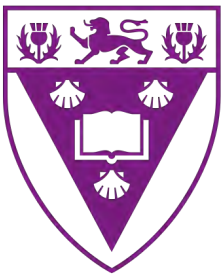


RHODES UNIVERSITY

Where leaders learn

Inhibitor search and variant analysis of Acetylcholinesterase

A minor thesis submitted in partial fulfillment of the degree

of

Master of Science

by

Coursework/Thesis

in

Bioinformatics and Computational Molecular Biology

in the Department of Biochemistry and Microbiology

Faculty of Science

by

Harnaud Ras

December 2019

Abstract

Acetylcholinesterase (AChE) inhibition is used to treat Alzheimer's disease by increasing the availability of acetylcholine to carry nerve signals in the brain. The response to this treatment varies widely, which may be due to altered affinity to the current drugs caused by genetic variation. Various negative side-effects limit their use. As this is one of the only available therapeutic drug targets to treat Alzheimer's disease, decreasing the negative effects is of great importance. AChE is involved in biological processes that occur after acute ischemic stroke. Stroke is the third leading cause of death worldwide, and 87% of all stroke cases belong to ischemic stroke. AchEI (cholinesterase inhibitors) have been suggested to have properties that lower the risk of stroke. AChE is one of 15 verified drug targets under study for treatment of stroke. In addition to Alzheimer's disease and stroke, Lewy body disease (LBD) may be treated using cholinesterase inhibitors. The goals of this study are to find inhibitors that can potentially be used to treat Alzheimer's disease and/or stroke and to investigate variants which may affect protein dynamics and function. Two variants were analyzed, P247L and T229S. Molecular simulation of the P247L variant resulted in a disruption in protein dynamics in comparison to the wild-type. A total of 5728 molecules were screened and 10 nanosecond simulations were used to narrow down the set of compounds. The four best performing molecules were simulated for 10 nanoseconds. MM-PBSA was performed to identify molecules with high binding free energies.

Declaration

The research described in this thesis was carried out as part of the one-year MSc course-work and research thesis programme in Bioinformatics and Computational Molecular Biology. The research part of the course was done between 15 July and 30 November 2018 under the supervision of Prof Özlem Tastan Bishop.

I, Harnaud Ras, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature .....

Date .....

Acknowledgements

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. The opinions expressed and derived conclusions are those of the author and are not necessarily to be attributed to the NRF.

The Center for High Performance Computing (CHPC) in Cape Town, South Africa, is acknowledged for providing computational time on high performance computing clusters.

Phillip Kimuda, a fellow Rhodes Bioinformatics student is acknowledged for supplying the structure files of a filtered set of compounds from the ZINC15 database which was used for screening during this project.

Contents

List of Figures

List of Tables

List of Web Servers and Applications

1. Automatic Topology Builder (ATB)

   https://atb.up.edu.au

2. ensemble genome browser

   https://www.ensembl.org

3. Jalview

   www.jalview.org

4. Molinspiration Cheminformatics

   www.molinspiration.com

5. PyMOL

   https://pymol.org

6. RCSB

   https://www.rcsb.org

7. Uniprot

   https://www.uniprot.org

8. VMD

   http://www.ks.uiuc.edu/Research/vmd/

List of Abbreviations

Abbreviation	Phrase
ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
APO	Apo protein - protein with no inhibitor or ligand bound to it
ATB	Automated Topology Builder
BC	Betwenness Centrality
PRiMA	Proline-rich Membrane Anchor
ColQ	Collagenic tail peptide
DNA	Deoxyribonucleic acid
DOPE	Discrete Optimized Protein Energy
FDA	Food and Drug Administration
GROMACS	GROningen MAchine for Chemical Simulations
GWAS	Genome Wide Association Studies
MM-PBSA	Molecular Mechanics Poisson Boltzmann Surface Area
MSA	Multiple Sequence Alignment
ns	nanoseconds
PCA	Principle Component Analysis
PDB	Protein Data Bank
RCSB	Research Collaboratory for Structural Bioinformatics
Rg	Radius of gyration
rhAChE	recombinant human Acetylcholinesterase
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SVM	Support Vector Machine
VMD	Visual Molecular Dynamics

Chapter 1
Literature Review and Aim

1.1 The basic physiology behind stroke and Alzheimer's

1.1.1 Basic physiology behind stroke

Stroke and Alzheimer's are worth looking at together since stroke increases the risk of Alzheimer's. Two types of stroke are encountered. The most common form of stroke is ischemic stroke, which is caused by the closing or blockage of a major cerebral artery. The second type of stroke is hemorrhagic stroke, which is the result of bleeding either in or on the brain. (Zhang et al., 2003). There is one drug that is currently FDA approved to be used to treat acute ischemic stroke. The drug is a thrombolytic peptide and is called rt-PA (recombinant tissue plasminogen activator). Large numbers of neurons, synapses and myelinated fibers die during and immediately after the onset of ischemic stroke, therefore it is necessary to treat the patient as soon as possible to mitigate damage. For this reason, the rt-PA thrombolytic drug is only effective if administered within 3 hours (Neumann-Haefelin et al., 2002). It has been estimated that on average, for every hour that the stroke remains untreated, 3.6 years' worth of brain matter dies. A series of events is put into motion by an acute ischemic stroke that ultimately leads to the destruction of neural cells. Various drugs have been developed that can act on mechanisms in this cascade that helps to protect neurons from damage. A large problem with the drugs that have been tested to treat stroke so far is that they increase the risk of hemorrhage significantly (Wardlaw et al., 1997). It appears that many thrombolytic compounds have this problem. Other reasons that ischemic stroke may not be treated by rt-PA, is that it is difficult to administer treatment quick enough (within 3 hours) and the type of stroke has to be confirmed first as thrombolytic agents would make an hemorrhagic stroke worse. (Green, 2009)

A stroke is followed by several biological responses. This includes the ischemic cascade and restorative responses. Some regions in the brain are salvageable for a small

amount of time after the stroke. Few patients receive treatment to recover these regions because of the small window of time. Therapy that can be administered in the weeks after a stroke will be accessible to a larger portion of patients. Drugs that are already in use for other neurological diseases could be repurposed to aid in recovery after stroke (Pardridge, 2007). Stroke sufferers may vary in terms of their mental and physical health before the occurrence of the stroke. As a result of physical and genetic differences between individuals, stroke is a highly variable occurrence and so one drug is not sufficient for all cases of stroke. Various drugs may be used to aid in the recovery process. Many genes are involved in the processes that occur after an individual suffers a stroke and so certain drugs may perform better than others, according to the specific case. The chances that a drug will benefit a patient may depend on genetic variation. Variation such as a certain polymorphism may theoretically prevent a drug from functioning correctly. The weeks and even months after stroke occurs are important for neural recovery. This means that the treatment of stroke is time sensitive and treatment is only viable for a limited amount of time after stroke. Time sensitivity dictates that drugs must be used at the correct time as administering drug at the wrong time may have negative effects or exacerbate the condition. (Cramer, 2015)

1.1.2 Symptoms and causes of Alzheimer's

Alzheimer's is a neurodegenerative disorder that involves important high-functioning brain areas such as the hippocampus and neocortex. Formations in the brain such as neurofibrillary tangles, Beta-amyloid plaque deposits and the loss of neural synapses, contribute towards the development of the disease. Alzheimer's is a major cause of dementia which increases the risk of ischemic stroke as well. (Francis et al., 1999). By 2050, 131.5 million people are predicted to suffer from dementia. The costs on patient care are high and families' lives are disrupted, making the development of treatment for this disease a high priority (Tan et al., 2018).

Most approved drugs that treat symptoms of Alzheimer's disease are cholinesterase inhibitors, with the exception of memantine which is a NMDA (N-methyl-D-aspartate)

receptor antagonist (Noetzli and Eap, 2013). Acetylcholinesterase inhibitors help acetylcholinergic function. The fact that acetylcholine is critical for learning and memory, indicates that this neurotransmitter is not being used properly in the case of Alzheimer's disease. This is called the "cholinergic hypothesis of Alzheimer's disease". Along with the cholinergic hypothesis, other factors play a part in developing the disease such as the degeneration of neurons and synapses in brain regions that are central to the development of memory and learning (Francis et al., 1999). AChE (Acetylcholinesterase) inhibitors are used for treatment because they fix the acetylcholine deficit by reducing it's degradation by acetylcholinesterase (Grossberg, 2003). Pharmacogenetic studies have focused on the enzymes that metabolize AChE inhibitors. These enzymes determine the amount of inhibitor available in the plasma and so polymorphisms that alter the functionality of these metabolizers will affect the effectiveness of the treatment (Noetzli and Eap, 2013).

1.2 Protein targets with potential to treat stroke or Alzheimer's

1.2.1 Plasminogen activator inhibitor-1 (PAI-1)

Plasminogen activator inhibitor-1 is one potential target protein for the treatment of ischemic stroke. PAI-1 inhibits fibrinolysis which is instrumental in recovery after stroke. (J.-Q. Liu et al., 2017) Fibrinolysis is a biological process that breaks down blood clots.

1.2.2 Acetylcholinesterase (AChE)

Acetylcholinesterase (AChE) is a potential drug target for anti-stroke therapy. Acetylcholinesterase is responsible for terminating nerve impulses at synapses that use acetylcholine neurotransmitters. It does this by hydrolyzing the acetylcholine, breaking it up into acetyl and choline. Thus the neurotransmitter is prevented from carrying on its signal. AchEI (cholinesterase inhibitors) have been suggested to have properties that lower the risk of stroke (Tan et al., 2018). Cholinesterase inhibitors will decrease the breakdown of the neurotransmitter, keeping the nerve impulse going. Acetylcholine is

important for the functioning of the declarative memory system. Acetylcholinesterase, which functions to break down acetylcholine, is therefore an important neuropharmacological target (Gais and Born, 2004). AChE is one of 15 verified drug targets under study for treatment of stroke (J.-Q. Liu et al., 2017). Various studies screen or dock compounds to this protein with the goal of finding new inhibitors that can potentially be used to treat symptoms of dementia and stroke. Compounds that can provide the benefits of currently approved drugs without any of the negative effects, may prove useful for the treatment of Alzheimer's and stroke. Diseases for which acetylcholinesterase inhibitors may be of therapeutic use include:

- Alzheimer's
- Acute ischemic stroke
- Lewy body disease (LBD) (Lam et al., 2009)
- Vascular dementia

These diseases are characterized by memory impairment as well as hallucinations.
The human acetylcholinesterase gene appears on chromosome 7q22.1 and is 7.4 kb in length. Exons 2 to 6 in this gene are translated and variants have been identified in these regions (Lockridge et al., 2016). Acetylcholinesterase plays a critical biological role and is therefore highly conserved in sequence and structure. It was believed that this gene had few polymorphisms but later it was suggested that this gene had and average amount of polymorphisms (Yue and Moult, 2006a). The human AChE gene codes for a protein that has three isoforms. The isoforms have similar catalytic attributes even though the quaternary structure of the final protein differs. Of the isoforms, only one occurs in the brains of humans. This form is known as the AChE-T variation of the protein. The various isoforms are generated by alternative splicing. Acetylcholinesterase hydrolyses its substrate at a high rate even though its active site opens and closes rapidly for every molecule that enters. The functional unit of AChE-T occurs as 4 sub-units that are connected with either a collagenous protein COlQ or a transmembrane protein, which anchors the functional Unit to a membrane.

Alternative splicing creates alternate C-terminal structures for the protein depending on what membrane it is bound and the location in the body the protein is functioning (Anglister and Silman, 1978). This alternative C-terminal tail produces the three isoforms of AChE. For the AChE-T isoform occurring in brain and muscle tissue, up to three tetramers can be connected to a collagenous protein that runs through the middle of the tetramer and consists of three strands (Vigny et al., 1978). COlQ includes a proline-rich region near its N-terminal end which is involved in attachment to AChE-T subunits and two cysteine residues that are next to this region form disulfide interactions with cysteine residues from the C-terminal end of the acetylcholinesterase subunits (Bon et al., 1997), (Dvir et al., 2010). Three isoforms of AChE are encountered. The differences occur at the C-terminus of the variants, where alternative splicing leads to different ending sequences (Soreq et al., 1990). This 'extra' region which is added to the main body of the protein, is 40 residues in length (Grisaru et al., 1999). The core of the protein that can be found in all of its isoforms, is 543 residues in length. All three isoforms have the same catalytic properties. (Soreq et al., 1990). The active site contains three important residues known as the catalytic triad. These residues aid in hydrolysis of acetylcholine. There is a second site in the protein structure that is important for the interaction with small molecules known as the peripheral anionic site (PAS). The variant of the protein that is mostly found in brain and muscle tissues, is known as the AChE-T or AChE-S variant (Grisaru et al., 1999). The motions of the tetramer functional unit cause the active sites of certain subunits to be momentarily obstructed or occluded. On average the active sites are estimated to be unobstructed 80% of the time. Residues 341 and 286 at the entrance of the active site are commonly obstructed by a neighboring AChE subunit. The subunits fluctuate relative to each other and the motion of two monomers has been described as a shearing motion (Gorfe et al., 2008).

1.2.3 Approved inhibitors that target Acetylcholinesterase and butyrylcholinesterase

There are three approved drugs that target either Acetylcholinesterase or Butyrylcholinesterase (BChE). These compounds are Rivastigmine, Donepezil and Galantamine. These differ in their application depending on the severity of the Alzheimer's case. Rivastigmine is hydrolyzed by AChE like its native substrate ACh. This compund has no selectivity difference between AChE or BuChE. Rivastigmine inactivates both these enzymes for a few hours. Donepezil is a reversible inhibitor that prefers to bind to AChE over BuChE by 300 times. It is not a competitive inhibitor. Galantamine binds competitively to AChE and prefers it 50 fold over BuChE. It also modulates nicotinic receptors which enhances the function of the cholinergic system (Wilkinson et al., 2004).

The effectiveness of these drugs vary widely between patients and negative side effects like vomiting and diarrhea are experienced also at a variable rate. This is a multifactorial problem and genetic variability likely plays a role in this variance although no specific subgroups are known that respond more or less favorably to the medication (Noetzli and Eap, 2013).

1.3 Genetic variation of cholinesterase family of enzymes

One non-synonymous variant stands out in terms of its frequency. This is the Histidine to Asparagine substitution at position 353 (canonical sequence). This variant occurs at a rate of over 40%. Although this mutation might not effect binding of a drug compound it is responsible for the YT-2 blood group phenotype and is an important factor when considering compatibility of donating blood or organs (Hasin et al., 2004). This variant has no effect on the catalytic properties of AChE (Bartels and Zelinski, 1993).

Mutations in acetylcholinesterase that are deleterious are rare since this is a criti-

cal protein in the functioning of any animal. Deleterious variants of this protein express a heterozygous pattern of inheritance (Valle et al., 2011). Butyrylcholinesterase (BChE) is of the cholinesterase family of proteins along with AChE. This enzyme is more prone to deleterious mutations and therefore may also show more variation in binding to cholinesterase inhibitors than AChE. BChE is able to protect AChE from organophosphate poisoning by acting as a scavenger to remove the nerve agents. This means that variation in the BChE protein may lead to differences in susceptibility of an individual to poisoning from nerve agents (Lockridge, 2015).

1.4 Therapeutic benefit of molecules that reversibly inhibit acetylcholinesterase

A variety of diseases exist that involve the acetylcholine receptor, such as vascular dementia, Alzheimer's, and Lewy body disease. Understanding the interaction on inhibitors with this enzyme can enable the development of better therapeutic drugs to treat these diseases. The acetylcholinesterase enzyme is also the target of nerve agents, which are organophosphates that irreversibly inhibit this enzyme. Molecules that enable acetylcholine to be hydrolyzed while blocking nerve agents from reacting with the enzyme may provide treatment options or preventative methods for organophosphate poisoning (Cheung et al., 2013). Peripheral site inhibitors may block only the nerve agent organophosphate while leaving enough space for the substrate, acetylcholine to be hydrolyze. One such molecule is Dihydrotanshinone (DHI), which may be used to find other molecules that selectively bind to the peripheral active site (PAS) of the protein (Beri et al., 2013).

1.5 Identification of effective drug molecules

1.5.1 Potential drugs that target AChE

There are many different available compounds that inhibit the protein and the compounds are assisted in reaching the active site by an electrostatic forces. TerritremB is a potent inhibitor of this enzyme. TerritremB binds both to the peripheral anionic cite

and the central anionic site, which contains the catalytic triad. This inhibitor initiates shifts the backbone positions of residues that form the active site gorge. Inhibitors may bind to the peripheral anionic site (PAS), the central anionic cite (CAS), or both (Cheung et al., 2013). (Felder et al., 1997).

Searches for drugs that have the potential to treat vascular dementia and acute ischemic stroke are being done. Most drugs have many negative side effects which makes finding a drug with low toxicity a priority. Having detailed knowledge of how the drug functions to treat stroke, makes it less likely that unknown harmful side effects will be experienced. Many anti stroke drugs function by preventing or reducing blood blockages, e.g. thrombolytics and anti-coagulants. Others protect neural cells from degenerating or being damaged. These are called neuroprotective agents. (J.-Q. Liu et al., 2017).

In a the Chinese medicinal plants study that isolated the top 1 % of anti-stroke compounds, 35 compounds remained after ADMET testing which was done to verify that the drug lacked overly harmful properties. Of these compounds, 9 were able to bind to more than one anti-stroke target while the other 26 compounds each were found to bind to a specific target each. These are known as single target compounds, and these 26 have not been exhaustively used in research which tries to discover anti-stroke compounds. Docking the compounds from anti-stroke plants were done using the ligand present in the crystal structure as a reference. The docking score of the novel compound was compared to the score of the original ligand and the docking was taken as significant if it had a similar or better score as ligand that was originally in the crystal structure. The compounds were verified by comparing their structure to that of known anti-stroke drugs. Further validation can be done by investigating the interaction between ligand and receptor in more detail by using molecular dynamics. Some attributes to consider when looking for anti-stroke compounds are: penetration of the blood brain barrier (BBB), human intestinal absorption (HIA), binding to plasma protein, hepatoxicity, and solubility (J.-Q. Liu et al., 2017).

Many drug lead searches depend on finding drugs that exhibit very high binding affinity to a target. There are factors other than binding affinity which determine the efficacy of a therapeutic drug.

Ligand efficacy may depend on residence time instead of merely binding affinity. Residence time is the amount of time that the ligand stays bound to the enzyme that it modulates or inhibits. It is the life time of the drug-target complex. The residence time determines how long the drug and the resulting effects will be active. Drug discovery attempts are often halted due to drug leads not performing well in an open system compared to a closed system. An open system like the human body, introduces factors that affect the concentration of the drug target. Measuring the half-life of the ligand-drug complex is possible in an open system, but not a closed system. A closed system limits the concentrations of the drug and enzyme target to a stat of equilibrium. The amount of time that the ligand-receptor complex is active in an open system is an indication of the efficacy of the drug (Lu and Tonge, 2010).

1.6 Effects of SNP's on protein structure and function

The vast majority of variation in human genetics, and estimated 90%, are single nucleotide polymorphisms (SNP's). SNP's refer to a single nucleotide change in the DNA sequence which may or may not result in the production of an alternate amino acid. Many of these SNP's are involved in diseases (Collins et al., 1998). The effect of a non-synonymous variant on the structure of the protein is difficult to analyze although predictions of the functional change that it will produce, can been made. The impact of an amino acid change on the 3D structure of the protein is critical when doing drug discovery as a new model that incorporates the change is required for use in molecular dynamic simulations. A model that incorporates the altered amino-acid can be created by editing an existing protein structure that is available in the protein data bank (D. Wang et al., 2015).

Non-synonymous SNP's are less prevalent than SNP's that code for an identical amino

acid. This is thought to be as a result of selection against mutant proteins that likely have deleterious properties introduced by the non-synonymous SNP's. Much of the variation in proteins may still be due to non-synonymous SNP's.

Some amino-acid changes have been shown to effect the way that the protein responds to a particular drug, including altering the toxicity of the drug and the binding affinity. Examples where amino-acid changes may influence response to drugs include the B2-adrenergic receptor and cytochrome P450. It was estimated that between 26% and 32% of all of the non-synonymous SNPs that occur throughout the human genome will have functional consequences (Chasman and Adams, 2001).

The non-covalent interactions between the protein and the ligand play and important role in the successful binding of the compound to the receptor. Van der Waals forces are small but add up to a large portion of the binding free energy of the protein ligand complex. This enables the molecule to bind to and inhibit the protein. These non-covalent interactions such as electrostatic forces and van der Waals interactions may be altered by a change of amino-acid which normally acts on the bound molecule (Yue and Moult, 2006b). The energies of these bonds and forces should be calculated when deciding whether the amino-acid substitution has an effect on the binding on the ligand to the mutant protein. Autodock Vina can be used to calculate the amount of hydrogen bonds that form between the two interacting molecules when doing a molecular docking experiment. Hydrogen bonds are an important non-covalent force to consider the effects of a mutation of the receptor protein (Nagasundaram et al., 2015).

1.7 predicting the effects of SNP's on structure and function

Proteins contain hotspot-surfaces which are the spots that interact with other proteins. The variation introduced by amino-acid substitution may alter these surfaces by changing the folding of the protein, the electrostatic properties of the spot and the preference of ligands to bind to the surface (Nagasundaram et al., 2015). Analysis of

such a mutation occurring in a cancer therapy target has been performed. This target is used to reduce the growth of tumors (Carlson et al., 1996). The effect on the ability of the CDK4 gene to bind with Cyclin-D1 proteins was considered and the impact on drug binding was analyzed. Virtual screening was used to find compounds that would be suitable to bind to the target and 5 of its variants, each containing a SNP. Molecular dynamics was performed to gain detailed information on the interactions effected by the structural changes. Investigations were done examining the effect of nsSNP's on the activity of flavopiridol which is an inhibitor of the CDK4 protein. The flavopiridol drug binds to the ATP binding site of the protein. It was found that for the variants, the drug did not bind to the same site as in the wild type protein. This is an example where single amino-acid substitutions affected the binding affinity of a drug to the protein. The wild type protein delivered the highest binding score compared to the 5 of its variants. The wild type protein delivered a score of -8.8kcal/mol while the energies after docking to the variants ranged from -7.1 to -7.7kcal/mol. This is a significant change in the binding affinity between the reference and altered structures. Although the amino-acid change did not seem to affect any of the residues in the ATP binding groove, it hindered the drug from binding. (Nagasundaram et al., 2015)

1.8 Databases containing SNP information

Ensemble hosts information from many different databases that relate SNP information to phenotype changes and links protein information to a SNP (Hubbard, 2002). This information can be queried using the BioMart tool. Tools that focus on the structural impact of a specific SNP include PinSnps and LS-SNP-PDB (Ryan et al., 2009). These tools help to visualize the region of amino-acid substitution. GWAS studies can be followed up with further investigation to determine the role that the SNP plays in disease by looking at the changes in structure and sequence. Some tools have been developed that do this analysis. Some take into account only sequence information to predict whether it will have a negative impact and possibly cause the disease. Conserved amino-acids tend to be important for biological function. This is the centrality-lethality principle where conserved sequences are likely central to the

function of the organism and thus disruptions in these sequences or genes will likely affect many other systems in the organism that are required for the organism to function. Other tools focus on using structural information as the structure of a protein determines its function.

With a growing number of SNP prediction tools it is useful to determine which tools are most accurate. A consensus can be used which is given by programs such as PredictSNP and Meta-SNP. This can then help to decide on the most likely functional change that the SNP will produce. A popular method in the structural analysis of proteins is homology modeling. The SNPs can be incorporated into the structure when modeling and their effects on the protein can be analyzed by further processes such as docking and molecular dynamic simulations. High throughput screening of compounds allows researchers to quickly narrow down possible drug compounds that are worth studying. A SNP will effect docking results if the SNP is an amino-acid which interacts with the ligand at the active site. (David K. Brown and Bishop, 2017)

1.9 The role of synonymous SNPs in heterogeneity of drug response

The increased rate of sequencing methods, and developments in the fields of bioinformatics and many others, combined with high throughput screening molecular docking, has increased interest and viability of personalized medicine. (The International SNP Map Working Group, 2001). There are a large number of polymorphisms in the human genome. This prompts researchers to neglect synonymous polymorphisms in favor of non-synonymous SNPs as it is less likely that silent polymorphisms have phenotypic consequences (Sauna et al., 2007). Single nucleotide polymorphisms (SNPs) are defined as having a minimum allele frequency of 1%. This means that the least frequent allele needs to occur at least at 1% in a population. Because of the frequency of SNPs in the human genome, they have been categorized according to significance. In this classification, synonymous SNP's are not classified as important as they are less likely to result in phenotypic alterations (Risch, 2000). Synonymous SNPs may affect protein expression. Various mechanisms may bring about these variations in expression.

Synonymous SNPs may affect the stability of mRNA, the use of rare codons which effects translation, and alternate splicing. These mechanisms may lead to change in the amounts of the protein being produced. These changes may mean that variation in drug response is possible due to silent polymorphisms that alter stability of mRNA (Sauna et al., 2007)

Investigation into the influence of SNP's in genes related to acetylcholine production and metabolism on the effectiveness of Acytelcholinesterase inhibitors, concluded that the rs2571598 – AA genotype of AChE lead to better response to Rivastigmine treatment. This SNP is an intron variant and does therefore not code for a different amino acid. Other genes examined include butyrylcholinesterase (BChE) and Choline acetyltransferase (ChAT), which produces acetylcholine. SNP's examined in these genes did not lead to a significant change in response to treatment (Scacchi et al., 2009)

1.10 Effects of structural variations on binding of inhibitors to proteins

Non-synonymous substitutions may alter drug response through a variety of mechanisms. It may alter a gatekeeper residue. This residue allows the substrate or drug to enter the active site but the variant gatekeeper no longer performs properly. The substitution may change the binding pocket of the protein, by changing the shape of the pocket. The stability of the protein may be altered by variations that occur in the protein. A drug requires the protein to be in a specific conformation, the SNP might prevent the protein from reaching this conformation or from remaining in this conformation long enough. SNP's may change the thermal stability of the protein. Thermodynamic instability may prevent a protein from performing its desired function. Drugs may be used to try to alter the thermodynamic stability of the protein (Lahti et al., 2012). Conserved protein regions tend to be functionally important. This tells us that amino-acid changes in regions that are highly conserved will have a greater probability of having deleterious effects. The sequence based approaches

to predict functional effect of SNPs works off of this principle and thus alignment of homologous sequences give us information of whether the SNP has occurred in a conserved region. If the altered amino acid shares physicochemical properties with the wild type amino acid, it is less likely to have deleterious effects. As the SNP's under study are not necessarily associated with disease, they may not be selected against by natural selection but may still effect the efficacy drugs.

