IDENTIFICATION OF SANCDB COMPOUNDS AGAINST G2019S AND I2020T VARIANTS OF LEUCINE-RICH REPEAT KINASE 2 (LRRK2) FOR THE DEVELOPMENT OF DRUGS AGAINST PARKINSON'S DISEASE

A thesis submitted in partial fulfilment of the requirements for the degree

Of

Master of Science in Bioinformatics and Computational Molecular Biology (Coursework and Thesis)

Of

RHODES UNIVERSITY, SOUTH AFRICA
Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY AND MICROBIOLOGY
Faculty of science

By

BERTHA BAYE
18B3036

JANUARY 2019

DECLARATION

I, **Bertha Cinthia Baye**, declare that this thesis entitled, the identification of SANCDB compounds against the G2019S and I2020T variants of Leucine-rich repeat kinase 2 (LRRK2) for the development of drugs against Parkinson's disease, submitted to Rhodes University is solely my own research work. I have acknowledged all authors' concepts and referenced direct quotations from their works. I also declare that this thesis has never been submitted to any different institution for whatever degree.

Signature …….

Date…….

ACKNOWLEDGEMENTS

I would like to take full advantage of this platform and show my gratitude to the people who helped me in this work. I would like to thank my wonderful supervisor Dr. Vuyani Moses for guiding me and making sure this program went smoothly and that l complete this MSc. thesis in due time. I also would like to thank Prof Özlem Tastan Bishop for the opportunity she granted me to pursue this degree and for her strong spirit and encouraging words.

I would like to thank my big family as well, my father (Denford Baye) and mother (Ruth Baye) for their financial support and encouragements, l also would like to thank my siblings Vanessa, Nicollette, Adolyn and Denford Junior for their phone calls and laughter. A special thank you goes to Malesela Nong for always being there for me, believing in me and making sure l stay on my toes. I would like to thank my best friend Dakalo Nemauluma for the friendship and proofreading my work.

I also want to thank the National Research Foundation (NRF) for funding my studies throughout my one-year MSc. degree. This work is supported by the National Research Foundation (NRF) South Africa (Grant Number 105267) allocated to Prof Özlem Tastan Bishop.

To the lecturers that taught me during coursework of this program, I am truly grateful it is your dedication and leadership that made this all possible. A big thank you to Dr. Rowan Hatherley, Dr. Vuyani Moses, Mr. Jeremy Daxter, Dr. Kevin A. Lobb, and Ms. Caroline Ross.

I would like to express my gratitude to all Research Unit in Bioinformatics (RUBi) members for all the help and friendship especially to Arnold, Phillip, Thommas, Jessica, Margaret, and Allan, it was truly a great pleasure getting to know all of you.

Above all, I would like to thank the Lord Almighty for being my strength and guide, for keeping His eye upon me without blinking and blessing my journey.

ABSTRACT

Parkinson's disease is a type of movement disorder that occurs when nerve cells in the brain stop producing dopamine. It is the second neurodegenerative disease affecting 1-2% of people above the ages of 65 years old. There is a worldwide prevalence of 7 to 10 million affected people of all cultures and race. Studies have shown that mutation that causes Parkinson's disease result in increased kinase activity. The c.6055 G > A in exon 41 is the most prevalent LRRK2 variation which causes a substitution of glycine to serine in G2019S in the highly activated loop of its MAP kinase domain. The LRRK2 G2019S variant is the most common genetic determinant of Parkinson's disease identified to date. This work focused on building accurate 3D models of the LRRK2 kinase domain, that were used for large-scale *in silico* docking against South African natural compounds from the South African Natural Compounds Database (SANCDB; https://sancdb.rubi.ru.ac.za/). Molecular docking was performed to identify compounds that formed interactions with the active site of the protein and had the lowest binding energy scores. Molecular dynamics simulations showed different movements of the protein-ligand complexes and behavioural difference of the wildtype and the variants, all three structures proved to be compact. Network analysis was done to study residue interactions, contact maps, dynamic cross correlations, average BC and average L were used to study the residue interactions and general residue contribution to the functioning of the protein.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

TABLE OF AMINO ACIDS

Amino acid name	Three letter code	One letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

LIST OF WEBSERVES, SOFTWARE TOOLS AND CLUSTERS

AutoDock Vina……. vina.scripps.edu

BioEdit……www.mbio.ncsu.edu

DISCOVERY STUDIO VISUALISER 4……..www.3dsbiovia.com

Jalview…….…www.jalview.org

LIGPLOTPLUS ……..https://www.ebi.ac.uk

MODELLER v9.19…….…. https://salilab.org/modeler

NCBI…….…..... https://blast.ncbi.nlm.nih.gov/Blast.cgi

PRIMO…….. https://primo.rubi.ru.ac.za

PROCHECK……...…….…https://www.ebi.ac.uk

PROSA…….https://prosa.services.came.sbg.ac.at/prosa.php

PYMOL……...https://pymol.org/2/

QMEAN……..https://swissmodel.expasy.org/qmean

T-COFFEE……...……...…. https://www.tcoffee.org

UNIPROT……...…. https://www.uniprot.org

CARBON……...

CHPC……. https://users.chpc.ac.za

YODA…….

LIST OF ABBREVIATIONS AND ACRONYMS

AMPK	5' AMP-activated protein kinase
CaMKK	Components of a Calmodulin-dependent protein kinase cascade
CNS	Central Nervous System
G2019S	Gly2019Ser
I2020T	Ile2020Thr
LRR	Leucine-Rich Repeat
LRRK2	Leucine-rich repeat kinase 2
MD	Molecular Dynamics
MPP	Myelin Protein Peripheral
NAADP	Nicotinic Acid Adenine Dinucleotide Phosphate
OHDA	Hydroxy Dopamine
PD	Parkinson’s disease
RG	Radius of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SANCDB	South African Natural Compound Database

CHAPTER ONE: Introduction

1.1 Parkinson's disease

Parkinson's disease is the second most common neurodegenerative disease affecting 1-2% of people above the ages of 65 years old (Gilsbach and Kortholt 2014). Parkinson's disease is a type of movement disorder that results when nerve cells in the substantia nigra of the midbrain stop producing enough chemical called dopamine. This disease is an autosomal dominant and genetically transmitted and it is estimated that 15-25% of people with Parkinson's disease have relatives with the disease too, however, exposure to certain chemicals such as pesticides and herbicides used in farming and traffic or industrial pollution might also contribute to the condition (Fox and Reno 2014; Rewar 2015).

Parkinson's disease is a sporadic condition of unknown causes in most patients, however in other cases the disease is inherited as a highly penetrant Mendelian trait (Di Fonzo *et al.* 2006). Mutations that occur in the following genes SNCA (α-synuclein), PARK2 (parkin), PARK7 (DJ-1), PINK1 and LRRK2 (Leucine-rich repeat kinase 2) can cause a Parkinsonism that resembles idiopathic Parkinson's disease and mutations in the SNCA and the LRRK2 genes causes autosomal dominant forms whereas the rest of the genes causes autosomal recessive forms (Di Fonzo *et al.* 2006; Healy *et al.* 2008). There are two mutations found in the Roc domain, one found in the COR domain, and two found in the kinase domain, two variants that act as risk factors for sporadic Parkinson's disease have been identified, one in the COR domain and one in the WD40 repeats (Cookson 2016).

Genetic factors play a role in disease's pathogenesis, in which the LRRK2 mutations are associated with autosomal dominant with incomplete penetrance which produces α-synuclein-type neuropathology and the G2019S and I2020T are the mutant variants (Kestenbaum and Alcalay 2017; Xiong *et al.* 2017). The mitochondrial variants and haplogroups are thought to modify the risk of developing Parkinson's disease by decreasing or increasing the penetrance of genetic variants in nuclear disease-genes also referred to as PARK genes (García *et al.* 2014). In the catalytic core, Roc(COR) and the kinase domain mutations with confirmed disease co-segregation are found here, and this causes increased kinase and decreased GTPase activity (Guaitoli *et al.* 2016).

There is also drug induced Parkinsonism whereby after taking certain treatment such as antipsychotic medication symptoms of Parkinson's start appearing, these usually improve after medication has been stopped (Rewar 2015). There is a worldwide prevalence of 7 to 10 million affected people of all

cultures and race[1]. This chronic disease was named after Doctor James Parkinson who was the first to describe it in 1817 (Goetz 2011). Parkinson's disease is a progressive, degenerative neurological condition that results into a person losing control over their body movement with association of resting tremor, bradykinesia and muscular rigidity (Goldwurm *et al.* 2005; García *et al.* 2014). When a person has this disease, the brain loses control and co-ordination of movements of the muscles in different parts of the body. In 5-10% of people affected by Parkinson's disease they have an early onset that begins before the age of 50 of which this disease usually affects people around the age of 60. There is also a young onset Parkinson's in which people below the age of 50 are affected and there is also juvenile Parkinsonism in which symptoms appear in people below the age of 20.

Mostly, human males are more affected by Parkinson's disease than females (Keyser 2011). The disease starts off by affecting one part of the body, before both sides are affected. There is usually trembling (tremor) of the hands, arms, legs, jaws and the face, stiffness of the arms, legs, and trunk and then slowness of movement, poor balance and coordination (Fox and Reno 2014). Symptoms can get worse causing the patients to have trouble walking, talking or performing normal tasks. There are also other complications that come along such as depression, sleeping problems, and trouble to chew or swallow. Neurological examinations and a medical history check are used to diagnose this disease; there are no lab tests that can be done. A change in the DNA or genetic material that makes up a gene can cause a gene mutation; in this case mutation of the LRRK2 gene can cause Parkinson's disease(Liu *et al.* 2018).

The progressive degeneration of the nerve cells in the middle area of the brain causes the symptoms of Parkinson's disease, this results in the lack of dopamine. Dopamine acts as a chemical messenger necessary for smooth controlled movements and symptoms will only appear when about 70% of the chemical has stopped working normally. The biological function of LRRK2 and how it contributes to Parkinson's disease is not well understood (Guaitoli *et al.* 2016). Parkinson's disease belongs to a group of conditions called motor system disorders which occur due to loss of dopamine-producing brain cells (Rewar 2015). The substantia nigra sends messages down nerves in the spinal cord to help control the muscles of the body, signals are sent between the brain cells, nerves and the muscles by neurotransmitters (Cho *et al.* 2017). This disease is caused by a loss of nerve cells in part of the brain called the substantia nigra which is the nucleus in the midbrain that is considered part of the basal ganglia, most of the dopamine neurons originate in this area of the brain (Cho *et al.* 2017). Currently,

[1] www.parkinsonassociation.org

there are no published longitudinal studies discussing the disease's progression rate (Kestenbaum and Alcalay 2017).

Dopamine plays a vital role in regulating the movement of the body and a reduction in dopamine is responsible for many of the symptoms of Parkinson's disease (Keyser 2011). Low concentrations or lack of dopamine will cause the person to lose control of their movements. The Parkinson's disease symptoms are grouped as motor and non-motor. The non-motor symptoms are more challenging because they come with pain, depression, memory loss, and sleep disorders. Muscle stiffness can create tension in the tendon, leading to structural adjustment and postural instability is often experienced in the late stages of Parkinson's disease (Rewar 2015). 7 of the 80 missense variants in the LRRK2 gene are associated with dominantly inherited Parkinson's disease and these missense are found in functionally important sites with G2019S being the most prevalent (Landoulsi *et al.* 2017).

There is a Lewy body pathology association with Parkinson's disease in the majority of LRRK2 cases and a neurofibrillary tangle pathology similar to the one observed in progressive supranuclear palsy or TDP-43 pathology in minority cases (Price *et al.* 2018). LRRK2 plays a role in signal transduction and variants that occur on this protein will cause disquiets in the signal transduction that exists in the heart of the protein dysfunction in Parkinson's disease (Price *et al.* 2018). The two largest mutations in LRRK2 gene with highly variable frequencies depending on geographic location and ethnic group are G2019S and R1441 codon, p.G2385R and p.R1628P single nucleotide polymorphisms are population specific and are mainly found in Asians (Cornejo-Olivas *et al.* 2017).

(Mattea *et al.* 2018), have figured out a method of expressing and crystallising the central part of LRRK2 including the GTPase, Kinase domain and WD40 domain, however the crystals diffracted poorly even after comprehensive optimisation limiting the diffraction and they plan to crystallise the protein under microgravity conditions which improve the diffraction properties of protein crystals. Pathogenic mutations of LRRK2 are found on the catalytic core of LRRK2 suggesting a role in altering enzymatic function in Parkinson's disease pathogenesis, *in vitro* kinase assays based on protein autophosphorylation or phosphorylation of generic substrate showed consistent increase in kinase activity for G2019S mutation (De Wit, Baekelandt and Lobbestael 2018). So, as a result, there currently no available crystal structures for the kinase domain.

According to Eyers PA (Eyers 2018), there are many factors that contribute to the localised neuronal death that occurs in the substantia nigra of Parkinson's disease brain, these include protein folding/aggregation defects generating α-synuclein-rich Lewy bodies, abnormal protein

phosphorylation/ubiquitination, aberrant intracellular trafficking and oxidative stress coupled to mitochondrial dysfunction. To take advantage of druggable lesions in different patient groups, pharmacological interventions that counteract effects of dopamine depletion in Parkinson's disease would be beneficial (Eyers 2018). There are factors that contribute to the death of dopaminergic and non-dopaminergic cells in the brains of people suffering from Parkinson's disease which are genetic mutations, abnormal handling of misfolded proteins by the ubiquitin-proteasome and the autophagy-lysosomal systems, increased oxidative stress, mitochondrial dysfunction, inflammation and other pathogenic mechanisms (Jankovic 2008). Non-motor symptoms can arise from the disease, disturbing the quality of life and as most treatments are geared towards the management and relief of motor symptoms in patients, complications do arise as the disease progresses (Rana *et al.* 2015). Historically, development of therapy was based on the empirical observations to rational designs based on growing knowledge of anatomy, biochemistry, and physiology of the basal ganglia (Goetz 2011). There is a unique 280 kDa protein that belongs to the ROCO protein family and is encoded by LRRK2 gene that present guanosine triphosphate (GTPase) and C-terminal of Ras (COR) domain most often in conjunction with a kinase domain (West *et al.* 2007).

Most Parkinson's disease cases are idiopathic, to advance knowledge in the aetiology of idiopathic disease and help identify targets for the development of biomarker and novel treatments, one has to study the disease mechanism (Mir *et al.* 2018). This neurodegenerative disorder is characterized by insoluble proteinaceous Lewy body and Lewy neurite inclusions that accumulate within multiple affected neuronal populations and they primarily contain α-synuclein (α-syn) which are the small aggregate-prone protein that plays a central role in idiopathic and familial forms of Parkinson's disease (Schapansky *et al.* 2018). Common themes have been identified in mechanisms of Parkinson's disease pathology including protein degradation, mitochondria and apoptotic pathways (Gao *et al.* 2018). 28 genetic risk variants at 24 loci with nonfamilial Parkinson's disease where mentioned in Genome-wide association studies (GWAS), in which LRRK2 is hereditary pinpointing a shared molecular pathway driving pathogenesis in both familial and non-familial Parkinson's disease (Steger *et al.* 2016).

Parkinson's disease incidence increases with age from 1% to 2% of people older than 60 years up to 4% of people older than 80 years (De Wit, Baekelandt and Lobbestael 2018). Evidence has proved that the abnormal regulation of gene expression and protein translation may contribute to Parkinson's disease development and duplication or triplication of the α-synuclein gene that then increases substance protein levels can cause genetic Parkinson's disease (Dorval and Hébert 2012). D1, D2, D3, D4, and D5 are the five subtypes of dopamine receptors and they form part of the large G-protein

coupled receptor superfamily (Ayano 2016). D1 and D5 are similar in structure and drug sensitivity and are referred to as “D1like” class of receptors, D2, D3, and D4 are also similar in structure and are known as “D2like” group (Ayano 2016). Dopamine is produced from the amino acid tyrosine which is produced in the liver from phenylalanine through the phenylalanine hydroxylase process and taken to the brain through an active transport mechanism (Ayano 2016). Tyrosine is then converted to dopamine in dopamine-containing neurons.

1.1.1 Diagnosis

A proper and correct diagnosis should be done in order for the patient to get the required counselling and therapeutic management (Berardelli *et al.* 2013). Having one of the symptoms is no reason for concern but if these small changes like handwriting, loss of smell, facial masking, stooped posture, slowed and stiff movements, tremor, frozen shoulder, change in voice and sleep disturbances occur then one should seek help.

Magnetic resonance imaging uses magnetic currents to create images of the brain and computerised tomography includes a series of X-rays that are passed through different directions providing the anatomical view of the brain, these scans can help detect loss of dopamine in the brain (Rewar 2015). DaT Scan (Dopamine transporter Scan) is an FDA approved imaging technique that helps view the dopamine system in the brain, a radioactive dye is injected into the body and then the dye binds to dopamine-releasing neurons, the specialised camera records the signals. A low signal shows that there are few dopamine-producing neurons itself to diagnose the disease (Rewar 2015). Genetic testing in a form of a fast and accurate genotyping method is done to screen G2019S and I2020T mutations in people who might be having Parkinson's disease with real-time polymerase chain reaction (PCR) including TaqMan assay and high resolution melting technique being some of the early genotyping routines that are done in diagnosis (Landoulsi *et al.* 2017). There are different clinical diagnostic criteria used for Parkinson's disease diagnosis which are classifications of motor signs combined with the absence of incompatible or atypical signs that were introduced by Queen Square Brain Bank (QSBB) (Berardelli *et al.* 2013). The UK Parkinson's disease Society Brain Bank (UKPDSBB) and the National Institute of Neurological Disorders and Stroke (NINDS), have implemented a diagnosis based on the identification of Parkinsonian symptoms, the absence of exclusion criteria and the presence of positive criteria (Lingor *et al.* 1999).

1.1.2 Current treatment

There is currently no cure for Parkinson's disease, and it is the second most common neurodegenerative disease in the world. Different medications are used to help the symptoms of Parkinson's disease whereas, surgery and deep brain stimulation can help severe cases. Modern treatment development projects start with the identification of a target protein and validation of its involvement in the disease progress (Guaitoli *et al.* 2016). However, there are treatments to make symptoms a little bearable (Rewar 2015). Recently anti-Parkinson drug safinamide has been approved and it inhibits monoamine oxidase B also including sodium and calcium channels that contribute to its unique properties. Treatment of Parkinson's disease depends on dopamine precursor levodopa, to overcome the decrease in dopamine levels in the brain (Jaiteh *et al.* 2018).

Levodopa is combined together with other medication such as benserazide or carbidopa to stop it from being broken down in the bloodstream before it gets to the brain, the nerve cells of the brain absorb it and convert it to dopamine (Keyser 2011). There is still a great need for effective treatments because the levodopa gradually losses its efficiency and causes side effects such as dyskinesia. Non-dopaminergic members of the GPCR superfamily and enzymes involved in the degradation of monoamine neurotransmitters are alternative targets for the development of anti-Parkinson drugs (Jaiteh *et al.* 2018). The A2A antagonist 8-(3-chlorostyryl) caffeine (CSC) showed promising neuroprotective effects in experimental Parkinson's disease models. However, it has low solubility and sensitivity to light-induced degradation making it undesirable.

The following can be exercised to improve way of living: physiotherapy (helps relieve muscle stiffness and joint pain), occupational therapy (ensure home is safe and properly set up), speech therapy (improves the clarity and volume of speech), supportive therapy (help one cope with the disease) and medication (improve the main symptoms of the disease) (Rewar 2015).

Some studies show that higher concentrations of urate decrease the risk of having Parkinson's disease. Urate circulates in high concentrations in human plasma and it is considered a powerful antioxidant. Laboratory experiments have shown that urate attenuates MPP+ toxicity in dopaminergic neurons and 6-OHDA on dopaminergic cells (Hughes *et al.* 2018). In the pathology of Parkinson's disease there is oxidative stress and urate is a powerful antioxidant responsible for antioxidant capacity in human plasma, therefore high urate levels in the plasma could be protective against Parkinson's disease and prospective studies show an increased risk of developing Parkinson's disease in people with lower urate levels (Hughes *et al.* 2018). Studies have shown that people with low urate levels have a high risk of developing Parkinson's disease (Hughes *et al.* 2018).

According to (Melrose 2008), in L-DOPA induced dyskinetic marmosets after 1-methyl-4-phenyl-1,2,3,6-tetra hydro pyridine (MPTP) treatment, there is increased LRRK2 mRNA. Between membranous and membrane-bound organelles and LRRK2, there is an association that is believed to play a crucial regulatory role in synaptic function, possibly by regulation of vesicle synthesis or transport regulation of membranous structures (Melrose 2008). Met-analysis treatment with COMT inhibitors combined with L-dopa proved to have significant control in comparison to L-dopa alone, however over 7 months results are lacking and hepatotoxicity is rare with potential lethal; side effects associated with tolcapone (Levine *et al.* 2003).

Surgery is usually performed on patients that respond to medication but suffer severe side effects, the procedure includes pallidotomy or thalamotomy, deep brain stimulation in which an electrode is placed in the globus pallidus, thalamus or subthalamic nucleus to stimulate their function and tissue transplant (Levine *et al.* 2003). An oral supplement called Coenzyme Q10 an essential co-factor in the electron transport chain is suggested to have neuroprotective properties that can delay or slow Parkinson's disease progression (Grosset, Macphee and Nairn 2010). The medication that is currently available does not halt or hinder the progression of the disease but only provide symptomatic relief and usually has undesirable side effects (Gao *et al.* 2018).

1.2 LRRK2 protein

According to (Landoulsi *et al.* 2017), the LRRK2 gene is located on chromosome 12q12 with 51 exons, it encodes a highly conserved 2527 amino acid multi-domain protein having kinase and GTPase functions. The c.6055 G > A in exon 41 is the most prevalent LRRK2 mutation which causes a substitution of glycine to serine in G2019S in the highly activated loop of its MAP kinase domain (Mata *et al.* 2006).

The LRRK2 mutations are associated with autosomal dominant Parkinson's disease with incomplete penetrance, this penetrance varies among variants and mutations with Gly2385Arg variant the penetrance is very low, among carriers of lle2020Thr mutation it is very high and other studies have shown that Gly2019Ser mutation is very controversial ranging from 24-100% risk (Healy *et al.* 2008; Kestenbaum and Alcalay 2017). LRRK2 is one of the two vertebrate LRRKs which show complementary expression in the brain. The role of LRRK2 kinase domain in regulating intrinsic GTPase activity is not yet known and it is known that intrinsic GTPase activity may modulate kinase activity and neurotoxicity (West *et al.* 2007). LRRK2 functions primarily as a scaffolding protein and it contain serine-threonine phosphorylation (kinase activity) and guanine triphosphate hydrolysis (GTPase activity) enzyme activities, leading to the suggestion that this might be a conventional signalling protein in GTPases Ras or Rac or functions as a protein kinase (Berwick and Harvey 2013).

The LRRK2 gene is located on chromosome number 12 positions 40.618.933 to 40.761.567 and it has the length coding sequence of 7581 nucleotides and has a Genbank ID of NM_198578 (RCSB PDB)[2]. The LRRK2 gene positively regulates autophagy through a calcium-dependent activation of the CaMKK/AMPK signalling pathway. The process involves activation of nicotinic acid adenine dinucleotide phosphate (NAADP) receptors, an increase in lysosomal pH, and calcium release from

[2] http://www.rcsb.org/pdb/protein/Q17RV3?evtc=Suggest&evta=ProteinFeature%20View&evtl=autosearch_SearchBar_querySuggest

lysosomes (UniProtKB). Together with RAB29, plays a role in the retrograde trafficking pathway for recycling proteins, such as mannose 6 phosphate receptor M6PR, between lysosomes and the Golgi apparatus in a retromer-dependent manner. LRRK2 regulates neuronal process morphology in the intact central nervous system (CNS) and plays a role in synaptic vesicle trafficking (Islam *et al.* 2016). Phosphorylates PRDX3, has GTPase activity may play a role in the phosphorylation of proteins central to Parkinson disease.

A mutation of the gene leucine-rich repeat kinase 2 (LRRK2) which contains both kinase and GTPase domains whose activity is critical for signal transduction and will cause Parkinson's disease (Ho *et al.* 2015). Pharmacological inhibition of kinase activity helps pathogenic phenotypes such as neurocytotoxicity and defective neurite outgrowth; G2019S pathogenic mutation increases kinase activity and affects various pathogenic phenotypes such as increased neuronal cytotoxicity and protein aggregation, decreased neurite length and changes in the autophagy rate (Ho *et al.* 2015). Due to the relevance of LRRK2 kinase activity to Parkinson's disease pathogenesis, this makes the protein a therapeutic target. According to (Cookson 2016), variation around LRRK2 includes a relatively common variant G2019S that increases the risk of Parkinson's disease and there are also non-coding variants that increase the risk of sporadic Parkinson's disease. Studies show that LRRK2 plays a role in vesicle trafficking, autophagy, mitochondrial dysfunction, and inflammation; it is also a translational regulator (Ho *et al.* 2015).

Age-related factors may play a role in contributing to Parkinson's disease pathogenesis, the advanced glycation end products (AGEs) that arise from the reaction of sugars with certain amino acids or fats and their interaction with RAGE (receptor of AGEs) is involved in the G2019S mutation progressive neuronal loss (Cho, Xie and Cai 2018). Elevated kinase activity resulting from the G2019S mutation of the LRRK2 protein was observed when the kinetic kinase assays that measure phosphorylation of generic peptide substrates was measured (Liu *et al.* 2014). The causative mutations such as G2019S localize to the enzymatic core of LRRK2 spanning the GTPase to kinase domain and alter LRRK2 enzymatic function, this increases the kinase activity with mutations in the ROC/COR domains demonstrating reduced GTPase activity (Price *et al.* 2018).

The LRRK2 protein has been linked to cancer with the Parkinson's disease G2019S mutation associated with specific cancers such as non-skin and hormonal cancers (Agalliu *et al.* 2015). Many studies have tried to investigate the role of both rare and common LRRK2 variants since the discovery of LRRK2 mutations as a major cause of autosomal dominant Parkinsonism in the neurodegenerative disorders and the pathophysiology of Parkinson's disease (Purlyte *et al.* 2017; Ng *et al.* 2018).

LRRK2 is a large, multi-domain protein kinase that consists of an armadillo repeat domain (residues 150-510), an Ankyrin domain (residues 690-860), leucine-rich repeats (residues 984-1278), a ROC-type GTPase domain (residues 1335-1510) that closely resembles a Rab GTPase and is associated with the COR domain (C-terminal of Roc, residues 1511-1878), a serine-threonine protein kinase domain (residues 1879-2138) (Purlyte *et al.* 2017b).

LRRK2 regulates intracellular vesicular traffic, however, when there are variations on the protein, trafficking defects which contribute to Parkinson's disease pathogenesis. LRRK2 translocates to membranes following Toll-like receptor stimulation of immune cells, secondly, LRRK2 regulates the autophagosome/lysosome system and thirdly LRKK2 phosphorylates three Rab GTPase (Rab8A, Rab10, and Rab12) this reduces their binding affinity for different Rab regulatory proteins (Steger *et al.* 2017). It has been observed that the Gly2019 residue is extremely conserved in the human kinase domains and also in all dardarin homologs, this supports the pathogenic role of the p.G2019S mutation (Di Fonzo *et al.* 2006).
A small N-terminal lobe and a larger C-terminal lobe are connected by a hinge-like region in catalytic domains of protein kinases to form a cleft in which Mg_{2+}C-ATP and the protein substrate bind (Mata *et al.* 2006). G2019S is found in the highly conserved activation loop of the kinase domain and this variation occurs in 3-6% of familial Parkinson's disease and in 1-2% of sporadic Parkinson's disease globally (Landoulsi *et al.* 2017). Between the conserved tripeptide motifs, DF/DG and APE is an active segment of about 20-35 residues within the large C-terminal lobe and most protein kinases require phosphorylation of the activation segment for activity. The activation segment is believed to adopt an active conformation upon phosphorylation enabling substrate access and catalysis to take place. The kinase activity of LRRK2 regulates its function, however, it is still not well known whether increased kinase activity is responsible for pathogenesis in PARK8 patients (Berwick and Harvey 2013).

According to the model created by Guaitoli G et al (Guaitoli *et al.* 2016), N-terminal ankyrin and LRR repeats are involved in the intramolecular regulation of the biological activity of the protein. Changes in the gene expression networks that occur during aging and mis-regulation of the pathways can cause serious biological consequences and understating LRRK2 gene expression is of importance due to its role in normal and pathological aging processes (Dorval and Hébert 2012). The penetrance of LRRK2 G2019S is incomplete and age-dependent hence it is also suggested that age-related factors could also contribute to G2019S LRRK2-linked Parkinson's disease pathogenesis (Cho, Xie and Cai 2018). Between the ankyrin domain and leucine-rich repeat region is a Ser residue (Ser910, Ser935, Ser955, Ser973) where LRRK2 is phosphorylated, which plays a role in regulating binding and

cytosolic localization (Purlyte *et al.* 2017b). According to (Rassu *et al.* 2017), LRRK2 plays a part in control of DRD1 and DRD2 trafficking both in cellular and animal models. It is of particular interest to understand the effects of LRRK2 on gene expression due to its role in both normal and pathological aging processes (Dorval and Hébert 2012). Figure 1.1 below shows the LRRK2 kinase domain model and the kinase domain represented in blue, the active site consists of the following residues DYGIAQYCCRMGIKTSEGTPGRAPE from residue, 2017G until 2042E.

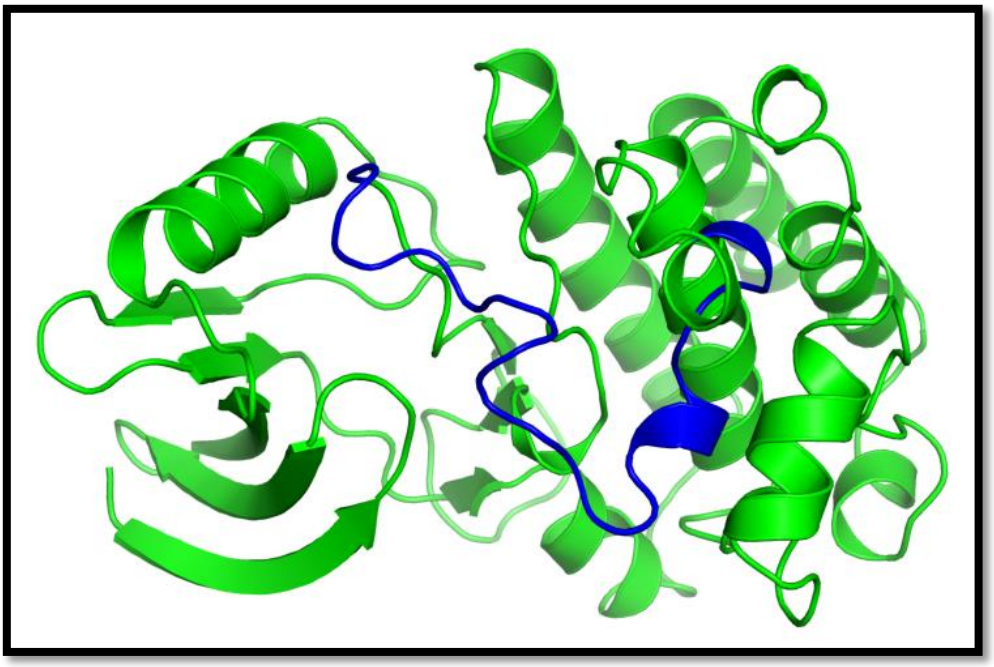


Figure 1. 1: LRRK2 kinase domain model and the blue part is the active site viewed in PyMOL.

1.2.1 Structural information and virtual screening

There currently is no LRRK2 crystal structure in the databases and hence there was need for modeling the structure specifically the domain of interest, figure 1.1 shows the 3D structure of LRRK2 kinase domain and the active site of the protein is highlighted by the blue colour. The LRRK2 protein, also termed dardarin, belongs to the ROCO group within the Ras/GTPase superfamily, characterized by the presence of several conserved domains: a Roc (Ras in complex proteins) and a COR (C-terminal of Roc) domain, together with a leucine-rich repeat region, a WD40 domain, and a protein kinase catalytic domain (Di Fonzo *et al.* 2006). The role of the LRRK2 kinase domain in regulating intrinsic GTPase activity which may modulate kinase activity and neurotoxicity is not yet known. The study of missense mutations linked to the development of the disease will help us understand the pathogenicity of LRRK2 and whether kinase activity or other activities present a viable therapeutic target (West *et al.* 2007). GTPase and kinase domains are the two enzymatic domains of the LRRK2 protein (Xiong *et al.* 2017) and the N-terminal Ankyrin, armadillo and namesake leucine-rich repeat

(LRR) together with the C-terminal WD40 domain are the four predicated solenoid domains that are involved in protein-protein interactions (Guaitoli *et al.* 2016b; Price *et al.* 2018). The arrays of pathways and interactions with which LRRK2 are linked is partly due to the intrinsic properties of the protein. Ankyrin repeats are found in molecules with cell-cell signaling functions and typically facilitate protein interactions or help in protein recognition (Price *et al.* 2018). N1437H, R1441C, R1441G, R1441S, Y1699C, G2019S and I1699C are the least missense variants that are pathogenic whilst fifty remains of undetermined significance (Cornejo-Olivas *et al.* 2017; Steger *et al.* 2017). The S1292 autophosphorylation site is thought to be having a physiological and direct marker of LRRK2 kinase activity (Kluss *et al.* 2018). To understand the dynamics and behaviour of protein-ligand interactions and predict the effect of variations on function and structure of different proteins molecular dynamic simulation technique is used (Illinois). (Kumar *et al.* 2019), used GROMACS 5.14 package to perform a molecular dynamics analysis of the native and the G2029S variant kinase domain of LRRK2 protein. (Bhayye, Roy and Saha 2018), carried out a molecular dynamic simulation on both the wildtype and variant LRRK2 kinase domain with and without ATP bound to the active site using the Desmond package.

