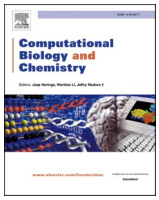
Computational Biology and Chemistry 120 (2026) 108653

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Sequence- and structure-based bioinformatic screening for potential *Theileria parva* transport-related proteins

Nikolaos Kotsovolos [a,*], Leonhard Schnittger [b,c], Kgomotso Sibeko-Matjila [a]

[a] Department of Veterinary Tropical Diseases, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, South Africa
[b] Instituto de Patobiología Veterinaria, Centro de Investigaciones en Ciencias Veterinarias y Agronómicas (CICVyA), INTA-Castelar, Hurlingham, 1686, Argentina
[c] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ciudad Autónoma de Buenos Aires C1033AAJ, Argentina

ARTICLE INFO

Keywords:
Theileria parva
Bioinformatics
Transport protein
Protein structural homology
Protein sequence homology
Protein annotation

ABSTRACT

As an obligate intracellular parasite, *Theileria parva* is strictly dependent on its host for nutrient acquisition. Transport proteins are expected to play a crucial role in the influx of essential nutrients to sustain the parasite's rapid growth. Unfortunately, the *T. parva* transportome is still not comprehensively elucidated, and plagued by the presence of uncharacterized proteins. In this study, we employed a combination of approaches including sequence orthology and structural similarity to identify 188 proteins predicted to be involved in transport-related processes. Among these, 24 were uncharacterized proteins, and 17 of them could be assigned a tentative annotation. Furthermore, the localization of these 188 proteins was investigated, resulting in their assignment to seven cellular compartments. Screening of the proteomes of other *Theileria* species, *T. annulata*, *T. orientalis*, and *T. equi* revealed that all 188 proteins were present in both transforming and non-transforming *Theileria* parasites. Among the 188 potential transport-related proteins, 45 were associated with transmembrane transport and most of them (87 %) are conserved across phylum Apicomplexa.

1. Introduction

Theileria parva is an intracellular protozoan parasite of cattle and buffalo, transmitted by the *Rhipicephalus appendiculatus* and *R. zambeziensis* ticks, and is distributed across eastern, central, and southern Africa where the vector is endemic (Blouin and Stoltsz, 1989; Nene et al., 2016). This parasite belongs to phylum Apicomplexa, class Aconoidasida, which contains piroplasmids such as *Theileria* and *Babesia* and haemosporidians such as *Plasmodium*. Infection in buffalo is typically asymptomatic, but often lethal in cattle, manifesting itself as East Coast fever, January disease, and Corridor disease (Nene et al., 2016). Infected ticks release sporozoites into the bovine bloodstream, which infect lymphocytes and develop into multinucleated schizonts (Fawcett et al., 1984; Nene et al., 2016). The primary pathology of *T. parva* infection arises from the cancer-like proliferation of schizont-infected lymphocytes, and the subsequent lysis of infected and uninfected cells (Dobbelaere and Heussler, 1999; Nene et al., 2016). Interestingly, this host cell transformation is not a universal trait among *Theileria* species; *T. parva, T. annulata, T. lestoquardi, T. taurotragi,* have this ability, while *T. orientalis, T. mutans, T. velifera, T. equi, and T. haneyi* do not (Bishop et al., 2020; Schnittger et al., 2022, 2012). *Theileria* parasites appear to have reduced metabolic pathways, potentially indicating an increased dependence on nutrient scavenging from the host (Hayashida et al., 2012). Given the rapid proliferation of infected lymphocytes and the parasites within them, it can be expected that the schizonts will be undergoing considerable metabolic activity and have a substantial nutritional requirement to sustain the increased rate of replication.

Membrane transport proteins, commonly referred to as transport proteins or transporters, are integral membrane proteins that facilitate the transport of polar substances across membranes (Staines et al., 2017). In the case of intracellular parasites, transport proteins are essential for nutrient acquisition from the host cell. In addition, transport proteins play important roles such as the removal of drugs and metabolic wastes and the maintenance of ion concentrations (Staines et al., 2017). In fact, mutations in transport proteins have been demonstrated to be involved in drug resistance either by reducing the importation of a drug or by increasing its exportation (Mäser et al., 1999; van Schalkwyk et al., 2016). A recent review article aiming to compile a comprehensive list of *Plasmodium falciparum* transport proteins (including members of transport complexes directly involved in substrate translocation) obtained a list of 197 such proteins (Wunderlich, 2022). Of these, 99 were considered essential in the blood

* Corresponding author.
E-mail address: u14196973@tuks.co.za (N. Kotsovolos).

https://doi.org/10.1016/j.compbiolchem.2025.108653
Received 1 April 2025; Received in revised form 21 July 2025; Accepted 24 August 2025
Available online 29 August 2025

stage by their respective sources, while a further three were deemed essential in other stages (Wunderlich, 2022). Hence, transporters may serve as attractive targets for drugs against *Plasmodium*, and likely apicomplexans in general.

Unfortunately, as hypothetical proteins with unknown identities and functions are prevalent in annotated proteomes, it is not currently feasible to have comprehensive lists of transmembrane transport proteins, especially in understudied species. Hypothetical proteins can be defined as proteins predicted to exist from nucleic acid sequence data, but have not been experimentally confirmed at the protein level (Lubec et al., 2005). Most hypothetical proteins are "conserved hypothetical proteins", which can be defined as proteins of unknown function, but with orthologues in a large phyletic distribution (Galperin and Koonin, 2004; Roberts, 2004). Annotation of unknown genes typically relies on the transfer of functional information from well-studied reference genes, often through the use of evolutionary relationships (e.g. typically a orthologous gene of a closely related model organism) (Stamboulian et al., 2020). While orthologous genes (arising from speciation events) have traditionally been considered superior for functional annotation i.e. the 'orthologue conjecture' (Nehrt et al., 2011), the inclusion of paralogous genes (arising from gene duplications) was found to be beneficial for annotation (Stamboulian et al., 2020). An orthologous group of genes, all of which evolved from a single gene after a particular speciation event, includes both orthologues and in-paralogues (Gabaldón and Koonin, 2013), may thus be more informative for functional annotation than a collection of purely orthologous genes.

Approximately 30–40 % of genes in newly sequenced bacterial genomes remain uncharacterized, with likely higher proportions in eukaryotic genomes, therefore, there is an urgent need to elucidate the functions of such genes and their associated proteins (Galperin and Koonin, 2004). Protozoa also suffer from high proportions of uncharacterized genes, with approximately 51 % (Yucesan et al., 2021), 50 % (Aguttu et al., 2021), 40 % (Berná et al., 2021) of *Toxoplasma gondii*, *Plasmodium falciparum*, and *Neospora caninum* genomes comprising of genes encoding hypothetical proteins, respectively. Notably, the proportion of such proteins in newly sequenced genomes is decreasing, as more sophisticated annotation methods are developed, and reference databases have grown to include novel characterized proteins. The fact remains however, that the high number of hypothetical and uncharacterized proteins prevents the establishment of a true understanding of biological systems, even in the case of well-studied model organisms (Galperin and Koonin, 2004). In *T. parva*, approximately 39 % of the encoded proteins of the proteome did not have any detectable Pfam domains in 2020 when a reannotation study was performed (Tretina et al., 2020), and thus their functions are unknown. This high percentage could be attributed to the fact that the majority of proteins of apicomplexan species are highly diverse and specific to the apicomplexan phylum, with many of those forming genus- and species-specific families (Wasmuth et al., 2009). Thus, those proteins could not be assigned meaningful annotations by using well-studied but distantly related model organisms, and would instead have to wait until taxonomically restricted protein families from understudied species are included into reference databases, such as the piroplasmid-specific RAP-1 family domain (PF03085) (Bastos et al., 2021).

It is expected that the uncharacterized proteins (including "hypothetical" and "[putative] integral membrane" proteins) of the *T. parva* proteome contain transport proteins. Identifying a potential transportome of the parasite can yield numerous benefits, such as a better understanding of its biology. In addition, it provides a pool of proteins that can be considered for the discovery of novel drugs, necessitated by increasing reports of drug resistance (Mhadhbi et al., 2010), as well as high costs of current treatments. Thus, this study aimed to use various *in silico* methods to identify, and potentially assign functions to uncharacterized, *T. parva* transporters.

2. Methods

2.1. Definitions and search strategy

For the purposes of this study, potential transport-related proteins (PTPs) are defined as proteins that are potentially involved in transport processes, either directly or indirectly. Furthermore, uncharacterized proteins refer to those that are annotated as "hypothetical protein", "uncharacterized protein" or "[putative] integral membrane protein". To identify PTPs in the *T. parva* proteome, a search on TransportDB 2.0 was conducted and an orthology-based approach was employed by searching databases EggNOG6, OrthoDB and OrthoMCL for orthologues to *T. parva* proteins within the Apicomplexan class Aconoidasida. Based on the annotations assigned to their orthologues, *T. parva* proteins were designated as PTPs. Various functional properties were predicted for these PTPs to provide insight on their functions, especially for those that were uncharacterized on NCBI and UniProt. Structural similarity searches of the AlphaFold-predicted structures of the *T. parva* PTPs against the Protein Data Bank (PDB) were performed using the DALI and FoldSeek servers. The PTPs were assigned to different subcellular localizations. Finally, the identified PTPs were queried on BLASTP against other transforming and non-transforming *Theileria* species for comparison. A flow chart of the *in silico* approach followed in this study is presented in Fig. 1.

The tools used to obtain protein information, predict properties for functional annotation, identify orthologues and perform structural comparisons are listed in Table 1. Default settings were used unless otherwise specified. Furthermore, Venn diagrams were generated using https://bioinformatics.psb.ugent.be/webtools/Venn/, while Maestro from the Schrodinger software suite (2024–1) (https://newsite.schrodinger.com/) was used to visualize and compare protein structures.

2.2. Sequence retrieval and identifier mapping

The NCBI was used to obtain protein sequences for *T. parva* strain Muguga. The *T. parva* genome was initially annotated in 2005 (Gardner et al., 2005), and re-annotated in 2020 (Tretina et al., 2020). At the start of this study, the GenBank annotation (GCA_000165365.1) was more recently updated than the RefSeq annotation (GCF_000165365.1). Thus, protein-coding sequences were retrieved from the GenBank annotation. Hereafter, this proteome of 4051 proteins is referred to as the *T. parva* 'GenBank proteome'. The protein sequences were also used for various purposes, including orthologue screening and functional annotation.

A protein ID mapping system was created to correlate the different identifiers for each protein across NCBI and UniProt databases. The details are described in Appendix A. Briefly, each entry was indexed by NCBI gene ID, and subsequently linked to their associated GenBank and RefSeq protein accession numbers, old (with the format: TP00_0000) and new (with the format: TpMuguga_00g00000) NCBI locus tags, as well as UniProt IDs. This mapping system was used to correlate the identifiers of PTPs from the orthology analyses as well as outputs from the annotation analyses.

2.3. Identification of transport-related proteins

An orthology-based analysis was performed to identify PTPs by virtue of their sequence similarity to transporters in other Aconoidasida species. Proteins already annotated as transporters in *T. parva* were also included.

