence of the hydrophobic interactions with the

adamntayl part during the simulation. An analysis of 9 conformations taken at 30 ns intervals from the entire simulation revealed that there were at least two residues forming pi-alkyl interactions with the adamntayl group of **1a** in each conformation. Figure 3-11 shows one of these conformations.

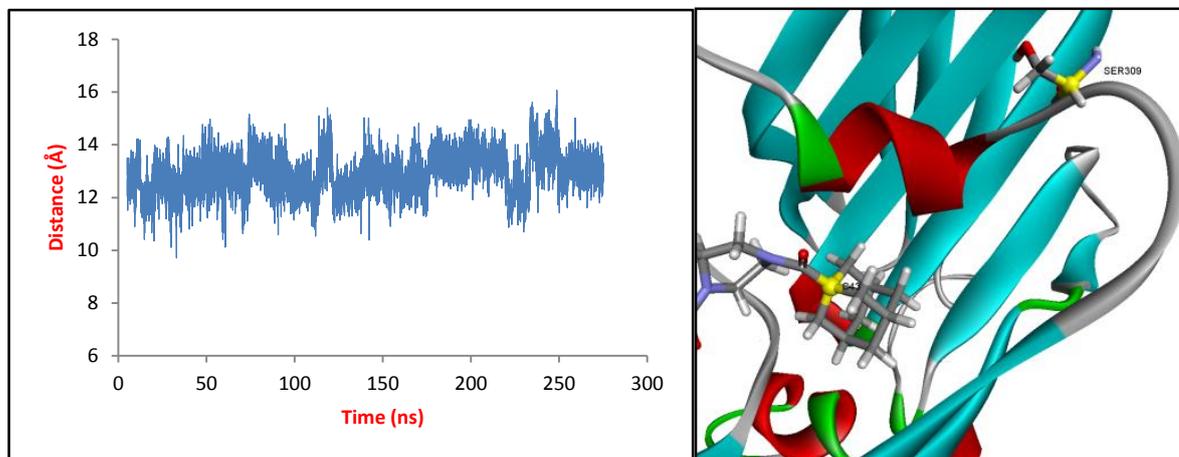


Figure 3-10: Distance between the lipophilic part and SER309 during the simulation.

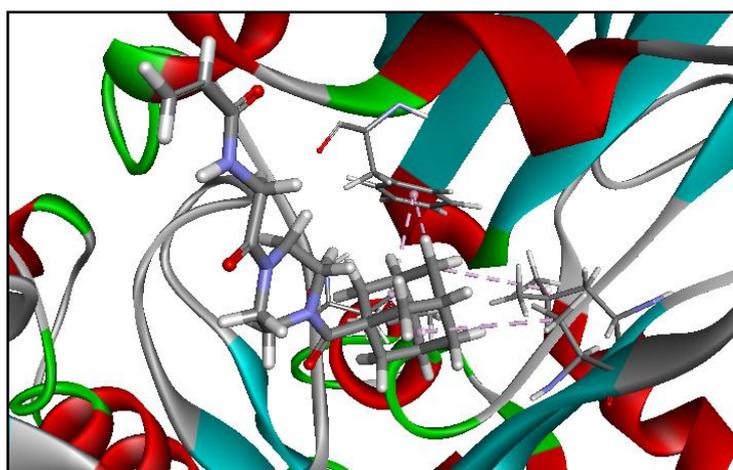


Figure 3-11: Hydrophobic interactions stabilising the adamantyl group in the hydrophobic loop in **1a** trajectory, even following the leaving of the warhead. The frame is from 200 ns of the trajectory.

### 3.2.1.4 TG2

The presence of **1a** did not trigger noticeable conformational change on the time scale of the simulation. This was observed from measuring the RMSD for the protein and comparing it to a measurement done on a trajectory of the same length for the MD simulation on empty TG2. Figure 3-12 shows the 2 RMSD measurements and indicates that no large differences occurred upon the introduction of **1a** into TG2.

The atomic fluctuation scores for TG2 residues in this simulation are presented in Figure 3-13. Only the peak with red points has residues that are close to the active site. The peak belongs to a helix and the high fluctuation is the result of coiling and uncoiling events. The helix is shown in Figure 3-14. It was missing in the original crystal structure. The helix was

not very stable possibly due to the presence of 2 glutamate residues (GLU319 and GLU322) that may be repelling each other causing the un-coiling of the helix. There is also a glycine residue in this helix which allows greater flexibility (Lehninger et al. 2000).

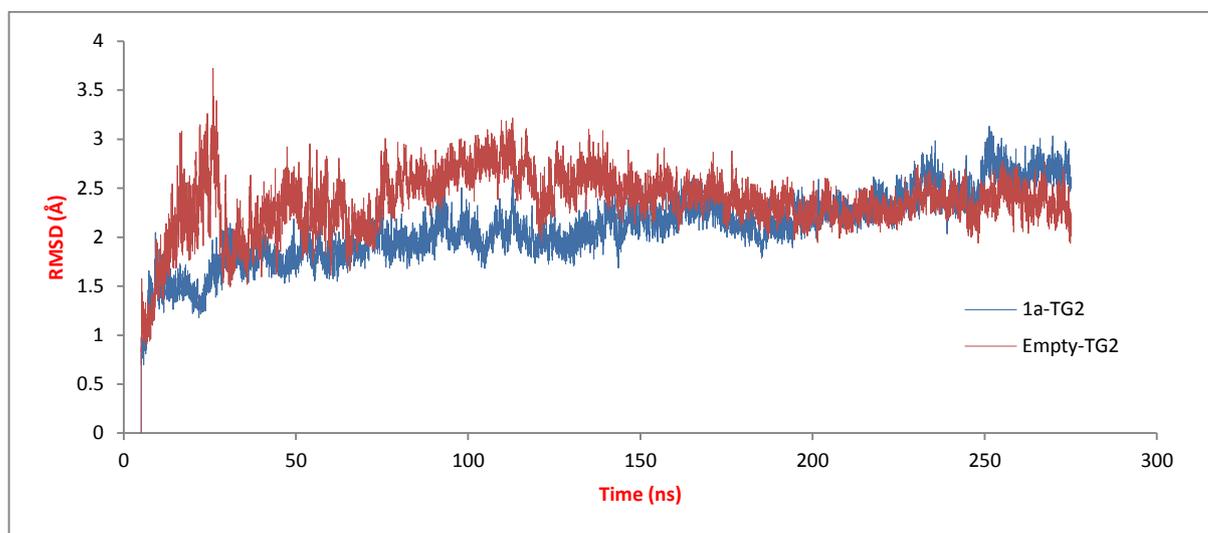


Figure 3-12: RMSD for TG2 in the presence and absence of compound 1a.

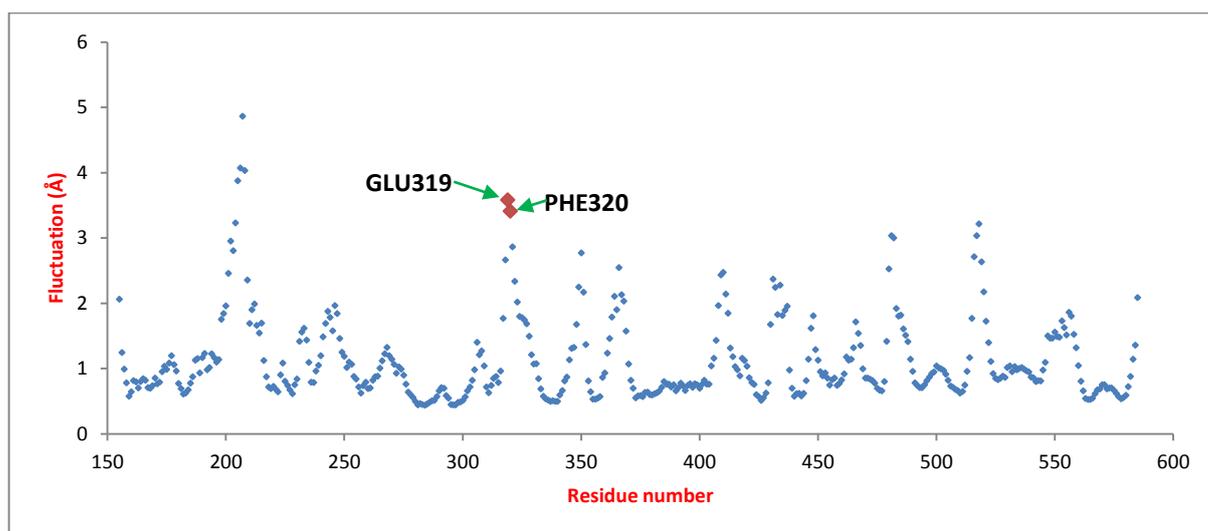


Figure 3-13: Atomic fluctuation for TG2 residues in the trajectory of 1a.



Figure 3-14: The high fluctuating helix from 1a trajectory, coloured red, while the rest of TG2 is green and showing CYS277 to appreciate the location of the helix.

### 3.2.1.5 Temperature and energy

The stability of the simulation was reflected in the changes in the temperature and total energy of the system. These two quantities were almost constant throughout the simulation after the initial relaxation and are shown in Figure 3-15.

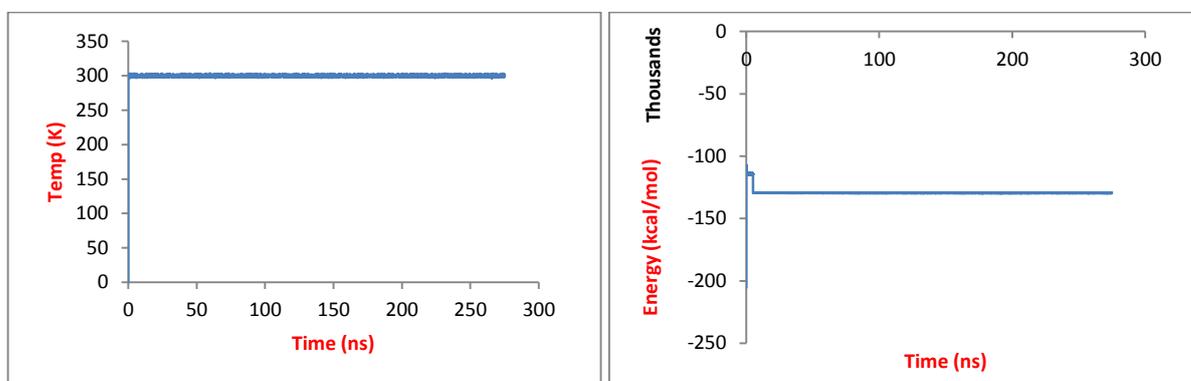


Figure 3-15: Changes in temperature and total energy during the simulation of 1a.

## 3.2.2 Trajectory of 1b-30ns-2

In the starting conformation, there was a hydrogen bond between ASN333 side chain and **1b** and the distance between SG of CYS277 and the electrophilic carbon of **1b** was 3.3 Å. There were no hydrophobic interactions involving the lipophilic part of **1b**.

### 3.2.2.1 Warhead position

The warhead of **1b** remained in the tunnel close to the sulphur atom of CYS277 for the entire simulation. The distance between the two averaged at 5.7 Å. Figure 3-16 shows a plot of this distance with time. The graph also shows the RMSD for **1b** in this simulation, and its low values also indicate that there was no significant motion involving compound **1b** during the simulation.

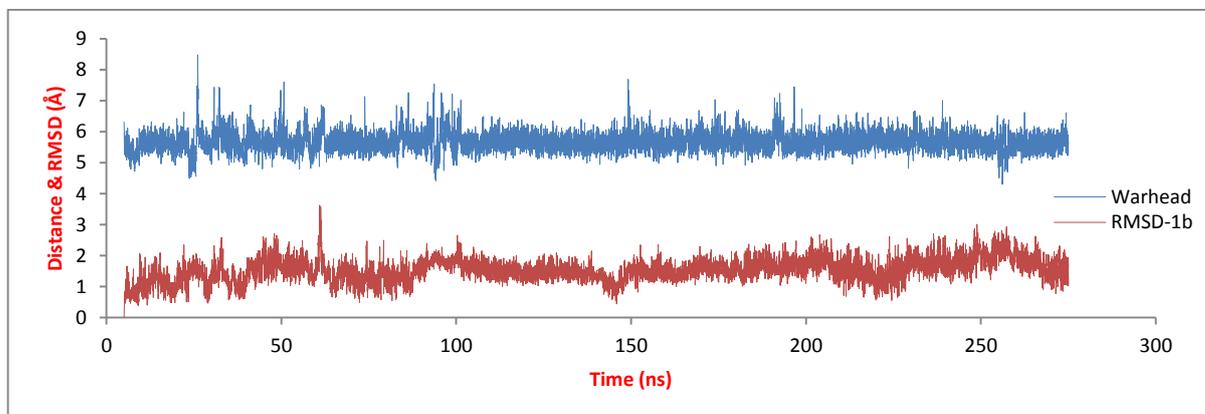


Figure 3-16: A graph of the distance between the warhead carbon of 1b and SG of CYS277, with the RMSD for 1b during the MD simulation.

### 3.2.2.2 Hydrogen bonds

Again, default VMD settings were used for hydrogen bond analysis and hydrogen bonds with 3 residues were found, ARG317 (side chain), ASN333 (side chain and backbone) and PHE334 (backbone) (Figure 3-17). In total, bonds with these residues were present for about 33% of simulation time. A deeper look into the time-dependent behaviour of these bonds showed that the ones with ASN333 and PHE334 (which were present for 28% of the simulation according to VMD) were probably responsible for keeping the warhead close to CYS277. When the hydrogen bond with ARG317 forms, a pocket in TG2 is formed that closes on the lipophilic part of **1b** (Figure 3-18). This bond persisted for 7% of the simulation time and was most evident between 75 and 160 nanoseconds of the simulation.

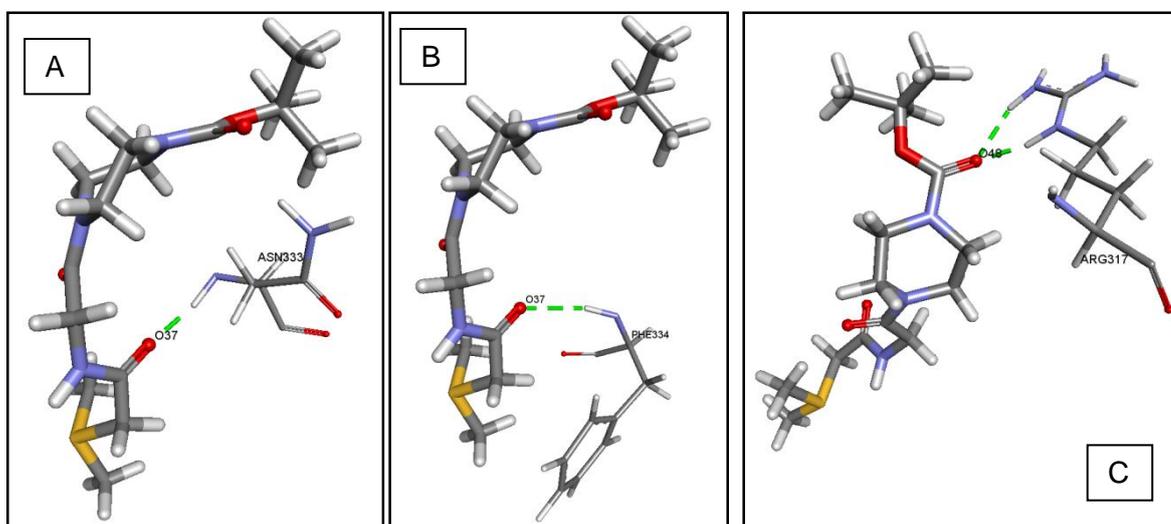


Figure 3-17: Hydrogen bonds in **1b** trajectory with ASN333 (A), PHE334 (B) and ARG317 (C) showing the acceptor atoms on **1b**.

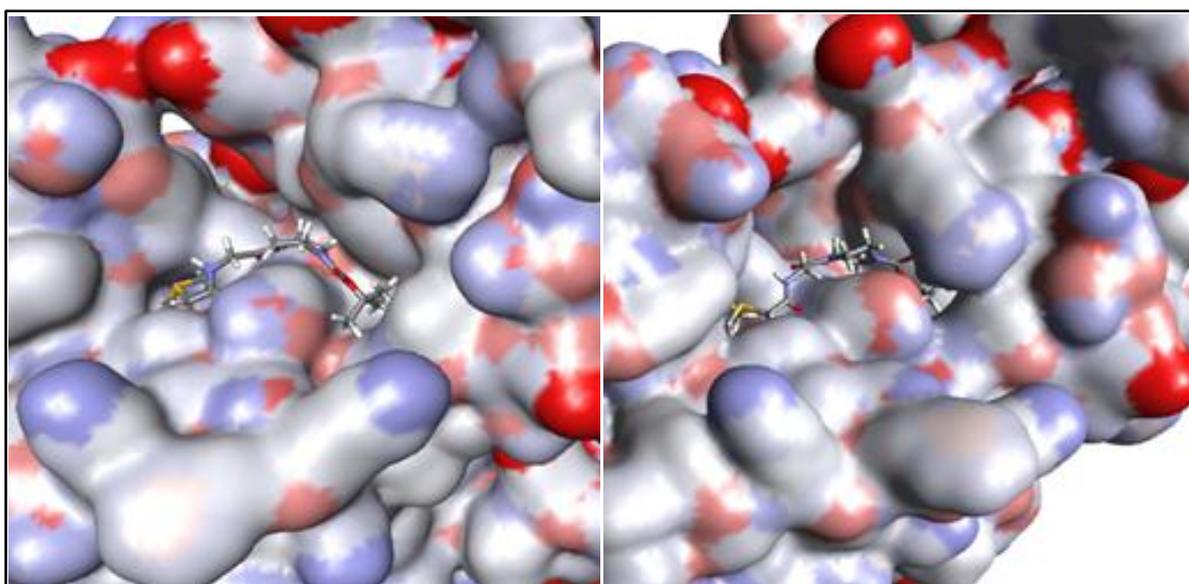
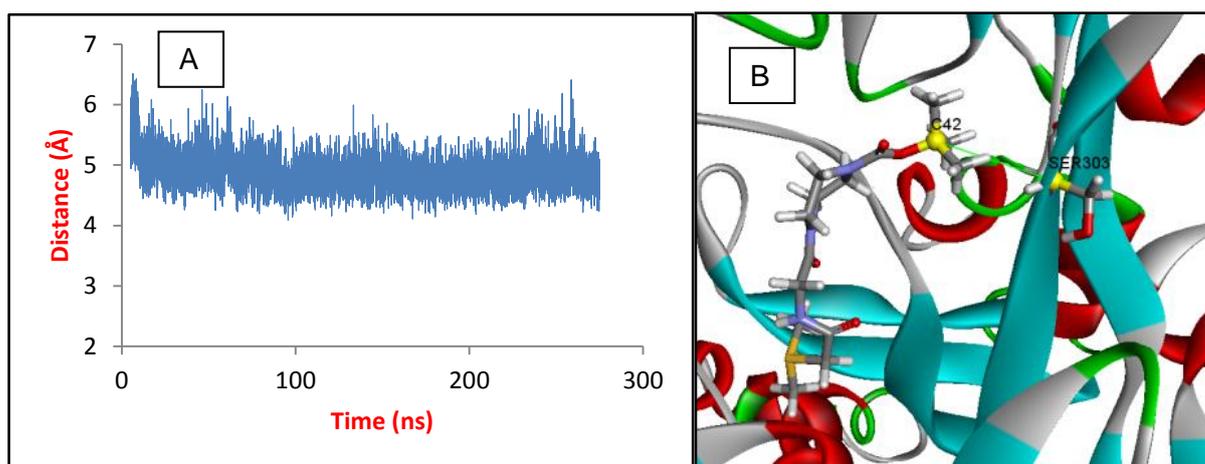


Figure 3-18: Pocket formed when there is H-bonding with ARG317 (right) and absence of the pocket when the interaction is not present (left).

### 3.2.2.3 The lipophilic part

The reference atom of **1b** that was used as an indication for the position of the tertiary butyl lipophilic part of **1b** within the hydrophobic region was C42 (Figure 3-19B). SER303 was used instead of SER309 as the counter protein reference part because SER309 showed more movement during the simulation than SER303, and thus it could not be used as a reference. This is evident from the atomic fluctuation values for the two residues; SER303 had a value of 0.77 Å while SER309 showed atomic fluctuation of 1.73 Å.

In the starting conformation, the distance between C42 of **1b** and CA of SER303 was 7.8 Å. In the simulation, it averaged around 4.9 Å. Figure 3-19A shows a graph of the change of this distance with time during the simulation and it can be seen that the distance has decreased after the start of the simulation indicating that the tertiary butyl group moved more inside the hydrophobic loop during the simulation. Figure 3-20 shows a comparison between the position of the tertiary butyl group between the starting conformation (A) and conformation at 65 nanoseconds (B).



**Figure 3-19: A: distance between C42 of **1b** and SER303 in the **1b** trajectory. B: the same distance on a PDB structure.**

An analysis of 9 conformations taken at 30 nanoseconds intervals showed at least 2 hydrophobic interactions between residues from the hydrophobic loop within TG2 and at least one of the methyl groups of the tertiary butyl lipophilic part of **1b** in 8 of the conformations. These interactions together with the hydrogen bond with ARG317 and the subsequent formation of the closed pocket around the tertiary butyl group are the likely reasons behind the persistence of the lipophilic part within the hydrophobic loop (Figure 3-21).

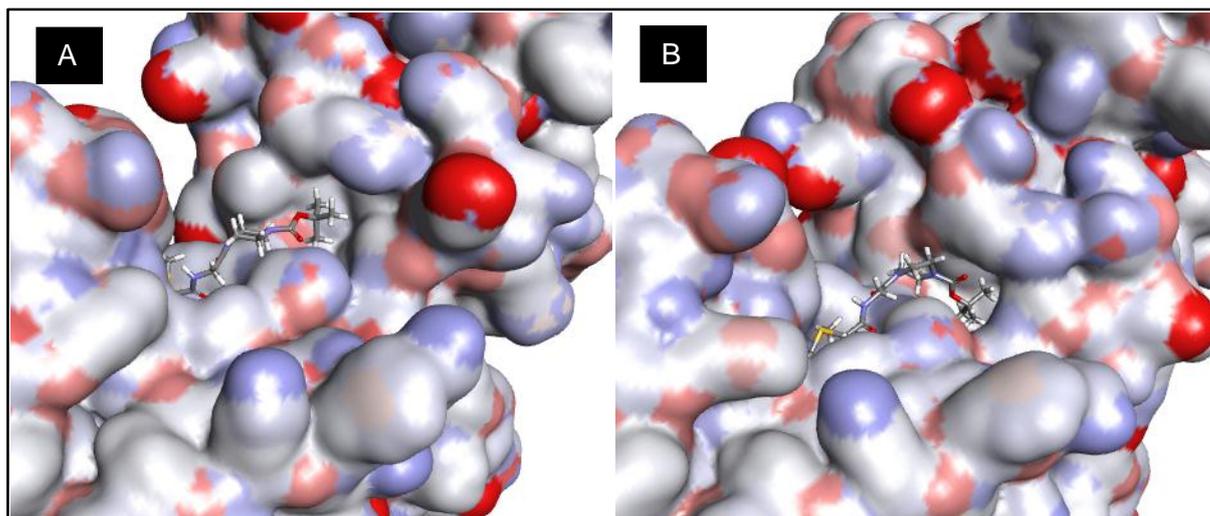


Figure 3-20: How the tertiary butyl group of **1b** moved further into the hydrophobic loop from the starting conformation (A) to the conformation at 65 ns (B).

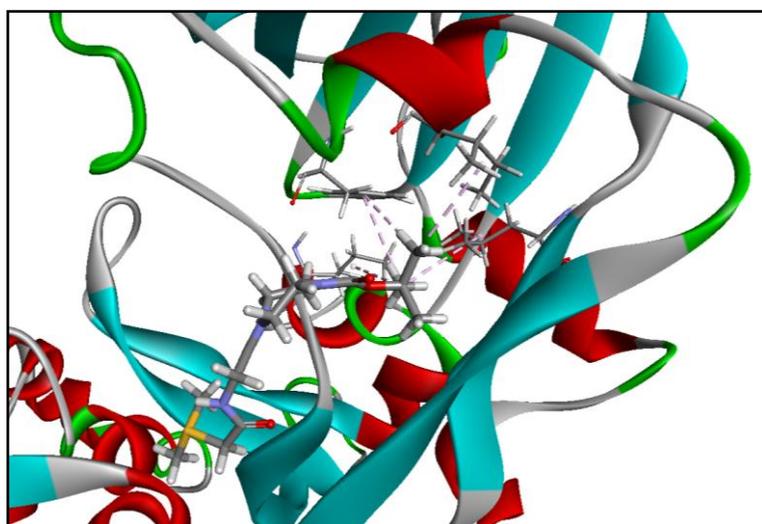


Figure 3-21: Example of hydrophobic interactions involving the tertiary butyl part of **1b** in the hydrophobic loop of TG2.

Pi-cation interactions were noticed to exist, which involved the positively charged sulfonium ion of **1b** warhead. The most prevalent of these was the one with the benzene ring PHE334. Because the VMD program cannot measure the distance between an atom and the centre of a ring, distances were measured between the positively charged sulphur atom of **1b** and C1 of the benzene ring of PHE334 (CZ) and then with C4 (CG) of the same ring and an average was calculated (Figure 3-22). Although this could not be used as an ultimate judgement for the presence of the bond as it does not take into account the orientation of sulphur atom with respect to the ring, it does give an insight on the possibility of the formation of this interaction based on distance.

The distance between the positively charged sulphur of **1b** and the benzene ring of PHE334 averaged around 5.1 Å (Figure 3-23). This bond with PHE334 was present in 7 of the 9 conformations taken at every 30 nanoseconds. In the remaining 2 conformations, there was instead a pi-sulphur interaction involving TRP241 and/or TRP332 (Figure 3-24). Therefore, it

can be said that there was always a pi-sulphur interaction that, along with hydrogen bonds with ASN333 and PHE334, has contributed to the stabilisation of the warhead within the catalytic tunnel. Ringer et al. (2007) studied the optimum conditions for favourable pi-sulphur interactions by mining in the data of the Protein Data Bank. Their observations were dependent on the orientation of hydrogen atoms attached to the sulphur and in one of the favourable configurations, they found that the optimum distance between the sulphur atom and the centre of the benzene ring is 5.5 Å. This further confirms the persistence of the interaction with PHE334 based on the graph in Figure 3-23.

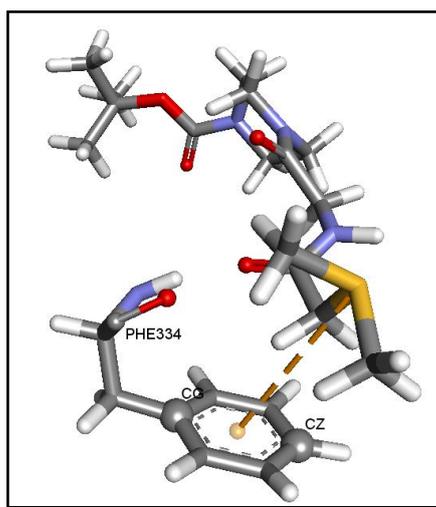


Figure 3-22: Pi-cation interaction between **1b** and PHE334, showing CG and CZ atoms used to define a distance between **1b** and the benzene ring.

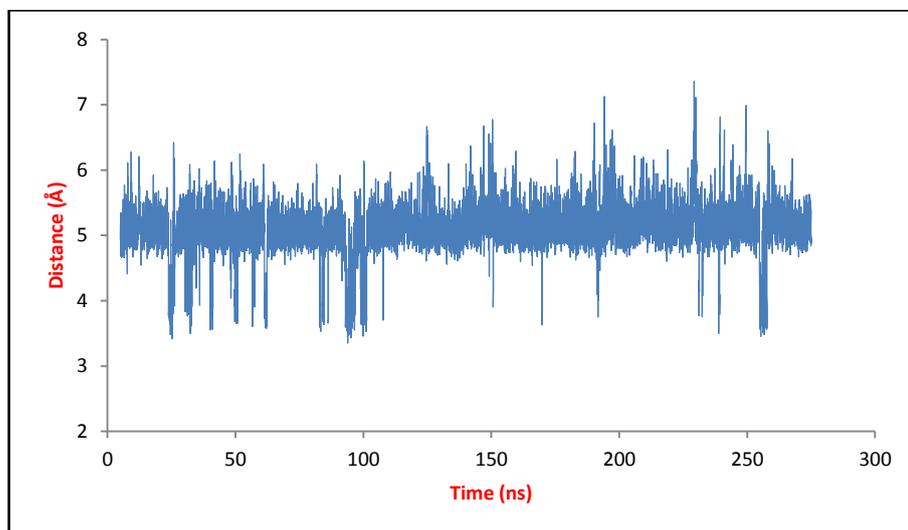


Figure 3-23: Plot of the distance between sulphur atom of **1b** and benzene ring of PHE180.

#### 3.2.2.4 TG2

As in **1a** trajectory, the presence of **1b** within TG2 active site in this simulation did not cause significant changes in the enzyme itself. This was confirmed by comparing the RMSD values of this run to those of the simulation applied on TG2 alone for the same simulation time (Figure 3-25). In fact, RMSD plots show that the presence of the compound stiffens the enzyme and limits its overall conformational changes. This also applies to compound **1a**

(Figure 3-12). This is consistent with the data that suggest that the covalent inhibitors of TG2 when bind to the enzyme lock it in its open conformation (Pinkas et al. 2007). Atomic fluctuation analysis did not show any significant peaks with all residues having values less than 3 Å, with the exception of 3 residues that were a part of a loop located on the other side of the active site.

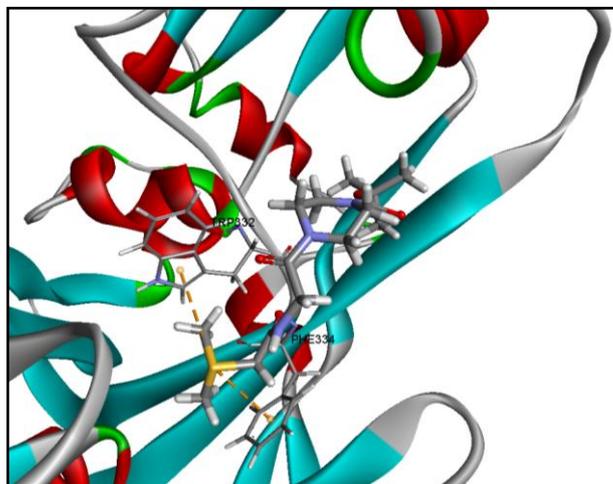


Figure 3-24: Example of pi-sulphur interactions involving TRP332 and PHE334 with the positively charged sulphur atom of 1b.

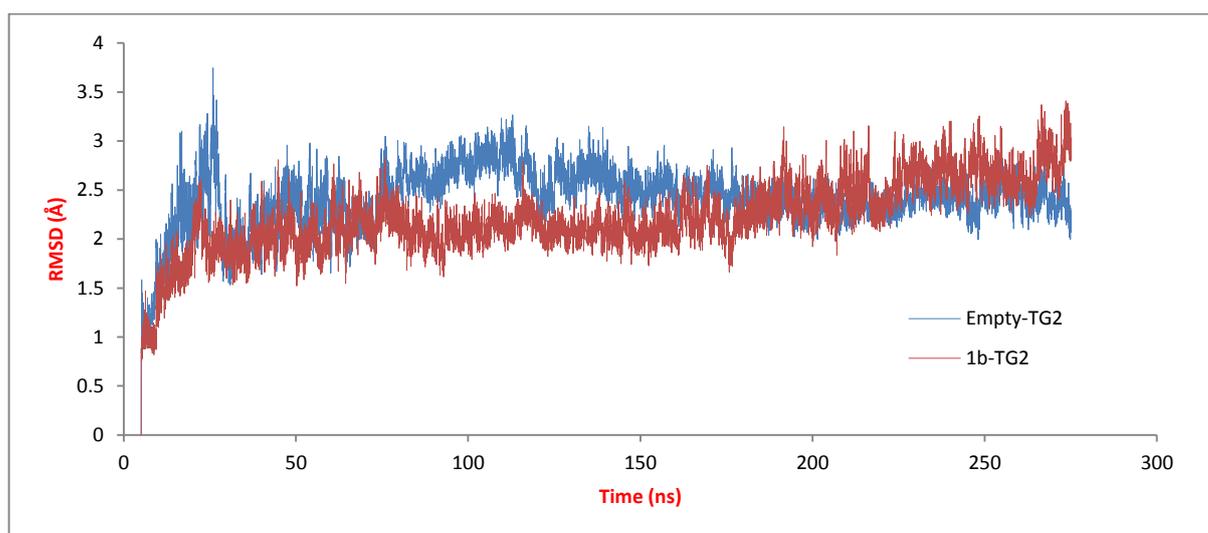


Figure 3-25: RMSD of TG2 alone in 1b trajectory compared to that applied on an empty TG2.

There was a loop, however, consisting of residues ASP306, GLN307, ASN308, SER309, ASN310 and LEU311, that showed a movement toward the back and inward to the hydrophobic region of the enzyme (Figure 3-26). The atomic fluctuation values for the 6 residues were all below 2.0 Å. The hydrophobic interactions with the tertiary butyl group of **1b** and hydrogen bonds between **1b** and ARG317 may be responsible for the movement of the loop and the subsequent enclosure of the hydrophobic loop for the tertiary butyl group of **1b**. It was therefore decided to add the residues of this loop to the definition of active site used in the docking experiments.

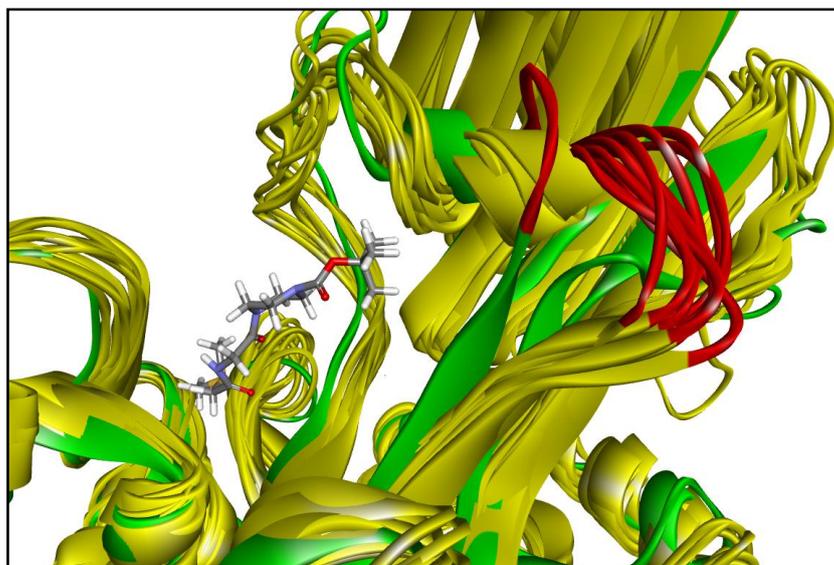


Figure 3-26: The loop between ASP152 and LEU157 (red) and how it moved from the starting conformation (green) compared to 9 conformations taken from the simulation at 30 ns intervals.

### 3.2.2.5 Temperature and total energy

Again in this trajectory the temperature and total energy during the simulation were stable as shown in Figure 3-27.

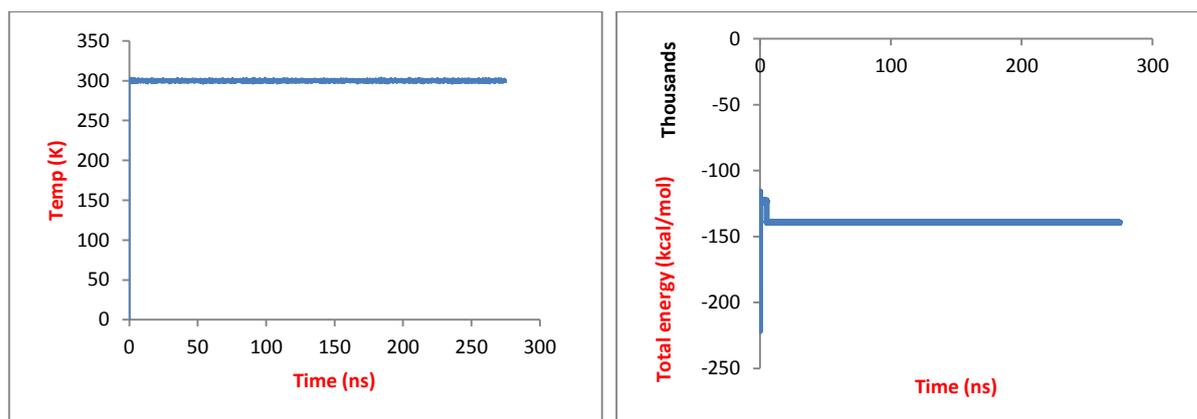


Figure 3-27: Temperature and total energy graphs for the trajectory of 1b.

### 3.2.3 Trajectory of 1c-30ns-14

In the starting conformation for this MD run, the distance between the warhead carbon atom of **1c** and SG atom of CYS277 was 3.7 Å. There were hydrogen bonds with CYS277 (backbone) and ASN333 (side chain). There were also pi-alkyl interactions between the piperazine ring of **1c** and TRP241, TRP332 and PHE334 (Figure 3-28). However, the naphthyl group, being the lipophilic part, was not in the hydrophobic region of TG2. No better pose for compound **1c** in TG2 active site could be obtained and therefore this one was used.

During the simulation, the lipophilic part of the inhibitor started to move in the direction of the hydrophobic loop of TG2 but it did not get fully inside the loop. This started to happen at

around 32.5 nanoseconds of the simulation. However, at approximately the same time, the warhead started to move away from CYS277 leaving the catalytic tunnel. The graph in Figure 3-29 shows the change in the warhead position relative to SG of CYS277. An analysis of hydrogen bonds showed that there were bonds at the beginning of the simulation up to the point when the warhead left the tunnel. The bonds were mainly with GLN276 and CYS277 (Figure 3-30). The disappearance of the bonds may be responsible for the warhead leaving the catalytic tunnel.

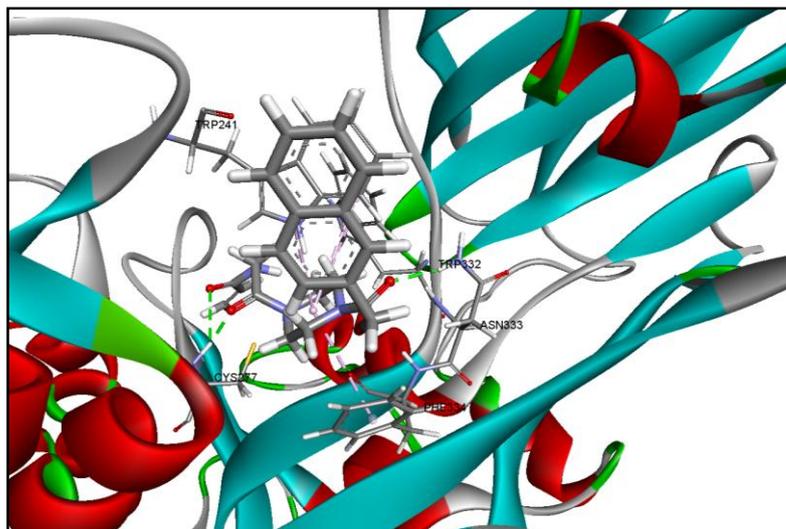


Figure 3-28: Interactions between 1c and TG2 active site in the starting conformation.

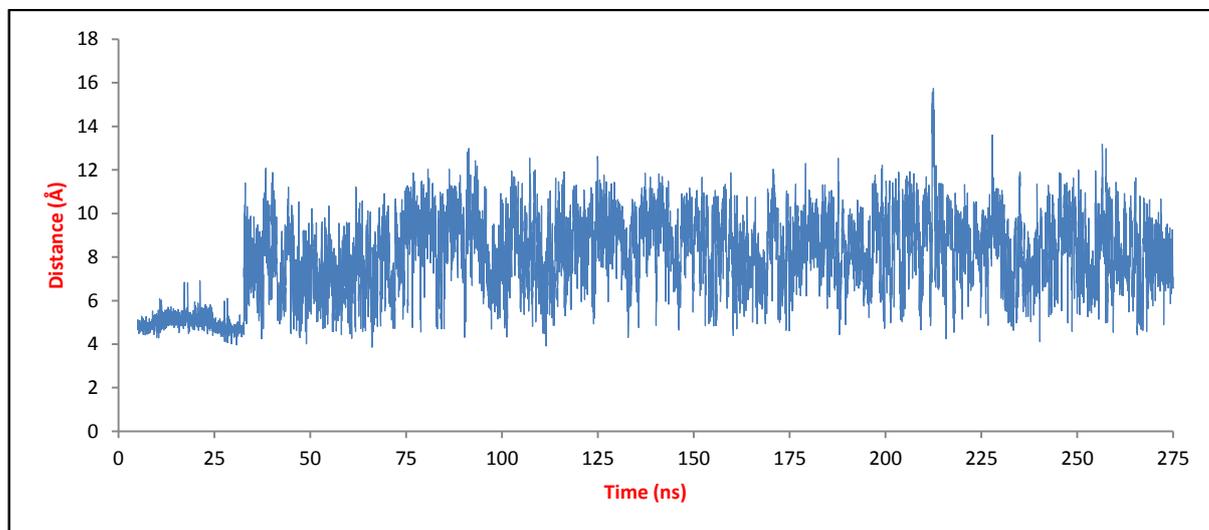


Figure 3-29: Time plot for the distance between warhead of 1c and sulphur atom of CYS123 showing the increased distance after 32 ns.

### 3.2.4 Trajectory of 1d-95ns-9

In the starting conformation, the distance between the warhead carbon atom and CYS277 SG was 3.7 Å. No hydrogen bonds were detected but there were some pi-interactions between TG2 residues and piperazine ring and sulphur atom of **1d**. As in the **1c** starting conformation, the lipophilic part of **1d** was not located in the hydrophobic loop of TG2. In the

simulation, the warhead of **1d** left the catalytic tunnel and moved away during the equilibration phase whereupon the distance between the warhead and the sulphur atom of CYS277 was 8.5 Å at the first conformation of the production phase. Therefore, this trajectory was considered a failure.

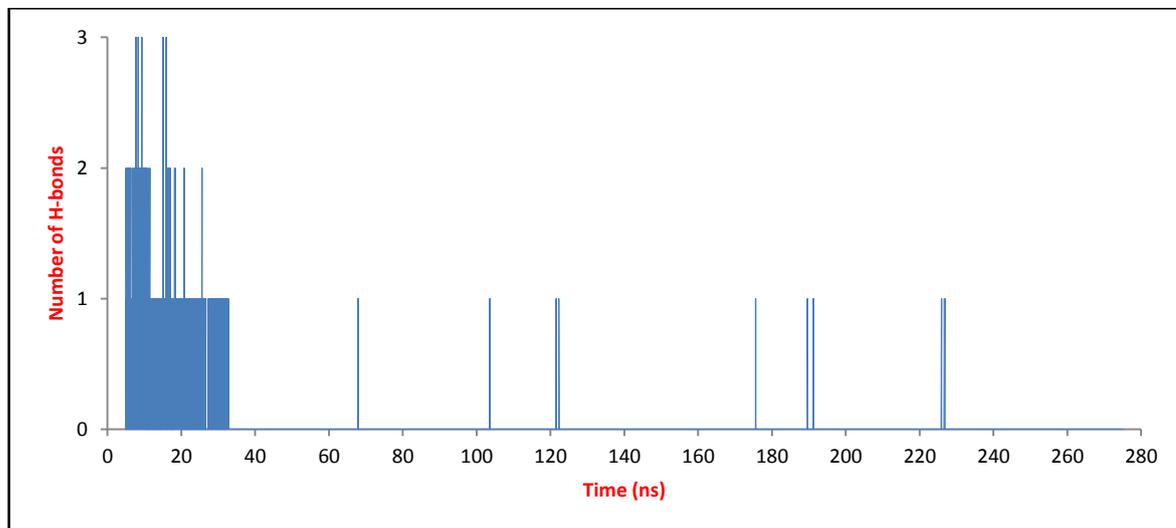


Figure 3-30: Number of hydrogen bonds with **1c** showing their disappearance after 32 ns.

### 3.2.5 Trajectory of **1e**-40ns-10

In the starting conformation, the distance between the warhead and CYS277 was about 3 Å. There was a single hydrogen bond with GLN276 side chain and a pi-alkyl interaction involving the adamantyl group of **1e** with ARG317 (Figure 3-31).

The warhead remained in the catalytic tunnel close to CYS277 for about 187 of the 275 nanoseconds of the simulation time (Figure 3-32). Analysing the hydrogen bonds in the trajectory showed that a bond with GLN276 was present up to the time at which the warhead left the tunnel. Figure 3-33 is a graph representing the frequency of this hydrogen bond. The disappearance of this bond may have been the reason for causing the warhead to leave the active site tunnel.

### 3.2.6 Trajectory of **1f**-80ns-6

In the starting conformation, the distance between the warhead and the sulphur atom of CYS277 was 5.7 Å. There were 2 hydrogen bonds with CYS277 side chain and GLN324 side chain. There were also pi-alkyl interactions between the piperazine ring of **1f** and TRP214 and TRP332. However, the lipophilic part (dansyl group) was not in the hydrophobic region of TG2. During the simulation, the warhead moved further away from CYS277 and the dansyl group failed to enter inside the hydrophobic loop of TG2.

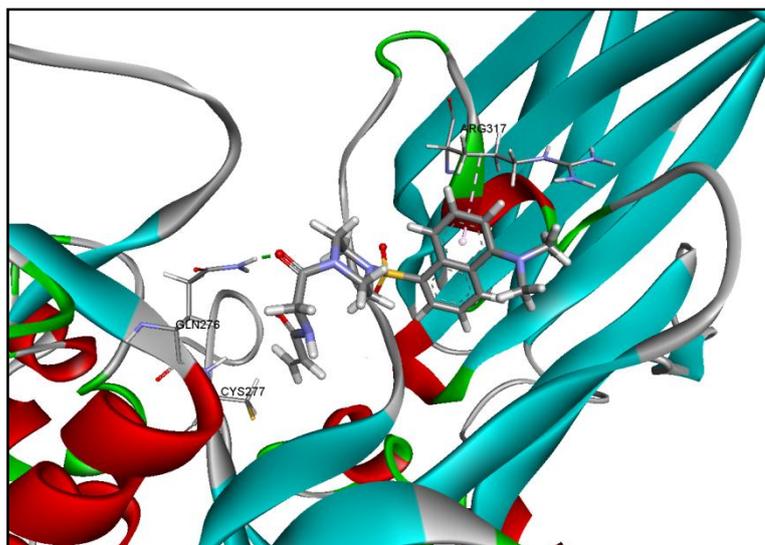


Figure 3-31: Interactions between 1e and TG2 active site in the starting conformation.

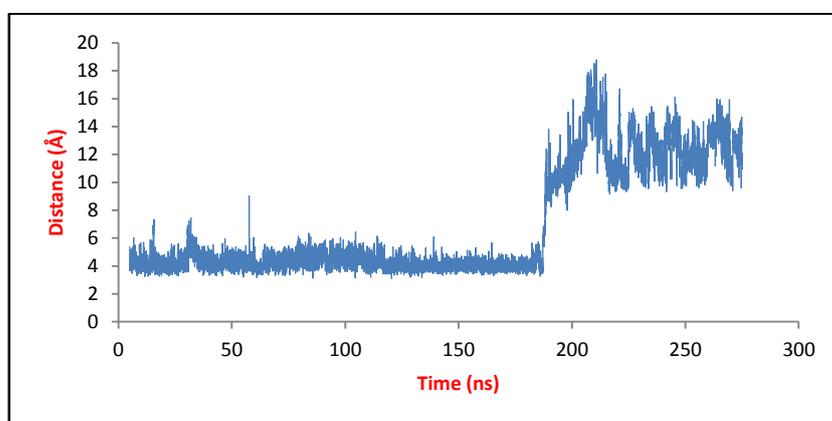


Figure 3-32: Graph for the distance between warhead of 1e and sulphur of CYS123 showing the increased distance after 187 ns.

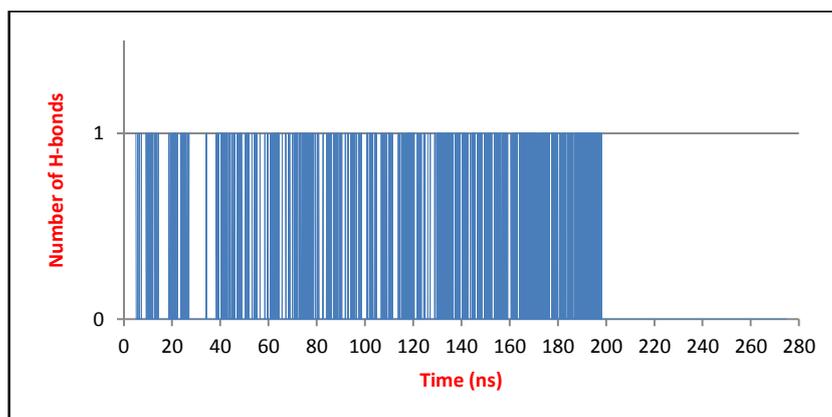
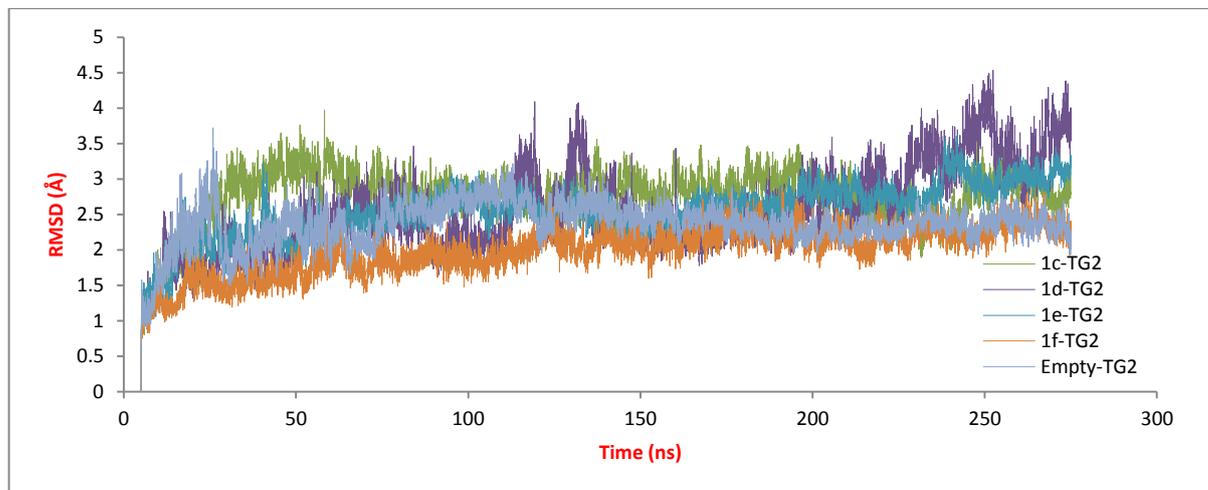


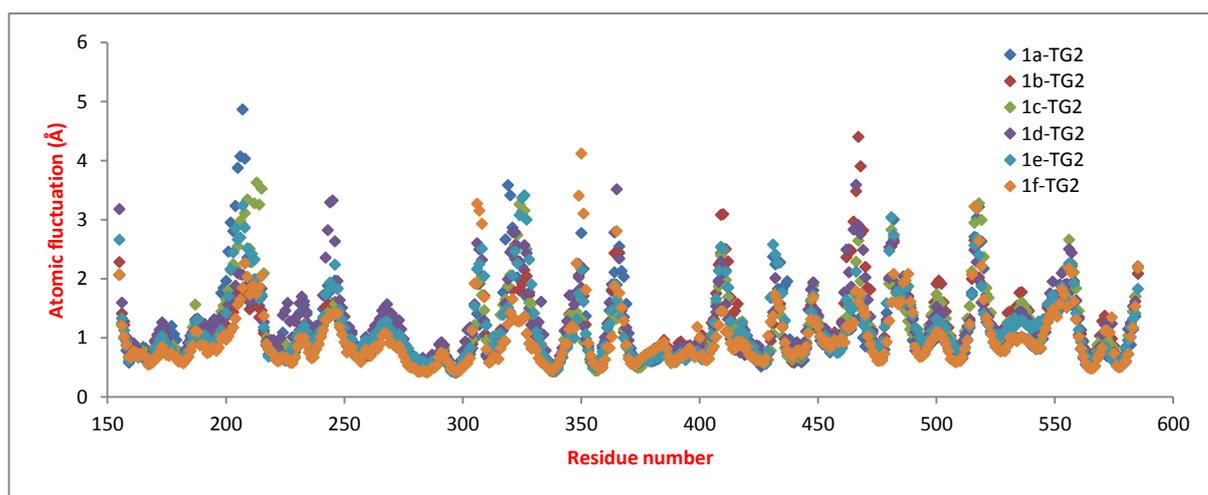
Figure 3-33: Frequency of the H-bond between 1e and GLN276 showing the disappearance of the bond after 187 ns.

The temperature and total energy for the last four simulations (compounds **1c**, **1d**, **1e** and **1f**) were stable throughout and very similar to graphs presented in figures Figure 3-15 and Figure 3-27. Also, none of these compounds caused any large scale changes in TG2. This was evident from the RMSD measured for TG2 only in each of the 4 simulations and compared to that from the simulation on empty TG2 for the same simulation time (Figure 3-34). The atomic fluctuations for TG2 residues in the 6 simulations have also been compared. They also showed that there were similar modes of motions in the 6 simulations.

Figure 3-35 contains the atomic fluctuation values for the 6 simulations and it can be seen the locations of the peaks are matching.



**Figure 3-34:** RMSD graphs for TG2 in the simulations for compounds **1c**, **1d**, **1e** and **1f** compared to the simulation of empty TG2. A detailed description of empty TG2 MD is presented later in section 3.9 of this chapter.



**Figure 3-35:** Atomic fluctuations for TG2 residues in the 6 MD simulations.

Several observations can be drawn from looking at the trajectories of the complexes. Hydrogen bonds are essential in keeping the warhead in the catalytic tunnel close to SG of CYS277. This was very clear with the trajectories of **1a** and **1c** where discontinuation of a hydrogen bond was associated with the warhead leaving the tunnel. ASN333 is probably the most important residue in this respect, being positioned conveniently to bond with the central, or linker, part of the inhibitors leading the warhead to the tunnel. GLN276 is another example, being close to CYS277, where a hydrogen bond with this residue may provide a means for keeping the warhead in the catalytic tunnel (trajectories for **1c** and **1e** are examples for the significance of this residue). Pi-cation interactions involving the positively charged sulphur atom of the inhibitors containing sulfonium ion warheads and the aromatic

rings of the residues close to CYS277, especially PHE334, also played significant role in maintaining the pose of compound **1b**.

Although the warhead left the catalytic tunnel in compounds **1a** and **1e**, both of which were appropriately posed at the start of their simulations, the lipophilic parts of the 2 compounds remained in the hydrophobic loop of TG2 active site. This indicates that the preservation of the lipophilic part in its starting location is easier than preserving the warhead. The reasons may include the magnitude of lipophilic interactions in the region and the shape of the hydrophobic loop and its ability to enclose the lipophilic part of the inhibitors. This enclosure was particularly evident with the tertiary butyl lipophilic part of compound **1b** (Figure 3-20).

### 3.3 Docking into TG2 Models from MD Trajectories of Complexes

Models of the active site were extracted at 30-nanosecond intervals from the six MD trajectories of the complexes from the initial docking to perform docking experiments starting from the 35<sup>th</sup> nanosecond. A total of 9 active site models were generated from each of the 6 trajectories to perform dockings on. Again, only the 6 active compounds from Badarau et al. (2015) were used at this stage. CAChe docking was used with models from the trajectories of **1b**, **1c** and **1e** but no good results were obtained with any model. As a result, CAChe was not used on models from the other trajectories. Instead, GOLD was used with models from all the trajectories. The GOLD settings were the same as those used for the first stage with the following modifications: early termination was allowed and default settings for docking speed with 100% efficiency were applied.

The results of docking on models from each of the six trajectories will be presented separately based on the trajectory. Within the individual trajectories, only results from models that have shown good docking will be presented, except when there are no good results. The complexes were named as in the initial docking. Figures will show the pose of the compounds within the active site with the distance between the warhead and the sulphur atom (SG) of CYS277 as a red solid line. The compounds will be in a stick form except the warhead atom which will be shown as a ball. Hydrogen bonds will be shown as green dotted lines and pi interactions as dotted light pink (pi-alkyl), dotted dark pink (pi-pi) or dotted orange (pi-cation) lines.

During the analysis certain criteria had been set to define a good docking complex and from which to select valid active site models. These criteria included:

1. An appropriate bent pose within the active site for the active compounds with the warhead in the catalytic tunnel containing the active site cysteine residue (CYS277) and the lipophilic part in the hydrophobic region of the active site (Badarau, Mongeot, et al. 2013; Prime, Brookfield, et al. 2012).
2. A ranking in the top 4 places out of the 20 solutions generated for that compound based on GoldScore scoring function first and then the CHEMPLP function. In both scoring functions, the higher (more positive) the score, the better is the complex because the score is taken as the negative of the sum of multiple energy terms (CCDC Software Limited 2013).
3. The presence of hydrogen bonds between the ligand and some key residues in the active site especially TRP241, GLN276, TRP332, ASN333 and PHE334.

### 3.3.1 Docking into TG2 models from the 1a complex trajectory

The warhead of the different active compounds was close to CYS277 in many cases for dockings on models from **1a** trajectory. However, the lipophilic part of the active compounds was not under the hydrophobic loop of TG2 in the majority of these cases. Figure 3-36 shows an example of such pose for compound **1b** in the model taken at 65 ns from this trajectory.

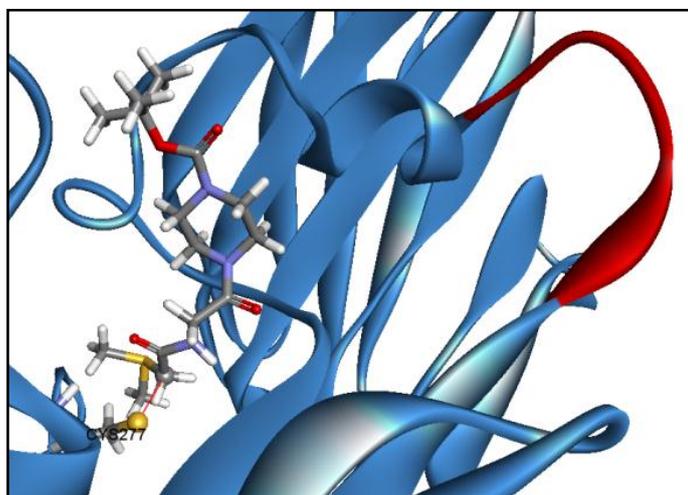


Figure 3-36: Example pose of **1b** with the lipophilic part outside the hydrophobic loop (coloured red).

The model taken at 95 nanoseconds (95ns) gave good results in terms of poses and scores and was therefore selected as a valid model to be taken for further investigation. Table 3-3 shows the docking results of the 6 active compounds on 95ns model. Figure 3-37 presents the pose of compound **1b** as an example of complexes generated in this model. From Table 3-3, it can be seen that a hydrogen bond with ASN333 was present in **1b** and **1e** complexes. In all the compounds, there were hydrophobic interactions stabilising the lipophilic part in the

hydrophobic loop. In case of the sulfonium ion warhead (**1b** and **1f**), there were pi-cation interactions with TRP241 and/or PHE334.

Complex	Distance (Å)	H-bonds	GoldScore*	Rank	CHEMPLP*	Rank
<b>1b</b> -95ns-9	6.1	ASN333	49	2 <sup>nd</sup>	38	14 <sup>th</sup>
<b>1b</b> -95ns-18	6.3	ASN333	47	3 <sup>rd</sup>	52	3 <sup>rd</sup>
<b>1c</b> -95ns-2	5.5	HIS335	56	2 <sup>nd</sup>	46	11 <sup>th</sup>
<b>1c</b> -95ns-12	4.3	CYS277 PHE334	47	4 <sup>th</sup>	61	4 <sup>th</sup>
<b>1e</b> -95ns-1	4.3	GLN276 ASN318 ASN333	57.07	1 <sup>st</sup>	49	4 <sup>th</sup>
<b>1e</b> -95ns-4	3.7	GLN276 ASN318 PHE334	57.03	2 <sup>nd</sup>	39	12 <sup>th</sup>
<b>1f</b> -95ns-13	4.5	GLN276 ASN318	52	9 <sup>th</sup>	31	12 <sup>th</sup>

Table 3-3: Docking results for active compounds into the 95ns model. \*GOLD scoring functions are unitless (Yang 2008).

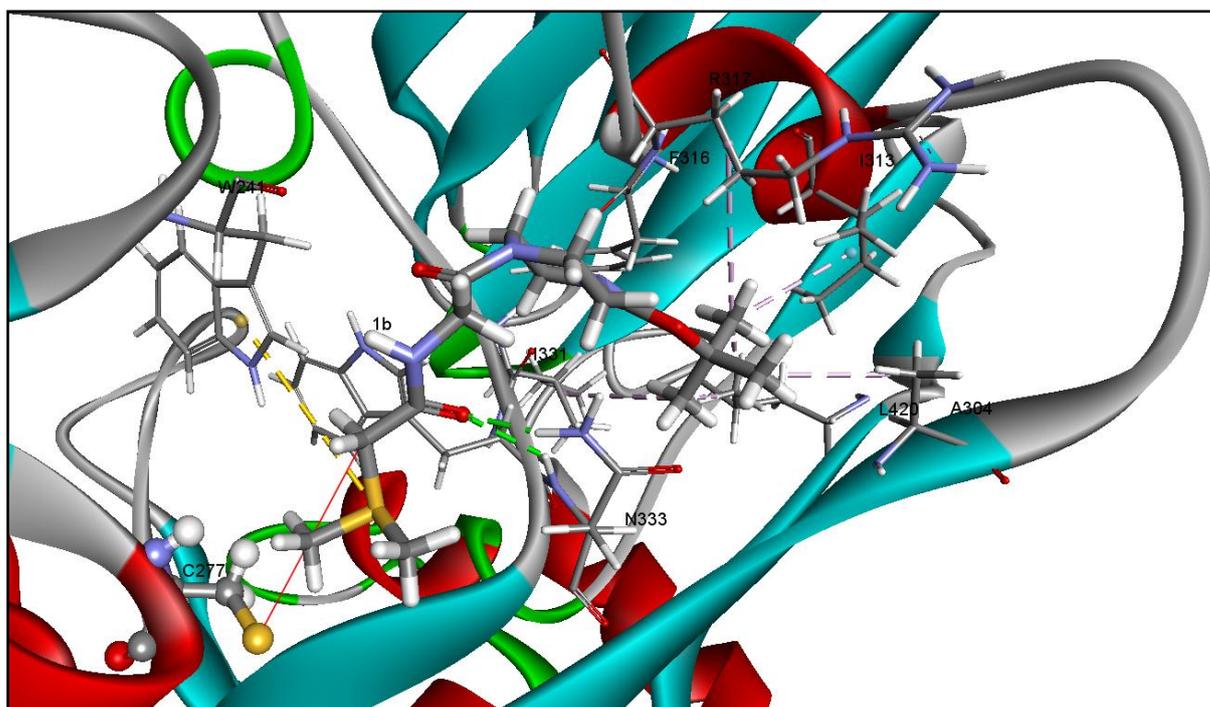


Figure 3-37: Poses of **1b** in the 95ns model showing interactions with TG2 residues.

ASN333 was the only residue among those in Table 3-3 to form hydrogen bonds with its side chain atoms as well as backbone atoms (Figure 3-37). The remaining 3 residues participated in either side chain or backbone atoms in hydrogen bonds with the compounds, but not both; GLN276 side chain, ASN318 backbone and PHE334 backbone. As a result of the good docking results obtained with model 95ns, two additional models were selected from this trajectory which are close to 95ns and were taken at 90 (90ns model) and 100 (100ns model) nanoseconds. The docking results on these two models are presented in Table 3-4

and Table 3-5. Figure 3-38 gives examples for the poses in 90ns model and Figure 3-39 examples for 100 ns model.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1a-90ns-8</b>	3.5	None	47	2 <sup>nd</sup>	46	13 <sup>th</sup>
<b>1a-90ns-14</b>	3.5	TRP241	46	4 <sup>th</sup>	55	6 <sup>th</sup>
<b>1b-90ns-6</b>	7.4	ASN333	41	16 <sup>th</sup>	47	2 <sup>nd</sup>
<b>1d-90ns-8</b>	5.2	None	57	1 <sup>st</sup>	62	1 <sup>st</sup>
<b>1d-90ns-20</b>	6.5	PHE334	55	2 <sup>nd</sup>	0.33	19 <sup>th</sup>
<b>1e-90ns-11</b>	4.1	ASN333	48	1 <sup>st</sup>	32	13 <sup>th</sup>
<b>1e-90ns-17</b>	3.1	GLN276 ASN318 PHE334	47	2 <sup>nd</sup>	23	18 <sup>th</sup>

**Table 3-4: Docking results for active compounds into the 90ns model.**

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1a-100ns-1</b>	5.2	None	46	1 <sup>st</sup>	49	9 <sup>th</sup>
<b>1a-100ns-4</b>	4.9	None	45	2 <sup>nd</sup>	53	3 <sup>rd</sup>
<b>1b-100ns-3</b>	5.9	ASN333 PHE334	46	4 <sup>th</sup>	58	1 <sup>st</sup>
<b>1e-100ns-1</b>	3.5	GLN276	60	1 <sup>st</sup>	46	4 <sup>th</sup>
<b>1e-100ns-8</b>	3.7	GLN276	55	3 <sup>rd</sup>	56	1 <sup>st</sup>
<b>1e-100ns-9</b>	3.6	None	59	2 <sup>nd</sup>	50	2 <sup>nd</sup>
<b>1f-100ns-4</b>	5.1	GLN276 ASN318	54	4 <sup>th</sup>	51	3 <sup>rd</sup>

**Table 3-5: Docking results for active compounds into the 100ns model.**

In 90ns and 100ns models, again there were hydrogen bonds with ASN333 (side chain and backbone) and PHE334 (backbone), in addition to TRP241 (side chain), GLN276 (side chain) and ASN318 (side chain). With the exception of ASN318, all the other mentioned residues are important in stabilising the warhead within the catalytic tunnel (Badarau, Mongeot, et al. 2013; Pinkas et al. 2007). In **1b** and **1d** complexes, the pi-cation interaction was also present between the sulfonium ion sulphur and the phenyl group of PHE334, TRP241 or TRP332 (Figure 3-38 and Figure 3-39). Hydrophobic interactions involving different lipophilic parts were also present. The residues participated the most were ALA304, ILE313, ILE331 and LEU420. In the 100ns model, there were pi-pi interactions between the dansyl group of **1e** and **1f** and the side chain of PHE316 (Figure 3-38 and Figure 3-39).

The 95ns model gave some good results when tried first with the original definition of the active site. The results presented above were obtained when the active site definition of **1b** trajectory was used. The same definition was also used for dockings on 90ns and 100ns models. The better docking results using this definition may give an indication to the significance of the 306-311 residue region (the hydrophobic loop of the active site) of TG2 for the activity of inhibitors. They also emphasise the importance of the lipophilic part of the

inhibitors and its pose within the hydrophobic loop of TG2 active site for an appropriate docking complex.

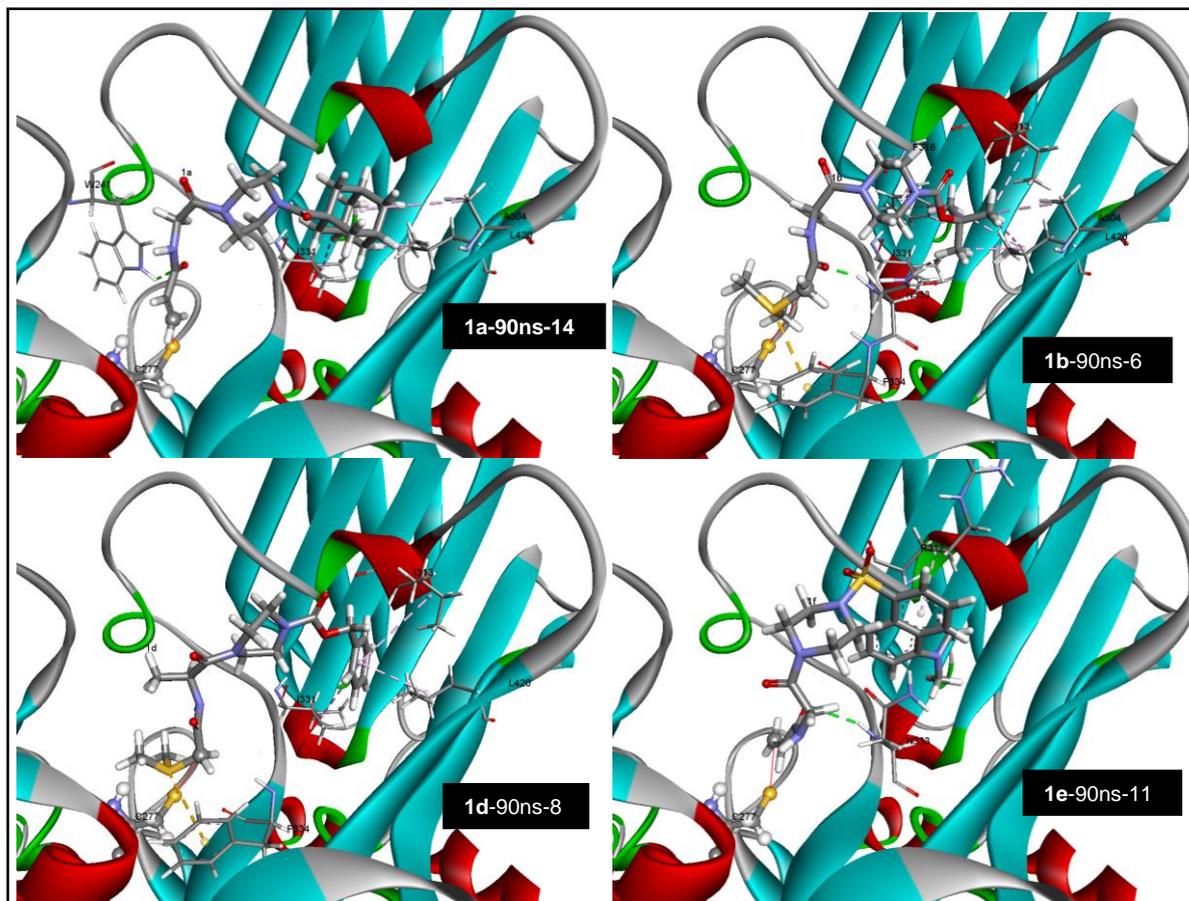


Figure 3-38: Poses of 1a, 1b, 1d and 1e in the 90ns model showing interactions with TG2 residues.

### 3.3.2 Docking into TG2 models from the 1b complex trajectory

Residues 306-311 were added to active site definition used in models of the **1b** trajectory. CAChe was used for docking on models from this trajectory, but out of the 216 docking attempts (4 dockings per compound for 6 compounds in 9 active site models), only 7 plausible poses were obtained. Figure 3-40 shows 2 of these 7 good poses for compounds **1d** and **1c** in the models taken at 65 ns and 95 ns respectively.

When GOLD was used, 3 models for the active site taken from this trajectory produced good docking results and were considered valid. These were 65ns, 155ns and 245ns. Their results will be presented separately. The other models also produced some good results but not as much. In all cases, the best results were obtained for compound **1b**.

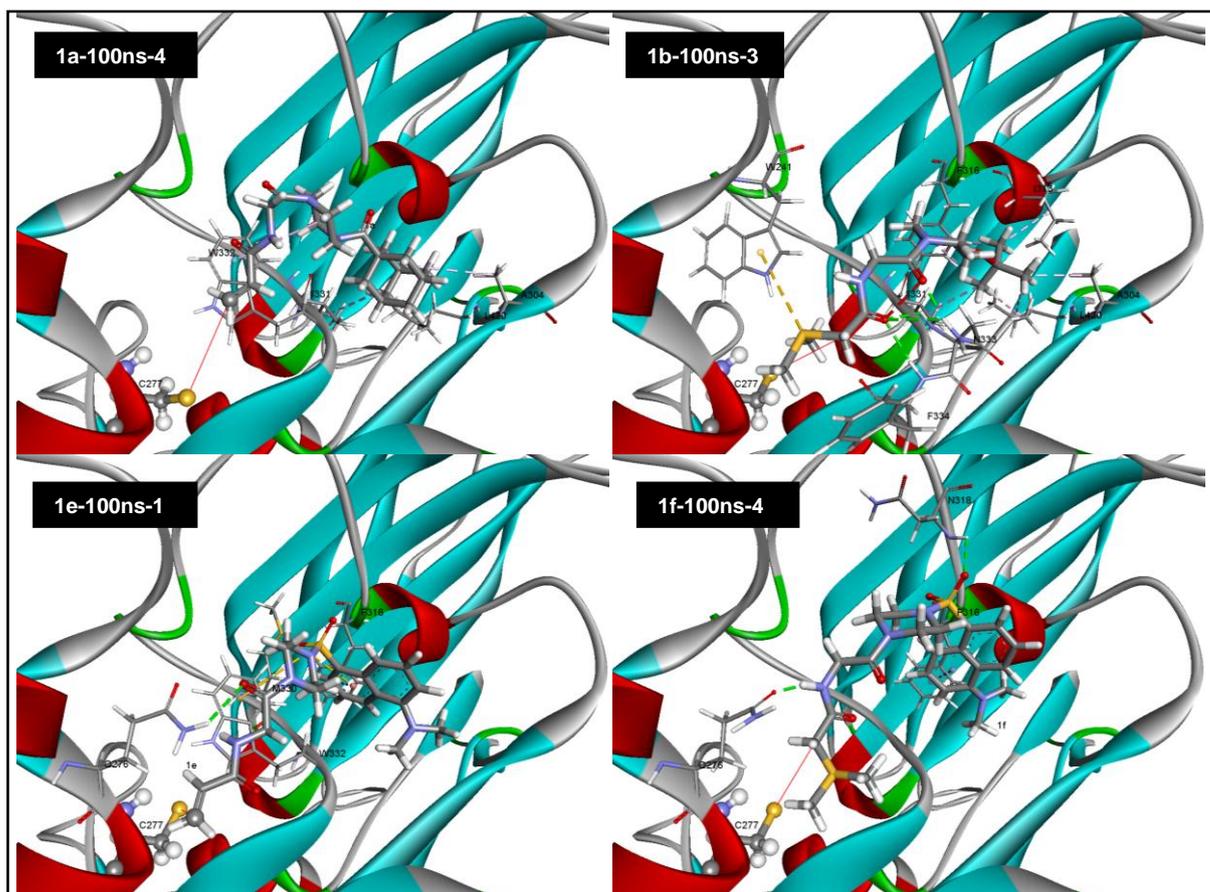


Figure 3-39: Poses of 1a, 1b, 1e and 1f in the 100ns model of the 1a trajectory showing interactions.

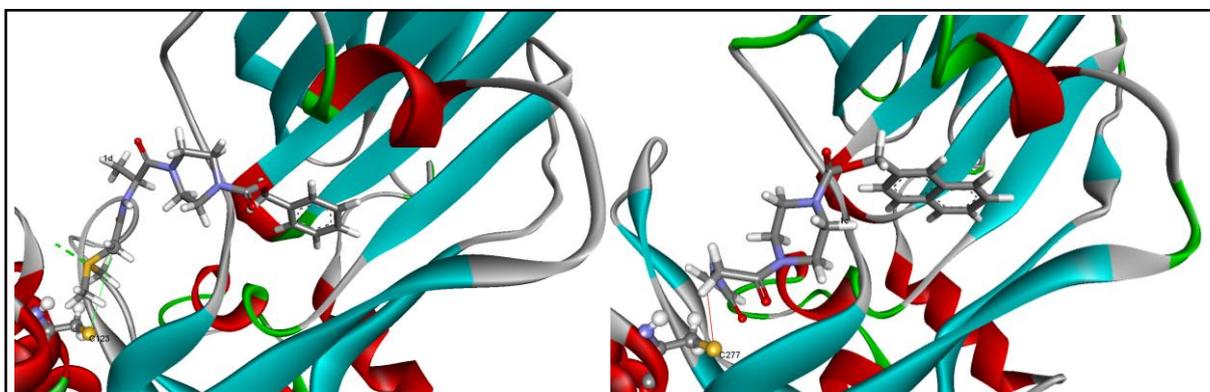


Figure 3-40: Poses of 1d and 1c in 65ns and 95ns models of the 1b trajectory after CAChe docking.

### 3.3.2.1 65ns model from the 1b trajectory

Table 3-6 represents a summary of the good docking results obtained with this model. The poses for these results are presented in Figure 3-41.

It can be seen from Table 3-6 and Figure 3-41 that a hydrogen bond with ASN333 is important for a good ligand pose. Pi interactions between residues from the hydrophobic loop of TG2 and the lipophilic parts of the various ligands were also evident. Finally, regarding the sulfonium ion warhead in **1b** and **1d**, pi-cation interactions were also noticed

between TRP332 and HIS335 from TG2 and **1b** and between PHE280 and HIS335 of TG2 and **1d**.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1b</b> -65ns-13	5.9	ASN333	66.32	1 <sup>st</sup>	51	12 <sup>th</sup>
<b>1b</b> -65ns-18	6.3	ASN333 PHE334	66.3	2 <sup>nd</sup>	50	13 <sup>th</sup>
<b>1d</b> -65ns-12	4.2	CYS277 HIS335	63	8 <sup>th</sup>	40	17 <sup>th</sup>
<b>1e</b> -65ns-20	4.5	GLN276 ASN333 PHE334	68	2 <sup>nd</sup>	78	1 <sup>st</sup>

Table 3-6: Docking results for active compounds into the 65ns model.

