UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Biosciences, Fisheries and Economics
Department of Arctic and Marine Biology

In Silico Screening for inhibitors against Apicoplast Phosphate Translocator from *Toxoplama gondii*

Muhammad Shamsuzzaman

Bio-3950, Master’s thesis in Biology
May 2019

Acknowledgement

The master thesis was written at the Microorganisms and Plant research group, Department of Arctic and Marine Biology, Faculty of Biosciences, Fisheries and Economics at Uit The Arctic University of Norway in collaboration with Professor Ingebrigt Sylte, Medical Pharmacology and Toxicology research group, Department of Medical Biology, Faculty of Health Sciences, UiT Arctic University of Norway.

First of all, I am grateful to the Almighty Allah for giving me strength and patience to finish the task.

In this entire journey of my thesis, there are so many people to give thanks to. My two supervisors Professor Karsten Bruno Fischer and Professor Ingebrigt Sylte will come first in the list, who were always kind and helpful towards me. In every step of this work, starting from the study design to the end of writing the thesis, they were always there for me.

Then, I would like to give a special thanks to Linn Evenseth for her assistance in understanding the methods in the laboratory. As well as being helpful, she was also an inspiring character for me.

I am also grateful to Imim Wushur, Mari Gabrielsen and Kurt Kristiansen for their help throughout the work. I would also like to mention Fatema Rahman and Kamil Piotr Wiejaczka for their inspiration and suggestions.

Finally, my gratitude goes to my parents and my brother, who were always beside me.

ABSTRACT

Apicomplexa parasites, including *Toxoplasma gondii* and *Plasmodium falciparum*, contain a secondary endosymbiosis-derived plastid like organ, called apicoplast, which is an anabolic hub. This apicoplast is fueled by phosphate translocator (APT), which transport phosphorylated sugar molecules in exchange of inorganic phosphate. Disruption of APT in *T. gondii* was found to be lethal for parasite. Beside this, its's plastidic nature and location in apicoplast, made it an ideal drug target.

In this study two homology models of TgAPT were used for predicting putative inhibitors against this protein by combining ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) approaches. Before doing the actual screening, a homology model of another APT, called PfoTPT from *P. falciparum* was generated to compare the binding pocket and the binding of known ligands by docking. The binding pocket of TgAPT was also compared with other plastidic phosphate translocator classes. The comparison revealed that there was only one amino acid different between two APTs, but several differences between the APTs and pPT classes and these differences are assumed to contribute to differences in substrate recognition and binding. Then, known substrates, non-substrates and inhibitors were docked in two TgAPT models and PfoTPT model. The non-substrates are those which are not usually transported, nor they inhibit the transport process. The PfoTPT model did not show good result in terms of scoring and rank ordering of compounds. Of the two TgAPT models, TgAPT_5y79 showed comparatively better result, so induced fit docking (IFD) was done in this model with 3-phosphoglyceric acid (3-PGA), phosphoenol pyruvate (PEP), pyridoxal-5-phosphate (PLP) and 2,4,6- trinitrobenzene sulfonate (TNBS) for generating better conformation. Then one of the poses generated with 3-PGA IFD was selected for the SBVS approach.

In VS approach, analogs of substrates and inhibitors were retrieved from PubChem database and docked into the IFD generated pose. From this docking, 318 compounds were sorted from different analog groups and compounds of each group were clustered by hierarchical clustering. Finally, 29 compounds were predicted as putative inhibitor of TgAPT based on the docking score and their interaction with the protein. These compounds will be tested in vitro for the inhibition potential.

Table of Content

List of Table

List of Figure

List of abbreviation

2D	Two dimensional
3-PGA	3-phosphogyceric acid
ATP	Adenosine triphosphate
APT	Apicoplast Phosphate Translocator
B. bovis	*Babesia bovis*
BLAST	Basic Local Alignment Search Tool
DHAP	Dihydroxy Acetone phosphate
DMAPP	Dimethylallyl diphosphate
DOXP	1-deoxy-D-xylulose-5-phosphate
Dxr	DOXP reductoisomerase
FASI	Fatty acid synthase I
FASII	Fatty Acid Synthase II
Fos	Fosmidomycin
Frc-6-p	Fructose -6- phosphate
Gly-3-P	Glyceraldehyde 3-phosphate
GPT	Glucose-6-Phosphate/phosphate translocator
G. Sulphuraria	*Galderia sulphuraria*
GsTPT2	*Galderi sulphuraria* triose phosphate translocator 2
H-bond	Hydrogen bond
HTS	High throughput screening
IPP	Isopentenyl diphosphate
LBVS	Ligand based virtual screening
MEP	Methylerythritriol phosphate
mM	Milimolar
MEV	Mevalonate
MM	Molecular mechanics
NADPH	Nicotinamide adenine dineucleotide phosphate
NMR	Nuclear magnetic resonance
PDB	Protein data bank
PEP	Phosphoenolpyruvate
PfoTPT	*Plamodium falciparum* outer triose phosphate translocator

P. falciparum	*Plasmodium falciparum*
PLP	pyridoxal 5’ phosphate
PP	Pyrophosphate
PPT	Phosphoenolpyruvate/phosphate translocator
pPT	Plastidic Phosphate Translocator
QM	Quantum mechanics
4-SBD	4-Sulfobenzenediazonium
SBVS	Structure based virtual screening
T. gondii	*Toxoplasma gondii*
TgAPT	*Toxoplasma gondii* apicoplast phosphate translocator
TPT	Triosephosphate/phosphate translocator
TNBS	2,4,6-trinitrobenzenesulfonate
VS	Virtual screening

1 INTRODUCTION

1.1 Apicomplexa parasites

Apicomplexa is a large phylum that consists of single-celled obligate intracellular parasitic protozoans. The defining characteristic of the members of this phylum is a group of organelles found in one end of the cell called apical complex. This complex, which gives the name apicomplexa, plays a crucial role during invasion of host cell (Katris et al., 2014).

This phylum includes a large spectrum of species, more than 6000 of which are named and even more than 6000 that are unnamed (Adl et al., 2007). Because of this versatility there is a wide range of hosts of this protist group including invertebrates, reptiles, amphibians, and mammals including humans (Duszynski, Wilson, J. Upton, & D. Levine, 1999). Two of the most important species are *Plasmodium falciparum (P. falciparum)* and *Toxoplasma gondii (T. gondii),* the causative agent of malaria and toxoplasmosis (Seeber & Steinfelder, 2016). The protein of interest in this study is from *T. gondii.*

1.2 Toxoplasma gondii

T. gondii is an opportunistic pathogen that is capable of infecting warm-blooded animals (Innes, 2010). In a statistic it was seen that the infection rate of *T. gondii* in the world population is up to 50% without showing any symptoms (Flegr, Prandota, Sovičková, & Israili, 2014). Even in Norway this parasite is widespread, especially in cats and sheep (Institute, 2016). Although in humans this infection is not apparently harmful, but chronic infection for lifetime can occur. There is also a chance of formation of cysts in host’s brain, skeletal muscle, heart and other vital organs. Additionally, *T. gondii* infection can lead to retinitis retinae, encephalitis and even death in immunodeficient patient (Jensen et al., 2015).

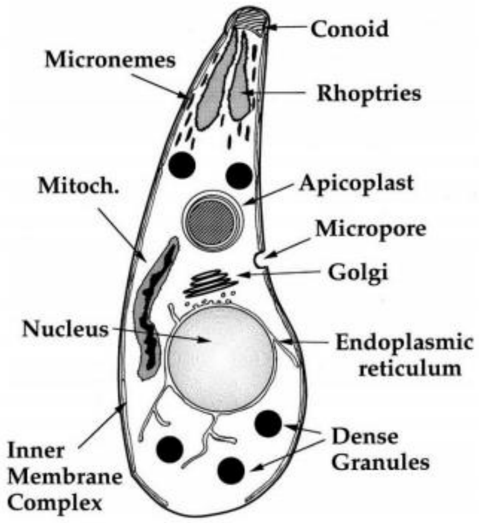


Figure 1. 1: Different organelles of *T. gondii* (Black & Boothroyd, 2000)

T. gondii is also considered as a model organism for studying Apicomplexan biology because of the well-established methodologies to study this organism (Kim & Weiss, 2004). The studies have shown some unique characteristics and organelles in more detail. For example, the apicoplast was identified as a whole organelle by *in-situ* hybridization studies in *T. gondii*, even though the presence was noticed before (McFadden, Reith, Munholland, & Lang-Unnasch, 1996). Later, this organelle got more attention for being the site of metabolic pathways different from the vertebrate hosts and these pathways can be a potential target for new chemotherapeutics (Seeber, Feagin, & Parsons, 2014).

1.3 The apicoplast

The apicoplast is a vestigial plastid found in species of Apicomplexa. It has evolved by secondary endosymbiosis, which is indicated by the presence of three or four membranes surrounding it (Gould, Waller, & McFadden, 2008). These membranes represent their origin from different sources, for example the outer membrane is similar to the host endomembrane system, the second membrane resembles the plasma membrane of the second endosymbiont and the third and fourth correspond to the envelope membranes of the primary plastid (Roos, Kissinger, Fast, & Keeling, 2001), 2001). There is now clear-cut evidence suggesting that the second endosymbiont was a red alga (Liting Lim & McFadden, 2010). In contrast to the photosynthetic algal plastid, the apicoplast is non-photosynthetic. So, the question comes why are the parasites investing energy on maintaining this organelle?

To answer this question, the function of plastids not involved in photosynthesis was looked at. It turned out that non-photosynthetic plastids are involved in the biosynthesis of various metabolites and from this, the conclusion was drawn that apicoplasts work in a similar manner in the Apicomplexa parasites. The theory behind that was that earlier in the symbiosis event, when the endosymbiont still had full photosynthetic capacity, the host started to make itself dependent on the symbiont for various metabolites which they got without or by the expense of little energy. Another probable reason was that, by this manner they could manage to accumulate biosynthesis redundancy. So, later despite the loss of photosynthetic capacity, the hosts are maintaining this organelle for the metabolites produced in the apicoplasts (Striepen, 2011). Now, the next question comes, what is the actual function of apicoplast?

Initially, this question was not answered by biochemical approaches, but rather by bioinformatic and genomic approaches. Genome sequence of many species including *Plasmodium, Toxoplasma and Cryptosporidium* have fueled this type of studies (Striepen, 2011). From these studies, about 500 apicoplast proteins were predicted (Ralph et al., 2004) and three major anabolic pathways (fatty acid synthesis, isoprenoid synthesis and part of the haem synthesis pathway), usually found in chloroplasts, were also found in apicoplasts of *Plasmodium* and *Toxoplasma* (Seeber & Soldati-Favre, 2010). Interestingly, there are differences in metabolic pathways within the species of *Apicomplexa* (Figure 1.2). This indicates that the apicoplast function is not rigid, rather the metabolites acquisition by the host from the apicoplast depend on the specific tissue or cell the parasites occupy (Striepen, 2011)

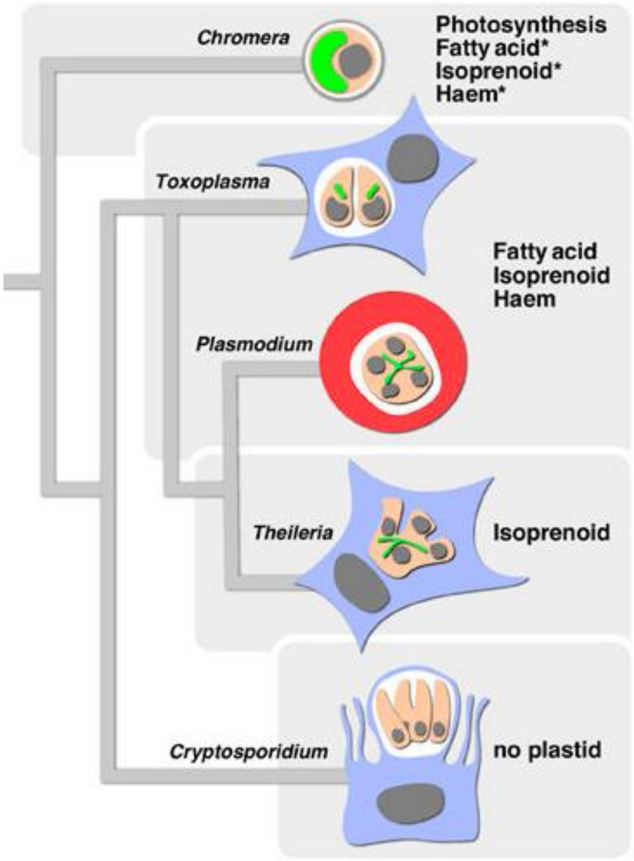


Figure 1. 2: **Evolutionary relationship and adoption of biosynthetic pathways among different members of Apicomplexa (Striepen, 2011)**

1.4 Apicoplast metabolism and potential drug targets

Anabolic pathways residing in apicoplasts are of divergent evolutionary origin from those in human cells, so have different biochemical mechanisms. These differences are making molecules involved in those pathways potential drug targets. An overview of the apicoplast metabolism is given in figure 1.3 and the anabolic pathways are discussed below:

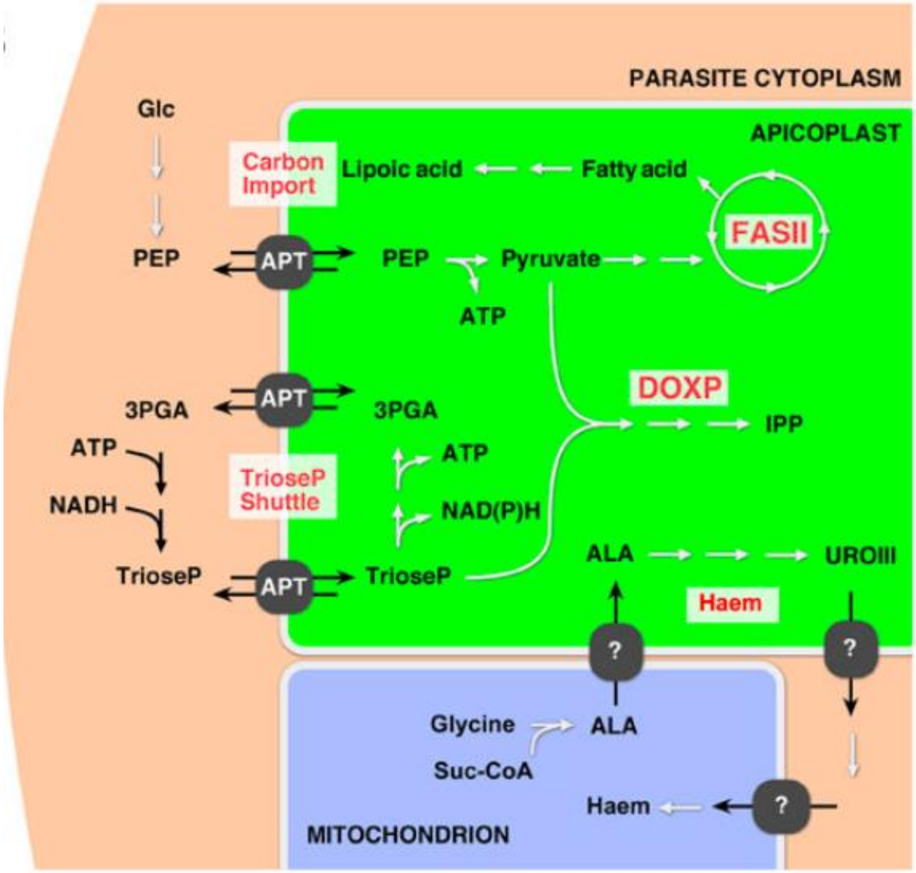


Figure 1. 3: **Overview of apicoplast metabolism. Three pathways (FASII, DOXP and haem) are shown here. FasII and DOXP directly depend upon the imported sugars from the cytoplasm. The haem synthesis is distributed in apicoplasts and mitochondria.** ALA, aminolevulinic acid; Glc, glucose; PEP, phosphoenolpyruvate; suc-CoA, succinyl-CoA; UROIII, uroporphyrinogen-III. (Striepen, 2011)

1.4.1 Fatty Acid Biosynthesis

Fatty acids are one of the fundamental components in any living cell because of their role as membrane building blocks, energy storage molecules, precursors for second messenger and co-factors. In comparison to other organisms, Apicomplexans are much more in need of fatty acids because of their additional membrane-bound organelles like micronemes, rhoptries, dense granules, pellicular complex, the apicoplast and the growing parasitophorous vacuole membrane beside the regular organelles like nucleus, endoplasmic reticulum, golgi body etc. (Bisanz et al., 2006; Charron & Sibley, 2002; Coppens & Vielemeyer, 2005). Fatty acids are generated in two different ways by living organisms. The FASI pathway, which is used by eukaryotic cells, combine all enzymatic steps in one multifunctional protein and produce Palmitate (C_{16}) as end product. On the other hand, the FASII pathway, where individual steps are carried out by separate protein entities, generates C_8 and longer fatty acids (Seeber & Soldati-Favre, 2010). Apicoplasts use the second pathway. Interestingly, *Theileria sp.* and *B. bovis* lack this machinery for fatty acid synthesis and they acquire fatty acids from the host (Seeber & Soldati-Favre, 2010).

Initially, molecules involved in this pathway were thought to be potential targets for drugs because of its essence for the survival of parasites as proved in *T. gondii* (Mazumdar, H Wilson, Masek, A Hunter, & Striepen, 2006). The FASII enzyme FabI inhibitor Triclosan was tested and found to be efficient against *Plasmodium* and *Toxoplasma* (Striepen, 2011). But in other studies, it was found that Triclosan is also efficient in a FabI mutant as well as in *Theileria* and *B. bovis* (Seeber & Soldati-Favre, 2010), which indicates Triclosan off-target activity and hence negates the possibility of using Triclosan as drug against these parasites.

1.4.2 Isoprenoid Biosynthesis

Isoprenoids are a diverse group of nuclear compounds with more than 23000 known structures (Holstein & Hohl, 2004). The diversity represents their diverse roles in biological activities, such as cell signaling, modification of proteins and tRNAS and synthesis of Ubiquinone (Seeber & Soldati-Favre, 2010). The starting compounds of this pathway are isopentenyl diphosphate and its isomer dimethylallyl diphosphate, which can be achieved in two ways, which are the 1-deoxy-D-xylulose-5-phosphate (DOXP) pathway, also known as methylerythritriol phosphate (MEP) pathway, and the mevalonate pathway (MEV) (Eisenreich, Bacher, Arigoni, & Rohdich, 2004; Lichtenthaler, 1999; Rohmer, 1999). The former one is generally used by eubacteria and plants and the latter one is used by archaebacteria and eukaryotes. Interestingly, the majority of plants and a few bacteria possess both pathways (Kirby & Keasling, 2009; Rohmer, 1999). Apicomplexans only possess the entire set of genes coding for the bacterial DOXP pathway (Clastre et al., 2007; Grauvogel, Reece, Brinkmann, & Petersen, 2007). As these genes are not found in human cells, the enzymes of this pathway are potential drug targets. From this idea the antibiotic fosmidomycin (Fos), which is a DOXP reductoisomerase (Dxr) inhibitor, was tested against *P. falciparum* and found to inhibit the growth of the malaria parasite in blood stages (Jomaa et al., 1999; Oyakhirome et al., 2007). But this compound was found to not inhibit the growth of *T. gondii* despite having structural and binding site similarities of TgDxr with PfDxr.

One of the reasons might be the poor uptake of the drug by *T. gondii* infected cell (Seeber & Soldati-Favre, 2010).

1.4.3 Haem biosynthesis

Haem is well known for its role in binding O_2 in hemoglobin as well as co-factor for several enzymes. Haem biosynthesis in apicomplexans is unique as it is partly located in mitochondria, apicoplasts and cytosol. On the other hand, in animals it is localized in mitochondria and in plants in plastids (Heinemann, Jahn, & Jahn, 2008; Layer, Reichelt, Jahn, & Heinz, 2010; Seeber & Soldati-Favre, 2010; Tanaka & Tanaka, 2007). This complex sub-cellular localization reflects the evolutionary mosaic with its origin from different sources. The potential of this pathway to be a pathway for drug interference was tested by using an inhibitor against one of the enzymes of this pathway in *T. gondii* and at high concentration the parasite was killed. But there is a lot to be done to elucidate the therapeutic potential of this pathway against these parasites (Striepen, 2011).

1.5 Apicoplast Phosphate Translocator

In the above description, some pathways for drug interference are described, but actually many more studies were done and are still being done to find a suitable way of inhibiting the function of apicoplasts (Fleige, Limenitakis, & Soldati, 2010; Goodman, Su, & McFadden, 2007; Lizundia, Werling, Langsley, & Ralph, 2009; Moreno & Li, 2008). Most of the work is focusing on internal processes of apicoplasts, which can be disadvantageous in a sense that if any potential inhibitor of any of the pathways is found, that inhibitor must overcome the barrier of four layers of membranes to reach to the target. In this case, Apicoplast Phosphate Translocator (APT) turned out to be interesting target for therapeutic intervention, which act as a link between the apicoplast metabolism and the cytoplasmic metabolism.

As discussed above, apicomplexa had to feed their apicoplasts with carbon sources, energy and reduction equivalents upon loss of photosynthesis. This supply is done by the APTs, which are members of a larger family of plastid phosphate translocators (pPT) (Striepen, 2011). These proteins act as antiporters and exchange inorganic phosphate for phosphorylated sugars of C3, C5 or C6 lengths (Brooks et al., 2010). In higher plants there are different pPTs for translocating different substrates, for example the triose phosphate / phosphate translocator (TPT) transports triose phosphate (Knappe, Flügge, & Fischer, 2003). Similarly, phosphoenolpyruvate transporter (PPT), xylulose-5-phosphate transporter (XPT) and glucose-6-phosphate transporter (GPT) transport phosphoenolpyruvate, xylulose-5-phosphate and glucose-6-phosphate, respectively (Eicks, Maurino, Knappe, Flügge, & Fischer, 2002; Fischer et al., 1997; Kammerer et al., 1998). In contrast, APTs have wider substrate specificity than the pPTs in higher plants and that may be the reason for fewer transporters in apicomplexa. In Plasmodium two PTs were found, which are differentially located in the outer and inner membranes of the apicoplast, and that's why they are named PfoTPT and PfiTPT, respectively. But in the periplastid membrane no such protein is identified yet, and it is suggested that these two proteins work in tandem to import the sugar into the apicoplast (Mullin et al., 2006). On the other hand, *Toxoplasma gondii* and *Theileria spp* have only one transporter (Fleige, Fischer, Ferguson, Gross, & Bohne, 2007). For *T. gondii* this transporter is called *T. gondii* apicoplast phosphate translocator (TgAPT) which is located in multiple membranes of the apicoplast (Fleige et al., 2007).

Among the APTs, PfiTPT, PfoTPT and TgAPT have been studied in more detail compared to others and their substrate specificities are determined *in vitro*. These studies revealed that they transport triose phosphate, 3-PGA, PEP and Pi, but not glucose-6-P with different substrate preference (Brooks et al., 2010; L. Lim, Linka, Mullin, Weber, & McFadden, 2010). These substrates enter the apicoplast and then act as precursors for different pathways. Now the question comes, how are these proteins different from other subtypes of pPTS and how do they accommodate compounds phosphorylated both in C-2 and C-3 in the same binding pocket? To answer this question, structural data are required, which is not available so far.

1.6 Disruption of APT and its consequences

One of the ways of determining the importance of a protein is to "knock out" the corresponding gene and observe the resulting phenotype. This was done for TgAPT and it was found that the parasite died rapidly. This death was thought to be linked with deprival of the apicoplast of metabolites required for anabolic pathways, specially FASII and isoprenoid biosynthesis (Brooks et al., 2010). In another study, mutation of Pb-TPT, which is one of the two APTs in *Plasmodium berghei* caused death of the parasite, while mutation of other (Pb-PPT) caused defect in the growth of the parasite (Banerjee, Jaijyan, Surolia, Singh, & Surolia, 2012). This phenomenon of

APT disruption leads to the idea of finding inhibitors against this protein to develop drugs against these parasites.

1.7 Virtual Screening: A Modern Drug Development Tool

In Late 1980 and early 1990 progress in experimental high throughput screening (HTS) and combinatorial chemistry created an excitement among the scientific community about launching significant amount of drug to the market. But due to low hit rates and significant costing reduced the euphoria (Lahana, 1999). So, it became necessary to develop new methods, which lead to the rise of virtual screening (VS). In contrast to HTS, which is mostly technology driven, VS uses computer programs to predict the binding of ligands to macromolecular targets like protein, DNA or RNA. There are two main approaches for virtual screening: Ligand based virtual screening (LBVS) and structure-based virtual screening (SBVS).

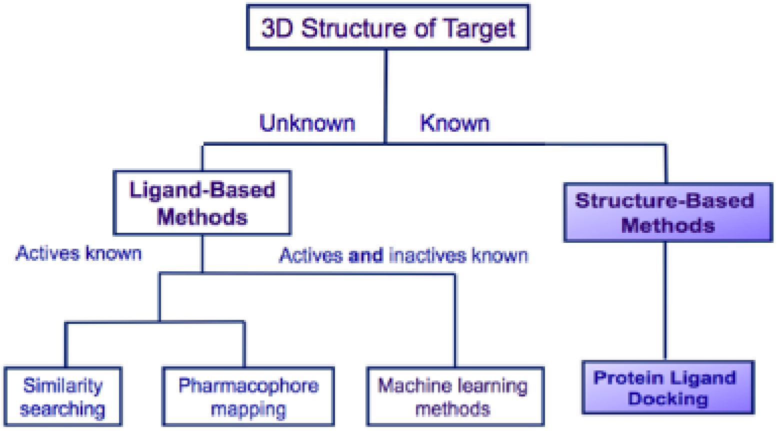


Figure 1. 4: Overview of virtual screening approaches. (modified from Gillet, 2013)

1.7.1 Ligand-based virtual screening

It is assumed that compounds with similar structures tend to have similar biological properties. Based on this principal, this approach is using the structures of active ligands for the target protein to derive potential active compounds.

There are several ways to derive structurally similar compound, which include pharmacophore mapping, machine learning methods and similarity method (Fig. 1.4). A pharmacophore is a set of structural features responsible for the compound functionality. In pharmacophore-based search, such a set is derived from the active compounds of the target protein and then it is used to find new compounds with similar features.

Machine learning methods are using the knowledge of known actives and known inactives to predict a model, which is then used to search for new compounds (Gillet, 2013).

Similarity based method uses active compounds of the target as reference structure and a search is done to find similar compounds of the reference structure. There are several ways of measuring similarity, which are categorized into two groups: molecular descriptors and similarity coefficients (Gillet, 2013). Molecular descriptors include physicochemical properties, two dimensional (2D) and three dimensional (3D) properties. Among these methods 2D fingerprinting was found to be most effective (Duan, Dixon, Lowrie, & Sherman, 2010). Similarity coefficients measure similarity between two sets of molecular descriptors.

1.7.2 Structure-based virtual screening

SBVS is an *in-silico* study of predicting ligands against a known target, whose 3D structure is available. This method includes several steps which are given in figure 1.5. In short, the target structure is prepared by choosing the binding site, selection of most relevant target structures, incorporating receptor flexibility, suitable assignment of protonation states and consideration of water molecules in the binding site. Then the ligands are prepared and docked in the target structure, ranked in order based on a scoring function, and final best possible hits are selected by more careful examination (Lionta, Spyrou, Vassilatis, & Cournia, 2014).

