Functional annotation hypothetical proteins: a world to be explored in drug development in Trypanosomatids

Raissa Santos de Lima[1], Ana Carolina Silva Bulla[1], Manuela Leal da Silva[1,2*]

[1]Programa de Pós-Graduação em Biologia Computacional e Sistemas, Instituto Oswaldo Cruz, Manguinhos, Rio de Janeiro, RJ, Brazil.

[2]Instituto de Biodiversidade e Sustentabilidade NUPEM, Universidade Federal do Rio de Janeiro, Macaé, RJ, Brazil.

*Corresponding author:

E-mail: manuela@macae.ufrj.br

Author contributions: Conceived and designed the experiments: RSLR MLS. Performed the experiments: RSLR ACSB. Analyzed the data: RSLR ACSB. Wrote the paper: RSLR ACSB MLS.

Abstract (300 words)

Hypothetical proteins can provide an alternative pathway for finding potential targets in the development of new drugs due to the fact that many Neglected tropical diseases are caused by Trypanosomatids (Chagas Disease, Leishmaniasis, and Human African Trypanosomiasis). In this work, we focus on applying functional prediction methods based on both sequence and structure to analyze the hypothetical proteins of the pathogenic agents that cause these diseases: *T. cruzi* (Tcr), *T. brucei brucei* (Tbr), *T. brucei gambiense* (Tbg), *L. infantum* (Lif), *L. donovani* (Ldo), and *L. braziliensis* (Lbz). By consulting databases and servers, we have predicted functional domains for twenty-six proteins in Tcr, thirteen in Tbr, fifteen in Tbg, ten in Lif, and one in both Ldo and Lbz. With the goal of developing multi-target therapies, we grouped the domains according to how they are shared among the organisms and investigated those that are shared among more species. By examining the existing literature using specific search strategies, we described what has already been reported for these domains and also analyzed protein structures and sequences, describing mutations among the species and potential drug sites. The published works have unveiled that some of these domains are non-essential for trypanosomatids, like the TRX domain, while others demand further investigation due to a lack of information about metabolic processes (UFC1, Ufm1, ACBP, AAA 18, and Fe-S). Although, we have identified three noteworthy domains that hold promise as targets: TPR, which plays a crucial role in the ciliogenesis process; Nuc deoxyrib tr, essential in purine recycling and recovery mechanisms; and MIX, important for protein targeting and the assembly of complexes such as COX. These three domains are

promising targets for drug development due to their conservation, their potential to affect multiple species and their exclusivity.

Introduction

Trypanosoma parasite species can infect several animals causing neglected diseases such as Chagas disease (CD) (*Trypanosoma cruzi*), Human African Trypanosomiasis (HAT) (*Trypanosoma brucei*) and Leishmaniasis (genus Leishmania) [1,2]. From an epidemiological standpoint, these diseases' incidence rates are alarmingly high, with millions of cases reported annually [3–5]. Furthermore, HAT was close to eradication in the 60s but the disease has since experienced a resurgence in reason of political problems, and a decline in control programs [5,6]. Besides that, many of these diseases have a wide geographical distribution due to infected individuals immigrating to non-endemic countries, and leishmaniasis in particular has three endemic foci: India, East Africa, and Brazil [3,7,8].

Aside from the challenge of dissemination, the treatment has several problems, mainly regarding efficacy and therapeutic adherence due to the high toxicity of the drugs used [9–11]. Furthermore, there are already cases of resistance that demonstrate the promiscuity not only of drugs but also of current targets for diseases caused by trypanosomatids [12,13]. As a result, there is a pressing need to take the initiative to develop new drugs, and one promising approach is the drug repositioning strategy. This process is highly cost-effective since it reduces drug development time, optimizing pre-clinical trials by utilizing drugs with known possible adverse effects [14,15]. Thus, efforts

66 are required to identify new molecular targets for new drugs and their implementation as
67 effective treatments.

68 Drugs can have different molecular targets, one of the most common being
69 proteins, and with the increasing popularity of sequencing techniques, there has been a
70 rise in the use of bioinformatics tools, including those for protein analysis [16]. Despite
71 the availability of predicted proteins in genomes, there is a limited amount of functional
72 details data for these molecules. The reason for that is the sheer volume of data that
73 experimental functional characterization produces, which is difficult to manage. So the
74 functional prediction approach comes in as a solution, and it has been constantly
75 improved [17].

76 There are two main approaches to protein function prediction and the first one is
77 the SEquence-Based Functional Prediction (SEBFP), which assumes that the sequence
78 contains a conserved "functional signature" that can be identified even in a small part of
79 the sequence [18]. Consequently, if two proteins have this conserved "signature" it is
80 possible to infer the protein function [19]. The second approach is Structure-Based
81 Functional Prediction (STBFP) which has two premises, the first one is that all the
82 essential information about a protein's structure is contained in its amino acid sequence
83 [20]. The second premise is that proteins have a specific native folding state or a set of
84 native states, and this structure is closely related to the catalytic protein function [21].

85 The primary aim of this study is to employ a hybrid approach, merging SEBFP
86 and STBFP strategies, to predict the functional domains of hypothetical proteins (HP) as
87 depicted in Fig 1. Our goal is to achieve more robust results compared to conventional

linear methodologies and to identify new targets for drug development that are applicable across multiple neglected diseases caused by these parasites.

Fig 1. Scheme of the functional prediction methodology. Through the utilization of a diverse range of computational tools and databases, we conducted an extensive analysis of both sequence and structural data to identify functional domains within the hypothetical proteins (HP). As illustrated in the sequence and structure, an exemplary functional domain is highlighted in orange, showcasing its conservation across various protein data levels. For the 3D protein representation, the PDBids were used: 3UM8, 4TU5, 6WEP, and 3Q1H.

To demonstrate our approach, we applied SEBFP and STBFP using various databases, identifying a total of sixty-six protein domains for the organisms *T. cruzi* (Tcr), *T. brucei brucei* (Tbr), *T. brucei gambiense* (Tbg), *L. infantum* (Lif), *L. donovani* (Ldo), and *L. braziliensis* (Lbz). Notably, nine of these domains were shared by four species (Tcr, Tbr, Tbg, and Lif), which we have provided detailed descriptions for. With a focus on identifying promising multi-targets, we conducted sequence and structural analyses in our pursuit of potential druggable sites for the three domains (TPR, Nuc deoxyrib tr, and MIX) that hold promising functions.

Materials and methods

Proteomes selection

According to the disease, we selected the proteomes of the following species *Trypanosoma cruzi* for CD, *Trypanosoma brucei* for HAT and *Leishmania* spp for

109 Leishmaniasis. To determine the strain among several deposited in the National Center
110 for Biotechnology Information (NCBI) databases [22], we first selected the most widely
111 reported strain based on literature data. Additionally, we considered the strains with the
112 highest assembly level available in the NCBI. After the selection process, using a script
113 that analyzes the identifiers, we separate the proteins annotated as HP from the others.

114 Large-scale comparative modeling

115 The multiFASTA file containing only the HP sequences was submitted to the
116 MHOLline web server for structural modeling [23]. MHOLline is a biological workflow
117 capable of performing three-dimensional modeling for different proteins at the same time.
118 The structure prediction process starts with searching for templates by aligning protein
119 sequences with structures in the Protein Data Bank (PDB) using the Basic Local
120 Alignment Search Tool (BLAST) [24]. Monomeric 3D models are generated with
121 MODELLER [25] for only results with e-value ≤ 10e-5, identity ≥ 0.25, and Length
122 Variation Index (LVI) ≤ 0. Subsequently, MHOLline classifies these models into five
123 groups among Very High and Very Low using the module called Filters, which is based
124 on the identity and LVI values. In this study, we selected those that were classified
125 between Very High and Good (Very High: ID ≥ 75% and LVI ≤ 0.1%; High: ID ≥ 50%
126 and <75% and LVI ≤ 0.1%; Good: ID ≥ 35% and < 50% and LVI ≤ 0.3%). Hence, only
127 the protein sequences that met the evaluation criteria were deemed suitable for the
128 functional prediction stage.

Functional prediction and description

As previously mentioned, we employed a combined SEBFP and STBFP approach to predict the selected HP sequence's functional domains. In total, nine databases and six web servers were utilized, as shown in Fig 2. To ensure the reliability of the functional prediction, we defined that the same domain must be predicted in at least eight of the SEBFP tools and two of the STBFP for functional inference. Similar strategies have been previously adopted by Shahbaaz, M., Hassan, M. I., & Ahmad, F. (2013) and Mazandu, G. K., & Mulder (2008).

Fig 2. Workflow containing databases and servers used in SEBFP and STBFP. Databases are represented as cylinders and the web servers as rectangles.

Consulting the literature, the functional domains were detailed, particularly through the use of bibliographic databases such as Web of Science, Google Scholar, The Lens, and PubMed. The search strategy employed was using keywords such as the functional domain acronym, disease's name, and the parasite species, for example, "TRX AND Chagas Disease AND Trypanosoma cruzi". Additionally, to find targets that are crucial for the parasite's metabolic pathways, we consulted the Kyoto Encyclopedia of Genes and Genom3es (KEGG) database [26] using the EC number or the functional domains acronym in the.

Sequence analysis

Sequence analysis began using tools that contain details about domains and families, which included: Pfam [27], Simple Modular Architecture Research Tool (SMART) [28], PROSITE [29], Conserved Domains Database (CDD) [30], Search Tool

151 for Retrieval of Interacting Genes/Proteins (STRING) [31] and the CATH [32] and
152 HMMER [33] web servers. These tools were used with default settings and allow the
153 identification of homologous sequences by aligning them with information in various
154 databases. This type of approach, the homologs identification, or evolutionary
155 relationships, is widely used for the structural and functional annotation of proteins [32–
156 34].

157 The proper characterization of a protein involves determining not only its function
158 but also its cellular localization. This information is crucial for understanding its role in
159 cellular processes and for the development of targeted therapies. To determine the
160 subcellular localization we used the Psort (Prediction of Protein Sorting Signals and
161 Localization Sites in Amino Acid Sequences) web server [35], more specifically WoLF
162 PSORT with default parameters. This tool predicts protein sorting signals and localization
163 sites based on amino acid sequences. Lastly, we consulted UniProtKB, where is possible
164 to search for similar sequences with predicted functions using the integrated BLAST tool
165 and we consulted the curated Swiss-Prot sequence database with default parameters
166 [36,37].

167 To further explore sequence similarity with other sequences with other already
168 known sequences deposited in data banks, we used the Basic Local Alignment Search
169 Tool (BLAST) tool with the Non-Redundant (NR) [24] database and the TriTrypDB
170 database [38], both with the default search parameters. The TriTrypDB database contains
171 sequence data for the Trypanosomatidae family species and sources its information from
172 banks such as the DNA Data Bank of Japan (DDBJ) and the European Molecular Biology
173 Laboratory (EMBL) [39].

Among species, the domain conservation, as well as mutations were analyzed using the UGENE alignment tool, specifically with the MUtiple Sequence Comparison by Log-Expectation (MUSCLE) algorithm [40]. MUSCLE realigns the sequences repeatedly, considering a guiding phylogenetic tree to generate the distance matrix [40–42]. Given that we were analyzing organisms from the same family, considering the phylogenetic relationships when assembling the matrix was crucial for the alignment quality [43].

Finally, in addition to the aforementioned analyses, we identified the signal peptide and transmembrane regions. First, for signal peptide detection, we used the SignalP web server, which can predict not only the signal peptide's presence but also their cleavage sites [44]. For the transmembrane regions encounter, the DeepTMHMM web server was used, which scores the protein region probability being inside, in the transmembrane region, or outside the membrane [45]. We used both servers with the default search parameters.

Structure analysis

STBFP analyses started by utilizing the structures generated by MHOLline to search the SCOP server. This server classifies the protein based on its structure and provides a structural relationship description among known proteins [46]. Besides that, we used the VAST server, which compares the spatial coordinates of an input 3D structure with other structures deposited in the Molecular Modeling Database (MMDB) of NCBI [47]. Furthermore, VAST can identify distant homologs that may not be recognized through a simple sequential comparison [48].

196 Another server utilized was the Flexible structure AlignmentT by Chaining
197 Aligned fragment pairs allowing Twists (FATCAT), which searches for similar structures
198 in databases such as PDB and SCOPe through a structure alignment algorithm [49].
199 Lastly, the Distance matrix ALIgnment server (Dali) was used also for comparison
200 structures in the PDB database but it utilizes three different ways: the heuristic search (all
201 structures in PDB), the exhaustive search (a PDB subset), and an AF-DB hierarchical
202 lookup (AlphaFold subset) [50]. Both FATCAT, Dali, and VAST servers perform
203 alignment based on calculating of Root Mean Square deviation (RMSD) using α carbons.

204 Regarding the RMSD, the Pymol (version 1.9) program [51] was utilized to
205 calculate it for the models generated by MHOLline with the templates used. In addition,
206 we examined the conservation of amino acid residues that compose the functional
207 domain, as well as the presence of mutations among the species in the structures. To
208 evaluate regions without clearly defined secondary structures, we referred to the
209 AlphaFold structure database [52], which provides information on models generated by
210 artificial intelligence [53], including details on species. Moreover, we utilized the
211 Quick2D web server, a tool that includes nine predictors of secondary structures, even
212 regions of disorder, to facilitate reaching a consensus on the structure in the region of
213 interest [54].

214 Modeling homo-oligomeric structures and validation and 215 binding site identification

216 The final structures were remodeled considering the oligomeric state through
217 AlphaFold2 on ColabFold [55] pipeline (num_relax: 5, template mode: none, msa_mode:

MMSeq2 (uniRef + environmental), num_recycle: 3), relaxing the structure with an amber force field, considering the sequences in table S4. To verify the structural quality of the built models, the models were submitted to MolProbity [56] and ProSA – web [57]. To access the protein druggability, the structures were submitted to DogSiteScorer [58] server for prediction of potential pockets on the proteins (score between 0 and 1).

Results and Discussion

Selection and large-scale 3D modeling

Using the NCBI database, we conducted a search of organism's proteomes, specifically: *Trypanosoma cruzi* (Tcr) for CD; *Trypanosoma brucei brucei* (Tbr), and *Trypanosoma brucei gambiensi* (Tbg) for HAT; *Leishmania infantum* (Lif), *Leishmania donovani* (Ldo) *Leishmania braziliensis* (Lbz) for Leishmaniasis. Lif and Ldo species were selected due to their association with the most severe clinical form of the disease, visceral leishmaniasis (VL), which is more prevalent in Brazil [3,7].