According to (Xiong *et al.* 2017), increased kinase activity for autophosphorylation and hyperphosphorylation of LRRK2 kinase substrates prevailed in the disease that caused LRRK2 GS mutation and the LRRK2 kinase inhibitors G2019S/D1994A mutants reduced the LRRK2 GS-mediated toxicity. The Roco4 G1179S kinase structure revealed a mechanism for the activation of G2019S mutation in the LRRK2 and humanised versions of this structure version of the structure support the structure and assessment of LRRK2 specific inhibitors (Guaitoli *et al.* 2016a). The catalytical domain of the Serine-Threonine kinase, Leucine-Rich Repeat Kinase 2 STKs catalyse the transfer of the gamma-phosphoryl group from ATP to serine-threonine residues on protein substrates.

1.2.2 G2019S variant

The G2019S mutation which is the most prevalent mutation which is in the activation loop of the kinase domain increases kinase activity, it has also been discovered that R1441C/G mutation in the GTPase domain also influence kinase activity. The G2019S variation cause alterations in vesicle trafficking, neurite outgrowth, autophagy, cytoskeletal dynamics *in vitro* (Xiong *et al.* 2017). The genetic variant G2019S rs34637584 within the LRRK2 is located in the kinase domain (García *et al.* 2014). Several tool compounds have been developed to target the kinase activity of LRRK2 as a therapeutic strategy for Parkinson's disease since it was discovered that variation on G2019S causes increased kinase activity that is linked to neuronal toxicity (Kluss *et al.* 2018). LRRK2 is composed of a unique arrangement of conserved protein domain motifs that harbor pathogenic mutations in addition to the protein kinase domain, in mature neurons overexpression of G2019S mutant LRRK2 protein results into toxicity that can be blocked with mutations in conserved residues in the kinase domain that terminate kinase activity (Liu *et al.* 2014). LRRK2 is a member of Roco superfamily proteins, a novel multi-domain family of RAS-like G-proteins, expression of G2019S mutant leads to alteration of DRD1 trafficking in which DRD1 and DRD2 are the most abundant dopamine receptors in the CNS and belong to D1 or D2-class dopamine receptors (Rassu *et al.* 2017). LRRK2 protein is involved in regulation of signal transduction cascades due to the presence of Roc and MAPKKK domains, the G2019S mutation occurs at the MAPKKK domain and one major and two extremely rare haplotypes are found in carriers of G2019S (Keyser 2011). The LRRKs proteins are also classified as ROCO proteins because they contain the ROC (Ras of complex proteins) that falls under the GTPase domain followed by the COR (C-terminal of ROC) domain that is of unknown function. LRRKs have a catalytic kinase domain and protein-protein interaction motifs which includes a WD40 domain, LRRs, and ankyrin (ANK) repeats as presented in figure 1.2. LRRKs contains both GTPase and kinase activities, with the ROC domain acting as a molecular switch for the kinase domain, cycling between a GTP-bound state which drives kinase activity and a GDP-bound state which decreases the activity. The LRRK2 subfamily is part of a larger superfamily that includes

the catalytic domains of other STKs, protein tyrosine kinases, RIO kinases, aminoglycoside phosphotransferase choline kinase, and phosphoinositide 3-kinase (Melrose 2008). I2020T mutation is known to both increase and reduce kinase activity (West *et al.* 2007).

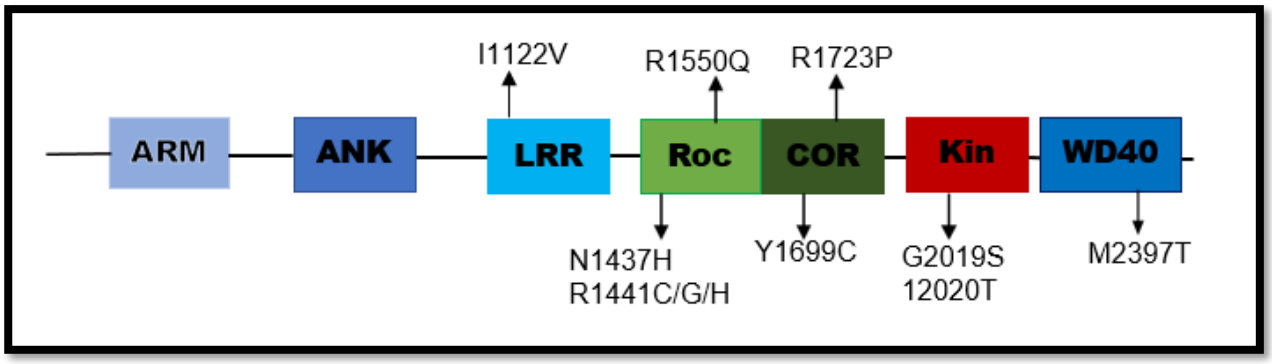


Figure 1. 2: LRRK2 protein with its 7 domains and some variants marked (Berwick and Harvey 2013)**.**

1.3 Kinase domain

LRRK2 is a protein kinase gene which signals switches that direct functional processes and protein communications in cellular networks and signal transduction pathways. It is regulated by different mechanisms such as phosphorylation of the activation loops, autoinhibition and allosteric activation by protein binding partners that enable the kinase domain to adopt a catalytically competent conformation and attain activity (Tse and Verkhivker 2015). In non-neuronal and neuronal cells LRRK2 is bound to GTP and lacks a GTPase activity, GTP binding takes part in the kinase activity and phosphorylation of the protein (Ito *et al.* 2007). A phylogenetic tree was used to calculate the similarity between RIP and mixed lineage kinases that form part of the tyrosine kinase-like (TKL) branch of the human kinome showed to the LRRK2 kinase domain (Gloeckner *et al.* 2006). These kinase families are involved in stress-induced cell signaling and mediate apoptosis. Within the LRRK2 are the Ras GTPase-like (Roc) and MAPKKK domain which raises the intramolecular signaling in which the activated (GTP-bound) GTPase directly stimulates kinase activity (Guo *et al.* 2007). Phosphorylation of generic peptide substrates is measured by kinetic kinase assays and show enhanced kinase activities caused by G2019S mutation (Liu *et al.* 2014). Through phosphorylation of RAB35 LRRK2 can regulate α-synuclein propagation in a kinase activity-dependent manner and LRRK2 kinase activity can be targeted for the purpose of reducing synucleinopathy lesions (Bae *et al.* 2018). Research shows that in early life increased LRRK2 activity can protect against opportunistic pathogenic infections, however, it later increases the risk of developing Parkinson's disease, a concept known as antagonistic pleiotropy (Alessi and Sammler). The ROC domain found in LRRK2 shares sequence homology with all five subfamilies of the Ras-related small GTPase

superfamily (Ras, Rho, Rab, Sar/Arf and Ran) including conservation of amino acids involved in GTP binding and hydrolysis (Guo *et al.* 2007). A unique arrangement of conserved protein domain motifs that also harbor pathogenic mutations is found on both sides of the kinase domain (Bae *et al.* 2018).

By cycling between GTP-bound (active) and GDP-bound (inactive) conformations Ras-related GTPases serves as molecular switches to regulate diverse cellular functions (Guo *et al.* 2007). Members of the Rab GTPase family including Rab8A, Rab10 and Rab29 are substrates for LRRK2, subset of 14 Rab proteins (Rab3A/B/C/D, Rab5A/B/C, Rab8A/B, Rab10, Rab12, Rab29, Rab35, and Rab43) are potential direct substrates for LRRK2 (Purlyte *et al.* 2017b). There is currently no evidence of protein substrates being dependent on LRRK2 kinase activity *in vivo,* using autophosphorylation-specific antibodies measurements of LRRK2 autophosphorylation *in vitro* or directly from cell lysates, show that pathological mutations in GTPase or kinase domain activate the proportion of autophosphorylated LRRK2 (Guo *et al.* 2007). Thr72 for Rab8A and Thr73 for Rab10 which are found at the Rab protein phosphorylation site are found within the effector-binding switch II motif. These block the binding to Rab GDP-dissociation inhibitor (GDI) that is required for Rab protein membrane delivery and recycling phosphorylation also inhibits binding of Rab8A to Rabin-8, its cognate guanine nucleotide exchange factor (GEF) (Purlyte *et al.* 2017b). Guanine nucleotide exchange factors (GEFs) facilitate GTP-binding and subsequent effector interaction and downstream signaling, GTPase activating proteins (GAPS) increase the intrinsic rate of GTP hydrolysis to terminate signaling and GDP dissociation inhibitors (GDIs) stabilize an inactive GDP-bound pool of the protein (Guo *et al.* 2007).

LRRK2 is expressed in different parts of the brain including the substantia nigra pars compacta, striatum, hippocampus, cortex, olfactory bulb, also in neurons and lastly in astrocytes and microglia in which it is associated with inflammatory processes related to Parkinson's disease (Russo *et al.* 2018). R1441C and I2020T mutations enhance the proportion of enzyme in an active state without affecting other kinetic parameters (Bae *et al.* 2018). At the cluster of Ser residues found between the ankyrin domain and leucine-rich repeat region (Ser910, Ser935, Ser955, and Ser973) that is where LRRK2 is phosphorylated and plays a role in regulating 14-3-3 binding and cytosolic localisation. These sites are controlled by LRRK2 kinase activity and often become dephosphorylated in response to diverse LRRK2 inhibitors and to assess the *in vivo* efficacy of LRRK2 inhibitors dephosphorylation of these residues especially Ser935 has been monitored (Purlyte *et al.* 2017b). Bae E-J et al (Bae *et al.* 2018), suggested that to reduce synucleinopathy LRRK2 kinase activity can be targeted. Phylogenetic analysis revealed that the MAPKKK domain resembles mixed lineage kinases, which

are part of tyrosine kinase-like branch of the human kinome which commonly has both Ser/Thr and Tyr kinase activity (Guo *et al.* 2007). It was discovered that at a molecular level LRRK2 negatively regulates protein kinase A (PKA) activity causing an increase of PKA-mediated phosphorylation and consequent accumulation of NF-kB inhibitory subunit p50 in the nucleus that then leads to repression of NF-kb target genes (Russo *et al.* 2018). (Purlyte *et al.* 2017b) demonstrated that RAb29 stimulates kinase activity by assessing autophosphorylation of Ser1292 as well as phosphorylation of LRRK2 substrates such as Rab10, they also mentioned that pathogenic mutants that bind GTP with higher affinity are activated by Rab29 and the interaction is between the N-terminal ankyrin domain. LRRK2 has been shown to influence kinase activity and several pathogenic mutations or variants increase kinase activity (Guo *et al.* 2007).

1.5 Project Motivation

Parkinson's disease remains as one of the most important autosomal dominant and genetically transmitted diseases worldwide. As a result, it is important to identify possible drug targets for the management and the treatment of the disease. The LRRK2 has proved to be a promising drug target for Parkinson's disease treatment. As such, new compounds against LRRK2 need to be identified using *in-silico* approaches. Screening natural compound databases such as the South African Natural Compounds Database (SANCDB) has the potential to provide compounds that can be used to inhibit LRRK2 thus treating Parkinson's disease.

1.5.1 Project knowledge gap

The reason for increased kinase activity in variant G2019S is not well known and there is no cure for Parkinson's disease. There currently is no Leucine-Rich Repeat Kinase 2 (LRRK2) crystal structure available in the databases, much is not well known or understood about the variants (G2019S and I2020T). There has not been any *in silico* virtual screening of SANCDB natural compounds studies or molecular dynamics to understand the movement of the macromolecule and ligands.

1.5.2 Problem statement

Levodopa and other dopaminergic medications are taken by patients suffering from the disease show dopa-resistance as time goes by and the drugs have side effects such as psychosis, motor fluctuation and dyskinesias and therefore the need for a more effective, safe and affordable drug targeting the active site for treatment of Parkinson's disease (Jaiteh *et al.* 2018).

1.5.3 Research hypothesis

Viable 3D homology models will be created for the wildtype and variants of the kinase domain. *In silico* studies will be used to identify natural compounds from SANCDB that can be used to inhibit variant LRRK2. Molecular dynamics simulation will provide insights to structural dynamics of the wildtype and variants of the kinase domain.

1.5.4 Project methodology

Figure 1.3 below, explains the flow of the project. In which NCBI and UniProt databases were used for data retrieval of the LRRK2 kinase domain sequence. LRRK2 kinase domain models were created in homology modeling using MODELLER v9.19. Ligands from the South African Natural Compound Database (SANCDB) were used for *in silico* molecular docking using AutoDock-Vina and PyMOL and Discovery studio visualiser were used to visualise the structures. The physical movements of the molecules were studied in molecular dynamics using GROMACS (Amber03) and ACPYPE for creating the topology files. Finally, the study will focus on the network analysis of the

structures. Figure 1.3 below shows the methodology that was used to perform the project, each stage has been mentioned together with the tools or packages that were used.

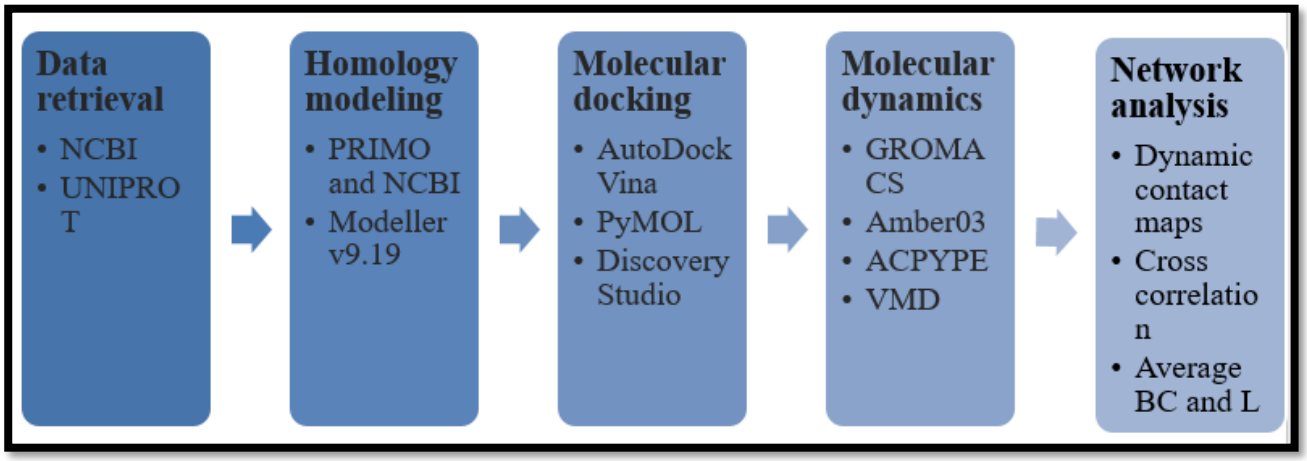


Figure 1. 3: Steps taken in the project.

1.6 Aims and objectives

1.6.1 Aims

The main aim of this work is to identify inhibitory compounds that are effective against the variant LRRK2 and not affect the wildtype protein and to dock models of natural compound substrates to inhibit variation of LRRK2. Since there is currently no available crystal structure for this enzyme, homology modeling will be performed to create a viable model of the LRRK2 kinase domain. Once this model is created docking will be performed followed by molecular dynamics (MD) simulations. The best hits will be selected for possible *in vitro* assays.

1.6.2 Objectives

1. Retrieve data which includes available crystal structures and sequences.

2. To model representative structures of the human LRRK2 kinase domain wildtype and variants.

3. Screen the SANCDB to identify possible inhibitory compounds against LRRK2 variants and the wildtype.

4. Perform Molecular dynamics simulation on the LRRK2 protein to test the effect of the identified compounds on protein dynamics.

5. Perform network analysis on wildtype and variants to assess the effect of variation on protein dynamics.

CHARPTER TWO: Homology modeling and Structural Analysis

2.1 Introduction

Sequence alignment is a process of comparing two or more sequences in a pairwise fashion, following the order of residues/nucleotide in the sequence (Xiong 2006). This method is an important process when comparing sequences because it allows inferring functional activity, assesses homology, performs phylogenetic analysis and offers the guide for homology modeling (Daugelaite, O' Driscoll and Sleator 2013). Sequence similarity looks at residues that share similar physiochemical properties. A good sequence alignment should have few gaps as possible. A sequence sharing less than 30% sequence identity with its homolog will result in a structure having more errors such as incorrect side chain packing, gaps in sequence alignment or disordered loop regions (Krieger, Nabuurs and Vriend 2003). The sequence alignment and template structure quality will determine the quality of the homology model. There are different types of search tools including BLASTP which is used to retrieve homologous sequences by applying the reverse-BLAST procedure (Yang and Tung 2006). It queries a protein sequence against a protein database for similar sequences, this BLAST increases the chances of retrieving a true homolog of the sequence (Pedersen).

Alignment can be performed through a command line that accepts the information of the sequence. A fasta format of the query sequence or its accession number is submitted to BLAST (Basic local Alignment search tool) and the tool also allows one to specify the algorithm parameters. Another way of inputting data is to upload files from one's computer or paste sequences to a dialogue box for alignment to be done in an online webserver. Jalview or Bioedit which are alignment tools can be used to view the alignment results, and they show them as coloured images based on physicochemical properties or conservation (Alzohairy 2014).An ideal homologous sequence should have a high sequence identity of at least 30% or higher, high overall sequence coverage and an expectation score close to zero compared to the query sequence. The expectation score or E-value is the expected number of errors per query. In PRIMO and NCBI the alignment profiles are scored based on residue similarity; to facilitate this scoring, amino acid by amino acid equivalence tables are constructed, gap penalties introduced, and complex mathematical algorithms applied (Rose *et al.* 2011).

2.1.1 Homology modeling

There currently is no crystal structure of LRRK2 available in the databases and no 3D structure of the kinase domain, therefore there is a need to create models using homology modeling. The LRRK2 structure has not been solved due to technical issues of purifying sufficient quantities of soluble full-

length recombinant human LRRK2 domains (Phu *et al.* 2017). Homology modeling is when the amino acid sequence of the protein of interest is used to construct an atomic-resolution model using the 3D structure of a related homologous protein referred to as a template. In this chapter, the focus will be on generating high-quality LRRK2 kinase domain models, the wildtype models, variant 1 and variant 2 models. Homology modeling, also known as comparative modeling identifies one or more known protein structures that resemble the query sequence and produces an alignment that maps residues in the query sequence to the residues in the template residues (Fernandez-Fuentes *et al.* 2007). Protein structures are more conserved than protein sequence amongst homologues.

Homology modeling is based on the principle that the protein structure is less susceptible to evolutionary change hence more conserved than the affiliated sequence (Brown and Tastan Bishop 2017). NMR spectroscopy, X-ray crystallography and electron microscopy are techniques used to resolve protein structures (Masso). X-ray crystallography has short wavelengths allowing more detail to be captured during imaging, here proteins are grown into large crystals and x-rays are fired at the crystals producing a diffraction pattern (Cme, Bmi and Dror 2015). In solution NMR the proteins are solved in solution, relies on nuclear Overhauser effect (NOE) and then proteins are labelled with radioisotopes ^{13}C and ^{15}N (Chou and Sounier 2013). NMR produces a lot of models and since proteins are mobile these numerous models will fit all restraints, allowing it to build models that fit the data (Edwards 1992). Electron microscopy measures the electrons scattering from contact with the sample and to avoid non-specific contact it is then performed *in vacuo* (Bai, McMullan and Scheres 2015). Using this method, the smaller the wavelength the better the resolution and it is the best for studying macromolecular assemblies that are too large for solution NMR and X-ray crystallography. Electron microscopy has the potential to be equivalent to X-ray crystallography and damage to the protein is reduced by using negative staining and lower intensity electron beams.

Protein Data Bank (PDB) is an archive for experimentally determined protein structures and despite the advancement in structural solvation technologies, there is still a pool of proteins not yet solved, reason being solvation experiments are expensive and time-consuming (Brown and Tastan Bishop 2017). Homology modeling serves as a technique that covers this gap and maybe the current reliable option remaining when experimental methods are unsuccessful. There are stages that are very important in homology modeling; template identification, sequence alignment, model building and model validation (Krieger, Nabuurs and Vriend 2003). These steps involve numerous quality checks and refinements for maximum accuracy of the models created. Webservers like HHpred, SWISS-MODEL and PRIMO are available to automate this process. These models will be used for further analysis in molecular docking (CHAPTER 3), molecular dynamics (CHAPTER 4) studies and

network analysis (CHAPTER 5). To understand the biological functioning of the protein of interest 3D structures can be used.

2.1.2 Template identification and Selection

This is the crucial step of homology modeling because there is need to model interactions within and between atoms in the model accurately and selection of the wrong template will result in an unreliable protein model due to incorrect fold assignment and result in incorrect interactions being inferred(Masso 2000). Select one or more templates that are appropriate for the modeling problem. There are factors that should be considered when selecting templates, for example, select the structure with the highest sequence identity, consider sequence coverage and secondary structure match between the template and target (Fiser 2014). An ideal template should cover majority length of the residues in the sequence, have a sequence identity of 30% and above, and again have a consistent secondary structure match with the target.

The optimal use of several templates increases the model accuracy, this allows the templates to be aligned with different domains of the target with little overlap between them in case the modeling procedure construct a homology-based model of the entire target sequence (Fiser 2014). It is necessary to consider the coverage as well, especially of the residues of interest. In this project, templates were chosen from PRIMO and NCBI, which uses an integrated database retrieval system to identify potential templates from the PDB database. In NCBI the results are represented as a visual format of alignment bars and column-column residue match and it then ranks the templates found based on identity and parameters that were set before blasting (Sharma *et al.* 2018). In PRIMO potential hits are arranged based on the most probable match to the least. The quality of these templates is further accessed by considering the resolution, R-value, completeness and global quality metrics (Fiser 2014).

BLAST is the tool that is used the most for template searching, however, it is not reliable when it comes to identifying homologs with sequence identity less than 30% (Xiong 2006). The E-value of the template chosen should be close to zero as possible so that the right model is created. The quality of the templates will determine the quality of the models created hence it is important to evaluate the templates selected since the backbone atoms of the template are copied to the target protein. Validation tools like ProSA (Wiederstein and Sippl 2007a), POCHECK (Laskowski *et al.* 1993) or QMEAN (Benkert, Ku and Schwede 2009) can be used to further check the quality of the templates.

Template selection was done using PRIMO and NCBI webserver. The best template for modeling the human LRRK2 kinase domain was selected. Modeling was done in the RUBi cluster. For each alignment profile, 100 models were built using MODELLER version 9.19 with very slow refinement. Generated structures were sorted according to DOPE (Discrete Optimized Protein Energy) Z-score. The top three models of each protein conformation as indicated in Table 2.1 below proceeded to structural quality evaluation, although some models had a Z-DOPE score bigger than -0.5 which is unfavourable. QMEAN (Benkert, Ku and Schwede 2009), ProSA (Wiederstein and Sippl 2007a) and PROCHECK (Laskowski *et al.* 1993) tools were used to evaluate the produced models. The GA341 score which is a function of the model combined statistical potentials z-score, compactness and percentage sequence identity of the alignment used to build the model. The GA341 score assesses the model reliability and considers the sequence identity between template and query. The best quality model from the top three were identified. Models were visualised in PyMOL visualisation program. The best model of the three structures proceeded to dock simulation studies.

Table 2. 1: The best top three model for LRRK2 (kinase domain) wildtype, variant 1 (G2019S) and variant 2 (I2020T).

Structure name	Model ID:	Z-DOPE score
	Q5S.B99990022.pdb	-0.53
Wildtype	Q5S.B99990057.pdb	-0.51
	Q5S.B99990028.pdb	-0.46
	Q5S.B99990022.pdb	-0.55
Variant 1(G2019S)	Q5S.B99990022.pdb	-0.49
	Q5S.B99990022.pdb	-0.48
	Q5S.B99990022.pdb	-0.52
Variant 2 (I2020T)	Q5S.B99990022.pdb	-0.52
	Q5S.B99990022.pdb	-0.51

2.1.3 Template-target sequence alignment

In this step, the sequence of the selected template is aligned with the sequence of the target structure. The model created depended on the alignment used, it is important to consider multiple sequence

alignment programs that can include structural information of the template, exploit evolutionary signals and profiles, structure-aware, context-aware taking the secondary structure or trans-membrane which are under presented in PDB (Feenstra 2017). Multiple sequence alignment reveals clearly the conserved regions that are significant to the target protein. In PRIMO, TCOFFEE alignment tool was used.

2.1.4 Protein modeling

In this step, the target protein is modelled based on the given template-target alignment structure. There are different types of software tools that can be used for the model building including MODELLER, HH-suite, Foldit, Raptor X and many more (Nikolaev *et al.* 2018). To generate template-based alignments MODELLER which uses a restraint-based approach for comparative modeling (Sali 2013), was considered in this project. In which the known template is aligned with the target sequence and spatial features such as Cα-Cα distances, hydrogen bonds, main chain and sidechain dihedral angles are transferred to the target from the template (Feenstra 2017). Backbone structure is generated before processing the side chain and loops when building a model (Jonathan and Meier 2014). MODELLER will use the PDB file of the template, a parameter file containing Python written commands, and template-target alignment .PIR file as input. An example of a parameter file used for modeling is shown in [Appendix 1]. The python script will contain the modeling parameters such as the number of models to be created and the level of refinement. The refinement level will set the time spent to build a single structure, together with the accuracy and quality of produced structures (Vyas, Garg and Iansavichus 2012).

Alignment in the .PIR file format is like the FASTA alignment file format, with an additional sequence structure/structure description line between the header and the sequence as shown in [Appendix 2.1 to 2.3]. The sequence/structure description line denotes the sequence/structure name, start and end protein chains, and the beginning and end positions of residues in the aligned sequences. It is recommended to check the start and end residue positions specified in this file with similar positions in the PDB file for accuracy because of residue one of the template sequence in the alignment .PIR file may not be the first amino-acid that formulates the complete protein sequence in the PDB file. An asterisk "*" sign always marks the end of each sequence. From the several models that are created by MODELLER select the best model based on model quality assessment programs or any additional knowledge about the structure or function of the target protein and perform model refinement if there is a need (Feenstra 2017).

2.2 Model validation

This step is to determine if the created models are a reliable approximation of the near native conformation, using different model quality assessment programs for model validation (Feenstra 2017). MODELLER evaluates the models created using the Z-DOPE score. The Z-DOPE score stands for discrete optimised protein energy which is a statistical potential based on an improved reference state that corresponds to noninteracting atoms in a homogeneous sphere with the radius dependent on a sample native structure, giving it the finite and spherical shape of the native structures (Shen and Sali 2006). The Z-DOPE score should be less than -0.5 for the model to be considered reliable. Validation scoring may be specific to structure prediction for example it can check if the secondary structure helix and strand versus loop, has a similar ratio known in protein structures, if there is a missing single helix or β-strand on a model with a sequence of 200 amino acid, it means this model does not resemble the true protein structure (Feenstra 2017). The quality of the models created can be evaluated globally or locally in which global evaluation programs determine the quality of the entire length of the structure whereas local evaluating tools assign quality scores to each residue in the structure. Local evaluation tools include ANOLEA, PROCHECK, ProSA, QMEAN and Verify-3D. local evaluation tools present results as graphical pictures making it easy to spot problematic regions in the structure and often these are loops and areas with large insertions or deletions that usually lack template structural information.

2.2.1 QMEAN

QMEAN stands for Qualitative Model Energy ANalysis and it is an evaluation tool that calculates both global and local quality. It also assists in the model selection and identification of the problematic regions for subsequent refinement (Benkert, Ku and Schwede 2009). QMEAN is different from other model quality assessment servers in that it includes a more detailed distance-dependent all-atom interaction potential as well as a torsion angle potential over three consecutive residues, the residue-level distance-dependent pairwise potential is based on Cβ atoms (Benkert, Tosatto and Schomburg 2008; Benkert, Ku and Schwede 2009). The identification of the native structures out of a variety of decoy sets depends on the torsion angle potential over three residues meaning the torsion angle potential describes the propensity of a certain amino acid for a certain local geometry considerably better than the single residue torsion angle potential (Benkert, Tosatto and Schomburg 2008). QMEAN scoring function is a valuable tool in model quality assessment, it distinguishes acceptable models from unacceptable models and identifies the native structure among decoys generated by a variety of methods (Benkert, Tosatto and Schomburg 2008).

2.2.2 ProSA

To understand the biological process at a molecular level, a structural model of that protein is needed. ProSA recognises errors in experimentally determined structures, theoretical models and protein engineering (Wiederstein and Sippl 2007b). The overall quality score calculated by ProSA for a specific input structure is displayed in a plot that shows the scores of all experimentally determined protein chains, the plot of local quality scores will show the problematic parts of the model and map them on a display of the 3D structure using colour codes (Wiederstein and Sippl 2007b). ProSA is very strict, it detects very minor errors such as distorted geometry of hydrogen-bonded residues and it evaluates nearly-native homology models than for the fold-recognition models that are plagued by local errors and this tool relies on empirical energy potentials derived from pairwise interactions observed in high-resolution protein structures (Pawlowski *et al.* 2008).

2.2.3 PROCHECK

PROCHECK aims to assess how normal or unusual the geometry of the residues in a protein structure in comparison to the stereochemical parameters derived from well-refined, high-resolution structures. PROCHECK can clean-up the coordinate file of any mislabelled atoms and relabel in accordance to the IUPAC naming conventions and the results produced by this program consist of a number of plots including Ramachandran plots, together with a detailed residue by residue listing (Laskowski *et al.* 1993). This program assesses the quality of protein structures in the process of being solved and also of existing structures and of those modelled on known structures (Laskowski *et al.* 1993). PROCHECK uses the stereochemical considerations alone, both to provide an overall assessment of the stereochemistry of the given structure and highlight regions that may need further investigation (Laskowski *et al.* 1993). The conformation of the backbone of non-terminal amino acid residues is determined by three torsion angles Phi, Psi and Omega, which are represented in a scatter plot called the Ramachandran plot.

2.3 Methodology

The LRRK2 kinase domain sequence was retrieved from UniProt and it had a length of 260 amino acids from position 1879-2138 of the amino acids in LRRK2. Sequence blast was performed in PRIMO and NCBI and the PDB database was used, two templates were selected to cover up for gaps not covering the sequence of interest. The target sequence was aligned to the template sequence using T-COFFEE in the 3D-COFFEE mode in PRIMO web tool, there were no missing residues, or any residues modified. PIR files were created for modeling, a python script was used to create 100 models using very slow refinement. The models were arranged based on their Z-DOPE score starting with the lowest energy score as the first one. Protein models for LRRK2 (kinase domain) were generated using MODELLER v9.19 and the following steps were used in generating the models:

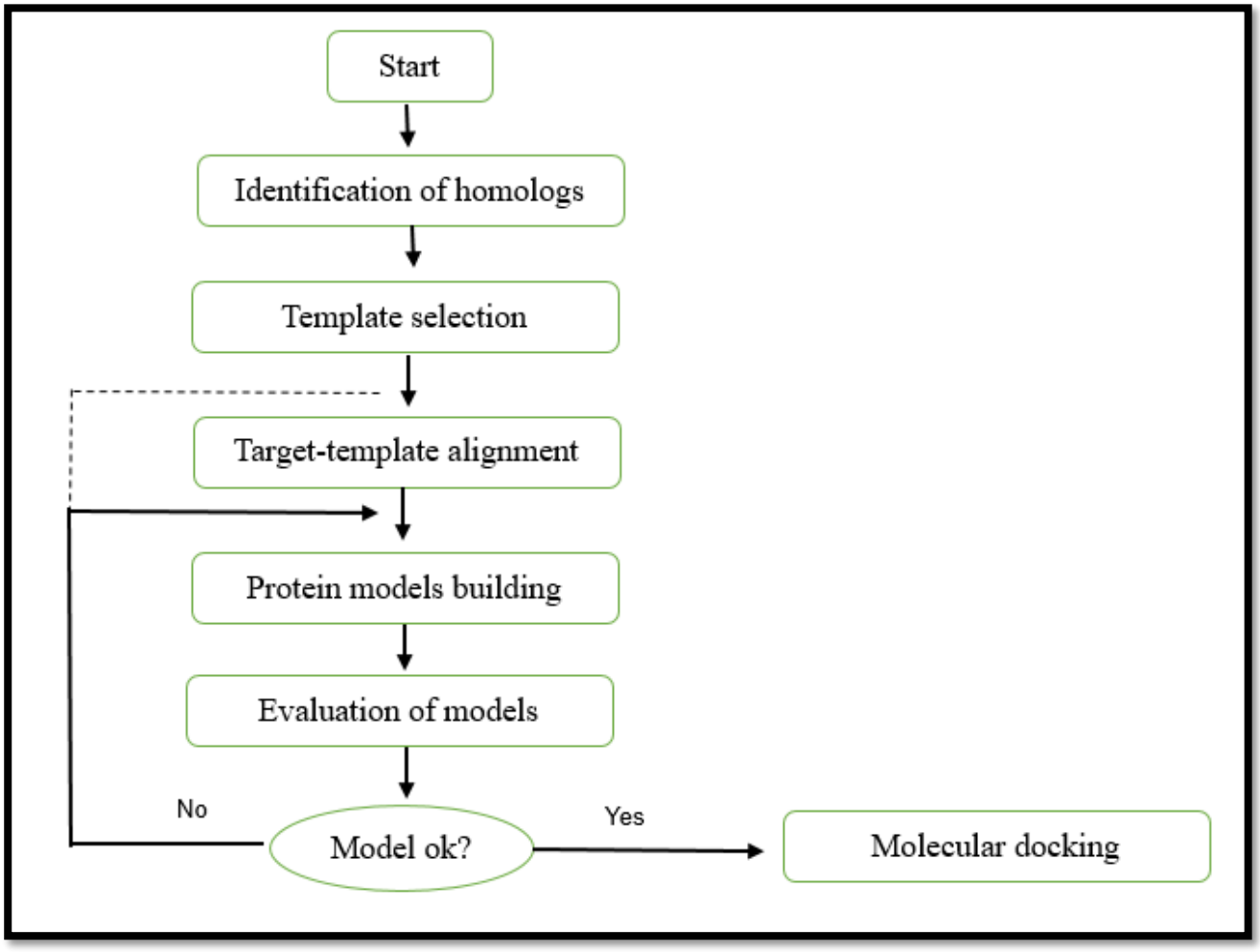


Figure 2. 1: A flow diagram of homology modeling steps used in this chapter, the arrows show the flow direction of the steps (Schwede *et al.* 2003)**.**

Molecular docking mentioned in figure 2.1, was performed in CHAPTER 4.

