




OPEN

A bioinformatics approach to the identification of novel deleterious mutations of human TPMT through validated screening and molecular dynamics

Sidharth Saxena¹, T. P. Krishna Murthy¹ , C. R. Chandrashekar¹, Lavan S. Patil¹, Abhinav Aditya¹, Rohit Shukla², Arvind Kumar Yadav², Tiratha Raj Singh², Mahesh Samantaray³ & Amutha Ramaswamy³

Polymorphisms of Thiopurine S-methyltransferase (TPMT) are known to be associated with leukemia, inflammatory bowel diseases, and more. The objective of the present study was to identify novel deleterious missense SNPs of TPMT through a comprehensive in silico protocol. The initial SNP screening protocol used to identify deleterious SNPs from the pool of all TPMT SNPs in the dbSNP database yielded an accuracy of 83.33% in identifying extremely dangerous variants. Five novel deleterious missense SNPs (W33G, W78R, V89E, W150G, and L182P) of TPMT were identified through the aforementioned screening protocol. These 5 SNPs were then subjected to conservation analysis, interaction analysis, oncogenic and phenotypic analysis, structural analysis, PTM analysis, and molecular dynamics simulations (MDS) analysis to further assess and analyze their deleterious nature. Oncogenic analysis revealed that all five SNPs are oncogenic. MDS analysis revealed that all SNPs are deleterious due to the alterations they cause in the binding energy of the wild-type protein. Plasticity-induced instability caused by most of the mutations as indicated by the MDS results has been hypothesized to be the reason for this alteration. While in vivo or in vitro protocols are more conclusive, they are often more challenging and expensive. Hence, future research endeavors targeted at TPMT polymorphisms and/or their consequences in relevant disease progressions or treatments, through in vitro or in vivo means can give a higher priority to these SNPs rather than considering the massive pool of all SNPs of TPMT.

Human Thiopurine S-methyltransferase (TPMT) is a protein-coding gene on chromosome 6p22.3 that codes for the enzyme Thiopurine S-methyltransferase. It is a monomer that belongs to the class I-like SAM-binding methyltransferase superfamily and is composed of 245 amino acid residues. This cytosolic enzyme catalyzes the S-methylation of thiopurines such as azathioprine, 6-thioguanine, 6-mercaptopurine, etc. Thiopurines were originally developed in the 1950s for the treatment of acute myeloid leukemia in children by Gertrude Elion and George Hitchings¹, which eventually won them the Nobel Prize in Physiology or Medicine in 1988. These thiopurine drugs are antimetabolites and immunomodulators and are used extensively as anticancer and immunosuppressive agents in the treatment of rheumatoid arthritis, inflammatory bowel diseases (IBD) including Crohn's disease and ulcerative colitis, acute lymphoblastic leukemia (ALL), and other autoimmune disorders. Although about two-third of patients do not suffer from any notable side effects throughout their treatments, the rest often cannot continue their treatments or have their treatments greatly modified due to cytotoxic side effects and an increased risk of malignancies. Most patients with thiopurine-induced hepatotoxicity have a high concentration of 6-methylmercaptopurine ribonucleotide or 6-MMPR (formed by TPMT) one week after treatment initiation. There is also evidence that elevated levels of 6-methylmercaptopurine or 6-MMP (formed by TPMT) are

¹Department of Biotechnology, Ramaiah Institute of Technology, Bengaluru, Karnataka 560054, India. ²Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Solan, Himachal Pradesh 173234, India. ³Department of Bioinformatics, Pondicherry University, Pondicherry 605014, India. ✉email: tpk@live.in

correlated with hepatotoxicity². To address these side effects, improved drug delivery mechanisms such as nano-formulations and delayed-release tablets are in development, none of which are currently in clinical practice³.

Akin to many immunosuppressants, the mechanism of action of thiopurines involves the inactivation of critical T-cell processes leading to inflammation. Unfortunately, the exact mechanism of action of thiopurines and their effects are not yet comprehensively understood. A combination of many different factors has been proposed for their cytotoxicities, such as the inhibition of de novo purine synthesis (DNPS), alterations in DNA methylation state, disruption of guanosine-triphosphate (GTP) signaling, and the incorporation of thioguanine nucleotides (TGN) as bases into DNA. The incorporation of thioguanine nucleotides into DNA causes the induction of apoptosis⁴.

Around 99.9% of the DNA sequence is identical between any two human genomes selected at random. The variation in 0.1% of the DNA sequence comprises genetic alterations known as polymorphisms. One such form of these polymorphisms involves the alteration of one nucleotide and is aptly referred to as a single nucleotide polymorphism (SNP). Nearly 90% of all human DNA polymorphisms are attributed to SNPs and hence, they are by far the most common form of genetic variation that occurs in the human genome⁵. In fact, an SNP occurs once every 300 base pairs on average in the human genome⁶. SNPs can result in a change of the encoded amino acids if they are non-synonymous or can be silent if they are synonymous or simply occur in the non-coding regions without consequence. Non-synonymous SNPs (nsSNPs) can alter the function of the protein products of genes if they occur in the coding regions, which can be especially dangerous if the gene holds biological significance⁷.

One of the major reasons for the cytotoxic side effects of thiopurine drugs is TPMT deficiency in patients and most TPMT polymorphisms often reduce TPMT activity⁸. Furthermore, it has already been established that several alleles of TPMT are in fact, disease-associated with respect to leukemia^{9–13}, IBD^{14–19}, and more. Such polymorphisms can also severely affect protein stability²⁰ and cause adverse drug reactions²¹. Therefore, SNPs of TPMT can be crucial in the disease progression and treatments associated with leukemia, IBD, and more. This warrants the identification of novel SNPs of TPMT which are deleterious in nature. While the conduction of *in vitro* and/or *in vivo* methods for the identification of novel deleterious nsSNPs from the massive pool of all nsSNPs is more conclusive than *in silico* methods, they are often challenging and expensive protocols. As a result, the present work is focused on the identification of novel deleterious missense mutations of TPMT through *in silico* means. These SNPs can later be given a higher priority by future researchers interested in identifying TPMT variants and their potential consequences through *in vitro* and/or *in vivo* means. Although to validate the initial *in silico* screening process, deleterious SNPs of TPMT identified through *in vitro* or *in vivo* means were also subjected to this screening process. The screening process was able to accurately identify a significant percentage of these pre-identified SNPs, as discussed in “Screening of nsSNPs” section.

Similar *in silico* studies have been conducted for several genes in the past to understand the molecular mechanisms of disease-causing mutations^{22–28}. In fact, an *in silico* study for TPMT was also carried out in the past²⁹, but that study only considered a few established mutations and did not aim to find novel deleterious SNPs of TPMT. It also considered far fewer tools for the *in silico* protocol than the present study. Meanwhile, the present study aimed at identifying novel deleterious missense nsSNPs by screening all missense nsSNPs of TPMT listed in the dbSNP database in an attempt to identify those SNPs from the database that are extremely deleterious to both the structure and the natural functioning of the protein. The usage of computational tools for the classification of variants is at the heart of an *in silico* based approach for the identification of novel deleterious mutations. The tools chosen for this study were based on the variety of techniques used by the various screening tools in order to arrive at an unbiased consensus-based result with respect to the classification of deleterious TPMT variants. The screened SNPs were then subjected to further *in silico* analyses such as conservation analysis, interaction analysis, oncogenic analysis, phenotypic analysis, structural analysis, and post-translational modification (PTM) analysis.

Molecular dynamics simulations (MDS) simulate the motions of every single atom in a protein or any other molecular system over time-based on the classical equations of motion governing interatomic interactions. Several important biomolecular processes, such as protein folding, ligand binding, and conformational change can be studied through MDS. Most importantly, they can simulate how a particular biomolecule will respond to mutations^{30,31}. Hence, MDS was also performed in order to assess the impact of the mutations with respect to the wild-type protein at an atomic level and from a dynamic perspective as opposed to the static protocol of other structural analysis tools. Several of the aforementioned *in silico* studies have also employed MDS to gain a greater understanding of the deleterious nature of the screened mutations. Finally, a consensus-based result was obtained which is at the heart of the *in silico* pipeline of the present study. Apart from this, two other factors make the present study unique. Firstly, the utilization of a consensus-based approach for the screening of deleterious mutations is based on a lot of different computational tools that possess different pipelines to predict the nature of SNPs. Secondly, the validation of this *in silico* screening pipeline by applying it to the SNPs associated with TPMT alleles that have already been proven to be deleterious by experimental methods in prior research efforts and assessing how many of them are redetected through this pipeline. Results showed that five novel mutations, namely W33G, W78R, V89E, W150G, and L182P can have a deleterious effect on TPMT and that W33G is the most deleterious among them.

Materials and methods. Collection of data. The NCBI dbSNP database was utilized to obtain the missense nsSNPs of the human TPMT gene (access date: July 26, 2021)³². dbSNP database contains more than 650 million human SNPs (Build 152, November 2018) and includes clinically relevant SNPs from the ClinVar database as well. The UniProt database was used to obtain the sequence for the same (UniProt ID: P51580)³³. The structure of the TPMT protein with PDB ID 2H11³⁴ was obtained through the Protein Data Bank. The overall protocol followed in the present study has been visually presented in Fig. 1.

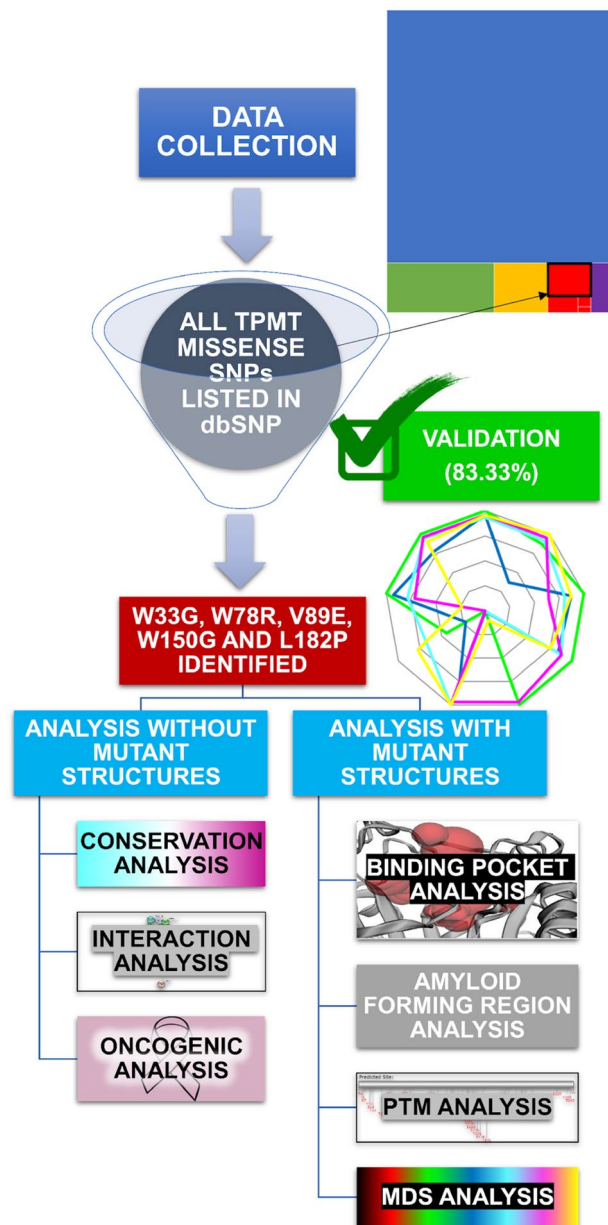


Figure 1. Visual representation of the methodology followed in the present study (The rectangle on the top right represents the pool of all TPMT SNPs. Please refer to Figure S1 for more details).

Screening of SNPs through sequence-based tools. 12 sequence-based tools were utilized to analyze the functional consequences of the missense nsSNPs of TPMT, including Protein Variation Effect Analyzer (PROVEAN)³⁵, Protein Analysis Through Evolutionary Relationships (PANTHER)³⁶, Polymorphism Phenotyping v2 (PolyPhen-2)³⁷, Mutation Assessor³⁸ and Sorting Intolerant From Tolerant (SIFT)³⁹. Apart from these tools, consensus-based tools such as Meta-SNP⁴⁰ and PredictSNP1⁴¹ were utilized in the screening process as well. Furthermore, SuSPect⁴², PMut⁴³, SNAP2⁴⁴, PhD-SNP⁴⁵ and SNPs&GO⁴⁶ were used as well during the screening process.

Screening of SNPs through structure-based tools. The stability changes imparted on the protein structure as a consequence of the nsSNPs identified to be dangerous (by at least nine out of the 12 aforementioned sequence-based tools), were assessed using structure-based tools. Eight structure-based tools were used, including CUPSAT⁴⁷, DUET⁴⁸, I-Mutant 3.0⁴⁹, MUpro⁵⁰, INPS-MD⁵¹, PoPMuSiC⁵², HoTMuSiC⁵³ and SNPmuSiC⁵⁴. The melting point of TPMT was inputted as 319.95 K for HoTMuSiC⁵⁵. Furthermore, for the screening process, an SNP was considered deleterious by the three MuSiC tools only if the SNP was predicted to “Strongly Decrease” the stability of the protein (A prediction score close to or greater than 4, close to or lesser than -10 and close to or greater than 1 for PoPMuSiC, HoTMuSiC and SNPmuSiC, respectively imply that the particular SNP

will “Strongly Decrease” the protein stability). This was done in order to identify the most deleterious missense SNPs among all of the missense SNPs.

Validation of the screening pipeline. Many SNPs of TPMT have already been proven to be deleterious by prior research efforts through experimental means. To validate and assess the *in silico* screening pipeline utilized in this study, the same screening methodology was applied for these pre-identified SNPs as well, albeit with one small difference. The difference is that the SNPs predicted to “Decrease” protein stability by the three MuSiC tools were considered deleterious, as opposed to the ones predicted to “Strongly Decrease” the protein stability. This was done because, in the validation phase, SNPs already proven to be deleterious were to be re-identified rather than finding the most deleterious SNPs among them.