### 3.3.2.2 155ns model from the **1b** trajectory

The results obtained from docking on this model are presented in Table 3-7 and Figure 3-42. Hydrogen bonds were not very frequent in this model but there were pi-interactions between the piperazine ring of the compounds with TRP241 (for **1c**, **1e** and **1f**) and with PHE334 (for **1f**). These interactions may have served the function of keeping the warhead within the catalytic tunnel close to CYS277.

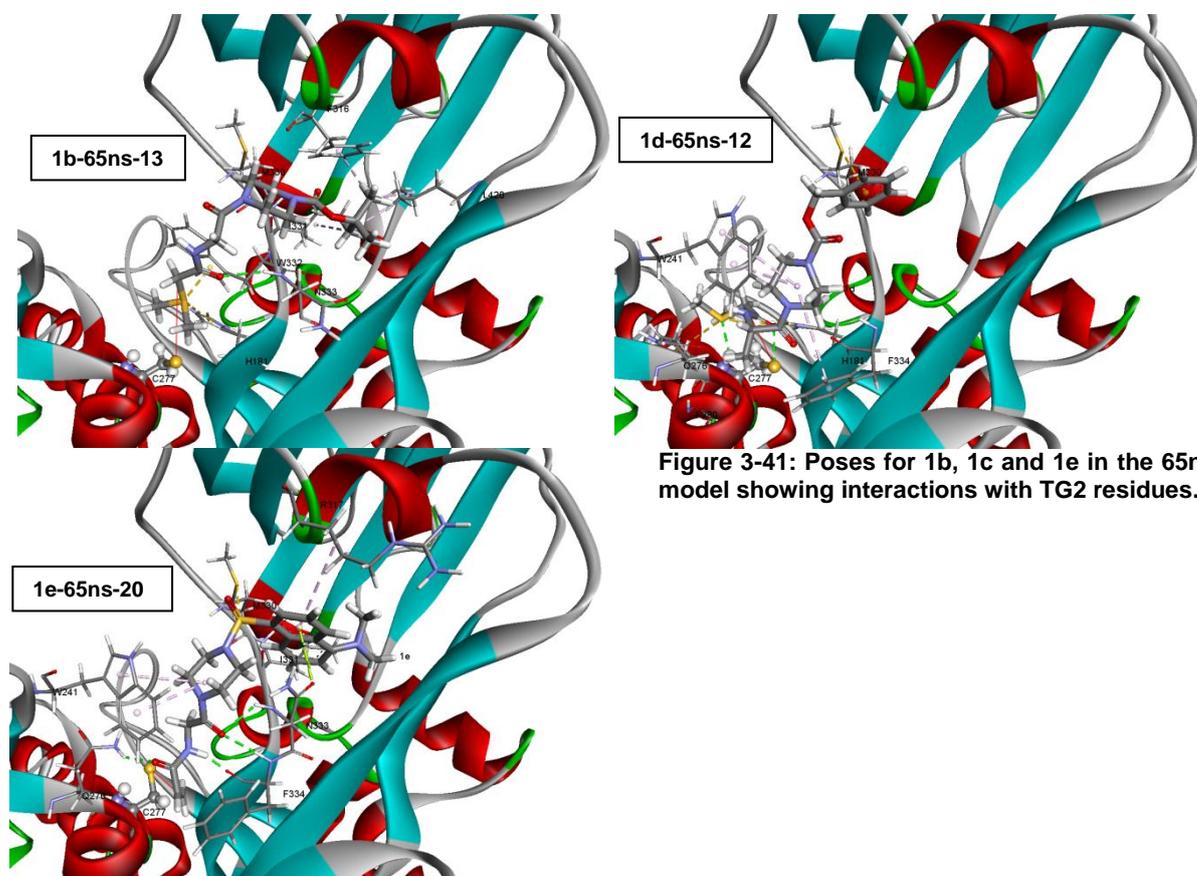


Figure 3-41: Poses for **1b**, **1c** and **1e** in the 65ns model showing interactions with TG2 residues.

Complex	Distance	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1c-155ns-16</b>	4.9	GLN276 ASN333	73	3 <sup>rd</sup>	86	1 <sup>st</sup>
<b>1d-155ns-19</b>	6.4	None	66	4 <sup>th</sup>	23	19 <sup>th</sup>
<b>1e-155ns-2</b>	4.4	None	70	3 <sup>rd</sup>	67	4 <sup>th</sup>
<b>1f-155ns-6</b>	4.2	CYS277	76	1 <sup>st</sup>	83	1 <sup>st</sup>

Table 3-7: Docking results for active compounds into the 155ns model.

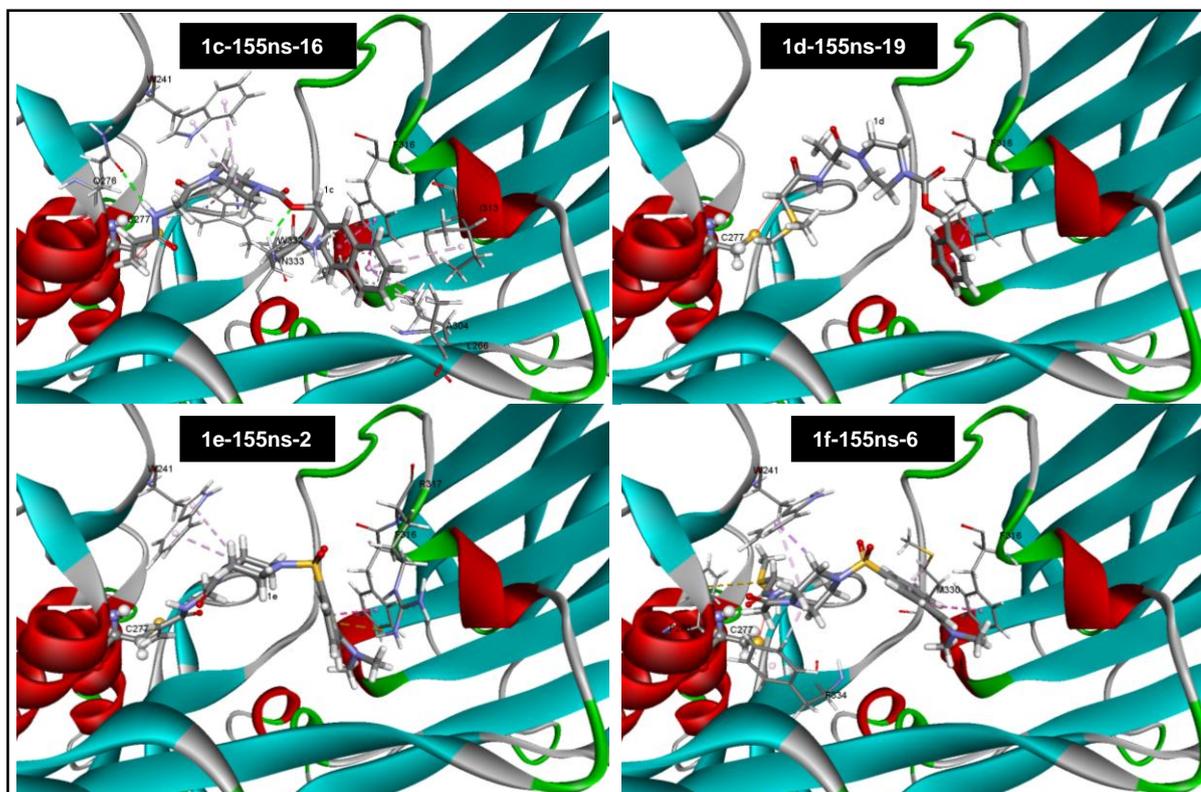


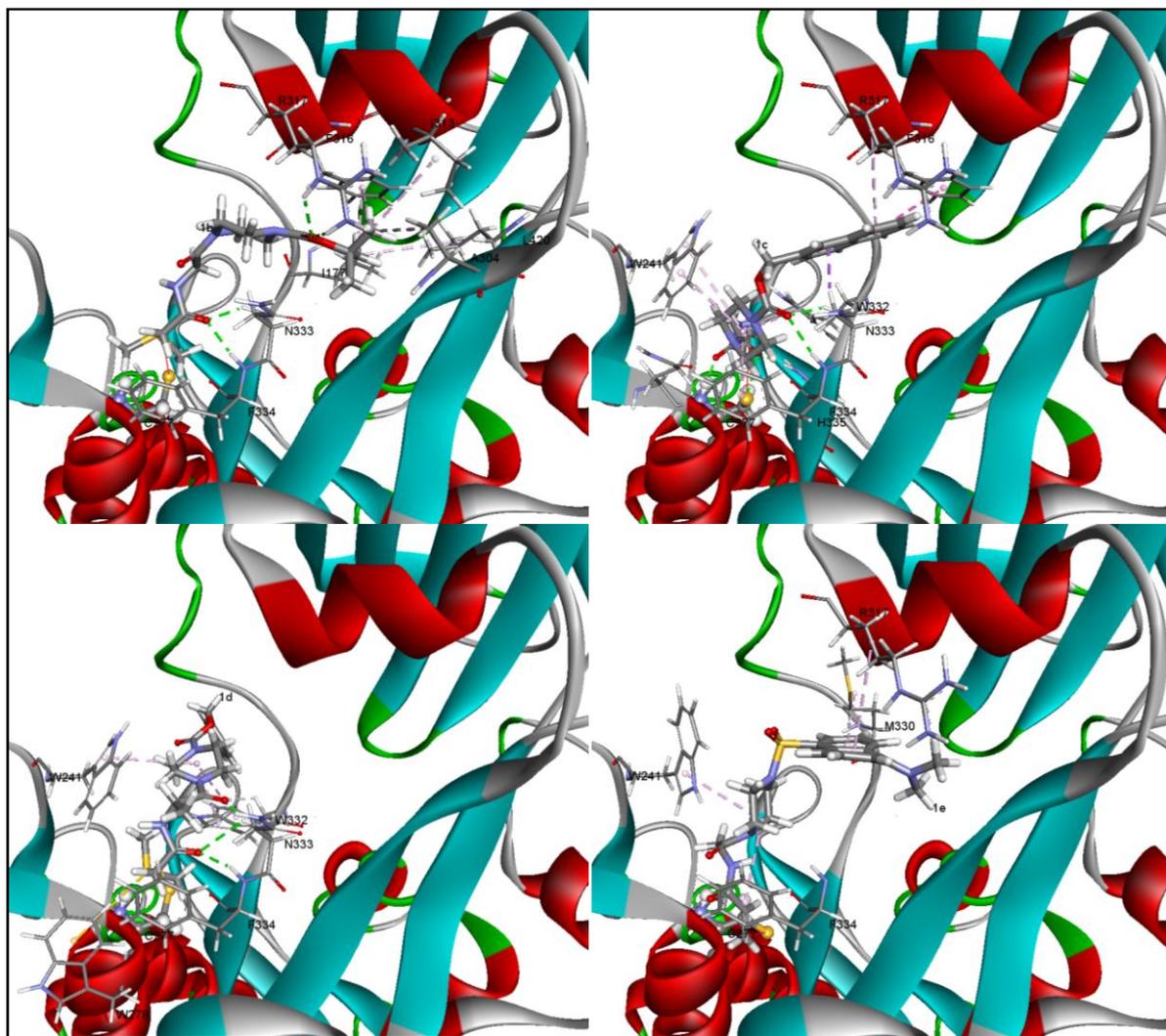
Figure 3-42: Poses for 1c, 1d, 1e and 1f in the 155ns model showing interactions with TG2 residues.

### 3.3.2.3 245ns model from the 1b trajectory

The results for the docking of the 6 active compounds on 245ns model are summarised in Table 3-8 and Figure 3-43.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1b-245ns-9</b>	6.8	ARG317 ASN333 PHE334	63	2 <sup>nd</sup>	73	1 <sup>st</sup>
<b>1b-245ns-2</b>	6.6	ASN333	69	1 <sup>st</sup>	62	5 <sup>th</sup>
<b>1c-245ns-4</b>	5.5	CYS277 ASN333 PHE334 HIS335	70	1 <sup>st</sup>	70	7 <sup>th</sup>
<b>1d-245ns-16</b>	6.4	ASN333 PHE334	67	2 <sup>nd</sup>	71	2 <sup>nd</sup>
<b>1e-245ns-11</b>	5.1	None	67	1 <sup>st</sup>	73	1 <sup>st</sup>

Table 3-8: Docking results for active compounds into the 245ns model.



**Figure 3-43: Poses for 1b, 1c, 1d and 1e in the 245ns model showing interactions with TG2 residues.**

Again, hydrogen bonds and pi-interactions were present and participating in positioning both the warhead and the lipophilic part of the active compounds. It can be seen from Figure 3-43 that the lipophilic part of **1d** was not located under the hydrophobic loop of TG2. In fact, this conformation has been adopted by compounds **1a** and **1d** in many cases of the dockings on the valid models. In this particular case of **1d** in the 245ns model, the conformation was ranked second in both scoring functions and the warhead was somewhat close to CYS277. No hydrophobic interactions were detected between the terminal phenyl group of **1d** and TG2 residues where it was docked but it seems that the pi-interaction with TRP332 with the piperazine ring of **1d** was responsible for positioning the lipophilic part.

### 3.3.3 Docking into TG2 models from the 1c complex trajectory

CAChe and GOLD were used for docking on models from this trajectory. Both programs were unable to produce plausible poses for the active compounds in the TG2 active site. This was true for the pose of the warhead and the lipophilic part of the ligands. The same

poor results were obtained after using the definition of the active site adopted from the trajectory of **1b**.

### 3.3.4 Docking into TG2 models from the **1d** complex trajectory

GOLD was used for docking into models from the **1d** trajectory using the original definition for the active site. No good results were obtained. Both the warhead and the lipophilic part of the active compounds failed to dock into their appropriate places in TG2. Active site definition from the **1b** trajectory was then tried in models 35ns, 95ns, 155ns, 215ns and 275ns as representatives for models from **1d** trajectory. Again, no good results were obtained.

### 3.3.5 Docking into TG2 models from the **1e** complex trajectory

CAChe was used for docking into models from this trajectory but no good results were obtained. In very few cases, the lipophilic part was able to position itself in the hydrophobic region but the warheads of the 6 compounds failed to dock close to the sulphur atom of CYS277. Figure 3-44 is an example for compound **1b** in the model taken at 155 nanoseconds, where the distance between the warhead and SG of CYS277 was about 9 Å.

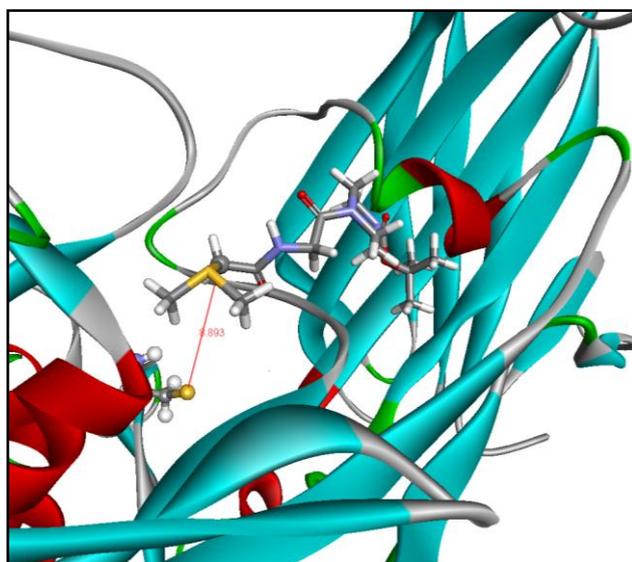
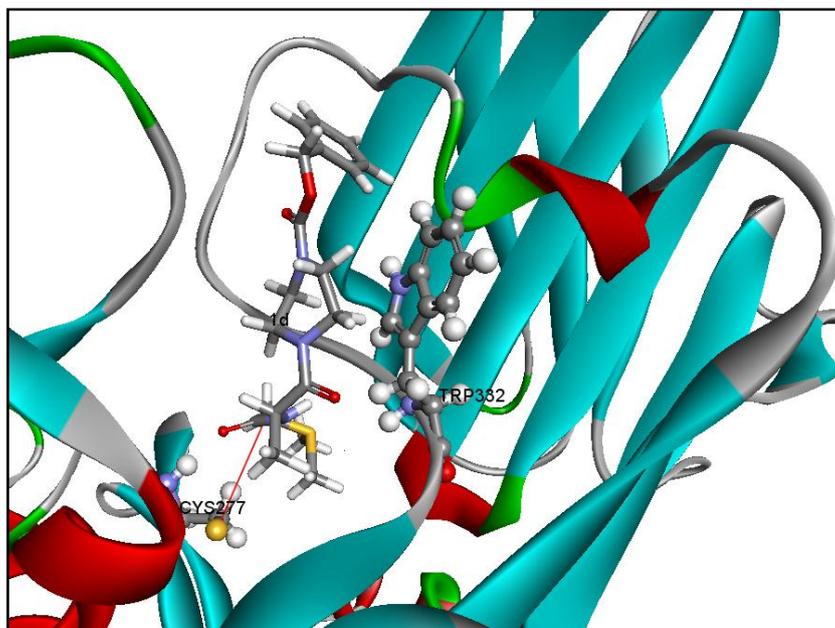


Figure 3-44: Pose of **1b** in the 155ns model of **1e** trajectory after CAChe docking.

For GOLD dockings, the results were similar to those obtained on models from the **1a** trajectory in that the warhead managed to get close to CYS277 but the lipophilic part failed to be positioned under the hydrophobic loop. The lipophilic part was pointing out of the protein in the majority of cases. In Figure 3-45, a representative of docking on models from this trajectory is shown; compound **1d** is docked into the 125ns model where the warhead was about 4.8 Å away from CYS277 but the lipophilic part is not in place. The conformation of TRP332 is shown as a possible cause for this.



**Figure 3-45:** The pose of **1d** in the 125ns model of **1e** trajectory showing the distance with CYS277 (4.8 Å) and the conformation of TRP332.

The conformation of TRP332 changed during the MD simulation in such a way as to block the hydrophobic loop in TG2 and prevent the access of the lipophilic part of the 6 active compounds to this region (Figure 3-46). Although TRP332 was one of the residues that were selected to be treated as flexible during the docking, it appears that a movement to allow the lipophilic part to dock well was outside the range of the flexibility offered by GOLD. This flexibility was based upon The Penultimate Rotamer Library (Lovell et al. 2000) which includes the most commonly observed side chain conformations for the natural amino acids. Therefore, this orientation of TRP332 was responsible for the inappropriate poses for the active compounds.

### 3.3.6 Docking into TG2 models from the **1f** complex trajectory

Again, no good results were obtained for the 6 active compounds when docked into the models taken from this trajectory. In some cases, the warhead was able to dock to a position that is close to CYS277, but the lipophilic part was never in the hydrophobic region of TG2. The reason behind this seems to be the orientation of TRP332. This residue, as in the models taken from **1e** trajectory, was blocking the approach of the lipophilic part to the hydrophobic loop of TG2. Figure 3-47 presents an example of the poses on models from **1f** trajectory. The figure shows the pose of **1d** in the 245ns model with the conformation of TRP332.

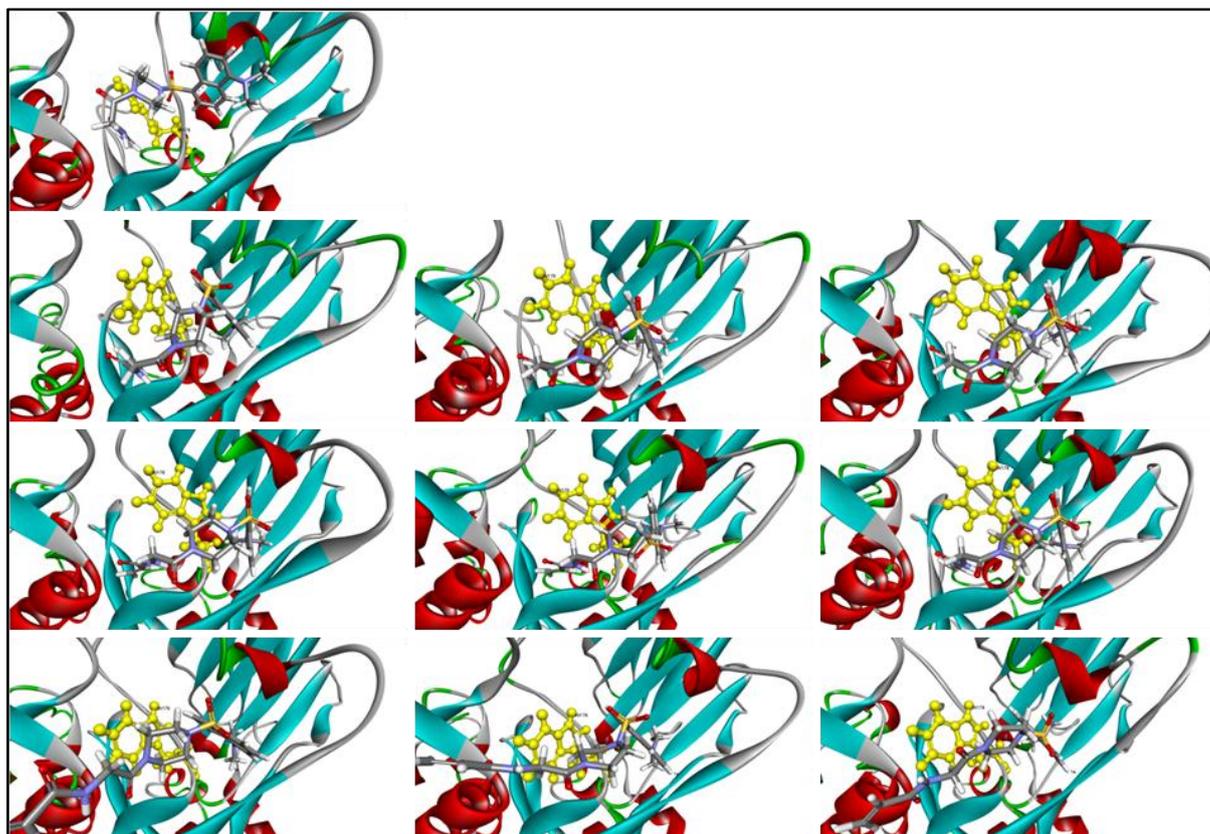


Figure 3-46: Orientation of TRP332 in the 1e trajectory compared to the starting conformation (top left). The other conformations were taken at 30 ns intervals from the trajectory. TRP332 is shown in yellow as ball and stick.

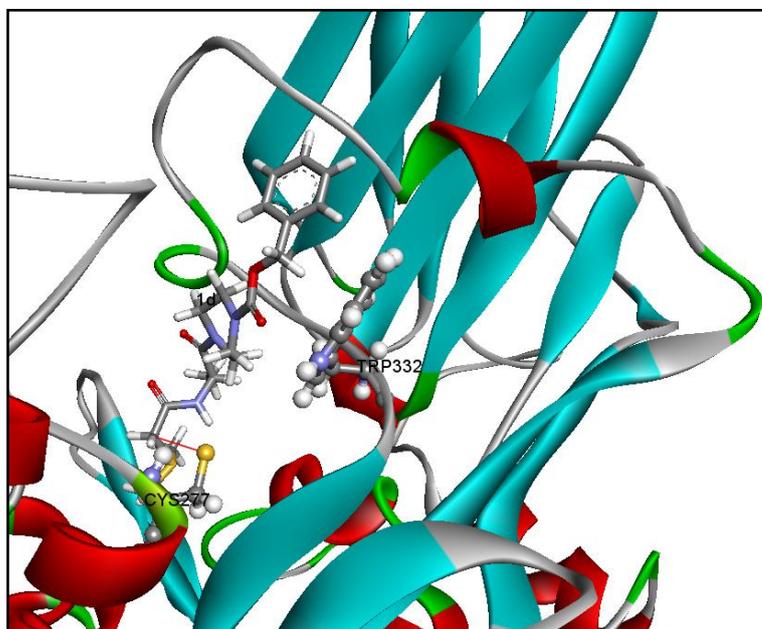


Figure 3-47: The pose of 1d in the 245ns model of 1f trajectory. The distance with CYS277 is shown (3.3 Å) along with the conformation of TRP332 to show a possible cause for the pose of the lipophilic part.

The criteria set at the start of this section (3.3) for the definition of an acceptable docking complex have been confirmed with the results obtained from the 6 models of the first two trajectories. Hydrogen bonds involving the central region (the linker) of the inhibitor molecules are important for generating the required bent conformation of the ligand, as are

the lipophilic interactions at the hydrophobic loop of TG2. It has also been noted that pi-cation interactions involving the sulfonium ion of compounds **1b**, **1d** and **1f** may be able to compensate for the lack of hydrogen bonding to the linker region in positioning the warhead to CYS277. This role has also been played by lipophilic pi interaction involving the piperazine ring of some compounds. Those observations are consistent with the ones made during the analysis of the MD trajectories on the docked complexes in section 3.2.

Compound **1e** was the only compound that has managed a good pose in all the selected valid models in this stage of docking. This should not be surprising, knowing that **1e** is the most potent compound within the group. The compound failed, however, to exhibit the formation of hydrogen bonds with TG2 residues in 2 models; 155ns and 245ns. The second best performance in terms of the correct poses was achieved by compound **1b**, which docked appropriately in all the models with the exception of 155ns. That was combined with passing the other criteria as well, including the formation of hydrogen bonds and the good ranking. Since all the compounds share the same linker region, and each 3 compounds share the same warhead, it can be said that the lipophilic part of the inhibitor in this group of compounds was responsible for the different docking results. Therefore, a combination of a dansyl group with an acrylamide warhead (compound **1e**), and a combination of a tertiary butyl group and a sulfonium ion warhead (compound **1b**) (Table 3-1) may offer superiority to the potential of the compounds to dock well in the active site of TG2.

Due to the significance of the lipophilic interactions in the docking complexes that proved to be valid, the presence of such an interaction has been added to the criteria defining a good docking complex; besides the appropriate bent conformation, the good rank with one or both of the used scoring functions and the presence of hydrogen bonds with the key residues in the active site of TG2.

Despite the good results obtained with the 6 models and from 6 different compounds with 2 types of warheads, the models need to be tested on a larger set of compounds, and some inactive compounds should be tested as well, in order to determine if the models could discriminate between potent and inactive compounds. This was the rationale of the next stage of this work.

## 3.4 Validation of Active Site Models

The six active site TG2 models selected from the previous stage were validated by docking experiments and short MD simulations. The validation process included:

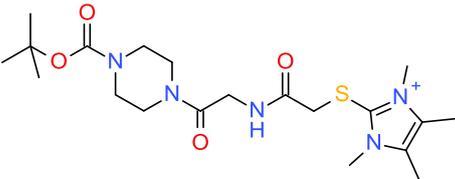
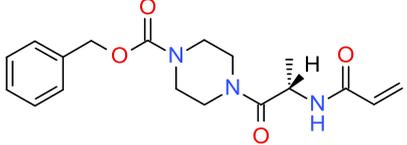
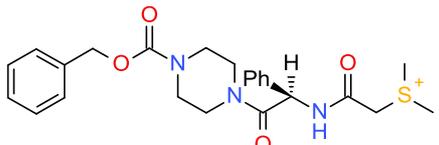
1. Performing a docking experiment with GOLD that is similar to the previous one with turning off early termination and adding 3 inactive compounds to the test set.
2. Performing a docking experiment with GOLD similar to the one described in step 1 with the same test compounds but with having 25 water molecules within the active site of TG2.
3. Applying a GOLD docking experiment identical to that in step 1 with a new set of 15 active and 5 inactive TG2 inhibitors.
4. Applying docking experiment using GOLD on the 6 models with settings similar to those in step 1 on a set of compounds that includes the original 6 active compounds, the 3 inactive compounds introduced in step 1 and the 20 active and inactive compounds from step 3. This experiment was repeated 3 times.
5. Performing 5-nanosecond MD simulations on selected docking complexes of the 6 valid models from the above experiments and the previous docking results.
6. Performing docking with GOLD using settings as those in experiment 1 on a set of 3 active compounds (different chemical type to the initial set) from the work by Prime, Andersen et al. (2012).

### 3.4.1 Validation process 1 (3 inactive compounds)

A valid model should not only dock active compounds in a good manner, but it should also give bad docking results for inactive compounds. This was the rationale behind adding three inactive compounds to the test set [compounds **2a**, **2b** and **2c** (Table 3-9) (Badarau et al. 2015; Griffin et al. 2014)]. The results will be presented by models.

#### 3.4.1.1 Models from the 1a trajectory

The results of docking the 6 active compounds and the 3 inactive ones into models 90ns, 95ns and 100ns from the **1a** trajectory are summarised in Table 3-10. An analysis of the docking complexes from this trajectory showed that, although some complexes in models 90ns and 100ns did not contain any hydrogen bonds, the hydrophobic interactions involving the lipophilic part in the hydrophobic loop of TG2 were evident in all the complexes except that of **1e** in the 90ns model. Probably that is why the complex of **1e** (**1e**-90ns-17) failed to achieve a good rank in the two scoring functions.

Compound	Structure	IC <sub>50</sub> (μM)
<b>2a</b>		400
<b>2b</b>		> 100
<b>2c</b>		> 100

**Table 3-9: Inactive compounds used in validation process 1.**

There were also pi-interactions with the piperazine ring of all the compounds in models 95ns and 100ns except for **1e** and **1f** in 95ns and **1f** in 100ns model. In those two compounds, there were pi-sulphur interactions between the sulphur atom of the sulphonamide group and the phenyl group of TRP332 or a hydrogen bond between the sulphonamide group and ASN318. Other pi-sulphur interactions were recorded involving the sulfonium ion of **1b**, **1d** and **1f** in the 3 models.

Compound **2c** achieved a good pose in the 3 models with good GoldScore ranking in the 95ns and 100ns models. The good rank is probably offset by the lack of hydrogen bonds in the 2 complexes. The GoldScore scoring function is the sum of multiple energy terms, one of which is van der Waals interactions between the ligand and the protein. Given the large size of **2c** molecule, the van der Waals component contributed much to the score giving rise to the high scores associated with **2c**. For **2b** complex in the 90ns model, however, a rank of 4 out of 20 would mean a favourable pose. This was the only pose for an inactive compound that achieved all the criteria for an acceptable docking complex.

### 3.4.1.2 Models from the 1b trajectory

The results of docking of the 9 test compounds on models 65ns, 155ns and 245ns of the **1b** trajectory are presented in Table 3-11. It can be clearly seen from this table that the hydrogen bonding profile for all the models is much better than in the models from the **1a** trajectory. This may contribute to the generally better scores for the active compounds in models from the **1b** trajectory than from the **1a** trajectory. Only one inactive compound achieved a good pose and score, **2b-245ns-4**, but in this complex, no interactions involving

the lipophilic part of the molecule were recorded. This may indicate that the pose of the lipophilic part may change easily.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1a-90ns-1</b>	3.7	None	49	2 <sup>nd</sup>	19	19 <sup>th</sup>
<b>1a-90ns-7</b>	3.5	None	47	7 <sup>th</sup>	57	1 <sup>st</sup>
<b>1b-90ns-2</b>	6.6	ASN333	51	1 <sup>st</sup>	53	3 <sup>rd</sup>
<b>1b-90ns-1</b>	6.4	None	49	2 <sup>nd</sup>	42	10 <sup>th</sup>
<b>1d-90ns-5</b>	5.3	None	58	1 <sup>st</sup>	60	1 <sup>st</sup>
<b>1d-90ns-3</b>	5.9	None	53	4 <sup>th</sup>	60	2 <sup>nd</sup>
<b>1e-90ns-17</b>	3.3	ASB318 ASN333 PHE334	43	12 <sup>th</sup>	39	8 <sup>th</sup>
<b>2b-90ns-8</b>	3.3	TRP241	54	4 <sup>th</sup>	64	8 <sup>th</sup>
<b>2c-90ns-8</b>	6.3	ASN318	46	19 <sup>th</sup>	15	20 <sup>th</sup>
<b>1a-95ns-14</b>	5.2	TRP332	50.5	2 <sup>nd</sup>	50	6 <sup>th</sup>
<b>1a-95ns-15</b>	4.3	TRP332	50	3 <sup>rd</sup>	50	5 <sup>th</sup>
<b>1b-95ns-6</b>	5.9	TRP332 ASN333	50	4 <sup>th</sup>	41	7 <sup>th</sup>
<b>1c-95ns-2</b>	4.7	CYS277 PHE334	53	2 <sup>nd</sup>	50	8 <sup>th</sup>
<b>1d-95ns-4</b>	6.6	ASN333	47	5 <sup>th</sup>	42	5 <sup>th</sup>
<b>1e-95ns-15</b>	3.7	GLN276 ASN318 ASN333	55	1 <sup>st</sup>	48	5 <sup>th</sup>
<b>1e-95ns-2</b>	3.9	None	53	2 <sup>nd</sup>	49	4 <sup>th</sup>
<b>1f-95ns-20</b>	5.8	GLN276 ASN318	61	1 <sup>st</sup>	43	5 <sup>th</sup>
<b>1f-95ns-13</b>	5.8	ASN318 ASN333	53	4 <sup>th</sup>	43.5	4 <sup>th</sup>
<b>2c-95ns-3</b>	5.6	None	61	3 <sup>rd</sup>	4	20 <sup>th</sup>
<b>1b-100ns-9</b>	5.8	ASN333	51	1 <sup>st</sup>	52	6 <sup>th</sup>
<b>1b-100ns-15</b>	5.7	ASN333	50	2 <sup>nd</sup>	47	13 <sup>th</sup>
<b>1c-100ns-6</b>	4.0	None	54	2 <sup>nd</sup>	72	2 <sup>nd</sup>
<b>1e-100ns-3</b>	3.9	None	60	1 <sup>st</sup>	47	1 <sup>st</sup>
<b>1e-100ns-16</b>	4.2	GLN276	51	7 <sup>th</sup>	56	1 <sup>st</sup>
<b>1f-100ns-16</b>	3.2	GLN276	52	6 <sup>th</sup>	82	2 <sup>nd</sup>
<b>2c-100ns-1</b>	5.5	None	65	1 <sup>st</sup>	57	4 <sup>th</sup>

**Table 3-10: Docking results for validation process 1 into models from the 1a trajectory.**

Regarding active compounds, pi-interactions involving different lipophilic parts of the ligands were evident in all the models. The piperazine ring of the active compounds was also involved in pi-interactions with active site residues, mainly TRP241 and TRP332. In models 65ns and 245ns, the lipophilic part of compounds **1a** and **1f** (65ns) and **1d** and **1f** (245ns) failed to locate in the hydrophobic region of TG2 in the complexes mentioned in Table 3-11. As it can be seen in Figure 3-48, this may be attributed to the orientation of the side chain of ASN333 which may be acting as a blocker to the access to the hydrophobic loop. The complex of **1b** is shown for comparison.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
<b>1a-65ns-16</b>	5.0	GLN276 ASN333 PHE334	52	12 <sup>th</sup>	75	1 <sup>st</sup>
<b>1b-65ns-7</b>	5.8	None	57	12 <sup>th</sup>	30	20 <sup>th</sup>
<b>1c-65ns-4</b>	4.9	GLN276 CYS277 PHE334	68	4 <sup>th</sup>	92	1 <sup>st</sup>
<b>1e-65ns-2</b>	2.9	CYS277	60	14 <sup>th</sup>	48	13 <sup>th</sup>
<b>1f-65ns-13</b>	5.1	CYS277 ASN333 PHE334 HIS335	75	2 <sup>nd</sup>	67	3 <sup>rd</sup>
<b>1b-155ns-10</b>	6.2	ARG317 ASN333 PHE334	65	3 <sup>rd</sup>	69	1 <sup>st</sup>
<b>1c-155ns-17</b>	4.7	ASN333 PHE334	76	1 <sup>st</sup>	89	2 <sup>nd</sup>
<b>1c-155ns-11</b>	5.4	GLN276	74	3 <sup>rd</sup>	94	2 <sup>nd</sup>
<b>1d-155ns-4</b>	5.2	GLN276 ASN333	58	11 <sup>th</sup>	64	6 <sup>th</sup>
<b>1e-155ns-3</b>	5.0	PHE334	74	1 <sup>st</sup>	73	2 <sup>nd</sup>
<b>1e-155ns-11</b>	5.2	PHE334	73	2 <sup>nd</sup>	77	1 <sup>st</sup>
<b>1b-245ns-20</b>	6.1	ARG317 ASN333 PHE334	63	1 <sup>st</sup>	66	3 <sup>rd</sup>
<b>1b-245ns-15</b>	5.6	ASN333 PHE334	61	3 <sup>rd</sup>	63	4 <sup>th</sup>
<b>1c-245ns-13</b>	4.9	TRP332 PHE334	66	1 <sup>st</sup>	84	1 <sup>st</sup>
<b>1d-245ns-15</b>	3.5	CYS277 ASN333	71	2 <sup>nd</sup>	76	1 <sup>st</sup>
<b>1e-245ns-20</b>	4.8	PHE334	64	4 <sup>th</sup>	61	6 <sup>th</sup>
<b>2b-245ns-4</b>	3.9	None	58	2 <sup>nd</sup>	71	1 <sup>st</sup>

Table 3-11: Docking results for validation process 1 into models from the 1b trajectory.

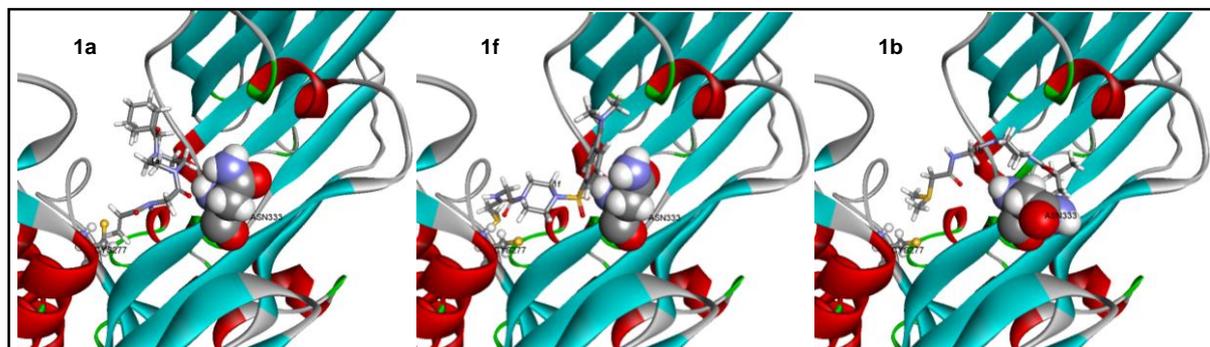


Figure 3-48: Poses of compounds 1a, 1b and 1f in model 65ns showing the role of ASN333 in blocking the hydrophobic region of TG2 active site. 1b is shown for comparison.

### 3.4.2 Validation process 2 (water docking)

GOLD is implemented with the capability of performing a docking with the presence of water molecules within the active site. The program can accommodate up to 25 water molecules and it gives the option to turn them on or off during the docking or to allow GOLD to decide whether a specific water molecule is turned on or off. There is also the option of keeping water molecules fixed or allowing them to spin about their axes (CCDC Software Limited 2013; Verdonk et al. 2005). For this experiment, GOLD was given the option to toggle which water molecules to be used. The molecules were also allowed to spin. The models were selected from the trajectory containing 25 closest water molecules to CYS277 at the same time points at which the water-free models were selected. Figure 3-49 shows model 90ns with the water molecules.

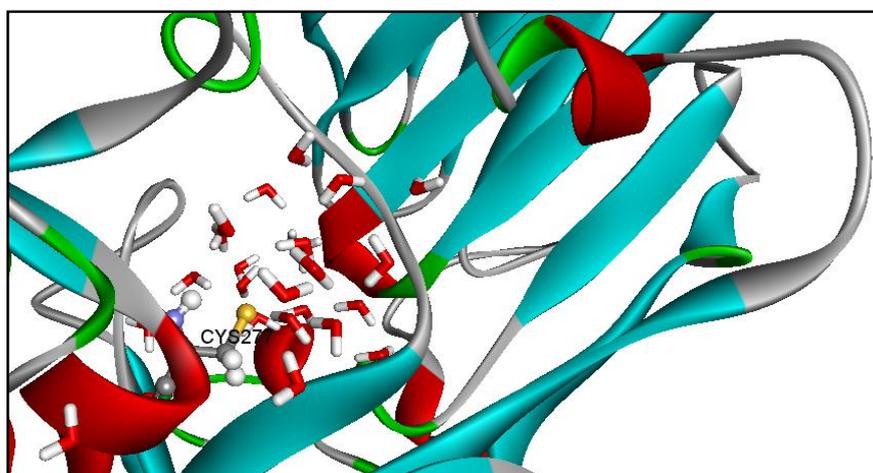


Figure 3-49: Model 90ns showing the water molecules in the active site.

#### 3.4.2.1 Models from the 1a trajectory

Compounds **1a**, **1b** and **2b** achieved good poses and ranks in all the 3 models, compound **1c** in the 100ns model, **1d** in the 90ns and the 100ns, **1e** in the 95ns and the 100ns, **1f** in the 95ns and **2c** in the 95ns model. The poses in general were similar to those obtained in validation process 1 with all the important interactions, except for compound **2b**, which either missed the interactions in the lipophilic region or within the catalytic tunnel. Compound **2c** did form a network of interactions within the active site but its rank within 95ns model was bad.

Surprisingly, in no complex for all the compounds were there hydrogen bonds involving water molecules. This happened despite the presence of a minimum of 4 water molecules and a maximum of 11 molecules chosen by GOLD to be involved in the docking (Figure 3-50). However, it is been explained that a reasonable binding mode may be found by GOLD that does not involve any interaction with key structural water molecules (Verdonk et al.

2005). Keeping this fact in mind, it will not be surprising to lack hydrogen bonds considering that these molecules were not structural; they were added by AMBER during MD simulations.

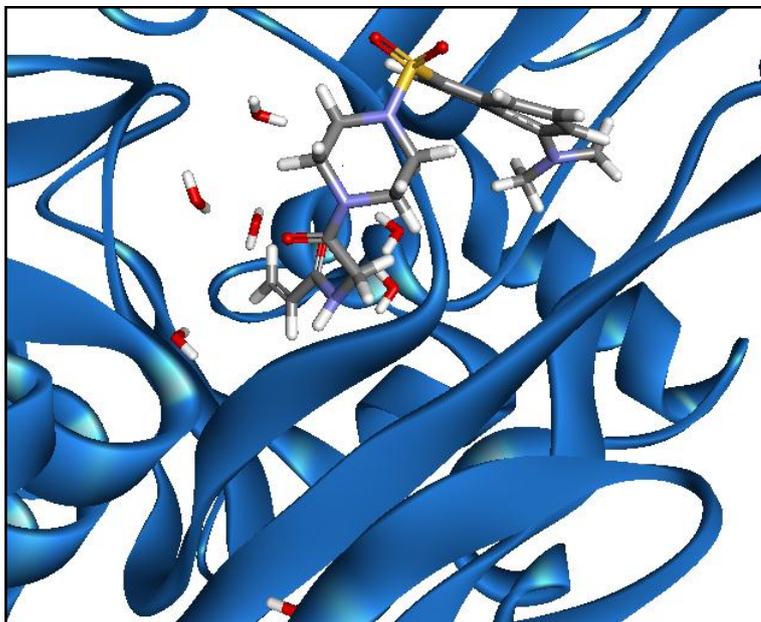


Figure 3-50: The seven water molecules chosen by GOLD for docking of compound **1e** in the 100ns model. None of them is close enough to allow for hydrogen bonding.

#### 3.4.2.2 Models from the **1b** trajectory

Similar results to what was seen with models from the **1a** trajectory were obtained on these models in terms of good poses and ranks in addition to the lack of hydrogen bonds with water molecules. However, Compound **1b** did form hydrogen bonds with water molecules in models 155ns and 245ns (Figure 3-51). Only **2c** of the inactive compounds achieved a good pose in 155ns model but the rank was bad. There was a minimum of no water molecules in some complexes and a maximum of 3 molecules in other complexes in models from the **1b** trajectory.

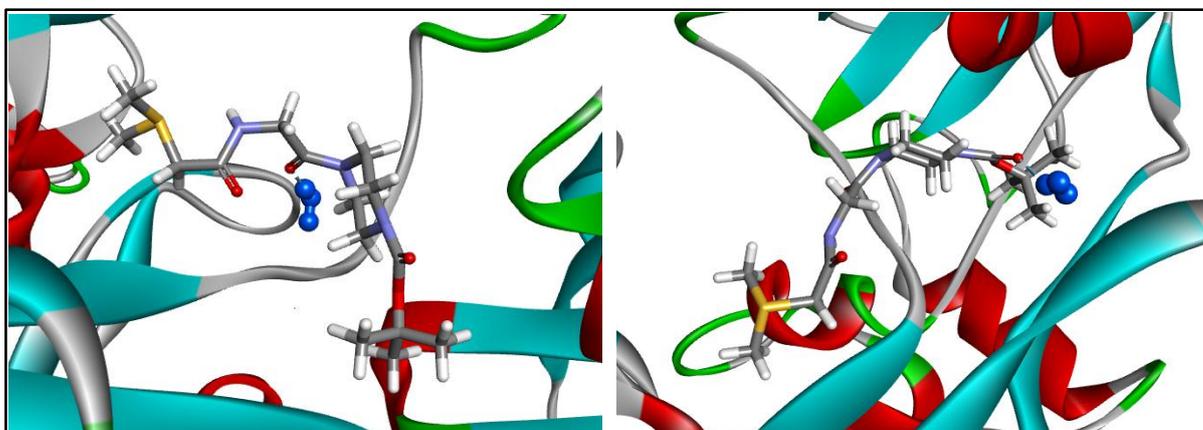
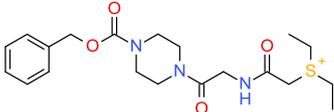
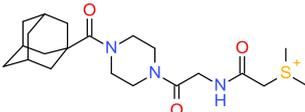
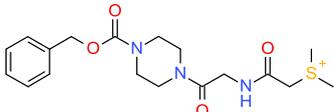
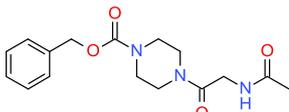
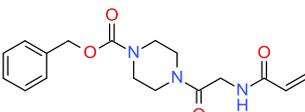


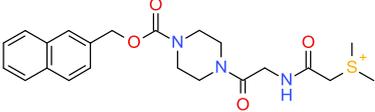
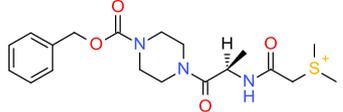
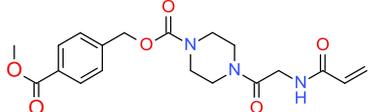
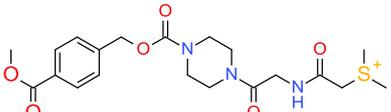
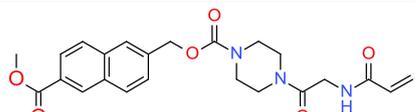
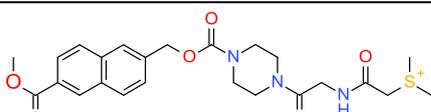
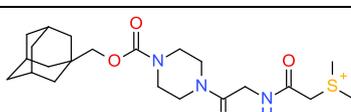
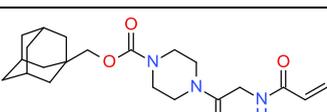
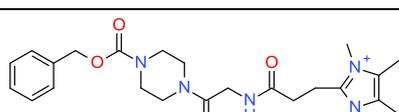
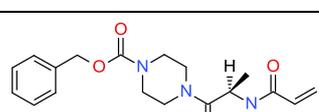
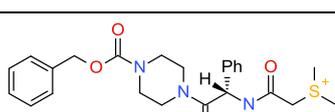
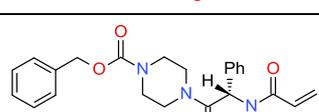
Figure 3-51: Hydrogen bonds between compound **1b** and water in models 155ns (left) and 245ns (right).

Although many of the 25 water molecules originally present in each of the 6 models were displaced during the docking, leaving between zero and 11 molecules, the displacement did not have a great effect on the score. In other words, there was not significant change in the scores of the inhibitors when there was a displacement of water molecules. The gain in energy upon displacement of water or the so-called desolvation effect is one of two mechanisms by which water can affect the docking of ligands into proteins. The other mechanism involves mediating hydrogen bonds with the ligands (Ladbury 1996). Since these 2 mechanisms did not prove to offer any advantage with the docking of TG2 inhibitors, the approach of water docking is probably not appropriate for TG2. This is also confirmed from the absence of any water molecules within the active site of TG2 in the original crystal structure [PDB 2Q3Z, (Pinkas et al. 2007)].

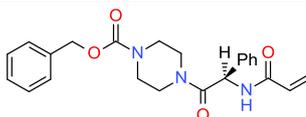
### 3.4.3 Validation process 3 (more test compounds)

Twenty additional compounds also adopted from Badarau et al. (2015) and the associated patent by Griffin et al. (2014) were used for this stage of testing. 15 of these compounds were active with  $IC_{50}$  values in the range of 0.006-6.8  $\mu$ M and 5 compounds were inactive with  $IC_{50}$  values above 100  $\mu$ M. The compound numbers, structures and  $IC_{50}$  values for these compounds are listed in Table 3-12.

Compound	Structure	$IC_{50}$ ( $\mu$ M)
3a		1.85
3b		0.89
3c		1.4
3d		0.008
3e		4.25
3f		5.925

3g		0.0059
3h		1.07
3i		6.8
3j		6.3
3k		3.3
3l		2.1
3m		1.5
3n		0.775
3o		1.625
4a		>100
4b		>100
4c		>100
4d		>100

4e



&gt;100

Table 3-12: Active and inactive compounds used in validation process 3.

### 3.4.3.1 Models from the 1a trajectory

A summary of the results of the docking of compounds used in validation process 3 is presented in Table 3-13. All of the presented complexes have achieved the first criterion of a good docking complex which is the bent conformation within TG2 active site with the warhead pointing in the direction of CYS277.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank	Lipo*
<b>3b</b> -90ns-5	5.2	GLN276 ASN318	52	1 <sup>st</sup>	49	6 <sup>th</sup>	Yes
<b>3c</b> -90ns-20	5.6	ASN333	49	4 <sup>th</sup>	46	8 <sup>th</sup>	Yes
<b>3d</b> -90ns-6	3.7	ASN333 PHE334	42	3 <sup>rd</sup>	49	1 <sup>st</sup>	Yes
<b>3e</b> -90ns-1	4.0	GLN276 ASN318 ASN333 PHE334	42	11 <sup>th</sup>	52	4 <sup>th</sup>	Yes
<b>3h</b> -90ns-10	3.7	GLN276 ASN333	54	3 <sup>rd</sup>	62	1 <sup>st</sup>	No
<b>3n</b> -90ns-14	6.2	ASN333	44	19 <sup>th</sup>	52	10 <sup>th</sup>	Yes
<b>3o</b> -90 ns-3	3.8	TRP241 GLN276 CYC277 ASN318	54	1 <sup>st</sup>	57	5 <sup>th</sup>	Yes
<b>4b</b> -90ns-8	3.0	None	50	4 <sup>th</sup>	34	17 <sup>th</sup>	Yes
<b>4c</b> -90ns-7	5.8	GLN276	63	2 <sup>nd</sup>	44	17 <sup>th</sup>	No
<b>4d</b> -90ns-18	3.4	ASN333	52	10 <sup>th</sup>	39	14 <sup>th</sup>	Yes
<b>4e</b> -90ns-1	3.4	GLN276 ASN333	48	19 <sup>th</sup>	39	20 <sup>th</sup>	Yes
<b>3a</b> -95ns-19	6.6	TRP332 ASN333	52	6 <sup>th</sup>	38	16 <sup>th</sup>	Yes
<b>3b</b> -95ns-12	5.7	None	57	2 <sup>nd</sup>	67	1 <sup>st</sup>	Yes
<b>3d</b> -95ns-7	6.3	GLN276 TRP332 ASN333	44	1 <sup>st</sup>	33	10 <sup>th</sup>	Yes
<b>3e</b> -95ns-16	3.6	GLN276 ASN333	46	3 <sup>rd</sup>	60	1 <sup>st</sup>	No
<b>3f</b> -95ns-7	4.3	GLN276 ASN333	46	1 <sup>st</sup>	23	18 <sup>th</sup>	No
<b>3g</b> -95ns-3	3.2	GLN276 ASN318 PHE334	56	2 <sup>nd</sup>	23	16 <sup>th</sup>	Yes
<b>3i</b> -95ns-10	5.7	GLN276	55	1 <sup>st</sup>	30	12 <sup>th</sup>	Yes
<b>3n</b> -95ns-6	6.3	ASN333	56	2 <sup>nd</sup>	47	15 <sup>th</sup>	Yes

<b>3o-95ns-1</b>	4.1	ASN276 PHE334	52	2 <sup>nd</sup>	63	4 <sup>th</sup>	Yes
<b>4b-95ns-16</b>	3.5	GLN276 ASN333	50	1 <sup>st</sup>	59	1 <sup>st</sup>	No
<b>4c-95ns-3</b>	5.4	ASN333	57	8 <sup>th</sup>	32	17 <sup>th</sup>	Yes
<b>4d-95ns-12</b>	3.5	TRP241	48	13 <sup>th</sup>	33	18 <sup>th</sup>	Yes
<b>4e-95ns-19</b>	3.7	GLN276	56	1 <sup>st</sup>	42	13 <sup>th</sup>	No
<b>3a-100ns-7</b>	6.5	TRP241 ASN333	56	1 <sup>st</sup>	45	8 <sup>th</sup>	Yes
<b>3b-100ns-16</b>	6.8	None	52	2 <sup>nd</sup>	64	5 <sup>th</sup>	Yes
<b>3d-100ns-10</b>	4.4	ASN333	42	2 <sup>nd</sup>	49	3 <sup>rd</sup>	Yes
<b>3e-100ns-15</b>	3.2	GLN276 TRP332 ASN333 PHE334	46	2 <sup>nd</sup>	46	13 <sup>th</sup>	Yes
<b>3f-100ns-13</b>	3.4	GLN276 ASN332 PHE334	47	3 <sup>rd</sup>	50	9 <sup>th</sup>	Yes
<b>3g-100ns-6</b>	3.7	PHE334	54	2 <sup>nd</sup>	59	1 <sup>st</sup>	Yes
<b>3i-100ns-20</b>	5.6	None	59	1 <sup>st</sup>	54	5 <sup>th</sup>	Yes
<b>3j-100ns-10</b>	3.9	GLN276	42	9 <sup>th</sup>	12	19 <sup>th</sup>	Yes
<b>3n-100ns-19</b>	5.2	GLN276 ASN333	54	2 <sup>nd</sup>	75	1 <sup>st</sup>	Yes
<b>3o-100ns-7</b>	4.4	TRP241	50	5 <sup>th</sup>	74	1 <sup>st</sup>	Yes
<b>4b-100ns-6</b>	3.4	GLN276	51	1 <sup>st</sup>	52	5 <sup>th</sup>	Yes
<b>4d-100ns-7</b>	3.1	None	43	16 <sup>th</sup>	57	5 <sup>th</sup>	Yes
<b>4e-100ns-17</b>	4.1	None	56	2 <sup>nd</sup>	50	8 <sup>th</sup>	Yes

**Table 3-13: Docking results of validation process 3 into models from the 1a trajectory.\* “Lipo” column in this table and the next ones determines the presence “Yes” or absence “No” of interactions between the lipophilic part of the inhibitors and the hydrophobic loop of TG2.**

A closer look at the distances in Table 3-13 shows that compounds with sulfonium warheads (**3a**, **3b**, **3c**, **3h**, **3i**, **3k**, **3n**, and **4c**) achieved SG-warhead distances that are generally larger than 5 Å. This can be attributed to the physical size of the warhead in these compounds. For the 90ns model, 7 active compounds out of the 15 gave good poses, 5 of which passed the 4 criteria for a good docking complex. Compound **3h** did not have any interactions with the hydrophobic loop of TG2 and compound **3n** failed to rank high in any of the scoring functions. In the same model, 4 of the 5 inactive compounds posed well in the active site but none of them passed the 4 criteria; **4b** with no hydrogen bonds, **4c** with no lipophilic interactions and **4d** and **4e** by failing to achieve high ranks in either scoring function.

For the 95ns model, 9 active compounds achieved good poses but only 5 passed all the validation criteria; compound **3a** was not ranked high, compounds **3e** and **3f** did not show interactions in the hydrophobic loop and compound **3b** had no hydrogen bonds with TG2. The latter did have pi-cation interactions between the sulfonium ion and the side chains of TRP241 and TRP332 (Figure 3-52A). This type of interaction was also observed in **3i** and **3n**

and may further contribute to directing the warhead to CYS277. Four of the 5 inactive compounds had good poses but, as in 90ns model, none passed all the criteria. In the 100ns model, 10 active compounds had good poses within the active site of TG2 and 7 of them passed all the validation criteria. **3b** and **3i** did not have hydrogen bonds but had pi-cation interactions with TRP241 which leaves only **3i** as a failing active compound. Only 3 inactive compounds posed well and only one of them managed to pass all the criteria with the other two missing the hydrogen bonds. Figure 3-52B is an example pose for compound **3o** for this group of models.

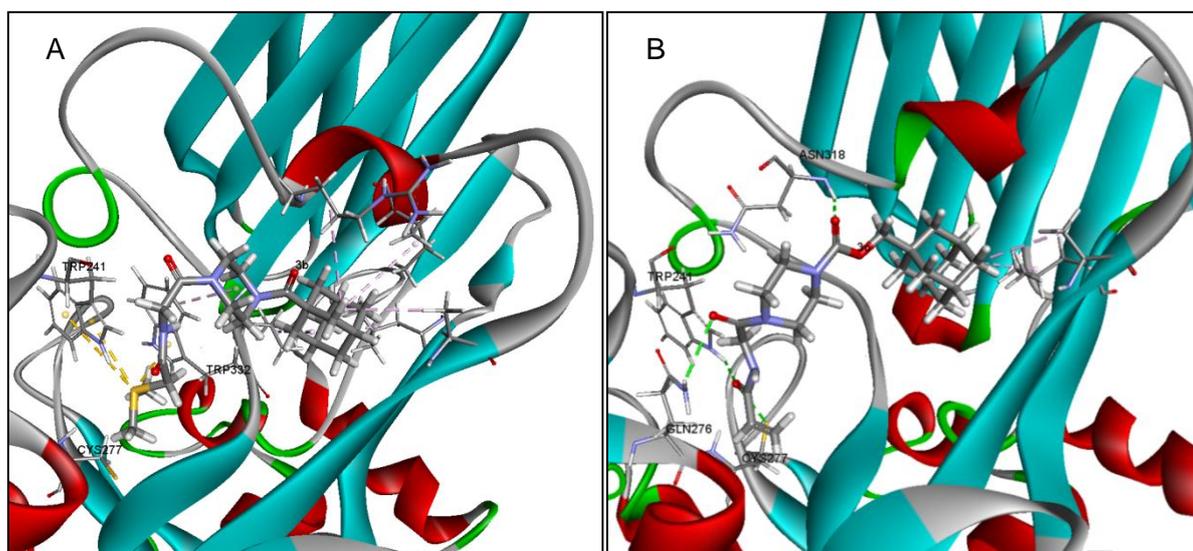


Figure 3-52: A: Pose of compound **3b** in 95ns model showing the pi-sulphur interaction. B: Pose of compound **3o** in 90ns model as an example.

### 3.4.3.2 Models from the 1b trajectory

A summary of docking results of all the compounds is presented in Table 3-14. Again, in all the complexes in the table, the compounds have achieved the criterion of the appropriate bent pose within the active site of TG2. The SG-warhead distances for compounds having sulfonium ion warheads are also generally larger than those of acrylamide compounds. For the 65ns model, 7 out of 15 active compounds gave good poses. Three of the 7 passed all the validation criteria and 4 did not, by either not having hydrogen bonds or not ranking high in the scoring functions. Three inactive compounds achieved good poses in this model but none passed the 4 criteria. In fact, all of them passed only 2 and failed with 2 criteria.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank	Lipo
<b>3c-65ns-13</b>	6.5	ASN333 PHE334	72	1 <sup>st</sup>	48	16 <sup>th</sup>	Yes
<b>3d-65ns-20</b>	6.6	ASN333 PHE334	57	2 <sup>nd</sup>	45	11 <sup>th</sup>	Yes
<b>3f-65ns-1</b>	5.3	None	62	5 <sup>th</sup>	66	3 <sup>rd</sup>	Yes

<b>3i-65ns-8</b>	4.7	GLN276 CYS277	68	5 <sup>th</sup>	66	7 <sup>th</sup>	Yes
<b>3j-65ns-2</b>	5.7	GLN276 CYS277 ASN333	82	1 <sup>st</sup>	95	1 <sup>st</sup>	Yes
<b>3m-65ns-11</b>	6.3	TRP332	65	17 <sup>th</sup>	69	12 <sup>th</sup>	Yes
<b>3o-65ns-11</b>	3.6	None	67	1 <sup>st</sup>	63	6 <sup>th</sup>	Yes
<b>4a-65ns-6</b>	3.3	GLN276 ASN333	62	15 <sup>th</sup>	75	5 <sup>th</sup>	No
<b>4b-65ns-16</b>	3.2	ASN333 PHE334	61	10 <sup>th</sup>	67	9 <sup>th</sup>	Yes
<b>4d-65ns-5</b>	4.9	None	69	3 <sup>rd</sup>	66	9 <sup>th</sup>	No
<b>3a-155ns-8</b>	7.1	ARG317 ASN333 PHE334	71	3 <sup>rd</sup>	62	12 <sup>th</sup>	Yes
<b>3c-155ns-19</b>	3.9	TRP241 HIS335	62	7 <sup>th</sup>	9	19 <sup>th</sup>	Yes
<b>3d-155ns-5</b>	4.5	ARG317 ASN333 PHE334	58	1 <sup>st</sup>	75	1 <sup>st</sup>	Yes
<b>3e-155ns-14</b>	4.3	TRP332	57	9 <sup>th</sup>	57	9 <sup>th</sup>	Yes
<b>3f-155ns-19</b>	4.6	None	59	8 <sup>th</sup>	72	2 <sup>nd</sup>	Yes
<b>3g-155ns-12</b>	5.1	PHE334	73	4 <sup>th</sup>	71	7 <sup>th</sup>	Yes
<b>3m-155ns-14</b>	5.2	CYS277 ARG318 HIS335	76	4 <sup>th</sup>	80	3 <sup>rd</sup>	Yes
<b>4b-155ns-12</b>	5.0	None	68	2 <sup>nd</sup>	18	20 <sup>th</sup>	Yes
<b>4d-155ns-20</b>	4.9	ASN333	57	12 <sup>th</sup>	65	8 <sup>th</sup>	Yes
<b>4e-155ns-2</b>	5.2	ASN333	56	17 <sup>th</sup>	41	20 <sup>th</sup>	No
<b>3d-245ns-18</b>	5.2	CYS277 ASN333	57	1 <sup>st</sup>	60	3 <sup>rd</sup>	Yes
<b>3e-245-18</b>	3.6	CYS277 ASN333 HIS335	61	3 <sup>rd</sup>	34	17 <sup>th</sup>	Yes
<b>3f-245ns-2</b>	4.3	PHE334	62	2 <sup>nd</sup>	33	20 <sup>th</sup>	No
<b>3g-245ns-19</b>	2.9	TRP241	58	11 <sup>th</sup>	45	16 <sup>th</sup>	Yes
<b>3h-245ns-9</b>	4.2	GLN276 PHE334	79	2 <sup>nd</sup>	70	9 <sup>th</sup>	Yes
<b>3i-245ns-2</b>	4.6	None	63	7 <sup>th</sup>	81	1 <sup>st</sup>	No
<b>3j-245ns-15</b>	6.3	CYS277 TRP332 ASN333 HIS335	67	2 <sup>nd</sup>	65	4 <sup>th</sup>	Yes
<b>3k-245ns-5</b>	3.4	CYS277 ARG317	73	3 <sup>rd</sup>	77	2 <sup>nd</sup>	Yes
<b>3o-245ns-1</b>	3.8	PHE334	63	2 <sup>nd</sup>	65	9 <sup>th</sup>	Yes
<b>4b-245ns-18</b>	4.9	GLN276	61	3 <sup>rd</sup>	76	1 <sup>st</sup>	Yes

Table 3-14: Docking results of validation process 3 into models from the 1b trajectory.

In the 155ns model, 7 active compounds and 3 inactive compounds achieved the required pose within TG2 active site. 4 active compounds succeeded in passing the 4 criteria for a good docking complex and 3 active compounds passed only 3 criteria missing either the hydrogen bonds or the top ranking in the scoring functions. The 3 inactive compounds failed to pass the 4 criteria missing the hydrogen bonds, top ranks and/or the interactions involving the lipophilic part of the inhibitor. The performance of model 245ns was the best in this group, where 9 active compounds and only 1 inactive compound passed the good pose criterion. 6 active compounds passed the 4 criteria, 2 passed 3 criteria and one compound (**3i**) passed only 2. Compound **3i** lacked interactions with the lipophilic part and hydrogen bonds but there was a pi-sulphur bond between the sulfonium ion and PHE280 that may compensate for a lack of hydrogen bonds. The only one inactive compound that posed well, however, has passed the 4 criteria for a good docking complex.

It should also be mentioned that in most complexes obtained in this process, there were pi interactions involving the piperazine ring in the centre of the inhibitors. TG2 residues on the other side of these interactions were mainly the bridging tryptophan residues, TRP241 and TRP332 (Figure 3-53). These interactions were present in almost all the complexes except those from model 90ns of the **1a** trajectory. These pi-interactions are weaker than the hydrogen bonds, but they can contribute to maintaining the bent conformation of the inhibitor required for its activity.

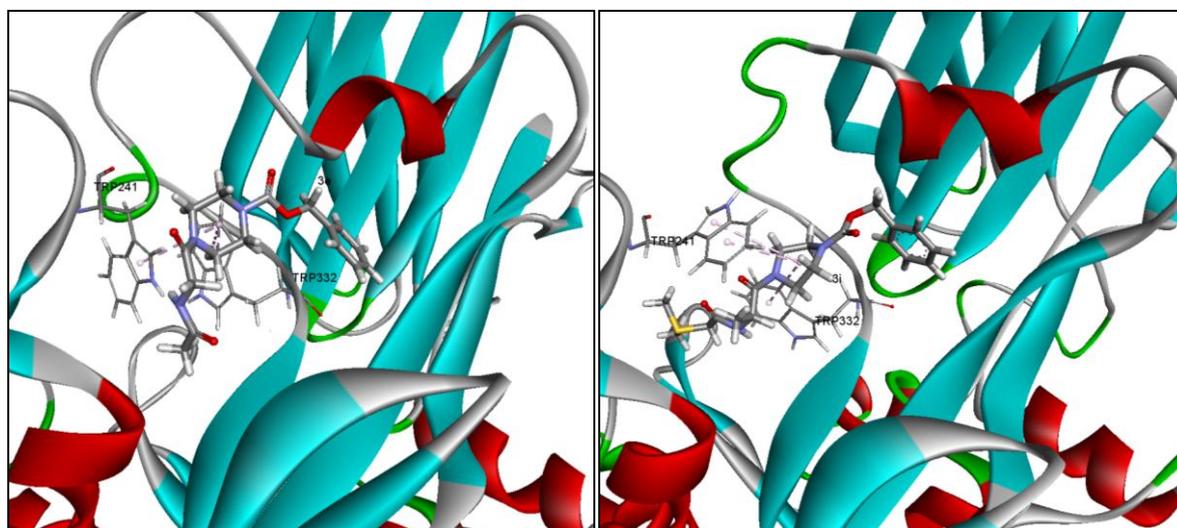


Figure 3-53: Poses of compounds **3e** (left) and **3i** (right) in the 95ns and 245ns models respectively, showing the pi interactions with the piperazine ring.

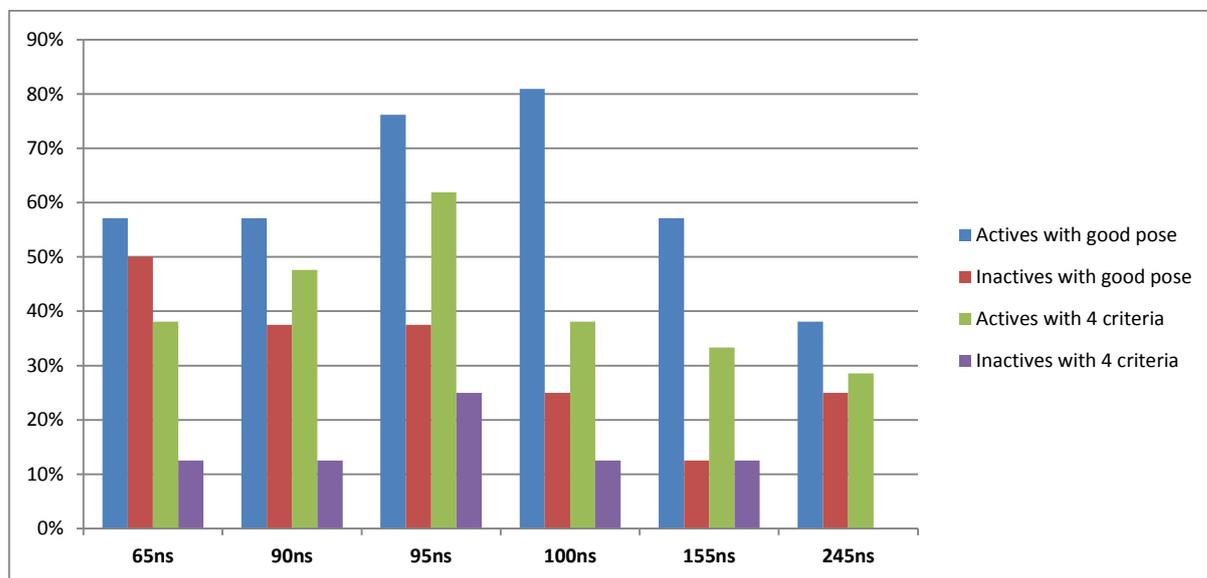
#### 3.4.4 Validation process 4 (all compounds)

A total of 21 active and 8 inactive compounds were docked into the 6 models using the default settings in GOLD with flexible treatment of the selected 10 active site residues. It was

performed as a triplicate. The rationale behind this experiment was to assess the performance of the models in testing such a comparably large set of TG2 inhibitors. The performance of each of the model will not be discussed in detail as it has been presented previously; rather a more concise presentation will be used. The docking attempt that gave the best results was considered, even if that was different for each model. The best result, rather than the average, was considered because GOLD uses a genetic algorithm as the docking method (Jones et al. 1997) and this is a stochastic method that works by generating a set of random initial solutions, scoring them and then modifying and scoring accordingly. Table 3-15 contains a summary of the percentages of active and inactive compounds that achieved good poses and that passed the 4 validation criteria in the best docking attempt for each model. The same information is also presented in Figure 3-54.

Model	% Actives / good pose	% Inactives / good pose	% Actives / 4 criteria	% Inactives / 4 criteria	Docking Attempt
65ns	57.14	50	38.10	12.5	3 <sup>rd</sup>
90ns	57.14	37.5	47.62	12.5	1 <sup>st</sup>
95ns	76.19	37.5	61.90	25	1 <sup>st</sup>
100ns	80.95	25	38.10	12.5	2 <sup>nd</sup>
155ns	57.14	12.5	33.33	12.5	1 <sup>st</sup>
245ns	38.10	25	28.57	0	2 <sup>nd</sup>

**Table 3-15: Performance of the 6 active site models in validation process 4. Percentages are approximated to the nearest hundredth.**



**Figure 3-54: A graphical presentation for information from Table 3-15.**

With the exception of 245ns, all the models scored more than 50% in capturing the bent conformation and it was model 95ns that performed the best in terms of achieving the 4 criteria. This was true for active as well as inactive compounds. Although model 245ns did not score high for good poses, no inactive compounds passed the 4 criteria in this model.

Model 90ns had the lowest difference between the percentage of active compounds with good pose and those that passed the 4 criteria. This would make 90ns suitable choice for docking after performing some virtual screening on 95ns and 100ns models.

The scores from GoldScore and CHEMPLP in this experiment have been analysed. The best score for each of the active compounds in every model for the 2 scoring functions have been plotted against the IC<sub>50</sub> values for the compound. No useful correlation could be obtained for all the models; not when IC<sub>50</sub> was plotted against the scores, nor when the logarithm of IC<sub>50</sub> was used, nor when the scores were divided by the molecular weight (to normalise the pairwise interactions during the scoring process). The R<sup>2</sup> value did not exceed 0.1 in any case. Furthermore, the scores for the inactive compounds were not very much different from those of active compounds; sometimes they were even better. Table 3-16 presents the best scores obtained for active and inactive compounds in model 95ns from this experiment. Figure 3-55 shows the correlation between biological activity and the scores of GoldScore and CHEMPLP for 95ns model active compounds as an example.

Compound	IC <sub>50</sub> (μM)	GoldScore	CHEMPLP
<b>1a</b>	0.125	52	57
<b>1b</b>	0.273	53	46
<b>1c</b>	0.44	52	60
<b>1d</b>	0.7	50	11
<b>1e</b>	0.0061	59	44
<b>1f</b>	0.38	60	22
<b>3c</b>	1.4	54	52
<b>3d</b>	0.008	44	27
<b>3e</b>	4.25	45	50
<b>3f</b>	5.925	44	60
<b>3g</b>	0.0059	59	26
<b>3h</b>	1.07	50	62
<b>3i</b>	6.8	60	28
<b>3j</b>	2.1	50	38
<b>3n</b>	0.775	64	62
<b>3o</b>	1.625	56	56
<b>2c</b>	100	64	16
<b>4b</b>	100	50	51
<b>4d</b>	100	42	35
<b>4e</b>	100	59	13

**Table 3-16: Scores of the active and inactive compounds in model 95ns from validation process 4.**

The fact that the activity of these inhibitors requires covalent bonding with TG2 is probably the reason behind these poor correlations. GOLD was able to produce the bent conformation

and to rank the solutions of individual compounds based on this conformation along with hydrogen bonding and lipophilic interactions but the program was unable to rank the compounds based on their activities.

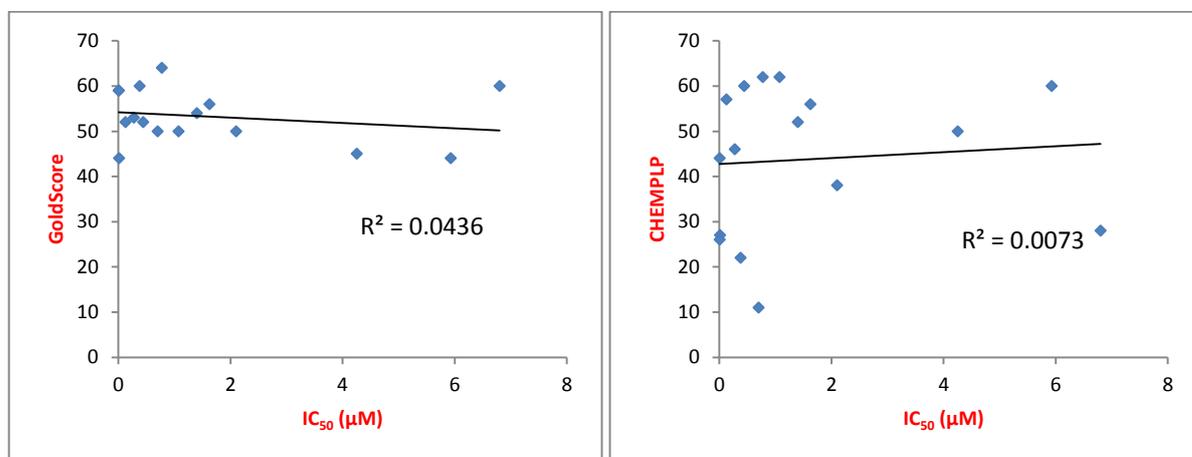


Figure 3-55: Correlation between scores for active compounds and biological activity in 95ns model, left is GoldScore and right is CHEMPLP.

### 3.4.5 Validation process 5 (5-ns MD simulations)

The rationale behind this experiment was to ensure that the different poses obtained during various docking experiments were valid. In other words, to show whether the active compounds will retain their good poses and whether a good pose within the active site for inactive compounds will change after 5 nanoseconds of a MD simulation.

The docking complexes used in this experiment were taken from the primary docking study during the selection of the valid models (section 3.3) and from validation processes 1, 2, 3 and 4. Similar settings for the previous MD simulations were used at this stage; the antechamber program was used first, and then minimisation and 3-stage MD simulation were applied. Only the production phase (MD2\_fast) was taken to investigation. The length of this stage was set to 5 nanoseconds. A similar approach has been used by Badarau et al. (2013).

The results will be presented in the form of separate tables for each valid model. The tables will contain, in addition to the complex name, a code for the performance of the compound during the MD. **1** is a simulation for a compound in which both the warhead and the lipophilic part maintained their initial positions during the simulation. **2** is a simulation in which the warhead has left its place next to CYS277 and **3** is a simulation in which the lipophilic part has left the hydrophobic loop of TG2. The distance between the electrophilic carbon and SG of CYS277 will be presented in the starting conformation and its average during the MD simulation. This will not be presented for simulations in which the lipophilic part has left the

hydrophobic loop. Hydrogen bonds will be presented for the starting conformation and during the MD simulation as the interacting residues and the percentage of time for their persistence.