2.3.1 Protein Sequence Retrieval and template selection

LRRK2 accession No: Q5S007 sequence was searched and retrieved from UniProtKB database[3]. Using the retrieved sequence, a BLASTP search was carried out to recover LRRK2 homologous sequences. Sequences of accession number: Q5S007 were retrieved from NCBI website[4]. LRRK2 sequence was further used to search homologous sequences of Parkinson's disease origin. Reverse BLAST was carried out against the sequence retrieved, to establish true orthologs. Structural and homology prediction of potential modeling templates was performed on LRRK2 sequence using PRIMO and NCBI webservers. There was no crystal structure available in the databases. Template search was done in PRIMO using default parameters. Several 3D structures came up as the search output using BLAST, two best templates were chosen based on their good PDB validation matrices, high resolution, high sequence identity to the query sequence, good coverage of the structure to the query sequence, completeness of the protein structure and an E-value close to zero as possible.

2.3.2 Template-target sequence alignment

Template-target alignment was performed in PRIMO using T-COFFEE in 3D-COFFEE mode, which considers the structural information of the proteins during alignment. Two templates were used to prevent gaps in the alignment and ensure total sequence coverage.

2.3.3 Protein Modeling

MODELLER v9.19 was used to create the models for the wildtype protein, variant 1 (G2019S) and variant 2 (I2020T). One hundred models were produced based on the sequence input and selected templates. Very slow refinement was used, and a python script was used to rank the models created based on Z-DOPE score. The top three models for each protein structure which had a Z-score of -0.5 and below were chosen and considered for further evaluation.

2.4 Evaluation and Validation of Models

ProSA, PROCHECK and QMEAN were the validating webserver tools used. The best model for each structure that had passed the evaluating tools was selected for further structural analysis using PyMOL version 1.8 and Discovery studio version 3.6.

[3] (http://www.uniprot.org/uniprot/Q5S007)
[4] (http://www.ncbi.nlm.nih.gov/)

2.5 Results and Discussions

LRRK2 (kinase domain) consists of 260 amino acids and the protein consists of one chain. The following (Figure 2.2), shows the sequence of the kinase domain of LRRK2 retrieved from UniProt and the residues that were changed to match the variants structure. The first sequence is the wildtype with Glycine (G) on position 140 and Isoleucine (I) on position 141, followed by the variant 1 with Glycine (G) on position 140 changed to Serine (S) and lastly, variant 2 with Isoleucine (I) on position 141 changed to Threonine (T). The LRRK2 kinase domain sequence numbers resiudes from 1879-2138, however in figure 2.2 below the number has changed to 1-261, marking regions of interest as G140 and I141.

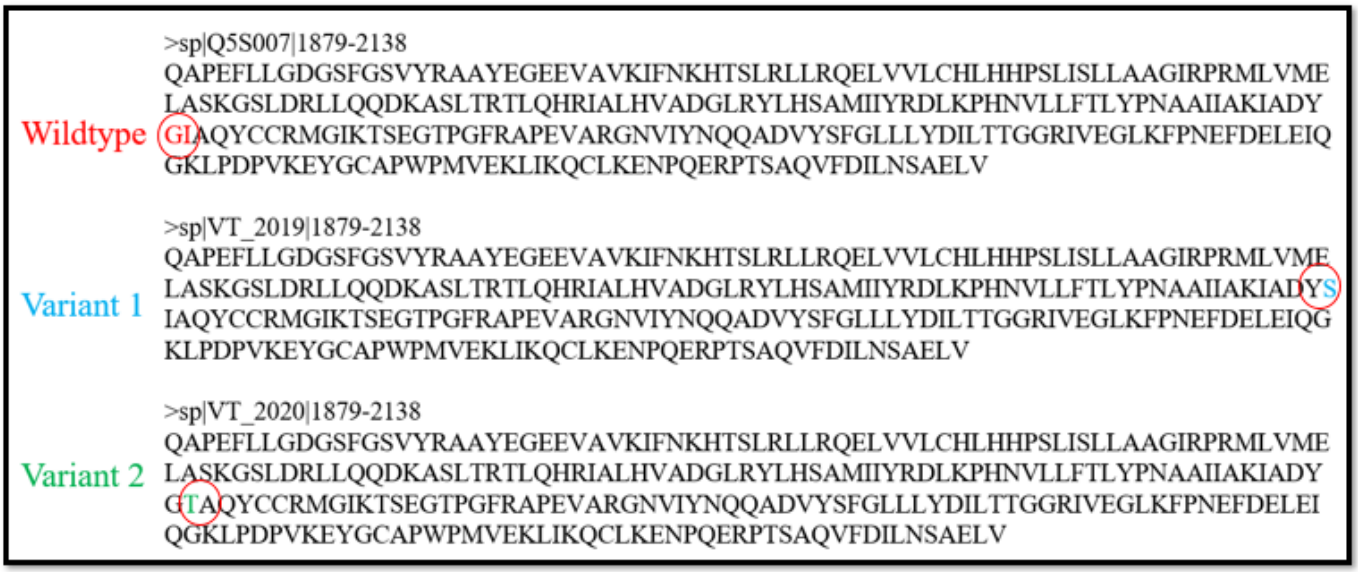


Figure 2. 2: The retrieved sequence of the three structures.

Table 2. 2: Template selection and validation table.

Organism	Templates quality				
	Template	Resolution (Å)	Coverage (%)	R-value	Identity (%)
Homo sapiens	4UY9	2.81Å	4-258 (98)	0.23	32
Homo sapiens	1FVR	2.20Å	6-255 (96)	0.23	32

Templates MLK1 kinase domain (4UY9) and TIE2 kinase domain (1FVR) for the target structure were identified in PRIMO[5]and NCBI website[6]. Table 2.2 above, shows the identity, quality and additional information about the two templates selected. Both templates had an identity above 30% and an R-value close to 0 in Table 2.2 which is desirable. The template is the structure that was used to construct the models. Coverage shows the sites on the target protein that are covered or represented. The R-value shows how well the model fits the diffraction data. Since R-value can be biased an R-free value is introduced to fix this problem. Resolution is the smallest distance between crystal lattice planes that is resolved in the diffraction pattern, measured in angstroms (Å). A high resolution is resembled by a low value and shows better quality and a low resolution is resembled by a high value showing poor quality. Identity shows how similar the template sequence resembles the target sequence. Figures 2.3 and 2.4 shows, the cartoon structure of the crystal structures of (1FVR and 4UY9) templates retrieved from RCDB protein data bank viewed in PyMOL.

[5] https://primo.rubi.ru.ac.za
[6] https://www.ncbi.nlm.nih.gov/Blastp.cgi

Figure 2. 3: 4UY9 cartoon structure viewed in PyMOL.

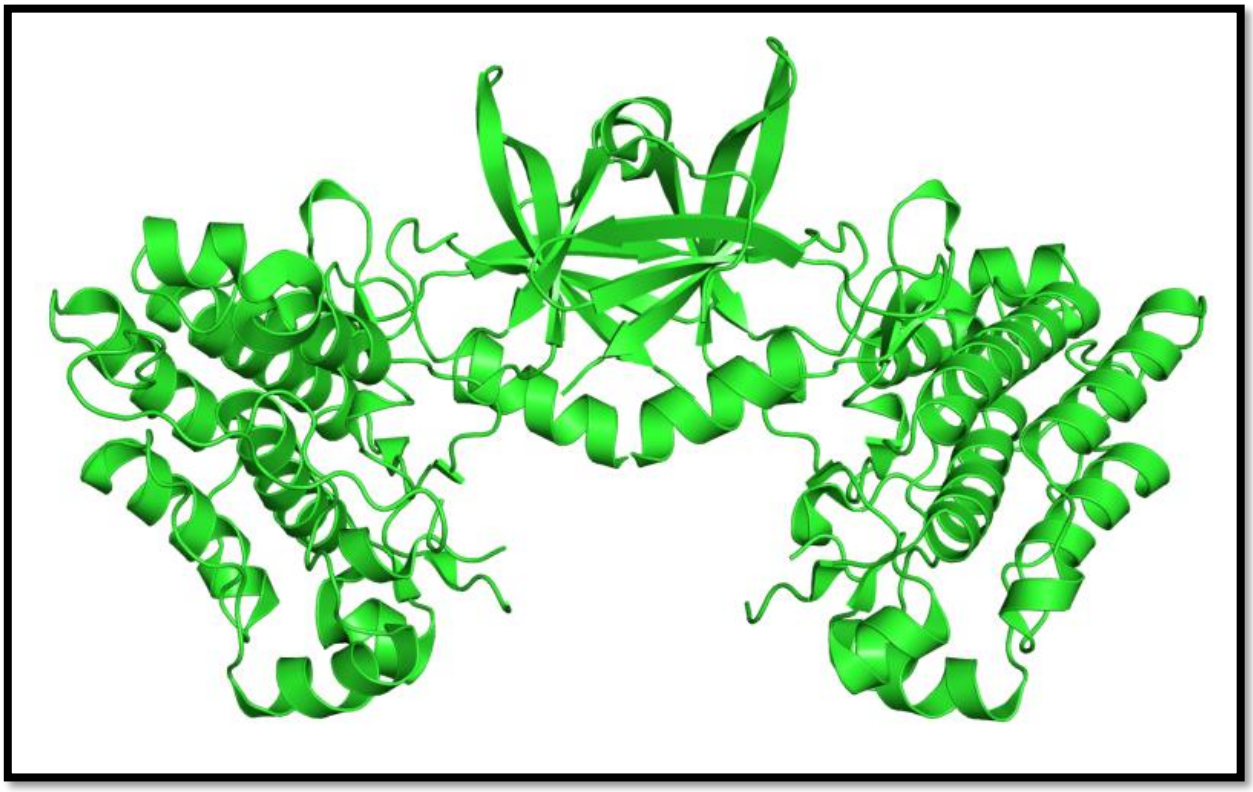


Figure 2. 4: 1FVR cartoon structure viewed in PyMOL.

Table 2.3 below, shows the Z-DOPE score for the three models and the templates. ProSA, PROCHECK and QMEAN were used as validating tools. Three tools were used because they have different algorithms and to make sure the models are good indeed. The template values were within the range of native proteins of similar size in the PDB, suggesting an acceptable overall quality. The plots showed good stereochemistry properties of the structures and the models were approved by these three validating tools. A Z-DOPE score of less than -0.5 is considered good, in PROCHECK the residues of interest should at least fall in the most favoured or allowed region and the QMEAN value should be between 0 and 1 to be considered acceptable.

Table 2. 3: Model evaluation of the templates and the best 3D models.

Model	**Z-DOPE score**	**ProSA**	**PROCHECK (%) Regions**			**QMEAN**
			Most Favoured	Allowed	Disallowed	
WT_57	-0.52	-5.59	86.70	11.10	1.80	0.55
VT_G2019S	-0.55	-5.49	85.90	10.60	1.80	0.59
VT_I2020T	-0.55	-5.29	87.20	9.70	2.20	0.59
4UY9	-1.69	-8.75	86.70	9.70	2.20	0.74
1FVR	-1.29	-5.94	90.0	8.40	1.60	0.70

PROCHECK assess the overall stereochemical quality of a protein structure through a local evaluation (Laskowski *et al.* 1993). This tool is found on the SWISS-MODEL server and it provides a Ramachandran plot that gives a landscape view of the distribution of torsion angles of a protein structure. 90% of the residues should fall in the most favoured region for the structure to be considered of acceptable quality. The evaluating tools showed that the templates chosen were reliable for homology modeling. ProSA had values less than -0.5 for all three models, the residues of interest were falling in the most favoured region in PROCHECK and the QMEAN values were falling between 0 and 1, the evaluating tools validated the models that were created. Figures 2.6-2.11 shows the results from the three evaluation tools. It was found that all three evaluation programs accepted the models created. Figure 2.5 below shows the structural similarity between protein of interest and

the templates used for modeling, cyan represents 4UY9, green represents 1FVR templates and magenta colour represents LRRK2 (kinase domain) protein.

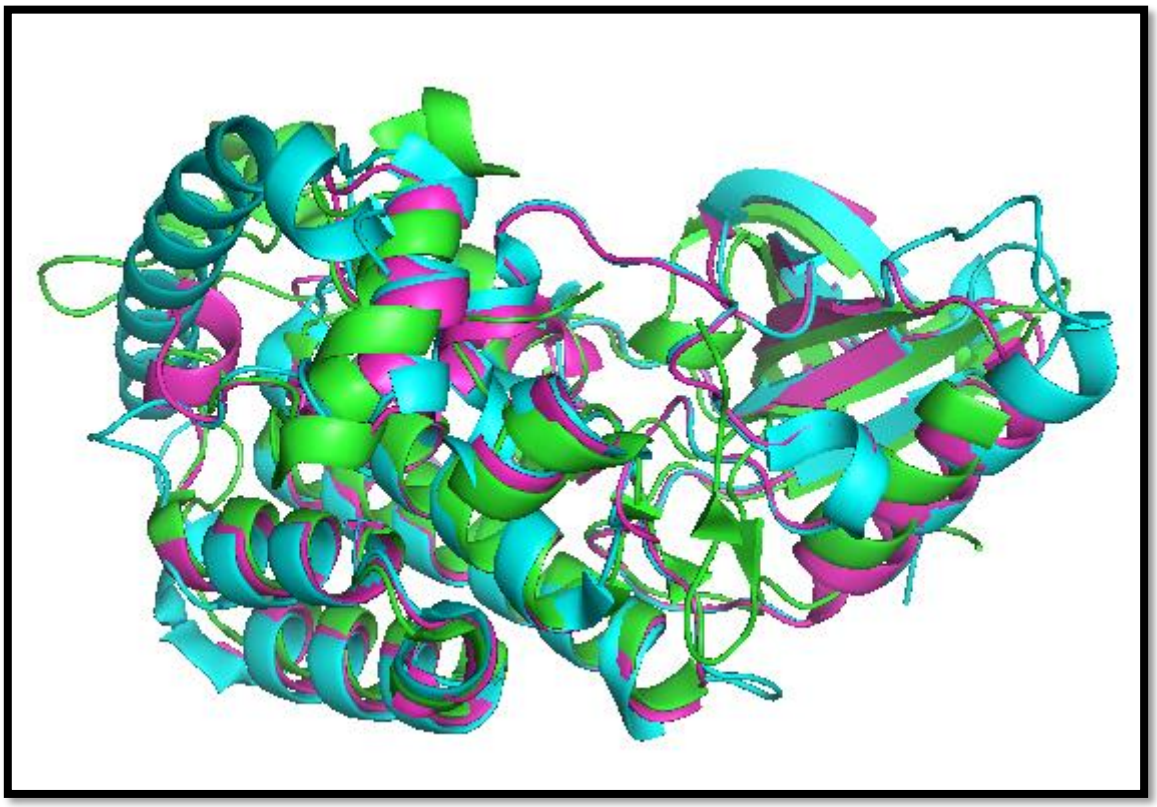


Figure 2. 5: Structural alignment between the models and the templates viewed in PyMOL, showing structural similarity, green for 1FVR, cyan for 4UY9 and magenta for LRRK2.

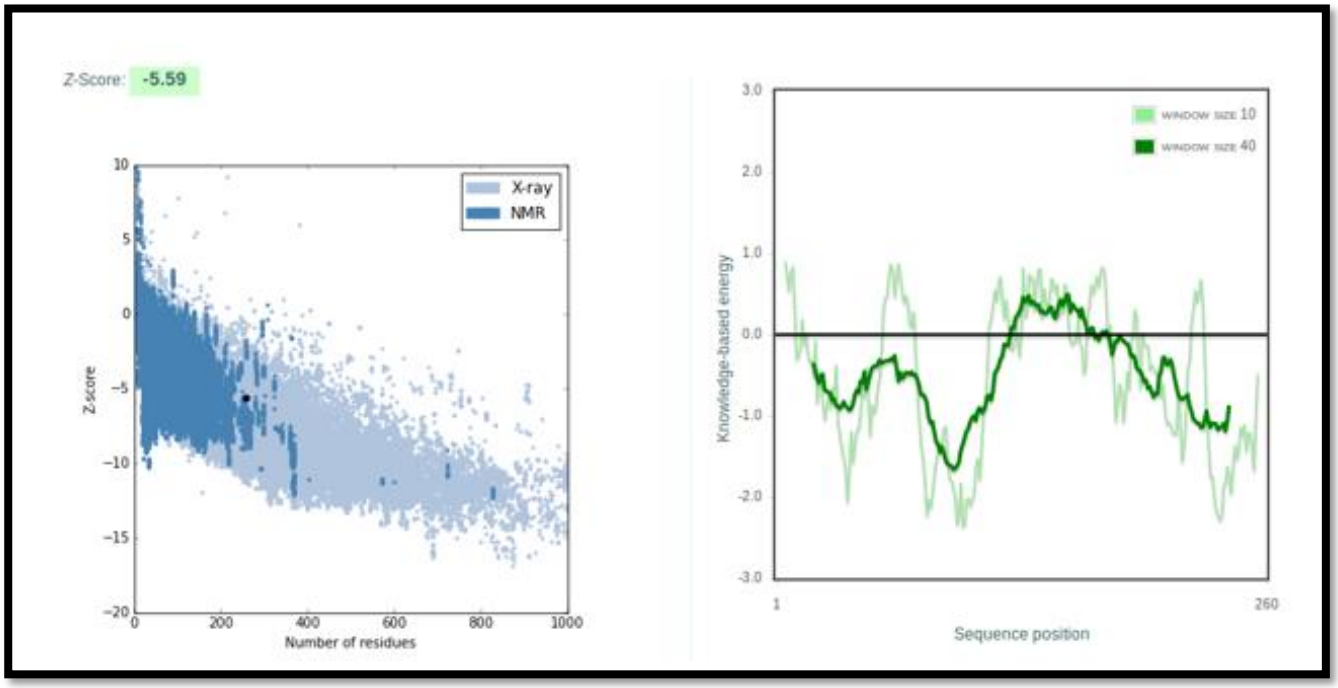


Figure 2. 6: Plots from ProSA evaluating tool for the wildtype Kinase domain model.

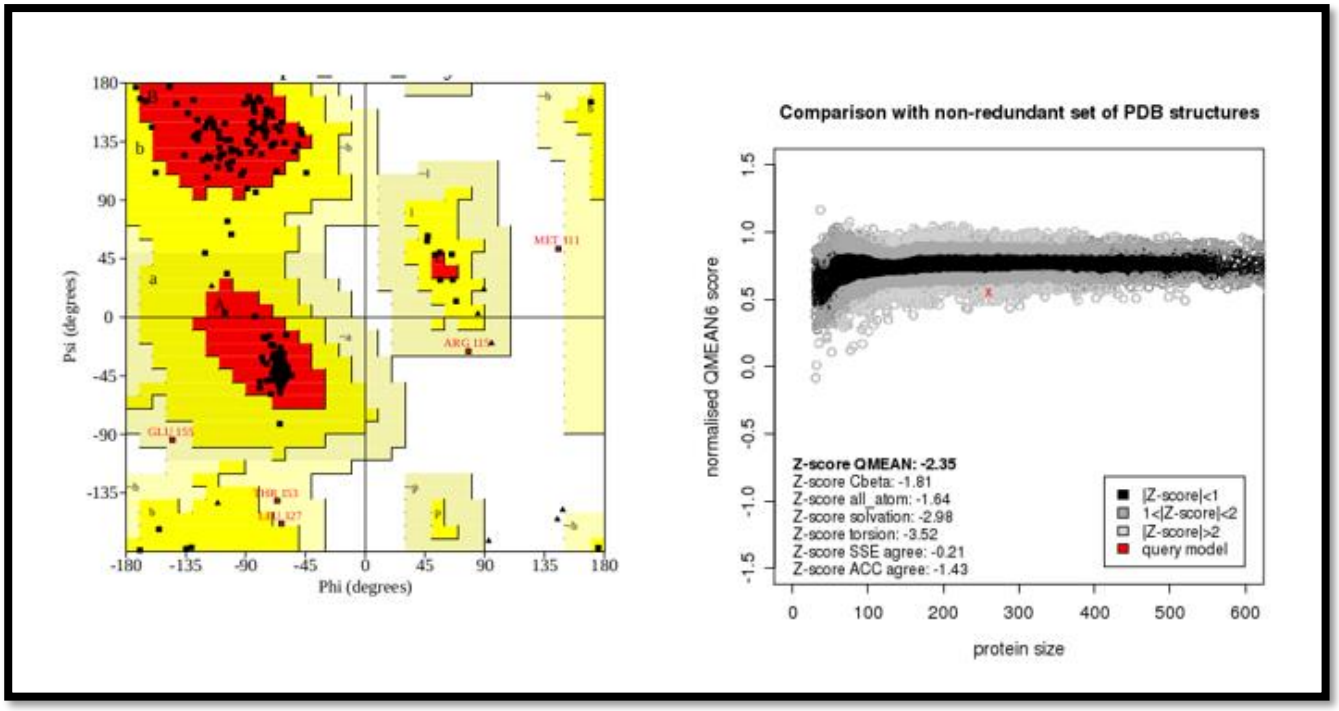


Figure 2. 7: Plots from PROCHECK and QMEAN evaluating tools for the wildtype Kinase domain model.

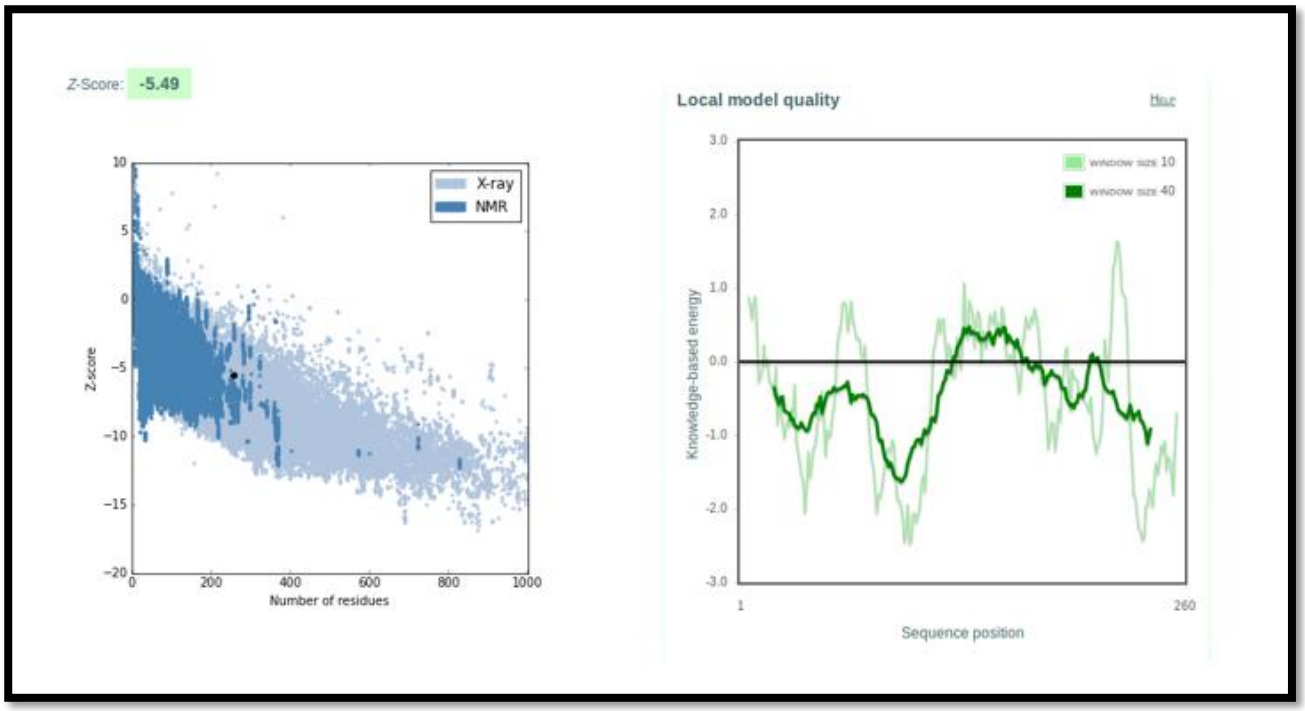


Figure 2. 8: Plots from ProSA evaluating tool for the variant 1 G2019S Kinase domain model.

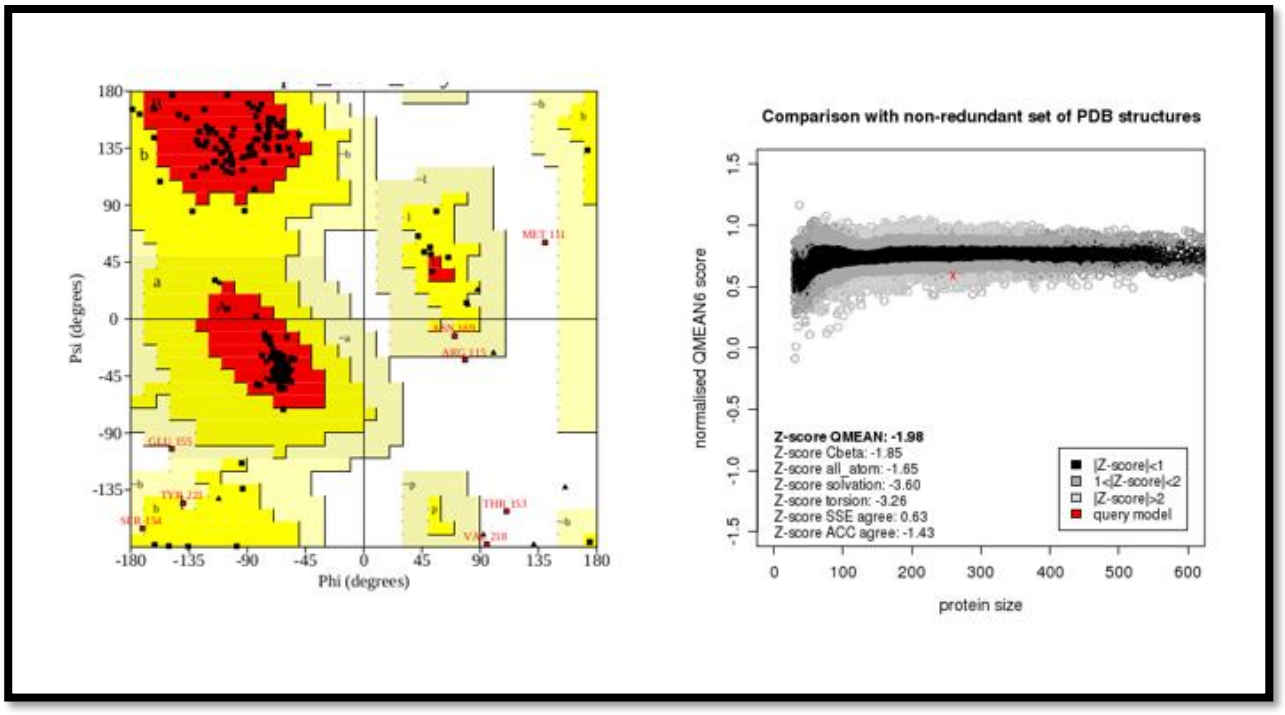


Figure 2. 9: Plots from PROCHECK and QMEAN evaluating tools for the variant 1 G2019S Kinase domain model.

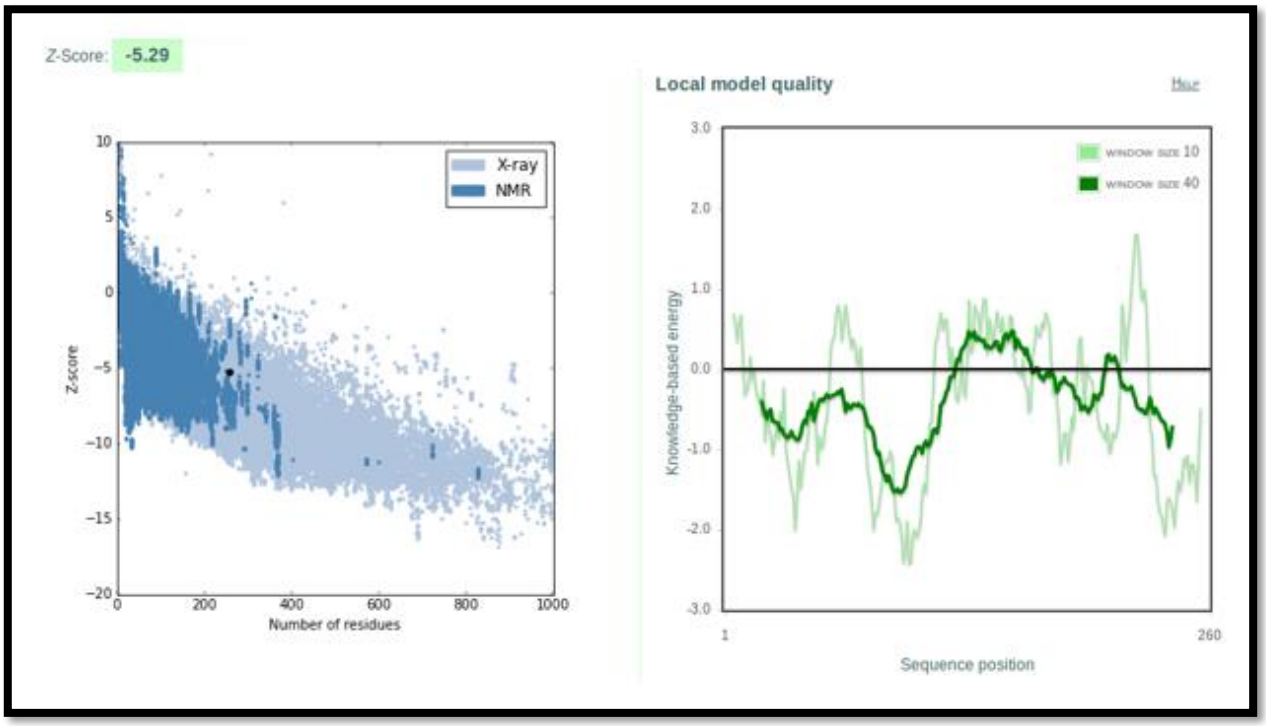


Figure 2. 10: Plots from ProSA evaluating tool for the variant 2 I2020T Kinase domain model.

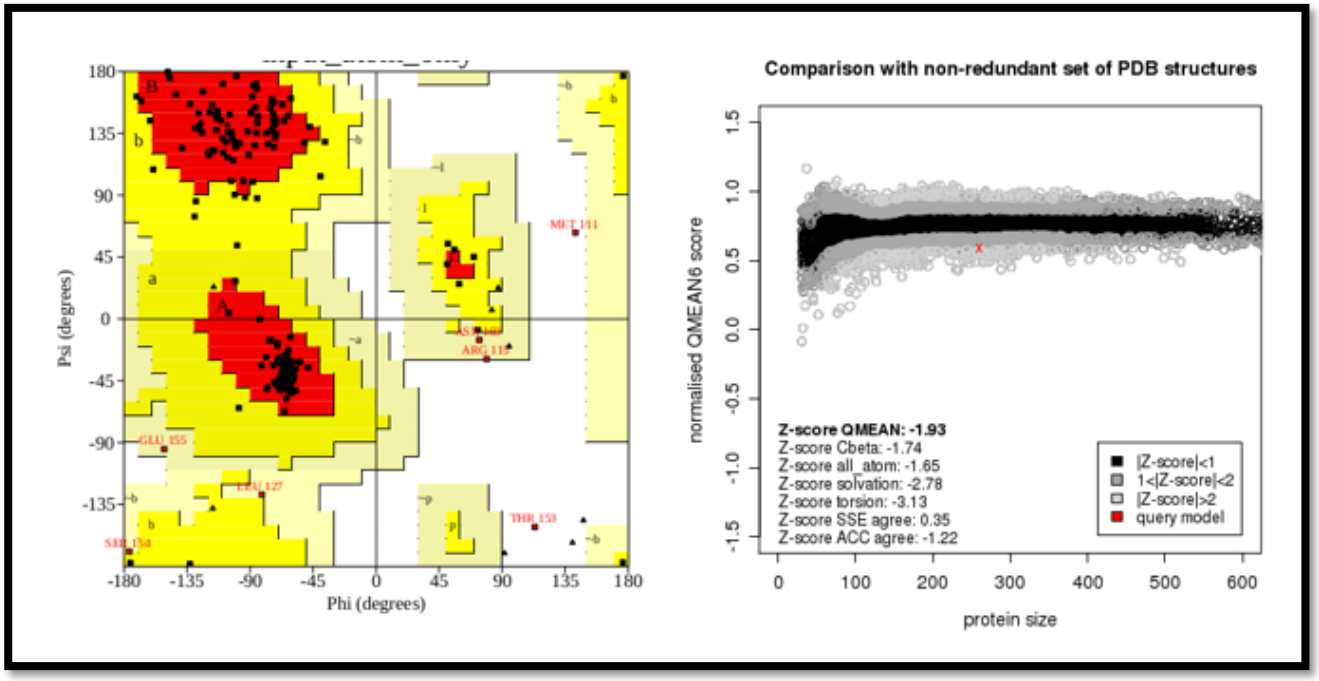


Figure 2. 11: Plots from PROCHECK and QMEAN evaluating tools for the variant 2 I2020T Kinase domain model.