Sequence based methods useful for the majority of predictions (83%) as the amount of sequences is immense and the number of solved structures is little in comparison (14%) (Ng and Henikoff, 2006). Mutations may not always be neutral or deleterious, in some cases the mutant form of a protein may be beneficial to the function of the protein. An example of such a case is sickle cell anemia. Prediction should be done with bone structural and sequence information. Crystal structures are not completely reliable on its own as it is isolated from effects outside of the crystal. Sequence methods have continually increasing sets of information to base predictions off of as the sequencing capacity and rate is increasing rapidly (Z. Wang and Moult, 2001). The free binding energy of ligand-receptor interactions is determined by the changes in entropy and enthalpy when the molecule binds or tries to bind. The enthalpy refers to the ability of two molecules to form stabilizing forces between them such as hydrogen bonds and salt bridges. Entropy refers to how well the ligand fits in the active site and whether it has freedom to move around until it finds an optimal position. The stabilizing forces that are able to form are determined by the shape complementarity between the molecules as well as the types of interactions that can occur between the specific molecules. Are the molecules physicochemical complementary? (Lahti et al., 2012) Bonds form between certain atoms when they are at certain optimal positions from another. SNP's that because amino-acid substitution may disrupt molecules that are important for forming these stabilizing forces or for creating an optimal shape for binding. This is especially likely if the mutation occurs in one of the residues that is active in the active site. Protein receptors may take on different conformations depending on which ligand is bound to it. This is known as induced fit. These conformational

changes occur open binding of the small molecule to the receptor. This means that the crystal structure that is being used that includes an inhibitor may be in its specific conformation because the ligand induced it to change into such a conformation. Proteins occur naturally in varied equilibrium conformations. Ligands may bind to the proteins that are already close to its desired conformation. (Lahti et al., 2012) Further induced conformational change can then take place. This suggests that different drugs will have different preferences in the conformations that they bind to. A SNP that induces a large overall conformational change may therefore effect the binding of a ligand even if it is not necessarily in the active site of the protein. Allosteric sites are less conserved than orthosteric or active sites. Structural variation is more common in allosteric sites. (Sadowsky et al., 2011) One would expect to find more missense SNP's in allosteric sites which modulate the activity of the protein. Drugs that bind to the allosteric site of the protein may either modulate the functioning of the active site by changing its affinity for the natural inhibitor. An allosteric drug may also change the protein to be inactive or active by producing a conformational change. It can be seen that allosteric sites are important in managing the reactivity and conformation of a target protein. (Schwartz and Holst, 2007) Annotations from databases such as Uniprot are useful when investigating structural significance of SNP's as they include information on areas of a protein that are involved in ligand binding and interactions with proteins. The position of the amino-acids substitution gives information such as the change in the free energy of the amino acid and the solvent accessibility among others. These features are used by prediction programs that rely on the structural information. These programs also use carbon-beta density and the crystallographic B factor of a protein structure. These prediction programs are tested by using separate datasets of amino-acid substitutions that either are known to be neutral or known to be deleterious. The error rate can be estimated by observing the amount of incorrect predictions. Information on the error rates of various tools should be considered before using the tool or to deciding on the confidence of the results. (Ng and Henikoff, 2006)

1.11 Aims and objectives

1.11.1 Aim

The aim is to identify potential inhibitors of the actylcholinesterase. A separate aim is to determine the influence of two separate SNP's on the dynamic properties of AChE.

1.11.2 Objectives

The objectives include running a standard high throughput screening of the normal AChE protein to search for potential inhibitor leads. This will incorporate the South African natural Compound Database. This is done because inhibitors may have various applications, such as the treatment of Alzheimer's, enzyme reactivation after organophosphate poisoning as well as treatment of vascular dementia and acute ischemic stroke. Assesment will be done on the influence on protein dynamics after introduction of SNP's. This will be done running molecular dynamics simulations of the protein with incorporated SNP's. Two SNP's will be analysed. SNP's may influence the effectiveness of drug by changeing their affinity to the protein.

1.12 Project layout

The following chapters start with the methods and materials. The results are together with their discussion in the chapter that follow the materials and methods. The different sections of the project include modelling, high throughput screening and molecular dynamic simulation. Modelling was required to complete the structure of the protein to be able to run the other simulations or tests. This was thus the first step. The molecular docking and molecular dynamic simulation parts also follow on each other. Potential compounds from the docking section were used to run molecular dynamic (MD) simulations, to further asses their affinity to bind to the protein. Lastly, SNP impact prediction tools were used and the scores compared to the results from MD simulations of SNP's. The results from the impact prediction tools did not always agree with results from the MD simulations.

Chapter 2
Materials and Methods

2.1 Modelling/Fixing Structures and Multiple Sequence Alignment

2.1.1 Introduction

Modelling was required to complete the rhAChE crystal structure selected for the project. This structure has the PDB ID: 4m0f and contains missing residues. rhAChE refers to recombinant acetylcholinesterase which means that the enzyme was engineered synthesised in a lab. The protein sequence was aligned to the canonical sequence which includes the residues not present in the downloaded structure. The canonical sequence, downloaded from Uniprot, is equivalent to the SEQRES sequence from the 4m0f pdb file. Modelling was done using MODELLER which functions on the principle of satisfying spacial restraints (Šali and Blundell, 1993).

There are frequently errors in the residue numbering between the pdb structure and the sequence of the gene given by databases such as Uniprot (Ryan et al., 2009). This can be fixed by aligning the protein sequence from the PDB file and the sequence of the entire protein. The residues can then be renumbered. The location of the amino acid change would then be clear. The modelling software, MODELLER, allows users to renumber residues by starting at a specific number (Šali and Blundell, 1993). Residues are often missing from solved structures of protein as crystallization of the enzyme may have been incomplete or flawed. The sequence of the protein is used to identify the type of residues that are missing. The location of these residues reletive to the others is then determined by the modeling software. This pdb structure was chosen because it is the only structure with a completely accessible catalytic cite. This is the best structure to perform docking experiments and to simulate ligands in the active site during molecular dynamics.

None of the solved structures of acetylcholinesterase observed in the pdf file archives included the 40 residue long c-terminal region in the experimentally determined crystal structure. This prevented the investigation of SNP's that fall in this region as there is no experimental data to accurately model a SNP in that region. Two of the suggested SNP's for this project fell in the c-terminal region and were also silent SNP's. This meant that they could not be investigated in this manner as they would not result in a structural change in the protein. The protein was not simulated together with its c-terminal tail in this experiment. The connection to the collagen anchor was also not simulated. A nucleotide change on the end of the protein sequence would in any case not significantly affect molecular dynamic simulations. This means that even if the SNP's which were suggested were not silent, they would still not significantly affect the protein dynamics. SNP with rs number rs17886728 is an intron variant. His353Asn is known as the yt-blood group variant and has been identified to not affect the catalytic properties of the protein. Last of the suggested SNP's to investigate is rs1799805 which is also listed as a noncoding transcript variant in dbSNP. Two other SNP's that result in altered protein structure were thus investigated. It was thus decided to simulate only the two SNP's that fell near the center of the protein core. Time constraints of half a year prompted the concentration on only two single nucleotide polymorphisms to investigate them at reasonable depth.

Along with filling in missing amino-acids, modeling was utilized to introduce single nucleotide variations into the structure. The ensemble variant table was used to select SNP's to test. The selected variants are rs143875983, P247L missense variant and rs202183011, T229S, also a missense variant. These variants are found at frequencies of less than 0.001 according to ensemble.com's variant table. Variants were selected for being the closest missense variants to the active site and being central in the enzyme. If ligands bound to the enzyme, variants close to the active site would stand a higher chance of disrupting the normal binding to the protein.

The ATOM records of PDB files can be used to identify the location of missing residues. The sequence derived from the ATOM section of the pdb was aligned to the canonical sequence of the protein in Uniprot. This allowed the renumbering of the PDB residues according to the numbering in the canonical sequence to determine the position of the desired SNV in both sequences. These sequences were aligned using MUSCLE (Edgar, 2004). The resulting alignment was then used to create a PIR file, required input for MODELLER.

2.1.2 Structure Preparation

The enzyme structure used in this research contains a complex of human recombinant acetylcholinesterase (rhAChE) with territremB, a potent and irreversible inhibitor. Recombinant proteins are expressed through an expression system other than the natural expression system in the human body. This enables expressing desired amounts of the protein for x-ray crystallography. This crystal structure was used for docking as the active site is accessible once the crystalized ligand is removed. This allowed the docking of potential inhibitors to determine whether it would bind to the site effectively. The crystal structure, PDB ID: 4m0f, contains missing residues at positions 259-264 and 495-497 in chain A. Chain B contains missing residues at positions 260-261 and 493-494. In addition, residues 2 and 3, as well as 543, are missing from both chains.

2.1.3 Modelling

A python script was prepared to run MODELLER with the appropriate settings. The script to start MODELLER can be seen in C. The missing residue position numbers were selected, which tells the program that these coordinates may be shifted. By doing this, the missing residue positions were unrestrained for modeling. This allows the backbone coordinates to be shifted to satisfy spacial restraints the best. All of the residues of known position were restricted in the movement of their backbone coordinates. This approach was used as it resulted in increased z-DOPE scores of the created models. Loop refinement was not used. 100 models were created and the

z-DOPE score for each model was calculated. The model with the best z-DOPE score was selected for further use. The more negative the z-DOPE score, higher the quality of the model and the more likely it is that the model is native-like. This score is suited to globular proteins and not very suitable to validate transmembrane proteins. This score is a statistical score and it is dependent on atomic distances. It was developed using a set of native-like structures. Verify-3D is a model quality assessment tool that compares the structure, either a model or experimentally determined, with its amino-acid sequence. A 3D-1D score is calculated indicating the compatibility of the 3D location of amino acids to its sequence position. 3D-1D scores for the 20 possible amino-acids are used to determine how well the amino-acid from the model is suited to its 3D environment (Eisenberg et al., 1997).

2.1.4 Model Improvement

The results of a simulation that included the co-crystalized ligands indicated that chain B of the enzyme was performing worse than chain A. This prompted investigation of the model quality for chain B. It was found that symmetry restraints that were used to model the homodimer tried to force symmetry between the two chains which lead to a decrease in model accuracy. The restraints option was removed from the MODELLER script and the resulting models showed an increase in z-DOPE score and verify-3D showed an increased score of 97.40 % over 94.34%. Most of the improvement was due to adjustments to chain B of the dimer. 50 models were produced and the best selected for further use. It should be noted that the first set of compounds from SANCDB were simulated using the models that were using symmetry constraints which were only later discovered to negatively impact model quality. The ZINC15 subset of molecules were simulated using the improved rhAChE model.

2.1.5 Multiple Sequence Alignment

In order to examine how well the altered residues from the SNV's are conserved, multiple sequence alignment was performed. 18 acetylcholinesterase protein sequences from 18 organisms were downloaded from Uniprot. Tcoffee with defaults was used for MSA.

Table 2.1: MSA: The organisms are listed in order from top to bottom, corresponding to the order used in figure 3.3

no.	UniProt ID	Organism
1	P22303	Homo sapiens
2	P21836	Mus Musculus
3	P07140	Drosophila melanogaster
4	P37136	Rattus norvegicus
5	P07692	Torpedo marmorata
6	Q9DDE3	Danio rerio
7	P04058	Tetronarce californica
8	Q92035	Bungarus fasciatus
9	P38433	Caenorhabditis elegans
10	P23795	Bos taurus
11	O62763	Felis catus
12	P36196	Gallus gallus
13	Q869C3	Anopheles gambiae
14	Q27459	Caenorhabditis briggsae
15	Q86GC8	Culex pipiens
16	O42275	Electrophorus electricus
17	Q27677	Leptinotarsa decemlineata
18	P56161	Anopheles stephensi

2.2 Single Nucleotide Variant (SNV) Impact Predictions and Analysis of SNV Simulations

2.2.1 Introduction

The degree of conservation of a residue gives an indication of its functional significance and whether a mutation of the residue will is likely to destabilize the structure. Although it is not always the case that a single nucleotide that is more conserved correlates with higher functional importance, this is the general trend (Ng and Henikoff,

2006). In silico predictions of the pathogenicity of a SNV is widely used. dbNSFP is an exhaustive database containing SNV predictions for non-synonymous variations of the human genome. This database contains prediction scores from a wide range of prediction tools and algorithms and also conservation scores. It has been used in conjunction with machine learning to develop gene specific prediction scores (X. Liu et al., 2016). For the purposes of this study, dbNSFP was used to extract prediction scores for the SNV's that were analyzed through molecular dynamics.

The properties of the environment surrounding the protein should be taken into consideration when deciding on the likely consequences of an amino-acid substitution. The lipid and water content around the protein interacts with the residues on the surface of the protein. Membrane and soluble proteins differ in the amino-acids they prefer on their surface. Membrane proteins, surrounded by a lipid environment will prefer hydrophobic residues on the surface and hydrophilic in the interior. Soluble proteins prefer hydrophilic amino-acids on the surface and hydrophobic residues towards the center of the protein (Betts and Russell, 2003).

2.2.2 Steps

- Introduce SNV's through modelling using MODELLER v9.19 (described in chapter 2)
- Simulate two protein variants using molecular dynamics (described in chapter 5: Molecular Dynamics)
- Analyze the results delivered by the molecular dynamics trajectories to make inferences about the consequences of the amino-acid substitution.

2.2.3 Amino Acid Properties of SNV's Under Investigation

The first substitution discussed here is the Proline to Leucine substitution. Leucine is hydrophobic and prefers the alpha helical secondary structure to beta sheet structure. In this enzyme, Leucine 216 (247 in canonical protein sequence) is located in a beta sheet in the model.

Proline is restricted in the conformations it can take as its side chain is bound to its backbone twice making it rigid. It is a small amino acid. It does not seem likely that the Leucine will be a good substitute for Proline. While proline is hydrophobic like Leucine, it is unable to take on many conformations that other amino acids are able to.

The second variant examined is a substitution of threonine with serine. These are both small amino acids and the only difference between the two is that serine has a hydrogen group where threonine has a methyl group (Betts and Russell, 2003). From this information it is predicted that the T229S variant (198 in model) will have a significant influence on the stability or dynamics of the protein.

2.3 High Throughput Screening

2.3.1 Introduction

High throughput screening refers to the process of docking a library of potential inhibitors to a target protein. This is used often for the identification of drug leads. Various software exists for docking and their performance varies depending on which complex is being docked. The docking tool used in this study was Autodock VINA, which uses a genetic algorithm approach to simulate natural genetic variation (Trott and Olson, 2009). Autodock tools was used to identify rotatable bonds and to generate tautomeric states for each molecule in the library. Blind docking was performed to chain A of the rhAChE structure where missing loops were introduced through modeling. This type of docking does not force the molecule to bind to the active site, but scans the entire space for areas that are conducive to good binding of the molecule. This is more similar to natural conditions than targeting the active site only.

2.3.2 South African Natural Compound Database (SANCDB)

The South African Natural Compounds Database (SANCDB) contains 623 natural compounds (Hatherley et al., 2015). This database was compiled and is maintained by Rhodes University, Grahamstown, South Africa. Many compounds in this database

are large which decreases the likelihood that they may pass through the blood brain barrier (BBB) and increases the likelihood that they contain qualities that prevent them from being utilized as drugs. However substructures that show high binding affinity can be used as a template to search other databases for potential inhibitors.

2.3.3 ZINC15 Subset

A subset of 5105 compounds retrieved from the ZINC15 database was also screened against the acetylcholinesterase enzyme. ZINC15 is a database of commercially available compounds for screening (Sterling and Irwin, 2015). This subset contains smaller molecules that can pass through the blood brain barrier and do not contain any violations of Lipinski's rule of 5. Each set of compounds was docked blindly to chain A of the crystal structure. 10 compounds from the results were selected. Targeted docking was performed of these compounds to chain B of the dimer to find coordinates to simulate the dimer with a molecule in each chain. The docking was targeted to the active site of the second chain.

Table 2.2: AUTODOCK VINA parameters

experiment	**box size**	**exhaustiveness**
blind docking	60 angstroms cubed	124
targeted docking	25 angstroms cubed	64

The center coordinates used when docking to chain A of the homodimer was: x: 5.356, y: -55.309, z: -30.815

The center coordinates used when docking to chain B of the homodimer was: x: -2.922, y: -40.018, z: 30.861

2.3.4 South African Natural Compounds Database (SANCDB)

623 compounds from the South African Natural Compounds Database (Hatherley et al., 2015) were docked to the model of the AChE monomer.

Ligand Preparation The SANCDB compounds were prepared for docking using Autodock tools to generate torsional angles and charges for the molecules and to assign Autodock atom types to each atom. Hydrogens were added to the molecules. The receptor was prepared using Autodock Tools. The exhaustiveness was set to 124 and the number of cpu's to 4. The box size was set to 47.25 angstroms cubed. This was later extended to 60 angstroms cubed in a second docking experiment to verify that the box was big enough to provide accurate blind docking results. Energy range was kept at 4.

2.3.5 ZINC15 Subset

The ZINC15 subset was the result of sorting through a larger compilation of compounds from the ZINC15 database. The molecules were filtered according to their ability to pass through the blood brain barrier, their molecular mass, and their violations of Lipinski's rule of 5. This molecule subset is ideal for screening to an enzyme that is located at cholinergic synapses in the brain.

Ligand Preparation The ZINC15 compounds were prepared in the same fashion as the SANCDB compounds and the same receptor was used. The center coordinates for the search space was identical to the SANCDB screening experiment. The box size was 60 angstroms cubed and the exhaustiveness was set to 124. The rest of the parameters were identical to the SANCDB docking experiment.

2.4 Molecular Dynamics Simulations

2.4.1 Introduction

Molecular dynamics can be used to analyze the influence of an amino-acid change on the dynamic motion of a protein. The amino-acid may propagate change throughout the molecule during the simulation. This type of simulation can be used to determine whether a protein structure will be stable after introduction of a SNV.
Molecular Dynamics is performed by calculating the instantaneous forces between the atoms in a system. The displacement that these forces produce is taken into account and the new instantaneous forces are calculated (Lahti et al., 2012). The system consists of a solvent, usually water molecules, and the protein or protein-ligand complex. Varying structures or chemicals can be added to the system where the goal is most often to replicate natural conditions.

While the accuracy of virtual screening is restricted by a rigid receptor, molecular dynamics accounts for the flexibility of the protein and the protein side-chains. Molecular dynamics was used to investigate the conformational changes of the protein, using the crystal structure as a starting structure, and how this conformational change varies between variants of the protein. Different ligands, if able to bind to the protein, may induce different conformational changes in the protein. In the case of protein-ligand complexes, molecular dynamics explores the ability of a ligand to stay bound to the protein. Various analysis can be done on the trajectories that molecular dynamic simulations produce. This includes MM-PBSA, and free binding energy calculation, as well as network analysis that includes betweenness centrality and average shortest distance (L). MM-PBSA (molecular mechanics Poisson Boltzmann Surface Area) is a popular method to calculate free energies of ligand-protein complexes. This augments molecular dynamic simulations that aim to identify ligands with high affinity. Residue specific contributions towards binding energy was calculated using the g_mmpbsa package which is compatible with GROMACS. (Kumari et al., 2014)

2.4.2 Molecular Dynamics System Overview

While the biologically active form of AChE at cholinergic synapses is a tetramer tethered to a membrane, the simulations were conducted using a homodimer that was not membrane-bound. The enzyme's connection with the PRIMA (proline rich membrane anchor) is not taken into account. The C-terminal collagen tail of the enzyme is also not present during the simulation. Two monomers were simulated, as the interaction between the monomers can then be taken into account to some extent.

2.4.3 Force Field Selection

Investigation of the consistency and error of different force fields in determining the binding affinity of molecules to AChE was done (Tam et al., 2018). The best combination of force field and water was the GROMOS 43a1 force field in combination with the SPC/E water model. This combination delivered a correlation of -0.88 with experimental data as well as the smallest error. This is a united-atom force field. The force field that was selected for our study was the GROMOS 96 54a7 force field which is similarly a united-atom force field albeit not the exact force field tested in that study. The force fields are fundamentally the same as they are both united-atom force field. The latter is an updated version of the force field that performed the best. Since force field updates are aimed to improve accuracy and increase performance it was deduced that the updated force field was a good option to choose. The topology generating tool was compatible with the GROMOS 96 54a7 force field. Since the tool was specialized to create topology files for this force field it was used instead of the GROMOS 43a1 force field. The ability of a united-atom force field to simulate the interaction of AChE with small molecules was shown to be on par and better than all-atom force field (Tam et al., 2018). It is a good choice as it uses less computational time with no negative effects on accuracy. In addition, ATB (automated topology builder) provides a good pipeline to generate topologies of ligands that are compatible with the GROMOS 96 54a7 force field. This force field allows for longer simulations due to less computation required as less interactions have to be calculated. This force field is known as a course-grained force field and methyl groups, which consist of a

carbon atom bound to three hydrogens, are merged and treated as one interaction site by the force field.

2.4.4 Topology Generation

Topological information is required for the ligand in order to simulate the ligand molecule along with its target enzyme. GROMOS, along with most other force fields, do not include parameters for non-protein atoms. The topology is arguably the most important factor determining the accuracy of the simulation as this determines how the interaction between the ligand and protein is calculated. With the objective of analyzing the ability of the ligand to stay within the active site, this interaction is integral. This topology can be generated using various methods and tools. The topology file includes information such as the charge of atoms in the molecule as well as bonds and angles between the atoms. ATB (automated topology builder) was used to generate topologies for the ligands (Malde et al., 2011).

ATB provides a web server where molecules may be uploaded for topology generation on the condition that the topology is available to other users. ATB is able to generate topology files for molecules that are compatible with the GROMOS 96 54a7 force field and GROMACS. ATB is also able to provide topologies for other force fields in the GROMOS family. To use the webserver, a molecule must be protonated and uploaded to the server. Any inconsistencies regarding the charge or protonation of the molecule results in discontinuation of the topology generation. The results produced include a topology (.itp) file and the coordinates in pdb format. Both all-atom and course-grained versions of these results are available. The course grained results were downloaded as the GROMOS 96 force field requires this format. The pdb coordinate file was then converted to .gro format using the 'editconf' functionality of GROMACS. The server selects which level of accuracy to use by examining the size of the molecule. Molecules of over 50 atoms are taken only to an 'initial guess' topology. Molecules of below 50 atoms are taken to geometry optimized topology generation using quantum mechanics. The SANCDB compounds were all above 50 atoms and the ZINC compounds all below 50, thus the ZINC set was provided a more accurate topology for the

simulation. Hydrogens were added using either Autodock tools, Babel or Discovery studio. In many cases the hydrogens were not added correctly by Autodock tools. Hydrogens were added instead of including double bonds where appropriate. Adding hydrogens using babel fixed this issue except for one case where only Discovery studio added hydrogens correctly.

2.4.5 Molecular Dynamics System Setup

The protein dimer coordinates were converted to .gro format using the pdb2gmx function of GROMACS. This generates a topology file where the information for the ligand molecules can be added. This information includes directions to each ligand's topology file and position restraint file. The protein dimer consists of two chains, each made up of 539 amino-acids. This amounts to a total of 1078 amino-acids. The coordinate information for the ligands for each chain were added to the protein coordinates, adjusting the atom count by adding the number of atoms present in the two ligands, to the number of atoms present in the protein dimer. Periodic boundary conditions were prepared for the system. A triclinic shape was selected and the distance between the protein dimer and the side of the box was set to 1.25nm. The system was solvated using the SPC/E water model which is a 3 point water model. Ions were added to neutralize the system. 16 Sodium atoms were added, replacing 16 solvent molecules. This neutralized the -16 net charge in the system. Each system was then minimized using an indefinite amount of steps. The minimization was set to terminate when the free energy value converges. Energy minimization ensures that steric clashes are cleared.

2.4.6 System Equilibration

In systems where ligands were present, the ligands were restrained for the NVT equilibration step. V-rescale temperature coupling was used and this was specified in the nvt.mdp file. An example of this file has been included as Appendix B. NVT refers to a constant number of particles, volume and temperature and this equilibration is also called an isothermal-isochoric ensemble. The NVT equilibration for each simulation

was performed for 50000 steps which is equivalent to 100ps. Steps are set to occur every 2 femtoseconds (fs) which results in 100ps when 50000 steps are simulated (2 * 50000). Two temperature coupling groups were created. The protein dimer and two ligand molecules were couples as a group. The solvent ion molecules were couples as another group. Pressure coupling was not activated during NVT equilibration. The temperature was equilibrated to 300 kelvin. Pressure equilibration was performed after temperature equilibration using the Berendsen isotropic ensemble. The pressure equilibration was run for 50000 steps, the same length as the temperature equilibration. Ligands were not restrained for the NPT step where pressure is equilibrated. The system was set up on a local machine using GROMACS v16.4 and uploaded to chpc's lengau cluster. The NVT and NPT steps as well as the production run was performed using GROMACS v16.1 on the cluster. For 10 nanosecond simulations, the SMP nodes was used, which allowed each simulation to use 24 cpu's. 100ns simulations or longer were performed using 10 nodes, each using 24 cpu's. This means that 240 cpu's were used to simulate the 100ns simulation.

