Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Jesús Alvarado Valverde

Born in: Mexico City, México

Oral examination: 20th March 2023

Computational prediction of Short Linear Motif candidates in the proteome of the Apicomplexan parasite *Toxoplasma gondii*

Referees: Prof. Dr. Detlev Arendt

Prof. Dr. Nina Papavasiliou

Acknowledgements

I would like to thank the European Molecular Biology Laboratory for offering me a unique and extraordinary place to not only perform exciting research and collaborative projects but also to develop multiple friendships and memorable life experiences. I am thankful to Dr. Toby J. Gibson for giving me the chance to undertake a Ph.D. project in his group and as his last student. I admire his knowledge, wit and wisdom on molecular biology, but also his scientific career, philosophy and principles. To all Gibson team members with who I shared multiple ideas and deep conversations during the last 4 years. To Hugo and Lena for welcoming me into the lab and for sharing their knowledge and experiences. To Balint, Laszlo and Renato for all the scientific discussions and feedback on my project. To Manjeet, Malvika and Mark for their friendliness and help. To Zsofia, Macarena and Nico for their support and enthusiasm during the last months, and to all the people that I meet through the Gibson team.

I would like to give special thanks and acknowledgements to the members of the Protein Expression and Purification facility for their contribution in my project and for their help in carrying out the experimental part of this thesis. To Dr. Karine Lapouge and Dr. Arne Boergel for guiding me through the process with the most openness and enthusiasm, and for their input in the writing of this thesis.

I want to thank my partner Hendrik for believing in me and for all his encouragement. I feel very lucky to have found you and for the things we live and shared every day. I am thankful to all the friends that have been there with me on this journey since the beginning. To Alberto and Gilberto for their cheerfulness and support, and for all the scientific, political and life discussions we had throughout these years. To Karolina, Anna, Lucia and David for including me in their lives and for listening to me. To Abiram, Sebastian, Ana, Aline, Carlos, Javier and Ned, for all the dinners, adventures and parties we have shared. To Agata, Kevin, Andrea, Matteo, Max, Lea, Maxime and all the friends that made the PhD a great scientific but also an exciting life experience.

I would also like to thank the EMBL Staff Association for all the teamwork, training and mentorship. To the Equality, Diversity and Inclusion committee for letting me be part of a forum of people dedicated to improving scientific environments and practices. To the Predoc reps with who I exchanged multiple ideas and important projects. Thank you all for the opportunity to work together to help our colleagues, as well as to improve and maintain EMBL as the great place it is.

Finalmente, le agradezco a mi familia por todo su apoyo, por creer en mí y ser mis más grandes admiradores. A mis padres por haber impulsado mi carrera y mis proyectos, por escucharme y darme su apoyo desde la distancia. A Emanuel y a Juan por su amistad incondicional. A los señores Luis y Rosy Comadurán por su sabiduría y consejos. A mi tía Laura Linares sin cuya confianza y generosidad no podría estar aquí. A mi abuela Eva, de quien no me pude despedir, por el amor y cariño que me mostró en vida. Y a todas aquellas personas de las que he aprendido sobre la vida.

Table of Contents

List of Figures

1. Introduction

2. Prediction pipeline

3. Motif match filter Development

4. Match candidates

5. Experimental validation of motifs

6. Discussion

Annex

List of Tables

2. Prediction pipeline

3. Match filtering

4. Match candidates

5. Experimental validation of motifs

List of Acronyms

A
AC: Accessibility confidence

B
BioID: Proximity-dependent Biotin Identification
BOAS: Beads-on-a-string structures

C
CAMLG: Calcineurin Activator Calcium-Modulating Ligand
CD40: Tumor necrosis factor protein CD40
CLV: Cleavage motif
COVID-19: Coronavirus disease 2019

D
DEG: Degradation motif, Degron
DOC: Docking motif
DSSP: Dictionary of Secondary Structure of Proteins

E
EBH: End binding homology domain
EGF: Epidermal Growth Factor
ELM: Eukaryotic Linear Motif resource

G
GO: Gene Ontology
GRA: Dense granule protein

H
HIV: Human immunodeficiency virus
hTRAF6: heterologous expressed TRAF6 domain
HyperLOPIT: Hyperplexed localization of organelle proteins by isotope tagging

I

IDP: Intrinsically disordered protein
IDR: Intrinsically disordered region
IRG: Immune related GTPases
IL-1R: interleukin-1 receptor
ITC: Isothermal titration calorimetry
IVN: Intra-vacuolar network

K

K_D: Dissociation constant

L

lDDT-Cα: Local distance difference test Cα
LIG: Ligand motif

M

MATH: Meprin And TRAF-Homology domain
MassSpec, MS: Mass Spectometry
MBP: Maltose binding protein
MIC: microneme protein
MISHIP: Microtubule and SH3 domain-interacting protein
MJ: Moving Junction
MOD: Modification motif
Motif candidate: a motif match with supportive information
Motif instance: a motif example validated experimentally
Motif match: a motif hit found using a REGEX model
MSA: Multiple Sequence Alignment
MST: Microscale thermophoresis
MT: Microtubule

N

NF-kB: Nuclear Factor kB
NFAT4: Nuclear Factor of Activated T cells 4

P

PDB: Protein Data Bank database
PEPCore: Protein expression and purification core facility
PEXEL: Plasmodium Export Element motif
PI: Isoelectric point
pLDDT: predicted local-distance difference test
PPI: protein-protein interaction

PSSM: Position specific scoring matrix
PTM: Post-translation modification
PV: Parasitophorous vacuole

R

REGEX: Regular expression
RLC: Relative Local Conservation
RON: Rhoptry neck protein
ROP: Rhoptry protein

S

SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
Seq_id: Sequence identifier
SLiM: Short Linear Motif
SP: Signal peptide

T

TJ: Tight Junction
TLN: Toxolysin protein
TM: Transmembrane
TNF: Tumor Necrosis Factor receptor protein
TRAF: TNF receptor-associated factor
TRG: Targeting, trafficking motif

U

UniProt: Universal Protein database
UPS: Ubiquitin proteasome system

V

VEuPathDB: Eukaryotic Pathogen, Vector and Host Informatics resource

+TIP: microtubule plus-end tracking protein

Abstract

Toxoplasma gondii is a unicellular parasite of the Apicomplexan family with the unique ability to infect a wide spectrum of warm-blooded animals, including mammals and birds. Host infection is established by distinct secreted proteins that interact with the cellular machinery and signaling networks of the host cells, hijacking their immune response and subverting cellular processes to their advantage. Short linear motifs (SLiMs) are small functional modules within protein sequences known to mediate protein-protein interaction between parasite and host proteins. By integrating SLiM information with sequences, structural, and experimental data I developed a computational pipeline to identify motif candidates relevant for *T. gondii* infection. Among these candidates, I identified motifs in microneme, rhoptry, and dense granule proteins that potentially link them to processes like cell attachment, nuclear targeting and cytoskeleton rearrangements. As a proof of concept, the protein-protein interaction of a group of motif candidates related to the innate immune response were tested experimentally in collaboration with the EMBL Protein expression and purification facility. This provided proof of binding and affinity measurements for some of them, and showed that the pipeline is able to identify true binding motifs. Taken together, I developed a computational pipeline that can potentially predict motif candidates relevant for *T. gondii* infection and provide a resource for further experimental validation and understanding of parasite infection.

Zusammenfassung

Toxoplasma gondii ist ein unizellulärer Parasit aus der Familie der Apicomplexa, welcher sich durch ein breites Infektionsspektrum auszeichnet, das unterschiedliche Warmblütler wie Säugetiere und Vögel einschließt. Im Rahmen der Infektion sezerniert der Erreger Proteine, welche die Stoffwechselprozesse sowie die Signalnetzwerke der Wirtszelle zugunsten der parasitären Vermehrung beeinträchtigen. Die Parasit-Wirt-Interaktion wird auf Proteinebene u.a. durch funktionale Module innerhalb der Proteinsequenz vermittelt, welche als sog. *short linear motifs* (SLiMs) bezeichnet werden. Durch die Integration von Informationen von SLiMs mit Daten zu Proteinsequenzen und -strukturen sowie Ergebnissen veröffentlichter experimenteller Daten (z.B. Massenspektrometrie) habe ich ein bioinformatisches Verfahren entwickelt, um Motiv-Kandidaten mit Relevanz für die Infektion mit *T. gondii* vorherzusagen. Unter diesen finden sich beispielsweise Motive in Mikronemen, Rhoptrien sowie Dichte-Granula Proteinen, welche potentiell mit Wirtsproteinen interagieren können mit Funktion in Zelladhäsion, Kerntransport und Zytoskelettumbau. Die vorhergesagten Protein-Protein-Interaktionen konnten ferner für Motiv-Kandidaten mit Bezug zur angeborenen Immunabwehr in Zusammenarbeit mit der EMBL *Protein Expression and Purification Core Facility* durch Affinitäts-Assays experimentell validiert werden. Zusammengefasst habe ich in der vorliegenden Arbeit ein bioinformatisches Verfahren zur Vorhersage von Motiv-Kandidaten entwickelt, als potentielle Grundlage für eine zeit- und kosteneffiziente experimentelle Untersuchung von Mechanismen der *T. gondii* Infektion.

CHAPTER 1
Introduction

1.1. Introduction

Despite the steady accumulation of genomic and proteomic data for unicellular eukaryotic parasites, the molecular mechanisms underlying their infection processes are still the subject of active research. A thorough understanding of these mechanisms is essential for the development of effective prevention approaches as well as pathogen specific treatments, an important goal due their role of parasites as major threat not only to humans but also for livestock.

An important concept to understand parasite infections is the so-called parasite-host interface, which is the entirety of interactions between a parasite and its host. While there is plenty of focus on microscopic and genetic research for parasites, there are steps that need to be covered to understand infection at a systemic level. Over the years, there has been research on some of these interfaces, mainly focusing experimental setups on a handful of essential proteins with key roles, but few efforts have been made to the collective effects of parasite molecules during infection. Although this approach allows a genotype-phenotype correlation after knocking out a subset of preselected genes and there are now high-throughput experiments to determine the generate high-throughput screens (e.g. CRISPR/Cas9 screens), this is not enough to determine how their different protein products interact with host components and contribute to infection. Secreted proteins have a unique role in interreacting with the host cell components as they repurpose their processes to their advantage. In this regard, they offer an interesting vantage point for investigating parasite-host interactions and to understand wider parasitic strategies

Protein-protein interactions via Short Linear Motifs (SLiMS) are commonly used by viruses and bacteria to infect their hosts and rewired their processes to escape immune recognition, sequester nutrients and multiply their numbers. Unlike these systems, there have been small efforts on determining the overall motif presence and their roles in eukaryotic parasite proteomes. With the current advancement in parasitology, there is now enough data to determine this protein-protein interaction landscape and start building a systems picture of their infection process.

This thesis delves into the *in-silico* prediction of short linear motif candidates in the secreted proteins of the widely distributed apicomplexan parasite *Toxoplasma gondii*. For this, I developed a discovery pipeline that takes advantage of publicly available data and some of the latest developments in protein structural determination (Chapter 2). I chose a set of filtering criteria to determine the best predictions (Chapter 3), to then explore different ways to select motif candidates for experimental validation (Chapter 4). In collaboration with the EMBL Protein expression and purification facility, we evaluated the binding potential of a set of motifs candidates in secreted proteins that are linked to the innate immune response (Chapter 5). Finally, I discuss the advantages and limitations, perks and advantages of these approaches, as well as their potential relevance for parasite infection research (Chapter 6).

1.2. Apicomplexan parasites

Apicomplexans are major agents of disease for humans and livestock

Apicomplexans are unicellular eukaryotes that live an obligate intracellular parasitic lifestyle and are major human and livestock agents of disease. The infectious disease with the highest number of parasite-related deaths in humans is malaria, caused by species of the apicomplexan parasite *Plasmodium* which is transmitted by *Anopheles* mosquitos. In 2020 *Plasmodium* species, particularly *P. falciparum*, caused over 241 million malaria cases in the world of which 627,000 ended up in death (Monroe et al., 2022). *Cryptosporidium*, another apicomplexan with a worldwide distribution, is a contributor to infant mortality by being a leading cause of diarrhea and malnutrition in children (Dhal et al., 2022). Different parasite species of the genera *Babesia* and *Theileria* are

transmitted by ticks and affect horses and cattle (Almazán et al., 2022), while species of *Eimeria*, the parasite causing Coccidiosis in birds, generates major economic losses in the poultry industry (Zaheer et al., 2022). *Toxoplasma gondii*, the parasite causing Toxoplasmosis, is believed to be present in one-third of the world population. Although it is not as deadly as some other Apicomplexans, its wide prevalence represents a major health burden as it can lead to more serious disease forms when present with other conditions like cancer and AIDS (Hakimi et al., 2017).

Apicomplexans are a diverse group of unicellular parasites

Beyond their roles as pathogens, Apicomplexans are a widespread and diverse family of protozoans in the environment, infecting different metazoan organisms – some of them being symbionts of corals. Currently, there are more than 5,000 apicomplexan species described. It is estimated that there is one apicomplexan per invertebrate species (both land and marine ones) with many others thought to be undiscovered yet (Mathur et al., 2021).

In the eukaryotic phylogenetic tree, Apicomplexans are placed within the *Myzozoa* group together with Dinoflagellates, their closest algal relatives. There they are included in the larger protozoan grouping of the Alveolates, sharing the characteristic alveolar sac network that supports their cellular outer membrane. Alveolates are then placed within the SAR supergroup together with Stramenopila and Rhizaria. The origin of Apicomplexans has been hypothesized to be related to a secondary event of endosymbiosis, where a protozoan organism engulfed a red algal cell which later develop into a remnant organelle termed the apicoplast (Janouškovec et al., 2010). Closely related organisms like the coral-associated algae *Chromera velia* and *Vitrella brassicaformis* still retain the photosynthetic capacity of this organelle. In Apicomplexans, the adoption of a parasitic lifestyle meant major losses of photosynthetic, ribosomal and other metabolic functions, while undergoing an expansion in specialized genes for infection and the interaction with their diverse hosts (Mathur et al., 2021; Woo et al., 2015).

Large families of Apicomplexa are the Gregarines, Cryptosporidians, Hematozoans and Coccidians **Figure 1.1.** They differ genetically and morphologically, but these groups shared features like life-style and type of hosts, some completing their life cycle within

one host (monoxenic), and others requiring two types of hosts to complete their life cycle (dixenic). Their diverse combination of lifestyles and host tropism make Apicomplexans fascinating organisms to study host-parasite interactions. Being unicellular also means that their morphological and physiological flexibility is regulated at the molecular level, so they offer an opportunity to understand these different strategies at this level.

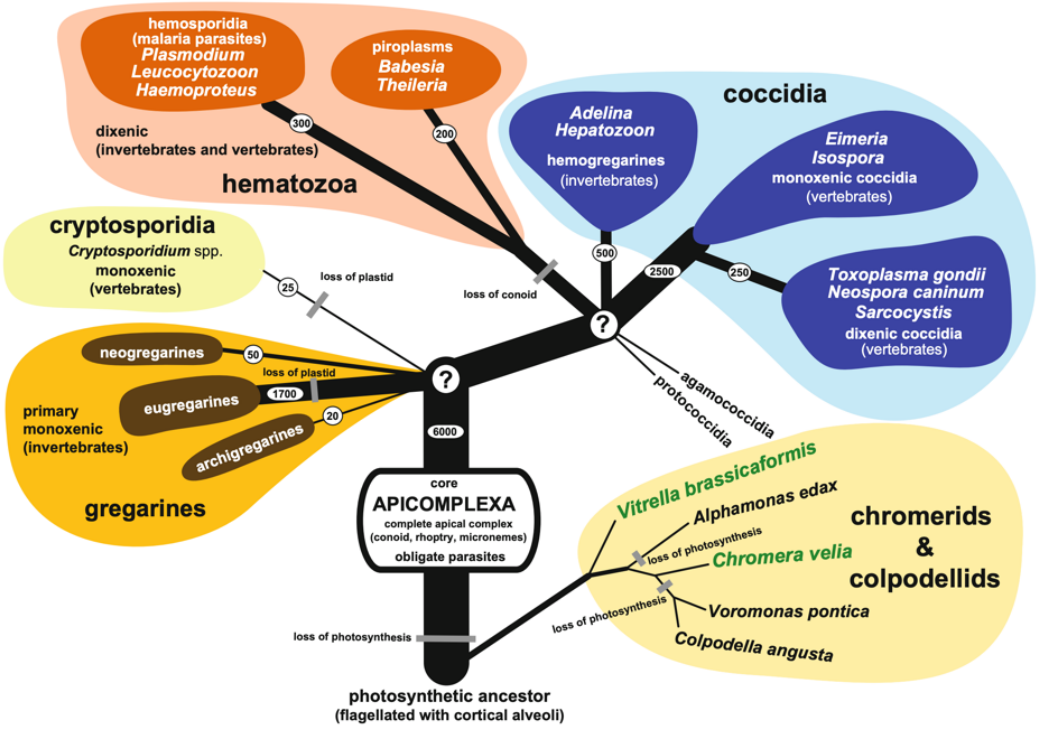


Figure 1.1 Apicomplexa tree showing its major phylogenetic groups. Major apicomplexan groups with representative organisms and lifestyle highlights (Votýpka et al., 2017).

1.3. *Toxoplasma gondii:* a successful parasite

Toxoplasma gondii is a successful Apicomplexa

Toxoplasma gondii (from the Greek 'toxo' arc, & 'plasmos' life) stands out as a successful Apicomplexan for being able to infect any nucleated cell from any warm-blooded animal, including land and marine mammals as well as birds. It was first discovered in 1908 during Nicolle and Manceaux studies on parasites in rodents. Since then, *T. gondii* has become a model for cell biology and apicomplexan organisms, not only because of its medical and veterinary relevance, but also due to the organism amenability for genetic manipulation (Dubey, 2021).

Toxoplasma gondii belongs to the Sarcocystidae clade

Toxoplasma gondii belongs to the apicomplexan family *Coccidae* and within it, to the cyst-forming *Sarcocystidae* clade, which is comprised of parasites with an obligatory two-host life cycle, in which the principal one is a carnivore (Dahlgren et al., 2008). The *Toxoplasma* genus shares a phylogenetical sister group with species of the genus *Hammondia* (*H. hammondi*), *Neospora* (*N. caninum*), *Besnoitia* (*B. besnoiti*) and *Cystoisospora* (*C. suis*) **Figure 1.2**. Unlike *Toxoplasma*, not all its closest relatives have been properly characterized as the diseases they cause are still understudied and their complete life cycles have not been completely elucidated or explored (Olias et al., 2011).

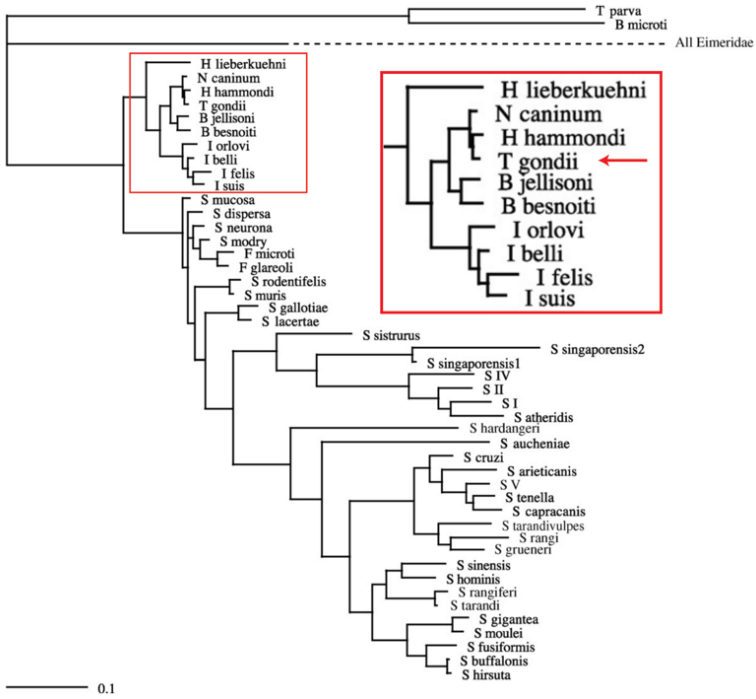


Figure 1.2 Sarcocystidae phylogeny showing *Toxoplasma* closest relatives. Phylogenetic tree created using all SSU rRNA (Small Subunit ribosomal RNA) gene sequences, adapted from (Dahlgren et al., 2008). Isospora genus has been renamed Cystoisospora.

While *Toxoplasma* only counts with one single species, numerous different strains have been isolated and studied. These strains vary according to their population geographical distribution and the pathology they cause to their hosts. Most of the strain isolates from human, domestic and wild animals come from a set of three clonal lineages: type I, II and III. Nevertheless, there are strains of *Toxoplasma* that show evidence of greater genetic recombination, e.g. in South America and Asia (Galal et al., 2022; Hakimi et al., 2017).

Toxoplasma gondii has a two-host life cycle with different cell forms

During their life cycle, *Toxoplasma* parasites thrive and multiply in an heteroxenous two-host cycle: with an intermediate host where they reproduce asexually and a definitive one where they reproduce sexually **Figure1.3**. During asexual replication, *Toxoplasma* cell division is characterized by the production of two daughter cells within the mother cell. This process is termed Endodyogeny and is the simplest form of multinuclear replication, with other Apicomplexans being able to produce more than two daughter cells from a single one (White & Suvorova, 2018). *Toxoplasma* goes through different cell forms during their asexual cycle: an infective one termed Tachyzoites and a chronic infective form termed Bradyzoite**.** Tachyzoites are the agents causing an acute infection going through cycles of replication and invasion, while Bradyzoites cause a chronic infection where they lay dormant as cysts in non-dividing cells like neurons and myocytes (English & Striepen, 2019).

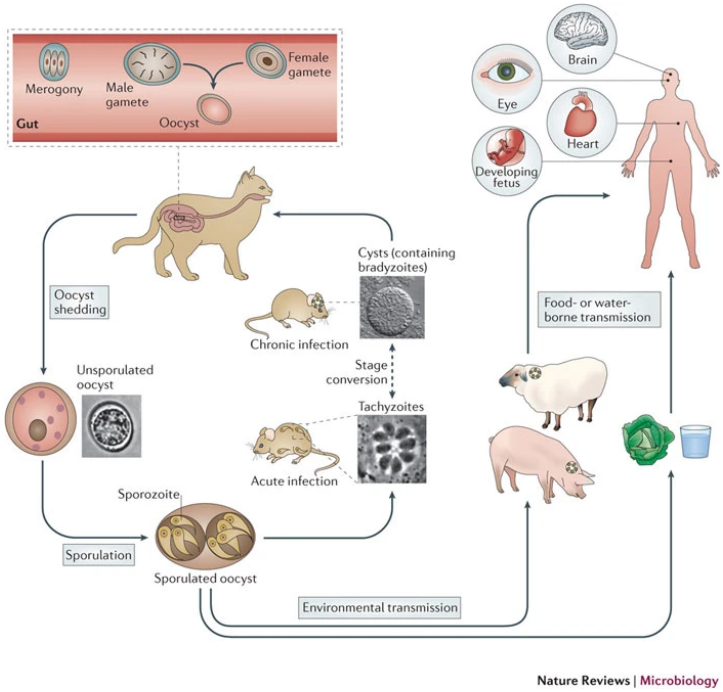


Figure 1.3 Summary of *Toxoplasma gondii* life cycle. (Hunter & Sibley, 2012)

All members of the *Felidae* family are definitive hosts, including domestic cats. *Toxoplasma* reproduces sexually in their gut epithelial cells by perceiving a systemic excess of linoleic acid, as all felines lack the delta-6-desaturase enzyme activity required for its metabolism. It is here that parasite cells will develop into further different cell forms: pre-sexual Merozoites forms and two sexual forms: microgametes, a male form

displaying flagellar motility towards the macrogametes, the female form which remains in epithelial cells. After fertilization, the zygote will turn into Oocysts that will be excreted into the environment and mature as infectious Sporozoites (English & Striepen, 2019; Martorelli Di Genova et al., 2019).

In a naive intermediate host, *Toxoplasma* infects the cells of different tissues and multiplies by asexual reproduction. When a given host mounts an immune response, *Toxoplasma* tachyzoites will then go into the Bradyzoite form remaining as cysts within muscle and nerve tissues that would be consumed by a predator, which then will become infected and Bradyzoites will become active again. In nature, *Toxoplasma* plays a role in predator-prey cycles between felines and their prey (e.g. rodents and birds). Small prey is susceptible to uncontrolled parasite multiplication and further tissue and systemic damage which may result in behavioral changes that increase their chances of being captured by their predator (Berdoy et al., 2000).

Humans are indirect hosts but still susceptible to infection

Humans are only indirect hosts getting infected by consuming food or water contaminated with oocysts or by eating undercooked or raw meat containing tissue cysts. It is estimated that one-third of the human population has been infected with *Toxoplasma* and has cysts within their cells (Montoya & Liesenfeld, 2004). Their immune system can control a *Toxoplasma* infection and it only causes health complications when they are immunocompromised, ending up in retinal damage or encephalitis, or during pregnancies, causing fetal malformation or miscarriage. *Toxoplasma* parasites from Type II strains are the most commonly associated with human infections while Type I are the most lethal to mice. It is worth mentioning that Humans display an immune response to *Toxoplasma* infection different to that of mice (Hakimi et al., 2017).