### 3.4.5.1 Models from the trajectory of 1a

Table 3-17, Table 3-18 and Table 3-19 contain the results of validation process 5 on the 3 models from the trajectory of 1a.

Complex	Performance	Warhead position (Å)		Hydrogen bonds		
		Starting	MD average	Starting conformation	MD simulation	
					Residue	%
1a-90ns-8	2	3.5	15.4			
1b-90ns-4	1	6.7	5.7	GLN276 ASN333	ASN318	3.5
					ASN333	2
					PHE334	13
1d-90ns-5	1	5.3	5.6	None	ASN333	10
					PHE334	4
1e-90ns-17	1	3.1	5.8	GLN276	ARG317	0.5
					ASN333	1.5
					PHE334	2
3b-90ns-5	1	5.2	6.2	GLN276	ASN318	1
				ASN318	ASN333	29
3c-90ns-20	1	5.6	6.3	ASN333	ASN318	4
					ASN333	1.5
3d-90ns-6	1	3.5	6.1	ASN333	GLN276	11
				PHE334	ASN333	11
					PHE334	2.5
3e-90ns-10	2	3.7	12			
3f-90ns-2	1	3.6	3.7	GLN276		
				ASN333	ASN333	34
				PHE334		
3h-90ns-10	3					
3j-90ns-1	2	6.1	10.3			
3o-90ns-3	1	3.8	5.8	TRP241		
				GLN276	ASN318	10
				CYS277	ASN333	43
				ASN318		
2b-90ns-8	2	3.3	10.6			
4b-90ns-8	2	3	6.5			

<b>4c-90ns-2</b>	<b>3</b>					
<b>4d-90ns-19</b>	<b>1</b>	3.4	6.2	ASN333	GLN276	15

Table 3-17: Results of validation process 5 on the 90ns model of the trajectory of 1a.

Complex	Performance	Warhead position (Å)		Hydrogen bonds		
		Starting	MD average	Starting conformation	MD simulation	
					Residue	%
<b>1a-95ns-10</b>	<b>1</b>	4.6	4	TRP332	GLN276	18
<b>1b-95ns-6</b>	<b>1</b>	6	6.3	ASN333	ASN333	21
<b>1c-95-2</b>	<b>1</b>	4.9	4.3	CYS277	TRP241	9.5
				PHE334	GLN276	25
					ASN333	1.5
<b>1d-95ns-4</b>	<b>1</b>	6.6	5.5	179	PHE334	29
<b>1e-95ns-17</b>	<b>1</b>	3.5	5.5	GLN276	None	
<b>1f-95ns-20</b>	<b>1</b>	5.8	6.1	ASN318	ASN333	53
					GLN276	88
					ASN318	34
<b>3a-95ns-19</b>	<b>1</b>	6.6	7.8	TRP332	ASN333	17
<b>3b-95ns-16</b>	<b>1</b>	6.6	6.1	None	GLN276	3.5
<b>3c-95ns-17</b>	<b>1</b>	5.8	6.2	GLN276	ARG317	9.5
				ASN333	TRP332	13
					ASN333	5.5
<b>3d-95ns-1</b>	<b>2</b>	6.3	Left during equilibration			
<b>3e-95ns-10</b>	<b>2</b>	3.7	8.5			
<b>3f-95ns-11</b>	<b>1</b>	4.1	3.7	ASN333	GLN276	41
<b>3g-95ns-8</b>	<b>2</b>	3.2	10.4			
<b>3i-95ns-5</b>	<b>1</b>	5.7	8.1	ASN333	GLN324	37
<b>3l-95ns-19</b>	<b>2</b>	4.1	7.7			
<b>3n-95ns-6</b>	<b>1</b>	6.3	6.6	ASN333	GLN276	5.5
					ARG317	11
<b>3o-95ns-1</b>	<b>1</b>	4.1	4.6	GLN276	GLN276	13
				PHE334	ASN318	8
<b>2c-95ns-3</b>	<b>1</b>	5.6	6.5	None	ARG317	1
<b>4b-95ns-12</b>	<b>2</b>	4.0	8.7			
<b>4e-95ns-19</b>	<b>2</b>	3.7	5.8			

Table 3-18: Results of validation process 5 on the 95ns model of the trajectory of 1a.

Complex	Performance	Warhead position (Å)		Hydrogen bonds		
		Starting	MD average	Starting conformation	MD simulation	
					Residue	%
1a-100ns-5	1	5.3	4.1	None	GLN324	4.5
					ASN333	1
1b-100ns-6	1	5.7	6.3	ASN333	ASN318	3.5
					ASN333	5
1c-100ns-6	1	4.0	3.7	GLN276	GLN276	5.5
					ASN333	16
1d-100ns-16	1	5.5	3.6	None	ASN333	22
1e-100ns-3	2	3.9	6.5			
1f-100ns-4	1	5.1	5.8	GLN276	ARG317	0.5
					ASN333	1.5
					PHE334	2
3a-100ns-17	1	7.3	7.3	GLN324	GLN276	3.5
					GLN324	2.5
3b-100ns-19	1	5.8	5.9	GLN276	ASN318	46
3c-100ns-1	2	6.7	8.1			
3d-100ns-7	1	4.2	6.4	GLN276		
				ASN333	ASN333	46
				PHE334		
3e-100ns-15	1	3.2	6.1	TRP241		
				GLN276	ASN333	19
				ASN333	PHE334	6
				PHE334		
3f-100ns-13	1	3.4	3.9	GLN276		
				ASN333	GLN276	49
				PHE334		
3g-100ns-6	1	3.7	3.8		GLN276	17
				PHE334	GLN324	4
					PHE334	7
3i-100ns-10	1	5.8	6.5	GLN324	ASN333	37
3j-100ns-10	2	3.9	12.4			
3n-100ns-19	1	5.2	5.7	GLN276	ASN333	41
				ASN333	PHE334	1.5
2c-100ns-11	3					
4b-100ns-6	1	3.4	3.8	GLN276	GLN276	60
4e-100ns-2	2	3.7	5.5			

Table 3-19: Results of validation process 5 on the 100ns model of the trajectory of 1a.

In the 90ns model, a total of 16 simulations were run, 12 for active compounds and 4 for inactive compounds. Of those, 8 active compounds and 1 inactive compound maintained their initial pose by the end of the simulation. For 95ns model, there were simulations for 17 active compounds and 3 inactive compounds from which 13 active and 1 inactive compounds managed to pass the simulation (performance code 1). In the 100ns model, there were 16 active compounds and 3 inactive compounds at the start. 13 active compounds and only one inactive compound finished the simulation successfully (performance code 1). In general, the three models performed well in terms of maintaining the pose of the active compounds within the active site of TG2 for 5 nanoseconds. At the same time, inactive compounds, for the most part, failed to stay in the active site.

When the hydrogen bonds are considered, it was noted that ASN333 was by far the most common in this sense. This was seen in the 3 models from the trajectory of **1a**. That only confirms the importance of this residue in positioning the inhibitors within TG2 active site and in maintaining the correct orientation. The amide group in the side chain has the ability to act as a hydrogen bond donor and acceptor (Johansson et al. 1974) and thereby increases the chance for the formation of hydrogen bonds with suitable groups on the other side. The location of this residue within the TG2 active site gives it an excellent opportunity to govern the orientation of the inhibitor (Figure 3-56).

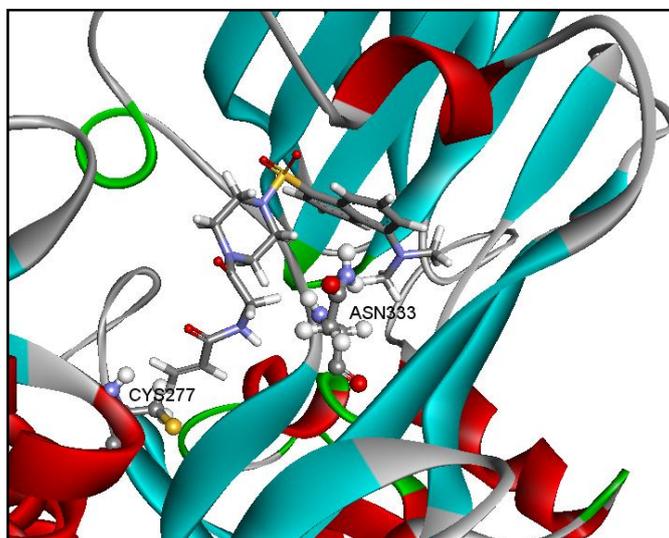


Figure 3-56: The location of residue ASN333 relative to the active site of TG2 and showing a docked inhibitor and CYS277 for reference.

The residue is located at the bridge between the hydrophobic loop, that accommodates the lipophilic part of TG2 inhibitors, and the catalytic tunnel that contains CYS277 and in the bent conformation of the inhibitors; ASN333 will be conveniently located opposite to the piperazine ring in the middle of all the compounds used so far in this work. There are 2 carbonyl groups on either side of the piperazine ring in most of the compounds (all compounds except **1e**, **1f** and **3g**) and these carbonyl groups were involved in the hydrogen

bonds with ASN333. This did not prevent ASN333 from forming hydrogen bonds with other potential groups in the inhibitor molecules that are closer to the warhead.

GLN276 and PHE334 have also been involved in high percentage of hydrogen bonds recorded during these simulations. GLN276 operates on the warhead end of the inhibitor while PHE334 formed hydrogen bonds with the central and the warhead parts of the inhibitors. There were also hydrogen bonds with ARG317 and ASN318 to a lesser extent. The latter formed hydrogen bonds mainly involving the central part of the inhibitors from the side of the lipophilic warhead.

Of all the compounds, whether active or inactive, that failed to maintain their initial positions during the simulations, only in 3 compounds; this failure was manifested by the exit of the lipophilic part of the inhibitor outside the hydrophobic loop of TG2. This may indicate that, with the correct lipophilic parts, it is easier to keep this end of the inhibitor in place than to keep the warhead in the catalytic tunnel. It also indicates that for all the tested compounds, the various lipophilic parts were appropriate to serve the function of correctly posing TG2 inhibitors. The fact that these inhibitors act by forming covalent bonds should not be forgotten; possibly if these simulations were able to account for the formation of covalent bonds, a different performance of the warheads may have been observed.

### 3.4.5.2 Models from the trajectory of 1b

The results of validation process 5 on models from this trajectory are presented in Table 3-20, Table 3-21 and Table 3-22. The performance of these models was worse than the models of the trajectory of 1a. This was especially true for the 245ns model in which 15 simulations were run for active compounds and only 6 maintained their initial positions after the simulation. The 155ns model was the 2<sup>nd</sup> worst after 245ns and the 65ns model was the best after 8 of the 11 tested active compounds succeeded in maintaining the orientation required for activity against TG2. Of the 3 inactive compounds, 2 left the active site in the 65ns model.

Complex	Performance	Warhead position (Å)		Hydrogen bonds		
		Starting	MD average	Starting conformation	MD simulation	
					Residue	%
1b-65ns-18	1	6.0	5.8	ASN333	ASN333	0.5
				PHE334	PHE334	5.5
1c-65ns-4	1	4.9	4.7	GLN276 CYS277	None	

				PHE334		
<b>1e-65ns-20</b>	<b>1</b>	4.5	5.1	GLN276		
				ASN333	GLN376	32
				PHE334	ASN333	16
<b>1f-65ns-8</b>	<b>1</b>	6.2	6.3	ASN333	ARG317	7.5
					ASN333	3
					PHE334	3
<b>3a-65ns-8</b>	<b>3</b>					
<b>3c-65ns-13</b>	<b>1</b>	6.5	6.1	ASN333	TRP241	33
				PHE334	ASN333	1
<b>3d-65ns-4</b>	<b>2</b>	7.3	8.5			
<b>3i-65ns-8</b>	<b>1</b>	4.7	5.6	GLN276	GNL276	7.5
				CYS277	ARG317	5
<b>3j-65ns-2</b>	<b>1</b>	5.7	3.8	GLN276	GLN276	1.5
				CYS277	ASN333	44
				ASN333		
<b>3m-65ns-2</b>	<b>1</b>	6.2	6.9	CYS277	ASN333	1.5
				ASN333	ARG317	0.5
				PHE334		
<b>3o-65ns-13</b>	<b>2</b>	4.8	8.6			
<b>2a-65ns-17</b>	<b>2</b>	5.4	8.7			
<b>2c-65ns-9</b>	<b>1</b>	4.2	6.4	GLN276		
				HIS335	ASN333	11
<b>4d-65ns-5</b>	<b>2</b>	4.9	6.2			

Table 3-20: Results of validation process 5 on the 65ns model of the trajectory of 1b.

Complex	Performance	Warhead position (Å)		Hydrogen bonds		
		Starting	MD average	Starting conformation	MD simulation	
					Residue	%
<b>1b-155ns-10</b>	<b>1</b>	6.2	5.7	ARG317	ARG317	15
				PHE334	ASN333	9
					PHE334	4.5
<b>1c-155ns-9</b>	<b>2</b>	3.7	5.1			
<b>1d-155ns-4</b>	<b>1</b>	5.2	4	GLN276	GLN276	35
				ASN333	ARG317	1
					ASN333	1.5
				PHE334		36
<b>1e-155ns-1</b>	<b>1</b>	3.1	4.7	GLN276	GLN276	28
<b>1f-155ns-6</b>	<b>1</b>	4.2	3.5	CYS277	GLN276	7

					CYS277	22
<b>3a</b> -155ns-8	<b>2</b>	7.1	12			
<b>3d</b> -155ns-5	<b>1</b>	4.3	6.8	ARG317	ARG317	21
				ASN333	ASN333	15
				PHE334	PHE334	4
<b>3e</b> -155ns-14	<b>1</b>	4.3	5.7	TRP332	ASN333	21
					PHE334	14
<b>3f</b> -155ns-10	<b>2</b>					
<b>3g</b> -155ns-12	<b>1</b>	4.8	3.8	PHE334	GLN276	16
					ASN333	10
					PHE334	39
<b>3h</b> -155ns-3	<b>1</b>	3.4	5.7	HIS335	GLN276	42
<b>3i</b> -155ns-3	<b>2</b>	3.3	6.6			
<b>3j</b> -155ns-17	<b>2</b>	5.0	8.5			
<b>3m</b> -155ns-14	<b>2 and 3</b>	6.5				
<b>4b</b> -155ns-1	<b>2</b>		Left during equilibration			

Table 3-21: Results of validation process 5 on the 155ns model of the trajectory of 1b.

Complex	Performance	Warhead position (Å)		Hydrogen bonds		
		Starting	MD average	Starting conformation	MD simulation	
					Residue	%
<b>1b</b> -245-12	<b>1</b>	5.9	4.9	ARG317	TRP241	18
				ASN333	ARG317	19
				PHE334	ASN333	17
					PHE334	9
<b>1c</b> -245ns-13	<b>2</b>	4.9	10.5			
<b>1d</b> -245ns-18	<b>1</b>	6.4	6.2	TRP241		
				GLN276	ASN333	2
				ASN333		
<b>1e</b> -245ns-15	<b>2</b>		Left during equilibration			
<b>3d</b> -245ns-18	<b>2</b>	3.7	6.4			
<b>3e</b> -245ns-8	<b>1</b>	3.5	3.7	TRP332	CYS277	3
<b>3f</b> -245ns-9	<b>2</b>		Left during equilibration			
<b>3g</b> -245ns-15	<b>2</b>		Left during equilibration			
<b>3h</b> -245ns-9	<b>2</b>		Left during equilibration			
<b>3i</b> -245ns-5	<b>1</b>	4.2	5.6	HIS335	ASN333	13
<b>3j</b> -245ns-15	<b>2</b>					
<b>3k</b> -245ns-5	<b>1</b>	3.4	5.3	CYS277	GLN276	1
				ARG317	ASN333	10
<b>3m</b> -245ns-	<b>1</b>	5.0	3.8	CYS277	CYS277	40

17				ASN333	ASN333	11
<b>3n</b> -245ns-6	<b>2</b>	3.2	6.9			
<b>3o</b> -245ns-2	<b>2</b>	3.5	4.7			
<b>2b</b> -245ns-4	<b>2</b>	3.9	6.4			
<b>4a</b> -245ns-4	<b>2</b>	3.5	8.2			
<b>4b</b> -245ns-18	<b>1</b>	4.9	4.8	GLN276	GLN276	27
					ASN333	1

**Table 3-22: Results of validation process 5 on the 245ns model of the trajectory of 1b.**

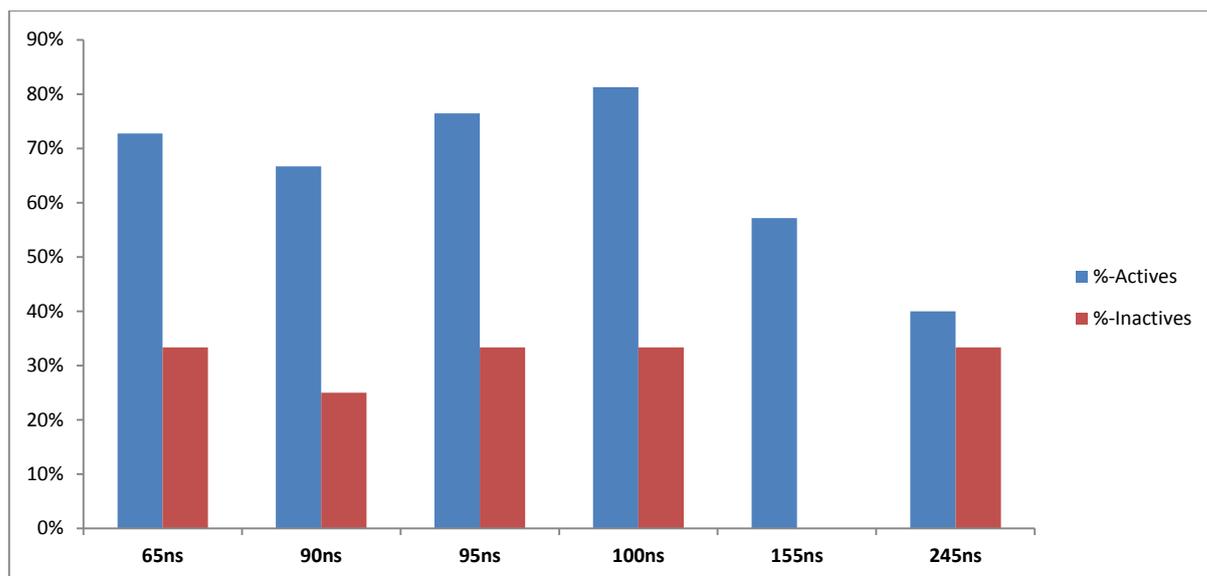
Regardless of the performance, the results of the simulations in these models were similar to those of the models from the trajectory of **1a**. This is particularly true for the hydrogen bonding profile, where ASN333 was the dominant hydrogen bond forming residue with these models, followed by GLN276, PHE334, ARG317 and ASN318. Again, it was the warhead that left the catalytic tunnel more easily than the lipophilic part. This happened with all the complexes that failed to maintain the initial orientation in TG2 active site in models 155ns and 245ns. Only the lipophilic part of compound **3a** left the hydrophobic loop during the simulation in 65ns model.

The results of validation process 5 are summarised in Table 3-23 which shows the number of complexes for active and inactive compounds that the experiment started with in each model and the number of complexes of both active and inactive compounds that maintained their positions after the simulations. Figure 3-57 shows this information as a percentage for each of the 6 models.

Model	Starting number		Passing number	
	Actives	Inactives	Actives	Inactives
65ns	11	3	8	1
90ns	12	4	8	1
95ns	17	3	13	1
100ns	16	3	13	1
155ns	14	1	8	0
245ns	15	3	6	1

**Table 3-23: Summary of the performance of the 6 models in validation process 5.**

Table 3-23 and Figure 3-57 showed that the trend followed for the performance of the models was similar to that observed in validation process 4 (Table 3-15 and Figure 3-54). Models 100ns and 95ns from the trajectory of **1a** performed the best in terms of the active compounds, followed by 65ns and 90ns which performed almost the same and 245ns was the worst. With regard to the inactive compounds, all the models were similar, where no more than one inactive complex was able to maintain its original orientation after the simulation in any of the 6 models.



**Figure 3-57: Percentages of active and inactive compounds that passed the 5-ns MD simulations without a significant change in their starting positions in each of the 6 active site models of TG2.**

### 3.4.5.3 Binding free energy calculations

As mentioned in the 'Methods' section, binding energy was calculated for these trajectories, using 2 models offered by AMBER which are MM/GBSA (Generalized Born Surface Area) (Onufriev et al. 2004) and MM/PBSA (Poisson Boltzmann Surface Area) (Srinivasan et al. 1998). This calculation was not performed on all the simulations, only the complexes that maintained their initial positions were considered, except for a few complexes for inactive compounds to investigate the effect of leaving the active site on the binding energy. The results will be presented as tables for each model. Different correlations with biological activities of the compounds, expressed as  $IC_{50}$ , were also tried and will be presented following the tables. Table 3-24 and Table 3-25 show binding free energies for the models from **1a** and **1b** trajectories respectively. Complexes in the tables marked with \$ are for inactive compounds, and those marked with \$\* are for inactive compounds that failed to maintain their starting positions at the end of the simulations.

Complex	Energy (kcal/mol)				$IC_{50}$ ( $\mu M$ )
	GB	PB	PM3 GB	DFTB GB	
<b>90ns Model</b>					
<b>1b-90ns-4</b>	-41.05	0.20	-28.87	-37.83	0.273
<b>1d-90ns-5</b>	-33.33	2.57	-21.28	-32.03	0.7
<b>1e-90ns-17</b>	-50.04	2.57	-44.15	-45.61	0.0061
<b>3b-90ns-5</b>	-41.05	6.77	-33.62	61.80	0.89
<b>3c-90ns-20</b>	-29.05	6.53	-18.53	-28.34	1.4
<b>3d-90ns-6</b>	-32.43	7.36	-27.92		0.008
<b>3f-90ns-2</b>	-40.42	2.04	-33.13	-37.39	5.925
<b>3o-90ns-3</b>	-46.66	-6.46	-47.11	-45.71	1.625

<b>2b-90ns-8<sup>\$*</sup></b>	-19.57	5.39	-18.01	-19.40	100
<b>4c-90ns-2<sup>\$*</sup></b>	-41.29	10.50	-30.96	-40.73	100
<b>4d-90ns-19<sup>\$</sup></b>	-30.47	5.63	-28.49	-31.02	100
<b>95ns Model</b>					
<b>1a-95ns-10</b>	-41.09	3.96	-37.23	-37.39	0.125
<b>1b-95ns-6</b>	-30.34	8.06	-20.19	244.88	0.273
<b>1c-95-2</b>	-41.79	0.11	-33.93	-39.76	0.44
<b>1d-95ns-4</b>	-27.83	10.37	-22.19	-27.17	0.7
<b>1e-95ns-17</b>	-36.24	5.38	6.00	-36.24	0.0061
<b>1f-95ns-20</b>	-57.52	0.59	-32.08	-53.46	0.38
<b>3a-95ns-19</b>	-33.54	6.32	-25.22	503.60	1.85
<b>3b-95ns-16</b>	-35.65	7.84	-28.11	-33.36	0.89
<b>3c-95ns-17</b>	-37.04	6.58	-25.47	-34.70	1.4
<b>3f-95ns-11</b>	-31.83	7.38	-28.59	-30.67	5.925
<b>3i-95ns-5</b>	-26.15	12.26	-20.57	289.11	6.8
<b>3n-95ns-6</b>	-41.06	7.74	-36.00	-39.18	0.775
<b>3o-95ns-1</b>	-35.69	11.77	-35.76	-33.65	1.625
<b>2c-95ns-3<sup>\$</sup></b>	-30.37	19.75	-25.67	-30.98	100
<b>4b-95ns-12<sup>\$*</sup></b>	-23.55	9.52	-22.10	-22.66	100
<b>4e-95ns-19<sup>\$*</sup></b>	-28.58	7.81	-27.02	-29.07	100
<b>100ns Model</b>					
<b>1a-100ns-5</b>	-35.68	7.62	-32.83	-32.91	0.125
<b>1b-100ns-6</b>	-33.70	8.79	-25.19	-31.89	0.273
<b>1c-100ns-6</b>	-33.26	3.10	-28.38	-32.24	0.44
<b>1d-100ns-16</b>	-50.65	-5.41	-38.02	-52.57	0.7
<b>1f-100ns-4</b>	-39.66	2.80	-26.57	-37.22	0.38
<b>3a-100ns-17</b>	-33.90	9.61	-27.38	-32.49	1.85
<b>3b-100ns-19</b>	-42.24	3.49	-31.71	-40.18	0.89
<b>3d-100ns-7</b>	-24.97	7.16	-22.24		0.008
<b>3e-100ns-15</b>	-24.12	6.38	-24.01	-24.25	4.25
<b>3f-100ns-13</b>	-32.96	9.45	-29.56	-30.32	5.925
<b>3g-100ns-6</b>	-52.77	1.32	-36.94		0.0059
<b>3i-100ns-10</b>	-29.54	7.91	-20.58	78.41	6.8
<b>3n-100ns-19</b>	-40.16	0.27	-32.18	168.24	0.775
<b>2c-100ns-11<sup>\$*</sup></b>	-38.64	7.94	-29.51	-37.55	100
<b>4b-100ns-6<sup>\$</sup></b>	-36.32	7.63	-33.43	-34.96	100

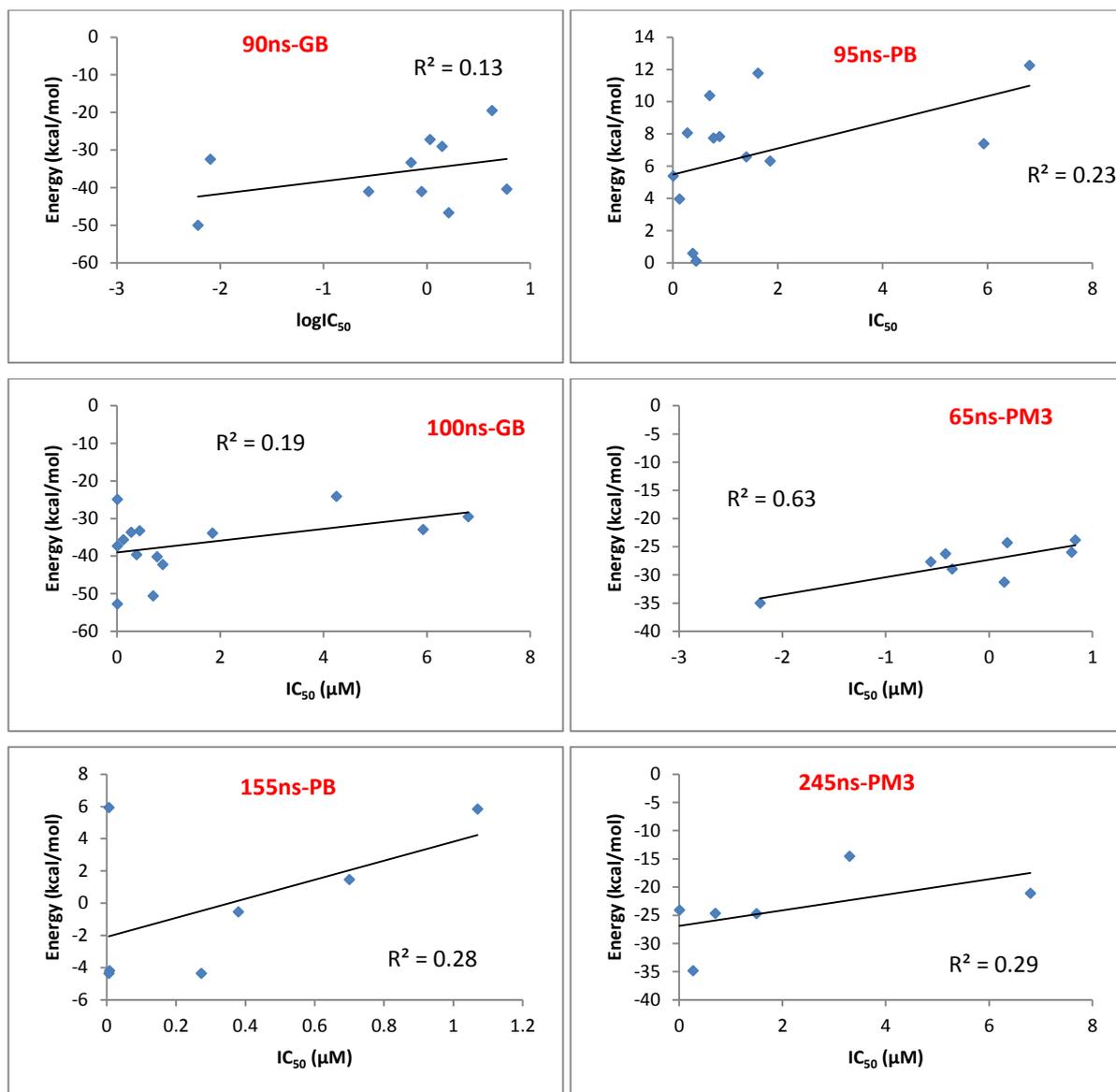
Table 3-24: Binding free energy values for the complexes in the 5-ns MD simulations in the models from the trajectory of 1a.

Complex	Energy (kcal/mol)				IC <sub>50</sub> (μM)
	GB	PB	PM3 GB	DFTB GB	
<b>65ns Model</b>					
<b>1b-65ns-18</b>	-33.78	3.83	-27.71	-31.13	0.273
<b>1c-65ns-4</b>	-32.22	7.03	-28.99	-31.00	0.44
<b>1e-65ns-20</b>	-41.04	-2.78	-35.01	-41.31	0.0061

<b>1f</b> -65ns-8	-38.26	3.49	-26.25	93.24	0.38
<b>3c</b> -65ns-13	-37.96	-3.40	-31.28	-35.91	1.4
<b>3i</b> -65ns-8	-32.42	4.03	-23.81	-30.52	6.8
<b>3j</b> -65ns-2	-33.87	2.72	-26.01	-30.07	6.3
<b>3m</b> -65ns-2	-36.32	10.46	-24.31	96.38	1.5
<b>2c</b> -65ns-9 <sup>§</sup>	-25.10	7.69	-22.11	-32.00	100
<b>4d</b> -65ns-5 <sup>§*</sup>	-38.71	0.52	-33.05	-37.53	100
<b>155ns Model</b>					
<b>1b</b> -155ns-10	-39.37	-4.36	-27.29	-36.15	0.273
<b>1d</b> -155ns-4	-38.27	1.46	-27.18	-41.67	0.7
<b>1e</b> -155ns-1	-29.95	5.93	-26.24	-34.05	0.0061
<b>1f</b> -155ns-6	-52.39	-0.53	-31.72	-46.37	0.38
<b>3d</b> -155ns-5	-39.76	-4.19	-36.15		0.008
<b>3g</b> -155ns-12	-35.12	-4.36	-31.01		0.0059
<b>3h</b> -155ns-3	-43.17	5.84	-30.67	77.02	1.07
<b>245ns Model</b>					
<b>1b</b> -245-12	-42.03	-5.23	-34.82	-40.45	0.273
<b>1d</b> -245ns-18	-29.13	9.63	-24.65	-30.33	0.7
<b>3e</b> -245ns-8	-25.35	4.58	-24.06	-23.30	0.0061
<b>3i</b> -245ns-5	-30.65	-2.27	-21.11	-29.99	6.80
<b>3k</b> -245ns-5	-24.09	15.42	-14.51	-23.82	3.30
<b>3m</b> -245ns-17	-48.49	-1.93	-24.70	-42.65	1.5
<b>4b</b> -245ns-18 <sup>§</sup>	-37.94	2.01	-31.62	-34.73	100

**Table 3-25: Binding free energy values for the complexes in the 5-ns MD simulations in the models from the trajectory of 1b.**

A short glance at Table 3-24 and Table 3-25 shows that there is no general trend in the values of the energy within a single model and between the models. Inactive compounds have achieved binding free energy values that are better than those of many of the active compounds. When correlations with biological activity were investigated, the graphs in Figure 3-58 were produced. A correlation was tried between  $IC_{50}$  values and each of the 4 energy values in Table 3-24 and Table 3-25 and the same was repeated with the logarithm of the  $IC_{50}$  values. The correlations were only tested for the active compounds. For the 65ns model, the best correlation obtained was between  $\log IC_{50}$  and GB calculated with PM3. In 90ns model, this was the correlation between the log of  $IC_{50}$  and conventional GB.  $IC_{50}$  against conventional PB gave the best correlation in 95ns. In 100ns, it was  $IC_{50}$  against conventional GB that gave the best correlation. 155ns was like 95ns ( $IC_{50}$  versus PB) and 245ns had its best correlation between  $IC_{50}$  and GB PM3 (Figure 3-58).



**Figure 3-58:** Graphs for best correlation of a binding free energy value and biological activity as TG2 IC<sub>50</sub> for the 6 models.

Figure 3-58 would give an indication that the models from the trajectory of **1b** gave better correlations than the models from the trajectory of **1a**. This is true if the R<sup>2</sup> values were the only determinants, and they were not; the numbers of complexes for which the correlations were studied should also be considered (Table 3-24 and Table 3-25) and these numbers were higher for the models from **1a** trajectory than for the second set of models.

Another approach was considered that involved combining the scores from all the models and dividing the compounds according to the warhead into sulfonium ions and acrylamides. Since more than one copy of a single compound may have been included in the combination, an average was taken for such compounds. This was then correlated to biological activity. Table 3-26 shows the average energy values for 10 sulfonium ion active compounds. These were correlated with IC<sub>50</sub> and its logarithm, and the best R<sup>2</sup> value was

obtained by plotting  $IC_{50}$  versus GB (0.43) followed by PM3-GB (0.39) (Figure 3-59). When the data for compound **3i** were excluded from the comparisons for being the least active,  $R^2$  increased to 0.56 for GB and 0.70 for PM3-GB. More interestingly,  $R^2$  for PB versus  $IC_{50}$  jumped from 0.18 to 0.84 after removing **3i** (Figure 3-60). Compound **3i** is identical to compound **1d**, one of the most active compounds used from the start, with the exception of the use of D-alanine instead of L-alanine in the amino acid linking between the piperazine ring and the acrylamide warhead resulting in about 10 times less activity against TG2 ( $IC_{50}$  of 6.8  $\mu$ M for **3i** compared to 0.7  $\mu$ M for **1d**).

Compound	Average Energy (kcal/mol)				$IC_{50}$ ( $\mu$ M)
	GB	PB	PM3 GB	DFTB GB	
<b>1b</b>	-36.71	1.88	-27.34	11.24	0.273
<b>1d</b>	-35.84	3.72	-26.66	-36.75	0.70
<b>1f</b>	-46.96	1.59	-29.16	-10.95	0.38
<b>3b</b>	-39.65	6.03	-31.14	-3.91	0.89
<b>3c</b>	-34.68	3.24	-25.10	-32.98	1.40
<b>3h</b>	-43.17	5.84	-30.67	77.02	1.07
<b>3i</b>	-29.69	5.48	-21.52	76.75	6.80
<b>3k</b>	-24.09	15.42	-14.51	-23.82	3.30
<b>3m</b>	-42.41	4.26	-24.51	26.87	1.50
<b>3n</b>	-40.61	4.00	-34.09	64.53	0.78

**Table 3-26: Average binding free energy values for compounds with sulfonium ion warhead taken from all the 6 models.**

It can therefore be said that for sulfonium ion inhibitors of TG2 with  $IC_{50}$  values of less than 3.5  $\mu$ M, molecular mechanics Poisson Boltzmann surface area (MM/PBSA, or PB) as a method for calculating the binding free energy, would give the best affinity ranking compared to other models such as MM/GBSA and PM3/GBSA. Such meaningful correlations could not be obtained with the acrylamide based inhibitors.

The results obtained from binding free energy calculations using MM/PBSA and MM/GBSA are not meaningful in all systems despite the popularity of the method in research, where 100-200 papers have been published each year in the past five years. This may be due to a variety of reasons including the lack of proper representation for entropy in the calculations, incorrect representation of charges and lacking the effects of binding site water molecules, in addition to the dependency of the results on the system being studied (Genheden & Ryde 2015). Having said that, it should also be noted that the results from these calculations could give useful correlations. In general, MM/PBSA is more rigorous and more computationally expensive than MM/GBSA but is not necessarily better in terms of results (Hou et al. 2011).

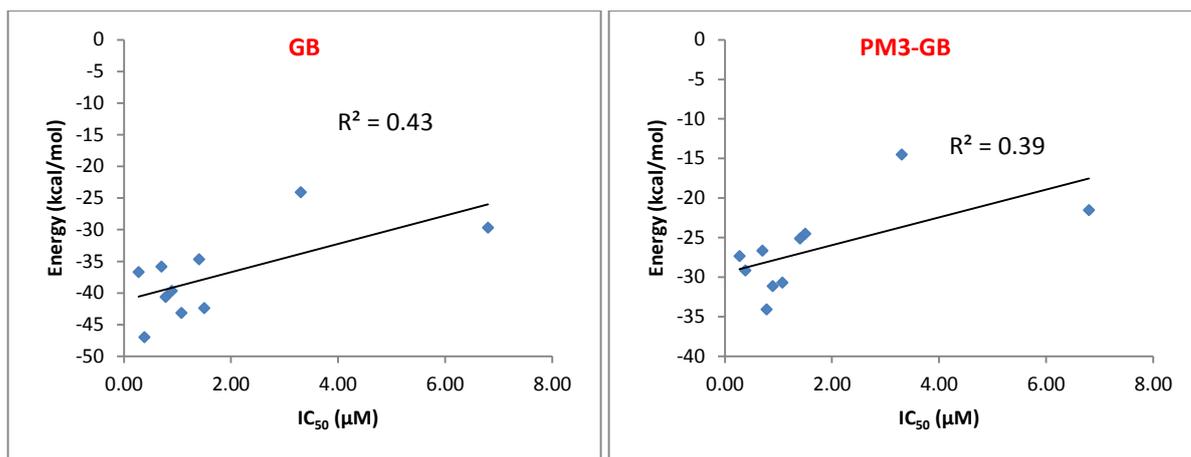


Figure 3-59:  $IC_{50}$  versus GB (left) and PM3-GB (right) for 10 sulfonium ion compounds in Table 3-26.

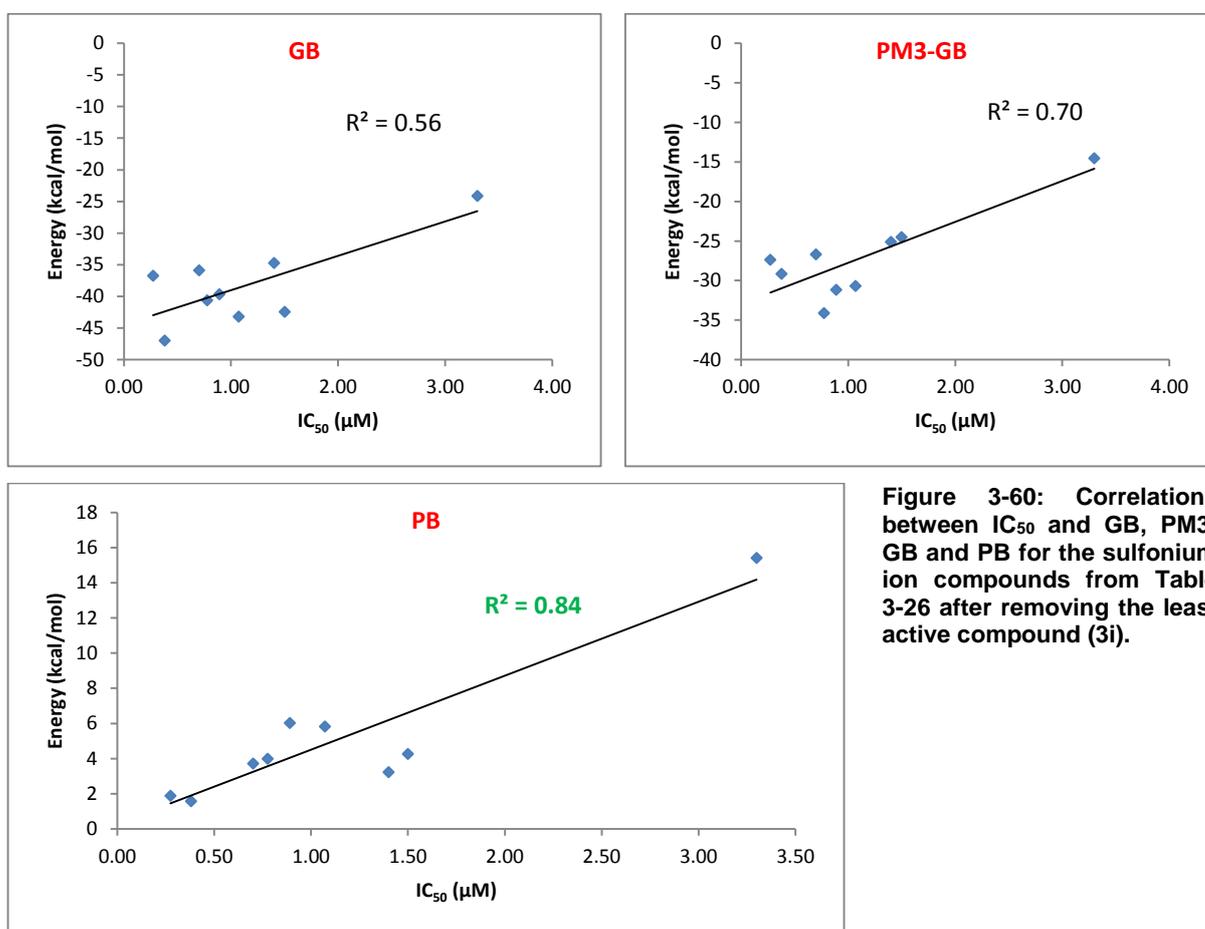


Figure 3-60: Correlations between  $IC_{50}$  and GB, PM3-GB and PB for the sulfonium ion compounds from Table 3-26 after removing the least active compound (3i).

Changes in the conformation of the complex due to entropy can produce large fluctuations in MD and thereby render the calculated free energies unstable. This can be avoided or minimised by including as many frames from the trajectories as possible (Hou et al. 2011). During the calculations of binding free energy applied on the trajectories of the 5ns-MD of TG2 inhibitors, all the frames were included in the calculations.

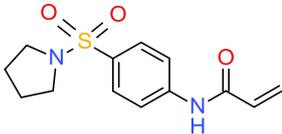
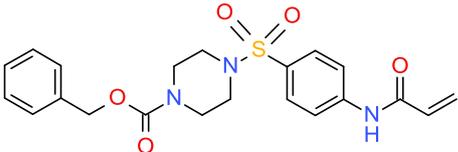
$IC_{50}$  is the concentration of the inhibitor required to reduce the rate of the reaction by one half, and in general, it is not a very good measure of affinity. This is mainly because of its

great dependence on the conditions of the measurement and the concentration of the inhibitors (Ajay & Murcko 1995). The difference in the compounds used to obtain the above mentioned correlations between  $IC_{50}$  and binding free energy values, is that the  $IC_{50}$  values were obtained using the same experimental procedures including temperatures, concentrations and incubation times (Badarau et al. 2015; Wang et al. 2012). It is worth to mention here that the paper that was the source for the compounds used in the entire thesis (Badarau et al. 2015) did not contain any free energy measurements of the tested compounds.

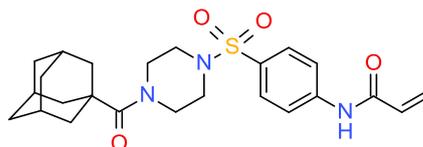
Another important point in this regard is the values of the free energies obtained; these should not be viewed as indicators of the affinity of any of the inhibitors as these values only measure the non-bonded interactions and the real activities of TG2 inhibitors involve the formation of a covalent bond with SG of CYS277 that was not accounted for. Therefore, the values would represent a measure of the affinity of the inhibitors to be inside the active site of TG2, making them ready to inhibit the enzyme by forming the covalent bond.

### 3.4.6 Validation process 6 (more active compounds)

Three active compounds were used in this validation process with GOLD settings identical to those used in validation process 1. The compounds (**i10**, **i11** and **i12**) were adopted from Prime, Andersen et al. (2012) (section 1.4.2.2, Table 3-27). The results on all the 6 valid models are presented in Table 3-28. The active compounds all had acrylamide as the warhead. An analysis of the docked complexes showed good poses for compounds **i11** and **i12** similar to those obtained in the previous test set. Compound **i10**, however, could not have its pyrrolidine ring buried within the hydrophobic loop of TG2 in all the models. As shown in Figure 3-61, the compound assumed a straight conformation; possibly due to its short linker (the part of the molecule connecting the warhead to the lipophilic part).

Compound	Structure	$IC_{50}$ ( $\mu$ M)
<b>i10</b>		1.90
<b>i11</b>		0.120

i12



0.010

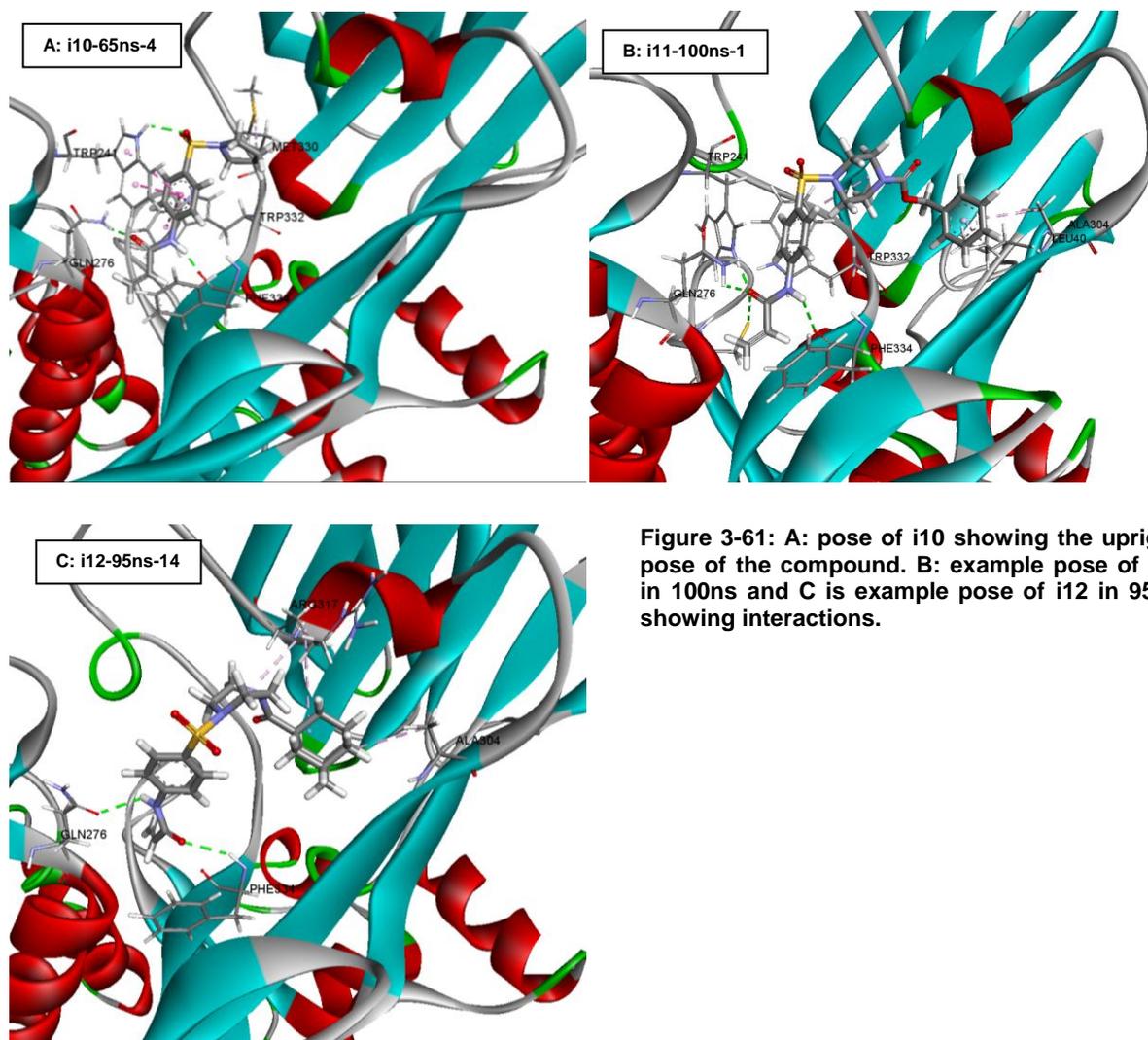
Table 3-27: Structures of compounds from (Prime, Andersen, et al. 2012) used in validation process 6.

Complex	Distance (Å)	H-bonds	GoldScore	Rank	CHEMPLP	Rank
Models of the trajectory of <b>1a</b>						
<b>i11-90ns-4</b>	6.2	CYS277	58	3 <sup>rd</sup>	37	16 <sup>th</sup>
<b>i12-90ns-17</b>	4.9	ASN333	53	1 <sup>st</sup>	55	2 <sup>nd</sup>
<b>i10-95ns-4</b>	4.1	CYS277 PHE334	45	1 <sup>st</sup>	49	2 <sup>nd</sup>
<b>i11-95ns-1</b>	4.0	TRP241 CYS277	53	12 <sup>th</sup>	47	5 <sup>th</sup>
<b>i12-95ns-14</b>	4.1	GLN276 PHE334	53	1 <sup>st</sup>	66	1 <sup>st</sup>
<b>i10-100ns-20</b>	3.1	TRP241 GLN276 PHE334	44	11 <sup>th</sup>	53	7 <sup>th</sup>
<b>i11-100ns-1</b>	3.7	TRP241 GLN276 CYS277 PHE334	61	1 <sup>st</sup>	57	4 <sup>th</sup>
<b>i12-100ns-4</b>	3.9	None	53	2 <sup>nd</sup>	59	3 <sup>rd</sup>
Models of the trajectory of <b>1b</b>						
<b>i10-65ns-4</b>	4.8	TRP241 GLN277 PHE334	63 2 <sup>nd</sup>	2 <sup>nd</sup>	58	1 <sup>st</sup>
<b>i11-65ns-13</b>	5.4	GLN276 CYS277	70	4 <sup>th</sup>	76	1 <sup>st</sup>
<b>i12-65ns-19</b>	3.9	None	76	1 <sup>st</sup>	73	1 <sup>st</sup>
<b>i10-155ns-11</b>	4.7	GLN276 PHE334	61	1 <sup>st</sup>	62	2 <sup>nd</sup>
<b>i10-245ns-12</b>	4.9	GLN276 PHE334	58	3 <sup>rd</sup>	50	5 <sup>th</sup>
<b>i11-245ns-13</b>	5.2	TRP241 MET330 HIS335	64	15 <sup>th</sup>	67	4 <sup>th</sup>
<b>i12-245ns-4</b>	4.8	GLN276 ASN333 PHE334	61	11 <sup>th</sup>	76	2 <sup>nd</sup>

Table 3-28: Docking results of validation process 6 on all valid models.

Regarding the interactions, as shown in Table 3-28, hydrogen bonds were present in all the models. Pi-interactions involving the lipophilic parts of the inhibitors were also evident in all the complexes. There were additional pi-interactions involving the piperazine and benzene rings in **i11** and **i12** and the benzene ring in **i10**. Figure 3-61 represents examples for complexes of **i11** and **i12**.

This experiment was performed in order to check the validity of the models in testing compounds from another source. Although the selected pool of ligands is not very big, nor is representative for all the chemical types of TG2 inhibitors, the results of the experiment do show that compounds with different lipophilic parts and different linkers can still dock well into the active site of TG2 models developed in the course of this work.



**Figure 3-61:** A: pose of i10 showing the upright pose of the compound. B: example pose of i11 in 100ns and C is example pose of i12 in 95ns showing interactions.

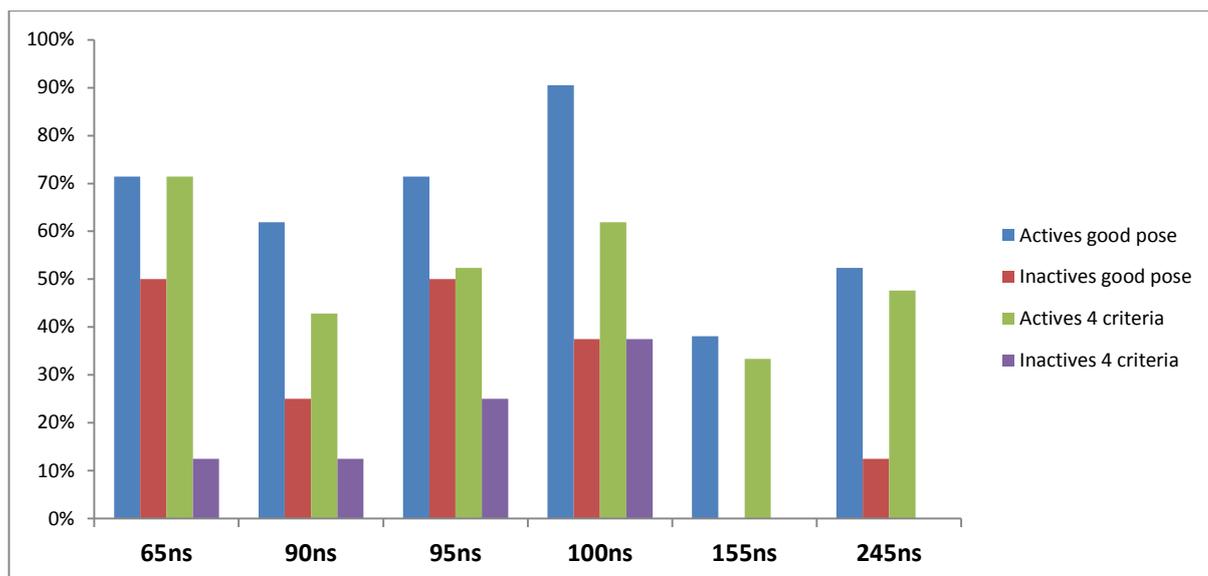
### 3.5 Inflexible Docking

Another experiment was run using GOLD that involved docking the 21 active and 8 inactive compounds used during validation process 4 into the six valid models of TG2 active site using different docking settings. This time, no residues were treated as being flexible during the docking and the search efficiency was set to “Very Flexible”. In the previous dockings, as well as in this one, the genetic algorithm (GA) speed was set to automatic, where GOLD will determine the number of genetic operations performed for each docking based on the size and flexibility of the ligands. In previous dockings, the search efficiency was set to default in which the program will apply optimum settings for each ligand. In this docking, the search efficiency was set to “Very Flexible”, where the efficiency will increase to 200% of the default (CCDC Software Limited 2013). This was applied to compensate for the inflexible treatment of the active site residues. Increasing the search efficiency like that would slow the docking but not as much as allowing 10 residues of the active site to be flexible. Therefore, the main objective of this experiment was to determine whether some faster settings of the docking experiment will produce as reliable results as the slower settings.

The results of this docking experiment will be presented as those for validation process 4 and are summarised in Table 3-29 and shown in Figure 3-62. The bent conformation was obtained in many complexes and the hydrogen bonding profile was similar to what has been observed during the previous experiments with dominance from ASN333 followed by PHE334, GLN276 and the other active site residues seen previously forming hydrogen bonds with the inhibitors. If only the pose of active compounds was to be considered, then model 100ns would have been the best. This model performed the worst, however, in terms of inactive compounds achieving the 4 criteria for good docking. If this were the most important determinant (achieving 4 criteria), then it is model 65ns that performed the best.

Model	% Actives / good pose	% Inactives / good pose	% Actives / 4 criteria	% Inactives / 4 criteria
65ns	71.43%	50%	71.43%	12.5%
90ns	61.90%	25%	42.86%	12.5%
95ns	71.43%	50%	52.38%	25%
100ns	90.48%	37.5%	61.90%	37.5%
155ns	38.10%	0%	33.33%	0%
245ns	52.38%	12.5%	47.62%	0%

**Table 3-29: The results of inflexible docking on the individual models showing percentages of good poses and passing the 4 criteria in each model.**



**Figure 3-62: Performance of the 6 models during inflexible docking showing the percentages of good poses and passing the 4 criteria for active and inactive compounds.**

In summary, rigid docking in a system involving TG2 and its irreversible inhibitors may be able to produce the correct binding mode but its ability to distinguish between active and inactive compounds is poor. It has been stated that treating the receptor as a rigid body has its disadvantages including incorrect prediction of correct binding modes and unreliable scores (Mohan et al. 2005). Therefore, the time saved by using inflexible docking should not motivate the generalisation of the method, especially that GOLD allows only for 10 residues to be treated as flexible which should not increase the computation time substantially. Particularly with TG2, flexible residues should be employed whenever possible, at least residue 333 (ASN333), whose conformation has been proven to be very important for successful achievement of the bent conformation of the inhibitors within TG2 active site (section 3.3.6).

### 3.6 Additional Scoring Functions

The GOLD docking program offers the possibility of using one of 4 scoring functions as the primary function along with one optional additional function to be used for rescoring the solutions obtained with the primary function (CCDC Software Limited 2013). GoldScore was used as the primary function and CHEMPLP as the rescoring function in all the dockings up to this stage. To validate the choice of GoldScore as the primary function, 3 additional dockings were performed using each one of the other available scoring functions as the primary and only scoring function. The compounds and docking settings used at this stage were the same as those used in the initial docking when choosing the active site models (6 active compounds from Badarau et al. (2015) and default GOLD settings with 10 flexible residues and early termination).

### 3.6.1 ASP scoring function

The Astex Statistical Potential (ASP) scoring function (Mooij & Verdonk 2005) is an atom-atom potential which was derived from analysing existing ligand-protein complexes. It resembles functions such as PMF and DrugScore (Mooij & Verdonk 2005). With this scoring function, no acceptable pose for any of the 6 compounds was obtained in 4 models; 65ns, 90ns, 155ns and 245ns. In the remaining 2 models, one compound in each gave only one acceptable pose. These were compounds **1c** in 95ns model and compound **1e** in 100ns model. They were ranked 8<sup>th</sup> and 6<sup>th</sup> respectively out of the resulting 20 solutions. A scoring function that estimates binding energy by examining the distances and angles between atoms from the ligand and the receptor active site is, therefore, not appropriate for the TG2 system.

### 3.6.2 ChemScore

ChemScore (Eldridge et al. 1997; Baxter et al. 1998) is an empirical scoring function that calculates the binding energy as a sum of multiple empirical terms derived from a training set of 82 ligand-protein complexes from the Protein Data Bank. The terms include binding free energy, hydrogen bonding, metal, torsional and lipophilic terms. Only 4 compounds managed to achieve the bent conformation within TG2 active site using this scoring function. These were **1d** in 90ns model, **1b** in 95ns model and **1c** and **1e** in 100ns model. Bad results obtained with ChemScore scoring function may be attributed to the inherent deficiencies of empirical scoring functions in general. Among these are the dependence of the function on approximations and the lack of negative data in the training set (Eldridge et al. 1997).

### 3.6.3 CHEMPLP

CHEMPLP (Korb et al. 2009) is another empirical function that includes terms such as piecewise linear potential for shape complementarity, and a heavy atom clash potential for the ligand atoms. The function also utilises some ChemScore terms such as hydrogen and metal bonding and torsional terms. In the studied TG2 system, the performance of CHEMPLP was better than that of ASP or ChemScore but not as good as that of GoldScore. The performance of this function was better than the previous two with 8 complexes in the 6 models having the right bent conformation. These were **1a** and **1d** in 90ns model, **1a**, **1b** and **1c** in 95 ns model, **1a** and **1c** in 100ns model and **1b** in 155ns model.

Being an empirical function is probably the reason of failure of CHEMPLP to perform as well as GoldScore with TG2 inhibitors. However, this function was suitable for rescoring when used with GoldScore as the primary function and was, in some cases, able to rank the

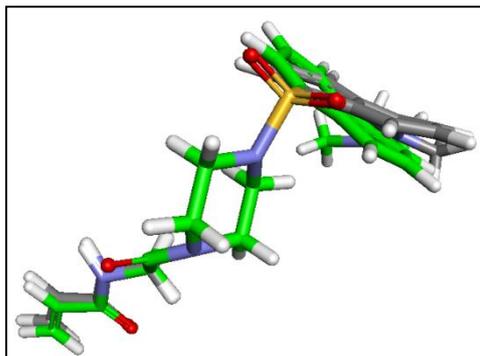
solutions of a single ligand in the same manner as GoldScore. It should also be noted that all the inhibitors studied here are covalently bound to TG2 and this has not been accounted for during the docking; all the rankings and differentiations were made on the assumptions of the bent conformations and the interactions involving the lipophilic part and the central part of the ligand.

GoldScore (Verdonk et al. 2003) is a scoring function that is based on a force field and has terms for hydrogen bonds and van der Waals interactions between the ligand and the receptor, in addition to ligand internal van der Waals and torsional energies. The performance from this function was simply much better than any of the other 3 functions offered by GOLD. This is very obvious from the results presented so far in this work. The superior performance was true despite the fact that GoldScore did have flaws manifested in giving good scores and ranks for inactive compounds and failing to rank the compounds correctly based on their scores. The good performance may be due in part to the importance of hydrogen bonding for TG2 inhibitors, which the function was able to demonstrate for many of the active inhibitors during the various stages of the docking as it has a separate term for hydrogen bonds between the ligand and the protein. Furthermore, the ligands used throughout this work generally have high molecular weights ranging from 239 (**1a**) to 488 (**3m**) and this can increase the contribution from the term describing van der Waals interactions between the ligand and the protein as well as the term for these interactions within the ligand itself during individual dockings.

### 3.6.4 Rescoring

The GOLD docking program is equipped with the ability to rescore an already docked conformation using any of the scoring functions available in the program. There is also the option of performing local minimisation on the ligand to enhance the score. The rescoring utility was applied on the valid models using complexes generated during different experiments of the validation process. Only active compounds were included in this experiment and the selected complexes were for compounds that achieved at least 3 of the criteria for a good docking complex.

The complexes were rescored with the four scoring functions including GoldScore to examine the effect of local minimisation on the already generated poses and scores. Rescoring was performed with and without local minimisation. The minimisation process did not produce significant changes; in these complexes it usually produced minor changes in the conformation of the ligand especially involving the lipophilic part (Figure 3-63).



**Figure 3-63: Minimised (grey carbons) and non-minimised (green carbons) conformations of compound 1e before rescoring with GOLD.**

In 4 of the 6 models (65ns, 90ns, 95ns and 245ns), no useful correlation could be found between the scores obtained with any function during the rescoring process and biological activity expressed as  $IC_{50}$ . In many of the cases in these models, the opposite of the trend that should ideally be seen was produced; there was a decline in the score with the increase in biological activity (Figure 3-64).

For the remaining 2 models (100ns and 155ns), the correct trend of scores was obtained but the correlation was not very high. The best correlation obtained for model 100ns was the minimised GoldScore ( $R^2 = 0.35$ ) and for 155ns was the non-minimised ChemScore ( $R^2 = 0.31$ ) (Figure 3-65). Although these correlations are poor, they are the best correlations obtained so far between docking scores and  $IC_{50}$  values in the line of this work. GoldScore has already been shown to be superior to the other functions in predicting the correct binding mode and the hydrogen bond profile; and here it has been shown that the performance of GoldScore with TG2 inhibitors can be improved by applying local minimisation on the docked ligand conformation. This further validates the choice of GoldScore as the primary scoring function in the dockings performed with the selected TG2 inhibitors.

The results from ChemScore were good without the local minimisation, therefore the application of the local minimisation during rescoring should not always be attempted even though it would sound reasonable. It can be also concluded that ChemScore, though was unable to predict the correct binding mode when used as the primary function (section 3.6.2), did manage to achieve the correct trend in the scores based on the activities giving an indication that using ChemScore for rescoring solutions obtained with GoldScore is a possible approach for enhancing the ranking when it comes to docking inhibitors of TG2.

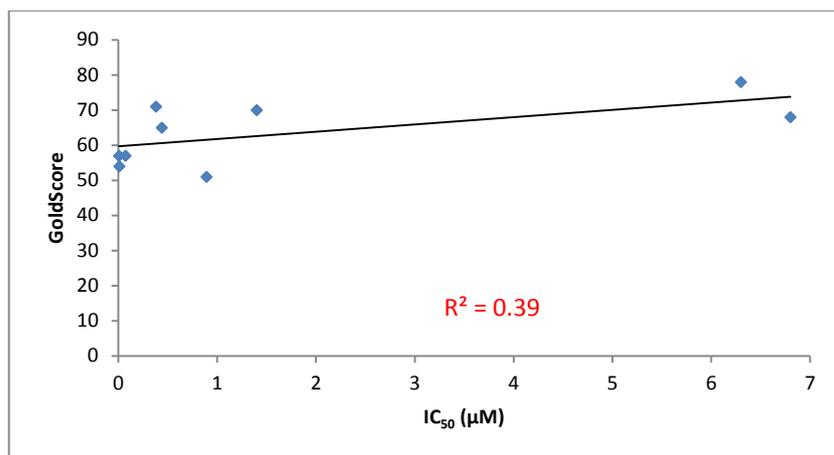


Figure 3-64: Correlation between TG2 IC<sub>50</sub> and GoldScore minimised rescoring for model 65ns with the inversed trend; lower scores for more active compounds (GOLD docking scores are unit-less).

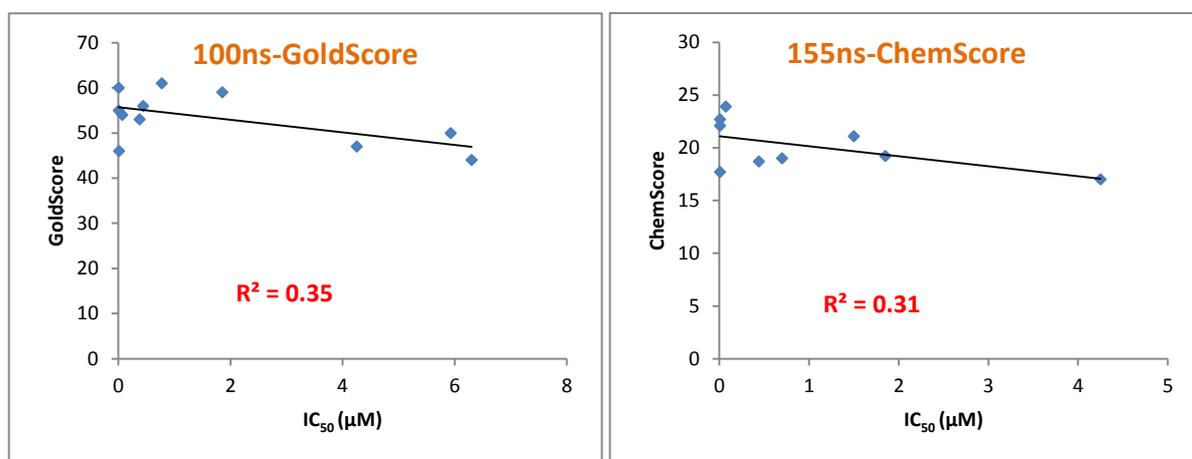


Figure 3-65: Correlations of minimised GoldScore and non-minimised ChemScore with IC<sub>50</sub> in models 100ns and 155ns respectively during the rescoring process. GOLD scores are without units.

The rescoring approach in GOLD with minimisation of the top ranked solutions has been found to improve the ranking of the docked ligands and the overall performance of the scoring functions (Perola et al. 2004). However, because rescoring did not produce reliable results with all the 6 models, it cannot be said that it should routinely be used for this set of compounds with TG2.

## 3.7 Covalent Docking and MD

### 3.7.1 Docking

The functionality of covalent docking in GOLD requires setting the ligand in a special way to allow for the formation of the covalent bond. This includes adding a link atom to the ligand and the protein at the point at which the bond would be formed. The link atom used was the sulphur atom of the active site cysteine residue. On the compounds, a sulphur atom was added to the electrophilic carbon and was set as the link atom. For sulfonium ion compounds, this would have required removing the sulfonium ion entirely and replacing it

with a sulphur atom. Since this removal was going to change the size of the sulfonium compounds considerably, this compound set was not considered for this experiment.

A total of 7 active and 4 inactive acrylamide compounds was used in this experiment. The results are presented in the form of percentages and column graphs as in validation process 4 in Table 3-30 and Figure 3-66. The bent conformation (Figure 3-67) with the required interactions and the top ranking were achieved for many of the active compounds during the docking. The inactive compounds, however, managed to dock well with the bent conformation at higher rates than what was seen in normal docking. This cannot be surprising because the program is forcing the covalent bond to be formed and the bent conformation for the remainder of the compound structure is probably the best it can adopt if the covalent bond was already there. These inactive compounds, however, did not always succeed with regard to the other criteria and hence the low rates in the last column of Table 3-30. This may mean that the models (except possibly 90ns) will adopt, with the right conformation, any potential compound having the necessary parts (the lipophilic part and a linker in addition to the warhead) but will be more selective towards active compounds when all the criteria are considered.

Model	% Actives / good pose	% Inactives / good pose	% Actives / 4 criteria	% Inactives / 4 criteria
65ns	85.71	50	42.86	25
90ns	100.00	75	85.71	75
95ns	85.71	75	71.43	0
100ns	85.71	100	57.14	25
155ns	85.71	25	71.43	0
245ns	57.14	50	28.57	0

Table 3-30: Results of covalent docking using acrylamide compounds on the 6 models.

### 3.7.2 Molecular dynamics

The molecular dynamic simulations applied to the covalent complexes were analysed visually by VMD program to investigate whether or not the bent conformation was preserved with time. This was confirmed by measuring the RMSD of the ligand alone. The hydrogen bonds formed during the simulation were also analysed using the CPPTRAJ program of the AmberTools package, and the default CPPTRAJ parameters for hydrogen bonds were employed (angle cut-off was 135° and distance cut-off of 3 Å). Two docking complexes for each of the 7 active compounds and 2 docking complexes for 3 of the 4 inactive compounds (all except **2b**) were run through this stage of the experiment.

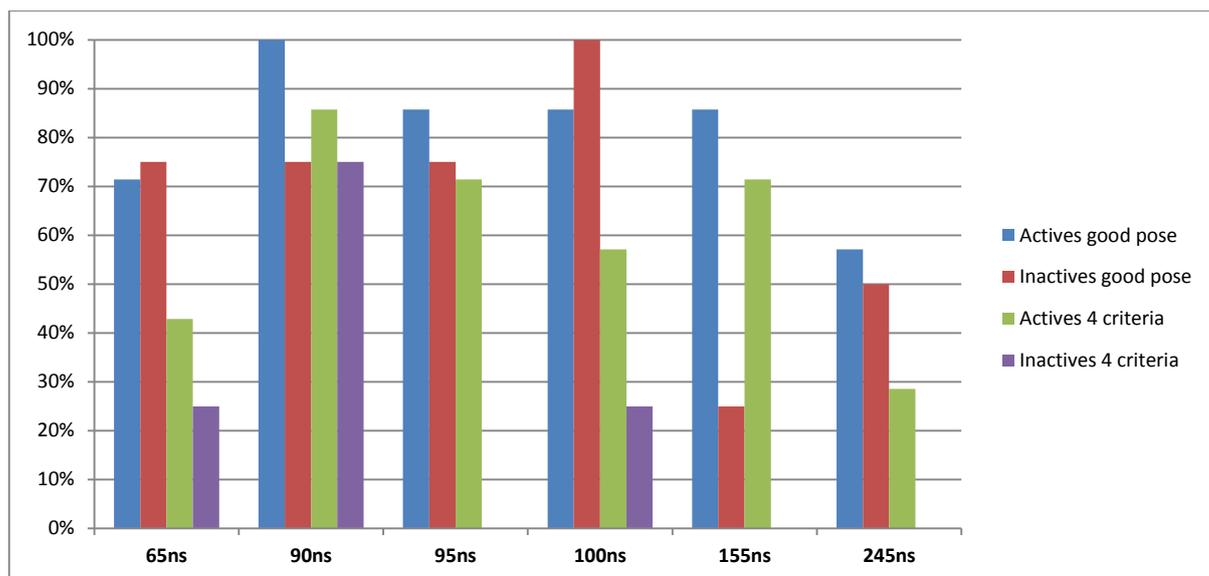


Figure 3-66: Performance of the 6 models in covalent docking experiment.

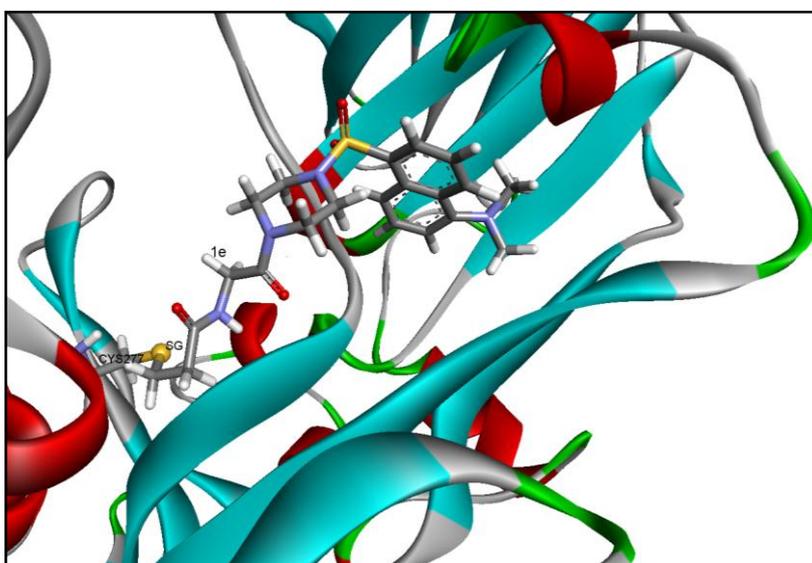
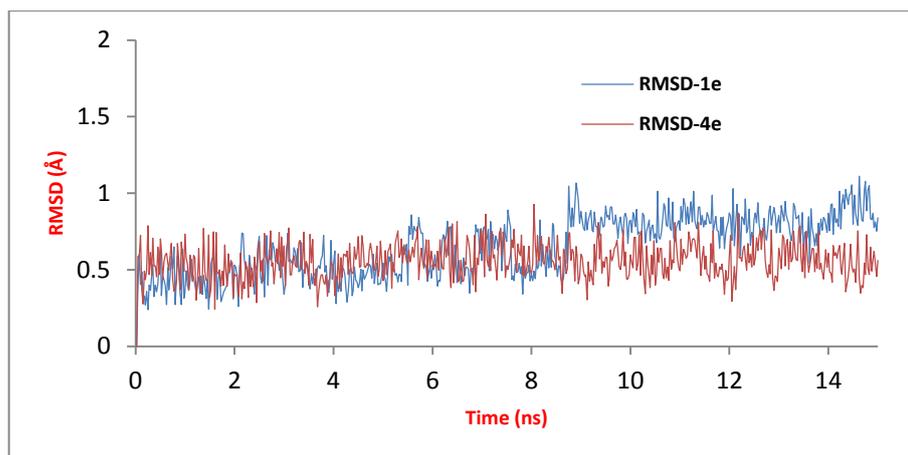
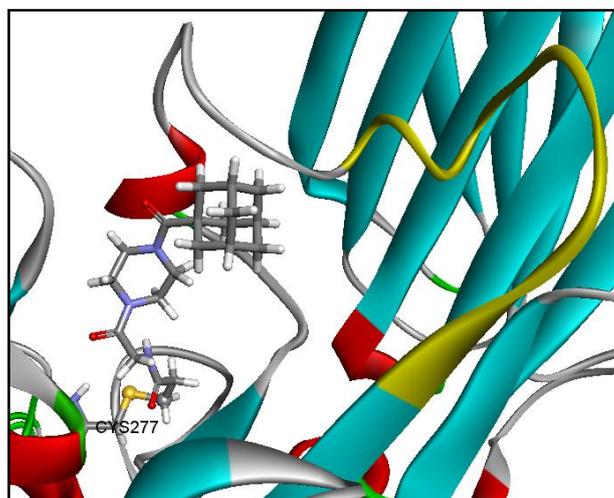


Figure 3-67: The pose of compound 1e in model 155ns as a representative for covalent docking.

In total there were 20 simulations for 10 compounds, each presented a simulation time of 15 ns. The behaviour of the individual compounds during these simulations was very similar, regardless of the activity of the compound. This behaviour can be described very briefly by saying that everything remained the same; the ligand maintained its bent conformation within the active site for the length of the simulation. This was confirmed by the very similar graphs for RMSD, for active and inactive compounds (Figure 3-68). The exception was in one of the simulations for each of these 3 active compounds, **1a**, **3I** and **3o**. The lipophilic part of the compounds left the hydrophobic loop of TG2 active site causing the compound to stand straight. This movement did not have any effect of the position of the warhead or the covalent bond between the warhead and CYS277 (Figure 3-69).