2.6 Chapter summary

Three reliable LRRK2 (kinase domain) models were successfully modelled using MODELLER v9.19. The overall model evaluation using ProSA, PROCHECK and QMEAN showed that the created models were reliable, however, some loops showed errors which is quite common since these regions exhibit most flexibility in a protein. These models were both 'locally in residue level' and 'globally as overall structure' statistically reliable. These models have proved satisfactory for further studying in drug discovery in the effort to identify new Parkinson's disease therapies.

CHAPTER THREE: High throughput virtual screening

3.1 INTRODUCTION

In this chapter, the generated models of LRRK2 (kinase domain) wildtype, the G2019S variant model and the I2020T variant model are exposed to high throughput virtual screening. The 623 small molecule compounds retrieved from South African Natural Compound Database (SANCDB) (Hatherley *et al.* 2015), were used to screen against each protein structure. Molecular docking will be discussed in this chapter, how it operates and virtual screening. The focus will also be given to the methodology that was used to screen each protein conformation in detail and finally discuss the results and findings obtained.

3.1.1 Molecular Docking

Molecular docking is a computational method used to predict the interaction of two molecules generating a binding model (Prieto-Martínez, Arciniega and Medina-Franco 2018). This computational procedure predicts the non-covalent binding of macromolecules (receptor) and a small molecule (ligand) starting with their unbound structures obtained from Molecular dynamic simulations or homology modeling. The human genome project through determining sequence of nucleotide base pairs that make up human DNA, has resulted in much data that can be analysed for high-throughput protein purification, crystallography and nuclear magnetic resonance spectroscopy techniques (Meng *et al.* 2011).

The binding affinity and the 3D structure of the complex in molecular docking predict whether the two molecules interact, for modern drug discovery. Molecular mechanics is the basis for most docking programs which involves the description of a polyatomic system using classical physics, to narrow down the differences between experimental data and molecular mechanics predictions parameters such as charges, torsional and geometrical angles are used (Prieto-Martínez, Arciniega and Medina-Franco 2018). The prediction of binding of small molecules to proteins is of importance since it is used to screen virtual libraries of drug-like molecules to obtain leads for further drug development, this method can also be used to predict the bound conformation of known binders when the experimental holo-structures are unavailable.

The docking function defining the energetics of the system is the foundation for the methodology and the docking algorithms attempts to optimize this target function The docking function goal is to discriminate between the manifold of true solutions usually defined as a root mean square deviation

(RMSD) of 2.0Å and false solutions or structures not properly docked (Wu *et al.* 2003). It is important to note that molecular docking has its limitations like the lack of standardised docking flow for testing and validating the results and not all docking algorithms are suited for any given system. It therefore, should be used with other computational and experimental methods (Prieto-Martínez, Arciniega and Medina-Franco 2018). Virtual screening is more direct and rational drug discovery approach since it aims to understand the molecular basis of a disease and utilises the knowledge of three dimensional structure of the biological target in the process and has the advantage of low cost and effective screening (Meng *et al.* 2011).

3.1.2 Virtual screening

Molecular docking is a typical choice for receptor structure-centric virtual screening when the structure of a protein target is available. Small molecules are fitted into the structure of receptor proteins evaluating their binding affinity using scoring systems usually constituted by semi-empirical potential functions and the advantage of virtual screening tools is that they can provide the binding mode of a molecule in a given target protein as well as the binding affinity (Lee and Zhang 2012). The primary objective in virtual screening is to access large libraries of compounds and rank active compounds ahead of inactive compounds and this degree is known as 'enrichment' (Madhavi Sastry *et al.* 2013). The generation of libraries of stereochemically good quality compounds demand a significant investment of resources and time, however, the use of *in silico* high throughput techniques for conducting screening protocols has been found to be cost effective and time-saving in comparison with wet-lab or experimental high-throughput techniques (Brown and Tastan Bishop 2017).

Virtual screening techniques require expert knowledge and extensive infrastructure and is not available to be used by many medically and biologically oriented investigators (Irwin and Shoichet 2005). There are a number of docking programs that can be sued for example AutoDock, MCDOCK, ICM-dock, DOCK, FlexX, Surflex, ICM, Glide, Ligandfit, Cdocker, FRED, MOE-Dock, LeDock, AutoDock-Vina, rDock, UCSF Dock, QXP, GOLD, CHARMM and many more (Bursulaya *et al.* 2004). There are several free small molecules that are available in databases but they are all not entirely acceptable for docking hence it is always good to double check with literature and be cautious when using these ligands (Irwin and Shoichet 2005).

Virtual screening can be divided into two key concepts: ligand-based virtual screening (LBVS) which operates in the absence of a target protein structure (receptor) but the active compound as a model for the screening procedure, it also comprises of 2D or 3D similarity screening, small compounds based pharmacophore screening among others and the second is structural-based virtual screening (SBVS)

which involves docking of candidate compounds into a receptor (protein) structure and it comprises of molecular docking, 3D protein structure-based pharmacophore and de nova design. Autodock vina uses an empirical scoring function which is inspired by the X-score function to estimate the ligand-receptor affinity (Quiroga and Villarreal, 2015). Scoring functions are mathematical functions used to approximately predict the binding affinity between two molecules after molecular docking, scoring functions are developed based on physical atomic interactions (Huang,2010).

The general concept of one ligand one receptor for a specific biological response is inadequate because we usually see several drugs being administered to patients to treat a disease. Multi-drugs interact with their respective targets, causing a series of biological responses. It is also believed that multi-target drugs can present improved efficacy, safety and even synergistic activity in differing disease-related targets (Scotti *et al.* 2015).

3.2 Chapter Objectives

To identify compounds from the SANCDB that can bind to the active site of the LRRK2 protein and form hydrogen bonds.

3.3 Steps involved in molecular docking

Receptor and ligand preparation, visualization and docking simulation were performed using AutoDock-Vina, PyMOL, Discovery studio Visualiser and Ligplotplus. Preparation of the ligand involved the addition of polar hydrogen molecules and defining rotational bonds used for flexible docking. Set various parameters during simulation such as the appropriate pH, structure optimization and partial charge calculations using molecular mechanics or semi-empirical quantum chemical methods, save the ligands in *pdbqt* format.

In protein preparation crystallised water are removed, add polar hydrogens and save the structure in *pdbqt* format. In setting up the simulation box, select the known binding site through a co-crystallised ligand unless performing blind docking then there is no need to select a site. Select the center mass of the protein, select the coordinates of the box center. Analyse the binding affinities of the dockings in which lower negative values are stronger ligand binding energies. PyMOL, Discovery studio visualizer, Ligplotplus and other programs can be used to analyse the docked results, which is important for ranking a ligand that might be a potential hit compound. A high throughput virtual screening was performed against the best LRRK2 models, using SANCDB compounds. These proteins were called receptors as this is the general term in docking. High throughput virtual screening was performed on LRRK2 protein structures modelled using novel molecules from SANCDB.

Receptor preparation involved elimination of water molecules, the addition of polar hydrogens and allocation of partial charges using AutoDockTools. Rigid protein structure was used. Flexible ligands were used with the varying number of torsions as assigned by AutoDockTools. Docking validation is done to control the docking ability of AutoDock-Vina to produce correct poses evaluated. Docking was done between a small molecule and a macromolecule, for example, protein-ligand docking, in many drug discovery applications. Experimental parameters such as charges, torsional and geometrical angles are used to narrow down the difference between experimental data and molecular mechanics predictions. Even though there are many robust docking programs available, one should bear in mind that not all docking algorithms are suited for any given system. It is generally advisable to use more than one docking program: different studies have shown that, overall, taking a consensus from various docking protocols yields better assessment of protein-ligand interactions and more reliable pose ranking, however this study used Autodock vina only because it is free for academic purposes and fast (Ferreira *et al.* 2015).

3.4 METHODOLOGY

Virtual screening was performed on the three structure models that were created (as mentioned in CHAPTER 2) using 623 SANCDB compounds. The protein structures will be referred to as 'receptors' and the SANCDB compounds will be referred to as ligands. Docking was performed using AutoDock-Vina. Figure 3.1 below, shows the steps of molecular docking. In which the 623 SANCDB compounds were docked to all three LRRK2 structures using AutoDockTools and python scripts. PyMOL and Discovery studio Visualiser was used to visualise the docking results.

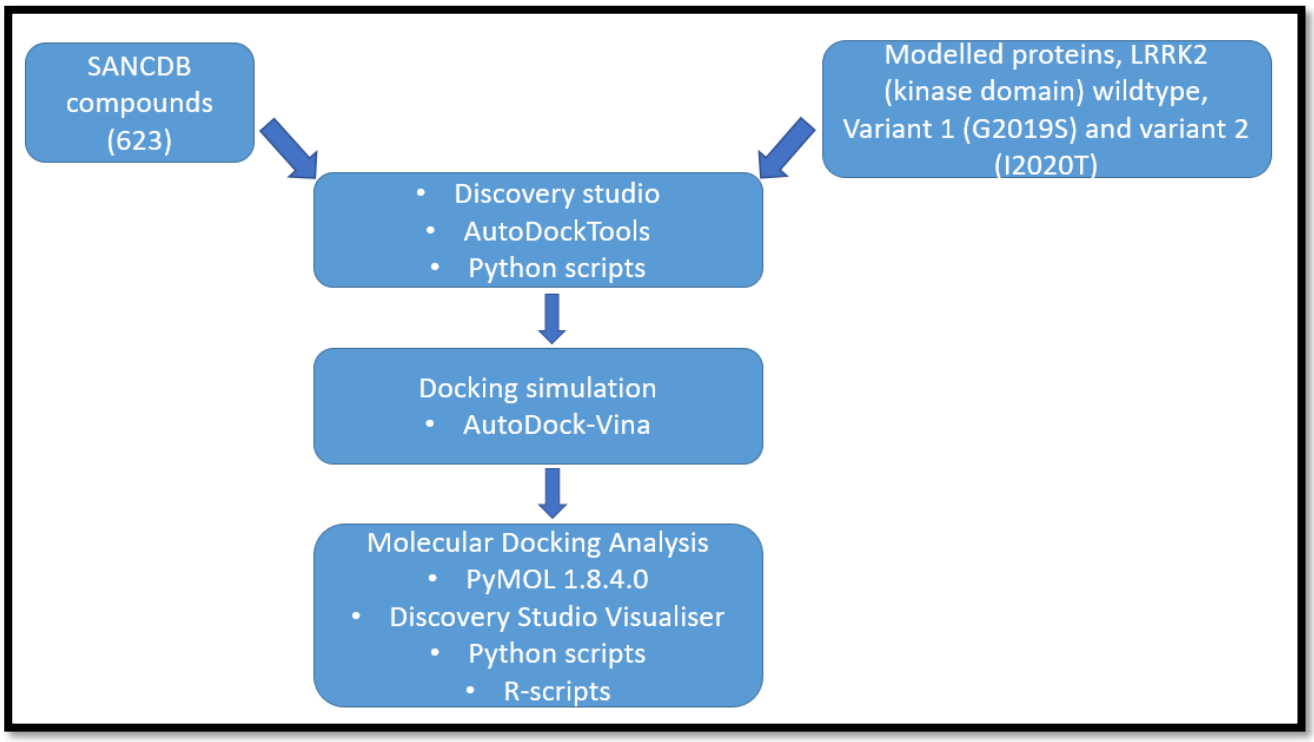


Figure 3. 1: Molecular docking steps and visualising tools used.

Figure 3.2 below, shows the molecular docking parameters that were used. Blind docking was performed in YODA cluster using AutoDock-Vina for the three structures, although the active site of the protein is known blind docking was performed, so not to restrict the ligands from binding to desired regions. AutoDock-Vina plugin in PyMOL was used to design the docking box and centres. The number of steps in a run is determined heuristically depending on size and flexibility of the ligand and the flexible side chains and the number of runs is set by the exhaustiveness parameter. Blind docking was performed and table 3.1 shows the parameters that were used to create a box that covered the whole protein before performing molecular docking.

Table 3. 1: Molecular docking parameters.

Box parameter	Value
Center x	10.007
Center y	-23.906
Center z	36.051
Size x	52.289
Size y	65.091
Size z	57.915
Energy range	4
CPU	4
Exhaustiveness	96

3.5 Docking analysis

Results that came from AutoDock-Vina were in form of binding geometries denoted as binding energies (kcal/mol) and these energies were ranged using a Python script. PyMOL was used to visualise the docked results. Three factors were considered for choosing ligands of interest, the best ligands were selected based on how close they were to the active site, high binding affinity (lowest binding energy) and the size of the ligands.

Reason for this selection criteria was the closer the ligand is to the active site increases the chances of its interaction with the active site residues. The lower or more negative the binding energy is better because this is the minimum energy required to disassemble a system of particles into separate parts. The smaller the ligand the better because it makes it easy to work with and for the possible drug to pass through the blood-brain barrier. A Python script was used to calculate the docked distance between the docked ligand and the active site and bond interactions formed by the receptor-ligand were obtained using Ligplotplus and Discovery Studio Visualiser. The top three ligands for each structure were chosen based on the three factors that were mentioned above and these were considered the best hit compounds. Microsoft Excel as seen in Appendix 3.1 and 3.2 and R package were used for data management.

To evaluate drug-likeness or determine if these compounds with certain pharmacological or biological activity have chemical properties and physical properties that would make it orally active drug in humans. Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio) website[7] was used to calculate the Lipinski rule of five.

Lipinski rule of five calculates the drug-likeness of the ligands[8]. The rule states that an orally active drug has no more than one violation of the five criteria's in Table 3.2 below. Drug-likeness is a complex balance of various molecular properties and structural features which determine whether a molecule is a drug or non-drug. The molecular weight should be less than 500. Log P value should be less than 5. The number of hydrogen bonds should be less than 10. Three ligands out of the 623 natural compounds were selected for each protein structure based on low binding affinity, size of the ligand and how close it was to the active site of the protein. Generally, the rule approved the chosen ligands and these ligands may be ideal drug candidates for further lead optimisation experiments in *in vitro* investigations.

Table 3. 2: Lipinski rule of five and the chosen ligands passed.

ID	Moleculer weight (daltons)	logP	Hydrogen bonds donors	Hydrogen bonds acceptor	Lipinski violation
Wt_237	392.53	5.48	0	5	1
Wt_525	472.66	3.78	2	5	0
Wt_595	488.7	5.08	0	5	1
Vt_19_101	290.27	0.51	5	6	0
Vt_19_458	162.18	-0.58	2	4	0
Vt_19_467	154.12	0.86	3	4	0
Vt_20_116	337.41	1.84	1	6	0
Vt_20_279	536.65	1.06	5	9	1
Vt_20_368	288.30	3.36	2	5	0

[7] http://www.scfbio-iitd.res.in/software/drugdesign/lipinski.jsp#anchortag
[8] www.sciencedirect.com

3.6 Results and discussion

Natural compounds have more diversified properties and 623 natural compounds from SANCDB[9] were used for docking in AutoDock-Vina. A large box size that covered the entire protein was used for blind docking and this process was used to identify possible binding sites. PyMOL was used to view how the ligands docked on the structures and active site ligands were of interest. Most ligands that showed lowest binding energies were a little far from the active site which was not desirable. AutoDock-Vina was used because of its accuracy in reproducibility and 623 natural compounds in SANCDB were used due to their diversity properties. Figures 3.2, 3.3 and 3.4, shows the three protein docking results viewed in PyMOL and it is shown as a grey surface body and the active site as the green shading. These are the docked results of all 623 compounds on the three structures and the green part represents the active site of the structure. The surface protein structures were generated in PyMOL surface compounds in which the grey bodies represent the structures and the multi-coloured sticks represent the different ligands.

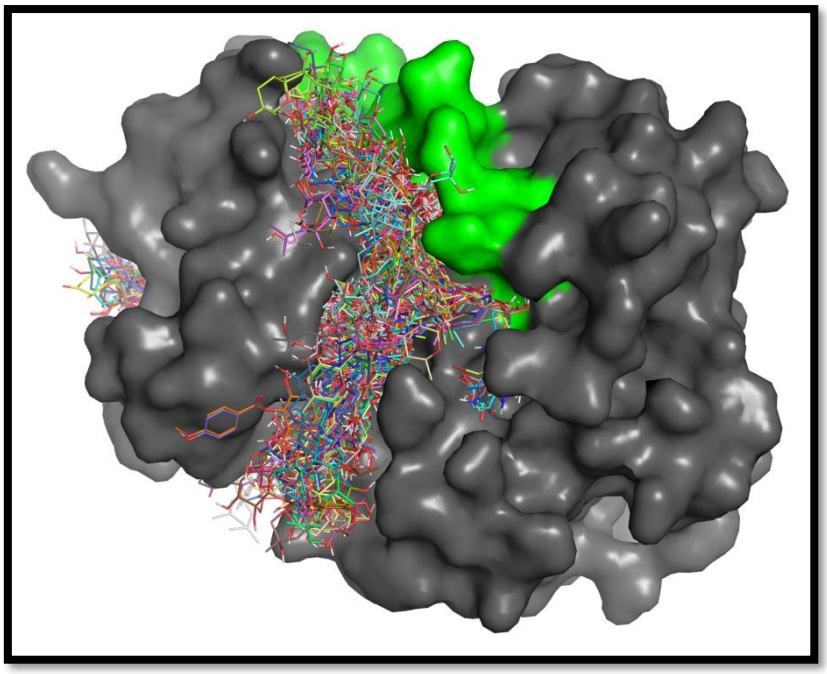


Figure 3. 2: SANCDB (623) ligands docked to wildtype.

[9] (https://sancdb.rubi.ru.ac.za/)

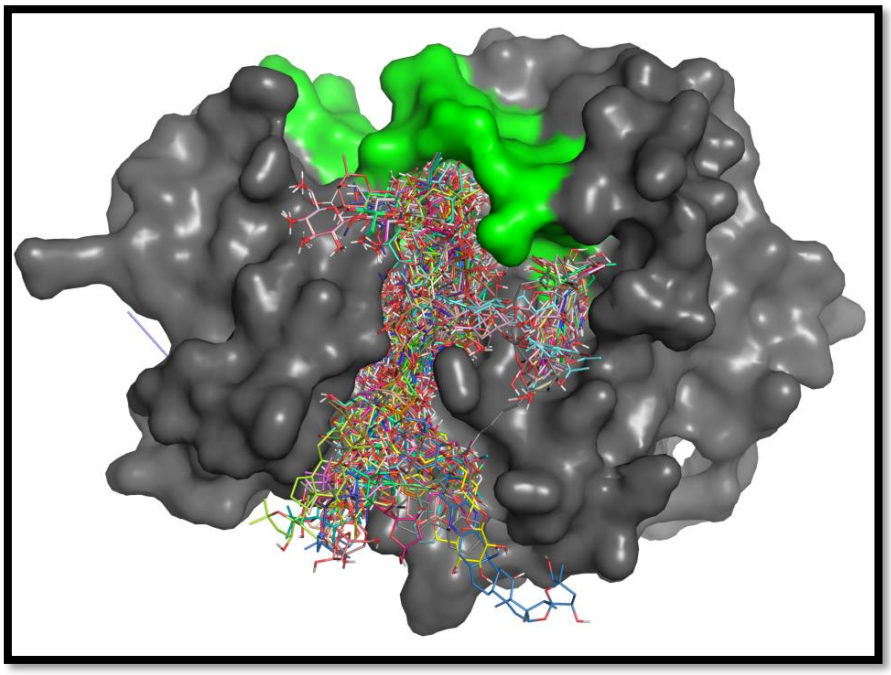


Figure 3. 3: SANCDB (623) ligands docked to variant 1 (G2019S).

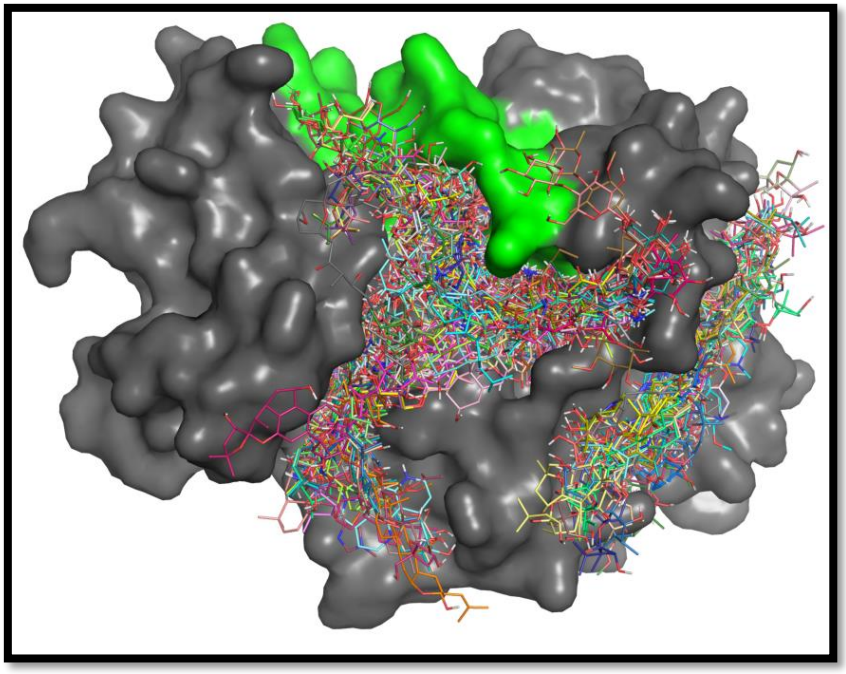


Figure 3. 4: SANCDB (623) ligands docked to variant 2 (I2020T).

3.6.1 Screening the ligands

As Figures 3.2, 3.3 and 3.4 above, had shown, most of the ligands bound further away from the active site which made the screening process a little easy. The study aimed to remain with at least 3 best ligands that were going to be further studied in Molecular dynamics in (CHAPTER 4). In post-docking analysis filtering candidate, SANCDB ligands were based on the potential and empirical scoring functions estimated by AutoDock-Vina. The following flow shows the steps taken in filtering the ligands. The ligands were ranked in Figure 3.5 below, based on the binding energy, interaction with the protein structures, falling in the active site region, ligands forming hydrogen bonds with the protein structures and druggability based on the Lipinski rule of five results:

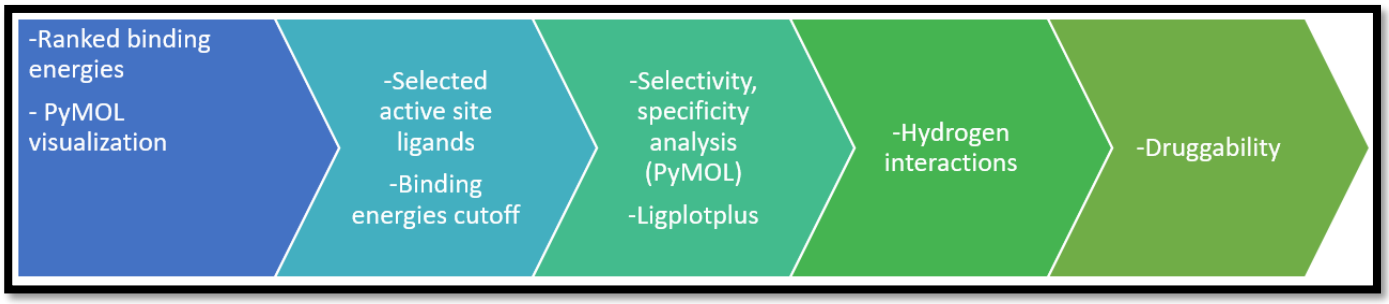


Figure 3. 5: The summary flow diagram of how docked ligands were screened or filtered, the numbers and arrows show the flow.

3.6.2 Binding energies and ligand visualisation

Generally, the ligands proved quite stable in the binding pockets because they showed a high binding affinity (lowest binding energies). Red represents low binding-energy scores meaning there is likely binding interaction and blue represent high binding-energy meaning improbable interaction. Ligands binding with low energy scores are not target specific or were generally big, hence were not considered in this study. The heatmap in Figure 3.6 below, shows the binding energies of the three structures ordered in a descending order in a Microsoft excel sheet. Natural compounds docked at the active sites were considered for further analysis for each protein structure.

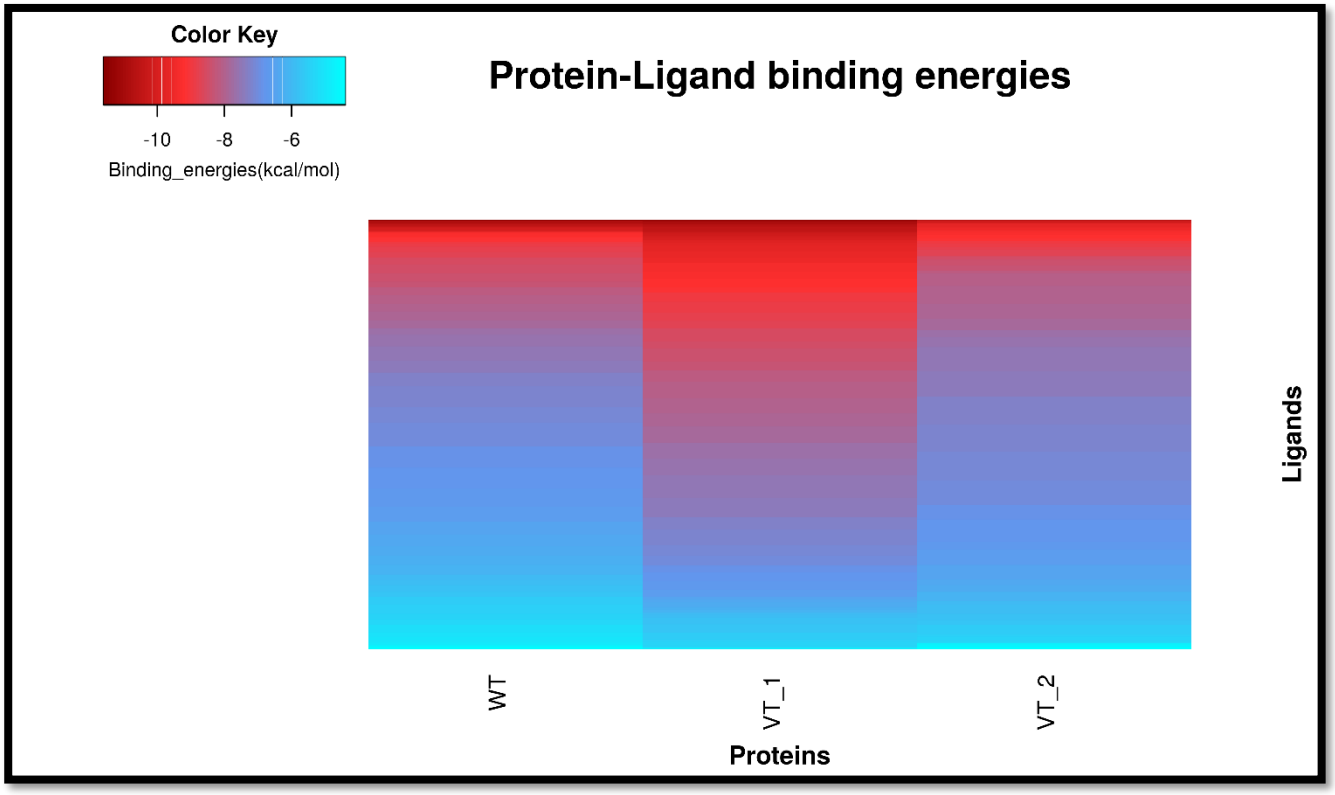


Figure 3. 6: This is a heatmap created in R studio of all the binding energies for each protein structure and the key is also provided in this figure.

3.6.3 Residues and bond interactions of hit compounds

For visualisation of the residues and bond interaction, Discovery studio visualiser and Ligplotplus was used. Three best ligands on each protein structure were selected based on how they were interacting with the active site, the binding energy and the size of the ligand. In Figure 3.7, the grey body represents the structure wildtype protein, the green part represents the active site, the yellow ligand represents the SANC00525, the blue ligand represents SANC00237 and the red ligand represent SANC00595 compound. It was observed that all the ligands were lying in the active site grooves, in the same position and had tails sticking out of it due to their size and they might not be small enough to pass the blood brain barrier.

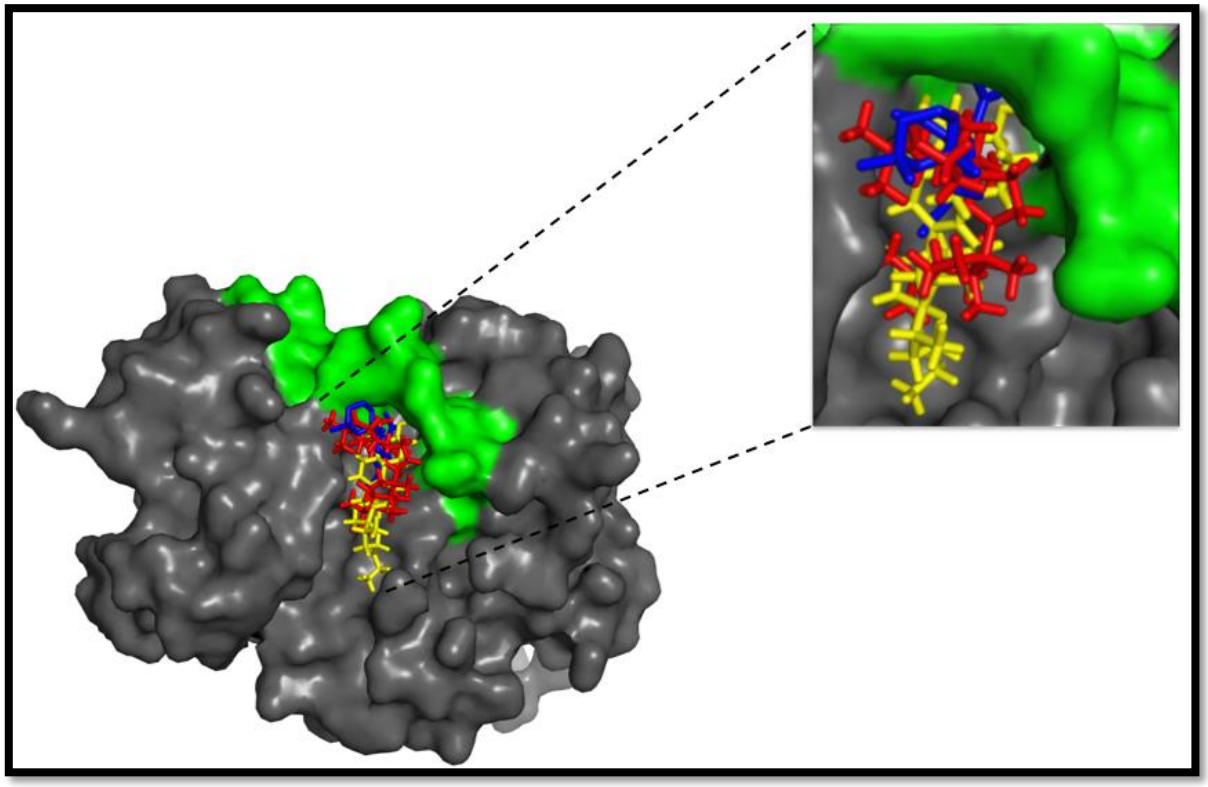


Figure 3. 7: A wildtype surface representation showing the best three ligands bound close to the active site.

In Figure 3.8 below, the grey body represents the structure variant 1 (G2019S), the green part represents the active site, the yellow ligand represents the SANC00101, the blue ligand represents SANC00458 and the red ligand represent SANC00467 compound. These compounds were small and lied in the active site groove, in the same position.