1.4. Infection process at the cellular and molecular level

Toxoplasma apical complex is required for motility and infection

Like all members of the SAR phylogenetic group, Apicomplexans have an inner membrane complex, an extra layer of membranous sacks underneath their plasma membrane. This makes them especially sturdy and resistant to cell breakage. However,

the name defining structure is the so-called apical complex, after which they are named. The apical complex in *Toxoplasma* is comprised of multicomponent structures organized around the conoid, a mobile set of fibers in a spiral arrangement. The conoid then associates with the preconoidal and polar rings at both of its ends, as well as with two intraconoid microtubules (MT). They work together to fulfill different cellular functions, from gliding motility to infection and reproduction. When apicomplexan cells divide the conoid and overall apical complex are the first structures to be formed in daughter cells, even before cell division. And when parts of the conoid are altered the parasite can no longer move properly or even infect cells (Hu et al., 2006). The conoid also serves as the base for microtubules radiating from it and through the cell towards its distal part, giving the parasite its shape (Dos Santos Pacheco et al., 2022).

Toxoplasma apical complex contains organelles for infection

Toxoplasma secretes proteins in a sequential fashion to infect host cells. The secretion of these proteins is mainly organized at the apical complex through the conoid and a series of specialized secretory organelles: the micronemes, rhoptries and dense granules **Figure 1.4**. Micronemes are membrane-bound organelles which contain proteins involved in host cell attachment, most of these proteins, but not all, are termed MICs (Rastogi et al., 2019). Rhoptries are membrane-bound organelles that contain proteins involved in protein secretion and host cell invasion. They have a bulbous shape with a thin neck and a globular part. The number and shape of rhoptries vary among the Apicomplexa, *Toxoplasma* tachyzoites normally count with 12 rhoptries (Boothroyd & Dubremetz, 2008). The proteins localized in the rhoptry neck are usually refer as RONs and the ones contained in the bulb globular part ROPs. Once they are inside their host, *Toxoplasma* cells continue to secrete proteins from membraneless organelles distributed in the cytosol called the Dense Granules. Proteins from these organelles are usually involved in niche establishment and signaling disruption, most proteins from this organelle are termed GRAs (Rastogi et al., 2019).

When approaching a potential host cell, microneme proteins are secreted into the parasite cell outer membrane surface, in an intracellular calcium-dependent manner. MICs then interact with host cell surface receptors and proteins to ensure cell attachment. Upon detecting attachment, *Toxoplasma* injects the contents of its rhoptries into the host

cytoplasm. This is achieved by the formation of a rosette pore that creates a parasite-host channel. Then it has been shown that structures called apical vesicles participate in the discrete coupling of the channel with a rhoptry and coordinate the secretion of their contents (Sparvoli et al., 2022).

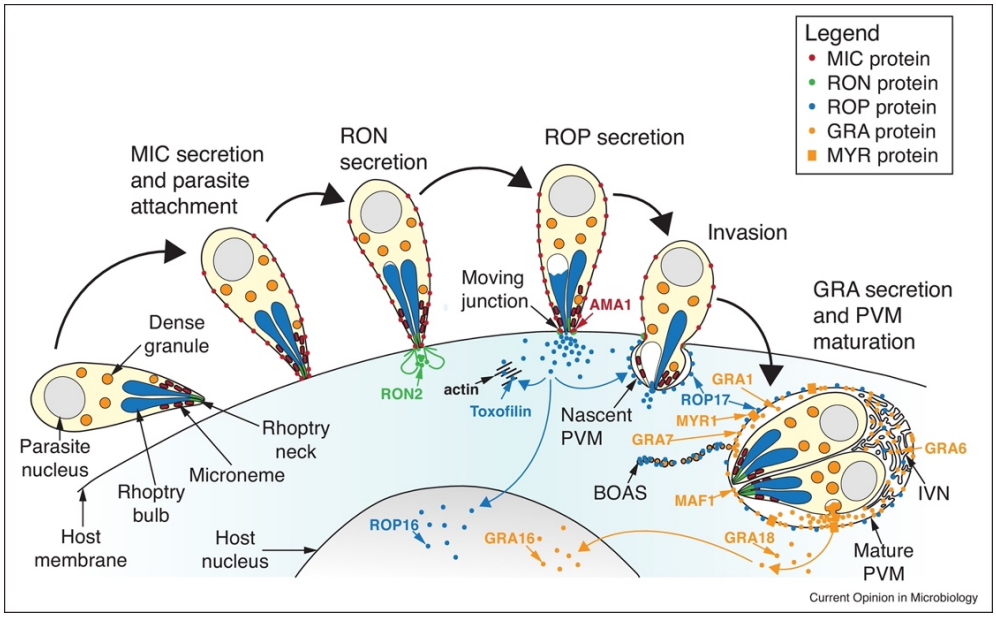


Figure 1.4 Summary of Toxoplasma gondii organelles and invasion cycle
(Rastogi et al., 2019)

Once attached and connected, the secreted MIC and RON proteins are involved in creating the Moving Junction (MJ) (also tight junction, TJ), a protein complex that will physically link the *Toxoplasma* cell with the host membrane. *Toxoplasma* is able to get a better grip through the MJ to then push itself inwards via an actin-myosin motor system creating an invagination into the host cell. After pulling the MJ to its posterior end, the invagination is closed giving rise to the so-called Parasitophorous Vacuole (PV). In this enclosed protective environment, the parasite will reside, grow and reproduce separated from the host cytoplasm as host lysosomes and endosomes are unable to fuse with it (Hakimi et al., 2017).

Within the host cell, *Toxoplasma* continues to secrete proteins contained in the Dense Granules. These proteins are now thought to be secreted from annuli, dedicated pores around the apical end of the parasite cell (Koreny et al., 2022). The proteins can be further secreted into the PV to remodel it or until to the host cell, through an undetermined translocon, to interfere with host processes. The PV remodeling includes the

development of membrane invaginations that will give rise to a PV network. This network increases the internal surface area of the vacuole, driving molecules passively from the host cytosol into its canals, i.e., it turns the vacuole into a sponge within the host. Further dense granule proteins decorate the PV protecting it from host-targeted degradation and connecting different PVs from different *Toxoplasma* infections. Dense granule proteins are also exported to the host nucleus where they interfere with transcriptional programs related to the host immune response (Hakimi et al., 2017).

Toxoplasma effectors interact with host proteins and rewire cell signaling

As could be imagined, during all steps in the *Toxoplasma gondii* infection, secreted proteins are involved in associating the parasite with host structures and processes. MICs interact with human receptors and membrane proteins, ROPs and RONs with the host membrane and cytoskeleton, while GRAs interact with host kinases and transcription factors. Over the years there has been substantial research on host-parasite protein-protein interactions, identifying several secreted proteins and their role at each step, but the nature and structural details of most of such interaction interfaces are still the subject of active research.

1.5. Protein-protein interactions

Protein structure is defined by its sequence

A protein is able to achieve its native structure spontaneously through the physicochemical composition and interaction of its amino acid residues. This means that the amino acid sequence contains all the necessary information for a protein to fold. Protein structure can then be categorized in different levels. The primary level is determined by the sequential order of its amino acids. The secondary level is formed by the intramolecular interactions among amino acids, these give rise to the so-called beta-strands and alpha-helices. The tertiary level is formed by the arrangement of secondary structure patterns into defined independent folds or globular folded domains. Finally, the quaternary level is comprised of the interaction of proteins with defined folds into bigger complexes and structures (Lesk, 2010). Overall, the function of a protein will be dependent on any of these structural levels and the interactions between them.

Unstructured proteins can have different cellular functions

Protein disorder is defined as the lack of a stable 3D structure under physiological conditions given the intrinsic properties of the amino acid sequence of a protein. Intrinsically disordered proteins are enriched in charged (R, K, E, D) and structure disruptive residues, e.g. prolines are too rigid to promote secondary structure, while glycine's are too flexible. On the other hand, disordered proteins are usually depleted from hydrophobic amino acids. This composition prevents them to forming stable secondary or tertiary structures and allows them to have greater flexibility as well as vastly greater surface area. Depending on the extent of the polypeptide regions we can classify them as intrinsically disordered regions (IDR), or as completely Intrinsically disordered proteins (IDP) (Habchi et al., 2014).

Disorder in protein can have several cellular functions. IDR can be linkers between ordered domains, and be enriched in post-translational modifications (PTMs) that serve as signals for further protein interactions or change the physicochemical properties of the protein. IDPs can have functions more related to their emergent physical properties. By their mere size, they can create spatial constrictions, e.g., by having a large conformational space they can be employed to sense the curvature of membranes (Quaglia et al., 2022). By binding to different proteins, they serve as scaffolds for forming larger multiprotein complexes. These complexes can accumulate and further form condensates with different properties from the cytosol. These biological condensates are now being characterized and their roles in biology are being defined. From forming membrane-less organelles to aggregates of proteins in response to environmental queues (Bratek-Skicki et al., 2020).

Proteins act collectively to carry out their functions

Proteins display a wide range of functions within organisms, from forming rigid structures that give cells their shape, the transport and enzymatic transformation of molecules, to the processing of complex environmental and internal signals. While performing their functions, protein rarely act alone. Proteins work in groups and interact with other macromolecules (carbohydrates, lipids, nucleic acids) and with other proteins to accomplish their functions.

Protein-protein interactions take place between different protein modules

The interactions between different proteins are regulated and mediated by different physicochemical properties (e.g. hydrophobicity, charges). The interaction could happen at any of the aforementioned levels. A domain interacts with a secondary structure or primary interface. There are also multicomplex interactions, domain-domain interactions. Protein domain-domain interactions are mediated by the physicochemical properties of their tertiary structure. They are the most well-studied as their stability allowed for them to be crystallized and their structure determined. Domains can also interact with smaller molecules such as peptides, carbohydrates and lipids, but also with small functional regions within proteins. Because of their intrinsic accessibility, IDPs and IDRs have the potential to interact with numerous and different binding partners, mainly through the presence of small binding interfaces termed motifs (Wright & Dyson, 2015).

1.6. Short linear motifs

SLiMs are dynamic modules for protein-protein interactions

Short Linear Motifs (SLiMs) are small functional modules within proteins characterized by different properties. They are peptide regions typically comprised of 5 to 10 consecutive amino acids. They are mostly found in IDPs but some might also exist in the ordered regions of proteins. SLiMs tend to lack a secondary structure when unbound, but many acquire one by induce fit upon binding. Being in disordered regions means that they are readily accessible for interaction and even though IDR sequence similarity is usually not strongly conserved among homologous proteins, the motifs are often relatively more conserved than their surroundings. Despite this, motifs can also display fast evolutionary dynamics. As disordered regions are free from the high selective pressures that maintain stable protein structures, they display a higher evolutionary rate that allows them to move along the protein sequence and explore more residue combinations, thus it is possible to converge into the short amino acids combinations of motifs. (From here onwards Motif is the word I used for SLiMs unless otherwise specified.)

SLiMs tend to have low affinity for their binding partners, because of their short nature and because of their particular amino acid composition. This low affinity allows them to engage in transient and reversible interactions. These fast and short-lived contact events

are necessary for the cell to have dynamic signaling events (Davey, Van Roey, et al., 2012). Motifs also work cooperatively; proteins could contain multiple copies of the same motif allowing for multiple binding events even if their affinity is low. Proteins could have multiple different motifs allowing them to bind multiple protein partners and to have the potential of becoming scaffolds for multiprotein complexes. It has been estimated that there are millions of PTM sites and hundreds of thousands of binding motifs to discover only in the human proteome (Tompa et al., 2014). This, together with the interaction multiplicity of proteins means that motif-domain interactions increase the overall protein-protein interaction space, adding more complexity and robustness to cellular systems (Kitano, 2004).

SLiMs have many different roles in the cell

In the cell, motif-based interactions play important roles in different biological processes. They serve as signal transduction modules, in the formation of protein complexes, in the post-translational processing of proteins, in protein abundance homeostasis and their sorting across different organelles and cell compartments. Based on their function, motifs can be classified into 6 types. Ligand (LIG) motifs are simple binding modules that facilitate the formation of larger protein complexes. Docking (DOC) motifs serve as recruiting modules that allow proteins to be recognized by enzymes and then be modified in a different site. Targeting and trafficking (TRG) motifs regulate the localization of proteins to a defined subcellular location by interacting with specific protein mediators and transporters. Degron (DEG) motifs are a specific subset of docking motifs that control the degradation of proteins by recruiting ubiquitin ligases to polyubiquitinate them and then target them for proteasomal degradation. Modification (MOD) motifs serve as the recognition modules for the addition or removal of post-translational modification biochemical groups. And finally, Cleavage (CLV) motifs are the modules recognized by proteases to carry proteolytic cleavage (Tompa et al., 2014).

Motif discovery and testing

Motifs have been discovered by research on PPIs and by using a variety of experimental methods. For example, they appeared to researchers as similar amino acid regions in sequence alignments of unrelated proteins that interact with the same protein or that were exported to the same cellular compartment. Motifs have been characterized by site-

directed mutagenesis, where segments of the protein are deleted to test its binding ability to their binding proteins. They repeat this process to narrow down the minimal protein regions necessary, thus finding the respective motifs. Once identified, researchers do single residue mutations to find the ones that abrogate the interactions and then infer the physicochemical interactions between these key amino acids and their binding domains. Motifs are commonly named using the one letter code of these key residues. These principles have been refined through structural research when small peptides of the protein containing the motifs are crystalized together with the binding domain. Another way to characterize motifs is through peptide-binding arrays, in which the relative amino acid preferences at each position of a motif is defined. Motifs can also be predicted computationally when knowing their amino acid composition. As they are short and their physicochemical principles might be known, we can create simpler models that capture them. Bioinformatic approaches are also helpful when deciding which motifs to test experimentally and how to test them, by combining as much data as possible to decide the most cost and effort-effective way (Gibson et al., 2015).

1.7. Host-parasite interface and motif hijacking

Viruses use motifs to hijack cell processes

There has been increasing research dedicated to defining the role of motifs during viral and bacterial infections. Viruses need to interfere with host cell processes to make them amenable to infection as soon as they are invaded so interference with PPIs is a common target. This is required for different purposes: from viral entry, controlling the degradation of its proteins, rewiring cell signaling to modulate host immune responses, recruit the membrane remodeling machinery to egress from the cell. Viruses can even bind to host domains outcompeting their binding partners by evolving similar motifs with higher affinity in a process called Motif mimicry (Davey et al., 2011). Examples of viral mimicry include the use of the PTAP motif for HIV virion budding. The HIV-1 gag protein contains a PTAP motif that binds to the TSG101 protein that recruits the ESCRT machinery to complete membrane abscission and release virion particles (Bieniasz, 2006). There are also motif instances in the proteins of the SARS-CoV-2, the pathogen causing Coronavirus disease 2019 (COVID-19). These identified motifs link the viral

proteins to cell attachment via integrins, autophagy and receptor mediated endocytosis (Mészáros et al., 2021)

Bacteria use motif mimicry to infect humans

Some bacterial groups and species also coexist within host environments and cells. They too interact with host cellular structures, and many species like *Chlamydophila caviae* and *Rickettsia rickettsia* even proliferate intracellularly. Some bacteria can inject proteins into their host via specialized secretion systems. These effector proteins then disrupt cellular processes enzymatically or through PPIs to exploit host resources. Part of these strategies also includes motif mimicry. As with viruses, bacteria interfere with their host during different infection stages. By using motif mimics bacterial proteins are able to interact with host receptors, undergo different post-translation modifications, remodel actin filaments, interfere with kinase signaling networks and target proteins to specific organelles (Sámano-Sánchez & Gibson, 2020).

Apicomplexans use motifs

As Apicomplexans are Eukaryotes it is expected that they use different types of motifs for their own cell biological processes but there is also evidence that they have their unique motifs and that they also use motif mimicry. An example of the former is the PEXEL (*Plasmodium* export element) motif used by species of *Plasmodium* parasites to export proteins from their endoplasmic reticulum into the cytoplasm of red blood cells. An example of the latter is the PxpxPR and SxIP motifs present in the parasite *Theileria annulata* protein MISHIP (microtubule and SH3 domain-interacting protein), to associated its cells with microtubules of the mitotic spindle to ensure daughter cell distribution within host during replication (Huber et al., 2018).

Toxoplasma use motifs during infection

A form of the PEXEL motif has also been reported for *Toxoplasma*, and even though this motif does not guarantee that the proteins will be exported into the host cytoplasm, it still plays a role in their maturation processing (Coffey et al., 2015). Work by (Guérin et al., 2017) showed that cytoskeletal-related motifs are involved in the cell entry process of *Toxoplasma* into host cells. They identified that rhoptry proteins from the MJ can engage with the host cytoskeleton through short linear motifs. The PxpxPR motifs

present in RON2, RON4 and RON5 are able to bind to the SH3 domain of CIN85, an adaptor protein involved in endocytosis, vesicle trafficking and cytoskeletal rearrangement **Figure 1.5** (Guérin et al., 2017). The role of the dense granule protein GRA14 in host cytosol material sequestration has been linked to the presence of two ESCRT system motifs, PTAP and LYPxL, in its C terminus (Rivera-Cuevas et al., 2021). Furthermore, the *Toxoplasma* proteome has a higher proportion of IDPs than other organisms **Figure 1.6** (Pancsa & Tompa, 2012). Altogether, this indicates that there is a high probability to find more motif examples in *Toxoplasma* that would be involved in the different stages of its life cycle, and could potentially explain its success in adapting and infecting a wide variety of hosts and cells types.

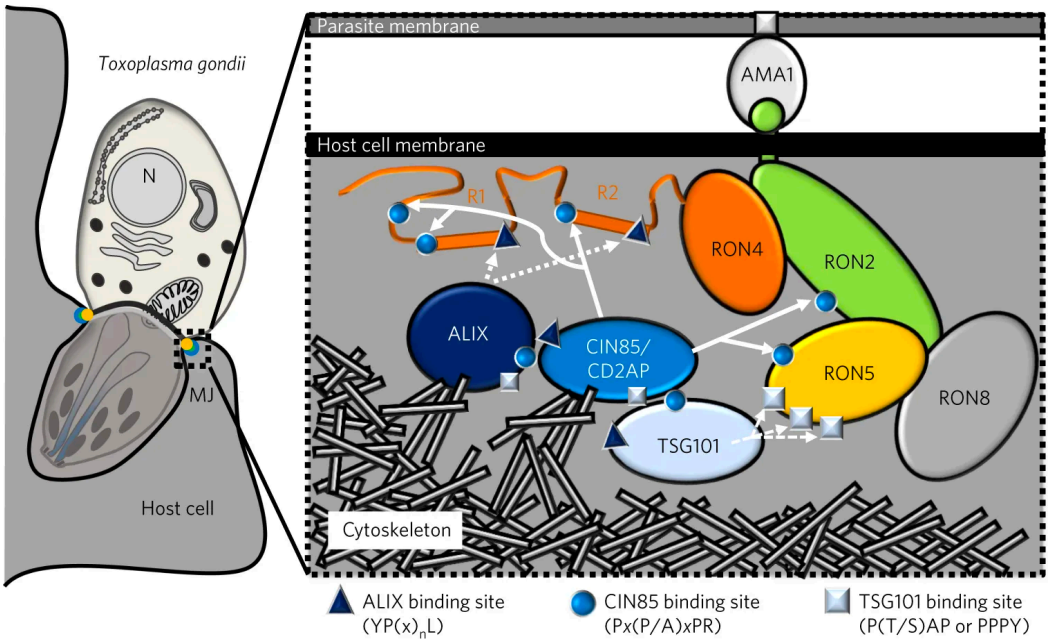


Figure 1.5 Moving junction proteins in *T. gondii* invagination process
(Guérin et al., 2017)

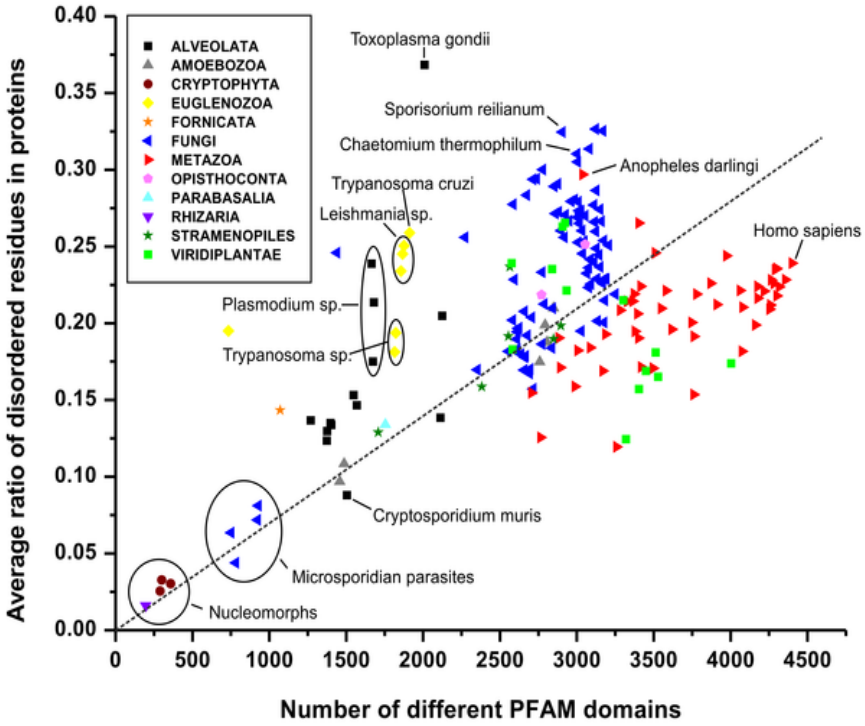


Figure 1.6 Disorder distribution among different organisms. The average ratio of disorder residues in proteins is plotted against the number of PFAM domains in an organism. *Toxoplasma gondii* stands out as an outlier with higher disorder levels of protein residues (Pancsa & Tompa, 2012).

Hypothesis & Aims

The aims of the study are (1) to establish a computational pipeline for an unbiased and systematic *in silico* prediction for true instances of short linear motifs in secreted proteins of *Toxoplasma gondii.* These predictions motifs will then be used (2) to determine groups of candidates based on their biological context and their potential binding domains in host proteins, and (3) to validate them experimentally using binding assays. Despite the possible application of this approach for time and cost-efficient parasite research *in vitro* and *in vivo* this might also help to answer the question, whether and to what extend *Toxoplasma gondii* uses short linear motifs in secreted proteins as an efficient way to adapt to a wide variety of different host and tissues

CHAPTER 2
Prediction Pipeline

A pipeline for motif discovery should be integrative and flexible

A pipeline for motif discovery can be built by combining motif models that predict sequence matches together with structural and conservation analyses and by integrating different publicly available data. Even though pipelines of this type have been applied to discover motifs in different organisms, from viruses to bacteria and animals, fewer efforts have been applied for specific eukaryotic parasites or apicomplexans. The development of such a pipeline for parasites poses challenges arising from the specific nature of such organisms, like highly derived genomes and unique proteins, from the quality and quantity of data, like low-quality sequencing or few related species characterized, and also from intrinsic complexities related to the diverse nature of motifs, some requiring post translational modifications or functioning in specific cellular contexts. In order to overcome these barriers, the construction of the pipeline should allow for the addition of data as it becomes available while allowing the flexible use of queries to select relevant candidates for further experimental testing. In the following sections I described how motif models can be used for motif discovery in the proteome of *Toxoplasma gondii*, and how to arrange a computational pipeline that effectively integrates structural, evolutionary and experimental data necessary to support the functionality of the motifs discovered.

2.1. Motif Models

Motifs can be represented by sequence pattern models

Motifs can be predicted computationally by examining the amino acid sequence of proteins. Given the short and linear nature of SLiMs their amino acid composition can be

captured by motif models. Motif models are usually represented by regular expressions (REGEX) or Position specific scoring matrices (PSSM). REGEX are character patterns used for queries within strings, they are widely employed when parsing text and in coding. PSSMs are probabilistic models that represent the frequency of specific characters in every position of a sequence pattern (Krystkowiak et al., 2018; O'Shea et al., 2013). Both types of models can be used to define and represent motifs based on different amounts of information. Surveys through any of these models have advantages and disadvantages so their application should be adapted to the objectives of the motif queries.

REGEX aim to capture the key residue composition of motifs aided by structural and conservation information. They can be easily created using a set of sequence examples. PSSMs rely on a larger set of high-quality examples to capture information from the specific motif composition with higher statistical confidence. Furthermore, motif searches using PSSMs give matches that can be ranked and classified based on the similarity to the reference model, while searches using REGEX models only give us exact pattern matches with equal values between them. Overall, REGEX are easier to create and implement but PSSM specificity makes them powerful tools to find relevant matches and discover new motifs (Krystkowiak et al., 2018). The lack of examples for many types of motifs makes the usage of PSSMs limited, so the flexibility and practicality of REGEX models make them more suitable for initial exploratory queries such as the ones in this work. REGEX models are also convenient when searching for motifs in parasites such as *Toxoplasma* because there are fewer characterized examples of parasite motifs.