2.4.7 g_mmPBSA

MM-PBSA calculations are used to break down the ligand binding energy into different components. MM-PBSA calculations may use a snapshot of a trajectory as input. In the case of the territremB complex simulation, the time frame from 95ns to 100ns at the end of the simulation was sampled for one set of calculations and the time from 90ns to 100ns for another round of calculations. Both 5ns trajectory samples and 10ns trajectory samples were used for separate MM-PBSA calculations. Each of the 5 tested protein-ligand complexes were therefore calculated twice with different trajectory lengths as input. The results are presented in tables 3.9 and 3.10.
The free binding energy is the average molecular mechanic potential energy, the entropic contribution and the energy of solvation. The sum of these three terms will give the binding free energy of the complex (Kollman et al., 2000). This can be summed

up in the following formula:

$$G_x = \langle E_{MM} \rangle - TS + \langle G_{solvation} \rangle \tag{1}$$

Formula 1 is used to calculate the free energy of individual items in the system ie. complex, ligand and protein. Where x refers to the item of the system and TS is the entropy. The molecular mechanics potential energy is given by E_{MM} and this is assumed to be in the absence of external forces. This equation can be used to calculate the complex, protein and ligand free energies, each separately. This can then be used to determine the free binding energy from interaction between the protein and ligand. This is given by the formula:

$$\Delta G_{binding} = G_{complex} - (G_{protein} + G_{ligand}) \tag{2}$$

The solvation free energy can be split into polar and nonpolar contributions. The nonpolar contribution to the solvation free energy consists of van der Waals forces and forces generated by cavity formation. The nonpolar contribution can be calculated using a range of methods. The method used here was the SASA (Solvent Accessible Surface Area) model. This method calculates only the energy of cavity formation and does not take into the van der Waals forces of attraction and repulsion. The SASA model assumes that the solvent accessible surface area is dependent linearly on the nonpolar solvation free energy (Kumari et al., 2014).

To determine whether the alternate subunit contributes towards interaction with the ligand eg. chain B influencing ligand bound to active site of chain A, was calculated taking both chains of the dimer into account. Once it was confirmed that only the one subunit is involved in the interaction with the ligand the calculations could be performed using only the chain to which the ligand has bound. This speeds up the MM-PBSA calculation.

MM-PBSA calculations were performed for a total of 5 protein-ligand complexes. All of the calculations were performed using the SASA model and were conducted separately for each chain. The molecules were selected based on stable RMSD values. The last 5

ns of their 100ns simulations were used for the calculation except for ZINC945 where the 5ns between 55 and 60ns were sampled.

2.4.8 Principal Component Analysis (PCA)

PCA analysis of protein motions plots the vectors of the most relevant motions. The motion is condensed to one vector instead of three for the x, y and z components of the motion. This is essentially a simplified representation of the most prominent protein movements. This can be used to find out at what time during the molecular mechanic (MM) simulation important changes occur in the major motions of the protein. This would be useful if one wishes to determine the duration of a specific motion that occurs within the protein. The internal motions are revealed by PCA, which is harder to analyze since these motions are hidden by the surface residues.

Principle component analysis was done using Cartesian coordinate ensembles from MD simulations as input. The PCA uses eigenvectors of the motions that explain the highest amount of variance. Vectors have both direction and magnitude. MODE-TASK was used to generate PCA plots of the 100ns trajectories from simulations of wild type dimers as well as the variants T229S and P247L (Ross et al., 2018). The five components that explain the most of the variance in the Cartesian coordinates is generated. The top 2 components are plotted and time is represented by a color map. PC1 is plotted on the x-axis and PC2 is plotted on the y-axis.

The xyz dimensions of the principle component are condensed into one dimension. This changes the 3D vector to a one dimensional vector that best explains the motion. This one dimensional vector (PC1) becomes the x-axis and the other (PC2) becomes the y-axis. This elucidates the relationships between these important motions.

2.4.9 Network Analysis: Betweeness Centrality (BC) and Average Shortest Distance (L)

Network analysis was conducted using scripts from the MD-TASK collection (David K Brown et al., 2017). The trajectory of the GROMACS simulation was reduced to the c-beta and c-alpha atoms. This was achieved using VMD - Visual Molecular

Dynamics (Humphrey et al., 1996). Betweenness centrality is the importance of a residue in terms of facilitating 'communication' throughout the protein. A high value indicates that the residue facilitates the transmission of force between distant residues. Long distance effects like this appear when molecules bind to allosteric sites as these sites modulate the conformation of residues at a distance from the binding molecule. The BC of a node is calculated by calculating the number of shortest paths from all nodes to all others that goes through the original node. Average BC was calculated using the dimer structure. The molecular dynamics simulation trajectory of the dimer simulation was used as input to MD-TASK. Residues residing at the interface between the subunits of the dimer should have high values of centrality. If interactions within a single subunit are of interest, calculating betweenness centrality using only the one subunit is more informative.

Average betweenness centrality and average shortest distance are inversely correlated. Average shortest distance of a residue is the sum of the shortest paths to that residue devided by the total residues minus one (David K Brown et al., 2017).

Chapter 3
Results and Discussion

3.1 Results Overview

The results are separated into four parts. Modelling and sequence alignment are presented and discussed first followed by the SNV impact predictions. Thereafter the high throughput screening is examined and then the molecular dynamics of protein-ligand complexes. Modelling is discussed first as this is required for the other methods. Molecular dynamics was used for different purposes which include assessing the impact of introduced SNP's of dynamics and calculating the binding affinity of potential therapeutic inhibitors.

3.2 Modelling and Sequence Alignment

The modelling process resulted in a complete homodimer structure of rhAChE, suitable for use in molecular dynamic simulations. Two measures were used to validate the accuracy of the model. These were verify-3D and the z-DOPE score. As only a small part of the structure was being modeled, there was not a large decrease in z-DOPE score between the template and the model. The validation scores for each respective model, indicated that the structure was of high accuracy. Investigating the effects of SNV's on protein dynamics is a sensitive problem and all other variance should be eliminated, so that any change will be a result of the SNV and not another variable.

Table 3.1: Protein model evaluation scores

model version	**z-DOPE score**	**verify-3D score**
wild-type(wt)	-2.037	97.44
T229S	-2.028	97.44
P247L	-2.030	97.44

The z-DOPE score indicates how native-like the model is. This score varied by a very small margin between the final model for each variant. A disruptive variant which

does not fit well in the protein, will be placed in positions that are less biologically likely or viable. Each variant and wild-type model delivered identical Verify-3D scores

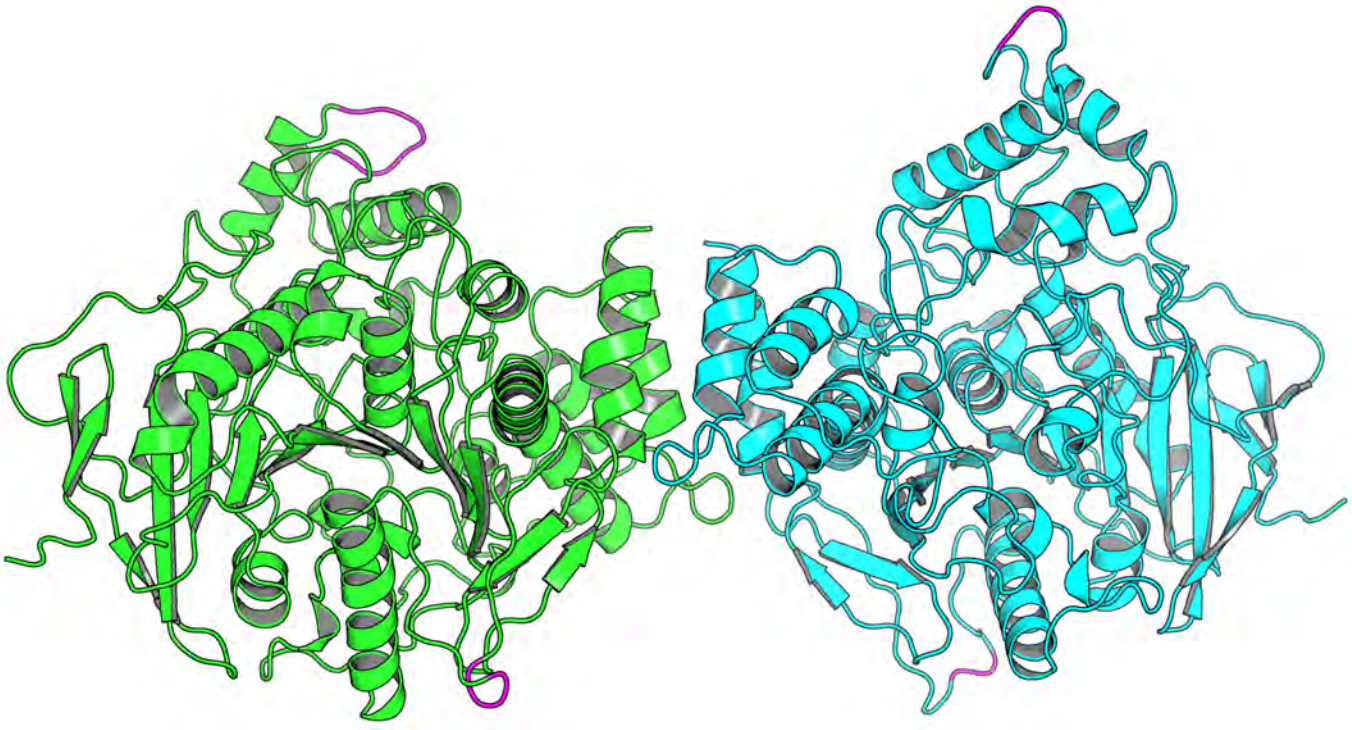
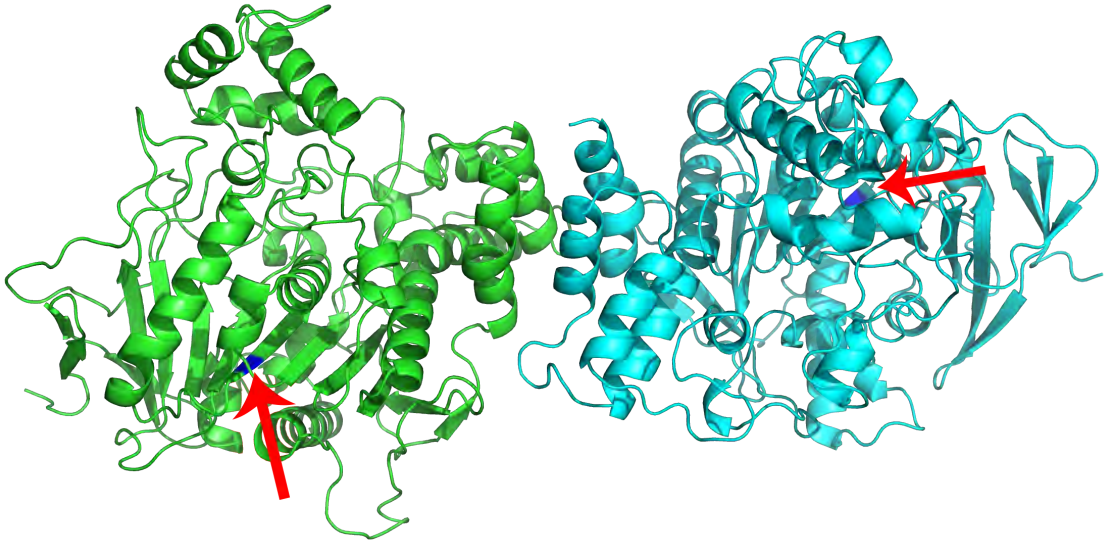


Figure 3.1: Figure 3.1 was created using PyMol. The loop regions that were inserted using MODELLER are indicated in magenta. Chain A, coloured in green, had more residues missing than chain B (cyan).

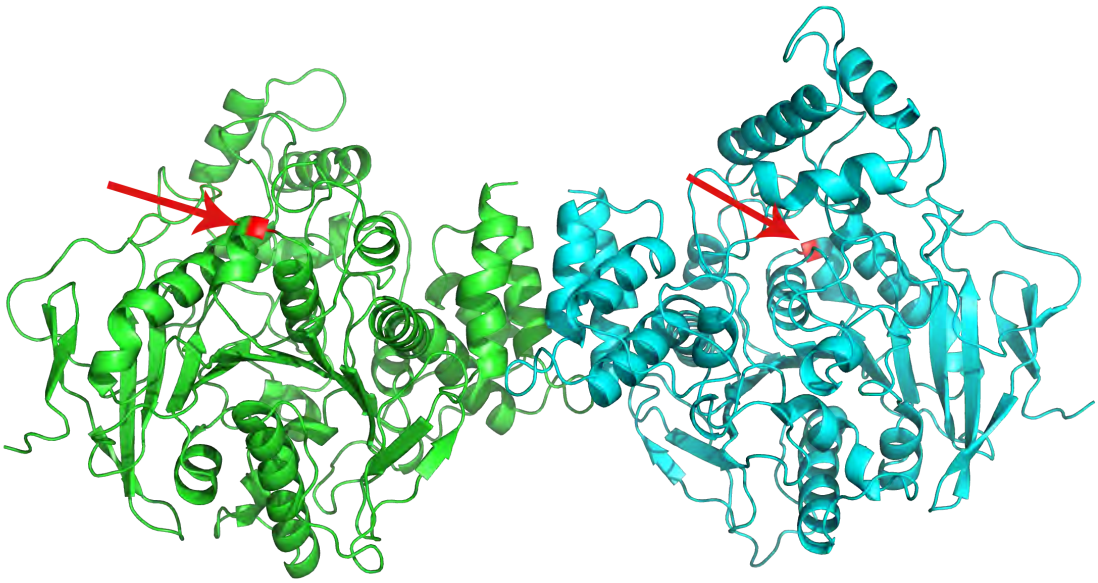
A total of 9 missing residues were inserted into chain A. These made a loop of 6 and 3 residues respectively. A total of 4 residues were missing from chain B. This consisted of gaps of 2 residues each. Mostly the same residues are missing from both chains, with chain B having residues present that were absent from chain A. The missing loops are relatively short in length and given the z-Dope scores that were generated after including the loops, the structure should perform accurately enough in molecular dynamic simulations. The missing residues have no effect on the Autodock Vina experiment as the residues are not near the active site, which is the most important region in the docking experiment.

Loops that are less than 10 residues in length can be modelled fairly accurately using Modeller software. The larger the gap the less accurate it will be as there are less nearby residues to base the position off of. Given that all of the gaps in this pdb structure were less than 10 residues in length, the model is sufficient to use for molec-

ular dynamics and docking simulations. The z-Dope scores in Table 3.1 above are all under -2, indicating that they are all high quality.



(a)



(b)

Figure 3.2: Models of variants. Figure 3.2 was rendered using PyMol. The position of the substituted amino-acid in each dimer is indicated in color. (a) The T229S variant (threonine to serine) where the mutated residue is coloured in blue. This variant falls within a beta-sheet secondary structure in the amino-acid sequence. (b) The P247L variant (proline to leucine) where the mutated residue is coloured in red. This variant falls within a helical secondary structure in the amino-acid sequence. Chain A of the homodimer is coloured green and chain B in cyan.

The secondary structure that the substituted nucleotide forms part of is significant as certain amino acids prefer to for part of specific secondary structures. If the variants fell merely within a loop structure the substitution would likely not create a large impact. The Leucine nucleotide in the case of the P247L variant, favors the alpha

helix conformation.

The important function that acetylcholinesterase plays in the human brain has resulted in it being very well conserved. There are not many SNP's that could impact the protein negatively that would not die out. The two variants were selected as prediction scores indicated that they had a good probability of having a deleterious effect. This in combination with being closest significant SNP's to the active site, led to their selection.

Models of the variants were required to analyze the influence of a variant on protein dynamics. Separate models were created for each variant under investigation. The model could then be used in molecular dynamics simulations to analyze the stability of the protein. The location of the residue change can be observed to see whether it is close to the active site or whether it will affect the active site and by extension the binding affinity of ligands to that active site. Other factors to consider is whether the stability of the structure is altered and whether the interaction of the protein with other proteins is altered.

The inserted SNP's both occur towards the center of the enzyme. They are however not adjacent to the catalytic site of the protein or part of the catalytic triad. It is thus not likely to interfere with the catalytic process of the enzyme but could create a shift in the dynamics of the protein.

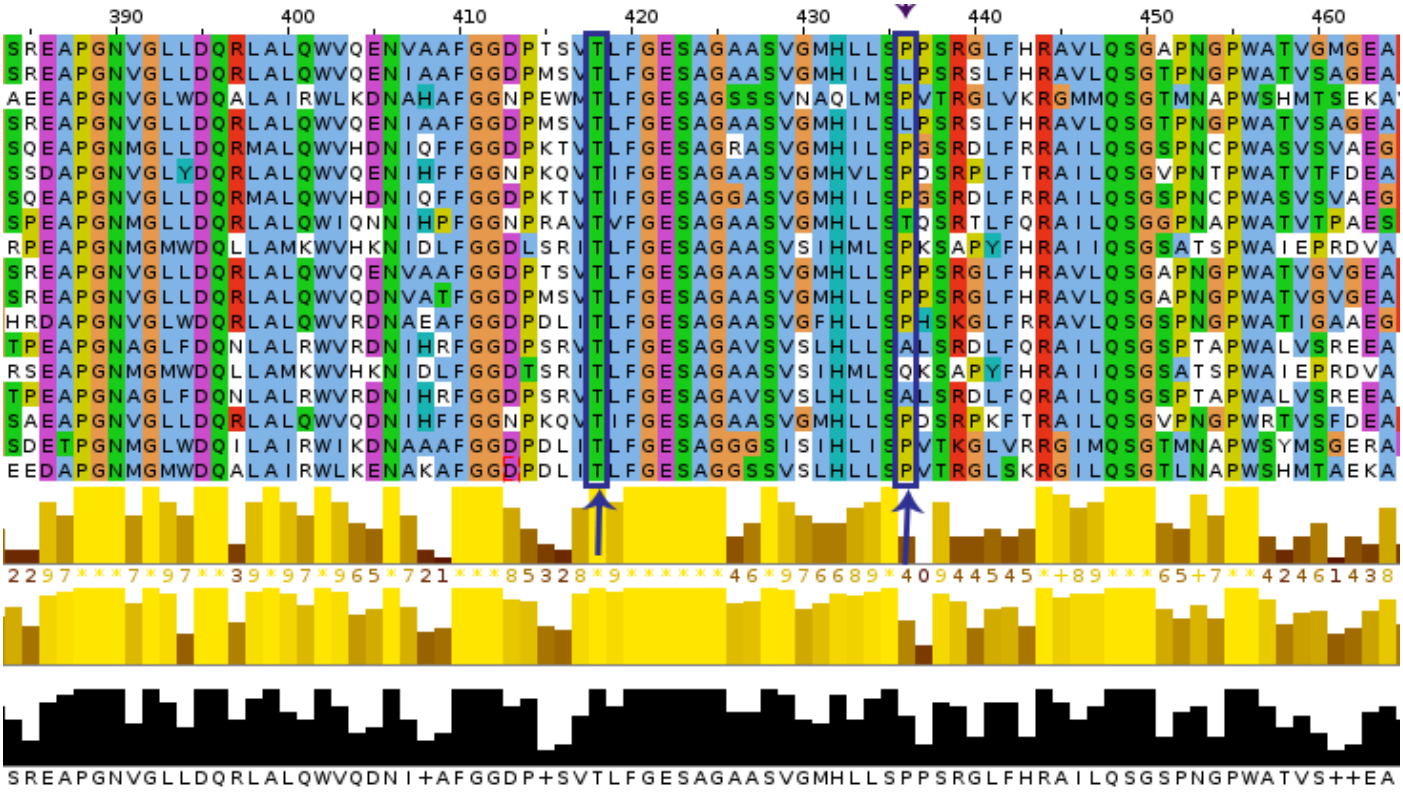


Figure 3.3: Multiple sequence alignment of 18 acetylcholinesterase sequences. Two amino acid locations are of interest and are surrounded by a dark box. This figure was produced using Jalview. The sequence below the black histogram represents the consensus sequence. The black histogram above the consensus sequence indicates the confidence of the consensus call.

The two residues for which the variants were investigated are moderately to highly conserved. The Threonine which is replaced by a Serine residue in the case of a T229S variant is conserved across all 18 sequences that were used for the alignment. This indicates a high degree of conservation. The Proline which is replaced by leucine in the P247L variant appears in 12 of the 18 canonical protein sequences. In two sequences this position is held by a leucine residue, as in the case of the P247L variant. Assuming that a residue that is more conserved is more functionally significant it can be predicted that the T229S variant will result in a larger conformational variance. This was however not found when analyzing the molecular dynamic simulation, in which the P247L variant resulted in a larger predicted change in the dynamics of the protein relative to the wild-type protein.

Table 3.2: MSA: The organisms are listed in order from top to bottom, corresponding to the order used in figure 3.3

no.	UniProt ID	Organism
1	P22303	Homo sapiens
2	P21836	Mus Musculus
3	P07140	Drosophila melanogaster
4	P37136	Rattus norvegicus
5	P07692	Torpedo marmorata
6	Q9DDE3	Danio rerio
7	P04058	Tetronarce californica
8	Q92035	Bungarus fasciatus
9	P38433	Caenorhabditis elegans
10	P23795	Bos taurus
11	O62763	Felis catus
12	P36196	Gallus gallus
13	Q869C3	Anopheles gambiae
14	Q27459	Caenorhabditis briggsae
15	Q86GC8	Culex pipiens
16	O42275	Electrophorus electricus
17	Q27677	Leptinotarsa decemlineata
18	P56161	Anopheles stephensi

3.3 Single Nucleotide Variant (SNV) Effect Predictions and Analysis of SNV Simulations

Predicting the effect of the two variants that were chosen revealed the following. The scores evaluated are from tools that give a consensus of many other prediction tools.

Table 3.3: Variant effect predictions. Lower score indates a lower chance of being deleterious

SNV	rs number	ref:aa	alt:aa	position in model	MetaSVM	MetaLR
P247L	rs143875983	Proline (P)	Leucine (L)	216	-0.3327	0.3535
T229S	rs202183011	Threonine (T)	Serine (S)	198	0.5973	0.6979

Table contains Amino acid positions and pathogenicity prediction scores from two prediction tools. A lower score indicates a lower probability of being deleterious for

both prediction models. MetaSVM was developed using support vector machines (SVM) and incorporates scores from 9 independent prediction tools as well as the maximum frequency observed in the 1000 genomes project. MetaLR is a Logistic regression (LR) prediction score and was developed using the same data as MetaSVM (Dong et al., 2015).

The scores predict that the T229S variant has a larger probability of being deleterious than the P247L variant. In contradiction to this prediction score, the porcupine plots that were generated for the motions of each variant suggests that the P247L variant will lead to a larger change in protein dynamics than the T229S variant. The shearing motion is disrupted to a greater extent by the P247L variant than the T229S variant. If this motion is important for the functioning of the protein, it is predicted that the variant that disrupts this motion the most will lead to a malfunctioning version of the protein. Taking this into consideration, the observed change in motion is not in agreement with the prediction scores. A variation that prevents the subunits from obstructing each other periodically could lead to stronger drug effects since more of the drug molecule will be able to enter the active site.

3.4 Molecular Dynamic Simulation of Models

The trajectories from molecular dynamics simulations can be used to extract various information. The Information was used to create a porcupine plot as this gives an overview of the magnitude and direction of the protein during the simulation. The root mean square deviation (RMSD) was used to create graphs comparing this metric between different versions of the model. RMSF was compared in the same way. The following figures present these molecular dynamics results.

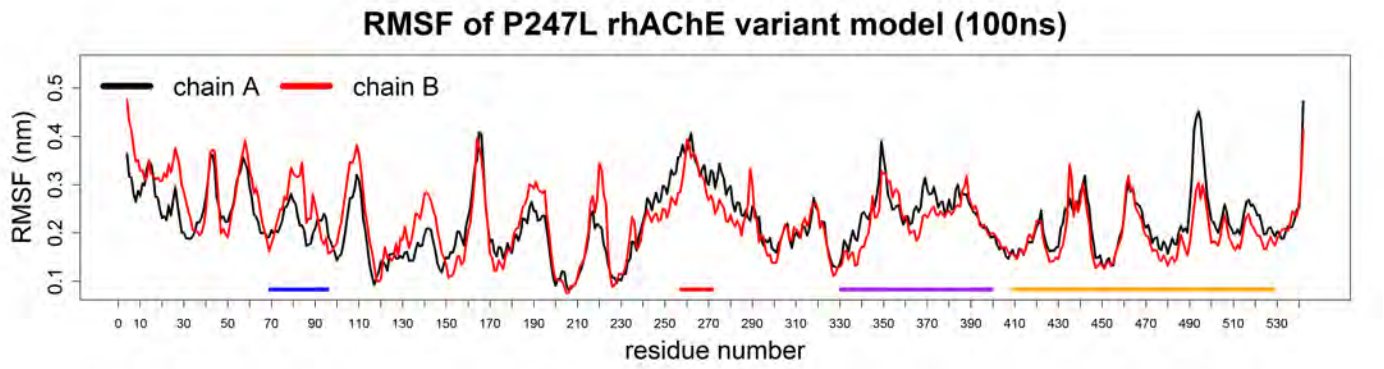


Figure 3.4: RMSF by chain for the P247L variant MD simulation. Areas between disulfide bonds are: blue (69:96), red (257:272), orange (409:529). The region in purple is a region that displays large movements and is next to the active site region.