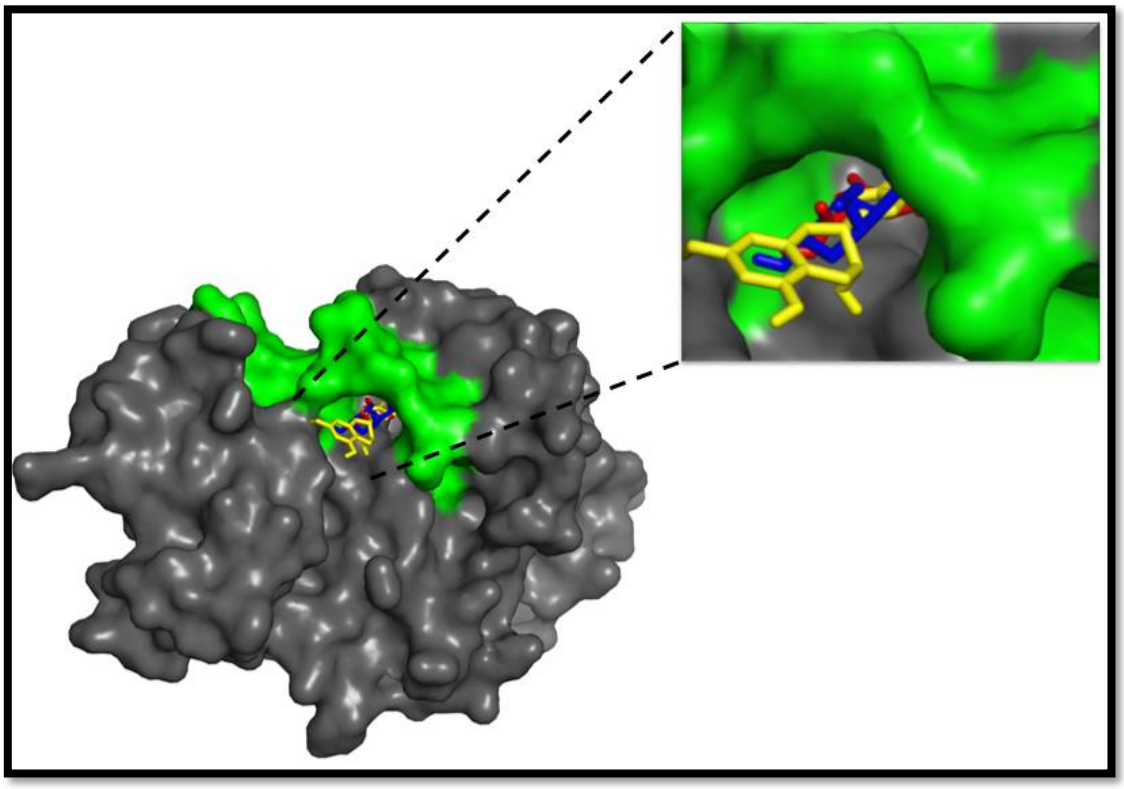


Figure 3. 8: Variant 1 surface representation showing the best three ligands bound close to the active site.

In Figure 3.9 below, the grey body represents the structure variant 2 (I2020T), the green part represents the active site, the red ligand represents the SANC00116, the blue ligand represents SANC00279 and the yellow ligand represent SANC00368 compound. SANC00279 and SANC00368 bound to the same position in the active site groove. SANC00116 was the smallest ligands in comparison to the other two ligands and it bound at a different position as well but still interacted with the active site.

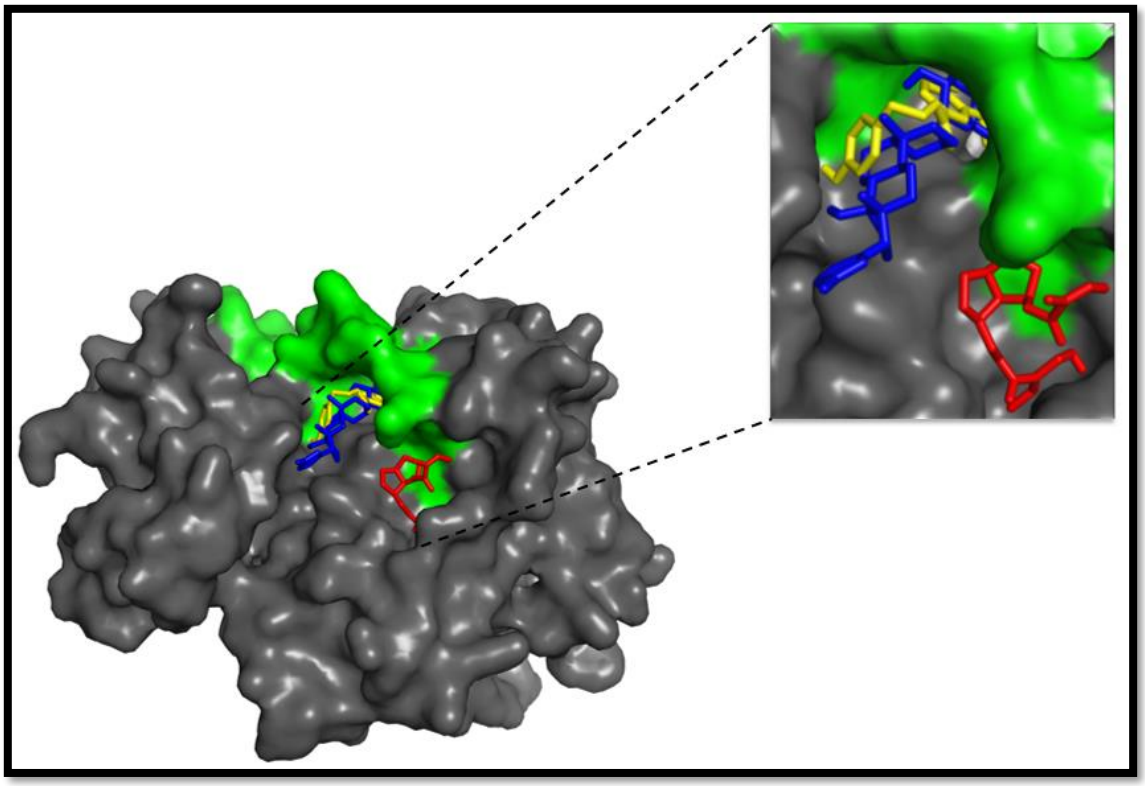


Figure 3. 9: Variant 2 surface representation showing the best three ligands bound close to the active site.

In Figure 3.10 below, the grey body represents the structure wildtype, the green part represents the active site, the yellow ligand represents the SANC00101, the cyan ligand represents SANC00116, the magenta ligand represents SANC00279, the orange ligand represents SANC00368, blue ligand represents SANC00458 and the red ligand represent SANC00467 compound. Figure 3.10 was aimed to see how the best ligands in the two variants were also behaving in the wildtype. It was observed that these ligands had higher binding energies, were far from the active site and were not forming hydrogen bonds with the active site as they did in the variant structures except SANC00279 which made a hydrogen bond with THR157, ASP116, TYR172 and ARG148, it also forms Pi-Alkyl bonds with HIS120. Which again shows us that these ligands docked on different positions of interest in all structures. Figures 3.11, 3.12 and 3.13, represents the 2D structure of the ligands, the species in which they come from and their scientific names.

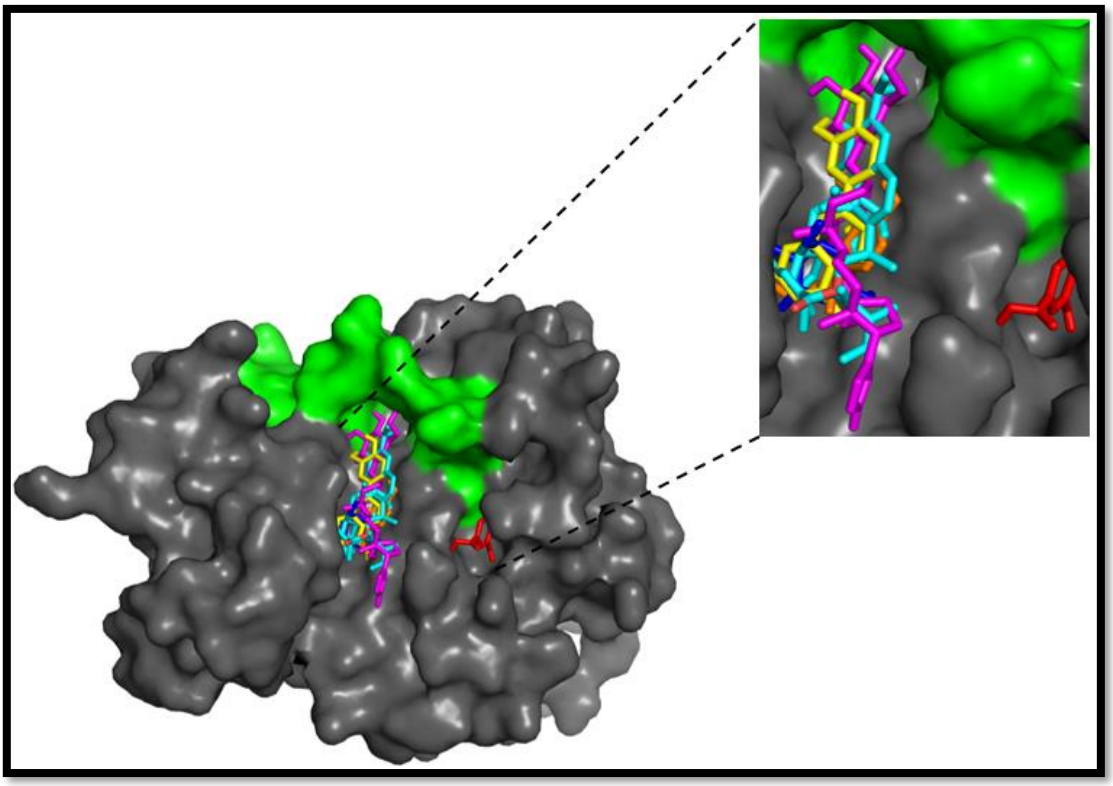


Figure 3. 10: Wildtype surface representation showing the variants best ligands bound to it.

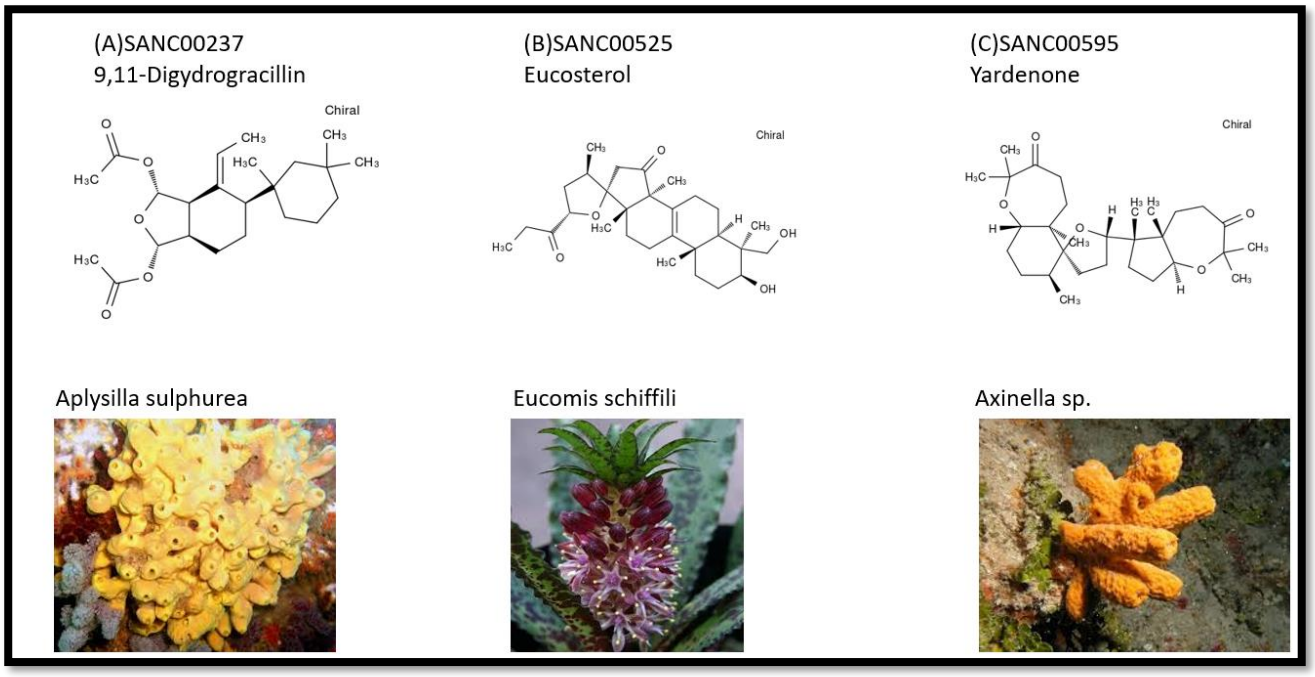


Figure 3. 11: The 2D structure of the three best ligands for the wildtype model with its identity name, entry name and the organism in which it comes from.

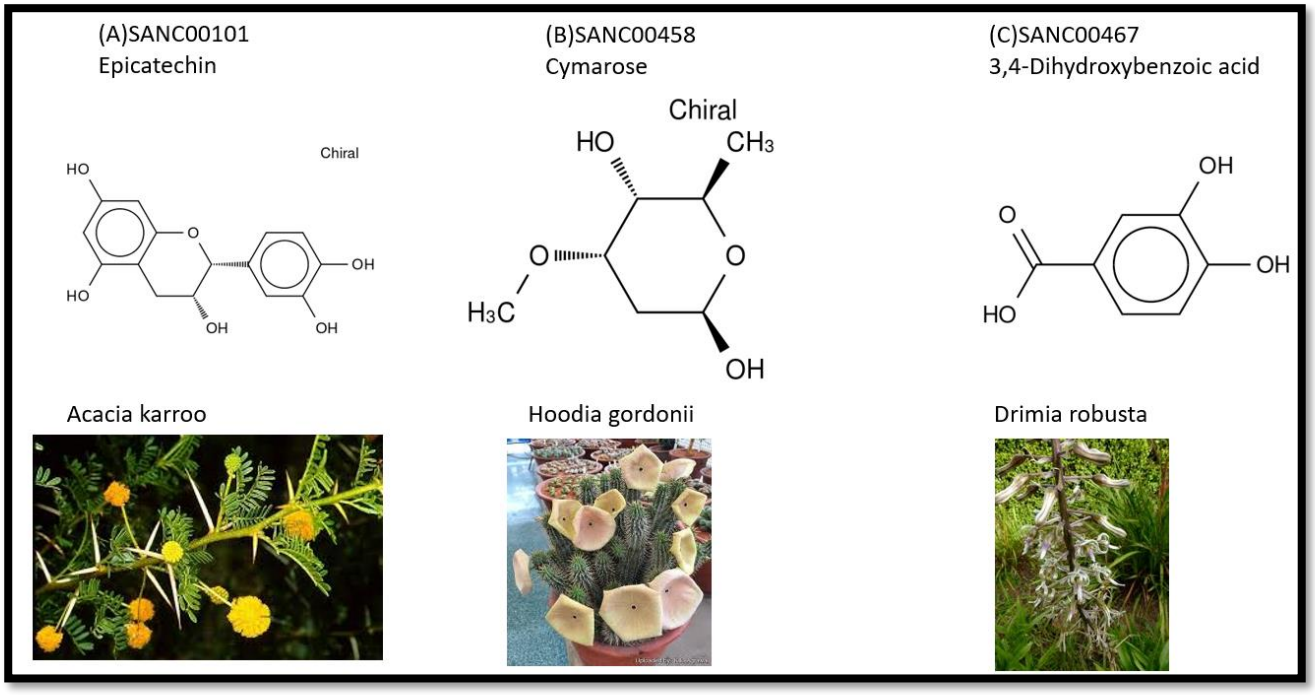


Figure 3. 12: The 2D structure of the three best ligands for the variant 1 (G2019S) model with its identity name, entry name and the organism in which it comes from.

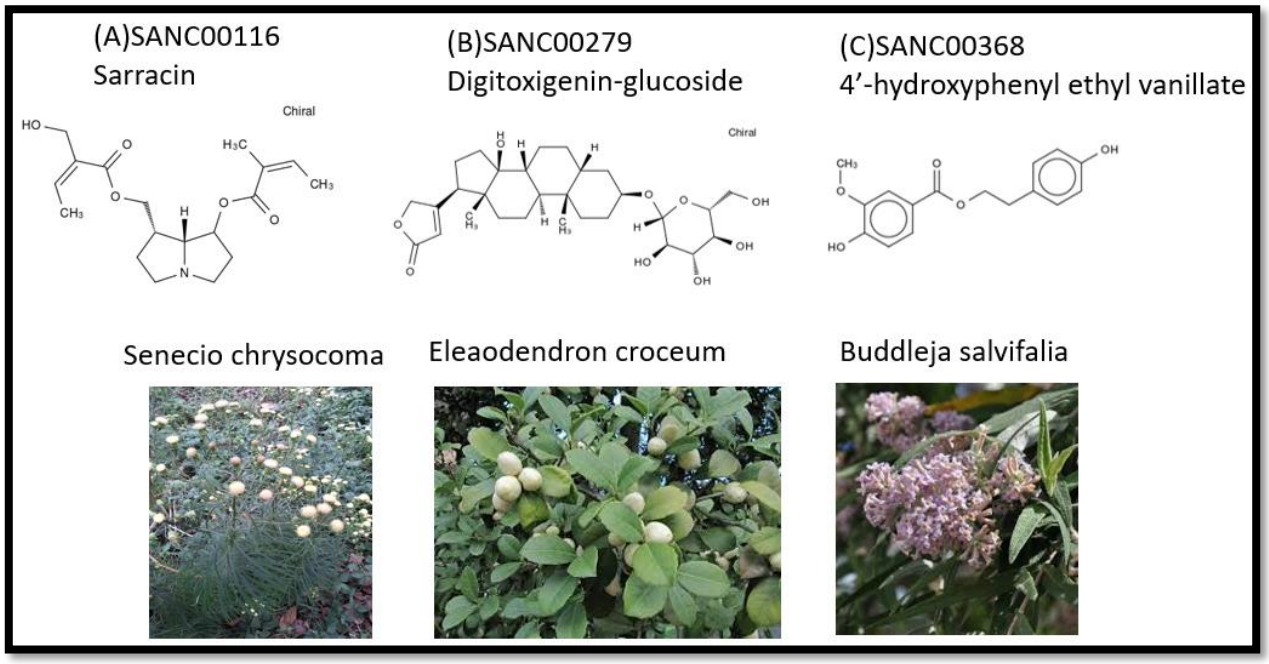


Figure 3. 13: The 2D structure of the three best ligands for the variant 2 (I2020T) model with its identity name, entry name and the organism in which it comes from.

3.6.4 Hydrogen bond interactions

Different tools were used to visualise the bond interactions because each tool, projects or interprets these interactions differently. The carbon-hydrogen bond (C-H bond) is the bond between carbon and hydrogen atoms that can be found in many organic compounds (Merrill and Madix 1991). This bond is a covalent bond that shares its outer valence electrons with up to four hydrogens and a bond length of about 1.09Å and a bond energy of 413 kJ/mol making it stable. A conventional hydrogen bond is a partially electrostatic attraction between a hydrogen (H) atom which is bound to a more electronegative atom or group such as nitrogen (N), oxygen (O) or fluorine (F) the hydrogen bond donor and another adjacent atom bearing a lone pair of electrons the hydrogen bond acceptor (Weinhold and Klein 2015). van der Waals is distance-dependent interactions between atoms or molecules, they are not a result of any chemical electronic bond and they are comparatively weak and more susceptible to being perturbed (Huang *et al.* 2017). Discovery studio visualiser and ligplotplus uses a cut off distance of 3Å between the receptor and the ligand in order to consider the bond a hydrogen bond or not.

In Figure 3.14 below, amino acid side chains interacting with ligands are represented in ball and stick. The ligand is represented in stick form. Dotted lines represent interacting bonds. SANC00237 formed van der Waals forces with Ser154, Thr153, Thr157, Arg148 and Met149 and conventional hydrogen bonds with residues Arg161 and Lys118 although they do not form part of the active site. SANC00525 formed van der Waals with Ser154, Thr157 and Glu155 and conventional hydrogen bonds with Thr153 and Lys152, and all these residues fall in the active site. SANC00595 formed van der Waals with Thr153, Se154, Glu155, Thr157 and Arg148, however, there were no conventional hydrogen bonds formed.

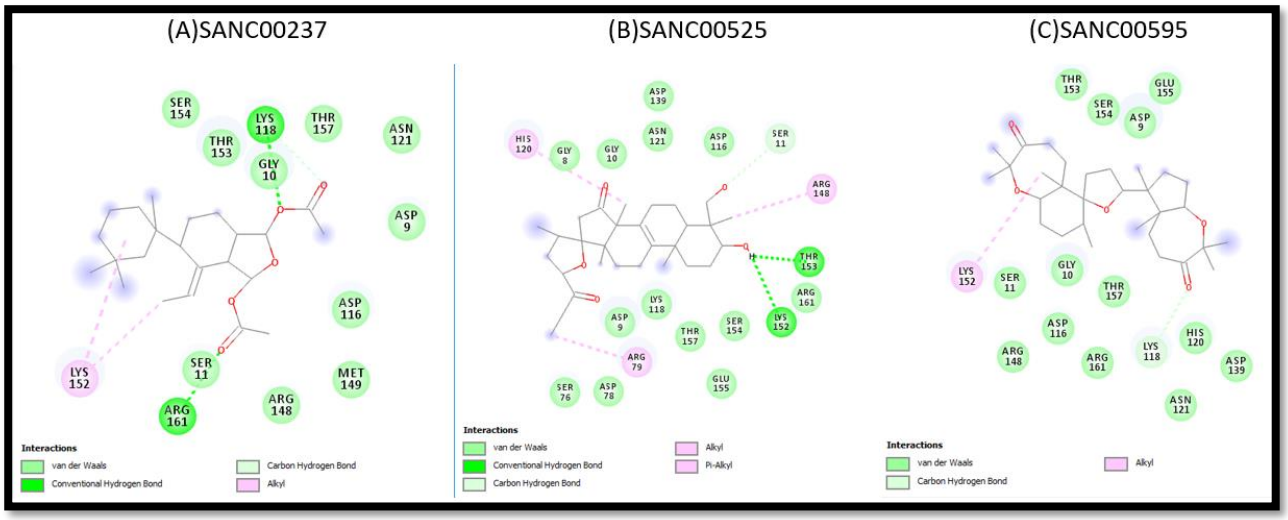


Figure 3. 14: This is the wildtype hydrogen bond interaction viewed in Discovery studio.

In Figure 3.15 below, the spoked red arcs represent the protein residues involved in hydrophobic contacts and the spheres that are joined by green dotted lines show the hydrogen bond and its length. Atoms indicated in black represent Carbon, blue for Nitrogen, red for Oxygen and yellow for Sulphur. In SANC00237 Lys152, met149 and Arg148 are residues involved in hydrophobic contacts. Residues in SANC00525 Thr157, Thr153, Arg148, Ser154 and Lys152 were involved in hydrophobic contacts. SANC00595 had Lys152, Thr153, Ser154 and Thr157 residues involved in hydrophobic contacts.

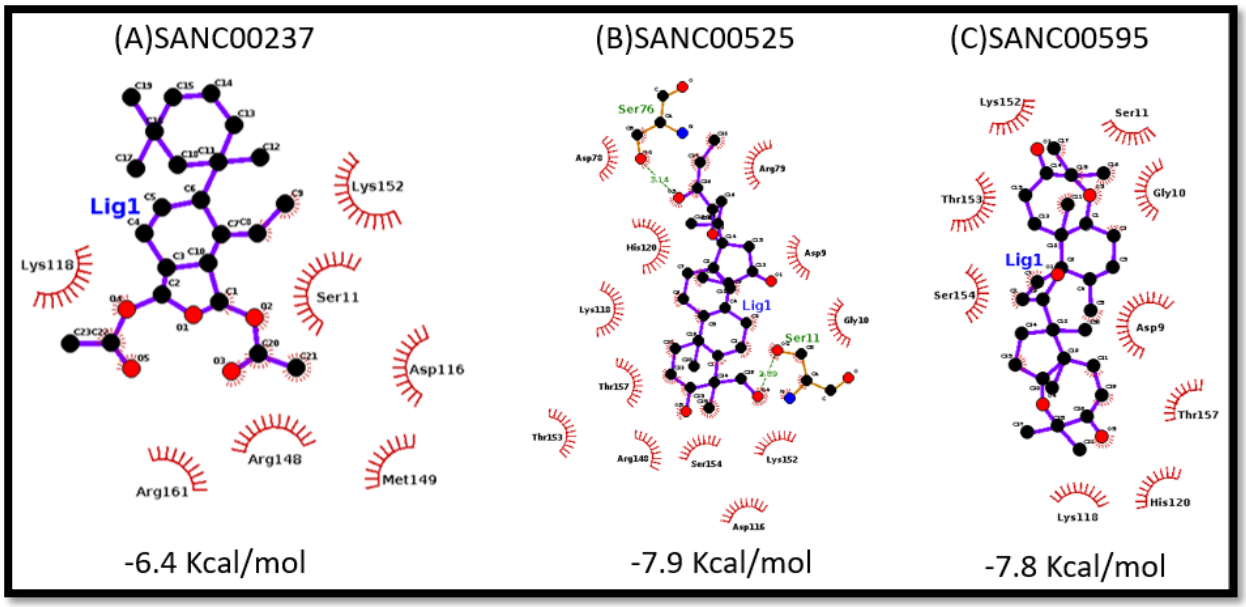


Figure 3. 15: A Ligplotplus visulisation of the hydroogen bonds between the wildtype and the ligands of interest.

In Figure 3.16 below, amino acid side chains interacting with ligands are represented in ball and stick. The ligand is represented in stick form. Dotted lines represent interacting bonds. SANC00101 formed van der Waals connections with Lys152, Met149, Thr153, Ser154 and conventional hydrogen bonds with residues Thr157 forms part of the active site. SANC00458 formed van der Waals with Met149 and conventional hydrogen bonds with Thr153, Ser154 and Thr157, and all these residues fall in the active site. SANC00467 formed van der Waals with Ser154, Thr157, Thr153, Lys152 and Met149 and formed conventional hydrogen bonds with Arg161 and Arg115 residues that do not form part of the active site.

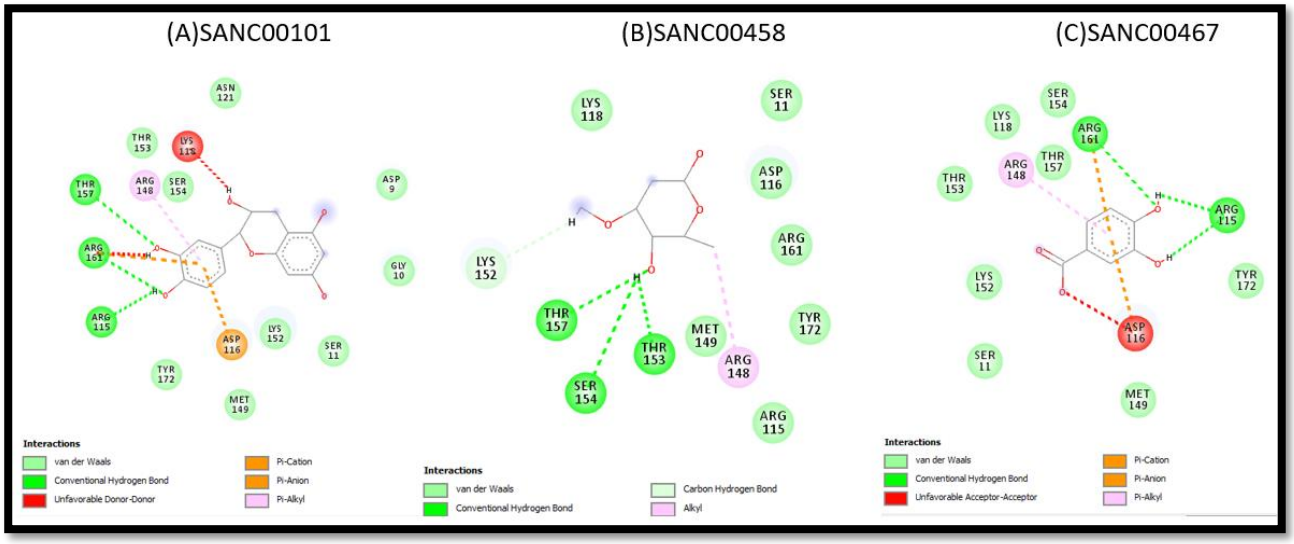


Figure 3. 16: This is a variant 1 (G2019S) hydrogen bond interaction viewed in Discovery studio.

In Figure 3.17 below, the spoked red arcs represent the non-ligand residues involved in hydrophobic contacts and the spheres that are joined by green dotted lines show the hydrogen bond and its length. In SANC00101 Thr153, Arg148 and Lys152 are residues involved in hydrophobic contacts and Arg161 and Thr157 formed hydrogen bonds with the ligand. Residues in SANC00458 Lys152 and Arg148 were involved in hydrophobic contacts and Thr153, Ser154 and Thr157 formed hydrogen bonds with the ligand. SANC00467 had Lys152 and Thr157 residues involved in hydrophobic contacts and residue Arg148 formed a hydrogen bond with the ligand.

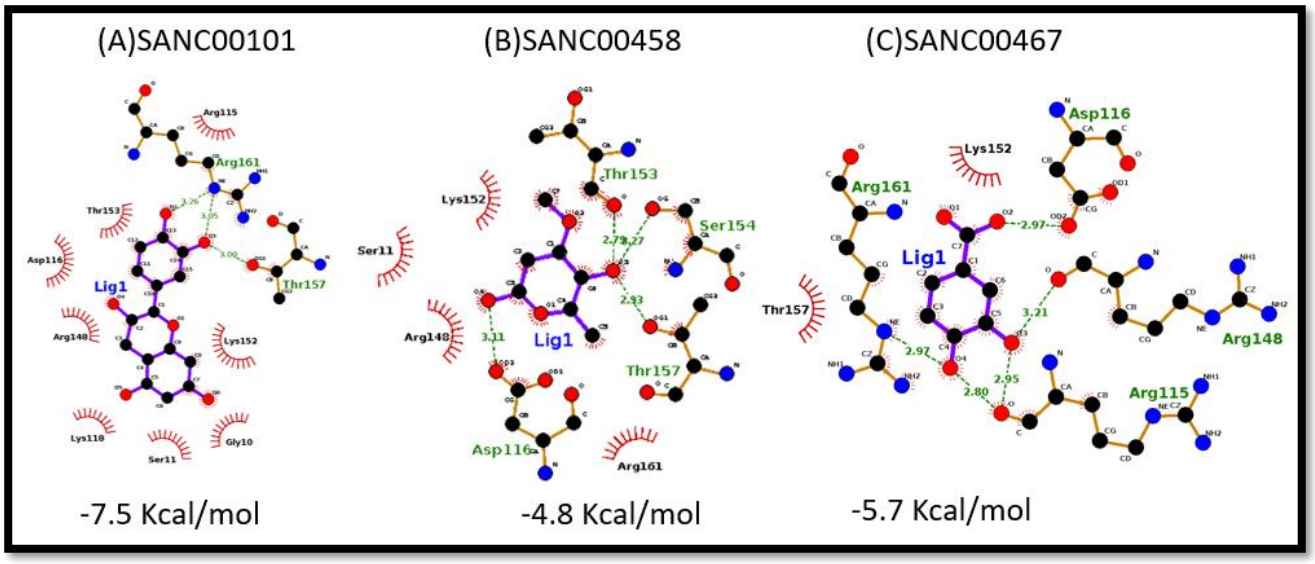


Figure 3. 17: Ligplotplus visualisation of the hydrogen interactions between variant 1 and ligands of interest.

In Figure 3.18 below, amino acid side chains interacting with ligands are represented in ball and stick. The ligand is represented in stick form. Dotted lines represent interacting bonds. SANC00116 formed van der Waals forces with Gly156, Glu155 and Thr157 and conventional hydrogen bonds with residues Asp187, Gly193 and Arg194 although they do not form part of the active site. SANC00279 formed van der Waals with Phe160, Pro158, Met149, Thr153, Lys152, Ser155, Glu155 and Gly150 and conventional hydrogen bonds with Thr157 and Arg148, and all these residues fall in the active site. SANC00368 van der Waals with Thr153, Met149, Thr157, Phe160, Ser154 and Lys152, conventional hydrogen bonds with Asp1399 and Arg115 although they do not fall in the active site.

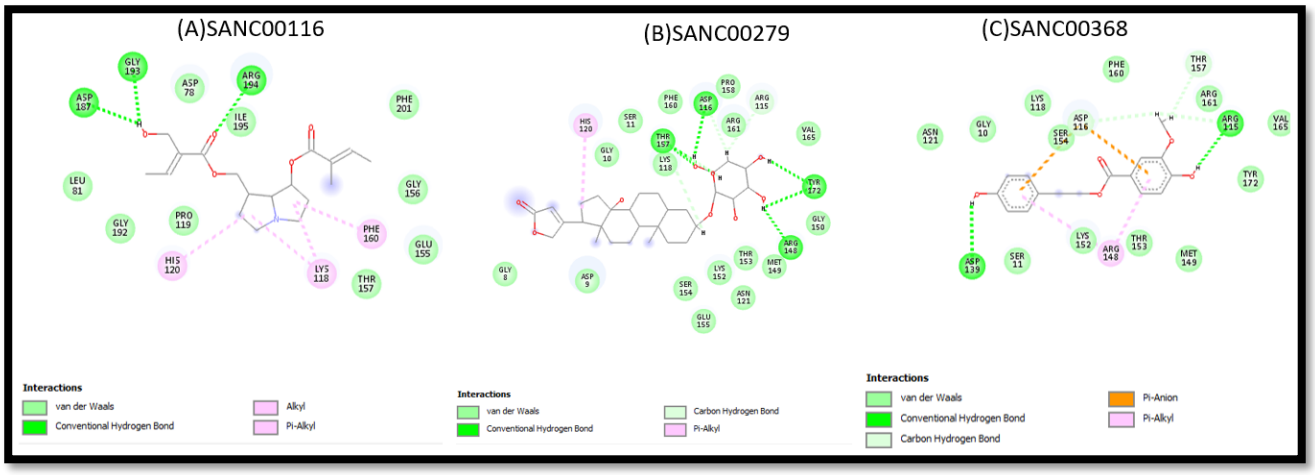


Figure 3. 18: This is the hydrogen bond interaction viewed in Discovery studio.