REGEX models find more motifs but need supportive information

While having advantages, REGEX models also pose some challenges. By generalizing the sequence composition of the motif, they can be too permissive or over-predictive, finding more motifs matches than could be functional, e.g. they might contain residues that disrupt binding or lie within hidden inaccessible parts of a protein. As mentioned above, we also do not get a way to discern between different REGEX exact matches. Here, I refer to spurious or non-relevant matches false positives. In order to reduce the overprediction problem and enrich the proportion of true positive motifs in datasets, the ones which might be functional, we have to complement them with supporting information. This information could be different, we can add details on the motif structural context like

their potential accessibility for interaction, e.g. if they are present in a disordered region, their conservation in homologous proteins in other species, and protein experimental information like their expression levels and cellular location.

ELM is the reference database for motif research

The Eukaryotic Linear Motif (ELM) resource is a publicly available manually curated motif database. It is the most representative collection of motif examples and knowledge from the literature. It was started in the early 2000's and it has continuously served as reference for motif function and research for almost 20 years (Puntervoll, 2003). The ELM database has steadily grown through the years to reach more than 3,900 validated motif examples that serve as basis for defining more than 300 models of different motifs (Kumar et al., 2022).

ELM motifs contain a range of supportive information

The ELM database is organized around motif classes, which represent functionally independent motifs backed by a set of literature references and instances, motif examples occurring in a protein and whose function has been previously validated by experimental methods. Classes contain a series of descriptions, database references, and gene ontology terms. All contribute to the definition of the motif and its associated properties, such as amino acid composition, binding characteristics, context and functionality (Gouw et al., 2020).

The descriptions summarize the motif function, its cellular role and the structural details of its interactions. There are links to the Protein Data Bank (PDB) when the instances are added from structural experiments. Gene ontology terms are provided for the biological processes the motif-domain interactions are involved in, for the cellular compartments where proteins containing the motif and its binding domain interact, and for the molecular function of the motif. Based on their function, motifs are then categorized in 6 major types: cleavage, degron, docking, ligand, modification and targeting (Chapter 1.4).

ELM classes have SLiM models in the form of REGEX

The class contains a unique model for the motif in the form of a REGEX. These REGEX are constructed based on different instances, protein alignments and available protein

structures (Gouw et al., 2020). The REGEX is associated with an estimation of how complex or information rich the pattern is, based on amino acid frequencies of protein disorder regions. This estimation mainly gives us an idea of how rare the motif pattern is, and should not be confuse with the likelihood of the motif being functional. There are simple motif patterns known to be functional.

ELM contains a prediction tool

The ELM database possesses a prediction tool that uses REGEX models from all ELM classes to look for motif instances in individual protein sequences. The prediction tool takes into account the protein context, its architecture and taxonomic range (Puntervoll, 2003). It is useful for looking at matches in individual proteins but it is a painstaking and impractical task to check protein by protein for a whole proteome. Instead, computational scripts can be developed to search for motifs in full proteomes by using the full set of ELM motif models.

SLiM models are retrieved from ELM for proteome motif survey

ELM has different available datasets that can be downloaded for further use. In order to develop the motif discovery pipeline, I downloaded the full set of motif classes that contain the motif code name and their REGEX. These manually curated models, backed by supportive information, are the first step reference for finding motifs in *Toxoplasma gondii,* proteins.

2.2. Structural context of motifs

Motifs reside in regions available for interaction

In order for protein domains to recognize and interact with SLiMs, they should be fully accessible, so one would expect to find them outside globular domains, transmembrane regions or other ordered structures. My motif survey will give us matches that might locate in these types of protein regions and therefore I had to consider their localization within the protein. It has also been observed that motifs are more frequently found in intrinsically disordered regions (Davey, Van Roey, et al., 2012), so I would filter out false positive matches by taking both protein disorder and architecture into account.

Differences among disorder predictors

There are many different computational programs to predict protein disorder. These predictors vary according to their basic principles*. Ab initio* predictors use the intrinsic amino acid composition of disordered proteins as well as their physicochemical properties to infer disorder tendencies. Machine learning predictors use different sets of disordered regions and proteins in order to classify and predict them. And finally meta-predictors combine the outputs of the previous types of predictors to arrive to a defined consensus (Katuwawala et al., 2020a).

IUPred is a practical predictor of protein disorder

Among *ab-initio* predictors, IUPred has been made freely available to add to bioinformatic pipelines, is easy to implement and to use, and has a good performance among predictors. IUPred takes the propensity of amino acids to interact with each other to then calculate the probability of single residues within a sequence to be in a disordered state. It uses a reference set of globular folded domains to construct a force field, then it considers the potential interactions of the central amino acid within a character window with their surrounding residues to make the calculation (Dosztányi et al., 2005; Erdős et al., 2021; Mészáros et al., 2018).

IUPred values over the whole length of the amino acid sequence from proteins identify well different protein regions, e.g., low values intervals identify globular domains and transmembrane regions. For this reason, the values are also a reference for protein architecture. The IUPred *v*alue of 0.4 can be used as a threshold to define whether the given residue is in a disorder context or not, e.g., individual amino acids values above it will be considered disordered as in the accessibility classification in (Benz et al., 2022).

Structural annotations complement disorder predictions

In order to complement my disorder assessments, I sought to retrieve information of known protein domains for *Toxoplasma gondii* from the UniProt (Universal Protein) database. UniProt is a highly used resource for the retrieval of high-quality protein data, as it integrates different experimental information and computational predictions and annotations (The UniProt Consortium et al., 2023). For this pipeline I took as reference the *Toxoplasma gondii* strain ME49 and through UniProt I retrieved domain mappings

from PROSITE, a specialized resource for protein families and domain annotations from reference databases (Sigrist et al., 2012). The retrieved table contained the domain names and their respective site mappings for 3,230 proteins. **Annex 2.1**

AlphaFold predicted structures provide an opportunity for protein architecture determination

The AlphaFold2 software has made a revolution in the protein structure determination field, moving it closer to solving the protein folding problem. It predicts the structure of a given amino acid sequence using a neural network algorithm that exploits available protein sequences and experimentally determined structures. It uses correlated residue pair information derived from multiple sequence alignments to derived their spatial and evolutionary relationships. These pairs are then used to produce a 3D coordinate prediction using physical and geometric information derived from PDB data (Jumper et al., 2021). This method has been used to predict almost all protein sequences in the UniProt database. Both the method and a predicted structures AlphaFold database have been made publicly available, allowing researchers to exploit them for structural analysis (Varadi et al., 2022). I retrieved all available *Toxoplasma gondii* strain ME49 predicted structures from the AlphaFold database, which in July 2022 consisted of 6,898 PDB files.

ColabFold provides a way to predict missing structures from the AlphaFold database

Not all *Toxoplasma gondii* proteins have structure predictions in the latest release of the database, missing around of 1,000 proteins. These are mainly large proteins of more than ~1,200 residues. Instead of waiting for another release I took advantage of the available options to predict some of the missing protein structures through the open access AlphaFold2 software ColabFold. ColabFold stands out for being a fast and easy to use software through Google Colab (Mirdita et al., 2022). It is able to predict both monomers and protein complexes and uses the mmseqs2 (Many-against-Many searching) software to search for similar sequences and align them. It accomplishes this in a faster manner than the classic AlphaFold pipeline, which relies on hidden Markov models that can take longer to generate sequence alignments. Using ColabFold batch option, I predicted a further group of ~60 protein structures of more than 1,200 residues, using 3 rounds of recycles and keeping only one model. Protein structures with more than 1,500 residues were still problematic to predict, so these few longer proteins represent a good

proportion of the ones that can be easily predicted, and they have the potential to contain more motifs.

AlphaFold structures can be used to predict disorder and accessibility in proteins

Predictions derived from the AlphaFold2 software have an accuracy score termed the predicted local-distance difference test (pLDDT) score on a scale from 0-100. This score is calculated using the local distance difference test Cα (lDDT-Cα) (Jumper et al., 2021). The IDDT-Cα test calculates the difference of a protein model to a reference set of structures using all pairs of Cα atoms and also estimates its stereochemical plausibility, i.e., that all bond lengths and angles are possible (Mariani et al., 2013). It has been observed that low pLDDT values correlate well with intrinsically disordered regions. This relation meant that AlphaFold could be used as a disorder predictor. And effectively, AlphaFold performed well and with high accuracy among disorder predictors when tested on a reference set of IDRs from the CAID database. From this it was estimated that a pLDDT value of 50 could serve as a threshold to infer whether a residue is in a disorder context (Tunyasuvunakool et al., 2021).

In order to employ and access the pLDDT values of the AlphaFold predictions I had to first retrieve them from their PDB files. In the PDB format, they are annotated in the column designated for the temperature factor, usually occupied by the B-factor score. I scripted a python program that takes the individual PDB files of *Toxoplasma* protein structures, retrieves the pLDDT values of each residue and saves them together in a text file in a python list format for later usage. **Annex 2.2**

The AlphaFold and ColabFold structures can be further exploited to approximate the accessibility of different protein segments and asses their availability for interaction. For example, there are regions with low quality scores but buried within domains that would not be available for interaction. Residue accessibility determination is accomplished by using the Dictionary of Secondary Structure of Proteins (DSSP) software. DSSP takes a PDB structure to assign secondary structure to its residues based on their atomic coordinates and hydrogen bonding patterns. It also gives a total solvent accessibility score which is a proxy for how many water molecules are in contact with a given residue

(Joosten et al., 2011; Kabsch & Sander, 1983). Assuming that the residue is buried in a folded domain, this value would be low, and when exposed to the solvent, higher.

Each amino acid has an intrinsic maximum solvent accessibility potential, larger amino acids could be in contact with more water molecules when exposed at the protein surface. In order to compare and employ accessibility values from different residues in a protein their scores should be normalized using maximum accessibility values for each amino acid. These values are usually derived empirically by analyzing the accessibility of residues in a large set of crystal structures. For this pipeline I used the theoretical normalization scale from (Tien et al., 2013) which models the maximum accessibility of a residue X based on all possible conformation of Gly-X-Gly tripeptides. I ran DSSP on each AlphaFold and ColabFold PDB file and scripted a python program that takes individual DSSP table files and retrieves the accessibility values of each residue and normalize them using the (Tien et al., 2013) scale. As for the pLDDT script, this one saves the accessibility scores together in a text file with a python list format for later usage. **Annex 2.3**

2.3. Motif conservation

Motif conservation in multiple sequence alignments (MSA) gives us a way to assess their functionality assuming that their conservation implies functional importance. To look for motif conservation I first collected protein sequence data of high quality to produce the MSAs for *Toxoplasma gondii* proteins that I analyzed and integrated into my pipeline.

The VEuPathDB resource is the reference database for eukaryotic pathogens

In order to collect protein sequences, and further information, I made use of the VEuPathDB (Eukaryotic Pathogen, Vector and Host Informatics) resource. VEuPathDB is the most known and used resource for the research of eukaryotic pathogens (protists and fungi), their hosts and invertebrate vectors. It combines large omics datasets as well as different data mining, visualization, and bioinformatic tools (Amos et al., 2022). There are fourteen interconnected databases that shared the same functionality and tools. Their power and usability depend on the extent and amount of experimental and research work done for every parasite family (genomes, RNA seq, Mass Spec and phenotypic data).

Parasitology researchers consistently refer to these databases, making them the standard for data collection and setting guidelines to report results.

Toxoplasma and other Sarcocystidae proteomes are downloaded from ToxoDB

ToxoDB is a member of the VEuPathDB databases that specializes in *Toxoplasma gondii*. It contains information on different *T. gondii* and *Sarcocystidae* species strains, encompassing 37 organisms. Throughout the analyses, *T. gondii* strain ME49 serves as the reference strain and *Sarcocystidae* species, as it is one of the most studied strains and from which most high-quality data is available. Additionally, I focused on 4 more *Toxoplasma* strains (from different *Toxoplasma* types, Section 1.III) and the 4 most closely related *Sarcocystidae* species **Table 2.1**. I retrieved the annotated proteomes from all listed organisms in fasta format.

Toxoplasma strains	Sarcocystidae species
Toxoplasma gondii ME49 (Type II) (reference)	
• Toxoplasma gondii GT1 (Type I) • Toxoplasma gondii MAS • Toxoplasma gondii VEG (Type III) • Toxoplasma gondii VAND	• Hammondia hammondii • Neospora caninum • Besnoitia besnoitii • Cystoisospora suis Wien I

Table 2.1 Toxoplasma gondii strains and Sarcocystidae species used in conservation analysis.

Orthologous groups are created through BLAST

In order to create protein orthologous groups, I carried out a BLAST search with default parameters to search for homologs for each *T. gondii* M49 protein among the proteomes of the eight additional species. In order to create comparable protein sequence groups with a similar number of organisms I retrieve the sequence with the lowest E-value from the result table from each protein selected for each strain and species. As proteins from *Toxoplasma* secretion organelles tend to be less conserved outside the *Sarcocystidae* group (Barylyuk et al., 2020), this will make the conservation comparison of different proteins a simpler task.

Clustal Omega is used to produce sequence alignments

In order to do the MSAs I used Clustal Omega, which is a fast, accurate and scalable sequence aligner. Clustal Omega creates a progressive alignment using a guide tree that

clusters similar sequences to then align them using Hidden Markov Models (Sievers et al., 2011). Its compatibility with Jalview, a highly used alignment viewer and editor was also appealing to implement in my pipeline (Clamp et al., 2004; Procter et al., 2021). Jalview facilitated later visual inspection of the alignments. Overall, I counted with a set of MSAs from *Toxoplasma* protein for later analysis.

The Relative Local Conservation is not practical for just a few related sequences

In motif discovery and research, conservation is usually calculated using the Relative Local Conservation (RLC) score. This metric makes it easier to spot small regions of conservation such as a functional motif. The RLC achieves this by comparing the single position conservation of a residue in a MSA with that of the surrounding positions (Davey, Cowan, et al., 2012). This type of metric requires at least 10's of homologous sequences with enough diversity, i.e., that they come from sufficiently distantly related species or paralogous proteins. *Toxoplasma* strains are mainly clonal and there are few related *Sarcocystidae* species identified and characterized at the genome sequence level. Furthermore, when analyzing secreted proteins (the ones I am primarily interested in) we also end up dealing with highly derived sequences or apparently unique and new proteins that would be missing in other species. For this reason, I expected orthologous groups of secreted protein to be small and to mainly contain *Toxoplasma* strains. Overall, using standard metrics in this pipeline, such as the RLC, is unfortunately not possible and I had to approach the conservation in a different way as described below.

2.4. Experimental evidence

To further back the validity of the motif matches we can enrich them with published experimental data. These can back up our candidates when we inspect them and help us analyze relevant groups of proteins. Information on proteins like expression levels and location evidence is essential to consider which proteins to analyze, e.g. some protein sequences might only be predicted from genomic data but they might never be expressed at any point of the organism lifecycle. And at the residue level, information of the post-translational modification state could back up the function of certain motif classes.

ToxoDB search strategies are used for information retrieval

Information for *Toxoplasma gondii* is vast compared to other apicomplexans and unicellular parasites, given that it is a model organism. ToxoDB is the hub for deposition, retrieval and analysis of these data. I was able to retrieve different *Toxoplasma* data through database search strategies. These are powerful and easy to use tools that facilitate the integration of different data types and datasets, e.g., the expression levels of different genes or proteins, with transcriptomics or mass spectrometry data respectively. ToxoDB also allows for REGEX based motif searches, but at the moment these searches do not handle motif match results as individual data types on which we could add more information into and build further strategies. For this reason, I had to download the information from the database and integrate it into my pipeline.

Mass spectrometry data tells us whether a protein is actually expressed

If there is no protein ever expressed motif matches in those will represent false positives, or if they are not present in Tachyzoites the motif matches will not be relevant during infection. I used the expression evidence tool to know whether or not a certain protein had mass spectrometry data as evidence of being expressed. ToxoDB has a collection of MassSpec data from almost 70 separate experiments and samples for *T. gondii* ME49.

Cellular location information is useful when inferring motif functionality

Knowing the context in which motifs will be functional is important to discern whether or not our matches are true positives, e.g. an extracellular protein will only interact with its binding motifs if they are present in extracellular proteins. Cellular elements are highly regulated in space and time, so having an idea of where in the cell a protein is located already offers a better idea about its function and potential interactors. In the case of effectors, we are interested to know if they are located in secretory organelles.

The HyperLOPIT method predicts subcellular location

It has been a challenging task to uncover the location of proteins in *Toxoplasma gondii*, as well as in other parasites. The location of a set of key proteins has been mainly determined by immunofluorescence microscopy experiments. For a larger set of proteins, approaches like homology transfer can be applied but their highly divergent proteins complicate this task. Predicting the presence of Signal Peptides (SP) to know the

secretory potential of proteins is also desirable but it is then hard to infer in which specific secretory organelle they locate and there are examples of proteins lacking apparent SP that are still reported to be located in secretory organelles (Barylyuk et al., 2020). This problem has seen a lot of progress thanks to the hyperplexed localization of organelle proteins by isotope tagging (HyperLOPIT) applied to *Toxoplasma gondii* Tachyzoites (Barylyuk et al., 2020).

This HyperLOPIT method is able to define the subcellular location of proteins based on profiles constructed from their abundance distribution in different fractionation pellets. In order to obtain protein fractions, Toxoplasma tachyzoites are disrupted by cavitation and the contents are then centrifuged. The different fractions are separated and the abundance of isobaric tagged proteins at each fraction is measured. These measurements are then arranged to build protein abundance profiles. The profiles of previously characterized proteins with known locations are used as the basis to build reference profiles for each location. These reference models are subsequently compared to the rest of the profiles and statistical analysis is used to cluster protein profiles and assign them to distinct subcellular locations (Barylyuk et al., 2020). Ultimately, through HyperLOPIT around 5,000 *Toxoplasma* Tachyzoite proteins were assigned to around 20 subcellular locations and the dataset was deposited in ToxoDB.

I retrieved a list of all *Toxoplasma gondii* proteins using the Mass Spec. evidence search together with their HyperLOPIT location predictions from ToxoDB. The filter of minimum number of unique peptide sequences was 1, i.e., to retrieve proteins with at least 1 mass spectrometry peptide mapped to it **Figure 2.1.** In this step I downloaded the table of protein IDs for 5,800 proteins together with their Product Description and HyperLOPIT predicted locations.

BioID provides location evidence for Bradyzoite proteins

I obtained secretion evidence of a group of 71 proteins from Dense Granule of a Bradyzoite stage experiment by Proximity-dependent Biotin Identification (BioID) (Nadipuram et al., 2020). Here they carried out a BioID assay using the bradyzoite upregulated protein MAG1. They created a fusion of MAG1 with the biotin ligase BirA which biotinylates nearby proteins interacting or localizing with MAG1 to then be purified

and identified by mass spectrometry. This experiment added some Bradyzoite stage proteins to the location information as most of them are not identified by the HyperLOPIT experiment, mainly because they were performed in different *Toxoplasma* stage forms. I downloaded the Supplementary information and formatted the data for further use in the pipeline.

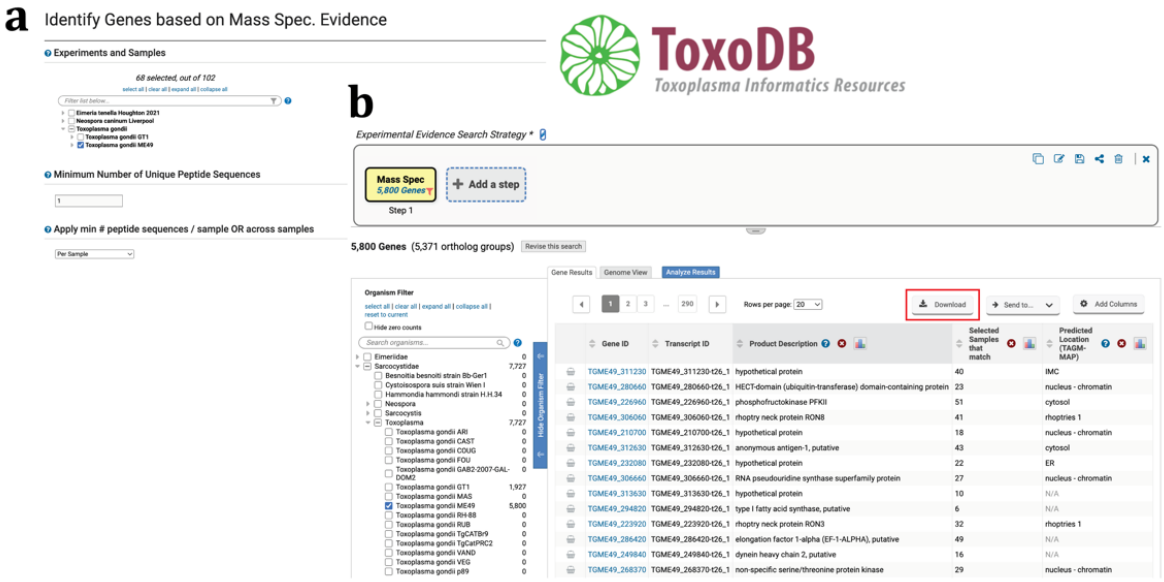


Figure 2.1 ToxoDB Mass Spec. Evidence tool search strategy. a. ToxoDB Mass Spec. evidence search tool parameters selecting 68 experiments and samples and marking the minimum of unique peptides. **b.** Following filter by organism (*Toxoplasma gondii* strain ME49) and addition of LOPIT location information.

Phosphorylation sites complement motif predictions

Phosphorylation evidence could also be coupled with motif matches to assess whether or not the motif match is functional. Some motifs like the 14-3-3 binding sites depend on phosphorylation to be functional. For this, I downloaded phosphosite data from separate experiments deposited in ToxoDB. From the three main sources, the first one surveyed the phosphosites targeted by calcium-dependent kinases in tachyzoites, and identified ~546 phosphosites from more than 300 *Toxoplasma* proteins (Nebl et al., 2011). A second surveyed the phosphoproteome of *Toxoplasma* tachyzoites that were either intracellular or free in the host material and identified 10,000 phosphosites (Treeck et al., 2011). And the third surveyed for phosphoproteome of *Toxoplasma* and identified 2,296 phosphosites (Beraki et al., 2019). All combined amount to 22,850 phosphorylation sites from 3,195 unique proteins. I then formatted the information from representing

combined phosphorylation data for a single protein to individual protein phosphorylation site information in order to allow their mapping to motif matches **Annex 2.4.**

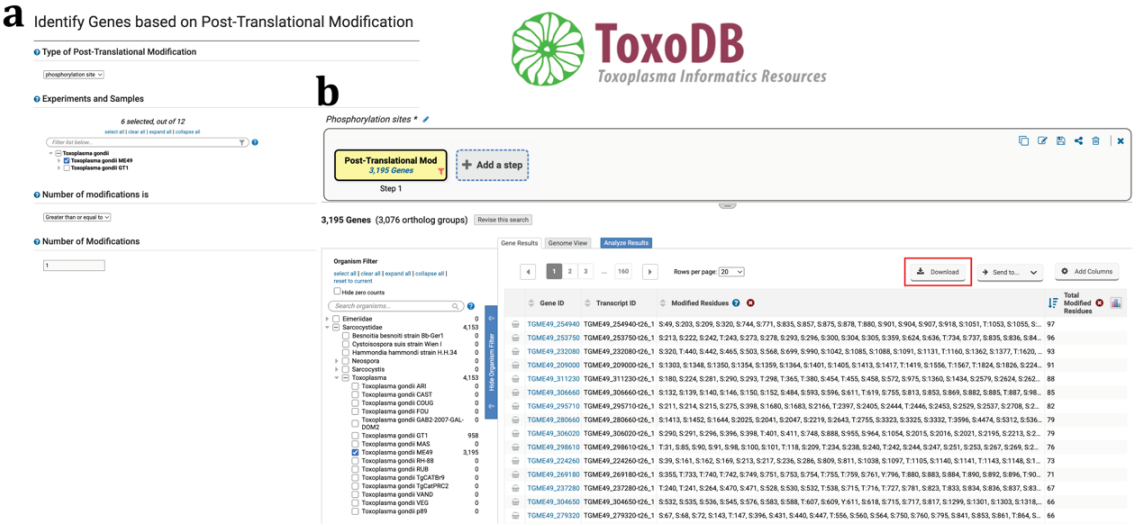


Figure 2.2 ToxoDB PTM search tool search strategy. a. ToxoDB Post-Translational Modification search tool parameters selecting 6 experiments and samples and marking the minimum of modifications. **b.** Following filter by organism (*Toxoplasma gondii* strain ME49).

Data and Software overview

In **Table 2.2** we can observe a summary of the data and software collected in order to assemble the motif prediction pipeline. From these, it is already clear that we do not have data for all the proteins in the reference proteome, e.g., the number of AlphaFold2 predictions are limited by computational resources and protein size, while others could be related on experimental setups or on biological reasons, e.g., not all proteins are expressed or are phosphorylated at a given experiment or in any condition.