Conservation analysis. The evolutionary conservation analysis of TPMT residues was conducted through the ConSurf server⁵⁶. Furthermore, multiple sequence alignment (MSA) was performed through various tools including Clustal Omega, Kalign, MUSCLE, T-Coffee and MAFFT⁵⁷ for TPMT sequences belonging to different species including *Homo sapiens* (P51580), *Gorilla gorilla gorilla* (Q3BCR3), *Pan troglodytes* (Q3BCR8), *Pongo abelii* (A0A0B4J2I0), *Rhinopithecus bieti* (A0A2K6KP23), *Chlorocebus sabaues* (A0A0D9R5L7), *Macaca mulatta* (F6TPF9), *Callithrix jacchus* (F6PNF2), *Vicugna pacos* (A0A6I9HY34) and *Panthera tigris* (Q3BCR2). Apart from MSA tools and ConSurf, NLM’s (National Library of Medicine) CDD (Conserved Domain Database)⁵⁸ was also considered. The “Type Selection” parameter was chosen as “the most diverse members” which identifies the most dissimilar TPMT sequences, as determined from the domain model MSA.

Interaction analysis. The STRING database⁵⁹ was used to conduct the interaction analysis for TPMT. Interaction analysis is important as it sheds light on the consequences of TPMT mutants on the biological processes that involve the interaction network of TPMT. This is because, in the presence of highly deleterious mutations, the structure and function of TPMT will be impaired. Since all proteins in the interaction network must function appropriately for the associated biological processes to occur in the desired manner, a dysfunctional TPMT may very well hinder or cause a lot of problems for said biological processes.

Oncogenic analysis and phenotypic analysis. To assess the oncogenicity of the five missense SNPs, CScape⁶⁰, CScape-somatic⁶¹, and Dr Cancer⁶² were utilized. The phenotypic consequences of the five mutations were predicted using the FATHMM (Functional Analysis through Hidden Markov Models) tool⁶³. The “Inherited Disease” section of the tool was utilized, where the weighted algorithm was used as the prediction algorithm. Furthermore, the “Phenotypic Associations” parameter was selected as “Human Phenotype Ontology”.

Modelling of TPMT mutants and assessing the structural effects of mutations. Modeling of the TPMT mutants (W33G, W78R, V89E, W150G, and L182P) was performed by using the PyMol software⁶⁴. The HOPE (Have Your Protein Explained) server⁶⁵ was used to visualize the mutations and also to gain some insight into the structural consequences caused by the mutations. NetSurfP-2.0⁶⁶ was used to identify alterations in the secondary structure of the protein that are caused by the five deleterious mutations. It utilized the amino acid sequences of the wild-type protein and mutant proteins to generate the results. Furthermore, the alterations caused to the binding pockets of the wild-type protein were identified using the CASTp 3.0 server⁶⁷. AmylPred 2⁶⁸ was also used for identifying amyloid-forming regions that may be associated with several conformational diseases, called amyloidoses. Furthermore, MDS analysis was performed to validate and expand upon the structural analysis.

Post-translational modification analysis. The Group-based Prediction System (GPS) 5.0 software⁶⁹ was used for PTM analysis which can predict the PTM sites within a protein. This software was used to predict the tyrosine kinase phosphorylation sites along with the serine/threonine kinase phosphorylation sites of TPMT.

Molecular dynamics simulations. The crystal structure of human TPMT (2H11) was subjected to MDS through GROMACS version 2019.4⁷⁰. The GROMOS9653a6 force field was used for the generation of the protein parameters. The gmxditconf tool was used to build the cubic simulation box. The steepest descent algorithm was used to vacuum minimize the processed setup for 1500 steps. The simple point-charge (SPC) water model was used to perform solvation by using the gmxsolvate tool. The system was electro-neutralized through the gmxcgenion tool, following which the steric clashes were removed. After this, energy minimization was performed to optimize the structure. System equilibration was performed through NVT equilibration and NPT ensemble. 1 ns of NVT equilibration was performed where the system was heated up to 300 K, to stabilize the temperature of the system. Furthermore, 1 ns of NPT ensemble was performed to stabilize the pressure and density of the system. After system equilibration, each structure was subjected to MDS over a simulation time of 100 ns.

As mentioned in the introduction section, MDS can be used to analyze the consequences of mutations at an atomic level through dynamic simulations of protein structure alterations. The MDS parameters including Root-mean-square deviation (RMSD), Root-mean-square fluctuations (RMSF), radius of gyration (Rg), solvent-accessible surface area (SASA) and intramolecular hydrogen bonding help us understand various aspects of protein structure alterations such as structural alterations in flexibility, rigidity, compactness and more. Trajectory analysis of the wild-type protein and the five mutant proteins was done using various GROMACS analysis tools. RMSD, RMSF, Rg, SASA and intramolecular hydrogen bonding, were calculated using the gmrxrms, gmrxrmsf,

gmx gyrate, gmx sasa and gmx hbond tools, respectively. Secondary structure analysis (SSA) was performed through the do_dssp tool. Principal component analysis (PCA) was performed using the gmx ana eig and gmx covar tools. The Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) method was utilised to calculate the binding energy and its components including van der Waals, electrostatic energy, polar solvation energy and SASA energy using the g_mmpbsa tool⁷¹. The relevance and interpretation of all of the aforementioned MDS parameters have been discussed in “[Molecular dynamics simulations analysis](#)” section.

Results and discussion

SNP distribution. In all, 10,305 SNPs of TPMT were obtained from dbSNP. SNPs in the intronic region made up 83.55% (8610) of all SNPs. 7.64% (787) and 3.86% (398) of the SNPs were present in the 3'UTR and 5'UTR regions, respectively. Furthermore, all forms of coding SNPs accounted for merely 3.06% of all SNPs. Among coding SNPs missense SNPs, synonymous SNPs, frameshift SNPs and nonsense SNPs accounted for 2.07% (213), 0.69% (71), 0.19% (20) and 0.11% (11), respectively. Other SNPs accounted for the remaining 1.89% (195). The present study has considered only missense SNPs of TPMT. The distribution of TPMT SNPs is shown in Figure S1.

Screening of nsSNPs. All missense SNPs of TPMT present in the dbSNP database were screened using sequence-based and structure-based tools to identify the most deleterious ones among them. Prediction results of all missense SNPs are provided in Table S1 and S2.

For sequence-based screening, the aforementioned 12 sequence-based tools were used, and SNPs found to be deleterious by at least nine out of the 12 tools were considered for further analysis. 53 such SNPs were found, which are reported in Table 1. These 53 missense SNPs were then subjected to structure-based screening using eight structure-based tools. Missense SNPs identified as deleterious by at least six of these tools were considered for further analysis. Six such SNPs were identified including W33G, W78R, V89E, W150G, L182P, and Y240S. However, owing to the fact that Y240S has already been proven to be deleterious^{9,19,20}, it was not considered for further analyses. Hence, five novel extremely deleterious missense SNPs of human TPMT were identified at the end of the screening process which was then subjected to further analyses. The results of the structure-based screening are presented in Table 2. The rationale behind considering the deleterious prediction of any 9 sequence-based tools and any 8 structure-based tools, many of which have varying approaches for variant classification is to get an unbiased result. If an SNP is predicted to be deleterious by 9 different sequence-based tools and 8 different structure-based tools, several of which have varying approaches, then the probability of it being dangerous is high.

Validation of the screening pipeline. It is a well-known fact that in silico predictions and analyses are not 100% accurate and are not as conclusive as in vitro or in vivo experiments. However, that does not mean that in silico predictions should be completely disregarded. In order to validate the in silico methodology utilized in this study, other human TPMT missense SNPs already identified as deleterious through experimental means were subjected to the same screening pipeline, with a minor difference as highlighted in “[Validation of the screening pipeline](#)” section. The results of the screening process on missense SNPs of TPMT, pre-identified as deleterious experimentally in other studies^{9,11,15,20,72–92} are summarised in Table 3 and are presented more comprehensively in Table S3.

13 out of the 17 (76.47%) very dangerous mutations, excluding M1V, were re-detected using the screening pipeline. M1V was not considered as that residue is not present in the structure of TPMT (2H11) and hence, structure-based screening could not be performed for M1V. Furthermore, with respect to extremely deleterious mutations which were proven to be very dangerous by one or more studies and/or dangerous by multiple studies, 10 out of 12 (83.33%) of them were re-detected. Figure 2 visually represents the accuracy of the screening pipeline.

ConSurf, which was used for evolutionary conservation analysis, predicts colour scores for residues that can range between 1 and 9 and are a measure of the conservation level of that particular residue. A residue can be considered as extremely variable if it has a score of “1” and conversely, it can be considered as extremely conserved if it has a score of “9”. It was observed that while high conservation of the residues which are predicted to harbor SNPs provides further validation of their deleterious nature, it is not completely conclusive. This is due to the fact that in the case of five SNPs (C132Y, G144R, C212R, Y240C and Y240S) pre-identified as extremely deleterious and also re-detected by the screening pipeline, the color scores predicted by ConSurf were lower than or equal to 5. However, nine SNPs pre-identified as extremely dangerous (L49S, L69V, G71R, A80P, Y107D, A154T, R163C, R163P and Y166C) and also re-detected using the screening pipeline did have color scores predicted as greater than or equal to 6. Hence, while conservation analysis is good as a secondary form of validation of the deleterious nature of screened SNPs, it is not guaranteed that if an SNP is at a relatively less conserved residue it will be benign in nature. In fact, in a similar study on the GOT1 gene, an SNP (L345P) was predicted to be highly deleterious despite having a color score (predicted by ConSurf) of 1 and its deleterious nature was reflected in further in silico analyses of that study⁹³.

Conservation analysis. The level of conservation of the amino acid residues provides a rough estimate regarding the level of damage that dangerous mutations can cause to the functioning of the protein and its structure. As mentioned before, they are not completely conclusive but do provide a good form of secondary validation of the deleterious nature of the SNP in question. The Bayesian method was used to obtain the ConSurf results, and the protein structure (2H11) was given as the input. 10 sequences of TPMT including human TPMT (P51580) were considered for MSA, and scores ranging from 1 to 10 were assigned based on how many

| Variant ID | SNP | PROVEAN | PANTHER | Polyphen-2 | Mutation Assessor | Meta-SNP | PMut | PredictSNP1 | SNAP2 | SuSpect | SNPs&GO | PhD-SNP | SIFT |
|--------------|-------|-------------|-------------------|-------------------|-------------------|----------|---------|-------------|---------|-------------|---------|---------|-------------|
| rs72552741 | W33G | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs200695400 | F40S | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1446592306 | F40L | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs757081801 | H41R | Deleterious | Probably damaging | Possibly damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs762702707 | H46R | Deleterious | Probably damaging | Probably damaging | Medium | Neutral | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs72552740 | L49S | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs757971326 | H52D | Deleterious | Possibly damaging | Possibly damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1236222449 | F66I | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Neutral | Neutral | Disease | Neutral |
| rs1368439131 | P68L | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1784276391 | P68S | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs200591577 | L69V | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs200591577 | L69F | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs777686348 | G71R | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs778578091 | E75K | Deleterious | Possibly damaging | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1200214781 | M76K | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1256618794 | W78C | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs753277177 | W78R | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1800462 | A80P | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs111901354 | R82G | Deleterious | Probably benign | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Neutral | Disease | Disease | Deleterious |
| rs111901354 | R82W | Deleterious | Probably benign | Probably damaging | Medium | Disease | Neutral | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1293957844 | G83V | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1235431245 | H84D | Deleterious | Probably damaging | Probably damaging | High | Disease | Neutral | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1582044292 | H84L | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Neutral | Deleterious | Effect | Deleterious | Disease | Disease | Neutral |
| rs753545734 | G88S | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs753545734 | G88C | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1784191846 | V89E | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1681788109 | E90V | Deleterious | Possibly damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1474060016 | L105R | Deleterious | Possibly damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Neutral | Disease | Disease | Deleterious |
| rs886061266 | I112N | Deleterious | Probably benign | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Neutral | Disease | Disease | Deleterious |
| rs1396619437 | S129P | Deleterious | Probably benign | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Neutral | Disease | Disease | Deleterious |
| rs72552738 | C132Y | Deleterious | Probably damaging | Possibly damaging | Medium | Disease | Disease | Deleterious | Neutral | Deleterious | Disease | Disease | Deleterious |
| rs72552737 | G144R | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Neutral | Deleterious | Effect | Neutral | Disease | Disease | Deleterious |
| rs1310627040 | I149T | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1447033392 | W150G | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |

Continued

| Variant ID | SNP | PROVEAN | PANTHER | Polyphen-2 | Mutation Assessor | Meta-SNP | PMut | PredictSNP1 | SNAP2 | SuSpect | SNPs&GO | PhD-SNP | SIFT |
|--------------|-------|-------------|-------------------|-------------------|-------------------|----------|---------|-------------|---------|-------------|---------|---------|-------------|
| rs1354851110 | D151Y | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1408113946 | R152T | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1800460 | A154T | Deleterious | Probably benign | Possibly damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1158437171 | L155S | Deleterious | Possibly damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs112339338 | R163S | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs112339338 | R163C | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs201695576 | Y166C | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1386533390 | L182P | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1783991755 | V184D | Deleterious | Probably benign | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Neutral | Disease | Disease | Deleterious |
| rs747307984 | L185R | Deleterious | Probably benign | Possibly damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1554137341 | Y187C | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs758437011 | G194S | Deleterious | Probably damaging | Probably damaging | High | Neutral | Disease | Deleterious | Effect | Deleterious | Neutral | Disease | Deleterious |
| rs79901429 | I204T | Deleterious | Probably benign | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Neutral | Disease | Deleterious |
| rs761626260 | L207W | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Neutral | Deleterious | Effect | Deleterious | Disease | Disease | Neutral |
| rs377085266 | C212R | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Neutral |
| rs780065109 | G231V | Deleterious | Probably damaging | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs781105138 | L235P | Deleterious | Probably benign | Probably damaging | Medium | Disease | Disease | Deleterious | Effect | Deleterious | Disease | Disease | Deleterious |
| rs1142345 | Y240C | Deleterious | Probably damaging | Probably damaging | High | Disease | Disease | Deleterious | Effect | Deleterious | Neutral | Disease | Deleterious |
| rs1142345 | Y240S | Deleterious | Probably damaging | Probably damaging | Medium | Neutral | Disease | Deleterious | Neutral | Deleterious | Disease | Disease | Deleterious |