The organisms and their respective strains were selected considering literature data and the proteomes with the highest assembly level deposited in the NCBI database (Table 1). Once the strains were chosen, the hypothetical proteins (HP) were separated from the others and submitted to the MHOLline server. This resulted in the generation of 3D models for 2928 proteins, ouch of which 111 models qualified for the functional prediction stage due to their classification between Very High and Good (Very High: ID ≥ 75% and LVI ≤ 0.1%; High: ID ≥ 50% and <75% and LVI ≤ 0.1%; Good: ID ≥ 35% and <50% and LVI ≤ 0.3%). Lastly, we consulted sequential and structural databases to infer functional domains for these proteins.

240 **Table 1. Correlation of species, numbers of proteins, and 3D models**

Species	Strains	NCBI Code	TP	HP	Gen 3D	Models selected
T. cruzi	CLBrener	GCA_000 209065.1	19.608	11.171	1.421	40
T. brucei brucei	TREU927	GCA_000 002445.1	8.758	5.711	658	20
T. brucei gambiensi	DAL972	GCA_000 210295.1	9.822	6.611	661	23
L. infantum	Jpcm5	GCA_000 002875.2	8.15	5.245	648	21
L. donovani	LdCl	GCA_003 719575.1	8.632	3.527	90	3
L. braziliensis	MHOM	GCA_900 537975.1	8.244	3.373	98	4

241 [a]TP - Total of proteins

242 [b]HP - Hypothetical protein

243 [c]Gen 3D - 3D models generated

244

245 # Annotation of HPs

246 The nomenclature used to describe the results was based on the user in the
247 following databases: InterPro [59], Pfam [27], SMART [28], PROSITE [29], and CDD
248 [60] (S1 Table). To predict the functional domains, we established the following criterion:
249 the same function must be predicted in at least eight SEBFP tools and two STBFP tools.
250 Using sequence and structure analysis tools, we inferred 26 functional domains for Tcr,
251 13 for Tbr, 16 for Tbg, 10 for, Lif, 1 for Ldo, and 1 for Lbz, totaling 67 proteins from the
252 initial 111 (S2 Table).

253 By analyzing SEBFP and STBFP results together, it was observed that some of
254 the annotated domains were repeated within the same species and also interspecies (Fig

255 3). Specifically, we found that nine predicted domains were shared among four species
256 of trypanosomatids, namely Tcr, Tbr, Tbg, and Lif. Therefore, we narrowed our focus to
257 these shared domains, as our goal is to identify multi-targets for trypanosomatids.

258 **Fig 3. Venn diagram illustrating the shared functional domains among**
259 **trypanosomatid species.** Each circle in the diagram represents a distinct trypanosomatid
260 species, and the predicted functional domains are denoted by bold abbreviations. Our
261 analysis primarily concentrates on the centrally located shared domains, which have the
262 highest number of species located in the middle of the Venn plot.

263 Multi-target functional domains identified

264 Ubiquitination: UFC1 and Ufm1

265 In our research, we employed a combination of SEBFP and STBFP methodologies
266 to uncover various functions. However, as we delved into the specifics of each function,
267 it became evident that there were significant gaps in our understanding. For instance, we
268 identified the UFC1 and Ufm1 domains, both of which play roles in the ubiquitination
269 process—a post-translational protein modification system utilizing multiple ubiquitin
270 molecules to regulate cellular functions [61].

271 The UFC1 domain is a member of the E2-type proteins (conjugators) in the
272 ubiquitination cascade. It plays a vital role in receiving a specific type of ubiquitin (Ub),
273 known as Ubiquitin fold modifier 1 (Ufm1), from Uba5 (E1-type activators) through a
274 thiolester bond mediated by a cysteine [62]. This process is crucial for central nervous
275 system (CNS) development in humans, and bi-allelic mutations in UFC1 have been linked
276 to compromised neurological phenotypes, leading to severe encephalopathy and

277 progressive microencephaly cases [63]. Mutations like Arg23Gln can affect binding to
278 E1 or E3 proteins (ligands), while Thr106Ile impairs the UFC1-Ufm1 complex formation
279 [63].

280 In trypanosomatids, specifically *T. brucei*, research has shown that the
281 glycoproteins on their surface rely on the ubiquitination process, particularly Invariant
282 Surface Glycoproteins (ISG). After undergoing ubiquitination, these proteins are
283 internalized and degraded, promoting cell surface renewal. This implies that
284 ubiquitination serves as a general mechanism for regulating surface proteins with
285 transmembrane domains [64]. However, research also reveals variations in the machinery
286 involved in this process, such as the absence of orthologs for E3 proteins in the *T. brucei*
287 genome, suggesting alternative mechanisms or factors may be in play in these eukaryotes
288 [65].

289 Regarding the Ufm1 domain, a study conducted by Yoo et al. in 2014 found that
290 certain ubiquitin ligases are overexpressed in tumors, promoting cell proliferation. This
291 study also identified interactions between estrogen receptors (ERα) and Activating signal
292 cointegrator 1 (ASC1) molecules through the Ufm1 domain, leading to the activation of
293 ERα target genes associated with excessive cell proliferation and breast cancer [66]. The
294 significance of ubiquitylation in the human body underscores the need for further research
295 into UFC1 and Ufm1, especially in the development of targeted inhibitors for specific
296 members of the E3 family [67]. Meanwhile, the proteasome pathway, which is essential
297 for the degradation of cytoskeletal proteins in *T. cruzi*, shows promise as a drug
298 development target. Studies suggest that inhibiting the proteasome can disrupt the
299 evolutionary cycle of *T. cruzi*, making it vital to explore this pathway for new drug targets

300 [68–70]. Hence, it is crucial to persist in the investigation of this pathway to unearth novel
301 drug targets.

302 Electron transfer: Fe-S cluster

303 The Fe-S domain assumes a pivotal role in modulating transcriptional and
304 translational processes within genes [71]. Originating during the prebiotic era on Earth,
305 this domain finds its basis in the abundance of iron and sulfur, coupled with their
306 structural versatility and chemical reactivity, facilitating the formation of Fe-S clusters
307 (iron and sulfur). These clusters function as "molecular switches," capable of influencing
308 gene regulation at multiple levels [72]. Their intrinsic chemical properties make them
309 active participants in various biological processes, encompassing photosynthesis, DNA
310 replication and repair, respiration, nitrogen fixation, and more. Consequently, proteins
311 containing Fe-S clusters are ubiquitous across all organisms [73,74].

312 In humans, the cytosolic Fe-S cluster assembly (Cia) targeting complex is
313 constituted by Cia1, MMS19, Cia2A, and Cia2B (also known as Fam96b) [75]. Notably,
314 Cia2A and Cia2B hold significance in iron homeostasis and participate in the maturation
315 of the iron regulatory protein (IRP1) [76]. Fam96a, also recognized as Cia2A in
316 mammals, has demonstrated its role in reducing gastrointestinal stromal tumors, where
317 its interaction with apoptotic peptidase activation factor 1 heightened tumor sensitivity to
318 apoptosis processes [77,78]. Additionally, Fam96a depletion has been linked to organ
319 damage protection and increased survival rates for sepsis patients by reducing the
320 secretion of pro-inflammatory cytokines by macrophages [79].

321 In trypanosomatids, like *T. brucei*, the presence of the Cia complex has been
322 acknowledged; however, the precise components for Cia in *T. brucei* remain uncertain,

323 and it's been suggested that these components may have influenced the divergence of the
324 Kinetoplastea class from other eukaryotes [80]. The ISC machinery, responsible for the
325 maturation of cytosolic and nuclear Fe-S proteins, has been proposed to directly impact
326 the formation of flagella in these parasites, particularly under conditions affecting
327 glycolysis and mitochondrial function [81,82]. *Leishmania* spp. also possess Cia
328 components with variations in their mechanisms. Given the conservation and essential
329 role of the Cia pathway in Fe-S cluster assembly within trypanosomatids, it presents a
330 promising source of potential pharmaceutical targets, with Fam96a being one of the
331 identified candidates in *T. brucei* [80] and detected via STBFP analysis (S2 Table).

332 Lipid metabolism: ACBP

333 ACBPs, in general, function by binding to acyl-CoAs, thereby preventing their
334 hydrolysis by hydrolases [83]. This action facilitates the efficient transport of acyl-CoAs
335 to cellular acylation machinery. In humans, ACBPs, also known as ACBDs, have been
336 linked to various cellular processes, including lipid metabolism and cell signaling [84].

337 In trypanosomatids, ACBPs are believed to protect against inhibition by acyl-CoA
338 esters of acetyl-CoA carboxylase and long-chain mitochondrial adenine nucleotide
339 translocase [85]. These proteins also play a role in the transport and donation of acyl-
340 CoAs for β-oxidation in mitochondria and the synthesis of glycerolipids in microsomes
341 [86].

342 Experimental studies on trypanosomatids, particularly *T. brucei*, have revealed
343 the significant role of ACBP in the biosynthesis of glycosylphosphatidylinositol (GPI).
344 Specifically, ACBP donates myristoyl-CoA for fatty acid remodeling reactions, causing
345 a conformational change in fatty acids [87]. This alteration allows them to exit the

346 endoplasmic reticulum and anchor to Variant Surface Glycoproteins (VSG), which is
347 crucial for the formation of the VSGs [88]. Once anchored, VSGs undergo further
348 processing in the Golgi complex, including the addition of galactoses, before being
349 transported to the plasma membrane [85].

350 Furthermore, trypanosomatids rely on GPI anchors for various essential functions,
351 such as evading the host complement system in *T. brucei*, facilitating host cell interaction
352 and sialic acid uptake in *T. cruzi*, and protecting against extracellular hydrolases while
353 interacting with the insect vector's midgut in *Leishmani*a spp. These membrane
354 components are pivotal for various critical processes, highlighting the importance of
355 understanding the functionality of domains like ACBP in these parasites [89].

356 Energy maintenance: AAA18

357 AAA 18 domain, also known as adenylate kinase (AK), plays a role in regulating
358 cellular energy management by maintaining adenine nucleotide homeostasis. AAA 18
359 catalyzes the reversible conversion of ATP and AMP molecules into two ADP molecules,
360 contributing to the balance of adenine nucleotides and cellular energy charge, thereby
361 supporting optimal cellular functions and metabolism [84].

362 Studies have revealed the presence of multiple isoforms of AAA 18 with diverse
363 subcellular localizations and functions. In *T. cruzi*, a cytosolic isoform of AAA 18 has
364 been identified, and it plays a crucial role in the energy metabolism of the parasite,
365 particularly during the growth of the epimastigote form. This suggests a potential
366 correlation between the activity of AAA 18 and the energy demands of the parasite.
367 Interestingly, the expression of the cytosolic AAA 18 isoform decreases when another
368 phosphotransferase, arginine kinase, is overexpressed, indicating a potential functional

369 connection between these two enzymes and a compensatory mechanism for maintaining
370 the parasite's energy homeostasis [90].

371 Furthermore, the number of AAA 18 isoforms varies among different species. In
372 the case of *T. brucei*, Ginger et al. (2005) conducted a study that revealed seven isoforms
373 of AAA 18 in this organism. The existence of multiple isoforms suggests a form of
374 metabolic adaptation, leading to increased complexity in energy homeostasis reactions.
375 These isoforms are expressed in both procyclic trypanosomes found in the midgut of the
376 tsetse fly and metacyclic trypanosomes located in the bloodstream, making them
377 particularly relevant in the context of parasite infections [90]. Additionally, in mammals,
378 several AAA 18 isoforms have been associated with cancer, such as the overexpression
379 of AK2 in lung cancer patients, which is linked to the reprogramming of energy
380 metabolism—an essential aspect of cancer progression [91].

381 Despite our understanding of the functions of ACBP and AAA 18 domains, many
382 mechanisms related to these domains remain enigmatic. One intriguing area of research
383 is the relationship between ACBPs and peroxisomes, given the critical role of
384 peroxisomes in various neurological disorders like Alzheimer's, autism, and amyotrophic
385 lateral sclerosis [92]. With these proteins playing a significant role in humans, concerns
386 arise when targeting them with drugs due to potential systemic consequences. Therefore,
387 further research is essential to enhance our understanding of the selectivity and
388 implications of targeting ACBPs and AAA 18 in the context of human health.

389 **Purine metabolism: TRX**

390 TRX domain is a pivotal component of the purine metabolism pathway. Purines
391 play a crucial role in the synthesis of signaling molecules like cyclic adenosine 3,5-
392 monophosphate (cAMP) and are integral to the cellular energy system, including
393 adenosine triphosphate (ATP). Furthermore, they participate in the production of RNA
394 and DNA, working in tandem with pyrimidines [93,94]. However, experiments conducted
395 on *T. brucei* cultures have raised questions about the indispensability of this protein due
396 to the presence of a second TRX enzyme, TRX2 [95].

397 Significant differences in conservation between TRX2 and TRX1 exist among
398 various trypanosomatid species. TRX1 exhibits a lower degree of conservation, with
399 approximately 32% identity between *T. cruzi* and *T. brucei* [95]. However, our alignment
400 results indicate a higher identity of 50% between these species, with the lowest being
401 44% between Tcr and Lif. This supports the hypothesis that the domain we've identified
402 is TRX1. Studies have revealed that the absence of TRX2 leads to compromised parasite
403 proliferation in *T. brucei*, while the deficiency of TRX1 doesn't impact parasite
404 proliferation [95]. Consequently, it appears that TRX1 may not be a promising target for
405 drug development.

406 In the upcoming sections, we will explore several protein domains that we have
407 identified as promising candidates for multi-target studies in the development of novel
408 drugs. These domains play critical roles in Trypanosomatids, making them valuable
409 subjects for therapeutic investigation. To gain a comprehensive understanding of these
410 domains, we conducted extensive analyses, including the generation of their oligomeric
411 structures using AlphaFold2, rigorous evaluation of quality parameters, and the prediction
412 of druggable sites. Detailed sequences of these domains are available in the S3 table,
413 which will aid in further elucidating their potential as therapeutic targets.