	Reference	Structural	Conservation	Experimental
Data	-T. gondii ME49 (8,322 protein sequences) -ELM motif classes (318 REGEX models)	-AlphaFold (6,898 structures) -Prosite Domains (3,230 proteins)		-Mass Spec & LOPIT location (5,800 proteins) -Phosphorylation (22,850 phosphosites in 3,195 proteins)
Software		-IUPred -DSSP -ColabFold	-Clustal Omega	

Table 2.2 Data and software collected for the motif prediction pipeline.

3. Prediction pipeline

I was able to integrate all of the previous structural, conservation and experimental information and software in a pipeline by combining both Python and R programming. The principal aim was to determine motif matches in *Toxoplasma gondii* proteins and then add further information features that will later help me retain more meaningful matches and filter out potential false positives **Figure 2.3**.

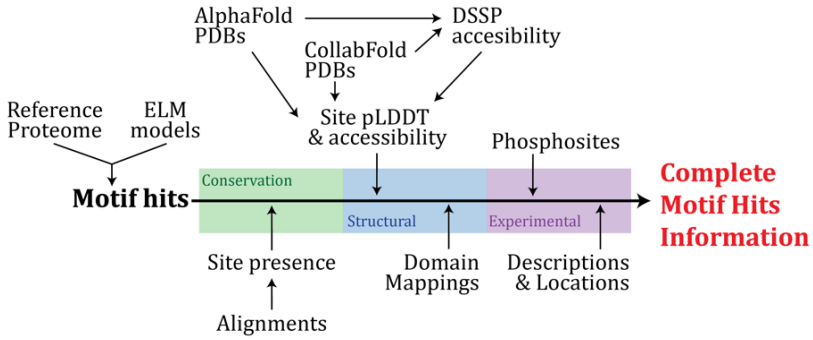


Figure 2.3 Motif matches discovery pipeline.

Initial motif matches are retrieved together with disorder scores

My workflow starts with a python script that gets an initial set of motif matches in a reference proteome by using a list of motif regular expressions. During this step the disordered state of each motif match is also determined. This is done by taking whole protein IUPred-long scores, selecting the values of the amino acids in the motif and averaging them. I then used an IUPred score threshold of 0.4 to define a disorder or order context value. The script gives as a result a complete list of all motif matches in all the proteomes together with their features and IUPred score. In this case I used the reference proteome of *Toxoplasma gondii* strain ME49, which contains the fasta sequences of its 8,322 proteins in FASTA format. For the Motif REGEX, I downloaded all 318 available ELM motif classes **Annex 2.5**.

Conservation of motifs can be approximated by assessing motif presence in MSA

The python script 'MotifMatches_Sites.R' uses this previous file to retrieve match sites, reformats them and then the script 'MotifMatches_InAlignments.py' look for the matches in a set of alignments. As a way to spare computational efforts, I only carried MSAs carried for proteins with expression evidence having a total set of 5,800 MSAs files. Instead of

looking for conservation metrics, I assess the presence of a given motif in a region of the alignments in sequences of other *Toxoplasma gondii* strains and *Sarcocystidae* species.

The software takes a MSA file and our list of Motif Hits, it transforms the MSA into a python list in which the sequence ID is the key and the alignment sequence is a chain of characters, this includes the gap character. Then it creates an equivalent list without the gaps having both lists allowed us to trace back the position of residues in any sequence between alignment and complete sequence. Afterward, it takes a motif hit and looks for its position in the reference strain, in this case *Toxoplasma gondii* ME49, then looks for the presence of the same motif around the same region or vicinity, in this case I used 15 amino acids **Figure 2.3**, **Annex 2.6.**

As I considered 5 strains and 5 species (**Table 2.1**) I did the calculation by taking the presence of other strains and dividing it by the number of extra species or strains apart from *Toxoplasma gondii* ME49. This means that values with 0 indicate the presence in just the reference strain, values of 0.25 mean that the motif is present in 1 additional species or 1 strain. And a value of 1 means that the motif is present in all extra species or strains. In this step I started using a motif key comprised of the protein sequence, the motif class identifier and the match hit number. This key made it possible to identify individual motifs for later data integration **Annex 2.7**.

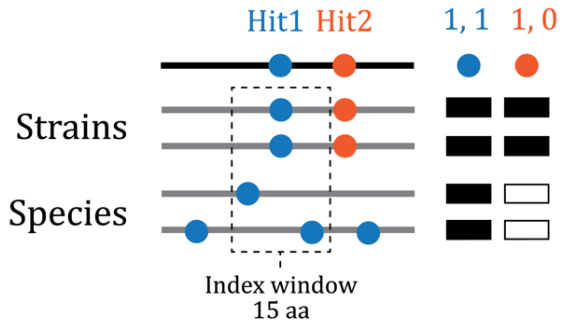


Figure 2.4 Motif match presence evaluation. The diagram captures how motif presence is obtained from protein alignments. The presence of motif hits from REGEX models (blue and orange) in sequences of different *Toxoplasma* strains and *Sarcocystidae* species is compared and the proportion of each is retained as a conservation score.

Further structural features are obtained from AlphaFold predicted structures

A further python script uses the motif matches to look for them in AlphaFold and ColabFold predicted structures and extract their pLDDT values and DSSP accessibility

values. A motif match is found in a PDB file, its pLDDT values are averaged and saved. The same procedure takes place by finding a motif match in a DSSP format result table and its accessibility values are extracted. After obtaining the normalized accessibility the scores of each motif are averaged and saved. The output of this script will give a table with all motif matches and their pLDDT and accessibility values. Annex8

Phosphosites and domain data are mapped to motif matches

Experimentally defined phosphosites collected from ToxoDB were combined with the motif matches from the motif survey. Each phosphosite had a seq_id (sequence identifier) that was used to check for motifs in that protein, the exact position of the phosphosite was kept and if it was localized between the start and end positions of the motif, the phosphosite was added to that particular motif. In the end the program gave a table with the motif match keys, the number of phosphosites and the exact residue and position of each one. **Annex9**

In order to determine whether a motif landed within known domains I scripted a python program that coupled the different motif match sites to the domain information previously retrieved. Here the position of a motif match was compared to the extension of the domains of the same protein and if the motif was localized within it, the number and information of the domains were added. **Annex10**

Motif data is integrated in a multicomponent dataset

At the end using an R script I combine the initial results table of motifs matches together with the different information tables. Using the common protein identifiers and through a motif match key I created a multicomponent dataset containing conservation, structural and experimental information **Table 2.3**. The overall pipeline and its details are depicted in **Figure 2.5.** Single motif matches were enriched with specific data that I later used in further analyses and filtering steps **Figure 2.6**. Annex11

Overall, using the discovery pipeline described above I was able to retrieve 2,472,290 motif matches in all the 8,322 proteins of *Toxoplasma gondii* ME49 proteome using the set of 318 motif classes from ELM and enrich them with different structural, conservation

and experimental information. I also deposited all computational scripts and data in a GitLab repository for later sharing:

- https://github.com/JesnsAV/Toxo_SLiMs_2

ID	Protein	Motif info.	Conservation	Structural	Experimental
-key	-Seq_id -Protein_ID -Product description	-Motif_Name -Match_N -Motif_Instance -Motif_sSite -Motif_Type	-Presence_str -Presence_spc	-Motif_Disorder -Dis_context -Mean_pLDDT -Mean_acc -Motif_AC -Doms_num -Doms_name	-LOPIT_MAP -LOPIT_location -Organelle -Location_Evidence -Secretion_Evidence -modsNum -modsSites

Table 2.3 Information types of final motif matches results.

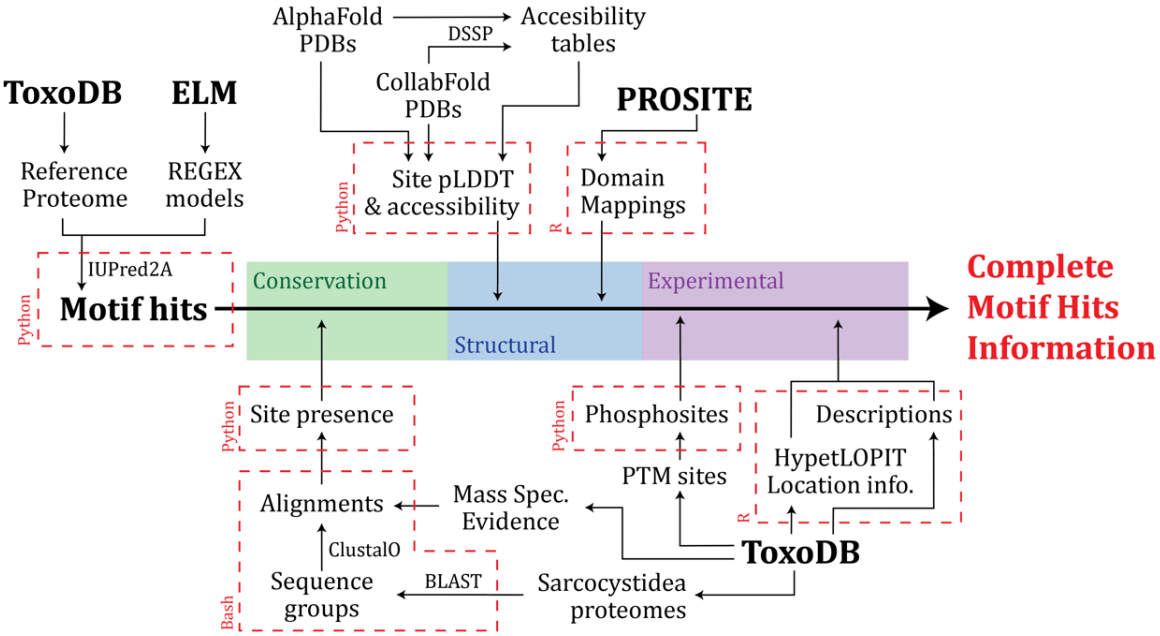


Figure 2.5 Motif matches discovery pipeline. Databases and main result tables in Bold. The main programming language of each data processing step is encircled in red.

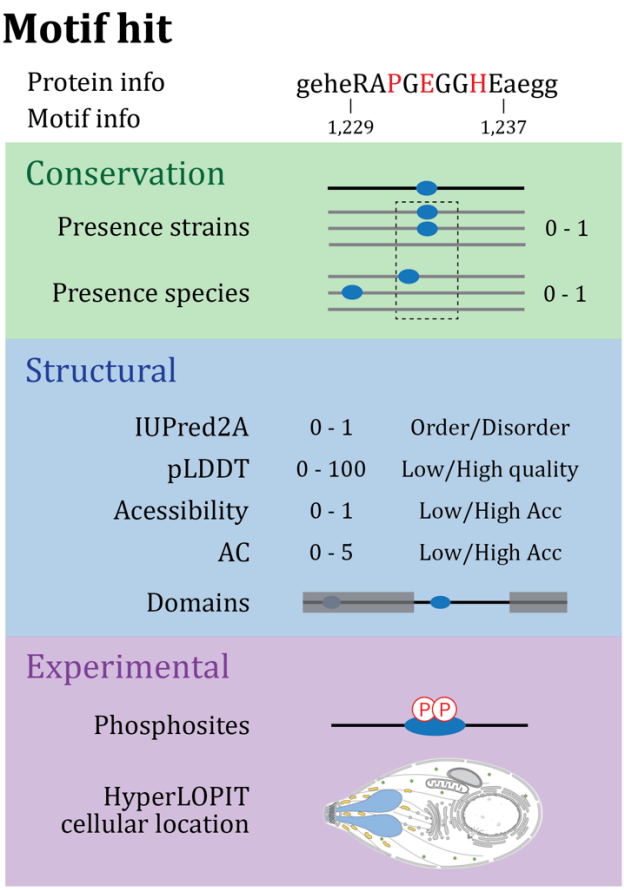


Figure 2.6 Motif match information summary. Different attributes of a motif match from my discovery pipeline. The ranges of the different scores are indicated as well as evaluation derived from the filtering threshold for each one. At the top the motif match is indicated with capital letters and the key interacting residues in red.

CHAPTER 3
Motif Match Filter Development

After obtaining an initial result table the following goal is to filter out matches that are obviously False Positive based on the available motif class data and the structural and conservation information of each match. Unfortunately, there are not enough known *Toxoplasma* motif instances in the ELM that I could use to benchmark and establish a filtering procedure using them as a reference (only 14 instances in 6 proteins). All ELM motif instances have also not been used together with AlphaFold predictions to evaluate ideal parameter filters. Instead of relying on a validated reference set, I filtered motif matches by first assessing their taxonomic logic, then exploring the distribution and trends of their different scores and finally combining them to reach a filtered set of motifs matches. Single filters are difficult to apply to all motif classes so combinations of them that take into account the biology and characteristics of different types of motifs are helpful before narrowing down to a defined list of candidates to test experimentally. In the following sections I explain the logic for the taxonomy filters, the properties of the different scores from the motif discovery pipeline table, and how I use them to reach a filtered set of motif matches.

3.1. Motif class taxonomy filter

The ELM classes include a Taxonomy term that indicates the organismal range of the motif-domain interaction. These ranges are based on the motif instance organisms and alignments created during the annotation process, e.g. I annotated the LIG_LYPXL_yS_3 class into the ELM database based on two motif instances from *Saccharomyces cerevisiae* and I developed the REGEX pattern using MSAs created with other yeast species from the Saccharomyceta clade, thus this class only applies to organisms in this clade. Nevertheless, terms in ELM are not always precise, e.g., sometimes they state they are

valid in Eukaryotes, which then would include Apicomplexans, but in reality, the domain required to interact with the motif is completely absent from this clade. For this reason, I inspected every ELM class to reevaluate their taxonomy range and assess if they would apply for *Toxoplasma gondii*. I divided Motif Classes by three criteria: does the Taxonomic range includes *Toxoplasma*? Is this range backed by the interacting domain taxonomic distribution in InterPro for *Toxoplasma*? And does this range include Vertebrates (mainly focusing on Mammals and Birds)? By evaluating this manually I was able use a binary/logical evaluation of the criteria I defined 8 groups of motif classes **Table 3.1**.

T. gondii		***Vertebrates***			
ELM	**InterPro**	**ELM**	**Num (%)**	**Logic**	**Code**
1	1	1	126 (41%)	Present in both groups	1
1	1	0	3 (0.9%)	Present only in T. gondii	2
1	0	1	23 (7.5%)	Not valid in T. gondii (low precision)	3
1	0	0	0	Not valid in T. gondii (low precision)	4
0	1	1	72 (23.4%)	Present in both with reservations	3
0	1	0	17 (5.5%)	Valid in T. gondii with reservations	4
0	0	1	59 (19.2%)	Not present in T. gondii	3
0	0	0	7 (2.2%)	Not present in T. gondii or Vertebrates	4

Table 3.1 Taxonomy presence group logic table. The validity of the motif classes is represented by 1 and 0 depending on whether they apply to Vertebrates and Toxoplasma in ELM, and in InterPro for the later. The logic column explains what would be the conclusion for such combination of values.

In the groupings based on the taxonomic validity around 41% of motif classes cover both *Toxoplasma* and Vertebrates. All of the matches from these classes can then be included in further analysis without reserves. Less than 1% of motif classes apply only to *Toxoplasma*, specifically 2 Classes based on apicomplexan instances LIG_LIR_Apic_2 and TRG_Pf-PMV_PEXEL_1 and one that includes all Eukaryotes but specifically excludes Vertebrates DEG_CRL4_CDT2_2. There were several Classes with conflicting annotation regarding the taxonomic extent of their domains. A total of 89 Classes excluded *Toxoplasma* but the motif-binding domain is actually present in some of its proteins, representing a binding potential. And 23 Classes stated that they include *Toxoplasma* but the motif-binding domain is absent in any of its proteins, which means that these taxonomic ranges are not precise enough. I consider the ELM criteria to have precedence even if the InterPro information gives us an idea of potential interactions.

With this we can say that there is a total of 154 classes absent in *Toxoplasma* but present in Vertebrates. Motif from these classes would not be functional in *Toxoplasma* proteins, unless they are present in in the right context, e.g., matches in secreted proteins from these classes would have the potential to bind host proteins when present in the same cellular location and could probably be involved in infection. For this reason, I kept the matches from these 154 classes in proteins with evidence to be in secretory organelles. Finally, there were 24 Classes not covering any of the groups and thus I filter out the matches derived from them. The Logic and proportion of the classes and matches are summarized in **Table3.2.** From these I only take Matches from groups with the code 1 and 2, and from 3 if they are in proteins reported to be in Secretory Organelles. This reduced the overall 2,472,290 raw motif matches to 1,396,026 motif matches (56.46%), so with this step I filtered almost half of the originally matches.

Code	Logic	Classes (%)	Matches (%)
1	Present in both (Retained)	126 (41%)	**1,337,742 (54.1%)**
2	Only in Toxoplasma (Housekeeping)	3 (0.9%)	**35,415 (1.4%)**
3	Only in Vertebrates (Omitted or Infection)	154 (50.1%)	782,158 (31.6 %) **22,869 (0.9%) Sec Org**
4	Absent in both (Omitted)	24 (7.8%)	316,975 (12.8%)

Table 3.2 Taxonomy filtering groups logic. In bold groups of matches that were retained for further analysis**.**

3.2. Structural scores filters

IUPred disorder scores

IUPred scores of the motif matches in the result table have a broad distribution **Figure3.1a**, from a peak at really low values to an extended distribution going down at high disorder scores. This firstly indicates that motifs do not have an intrinsic property of being in any order or disorder state, as there is a continuum of values from low to high, even across the IUPred 0.4 value disorder threshold. This also highlighted the need of more information to differentiate between True and False positives, especially among the motifs around the threshold.

IUPred score distributions vary according to Motif Type

The IUPred scores display different distribution trends when divided according to Motif type **Figure 3.1.b**. This could be explained by the basic residue composition of the motif models, e.g., if certain types of motifs have more prolines, they will more likely be located in disorder regions and have a higher IUPred score. LIG and DOC motif models have a higher density in low IUPred scores. On the other hand, CLV motifs show an opposite trend, with most hits having high IUPred scores. There is a higher quantity of motif matches on proteins from secretory organelles with high IUPred scores compared to proteins from other cellular locations **Figure 3.1.c.**

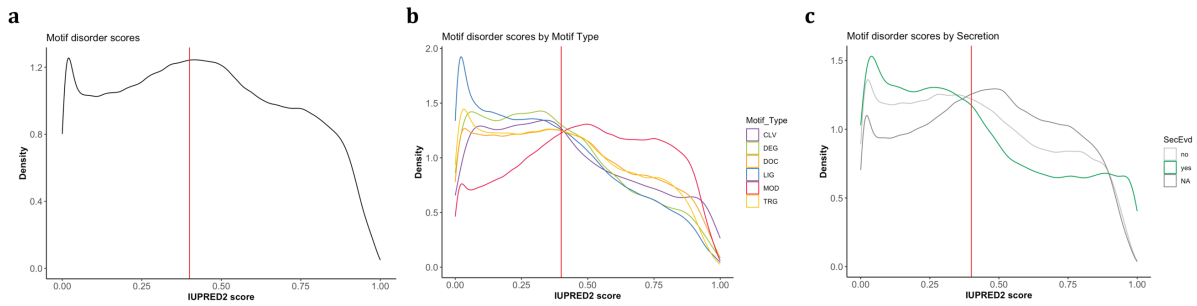


Figure 3.1 Disorder score distributions of the different motif matches. Density plots of the mean IUPred disorder values for **a.** all motif matches, **b.** all motif matches divided by motif class type, and **c.** divided by evidence of being in secretory organelles. The 0.4 IUPred value threshold is represented with the vertical red lines.

AlphaFold pLDDT scores

As mentioned before in the section 2.II, AlphaFold pLDDT values correlate well with disorder, so I inspected their potential to complement IUPred scores when filtering out False Positives. Not all sequences in the *Toxoplasma* proteome have an AlphaFold structure prediction, mainly because of their amino acid length exceeding the AlphaFold length cut-off. Because of this not all motif matches have AlphaFold and DSSP scores, actually only 50% of all matches have available scores. This might be a correlation between protein length and the number of motif matches they have, i.e., longer proteins without predictions are more likely to contain more motif matches (not shown).

Unlike the IUPred scores, the AlphaFold pLDDT score distribution is more bimodal with two peaks in distant parts of the distribution **Figure 3.2.a**. When splitting the distribution by IUPred disorder context, IUPred values split by a 0.4 threshold, there are not two disconnected peaks but instead two distributions that overlap on values. The motifs in

disorder context distribution have a high peak in low pLDDT values with a long shoulder towards the high pLDDT values. The motifs in order context distribution still has two peaks, a higher one on high pLDDT values with a lower peak in low values. The distributions pLDDT values divided by motif type and by disorder context are also different. In the order pLDDT value distribution, the motif type that has higher values are TRG and LIG motifs, while the ones with lower pLDDT values are from MOD, DOC and DEG types **Figure 3.2.b**. This could mean that if I would filter out all motifs in order context, we will also filter some of the MOD and DOC motifs with low pLDDT values. In the motif in disorder distribution, MOD, DEG and DOC are still the ones with lower pLDDT values, while TRG the higher ones, but all of them have their peak in the lower values **Figure 3.2.c**.

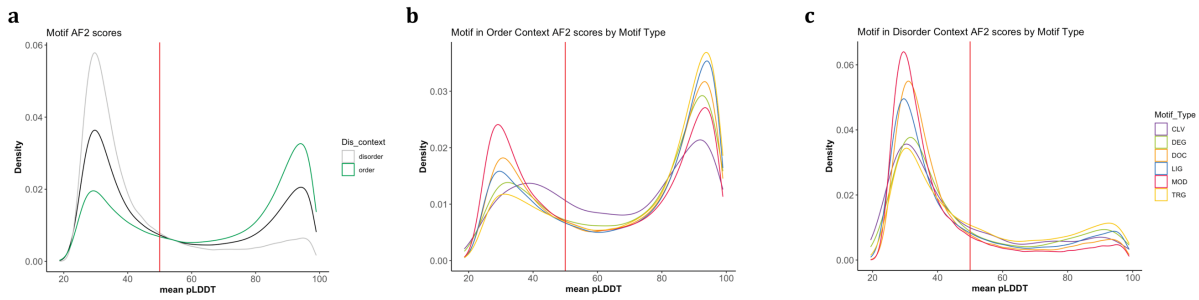


Figure 3.2 AlphaFold pLDDT confidence score distribution of the different motif matches. Density plots of the mean pLDDT confidence values for **a**. motif matches in different disorder context, **b.** motif matches divided by motif class type in order context, and **c.** in disorder context. The 50 pLDDT confidence value threshold is represented with the vertical red lines.

DSSP accessibility scores

Motif matches with higher accessibility scores are more abundant, displaying a distribution with a peak at high accessibility and a broad shoulder towards the lower ones **Figure 3.3.a**. While there is a higher number of motifs in order context with low accessibility, there are still some with high accessibility values **Figure 3.3.b**. The motifs in order context with high accessibility are of the DOC and MOD type. DSSP accessibility scores of motif matches in disorder context have a lower density on lower accessibility scores and most types show the same distribution **Figure 3.3.c**. I used a relative accessibility threshold of 0.36 derived from (Rost & Sander, 1994) to divide values.

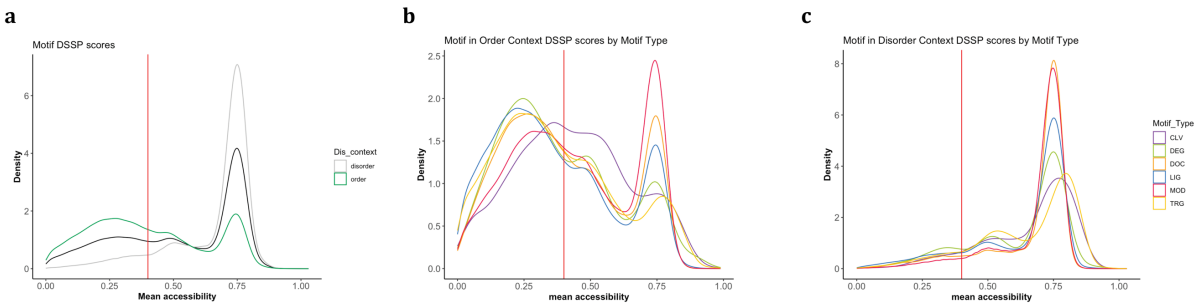


Figure 3.3 DSSP accessibility score distributions of the different motif matches. Density plots of the mean DSSP accessibility values of **a.** motif matches in different disorder context, **b.** motif matches divided by motif class type in order context, and **c.** and in disorder context. **a.** black line represents all motif matches, the green line motifs in order context and the grey one the ones in disorder. The 0.36 accessibility value threshold is represented with vertical red lines.

Combined accessibility and disordered scores

Analyzing the distribution of values of different metrics can aid us to spot relevant groupings and define optimal thresholds to filter out possible false positives. Motif matches with high IUPred scores also tend to have low AlphaFold pLDDT values, **Figure 3.4.a**. The high pLDDT values have a peak on low IUPred, these two properties would clearly point that this peak represents ordered well defined regions. While the low pLDDT peak, the distribution is more extended, and the group of low pLDDT is not only populated with high IUPred scores.