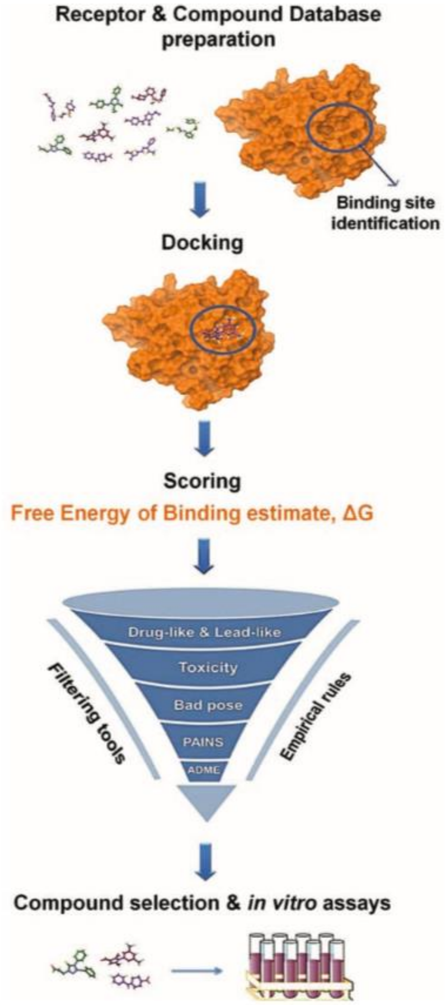

Receptor & Compound Database preparation
Binding site identification
Docking
Scoring
Free Energy of Binding estimate, ΔG
Drug-like & Lead-like
Toxicity
Bad pose
PAINS
ADME
Filtering tools
Empirical rules
Compound selection & *in vitro* assays

Figure 1. 5: **Workflow of SBVS (Lionta et al., 2014).**

1.8 3D structure of the target

3D structure is an essential part for VS, although in very rare case VS can be executed without 3D structure of DNA or RNA (Klebe, 2006). But, for proteins, it is must and real structure can be gained by X-ray crystallography or nuclear magnetic resonance (NMR) methods. Although these are most reliable, but it is not always possible to obtain the 3D structure of the desirable protein for various reasons. As an alternative molecular modelling can be applied to generate models.

1.8.1 Molecular modelling

By definition, molecular modelling is a way of mimicking the behavior of molecules or molecular system. Because of its usefulness, it has become popular in various fields to study 3D structures from small molecular system to large biomolecules including proteins. The main feature of this method is to generate a description of the atoms of a molecular system and there are two main ways for doing that: 1) Molecular mechanics (MM) and 2) Quantum mechanics (QM) (Chen & Houk, 1998).

In the MM approach, each atom of the system is considered as a particle and the interactions are describe by spring-like interactions and van der Waals and electrostatic forces (Cannon, 1996). The mathematical expression is called ‘Potential Energy Function (Etot)’, which takes into account the bonded (Ebonded) and non-bonded (Enon-bonded) atomic interactions. The bonded term computes the deviation of bond lengths (b), bond angles (θ) and torsion angles (φ) away from equilibrium values (Eq. 1) and non-bonded term describe van der Waals force and electrostatic interaction (Eq. 2) (Bordner, 2012).

Etot = Ebonded + Enon-bonded

$$E_{\text{bonded}} = \sum_{\text{bonds}} C_b \left(b - b_0\right)^2 + \sum_{\text{angles}} C_\theta \left(\theta - \theta_0\right)^2 + \sum_{\text{dihedrals}} C_\phi \left(1 + \cos(n\phi + \delta)\right) + \sum_{\text{impropers}} C_\alpha \left(\alpha - \alpha_0\right)^2 . \quad (1)$$

$$E_{\text{nonbonded}} = \sum_{\text{nonbonded}} \varepsilon_{ij} \left[\left(\frac{r_{ij}}{r_{ij}^{\min}} \right)^{-12} - 2 \left(\frac{r_{ij}}{r_{ij}^{\min}} \right)^{-6} \right] + \frac{q_i q_j}{\varepsilon r_{ij}} . \qquad (2)$$

The first three terms in Eq. 1 represents the energy of bond stretching, angle bending, rotation of torsion angle and the last term is used to maintain planarity of peptide bonds and aromatic rings in protein structures. In Eq. 2 the first term represents van der Waals energy and the last term represents electrostatic energy. The suffix i, j represents atoms (fig 1.6) (Bordner, 2012)

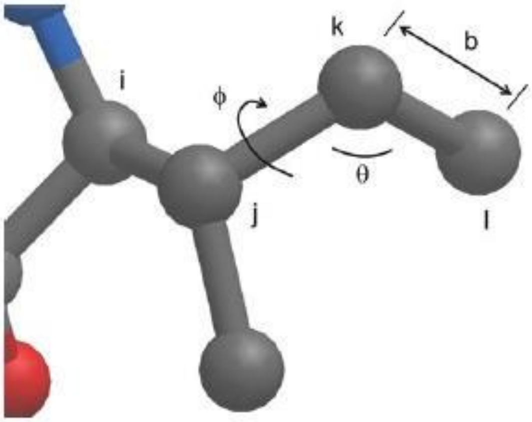


Figure 1. 6: **Bonded interaction variables for the bond length (*b*), bond angle (q), and dihedral angle (f) as seen in Eq. 1 (Bordner, 2012).**

This approach is valid for doing energy minimization, energy calculation of specific conformation, generating different conformation, identifying best conformation and molecular motion.

In QM, the movement of electrons relative to nucleus are also included, which made it possible to derive properties that depend upon the electronic distribution. As a result, this approach has higher accuracy of geometry and energy calculation than the MM. The problem with this method is that it is time consuming and limited to small molecules (Chen & Houk, 1998).

1.8.2 Protein modelling

There are three different ways for constructing 3D models of proteins: 1) Homology modelling, 2) Threading/ fold recognition and 3) *Ab-initio* methods.

Homology modelling is used when the structure of a similar (homologues) protein (template) to the target is available. Using the structure of the template, the structure of the target protein can be constructed (Krieger, B Nabuurs, & Vriend, 2003).

The next method is threading, which is applicable when there is no detailed structure of a specific homologue available, but only homologous proteins with low similarity with the target. In this case the sequence of the unknown target protein is compared with available structures with low similarity in the PDB database and then the best fitting structure is selected (Forster, 2002).

The *ab-initio* method is used, when there is no template available. So, local fold of a sequence is predicted by computational method and then compared with other protein sequences. In the end, the whole protein is modelled. This method is suitable for smaller proteins with less than 85 amino acids (Bradley, Misura, & Baker, 2005).

1.8.2.1 Homology modelling

The basis of homology modelling relies on two observations:

1) The 3D structure of a protein is determined by its sequence (Epstein, 1964)
2) The fact that during evolution structural changes evolve much slower than changes in sequence, such that not only similar sequences but also related sequences fold into similar structures (Chothia & Lesk, 1986; Sander & Schneider, 1991).

For homology modelling, a 3D structure of a similar (homologues) protein is required, which can be used to build the model of target protein. The higher similarity between template and target, the better chance for a good model to be built. But this similarity limit can vary among protein types. For example, for soluble proteins 30% similarity is considered as the borderline, but more than 50% is believed to produce high accuracy model. But for membrane protein the similarity between template and target can be very low (even less than 20%), but their structural identity can be high in transmembrane regions and the active site. So, using a structure of low similarity, it is still possible to generate model having reliable transmembrane region and active site (Ravna & Sylte, 2012).

There are several steps in homology modelling which are shown in the schematic diagram below:

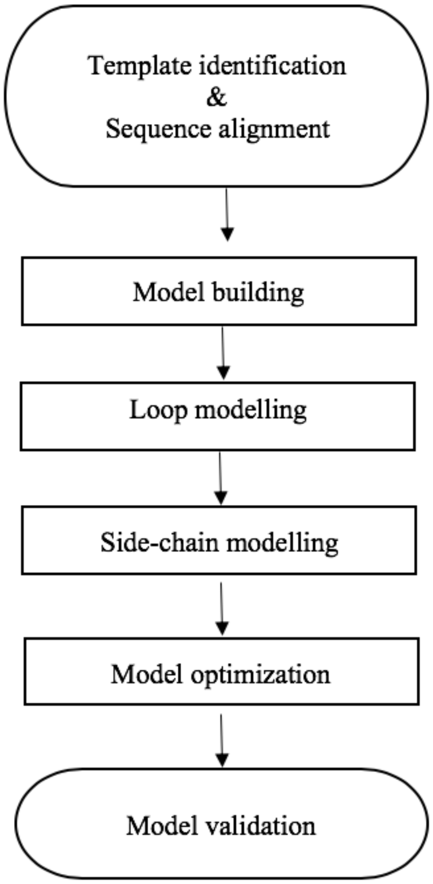


Figure 1. 7: **Schematic diagram of homology modelling protocol.**

1.8.2.1.1 Template identification and sequence alignment

A template can be the structure of a protein, which sequence fall into the 'safe' zone compared to the target sequence in terms of similarity. In practice, one can take the sequence of the target and using it as query sequence, make search for similar protein structures in any BLAST (Basic Local Alignment Search Tool) server and obtain hits with corresponding alignments. Sometimes some regions are found which are not so similar, and in that case the two sequences are aligned with other homologue sequences to fix regions of low similarity. This method is termed multiple sequence alignment (Krieger et al., 2003).

1.8.2.1.2 Backbone generation

When the alignment is ready, it is possible to create the model. It is done by copying the coordinates of the template to the new structure, according to the alignment. For identical residues, the side chain of the residues can be included (more rigid side chain as rotamer are conserved), but if the residues are different only the backbone coordinate (N, Cα, C and O) can be copied (Krieger et al., 2003).

1.8.2.1.3 Loop modelling

Homologous proteins contain gaps, when aligned due to insertion and deletion in either of the sequences, which is referred to as loops. These loops are important in both structural and functional aspects. But it is very difficult to predict the loop conformation. There are two main approaches for loop modelling:

1) Knowledge-based: Searches the PDB database for loops with matching residues to the target.
2) Energy-based: an ab-initio approach to predict the fold and then the energy function is used to judge the quality, which is then minimized to possible best conformation (Krieger et al., 2003)

1.8.2.1.4 Side-chain modelling

As mentioned before, side chain can be obtained from the template in case of identical residues or need to be generate by ab-initio modelling. Naturally, protein side chains exist in limited number of low conformations, called rotamer. During modelling this rotamer is selected based on the sequence and then the backbone coordinates and the quality is assessed.

1.8.2.1.5 Model optimization

To have a model of high accuracy, it is required to have a correct backbone, which is dependent on correct side chain rotamer and packing. The rotamer prediction in turn depends on correct backbone. So, several steps of rotamer prediction and energy minimization is done until the whole structure is optimized. The energy function is very important for this step (Krieger et al., 2003).

1.8.2.1.6 Model validation

It is almost obvious that errors will be introduced in the model structure, therefore it is required to validate the model before using it for structural predictions. This can be done by uploading the model to the structure analysis and verification server (SAVES; http://nihserver.mbi.ucla.edu/SAVES/) to check the stereochemical quality of the model.

Another approach is to dock known binders and non-binders in the model and check how good the model is distinguishing between them, which is a test of the accuracy of the binding site region.

1.9 Docking

After the development of the first algorithm for molecular docking, it became a popular tool in predicting conformations of small molecule ligands with the binding site of the target, with a degree of accuracy. This process includes two steps: exploration of potential binding conformation of the ligand and predicting interaction energy associated with each conformation, termed as scoring (Ferreira, Dos Santos, Oliva, & Andricopulo, 2015).

In the conformational search, the degrees of freedom of the ligand, which is defined by the torsional, translational and rotational parameters, are increasingly modified. To detect suitable binding modes, the conformational search are using both systematic and stochastic search algorithms (Ferreira et al., 2015).

In systematic search, the conformation changes gradually and the energy landscape is explored for each conformations. After numerous search the minimum energy solution is selected as the most likely binding mode (Sousa, Fernandes, & Ramos, 2006). The problem with this is that number of possible combinations grows exponentially with the increasing degrees of freedom of the ligand, which leads to combinatorial explosion. Docking tools have their own strategy to handle this problem (Ferreira et al., 2015).

In a stochastic method, conformations of the ligands are generated randomly until a low energy conformer is obtained. In contrast to systematic search, which is prone to select local energy minimum, stochastic method has higher chance of finding a global energy minimum (Zsoldos, Reid, Simon, Bashir Sadjad, & Johnson, 2007).

1.10 Scoring

The scoring functions estimate the binding energy by taking into account the physical chemical phenomenon like intermolecular interactions, desolvation and entropic effects, which are involved in ligand-target binding. So, the greater the number of considered parameters is, the closer the scoring functions are towards accuracy and reality (Ferreira et al., 2015). But due to the computational costs involved, the scoring functions have to maintain the balance between speed and accuracy. Scoring functions are categorized as follows:

Force-field based approach which takes into account the bonded and non-bonded interactions like van der Waals, electrostatic interaction and hydrogen bonding between all atoms of the binding partners in the complex. Solvation and entropic effects are also considered but not explicitly (Ferreira et al., 2015).

Emperical scoring functions are based on counting the number of various interactions like hydrogen bonding, ionic and apolar interactions. It also considers the desolvation and entropic effects. These functions were found to be effective for several protein ligand complexes (Lionta et al., 2014).

Knowledge-based function use statistical observations of intermolecular contacts in receptor-ligand, whose structural conformations are established (Lionta et al., 2014).

Although scoring functions are widely used to calculate the binding energies, it is also accepted that they usually fail to rank compound in proper order, and it is still a challenge to choose the correct binding pose as the top ranked one (Ferreira et al., 2015).

1.11 Aim of the study

From the above discussion it is seen that the apicoplast is a metabolic hub in Apicomplexa parasites, which is fueled by the APT. Due to the plastidic nature of APT, it is a potential drug target. Among the APTs. TgAPT was studied best and its potential as a drug target was tested by disrupting the APT gene, which lead to the quick death of the parasite. So, finding inhibitors against this protein will not only help to develop drugs against *T. gondii*, but also against *P. falciparum* as the APTs of these organisms have significant similarity.

There is no crystal structure of TgAPT available. But two 3D structures of a TPT from *Galderia sulphuraria* that was co-crystallized with two substrates (phosphate and 3-PGA) were published by Lee et al. (2017). Based on these structures, two homology models of TgAPT were generated in previous work (Vold, 2018) and named TgAPT_5Y78 and TgAPT_5y79. The models were optimized and validated. In this study, these two models will be used to:

1) Predict potential inhibitors of TgAPT for in-vitro testing, using a ligand-based and structure-based virtual screening approach.
2) Elucidate the binding site differences between TgAPT and PfoTPT, which has similar substrate specificity. For this a homology model of PfoTPT will be generated and compared with the TgAPT models.
3) Compare the binding site of TgAPT with the binding site of other pPT subtypes and relate the differences to the differences in substrate specificity.

2 METHODS

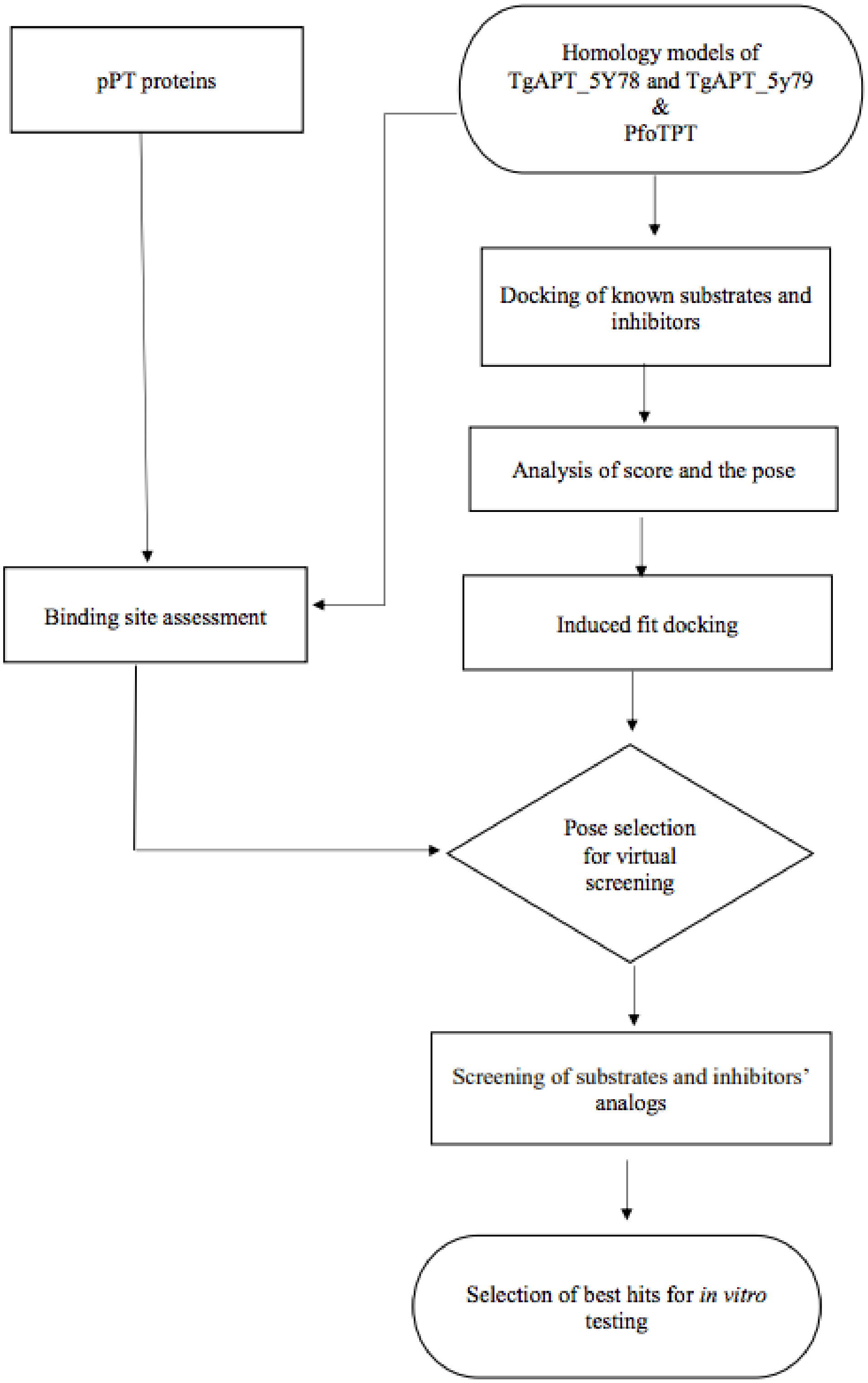

pPT proteins
Homology models of TgAPT_5Y78 and TgAPT_5y79 & PfoTPT
Docking of known substrates and inhibitors
Analysis of score and the pose
Binding site assessment
Induced fit docking
Pose selection for virtual screening
Screening of substrates and inhibitors' analogs
Selection of best hits for *in vitro* testing

Figure 2. 1: Schematic diagram of the workflow

2.1 Structure import into Maestro workspace

As already mentioned, there are two homology models of TgAPT generated by using triose phosphate transporter structure of *G. sulphuraria* as template. To view the previously generated models of TgAPT, the Schrodinger Maestro program was used on a Computer based on the Linux operating system. Before importing the files, the working directory was set to a desired location and the project was saved by a specified name “TgAPT_project”. After that from the “Import Structure” option under the “File” menu two models of TgAPT named “TgAPT_5y78” and TgAPT_5y79” were imported into the workspace from the specified folder. Only one of the structures will be appear on the screen, other one remained in the entry list.

2.2 Renumbering the Models

The template sequence was shorter than the target sequence, and during model generation, proper alignments of amino acids 1-38 amino at the N-terminal with the template was not obtained and these amino acids were therefore not present in the 3D TgAPT models. As a result, the 39th amino acid of the original sequence was numbered as 1, which created some confusion to track the important amino acids described in the literature. The sequences of the homology models were therefore renumbered starting with amino acid 39. For doing this, the 3D models were imported into the Schrödinger workspace and then opened from the task menu “Multiple Sequence Viewer”. The model sequence was now displayed on the screen in addition to the 3D model, and the renumbering option in the Edit panel was used to renumber the 3D model starting from amino acid 39 of the TgAPT sequence.

2.3 Binding site comparison of TgAPT with other pPT

As TgAPT has a unique substrate specificity, it was quite interesting to know the difference of the binding pocket of this protein with other phosphate translocators. In this process, only TgAPT_5y79 was used and firstly, comparison was done with the pPT classes: TPT, PPT, GPT and XPT and then specific comparison was done with a translocator from another species of Apicomplexa phylum, which reside in the outer membrane of apicoplast of *Plasmodium falciparum.*

2.3.1 Comparison with subclasses

Before doing the comparison, it was necessary to identify amino acids in the binding pocket of TgAPT. For this, the crystal structure of GsTPT2 5y79, which was co-crystallized with 3-PGA, was imported into maestro suite. Then the co-crystallized ligand 3-PGA was selected, and the selection was expanded to 5 Å outside the ligand. By this way the amino acids within this region was selected and labelled, which resembles the binding pocket of the crystallized protein. Then the amino acids in TgAPT protein at these locations were detected by pairwise alignment. By selecting and superimposing these amino acids on the crystal protein the binding pocket was confirmed on TgAPT protein. Finally, the amino acids in the members of different subclasses at those specific locations were detected from an in-house sequence alignment (Appendix 1)

2.3.2 Comparison with PfoTPT

To do this study, a model of PfoTPT was prepared and then the binding site of the protein was examined and compared with TgAPT. The methods are described below.

2.3.2.1 Homology Modelling

For this method, 'Prime Structure Prediction Wizard' in Maestro suite was used. The steps followed in this method was according to the Prime user manual (Prime, 2019). Firstly, structure prediction wizard was opened from the task menu. The sequence of PfoTPT, triose phosphate transporter [Plasmodium falciparum 3D7] (accession no. XP_001351641.1) was derived from the NCBI database. After that, the sequence of the template structure was inserted from the workspace, which was the crystal structure with PDB-code 5Y79. This protein has two identical chains, and chain B was used for this task. In the next step, for the sequence's alignment Prime STA (Single Template Alignment) method was chosen as the sequence identity was low (33%) between the template and target sequences. This alignment approach takes into account secondary structure matching along with sequence matching, which allows to generate better alignment in regions of low sequence conservation. By manual editing residue 1-35 of PfoTPT were cropped and some other changes were done to make the alignment look like the sequence alignment provided by Karsten Fischer (Appendix 1). Then the structure was built. After finishing the model building, loop refinement was done according to default setting of the 'Prime Loop Refinement' tool as the loop length was less than six amino acid. In this setting the loop is reconstructed using the backbone dihedral library, by building up half from each direction. By this way many loops were generated which then were clustered, and representatives of each cluster were selected. These loops are then ranked by assigning scores. Scores were assigned by the following procedures: side chains are re-added to the representatives. The loops and side chain were then energy minimized. Finally, the best scoring loop structures were returned. After that, the refined regions were energy minimized. Apart from this, steric clashes, bond length and bond angle deviations were updated through protein preparation tool. Finally, the energy minimized structures were exported as PDB files.

2.3.2.2 Binding site detection

The amino acids in the binding site of these two new models were examined as described in section 2.3.1.

2.4 Searching for known substrates and inhibitors of TgAPT

Before doing a virtual screening with a homology model it is important to know the reliability of the model is in terms of interactions with known substrates and inhibitors. Due to limited experimental binding data it was not possible to know exactly how the interaction would be, which means how the substrates fit into the binding pocket, which amino acids in the protein interact with the ligand and why non-binders do not bind and inhibitors inhibit the function of the protein. But still it is possible to make an assumption based on the experiments done *in-vitro* on this protein. From this idea literature search was done to learn about the active and inactive compounds which means compounds that are substrates, inhibitors, or not bind at all to the transporter. After generating a list, a prediction was made about the rank order of the different substrates based on their experimental affinity towards the transporter.

2.5 Substrate and inhibitors docking in the homology models:

One of the prescribed ways of testing a homology model is to dock known compounds of that protein into the model and check whether the result reconstruct experimental observations. That was done in the present study, and the procedure is explained below.

2.5.1 Ligand preparation

2D structures of selected substrates and inhibitors were downloaded from the Pubchem database (https://pubchem.ncbi.nlm.nih.gov/) in "Sdf" format and included as entries into the Maestro workspace. After that, from the task menu of Maestro suit "Ligprep" program was opened. The structures of the substrates and inhibitors were used as input file by selecting these structures in the workspace and choosing Workspace in the "use structures from" option. For ionization Epik was chosen, which predict not only the ionization state but also the energetic state associated with them. The pH range for generation of states were 7.0 ± 2.0. Keeping the 'Desalt' option, while generate tautomer option was deselected. For stereoisomer computation the specified chirality of the input ligands was kept. The output file was saved as "Ligands.3d.mae".

2.5.2 Protein preparation

Before virtual screening it is required that a protein is prepared by fixing missing atoms and side chain, assigning bond order and formal charges, optimizing H-bond network and minimization of the structure. For this purpose, Protein Preparation Wizard in the Maestro Suit was used.

The homology models of TgAPT were imported into the Maestro workspace and then the Protein Preparation wizard was turned on from the favorite toolbar of Maestro. In this program there are three tabs named 'Import and Process', 'Review' and 'Refine'. By 'Import and Process' tab target protein is imported, if that is not already done and then basic structural fixation is done. There are several options under this tab, of them 'Assign bond order', 'Add hydrogen', Create zero-order bonds to metal', 'Create disulfide bond' options were chosen. Also, water molecules beyond 5 Å of het groups were deleted and het states were generated within 7 ± 2 pH. The function of 'Review' tab is to delete unwanted side chain and fix and delete het groups. Only for preparing the crystal structure this tab was used to delete one of the chains of the protein and remove the detergent molecules. Under the tab 'Refinement', optimization of H-bond network is done by reorientation of OH group, H_2O molecules, amide group in Asparagine (Asn) and Glutamine (Gln), imidazole ring in Histidine (His), predicting the protonation stage of His, Asn, Gln as well as tautomeric states of His. After the optimization of H-bond, the structure was minimized by selecting all-atom minimization with a termination criterion based on the root-mean-square deviation (RMSD) of 3 Å of the heavy atoms relative to their initial location.

2.5.3 Binding site prediction by Sitemap

Information about the binding pocket is required for docking, but the pocket was not defined in the homology models of TgAPT, and it was necessary to predict the pocket. Prediction was done by the Sitemap program in Maestro. For the prediction, the default setting in Sitemap was used. In short, 'Identify top-ranked potential receptor binding sites' was selected, for hydrophobicity definition 'More restrictive' and for grid 'Standard' option was chosen. The site map was cropped at 4 Å from nearest site point. Using this setting 5 sites were reported. The most realistic was selected based on similarity with the binding site of the template structure (GsTPT2).

2.5.4 Docking

For docking the Glide docking tool was used. In this tool, the binding site of the protein has to be prepared as grid before docking. In this study the binding site was selected from the prediction by 'Sitemap'. During receptor grid generation the van der Waals scaling factor was set to 1.0 and partial charge cutoff value was 0.25. After grid generation the ligands (prepared by Ligprep) were allowed to dock. There are three modes of docking in Glide: high throughput virtual screening (HTVS), standard precision (SP) and extra precision (XP), which differ in sampling ligand degrees of freedom and the scoring function employed. HTVS and SP uses the same docking algorithm and scoring function, but HTVS consider lesser ligand degrees of freedom and reduces final torsional refinement and sampling (Repasky, Shelley, & Friesner, 2007). On the other hand, XP does more extensive sampling than SP and employs a harder scoring function with greater requirement for protein-ligand shape complementarity. In this study, SP docking was used.

2.6 Induced fit docking in TgAPT_5y79

As the protein was kept rigid in the initial docking, it is possible that true substrates and inhibitors would score better if the amino acids in the receptor were allowed some movements, which may give better protein conformations for the SBVS process. This possibility was tested through "Induced Fit Docking" in Schrodinger Maestro Program Suite which was opened from the Task menu. In the "Ligands to be Docked" option, two of the substrates and two of the inhibitors were tried. Of the two substrates, one was phosphorylated at C-3position, which is 3-PGA and another one was phosphorylated at C-2 position which is PEP. Among the inhibitors one was phosphate containing, which is pyridoxal phosphate (PLP) and another one was sulfate containing, which is trinitro benzene sulfonate (TNBS). Before docking, they were prepared by "Ligprep". To define the amino acids for the grid box center, several docking using 3-PGA as ligand and trying different combination of amino acids in each run. In the first combination, 6 amino acids were chosen which were His 126, Lys 145, Ser 204, Arg 207, Tyr 287 and Arg 311 residing in TM respectively. In the second combination again 6 amino acids were chosen replacing Ser 204 with Asn 307. And finally, only 4 amino acids were selected which are His 126, Lys 145, tyr 287 and Arg 311. Based on the best result, other three run were done with the remaining three ligands for induced fit.

2.7 Virtual screening

2.7.1 Pose selection for virtual screening

After generation of several protein conformations (poses) by induced fit docking, the next challenge was selecting suitable poses for virtual screening. For this each pose was inspected visually. In this inspection several things were considered. For example, docking score, ligand position relative to the target, and especially the position of the phosphate group, and that the amino acids were interacting with the phosphate. In addition, similarity with the ligand interactions observed in the template crystal structures were considered. When an interesting pose was found, the whole binding pocket was superimposed with the binding pocket of original homology model as well as the crystal structure to see the structural change of amino acids during induced fit. After selection of eight poses, the initial set of substrates and inhibitors were docked in those poses. Finally, the one, which produced comparatively better score than the initial docking and maintained the affinity order of the substrate most similar to their experimental affinity values, was chosen for the actual virtual screening.