In Figure 3.19 below, the spoked red arcs represent the non-ligand residues involved in hydrophobic contacts and the spheres that are joined by green dotted lines show the hydrogen bond and its length. In SANC00116 Glu155, Phe160, Thr157 and Gly156 are residues involved in hydrophobic contacts and Ile195, Gly193, Asp187 and Arg194 formed hydrogen bonds. Residues in SANC00279 Phe160, Gly150, Thr153, Thr157, Ser154 and Lys152, were involved in hydrophobic contacts and formed hydrogen bonds with Arg148 and Met149. SANC00368 had Met149, Thr153, Lys152 and Thr157 residues involved in hydrophobic contacts and hydrogen bonds were formed with Arg148.

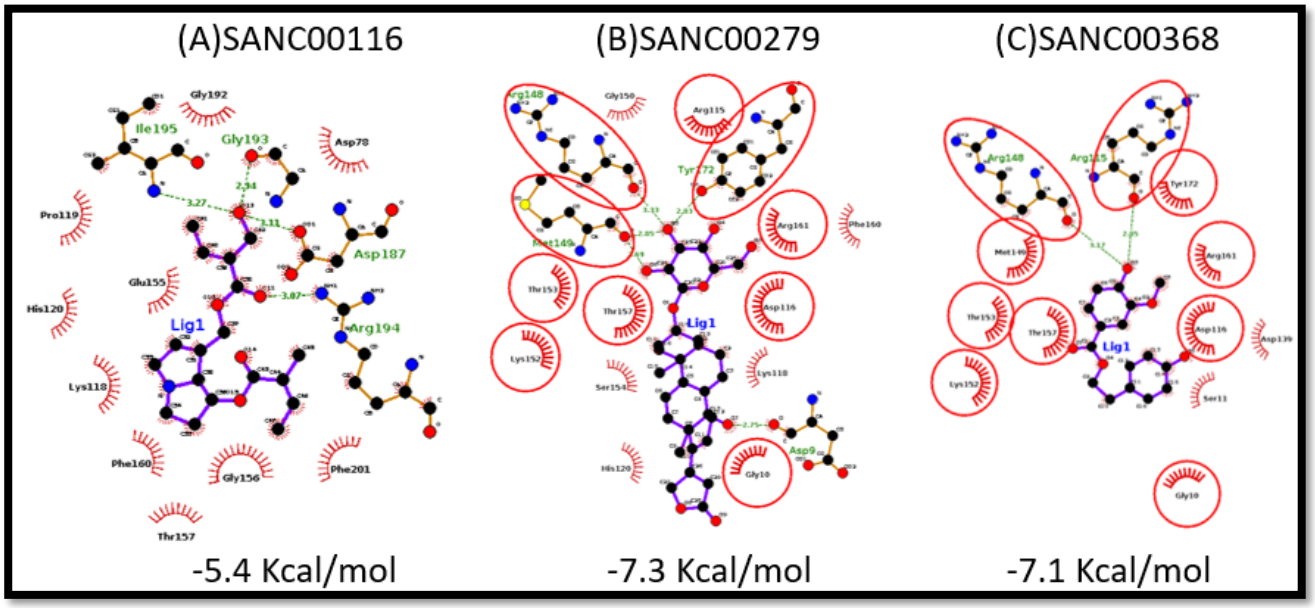


Figure 3. 19: Ligplotplus visualisation of the hydrogen bonds between variant 2 and ligands of interest.

3.7 Chapter summary

Virtual screening approach of the 623 SANCDB compounds on the three structures was successful in identifying high affinity binders. Molecular docking was performed in the RUBi cluster using AutoDock-Vina. The compounds bound differently in all three structures; however, the positions were on almost the same groove of the structures. The compounds had different sizes in terms of length hence they bound differently besides the different chemical compositions.

Due to difference in sizes some ligands were long and stretched all the way to the active site region, however with majority of its body outside this region. The ligands were selected based on size, interaction with the active site and its binding energy. The binding energies of the docked results were then sorted in Microsoft excel, ranking from the highest binding energy to the lowest. The natural compounds generally behaved differently from each other in all three structures. The three structures showed multiple inhibitor binding sites and the ligands were binding on almost the same groove. Molecular docking proved that indeed LRRK2 can be selectively inhibited using small natural compounds. The heatmap showed that variant 1 was the most susceptible to compound inhibition. Three best ligands for each structure were selected for further studying based on their binding energy and binding site. Some ligands interacted with the active site and hence were subjected for further studying in CHAPTER 4. The ligands of choice for each structure formed hydrogen interactions with the active site and the active loop or conserved region of the structures, Figures 3.14 till 3.19 shows the interactions. These interactions can be further studied to see if they do change the behaviour or functioning of the proteins. Ligplotplus and Discovery studio visualiser were used to study the protein-ligand interactions.

CHAPTER FOUR: Molecular dynamics simulations

4.1 Molecular dynamics

Molecular dynamics is a technique that is used to understand the movement of atoms of macromolecules in the simulated cell environment. The simulations produce trajectories depicting the motions of atoms by presenting the atomic coordinates at specified time intervals allowing for investigation into changes over time (Brown *et al.* 2017). This process is based on the molecular mechanic's principle which functions by applying potential energy functions. Molecule-level behavior determines the macroscopic properties, the simulation of a system at atomic level can show the quantitative or qualitative information of the macro-molecules and this simulation is done in a process called Molecular dynamics that calculates the motion of the atoms in a molecular assembly using Newtonian dynamics to determine the net force and acceleration experienced by each atom (Illinois). According to (Berendsen 1999), molecular dynamics is a mature technique for classical simulations of all-atom systems in the nanosecond time range and it is in its infancy in reaching reliably into longer time scales.

Molecular dynamics by providing spatial and temporal resolution that are not available in experiments have expanded greatly the scope of chemistry and several other fields, with better force fields simulations, have become more accurate they easily sample molecular motions on the μs scale and ensemble techniques allow one to study millisecond scale processes such as protein folding (James *et al.* 2015). To understand the properties of assemblies of molecules in terms of their structure and the microscopic interactions between the computer simulation is performed and there are two main simulation techniques molecular dynamics (MD) and Monte Carlo (Allen 2004).

A molecule is a series of charged points (atoms) joined together by springs (bonds) and a force field which is a collection of equations and associated constants designed to reproduce molecular geometry and selected properties of tested structures are used to describe the time evolution of bond lengths, bond angles and torsions, also non-bonding van der Waals electrostatic interactions between atoms (Illinois). Molecular dynamics can be performed in many different software packages like GROMACS, AMBER, NAMD, CHARMM, LAMMPS and Desmond. The scientific roots of the technique in molecular dynamics trace back to polymer chemistry and structural biology in the 1970s used to study the physics of local molecular properties flexibility, distortion and stabilisation and stabilises the X-ray structures of proteins on short time scales (Pronk *et al.* 2013).

Molecular dynamics is the only method that shows the complete molecule. This process is used to see what biomolecules do and their role in the biological system. Using simulations, one can see how the protein aggregates at the atomistic level and describe the adsorption and diffusion mechanisms. The trajectories in molecular dynamics of molecules and atoms are determined by numerically solving the classical equations of motion (Newton's equations) for a system of interacting particles, where forces between the particles and potential energy are defined by molecular mechanics force fields (Ossowska).

Computer simulations were used to see the physical movements of the molecules. There was modeling of molecular systems by applying potential energy functions to study molecular Mechanics. The atoms were represented as spheres and the bonds as springs and then mathematical functions were used to model the system. The forces during simulation were calculated from a force field that contains information about the potential energy of the involved molecules. Molecular dynamics (MD) has been merged as an efficient method for simulation of different biotechnology related phenomena in nanoscales and it accurately models fully-atomistic interactions between biomolecules and surfaces in nanoscales (Sarmadi, Shamloo and Mohseni 2017)**.**

The accuracy of its potential energy function is one of AMBER's strengths, which pertains to conformational and non-bonded energies, not high-frequency intramolecular motions (Pearlman *et al.* 1995). All-atom molecular dynamics was studied to explore the stability and conformational flexibility (global and local) of the protein and ligand systems (Musyoka *et al.* 2016). The force field parameters for all ligands studied were not available in AMBER03 force field used in GROMACS MD simulations, an AnteChamber Python Parser interface (ACPYPE) was used to parameterise the required topologies, atomic types, and charges. The generated GROMACS compatible files for the protein and ligands were then merged, solvated, minimised and equilibrated. The stability of docked complex and the binding pose obtained in docking studies are widely used to be verified by molecular dynamics simulation studies (Anusuya and Gromiha 2017). Measuring the root mean square deviation (RMSD) of each trajectory obtained in molecular dynamics simulation with respect to their position in the reference frame is used to observe the stability of the docked complex.

4.2 Methodology

Molecular dynamics simulations were done for the docked results of the three structures using GROMACS v2016.4 (Groningen Machine for Chemical Simulations) package with AMBER03 force field parameters. AMBER ACPYPE (a python interface to Antechamber that writes GROMACS topologies) was used to create the topology files of the ligands. The models were then solvated using spc216 water molecules in a triclinic box having edges at 1.5 nm from the molecular structure. Ions 1 positive (Na) and 1 negative (Cl) were added to neutralise the system.

To remove the steric conflicts between atoms of protein and water molecules having a maximum step of 5000 with steepest descent integrator, energy minimisation was performed on the solvated system as seen in Appendix 3.3. NVT ensemble (constant number of particles, volume and temperature) referred to as isothermal-isochoric, here the temperature of the system can reach a plateau at a desired value. Energy minimised model was subjected to position restrained molecular dynamics with NPT ensemble keeping the number of particles (N), system pressure (P) and temperature (T) a constant parameter, at 50 000 steps for 100ps time. The energy minimisation was performed to clear out all steric clashes prior to simulation. The final molecular dynamics simulation was performed on the CHPC cluster at 100ns. The trajectories were analysed using root mean square deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (RG). RMSD indicates the convergence of the structure towards an equilibrium state, calculated using gmx fms, RMSF records for each atom the fluctuation about its average position, giving insight into the flexibility of regions of the peptide calculated using gmx rmsf and the radius of gyration measures the shape of the molecule at each point calculated using gmx gyrate Figure 4.1 below, shows the steps that were taken to perform molecular dynamics in which GROMACS AMBER03 and ACPYPE tools were used to create the topology files for the protein and the ligands. The system was the solvated using spc216 and there was ion addition to equilibrate the system, followed by energy minimisation to reduce steric clashes and finally the production run was performed in the CHPC cluster.

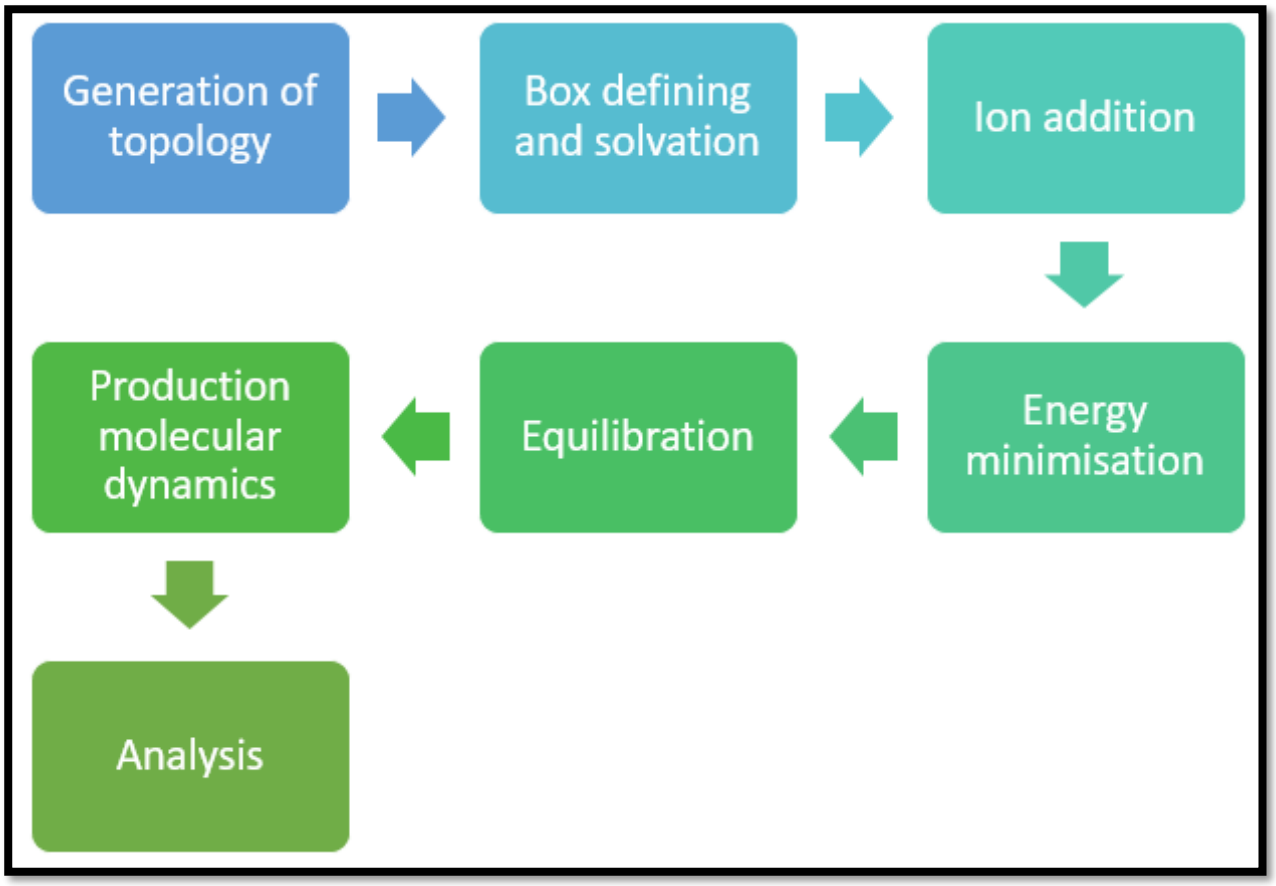


Figure 4. 1: Steps taken in molecular dynamics (Abraham *et al.* 2015)**.**

4.3 Results and discussion

The root mean square deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (RG) was calculated to analyse the molecular dynamics trajectories. RMSD was used to calculate the deviation of the structures at 100ns, RMSF was used to calculate the fluctuation of each residue and the radius of gyration was used to calculate the compactness of the structures. Figure 4.2 below, show the thermodynamic (temperature) results showing that the system reached convergence.

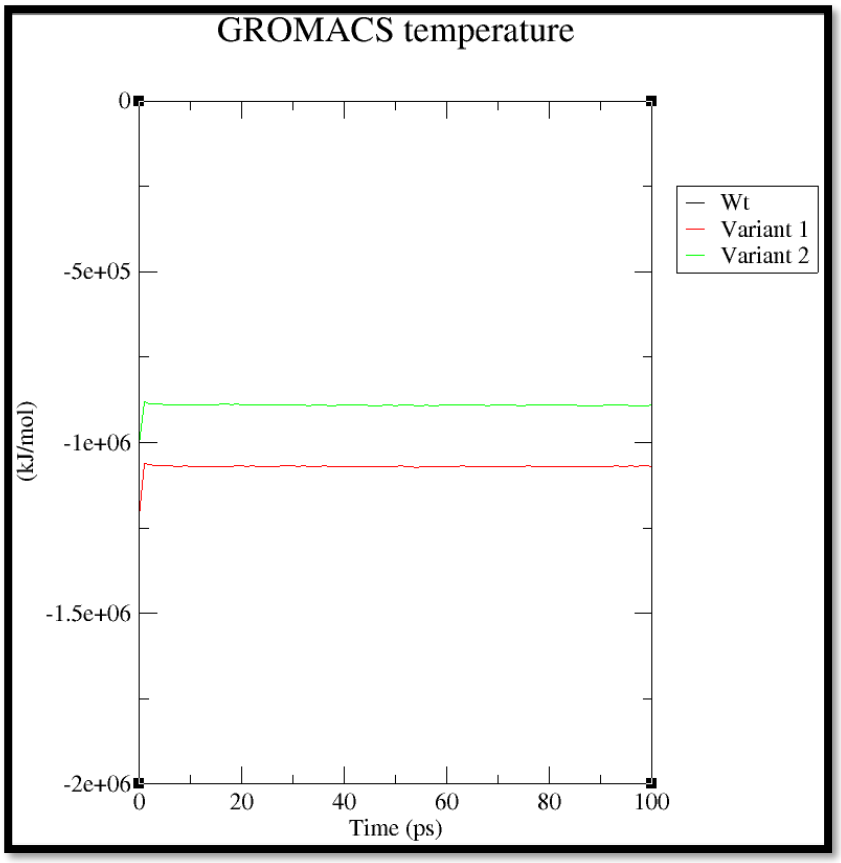


Figure 4. 2: Thermodynamics graph viewed in xmgrace.

In Figure 4.3 below, black represents wildtype apo, magenta represents variant 1 and blue represents variant 2 apo. Figure 4.3 below shows that the RMSD of the apo structures stabilised just after 20ns and the deviations of the three structures were different with variant 1 showing the highest deviation of 0.45 nm, generally the structures were quite stable. The RMSF showed a low fluctuation at residues of interest (G2019S and I2020T), however, there were peaks spotted in the variant 1 and variant 2, with variant 1 showing the highest and this revealed that indeed the variants do affect the behaviour of the (LRRK2 kinase domain) protein. In variant 2, the highest peak was observed at residue 2077 with 0.54nm and in variant 1 the highest peak was spotted at residue 2104 with 0.32nm. The radius of gyration was a straight line showing that the structures were quite compact.

4.3.1 RMSD, RMSF and radius of gyration of the three free (apo) structures

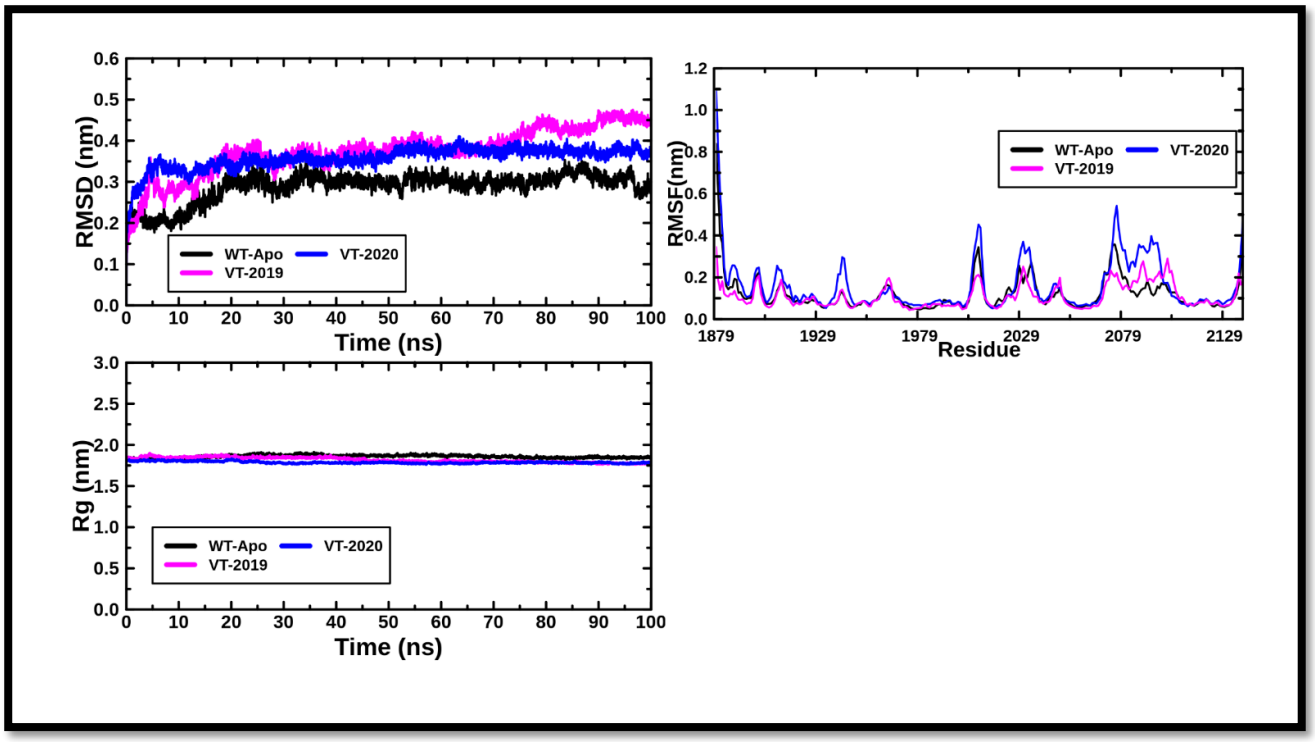


Figure 4. 3: The RMSD, RMSF and radius of gyration for the wildtype apo, variant 1 (G2019S) apo and variant 2 (I2020T) apo.

The first graph vertically is the RMSD and the second is the RMSF and the last one is the radius of gyration graph. In Figure 4.4 below, horizontal black represents the wildtype apo, green represent ligand SANC00237, blue represents ligand SANC00525 and magenta represents the ligand SANC00595. The RMSD of the wildtype showed the deviation stabilising after 15 ns, the deviations with the ligands bound to the protein were generally higher in comparison to the wildtype apo, showing that these ligands do alter the behaviour of the wildtype protein. The residues of interest (G2019S) and I2020T) were lower in both the apo and when the ligands were bound, SANC00237 had a higher fluctuation of 0.5nm at residue 2031. The radius of gyration showed compactness even when the ligands were bound to the apo. The second graph horizontally black represents variant 1, green represent SANC00101, blue represents SANC00458 and magenta represents SANC00467. In the second vertical graphs, the RMSD of the variant 1 was higher in comparison to when ligands were bound to it with an RMSD value of 0.5nm. Variant 1 also had the highest fluctuation peak at 0.59nm of residue 2078 and the residues of interest had low fluctuations. The radius of gyration showed compactness and was a straight line. The last graph black represents variant 2, green represents SANC00116, blue represents SANC00279 and magenta represents SANC00368. The last vertical graph showed an RMSD deviation stabilising after 10ns both the apo and when ligands were bound

to the apo showed similar deviations. The RMSF of SANC00116 showed higher fluctuations with the high peak of 0.5nm. The residues of interest like in other structures were low as well and the radius of gyration showed compactness with a straight line even when ligands were bound to the apo.

4.3.2 RMSD, RMSF and radius of gyration of the three structures with ligands bound

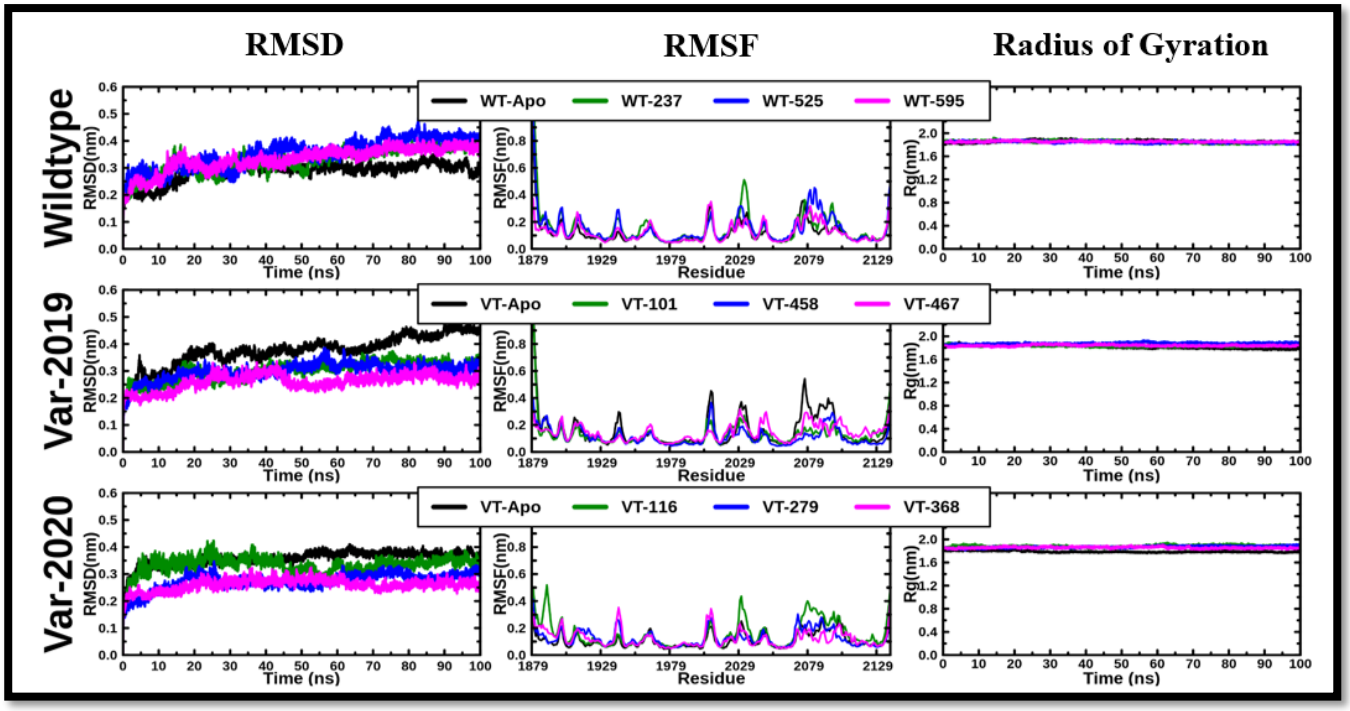


Figure 4. 4: The RMSD, RMSF and radius of gyration for the wildtype apo, variant 1 (G2019S) apo and variant 2 (I2020T) apo.

4.4 Chapter summary

Molecular dynamics helps one to understand time-dependent behaviour of the proteins, resulting into the discovery of important biological phenomena. In this study, the probable effects of the variants on LRRK2 kinase domain and the effect of ligands bound to the three structures were studied.

Generally, there was difference in the RMSD behaviour of the three structures, with the deviation being higher in variant 1 and variant 2. Variant 1 apo had a RMSD of 0.45nm and variant 2 had a RMSD of 0.3nm. The three structures showed large conformational changes after 20ns as seen in the RMSD of the C-alpha atoms in the structures increases. The fluctuations in the RMSF proved to be higher in variant 1 and variant 2 apo structures. The radius of gyration was stable in all structures. The RMSD of all three structures started increasing after 20ns showing large conformational changes. The RMSD of the wildtype with ligands was high when the ligands were bound in comparison to the wildtype apo, with SANC00525 being the highest in deviation. The residues of interest (G2019 and I2020) had a low fluctuation and SANC00525 had the highest RMSF of 0.9nm.

Variant 1 had a high deviation in comparison to when the ligands were bound to it and the RMSD started increasing after 20ns. Variant 1 apo showed the highest fluctuation in the RMSF compared to when the ligands were bound to it and the fluctuations of the residues of interest were low. Variant 2 proved to be of a higher deviation in comparison to when the ligands were bound to it and the RMSD increased after 20ns. The residues of interest had a low fluctuation and when SANC00116 was bound to variant 2 higher RMSF of the different residues in comparison to when the other ligands were bound and the apo were observed. The radius of gyration of all the structures when then ligands were bound to it proved to be stable showing compactness. The behavioural different of the three structures was studied successfully and there was a difference in the structure movements when there was a variant or when there were ligands bound to the apo proteins. No major changes were observed from Figure 4.4 above, as the simulations for RMSD and RMSF were made calculated for both the apo protein and the protein-ligand complex. The regions of high fluctuation were loop regions, which are responsible for the protein's shape, dynamics and physiochemical properties.

CHAPTER FIVE: Network analysis

5.1 Network Analysis

MD-Task analyses molecular dynamics trajectories using python scripts that fall in three categories Residue Interaction Network (RIN) analysis, Perturbation Response Scanning (PRS) and Dynamic Cross-Correlation (MD-TASK Documentation 2018). MD-Task is a technique that is used for further analysis of molecular dynamics trajectories, it includes dynamic residue network (DRN which combines the RINs of the frames in a molecular dynamic trajectory) analysis, PRS and dynamic cross-correlation (DCC) none of which are found in commonly used MD packages (Brown *et al.* 2017). Network analysis is used to study protein aspects like protein structure flexibility, folding of protein domains, recurring structural patterns, key residue fluctuation and side-chain clusters (Amitai *et al.* 2004).

Each residue in the protein is a node in a RIN, an edge between two nodes exists if the Cβ atoms (Cα for glycine) of the residues are within a user-defined cut-off distance (6.5-7.5Å) of each other, a DRN is also constructed and is used to calculate the changes in betweenness centrality (BC) and average shortest path(L) to residues over the trajectory (Brown *et al.* 2017). MD-Task analyses the MD trajectories in three ways Betweenness centrality (BC), average shortest path (L) and residue contact map. Betweenness centrality (BC) is a measure of how important a residue is for communication within a protein, the BC of a node is equal to the number of shortest paths from all nodes to all others that pass through that node. Average shortest path (L), is the sum of the shortest paths to that residue, divided by the total number of residues less one, it describes how accessible a residue is within the protein. Residue contact map yields a network diagram with the residue of interest by monitoring the interactions of a residue throughout a simulation at the centre and residues that it interacts with arranged around it and edges between the residue of interest and the other residues are weighted based on how often the interaction exists.

Mathematical branches known as graph theory are used to analyse residue interaction networks (RIN). To analyse and characterize protein structures residue interaction network based on network theory can be used. This analysis is mathematical chemistry and is rapidly moving towards biological network analysis of systems biology (Hu *et al.* 2014). For protein activity the interactions of protein residues within and between functional sites are crucial. Small-world networks are characterized by both clusters of local interactions and ‘long-range’ interactions between different clusters in which these networks include a small number of central nodes that are hubs through which many nodes can indirectly connect (Amitai *et al.* 2004). The protein backbone plays an important role in determining

the small-world properties which is of interest to study the network properties of proteins with different connectivities (Hu *et al.* 2014).

In seeking to understand whether central nodes of protein residue interaction networks corresponded to functional residues, the closeness of a residue to other residues on the network was found to characterize many functional protein sites (Amitai *et al.* 2004). The path in which the two concerned nodes are connected by the smallest number of intermediate nodes is considered the shortest path, then the path that indicates the expected distance between two nodes is known as the average shortest path length or characteristic path length and finally the maximum length of shortest paths between two nodes is called the network diameter (Hu *et al.* 2014).

Studying together the small world network properties with network parameters provides a better understanding of protein folding and stability, identifying functional residues and hotspot residues, mutations and allosteric communication analysis, detecting protein decoys and interactions with protein complexes (Amitai *et al.* 2004). Stability is the ability of the substance to remain unchanged over time under stated or reasonably expected conditions such as change in temperature. The molecular dynamics trajectories can be extremely large and MD-TASK tools only require the alpha and beta carbon atoms to be present, hence it is essential to reduce the trajectories to save space and improve performance. When reducing the trajectories, it is very important to perform the same reduction in the topology PDB file, the trajectory and topology file should have the exact same number of atoms to avoid errors.

Complex systems can be analysed as networks of interactions between the system components, analyzing the network can then characterize the whole system and its individual components (Amitai *et al.* 2004).

5.2 Methodology

Residue interaction network was constructed by considering a Cα as nodes that are connected by non-covalent interactions. The graphs are then either unweighted or weighted in which edges are defined based on predefined cut off such as the strength of the interaction or distance. Contact maps were then used to generate similar abstracted representations of protein structures. The residue interaction networks, and protein contact maps were specified by two parameters, in which each amino acid residue is represented by a vertex and edges represented contacts between atoms that had at least one Van der Waals interaction. All proteins were modelled using undirected graphs in which the amino acids are represented as nodes and their interactions as edges.

Contact map is a network with weighted edges depicting how often residues are interacting with the selected residue over the course of the simulation. The following command is used to create the contact maps for SNP analysis:

contact_map.py <options> --trajectory <trajectory> --topology <b file> (1)

A trajectory file, topology file, residue, threshold and prefix are required as input to create contact maps. A trajectory file from a molecular dynamic's simulation converted into a DCD format was used. The topology file which is a PDB reference file for the trajectory was also used. A distance threshold of 5.0 Angstroms was used. Four trajectories were given wildtype.dcd, variant_1.dcd and variant_2.dcd, to build contact maps around position Gly2019, Ser2019, Ile2020 and Thy2020. A contact map in PDF together with a CSV file containing all the calculated values were produced.