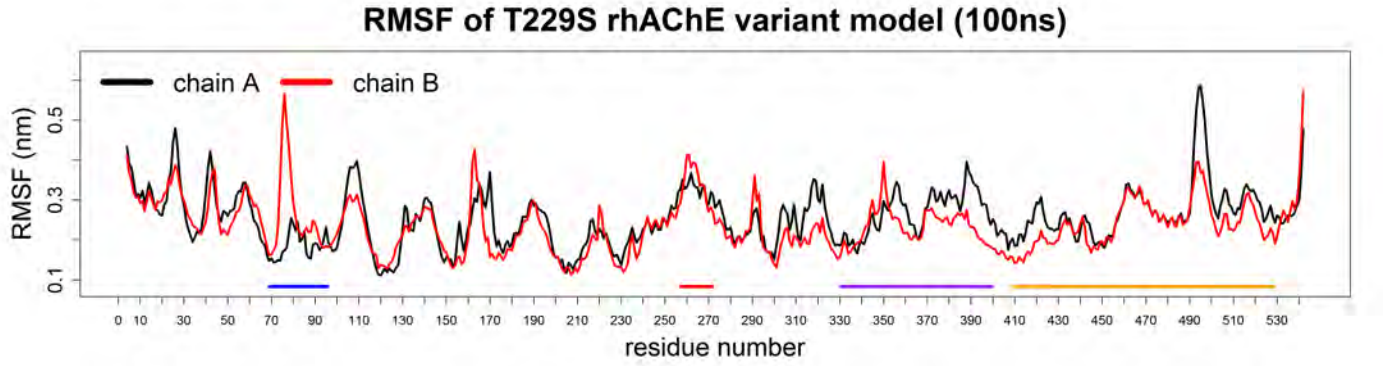
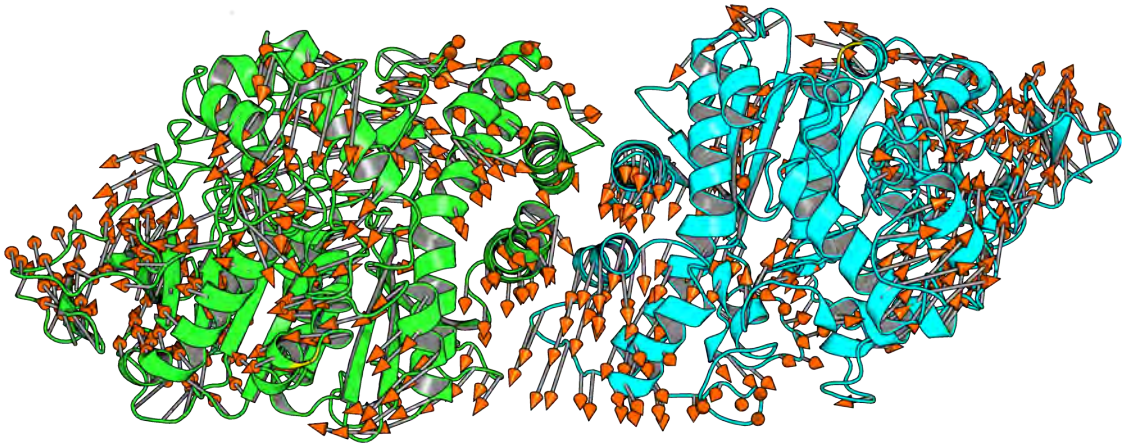
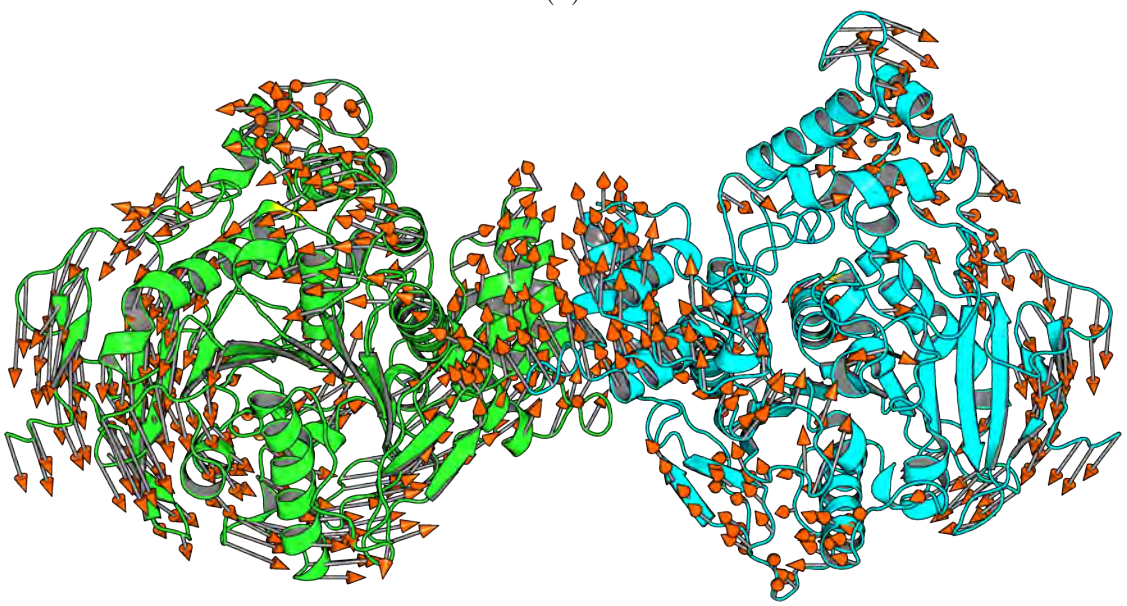


Figure 3.5: RMSF by chain for the T229S model MD simulation. Areas between disulfide bonds are: blue (69:96), red (257:272), orange (409:529). The region in purple is a region that displays large movements and is next to the active site region.

The RMSF graphs in Figures 3.4 and 3.5 were compiled using R. RMSF for each residue is averaged over the course of a 100ns molecular dynamics trajectory. Variant T229S is named according to the canonical acetylcholinesterase human sequence. The location of this varied residue in the model is 198.



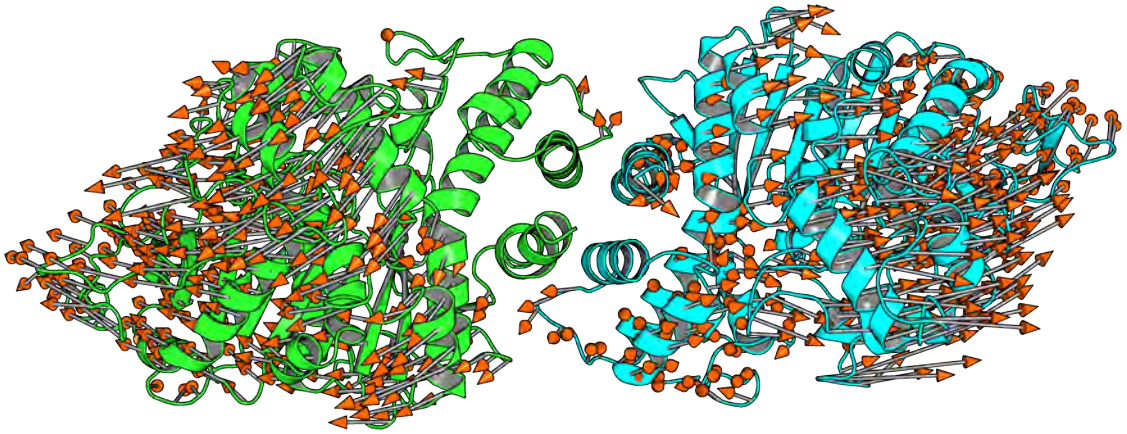
(a)



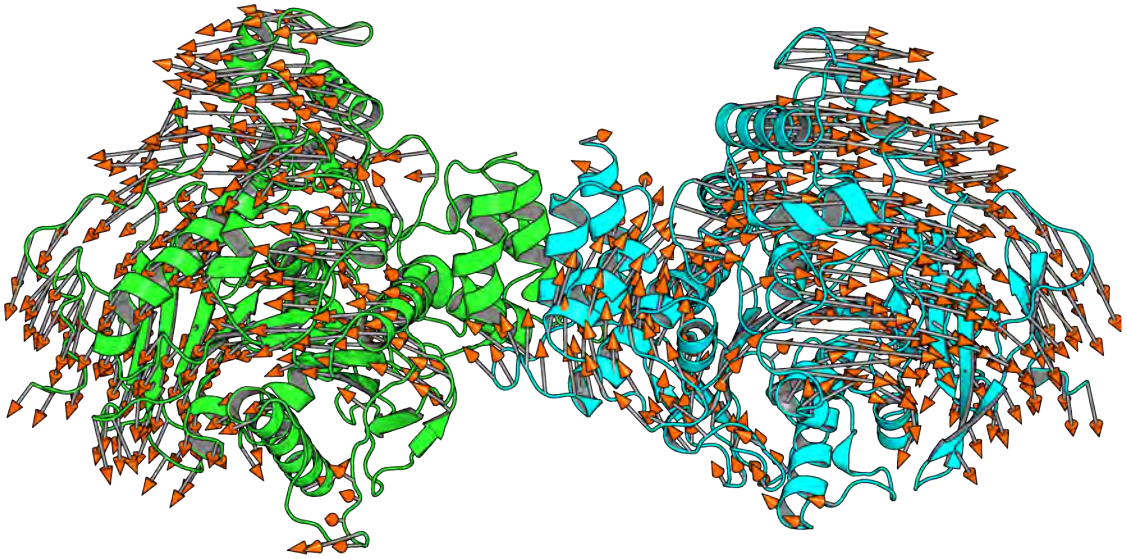
(b)

Figure 3.6: Porcupine plot of P247L variant model. Both images are of the same enzyme, from different angles to better analyze the vectors. (a) - topview. (b) - sideview. The images were rendered using PyMol and the modevectors script. This plot shows the magnitude and direction of the displacement of residues during the simulation.

The motion of the protein is disrupted as seen by comparing to the porcupine plot of the wild type protein dimer in Figure 3.15. The change in the residue motion is most noticeable in the first image (a) Wild type apo protein. This view reveals distinct movement of subunits in opposite directions. Here, motion has been disrupted and the residues do not show uniform displacement as in the wild-type protein.



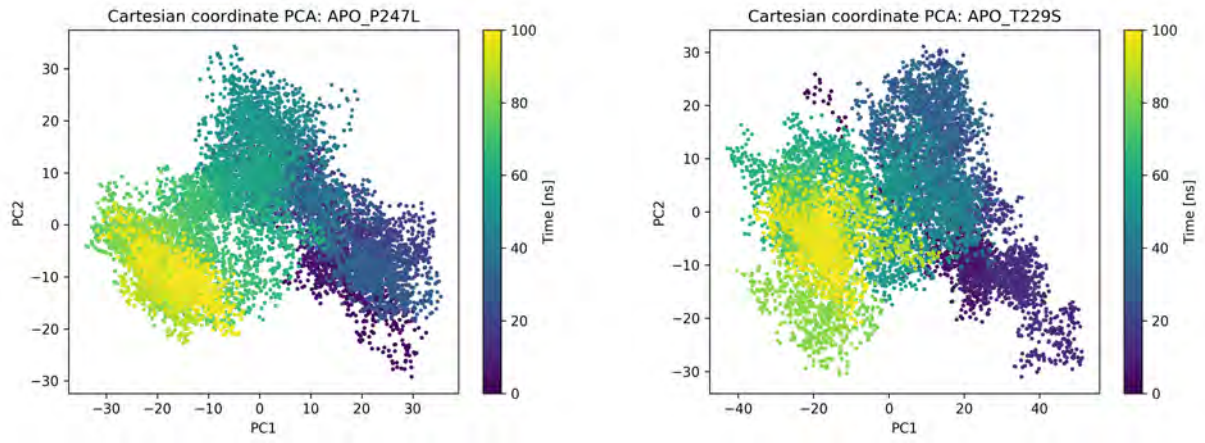
(a)



(b)

Figure 3.7: Porcupine plot for the T229S variant simulation. The porcupine plots were generated using data from a 100ns simulation and the modevectors script which is compatible with PyMol. (a) - topview. (b) - sideview.

The plot of the simulation of the T229S variant appears similar to the wild-type version (Figure 3.16), unlike the porcupine plot of the P247L variant. Threonine is similar to Serine, they are both small amino acids. Similar amino acids such as these can normally substitute for one another without as much harm as very dissimilar amino acids.



(a) (b)

Figure 3.8: P247L and T229S variant PCA. (a) - P247L variant PCA. (b) - T229S variant PCA. created using ensemble of atom coordinates in MODE-TASK (Ross et al., 2018)

For the PCA plot for P247L 100ns trajectory, PC1 explains 34.34 % of variance and PC2 explains 16.13 % of variance. For the PCA plot for (b) T229S PCA, PCA1 accounted for 50.00 % of variance and PC2 accounted for 15.00 % of variance. 100ns trajectory information from MD simulations were used as input to create ensemble of atom coordinates.

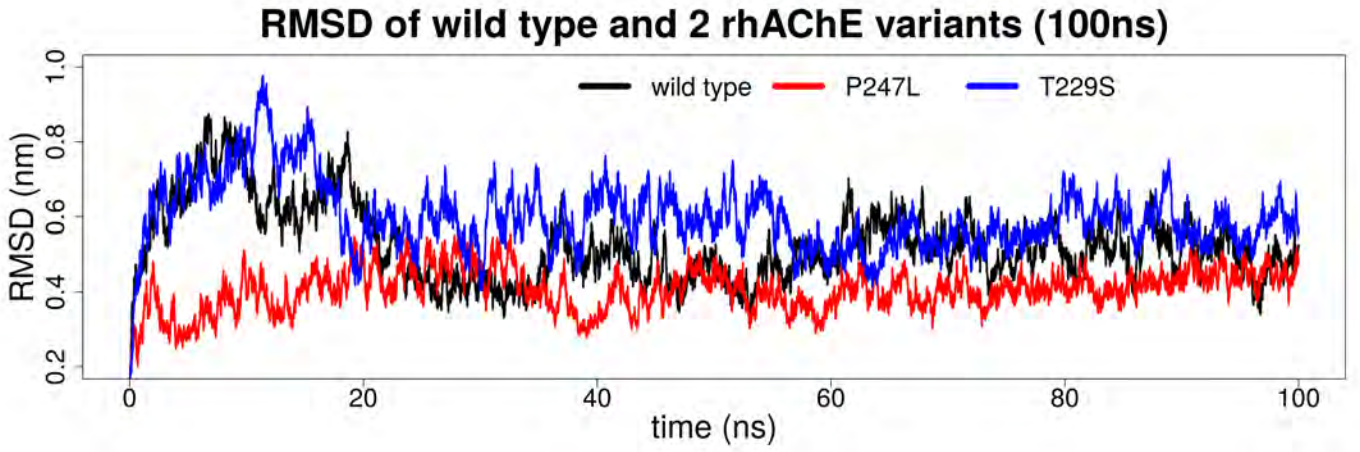


Figure 3.9: RMSD for variants of AChE

RMSD was calculated using the backbone of the protein for fitting. RMSD is close to converging at 100ns which indicates stabilization. The porcupine plot of the T229S SNV indicates similar direction of motion to the wild type but with greater magnitude. This is supported by the RMSD graph which shows greater fluctuation for the T229S variant than the wild type protein. Results from the porcupine plot thus agree with

the molecular dynamics simulation results regarding the level of fluctuation of the T229S variant compared to the wild type. The P247L variant (red line) delivered the lowest RMSF values, this correlates with the porcupine plot. The porcupine plot in figure 3.7 above has arrows pointing in scattered directions and at a lower density than the wild type.

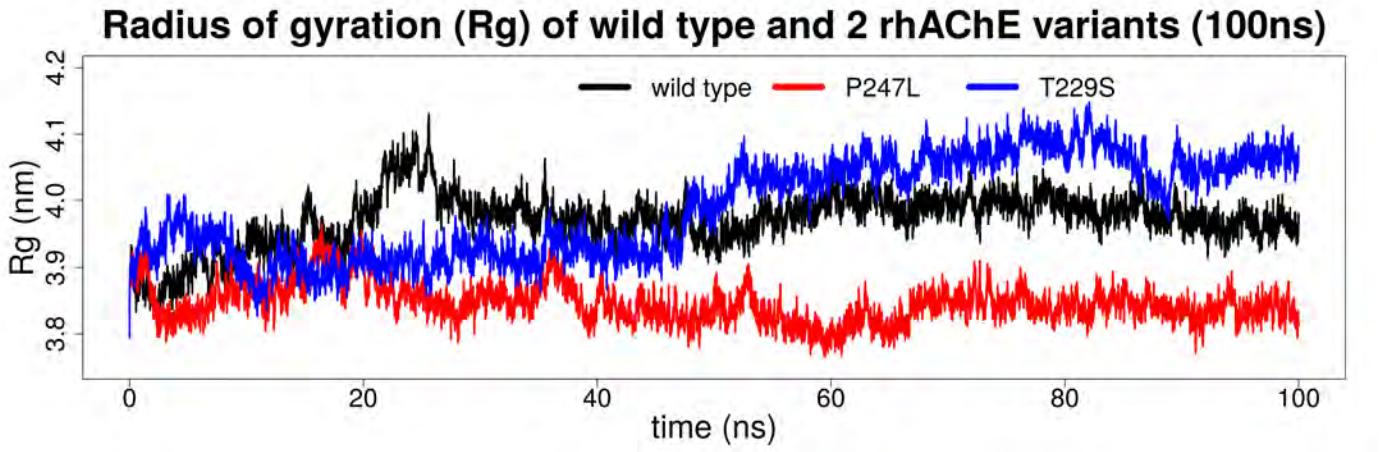


Figure 3.10: Radius of gyration (Rg) for wild type and AChE variants

Radius of gyration is an indication of the compactness of the protein. The two variants seem to have opposite effects on the Rg of the protein. While the P247L variant results in a more compact and static structure, the T229S variant results in a more dynamic and expanded conformation. The compactness induced by the P247L variant may prevent the substrate or ligands from binding to the active site as fluently as the other versions of the protein. Acetylcholinesterase occurs as a tetramer, bound by a collagen membrane anchor. In this configuration the four subunits move relative to each other in a shearing motion. This motion could be hindered by a substitution that compacts the enzyme. Based on this information the P247L variant is predicted to be more deleterious than the T229S variant in contrast with variant effect predictions in Table 3.2. This is merely a prediction and could guide further studies that wish to investigate this enzyme.

3.5 High Throughput Screening with Autodock VINA

3.5.1 Docking Validation

Docking validation is required to confirm that the docking analysis is functioning as intended. The most common method is re-docking. This is the process of removing a co-crystalized ligand from a structure and docking this molecule back to the structure. The molecule is docked to the same region and in the same pose as was originally observed in the crystal structure. The molecule used to re-dock was territremB and this was docked successfully back into the same position as in the crystal structure. Some minor variance is observed. The high throughput screening was continued using the same method used for validation.

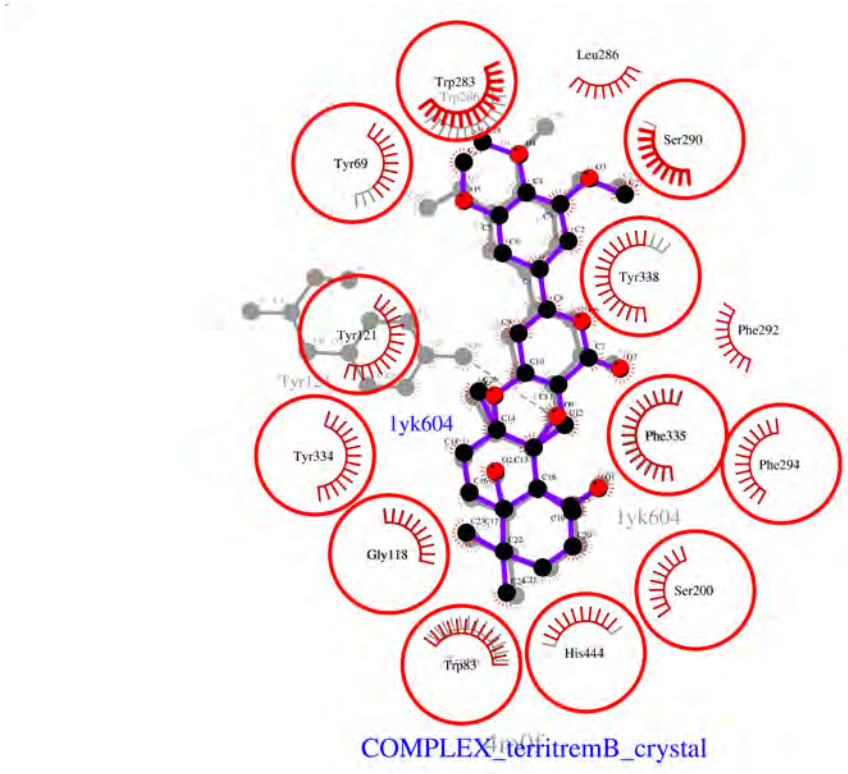


Figure 3.11: TerritremB docking validation results

Figure 3.11 indicates the overlap of the interaction plot for the original crystal structure PDB ID: 4m0f (greyed out image) with the interaction plot for the newly docked coordinates used for validation (coloured plot). The plot was generated using ligplot+ v1.4.5. Docking is a static process and the enzyme will need to take on a particular conformation to enable a compound to enter the active site. The structure is well suited to the docking of territrem B as it was crystalized together with the compound.

The docking validation confirms that the parameters are set correctly.

In the redocking experiment in figure 3.11 above, the newly redocked ligand does not indicate a hydrogen bond between the ligand and the enzyme. The positions are very similar and this discrepancy can be put down to a small distance shift which took the bond out of the cutoff range.

3.5.2 South African Natural Compounds Database

The results for the 623 compounds were sorted by their free binding energy and the distance of the compound from a central atom of the co-crystalized ligand. This was to separate compounds that did not bind to the active site from those that did. Sorting in this fashion allowed selecting compounds according to their binding energy and the proximity to the peripheral anionic site. 10 compounds were selected for further analysis using molecular dynamics. These compounds were selected based on their VINA binding scores, the number of hydrogen bonds predicted, the number of residues interacting that also interacted with territremB, the proximity to the peripheral anionic site.

Table 3.4: 10 selected molecules from the SANCDB set for molecular dynamics

compound	**initial binding energy (kcal/mol)**	**no. hydrogen bonds at pose**	**binding energy at 10ns (chain A)**	**binding energy at 10ns (chain B)**
S1	-12.4	5	-11.02	-11.34
S2	-10.5	2	-11.79	-8.50
S3	-11.3	2	-9.71	-9.36
S4	-11.0	4	-8.79	-8.97
S5	-10.5	3	-8.92	-7.37
S6	-10.6	4	-8.15	-7.63
S7	-11.9	0	-10.12	-10.64
S8	-10.7	2	-8.92	-9.77
S9	-10.5	2	-	-
S10	-10.9	4	-	-

Table 3.5: Identification information for the selected 8 SANCDB compounds

no.	SANC ID	compound SMILES
S1	SANC00347	CC2(C)C=Cc1cc5c(cc1O2)OC6Oc4cc3OC(C)(C)C=Cc3c(O)c4C(=O)C56O
S2	SANC00374	[H][C@@]35Cc1ccc(OC)c(O)c1c4c2OCOc2cc(CCN3C)c45
S3	SANC00152	C=C1CC=CC(C)(OC(C)=O)CCC2C(C(=O)C/C=C(C)/C)=COC(OC(C)=O)C12
S4	SANC00233	C[C@H](CCCC(C)C)[C@H]1CC[C@H]([C@]1(C)CCO)C2=CC(=O)C3=C[C@@H](CC[C@@]3(C2=O)C)O
S5	SANC00237	CC(=O)O[C@H]1O[C@@H](OC(C)=O)[C@@H]2CC[C@@H](C(=CC)/[C@H]12)C3(C)CCCC(C)(C)C3
S6	SANC00230	[H][C@@]12[C@H](OC(C)=O)[C@@H](OC(C)=O)C(C)=C(CC/C(C)=C/CO)C1(C)C[C@@H](OC(C)=O)CC2(C)C
S7	SANC00703	[H][C@]3([C@H](C)/C=C/C(CC)C(C)C)CC[C@@]4([H])[C@]2([H])CCC1=CC(=O)CC[C@]1(C)[C@@]2([H])CC[C@]34C
S8	SANC00706	C/C(=C[C@H](C(C)C)N([CH3])C(=O)[C@@H](NC(=O)[C@@H](N[CH3])C(C)(C)c1cn([CH3])c2ccccc12)C(C)(C)C)C(=O)O

Table 3.6: Identification information for the 2 SANCDB compounds that failed Lipinski's rule of 5.

no.	SANC ID	compound SMILES
S9	SANC00512	[H][C@@]5(O[C@H]4C[C@@]3([H])[C@]2([H])CC=C1C[C@@H](O)CC[C@]1(C)[C@@]2([H])[C@H](O)C[C@]3(C)[C@@]4([H])[C@H](C)[C@@H](O)C/C=C(C)/C)O[C@@H](C)[C@H](O)[C@@H](OC(C)=O)[C@H]5OC(C)=O
S10	SANC00548	[H][C@]2([C@H](C)CC[C@H](O)C(C)(C)O[C@@H]1O[C@H](CO)[C@@H](O)[C@H](O)[C@H]1O)CC[C@]5(C)[C@]2(C)CC[C@@]36C[C@@]34C(=O)C=CC(C)(C)[C@]4([H])C[C@H](O)[C@@]56[H]

The molecules were given codenames from S1 to S10 for simplification. Molecules S9 and S10 failed two Lipinski's rule of 5 tests and were not used for molecular dynamics simulations. It is clear when looking at the SMILES format that the two molecules who did not perform well in the Lipinski's rule of 5, are significantly larger than the

other molecules in the tables.

In Table 3.4 above, in all cases except one, the calculated binding energy decreased during the molecular dynamic simulation. The exception is molecule S2 where the docking posed yielded a binding energy value of 10.5 kcal/mol and by the end of the simulation of the molecule in chain A, the energy had risen to -11.79 kcal/mol. The results that were generated using the poses by the end of the simulation should reflect the binding energy more accurately. The Autodock Vina procedure forces the docking position whereas the molecular dynamics simulation takes into account the movement of sidechains. Although molecular dynamic simulations are useful for lead searching, it would take laboratory physical experiments to confirm and determine the accuracy of these predictions.