2.7.2 Analog Search

Structural analogs of the inhibitors were downloaded from the Pubchem database. Firstly https://Pubchem.ncbi.nlm.nih.gov/# was accessed. In this page, several options were available on the right side and from there "Structure search" was selected. A new page appeared, from where identity/similarity was chosen. Under this option, there were three ways to define the target compound, for this study SMILES codes were used. Then, there were options to choose expected similarity of the compound. Similarity is measured by using Tanimoto equation and PubChem dictionary-based binary fingerprint. This fingerprint consists of a series of chemical substructures, termed as ‘Keys’. Each key represents presence or absence of a particular substructure in a molecule. These substructure keys do not take into account the stereochemical and isotopic variations. This is how these binary keys provide a chemical structure with a fingerprint. The degree of similarity is then selected by threshold parameter. For compounds like Pyridoxal phosphate, DIDS, 2,4,6-trinitrobenzene sulfonate and 4-sulfobenzenediazonium 80% threshold of similarity and for phenylglyoxal 90% similarity was chosen. Using these parameters search was done, and after that the structures were downloaded in SDF format.

2.7.3 Docking the analogs

Like the previous steps, the analogs were prepared with “Ligprep” and then docked on the protein conformations derived from 2.7.1 following the same procedure as described in 2.4.5. After that, a threshold score was set for each analog group. This threshold score was set considering the following: highest and lowest scoring value, score of the parent compound and number of compounds above the threshold score. For example- in the 3-PGA analog group, the highest scoring compound scored at -11.66 Kcal/mol, lowest scoring compound score -0.25 Kcal/mol and 3-PGA itself scored -9.90. It was found that if the threshold score for this analog group is set to -9.0 Kcal/mol, a reasonable number of compounds can be extracted for visual inspection. For other analog groups threshold score was set in a similar manner. Compounds scoring above that threshold were exported as separate entries. There were some repeats of the same structure which were discarded. These isolated compounds were then examined and sorted by structural clustering.

Table 2. 1: Threshold scores for selection of compounds for clustering.

Parent compound	Score of selection threshold (kcal/mol)
3-PGA	-9.0
Gly-3-P	-7.0
PEP	-7.0
PLP	-8.0

DIDS	-7.0
TNBS	-7.0
4-SBD	-7.5
Phenylglyoxal	-6.0

2.8 Clustering:

According to Similar Property Principle by Johnson and Maggiora (1990) molecules having similar structure are likely to possess similar properties. That is the reason for that clustering provides with the possibility to cover bigger spectrum of compounds by allowing to choose one or two compounds in a cluster, as a representative for the whole cluster. For the clustering, the “Canvas” program in the “Schrodinger Suite” was used. Firstly the “Canvas” program was opened from the terminal and then the isolated structures of one of the analogs were imported. Then their hashed binary fingerprints were created by ‘Binary Fingerprint’ option in the ‘Application’ Menu. In this study among the various types of hashed fingerprints ‘Molprint2D’ type was generated. It was incorporated in the program. Next, using the ‘Hierarchical Clustering’ application the compounds were clustered based on the fingerprint that was generated using ‘Tanimoto Similarity’ metric and in the ‘Cluster Linkage Method’ ‘Average’ was chosen. The dendogram of the cluster was opened and some adjustments were done, such as reducing or increasing the number of clusters to make it convenient for further analysis. Finally, the structures in the clusters were exported in a separate file. Similar things were done for rest of the analogs.

2.9 Sorting out compound from clusters:

The clusters created in the previous step were imported in ‘Maestro’. These structures were then examined visually to look at their size, docked position, interaction and the score. By this manner one with the better score and interaction in a cluster, was considered for in vitro screening.

3 RESULT

3.1 Homology Modelling of PfoTPT

Sequence alignment showed 33% similarity between the template (GsTPT2) and the target (PfoTPT) (Appendix 2), which can be considered acceptable as these are membrane proteins. Based on this alignment one model was built (figure 3.1). Similar to the template this model has 10 transmembrane helices and the loops were predicted by the program. After the model built some of the loop regions were refined and these regions are residues 102-106, 217-220, 248-252, 261-264.

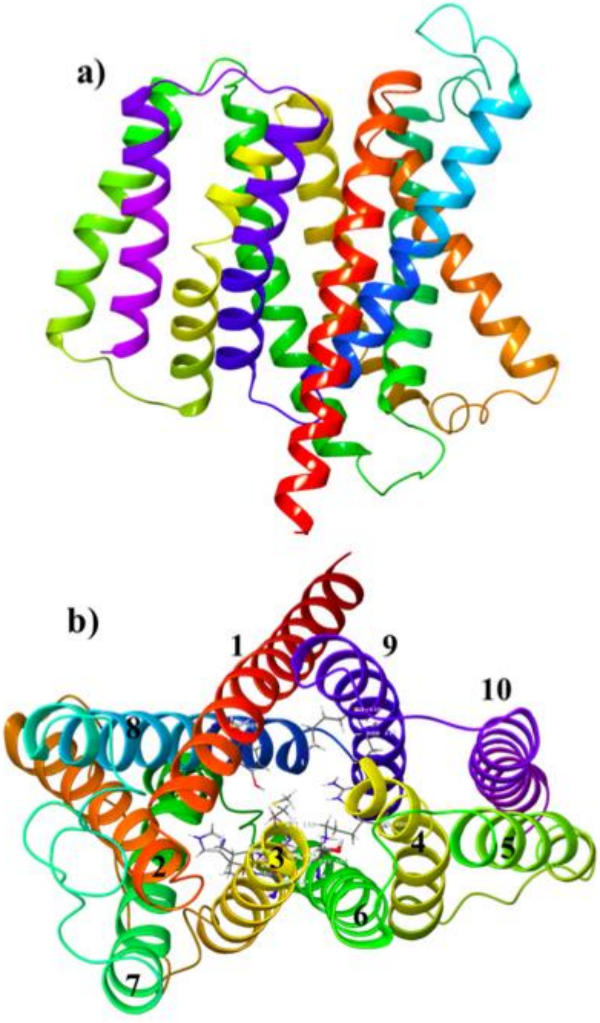


Figure 3. 1: **Backbone of homology model of PfoTPT a) Side view. b) Top view**

To evaluate the model, the model was superimposed on the template and rmsd between template and target was found to be 7.43. The PDB format of the model was uploaded to SAVES server for further evaluation. The results are given below:

Table 3. 1: Result of homology model verification tool

Evaluation tool	result
Verify	53.09% of the residues have averaged 3D-1D score >= 0.2 Fail
Errat	Overall quality factor: 93.64
Prove	Buried outlier protein atoms total from 1 Model: 4.6% Warning
Procheck	Out of 8 evaluations • Errors: 2 • Warning: 4 • Pass: 2

The verify tool determine how compatible the 3D model is with its own sequence by predicting a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing it to high quality experimental 3D structures (Mannhold, Kubinyi, & Timmerman, 2008). Compared to the expectation, which 80% of the residues scoring =>0.2 in 3d/1D profiling, this model scored 53.09%.

According to ERRAT the overall quality factor is 93.64, which is an indication of structure reliability (Colovos & Yeates, 1993).

The Prove tool calculates the z-score deviation of the model from the highly resolve PDB-structures based on the volume of the atoms, where atoms are treated as hard spheres (Pontius, Richelle, & Wodak, 1996). A model pass this when its score is <1 %, here the model scored 4.6 % which is a warning. Scores > 5% is considered as failing.

PROCHECK verifies the stereochemical quality of a protein (Laskowski, Macarthur, Moss, & Thornton, 1993) and this model passed in two, got warning in 3 and failed 2, out of 8 evolutions. As for the Ramachandran plot, which was also passed, showed 94.4% of the residues were in most favored regions, 4.9% in additional allowed regions and 0.7% were in disallowed regions (fig 3.2).

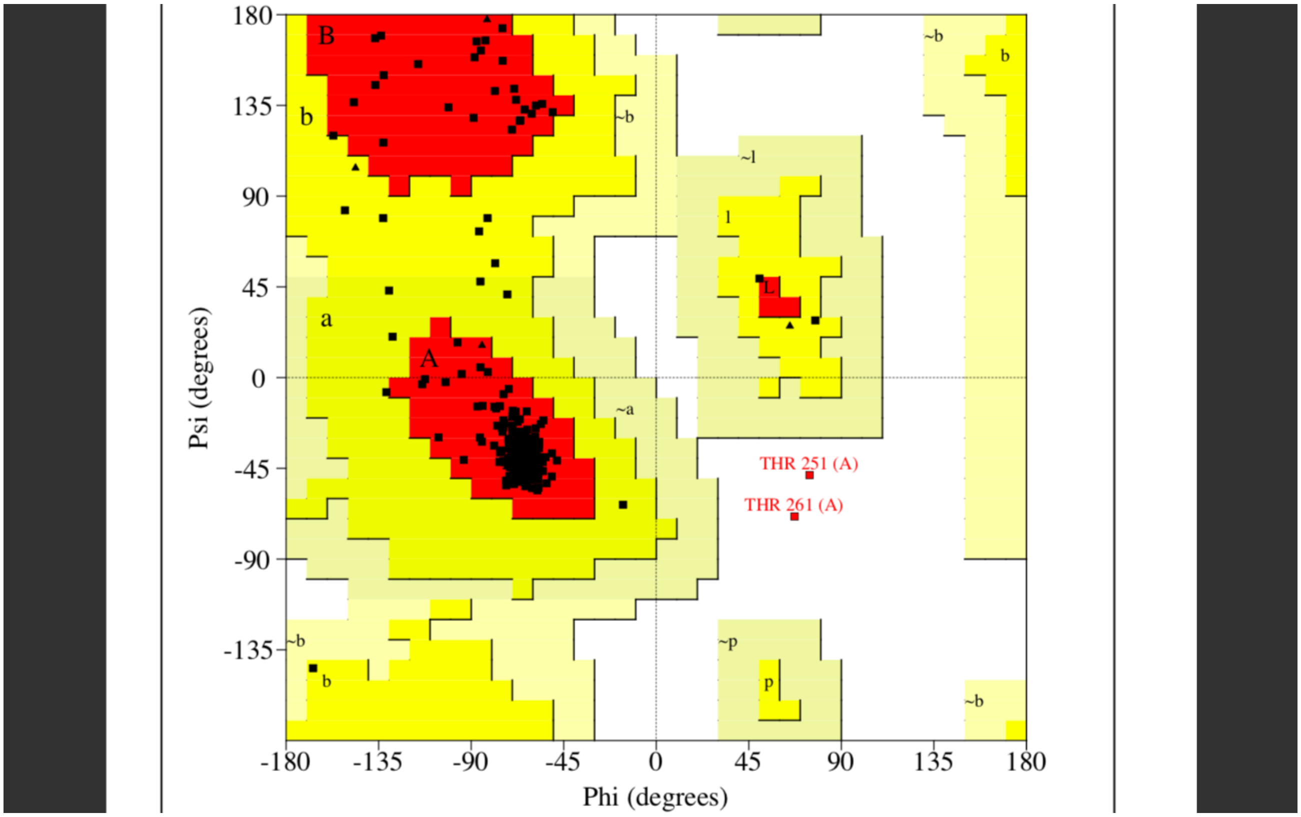


Figure 3. 2: Ramachandran plot of PfoTPT model

3.2 Comparison of binding site

The binding site comparison is presented in three parts:1) Between different classes (TPT, PPT, GPT and XPT), 2) Between TgAPT and PfoTPT and 3) Between two APT and all other classes. A number of sequences from higher plants in each subclass was included in this study. For example: 26 sequences of the TPT class, 30 of the PPT class, 19 of the GPT and 3 of the XPT class.

Twenty-three amino acids were found within 5 Å of co-crystallized ligand 3-PGA in the template structure 5y79, (Table 3.2) and these amino acids are distributed within in 7 helices (1,2,3,4,6,8,9 helices). Similarity and differences in these positions between different classes and TgAPT and PfoTPT were determined as described in the method. For the ease of presentation, the comparison is done by classifying these amino acids into three groups: Phosphate recognizing, Carbon body of the ligand recognizing and amino acids with no binding role. Another thing to note here is that the amino acid position of GsTPT2 will be considered as anchor and others will be compared relative to them.

Table 3. 2: Amino acids within 5Å of co-crystallized ligand 3-PGA in GsTPT2 and corresponding amino acids in TgAPT, PfoTPT, TPT, PPT, GPT and XPT respectively. Yellow labels indicate differences from the corresponding GsTPT2 residues and the grey one in PfoTPT is the only difference between TgAPT and PfoTPT. In some cases, there were differences among the compared sequences of the same class, which is showed by mentioning the number of species (sp.) the difference is presen**t**

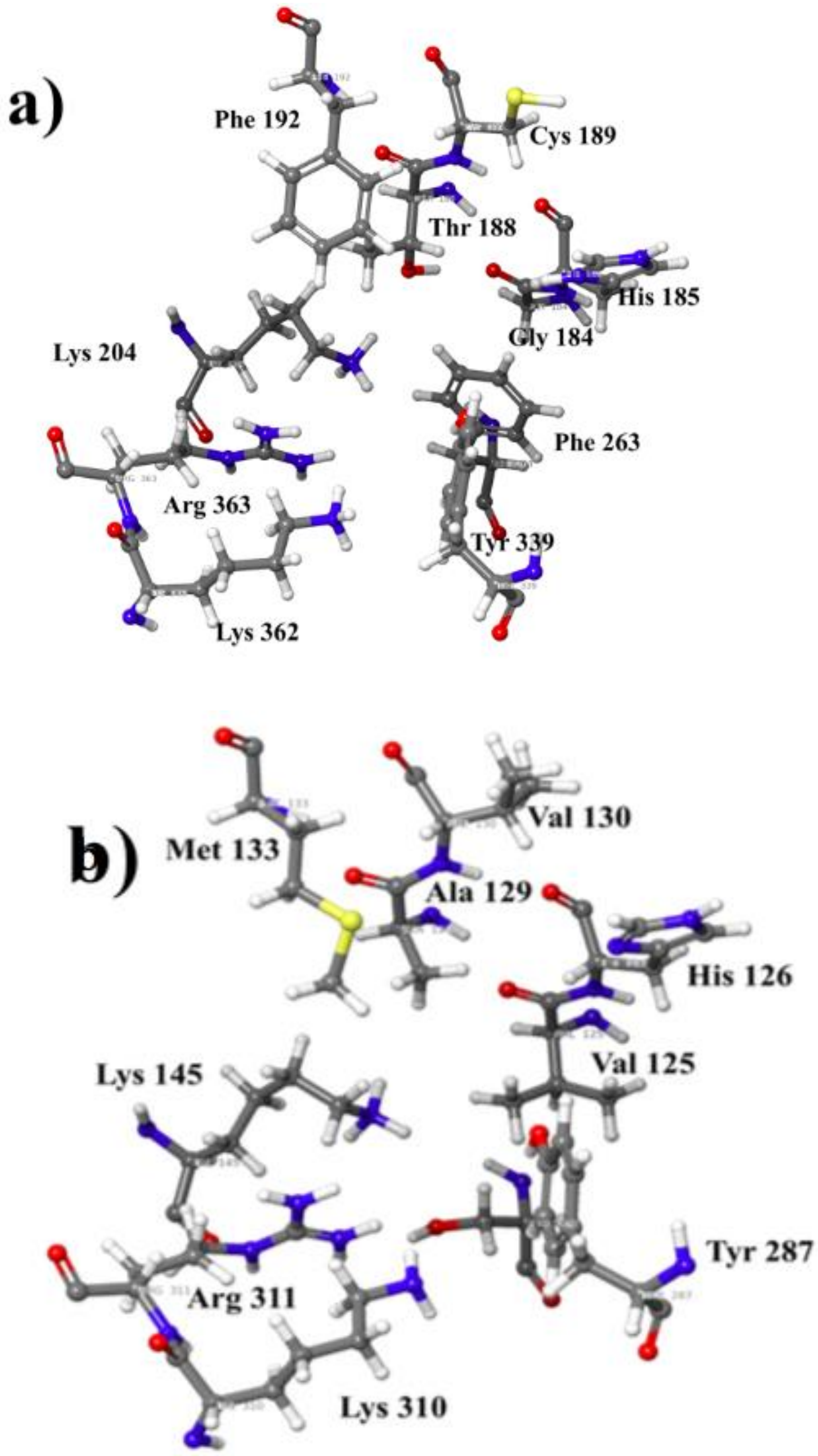
GsTPT2	TgAPT	PfoTPT	TPT	PPT	GPT	XPT
Trp 116	Trp	Trp	Trp	Trp	Trp	Trp
Asn 120	Asn	Asn	Asn	Asn	Asn	Asn
Phe 123	Tyr	Tyr	PHe	Phe/ tyr (3 sp.)	PHe	PHe
Asn 124	Asn	Asn	Asn	Asn/ His (only 1 pr.)	Asn	Asn
Asn 127	Asn	Asn	Asn	Asn	Asn	Asn
Gln 144	Gln	Gln	His	Gln	Ser	Gln
Gly 184	Val	Val	Gly	Gly	Gly	Gly
His 185	His	His	His	Asn/ Thr (only 1 sp.)	His	His

Thr 188	Ala	Ala	Ser/ Thr (3 sp.)	Thr	Ala	Ala
Cys 189	Val	Val	Asn/Thr (3 sp.)	Asn	Thr	Cys
Phe 192	Met	Met	Phe	Leu (Ile 1 sp.)	Met	Phe
His 201	His	His	His	His	His	His
Lys 204	Lys	Lys	Lys	Lys	Lys	Lys
Glu 207	Glu	Glu	Glu	Glu	Glu	Glu
Ser 259	Ser	Ser	Ser / Ala (only 2sp.)	Ser /Cys/ Ala (1& 1 sp.)	Ser	ser
Phe 263	Ser	Ser	Phe	Asn/ Phe (Phe 8 sp.)	Phe	Phe
Arg 266	Arg	Arg	Arg	Arg	Arg	Arg
Tyr 336	Tyr	Tyr	Tyr /Phe (1 sp.)	Phe / Leu (1 sp.)	Tyr	Tyr
Tyr 339	Tyr	Asn	Tyr /Asp (1 sp.)	Tyr	Tyr	Tyr
Asn 340	Asn	Asn	Asn	Gln	Asn	Asn
Asn 359	Asn	Asn	Asn / Ser (1 sp.)	Asn	Asn	Asn
Lys 362	Lys	Lys	Lys	Lys	Lys	Lys
Arg 363	Arg	Arg	Arg	Arg	Arg	Arg

Phosphate recognizing residues, which are Lys 204, Lys 362 and Arg 363 in GsTPT2 are found to be conserved in all sequences examined here, most probably because of all proteins of the family have phosphates common as their substrates. For the amino acids involved in other part of substrate recognition or the amino acids close to them some differences were observed, which might contribute to the different substrate recognition by different classes. For example: in place of His 185, PPT contain Asn/ Thr, For Thr 188, TPT has Ser or Thr, GPT and XPT has Ala. And instead of Phe 263 in GsTPT2, the PPTs has Asn in some Phe in some.

There are also differences between classes in the amino acids not directly involved in ligand interaction. For example: proteins of TPT class contain His instead of Gln 144, Ser/ Thr for Thr 188, Asn/ Thr for cys 189 compared to the GsTPT2 protein. Similarly, PPT proteins has Phe/Tyr in place of Phe 123, Asn in place of Cys 189, Leu instead of Phe 192, Phe in place of Tyr 336 and Gln in place of Asn 340. In case of GPT, it has Ser in place of Gln 144, Thr for Cys 189. Among all classes XPT has the most similar binding pocket structure to GsTPT2 and differs only in one position, which is Ala for Thr 188. This difference may allow XPT to accommodate xylulose.

As TgAPT and PfoTPT reside in the apicoplast and have the same substrate specificity (Brooks et al., 2010; L. Lim et al., 2010), it was expected that they have a very similar binding pocket. It was found that only one amino acid is different and that is an Asn in PfoTPT where TgAPT has a Tyr, while other amino acids in the binding pocket are similar in these two proteins. When these two proteins were compared with others it was found that the phosphate recognizing residues are similar, but there are differences in some positions. For example- for Phe 123. Gly 184, Thr 188, Cys 189, Phe 192 and Phe 263 in GsTPT2, TgAPT and PfoTPT both have Tyr, Val, Ala, Val, Met and Ser, respectively. Except for Thr 188 to Ala, the other differences with GsTPT2 are unique for these two proteins.


a)
Phe 192
Cys 189
Thr 188
His 185
Gly 184
Lys 204
Phe 263
Arg 363
Tyr 339
Lys 362
b)
Met 133
Val 130
Ala 129
His 126
Val 125
Lys 145
Tyr 287
Arg 311
Lys 310

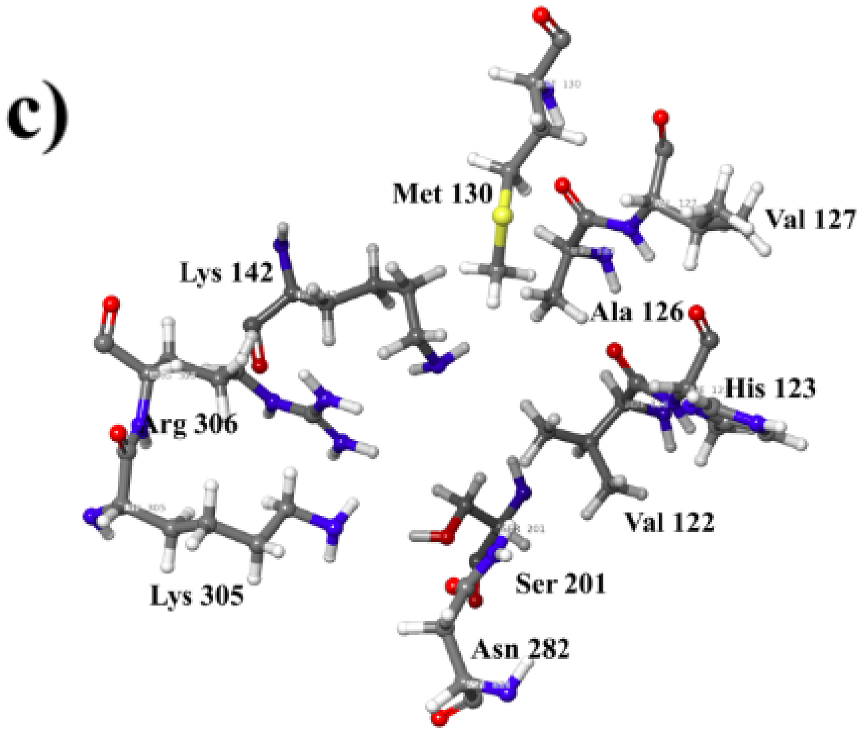
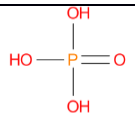
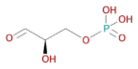
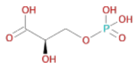
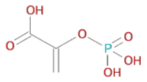
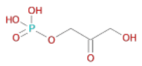
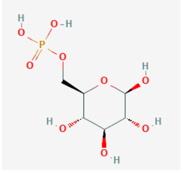


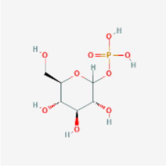
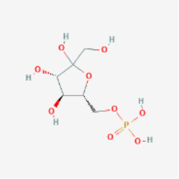
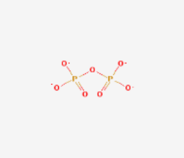
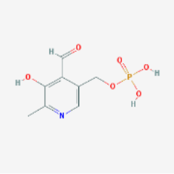
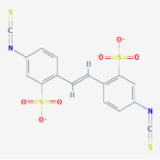
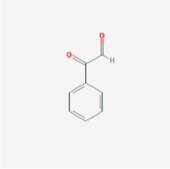
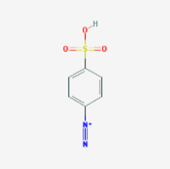
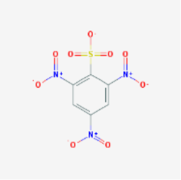
Figure 3. 3: **Differences of residues in the binding site of a) GsTPT2 b) TgAPT_5y79 c) PfoTPT**

3.3 Known substrates and inhibitors of TgAPT

Previous work regarding this protein and other members of this protein family detected several known substrates and inhibitors, which were sorted into three categories: 1) Substrate, 2) Known non-transported compound and 3) Inhibitors (Table 3.3). Substrates are those compounds, which are transported through this protein. Known non-transported are usually not transported, but in experimental condition they might be transported. For example- glucose-1-phosphate (Glc-1-P) and fructose-6-phosphate (Frc-6-P) are not transported either *in vitro* or *in vivo*, but glucose-6-phosphate (Glc-6-P) *can* be transported *in vitro*, but not *in vivo* (Brooks et al., 2010). The reason for that Pyrophosphate is considered in this group is that Lee et al. reported that the binding pocket of GsTPT2 cannot accommodate two phosphate group at the same time (Lee et al., 2017). From this, it was assumed that pyrophosphate was not supposed to be accommodated in TgAPT binding pocket either. The third category are inhibitors, which inhibit the transport process. Experiments of substrates and the non-transported compounds have been done on this protein (Brooks et al., 2010), and K_i values of substrates based on inhibition assay of phosphate transport are given in table 3.3. From this, the compounds can be ranked according to TgATP binding affinities, which is 3-Phosphoglyceric acid (3-PGA) > Triose phosphate > phohosphoenol pyruvate (PEP). For PfoTPT, the rank order is different, cause it has higher affinity for PEP as reflected on K_i value for competitive inhibition of [32 P]-Pi is 0.22 ± 0.03 mM, whereas for DHAP and 3-PGA that value is 1.53 ± 0.03 mM and 3.72 ± 0.40 mM, respectively (L. Lim et al., 2010). This can be an indication for how the rank order of scoring should ideally be when these molecules are docked. For inhibitors no published affinity data was found for TgAPT and PfoTPT, but some amino acid reagents like pyridoxalphosphate, 2,4,6-trinitrobenzene-sulfonate, which reacts with lysine residue, phenylglyoxal which reacts with Arginine residue and 4-sulfobenzenediazonium, which reacts with histidine and tyrosine were found to inhibit the phosphate translocation process of chloroplast phosphate translocator (Kenny, 1981). Therefore, these reagents are also considered as inhibitors of this TgAPT also. Beside this, 4,4’-diisothyanocyanostilbene-2, 2’-disulfonic acid (DIDS) is an inhibitor of this process (Gross, Brückner, Heldt, & Flügge, 1990).

Table 3. 3:Known substrates and inhibitors of TgAPT with the structures and docking score in the two homology models and the crystal structure. TP= triose phosphate, 3-PGA= 3-phosphoglyceric acid, PEP= phosphoenol pyruvate, DHAP= dihydroxy acetone phosphate, Glc-6-p = glucose-6-phosphate, Frc-6-p= fructose-6-phosphate, PP= pyrophosphate, PLP= pyridoxal 5’ phosphate, DIDS= Diisothiocyanostilbene-2, 2’ disulfonate, 4-SBD= 4-Sulfobenzenediazonium, TNBS= 2,4,6-trinitrobenzenesulfonate

Compound type	Name	Structure	Km or Ki In TgAPT	Score (Kcal/mol)			Reference
				TgAPT_5y78 model	TgAPT_5y79 model	PfoTPT	
Substrate	Phosphate		K_m = 1.39 ± 0.28 mM	-5.28	-6.30	-4.29	(Brooks et al., 2010)
	TP		K_i = 1.63 ± 0.26 mM	-6.53	-6.62	-6.84	
	3-PGA		K_i = 1.33 ± 0.49 mM	-6.49	-7.41	-6.56	
	PEP		K_i = 1.65 ± 0.52 mM	-5.69	-7.22	-5.72	
	DHAP			-6.80	-5.92	-5.79	
Known non-transported	Glc-6-P	 H O O-H P O O O H O H O O H O H		-6.15	-7.64	-7.48	(Brooks et al., 2010)

	Glc-1-P			-6.12	-5.16	-6.04	
	Frc-6-P			-6.48	-5.97	-7.23	
	PP			-6.89	-8.61	-6.90	
Inhibitors	PLP			-5.61	-6.89	-7.51	(U.-I. Flügge & Heldt, 1977)
	DIDS						(Gross et al., 1990)
	Phenylglyoxal			-6.37	-6.57	-6.64	(Kenny, 1981)
	4-SBD			-4.72	-5.03	-5.02	(Ulf.I. Flügge & Heldt, 1976)
	TNBS			-5.86	-5.88	-5.55	(Ulf Ingo Flügge & Heldt, 1978)

3.4 Docking of substrates and inhibitors

In order to assess the model quality and have an idea about the protein ligand interaction docking of the compounds in table 3.3 was done in the models. Among the homology models, the scoring and ranking of the compounds varied greatly. C-3 substrates comparatively scored better than the C-6 compounds and the inhibitors in TgAPT models. On the other hand, C-6 compounds scored better in PfoTPT models.