Correlation heatmap is a PNG heatmap depicting the dynamic correlation between atoms in the trajectory. Analysis of the trajectories of the system will enable the calculating of the dynamic correlation between all atoms within the molecule, the degree in which they move together. The following command is used to calculate the dynamic cross-correlation:

calc_correlation.py --step 100 --prefix example_corr --trajectory example_small.dcd -- (2)
topology example_small.pdb --lazy-load

The following files and steps are required for one to create the dynamic cross-correlation heatmaps a trajectory, topology files, step, prefix and lazy load. The trajectory and topology file are defined above, 100 steps were used to iterate through the trajectory frames which calculates the frames to be skipped. Prefix is the output name and a lazy load is used to load trajectory frames in a memory efficient manner for large trajectories.

The shortest path is identified as the path through which the two concerned nodes are connected by the smallest number of intermediate nodes. The characteristic path length L which is the average shortest path length indicates the expected distance between two connected nodes. The number of times a node is included in the shortest path between each of nodes, normalized by the total number of pairs is known as the betweenness centrality B. The following equation is used to calculate the betweenness centrality:

$$B_k = \Sigma_{s \neq n \neq t}(\sigma_{st}(k) / \sigma_{st}) \qquad (3)$$

Whereby (s) and (t) are nodes in the network other than (k), $\sigma_{st,}$ represents the number of shortest path from (s) and (t), and σ_{st} (n), the number of shortest paths from (s) and (t) on (k) lies. The betweenness centrality of a node reflects the amount of control the node exerts over the interactions of other nodes in the network. The reciprocal of the average shortest path length is the closeness centrality (C) which is calculated as follows:

$$C_k = (x\text{-}1) / \Sigma_{k \varepsilon \underline{U, k \neq m}} L(m,k) \qquad (4)$$

In which (U) is the set of all nodes and (x) is the number of nodes in the network. Closeness centrality measures how fast information spreads from a given node to the other reachable nodes in the network.

The following files and steps are required for calculating average BC and L a trajectory, topology file, ligands, threshold, step, generate plots, calculate BC, calculate L, discard graphs and lazy load. A threshold of 7.0 was used, 100 steps to iterate through the trajectory frames was used. The following command was used to calculate average BC and L:

calc_network.py --topology wt.pdb --threshold 7.0 --step 100 --generate-plots --calc- qa BC --calc-L --discard-graphs --lazy-load wt.dcd (5)

The command above will calculate the network for every 100th frame in the trajectory. The trajectory will be iterated through and the frames loaded one at a time and then discarded once the network for the frame has been calculated. The average shortest path for each residue in each frame and the betweenness centrality of each residue in each frame was calculated. For each frame analysed in BC matrices and avg_L matrices, and N x 1 matrix is produced, where N is the number of residues in the protein and each value represents the BC and L for the residue at that index. The trajectory had 260 residues; 125001 frames interval was set for the trajectories.

5.3 Results and discussions

The residue contact maps allow one to determine how often (number of times per unit time), throughout the trajectory, a residue interact with the surrounding residues. The contact map is generated at the position of the variant or SNP and is compared to the same position in the wildtype protein to determine whether the variant or SNP affect the immediate interactions at that position. The center residue is the residue of interest and the green ball structures surrounding this center structure are the residues contacting the center residue. The grey line shows how often the residues are interacting, meaning the thicker the grey line the more residues interact.

In Figure 5. 1 below, the contact map of the behaviour of Glycine of position 2019 is studied in which it was observed that the residue formed frequent interactions with Cys146, Cys147, Ala143, Ile143 and Tyr140 which formed part of the active site. However, in Figure 5.2 below, a change in behaviour is observed in which there is now less contact of the Serine 2019 with other residues in general. When there is this variation it is observed that the Ser142 which is Serine 2019 interacts more frequently with Val45 which did not even form interaction in Figure 5.1. In Figures 5.1 and 5.2, the two residues of interest interacted with ILE112, GLU42 and ALA143. The MD task package numbers residues starting from 1until the end of the sequence, the number after the 3-letter code of the amino acid represents the position of the amino acid in the sequence and it does not number the residues based on their original numbering. The contact maps in figure 5.1-5.4 shows the numbering of the residues starting from 0 up until 260, representing the residue number of the LRRK2 kinase domain starting from 1879-2138 and the residues of interest G2019S and I2020T being at position 141 and 142 respectively.

Figure 5. 1: Residue contact map for Glycine 2019.

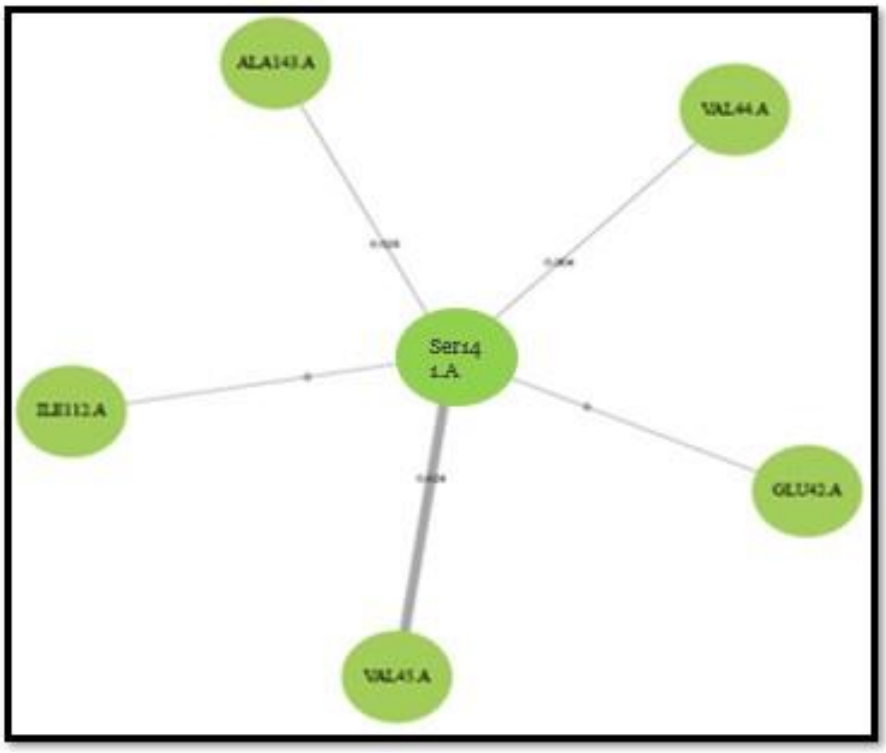


Figure 5. 2: Residue contact map for Serine 2019.

In Figure 5.3 below, Isoleucine 2020 referred to as Ile142 was the residue of interest. It formed frequent interactions with Tyr140, Ala143, Glu42, Gly141, Ile112, and Val45. However, once again it was observed that when there is variation, interaction with other residues reduces. When there is Threonine 2020 it forms frequent interactions with Glu42, Val44, Gly141, Gln41, Ala143, and Val45. In Figure 5.3 and Figure 5.4, the residues of interest interacted with TYR140, GLU42, CYS147, CYS146, ALA143, VAL45, GLN41 and LEU46. In both G2019S and I2020T these similar interactions were observed ALA43 and GLU42. Variant 1 and variant 2 indeed affect the dynamics of this protein as observed from the contact maps,

Figure 5. 3: Residue contact map for Isoleucine 2020.

Figure 5. 4: Residue contact map for Threonine 2020.

The dynamic cross-correlation tool produces an N X N heatmap, where N is the number of (alpha carbon) atoms in the system and each element corresponds to the dynamic cross-correlation between each i, j atom, the i and j represent each atom in the sequence. The correlations are calculated between -1 and 1, where 1 is the complete correlation, -1 is complete anti-correlation and 0 is no correlation. It was observed in Figures 5.5 to 5.7, that there was more correlation when there was variation in comparison to the wildtype. This finding suggests that the variants have the potential to affect the function of the protein. The dotted lines in Figures 5.5 to 5.7 represents the regions of interest 140 and 141. The x-axis and y-axis of the cross-correlation graph starts from 0 and end at 260 representing the residue number of the LRRK2 kinase domain starting from 1879-2138 and the residues of interest G2019S and I2020T being at position 140 and 141 respectively.

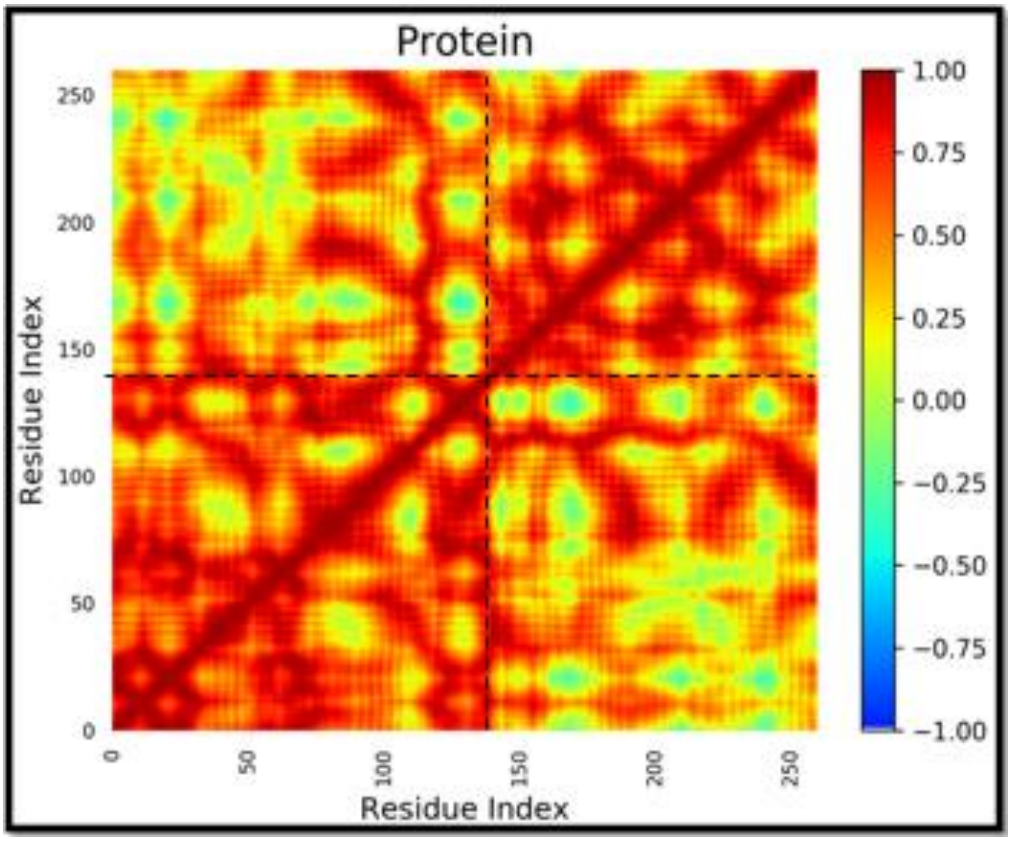


Figure 5. 5: Wildtype dynamic cross correlation.

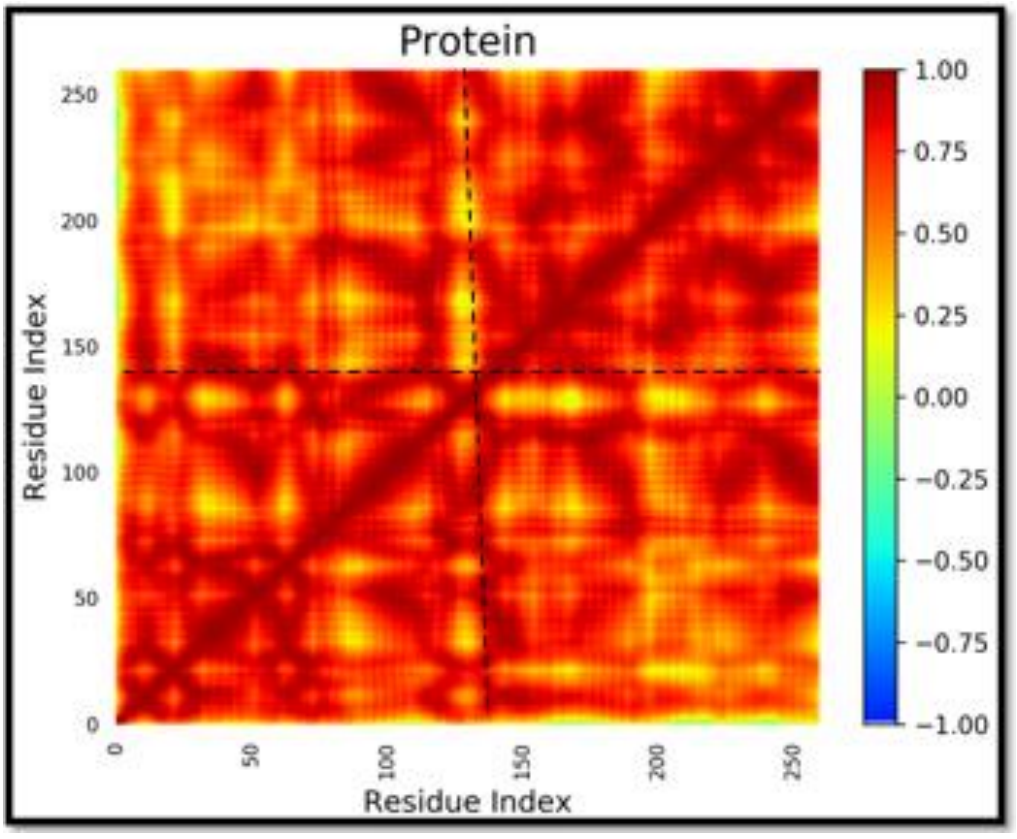


Figure 5. 6: Variant 1 G2019S dynamic cross correlation.

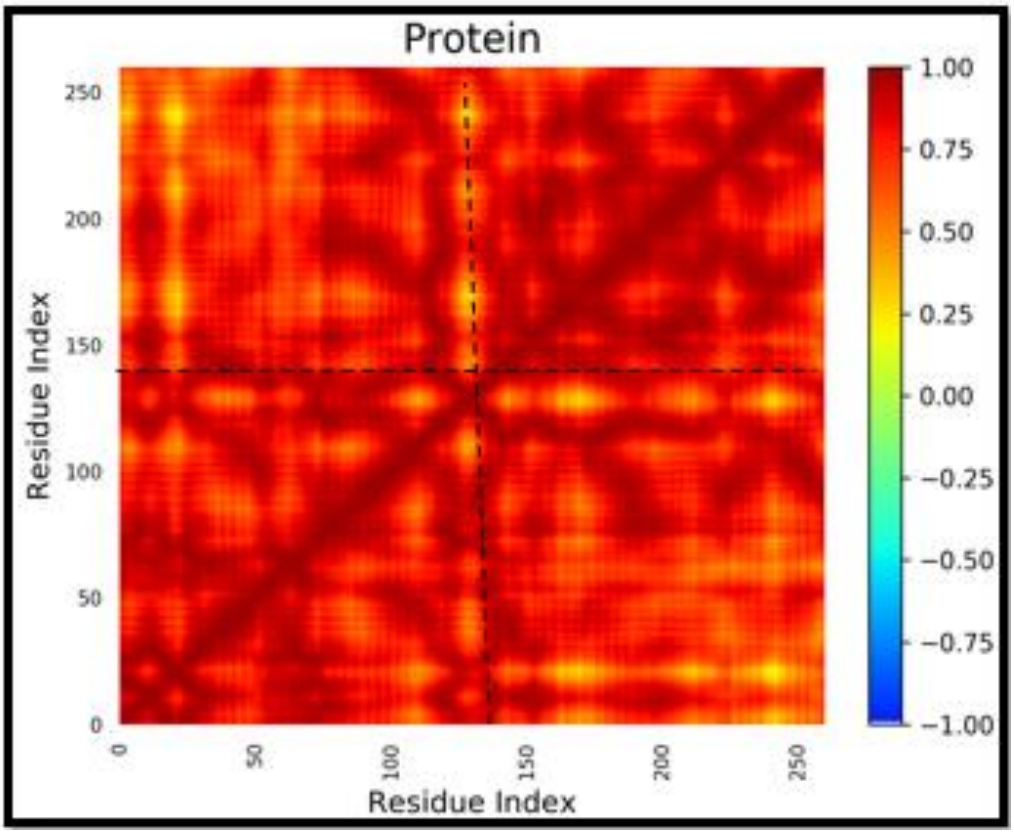


Figure 5. 7: Variant 2 I2020T dynamic cross correlation.

The x-axis of the average BC and average L, graph starts from 0 and end at 260 representing the residue number of the LRRK2 kinase domain starting from 1879-2138 and the residues of interest G2019S and I2020T being at position 140 and 141 respectively. In Figures 5.8, 5.9 and 5.10 below, the values from different trajectories are plotted on the same set of axes. In Figures 5.8, 5.9 and 5.10 below, residue 1930 has the high BC implying that it might be important for controlling inter-domain communication in the protein. It was also observed that the residues of interest Gly2019Ser and Ile2020Thr at positions 140 and 141 respectively of the x-axis of the average BC graphs have an average BC which is also quite high implying that they too maybe important for inter-domain communication in the protein. The broken line on the graph in figures 5.8-5.13 marks the residues of interest. The residues of interest in figures 5.8-5.10, shows high average BC, indicating that the residues are important for communication within the LRRK2 protein.

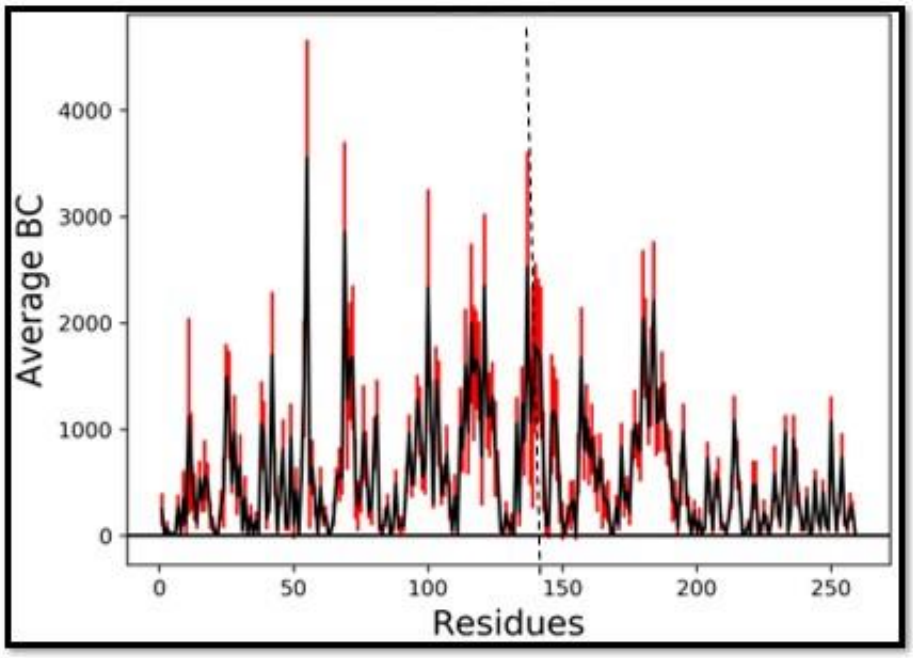


Figure 5. 8: Average BC for the wildtype structure.

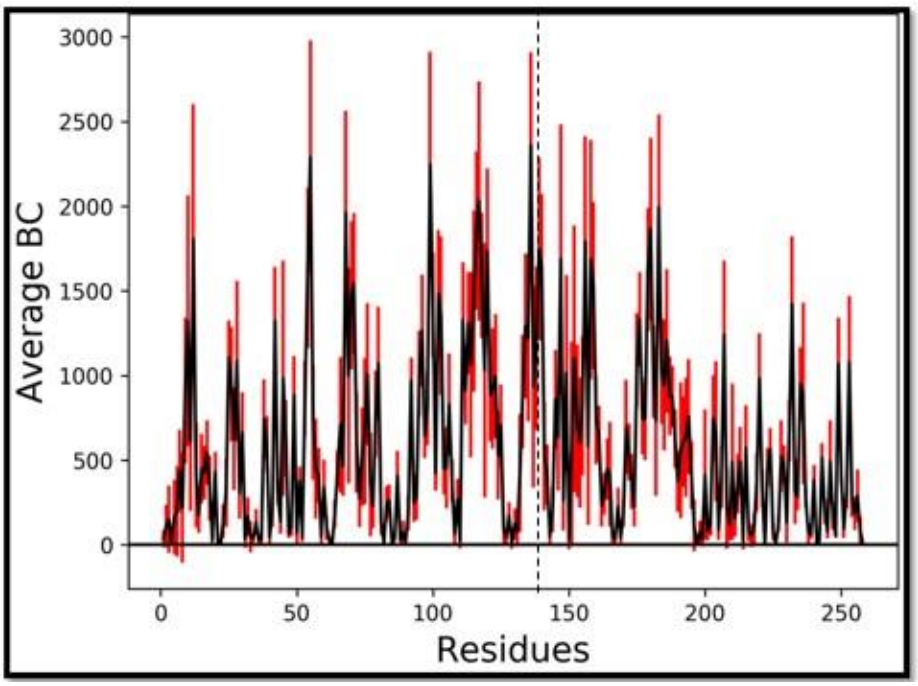


Figure 5. 9: Average BC for the variant 1 (G2019S) structure.

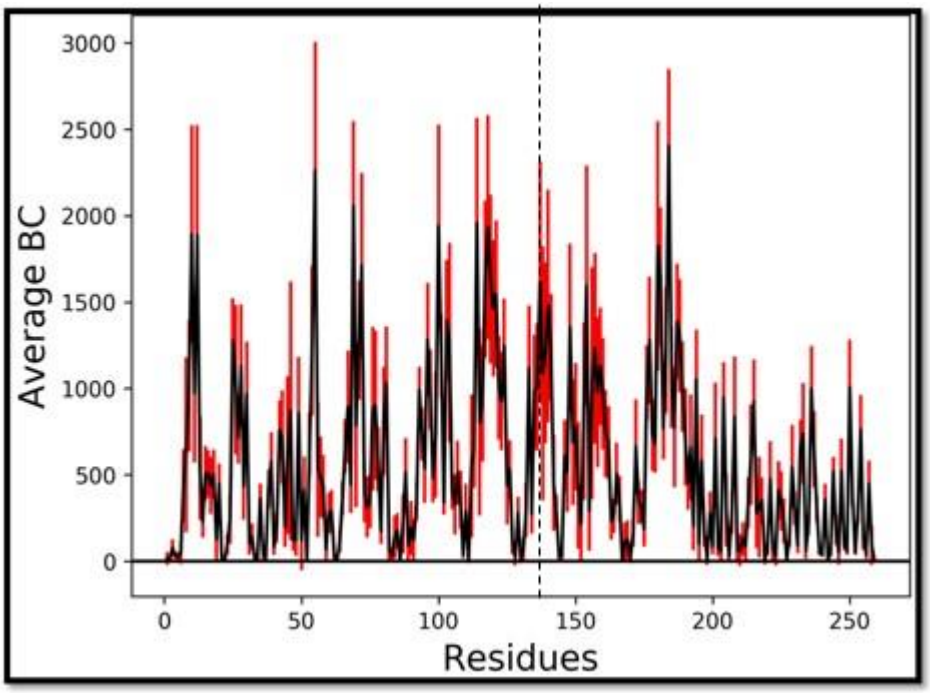


Figure 5. 10: Average BC for the variant 2 (I2020T) structure.

Figures 5.11 to 5.13 below, represents the average L of the three structures showed the accessibility of the residue of interest within the protein. Variant 1 and variant 2 cause a change in the average L of several residues in the protein including residues of interest at positions 140 and d141, indicating that the variants might have an important effect on the protein function. Residues which had a high delta L value can steer the conformational changes of the protein (Amusengeri and Tastan Bishop 2019).

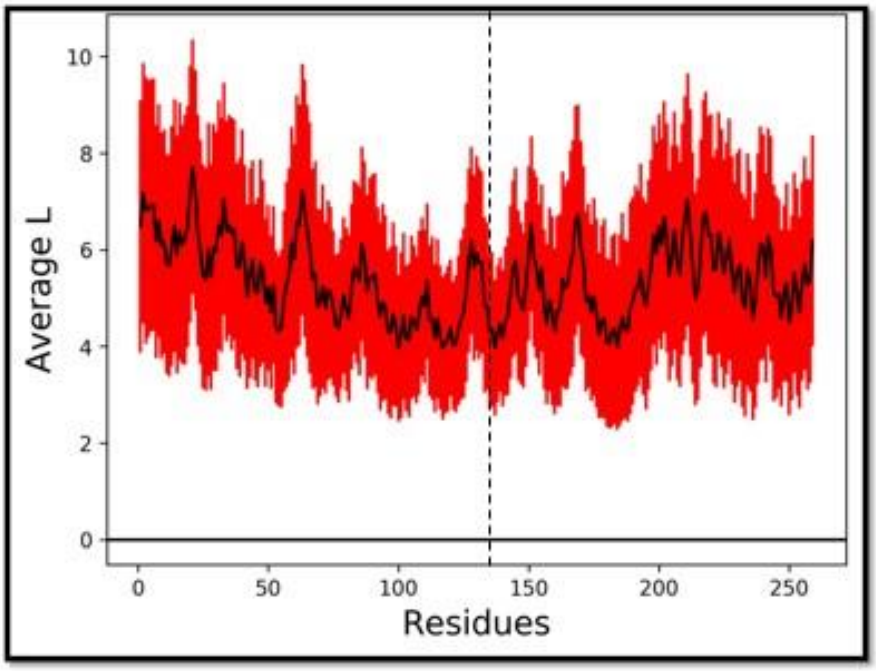


Figure 5. 11: Average L for the wildtype structure.

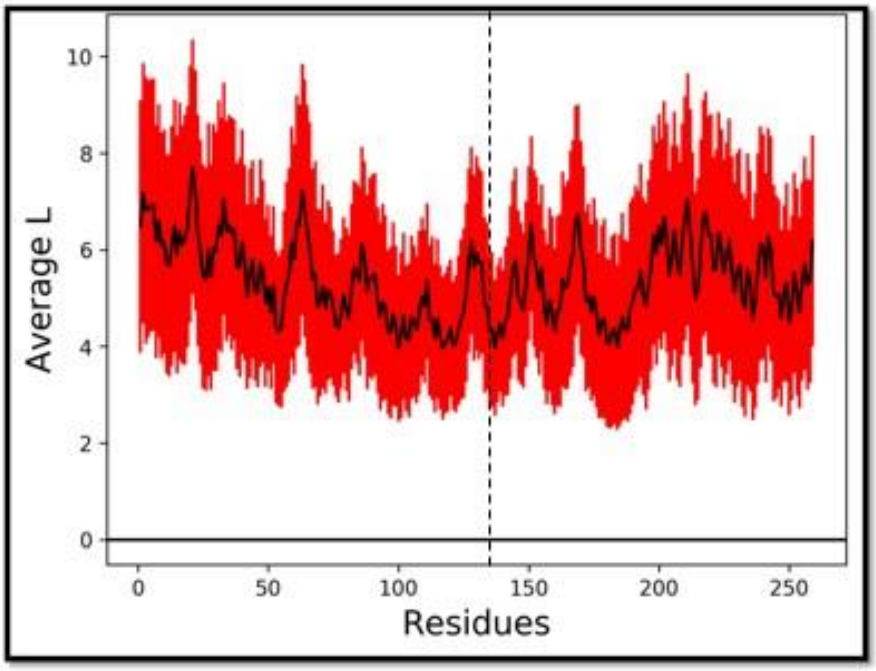


Figure 5. 12: Average L for variant 1 (G2029S).

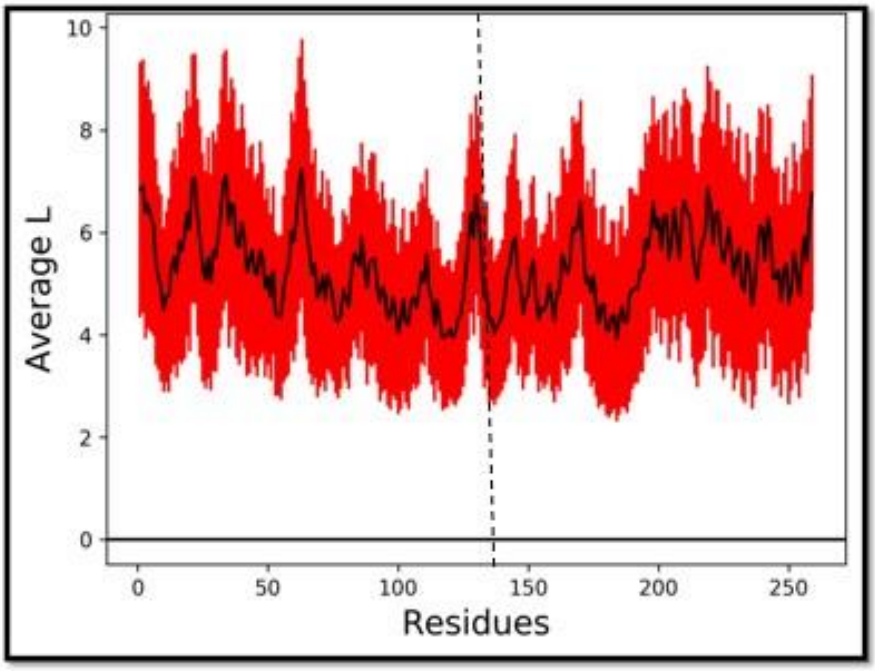


Figure 5. 13: Average L for variant 2 (I2020T).

5.4 Chapter summary

Molecular dynamics trajectories were reduced to save space and improve performance for network analysis. This was achieved by course grading the system by retaining only the alpha and beta carbon atoms of the protein. MD-Task only requires the alpha and beta carbon atoms to be present in order for the tool to perform all necessary calculations (MD-TASK Documentation 2018). Network analysis was performed in RUBi cluster and R-script was used to create the graphs.

The contact maps Figures 5.1 until 5.4 showed how the residue of interest interacted with the other residues of the protein and Figures 5.2 and 5.4 show loss of residue interaction whenever there was a variation. Figures 5.1 and 5.2, showed these common residue interactions ILE112, GLU42 and ALA143. Figures 5.3 and 5.4, showed the following common interactions TYR140, GLU42, CYS147, CYS146, ALA143, VAL45, GLN41, GLY141 AND LEU46. Residues ALAL143 and GLU42 were the common residues to form interaction when all four (GLY140. SER140, ILE141 and THR141) residues of interest were being analysed. Figures 5.5 until 5.7 showed complete correlation between the residues in the variants in comparison to the wildtype and the residues of interest that were highlighted by dotted lines showed complete correlation in all three structures.

The betweenness centrality (average BC), was not clear as to whether the residues of interest (G2019S and I2020T) were important for controlling communication within the protein. Residue 1930 at position 50 in Figures 5.8 until 5.10, had a high BC revealing that it might be important for controlling inter-domain communication in a protein. The average shortest path (L) as seen in Figures 5.11 until 5.13 proved that the variants had an important effect on the protein function, as there was a difference in the pattern of the three graphs. Residues at position 60 (1939) in Figures 5.11 and 5.13 proved to be the highest and residue 200 (2079) in Figure 5.12 proved to be the highest. Figures 5.11 until 5.13, showed a low average L value for the residues of interest 140 and 141 showing a high availability of the residue for signal transduction. Average L shows the mean topological spread of all residues from the residue of interest by considering the shortest path to every other residue in the network. The residues behaved differently in both average BC and average L including the residues of interest. It is important to compare the changes in average BC and L between the simulation of the wildtype and the variants because they can provide interesting information about the difference in intra-protein communication that affects the functions of the protein (MD-TASK Documentation 2018). Network analysis was performed successfully.

CHAPTER SIX: Findings and future prospects

6.1 Project conclusions and future prospect

Parkinson's disease remains the second most dangerous neurodegenerative disease in the world, hence it was of interest to study. Leucine-rich repeat Kinase 2 is a large protein of 2527 amino acid and there is no crystal structure available in any database so far. There was a need to build a model of the kinase domain of the protein since it consists of the active site and the variations of interest are found on this domain.

SANCDB natural compounds were used for blind docking on the three structures. It was confirmed that the Kinase domain of LRRK2 protein can be selectively targeted using inhibitor compounds. The Lipinski rule of five showed that the compounds had satisfactory selectivity. Only three best ligands were chosen for each structure and they formed hydrogen bonds with the active site and residues of the conserved loop. Molecular dynamics was performed to observe the behavior of the three structures with and without ligands bound and to study the behaviour of each residue especially the residues of interest. The compactness of the structures was also studied. To further study the behavior of these structures network analysis was performed in which the interaction of residues of interest was observed and there was loss of interaction when there was a variation. The dynamic cross-correlation showed complete correlation in the variants, in comparison to the wildtype. The residues important for controlling inter-domain communication in a protein were studied through calculating BC and average BC. The shortest path (L) showed that the variants might cause conformation changes in the protein.

The study was performed successfully, and the necessary steps and stages were followed. For further work re-docking and re-analysis of the results is recommended. Modeling of the entire LRRK2 protein or creation of the crystal structure will help in furthering the study of this protein and understand its properties and mechanism. Further wet laboratory tests might be necessary to add and prove current work.

REFERENCES

Abraham MJ, Murtola T, Schulz R *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 2015;**1**–**2**:19–25.