To this aim, three databases, namely EggNOG6, OrthoMCL and OrthoDB, were used to search for orthologous groups (i.e., a set of proteins orthologous to each other) at the Aconoidasida level, specifically those containing at least one protein from *T. parva*. For the EggNOG 6 database, the proteins from the appropriate orthologous groups were extracted from the "e6.og2seqs_and_species.tsv" file. As the proteins contained within these orthologous groups were indexed by

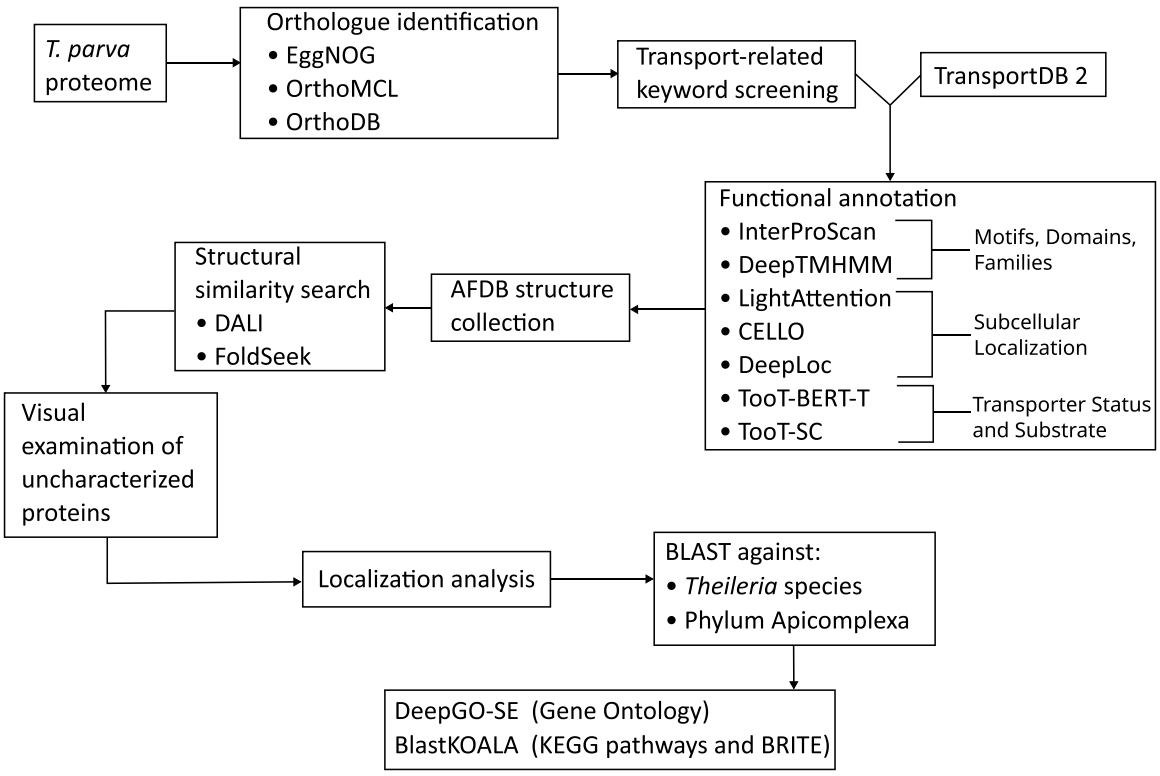


Fig. 1. Flow chart of the *in silico* approach used for identification of potential transport-related proteins (PTPs) of *T. parva*.

UniProt ID, annotations for each of the proteins of interest were obtained from UniProt. The analysis on the OrthoMCL website returned proteins from all species in the database, therefore a filter was applied to retain only proteins within the Aconoidasida (specifically the following genera: *Theileria, Babesia, Cytauxzoon, Plasmodium, Haemoproteus, Hepatocystis*). The OrthoMapper tool from OrthoLoger 3.0.5 was used to retrieve the proteins from the appropriate OrthoDB orthologous groups. Subsequently, the odb11v0_genes.tab file from the OrthoDB website was used to assign annotations to the retrieved proteins.

The identity of each *T. parva* protein was inferred by collecting the annotations of all proteins present in its respective orthologous group. These annotations were screened using a keyword-based approach. If any annotation associated with a particular *T. parva* protein contained any of the keywords of interest, the respective protein was designated as a PTP. The keywords used were: "transport", "carrier", "traffic", "MFS", "facilitator", "transloc", "portin", "channel", "porin", "porter" and "pump".

Additionally, TransportDB 2.0 was searched for *T. parva* transporters to supplement the orthology-derived analysis. This database provides annotations for predicted transporters across sequenced genomes. The proteins identified as transporters were added to the list of PTPs identified by the orthology-based approach.

2.4. Functional annotation of potential transporters

A range of tools was used to predict various properties for the *T. parva* PTPs in order to gain insight into their function and localization.

Motifs, domains and families: To identify protein motifs, domains and families of PTP sequences, the standalone InterProScan 5.65–97.0 tool was used with pre-calculated values (default) as well as locally computed values (-dp flag). The associated functions of identified protein motifs, domains and families were then used to infer the functions of the PTPs.

Transmembrane domains: DeepTMHMM 1.0.24 was installed locally and used to predict transmembrane domains in the PTPs. The software classifies transmembrane proteins into five categories, namely: alpha-helical transmembrane protein without signal peptide (TM), alpha-helical transmembrane protein with signal peptide (SP+TM), beta-barrel transmembrane protein (Beta), globular protein with signal peptide (SP), globular protein without signal peptide (Glob).

Subcellular localization: The subcellular localization of the PTPs was predicted using CELLO 2.5, DeepLoc 2.0 and LightAttention. On the CELLO 2.5 server, the prediction of subcellular localization was done using the Eukaryote and Protein options. The DeepLoc 2.0 server's "high-quality" model was used to predict subcellular localization. The use of LightAttention was a two-step process. First, embeddings were generated by the Bio-Embeddings package using the LightAttention pipeline provided in the package. Subsequently, LightAttention inference was performed using the configuration file included with the tool.

Transporter status and substrate: Potential transport-related proteins were confirmed by TooT-BERT-T, a tool that discriminates between transporters and non-transporters. Furthermore, TooT-SC was used to predict the substrate class of the PTPs.

2.5. Structure-based comparison

AlphaFoldDB (AFDB), a database of predicted protein structures, was used to obtain structural models for the PTPs. The AFDB is indexed by UniPort ID, thus the GenBank protein accession numbers were converted accordingly. The UniProt servers were used to obtain UniProt IDs, UniProt annotations and AFDB URLs for PTPs. UniProt IDs were queried against the AFDB, from which the.pdb structure URL was obtained and the structure subsequently downloaded. The average pLDDT score (predicted local distance difference test) was calculated as the average of the pLDDT scores assigned to each amino acid of the protein structure.

The retrieved protein structures were submitted to the FoldSeek and DALI servers to perform structural similarity searches, identifying other proteins with comparable structures. In the interest of increasing sensitivity, these structural similarity searches were performed on all available AFDB structures, regardless of their pLDDT quality scores. In

Table 1
A list of tools used in this study. Where applicable, Uniform Resource Locators (URLs) and/or Research Resource Identifiers (RRIDs) are included. Dates are in DD/MM/YYYY format.

Tool/Database	Function	References, URLs, RRIDs	Date accessed or last updated
NCBI	Protein sequences Annotations Accessions and identifiers	https://www.ncbi.nlm.nih.gov/ RRID:SCR_006472	11/09/2023 (GenBank proteome)
UniProt	Annotations Accessions and identifiers	(Nightingale et al., 2017) https://www.uniprot.org/ RRID:SCR_002380	28/12/2023 (EggNOG annotations) 15/02/2024 (Potential transporter annotations and AFDB links)
InterProScan 5.65–97.0	Prediction of protein properties such as functional motifs, domains, and family classification	(Jones et al., 2014) https://www.ebi.ac.uk/interpro/ RRID:SCR_005829	N/A
DeepTMHMM 1.0.24	Transmembrane domain prediction	(Hallgren et al., 2022) https://dtu.biolib.com/DeepTMHMM RRID:SCR_025039	N/A
CELLO 2.5	Localization prediction	(Yu et al., 2006) http://cello.life.nctu.edu.tw/ RRID:SCR_011968	13/12/2023
Bio-Embeddings	Protein sequence embedding	(Dallago et al., 2021) https://github.com/sacdallago/bio_embeddings	Latest GitHub version as of November 2023
Light Attention	Localization prediction	(Stärk et al., 2021) https://github.com/HannesStark/protein-localization	Latest GitHub version as of November 2023
DeepLoc 2.0	Localization prediction	(Thumuluri et al., 2022) https://services.healthtech.dtu.dk/services/DeepLoc−2.0/	16/05/2024
TooT-BERT-T	Transporter prediction	(Ghazikhani and Butler, 2023) https://github.com/bioinformatics-group/TooT-BERT-T	Latest GitHub version as of November 2023
TooT-SC	Transporter Substrate Prediction	(Alballa and Butler, 2022) https://github.com/bioinformatics-group/TooT-SC	Latest GitHub version as of November 2023
EggNOG 6	Orthologue search	(Hernández-Plaza et al., 2023) http://eggnog6.embl.de/ RRID:SCR_002456	"e6.og2seqs_and_species.tsv" file (last updated 20 September 2022)
OrthoMCL	Orthologue search	(Li et al., 2003) https://orthomcl.org/orthomcl/app RRID:SCR_007839	Release 6.19, Updated 28/11/2023, Accessed 16/12/ 2023
OrthoDB + Orthologer	Orthologue search	(Kuznetsov et al., 2023) https://www.orthodb.org/ https://orthologer.ezlab.org/ RRID:SCR_011980	24/12/2023
TransportDB 2	Transporter identification	(Elbourne et al., 2017) https://www.membranetransport.org/transportDB2/index.html RRID:SCR_005643	22/1/2024 (*T. parva*) 02/02/2024 (*Babesia equi* and *P. falciparum*)
BLAST+	Sequence similarity search	https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html RRID:SCR_004870	N/A
Protein Data Bank	Experimental protein structures	https://www.rcsb.org/ RRID:SCR_012820	N/A
AlphaFoldDB	Predicted protein structures	(Varadi et al., 2022) https://alphafold.ebi.ac.uk/ RRID:SCR_023662	15/02/2024
DALI	Structural similarity search	(Holm, 2022) http://ekhidna2.biocenter.helsinki.fi/dali/ RRID:SCR_013433	29/02/2024–02/03/2024
FoldSeek	Structural similarity search	(van Kempen et al., 2024) https://search.foldseek.com/search	29/02/2024–02/03/2024
DeepGO-SE	Gene ontology term prediction	(Kulmanov et al., 2023) https://github.com/bio-ontology-research-group/deepgo2	23/08/2024
GOATools	Completion of missing gene ontology terms	(Klopfenstein et al., 2018) https://github.com/tanghaibao/goatools	28/08/2024
KEGG BlastKOALA	Pathway and BRITE term prediction	(Kanehisa et al., 2016) https://www.kegg.jp/blastkoala/	09/07/2025

the case of FoldSeek, the default "3D1/AA" mode was used to obtain hits from the structure database, pdb100. The analysis in DALI used a Z-Score cutoff of 8, since a Z-Score above 8 suggests probable homology between two proteins, while structures with a Z-Score above 20 suggests that two proteins are "definitely homologous" (Holm et al., 2006). The cutoff for FoldSeek was set at 50 % for a probability of a true positive match (van Kempen et al., 2024). This relatively low cutoff was considered acceptable due to the lower sensitivity (86 %) of FoldSeek relative to DALI (van Kempen et al., 2024). Uninformative hits (i.e. multiple hits from the same PDB file) were removed as described below prior to analysis. For the DALI results, which returned descriptions for each chain, only the first was retained when multiple chains from the same PDB entry had the same description. In the case of FoldSeek, only one hit per PDB file was retained since the hits returned the title of the PDB file.

The results were analysed based on the same keywords used for orthology analysis. A number of values representing the prevalence of keyword matches were calculated, including the rank of the first keyword match, and the percentage of hits that matched the keywords. For the purposes of this analysis, a query protein was considered to have

a “hit” when it was structurally similar to a PDB protein and a “match” when the respective hit had a description/annotation matching at least one of the applied keywords. The primary tool used for classification was DALI, with support from FoldSeek. A protein was considered to have structural support for transporter-related function if it had a structural match with a good DALI and/or FoldSeek score. A detailed criterion is presented in the Appendix B.

Proteins predicted as transporters by TooT-BERT-T and/or passed the structural similarity comparison were retained for further analysis. Thus, proteins with support only at the orthology level were excluded.

2.6. Structural analysis of uncharacterized proteins

Predicted transporters with good structural matches, but annotated as “hypothetical”, “uncharacterized” or “integral membrane” proteins by UniProt and NCBI, were further analysed to determine additional structural characteristics by superimposing the AFDB structure and the best structural match (usually the top DALI match). The analysis especially focused on protein topology and amino acids believed to be involved in transporter function. When available, the publication associated with the relevant structure on the PDB was used to identify residues important for transport activity. In some cases, FoldSeek was used to search a PTP structure against datasets containing predicted structures (“afdb-proteome”) for possible annotation of the hypothetical protein.