Table 1. List of missense nsSNPs predicted to be deleterious by at least nine out of 12 sequence-based tools.

| Variant ID | SNP | CUPSAT | DUET | | | I-Mutant | MUpro | | INPS-MD | PoPMuSiC | HoTMuSiC | SNPMuSiC |
|--------------|-------|---------------|---------------|---------------|---------------|----------|----------|----------|----------|-------------------|-------------------|-------------------|
| | | | mCSM | SDM | DUET | | SVM* | NN* | | | | |
| rs72552741 | W33G | Destabilising | Destabilising | Destabilising | Destabilising | Decrease | Decrease | Decrease | Decrease | Strongly Decrease | Strongly Decrease | Strongly Decrease |
| rs753277177 | W78R | Destabilising | Destabilising | Destabilising | Destabilising | Decrease | Decrease | Decrease | Decrease | Decrease | Decrease | Strongly Decrease |
| rs1784191846 | V89E | Destabilising | Destabilising | Destabilising | Destabilising | Decrease | Decrease | Decrease | Decrease | Strongly Decrease | Strongly Decrease | Decrease |
| rs1447033392 | W150G | Destabilising | Destabilising | Destabilising | Destabilising | Decrease | Decrease | Decrease | Decrease | Strongly Decrease | Strongly Decrease | Strongly Decrease |
| rs1386533390 | L182P | Destabilising | Destabilising | Destabilising | Destabilising | Decrease | Decrease | Decrease | Decrease | Strongly Decrease | Strongly Decrease | Decrease |
| rs1142345 | Y240S | Destabilising | Destabilising | Destabilising | Destabilising | Decrease | Decrease | Decrease | Decrease | Strongly Decrease | Decrease | Decrease |

Table 2. List of missense nsSNPs predicted to be deleterious by at least six out of eight structure-based tools (As Y240S has already been identified to be deleterious, it is not a novel deleterious SNP and hence, it was not considered for further in silico analyses). *SVM = Support Vector Machines Method, NN = Neural Networks Method.

| Variant ID | SNP | Sequence-based tools (12) | | | | | | | | | | | | Structure-based tools (8) | | | | | | | | Tools (20) | Nature | |
|-------------|-------|---------------------------|----|-----|-----|----|----|-----|----|----|----|----|----|---------------------------|-----|-----|-----|-----|-----|-----|-----|------------|--------|--|
| | | PV | PT | PP2 | MA | MS | PM | PS1 | S2 | SP | SG | PS | ST | C | D | IM | MP | IMD | PPM | HTM | SM | | | |
| rs9333569 | M1V | N | D | N | NA* | N | N | N | N | N | N | N | D | NA* | NA* | NA* | NA* | NA* | NA* | NA* | NA* | NA* | NA* | D ⁹¹ , VD ⁷⁷ |
| rs72552742 | E28V | D | N | N | N | N | D | N | D | N | D | D | D | D | N | N | D | N | D | D | N | N | 9 | D ^{77,80,81} |
| NA** | Q42E | N | N | N | N | N | N | N | N | N | N | N | N | D | N | N | N | N | D | D | N | N | 3 | D ^{77,81} |
| rs72552740 | L49S | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | 20 | VD ^{77,80,81} |
| rs200591577 | L69V | D | D | D | D | D | D | D | D | D | D | D | D | D | N | D | D | D | D | D | N | N | 18 | VD ^{75,83} |
| rs777686348 | G71R | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | N | D | D | D | D | N | 19 | VD ^{77,81} |
| rs281874771 | A73V | D | D | D | N | D | D | D | D | D | N | N | D | D | N | N | N | N | D | D | D | N | 14 | VD ⁸⁰ |
| rs1800462 | A80P | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | D | 20 | VD ^{77,80,81} |
| rs111901354 | R82W | D | N | D | N | D | N | D | D | D | D | D | D | D | D | N | N | D | D | D | N | N | 14 | VD ⁷⁴ |
| NA** | Y107D | D | D | D | N | D | D | D | D | N | D | D | D | D | D | D | D | D | D | D | D | D | 18 | VD ⁷⁶ |
| rs115106679 | E114K | N | N | D | N | N | N | N | N | N | N | N | D | D | N | D | D | N | D | D | N | N | 7 | VD ^{74,86} |
| rs200220210 | S125L | D | N | N | N | N | N | N | N | N | N | N | D | N | N | N | N | N | D | D | D | N | 5 | D ^{77,80,81} |
| rs72552738 | C132Y | D | D | N | N | D | D | D | N | D | D | D | D | D | D | D | D | D | D | D | D | D | 16 | D ^{77,81} , VD ⁸⁰ |
| rs72552737 | G144R | D | D | D | N | D | N | D | D | N | D | D | D | D | D | D | N | D | D | D | D | D | 16 | D ^{77,80,81} , VD ⁷⁵ |
| rs1800460 | A154T | D | N | N | D | D | D | D | D | D | D | D | D | N | D | D | D | N | D | D | D | D | 16 | VD ^{11,75,76,80,87,89} |
| rs112339338 | R163C | D | D | D | D | D | D | D | D | D | D | D | D | N | D | D | D | N | D | D | D | D | 18 | VD ⁷⁴ |
| rs144041067 | R163H | D | D | D | N | D | N | D | D | D | N | D | N | D | D | D | D | D | D | D | N | N | 15 | VD ^{72,77,81} |
| NA** | R163P | D | D | D | D | D | D | D | D | D | D | D | D | N | D | D | D | D | D | D | D | D | 19 | VD ^{75,77} |
| rs201695576 | Y166C | D | D | D | D | D | D | D | D | D | D | D | D | N | D | D | D | D | D | D | D | D | 19 | D ⁷⁹ |
| rs74423290 | A167G | D | N | N | N | N | D | D | D | D | N | D | N | D | D | D | D | D | D | D | D | D | 15 | D ⁹¹ |
| rs72556347 | F208L | N | D | N | N | N | N | N | N | N | N | D | N | D | D | D | D | D | D | D | D | D | 10 | D ⁹⁰ |
| rs377085266 | C212R | D | D | D | N | D | D | D | D | D | D | D | N | D | D | D | D | D | D | D | D | D | 18 | VD ⁸³ |
| rs150900439 | K238E | N | N | N | N | N | N | N | N | N | N | N | D | N | N | D | N | D | D | D | N | N | 5 | D ^{75,77} |
| rs1142345 | Y240C | D | D | D | D | D | D | D | D | D | N | D | D | D | D | D | N | D | D | D | D | D | 19 | VD ^{76,77,78,82,87} |
| rs1142345 | Y240S | D | D | D | N | N | D | D | N | D | D | D | D | D | D | D | D | SD | D | D | D | D | 18 | VD ⁸⁵ |

Table 3. Screening of pre-identified SNPs to validate the screening process (SNPs that would get selected as per the screening process utilised in this study are highlighted in bold. PV = PROVEAN, PT = PANTHER, PP2 = PolyPhen-2, MA = Mutation Assessor, MS = Meta-SNP, PM = PMut, PS1 = PredictSNP1, S2 = SNAP 2, SP = SuSPect, SG = SNPs&GO, PS = PhD-SNP, ST = SIFT, C = CUPSAT, D = DUET, IM = I-Mutant, MP = MUpro, IMD = INPS-MD, PPM = PoPMuSiC, HTM = HoTMuSiC, SM = SNPmuSiC, VD = Very Dangerous, D = Dangerous, N = Neutral). *NA = Not Available (Residues 1 to 16 are absent in the PDB file 2H11, hence structural analysis tools could not be used for this SNP. Mutation Assessor also did not yield any result for this SNP). **NA = Not Available (This pre-identified SNP is not annotated in dbSNP).

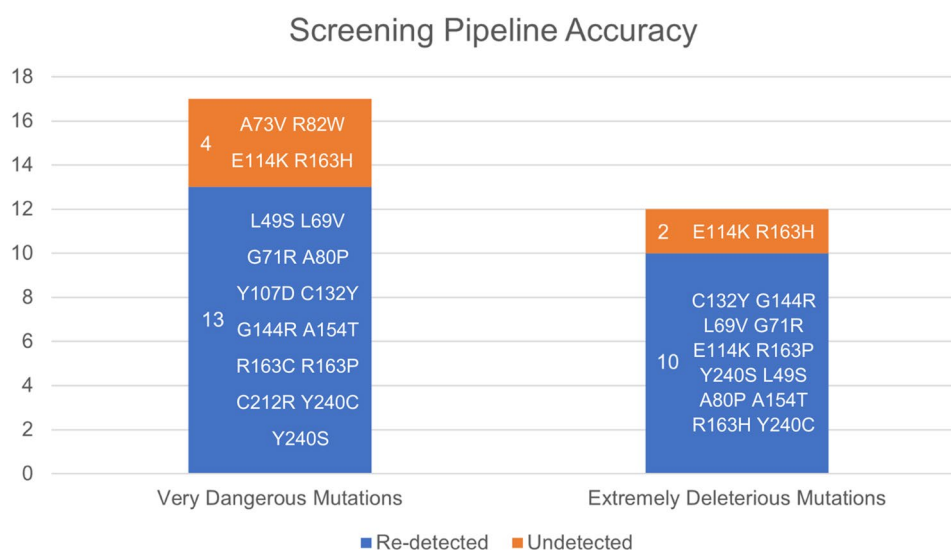


Figure 2. Accuracy of the in silico screening pipeline with respect to the pre-identified deleterious missense mutations of human TPMT.

sequences had the same amino acid residue at the residue location under consideration. For instance, the amino acid “W” or tryptophan was conserved at residue 33 across all 10 sequences and hence, was assigned a score of 10 for MSA. If an amino acid was at a particular residue location only for human TPMT and none of the other nine TPMT sequences, it would be assigned a score of 1. As for CDD, 13 of the most dissimilar sequences of TPMT were considered for MSA, and scores ranging from 1 to 13 were considered, assigned similarly to MSA depending on how many sequences had that particular amino acid conserved at the given residue location.

The ConSurf results showed that four of the five screened mutations, W33G, W78R, V89, and W150G were present at residues that were highly conserved as all of them were on residues possessing a color score of 8 or 9. Residue 182 (where the L182P mutation occurs) had a slightly above-average conservation score of 6. The color scores corresponding to all amino acid residues of TPMT as predicted by ConSurf are shown in Table S4. They are structurally represented for chain A of TPMT (2H11) in Figure S2 as a still image and as an animation for better visualization through the POLYVIEW-3D server⁹⁴ in Animation 1. The conservation analysis results obtained through the five MSA tools, ConSurf and CDD with respect to the residues associated with the five screened SNPs are presented in Table S5. All of them were conserved across the 10 aforementioned species whose TPMT sequences were considered for MSA. In addition, CDD which identified the most diverse TPMT sequences and performed MSA found that all but one SNP (W150G) were at residues that were conserved across at least seven of the most dissimilar TPMT sequences indicating that those residues have stood the test of time and evolution to remain a part of the amino acid sequence of the protein, presumably due to their important contribution to the desired functioning or maintenance of structural stability of the protein. It is worth noting that this does not invalidate the candidature of W150G as a potentially deleterious missense SNP of TPMT, since despite the CDD results, residue 150 had a color score of 8 as per ConSurf results. Furthermore, conservation level is not an absolute indicator, but rather a good general indicator of deleterious nature.

As per ConSurf, 108 residues (46.55%) possessed color scores that were greater than or equal to 6, and 47 (18.1%) among them were highly conserved having a conservation score of 8 or 9. Among these 108 residues, the largest conserved region consisted of residues 64–110 (except residues 85, 91, 94, 97, 98, 101, 104, 106, and 108) having 38 residues with color scores greater than or equal to 6 and 20 highly conserved residues having color scores greater than or equal to 8. The regions covering the ligand (S-adenosyl-L-methionine or SAM) attachment site are from residues 29 to 40 and 134 to 135. Hence, ligand binding does not explain the high conservation level of this region. Further investigation is required to deduce the high level of conservation observed in this region.

Interaction analysis. The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database was used to identify the interaction network of human TPMT. The output was given in the form of edges and nodes that represent interactions and proteins, respectively, as shown in Figure S3. All of the predicted functional partners of TPMT and their respective scores are shown in Table S6. These scores indicate how likely an interaction is to be true, as evaluated by STRING.

The interaction analysis revealed that the TPMT interaction network is primarily involved in the biosynthesis of guanosine monophosphate (GMP) and guanosine triphosphate (GTP), catabolism of deoxyribonucleoside triphosphate (dNTPs) and the biosynthesis as well as degradation of purines. GMP is a nucleotide that is utilized as a monomer in RNA. GTP is an extremely essential nucleotide in the body serving as a medium of energy transfer in the cell, similar to adenosine triphosphate (ATP), albeit more specific than its more universal counterpart, ATP. It is also important in signal transduction, most prominently in G-proteins. Furthermore, cyclic GMP (cGMP) which is derived from GTP, is a very important second messenger in the body involved in key biological processes, including but not limited to glycogenolysis, cellular apoptosis, vasodilation, phototransduction in the eye, and more. Due to the importance of the network with regard to purines and dNTPs, it is very important for the regular functioning of most cellular processes. Interestingly, the network is also involved in lymphocyte proliferation reaffirming an immune response association which is in stark contrast to the importance of TPMT in the methylation of thiopurine drugs that treat autoimmune disorders like Crohn’s disease and rheumatoid arthritis.