414 ## Cyclogenesis: IFT - TPR

415 The tetra-trico peptide repeat domain (TPR) is a highly conserved structure
416 comprised of multiple repeats of 34 hydrophobic amino acids, found in various organisms
417 [96]. Its involvement in critical processes, such as transcriptional inhibition, stress
418 response, and mitochondrial protein transport, positions it as a promising target for drug
419 development [97–99].

420 Our evaluation of the TPR's conservation across trypanosomatid species revealed
421 high levels of identity, ranging from 73% (between Lif and Tbg/Tbr) to 98% (between
422 Tbg and Tcr). Furthermore, the human isoforms of IFT70 (A and B) showed an identity
423 range of 51% to 54% (Fig 4A). While the isoforms share a considerable degree of
424 similarity, it's essential to emphasize the significance of the IFT train's involvement in
425 various processes, which we will describe in detail below.

426 **Fig 4. TPR - Sequential and structural analysis.** (A) Shows the alignment of
427 trypanosomatid and human (Hsa) sequences, accompanied by a graphical summary
428 revealing the percentage of identity. Notably, the proline-rich region is highlighted,
429 demonstrating its conservation across species. The alignment matrix below employs
430 color-coding within the UGENE software, indicating dissimilarity percentages: white
431 (0%), red (10%), orange (25%), green (50%), blue (70%), and gray (90%). (B) Illustrates
432 the alignment between the 3D model generated by MHOLline for the Lif sequence (dark
433 blue) and the 4UZY template (gray). The TPR domain is depicted in pink, while the IFT52
434 protein, resolved with the PDBid: 4UZY structure, is represented in orange. (C) Results
435 for the druggability prediction where in a blue cartoon is IFT70 of *L. infantum* (Lif) and
436 the druggability score and corresponding surface area values are color-coded: green (0.84,

437 888.13 Å²), orange (0.81, 3311.56 Å²), purple (0.78, 878.57 Å²), pink (0.76, 746.52 Å²),
438 and blue (0.76, 605.84 Å²).

439 The TPR domain's versatility in cellular location, being found in various
440 compartments such as the nucleus, cytoplasm, and mitochondria [96], was confirmed in
441 our Psort prediction for trypanosomatid species. Tcr, Tbr, and Lif were predicted to be
442 mitochondrial, while Tbg was cytoplasmic. The TPR domain's role in mediating protein-
443 protein interactions and assembly of protein complexes was underscored by our
444 Structure-Based Functional Prediction (STBFP) results, which identified the IFT70
445 protein (PDBid: 4UZY) [100]. IFT70 is present in the IFT train which is a crucial
446 component of the intraflagellar transport protein complex responsible for cyclogenesis,
447 an evolutionarily conserved transport process involving the bidirectional movement of
448 particles within cilia [101].

449 As noted in the structural analyses, the 4UZY template corresponds to the IFT70
450 protein complexed with part of IFT52 from the organism *Chlamydomonas reinhardtii*
451 (PDBid: 4UZY). The interaction between IFT70 and the IFT52 protein occurs through a
452 proline-rich region [102], and we observed that most of these regions are conserved, even
453 in humans (Fig 4A). Therefore, the tertiary and primary structures exhibited significant
454 similarity to a functionally characterized protein (Fig 4 A and B), reinforcing the
455 connection between structure and functionality in proteins.

456 To enhance the accuracy of our structure prediction, we employed AlphaFold2 to
457 generate new models for each species, taking into account the oligomeric states of the
458 proteins. The results for IFT70 can be found in S1 Fig and to ensure the quality of these
459 models, conducted evaluations of the models using MolProbity [56] (Clash score,

460 MolProbity score, and Ramachandran favored regions), as well as ProSa [57] (as shown
461 in S4 Table). Notably, the best IFT70 model was identified for the *L. infantum* (Lif)
462 organism (accessible via this ZENODO link: https://doi.org/10.5281/zenodo.10072065).
463 Consequently, we employed this model to predict potential binding sites using
464 DoGSiteScorer [58], which assigns druggability scores within the range of 0 to 1. Higher
465 scores indicate more favorable druggable pockets, guiding our search for potential drug-
466 binding sites. The results for the Lif model indicate the presence of five promising regions
467 (Fig 4 C), with the majority of them located within the IFT52 binding site, reaffirming
468 the crucial role of this region in the functionality of IFT70.

469 Furthermore, the significance of the IFT52 protein in stabilizing the IFT-B1
470 complex and its direct involvement in cilia formation has been well-documented in
471 studies across various organisms [103,104]. For instance, the study by HUET et al. in
472 2019 demonstrated that other proteins within the IFT train, such as IFT172 (IFT-B2),
473 were pivotal for proper cyclogenesis and flagellar organization in *T. brucei*. These
474 findings underscore the indispensable role of the IFT train in cyclogenesis and emphasize
475 the critical contributions of individual proteins, like IFT52, in maintaining the complex's
476 stability and functionality.

477 Despite the significant progress in understanding the IFT train, questions remain
478 regarding the interactions and subunits involved in complex stabilization. However, the
479 potential of the IFT70 protein and other IFT train proteins as targets for drug development
480 in trypanosomatids is promising, given their pivotal role in critical cellular processes and
481 preliminary in vitro study results. Further investigations may reveal new drug
482 development opportunities.

483 **Nucleosides transferases: Nuc deoxyrib tr**

484 Nucleoside 2'-deoxyribosyltransferases (NDTs), also known as Nuc deoxyrib tr
485 according to domain databases' nomenclature, constitute a protein family that mediates
486 the transfer of the 2'-deoxyribosyl group between 2'-deoxyribonucleosides (donor) and
487 nucleobases (acceptor) [105]. These enzymes belong to the class of Nucleoside
488 Phosphorylases (NPs), which are further divided into ribo- and 2'-deoxyribonucleosides.
489 The class encompasses various subclasses, including type I NDTs (PDT), responsible for
490 deoxyribotransfer between purines only, and type II NDTs (NDT), which enable transfer
491 between purines and/or pyrimidines [106]. NDTs have been identified in various species,
492 including parasitic ones like *T. brucei*, *L. major*, and *L. mexicana*, as well as in
493 lactobacillus species like *Lactobacillus helveticus* [107,108].

494 Bosch et al., 2006 conducted a comparative analysis of NDTs among different
495 trypanosomatid species, including *T. cruzi*, *L. major*, *L. infantum*, and *T. brucei*, along
496 with a sequence from Lactobacillus helveticus. The study revealed that the percentage of
497 identity between *T. brucei* and other species ranged from 84.2% (*T. cruzi*) to 89.5% (*L.*
498 *major* and *L. infantum*), and only 21% with *L. helveticus*. Our analysis of predicted NDT
499 domain sequences showed identity ranging from 65% between Tcr and Lif to 100%
500 between Tbg and Tbr (Fig 5A).

501 Bosch's work also highlighted conserved regions among trypanosomatids, which
502 included key residues within the 2'-deoxyribose binding site, such as Ala15, Gly16,
503 Val19, Phe20, Pro46, Thr47, Glu50, Ile57, Arg58, Asn61, Arg125, Glu127, Phe129,
504 Asn134, Leu135, and Met136 [108]. Comparing these regions among the generated
505 structures, it was evident that most of them were conserved across species, with only a

few mutations observed, such as Thr47Val in Tcr and Lif; Glu50Ile and Arg125Ser in Lif; Glu50Met and Arg125Thr in Tcr (Fig 5B). The RMSD values between the mold and the models ranged from 0.467 Å (Tcr) to 0.611 Å (Lif).

Fig 5. NDT - Sequential and structural analysis. (A) The alignment reveals the comparison between trypanosomatid and human sequences. Notably, the highlighted region contains three amino acid mutations. The right side of the figure displays the aligned residues: Ala15, Gly16, Val19, Phe20, Pro46, Thr47, Glu50, Ile57, Arg58, Asn61, Arg125, Glu127, Phe129, Asn134, Leu135, and Met136. Among these residues, three mutations were observed: Thr47Val, Glu50Met, and Arg125Thr, as indicated in bold. The numbering of *T. cruzi* amino acids was adjusted to account for the eight fewer amino acids compared to the template. (B) The alignment between the predicted model for the Lif protein (light blue) and the template with PDBid: 2F62 (chain A in pink and chain B in gray). This template was selected due to having more mutations and the lowest RMSD. The aligned residues match those presented in part (A) and highlight the conserved regions among the sequences. (C) Results for the druggability prediction where Nuc deoxyrib tr Lif is in red (chain A) and blue (chain B) in a blue and the druggability score and corresponding surface area values are color-coded: orange 0.81 (1649.44 Å²); purple 0.57 (354.17 Å²); green 0.43 (231.66 Å²); blue 0.4 (264.39 Å²); pink 0.35 (213.49).

Following the utilization of AlphaFold2 and the subsequent evaluation of model quality (refer to S1 Fig and S4 table for details), we proceeded to select the Lif structure for predicting druggable sites. Among the identified sites, the orange region (Fig 5C) scored the highest, signifying its potential druggability. Notably, these druggable sites are

528 concentrated within a region that contains key amino acids critical for the proper
529 functioning of this protein.

530 Analysis in parasites such as *Trypanosoma brucei*, *Trypanosoma cruzi*, and
531 *Leishmania major*, showed that Nucleoside 2'-deoxyribosyltransferases (NDTs) play a
532 crucial role in the processes of recycling and recovery of purines, as well as in the de novo
533 biosynthesis of purines [108]. These parasites employ various enzymes to remove ribose
534 or deoxyribose from purine nucleosides, making it challenging to determine which
535 enzymes are essential in the purine turnover pathway.

536 In contrast, humans rely on Deoxynucleoside 5-monophosphate N-glycosidase
537 (DNPH1 – EC 3.2.2.) for similar functions, but differences in substrate specificity and
538 essentiality differentiate it from NDTs [109]. Furthermore, it has been suggested that
539 DNPH1 is not an essential gene, as many organisms can survive without it. These findings
540 support the idea that NDT and DNPH1 enzymes are distinct. To further compare the two,
541 we aligned the predicted NDT domain sequences of various species with the DNPH1
542 sequence of Homo sapiens (UniProtKB code: O43598). The identities between these
543 sequences ranged from 16% (between Lif and Hsa) to 22% (between Tcr and Hsa),
544 confirming their significant differences (Fig 5A).

545 Given the precariousness of enzymes involved in de novo purine biosynthesis in
546 *T. brucei* and their reliance on the host for purine uptake, inhibiting NDT, an enzyme
547 crucial for their processes, could significantly impact trypanosome replication [108]. In
548 light of the similarity among trypanosomatid sequences and the absence of NDT in
549 humans, NDT presents a promising therapeutic target for further exploration.

550 **Mitochondrial membrane: MIX**

551 The MIX domain is a unique protein structure anchored to the mitochondrial
552 membrane and found exclusively in kinetoplastids [110], a group of organisms that
553 includes Trypanosoma and Leishmania species. In these kinetoplastids, the precise
554 mechanisms of mitochondrial protein import and processing are not fully understood. One
555 approach to gaining insights into the function of the MIX domain is to study its
556 association with the Cytochrome C Oxidase (COX) multiprotein complex, as proposed
557 by Gorman et al. in 2011 [111].

558 Cytochrome C Oxidase (COX) plays a crucial role in mitochondrial cellular
559 respiration, driving oxidative phosphorylation to generate ATP molecules. This energy
560 production process involves various protein complexes (I, II, III, IV, and V) [112], with
561 COX representing complex IV in trypanosomatids [111]. Notably, the COX complex in
562 these organisms exhibits distinctive features in 8 of the 10 smaller subunits, for which no
563 counterparts are identified outside the Kinetoplastida group [113]. Early hypotheses
564 suggested that the MIX domain may be instrumental in directing or assembling COX
565 components [111].

566 The MIX domain comprises small sequences characterized by a conserved
567 primary structure, facilitating their targeting of mitochondria in parasitic and yeast cells
568 [110]. In their 2011 study, Gorman et al. compared the MIX sequence from *Trypanosoma*
569 *brucei* with that of other trypanosomatids, specifically *Leishmania major* and
570 *Trypanosoma cruzi*, revealing identity percentages ranging from 75% to 66%,
571 respectively. In the present study, comparisons of sequences predicted to contain this
572 domain show identity percentages ranging from 67% to 99% (Fig 6A).

573 **Fig 6. MIX - Sequential and structural analysis.** (A) This section features the alignment
574 of trypanosomatid sequences, with a specific emphasis on critical amino acid residues
575 (Arg86, Glu87, and Glu115) that play a pivotal role in protein dimerization. (B) The
576 structural alignment showcases the model sequence CAM65809 from *L. infantum* in light
577 blue, contrasted with the template PDBid: 3HA4 (F chain in pink and H chain in gray).
578 Notably, this template contains the sole mutation found in Asn158Lys, represented in
579 bold. Additionally, predictions from the DeepTMHMM web server are illustrated,
580 highlighting the transmembrane region in vibrant yellow and the extracellular region in
581 regal purple. A display of amino acid residues is provided on the right side of the image,
582 featuring elements like salt bridge formers (Arg86, Glu87, and Glu115), as well as
583 residues that engage in hydrogen bonds without mediation by water molecules (Gln80,
584 Glu160, Pro157, Gly159, and Asn163). (C) The results of druggability assessments for
585 Mix Lif are presented in red (chain A) and blue (chain B), utilizing a cartoon
586 representation (excluding the initial 70 amino acids). Druggability scores and their
587 corresponding surface area values are color-coded: green 0.87 (602.82 Å²); orange 0.81
588 (1446.28 Å²); purple 0.8 (772.76 Å²); pink 0.64 (404.49 Å²); blue 0.6 (369.83 Å²).

589

590 Moreover, Gorman et al. in 2011 identified three critical regions for the MIX
591 domain's structural integrity in various trypanosomatid species, using *Leishmania major*
592 (Lma) as the model organism. The first region involves three amino acid residues (Arg86,
593 Glu87, and Glu115) essential for dimerization. The second region consists of amino acid
594 residues that form salt bridges (Arg89, Glu90, and Glu118), while the third region

595 comprises amino acid residues that engage in hydrogen bonding without water molecule
596 mediation, including Gln80, Glu160, Pro157, Asn158, Gly159, and Asn163.