2.7. Conservation of PTPs in Theileria species and among apicomplexa

To determine the distribution of PTP homologues across genus *Theileria*, BLASTP protein databases were individually created using the RefSeq genomes of a transforming *Theileria* species, *T. annulata* (GCF_000003225.4), and two non-transforming, *T. orientalis* strain Shintoku (GCF_000740895.1) and *T. equi* strain WA (GCF_000342415.1). BLASTP was used to query the PTPs against these databases with an E-value threshold of 0.05, for identification of orthologues in the other *Theileria* species. These species were selected for this analysis as they were the only *Theileria* species with annotated genomes on the NCBI.

To determine the conservation of PTPs among Apicomplexa, a local installation of BLASTP 2.12.0 was used to query the PTPs against the official NCBI “refseq_proteome” database and restrict the results to the taxon Apicomplexa, using an E-value threshold of 0.05. The query was performed on 08/07/2025 using the following options: -remote -evalue 0.05 -db refseq_protein -entrez_query "Apicomplexa [organism]".

2.8. Filtering of high-confidence transmembrane transport proteins

To identify proteins with more direct roles in transmembrane (TM) transport, the PTPs were further analysed using gene ontology (GO) terms. DeepGO-SE was used to predict GO terms with the default threshold of 0.1. GOATools was used to infer parental terms that were missing. The high-confidence set contained proteins that had at least one TM helix, had the “transmembrane transporter activity” molecular function GO term (GO:0022857), and were supported by both TooT-BERT-T and structural similarity, with the latter limited to proteins that had AFDB structures with an average pLDDT score ≥ 70. In general, pLDDT scores above 70 indicate a correct local backbone prediction, and the same threshold as a reference for the average pLDDT was used.

2.8.1. Pathway analysis

The KEGG (Kyoto Encyclopedia of Genes and Genomes) database was used to assign pathways and BRITE (a classification system within KEGG incorporating more diverse biological data than the pathways system) terms to the high-confidence subset of the PTPs. To achieve this, the sequences of the PTPs were queried on BlastKOALA against the “Eukaryotes” reference dataset.

3. Results and discussion

3.1. Initial screening for transport-related proteins

Of the entire 4051 proteins in the *T. parva* proteome, 217, 259, 121, and 121 were detected as PTPs by EggNOG, OrthoMCL, OrthoDB and TransportDB respectively. In total, there were 334 proteins supported by at least one of these tools and their distribution between the different tools is shown in Fig. 2A (and detailed in Table C.1). Approximately 20 % (68/334) of these proteins were detected as PTPs by all four tools. About 44 % (146/334) were flagged by only a single tool, with 81 proteins being attributed to OrthoMCL, 44 to EggNOG, 20 to TransportDB, and only one to OrthoDB. In general, there was a good agreement between proteins detected by EggNOG and OrthoMCL, accounting for almost half of the proteins detected as PTPs (163/334).

As only one protein was unique to OrthoDB, this may be a consequence of a shortcoming in the approach used for assigning annotations to proteins in the OrthoDB orthologous groups. The annotations were obtained from the odb11v0_genes.tab file, though in many cases, rather than containing an annotation, protein accession numbers to various databases were given instead. The reduced number of annotations that could be extracted for each orthologous group may have made it less likely for keyword matches to occur, thus reducing the sensitivity of the analysis for this database.

3.2. Structure-based comparison and TooT-BERT-T prediction

As the structure of a protein tends to be most conserved, structural similarity may indicate evolutionary relationships which are lost at the sequence level (Lubec et al., 2005). However, structural similarity does not guarantee evolutionary relationship, as convergent evolution can lead to unrelated proteins of similar structures (Lubec et al., 2005). Nevertheless, it can serve as convincing evidence in combination with sequence orthology data.

The UniProt search for the 334 PTPs returned 201 entries, of which 181 had AFDB structures (Fig. 2B). These proteins were subjected to DALI and FoldSeek analyses to find experimentally resolved structures from the PDB, with which they share structural similarity. DALI and FoldSeek returned hits for 156 and 147 proteins respectively. Only 108 of these PTPs were structurally similar to proteins labelled with transport-related terms, based on the criteria described in Appendix B. Of the 334 PTPs, 160 were supported by transporter prediction through TooT-BERT-T, comprising 80 proteins supported only by TooT-BERT-T and 80 PTPs supported by both TooT-BERT-T and structural similarity analysis. Furthermore, 28 transporter proteins were predicted by structural analysis only (Fig. 2C, detailed in Table C.2). Consequently, a total of 188 proteins were subjected to further analysis, incorporating both TooT-BERT-T and structural similarity analyses.

3.3. Annotation of uncharacterized potential transport-related proteins

Within the 334 PTPs, 80 were uncharacterized proteins. Of these, 24 were part of the 188 proteins supported by TooT-BERT-T and structural similarity, which were assigned tentative annotations. Following the annotation analysis, these 24 proteins could be categorized based on four possible transporter types, namely metal, Major Facilitator Superfamily (MFS), nucleotide-sugar, and hENT1(human Equilibrative Nucleoside Transporter 1)-like (Fig. 2D). Notably, besides non-transporters which made up 42 % of the analysed proteins, most of them were associated with hENT1-like transporters (29 %). The tentative annotations for these 24 uncharacterized proteins and three related characterized proteins are given in Table D.1.

To assist in the annotation of the uncharacterized proteins, FoldSeek was used to perform additional structural similarity searches against the datasets available to it, including the “afdb-proteome” dataset of predicted structures and the “pdb100” dataset of experimentally resolved

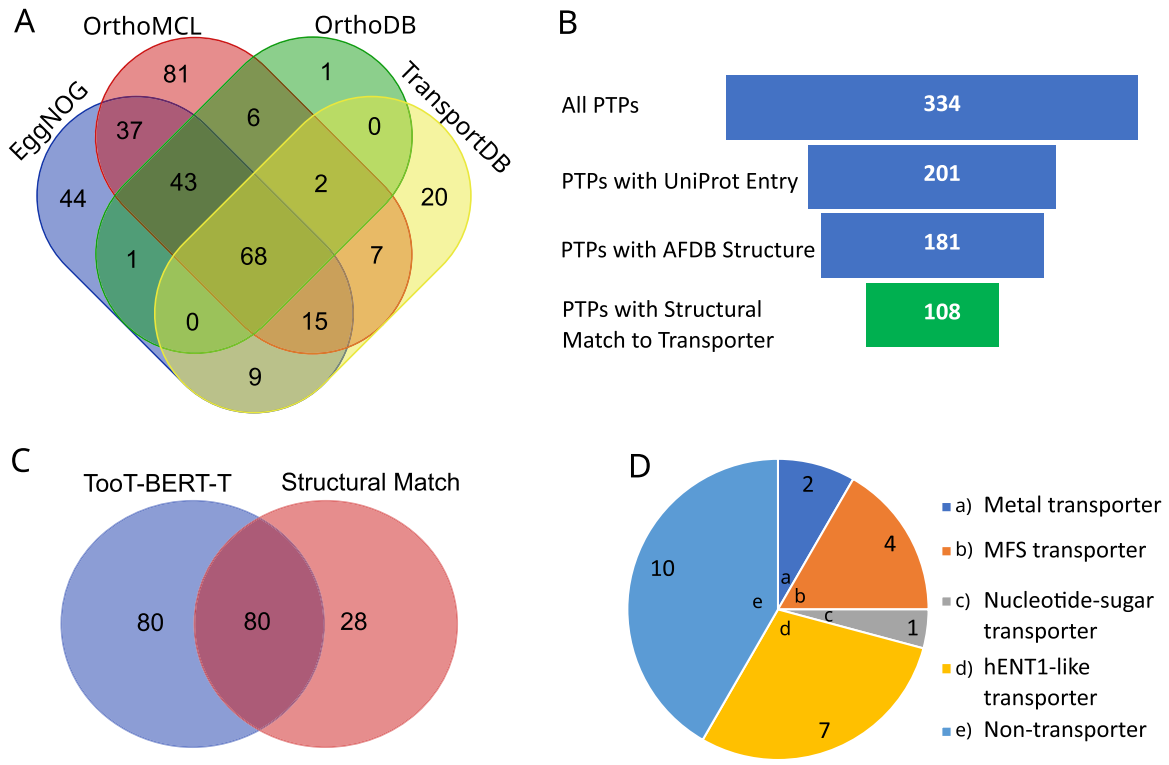


Fig. 2. Bioinformatic analysis of potential transport-related proteins (PTPs). (A) Distribution of 334 PTPs between tools used for orthology analysis (EggNOG, OrthoMCL, OrthoDB) and TransportDB. (B) Number of PTPs detected at each stage of the structural similarity analysis. (C) Summary of the PTPs which had support for a transporter designation based on the TooT-BERT-T transporter prediction and/or the structural similarity analysis. (D) Classification of the 24 uncharacterized PTPs into tentative categories. MFS, Major Facilitator Superfamily; hETN1, human Equilibrative Nucleoside Transporter 1; AFDB, AlphaFold Database.

structures. Furthermore, some proteins did not have up-to-date UniProt records, but had older versions available. For example, EAN32839.2 did not exist on UniProt, but EAN32839.1 did. In these cases, detailed structural analysis using DALI and manual inspection of superimposed structures was not performed; instead, only FoldSeek queries were performed to retrieve the annotations of structurally similar proteins. All FoldSeek results reported below had a probability score of 1.

Metal transporters: Two uncharacterized transporters could be designated as metal transporters, namely EAN31293.2 and EAN32839.2. Neither of these proteins had entries on UniProt, though the previous versions of these protein records (eg, EAN31293.1) could be found. Both proteins had TooT-BERT-T support.

EAN31293.2 is a 397 amino acid (aa) protein with two DeepTMHMM predicted TM helices. On NCBI, it had a region described as "*Saccharomyces cerevisiae* inner mitochondrial membrane Mg^{2+} transporters Mfm1p and Mrs2p-like family". Its predecessor on UniProt (EAN31293.1) was annotated as "EF-hand domain-containing protein". OrthoDB, OrthoMCL and EggNOG all returned "magnesium transporter" from various parasites, including *Plasmodium, Theileria* and *Babesia* species, as orthologues to this protein. A FoldSeek analysis was performed on the previous version, and the top pdb100 hit was "*Saccharomyces cerevisiae* Mrs2 magnesium channel (PDB 3RKG)", with sequence identity of 15.1 %. The top hit in the afdb-proteome database was the "*Plasmodium falciparum* magnesium transporter" (UniProt A0A143ZWJ6, 468 aa) with a sequence identity of 28.5 %. Thus, the orthology and structural similarity data both supported magnesium transport as the annotation for this protein. Additionally, InterProScan Gene3d, PANTHER and CDD returned magnesium transport related hits. TooT-SC predicted inorganic cation as the substrate class with a score of 0.99.

EAN32839.2 is a 501 aa protein with three DeepTMHMM predicted TM helices, and its predecessor was annotated as "CNNM transmembrane domain-containing protein" on UniProt. The NCBI entry contained a region described as "Hemolysin-related protein, contains CBS domains". The InterProScan PANTHER description was "METAL TRANSPORTER CNNM". Orthologues detected from EggNOG, OrthoMCL, and OrthoDB were annotated with "Protein MAM3", "CNNM transmembrane domain-containing protein" (in one case "Metal transporter CNNM4"), and "CBS domain-containing protein". FoldSeek analysis of the predecessor returned "murine magnesium transporter CNNM2 (PDB 4P1G)" as the top pdb100 hit with a sequence identity of 30.9 %. In afdb-proteome, the top hit was *Drosophila melanogaster* "unextended protein" (UniProt A0A0B7P9G0, 834 aa) with a sequence identity of 30 %. This *Drosophila* protein had a Gene Ontology annotation on UniProt for "magnesium ion transmembrane transporter activity". The second hit was to *Dania rerio* "Cyclin and CBS domain divalent metal cation transport mediator 4b" (UniProt F1Q7I7, 778 aa) with a sequence identity of 29.4 %, a protein annotated as "metal transporter" on UniProt. However, these hits were significantly longer than EAN32839.2. Based on the InterProScan results, orthology and structural similarity, it is possible that EAN32839.2 is a metal transporter. TooT-SC supported this classification, with a score of 0.97 for inorganic cation substrate classification.

