IN-SILICO INVESTIGATION OF THE EFFECTS OF GENETIC MUTATIONS ON THE STRUCTURAL DYNAMICS OF THIOPURINE S-METHYLTRANSFERASE AND THEIR IMPLICATIONS ON THE METABOLISM OF 6-MERCAPTOPURINE

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

of

Master of Science in Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY AND MICROBIOLOGY

Faculty of Science

by

Rehema Mukami Mwaniki

22M0005

FEBRUARY 2023

ABSTRACT

Thiopurine S-methyltransferase (TPMT) is a cytosolic enzyme that catalyzes the S-methylation of aromatic and heterocyclic sulfhydryl compounds such as 6-mercaptopurine (6MP), 6-thioguanine (6TG) and azathioprine (AZA) which is first converted to 6MP through reduction by glutathione S-transferases (GST). The compounds, generally referred to as thiopurines, are immunosuppressants used to treat childhood acute lymphoblastic leukemia (ALL), autoimmune disorders and transplant rejection. Thiopurines are prodrugs which require metabolic activation to give thioguanine nucleotides that exert their cytotoxic effects by incorporation into DNA or inhibiting purine synthesis. The methylation reaction by TPMT utilizing S-adenosylmethionine (SAM) as the methyl donor prevents their conversion to these toxic compounds. The catalytic activity of TPMT in metabolising these compounds has been associated with occurrence of genetic variations. The variations that result to missense mutations cause amino-acid changes and in turn alter the polypeptide sequence of the protein. This could alter functionality and structural dynamics of the enzyme. This study sought to understand the underlying mechanism by which 7 specially selected mutations impede metabolic activity of the enzyme on 6-MP using *in silico* techniques. VAPOR and PredictSNP were used to predict the effects of single nucleotide polymorphisms (SNPs) on the stability and function of the enzyme. Of the 7 mutations, only H227Q was predicted to be functionally benign while the rest (L49S, L69V, A80P, R163H, R163C and R163P) were predicted to be deleterious or associated with disease. All the SNPs were predicted to destabilize the enzyme. Molecular dynamics (MD) simulations were preformed to mimic the behaviour of the apo, holo and drug-bound WT and mutant enzymes in vivo. This was followed by post-MD analysis to identify changes in the local and global motions of the protein in the presence of mutations and changes in intra-protein communication networks through contact map and centrality metrics calculations. RMSD and Rg analyses were performed to assess changes in global motions and compactness of the enzyme in the apo, holo and drug-bound states and in the presence of mutations. These revealed that binding of the ligand had a stabilizing effect on the WT enzyme evident from more steady trends from the analyses across trajectories in the holo and drug-bound enzymes compared to the apoenzyme. The occurrence of mutations had an effect on the global motions and compactness of the enzyme across the trajectories. Most mutations resulted in destabilized systems and less compact structures shown by unsteady RMSD and Rg across trajectories respectively. The drug-bound systems appeared to be more stable in most of the systems meaning that the binding of 6MP stabilized the enzyme regardless of the presence of a mutation. RMSF analysis recorded local

changes in residue flexibility due to the presence of mutations in all the systems. All the drug-bound mutant systems lost flexibility on the αA helix which caps the active site. This could have an effect on drug binding and result to defective drug metabolism. The A80P mutation resulted to a more rigid structure from both global and local motions compared to the WT enzyme which could be associated with its nearly loss of function in vivo and in vitro. Dynamic cross correlation calculations were performed to assess how the atoms moved together. Correlated, anti-correlated and areas of no correlations were recorded in all the systems and in similar places when compared to each other. This meant that occurrence of mutations had no effect on how the atoms moved together. Contact map analysis showed that occurrence of mutations caused changes in interactions around the positions where the mutations occurred, which could have an effect on protein structural dynamics. The A80P substitution which occurred on the surface away from the binding site was identified as an allosteric mutation that resulted to changes in the catalytic site. Contact maps for the drug-cofactor complex in the mutant systems in comparison with the WT protein revealed changes that could suggest reorientation of the drug at the catalytic site. This could be an implication to altered drug metabolism. Eigenvector centrality (*EC*) and betweenness centrality *(BC)* for the most equilibrated portions of the trajectories were calculated for all the studied systems to identify residues connected to the most important residues and those that were spanned the most in shortest paths connecting other residues. Areas that scored highest in these metrics where mostly found in regions surrounding the catalytic site. Top 5% centrality hubs calculations showed loss of major hubs due to mutations with gaining of new ones. This means that mutations affected communication networks within the protein. The gained hubs were in areas close-by the lost ones which could have been an attempt of the protein to accommodate the mutations. Persistent top 5% *BC* hubs were identified at positions 90 and 151 while one persistent top 5% *EC* hub was identified at position 70. This positions play important roles in shaping the catalytic site and are in direct contact with the ligands. It was concluded that *in silico* techniques and analysis applied in this study revealed possible mechanisms in which genetic variations affected the structural dynamics of TMPT enzyme and affected 6MP metabolism.

DECLARATION

I, Rehema Mukami Mwaniki, declare that this thesis entitled, *In-silico investigation of the effects of genetic mutations on the structural dynamics of thiopurine s-methyltransferase and their implications on the metabolism of 6-mercaptopurine.* submitted to Rhodes University is solely my own research work. I have acknowledged all authors' concepts and referenced direct quotations from their works. I also declare that this thesis has never been submitted to any different institution for whatever degree.

Signature .....

Date 4/7/2023

DEDICATION

To my parents, Mr and Mrs Mwaniki. You have taught to be limitless in my aspirations because nothing I can perceive I can't achieve.

ACKNOWLEDGEMENTS

I would like to extend my since gratitude to my supervisor Prof Ozlem Tastan Bishop for offering me the opportunity to learn at RUBi and work on this project. Your encouragement, patience and guidance shed light to my often flickering hope of completion. You were always behind me ensuring that I waded through the project with the motive to publish. You are a mentor! To my co-supervisor Dr Taremekedzwa Allan Sanyanga, I might not have been the best student but your unwavering help, support and suggestions throughout the course of this project always thrilled me beyond appreciation. Tons of gratitude to all the postgraduate lectures: Dr Olivier Sheik Amamuddy, Dr Kevin A.Lobb and Mr. Jeremy Baxter. To my MSc classmate Yamkela Ntantiso, you were an amazing study buddy, you made the long nights of toiling to meet up assignment deadlines bearable. A standing ovation to the RUBi team: Lillian, Chiraditzo, Jessika, Victor, Zacharia, Tendai, Queennie, Fisayo and Wade. Thanks for the company, holistic uplifts and smiles. You made research sound fun, and sure it was! Special mentions to Lillian Mbaisi and Dr Bilinga M Tendwa for your invaluable friendship and ensuring that I was keeping sane and sensible.

My God sent family while in South Africa, Dr Lorna Chebon Bore and Nelson Kibichii Bore, your encouragement, love and consistent prayers held me in faith and nourished my heart. My other family Iris Kuja and Dr Stanley Kuja, I could not have made it to the finish line without your support and advise. You made life easy for me.

To my champions, my loving parents, you supported me financially, spiritually and emotionally. Even though miles separated us, your cherishing love kept me moving. Your strong hands continue to hold me in every step I take. My sister and my brother, thank you for blazing the light for me to follow and challenging me to dare chase after my dreams. Dr Harris Onywera, thank you for being my go-to person whenever my system was crashed, you knew the correct code to run in me ensuring I compiled properly.

As far as the heights reach from the depths, as far as the east is from the west, so far Your grace has carried me. When my courage ended, my heart always found strength in Your presence. *My God You're faithful ...all Your promises are yes and amen!*

Table of Contents

List of Figures

LIST OF TABLES

LIST OF EQUATIONS

LIST OF SUPPLEMENTARY MATERIAL

LIST OF WEB SERVERS

Ensembl	**https://www.ensembl.org/index.html**
PharmGKB	**https://www.pharmgkb.org/**
H++	**http://biophysics.cs.vt.edu**
HUMA	**https://huma.rubi.ru.ac.za**
RCSB	**https://www.rcsb.org**
Uniprot	**https://www.uniprot.org/**
VAPOR	**https://huma.rubi.ru.ac.za/#vapor**
PredictSNP	**https://loschmidt.chemi.muni.cz/predictsnp/**

LIST OF ABBREVIATIONS

6MP	6-mercaptopurine
ALL	Acute Lymphoblastic Leukaemia
AMBER	Assisted Model Building with Energy Refinement
AMLs	Acute Myeloid Leukemias
ATP	Adenosine Triphosphate
BC	Betweenness Centrality
CC	Closeness Centrality
CHARMM	Chemistry at Harvard Macro-molecular Mechanics
CHPC	Center for High Performance Computing
CNS	Central Nervous System
CPU	Central Processing Unit
CSV	Comma Separated Value
DC	Degree Centrality
DCC	Dynamic Cross Correlation
DNA	Deoxyribonucleic Acid
DRN	Dynamic Residue Network
EC	Enzyme Commission
EC	Eigenvector Centrality
FASTA	Fast-all
GROMACS	Groningen Machine for Chemical Simulations
GROMOS	Groningen Molecular Simulation
GTP	Guanosine Triphosphate
HIF-1α	Hypoxia Inducible Factor 1α
HK	Hexokinase

HPRT	Hypoxanthine Guanine Phosphoribosyltransferase
IMPDH	Inosine Monophosphate Dehydrogenase
L	Average Shortest Path
MD	Molecular Dynamics
mTOR	Mammalian Target of Rapamycin
NCBI	National Center for Biotechnology Information
NR4A	Nuclear Receptor Subfamily 4A
nsSNP	Non-synonymous Single Nucleotide Polymorphism
OPLS	Optimized Potentials for Liquid Simulations
OXPHOS	Oxidative Phosphorilation
PANTHER	Protein Analysis Through Evolutionary Relationships.
PBC	Periodic Boundary Conditions
PDB	Protein Data Bank
PhD-SNP	Predictor of Human Deleterious SNPs.
pI	Isoelectric Point
PI3K	Phospatidylinositol Kinase
PolyPhen	Polymorphism Phenotyping
PPI	Protein-protein Interactions
PPP	Pentose Phosphate Pathway
PROVEAN	Protein Variation Effect
RCSB	Research Collaboratory for Structural Bioinformatics
Rg	Radius of Gyration
RIN	Residue Interaction Networks
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction

SAH	S-adenosylhomocysteine
SAM	S-adenosylmethionine
SAM-MTs	S-adenosylmethionine Methyltransferases
SGNA	Statistically Guided Network Analysis
SIFT	Sorting Intolerant From Torelant
SMP	S-methyl Mercaptopurine
SN_2	Nucleophilic Reaction
SNAP	Screening for Non-acceptable Polymorphisms
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SVM	Support Vector Machines
TGNs	Thioguanine Nucleotides
TIMP	Thioinosine mono-phosphate
TPMT	Thiopurine S-methyl Transferase
VSC	Visual Studio Code

TABLE OF AMINO ACIDS

Full Name	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

"Do not go where the path may lead, go instead where there is no path and leave a trail"

CHAPTER ONE

LITERATURE REVIEW

1.1 Acute Lymphoblastic Leukaemia

1.1.1 What is Acute Lymphoblastic Leukaemia?

Acute Lymphoblastic Leukaemia or Acute Lymphocytic Leukaemia (ALL) is a type of cancer resulting from immature lymphocytes in the bone marrow [1]. It encompasses a group of lymphoid neoplasms that morphologically resemble B-lineage and T-lineage precursor cells [2].

1.1.2 Etiology of ALL

The development of leukaemia requires a hematopoietic stem cell [3] or one of its committed progenitors to escape normal mechanisms of homeostatic control that regulate growth-factor signalling, differentiation, apoptosis, and self-renewal [4,5]. ALL presents primarily as a de novo disease [2,6]. The exact pathogenetic events leading to the development of ALL are not known [7]. It is thought to occur after damage to DNA causes lymphoid cells to undergo uncontrolled growth and spread throughout the body [8]. Some genetic and environmental factors have been related to ALL. A few cases (<5%) have been associated with inherited predisposing genetic syndromes such as Down syndrome, Bloom syndrome, neurofibromatosis type I, and ataxia-telangiectasia [7]. Additionally, exposure in utero to ionizing radiation, pesticides, solvents or to specific chemotherapeutic drugs has been linked to an increased risk of childhood ALL [7,9]. The clinical onset of ALL is most often acute with a small percentage of cases evolving insidiously over several months [7]. The most common symptoms include fever, fatigue and lethargy, bone and joint pain, and a bleeding diathesis [2].

1.1.3 Epidemiology

Most ALL cases occur in children, with an incidence of 3-4/100,000 in patients below the age of 14 and ~1/100,000 in patients older than 15 years, in the United States. In children, ALL represent 75% of all acute leukemias, which in turn represents 34% of all cancers in this age group. The peak incidence of ALL is between 2-5 years of age. This percentage is much lower in adults, in whom acute myeloid leukemias (AMLs) and chronic lymphocytic leukemias are more common. There is a

slight male predominance in all age groups and a significant excess incidence among Caucasian children [2,10] in the United States. Globally, ALL typically occurs more often in Caucasians, Hispanics, and Latin Americans than in Africans [11].

1.1.4 Diagnosis and Treatment

In diagnosis, cytochemical staining techniques have often been used [2]. With the development of immunophenotyping techniques, the frequency with which cytochemical staining has been used is decreasing [2]. Immunophenotyping of leukaemic lymphoblasts by flow cytometry is essential to establishing the correct diagnosis and defining cell lineage [7]. Chromosomal analysis and other highly specific and sensitive techniques–such as reverse transcriptase polymerase chain reaction (RT-PCR), fluorescence in-situ hybridisation, and flow cytometry–are used to detect specific fusion transcripts, gain or loss of cellular DNA content, or specific chromosomes with prognostic or therapeutic relevance [7].

Treatment of ALL aims to induce a lasting remission. Clinical remission can now be induced in 96-99% of children and 78-93% of adults [12]. Possible treatments include chemotherapy, steroids, radiation therapy, intensive combined treatment (including bone marrow or stem cell transplants), targeted therapy, and/or growth factors [13]. Chemotherapy is the initial treatment of choice, where most people with ALL receive a combination of medications. It consists of three phases: remission induction, intensification, and maintenance therapy [14].

i. **Remission-induction phase**

The goal of remission-induction treatment is to eradicate more than 99% of the initial leukaemic cell burden and to restore normal haemopoiesis and healthy performance status. This approach typically includes the administration of a glucocorticoid (prednisone or dexamethasone), vincristine, and at least a third drug (asparaginase, anthracycline, or both) [7]. In adults, this also includes Central nervous system (CNS) prophylaxis therapy. This involves cranial radiation therapy plus intrathecal (IT) methotrexate, high-dose systemic methotrexate and IT methotrexate without cranial radiation therapy and IT chemotherapy alone [14].

ii. **Consolidation (intensification) treatment**

With the restoration of normal haemopoiesis and body function, intensification treatment is generally used to eradicate drug-resistant residual leukaemic cells, thus reducing the risk of relapse

[7]. CNS prophylaxis in adults sometimes continues even to this phase [14]. Frequently used strategies include high-dose methotrexate plus mercaptopurine, re-induction treatment with the same agent that was given initially, frequent pulses of vincristine and corticosteroid plus high-dose asparaginase for 20–30 weeks, and an augmented regimen consisting of reinduction treatment and additional doses of vincristine, asparaginase, and intravenous methotrexate during periods of myelosuppression [7].

iii. **Maintenance therapy**

Patients with ALL need continuation treatment to prevent or forestall relapse. Although about two-thirds of childhood cases can be treated successfully with only 12 months of therapy, they cannot be identified prospectively with any degree of certainty [15]. Hence, all patients receive chemotherapy for 2-2.5 years. Daily mercaptopurine and methotrexate every week constitute the backbone of continuation regimens [8].

1.1.5 Metabolism of 6-Mercaptopurine and its mode of action

Thiopurines were originally developed in the 1950s for the treatment of ALL in children by Gertrude Elion and George Hitchings [16,17]. The first step in the intracellular activation of 6-mercaptopurine (6MP) involves being catalyzed by hypoxanthine-guanine phosphoribosyltransferase (HPRT), which yields thioinosine monophosphate (TIMP) [18]. Subsequently, TIMP is converted to thioguanine triphosphate nucleotides by inosine monophosphate dehydrogenase (IMPDH) [19], which are incorporated into DNA [20]. Thiopurine s-methyltransferase (TPMT) is responsible for the last step in the formation of methylated TIMP, which inhibits de novo purine synthesis and leads to the lack of endogenous nucleotides needed for replication [21]. Figure 1.1 illustrates the general metabolic pathway of 6MP with a secondary metabolic pathway for 6MP by TPMT [22]. This secondary metabolic pathway leads to subsequent conversion of 6MP into inactive S-methyl-mercaptopurine thereby decreasing the amount of pro-drug available for active thioguanine nucleotides (TGN) [20].

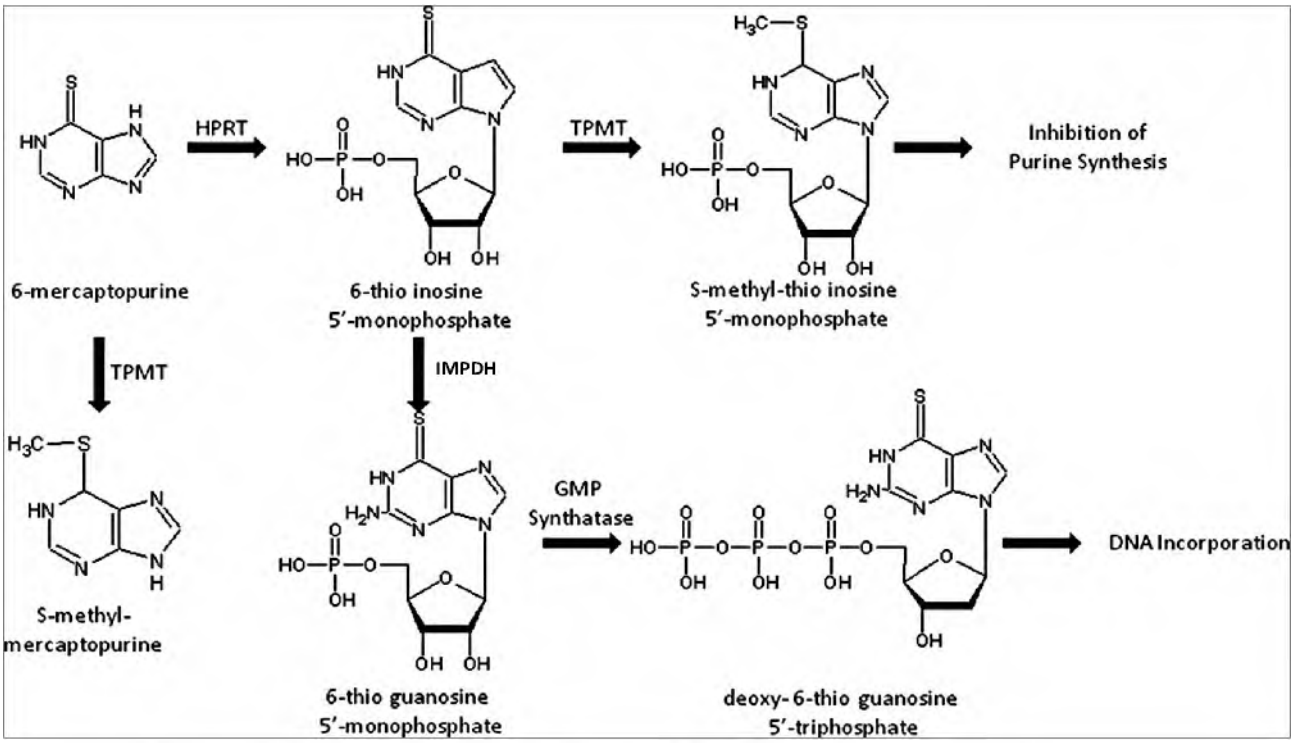


Figure 1.1: Pathway of thiopurine metabolism (Adapted from Murugesan et al., 2010).

Numerous mechanisms are involved in the antiproliferative effects of 6MP, which include:

- Its incorporation into nucleic acids chains (leading to DNA damage and apoptosis) and its ability to inhibit de novo purine synthesis [23–26], which is crucial for lymphocyte proliferation [25,27].
- Its inhibition of de novo biosynthesis of adenosine triphosphate (ATP) and guanosine triphosphate (GTP) [28]. Fernández-Ramos [29] demonstrated that 6MP promotes energetic failure in proliferating T-cells by causing a rapid depletion of ATP.
- Evidence indicates that 6MP inhibits the phosphatidylinositol-3-kinase (PI3K)/mammalian target of the rapamycin (mTOR) signalling pathway [30], one of the most commonly activated signalling pathways in cancer, leading to cell proliferation, survival, and differentiation [23]. This suggests that the drug might interfere with metabolic checkpoints and impact metabolic reprogramming in normal T cells and cancer cells [31].
- In relation to its possible role in cell metabolic reprogramming, 6MP regulates the activity of members of the orphan nuclear receptor NR4A family, which act as key transcriptional regulators of glucose and lipid metabolism [32].

- 6MP modifies the transcriptional activity of hypoxia-inducible factor 1α (HIF-1α) [33] and inhibits enzymes involved in glycolysis, such as phosphofructokinase 2 (PFK2) and hexokinase (HK) [34].

1.1.5.1: Role of metabolic pathways inhibited by 6MP in cancer

Metabolic reprogramming promotes glycolysis, glutaminolysis, and biosynthesis of nucleotides and lipids to support cell growth and proliferation [31,35] which is a major feature of cancer cells. To produce ATP, proliferating cells shift glucose metabolism from oxidative phosphorylation (OXPHOS) to aerobic glycolysis [36–38]. The process far is less efficient than OXPHOS [31,35,36] but generates biosynthetic precursors through the pentose phosphate pathway (PPP) [39] thereby facilitating the production of a pool of purine and pyrimidine nucleotides to support cancer cell growth and proliferation [29,35]. AMP-activated protein kinase (AMPK), mTOR and the oncogenes Myc and HIF-1α are critical regulators of metabolic reprogramming in proliferating cells and are referred to as metabolic checkpoints [29,35]. Although they are effective drugs, thiopurines have a relatively narrow therapeutic index and are capable of causing life-threatening drug-induced toxicity [40,41].

1.1.6 Pharmacogenomics in ALL treatment

Pharmacogenomics is the study of how human genetic information impacts drug response and aims to improve efficacy and reduce side effects [42]. Genetic polymorphisms of drug transporters, receptors, targets, and drug-metabolising enzymes can affect the effectiveness and toxic effects of antineoplastic drugs [4,7]. Key genetic variations need to be combined with clinical variables and responses to therapy to avoid over- or under-treatment [43]. Such variations could be in genes encoding for drug metabolising enzymes like thiopurine S-methyltransferase (TPMT), glutathione S-transferase, cytochrome P450 3A4 and methylene-tetrahydrofolate reductase [2]. Despite the promise of pharmacogenetic studies to enhance treatment outcome in ALL, only polymorphisms and the activity of TPMT—an enzyme that catalyses S-methylation (inactivation) of thiopurines—have been useful in clinical practice [44–46]. About 10% of the total population inherit one wild-type gene encoding thiopurine methyltransferase and one non-functional variant allele, resulting in intermediate enzyme activity, whereas 1 in 300 people inherit two non-functional variant alleles with no enzyme activity. When treated with conventional doses of thiopurines, up to half of the patients with the heterozygous deficiency and all homozygous-deficient patients develop haemopoietic toxic effects, which can be fatal in the homozygous group [44]. The enzyme

deficiency also confers a high risk of developing therapy-related acute myeloid leukaemia [44] and radiation-induced brain tumours, in the context of intensive thiopurine treatment [47]. Conversely, patients with high levels of enzyme activity might be at greater risk of relapse owing to decreased exposure of leukaemic cells to active drug metabolites [46].

1.2 Thiopurine S-methyltransferase (TPMT)

TPMT (EC 2.1.1.67, 245 amino acids, 28180 g mol^{-1}), first described by Remy [48] is a cytosolic single-domain protein that belongs to Class I of the S-adenosylmethionine dependent methyltransferases (SAM-MTs) [49]. This class of SAM-MTs have a highly conserved Rossman-like α/β core fold topology consisting of a central 7-stranded β-sheet flanked by 3 helices on each side [49,50]. Some deviations from the consensus topology along with multiple insertions of structural elements are evident. Structural deviations from the consensus topology are common and found in similar locations [51]. The TPMT structure shown in Figure 1.2 is a monomeric protein composed of a nine β sheets core flanked on each side by three helices (αB, αC, αD and αE, αF, αG). Two additional helices, αA and αH, are inserted into the methyltransferase fold [52] as seen in Figure 1.2a & b.

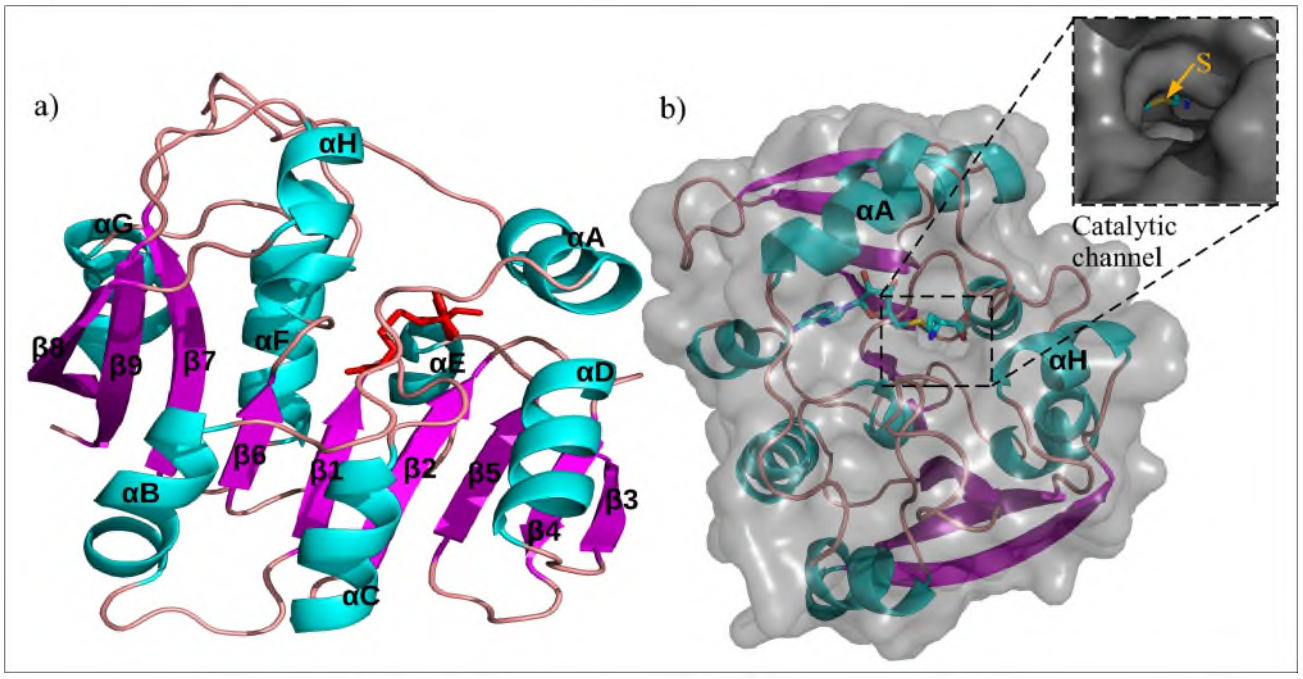
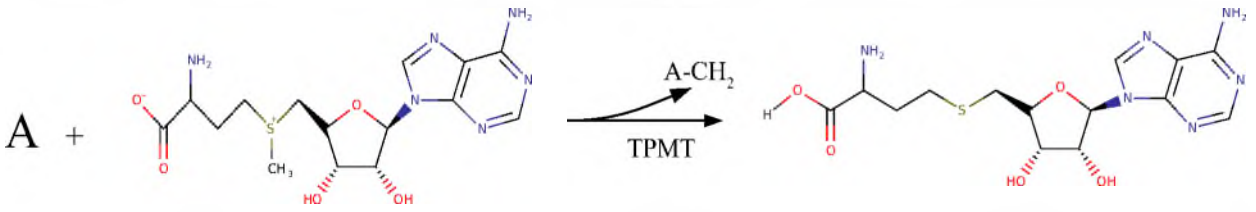


Figure 1.2: a) Topology of TPMT. The cartoon model of the enzyme (2BZG) with SAH (red sticks) generated on Pymol. The alpha helices are coloured cyan and labelled αA-H and the beta sheets are coloured magenta and labelled β1-9. **b)** The surface outline of the cartoon model showing the catalytic cavity. SAH is seen represented as cyan sticks buried inside the catalytic cavity with the sulphur atom (yellow) that transfers its methyl group to a methyl acceptor pointing towards the opening of the channel

The TPMT reaction takes two substrates, a methyl donor, S-adenosyl-L-methionine (SAM) and a methyl acceptor (A), then generates S-adenosyl-L-homocysteine (SAH) and a methylated product as shown in the reaction equation in Figure 1.3 [53].



SAM SAH

Figure 1.3: The conversion of SAM to SAH by TPMT with the transfer of a methyl group to a methyl acceptor as drawn on MarvinSketch.

A methyl group is transferred from the sulphur atom of SAM to the sulphur atom of the accepting compound [52] by a nucleophilic reaction-like mechanism where the sulfur atom of the acceptor is the nucleophile attacking the methyl group on SAM [54]. 6-Mercaptopurine (6MP), 6-thioguanine (6-TG), thioguanine monophosphate (6-TGMP) and thioinosine mono-phosphate (6TIMP) are the only four identified methyl acceptors of TPMT [55].

The protein thus has two site receptors: the first site is for the co-substrate SAM surrounded by residues in helices αA, αE, and β strands 1 and 2 and is completely buried in the structure by αA which caps the active site, as seen in Figure 4b. Thiopurine binding site is surrounded by residues in helices αA, αI, αH and the β7–αG loop and is capped by αA (Figure 1.4b) [53]. Wu et al. identified a small water channel flanked by F40 and P196 (Figure 4a) that facilitated binding of 6MP onto the active site [52].