597 In our own work, we compared the same Lma sequence used by Gorman et al. in
598 2011 and found identity percentages ranging from 56% (between Lma and Tbg/Tbr) to
599 77% (between Lma and Lif). The first two regions remained conserved, while we
600 identified a single mutation in the third region, specifically in Lif: Asn158Lys (Fig 6B).
601 This mutation may impact the formation of chemical bonds due to the distinct
602 physicochemical characteristics of the lysine radical, suggesting the need for further
603 investigation to understand its effects on hydrogen bond formation.

604 In our structural analyses, the STBFP results and the template used led to the
605 structure PDBid:3HA4, representing a MIX protein from the *L. major* organism [114].
606 However, we observed that the generated models exhibited a significant region without a
607 clearly defined secondary structure, encompassing residues 1 to 70. To address this issue,
608 we employed the transmembrane region predictor, DeepTMHMM, based on prior
609 descriptions of the MIX domain. The results indicated a consensus for the Lif and
610 EAN89737 species of Tcr, while transmembrane regions were not predicted for Tbg and
611 Tbr (Table 2). This provides valuable insights into the structural characteristics of the
612 MIX domain in different trypanosomatid species.

613 **Table 2. Results DeepTMHMM**

Organism	Regions	Prediction
Lif	1 – 22	I
	23 – 32	Tmhelix
	33	O

	34 – 43	Tmhelix
	44 –T5	I
Tbg	1 – 198	I
Tbr	1 – 198	I
Tcr	1 – 27	I
	28 – 37	Tmhelix
	38	O
	39 – 48	Tmhelix
	49 – 198	I

614 [a]i: Intracelular region.

615 [b]TMhelix: Transmembrane region.

616 [c]o: Extracellular region.

617 Uboldi et al. (2006) postulated the existence of an N-terminal amphipathic α-helix
618 within the MIX domain, followed by a subsequent transmembrane region, with the
619 transmembrane portion typically being laterally inserted into the mitochondrial
620 membrane. In the case of *L. major*, this transmembrane region consists of the initial 45
621 residues [110]. The DeepTMHMM results for proteins that possess the MIX domain, even
622 those with unknown functions, align with this characteristic.

623 In terms of the structural attributes of MIX, it comprises seven α-helices that
624 collectively form the coiled domain [114]. This pattern of folding is consistent with our
625 generated models, and the root mean square deviation (RMSD) values derived from the
626 comparison of these models with the template (PDBid: 3HA4) ranged from 2.069 Å (Tbr)
627 to 2.493 Å (Lif). It is worth noting that the variability in RMSD values is likely associated

628 with the N-terminal segment (residues 1-45), for which the template lacks a well-defined
629 structure (Fig 6B).

630 In summary, the MIX domain is universally expressed across all evolutionary
631 stages of Leishmania, including the amastigote form residing in the mammalian host.
632 Intriguingly, both in vitro and in vivo experiments have demonstrated that the deletion of
633 two MIX gene alleles leads to a reduction in infectivity and virulence [110]. While initial
634 assessments in *T. brucei* suggested that COX activity was solely required in the
635 epimastigote form, indicating that the interaction between MIX and COX was specific to
636 this developmental stage [110,114], subsequent research has unveiled the presence of
637 MIX in the blood stage. This suggests that MIX could engage with other proteins and
638 catalyze additional reactions [111]. Given the unique characteristics of MIX within the
639 Kinetoplastida group and its pivotal roles in diverse processes, it emerges as a promising
640 molecular target for the development of novel drugs.

641 # Conclusion

642 Based on the SEquence-Based Functional Prediction (SEBFP) and Structure-
643 Based Functional Prediction (STBFP) results, we were able to predict functional domains
644 for 67 trypanosomatid proteins, of which nine domains were shared among the majority
645 of trypanosomatid species. These domains were analyzed, by consulting bibliographic
646 resources, and we were able to clarify their functions and their existing correlations with
647 trypanosomatids.

648 Among the domains analyzed, we found those that had several isoforms, not being
649 crucial for the parasites under study, TRX, ACBP, and AAA 18 [95,115][115].

650 Additionally, the complexity of the pathways in which they are inserted, along with the
651 lack of detailed information on all the inherent metabolic processes, currently hinders the
652 development of inhibitors for these targets.

653 In addition to the aforementioned domains, we also identified two domains that
654 are integral to the ubiquitination process, a crucial mechanism for protein degradation.
655 Research has shown that mutations in UFC1, for instance, can result in neurological
656 impairment [63]. Similarly, Ufm1 can recruit transcriptional coactivators in humans,
657 contributing to excessive cell proliferation and ultimately leading to conditions such as
658 breast cancer [66]. Both domains possess complex intricacies that need to be addressed
659 before considering them as therapeutic targets. Nonetheless, the proteasome, another
660 component of the ubiquitination pathway, has been studied extensively due to its
661 significant role in *T. cruzi* [116]. Inhibitors of the proteasome have already been tested,
662 and the results demonstrate impairment of the infective evolutionary form, indicating that
663 a study considering the proteasome as a therapeutic target could yield promising results.

664 Lastly, we identified domains that can be essential for the survival and virulence
665 of the trypanosomatids under study, namely Nuc deoxyrib tr, MIX, and TPR. Nuc
666 deoxyrib tr is critical in de novo purine biosynthesis, and studies on *T. brucei* have shown
667 that these parasites do not produce an abundance of such proteins, making them heavily
668 reliant on the catalytic activity of this domain for survival [108]. Regarding the MIX
669 domain, this is exclusive to kinetoplastids and *in vitro* and *in vivo* tests demonstrate that
670 its deletion led to a reduction in infectivity and virulence [111]. Furthermore, we also
671 discovered the TPR domain, specifically the IFT70 protein, which is involved in a protein
672 complex that facilitates intraflagellar transport. IFT70 is a vital member of the IFT train,

673 and studies on *T. brucei* have confirmed that the train is necessary for the transportation
674 of ciliary assembly precursors from the cytoplasm into the cell body [103]. Therefore, we
675 believe these targets hold promise for multi-target studies aimed at developing innovative
676 therapeutic approaches.

677 # Acknowledgments

678 RSLR expresses sincere gratitude to the Graduate Program in Computational and Systems
679 Biology at Instituto Oswaldo Cruz for their unwavering support. Furthermore, she would
680 like to extend her heartfelt appreciation to Instituto Oswaldo Cruz and the Capes program
681 for generously providing financial assistance (No. 88887.596358/2020-00, No.
682 88887.801809/2023-00).

683 # References

684 [1] Rasoanoro M, Ramasindrazana B, Goodman SM, Rajerison M,
685 Randrianarivelojosia M. A review of Trypanosoma species known from Malagasy
686 vertebrates. Malagasy Natiora 2019;13:65–75.
687 [2] Thompson CK, Godfrey SS, Thompson RCA. Trypanosomes of Australian
688 mammals: A review. Int J Parasitol Parasites Wildl 2014;3:57–66.
689 https://doi.org/10.1016/j.ijppaw.2014.02.002.
690 [3] Fahrion A, Gasimov E, Joseph S, Grout L, Allan M, Postigo JR. Surveillance of
691 leishmaniasis in the WHO European Region. Rev Epidemiol Sante Publique
692 2018;66:S394. https://doi.org/10.1016/j.respe.2018.05.429.
693 [4] WHO. Neglected tropical diseases 2017.

694 [5] Imran M, Khan SA, Alshammari MK, Alqahtani AM, Alanazi TA, Kamal M, et
695 al. Discovery, Development, Inventions and Patent Review of Fexinidazole: The First
696 All-Oral Therapy for Human African Trypanosomiasis. Pharmaceuticals 2022;15.
697 https://doi.org/10.3390/ph15020128.
698 [6] Capewell P, Atkins K, Weir W, Jamonneau V, Camara M, Clucas C, et al.
699 Resolving the apparent transmission paradox of African sleeping sickness. PLoS Biol
700 2019;17:1–6. https://doi.org/10.1371/journal.pbio.3000105.
701 [7] Abdellahi L, Iraji F, Mahmoudabadi A, Hejazi SH. Vaccination in leishmaniasis:
702 A review article. Iran Biomed J 2022;26:1–35. https://doi.org/10.52547/ibj.26.1.35.
703 [8] Pires M, Wright B, Kaye PM, da Conceição V, Churchill RC. The impact of
704 leishmaniasis on mental health and psychosocial well-being: A systematic review. PLoS
705 One 2019;14:e0223313. https://doi.org/10.1371/journal.pone.0223313.
706 [9] Franco J, Scarone L, Comini MA. Drugs and Drug Resistance in African and
707 American Trypanosomiasis. vol. 51. 1st ed. Elsevier Inc.; 2018.
708 https://doi.org/10.1016/bs.armc.2018.08.003.
709 [10] De Fuentes-Vicente JA, Vidal-López DG, Flores-Villegas AL, Moreno-
710 Rodríguez A, De Alba-Alvarado MC, Salazar-Schettino PM, et al. Trypanosoma cruzi:
711 A review of biological and methodological factors in Mexican strains. Acta Trop
712 2019;195:51–7. https://doi.org/10.1016/j.actatropica.2019.04.024.
713 [11] Field MC, Horn D, Fairlamb AH, Ferguson MAJ, Gray DW, Read KD, et al. Anti-
714 trypanosomatid drug discovery: An ongoing challenge and a continuing need. Nat Rev
715 Microbiol 2017;15:217–31. https://doi.org/10.1038/nrmicro.2016.193.
716 [12] Pradhan S, Schwartz RA, Patil A, Grabbe S, Goldust M. Treatment options for
717 leishmaniasis. Clin Exp Dermatol 2022;47:516–21. https://doi.org/10.1111/ced.14919.

718 [13] Zingales B. Trypanosoma cruzi genetic diversity: Something new for something
719 known about Chagas disease manifestations, serodiagnosis and drug sensitivity. Acta
720 Trop 2018;184:38–52. https://doi.org/10.1016/j.actatropica.2017.09.017.
721 [14] Deotarse PP, Jain AS, Baile MB, Kolhe NS, Kulkarni AA. Drug Repositioning :
722 A Review. Int J Pharma Res Rev 2018;4:51–8.
723 [15] Luo H, Li M, Yang M, Wu FX, Li Y, Wang J. Biomedical data and computational
724 models for drug repositioning: A comprehensive review. Brief Bioinform 2021;22:1604–
725 19. https://doi.org/10.1093/bib/bbz176.
726 [16] Sadowski MI, Jones DT. The sequence-structure relationship and protein function
727 prediction. Curr Opin Struct Biol 2009;19:357–62.
728 https://doi.org/10.1016/j.sbi.2009.03.008.
729 [17] Radivojac P, Clark W, Oron T, Schnoes A. A large-scale evaluation of
730 computational protein function prediction. Nature 2013;10:221–7.
731 https://doi.org/10.1038/nmeth.2340.A.
732 [18] Araújo CL, Blanco I, Souza L, Tiwari S, Pereira LC, Ghosh P, et al. In silico
733 functional prediction of hypothetical proteins from the core genome of Corynebacterium
734 pseudotuberculosis biovar ovis. PeerJ 2020;8:1–24. https://doi.org/10.7717/peerj.9643.
735 [19] Shahbaaz M, Hassan MI, Ahmad F. Functional annotation of conserved
736 hypothetical proteins from Haemophilus influenzae Rd KW20. PLoS One 2013;8.
737 https://doi.org/10.1371/journal.pone.0084263.
738 [20] VERLI H. Bioinformática da Biologia à Flexibilidade Molecular. vol. 53. 2014.
739 https://doi.org/10.1017/CBO9781107415324.004.
740 [21] Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use
741 sequence and structure information to predict protein function. PLoS Comput Biol

742 2008;4. https://doi.org/10.1371/journal.pcbi.1000160.
743 [22] Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, et al. Database
744 resources of the National Center for Biotechnology Information. Nucleic Acids Res
745 2016;44:D7–19. https://doi.org/10.1093/nar/gkv1290.
746 [23] Capriles PVSZ, Guimarães ACR, Otto TD, Miranda AB, Dardenne LE, Degrave
747 WM. Structural modelling and comparative analysis of homologous, analogous and
748 specific proteins from Trypanosoma cruzi versus Homo sapiens: Putative drug targets for
749 chagas' disease treatment. BMC Genomics 2010;11. https://doi.org/10.1186/1471-2164-
750 11-610.
751 [24] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment
752 search tool. J Mol Biol 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-
753 2.
754 [25] Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M, et
755 al. Comparative protein structure modeling using Modeller. vol. Chapter 5. 2006.
756 https://doi.org/10.1002/0471250953.bi0506s15.Comparative.
757 [26] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for
758 understanding genome variations in KEGG. OXFORD Acad Nucleic Acids Res 2018.
759 [27] El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam
760 protein families database in 2019. Nucleic Acids Res 2019;47:D427–32.
761 https://doi.org/10.1093/nar/gky995.
762 [28] Letunic I, Khedkar S, Bork P. SMART: Recent updates, new developments and
763 status in 2020. Nucleic Acids Res 2021;49:D458–60.
764 https://doi.org/10.1093/nar/gkaa937.
765 [29] Sigrist CJA, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and

766 continuing developments at PROSITE. Nucleic Acids Res 2013;41:344–7.
767 https://doi.org/10.1093/nar/gks1067.
768 [30] Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al.
769 CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res
770 2020;48:D265–8. https://doi.org/10.1093/nar/gkz991.
771 [31] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al.
772 STRING v11: Protein-protein association networks with increased coverage, supporting
773 functional discovery in genome-wide experimental datasets. Nucleic Acids Res
774 2019;47:D607–13. https://doi.org/10.1093/nar/gky1131.
775 [32] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al.
776 CATH: Increased structural coverage of functional space. Nucleic Acids Res
777 2021;49:D266–73. https://doi.org/10.1093/nar/gkaa1079.
778 [33] Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server:
779 2018 update. Nucleic Acids Res 2018;46:W200–4. https://doi.org/10.1093/nar/gky448.
780 [34] Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology
781 search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res
782 2013;41. https://doi.org/10.1093/nar/gkt263.
783 [35] Peabody MA, Laird MR, Vlasschaert C, Lo R, Brinkman FSL. PSORTdb:
784 Expanding the bacteria and archaea protein subcellular localization database to better
785 reflect diversity in cell envelope structures. Nucleic Acids Res 2016;44:D663–8.
786 https://doi.org/10.1093/nar/gkv1271.
787 [36] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al.
788 UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res
789 2021;49:D480–9. https://doi.org/10.1093/nar/gkaa1100.