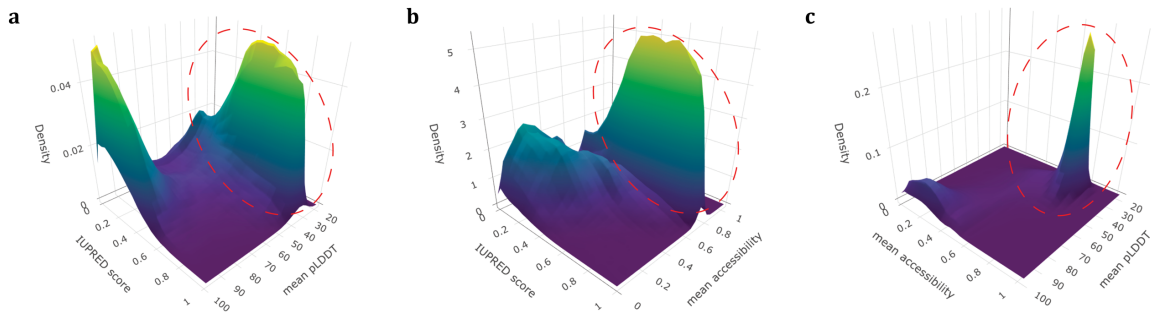


Figure 3.4 Combined score distributions of the different motif matches. 3D density plots of the combined structural values **a**. mean IUPred vs mean pLDDT values, **b**. IUPred vs mean accessibility values, and **c.** mean pLDDT vs mean accessibility. Desirable groups of motif match scores are encircled with a red dotted line.

While comparing the IUPred scores with the DSSP accessibility values, their combined distribution shows a division between high and low accessibility matches, but there is not a clear division along the IUPred values **Figure 3.4.b**. There are indeed more matches with low IUPred and low accessibility. And when analyzing the relation of both

accessibility and pLDDT values, there are more motif matches with low pLDDT and high accessibility, while there is a broad distribution of the rest of the values **Figure 3.4.c**. Taking this into account it follows that I would need to filter out motif matches with high pLDDT values as this metric is negatively correlated with accessibility (R= -0.84) and IUPred scores (R= -0.48), while these two are positively correlated (R=0.58).

Instead of just using the pLDDT values, I combined both Accessibility and pLDDT scores into a single one by first scaling the pLDDT scores between 0 and 1, instead of 0 to 100, and then dividing the accessibility score by this number. $AC = \frac{accessibility}{pLDDT*0.01}$. This new score which I refer to as Accessibility confidence (AC) means that high accessibility scores with high quality pLDDT values will be punished and lowered, while high accessibility scores with low pLDDT ones will be scored higher. Motif matches in an ordered context have lower AC scores while the ones in disordered context have a two-peak distribution **Figure 3.5.a**. Motifs in disordered context of the TRG and CLV types tend to have low AC values **Figure 3.5.b**. There are significantly more motif matches with low IUPred scores and low AC, while after a certain AC value there is a broad peak with high disordered values **Figure 3.5.c**. Taking these two groups in consideration we can say that there is a useful cutoff for motifs in proteins with predicted AlphaFold structures. The combination of pLDDT and accessibility score in the Accessibility confidence of 0.8**,** the AC value for the accessibility threshold 0.36 from (Rost & Sander, 1994) and the pLDDT value of 50, could be used to filter out possible false positive.

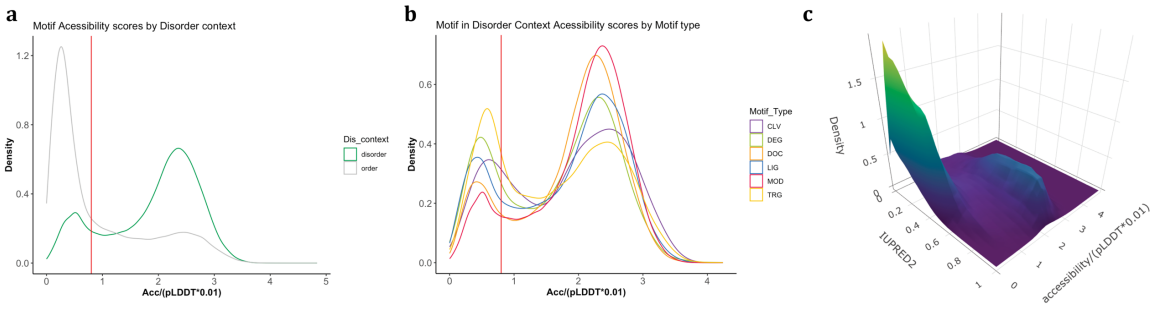


Figure 3.5 Combined accessibility score distributions of the different motif matches. Density plots of the combined accessibility values for **a.** motif matches in different disorder context, **b.** motif matches divided by motif class type in disorder context, and **c.** the 3D density plot of mean IUPred vs the combined accessibility score. The 0.8 combined accessibility value threshold is represented with the vertical red lines.

3.3. Motif conservation filters

When checking the conservation, we should remember that values are fractions of the extra 4 species or 4 strains, meaning that they go discretely from 0, 0.25, 5, 0.75 and 1. There are more matches having at least one or two more species than *Toxoplasma*, most probably *Hammondia* and *Neospora* (not calculated) **Figure 3.6**. After this most motifs are present in at least three species, then in all 5 species and finally there are few motifs present only in *Toxoplasma*. Considering strains, most of the motif matches are present in all 5 *Toxoplasma* strains (70%), while there are only a few that are present in less than 4 strains.

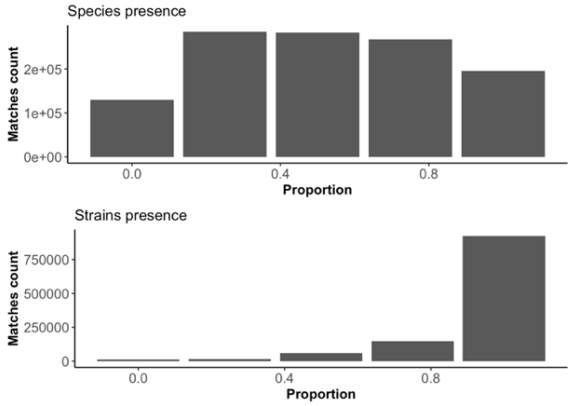


Figure 3.6 Motif match presence score distributions. Histogram plots of the presence distribution in *Toxoplasma* strains and in *Sarcocystidae* species**.**

When analyzing motif matches present together with other motif characteristics, there are no outstanding trends. For example, the motif matches that are present in more species are more likely to be in an order context, and when they are present in less than 2 species the more disorder, they are **Figure 3.7.a**. In the case of strain presence, the proportion of order and disorder state remains similar. Considering the AC score, low and high accessibility defined by the 0.8 threshold, the motif matches present in more species have lower accessibility confidence, while among strains the proportions are mostly similar **Figure 3.7.b**. There are no apparent trends between motif presence and the motif types for both species and strains, there only seem to be more LIG motifs in motif matches only present in the reference species **Figure 3.7.c**. And finally, there is a higher proportion of motif matches in secreted proteins present in fewer species, while there are more motif matches in non-secreted proteins **Figure 3.7.d.**

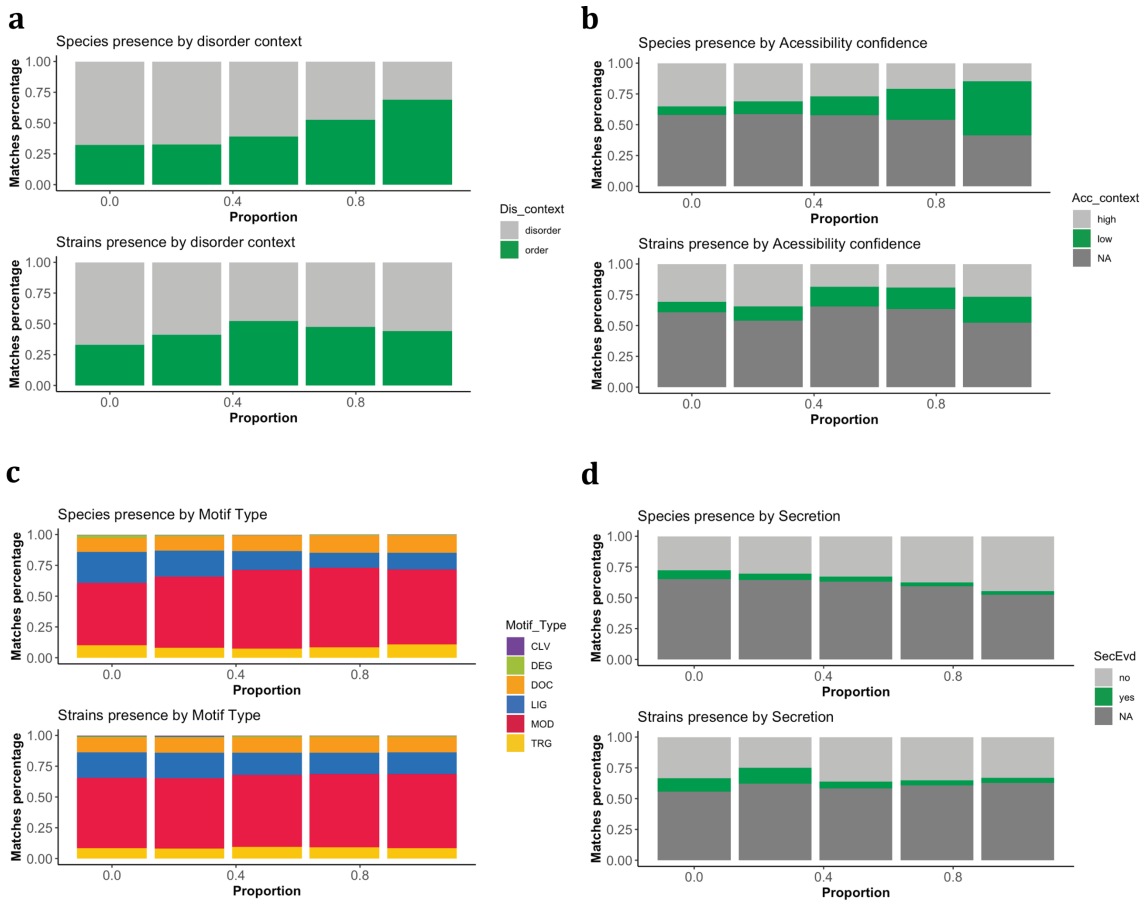


Figure 3.7 Breakdown of motif match presence score distributions. Histogram plots of the presence distribution in *Toxoplasma* strains and in *Sarcocystidae* species representing the proportions of **a.** matches in different disorder context, **b.** combined accessibility confidence, **c.** by different motif types, and **d.** by evidence of being in proteins of secretory organelles**.**

3.4. Combined filters

Considering all combinations of filter thresholds and criteria, there are 341,626 motif matches fulfilling the IUPred threshold, being present in at least one species, absent in only one strain and being outside of domains and only not surpassing the desired AC threshold. There is a further group of 158,129 motif matches fulfilling all criteria **Figure 3.8**. Both groups represented a 24.47% and 11.32% respectively from the post-taxonomic 1,396,026 filtered match total. I decided to keep two groups: the one with motif matches that covered all the criteria and another one with motif matches that did not fulfill the AC threshold because they were located in proteins without an AlphaFold or ColabFold structure prediction. I only produced MSAs for proteins known to be expressed having values for strain and species presences means the ones with them, also have evidence for expression. In the end I reached a total of 476,224 motif matches from 211 motif classes in all the 5,211 proteins with alignments.

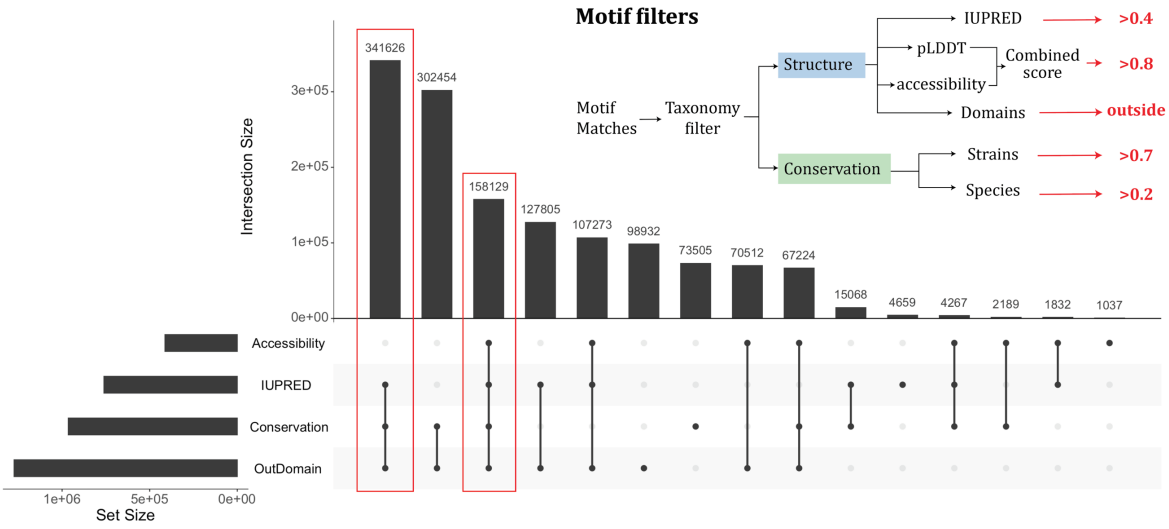


Figure 3.8 Combined filters for the final list of motif matches. Upset plot showing the different amount of motif matches covering different value criteria (on the top right corner). Final selections of motif matches are marked with a red box.

3.5. Cellular location

Protein subcellular location evidence is another line of evidence that facilitates linking motif matches with specific biological processes. In the case of this project, I am primarily interested in proteins that are located in secretory organelles (micronemes, rhoptries and dense granules) and have the potential to interact with host cell components. Knowing the location of other compartments can also help us filter out motifs models that do not correspond to functions associated with them. Based on the HyperLOPIT and BioID location data, most of the matches reside on proteins from the nucleus, followed by proteins in the cytosol and thirdly in the ER **Figure 3.9.a**. Matches from secretory organelles do not amount to a substantial part of the total, but among them the dense granule and rhoptries have the largest set of motif matches. Most motif matches from different locations shared similar proportions of disorder context (at around 60%). The cellular locations that have a larger motif proportion in disorder context are the apical region and the nuclear proteins, and for motifs mostly in ordered context are the proteosome proteins. When filter for disorder we will then enrich motifs in proteins from specific organelles. In the case of motif types, there are no big differences among locations. They mostly maintain the same proportion of motif types **Figure 3.9.b**. A clear exemption is the motif matches from secretory organelles which were the only ones holding CLV type motifs based on the taxonomic filter. Motifs in ribosomal and nucleolar proteins also have a higher proportion of TRG motifs.

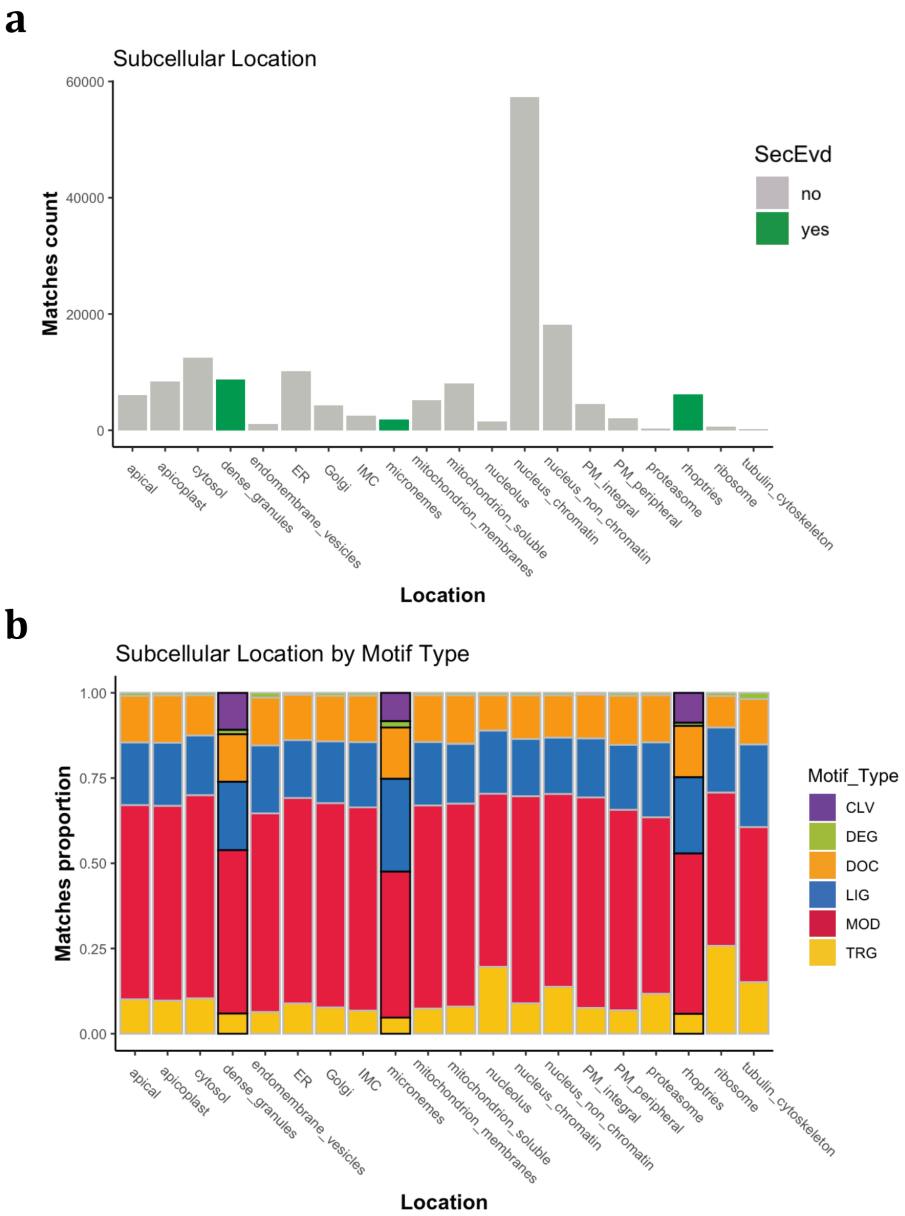


Figure 3.9 Distribution of final motif matches across cellular locations. Histograms show **a.** the total amount of motif matches from different cellular location, and **b.** the proportion of motif matches from different types from different cellular locations. Matches from secretory organelles is marked in green in **a.** and by a black border in **b**.

The filter matches and plots in this chapter were produced using R-markdown and deposited together with the resulting tables in a GitHub repository:

- https://github.com/JesnsAV/Toxo_SLiMs_2

CHAPTER 4
Motif Candidates

In the following sections I explore different set of motifs that were obtained by focusing on specific functional groups of motifs and only on the ones located in proteins from *Toxoplasma gondii* secretory organelles. Further filters are difficult to apply to all motif classes so combinations of them that take into account the biology and characteristics of different types of motifs are helpful before reaching a defined list of candidates to test experimentally. Another way to identify possible true motif matches is to select groups of motif matches with related functions that are found in the same protein. I applied different criteria to select the groups, from the functional overlap of motifs as well as their presence in the appropriate organelle to be functional. I provide different lists of matches to highlight their numbers as well as some illustrations of their location within the proteins.

4.1. Motifs in secretory organelle proteins

After filtering, I counted 476,224 motif matches from 211 motif classes. Using the cellular location information, we can observe that 17,733 matches are in proteins from secretion organelles, only 3.72%. The split among the three secretion organelles can be observed in **Figure 4.1**, most of the matches are from dense granule proteins, while the minority are in microneme proteins. This distribution can be related to the levels of disorder from proteins in those organelles **Supplementary Figure 4.1.** I expect that the number of matches in microneme proteins also might be lower as they are mainly extracellular proteins and only a handful of motif classes are annotated to be functional in that context.

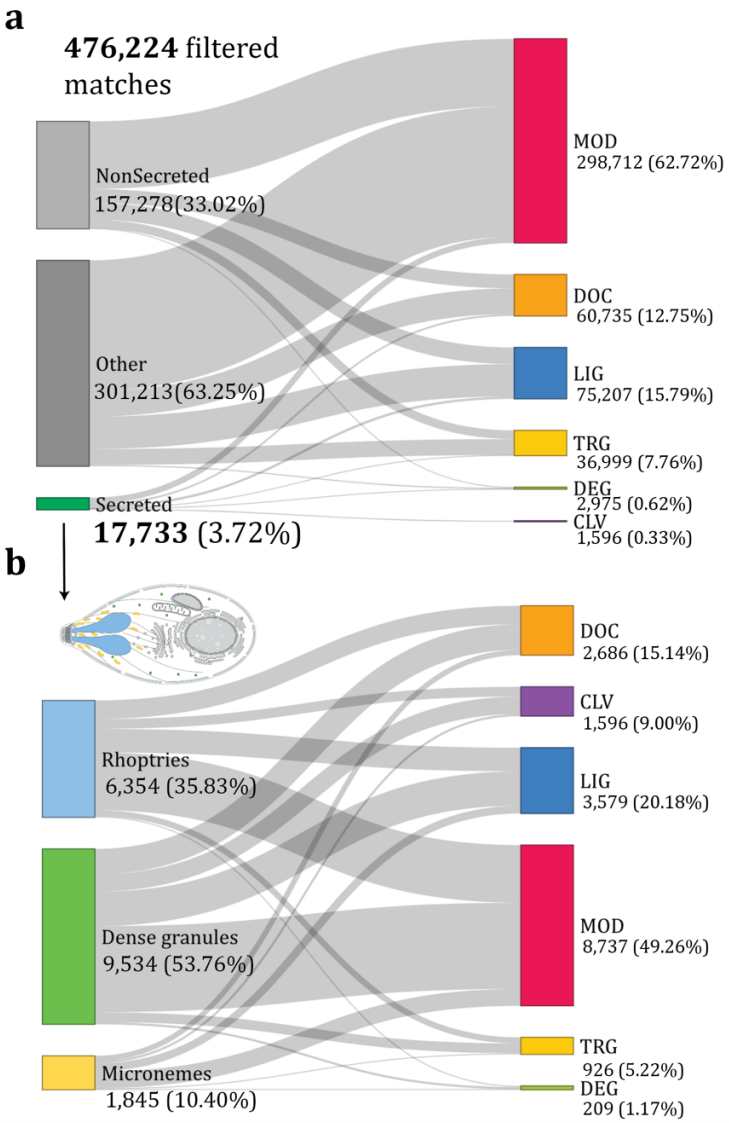


Figure 4.1 Breakdown of motif match types across secretion organelles. Sankey ploy of the different amount of motif matches in secretory organelles and their corresponding motif types **a**. compared with non-secreted proteins and in **b**. among themselves.

4.2. RGDs and integrin binding motifs

Integrin binding motifs are among the motifs that are functional in an extracellular context. Furthermore, Integrin binding is a common strategy for pathogens to become attached to the cell membrane. The most basic integrin binding motif is the RGD motif, which can bind to 8 of the different integrin α and integrin β heterodimers. ELM contains 5 motif classes of the RGD motif and its variants. In *Toxoplasma* we would expect to find RGD motifs in extracellular proteins with transmembrane regions, as well as in microneme proteins, that would aid the parasite in host cell attachment. As the RGD

motifs are short in nature, having only 3 key residues, and are usually present in loops between well-ordered domains, the evaluation of disorder might not be useful for these motifs. When analyzing the different matches and before filters we get 86 matches from 3 classes. Inspecting which ones are located in microneme proteins we get 6 matches for 2 classes in 6 different proteins **Table 4.1.**

Motif class name	Matches	Micronemes
LIG_INTEGRIN_RGD_1	**60**	3 (3)
LIG_INTEGRIN_ISODGR_2	**26**	3 (3)
LIG_INTEGRIN_RGD_TGFB_3	**1**	0
	86	6 (6)

Table 4.1 Integrin binding motif matches. Under the organelle column the number of matches is indicated together with their corresponding proteins in parenthesis.

5 out of these 6 RGD containing micronemes proteins have some experimental evidence. Toxolysin TLN4 is a zinc metalloproteinase localized within the micronemes which has a role in motility, invasion and parasite egress. It is a large protein that is processed multiple times before being secreted in a calcium-dependent manner. Its smaller fragments are though to remain bound in a larger complex (Laliberté & Carruthers, 2011). MIC12 contains several Epidermal Growth Factor (EGF) like domain repeats. M2AP is responsible for the correct trafficking of MIC2 (Liu et al., 2017). MIC8 is a transmembrane microneme protein that is exported to the cell surface upon an increase in intracellular calcium, it is essential for infection and it is involved in the formation of the moving junction (Kessler et al., 2008).

Besides TLN4, these motifs are present in all strain proteins, but seem to be absent in different species **Table 4.2**. Almost all the motifs seem to have high accessibility either by their AC score (for the ones with available AlphaFold structures) or as observed in their predicted structures **Figure 4.2** (also including predicted segments from ColabFold). The RGD motif in MIC12 is localized between a pair of EGF repeats, these are thought to be exposed as the protein adopts an extracellular extended form in a calcium-dependent manner. These motifs might represent another way for *Toxoplasma* proteins to interact with membranes of different host cell types, as integrins are expressed in many different tissues, or as a possible way to interact between themselves, MIC12 interacts with MIC2 which has integrin-like domains (Liu et al., 2017).