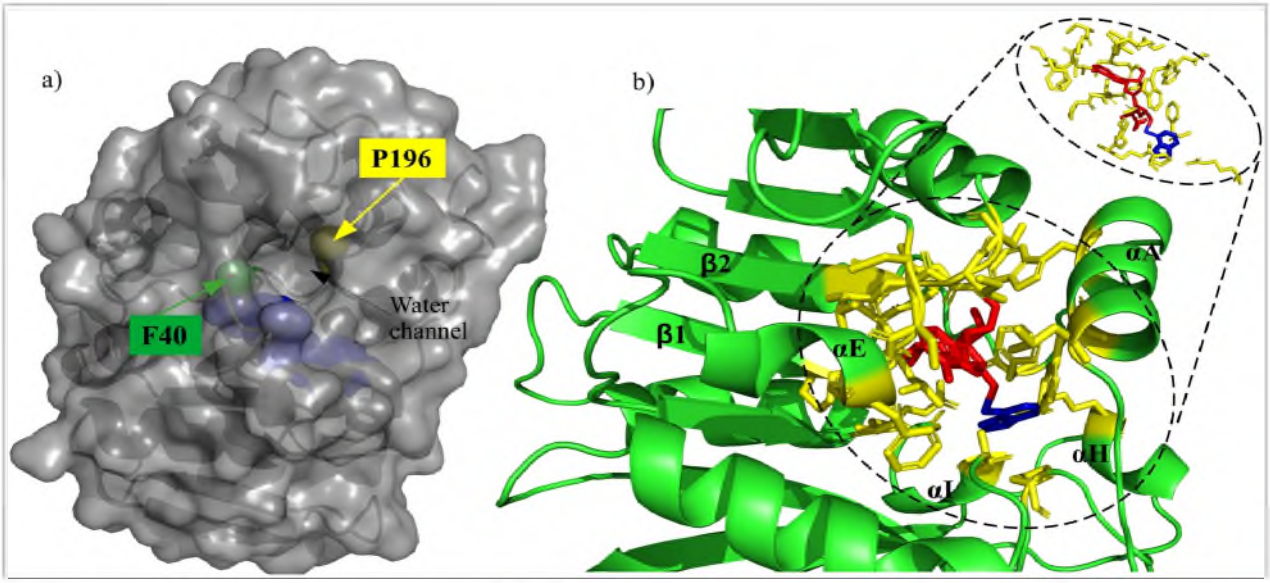


Figure 1.4: a) A surface representation of TPMT with its cartoon structure underneath. The location of F40 (green) and P196 (yellow) flanking the water channel have been identified by their respective locations and single letter representations. SAH is seen buried into the enzyme and coloured blue in sphere representation. **b)** Secondary structures surrounding SAH and 6MP. SAH is shown as red sticks while 6MP is shown as blue sticks. Surrounding residues are represented as yellow sticks and secondary structures as green cartoon.

TPMT catalyzes the S-methylation of aromatic and heterocyclic sulfhydryl compounds or thiopurines [52] as a phase II enzyme [56]. The compounds are widely used to treat ALL, inflammatory bowel disease, autoimmune disorders and organ transplant recipients to minimize graft rejection [53]. These compounds are pro-drugs which are extensively metabolized to give thioguanine nucleotides (TGNs) that exert their cytotoxic effects by incorporation into DNA or inhibiting purine synthesis [52]. Specifically, 6MP is widely prescribed to block tumour growth in childhood ALL [29] as previously seen.

1.2.2 Single nucleotide polymorphisms in the TPMT gene

TPMT genetic polymorphisms represent a striking example of the importance of pharmacogenetics for individualized drug therapy [40]. The TPMT gene, located in chromosome 6p22.3, is approximately 34 kb in length and consists of 10 exons and 9 introns [21]. Around 46 allelic variants of the gene have been reported to the TPMT nomenclature committee and assigned allele names [41,57,58]. These variants result in missense mutations and more than half of them have been associated with reduced or deficient enzymatic activity [19,40]. Among the variants, the most frequent one is TPMT*3A, which is common in Caucasians (5 % frequency). The second most frequent variant is TPMT*3C, which is more common in Asian and African populations (2 %

frequency). Some other variants which show considerable presence in different groups are TPMT*3B, TPMT*3D, TPMT*2, TPMT*5, TPMT*6 and TPMT*8 [56,59,60]. The distribution of these alleles has been illustrated in Figure 1.5.

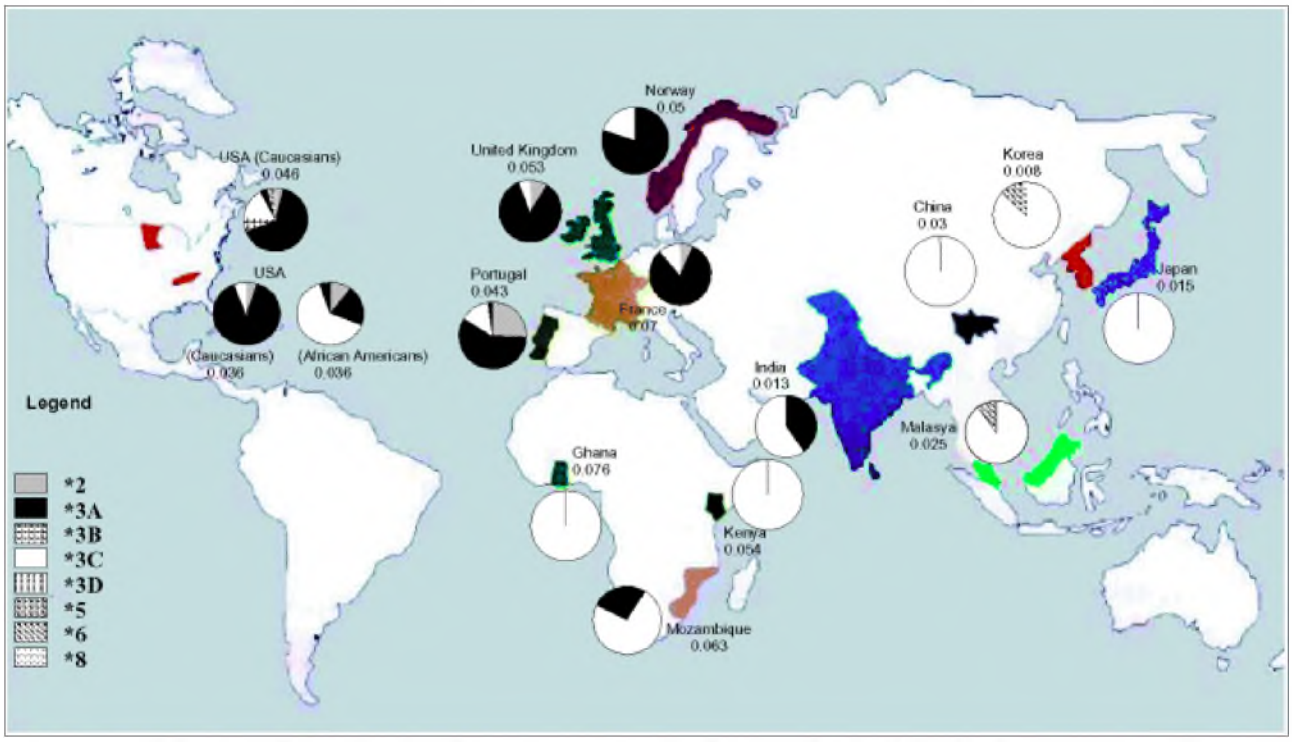


Figure 1.5: Pattern of the global distribution of the most common TPMT mutant alleles. For each population, the overall frequency of deficient alleles is indicated below the corresponding label (Adapted from Alves et al., 2004).

Table 1.1 summarizes the pathogenic features associated with these common alleles.

Table 1.1: The most common TPMT alleles, the resulting amino-acid change and pathogenetic features.

Allelic variant	Variant SNP	Amino acid change	Pathogenic feature
TPMT *2	G238C	Ala80Pro	100-fold decrease in TPMT activity and very low levels of protein in vitro.
TPMT *3A	G460A A719G	Ala154Thr Tyr240Cys	400-fold decrease in levels to no detectable enzyme activity and aberrant protein folding and aggresome formation in yeast cells. Deficient activity in vivo.
TPMT *3B	G460A	Ala154Thr	4-fold decrease in protein levels with no catalytic activity.
TPMT *3C	A719G	Tyr240Cys	Decreased enzymatic activity.
TPMT *5	T146C	Leu49Ser	Nearly complete loss of function due to local changes in the active site rather than reduced structural stability
TPMT *6	A539T	Tyr180Phe	Reduced activity.
TPMT *8	G644A	Arg215His	Intermediate activity in vivo and in vitro.

Residues in conserved positions are considered to be important for protein structure and function [55]. Substituting these residues, especially with residues with different physicochemical properties might lead to deleterious effects [55,61]. Other factors such as proximity to the active site and the residue's role in protein folding have been used to explain the effects of amino acid substitutions on enzyme activity and dynamics [17,51]. The exact mechanism underlying these deleterious effects is still not fully elucidated [62].

1.3 Problem Statement

Childhood ALL is the most common cancer in children. The purine analogues mercaptopurine and thioguanine are drugs that have been used successfully as cytotoxic agents in the treatment of leukaemias for more than 50 years [22]. Because the therapeutic window between toxicity and efficacy is narrow, thiopurine drugs can potentially have life-threatening toxicity [41]. The occurrence of deleterious polymorphisms in the drug metabolising enzyme, TPMT, makes this therapeutic window even more narrow and poses a high risk of serious adverse drug reactions [21,60]. Identifying the presence of record and/or novel variants is important in enhancing individualized therapy. Even more important is understanding the mechanisms by which these variations affect enzymatic activity at a molecular level. Studies have been carried out to functionally characterize variant alleles in TPMT in vitro and in vivo [19,40,61,63–65]. Only a few reports illustrating the effects of mutations on TPMT enzyme structure and function using computational approaches exist [17,55,66]. Understanding how mutations might affect inter-residue interactions, ligand binding, structure dynamics and thus enzyme function is paramount in getting the full picture of the role of pharmacogenomics in drug metabolism during ALL chemotherapy and TPMT enzyme kinetics. This study will apply in silico techniques to decipher these effects.

1.4 Hypothesis

This study hypothesizes that the effects of genetic variations in TPMT can be identified through *in silico* techniques by incorporation of molecular dynamics (MD) for use in precision medicine and drug discovery.

1.5 Aim

The study aims to apply in silico techniques to analyse the effects of mutations on the structural dynamics of the TPMT enzyme and functional implications on the metabolism of 6MP.

1.6 Objectives

- To retrieve genetic variation data from Human Genetic Mutation databases and protein sequence and structure of TPMT from Universal Protein Resource (UniProt) and Protein Data Bank (PDB) respectively.

- To identify clinically important variants and introduce the resulting mutations into the wild-type (WT) apo, holo and drug-bound protein structures.
- To perform MD simulations on these structures and analytically compare the trajectory calculations in both wild-type and mutant protein.
- To carry out dynamic residue network (DRN) analysis and to identify the important residue interactions which are affected in presence of mutations and compare with WT systems.
- To explicate the impact of selected mutations on the structural dynamics of TPMT and their functional implications on 6MP metabolism and toxicity.

CHAPTER TWO

VARIANT ANALYSIS AND STRUCTURAL MAPPING OF 6MP ONTO THE ACTIVE SITE OF HUMAN TPMT

2.1 Introduction

Single-base substitutions (SBSs) resulting from point mutations can alter pre-mRNA processing, amino-acid sequence and the ultimate structure of the protein [67,68]. When these point mutations or single nucleotide variations (SNVs) are observed with at least 1% frequency within a population, they are then referred to as single nucleotide polymorphisms (SNPs) [69]. SNPs are the most common type of genetic variation [70] and occur normally throughout the human DNA sequence at a frequency of 1 on every 300 nucleotides [71].

When occurring in the gene coding regions, SNPs can be synonymous also known as silent (that is, not causing a change in the amino acid) or non-synonymous (when the amino acid is altered) [72]. Non-synonymous SNPs (nsSNPs) can either result in an amino acid substitution known as missense mutation which can alter protein function and structure dynamics, or cause the introduction of a stop/start codon referred to as nonsense mutation. This results in a truncated/extended polypeptide sequence [71].

SNPs that occur in functionally important residue positions such as binding sites or conserved regions are speculated to have deleterious effects on protein structural dynamics and function [40,65,73].

There is a need to effectively and efficiently identify functionally important variants which may be deleterious or disease-causing and to identify their molecular effects [74]. Predicting the effects of SNPs is important in gaining insight into their impact on gene function and the protein's structure-function relationship [62,71]. Various tools exist for such analysis [71,75,76]. Such tools include and are not limited to PROVEAN (Protein Variation Effect Analyzer) [77], PolyPhen-2 [78], PhD-SNP [79], SIFT (Sorting Intolerant From Tolerant) [80], which predict the functional impact of amino acid substitutions [75,81,82], I-Mutant [76] and MuPro [83] which predict the effect of the substitutions to protein stability.

Methods in the existing literature for SNP prediction have used a large range of structure-, sequence-based or both approaches to separate deleterious from neutral SNPs [84]. Protein sequence-based predictions identify important residues using sequence homology and evolutionary conservation information [85,86]. The structure-based approach concentrates on the protein's mechanistic attributes and stability, providing an understanding of the phenotypical effect of the SNPs. SIFT [80,82], SNAP [87], MutPred [74] and PANTHER [88] are examples of prediction tools that use the sequence-based approach, while ERIS [89] and CUPSAT [90] use the structure-based approach. Sequence-based methods could have an advantage over structure-based approaches as protein sequences are readily available while crystal structures are not readily available for all proteins. An integration of both approaches as used in PROVEAN [77], PolyPhen [78] and Mupro [83] have the advantage of cross-referencing the results from both approaches, thereby providing wider coverage of the different aspects of SNP analysis. As different tools use different algorithms and different datasets to perform data training, none of the methods is completely accurate nor do they return similar results [84]. The choice in training data can significantly affect SNP classification and estimated error rates [84]. To solve this problem, the best approach is to get a consensus from several different prediction tools before proceeding with further SNP analysis. Web servers such as variant analysis portal (VAPOR) [91] and PredictSNP [92] can be used for comprehensive prediction results [93].

VAPOR, a program in Human Mutation Analysis (HUMA) [91] web server, compiles prediction from PROVEAN, Polyphen-2, PhD-SNP, I-Mutant2.0 and Mupro to generate a consensus table as output. PredictSNP on the other hand is a consensus classifier that integrates tools such as MAPP [94], nsSNPAnalyzer [95], PANTHER [88], PhD-SNP [79], PolyPhen-1 [96], PolyPhen-2 [75], SIFT [80] and SNAP [87]. The six best-performing of these tools are used to give a consensus prediction stating whether a variant is deleterious or neutral from a score calculated as follows [92]:

$$PredictSNP\ score = \frac{\sum_{i=1}^{N} (\delta_i \cdot S_i)}{\sum_{i=1}^{N} S_i},$$

Equation 2.1: Calculation of PredictSNP score.

N is the number of integrated tools, δ_i represents the overall prediction (+1 for the deleterious prediction, -1 for the neutral prediction) and S_i expresses the transformed confidence scores [92].

The output value of the PredictSNP score belongs to the continuous interval <-1, +1> [92]. The mutations are considered to be neutral for the values in the interval <-1, 0> and deleterious for the values in the interval 0, +1> [92].

2.1.1 Protein Variation Effect (PROVEAN)

The principle of PROVEAN was described by Choi [76]. The tool collects a set of homologous and distantly related sequences from the NCBI database using BLASTP and clusters them based on 80% sequence identity. For each sequence in the supporting sequence set, a delta alignment score is computed using the BLOSUM62 substitution matrix and gap penalties of 10 for opening and 1 for extension. Within each cluster, an average delta score is computed. The averaged delta scores are then again averaged among all clusters to remove bias. This unbiased averaged delta score is the final PROVEAN score calculated using the equation below:

$$PROVEAN\ score = \frac{1}{N}\sum_{c=1}^{N}\left(\frac{1}{N_c}\sum_{i=1}^{N_c}\Delta_{c,i}\right),$$

Equation 2.2: Determination of the PROVEAN score.

where N is the number of clusters in the supporting set, N_c is the number of supporting sequences in the c-th cluster, and $\Delta_{c,i}$ is the delta score with respect to the i-th supporting sequence in the c-th cluster. A low delta score means that the variant has deleterious effects on the protein function while a high score means that the variant has a neutral effect.

2.1.2 Polymorphism Phenotyping V2 (PolyPhen-2)

PolyPhen-2 also known as PPH2 uses eight sequence-based and three structure-based predictive features, which are selected automatically by an iterative greedy algorithm [75]. These features involve a comparison of a property of the wild-type allele and the corresponding property of the mutant allele [75]. The most informative predictive features characterize how likely the two human alleles are to occupy the site given the pattern of amino-acid replacements in the multiple-sequence alignment; how distant the protein harbouring the first deviation from the human wild-type allele is from the human protein; and whether the mutant allele originated at a hypermutable site [97]. The functional importance of an allele replacement is predicted from its features by a naive Bayes classifier [75,77].

2.1.3 I-Mutant

I-Mutant uses support vector machines (SVMs) to predict the direction (ΔΔG sign) of the protein stability changes and the associated ΔΔG values [76]. A low or negative ΔΔG indicates that the SNP is destabilizing while a high or positive ΔΔG indicates that the SNP is stabilizing. It utilizes information from both the sequence and tertiary structure.

2.1.4 Predictions of Protein Stability Changes Upon Mutations (MuPro)

MuPro like I-Mutant uses SVM-based methods to predict the stability changes for single-site mutations in two contexts taking into account structure-dependent and sequence-dependent information respectively [83]. The first framework focuses on predicting the ΔΔG sign as described by Capriotti et al [76]. The second regression context directly predicts the ΔΔG resulting from single-site mutations [83]. If the energy change ΔΔG is positive, the mutation increases stability; if negative, the mutation is destabilizing [83].

2.1.5 PhD-SNP

PhD-SNP uses SVMs to predict whether a nsSNP is neutral or is involved in the insurgence of human genetic disease [79]. It utilizes different algorithms to do this which include, a baseline probability-based predictor as a benchmark to overperform, a single sequence SVM method (SVM-Sequence) that discriminates disease-related mutations based on the local sequence environment of the mutation at hand and a sequence-profile-based SVM (SVM-Profile) [79]. SVM-Sequence and SVM-Profile are cast in a unique workflow with a decision tree method (HybridMeth) that allows adopting either SVM-Sequence or SVM-Profile depending on the presence or absence of a sequence profile of the sequence at hand [79].

2.1.6 Sorting Intolerant From Tolerant (SIFT)

SIFT is an algorithm that utilizes sequence homology to predict the potential impact of amino acid substitutions on protein function [80]. In addition to conservation, it also considers the physicochemical properties of the amino acids involved to make a probability calculation [82]. Based on the amino acids appearing at each position in an alignment, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated [82]. If the normalized value (SIFT score) is less than the cutoff, the substitution is predicted to be deleterious [80,82]. The score ranges from 0 to 1. A value of between 0 and 0.05 is predicted to affect protein function [98,99].

2.1.7 Screening for Non-acceptable Polymorphisms (SNAP)

SNAP is a neural network-derived tool that predicts the functional effects of nsSNPs using sequence-based computationally acquired information alone [87]. It incorporates factors such as evolutionary information and predicted aspects of protein structure (secondary structure, solvent accessibility) among others [87].

2.1.8 Multivariate Analysis of Protein Polymorphism (MAPP)

MAPP considers the physicochemical variation present in a column of a protein sequence alignment and, based on this variation, predicts the impact of all possible amino acid substitutions on the function of the protein [94].

2.2 Methodology

2.2.1 Data retrieval

Genetic variation data was obtained from the Ensembl genome browser [100,101] which contains a collection of over 6000 mutations validated from dbSNP [102]. The human gene TPMT (ENSG00000137364) was selected and a list of variants retrieved. The data was filtered to contain only the variants with missense mutations and on the basis of clinical relevance (drug response and pathogenicity). This resulted in a list of 170 missense variants which were downloaded as a comma-separated value (CSV) file. The SNPs were then validated ensuring the existence of evidence information and that the residues undergoing mutation as per the data were the correct ones on the various positions in the WT protein. A python script was written to do this by reading the CSV file then checking for evidence information at the "Evidence" column of the file. The script also compared the mutated residue with that of the WT at the respective position confirming the correctness of the residue. The FASTA sequence of the WT TPMT was retrieved from Universal Protein Resource (UniProt) (UniProt ID: P51580) and was used for this validation.

2.2.2 Variant Analysis and selection

The FASTA sequence of the WT TPMT protein and the list of 170 validated SNPs were submitted as input to VAPOR [91] to predict the effects on stability and functional impact on protein structure. To perform these predictions, the portal uses I-Mutant 2.0 [76] and MuPro [83] for effect on stability and PROVEAN [77], PolyPhen-2 and PhD-SNP [79] for functional impact, then generate a

consensus table from all the tools. The obtained results were cleaned and parsed out using a python script then saved as a CSV file (Table S1). Mutations from 5 residue positions were selected based on these predictions, their effect on catalytic activity, location on the protein structure and conservation. Information from 46 variants reported to the TPMT nomenclature committee [57,103], clinical data from Ensembl and Pharmacogenomics Knowledgebase (PharmGKB) [104,105] and in vivo, in vitro and in silico studies [40,52,55,60,106,107] were considered when making the selection. One position had three occurrences of residue substitutions where one was a novel variant unreported on Ensembl. This resulted in a total of 7 SNPs. These SNPs were reanalyzed on PredictSNP [92] to get a wider range of predictions. For this analysis, PredictSNP drew a consensus using MAPP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP.

2.2.3 Mapping 6MP onto the active site of the human TPMT

To investigate the pharmacogenetics of TPMT in 6MP metabolism, the X-ray co-crystal structure of TPMT complexed with S-adenosylhomocysteine-6-mercaptopurine (SAH(CH_2)6MP) [54] was needed. However, no such crystal structures had been reported for human TPMT owing to the poor solubility of 6MP in water [52,54]. The X-ray crystal structure of murine TPMT (PDB ID: 3BGD, resolution 2.00 Å) in complex with SAH and 6MP [53] shares a 77% sequence identity with the human sequence [54]. The drug coordinates from this structure were copied (mapped) onto the human protein by adaptation of the potential binding modes previously suggested by crystallography [52] as described by Mokmak [54] and as detailed below.

The 3D crystal structure of human TPMT bound to the ligands SAH and B3P (PDB ID: 2BZG, Chain A; Resolution: 1.58 Å) [52] and the murine TMPT crystal structure aforementioned were downloaded from the RCSB Protein Data Bank [108]. The structures were visualized on Maestro-v13.2 software [109] where the mapping would be done. Prior to mapping, all the crystallographic water molecules in both structures, the B3P ligand in the human protein and extra chain B in the murine structure were deleted. Three selenomethionine residues at positions 76, 148 and 170 in the human protein were converted to methionine. The two structures were then superimposed aligning them together along with the SAH molecules bound to them. The coordinates of 6MP from the murine structure were then mapped onto the human structure. This was done by addition of a methyl group to the sulfur atom of SAH bound to the human protein as seen in Figure 2.1b); the methyl carbon atom was then used as a reference to form a bond with the sulfur atom of 6MP (Figure 2.1c), thus forming SAH (CH_2)6MP complex (Figure 2.1d). The coordinates for this

complex were saved and renamed SMP then renumbered 300 with the assignment of an atomic charge of +1 as reported by Mokmak [54]. This structure represented the human TPMT protein complexed with the drug 6MP and was used as the initial WT TPMT-SMP working structure.

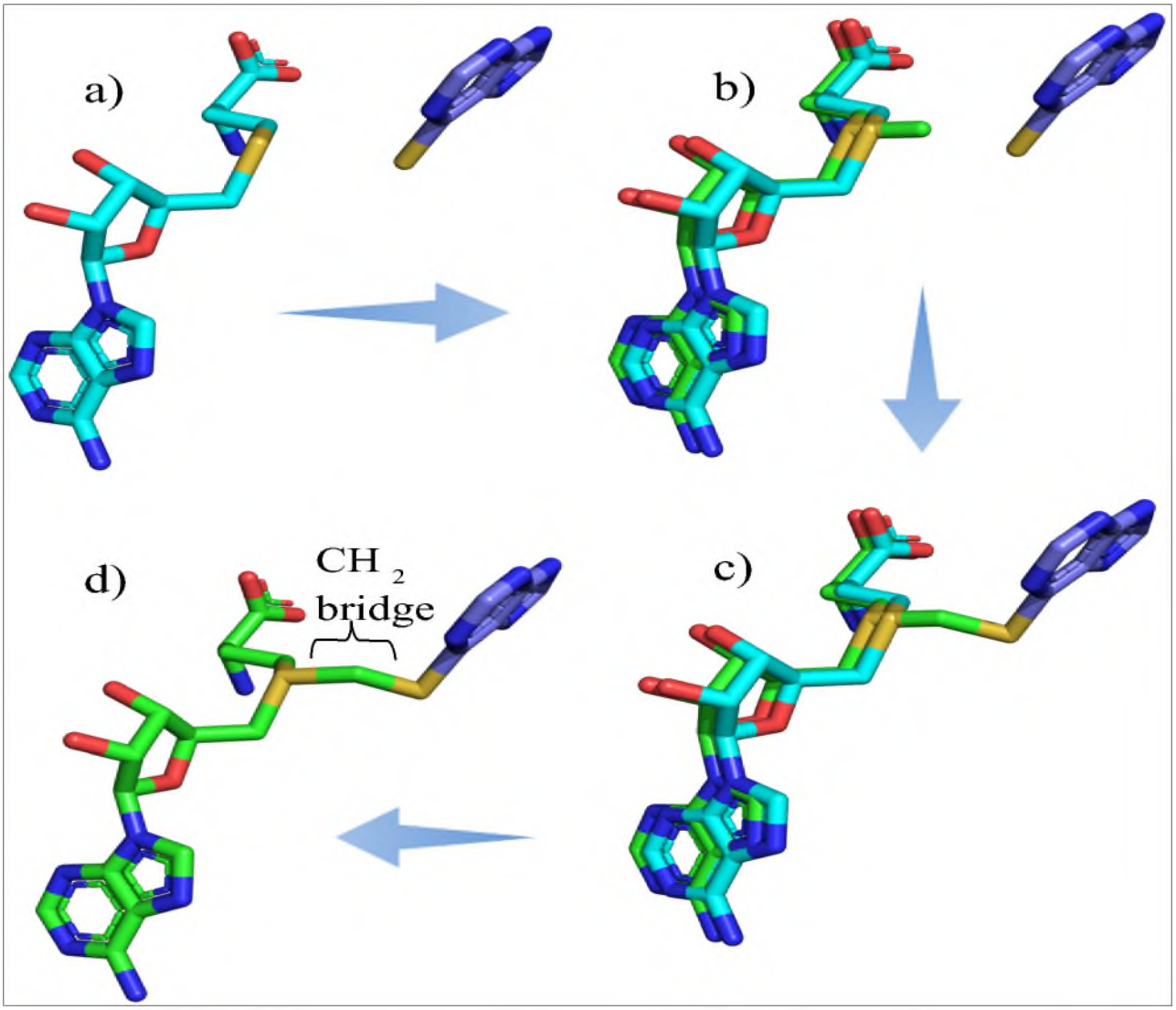


Figure 2.1: The diagrammatic representation of the steps involved in the mapping of the SAH-6MP complex into the active site of the WT protein. **a)** shows the orientation of SAH and 6MP while in the active site of the murine protein represented as cyan-coloured and blue-coloured sticks respectively. **b)** shows how SAH from the human protein (green-coloured sticks) and that of the murine aligned after superimposing the two proteins. A methyl group was added to the sulphur atom of SAH in the human protein. The sulphur atoms in all the structures were coloured orange. In **c)**, the carbon atom of the added methyl group was used as a reference to form a bond with the sulphur atom of 6MP. **d)** is the final SMP complex.

2.2.3.3 Generation of mutant complexes

For further variant analysis in MD simulations, mutant protein structures of TPMT in complex with SAH and SMP were generated. This was done by mutating the WT residues to their respective variant residues at specific amino acid positions using Maestro-v13.2. The mutation numbering was based on the initial PDB file of 2BZG (17-245). Listing S1 represents the FASTA sequences for the WT human, murine and variant enzymes used in the study.

2.3 Results and discussions

2.3.1 Variant Analysis and selection

To predict the impact of residue substitutions on the structure and function of the enzyme, variant analysis was done on VAPOR for 170 variants (Table S1). The predictions aided in selecting SNPs for the study. According to Table 2.1, MuPro and I-Mutant analysis from VAPOR predicted the selected SNPs as having a destabilizing effect. Functional analysis from PROVEAN predicted the SNPs as being deleterious, PolyPhen-2 analysis predicated all the SNPs with exception of H227Q, which was predicted benign, as probably damaging; and PhD-SNP analysis predicted L49S, L69V and H227 as neutral while A80P, R163H and R163C were predicted to be associated with disease.

Table 2.1: Variant analysis results from VAPOR analysis

SNP	PROVEAN	PPH2	I-Mutant2.0 DDG	MuPro Sign	PhD-SNP
L49S	-5.098	probably damaging	-1.50	Decrease	Neutral
L69V	-2.747	probably damaging	-1.44	Decrease	Neutral
A80P	-4.420	probably damaging	-0.37	Decrease	Disease
R163H	-3.150	probably damaging	-1.35	Decrease	Disease
R163C	-5.355	probably damaging	-0.66	Decrease	Disease
H227Q	-3.950	benign	-0.19	Decrease	Neutral

Taking consideration of VAPOR predictions, clinical data from PharmGKB and information from literature, a total of 7 SNPs were selected for the study. These SNPs can be classified in groups of those that severely reduce enzymatic activity, those that just cause a reduced activity and those that cause an intermediate-minimally reduced effect on activity. No mutations have been reported to cause an increase in enzymatic activity. Table 2.2 gives a summary of the effects of these mutations according to classification drawn in this study. In vivo information on the effect of enzymatic activity was as reported from patients being tested in health facilities.

Table 2.2: A list of selected mutations with information from clinical data that were used for further analysis.

CLASSIFICATION	SNP	EFFECT ON ENZYME ACTIVITY/ PATHOGENIC FEATURE	TYPE OF EXPERIMENTS /DATA	TPMT NOMENCLATURE	REFERENCE
Severely reduced activity	L49S	Nearly complete loss of function due to local changes in the active site rather than reduced structural stability.	In vivo, in vitro, in silico	TPMT*5	[40,54,59,105]
	A80P	100-fold decrease in TPMT activity and very low levels of protein in vitro. TPMT deficiency.	In vivo, in vitro.	TPMT*2	[40,55,60,110]
Reduced activity	L69V	Reduced enzymatic activity in vivo and no detectable activity in vitro.	In vitro, in vivo	TPMT*21	[19,60]
	R163P	Reduced enzymatic activity in vivo with decreased protein levels relative to WT in vitro.	In vitro, in vivo, in silico	TPMT*22	[19,54,106]
	R163C	Reduced enzymatic activity both in vivo and in vitro.	In vitro, in vivo, in silico	TPMT*33	[54,111]
Intermediate-minimally reduced activity	R163H	Intermediate activity in vivo with reduced clearance in vitro.	In vitro, in vivo, in silico	TPMT*16	[54,64,106]
	H227Q	Minimally reduced activity, 98% of the WT	In vitro, in vivo	TPMT*7	[40,52,112]

The selected SNPs were reanalyzed on PredictSNP (Table 2.3) to compare results from the two web servers and get a wider range of predictions.

Table 2.3: Variant analysis results from PredictSNP

SNP	PredictSNP prediction	MAPP prediction	PhD-SNP prediction	PolyPhen-1 prediction	PolyPhen-2 prediction	SIFT prediction	SNAP prediction
L49S	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious
L69V	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious
A80P	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious
R163H	deleterious	deleterious	deleterious	deleterious	deleterious	neutral	deleterious
R163C	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious
R163P	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious	deleterious
H227Q	neutral	neutral	neutral	neutral	neutral	neutral	neutral

The substitutions, L49S, L69V, A80P, R163H, R163C and R163P were predicted to be deleterious from PredictSNP consensus (Table 2.3). All the tools had an output of deleterious for these 6 SNPs except R163H which was neutral from SIFT analysis. H227Q was predicted to be neutral from the analysis by all tools with the consensus prediction from PredictSNP also being neutral. PhD-SNP prediction for L49S and L69V from VAPOR (neutral effect) differed with that of PredictSNP (deleterious effect). This could result from running the prediction tool from different web servers. The R163P variant was not retrieved on Ensembl during data retrieval from the database as it had not been reported on dbSNP, and therefore VAPOR analysis was not performed on it. Reports of this variant on PharmGKB [104,111] indicate that it is associated with decreased expression of the resultant enzyme when assayed within COS-7 cells compared to the wildtype [19]. In general, all the selected mutations except H227Q were predicted to be deleterious.