Suffice it to say that the TPMT interaction network is a very important network involved in several key biological processes. All major biological processes in which the TPMT interaction network is involved, as per the STRING database are shown in Table S7. As mentioned before, TPMT polymorphisms can severely alter enzyme activity and the appropriate conduction of these key biological processes that the network is involved in, are subject to the deleterious consequences of these dangerous polymorphisms.

Oncogenic analysis and phenotypic analysis. As mentioned before, TPMT polymorphisms are known to be associated with cancers. Hence, to predict the oncogenic nature of the screened SNPs, CScape and CScape-somatic were used. The former predicts the oncogenic nature of somatic point mutations that occur in the coding regions of the cancer genome with an accuracy of 92%. The input is given using the format: chromosome, position, reference, mutant; as per the GRCh38 assembly for one or multiple mutations. The output is in the form of p-values which can lie between 0 and 1 and values above 0.5 are considered deleterious. On the other hand, values below 0.5 are considered benign. It predicted that all five mutations are oncogenic and W33G with high confidence. CScape-somatic is a tool that can differentiate between mutations that can be cancer drivers and mutations that can be passenger variants. Cancer drivers occur fairly early in the development of the tumor. Meanwhile, passenger variants accumulate at later stages after a tumor starts to grow and usually correspond to low or no oncogenicity. The input and accuracy of CScape-somatic are identical to CScape, except for the fact that the GRCh37 assembly is used in the input. Akin to CScape, the predictions of CScape-somatic are again given as p-scores that lie between 0 and 1. Mutations that possess p-scores above 0.5 are predicted to be cancer drivers. Conversely, mutations that possess p-scores 0.5 are predicted to be passenger variants. CScape-somatic

| Variant ID | SNP | CScape | | | CScape-somatic | | |
|--------------|-------|---------------------------|--------------|-----------------|---------------------------|--------------|--------------|
| | | Input & Assembly | Coding Score | Message | Input & Assembly | Coding Score | Message |
| rs72552741 | W33G | 6,18,149,031,A,C & GRCh38 | 0.898663 | Oncogenic (HC*) | 6,18,149,262,A,C & GRCh37 | 0.907987 | Driver (HC*) |
| rs753277177 | W78R | 6,18,147,824,A,G & GRCh38 | 0.695307 | Oncogenic | 6,18,148,055,A,G & GRCh37 | 0.820811 | Driver |
| rs1784191846 | V89E | 6,18,143,696,A,T & GRCh38 | 0.807748 | Oncogenic | 6,18,143,927,A,T & GRCh37 | 0.781448 | Driver |
| rs1447033392 | W150G | 6,18,139,009,A,C & GRCh38 | 0.860191 | Oncogenic | 6,18,139,240,A,C & GRCh37 | 0.728649 | Driver |
| rs1386533390 | L182P | 6,18,133,839,A,G & GRCh38 | 0.519117 | Oncogenic | 6,18,134,070,A,G & GRCh37 | 0.841487 | Driver |

Table 4. : Oncogenic nature of mutations predicted using CScape and CScape-somatic. *HC = High Confidence.

predicted that all five mutants are cancer drivers and W33G with high confidence. The results obtained from both oncogenic tools are presented in Table 4.

In addition to CScape and CScape-somatic, Dr Cancer was also used for the assessment of the oncogenic potential of SNPs. It is a slightly older tool than the CScape tools that utilize a disease-specific machine learning approach to predict whether a missense SNP is associated with cancer or not. It does so by utilizing multiple methods including SEQPROF where the sequence and profile at the mutated position are the input for the Support Vector Machine (SVM), SVM-GOS where the GOS (Gene Ontology Score) of the mutated sequence is the input for the SVM, and finally, SEQPRFGO where the input for the SVM is the input provided for the two prior methods. The first method, SEQPROF predicted that all five mutations are cancer-associated while the other two methods did not predict any of the mutations to be cancer-associated. The results pertaining to the five mutations obtained from Dr Cancer are shown in Table S8.

Functional Analysis through Hidden Markov Models (FATHMM) server was used for phenotypic analysis as it predicts the phenotypic consequences of mutations in the human genome. The input given was the Uniprot ID of TPMT along with the amino acid substitutions associated with the five deleterious missense SNPs. The tool classified only L182P as a dangerous mutation, associated with abnormalities of the head, neck, face, mouth, and philtrum. The results obtained from the FATHMM server are presented in Table S9.

Structural effects of deleterious missense SNPs. W33G at residue 33 is present at a ligand binding site and will undoubtedly have an effect on the ligand binding process and subsequent functioning of the protein. V89E at residue 89 and W150G at residue 150 is extremely close to the ligand binding sites at residues 90 and 150 respectively and hence, they may also impact the ligand binding process. The HOPE server was used to visualize the five highly deleterious missense SNPs, as shown in Animation 2. It was found that except V89E, all mutant residues were smaller in size when compared to their wild-type counterparts. Hence, with respect to W33G, W78R, W150G, and L182P, it was predicted that the mutant residues would not fit in the core of the protein since they are larger than the respective wild-type counterparts. All structural consequences of the mutations as predicted by the HOPE server have been presented in Table S10.

The secondary structure alterations caused by the mutations, as obtained from NetSurfP-2.0 are shown in Figure S4. It was observed that L182P caused the helix to start from residue 47 instead of residue 46 in the wild-type structure of TPMT (2H11). However, NetSurfP-2.0 predicted this coil to start at residue 47 itself, which is incorrect. Regardless, if the predicted change does occur, it might be damaging due to the change induced with respect to the helix. W33G and L182P replaced a coil with a strand at residue 122. However, in the wild-type structure of TPMT (2H11), this residue is already under the conformation of a strand. This is because NetSurfP-2.0 did not correctly predict the confirmation at residue 122 for the wild-type sequence of TPMT. Since the wild-type structure of TPMT already has that conformation at residue 122, if the mutation causes the predicted change, it will make no difference. A similar change was observed with respect to W33G and V89E, but here the strand was replaced by a coil at residue 68. However, the wild-type structure of TPMT (2H11) already has a coil at residue 68 which was not predicted by NetSurfP-2.0. Hence, this predicted change will also cause no damage. Yet another change akin to the aforementioned changes was observed where L182P replaced a strand with a coil at residue 187, but in the wild-type structure of TPMT, this residue is already in the confirmation of a coil. Hence, this predicted change will also cause no damage. Due to the inconsistencies between the secondary structure prediction of wild-type TPMT by NetSurfP-2.0 and the actual wild-type structure of TPMT, it is difficult to draw any concrete conclusions regarding the secondary structure changes. Hence to infer some credible changes caused by the mutations in the secondary structure and other structural aspects of the protein, MDS was carried out for a period of 100 ns, the results of which are discussed in “Molecular dynamics simulations analysis” section.

The CASTp 3.0 server was used to assess the alterations in the binding pockets caused by the mutations. The results for the same are shown in Table S11. It was observed that W33G and W150G vastly increased the size of the largest binding pocket of TPMT which will undoubtedly impact ligand binding as residue 33 is a ligand binding site. W78R and V89E did not impact the binding pocket dimensions significantly. L182P introduced a few pockets that are absent from the wild-type protein albeit of a very minuscule size that is unlikely to impact

the binding process much. However, it also introduced a larger binding pocket that is absent from the wild-type protein. This might affect the binding process.

Finally, AmylPred 2 was used to identify amyloid-forming regions in TPMT. Amyloid refers to proteinaceous and abnormally fibrous extracellular deposits. It is insoluble, mostly possesses a β -sheet conformation and is associated with a variety of diseases called amyloidoses, such as Alzheimer's disease, type 2 diabetes and more. They are usually progressive disorders associated with high morbidity and mortality⁹⁵. The amyloid-forming regions predicted by AmylPred 2 for TPMT include residue groups 64–69, 86–89, 100–101, 127–138, 147–149, 154–158, 170–175, 178–187, 207–217 and 231–243. W150G is just one residue away from the predicted amyloid region comprised of residues 147–149, while V89E and L182P fall under these regions with residues 86–89 and 178–187, respectively. Hence, they may influence these predicted amyloid-forming regions and worsen disease progression.

Post-translational modification analysis. GPS 5.0 software was used to predict the PTM sites of TPMT. The predicted PTM sites and their distribution are shown in Figure S5. While none of the mutations are at PTM sites, W33G is near the PTM site T38, V89E is in proximity to S92 and S85. Furthermore, L182P is in close proximity to the PTM site Y180. It is very much possible that due to their proximity to PTM sites, W33G, V89E and L182P can impact the phosphorylation of the protein, which is essential for the natural functioning of the protein.

Molecular dynamics simulations analysis. MDS was performed to identify the structural consequences of the mutations in a dynamic fashion and at an atomic level. Seven systems (Wild-Type Apo protein (WT), Wild-Type protein with SAH ligand attached (WT-SAH), W33G, W78R, V89E, W150G, and L182P) were generated for MDS, following which MDS was performed for a simulation time of 100 ns through GROMACS. All of the 10,000 frames from the 100 ns of simulation time were recorded for the WT, WT-SAH, W33G, W78R, V89E, W150G, and L182P systems and are shown in Animations 3, 4, 5, 6, 7, 8, and 9 respectively. From Animation 4 (WT-SAH) and Animation 5 (W33G) it can be clearly seen that the ligand interactions upon stabilization do not resemble each other. As mentioned before, W33G occurs at a ligand binding site and is undoubtedly the reason for the difference in ligand interactions between the W33G mutant and the wild-type protein. From Animation 6 (W78R) it is apparent that the ligand interactions are different from the wild-type protein, however not to the extent observed in W33G. The difference is even less apparent in V89E (Animation 7). However, from Animation 8 (W150G), it can be clearly seen that the ligand interactions in the W150G mutant, upon stabilization, are very different from the wild-type protein. Furthermore, from Animation 9 (L182P), it can also be seen that the ligand interactions are different from the wild-type protein. These results are in line with the CASTp 3.0 results where W33G and W150G affected the largest binding pocket of TPMT and L182P introduced and altered several pockets, while W78R and V89E did not have much impact on the binding pockets. Trajectory analysis and MM-PBSA analysis were performed in order to quantify the MDS results.

RMSD analysis. RMSD was calculated for all seven protein systems to assess their respective stabilities. RMSD as a parameter is useful for assessing the stability of the protein relative to its conformation. Smaller RMSD deviations correspond to a more stable protein structure. A plot of RMSD against time for all systems is shown in Fig. 3A. The apo form of the wild-type protein and the ligand-bound form had average RMSD values of 0.301 nm and 0.353 nm, respectively. From Fig. 3A, it can be seen that the apo form of TPMT is more stable than the ligand-bound TPMT and the average RMSD values reflect the same. Meanwhile, the mutants W33G, W78R, V89E, W150G and L182P had average RMSD values of 0.317, 0.392, 0.335, 0.331 and 0.291 nm, respectively. It was also observed that the W150G mutant showed the most RMSD fluctuations till the first 50 ns. Interestingly, the L182P mutant protein appeared to possess higher stability than both the apo form and ligand-bound wild-type conformations of the protein. Furthermore, only W78R had an average RMSD value higher than ligand-bound TPMT, implying all other mutations are more stable than ligand-bound TPMT as per RMSD analysis. The RMSD results indicated that all systems attained stability after 30 ns. Hence, the last 70 ns of the trajectory were considered with respect to the calculations pertaining to the rest of the MDS parameters.

RMSF analysis. RMSF was used to assess the flexibility and stability of the seven systems. High RMSF values usually signify greater flexibility during MDS. A plot showcasing the RMSF values against all residues for all systems is shown in Fig. 3B. The apo form of the wild-type protein and the ligand-bound form had average RMSF values of 0.120 nm and 0.132 nm, respectively, again indicating that the apo form of TPMT is more stable than the ligand-bound TPMT. Meanwhile, the mutants W33G, W78R, V89E, W150G, and L182P had average RMSF values of 0.108, 0.123, 0.100, 0.148, and 0.108 nm, respectively. Interestingly, mutants W33G, V89E, and L182P appear to be more stable than both the apo form and ligand-bound wild-type conformations of the protein. Furthermore, only W150G had an average RMSF higher than ligand-bound TPMT, implying all other mutations are more stable than ligand-bound TPMT at least in the 100 ns of simulation time and as per RMSF.