Sequence ID	Protein Description	Site	Context	Strains	Species	AC
LIG_Integrin_RGD_1				**RGD**		
TGME49_293770	Chitinase-like Protein CLP1	458	disorder	1.00 (4)	0	3.93
TGME49_206510	Toxolysin TLN4	1452	disorder	0.75 (3)	0	-
TGME49_267680	Microneme protein MIC12	555	order	1.00 (4)	0	-
LIG_Integrin_isoDGR_2				**NGR**		
TGME49_205680	Hypothetical protein	134	disorder	1.00 (4)	0.25 (1)	2.15
TGME49_214940	MIC2 associated protein M2AP	320	disorder	1.00 (4)	0	1.69
TGME49_245490	Microneme protein MIC8	247	order	1.00 (4)	0.5 (2)	0.59

Table 4.2 Microneme proteins containing integrin binding motif matches. Under the additional strains and species columns their presence proportion is indicated and in parenthesis their number.

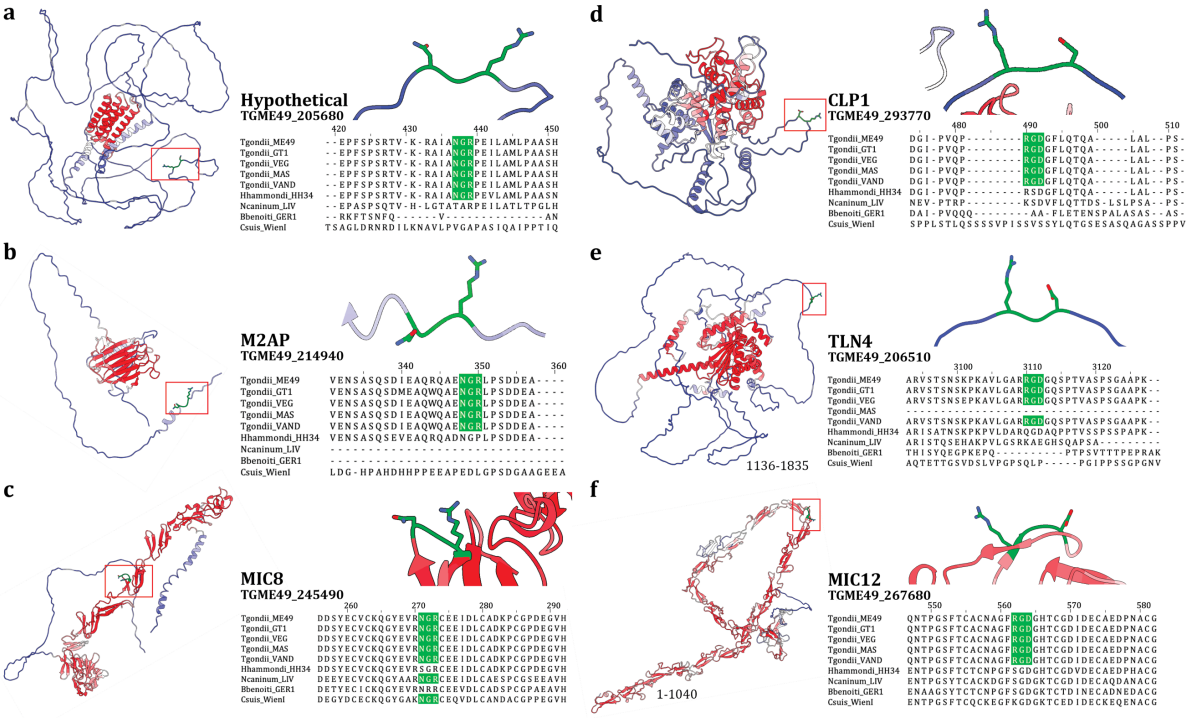


Figure 4.2 Microneme RGD candidate integrin binding motif matches. a-d. AlphaFold and **e-f**. ColabFold structure predictions, with their corresponding alignment motif presence. Red coloring in the structures represent high pLDDT values while Blue represents lower ones. The protein region numbering of amino acids regions is given to partial ColabFold models. Structure visualizations were produced in ChimeraX and motif alignments in Jalview.

4.3. PDZ signaling

PDZ domains are globular protein modules involved in various signaling and regulatory processes in the cell. They are mainly part of the membrane proteins and tend to form multicomponent complexes. They bind to specific motifs, usually located at the C-

terminal end of proteins. There are 5 different PDZ motif classes in the ELM database, through the filtered set of motif matches I obtained a total of 183 matches for 3 of them **Table 4.3.** These contain a total of 10 matches in 8 different Rhoptry and Dense Granule proteins, most of which 8 of them are labelled as hypothetical and two of them have been previously characterized **Table 4.4.**

Motif class name	Matches	Rhoptries	Dense granules
LIG_PDZ_CLASS_1	**5**	2 (2)	2 (2)
LIG_PDZ_CLASS_2	**83**	1 (1)	0
LIG_PDZ_CLASS_3	**106**	3 (3)	2 (2)
	183	6 (6)	4 (4)

Table 4.3 PDZ domain binding motif matches. Under the organelle columns the number of matches is indicated together with their corresponding proteins in parenthesis.

Most matches have a high presence in *Toxoplasma* strains and some are conserved in different species too. RON4 is a rhoptry protein involved in the formation of the Moving Junction, which has previously been reported to contain motifs that allow it to interact with cytoskeletal and membrane-associated proteins (described below) (Guérin et al., 2017). ROP15 is an active kinase that has been shown to be differentially expressed across *Toxoplasma* development (Wang et al., 2017). Neither have been reported to contain PDZ binding motifs, but the presence of the motif in other organisms (when not present they have small sequence variations) offers the potential to test this further **Figure 4.3**.

Sequence ID	Protein Description	Organelle	Strains	Species
LIG_PDZ_Class_1				
TGME49_211290	Rhoptry protein ROP15	Rhoptries	1.00 (4)	0.25 (1)
TGME49_229010	Rhoptry neck protein RON4		1.00 (4)	0.50 (2)
TGME49_203290	Hypothetical protein	Dense granules	1.00 (4)	0.50 (2)
TGME49_247440	Hypothetical protein		1.00 (4)	0.25 (1)
LIG_PDZ_Class_2				
TGME49_253100	Hypothetical protein	Rhoptries	1.00 (4)	0.50 (2)
LIG_PDZ_Class_3				
TGME49_229500	Hypothetical protein	Rhoptries	1.00 (4)	0.75 (3)
TGME49_230350	Hypothetical protein		0.75 (3)	0.50 (2)
TGME49_294630	Hypothetical protein		0.75 (3)	0.25 (1)
TGME49_202620	Hypothetical protein	Dense granules	1.00 (4)	0.25 (1)
TGME49_315910	Hypothetical protein		1.00 (4)	0.25 (1)

Table 4.4 Secretory organelle proteins containing PDZ domain binding motif matches. Under the additional strains and species columns their presence proportion is indicated and in parenthesis their number.

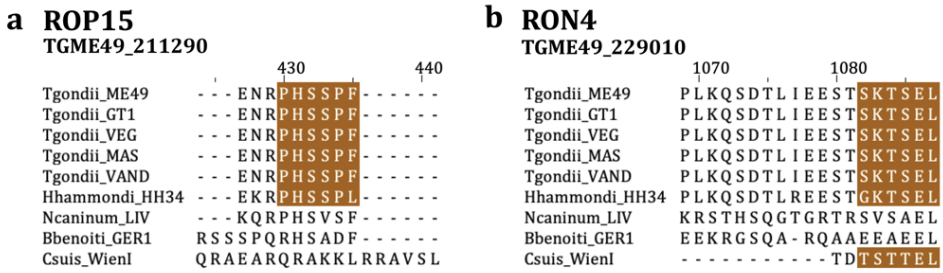


Figure 4.3 PDZ domain binding motif matches. a-b. Alignments of the two PDZ motif matches from characterized rhoptry proteins. Alignments produced in Jalview.

4.4. Nuclear targeting

Nuclear localization and export signals are small peptide regions present in nuclear proteins and some proteins that transverse through the nuclear pore complex. They are recognized by importin and exportin proteins. There are 6 targeting motif classes in the ELM database related to nuclear transport and, through the filtered set of motif matches, I obtained 5,989 overall matches across different *Toxoplasma* proteins **Table 4.5**. I would expect that most of the matches would be present in the parasite's own nuclear proteins. Nevertheless, it would be less clear why would there be present in protein of secretion organelles. In fact, I obtained 63 matches in 26 rhoptry proteins and 118 matches in 40 dense granule proteins. Some of these proteins contain multiple instances of the same or different motifs **Table 4.6,** up to 18 matches in the case of the hypothetical rhoptry protein TGME49_304720.

Motif class name	Matches	Rhoptries	Dense granules
TRG_NLS_BIPARTITE_1	**814**	4 (4)	12 (10)
TRG_NLS_MONOCORE_2	**1,151**	10 (10)	14 (10)
TRG_NLS_MONOEXTC_3	**2,070**	20 (16))	35 (26)
TRG_NLS_MONOEXTN_4	**2,804**	28 (21)	43 (28)
TRG_NES_CRM1_1	**252**	1 (1)	7 (6)
LIG_NRBOX	**6**	-	7 (4)
	7,097	63 (26)	118 (40)

Table 4.5 Nuclear targeting motif matches. Under the organelle columns the number of matches is indicated together with their corresponding proteins in parenthesis.

There are several other characterized proteins that contain nuclear localization signals **Table 4.7**. Even though we can say that they are in an accessible context for interaction and that they are present in different organisms, not all of them might be functional. From

the ones that can be functional we have ROP16, a rhoptry kinase already known to be imported to the nucleus via NLS and subvert gene transcriptional programs. Specifically it has been tested that ROP16 phosphorylates STAT3 and STAT6 proteins in order to inhibit cell death, these proteins are also associated with IL-4 and IL-6 production (Saeij et al., 2007; Zhu et al., 2019). I also found matches in GRA6, a dense granule protein known to interact with NFAT4 (Nuclear Factor of Activated T cells 4) and with calcineurin activator calcium-modulating ligand (CAMLG), but so far it has only been observed to be localized in the PV (Ma et al., 2014).

Sequence ID	Protein Description	Matches	Motif class name
Rhoptries			
TGME49_242820	Hypothetical protein	1	TRG_NLS_MonoCore_2
		2	TRG_NLS_MonoExtC_3
		2	TRG_NLS_MonoCore_2
TGME49_225200	Hypothetical protein	1	TRG_NLS_Bipartite_1
		1	TRG_NLS_MonoCore_2
		2	TRG_NLS_MonoExtC_3
		5	TRG_NLS_MonoExtN_4
Dense granules			
TGME49_304720	Hypothetical protein	2	TRG_NES_CRM1_1
		4	TRG_NLS_MonoCore_2
		4	TRG_NLS_MonoExtC_3
		5	TRG_NLS_MonoExtN_4
		1	LIG_NRBOX
TGME49_282170	Hypothetical protein	1	TRG_NLS_Bipartite_1
		1	TRG_NLS_MonoCore_2
		3	TRG_NLS_MonoExtC_3
		3	TRG_NLS_MonoExtN_4
TGME49_215360	Hypothetical protein	1	TRG_NLS_Bipartite_1
		2	TRG_NLS_MonoCore_2
		2	TRG_NLS_MonoExtC_3
		2	TRG_NLS_MonoExtN_4
TGME49_304955	Serine/threonine specific protein phosphatase	2	TRG_NLS_Bipartite_1
		1	TRG_NLS_MonoExtC_3
		2	TRG_NLS_MonoExtN_4
TGME49_262400	Lipase	1	TRG_NLS_Bipartite_1
		1	TRG_NLS_MonoCore_2
		1	TRG_NLS_MonoExtC_3
		2	TRG_NLS_MonoExtN_4
TGME49_288000	Hypothetical protein	2	TRG_NLS_MonoExtC_3
		3	TRG_NLS_MonoExtN_4
TGME49_247440	Hypothetical protein	1	TRG_NLS_Bipartite_1
		1	TRG_NLS_MonoCore_2
		1	TRG_NLS_MonoExtC_3
		2	TRG_NLS_MonoExtN_4
TGME49_217680	Hypothetical protein	2	TRG_NLS_Bipartite_1
		3	TRG_NLS_MonoExtN_4

Table 4.6 Secretory organelle proteins containing the most nuclear targeting motif matches. Top 10 proteins from secretory organelles containing at least 5 nuclear targeting motif matches.

There are other types of secreted protein that have been characterized and for which the presence of NLS might not be functional **Table 4.7**. ROP5 is a pseudokinase related to disease virulence that localizes to the PV and is known to interact with ROP18 and GRA7. GRA7 is a transmembrane protein that localizes to the PV membrane where it interacts with IRG (immune related GTPases) and other ROP proteins (Hakimi et al., 2017). It has a NLS but has a low potential to be transported into the nucleus. Other RON proteins like RON1, RON2, RON4, RON5 and RON8 are known to localize to the MJ and the host cytoskeleton, so the presence of NLS would not be functionally relevant (Guérin et al., 2017). Some other *Toxoplasma* proteins known to localize to the nucleus like GRA24 and GRA16 might not need to have these signals to be transported into the nucleus, instead they are associated with host proteins that have these signals, like p38 MAPK and HAUSP, and then be transported as a complex.

Sequence ID	Protein Description	Sequence ID	Protein Description
Rhoptries		**Dense granules**	
TGME49_305590	ABC transporter transmembrane region domain containing protein	TGME49_227280	Dense granule protein GRA3
TGME49_300100	Rhoptry neck protein RON2	TGME49_228170	Inner membrane complex protein IMC2A
TGME49_243730	Rhoptry protein ROP9	TGME49_209755	MAG2
TGME49_269885	Rhoptry metalloprotease toxolysin TLN1	TGME49_262400	Lipase
TGME49_299060	Sodium/hydrogen exchanger NHE2	TGME49_264660	SRS44/CST1
TGME49_261440	ARM repeats containing protein	TGME49_304955	Serine/threonine specific protein phosphatase
TGME49_236860	Haloacid dehalogenase family hydrolase domain containing protein	TGME49_320490	N acyl phosphatidyl-ethanolamine hydrolysing phospholipase D family protein
TGME49_229010	Rhoptry neck protein RON4	TGME49_208070	Inositol polyphosphate kinase
TGME49_310010	Rhoptry neck protein RON1	TGME49_208450	Protease inhibitor PI2
TGME49_262730	**Rhoptry protein ROP16**	TGME49_203310	Dense granule protein GRA7
TGME49_291960	Rhoptry kinase family protein ROP40 incomplete catalytic triad	TGME49_237500	Protein phosphatase 2C domain containing protein
TGME49_306060	Rhoptry neck protein RON8	TGME49_275440	Dense granule protein GRA6
TGME49_308090	Rhoptry protein ROP5	TGME49_310780	Dense granule protein GRA4

Table 4.7 Characterized secretory organelle proteins containing nuclear targeting motif matches. Previously characterized motifs in bold.

4.5. Phosphomotifs

By having information on the phosphorylation of motifs we can infer the functionality of phospho-motifs, motifs that require this PTM to be recognized by its binding partners or which binding potential is regulated by it. There are around 18 phospho-motifs in ELM and through the filtered set of motif matches I obtained more than 100,000 matches for 16 of them Table 4.8. Out of ~8,000 matches, just 8.3% have evidence for being modified. There was a total of 129 motif matches located in proteins from secretion organelles. 50 matches were located in 23 rhoptry proteins, including 5 RONs and 5 ROPs, while there were 79 matches in 35 dense granule proteins, including 6 GRA proteins. There was a total of 30 hypothetical proteins, 8 from rhoptries and 22 in dense granules.

Motif class name	Filtered	+ PTM	Proportion	Rhoptries	Dense granules
Degrons					
DEG_SCF_FBW7_1	803	92	0.113	-	-
DEG_SCF_FBW7_2	246	24	0.089	-	-
DEG_SCF_TRCP1_1	7	2	0.286	1 (1)	1 (1)
Docking					
DOC_AGCK_PIF_1	18	1	0.059	-	-
DOC_CKS1_1	3,124	272	0.084	-	1 (1)
DOC_WW_Pin1_4	41,583	4,444	0.110	18 (11)	31 (16)
DOC_PP2A_B56_1	1,728	40	0.026	-	-
Ligand					
LIG_14-3-3_CanoR_1	20,458	1,925	0.096	12 (11)	20 (16)
LIG_14-3-3_CterR_2	69	1	0.017	-	-
LIG_BRCT_BRCA1_1	6,324	397	0.059	1 (1)	4 (4)
LIG_BRCT_BRCA1_2	104	8	0.070	-	-
LIG_FHA_1	12,534	581	0.046	7 (6)	5 (5)
LIG_FHA_2	16,000	801	0.050	10 (10)	15 (9)
LIG_GSK3_LRP6_1	3	0	0	-	-
LIG_PTB_Phospho_1	4	0	0	-	-
LIG_TYR_ITIM	3	0	0	-	-
TOTAL=16	**103,008**	**8,588**	**0.083**	50 (23)	79 (35)

Table 4.8 Phosphomotifs matches from different classes. Under the organelle columns the number of matches is indicated together with their corresponding proteins in parenthesis.

4.6. The ESCRT membrane remodeling system

The ESCRT system is mainly known to be a cargo-recognition and membrane-deformation machine. It is involved in different biological processes such as the formation of multivesicular bodies, cell abscission and viral budding, exosome secretion and autophagy (Henne et al., 2011). Structurally it is comprised of different complexes termed ESCRT-0, -I, -II, III and Vps4. Linear motifs play an important role in how the

different ESCRT complexes interact with each other and with other proteins. The PTAP motif allows ESCRT-0 proteins to interact with the UEV domain of TSG101 of the ESCRT-I complex whereas the YPxL motifs link ESCRT-III to the adaptor protein Alix, which in turn stabilizes ESCRT-III filaments and recruits deubiquitinating enzymes (Henne et al., 2011). There are 6 different ESCRT system related motif classes in the ELM database, through the filtered set of motif matches I obtained a total of 91 matches for 5 of them **Table 4.9.** These included a total of 14 matches in 13 different Rhoptry and Dense Granule proteins. It is relevant to highlight that from the 4 motif classes the PTAP motif is not valid within *Toxoplasma* cells as the UEV domain that binds to the motif is absent from its proteome, so it should be worth exploring the potential functionality from these matches in proteins in secreted organelles.

Motif class name	Matches	Rhoptries	Dense granules
LIG_PTAP_UEV_1	**10**	2 (2)	6 (6)
LIG_LYPXL_S_1	**43**	4 (3)	1 (1)
LIG_LYPXL_L_2	**1**	-	-
LIG_LYPXL_SIV_4	**12**	-	1 (1)
DOC_MIT_MIM_1	**25**	-	-
	91	6 (5)	8 (7)

Table 4.9 ESCRT system related motif matches. Under the organelle column the number of matches is indicated together with their corresponding proteins in parenthesis.

Similarly to previous motif match groups most of the proteins the ESCRT system related motifs localized in are labelled as hypothetical **Table 4.10** but some have been previously characterized: for example RON4 and GRA14 have been already reported to contain these motifs in their sequences (Guérin et al., 2017; Rivera-Cuevas et al., 2021). The presence of the LYPxL motif in the RON8 protein might have a similar role to that of RON4 as both are localized to the MJ. ROP13 is a rhoptry protein known to be exported to the host cytosol. ROP13 deletion results in a small *Toxoplasma* growth defect whereas overexpression is toxic to the host cell. Even though its host interactors are unknown, there is little evidence that the presence of the PTAP motif in ROP13 would allow it to associate with membrane remodeling as it is mainly soluble (Turetzky et al., 2010).

Of the dense granule proteins, GRA15 is also a highlight as it is known to associate with the PVM and to interact with NF-κB (more in Chapter 5). On the other hand, GRA14 stands out as the only protein from my filtered match list to contain both PTAP and LYPxL motifs

Figure 4.4. It is a transmembrane protein that associates with the intravacuolar network (IVN) and localizes onto structures termed BOAS (beads-on-a-string). The IVN is an elaborate network of membranous nanotubules that form multiple lumens in the PV, while BOAS seem to form connections between different PV in the same host cell or between the PV and the host nucleus (Rastogi et al., 2019). The two motifs are present in the GRA14 C-terminus, which faces the cytosol (Rome et al., 2008) Recent work by (Rivera-Cuevas et al., 2021) showed that the GRA14 C-terminus containing the motifs was enough to produce exosomes in a similar manner as the HIV gag protein, and in its absence, *Toxoplasma* ability to sequester host cytosolic contents was also reduced. This work proposed a role for GRA14 as the recruiter of the ESCRT system to aid in the creation of the endosome at the PVM that will sequester cytosolic contents and then carry them to the plasma membrane.

Sequence ID	Protein description	Matches	Motif class name
Rhoptries			
TGME49_229010	**Rhoptry neck protein RON4**	**2**	
TGME49_279420	Hypothetical protein	1	LIG_LYPXL_S_1
TGME49_306060	Rhoptry neck protein RON8	1	
TGME49_203990	Rhoptry protein ROP12	1	LIG_PTAP_UEV_1
TGME49_312270	Rhoptry protein ROP13	1	
Dense granules			
TGME49_239740	**Dense granule protein GRA14**	**1** **1**	**LIG_LYPXL_S_1** **LIG_PTAP_UEV_1**
TGME49_294970	Hypothetical protein	1	
TGME49_275470	GRA15	1	
TGME49_203290	Hypothetical protein	1	LIG_PTAP_UEV_1
TGME49_304720	Hypothetical protein	1	
TGME49_247440	Hypothetical protein	1	
TGME49_270240	MAG1	1	LIG_LYPXL_SIV_4

Table 4.10 Secretory organelle proteins containing ESCRT system related motif matches. Previously characterized motifs in bold.

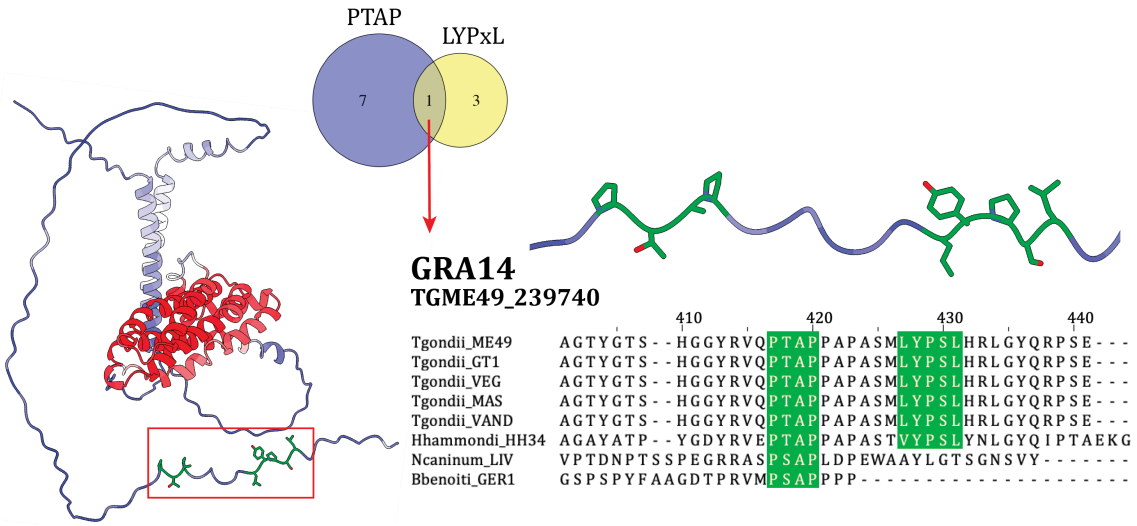


Figure 4.4 GRA14 ESCRT related motif matches. ESCRT motif match overlap between proteins from secretion organelles. AlphaFold structure prediction of GRA14, with corresponding alignment motif presence. Red coloring in the structures represent high pLDDT values while Blue represents lower ones. Structure visualization was produced in ChimeraX and the motif alignments in Jalview.

4.5. Parasite entry and the host cytoskeleton

As mentioned in the introduction, *Toxoplasma* is able to enter host cells via a MJ, a multicomponent protein complex that is linked to the host cytoskeleton through different short linear motifs (Guérin et al., 2017). This makes it interesting to look for further motif instances in secreted proteins that can give us a perspective of which ones are potentially used during infection. In the ELM database there are around 20 motif classes that mediate the interaction with proteins from the cytoskeleton, some specific to the actin filaments and others with microtubules. Using the list of filtered matches, I identified 661 matches related to actin regulation from which a total of 43 are located in proteins from secretory organelles **Table 4.11**. I identified 2,184 related to microtubules, from which a total of 63 matches reside in 48 proteins from rhoptries and dense granules **Table 4.11.**

As in the case of the proteins of the MJ, these motifs are often found in together in the same cellular protein, so it made sense to look for linked pairs that would indicate a functional overlap and further back the role of the matches together. I found 27 instances of paired overlaps between motifs related to interactions with cytoskeletal proteins, some occurring specifically in proteins from secretory organelles **Table 4.12**. Most of these overlaps happen together with the WIRS, PP2A, PxPxPR and WH2 motifs.