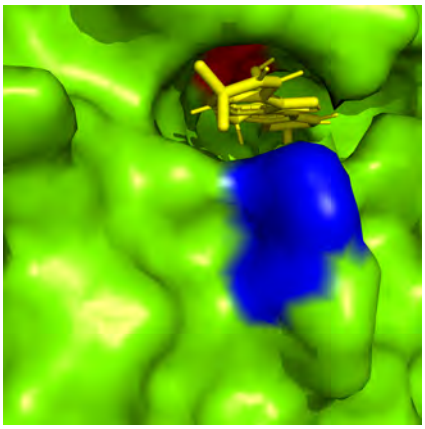
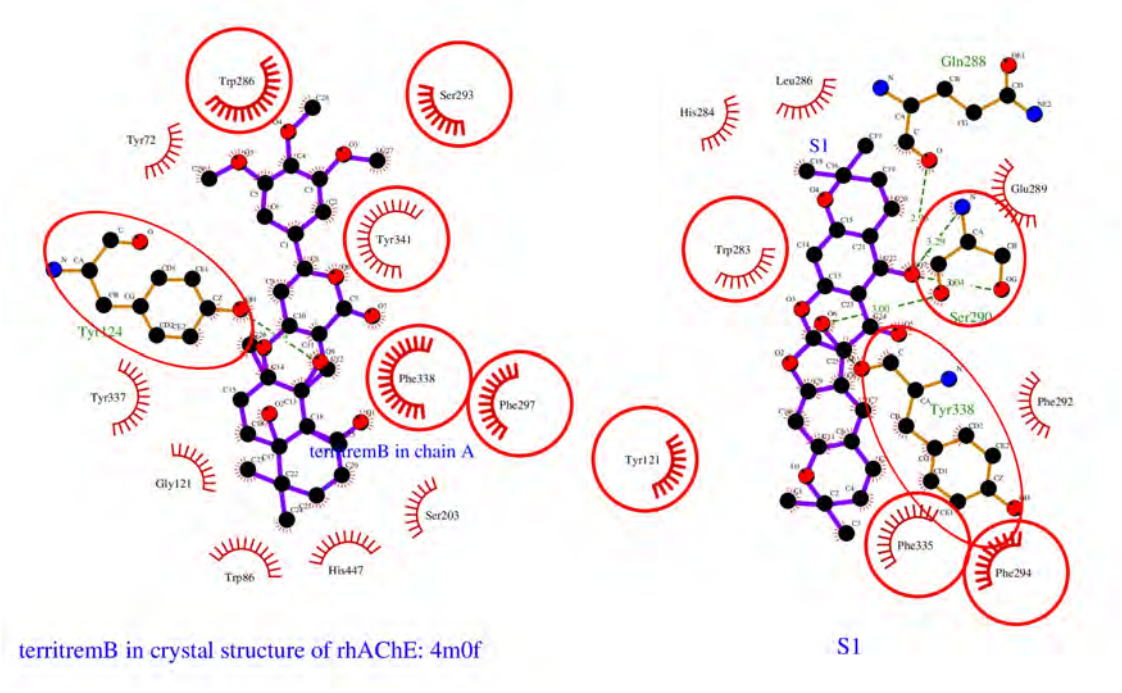
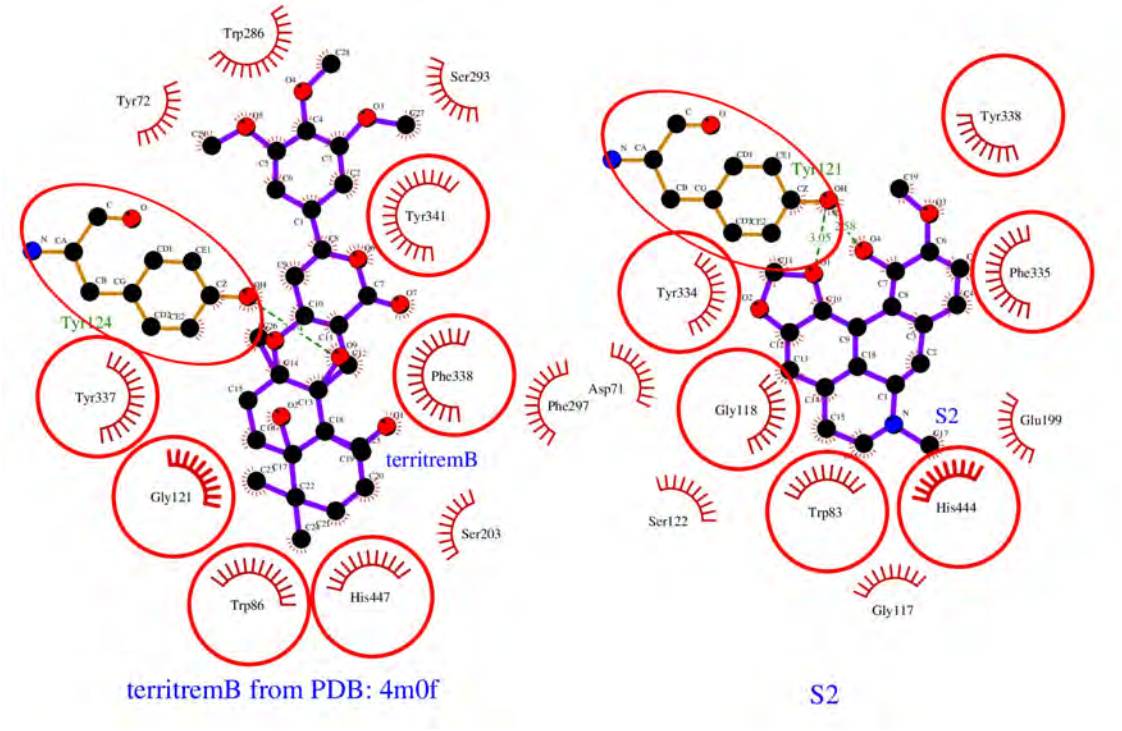


Figure 3.12: visualization of best performing molecule at 100ns of simulation

Figure 3.12, rendered using PyMol, depicts the active site of the enzyme. The residues that line the tunnel of the active site each contribute either favorably or unfavorably towards the overall binding energy. Proline 344 in blue contributed most negatively to binding energy. Arginine 296 in red contributed most unfavorably towards the binding energy



(a)



(b)

Figure 3.13: SANCDB ligand interaction diagrams. (a) S1 compared to territremB from the crystal structure using interaction plots. (b) S2 compared to territremB

Figure 3.13 includes ligand interaction diagrams of two molecules that were selected for further analysis using 100ns molecular dynamic simulations and MM-PBSA. The results of this further analysis are presented in the Molecular dynamics results section. The plots represent hydrogen bonds using a stippled green line and van der Waal interactions are displayed by indicating the residue involved surrounded by a

read half-circle. Residues that interact with both molecules are fully circled in red. Residue numbering may be inconsistent between the two molecule interaction plots due to alternative numbering that was introduced through modeling. The interaction diagrams were created using ligplot+ v1.4.5 Laskowski and Swindells, 2011.

Compounds 512 and 548 from the South African Natural Compounds database had both scored positive for 2 violations of Lipinski's rule of 5 and so these compounds were not taken further for molecular dynamics as they would not be suitable drugs even if they showed high affinity to the target. These compounds were the largest compounds which makes them unlikely to be able to efficiently pass through the blood brain barrier (BBB) where the target enzyme is located. Therefore the other 8 compounds were selected for 10ns molecular dynamic simulations to further asses their affinity to bind to the protein.

3.5.3 ZINC15 Subset

Table 3.7: 10 selected compounds from ZINC15 subset screening

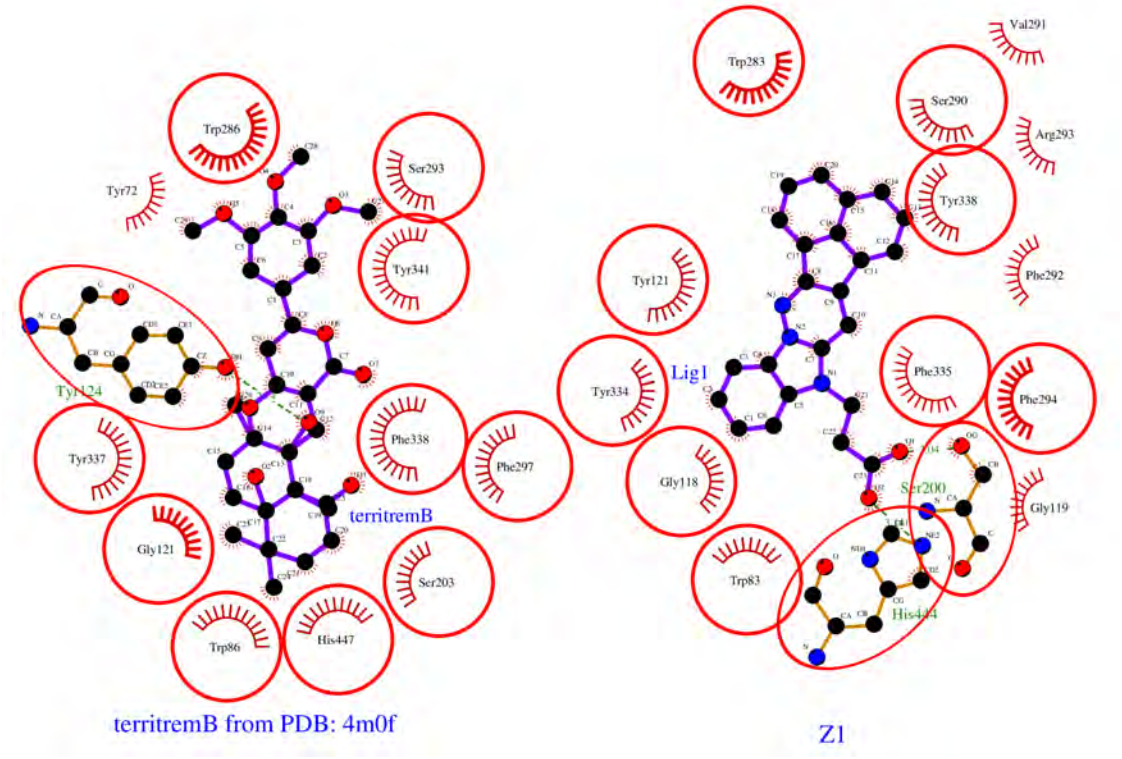
compound	**binding energy (kcal/mol)**	**no. hydrogen bonds at pose**	**initial binding energy at 10ns (chain A)**	**binding energy at 10ns (chain B)**
Z1	-12.9	2	-11.18	-11.12
Z2	-12.4	2	-7.35	-7.74
Z3	-12.5	1	-8.90	-7.90
Z4	-12.5	4	-9.03	-7.76
Z5	-12.2	3	-9.04	-10.39
Z6	-12.1	2	-10.72	-10.65
Z7	-11.8	2	-9.89	-8.44
Z8	-12.1	4	-9.38	-9.60
Z9	-12.4	4	-9.56	-8.77
Z10	-11.4	1	-11.53	-10.22

Table 3.8: Structure information for the selected 10 ZINC compounds

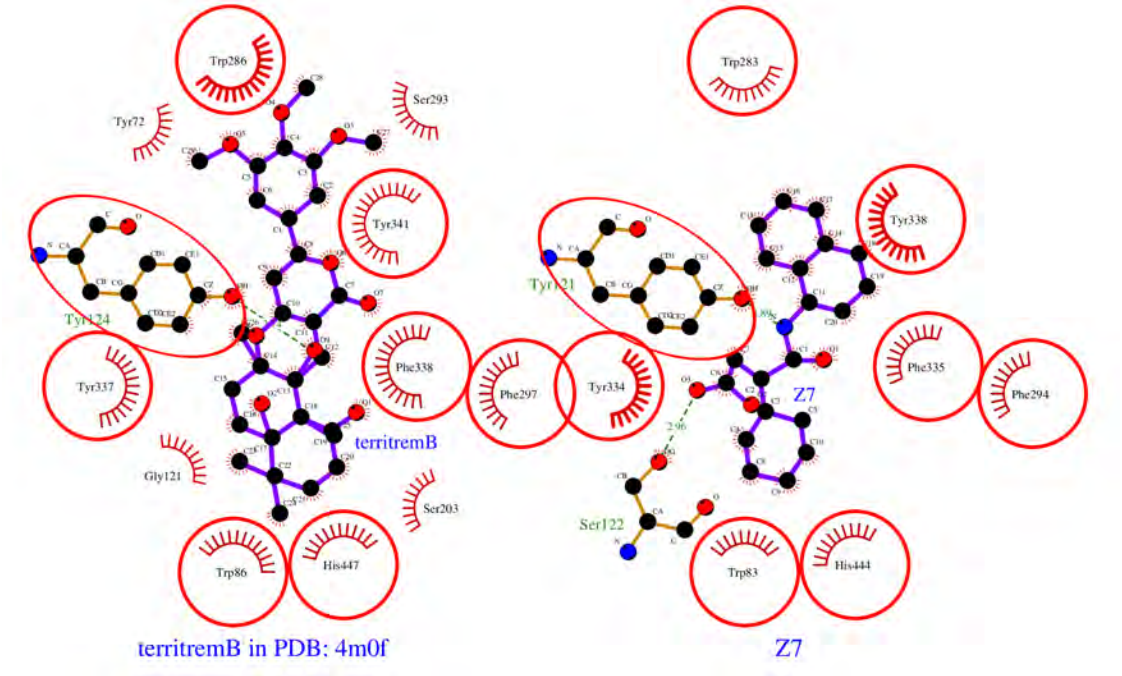
no.	ZINC ID	compound SMILES
Z1	ZINC628966	O=C(O)CCn1c2ccccc2[n+]2nc3c(cc12)-c1cccc2cccc-3c12
Z2	ZINC816436	[H][CC(=O)Nc1ccc(NS(=O)(=O)c2ccc3c4c(cccc24)C(=O)N3C)cc1
Z3	ZINC679638	Cn1c(=O)c2ncn(CC(=O)Nc3ccc4c5c(cccc35)CC4)c2n(C)c1=O
Z4	ZINC00361569	Cc1ccccc1NC(=O)Cn1c(-c2nonc2N)nc2ccccc21
Z5	ZINC00679618	CCOC(=O)c1ccc(NC(=O)Cn2c(-c3nonc3N)nc3ccccc32)cc1
Z6	ZINC00631389	O=C(CSc1nnc2ccccn12)N1CCC(c2noc3ccc(F)cc23)CC1
Z7	ZINC00310453	O=C1C[C@@H](C(=O)Nc2cccc3ccccc23)C2(CCCCC2)O1
Z8	ZINC02889230	NS(=O)(=O)c1ccc(NC(=O)CSc2nnc3ccc4ccccc4n23)cc1
Z9	ZINC00677623	Cc1ccc(NC(=O)Cn2c(-c3nonc3N)nc3ccccc32)cc1F
Z10	ZINC00359857	O=C1NN(c2ccc(Cl)cc2)C(=O)/C1=C/c1ccc2c(c1)OCO2

Binding energy values are given in kcal/mol. Binding energy scores at 10ns of the simulation were calculated using the Autodock VINA scoring function. This was calculated for the molecule in chain A and chain B separately. The 10 selected ZINC compounds were given codes to simplify their naming.

The compounds downloaded from the ZINC database displayed the same trend as the compounds from the SANC database. This trend is the decrease in calculated binding energy between the position of the compound at the start and the end of the simulation. As the simulation progresses the compound and enzyme would theoretically shift into more energetically favorable positions. This indicates that autodock VINA has a tendency to overestimate the true binding energy between the molecule and the enzyme. It should be noted that the molecules are docking to and enzyme with a completely open active site, which may be behind the high autodock VINA scores.



(a)



(b)

Figure 3.14: Ligand interaction diagrams from ZINC subset. (a) Z1 compared to territremB from the crystal structure using interaction plots. (b) Z7 compared to territremB.

Figure 3.14 contains ligand interaction diagrams of two molecules that were selected for further analysis using 100ns molecular dynamic simulations and MM-PBSA. The results of this further analysis are presented in the Molecular dynamics results section. The plots represent hydrogen bonds using a stippled green line and wan der

Waal interactions are displayed by indicating the residue involved surrounded by a read half-circle. Residues that interact with both molecules are fully circled in red. Residue numbering may be inconsistent between the two molecule interaction plots due to alternative numbering that was introduced through modeling. The interaction diagrams were created using ligplot+ (Laskowski and Swindells, 2011)

An observation can be made that the selected compounds from ZINC showed higher binding energy scores than the selected SANC compounds. The SANCDB compounds however, had higher number of residues in common with territremB when it binds. This seems to be caused by the larger size of the SANCDB compounds.

3.6 Molecular Dynamics of Enzyme and Enzyme-ligand Complexes

The stability of ligand-protein complexes determine their suitability as a drug. Examining the dynamic trajectories of these complexes were used to calculate free binding energy as well as the RMSD and RMSF of the complexes. The apoprotein simulation is also evaluated as this is a reference. The MMPSA results are complimentary to the Autodock Vina results as both these tools aim to calculate the binding free energy of the protein-ligand complex. Molecular dynamics has thus been used here as a compliment and continuation of the molecular docking experiment.

3.6.1 APO AChE

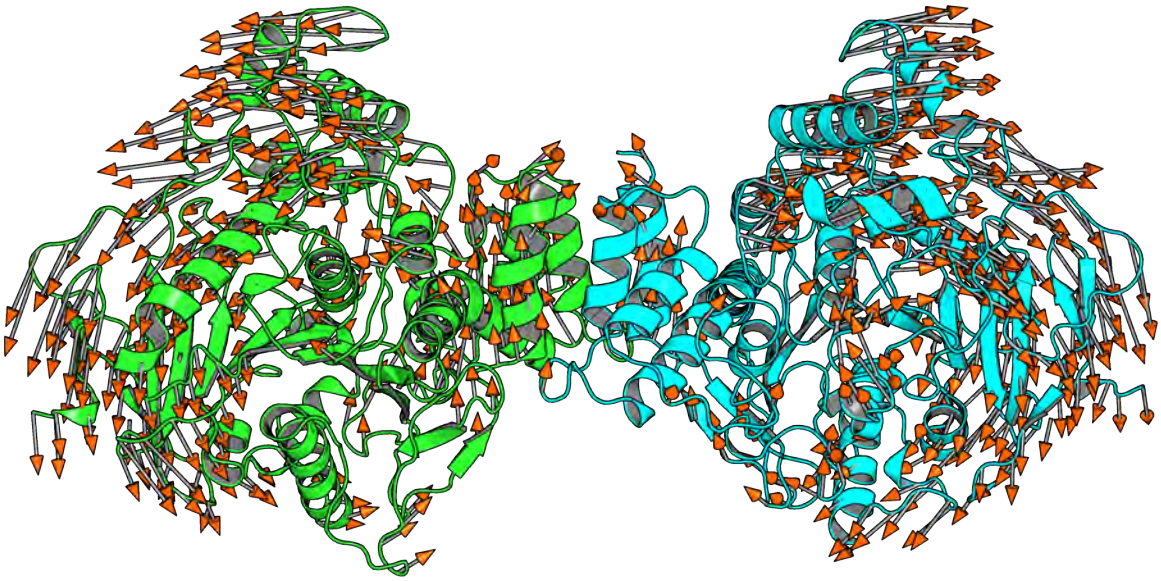


Figure 3.15: APO rhAChE porcupine plot created using modevectors in PyMol

This plot can be compared to similar plots of other protein variant simulations to reveal the discrepancy in motion between the two. This plot reveals a circular pattern in the protein movement which may be linked to the functioning of the enzyme.

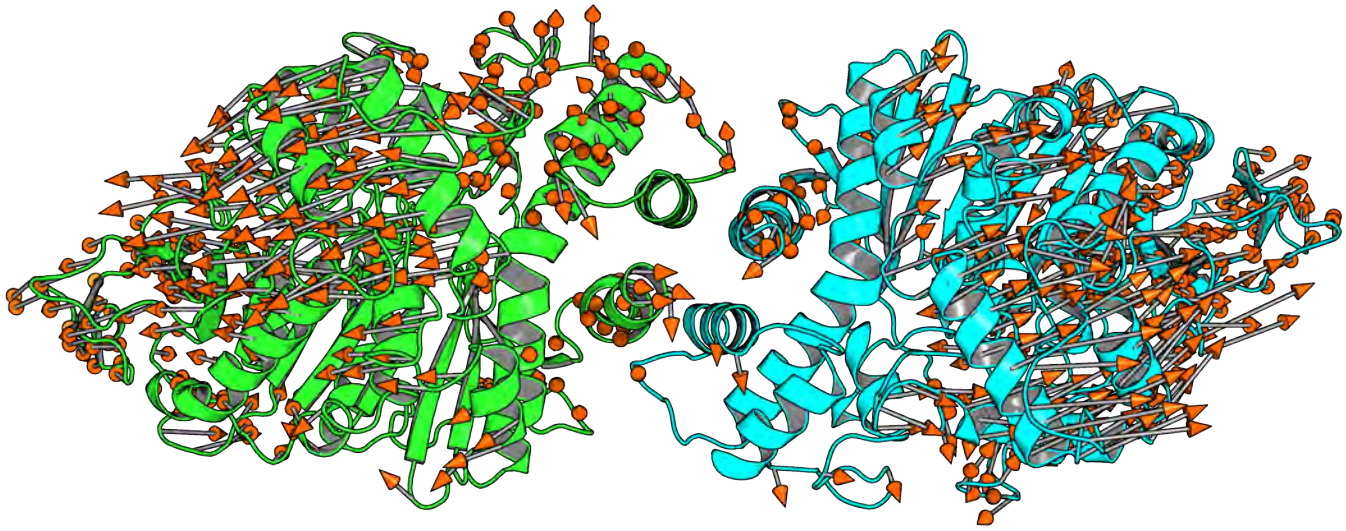


Figure 3.16: Porcupine plot of wild type rhAChE dimer created using modevectors in PyMol

The images in Figures 3.15 and 3.16 are rotated approximately 90 degrees and are the same porcupine plot. Porcupine plots were generated using modevecors PyMol script.

The coordinates at time 0ns was used as the starting positions and the ending positions are the coordinates from time 100ns. This information is required when using the modevectors script which generates the porcupine plots. The motions shown in the porcupine plots above resembles a shearing motion between the two subunits of the protein dimer. This agrees with a previous investigation of motions in the tetramer formation (Gorfe et al., 2008). This motion may lead to an active site that is more accessible to the substrate. In the tetramer formation, this motion may be implicated in the electrostatic anionic force that guides the substrate into the active site towards the catalytic triad in the central anionic site. Arrow length indicates magnitude of displacement in the direction show. Longer molecular dynamic simulations or targeted simulations may provide more insight into the extent of these motions and their frequency.

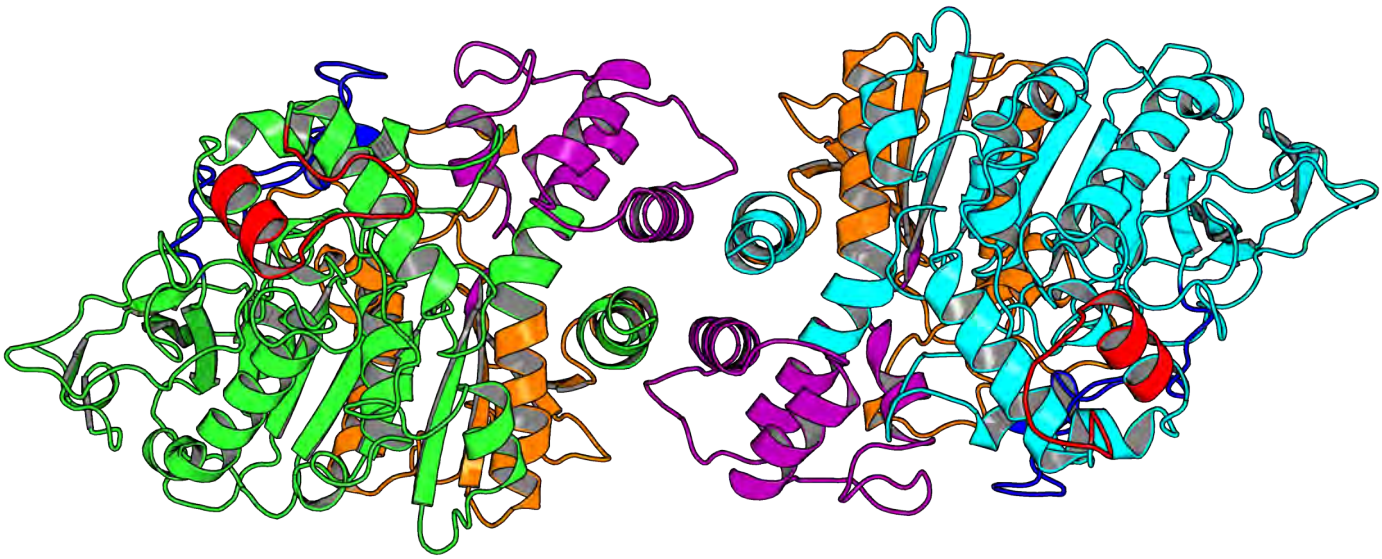


Figure 3.17: rhAChE structure coloured for reference. Areas between disulfide bonds are: blue (69:96), red (257:272), orange (409:529). The region in purple is a region that displays large movements and is next to the active site region.

Areas between disulfide bonds are colored. Disulfide loops have no special attribute themselves, it is the disulfide bond that forms between two cysteine residues that is structurally significant. The area that is 'pinched' off by the disulfide bond is known as a disulfide loop. Three of these strong bonds occur within the acetylcholinesterase enzyme. The porcupine plot for this protein can be seen in figure 3.16.

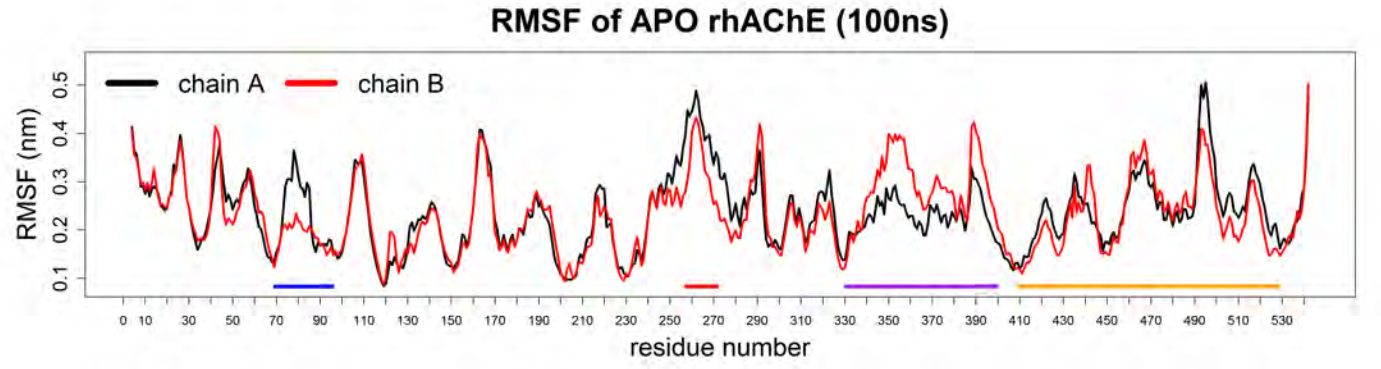


Figure 3.18: RMSF for residues in rhAChE apoprotein. Areas between disulfide bonds are: blue (69:96), red (257:272), orange (409:529). The region in purple is a region that displays large movements and is next to the active site region.