All the compounds except DIDS were docked in all models. One possible reason for that DIDS did not dock, may be the size (table 3.3). Among the TgAPT homology models, the compounds scored best in TgAPT_5y79 model. In both TgAPT models the highest and lowest scoring compounds were the same, which were PP and 4-SBD respectively. But, the expected rank order of scoring of the substrates was not found in any of the models. According to scoring values for 5y78 model c-3 compounds were ranked as DHAP (2nd), Gly-3-P (3rd), 3-PGA (5th), PEP (10th) and for 5y79 it is 3-PGA (3rd), PEP (4th), Gly-3-P (6th), DHAP (10th), none of which match with the expected 3-PGA > Triose-P > PEP. The C-6 compounds appeared as Frc-6-P (4th), Glc-6-P (7th), Glc-1-P (8th) for TgAPT_5y78 model and Glc-6-P (2nd), Frc-6-P (9th), and Glc-1-P (12th) for TgAPT_5y79 model. For the inhibitors, the order was: Phenylglyoxal (6th), TNBS (9th), Pyridoxal-5-P (11th), 4-SBD (12th) for the TgAPT_5y78 model and Pyridoxal Phosphate (5th), Phenylglyoxal (7th), TNBS (11th) and 4-SBD (13th) for the 5y79 model. In comparison to TgAPT models, the PfoTPT model had quite different results. The order of the C-3 compounds were Gly-3-P (5th), 3-PGA (7th), DHAP (9th), PEP (10th), and for C-6 compounds Glc-6-P (2nd), Frc-6-P (3rd), Glc-1-P (8th), while for inhibitors the order was Pyridoxal phosphate (1st), Phenylglyoxal (6th), TNBS (11th), 4-SBD (12th).

Next, the docking position of 3-PGA in these models were compared with the crystal structure template. It was found that, in none of the model, 3-PGA was docked in a position that was completely matching with the bound 3-PGA in crystal structure 5y79 (figure 3.3). From overall assessments, it was seen that TgAPT_5y79 scored better and showed better result in terms of differentiating between substrates and non- substrates than the TgAPT_5y78, although there was some ambiguity. But, the docking position of 3-PGA did not completely match with the crystal structure, which created some confusion. So, induced fit docking was done to identify better conformations of TgAPT for the interactions with the tested compounds.

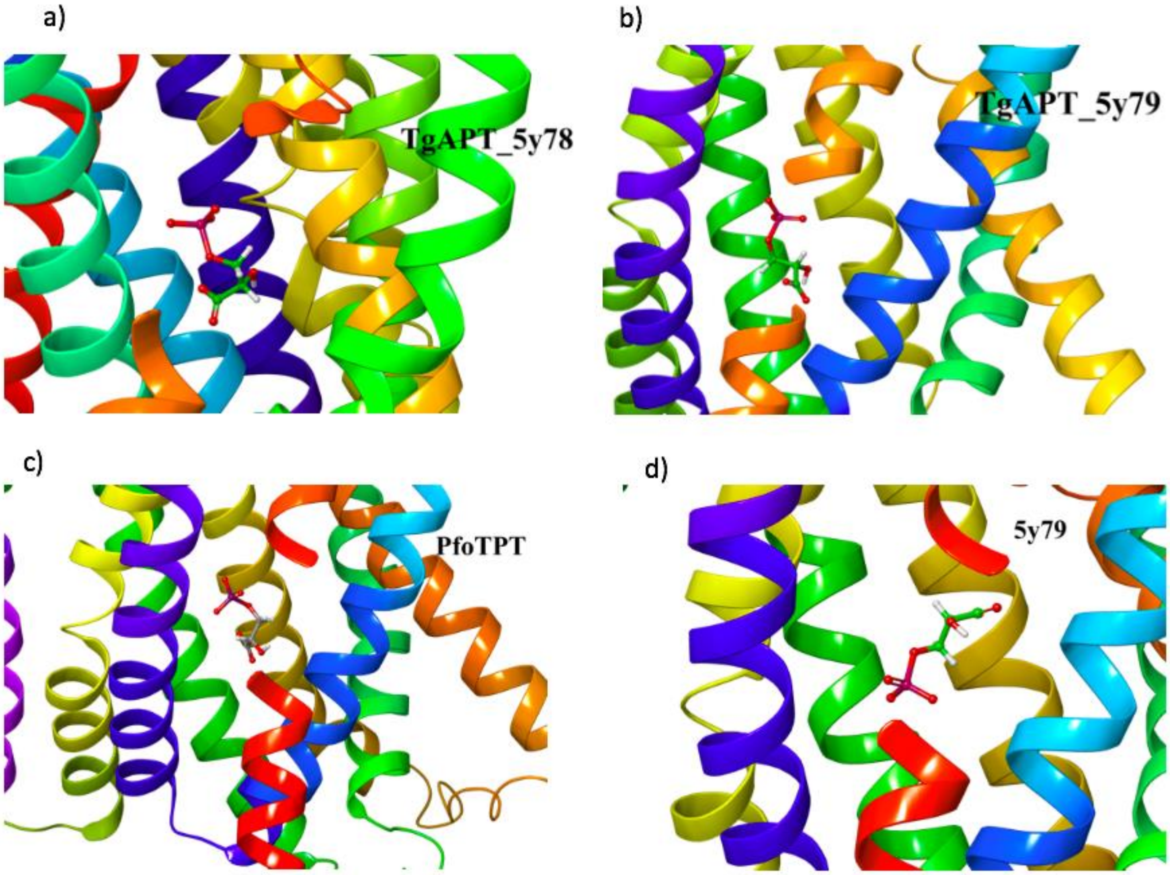


Figure 3. 4: **Position of 3-PGA in a) TgAPT_5y78 b) TgAPT_5y79 c) PfoTPT model d) Crystal structure 5y79.** In homology models it is docked position and in the crystal structure it is the crystallized bound position.

3.5 Induced fit in TgAPT_5y79

As mentioned in the method section, three combination of amino acids were tested for grid generation. It was found that grid generated with the four amino acids (His 126, Lys 145, Tyr287, Arg 311) produced better result in terms of scoring and docking position (result not shown here). Using this combination, induced fit was done on TgAPT with 3-PGA, PEP, PLP and TNBS. From these four runs, 6 poses from 3-PGA, PEP and TNBS were selected initially based on scoring value and docked pose. Their pose and interactions with neighboring amino acids are shown in fig 3.4, 3.5 and 3.6. From PLP docking, better poses were not obtained after induced fit docking.

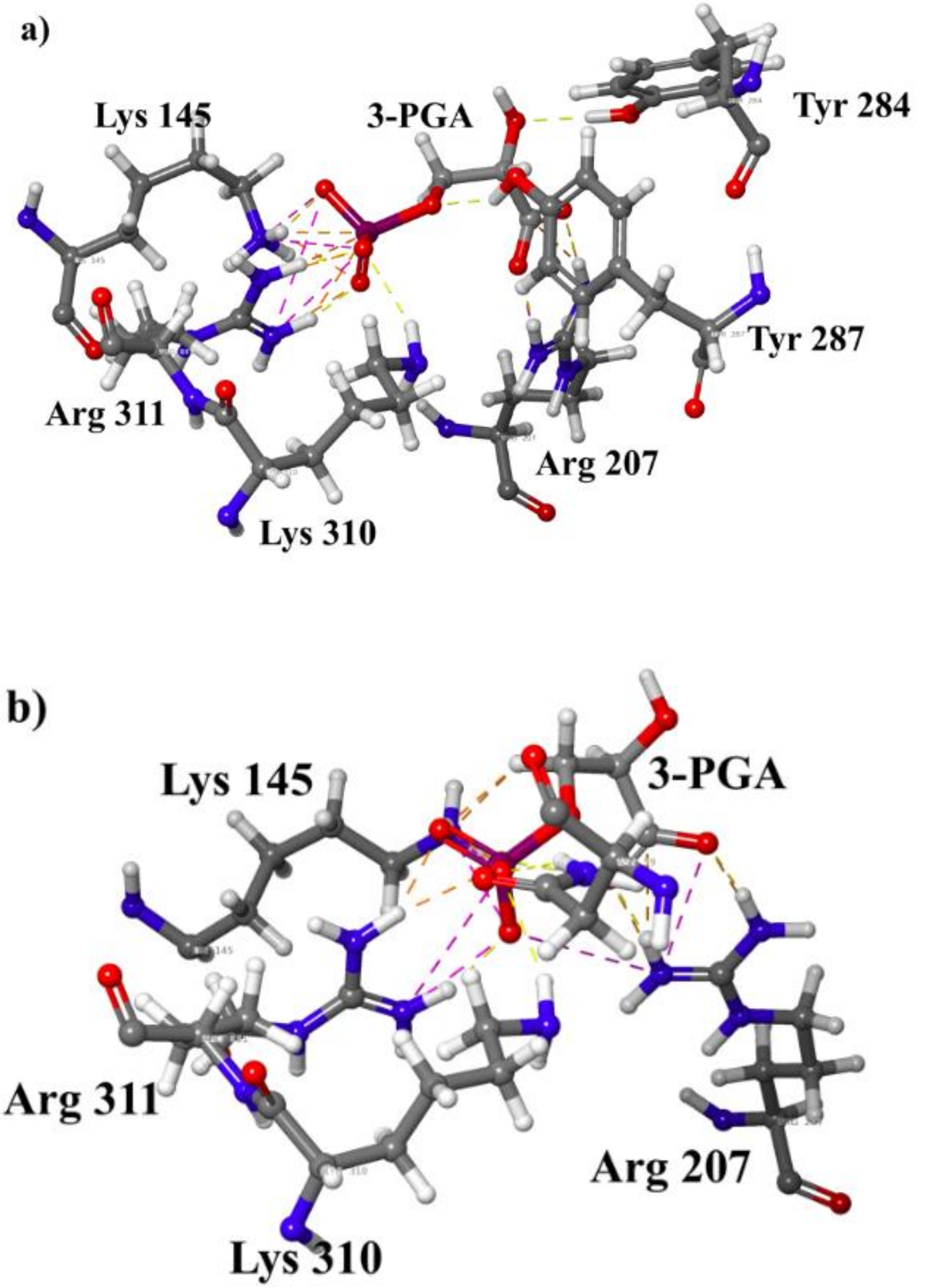


Figure 3. 5: **Two poses selected from induced-fit docking with 3-PGA.** a) 3-PGA_A:2, Lys 145, Arg 311, Lys 310 and Tyr 287 interacted with the Phosphate group, Tyr 284 form H-bond with OH of C-2 and Arg 207 interacted with carboxyl group b) 3-PGA_B_2, Close to 3-PGA_A_2 interaction, but no, Tyr 284 or Tyr 287 interaction.

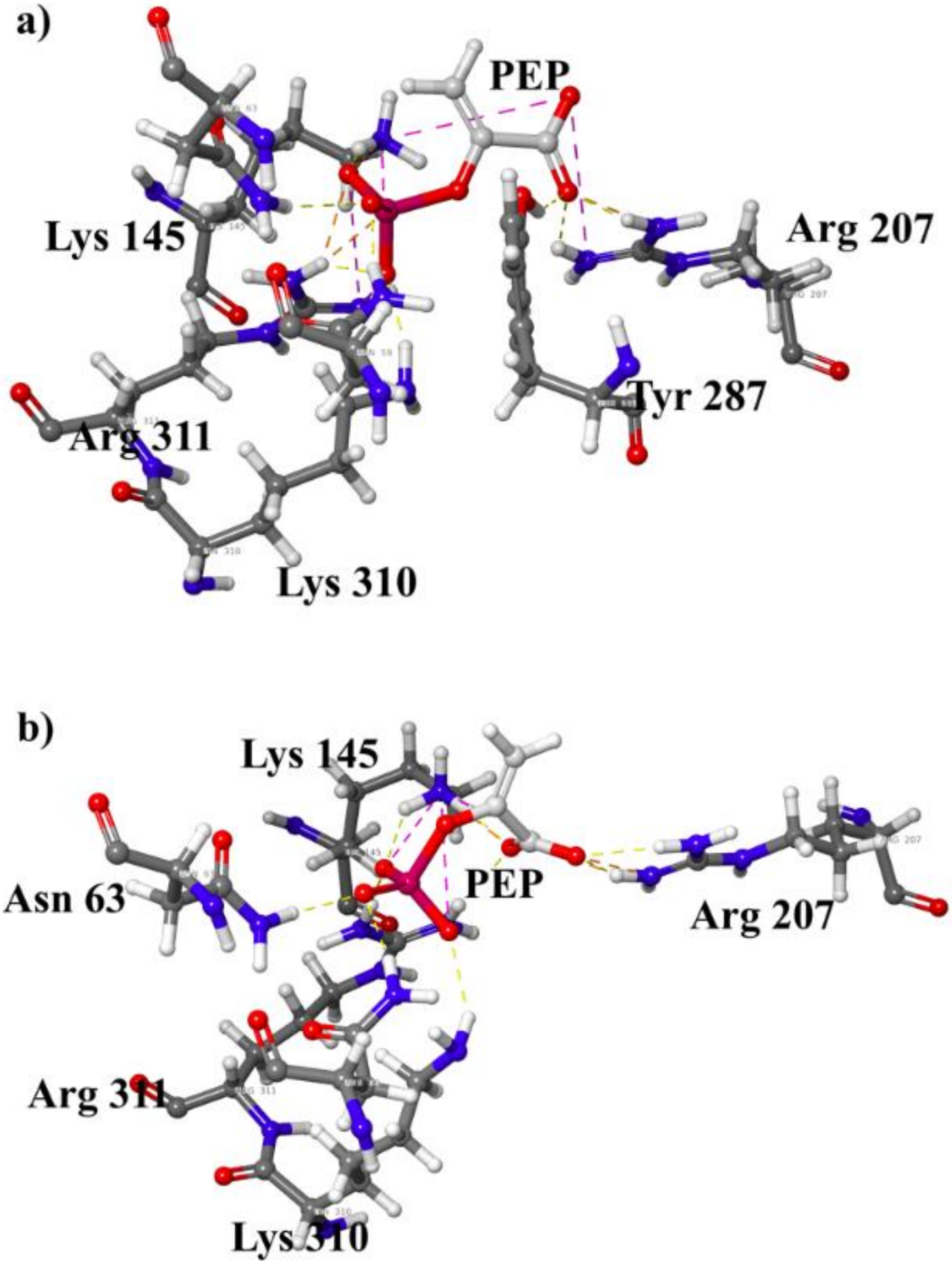


Figure 3. 6: **Two potential poses from induced fit docking with PEP.** a) PEP_A_3, Phosphate group has interaction with Asn 63 and Asn 59 along with Lys 145, Lys 310 and Arg 311, one of the Carboxyl O interact with Arg 207 and Tyr 287 and another O form salt bridge with Lys 145 b) PEP_A_11, it has also similar interaction except Tyr 287 interaction is missing

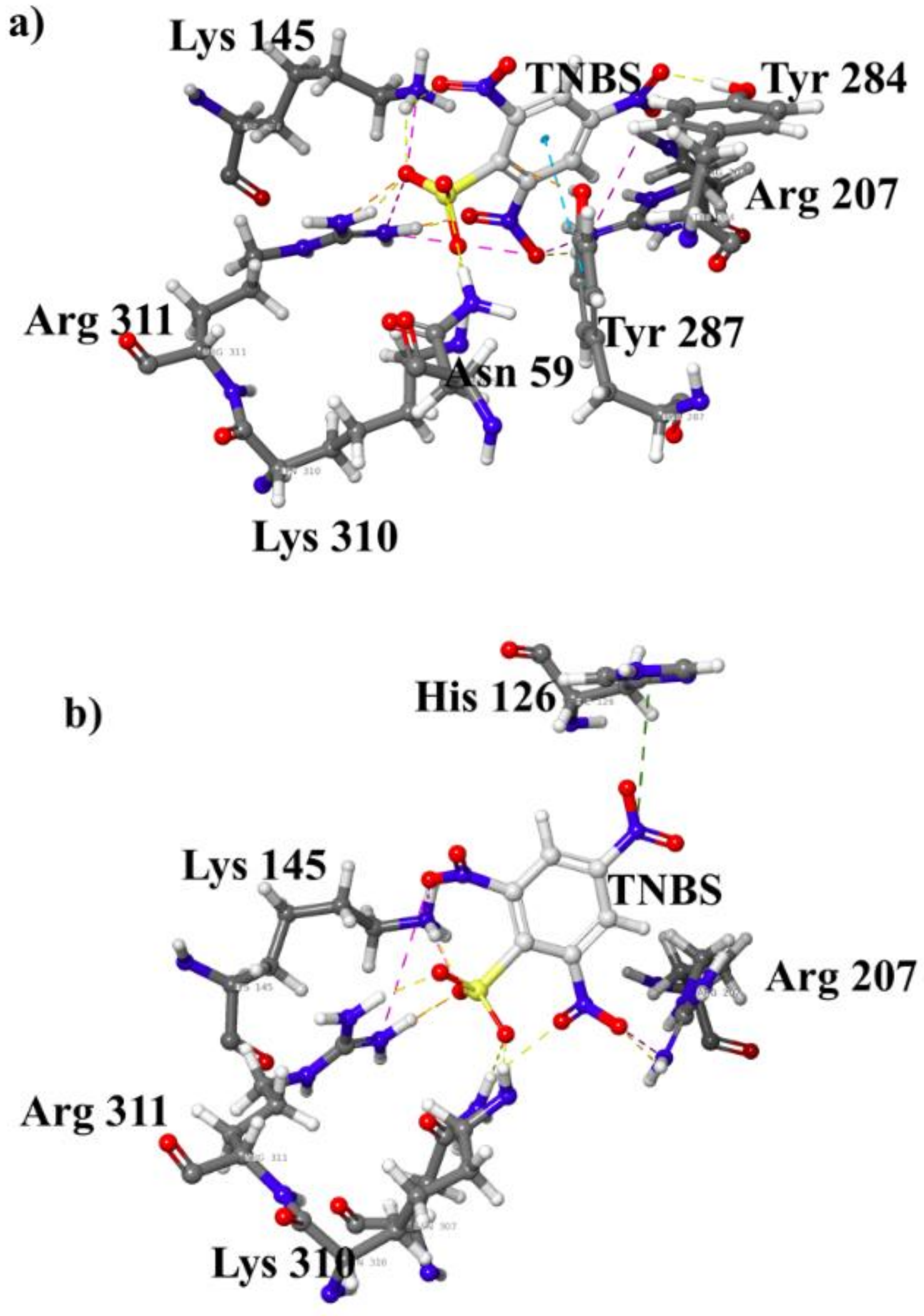


Figure 3. 7: Poses from induced fit docking with TNBS a) TNBS_1, Lys 145, Lys 310 and Arg 311 interacted with Sulphate group. Tyr 284, Arg 207, Arg 311 and Lys 145 inte b) ed with the nitro group and Tyr 287 also showed pi-pi stacking interaction with the Benzene ring. b) TNBS_8, Lys 145, Arg 311 and Asn 307 interacted with both sulphate group and one of the nitro group as well. Arg 207 interacted with nitro group and Tyr 287 form similar Pi-Pi stacking same as previous one.

The poses selected after induced fit were mainly based on protein-ligand interactions. In the selected, the phosphate group of 3-PGA and PEP and the sulfate group of TNBS were recognized almost in a similar manner by Lys 145, Lys 310 and Arg 311 and the other part of the substrates were extended to the opposite side (fig 3.4, 3.5 and 3.6). Like the phosphate recognizing residues there was another residue which was found to be interacting with all the compounds, which was Arg 207. Beside this, Tyr 287 and Tyr 284 were also seen to be common substrate binding residues.

Other interesting points to be observed were movements of side chains of amino acids in these selected poses compared to the initial positions in the model. In 3-PGA_A_2 conformation some movements of the side chain of Arg 207, Tyr 284, Asn 288, Tyr 287, Asn 307, His 122, Tyr 62 were seen. In 3-PGA_B_2 Tyr 62, His 122, Lys 145, Arg 207 and Tyr 287 side chains changed their position. In PEP_A_3 Arg 207 and Asn 288 and in PEP_A_11 Tyr 62, His 142, Lys 145, Arg 207, Tyr 284 and Asn 307 side chain movement was visible. In TNBS_1 His 122, Arg 207 and Tyr 287 and TNBS_8 His 122, Lys 145, Arg 207 and Asn 307 moved their side chains a bit.

3.6 Selection of docking pose for virtual screening

As mentioned in the method section, the substrates and inhibitors were re-docked in the selected poses after induced fit. The results are presented in table 3.4. After re-docking, one of the models with 3-PGA, 3-PGA_A_2, scored highest and had a better rank order of the compounds than the others. Although the rank order of the compounds is slightly deviating from the rank order from experimental studies (PEP ranked 2nd and Gly-3-P was 3rd), it was still the best ranking among the obtained model conformations. Interestingly DIDS also could dock in this model. Both models with PEP have some positives and negatives. PEP_A_3 model scored good, but in ranking frc-6-P came before Gly-3-P, which is a drawback of this model. Then, PEP_A_11 scored less than the two models with 3-PGA and PEP_A_3 and Glc-6-P and Frc-6-P scored better than PEP in this model. Of the six models tested here, these three models had good scoring value and closer to actual rank order of the compounds. Other models selected from induced fit (3-PGA_B_2, TNBS_1 and TNBS_8) had more deviant ranking of the compounds, although 3-PGA_B_2 scored good. In overall comparison, 3-PGA_A_2 showed the best results and that model conformation (docking pose) was therefore selected for virtual screening.

Table 3. 4: Docking result of compounds in poses selected from induced fit docking

Compounds name	Rank order based on experimental value	Score (Kcal/mol) in 3-PGA based poses		Score (Kcal/mol) in PEP based poses		Score (Kcal/mol) in TNBS based poses	
		A_2	B_2	A_3	A_11	_1	_8
3-PGA	1st K_i = 1.33 ± 0.49 mM	-9.90 Rank: 1st	-9.26 Rank: 1st	-9.89 Rank: 1st	-8.95 Rank: 1st	-7.67 Rank: 3rd	-7.67 Rank: 2nd
PEP	3rd K_i = 1.65 ± 0.52 mM	-9.58 Rank: 2nd	-7.51 Rank: 5th	-9.34 Rank: 2nd	-7.78 Rank: 7th	-6.50 Rank: 9th	-6.88 Rank:8th
Pyrophosphate		-8.49 Rank: 3rd	-8.17 Rank: 2nd	-8.30 Rank: 3rd	-8.08 Rank: 6th	-7.79 Rank: 2nd	-7.30 Rank: 5th
Gly-3-P	2nd K_i = 1.63 ± 0.26 mM	-7.96 Rank: 4th	-7.72 Rank: 4th	-7.32 Rank: 5th	-8.54 Rank: 3rd	-8.05 Rank: 1st	-6.94 Rank: 6th
Pyridoxal-5-P		-7.78 Rank: 5th	-6.91 Rank: 7th	-5.18 Rank: 11th	-8.14 Rank: 5th	-5.52 Rank: 12th	-6.90 Rank: 7th
Glc-1-p		-7.53 Rank: 6th	-6.71 Rank: 9th	-7.20 Rank: 7th	-7.22 Rank: 9th	-7.56 Rank: 5th	-7.32 Rank: 4th

DHAP		-7.38 Rank: 7th	-6.87 Rank: 8th	-6.46 Rank: 8th	-8.54 Rank: 2nd	-6.94 Rank: 7th	-6.78 Rank: 9th
Glc-6-p		-7.26 Rank: 8th	-7.21 Rank: 6th	-6.61 Rank: 6th	-8.47 Rank: 4th	-6.89 Rank: 8th	-7.92 Rank: 1st
Phosphate	K_m = 1.39 ± 0.28 mM	-7.10 Rank: 9th	-6.08 Rank: 10th	-6.05 Rank: 9th	-6.11 Rank: 11th	-6.25 Rank: 10th	-6.369 Rank: 11th
Frc-6-p		-6.49 Rank: 10th	-7.91 Rank: 3rd	-7.58 Rank: 4th	-7.29 Rank: 8th	-7.25 Rank: 6th	-6.31 Rank: 12th
TNBS		-5.89 Rank: 11th	-6.01 Rank: 11th	-5.17 Rank: 12th	-5.19 Rank: 13th	-7.56 Rank: 4th	-7.59 Rank: 3rd
Phenylglyoxal		-5.77 Rank: 12th	-5.89 Rank: 12th	-6.04 Rank: 10th	-6.47 Rank: 10th	-5.78 Rank: 11th	-6.44 Rank: 10th
4-SBD		-5.14 Rank: 13th	-4.56 Rank: 13th	-4.88 Rank: 13th	-6.47 Rank: 12th	-5.32 Rank: 13th	-5.78 Rank: 13th
DIDS		-3.93 Rank: 14th					

3.7 Virtual screening

In this study, ligand-based screening was done by searching the analogs of the substrates and inhibitors in the PubChem database and structure-based screening was done by docking the analogs in the TgAPT_5y79 model. Interestingly, the ligand-based step showed that inhibitors that contain ring structure had more analogs than the substrates with linear structure.

Table 3. 5: Analog search, docking and filtering of compounds. Similarity threshold is the parameter to dictate structural similarity between analogs and the parent compound. To sort the compounds from large number of docked analogs, Threshold scoring value was used and filtered compounds are number of compounds scored above the threshold value.

Compounds name	Similarity threshold	Downloaded analogs	After Ligprep	Threshold score	Filtered compounds
3-PGA	80%	469	2400	-9.0	90
Gly-3-P	80%	166	578	-7.0	16
PEP	80%	64	128	-7.0	15
PLP	80%	2267	11293	-8.0	40
DIDS	80%	1883	3626	-7.0	36
TNBS	80%	1848	2312	-7.0	25
4-SBD	80%	3267	4572	-7.5	40
Phenylglyoxal	90%	1426	1965	-6.0	56

Of all the compounds 4-SBD has the highest number of analogs within the search criteria. After Ligprep modification, PLP analogs gave the highest number of compounds. In the process of sorting out compounds for closer inspection, cut-off scoring values were used. Although scoring value is not an ideal parameter to justify affinity between protein and ligand, it is still a widely accepted tool to predict protein ligand interaction during virtual screening. As, it was not possible to inspect all the protein-ligand complex visually, therefore threshold scores were used here for initial screening. Different threshold scores for different set of analogs were used because those sets scored differently during docking and the parent compounds also had different scores. In some cases, comparatively lower score was used to have considerable number of compounds in that set. So, in this manner highest threshold score was used for 3-PGA analogs, which is -9.0 kcal/mol and above that score there were 90 compounds, which is the largest amount among all analog sets. For Gly-3-P and PEP analogs the same cut off score was used (-7.0 kcal/mol), and 16 and 15 compounds were filtered from each group respectively. For inhibitors, the highest number of compounds were analogs of both PLP and 4-SBD group (40 from each) above -8.0 kcal/mol and -7.5 kcal/mol cut-off score respectively. For DIDS and TNBS -7.0 kcal/mol cut-off score was used, giving 36 and 25 analog compounds. Finally, for Phenylglyoxal lowest cut-off score was used (-6.0 kcal/mol) and from this group 56 were sorted out.

3.8 Clustering

Isolated compounds in the previous step were clustered based on 2D similarity. Usually, one compound in a cluster represents some common features of all the compounds in that cluster. By this manner choosing one compound from a cluster can give an idea about the activity of that cluster. Output of clustering is summarized in table 3.6.