Radius of gyration analysis. The radius of gyration (R_g) of a system indicates the root-mean-square distance of the atoms of the protein in the system from the axis of rotation and is useful to assess protein structure compactness. Low R_g values indicate a more compact and rigid protein structure, while high R_g values indicate that the protein is less compact and flexible. A plot showcasing the variation of R_g with respect to time for all systems is shown in Fig. 3C. The apo form of the wild-type protein and the ligand-bound form had average R_g values of 1.704 nm and 1.726 nm, respectively. Hence, the apo form of TPMT is less flexible than the ligand-bound form of TPMT. Meanwhile, the mutants W33G, W78R, V89E, W150G, and L182P had average RMSF

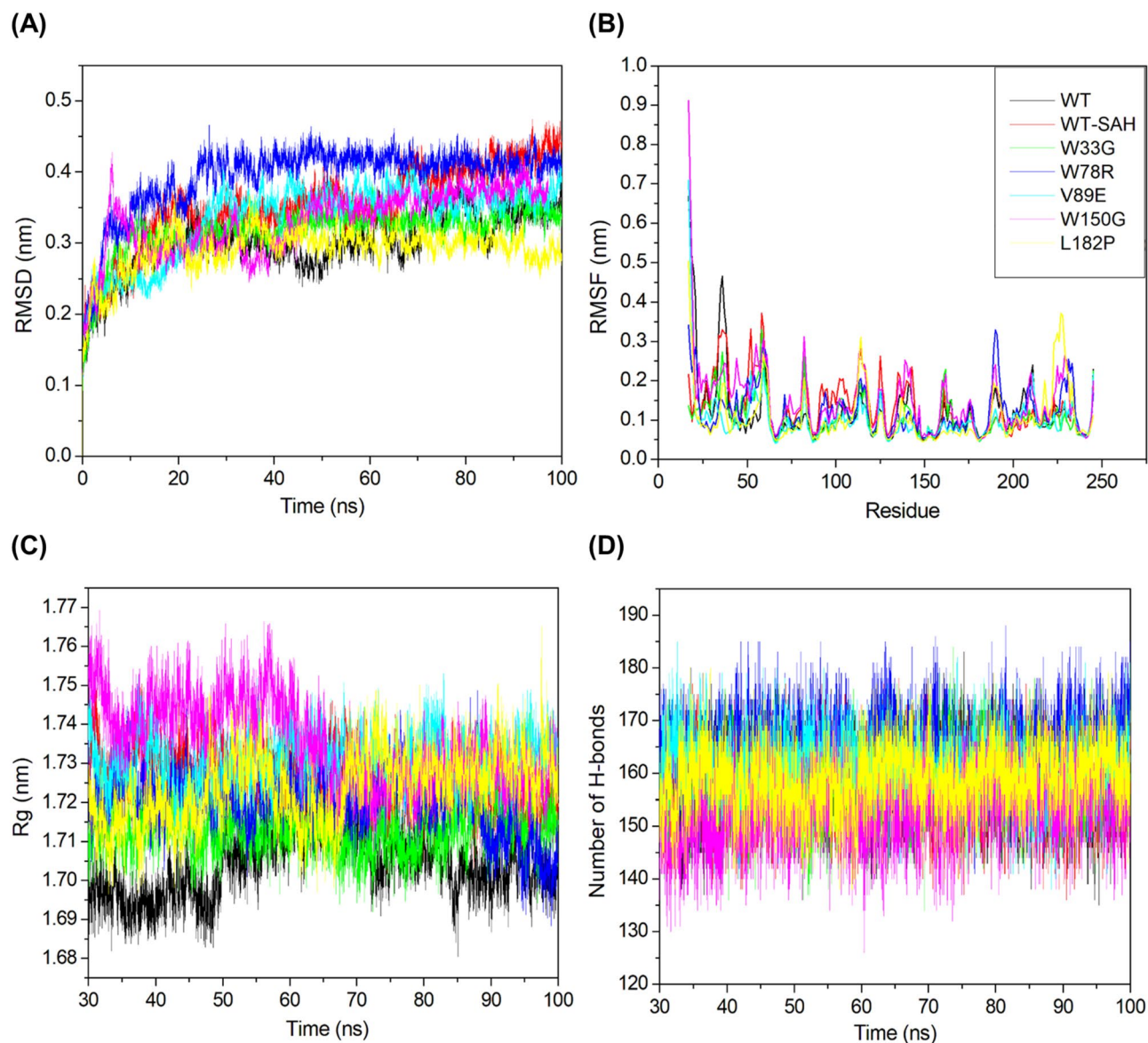


Figure 3. RMSD, RMSE, Rg and H-bond analysis results.

values of 1.713, 1.721, 1.730, 1.735, and 1.724 nm, respectively. The apo form of the protein was found to be the most stable among all systems. V89E and W150G were the only systems that were less compact and more flexible than ligand-bound TPMT, implying all other mutations are more stable as per the Rg values.

Hydrogen bonding analysis. Hydrogen bonds are important for maintaining the stability of the protein structure. A greater number of average hydrogen bonds correspond to a more stable protein structure, and conversely, a fewer number of average hydrogen bonds correspond to a less stable protein structure. A plot showcasing the number of intramolecular hydrogen bonds against time for all systems is shown in Fig. 3D. The apo form of the wild-type protein and the ligand-bound form had 158 and 157 hydrogen bonds on average, respectively. While this does mean that, yet again ligand-bound TPMT is less stable than the apo form of TPMT, the difference is very minuscule. Meanwhile, the mutants W33G, W78R, V89E, W150G, and L182P had 158, 165, 160, 153, and 159 hydrogen bonds on average, respectively. W33G and L182P seemed to have a negligible effect on the hydrogen bond formation. However, apart from W150G, the average number of hydrogen bonds was more than or equal to the average hydrogen bonds in the apo form of TPMT, implying that all other mutations are more stable than the apo form of TPMT.

SASA analysis. SASA as a parameter is useful for assessing the portion of the protein surface that is exposed to the water solvent. A plot of SASA against time for all systems is shown in Fig. 4A, while a plot of SASA against all residues for all systems is shown in Fig. 4B. The apo form of the wild-type protein and the ligand-bound form had average SASA values of 117.414 nm² and 124.768 nm², respectively. Their average residual SASA values were found to be 0.512 nm² and 0.545 nm², respectively. Both parameters indicate that the ligand-bound TPMT has a

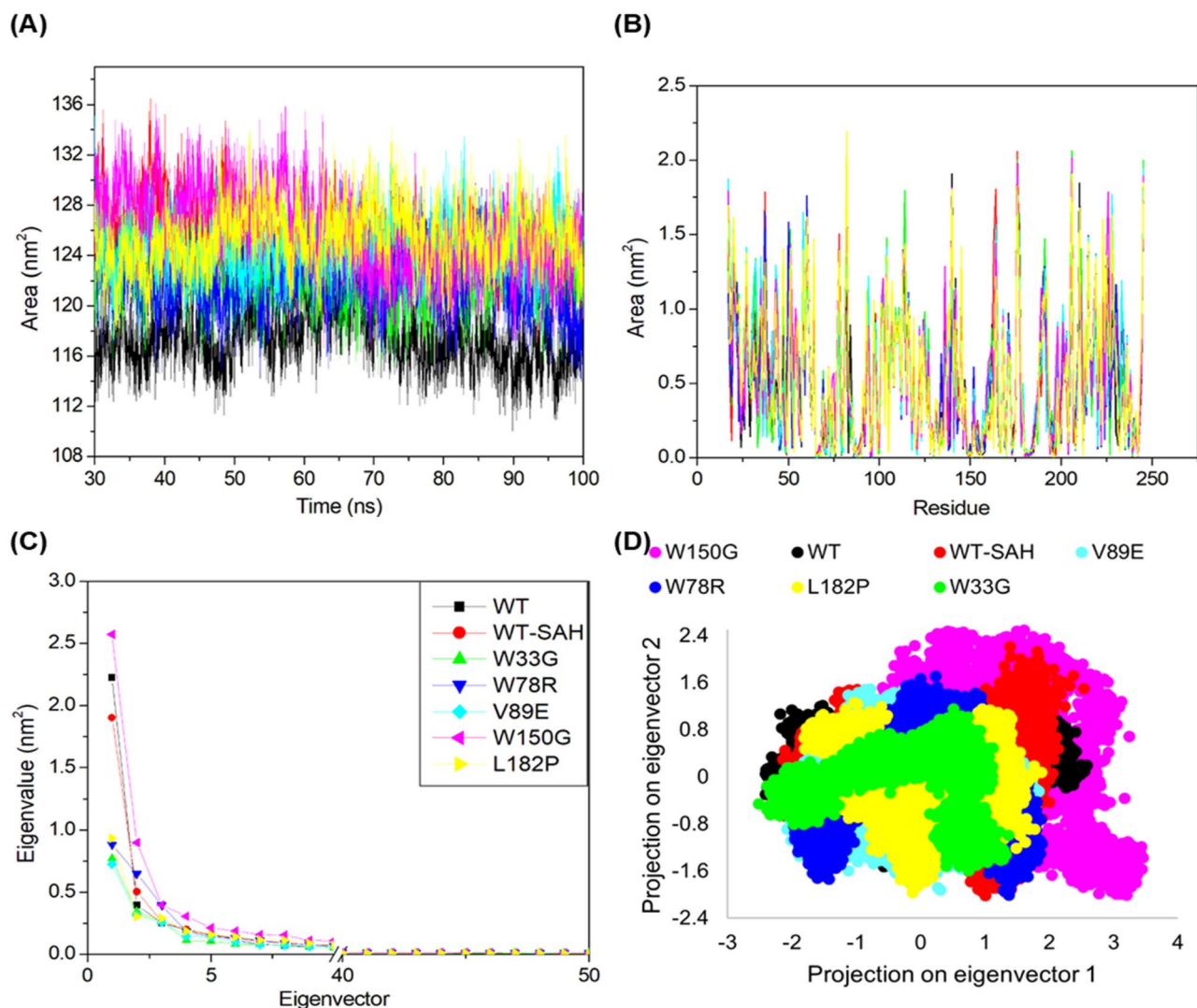


Figure 4. SASA analysis and PCA results.

hydrophobic core that is more exposed to the solvent than the apo form of TPMT than the apo form of TPMT. This could be essential for the appropriate functioning of the protein. Meanwhile, the mutants W33G, W78R, V89E, W150G, and L182P had average SASA values of 122.214, 121.862, 124.490, 126.017, and 125.385 nm², respectively, and average residual SASA values of 0.534, 0.532, 0.543, 0.550 and 0.557 nm², respectively. Hence, mutants W33G, W78R, and V89E have more hydrophobic cores that are more protected from the solvent when compared to the ligand-bound form of TPMT. However, this could be detrimental as a certain degree of flexibility is important for the normal functioning of the protein. On the other hand, mutants W150G and L182P appeared to have more hydrophobic cores that are more exposed to the solvent than ligand-bound TPMT as per the SASA analysis.

This is in agreement with respect to the results from prior analyses where the apo form of TPMT was always found to possess more stability than ligand-bound TPMT, however excess stability or plasticity of the protein can actually hamper its natural functioning and a certain degree of flexibility is needed for the protein to bind to the ligand and perform its intended function⁹⁶. From the results obtained thus far, it would seem that most mutations increase the plasticity of the protein structure and thereby reduce its functional capabilities. In fact, in another study where MDS was performed for the Y240S mutant of TPMT, the RMSD graph of the Y204S mutant showed a lesser average value than wild-type TPMT, despite the fact that in the same study Y204S was proven to be very deleterious through *in vivo* techniques¹⁹. Another study found higher SASA values in the three most well-known and deleterious TPMT alleles (A80P, A154T, and Y240C) than the SASA values of WT-TPMT⁹⁷. Hence, it is not unprecedented for enhanced stability resulting from missense mutations to have a negative effect on TPMT. Hence, these results are in line with prior mutational studies of TPMT. To further validate this phenomenon of functional hampering due to enhanced plasticity in the case of TPMT polymorphisms, the MM-PBSA method was utilized to assess the binding energy and its components for all six ligand-bound systems in “[Binding energy analysis](#)” section.

PCA. PCA was also performed in order to assess the collective motions of the seven systems. The motions of the C α atoms of a given protein were the basis of PCA. Eigenvalues were used to quantify the atomic contribution to the collective motion and eigenvectors were used to quantify the overall direction of motion of atoms. The eigenvalues were plotted in decreasing order against the respective eigenvector for all seven systems in Fig. 4C. As is evident from Fig. 4C, apart from W150G, the eigenvalues of all mutations possessed amplitudes lower than that of the apo form and the ligand-bound form of TPMT, implying more rigidity in the collective motions upon mutations. This is in line with prior analyses as W150G was the only mutation that was predicted to be unstable by three parameters (RMSF, Rg, H-bonds, SASA). The first two eigenvectors were plotted oppositely in phase space where each of the continuum spectra represents the correlated motions. The analyzed clusters of all seven systems were plotted and are depicted in Fig. 4D. As is evident from prior analyses, most mutations with the exception of W150G, would be expected to reduce the collective motion of the system due to the enhanced plasticity. The PCA results line up with this expectation as it can be seen in Fig. 4D that apart from W150G, the collective motions of all mutations are more restricted than both the apo form and the ligand-bound form of TPMT, with W33G having the most restricted motion. This lines up with the Rg results where W33G was found to be the most rigid mutation system. Hence, it was observed that none of the mutant systems developed motion clusters that were similar to that of the apo form or ligand-bound form of TPMT, reaffirming the fact that they have a deleterious impact on TPMT.

Secondary structure analysis. SSA was performed for all systems, and the results for the same are presented in Figure S6. In the apo form of the protein, a 5-turn helix conformation exists between residues 17 and 27 after 85 ns. However, in the WT-SAH system, the residues were not in this conformation after 85 ns. Furthermore, close to residue 177, WT-SAH replaced the β -sheet conformation observed in the apo form, with bends. It also replaced the α -helix near residue 137 with turns for a brief period but stabilized similarly to the apo form near the end of simulation time. Most notably, the α -helix between residues 222 and 237 was completely replaced by bends in the WT-SAH system over the entire simulation time.