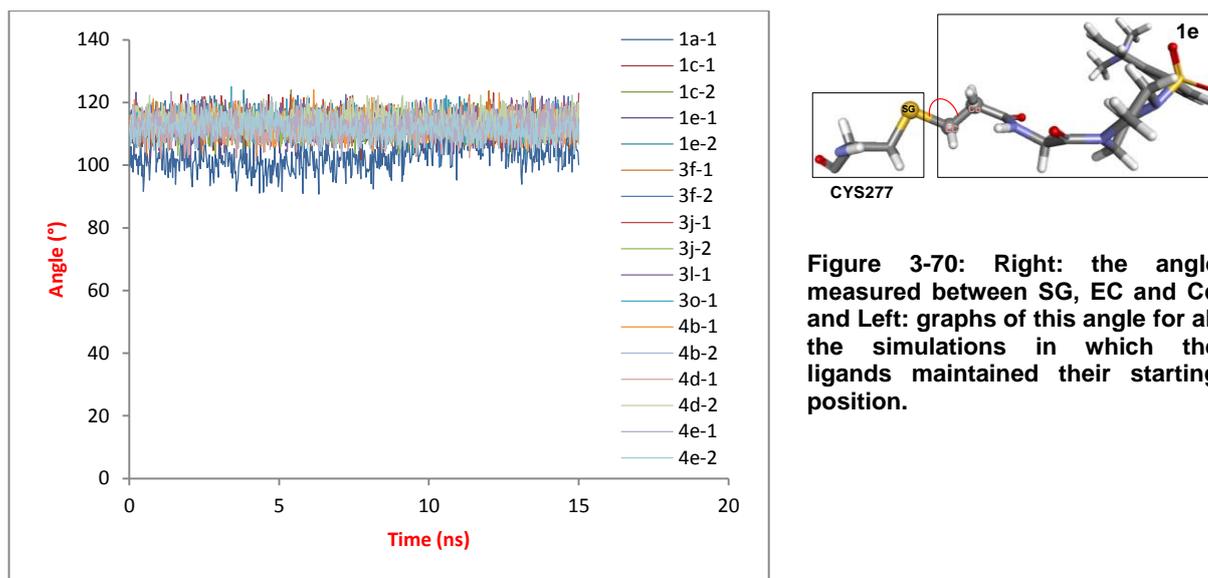


**Figure 3-68:** RMSD for the ligand only during the 15-ns covalent MD simulations for active (1e) and inactive (4e) compounds.



**Figure 3-69:** Compound 1a from one of its simulations in which the lipophilic part of the compound straightened in the active site. The hydrophobic loop of TG2 is coloured yellow while the rest of the protein is coloured by secondary structure. The covalent bond between 1a and CYS277 is shown.

The angle between the sulphur atom of CYS277 (SG), and the electrophilic carbon (EC) and the alpha carbon ( $C\alpha$ ) of the acrylamide bond of the compounds (Figure 3-70 Right) has been measured for all the simulations where the lipophilic part did not leave the hydrophobic loop of TG2. The graph in Figure 3-70 was produced which clearly shows that the angle oscillated over a short range of degrees and that the oscillation was similar for all the simulations. The average value for the angle across all the simulations was  $112.3^\circ$ . This value is consistent with that recorded in GAFF (AMBER General Force Field for organic molecules) (Wang et al. 2006; Wang et al. 2004) for c3-c3-ss which is  $112.690^\circ$  where c3 is an  $sp^3$  hybridised carbon and ss is  $sp^3$  hybridised sulphur. If the angle was considered one of the parameters of the covalent bond, and it is, then these results can be used to confirm that all the simulations preceded without any problems with the covalent bond and that the parameters created for the bound compounds using GAMESS and RESP worked properly during the simulations.

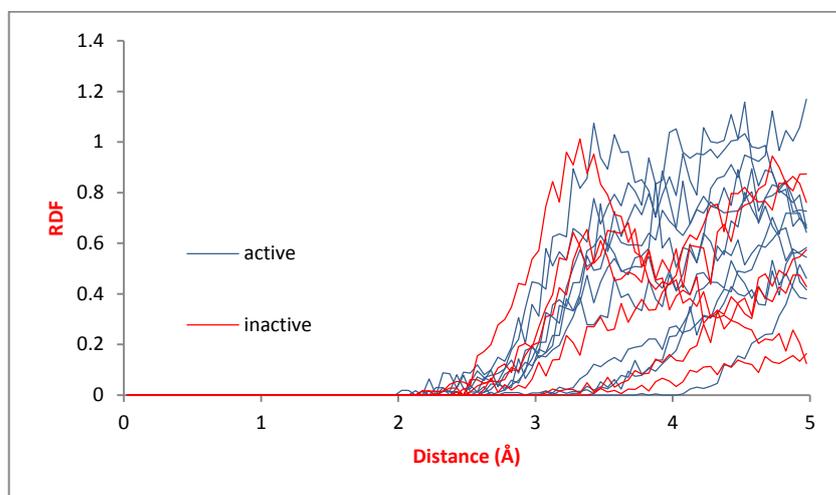


**Figure 3-70:** Right: the angle measured between SG, EC and C $\alpha$  and Left: graphs of this angle for all the simulations in which the ligands maintained their starting position.

The hydrogen bonding profile was not very meaningful in terms of discriminating between active and inactive compounds. In fact, the inactive compound **4e** showed the best profile where in one of the 2 simulations for this compound there were hydrogen bonds with ASN333 for 75% of simulation time, with PHE334 for 67%, with TRP332 for 53% and with GLN276 for 45% of simulation time. However, in most of the cases there was an abundance of hydrogen bonds between the acrylamide compound and one or more of those 4 residues of TG2. This could be due to the presence of the covalent bond. In the previous dockings and MD simulations when there was no covalent bond, the objective was to get the warhead close to CYS277 to aid in the formation of the covalent bond and hydrogen bonds were one of the tools in achieving this goal. In the case of covalent MD, the covalent bond was already in place and therefore the compounds were in ideal positions for the formation of hydrogen bonds.

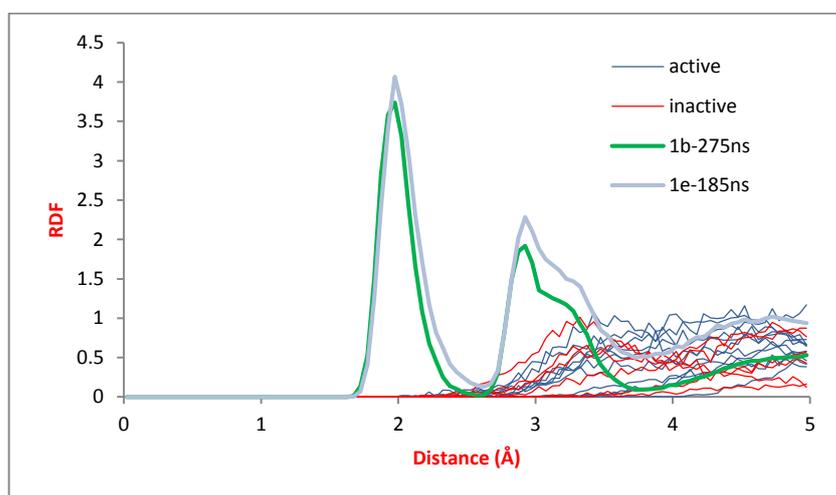
Another parameter that was analysed was the radial distribution function (RDF) of water molecules around the covalent bond. This function measures the probability of finding water molecules within a predefined distance from a specified atom mask. In other words, it measures the density of water as a function of distance. It was measured using CPPTRAJ to account for all water molecules that are within 5 Å of SG of CYS277. The rationale was to test whether the density of water around the covalent bond would be different between active and inactive compounds. This is based on the assumption that water can serve as a proton shuttle or a source of proton during the formation of the bond. The results are shown in Figure 3-71 in which the RDF is presented for active (blue) and inactive (red) compounds. The graph shows the RDF or the probability as a function of the distance. There were no water molecules within the 2 Å of SG and that is reasonable as anything nearer will be within the van der Waals radii of the atoms, and may result in clashes. RDF starts to rise thereafter

with the distance without any possible distinction between active and inactive compounds. In any case, probability did not exceed 1 except in the simulations for 2 active compounds and 1 inactive compound.



**Figure 3-71:** The radial distribution function of water molecules around the covalent bond during the simulations of active and inactive compounds.

When this was compared to the simulations applied initially on compounds **1b** and **1e**, the graph in Figure 3-72 was produced. The calculations for compound **1e** was performed up to the 185<sup>th</sup> nanosecond because the warhead had left the catalytic tunnel after that and for the 2 compounds it was performed on the SG atom as well. For the simulations containing active compounds within the active site, RDF was highest at 2 Å and then at 3 Å, indicating that if SG were free (not involved in a bond), water molecules can get very close.



**Figure 3-72:** The same RDF graphs as in Figure 3-71 with the addition of RDF graphs for the initial simulations of compounds **1b** and **1e**.

In conclusion, water molecules do not come within 2 Å of SG and EC if there was a bond between the 2 atoms, and after this distance, some water molecules may approach but not more than very few. Without the bond being there, there is much higher probability for water molecules to come within a 2-Å distance from SG. RDF tends to move toward unity at higher distances, approaching the actual density of water molecules.

Regarding the parameter production, the cap added to the ligand during GAMESS calculation was chosen because it will produce, with the acrylamide compound, a product that is very similar to the acrylamide-CYS277 complex built during these simulations. Another reason for choosing this particular cap is that it is available in the AMBER Force Field as a part of the side chain of the amino acid methionine and therefore the charges of its atoms are already available. Those charges were used during the first stage of charge fitting with RESP to fix the charges of the cap and distribute the rest of the electrostatic potential calculated by GAMESS on the atoms of the uncapped acrylamide compound.

One simple conclusion can be drawn from the results of the covalent docking and MD. The difference between active and inactive TG2 inhibitors lies in their ability to place the warhead close to CYS277. Once these are close to each other, the covalent bond will be formed. This is at least true to this set of acrylamide compounds; they all had the same warhead and when the covalent bond was forced on the system, they all behaved in a similar way. The behaviour of water molecules as compared to that when there was no covalent bond may give an indication that water can serve as a proton source or a proton shuttle during the reaction but once the reaction has finished and the covalent bond was formed, then water molecules may leave their places next to SG.

### 3.8 Analysis of the Valid Active Site Models of TG2

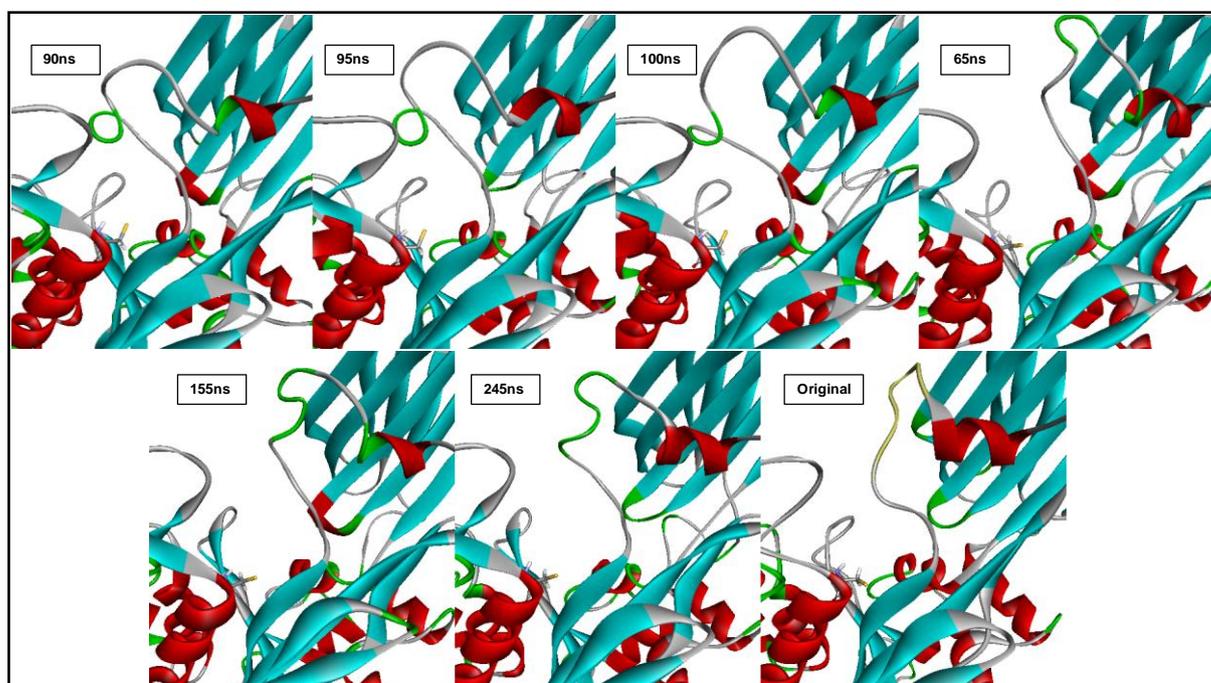
The conformations that were selected from the MD trajectories as being valid models for the docking of TG2 inhibitors, and were then validated for this role, have been analysed to examine the differences between them and between the conformation of TG2 in the original crystal structure (2Q3Z) (Pinkas et al. 2007). The analysis involved several measurements which are presented in this section.

The root-mean-squared-deviations (RMSD) of the models was measured using the structure of 2Q3Z as a reference. The results for the individual models are presented in Table 3-31 which shows that the models had very comparable values for their RMSD if their source is considered; in other words, models from **1a** trajectory (90ns, 95ns and 100ns) had an average RMSD of about 3.0 Å while those from the trajectory of **1b** (65ns, 155ns and 245ns) showed a value of about 2.5 Å. These values are slightly different from the values of the RMSD measurement performed on the trajectories in section 3.2 (Figure 3-12 and Figure 3-25) but this can be expected since the reference frame in the 2 measurements was not the same. In section 3.2, the reference was the first frame in the production phase.

Model	RMSD (Å)
90ns	3.1
95ns	3.1
100ns	2.9
65ns	2.4
155ns	2.5
245ns	2.6

**Table 3-31: RMSD values of the 6 models using the original crystal structure as a reference.**

It should be noted here that the structure that was used as a reference had the missing residues added and minimised using CAChe only while in the models, the added residues would have had the time to relax and adjust themselves to their positions. This was especially important for the loop between residues 319 and 327, because it consisted of the closest missing residues to the active site. Other added residues were in locations farther away from the active site. In this loop, 2 distinct patterns were noticed; the behaviour in the models from the trajectory of **1a** in which the loop moved to the inside of the active site to become closer to the tunnel and CYS277, and the behaviour in the models from the trajectory of **1b**, where the loop was pointing to the outside of TG2 creating an opening in the active site (Figure 3-73). The different orientation of the loop between the 2 sets of models may be responsible for the slightly better performance of the models from the trajectory of **1a** in posing the compounds with the correct bent conformation, for the models from the trajectory of **1b** would have had more space to accommodate the lipophilic part; the space created by the added loop and the hydrophobic loop.



**Figure 3-73: The conformation of the loop between residues 319 and 327 in the 6 valid models compared to the original. The loop is coloured yellow in the “Original” picture and by secondary structure in the rest and CYS277 is shown to give an indication of the active site.**

This was seen during the process of selection of the valid models, one example is the results of docking on model 245ns in section 3.3.2.3. The justification is that in the models from **1a** trajectory, the loop orientation will make it very difficult for the lipophilic part of the inhibitors to go anywhere except in the hydrophobic loop of the active site while in the models from **1b** trajectory, there is the space created by the orientation of the added loop that can provide additional accommodation for the lipophilic part of the inhibitors. Although the orientation of the loop in the original structure is closer to the one in the models from **1b** trajectory, this should not make the orientation more realistic, as the loop contains added residues and without any reference structure to compare to, there is no way of telling that the loop has relaxed enough to be realistic. The atomic fluctuation values for the loop residues in the 2 trajectories from which the models were taken (Table 3-32) are comparable with the exception of the 1<sup>st</sup> three residues and cannot be used to judge which loop is better.

Residue	Atomic fluctuation in <b>1a</b> -trajectory (Å)	Atomic fluctuation in <b>1b</b> -trajectory (Å)
319	3.5835	1.4738
320	3.4112	1.3913
321	2.8635	1.8123
322	2.3366	2.5744
323	2.0198	2.2968
324	1.7981	1.8178
325	1.7782	1.9103
326	1.7456	2.1469
327	1.6844	2.046

**Table 3-32: Atomic fluctuation values of the added loop from Figure 3-73.**

One additional observation that can be easily spotted from Figure 3-73 is the orientation of the SG atom in the catalytic cysteine residue. In the original structure, and in the models from the trajectory of **1b**, SG was facing to the outside of the protein while in models from the trajectory of **1a** it was facing to the inside of the protein. This did not have any major effects on the dockings as manifested by the similar distances recorded in the docking tables between the warhead and this atom. The inclusion of CYS277 within the 10 residues to be treated as flexible during GOLD dockings may have minimised the effect of this difference.

Despite the importance of residue ASN333 in the dockings of TG2 inhibitors, its conformation was not very different between the different models (Figure 3-74). Again it was one of the 10 flexible residues during GOLD docking.

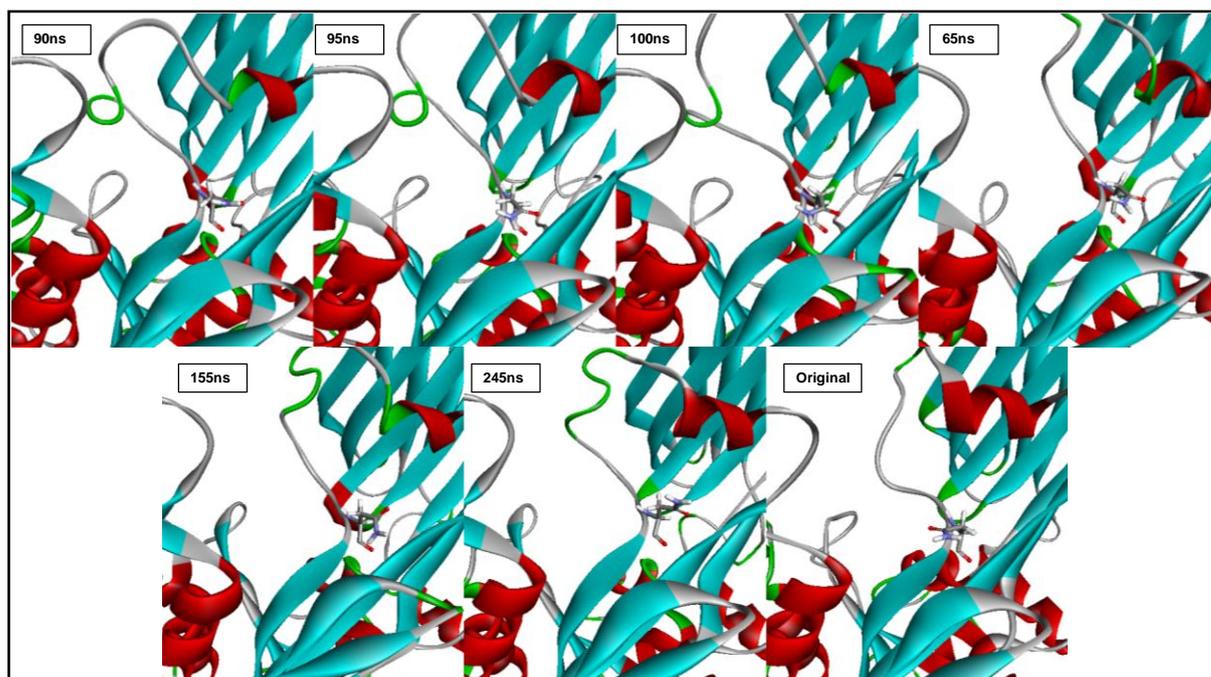


Figure 3-74: The orientation of ASN333 in the 6 models compared to the original.

The same cannot be said about the bridging tryptophan residues (TRP241 and TRP332) (Figure 3-75). In the original crystal structure, these residues formed a bridge to the catalytic tunnel through being stacked in the way presented in Figure 3-75, with the five-membered ring of the indole side chain facing each other on the 2 residues and the benzene rings are parallel to each other and the inhibitor is supposed to position itself between the 2 residues (Pinkas et al. 2007). Such orientation of the bridging tryptophan residues is more manifested in the models extracted from the trajectory of **1a** than in those from the trajectory of **1b**. This may contribute to the slightly better docking results in the former set of models as they offer more realistic representation on the approach to the active site than the second set of models. It should be emphasised, however, that both residues were allowed to be flexible during the dockings, minimising the effect of this difference.

Finally, regarding a justification for the different docking results between the models, GOLD uses genetic algorithm as its docking method and this is a random searching algorithm that operates by generating arbitrary poses for the ligands within the active site and scoring them. This may result in different docking solutions for the same set of ligands being docked in the same protein active site models using the same settings. Such different solutions have actually been observed during the running of validation process 4 (section 3.4.4) when 3 identical attempts of docking were performed on the 6 models and yet different results were obtained with each attempt. As a result, although the performance of the models from the trajectory of **1a** seems slightly better than that of the models from the trajectory of **1b**, all the

six models are important when it comes to testing potential compounds as inhibitors of TG2, if GOLD random algorithm was to be used for the docking.

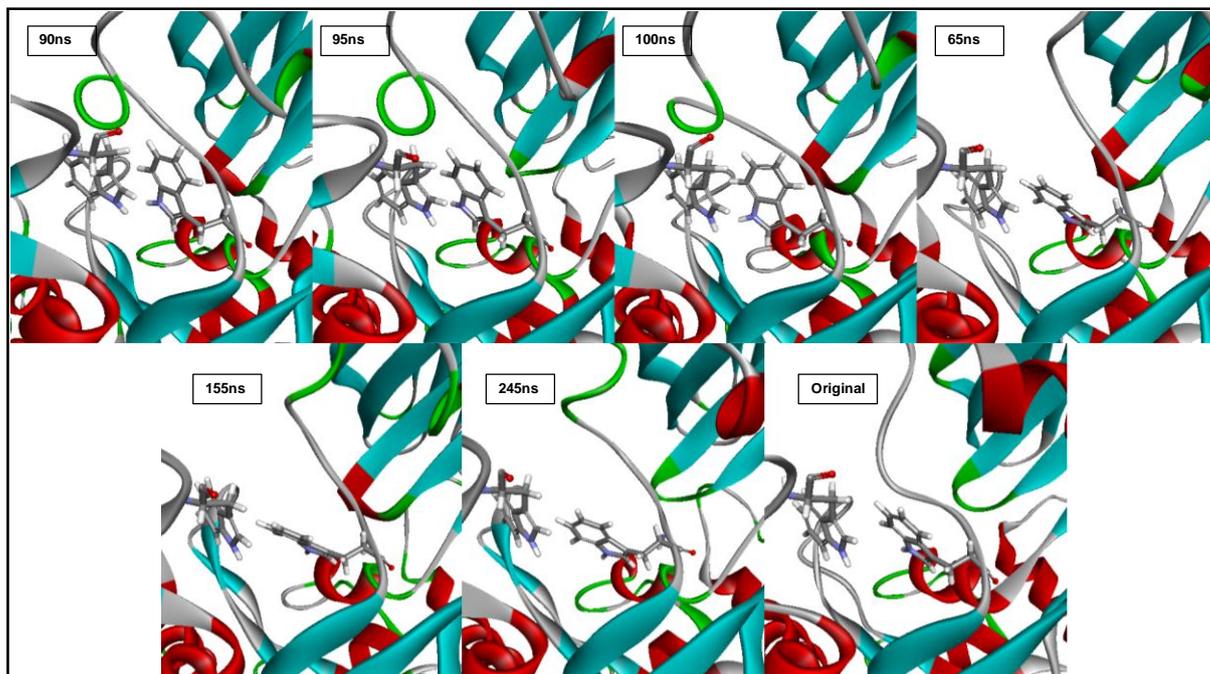


Figure 3-75: The orientation of the bridging tryptophan residues in the 6 models along with their orientation in the original crystal structure.

### 3.9 Analysis of the MD Trajectory of Empty TG2

The initial MD simulation that was applied on TG2 crystal structure to generate the initial models for docking was repeated twice with  $ig=-1$  that sets the random seed for the initial velocity based on the current date and time. The three simulations were performed to collect some data on TG2 on its own, without any inhibitors during MD simulations. The results are presented below. The 3 runs were stable as reflected by the change in temperature and total energy of the system with time. Graphs of these changes are shown in Figure 3-76.

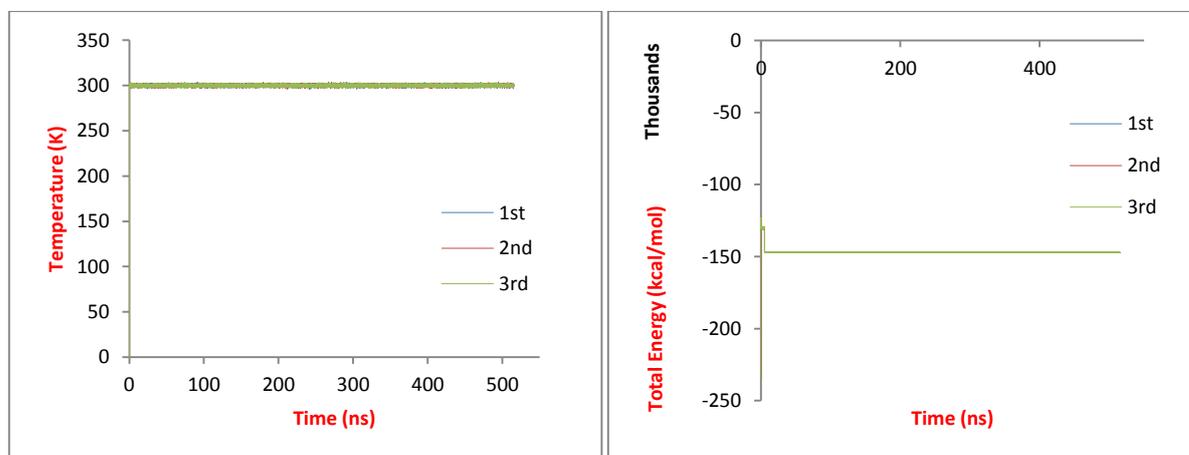
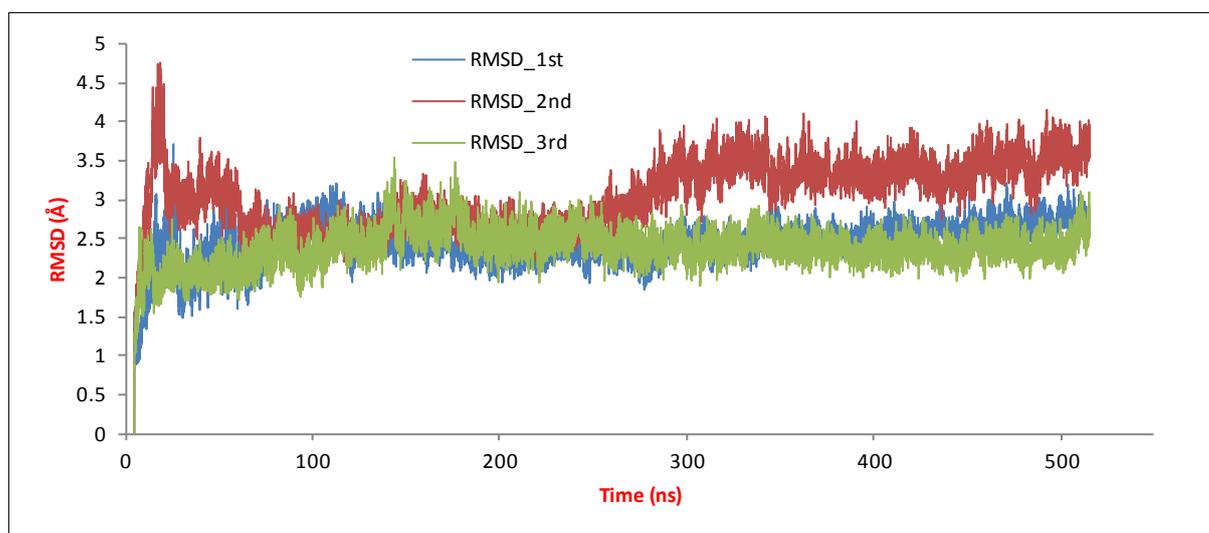


Figure 3-76: Energy and temperature graphs for the 3 simulations applied to TG2 for 500 ns.

### 3.9.1 RMSD and PCA

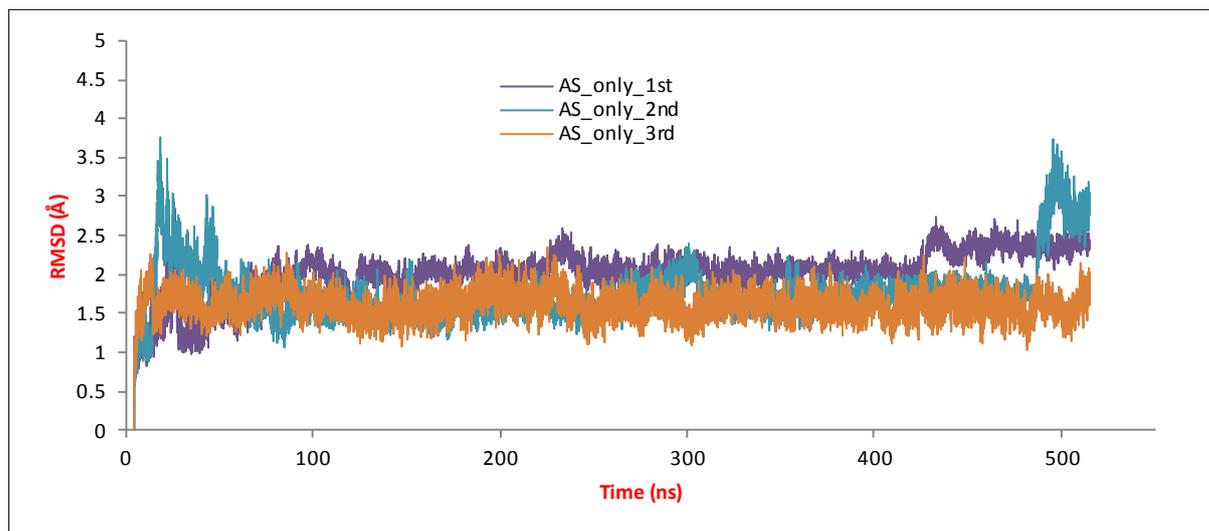
The stability of the simulations was also expressed by the RMSD of the runs. Figure 3-77 shows the RMSD for the 3 runs calculated on the production phases and using the first frame as a reference. RMSD calculations were performed on the backbone atoms only using the program CPPTRAJ from the AmberTools Package (Roe & Cheatham 2013). The RMSD for the 1<sup>st</sup> and 3<sup>rd</sup> runs was stable between 2 and 2.5 Å. In the 2<sup>nd</sup> run there was some fluctuation at the start followed by stable run at RMSD between 2 and 2.5 Å until the half of the simulation time. After that, the RMSD was stable at values between 3 and 3.3 Å. In all the cases, RMSD was almost stable during the production phase. This measurement was followed by calculating the RMSD for the residues used in defining the active site of TG2 during docking. Figure 3-78 shows this RMSD and it can be seen that the part of TG2 that constitutes the active site was moving less than the remainder of the protein. Figure 3-78 also shows that the first 100 ns of the 1<sup>st</sup> run was almost as stable as the rest of TG2, and this was the part of the trajectory from which the models for the initial docking were taken from.



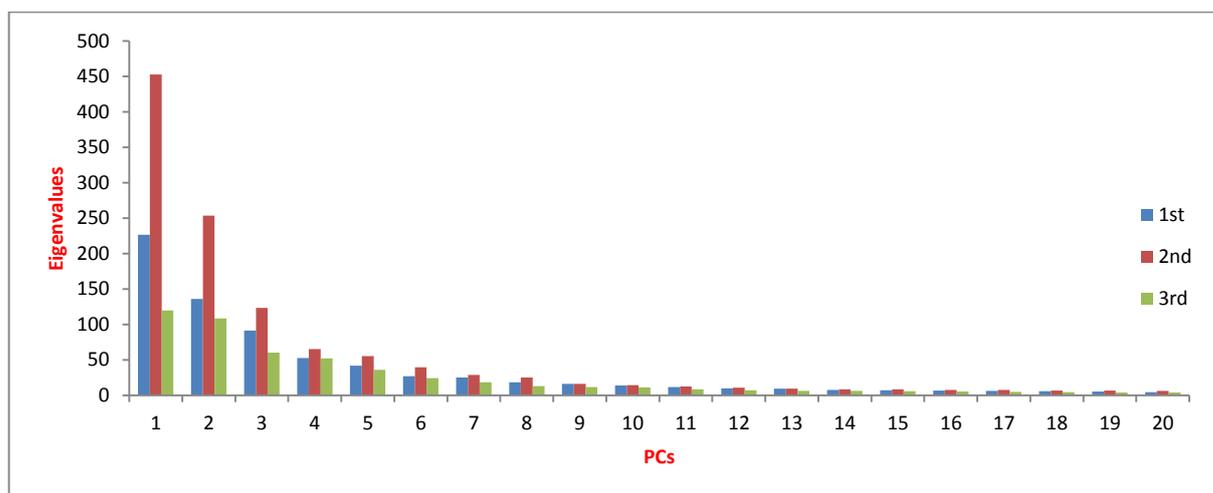
**Figure 3-77: RMSD curves for TG2 from the 3 runs.**

Following the RMSD calculations, principal component analysis (PCA) was performed using the program pyPcazip (Shkurti et al. 2016), which is a program used to compress and analyse MD simulations data. The analysis was applied on the trajectories after stripping water molecules, and included the backbone alpha carbon atoms only. The first 20 principal components (PCs) for the 3 runs are presented Figure 3-79. As shown from the RMSD graphs, it appears that the 2<sup>nd</sup> run had the highest fluctuation. This is evident in Figure 3-79 as well. When the 1<sup>st</sup> PC was plotted versus the 2<sup>nd</sup> PC, Figure 3-80 was produced. The figure shows that the 3<sup>rd</sup> run was the least flexible. However, the majority of the points from

the 3 runs were approximately in the same region, indicating similar coverage of the available conformational space in the 3 runs.



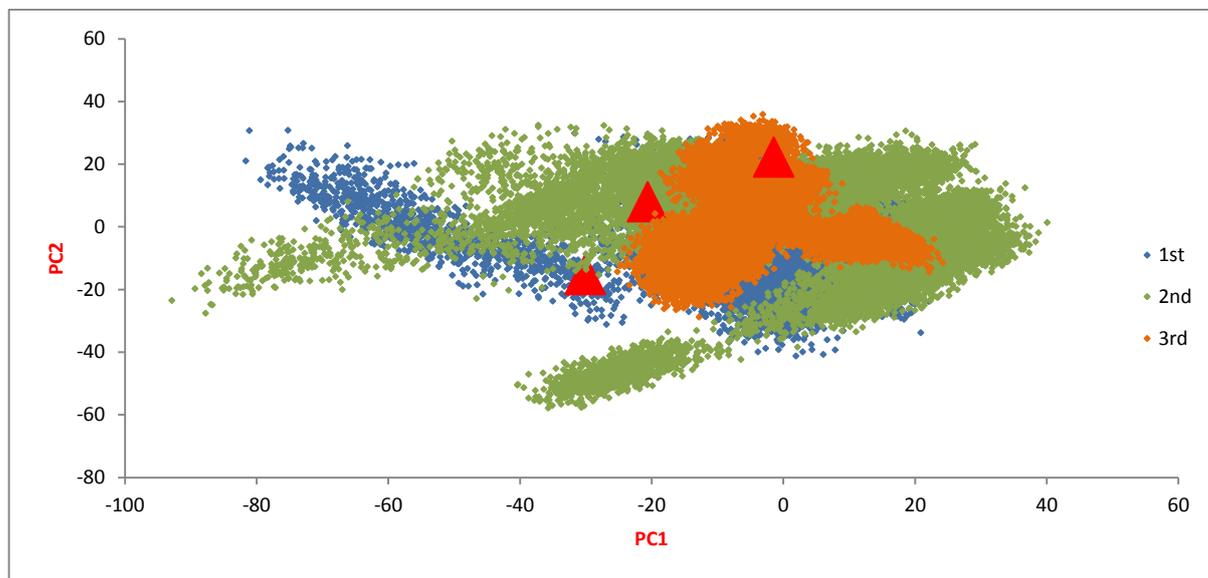
**Figure 3-78:** RMSD of the active site residues only for the 3, 500-ns runs on empty TG2.



**Figure 3-79:** The first 20 principal components (PCs) generated by pyPcazip for the 3 MD runs with their eigenvalues.

pyPcazip has the ability to derive from the principal components, a trajectory composed of 20 frames that represents the motion of the protein in that component. This feature was used to investigate which parts of the protein moved the most in the 3 runs by generating such trajectories for the 1<sup>st</sup> PC. It was noted that in the 1<sup>st</sup> run, it was the loop between residues PHE320 and GLY325 that moved the most. These 6 residues were missing in the original crystal structure and were added before starting the simulations (Figure 3-81A). Although close to the active site of TG2, this loop is not involved in the inhibition mechanism at all and its residues are not part of the active site definition used in the docking experiments. The effect of this loop on the valid active site models of TG2 has been discussed in more detail in section 3.8. For the 2<sup>nd</sup> run, it was the loop composed of residues 449-456 that had the

highest motion. The loop was originally an alpha helix that opened. This loop-opening event has not happened in the 1<sup>st</sup> or 3<sup>rd</sup> runs. In any case, it is away from the active site (Figure 3-81B). In the 3<sup>rd</sup> run, like the 1<sup>st</sup> run, it was an added loop originally missing from the crystal structure composed of residues 407-413. Again, this loop was not in the immediate vicinity of the active site.



**Figure 3-80: The first versus second PCs for the 3 runs.**

The results shown in Figure 3-81 were consistent with the atomic fluctuation data of the residues of TG2. This was calculated through measuring the average change in RMSD of the backbone atoms over time per residue. The atomic fluctuation calculation was performed using CPPTRAJ. The results are shown in Figure 3-82. Although the peak heights are different between the runs, it is obvious that the locations of the peaks are the same, indicating that they are the same regions of TG2 that are being more flexible than the others even though the extent of the flexibility is different between the runs. Also, the highest peak in each run corresponds to the same region in TG2 as the region observed from PCA in Figure 3-81.

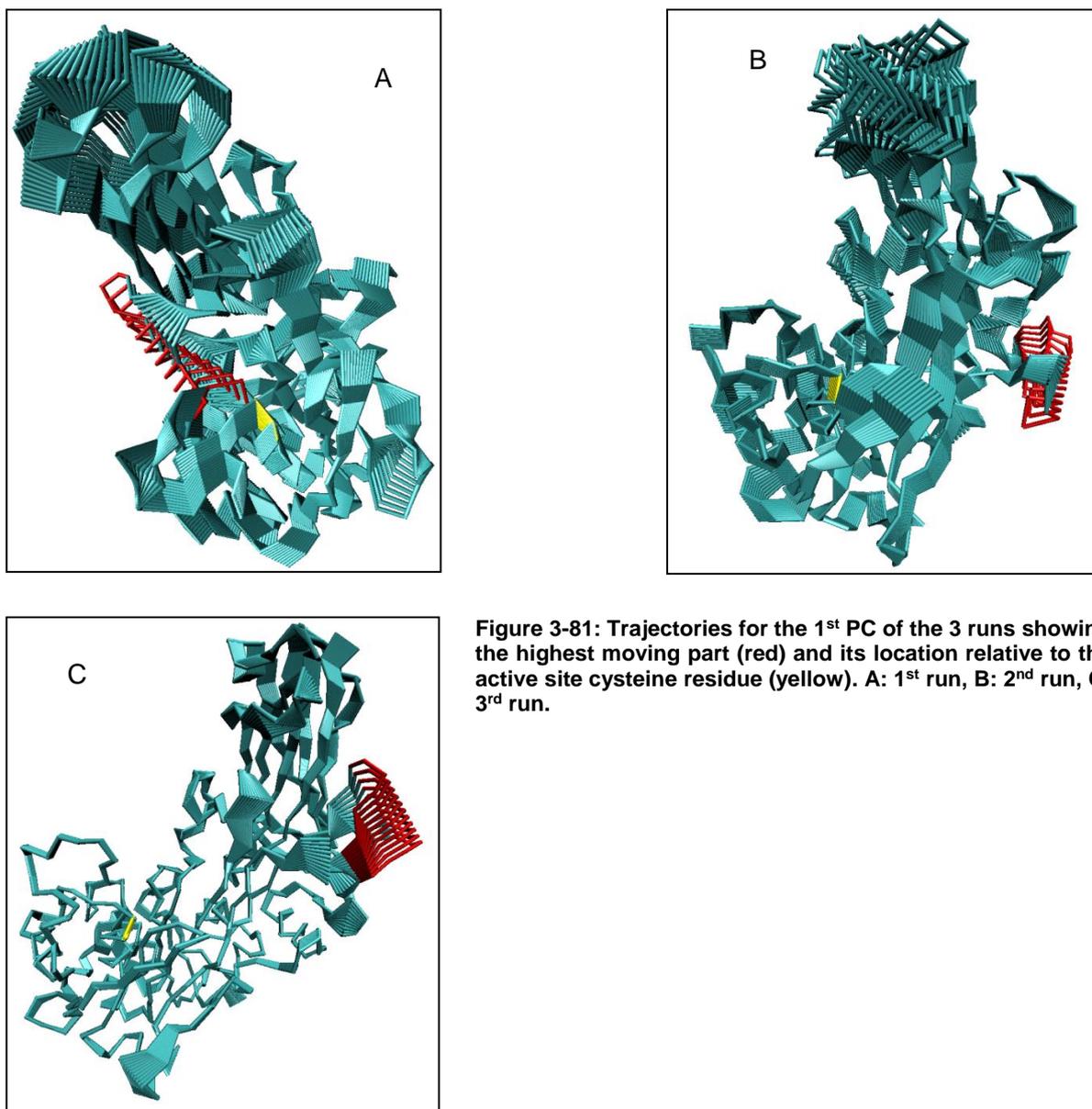


Figure 3-81: Trajectories for the 1<sup>st</sup> PC of the 3 runs showing the highest site moving part (red) and its location relative to the active site cysteine residue (yellow). A: 1<sup>st</sup> run, B: 2<sup>nd</sup> run, C: 3<sup>rd</sup> run.

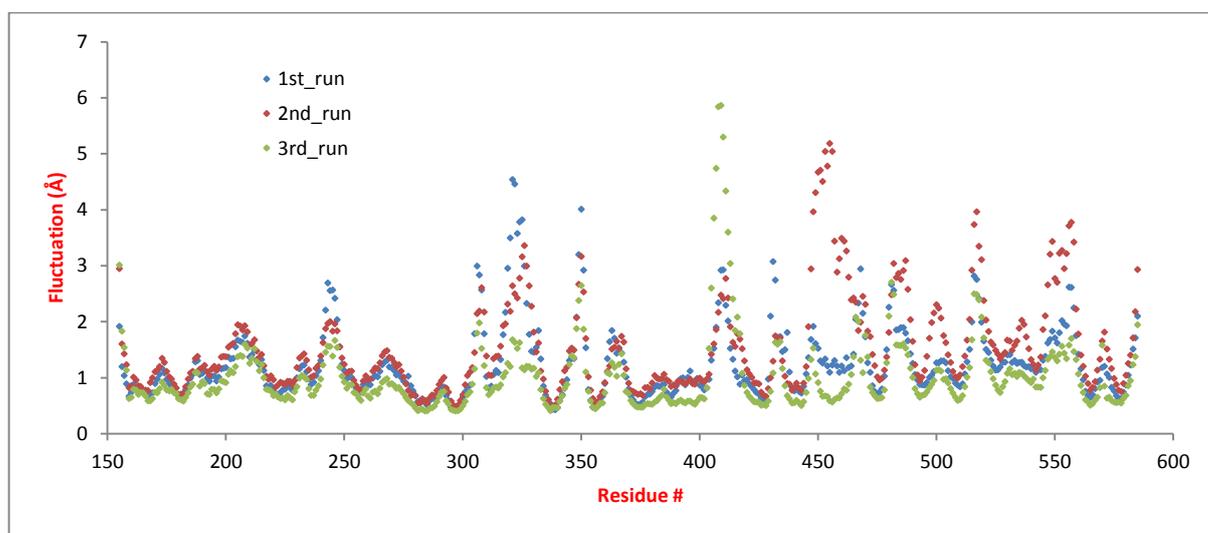


Figure 3-82: Atomic fluctuation for the individual residues of TG2 for the 3 runs.

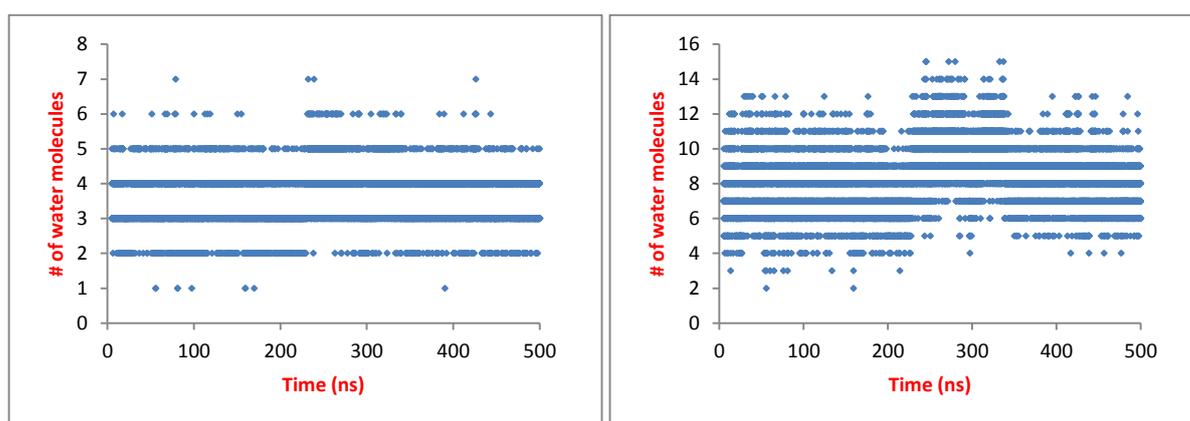
### 3.9.2 Water behaviour

How water molecules were moving within the active site of TG2 was investigated by using the “watershell” command of the CPPTRAJ program. This command calculates the number of water molecules in the 1<sup>st</sup> and 2<sup>nd</sup> solvation shells as a function of the frames in the trajectory with respect to a reference atom mask. The CPPTRAJ default dimensions of the shells were used, which are 3.4 Å for the 1<sup>st</sup> shell and 5 Å for the 2<sup>nd</sup> shell, and solvation shells were calculated with reference to the SG atom of CYS277. Table 3-33 shows the average number of water molecules in the 2 shells computed for the 3 runs. Again, the 1<sup>st</sup> and 3<sup>rd</sup> runs appear to be very similar having an average of  $\approx 3.5$  water molecules within 3.4 Å and 8 water molecules within 5 Å across the simulations. Because the numbers of the 1<sup>st</sup> and 3<sup>rd</sup> runs were similar, those runs will be considered for comparisons.

	1 <sup>st</sup> Run	2 <sup>nd</sup> Run	3 <sup>rd</sup> Run
1 <sup>st</sup> Shell Average	3.5	2.6	3.4
2 <sup>nd</sup> Shell Average	8.2	4.8	8.0

**Table 3-33: Average values for the number of water molecules in the 1<sup>st</sup> and 2<sup>nd</sup> solvation shells in the 3 MD runs applied on empty TG2.**

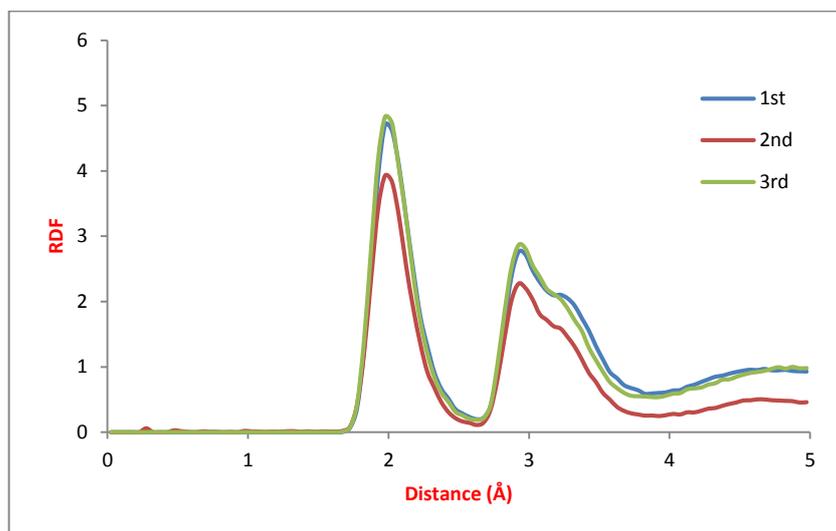
Figure 3-83 is graphical representation for the number of water molecules in the 1<sup>st</sup> and 2<sup>nd</sup> shells during the 1<sup>st</sup> MD run and it shows that within the 1<sup>st</sup> shell, there was always at least one water molecules within 3.4 Å of SG (this occurred in 9 frames). The maximum was 7 residues (4 frames only), and for the majority of time, the number was between 3 and 5, explaining the average values shown in Table 3-33. The same applies to the graph of the 2<sup>nd</sup> shell. This indicates that there were always few water residues in the immediate vicinity of SG in the active site. To investigate how that is related to distance from SG, RDF was used.



**Figure 3-83: Water molecules in the 1<sup>st</sup> (left) and 2<sup>nd</sup> (right) solvation shells of the 1<sup>st</sup> MD run as representative of distribution of water molecules in the 3 runs.**

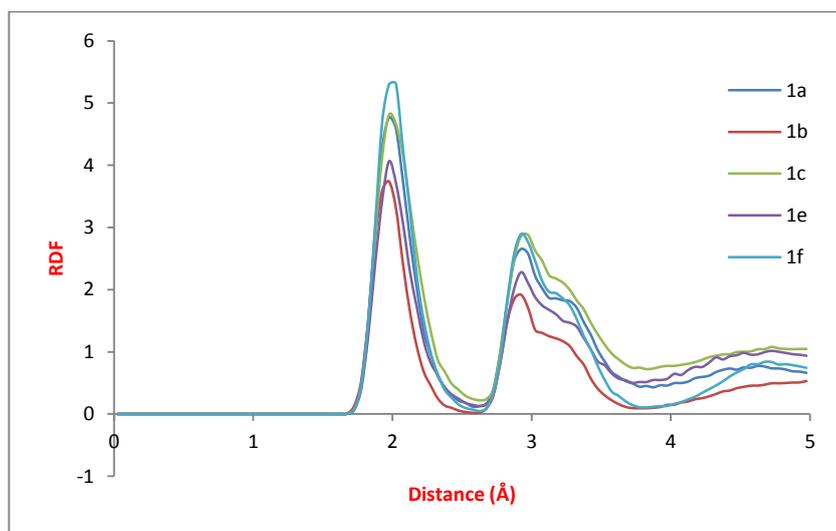
The radial distribution function (RDF) was used to inspect how the density of water molecules around SG changed as a function of distance from SG. The results of the calculation are presented in Figure 3-84, in which a similar pattern can be observed within

the 3 runs. Within 2 Å of SG, there was 5 times higher probability of finding water molecules than everywhere else in the 1<sup>st</sup> and 3<sup>rd</sup> runs and 4 times in the 2<sup>nd</sup> run. The probability decreases to about 3 within 3 Å of SG in the 3 runs. RDF goes down after 3 Å to around 1. This means that most of the water molecules seen in Figure 3-83 are between 2 and 3 Å and there are no water molecules that are less than 2 Å away from SG, due to atom radii restrictions.



**Figure 3-84: RDF of water molecules relative to SG atom of CYS277 in the 3 MD runs applied to empty TG2.**

The similar water behaviour in the 3 runs strengthens the assumption that this is how water molecules distribute themselves around the catalytic cysteine. The same calculation of water RDF was applied on the 275-ns trajectories of the active compounds. Only compound **1d** trajectory was not considered due to the compound leaving the active site before the start of the production phase. Figure 3-85 shows RDF values for the 5 trajectories. It can be seen that the behaviour is very similar in the 2 cases, indicating that the presence of the inhibitors within the active site of TG2 did not affect the density of water around SG of CYS277.



**Figure 3-85: RDF of water molecules in the 275-ns trajectories of the active compounds.**

It is true that the trajectories in Figure 3-85 had inhibitors, but it is only the one for compound **1b** that managed to finish with the inhibitor in its starting pose; in the rest, the warhead has left the catalytic tunnel at some point during the simulation. The compound **1e**-simulation had the compound in its appropriate conformation for about 185 ns of the simulation time. A closer look at the graph in Figure 3-85 shows that within 2 Å, the probability of finding water was lower in the trajectories for compounds **1b** and **1e**. Compared to an RDF of 5 in the 1<sup>st</sup> and 3<sup>rd</sup> runs on empty TG2 and in the simulations for complexes in which the warhead has left the catalytic tunnel, this may actually mean that the presence of the inhibitor can reduce the density of water in the vicinity of SG, but it is until the covalent bond is formed, that all the water molecules within 2 Å will all be pushed away. This was shown in section 3.7 when discussing covalent docking and MD.

An analysis of the number of water molecules in the 1<sup>st</sup> solvation shell was also performed on the trajectories of the complexes and showed that the lowest average was 1.9 molecules and that was for the trajectory of compound **1b** and the highest was for compound **1c** with an average of 3.6 water molecules. For the remaining 3 compounds, the averages were 3.1, 2.9 and 3 for compounds **1a**, **1e** and **1f** respectively. The lowest average for compound **1b** is compatible with the RDF for the trajectory of that compound and compound **1e** had the 2<sup>nd</sup> lowest average. Although the averages for the other simulations are not quite similar to the averages of the empty TG2 runs, the fact that they are higher than the averages of simulations in which the warhead managed to stay close to SG indicates that the presence of the warhead can push some water molecules away.

The distribution of water molecules in the hydrophobic loop of TG2 has also been analysed. This has been calculated as RDF with reference to the alpha carbon atom of residue ILE331 (Figure 3-86), which is one of the residues involved in the hydrophobic interactions with the lipophilic part of the inhibitors. The results are shown in Figure 3-87. Although residue ILE331 on its own does not represent the hydrophobic loop, it is one of the best choices available. RDF graphs for the 3 runs are very similar and show that there are no water molecules in the immediate vicinity of ILE331. RDF shows that water molecules start to appear at around 3 Å from ILE331 but at very low density when compared to SG of CYS277 (RDF has a value of  $\approx 1$  at 3 Å which is the normal density of water). This was confirmed by calculating the numbers of water molecules residing in the 1<sup>st</sup> solvation shell which averaged around 0.4, 0.6 and 0.7 water molecules for the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> MD runs respectively.

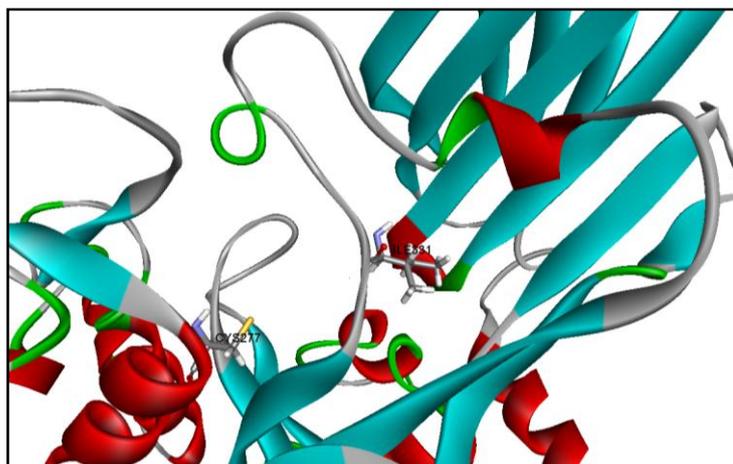


Figure 3-86: The location of ILE331 used as reference for the RDF of water molecules in the hydrophobic loop of TG2.

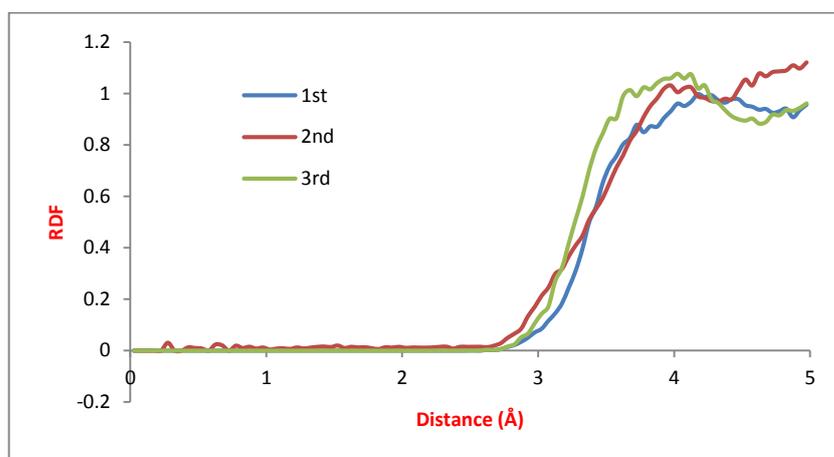


Figure 3-87: RDF of water molecules in the 3 runs in the hydrophobic loop of TG2.

The absence of water molecules from the hydrophobic loop in the simulation of the empty TG2 may provide an explanation for the apparent ease with which the lipophilic parts of the compounds maintained their positions within the hydrophobic loop, when compared to the warheads in the 275-ns and in the 5-ns simulations. There were much more cases for the leaving of the warhead than for the lipophilic part in the 5-ns simulations (section 3.4.5). In the 275-ns simulations, no lipophilic part left its place.

A very similar pattern was seen when inspecting the RDF graphs for the density of water around ILE331 in the 5 trajectories for the compounds. The results are shown in Figure 3-88. The absence of water until 3 Å from ILE331 is very clear, followed by the slight increase in the density. Compound **1f** trajectory showed the least density of water despite the fact that the simulation for this compound started with the lipophilic part outside the hydrophobic loop and continued like that for the remainder of the simulation.

In general, the results of RDF analysis in the hydrophobic loop of TG2 show that it is highly unlikely for water molecules to reside underneath the loop. The presence of the lipophilic part of the inhibitors within the hydrophobic loop does not appear to reduce the probability of finding water molecules in the loop, as it would be expected from the physical occupancy,

since in some cases when there was a compound, the probability was actually higher than it was in empty TG2 (trajectory for compound **1b**).

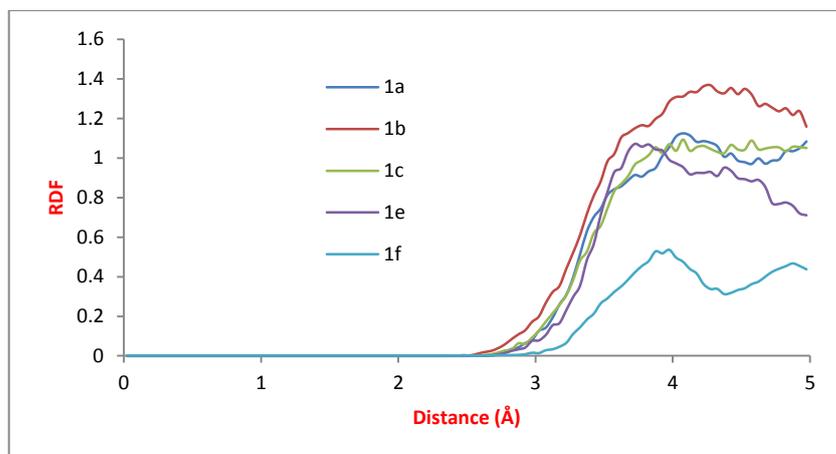


Figure 3-88: RDF graphs for the hydrophobic loop in the trajectories of complexes.

### 3.10 Conclusions

The main goal of the work at this point was to derive active site models of TG2 that could be used to test potential inhibitors of TG2 and to validate the models. The initial MD simulation applied to TG2 was proven to be stable and did succeed in providing starting structures for MD applied on the complexes in the next stage. The MD simulations applied to the complexes finished appropriately only for compound **1b** in terms of maintaining the initial pose of the compound; but in all the simulations, the importance of hydrogen bonding with key residues in the active site (GLN276, ASN333 and PHE334) and of the hydrophobic interactions involving the lipophilic parts of the inhibitors has been shown. Although the MD simulation of compound **1a** failed to keep the compound in the active site, the trajectory provided 3 conformations that passed the 1<sup>st</sup> stage of testing which involved docking 6 active inhibitors of TG2. This supplied 3 of the 6 active site models that went to the next stage and the remaining 3 were extracted from the trajectory of compound **1b**. The dockings at this stage also signified the importance of hydrogen bonding and hydrophobic interactions.

The six models were validated next through various experiments that involved docking of inactive compounds, docking with water in the active site, expanding the compound test set, applying short MD simulations that involved calculating the binding free energies and finally docking of more active compounds. Even though the results were not always ideal, they gave the suggestion that the models can detect compounds based on their ability to inhibit TG2. The results also enabled the fabrication of criteria required to define a good docking complex for TG2 inhibitors. A positive correlation was also obtained when the binding free energy was related to the biological activity for compounds containing the sulfonium ion warhead.

The choice of GoldScore as the primary scoring function during the GOLD dockings performed was justified by testing the performance of the other functions which proved to be inferior to GoldScore. No scoring function, however, was able to rank the compounds based on their biological activity, although with GoldScore, this was slightly improved by rescoring after local minimisation of the docked pose.

Covalent docking and MD proved that active and inactive inhibitors of TG2 behave similarly if the covalent bond was already there. They also proved that no water molecules can reside next to the covalent bond during MD simulations.

The analysis performed on TG2 conformations of the 6 valid models succeeded in explaining the better results obtained with the models from **1a** trajectory than those from **1b** trajectory and showed that it is better to use the 6 models when trying to test some new compounds as inhibitors of TG2 if GOLD was to be used as the docking program.

The behaviour of water molecules in the MD simulations has shown that there could be few water molecules residing very close to SG of CYS277, whether there was an inhibitor in the active site or not. The analysis also showed that very few to no water molecules can be in the hydrophobic loop of TG2 which may explain why it was easier to keep the lipophilic part in place than to keep the warhead.

To conclude, the models generated during this work have been validated and proven to be the best available choices for testing TG2 inhibitors. More work is required to investigate the mechanism of inhibition to fill some of the gaps in the work so far related to the failure to relate some measured quantity to biological activity. This was the main rationale for the next chapter of the thesis.

# **Chapter 4**

# **Quantum Mechanical**

# **Experiments**

## 4 Quantum Mechanical Experiments

This chapter covers the results of the experiments performed with partial or total QM treatment to investigate the mechanism by which the various studied inhibitors react with TG2. Partial treatment was employed through a combined QM/MM approach with umbrella sampling on selected complexes of TG2 with acrylamide based inhibitors. Total QM treatment was applied on versions of TG2-sulfonium ion inhibitors complexes in which all the system was removed with the exception of the inhibitor and the catalytic cysteine residue and QM reaction path experiments were performed.

### 4.1 Umbrella Sampling

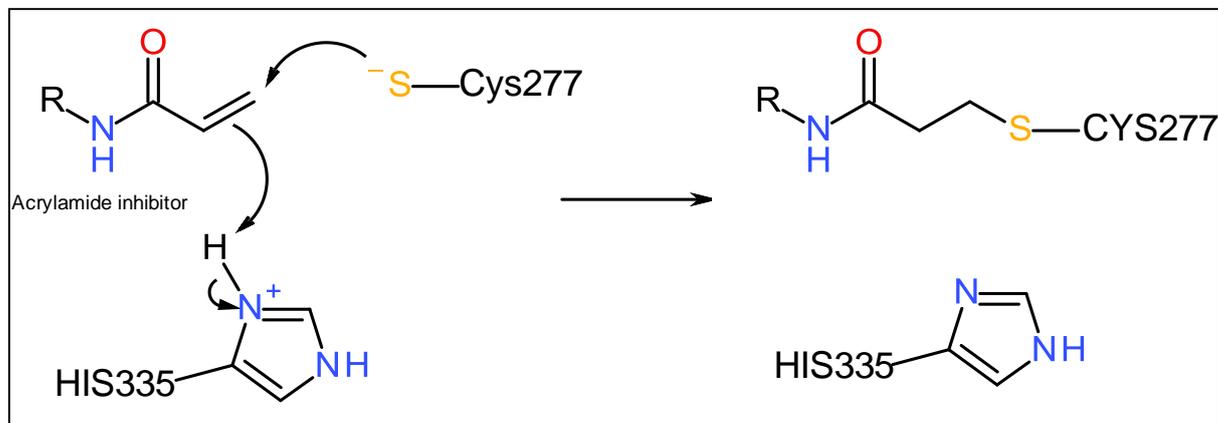
#### 4.1.1 Introduction

Umbrella sampling (US) is an enhanced sampling technique in molecular dynamics that permits the exploration of events which would otherwise take infinitely long time to occur. Candidate events would ideally involve crossing high energy barriers where conventional MD will only result in trapping the system in an energy minimum with no possibility of crossing the barrier. US applies an external biasing potential over a predefined reaction coordinate (RC) to drive the event to completion. The change in RC should be small to ensure a quasi-static process. Thus US spreads the change in RC over windows and for successful sampling, the windows must overlap (Kästner 2011). During the sampling, the change in RC in each window is recorded to represent the probability distribution along the reaction coordinate in that window. This distribution can then be converted to the potential of mean force (PMF) using the weighted histogram analysis method (WHAM) (Hub et al. 2010; Kumar et al. 1995).

The next section introduces the results of the umbrella sampling applied to simulate the reaction between TG2 inhibitors with the acrylamide warhead and TG2 active site cysteine (CYS277) (Figure 4-1). The inhibitors are compounds **1a**, **1c**, **1e** (Table 3-1, p86), **3j**, **3l** and **3o** (Table 3-12, p125) from the work of Badarau et al. (2015) and the associated patent (Griffin et al. 2014). These compounds had a range of activity expressed as  $IC_{50}$  from 0.006 to 6.3  $\mu$ M.

The aim of the study is to use the US-simulated reaction to find some type of correlation with biological activity manifested as  $IC_{50}$  values for the tested acrylamide inhibitors. Since the formation of bonds was involved, US was performed with quantum mechanical treatment of the reaction centre (the acrylamide inhibitor as the electrophile, CYS277 as the nucleophile

and HIS335 from TG2 as the source of the proton), while the remainder of the protein and the water box were treated molecular mechanically. PM3 (Parametric Method Number 3) (Stewart 1989) and SCC-DFTB (self-consistent-charge density-functional tight-binding) (Elstner et al. 1998) were used as the QM methods.



**Figure 4-1:** The reaction of inhibition of TG2 by acrylamide inhibitors.

Two possible mechanisms for the reaction were explored; the first involves a single stage only during which the bond between nucleophilic sulphur of CYS277 (SG) and the electrophilic carbon (EC) on the acrylamide, and the bond between a proton from a nearby histidine residue (HD1) and the alpha carbon of the double bond of the acrylamide (C2) both form simultaneously. The second mechanism involved the nucleophile-electrophile reaction first, followed by the protonation stage. The choice of HIS335 as the proton source is justified in a later section of this chapter (4.1.2.4). The starting structures were frames taken from the 5-ns MD simulations (**1a**, **1c** and **1e**) or docking complexes from validation processes 3 (more test compounds) and 4 (all compounds) (**3j**, **3l** and **3o**).

### 4.1.2 Results and Discussion

The overlap between the windows of all the simulations was analysed by plotting the time series files where the Y-axis represented the change in RC across the simulation and the X-axis the time of a single window in picoseconds. This was used as a measure to the success of the simulation. Other measures used to judge the quality of an individual simulation were the actual formation of the 2 bonds; SG-EC and HD1-C2, and the fact that the compound maintained its general pose within TG2 active site at the end of the simulation. In addition, RMSD of TG2 in the different simulations was measured. Figure 4-2 has been presented in the Methods (section 2.2.2.1) and is presented here as a reminder of the distances used to define reaction coordinates used in the 2 mechanisms.

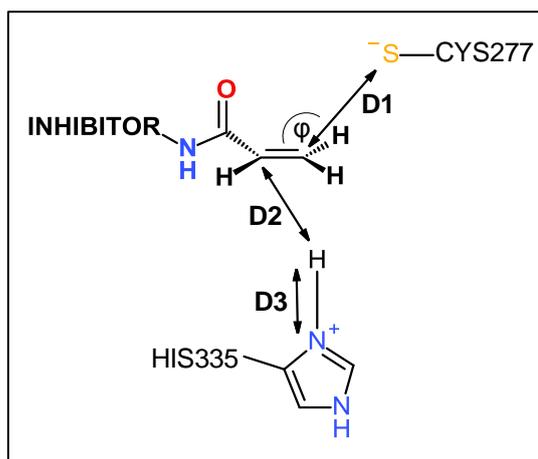
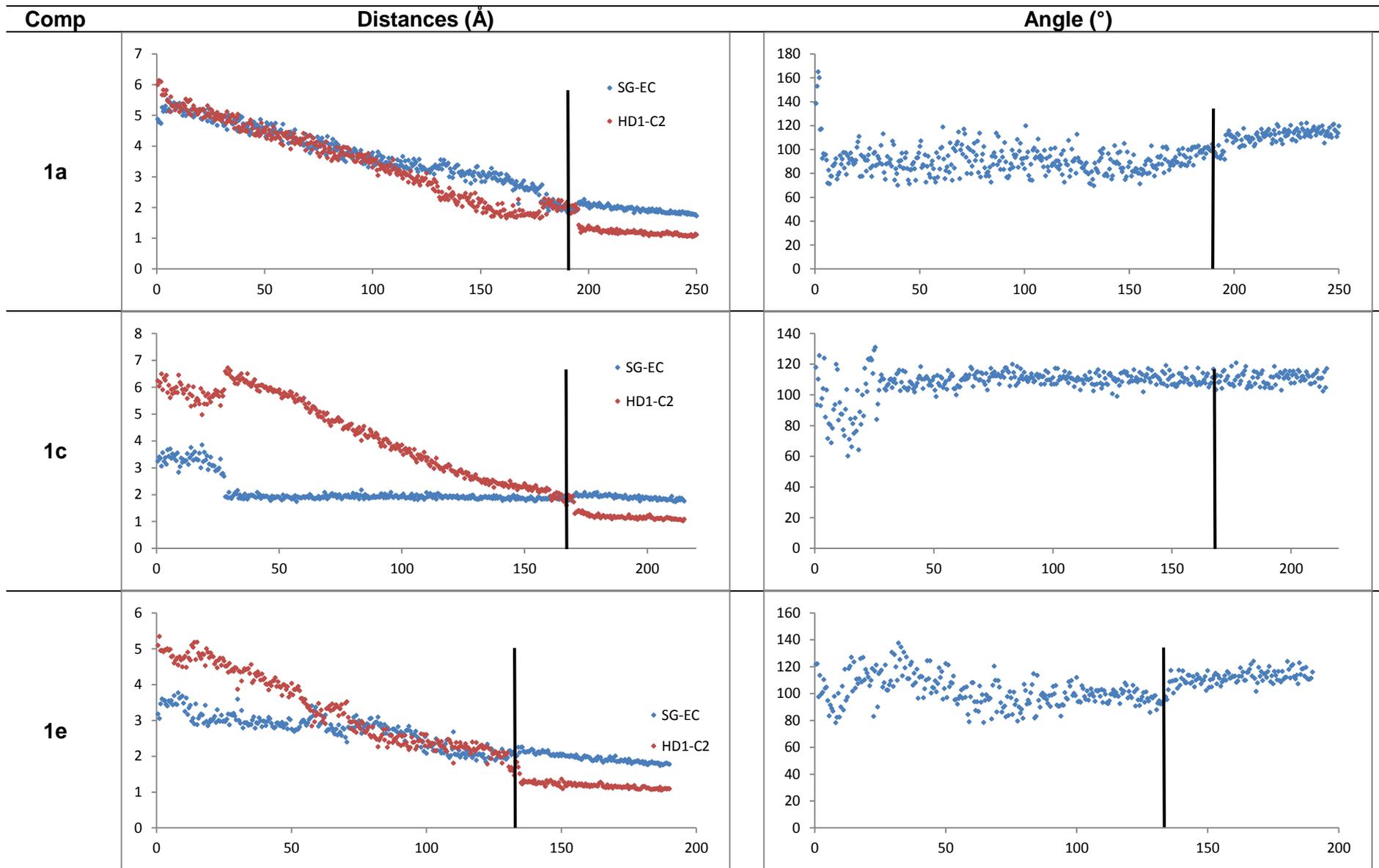


Figure 4-2: Distances used during umbrella sampling for the 1-stage and 2-stage simulations.

#### 4.1.2.1 Single stage simulations

In these simulations, the reaction coordinate (RC) was set as a generalised distance coordinate comprising the sum of D1 and D2 in Figure 4-2. This definition of the RC allows for the whole reaction to proceed in a single stage. SCC-DFTB was unable to drive the reaction to completion in compound **1c**, and for the remaining 5 compounds, no useful correlation with the biological activity could be obtained from the US simulations. For these reasons, US using SCC-DFTB on acrylamide compounds assuming 1 stage for the reaction was not considered. Such anomalies did not occur when PM3 was set as the QM method. The simulations with all the compounds completed to the assigned value of the reaction coordinate, the 2 bonds formed, and the compounds maintained their original pose within TG2. In addition, a sufficient degree of overlap between the windows was observed.

The trajectories of the simulations on individual compounds have been analysed. The distances involved in RC have been measured using VMD program (Humphrey et al. 1996) along with the SG-EC-C2 angle. The analysis showed that the 2 distances were declining together but the SG-EC bond would form first and that the angle would change from its starting value to a value between 94° and 115° at the barrier height and maintains that for the rest of the simulation (Table 4-1). The dihedral angle of the attack of SG on the plane of EC (EC, C2 and the carbonyl carbon) has also been measured. The discontinuities seen with compounds **1c** and **3o** in Table 4-1 were not flaws in the way in which the RC has changed during the US simulation for these compounds. Rather, the SG-EC bond was formed suddenly and early in the simulation, where the SG-EC distance jumped from a value of 2.7 Å in one frame of the trajectory for the simulation of compound **1c** to a value of 1.9 Å in the next frame. This sudden change was accompanied by a temporary corresponding increase in the HD1-C2 distance from 5.9 Å to 6.6 Å to keep the change in the overall RC consistent.



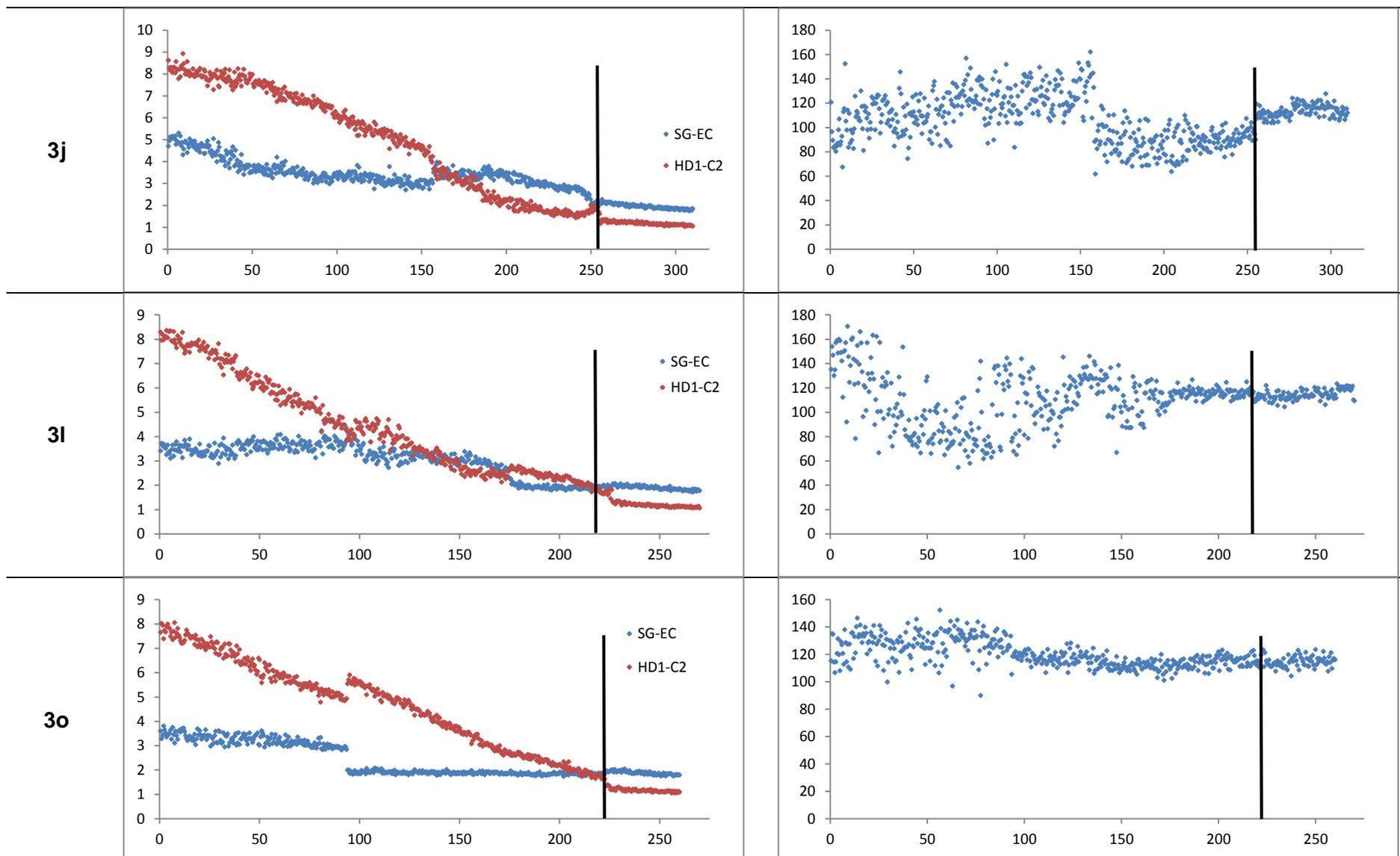
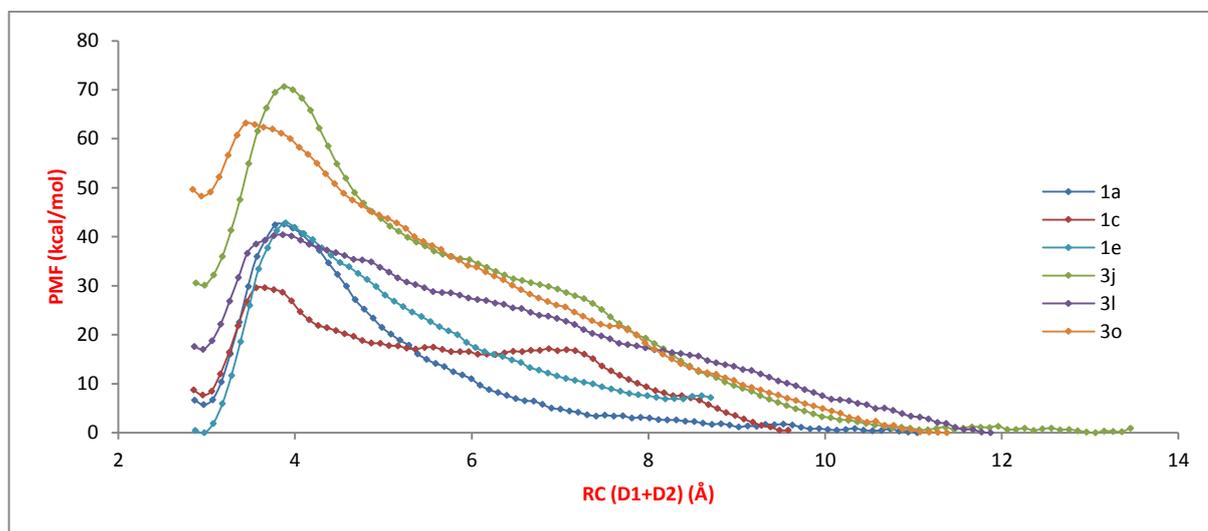


Table 4-1: Distances involved in the RC for the 1-stage simulations and the corresponding angles for the SG-EC-C2 bond. X-axis is time in ps and Y-axis is in Å for the RC distances and degrees (°) for the angle. The black line represents the point in the simulation corresponding to the barrier height on the PMF graph.