Table 3. 6: Clustering of sorted compounds from analogs docking

Analogs of	Merging distance	No. of clusters	Largest cluster
3-PGA	0.79	11	Cluster no: 3 Compounds: 44
Gly-3-P	0.82	3	Cluster no: 1 Compounds: 11
PEP	0.81	5	Cluster no: 1 and 2 Compounds: 5
PLP	0.80	7	Cluster no: 6 Compounds: 23

DIDS	0.68	5	Cluster no: 4 Compounds: 18
TNBS	0.75	7	Cluster no: 4 Compounds: 12
4-SBD	0.79	8	Cluster no: 4 Compounds: 27
Phenylglyoxal	0.81	9	Cluster no: 3 Compounds: 18

During clustering, the cluster number varies with the merging distance. It is a parameter to justify how similar the compounds are. If two compounds form a cluster with a short merging distance, that means those two compounds are highly similar. So, increasing the merging distance will reduce the number of clusters. In this study, relatively longer merging distances were used, so, a smaller number of clusters were generated. Longest merging distance was used for the Gly-3-P analog group and smallest for the DIDS group which were 0.82 and 0.68 respectively. In that way these two-group generated 3 and 5 clusters, respectively. After Gly-3-P, both PEP and Phenylglyoxal group with the same merging distance 0.81 generated 5 and 9 clusters respectively. 3-PGA and 4-SBD both groups also had the same merging distance of 0.79 and generated 11 and 8 clusters in respective manner. Then PLP group produced 7 clusters with a merging distance of 0.80 and TNBS group produced 7 clusters with 0.75 merging distance. Interestingly, almost all groups, except TNBS had 1 cluster with the majority of compounds. For example, in 3-PGA, PLP and 4-SBD groups all had 1 cluster containing 44, 23 and 27 compounds out of 90, 40 and 40 compounds in those particular group. In Gly-3-P group 11 compounds were in the same cluster out of 16 compounds. Phenylglyoxal and DIDS group also had large clusters with 18 and 18 compounds out of 56 and 36 compounds in respective group.

3.9 Predicting compounds for *in vitro* testing

As mentioned in the method only the best scoring compounds with good interactions were suggested for *in vitro* testing. To be more specific, compounds within the first 10 in terms of docking score in each group were chosen. For the ease of presentation, the compounds will be addressed according to their ranking within the particular group.

Total 29 compounds are selected of which 6 were from PLP group, 5 from 4-SBD group, 4 from each of DIDS, 3-PGA and PEP group, 3 from TNBS and Phenylglyoxal group. These compounds were selected from 5 clusters of the PLP group, 4 from each of DIDS, 4-SBD and PEP, 3 from TNBS and 2 from phenylglyoxal group.

Among the 3- PGA anlogs 1st and 2nd best scoring compounds were from cluster 4, 3rd best from cluster 5 and 7th from cluster 3 were selected. In these selected clusters 4, 5 and 3 there were 18, 18, 2 and 44 compounds respectively. Of the PEP analogs 1st, 2nd, 5th and 6th compound was chosen from cluster no. 3, 5, 2 and 4 and in these respective clusters number of compounds were 1, 2, 5 and 2.

Among the predicted inhibitors, PLP analogs were the highest scoring compounds. Of them, the 1st was from cluster 7 and the 2nd from the same cluster. Other than these two, 3rd, 5th, 6th and 10th from cluster 1, 4, 6 and 3 were also chosen and these clusters contained 1, 5, 23 and 2 compounds respectively. The analogs of DIDS did not scored as good as those of PLP, but still considerable for *in vitro* testing and from this group 1st, 2nd, 4th and 6th from cluster 1, 2, 4 and 3 were suggested. In these clusters the number of compounds was 6, 6, 18 and 4 respectively. Among the 5 compounds from 4-SBD group, 1st and 2nd best scoring compounds were from 4th cluster where 27 compounds gathered. Other than these, 4th, 5th and 9th were selected which were stacked in cluster no. 1, 8 and 6 along with 2, 3 and 3 compounds. From TNBS analogs 1st, 2nd and 3rd best scoring molecule from cluster no. 4, 3 and 7 were chosen and in these clusters, there were 12, 6 and 2 molecules respectively. Phenylglyoxal group had the lowest scoring analogs, but, still 3 compounds were selected. Of them 1st and 2nd were from cluster 3, which contain 18 compounds and the 3rd one is from 5th cluster which contain 6 compounds. List of the selected compounds are given in table 3.7:

Table 3. 7: List of predicted compounds

Analog of	Ranking of the compound	Score	Cluster no. Of the compound	Total compounds in the cluster
3-PGA	1st	-11.66	4	18
	2nd	-10.52	4	18
	3rd	-10.5	5	2
	7th	-10.36	3	44
PEP	1st	-10.80	3	1
	2nd	-10.11	5	2
	5th	-9.58	2	5
	6th	-9.57	4	2
PLP	1st	-10.26	7	7
	2nd	-9.71	7	7
	3rd	-9.15	1	1
	5th	-8.84	4	5
	6th	-8.77	6	23
	10th	- 8.65	3	2
DIDS	1st	-7.95	1	6
	2nd	-7.93	2	6

	4th	-7.78	4	18
	6th	-7.64	3	4
4-SBD	1st	-8.86	4	27
	2nd	-8.53	4	27
	4th	-8.39	1	2
	5th	-8.29	8	3
	9th	-8.08	6	3
TNBS	1st	-8.43	4	12
	2nd	-7.80	3	6
	3rd	-7.75	7	2
Phenylglyoxal	1st	-7.93	3	18
	2nd	-7.54	3	18
	3rd	-7.05	5	6

4 DISCUSSION

The purpose of this study was to predict potential inhibitors of TgAPT. There were two homology models of this protein available from a previous study and for better understanding of the binding site, a comparison was done with different subclasses of Phosphate translocator and a similar protein from Plasmodium falciparum apicoplast. In this process a homology model of PfoTPT was generated. Known substrates and inhibitors were docked in the TgAPT models and PfoTPT model. A reliable conformation (docking pose) of TgAPT was generated by induced fit docking and then analogs of the substrates and inhibitors were docked in that conformation. From this docking, compounds were predicted as inhibitors based on the scoring value and interaction modes.

4.1 Homology Modelling

Template for the model generation of the PfoTPT model was the same as the template for the TgAPT model (PDB id: 5y79). Sequence alignment showed it has 33% similarity with the template, which is less than for the TgAPT sequence. But it is above 30% similarity, which is the minimum requirement for a generated homology model to be comparable with an X-Ray structure of low resolution (Xiang, 2006). Actually as this APT is a membrane protein, it has little bit wider similarity window for generating a considerable good model (Ravna & Sylte, 2012). When the PfoTPT model was evaluated with different tools, it was evaluated to pass in some, got warning in some and failed in some.

Of the verification tools Verify3D assessed the compatibility of a 3D model with its own amino acid sequence (1D) using a 3D profile, computed from the atomic coordinates and then score is given for each amino acid based on a probability of observing that particular amino acid in the environment observed in the protein structure. The PfoTPT model failed this test as the model had only 53.09% residues averaged 3D → 1D score => 0.2, whether in case of a good model at least 80% residues do that. In the window plot, in some regions, residues are found to score below zero, which are residue 121 to 141 and 211 to 231. This indicates that the conformation is not correct in these region (Mannhold et al., 2008). Actually, both of these are in loop region, which is the most notorious region of a protein. And most of the part of helix 2 and helix 3 scored below 0.2, which indicates their lower conformation than the standard. It might because of the difference between template and target, which can lead to alignment error which leads to problem in identifying structurally equivalent residues despite their presence (Petrey & Honig, 2005). The Verify3D tool and other verification tools have been developed for checking the quality of 3D structures of soluble proteins and not membrane proteins. Most membrane proteins have in general more amino acids in helixes than soluble proteins and some evaluation tools may be a bit misleading for membrane proteins.

ERRAT calculates the statistical organizations of particular type of atom relative to each other and hence gives a ‘Overall Quality Factor’ for non-bonded interactions (Colovos & Yeates, 1993). By scoring 93.64 in this, the PfoTPT model passed the validation.

Another evaluation tool PROVE, which calculates z-score deviation for the protein by calculating the volume of the atom considering them as hard spheres (Pontius et al., 1996). The model here got warning in this tool as it has 60 outlier buried atoms (4.6%), which means these atoms have volume more than 3.0 standard deviation away from the mean of their particular type.

The PROCHECK gives an idea about the stereo-chemical characteristic of the protein model based on Ramachandran plot (Laskowski et al., 1993). The Ramachandran Plot showed that PfoTPT model has 94.4% and 4.9% in the favored and additional allowed region, which is an indication of a good model because a good model is supposed to have more than 90% of residues in allowed and favored region (Laskowski et al., 1993).

From overall assessment, it can be said that the quality of the PfoTPT model is satisfactory and can be used for docking and predicting protein-ligand interactions.

4.2 Comparison of binding sites among Phosphate translocators

From the alignment (Appendix 1) it is seen that the proteins in this class are mostly conserved in their binding pocket, but there are differences, which may explain differences subtype substrate specificity. A common feature of the substrates of these proteins is the phosphate, so, there should be commonality in the recognizing of phosphate group(s) between the transporters. In this study total 78 sequences of different subtypes from higher plants were compared and all of them contained the same amino acids that recognizes phosphate. Actually, not only the core phosphate binding, also the vicinities of these residues are conserved, which is consistent with a previous study (Lee et al., 2017). On the other hand, residues near the sugar moiety showed major differences. These differences along with their possible role in substrate recognition is discussed below in comparison with the GsTPT2 structure.

Although GsTPT2 and TPT proteins transport similar substrates there are three residues different between these two proteins. Of them two differences are not common, but tolerable, which are Gln 144 and Cys 189 in GsTPT2, which corresponds to His and Asn/Thr respectively in TPT. Another is Thr 188 to Ser in TPT. Thr 188 was found to be involved in hydrophobic interaction with the substrate (Lee. 2017), but transformation of this into Ser in TPT did not affect the substrate specificity. The reason for that must be the similarity between Ser and Thr.

The PPT binding site has 6 differences with the GsTPT2 binding site. Of them His 185 to Asn and Phe 263 to Asn in PPT are the two most important substitution compared to GsTPT2. His is chemically unique and involved in substrate binding. Replacing it with Asn surely affects the protein behavior. Similarly, Phe is a big aromatic amino acid, which may cause steric clashes with the branched side chain of PEP causing lower preference of PEP in other Phosphate translocators. So, replacing it with Asn in PPT allows the protein to accommodate PEP in the widened binding pocket. Other differences, Cys 189 to Asn, Phe 192 to Leu, Tyr 336 to Phe, Asn 340 to Gln may also have some impact on PEP transport.

GPT binds the largest substrates and should have a wider binding pocket than the others. From the binding site analysis, it was seen that two differences of GPT compared to GsTPT2 cause the widening. One is Ser in place of Gln 144 and another one is Ala in place of Thr 188. Ser and Ala both are smaller than Gln and Thr, respectively. Thr 188 was also seen to play a role in substrate recognition by hydrophobic interaction, so, transforming it into a very nonreactive Ala will also have some effect in GPT. Other than this, Cys 189 to Thr in GPT, which is a substitution into a similar amino acid, and Phe 192 to Met, which is a change of an aromatic hydrophobic to an aliphatic hydrophobic amino acid, may also contribute to differences in substrate specificity.

For XPT, only one amino acid difference was found with GsTPT2 and that was Thr 188 to Ala and it seems that this change is enough to accommodate Xylulose in the XPT binding pocket.

TgAPT and PfoAPT, both transport triose phosphates and PEP in natural conditions and also Glucose 6 Phosphate in experimental conditions although in low quantity (Brooks et al., 2010) From this, we can be assume that the binding pocket of these two proteins should have some commonality with TPT, PPT and GPT, especially the substrate recognizing residues. In this study, some of the characteristic residues are predicted above and in the following their commonality with two APTs will be mentioned.

In TPT, His 185 was assumed to be one of the substrate binding residues and in both APTs this His is present. In both APTs, like PPT, Phe 263 (GsTPT2 numbering) is transformed into a polar amino acid, although it is Ser in APTs and Asn in PPT. It was seen that this amino acid is very crucial in PEP recognition (Lee et al., 2017). So, substitution of this amino acid in both PPT and APTs resemble their similarity. The APTs also show similarity with GPT in position corresponding to Thr 188 and Phe 192 of GsTPT2, where both the APTs and GPT contain Ala and Met respectively. However, there are some residues which are unique to the APTs. For example: Phe 123, Gly 184, Cys 189 in GsTPT2 to Tyr, Val and Val, respectively, in both APTs. The first two are comparatively similar type of amino acid, although Val is larger than Gly, but in place of Cys 189 the uniqueness to APTs is that other PPTs contain polar residue on this site, whether Val in non-polar.

TgAPT and PfoTPT transport the same substrates, but with different affinity (Lim et al., 2010) and therefore should have some difference. Interestingly, only one amino acid was found to be different in these two proteins and that is in position of Tyr 339 in GsTPT2, where TgAPT contains Tyr like the others, but pfoTPT contains Asn. This Tyr 339 was seen to interact with the phosphate group by a hydrogen bond, so, substitution of with Asn in PfoTPT should be influential. Probably, it can give some idea about the differences in substrate affinity.

Based on the alignment (Appendix 1), it was tried to give an overview of the similarity and differences in the binding pocket of different PTs and correlate the differences with their substrate specificity. But, the role of the amino acids was mostly assumption as the structural data was not available. So, to identify the specific role of important amino acids, site-directed mutagenesis and subsequent transportation experiment should be done.

4.3 Docking of known substrates and inhibitors in the homology models

In order to get some idea about the behavior of homology models, known substrates and inhibitors were docked in the models. The output showed some unexpected result. According to Lee et al. 2017, the binding pocket of Phosphate translocator cannot accommodate two phosphate groups at the same time, but pyrophosphate not only docked, but also was among the high scoring compounds in all the models. To exclude the probability of having problem with the model, the same compounds were docked in the crystal structure 5y79, which was used as template for homology modelling, and a similar result was obtained found in case of pyrophosphate (Data not shown here). This result produced an ambiguity, but *in vitro* experiment may give a better insight of this problem.

The next problem of the docking was the ranking of the compounds. Experimentally, it was seen that TgAPT and PfoTPT does not readily transport Glc-6-P and other hexose phosphates (Brooks et al., 2010; L. Lim et al., 2010). But in the docking, Frc-6-P and Glc-6-P scored better than the actual substrate 3-PGA and PEP in TgAPT_5y78 model and Glc-6-P scored better than all other substrates in TgAPT_5y79 model. In PfoTPT model, Glc-6-P and Frc-6-P came in the 1st and 2nd position. Actually predicting binding affinities and rank them in order is one of the biggest methodological challenges in docking (Leach, Shoichet, & Peishoff, 2006). The scoring functions also have a lot of limitations. Using additional scoring functions can give better result. One of the reasons for incorrect prediction can be the condensed phases of biology in which it occurs and the degree of freedom of biomolecules (van Gunsteren & Berendsen, 1990). Beside this, the accuracy of the homology model itself might be an issue.

Another important purpose of docking was not fulfilled either, which is to predict the correct binding mode of the ligand. None of the models could perfectly generate a pose of 3-PGA similar to the crystallized pose in the template structure. The interaction found did not completely match with the crystal structure. The ionic interaction of His 126 with the carboxyl group of C-1 was not seen in any of the poses, although phosphate recognition was similar. During docking the receptor was considered rigid and the ligands were allowed flexibility, but in reality, during ligand recognition and binding the protein conformation is changed very often. This can be a factor for the ligand not to be docked in the expected position.

One point to be noted is the low scoring value of phosphate in all the models. There is a water molecule present in the binding pocket, which plays a crucial role during phosphate binding (Lee et al., 2017). That water molecule was not considered in the docking and may be a probably for that phosphate has low score in all models.

Overall, the docking gave an idea about the behavior of the models and TgAPT_5y79 showed better result than other models. So, this model was chosen for induced fit docking to find a better conformation for the actual screening.

4.4 Induced fit docking and selection of pose for virtual screening

The purpose of the induced fit docking was to find a conformation for virtual screening of unknown compounds by allowing some flexibility in the binding site of TgAPT_5y79 model. Although the goal was to find a pose that will mimic the interaction in the crystal structure, but none of the conformations could do that. So, from the binding modes which were close to those in the template, complexes were selected for further inspection.

In the selected poses, phosphate binding Lys 310 and Arg 311 did not move in regard to their position in the initial model. Only Lys 145 of the three phosphate binding residues moved during induce fit. On the other hand, Arg 207, which has been found to interact with ligands via ionic interaction or salt bridges, moved the side chain in every pose. Other than this, His, Tyr and Asn of different positions changed their side chain positions in the different poses. Observing these movements, the next question was which model conformation should be picked for the virtual screening.

To find the answer, initial substrates and inhibitors were docked again in each pose, and then based on their scoring values, interactions in the binding pocket and rank order of affinities, the 3-PGA_A_2 model was selected. In this model all compounds were docked including DIDS, which did not dock in any other, so it resembles the ability of the model to dock larger compounds beside smaller compounds. After that only PEP_A_3 had close result to 3-PGA_A_2 model, but ranking order was slightly distorted as PEP scored a bit lower and Frc-6-P scored higher and was ranked 4th. In other models, ligands had lower score and more deviant scoring rank order of the compounds than expected.

Although only based on this, it is not wise to exclude the possibility of other poses, but due to shortage of time and also for simplifying the study, only one model was chosen for the screening step.

4.5 Virtual screening

In this study structural analogs of both substrates and inhibitors of TgAPT protein were used in search for potential new inhibitors. The idea of using inhibitor analog came from the fact that structurally similar compounds most often possess similar functional activity. So, among the analogs of an inhibitor, it is possible that some compounds will have the same function as the parent compound with better affinity. On the other hand, from the analogs of substrates, molecules with better interaction and affinity with the protein can also be extracted, which will bind to the binding site, but not be transported.

In this study, out of 26874 docked molecules, 318 were extracted from the different groups of analogs based on their scoring values. Of them, most were analogs of 3-PGA. A reason for these compounds had better scores than others, may be that this pose was selected from induced fit docking of 3-PGA. From the other two substrate analog groups few compounds were extracted although lower cut-off scores were used. Actually, these two substrates had a smaller number of analogs in the first place.

For the inhibitors, the highest number of compounds were analogs of PLP and 4-SBD (40 from each group), using the cut-off score -8.0 Kcal/mol and -7.5 Kcal/mol for the respective groups. One interesting thing to notice was that, 4-SBD itself had low score, but its analogs scored better and even giving a threshold score higher than TNBS, DIDS, Phenylglyoxal, PEP and Gly-3-P, and a larger number of compounds were possible to extract. From the TNBS and DIDS groups also a considerable number of molecules were obtained. These two inhibitors had a quite similar number of analogs, but after Ligprep preparation the number of DIDS analogs increased since DIDS analogs have a higher number of enantiomers. So, although same threshold score was used, number of molecules in the DIDS group was higher.

For the phenylglyoxal group the lowest cut-off score of all was used, and 56 compounds scored better than the threshold. Although a good number of analogs were found under 90% similarity, but within this limit not that much modification can be expected compared with phenylglyoxal to increase the interaction capability.

4.6 Clustering of selected analogs

It is a common notion that similar structure tends to have similar properties. On the contrary, it is also true that slight change in the structure can lead to functional change of that compound (Zahoránszky et al., 2009). Despite this possibility, the extracted analogs were clustered based on their similarity. There are several fingerprint methods, which are used to cluster compounds. As for the compounds whose correct fingerprint type is uncertain, it is recommended to use MOLPRINT2D, which was used here (Duan et al., 2010; Sastry et al., 2010). Then the ligands were clustered by hierarchical clustering with relatively longer merging distance. In case of long merging distance there is a chance of less similar compounds getting into one cluster which can increase the chance of identifying more dissimilar to known compounds as inhibitors.

The clustering result showed that majority of the compounds of an analog group gather into one cluster, which indicates that most of the compounds are close to each other. And rest of the compounds get separated into several clusters, which is a reflection of their relative dissimilarity. For example, for the PLP and 4-SBD analog groups that both contained 40 compounds, 23 and 27 compounds from the respective groups gathered into one cluster, while the rest of the compounds dispersed into 6 and 7 clusters respectively. For DIDS analogs having lowest merging distance, out of 36 molecules 18 were gathered into one cluster and rest of the compounds diffused into 4 clusters indicating that there are significant differences among them. TNBS and Phenylgloxal analogs also have variation in their structure as seen in their cluster number. In case of 3-PGA, although it shows that there are 11 clusters, but 80 out of 90 compounds gathered into 3 clusters, which justifies their structural similarity. PEP and Gly-3-P analogs also had more similar compounds than dissimilar ones.

4.7 Selection of compounds

The docking score and the mode of interactions with the binding site were the main criteria for selecting compounds for experimental testing. Amino acids proved to be involved in phosphate recognition and binding are Lys 145, Lys 310, Arg 311, while His 126, Arg 207 and Tyr 287 were found important for recognizing and binding with the rest of the ligand. Site directed mutagenesis studies have confirmed their importance in the transport (Lee et al., 2017; Takemoto et al., 2018). Following this pattern of interaction and scoring values, 29 compounds were suggested for experimental testing.

One common feature of the selected compounds are negative ions contributed by mostly Oxygen atoms, which lead to the ionic interaction with the positively charged amino acids, especially Lys 145 and 310, Arg 207 and 311. That might be one reason for that more negative ions scored highest.

Closer inspection also revealed that all the selected compounds were found to have common interaction with Lys 145, Lys 310, Arg 311, Arg 207 and Tyr 287. Beside this some of the analogs were found to have additional interaction, for example- 4-SBD, DIDS and TNBS analogs showed interaction with Asn 288. In addition to the good interactions, some bad interactions were found in some of the analogs of DIDS and TNBS having bad contact with His 126 and Tyr 284.

Some of the clusters containing low scoring compounds were not selected despite having the probability of a good candidate scoring low in docking experiment. Keeping all these drawbacks under consideration compounds were finally selected. The compounds will be experimentally tested for their inhibition potential.

5 CONCLUSIONS

The main purpose of this study was to predict potential inhibitors against TgAPT. Before going to the actual screening, the binding site of TgAPT was compared with PfoTPT and pPT classes. This comparison revealed that the binding pocket of these protein share mostly common residues. There are also differences as well, which might contribute to their substrate specificity. Interestingly, TgAPT and PfoTPT were found to be very similar in their binding site, differ in only one amine acid. These both proteins have similar substrate specificity, so it can be assumed that inhibitor of TgAPT possibly inhibit PfoTPT.

Through LBVS and SBVS approaches, 29 compounds were finally predicted as potential inhibitor. These compounds showed good scoring with a 3D model of TgAPT, but it is not possible to be sure that all of them are true TgAPT binders.   To explore the possibility of these predicted compounds they needed to be tested experimentally.

7 FUTURE DIRECTION

Experimental tests will be done in the laboratory of Prof. Eva Pebay-Peyroula (University of Grenoble, France) who is a collaborator in this project. Inhibition is shown by two different experimental strategies. First, the compounds are analyzed for inhibition in biochemical transport assays of the APT protein which is integrated into artificial liposomes. If these results are found to be positive, then their ability to inhibit growth or kill the parasite is directly determined in cell cultures of host and Toxoplasma cells. In the case of a positive outcome from this test these compounds will be also tested in cell cultures with Plasmodium as parasite. By this manner, a foundation can be set for development of drug against toxoplasmosis and even malaria in next stage.

REFERENCES

Adl, S. M., Leander, B. S., Simpson, A. G. B., Archibald, J. M., Anderson, O. R., Bass, D., . . . Spiegel, F. (2007). Diversity, Nomenclature, and Taxonomy of Protists. *Systematic Biology, 56*(4), 684-689. doi:10.1080/10635150701494127

Banerjee, T., Jaijyan, D. K., Surolia, N., Singh, A. P., & Surolia, A. (2012). Apicoplast triose phosphate transporter (TPT) gene knockout is lethal for Plasmodium. *Mol Biochem Parasitol, 186*(1), 44-50. doi:10.1016/j.molbiopara.2012.09.008

Bisanz, C., Bastien, O., Grando, D., Jouhet, J., Maréchal, E., & Cesbron-Delauw, M.-F. (2006). &lt;em&gt;Toxoplasma gondii&lt;/em&gt; acyl-lipid metabolism: &lt;em&gt;de novo&lt;/em&gt; synthesis from apicoplast-generated fatty acids versus scavenging of host cell precursors. *Biochemical Journal, 394*(1), 197. doi:10.1042/BJ20050609

Black, M. W., & Boothroyd, J. C. (2000). Lytic cycle of Toxoplasma gondii. *Microbiology and molecular biology reviews : MMBR, 64*(3), 607-623.

Bordner, A. J. (2012). Force Fields for Homology Modeling. In A. J. W. Orry & R. Abagyan (Eds.), *Homology Modeling: Methods and Protocols* (pp. 83-106). Totowa, NJ: Humana Press.

Bradley, P., Misura, K. M., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science, 309*(5742), 1868-1871.

Brooks, C. F., Johnsen, H., van Dooren, G. G., Muthalagi, M., Lin, S. S., Bohne, W., . . . Striepen, B. (2010). The phosphate translocator is the source of carbon and energy for the Toxoplasma apicoplast and essential for parasite survival. *Cell host & microbe, 7*(1), 62-73. doi:10.1016/j.chom.2009.12.002

Cannon, J. G. (1996). An Introduction to Medicinal Chemistry By Graham L. Patrick. Oxford University Press, New York. 1995. xiv + 336 pp. 19.5 × 25 cm. ISBN 0-19-855872-4. $59.00. *Journal of Medicinal Chemistry, 39*(20), 4131-4132. doi:10.1021/jm960427+

Charron, A. J., & Sibley, L. D. (2002). Host cells: Mobilizable lipid resources for the intracellular parasite Toxoplasma gondii. *Journal of Cell Science, 115*(15), 3049-3059.

Chen, J., & Houk, K. N. (1998). Molecular Modeling: Principles and Applications By Andrew R. Leach. Addison Wesley Longman Limited: Essex, England, 1996. 595 pp. ISBN 0-582-23933-8. $35. *Journal of Chemical Information and Computer Sciences, 38*(5), 939-939. doi:10.1021/ci9804241

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal, 5*(4), 823-826.

Clastre, M., Goubard, A., Prel, A., Mincheva, Z., Viaud-Massuart, M.-C., Bout, D., . . . Laurent, F. (2007). The methylerythritol phosphate pathway for isoprenoid biosynthesis in coccidia: Presence and sensitivity to fosmidomycin. *Experimental Parasitology, 116*(4), 375-384. doi:https://doi.org/10.1016/j.exppara.2007.02.002

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein science : a publication of the Protein Society, 2*(9), 1511-1519. doi:10.1002/pro.5560020916

Coppens, I., & Vielemeyer, O. (2005). Insights into unique physiological features of neutral lipids in Apicomplexa: from storage to potential mediation in parasite metabolic activities. *International Journal for Parasitology, 35*(6), 597-615. doi:https://doi.org/10.1016/j.ijpara.2005.01.009

Duan, J., Dixon, S. L., Lowrie, J. F., & Sherman, W. (2010). Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling, 29*(2), 157-170. doi:https://doi.org/10.1016/j.jmgm.2010.05.008

Duszynski, D., Wilson, W., J. Upton, S., & D. Levine, N. (1999). *Coccidia (Apicomplexa: Eimeriidae) in the Primates and the Scandentia* (Vol. 20).

Eicks, M., Maurino, V., Knappe, S., Flügge, U.-I., & Fischer, K. (2002). The plastidic pentose phosphate translocator represents a link between the cytosolic and the plastidic pentose phosphate pathways in plants. *Plant physiology, 128*(2), 512-522. doi:10.1104/pp.010576

Eisenreich, W., Bacher, A., Arigoni, D., & Rohdich, F. (2004). Biosynthesis of isoprenoids via the non-mevalonate pathway. *Cellular and Molecular Life Sciences CMLS, 61*(12), 1401-1426. doi:10.1007/s00018-004-3381-z

Epstein, C. J. (1964). Relation of Protein Evolution to Tertiary Structure. *Nature, 203*(4952), 1350-1352. doi:10.1038/2031350a0

Ferreira, L. G., Dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular Docking and Structure-Based Drug Design Strategies. *Molecules, 20*(7), 13384-13421.