790 [37] Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res
791 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
792 [38] Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et
793 al. TriTrypDB: A functional genomic resource for the Trypanosomatidae. Nucleic Acids
794 Res 2009;38:457–62. https://doi.org/10.1093/nar/gkp851.
795 [39] Amos B, Aurrecoechea C, Barba M, Barreto A, Basenko EY, Bażant W, et al.
796 VEuPathDB: The eukaryotic pathogen, vector and host bioinformatics resource center.
797 Nucleic Acids Res 2022;50:D898–911. https://doi.org/10.1093/nar/gkab929.
798 [40] Okonechnikov K, Golosova O, Fursov M, Varlamov A, Vaskin Y, Efremov I, et
799 al. Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics 2012;28:1166–7.
800 https://doi.org/10.1093/bioinformatics/bts091.
801 [41] Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of
802 sequences with insertions. Proc Natl Acad Sci U S A 2005;102:10557–62.
803 https://doi.org/10.1073/pnas.0409137102.
804 [42] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high
805 throughput. Nucleic Acids Res 2004;32:1792–7. https://doi.org/10.1093/nar/gkh340.
806 [43] Naser-Khdour S, Minh BQ, Zhang W, Stone EA, Lanfear R, Bryant D. The
807 Prevalence and Impact of Model Violations in Phylogenetic Analysis. Genome Biol Evol
808 2019;11:3341–52. https://doi.org/10.1093/gbe/evz193.
809 [44] Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O,
810 Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural
811 networks. Nat Biotechnol 2019;37:420–3. https://doi.org/10.1038/s41587-019-0036-z.
812 [45] Hallgren J, Tsirigos KD, Damgaard Pedersen M, Juan J, Armenteros A, Marcatili
813 P, et al. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural

814 networks. BioRxiv 2022:2022.04.08.487609.
815 [46] Hubbard TJP, Murzin AG, Brenner SE, Chothia C. SCOP: A Structural
816 Classification of Proteins database. Nucleic Acids Res 1997;25:236–9.
817 https://doi.org/10.1093/nar/25.1.236.
818 [47] Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, et
819 al. MMDB and VAST+: Tracking structural similarities between macromolecular
820 complexes. Nucleic Acids Res 2014;42:297–303. https://doi.org/10.1093/nar/gkt1208.
821 [48] Madej T, Marchler-Bauer A, Lanczycki C, Zhang D, Bryant SH. Biological
822 Assembly Comparison with VAST+. Struct Bioinforma 2020;2112.
823 [49] Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. FATCAT 2.0: Towards a
824 better understanding of the structural diversity of proteins. Nucleic Acids Res
825 2020;48:W60–4. https://doi.org/10.1093/NAR/GKAA443.
826 [50] Holm L. Using Dali for protein structure comparison. Struct Bioinforma Methods
827 Protoc Methods Mol Biol 2020;2112:29–42.
828 [51] DeLano WL. The PyMol molecular graphics system 2002.
829 [52] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly
830 accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
831 https://doi.org/10.1038/s41586-021-03819-2.
832 [53] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al.
833 AlphaFold Protein Structure Database: Massively expanding the structural coverage of
834 protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50:D439–
835 44. https://doi.org/10.1093/nar/gkab1061.
836 [54] Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein
837 Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinforma

838 2020;72. https://doi.org/10.1002/cpbi.108.
839 [55] Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M.
840 ColabFold: making protein folding accessible to all. Nat Methods 2022;19:679–82.
841 https://doi.org/10.1038/s41592-022-01488-1.
842 [56] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al.
843 MolProbity: More and better reference data for improved all-atom structure validation.
844 Protein Sci 2018;27:293–315. https://doi.org/10.1002/pro.3330.
845 [57] Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition
846 of errors in three-dimensional structures of proteins. Nucleic Acids Res 2007;35:407–10.
847 https://doi.org/10.1093/nar/gkm290.
848 [58] Volkamer A, Kuhn D, Rippmann F, Rarey M. Dogsitescorer: A web server for
849 automatic binding site prediction, analysis and druggability assessment. Bioinformatics
850 2012;28:2074–5. https://doi.org/10.1093/bioinformatics/bts310.
851 [59] Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al.
852 The InterPro protein families and domains database: 20 years on. Nucleic Acids Res
853 2021;49:D344–54. https://doi.org/10.1093/nar/gkaa977.
854 [60] Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, et
855 al. CDD: Conserved domains and protein three-dimensional structure. Nucleic Acids Res
856 2013;41:348–52. https://doi.org/10.1093/nar/gks1243.
857 [61] Kimura Y, Tanaka K. Regulatory mechanisms involved in the control of ubiquitin
858 homeostasis. J Biochem 2010;147:793–8.
859 [62] Komatsu M, Chiba T, Tatsumi K, Lemura SI, Tanida I, Okazaki N, et al. A novel
860 protein-conjugating system for Ufm1, a ubiquitin-fold modifier. EMBO J 2004;23:1977–
861 86. https://doi.org/10.1038/sj.emboj.7600205.

862 [63] Nahorski MS, Maddirevula S, Ishimura R, Alsahli S, Brady AF, Begemann A, et
863 al. Biallelic UFM1 and UFC1 mutations expand the essential role of ufmylation in brain
864 development. Brain 2018;141:1934–45. https://doi.org/10.1093/brain/awy135.
865 [64] Chung WL, Leung KF, Carrington M, Field MC. Ubiquitylation is required for
866 degradation of transmembrane surface proteins in Trypanosomes. Traffic 2008;9:1681–
867 97. https://doi.org/10.1111/j.1600-0854.2008.00785.x.
868 [65] Leung KF, Riley FS, Carrington M, Field MC. Ubiquitylation and developmental
869 regulation of invariant surface protein expression in trypanosomes. Eukaryot Cell
870 2011;10:916–31. https://doi.org/10.1128/EC.05012-11.
871 [66] Yoo HM, Kang SH, Kim JY, Lee JE, Seong MW, Lee SW, et al. Modification of
872 asc1 by ufm1 is crucial for erα transactivation and breast cancer development. Mol Cell
873 2014;56:261–74. https://doi.org/10.1016/j.molcel.2014.08.007.
874 [67] Huang X, Dixit VM. Drugging the undruggables: Exploring the ubiquitin system
875 for drug development. Cell Res 2016;26:484–98. https://doi.org/10.1038/cr.2016.31.
876 [68] Hochstrasser M. Evolution and function of ubiquitin-like protein-conjugation
877 systems 2000;2:153–7.
878 [69] González J, Ramalho-Pinto FJ, Frevert U, Ghiso J, Tomlinson S, Scharfstein J
879 ulio, et al. Proteasome activity is required for the stage-specific transformation of a
880 protozoan parasite. J Exp Med 1996;184:1909–18.
881 [70] Kurup SP, Tarleton RL. The trypanosoma cruzi flagellum is discarded via
882 asymmetric cell division following invasion and provides early targets for protective
883 CD8+ T cells. Cell Host Microbe 2014;16:439–49.
884 https://doi.org/10.1016/j.chom.2014.09.003.
885 [71] Ayala-Castro C, Saini A, Outten FW. Fe-S Cluster Assembly Pathways in

886 Bacteria. Microbiol Mol Biol Rev 2008;72:110–25. https://doi.org/10.1128/mmbr.00034-
887 07.
888 [72] Kispal G, Csere P, Prohl C, Lill R. The mitochondrial proteins Atm1p and Nfs1p
889 are essential for biogenesis of cytosolic Fe/S proteins. EMBO J 1999;18:3981–9.
890 https://doi.org/10.1093/emboj/18.14.3981.
891 [73] Mettert EL, Kiley PJ. How Is Fe-S Cluster Formation Regulated? Annu Rev
892 Microbiol 2015;69:505–26. https://doi.org/10.1146/annurev-micro-091014-
893 104457.How.
894 [74] Py B, Barras F. Building Fe-S proteins: Bacterial strategies. Nat Rev Microbiol
895 2010;8:436–46. https://doi.org/10.1038/nrmicro2356.
896 [75] Basu S, Netz DJ, Haindrich AC, Herlerth N, Lagny TJ, Pierik AJ, et al. Cytosolic
897 iron-sulphur protein assembly is functionally conserved and essential in procyclic and
898 bloodstream Trypanosoma brucei. Mol Microbiol 2014;93:897–910.
899 https://doi.org/10.1111/mmi.12706.
900 [76] Stehling O, Mascarenhas J, Vashisht AA, Sheftel AD, Niggemeyer B, Rösser R,
901 et al. Human CIA2A-FAM96A and CIA2B-FAM96B integrate iron homeostasis and
902 maturation of different subsets of cytosolic-nuclear iron-sulfur proteins. Cell Metab
903 2013;18:187–98. https://doi.org/10.1016/j.cmet.2013.06.015.
904 [77] Schwamb B, Pick R, Fernández SBM, Völp K, Heering J, Dötsch V, et al.
905 FAM96A is a novel pro-apoptotic tumor suppressor in gastrointestinal stromal tumors.
906 Int J Cancer 2015;137:1318–29. https://doi.org/10.1002/ijc.29498.
907 [78] Yin A, Luo Y, Chen W, He M, Deng JH, Zhao N, et al. FAM96A Protects Mice
908 From Dextran Sulfate Sodium (DSS)-Induced Colitis by Preventing Microbial Dysbiosis.
909 Front Cell Infect Microbiol 2019;9:1–14. https://doi.org/10.3389/fcimb.2019.00381.

910 [79] Yin A, Chen W, Cao L, Li Q, Zhu X, Wang L. FAM96A knock-out promotes
911 alternative macrophage polarization and protects mice against sepsis. Clin Exp Immunol
912 2021;203:433–47. https://doi.org/10.1111/cei.13555.
913 [80] Lukeš J, Basu S. Fe/S protein biogenesis in trypanosomes - A review. Biochim
914 Biophys Acta - Mol Cell Res 2015;1853:1481–92.
915 https://doi.org/10.1016/j.bbamcr.2014.08.015.
916 [81] Sipos K, Lange H, Fekete Z, Ullmann P, Lill R, Kispal G. Maturation of cytosolic
917 iron-sulfur proteins requires glutathione. J Biol Chem 2002;277:26944–9.
918 https://doi.org/10.1074/jbc.M200677200.
919 [82] Long S, Changmai P, Tsaousis AD, Skalický T, Verner Z, Wen YZ, et al. Stage-
920 specific requirement for Isa1 and Isa2 proteins in the mitochondrion of Trypanosoma
921 brucei and heterologous rescue by human and Blastocystis orthologues. Mol Microbiol
922 2011;81:1403–18. https://doi.org/10.1111/j.1365-2958.2011.07769.x.
923 [83] Milne KG, Güther MLS, Ferguson MAJ. Acyl-CoA binding protein is essential
924 in bloodstream form Trypanosoma brucei. Mol Biochem Parasitol 2001;112:301–4.
925 https://doi.org/10.1016/S0166-6851(00)00369-8.
926 [84] Konrad M. Molecular analysis of the essential gene for adenylate kinase from the
927 fission yeast Schizosaccharomyces pombe. J Biol Chem 1993;268:11326–34.
928 https://doi.org/10.1016/s0021-9258(18)82128-7.
929 [85] Milne KG, Ferguson MAJ. Cloning, expression, and characterization of the acyl-
930 coA-binding protein in African trypanosomes. J Biol Chem 2000;275:12503–8.
931 https://doi.org/10.1074/jbc.275.17.12503.
932 [86] Lee SH, Stephens JL, Englund PT. A fatty-acid synthesis mechanism specialized
933 for parasitism. Nat Rev Microbiol 2007;5:287–97. https://doi.org/10.1038/nrmicro1617.

934 [87] Smith TK, Bütikofer P. Lipid metabolism in Trypanosoma brucei. Mol Biochem
935 Parasitol 2010;172:66–79. https://doi.org/10.1016/j.molbiopara.2010.04.001.
936 [88] Mehlert A, Bond CS, Ferguson MAJ. The glycoforms of a Trypanosoma brucei
937 variant surface glycoprotein and molecular modeling of a glycosylated surface coat.
938 Glycobiology 2002;12:607–12. https://doi.org/10.1093/glycob/cwf079.
939 [89] McConville MJ, Mullin KA, Ilgoutz SC, Teasdale RD. Secretory Pathway of
940 Trypanosomatid Parasites. Microbiol Mol Biol Rev 2002;66:122–54.
941 https://doi.org/10.1128/mmbr.66.1.122-154.2002.
942 [90] Bouvier LA, Miranda MR, Canepa GE, Alves MJM, Pereira CA. An expanded
943 adenylate kinase gene family in the protozoan parasite Trypanosoma cruzi. Biochim
944 Biophys Acta - Gen Subj 2006;1760:913–21.
945 https://doi.org/10.1016/j.bbagen.2006.02.013.
946 [91] Klepinin A, Zhang S, Klepinina L, Rebane-Klemm E, Terzic A, Kaambre T, et al.
947 Adenylate Kinase and Metabolic Signaling in Cancer Cells. Front Oncol 2020;10:1–9.
948 https://doi.org/10.3389/fonc.2020.00660.
949 [92] Berger J, Dorninger F, Forss-Petter S, Kunze M. Peroxisomes in brain
950 development and function. Biochim Biophys Acta - Mol Cell Res 2016;1863:934–55.
951 https://doi.org/10.1016/j.bbamcr.2015.12.005.
952 [93] Hubert L, Sutton VR. Disorders of purine and pyrimidine metabolism. Elsevier
953 Inc.; 2017. https://doi.org/10.1016/b978-0-12-802896-4.00009-2.
954 [94] Yin J, Ren W, Huang X, Deng J, Li T, Yin Y. Potential Mechanisms Connecting
955 Purine Metabolism and Cancer Therapy. Front Immunol 2018;9:1–8.
956 https://doi.org/10.3389/fimmu.2018.01697.
957 [95] Currier RB, Ulrich K, Leroux AE, Dirdjaja N, Deambrosi M, Bonilla M, et al. An