Motif class name	Matches	Rhoptries	Dense granules
ACTIN			
LIG_ACTIN_RPEL_3	**5**	-	-
LIG_ACTIN_WH2_1	**1**	-	-
LIG_ACTIN_WH2_2	**542**	7 (5)	4 (4)
LIG_ACTINCP_TWFCPI_2	**34**	1 (1)	-
LIG_IBAR_NPY_1	**1**	1 (1)	-
LIG_PROFILIN_1	**47**	-	-
LIG_SH3_CIN85_PXPXPR_1	**10**	8 (5)	1 (1)
LIG_WRC_WIRS_1	**21**	9 (7)	12 (8)
TOTAL	**661**	26 (17)	17 (12)
MICROTUBULES			
DEG_APCC_TPR_1	**13**	-	-
DOC_PP2A_B56_1	**1,728**	20 (18)	27 (20)
LIG_CAP-GLY_1	**5**	1 (1)	-
LIG_DYNEIN_DLC8_1	**37**	-	1 (1)
LIG_FAT_LD_1	**2**	2 (2)	-
LIG_KLC1_WD_1	**314**	2 (2)	3 (3)
LIG_KLC1_YACIDIC_2	**2**	-	2 (2)
LIG_SXIP_EBH_1	**83**	3 (2)	1 (1)
TOTAL	**2,184**	28 (23)	35 (25)

Table 4.11 Actin and microtubule related motif matches. Under the organelle column the number of matches is indicated together with their corresponding proteins in parenthesis.

Motif 1	Motif 2	Protein overlaps	Rhoptries	Dense granules
Actin				
LIG_Actin_WH2_2	LIG_WRC_WIRS_1	**2**	1	1
LIG_SH3_CIN85_PxpxPR_1	LIG_WRC_WIRS_1	**1**	1	-
Microtubules				
DOC_PP2A_B56_1	LIG_KLC1_WD_1	**2**	1	1
DOC_PP2A_B56_1	LIG_KLC1_Yacidic_2	**1**	-	1
LIG_SxIP_EBH_1	DOC_PP2A_B56_1	**2**	1	1
Actin & microtubules				
LIG_Actin_WH2_2	LIG_SxIP_EBH_1	**1**	1	-
LIG_Actin_WH2_2	DOC_PP2A_B56_1	**4**	3	1
LIG_Actin_WH2_2	LIG_FAT_LD_1	**1**	1	-
LIG_ActinCP_TwfCPI_2	DOC_PP2A_B56_1	**1**	1	-
LIG_SH3_CIN85_PxpxPR_1	LIG_SxIP_EBH_1	**1**	1	-
LIG_SH3_CIN85_PxpxPR_1	DOC_PP2A_B56_1	**1**	-	1
LIG_SH3_CIN85_PxpxPR_1	LIG_KLC1_WD_1	**1**	-	1
LIG_SH3_CIN85_PxpxPR_1	LIG_FAT_LD_1	**1**	1	-
LIG_WRC_WIRS_1	LIG_KLC1_WD_1	**1**	1	-
LIG_WRC_WIRS_1	LIG_KLC1_Yacidic_2	**1**	-	1
LIG_WRC_WIRS_1	DOC_PP2A_B56_1	**6**	2	4
		27	13	6

Table 4.12 Pairwise protein overlap between motif cytoskeleton related motif classes. The overlaps indicate how many proteins have motif matches from different cytoskeleton related classes.

There are 6 proteins that contained the larger number of different motifs, 3 from rhoptries and 3 from dense granules **Table 4.13** & **Figure 4.5**. Some of these matches were located in RON4 and have been previously characterized together with the ones for RON2 and RON5 (Guérin et al., 2017), but in that work the first PxPxPR and ESCRT-related LYPxL motifs in RON4 were missed as well as all the instances of the SxIP motif **Figure 4.5.c**. In the cell, the SxIP motif is present in different microtubule-associated proteins, like the +TIP (microtubule plus-end tracking) proteins, and mediates their interaction with EBH (end binding homology) domain containing proteins, like EB1 (Honnappa et al., 2009). Its presence in RON4 and RON8 further support its role in linking the MJ with the host cytoskeleton. If proven to be functional these motifs would complement the previously proposed models. On the other hand, the presence of multiple related motifs in the other 4 uncharacterized proteins will also advance the view of how *Toxoplasma* uses the host cytoskeleton for invasion. This might be a general strategy for other parasites (Havrylov et al., 2010).

4.6. Ubiquitin proteasome system

The homeostasis of protein levels in the cell is a complex and highly regulated process. The ubiquitin proteasome system (UPS) is responsible for degrading and recycling most of the cellular proteins. Protein degradation via the 26S proteasome depends on the correct recognition of target proteins with proper degradation signals. From these signals, the polyubiquitination of proteins is the canonical one. E3 ligases are the enzymes responsible for adding polyubiquitin chains to proteins, they recognize their substrates via specific degradation motifs or Degrons (Mészáros et al., 2017). Degrons can be subverted during infection in order to disrupt degradation processes and extend the lifetime of proteins from parasites, as has been observed in viruses (Davey et al., 2011). Thus, finding examples of them in secreted proteins could provide evidence for this strategy in *Toxoplasma*. ELM counts with different motif classes describing degrons from which 11 classes are specifically related to proteasomal targeting. Through my pipeline I obtained a total of 2,206 motif matches from 7 of those classes **Table 4.14**. From these 14 motif matches corresponded to 154 rhoptry proteins and 36 to 28 dese granule proteins, all from 5 different classes.

Sequence ID	Protein Description		Motif class name	Strains	Species	Site
Rhoptries						
TGME49_306060	Rhoptry neck protein RON8	1	LIG_Actin_WH2_2	1.00 (4)	0.50 (2)	1,615
		1	DOC_PP2A_B56_1	1.00 (4)	0.50 (2)	1,544
		1	LIG_SxIP_EBH_1	1.00 (4)	0.50 (2)	2,080
		1	LIG_LYPXL_S_1	1.00 (4)	1.00 (4)	729
TGME49_279420	Hypothetical protein	2	LIG_Actin_WH2_2	1.00 (4)	0.50 (2)	436
				1.00 (4)	0.25 (1)	1,179
		1	LIG_FAT_LD_1	1.00 (4)	0.75 (3)	142
		1	LIG_LYPXL_S_1	1.00 (4)	1.00 (4)	1,014
TGME49_229010	Rhoptry neck protein RON4	4	LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	0.25 (1)	28
				1.00 (4)	**0.50 (2)**	**114**
				1.00 (4)	**0.25 (1)**	**130**
				1.00 (4)	**0.25 (1)**	**259**
		2	LIG_SxIP_EBH_1	1.00 (4)	0.25 (1)	156
				1.00 (4)	0.25 (1)	285
		3	LIG_LYPXL_S_1	1.00 (4)	0.25 (1)	34
				1.00 (4)	**0.50 (2)**	**169**
				1.00 (4)	**0.75 (3)**	**298***
Dense granules						
TGME49_288000	Hypothetical protein	1	LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	0.25 (1)	57
		1	DOC_PP2A_B56_1	1.00 (4)	0.50 (2)	498
		1	LIG_KLC1_WD_1	1.00 (4)	0.50 (2)	1,431
TGME49_304720	Hypothetical protein	1	LIG_WRC_WIRS_1	1.00 (4)	0.50 (2)	1,214
		3	DOC_PP2A_B56_1	1.00 (4)	0.75 (3)	978
				1.00 (4)	0.50 (2)	1,038
				0.75 (3)	0.25 (1)	5,555
		1	LIG_KLC1_Yacidic_2	0.75 (3)	0.75 (3)	5,965
		1	LIG_PTAP_UEV_1	1.00 (4)	0.25 (1)	2,032
TGME49_304955	Serine/threonine specific protein phosphatase	1	LIG_Actin_WH2_2	1.00 (4)	1.00 (4)	2,110
		4	LIG_WRC_WIRS_1	0.75 (3)	0.75 (3)	455
				1.00 (4)	0.25 (1)	1,432
				1.00 (4)	0.75 (3)	2,065
				1.00 (4)	0.25 (1)	2,351
		1	DOC_PP2A_B56_1	1.00 (4)	1.00 (4)	2,264

Table 4.13 Secretory organelle proteins containing ESCRT system related motif matches. Previously characterized motifs in bold. * Motif match was not present in the filtered table due to a low AC score but is well conserved, in a disorder context and has been previously tested. Under the additional strains and species columns their presence proportion is indicated and in parenthesis their number.

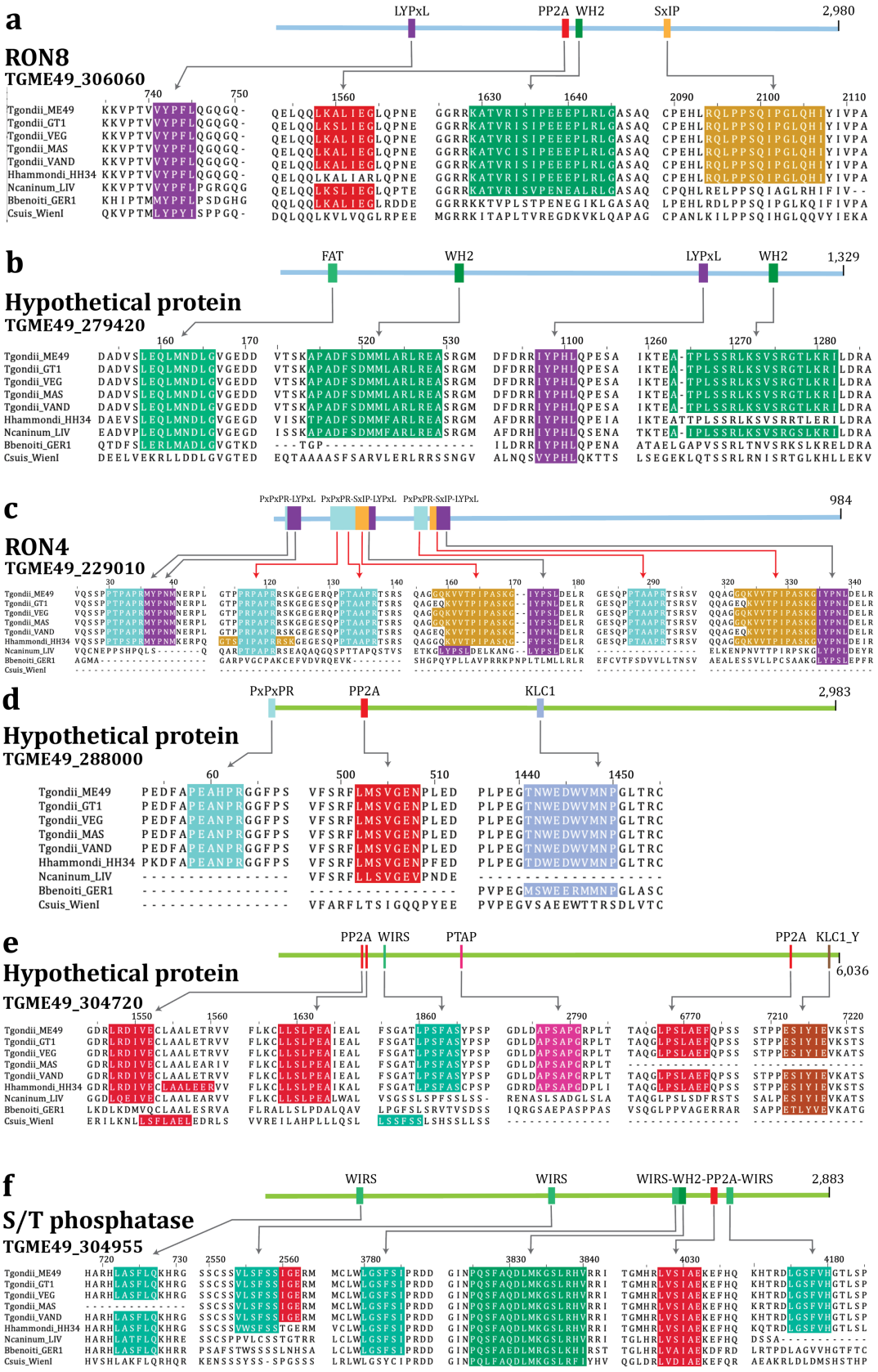


Figure 4.5 Cytoskeletal related motif matches. a-f. Diagram and alignments of different motifs in proteins from secretory organelles. Previously characterized motif instances in red. Alignments produced in Jalview.

Class name	Matches	Rhoptries	Dense granules
DEG_APCC_DBOX_1	**1,504**	9 (9)	20 (18)
DEG_APCC_KENBOX_2	**269**	-	5 (5)
DEG_APCC_TPR_1	**13**	-	-
DEG_COP1_1	**3**	-	2 (2)
DEG_KELCH_ACTINFILIN_1	**3**	-	-
DEG_KELCH_KEAP1_1	**6**	2 (2)	3 (3)
LIG_APCC_ABBA_1	**408**	3 (3)	6 (5)
	2,206	14 (14)	36 (28)

Table 4.14 Proteasome related motif matches. Under the organelle column the number of matches is indicated together with their corresponding proteins in parenthesis.

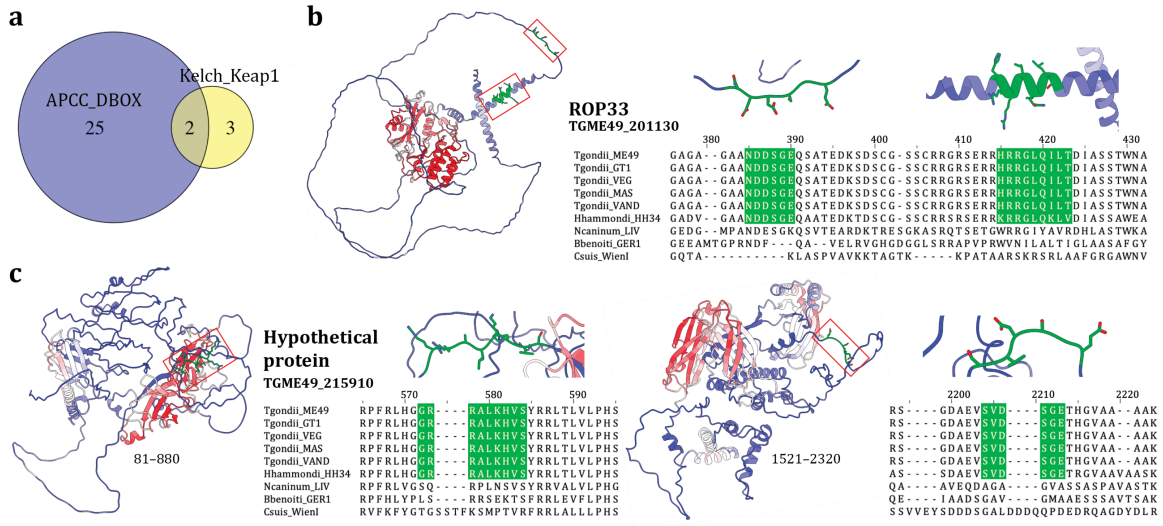


Figure 4.6 Proteins containing proteasome motif matches. a Proteasome motif match overlap between secretory proteins **b**. AlphaFold and **c**. ColabFold structure predictions, with their corresponding alignment motif presence. Red coloring in the structures represent high pLDDT values while Blue represents lower ones. The protein region numbering is given to partial ColabFold predicted models. Structure visualizations were produced in ChimeraX and motif alignments in Jalview.

From the 279 motif classes used in the pipeline I obtained matches for 263 of them, which meant 16 classes did not contain any motif match. There were numerous matches for the PEXEL motif (4,614), but their presence should be paired with a correct positioning close to the N-terminus and a possession of a SP (signal peptide). These considerations reduce the number of matches considerably. My pipeline was also able to recapitulate previously identified motifs with experimental evidences **Table 4.15,** and in some cases adds further instances or complement them with motif of related functions (Chapter 4.5). This

set of true instances offers a small control to the effectiveness of the pipeline. From the 14 ELM instances the pipeline was able to recapitulate 8. For the case of matches in RON5, the three instances were missed due to the presence score in other species and the IUPred, for the RON4 a LYPxL motif was missed due to its AC score. In the case of the GRA24 motif, which is actually based on the *Toxoplasma* dense granule, both instances were not present in other Sarcocystidae species. In reality this might hint that the REGEX expression is too restrictive, it has a fixed Gly, and could be reevaluated considering a better alignment. Overall, I found many interesting motif candidates that could be taken to carry out motif-domain binding experiments to complement the secreted proteins functions or assign one to hypothetical proteins. In fact, I identified 901 hypothetical proteins containing filtered matches, these offer a venue to explore and understanding their function.

Sequence ID	Protein Description	Motif Class	Strains	Species	IUPred	AC
Rhoptries						
TGME49_311470	Rhoptry neck protein RON5	LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	**0**	0.745	-
		LIG_WW_1	1.00 (4)	0.75 (3)	0.952	-
		LIG_PTAP_UEV_1	1.00 (4)	0.25 (1)	**0.365**	-
		LIG_PTAP_UEV_1	1.00 (4)	1.00 (4)	**0.387**	-
TGME49_300100	Rhoptry neck protein RON2	LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	0.25 (1)	0.787	-
TGME49_229010	Rhoptry neck protein RON4	LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	0.50 (2)	0.976	2.60
		LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	0.25 (1)	0.968	2.09
		LIG_SH3_CIN85_PxpxPR_1	1.00 (4)	0.50 (2)	0.959	2.74
		LIG_LYPXL_S_1	1.00 (4)	0.50 (2)	0.870	1.75
		LIG_LYPXL_S_1	1.00 (4)	0.75 (3)	0.702	**0.59**
Dense Granules						
TGME49_239740	Dense granule protein GRA14	LIG_LYPXL_S_1	1.00 (4)	0.25 (1)	0.557	1.97
		LIG_PTAP_UEV_1	1.00 (4)	0.75 (3)	0.613	1.77
TGGT1_230180*	Dense granule protein GRA24	DOC_MAPK_GRA24_9	0.75 (3)	**0**	0.611	1.68
		DOC_MAPK_GRA24_9	0.75 (3)	**0**	0.710	1.54

Table 4.15 Pipeline scores for *Toxoplasma gondii* motif instances in ELM. Under the additional strains and species columns their presence proportion is indicated and in parenthesis their number. In bold the values that prevented the instance to be captured in the pipeline. *Instance from *T. gondii* GT1 strain but with the scores of the ME49 strain.

CHAPTER 5
Motif candidate experimental validation

5.1. Selection of assays and candidates

Dr. Toby Gibson and I decided to focus on the LIG type motif LIG_TRAF6_MATH_1 to test the potential binding of some of the different sets of candidates produced by my pipeline. There is a range of different general and specific experimental methods to test each type of motifs (Gibson et al., 2015), so we sought to carry a combination of *in vitro* binding assays and mutational analysis as they were more suited to this type of motif. We then joined efforts with the European Molecular Biology Laboratory (EMBL) Protein expression and purification core facility (PEPCore) staff to carry initial testing of the binding affinity of peptides containing candidate motifs and their binding domain.

TRAF6 is involved in different cellular processes

The **LIG_TRAF6_MATH_1** motif mediates the interaction with the TRAF-C domain of the TRAF6 protein (also known as MATH, Meprin And TRAF-Homology domain PFAM: PF00917). It is member of the Tumor Necrosis Factor (TNF) receptor associated factor (TRAF) protein family and is involved in the canonical activation of the transcription factor NF-κB in the innate immune response. TRAF6 has other physiological roles in the differentiation of cell types involved in bone homeostasis, lymph node development, T-cell maturation, and the homeostasis of Schwann and glial cells (Yamamoto et al., 2021). Most of the proteins of the TRAF family share a conserved C-terminal TRAF-C domain, from which TRAF6 has the most divergent one displaying different substrate binding specificity. The E3 ligase domain of TRAF6 mediates the addition of specific Lys63-linked polyubiquitin chains which serves as a signal for interaction and not for proteasomal destruction. TRAF6 recognizes its interactors through a linear motif which is present in CD40 proteins (the Tumor necrosis factor protein CD40) and IL-1R (interleukin 1 receptor) (Z. Shi et al., 2015). Viruses are known to use TRAF6 motif containing proteins

to stimulate NF-κB signaling (Heinemann et al., 2006), while others like the bovine herpesvirus can even use them to target TRAF6 for proteasomal degradation (J.-H. Shi & Sun, 2018). Both ways exemplified how this motif can be subverted to modulate the immune response.

The binding properties of the TRAF6 motif

The TRAF6 motif adopts a secondary structure upon binding. It adds a β strand to its binding domain in a process termed β-augmentation which imparts rigidity to the final conformation of the complex. The core TRAF6 motif was firstly described as **PxExx[DE(Ar)]**, where x denotes any amino acid and Ar an aromatic one (Darnay et al., 1999). The first and third position are the most conserved having a Proline and Glutamic acid respectively. The +1 Pro in the motif has been shown to interact with Phe471 and Tyr473 of the human TRAF6, while the +3 Glu establishes a H-bond with Ala458. The sixth position of the motif is more variable having either a negatively charged amino acid or an aromatic one. It has been observed that each type of +6 residue displays a different mode of interaction but both are important for the overall binding of the motif (Huang et al., 2018; Ye et al., 2002). The non-conserved residues in the motif are variable. The ones in the middle cannot be Pro because they would prevent β-augmentation, as their ring structure is too bulky and they lack the peptide NH group to make the H-bond. The non-conserved positions at both sides of the motif can be more flexible. The ELM **LIG_TRAF6_MATH** model captures these properties in the following REGEX:

```
..P[^P]E[^P].[FYWHDE].
```

TRAF6 motif candidates in Toxoplasma proteins

The ELM database contains 4 distinct TRAF motif classes, all of which yield matches in the *Toxoplasma* proteins. From my filtered results, I identified a total of **370** motif matches from these four classes, of which 154 were in 52 rhoptry proteins and 167 in 61 dense granule proteins **Table 5.1**. For the TRAF6 motif I retrieved a total of **55** motif matches between proteins from both secretory organelles. Most of the motifs were present in previously characterized proteins **Table 5.2**, but also in a total of 15 proteins labelled as hypothetical.

Motif class name	Matches	Rhoptries	Dense granules
LIG_TRAF2_1	**287**	127 (48)	124 (55)
LIG_TRAF2_2	**3**	1 (1)	2 (2)
LIG_TRAF4_MATH_1	**13**	3 (3)	9 (8)
LIG_TRAF6_MATH_1	**67**	23 (15)	32 (23)
	370	154 (52)	167 (61)

Table 5.1 NF-κB signaling related motif matches. Under the organelle column the number of matches is indicated together with their corresponding proteins in parenthesis.

Sequence ID	Protein Description		Sequence ID	Protein Description	
Rhoptries			**Dense granules**		
TGME49_297960	**Rhoptry neck protein RON6***	**17**	TGME49_264660	SRS44/CST1	3
TGME49_258660	Rhoptry protein ROP6	3	TGME49_304955	serine/threonine specific protein phosphatase	3
TGME49_310010	Rhoptry neck protein RON1	2	TGME49_217680	Hypothetical protein	3
TGME49_261750	**Rhoptry neck protein RON10**	**2**	TGME49_202780	Rhoptry kinase family protein ROP25**	2
TGME49_305590	ABC transporter transmembrane region domain containing protein	2	TGME49_306890	Hypothetical protein	2
TGME49_235130	Transmembrane protein	2	**TGME49_203310**	**Dense granule protein GRA7**	**1**
TGME49_300100	Rhoptry neck protein RON2	1	**TGME49_275470**	**GRA15**	**1**
TGME49_229010	Rhoptry neck protein RON4	1	TGME49_208450	Protease inhibitor PI2	1
TGME49_311470	Rhoptry neck protein RON5	1	TGME49_240090	Rhoptry kinase family protein ROP34, putative**	1
TGME49_308810	Rhoptry neck protein RON9	1	TGME49_269920	Phosphatidylserine decarboxylase	1

Table 5.2 Characterized secretory organelle proteins containing the most NF-κB signaling related motif matches. Protein selected for further experimental testing in bold. * Protein included due to its high number of matches but not present in filtered results. ** Protein located in dense granules in HyperLOPIT experiment.

From the list of TRAF6 motif candidates, I selected a series of candidates for further inspection. As the TRAF6 motif was known to appear multiple times in proteins, RON6 became an appealing candidate as it contains 17 matches of the motif. All these motifs showed low presence in different organisms only being present in half of the strains and rarely in any species **Table 5.3.** All matches were located in an intrinsically disordered region between residues 1210-1557 of the RON6 protein sequence. This region did not display any sign of secondary structure as seen in the ColabFold structure prediction **Figure 5.1.a**. The high motif number and their accessibility, as well as the available experimental evidence of its expression throughout the *Toxoplasma* infection cycle was enough to make it an appealing candidate.