Fischer, K., Kammerer, B., Gutensohn, M., Arbinger, B., Weber, A., Häusler, R. E., & Flügge, U. I. (1997). A new class of plastidic phosphate translocators: a putative link between primary and secondary metabolism by the phosphoenolpyruvate/phosphate antiporter. *The Plant cell, 9*(3), 453-462. doi:10.1105/tpc.9.3.453

Flegr, J., Prandota, J., Sovičková, M., & Israili, Z. H. (2014). Toxoplasmosis--a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries. *PloS one, 9*(3), e90203-e90203. doi:10.1371/journal.pone.0090203

Fleige, T., Fischer, K., Ferguson, D. J. P., Gross, U., & Bohne, W. (2007). Carbohydrate metabolism in the Toxoplasma gondii apicoplast: localization of three glycolytic isoenzymes, the single pyruvate dehydrogenase complex, and a plastid phosphate translocator. *Eukaryotic cell, 6*(6), 984-996. doi:10.1128/EC.00061-07

Fleige, T., Limenitakis, J., & Soldati, D. (2010). *Apicoplast: keep it or leave it* (Vol. 12).

Flügge, U.-I., & Heldt, H. W. (1977). Specific labelling of a protein involved in phosphate transport of chloroplasts by pyridoxal-5′-phosphate. *FEBS Letters, 82*(1), 29-33. doi:10.1016/0014-5793(77)80878-8

Flügge, U. I., & Heldt, H. W. (1976). Identification of a protein involved in phosphate transport of chloroplasts. *FEBS Letters, 68*(2), 259-262. doi:10.1016/0014-5793(76)80449-8

Flügge, U. I., & Heldt, H. W. (1978). Specific labelling of the active site of the phosphate translocator in spinach chloroplasts by 2,4,6-trinitrobenzene sulfonate. *Biochemical and Biophysical Research Communications, 84*(1), 37-44. doi:https://doi.org/10.1016/0006-291X(78)90259-0

Forster, M. J. (2002). Molecular modelling in structural biology. *Micron, 33*(4), 365-384. doi:https://doi.org/10.1016/S0968-4328(01)00035-X

Goodman, C. D., Su, V., & McFadden, G. I. (2007). The effects of anti-bacterials on the malaria parasite Plasmodium falciparum. *Molecular and Biochemical Parasitology, 152*(2), 181-191. doi:https://doi.org/10.1016/j.molbiopara.2007.01.005

Gould, S. B., Waller, R. F., & McFadden, G. I. (2008). Plastid Evolution. *Annual Review of Plant Biology, 59*(1), 491-517. doi:10.1146/annurev.arplant.59.032607.092915

Grauvogel, C., Reece, K. S., Brinkmann, H., & Petersen, J. (2007). Plastid Isoprenoid Metabolism in the Oyster Parasite Perkinsus marinus Connects Dinoflagellates and Malaria Pathogens—New Impetus for Studying Alveolates. *Journal of Molecular Evolution, 65*(6), 725-729. doi:10.1007/s00239-007-9053-5

Gross, A., Brückner, G., Heldt, H. W., & Flügge, U.-I. (1990). Comparison of the kinetic properties, inhibition and labelling of the phosphate translocators from maize and spinach mesophyll chloroplasts. *Planta, 180*(2), 262-271. doi:10.1007/bf00194006

Heinemann, I. U., Jahn, M., & Jahn, D. (2008). The biochemistry of heme biosynthesis. *Archives of Biochemistry and Biophysics, 474*(2), 238-251. doi:https://doi.org/10.1016/j.abb.2008.02.015

Holstein, S. A., & Hohl, R. J. (2004). Isoprenoids: Remarkable diversity of form and function. *Lipids, 39*(4), 293-309. doi:10.1007/s11745-004-1233-3

Innes, E. A. (2010). A Brief History and Overview of Toxoplasma gondii. *Zoonoses and Public Health, 57*(1), 1-7. doi:10.1111/j.1863-2378.2009.01276.x

Institute, N. V. (2016). Toxoplasmosis. Retrieved from https://www.vetinst.no/sykdom-og-agens/toksoplasmose-toxoplasma-gondii

Jensen, K. D. C., Camejo, A., Melo, M. B., Cordeiro, C., Julien, L., Grotenbreg, G. M., . . . Saeij, J. P. J. (2015). Toxoplasma gondii superinfection and virulence during secondary infection correlate with the exact ROP5/ROP18 allelic combination. *mBio, 6*(2), e02280-e02280. doi:10.1128/mBio.02280-14

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., . . . Beck, E. (1999). *Inhibitors of the Nonmevalonate Pathway of Isoprenoid Biosynthesis as Antimalarial Drugs* (Vol. 285).

Kammerer, B., Fischer, K., Hilpert, B., Schubert, S., Gutensohn, M., Weber, A., & I Flügge, U. (1998). *Molecular Characterization of a Carbon Transporter in Plastids from Heterotrophic Tissues: The Glucose 6-Phosphate/Phosphate Antiporter* (Vol. 10).

Katris, N. J., van Dooren, G. G., McMillan, P. J., Hanssen, E., Tilley, L., & Waller, R. F. (2014). The apical complex provides a regulated gateway for secretion of invasion factors in

Toxoplasma. *PLoS pathogens, 10*(4), e1004074-e1004074. doi:10.1371/journal.ppat.1004074

Kenny, J. (1981). Function and molecular aspects of biomembrane transport: Edited by E Quagliariello, F Palmieri, S Papa, and M Klingenberg. Elsevier/North Holland Biomedical Press, Amsterdam and New York. 1979. $68.25. ISBN 0-444-80149-9. *Biochemical Education, 9*(1), 32-32. doi:10.1016/0307-4412(81)90071-6

Kim, K., & Weiss, L. M. (2004). Toxoplasma gondii: the model apicomplexan. *Int J Parasitol, 34*(3), 423-432. doi:10.1016/j.ijpara.2003.12.009

Kirby, J., & Keasling, J. D. (2009). Biosynthesis of Plant Isoprenoids: Perspectives for Microbial Engineering. *Annual Review of Plant Biology, 60*(1), 335-355. doi:10.1146/annurev.arplant.043008.091955

Klebe, G. (2006). *Virtual ligand screening: Strategies, perspectives and limitations* (Vol. 11).

Knappe, S., Flügge, U.-I., & Fischer, K. (2003). Analysis of the plastidic phosphate translocator gene family in Arabidopsis and identification of new phosphate translocator-homologous transporters, classified by their putative substrate-binding site. *Plant physiology, 131*(3), 1178-1190. doi:10.1104/pp.016519

Krieger, E., B Nabuurs, S., & Vriend, G. (2003). *Homology Modeling* (Vol. 44).

Lahana, R. (1999). How many leads from HTS? *Drug Discovery Today, 4*(10), 447-448. doi:https://doi.org/10.1016/S1359-6446(99)01393-8

Laskowski, R., Macarthur, M. W., Moss, D. S., & Thornton, J. (1993). *PROCHECK: A program to check the stereochemical quality of protein structures* (Vol. 26).

Layer, G., Reichelt, J., Jahn, D., & Heinz, D. W. (2010). Structure and function of enzymes in heme biosynthesis. *Protein science : a publication of the Protein Society, 19*(6), 1137-1161. doi:10.1002/pro.405

Leach, A. R., Shoichet, B. K., & Peishoff, C. E. (2006). Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. *Journal of Medicinal Chemistry, 49*(20), 5851-5855. doi:10.1021/jm060999m

Lee, Y., Nishizawa, T., Takemoto, M., Kumazaki, K., Yamashita, K., Hirata, K., . . . Nureki, O. (2017). Structure of the triose-phosphate/phosphate translocator reveals the basis of substrate specificity. *Nature Plants, 3*(10), 825-832. doi:10.1038/s41477-017-0022-8

Lichtenthaler, H. K. (1999). THE 1-DEOXY-D-XYLULOSE-5-PHOSPHATE PATHWAY OF ISOPRENOID BIOSYNTHESIS IN PLANTS. *Annual Review of Plant Physiology and Plant Molecular Biology, 50*(1), 47-65. doi:10.1146/annurev.arplant.50.1.47

Lim, L., Linka, M., Mullin, K. A., Weber, A. P., & McFadden, G. I. (2010). The carbon and energy sources of the non-photosynthetic plastid in the malaria parasite. *FEBS Lett, 584*(3), 549-554. doi:10.1016/j.febslet.2009.11.097

Lim, L., & McFadden, G. I. (2010). The evolution, metabolism and functions of the apicoplast. *Philosophical Transactions of the Royal Society B, 365*(1541), 749-763. doi:10.1098/rstb.2009.0273

Lionta, E., Spyrou, G., Vassilatis, D. K., & Cournia, Z. (2014). Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Current Topics in Medicinal Chemistry, 14*(16), 1923-1938. doi:10.2174/1568026614666140929124445

Lizundia, R., Werling, D., Langsley, G., & Ralph, S. A. (2009). Theileria apicoplast as a target for chemotherapy. *Antimicrobial agents and chemotherapy, 53*(3), 1213-1217. doi:10.1128/AAC.00126-08

Mannhold, R., Kubinyi, H., & Timmerman, H. (2008). *Molecular Modeling: Basic Principles and Applications* (Vol. 5): John Wiley & Sons.

Mazumdar, J., H Wilson, E., Masek, K., A Hunter, C., & Striepen, B. (2006). Apicoplast fatty acid synthesis is essential for organelle biogenesis and parasite survival in Toxoplasma gondii. *Proceedings of the National Academy of Sciences of the United States of America, 103*(35), 13192-13197. doi:10.1073/pnas.0603391103

McFadden, G. I., Reith, M. E., Munholland, J., & Lang-Unnasch, N. (1996). Plastid in human parasites. *Nature, 381*(6582), 482-482. doi:10.1038/381482a0

Moreno, S. N. J., & Li, Z.-H. (2008). Anti-infectives Targeting the isoprenoid pathway of Toxoplasma gondii. *Expert Opinion on Therapeutic Targets, 12*(3), 253-263. doi:10.1517/14728222.12.3.253

Mullin, K. A., Lim, L., Ralph, S. A., Spurck, T. P., Handman, E., & McFadden, G. I. (2006). Membrane transporters in the relict plastid of malaria parasites. *Proceedings of the National Academy of Sciences, 103*(25), 9572. doi:10.1073/pnas.0602293103

Oyakhirome, S., Issifou, S., Pongratz, P., Barondi, F., Ramharter, M., Kun, J. F., . . . Kremsner, P. G. (2007). Randomized controlled trial of fosmidomycin-clindamycin versus sulfadoxine-pyrimethamine in the treatment of Plasmodium falciparum malaria. *Antimicrobial agents and chemotherapy, 51*(5), 1869-1871. doi:10.1128/AAC.01448-06

Petrey, D., & Honig, B. (2005). Protein structure prediction: inroads to biology. *Mol Cell, 20*(6), 811-819. doi:10.1016/j.molcel.2005.12.005

Pontius, J., Richelle, J., & Wodak, S. J. (1996). Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology, 264*(1), 121-136. doi:https://doi.org/10.1006/jmbi.1996.0628

Ralph, S. A., van Dooren, G. G., Waller, R. F., Crawford, M. J., Fraunholz, M. J., Foth, B. J., . . . McFadden, G. I. (2004). Metabolic maps and functions of the Plasmodium falciparum apicoplast. *Nature Reviews Microbiology, 2*, 203. doi:10.1038/nrmicro843

https://www.nature.com/articles/nrmicro843#supplementary-information

Ravna, A., & Sylte, I. (2012). *Homology Modeling of Transporter Proteins (Carriers and Ion Channels)* (Vol. 857).

Repasky, M. P., Shelley, M., & Friesner, R. A. (2007). Flexible Ligand Docking with Glide. *Current Protocols in Bioinformatics, 18*(1), 8.12.11-18.12.36. doi:10.1002/0471250953.bi0812s18

Rohmer, M. (1999). The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants†. *Natural Product Reports, 16*(5), 565-574. doi:10.1039/A709175C

Roos, D. S., Kissinger, J. C., Fast, N. M., & Keeling, P. J. (2001). Nuclear-Encoded, Plastid-Targeted Genes Suggest a Single Common Origin for Apicomplexan and Dinoflagellate Plastids. *Molecular Biology and Evolution, 18*(3), 418-426. doi:10.1093/oxfordjournals.molbev.a003818

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics, 9*(1), 56-68. doi:10.1002/prot.340090107

Sastry, M., Lowrie, J. F., Dixon, S. L., & Sherman, W. (2010). Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *Journal of Chemical Information and Modeling, 50*(5), 771-784. doi:10.1021/ci100062n

Seeber, F., Feagin, J. E., & Parsons, M. (2014). Chapter 9 - The Apicoplast and Mitochondrion of Toxoplasma gondii. In L. M. Weiss & K. Kim (Eds.), *Toxoplasma Gondii (Second Edition)* (pp. 297-350). Boston: Academic Press.

Seeber, F., & Soldati-Favre, D. (2010). Chapter 5 - Metabolic Pathways in the Apicoplast of Apicomplexa. In K. W. Jeon (Ed.), *International Review of Cell and Molecular Biology* (Vol. 281, pp. 161-228): Academic Press.

Seeber, F., & Steinfelder, S. (2016). Recent advances in understanding apicomplexan parasites. *F1000Research, 5,* F1000 Faculty Rev-1369. doi:10.12688/f1000research.7924.1

Sousa, S. F., Fernandes, P. A., & Ramos, M. J. (2006). Protein–ligand docking: Current status and future challenges. *Proteins: Structure, Function, and Bioinformatics, 65*(1), 15-26. doi:10.1002/prot.21082

Striepen, B. (2011). *The apicoplast: A red alga in human parasites* (Vol. 51).

Takemoto, M., Lee, Y., Ishitani, R., & Nureki, O. (2018). Free Energy Landscape for the Entire Transport Cycle of Triose-Phosphate/Phosphate Translocator. *Structure, 26*(9), 1284-1296 e1284. doi:10.1016/j.str.2018.05.012

Tanaka, R., & Tanaka, A. (2007). Tetrapyrrole Biosynthesis in Higher Plants. *Annual Review of Plant Biology, 58*(1), 321-346. doi:10.1146/annurev.arplant.57.032905.105448

van Gunsteren, W. F., & Berendsen, H. J. C. (1990). Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition in English, 29*(9), 992-1023. doi:10.1002/anie.199009921

Xiang, Z. (2006). *Advances in Homology Protein Structure Modeling* (Vol. 7).

Zahoránszky, L. A., Katona, G. Y., Hári, P., Málnási-Csizmadia, A., Zweig, K. A., & Zahoránszky-Köhalmi, G. (2009). Breaking the hierarchy - a new cluster selection mechanism for hierarchical clustering methods. *Algorithms for Molecular Biology, 4*(1), 12. doi:10.1186/1748-7188-4-12

Zsoldos, Z., Reid, D., Simon, A., Bashir Sadjad, S., & Johnson, A. (2007). *eHiTS: A new fast, exhaustive flexible ligand docking system* (Vol. 26).