958 essential thioredoxin-type protein of Trypanosoma brucei acts as redox-regulated
959 mitochondrial chaperone. vol. 15. 2019. https://doi.org/10.1371/journal.ppat.1008065.
960 [96] Zeytuni N, Zarivach R. Structural and functional discussion of the tetra-trico-
961 peptide repeat, a protein interaction module. Structure 2012;20:397–405.
962 https://doi.org/10.1016/j.str.2012.01.006.
963 [97] Young JC, Barral JM, Hartl FU. More than folding: localized functions of
964 cytosolic chaperones. Trends Biochem Sci 2003;28.
965 [98] Baker MJ, Frazier AE, Gulbis JM, Ryan MT. Mitochondrial protein-import
966 machinery: correlating structure with function. Trends Cell Biol 2007;17.
967 [99] Brocard C, Hartig A. Peroxisome targeting signal 1: Is it really a simple
968 tripeptide? Biochim Biophys Acta 2006;1763:1565–73.
969 [100] Takei R, Katoh Y, Nakayama K. Robust interaction of IFT70 with IFT52–IFT88
970 in the IFT-B complex is required for ciliogenesis. Biol Open 2018;7.
971 https://doi.org/10.1242/bio.033241.
972 [101] Taschner M, Kotsis F, Braeuer P, Wolfgang Kuehn E, Lorentzen E. Crystal
973 structures of IFT70/52 and IFT52/46 provide insight into intraflagellar transport B core
974 complex assembly. J Cell Biol 2014;207:269–82. https://doi.org/10.1083/jcb.201408002.
975 [102] Guido RVC, Andricopulo AD, Oliva G. Drug design, biotechnology and
976 medicinal chemistry: applications to infectious diseases. Estud Avançados 2010;24:81–
977 98. https://doi.org/10.1016/j.neuroscience.2012.06.055.
978 [103] Wachter S, Jung J, Shafiq S, Basquin J, Fort C, Bastin P, et al. Binding of IFT22
979 to the intraflagellar transport complex is essential for flagellum assembly. EMBO J
980 2019;38:1–18. https://doi.org/10.15252/embj.2018101251.
981 [104] Bertiaux E, Mallet A, Rotureau B, Bastin P. Intraflagellar transport during

982 assembly of flagella of different length in Trypanosoma brucei isolated from tsetse flies.
983 J Cell Sci 2020;133. https://doi.org/10.1242/jcs.248989.
984 [105] Cardinaud R, Holguin J. Nucleoside deoxyribosyltransferase-II from
985 Lactobacillus helveticus. Substrate specificity studies. Pyrimidine bases as acceptors.
986 Biochim Biophys Acta - Enzymol 1979;568:339–47.
987 [106] Fresco-Taboada A, De La Mata I, Arroyo M, Fernández-Lucas J. New insights on
988 nucleoside 2′-deoxyribosyltransferases: A versatile biocatalyst for one-pot one-step
989 synthesis of nucleoside analogs. Appl Microbiol Biotechnol 2013;97:3773–85.
990 https://doi.org/10.1007/s00253-013-4816-y.
991 [107] Fernández-Lucas J, Fresco-Taboada A, Acebal C, De La Mata I, Arroyo M.
992 Enzymatic synthesis of nucleoside analogues using immobilized 2′-
993 deoxyribosyltransferase from Lactobacillus reuteri. Appl Microbiol Biotechnol
994 2011;91:317–27. https://doi.org/10.1007/s00253-011-3221-7.
995 [108] Bosch J, Robien MA, Mehlin C, Boni E, Riechers A, Buckner FS, et al. Using
996 fragment cocktail crystallography to assist inhibitor design of Trypanosoma brucei
997 nucleoside 2-deoxyribosyltransferase. J Med Chem 2006;49:5939–46.
998 https://doi.org/10.1021/jm060429m.
999 [109] Reintamm T, Aas-Valleriani N, Kelve M. Evolutionary distribution of
1000 deoxynucleoside 5-monophosphate N-glycosidase, DNPH1. Gene 2019;683:1–11.
1001 https://doi.org/10.1016/j.gene.2018.10.005.
1002 [110] Uboldi AD, Lueder FB, Walsh P, Spurck T, McFadden GI, Curtis J, et al. A
1003 mitochondrial protein affects cell morphology, mitochondrial segregation and virulence
1004 in Leishmania. Int J Parasitol 2006;36:1499–514.
1005 https://doi.org/10.1016/j.ijpara.2006.08.006.

1006 [111] Zíková A, Panigrahi AK, Uboldi AD, Dalley RA, Handman E, Stuart K. Structural
1007 and functional association of Trypanosoma brucei MIX protein with cytochrome c
1008 oxidase complex. Eukaryot Cell 2008;7:1994–2003. https://doi.org/10.1128/EC.00204-
1009 08.
1010 [112] Fitzpatrick F. Cyclooxygenase Enzymes: Regulation and Function. Curr Pharm
1011 Des 2005;10:577–88. https://doi.org/10.2174/1381612043453144.
1012 [113] Speijer D, Breek CKD, Muijsers AO, Groenevelt PX, Dekker H, De Haan A, et
1013 al. The sequence of a small subunit of cytochrome c oxidase from Crithidia fasciculata
1014 which is homologous to mammalian subunit IV. FEBS Lett 1996;381:123–6.
1015 https://doi.org/10.1016/0014-5793(96)00095-6.
1016 [114] Gorman MA, Uboldi AD, Walsh PJ, Tan KS, Hansen G, Huyton T, et al. Crystal
1017 structure of the Leishmania major MIX protein: A scaffold protein that mediates protein-
1018 protein interactions. Protein Sci 2011;20:1060–8. https://doi.org/10.1002/pro.631.
1019 [115] Burton M, Rose TM, Færgeman NJ, Knudsen J. Evolution of the acyl-CoA
1020 binding protein (ACBP). Biochem J 2005;392:299–307.
1021 https://doi.org/10.1042/BJ20050664.
1022 [116] Bijlmakers MJ. Ubiquitination and the Proteasome as Drug Targets in
1023 Trypanosomatid Diseases. Front Chem 2021;8:1–18.
1024 https://doi.org/10.3389/fchem.2020.630888.
1025

1026 # Supporting Files

1027 **S1 Fig. AlphaFold2 prediction.** Presented in both cartoon and surface views. These
1028 models pertain to the TPR, Nuc deoxyrib tr, and MIX domains, covering various species,

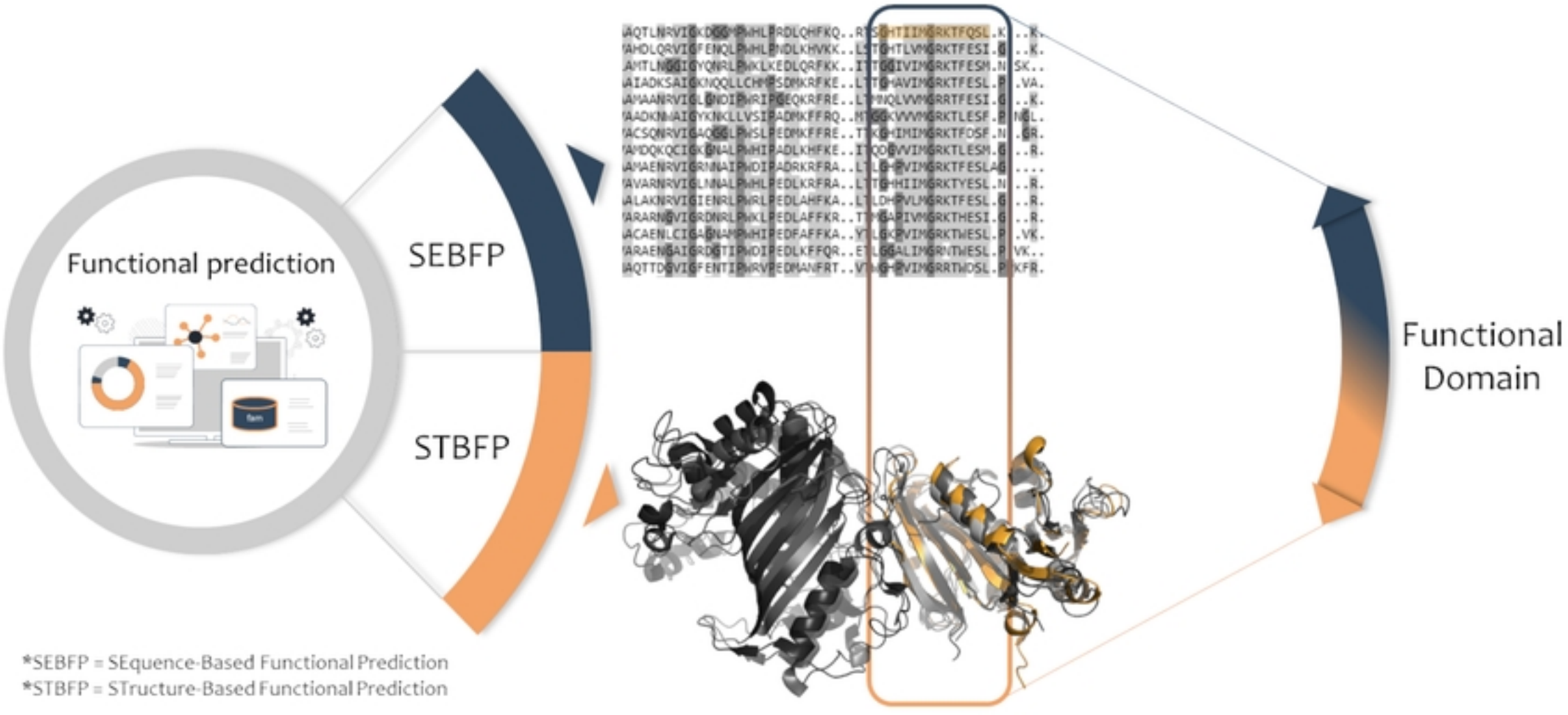
namely *L. infantum* (Lif), *T. brucei gambiensi* (Tbg), *T. brucei brucei* (Tbr), and *T. cruzi* (Tcr). In this representation, monomeric TPR models are depicted in vibrant purple, offering insights into the individual subunits. Dimeric structures are presented with chain A in a distinct orange shade and chain B in a contrasting blue hue, allowing for clear differentiation between the two subunits. Arrows within the figures highlight regions where MIX transmembrane segments are predicted, indicating potential areas of interest for further study and analysis.

S1 Table. Domain index.

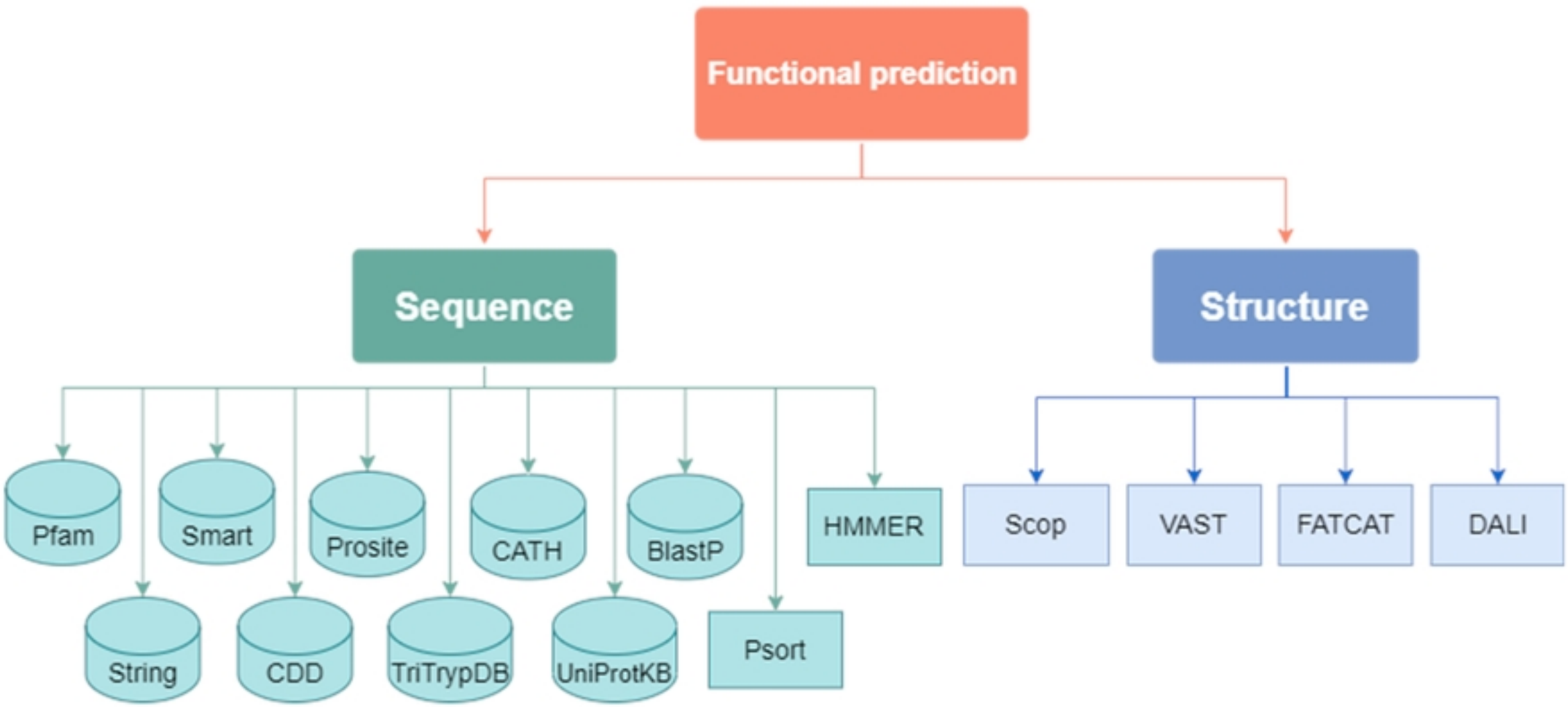
S2 Table. Functional prediction results. The abbreviations in the table mean WR – Without Results; UP - Uncharacterized Protein; UCP - Uncharacterized Conserved Protein.

S3 Table. Proteins/Domains sequences.