The regions of colour in Figure 3.18 have been mapped to an image of the apoprotein rhAChE dimer in Figure 3.17. The coloured regions indicate the residues that form a disulfide loop as they fall between a disulfide bond which occurs only between two cysteine residues. Disulfide bonds are very strong and are responsible for holding

the core formation of the protein intact. The residues that fall on the outside of a disulfide bond form a kind of loop. This loop does not poses any special motif, it is the bond that is of significance. Coloured regions: blue indicates residues 69 to 96, green indicates residues 257 to 272 and orange marks residues 409 to 529. Notably the region spanning residues 330 to 400 (purple) shows variance in its motion between the subunits. Subunit A experiences more overall displacement, but less fluctuation in this region. Subunit B experiences less overall displacement but more fluctuation in this region. The same trend is observed for the disulfide loop between residues 69 and 96 shown in blue. For this region subunit A experiences higher fluctuation but less overall displacement. Subunit B experiences lower fluctuation but higher overall displacement during the simulation. Higher fluctuation in this case results in holding the region in place. Displacement is revealed by the arrows in the porcupine plot in figure 3.16.

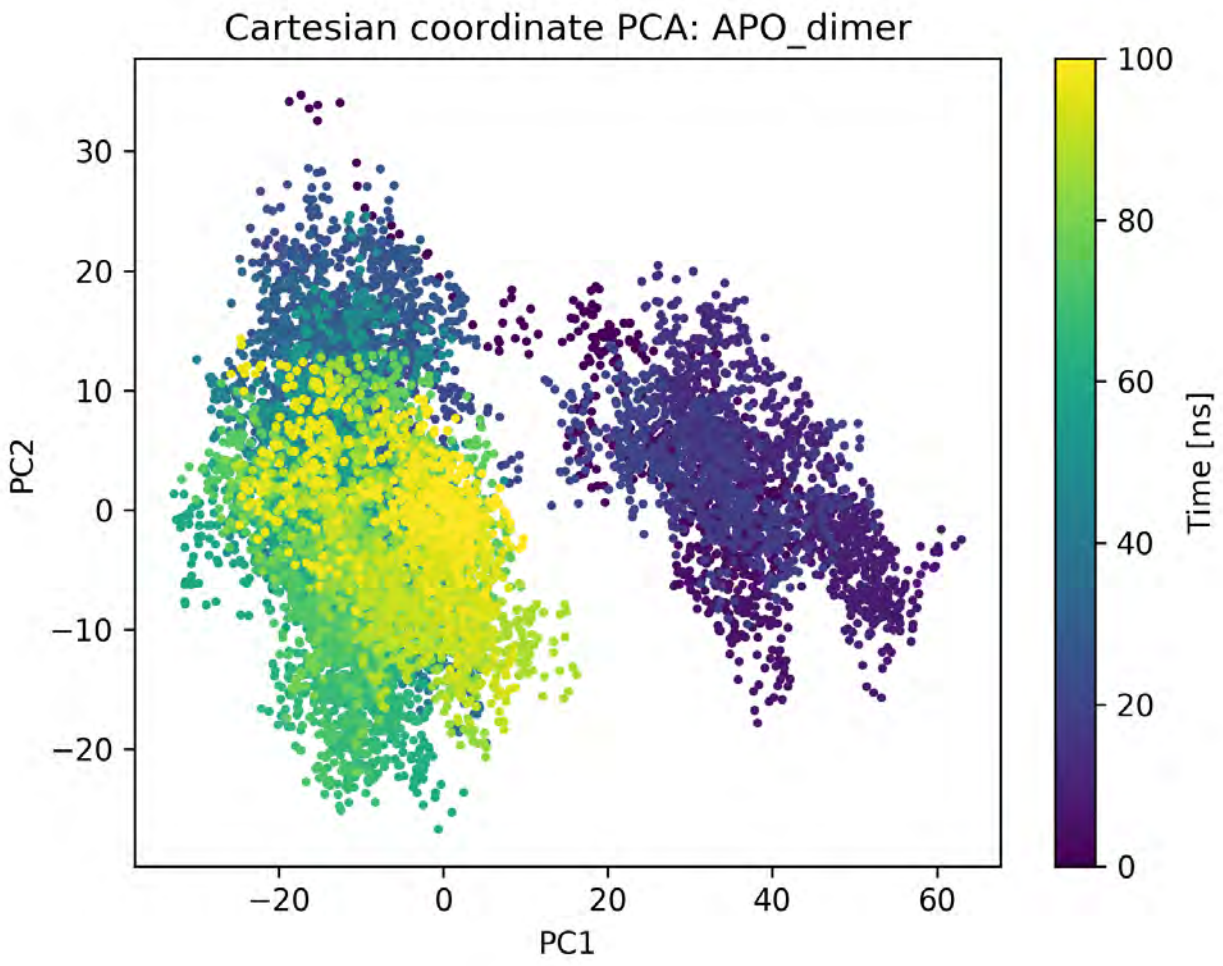


Figure 3.19: PCA analysis of rhAChE apoprotein created using MODE-TASK (Ross et al., 2018)

The Cartesian coordinate ensemble used for PCA was generated by performing a 100ns molecular dynamics simulation. PC1 explains 50% of the variance in motion and PC2 explains 10.6% of the variance.

The PCA plot indicates that the direction of the most prominent movements at the end of the simulation significantly different from the start of the simulation. This indicates that during the simulation the most important motion undergoes a distinct change in direction around the 25ns mark of the simulation. A more verbose explanation of what the PCA plot consists of, is provided in the PCA section of materials and methods.

3.6.2 TerritremB

In order to validate the molecular dynamic simulation, the territremB molecule that was co-crystalized with the enzyme was simulated. This inhibitor strongly, but covalently binds to AChE. The molecule shows high binding affinity values according to the VINA docking results(-13.69) and is expected to deviate little in the molecular dynamic simulation.

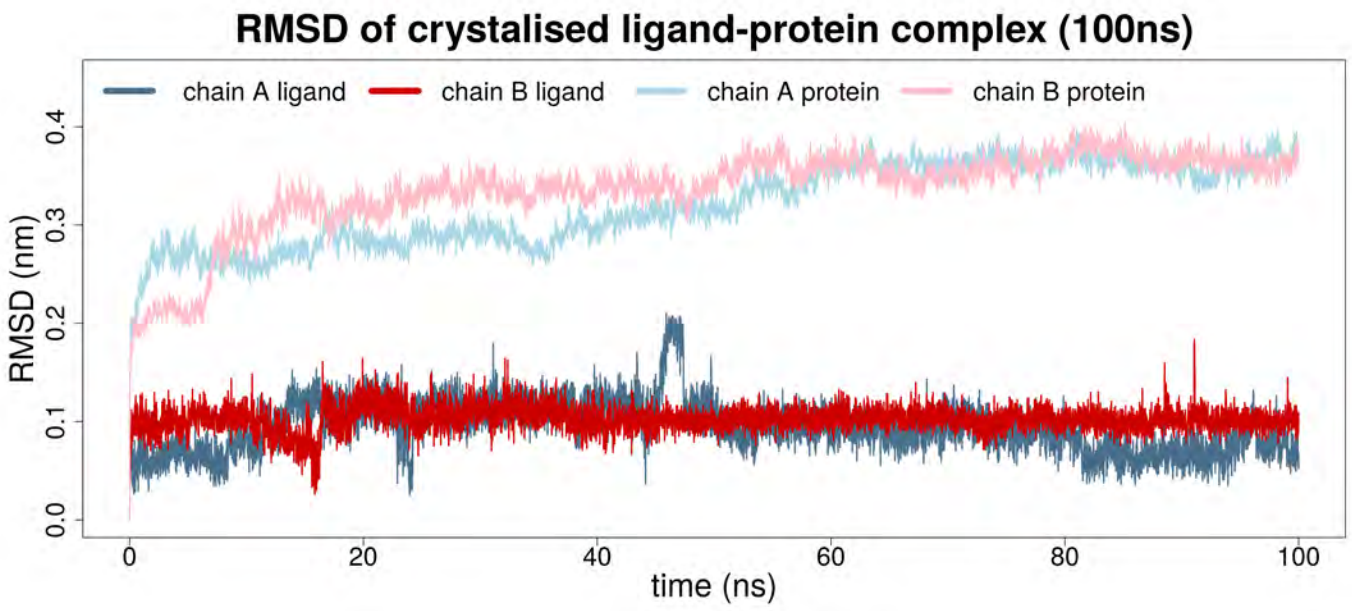


Figure 3.20: RMSD of territremB in complex with rhAChE dimer. Graph generated using R

TerritremB was used as a template molecule to compare to potential therapeutic in-

hibitors which bind to the peripheral anionic site of the acetylcholinesterse enzyme.

The protein RMSD rises quickly as the simulation starts and then stabilizes. The ligands that are simulated together start and end up in about the same position. This is expected as this is the ligand that the protein is crystalized with and it should fit very well.

3.6.3 S1

Molecule S1 illustrated high binding affinity to rhAChE according to docking and molecular dynamic analysis. It resulted in the highest free binding energy value calculated using MM-PBSA.

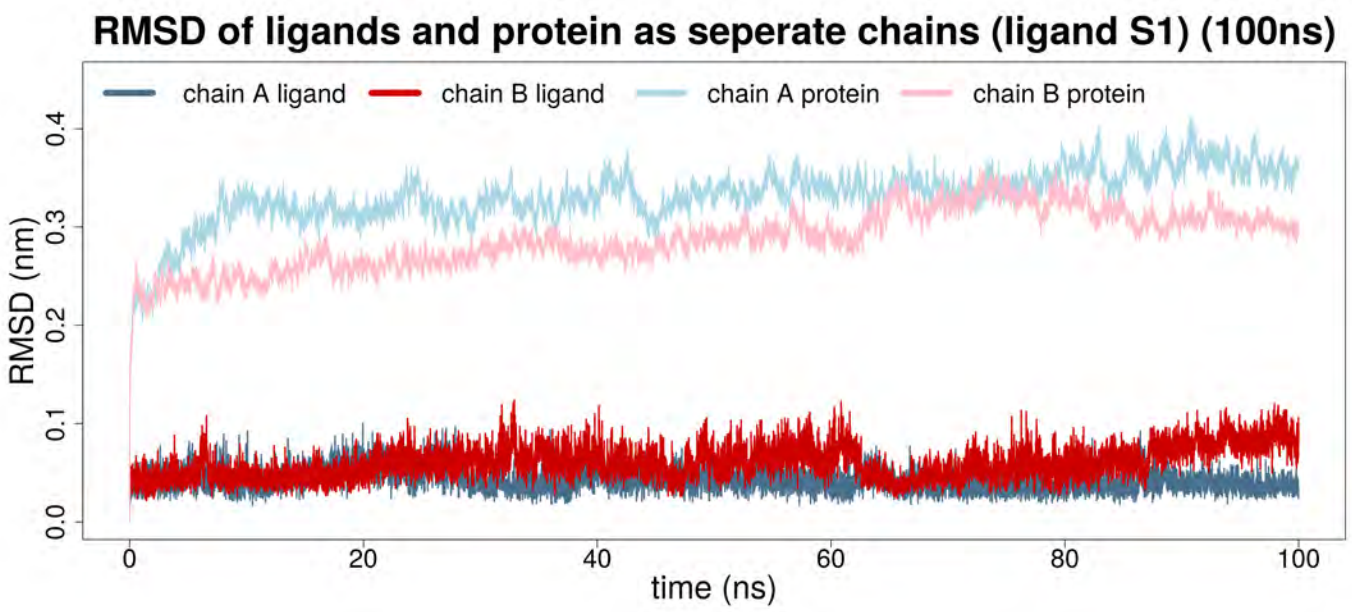


Figure 3.21: RMSD of protein and ligand complex for molecule S1. Graph generated using R

The molecule S1 delivered the highest MM-PBSA binding energy calculation results. Discrepancy between its performances in different chains of the protein can be seen. It can be seen in Figure 3.21 that the molecule deviates less in chain A of the protein than chain B. Less deviation may not always indicate a stronger binding energy however, as the MMPBSA calculations resulted in higher calculated affinity in chain B. This can be seen in Table 3.9 and Table 3.10. The net binding energy was approximately 22 kj/mol stronger in chain B of the protein. One should thus be weary of predicting

a high binding affinity from only RMSD information.

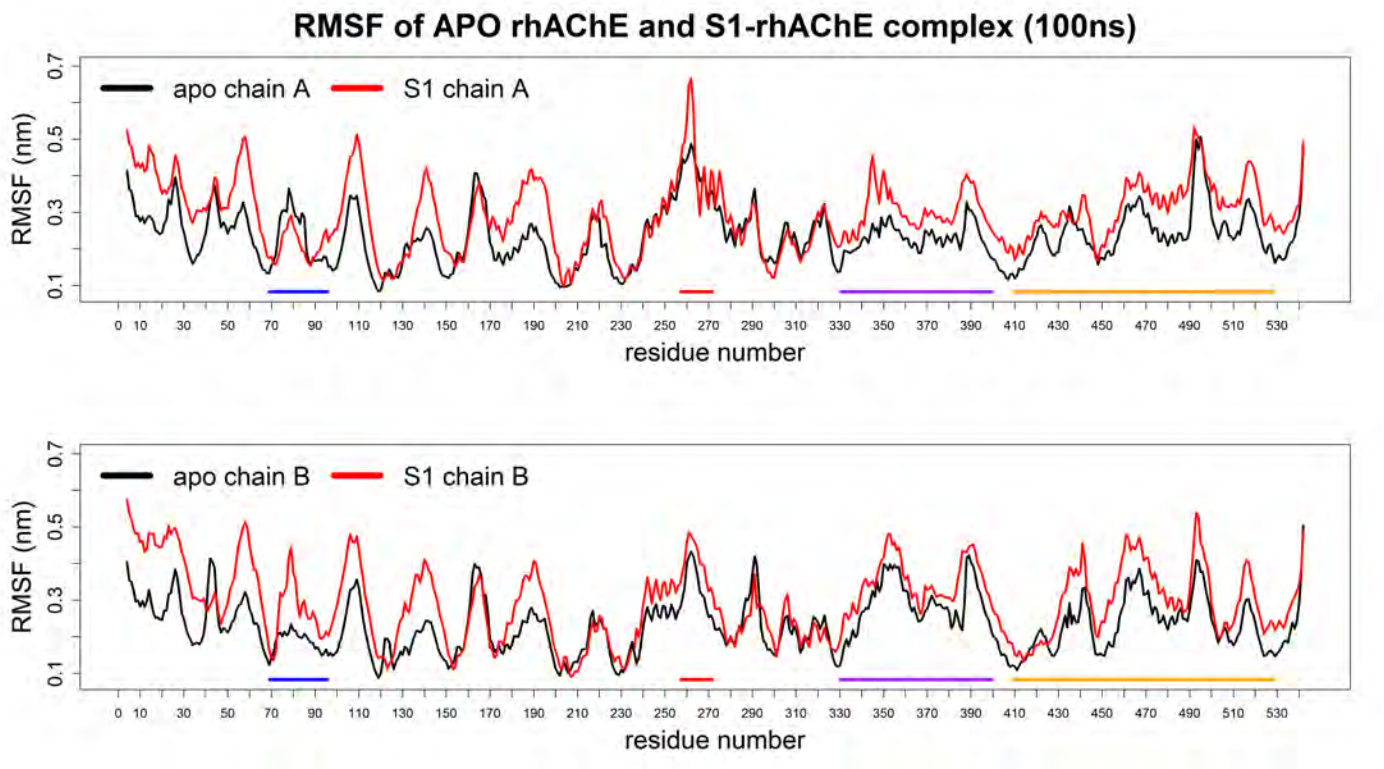


Figure 3.22: RMSF of Apo AChE and AChE-S1 complex. Areas between disulfide bonds are: blue (69:96), red (257:272), orange (409:529). The region in purple is a region that displays large movements and is next to the active site region.

Figure 3.22 includes the RMSF for each chain of the rhAChE protein. This includes the apoprotein and the protein that was simulated with the docked S1 molecule in the active site. S1 was the best performing molecule according to MM-PBSA calculations. Overall, the protein-ligand complex resulted in substantially higher RMSF values than the Apo protein. The entire length of the protein appears to be compensating for the interaction with the ligand.

3.6.4 South African Natural Compounds Database

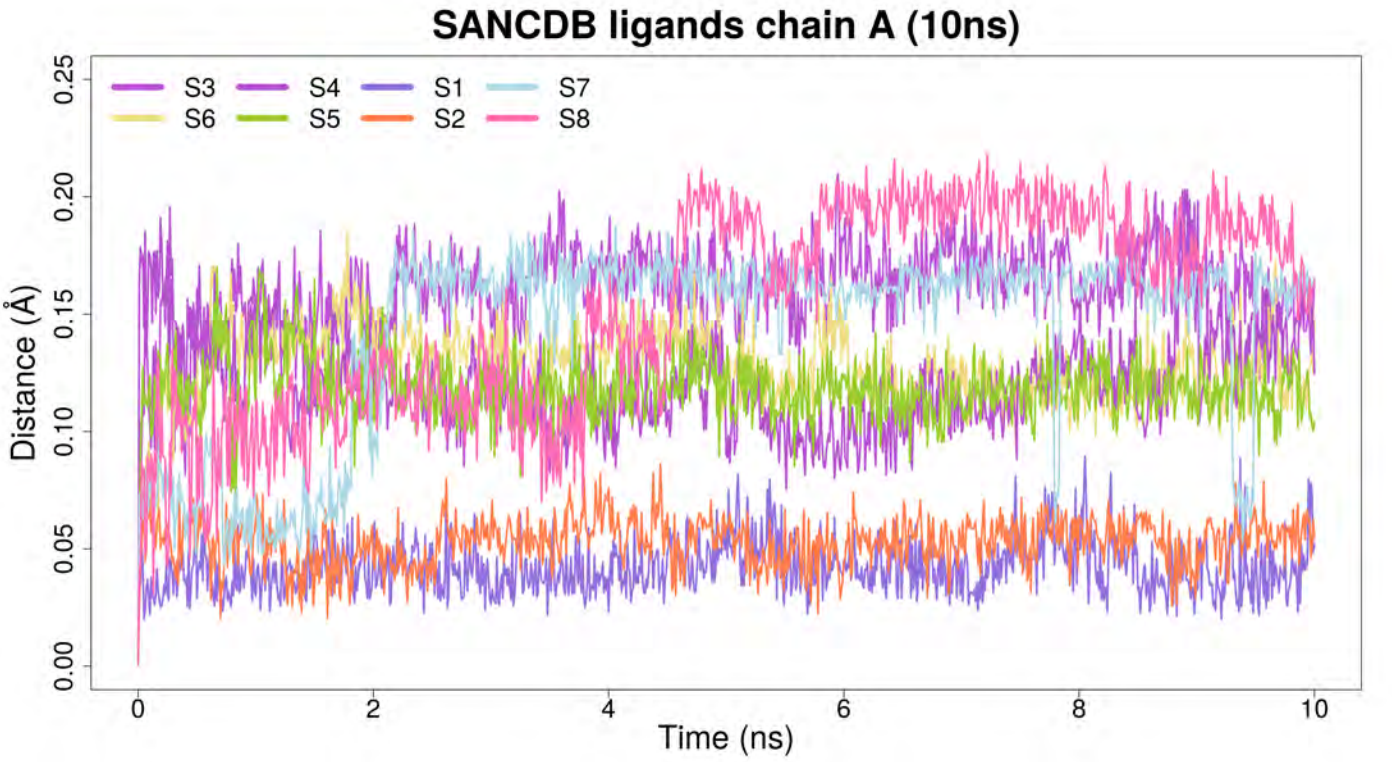


Figure 3.23: RMSD of ligands docked to chain A of homodimer created using R

Two ligands stand out in the 10 nanosecond molecular dynamics simulations. These are ligands S1 and S2 in the figures. These molecules show a relatively stable RMSD throughout the simulation. The RMSD remains close to 0.05nm. For chain B below, which was found to be less accommodating to most ligands than chain A, the same two compounds perform the best, regardless of the seemingly lower affinity of most ligands to this chain in this particular conformation.