Appendix 1: Sequence comparison between the phosphate translocators

```
SoTPT  A-ASGS----- -S-GEAKT---- -GFLEKYPAL VTGSFFFMWY FLNVIFNILN KKIYNYFPYP--   50
McTPT  .A.E..----- D.A....V---- -...Q..... ...F...... .......... ..........--   53
VvTPT  .ST.SPAEGS- D.A.D..I---- -...D..... ...F...... .......... ..........--   57
VvTPT2 .A.AADADG-- --VV.PA----- KSLS.RF... ......MT.. .S.IV..... ..V.......--   54
VvTPT3 .A.AADADG-- --VTKPS----- KS.A..F.V. ...F...... .......... ..V.......--   54
PsTPT  .T.G.N----- D.A..E.VAP-- V..FSR.... T..F...T.. .......... ..........--   56
AtTPT  .-.--AAEGG- DTA.D..V---- -.......W. ...F...... .......... ..........--   54
BoTPT  .-..--AEGG- D.A..T.V---- -...G...W. ...F...... .......... ..........--   54
PtTPT1 .A..SPAEGS- D.S.DG.VAP-- I..F..N... ...F...... .......... ..........--   60
PtTPT2 .A..SPAEGS- D.S.DG.VAP-- V..FD..... ...F...... .......... ..........--   60
RcTPT1 .A..SPAEGS- D.S.DKVAPV-- -..F...... ...F...... .......... ..........--   59
RcTPT2 .A.-ADAEGH- ---V.PAA---- KS.G.RF... ...F...... .......... ..V.......--   54
StTPT  .A..S.AGSS- D.S.D..V---- -..FN.AT-. T..F...... .......... ..........--   56
NtTPT  TA..SPAEGS- D.A.D..V---- -..FN.AT-. I..F...... .......... ..........--   56
FpTPT  .T..------- D.A.D.--AP-- V..FA...F. ...F...... .......... ..........--   52
FtTPT  .T..------- D.A.D.--AP-- V...A...F. ...F...... .......... ..........--   52
TaTPT  .S.EP------ --A....S-P-- -.L......I T..F...... .......... ..........--   51
OsTPT1 .-.TS------ ---....PA--- -......... I..F...... .......... ..........--   49
OsTPT2 .SS.S--SSL- D.T....PV--- -..A.R.... ...F...... .......... ...FD.....--   56
SbTPT  .A.-------- E.A....----- -......... ...F...... .......... ..........--   49
SbTPT2 .A..SG----- -.A.D.EP---- Q..A.R..T. ...F...L.. .......... ...FD.....--   53
ZmTPT1 .A.E------- -.A....S---- V......... ...F...... .......... ..........--   51
ZmTPT2 .A.-------- ---....S---- V......... ...F...... .......... ..........--   48
ZmTPT3 .G..SG----- -PA....P---- Q..A.R.... ...F...L.. .......... ...FD.....--   53
PsiTPT .GTAD-AEGDE VFISSGLDKPS- QS.AD...W. I..F...... L......... ..........--   61
PpTPT  .S..D.SGDDP AEVAKEKKEEA- Q...A..... ...F...A.. ........M. ..........--   62
McPPT  .TSV--PES-- AGAD..------ PKAGGIGKT. EL.LL.GF.. LF.IY...Y. .QVLKV.H..--   53
VvPPT1 .SSV--PES-- AGES.------- -KSGNLVQT. QL.LL.GL.. LF.IY...Y. .QVLKVY.F.--   51
VvPPT2 .SSV--PEN-- AEET.------- -KSSNLGGI. QL..M.AI.. L..IY...F. .Q.LKVY.F.--   51
PsPPT  .T----SES-- AAES-------- ADSSSLLKT. QL..L.GL.. LF.IY...Y. .QVLKACHF.--   49
GmPPT1 .AE.AVPES-- APVE-------- ---NPLFKT. EL.AL.GL.. LF.IY...Y. .QVLKA.H..--   50
```

```
GmPPT2 .AS.I-PDA-- R.DE-----P-- AKTSDFLKTF QL.AM.AT.. L..IYY..Y. .QVLKVY.F.--   53
AtPPT1 .ATAV-PES-- AEE.D------- -NSGKLTKV. EL.LL.A... LF.IY...Y. .QVLKALHA.--   52
AtPPT2 .T--V-PEN-- VGGD-------- LESGSLVKG. KL.GM.GV.. L..IYY..F. .QVLRVY...--   52
BoPPT  .ATAV-PEE-- -GE--------- -.SGKMTKV. EL.LL.A... LF.IY...Y. .QVLKALHA.--   49
BnPPT  .ATAV-PEN-- AEE..------- -.SGKMTKV. EL.LL.A... LF.IY...Y. .QVLKALHA.--   52
PtPPT1 .T.V--PES-- AGE.DE------ -KSS-LVKT. EL.LL.GL.. LF.IY...Y. .QVLKV..N.--   51
PtPPT2 .T.V--PES-- AGE.KE------ -KSS-LTKT. EL.LL.GL.. LF.IY...Y. .QVLRV..N.--   51
RcPPT1 .T.V--PES-- AGES.------- -KSSSMIKT. EL.LL.GL.. LF.IY...Y. .QVLKV..N.--   51
RcPPT2 .ASV--PES-- T.QN.------- --TSDLARII QLAAM.GI.. L..IYY..F. .QVLKVY.F.--   50
NtPPT1 .-T.V-PES-- AGEA-----P-- -KSKPLTDT. .L..L.GL.. LF.IY...Y. .QVLKA.H..--   51
NtPPT2 VTS.E.PEI-- SAGE.E--PP-- -KSKPLADT. .L..L.GL.. IF.IY...Y. .QVLKT.H..--   55
FtPPT1 .ASV--PDK-- ADDGDAAALG-- -KS-KLVDT. FL..M.GL.. LF.IY...Y. .QVLKVL.S.--  107
FtPPT2 DSVVSRAAAS- ETSD.−SANP-- -AE--ISRI. QLAAM.GV.. L..IY...F. .QVLKV....--   56
OsPPT1 .ATAA------ AA-...GAEE-- --GGGLAKT. QL.AL.GL.. LF.IY...Y. .QVLKV....--   52
OsPPT2 .CGAA------ AGDAK.EE---- -EESGLAKT. QL.AL.GL.. LF.IY...Y. .QVLKV....--   52
OsPPT3 .VTARVAAAEA PLPADDADAAAG RERGALAETA QL.AMIVA.. L..IY...Y. .QVLQPL.F.--   63
OsPPT4 .AVATAAAAS- PPAEGGGKANGG AVAGGISRTV QL.AMILV.. L..IY...F. .LVLKSV.F.--   62
SbPPT1 .A.A.KVAAA- DTA..E------ -AGGGLAKT. QL.AL.GL.. LF.IY...Y. .QVLKVL...--   55
SbPPT2 .AGDAVAAPS- ---A.E------ --GGGFMKT. WL..L.GL.. LF.IY...Y. .QVLKV....--   51
SbPPT3 .V.AAAAA-S- VPADD.SAAAVT GDRGGIAATA QL.AMIVA.. L..IY...Y. .QVLGAL.L.LP   64
SbPPT4 .G.AAAA--S- PPAAGKPE---- -.AAGISRT. QL.AMILV.. L..IY...Y. .LVLKAI.F.--   55
ZmPPT1 .ASA------- GEE--------- -AGGGLAKT. QL.AL.GL.. LF.IY...Y. .QVLKVL...--   46
ZmPPT2 .A.D.A----- VEE--------- -AGGGLVKT. QL..L.GL.. LF.IY...Y. .QVLKVL...--   48
ZmPPT3 .AGDAVAAPK- AEE--------- --GGGLMKT. WL..L.GL.. LF.IY.H.Y. .QVLKV....--   51
PpPPT                              LAET. QL..L.GL.. MF.IC...Y. .QVLKV....--
McGPT1 .YEAD-GSEP- -IKP.PVPVP-- -IPG.AARKV KI.IY.AV.W A...V...Y. ..VL.A....--   57
McGPT2 .YEANRSQPLD -INI.L---P-- SVKS.TAKRV KI.IY.AT.W A...V...Y. ..VL.A....--   57
VvGPT1 .YEAE-RSQPL DLNI.LS-DQ-- EARS.AAQK. KI.IY.AT.W A...V...Y. ..VL.A....--   59
VvGPT2 .YEAD-RSEP- VES-DVVK---- -.RS.AAKKV KI.LY.AT.W A...V...Y. ..VL.A....--   55
PsGPT  .YEAD------ --RS.VEGGD-- GTPS.AAKKV KI.IY.AT.W A...V...Y. ..VL.AY...--   53
GmGPT  .YEAD------ --RS.VEGA--- STPS.AAKKV KI.IY.AT.W A...V...Y. ..VL.AY...--   52
MtGPT  .YEADRSQPLE -INIDIAGEQ-- -----AAQK. KI.LY.AT.W A...V...Y. ..VL.A....--   55
AtGPT1 .YEAD-RSEP- HPIGDDAAAA-- ETKS.AAKK. KI.IY.AT.W A...V...Y. ..VL.AY...--   59
AtGPT2 .YEAD-RSRP- LDI-NIEL-P-- DEQSAQ--K. KI.IY.AT.W A...V...Y. ..VL.A....--   54
ThGPT  .YEAD-RSQPI EIGI.IS----- DEQSRQ--KV KI.IY.AT.W A...V...Y. ..VL.A....--   54
StGPT  .YEASQPQ--- SIPIDIEFGQ-- EAQAAATQK. KI.LY.AT.W A...V...Y. ..VL.A..F.--   58
```

```
HaGPT  .YEA------- --GGDVV----- -ENT.AAKRV KI.FY.AT.W ......X.Y. ..VL.A....--   48
TaGPT  .S.AD----DK E.KA.VLPA--- --SS.AAQK. KISIY.AT.W A.......Y. ..VL.A....--   54
OsGPT1 .S.AD----DK E.KT.VVPV--- --RS.AAQK. KISIY.AT.W A.......Y. ..VL.A....--   55
OsGPT2 SATAD-GARP- VEAAP.GAAP-- ---E.AARRA KI.VY.AT.W A.......Y. ..VL.A....--   57
SbGPT  .S.AD----DK E.KTK.VPV--- --QS.GAQR. KISIY.AT.W A.......Y. ..VL.A....--   54
ZmGPT  .S.AD----DK E.KTQVVPV--- --QS.GAQR. KISIY.AT.W A.......Y. ..VL.A....--   54
PsiGPT .YEAS.SDLVS D.DV.EEVLSEN PSPQAAAQR. KI.IY.VA.W T...V...Y. ..VL.A....--   63
PpGPT  .YPE.---TPK VGDV.------- -VPKPAMRRV KI.IY.AT.W A...V...Y. ..VL.V..F.--   52
VvXPT  VAKAA------ ----.FEGES-- EVS-KPNKT. QL.IV.G... .Q.IV...Y. ..VL.L..F.--   50
GmXPT  IVKAA------ -SEANPEGENV- APTEPNSKN. KL.LV.GL.. .Q.IV...Y. ..VL.I..F.--   55
AtXPT  .AV-..SDSN- PDEKSDLGEA-- EKKEKKAKT. QL.IV.GL.. .Q.IV...F. ..AL.V....--   59
GsTPT2 AAV---DKSE- SGGSPQKSSV-- GVSPTLVHT. KV.FY.FL.. .F.F....A. .RTL.MWK..--
PfoTPT TFPITINEGYS DNVGDNKLKSK- GIYHKLFEK. KLALL.LT.. T...LY.VD. ..AL.MVKL.--
TgAPT  QYGTVSTGGAR PAKDLESQASP  ASGDQTAFYA QL.VMLLF.. A...MY.LD. .LALIML.L.--
```

```
SoTPT   YFVSVIHLFV GVVYCLASWS VGLPKRAPMD -----SKLLKLLIPV AVCHAIGHVT SNVSFAAVAV  110
McTPT   ........L. ..I...V..A ........I. -----GN........ .L...L.... ..........  113
VvTPT   .......... ......V..G ........I. -----.N........ .....L.... ..........  117
VvTPT2  R..AF...L. ..I...VC.. L.......I. -----KEF.L..T.. .F...L...M T.....S...  114
VvTPT3  ....L...L. ..A...V..A .......... -----KE..L..T.. .L...L...M ..........  114
PsTPT   ........A. ......V..T ........I. -----GN........ .....L.... ..........  116
AtTPT   .......... ......I... ........I. -----.N...V.... .....L.... ..........  114
BoTPT   .......... ......V... ........VN -----.DI..V.... .......... ..........  114
PtTPT1  .......... ......V..T ........I. -----.N..K..... .....L.... ..........  120
PtTPT2  .......... ......V..A .......... -----.N........ .....L.... ..........  120
RcTPT1  ........L. ......V..A ........I. -----.N........ .....L.... ..........  119
RcTPT2  ........L. ......T..G F.......I. -----RD..V..T.. .C...L...M ..........  114
StTPT   ........A. ......V..T ........I. -----.TQ....T.. .F...L.... ........R.  116
NtTPT   ........A. ......I..T ........I. -----.TQ....T.. .F...L.... ..........  116
FpTPT   ....A...A. ......GG.A .......... -----.N........ .F...L.... ..........  112
FtTPT   ........A. ......G..T ........V. -----.NI....... GF...L.... ..........  112
TaTPT   ........L. ......L..A ........IN -----AT.....F.. .L...L.... ......T...  111
OsTPT1  ........L. ......L..A ........IN -----ST.....F.. .L...L..A. ......T...  109
OsTPT2  .....S..L. ..L...VG.. F.......IN -----.TV....F.. .......... .T........  116
SbTPT   ....L...V. ..A...VG.. ........IN -----AN.....F.. .L..G..... ..........  109
SbTPT2  .....S...I ..L...IG.. F.I.....IN -----.T...Q.L.. .......... .T........  113
ZmTPT1  ....L...V. ......I... ........IN -----GT.....F.. .L..G...I. ..........  111
ZmTPT2  ....L...V. ......I... ........IN -----GT.....F.. .L..G...I. ..........  108
ZmTPT3  .....S...I ..L...IG.. F.I.....IN -----.T...QLV.. .......... .T........  113
PsiTPT  ........V. ..A...V... L.......I. -----KE..L..T.. .I...L...M T.....T...  121
PpTPT   ....A...A. ......I..M L.Y.....I. -----KE.FMM.... SI...L...M T.........  122
McPPT   VT.T..QFA. .S.LVGLM.L FN.Y..PKIS -----MGQ.AAIL.L ..V.TL.NLF T.M.LGK...  113
VvPPT1  VT.T.VQFA. .T.LVILM.G LN.Y..PKIS -----.SQ.VAIL.L ..V.TL.NLF T.M.LGK.S.  111
VvPPT2  AT.TAFQFGC .T.LVILM.A FN.Y..PKIS -----KSQFSGILIL ..T.TM.NLL T.L.LRK...  111
PsPPT   VT.T.VQFA. .T.LVSVM.A LN.Y..PKIN -----GAM.AAIF.L .IV.TL.NLF T.M.LGK...  109
GmPPT1  VT.T.VQFA. .T.LVAFM.G LN.Y..PKLS -----GAM.GAIL.L .AV.TL.NLF T.M.LGK...  110
GmPPT2  ATITAFQFGF ASLVINLV.T LN.HP.PSIS -----GSQFAAIL.L ..A.TM.NLL T.I.LGK...  113
AtPPT1  MT.TLVQFA. .S.LITIM.V LN.Y..PKIS -----GAQ.AAIL.L ..V.TL.NLF T.M.IGK.S.  112
AtPPT2  AT.TAFQ.GC .TLMIAIM.L LK.HP.PKFS -----PSQFTVIVQL ..A.TL.NLL T...LGR.N.  112
BoPPT   MT.TLVQFA. .S.LITFM.A LN.Y..PKIS -----AAQ.AAIL.L ..V.TL.NLF T.M.LGK.S.  109
BnPPT   MT.TLVQFA. .S.LITFM.A LN.Y..PKIS -----AAQ.AAIL.L ..V.TL.NLF T.M.LGK.S.  112
```

```
PtPPT1 VT.TAVQFA. .T.LVVFM.T FN.Y.KPKIS -----GAQ.AMIL.L ..V.TL.NLF T.M.LGK...  111
PtPPT2 VTITAAQFT. .T.LVACM.T FN.Y.KPKVS -----GAQ.AAIL.L ..V.TL.NLF T.M.LGK...  111
RcPPT1 VTITLAQFA. .T.LVTLM.T FN.Y..PKIT -----LAQ.AAIL.L .FV.TL.NLF T.M.LGK...  111
RcPPT2 AT.TAFQCGC .TLMIIIT.A LN.YHKPKLT -----RSQFTAIL.L ..A.TM.NLL T.I.LGK...  110
NtPPT1 VT.TLVQFR. .S.LVILM.T LN.Y..PKIS -----GAQ.VAIL.L ..V.TL.NLF T.M.LGK...  111
NtPPT2 VTITLAQ.A. .TILVIFM.T SN.Y..PKIS -----GAQ.AAIL.L ..V.TL.NLF T.M.LGK.S.  115
FtPPT1 ITITT.QFGI .G.IV.LM.. LN.H..PKVS -----GAQ.LAIL.L .MV.TL.NLF T.M.LGK...  167
FtPPT2 TTITTFQFGC .TLMV.IM.T LR.HPIPKFY -----KSQMVPVLVL ..A.TM.NLL T...LGK...  116
OsPPT1 INITNVQFA. .T.IA.FM.I T.IL..PKIS -----GAQ.AAIL.L .MV.TM.NLF T.M.LGK...  112
OsPPT2 INITTVQFA. .T.VA.FM.I T.ILR.PKIS -----GAQ.FAIL.L ..V.TM.NLF T.M.LGK...  112
OsPPT3 .TITAFQ.AF .SFVIFLM.A LKLHPAPRIS -----ISQ.AKIA.L .AG.ML.T.F T.M.LSK...  123
OsPPT4 .TITTFQFAS .SFFITLM.L LN.HPKPRLS -----LGQYAKIL.L .LV.TM.N.F T.M.LGK...  122
SbPPT1 INITTVQFA. .SAIA.FM.I T.IL..PKIS -----GAQ.FAIL.L .IV.TM.NLF T.M.LGK...  115
SbPPT2 INITEAQFA. .S.VS.FF.T T.II..PKIS -----GAQ.AAIL.L .IV.TM.NLF T.M.LGK...  111
SbPPT3 .TITAFQ.AF .SLLIFLM.A TR.HPVPRLS -----AAQ.GKIA.L ..G.ML.T.F T.M.LGK...  124
SbPPT4 .TITTF.FAS .SFFITLM.L LN.HPKPRLS -----L.QYAK.L.L .LI.ML.N.F T.M.LGK...  115
ZmPPT1 INITTVQFA. .SAIA.FM.I T.IH..PKIS -----GAQ.FAIL.L .IV.TM.NLF T.M.LGK..M  106
ZmPPT2 INITTVQFA. .SAIA.FM.I T.IL..PKIS -----GAQ.FAIL.L .IV.TM.NLF T.M.LGK...  108
ZmPPT3 INITEVQFA. .T.AA.FM.I T.II..PKIS -----GAQ.VAIL.L .IV.TM.NLF T.M.LGK...  111
PpPPT  ITITSLQFA. .A.IA.LT.F S..H..PQIS -----LAQ...IL.L .CV.TL.NLF T.M.LGK...
McGPT1 WLT.TLS.AM .SLIMVV..A TRIAEAPNT. -----.DFW.A.L.. ..A.T....A AT..MSK...  117
McGPT2 WLT.TLS.AA .SLMM.I..A SRVAHPPKT. -----LQFW.S.L.. ..A.T....A AT..MSK...  117
VvGPT1 WLT.TLS.AT .SLMM.I..A .RIAEPPKT. -----LDFW.T.F.. ..A.T....A AT..MSK...  119
VvGPT2 WLT.TLS.AT .SLMM.I..A TRIAETPKT. -----FAFW.T.F.. ..A.T....A AT..MSK...  115
PsGPT  WLT.TLS.AC .SLMM.I..A TRIAEAPKT. -----LEFW.T.F.. ..A.T....A AT..MSK...  113
GmGPT  WLT.TLS.AC .SLMM.IX.A T.IAEAPKT. -----PEFW.S.F.. ..A.T....A AT..MSK...  112
MtGPT  WLT.TLS.AA .SLIM.I..A TRVAEAPKVN -----LEFW.A.F.. ..A.T....A AT..MSK...  115
AtGPT1 WLT.TLS.AA .SLMM.I..A ..IVETPKT. -----FDFW.T.F.. ..A.T....A AT..MSK...  119
AtGPT2 WLT.TLS.AC .SLMM.V..A TRIADAPKT. -----LEFW.T.F.. ..A.T....A AT..MSK...  114
ThGPT  WLT.TLS.AC .SLMM.V..V TRVAEAPKT. -----LDFW.T.F.. ..A.T....A AT..MSK...  114
StGPT  WLT.TLS.AA .SLMM.V..A TKIAETPKT. -----FDFW.A.F.. ..A.T....A AT..MSK...  118
HaGPT  WLT.TLS.AA .SAIM.V..A SKVAEPPNT. -----VEFW.A.F.. .LA.T....A AT..MSK...  108
TaGPT  WLT.TLS.AC .SAMM.F..V TC.VEAPKT. -----LDFW.A.F.. ..A.T....A AT..MSK...  114
OsGPT1 WLT.TLS.AC .SAMM.V..A TR.VEAPKT. -----LDFW.V.F.. ..A.T....A AT..MSK...  115
OsGPT2 WLT.TLS.AA .SAIM....A TRIAEAPAT. -----LDFW.A.S.. .IA.T....A AT..MAK...  117
```

```
SbGPT  WLT.TLS.AC .SAMM.F..A TR.VEAPKT. -----LDFW.V.F.. ..A.T....A AT..MSK...  114

ZmGPT  WLT.TLS.AC .SAMM.F..A TR.VEAPKT. -----LDFW.V.F.. ..A.T....A AT..MSK...  114

PsiGPT WLT.TLS.A. .SLMMWV..A TR.VDAPDT. -----LEFW.A.A.. ..A.T....A AT..MSK...  123

PpGPT  WLT.TLS.AA .SAIM.I..A LRIVPAPDV. -----VEFW.G.A.A .LA.T....A AT..MSK...  112

VvXPT  WLLASFQ... .S.WM.IL.. FK.QPCPKIS -----KPFIVA.LGP .LF.T...IS AC...SK...  110

GmXPT  WLLASFQ... .SIWM.VL.. LK.QPCPKIS -----KPFIIA.LGP .LF.T...IS AC...SK...  115

AtXPT  WLLASFQ..A .SIWM.VL.. FK.YPCPKIS -----KPFIIA.LGP .LF.T...IS AC...SK...  119

GsTPT2  WVL.T.Q.G. .AL..TFL.V L..RTKPNVS -----K..I.A..WP SLG.TL..AA TCM..SL..I

PfoTP  WFI.SMQ.Y. .WIFIFIY.I S.MK.IPKIY SYDIFIRN---ILIQ S...IFV.FG AVMAMS.TS.

TgAPT  WT.STFQ..F .WLFFGFA.A T..RPVPRIH TTELFVTR---IA.Q GL..FFV.IG AVI.MGCG..
```

```
SoTPT   SFTHTIKALE PFFNAAASQF VLGQ-SIPITL WLSLAPVVIG VSMASLTELS FNWLGFISAM  170
McTPT   .......... .....S.... I...-P..... ........L. .A........ ...T......  173
VvTPT   .......... .......... I...-...L.. ........L. .......... ...I......  177
VvTPT2  .......... .......... ...H-Q..FP. ........F. .......... ...T......  174
VvTPT3  .......... .......... ...H-Q..FS. .......... .......... ...T......  174
PsTPT   .....V.... .......... I...-...... .......... .......... ..........  176
AtTPT   .......... .......... IM..-.....I ........L. .A........ ..........  174
BoTPT   .......... .....S.... L...-P..... ........L. .A........ ..........  174
PtTPT1  .......... .......... ....-...... ....L...L. .......... ...T......  180
PtTPT2  .......... .......... I...-Q..... ........L. ..V....... ...T......  180
RcTPT1  .......... .......... I...-...... .......... .......... ...I......  179
RcTPT2  .......... ...S...... ...H-Q..LS. .......... .......... ...T......  174
StTPT   .....V.... .......... I...-Q..LA. .......... .......... ......T...  176
NtTPT   .......... .....S.... I...-Q..LA. ........L. .......... ..........  176
FpTPT   .......S.. .......... I...-...... .......... .......... ..........  172
FtTPT   .......... .......... ....-....S. .......... .......... ..........  172
TaTPT   ..A....... .......T.. ....-TV.LS. ........L. .......... .S.K...N..  171
OsTPT1  ..A....... .L.....T.. ....-TV.LS. ........L. .......... .S.K...N..  169
OsTPT2  ..A....... .......... I...-QV.L.. .......... .......... ...T..VN..  176
SbTPT   ..A....... ...S...T.. I...-QV.LS. .M........ .......... ...T...N..  169
SbTPT2  ..A....... .......... I...-PV.L.. ....V...V. ..V....... ...T...N..  173
ZmTPT1  ..A....... ...S...T.. I...-QV.FS. .......... .......... ...T...N..  171
ZmTPT2  ..A....... ...S...T.. I...-QV.FS. .......... .......... ...T...N..  168
ZmTPT3  ..A....... .......... I...-PV.L.. ....V..... ..V....... ...T...N..  173
PsiTPT  .......... .....S.... ....-Q..F.. ........L. .......... ...T......  181
PpTPT   .......... ...S...... ....-..SLP. ....T.I.L. .....M.... ...K......  182
McPPT   ........M. ...SVVL.AM F..E-RPTPWV V...L.I.G. .AL..I..A. ...S..T...  173
VvPPT1  ........M. ...SVVL.AM F..E-FPT.WV LS..L.I.G. .AL..A..A. ...S..W...  171
VvPPT2  ........M. ...TVVLATL F..E-KPTLPI VS..V.I.G. .AL..F..S. ...T..W...  171
PsPPT   ........M. ...SVIL.AM F..E-RPT.WV IG..V.I.G. .AL..V..A. ...A..W...  169
GmPPT1  ........M. ...SVVL.AM F..E-FPTPWV VG..V.I.G. .AL..V..A. ...A..W...  170
GmPPT2  ........M. ...TVVL.AL L..E-MPTFWV VS..V.V.G. .AL..M..V. ...I..TT..  173
AtPPT1  ........M. ...SVLL.AM F..E-KPTPWV LGAIV.I.G. .AL..IS.V. ...A..S...  172
AtPPT2  ........M. ...TVLL.VL L..E-WPSLWI VC..L.I.A. ..L..F..A. ...I..C...  172
BoPPT   ........M. ...SVVL.AM F..E-VPTPWV IG.II.I.G. .AL..V..V. ...A..L...  169
BnPPT   ........M. ...SVVL.AM F..E-VPTPWV IG.II.I.G. .AL..V..V. ...A..L...  172
```

```
PtPPT1 ........M. ...SVVL.AM F..E-MPTLWV VG..L.I.G. .AL..V..A. ...A..W...  171
PtPPT2 ........M. ...SVVL.AM F..E-MPTLWV VG.II.I.G. .AL..V..A. ...A..W...  171
RcPPT1 ........M. ...SVIL.AM F..E-MPT.WV VG..V.IMG. .AL..A..A. ...A..W...  171
RcPPT2 ........M. ...TVLFASL F..E-RPSFWV LS..V.I.G. .AL..F..S. ..LT..C...  170
NtPPT1 ........M. ...SVVL.AM F..E-FPT.WV MS..V.I.G. .AL.....A. ...A..W...  171
NtPPT2 ........M. ...SVVL.AM F..E-FPTLWV IS..V.I.G. .GL.....A. ...A..W...  175
FtPPT1 ........M. ...SVLL.AM F..E-MPTPWV VG..L.IAG. .AL..M..A. ...A..W...  227
FtPPT2 ........M. ...TVVF.VL L.SE-RPTLWV FS..V.I.A. .AL..F..A. ...I..G...  176
OsPPT1 ........M. ...SVLL.AL F..E-MPTPFV V...V.I.G. .AL.....A. ...A..W...  172
OsPPT2 ........M. ...SVLL.AI F..E-LPTVWV I...L.I.G. .AL.....A. ...A..W...  172
OsPPT3 ........S. ...TVLL.A. F..E-TPSLLV LG..V.I.G. .AL....... ...I..W...  183
OsPPT4 ........M. ...SVLL.VL F..E-TPSFLV LG..V.I.G. .VL..M..V. ...I..W...  182
SbPPT1 ........M. ...SVLL.AI F..E-LPTPWV V...L.I.G. .AL..L..A. ...A..W...  175
SbPPT2 ........M. ...SVLL.AI F..E-FPTVWV VA..L.I.G. .AL..L..A. ...I..W...  171
SbPPT3 ........S. ...TVVL.AL F..E-VPSLPV LG..V.I.G. .AL..F..V. ...T..W...  184
SbPPT4 .....V..M. ...SVLL.VL F..Q-TPSLLV LG..V.V.G. .VL..M..V. ...I..W...  175
ZmPPT1 ........M. ...SVLL.AI F..E-LPTPWV V...L.I.G. .AL.....A. ...A..W...  166
ZmPPT2 ........M. ...SVLL.AI F..E-LPTPWV V...L.I.G. .AL.....A. ...A..W...  168
ZmPPT3 ........M. ...SVIL.AI F..E-LPT.WV VS..L.I.G. .AL.....A. ...A..W...  171
PpPPT  ........M. ...SVLL.AL F..D-MPNPMV VAT.V.I.G. .AL.....A. ...A..L...
McGPT1 ....I..SA. .A.SVLV.R. F..E-.FAAGV YW..V.IIG. CAL.AV...N ..MI..MG..  177
McGPT2 ....I..SG. .A.TVLV.R. L..D-TF.MPV YM..I.IIG. CAL.AV...N ..MI..MG..  177
VvGPT1 ....I..SG. .A.SVLV.R. L..E-TF.VPV YF..L.IIG. CAL.AV...N ..MT..MG..  179
VvGPT2 ....I..SG. .A.SVLV.R. L..E-.F.TSV YF..I.IIG. CAL.AV...N ..MI..MG..  175
PsGPT  ....I..SG. .A.SVLV.R. I..E-TF.VPV Y...L.IIG. CAL.AV...N ..MI..MG..  173
GmGPT  ....I..SG. .A.SVLV.R. L..E-.F.VPV Y...I.IIG. CAL.AV...N ..MI..MG..  172
MtGPT  ....I..SG. .A.SVLV.K. L..E-AF.LQV Y...L.IIG. CAL.AV...N ..MI..MG..  175
AtGPT1 ....I..SG. .A.SVLV.R. I..E-TF.TSV Y...I.IIG. CALSA....N ..MI..MG..  179
AtGPT2 ....I..SG. .A.SVLV.R. FM.E-TF.LPV Y...L.IIG. CAL.AI...N ..IT..MG..  174
ThGPT  ....I..SG. .A.SVLV.RL F..D-TF.LPV Y...L.IIG. CAL.AV...N ..MI..MG..  174
StGPT  ....I..SG. .A.SVLV.RL -..E-TF.LPV Y...L.IIG. CGL.AI...N ..LI..MG..  178
HaGPT  ....I..SG. .A.SVLV.R. I..E-TF.TSV Y...L.IIG. CGL.A....N ..MT..MG..  168
TaGPT  ....I..SA. .A.SVLV.R. I..E-.F.MPV Y...L.IIG. CGL.AA...N ..MI..MG..  174
OsGPT1 ....I..SA. .A.SVLV.R. L..E-TF.VPV Y...L.IIG. CGL.AV...N ..MV..MG..  175
OsGPT2 ....I..SG. .A.SVLV.R. F..E-HF.APV YF..L.IIG. CAL.AI...N ..MI..MG..  177
```

```
SbGPT   ....I..SA. .A.SVLV.R. I..E-TF.VPV Y...L.IIG. CAL.AV...N ..MV..MG..  174

ZmGPT   ....I..SA. .A.SVLV.R. F..E-TF..PV Y...L.IIG. CAL.AV...N ..MV..MG..  174

PsiGPT  ....I..SA. .A.SVLV.R. I..E-.F.MPV Y...L.IIG. CAL.AA...N ..MT..MG..  183

PpGPT   ....I..SA. .A.SVIIQRL L..E-DF.LPV Y...L.I.G. CGL.AA...N ..MT..VG..  172

VvXPT   ....V..SS. .V.SVIF.T- I..DNTY.LRV ...IL.I.L. C.L.AV..V. ..LQ.LWG.L  170

GmXPT   ....V..SA. .V.SXMF.S- ...D-KY..QV ...IL.I.L. C.L.AV..V. ..VQ.LWC.L  174

AtXPT   ....V..SA. .V.SVIF.S- L..D-.Y.LAV ...IL.I.M. C.L.AV..V. ..LG.LSG..  178

GsTPT2  ....VV.SA. .V.G.VG.AL ...EFFHPLT- Y.T.V.I.S. .ALSAA...T .T.T...T..

PfoTPT  ....VV..C. .V.T.IF.IL L.K.YLK.NK- YIA.LII.G. .VC..MK..H .T.IA.WC.T

TgAPT   ....IV..S. .VLT.LL.GL A.H.VFSWQT- Y...V.I.A. .I...V.... .T.KA.GC.L
```

```
SoTPT  ISNVSFTYRS LYSKKAMT-DM -------DST--NIYA YISIIALFVC ---LPPAIIV EGPQLMKHGF NDAIAK---  230
McTPT  ...I...... I.......-.. -------...--.V.. ..T....... ---I...L.I .....I.Y.. ......---  233
VvTPT  ...I...... I.......-.. -------...--.... .......I.. ---I...L.. .......... ......---  237
VvTPT2 VA.FA..... ..L.....-G. -------..A--.VC. .TAM...VF. ---F...LLI D.....Q... R.....---  234
VvTPT3 ...IA..... I.......-G. -------...--.V.. .T.....LF. ---I...VLI ......QY.. R.....---  234
PsTPT  ...I...... I.......-.. -------...--.... .......I.. ---I...L.I ...T.L.T.. ......---  236
AtTPT  ...I...... IF......-.. -------...--.V.. .......... ---I...... ...K.LN... A.....---  234
BoTPT  ...I...... IF......-.. -------...--.V.. .......... ---....... .....L.... ......---  234
PtTPT1 ...I...... ........-.. -------...--.... .......... ---I....L. .....I.... ......---  240
PtTPT2 ...I...... I.......-.. -------...--.... .......I.. ---I.....L .....I.... S.G...---  240
RcTPT1 ...I...... I.......-.. -------...--.... .......... ---I.....F .......Y.. ......---  239
RcTPT2 ...IA..... I.......-G. -------...--.V.. .......LF. ---I...VLI ...K..QY.. R...S.---  234
StTPT  ...I...... I.......-.. -------...--.V.. .......IF. ---...S.FI .....LQ... ......---  236
NtTPT  ...I...... I.......-.. -------...--.V.. .......I.. ---I.....I .....LQ... A.....---  236
FpTPT  ...I...... I.......-.. -------...--.L.. .....S.LF. ---I.....L .....L.... S.....---  232
FtTPT  ...I...... I.......-.. -------...--.L.. .......LF. ---I...VLF .....L.... ......---  232
TaTPT  ...I...... I.......-.. -------...--.V.. .......V.. ---I...L.I ......QY.L ......---  231
OsTPT1 ...I...... I.......-.. -------...--.V.. .......V.. ---I...V.I .....VQY.L ......---  229
OsTPT2 ...I...L.. V.......-.. -------...--.L.. .......L.. ---I.....I .....VQ... K.....---  236
SbTPT  ...I...... I.......-.. -------...--.V.. .......I.. ---I.....F ......S... S.....---  229
SbTPT2 ...I...... I.......-.. -------...--.L.. .......... ---I...L.I ......Q... K...G.---  233
ZmTPT1 ...I...... I.......-.. -------...--.V.. .......I.. ---I...L.F ...K..Q... S.....---  231
ZmTPT2 ...I...... I.......-.. -------...--.V.. ...?...I.. ---I...V.F ...R..Q... S.....---  228
ZmTPT3 ...I...... I.......-.. -------...--.L.. .......... ---I.....I .....VQ... K.....---  233
PsiTPT ...IA..... I.......-G. -------...--.V.. ........F. ---......I ...K..QS.. A.....---  241
PpTPT  TA..A....N I.......-G. -------...--.L.. .....S.AL. ---I.....I ...A.LNS.. S...T.---  242
McPPT  A...TNQS.N VL...L.V-KK DVD----QESMD..T- LF...TVMSF ILLA.A.YFM..VKFTPTYLEA.G-----  238
VvPPT1 A..LTNQS.N VL...F.I-KK -------EDSLD..T- LF...TIMSF ILLA.VS.FM..INFTPSYLQS.G-----  233
VvPPT2 A..LTNQS.N VF...F.V-NK -------EEALDT.N- LF.V.TVISF LLCT.V..FI ..IKFTPSYL QF.ASQG--  236
PsPPT  A...TNQS.N VL...V.V-KQ -------EESLD..T- LF...TIMSF FLLA.A..FM ..VKFTPAYL QS.G-----  231
GmPPT1 A...TNQS.N VL...A.V-NK -------EDSMD..T- LF...TVMSF FLLA.V..FM..VKFTPAYLQS.G-----  233
GmPPT2 A...TNQS.N VL...L.T-NE -------EETLD..N- LY.V.TIISF LLLV.C..L. ..VKFSPSYL QS.ASQG--  238
AtPPT1 A..LTNQS.N VL...V.V-KK -------.DSLD..T- LF...T.MSL VLMA.VTFFT ..IKFTPSYI QS.G-----  234
AtPPT2 A...TNQS.N VL...F.V-GK --------DALD..N- LF...TIISF ILLV.L..LI D.FKVTPSHL QV.G-----  233
BoPPT  A..LTNQS.N VL...V.V-KK -------.DSLD..T- LF...T.MSL FLMA.VTFFS ..IKFTPSYI QS.G-----  231
BnPPT  A..LTNQS.N VL...V.V-KK -------.DSLD..T- LF...T.MSL FLMA.VTFFS ..IKFTPSYI QS.G-----  234
```

```
PtPPT1 A..LTNQS.N VL...V.V-KN -------EESMD..T-LF...TIMSL VLLA.VT.FM ..VKFTPAYL QS.G-----  233
PtPPT2 A..LTNQS.N VL...V.L-KK -------EESMD..T-LF...TIMSF ILLA.VT.FM ..VKFTPAYL QSVG-----  233
RcPPT1 A..LTNQS.N VL...V.V-KK -------EDSID..T-LF...TIMSF FLLT.V.L.M ..VKFTPAYLQS.G-----  233
RcPPT2 A...TNQS.N VL...F.V-SK -------EEALD.VN-LF.V.TIISF ILLA.T.VVM ..IKFTPSYLQS.ANHG--  235
NtPPT1 A..LTNQS.N VL...F.V-RK -------EDSLD..T-LF...TIMSF FLLA.Y.FFA ..VKFTPAYLEA.G-----  233
NtPPT2 AC.LTNQS.N VL...F.V-RK -------EESLD..T- LF...TIMSF ILLA.F.FFM ..VKFTPAYL EASG-----  237
FtPPT1 A.....QS.N VL...F.V-KK -------EESLD..S- LF.VMTIMSF FLLA.V.FF. ..LT.SPAYL QS.G-----  289
FtPPT2 AA..TNQT.N VL...F.I-RK -------EEALD..N- LF.VMTILSF LFLI.I.VCL ..FK.TPEYM QF.ASQG--  241
OsPPT1 A...T.QS.N VL...L.V-KK -------EESLD..T- LF...TVMSF FLLA.VTLLT ..VKVTPTVL QS.G-----  234
OsPPT2 A...T.QS.N VL...L.V-KK -------EESLD..N- LF...TVMSF FLLA.V.FLT ..IKITPTVL QS.G-----  234
OsPPT3 A..LLYQS.N VL...LLG-GE -------EEALDD.N- LF..LTILSF LLS..LMLFS ..VKFSPGYL RSTG-----  245
OsPPT4 A..LTNQS.N VF...LLA-.K -------EETLDD.N- LF..MTVMSF LLSA.LMLS. ..IKFSPSYL QSNG-----  244
SbPPT1 A...T.QS.N VL...L.V-KK -------EESLD..N- LF...TVMSF FLLA.VTLLT ..VKVSPAVLQS.G-----  237
SbPPT2 A...T.QS.N VL...L.V-KK -------EESLD.LN- LF...TVMSF FVLA.VTFFT ..VKITPTFL QS.G-----  233
SbPPT3 A..LTNQS.N VL...LLAG.K --------DVMDD.N- LF.V.TVLSF LLSC.LMFFA ..IKFTPGYL QSTG-----  245
SbPPT4 A..LTNQS.N V....ILA-.K -------EDSLDD.N- LF...TIMAF LLSA.LMLS. ..IKFSPSYL QS.G-----  237
ZmPPT1 A...T.QS.N VL...L.V-KK -------EESLD..N- LF...TVMSF FLLA.VTLLT ..VKVSPAVL QS.G-----  228
ZmPPT2 A...T.QS.N VL...L.V-KK -------EESLD..N- LF...TVMSF FLLA.VTLLT ..VKVSPAVL QS.G-----  230
ZmPPT3 A...T.QS.N VL...L.V-KK -------EESLD.LN- LF...TVMSF FLLA.VTFFT ..VKITPTFL QS.G-----  233
PpPPT  A..VT.QS.N VL...F.V-KK -------EGSLD..N- LF...TVMSF FLL..VTFF. ..VKFTPSAL AASG-----
McGPT1 ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL..LS.VLL ---T.F.LY. ...KMWAA.W DK.VSD---  239
McGPT2 ....A.VF.N IF...G.N-GQ ------SV.GM-.Y.. CL.MLS.LLL ---T.F..A. ....VWAA.W QK.VSQ---  239
VvGPT1 ...LA.VF.N IF..RG.K-GK ------SVGGM-.Y.. CL.MLS.LIL ---T.F..A. ....MWAA.W QK..SQ---  241
VvGPT2 ...LA.VF.N IF..RG.K-GK ------SV.GM-.Y.. CL..MS.LIL ---T.F..A. ....MWAA.W QN.VSQ---  237
PsGPT  ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL..LS.AIL ---T.F..A. ...AMWAA.W QT.LSE---  235
GmGPT  ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL..LS.AIL ---T.F..A. ....MWAA.W QT.MSQ---  234
MtGPT  ....A.VF.N IF...G.K-G. ------SV.GM-.Y.. CL..LS.LLL ---T.F..A. ...TMWAA.W QT.------  237
AtGPT1 ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL.MLS.LIL ---T.F..A. ....MWVD.W QT.L.T---  241
AtGPT2 ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL.MMS.VIL ---T.FS.A. ....MWAA.W QN.VSQ---  236
ThGPT  ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL.MMS.LIV ---T.F..A. ....VWAA.W QN.VSE---  236
StGPT  ...LA.VF.N IF...G.K-GK ------SVGGM-.Y.. CL.MMS.LIL ---I.F..A. ....VWAL.W QN.VSQ---  240
HaGPT  ...LA.VF.N IF..RG.K-GK ------SV.GM-.Y.. CL.MLS.LIL ---T.F..A. ...KMWAA.W QN.VTE---  230
TaGPT  ...LA.VF.N IF..RG.K-GK ------SV.GM-.Y.. CL..MS.VIL ---T.FA.AM ....MWAA.W QK.L.D---  236
OsGPT1 ...LA.VF.N IF..RG.K-GK ------SV.GM-.Y.. CL..MS.VIL ---T.FA.AM ....MWAA.W QK.L.E---  237
OsGPT2 ...LA.VF.N IF...G.K-GK ------SV.GM-.Y.. CL.MLS.VIL ---L.FAFAM ...KVWAA.W QK.VAE---  239
SbGPT  ...LA.VF.N IF..RG.K-GK ------SV.GM-.Y.. CL..MS.VIL ---T.F..AM ....MWAA.W QK.L.E---  236
```

```
ZmGPT   ...LA.VF.N IF..RG.K-GK ------SV.GM-.Y.. CL..MS.VIL ---T.FA.AM ....MWAA.W QK.L.E---  236

PsiGPT ...LA.VF.N IF...G.KAGK ------SVGGM-.Y.. CL.MMS.ALL ---T.F.FA. ....AWAA.W QE.LRA---  246

PpGPT  V..IA.VF.N IF...G..SGK ------SVGGM-.Y.. CL.MMS.VFL ---T.F..A. ...KSWTA.W DA.NLT---  235

VvXPT   ...VG.VL.N IY..RSLE-SF ------KEVNGL.L.G W....S.LYL ---F.VA.F. ..T.-WIE.Y HR..QA---  232

GmXPT   ...VG.VL.N IY..RSLQ-NF ------KEVDGL.L.G W.T.LS.LYL ---F.V..F. ..S.-WIP.Y YK..EA---  236

AtXPT  ....G.VL.N I...RSLQ-SF ------KEIDGL.L.G C...LS.LYL ---F.VA.F. ..SH-WVP.Y HK...S---  240

GsTPT2  ....A.VT.N IT..FT.V-DF KNEKTL-IAQ--.T.. L.T..SF.ME ---..F.LLM ..F----PPL VS...G---

PfoTPT L..FGSSI.. I.A..M..-QK SLIGENLNAS--.... F.T..SALIS ---..LVLAF ..KETYNFLV .YQGTN---

TgAPT  V.ALGSSA.A VFA.L..A-.R KQVGENLS.A--.M.. LLT.V.SL.S ---..L..FA ..AKVAAV-W EACTGPDSP
```

```
                         VI
         _ _ _ ________________________ _ _ _

SoTPT   VG-LTKFISDL FWVGMFYHLY NQLATNTLER VAPLTHAVGN VLKRVFVIGF SIIAFGNKIS  290
McTPT   ..-.....T.. .......... .......... .......... .......... ...I......  293
VvTPT   ..-....L... .......... .......... .......... .......... ..LV......  297
VvTPT2  ..-.A.LV... ....L.F..D ....VS.... .S.......S .....V..VL .T.V.....T  294
VvTPT3  ..-....L... ..I....... .......... .......... .......... ..VI......  294
PsTPT   ..-.V..V... .......... ..V....... .......... .......... ...I......  296
AtTPT   ..-M....... .......... .......... .......... .......... ..VI......  294
BoTPT   ..-M....... .......... ..L....... .......... .......... ..VI......  294
PtTPT1  ..-........ .......... .......... .......... .......... ..LI......  300
PtTPT2  ..-........ .......... .......... .......... .......... ..VI......  300
RcTPT1  ..-T....T.. .......... .......... .......... .......... ..VV......  299
RcTPT2  ..-.F..V... ..I....... ..V....... .......... .......... ..VV...R..  294
StTPT   ..-....VT.. .......... ..V....... .......... .......... ..VI......  296
NtTPT   ..-....VT.. .......... ..V....... .......... .......... ...V......  296
FpTPT   ..-M....... .......... ....I..... .......... .......... ...V......  292
FtTPT   ..-MI...... .......... ..I....... .......... .......... ...V......  292
TaTPT   ..-M...V... .L..L..... ..I....... .......... .......... ...I.....T  291
OsTPT1  ..-M...V... .L..L..... ..I....... .......... .......... ...V...R.T  289
OsTPT2  ..-.A.LV.N. LV..L..... ..V....... .T........ .......... .........T  296
SbTPT   ..-....V... VL..L..... ..I....... .......... .......... ..VV......  289
SbTPT2  ..-...L..NF .V..L..... ..V....... ....S..I.. .......... ...V.....T  293
ZmTPT1  ..-....V... .L..L..... ..I....... .......... .......... ...V......  291
ZmTPT2  ..-....V... .L..L..... ..I....... .......... .......... ..VV......  288
ZmTPT3  ..-...L..NF .V..L..... ..V....... .......I.. .......... .........T  293
PsiTPT  ..-.V..L... .......... .......... .......... .......... ...V...R..  301
PpTPT   ..-MQ..L... .......... ....N..... .......... .......... ..VV......  302
McPPT   LN-VQQVYMKS .LAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VS .V.V.RTAVN  298
VvPPT1  LN-MGQIYKRS LIAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VT .VLF.RTPV.  293
VvPPT2  LN-VRELCVRS LLA.ICF.S. Q.VSYTI.QM .S.V..A... CV...V..IS .V.F.QTPA.  296
PsPPT   LN-VRQVYTRS LLAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VS .V.I.KTPV.  291
GmPPT1  .N-VRQLYIRS LLAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VS .V.F.QTPV.  293
GmPPT2  LN-VRELCVRS VLAAFCF.A. Q.VSHMI.QM .S.V..S... CV...V..VS .V.F.QIPV.  298
AtPPT1  .N-VKQIYTKS LIAALCF.A. Q.VSYMI.A. .S.V..S... CV...V..VS .V.F.KTPV.  294
AtPPT2  LS-VKE.CIMS LLA.VCL.S. Q.VSYMI..M .S.V..S... CV...V..TS ..LF.KTPV.  293
BoPPT   .N-VQQIYTKS LIAALCF.A. Q.VSYMI.A. .S.V..S... CV...V..VS .V.F.KTPV.  291
```

```
BnPPT  .N-VQQIYTKS LIAALCF.A. Q.VSYMI.A. .S.V..S... CV...V..VS .V.F.KTPV.  294
PtPPT1 LN-VKQVYTRS LIAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VS .VFF.KTPV.  293
PtPPT2 LN-VKEVYTRA .LAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VS .VLF.KTPV.  293
RcPPT1 LN-VKEVYIRS LLAALCF.A. Q.VSYMI.Q. .S.V..S... CV...V..VS .VLF.RTPV.  293
RcPPT2 LN-VRELCVRA LIA.FCF.S. Q.VSYLI.QM .N.VS.A... SV...V..VS .V.F.QIP..  295
NtPPT1 .N-VNQLYTRS LIAALCF.A. Q.VSYMI.Q. .S.V..SL.. CV...V..VT .VLF.RTPV.  293
NtPPT2 LN-VNQIYTRS LLAALCF.A. Q.VSYMI... .S.V..S... CV...V..VT .VLF.RTPV.  297
FtPPT1 LN-VQQIYVRS LLAAICF.A. Q.VSYMI.Q. .S.V..S... CV...V..VT .VLF.RTPV.  349
FtPPT2 .N-VRELCVRA LLT.ICF.S. Q.VSYMI..M .S.V....A. CV...V..VS .V.F.RTPVT  301
OsPPT1 LN-.KQIYTRS LIAAFCF.A. Q.VSYMI.A. .S.V..S... CV...V..VT .VLF.RTPV.  294
OsPPT2 LN-VKQVLTRS LLAALCF.A. Q.VSYMI.A. .S.V..S... CV...V..VT .VLF.RTPV.  294
OsPPT3 LN-.QELCVRA ALA.FCF.G. QK.SYLI.A. .S.V..S.A. CV...V..VA .VLF.RTP..  305
OsPPT4 .N-.QELCMKA ALA.TCF.F. Q.VSYSL.A. .S.V..S.A. CV...V..VS .VLF.RTP..  304
SbPPT1 LN-.KQVYTRS LIAAFCF.A. Q.VSYMI.A. .S.V..S... CV...V..VT .VLF.RTPV.  297
SbPPT2 LN-VNQVLTRS LLA.LCF.A. Q.VSYMI.AM .S.V..S... CV...V..VT .VLF.RTPV.  293
SbPPT3 LN-.QELCVRA ALA.LCF.G. QK.SYLI.S. .S.V..S.A. CV...V..VS .VLF.STP..  305
SbPPT4 .S-VKELCVRA ALA.TCFYF. Q.VSYSL.A. .S.V..S.A. SL...V..VS .VLF.RTP..  297
ZmPPT1 LN-.KQVYTRS LIAACCF.A. Q.VSYMI.A. .S.V..S... CV...V..VT .VLF.RTPV.  288
ZmPPT2 LN-.KQIYTRS LIAACCF.A. Q.VSYMI.A. .S.V..S... CV...V..VT .VLF.RTPV.  290
ZmPPT3 LN-VNQVLTRC LFA.LCF.A. Q.VSYMI.AM .S.V..S... CV...V..VT .VLF.RTPV.  293
PpPPT  LD-VKVVVTRA LIA.LCF.A. Q.VSYMI.AK .T.V..S... CV...V..VT .VLF.RTPV.
McGPT1 I.-S-N..WW. TAQSV..... ..VSYMS.DE IS...FSI.. TM..IS..VS ...I.RTPVQ  298
McGPT2 I.-P-N.VWWV AAQSI..... ..VSYMS.DQ IS...FS... TM..IS..VS ...I.HTP.Q  298
VvGPT1 I.-P-N..WWV AAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VS ...I.HTPVQ  300
VvGPT2 I.-P-H.VWWV AAQSV..... ..VSYMS.DE IS...FSI.. TM..IS..VS ...I.HTPVQ  296
PsGPT  I.-P-Q..WWV AAQSI..... ..VSYMS.DE IS...FSI.. TM..IS..VS ...I.HTP.Q  294
GmGPT  I.-P-Q..WW. AAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VS ...I.HTPVQ  293
MtGPT  -------NWWV AAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VS ...I.HTP.Q  296
AtGPT1 ..-P-Q.VWWV VAQSV..... ..VSYMS.DQ IS...FS... TM..IS..VS ...I.RTPVQ  300
AtGPT2 ..-P-N.VWWV VAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VA ...I.HTP.Q  295
ThGPT  I.-P-N.VWWV AAQSV..... ..VSYMS.DQ IS...FS... TM..IS..VA ...I.HTP.R  295
StGPT  I.-P-N..WWV VAQSV..... ..VSYMS.NE IS...FSI.. TM..IS..VS ...I.QIPIQ  299
HaGPT  I.-P-H..WWV AAQSI..... ..VSYMS.DE IS...FSI.. TM..IS..VS ...I.HTPVQ  289
TaGPT  ..-P-NVLWWI GAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VS ...I.RTPVR  295
OsGPT1 ..-P-DVVWWV AAQSV..... ..VSYMS.DE IS...FSI.. TM..IS..VS ...I.HTPVR  296
```

```
OsGPT2 I.-P-N.VWWV AAQSV..... ..VSYMS.DE IS...FSI.. TM..IS..VA ...I.HTPVQ  298

SbGPT  ..-P-NV.WWI AAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VS ...I.HTPVR  295

ZmGPT  ..-P-NVVWWI AAQSV..... ..VSYMS.DQ IS...FSI.. TM..IS..VS ...I.HTPVR  295

PsiGPT I.-P-Q.VWWV AAQSV..... ..VSYMS.NE IS...FSI.. TM...S...S ...I.RTEVR  305

PpGPT  ..-P-.IFWWV VAQSV..... ..VSYMS.NE IS...FSI.. TM...T..VS ...I.HTQVQ  294

VvXPT  ..KP.T.YIWV MLS.V..... ..SSYQA.DD IS...FS... TM...V..VA T.LV.R.PVK  293

GmXPT  I.KAST.YTWV LVS.V..... ..SSYQA.DE IS...FS... TM...V..VS .VLV.R.PVR  297

AtXPT  ..TPST.YFWV LLS.V..... ..SSYQA.DE IS...FS... TM...V..IS TVLV.R.PVR  301

GsTPT2  .S-KA.LFGSI MFCSL..... .EVSYLC.DN .S.VSFSI.. TI...II.FG ..LV.RTPVT

PfoTPT YT-FKDV.FKI ILS..W.YFN .EV.FMC... .NQI...LA. SI...VI.VS ...I.KTQ.T

TgAPT  WT-GQQI.AK. CFS.LW.YM. .EV.YLC..K INQV....A. T....VI.VA .VLF.QTPVT
```

```
SoTPT  TQTAIGTSIA IAGVALYSLI KAKMEEEKRQ -----MKST  324
McTPT  .......... .....I..F. .G........ -----K.AA  327
VvTPT  ...G...CV. .....M..F. .......... -----L..A  331
VvTPT2 .......A.. .T...I.... R.N....NQN -----AAA.  328
VvTPT3 R..G...A.. .....I.... ..NI..Q..K --AAVTPAS  331
PsTPT  ...G...G.. ........F. ..QI...... -----A.AA  330
AtTPT  ...G...G.. .....CT.I. ...I...... -----G.KA  328
BoTPT  ...G...G.. ........V. ...I...... -----G.TA  328
PtTPT1 ...G...G.. .....T..Y. .........R -----G.AA  334
PtTPT2 ...G...AV. .....T..Y. ...L...... -----G.AA  334
RcTPT1 ...G...C.. .....M..FL ...I...... -----G.TA  333
RcTPT2 ...G...A.. .....M.... ..N...Q..K --AAIAPAS  328
StTPT  ...G...C.. .....I..F. .......... -----K.AA  330
NtTPT  ...G...C.. ........F. .......... -----K.AA  330
FpTPT  .......... .....V.... ...I.....G -----L..A  326
FtTPT  .......... .....I.... ..RI.....R -----...A  326
TaTPT  ...G...CV. ........Y. ...I.....- -----A.AA  325
OsTPT1 ...G...C.. ........Y. ...I.....- -----A..A  323
OsTPT2 ...G...C.. ........Y. ...I....T. -----...A  330
SbTPT  ...G...... ........Y. ...I.....- -----K..A  322
SbTPT2 ...G...... .S......F. ...I....K. -----I..A  327
ZmTPT1 ...G...... .....M..Y. ...I.....- -----K..A  325
ZmTPT2 ...G...... .....M..Y. ...I.....- -----K..A  322
ZmTTP3 ...G...... VS......F. ...I..-.K. -----I..A  326
PsiTPT .......... .....I..F. ..QL.....K AVPPSPRAS  340
PpTPT  ...G...A.. .G......F. ..RQ..A.IA -----K.AA  336
McPPT  PIN.L..AV. L...F...-- RV.R--I.AK -----A.EA  328
VvPPT1 PVNSL..GV. L...F...-- RV.R--I.-- -----P.TA  321
VvPPT2 PINSL..GV. LV..F...-- R..R--M.PK -----P.AA  326
PsPPT  PVN.L..AVG L...F...-- RV.R--I.SK -----P.AV  321
GmPPT1 PVN.F..A.. L...F...-- RV.R--I.AK -----P.TA  323
GmPPT2 PVNTL..GL. LV..F...-- R..R--I.SV -----Q.TN  328
AtPPT1 PVN.F..G.. L...F...-- RV.G--I.PK -----P.TA  324
AtPPT2 PLNS...AT. L...Y...-- R..RVQV.PN -----P.MS  325
BoPPT  PVN.F..G.. L...F...-- RV.R--I.PK -----P.TA  321
BnPPT  PVN.F..G.. L...F...-- RV.R--I.PK -----P.TA  324
```