Major facilitator superfamily transporters: Four proteins had support for the MFS classification, EAN34036.2, KAF5153638.1, EAN33599.1, and EAN31178.1. UniProt only had records for the previous versions of EAN34036.2 and KAF5153638.1.

EAN34036.2 is a 441 aa protein with 12 predicted DeepTMHMM TM helices, and its predecessor was marked obsolete on UniProt. InterProScan Gene3D and SUPERFAMILY returned MFS annotations, while the PANTHER description was "transporter, putative-related". EggNOG found orthologues annotated as "Metabolite/drug transporter", while OrthoMCL found "Metabolite/drug transporter", "mitochondrial import inner membrane translocase", and "rna polymerase subunit". An AFDB structure of the previous version existed, allowing for FoldSeek analysis that resulted on the top pdb100 hit of *Escherichia coli* "xylose transporter

XylE" (PDB 4JA4), with sequence identity of 11.9 %. The top afdb-proteome hit was to *Dictyostelium discoideum* "MFS domain-containing protein" (UniProtID Q8T2G9, 498 aa), with sequence identity of 12.4 %. Thus, the InterProScan, orthology and structural data suggested MFS transporter activity. TooT-SC predicted inorganic cation as the substrate class with a score of 0.88, which agrees with the proton-symporter nature of XylE (Zhao et al., 2020).

KAF5153638.1 is a 523 aa protein equivalent to XP_765021.2, which on the NCBI was annotated as "putative integral membrane protein". This protein has 12 DeepTHMHH helices. EggNOG and OrthoMCL identified orthologues annotated as "transporter [putative]" from various *Plasmodium* species. The KAF5153638.1 locus (TpMuguga_02g02510) has a gene range that appears to fully encompass the loci of the obsolete XP_765020.1 and XP_765021.1 (TP02_0454 and TP02_0454 respectively). Both XP_765020.1 (UniProt Q4N536, 154 aa) and XP_765021.1 (UniProt Q4N535, 357 aa) had UniProt entries, thus FoldSeek analysis could be performed. Both old UniProt entries were uncharacterized. There were no good FoldSeek results across any dataset for XP_765020.1. While for XP_765021.1, the top pdb100 hit was to the rat GLUT5 transporter (PDB 4YBQ, sequence identity 10.8 %), an MFS sugar transporter, and the top afdb-proteome hit was *Shigella dysenteriae* "Low-affinity L-arabinose transport system proton symport protein" (UniProt Q32C80, 409 aa, sequence identity 14.5 %). UniProt's entry for the L-arabinose transport system, PANTHER, CDD, Gene3D, ProSite, and SUPERFAMILY all report MFS-related hits. Since KAF5153638.1 fully encompassed XP_765021.1, and has a total of 12 predicted helices, it is possible that it is an MFS transporter, potentially of sugar-like substrates.

EAN33599.1 is a 640 aa protein with 12 DeepTMHMM predicted TM helices. It had EggNOG orthologues in *Plasmodium* annotated as MFS transporters, and InterProScan SUPERFAMILY classification placed it in the MFS transporter family. Consistently, the AFDB structure of EAN33599.1 (average pLDDT of 75.0) was similar to MFS transporters, according to both DALI and FoldSeek results. The best DALI hit was the *Syntrophobacter fumaroxidans* monocarboxylate transporter (SfMCT), a bacterial MFS transporter (PDB 6G9X-B, Z-Score 25.0, RMSD 2.8). The top FoldSeek result in the pdb100 dataset was also a structure of SfMCT (PDB 6HCL-B), with a sequence identity of 10.8 %. However, there was almost no conservation of key amino acids between EAN33599.1 and SfMCT (Bosshart et al., 2019), therefore, monocarboxylate transport activity similar to SfMCT could not be inferred. Additionally, TooT-SC predicted "amino acids and derivatives" as the substrate for EAN33599.1, with a score of 0.89. This classification is consistent with the identity of the *P. falciparum* orthologue detected by EggNOG (UniProt Q88II64, ORF PF3D7_1129900), which is a 609 aa *P. falciparum* Apicomplexan Amino Acid Transporter 4 (PfApiAT4) (Wichers et al., 2021). Furthermore, FoldSeek results from the "CATH50" dataset returned PfApiAT8 and PfApiAT4 as the top hits, with sequence identities of 20.2 % and 15.8 % respectively. Thus, multiple data suggested that this protein is an MFS transporter, and possibly an amino acid transporter.

EAN31178.1 is a 439 aa protein with 11 predicted DeepTMHMM TM helices. The EggNOG and OrthoMCL orthology analyses both identified nucleoside transporters as orthologues, with the latter also identifying a "CLN3 protein domain-containing protein". The CLN3 protein's domain structure resembled that of the major facilitator superfamily, and had some sequence similarity to Equilibrative Nucleoside Transporter (ENT) proteins (Mirza et al., 2019). The lack of structural matches in DALI and FoldSeek could potentially be attributed to the very low quality (average pLDDT score of 39.3) of the predicted AFDB structure. Thus, EAN31178.1 is possibly an MFS transporter, potentially a nucleoside transporter.

Nucleotide-sugar transporter: EAN31625.1 is a 354 aa protein with 10 predicted DeepTMHMM helices. Orthology analysis identified orthologues annotated as "Sugar phosphate transporter domain-containing protein", "GDP-fucose transporter 1", "Glideosome associated protein 40" (GAP40) and "metabolite drug protein". The OrthoDB orthologous group description was the glideosome-associated protein. Considering size, the GAP40 is approximately 450 aa, while the other orthologues were ~350 aa. The AFDB predicted structure exhibited high quality (average pLDDT score of 84.3). Both the DALI and FoldSeek pdb100 top hit was "triose-phosphate/phosphate translocator". The top FoldSeek afdb-proteome hit was a GAP40, followed by a "sugar phosphate domain containing protein"; however, the majority of subsequent hits were associated with GDP-fucose transporters. There was therefore evidence to support both a sugar-phosphate transporter and a GAP40 annotation, though the size of the protein suggested that the former was more likely. Notably, the afdb-proteome GAP40 hit had a relatively high sequence identity of 48.7 % when compared to the 15–20 % of the sugar transporter hits. The EAN31625.1 had its top DALI match to a triose-phosphate/phosphate translocator PDB 5Y78-B (Z-Score 23.9, RMSD 4.0), with good overall structural agreement. Despite this agreement, the lack of any important conserved residues made it unlikely that EAN31625.1 could function as a triose-phosphate transporter. The top four DALI hits were all the same triose-phosphate transporter, belonging to PDB 5Y78 and 5Y79. The fifth best hit was PDB 6QSK-B (Z-Score 17.6, RMSD 3.8), a structure of Vrg4, a yeast GDP-mannose transporter. There was generally good agreement between the 6QSK-B structure and that of EAN31625.1, except for some helices being tilted in different directions. There were some conserved residues between 6QSK-B and EAN31625.1 (Parker et al., 2019). The Tyr281 was replaced with Trp278, which still has a bulky side chain as required. Tyr28 was replaced with Phe29, resulting in loss of the ability to form H-bond to the substrate. Similarly, Lys289 being replaced by Val286 presumably abolishes the interaction with the substrate phosphate. On the other hand, Ser269 is found as Ser266, and Asn220 is found as the similar Gln211. Thus, EAN31625.1 may potentially have some ability to bind to nucleotide sugars, though the lack of conservation for the essential Tyr28 is concerning. The comparison between important residues in the reference proteins PDB 5Y78 and PDB 6QSK, and the uncharacterized EAN31625.1 is summarized in Table 2.

Human Equilibrative Nucleoside Transporter 1-like transporters: The proteins bearing similarity to the hENT1 were the largest subset of uncharacterized proteins (n = 7). As hENT1 is structurally similar to *Plasmodium falciparum* ENT1 (PfENT1) (Wang et al., 2023), the latter was also used as a reference to identify residues important for transporter function.

Two studies on PfENT1 found that Trp53, Gln135, Asp287, Arg291 and Ser290 were crucial for nucleoside transport activity (Dillague and Akabas, 2023; Wang et al., 2023). In PfENT1 and hENT1, the Gln residue was found to interact with the purine moiety, while Arg and Asn interacted with the ribose (Wang et al., 2023; Wright and Lee, 2019). Comparison of PfENT1 and hENT1 revealed that hENT1 had perfect counterparts for Trp53, Gln135, Asp287, and Arg291, while PfENT1's Ser290 was found as a Gly in hENT1. Also, while not being found strictly necessary for PfENT1 transport activity, Ser49 was believed to stabilize Gln135 (Dillague and Akabas, 2023; Wang et al., 2023), and was found as Thr in hENT1.

For a more comprehensive analysis, the set of uncharacterized proteins with structural similarity to hENT1 (n = 7) was expanded to include annotated proteins whose first hit was the hENT1 protein. Three annotated proteins were identified, resulting in a total of ten proteins (Fig. 3). For eight of these, the DALI Z-score exceeded 20, and the top FoldSeek hit was hENT1. Except for EAN32009.1, which was not supported by TooT-BERT-T, all other proteins from these eight had both TooT-BERT-T and structural support. A comparison of the key residues and some loop lengths between these eight proteins, hENT1, and PfENT1 is shown in Table 3. To confirm similarity to the PfENT1 protein which would also be used as a reference, we observed that seven of the ten hENT1-like proteins had a DALI hit to PfENT1 (PDB 7WN0 or 7WN1) with a Z-score of at least 20, while EAN33054.1, EAN32702.1, and EAN33799.1 had Z-scores of 18.2, 16.6, and 12.1 respectively.

Table 2
Comparison of the key amino acids in a triose-phosphate translocator and a GDP-mannose transporter to the *T. parva* protein EAN31625.1.

Reference	EAN31625.1	Function	Notes
Triose-phosphate translocator (5Y78-B) (Lee et al., 2017)			
Lys204	Ala126	Substrate recognition Ionic bond to phosphate	Conserved in higher plants plastidic phosphate translocators (pPTs) Lys204Ala greatly reduced transport
Lys362	Ser285	Substrate recognition Ionic bond to phosphate	Conserved in higher plants pPTs Lys362Ala greatly reduced transport
Arg363	Val286	Substrate recognition Ionic bond to phosphate	Conserved in higher plants pPTs Arg363Ala greatly reduced transport
Tyr339	Lys262	Substrate recognition H-bond to phosphate and carboxyl	Conserved in higher plants pPTs Tyr339Phe greatly reduced transport
His185	Leu106	Substrate recognition Ionic bond to carboxyl	His185Ala greatly reduced transport
GDP-mannose transporter (6QSK) (Parker et al., 2019; Parker and Newstead, 2017)			
Tyr28	Phe29	Essential, H-bond with GMP phosphate	Tyr28Ala abolished transport
Tyr281	Trp278	Ligand discrimination	Not strictly conserved in orthologues, but bulky side chain required.
Lys289	Val286	Recognition and Transport, salt bridge with GMP phosphate	Conserved in orthologues, part of GALNK motif for sugar discrimination
Ser266	Asn263	Possibly ligand discrimination	Ser266Ala greatly reduced GMP affinity, slightly reduced GDP-mannose affinity
Ser269	Ser266	Possibly ligand discrimination	Ser269Ala mutation did not significantly affect substrate affinity
Asn220	Gln211	Purine selectivity	
Asn221	His212	Purine selectivity, favours guanine	

Six of the eight proteins were divided into two groups based on structural traits. The first group, consisting of EAN32811.1, EAN31807.1 and EAN33784.1, had an acidic amino acid (Glu or Asp) instead of the Gln158, and contained counterparts for both Asp341 and Arg345. A counterpart for Thr25 was also found in two of the three proteins, EAN32811.1 and EAN33784.1. Relative to hENT1, these proteins had shorter loops between TM6–7 but longer loops between TM8–9 and TM10–11; they also contained an additional C-terminal helix modelled in a different orientation to the TM helices. The length of group 1 proteins ranged from 534 to 556 aa, and had between 11 and 13 DeepTMHMM helices, which is expected for ENT and MFS transporters. Furthermore, the analysis in FoldSeek against the afdb-proteome dataset, returned ENT proteins in all cases as the top characterized hit, with sequence identities of 10.8–12.6 %. The second group consisted of EAN32006.1, EAN32008.1 and EAN32009.1. In all three of these, Gln158 was conserved. Thr25 was found as Ser in EAN32006, but as non-polar amino acids in EAN32008.1 and EAN32009.1. All three had non-polar amino acids instead of Asp341, and either Ser or Thr instead of Arg345. This group had loops between TM8–9 and TM10–11 comparable to those found in hENT1. Also, the lengths of the loops between TM6–7 were relatively similar, except in the case of EAN32009.1, which had a considerably longer loop (approximately 90 aa vs 50–65 aa). Similar to the first group, all group 2 proteins had an additional helix at the C-terminal side, although EAN32006.1 and EAN32009.1 also had additional helices on the N-terminal side. Overall, the length of group 2 proteins ranged from 483 to 525 aa, and contained 11 DeepTMHMM helices, as expected for ENT and MFS proteins. They bear structural similarity to experimentally resolved ENT proteins on the PDB, as well as predicted structures of ENT proteins of the afdb-proteome datasete, with sequence identities in the range of 12.5–14.9 %. Furthermore, EAN32008.1 and EAN32009.1 were annotated as "MFS transporter superfamily" proteins by InterProScan SUPERFAMILY.