As stated in the Methods section (section 2.2.2.1), PMF was calculated from the time series files using WHAM. The results of PMF calculations for all the compounds are presented in Figure 4-3. It can be seen that a similar path has been followed by all the compounds during their US simulations and that the barrier height was reached at a roughly similar value for the reaction coordinate in all the compounds.



**Figure 4-3:** PMF graphs of the 6 acrylamide compounds generated from WHAM program. The graph starts from the right with the higher value from RC and ends to the left.

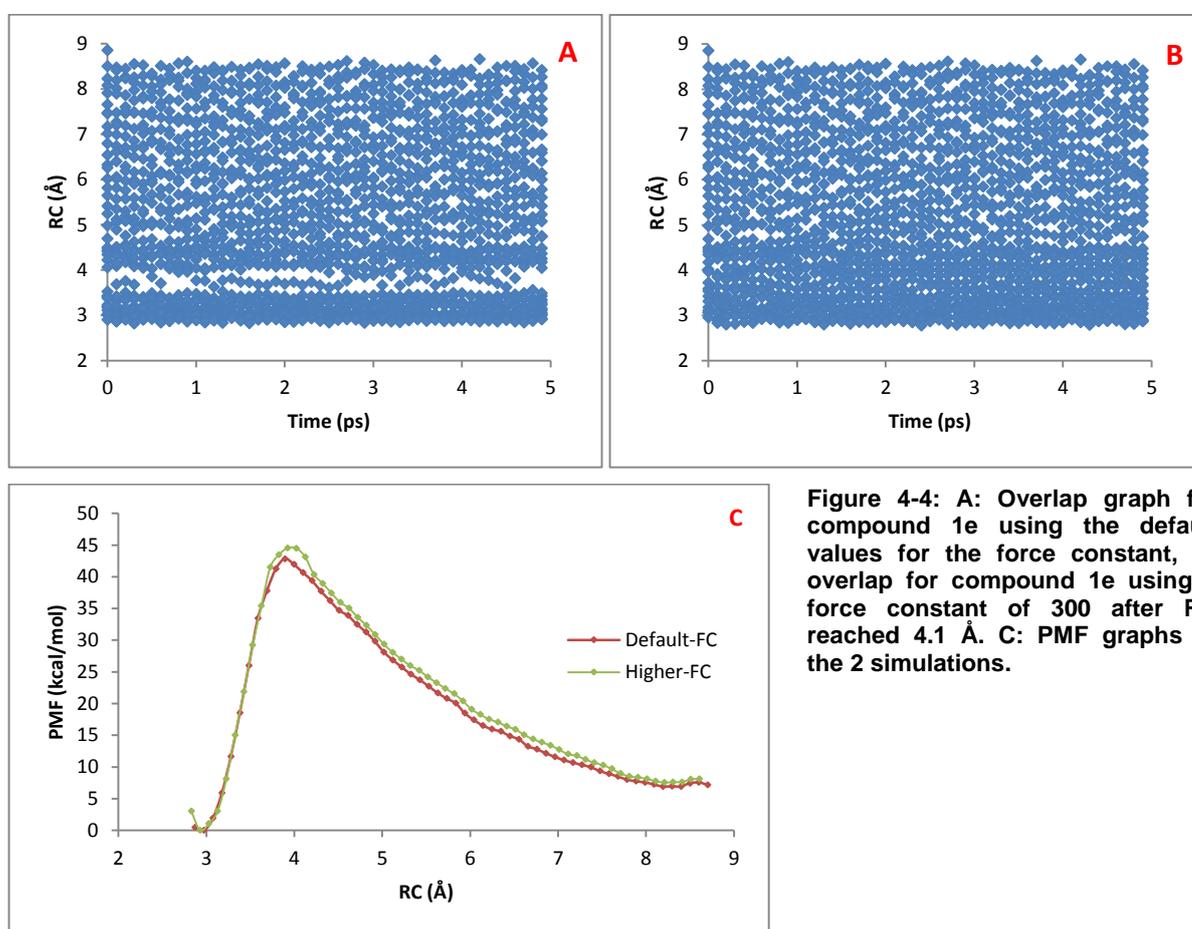
The values of the parameters measured from the US MD trajectories for the structures at the barrier heights of the individual PMF graphs are presented in Table 4-2. The dihedral angle is presented in its simplest positive form to facilitate the comparisons. These structures are the closest one can get to transition states (TS). It can be observed from Table 4-2 that at the barrier, the SG-EC bond was formed only in **1a** and **3o** (assuming  $\approx 1.8$  Å for S-C bond) while HD1-C2 bond was not formed in any of the tested compounds. The value of RC ranged from 3.5-3.9 Å, the angle from 94-115° and the dihedral angle had values between 56-87°.

Compound	TS parameters				
	SG-EC (D1) (Å)	HD1-C2 (D2) (Å)	RC (Å)	Angle ( $\varphi$ ) (°)	Dihedral (°)
<b>1a</b>	1.83	2.06	3.89	98.35	56
<b>1c</b>	1.94	1.61	3.54	111.51	72
<b>1e</b>	2.03	1.85	3.89	94.09	80
<b>3j</b>	2.15	1.71	3.85	94.47	61
<b>3l</b>	1.96	1.91	3.87	114.73	87
<b>3o</b>	1.83	1.63	3.46	110.71	73

**Table 4-2:** RC distances and SG-EC-C2 and dihedral angles for TS structures of the 6 compounds.

The fact that the 2 bonds have formed indicates that the chosen reaction coordinate could represent the reaction successfully. The end value for the RC was appropriate; the RC could not be pushed any further as this would have resulted in a shift in PMF scores, as there was

already a very slight shift at the end of the graphs (Figure 4-3). The overlap between the windows was checked by plotting the time series of the RC of all the windows using time values for the X-axis that correspond to a single window (5 ps) (Figure 4-4A). The gap seen in Figure 4-4A at  $\approx 4$  Å of overlap graph for compound **1e**, has been spotted in the overlap graphs for other compounds. It corresponds to the barrier height as seen in Table 4-2. This may be expected due to the formation of the bond at this point, which could occur with a sudden, and rough, change in the RC. A simulation for compound **1e** was run in which the force constant was increased from 100 to 300 kcal/mol.Å<sup>2</sup> starting from the value of 4.1 Å for RC to the end of the simulation. The gap has disappeared from the overlap graph but the PMF graph produced was almost exactly the same (Figure 4-4C). It is for this that the default force constant was considered, to avoid using higher than necessary force constant.



**Figure 4-4:** A: Overlap graph for compound **1e** using the default values for the force constant, B: overlap for compound **1e** using a force constant of 300 after RC reached 4.1 Å. C: PMF graphs of the 2 simulations.

The reaction is expected to proceed as a Michael addition; the acrylamide double bond acts as the acceptor being attached to an electron withdrawing group, and SG of CYS277 being the donor. This can be viewed as a similar mechanism to the nucleophilic addition to carbon-oxygen double bonds (Smith & March 2001). Hence, it may be possible to apply the Bürgi-Dunitz angle to this reaction. This is the angle of the attack of a nucleophile (Nu) to a carbonyl group carbon atom (Nu-C-O angle) and its optimal value was found to be 107°

(Bürgi et al. 1974). The same angle has actually been proposed for any reaction that involves a Michael addition (Sinnott 2007).

It is true that the angles in Table 4-2 do not all agree with the value of  $107^\circ$  stated in the literature for this angle but they are close. After the barrier height and the formation of the SG-EC bond, the angle keeps a value that is close to the values in Table 4-2 for the remaining of the simulation as can be seen from the angle graphs in Table 4-1. At this point the angle would represent that for an ordinary S-C-C bond. According to AMBER force field, such angle has a value of  $108.6^\circ$  in normal cysteine residue and by looking at the graphs in Table 4-1; it can be argued that the bond formed is behaving as a typical bond involving sulphur and  $sp^3$  carbon.

Another angle was measured and that was the dihedral angle between the attacking sulphur and the plane of EC, C2 and the carbonyl carbon (C) (Figure 4-5). The angle assumed very diverse distribution among the compounds and this can be attributed to MD oscillations. The results for 2 compounds (**1a** and **3j**) are presented in Figure 4-6 as examples of the diversity. Due to the variations, the simplest positive magnitude at the barrier height was considered and presented in Table 4-2. The common feature within those angles is that the 2 planes were virtually vertical to each other. This indicates that the attack of the nucleophilic SG of CYS277 was almost perpendicular to the plane containing the acrylamide warhead (Figure 4-5) or at least not in the plane of the hydrogen atoms attached to EC. Dihedral angle values are therefore consistent with the values of the SG-EC-C2 angles.

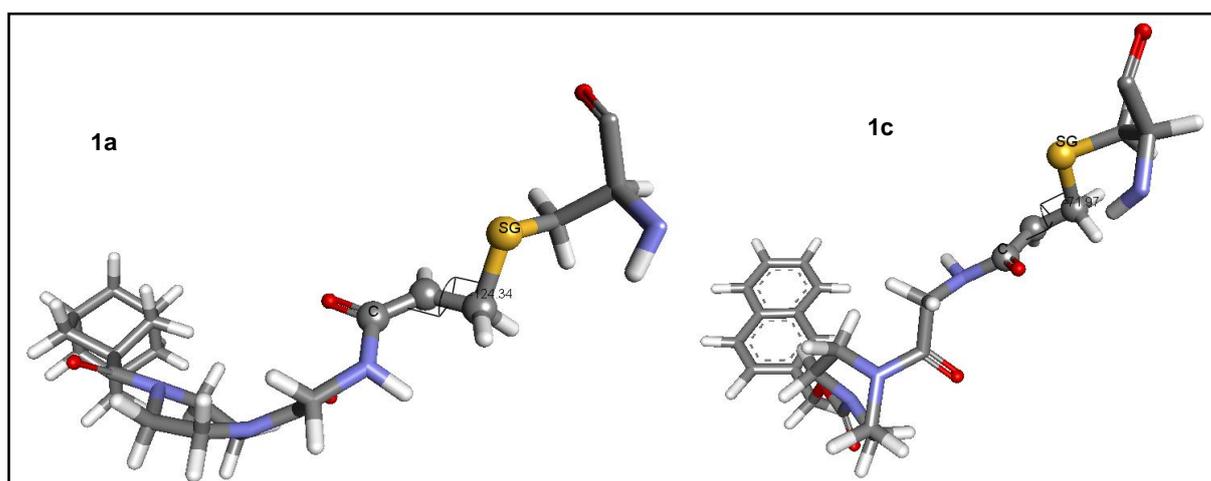
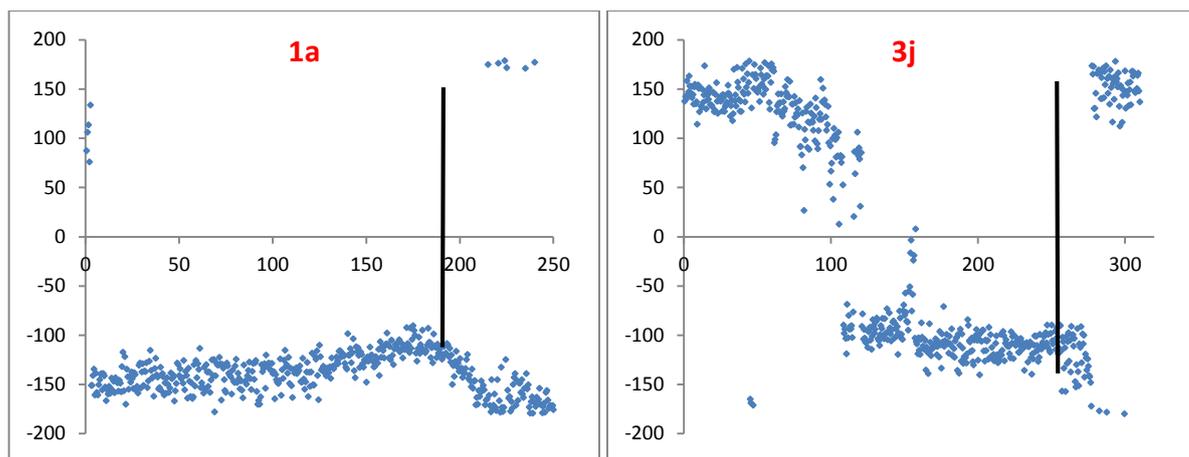
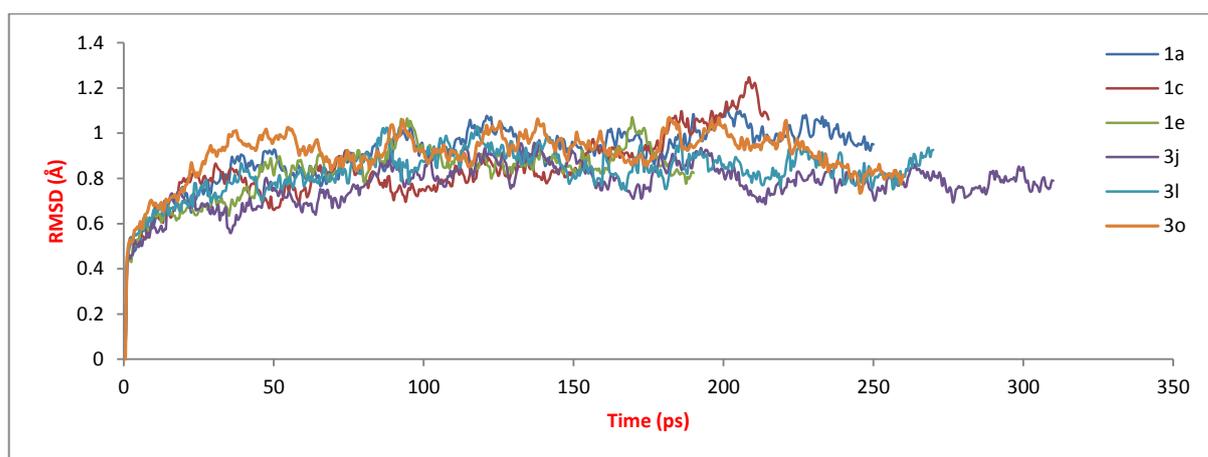


Figure 4-5: The dihedral angle that was measured during umbrella sampling between SG from CYS277 and the plane of EC, C2 and C of the compounds, showing two examples for the angle with compounds 1a and 1c.



**Figure 4-6:** Graphs of the dihedral angle in compounds 1a and 3j as representatives of the 6 compounds showing the diversity in the values of the angle. The black line vertical represents the barrier height in the PMF graph. X-axis is time in ps and Y-axis is dihedral angle in degrees ( $^{\circ}$ ).

The RMSD for TG2 in the simulations of the 6 compounds was measured as an indicator of the stability of those simulations, using the 1<sup>st</sup> frame of the trajectory as a reference. The results are presented in Figure 4-7 and imply that TG2 was stable. The RMSD in the 6 compounds had a maximum value of 1.25 Å.

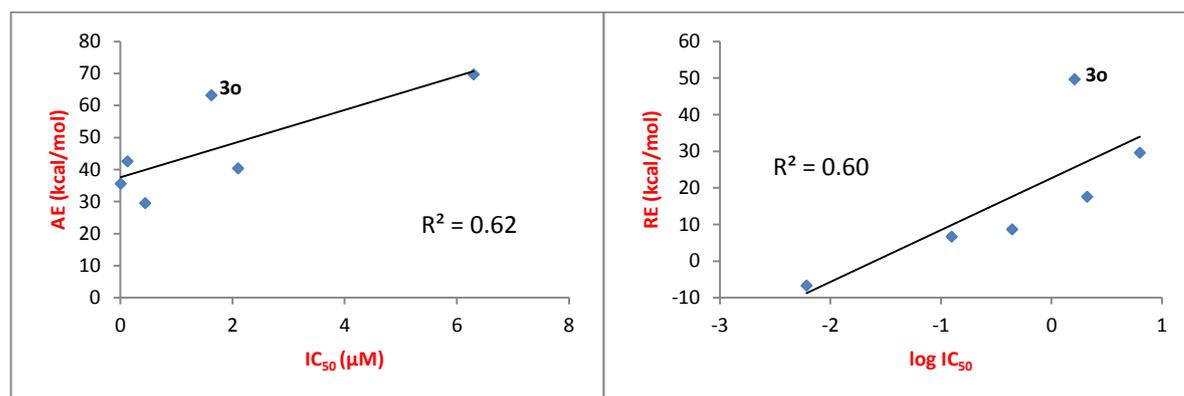


**Figure 4-7:** RMSD graphs for the simulations of the 6 acrylamide compounds.

Finally, the activation energy (AE) for the reaction was calculated, and was taken as the difference between the barrier height and the energy of the reactants; the latter being zero in most cases. The reaction energy (RE) was also computed by subtracting the energy of the reactants from that of the products. The results for the different compounds are summarised in Table 4-3. When the correlation between biological activity (as TG2  $IC_{50}$ ) and the AE was considered, a line with positive slope was obtained indicating an increase in AE with a decrease in biological activity. The line had an  $R^2 \approx 0.6$ . Interestingly, the reaction energy gave a similar line when plotted against the log value of the  $IC_{50}$  (Figure 4-8).

Compound	IC <sub>50</sub> (μM)	AE (kcal/mol)	RE (kcal/mol)
<b>1a</b>	0.125	42.55	6.66
<b>1c</b>	0.44	29.57	8.70
<b>1e</b>	0.0061	35.66	-6.70
<b>3j</b>	6.3	69.71	29.66
<b>3l</b>	2.1	40.44	17.57
<b>3o</b>	1.625	63.19	49.69

**Table 4-3: AE and RE (both in kcal/mol) for the 6 acrylamide compounds resulting from US simulations using PM3 as the QM method.**

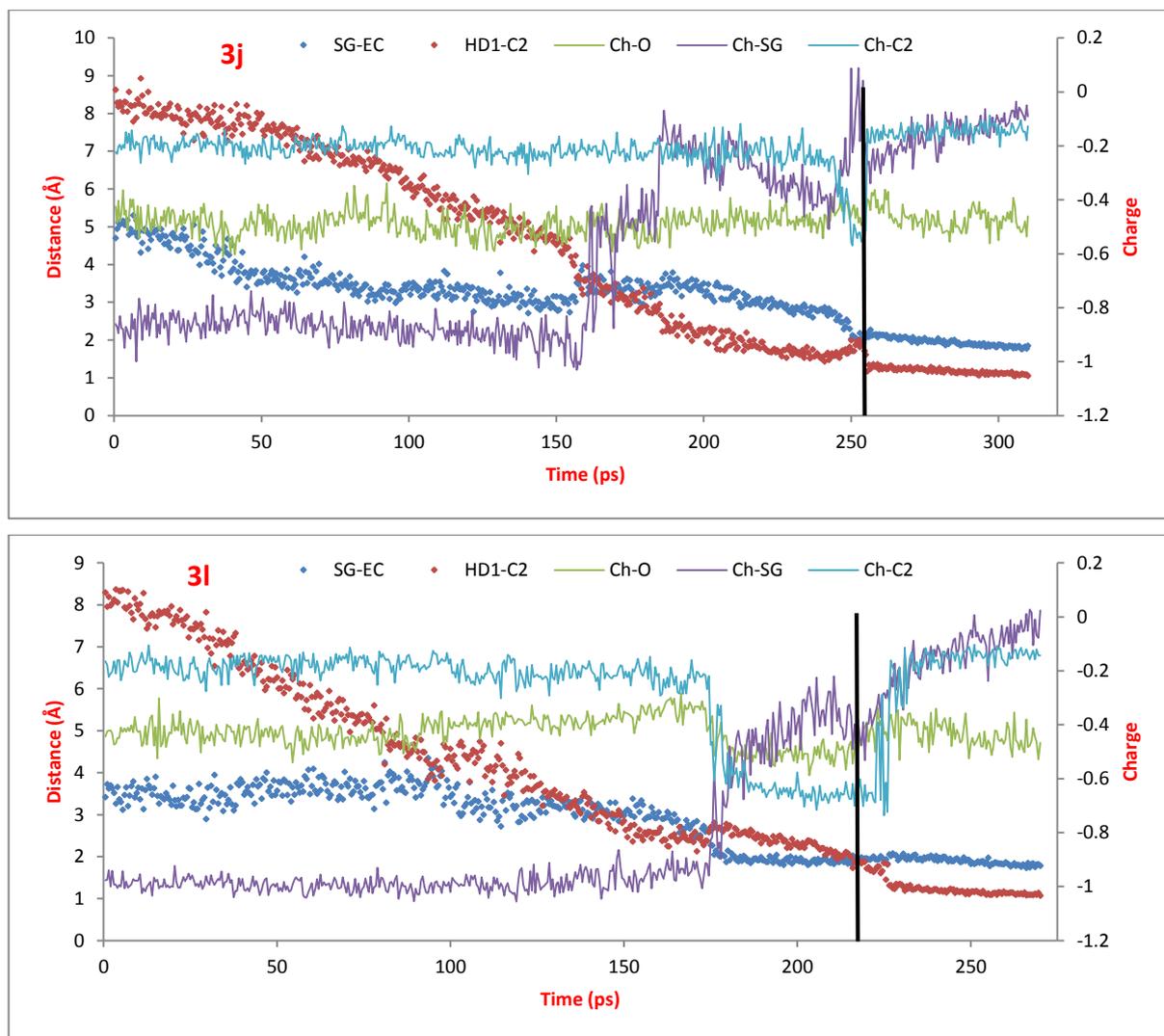


**Figure 4-8: left: IC<sub>50</sub> versus AE, right: log of IC<sub>50</sub> versus RE.**

The method was able to drive the reaction to completion and to produce AEs compatible, to some degree, with biological activity, with the appropriate trend; the AE was lower for the more active compounds. The charges of the atoms of the QM region were monitored during the simulations. Those of SG, C2 and the acrylamide carbonyl oxygen atom were recorded. The changes in these were not uniform across the compounds and could not be interpreted to explain a possible discrete mechanism for the reaction. In compound **3j**, for example, only the charge of SG changed considerably during the simulation while with compound **3l**, there was a decline in the charges of the acrylamide carbonyl oxygen and C2 around the TS (Figure 4-9).

The changes in the charges for compound **3j** suggest that the 2 processes (attack of the nucleophile and the protonation) happened at the same time with a few picoseconds during which SG-EC bond was there but the protonation has not happened yet. This explains the sharp decline in the charge of C2 at around 250 picoseconds, just before the barrier height. The graph of charges for compound **3l** suggests differently; the SG-EC bond formed first accompanied by a decline in the charges of C2 and the acrylamide carbonyl oxygen that continued until the protonation completed. The pattern observed with **3j** was observed in **1a** and **1e**, while that of **3l** was seen with **1c** and **3o**. By looking at distance graphs in Table 4-1 (the distances are also presented in Figure 4-9), it can be seen that the “**3j** charge pattern” was followed by the compounds in which the 2 bonds formed at the same time and the “**3l**

charge pattern” was adopted by the compounds in which the SG-EC bond was formed before the protonation (Figure 4-9).



**Figure 4-9:** Changes in the charges (lines, left axis, Ch-O, Ch-SG and Ch-C2) of compounds **3j** and **3l** along with change in distances for the bonds to be formed (dots, left axis SG-EC and HD1-C2) during the PM3, single-stage US simulations.

The charge pattern of **3j** implies a simple straightforward mechanism, in which all the events in the reaction happen almost simultaneously without a real chance for the participating atoms to have their charges altered. The change in the charges of C2 and the oxygen in compound **3l** is indicative of the formation of 1 of 2 possible intermediates, one in which the lack of the proton is manifested as a negative charge on C2 (Figure 4-10A) or one in which the broken double bond would migrate back to the carbonyl carbon with a shift of the carbonyl double bond to a negative charge on the oxygen (Figure 4-10B). A combination of the 2 intermediates is also a possibility.

It is because the non-uniformity in the charge graphs between the compounds and the consequent inability to decide a possible discrete mechanism for the inhibition that the 2-

stage simulation was tried. It should be noted, however, that a positive correlation between the reaction energies and biological activity (Figure 4-8) indicates that the method was able to predict the energies of the reactants and products correctly but was unable to predict the path followed accurately.

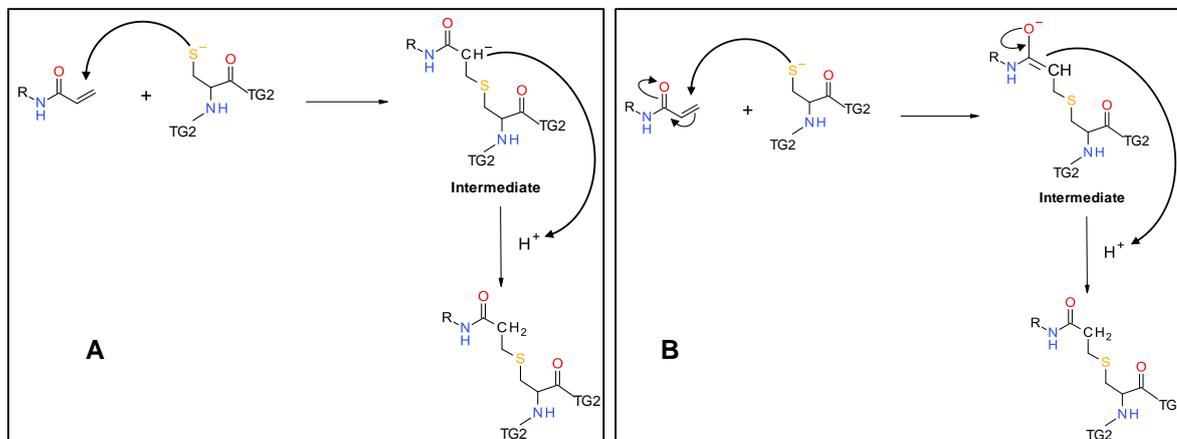


Figure 4-10: Possible intermediates formed during single-stage simulation of compound 31.

#### 4.1.2.2 Two-stage US simulations

The force constant of  $250 \text{ kcal/mol.Å}^2$  used during the 2-stage simulations was higher than that used previously. A lower value for the constant of  $50 \text{ kcal/mol.Å}^2$ , increased to  $100 \text{ kcal/mol.Å}^2$  towards the bond, was tried initially with PM3, but the simulation failed to drive the reaction to completion, especially in the protonation stage. Umbrella sampling has been used to describe similar reactions over 2 steps with such high force constant. Silva et al. (2015) investigated the inhibition of mycobacterial L,D-transpeptidase 2 by carbapenems, where the process involves a similar electrophile-nucleophile reaction between the negatively charged sulphur of an active site cysteine residue and the carbonyl carbon atom of a carbapenem. A proton is transferred in the second step from a nearby histidine residue to saturate a nitrogen atom that was originally attached to the carbonyl carbon of the carbapenem. They used a force constant of  $250 \text{ kcal/mol.Å}^2$  in the 2 steps (Silva et al. 2015).

PM3 was successful in forcing the 2 steps to completion with plausible structures and no change in the final pose of the acrylamide compounds. The 2 bonds eventually formed in all the compounds. However, no meaningful correlation could be made between the biological activity and any of the obtained energy values. SCC-DFTB, on the other hand, gave some correlation. Therefore, the results for SCC-DFTB will be presented with more detail while those from PM3 calculations will not be considered. The results from the 2 steps will be presented individually.

#### 4.1.2.2.1 The first step (formation of the SG-EC bond)

The RC for this stage was the distance between SG and EC (D1 in Figure 4-2). The overlap between the windows for this stage of the simulation was sufficient and Figure 4-11 is an example. At the end of this stage, the bond between SG and EC was present in all the cases, and the final conformation was used as a starting point for the 2<sup>nd</sup> step. Again, WHAM was used to calculate the PMF and the graphs produced are shown in Figure 4-12. All the PMF graphs produced had what appears like a lump or a saddle somewhere around 2 Å. This was selected to be the structure corresponding to the TS of the 1<sup>st</sup> stage (Figure 4-12).

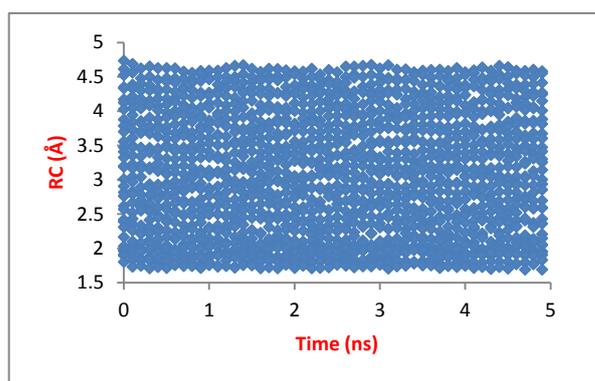


Figure 4-11: The overlap for the 1<sup>st</sup> step during the 2-step simulations for compound 1a as an indication of the degree of overlap achieved during this step.

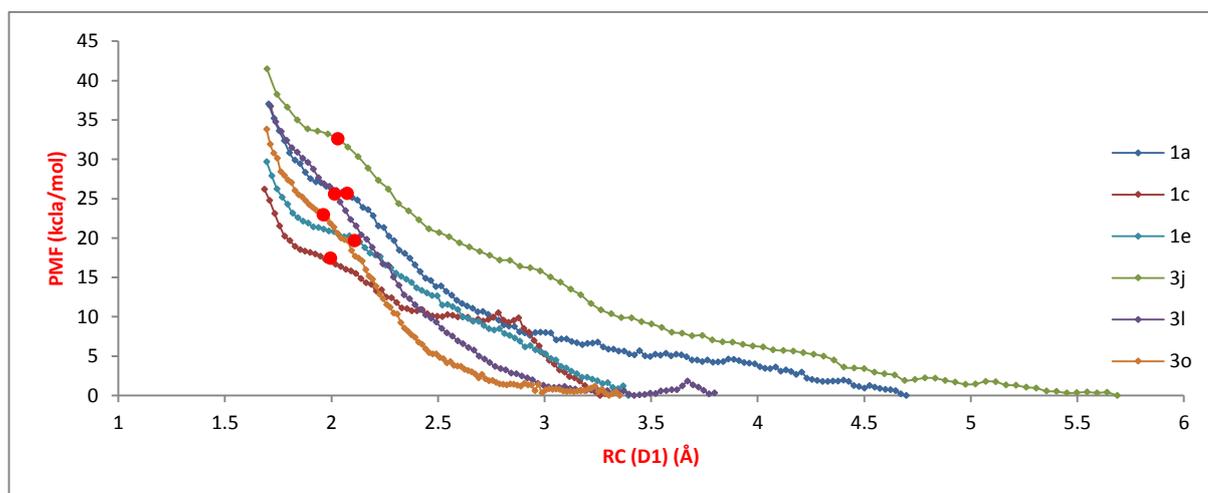
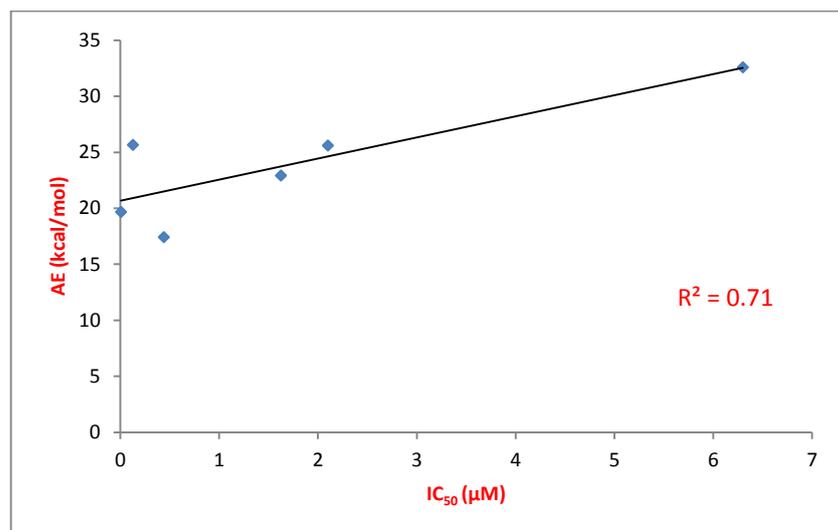


Figure 4-12: PMF graphs of the 6 acrylamide compounds during the 1<sup>st</sup> step of the 2-stage reaction. The red balls represent the TS position. The reaction proceeds from right to left.

When the charge of SG was monitored during the simulation, it was noted that the fully negative charge of SG became more positive, reaching a value of around -0.5 at the saddle point. The structures at the saddle point also had very similar values for the SG-EC distance and the SG-EC-C2 angle. The dihedral angle was also recorded at the saddle point of the 1<sup>st</sup> step. These transition state parameters are summarised in Table 4-4 along with the values of PMF at that point for each compound. The latter showed reasonable correlation with IC<sub>50</sub> producing a straight line with an R<sup>2</sup> of around 0.7 (Figure 4-13).

Compound	SG-EC (D1) (Å)	Angle ( $\varphi$ ) (°)	Dihedral (°)	Charge-SG	AE (kcal/mol)
<b>1a</b>	2.08	114	84	-0.49	25.67
<b>1c</b>	1.99	114	64	-0.40	17.45
<b>1e</b>	2.11	120	63	-0.53	19.68
<b>3j</b>	2.03	114	57	-0.55	32.61
<b>3l</b>	2.01	124	80	-0.59	25.61
<b>3o</b>	1.96	113	76	-0.48	22.95

**Table 4-4: Parameters for the structures at the saddle point of the PMF graph of the simulation of the 1<sup>st</sup> step of the 2-stage simulations.**



**Figure 4-13: Correlation between TG2 IC<sub>50</sub> and the energy at the saddle point of the PMF graphs of the 1<sup>st</sup> step of the 2-stage simulations for acrylamide compounds.**

The values of the angle of the attack of the nucleophile at the structures at the saddle point of the 1<sup>st</sup> step (Table 4-4) are also close to the value of Bürgi–Dunitz angle (107°) which describes the addition to carbon-oxygen double bonds, and can be applied to any reaction involving a Michael addition (Bürgi et al. 1974; Sinnott 2007). The values suggest that the approach of SG to EC followed the proposed route, where the nucleophile is expected to attack from above the plane of the electrophilic centre, rather than having the reacting species lining up in a straight line.

For the dihedral angle, again it was the simplest positive form (magnitude) that was recorded. The values are not actually right angles but they prove that the attack of the nucleophile was orthogonal to the plane of the acrylamide double bond rather than from the same plane of the bond. This further confirms that the reaction followed the appropriate route. It should be emphasised that during all the US simulations, saving was being performed every 0.5 ps and it is possible that the trajectory produced did not contain conformations corresponding to the cornerstone events during the simulation. In other words, the attack of the nucleophile may have happened at exactly the angle of Bürgi–Dunitz and at exactly right dihedral angle, but was missed during the saving; the saving point was not the same as the point of the attack. This may also explain the differences in bonds

lengths observed at TS structures between the compounds, both in the concerted single-stage and 2-step mechanisms.

#### 4.1.2.2.2 Second step (protonation of the initial Michael addition)

A generalised distance coordinate was used to represent the RC for this stage comprising the difference between D2 and D3 distances from Figure 4-2. Although the separation between the windows for this stage was 0.4 and 0.2 Å, the overlap achieved was reasonable (Figure 4-14). The graphs generated by WHAM for the PMF of this step showed some degree of lack of overlap in the sense that the windows were obvious in the graphs in the shape of curves (umbrellas) corresponding to each window (Figure 4-15). This is especially true for the first section of the 2<sup>nd</sup> step when the separation between the windows was 0.4 Å (at distances > 2 Å in Figure 4-15). The method was able to drive this stage to completion and the protonation of C2 was evident in all the compounds.

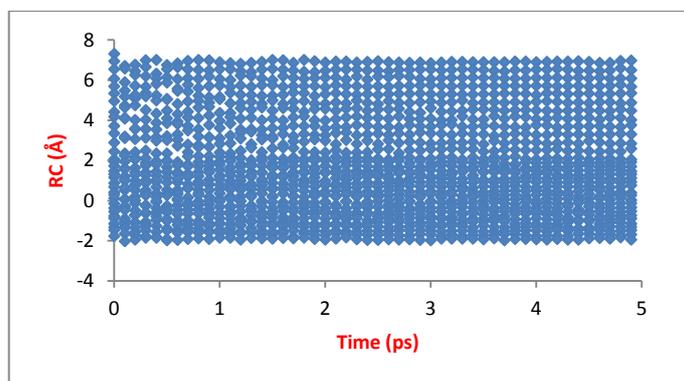


Figure 4-14: The overlap for the 2<sup>nd</sup> step during the 2-step simulations for compound 1e as an indication of the degree of overlap achieved during this step.

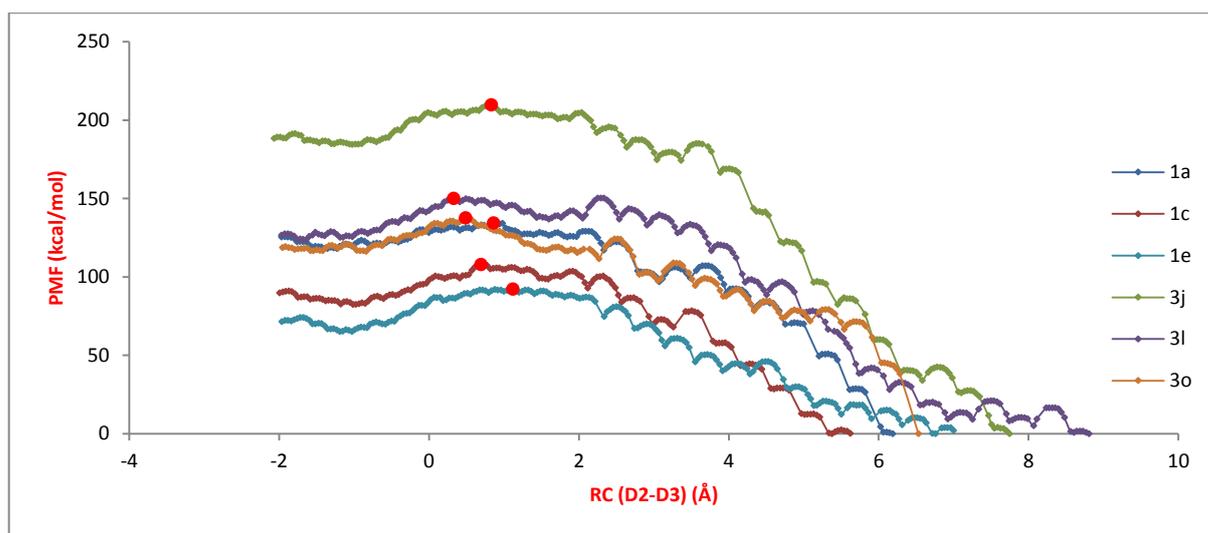


Figure 4-15: PMF graphs of the 6 compounds for the 2<sup>nd</sup> step of the 2-step US simulations using DFTB as the QM method. The red balls represent the TS. The reaction proceeds from right to left.

There was a common trend in the path followed by all the compounds for the 2<sup>nd</sup> step and this is manifest from Figure 4-15. The PMF curves start at the right side of the graphs accompanied by an increase in the PMF with decreasing values of RC until a maximum is

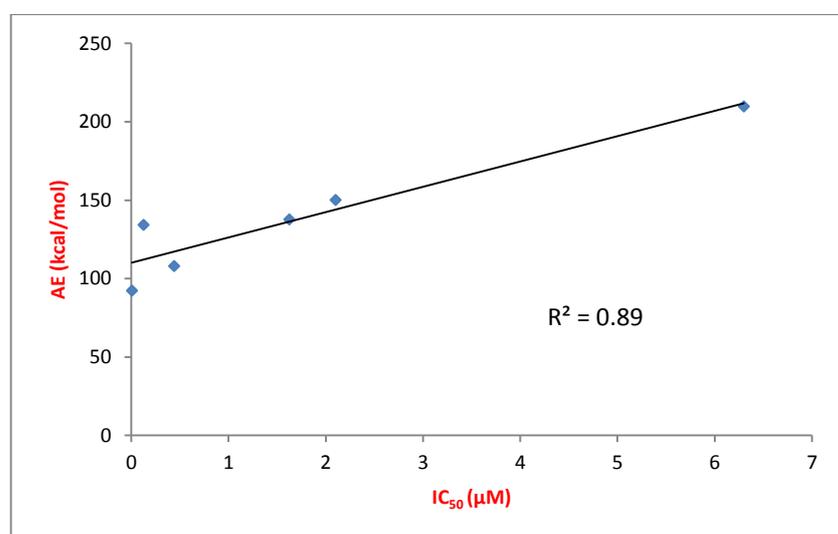
reached, followed by a decline in PMF score. The highest point on the graph was taken to be the transition state or the barrier height for the 2<sup>nd</sup> step. It is represented on Figure 4-15 by red balls.

TS structures did not occur at similar values of RC across the compounds. The RC assumed values ranging from 0.4-1.1 Å. In all the cases, the HD1-C2 bond was not formed and the HD1-ND1 bond (the bond connecting the proton to HIS335) was still present. The charge on the acrylamide carbonyl oxygen averaged around -0.7 which is more negative than its original charge signifying that HD1-C2 bond has not yet been formed and that the electron density shifting from the original double bond of the acrylamide was still on this oxygen atom. The charge of C2 was around -0.6 and this is much more negative than its starting value ( $\approx -0.2$ , details are presented later). This is another indication that HD1-C2 bond was not formed yet. The TS parameters for the 2<sup>nd</sup> step are presented in Table 4-5.

Compound	RC (Å)	HD1-C2 (D2) (Å)	HD1-ND1 (D3) (Å)	Charge-O	Charge-C2	AE (kcal/mol)
<b>1a</b>	0.85	1.94	1.10	-0.72	-0.60	134.27
<b>1c</b>	0.70	1.80	1.09	-0.57	-0.57	107.95
<b>1e</b>	1.13	2.19	1.06	-0.65	-0.61	92.24
<b>3j</b>	0.85	1.91	1.11	-0.70	-0.56	209.74
<b>3l</b>	0.34	1.48	1.09	-0.62	-0.56	150.15
<b>3o</b>	0.48	1.63	1.03	-0.67	-0.56	137.73

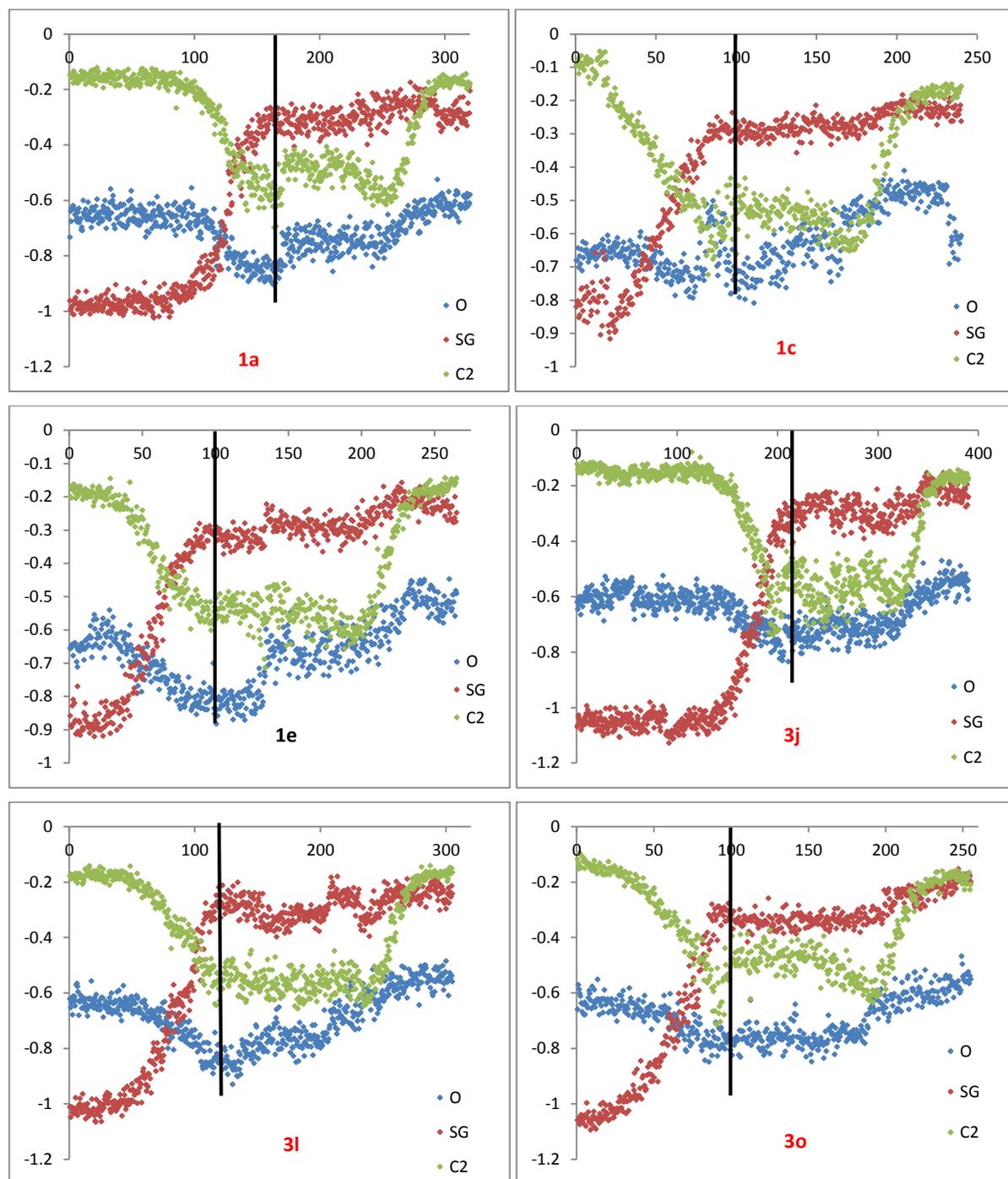
**Table 4-5: Parameters for the structures at the TS for the 2<sup>nd</sup> stage of US on acrylamides using DFTB as the QM method.**

The activation energy for the 2<sup>nd</sup> step correlated well with biological activity producing a straight line with an  $R^2$  of around 0.9 (Figure 4-16). The activation energies of the 2 steps during this 2-stage simulation of the reaction produced reasonable correlations with the biological activities for the tested compounds.



**Figure 4-16: Correlation between biological activity and AE for the 2<sup>nd</sup> step of simulations on acrylamides.**

When the charges of the reaction atoms were monitored during the 2-step simulations (both 1<sup>st</sup> and 2<sup>nd</sup> steps combined), the patterns of changes across the compounds were more comparable than what was seen in the single-stage simulations. There was a similar trend followed by all the compounds in terms of the charge behaviour around the reaction centre. For the individual compounds, the changes in charges across the individual simulations charges are shown in Figure 4-17.



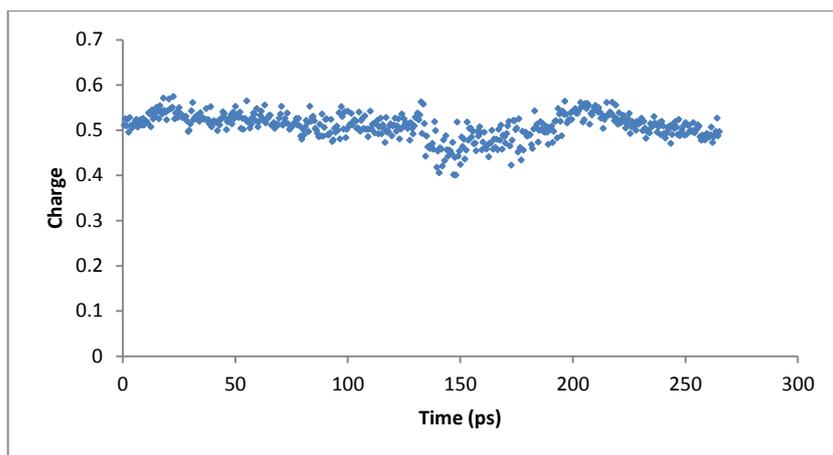
**Figure 4-17: Charges of the 6 compounds during the 2-step simulations. On the X-axis is the time in ps and on the Y-axis is the charge of the 3 monitored atoms (SG from CYS277, and C2 and carbonyl oxygen from the acrylamide warhead). The black vertical line marks the end of the 1<sup>st</sup> step.**

As it can be seen in Figure 4-17, the charges of C2 and O became more negative, while that of SG more positive as the 1<sup>st</sup> step progresses. At the end of the 1<sup>st</sup> step, C2 and O had their most negative charges. Again, this would suggest that any of the intermediates in Figure 4-10 may be the product of the 1<sup>st</sup> step. The oxyanion intermediate (Figure 4-10B), however, may be the most favourable because the charge of the oxygen is more negative than that of C2 at the end of the 1<sup>st</sup> step. In all the compounds, the acrylamide carbonyl oxygen atom had a charge of -0.8, which is closer to -1.0 than the charge of C2 in the compounds, the latter was ranging between -0.5-(-0.6) at the end of the 1<sup>st</sup> step.

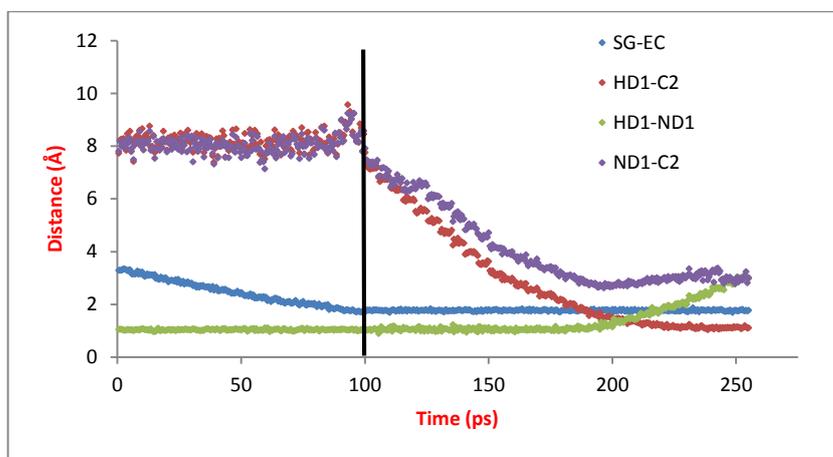
Silva et al. (2015), in their US study for the mechanism of inhibition of L,D-transpeptidase 2 (LDT) enzyme of mycobacteria by carbapenems, proposed a 2-stage mechanism. Their system and the acrylamide-inhibited TG2 are not exactly the same; LDT also has charged cysteine and histidine residues in the active site but the inhibitor has a beta lactam ring instead of the acrylamide for the electrophile. The intermediate at the end of the 1<sup>st</sup> stage, according to Silva et al. (2015), is an anionic stable compound bearing a negative charge on the N4 nitrogen atom of the beta lactam ring that was protonated during the 2<sup>nd</sup> stage. The PMF graph of the first stage showed that the formed intermediate has lower energy than the reactant or the TS of that stage, confirming it being an anion. In this study, the intermediate at the end of the 1<sup>st</sup> stage had higher energy than that of its TS (Figure 4-12). This, in addition to the charges, supports the formation of the oxyanion intermediate (Figure 4-10B).

The possibility of the 1,2-addition should not be overlooked. In this study, the 1,4 addition was proposed and investigated. This is because the double bond in the acrylamide is attached to an electron withdrawing group and this will induce the nucleophile to attach to the carbon away from the electron withdrawing group (Bernasconi 1989). In 1,2-addition, the nucleophilic addition to the acrylamide carbonyl carbon is also a possibility. The charge of this atom has been monitored and it did not show any change from its original value during the 2-stage simulation. It is true that the US was directed to drive the reaction to EC and not the carbonyl carbon, but the fact that the charge of the latter remained essentially constant during the simulations (Figure 4-18) may indicate the higher ease for directing the nucleophile to EC. Therefore, the 1,4 addition may be more likely with this system.

The changes in the distances followed very similar pattern across the compounds and were under the control of the reaction coordinates of the two stages at all times. Figure 4-19 is a representation for the distances changes during the simulations. The angle between SG-EC-C2 ( $\varphi$  in Figure 4-2) did not affect the 2<sup>nd</sup> step as the bond was already present, and therefore the angle maintained its value of that of a S-C-C bond.

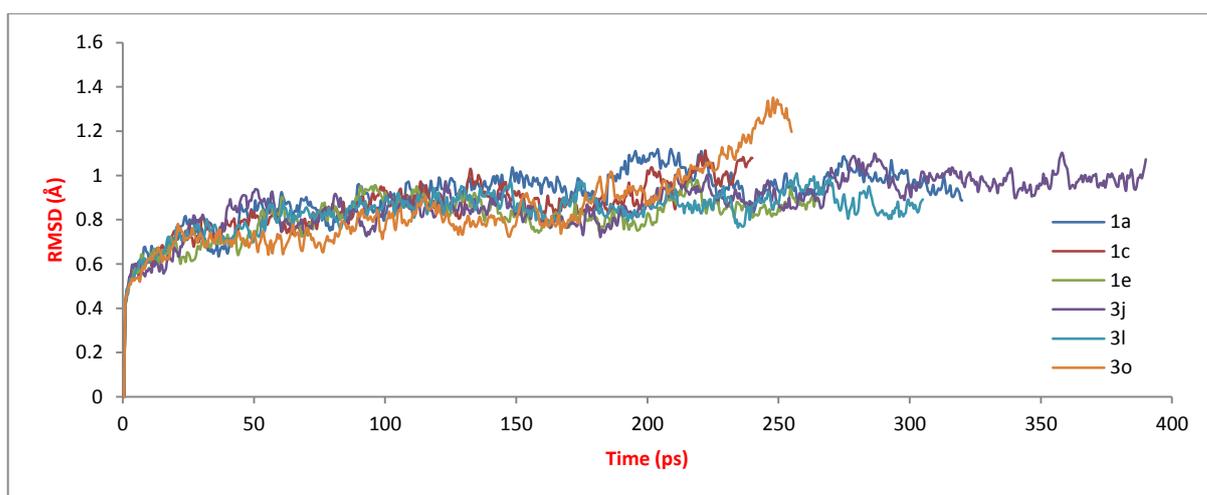


**Figure 4-18:** The change in the charge of the carbonyl carbon of the acrylamide warhead of compound 1e during the 2-step US experiment with DFTB.



**Figure 4-19:** Distances around the reaction centre in the 2-step simulations for compound 3o as a representative for the remaining compounds.

In the 2-stage simulations, the RMSD for the individual compounds showed a similar stable trend to what was seen during the 1-stage simulations with a maximum value of 1.35 Å (Figure 4-20). Both stages achieved good correlations with biological activity.



**Figure 4-20:** RMSD for the simulations of the 6 acrylamide compounds with the 2-step US.

The AE values for the 2<sup>nd</sup> stage (the protonation step) were generally higher than those from the 1<sup>st</sup> stage (the formation of the SG-EC bond) suggesting that the protonation is the rate

limiting step for this reaction. This may not be consistent with the literature kinetics for thiol Michael addition reaction, which states that it is the attack of the nucleophile that constitutes the rate limiting step (Nair et al. 2014). An explanation could be offered based on the proton source used for the 2<sup>nd</sup> stage in this work. It was more difficult for the US simulation to take the proton from HIS335 than it was for the method to form the bond in the 1<sup>st</sup> step. It is the relative rigidity of the location of HIS335 being part of the protein, compared to the more freely moving acrylamide approaching SG in the 1<sup>st</sup> step, that was possibly responsible for the higher AE values observed in the 2<sup>nd</sup> step. An alternative proton source may have been better in this regard, but the choice of HIS335 here is justified (section 4.1.2.4).

### 4.1.2.3 Water behaviour

The behaviour of water molecules around the reaction centre was investigated to examine the effect of desolvation on the energies of the different compounds during US simulations. This was performed using CPPTRAJ by measuring the number of water molecules in the 1<sup>st</sup> solvation shell around SG (3.4 Å from SG) and by calculating the radial distribution function (RDF) (section 3.7.2) of water molecules up to 5 Å around SG to inspect the change in density of water molecules as a function of distance.

#### 4.1.2.3.1 Single-stage simulations

For the single-stage simulations, RDF graphs for the individual compounds are presented in Figure 4-21. The probability of finding water molecules in the vicinity of SG followed a similar pattern to that observed during the analysis of empty TG2 conventional MD trajectories as well as the trajectories for 275-ns MD applied on the complexes (section 3.9). This probability was highest at 2 Å from SG followed by another peak at 3 Å. No water molecules existed in any simulation at a distance of less than 2 Å from SG.

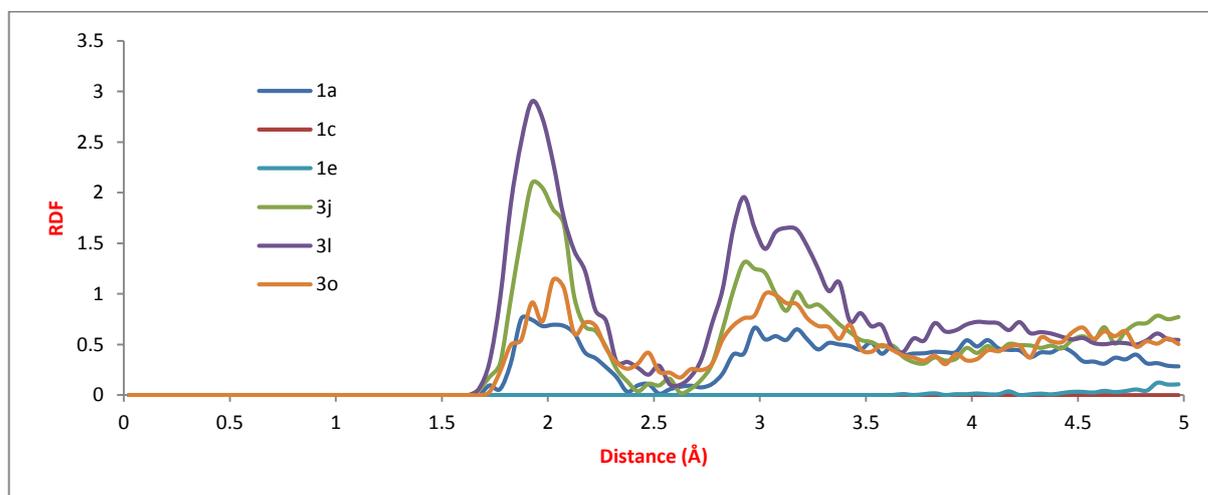


Figure 4-21: RDF graphs for the 6 compounds during the single-stage simulations.

In the simulations for compounds **1c** and **1e**, there were no water molecules at all within 5 Å of SG. This has happened since the start for these 2 compounds during the generation of files for MD where AMBER LEaP program did not place any water molecules within 5 Å of SG during the solvation of the complex for the 2 compounds. The absence of water in these 2 cases did not change with the application of initial minimisation, heating the system, equilibration or even QM/MM relaxation applied prior to the US simulations.

The numbers of water molecules found in the 1<sup>st</sup> solvation shell for the individual compounds' simulations are presented in the graphs in Figure 4-22. It should be noted here that the graphs in Figure 4-22 are continuous, so there cannot be 2 and 3 water molecules at the same time point, and they appear like that because there are too many time points in each graph. The final graph in Figure 4-22 represents the time from the 240<sup>th</sup> to the 260<sup>th</sup> picoseconds of the simulation of compound **3i** and was presented to clarify the continuity issue.

The lack of water molecules in the 1<sup>st</sup> solvation shell is very obvious for compounds **1c** and **1e** (**1e** not presented). For compounds **1a** and **3j**, it can be clearly seen that there were some water molecules (up to 2 in compound **1a** and up to 4 in compound **3j**) in the initial stages of the simulations and that these water molecules left the solvation shell after the barrier height and the formation of the SG-EC bond. In compound **3i**, there were between 2 and 4 water molecules in the solvation shell, up to 3 of them stayed within the shell even after the formation of the SG-EC bond. A visual inspection of the trajectory of **3i** showed that these molecules were beneath the reaction centre, deep in the catalytic tunnel and they were physically trapped there for the entire simulation (Figure 4-23). They did not affect the simulation in terms of desolvation effect as can be seen from the free energy of the reaction that fitted well with respect to biological activity (Figure 4-8).

In compound **3o**, solvation shell shows that there was a water molecule close to SG after the formation of the bond with EC. It has been confirmed that it was a single molecule and not multiple molecules moving in and out of the shell, by measuring the distance between SG and the oxygen atom of the water molecule in question over the entire simulation (Figure 4-24). This molecule was tightly bound even after the SG-EC bond formation which occurred relatively early in this simulation. A closer look at the trajectory showed the involvement of this water molecule in hydrogen bonds with CYS277 SG in addition to hydrogen bonds with the bridging tryptophan residues, TRP241 and TRP332 and backbone atoms of CYS277. Figure 4-25 represents the last frame in the trajectory and shows hydrogen bonds binding the water molecule to the active site.

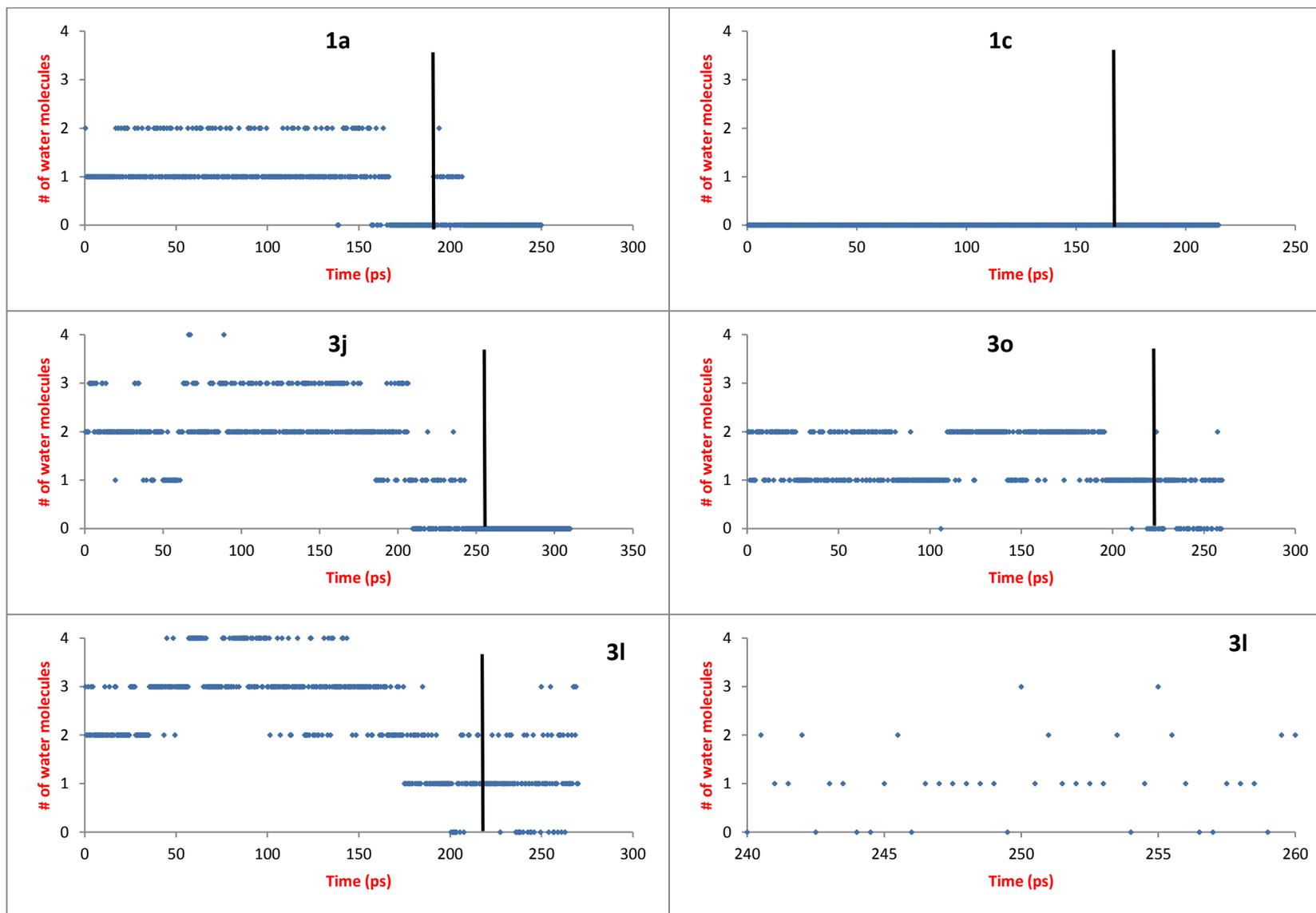


Figure 4-22: Graphs of the number of water molecules in the 1<sup>st</sup> solvation shell, as a function of simulation time all for single-stage simulations. Graph for 1e was identical to that of 1c and was not presented. The final graph represents 20 ps only from the original graph of 3l. The black line represents the barrier height.

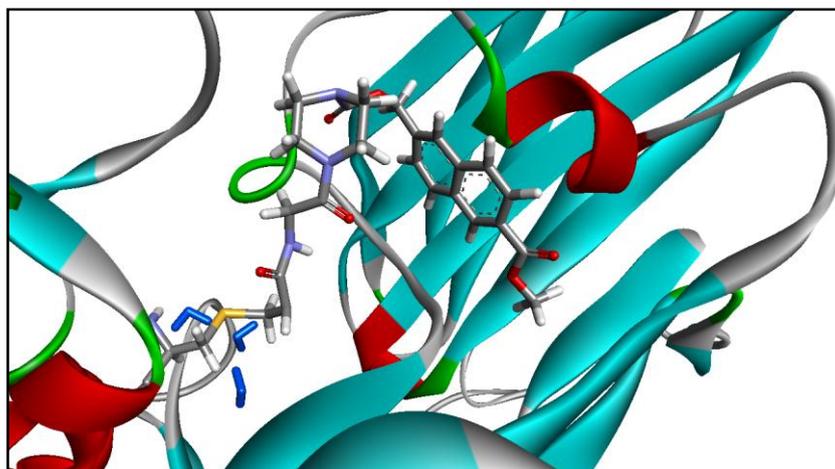


Figure 4-23: The 3 water molecules (coloured blue) that were trapped in the 1<sup>st</sup> solvation shell during the single-stage simulation of compound 3l, in a frame after the SG-EC bond has been formed.

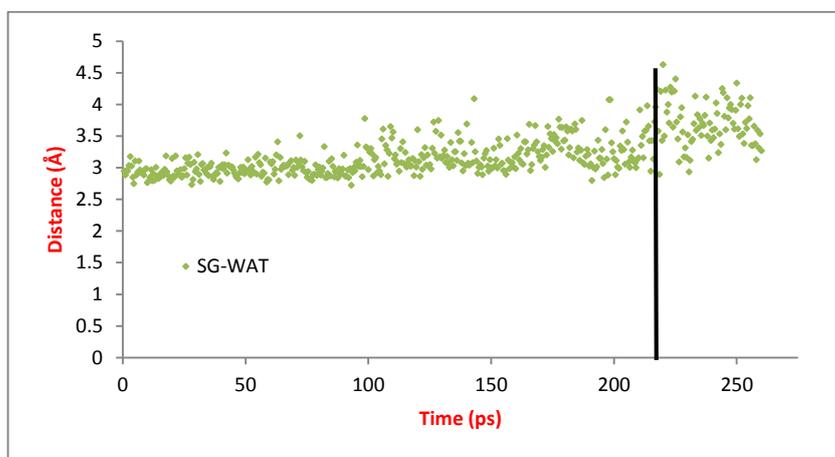


Figure 4-24: Distance between SG of CYS277 and the oxygen atom of the problematic water molecule during the single-stage simulation for compound 3o.

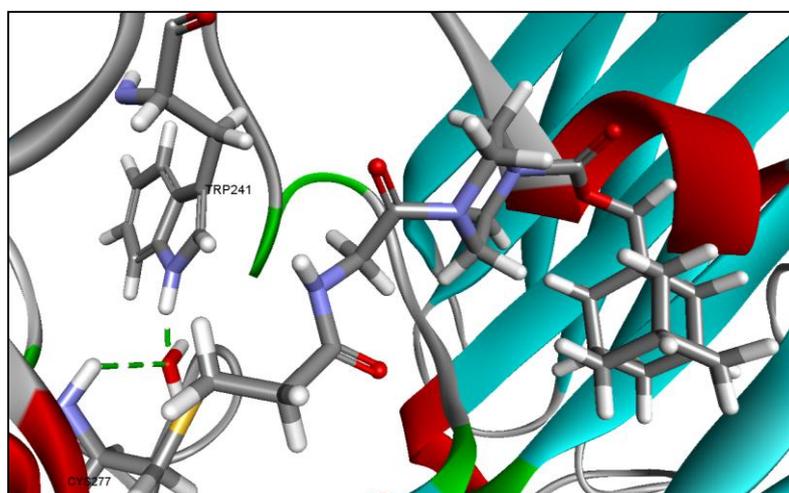


Figure 4-25: Hydrogen bonds with water molecule in compound 3o single-stage simulation.

The RDF of compound **3o** shows a similar probability at 2 and 3 Å because the problematic water molecule stayed in its relative starting position for the entire simulation and was moving in the range of 2-3 Å. The failure of compound **3o** active site to desolvate completely prior to the reaction may be the reason for the very high PMF reading for the compound during the single-stage simulations. The probably false high PMF value for **3o** could have resulted from the difficulty in driving the process to completion. Since SG-EC bond was formed early (Table 4-2), it can be assumed that the formation of the HD1-C2 bond was the

difficult part. The tightly bound molecule may have contributed to an increased rigidity of the **3o**-CYS277 adduct with the subsequent difficulty in the delivery of the proton to C2. If compound **3o** was not included in the correlation made between the activation energies obtained from the single-stage simulations and biological activity (Figure 4-8), then the  $R^2$  value for that correlation would jump to around 0.9 (Figure 4-26).

High RDF values seen at 2 Å in Figure 4-21, which indicate higher probability for finding water molecules at this distance, may be explained by the fact that the bond between SG and EC was not formed until the later stages of the simulations. Therefore, as a function of the entire simulation, there is higher probability to finding a water molecule within 2 Å, but if only the later sections of the simulations were considered in RDF analysis, then no such high probabilities would have been produced. This has been confirmed once for compound **3j** where the RDF was calculated for the final section of the simulation, only considering the frames occurring immediately before the barrier height and those after them to the end of the simulation (Figure 4-27). The result showed no water molecules within 4 Å, to be followed by an RDF approaching unity and the normal water density. This behaviour of water following the formation of the bond can be compared to the RDF measured during covalent MD in section 3.7.2 (Figure 3-71).

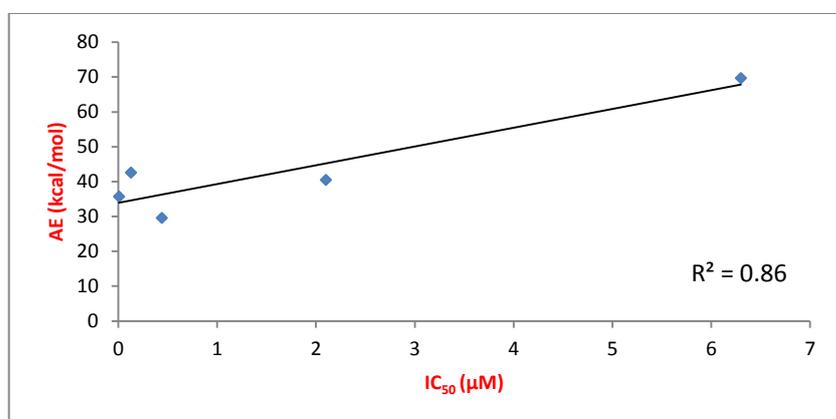


Figure 4-26: Correlation between biological activity as  $IC_{50}$  and AE for the acrylamide compounds, excluding compound **3o**, during the single-stage PM3 US simulations.

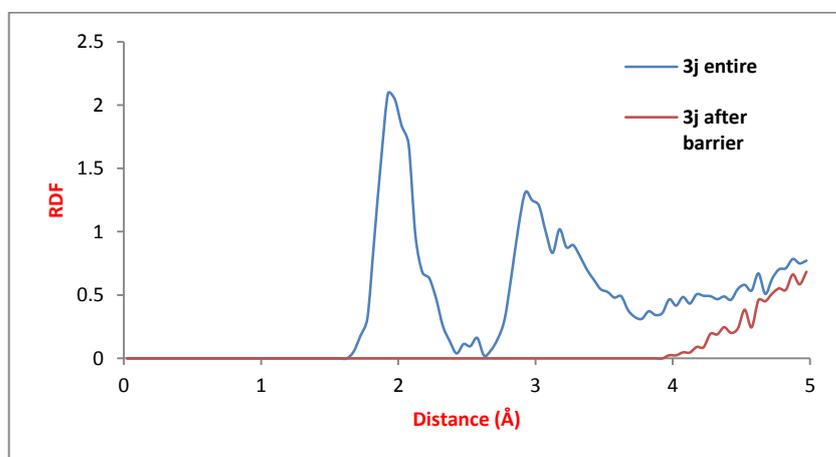


Figure 4-27: RDF graphs for the simulation of compound **3j**, showing the probability calculated for the entire simulation and that calculated for the final part of the simulation after the barrier.

#### 4.1.2.3.2 Two-step simulations

For the 2-step simulations, the solvation shell graphs for the compounds are presented in Figure 4-28. Again, there were no water molecules within 3.4 Å of SG in compound **1c** entire simulation (not shown). For the other compounds, the desolvation effect was not very clear. In compounds **3i** and **3o**, the simulation started with more than one water molecule in the solvation shell but the number decreased to only 1 molecule towards the end of the 1<sup>st</sup> step, showing some desolvation. This was not seen in compounds **1a** and **3j**, in which the number of molecules has remained the same or even increased after the formation of the bond. A possible explanation could be that at the start, SG was negatively charged and was attracting the water molecules and when it was neutralised through the formation of the bond, some water molecules left; at the same time, the charge of both C2 and the carbonyl oxygen on the compounds was becoming more negative before the protonation, and this could supply another source for the attraction of water molecules. By the end of the 2<sup>nd</sup> step, the protonation is established and the number of water molecules declined.

The simulation for compound **1e** started with no water molecules in the solvation shell but after the formation of the bond, at around the 150<sup>th</sup> picoseconds, a water molecule appeared in the shell and stayed there for about 75 ps. Visual inspection and hydrogen bond analysis with VMD showed that it was a single water molecule arriving in the vicinity of SG and that this molecule was involved in a hydrogen bond with the carbonyl oxygen of compound **1e** for 43% of simulation time, which makes it relatively close to CYS277. When the water molecule appeared in the solvation shell it was because the carbonyl oxygen of the compound moved closer to CYS277 during the 2<sup>nd</sup> step of the umbrella sampling and dragged the water molecule along. The last graph in Figure 4-28 is for the frequency of the hydrogen bond between the water molecule and the carbonyl oxygen of compound **1e**.

In summary, water molecules did provide some desolvation effect during the single-stage simulations, and to some extent in the 2-step simulations. The better desolvation effect observed during the single-stage simulation could be attributed to the lack of a negative charge in the reaction centre with the formation of the 2 bonds at relatively the same time, and therefore, the lack of attraction for water molecules with the subsequent desolvation. Capoferri et al. (2015) studied a similar system, irreversible epidermal growth factor receptor (EGFR) inhibition with acrylamide-based inhibitors with an approach combining steered MD with US at the DFTB level of theory. The inhibitors also act by alkylating an active site cysteine residue (CYS797). The authors proposed a single stage concerted mechanism for the inhibition and identified desolvation of water molecules as a key event in the process.