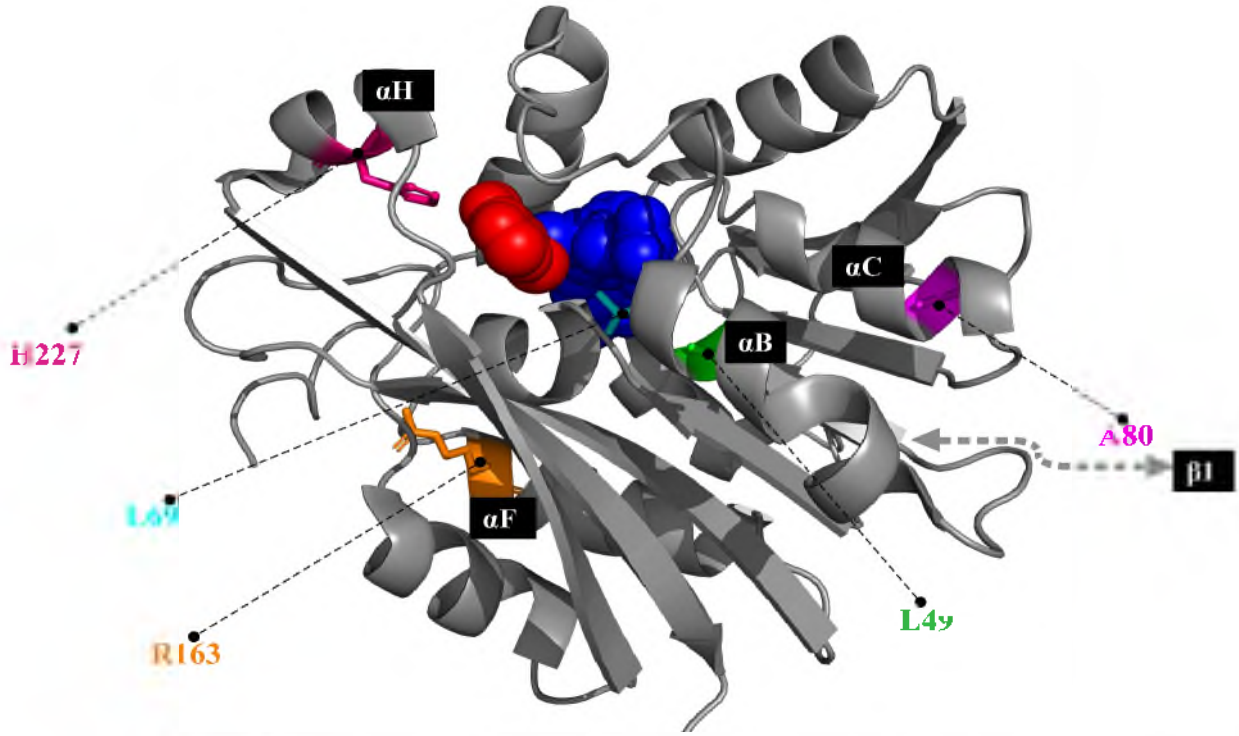


Figure 2.1: A diagrammatic illustration of the residue substitution locations. Each location is shown as sticks the same colour as the corresponding name. The secondary structures where these locations are found are labelled in white letters against a black background. SAH and 6MP are represented as spheres coloured blue and red respectively.

The H227Q mutation is located on αH, which is part of the methyltransferase fold (Figure 2.1) [52] and at a solvent-exposed position away from the binding site. A loss in the histidine imidazole ring (Figure 2.2a) could introduce inter-residue clashes which in turn could destabilize the protein. Additionally, histidine is a positively charged amino acid while glutamine lacks an ionizable side chain. The introduction of glutamine at this position would therefore interfere with local interactions and cause a disturbance in the structural dynamics of the protein. According to studies, the location of this mutation is not conserved meaning the residue could have minimal importance to protein function [40,52].

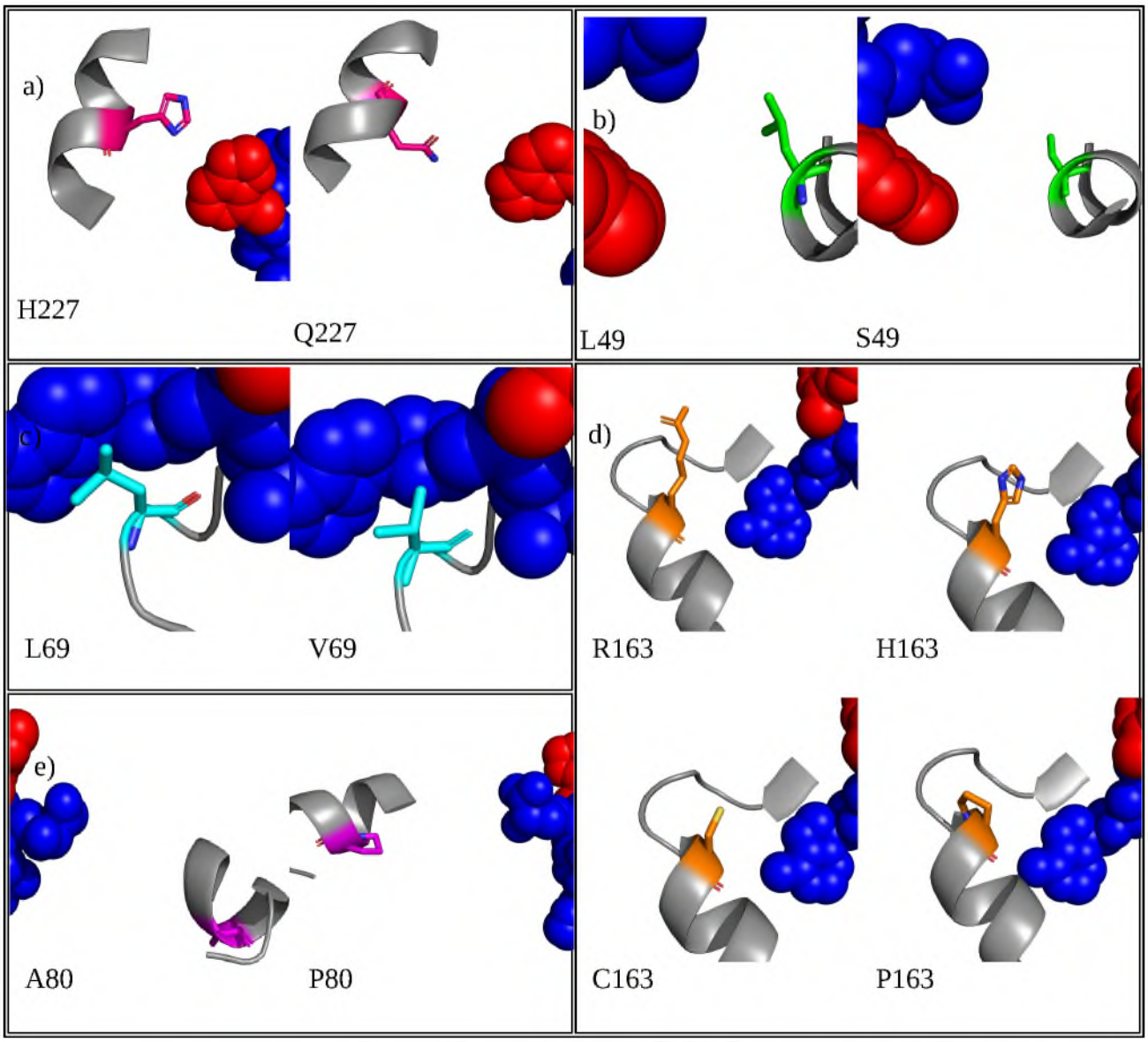


Figure 2.2: A representation of each of the substitutions. Residues are coloured differently for each substitution while the secondary structures they are found on are represented as grey-coloured cartoon. The substrate and co-substrate 6MP and SAH are presented as red and blue spheres respectively. **a)** is the change from His to Gln at position 227. The residues are presented as pink sticks. **b)** is the change from Leu to Ser at position 49 with the residues presented as green sticks. **c)** is the change from Leu to Val at position 69. Residues are coloured cyan and represented as sticks. **d)** is the representation of mutations at position 163. At the top left is the wildtype residue Arg, at the top right is the His substitution, at the bottom left is the Cys substitution and at the bottom right is the Pro substitution. The residues are presented as orange sticks. Finally, **e)** is the Ala80Pro substitution with the residues represented as sticks and coloured magenta.

Studies have shown that L49, L69, A80 and R163 residues are located at highly conserved locations [55,65,106].

Leucine is substituted at two positions surrounding the catalytic site. The first substitution, L49S, is located on the αB helix which is adjacent to the catalytic cavity while the second substitution, L69V, is located at the loop between αC and β1 in the catalytic channel where it interacts directly with the ligand (Figure 2.1). Both of these substitutions introduce small amino acids which might cause rearrangements around the catalytic site and affect catalytic activity (Figure 2.2 b-c). The L49S change introduces a polar residue in a less packed cavity which might introduce new polar interactions with neighbouring residues next to the catalytic site and as a result, affect protein flexibility and cause changes to the active site resulting in altered ligand binding.

The A80P substitution introduces a deformation into the last turn of helix αC due to the inflexible nature of proline (Figure 2.1 & 2.2e). The helix is next to secondary structures with residues involved in ligand binding and surrounding the catalytic site. Additionally, the distal end of the helix from the location of A80P has residues that interact directly with the ligand. The reduced catalytic activity from this substitution would likely be as a result of interfered amino acid interactions, around and on the secondary structure and protein misfolding. The position is on the surface far away from the binding site (Figure 2.1).

R163 is located on the αF at a position proximal to residues forming the ligand binding cleft (Figure 2.1). Three substitutions which include His, Cys and Pro have been reported at this site [17,19,55]. R163H (Figure 2.2d) involves a change from one hydrophilic amino acid to another. Both of these amino acids have ionizable side chains. The substitution could still bring about clashes with other close by residues owing to the different isoelectric points (pIs) and side chain pK_as. Arg has a pI of 10.76 with its side chain having a pK_a of 12.48 while His has a pI of 7.59 with its side chain having a pK_a of 6.00 [112]. The pK_a of a substance is the $-\log_{10}$ of its acid dissociation constant. It is a measure of the tendency of a group to give up a proton with that tendency decreasing by 10-fold as the pK_a increases by one unit [112]. The pI on the other hand is the pH at which a substance has a net charge of zero [112]. This means that at the protein's physiologic pH of 7.5 [52], Arg would likely be protonated and thus be positively charged while His would likely have a net charge of zero resulting to loss of interactions. Replacement with cysteine at this position introduces a smaller hydrophilic residue (Figure 2.2d) which could cause rearrangements. Cysteine also has an ionizable side chain (pKa of 8.18) and a pI of 5.07. At the the protein's physiologic pH, cysteine would be deprotonated as opposed to the protonated Arg in the WT protein which could cause a shift in interactions at this location and around the catalytic site. Substitution with proline at this position

could bring about changes in the secondary structure of the protein considering the structure of proline (Figure 2.2d) and its hydrophobic nature.

2.3.2 Mapping of 6MP onto the catalytic site of human WT TPMT

Maestro-v13.2 [73,109] was used for mapping 6MP onto the human TPMT. The final inter-sulfur distance between SAH and 6MP in the newly generated complex was 3.2 Å (Figure 2.3) which is a close approximate of the inter-sulfur distance between SAH and 6MP molecules in the murine TPMT structure (3.4 Å) (Figure 2.3a). This validates the correctness of the complex. The distances are approximately double the 1.82 Å length of a sulfur-carbon single bond confirming that the methyl group is directly transferred from SAM to 6MP in a nucleophilic reaction where the sulfur from 6MP is the attacking atom [52,54]. Figure S1 shows the ligand interaction diagram for the SMP complex formed from this mapping.

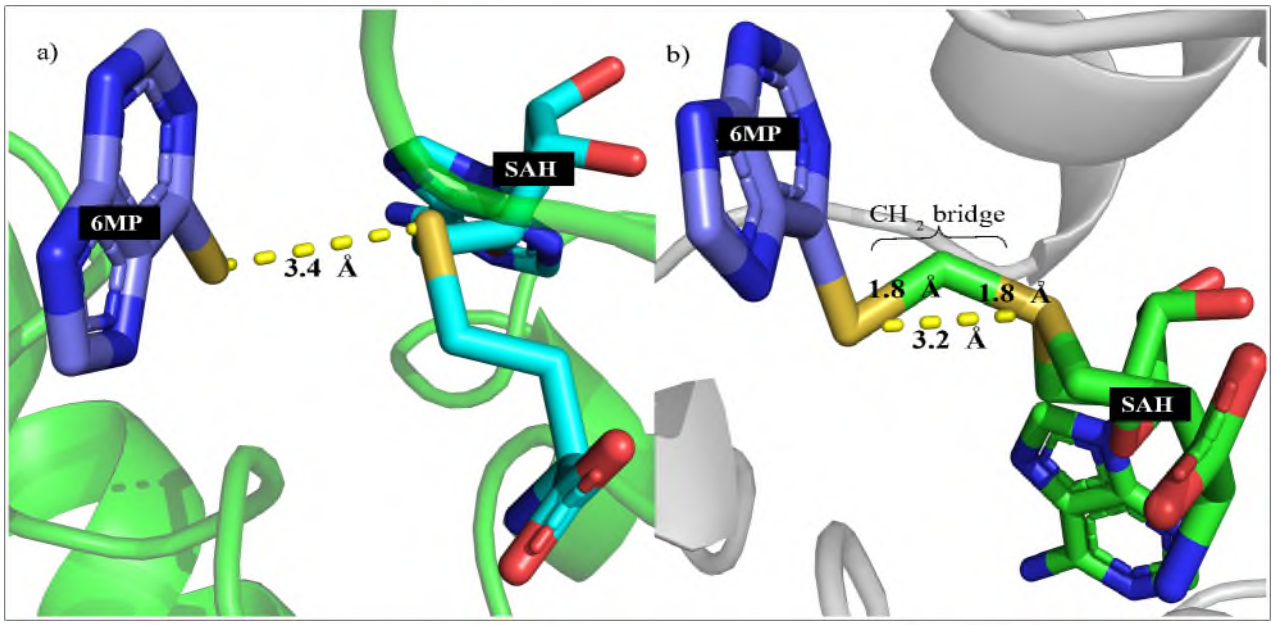


Figure 2.3: Generation of SAH(CH_2)6MP complex. **a)** is the orientation of 6MP and SAH as seen in the murine TPMT crystal structure. **b)** is a depiction of the complex formed between SAH and 6MP during methylation with the formation of a CH_2 bridge.

2.4 Conclusions

Most non-synonymous single nucleotide polymorphisms on the human TPMT have deleterious effects on catalytic activity. Several factors could play a role in how or why an amino acid change could affect enzymatic activity. Such factors include the location of the mutation in relation to the binding site which could interfere with ligand interactions directly if the residue is involved in ligand binding or indirectly if it causes local changes to the binding site by interfering with inter-

residue interactions. Another factor could be the introduction of an amino acid with different physicochemical properties such as hydrophobicity, presence of ionizable side chains, the polarity of the amino acids involved and the presence of charged side chains. These properties play a role in protein folding, function and protein interactions depending on the cellular location of the protein and each amino acid is strategically placed to ensure the proper functioning and final conformation of the protein. Another factor could be the size of the amino acids involved. Replacement of a small amino acid with a bigger amino acid and vice versa could cause shifts and rearrangements thus introducing inter-residue clashes within the protein. This could in turn have an effect on the structural dynamics of the protein. How conserved a residue is could also be used to explain the effect of its substitution. Highly conserved residues are believed to play a functional role in the protein and their substitution could have deleterious effects on the protein. SNPs that interfere with protein folding could mark it for rapid chaperon-mediated clearance from cells. Understanding the functional and structural impact of nsSNPs on TPMT is a step forward in gaining insight into how enzymatic activity is affected in their presence.

CHAPTER THREE

MOLECULAR DYNAMICS SIMULATIONS

3.1 Introduction

Molecular Dynamics (MD) simulations is a computational method that employs Newton's laws to evaluate the motions of water, ions, small molecules, and macromolecules or more complex systems, for example, whole viruses, to reproduce the behaviour of the biological environment, including water molecules and lipid membranes [54]. When using an empirical force field [113], such as GROMOS, OPLS, AMBER, CHARMM or MARTINI, the equations of motion are essentially Newton's second law [114]. Equation 3.1 is Newton's second law of motion which equates the force *(F)* of an object to its mass *(m)* and acceleration *(a)*.

$$ma = F$$

Equation 3.1: Newton's second law of motion.

Conservative forces are computed as the gradient of the energy of the system [115]. Molecular dynamics solve these equations by time integration [115]. These calculations predict how every atom in a protein or other molecular system will move over time, based on a general model of physics governing inter-atomic interactions [115]. These interactions could be bonded or non-bonded [116]. Bonded atomic interactions are a result of bond extension, angle bending and dihedral torsions of the bonded atoms and improper dihedrals [117]. Non-bonded interactions result from the interactions between Van der Waals and electrostatic forces [117]. After calculating the forces, the configuration is updated by applying those forces over a discrete time step [117]. To ensure numerical stability, the time steps in an MD simulation must be short, typically only a few femtoseconds (10^{-15}s) each [118]. Most of the events of biochemical interest, for example, functionally important structural changes in proteins, take place on time scales of nanoseconds, microseconds, or longer [113]. A typical simulation thus involves millions or billions of time steps. This fact, combined with the millions of interatomic interactions typically evaluated during a single time step, causes simulations to be very computationally demanding [119].

Simulations can predict how biomolecules will respond at an atomic level to perturbations such as mutation, phosphorylation, protonation, or the addition or removal of a ligand [119].

MD simulations have been widely applied to study structure-function relationships, disease pathways, and drug design [119]. Various open-source packages have been developed over time to perform MD simulations. Such packages include AMBER [118], CHARMM [120], NAMD [121], LAMMPS [122], Desmond [123] and GROMACS [124]. This study utilized GROMACS version 2018.2 and the AMBER99SB-ILDN force fields to study the effects of mutations on TPMT and how they affect 6MP metabolism.

3.1.1 GROMACS

GROningen MAchine for Chemical Simulations (GROMACS) is a free open-source software package commonly used to perform MD simulations for biomolecules [125]. It is made up of a set of programs for preparing, running and analyzing MD simulations. Each program is called from within a single binary, *gmx,* with a similar syntax among all the programs. Options are preceded by an hyphen (-) and the subsequent keyword or filename is the next argument on the command line [118].

3.1.2 AMBER force fields

A force field is built up from two distinct components; the set of equations used to generate the potential energies and their derivatives (the forces) and the parameters used in this set of equations [126]. AMBER force fields (AMBERff) introduce a set of parameters for all-atom simulations. Some characteristic features of this force fields include fixed partial charges on atom centres, explicit use of all hydrogen atoms, no specific functional form for hydrogen bonding, and dihedral parameters fit to relative quantum mechanical (QM) energies of alternate rotamers of small molecules. Some common AMBERff include ff94, ff96 and ff99 which are upgrades of each preceding ff in terms of improvement in conformation sampling with consideration of dihedral angles [127]. The AMBER99SB-ILDN applied in this study is a refined AMBER ff99SB with optimized x_1 torsion potentials for amino acid side chains [128]. The torsion potentials result from rotations around the dihedral angles.

3.2 Methodology

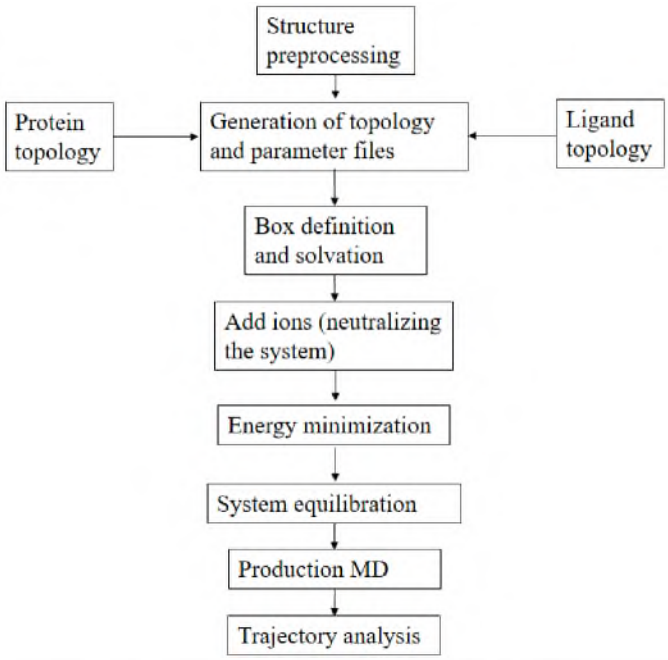


Figure 3.1: The general work-flow for protein simulations in GROMACS.

The general work flow for protein simulations in GROMACS involves:

i. Getting a starting structure e.g. from a database such as PDB or a homology modelled structure.

ii. Structure preprocessing - protonating to the desired pH.

iii. Preparing GROMACS parameters and topology files.

iv. Choosing an appropriate force field.

v. Solvating and adding ions to neutralize the system.

vi. Energy-minimising the system.

vii. Equilibrating temperature and pressure.

viii. Running the production MD.

ix. Analysing the trajectories.

3.2.1 Structure preprocessing

Before setting up parameters for MD runs, the protein systems needed to be preprocessed. This is because the accuracy of atomistic biomolecular modelling and simulation studies depends on the accuracy of the input structures [118]. Preprocessing involved removing hydrogen atoms, and

ligands if the analysis was being carried out on an apoprotein. The structures were then protonated at the protein's working pH of 7.5 on the H++ web server [129,130]. To protonate the ligands on the server, the HETATM records were modified to LIGAND. The apo protein was generated from the holo protein by removing the cofactor. After protonation, the topology ".top" and coordinate ".crd" files were downloaded from H++ and converted to a PDB file by merging them using the *ambpdb* command.

3.2.2 Generating topology and parameters for MD simulation

The AMBER99SB-ILDN force field and tip3p water model were used to create topology and introduce position restraints into the system. Topologies for SAH and SMP were generated separately using AnteChamber Python Parser intercafacE (ACPYPE) tool with a net charge of 0 for SAH and +1 for SMP. An octahedron simulation box was chosen to create periodic boundary conditions (PBC) and solvated explicitly using water molecules. The periodic box is an artificial construct that provides infinite space, at a low computational cost using an effective algorithmic trick. Molecules exiting the box on one side reappear on the opposite side such that if a molecule goes to the left, it is mirrored at the right of the box, and vice versa; if it goes to the bottom, it is mirrored at the top, and vice versa. This allows continuous simulation. The systems were then neutralized by replacing some of the solvent molecules with sodium cations (Na^+) or chloride anions (Cl^-) depending on the system's net charge at 0.15M salt concentration.

Energy minimization was performed to relax the structures and resolve unrealistic forces in the system that would result from repulsiveness of atoms that are too close and new interactions that form as a result of replacing solvent molecules with ions. This was done in 50,000 steps using the steepest descent algorithm until the systems converged to a maximum force of 1000/kJ/mol/nm. A two-phase equilibration in 100 ps was then performed on the systems to adjust to the required temperature and pressure in order to avoid system collapse. The first phase was conducted under *NVT* ensemble with a constant number of particles, volume and temperature at 300K temperature using the modified Berendsen thermostat algorithm at 0.1 ps time constant. The second phase was conducted under *NPT* ensemble, wherein the number of particles, pressure, and temperature was constant at 1 bar using the Parrinello-Rahman barostat at 2.0 ps time constant. In both ensembles, all bonds were constrained using the LINCS solver. Non-bonded interactions were allowed a maximum distance of 1.4 nm. Long-range electrostatic interactions were evaluated using the

Particle Mesh Ewald (PME) method and the Van der Waals interactions were handled using the Lennard Jones potential.

3.2.4 Production run

MD simulations were run at the Centre for High Performance Computing (CHPC) cluster using 10 CPU cores and one Nvidia Tesla v100 GPU for 100 ns with a 2 fs timestep. Trajectories were written out every 2 ps. Double runs for the WT systems in apo, holo and drug-bound conformations were performed as controls to ensure that all the parameters were correctly set and confirm the reproducibility of the results (Figure S2).

3.2.5 Trajectory analysis

Periodic boundaries were corrected for the systems to rewrap and recentre the protein systems within the simulation box avoiding PBC jumps using *gmx trajconv*. The corrected trajectories were then stripped off waters and ions using *cpptraj* [131] before analysis was performed. The trajectories were additionally visualized using VMD [132] to confirm the PBC corrections and ensure that the ligands remained in the active site. Calculations for the root mean square deviation (RMSD), root mean square fluctuation (RMSF), and radius of gyration (Rg) were then performed for all the systems using *gmx tools*.

RMSD is the average distance between atoms of superimposed proteins. The RMSD of certain atoms in a molecule with respect to a reference structure can be calculated by least-square fitting that structure to the reference structure $(t_2=0)$ [133] and subsequently calculating the RMSD according to Equation 3.1.

$$RMSD(t_1,t_2)=[\frac{1}{M}\sum_{i=1}^{N} m_i\|r_i(t_1)-r_i(t_2)\|^2]^{\frac{1}{2}} \quad ,$$

Equation 3.1: Calculation of RMSD.

Where $M=\sum_{i=1}^{N} m_i$ and $r_i(t)$ is the position of atom *i* at time *t*.

A constant continuous increase shows sustained conformational changes in the structure of the protein and can be used to analyse the stability of the protein. In this study, the RMSD for all the protein systems in apo, holo and drug-bound states were calculated after least-square fitting to the C_α.

The Rg of a protein is a measure of its compactness [134]. It is calculated as the mass-weighted geometric mean of each atom's distance from the centre of mass of the protein. Equation 3.2 describes this calculation.

$$R_g = \left(\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i}\right)^{\frac{1}{2}} ,$$

Equation 3.2 : Calculation of Rg.

Where m_i is the mass of atom i and r_i the position of the atom i with respect to the centre of mass of the molecule.

The stability and folding of a protein are known to be affected by the packing of amino acid residues, hence, the more compact the protein the lower its folding rate. A stably folded protein will likely maintain a relatively steady value of Rg. The Rg values from all the runs were calculated for the protein systems.

RMSF is a measure of local chain flexibility. It analyses the portions of the structure that are fluctuating from their mean structure the most and is measured as the distance between the portion and its average position [133] according to Equation 3.3.

$$RMSF = \sqrt{\langle (X_i - \langle X_i \rangle)^2 \rangle} ,$$

Equation 3.3: Calculation of RMSF.

Where X_i is the coordinates of particle i, and $\langle X_i \rangle$ is the ensemble average position of i. The RMSF for residues in the apo, holo and drug-bound systems were calculated in this study.

Further analysis was done to understand how the atoms moved together using dynamic cross-correlation (DCC) which is calculated as in Equation 3.4 [135].

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle} \cdot \sqrt{\langle \Delta r_j^2 \rangle}} ,$$

Equation 3.4: Calculation of dynamic cross-correlation.

Where Δr_i is the displacement from the average position of atom i, and $\langle\rangle$ is the time average over the whole trajectory. Trajectories were reduced using an in-house python script to keep only the C_α and C_β atoms. DCC was then calculated for each of the proteins with a time step size of 50 using

calc_correlation.py from MD-TASK suite [135]. Heatmaps that showed the correlation of the motions of Cα atoms within the protein over the trajectory frames were generated from the same script. In DCC analysis, atoms that moved together are regarded as correlated, when they move in opposite directions they are regarded as anti-correlated and if no relationship exists between their movement, they are then regarded as having no correlation [136].

3.3 Results and discussions

3.3.1 Trajectory analysis of the WT protein

To follow the global conformational changes of the WT enzyme in all states over the course of the 100ns MD simulations, the RMSD and Rg of the structures were analysed. The enzyme had a fairly steady RMSD and thus stable in the drug-bound and holo states compared to the apo state as seen from violin and line plots in Figure 3.2a.

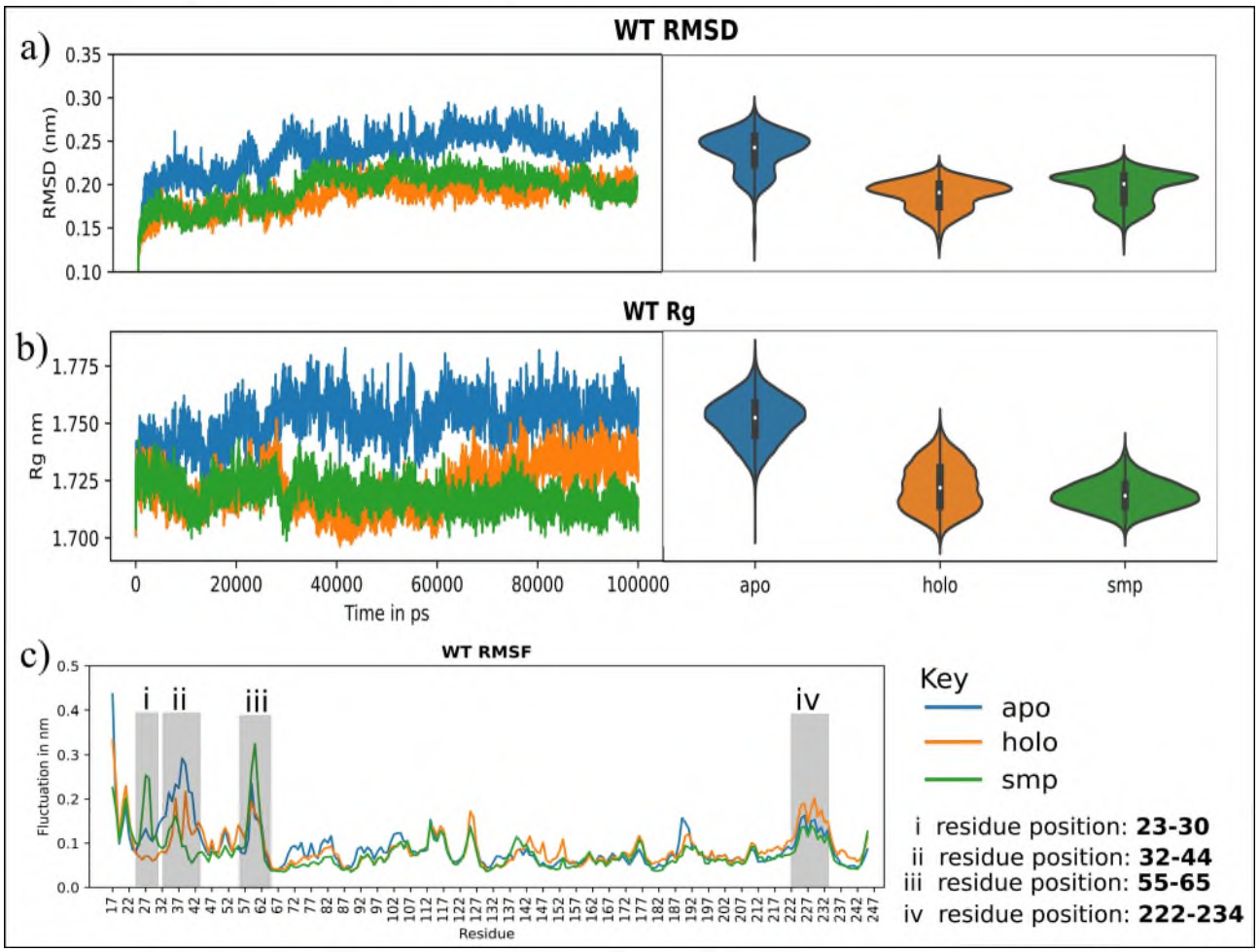


Figure 3.2: Plots for the WT enzyme trajectories and residue fluctuations in apo (blue), holo (orange) and drug (green) bound states. **a)** shows the RMSD line and violin plots for all the states. **b)** shows the Rg line

and violin plots. and c) is the RMSF line plots for all the systems. Areas with high RMSF are shaded grey and numbered i-iv.