The remaining two of the eight proteins, namely EAN33054.1 and EAN31530.1, could not be placed in either of the above groups based on overall structure similarity and key conserved amino acids. None of the key hENT1 residues were present in EAN33054.1, and only Asp341 and Arg345 were conserved in EAN31530.1. These proteins had FoldSeek hits in the afdb-proteome dataset to nucleoside transporter proteins with sequence similarity in the range of 10.2–13.7 %. Data from OrthoMCL suggested relation to group 1.

The two eliminated proteins were EAN32702.1 and EAN33799.1 (Fig. 3, indicated in italicised red font). In the case of EAN32702.1, the Z-Score to hENT1 was 16.7, but the top FoldSeek pdb100 hit was an MFS proton:xylose symporter, while EAN33799.1 had a DALI Z-score of 14.9 for hENT1, and no FoldSeek pdb100 hits. As EAN33799.1 and EAN32702.1 had only 8 and 7 TM helices respectively, they were considered unlikely to have nucleoside transport function.

Non-transporters: There were ten proteins that did not have convincing support for classification as transporters. EAN31968.1 and EAN33934.2 had some evidence supporting transporter function, but not enough to confirm their identity as transporters (Appendix E). EAN32476.1, EAN32765.1, EAN33420.1, and EAN33761.1 had almost no evidence to support transporter function. Finally, EAN34131.2, EAN33749.1, and EAN33155.1 had support as members of the Sec61 protein translocation complex, a vesicle transport complex, and a nuclear pore complex, respectively.

Finally, KAF5153161.1 is a 520 aa protein with 11 DeepTMHMM helices. Based on OrthoMCL, it clustered with group 1 hENT1-like proteins. However, as it did not have a UniProt entry and associated AFDB structure, it was not included in the structural similarity analysis, which explains its absence from Fig. 3 and Table 3. On the other hand, TooT-BERT-T supported the transporter designation. No useful information was available to further discern its identity. Thus, it is tentatively categorized as a non-transporter.

3.4. Localization of potential transport proteins

The subcellular localization of a protein not only provides context to better understand its function, but can also affect the accessibility of the protein to pharmaceuticals. The localization predictions of CELLO, DeepLoc, and Light Attention were compared and whenever possible, a consensus localization was assigned to PTPs for which at least two of the three tools agreed. Furthermore, the DeepTMHMM predictions were used to classify proteins as transmembrane (TM) or non-TM based on the presence or absence of TM helices, respectively (Fig. 4).

Inspection of the DeepTMHMM results for the 188 PTPs revealed that 60 % (n = 113) had TM helices. Of the 188 PTPs, the largest group (n = 54) was predicted to localize to the plasma membrane, and all those proteins had at least one TM helix. Transmembrane helices were also present in all lysosome/vacuole (n = 8) and Golgi (n = 3) proteins, as well as most endoplasmic reticulum (ER) proteins (11 of 12). Notably, the proteins predicted to localize to the cytoplasm (n = 28) and nucleus (n = 17), as well as the majority of mitochondrial PTPs (20 of 31) had no TM helices. There were 35 proteins with no consensus between the three localization prediction tools, most of which (n = 26) had TM helices. More detailed distribution of these PTPs into classes/types is discussed below, and the protein accession numbers are provided in the tables of Appendix F (Tables F.1 – F.8).

Plasma membrane: The 54 proteins predicted to localize to the plasma membrane were distributed in 16 protein categories (Fig. 5A). The ABC (ATP-Binding Cassette) family and MFS family proteins accounted for nearly half of these PTPs. Three plasma membrane PTPs are uncharacterized, including two MFS proteins, namely EAN32476.1

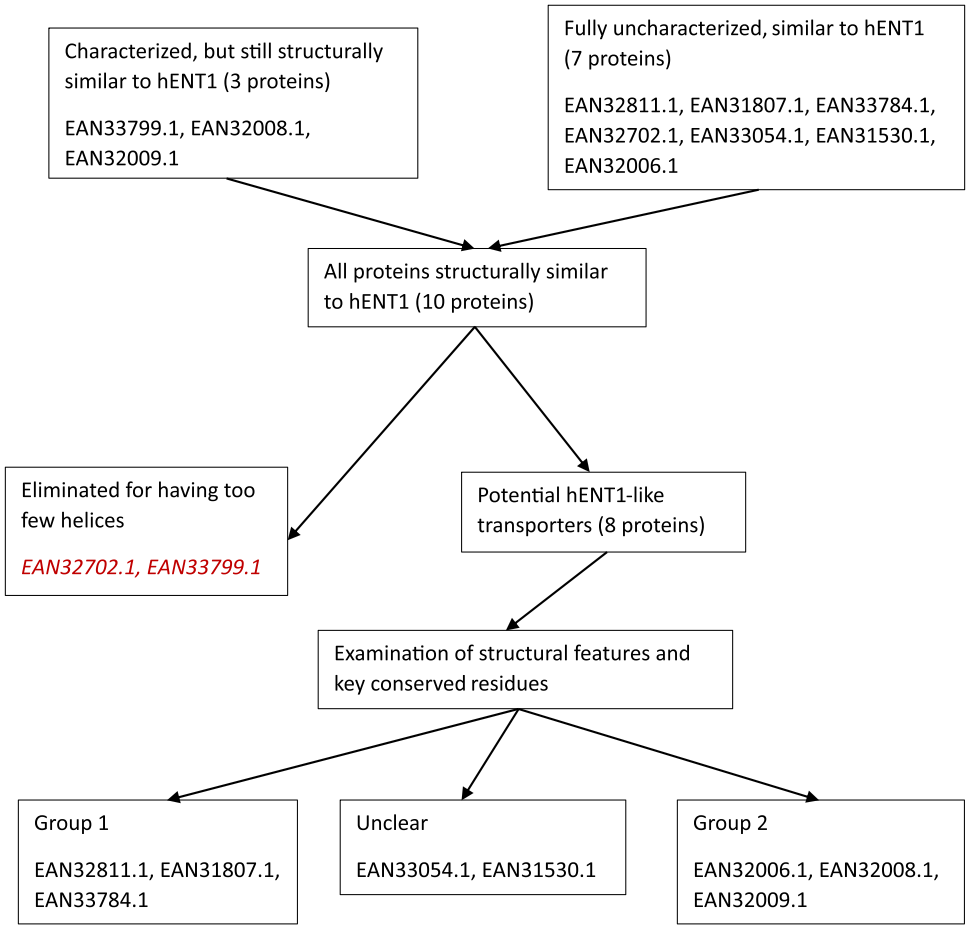


Fig. 3. Overview of the sorting of hENT1-like proteins into group 1 and 2 and an undefined group. Proteins in black font were part of the initial set of hENT1-like proteins and remained in the final set. Proteins in italicised red font were in the initial set but eliminated due to being too short and thus not containing the expected 11–12 transmembrane helices.

Table 3
Comparison of key amino acids in human and *Plasmodium* ENT1s, and eight hENT1-like PTPs.

Trait	hENT1 (6OB6)	PfENT1 (7WN1)	EAN31807.1	EAN32811.1	EAN33784.1	EAN32006.1	EAN32008.1	EAN32009.1	EAN33054.1	EAN31530.1
Average pLDDT score	N/A	N/A	70.9	68.0	69.0	79.2	79.8	83.4	69.5	74.8
Thr25	Thr25	Ser49	Pro34	Thr34	Thr34	Ser38	Ala41	Leu49	Ile45	Ala34
Trp29	Trp29	Trp53	Ile38	Ile38	Ile38	Ile42	Ile45	Ile53	Lys49	Leu38
Gln158	Gln158	Gln135	Glu136	Glu136	Asp136	Gln143	Gln146	Gln154	Lys155	Thr144
Asp341	Asp341	Asp287	Asp301	Asp300	Asp300	Leu306	Ala323	Gly353	Gly319	Asp309
Gly344	Gly344	Ser290	Gly304	Gly303	Gly303	Gly309	Ala326	Gly356	Gly322	Gly312
Arg345	Arg345	Arg291	Arg305	Arg304	Arg304	Ser310	Thr327	Thr357	Thr323	Arg313
Additional N-Terminal helix	No	No	No	No	No	Yes	Yes	Yes	Yes	No
Additional C-Terminal Helix	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Loop[b] TM 1–2	45	5	15	10	10	15	15	15	5	5
Loop[b] TM 6–7	60	30	30	30	30	55	65	90	30	30
Loop[b] TM 8–9	10	15	60	50	50	15	10	15	55	35
Loop[b] TM 10–11	5	10	40	35	30	10	5	5	15	15

[a] pLDDT: predicted local distance difference test
[b] Loops are identified by their flanking TM helices and their lengths are broad estimates and can vary by around ± 5 amino acids

and EAN34036.2, and a metal transporter, EAN32839.2.

In general, transport proteins localized to the parasite plasma membrane facilitate the acquisition of various nutrients, including nucleotides and amino acids, and are involved in the removal of waste products, such as lactic acid and ammonia (Meier et al., 2018). Moreover, access of drugs to transport proteins of the plasma membrane should be easier than to the intracellular compartments, as there are fewer membranes that need to be traversed. Furthermore, the utilization of these drugs will prevent their modification within the parasite cytoplasm and they will not be expelled from the cell by drug exporters (Meier et al., 2018).

Cytoplasm: The 28 cytoplasmic PTPs were distributed in 13 protein categories (Fig. 5B). The largest groups were the karyopherins and ATPase family proteins, with six members in each. Notably, a porin

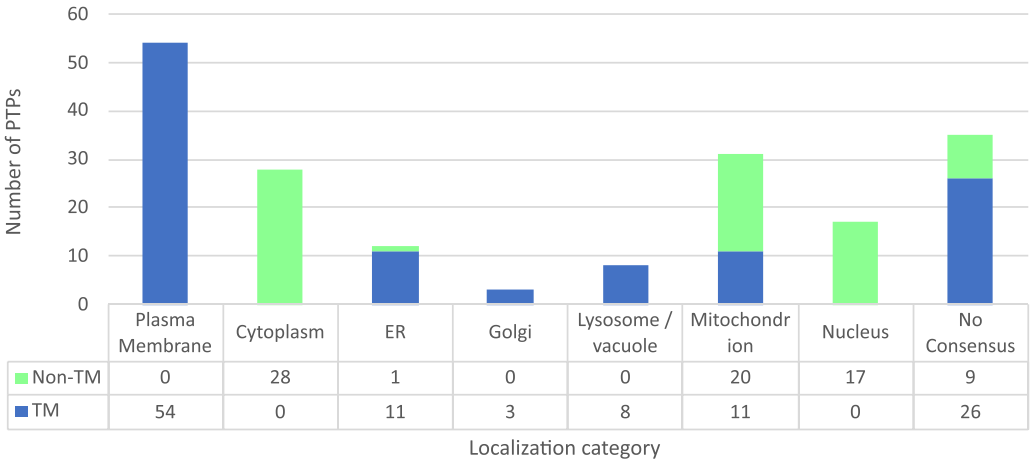


Fig. 4. Distribution of potential transport-related proteins (PTPs) into different localization classes, as well as distinction between transmembrane (TM) and non-transmembrane (Non-TM) proteins. ER, endoplasmic reticulum.