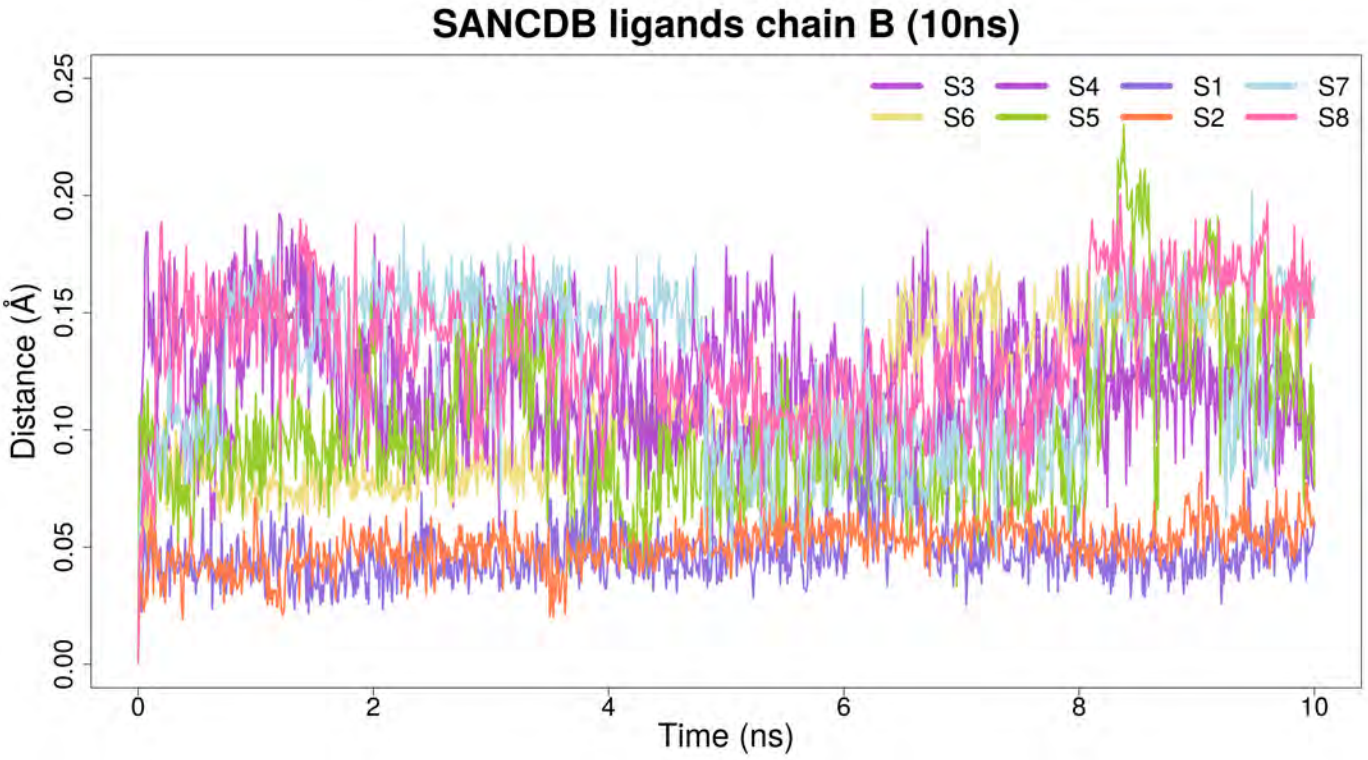


Figure 3.24: RMSD results of ligands docked to chain B of homodimer

3.6.5 RMSD of Selected Compounds From ZINC15 Database

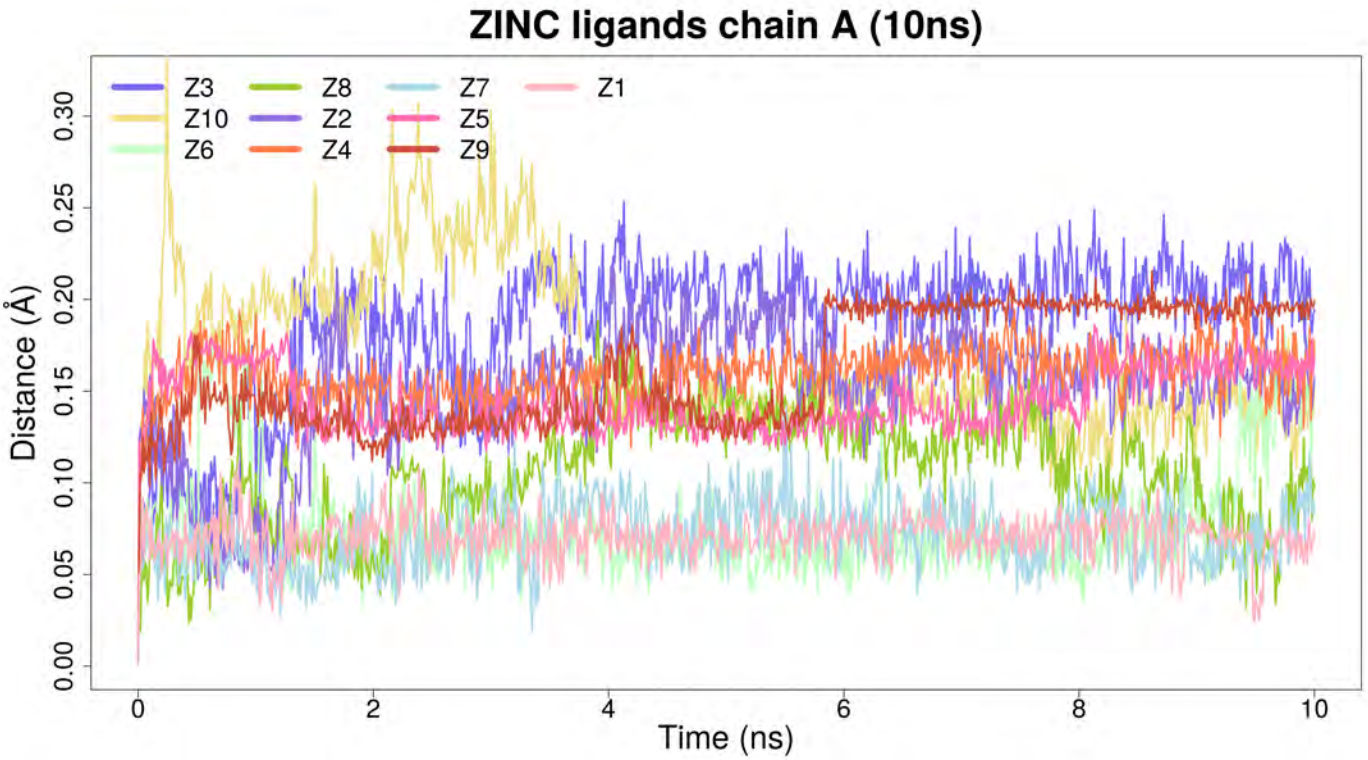


Figure 3.25: RMSD over 10ns for selected compounds from ZINC15

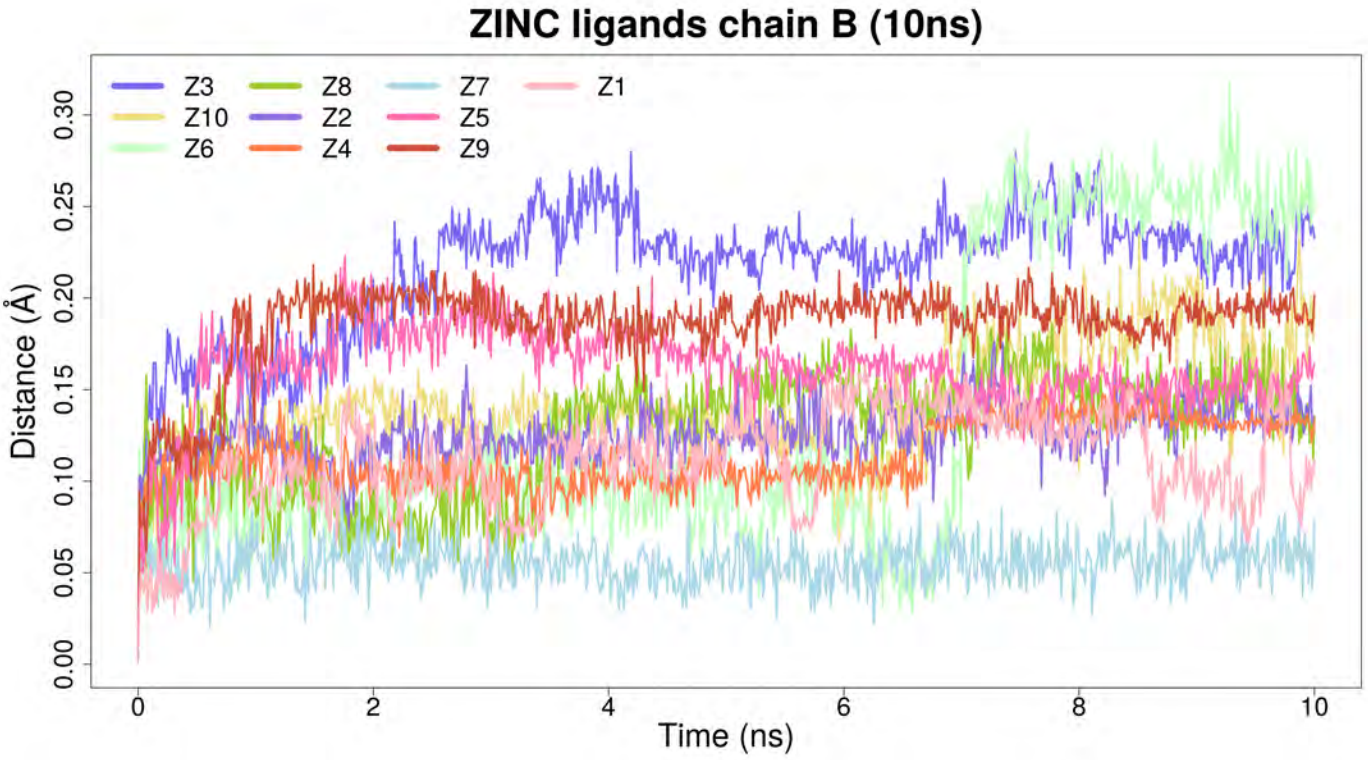


Figure 3.26: RMSD over 10ns for selected compounds from ZINC15

The RMSD values presented in Figures 3.25 and 3.26 indicate that certain ligands such as S5 are sporadic in chain B compared to chain A. Others such as Z1 and Z7 are consistent between the two chains of the dimer. The compounds labeled Z1 and Z7 were selected for further simulation. These compounds exhibited among the most consistent RMSD profiles of the 10 compounds that were examined. Some compounds showed highly variable RMSD between chain A and chain B of the enzyme, but overall descrimination between the low and high RMSD profiles were possible.

3.6.6 MM-PBSA (Molecular Mechanics Poisson Boltzmann Surface Area) Calculations

MM-PBSA calculations were performed to provide binding energy values that aid in determining whether the molecule is likely to bind and the strength of the bond. In each case, the last 5 nanoseconds of the simulation was sampled for the calculation.

Table 3.9: 5ns MM-PBSA results

Molecule	Chain	Van der Waals forces (kj/mol)	Electrostatic forces (kj/mol)	Polar solvation energy (kj/mol)	SASA energy (kj/mol)	net binding energy (kj/mol)
territremB	A	-281.553	-24.398	91.523	-24.290	-238.717
territremB	B	-272.422	-30.246	-24.743	24.745	-352.206
S1	A	-220.885	-9.782	80.371	-20.190	-170.478
S1	B	-258.848	-39.389	126.119	-22.124	-194.220
S2	A	-192.503	-31.190	104.528	-17.553	-136.699
S2	B	-179.499	-0.741	52.745	-15.950	-143.459
Z7	A	-172.460	-40.740	101.212	-18.119	-130.092
Z7	B	-164.889	-43.364	106.443	-16.562	-118.368
Z1	A	-193.659	-18.064	100.036	-20.045	-131.712
Z1	B	-168.611	-52.024	120.470	-18.838	-119.029

Ligand MM-PBSA free binding energy is broken down into individual components by g_mmpbsa. Energy values for 5 molecules are given. Each chain was calculated separately. Residues that contributed from the adjacent subunit to the subunit simulated in, were not taken into account. Each column represents a component that contributes towards the final binding energy. All energy values are given in kj/mol.

Table 3.10: 10ns MM-PBSA results

Molecule	Chain	Van der Waals forces (kj/mol)	Electrostatic forces (kj/mol)	Polar solvation energy (kj/mol)	SASA energy (kj/mol)	net binding energy (kj/mol)
territremB	A	-278.159	-25.521	91.271	-24.106	-236.545
territremB	B	-269.722	-28.319	-24.643	-24.642	-347.312
S1	A	-218.882	-10.951	78.749	-19.913	-170.995
S1	B	-257.795	-38.498	126.157	-22.258	-192.412
S2	A	-192.901	-32.659	105.385	-17.364	-137.546
S2	B	-178.791	-2.149	54.333	-16.002	-142.626
Z7	A	-172.177	-40.752	101.121	-18.032	-129.845
Z7	B	-163.434	-44.213	108.221	-16.493	-115.922
Z1	A	-194.979	-12.388	107.602	-20.063	-119.805
Z1	B	-170.264	-53.089	120.424	-18.738	-121.677

The calculation to generate the values in Table 3.10 above was done using the last 10ns of the simulation trajectory as the sample. The last 5ns is the same trajectory section that was used in Table 3.9. This table essentially added the 5ns to the calculation done previously. This was done to provide more robust results as a larger sample of time should provide a more accurate binding energy estimation. This can

be compared with results generated by using half the time to investigate the impact of lengthening the sampled time has on results.

There is little variation between the MM-PBSA free binding energy values calculated from 5ns, and the values calculated using 10ns of the trajectory. MM-PBSA calculations indicate that the co-crystalized ligand is the most effective molecule that was evaluated. This is expected as territremB binds non-covalently to acetylcholinesterase, but the combination of many interaction along its length, create a strong force (Cheung et al., 2013). The second most effective molecule is S1 with calculated free energies of -170.478 kj/mol and -194.220 kj/mol for chain A and chain B respectively. The docking score of this molecule was among the highest scores and the docking pose predicted that 5 hydrogen bonds could be formed with the enzyme. The binding free energies of the remaining three molecules are average compared to the top two compounds. The compounds from the SANCDB were allocated less refined topologies because of their larger size. TerritremB also was allocated an unrefined topology.

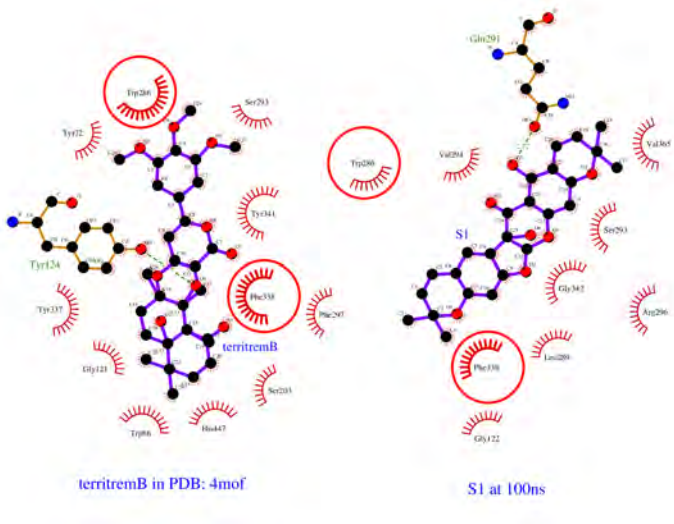
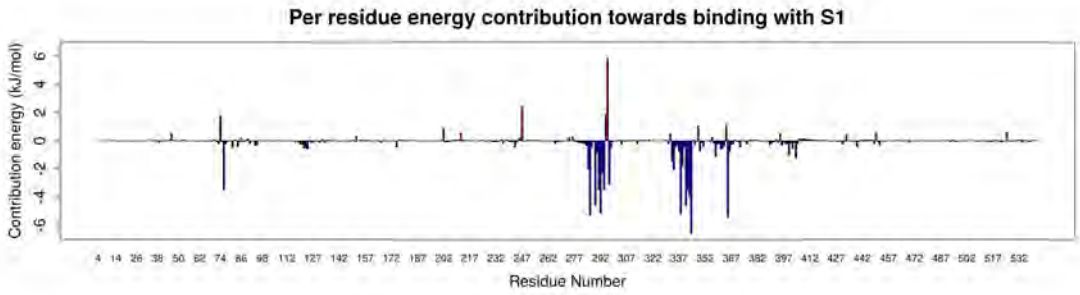
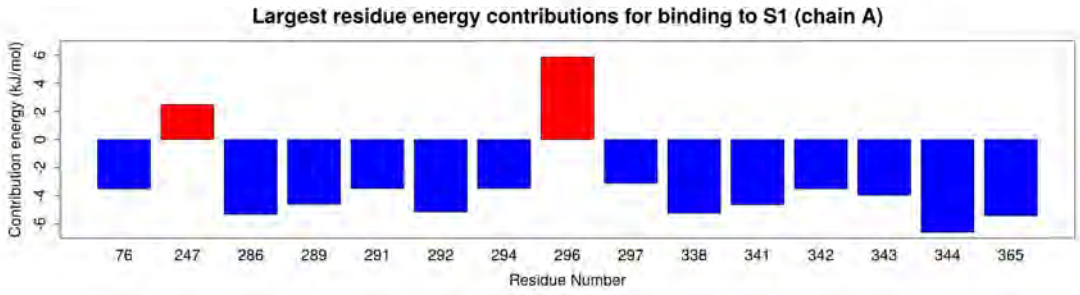


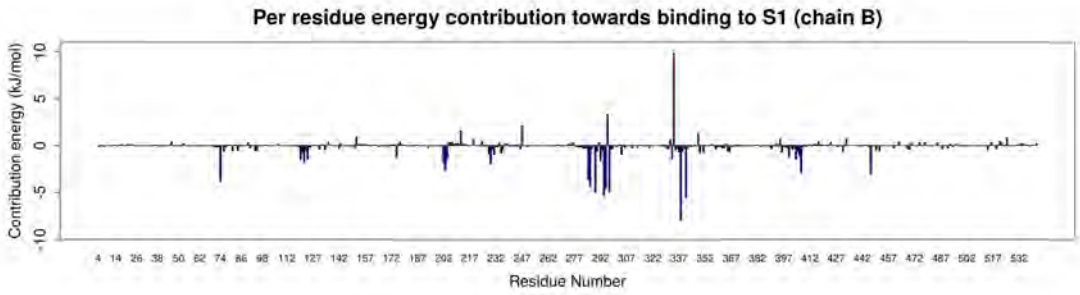
Figure 3.27: Interaction diagram of rhAChE-S1 complex at 100ns. Left: territrem B, right: SANC00347. This image was captured using ligplot+ (Laskowski and Swindells, 2011)

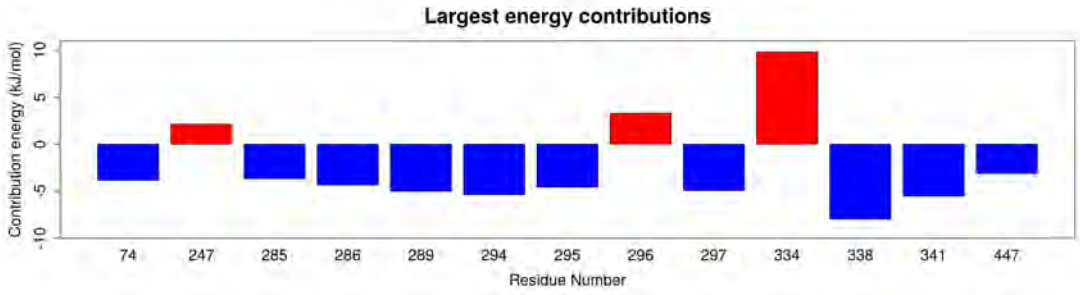
In the ligplot diagram in Figure 3.27 above, TerritremB on the left is included for comparison. The S1 molecule at 100ns has shifted and no longer interacts with many of the residues that it interacted with according to the Autodock VINA pose. S1 was rescored at its 100ns coordinates using the VINA score_only function, delivering -13.51 kcal/mol and -14.03 kcal/mol for chain A and B respectively.





(a)





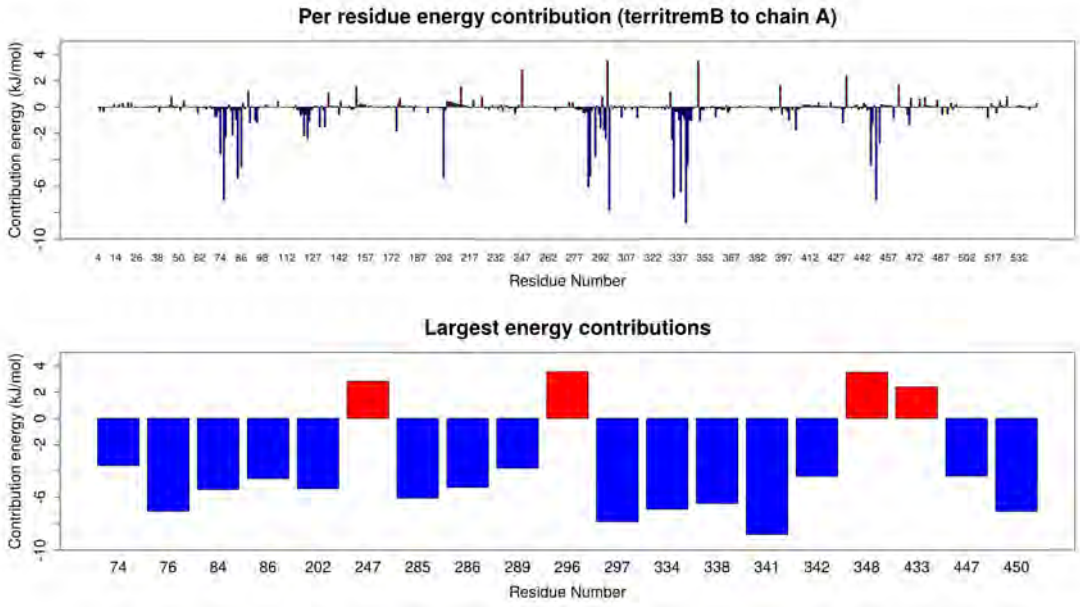
(b)

Figure 3.28: Per residue energy contribution histogram for S1 created using R. (a) - chain A of homodimer. (b) - chain B of homodimer

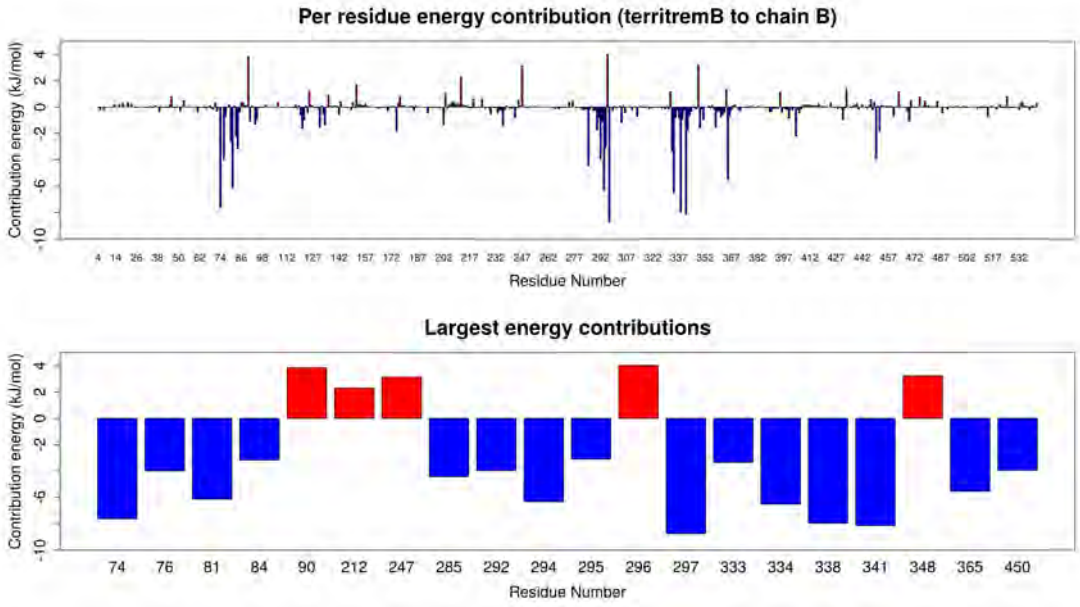
Figure 3.28 is a per residue energy contribution histogram for the number one ranked compound in terms of MM-PBSA binding energy results. The first plot gives an overview of regions of residues that contribute favorable or not favorable to the binding interaction. The second plot for each subfigure indicates the residues that contributed the most towards the binding energy, either positively (red) or negatively (blue). The first figure:(a) contains the binding energy values for chain A of the rhAChE dimer to the ligand bound to chain A. The second figure:(b) contains binding energy values for residues from chain B of the simulated homodimer enzyme.

Figure 3.29 below is a MM-PBSA histogram of contribution energies of individual residues to identify valuable residues taking part binding of the ligand. Each chain contained its own ligand during the molecular dynamic simulation. The energy value of the chain that the molecule was not bound to, was disregarded. The chains were treated separately for MM-PBSA calculations. The first figure:(a) contains residue contribution energies for the interaction of territremB with chain A of the enzyme. The second figure:(b) contains residue contribution energies for the interaction of a separate territremB molecule with chain B of the enzyme.

Figure 3.29 below indicates that residues 341 and 297 contributes the most towards strong binding of territremB. This residue may play a similar role in the binding to other molecules, such as the acetylcholine substrate. Residue 341 is one of the residues that are temporarily occluded by another subunit in the tetramer AChE assembly. The occlusion will likely negatively affect the binding of some molecules to the occluded active site (Gorfe et al., 2008).



(a)



(b)

Figure 3.29: MM-PBSA residue contribution histogram for territremB created using R. (a) - chain A of homodimer. (b) - chain B of homodimer

Chapter 4

Critical Discussion and Concluding Remarks

4.1 Introduction

There are a few limitations in this study and aspects which could be improved. This is discussed in this chapter.

4.2 Docking

A substantial number of molecules that were docked showed high binding affinity values. This made it difficult to choose which molecules to investigate in more detail. The active site of this enzyme is open only for a small amount of time (less than 3%) and alternates rapidly between open and closed states. This means that it is not certain whether the molecules that were docked, would reach the active site where the docking algorithm placed them. This is a limitation of this docking experiment as the binding affinity value assumes that the ligand can enter the site. The crystal structure used for docking was taken from a complex of the enzyme with territremB, which distorts the active site gorge. This leads to the active site gorge being wider than it would normally be. This may have biased the docking results as the active site may not naturally be as accessible. TerritremB simulation should be accurate as it was originally in the protein, meaning the binding conformation was experimentally determined. Docked molecules may be less accurate as they were not identified biologically to be able to enter the active site of the protein.

The drug and target concentrations were not determined. The residence time of the identified potential inhibitors are unknown. Residence time, in addition to binding affinity, is a significant factor determining the efficacy of a drug.

4.3 Molecular Dynamics

It is suggested that running multiple shorter independent simulations is more effective than long simulation. Long simulations can results in underestimating the uncertainty of the result. The molecular dynamics GROMACS engine allows simulations to use varying starting velocities with a random velocity generator. While the time here was not sufficient to run multiple simulations at the same length as was performed, it would have been possible to decrease the simulation time and running simulations with varying starting velocities. This would increase the statistical significance of the results and the conclusions that can be made (Genheden and Ryde, 2015).

The protein occurs in different structural assemblies such as tetramers, and monomers. The type of structural assembly being used may be important in analyzing the effects of SNP's and the binding of ligands to this protein. The functional unit of synaptic AChE forms a tetramer and is tethered to a membrane. The C-terminal tail of the acetylcholinesterase subunits were not incorporated into the model. Being able to include the membrane attachment as well as the tetramer would provide a simulation which more accurately simulates biological conditions. Available models of the AChE tetramer are not of a high resolution.

4.4 Variant Effect Predictions

The molecular dynamics results are consistent with expectations of amino-acid properties. The P247L variant resulted in disrupted protein motion while T229S did not. The T229S variant resulted in motions also present in the wild type. This method for analyzing the influence of single nucleotide variants may be used for other proteins and variants. The method was able to successfully discriminate between a variant which does not have a significant influence on protein dynamics and one that does.