The observed RMSDs in the various states could be used to explain conformational changes before and after ligand binding. Once it is bound, stability is paramount to maintaining the ligand in the binding site until catalysis is finished. The binding of the co-substrate therefore had a stabilizing effect on the protein. The stability was sustained even after the drug was bound into the active site. RMSD calculations were used to confirm consistency of the simulations in the WT duplicate runs. The results were represented as line graphs in Figure S2.

Rg measures the compactness of a structure along the period of an MD simulation. Compactness is related to the flexibility and stability of the structure. According to Figure 3.2a all the states maintained a relatively steady Rg meaning that the enzyme was stably folded. The apo state enzyme had a higher record of Rg which could explain the global flexibility changes within the structure as seen in RMSD analysis.

To identify local flexible areas in the different enzyme states over the course of the simulation. RMSF analysis was done. In Figure 3.2c, highly fluctuating regions have been shaded and numbered. The regions were mapped onto the structure as seen in Figure 3.3.

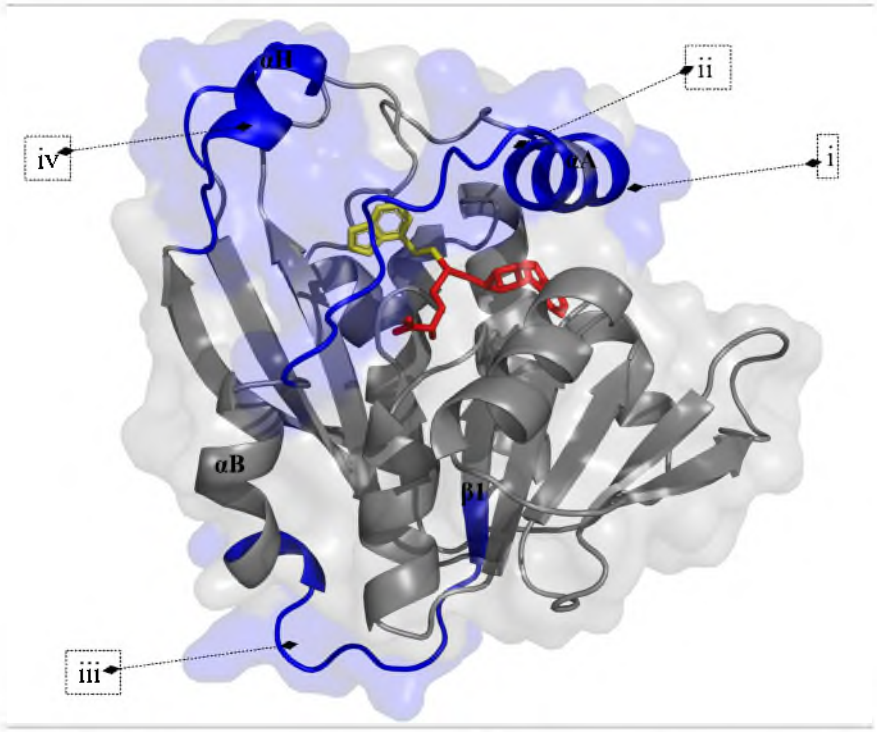


Figure 3.3: Mapping of the flexible areas (blue) in the WT protein. SMP is represented as sticks with 6MP coloured yellow and SAH red.

Region *i* was unique to the drug-bound enzyme. The region falls in the αA helix which caps the active site (Figure 3.3). The flexibility of this helix is important in allowing the substrate to get into the catalytic site and thus plays an important role in drug binding. Region *ii* was mostly in a loop region between αA and αB and (Figure 3.3) was more prominent in the apoenzyme than the holoenzyme, with minimal fluctuations in the drug-bound enzyme. The region contains important residues that form the catalytic channel including F40 (aids in drug binding). The transition in fluctuations in this region could be a demonstration of structural changes around the catalytic site as catalysis is going on. The area *iii* spans from αB into a loop then to the beginning of β1 sheet (Figure 3.3). These two secondary structures have residues interacting directly with both the drug and co-factor. The flexibility of the loop might therefore be important in ensuring functionality of the enzyme. The final region, *iv*, covers the αH helix from both ends (Figure 3.3). The helix is part of the methyltransferase fold and could play an important role in allowing substrate binding. As was observed, flexible regions are known to be loop areas, residues involved in ligand binding or those involved in conformational changes to allow ligand binding.

3.3.2 Trajectory analysis of the Mutant systems in comparison to the WT

3.3.2.1 RMSD analysis

To see global conformational changes of the mutant systems in comparison with the WT along the trajectories, RMSD for these systems were analysed in their apo, holo and drug-bound states.

Varying changes in RMSD were observed in all the systems according to Figure 3.4. The most stable systems in the apo systems compared to the WT included L49S, A80P, R163C and H227Q with the most instability being observed in the L69V, R163H and R163P. In the holo systems L49S, A80P and R163P remained stable while L69V, R163H, R163C and H227Q were unstable. As with the WT enzyme, the substitutions that sampled stable in this state therefore experienced the a stabilizing effect from co-factor binding. In the drug-bound systems, all the systems were stable except L49S therefore drug binding seemed to stabilize the systems. The L49S substitution became unstable after drug binding which could have been a result of altered structural integrity in areas close to or at the binding site.

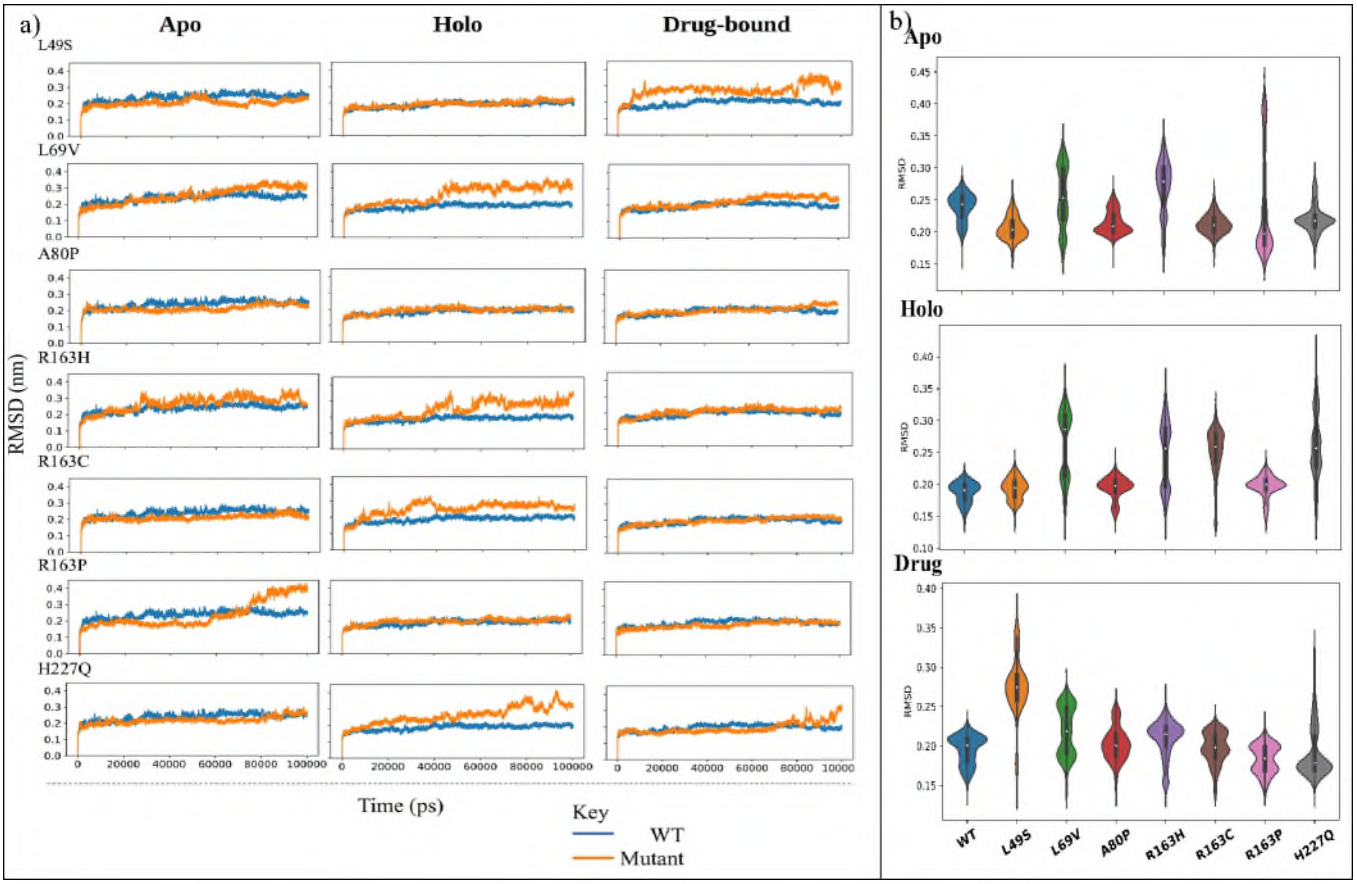


Figure 3.4: a) RMSD line plots for the WT (blue) and mutant (orange) apo systems. The x-axis is the time in ps and the y-axis is the RMSD in nm. **b)** RMSD violin plots for all the systems in apo, holo and drug-bound states.

Collective RMSD violin plots comparing the different states in each protein systems were plotted according to Figure 3.5 to get a better view the conformational changes that occurred before and after ligand binding as discussed above.

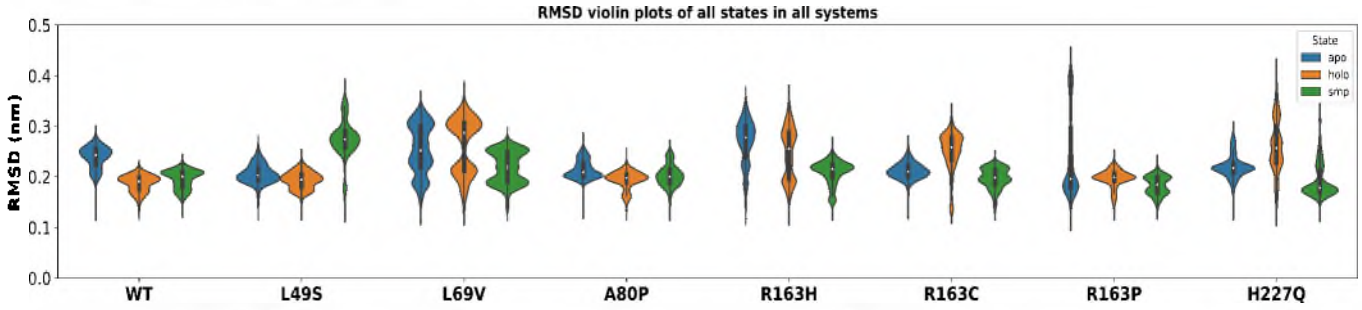


Figure 3.5: RMSD Violin plots of all the systems in all the analysed states giving a general overview of the conformational changes that occurred over the course of the MD simulation.

Using the WT as reference, RMSD sampling above or below what was recorded in the WT could be an indication of altered global conformation. Excessive stability associated with high rigidity could

affect ligand binding while instabilities associated with excessive flexibility could cause ineffective catalysis. Occurrence of mutations therefore had an effect on global enzyme motions which could be associated with reduced catalytic activity.

3.3.2.2 Rg analysis

Changes in the compactness of the enzyme in relation to the presence of mutations were investigated using Rg calculations. Deviating Rg trends were observed in similar mutant systems that had unsteady RMSD when compared to the WT according to Figure 3.6. This is an indication that compactness is related to stability of a protein. Mutations that were more stable in the RMSD analysis had a lower and more steady sampling in Rg analysis which means that the structures were more rigid which could result in alterations in functionality.

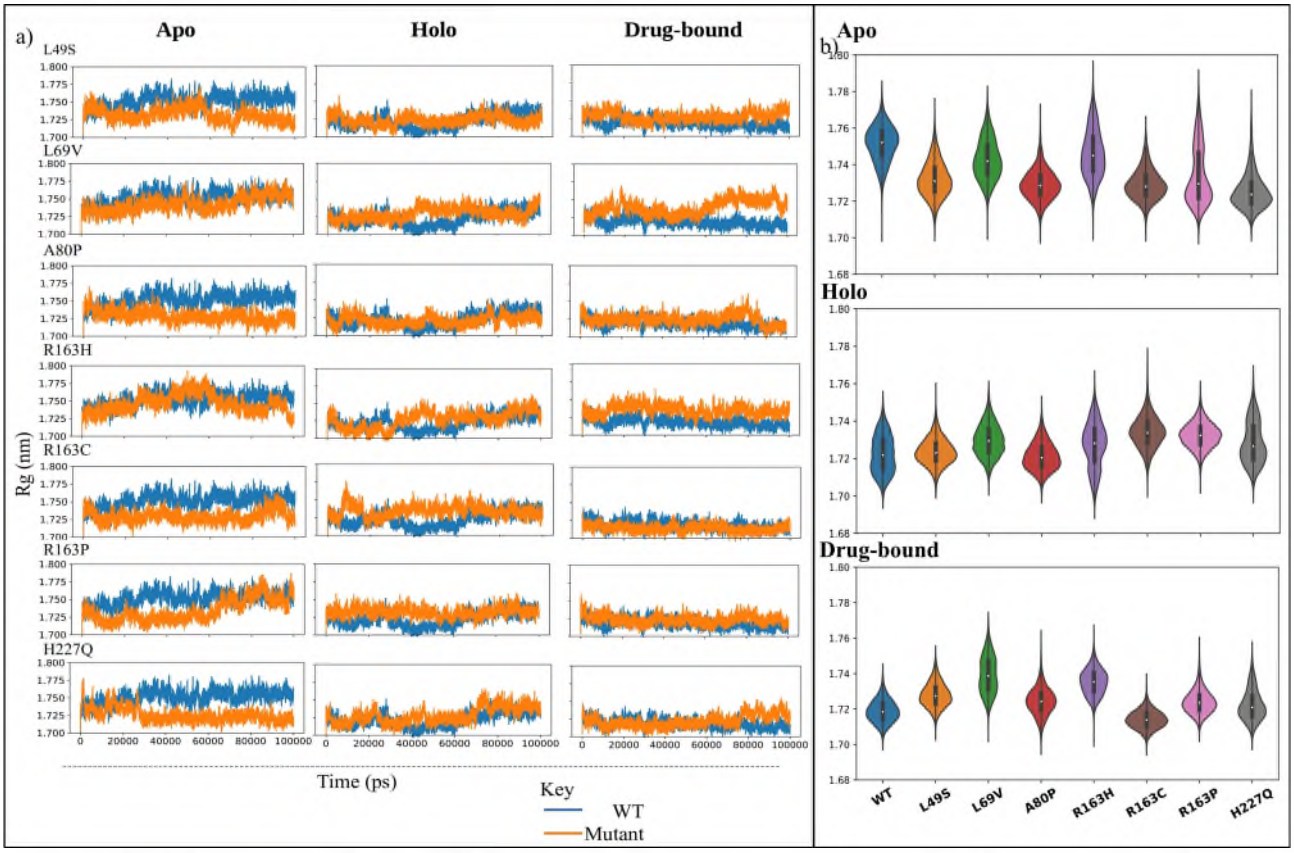


Figure 3.6: a) Rg line plots for the WT (blue) and mutant (orange) holo systems. The y-axis is the Rg in nm and the x-axis is time in ps. **b)** Rg violin plots for all the apo, holo and drug-bound systems.

3.3.2.2 RMSF analysis

To identify how flexible areas occurred and differed from the WT systems in the variant systems, RMSF analysis was done. Figure 3.7 shows how the various systems compared with the WT system.

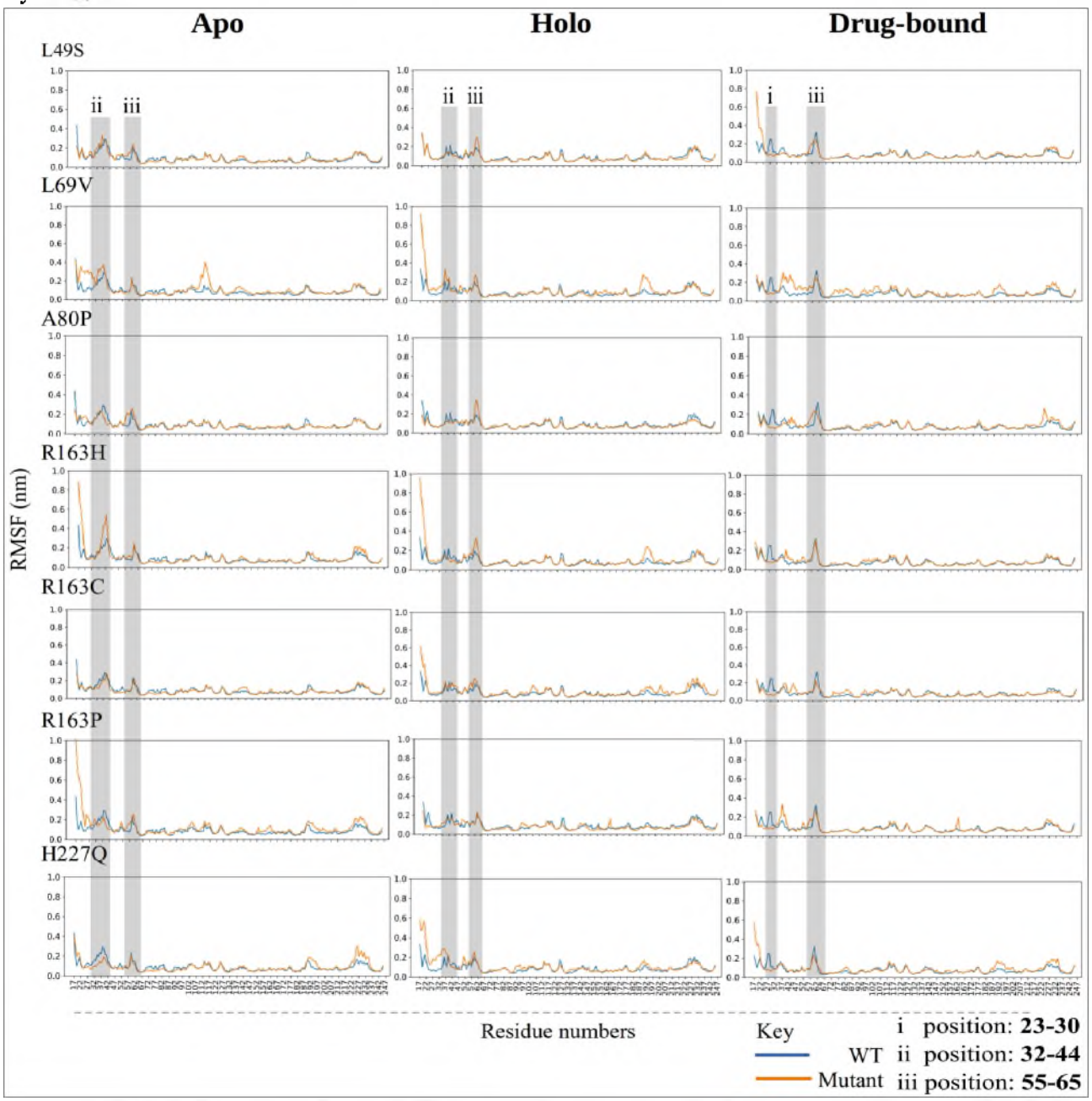


Figure 3.7: RMSF line plots for the WT (blue) and mutant (orange). The major regions as identified in section 3.3.1 have been shaded and labelled i-iii.

According to Figure 3.7, L69V and R163H recorded increased fluctuations in region *ii* compared to the WT in the apo conformation. These mutations together with R163P and H227Q had regions of increased fluctuations in other residue positions while the rest of the variants had relatively similar

areas of fluctuations when compared with the WT enzyme. This increased flexibility could be associated with instability as a result of the residue substitutions. In the L49S, A80P, R163H and R163P holo systems, a decrease in flexibility was observed in region *ii* (Figure 3.7) with H227Q and L69V showing an increase in flexibility. Decreased flexibility could affect association with the co-substrate while an increase in flexibility could confer changes in structural rigidity both of which could affect functionality. All the drug-bound variant systems recorded a decreased RMSF in region *i* which spans residue 23-30 as identified in section 3.3.1 and shaded in Figure 3.7. This region includes the αA helix which caps the active site. The loss in flexibility at this region could be detrimental to the association of the enzyme with the ligand. Flexibility in region *iii* was sustained in the apo and drug bound systems with an increase in flexibility being observed in L49S, L69V, A80P and R163H in the holo systems. As the area is a loop between two functionally important secondary structures, increased flexibility in this region compared to the WT could have functional implications on co-factor binding. It is worth noting that A80P recorded a global lower RMSF in all states compared to the WT which could explain the observed sampling in RMSD and more compact structures seen in Rg results. The substitution therefore resulted to a more rigid structure in both local and global motions.

3.3.3 Dynamic cross-correlation (DCC)

To understand how the moving atoms correlated with each other over the course of the MD simulation, DCC analysis was done using the C_β and glycine C_α atoms. From Figure 3.8, correlated, anti-correlated and areas with no correlation were present in all systems. The analysis revealed that occurrence of mutations had no significant effect in how the atoms moved together when compared to the WT. Similar patterns depicting correlation, anti-correlation and no correlation were observed in all the systems regardless of the state and presence of mutation. This could mean that the protein has a preserved structural state.

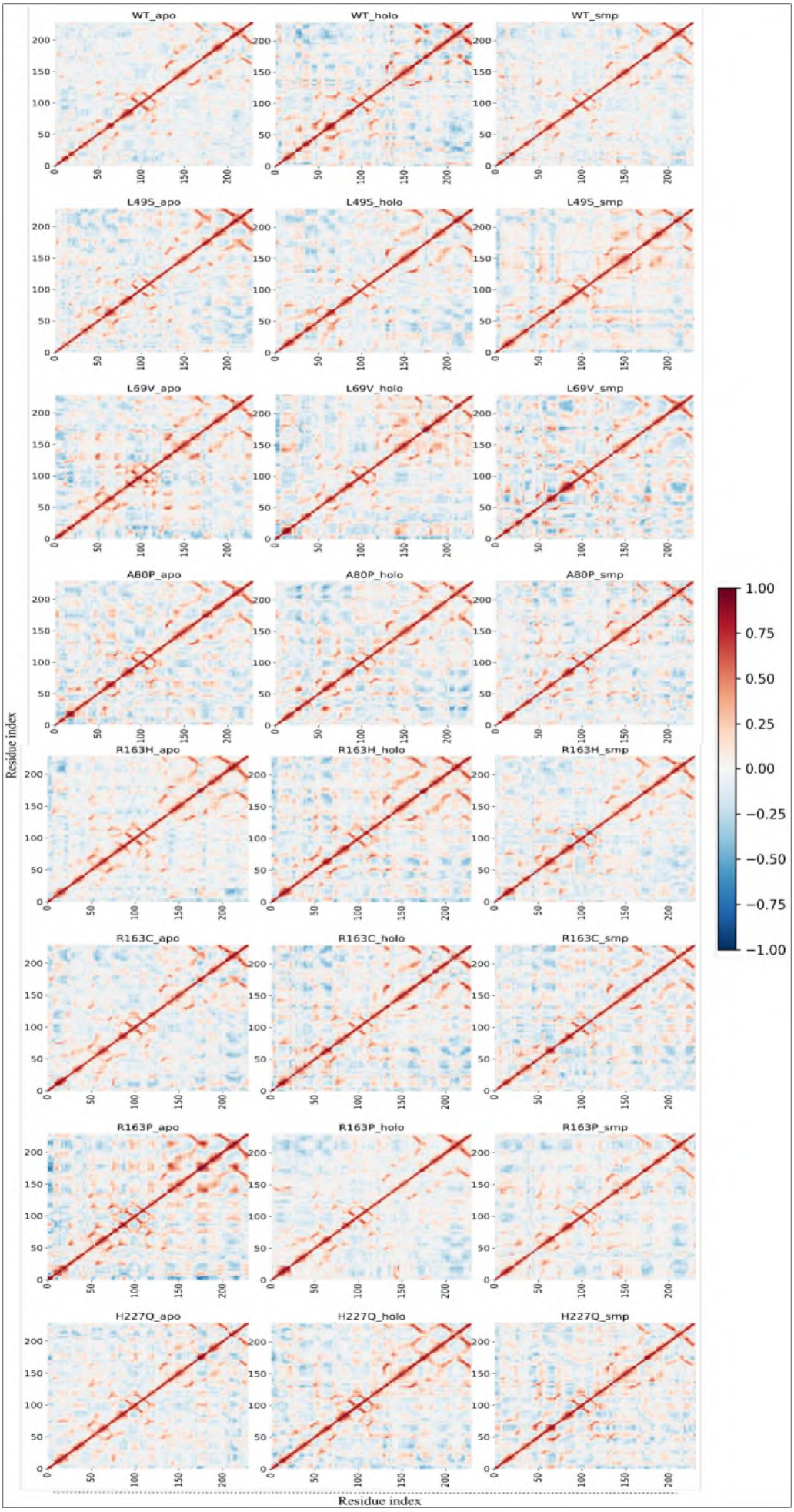

WT_apo
WT_holo
WT_smp
L49S_apo
L49S_holo
L49S_smp
L69V_apo
L69V_holo
L69V_smp
A80P_apo
A80P_holo
A80P_smp
R163H_apo
R163H_holo
R163H_smp
R163C_apo
R163C_holo
R163C_smp
R163P_apo
R163P_holo
R163P_smp
H227Q_apo
H227Q_holo
H227Q_smp
Residue index
Residue index
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 3.8: DCC analysis represented as heatmaps showing residue correlation. The *x* and *y*-axis annotate the residue index. Values moving towards 1 (red) represent areas of high correlation, zero (white) represents no correlation while the values moving towards -1 (blue) represent areas of anti-correlation.

3.5 Conclusion

In summary, the mutants with the most unsteady RMSDs in the apo systems were L69V, R163H and R163P while the rest kept a steady RMSD along the trajectories. In the holo systems, L69V, R163H, R163C and H227Q were unstable from RMSD analysis with the rest being stable while in the drug-bound systems, only L49S showed an unsteady RMSD. This same trend was observed in Rg analysis. For these variant enzymes, these could mean that their effect on catalytic activity was a result of global conformational changes in addition to residual flexibility changes observed in RMSF results. The A80P variant enzyme showed no global conformational changes compared to the WT and could have its severe effects on enzymatic activity stemming from local inter-residue changes resulting in rigid structures as seen in decreased areas of fluctuation in RMSF analysis. From DCC analysis, the occurrence of mutations had no effect on how the atoms moved together.

CHAPTER FOUR

DYNAMIC RESIDUE NETWORK ANALYSIS

4.1 Introduction

A network or a graph is a collection of objects connected by lines. The objects are called nodes or vertices while the connections between the objects are called edges or links which are drawn as lines between points [137]. Much of what is used in network science comes from the social sciences from the study of human relationships. Some relationships are reciprocated while others are not, some are stronger than others, some people influence more people and some form subgroups in diverse ways [138]. Network analysis is the evaluation of how nodes are related to one another [137]. Edges between nodes can be undirected (Figure 4.1a) or directed (Figure 4.1b). In an undirected network graph, each node is connected by an edge that does not show the origin and destination by way of an arrow. Relationships in this kind of graph are bidirectional. In a directed graph, on the other hand, each node is connected by an edge with an arrow indicative of the direction of the relationship. In directed graphs, relationships may not be reciprocated [137] Figure 4.1a & b are illustrative of these network graphs. Edges can carry additional information such as interaction strength and frequency [138].

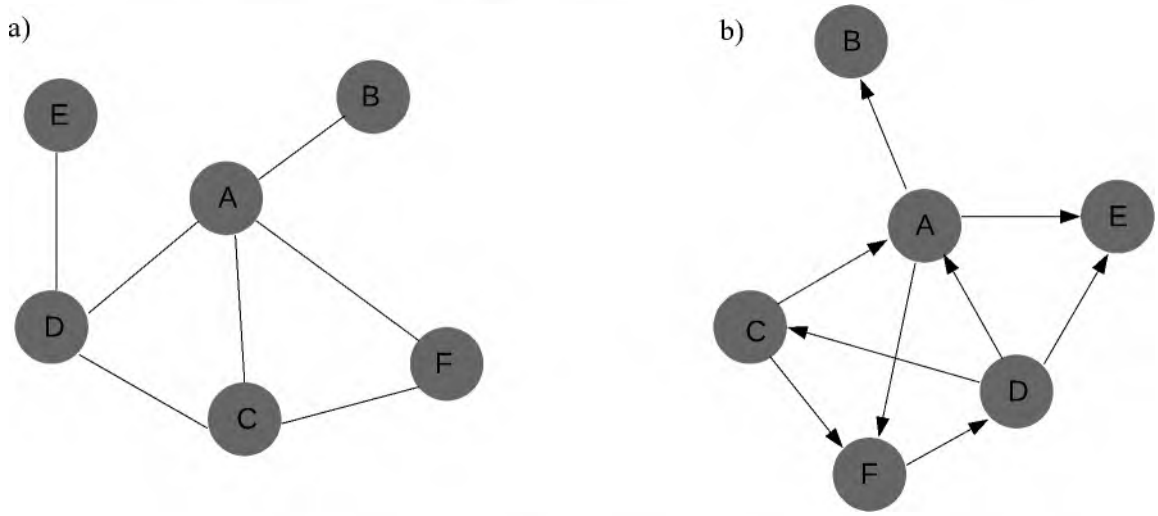


Figure 4.1: **a)** An undirected network graph. Nodes are represented as circles labelled A-F connected by lines (edges). **b)** A directed network graph. Nodes are represented as circles labelled A-F connected by directive arrows.

Interactome networks have been used in molecular biology research to provide a global picture useful in understanding how the interaction between and within molecules occurs and how that influences cellular behaviour [137,139]. Interactome networks have been applied to analyse protein interactions in cancer, to understand brain connectivity in conditions such as Alzheimers's' disease, and to study allostery in proteins among others [138]. In proteomics, several types of network analysis have been adapted. Such include protein-protein interaction (PPI) network analysis where nodes are proteins and edges can come from evidence of interaction or co-expression, intra-protein network analysis which involves residue network analysis (RIN) and dynamic residue network analysis (DRN) with nodes as residue representations while edges are mostly from residue-residue interactions then finally, statistically guided network analysis (SGNA) where groups of protein conformations are compared with edges based on statistical significance [138]. Nodes can be ranked or sorted according to their properties thus identifying central nodes that affect the topology of the network depending on the question being asked [139]. The importance of a node in a communication network is deduced by the term "centrality" [140]. Various centrality metrics can be calculated in such kind of analysis. Such metrics include:

- *degree centrality* (DC) which shows that an important node is involved in a large number of interactions,
- *closeness centrality (CC)* which indicates important nodes that can communicate quickly with other nodes,
- *betweenness centrality (BC)* which shows that nodes which are intermediate between neighbours rank higher and thus are important for communication,
- *eigenvector centrality (EC)* which ranks higher nodes that are connected to important neighbours,
- *average shortest path (L)* which is the average of the shortest path lengths from all other nodes to a given node, and many more.