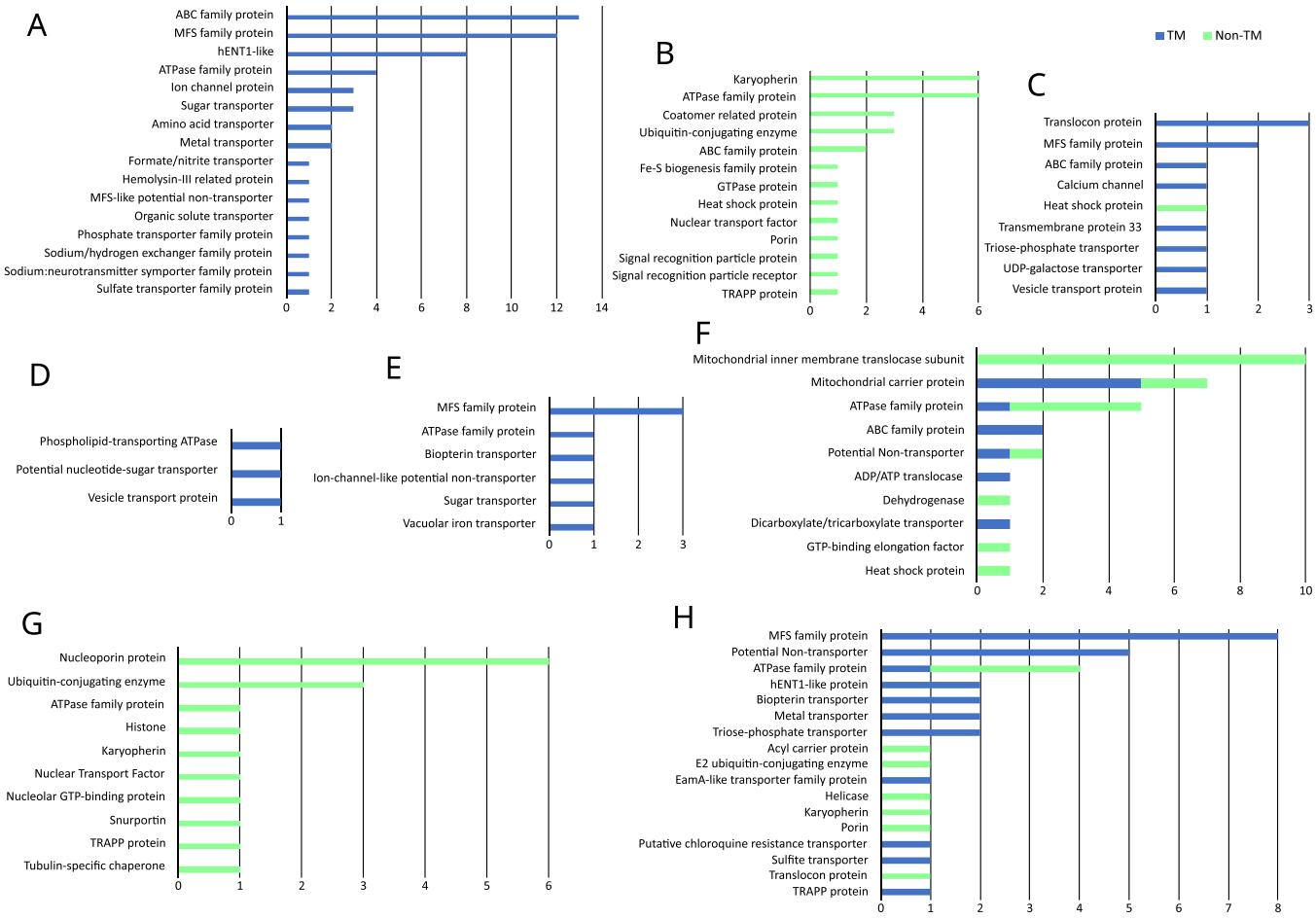


Fig. 5. Graphs showing the number of potential transport-related proteins (PTPs) assigned to various functional categories based on their annotations, with each graph representing a different subcellular localization. The x-axis is the number of PTPs per category, while the y-axis is the category label. Blue represents proteins with transmembrane (TM) helices, while light green represents those without. (A) Plasma membrane; (B) Cytoplasm; (C) Endoplasmic reticulum; (D) Golgi apparatus; (E) Lysosome/vacuole; (F) Mitochondrion; (G) Nucleus; (H) No consensus between tools. ABC, ATP Binding Cassette; MFS, Major Facilitator Superfamily; hENT1, human Equilibrative Nucleoside Transporter 1.

protein was included, and despite not having a TM helix predicted by DeepTMHMM, it may be a transmembrane protein with a beta-barrel structure, which is supported by the "membrane bound" solubility prediction of BioEmbeddings.

Endoplasmic reticulum: There were 12 proteins predicted to localize to the ER, distributed in nine categories (Fig. 5C). Seven of the nine protein categories consisted of a single PTP, while three proteins were associated with the translocon, and two with MFS protein categories. All 12 PTPs had at least one DeepTMHMM-predicted TM helix, except for a single heat shock protein.

Golgi: Only three PTPs were predicted to localize to the Golgi, namely a phospholipid-transporting ATPase, a vesicle transport protein, and a potential nucleotide-sugar transporter (EAN31625.1) (Fig. 5D).

Lysosome / vacuole: Eight proteins were predicted to localize to the lysosome/vacuole, distributed in six protein categories (Fig. 5E). The largest group being the MFS proteins (n = 3) and one of these is uncharacterized (KAF5153638.1).

Mitochondrion: The 31 proteins predicted to localize to the mitochondrion were distributed in 10 protein categories (Fig. 5F). The largest group was the "Mitochondrial inner membrane translocase subunit" (n = 10), followed by "Mitochondrial carrier protein" (n = 7) and "ATPase family protein" (n = 5). Notably, none of the mitochondrial inner membrane translocase subunits were predicted to have TM helices. In contrast, transmembrane helices were predicted in five of the seven mitochondrial carrier proteins and in only one of the five ATPases. All five ATPases were annotated as ATP synthase subunits.

Nucleus: There were 17 proteins predicted to localize to the nucleus, though none had DeepTMHMM-predicted TM helices. These PTPs were distributed in 10 protein categories (Fig. 5G). The biggest group was the nucleoporins (n = 6), followed by ubiquitin-conjugating enzymes (n = 3). One of the nucleoporins, EAN33155.1, was uncharacterized, although it may potentially be a Nup214-like protein (refer to Appendix E for more detail).

No consensus: There were 35 proteins for which a localization consensus could not be reached by the three tools, and these were distributed across 17 protein categories (Fig. 5H). The largest group of these PTPs was the MFS proteins (n = 8). Uncharacterized proteins contributed two members of the MFS proteins, one of the metal transporters, as well as all members of the "Potential Non-transporters" and the "hENT1-like proteins". The translocon protein was the potential non-transporter EAN34131.2, likely representing a Sec61 subunit. The transmembrane ATPase family protein, EAN32807.1, was annotated as a calcium transporting ATPase (on UniProt), and could therefore also be considered a metal transporter. The ATPase proteins further included two ATP synthase subunits. One of the non-transporters is KAF5153161.1, which was orthologous to the hENT1-like proteins but had no structure to further support its identity. The porin with no TM domain was annotated on UniProt as the β-barrel TOM40 protein.

The largest categories of PTPs are the ATPase, MFS, and ABC proteins. The ATPase family proteins were detected at all localizations except the ER. MFS proteins were found on the plasma membrane, ER, and lysosome/vacuoles. ABC proteins were found on the plasma membrane, cytoplasm, ER, and mitochondrion. Metal transporters were identified on the plasma membrane, ER and lysosome/vacuoles. Karyopherins were the largest group of proteins with supportive transport roles, and were detected in the cytoplasm and nucleus, consistent with their roles in nuclear import and export (Wing et al., 2022). Two vesicle transport proteins were also detected, one each at the ER and Golgi localizations.

Importantly, the PTP detection approach in the present study depended on the annotation of orthologous proteins using a keyword-based approach. Noteworthy, mistaken, incomplete, or ambiguous annotations are frequent and, consequently, proteins that are not true transmembrane transporters were identified as PTPs. This was evident in the localization analysis, where 40 % of the 188 investigated proteins did not have TM helices. Transporters involved in the transmembrane transport of substrates are expected to have transmembrane domains, as they need to form a hydrophilic pathway through the membrane for solutes to travel (Muth and Caplan, 2003). However, not all proteins identified in this study were complete transmembrane transport proteins, hence their lack of transmembrane domains. For example, some of these were non-transmembrane members of large transport complexes, such as the ATPase family proteins, others were only tangentially involved in transport, such as the heat shock proteins, and there were even proteins with non-transmembrane transport functions, such as vesicle transport proteins. Of course, some false positives were also detected, such as ubiquitin-conjugating enzymes (with orthologues annotated as "ubiquitin carrier enzyme"), and curiously a histone protein (which was included as one of its OrthoMCL orthologues in *Babesia bovis* was annotated as "putative Trafficking protein particle complex subunit 5"). These false positives could have resulted from unintended matches to the selected keywords, incorrect annotations of reference proteins, as well as incorrect inclusion of correctly annotated proteins in orthologous groups.

3.5. High-confidence transmembrane transport proteins

The 188 PTPs were further filtered to identify proteins involved in transmembrane transport. Among these, only 108 had structural matches to transport-related proteins (Section 3.2), of which 15 had AFDB structures below the pLDDT threshold.

The GO term prediction using DeepGO-SE revealed that 179 of the 188 PTPs had association with transport biological processes (GO:0006810) (Fig. 6, detailed in Table C.3). Of these, 70 % (125 of 179) had the molecular function for transmembrane transporter activity (GO:0022857), suggesting direct involvement in the transport of substrates across membranes. Notably, 108 PTPs with the transmembrane transporter activity also had at least one TM helix. When TooT-BERT-T and structural similarity support was considered, only 45 proteins had structural support with average pLDDT of at least 70 (listed in Table G.1).

3.6. KEGG pathway and BRITE term prediction

From the 45 high-confidence PTPs, only eight could be assigned KEGG pathways (Fig. 7A, detailed in Table G.2). Cellular processes and human diseases were the most common pathway categories. In the cellular processes category, the largest subgroup was "Transport and catabolism", predominantly related to the phagosome (map04145, 4 members). Among the human diseases, the "Infectious disease: bacterial" was the largest, with 4 members. In contrast, 13 of the 45 proteins could be assigned BRITE terms (Fig. 7B, detailed in Table G.3). The most common category was "Protein families: signaling and cellular processes", and its most dominant subgroup was "Transporters" (ko02000), containing seven members. Overall, while very few of the PTPs could be associated with KEGG pathways and BRITE classifications, those that could were generally involved in transport, catabolism, and infectious disease.