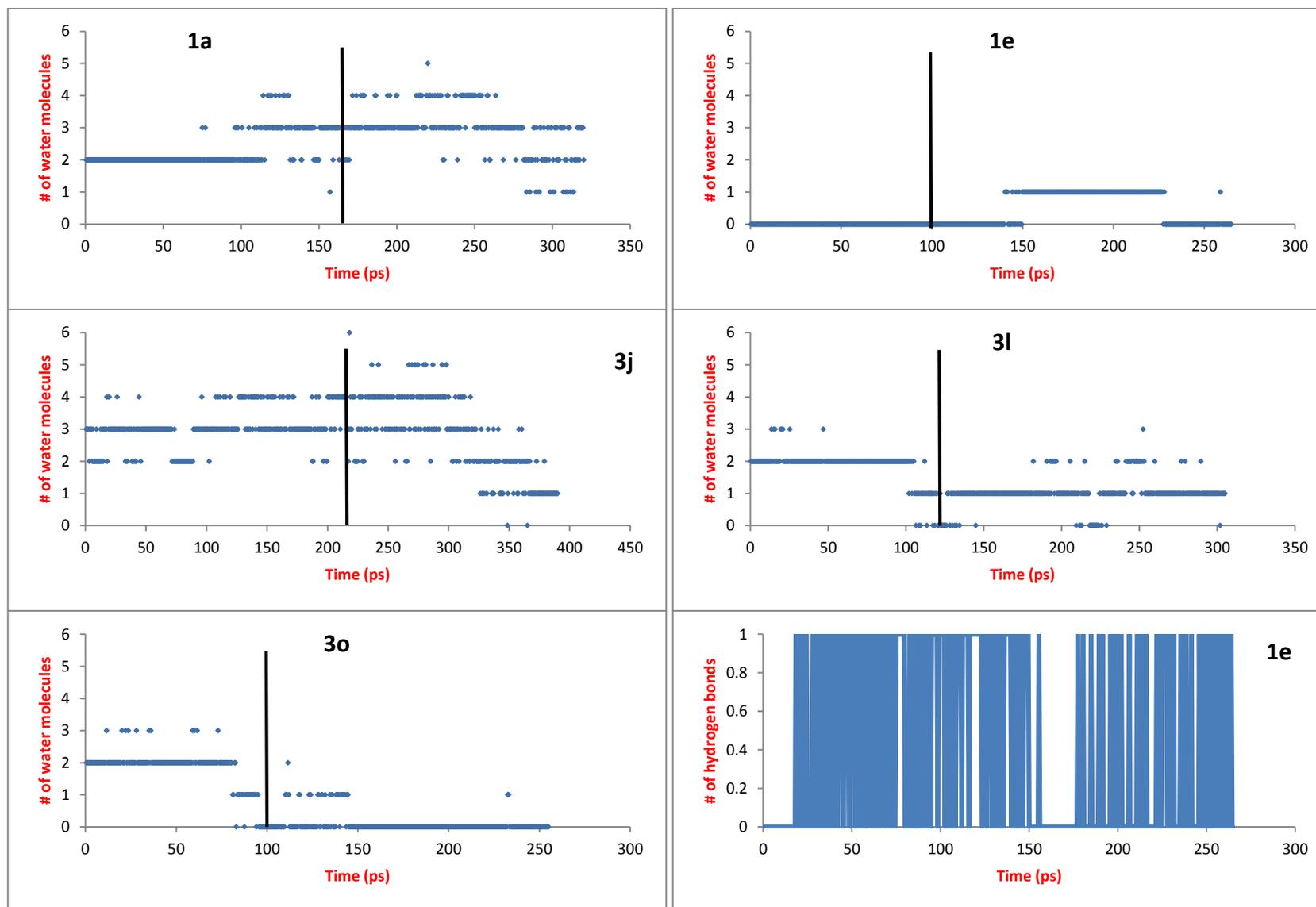


Figure 4-28: Graphs of the number of water molecules in the 1<sup>st</sup> solvation shell for 2-step simulations, 3.4 Å around SG of CYS277 as a function of simulation time all. The black line marks the end of the 1<sup>st</sup> step. The last graph is for hydrogen bonds in compound 5 simulation.

Desolvation was relatively obvious in the single-stage simulations but those simulations could not explain the charges of the reaction centre, which was the main reason for applying the 2-stage simulations. In the latter, some tightly bound water molecules remained close to CYS277 during the entire simulations, and it was shown that such molecules were attracted by negative charges either of SG or those created on the system after the formation of the SG-EC bond. Hydrogen bonds with some active site residues also had their role for keeping some water molecules tightly bound within the active site. The charges of the atoms at the reaction centre, however, supported the hypothesis of 2 consecutive steps for the reaction.

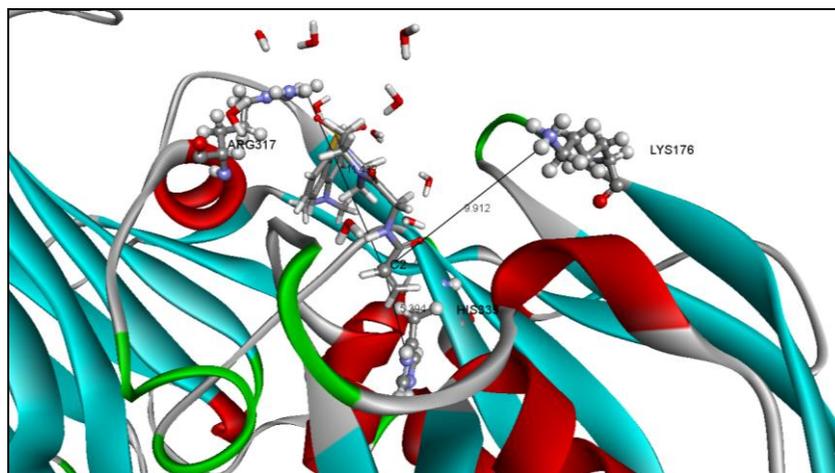
#### 4.1.2.4 Choice of proton source

It has been stated in the literature that there is a catalytic triad in TG2 consisting of CYS277, HIS335 and ASP358 (Pinkas et al. 2007; Iismaa et al. 2003). CYS277 is involved in the transamidation and deamidation reactions through the formation of a thio-ester bond with the amide group from a glutamine residue (Siegel & Khosla 2007). The role of the other two residues remains to be clarified. One possible contribution of HIS335 was suggested by Iismaa et al. (2003), which involves the interaction with CYS277 to form a thiolate-imidazolium ion pair and both residues become charged. Protonated histidine is a well-known proton donor in proteins (Liao et al. 2013). This and the fact that HIS335 is a constituent of TG2 catalytic triad were the main reasons behind using this residue as the proton donor to saturate the  $\alpha$  carbon of the double bond (C2) on the acrylamide compounds.

Lysine and arginine residues can act as proton donors, but in TG2 there are no lysine or arginine residues in the vicinity of the catalytic cysteine residue, where the protonation would occur. The proton from such a residue is at least 9 Å apart from C2 of the acrylamide and is positioned inappropriately, where supplying this proton would require folding the acrylamide compound and most probably would not allow for the formation of the SG-EC bond (Figure 4-29). That would make HIS335 the best option available. The latter was not always very close to C2 but is positioned perfectly with respect to CYS277 to allow for the protonation without distorting the formed SG-EC bond.

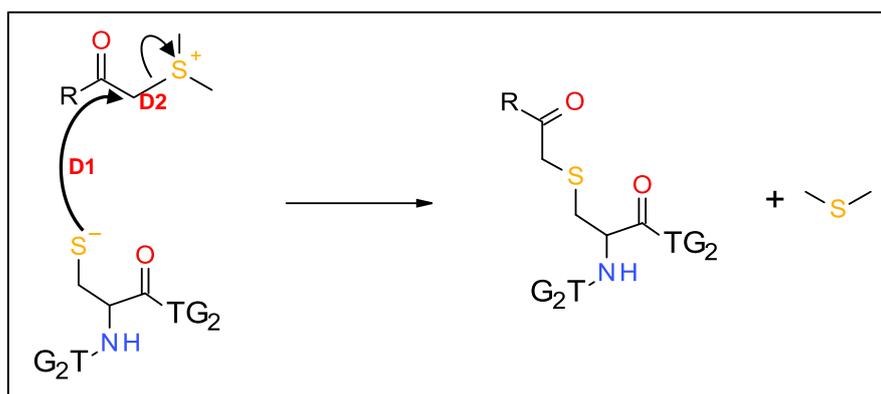
There were water molecules within the active site that may be closer than HIS335 to C2 of the acrylamide (Figure 4-29), as seen during the analysis of water behaviour. However, it was also confirmed that they were not uniformly distributed around the reaction centre between the compounds, and they were entirely missing in some cases (compounds **1c** and **1e**). Therefore, a single and similar reaction coordinate that involves a water molecule is not easy to define, but remains a possibility, either as explicit source for the proton or as a

proton shuttle between one of the distant protonated residues and C2 of the acrylamide inhibitors.



**Figure 4-29:** The closest arginine (ARG317) and lysine (LYS176) residues to C2 of the acrylamide within TG2 active site for compound 1e. The figure labels C2 of the compound along with the distances between the potential protons and C2.

For the inhibitors bearing sulfonium ion, a single-stage US mechanism was tried using a generalised distance coordinate as the difference between the bond to be formed (SG-EC), and the bond to be broken (EC-S+, the positively charged sulphur of the inhibitor) (D1-D2 in Figure 4-30), with PM3 and DFTB as the QM methods. The simulations managed to produce the appropriate products in which the bonds formed and broke correctly, but the associated PMF values gave an inverse trend when correlated to IC<sub>50</sub> values; higher PMF values for the more active compounds. For that, the approach was not followed any further and the resultant data are not presented in this work. Instead, the reaction mechanism of sulfonium ion inhibitors was investigated with CAChe using the intrinsic reaction coordinate (IRC).



**Figure 4-30:** The single-stage mechanism for the inhibition of TG2 with sulfonium ion inhibitors that was investigated with US simulation, showing the distances involved in the selected RC.

## 4.2 CAChe Reaction Path Experiments

### 4.2.1 Introduction

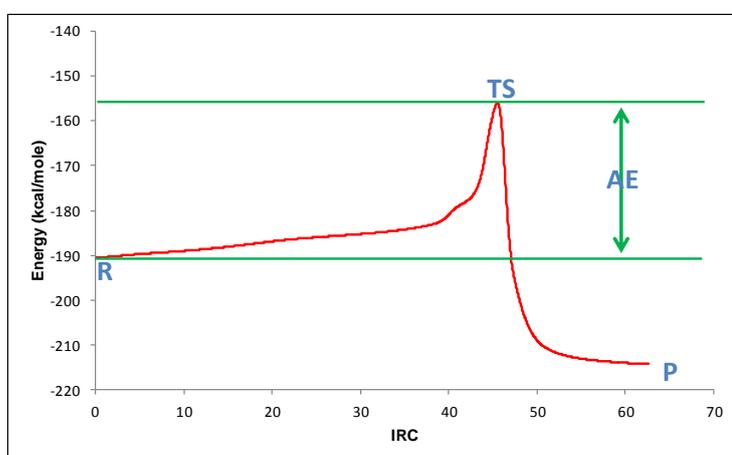
#### 4.2.1.1 CAChe reaction path

The MOPAC 2000 (Molecular Orbital PACkage) (Stewart 2002) within CAChe (Fujitsu Limited 2006) was used to study reaction paths. Studies were performed using quantum mechanical methods by having a transition state structure (TS). A “map reaction” experiment could be performed to identify the TS. This type of experiment within CAChe uses a semi-empirical method (such as AM1, PM3 or PM5) to minimise the system over a predefined reaction coordinate (RC) (a distance or an angle). The result is a potential energy surface (PES) for the system as a function of the change in RC. The minimum on the PES would correspond to a structure of the product and the nearest saddle point (hill) would represent the TS structure.

Once the TS is found, the structure needs to be refined by a conjugate gradient energy minimisation. This is a minimisation algorithm that operates to ensure moving down the hill on the PES all the time. It does so by using information obtained from the previous minimisation step, so that reversed progress is avoided (Höltje & Folkers 2008). The minimisation is followed by a verification process, which is performed by calculating the vibrational frequencies of the molecule through a FORCE calculation (a method that calculates the force matrix of the system by calculating the 2<sup>nd</sup> derivative of energy as a function of atomic displacements, and uses the force matrix for predicting the vibrational frequencies). A true TS structure would have only one negative vibration corresponding to only one imaginary mode of vibration. This is the vibration along the reaction coordinate (CAChe-Group 2006; Stewart 2002).

After the TS structure is verified, a “reaction path” experiment is applied. This is also a QM based method that explores the path from a TS structure to either reactants or products. The method utilises an intrinsic reaction coordinate (IRC) to calculate the path of the reaction (CAChe-Group 2006; Stewart 2002). The IRC is a minimum energy path using the steepest descent energy minimisation. The characteristic of this path is that it starts from a TS structure and moves towards either the reactants or the products, using mass-weighted Cartesian coordinates (Maeda et al. 2015). IRC was first proposed for the calculation of reaction paths by Fukui (1970).

The IRC reaction path method calculates the potential energy of the system along the IRC and annihilates the kinetic energy. The result is a collection of structures from reactants to products along the IRC with the potential energies of each. Typically, this would give a graph in which the energy starts at certain value (reactants) and increases along the IRC until reaching a maximum (TS) after which the energy drops and ends with a value that is lower than that of the reactant (Figure 4-31) (CACHe-Group 2006; Stewart 2002). This final value would correspond to the product. The difference between the energies of the reactant and the TS is the activation energy for that reaction; which, for our system, was hoped to correlate to the biological activity of the individual compounds. IRC is an established method for the calculation of the activation energy of chemical reactions (Ikuo & Ogawa 2014; Yoshimura et al. 2012; Baowei et al. 2007; Yang et al. 2003; Takai et al. 1998).



**Figure 4-31: A typical reaction path between reactants (R) and products (P) showing the activation energy.**

In CACHe, the method is inapplicable to large systems, and there is no functionality for splitting the system into QM and MM regions. As such, these experiments were applied to smaller systems including the ligand and CYS277. The compound set involved the active inhibitors bearing the sulfonium ion warhead. The compromise of not including the entire system was made because US simulations failed to produce informative correlations when applied to the sulfonium ion inhibitors.

#### 4.2.1.2 Starting structures

Table 4-6 shows the sources of the starting structures for each of the compounds. The table also shows whether the structure has been minimised or not. If minimisation was performed, then it would have involved only the compound and CYS277 while everything else in the system was constrained with CACHe locking function which freezes the coordinates of the locked atoms. The procedure used for the minimisation is MM2 (Allinger 1977) and was performed within CACHe. For the compounds whose starting structures were not minimised, this was because the minimised structures did not give valid map files from which to extract

the TS structure. It should be noted, however, that minimisation and MD must have been applied at one stage in acquiring these starting structures (Chapter 3).

Compound	Source of starting structure	Minimisation	IC <sub>50</sub> (μM)
<b>1b</b>	Frame at 98.895 ns of the 275 ns MD run	Yes	0.277
<b>1d</b>	Docking complex of <b>1d</b> in 245ns model	Yes	0.7
<b>1f</b>	Frame at 4.225 ns of the 5-ns MD on <b>1f</b> in 95ns model	No	0.38
<b>3b</b>	Docking complex of <b>3b</b> in 95ns model	Yes	0.89
<b>3c</b>	Frame at 1.775 ns of the 5 ns MD run on <b>3c</b> in 90ns model	Yes	1.4
<b>3h</b>	Docking complex of <b>3h</b> in 90ns model	No	1.07
<b>3m</b>	Docking complex of <b>3m</b> in 155ns model	Yes	1.5
<b>3n</b>	Frame at 4.925 ns of the 5-ns MD on <b>3n</b> in 100ns model	Yes	0.775

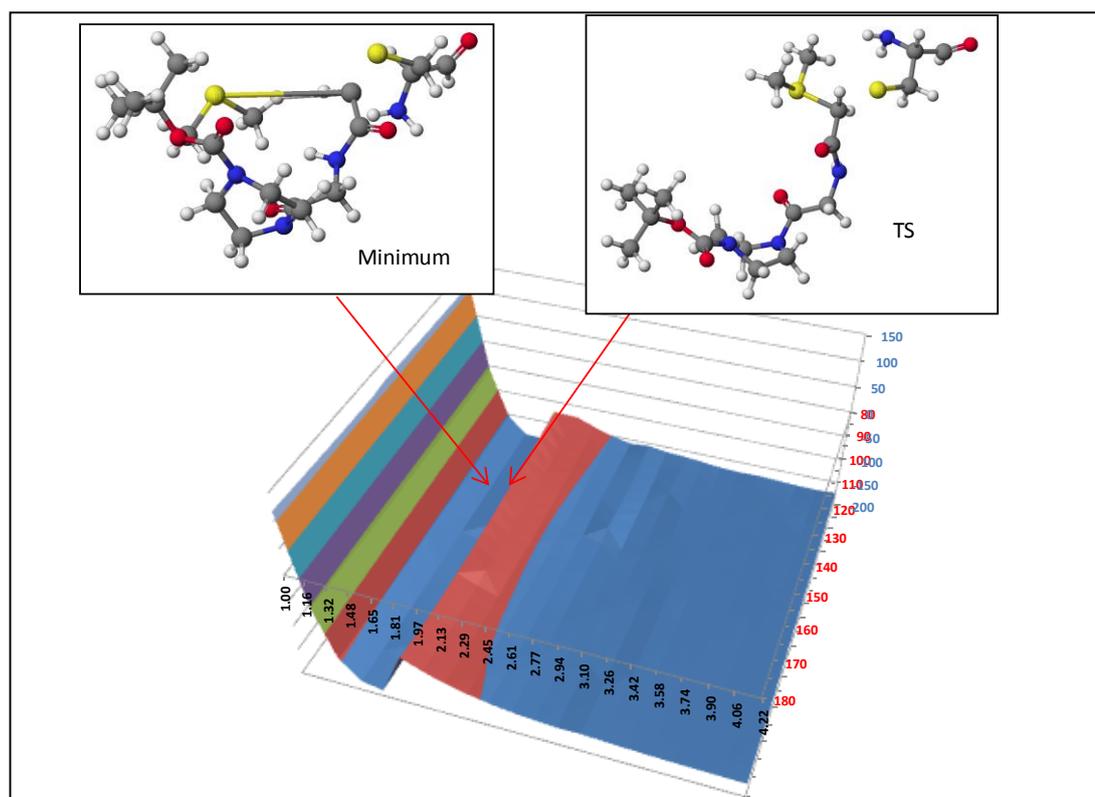
**Table 4-6: Source of the starting structures of sulfonium ion compounds. Their structures have been presented in Table 3-1 (p86) and Table 3-12 (p125).**

After the minimisation step (or the non-minimisation), all TG2 residues were deleted with the exception of CYS277, leaving just the reaction centre consisting of the inhibitor with the sulfonium ion and the nucleophilic active site cysteine residue. Therefore, the starting structures used for the reaction path experiments by CAChe had the effect of the presence of TG2 in the sense that the initial conformation of the compounds was a one produced within the enzyme. A conformation taken from a MD trajectory was preferred but if this did not produce a reasonable PES from which to extract a TS, then a docking complex has been used. The starting structures were chosen as having the shortest distance between EC of the ligand and SG of CYS277.

## 4.2.2 Results and Discussion

The results of the experiments on compound **1b** will be presented in detail while the results for the other compounds will be summarised. As mentioned in the methods, two reaction coordinates were used to generate the potential energy surface for the system; the distance between SG and EC, and the angle between SG, EC and the positively charged sulphur of the ligands (S+) (section 2.2.2.2.1). The distance of the bond to be formed (SG-EC) has been tried as a reaction coordinate on its own but failed to produce informative results, and the same applies for the two bonds to be formed and broken (SG-EC and EC-S+) (details are presented later in this chapter). For compound **1b**, the distance and angle values in the starting structure after minimisation were 4.225 Å and 88.1° respectively. The generated map file had 441 structures. Figure 4-32 shows a graph for the energies of these structures as functions of the reaction coordinates (PES).

The structure of the minimum (the lowest energy conformation on the PES) had values of 1.645 Å and 125° for the reaction coordinates (the Minimum in Figure 4-32) and an energy value of -195.133 kcal/mole. The chosen TS structure had the same value for the angle as the minimum but the distance was 1.80 Å and the energy -113 kcal/mole. The TS structure should ideally be the maximum on the potential energy surface during the movement from the reactants to the products. However, since the distance of the bond that is supposed to be formed was allowed to change from its original value to 1 Å, the highest energy conformation was not realistic as the distance between the two atoms forming the bond was very short.



**Figure 4-32: PES produced from map experiment on 1b with the minimum and TS. Black axis is distance (Å), red is angle (degrees) and blue is energy (kcal/mole).**

The chosen TS structure was refined with PM3. The refinement process produced a TS structure with values of 2.36 Å and 150° for the distance and angle coordinates respectively. In this process, the energy of the refined TS dropped to -156 kcal/mole (Figure 4-33). This refined TS was verified by a FORCE calculation to give only one negative vibration (Figure 4-34) indicating a true transition state.

When the verified TS was used as a starting point for the “reaction path” experiment, a typical reaction path was produced (Figure 4-35). This represents the IRC values for the conversion from reactants to products and going through the TS. The energy of the reactant was higher than that of the product but the energy of TS was the highest. The potential

energies of the reactant, the transition state and the product were -190.5, -156 and -214.3 kcal/mole respectively. The TS structure from the reaction path experiment had the same values for the energy, distance and angle as the TS obtained previously from the map reaction experiment after being refined.

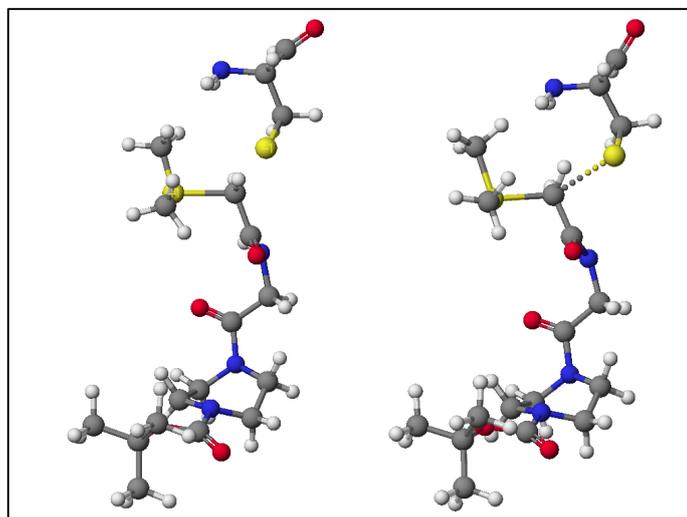


Figure 4-33: Left: TS structure before refinement. Right: TS structure after PM3 refinement. The dotted line is used by CAChe to represent weak bonds.

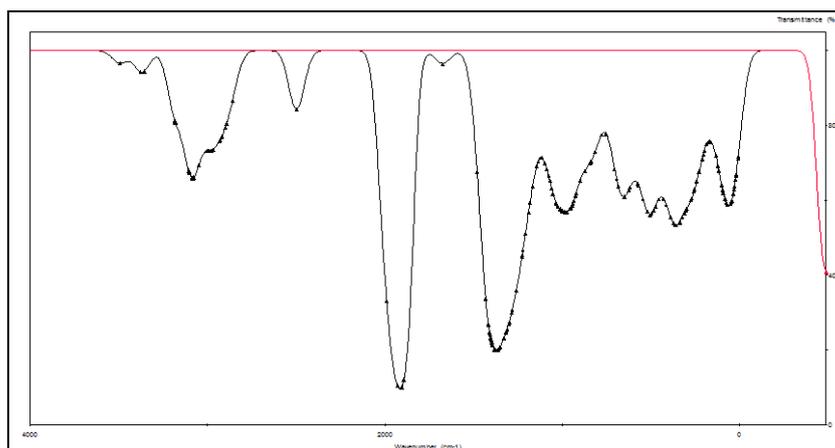


Figure 4-34: IR spectrum of refined TS showing ONE negative vibration (red).

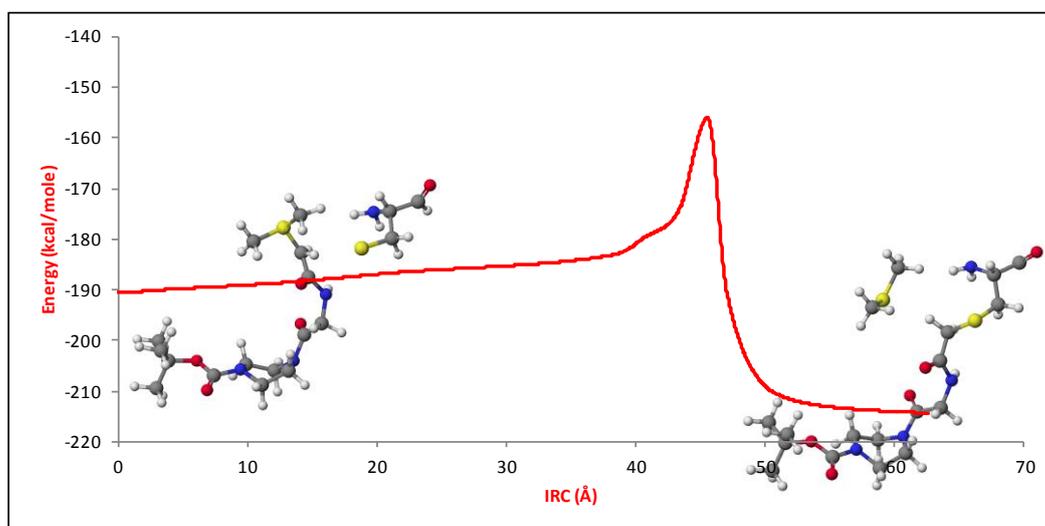


Figure 4-35: reaction path for 1b starting from reactants (left) to products (right).

The first set of experiments (map reaction, TS refinement and verification and reaction path) was then applied to the rest of the compounds and their results are presented in Table 4-7. The presented results include the values of the reaction coordinates in the TS structure for each compound along with the activation energy values. All the energy values have been obtained from “reaction path” experiments on TS structures that were verified by showing only one negative vibration in the FORCE calculation.

Compound	TS Coordinates				Energy (kcal/mol)		
	Distance (Å)		Angle (°)		TS	Reactant	Activation
	Start	End	Start	End			
<b>1d</b>	4.99	2.33	75.7	151	-123.10	-151.18	-28.07
<b>1f</b>	5.64	2.31	45.6	151	-75.70	-109.57	-33.87
<b>3b</b>	4.92	2.28	67	146	-126.00	-152.82	-26.82
<b>3c</b>	5.65	2.32	43	143	-117.92	-146.68	-28.76
<b>3h</b>	3.68	2.34	104.4	145	-98.32	-120.14	-21.81
<b>3m</b>	5.67	2.35	104.7	147	-181.76	-204.81	-23.05
<b>3n</b>	4.9	2.29	81.5	148	-173.63	-202.56	-28.94

**Table 4-7: Reaction coordinates and energy values for the remaining 7 compounds.**

It can be seen from the table that the values for the reaction coordinates are very similar between the different compounds. This indicates that the transition state that precedes the formation of the product would have these values for the coordinates if the warhead was an electrophilic carbon attached to a sulfonium ion being attacked by a nucleophilic thiolate ion of a cysteine residue. It can also be deduced from the table that the structure of the transition state around the reacting centre is not affected by the structure of the rest of the ligand molecule. In addition, it can be seen that despite the wide range of the starting values for the angle coordinate (43-104.7°), all the TS structures had values for this coordinate in a range of less than 8 degrees (143.6-151°). This implies that the nucleophile and electrophile should get into a semi-straight line before the reaction could occur.

The charges of the 2 sulphur atoms and the orders of the bonds to be formed and broken in the TS structures are donated in Table 4-8. It is evident from the table that all the ligands followed a similar path during reaction in terms of the charges and bond orders. Other quantities that were found to be very similar in all TS structures were the distance between SG and S+ which ranged between 4.1 and 4.2 Å and the length of the bond to be broken which ranged between 1.98 and 2.08 Å (Figure 4-36).

From the information presented so far, a proposed mechanism for the reaction of inhibitors having sulfonium ion warheads with the positively charged sulphur of the active site cysteine residue of TG2 may be drawn. Such a mechanism would involve the approach of SG to S+ accompanied by an increase in the charge of the first and a reduction in the charge of the latter. At the TS, an almost isosceles triangle is formed; at its sides are the bonds to be

formed and broken and the base is the distance between the 2 sulphur atoms (SG and S+). This is followed by the neutralisation of the charges of both sulphur atoms accompanied by the breaking of a bond (EC-S+) and the formation of a new bond (SG-EC) (Figure 4-37).

Compound	Partial Charge						TS Bond order	
	Reactant		TS		Product		Broken	Formed
	SG	S+	SG	S+	SG	S+		
<b>1b</b>	-0.88	0.75	-0.57	0.33	0.02	-0.02	0.53	0.38
<b>1d</b>	-0.69	0.71	-0.61	0.33	-0.01	-0.03	0.54	0.37
<b>1f</b>	-0.84	0.67	-0.59	0.32	0.00	-0.01	0.51	0.39
<b>3b</b>	-0.89	0.72	-0.57	0.27	-0.02	-0.01	0.47	0.43
<b>3c</b>	-0.89	0.69	-0.55	0.31	-0.02	0.02	0.53	0.39
<b>3h</b>	-0.86	0.67	-0.55	0.33	0.00	0.04	0.53	0.38
<b>3m</b>	-0.73	0.69	-0.57	0.33	-0.01	0.02	0.55	0.37
<b>3n</b>	-0.82	0.65	-0.56	0.30	-0.01	-0.03	0.51	0.39

Table 4-8: Charges and bond orders.

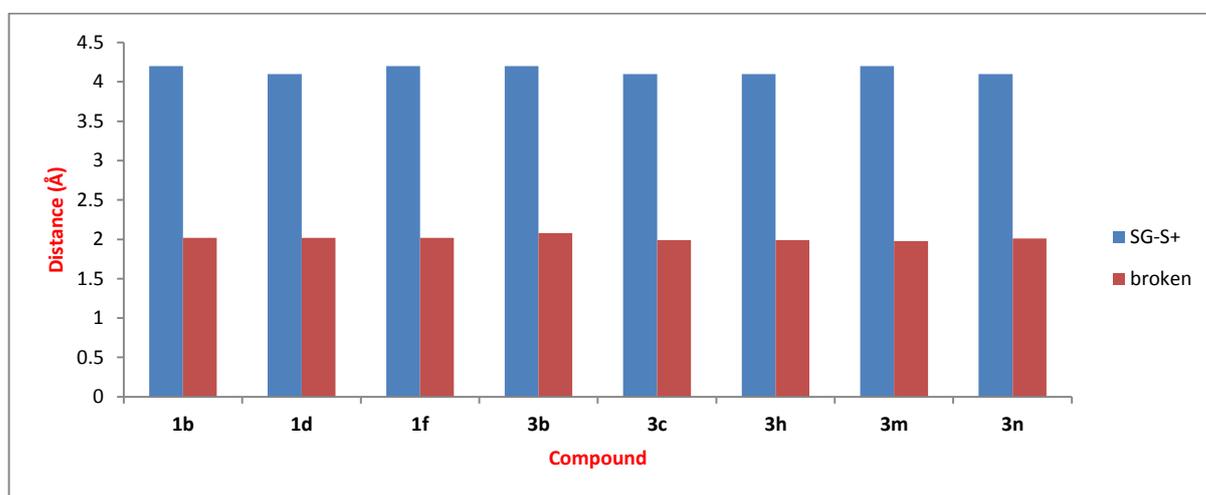
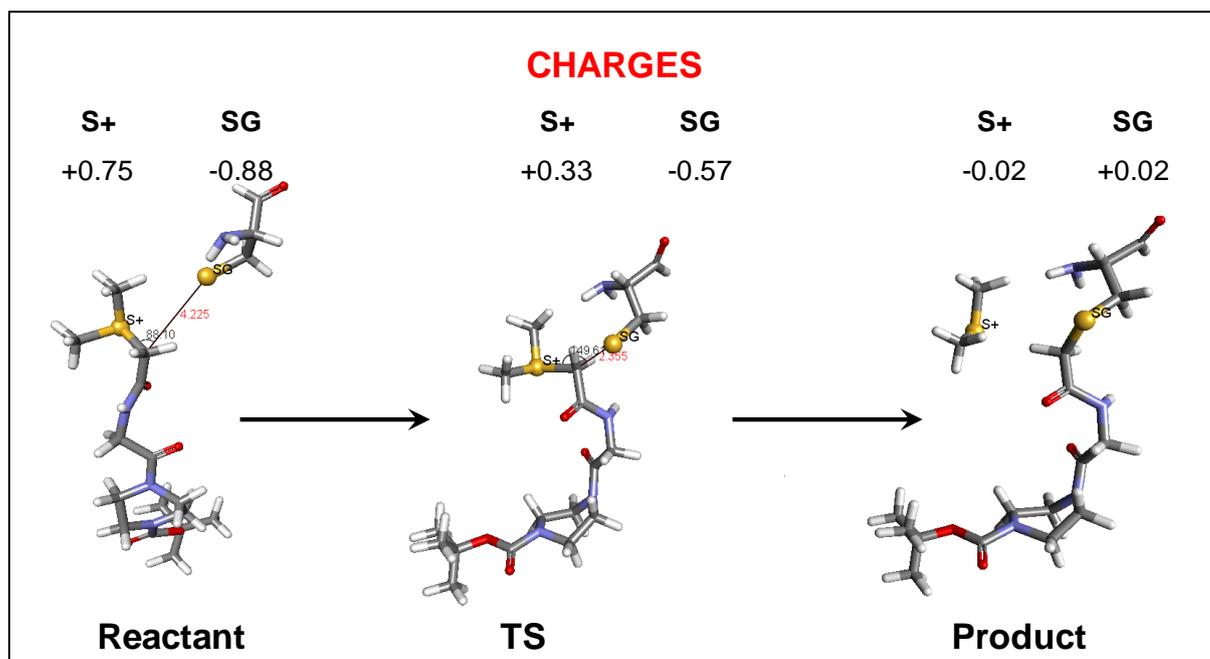


Figure 4-36: Distance between SG and S+ and of the bond to be broken in the 8 TSs.

These features are consistent with an  $S_N2$  reaction mechanism, where the nucleophile (SG) attacks the electrophile (EC) from the opposite side of the leaving group (S+), and at the transition state, the three reacting groups are in a straight line with an angle of  $180^\circ$  (Smith & March 2001). The range of angles seen in Table 4-7 for the transition states of the compounds, although very similar, does not reflect straight lines. However, angle values and the way they changed from their starting standards provide evidence that the system attempted to render the attack of SG from the opposite side of S+. The non- $180^\circ$  values for the angles may be explained by the attraction forces between the oppositely charged sulphur atoms (S+ from the sulfonium ion and S- from CYS277 thiolate ion) involved in the reaction, causing the two sulphur atoms to try to come close to each other with a resultant bending in the angle to the values reported in the table.



**Figure 4-37: Mechanism of the reaction including the change in charges and reaction coordinates between reactants, TS and products.**

The bond orders of the TS structures reported in Table 4-8 are also consistent with  $S_N2$  mechanism, where such a mechanism would involve the change in the hybridization of the electrophilic carbon from  $sp^3$  to  $sp^2$  at the transition state with a  $p$  orbital that is shared by the nucleophile and the leaving group (Smith & March 2001). This would result in partially formed bond with the nucleophile and a partially broken bond with the leaving group. The bond orders shown in the table indicate that the 2 bonds are weaker than a single bond, confirming a TS structure of an  $S_N2$  mechanism.

One last characteristic of an  $S_N2$  mechanism is that the three non-reacting atoms attached to the electrophilic carbon are nearly coplanar at the transition state to increase the chances of overlap between the segments of the  $p$  orbital of the electrophilic carbon with the nucleophile and the leaving group (Smith & March 2001). Figure 4-38 represents a different view for the TS of compound **1b**, and it shows EC with its 3 connected atoms. All the 4 atoms appear to be on approximately the same plane in space, again confirming the  $S_N2$  mechanism.

The values for the activation energy for the compounds were plotted against their  $IC_{50}$  values; and a near straight line was obtained. The  $R^2$  value for the line was 0.61 (Figure 4-39). If compounds **3c** and **3m** were removed from the curve, the  $R^2$  value would rise to 0.95 (Figure 4-40). This implies that the method works best in predicting the biological activity if the  $IC_{50}$  was equal to or less than 1  $\mu$ M.

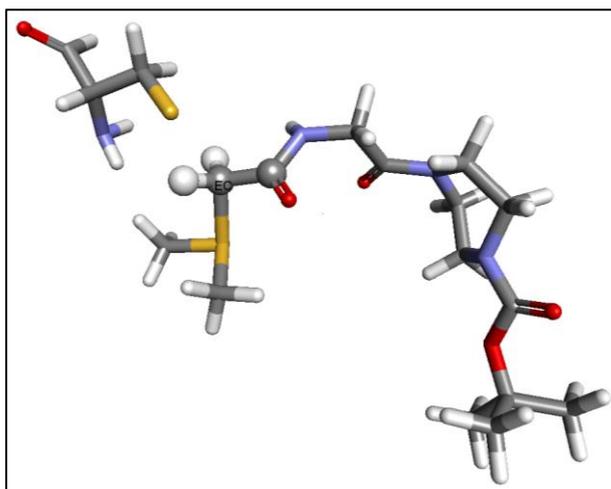


Figure 4-38: TS structure for compound 1b in the stick form, showing EC and the three non-reacting atoms attached to it in the ball form to confirm the coplanar geometry they adopt in accordance with  $S_N2$  mechanism.

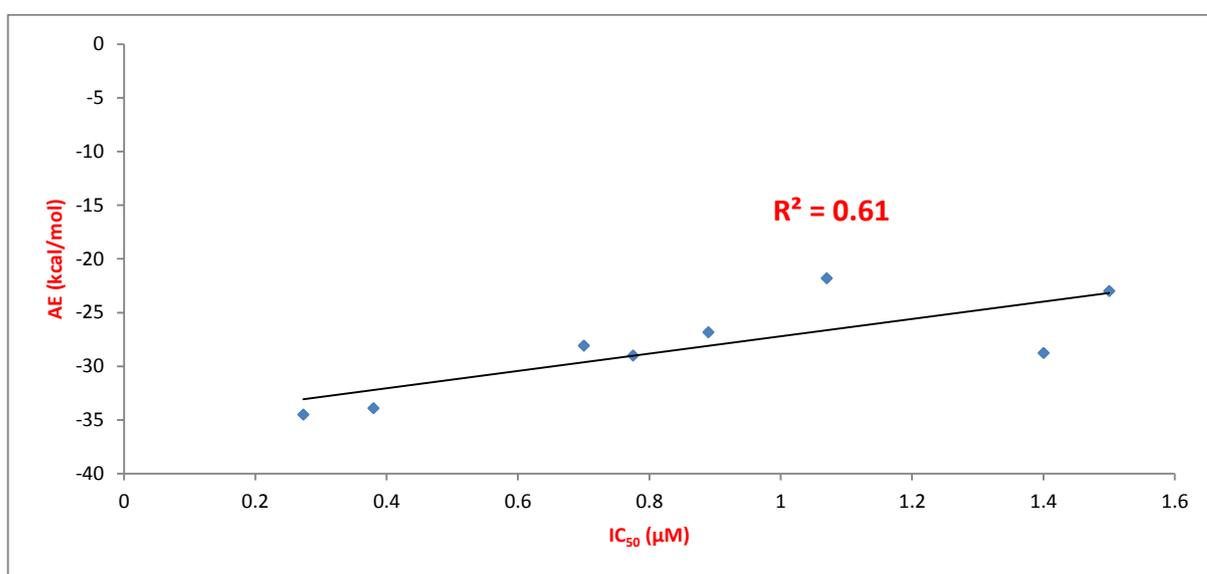


Figure 4-39:  $IC_{50}$  values for the 8 compounds versus their activation energies.  $IC_{50}$  values are given in Table 4-6.

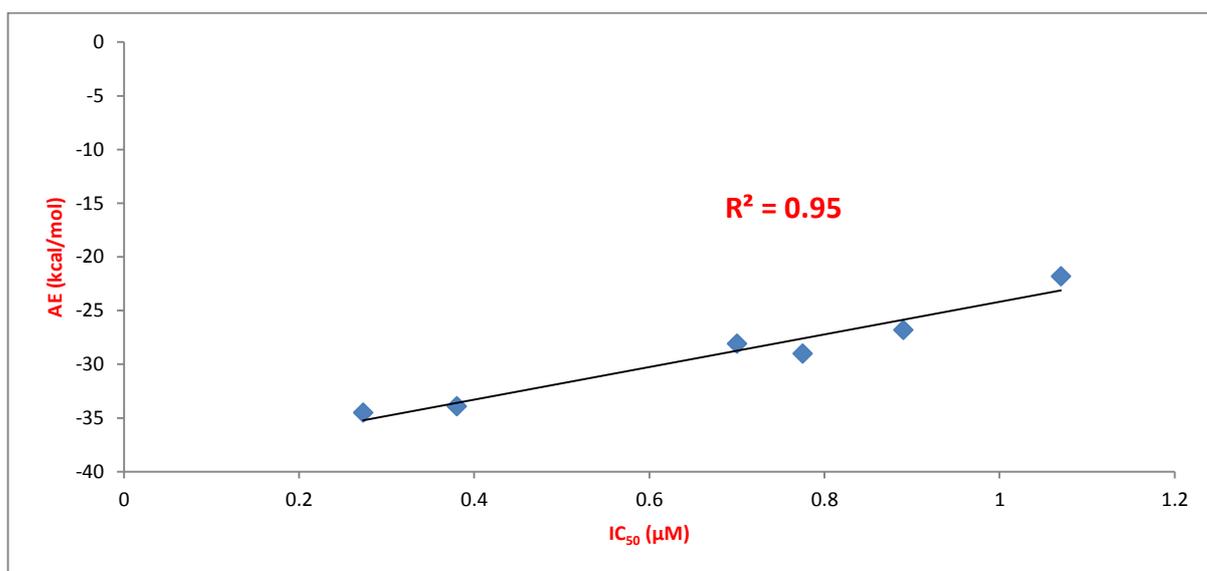


Figure 4-40: Correlation with biological activity for compounds with  $IC_{50}$  of 1  $\mu M$  or less.

The same protocol was also attempted on two inactive compounds (**2c** and **4c**). In both cases, a reasonable PES was obtained but proper TS could not be extracted. Multiple conformations were taken from the PES and refined to produce TSs with acceptable structures in terms of the distances and the angle but they could not be verified as true transition states, where there always was more than one negative vibration. Most probably, this was due to the fact that the resultant TS structures had high energy values, in the range of -70 to -85 kcal/mol.

A single label for the distance of the bond to be formed has been tried as a reaction coordinate for compound **1b** in the map reaction experiment and a TS structure was extracted from the resultant PES (Figure 4-41A). The TS was refined and verified to give one negative vibration but it could not produce a reasonable reaction path graph (Figure 4-41B) and hence no relevant structures for the product or the TS could be obtained. The distances for the two bonds to be formed and broken have also been tried as separate reaction coordinates in CAChe but the PES graph from the map reaction experiment was not meaningful; no reasonable TS structure could be extracted. This happens because the breakage of the EC-S+ bond cannot be controlled by a reaction coordinate, and it takes place immediately and suddenly after the SG-EC bond is formally created. The two failed experiments with the bond to be formed on its own and the two bonds separately were the justification for the use of the angle with the distance as the initial reaction coordinates.

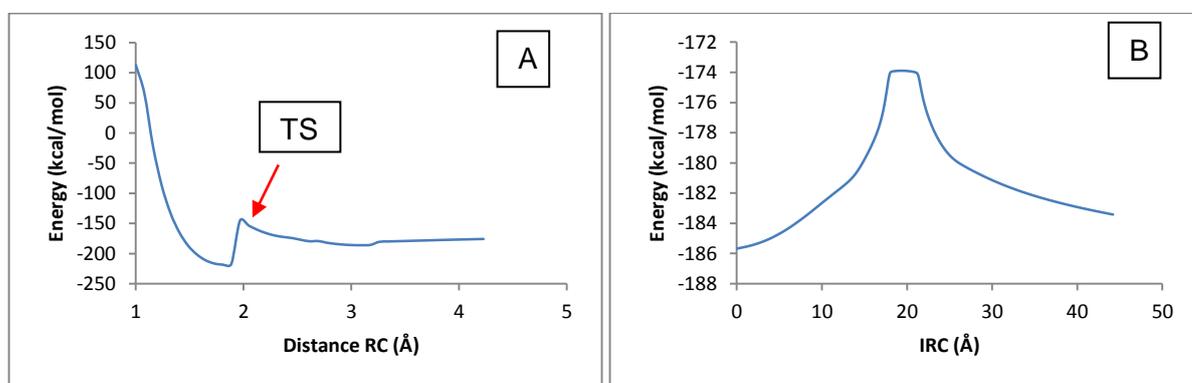


Figure 4-41: PES (A) and reaction path (B) graphs for compound **1b** when using **1** label as the reaction coordinate.

The IRC values in the reaction path graph shown in Figure 4-35 should not be interpreted as any of the original coordinates used in the beginning for the generation of the initial PES. The IRC represents the movements of all the atom in the system in angstroms in their path starting from the transition state and going to the reactants and to the products (Stewart 2002).

It has been shown during the course of this work that the structure of the transition state for the reaction of the inhibition of TG2 with compounds containing sulfonium ion is the same in

the area of the reaction and is not affected by the remaining structure of the ligand. Despite this observation, the method used was able to produce activation energies that could be correlated to biological activity. This implies that the conformation of the ligand in the starting structure was the factor that made the difference and this conformation is related to the structure of the entire ligand justifying the good correlation with biological activity. Such an outcome would compensate for not including the entire system in the calculations.

### 4.3 Conclusions

Quantum mechanical methods are important for modelling inhibitors whose activity involves the formation and breakage of covalent bonds. Umbrella sampling simulations combined with quantum mechanical treatment of the reaction centre were able to simulate the reaction between inhibitors carrying the acrylamide warhead and TG2 active site cysteine residue using HIS335 as a source for the proton required to saturate the alpha carbon of the acrylamide double bond. The reaction was simulated as a single concerted stage with PM3 level of theory and as 2 successive steps with DFTB as the QM method. The 2 mechanisms produced reasonable correlations with biological activities for the compounds and had supporting evidence. The single-stage simulations, for example, showed a more obvious desolvation effect happening near the formation of the bonds, when one or more water molecules left the active site. The charge distribution around the reaction centre, during the 2-stage simulations, was more consistent with the reaction proceeding as 2 separate consecutive stages, with the formation of an oxyanion intermediate at the end of the 1<sup>st</sup> stage. It can therefore be concluded that the studied reaction may proceed with either mechanism depending on the available evidence to support one of them.

Reaction path experiments using CAChe were able to generate typical reaction graphs for the transition from reactants to products and going through the transition state for inhibitors bearing the sulfonium ion warhead. The fact that the structures of the transition states from the 8 compounds were similar around the reaction centre gives these structures reliability in predicting the mechanism of the reaction. The latter would involve the formation of an isosceles triangle; at its base are SG and S+ prior to the dissociation of the sulfonium ion and the covalent inhibition of TG2. An S<sub>N</sub>2 reaction mechanism was proposed, and evidence for such mechanism in this reaction system was presented, in the form of the angles at the transition states, the orders for the bonds to be formed and broken and the plane involving the electrophilic carbon and the three atoms attached to it. The activation energies obtained and their good correlation with the biological activity further validate the method.

# **Chapter 5**

## **Allosteric Inhibition of TG2**

## 5 Allosteric Inhibition of TG2

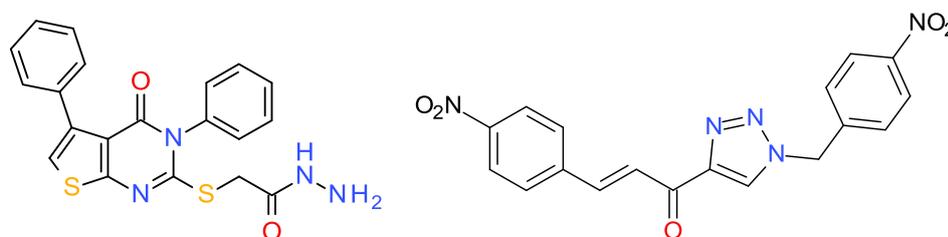
This chapter covers the computational work that was performed with respect to the allosteric inhibition of TG2. This involved docking experiments using proposed allosteric inhibitors, followed by MD simulations on the docked complexes. Accelerated MD was applied instead of conventional MD, because the change in TG2 that was expected from the binding of the allosteric inhibitors was supposed to involve the folding of a large portion of TG2, and such change is unlikely to be sampled using conventional MD.

### 5.1 Introduction

As mentioned in the general introduction (sections 1.2.2 and 1.2.3), TG2 is normally inactive by binding to GTP or GDP. The binding of these molecules induces a conformational change that involves the folding of the two C-terminal  $\beta$ -barrels onto the catalytic core, causing the unavailability of CYS277 for catalysis (closed, inactive TG2 conformation). Those natural regulators dissociate from TG2 in conditions associated with high calcium ion concentrations, with unfolding of the TG2 and the exposing of CYS277 for catalysis (open, active TG2 conformation) (Figure 1-2) (Jang et al. 2014; Liu et al. 2002; Pinkas et al. 2007).

The work in this chapter aimed at examining the effects of the binding of 2 proposed allosteric TG2 inhibitors on the TG2 open conformation, and whether this binding can induce the closure of the TG2 structure. This was achieved by docking the inhibitors into a predicted TG2 allosteric site and applying accelerated MD on the resultant complexes.

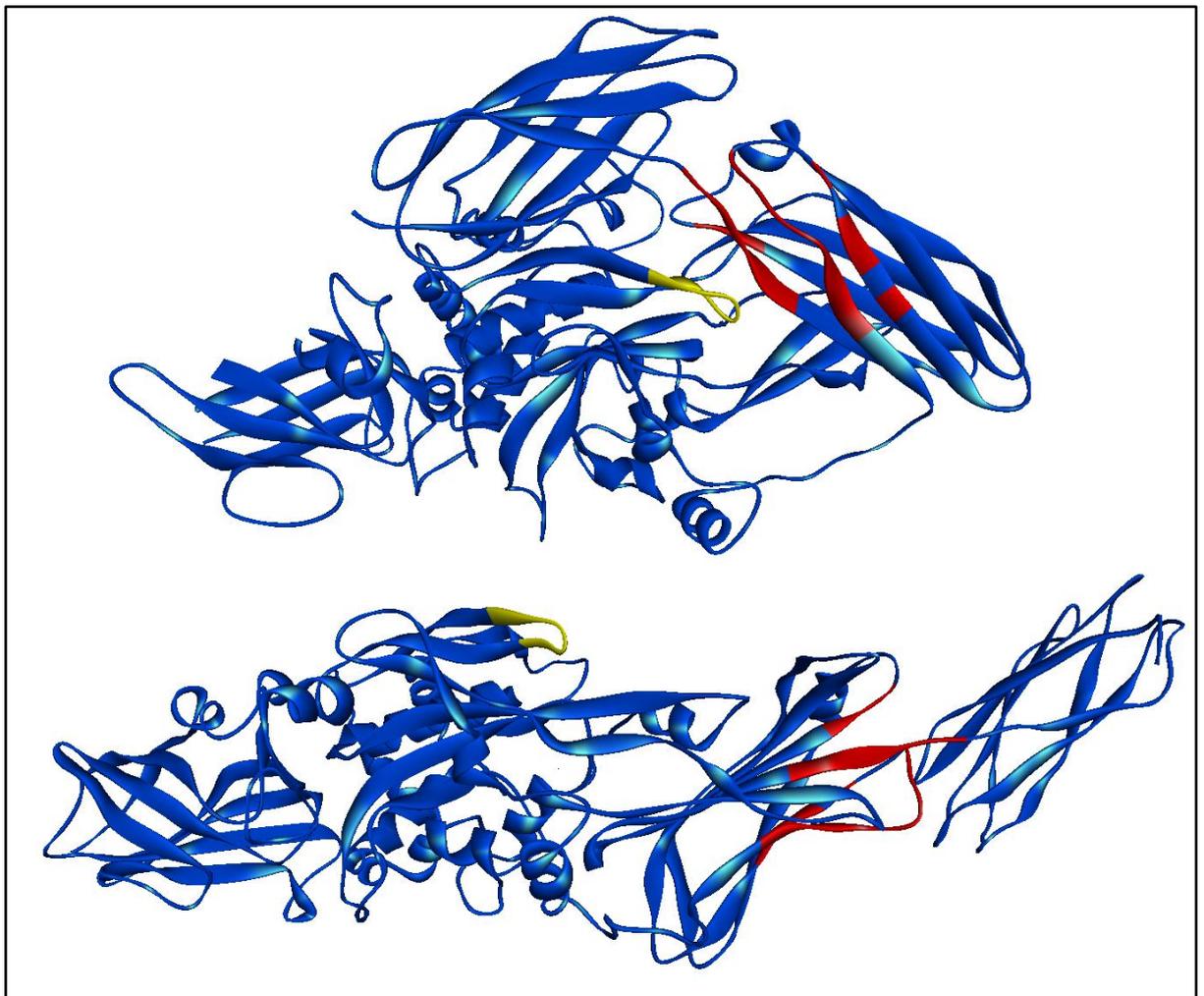
The 2 allosteric inhibitors used in this work were compound **i3** by Case & Stein (2007), which was proposed by the authors to bind to an allosteric site in TG2, and compound **i2** by Pardin, Roy et al. (2008) which was proven to be allosteric by Caron et al. (2012) (Figure 5-1) (section 1.4.1.2). The research associated with both compounds could not prove that the allosteric inhibitors would bind to the same site as GTP/GDP.



**Figure 5-1: Compounds i3 (left) and i2 (right) which were used as allosteric inhibitors.**

The GDP binding site is located in the cleft between the catalytic core and the first  $\beta$ -barrel. In the open conformation, these two regions are distinctly separate. In the crystal structure

1KV3 (Liu et al. 2002), in which TG2 is in closed conformation and complexed with GDP, there are 31 TG2 residues within 8 Å of the GDP molecule. Of these, 7 residues are on the catalytic core domain of TG2 while the remaining 24 residues are on the 1<sup>st</sup> β-barrel. In the open conformation of TG2 (2Q3Z), the 7 residues of the catalytic core are located on a distant region on TG2 from the 24 residues on the β-barrel (Figure 5-2). Furthermore, the largest section of GDP binding site from the 1<sup>st</sup> β-barrel is located on the surface of TG2. Consequently, it would be difficult to define a binding site for allosteric inhibitors in the open form of TG2 using the binding site definition from GDP-bound closed structures.



**Figure 5-2: TG2 closed, inactive (top) and open, active (bottom) conformations, showing the residues that are within 8 Å of GDP in the closed conformation (coloured yellow and red). Yellow residues are those on the catalytic core while red residues are those on the 1<sup>st</sup> β-barrel.**

Since it is the open form of TG2 that requires inhibition, it may be more appropriate to use a binding site for allosteric inhibitors that could be defined properly in the open conformation. This hypothesis was strengthened by the fact that no published research for allosteric inhibitors, to our knowledge, proved the binding site for the inhibitors to be the same as that

for GTP or GDP nor suggested an alternative binding site for allosteric inhibitors. Therefore, an approach has been employed to predict an alternative binding site for allosteric inhibitors.

A collaboration was initiated with Dr Blair Johnston and his PhD student Antony Vassileiou from the Strathclyde Institute of Pharmacy and Biomedical Sciences at the University of Strathclyde. They developed a computational tool for the prediction of allosteric binding sites in proteins. Their approach involves performing analysis on MD trajectories of the studied proteins; the analysis comprises residue fluctuation, accessible surface area, energy correlation and simple intra-sequence differences (SID). SID (Pritchard et al. 2003) is a bioinformatics tool built to help understand the topology of protein folding. The collaborators used their protocol on a training set of 40 proteins with known allosteric sites from the Protein Data Bank and the protocol was able to predict the allosteric site correctly in the majority of cases. They presented their work as a poster in the MGMS (Molecular Graphics and Modelling Society) Meeting in August 2014 in the University of Strathclyde. The meeting was held between 20<sup>th</sup> and 22<sup>nd</sup> of August, 2014 under the name Modelling Molecules and Materials – M<sup>3</sup> (<http://www.m3glasgow.org.uk/>).

## 5.2 Methods

### 5.2.1 Prediction of the allosteric binding site

A MD trajectory of the open conformation of TG2 was sent to the collaborators in the University of Strathclyde for the prediction of the alternative binding site of allosteric inhibitors. The MD trajectory comprised a run that was applied starting from the 2Q3Z (Pinkas et al. 2007) crystal structure after adding and minimising the missing residues, but this time without removing any residues from TG2; the whole protein was used. Similar settings to those described in section 2.2.1.1 (MD simulation on empty TG2) were used here; minimisation followed by heating, equilibration and finally production. The MD simulation was run for a total of 252.075 ns. This MD simulation was performed by Dr Dan Rathbone of the School of Life and Health Sciences in Aston University, using the servers whose details were mentioned in section 2.1.1.

The prediction protocol found a binding site composed of 10 residues and was located in a pocket surrounded by 3 helices at the end of the catalytic core domain of TG2. Before it was used in docking, residues were added at the beginning, middle and end of the segments of the predicted site to raise the number of residues involved to 27 (Figure 5-3).

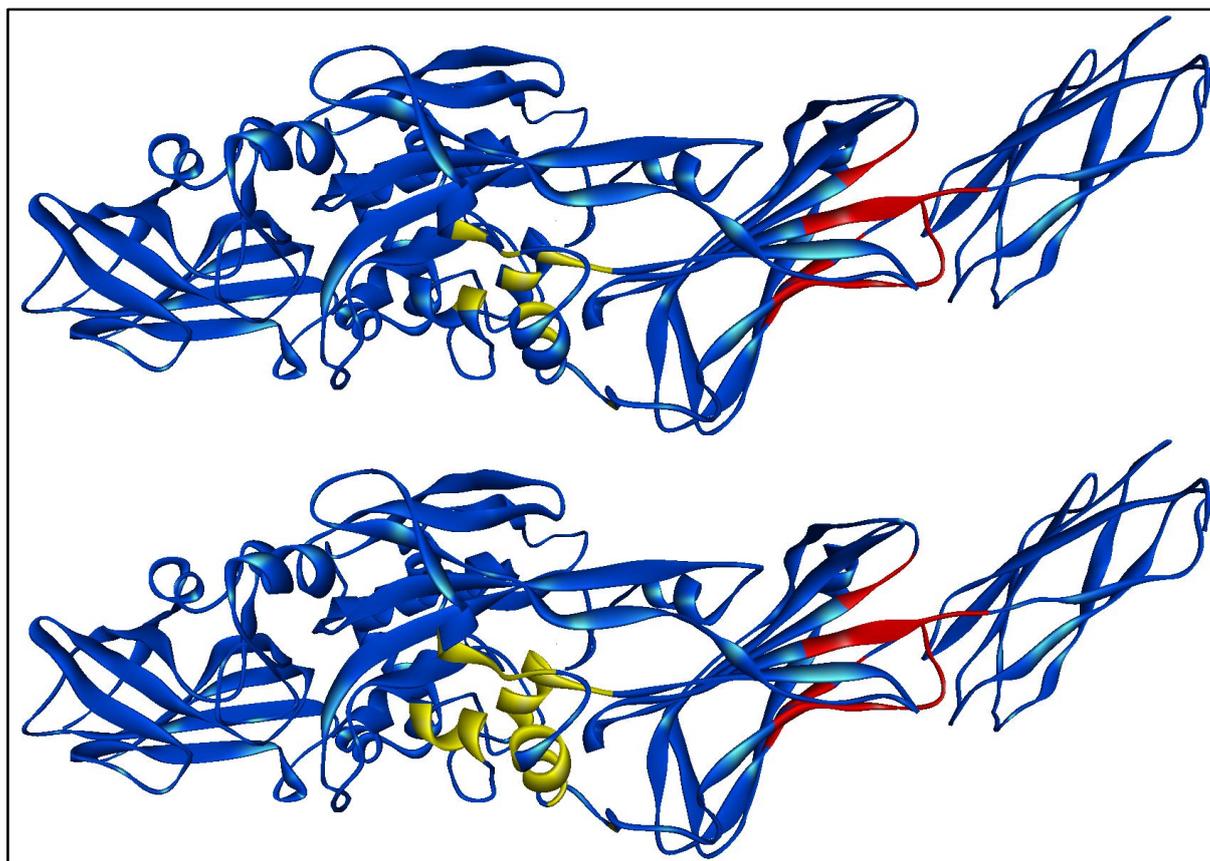


Figure 5-3: The predicted allosteric binding site by the Strathclyde protocol (top) versus the site used for docking purposes (bottom) after adding residues between the segments, both coloured yellow. GDP binding site is shown in red for comparison of the location of the 2 sites.

### 5.2.2 Docking

The GOLD program was used to run the docking experiments of compounds **i3** and **i2**. The TG2 structure used as the model for docking was the last frame from the conventional 252.075-ns MD run described previously. Similar settings to those used in the original dockings were used here, including the use of GoldScore as the primary scoring function and rescoring each pose with CHEMPLP. Early termination was turned off and 20 poses were generated for each of the 2 ligands. Ten residues from the core of the predicted site were chosen to be treated flexibly using the library option in GOLD. Default docking speed was used. The 27 residues from the predicted site used to define the binding site were (flexible residues are in **bold**): ARG377 ALA378 **ILE379** **LYS380** **GLU381** PRO391 PHE392 VAL393 PHE394 ALA395 **GLU396** VAL397 **ASN398** ALA399 THR442 TYR443 **LYS444** TYR445 PRO446 **GLU451** **GLU452** ARG453 **GLU454** ALA455 PHE456 THR457 **LEU462**.

### 5.2.3 Molecular dynamics

Molecular dynamics was applied on the last frame from the 252.075-ns conventional MD without any ligands. Two docking complexes for compounds **i3** and **i2** in the predicted

allosteric site have also been subjected to MD. The aim was to examine whether the presence of the ligand in the predicted site would induce the closing of TG2 conformation into the inactive form. Since this involves a large conformational change, accelerated MD (aMD) was used at this stage.

aMD was applied according to the procedure developed by Pierce et al. (2012). The systems were loaded into LEaP program for the preparation of the parameter and topology files after solvating with 8-Å octahedron of TIP3P water molecules. The systems were then minimised, heated and equilibrated using the same settings used previously and described in detail in section 2.2.1.1. The restart file from the equilibration stage was used as the starting structure for the aMD run. The aMD runs used the same settings used for the production runs previously in conventional MD with the addition of the “iamd” term which triggers aMD. The whole potential of the system was boosted with an additional boost for the torsions (iamd=3). The average potential energy obtained at the end of the equilibration phase was used to define the total boost (average potential energy + (0.16 kcal/mol.atom \* number of atoms)) and the average dihedral energy was used to define the torsion boost (average dihedral energy + (4 kcal/mol.residue \* number of the solute residues)). aMD was run over 30-ns trajectories to a total of 36 trajectories with a simulation time of just over 1 microsecond (unless otherwise specified).

In addition to empty TG2 and the two docking complexes, 3 supplementary systems were subjected to aMD. These extra systems consisted of TG2 in its open form (the same structure used for the docking of the 2 allosteric inhibitors) around which 24 copies of each of the 2 allosteric inhibitors and GDP were placed at a distance of at least 7 Å from the protein surface. These copies were added manually using Accelrys DS Visualizer v4.0 (Accelrys Software Incorporation 2013). Figure 5-4 shows the starting structures for these runs for compounds **i3** and **i2**.

Different analyses were performed on the resulting trajectories; mainly the RMSD and atomic fluctuations were measured using CPPTRAJ (Roe & Cheatham 2013), where the first frame of the production phase was always used as a reference and the calculations involved backbone atoms of TG2 residues only. Principal component analysis was also applied on some of the resultant trajectories.

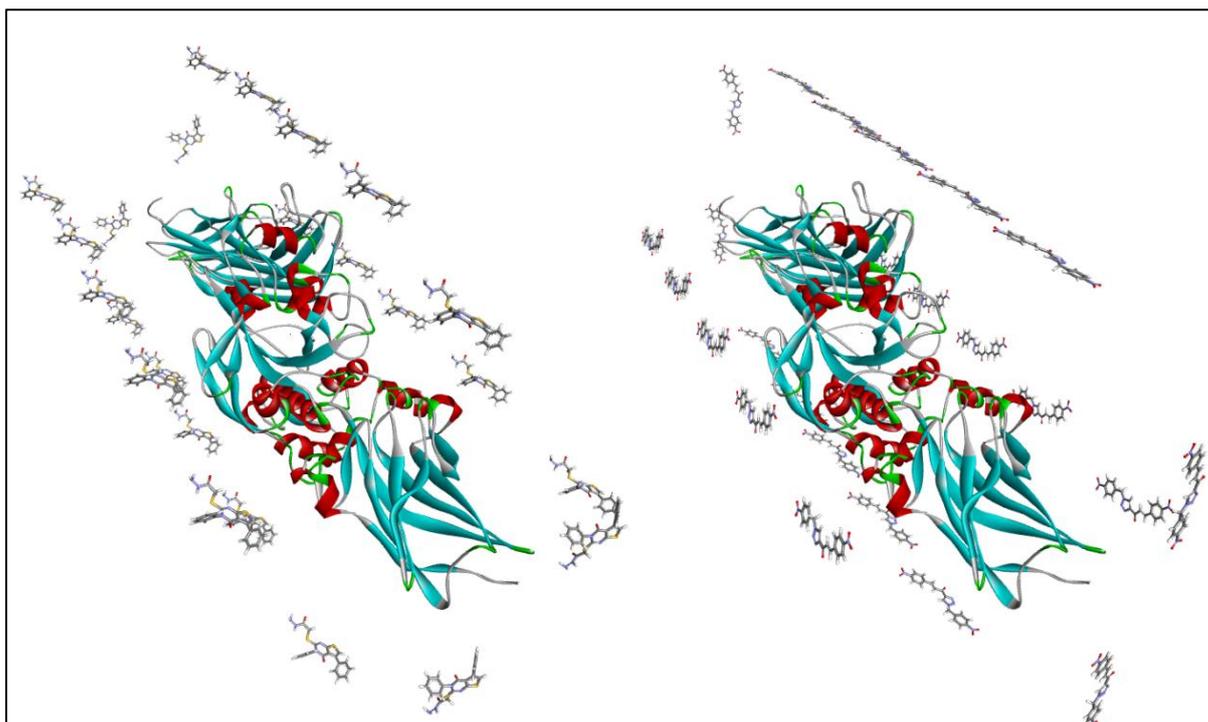


Figure 5-4: The starting structures for the aMD simulations on TG2 surrounded by 24 copies of i3 (left) and i2 (right).

## 5.2.4 Principal component analysis

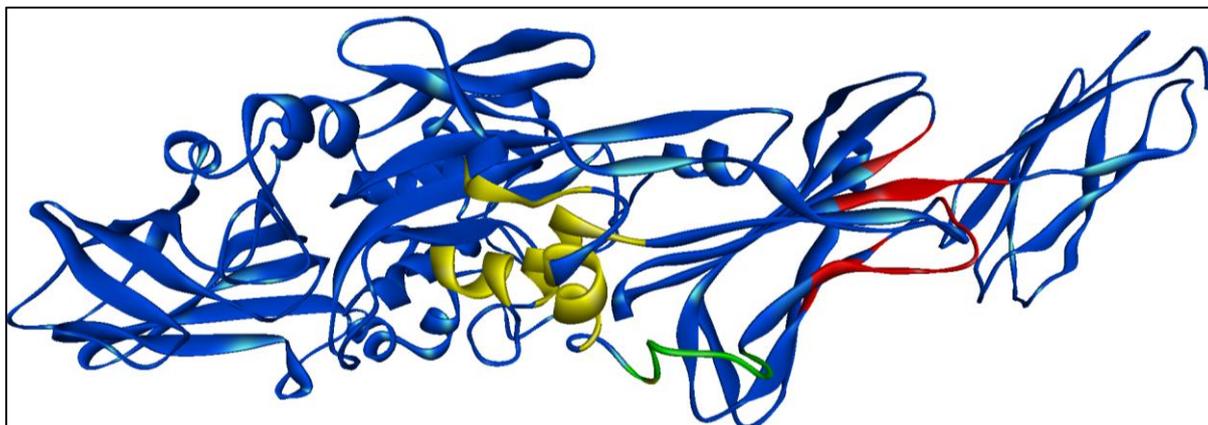
PCA was performed using the program pyPcazip (Shkurti et al. 2016). The trajectories were stripped of water and combined in a single trajectory for each individual simulation. The single trajectory was then treated with pyPcazip to calculate the principal components. Unless otherwise indicated, the top 2 components were considered. The program was also used to generate moving trajectories consisting of 20 frames each, to represent the motion in the protein represented by any principal component. The projections of each of the produced components can also be calculated by the program as a function of time. All PCA calculation were performed using only the alpha carbon atoms of TG2 residues.

## 5.3 Results and Discussion

### 5.3.1 Predicted allosteric binding site

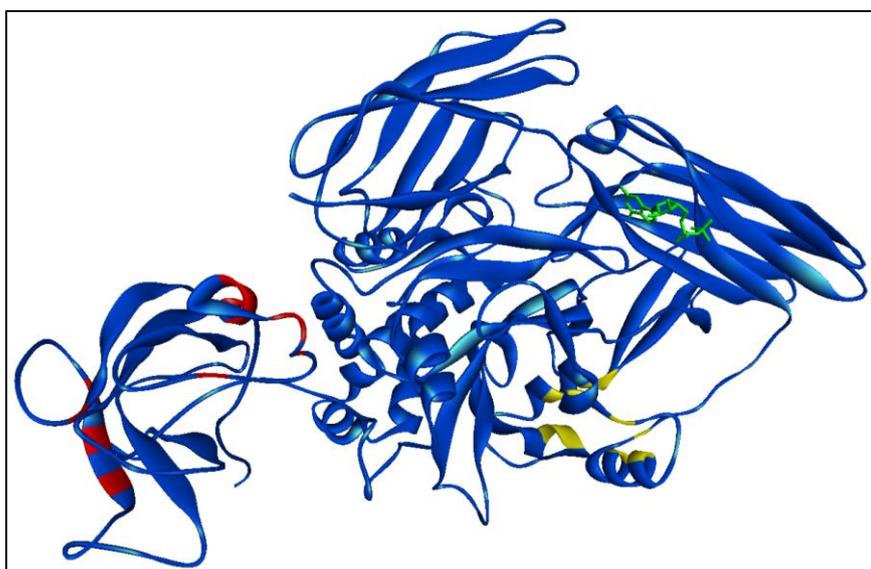
According to Jang et al. (2014), the catalytic core in TG2 ends with residue ASN460, and the 1<sup>st</sup>  $\beta$ -barrel starts with residue GLY472. The 2 domains are connected by a strand composed of 11 residues, at its end the 2 C-terminal barrels fold on the catalytic core in the closed conformation of TG2. As shown in Figure 5-3, the location of the predicted allosteric binding site at the end of the catalytic core seems reasonable to handle the job of triggering the folding of TG2, since it is positioned just before the folding segments. GDP binds in a site

located at the beginning of the 1<sup>st</sup> barrel, just after the folding segment (Figure 5-5). Therefore, in theory, the predicted site is as close to the location of the folding event as the GDP binding site, and may have the potential to serve a similar function.



**Figure 5-5:** TG2 in open conformation, showing the predicted allosteric binding site (yellow), the folding segment (green) and GDP binding site on the 1<sup>st</sup>  $\beta$ -barrel (red).

To further confirm the prediction, the prediction protocol from the Strathclyde group was additionally applied on a 100-ns MD trajectory of TG2 in the closed conformation [PDB 1KV3 (Liu et al. 2002), conventional MD run performed with the same settings used previously]. The aim was to examine the difference in prediction with the starting conformation of TG2. In this case, the predicted site was located in the N-terminal  $\beta$ -sandwich, far away from the catalytic core, the portion of TG2 that will close on the active site and the GDP binding site. This new site has been discarded and the one obtained from the open form of TG2 was used. The main evidence was based on proximity measures; the 1<sup>st</sup> predicted site was in the form of a pocket and was located just before the folding segments. The new predicted site, in addition, was on two separate regions separated by a wrapped sheet, indicating that it would be difficult to define a binding site (Figure 5-6).



**Figure 5-6:** Closed conformation of TG2 showing the binding site predicted from the trajectory of closed TG2 (red) versus the original prediction (yellow) and showing GDP in its binding site.

### 5.3.2 Docking results

Since the binding site used in these experiments is novel, there is no experimental or computational pose that could be used as a reference for an appropriate docking pose. Visual inspection was, thus, used to judge the 20 poses generated by GOLD, in addition to GoldScore ranking. One pose was selected for each compound, in which the compounds were more deeply embedded within the predicted allosteric site than the remaining solutions. Furthermore, the poses for the 2 compounds were ranked 2<sup>nd</sup> in GoldScore out of 20 solutions. In each pose, there was a network of interactions holding the compound in place, mainly in the form of Pi interactions (Figure 5-7). Therefore, even though there was nothing to compare to, these poses were deemed to be appropriate and were considered for aMD simulations to test whether the presence of allosteric inhibitors in this binding site would trigger the conformational change brought about by the binding of GTP and GDP.

### 5.3.3 Results of accelerated MD

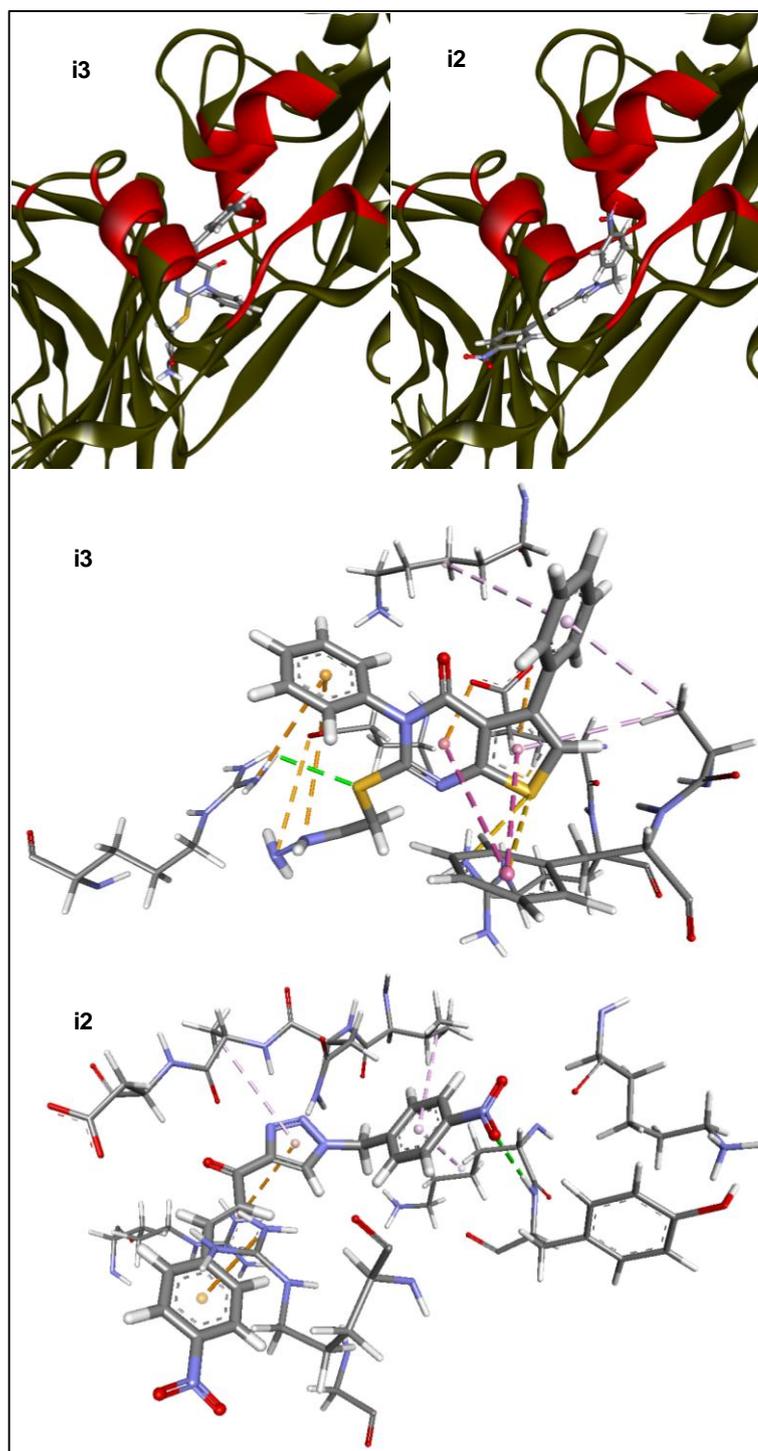
aMD is an enhanced sampling technique that enables the exploration of rare events which require passing high-energy barriers on the energy surface of the system, and the technique works by adding a boosting potential to the system when the energy falls below a predefined value (section 1.5.2.3.4) (Miao et al. 2014). It has been used to capture biomolecular events that involve relatively significant conformational changes such as the activation of the M2 muscarinic receptor, a G-protein coupled receptor (GPCR) that regulates heart rate, through the relocation of some residues and a tilting event involving one of the alpha helices of the protein (Miao et al. 2013). aMD has also been used specifically to investigate allosteric inhibition, one example involved interleukin-1 receptor (IL-1R) allosteric inhibitors in terms of binding sites within IL-1R as well as the conformational changes brought about by the binding (Yang 2015).