Nodes with very high scores in the centrality calculations are referred to as *hubs* [139]. Additionally, contact maps can be calculated for a node of interest with the edges carrying information about the strength and frequency of interaction with other nodes.

In this study, weighted contact maps for important residues, *EC* and *BC* were applied in DRN analysis.

4.2 Methodology

4.2.1 Residue contact map analysis

Residue contact maps were calculated to illustrate the degree of interaction at mutation positions and with the ligand. This was done to identify changes in residue-residue and ligand-residue interactions as a result of mutations in the equilibrated portion of the simulation. Weighted contact maps were therefore calculated for the last 50 ns using a step size of 1 frame and a cutoff Euclidean distance of 6.7 Å between the target residue and its immediate contacts using the *contact_map.py* tool from MDM-TASK-web suite [141]. Contact heat maps were then generated using the *contact_heatmap.py* tool from the same suite.

4.2.2 Dynamic residue network analysis

Dynamic residue network analysis was used to identify the key residues in intra-protein interactions in the WT and mutant systems. Scripts from MDM-TASK-web were used to calculate average *BC* and *EC* metrics for the last 50 ns of the simulations where the C_β atoms (C_α for glycine) of each residue were treated as nodes with the existence of a cutoff distance of ≤6.7 Å being treated as an edge.

Average *BC* measures how frequently a residue is traversed along the shortest paths connecting other residues in the protein thus its importance in the intra-protein's interaction network [140,142]. *BC* is calculated as per equation 4.1 where m denotes the overall number of frames, n shows the number of residues, $\delta(s,t|v)$ symbolises the number of shortest paths bridged between a residue v and other nodes s and t. $\delta(s,t)$ denotes the averaged shortest paths existing between residues s and t where s and t are part of the set V, which comprises the set of all nodes.

$$\bar{BC}(v)=\frac{1}{m}\sum_{i=1}^{m}\sum_{u=1}^{n-1}\frac{\delta(s_i,t_i|v_i)}{\delta(s_i,t_i)}$$

Equation 4.1: Calculation of *BC*.

EC is used to measure the influence of a node in a network based on the centrality of the neighbouring nodes [140,141]. Here, an adjacency matrix A is used to assign node importance by solving for the dominant unit eigenvector *EC* of the matrix. Equations 4.2a and 4.2b are applied in these calculations. In a), *EC* is the eigenvector and λ is the eigenvalue for the eigen decomposition of the adjacency matrix. In NetworkX, *EC* is solved using the power iteration method. In b), the

averaged *EC* for a residue *i* is determined from the average time of *m* frames. The converged eigenvector is a metric that recursively assigns importance, giving high centrality to nodes that have a high degree or those connected to high importance nodes.

$$A \cdot \overrightarrow{EC} = \lambda \cdot \overrightarrow{EC}$$

Equation 4.2a): Eigenvector decomposition.

$$\overline{EC} = \frac{1}{m} \sum_{k=1}^{m} EC_{ik}$$

Equation 4.2b): Calculation of average *EC*.

4.2.3 Identification of top 5% global high network centrality hubs

In order to identify residue hubs that were persistent in the various centrality measures, all related samples were combined in order to have a common scale. The metric-specific results for the apo, holo and drug-bound systems were concatenated into a single vector before being sorted in descending order while keeping track of residue indices in order to focus on nodes of the highest overall centrality. The top 5% of these values were extracted and used to obtain a binary matrix of the same dimensions as the original dataset that contained all the samples. Any residue positions that contained a hub were chosen from other samples. The results were presented as a heatmap.

4.3 Results and discussions

4.3.1 Contact map analysis

4.3.1.1 Mutant residue contact map analysis

To identify changes in interactions after mutation at the positions of interest, weighted contact maps (Figure S3-S5) were calculated for the native and mutant residues at that position for all the states. The csv files from the calculations were used to put the data together in heatmaps (Figure 4.2) for better visualization and analysis.

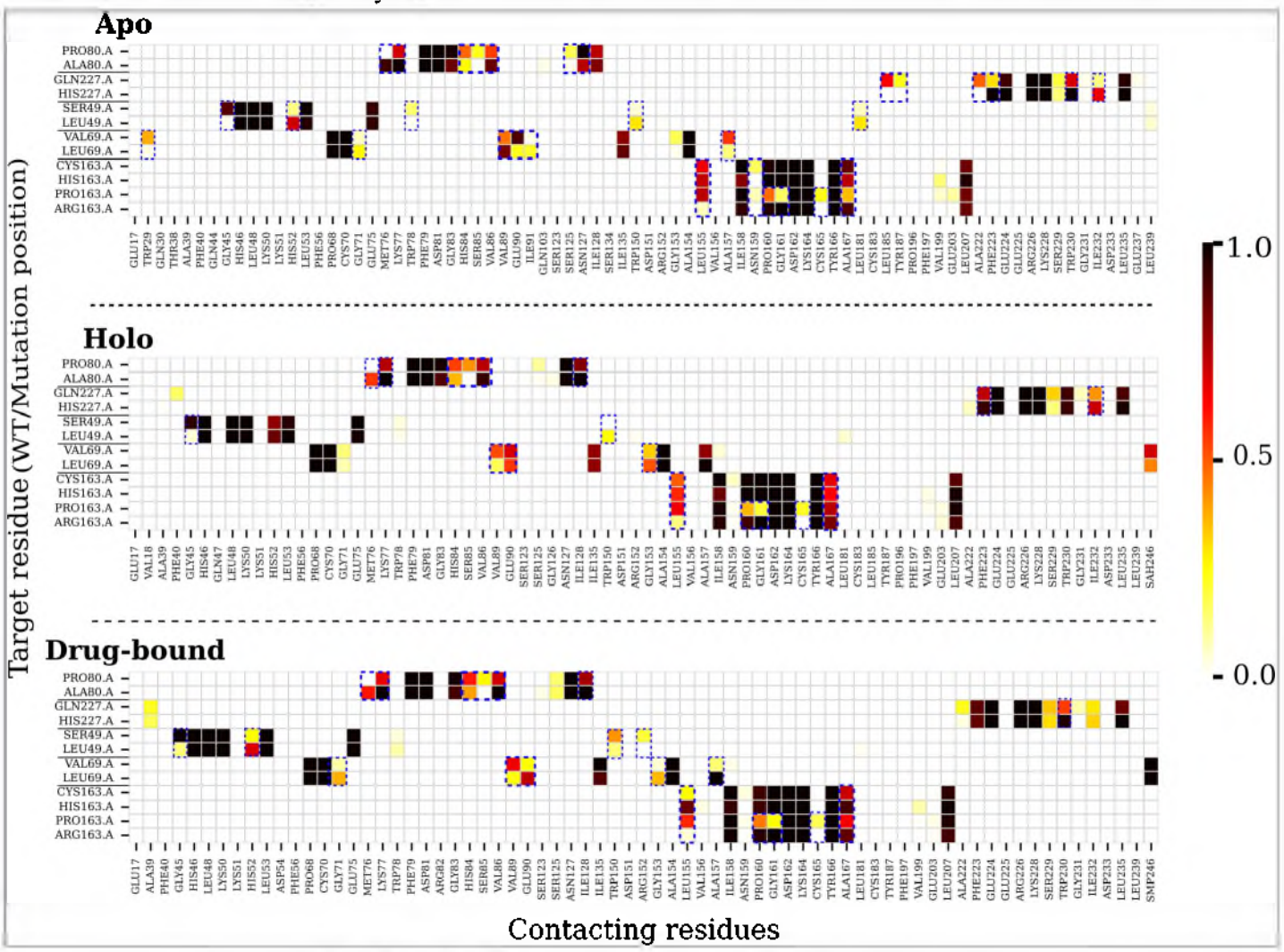


Figure 4.2: Contact heatmaps for WT and mutant enzymes at residue substitution positions in the apo, holo and drug-bound systems. Instances where changes in interactions have occurred have been surrounded with dashed blue lines. The *x*-axis represents the target residues in the various systems while the *y*-axis represents the contacting residues. A contact frequency scale that runs from 0-1 with colour intensifying from white through yellow to red then black depending on the increasing contact frequency is shown.

According to data in Figure 4.2, the substitution of Ala with Pro at position 80 leads to a loss in contacts with M76, K77 and V86 in all the states while gaining contacts with H84, S85, S125 and N127 in the apo state and H84 and S85 in the holo and drug-bound states (Figure 4.3c).

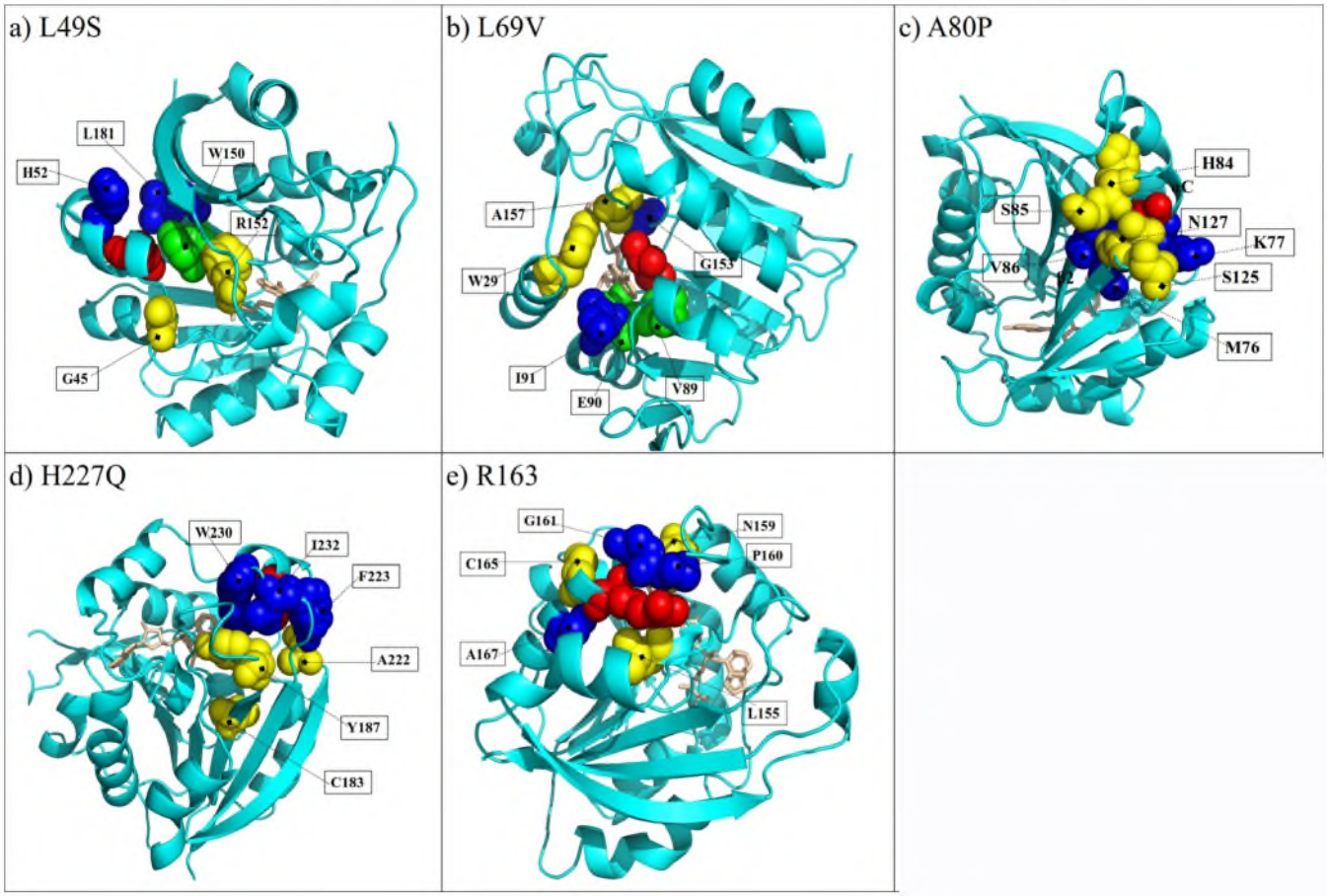


Figure 4.3: Interacting residues before and after substitution from all states in a single structure for individual mutant enzymes. Lost interacting residues are represented as blue spheres while gained interacting residues are represented as yellow spheres. Positions where interactions lost intensity in one state and gained in another state compared to the WT are represented as green-coloured spheres. Substitutions at position 163 are have their changing interactions represented together (in a single structure) because most of them happened in similar locations in all the states. The substitution position is represented as red-coloured spheres. Bound SMP is represented as wheat-coloured sticks.

As seen in Figure 4.3c, the lost contacts are located in αC and β2 whose distal ends have residues surrounding the catalytic site. This could result in downstream changes at the catalytic site also known as allosteric effects. The gained contacts are mainly in loop regions on solvent-exposed positions as is the substitution position. This could mean that the substitution resulted in increased inter-residue interactions among surface residues from reduced conformational entropy as a result of proline's pyrrolidine ring restricting N–C_α rotation and thus dihedral movements. In previous

studies, the introduction of a proline residue at a solvent-exposed position and at the end of helices as a mutation has been shown to increase the stability of proteins by reduction of conformational entropy [143–145]. This has been applied in protein bioengineering to enhance the stability of the proteins [143,145,146]. This variant appeared to have sustained stability in all the states in the RMSD analysis from chapter 3 (Figure 3.3). The observations from this analysis could be used to explain the sustained stability. Additionally, this could mean that the proline substitution alters the structural energetics of the enzyme, which could affect its normal functioning and/or mark it for chaperon-mediated clearance from the cell. The resulting reduced catalytic activity of the mutant enzyme could therefore be a result of disrupted structural dynamics, allosteric effects on the catalytic site and clearance from the cell.

The H227Q substitution resulted in a loss in interactions at F223, W230 and I232 and a gain in the C183, Y187 and A222 positions in the apo systems. Due to the similar hydrophobic natures of the gained and lost interactions these observations could be viewed as an attempt of the enzyme to accommodate the mutation. Loss of interaction was observed at F223 and I232 in the holoenzyme and W230 in the drug-bound system with no significant gain in interaction in these systems (Figure 4.2 & Figure 4.3d). This could affect the structural dynamics of the enzyme and thus effective functioning. RMSD analysis in chapter 3 revealed instabilities in the holo system while the apo and drug-bound systems remained fairly stable. With no compensatory gain in interactions for the holo system, observed instability was not unexpected. As not many changes were observed in the drug-bound system, the effect on enzymatic activity would be expected to be minimal and this agrees with available information on the minimally reduced enzymatic activity [40,52].

In L49S, loss in interactions was observed at H52, W150 and L181 in the apoenzyme, at W150 in the holoenzyme and at H52 in the drug-bound enzyme. A gain in interactions was observed at G45, W78 in the apo system, G45 in the holoenzyme and G45, W150 and R152 in the drug-bound enzyme where the W150 interaction intensified rather than lost intensity as seen in the apoenzyme (Figure 4.2 & Figure 4.3a). The gained interactions in the drug-bound enzyme could cause shifts in the catalytic site as the involved positions surround the catalytic cavity which would result in diminished catalytic activity. This observation is supported by previous studies which have shown that the mutation results in local changes in the catalytic site [40,53,55].

L69V resulted in a loss of interactions at I91, V89 in the apoenzyme, G153 in the holoenzyme and G153 and A157 in the drug-bound enzyme with gained interactions being at W29, G90, A157 in the apoenzyme and V89 in the holoenzyme and drug-bound enzyme (Figure 4.2). The higher number of

changing interactions in the apoenzyme could mean that the resulting enzyme was misfolded which could be attributed to increased clearance from the cell resulting in reduced enzymatic activity.

According to data on Figure 4.2: of the substitutions at position 163, R163P recorded a significant loss in interactions which occurred at P160 and G161 in all states and A167 in the apo and drug-bound state. Loss of interactions with A167 was additionally recorded in R163H in the apo system and R163C in the drug-bound system. R163P recorded a gain in interactions with C165 in all states while the rest did not. All the substitutions at this position recorded a gain in interactions with L155 in all states with R163H having a higher intensity of interaction followed by R163P and then R163C in the drug-bound enzyme. R163C also recorded a gain in interaction at N159 in the apo state. The occurrence of changes in inter-residue interactions in mutations occurring at this position could be attributed to reduced catalytic activity. The location has previously been reported to be highly conserved [17,55].

4.3.1.2 SMP contact map analysis

Further contact analysis for SMP was done across the mutant systems in comparison with the WT to identify any changes in interactions that might have occurred due to residue substitutions, in an attempt to further elucidate the impact of mutations on 6MP metabolism.

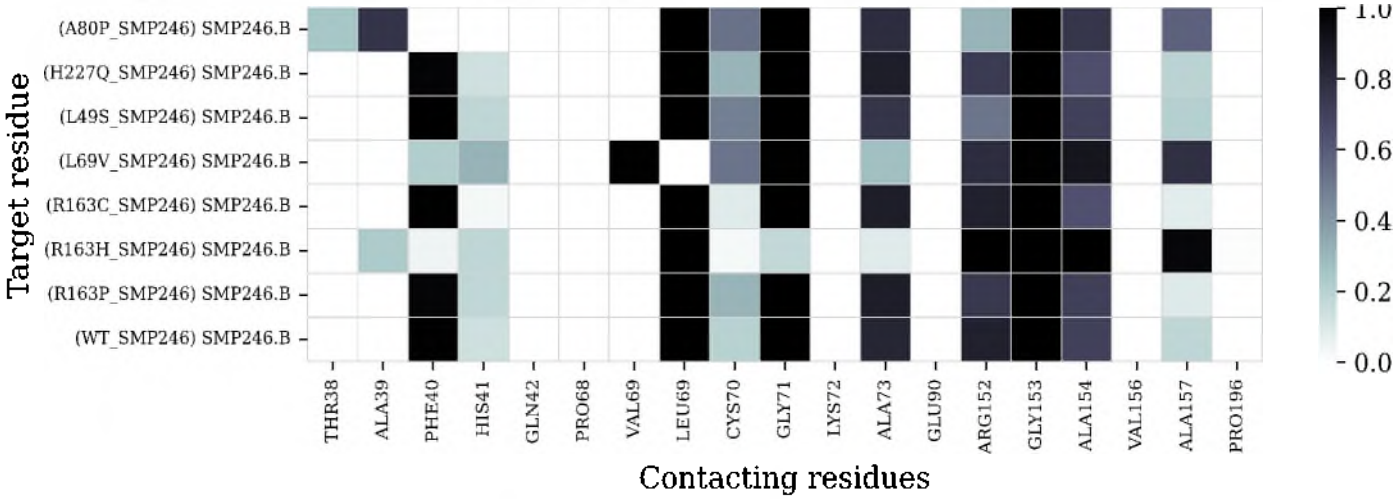


Figure 4.4: Residues interacting with SMP across the WT and mutant systems along the trajectories. The *y*-axis represents SMP in the various systems while the *x*-axis represents the interacting residues. Colour intensifies from white to black according to the frequency and strength of interactions and is represented by a scale of 0-1 where there is no interaction at 0 (white) while there is high interaction at 1 (black).

Data from Figure 4.4 suggests that L49S, R163C, R163P and H227Q did not lead to significant changes in SMP interactions. H227Q has been reported to be benign and cause a minimal effect on

catalytic activity. These observations are thus not unexpected. The observations at the L49S, R163P and R163P substitutions could mean that the reduced catalytic activity from the resulting mutant enzymes is not a result of affected SMP interactions.

According to Figure 4.4 SMP, loses interactions with F40 and H41 in the A80P substitution. F40 is directly in contact with the drug and is involved in the formation of the water channel that directs drug binding as mentioned in Chapter 1. This loss of these contacts might have dire effects on drug binding and thus its metabolism. New interactions are introduced at positions T38 and A39 as a result of this substitution. Interactions at positions C70 and A157 intensify more compared to the WT while that at R152 loses intensity. This might suggest reorientation of the drug at the binding site during methylation which might affect catalysis. The mutation is located far away from the binding site which additionally suggests an allosteric effect on drug metabolism. This could further confirm the findings discussed in the previous section (Section 4.3.1.1). The observations made for this substitution might explain the 100-fold loss in catalytic activity in the variant enzyme.

The L69V substitution leads to a loss of interactions at positions F40 and A73 and a gain in interactions at C70, A154 and A157. This might suggest affected ligand binding given the importance of F40 in drug binding. The changes might also cause changes in how SMP fits in the catalytic channel which has an affect catalytic activity. The mutation is located at the catalytic site where it interacts directly with the co-factor. The interactions at the position are sustained even with the occurrence of an SNV. The resulting loss in catalytic activity might thus be a result of altered interactions with other surrounding residues as a result of the substitution.

The R163H substitution leads to a loss of interactions at F40, C70 and A73 with intensified interaction at A157. This might lead to poor drug binding and how it orients itself at the binding site. These observations might explain the reduced catalytic activity of the variant enzyme.

4.3.2 Dynamic residue network analysis

4.3.2.1 Betweenness centrality (BC)

To investigate how often a residue is traversed by the shortest paths connecting other residues *BC* calculations were done and plotted as heatmaps (Figure 4.5).

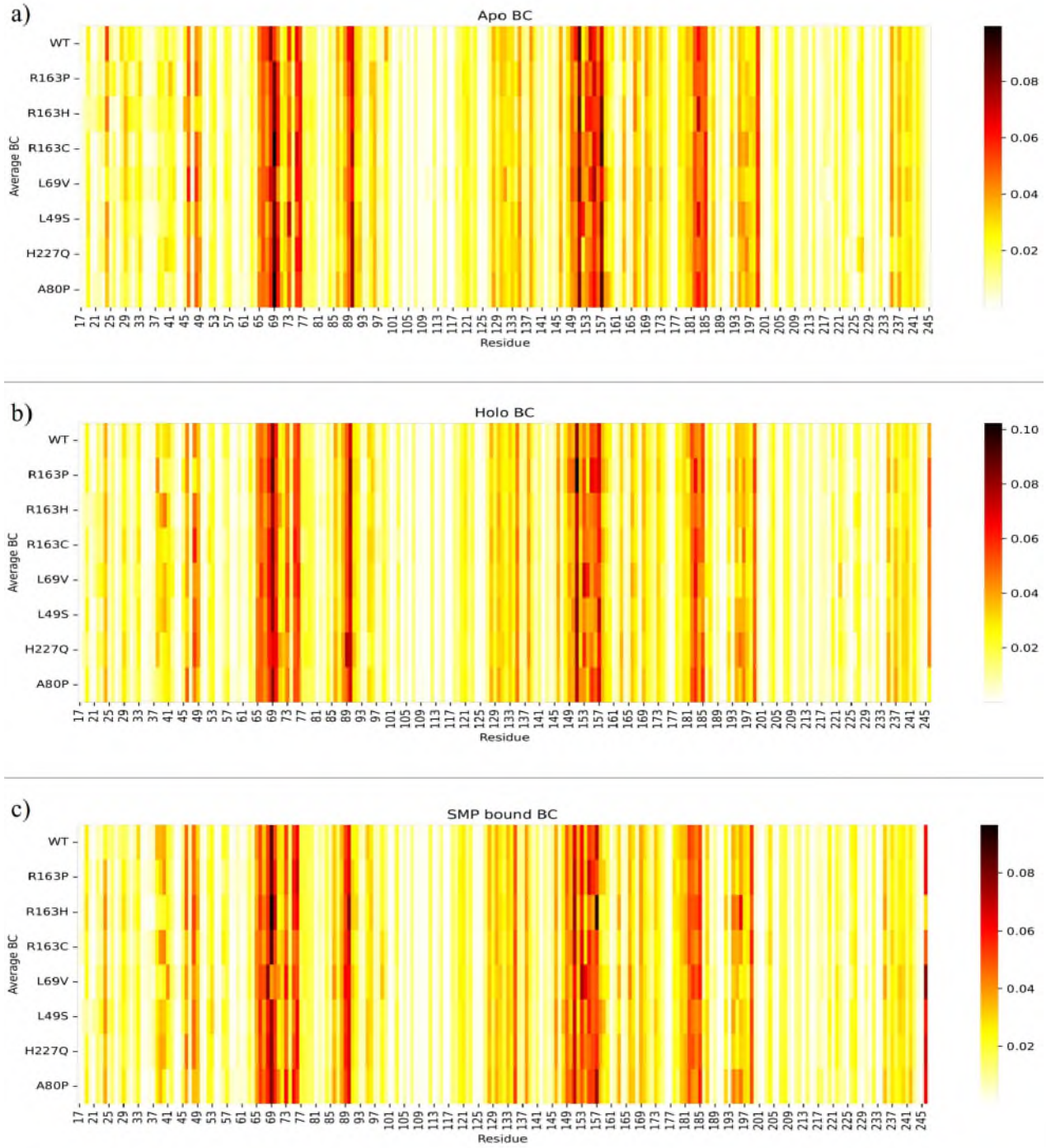


Figure 4.5: Heatmap representation of BC in all the systems in all the states. *The* y-axis represents the WT and mutant systems while the x-axis represents the residue positions. A scale that increases colour intensity

from white to black through yellow and red is used to represent an increase in *BC*. Darker colours show higher *BC* regions while lighter colours represent low *BC* regions.

According to figure 4.5, areas of high *BC* were observed from residue 65-76, 89-90, 149-157, 182-185 and 199 in all the states. All of these regions except 199 span through areas that line the catalytic channel with some of the residues being in direct contact with the ligand. These regions could thus be important in intra-protein communications that involve shaping the catalytic channel and conferring stability of the folded protein.

Persistent global top 5% *BC* hubs in all the studied systems were identified on residues 90 and 151 (Figure 4.6).

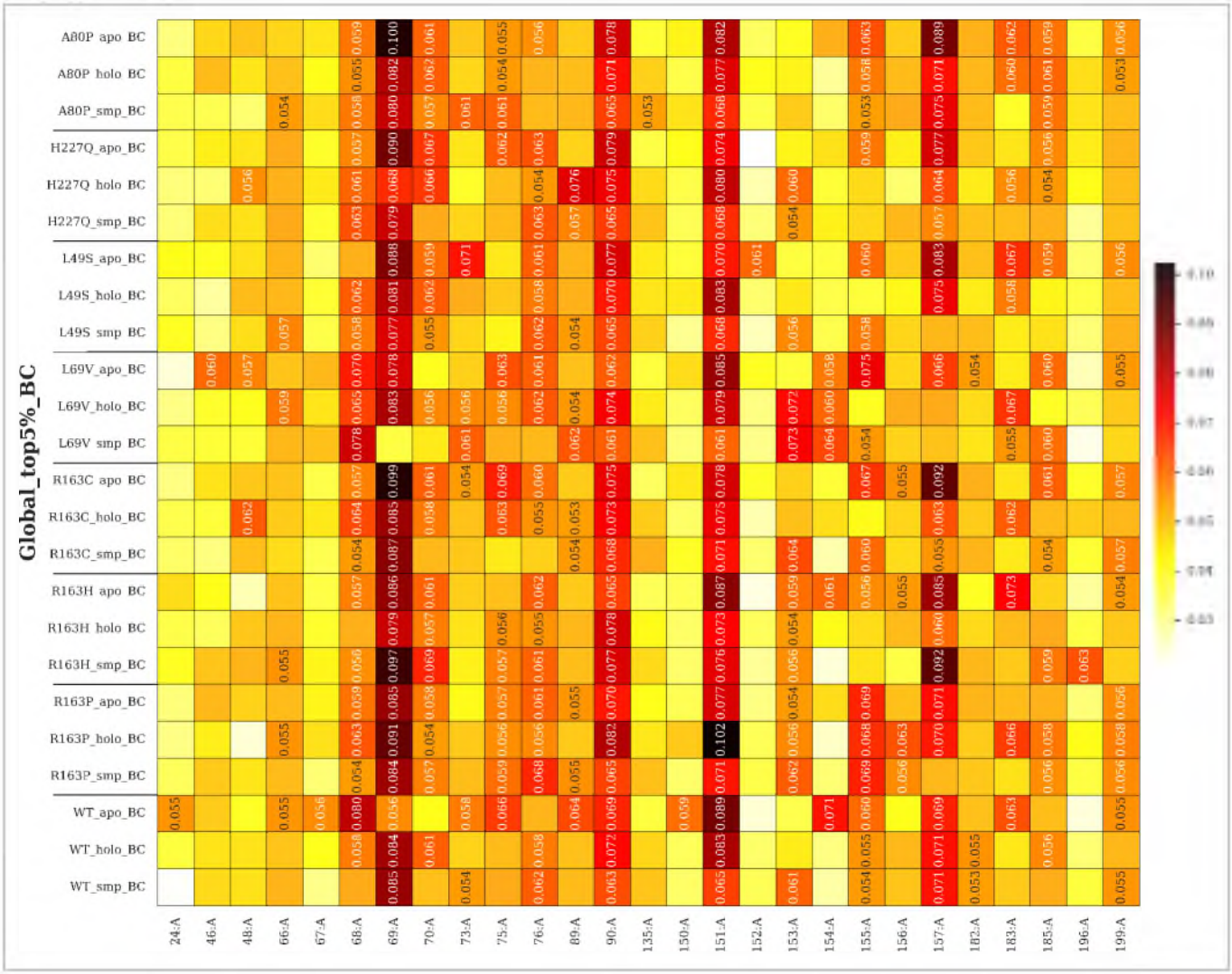


Figure 4.6: A heatmap of the potential hubs according to the global top 5% for averaged *BC* metric. The *x*-axis represents the protein residue positions while the *y*-axis is the WT and mutant enzymes in all the studied states. Detected hubs are marked with their centrality values while positions that were not identified as hubs are not but are still shown for comparison. Low to high centrality positions are coloured white to black.

Residue at position 90 lines the catalytic channel while 151 is next to position 152 which is in direct contact with the ligand. Position 69 was recorded as a *BC* hub in all the systems but was lost in the SMP-bound L69V variant. The residue at position 69 is also in direct contact with the ligand. Its loss as a hub in the drug-bound L69V mutation could mean that the substitution poses severe effects in intra-protein communications in the enzyme after drug binding. This could be a reason for the nearly undetectable enzymatic activity of the variant enzyme in experimental studies.

Figure 4.7 maps hubs in each individual mutant protein in comparison to the WT. The hubs seemed to be concentrated mainly around the catalytic site in all the systems. The occurrence of mutations resulted in the loss of some hubs with the gaining of new ones. This could have been in an attempt to accommodate the mutations in order to maintain the functionality of the enzyme. Figure S7a demonstrates the secondary structures where areas of high BC and top 5% BC hubs occurred while Table S3 represents all the identified hubs with the persistent hubs and top 5% hubs present in at least 12 studied systems being highlighted differently.