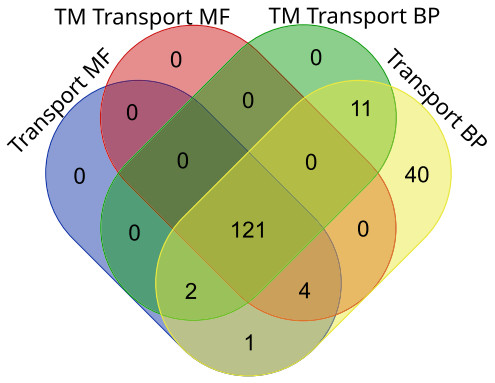


Fig. 6. Distribution of 179 of 188 potential transport-related proteins across different Gene Ontology terms. Transport MF, GO:0005215 "transporter activity" molecular function; TM Transport MF, GO:0022857 "transmembrane transporter activity" molecular function; TM Transport BP, GO:0055085 "transmembrane transport" biological process; Transport BP, GO:0006810 "transport" biological process.

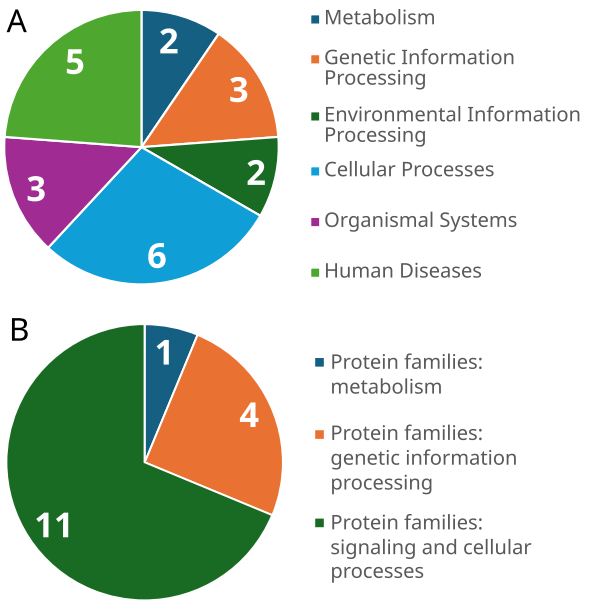


Fig. 7. Summary of the main categories for the KEGG pathway (A) and KEGG BRITE (B) terms that were assigned to eight and 13 of the 45 high-confidence potential transport-related proteins, respectively.

3.7. Transportome conservation within the genus Theileria and phylum Apicomplexa

All 188 PTPs detected in this study had BLASTP hits to at least one other *Theileria* species (*T. annulata, T. orientalis, T. equi*), with 178 having hits in all three (Fig. 8A, detailed in Table C.4). There were no PTPs that were exclusive to transforming species (*T. annulata* and *T. parva*). These observations suggest that these PTPs have roles that are not exclusive to transforming *Theileria* species. Furthermore, the broad distribution of these transporters within *Theileria* could be exploited in drug development, as the same or only slightly modified drugs could be used against multiple species.

The conservation analysis of the 188 PTPs in Apicomplexa revealed that seven proteins were exclusive to genus *Theileria* (Appendix H). Considering the 45 high-confidence PTPs, BLAST hits were detected in twelve genera, distributed across the orders Piroplasmida (*Theileria*, *Babesia*) and Haemosporidia (*Plasmodium*), and within the class Conoidasida (*Cryptosporidium*, *Cyclospora*, *Eimeria*, *Besnoitia*, *Cystoisospora*, *Neospora*, *Toxoplasma*, *Gregarina*, *Porospora*). Interestingly, only two of the 45 PTPs (EAN31530.1 and EAN31807.1, both hENT1-like proteins) had hits exclusively found within the *Theileria* genus. Majority of the remaining PTPs (39 of 45, 87 %) were found to be common between *Babesia*, *Plasmodium*, and Conoidasida (Fig. 8B, detailed in Tables C.5 and G.4), suggesting broad conservation within Apicomplexa. Curiously, one protein (EAN33647.1, MFS transporter on UniProt) had hits against only *Babesia* and the Conoidasida, suggesting conservation within Apicomplexa and possible deletion within *Plasmodium*. Finally, three proteins had hits against *Babesia* only, suggesting more specific functions within Piroplasmida. These were two group 2 hENT1-like proteins (EAN32006.1, EAN32008.1) and EAN32078.1 (amino acid transporter, UniProt).

4. Conclusion

We identified 334 proteins potentially involved in transport-related processes in the *T. parva* GenBank proteome, based on orthology to transport proteins within Aconoidasida and results from TransportDB. Of these, 188 were additionally supported by structural matches to transport proteins or by TooT-BERT-T, a transporter prediction tool. There were 24 uncharacterized proteins within this set, including potential MFS transporters, metal transporters, and nucleoside transporters. The majority of the 188 PTPs were predicted to be localized to the plasma membrane and the mitochondrion. Finally, it was determined that none of these PTPs were exclusive to transforming *Theileria* species, although seven were exclusive to the *Theileria* genus. A subset of 45 PTPs, most of which being conserved across phylum Apicomplexa, could play a more direct role in transmembrane transport. The remaining PTPs (146 of 334) will require further analysis to confirm their transport function. Knowledge of the *T. parva* transportome may assist in the identification of drug targets that can block the import of essential nutrients or the export of harmful metabolic waste. This information could also be helpful in elucidating routes for drug influx and efflux, which may be involved in resistance.

Ethical considerations

This study was approved by the Research Ethics Committee of the Faculty of Veterinary Science, University of Pretoria, with the reference number REC170–21.

Funding

NK would like to thank the University of Pretoria for a doctoral research bursary. The funder had no influence on any aspect of the

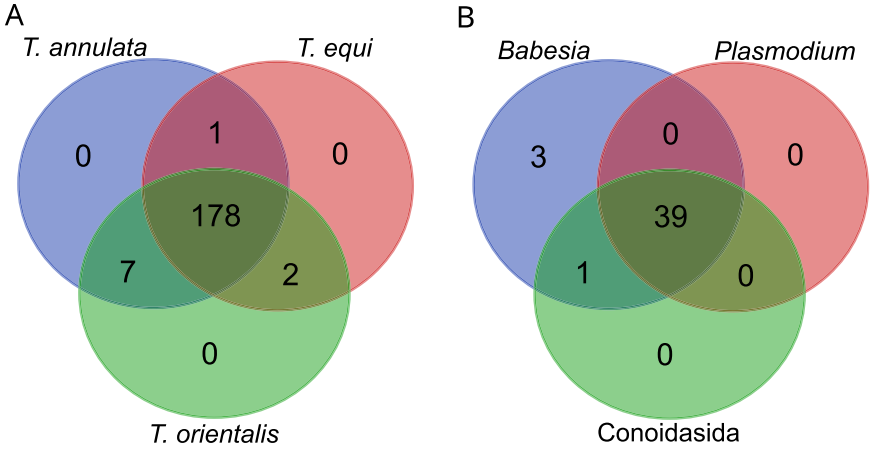


Fig. 8. Venn diagrams showing the presence of homologues of potential transport-related proteins (PTPs) within two different taxonomic groups. (A) shows the distribution of homologues of the 188 PTPs in other transforming (*T. annulata*) and non-transforming (*T. orientalis* and *T. equi*) species. (B) shows the distribution of homologues of the 45 high-confidence transmembrane PTPs across phylum Apicomplexa.

study.

CRediT authorship contribution statement

Leonhard Schnittger: Writing – review & editing, Supervision, Methodology. **Kgomotso Sibeko-Matjila:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Nikolaos Kotsovolos:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2025.108653.

References

Aguttu, C., Okech, B.A., Mukisa, A., Lubega, G.W., 2021. Screening and characterization of hypothetical proteins of *plasmodium falciparum* as novel vaccine candidates in the fight against malaria using reverse vaccinology. J. Genet. Eng. Biotechnol. 19, 103. https://doi.org/10.1186/s43141-021-00199-y.

Alballa, M., Butler, G., 2022. TooT-SC: Predicting Eleven Substrate Classes of Transmembrane Transport Proteins. https://doi.org/10.1101/2022.01.25.477715.

Bastos, R.G., Thekkiniath, J., Ben Mamoun, C., Fuller, L., Molestina, R.E., Florin-Christensen, M., Schnittger, L., Alzan, H.F., Suarez, C.E., 2021. *babesia microti* immunoreactive Rhoptry-Associated Protein-1 paralogs are ancestral members of the Piroplasmid-Confined RAP-1 family. Pathogens 10, 1384. https://doi.org/10.3390/pathogens10111384.

Berná, L., Marquez, P., Cabrera, A., Greif, G., Francia, M.E., Robello, C., 2021. Reevaluation of the *toxoplasma gondii* and *neospora caninum* genomes reveals misassembly, karyotype differences, and chromosomal rearrangements. Genome Res 31, 823–833. https://doi.org/10.1101/gr.262832.120.

Bishop, R.P., Kappmeyer, L.S., Onzere, C.K., Odongo, D.O., Githaka, N., Sears, K.P., Knowles, D.P., Fry, L.M., 2020. Equid infective theileria cluster in distinct 18S rRNA gene clades comprising multiple taxa with unusually broad mammalian host ranges. Parasit. Vectors 13, 261. https://doi.org/10.1186/s13071-020-04131-0.

Blouin, E.F., Stoltsz, W.H., 1989. Comparative infection rates of *theileria parva lawrencei* in salivary glands of *rhipicephalus appendiculatus* and *rhipicephalus zambeziensis*. Onderstepoort J. Vet. Res. 56, 211–213.

Bosshart, P.D., Kalbermatter, D., Bonetti, S., Fotiadis, D., 2019. Mechanistic basis of L-lactate transport in the SLC16 solute carrier family. Nat. Commun. 10. https://doi.org/10.1038/s41467-019-10566-6.

Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A.X., Yang, K.K., Min, S., Yoon, S., Morton, J.T., Rost, B., 2021. Learned embeddings from deep learning to visualize and predict protein sets. Curr. Protoc. 1, e113. https://doi.org/10.1002/cpz1.113.

Dillague, C., Akabas, M.H., 2023. Putative purine nucleoside interacting residues in the malaria parasite purine uptake transporter PfENT1 are critical for transporter function. PLOS ONE 18, e0293923. https://doi.org/10.1371/journal.pone.0293923.

Dobbelaere, D., Heussler, V., 1999. Transformation of leukocytes by *theileria parva* and *t. annulata*. Annu. Rev. Microbiol 53, 1–42. https://doi.org/10.1146/annurev.micro.53.1.1.

Elbourne, L.D.H., Tetu, S.G., Hassan, K.A., Paulsen, I.T., 2017. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. Nucleic Acids Res 45, D320–D324. https://doi.org/10.1093/nar/gkw1068.

Fawcett, D., Musoke, A., Voigt, W., 1984. Interaction of sporozoites of *theileria parva* with bovine lymphocytes *in vitro*. 1. early events after invasion. Tissue Cell 16, 873–884. https://doi.org/10.1016/0040-8166(84)90068-5.

Gabaldón, T., Koonin, E.V., 2013. Functional and evolutionary implications of gene orthology. Nat. Rev. Genet 14, 360–366. https://doi.org/10.1038/nrg3456.

Galperin, M.Y., Koonin, E.V., 2004. Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res 32, 5452–5463. https://doi.org/10.1093/nar/gkh885.

Gardner, M.J., Bishop, R., Shah, T., de Villiers, E.P., Carlton, J.M., Hall, N., Ren, Q., Paulsen, I.T., Pain, A., Berriman, M., Wilson, R.J.M., Sato, S., Ralph, S.A., Mann, D. J., Xiong, Z., Shallom, S.J., Weidman, J., Jiang, L., Lynn, J., Weaver, B., Shoaibi, A., Domingo, A.R., Wasawo, D., Crabtree, J., Wortman, J.R., Haas, B., Angiuoli, S.V., Creasy, T.H., Lu, C., Suh, B., Silva, J.C., Utterback, T.R., Feldblyum, T.V., Pertea, M., Allen, J., Nierman, W.C., Taracha, E.L.N., Salzberg, S.L., White, O.R., Fitzhugh, H.A., Morzaria, S., Venter, J.C., Fraser, C.M., Nene, V., 2005. Genome sequence of *theileria parva*, a bovine pathogen that transforms lymphocytes. Science 309, 134–137. https://doi.org/10.1126/science.1110439.