Agalliu I, San Luciano M, MirelmanMD A *et al.* Higher frequency of certain cancers in LRRK2 G2019S mutation carriers with Parkinson disease a pooled analysis. *JAMA Neurol* 2015;**72**:58–65.

Alessi BDR, Sammler E. LRRK2 kinase in Parkinson's disease. :1–3.

Allen MP. Introduction to Molecular Dynamics Simulation. 2004;**23**.

Alzohairy AM. همدقتملا هيويحلا هيتامولعملا) 8 (هددعتملا تاعباتتلا نزاوت ءانب) 3 (ويف لاچلا مادختسأب نيولتلاو ليدعتلا Building a Multiple Sequence Alignment (8) Jalview Basics in More details بريهزلا روصنم دمحأ / د. 2014, DOI: 10.13140/2.1.3365.8561.

Amitai G, Shemesh A, Sitbon E *et al.* Network Analysis of Protein Structures Identifies Functional Residues. 2004:1135–46.

Amusengeri A, Tastan Bishop Ö. Discorhabdin N, a South African Natural Compound, for Hsp72 and Hsc70 Allosteric Modulation: Combined Study of Molecular Modeling and Dynamic Residue Network Analysis. *Molecules* 2019;**24**:188.

Anusuya S, Gromiha MM. Quercetin derivatives as non-nucleoside inhibitors for dengue polymerase : molecular docking , molecular dynamics simulation , and binding free energy calculation. *J Biomol Struct Dyn* 2017;**1102**:1–15.

Ayano G. Dopamine: Receptors, Functions, Synthesis, Pathways, Locations and Mental Disorders: Review of Literatures. *J Ment Disord Treat* 2016;**2**:2–5.

Bae E-J, Kim D-K, Kim C *et al.* LRRK2 kinase regulates α-synuclein propagation via RAB35 phosphorylation. *Nat Commun* 2018;**9**:3465.

Bai X chen, McMullan G, Scheres SHW. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 2015;**40**:49–57.

Benkert P, Ku M, Schwede T. QMEAN server for protein model quality estimation. 2009;**37**:510–4.

Benkert P, Tosatto SCE, Schomburg D. for model quality assessment. 2008:261–77.

Berardelli A, Wenning GK, Antonini A *et al.* EFNS/MDS-ES recommendations for the diagnosis of Parkinson's disease. *Eur J Neurol* 2013;**20**:16–34.

Berendsen HJC. Molecular Dynamics Simulations : The Limits and Beyond *. 1999.

Berwick DC, Harvey K. LRRK2: an éminence grise of Wnt-mediated neurogenesis? *Front Cell Neurosci* 2013;**7**, DOI: 10.3389/fncel.2013.00082.

Bhayye SS, Roy K, Saha A. Molecular dynamics simulation study reveals polar nature of pathogenic mutations responsible for stabilizing active conformation of kinase domain in leucine-rich repeat kinase II. 2018:657–66.

Brown DK, Penkler DL, Amamuddy OS *et al.* Structural bioinformatics MD-TASK : a software suite for analyzing molecular dynamics trajectories. 2017, DOI: 10.1093/bioinformatics/btx349.

Brown DK, Tastan Bishop Ö. Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis: Analyzing Variation at the Protein Level. *Glob Heart* 2017;**12**:151–61.

Bursulaya BD, Totrov M, Abagyan R *et al.* Comparative study of several algorithms for flexible ligand docking. 2004:755–63.

Cho HJ, Xie C, Cai H *et al.* Parkinson's disease associated LRRK2 hyperactive kinase mutant disrupts synaptic vesicle trafficking in ventral midbrain neurons. *Eneuro* 2017;**12**:0964-17.

Cho HJ, Xie C, Cai H. AGE-induced neuronal cell death is enhanced in G2019S LRRK2 mutation with increased RAGE expression. *Transl Neurodegener* 2018;**7**:1–8.

Chou JJ, Sounier R. Electron Crystallography of Soluble and Membrane Proteins. 2013;**955**, DOI: 10.1007/978-1-62703-176-9.

Cme CS, Bmi B, Dror R. X- -ray crystallography. 2015:1–27.

Cookson MR. Structure, function, and leucine-rich repeat kinase 2: On the importance of reproducibility in understanding Parkinson's disease. *Proc Natl Acad Sci* 2016;**113**:8346–8.

Cornejo-Olivas M, Torres L, Velit-Salazar MR *et al.* Variable frequency of LRRK2 variants in the Latin American research consortium on the genetics of Parkinson's disease (LARGE-PD), a case of ancestry. *npj Park Dis* 2017;**3**:19.

Daugelaite J, O' Driscoll A, Sleator RD. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. *ISRN Biomath* 2013;**2013**:1–14.

Dorval V, Hébert SS. LRRK2 in transcription and translation regulation: Relevance for Parkinson's disease. *Front Neurol* 2012;**FEB**:2–7.

Edwards JC. Principles of NMR. *TrAC Trends Anal Chem* 1992;**11**:XVIII.

Eyers PA. Back to the future: new target-validated Rab antibodies for evaluating LRRK2 signalling in cell biology and Parkinson's disease. *Biochem J* 2018;**475**:185–9.

Feenstra KA. Structural Bioinformatics. 2017.

Fernandez-Fuentes N, Rai BK, Madrid-Aliste CJ *et al.* Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics* 2007;**23**:2558–65.

Ferreira LG, Santos RN, Oliva G *et al. Molecular Docking and Structure-Based Drug Design*

Strategies., 2015.

Fiser A. NIH Public Access. 2014:1–20.

Di Fonzo A, Tassorelli C, De Mari M *et al.* Comprehensive analysis of the LRRK2 gene in sixty families with Parkinson's disease. *Eur J Hum Genet* 2006;**14**:322–31.

Fox MJ, Reno GJ. Parkinson ' s Disease and Environmental Factors. 2014.

Gao Y, Wilson GR, Stephenson SEM *et al.* The emerging role of Rab GTPases in the pathogenesis of Parkinson's disease. *Mov Disord* 2018;**33**:196–207.

García S, López-Hernández LB, Suarez-Cuenca JA *et al.* Original article Low prevalence of most frequent pathogenic variants of six PARK genes in sporadic Parkinson's disease. *Folia Neuropathol* 2014;**1**:22–9.

Gilsbach BK, Kortholt A. Structural biology of the LRRK2 GTPase and kinase domains: implications for regulation. *Front Mol Neurosci* 2014;**7**:1–9.

Gloeckner CJ, Kinkl N, Schumacher A *et al.* The Parkinson disease causing LRRK2 mutation I2020T is associated with increased kinase activity. *Hum Mol Genet* 2006;**15**:223–32.

Goetz CG. The History of Parkinson's Disease: Early Clinical Descriptions and Neurological Therapies. *Cold Spring Harb Perspect Med* 2011;**1**:1–16.

Goldwurm S, Di Fonzo A, Simons EJ *et al.* The G6055A (G2019S) mutation in LRRK2 is frequent in both early and late onset Parkinson's disease and originates from a common ancestor. *J Med Genet* 2005;**42**:1–8.

Grosset DG, Macphee GJA, Nairn M. Diagnosis and pharmacological management of Parkinson's disease: Summary of SIGN guidelines. *BMJ* 2010;**340**:206.

Guaitoli G, Gilsbach BK, Raimondi F *et al.* First model of dimeric LRRK2: the challenge of unrevealing the structure of a multidomain Parkinson's-associated protein. *Biochem Soc Trans* 2016a;**44**:1635–41.

Guaitoli G, Raimondi F, Gilsbach BK *et al.* Structural model of the dimeric Parkinson's protein LRRK2 reveals a compact architecture involving distant interdomain contacts. *Proc Natl Acad Sci* 2016b;**113**:E4357–66.

Guo L, Gandhi PN, Wang W *et al.* The Parkinson's disease-associated protein, leucine-rich repeat kinase 2 (LRRK2), is an authentic GTPase thatstimulates kinase activity. *Exp Cell Res* 2007;**313**:3658–70.

Hatherley R, Brown DK, Musyoka TM *et al.* SANCDB: A South African natural compound database. *J Cheminform* 2015;**7**, DOI: 10.1186/s13321-015-0080-8.

Healy DG, Falchi M, O'Sullivan SS *et al.* Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol* 2008;**7**:583–90.

Ho DH, Kim H, Kim J *et al.* Leucine-Rich Repeat Kinase 2 (LRRK2) phosphorylates p53 and

induces p21WAF1/CIP1expression. *Mol Brain* 2015;**8**, DOI: 10.1186/s13041-015-0145-7.

Hu G, Yan W, Zhou J *et al.* Residue interaction network analysis of Dronpa and a DNA clamp. *J Theor Biol* 2014;**348**:55–64.

Huang B, Clark G, Navarro-Moratalla E *et al.* Layer-dependent ferromagnetism in a van der Waals crystal down to the monolayer limit. *Nature* 2017;**546**:270–3.

Hughes KC, Gao X, O'Reilly EJ *et al.* Genetic variants related to urate and risk of Parkinson's disease. *Park Relat Disord* 2018:0–1.

Illinois U. Introduction to Molecular Dynamics VMD : Visual Molecular Dynamics.

Irwin JJ, Shoichet BK. ZINC--A Free Databse of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model* 2005;**45**:177–82.

Islam MS, Nolte H, Jacob W *et al.* Human R1441C LRRK2 regulates the synaptic vesicle proteome and phosphoproteome in a Drosophila model of Parkinson's disease. *Hum Mol Genet* 2016;**25**:5365–82.

Ito G, Okai T, Fujino G *et al.* GTP binding is essential to the protein kinase activity of LRRK2, a causative gene product for familial Parkinson's disease. *Biochemistry* 2007;**46**:1380–8.

Jaiteh M, Zeifman A, Saarinen M *et al.* Docking screens for dual inhibitors of disparate drug targets for Parkinson's disease. *J Med Chem* 2018, DOI: 10.1021/acs.jmedchem.8b00204.

James M, Murtola T, Schulz R *et al.* ScienceDirect GROMACS : High performance molecular simulations through multi-level parallelism from laptops to supercomputers. 2015;**2**:19–25.

Jankovic J. Parkinson's disease: Clinical features and diagnosis. *J Neurol Neurosurg Psychiatry* 2008;**79**:368–76.

Jonathan A, Meier O. Probabilistic Protein Homology Modeling. 2014.

Kestenbaum M, Alcalay RN. Leucine-Rich Repeat Kinase 2 (LRRK2). 2017;**14**, DOI: 10.1007/978-3-319-49969-7.

Keyser RJ. Investigation of the genetic aetiology of Parkinson ' s disease in South Africa. *Present Dr* 2011.

Kluss JH, Conti MM, Kaganovich A *et al.* Detection of endogenous S1292 LRRK2 autophosphorylation in mouse tissue as a readout for kinase activity. *NPJ Park Dis* 2018;**4**:13.

Krieger E, Nabuurs SB, Vriend G. Chapter 25: homology modeling. *Struct Bioinforma* 2003;**44**:507–21.

Kumar A, Priya G, Doss C *et al.* Molecular insights of the G2019S substitution in LRRK2 kinase domain associated with Parkinson ' s disease : A molecular dynamics simulation approach. *J Theor Biol* 2019;**469**:163–71.

Landoulsi Z, Benromdhan S, Ben Djebara M *et al.* Using KASP technique to screen LRRK2 G2019S mutation in a large Tunisian cohort. *BMC Med Genet* 2017;**18**:1–6.

Laskowski RA, MacArthur MW, Moss DS *et al.* PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993;**26**:283–91.

Lee HS, Zhang Y. BSP-SLIM: A blind low-resolution ligand-protein docking approach using predicted protein structures. 2012:93–110.

Levine CB, Fahrbach KR, Siderowf AD *et al.* Diagnosis and Treatment of Parkinson's Disease : A Systematic Review of the Literature. *Ahrq* 2003;**03**:306.

Lingor P, Liman J, Kallenberg K *et al.* Diagnosis and Differential Diagnosis of Parkinson ' s Disease. *Intech* 1999:1–21.

Liu X wu, Li W, Han T *et al.* The finding of a new heterozygous mutation site of the SCN2A gene in a monozygotic twin family carrying and exhibiting genetic epilepsy with febrile seizures plus (GEFS+) using targeted next-generation sequencing. *Clin Neurol Neurosurg* 2018;**169**:86–91.

Liu Z, Galemmo RA, Fraser KB *et al.* Unique functional and structural properties of the LRRK2 protein ATP-binding pocket. *J Biol Chem* 2014;**289**:32937–51.

Madhavi Sastry G, Adzhigirey M, Day T *et al.* Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* 2013;**27**:221–34.

Masso M. Protein Structure Analysis Secondary Structure: Computational Problems Secondary structure characterization Secondary structure assignment Protein structure classification Secondary structure prediction.

Mata IF, Wedemeyer WJ, Farrer MJ *et al.* LRRK2 in Parkinson's disease: protein domains and functional insights. *Trends Neurosci* 2006;**29**:286–93.

Mattea S, Baptista M, Reichert P *et al.* Crystallizing the {Parkinson}'s {Disease} {Protein} {LRRK}2 {Under} {Microgravity} {Conditions}. 2018;**epub**:11 PP.

MD-TASK Documentation. 2018.

Melrose H. Update on the functional biology of Lrrk2. *Future Neurol* 2008;**3**:669–81.

Meng X-Y, Zhang H-X, Mezei M *et al.* Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des* 2011;**7**:146–57.

Merrill PB, Madix RJ. O-H and C-H Bond Activation in Cyclohexanol by Atomic Oxygen on Ag(110): Alkoxide Formation and Selective Dehydrogenation to Cyclohexanone. *Langmuir* 1991;**7**:3034–40.

Mir R, Tonelli F, Lis P *et al.* The Parkinson's disease VPS35[D620N] mutation enhances LRRK2-mediated Rab protein phosphorylation in mouse and human. *Biochem J* 2018;**475**:1861–83.

Musyoka TM, Kanzi AM, Lobb KA *et al.* Structure Based Docking and Molecular Dynamic Studies of Plasmodial Cysteine Proteases against a South African Natural Compound and its Analogs.

Nat Publ Gr 2016, DOI: 10.1038/srep23690.

Ng ASL, Ng EYL, Tan YJ *et al.* Case-control analysis of LRRK2 protective variants in Essential Tremor. *Sci Rep* 2018;**8**:8–11.

Nikolaev DM, Shtyrov AA, Panov MS *et al.* A Comparative Study of Modern Homology Modeling Algorithms for Rhodopsin Structure Prediction. *ACS Omega* 2018;**3**:7555–66.

Ossowska KK. Basic Principles of Molecular Dynamics (MD) Theory.

Pawlowski M, Gajda MJ, Matlak R *et al.* MetaMQAP : A meta-server for the quality assessment of protein models. 2008;**20**:1–20.

Pearlman DA, Case DA, Caldwell JW *et al.* AMBER , a package of computer programs for applying molecular mechanics , normal mode analysis , molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. 1995;**91**:1–41.

Pedersen AG. Pairwise Alignment and Database Searching.

Phu A, Nguyen T, Moore DJ *et al.* Leucine-Rich Repeat Kinase 2 (LRRK2). 2017;**14**:1–17.

Price A, Manzoni C, Cookson MR *et al.* The LRRK2 signalling system. *Cell Tissue Res* 2018:1–12.

Prieto-Martínez FD, Arciniega M, Medina-Franco JL. Acoplamiento Molecular: Avances Recientes y Retos. *TIP Rev Espec en Ciencias Químico-Biológicas* 2018;**21**:0–23.

Pronk S, Apostolov R, Shirts MR *et al.* Structural bioinformatics molecular simulation toolkit ´ rd Pa. 2013;**29**:845–54.

Purlyte E, Dhekne HS, Sarhan AR *et al.* Rab29 activation of the Parkinson's disease-associated LRRK2 kinase. *EMBO J* 2017a;**37**:e201798099.

Purlyte E, Dhekne HS, Sarhan AR *et al.* Rab29 activation of the Parkinson's disease-associated LRRK2 kinase. *EMBO J* 2017b;**37**:e201798099.

Rana AQ, Ahmed US, Chaudry ZM *et al.* Parkinson's disease: A review of non-motor symptoms. *Expert Rev Neurother* 2015;**15**:549–62.

Rassu M, Del Giudice MG, Sanna S *et al.* Role of LRRK2 in the regulation of dopamine receptor trafficking. *PLoS One* 2017;**12**:1–22.

Rewar S. A systematic review on Parkinson ' s disease (PD). 2015;**5674**:176–85.

Rose PW, Beran B, Bi C *et al.* The RCSB Protein Data Bank: Redesigned web site and web services. *Nucleic Acids Res* 2011;**39**:392–401.

Russo I, Benedetto G Di, Kaganovich A *et al.* Leucine-rich repeat kinase 2 controls protein kinase A activation state through phosphodiesterase 4. 2018;**8**:1–11.

Sali A. MODELLER: A Program for Protein Structure Modeling Release 9.12, r9480. *Rockefeller Univ* 2013.

Sarmadi M, Shamloo A, Mohseni M. Utilization of Molecular Dynamics Simulation Coupled with Experimental Assays to Optimize Biocompatibility of an Electrospun PCL / PVA Scaffold.

2017:1–18.

Schapansky J, Khasnavis S, DeAndrade MP *et al.* Familial knockin mutation of LRRK2 causes lysosomal dysfunction and accumulation of endogenous insoluble α-synuclein in neurons. *Neurobiol Dis* 2018;**111**:26–35.

Schwede T, Kopp J, Guex N *et al.* SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003;**31**:3381–5.

Scotti L, Júnior FJBM, Ishiki H *et al.* Docking Studies for Multi-Target Drugs Docking Studies for Multi-Target Drugs. 2015, DOI: 10.2174/13894501166661508251l.

Sharma S, Ciufo S, Starchenko E *et al.* The NCBI BioCollections Database. *Database (Oxford)* 2018;**2018**:1–8.

Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. 2006:2507–24.

Steger M, Diez F, Dhekne HS *et al.* Systematic proteomic analysis of LRRK2-mediated rab GTPase phosphorylation establishes a connection to ciliogenesis. *Elife* 2017;**6**:1–22.

Steger M, Tonelli F, Ito G *et al.* Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *Elife* 2016;**5**:1–28.

Tse A, Verkhivker GM. Molecular Dynamics Simulations and Structural Network Analysis of c-Abl and c-Src Kinase Core Proteins: Capturing Allosteric Mechanisms and Communication Pathways from Residue Centrality. *J Chem Inf Model* 2015;**55**:1645–62.

Vyas M V, Garg AX, Iansavichus A V. Shift work and vascular events : systematic review and. 2012;**4800**:1–11.

Weinhold F, Klein RA. Improved general understanding of the hydrogen-bonding phenomena: A reply. *Angew Chemie - Int Ed* 2015;**54**:2600–2.

West AB, Moore DJ, Choi C *et al.* Parkinson's disease-associated mutations in LRRK2 link enhanced GTP-binding and kinase activities to neuronal toxicity. *Hum Mol Genet* 2007;**16**:223–32.

Wiederstein M, Sippl MJ. ProSA-web : interactive web service for the recognition of errors in three-dimensional structures of proteins. 2007a;**35**:407–10.

Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 2007b;**35**:407–10.

De Wit T, Baekelandt V, Lobbestael E. LRRK2 Phosphorylation: Behind the Scenes. *Neurosci* 2018:107385841875630.

Wu G, Robertson DH, Iii CLB *et al.* Detailed Analysis of Grid-Based Molecular Docking : A Case Study of CDOCKER — A CHARMm-Based MD Docking Algorithm. 2003;**37554**.

Xiong J. *Essential Bioinformatics*., 2006.

Xiong Y, Neifert S, Karuppagounder SS *et al.* Overexpression of Parkinson's Disease-Associated Mutation LRRK2 G2019S in Mouse Forebrain Induces Behavioral Deficits and α-Synuclein Pathology. *Eneuro* 2017;**4**:ENEURO.0004-17.2017.

Yang JM, Tung CH. Protein structure database search and evolutionary classification. *Nucleic Acids Res* 2006;**34**:3646–59.

APPENDICES

Appendix 1: The MODELLER python script

```
# Comparative modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ() # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']
env.io.hetatm = True
allow_alternates=True

a = automodel(env,

    alnfile = 'mutant_2019.pir',       # alignment filename
    knowns = ('1fvr', '4uy9'),            # codes of the templates
    sequence ='2019')     # code of the target


a.starting_model= 1      # index of the first model
a.ending_model = 100      # index of the last model
                         # (determines how many models to calculate)
a.md_level = refine.very_slow
a.make()                 # do the actual comparative modeling

```

Appendix 2. 1: Target- template input .PIR alignment wildtype file

```
>P1;Q5S
sequence:Q5S::::::::
QAPEFLLGDGSFGSVYRAAYEGEEVAVKIFNKH------TSLRLLRQELVVLCHL-HHPSLISLLAAGIR--PRMLVMEL
ASKGSLDRLLQQDKA---------------SLTRTLQHRIALHVADGLRYLHSA---MIIYRDLKPHNVLLFTLY---PN
AAIIAKIADYGIAQYCCRMGIKTSEGTPGFRAPEVARGNVIYNQQADVYSFGLLLYDILTTGGRIVEGLKFPNEFDELEI
QGKLPDPVKEYGCAPWPMVEKLIKQCLKENPQERPTSAQVFDILNSAELV*

>P1;1fvr
structureX:1fvr::A::::::
IKFQDVIGEGNFGQVLKARIKKDGLRMDAAIKRM-----KEYRDFAGELEVLCKLGHHPNIINLLGACEHRGYLYLAIEY
APHGNLLDFLRKSRVLETDPAFAIANSTASTLSSQQLLHFAADVARGMDYLSQK---QFIHRDLAARNILVG--------
ENYVAKIADFGLSRGQEVY---VKKLPVRWMAIESLNYSV-YTTNSDVWSYGVLLWEIVSLGGTPYCGMTCAELYEKLPQ
GYR-LEK----PLNCDDEVYDLMRQCWREKPYERPSFAQILVSLNRMLEE*

>P1;4uy9
structureX:4uy9::A::::::
LTLEEIIGIGGFGKVYRAFWIGDEVAVKAARHDPDEDISQTIENVRQEAKLFAML-KHPNIIALRGVCLKEPNLCLVMEF
ARGGPLNRVLSGKR-----------------IPPDILVNWAVQIARGMNYLHDEAIVPIIHRDLKSSNILILQKVENGDL
SNKILKITDFGLAREWHRTTKMSAAGTYAWMAPEVIRASM-FSKGSDVWSYGVLLWELLTGE-VPFRGIDGLAVAYGVAM
NKLALPI----PSTCPEPFAKLMEDCWNPDPHSRPSFTNILDQLTTIEES*

```

Appendix 2. 2: Target- template input .PIR alignment variant 1 file

```
>P1;2019
sequence:2019::::::::
QAPEFLLGDGSFGSVYRAAYEGEEVAVKIFNKH------TSLRLLRQELVVLCHL-HHPSLISLLAAGI
R--PRMLVMELASKGSLDRLLQQDKA---------------SLTRTLQHRIALHVADGLRYLHSA---MIIYRDLKPHNV
LLFTL---YPNAAIIAKIADYSIAQYCCRMGIKTSEGTPGFRAPEVARGNVIYNQQADVYSFGLLLYDILTTGGRIVEGL
KFPNEFDELEIQGKLPDPVKEYGCAPWPMVEKLIKQCLKENPQERPTSAQVFDILNSAELV*

>P1;1fvr
structureX:1fvr::A::::::
IKFQDVIGEGNFGQVLKARIKKDGLRMDAAIKRM-----KEYRDFAGELEVLCKLGHHPNIINLLGACE
HRGYLYLAIEYAPHGNLLDFLRKSRVLETDPAFAIANSTASTLSSQQLLHFAADVARGMDYLSQK---QFIHRDLAARNI
LVG--------ENYVAKIADFGLSRGQEVY---VKKLPVRWMAIESLNYS-VYTTNSDVWSYGVLLWEIVSLGGTPYCGM
TCAELYEKLPQGYR-LE----KPLNCDDEVYDLMRQCWREKPYERPSFAQILVSLNRMLEE*

>P1;4uy9
structureX:4uy9::A::::::
LTLEEIIGIGGFGKVYRAFWIGDEVAVKAARHDPDEDISQTIENVRQEAKLFAML-KHPNIIALRGVCL
KEPNLCLVMEFARGGPLNRVLSGKR-----------------IPPDILVNWAVQIARGMNYLHDEAIVPIIHRDLKSSNI
LILQKVENGDLSNKILKITDFGLAREWHRTTKMSAAGTYAWMAPEVIRAS-MFSKGSDVWSYGVLLWELLTGE-VPFRGI
DGLAVAYGVAMNKLALP----IPSTCPEPFAKLMEDCWNPDPHSRPSFTNILDQLTTIEES*

```

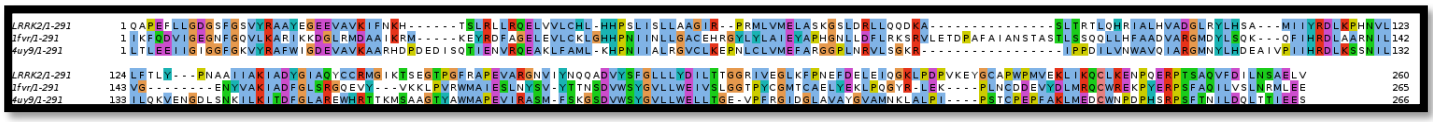
Appendix 2. 3: Target- template input .PIR alignment variant 2 file

```
>P1;2020
sequence:2020:::::::
QAPEFLLGDGSFGSVYRAAYEGEEVAVKIFNKH------TSLRLLRQELVVLCHL-HHPSLISLLAAGI
R--PRMLVMELASKGSLDRLLQQDKA---------------SLTRTLQHRIALHVADGLRYLHSA---MIIYRDLKPHNV
LLFTL---YPNAAIIAKIADYGTAQYCCRMGIKTSEGTPGFRAPEVARGNVIYNQQADVYSFGLLLYDILTTGGRIVEGL
KFPNEFDELEIQGKLPDPVKEYGCAPWPMVEKLIKQCLKENPQERPTSAQVFDILNSAELV*

>P1;1fvr
structureX:1fvr::A::::::
IKFQDVIGEGNFGQVLKARIKKDGLRMDAAIKRM-----KEYRDFAGELEVLCKLGHHPNIINLLGACE
HRGYLYLAIEYAPHGNLLDFLRKSRVLETDPAFAIANSTASTLSSQQLLHFAADVARGMDYLSQK---QFIHRDLAARNI
LVG--------ENYVAKIADFGLSRGQEVY---VKKLPVRWMAIESLNYS-VYTTNSDVWSYGVLLWEIVSLGGTPYCGM
TCAELYEKLPQGYR-LE----KPLNCDDEVYDLMRQCWREKPYERPSFAQILVSLNRMLEE*

>P1;4uy9
structureX:4uy9::A::::::
LTLEEIIGIGGFGKVYRAFWIGDEVAVKAARHDPDEDISQTIENVRQEAKLFAML-KHPNIIALRGVCL
KEPNLCLVMEFARGGPLNRVLSGKR-----------------IPPDILVNWAVQIARGMNYLHDEAIVPIIHRDLKSSNI
LILQKVENGDLSNKILKITDFGLAREWHRTTKMSAAGTYAWMAPEVIRAS-MFSKGSDVWSYGVLLWELLTGE-VPFRGI
DGLAVAYGVAMNKLALP----IPSTCPEPFAKLMEDCWNPDPHSRPSFTNILDQLTTIEES*
```

Appendix 2. 4: Sequence alignment between LRRk2 (kinase domain) and the two templates (1FVR and 4UY9) viewed in jalview



Appendix 3. 1: The wildtype docking energies and distance presented in Microsoft excel.

Ligand_list	Ligand_list	Ligand_list	Ligand_list	euc_distance	Energies	Energies
14.5305	wt_57.pdbqt_SANC00178_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00178_minRM1.vinaall.log	-11.6
18.4511	wt_57.pdbqt_SANC00479_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00479_minRM1.vinaall.log	-11.2
41.4401	wt_57.pdbqt_SANC00482_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00482_minRM1.vinaall.log	-10.8
41.4401	wt_57.pdbqt_SANC00484_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00484_minRM1.vinaall.log	-10.8
18.6518	wt_57.pdbqt_SANC00491_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00491_minRM1.vinaall.log	-10.8
19.3985	wt_57.pdbqt_SANC00447_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00447_minRM1.vinaall.log	-10.7
41.4401	wt_57.pdbqt_SANC00485_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00485_minRM1.vinaall.log	-10.6
34.055	wt_57.pdbqt_SANC00446_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00446_minRM1.vinaall.log	-10.5
15.5666	wt_57.pdbqt_SANC00478_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00478_minRM1.vinaall.log	-10.5
14.0295	wt_57.pdbqt_SANC00489_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00489_minRM1.vinaall.log	-10.5
19.5082	wt_57.pdbqt_SANC00490_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00490_minRM1.vinaall.log	-10.3
32.8543	wt_57.pdbqt_SANC00448_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00448_minRM1.vinaall.log	-10.2
41.4401	wt_57.pdbqt_SANC00483_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00483_minRM1.vinaall.log	-10.1
41.4401	wt_57.pdbqt_SANC00486_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00486_minRM1.vinaall.log	-10.1
41.4401	wt_57.pdbqt_SANC00487_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00487_minRM1.vinaall.log	-10.1
34.7008	wt_57.pdbqt_SANC00480_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00480_minRM1.vinaall.log	-10
41.4401	wt_57.pdbqt_SANC00481_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00481_minRM1.vinaall.log	-10
20.791	wt_57.pdbqt_SANC00449_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00449_minRM1.vinaall.log	-9.7
17.6854	wt_57.pdbqt_SANC00384_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00384_minRM1.vinaall.log	-9.4
41.4401	wt_57.pdbqt_SANC00685_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00685_minRM1.vinaall.log	-9.4
17.0509	wt_57.pdbqt_SANC00386_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00386_minRM1.vinaall.log	-9.3
16.0756	wt_57.pdbqt_SANC00488_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00488_minRM1.vinaall.log	-9.3
41.4401	wt_57.pdbqt_SANC00526_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00526_minRM1.vinaall.log	-9.3
41.4401	wt_57.pdbqt_SANC00669_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00669_minRM1.vinaall.log	-9.3
41.4401	wt_57.pdbqt_SANC00406_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00406_minRM1.vinaall.log	-9.2
41.4401	wt_57.pdbqt_SANC00417_minRM1.vinaall.			pdbqt	../log_files/wt_57.pdbqt_SANC00417_minRM1.vinaall.log	-9.2

Appendix 3. 2: The mutant G2019S docking energies and distance presented in Microsoft excel.

Distance	Energies					Energy
10.8249	mut_1956.pdbqt_SANC00101_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00101_minRM1.vinaa	-7.5
17.098	mut_1956.pdbqt_SANC00102_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00102_minRM1.vinaa	-8
12.5505	mut_1956.pdbqt_SANC00103_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00103_minRM1.vinaa	-7.4
15.3604	mut_1956.pdbqt_SANC00104_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00104_minRM1.vinaa	-6.8
12.4231	mut_1956.pdbqt_SANC00105_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00105_minRM1.vinaa	-7.9
12.5588	mut_1956.pdbqt_SANC00106_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00106_minRM1.vinaa	-7.5
18.3101	mut_1956.pdbqt_SANC00107_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00107_minRM1.vinaa	-8.1
11.8022	mut_1956.pdbqt_SANC00108_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00108_minRM1.vinaa	-5.7
10.4059	mut_1956.pdbqt_SANC00109_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00109_minRM1.vinaa	-6.4
11.817	mut_1956.pdbqt_SANC00110_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00110_minRM1.vinaa	-6.2
10.7497	mut_1956.pdbqt_SANC00111_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00111_minRM1.vinaa	-6.3
11.2369	mut_1956.pdbqt_SANC00112_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00112_minRM1.vinaa	-6.4
11.729	mut_1956.pdbqt_SANC00113_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00113_minRM1.vinaa	-6
13.4087	mut_1956.pdbqt_SANC00114_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00114_minRM1.vinaa	-6.2
11.774	mut_1956.pdbqt_SANC00115_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00115_minRM1.vinaa	-6
18.3101	mut_1956.pdbqt_SANC00116_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00116_minRM1.vinaa	-6.3
18.3101	mut_1956.pdbqt_SANC00117_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00117_minRM1.vinaa	-6.5
12.2254	mut_1956.pdbqt_SANC00118_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00118_minRM1.vinaa	-8.2
19.1598	mut_1956.pdbqt_SANC00119_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00119_minRM1.vinaa	-9.4
17.8154	mut_1956.pdbqt_SANC00120_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00120_minRM1.vinaa	-9.2
19.3572	mut_1956.pdbqt_SANC00121_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00121_minRM1.vinaa	-7.9
17.935	mut_1956.pdbqt_SANC00122_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00122_minRM1.vinaa	-8.2
18.8575	mut_1956.pdbqt_SANC00123_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00123_minRM1.vinaa	-8.3
15.7298	mut_1956.pdbqt_SANC00124_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00124_minRM1.vinaa	-8.3
15.2114	mut_1956.pdbqt_SANC00125_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00125_minRM1.vinaa	-8.2
16.2252	mut_1956.pdbqt_SANC00126_minRM1.vinaall.pdbqt				../log_files/mut_1956.pdbqt_SANC00126_minRM1.vinaa	-8.6

Appendix 3. 3: LRRK2 kinase domain in the triclinic solvated system.