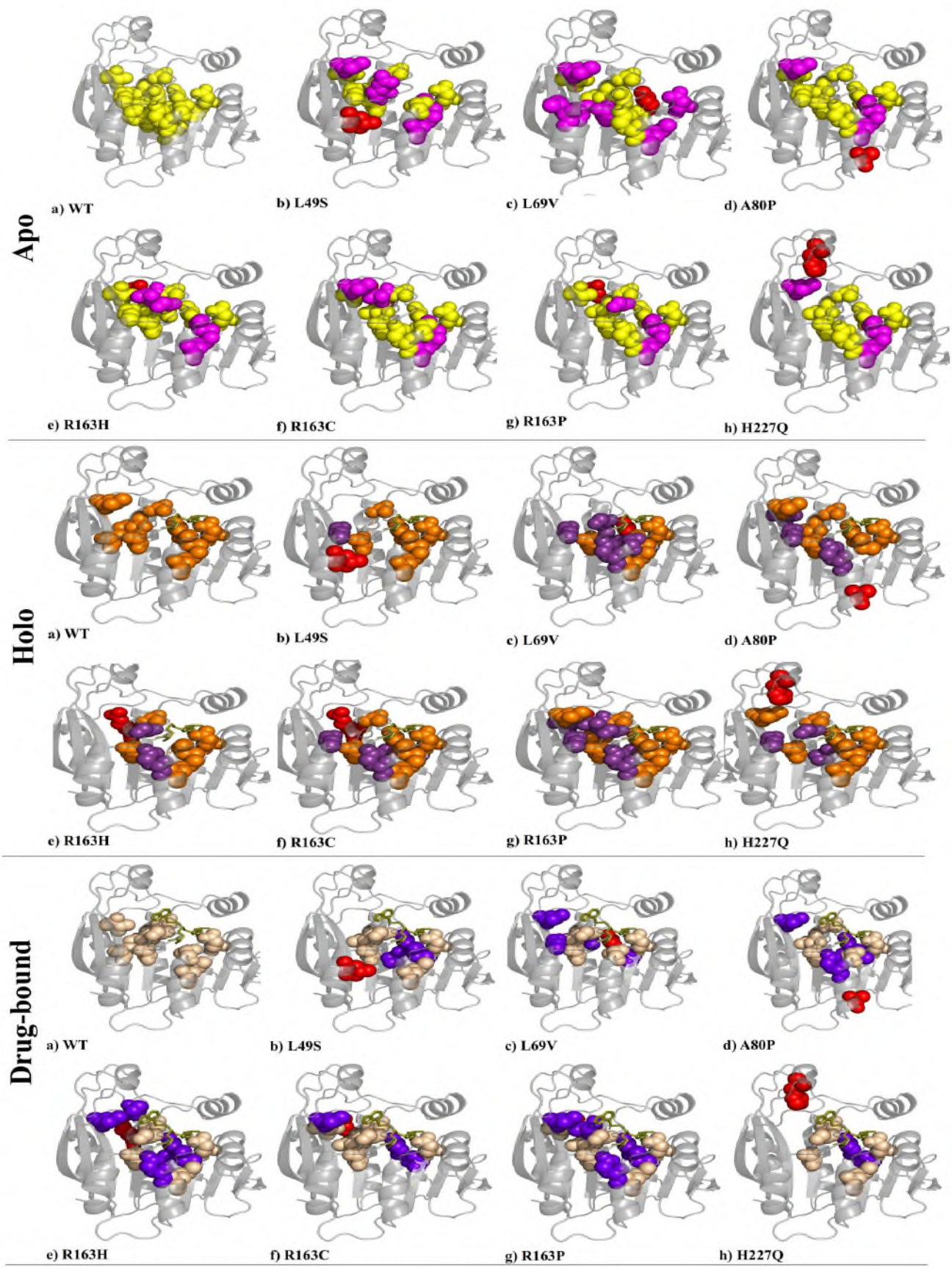


Figure 4.7: Structural mapping of global top 5% BC hubs in each of the systems in the apo, holo and drug-bound systems in comparison to the WT. In the apo systems, the WT hubs are represented as yellow spheres

while the hubs unique to each mutant system are magenta-coloured spheres. In the holo systems, the WT hubs are represented as orange spheres while the mutants are violet-purple spheres. In the drug-bound systems, the WT hubs are shown as wheat-coloured spheres while the mutant-specific hubs are purple-blue spheres. The mutant positions are represented as red spheres in all positions while SAH in holo and SMP in drug-bound systems are represented as deep-olive coloured sticks.

4.3.2.2 Eigenvector centrality (*EC*)

EC was calculated in order to identify which residues were connected to other important residues. Globally, regions of high *EC* were observed to lie on positions 65-76, 85-99, 118-121, 128-137 and 179-184 (Figure 4.8). The region 65-76 spans from β1 through a loop into αC (Figure S7b). Residues in this region surround the catalytic site and include L69 which is in direct contact with the ligand. The region 89-99 spans from β2-αD (Figure S7b) whose loop region residues also line the catalytic site. Region 118-121 falls on β4 (Figure S7b) which neighbours αD, β5 and β3 (Figure 1.2a). 128-137 spans from β5-αE (Figure S7b) with residues in interaction with the ligand while 179-184 falls on β7 (Figure S7b) which neighbours β9 (the last structure at the C-terminal) and β6 on both sides (Figure 1.2a). These regions were therefore connected to residues that form communication networks involved in ligand binding, formation of the catalytic channel and maintaining the structural integrity of the folded protein. The 179-184 region appeared to be a high *EC* region in the apo systems only with the exception of the three mutants occurring at position 163. This could offer an explanation for the structural changes that occur during catalysis as states change from apo to holo to the point of drug binding. Of the three mutants at position 163 that did not have the region as a high *EC* region, R163H which was the most affected seemed to have gained it as a high EC region in the holo and drug-bound systems. This could have been an attempt of the mutant enzyme to restore nativity after ligand binding to ensure functionality. The region was retained by H227Q in the holo state and L69V in the drug-bound state.

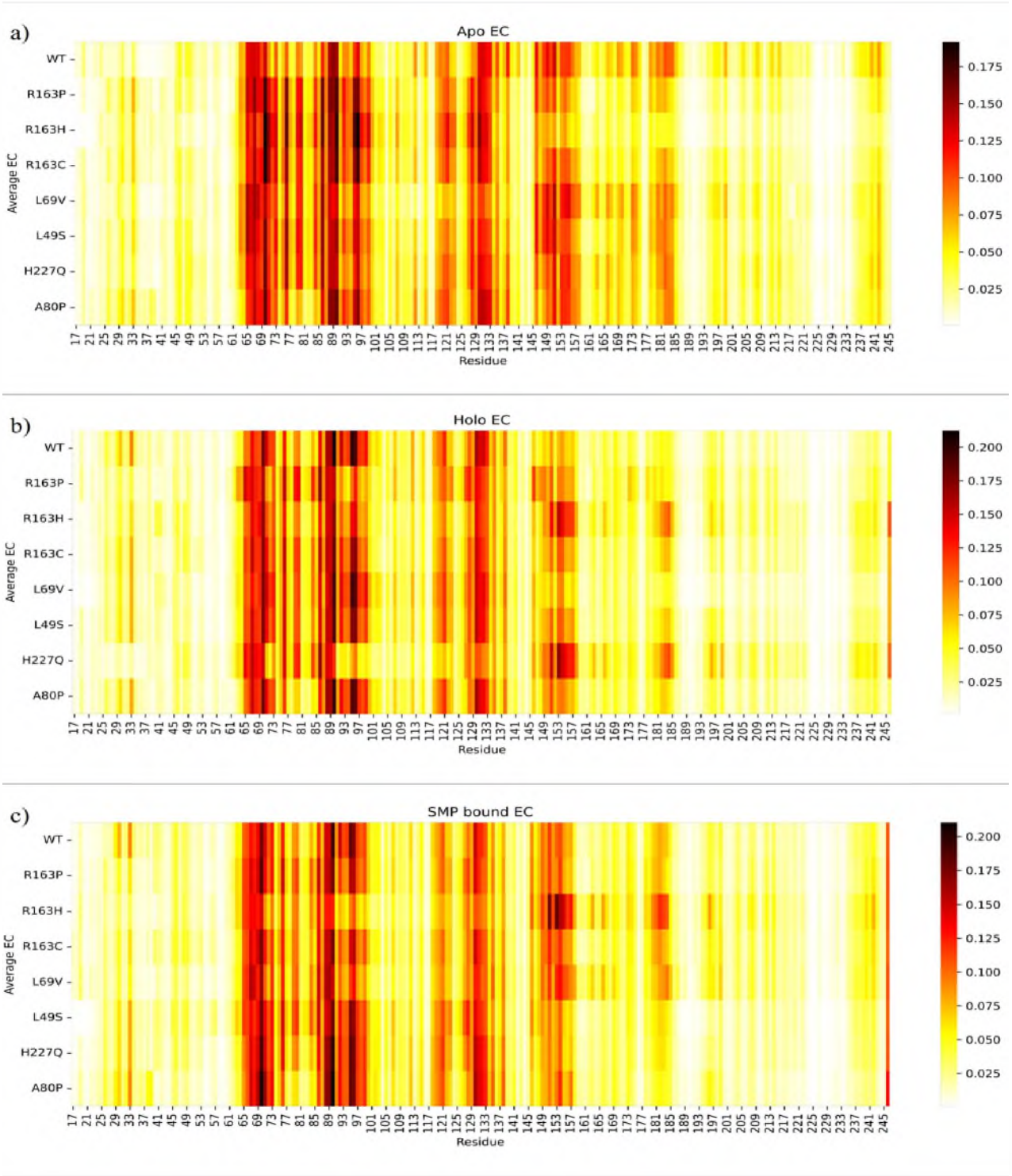


Figure 4.8: EC heatmaps of all the systems in all the states. The *y*-axis represents the WT and mutant systems while the *x*-axis represents the residue positions. A scale that increases colour intensity from white to black through yellow and red is used to represent the increase in *EC*. Darker colours show higher *EC* regions while lighter colours represent low *EC* regions.

After calculating the top 5% EC hubs in all the systems, only position 70 was identified as a persistent hub in all the studied systems (Figure 4.9). C70 forms non-covalent contacts with SAH

and is next to L69 which contacts SAH through a hydrogen bond and G71 which also forms non-covalent contacts with SAH (Figure S1).

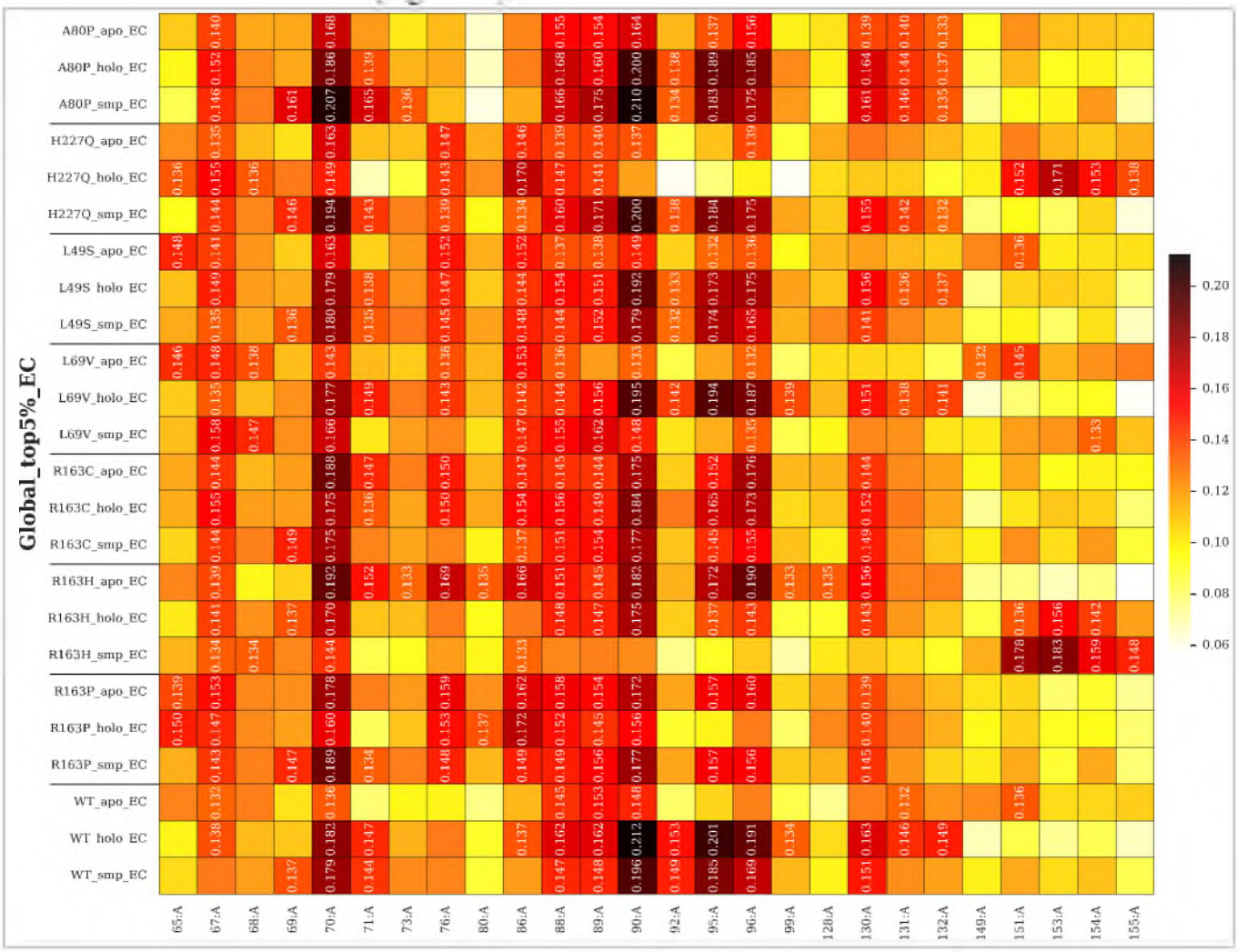


Figure 4.9: Heatmap representation of the global top 5% *EC* hubs in all the studied systems. The *x*-axis represents the protein residue positions while the *y*-axis is the WT and mutant enzymes in all the studied states. Detected hubs are marked with their centrality values. Low to high centrality positions are coloured white to black.

Mapping the identified hubs in individual structures (Figure 4.10) revealed a tendency of the hubs to concentrate on one side of the catalytic site in all the systems at positions mainly in β2, β5 and the beginning of αD. This could mean that the structures are important in maintaining protein dynamics and functionality. The occurrence of mutations had an influence on how the hubs shifted with the biggest shift being visible on the drug-bound R163H mutant. The loss of a hub often resulted in the gaining of a new one in a nearby region probably in an attempt to maintain functionality. The A80P mutant holoenzyme however recorded a loss in major hubs without gaining any new ones.

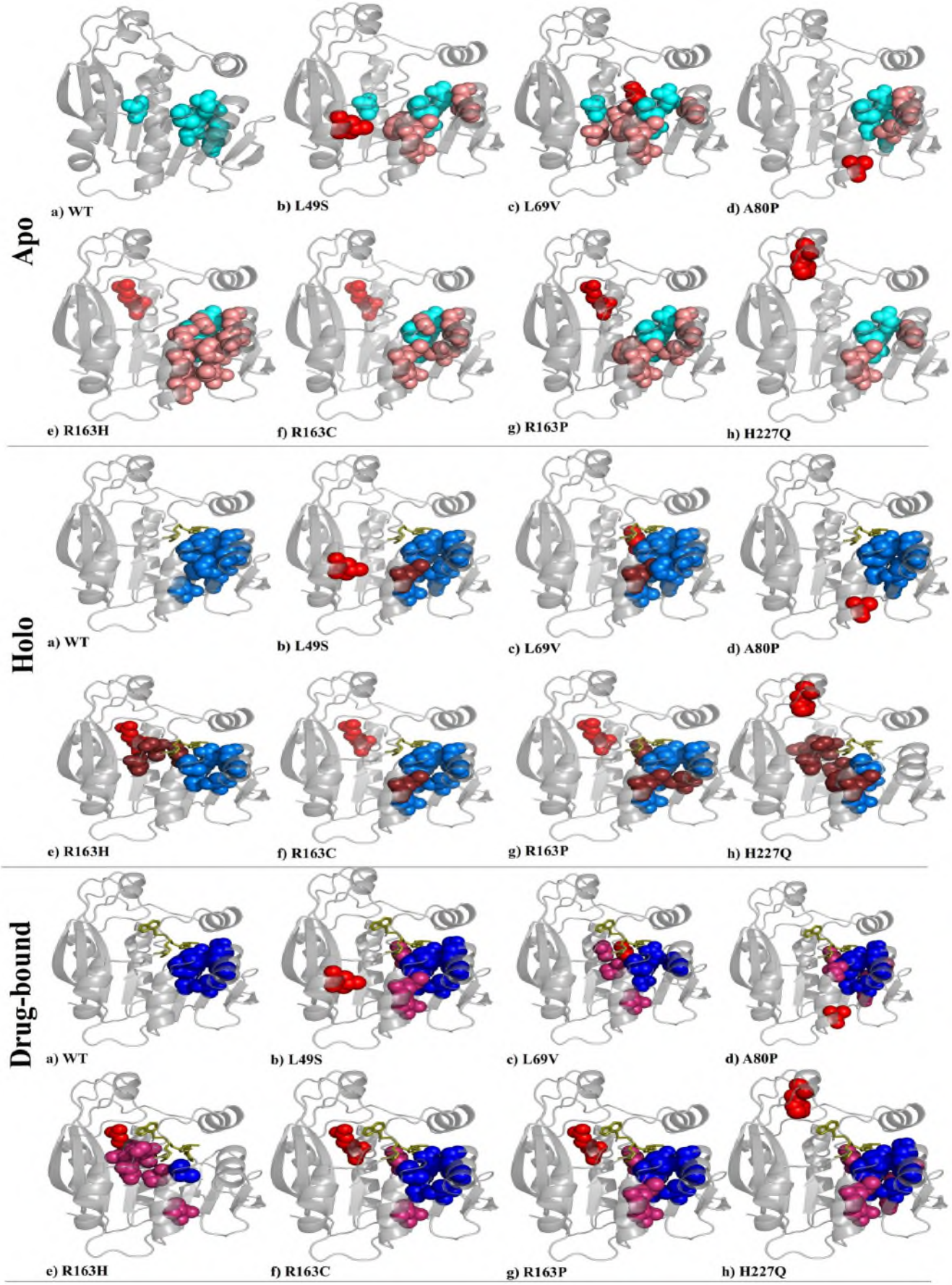


Figure 4.10: Structural mapping of the global top 5% EC hubs. WT hubs in the apo systems are represented as cyan spheres while those in the mutant systems are represented as salmon spheres. In the holo systems,

identified WT hubs are represented as marine-coloured spheres while mutant-specific hubs are represented as ruby-coloured spheres. In the drug-bound systems, WT hubs are shown as blue-coloured spheres while mutant-specific hubs are represented as warm-pink-coloured spheres. Mutant positions in all the systems are represented as red spheres while SAH in holo and SMP in the drug-bound systems are represented as deep-olive sticks.

4.4 Conclusion

Residue contact map analysis revealed major changes in interactions that occurred due to mutations at the various points of mutation. Analysis of SMP interactions revealed possible reorientation at the catalytic site as a result of A80P, R163H and L69V mutations with allosteric effect being noted in the A80P substitution. Changes in interactions could be attributed to the different physico-chemical properties involved or size of amino acid side chains. The occurrence of mutations also affected global intra-protein communication networks with the loss of important centrality hubs and the gaining of new ones in most of the systems. This could be associated with changes in protein dynamics which could affect catalytic activity. Table S3 highlights hubs that were common in at-least 12 of the 24 studied systems and persistent hubs in all the systems in either (BC/EC) of the analysed centralities.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Concluding remarks

The reduced catalytic activity of TPMT in metabolising 6MP, a chemotherapeutic agent in ALL treatment, has been associated with the occurrence of genetic variations. Variations that result in missense mutations cause amino-acid changes and in turn alter the polypeptide sequence of the protein. This could alter the structural dynamics of the enzyme and hence its functionality. This study sort to investigate the effects of seven amino-acid substitutions (L49S, L69V, A80P, R163H, R163C, R163P and H227Q) on the structural dynamics of TPMT and their implications on the metabolism of 6MP. The substitutions were selected based on their effect on catalytic activity, location on the protein structure and conservation. The information sourced in the selection criteria was from variant analysis predictions, clinical data on Ensembl and PharmGKB and information from literature. All the selected variants were predicted to be deleterious with the exception of H227Q which was predicted to be benign. Reports from studies indicate that positions 49, 69, 80 and 163 occurred in highly conserved regions while 227 was less conserved. This could explain why H227Q was mostly benign from variant analysis as most of the prediction tools use algorithms that apply multiple sequence alignment methods to perform analysis. Positions 80 and 227 occur on the surface of the enzyme's structure at solvent-exposed positions while 49, 69 and 163 are buried in the structure with 69 occurring at the binding site and in contact with the ligand.

RMSD and Rg analysis post-MD revealed that the binding of the co-factor stabilized the WT enzyme. Comparison of the WT with the mutant systems in RMSD and Rg analysis largely revealed that mutations affect structural global motions and compactness of the enzyme. Binding of the drug largely seemed to stabilize the mutant systems as observed in RMSD and Rg analysis. Correlated atomic movements were not significantly affected by the occurrence of mutations according to DCC analysis. Some mutations seemed to cause perturbations in the stability of the structure in the different states while others such as A80P seemed to cause consistent stability on the structure of the protein contrary to what was expected. This could have a deleterious effect on protein functionality as excessive rigidity might interfere with effective ligand binding. RMSF analysis showed changes

in local chain flexibility especially in the drug-bound systems in the presence of mutations in regions engaged in drug binding. This could be used to draw inference on altered 6MP metabolism.

In network analysis, weighted contact maps for the mutation positions and SMP in the WT and mutant systems were calculated. Further analysis was done to calculate *BC* and *EC* and to further identify global top 5% hubs for these centrality metrics. The occurrence of mutations appeared to affect interactions with the residue at the position of the mutation. The different physio-chemical properties of the amino-acids involved could have resulted in shifts in interactions. The H227Q substitution showed an attempt to accommodate the mutation by gaining interactions with close-by residues that had similar physico-chemical properties as the lost interactions. Attempts to understand why the A80P mutation appeared to induce structural stability in RMSD analysis revealed enhanced inter-residue interactions at the surface due to reduced conformational entropy. The mutation however showed an allosteric effect that could introduce changes at the catalytic site and thus can be classified as an allosteric mutation [147]. Analysis of SMP interactions in the presence of mutations in comparison with the WT revealed that A80P, R163H and L69V mutations affected how the ligand oriented itself at the catalytic site. This could have interfered with the effective methylation of the drug. The rest of the mutant systems did not lead to significant changes in SMP interactions therefore their reduced catalytic activity might not have been associated with changes in ligand-residue interactions.

In the analysis of the centrality metrics, global regions identified to be of high *BC* and *EC* spanned secondary structures that lined the catalytic cavity or were involved in conferring structural stability. Some regions overlapped in the two analyzed metrics. Such regions included 65-76 which was in the same location in both metrics, 89-90 in *BC* overlapped 85-99 in *EC* and 182-185 in *BC* overlapped with 179-184 in *EC*. The first two regions are found in secondary structures with residues involved in the formation of the catalytic site which could mean that they are part of functionally important communication networks. The last mentioned overlapping regions fall on β7 which interacts with β9 and β6 at the C-terminus which could mean that the residues in the region are involved in communication networks that confer structural integrity of the enzyme. Persistent hubs in the global top 5% *BC* calculations were identified at positions 90 and 151. The proximity of these positions to the catalytic site could mean that the two hubs lie in functionally important network paths. Position 70 was the only identified persistent hub in the global top 5% *EC* calculations. The residue C70 is involved in unbonded interactions with SAH and interacts with

L69 and G71 which are in direct contact with the ligands. This would explain its status as a persistent *EC* hub.

5.2 Future work

This study covered the effects of amino-acid substitutions on the structure and function of TPMT on 6MP metabolism by identifying changes in global motions and interaction networks applying *in silico* techniques. As only EC and BC calculations were done, more centrality value analyses such as *DC, L* and *CC* could be performed to calculate the number of neighbouring nodes around the node of interest, average shortest paths to the node and central nodes which are closer to other nodes. Further, contact map calculations around the identified persistent hubs to identify changes in communication around these important residues could be performed. In addition, perturbation response scanning could be performed to identify residues whose perturbation invokes conformational displacement within the protein. This study only looked into 7 variant alleles out of the reported 46 alleles and filled the gap in understanding how the mutant enzymes affect 6MP metabolism applying *in silico* techniques. The remaining ones can be studied in future work with the application of molecular dynamics.

References

1. Rehman A, Abbas N, Saba T, Rahman SI ur, Mehmood Z, Kolivand H. Classification of acute lymphoblastic leukemia using deep learning. Microsc Res Tech. 2018;81(11):1310–7.

2. Onciu M. Acute Lymphoblastic Leukemia. Hematol Clin NA [Online]. 2009;23(4):655–74. Available from: http://dx.doi.org/10.1016/j.hoc.2009.04.009

3. Speck NA, Gilliland DG. Core-binding factors in haematopoiesis and leukaemia. Nat Rev Cancer [Online]. 2002;2(7):502–13. Available from: https://doi.org/10.1038/nrc840

4. Pui C hon, Relling M V, Downing JR. Acute Lymphoblastic Leukemia. 2004;1535–48.

5. Loh ML, Rubnitz JE. TEL/AML1-positive pediatric leukemia: prognostic significance and therapeutic approaches. Curr Opin Hematol [Online]. 2002;9(4). Available from: https://journals.lww.com/co-hematology/Fulltext/2002/07000/TEL_AML1_positive_pediatric_leukemia__prognostic.13.aspx

6. Shivakumar R, Tan W, Wilding GE, Wang ES, Wetzler M. Biologic features and treatment outcome of secondary acute lymphoblastic leukemia - A review of 101 cases. Ann Oncol. 2008;19(9):1634–8.

7. Ching-Hon Pui, Leslie L Robison ATL. Acute lymphoblastic leukaemia. Methods Mol Biol. 2008;371:99–117.

8. Puckett Y, Chan O. Acute Lymphocytic Leukemia. In Treasure Island (FL); 2022.

9. Spector LG, Ross JA, Robison LL, Bhatia S. Epidemiology and etiology. In: Pui CH, editor. Childhood Leukemias [Online]. 2nd ed. Cambridge: Cambridge University Press; 2006 [accessed 2022 Dec 15]. p. 48–66. Available from: https://www.cambridge.org/core/books/childhood-leukemias/epidemiology-and-etiology/8925C88C85414C8ADBB0B536D76D7787

10. Abbasi S, Maleha F, Shobaki M. Acute lymphoblastic leukemia experience: Epidemiology and outcome of two different regimens. Mediterr J Hematol Infect Dis. 2013;5(1).

11. Urayama KY, Manabe A. [Genomic evaluations of childhood acute lymphoblastic leukemia susceptibility across race/ethnicities]. Rinsho Ketsueki. 2014 Oct;55(10):2242–8.

12. Pui C hon, Evans WE. TRATAMIENTO LLA nejm pui2006. Th e new engl J Med Rev. 2006;354(2):166–78.

13. Cordo' V, van der Zwet JCG, Canté-Barrett K, Pieters R, Meijerink JPP. T-cell Acute Lymphoblastic Leukemia: A Roadmap to Targeted Therapies. Blood Cancer Discov. 2021;2(1):19–31.

14. Larson R a. PDQ Adult Acute Lymphoblastic Leukemia Treatment. Natl Cancer Inst [Online]. 2014;(table 1):1–18. Available from: http://www.uptodate.com/contents/acute-lymphoblastic-leukemia-all-treatment-in-adults-beyond-the-basics?view=print#

15. Toyoda Y, Manabe A, Tsuchida M, Hanada R, Ikuta K, Okimoto Y, et al. Six months of maintenance chemotherapy after intensified treatment for acute lymphoblastic leukemia of childhood. J Clin Oncol Off J Am Soc Clin Oncol. 2000 Apr;18(7):1508–16.

16. Elion GB, Hitchings GH. The Synthesis of 6-Thioguanine. J Am Chem Soc [Online]. 1955 Mar 1;77(6):1676. Available from: https://doi.org/10.1021/ja01611a082

17. Saxena S, Krishna Murthy TP, Chandrashekhar CR, Patil LS, Aditya A, Shukla R, et al. A bioinformatics approach to the identification of novel deleterious mutations of human TPMT through validated screening and molecular dynamics. Sci Rep [Online]. 2022;12(1):1–21. Available from: https://doi.org/10.1038/s41598-022-23488-z

18. Zaza G, Cheok M, Krynetskaia N, Thorn C, Stocco G, Hebert JM, et al. Thiopurine pathway. Pharmacogenet Genomics [Online]. 2010 Sep;20(9):573–4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf

19. Ujiie S, Sasaki T, Mizugaki M, Ishikawa M, Hiratsuka M. Functional characterization of 23 allelic variants of thiopurine S-methyltransferase gene (TPMT*2 - *24). Pharmacogenet Genomics. 2008;18(10):887–93.

20. Murugesan R, Vahab S abdul, Satyajit P, Rekha R, Jyothi R, Rai P, et al. Thiopurine S-methyltransferase alleles, TPMT*2, *3B and *3C, and genotype frequencies in an Indian population. Exp Ther Med. 2010;121–7.

21. Zimdahl Kahlin A, Helander S, Skoglund K, Söderkvist P, Mårtensson LG, Appell ML. Comprehensive study of thiopurine methyltransferase genotype, phenotype, and genotype-phenotype discrepancies in Sweden. Biochem Pharmacol. 2019;164:263–72.

22. Sharma H, Wadhwa R. Mercaptopurine. In Treasure Island (FL); 2022.

23. Beck JT, Ismail A, Tolomeo C. Targeting the phosphatidylinositol 3-kinase (PI3K)/AKT/mammalian target of rapamycin (mTOR) pathway: an emerging treatment strategy for squamous cell lung carcinoma. Cancer Treat Rev. 2014 Sep;40(8):980–9.

24. Genet N. NUDT15 Polymorphisms Alter Thiopurine Metabolism and Hematopoietic Toxicity HHS Public Access Author manuscript. 2016;48(4):367–73.

25. Karran P, Attard N. Thiopurines in current medical practice: molecular mechanisms and contributions to therapy-related cancer. Nat Rev Cancer. 2008 Jan;8(1):24–36.

26. Chouchana L, Narjoz C, Beaune P, Loriot MA, Roblin X. Review article: the benefits of pharmacogenetics for improving thiopurine therapy in inflammatory bowel disease. Aliment Pharmacol Ther. 2012 Jan;35(1):15–36.

27. Dervieux T, Brenner TL, Hon YY, Zhou Y, Hancock ML, Sandlund JT, et al. De novo purine synthesis inhibition and antileukemic effects of mercaptopurine alone or in combination with methotrexate in vivo. Blood. 2002 Aug;100(4):1240–7.

28. Cara CJ, Pena AS, Sans M, Rodrigo L, Guerrero-Esteo M, Hinojosa J, et al. Reviewing the mechanism of action of thiopurine drugs: towards a new paradigm in clinical practice. Med Sci Monit Int Med J Exp Clin Res. 2004 Nov;10(11):RA247-54.

29. Fernández-Ramos AA, Marchetti-Laurent C, Poindessous V, Antonio S, Laurent-Puig P, Bortoli S, et al. 6-mercaptopurine promotes energetic failure in proliferating T cells. Oncotarget. 2017;8(26):43048–60.

30. Huang HY, Chang HF, Tsai MJ, Chen JS, Wang MJ. 6-Mercaptopurine attenuates tumor necrosis factor-α production in microglia through Nur77-mediated transrepression and PI3K/Akt/mTOR signaling-mediated translational regulation. J Neuroinflammation [Online]. 2016;13(1):1–20. Available from: http://dx.doi.org/10.1186/s12974-016-0543-5

31. Fernández-Ramos AA, Poindessous V, Marchetti-Laurent C, Pallet N, Loriot MA. The effect of immunosuppressive molecules on T-cell metabolic reprogramming. Biochimie. 2016 Aug;127:23–36.

32. Ordentlich P, Yan Y, Zhou S, Heyman RA. Identification of the antineoplastic agent 6-mercaptopurine as an activator of the orphan nuclear hormone receptor Nurr1. J Biol Chem. 2003 Jul;278(27):24791–9.