As mentioned in the 'Methods' section, aMD was applied on 2 docking complexes for **i2** and **i3**, on empty TG2 in the open form and on open TG2 surrounded by 24 copies of the 2 allosteric inhibitors and GDP. The results of these simulations will be presented individually.

#### 5.3.3.1 aMD on empty TG2 (Run 1)

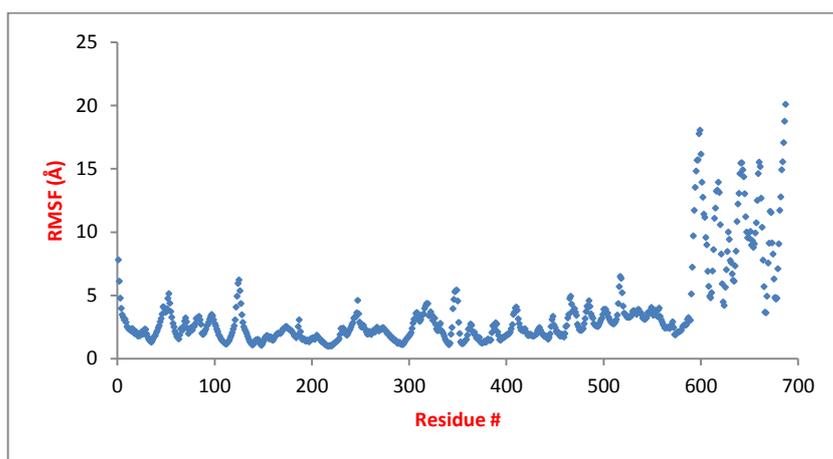
This aMD run was performed as a control for the other runs, to assess the behaviour of TG2 in an aMD simulation with the absence of any ligand anywhere in the protein. Run **1** was allowed to proceed for 1,000 ns (1  $\mu$ s). At the end, there was some kind of motion that involved the hinge region connecting the terminal  $\beta$ -barrel to the rest of TG2. The terminal

barrel showed the highest degree of fluctuation compared to other regions of TG2, and this is more clearly seen with the atomic fluctuations of the individual residues (Figure 5-8).

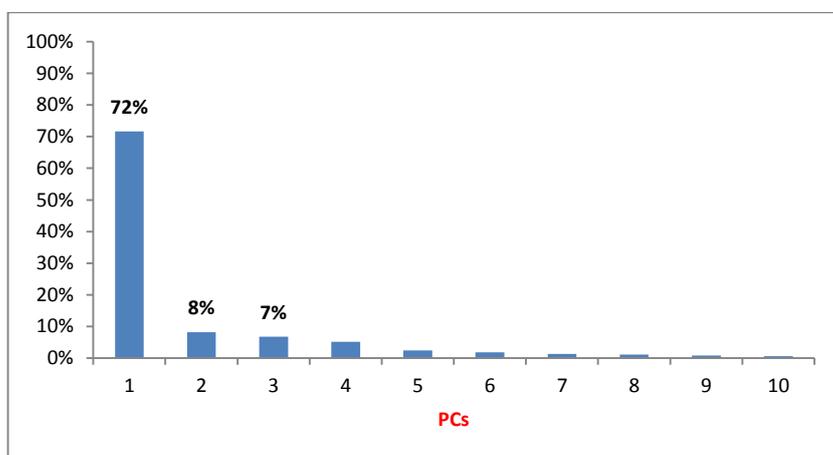


**Figure 5-7:** Docking complexes for compounds i3 and i2 in the predicted allosteric site of TG2. At the top are the poses within the predicted site which is coloured red while the rest of the protein is dark green. At the bottom are the interactions of each compound with residues in the predicted site. Green dotted lines are hydrogen bonds, light purple lines are Pi-alkyl interactions, dark purple lines are Pi-Pi interactions, orange lines are Pi-ion interactions.

The fluctuation of the terminal barrel produced bending in TG2 structure. The bending was mainly to the outside of TG2 structure and not towards the catalytic core, although some degree of bending to the inside was also observed. A better appreciation for the type of bending in the terminal barrel during Run 1 can be obtained by analysing the PCs of the run. When the trajectory of Run 1 was analysed with pyPcazip, ten principal components were produced. The 1<sup>st</sup> PC accounted for 72% of all the motion in TG2, while the 2<sup>nd</sup> and 3<sup>rd</sup> highest PCs contributed only 15%. Those PCs are shown in the graph presented in Figure 5-9.



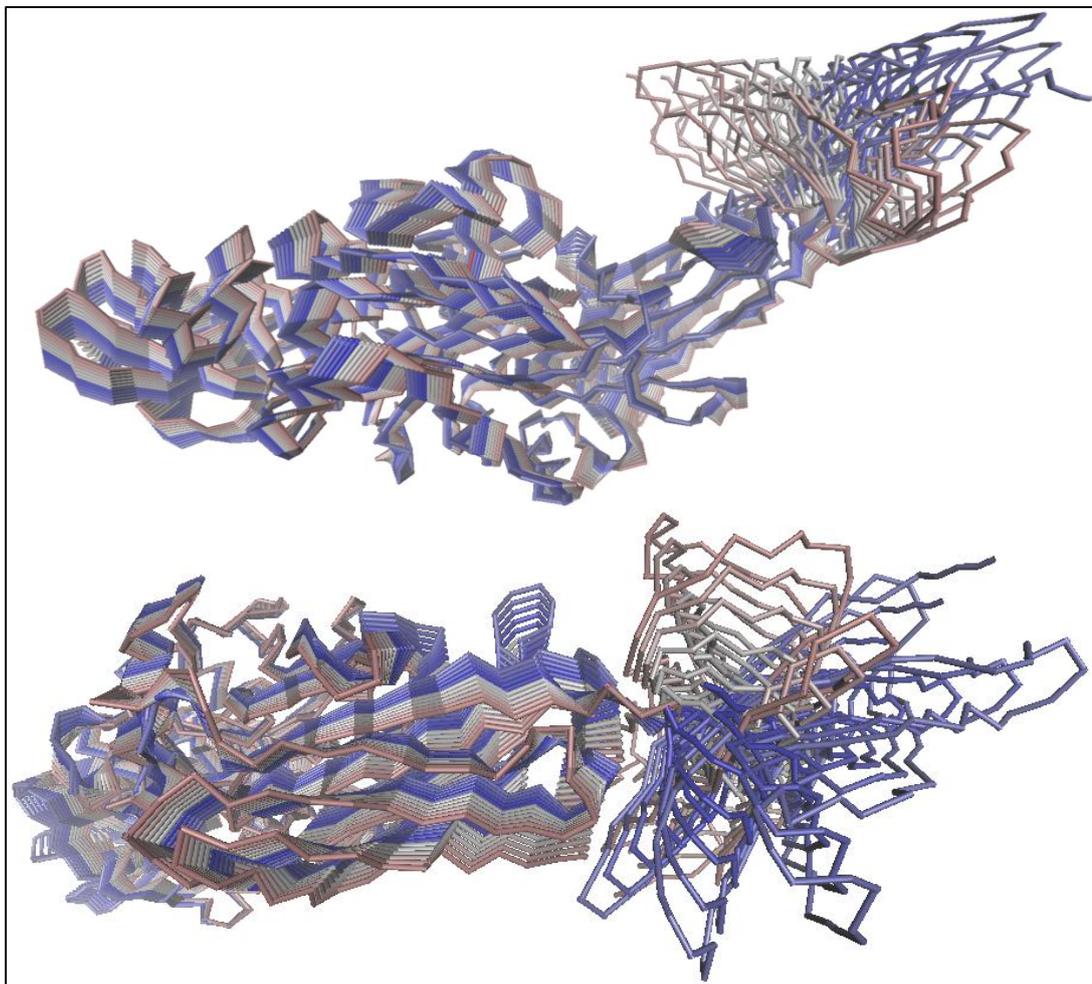
**Figure 5-8: Atomic fluctuation values for TG2 residues during Run 1.**



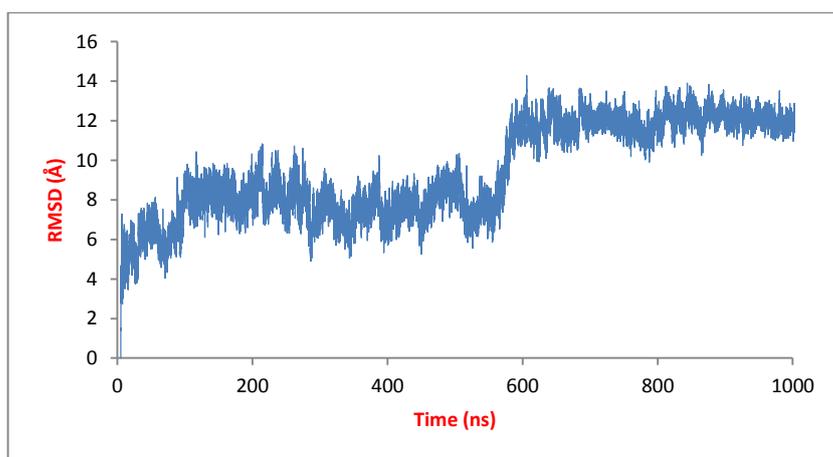
**Figure 5-9: PCs obtained by pyPcazip for the trajectory of Run 1.**

The animation produced for pyPcazip for the 1<sup>st</sup> PC is presented in Figure 5-10 and shows that the motion to the side was of a similar magnitude to the motion towards the catalytic core. This indicates that it may be difficult for TG2 to achieve the full bending required for the closure of the conformation with the subsequent inactivation of the enzyme. In any case, the bending event was not observed until more than half of the simulation was done. This was confirmed by visual inspection of the trajectory as well as by measuring the RMSD for TG2 backbone atoms for Run 1 (Figure 5-11). It was found that the 2<sup>nd</sup>  $\beta$ -barrel was oscillating

from the start of the simulation but the bending was not evident and permanent until around 560 ns of simulation time.



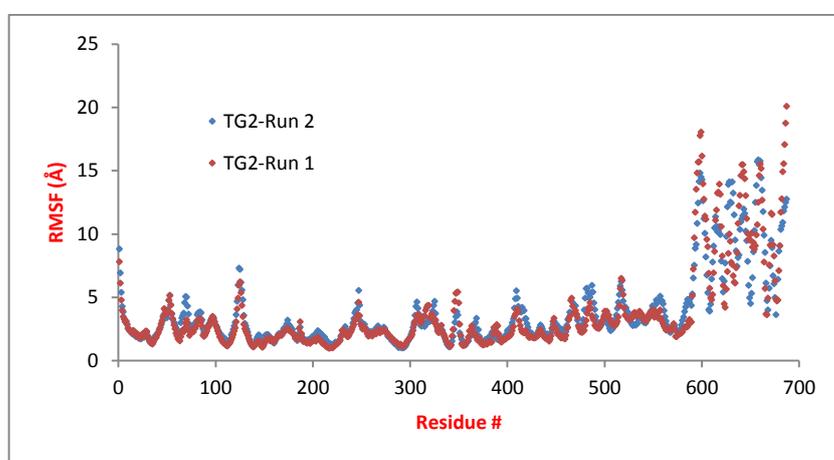
**Figure 5-10:** Animation of PC1 for Run 1 shown as a side view of TG2 (top) to view the bending towards that catalytic core and a back view (bottom) to view the bending to the sides.



**Figure 5-11:** RMSD for TG2 backbone atoms during Run 1.

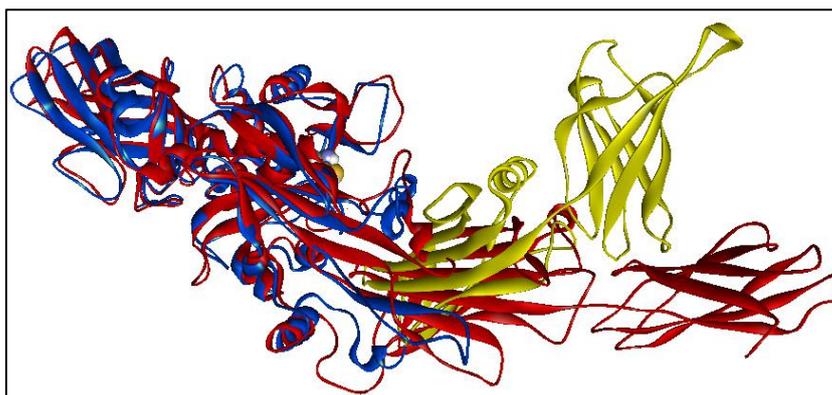
### 5.3.3.2 aMD on TG2 with 24 copies of GDP (Run 2)

This aMD simulation involved the same TG2 open conformation used as the basis for the allosteric work surrounded by 24 copies of GDP. The rationale was to see whether one of the GDP molecules would find its way to the original allosteric binding site, and if a copy did, will it be able to induce the closure of TG2 conformation? A docking complex was not considered a possibility since it is difficult to define a binding site that is located on the surface of the protein, which is the case with GDP binding site when TG2 is in the open conformation. The simulation was allowed to run for a total of 635 ns of simulation time. In the simulation, again, it was the 2<sup>nd</sup>  $\beta$ -barrel that showed the greatest amount of atomic fluctuation as manifested by the RMSF values presented in Figure 5-12.



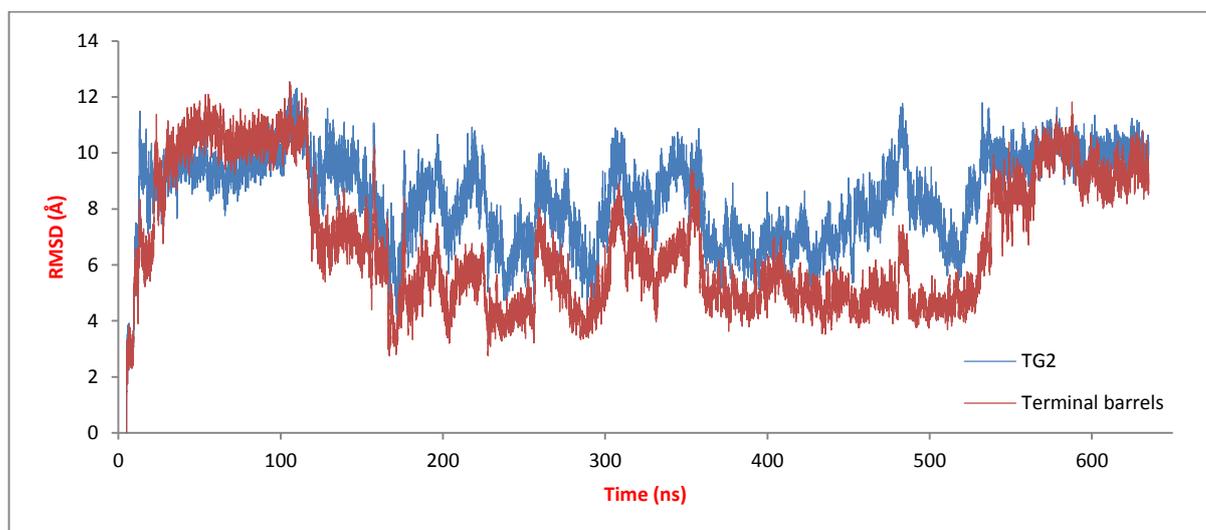
**Figure 5-12:** Atomic fluctuation values for TG2 residues during Run 2, when TG2 was surrounded by 24 copies of GDP, and those from Run 1 (empty TG2).

Visual inspection of the trajectory revealed that there was a considerable oscillation involving the terminal barrel at the start of the simulation, which did not result in any noticeable bending in the TG2 structure. Towards the end of the simulation, at  $\approx 530$  ns, a significant change in the conformation of TG2 was starting to develop. The change was manifested as a bending event that involved the hinge region connecting the two C-terminal  $\beta$ -barrels to TG2. There was not any noticeable change in the 2<sup>nd</sup> barrel with reference to the 1<sup>st</sup>. By the end of the simulation, the TG2 open conformation was transformed into a structure in which the two terminal barrels were straight with reference to each other and almost on a right angle with respect to the catalytic core and the N-terminal  $\beta$ -sandwich. A comparison between the structure at the start and the end of Run 2 is given in Figure 5-13, showing a clearer representation of the bending.



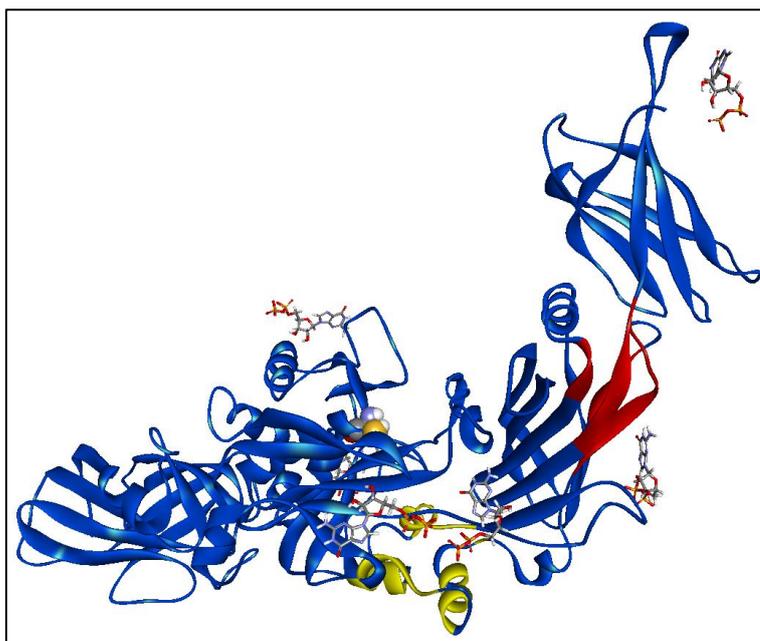
**Figure 5-13:** A comparison of TG2 structure at the end (blue and yellow) and start (red) of Run 2. The last two barrels are coloured yellow while the remainder of TG2 is blue in the “end” structure. CYS277 is shown as space filling to appreciate the direction of bending.

The RMSD of TG2 during Run 2 is given in Figure 5-14, and it can be used to explain the events that signified this run. In the first 100 ns there was a considerable oscillation that involved the 2<sup>nd</sup> barrel and resulted in high RMSD values for TG2 during this time. This was followed by a rather stable run for the next 400 ns. The bending event that resulted in the L-shaped TG2 structure was not very clear on the RMSD graph, perhaps because it happened very smoothly. This was the rationale for measuring the RMSD for the terminal barrels only, which is also presented in the same graph. This graph more clearly shows the bending event which started at around 530 ns and continued until the simulation was stopped.



**Figure 5-14:** RMSD for TG2 in Run 2, compared to that of the 2 terminal  $\beta$ -barrels during the same run.

With regard to the behaviour of GDP copies, it was noted that 18 out of the 24 copies wandered around TG2 structure during the entire simulation, and that only six copies managed to settle themselves within different distinct regions of TG2 structure and remain in their spots for relatively long durations of time. No important positions were taken by the 6 copies; the predicted allosteric site and the actual active site of TG2 did not attract any copy. The exception, however, was the original GDP binding site, to which one of the 24 copies of GDP was attracted (Figure 5-15).



**Figure 5-15:** The six GDP copies (in the stick form) that settled within distinct TG2 locations in Run 2. The predicted and original allosteric sites are shown in yellow and red respectively for reference, in addition to CYS277 (space filling).

The copy of GDP within the GDP binding site started to settle in the area at around 400 ns, and since this was the only important event related to the movement of the copies, it can be assumed that the event ultimately led to the bending of TG2 structure from the middle. There were 14 TG2 residues lying within 5 Å of the GDP copy at the end of Run 2. Six of them were identical to those lying within the same distance in the closed TG2 conformation in 1KV3 crystal structure (Liu et al. 2002).

In addition to the induced bending, the GDP copy in Run 2 showed some of the characteristic interactions holding GDP in place in the 1KV3 crystal structure (Liu et al. 2002), namely hydrogen bonds and electrostatic interactions between the phosphate groups of GDP and the positively charged side chains of ARG476 and ARG478 (Figure 5-16). An analysis of the hydrogen bonds in the region of the trajectory of Run 2 when GDP copy settled itself within the GDP binding site (from 400-635 ns of simulation time) showed very frequent hydrogen bonds with ARG476 and ARG478 (51% and 53% respectively of the analysed 235 ns).

The TG2-bending event seen in Run 2, which could be a reasonable predecessor for the closure of TG2 conformation, and the pattern of interactions seen in the run can give an indication that the copy of GDP has managed a plausible binding within TG2 to induce a conformational change, and that the aMD run was able to simulate the induced conformational change in TG2 to some extent.

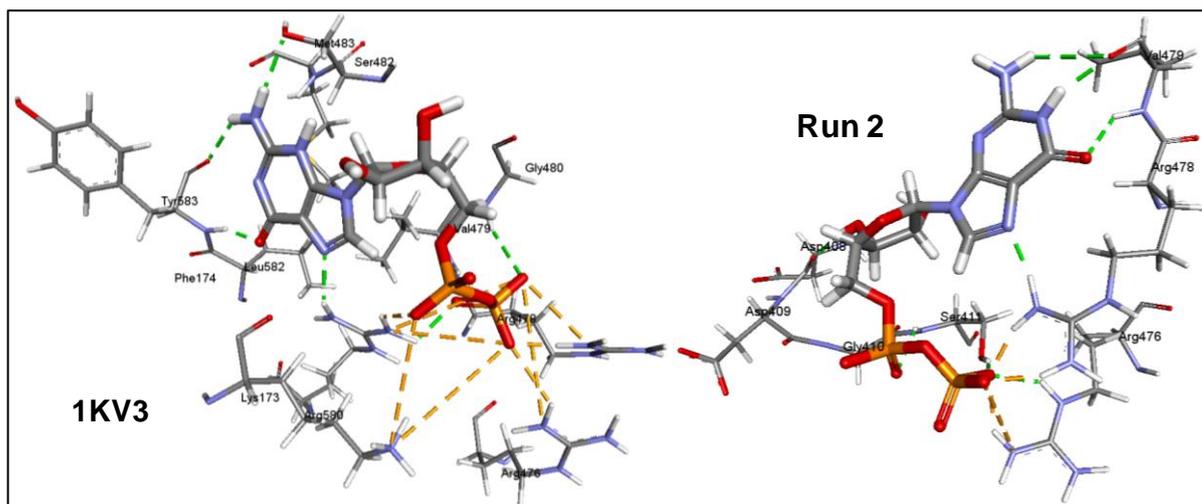


Figure 5-16: Interactions between GDP and TG2 residues, in 1KV3 crystal structure (left) and in the last frame in Run 2 (right), shown as dotted lines. Hydrogen bonds are green while electrostatic interactions are orange.

### 5.3.3.3 aMD on compound i2

#### 5.3.3.3.1 aMD on i2 docking complex (Run 3)

During the aMD run applied on the docking complex for compound **i2** (Figure 5-7), the compound moved outside its starting position after about 100 ns of simulation time only. The compound did not leave the predicted pocket entirely but was not as deep as it was when the simulation started. This was confirmed by measuring the RMSD values for the compound on its own using the 1<sup>st</sup> frame as a reference (TG2 was aligned against the 1<sup>st</sup> frame) (Figure 5-17), which showed that there was a considerable movement of the compound at around 100 ns. The simulation was discontinued after 395 ns because of this motion of **i2**. A pose comparison for the compound at the start and end of the simulation within the predicted site is presented in Figure 5-18.

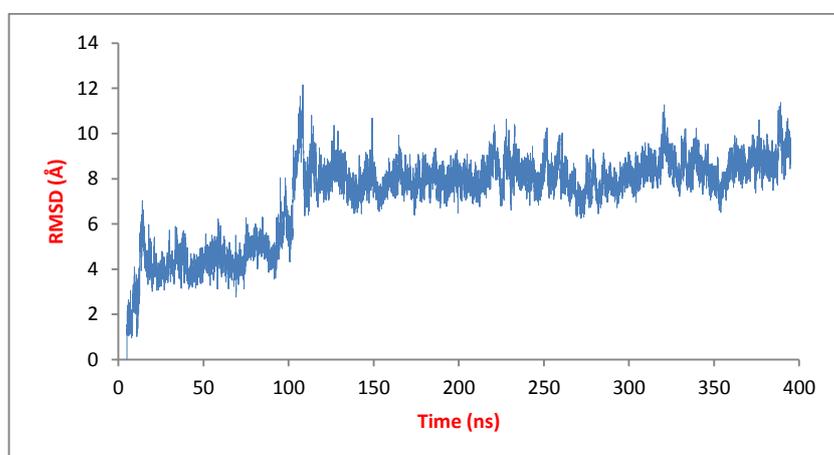
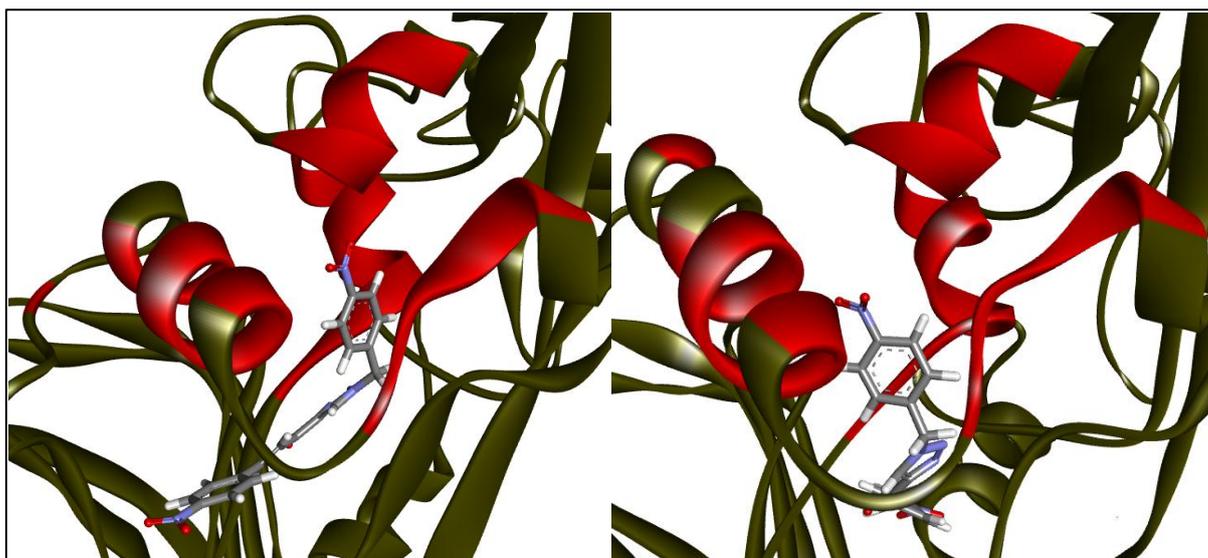
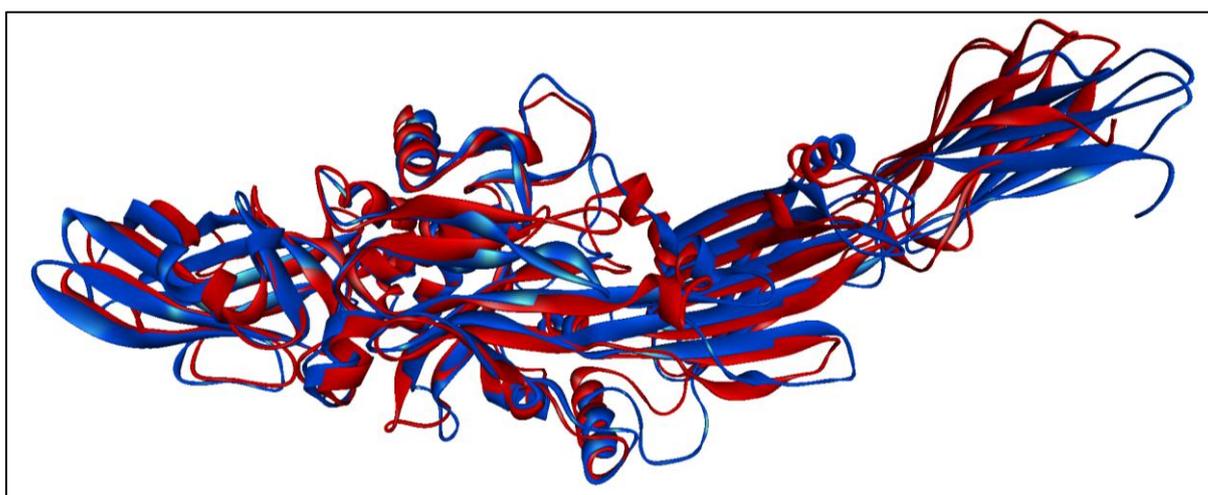


Figure 5-17: RMSD values for compound **i2** during Run 3.



**Figure 5-18:** A comparison between the pose of compound i2 at the start of Run 3 (left) and the pose after 395 ns of simulation time (right). TG2 is dark green and the predicted allosteric site is red.

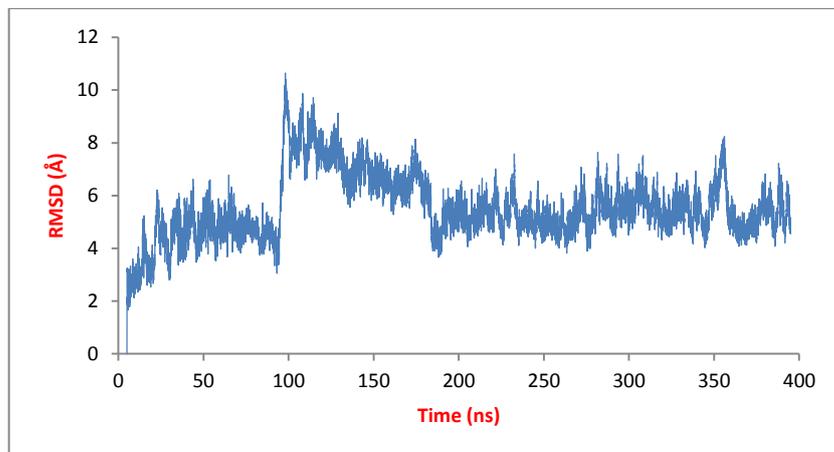
When TG2 itself was examined, it was observed that the enzyme maintained its open conformation without any significant change. This also encouraged the discontinuation of the simulation. A comparison between the starting and the final conformations of TG2 in this simulation is shown in Figure 5-19. When the RMSD of TG2 was calculated, it gave a similar insight in that the protein was relatively stable (Figure 5-20). There appears to be, however, a slight shift in the 2<sup>nd</sup>  $\beta$ -barrel. This was verified by a calculation of the atomic fluctuations (root-mean-squared-fluctuations, RMSF) of the individual residues (Figure 5-21) which proved greater flexibility in the final  $\approx 100$  residues of TG2 (2<sup>nd</sup>  $\beta$ -barrel).



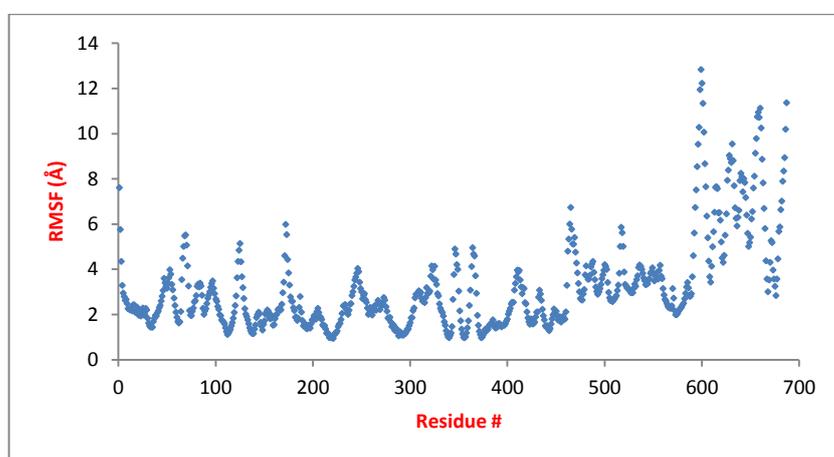
**Figure 5-19:** A comparison between the structure of TG2 at the start (blue) and the end (red) of Run 3.

Visual inspection of the trajectory with the VMD program showed that this movement was in the form of an oscillation affecting the 2<sup>nd</sup>  $\beta$ -barrel, where it moved to the inside of TG2 and then out again; sometimes very quickly between successive frames, which were 25 ps apart, and sometimes over extended periods such as the shift in the RMSD that happened around

the 100<sup>th</sup> ns. Because no permanent change in TG2 structure was observed, it may be assumed that **i2**, in the predicted site, does not have an influence on TG2 structure.



**Figure 5-20: RMSD for TG2 during Run 3.**



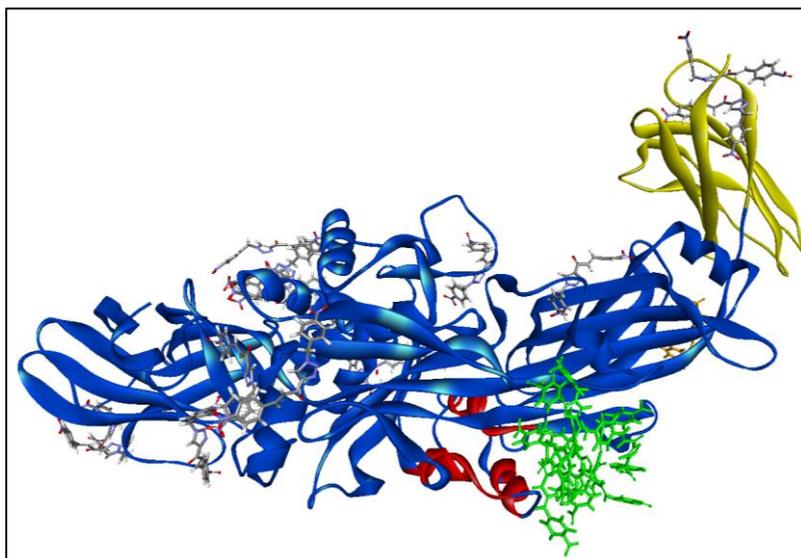
**Figure 5-21: Atomic fluctuation scores for TG2 residues in Run 3.**

### 5.3.3.3.2 aMD on TG2 with 24 copies of **i2** (Run 4)

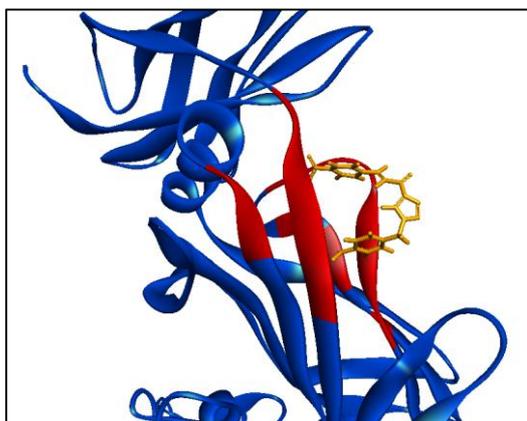
The rationale behind running this simulation was to examine the potential of the predicted allosteric site to be recognised by an allosteric inhibitor that is not in its immediate vicinity. Run 4 was allowed to continue for a total of 320 ns, at the end of which none of the copies was able to find its way into the predicted binding site, but 7 copies managed to locate themselves in the surrounding area of the site. It was also noticed that the terminal  $\beta$ -barrel was facing inwards, in the same direction of the folding in the closed conformation of TG2 (Figure 5-22).

The remaining copies were distributing themselves around the surface of the protein during the simulation, and none achieved a pose in which it was embedded within any region of the protein. One copy, however, (orange, Figure 5-22) was located in the section of the GDP binding site that is on the 1<sup>st</sup>  $\beta$ -barrel (Figure 5-23). This started to happen early in the course of the aMD simulation, where the copy started to move into the vicinity of the GDP binding site at around the 1<sup>st</sup> nanosecond of the production phase (6<sup>th</sup> nanosecond of

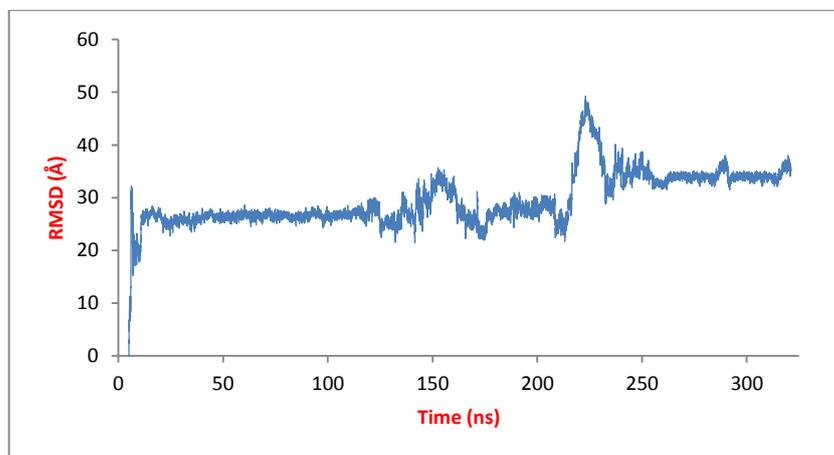
simulation time). The copy stayed in the vicinity for about 120 ns and then started to move in and out of GDP binding site for about 30 ns, at the end of which the copy settled within the pose shown in Figure 5-23. It is at this time that the 2<sup>nd</sup>  $\beta$ -barrel of TG2 started to fold toward the catalytic core. These events are better appreciated by inspecting the RMSD graph for this **i2** copy (Figure 5-24).



**Figure 5-22:** TG2 with the 24 copies of **i2** at the end of Run 4. The 7 copies that came close to the predicted site are coloured green, the predicted site is red and the 2<sup>nd</sup>  $\beta$ -barrel is coloured yellow. **i2** copy residing in GDP binding site is coloured orange.



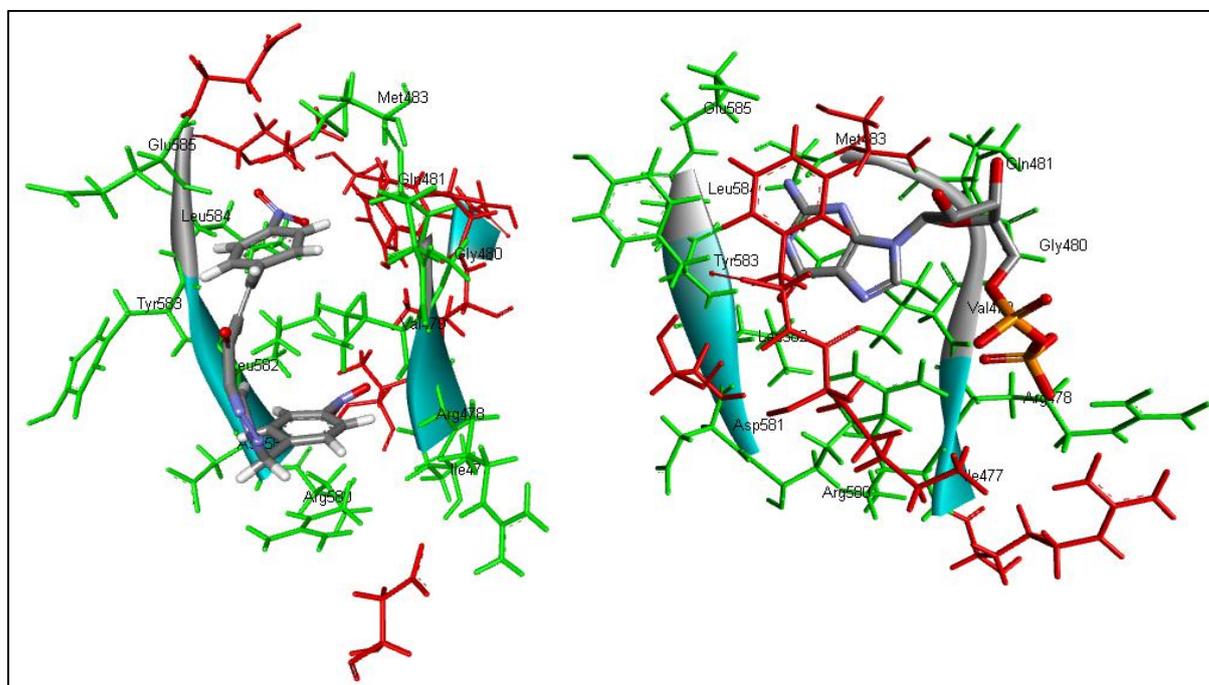
**Figure 5-23:** Another angle for TG2 from Figure 5-22 showing **i2** copy (orange) on the GDP binding site within the 1<sup>st</sup> barrel (coloured red).



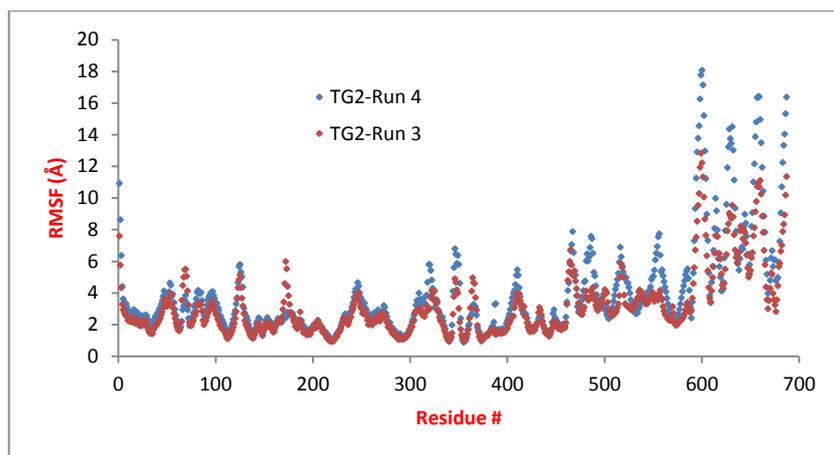
**Figure 5-24:** RMSD graph for **i2** copy that positioned itself in the GDP binding site during Run 4.

The shift in RMSD that occurred at around 225 ns (Figure 5-24) was caused by the copy of **i2** leaving the GDP binding site, but it went back in and stayed there for the rest of the simulation, where it was adopting the pose seen in Figure 5-23. The bending in TG2 structure was not affected by this motion and did not change once it was there.

To further confirm that the **i2** copy was residing in GDP binding site, DS Visualizer 4.0 was used to inspect the residues that are within 5 Å of GDP in 1KV3 crystal structure (Liu et al. 2002), and those that are within 5 Å of **i2** in the last frame of Run 4. There are 16 residues within 5 Å of GDP in 1KV3, 14 of them are located on the 1<sup>st</sup> barrel and 2 are on the catalytic core. There were 20 residues within 5 Å of **i2** copy in the last frame of Run 4, 12 of which are identical to those that surround GDP in 1KV3, confirming that **i2** was actually within a site that very much resembles that of GDP binding (Figure 5-25). When the atomic fluctuations of TG2 residues were measured, the residues for the 2<sup>nd</sup>  $\beta$ -barrel were found to have moved more when compared to the remaining residues in TG2 and when compared to the same region during the simulation applied on **i2** docking complex (Run 3) (Figure 5-26). This is consistent with the 2<sup>nd</sup> barrel moving towards the inside of TG2 and attempting to produce the closed inactive conformation of the enzyme. Although the end result was not a fully closed conformation as that in 1KV3, it was a move in the right direction.



**Figure 5-25:** Left: Residues within 5 Å of **i2** copy at the end of Run 4. Right: Residues within 5 Å of GDP in 1KV3 crystal structure. **i2** and GDP are coloured by element and TG2 residues are shown as stick and coloured green if they are common and red if they are not.



**Figure 5-26: Atomic fluctuation values for TG2 residues during runs 3 and 4.**

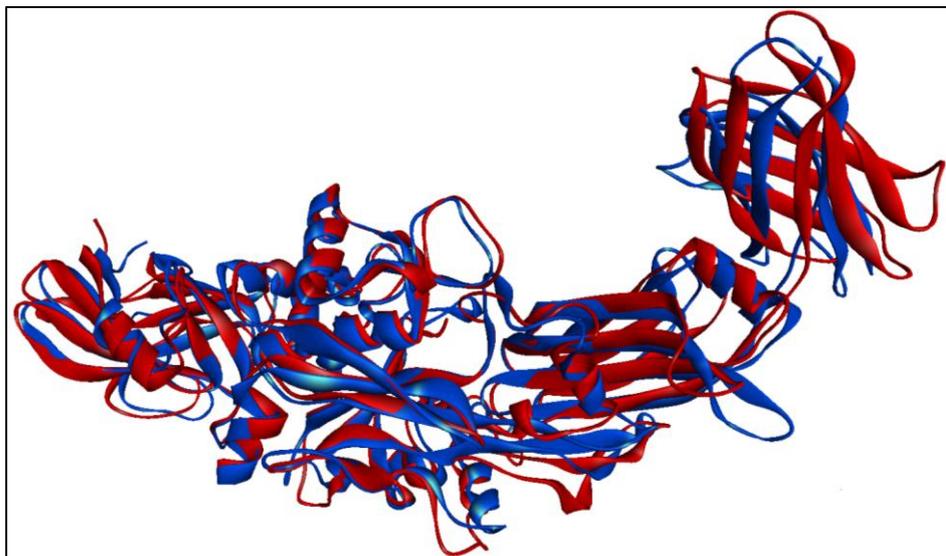
The results from the 1<sup>st</sup> simulation applied on the **i2** docking complex (Run 3) indicate that compound **i2** may not be a suitable binder in the predicted allosteric site or that the predicted site is not a true allosteric binding site. The results from the 2<sup>nd</sup> simulation give an indication that **i2** has the potential to reside within the binding site of GDP and by doing this, the compound triggered a conformational change in TG2 that, to some extent, resembles what is induced by the binding of GTP or GDP. This was the inspiration that motivated the application of an aMD simulation starting from the last frame of Run 4 after removing all **i2** copies with the exception of the copy located within the GDP binding site (next section).

#### 5.3.3.3.3 aMD on TG2 with 1 copy of **i2**

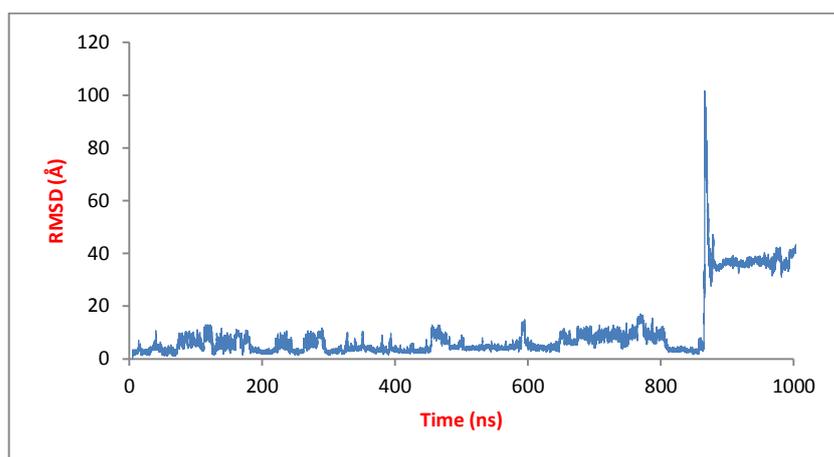
**i2** copies from the previous simulation (Run 4), with the exception of the copy in the GDP binding site, were deleted for 2 reasons; the first was to reduce the computational cost by reducing the size of the system and the associated water box, and the second was to examine whether the TG2 bending was actually triggered by the binding of **i2** in this site and not from the presence of another copy elsewhere. To further confirm that the bending was triggered by the binding of **i2** copy in the GDP binding site, another aMD simulation was started which was identical to the one described here with the deletion of all the **i2** copies including the one in the GDP binding site.

The simulation with **i2** copy (Run 5) was allowed to continue for 1,000 ns (1  $\mu$ s), and at its end, the conformation of TG2 did not change much from that at the end of the previous run (Run 4) (Figure 5-27). The copy of **i2** that started in the GDP binding site did not stay there for the entire simulation. **i2** molecule left the binding site after about 850 ns and was wandering around the surface of TG2 until it settled itself at some point on the surface in the middle of TG2 molecule. This event of **i2** leaving the GDP binding site did not affect the conformation of TG2 in terms of the bending. The RMSD of **i2** molecule on its own is shown in Figure 5-28, in which it can be clearly seen that there was a major motion involving **i2**

molecule. This motion was confirmed to be involving the exit of **i2** molecule from the GDP binding site by visual inspection of the trajectory with VMD.



**Figure 5-27:** TG2 structure at the end of Run 5 (red) compared to that at the end of Run 4 (blue).

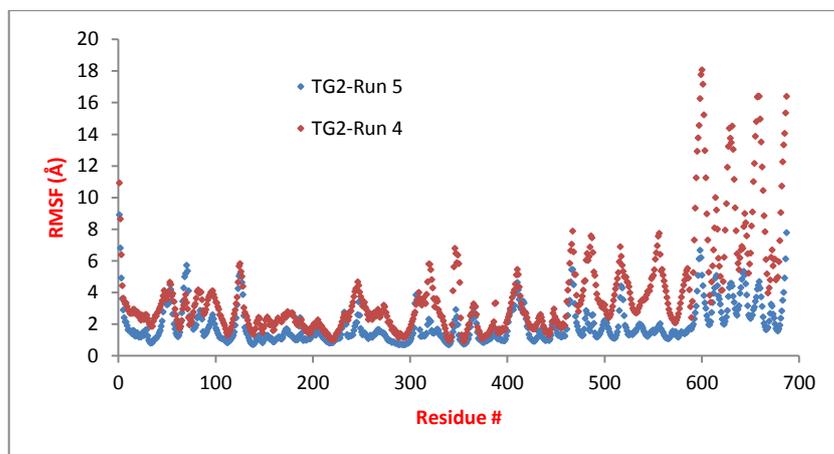


**Figure 5-28:** RMSD for **i2** during Run 5.

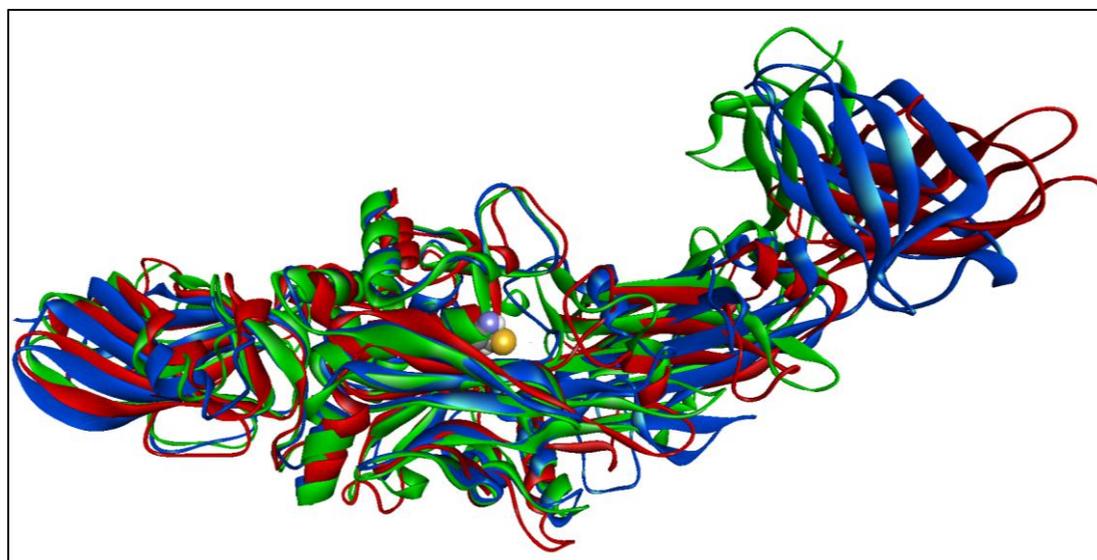
When the atomic fluctuations (RMSF) values for TG2 residues during this run were measured, the graph in Figure 5-29 was produced. The movement of the terminal barrel was much lower in Run 5 when compared to Run 4. This indicates that this barrel did not move much in this aMD simulation, and that the bent shape of TG2 produced during Run 4 was maintained during this run.

Another aMD run was applied on the same starting structure used for Run 5 (the last frame from the simulation on TG2 with 24 copies of **i2**, with only one **i2** copy which is the one positioned in the GDP binding site) without the **i2** copy (Run 6) and was also continued for 1,000 ns. The bending of TG2 at the junction between the 1<sup>st</sup> and the 2<sup>nd</sup>  $\beta$ -barrels changed during this simulation; instead of having the barrel facing towards the catalytic core and CYS277, the direction of bending was towards the outside of TG2. This was similar to the

orientation of the section involving the 2<sup>nd</sup>  $\beta$ -barrel seen during Run 1. A comparison in the direction of bending for this run with that of Run 1 and Run 4 is presented in Figure 5-30.



**Figure 5-29:** Atomic fluctuation values for TG2 residues in Run 5 compared to those from Run 2.



**Figure 5-30:** Conformation at the end of Run 6 (red) compared to that at the end of Run 4 (green), and the one at the end of Run 1 (blue), showing CYS277 to appreciate the direction of bending of the 2<sup>nd</sup>  $\beta$ -barrel.

This conformational change started to happen at around the 350<sup>th</sup> nanosecond, and the new form of TG2 was maintained for the remainder of the aMD simulation. The conformational change in TG2 observed during Run 6 may be better appreciated by examining the RMSD graph for the run. The graph is presented in Figure 5-31. RMSD for Run 5 is also presented. In Run 5, TG2 structure was mostly stable as the bent conformation of TG2 was preserved throughout the simulation. This has been confirmed by measuring the atomic fluctuations for the individual residues (Figure 5-29). For Run 6, the time and magnitude of the conformational change can be clearly seen in Figure 5-31. To confirm that the change in TG2 conformation mainly affected the 2<sup>nd</sup>  $\beta$ -barrel, atomic fluctuations for Run 6 were calculated. RMSF values are presented in Figure 5-32, which show that the residues of the terminal region of TG2 in Run 6 moved more than their counterparts in Run 5, and more than those in the remainder of TG2.

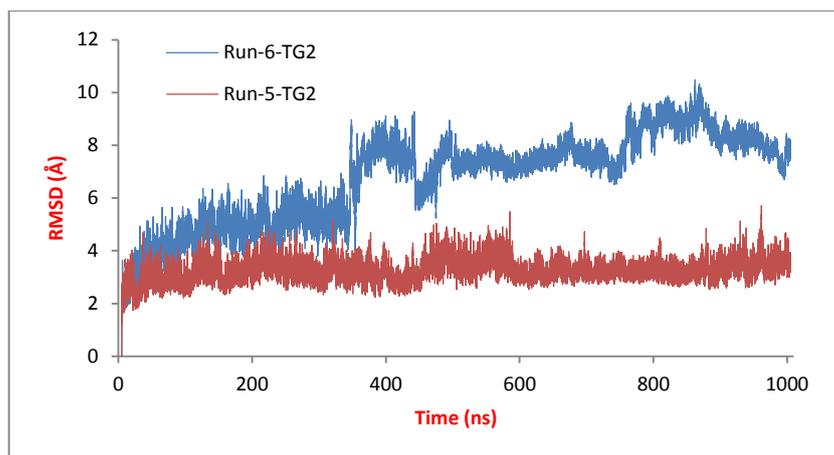


Figure 5-31: RMSD for Run 6, compared to that of Run 5.

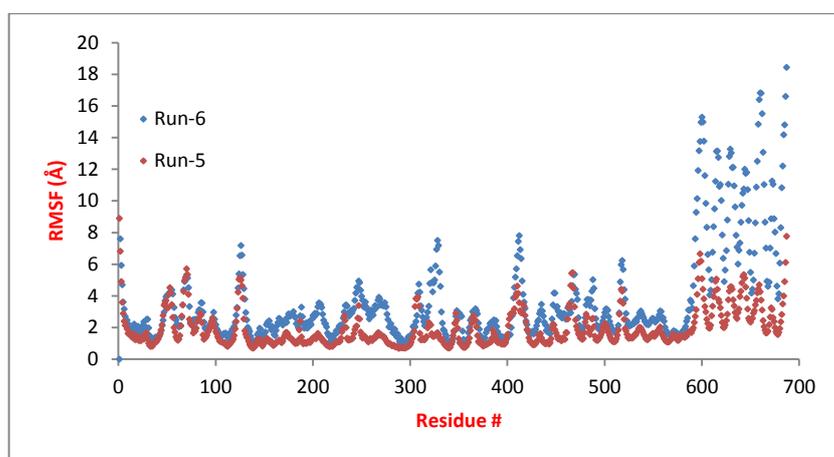


Figure 5-32: Atomic fluctuations for TG2 residues in Run 6 compared to those in Run 5.

Although the final result of Run 6 did not have TG2 in the open conformation, the change in the direction of the bending of the 2<sup>nd</sup>  $\beta$ -barrel towards the outside of the catalytic core and the persistence of this change throughout the simulation indicates that it was no longer possible for TG2 to be folded into the correct closed form. The same bending orientation has been noticed when aMD was applied on empty TG2 in Run 1, possibly indicating that it is a possible orientation active TG2 can adopt, but that cannot progress into a fully closed inactive structure.

The results from the 4 aMD simulations applied with **i2** can be used to draw some observations. No change in the conformation of TG2 was observed after Run 3, in which **i2** was docked into the predicted allosteric site. This, combined with the fact that the **i2** molecule did not maintain its pose within the predicted site, may indicate that the predicted site is not a suitable binding site for **i2**, or that **i2** is not an allosteric inhibitor. The latter is actually consistent with the research in which **i2** was first mentioned (Pardin, Roy, et al. 2008), where the compound was presented as a reversible competitive inhibitor of TG2. The aMD simulation with the 24 copies of **i2** (Run 4) showed something different. There was one copy that managed to position itself within the section of the GDP binding site that is located on the 2<sup>nd</sup>  $\beta$ -barrel of TG2. This positioning of the **i2** copy was associated with a change in

the general conformation of TG2, manifested as the bending of the terminal barrel towards the catalytic core of TG2.

Run **5** was performed to test whether the pose of **i2** within the GDP binding site could induce further bending in TG2 to result ultimately in closure of the enzyme conformation after removing the other 23 copies of **i2**. The simulation finished with the same general conformation of TG2 with which it started; no further bending was observed. Furthermore, the **i2** molecule left the GDP binding site, although this did not take place until late in the simulation. Run **6**, which was identical to Run **5** minus the **i2** copy, resulted in a change in the direction of bending of the terminal barrel; in Run **5** the possibility of complete closure of TG2 conformation was still there, while in Run **6**, further bending in the same direction would have resulted in a totally unrealistic enzyme structure. Since the only different parameter between Run **5** and Run **6** was the presence of the **i2** copy, it could be said that **i2** has the potential to induce large conformational change in TG2 structure, similar to that produced by GTP and/or GDP by binding to a similar site.

#### **5.3.3.4 aMD on compound i3**

##### **5.3.3.4.1 aMD on i3 docking complex (Run 7)**

This aMD run started with compound **i3** docked into the predicted allosteric binding site and the run was continued for a total of 1,000 ns of simulation time. The **i3** pose within the predicted site did not change much during this simulation; although it appeared to have been more deeply embedded in the site at the start of the simulation than it was at its end (Figure 5-33). The RMSD graph for **i3** in Run **7** is shown in Figure 5-34. In the same figure, there is a comparison with the RMSD for compound **i2** during Run **3**, and it can be observed that **i3** was more stable in its simulation than was **i2**. This indicates that **i3** is a more suitable binder in the predicted allosteric site than **i2**.

With regard to TG2 behaviour in Run **7**, it was detected that there was a bending in the enzyme structure, also involving the hinge between the 1<sup>st</sup> and 2<sup>nd</sup>  $\beta$ -barrels. It was similar to the bending observed during aMD on TG2 with 24 copies of **i2** (Run **4**). The barrel was bending towards the inside of TG2 to close on the catalytic core and CYS277. Figure 5-35 presents the structure of TG2 at the end of Run **7**, superimposed on TG2 structure at the start of the run, and it is clear from the figure that there was considerable bending in TG2 structure. The bending event started to appear in TG2 after about 320 ns of the simulation time, and persisted in the same form for the remainder of the simulation time. This is confirmed by inspecting the RMSD graph for TG2 during Run **7** (Figure 5-36), which

demonstrates that there was an increase followed by stabilisation in RMSD after 320 ns of simulation time.

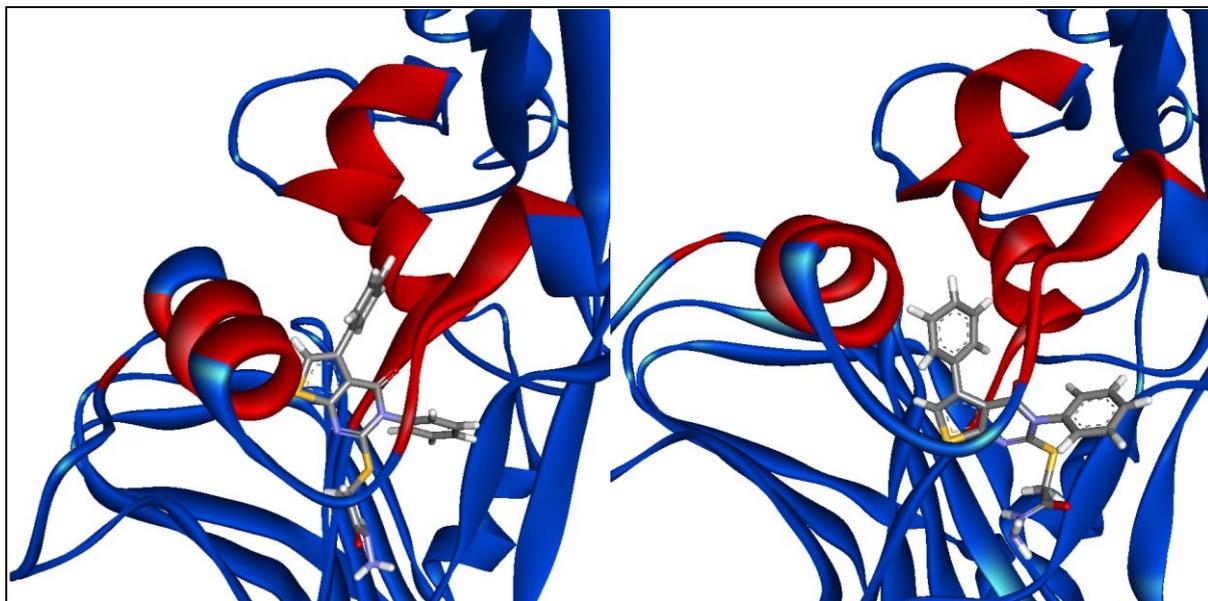


Figure 5-33: A comparison of the pose of i3 within the predicted site (red) between the docking complex (left) and the conformation at the end of the Run 7 (right).

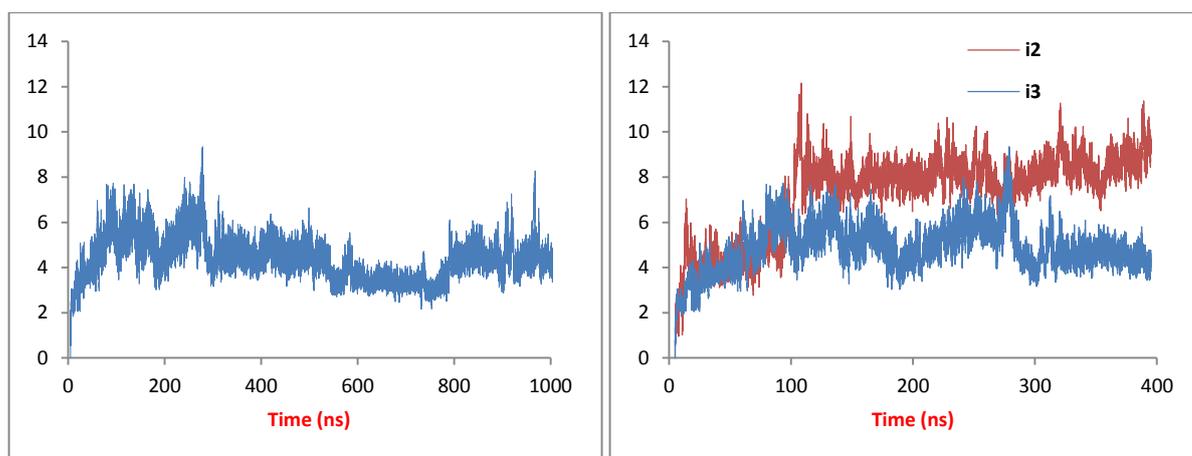


Figure 5-34: RMSD graphs, the Y-axis represents RMSD values in Å. On the left is RMSD of i3 during Run 7 and on the right is the RMSD of i3 in Run 7 compared to that of i2 in Run 3 for the same length of the simulation time.

To further confirm that the bending event was the major general motion of TG2 during Run 7, the atomic fluctuations of the individual residues of TG2 were recorded. The results are presented in Figure 5-37 where the RMSF values are shown in comparison with those from Run 4. The relatively large motion of terminal barrel during the simulation is obvious and can be easily be compared to that observed in Run 4.

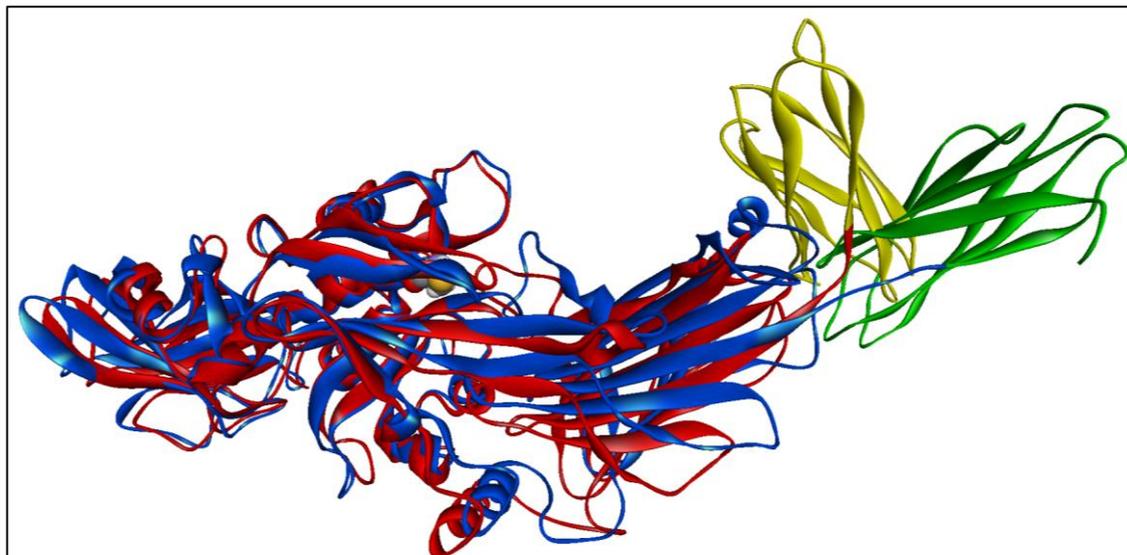


Figure 5-35: TG2 structure at the start of Run 7 (blue) and after 1  $\mu$ s (red), showing the bending of the terminal barrel (green for start and yellow for the end of the simulation), and showing CYS277 to appreciate the direction of bending.

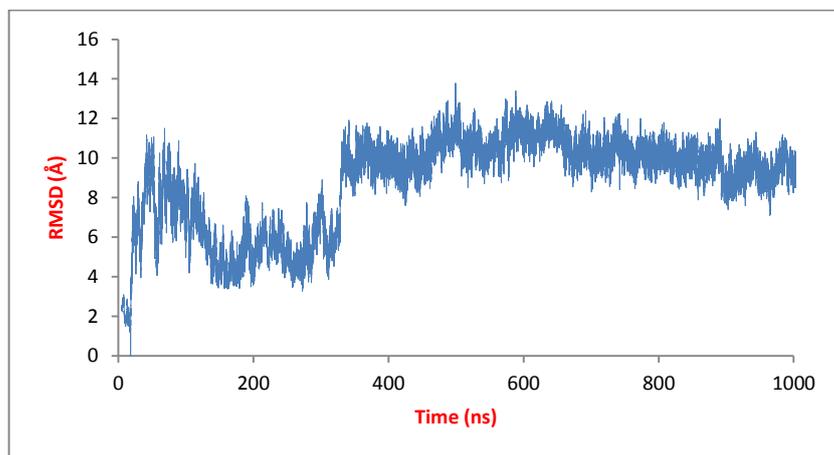


Figure 5-36: RMSD graph for TG2 residues in Run 7.

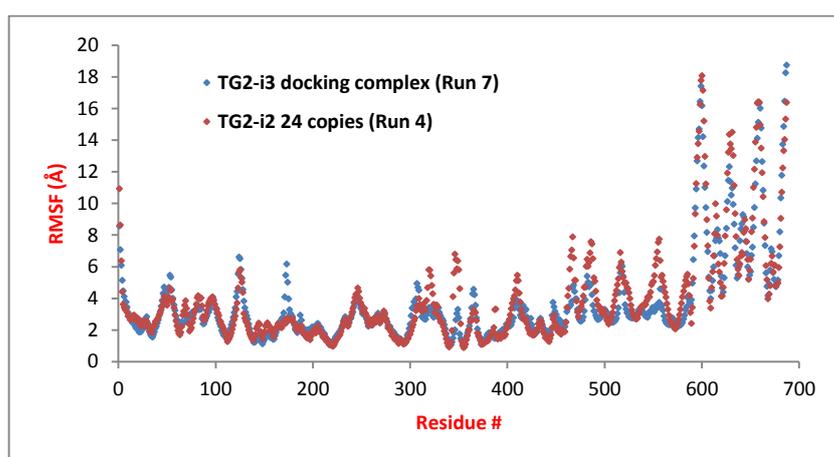
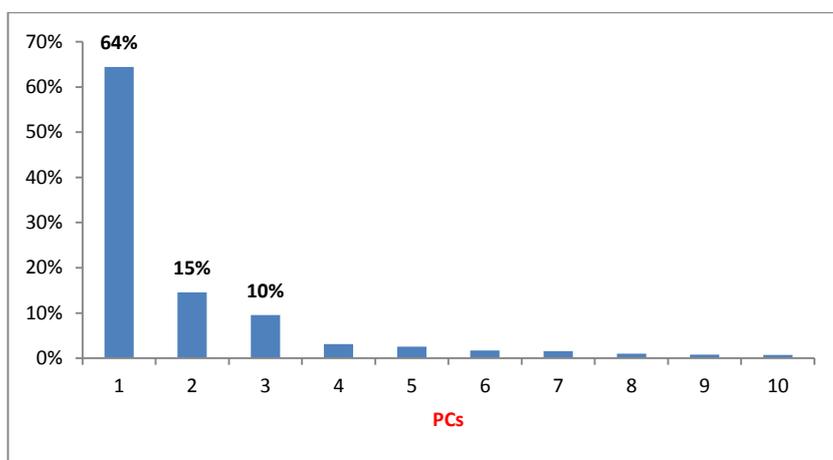


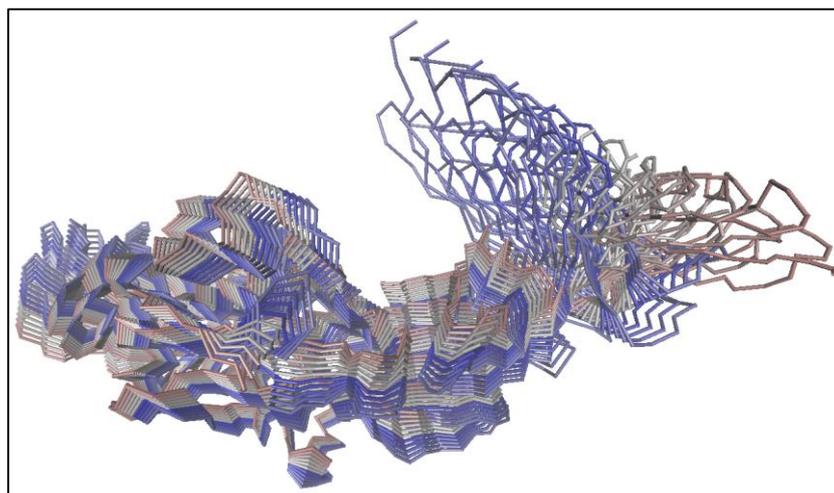
Figure 5-37: Atomic fluctuations for TG2 residues during Run 7, compared to those of Run 4.

Additional confirmation for the bending event was performed using PCA. pyPcazip was able to extract 10 PCs from this aMD simulation representing the most important modes of motion observed in the trajectory. These PCs are presented in Figure 5-38, which shows the percentage contribution of the major 3 PCs to the total motion in TG2 structure. The first PC

accounts for about two thirds of the motion performed by TG2 during Run 7 aMD simulation, and the second major PC contributes only  $\approx 15\%$ . The animation produced by pyPcazip for the first and most important PC of Run 7 is presented in Figure 5-39 which shows that the PC mainly involved the terminal barrel of TG2. When the frames are coloured by time step, the result is an open conformation of TG2 at the beginning of the simulation and a bent conformation involving the hinge connecting the 2<sup>nd</sup>  $\beta$ -barrel to the rest of TG2 at the end. The bending is the most important component of this PC since, as can be seen from the figure, the rest of the protein is mostly stable and have the same structure at the start and the end of the simulation.



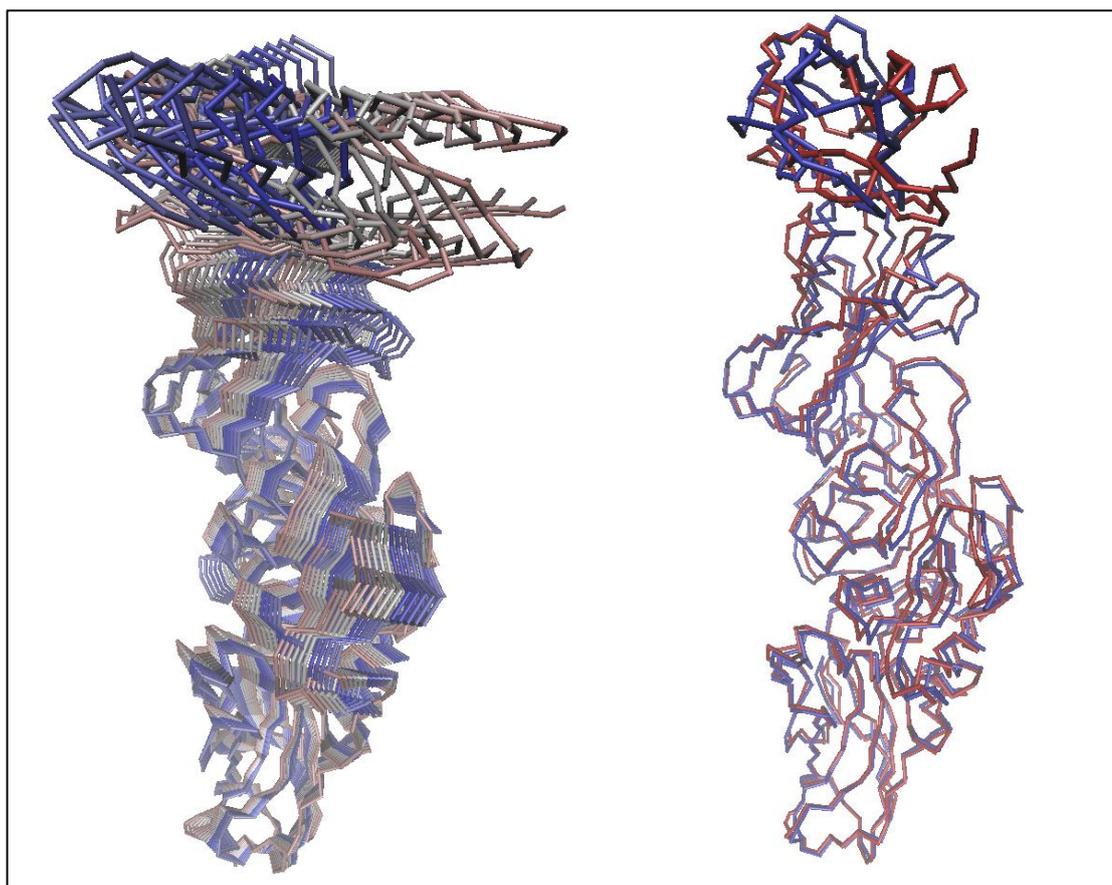
**Figure 5-38: PCs produced for Run 7 and the contribution of each to the total motion of TG2 during the run.**



**Figure 5-39: The animation produced by pyPcazip for the first PC of Run 7, showing the 20 frames, and the frames are coloured by time step, where red represent the start and blue the end of the simulation.**

Figure 5-39 clearly shows that the motion that accounted for 64% of total TG2 motion during Run 7 involved the terminal  $\beta$ -barrel and was in the direction that resulted in the bending of the TG2 in a manner that could result in the closure and subsequent inactivation of the enzyme. It should be noted, however, that there was a considerable oscillation involving the barrel and it is the frame with the darkest blue colour that represents the final form of TG2 in this simulation. This means that the bending did not go all the way in the fashion presented in Figure 5-39, rather to the frame with the darkest blue colour.