Appendices

A PIR alignment file for use with MODELLER

```
>P1;4m0f_ATOM
structureX:4m0f.pdb:4     :A :542    :B :::-1.00:-1.00
EDAELLVTVRGGRLRGIRLKTPGGPVSAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYP
GFEGTEMWNPNRELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSMNYRV
GAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASVGMHLLSPPSRGLFHRAVLQ
SGAPNGPWATVGMGEARRRATQLAHLVGCP------NDTELVACLRTRPAQVLVNHEWHVLPQESVFRFSFVPVV
DGDFLSDTPEALINAGDFHGLQVLVGVVKDEGSYFLVYGAPGFSKDNESLISRAEFLAGVRVGVPQVSDLAAEAV
VLHYTDWLHPEDPARLREALSDVVGDHNVVCPVAQLAGRLAAQGARVYAYVFEHRASTLSWPLWMGVPHGYEIEF
IFGIPLDPSRNYTAEEKIFAQRLMRYWANFARTGDPNEPRD---PQWPPYTAGAQQYVSLDLRPLEVRRGLRAQA
CAFWNRFLPKLLSA/
EDAELLVTVRGGRLRGIRLKTPGGPVSAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYP
GFEGTEMWNPNRELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSMNYRV
GAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASVGMHLLSPPSRGLFHRAVLQ
SGAPNGPWATVGMGEARRRATQLAHLVGCPP--TGGNDTELVACLRTRPAQVLVNHEWHVLPQESVFRFSFVPVV
DGDFLSDTPEALINAGDFHGLQVLVGVVKDEGSYFLVYGAPGFSKDNESLISRAEFLAGVRVGVPQVSDLAAEAV
VLHYTDWLHPEDPARLREALSDVVGDHNVVCPVAQLAGRLAAQGARVYAYVFEHRASTLSWPLWMGVPHGYEIEF
IFGIPLDPSRNYTAEEKIFAQRLMRYWANFARTGDPNEP--PKAPQWPPYTAGAQQYVSLDLRPLEVRRGLRAQA
CAFWNRFLPKLLSA*

>P1;4m0f_wt
sequence::     : :     : :::-1.00:-1.00
EDAELLVTVRGGRLRGIRLKTPGGPVSAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYP
GFEGTEMWNPNRELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSMNYRV
GAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASVGMHLLSPPSRGLFHRAVLQ
SGAPNGPWATVGMGEARRRATQLAHLVGCPPGGTGGNDTELVACLRTRPAQVLVNHEWHVLPQESVFRFSFVPVV
DGDFLSDTPEALINAGDFHGLQVLVGVVKDEGSYFLVYGAPGFSKDNESLISRAEFLAGVRVGVPQVSDLAAEAV
VLHYTDWLHPEDPARLREALSDVVGDHNVVCPVAQLAGRLAAQGARVYAYVFEHRASTLSWPLWMGVPHGYEIEF
IFGIPLDPSRNYTAEEKIFAQRLMRYWANFARTGDPNEPRDPKAPQWPPYTAGAQQYVSLDLRPLEVRRGLRAQA
CAFWNRFLPKLLSA/
EDAELLVTVRGGRLRGIRLKTPGGPVSAFLGIPFAEPPMGPRRFLPPEPKQPWSGVVDATTFQSVCYQYVDTLYP
GFEGTEMWNPNRELSEDCLYLNVWTPYPRPTSPTPVLVWIYGGGFYSGASSLDVYDGRFLVQAERTVLVSMNYRV
GAFGFLALPGSREAPGNVGLLDQRLALQWVQENVAAFGGDPTSVTLFGESAGAASVGMHLLSPPSRGLFHRAVLQ
SGAPNGPWATVGMGEARRRATQLAHLVGCPPGGTGGNDTELVACLRTRPAQVLVNHEWHVLPQESVFRFSFVPVV
DGDFLSDTPEALINAGDFHGLQVLVGVVKDEGSYFLVYGAPGFSKDNESLISRAEFLAGVRVGVPQVSDLAAEAV
VLHYTDWLHPEDPARLREALSDVVGDHNVVCPVAQLAGRLAAQGARVYAYVFEHRASTLSWPLWMGVPHGYEIEF
IFGIPLDPSRNYTAEEKIFAQRLMRYWANFARTGDPNEPRDPKAPQWPPYTAGAQQYVSLDLRPLEVRRGLRAQA
```

CAFWNRFLPKLLSA*

B NVT parameter file for GROMACS temperature equilibration

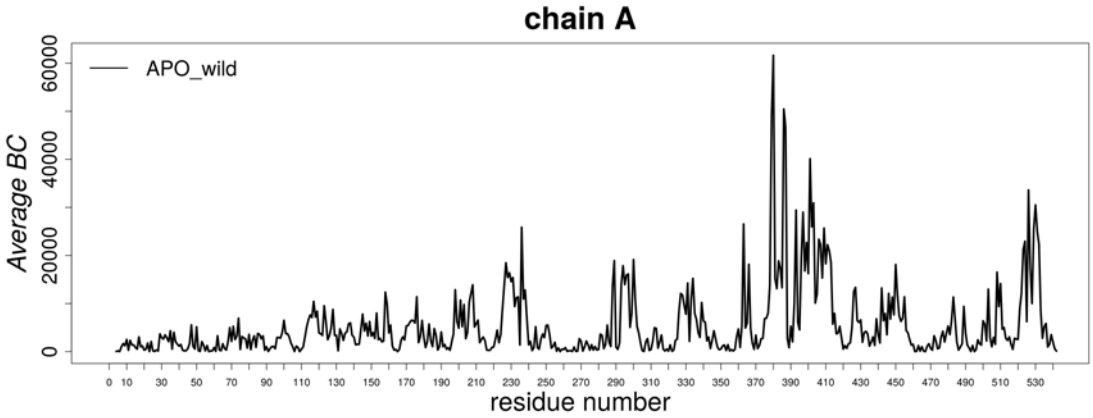
```
title                    = Protein-ligand complex NVT equilibration
define                   = -DPOSRES  ; position restrain the protein and ligand
; Run parameters
integrator               = md        ; leap-frog integrator
nsteps                   = 50000     ; 2 * 50000 = 100 ps
dt                       = 0.002     ; 2 fs
; Output control
nstenergy                = 500   ; save energies every 1.0 ps
nstlog                   = 500   ; update log file every 1.0 ps
nstxout-compressed       = 500   ; save coordinates every 1.0 ps
energygrps               = Protein A237 B237
; Bond parameters
continuation             = no        ; first dynamics run
constraint_algorithm     = lincs     ; holonomic constraints
constraints              = h-bonds   ; bonds to H are constrained
lincs_iter               = 1         ; accuracy of LINCS
lincs_order              = 4         ; also related to accuracy
; Neighbor searching and vdW
cutoff-scheme            = Verlet
ns_type                  = grid      ; search neighboring grid cells
nstlist                  = 20        ; largely irrelevant with Verlet
rlist                    = 1.2
vdwtype                  = cutoff
vdw-modifier             = force-switch
rvdw-switch              = 1.0
rvdw                     = 1.2       ; short-range van der Waals cutoff (in nm)
; Electrostatics
coulombtype              = PME       ; Particle Mesh Ewald for long-range electrostatics
rcoulomb                 = 1.2       ; short-range electrostatic cutoff (in nm)
pme_order                = 4         ; cubic interpolation
fourierspacing           = 0.16      ; grid spacing for FFT
; Temperature coupling
tcoupl                   = V-rescale                     ; modified Berendsen thermostat
tc-grps                  = Protein_A237_B237 Water_and_ions    ; two coupling groups - more accurate
tau_t                    = 0.1   0.1                     ; time constant, in ps
ref_t                    = 300   300                     ; reference temperature, one for each group, in K
; Pressure coupling
pcoupl                   = no        ; no pressure coupling in NVT
; Periodic boundary conditions
```

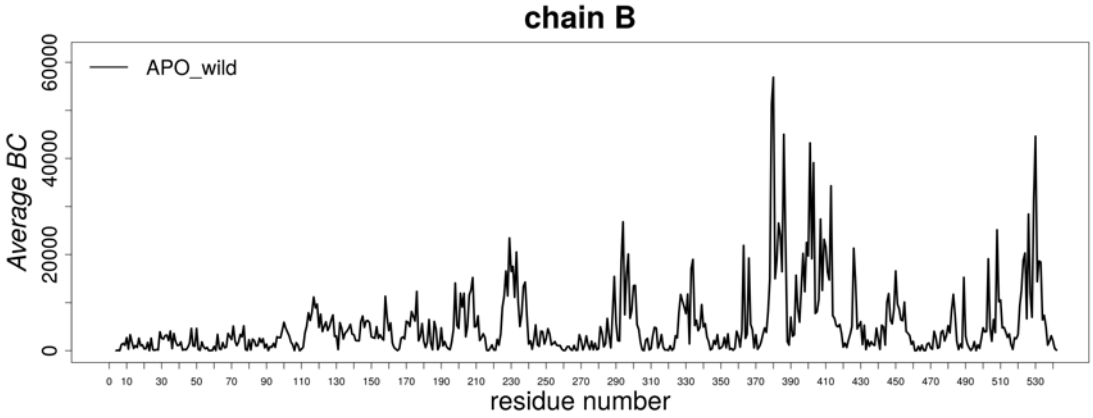
```
pbc                     = xyz       ; 3-D PBC
; Dispersion correction is not used for proteins with the C36 additive FF
DispCorr                = no
; Velocity generation
gen_vel                 = yes       ; assign velocities from Maxwell distribution
gen_temp                = 300       ; temperature for Maxwell distribution
gen_seed                = -1        ; generate a random seed
```

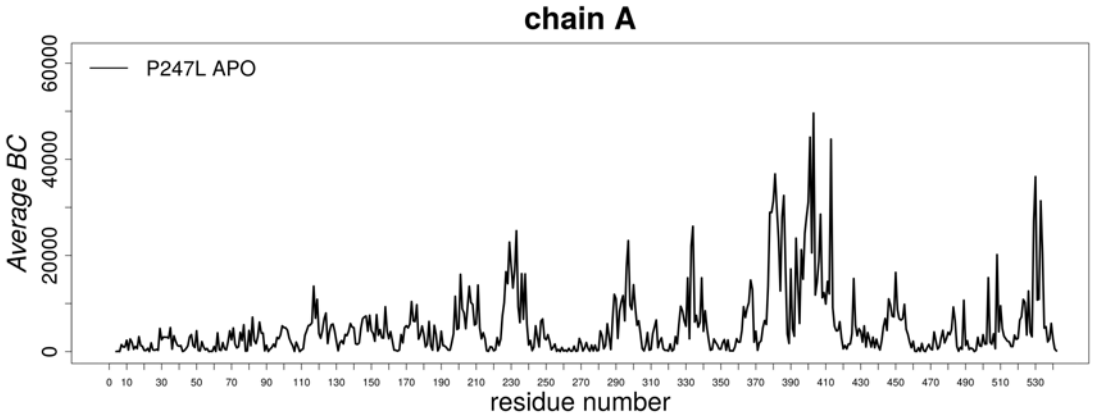
C Python script to run MODELLER (settings for best models)

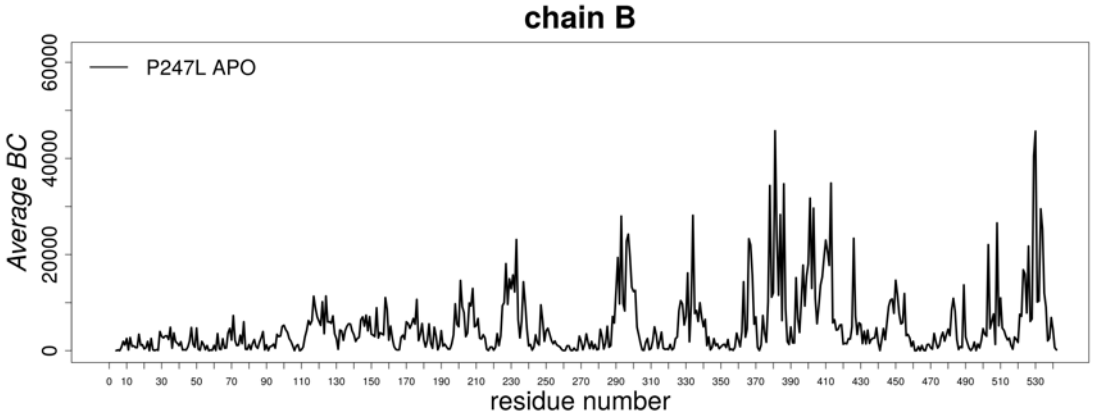
```
#!/usr/bin/python

from modeller import *
from modeller.automodel import *    # Load the automodel class

log.verbose()
env = environ()

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

class MyModel(automodel):
    def special_patches(self, aln):
        # Rename both chains and renumber the residues in each
        self.rename_segments(segment_ids=['A', 'B'],
                             renumber_residues=[4, 4])

    def select_atoms(self):
        return selection(self.residue_range('259:A', '264:A'),
                         self.residue_range('495:A', '497:A'), self.residue_range('260:B', '261:B'), self.r

a = MyModel(env, alnfile = 'dimer.pir',
            knowns = '4m0f_ATOM', sequence = '4m0f_wt')
a.starting_model= 1
a.ending_model  = 50

a.make()
```

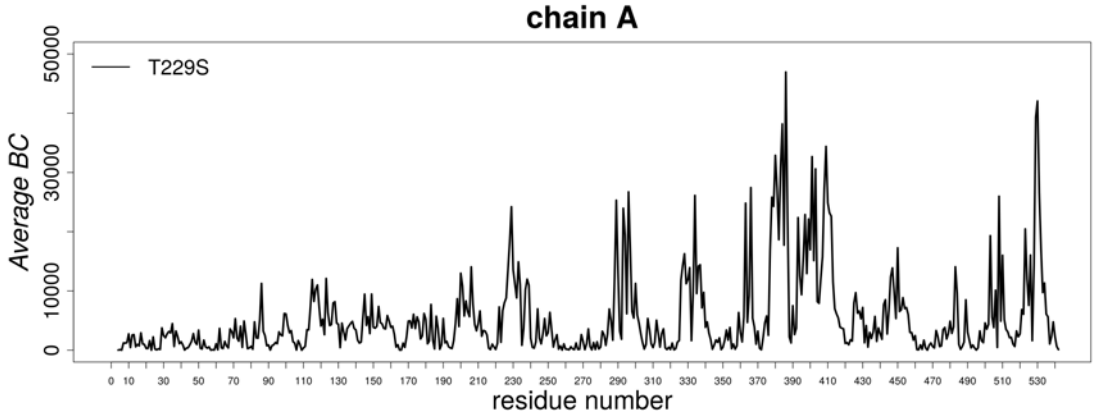
D Average Betweenness Centrality

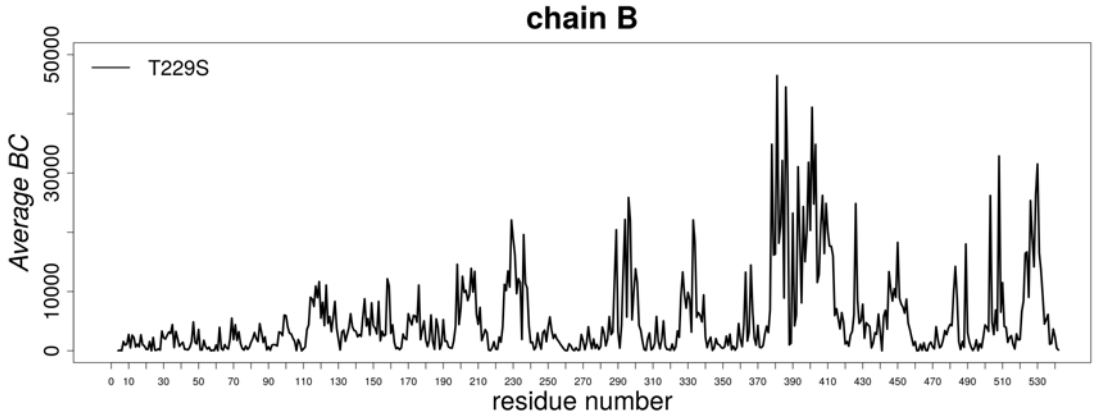









chain A
T229S
Average BC
0
10000
30000
50000
0 10 30 50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350 370 390 410 430 450 470 490 510 530
residue number


chain B
T229S
Average BC
0
10000
30000
50000
0 10 30 50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350 370 390 410 430 450 470 490 510 530
residue number

5 Bibliography

References

Anglister, L., & Silman, I. (1978). Molecular structure of elongated forms of electric eel acetylcholinesterase. *Journal of Molecular Biology*, *125*(3), 293–311.

Bartels, C. F., & Zelinski, T. (1993). Mutation at Codon 322 in the Human Acetyicholinesterase (ACHE) Gene Accounts for YT Blood Group Polymorphism. *American journal of human genetics*, *52*(5), 928–936.

Beri, V., Wildman, S. A., Shiomi, K., Al-Rashid, Z. F., Cheung, J., & Rosenberry, T. L. (2013). The Natural Product Dihydrotanshinone I Provides a Prototype for Uncharged Inhibitors That Bind Specifically to the Acetylcholinesterase Peripheral Site with Nanomolar Affinity. *Biochemistry*, *52*(42), 7486–7499.

Betts, M. J., & Russell, R. B. (2003). Amino Acid Properties and Consequences of Substitutions, 28.

Bon, S., Coussen, F., & Massoulié, J. (1997). THE POLYPROLINE ATTACHMENT DOMAIN OF THE COLLAGEN TAIL. *The Journal of Biological Chemistry*, *272*(5), 3016–3021.

Brown, D. K. [David K.], & Bishop, Ö. T. (2017). Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis: Analyzing Variation at the Protein Level. *Global heart*, *12*(2), 151–161.

Brown, D. K. [David K], Penkler, D. L., Sheik Amamuddy, O., Ross, C., Atilgan, A. R., Atilgan, C., & Tastan Bishop, Ö. (2017). MD-TASK: A software suite for analyzing molecular dynamics trajectories. *Bioinformatics*, *33*(17), 2768–2771.

Carlson, B. A., Dubay, M. M., Sausville, E. A., Brizuela, L., & Worland, P. J. (1996). Flavopiridol Induces Gj Arrest with Inhibition of Cyclin-dependent Kinase (CDK) 2 and CDK4 in Human Breast Carcinoma Cells, 7.

Chasman, D., & Adams, R. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation11Edited by F. Cohen. *Journal of Molecular Biology*, *307*(2), 683–706.

Cheung, J., Gary, E. N., Shiomi, K., & Rosenberry, T. L. (2013). Structures of Human Acetylcholinesterase Bound to Dihydrotanshinone I and Territrem B Show Peripheral Site Flexibility. *ACS Medicinal Chemistry Letters*, *4*(11), 1091–1096.

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation, 3.

Cramer, S. C. (2015). Drugs to Enhance Motor Recovery After Stroke. *Stroke*, *46*(10), 2998–3005.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137.

Dvir, H., Silman, I., Harel, M., Rosenberry, T. L., & Sussman, J. L. (2010). Acetylcholinesterase: From 3D structure to function. *Chemico-Biological Interactions*, *187*(1-3), 10–22.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. In *Methods in Enzymology* (Vol. 277, pp. 396–404).

Felder, C. E., Botti, S. A., Lifson, S., Silman, I., & Sussman, J. L. (1997). External and internal electrostatic potentials of cholinesterase models. *Journal of Molecular Graphics and Modelling*, *15*(5), 318–327.

Francis, P. T., Palmer, A. M., Snape, M., & Wilcock, G. K. (1999). The cholinergic hypothesis of Alzheimer's disease: A review of progress. *Journal of Neurology, Neurosurgery & Psychiatry*, *66*(2), 137–147.

Gais, S., & Born, J. (2004). Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation. *Proceedings of the National Academy of Sciences*, *101*(7), 2140–2144.

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, *10*(5), 449–461.

Gorfe, A. A., Chang, C.-e. A., Ivanov, I., & McCammon, J. A. (2008). Dynamics of the Acetylcholinesterase Tetramer. *Biophysical Journal*, *94*(4), 1144–1154.

Green, A. R. (2009). Pharmacological approaches to acute ischaemic stroke: Reperfusion certainly, neuroprotection possibly: Pharmacological approaches to acute ischaemic stroke. *British Journal of Pharmacology*, *153*(S1), S325–S338.

Grisaru, D., Sternfeld, M., Eldor, A., Glick, D., & Soreq, H. (1999). Structural roles of acetylcholinesterase variants in biology and pathology. *European Journal of Biochemistry*, *264*(3), 672–686.

Grossberg, G. T. (2003). Cholinesterase Inhibitors for the Treatment of Alzheimer's Disease: *Current Therapeutic Research*, *64*(4), 216–235.

Hasin, Y., Avidan, N., Bercovich, D., Korczyn, A., Silman, I., Beckmann, J. S., & Sussman, J. L. (2004). A paradigm for single nucleotide polymorphism analysis: The case of the acetylcholinesterase gene. *Human Mutation*, *24*(5), 408–416.

Hatherley, R., Brown, D. K., Musyoka, T. M., Penkler, D. L., Faya, N., Lobb, K. A., & Tastan Bishop, Ö. (2015). SANCDB: A South African natural compound database. *Journal of Cheminformatics*, *7*(1).

Hubbard, T. (2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*(1), 38–41.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, *14*, 33–38.

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., ... Cheatham, T. E. (2000). Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research*, *33*(12), 889–897.

Kumari, R., Kumar, R., Open Source Drug Discovery Consortium, & Lynn, A. (2014). *G_MMPBSA* —A GROMACS Tool for High-Throughput MM-PBSA Calculations. *Journal of Chemical Information and Modeling*, *54*(7), 1951–1962.

Lahti, J. L., Tang, G. W., Capriotti, E., Liu, T., & Altman, R. B. (2012). Bioinformatics and variability in drug response: A protein structural perspective. *Journal of The Royal Society Interface*, *9*(72), 1409–1437.

Lam, B., Hollingdrake, E., Kennedy, J. L., Black, S. E., & Masellis, M. (2009). Cholinesterase inhibitors in Alzheimer's disease and Lewy body spectrum disorders: The emerging pharmacogenetic story. *Human Genomics*, *4*(2), 91.

Laskowski, R. A., & Swindells, M. B. (2011). LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *Journal of Chemical Information and Modeling*, *51*(10), 2778–2786.

Liu, J.-Q., Dai, S.-X., Zheng, J.-J., Guo, Y.-C., Li, W.-X., Li, G.-H., & Huang, J.-F. (2017). The identification and molecular mechanism of anti-stroke traditional Chinese medicinal compounds. *Scientific Reports*, *7*(1).

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation*, *37*(3), 235–241.

Lockridge, O. (2015). Review of human butyrylcholinesterase structure, function, genetic variants, history of use in the clinic, and potential therapeutic uses. *Pharmacology & Therapeutics*, *148*, 34–46.

Lockridge, O., Norgren, R. B., Johnson, R. C., & Blake, T. A. (2016). Naturally Occurring Genetic Variants of Human Acetylcholinesterase and Butyrylcholinesterase and Their Potential Impact on the Risk of Toxicity from Cholinesterase Inhibitors. *Chemical Research in Toxicology*, *29*(9), 1381–1392.

Lu, H., & Tonge, P. J. (2010). Drug–target residence time: Critical information for lead optimization. *Current Opinion in Chemical Biology*, *14*(4), 467–474.

Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., ... Mark, A. E. (2011). An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *Journal of Chemical Theory and Computation*, *7*(12), 4026–4037.

Nagasundaram, N., Zhu, H., Liu, J., V, K., C, G. P. D., Chakraborty, C., & Chen, L. (2015). Analysing the Effect of Mutation on Protein Function and Discovering

Potential Inhibitors of CDK4: Molecular Modelling and Dynamics Studies. *PLOS ONE*, *10*(8), e0133969.

Neumann-Haefelin, C., Brinker, G., Uhlenküken, U., Pillekamp, F., Hossmann, K.-A., & Hoehn, M. (2002). Prediction of Hemorrhagic Transformation After Thrombolytic Therapy of Clot Embolism: An MRI Investigation in Rat Brain. *Stroke*, *33*(5), 1392–1398.

Ng, P. C., & Henikoff, S. (2006). Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annual Review of Genomics and Human Genetics*, *7*(1), 61–80.

Noetzli, M., & Eap, C. B. (2013). Pharmacodynamic, pharmacokinetic and pharmacogenetic aspects of drugs used in the treatment of Alzheimer's disease. *Clinical pharmacokinetics*, *52*(4), 225–241.

Pardridge, W. M. (2007). Drug Targeting to the Brain. *Pharmaceutical Research*, *24*(9), 1733–1744.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, *405*(6788), 847–856.

Ross, C., Nizami, B., Glenister, M., Sheik Amamuddy, O., Atilgan, A. R., Atilgan, C., & Tastan Bishop, Ö. (2018). MODE-TASK: Large-scale protein motion tools. *Bioinformatics*, *34*(21), 3759–3763.

Ryan, M., Diekhans, M., Lien, S., Liu, Y., & Karchin, R. (2009). LS-SNP/PDB: Annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics*, *25*(11), 1431–1432.

Sadowsky, J. D., Burlingame, M. A., Wolan, D. W., McClendon, C. L., Jacobson, M. P., & Wells, J. A. (2011). Turning a protein kinase on or off from a single allosteric site via disulfide trapping. *Proceedings of the National Academy of Sciences*, *108*(15), 6056–6061.

Šali, A., & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, *234*, 779–815.

Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., & Gottesman, M. M. (2007). Silent Polymorphisms Speak: How They Affect Pharmacogenomics and the Treatment of Cancer. *Cancer Research*, *67*(20), 9609–9612.

Scacchi, R., Gambina, G., Moretto, G., & Corbo, R. M. (2009). Variability of AChE, BChE, and ChAT genes in the late-onset form of Alzheimer's disease and relationships with response to treatment with Donepezil and Rivastigmine. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150B*(4), 502–507.

Schwartz, T. W., & Holst, B. (2007). Allosteric enhancers, allosteric agonists and ago-allosteric modulators: Where do they bind and how do they act? *Trends in Pharmacological Sciences*, *28*(8), 366–373.

Soreq, H., Ben-Aziz, R., Prody, C. A., Seidman, S., Gnatt, A., Neville, L., ... Lipidot-Lifson, Y. (1990). Molecular cloning and construction of the coding region for human acetylcholinesterase reveals a G + C-rich attenuating structure. *Proceedings of the National Academy of Sciences*, *87*(24), 9688–9692.

Sterling, T., & Irwin, J. J. (2015). ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, *55*(11), 2324–2337.

Tam, N. M., Vu, K. B., Vu, V. V., & Ngo, S. T. (2018). Influence of various force fields in estimating the binding affinity of acetylcholinesterase inhibitors using fast pulling of ligand scheme. *Chemical Physics Letters*, *701*, 65–71.

Tan, E. C., Johnell, K., Garcia-Ptacek, S., Haaksma, M. L., Fastbom, J., Bell, J. S., & Eriksdotter, M. (2018). Acetylcholinesterase inhibitors and risk of stroke and death in people with dementia. *Alzheimer's & Dementia*, *14*(7), 944–951.

The International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), 928–933.

Trott, O., & Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, NA–NA.

Valle, A. M., Radic, Z., Rana, B. K., Mahboubi, V., Wessel, J., Shih, P.-a. B., ... Taylor, P. (2011). Naturally Occurring Variations in the Human Cholinesterase Genes: Heritability and Association with Cardiovascular and Metabolic Traits. *Journal of Pharmacology and Experimental Therapeutics*, *338*(1), 125–133.

Vigny, M., Gisiger, V., & Massoulie, J. (1978). "Nonspecific" cholinesterase and acetylcholinesterase in rat tissues: Molecular forms, structural and catalytic properties, and significance of the two enzyme systems. *Proceedings of the National Academy of Sciences*, *75*(6), 2588–2592.

Wang, D., Song, L., Singh, V., Rao, S., An, L., & Madhavan, S. (2015). SNP2Structure: A Public and Versatile Resource for Mapping and Three-Dimensional Modeling of Missense SNPs on Human Protein Structures. *Computational and Structural Biotechnology Journal*, *13*, 514–519.

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–270.

Wardlaw, J., Warlow, C., & Counsell, C. (1997). Systematic review of evidence on thrombolytic therapy for acute ischaemic stroke. *The Lancet*, *350*(9078), 607–614.

Wilkinson, D. G., Francis, P. T., Schwam, E., & Payne-Parrish, J. (2004). Cholinesterase Inhibitors Used in the Treatment of Alzheimer???s Disease: The Relationship Between Pharmacological Effects and Clinical Efficacy. *Drugs & Aging*, *21*(7), 453–478.

Yue, P., & Moult, J. (2006a). Identification and Analysis of Deleterious Human SNPs. *Journal of Molecular Biology*, *356*(5), 1263–1274.

Yue, P., & Moult, J. (2006b). Identification and Analysis of Deleterious Human SNPs. *Journal of Molecular Biology*, *356*(5), 1263–1274.

Zhang, L.-F., Yang, J., Hong, Z., Yuan, G.-G., Zhou, B.-F., Zhao, L.-C., ... Wu, Y.-F. (2003). Proportion of Different Subtypes of Stroke in China. *Stroke*, *34*(9), 2091–2096.