33. Yoo YG, Na TY, Yang WK, Kim HJ, Lee IK, Kong G, et al. 6-Mercaptopurine, an activator of Nur77, enhances transcriptional activity of HIF-1alpha resulting in new vessel formation. Oncogene. 2007 May;26(26):3823–34.

34. Rodríguez-Enríquez S, Marín-Hernández A, Gallardo-Pérez JC, Carreño-Fuentes L, Moreno-Sánchez R. Targeting of cancer energy metabolism. Mol Nutr \& Food Res [Online]. 2009;53(1):29–48. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/mnfr.200700470

35. Phan LM, Yeung SCJ, Lee MH. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. Cancer Biol Med. 2014 Mar;11(1):1–19.

36. Herbel C, Patsoukis N, Bardhan K, Seth P, Weaver JD, Boussiotis VA. Clinical significance of T cell metabolic reprogramming in cancer. Clin Transl Med. 2016;5(1).

37. Heiden MG Vander, Cantley LC, Thompson CB. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. Encycl Neurosci. 2009;1029–1033.

38. Gatenby RA, Gillies RJ. Why do cancers have high aerobic glycolysis? Nat Rev Cancer. 2004 Nov;4(11):891–9.

39. Palsson-McDermott EM, O'Neill LAJ. The Warburg effect then and now: from cancer to inflammatory diseases. Bioessays. 2013 Nov;35(11):965–73.

40. Salavaggione OE, Wang L, Wiepert M, Yee VC, Weinshilboum RM. Thiopurine S-methyltransferase pharmacogenetics: Variant allele functional and comparative genomics. Pharmacogenet Genomics. 2005;15(11):801–15.

41. Li JJ, Sun C, Cai W, Li JJ, Rosen BP, Chen J. Insights into S-adenosyl-L-methionine (SAM)-dependent methyltransferase related diseases and genetic polymorphisms. Mutat Res - Rev Mutat Res [Online]. 2021;788(October):108396. Available from: https://doi.org/10.1016/j.mrrev.2021.108396

42. Karczewski KJ, Daneshjou R, Altman RB. Chapter 7: Pharmacogenomics. PLoS Comput Biol. 2012;8(12).

43. Pui C hon. Precision medicine in acute lymphoblastic leukemia. 2020;14(6):689–700.

44. Evans WE, Relling M V. Nature02626. 2004;429(May).

45. Relling M V, Pui CH, Cheng C, Evans WE. Thiopurine methyltransferase in acute lymphoblastic leukemia. Vol. 107, Blood. United States; 2006. p. 843–4.

46. Stanulla M, Schaeffeler E, Flohr T, Cario G, Schrauder A, Zimmermann M, et al. Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia. JAMA. 2005 Mar;293(12):1485–9.

47. Relling M V, Rubnitz JE, Rivera GK, Boyett JM, Hancock ML, Felix CA, et al. High incidence of secondary brain tumours after radiotherapy and antimetabolites. Lancet (London, England). 1999 Jul;354(9172):34–9.

48. Remy CN. Metabolism of Thiopyrimidines and Thiopurines. J Biol Chem [Online]. 1963;238(3):1078–84. Available from: http://dx.doi.org/10.1016/S0021-9258(18)81262-5

49. Blissing A. Thiopurine S-methyltransferase - characterization of variants and ligand binding. Thiopurine S-methyltransferase - characterization of variants and ligand binding. 2017.

50. Schubert HL, Blumenthal RM, Cheng X. Many paths to methyltransfer: A chronicle of convergence. Trends Biochem Sci. 2003;28(6):329–35.

51. Scheuermann TH, Keeler C, Hodsdon ME. Consequences of binding an S-adenosylmethionine analogue on the structure and dynamics of the thiopurine methyltransferase protein backbone. Biochemistry. 2004;43(38):12198–209.

52. Wu H, Horton JR, Battaile K, Allali-Hassani A, Martin F, Zeng H, et al. Structural basis of allele variation of human thiopurine-S-methyltransferase. Proteins Struct Funct Bioinforma [Online]. 2007;67(1):198–208. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21272

53. Peng Y, Feng Q, Wilk D, Adjei AA, Salavaggione OE, Weinshilboum RM, et al. Structural basis of substrate recognition in thiopurine S-methyltransferase. Biochemistry [Online]. 2008 Jun 1;47(23):6216–25. Available from: https://doi.org/10.1021/bi800102x

54. Mokmak W, Tongsima S, Jenwitheesuk E. Molecular dynamics simulation of a human thiopurine S-methyltransferase complexed with 6-mercaptopurine model. Bioinformation [Online]. 2009 Sep 5;4(2):59–62. Available from: http://www.bioinformation.net/004/001200042009.htm

55. Fazel-Najafabadi E, Vahdat Ahar E, Fattahpour S, Sedghi M. Structural and functional impact of missense mutations in TPMT: An integrated computational approach. Comput Biol

Chem [Online]. 2015;59:48–55. Available from: http://dx.doi.org/10.1016/j.compbiolchem.2015.09.004

56. Katara P, Kuntal H. TPMT Polymorphism: When Shield Becomes Weakness. Interdiscip Sci – Comput Life Sci. 2016;8(2):150–5.

57. Appell ML, Berg J, Duley J, Evans WE, Kennedy MA, Lennard L, et al. Nomenclature for alleles of the thiopurine methyltransferase gene. Pharmacogenet Genomics. 2013;23(4):242–8.

58. TPMT nomenclature [Internet]. 2019 [accessed 2022 Jul 30]. Available from: https://liu.se/en/research/tpmt-nomenclature-committee

59. Alves S, Rocha J, Amorim A, Prata MJ. Tracing the Origin of the Most Common Thiopurine Methyltransferase (TPMT) Variants: Preliminary Data from the Patterns of Haplotypic Association with Two CA Repeats. Ann Hum Genet [Online]. 2004 Jul;68(4):313–23. Available from: http://doi.wiley.com/10.1046/j.1529-8817.2004.00104.x

60. Wang L, Pelleymounter L, Weinshilboum R, Johnson JA, Hebert JM, Altman RB, et al. Very important pharmacogene summary: Thiopurine S-methyltransferase. Pharmacogenet Genomics. 2010;20(6):401–5.

61. Garat A, Cauffiez C, Renault N, Lo-Guidice JM, Allorge D, Chevalier D, et al. Characterisation of novel defective thiopurine S-methyltransferase allelic variants. Biochem Pharmacol. 2008;76(3):404–15.

62. Høie MH, Cagiada M, Beck Frederiksen AH, Stein A, Lindorff-Larsen K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. Cell Rep [Online]. 2022;38(2):110207. Available from: https://doi.org/10.1016/j.celrep.2021.110207

63. Otterness D, Szumlanski C, Lennard L, Klemetsdal B, Aarbakke J, Park-Hah JO, et al. Human thiopurine methyltransferase pharmacogenetics: Gene sequence polymorphisms. Clin Pharmacol Ther. 1997;62(1):60–73.

64. Wang L, Nguyen T V., McLaughlin RW, Sikkink LA, Ramirez-Alvarado M, Weinshilboum RM. Human thiopurine S-methyltransferase pharmacogenetics: Variant allozyme misfolding and aggresome formation. Proc Natl Acad Sci U S A. 2005;102(26):9394–9.

65. Hamdan-Khalil R, Gala JL, Allorge D, Lo-Guidice JM, Horsmans Y, Houdret N, et al. Identification and functional analysis of two rare allelic variants of the thiopurine S-methyltransferase gene, TPMT*16 and TPMT*19. Biochem Pharmacol. 2005;69(3):525–9.

66. Larif S, Ben Salem C, Soua Z, Hmouda H, Bouraoui K. Insight into TPMT*23 mutation mis-folding using molecular dynamics simulation and protein structure analysis. J Biomol Struct Dyn. 2013;31(10):1066–76.

67. Krawczak M, Thomas NST, Hundrieser B, Mort M, Wittig M, Hampe J, et al. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat. 2007 Feb;28(2):150–8.

68. Robert F, Pelletier J. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. Front Genet. 2018;9(October):1–11.

69. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. Science (80-). 1998;280(5366):1077–82.

70. Ahmad T, Valentovic MA, Rankin GO. Effects of cytochrome P450 single nucleotide polymorphisms on methadone metabolism and pharmacodynamics. Biochem Pharmacol [Online]. 2018;153(December 2017):196–204. Available from: https://doi.org/10.1016/j.bcp.2018.02.020

71. Nusrath A, Raiza P.T B. Review on Single Nucleotide Polymorphism Analysis Methods. Internaional J Eng Res Technol [Online]. 2015;3(30):1–4. Available from: http://pmd.ddbj.nig.ac.jp

72. Hunt R, Sauna ZE, Ambudkar S V, Gottesman MM, Kimchi-Sarfaty C. Silent (synonymous) SNPs: should we care about them? Methods Mol Biol. 2009;578:23–39.

73. Schaeffeler E, Zanger UM, Eichelbaum M, Asante-Poku S, Shin JG, Schwab M. Highly multiplexed genotyping of thiopurine S-methyltransferase variants using MALDI-TOF mass spectrometry: Reliable genotyping in different ethnic groups. Clin Chem. 2008;54(10):1637–47.

74. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.

75. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods [Online]. 2010;7(4):248–9. Available from: http://dx.doi.org/10.1038/nmeth0410-248

76. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33(SUPPL. 2):306–10.

77. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One. 2012;7(10).

78. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013;34(1):57–65.

79. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729–34.

80. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(W1):452–7.

81. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81.

82. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.

83. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins Struct Funct Genet. 2006;62(4):1125–32.

84. Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP prediction: Be mindful of your training data! Bioinformatics. 2007;23(6):664–72.

85. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics. 2005;21(10):2185–90.

86. Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006;356(5):1263–74.

87. Bromberg Y, Rost B. SNAP: Predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823–35.

88. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45(D1):D183–9.

89. Yin S, Ding F, Dokholyan N V. Eris: an automated estimator of protein stability. Nat Methods [Online]. 2007 Dec 29;4:466+. Available from: https://link.gale.com/apps/doc/A181818633/HRCA?u=anon~2c852558&sid=googleScholar&xid=b8c4aeb8

90. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: Prediction of protein stability upon point mutations. Nucleic Acids Res. 2006;34(WEB. SERV. ISS.):239–42.

91. Brown DK, Tastan Bishop Ö. HUMA: A platform for the analysis of genetic variation in humans. Hum Mutat. 2018;39(1):40–51.

92. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLoS Comput Biol. 2014;10(1):1–11.

93. Mahfuz AMUB, Khan MA, Deb P, Ansary SJ, Jahan R. Identification of deleterious single nucleotide polymorphism (SNP)s in the human TBX5 gene & prediction of their structural & functional consequences: An in silico approach. Vol. 28, Biochemistry and Biophysics Reports. 2021.

94. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005;15(7):978–86.

95. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005 Jul;33(Web Server issue):W480-2.

96. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: Server and survey. Nucleic Acids Res. 2002;30(17):3894–900.

97. Schmidt S, Gerasimova A, Kondrashov FA, Adzuhbei IA, Kondrashov AS, Sunyaev S. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet. 2008;4(11).

98. Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;12(3):436–46.

99. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863–74.

100. Ensembl [Internet]. [accessed 2022 Jul 25]. Available from: https://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Table?db=core;g=ENSG00000137364;r=6:18128311-18155077;t=ENST00000309983

101. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res [Online]. 2021;50(D1):D988–95. Available from: https://doi.org/10.1093/nar/gkab1049

102. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan;29(1):308–11.

103. Nomenclature C. TPMT allele nomenclature committee The table defines all the single-nucleotide polymorphisms (SNPs) in TPMT as of Nov 2022. 2022;

104. Whirl-Carrillo M, Huddart R, Gong L, Sangkuhl K, Thorn CF, Whaley R, et al. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. Clin Pharmacol Ther. 2021;110(3):563–72.

105. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92(4):414–7.

106. Wennerstrand P, Dametto P, Hennig J, Klingstedt T, Skoglund K, Lindqvist Appell M, et al. Structural characteristics determine the cause of the low enzyme activity of two thiopurine S-methyltransferase allelic variants: A biophysical characterization of TPMT*2 and TPMT*5. Biochemistry. 2012;51(30):5912–20.

107. Schaeffeler E, Eichelbaum M, Reinisch W, Zanger UM, Schwab M. Three novel thiopurine S-methyltransferase allelic variants (TPMT*20, *21, *22) - association with decreased enzyme function. Hum Mutat. 2006;27(9):976.

108. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow G V., et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic

and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 2021;49(1):D437–51.

109. Maestro, Schrödinger, LLC, New York, NY, 2021 [Internet]. Available from: https://www.schrodinger.com/citations#Maestro

110. Krynetski EY, Schuetz JD, Galpin AJ, Pui CH, Relling M V., Evans WE. A single point mutation leading to loss of catalytic activity in human thiopurine S-methyltransferase. Proc Natl Acad Sci U S A. 1995;92(4):949–53.

111. PharmGKB TPMT*22 [Internet]. [accessed 2022 Sep 10]. Available from: https://www.pharmgkb.org/haplotype/PA165819292

112. David L. Nelson MMC. Lehninger Principles of Biochemistry 5th Edition. 2008. 76,84.

113. Hernández-Rodríguez M, Rosales-Hernández MC, Mendieta-Wejebe JE, Martínez-Archundia M, Basurto JC. Current Tools and Methods in Molecular Dynamics (MD) Simulations for Drug Design. Curr Med Chem. 2016;23(34):3909–24.

114. Leach AR. Empirical Force Field Models. Molecular Modelling: Principles & Applications. 2001.

115. Konteatis ZD, Klon AE, Zou J, Meshkat S. Computational approach to de novo discovery of fragment binding for novel protein states [Internet]. 1st ed. Vol. 493, Methods in Enzymology. Elsevier Inc.; 2011. 357–380 p. Available from: http://dx.doi.org/10.1016/B978-0-12-381274-2.00014-5

116. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol [Online]. 2002 Sep;9(9):646–52. Available from: https://doi.org/10.1038/nsb0902-646

117. Vanommeslaeghe K, Guvench O, MacKerell AD. Molecular mechanics. Curr Pharm Des [Online]. 2014 May 31;20(20):3281–92. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1381-6128&volume=20&issue=20&spage=3281

118. Lemkul J. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. Living J Comput Mol Sci [Online]. 2019;1(1):0–53. Available from: https://www.livecomsjournal.org/article/5068-from-proteins-to-perturbed-hamiltonians-a-suite-of-tutorials-for-the-gromacs-2018-molecular-simulation-package-article-v1-0

119. Scott A. Hollingsworth, Ron O. Dror. Molecular dynamics simulation for all. Neuron. 2018;99(6)(1):1129–43.

120. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KMJ, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005 Dec;26(16):1668–88.

121. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem [Online]. 1983;4(2):187–217. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540040211

122. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005 Dec;26(16):1781–802.

123. Bowers KJ, Chow DE, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In: ACM/IEEE SC 2006 Conference (SC'06) [Online]. IEEE; 2006. p. 43–43. Available from: https://ieeexplore.ieee.org/document/4090217/

124. Plimpton S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. J Comput Phys [Online]. 1995;117(1):1–19. Available from: https://www.sciencedirect.com/science/article/pii/S002199918571039X

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

126. Abraham MWM, Van Der Spoel D, Lindahl, ErikApol E, Hess B, Apostolov R, Berendsen HJC, et al. GROMACS 2016 Reference Manual. GROMACS 2016 Ref Man [Online]. 2018; Available from: www.gromacs.org

127. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins Struct Funct Bioinforma [Online]. 2006 Nov 15;65(3):712–25. Available from: https://onlinelibrary.wiley.com/doi/10.1002/prot.21123

128. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinforma. 2010;78(8):1950–8.

129. H++ [Internet]. [accessed 2022 Aug 12]. Available from: http://newbiophysics.cs.vt.edu/H++

130. Anandakrishnan R, Aguilar B, Onufriev A V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Res [Online]. 2012 Jul 1;40(W1):W537–41. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks375

131. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J Chem Theory Comput [Online]. 2013 Jul 9;9(7):3084–95. Available from: https://pubs.acs.org/doi/10.1021/ct400341p

132. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J Mol Graph [Online]. 1996;14(1):33–8. Available from: https://www.sciencedirect.com/science/article/pii/0263785596000185

133. Kuzmanic A, Zagrovic B. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. Biophys J [Online]. 2010;98(5):861–71. Available from: http://dx.doi.org/10.1016/j.bpj.2009.11.011

134. Lobanov MY, Bogatyreva NS, Galzitskaya O V. Radius of gyration as an indicator of protein structure compactness. Mol Biol. 2008;42(4):623–8.

135. Brown DK, Penkler DL, Amamuddy OS, Ross C, Atilgan AR, Atilgan C, et al. MD-TASK: A software suite for analyzing molecular dynamics trajectories. Bioinformatics. 2017;33(17):2768–71.

136. Sanyanga TA, Bishop ÖT. Structural characterization of carbonic anhydrase viii and effects of missense single nucleotide variations to protein structure and function. Int J Mol Sci. 2020;21(8):1–20.

137. Chisanga D, Keerthikumar S, Mathivanan S, Chilamkurti N. Network tools for the analysis of proteomic data. Methods Mol Biol. 2017;1549:177–97.

138. Olivier B, Amamuddy S. Residue-Interaction-Networks. 2022. p. 1–14.

139. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, et al. Using graph theory to analyze biological networks. BioData Min. 2011;4(1):1–27.

140. Barozi V, Musyoka TM, Sheik Amamuddy O, Tastan Bishop Ö. Deciphering Isoniazid Drug Resistance Mechanisms on Dimeric Mycobacterium tuberculosis KatG via Post-molecular

Dynamics Analyses Including Combined Dynamic Residue Network Metrics. ACS Omega. 2022;7(15):13313–32.

141. Sheik Amamuddy O, Glenister M, Tshabalala T, Tastan Bishop Ö. MDM-TASK-web: MD-TASK and MODE-TASK web server for analyzing protein dynamics. Comput Struct Biotechnol J [Online]. 2021;19:5059–71. Available from: https://doi.org/10.1016/j.csbj.2021.08.043

142. Sheik Amamuddy O, Afriyie Boateng R, Barozi V, Wavinya Nyamai D, Tastan Bishop Ö. Novel dynamic residue network analysis approaches to study allosteric modulation: SARS-CoV-2 Mpro and its evolutionary mutations as a case study. Comput Struct Biotechnol J [Online]. 2021;19:6431–55. Available from: https://doi.org/10.1016/j.csbj.2021.11.016

143. Choi EJ, Mayo SL. Generation and analysis of proline mutants in protein G. Protein Eng Des Sel. 2006;19(6):285–9.

144. Arnórsdóttir J, Sigtryggsdóttir ÁR, Thorbjarnardóttir SH, Kristjánsson MM. Effect of proline substitutions on stability and kinetic properties of a cold adapted subtilase. J Biochem. 2009;145(3):325–9.

145. Watanabe K, Hata Y, Kizaki H, Katsube Y, Suzuki Y. The refined crystal structure of Bacillus cereus oligo-1,6-glucosidase at 2.0 Å resolution: Structural characterization of proline-substitution sites for protein thermostabilization. J Mol Biol. 1997;269(1):142–53.

146. Bogin O, Peretz M, Hacham Y, Korkhin Y, Frolow F, Kalb AJ, et al. Enhanced thermal stability of Clostridium beijerinckii alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic Thermoanaerobacter brockii alcohol dehydrogenase. Protein Sci. 1998;7(5):1156–63.

147. Tastan Bishop Ö, Musyoka TM, Barozi V. Allostery and Missense Mutations as Intermittently Linked Promising Aspects of Modern Computational Drug Discovery. J Mol Biol [Online]. 2022;434(17):167610. Available from: https://doi.org/10.1016/j.jmb.2022.167610

Supplementary material

Table S1: VAPOR parsed out results

#SNP	Provean	PPH2	I-Mutant2.0 DDG	MUpro Sign	PhD-SNP
G3D	-0.113	benign	-1.06	Decrease	Neutral
T4S	-0.282	benign	-0.21	Decrease	Neutral
I10M	-0.761	benign	-1.45	Decrease	Neutral
I10S	-1.494	benign	-2.16	Decrease	Disease
Y13C	-2.883	benign	-0.55	Decrease	Disease
S14L	-2.227	benign	0.39	Decrease	Neutral
D15A	-3.790	benign	-0.32	Decrease	Neutral
D15Y	-4.565	benign	0.20	Decrease	Neutral
T16S	-1.190	benign	-0.41	Decrease	Neutral
T16I	-1.426	benign	-0.00	Increase	Neutral
E17Q	-1.149	possibly damaging	-0.49	Decrease	Neutral
E17A	-1.599	benign	-0.48	Decrease	Neutral
E17K	-1.302	benign	-0.67	Decrease	Disease
V18G	-2.885	possibly damaging	-2.09	Decrease	Neutral
K20E	-1.833	benign	-0.65	Decrease	Neutral
N21K	-3.461	possibly damaging	0.03	Decrease	Neutral

L24P	-4.660	probably damaging	-1.26	Decrease	Disease
T25N	-3.006	possibly damaging	-0.22	Decrease	Neutral
K32R	-1.169	benign	0.15	Decrease	Neutral
V34M	-0.840	probably damaging	-0.31	Decrease	Neutral
G36D	-1.706	benign	-1.00	Decrease	Neutral
G36S	-1.213	benign	-1.11	Decrease	Neutral
F40L	-5.245	probably damaging	-0.98	Decrease	Neutral
F40S	-6.911	probably damaging	-1.50	Decrease	Disease
H41R	-7.030	probably damaging	0.44	Decrease	Neutral
E43D	-1.805	benign	0.23	Decrease	Neutral
Q44R	-1.660	benign	0.52	Decrease	Neutral
H46R	-6.811	probably damaging	0.42	Decrease	Neutral
Q47H	-2.454	possibly damaging	-0.32	Decrease	Neutral
L49S	-5.098	probably damaging	-1.50	Decrease	Neutral
K51N	-3.110	possibly damaging	0.58	Decrease	Neutral

H52R	-5.233	possibly damaging	0.41	Decrease	Neutral
H52D	-5.953	possibly damaging	0.37	Decrease	Neutral
L53S	-4.680	probably damaging	-1.33	Decrease	Neutral
T55I	-2.417	benign	-0.21	Decrease	Neutral
T55P	-3.234	benign	-0.65	Decrease	Neutral
F56L	-1.764	benign	-1.42	Decrease	Neutral
G59D	-5.660	benign	-0.83	Decrease	Disease
G62R	-4.710	probably damaging	-0.31	Decrease	Disease
G62E	-4.372	possibly damaging	-0.33	Decrease	Disease
L63P	-5.202	probably damaging	-1.45	Decrease	Disease
L63M	-1.270	probably damaging	-0.85	Decrease	Neutral
R64S	-5.238	probably damaging	-0.75	Decrease	Disease
V65L	-1.484	benign	-1.62	Decrease	Neutral
F66I	-5.435	probably damaging	-0.81	Decrease	Disease
P68L	-9.201	probably damaging	-0.33	Decrease	Disease

L69F	-3.687	probably damaging	-0.70	Decrease	Neutral
L69V	-2.747	probably damaging	-1.44	Decrease	Neutral
G71R	-7.244	probably damaging	-0.44	Decrease	Disease
A73S	-2.604	possibly damaging	-0.62	Decrease	Neutral
A73V	-3.547	probably damaging	-0.17	Increase	Disease
E75K	-3.492	probably damaging	-0.53	Decrease	Disease
M76I	-3.700	probably damaging	-0.77	Decrease	Neutral
M76K	-5.534	probably damaging	-1.52	Decrease	Disease
K77E	-3.598	probably damaging	-1.05	Decrease	Neutral
W78C	-11.831	probably damaging	-1.75	Decrease	Disease
W78R	-12.599	probably damaging	-1.14	Decrease	Disease
A80V	-3.153	probably damaging	-0.14	Decrease	Disease
A80P	-4.420	probably damaging	-0.37	Decrease	Disease

R82G	-3.428	possibly damaging	-1.02	Decrease	Disease
R82W	-2.391	probably damaging	-0.10	Decrease	Disease
R82Q	-0.840	benign	-0.24	Decrease	Disease
G83V	-8.676	probably damaging	-0.86	Decrease	Disease
H84L	-10.250	probably damaging	1.20	Increase	Disease
H84D	-8.360	probably damaging	0.22	Decrease	Disease
S85T	-1.193	benign	-0.20	Decrease	Neutral
G88C	-8.668	probably damaging	-1.50	Decrease	Disease
G88S	-5.779	probably damaging	-1.46	Decrease	Disease
I91M	-2.603	probably damaging	-1.19	Decrease	Neutral
I96M	-2.156	probably damaging	-1.04	Decrease	Neutral
Q97R	1.565	benign	0.37	Decrease	Neutral
E98D	-2.316	benign	-0.09	Decrease	Neutral
L105R	-5.561	probably damaging	-1.02	Decrease	Disease
S106F	-4.854	probably damaging	0.82	Decrease	Disease

E109K	-3.496	probably damaging	-0.18	Decrease	Neutral
E109V	-6.458	probably damaging	0.57	Decrease	Neutral
E109D	-2.591	probably damaging	-0.29	Decrease	Neutral
P111R	-5.465	probably damaging	-0.68	Decrease	Neutral
I112N	-6.090	probably damaging	-1.75	Decrease	Disease
E114K	-2.595	probably damaging	-0.49	Decrease	Neutral
I115T	-4.495	probably damaging	-1.97	Decrease	Neutral
P116A	-4.670	benign	-1.54	Decrease	Neutral
G117R	-6.218	probably damaging	-0.49	Decrease	Neutral
K119T	-5.503	probably damaging	-0.68	Decrease	Neutral
S123N	-2.572	benign	-0.85	Decrease	Neutral
S125L	-4.741	possibly damaging	-0.42	Decrease	Neutral
G126A	-4.584	possibly damaging	-1.45	Decrease	Neutral
G126R	-5.893	possibly damaging	-0.74	Decrease	Disease
N127S	-3.839	benign	-0.46	Decrease	Neutral

I128V	-0.955	possibly damaging	-1.07	Decrease	Neutral
S129P	-3.976	probably damaging	-0.76	Decrease	Disease
S134G	-2.550	possibly damaging	-1.64	Decrease	Neutral
D137Y	-7.301	possibly damaging	-0.28	Decrease	Disease
D137H	-5.438	probably damaging	-0.81	Decrease	Disease
L138F	-1.435	benign	-1.15	Decrease	Neutral
R140K	-0.857	possibly damaging	-0.94	Decrease	Neutral
T141I	-2.748	benign	-0.74	Decrease	Disease
I143T	-1.679	benign	None	Decrease	Neutral
K145Q	-2.224	benign	-1.01	Decrease	Neutral
I149T	-3.826	probably damaging	-2.41	Decrease	Disease
W150G	-11.070	probably damaging	-2.34	Decrease	Disease
D151Y	-7.202	probably damaging	-0.43	Decrease	Disease
R152T	-5.200	probably damaging	-0.51	Decrease	Disease
A154T	-2.135	probably damaging	-0.81	Decrease	Disease

L155S	-4.737	probably damaging	-2.74	Decrease	Disease
I158M	-2.225	possibly damaging	-1.54	Decrease	Neutral
I158V	-0.281	benign	-0.93	Decrease	Neutral
P160A	-6.613	probably damaging	-1.63	Decrease	Neutral
R163H	-3.150	probably damaging	-1.35	Decrease	Disease
R163C	-5.355	probably damaging	-0.66	Decrease	Disease
R163S	-4.214	probably damaging	-1.29	Decrease	Disease
Y166C	-7.722	probably damaging	-1.09	Decrease	Disease
A167G	-3.210	possibly damaging	-1.56	Decrease	Neutral
D168V	-3.274	benign	0.07	Decrease	Disease
T169A	-0.438	benign	-1.08	Decrease	Neutral
M170T	-3.799	possibly damaging	-0.65	Decrease	Neutral
S172P	-3.673	benign	-0.11	Decrease	Disease
G175E	1.463	benign	-0.46	Decrease	Disease
F178I	-1.614	probably damaging	-0.56	Decrease	Disease

F178L	-1.747	possibly damaging	-1.59	Increase	Neutral
Q179R	0.344	benign	0.22	Decrease	Neutral
Q179H	0.344	benign	-0.77	Decrease	Neutral
Q179E	-0.877	possibly damaging	0.15	Decrease	Neutral
Y180C	-6.200	benign	-0.91	Decrease	Disease
Y180F	-2.992	possibly damaging	-0.05	Decrease	Neutral
L181F	-3.136	possibly damaging	-0.80	Decrease	Neutral
L185R	-4.094	possibly damaging	-1.44	Decrease	Disease
Y187C	-7.407	probably damaging	-0.97	Decrease	Disease
D188N	-2.667	possibly damaging	-0.47	Decrease	Neutral
D188A	-5.514	probably damaging	-0.32	Decrease	Neutral
K191E	-2.601	possibly damaging	-0.56	Decrease	Neutral
H192L	-6.794	possibly damaging	0.70	Increase	Disease
G194S	-4.919	probably damaging	-1.16	Decrease	Neutral
P200L	-5.964	probably damaging	-0.04	Decrease	Disease

H201R	-1.075	benign	0.36	Decrease	Neutral
I204T	-3.066	possibly damaging	-2.20	Decrease	Neutral
L207F	-3.326	probably damaging	-0.83	Decrease	Neutral
L207W	-4.911	probably damaging	-1.50	Decrease	Neutral
F208L	-4.722	benign	-1.33	Decrease	Disease
C212R	-9.077	probably damaging	-0.15	Decrease	Disease
N213D	-0.742	benign	-0.12	Decrease	Disease
N213S	-1.881	benign	-0.14	Decrease	Neutral
R215C	-3.656	probably damaging	-0.99	Decrease	Disease
R215H	0.120	benign	-1.58	Decrease	Disease
C216Y	-4.701	possibly damaging	-0.26	Decrease	Disease
C216R	-5.987	benign	-0.16	Decrease	Disease
K219R	-1.676	benign	-0.38	Decrease	Neutral
V220F	-3.283	probably damaging	-1.05	Decrease	Disease
V220A	-2.322	benign	-1.15	Decrease	Neutral
A222D	-2.053	benign	-0.88	Decrease	Disease
E225K	-2.789	benign	-0.70	Decrease	Disease

R226Q	-2.298	possibly damaging	-0.64	Decrease	Neutral
H227Y	-3.244	possibly damaging	0.62	Decrease	Disease
H227Q	-3.950	benign	-0.19	Decrease	Neutral
G231A	-4.742	probably damaging	-0.82	Decrease	Disease
G231V	-7.007	probably damaging	-0.73	Decrease	Disease
I232V	-0.739	benign	-0.51	Decrease	Neutral
D233N	-3.506	probably damaging	-0.38	Decrease	Neutral
D233H	-5.049	probably damaging	-0.41	Decrease	Neutral
D233E	-2.716	benign	-0.05	Decrease	Neutral
C234S	-0.911	benign	-0.52	Decrease	Disease
C234Y	3.979	benign	0.07	Decrease	Disease
L235P	-4.262	probably damaging	-1.52	Decrease	Disease
K238E	-1.644	possibly damaging	-0.53	Decrease	Neutral
Y240C	-6.211	probably damaging	-0.67	Decrease	Disease
Y240H	-2.110	possibly damaging	-0.96	Decrease	Neutral

Y240S	-6.044	probably damaging	-0.95	Decrease	Disease
E244G	-2.809	possibly damaging	-0.45	Decrease	Neutral

Table S2: PredictSNP confidence level

neutral | deleterious | XX % expected accuracy | Expand all annotations

Mutation	PredictSNP	MAPP	PhD-SNP	PolyPhen-1	PolyPhen-2	SIFT	SNAP
L49S	87 %	86 %	73 %	74 %	81 %	79 %	72 %
L69V	87 %	84 %	86 %	74 %	68 %	79 %	72 %
A80P	87 %	77 %	88 %	74 %	81 %	46 %	72 %
R163C	87 %	86 %	88 %	74 %	81 %	79 %	89 %
R163H	76 %	72 %	88 %	59 %	81 %	53 %	85 %
R163P	87 %	88 %	88 %	74 %	81 %	79 %	89 %
H227Q	83 %	76 %	72 %	67 %	75 %	79 %	61 %

Listing S1: FASTA sequences for the WT human (P51580), murine (O55060) and variant enzymes used in the study.