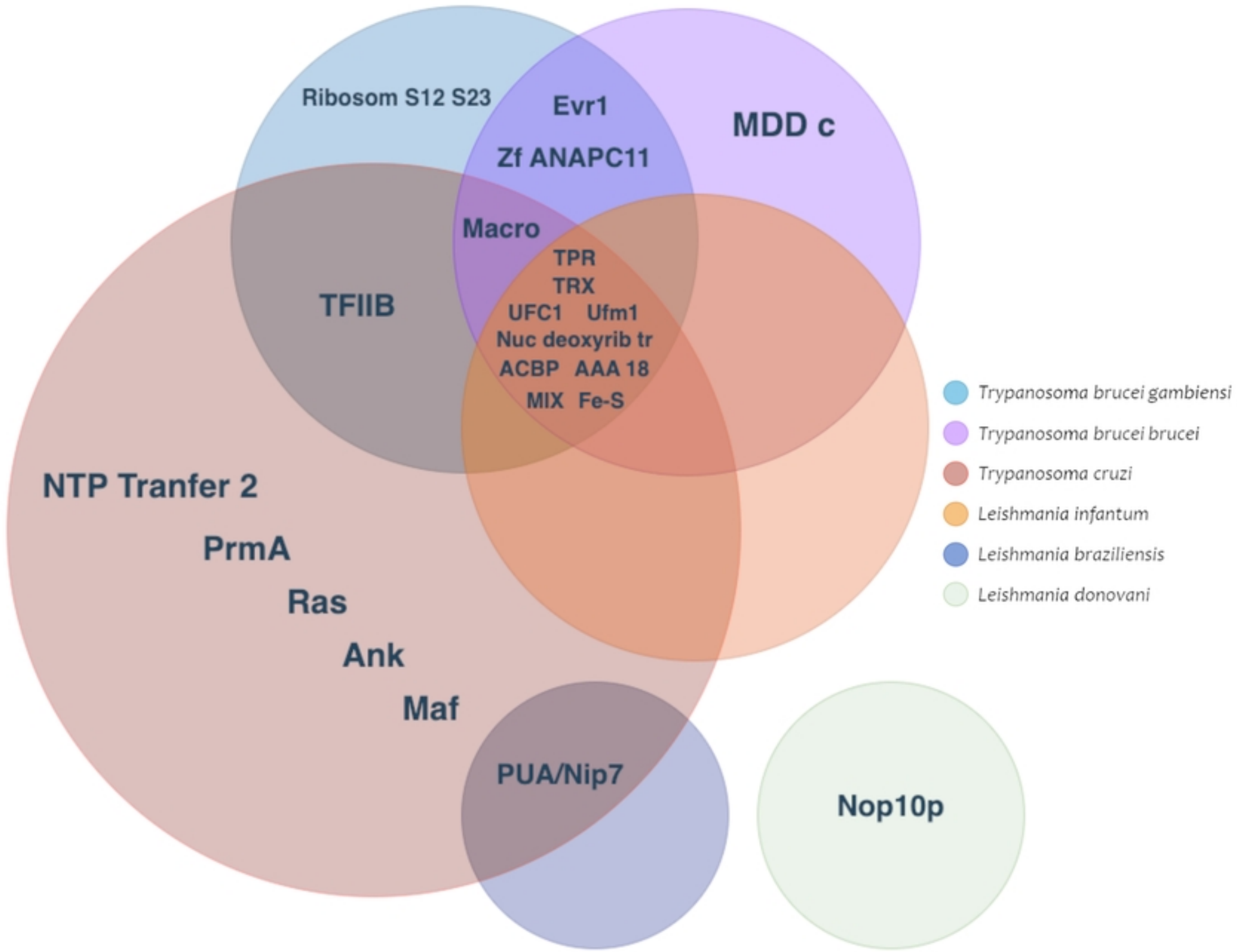
S4 Table. Oligomeric AlphaFold Models - Validation.



Manuscript Figure 1

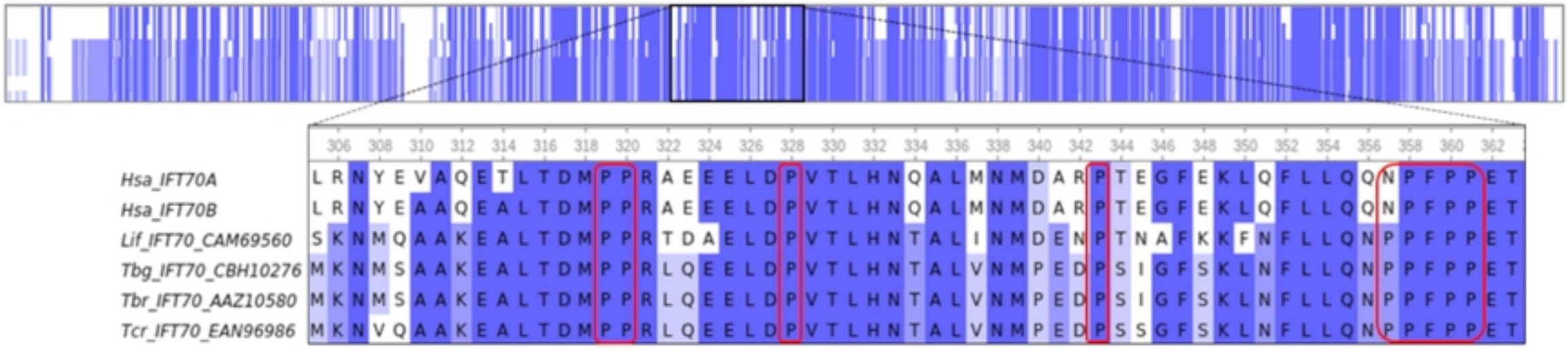


Manuscript Figure 2



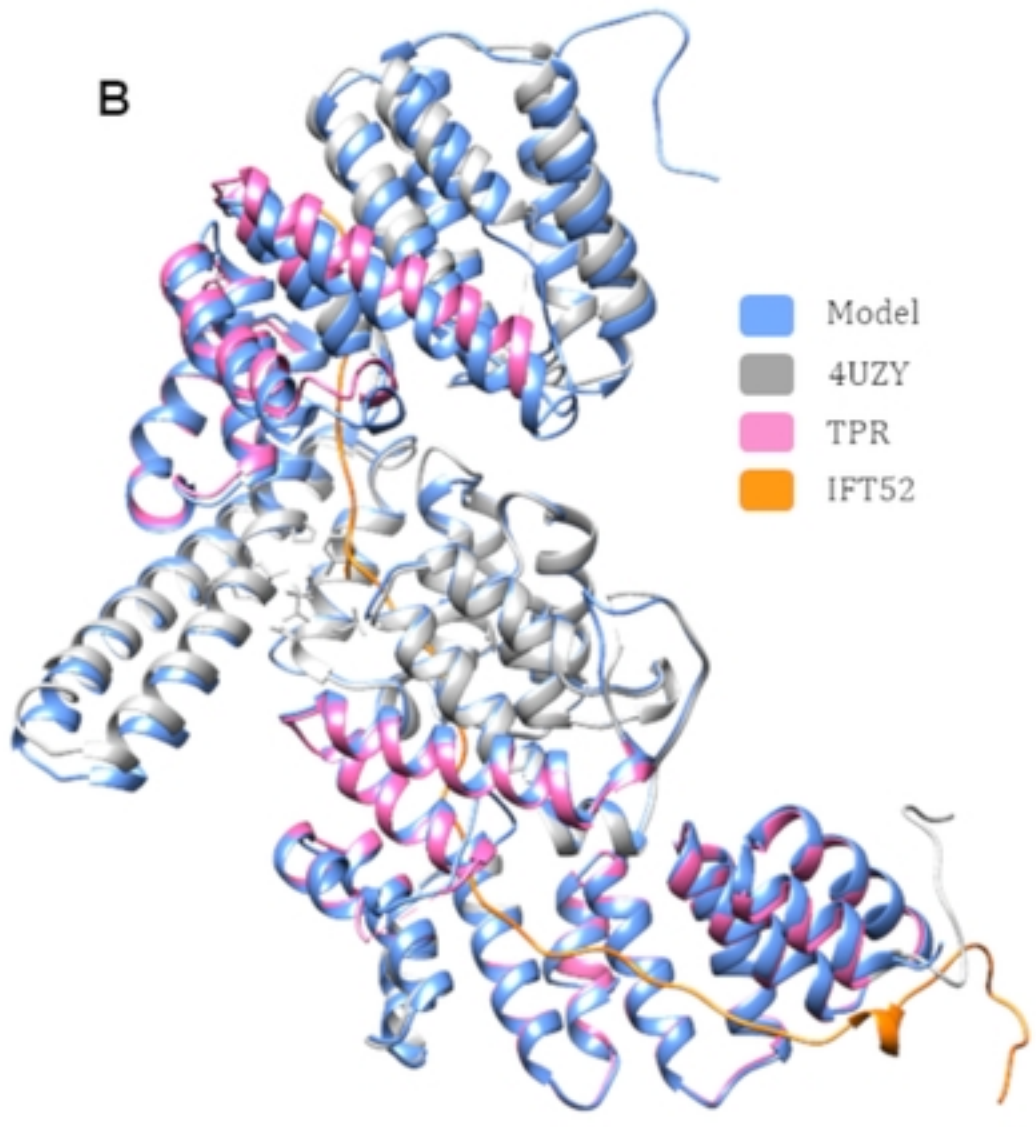
Manuscript Figure 3

A

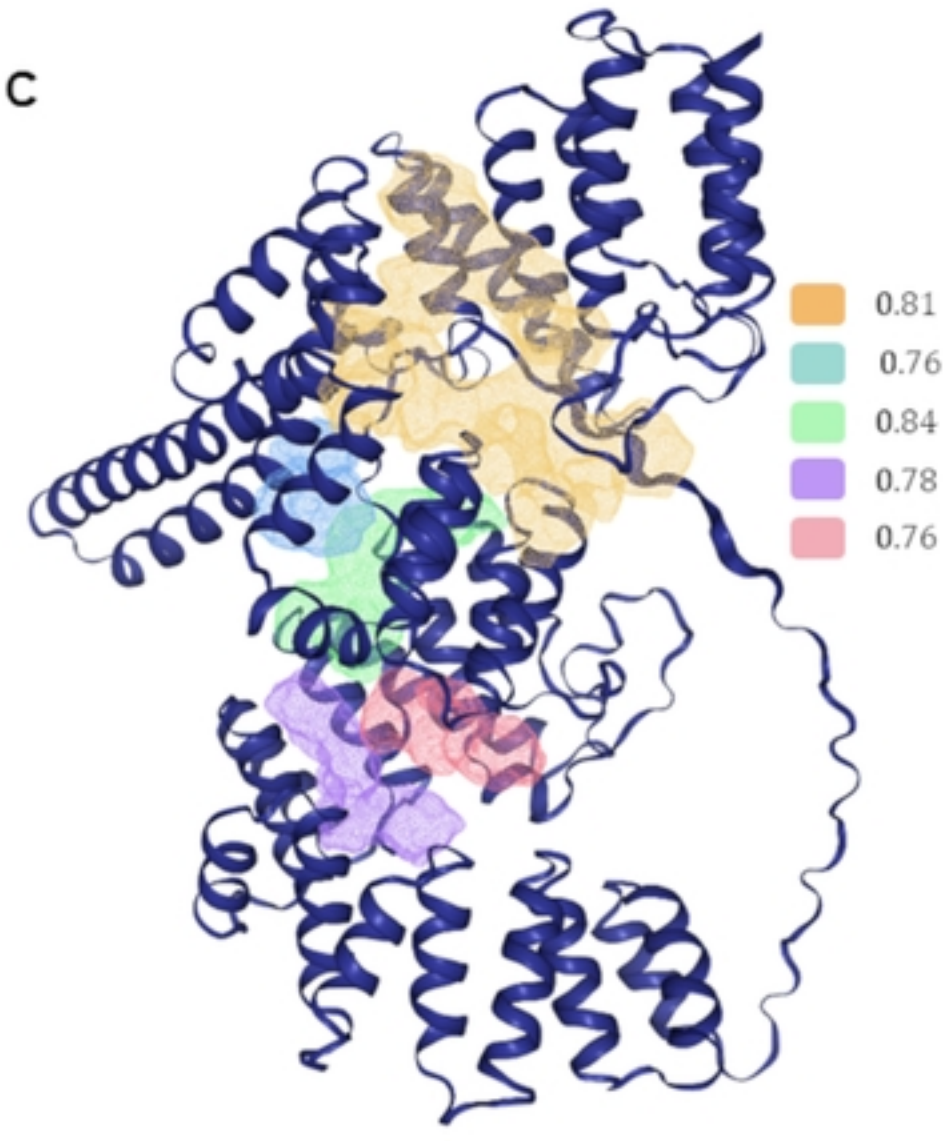


	Hsa_IFT70A	Hsa_IFT70B	Lif_IFT70_CAM69560	Tbg_IFT70_CBH10276	Tbr_IFT70_AAZ10580	Tcr_IFT70_EAN96986
Hsa_IFT70A	100%	96%	51%	52%	54%	51%
Hsa_IFT70B	96%	100%	52%	52%	54%	51%
Lif_IFT70_CAM69560	51%	52%	100%	74%	73%	75%
Tbg_IFT70_CBH10276	52%	52%	74%	100%	98%	90%
Tbr_IFT70_AAZ10580	54%	54%	73%	98%	100%	88%
Tcr_IFT70_EAN96986	51%	51%	75%	90%	88%	100%