Ghazikhani, H., Butler, G., 2023. TooT-BERT-T: A BERT Approach on Discriminating Transport Proteins from Non-transport Proteins, in: Fdez-Riverola, F., Rocha, M., Mohamad, M.S., Caraiman, S., Gil-González, A.B. (Eds.), Practical Applications of Computational Biology and Bioinformatics, 16th International Conference (PACBB 2022), Lecture Notes in Networks and Systems. Springer International Publishing, Cham, pp. 1–11. https://doi.org/10.1007/978-3-031-17024-9_1.

Hallgren, J., Tsirigos, K.D., Pedersen, M.D., Almagro Armenteros, J.J., Marcatili, P., Nielsen, H., Krogh, A., Winther, O., 2022. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. BioRxiv 2022.04. 08.487609.

Hayashida, K., Hara, Y., Abe, T., Yamasaki, C., Toyoda, A., Kosuge, T., Suzuki, Y., Sato, Y., Kawashima, S., Katayama, T., Wakaguri, H., Inoue, N., Homma, K., Tada-Umezaki, M., Yagi, Y., Fujii, Y., Habara, T., Kanehisa, M., Watanabe, H., Ito, K., Gojobori, T., Sugawara, H., Imanishi, T., Weir, W., Gardner, M., Pain, A., Shiels, B., Hattori, M., Nene, V., Sugimoto, C., 2012. Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of Theileria-Induced leukocyte transformation, 10.1128/mbio.00204-12 mBio 3. https://doi.org/10.1128/mbio.00204-12.

Hernández-Plaza, A., Szklarczyk, D., Botas, J., Cantalapiedra, C.P., Giner-Lamia, J., Mende, D.R., Kirsch, R., Rattei, T., Letunic, I., Jensen, L.J., Bork, P., von Mering, C., Huerta-Cepas, J., 2023. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res 51, D389–D394. https://doi.org/10.1093/nar/gkac1022.

Holm, L., 2022. Dali server: structural unification of protein families. Nucleic Acids Res 50, W210–W215. https://doi.org/10.1093/nar/gkac387.

Holm, L., Kääriäinen, S., Wilton, C., Plewczynski, D., 2006. Using dali for structural comparison of proteins, 5.5.1-5.5.24 Curr. Protoc. Bioinforma. 14. https://doi.org/10.1002/0471250953.bi0505s14.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kanehisa, M., Sato, Y., Morishima, K., 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726–731. https://doi.org/10.1016/j.jmb.2015.11.006.

van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., Steinegger, M., 2024. Fast and accurate protein structure search with foldseek. Nat. Biotechnol. 42, 243–246. https://doi.org/10.1038/s41587-023-01773-0.

Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., Tang, H., 2018. GOATOOLS: a python library for gene ontology analyses. Sci. Rep. 8, 10872. https://doi.org/10.1038/s41598-018-28948-z.

Kulmanov, M., Guzmán-Vega, F.J., Roggli, P.D., Lane, L., Arold, S.T., Hoehndorf, R., 2023. DeepGO-SE: Protein function prediction as Approximate Semantic Entailment. https://doi.org/10.1101/2023.09.26.559473.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E.V., Zdobnov, E.M., 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res 51, D445–D451. https://doi.org/10.1093/nar/gkac998.

Lee, Y., Nishizawa, T., Takemoto, M., Kumazaki, K., Yamashita, K., Hirata, K., Minoda, A., Nagatoishi, S., Tsumoto, K., Ishitani, R., Nureki, O., 2017. Structure of the triose-phosphate/phosphate translocator reveals the basis of substrate specificity. Nat. Plants 3, 825–832. https://doi.org/10.1038/s41477-017-0022-8.

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13, 2178–2189. https://doi.org/10.1101/gr.1224503.

Lubec, G., Afjehi-Sadat, L., Yang, J.-W., John, J.P.P., 2005. Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog. Neurobiol. 77, 90–127. https://doi.org/10.1016/j.pneurobio.2005.10.001.

Mäser, P., Sütterlin, C., Kralli, A., Kaminsky, R., 1999. A nucleoside transporter from *trypanosoma brucei* involved in drug resistance. Science 285, 242–244. https://doi.org/10.1126/science.285.5425.242.

Meier, A., Erler, H., Beitz, E., 2018. Targeting channels and transporters in protozoan parasite infections. Front. Chem. 6. https://doi.org/10.3389/fchem.2018.00088.

Mhadhbi, M., Naouach, A., Boumiza, A., Chaabani, M.F., BenAbderazzak, S., Darghouth, M.A., 2010. In vivo evidence for the resistance of theileria annulata to buparvaquone. Vet. Parasitol. 169, 241–247. https://doi.org/10.1016/j.vetpar.2010.01.013.

Mirza, M., Vainshtein, A., DiRonza, A., Chandrachud, U., Haslett, L.J., Palmieri, M., Storch, S., Groh, J., Dobzinski, N., Napolitano, G., Schmidtke, C., Kerkovich, D.M., 2019. The CLN3 gene and protein: what we know. Mol. Genet. Genom. Med. 7, e859. https://doi.org/10.1002/mgg3.859.

Muth, T.R., Caplan, M.J., 2003. Transport protein trafficking in polarized cells. Annu. Rev. Cell Dev. Biol. 19, 333–366. https://doi.org/10.1146/annurev.cellbio.19.110701.161425.

Nehrt, N.L., Clark, W.T., Radivojac, P., Hahn, M.W., 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLOS Comput. Biol. 7, e1002073. https://doi.org/10.1371/journal.pcbi.1002073.

Nene, V., Kiara, H., Lacasta, A., Pelle, R., Svitek, N., Steinaa, L., 2016. The biology of *theileria parva* and control of east coast fever – current status and future trends. TTP8-STVM Special Issue Ticks Tick. Borne Dis. 7, 549–564. https://doi.org/10.1016/j.ttbdis.2016.02.001.

Nightingale, A., Antunes, R., Alpi, E., Bursteinas, B., Gonzales, L., Liu, W., Luo, J., Qi, G., Turner, E., Martin, M., 2017. The proteins API: accessing key integrated protein and genome information. Nucleic Acids Res 45, W539–W544. https://doi.org/10.1093/nar/gkx237.

Parker, J.L., Newstead, S., 2017. Structural basis of nucleotide sugar transport across the Golgi membrane. Nature 551, 521–524. https://doi.org/10.1038/nature24464.

Parker, J.L., Corey, R.A., Stansfeld, P.J., Newstead, S., 2019. Structural basis for substrate specificity and regulation of nucleotide sugar transporters in the lipid bilayer. Nat. Commun. 10, 4657. https://doi.org/10.1038/s41467-019-12673-w.

Roberts, R.J., 2004. Identifying protein Function—A call for community action. PLoS Biol. 2, e42. https://doi.org/10.1371/journal.pbio.0020042.

van Schalkwyk, D.A., Nash, M.N., Shafik, S.H., Summers, R.L., Lehane, A.M., Smith, P.J., Martin, R.E., 2016. Verapamil-Sensitive transport of quinacrine and methylene blue via the *plasmodium falciparum* chloroquine resistance transporter reduces the Parasite's susceptibility to these tricyclic drugs. J. Infect. Dis. 213, 800–810. https://doi.org/10.1093/infdis/jiv509.

Schnittger, L., Rodriguez, A.E., Florin-Christensen, M., Morrison, D.A., 2012. Babesia: a world emerging. Infect. Genet. Evol. 12, 1788–1809. https://doi.org/10.1016/j.meegid.2012.07.004.

Schnittger, L., Ganzinelli, S., Bhoora, R., Omondi, D., Nijhof, A.M., Florin-Christensen, M., 2022. The piroplasmida babesia, cytauxzoon, and theileria in farm and companion animals: species compilation, molecular phylogeny, and evolutionary insights. Parasitol. Res. 121, 1207–1245. https://doi.org/10.1007/s00436-022-07424-8.

Staines, H.M., Moore, C.M., Slavic, K., Krishna, S., 2017. Transmembrane solute transport in the apicomplexan parasite plasmodium. Emerg. Top. Life Sci. 1, 553–561. https://doi.org/10.1042/ETLS20170097.

Stamboulian, M., Guerrero, R.F., Hahn, M.W., Radivojac, P., 2020. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. Bioinformatics 36, i219–i226. https://doi.org/10.1093/bioinformatics/btaa468.

Stärk, H., Dallago, C., Heinzinger, M., Rost, B., 2021. Light attention predicts protein location from the language of life. Bioinforma. Adv. 1, vbab035. https://doi.org/10.1093/bioadv/vbab035.

Thumuluri, V., Almagro Armenteros, J.J., Johansen, A.R., Nielsen, H., Winther, O., 2022. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res 50, W228–W234. https://doi.org/10.1093/nar/gkac278.

Tretina, K., Pelle, R., Orvis, J., Gotia, H.T., Ifeonu, O.O., Kumari, P., Palmateer, N.C., Iqbal, S.B.A., Fry, L.M., Nene, V.M., Daubenberger, C.A., Bishop, R.P., Silva, J.C., 2020. Re-annotation of the *theileria parva* genome refines 53% of the proteome and uncovers essential components of N-glycosylation, a conserved pathway in many organisms. BMC Genom. 21, 279. https://doi.org/10.1186/s12864-020-6683-0.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444. https://doi.org/10.1093/nar/gkab1061.

Wang, C., Yu, L., Zhang, J., Zhou, Y., Sun, B., Xiao, Q., Zhang, M., Liu, H., Li, Jinhong, Li, Jialu, Luo, Y., Xu, J., Lian, Z., Lin, J., Wang, X., Zhang, P., Guo, L., Ren, R., Deng, D., 2023. Structural basis of the substrate recognition and inhibition mechanism of *plasmodium falciparum* nucleoside transporter PfENT1. Nat. Commun. 14, 1727. https://doi.org/10.1038/s41467-023-37411-1.

Wasmuth, J., Daub, J., Peregrín-Alvarez, J.M., Finney, C.A.M., Parkinson, J., 2009. The origins of apicomplexan sequence innovation. Genome Res 19, 1202–1213. https://doi.org/10.1101/gr.083386.108.

Wichers, J.S., van Gelder, C., Fuchs, G., Ruge, J.M., Pietsch, E., Ferreira, J.L., Safavi, S., von Thien, H., Burda, P.-C., Mesén-Ramirez, P., Spielmann, T., Strauss, J., Gilberger, T.-W., Bachmann, A., 2021. Characterization of apicomplexan amino acid transporters (ApiATs) in the malaria parasite *plasmodium falciparum*. mSphere 6, e00743-21. https://doi.org/10.1128/mSphere.00743-21.

Wing, C.E., Fung, H.Y.J., Chook, Y.M., 2022. Karyopherin-mediated nucleocytoplasmic transport. Nat. Rev. Mol. Cell Biol. 23, 307–328. https://doi.org/10.1038/s41580-021-00446-7.

Wright, N.J., Lee, S.-Y., 2019. Structures of human ENT1 in complex with adenosine reuptake inhibitors. Nat. Struct. Mol. Biol. 26, 599–606. https://doi.org/10.1038/s41594-019-0245-7.

Wunderlich, J., 2022. Updated list of transport proteins in *plasmodium falciparum*. Front. Cell. Infect. Microbiol 12, 926541. https://doi.org/10.3389/fcimb.2022.926541.

Yu, C.-S., Chen, Y.-C., Lu, C.-H., Hwang, J.-K., 2006. Prediction of protein subcellular localization. Proteins Struct. Funct. Bioinforma. 64, 643–651. https://doi.org/10.1002/prot.21018.

Yucesan, B., Guldemir, D., Babur, C., Kilic, S., Cakmak, A., 2021. Whole-genome sequencing of a *toxoplasma gondii* strain from a turkish isolate using next-generation sequencing technology. Acta Trop. 218, 105907. https://doi.org/10.1016/j.actatropica.2021.105907.

Zhao, Z., Xian, M., Liu, M., Zhao, G., 2020. Biochemical routes for uptake and conversion of xylose by microorganisms. Biotechnol. Biofuels 13, 21. https://doi.org/10.1186/s13068-020-1662-x.