>TPMT_HUMAN

MDGTRTSLDIEEYSDTEVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKG
KSGLRVFFPLCGKAVEMKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFK
SSSGNISLYCCSIFDLPRTNIGKFDMIWDRGALVAINPGDRKCYADTMFSLLGKKFQYLLCVL
SYDPTKHPGPPFYVPHAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>TPMT_MOUSE

MSLDMKEHPDAEVQKNQVLTLEDWKEKWVTRHISFHQEQGHQLLKKHLDTFLKGQSGLR
VFFPLCGKAIEMKWFADRGHTVVGVEISEIGIREFFAEQNLSYTEEPLAEIAGAKVFKSSSGS
ISLYCCSIFDLPRANIGKFDRIWDRGALVAINPGDHDRYADIILSLLRKEFQYLVAVLSYDPTK
HAGPPFYVPSAELKRLFGTKCSMQCLEEVDALEERHKAWGLDYLFEKLYLLTEK

>WT_2BZG

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDRKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>L49S

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLSKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDRKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>L69V

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPVCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDRKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>A80P

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFPDRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLPR
TNIGKFDMIWDRGALVAINPGDRKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVPH
AEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>R163H

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDHKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>R163P

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDPKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>R163C

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDCKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERHKSWGIDCLFEKLYLLTEK

>H227Q

EVQKNQVLTLEEWQDKWVNGKTAFHQEQGHQLLKKHLDTFLKGKSGLRVFFPLCGKAVE
MKWFADRGHSVVGVEISELGIQEFFTEQNLSYSEEPITEIPGTKVFKSSSGNISLYCCSIFDLP
RTNIGKFDMIWDRGALVAINPGDRKCYADTMFSLLGKKFQYLLCVLSYDPTKHPGPPFYVP
HAEIERLFGKICNIRCLEKVDAFEERQKSWGIDCLFEKLYLLTEK

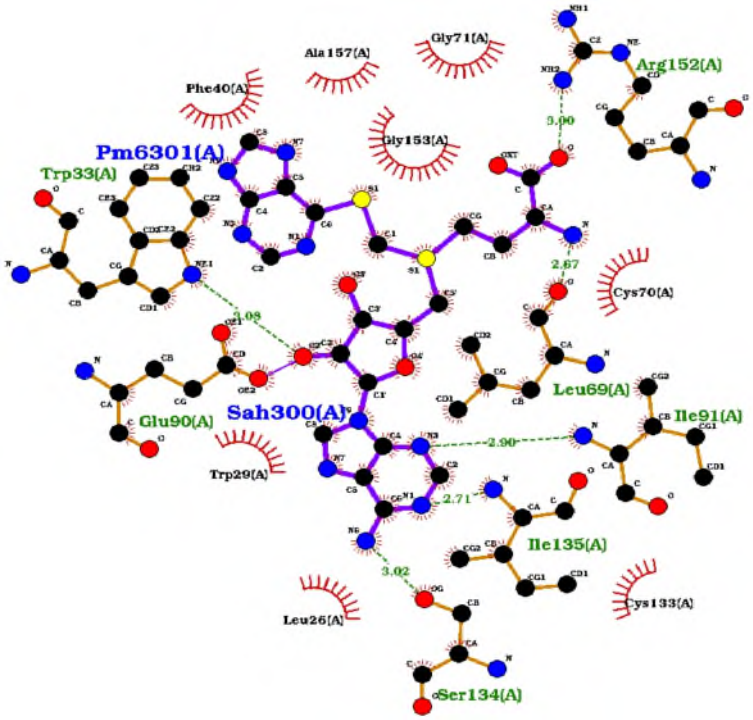


Figure S1: Ligand interaction diagram for SMP. SAH is labelled *Sah300(A)* while 6MP is labelled *Pm6301(A)*. The ligands and protein side chains are shown in ball-and-stick representation, with the ligand bonds coloured in purple. Hydrogen bonds are shown as green dotted lines, while the spoked arcs represent protein residues making non-bonded contacts with the ligands.

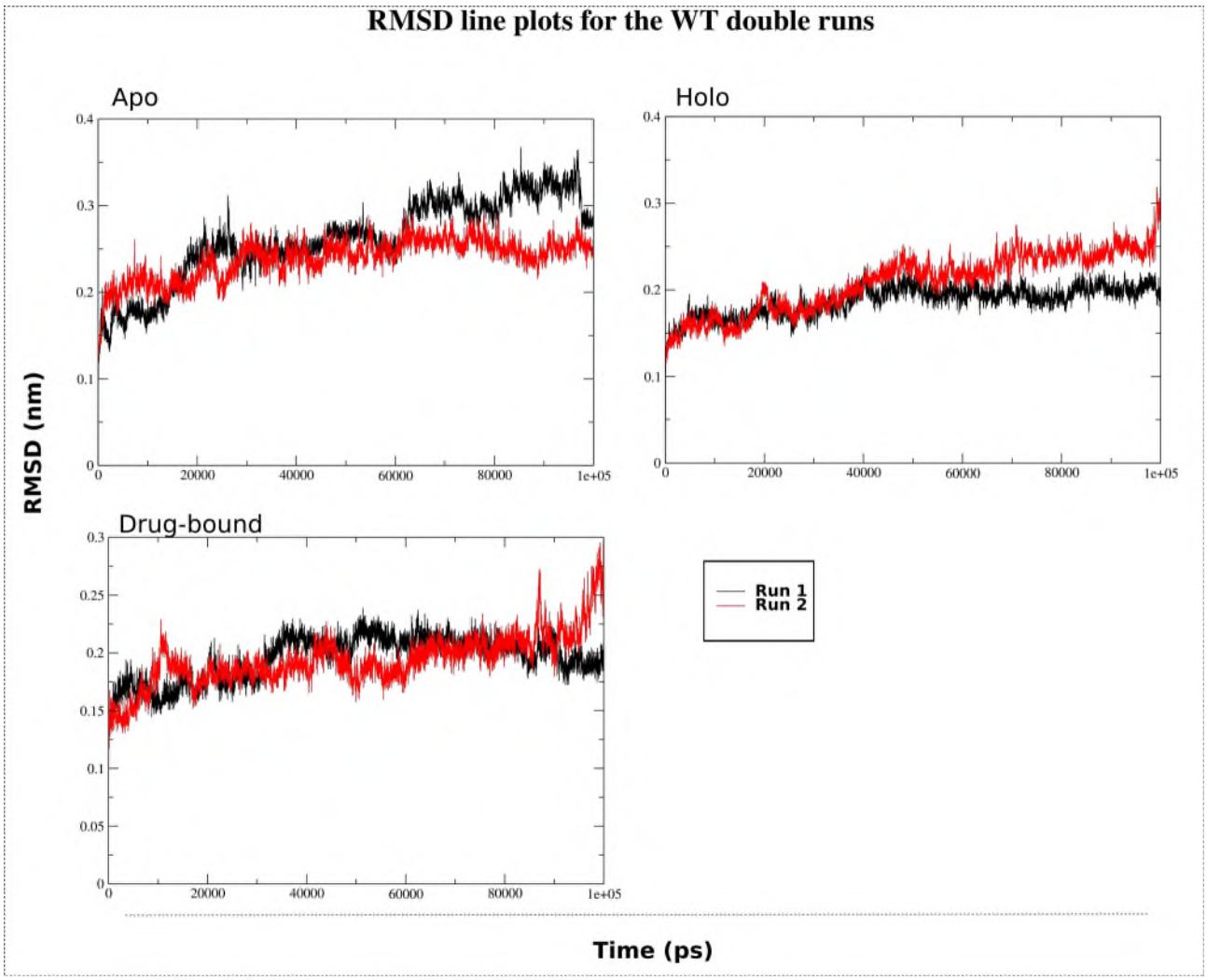


Figure S2: RMSD line plots for the WT double runs in all states.

Apo

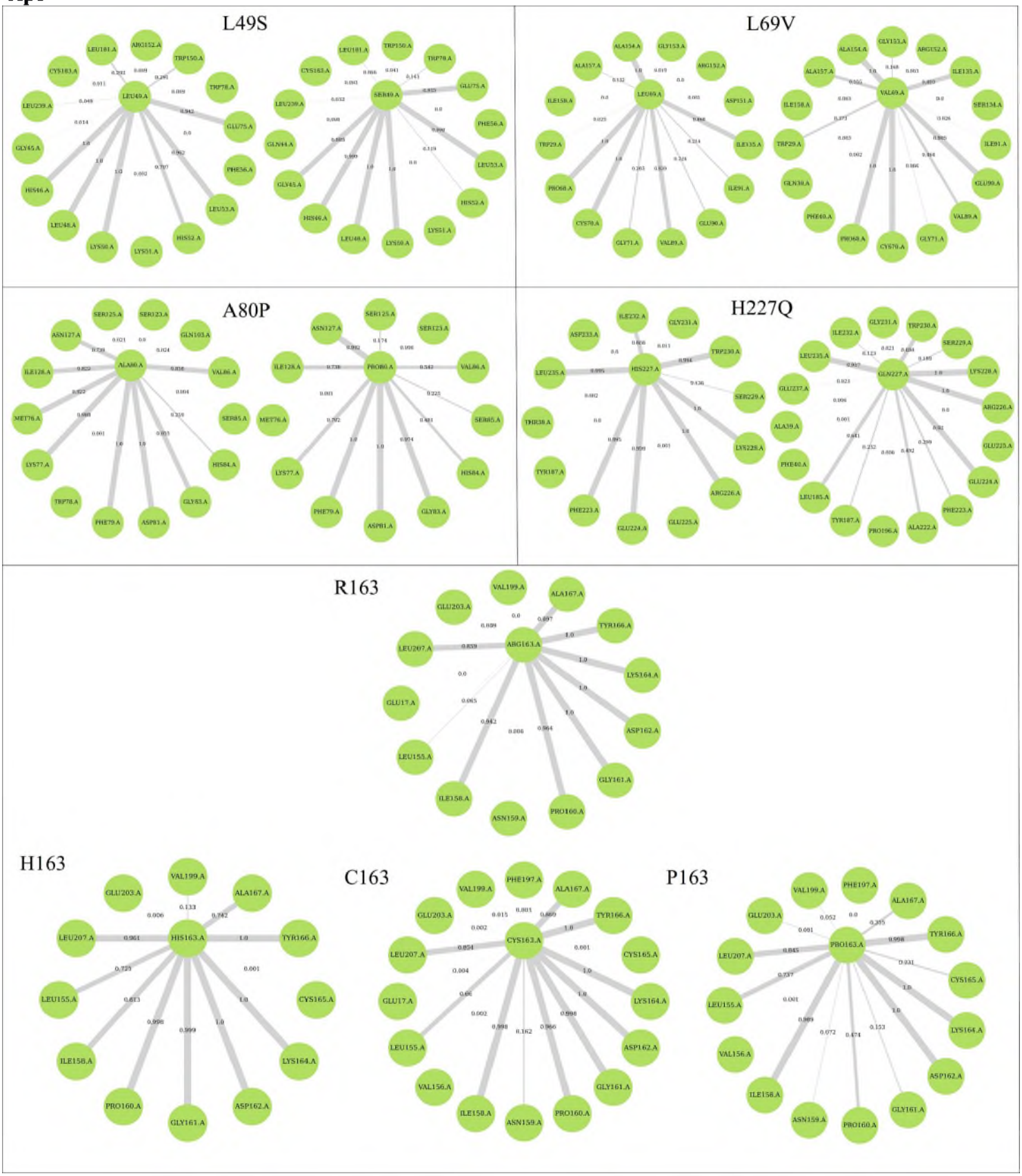


Figure S3: Weighted contact maps for mutant positions in the apo WT and mutant systems.

Holo

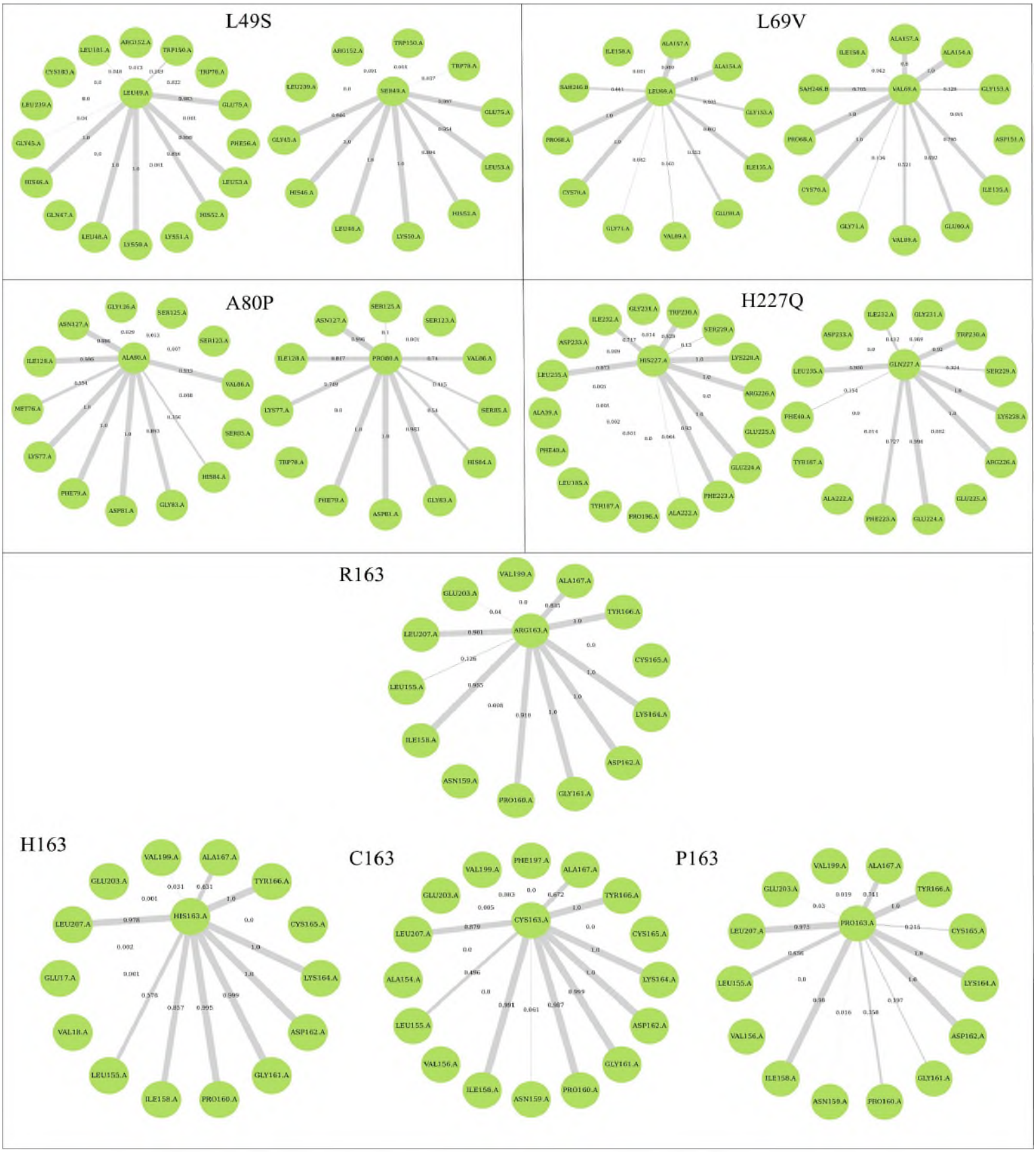


Figure S4: Weighted contact maps for mutant positions in the holo WT and mutant systems.

Drug-bound

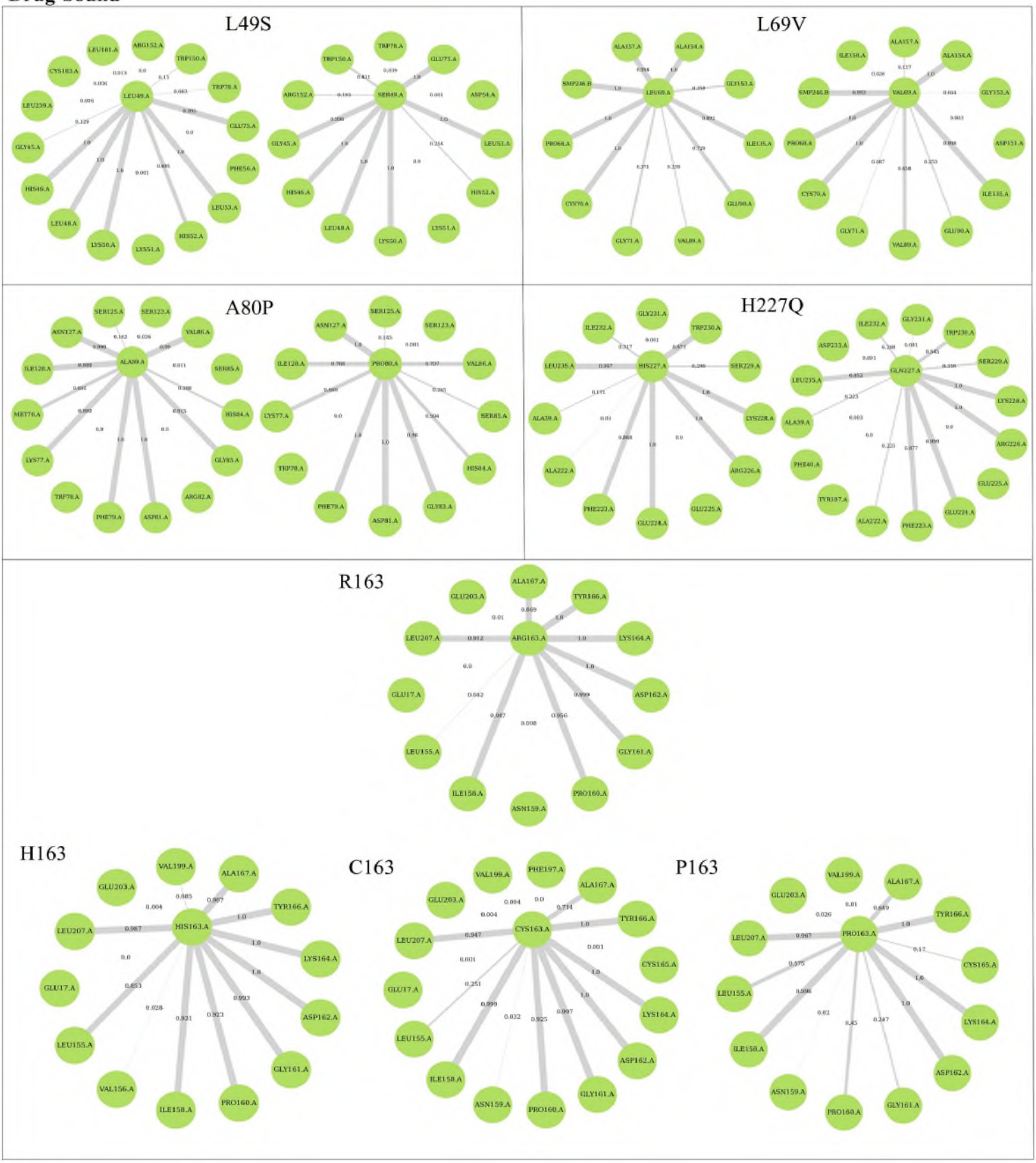


Figure S5: Weighted contact maps for mutant positions in the drug-bound WT and mutant systems.

SMP weighted contact maps

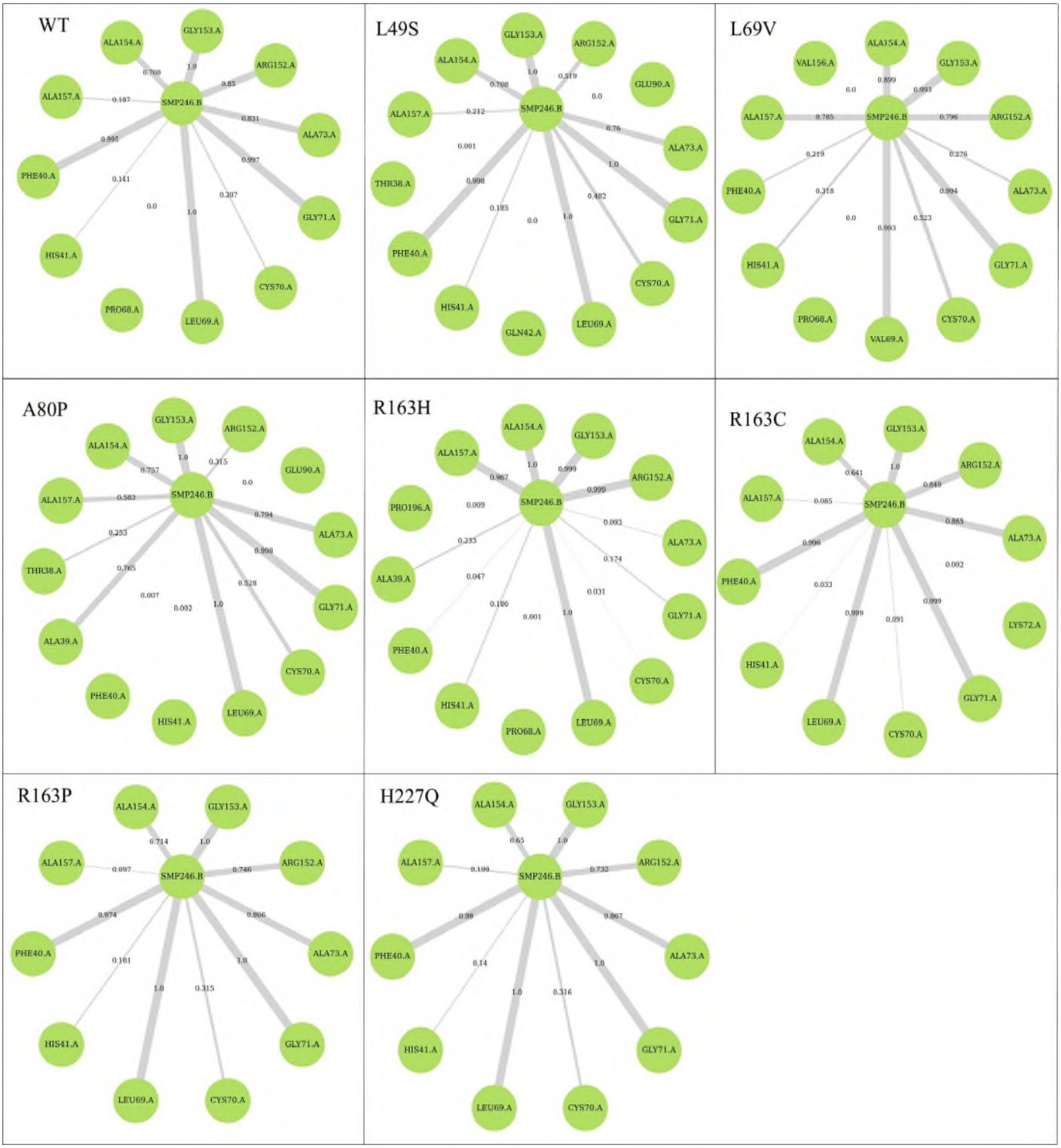


Figure S6: SMP weighted contact maps for the WT and mutant systems.

Table S3: BC and EC centrality hubs. Hubs found in at least 12 systems in either centralities are underlined and coloured blue while persistent hubs are underlined and coloured maroon.

Protein	State	Top 5% BC hubs	Top 5% EC hubs
A80P	Apo	68. 69. 70. 75. 76. 90. 151. 155. 157. 183. 185. 199	67. 70. 88. 89. 90. 95. 96. 130. 131. 132
	Holo	68. 69. 70. 75. 90. 151. 155. 157. 183. 185. 199	67. 70. 71. 88. 89. 90. 92. 95. 96. 130. 131. 132
	Drug-bound	66. 68. 69. 70. 73. 75. 90. 135. 151. 155. 157. 185	67. 69. 70. 71. 73. 88. 89. 90. 92. 95. 96. 130. 131. 132
H227Q	Apo	68. 69. 70. 75. 76. 90. 151. 155. 157. 185	67. 70. 76. 86. 88. 89. 90. 96
	Holo	48. 68. 69. 70. 76. 89. 90. 151. 153. 157. 183. 185	65. 67. 68. 70. 76. 86. 88. 89. 151. 153. 154. 155
	Drug-bound	68. 69. 76. 89. 90. 151. 153. 157	67. 69. 70. 71. 76. 86. 88. 89. 90. 92. 95. 96. 130. 131. 132
L49S	Apo	69. 70. 73. 76. 90. 151. 152. 155. 157. 183. 185. 199	65. 67. 70. 76. 86. 88. 89. 90. 95. 96. 151
	Holo	68. 69. 70. 76. 90. 151. 157. 183	67. 70. 71. 76. 86. 88. 89. 90. 92. 95. 96. 130. 131. 132
	Drug-bound	66. 68. 69. 70. 76. 89. 90. 151. 153. 155	67. 69. 70. 71. 76. 86. 88. 89. 90. 92. 95. 96. 130
L69V	Apo	46. 48. 68. 69. 75. 76. 90. 151. 154. 155. 157. 182. 185. 199	65. 67. 68. 70. 76. 86. 88. 90. 96. 149. 151
	Holo	66. 68. 69. 70. 73. 75. 76. 89. 90. 151. 153. 154. 183	67. 70. 71. 76. 86. 88. 89. 90. 92. 95. 96. 99. 130. 131. 132
	Drug-bound	68. 73. 89. 90. 151. 153. 154. 155. 183. 185	67. 68. 70. 86. 88. 89. 90. 96. 154
R163C	Apo	68. 69. 70. 73. 75. 76. 90. 151. 155. 156. 157. 185. 199	67. 70. 71. 76. 86. 88. 89. 90. 95. 96. 130
	Holo	48. 68. 69. 70. 75. 76. 89. 90. 151. 157. 183	67. 70. 71. 76. 86. 88. 89. 90. 95. 96. 130
	Drug-bound	68. 69. 89. 90. 151. 153. 155. 157. 185. 199	67. 69. 70. 86. 88. 89. 90. 95. 96. 130

R163H	Apo	68. 69. 70. 76. 90. 151. 153. 154. 155. 156. 157. 183. 199	67. 70. 71. 73. 76. 80. 86. 88. 89. 90. 95. 96. 99. 128. 130
	Holo	69. 70. 75. 76. 90. 151. 153. 157	67. 69. 70. 88. 89. 90. 95. 96. 130. 151. 153. 154
	Drug-bound	66. 68. 69. 70. 75. 76. 90. 151. 153. 157. 185. 196	67. 68. 70. 86. 151. 153. 154. 155
R163P	Apo	68. 69. 70. 75. 76. 89. 90. 151. 153. 155. 157. 199	65. 67. 70. 76. 86. 88. 89. 90. 95. 96. 130
	Holo	66. 68. 69. 70. 75. 76. 90. 151. 153. 155. 156. 157. 183. 185. 199	65. 67. 70. 76. 80. 86. 88. 89. 90. 130
	Drug-bound	68. 69. 70. 75. 76. 89. 90. 151. 153. 155. 156. 185. 199	67. 69. 70. 71. 76. 86. 88. 89. 90. 95. 96. 130
WT	Apo	24. 66. 67. 68. 69. 73. 75. 89. 90. 150. 151. 154. 155. 157. 183. 199	67. 70. 88. 89. 90. 131. 151
	Holo	68. 69. 70. 76. 90. 151. 155. 157. 182. 185	67. 70. 71. 86. 88. 89. 90. 92. 95. 96. 99. 130. 131. 132
	Drug-bound	69. 73. 76. 90. 151. 153. 155. 157. 182. 199	69. 70. 71. 88. 89. 90. 92. 95. 96. 130

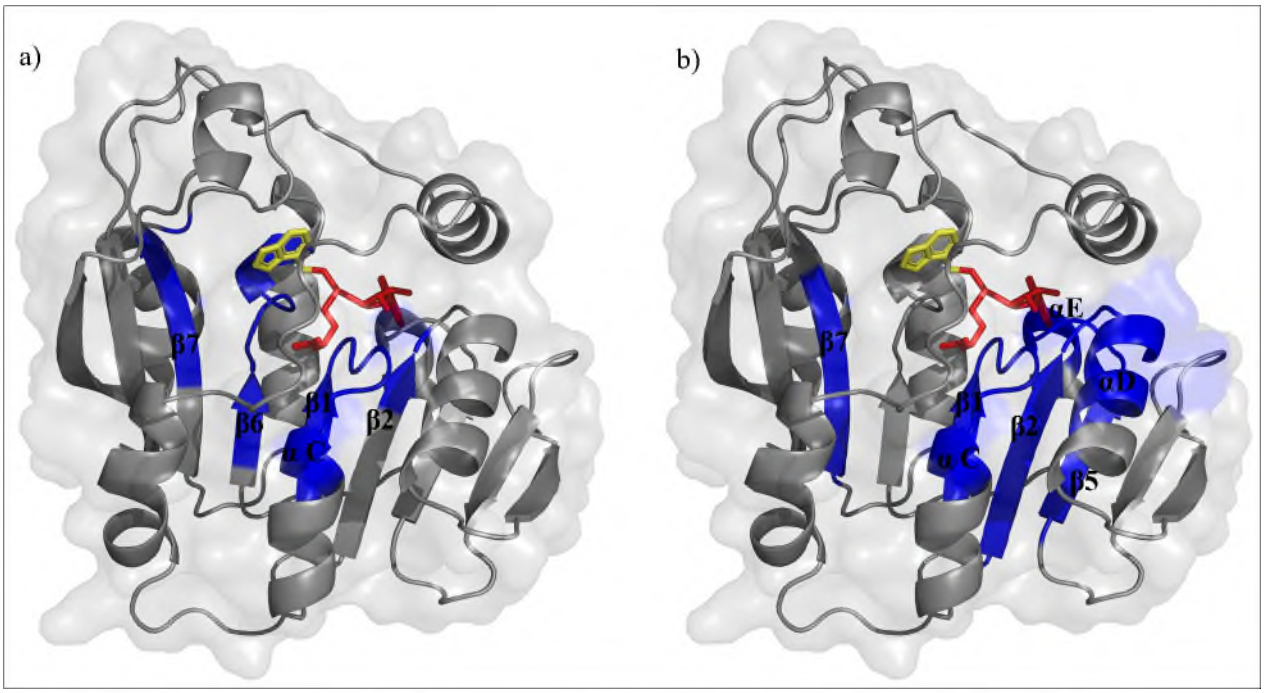


Figure S7: Cartoon representation of areas that sample high centrality measure and top 5% centrality hubs. The areas have been coloured blue and the secondary structures where they occurred have been labelled according to their names. SMP is represented as sticks with 6MP coloured yellow and SAH coloured red. **a)** represents areas sampled in BC calculations. while **b)** represents areas sampled in EC calculations.