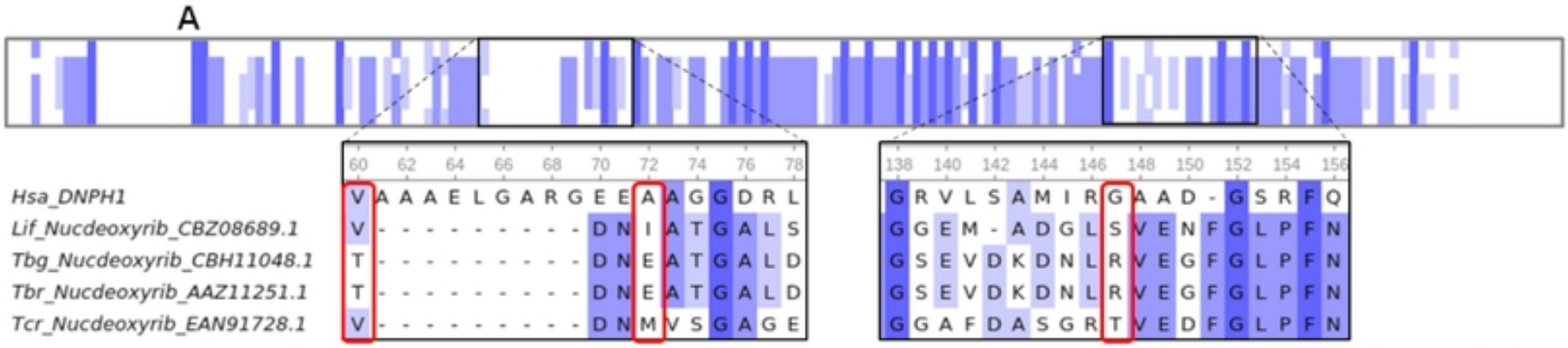
B



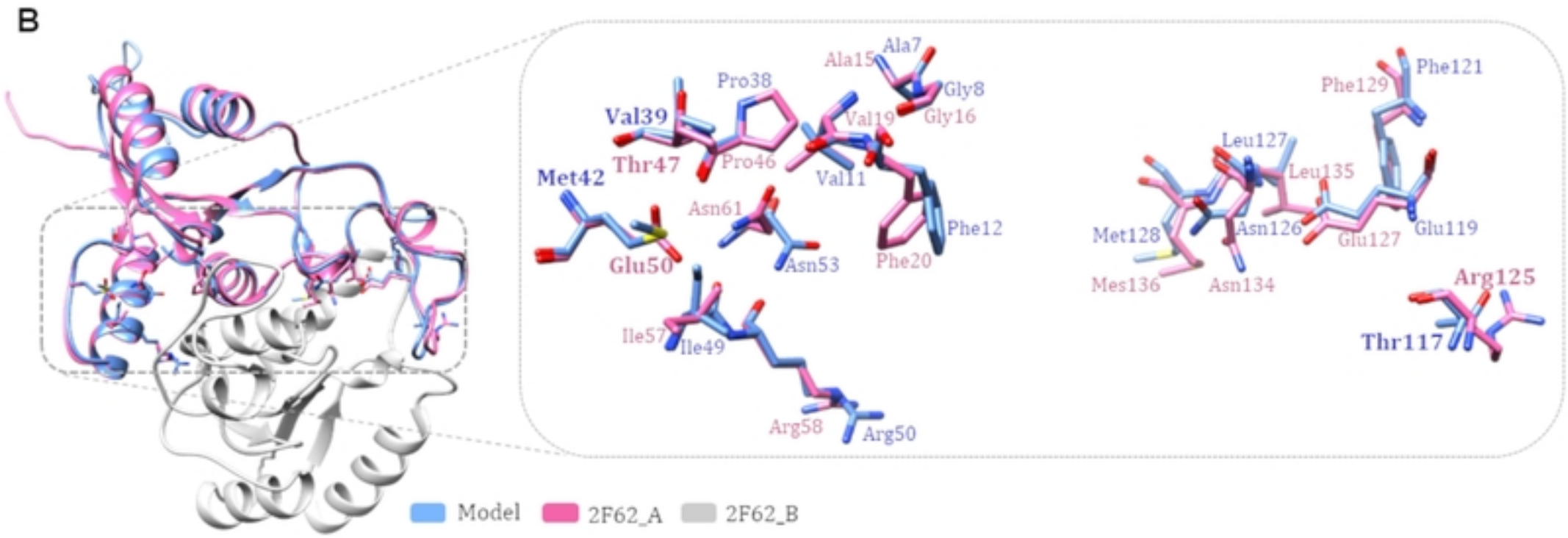
C

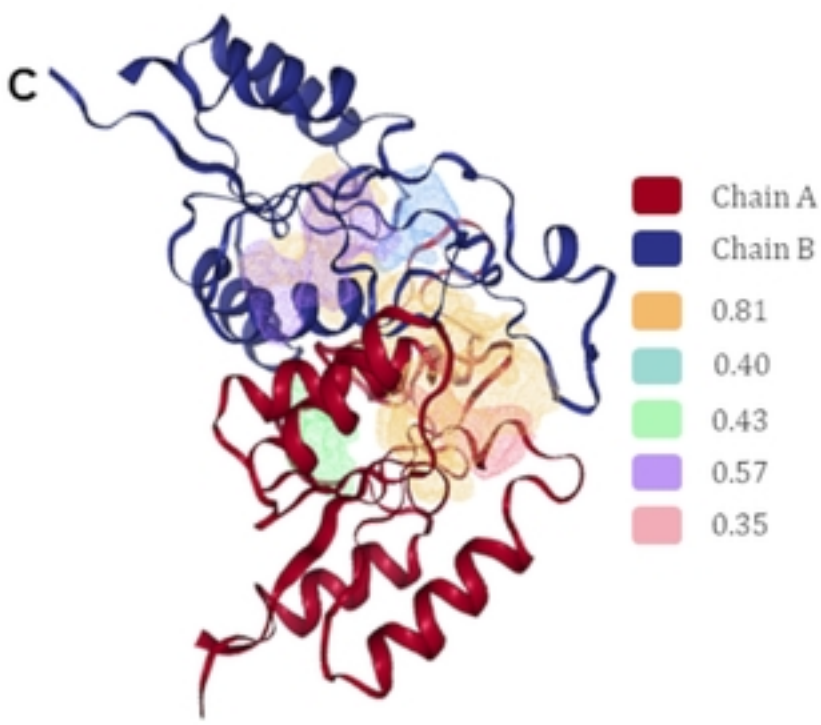


Manuscript Figure 4



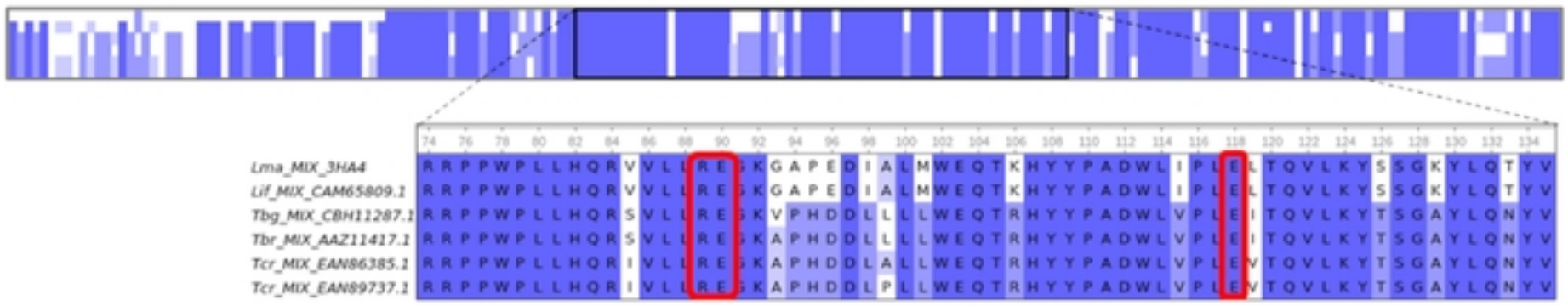
	Hsa_DNPH1	Lif_Nucdeoxyrib_CBZ08689.1	Tbg_Nucdeoxyrib_CBH11048.1	Tbr_Nucdeoxyrib_AAZ11251.1	Tcr_Nucdeoxyrib_EAN91728.1
Hsa_DNPH1	100%	16%	20%	20%	22%
Lif_Nucdeoxyrib_CBZ08689.1	16%	100%	72%	72%	65%
Tbg_Nucdeoxyrib_CBH11048.1	20%	72%	100%	100%	68%
Tbr_Nucdeoxyrib_AAZ11251.1	20%	72%	100%	100%	68%
Tcr_Nucdeoxyrib_EAN91728.1	22%	65%	68%	68%	100%





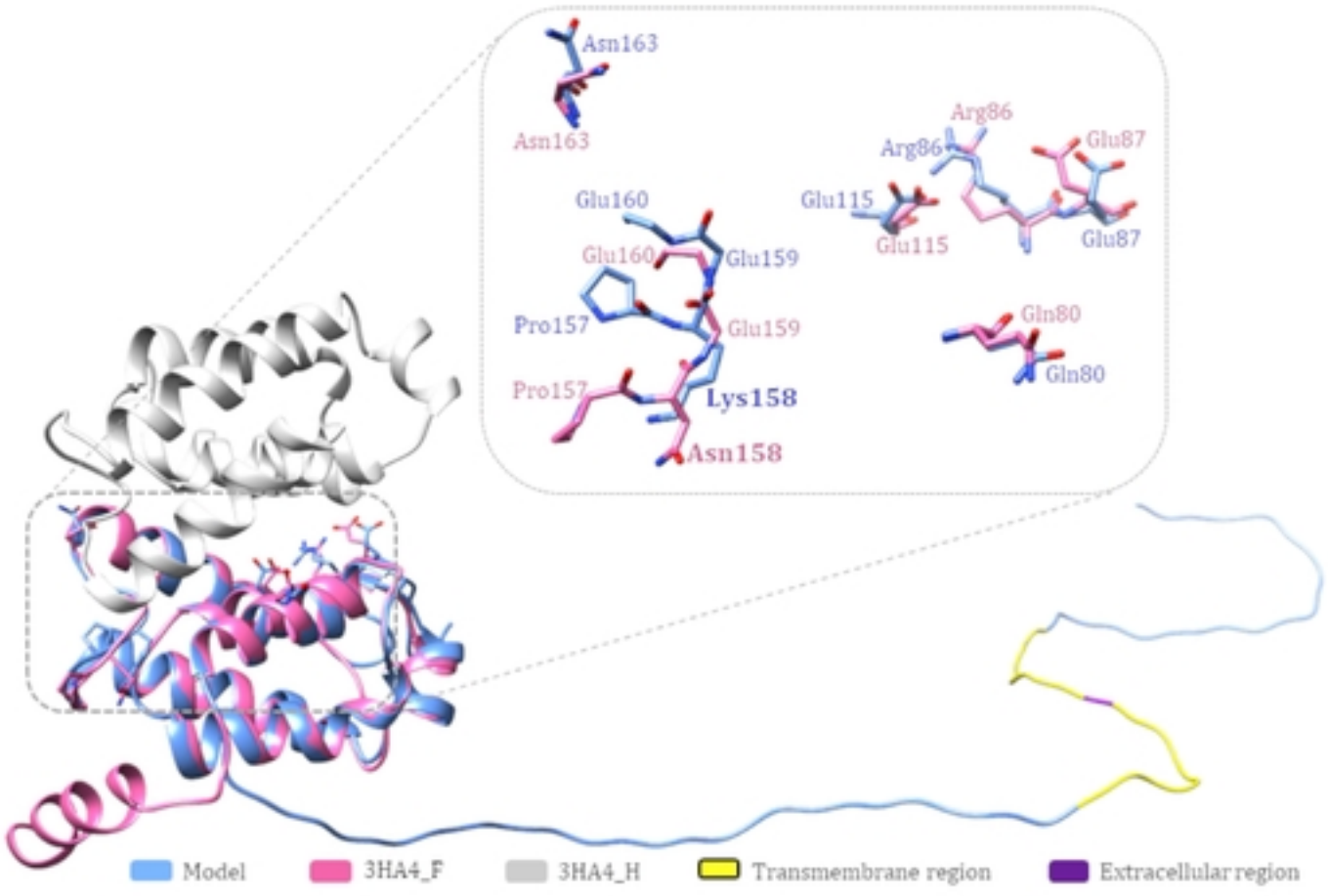
Manuscript Figure 5

A

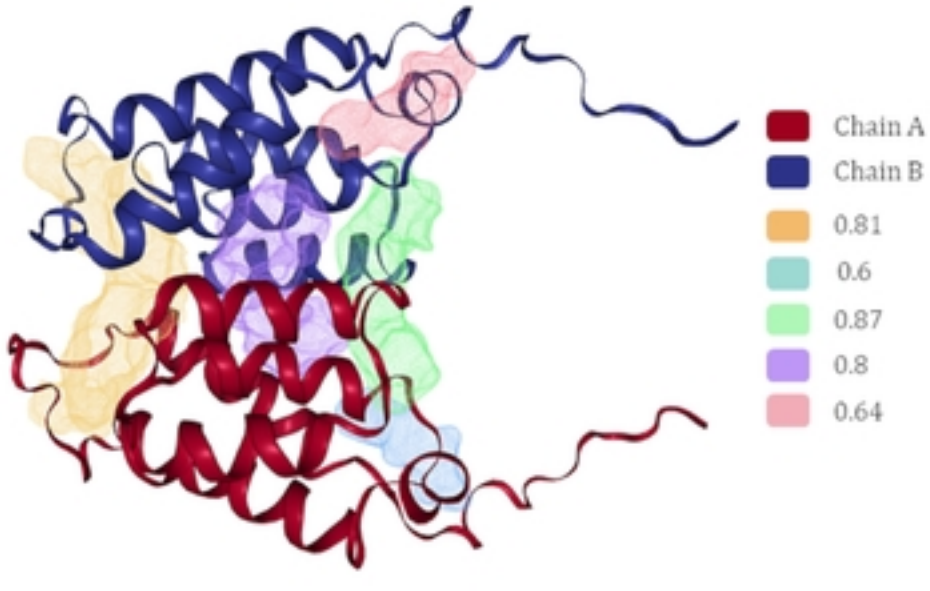


	Lma_MIX_3HA4	Lif_MIX_CAM65809.1	Tbg_MIX_CBH11287.1	Tbr_MIX_AAZ11417.1	Tcr_MIX_EAN86385.1	Tcr_MIX_EAN89737.1
Lma_MIX_3HA4	100%	77%	56%	56%	61%	60%
Lif_MIX_CAM65809.1	77%	100%	67%	67%	74%	73%
Tbg_MIX_CBH11287.1	56%	67%	100%	99%	83%	83%
Tbr_MIX_AAZ11417.1	56%	67%	99%	100%	83%	83%
Tcr_MIX_EAN86385.1	61%	74%	83%	83%	100%	99%
Tcr_MIX_EAN89737.1	60%	73%	83%	83%	99%	100%

B



C



Manuscript Figure 6