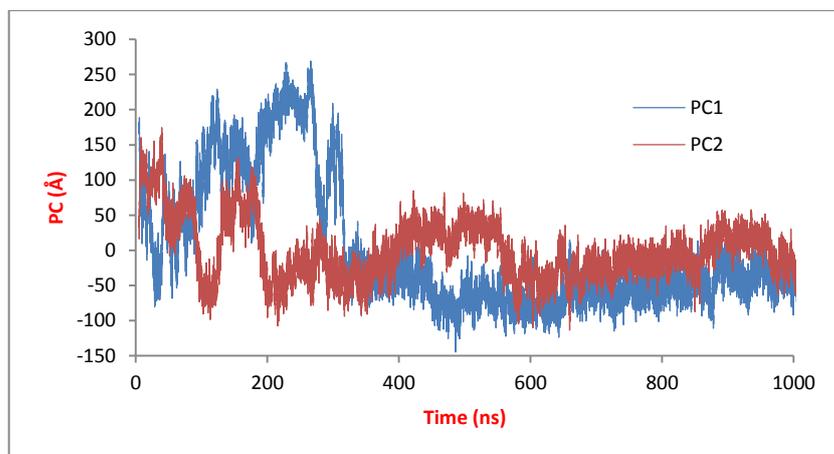
When the second PC was considered, it was noted that the motion represented by this PC also involved the terminal  $\beta$ -barrel. The difference was in the direction and intensity, where with the 2<sup>nd</sup> PC, the motion was in the form of oscillation of the terminal barrel to the sides of TG2 rather than towards the catalytic core of the enzyme. A visual inspection of the associated animation (shown in Figure 5-40) gives a clearer understanding of the mode of motion of TG2 terminal barrel presented by this PC; most importantly, its direction and the locations of the starting and end frames which were essentially in the same position at the middle of TG2 with reference to the direction of motion within this PC. This confirms that the 2<sup>nd</sup> PC involved only oscillation affecting the terminal  $\beta$ -barrel.



**Figure 5-40:** The animation produced by pyPcazip for the second PC of Run 7, showing the 20 frames (left) coloured by time step. The first and final frames are shown to the right.

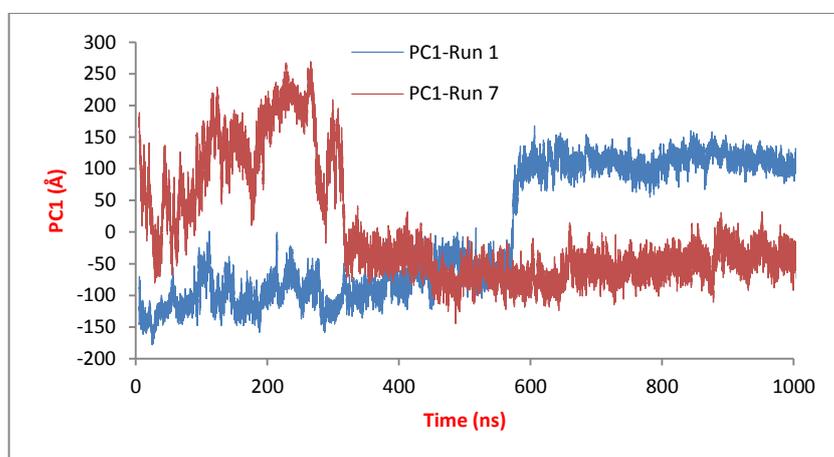
The analysis of the animations from the top two PCs was followed by investigating the projections of each PC into the general subspace as a function of time. The graph in Figure 5-41 was produced from which two important observations can be drawn. The first is the magnitude of PC1 when compared to PC2, especially in the first 300 ns of simulation time, from which the percentage contributions of the two PCs in Figure 5-38 can be better appreciated. The second observation is the time during the simulation at which PC1 started to stabilise, which is roughly the same time at which the bending event started to happen as shown in the RMSD for TG2 in this run (Figure 5-36). The latter observation confirms that

the bending event was actually presented by PC1 and it corresponds to the most important motion of TG2 during Run 7.



**Figure 5-41:** The projections of the top 2 PCs from Run 7 into the general subspace as a function of simulation time.

To examine the difference between the bending events observed in this run and in Run 1 (empty TG2), a comparison was made between the first and most important PC from the 2 runs. The result is displayed in Figure 5-42, which shows that the motion in Run 7 was more evident. The graph additionally shows that the bending in Run 7 lead to a stabilisation of the associated PC while in Run 1, the bending was associated with increased motion in the terminal barrel. This further confirms that the bending observed in Run 1 was not going to result in any further closure of TG2 conformation.



**Figure 5-42:** PC1 as a function of time for Run 1 and Run 7.

#### 5.3.3.4.2 aMD on TG2 with 24 copies of i3 (Run 8)

This simulation was identical to the aMD simulation in Run 4, except that the 24 copies were for compound **i3**, and the simulation was allowed to continue for 1,000 ns. There was no significant alteration in the general structure of TG2; the enzyme maintained its open, active conformation for the entire length of the aMD simulation, and this can be easily seen from the RMSD graph for the run which is presented in Figure 5-43.

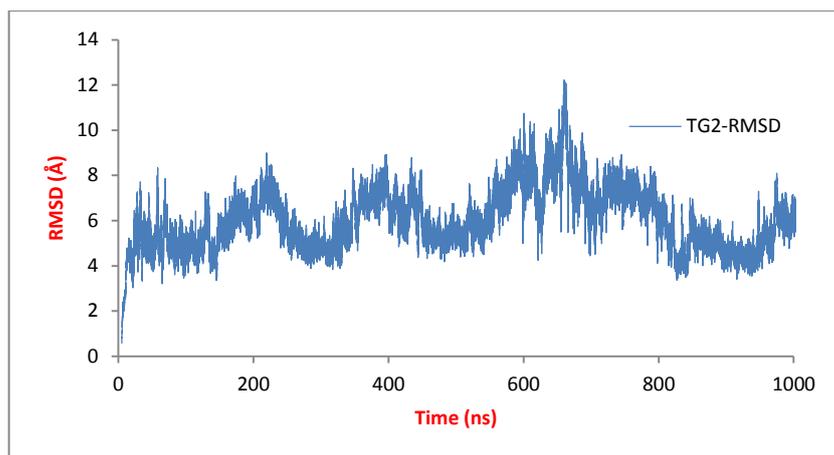


Figure 5-43: RMSD graph for TG2 during Run 8.

Although there was fluctuation in the RMSD graph, the values were generally ranging between 4 and 8 Å and averaged around 6 Å. A comparison of the TG2 general structure at the start and the end of this run is shown in Figure 5-44, and is compatible with the RMSD of the enzyme by stating that the run did not affect the general structure, specifically the conformation at the junction between the terminal  $\beta$ -barrel and the rest of TG2 structure.

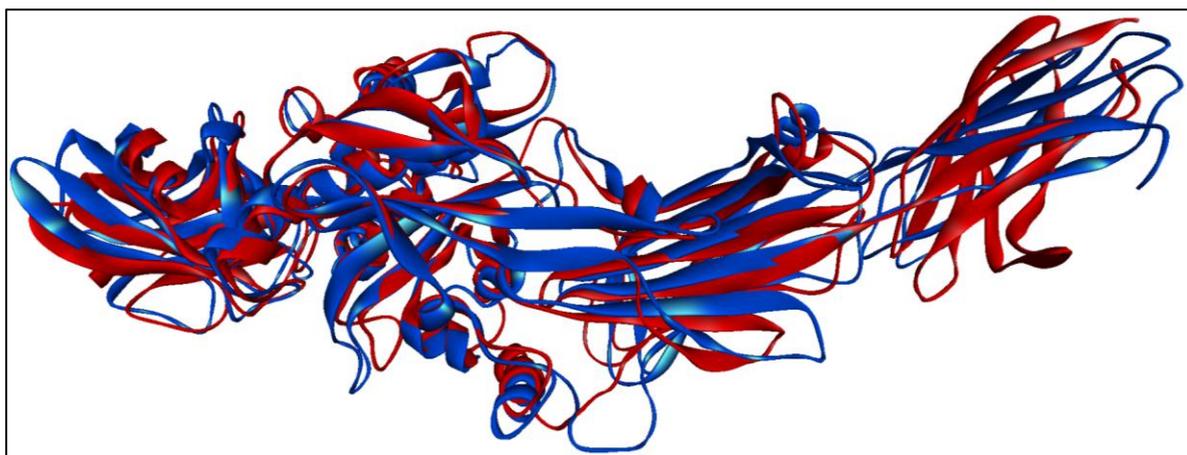


Figure 5-44: A comparison of TG2 structure at the start and the end of Run 8, showing the preservation of the open form of the enzyme.

Regarding the behaviour of the 24 copies of **i3**, no general pattern could be observed, and the motions were mostly random. In the final 200 ns of the simulation time, some copies were settling themselves within 2 major regions of TG2. Eight copies positioned themselves randomly within the N-terminal  $\beta$ -sandwich, distributing between the beta sheets of the sandwich. Six copies of **i3** were in the region of the catalytic core, specifically above the area between the catalytic tunnel and the hydrophobic loop of the active site of TG2. No copy, however, was close to CYS277, nor positioned in a manner similar to that of irreversible inhibitors (Figure 5-45). Neither the predicted allosteric site nor the GDP binding site were among the popular attractions for **i3** copies.

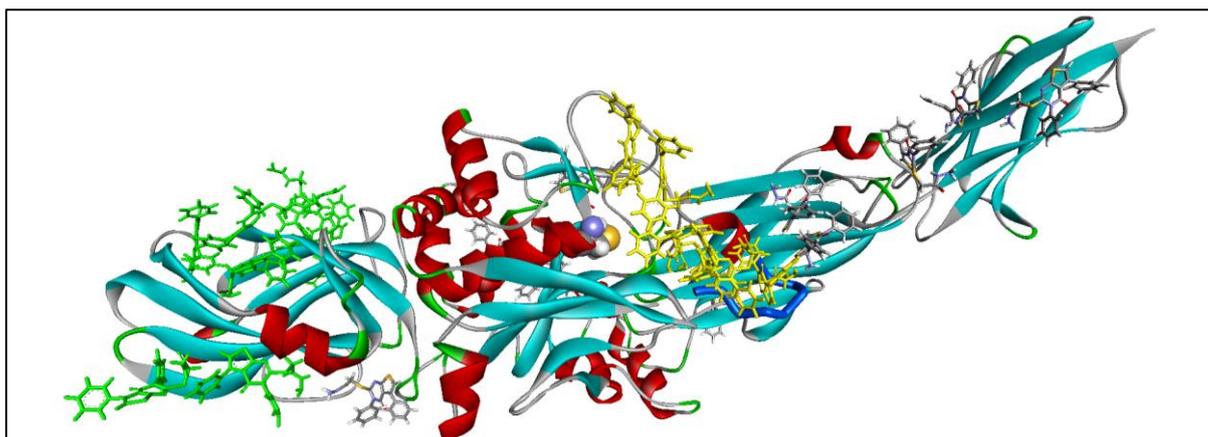


Figure 5-45: **i3** copies in the final frame in Run 8. Green copies are those within the N-terminal  $\beta$ -sandwich, and yellow copies are those close to TG2 active site. CYS277 is shown as space filling and the hydrophobic loop as blue tube for reference to the active site.

The results obtained from Run 7 (aMD on **i3** docking complex in the predicted site) can be used to draw the conclusion that **i3** has the potential to act as an allosteric inhibitor and to induce a conformational change in TG2, and that the predicted site could function as an alternative allosteric site within TG2. Run 8 (TG2 with 24 copies of **i3**) may not necessarily contradict Run 7, if it was explained on the basis of the nature of the predicted site; the site is not readily accessible and the approach of the ligand to the interior of the site may require an enhanced sampling technique. In addition, it may be argued that **i3**, unlike **i2**, cannot bind at the original allosteric site at which GDP binds.

The energy and the temperature change was followed during each of the discussed aMD simulations, and the results showed that all the simulations were stable with reference to these parameters. Example measurements of energy and temperature for Run 5 and Run 7 are presented in Figure 5-46.

## 5.4 Conclusions

During the work with the allosteric inhibitors, aMD was able to simulate some relatively large conformational changes in the structure of TG2 molecule. It was proven that these changes were induced by the binding of the allosteric inhibitors. The structure of the inhibitor had an effect on the preferred binding site, where **i2** did not produce any change in TG2 structure when it was bound at the predicted site, but triggered bending in the terminal  $\beta$ -barrel when it was bound at the section of the GDP binding site located in the terminal barrel after the compound was allowed to select a binding site during Run 4 (TG2 with 24 copies of **i2**). **i3** appeared to be a better binder within the predicted site in Run 7, where it induced a similar bending effect to that induced by **i2** during Run 4. GDP produced the best bending effect when compared to the other tested allosteric inhibitors, and the effect was induced by

binding at the original GDP binding site. This may mean that GDP has better potential of inducing the conformational change required for inactivating TG2 than **i2** or **i3**.

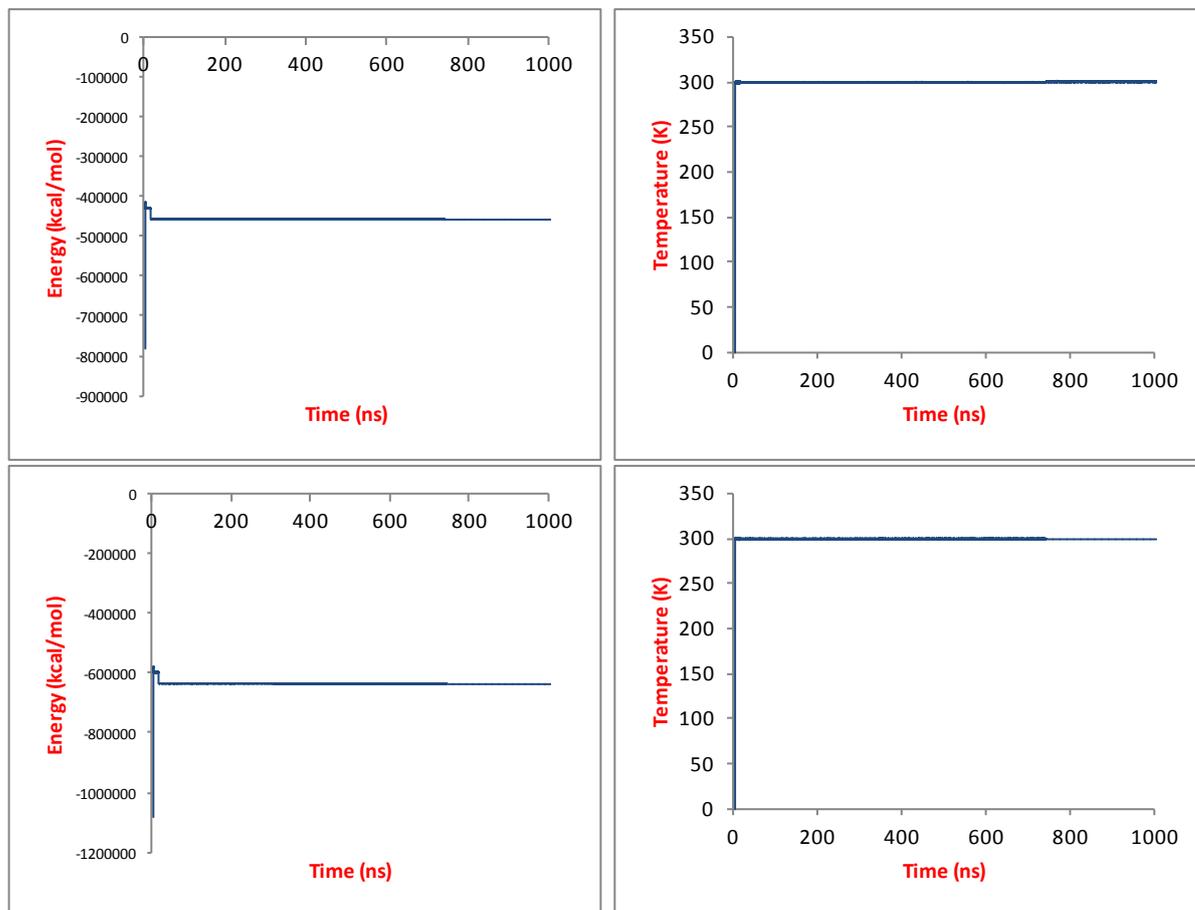


Figure 5-46: Energy and temperature during Run 5 (top) and Run 7 (bottom).

# **Chapter 6**

## **General Discussion and Conclusions**

## 6 General Discussion and Conclusions

In the first chapter (General Introduction), TG2 was thoroughly discussed in terms of its biological functions and its role in diseases, in addition to the available inhibitors. The facts that TG2 is not a vital enzyme and its possible role in a variety of pathological conditions present the enzyme as an attractive target for drug discovery efforts. To this end, the work that was presented in this thesis was performed with the aim of developing computational methods to help in the prediction of potential irreversible inhibitors for TG2. In the first part of the work, active site models of TG2 were developed and tested rigorously, and the same models were used in the second part of the work to study the mechanism of TG2 inhibition by irreversible inhibitors. The last part was independent of the previous parts and involved studying TG2 allostereism.

Molecular dynamics can be used to account for receptor flexibility during docking, either by supplying multiple structures of the protein or by relaxing a ligand docked structure (Hospital et al. 2015; B-Rao et al. 2009; Huang & Zou 2010). As shown in Chapter 3, MD was successful in improving the structure of TG2 at its active site to be able to achieve plausible docking complexes for a set of known inhibitors. This was particularly true for the active site models extracted from the MD simulations applied on the initial docking complexes for compounds **1a** and **1b**. The MD trajectories for the complexes themselves gave some useful information about the interactions between the compounds and TG2 active site, especially with regard to hydrogen bonding, where it has been shown that a hydrogen bond with ASN333 is essential for the maintenance of a good pose for an active inhibitor (bent conformation) within the active site of TG2.

The active site models extracted from MD trajectories were tested using the 6 most active TG2 inhibitors from the work by Badarau et al. (2015) (section 3.3). The criteria selected at this stage to define an appropriate docking complex were sufficient to judge the ability of the active site models to dock active TG2 inhibitors. The criteria were reasonable and all were important as proven during the later stages of the work. The most important is the bent pose; it is this pose that prompted the design of irreversible inhibitors with lipophilic and electrophilic ends connected by a linker. An ideal active irreversible TG2 inhibitor would adopt a bent pose within the active site of TG2, with the electrophilic end pointing in the direction of CYS277 and the lipophilic end being embedded within the hydrophobic loop of TG2 active site. Hydrogen bonds connecting the linker of the inhibitor to active site residues such as ASN333, PHE334 and others are important for maintaining the bent pose within the

active site. Similarly, lipophilic interactions stabilising the lipophilic end of the inhibitor within the hydrophobic loop of TG2 active site are also important.

The selected 6 active site models were validated in the next stage by multiple experiments (section 3.4). These experiments started with the docking of 3 inactive compounds plus the original 6 active compounds, in which the models were able to distinguish between the compounds according to their inhibitory effect on TG2. The model performance proved to be similarly good when the test set was extended using more active and inactive compounds (validation processes 3 and 4), and the criteria for defining a plausible docking complex sustained their usefulness at this stage as well. There were differences in the performance of the individual models, where some models were better in capturing the bent pose for active compounds. Other models were superior in preventing the inactive compounds from docking appropriately into the active site. In any case, the docking scores failed to provide any useful information in ranking the compounds according to their biological activities. The scores even failed with the discrimination between active and inactive compounds. This, however, was achievable through the application of the docking criteria set at the beginning of the process, which again proved successful and managed to further validate the selected active site models. It should be noted here that when using GOLD, GoldScore, and to a lesser extent CHEMPLP, were able to rank the bent pose of TG2 inhibitors within the top 4 solutions for the active compounds, and this top ranking was one of the criteria used to define a plausible docking complex. Similar model performance results were achieved when 5-ns MD simulations were applied on docking complexes.

The selection of GoldScore as the primary scoring function during all the dockings performed with GOLD was validated through experiments in which the other scoring functions available in GOLD were tried. The 3 other functions (ASP, ChemScore and CHEMPLP) failed to produce results comparable to those of GoldScore in posing the active compounds correctly within the active site of TG2. The success of GoldScore was attributed to the importance of hydrogen bonding in the dockings of TG2 inhibitors and the relatively high molecular weights of the tested TG2 inhibitors (more van der Waals contribution), which are both important parameters in GoldScore (Verdonk et al. 2003). Forcing a covalent bond to be formed between the inhibitors and CYS277 during covalent docking and MD (section 3.7) proved that it is the ability of the compound to place itself in the correct pose within the active site that would ultimately determine its activity. This was determined from the similar results obtained for active and inactive compounds during docking as well as MD. No water molecules could be found near the formed covalent bond, and the water density around the bond could not be used to differentiate between active and inactive compounds.

When the 6 validated models were analysed, some differences in the observed docking results could be explained; namely, the slightly better performance for the models extracted from the trajectory of **1a**. These have been attributed, for example, to the orientation of the added loop between residues 319 and 327, where the orientation of the loop was acting as a determinant of the space available for the lipophilic part of the inhibitors within the active site of TG2. Another example for the difference between the 2 sets of models was the arrangement of the bridging tryptophan residues (241 and 332), which showed the stacking parallel shape in the trajectory of **1a** that is characteristic of their orientation in the original crystal structure.

The 3 independent MD runs applied on empty TG2 for 500 ns showed a similar behaviour for TG2 residues, with the most mobile parts being the loops composed of the residues that had been missing in the original TG2 crystal structure (2Q3Z). This was confirmed through measuring RMSD, atomic fluctuations and the principal components. Regarding the behaviour of water, it was observed that when TG2 is simulated on its own, water molecules usually reside within the active site close to the catalytic cysteine residues, and their highest density was recorded to be at a 2-Å distance from SG atom of CYS277. The number of water molecules as well as their density was found to decrease when there was an inhibitor within the active site during the simulation, as was the case with the trajectory of **1b**. Within the hydrophobic loop of TG2 active site, the likelihood of finding water molecules was much lower, and this probably explains why it was relatively easier to keep the lipophilic part of the inhibitor within the loop than keeping the warhead within the catalytic tunnel.

The results from Chapter 3 have shown in different ways that the selected 6 active site models are valid for discriminating between active and inactive TG2 inhibitors, if the set criteria were considered properly. The work performed within the chapter, however, could not produce appropriate correlations between many of the measured quantities and biological activity expressed as TG2 IC<sub>50</sub> values. The best and most easily accessible example was the docking score obtained from different scoring functions. This failure was attributed to the fact that the activity of the different inhibitors is triggered by a covalent bond formation with TG2 active site cysteine residue, and most of the techniques used in the chapter cannot account for bond formation. The production of an informative correlation with biological activity and curiosity behind the possible mechanisms of inhibition of TG2 by the studied inhibitors motivated the work in Chapter 4. Furthermore, the work in Chapter 4 added to the validity of the models by showing that their complexes can represent the actual chemical reactions behind the inhibition.

The two classes of inhibitors, acrylamide- and sulfonium ion-bearing TG2 inhibitors, were subjected to two different methods to investigate the inhibition mechanism. Both of the methods involved quantum mechanical (QM) treatment of the reaction centre. Inhibition with acrylamide inhibitors was investigated through umbrella sampling simulations using the QM/MM approach of AMBER, while the CAChe intrinsic reaction coordinate (IRC) reaction path tactic was employed when studying the inhibition by sulfonium ion inhibitors.

The inhibition by the acrylamide compounds was proposed to proceed according to one of two possibilities; as a single concerted step (where the nucleophilic SG from TG2 reacts with the electrophilic carbon of the inhibitor and at the same time a proton from HIS335 of TG2 saturates the alpha carbon of the acrylamide or C2) or as a two-step process (the nucleophile and the electrophile react first and the protonation follows). In the concerted mechanism, PM3, as the QM method, was able to drive the reaction to completion and to produce a reasonable correlation between the resultant PMF values and TG2 IC<sub>50</sub> values of the inhibitors. A certain level of desolvation was also observed in this method, where 1 or 2 water molecules have left the vicinity of the reaction centre just before the formation of the two bonds. The pattern of the change in the charges of the reacting atoms, however, was not uniform across the compounds and could not be used to devise a distinct mechanism for the inhibition. This happened because two intermediates were suggested based on the charge distribution and each intermediate supports a different mechanism.

For the 2-step mechanism, SCC-DFTB performed better than PM3 as the QM method. A good correlation was obtained between the PMF values and TG2 IC<sub>50</sub> values for the compounds in the 2 steps, and the correlation was better for the second step ( $R^2$  of 0.71 and 0.89 for the 1<sup>st</sup> and 2<sup>nd</sup> steps respectively). The charge distribution for the atoms participating in the reaction was uniform throughout the compounds and could be used to suggest a mechanism for the inhibition. Such a mechanism would involve the formation of an oxyanion intermediate having a negative charge on the oxygen atom of the acrylamide carbonyl group. The behaviour of water molecules did not show an obvious desolvation event during the 2-stage approach. To sum up with umbrella sampling, the concerted mechanism could show some desolvation effect while the charges agreed more with the 2 stages. Since CYS277 lies in a tunnel in the TG2 active site and since there were no water molecules in the vicinity of this residue in the original crystal structure, the water effect may not be very important. With that, and the charge pattern, the 2-stage approach may give a better representation for the inhibition of TG2 with acrylamide-based inhibitors.

The CAChe based experiments adopted a different approach. They started by locating a transition state structure, refining and verifying it and then using the verified structure as a

starting point for a reaction path experiment. The latter follows the IRC to the reactants and the products. The TS structures produced for the 8 studied sulfonium ion-based TG2 inhibitors had very similar structure parameters, such as the lengths of the bond to be formed and the bond to be broken and the angle at the reaction point. They had similar values for the charges of the 2 sulphur atoms and the orders of the formed and broken bonds. All these indicate a similar TS structure for inhibitors having the same sulfonium ion warhead. Nevertheless, the produced activation energies correlated very well with the biological activities, especially when only the compounds with IC<sub>50</sub> values of  $\leq 1 \mu\text{M}$  were considered. This difference in activation energy despite the similar mechanism was attributed to the starting conformation of the compounds with respect to CYS277, and this in turn was dependent on the compound's ability to dock into the active site. A backside attack S<sub>N</sub>2 mechanism for this reaction was proposed and the values for the angle at the reaction centre and the bond orders agree with such mechanism.

With regard to the work that involved allosteric inhibition, the allosteric binding site predicted by the University of Strathclyde group has been preferred over the original GDP binding site during docking, for more than one reason. The relatively buried location of the predicted site at the end of the catalytic core, compared to the totally exposed, and torn apart (part on the catalytic core and part on the 1<sup>st</sup>  $\beta$ -barrel in the open conformation of TG2) GDP binding site was the main reason. The fact that the binding site for GDP has not been actually confirmed as the binding site for the reported allosteric inhibitors (Case & Stein 2007; Caron et al. 2012) was another reason. Both chosen allosteric inhibitors attained a plausible pose within the predicted site in terms of their settlement within the site as well as the rank their poses achieved with the GoldScore scoring function.

The predicted site did not prove very useful for the binding of **i2** molecule, where the compound failed to maintain the docked pose during the aMD run applied, in addition to the lack of any significant and noticeable conformational change within TG2 structure. For **i3**, the predicted site was better; the compound maintained more stable pose within the site and could induce a relatively significant conformational change in TG2, manifested as the movement of the hinge region connecting the terminal  $\beta$ -barrel towards the catalytic core, in the same direction as the actual bending in the inactive closed form of TG2. The bending event did not end in the closed form of TG2, but it produced a distinctly different conformation from the starting structure of TG2, and the direction of the bending was correct. It can, therefore, be assumed that this hinge movement was a first step towards complete closure of TG2 conformation to achieve the inactive form.

The aMD run performed with 24 copies of **i2** distributed around TG2 (Run **4**) showed that **i2** has the potential to bind at the GDP binding site to induce a conformational change in TG2 similar to that produced by **i3** in the predicted site. This has been confirmed by Run **5** (in which all **i2** copies have been deleted from the last frame in Run **4**, save the one in the GDP binding site) and Run **6** (the final frame from Run **4** in which all **i2** copies have been deleted). The conformational change brought about in Run **4** was preserved in Run **5**, while TG2 in Run **6** adopted a structure similar to that produced by empty TG2 in Run **1**, in which there was a bending but in the wrong direction. “Wrong direction” was used to describe a TG2 bending event where the terminal barrel was pointing towards the outside of TG2 relative to the catalytic core and CYS277, making it difficult to achieve the TG2 inactivating closure.

The most significant change in the conformation of TG2, and the closest change to the complete closure of TG2 confrontation was observed during Run **2** when TG2 was surrounded by 24 copies of GDP. One of the copies found its way to the original GDP binding site where it induced a bending that involved the connection between the 2 terminal  $\beta$ -barrels and the catalytic core of TG2, in a manner that would be expected to end in the closed inactive TG2 conformation, considering the direction of the bending. The magnitude of the bending was more easily noticeable when compared with the bending events obtained from the other runs. This run confirms the ability of GDP to inactivate TG2 by closing its structure and that the aMD was capable of simulating this effect to some extent.

As a final remark, the aim set at the start of the thesis is expected to have been achieved, where several computational methods have been employed for the purpose of the analysis of the activities of known TG2 inhibitors. As a result, these methods could be applied when the objective is to predict the activity of new potential inhibitors for the enzyme.

## References

- Accelrys Software Incorporation, (2012). Accelrys Discovery Studio Visualizer version 3.5.0.12158, California, USA.
- Accelrys Software Incorporation, (2013). Accelrys Discovery Studio Visualizer version 4.0.100.13345, California, USA.
- Aeschlimann, D. & Thomazy, V., (2000). Protein crosslinking in assembly and remodelling of extracellular matrices: The role of transglutaminases. *Connective Tissue Research*, 41(1), pp.1–27.
- Ajay & Murcko, M.A., (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *Journal of Medicinal Chemistry*, 38(26), pp.4953–4967.
- Akimov, S.S. & Belkin, A.M., (2001). Cell-surface transglutaminase promotes fibronectin assembly via interaction with the gelatin-binding domain of fibronectin: A role in TGFbeta-dependent matrix deposition. *Journal of Cell Science*, 114(16), pp.2989–3000.
- Akimov, S.S., Krylov, D., Fleischman, L.F. & Belkin, A.M., (2000). Tissue transglutaminase is an integrin-binding adhesion coreceptor for fibronectin. *Journal of Cell Biology*, 148(4), pp.825–838.
- van den Akker, J., van Weert, A., Afink, G., Bakker, E.N.T.P., van der Pol, E., Böing, A.N., et al., (2012). Transglutaminase 2 is secreted from smooth muscle cells by transamidation-dependent microparticle formation. *Amino Acids*, 42(2-3), pp.961–973.
- Alberts, I.L., Todorov, N.P. & Dean, P.M., (2005). Receptor flexibility in de novo ligand design and docking. *Journal of Medicinal Chemistry*, 48(21), pp.6585–6596.
- Allinger, N.L., (1977). Conformational analysis. 130. MM2. A hydrocarbon force field utilizing  $V_1$  and  $V_2$  torsional terms. *Journal of the American Chemical Society*, 99(25), pp.8127–8134.
- Badarau, E., Collighan, R.J. & Griffin, M., (2013). Recent advances in the development of tissue transglutaminase (TG2) inhibitors. *Amino Acids*, 44(1), pp.119–127.
- Badarau, E., Mongeot, A., Collighan, R., Rathbone, D. & Griffin, M., (2013). Imidazolium-based warheads strongly influence activity of water-soluble peptidic transglutaminase inhibitors. *European Journal of Medicinal Chemistry*, 66, pp.526–530.
- Badarau, E., Wang, Z., Rathbone, D.L., Costanzi, A., Thibault, T., Murdoch, C.E., et al., (2015). Development of potent and selective tissue transglutaminase inhibitors: Their effect on TG2 function and application in pathological conditions. *Chemistry & Biology*, 22(10), pp.1347–61.
- Baek, K.J., Kwon, N.S., Lee, H.S., Kim, M.S., Muralidhar, P. & Im, M., (1996). Oxytocin receptor couples to the 80 kDa G $\alpha$  myometrium family protein in human myometrium. *Biochemical Journal*, 315(3), pp.739–744.

- Baowei, W., Encui, Y. & Genhui, X., (2007). Theoretical study of reaction paths and transition states on conversion methane into C<sub>2</sub> hydrocarbons through plasma. *Chinese Journal of Chemical Engineering*, 15(1), pp.44–50.
- Baştuğ, T., Chen, P.C., Patra, S.M. & Kuyucak, S., (2008). Potential of mean force calculations of ligand binding to ion channels from Jarzynski's equality and umbrella sampling. *Journal of Chemical Physics*, 128(15), pp.1–9.
- Baxter, C. a., Murray, C.W., Clark, D.E., Westhead, D.R. & Eldridge, M.D., (1998). Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins: Structure, Function and Genetics*, 33(3), pp.367–382.
- Bayly, C., Cieplak, P., Cornell, W. & Kollman, P., (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *Journal of Physical Chemistry*, 97(40), pp.10269–10280.
- Belkin, A.M., (2011). Extracellular TG2: Emerging functions and regulation. *FEBS Journal*, 278(24), pp.4704–4716.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp.235–242.
- Bernardi, R.C., Melo, M.C.R. & Schulten, K., (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta - General Subjects*, 1850(5), pp.872–877.
- Bernasconi, C.F., (1989). Nucleophilic addition to olefins. Kinetics and mechanism. *Tetrahedron*, 45(13), pp.4017–4090.
- Bernassola, F., Federici, M., Corazzari, M., Terrinoni, A., Hribal, M.L., De Laurenzi, V., et al., (2002). Role of transglutaminase 2 in glucose tolerance: Knockout mice studies and a putative mutation in a MODY patient. *The FASEB Journal*, 16(11), pp.1371–1378.
- Boeyens, J.C.A. & Comba, P., (2001). Molecular mechanics: theoretical basis, rules, scope and limits. *Coordination Chemistry Reviews*, 212(1), pp.3–10.
- Boros, S., Åhrman, E., Wunderink, L., Kamps, B., De Jong, W.W., Boelens, W.C., et al., (2006). Site-specific transamidation and deamidation of the small heat-shock protein Hsp20 by tissue transglutaminase. *Proteins: Structure, Function and Genetics*, 62(4), pp.1044–1052.
- Boros, S., Wilmarth, P. a., Kamps, B., de Jong, W.W., Bloemendal, H., Lampi, K., et al., (2008). Tissue transglutaminase catalyzes the deamidation of glutamines in lens βB2- and βB3-crystallins. *Experimental Eye Research*, 86(2), pp.383–393.
- B-Rao, C., Subramanian, J. & Sharma, S.D., (2009). Managing protein flexibility in docking and its applications. *Drug Discovery Today*, 14(7-8), pp.394–400.
- Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., et al., (2009). CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), pp.1545–1614.

- Bürgi, H.B., Dunitz, J.D., Lehn, J.M. & Wipff, G., (1974). Stereochemistry of reaction paths at carbonyl centres. *Tetrahedron*, 30(12), pp.1563–1572.
- CAChe-Group, (2006). *CAChe User Guide: Molecular Modeling in Chemistry*, Oregon: Fujitsu Limited.
- Candi, E., Melino, G., Lahm, A. & Ceci, R., (1998). Transglutaminase 1 mutations in lamellar ichthyosis. Loss of activity due to failure of activation by proteolytic processing. *Journal of Biological Chemistry*, 273(22), pp.13693–13702.
- Candi, E., Oddi, S., Paradisi, A., Terrinoni, A., Ranalli, M., Teofoli, P., et al., (2002). Expression of transglutaminase 5 in normal and pathologic human epidermis. *Journal of Investigative Dermatology*, 119(3), pp.670–677.
- Capoferri, L., Lodola, A., Rivara, S. & Mor, M., (2015). Quantum mechanics/molecular mechanics modeling of covalent addition between EGFR-cysteine 797 and N -(4-anilinoquinazolin-6-yl) acrylamide. *Journal of Chemical Information and Modeling*, 55(3), pp.589–599.
- Carlson, H.A., (2002). Protein flexibility and drug design: How to hit a moving target. *Current Opinion in Chemical Biology*, 6(4), pp.447–452.
- Caron, N.S., Munsie, L.N., Keillor, J.W. & Truant, R., (2012). Using FLIM-FRET to measure conformational changes of transglutaminase type 2 in live cells. *PLoS ONE*, 7(8), p.e44159.
- Case, A. & Stein, R.L., (2007). Kinetic analysis of the interaction of tissue transglutaminase with a nonpeptidic slow-binding inhibitor. *Biochemistry*, 46(4), pp.1106–1115.
- Case, D.A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke, R.E., et al., (2012). AMBER 12, University of California, San Francisco, USA.
- CCDC Software Limited, (2013). GOLD Suite version 5.2.2, Cambridge, UK.
- Chica, R.A., Gagnon, P., Keillor, J.W. & Pelletier, J.N., (2004). Tissue transglutaminase acylation: Proposed role of conserved active site Tyr and Trp residues revealed by molecular modeling of peptide substrate binding. *Protein Science*, 13(4), pp.979–991.
- Cho, S.-Y., Choi, K., Jeon, J.-H., Kim, C.-W., Shin, D.-M., Lee, J.B., et al., (2010). Differential alternative splicing of human transglutaminase 4 in benign prostate hyperplasia and prostate cancer. *Experimental & Molecular Medicine*, 42(4), pp.310–318.
- Choi, K., Siegel, M., Piper, J.L., Yuan, L., Cho, E., Strnad, P., et al., (2005). Chemistry and biology of dihydroisoxazole derivatives: Selective inhibitors of human transglutaminase 2. *Chemistry & Biology*, 12(4), pp.469–475.
- Cole, J.C., Nissink, J.W.M. & Taylor, R., (2005). Protein-ligand docking and virtual screening with GOLD. In J. Alvarez & B. Shoichet, eds. *Virtual Screening in Drug Discovery*. Florida: Taylor & Francis CRC Press, pp. 379–411.
- Cooper, A.J.L., Jeitner, T.M., Gentile, V. & Blass, J.P., (2002). Cross linking of polyglutamine domains catalyzed by tissue transglutaminase is greatly favored with pathological-

- length repeats: Does transglutaminase activity play a role in (CAG)<sub>n</sub>/Qn-expansion diseases? *Neurochemistry International*, 40(1), pp.53–67.
- Cozzini, P., Kellogg, G.E., Spyraakis, F., Abraham, D.J., Costantino, G., Emerson, A., et al., (2008). Target flexibility: An emerging consideration in drug discovery and design. *Journal of Medicinal Chemistry*, 51(20), pp.6237–6255.
- Cramer, C.J., (2013). *Essentials of Computational Chemistry: Theories and Models* 2<sup>nd</sup> ed., John Wiley & Sons, Inc.
- Cross, S.S.J., (2005). Improved FlexX docking using FlexS-determined base fragment placement. *Journal of Chemical Information and Modeling*, 45(4), pp.993–1001.
- David, C.C. & Jacobs, D.J., (2014). Principal component analysis: A method for determining the essential dynamics of proteins. *Methods in Molecular Biology*, 1048, pp.193–226.
- Dedeoglu, A., Kubilus, J., Jeitner, T.M., Matson, S.A., Bogdanov, M., Kowall, N.W., et al., (2002). Therapeutic effects of cystamine in a murine model of Huntington's disease. *The Journal of Neuroscience*, 22(20), pp.8942–8950.
- Dieterich, W., Ehnis, T., Bauer, M., Donner, P., Volta, U., Riecken, E.O., et al., (1997). Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nature Medicine*, 3(7), pp.797–801.
- Dodson, M.L., Walker, R.C. & Lloyd, R.S., (2012). Carbinolamine formation and dehydration in a DNA repair enzyme active site. *PLoS ONE*, 7(2), p.e31377.
- Dorsett, H. & White, A., (2000). *Overview of molecular modelling and ab initio molecular orbital methods suitable for use with energetic materials.*, Salisbury: DSTO Aeronautical and Maritime Research Laboratory.
- Dørum, S., Arntzen, M.Ø., Qiao, S.-W., Holm, A., Koehler, C.J., Thiede, B., et al., (2010). The preferred substrates for transglutaminase 2 in a complex wheat gluten digest are peptide fragments harboring celiac disease T-cell epitopes. *PLoS ONE*, 5(11), p.e14056.
- Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G. V & Mee, R.P., (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*, 11(5), pp.425–445.
- Elfturi, F., (2014). *Computational investigation of InGaN alloys over the full composition range.* University of Strathclyde.
- Elstner, M., (2006). The SCC-DFTB method and its application to biological systems. *Theoretical Chemistry Accounts*, 116(1), pp.316–325.
- Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Frauenheim, T., et al., (1998). Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B*, 58(11), pp.7260–7268.
- Facchiano, F., Facchiano, A. & Facchiano, A.M., (2006). The role of transglutaminase-2 and its substrates in human diseases. *Frontiers in Bioscience*, 11, pp.1758–1773.

- Foulkes, W.M.C. & Haydock, R., (1989). Tight-binding models and density-functional theory. *Physical Review B*, 39(17), pp.12520–12536.
- Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., et al., (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), pp.1739–1749.
- Fujitsu Limited, (2006). CAChe WorkSystem Pro version 7.5.0.85, Tokyo, Japan.
- Fukui, K., (1970). Formulation of the reaction coordinate. *The Journal of Physical Chemistry*, 74(23), pp.4161–4163.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. & Bairoch, A., (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31(13), pp.3784–3788.
- Genheden, S. & Ryde, U., (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5), pp.449–461.
- Gore, J., Ritort, F. & Bustamante, C., (2003). Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22), pp.12564–12569.
- Goringe, C.M., Bowler, D.R. & Hernández, E., (1997). Tight-binding modelling of materials. *Reports on Progress in Physics*, 60(12), pp.1447–1512.
- Gotō, H. & Ōsawa, E., (1993). An efficient algorithm for searching low-energy conformers of cyclic and acyclic molecules. *Journal of the Chemical Society, Perkin Transactions*, 2(2), pp.187–198.
- Götz, A.W., Williamson, M.J., Xu, D., Poole, D., Le Grand, S. & Walker, R.C., (2012). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation*, 8(5), pp.1542–1555.
- Le Grand, S., Götz, A.W. & Walker, R.C., (2013). SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications*, 184(2), pp.374–380.
- Griffin, M., Casadio, R. & Bergamini, C.M., (2002). Transglutaminases: Nature's biological glues. *Biochemical Journal*, 368(2), pp.377–396.
- Griffin, M., Mongeot, A., Collighan, R., Saint, R.E., Jones, R.A., Coutts, I.G.C., et al., (2008). Synthesis of potent water-soluble tissue transglutaminase inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 18(20), pp.5559–5562.
- Griffin, M., Rathbone, D.L. & Badarau, E., (2014). Acylpiperazines as inhibitors of transglutaminase and their use in medicine.
- Grossfield, A., (2013). WHAM: the weighted histogram analysis method, version 2.0.9.
- Guan, W.J., Xia, K. De, Ma, Y.T., Liu, Y.T., Shi, Y.T., Jiang, H., et al., (2013). Transglutaminase 6 interacts with polyQ proteins and promotes the formation of polyQ

- aggregates. *Biochemical and Biophysical Research Communications*, 437(1), pp.94–100.
- Gundemir, S., Colak, G., Tucholski, J. & Johnson, G.V.W., (2012). Transglutaminase 2: A molecular Swiss army knife. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1823(2), pp.406–419.
- Halim, D., Caron, K. & Keillor, J.W., (2007). Synthesis and evaluation of peptidic maleimides as transglutaminase inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 17(2), pp.305–308.
- Hamelberg, D., Mongan, J. & McCammon, J.A., (2004). Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120(24), pp.11919–11929.
- Hartshorn, M.J., Verdonk, M.L., Chessari, G., Brewerton, S.C., Mooij, W.T.M., Mortenson, P.N., et al., (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4), pp.726–741.
- Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E., (2008). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3), pp.435–447.
- Hinchliffe, A., (2003). *Molecular Modelling for Beginners* 1<sup>st</sup> ed., West Sussex: Wiley.
- Höltje, H.-D. & Folkers, G., (2008). *Molecular Modeling Basic Principles and Applications* 2<sup>nd</sup> ed., Weinheim: Wiley-VCH Verlag GmbH.
- Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C., (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3), pp.712–725.
- Hospital, A., Goñi, J.R., Orozco, M. & Gelpi, J., (2015). Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 8, pp.37–47.
- Hou, T., Wang, J., Li, Y. & Wang, W., (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1), pp.69–82.
- Huang, S.Y. & Zou, X., (2010). Advances and challenges in protein-ligand docking. *International Journal of Molecular Sciences*, 11(8), pp.3016–3034.
- Huang, S.-Y. & Zou, X., (2007). Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 66(2), pp.399–421.
- Hub, J.S., De Groot, B.L. & Van Der Spoel, D., (2010). G-whams-a free Weighted Histogram Analysis implementation including robust error and autocorrelation estimates. *Journal of Chemical Theory and Computation*, 6(12), pp.3713–3720.

- Humphrey, W., Dalke, A. & Schulten, K., (1996). VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14(1), pp.33–38.
- Hwang, K.-C., Gray, C.D., Sivasubramanian, N. & Im, M.-J., (1995). Interaction site of GTP binding Gh (transglutaminase II) with phospholipase C. *Journal of Biological Chemistry*, 270(45), pp.27058–27062.
- Iismaa, S.E., Holman, S., Wouters, M. a, Lorand, L., Graham, R.M. & Husain, A., (2003). Evolutionary specialization of a tryptophan indole group for transition-state stabilization by eukaryotic transglutaminases. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22), pp.12636–12641.
- Ikuo, A. & Ogawa, H., (2014). Visualization of reaction mechanism by CG based on quantum chemical calculation - An approach to electronic laboratory textbook. *African Journal of Chemical Education*, 4(3), pp.22–33.
- Isralewitz, B., Baudry, J., Gullingsrud, J., Kosztin, D. & Schulten, K., (2001). Steered molecular dynamics investigations of protein function. *Journal of Molecular Graphics and Modelling*, 19(1), pp.13–25.
- Iwata, Y., Tago, K., Kiho, T., Kogen, H., Fujioka, T., Otsuka, N., et al., (2000). Conformational analysis and docking study of potent factor XIIIa inhibitors having a cyclopropanone ring. *Journal of Molecular Graphics & Modelling*, 18(6), pp.591–599.
- Jang, T.H., Lee, D.S., Choi, K., Jeong, E.M., Kim, I.G., Kim, Y.W., et al., (2014). Crystal structure of transglutaminase 2 with GTP complex and amino acid sequence evidence of evolution of GTP binding site. *PLoS ONE*, 9(9), p.e107005.
- Jankovic, J., (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4), pp.368–376.
- Jarzynski, C., (1997)(a). A nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14), pp.2690–2693.
- Jarzynski, C., (1997)(b). Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5), pp.5018–5035.
- Jeitner, T.M., Delikatny, E.J., Ahlqvist, J., Capper, H. & Cooper, A.J.L., (2005). Mechanism for the inhibition of transglutaminase 2 by cystamine. *Biochemical Pharmacology*, 69(6), pp.961–970.
- Jiang, W.G., Ye, L., Sanders, A.J., Ruge, F., Kynaston, H.G., Ablin, R.J., et al., (2013). Prostate transglutaminase (TGase-4, TGaseP) enhances the adhesion of prostate cancer cells to extracellular matrix, the potential role of TGase-core domain. *Journal of Translational Medicine*, 11(1), pp.269–280.
- Johansson, A., Kollman, P., Rothenberg, S. & McKelvey, J., (1974). Hydrogen bonding ability of the amide group. *Journal of the American Chemical Society*, 96(12), pp.3794–3800.
- John, S., Thiebach, L., Frie, C., Mokkaapati, S., Bechtel, M., Nischt, R., et al., (2012). Epidermal transglutaminase (TGase 3) is required for proper hair development, but not the formation of the epidermal barrier. *PLoS ONE*, 7(4), pp.1–12.

- Jones, G., Willett, P. & Glen, R.C., (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, 245(1), pp.43–53.
- Jones, G., Willett, P., Glen, R.C., Leach, a R. & Taylor, R., (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 267(3), pp.727–748.
- Jones, R.A., Kotsakis, P., Johnson, T.S., Chau, D.Y.S., Ali, S., Melino, G., et al., (2006). Matrix changes induced by transglutaminase 2 lead to inhibition of angiogenesis and tumor growth. *Cell Death and Differentiation*, 13(9), pp.1442–1453.
- Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L., (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2), pp.926–935.
- Junn, E., Ronchetti, R.D., Quezado, M.M., Kim, S.-Y. & Mouradian, M.M., (2003). Tissue transglutaminase-induced aggregation of  $\alpha$ -synuclein: Implications for Lewy body formation in Parkinson's disease and dementia with Lewy bodies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4), pp.2047–2052.
- van der Kamp, M.W. & Mulholland, A.J., (2013). Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry*, 52(16), pp.2708–2728.
- Karplus, M. & McCammon, J.A., (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9), pp.646–652.
- Kästner, J., (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6), pp.932–942.
- Katano, M., Okamoto, K., Arito, M., Kawakami, Y., Kurokawa, M.S., Suematsu, N., et al., (2009). Implication of granulocyte-macrophage colony-stimulating factor induced neutrophil gelatinase-associated lipocalin in pathogenesis of rheumatoid arthritis revealed by proteome analysis. *Arthritis Research & Therapy*, 11(1), pp.1–12.
- Kitchen, D.B., Decornez, H., Furr, J.R. & Bajorath, J., (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews. Drug Discovery*, 3(11), pp.935–949.
- Klock, C., Herrera, Z., Albertelli, M. & Khosla, C., (2014). Discovery of potent and specific dihydroisoxazole inhibitors of human transglutaminase 2. *Journal of Medicinal Chemistry*, 57(21), pp.9042–9064.
- Korb, O., Stützle, T. & Exner, T.E., (2009). Empirical scoring functions for advanced protein-ligand docking with PLANTS. *Journal of Chemical Information and Modeling*, 49(1), pp.84–96.
- Koskinen, P. & Mäkinen, V., (2009). Density-functional tight-binding for beginners. *Computational Materials Science*, 47(1), pp.237–253.

- Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H. & Kollman, P.A., (1995). Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry*, 16(11), pp.1339–1350.
- Kuramoto, K., Yamasaki, R., Shimizu, Y., Tatsukawa, H. & Hitomi, K., (2013). Phage-displayed peptide library screening for preferred human substrate peptide sequences for transglutaminase 7. *Archives of Biochemistry and Biophysics*, 537(1), pp.138–143.
- Ladbury, J.E., (1996). Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology*, 3(12), pp.973–980.
- De Laurenzi, V. & Melino, G., (2001). Gene disruption of tissue transglutaminase. *Molecular and Cellular Biology*, 21(1), pp.148–155.
- Leach, A.R., (2001). *Molecular Modelling: Principles and Applications* 2<sup>nd</sup> ed., Essex: Pearson Education Limited.
- Lehninger, A.L., Nelson, D.L. & Cox, M.M., (2000). *Lehninger Principles of Biochemistry* 3<sup>rd</sup> ed., New York: Worth Publishers.
- Lesort, M., Chun, W., Johnson, G.V.W. & Ferrante, R.J., (1999). Tissue transglutaminase is increased in Huntington's disease brain. *Journal of Neurochemistry*, 73(5), pp.2018–2027.
- Lesort, M., Lee, M., Tucholski, J. & Johnson, G.V.W., (2003). Cystamine inhibits caspase activity. Implications for the treatment of polyglutamine disorders. *The Journal of Biological Chemistry*, 278(6), pp.3825–3830.
- Liao, S.-M., Du, Q.-S., Meng, J.-Z., Pang, Z.-W. & Huang, R.-B., (2013). The multiple roles of histidine in protein interactions. *Chemistry Central Journal*, 7(1), pp.44–55.
- Liu, S., Cerione, R.A. & Clardy, J., (2002). Structural basis for the guanine nucleotide-binding activity of tissue transglutaminase and its regulation of transamidation activity. *Proceedings of the National Academy of Sciences of the United States of America*, 99(5), pp.2743–2747.
- Lonsdale, R., Harvey, J.N. & Mulholland, A.J., (2012). A practical guide to modelling enzyme-catalysed reactions. *Chemical Society Reviews*, 41(8), pp.3025–3038.
- Lorand, L., (1998). DRPLA aggregation and transglutaminase, revisited. *Nature Genetics*, 20(3), p.231.
- Lorand, L. & Conrad, S., (1984). Transglutaminases. *Molecular and Cellular Biochemistry*, 58(1-2), pp.9–35.
- Lovell, S.C., Word, J.M., Richardson, J.S. & Richardson, D.C., (2000). The penultimate rotamer library. *Proteins*, 40(3), pp.389–408.
- de Macédo, P., Marrano, C. & Keillor, J.W., (2000). A direct continuous spectrophotometric assay for transglutaminase activity. *Analytical Biochemistry*, 285(1), pp.16–20.

- de Macédo, P., Marrano, C. & Keillor, J.W., (2002). Synthesis of dipeptide-bound epoxides and alpha,beta-unsaturated amides as potential irreversible transglutaminase inhibitors. *Bioorganic & Medicinal Chemistry*, 10(2), pp.355–360.
- MackKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., et al., (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18), pp.3586–3616.
- Maeda, S., Harabuchi, Y., Ono, Y., Taketsugu, T. & Morokuma, K., (2015). Intrinsic reaction coordinate: Calculation, bifurcation, and automated search. *International Journal of Quantum Chemistry*, 115(5), pp.258–269.
- Maiur, L., Luciani, A., Giardino, I., Raia, V., Villella, V.R., D’Apolito, M., et al., (2008). Tissue transglutaminase activation modulates inflammation in cystic fibrosis via PPAR $\gamma$  down-regulation. *The Journal of Immunology*, 180(11), pp.7697–7705.
- Mangala, L.S., Fok, J.Y., Zorrilla-Calancha, I.R., Verma, A. & Mehta, K., (2007). Tissue transglutaminase expression promotes cell attachment, invasion and survival in breast cancer cells. *Oncogene*, 26(17), pp.2459–2470.
- Martin, A., Giuliano, A., Collaro, D., De Vivo, G., Sedia, C., Serretiello, E., et al., (2013). Possible involvement of transglutaminase-catalyzed reactions in the physiopathology of neurodegenerative diseases. *Amino Acids*, 44(1), pp.111–118.
- Mastroberardino, P.G., Iannicola, C., Nardacci, R., Bernassola, F., De Laurenzi, V., Melino, G., et al., (2002). ‘Tissue’ transglutaminase ablation reduces neuronal death and prolongs survival in a mouse model of Huntington’s disease. *Cell Death and Differentiation*, 9(9), pp.873–880.
- McCammon, J.A., Gelin, B.R. & Karplus, M., (1977). Dynamics of folded proteins. *Nature*, 267(5612), pp.585–590.
- Mehta, K., Fok, J., Miller, F.R., Koul, D. & Sahin, A.A., (2004). Prognostic significance of tissue transglutaminase in drug resistant and metastatic breast cancer. *Clinical Cancer Research*, 10(23), pp.8068–8076.
- Mehta, K., Kumar, A. & Kim, H.I., (2010). Transglutaminase 2: A multi-tasking protein in the complex circuitry of inflammation and cancer. *Biochemical Pharmacology*, 80(12), pp.1921–1929.
- Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M., (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design*, 7(2), pp.146–157.
- Miao, Y., Nichols, S., Gasper, P., Metger, V. & McCammon, J., (2013). Activation and dynamic network of the M2 muscarinic receptor. *Proceedings of the National Academy of Sciences of the United States of America*, 110(27), pp.10982–10987.
- Miao, Y., Sinko, W., Pierce, L., Bucher, D., Walker, R.C. & Mccammon, J.A., (2014). Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. *Journal of Chemical Theory and Computation*, 10(7), pp.2677–2689.

- Michel, J. & Essex, J.W., (2010). Prediction of protein-ligand binding affinity by free energy simulations: Assumptions, pitfalls and expectations. *Journal of Computer-Aided Molecular Design*, 24(8), pp.639–658.
- Miller, B.R., Mcgee, T.D., Swails, J.M., Homeyer, N., Gohlke, H. & Roitberg, A.E., (2012). MMPBSA.py: An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8(9), pp.3314–3321.
- Mills, M. & Andricioaei, I., (2008). An experimentally guided umbrella sampling protocol for biomolecules. *Journal of Chemical Physics*, 129(11), pp.1–13.
- Mishra, S., Melino, G. & Murphy, L.J., (2007). Transglutaminase 2 kinase activity facilitates protein kinase A-induced phosphorylation of retinoblastoma protein. *Journal of Biological Chemistry*, 282(25), pp.18108–18115.
- Mishra, S. & Murphy, L.J., (2006). The p53 oncoprotein is a substrate for tissue transglutaminase kinase activity. *Biochemical and Biophysical Research Communications*, 339(2), pp.726–730.
- Mishra, S. & Murphy, L.J., (2004). Tissue transglutaminase has intrinsic kinase activity. Identification of transglutaminase 2 as an insulin-like growth factor-binding protein-3 kinase. *Journal of Biological Chemistry*, 279(23), pp.23863–23868.
- Mishra, S., Saleh, A., Espino, P.S., Davie, J.R. & Murphy, L.J., (2006). Phosphorylation of histones by tissue transglutaminase. *Journal of Biological Chemistry*, 281(9), pp.5532–5538.
- Mitsutake, A., Sugita, Y. & Okamoto, Y., (2001). Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, 60(2), pp.96–123.
- Mohan, V., Gibbs, A.C., Cummings, M.D., Jaeger, E.P. & DesJarlais, R.L., (2005). Docking: successes and challenges. *Current Pharmaceutical Design*, 11(3), pp.323–333.
- Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C.R., (2008). Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British Journal of Pharmacology*, 153(Suppl 1), pp.S7–S26.
- Mooij, W.T.M. & Verdonk, M.L., (2005). General and targeted statistical potentials for protein-ligand interactions. *Proteins: Structure, Function and Bioinformatics*, 61(2), pp.272–287.
- Mortier, J., Rakers, C., Bermudez, M., Murgueitio, M.S., Riniker, S. & Wolber, G., (2015). The impact of molecular dynamics on drug design: Applications for the characterization of ligand-macromolecule complexes. *Drug Discovery Today*, 20(6), pp.686–702.
- Munoz-Sanjuan, I. & Bates, G.P., (2011). The importance of integrating basic and clinical research toward the development of new therapies for Huntington disease. *The Journal of Clinical Investigation*, 121(2), pp.476–483.
- Muszbek, L., Bereczky, Z., Bagoly, Z., Komáromi, I. & Katona, É., (2011). Factor XIII: A coagulation factor with multiple plasmatic and cellular functions. *Physiological Reviews*, 91(3), pp.931–972.

- Nagata, T., (2014). Molby version 1.0b3.
- Nair, D.P., Podgórski, M., Chatani, S., Gong, T., Xi, W., Fenoli, C.R., et al., (2014). The thiol-Michael addition click reaction: A powerful and widely used tool in materials chemistry. *Chemistry of Materials*, 26(1), pp.724–744.
- Nanda, N., Iismaa, S.E., Owens, W.A., Husain, A., Mackay, F. & Graham, R.M., (2001). Targeted inactivation of Gh/tissue transglutaminase II. *Journal of Biological Chemistry*, 276(23), pp.20673–20678.
- Nemes, Z., Petrovski, G., Aerts, M., Sergeant, K., Devreese, B. & Fésüs, L., (2009). Transglutaminase-mediated intramolecular cross-linking of membrane-bound  $\alpha$ -synuclein promotes amyloid formation in Lewy bodies. *Journal of Biological Chemistry*, 284(40), pp.27252–27264.
- Ng, H.W., Laughton, C.A. & Doughty, S.W., (2013). Molecular dynamics simulations of the adenosine A2a receptor: structural stability, sampling, and convergence. *Journal of Chemical Information and Modeling*, 53(5), pp.1168–1178.
- Nicholas, B., Smethurst, P., Verderio, E., Jones, R. & Griffin, M., (2003). Cross-linking of cellular proteins by tissue transglutaminase during necrotic cell death: A mechanism for maintaining tissue integrity. *Biochemical Journal*, 371(2), pp.413–422.
- Nissink, J.W.M., Murray, C., Hartshorn, M., Verdonk, M.L., Cole, J.C. & Taylor, R., (2002). A new test set for validating predictions of protein-ligand interaction. *Proteins*, 49(4), pp.457–471.
- Nurminskaya, M. V. & Belkin, A.M., (2013). Cellular functions of tissue transglutaminase. *International Review of Cell and Molecular Biology*, 294, pp.1–97.
- Odi, B.O. & Coussons, P., (2014). Biological functionalities of transglutaminase 2 and the possibility of its compensation by other members of the transglutaminase family. *The Scientific World Journal*, 2014, pp.1–13.
- Ohtake, Y., Maruko, A., Abe, S., Fukumoto, M. & Ohkubo, Y., (2006). Effect of retinoic acid-induced transglutaminase on cell growth in regenerating liver. *Biomedical Research*, 27(2), pp.75–80.
- Olsen, K.C., Sapinoro, R.E., Kottmann, R.M., Kulkarni, A.A., Iismaa, S.E., Johnson, G.V.W., et al., (2011). Transglutaminase 2 and its role in pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 184(6), pp.699–707.
- Onufriev, A., (2008). Implicit solvent models in molecular dynamics simulations: A brief overview. *Annual Reports in Computational Chemistry*, 4, pp.125–137.
- Onufriev, A., Bashford, D. & Case, D. a., (2004). Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins: Structure, Function and Genetics*, 55(2), pp.383–394.
- Oostenbrink, C., Villa, A., Mark, A.E. & Van Gunsteren, W.F., (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25(13), pp.1656–1676.

- Orio, M., Pantazis, D. a. & Neese, F., (2009). Density functional theory. *Photosynthesis Research*, 102(2), pp.443–453.
- Paquet, E. & Viktor, H.L., (2015). Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: A computational review. *BioMed Research International*, 2015, pp.1–18.
- Pardin, C., Gillet, S.M.F.G. & Keillor, J.W., (2006). Synthesis and evaluation of peptidic irreversible inhibitors of tissue transglutaminase. *Bioorganic & Medicinal Chemistry*, 14(24), pp.8379–8385.
- Pardin, C., Pelletier, J.N., Lubell, W.D. & Keillor, J.W., (2008). Cinnamoyl inhibitors of tissue transglutaminase. *The Journal of Organic Chemistry*, 73(15), pp.5766–5775.
- Pardin, C., Roy, I., Lubell, W.D. & Keillor, J.W., (2008). Reversible and competitive cinnamoyl triazole inhibitors of tissue transglutaminase. *Chemical Biology & Drug Design*, 72(3), pp.189–196.
- Park, M.K., Lee, H.J., Shin, J., Noh, M., Kim, S.Y. & Lee, C.H., (2011). Novel participation of transglutaminase-2 through c-Jun N-terminal kinase activation in sphingosylphosphorylcholine-induced keratin reorganization of PANC-1 cells. *Biochimica et Biophysica Acta*, 1811(12), pp.1021–1029.
- Park, M.K., You, H.J., Lee, H.J., Kang, J.H., Oh, S.H., Kim, S.Y., et al., (2013). Transglutaminase-2 induces N-cadherin expression in TGF- $\beta$ 1-induced epithelial mesenchymal transition via c-Jun-N-terminal kinase activation by protein phosphatase 2A down-regulation. *European Journal of Cancer*, 49(7), pp.1692–1705.
- Park, S., Khalili-Araghi, F., Tajkhorshid, E. & Schulten, K., (2003). Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *The Journal of Chemical Physics*, 119(6), pp.3559–3566.
- Perez, A., van Heeckeren, A.M., Nichols, D., Gupta, S., Eastman, J.F. & Davis, P.B., (2008). Peroxisome proliferator-activated receptor- $\gamma$  in cystic fibrosis lung epithelium. *American Journal of Physiology - Lung Cellular and Molecular Physiology*, 295(2), pp.L303–L313.
- Perola, E., Walters, W.P. & Charifson, P.S., (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Structure, Function and Genetics*, 56(2), pp.235–249.
- Pierce, L.C.T., Salomon-Ferrer, R., Augusto F de Oliveira, C., McCammon, J.A. & Walker, R.C., (2012). Routine access to millisecond time scale events with accelerated molecular dynamics. *Journal of Chemical Theory and Computation*, 8(9), pp.2997–3002.
- Pinkas, D.M., Strop, P., Brunger, A.T. & Khosla, C., (2007). Transglutaminase 2 undergoes a large conformational change upon activation. *PLoS Biology*, 5(12), pp.2788–2796.
- Prime, M.E., Andersen, O.A., Barker, J.J., Brooks, M.A., Cheng, R.K.Y., Toogood-Johnson, I., et al., (2012). Discovery and structure–activity relationship of potent and selective covalent inhibitors of transglutaminase 2 for Huntington's disease. *Journal of Medicinal Chemistry*, 55(3), pp.1021–1046.

- Prime, M.E., Brookfield, F.A., Courtney, S.M., Gaines, S., Marston, R.W., Ichihara, O., et al., (2012). Irreversible 4-aminopiperidine transglutaminase 2 inhibitors for Huntington's Disease. *ACS Medicinal Chemistry Letters*, 3(9), pp.731–735.
- Pritchard, L., Cardle, L., Quinn, S. & Dufton, M., (2003). 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 A2, chymotrypsin and carboxypeptidase A. *Protein Engineering*, 16(2), pp.87–101.
- Ramachandran, G.N. & Sasisekharan, V., (1968). Conformation of polypeptides and proteins. *Advances in Protein Chemistry*, 23, pp.283–437.
- Ranaghan, K.E. & Mulholland, A.J., (2010). Investigations of enzyme-catalysed reactions with combined quantum mechanics/molecular mechanics (QM/MM) methods. *International Reviews in Physical Chemistry*, 29(1), pp.65–133.
- Ratjen, F. & Döring, G., (2003). Cystic fibrosis. *The Lancet*, 361(9358), pp.681–689.
- Rauhavirta, T., Oittinen, M., Kivistö, R., Männistö, P.T., Garcia-Horsman, J.A., Wang, Z., et al., (2013). Are transglutaminase 2 inhibitors able to reduce gliadin-induced toxicity related to celiac disease? A proof-of-concept study. *Journal of Clinical Immunology*, 33(1), pp.134–142.
- Reisz, J. a., Bechtold, E., King, S.B., Poole, L.B. & Furdui, C.M., (2013). Thiol-blocking electrophiles interfere with labeling and detection of protein sulfenic acids. *FEBS Journal*, 280(23), pp.6150–6161.
- Ringer, A.L., Senenko, A. & Sherrill, C.D., (2007). Models of S/ $\pi$  interactions in protein structures : Comparison of the H<sub>2</sub>S–benzene complex with PDB data. *Protein Science*, 16(10), pp.2216–2223.
- Roe, D.R. & Cheatham, T.E., (2013). PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9(7), pp.3084–3095.
- Sakly, W., Thomas, V., Quash, G. & El Alaoui, S., (2006). A role for tissue transglutaminase in alpha-gliadin peptide cytotoxicity. *Clinical and Experimental Immunology*, 146(3), pp.550–558.
- Salomon-Ferrer, R., Case, D.A. & Walker, R.C., (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2), pp.198–210.
- Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S. & Walker, R.C., (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of Chemical Theory and Computation*, 9(9), pp.3878–3888.
- Salter, N.W., Ande, S.R., Nguyen, H.K., Nyomba, B.L.G. & Mishra, S., (2012). Functional characterization of naturally occurring transglutaminase 2 mutants implicated in early-onset type 2 diabetes. *Journal of Molecular Endocrinology*, 48(3), pp.203–216.

- Sarang, Z., Mádi, A., Koy, C., Varga, S., Glocker, M.O., Ucker, D.S., et al., (2007). Tissue transglutaminase (TG2) facilitates phosphatidylserine exposure and calpain activity in calcium-induced death of erythrocytes. *Cell Death & Differentiation*, 14(10), pp.1842–1844.
- Schaertl, S., Prime, M., Wityak, J., Dominguez, C., Munoz-Sanjuan, I., Pacifici, R.E., et al., (2010). A profiling platform for the characterization of transglutaminase 2 (TG2) inhibitors. *Journal of Biomolecular Screening*, 15(5), pp.478–487.
- Schmidt, M.W., Baldrige, K.K., Boatz, J.A., Elbert, S.T., Gordon, M.S., Jensen, J.H., et al., (1993). General atomic and molecular electronic structure system. *Journal of Computational Chemistry*, 14(11), pp.1347–1363.
- Schuppan, D., Junker, Y. & Barisani, D., (2009). Celiac disease: From pathogenesis to novel therapies. *Gastroenterology*, 137(6), pp.1912–1933.
- Seabra, G. de M., Walker, R.C., Elstner, M., Case, D.A. & Roitberg, A.E., (2007). Implementation of the SCC-DFTB method for hybrid QM/MM simulations within the Amber Molecular Dynamics Package. *The Journal of Physical Chemistry A*, 111(26), pp.5655–5664.
- Seabra, G. de M., Walker, R.C. & Roitberg, A.E., (2009). Are current semiempirical methods better than force fields? A study from the thermodynamics perspective. *The Journal of Physical Chemistry A*, 113(43), pp.11938–11948.
- Shkurti, A., Goni, R., Andrio, P., Breitmoser, E., Bethune, I., Orozco, M., et al., (2016). pyPcazip: A PCA-based toolkit for compression and analysis of molecular simulation data. *SoftwareX*, 5, pp.44–50.
- Siegel, M. & Khosla, C., (2007). Transglutaminase 2 inhibitors and their therapeutic role in disease states. *Pharmacology & Therapeutics*, 115(2), pp.232–245.
- Siegel, M., Strnad, P., Watts, R.E., Choi, K., Jabri, B., Omary, M.B., et al., (2008). Extracellular transglutaminase 2 is catalytically inactive, but is transiently activated upon tissue injury. *PLoS ONE*, 3(3), p.e1861.
- Silva, J.R. a., Govender, T., Maguire, G.E.M., Kruger, H.G., Lameira, J., Roitberg, A.E., et al., (2015). Simulating the inhibition reaction of *Mycobacterium tuberculosis* L,D-transpeptidase 2 by carbapenems. *Chemical Communications*, 51(63), pp.12560–12562.
- Sinnott, M., (2007). *Carbohydrate Chemistry and Biochemistry: Structure and Mechanism* 1<sup>st</sup> ed., Cambridge: RSC Publishing.
- Skjaerven, L., Martinez, A. & Reuter, N., (2011). Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins*, 79(1), pp.232–243.
- Smith, M.B. & March, J., (2001). *March's Advanced Organic Chemistry: Reactions, Mechanisms, and Structure* 5<sup>th</sup> ed., New York: Wiley-Interscience.

- Sohn, J., Kim, T.-I., Yoon, Y.-H., Kim, J.-Y. & Kim, S.-Y., (2003). Novel transglutaminase inhibitors reverse the inflammation of allergic conjunctivitis. *The Journal of Clinical Investigation*, 111(1), pp.121–128.
- Sollid, L.M. & Khosla, C., (2011). Novel therapies for coeliac disease. *Journal of Internal Medicine*, 269(6), pp.604–613.
- Souaille, M. & Roux, B., (2001). Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1), pp.40–57.
- Sousa, S.F., Fernandes, P.A. & Ramos, M.J., (2006). Protein–ligand docking: Current status and future challenges. *Proteins: Structure, Function, and Bioinformatics*, 65(1), pp.15–26.
- Spiwok, V., Sucur, Z. & Hosek, P., (2015). Enhanced sampling techniques in biomolecular simulations. *Biotechnology Advances*, 33(6), pp.1130–1140.
- Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P. a. & Case, D. a., (1998). Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *Journal of the American Chemical Society*, 120(37), pp.9401–9409.
- Stewart, J.J.P., (2002). MOPAC version 2.5.3, Fujitsu Limited, Tokyo, Japan.
- Stewart, J.J.P., (1989). Optimization of parameters for semiempirical methods I. Method. *Journal of Computational Chemistry*, 10(2), pp.209–220.
- Stewart, J.J.P., (2007). Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling*, 13(12), pp.1173–1213.
- Stote, R., Dejaegere, A., Kuznetsov, D. & Falquet, L., (1999). Setting up and running a molecular dynamics simulations. Available at: [http://www.ch.embnet.org/MD\\_tutorial/](http://www.ch.embnet.org/MD_tutorial/).
- Suan Li, M. & Khanh Mai, B., (2012). Steered molecular dynamics-A promising tool for drug design. *Current Bioinformatics*, 7(4), pp.342–351.
- Tagami, U., Shimba, N., Nakamura, M., Yokoyama, K.-I., Suzuki, E.-I. & Hirokawa, T., (2009). Substrate specificity of microbial transglutaminase as revealed by three-dimensional docking simulation and mutagenesis. *Protein Engineering, Design & Selection*, 22(12), pp.747–752.
- Takai, T., Senda, H., Lee, H.-H., Kuwae, A. & Hanai, K., (1998). Keto-enol tautomerism of mono-substituted phenylpyruvic acids as studied by NMR and PM3 calculation. *Spectroscopy Letters*, 31(2), pp.379–395.
- Taylor, R.D., Jewsbury, P.J. & Essex, J.W., (2002). A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design*, 16(3), pp.151–166.
- Tosco, A., Aitoro, R., Auricchio, R., Ponticelli, D., Miele, E., Paparo, F., et al., (2013). Intestinal anti-tissue transglutaminase antibodies in potential coeliac disease. *Clinical and Experimental Immunology*, 171(1), pp.69–75.

- Tuckerman, M.E. & Martyna, G.J., (2000). Understanding modern molecular dynamics: Techniques and applications. *The Journal of Physical Chemistry B*, 104(2), pp.159–178.
- Verdonk, M.L., Chessari, G., Cole, J.C., Hartshorn, M.J., Murray, C.W., Nissink, J.W.M., et al., (2005). Modeling water molecules in protein-ligand docking using GOLD. *Journal of Medicinal Chemistry*, 48(20), pp.6504–6515.
- Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W. & Taylor, R.D., (2003). Improved protein-ligand docking using GOLD. *Proteins*, 52(4), pp.609–623.
- Verma, A., Wang, H., Manavathi, B., Fok, J.Y., Mann, A.P., Kumar, R., et al., (2006). Increased expression of tissue transglutaminase in pancreatic ductal adenocarcinoma and its implications in drug resistance and metastasis. *Cancer Research*, 66(21), pp.10525–10533.
- Veza, R., Habib, A. & Fitzgerald, G.A., (1999). Differential signaling by the thromboxane receptor isoforms via the novel GTP-binding protein, Gh. *The Journal of Biological Chemistry*, 274(18), pp.12774–12779.
- Vinter, J.G. & Gardner, M., (1994). *Molecular Modelling and Drug Design* 1<sup>st</sup> ed., New York: CRC Press.
- Vreven, T., Byun, K.S., Komáromi, I., Dapprich, S., Montgomery, J.A., Morokuma, K., et al., (2006). Combining quantum mechanics methods with molecular mechanics methods in ONIOM. *Journal of Chemical Theory and Computation*, 2(3), pp.815–826.
- Walker, R.C. & Cheatham, T.E., (2010). Amber Tutorials. Available at: <http://ambermd.org/tutorials/basic/tutorial1/>.
- Walker, R.C., Crowley, M.F. & Case, D.A., (2008). The implementation of a fast and accurate QM/MM potential method in Amber. *Journal of Computational Chemistry*, 29(7), pp.1019–1031.
- Wang, D.-S., Dickson, D.W. & Malter, J.S., (2008). Tissue transglutaminase, protein cross-linking and Alzheimer's disease: Review and views. *International Journal of Clinical and Experimental Pathology*, 1(1), pp.5–18.
- Wang, J., Wang, J.L., Yang, X., Xia, K., Hu, Z.M., Weng, L., et al., (2010). TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain*, 133(12), pp.3510–3518.
- Wang, J., Wang, W., Kollman, P.A. & Case, D.A., (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics & Modelling*, 25(2), pp.247–260.
- Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. & Case, D.A., (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9), pp.1157–1174.
- Wang, Z., Collighan, R.J., Pytel, K., Rathbone, D.L., Li, X. & Griffin, M., (2012). Characterization of heparin-binding site of tissue transglutaminase: Its importance in

- cell surface targeting, matrix deposition, and cell signaling. *The Journal of Biological Chemistry*, 287(16), pp.13063–13083.
- Wang, Z. & Griffin, M., (2012). TG2, a novel extracellular protein with multiple functions. *Amino Acids*, 42(2), pp.939–949.
- Watts, R.E., Siegel, M. & Khosla, C., (2006). Structure-activity relationship analysis of the selective inhibition of transglutaminase 2 by dihydroisoxazoles. *Journal of Medicinal Chemistry*, 49(25), pp.7493–7501.
- Wilhelmus, M.M.M., Grunberg, S.C.S., Bol, J.G.J.M., Van Dam, A.M., Hoozemans, J.J.M., Rozemuller, A.J.M., et al., (2009). Transglutaminases and transglutaminase-catalyzed cross-links colocalize with the pathological lesions in Alzheimer's disease brain. *Brain Pathology*, 19(4), pp.612–622.
- Yang, C.Y., (2015). Identification of potential small molecule allosteric modulator sites on IL-1R1 ectodomain using accelerated conformational sampling method. *PLoS ONE*, 10(2), p.e0118671.
- Yang, E.-C., Hao, J.-K., Tang, T.-H., Wang, B.-W. & Cao, Y.-Y., (2003). The theoretical study of reaction paths and transition states on coupling reaction of methane through plasma. *Journal of Molecular Structure: THEOCHEM*, 626, pp.121–126.
- Yang, S., (2008). *Computer aided design of FabI inhibitors and structural models of a complex between tubulin and Gs-alpha*. University of Illinois.
- Yoshimura, K., Miyazaki, H., Matsuo, S., Hashimoto, H., Hori, T., Yamaguchi, Y., et al., (2012). A theoretical study on synthesis mechanism of *fac*-[Ir(ppy)<sub>3</sub>] from [Ir(acac)<sub>3</sub>] in the presence of Brønsted acids and water molecules. *Bulletin of the Chemical Society of Japan*, 85(2), pp.209–216.
- Young, D., (2001). *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*, New York: Wiley–Interscience.