All mutant comparisons regarding secondary structure are with respect to the WT-SAH system, as all mutant systems are also ligand-bound. It was observed that in all mutant systems, residues 17 to 27 seemed to lose a lot of their secondary structure conformations, which is problematic as residue 26 is a ligand binding site. With respect to the W33G system, it very notably replaced the α -helix conformation between residues 27 and 37 with a 5-turn helix conformation for nearly the entirety of the simulation time. This is probably due to the difference in properties between tryptophan and glycine as discussed in the HOPE results, especially glycine being more flexible than tryptophan. This is also in line with the CASTp 3.0 results where this mutation massively altered the largest protein binding pocket and was predicted to impact ligand binding. However, towards the end of the simulation time, the 5-turn helix conformation was replaced by turns. It is still very much possible that ultimately the conformation of that region would be a 5-turn helix in the W33G system, as that was the conformation conferred on the protein for the majority of the simulation time. The region in question covers three ligand binding sites at residues 26, 29, and 33. Hence, such a change could drastically impact ligand binding. The W33G system also brought back the α -helix conformation observed in the apo form of the protein between residues 222 and 237, which was absent in the WT-SAH system. As observed in prior analyses, W33G tends to enhance the plasticity of the protein. This was also observed in the variation of the secondary structure across time, wherein W33G had far lesser variation in the secondary structure conformations over time when compared to the WT-SAH system.

W78R on the other hand did introduce some more secondary structure variation in the initial residues and close to the mutation site, but reduced variation in other places. Furthermore, the α -helix between residues 222 and 237 as seen in the W33G system, and the apo form of the protein was also seen in the first 50 ns in the W78R system, post which it was largely replaced by turns. Similar to W78R, the V89E system seemed to reduce the structural variation when compared to the WT-SAH system, although not to the degree observed in the W33G system, probably due to the enhanced plasticity caused by the V89E mutation. Notably, the α -helix conformation near residue 137 as observed in the WT-SAH system for a majority of the simulation time, was absent from the V89E system. This could be problematic for ligand binding as the region near residue 133 contains three ligand binding sites at residues 133, 134 and 135. The α -helix between residues 222 and 237 did appear in the first 5 ns in the V89E system as well, however it was replaced by bends after that. W150G appeared to have a slightly increased variation in the secondary structure when compared to the WT-SAH system which is in line with prior MDS analyses where W150G was predicted to impart instability to the system. It also introduced a 5-turn helix conformation between residues 47 and 57 as opposed to an α -helix conformation in the WT-SAH system. The α -helix between residues 222 and 237 also appeared in the W150G system, albeit only for the first 5 ns similar to V89E post which it only appeared on a few occasions. Curiously, L182P also created an α -helix conformation between residues 222 and 237, which lasted for over 60 ns in the 100 ns of simulation time. Akin to other mutations, the L182P system imparted enhanced plasticity to the protein and as a result, showed a reduced variation in secondary structure conformations across the simulation time to a degree similar to that observed in the W33G system. This lines up with the PCA results where L182P showed a restricted collective motion to a degree similar to W33G.

This common trend of introducing the α -helix conformation between residues 222 and 237 in mutant systems could be due to the fact that its presence might be partially responsible for enhanced plasticity. Upon closer inspection, it was found that while this residue region was in fact not in the α -helix conformation in the 100 ns simulation time of the WT-SAH system, it was in the α -helix conformation in the final ligand-bound TPMT structure i.e., in the structure with PDB ID 2H11. Furthermore, this particular stretch of residues harbored 28 hydrogen bonds in the final structure. Figure S7 (generated using UCSF Chimera⁹⁸) shows the hydrogen bonds present in the final ligand-bound TPMT structure (2H11) between the stretch of residues from residue 221 to

| System | Van der Waals energy (kJ/mol) | Electrostatic energy (kJ/mol) | Polar solvation energy (kJ/mol) | SASA energy (kJ/mol) | Binding energy (kJ/mol) |
|--------|-------------------------------|-------------------------------|---------------------------------|----------------------|-------------------------|
| WT-SAH | -215.76 ± 11.86 | -47.82 ± 9.34 | 146.29 ± 12.92 | -21.48 ± 0.92 | -138.77 ± 14.68 |
| W33G | -0.004 ± 0.00 | 0.00 ± 0.00 | 16.21 ± 50.22 | -0.40 ± 3.47 | 15.81 ± 50.25 |
| W78R | -0.005 ± 0.00 | 0.00 ± 0.00 | 5.13 ± 61.70 | -0.47 ± 3.52 | 4.65 ± 61.61 |
| V89E | -0.005 ± 0.00 | 0.00 ± 0.00 | -17.84 ± 47.13 | 0.39 ± 3.64 | -17.45 ± 47.15 |
| W150G | -0.005 ± 0.00 | 0.00 ± 0.00 | -28.27 ± 41.29 | -0.52 ± 3.85 | -28.8 ± 41.57 |
| L182P | -210.10 ± 11.34 | -60.55 ± 10.00 | 189.25 ± 14.48 | -20.84 ± 0.87 | -102.24 ± 13.42 |

Table 5. Binding energy and its components including van der Waals, electrostatic, polar solvation and SASA energy for all six ligand-bound systems, calculated over the last 70 ns of simulation time.

residue 237. This suggests that the formation of the α -helix conformation between residues 222 and 237 occurs post 100 ns and is important for enhancing the stability of the protein, as it is present in the final conformation. As the mutations increase plasticity, it is logical to surmise that they might introduce this conformation very early in order to contribute to the plasticity of the protein.

Binding energy analysis. To further analyze the impact of the missense mutations on the protein with respect to ligand binding, the MM-PBSA method was used over the last 70 ns of simulation time with an interval of 100 ps. Various components of binding energy including van der Waals energy, electrostatic energy, polar solvation energy and SASA energy, as well as the binding energy itself, which is the cumulative sum of the aforementioned components, were calculated for all systems except the apo form of TPMT. The average values of these parameters are stated in Table 5. It was observed that apart from L182P, all mutations drastically increased the binding free energy when compared to ligand-bound TPMT with W33G having the most significant impact on binding free energy. These results are in line with prior MDS analyses where most mutations enhanced the plasticity of the protein which has likely hampered ligand binding as evident from the binding free energy values. Furthermore, these results are also in line with the secondary structure analysis and PCA results analysis where it was observed that W33G seemed to impart the most plasticity to the protein–ligand system.

A graph showcasing the residual contribution to binding energy at catalytic residues (residue number 26, 29, 33, 40, 69, 70, 90, 91, 133, 134, 135, 152 and 153) involving in ligand binding is presented in Figure S8. As is evident from the figure, most catalytic residues were unable to contribute to binding energy in all mutations except L182P. Interestingly, L182P shows a nearly threefold increase in binding energy at residue 152 when compared to wild-type ligand-bound TPMT. As shown in Figure S9, in the wild-type ligand-bound TPMT (WT-SAH) system the distance between residue 152 and the ligand is over 7 Å while in the L182P mutant system, the distance between residue 152 and the ligand is just over 3 Å. Furthermore, there are a lot of hydrogen bonds nearby, making the conformations very rigid as opposed to a single one observed at the ligand near residue 152 in the case of the WT-SAH system. This increased proximity and rigidity may cause repulsion and result in a threefold increase in binding energy at residue 152, as observed in the case of L182P. Figure 5 shows the change in ligand conformations over the simulation time for all six ligand-bound systems. It can be clearly seen that the ligand conformations are very similar at the start of the simulation (0 ns) and then from 20 ns onwards, they show variation with respect to the WT-SAH system. Animation 10 provides a 360° view of the change in ligand conformations over the simulation time for all six ligand-bound systems.

Relative deleterious nature of the screened deleterious TPMT missense SNPs. To summarise the present study and visually draw meaningful conclusions, the relative deleterious natures of the five highly deleterious missense SNPs were determined, as shown in Fig. 6. The relative deleterious natures of the five ns-SNPs were calculated through their respective absolute cumulative scores of sequence-based and structure-based tools, cumulative conservation scores at the mutated residues by ConSurf, CDD and all MSA tools, cumulative coding scores provided by both CScape and CScape-somatic (Oncogenic scores), changes in protein binding pockets with respect to volume (cumulative for W33G and L182P, due to the presence of multiple binding pocket changes), the distances to nearest amyloid-forming regions, the distances to nearest PTM sites, and the respective binding energies of the mutant systems calculated using the MM-PBSA method (MDS score) during MDS analysis.

Consider the oncogenic score where the cumulative score of CScape and CScape-somatic were considered. W33G was observed to have the highest cumulative oncogenic score among all mutations. Hence, to calculate the relative oncogenic score of say W78R, the cumulative oncogenic score of W78R (1.806) was divided by the cumulative oncogenic score of W33G (1.516) and multiplied by 100 which resulted in a relative oncogenic score for W78R as 83.942%.

This procedure was followed for all of the aforementioned categories for all mutations in order to unify the scale for all categories for all mutations for better interpretation and visualization. Due to the nature of the procedure, a score of 100 for a mutation in a given category implies that it has the most deleterious effect on TPMT with respect to that category when compared to the other four mutations. For example, W33G has an oncogenic score of 100, which implies that it is the most deleterious missense SNP of TPMT among the five screened mutations with respect to oncogenicity. It is worth noting that for the categories that account for the distance of mutations from their nearest PTM site and nearest amyloid-forming region, a very minor change was incorporated into

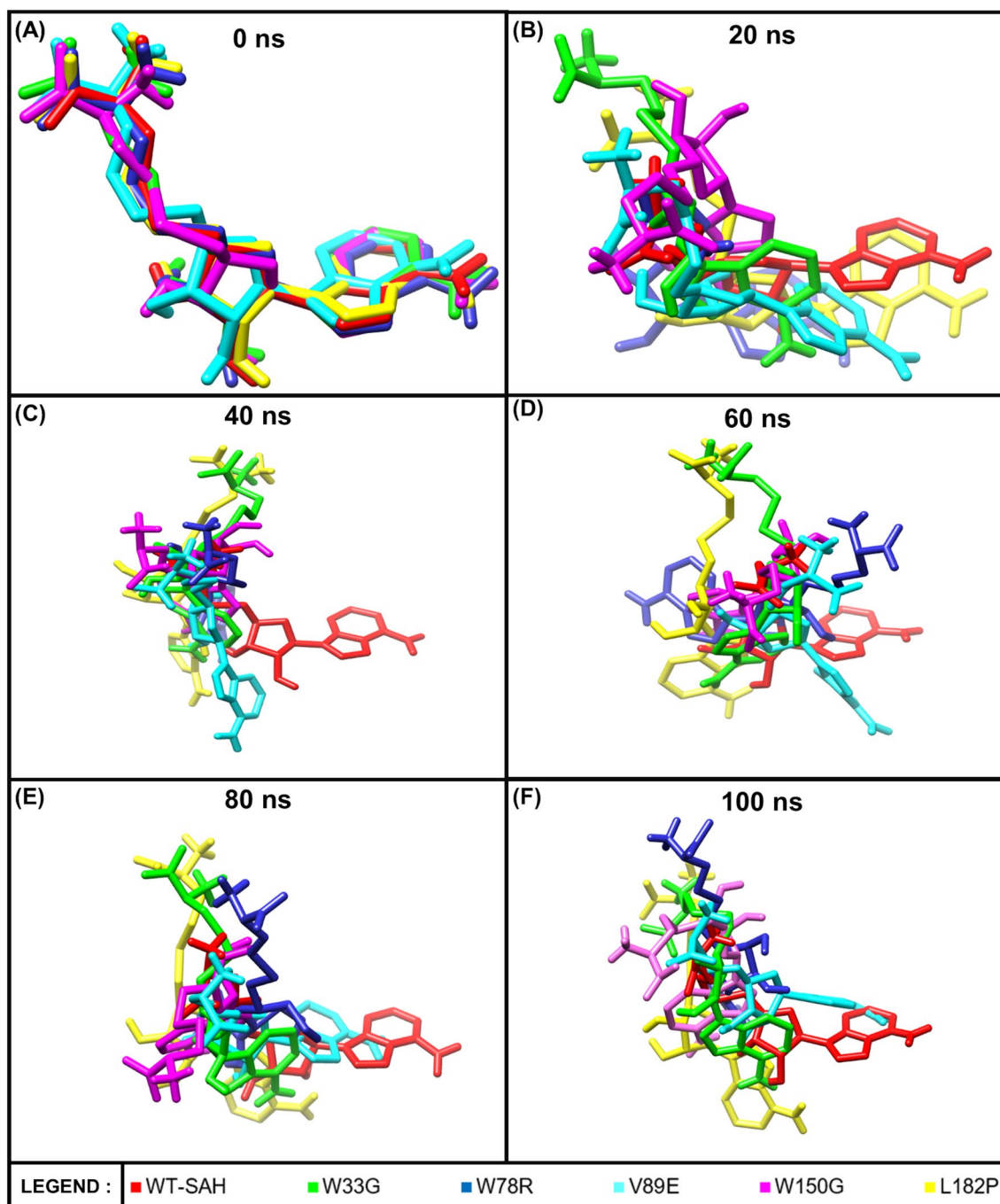


Figure 5. Ligand conformations for all six ligand-bound systems across 0, 20, 40, 60, 80 and 100 ns of simulation time.

the methodology. In the case of those two categories, after following the same aforementioned methodology to get the relative score, it was subtracted from 100 to get the final relative score. This was done because contrary to the other categories, a lower value indicates a closer distance to the nearest PTM site or amyloid-forming region which implies that the given mutation has a more deleterious effect. Finally, a category named “Total” was computed where the relative scores of all categories were added and then using the same procedure used in the other categories, the “Total” category itself was calculated relative to the highest value in the “Total” category (W33G). Thus, from Fig. 6 it can be seen that W33G is the most deleterious missense mutation of TPMT amongst the five deleterious SNPs followed by W150G, V89E, L182P and W78R in the order of decreasing deleterious nature.