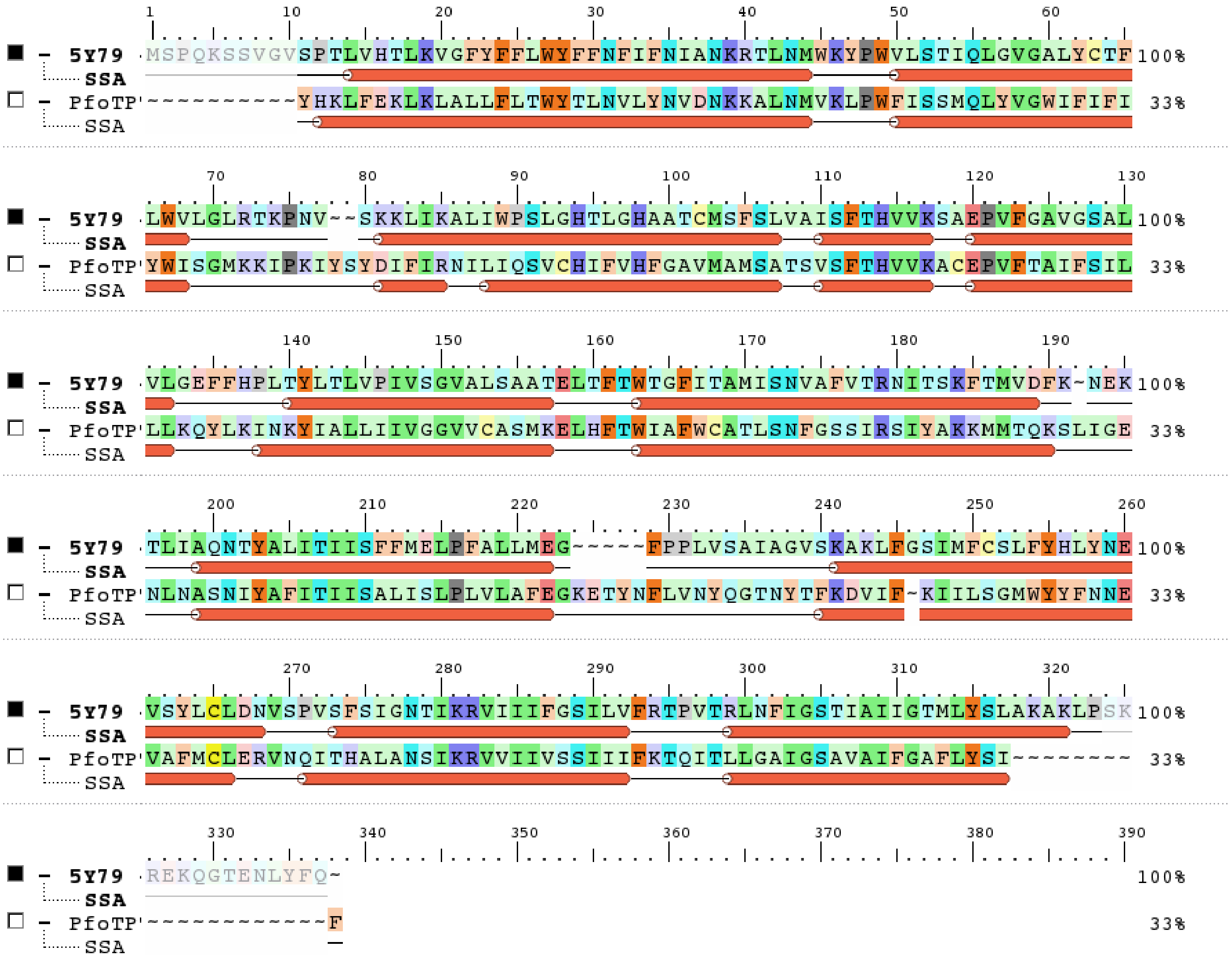
```
PtPPT1 PINSL..GV. L...F...-- RV.R--I.PK -----P.TA  323
PtPPT2 PINSL..G.. L...F...-- RV.S--I.PK -----P.TA  323
RcPPT1 PINSL..G.. L...F...-- RV.R--I.PK -----P.TA  323
RcPPT2 PVNSL..A.. L...F...-- R..R---.TP --PPMP.AS  327
NtPPT1 PINGL..GV. L...F...-- RV.R--I.PK -----A.TE  323
NtPPT2 PINT...GV. L...F...-- RV.G--I.PK -----P.TA  327
FtPPT1 PINS...GV. L...F...-- QV.R--L.-- -----P.KA  377
FtPPT2 PIN.L..GL. L...F...-- R..R--I.-- -----P.AA  329
OsPPT1 PINSL..GV. L...F...-- QL.R--L.PK -----P.TA  324
OsPPT2 PINSL..A.. L...F...-- QL.R--L.PK -----P.AA  324
OsPPT3 PVN.L..GV. LG..F...-- RL.R--T.-- -----P.NA  333
OsPPT4 PIN.L..GV. L...F...-- RF.K--A.PK -----A.TA  334
SbPPT1 PINSL..G.. L...F...-- QL.R--L.PK -----P.AA  327
SbPPT2 PINSL..A.. L...F...-- QL.R--L.PK -----P.TP  323
SbPPT3 PVN.L..GA. L...F...-- RLTR--T.-K -----P.DA  334
SbPPT4 PIN.L..GV. L...F...-- QF.K--L.PK -----T.AA  327
ZmPPT1 PINSL..G.. L...F...-- QL.R--L.PK -----P.TA  318
ZmPPT2 PINSL..G.. L...F...-- QL.R--L.PK -----P.AA  320
ZmPPT3 PINSL..A.. L...F...-- QL.R--L.PK -----P.TA  323
PpPPT  PVNGL..GL. LC.VFA..-- RV-----.SK ------.--
McGPT1 PVN.L.AA.. VF.TF...-- ---------. -----A.Q-  320
McGPT2 PVN.L.AA.. .L.TFI..-- ---------. -----A.V-  320
VvGPT1 PVN.L.AA.. .L.TK???-- --???????? -----????  3??
VvGPT2 PIN.L.AA.. .L.TF...QV ?????????? -----????  3??
PsGPT  PVN.L.AA.. VF.TF...-- ---------. -----A.Q-  316
GmGPT  PIN.L.AA.. .L.TF...-- Q..G-.VRLN -----LQD-  323
MtGPT  PNN.L.AA.. .L.TF...-- ---------. -----A.Q-  318
AtGPT1 PVN.L.AA.. .L.TF...-- ---------. -----A.L-  322
AtGPT2 PVN.L.AA.. .F.TF...-- ---------. -----A.Q-  317
ThGPT  PVN.L.AA.. .L.TFI.F-- ---------. -----VE--  316
StGPT  PIN.L.AA.. .L.TF...-- ---------. -----A.Q-  321
HaGPT  PIN.L.AA.. .F.TF...-- ---------. -----A.Q-  311
TaGPT  PVN.L.AA.. .F.TF...-- ---------. -----A.Q-  317
OsGPT  PVN.L.AA.. .L.TF...-- ---------. -----A.Q-  318
OsGPT2 PIN.L.AA.. .L.TFI..-- ---------. -----A.Q-  320
SbGPT  PVN.L.AA.. .L.TF...-- ---------. -----A.A-  317
ZmGPT  AVN.L.AA.. .L.TF...-- ---------. -----A.A-  317
```

```
PsiGPT PVNGL.AA.. .L.TF...-- ---------. -----A.Q-  316
PpGPT  PMN.V.AA.. .F.TF...QV LHHALPYFLA -RTELL.--  330
VvXPT  PLN.L.SA.. .F.TF...QA TS.KSPK.IE ----GEKSS  328
GmXPT  PLNGL.SA.. .L.TF...QA TS.------- -----K.A-  323
AtXPT  PLN.L.SA.. .C.TF...QA TA.KKKIEVG ----GD.KN  336
GsTPT2 RLNF..ST.. .I.TM....A ...LPS-..E -------.Q-
pfoTPT LLG...SAV. .F.AF...IF ---------- ---------
TgAPT  ALG.T.SFV. ...TLI...S ---------K -----T.YG
```

So	*Spinacia oleracea* (Caryophyllales, Amaranthaceae)
Mc	*Mesembryanthemum crystallinum* (Caryophyllales, Aizoaceae)
Cs	*Camellia sinensis* (Ericales, Theaceae)
Fp	*Flaveria pringlii* (Asterales, Asteraceae)
Ft	*Flaveria trinervia* (Asterales, Asteraceae)
Ha	*Helianthus annuus* (Asterales, Asteraceae)
St	*Solanum tuberosum* (Solanales, Solanaceae)
Nt	*Nicotiana tabacum* (Solanales, Solanaceae)
Bo	*Brassica oleracea* (Brassicales, Brassicaceae)
Bn	*Brassica napus* (Brassicales, Brassicaceae)
At	*Arabidopsis thaliana* (Brassicales, Brassicaceae)
Pt	*Populus trichocarpa* (Malpighiales, Salicaceae)
Rc	*Ricinus communis* (Malpighiales, Euphorbiaceae)
Ps	*Pisum sativum* (Fabales, Fabaceae)
Gm	*Glycine max* (Fabales, Fabaceae)
Mt	*Medicago trunculata* (Fabales, Fabaceae)
Vv	*Vitis vinifera* (Vitales, Vitaceae)
Zm	*Zea mays*
Os	*Oriza sativa*
Ta	*Triticum aestivum*
Sb	*Sorghum bicolor*
Psi	*Picea sitchensis*
Pp	*Physcomitrella patens*
Th	*Thelungiella halophila*

Gs	*Galderia sulphuraria*
Pf	*Plasmodium falciparum*
Tg	*Toxoplasma gondii*

Appendix 2: Sequence alignment between PfoTPT and GsTPT2 during homology modelling



```
         1         10        20        30        40        50        60
■ 5Y79   MSPQKSSVGVSPTLVHTLKVGFYFFLWYFFNFIFNIANKRTLNMWKYPWVLSTIQLGVGALYCTF 100%
  SSA
□ PfoTP  ~~~~~~~~~~YHKLFEKLKLALLFLTWYTLNVLYNVDNKKALNMVKLPWFISSMQLYVGWIFIFI  33%
  SSA

             70        80        90        100       110       120       130
■ 5Y79   LWVLGLRTKPNV~~SKKLIKALIWPSLGHTLGHAATCMSFSLVAISFTHVVKSAEPVFGAVGSAL 100%
  SSA
□ PfoTP  YWISGMKKIPKIYSYDIFIRNILIQSVCHIFVHFGAVMAMSATSVSFTHVVKACEPVFTAIFSIL  33%
  SSA

                   140       150       160       170       180       190
■ 5Y79   VLGEFFHPLTYLTLVPIVSGVALSAATELTFTWTGFITAMISNVAFVTRNITSKFTMVDFK~NEK 100%
  SSA
□ PfoTP  LLKQYLKINKYIALLIIVGGVVCASMKELHFTWIAFWCATLSNFGSSIRSIYAKKMMTQKSLIGE  33%
  SSA

             200       210       220       230       240       250       260
■ 5Y79   TLIAQNTYALITIISFFMELPFALLMEG~~~~~FPPLVSAIAGVSKAKLFGSIMFCSLFYHLYNE 100%
  SSA
□ PfoTP  NLNASNIYAFITIISALISLPLVLAFEGKETYNFLVNYQGTNYTFKDVIF~KIILSGMWYYFNNE  33%
  SSA

                   270       280       290       300       310       320
■ 5Y79   VSYLCLDNVSPVSFSIGNTIKRVIIIFGSILVFRTPVTRLNFIGSTIAIIGTMLYSLAKAKLPSK 100%
  SSA
□ PfoTP  VAFMCLERVNQITHALANSIKRVVIIVSSIIIFKTQITLLGAIGSAVAIFGAFLYSI~~~~~~~~  33%
  SSA

             330       340       350       360       370       380       390
■ 5Y79   REKQGTENLYFQ~                                                     100%
  SSA
□ PfoTP  ~~~~~~~~~~~~F                                                      33%
  SSA
```