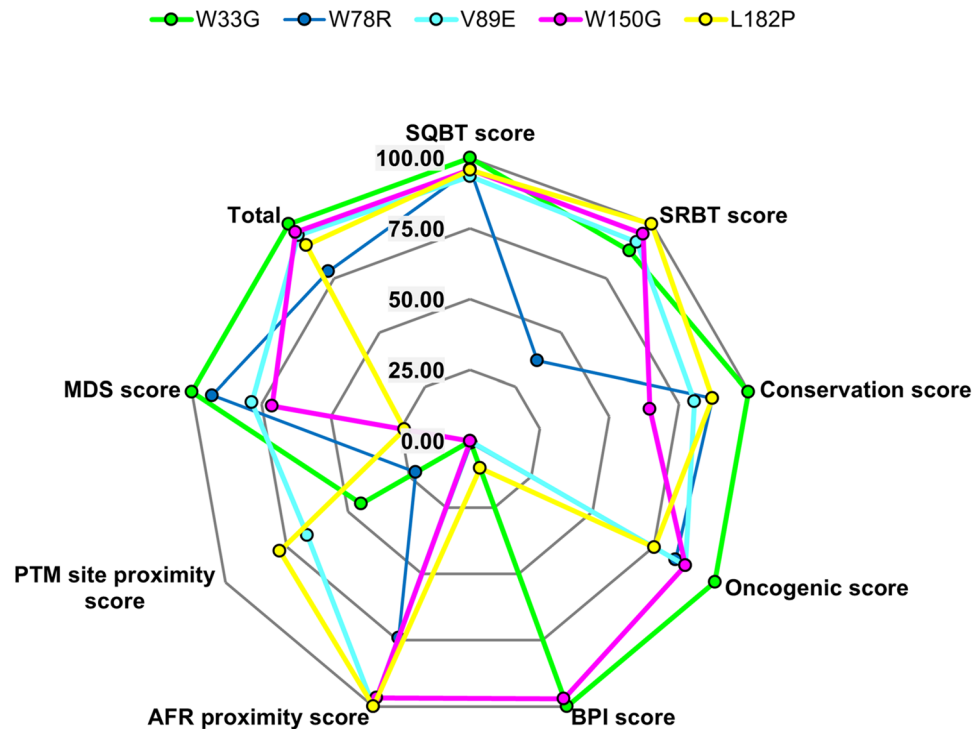


Figure 6. The relative deleterious nature of the five highly deleterious TPMT missense SNPs. A score of 100 for a mutation in a particular category indicates that the given mutation is the most deleterious with respect to that category (SQBT = Sequence-based tools, SRBT = Structure-based tools, BPI score = Binding pocket impact score, AFR = Amyloid forming region, PTM = Post-translational modification, MDS = Molecular dynamics simulations).

Conclusion

The missense SNPs of the TPMT gene have been analyzed in the present study in order to identify novel deleterious SNPs through in silico means, including the validation of the in silico screening pipeline. TPMT was chosen due to the known association of its polymorphisms with the disease progression and treatments of several dangerous diseases including leukemia and IBD (Crohn's disease and ulcerative colitis) (IBD). Since it is a known fact that in silico protocols cannot completely replace in vitro and in vivo experimental protocols which are usually more conclusive in nature, the present study has attempted to validate the screening pipeline by means of re-identifying known deleterious alleles of TPMT through the same. This was followed by the utilisation of a plethora of tools and analyses to gain a more comprehensive understanding of how the SNPs identified in the screening phase might be deleterious in nature. Among all missense SNPs of TPMT reported in the dbSNP database so far, five novel highly deleterious SNPs namely W33G (rs72552741), W78R (rs753277177), V89E (rs1784191846), W150G (rs1447033392) and L182P (rs1386533390) were identified in the present study. These SNPs have a high probability of being associated with the aforementioned diseases that are known to be associated with TPMT polymorphisms. The most deleterious SNP among the five was found to be W33G based on several different analyses. The authors believe that by giving a higher priority to these particular SNPs rather than considering the massive pool of all TPMT SNPs, future in vivo or in vitro research endeavors targeted at TPMT polymorphisms and/or their consequences in relevant disease progressions or treatments will greatly benefit from the present study.

Data availability

The SNP data utilized in this study are openly available in the dbSNP database at <https://www.ncbi.nlm.nih.gov/snp/?term=TPMT>. The STRING database and CDD (Conserved Domain Database), used in the present study are available at <https://string-db.org/> and <https://www.ncbi.nlm.nih.gov/cdd/> respectively. The protein sequence and structure used for TPMT in the present study, obtained from the UniProt database and PDB (Protein Data Bank) respectively are available at <https://www.uniprot.org/uniprotkb/P51580/entry#sequences> and <https://www.rcsb.org/structure/2H11> respectively. All data generated in the present study is present in the article and supplementary information.

Received: 13 January 2022; Accepted: 1 November 2022
Published online: 07 November 2022

References

- Elion, G. B. & Hitchings, G. H. The synthesis of 6-thioguanine. *J. Am. Chem. Soc.* **77**(6), 1676. <https://doi.org/10.1021/ja01611a082> (1955).
- Zakerska-Banaszak, O. *et al.* Cytotoxicity of thiopurine drugs in patients with inflammatory bowel disease. *Toxics* **10**(4), 151. <https://doi.org/10.3390/toxics10040151> (2022).
- Bayoumy, A. B. *et al.* Advances in thiopurine drug delivery: the current state-of-the-art. *Eur. J. Drug Metab. Pharmacokinet.* **46**(6), 743–758. <https://doi.org/10.1007/s13318-021-00716-x> (2021).
- Coulthard, S. A., McGarrity, S., Sahota, K., Berry, P. & Redfern, C. P. F. Three faces of mercaptopurine cytotoxicity in vitro: methylation, nucleotide homeostasis, and deoxythioguanosine in DNA. *Drug Metab. Dispos. Biol. Fate Chem.* **46**(8), 1191–1199. <https://doi.org/10.1124/dmd.118.081844> (2018).
- Brookes, A. J. The essence of SNPs. *Gene* **234**(2), 177–186. [https://doi.org/10.1016/S0378-1119\(99\)00219-X](https://doi.org/10.1016/S0378-1119(99)00219-X) (1999).
- Gebert, M., Jaśkiewicz, M., Moszyńska, A., Collawn, J. F. & Bartoszewski, R. The effects of single nucleotide polymorphisms in cancer RNAi therapies. *Cancers* **12**(11), 3119. <https://doi.org/10.3390/cancers12113119> (2020).
- Shastri, B. S. SNPs: impact on gene function and phenotype. In *Single Nucleotide Polymorphisms* (ed. Komar, Anton A.) 3–22 (Humana Press, Totowa, NJ, 2009). https://doi.org/10.1007/978-1-60327-411-1_1.
- Franca, R., Braidotti, S., Stocco, G. & Decorti, G. Understanding thiopurine methyltransferase polymorphisms for the targeted treatment of hematologic malignancies. *Expert Opin. Drug Metab. Toxicol.* **17**(10), 1187–1198. <https://doi.org/10.1080/17425255.2021.1974398> (2021).
- de CardosoCarvalho, D. *et al.* Association between the TPMT*3C (rs1142345) polymorphism and the risk of death in the treatment of acute lymphoblastic leukemia in children from the Brazilian Amazon Region. *Genes* <https://doi.org/10.3390/genes11101132> (2020).
- Davavala, S. K. *et al.* Prevalence of TPMT polymorphism in Indian patients requiring immunomodulator therapy and its clinical significance. *Indian J. Gastroenterol.* **33**(1), 41–45. <https://doi.org/10.1007/s12664-013-0374-6> (2014).
- Hedayati, M. *et al.* Association of TPMT (rs1800460) gene polymorphism with childhood acute lymphoblastic leukemia in a population from Guilan Iran. *J. Genet. Resour.* **6**(2), 142–147 (2020).
- Khera, S. *et al.* Prevalence of TPMT, ITPA and NUDT15 genetic polymorphisms and their relation to 6MP toxicity in north Indian children with acute lymphoblastic leukemia. *Cancer Chemother. Pharmacol.* **83**(2), 341–348. <https://doi.org/10.1007/s00280-018-3732-3> (2019).
- Zgheib, N. K. *et al.* NUDT15 and TPMT genetic polymorphisms are related to 6-mercaptopurine intolerance in children treated for acute lymphoblastic leukemia at the Children's Cancer Center of Lebanon. *Pediatr. Blood Cancer* **64**(1), 146–150. <https://doi.org/10.1002/psc.26189> (2017).
- Cao, Q., Zhu, Q., Shang, Y., Gao, M. & Si, J. Thiopurine methyltransferase gene polymorphisms in Chinese patients with inflammatory bowel disease. *Digestion* **79**(1), 58–63. <https://doi.org/10.1159/000205268> (2009).
- Colombel, J. F. *et al.* Genotypic analysis of thiopurine S-methyltransferase in patients with Crohn's disease and severe myelosuppression during azathioprine therapy. *Gastroenterology* **118**(6), 1025–1030. [https://doi.org/10.1016/s0016-5085\(00\)70354-4](https://doi.org/10.1016/s0016-5085(00)70354-4) (2000).
- Haglund, S., Lindqvist, M., Almer, S., Peterson, C. & Taipalensuu, J. Pyrosequencing of TPMT alleles in a general Swedish population and in patients with inflammatory bowel disease. *Clin. Chem.* **50**(2), 288–295. <https://doi.org/10.1373/clinchem.2003.023846> (2004).
- Lennard, L. TPMT in the treatment of Crohn's disease with azathioprine. *Gut* **51**(2), 143–146. <https://doi.org/10.1136/gut.51.2.143> (2002).
- Steponaitiene, R. *et al.* TPMT and ITPA genetic variants in Lithuanian inflammatory bowel disease patients: Prevalence and azathioprine-related side effects. *Adv. Med. Sci.* **61**(1), 135–140. <https://doi.org/10.1016/j.advms.2015.09.008> (2016).
- Zalizko, P. *et al.* Thiopurine S-methyltransferase genetic polymorphisms in adult patients with inflammatory bowel diseases in the Latvian population. *Ther. Adv. Gastroenterol.* **13**, 1756284820937426. <https://doi.org/10.1177/1756284820937426> (2020).
- Yan Ping Heidi, I. *et al.* One amino acid makes a difference—Characterization of a new TPMT allele and the influence of SAM on TPMT stability. *Sci. Rep.* **7**(1), 46428. <https://doi.org/10.1038/srep46428> (2017).
- Katarata, P. & Kuntal, H. TPMT polymorphism: when shield becomes weakness. *Interdis. Sci. Comput. Life Sci.* **8**(2), 150–155. <https://doi.org/10.1007/s12539-015-0111-1> (2016).
- Hasnain, M. J. U. *et al.* Computational analysis of functional single nucleotide polymorphisms associated with SLC26A4 gene. *PLoS One* **15**(1), e0225368. <https://doi.org/10.1371/journal.pone.0225368> (2020).
- Saxena, S. *et al.* In-silico analysis of deleterious single nucleotide polymorphisms of PNMT gene. *Mol. Simul.* <https://doi.org/10.1080/08927022.2022.2094922> (2022).
- Hossain, M. S., Roy, A. S. & Islam, M. S. In silico analysis predicting effects of deleterious SNPs of human RASSF5 gene on its structure and functions. *Sci. Rep.* **10**(1), 14542. <https://doi.org/10.1038/s41598-020-71457-1> (2020).
- Sinha, S. & Wang, S. M. Classification of VUS and unclassified variants in BRCA1 BRCT repeats by molecular dynamics simulation. *Comput. Struct. Biotechnol. J.* **18**, 723–736. <https://doi.org/10.1016/j.csbj.2020.03.013> (2020).
- Soltani, I. *et al.* Comprehensive in-silico analysis of damage associated SNPs in hOCT1 affecting Imatinib response in chronic myeloid leukemia. *Genomics* **113**(1), 755–766. <https://doi.org/10.1016/j.ygeno.2020.10.007> (2021).
- Stalin, A. *et al.* Computational analysis of single nucleotide polymorphisms (SNPs) in PPAR gamma associated with obesity, diabetes and cancer. *J. Biomol. Struct. Dyn.* <https://doi.org/10.1080/07391102.2020.1835724> (2020).
- Yadav, A. K. & Singh, T. R. Novel structural and functional impact of damaging single nucleotide polymorphisms (SNPs) on human SMYD2 protein using computational approaches. *Meta Gene* **28**, 100871. <https://doi.org/10.1016/j.mgene.2021.100871> (2021).
- Fazel-Najafabadi, E., Vahdat, E., Fattahpour, A. S. & Sedghi, M. Structural and functional impact of missense mutations in TPMT: an integrated computational approach. *Comput. Biol. Chem.* **59**(48), 55. <https://doi.org/10.1016/j.compbiolchem.2015.09.004> (2015).
- Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**(6), 1129–1143. <https://doi.org/10.1016/j.neuron.2018.08.011> (2018).
- Geng, H., Chen, F., Ye, J. & Jiang, F. Applications of molecular dynamics simulation in structure prediction of peptides and proteins. *Comput. Struct. Biotechnol. J.* **17**, 1162–1170. <https://doi.org/10.1016/j.csbj.2019.07.010> (2019).
- Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**(1), 308–311. <https://doi.org/10.1093/nar/29.1.308> (2001).
- Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100> (2021).
- Hong, W. *et al.* Structural basis of allele variation of human thiopurine-S-methyltransferase. *Proteins Struct. Funct. Bioinform.* **67**(1), 198–208. <https://doi.org/10.1002/prot.21272> (2007).
- Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**(16), 2745–2747. <https://doi.org/10.1093/bioinformatics/btv195> (2015).
- Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**(9), 2129–2141. <https://doi.org/10.1101/gr.772403> (2003).
- Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**(4), 248–249 (2010).

38. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**(17), e118–e118. <https://doi.org/10.1093/nar/gkr407> (2011).
39. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**(W1), W452–W457. <https://doi.org/10.1093/nar/gks539> (2012).
40. Capriotti, E., Altman, R. B. & Bromberg, Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genom.* **14**(3), S2. <https://doi.org/10.1186/1471-2164-14-S3-S2> (2013).
41. Bendall, J. *et al.* PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* **10**(1), e1003440. <https://doi.org/10.1371/journal.pcbi.1003440> (2014).
42. Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. E. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* **426**(14), 2692–2701. <https://doi.org/10.1016/j.jmb.2014.04.026> (2014).
43. López-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M. & Gelpi, J. L. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res.* **45**(W1), W222–W228. <https://doi.org/10.1093/nar/gkx313> (2017).
44. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genom.* **16**(8), S1. <https://doi.org/10.1186/1471-2164-16-S8-S1> (2015).
45. 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 (Oxford, England)* **22**(22), 2729–2734. <https://doi.org/10.1093/bioinformatics/btl423> (2006).
46. Capriotti, E. *et al.* WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genom.* **14**(3), S6. <https://doi.org/10.1186/1471-2164-14-S3-S6> (2013).
47. Parthiban, V., Gromiha, M. M. & Schomburg, D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **34**(suppl_2), W239–W242. <https://doi.org/10.1093/nar/gkl190> (2006).
48. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* **42**(W1), W314–W319. <https://doi.org/10.1093/nar/gku411> (2014).
49. Capriotti, E., Fariselli, P., Rossi, I. & Casadio, R. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinform.* **9**(2), S6. <https://doi.org/10.1186/1471-2105-9-S2-S6> (2008).
50. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Bioinform.* **62**(4), 1125–1132. <https://doi.org/10.1002/prot.20810> (2005).
51. Savojardo, C., Fariselli, P., Martelli, P. L. & Casadio, R. INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* **32**(16), 2542–2544. <https://doi.org/10.1093/bioinformatics/btw192> (2016).
52. Dehouck, Y., Kwasigroch, J. M., Gilis, D. & Rooman, M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinform.* **12**, 151. <https://doi.org/10.1186/1471-2105-12-151> (2011).
53. Pucci, F., Kwasigroch, J. M. & Rooman, M. Protein thermal stability engineering using HoTMuSiC. In *Structural Bioinformatics: Methods and Protocols* (ed. Gáspári, Z.) 59–73 (Springer US, New York, NY, 2020). https://doi.org/10.1007/978-1-0716-0270-6_5.
54. Ancien, F., Pucci, F., Godfroid, M. & Rooman, M. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Sci. Rep.* **8**(1), 4480. <https://doi.org/10.1038/s41598-018-22531-2> (2018).
55. Wennerstrand, P., Blissing, A. & Mårtensson, L.-G. In vitro protein stability of two naturally occurring thiopurine s-methyltransferase variants: biophysical characterization of TPMT*6 and TPMT*8. *ACS Omega* **2**(8), 4991–4999. <https://doi.org/10.1021/acsomega.7b00801> (2017).
56. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**(W1), W344–W350. <https://doi.org/10.1093/nar/gkw408> (2016).
57. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**(W1), W636–W641. <https://doi.org/10.1093/nar/gkz268> (2019).
58. Shennan, L. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**(D1), D265–D268. <https://doi.org/10.1093/nar/gkz991> (2020).
59. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**(D1), D607–D613. <https://doi.org/10.1093/nar/gky1131> (2019).
60. Rogers, M. F., Shihab, H. A., Gaunt, T. R. & Campbell, C. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci. Rep.* **7**(1), 11597. <https://doi.org/10.1038/s41598-017-11746-4> (2017).
61. Rogers, M. F., Gaunt, T. R. & Campbell, C. CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. *Bioinformatics* **36**(12), 3637–3644. <https://doi.org/10.1093/bioinformatics/btaa242> (2020).
62. Capriotti, E. & Altman, R. B. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**(4), 310–317. <https://doi.org/10.1016/j.ygeno.2011.06.010> (2011).
63. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutat.* **34**(1), 57–65. <https://doi.org/10.1002/humu.22225> (2012).
64. Schrödinger LLC. (2015). *The PyMOL Molecular Graphics System, Version-1.8*.
65. Venselaar, H., te Beek, T. A. H., Kuipers, R. K. P., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinform.* **11**(1), 548. <https://doi.org/10.1186/1471-2105-11-548> (2010).
66. Klausen, M. S. *et al.* NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins: Struct. Funct. Bioinform.* **87**(6), 520–527. <https://doi.org/10.1002/prot.25674> (2019).
67. Tian, W., Chen, C. & Liang, J. CASTp 3.0: computed atlas of surface topography of proteins and beyond. *Biophys. J.* **114**(3), 50a. <https://doi.org/10.1016/j.bpj.2017.11.325> (2018).
68. Tsolis, A. C., Papandreou, N. C., Ionomidou, V. A. & Hamodrakas, S. J. A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins. *PLoS One* **8**(1), e54175. <https://doi.org/10.1371/journal.pone.0054175> (2013).
69. Wang, C. *et al.* GPS 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. *Genom. Proteom. Bioinform.* **18**(1), 72–80. <https://doi.org/10.1016/j.gpb.2020.01.001> (2020).
70. Lindahl, E., Abraham, M. J., Hess, B., & van der Spoel, D. (2019). GROMACS 2019.4 Source code. <https://doi.org/10.5281/zenodo.3460414>
71. Kumari, R., Kumar, R. & Lynn, A. g_mmpbsa—a GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.* **54**(7), 1951–1962. <https://doi.org/10.1021/ci500020m> (2014).
72. Coelho, T. *et al.* Genes implicated in thiopurine-induced toxicity: Comparing TPMT enzyme activity with clinical phenotype and exome data in a paediatric IBD cohort. *Sci. Rep.* **6**(1), 34658. <https://doi.org/10.1038/srep34658> (2016).
73. Feng, Q. *et al.* Thiopurine S-methyltransferase pharmacogenetics: functional characterization of a novel rapidly degraded variant allele. *Biochem. Pharmacol.* **79**(7), 1053–1061. <https://doi.org/10.1016/j.bcp.2009.11.016> (2010).
74. Garat, A. *et al.* Characterisation of novel defective thiopurine S-methyltransferase allelic variants. *Biochem. Pharmacol.* **76**(3), 404–415. <https://doi.org/10.1016/j.bcp.2008.05.009> (2008).
75. Hamdan-Khalil, R. *et al.* In vitro characterization of four novel non-functional variants of the thiopurine S-methyltransferase. *Biochem. Biophys. Res. Commun.* **309**(4), 1005–1010. <https://doi.org/10.1016/j.bbrc.2003.08.103> (2003).
76. Hamdan-Khalil, R. *et al.* Identification and functional analysis of two rare allelic variants of the thiopurine S-methyltransferase gene, TPMT*16 and TPMT*19. *Biochem. Pharmacol.* **69**(3), 525–529. <https://doi.org/10.1016/j.bcp.2004.10.011> (2005).

77. Kham, S. K. Y., Soh, C. K., Aw, D. C. W. & Yeoh, A. E. J. TPMT*26 (208F→L), a novel mutation detected in a Chinese. *Br. J. Clin. Pharmacol.* **68**(1), 120–123. <https://doi.org/10.1111/j.1365-2125.2009.03405.x> (2009).
78. Kim, H.-Y. *et al.* Complete sequence-based screening of TPMT variants in the Korean population. *Pharmacogenet. Genom.* **25**(3), 143–146. <https://doi.org/10.1097/FPC.000000000000117> (2015).
79. Krynetski, E. Y. *et al.* A single point mutation leading to loss of catalytic activity in human thiopurine S-methyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **92**(4), 949–953. <https://doi.org/10.1073/pnas.92.4.949> (1995).
80. Lennard, L., Cartwright, C. S., Wade, R., Richards, S. M. & Vora, A. Thiopurine methyltransferase genotype-phenotype discordance and thiopurine active metabolite formation in childhood acute lymphoblastic leukaemia. *Br. J. Clin. Pharmacol.* **76**(1), 125–136. <https://doi.org/10.1111/bcp.12066> (2013).
81. Lindqvist, M. *et al.* Identification of two novel sequence variants affecting thiopurine methyltransferase enzyme activity. *Pharmacogenetics* **14**(4), 261–265. <https://doi.org/10.1097/00008571-200404000-00006> (2004).
82. Lindqvist, M. *et al.* Explaining TPMT genotype/phenotype discrepancy by haplotyping of TPMT*3A and identification of a novel sequence variant, TPMT*23. *Pharmacogenet. Genom.* **17**(10), 891–895. <https://doi.org/10.1097/FPC.0b013e3282ef642b> (2007).
83. Otterness, D. *et al.* Human thiopurine methyltransferase pharmacogenetics: gene sequence polymorphisms. *Clinical Pharmacol. Therapeut.* **62**(1), 60–73. [https://doi.org/10.1016/S0009-9236\(97\)90152-1](https://doi.org/10.1016/S0009-9236(97)90152-1) (1997).
84. Salavaggione, O. E., Wang, L., Wiepert, M., Yee, V. C. & Weinshilboum, R. M. Thiopurine S-methyltransferase pharmacogenetics: variant allele functional and comparative genomics. *Pharmacogenet. Genom.* **15**(11), 801–815. <https://doi.org/10.1097/01.fpc.0000174788.69991.6b> (2005).
85. Schaeffeler, E. *et al.* A novel TPMT missense mutation associated with TPMT deficiency in a 5-year-old boy with ALL. *Leukemia* **17**(7), 1422–1424. <https://doi.org/10.1038/sj.leu.2402981> (2003).
86. Schaeffeler, E. *et al.* Comprehensive analysis of thiopurine S-methyltransferase phenotype-genotype correlation in a large population of German-Caucasians and identification of novel TPMT variants. *Pharmacogenetics* **14**(7), 407–417. <https://doi.org/10.1097/01.fpc.0000114745.08559.db> (2004).
87. Schaeffeler, E., Eichelbaum, M., Reinisch, W., Zanger, U. M. & Schwab, M. Three novel thiopurine S-methyltransferase allelic variants (TPMT*20, *21, *22) - association with decreased enzyme function. *Hum. Mutat.* **27**(9), 976. <https://doi.org/10.1002/humu.9450> (2006).
88. Spire-Vayron de la Moureyre, C. *et al.* Detection of known and new mutations in the thiopurine S-methyltransferase gene by single-strand conformation polymorphism analysis. *Human Mutat.* **12**(3), 177–185 (1998).
89. Szumlanski, C. *et al.* Thiopurine methyltransferase pharmacogenetics: human gene cloning and characterization of a common polymorphism. *DNA Cell Biol.* **15**(1), 17–30. <https://doi.org/10.1089/dna.1996.15.17> (1996).
90. Tai, H., Krynetski, E., Schuetz, E., Yanishevski, Y. & Evans, W. Enhanced proteolysis of thiopurine S-methyltransferase (TPMT) encoded by mutant alleles in humans (TPMT*3A, TPMT*2): mechanisms for the genetic polymorphism of TPMT activity. *Proc. Natl. Acad. Sci. U.S.A.* **94**(12), 6444–6449 (1997).
91. Ujiiie, S., Sasaki, T., Mizugaki, M., Ishikawa, M. & Hiratsuka, M. Functional characterization of 23 allelic variants of thiopurine S-methyltransferase gene (TPMT*2 - *24). *Pharmacogenet. Genom.* **18**(10), 887–893. <https://doi.org/10.1097/FPC.0b013e3283097328> (2008).
92. Zimdahl Kahlin, A. *et al.* Comprehensive study of thiopurine methyltransferase genotype, phenotype, and genotype-phenotype discrepancies in Sweden. *Biochem. Pharmacol.* **164**, 263–272. <https://doi.org/10.1016/j.bcp.2019.04.020> (2019).
93. Saxena, S. *et al.* Structural and functional analysis of disease-associated mutations in GOT1 gene: An in silico study. *Comput. Biol. Med.* **136**, 104695. <https://doi.org/10.1016/j.compbiomed.2021.104695> (2021).
94. Porollo, A. & Meller, J. Versatile annotation and publication quality visualisation of protein complexes using POLYVIEW-3D. *BMC Bioinform.* **8**(1), 316. <https://doi.org/10.1186/1471-2105-8-316> (2007).
95. Rambaran, R. N. & Serpell, L. C. Amyloid fibrils: abnormal protein assembly. *Prion* **2**(3), 112–117. <https://doi.org/10.4161/pri.2.3.7488> (2008).
96. Shukla, R., Shukla, H. & Tripathi, T. Activity loss by H46A mutation in Mycobacterium tuberculosis isocitrate lyase is due to decrease in structural plasticity and collective motions of the active site. *Tuberculosis* **108**, 143–150. <https://doi.org/10.1016/j.tube.2017.11.013> (2018).
97. Rutherford, K. & Daggett, V. Four human thiopurine s-methyltransferase alleles severely affect protein structure and dynamics. *J. Mol. Biol.* **379**(4), 803–814. <https://doi.org/10.1016/j.jmb.2008.04.032> (2008).
98. Pettersen, E. F. *et al.* UCSF Chimera—a visualisation system for exploratory research and analysis. *J. Comput. Chem.* **25**(13), 1605–1612. <https://doi.org/10.1002/jcc.20084> (2004).

Acknowledgements

The authors thank Management, Ramaiah Institute of Technology, and the Jaypee University of Information Technology for providing the computational facilities necessary to perform the research studies. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

S.S.: Data curation, Investigation, Methodology, Roles/Writing—original draft; T.P.K.M.: Conceptualization, Supervision, Writing—review & editing Chandrashekhar; C.R.: Data curation, Investigation, Methodology; L.S.P.: Data curation, Investigation, Methodology, A.A.: Data curation, Investigation, Methodology, R.S.: Data curation, Investigation, Methodology, Roles/Writing—original draft; A.K.Y.: Data curation, Investigation, Methodology, Roles/Writing—original draft; T.R.S.: Conceptualization, Supervision, Writing—review & editing; M.S.: Data curation, Investigation, Methodology, Roles/Writing—original draft; R.A.: Conceptualization, Supervision, Writing—review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-23488-z>.

Correspondence and requests for materials should be addressed to T.P.K.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022