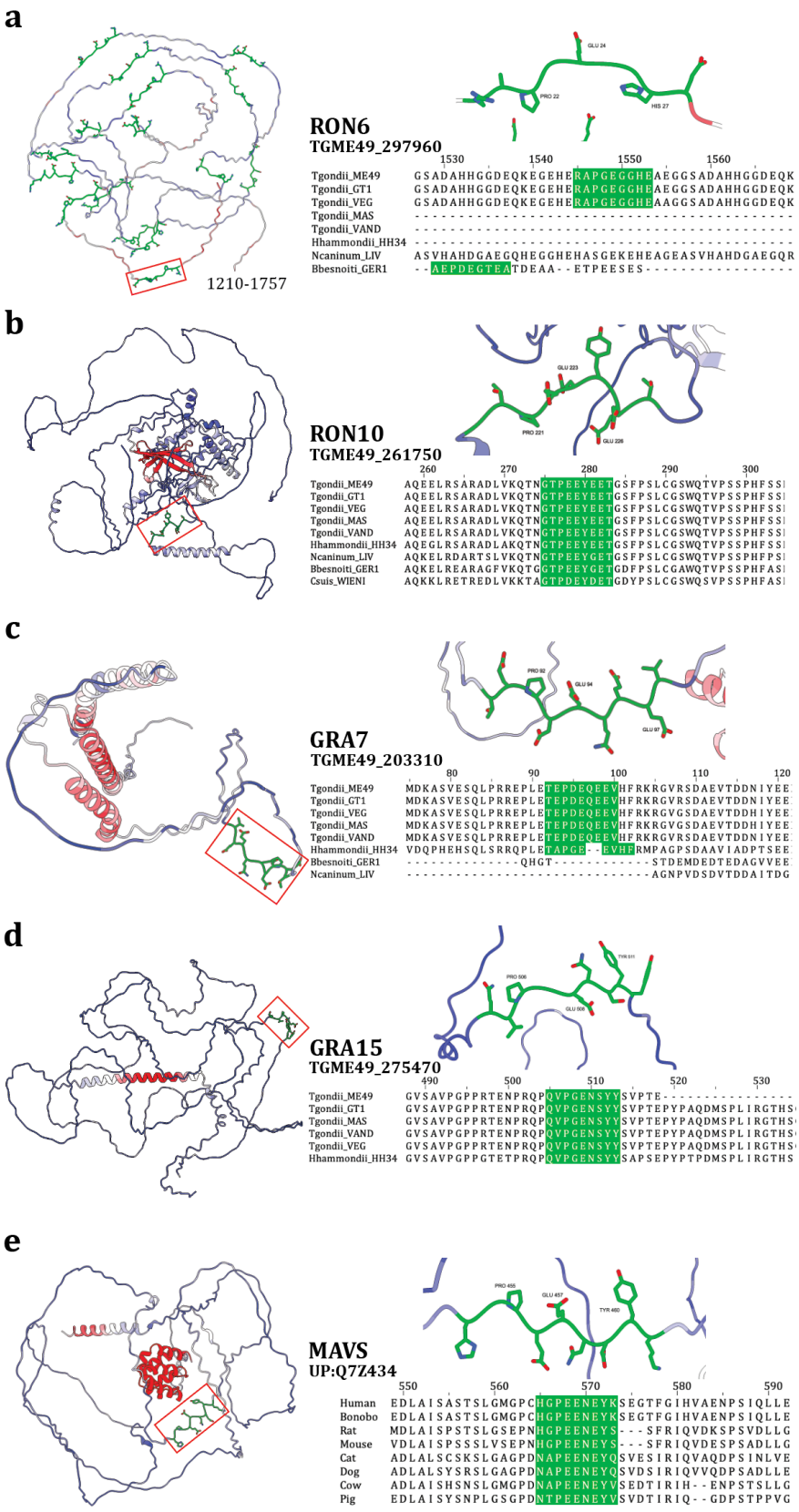


Figure 5.1 TRAF6 motif candidates in secreted proteins. a. ColabFold and **b-e.** AlphaFold structure predictions, with their corresponding alignment motif presence. Red coloring in the structures represent high pLDDT values while Blue represents lower ones. The protein region umbering of amino acids regions is given to partial ColabFold models. Structure visualizations were produced in ChimeraX and motif alignments in Jalview.

Rhoptry neck protein RON10 contained 2 TRAF6 motif matches. RON10 was originally characterized as a binding partner of RON9 in a complex independent from the MJ (Lamarque et al., 2012). RON10 motifs are completely present in all organisms, and RON9 also contains a TRAF6 motif. Dense granules proteins, GRA7 and GRA15 were appealing candidates because they have been previously reported to interact with TRAF6 and play a role in modulating the innate response during *Toxoplasma* infection, but the molecular mechanisms through which they do that are not completely understood (Alaganan et al., 2014; Hunter & Sibley, 2012; Ihara et al., 2020; Yang et al., 2016). However, in neither case has a direct binding affinity measurement to TRAF6 been reported, so testing the motif candidates *in vitro* would provide additional and complementary evidence, as well as structural details of their interaction with TRAF6. Both proteins are mostly disordered and their respective motif matches are present in all *Toxoplasma* strains and in *H. hammondii* sequences **Figure 5.1.c** & **Figure 5.1.d**.

Organism	Protein	Sequence ID	Motif instance	Range	Strains	Species
T. gondii ME49	**RON6**	TGME49_297960 (S8GK12)	RA**P**G**E**GG**H**E	1229-1237	0.5 (2)	0.25 (1)
	RON10	TGME49_261750 (S8F7R7)	GT**P**E**E**YE**E**T	219-227	1.00 (4)	1.00 (4)
	GRA7	TGME49_203310 (O00933)	TE**P**D**E**QE**E**V	88-98	1.00 (4)	0.25 (1)
	GRA15	TGME49_275470 (A0A125YGQ9)	QV**P**G**E**NS**Y**Y	502-512	1.00 (4)	0.25 (1)
H. sapiens (control)	**MAVS**	ENSG00000088888 (Q7Z434)	HG**P**E**E**NE**Y**K	453-461	-	-

Table 5.3 TRAF6 motif candidates in Secreted proteins. In the Sequence ID column both ToxoDB and Uniprot identifiers are included. Under the additional strains and species columns their presence proportion is indicated and in parenthesis their number.

5.2. TRAF6 domain expression

Several TRAF6 domain structures have previously been experimentally determined (Z. Shi et al., 2015). Even so, several labs have reported that it is hard to expressed, presenting problems such as aggregation in bacterial expression systems. After taking a closer look at the TRAF6 domain binding to CD40 structure (PDB: 1LB6), Dr. Gibson identified the presence of two cis-prolines in positions P385 and P425 which might

impact the ability of the domain to fold quickly and without chaperoning by cis-Proline isomerases. In order to increase protein yield, normal protocols tend to use optimized codons. Through these the coding sequence of a gene is modified by changing the codons to the most available tRNAs on a given system, in this case for *E. coli*, ultimately increasing translation speed.

Dr. Gibson's team is fully computational and lacks experimental facilities. Therefore, we sought to express and purify the TRAF6 domain with the help of Dr. Kim Remans and Dr. Arne Boergel from the EMBL Protein Expression and Purification core (PEPcore) facility. Dr Gibson and and Dr Kim Remans then suggested to avoid optimized codons to slower the translation pace and allow the TRAF6 domain to fold properly. I selected the TRAF6 residues 346-504 (Uniprot ID: Q9Y4K3) that contained the motif binding domain and Dr. Kim Remans designed the vector for heterologous expression with a construct that contained the TRAF6-C domain (hTRAF6 from now on) preceded by Histidine and MBP (Maltose binding protein) tag.

His-Tag - MBP-tag - 3c-cleavage site - hTRAF6
MKHHHHHHPMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAH
DRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELK
AKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEA
AFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGL
EAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDA
QTPGSLEVLFQ↓GPAQQCNGIYIWKIGNFGMHLKCQEEEKPVVIHSPGFYTGKPGYKLCMRLHLQLPTAQRCANYI
SLFVHTMQGEYDSHLPWPFQGTIRLTILDQSEAPVRQNHEEIMDAKPELLAFQRPTIPRNPKGFGYVTFMHLEAL
RQRTFIKDDTLLVRCEVSTRFD*

Dr. Boergel carried the heterologous expression of hTRAF6 construct. He transformed the His-tagged MBP hTRAF6 construct into E. coli BL21 DE3 RIL+. He incubated a 60ml culture overnight at 37°C. He then used the culture to inoculate 6L TBFB (Terrific Broth plus Phosphate Buffer) + 30μg/ml kanamycin + 35μg/ml chloramphenicol. He grew the culture until it reached a OD600 cell density of 0.8 and the expression induced with 0.5 mM IPTG. He incubated the culture overnight at 18°C, then harvested the cells by centrifugation at 4,500xg for 30min at 4°C, snap froze them in liquid nitrogen and stored the pellets at -20°C.

Dr. Boergel proceeded by purifying the hTRAF6 through nickel affinity chromatography, anion exchange chromatography, and size-exclusion chromatography. First, he

resuspended the cell pellets in 300ml Ni running buffer (50 mM Tris-HCl pH 8, 500 mM NaCl and 20 mM Imidazol) and added a combination of Sm-nuclease (final concentration 25µg/l), MgCl2 (final concentration 5mM) and cOmplete® protease inhibitor. He then lysed the cells by 5 rounds of microfluidizer and centrifuged the total lysate at 35.000 rpm for 30min at 4°C using a Beckmann Ti45 rotor. He loaded a 5ml-HisTrap column (Machery and Nagel®) with the supernatant, washed it with the same Ni running buffer and eluted using a gradient of 60ml from 0% to 100% Ni elution buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, and 350 mM Imidazol). He analyzed the different separation fractions using SDS-PAGE. He pooled the cleanest fractions with the most protein in them, added 2mg His-3C protease and dialyzed it against a buffer (25mM Hepes pH 7.5, 100 mM NaCl, 20 mM Imidazol) overnight at 4°C to cleave the tag from hTRAF6. Dr. Boergel continued the purification by loading the dialyzed protein on a 5ml Ni column coupled to an anion exchange 5ml Q-HP column (GE®). Through the first one he removed the cleaved tag and protease, and through the second one he removed DNA and other impurities. He then washed the columns with buffer A (25mM Hepes pH 7.5, 100 mM NaCl, 20 mM Imidazol), and eluted the Q-HP column with a step gradient from 0% to 100% with buffer B (25mM Hepes pH 7.5, 1 M NaCl). Finally, he concentrated the flow-through to 5ml and then purified it using a Superdex S75 16/60 with SEC buffer (25 mM Hepes pH7.5, 150 mM NaCl, 1 mM DTT). He pooled the cleanest fractions with the higher amount of hTRAF6 protein and concentrated them to 2.1 mg/ml (111µM). Finally, he flashed froze 38 aliquots of 50μl and stored them at -80°C.

5.3. TRAF6-peptide binding assays

In order to prepare for the binding assays, I retrieved extended versions of the *Toxoplasma* motif candidate sequences **Table 5.3** and as control the human MAVS TRAF6 motif, which was previously reported to bind TRAF6 with known affinity (Z. Shi et al., 2015)**.** Together with Dr. Gibson we designed different mutants to disrupt the potential binding of the peptides. We designed versions of all peptides where the glutamic acid at position +3 was mutated to a serine. These changes kept the PI of the peptides compatible with the next experimental steps. We ordered peptides labelled with 5-Carboxyfluorescine (5-Fluo) at the C-terminus from Biosyntan GmbH (https://www.biosyntan.de) in order to carry out Microscale thermophoresis (MST).

MST detects binding events between molecules by measuring the temperature induced change (after laser exposure) of the fluorescent target intensity (in our case the candidate labelled peptides) as a function of an increasing concentration of the non-fluorescent ligand (in our case hTRAF6). This technique was also appealing to us because it works better with lower amounts of protein than Isothermal Titration Calorimetry (ITC).

Dr. Karine Lapouge from the EMBL PEPCore was responsible for carrying out the binding assays between the purified hTRAF6 domain and the candidate peptides. She dissolved the lyophilised peptides in 25mM HEPES pH 7.5 and 150mM NaCl to a concentration of 2mM and the pH adjusted to 7.5 when necessary. She then performed the MST measurements using a Monolith NT.115 (NanoTemper Technology). She mixed two-fold serial dilutions of the hTRAF6 protein (99 μM) in a 1:1 ratio with 100nM peptides. She performed further titration measurements for all *Toxoplasma* peptides, and the human control MAVS, with two-fold serial dilutions of the hTRAF6 protein (99 μM) mixed in a 9:1 ratio with 500nM. All measurements were performed in triplicates and carried out at 25°C with a LED excitation power of 20% and a medium MST power. Finally, she analyzed the data assuming a 1:1 binding model using the MO.Affinity Software (NanoTemper Technology). Fluorescence anisotropy experiments were also performed by Dr Lapouge, but they were not conclusive enough due to very low anisotropy signal.

5.4. Motif binding results

The MST binding assay produced mixed results among out peptide candidates **Table 5.4**. Firstly, the RON6 motif and its mutant did not shown signs of binding. The high content of Gly inside the motif could affect its binding potential or it can be that the collective motif matches work cooperatively to interact with TRAF6. GRA7 also showed little evidence of binding giving hints that GRA7 might interact with TRAF6 using another protein interface. Only RON10 and GRA15 showed enough signal for binding. Based on the saturation Dr. Lapouge was able to derive binding affinities for GRA15 having a K_D of 27μM and RON10 having a K_D of 18μM **Figure 5.2**. Their respective mutants, as well as the one of the human control, lost binding with the E>S variation.

Protein	ID Nr.	Sequence	MW (g/mol)/PI	Conclusion
RON6	2.0mg	5-Fluo-HERAPG**E**GCHE-Amid	1532.5 / 6.0	No binding
RON6 mut	2.5mg	5-Fluo-HERAPG**S**GGHE-Amid	1490.5 / 6.0	No binding
RON10	3.6mg	5-Fluo-TNGTPE**E**YEET-Amid	1626.6 / 3.58	**Binds (18μM)**
Ron10 mut	2.0mg	5-Fluo-TNGTPE**S**YEET-Amid	1584.5 / 3.67	No binding
GRA7	2.6mg	5-Fluo-LETEPD**E**QEEV-Amid	1674.7 / 3.4	Very week binding
GRA7 mut	2.4mg	5-Fluo-LETEPD**S**QEEV-Amid	1632.6 / 3.45	No binding
GRA15	2.0mg	5-Fluo-QPQVPG**E**NSYY-Amid	1638.7 / 4.0	**Binds (27μM)**
GRA15 mut	2.1mg	5-Fluo-QPQVPG**S**NSYY-Amid	1596.6 / 5.52	No binding
MAVS	2.1mg	5-Fluo-PSHGPE**E**NEYK-Amid	1643.7 / 4.75	**Binds (3μM)**
MAVS mut	3.1mg	5-Fluo-PSHGPE**S**NEYK-Amid	1601.6 / 4.51	Week binding

Table 5.4 Results of TRAF6 motif candidates binding assay. E-S mutations in bold.

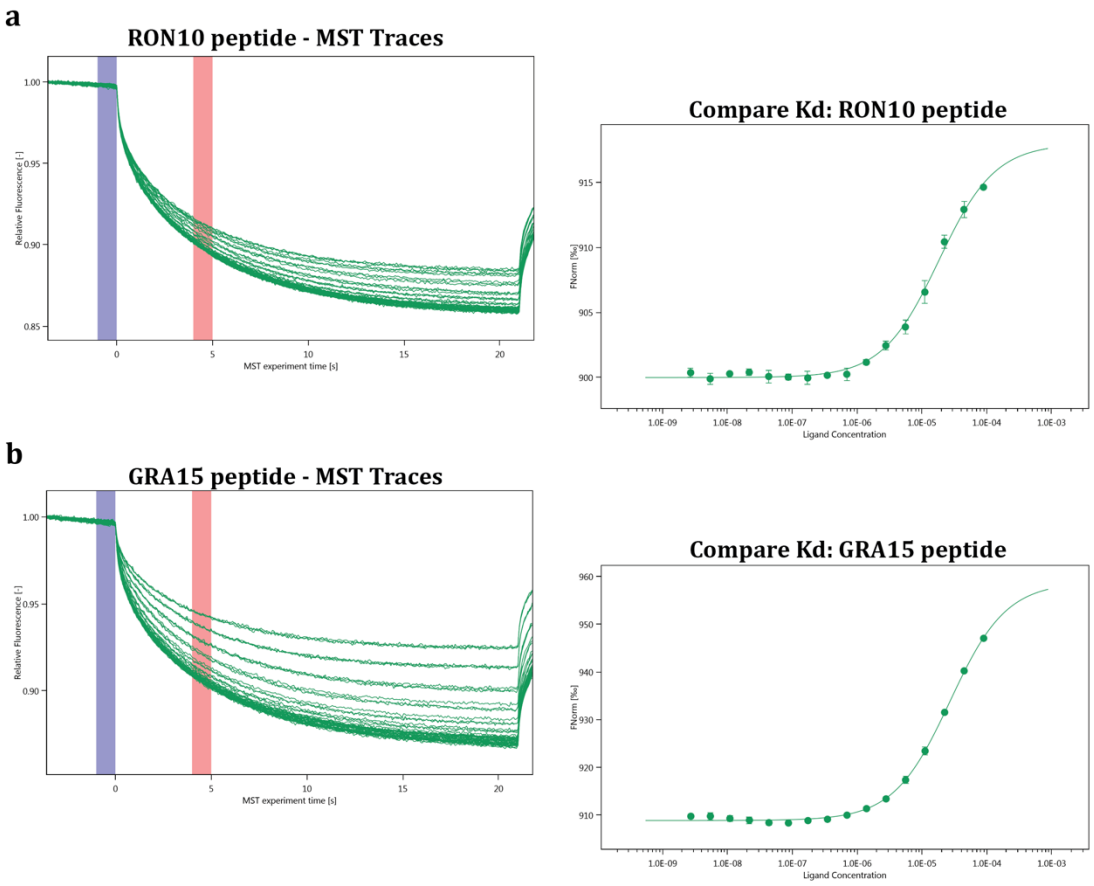


Figure 5.2 hTRAF6 domain binding to RON10 and GRA15 motif peptide. MST signal traces and dose-response curves of **a**. RON10 and **b**. GRA15 peptides binding to hTRAF6 (green lines). Figure was produced by integrating the plots generated by Dr. Karine Lapouge through NanoTemper Technology Software.

5.5. Binding experiment conclusions

In the end we were able to test the binding potential of 4 motif candidates, from which only 2 showed promising results. From an initial set of 205 motif matches in *Toxoplasma* proteins in secretory organelles, we filtered out around 67% to reach a set of 52 motif candidates from which we tested 4 (67 and 20 if we count the multiple matches in RON6), meaning that we provide binding evidence to half of our candidates. From the 17 matches in RON6, individual ones do not show strong binding but might work cooperatively. In the case of GRA15, the protein was already shown to interact with TRAF6 and we now provide more evidence that it is through this motif instance and have measured its binding affinity (as dissociation constant). An *in vivo* set of experiments could now be design in which GRA15 and RON10 are specifically mutated as in our setup, and then test the production of cytokines. One could also mutate the residues in the TRAF6 domain involved in the motif interaction and test if it colocalizes with GRA15 or RON10, or whether the innate immune signaling is affected.

CHAPTER 6
Discussion

6.1. Motif model power and limitations

REGEX models limitations

One of the first things to consider when predicting motifs *in silico* is the power and limitations of the motif models being used. REGEX models do not capture all properties of motif composition, e.g "free" positions represented with "." might still display a clear preference for specific amino acids. The variable presence of charged residues around key interacting amino acids in motifs can also modulate the affinity between them and their binding partners, a property termed motif fuzziness (Duro et al., 2015). This fuzziness is hard to capture through REGEX models, as we would have to extend them to include many optional combinations of charged amino acid positions. Secondly, matches obtained with REGEX models are not ranked and all have the same value as each other. This means that I cannot differentiate which motifs could have a higher or lower affinity to their binding partners. Therefore, when selecting motif candidates, we should take into account that there are still additional motif properties that REGEX models do not capture but will affect their binding potential. In these cases, one should still look carefully at the motif match composition and inspect if the flexible residues are compatible with the motif mode of binding.

ELM classes do not have the same annotation quality

Another source of uncertainty that comes with the use of REGEX motif models arises from the amount of information backing them. ELM models have different amount of sequence and structural evidence behind them **Figure 6.1**. The instance number, the examples based on which the expressions are constructed, varies among them: from classes with no instances to back them, to some having up to ~160 instances. In fact, the median number of instances for all motif classes in ELM is 7. Classes also vary on the number of

instances backed with experimentally derived structures, meaning that their description and REGEX development do not have the same level of structural details.

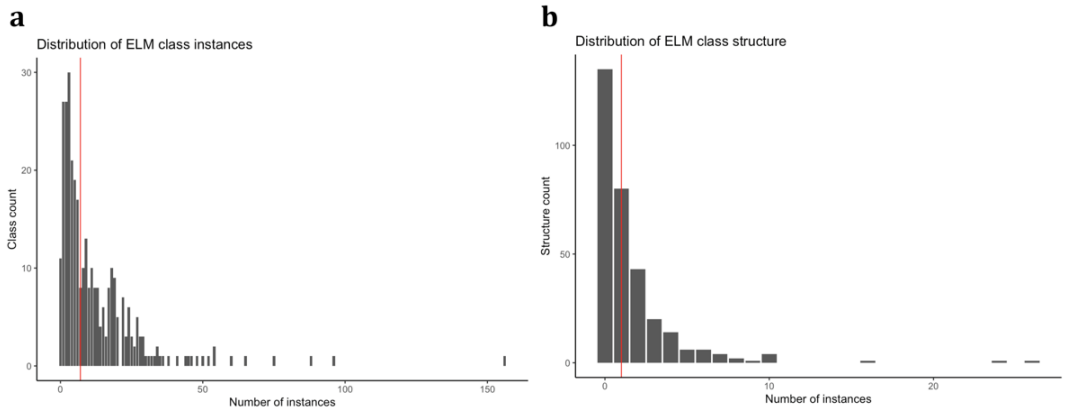


Figure 6.1 Number of ELM motif class instances and structures. Histograms of the number of **a**. instances and **b**. structures backing them. The median is marked with a red line.

The decision to use REGEX instead of PSSM was done because of the former's property for generalizing motif properties and requiring less sequence information, but also due to the unavailability of PSSM models for most known motifs. Additionally, length variations can be flexibly and precisely controlled for the REGEX model. For PSSM models to be reliable they require a higher number of reliable motif instances, which are currently unavailable for Apicomplexan parasites. It might be an option to predict a subset of motifs using only high-quality PSSM models. They could even be implemented in my pipeline to compare the predictions numbers and filtering processes. The variability of motif length, e.g., derived from the spacing between key interacting residues, is not properly captured by current PSSM software which requires researchers to create different matrices for motif instance groups of different lengths. REGEX models do not have this problem, as the expressions can easily incorporate flexible spacing between residues and logical statements. So, we can argue that while acknowledging their shortcomings and data variability, REGEX models are still insightful and flexible for fast and scalable motif searches in whole proteomes, especially when doing the first exploratory surveys.

Class taxa and further motif searches

ELM classes have evidence from specific taxa and whose accuracy might not be precise enough when analyzing proteomes of specific organisms. For this I relied on the

taxonomic annotation of the ELM database and complemented it with a manual search to filter out classes with examples and assessments in specific independent taxa, like Viridiplantae and Fungi, and the ones that would not apply to Apicomplexans or in the infection process (**Chapter 3.1**). There might be that the assessments from ELM and my pipeline are still not precise and certain motifs would still bind to other *Toxoplasma* proteins. Then, the class annotation might need to be revised and reevaluated with new instances and larger alignments. Some motifs can be reevaluated when analyzing different parasite-host pairs, e.g., the ones for Arthropoda might apply to secreted proteins in *Plasmodium* or *Babesia* that use different invertebrate hosts. Overall, the taxonomic annotation filter has a valuable power as it allowed me to focus on motifs potentially involved in infection, the ones that bind to domains absent in *Toxoplasma* but present in its hosts.

The Pipeline is able to use newly defined motifs and their variations

The definition of motif classes and the addition of instances to the ELM is a continuous process, so it is expected that there are more motifs in the literature that were not included in my pipeline. Motif models can also be reevaluated and refined, especially as new evidence and data are generated. In these cases, my pipeline can integrate motifs that have not been curated, e.g., in case there is debate on the definition of a motif model, one can create different versions of the REGEX and add them to the initial model table and compare the results. Nevertheless, the motifs covered by the ELM resource have been through an annotation process providing more reliability to my predictions. In the end I can say that these motif classes are a good and representative start and give a good perspective of the motif content in *Toxoplasma* proteins.

6.2. Variability of structural scores

Motif disorder calculation method

Within the pipeline I determined the disorder context of motif matches by averaging the values derived from IUPred. In certain contexts, this method might not be accurate, e.g., as IUPred uses an amino acid window to calculate individual scores, motifs in small disordered regions close to ordered ones tend to have lower IUPred values. This was the case for the RGD binding motifs (**Chapter 4.2**), and extra considerations and information

had to be taken in order to recover valuable motif candidates. It could be feasible to vary the calculation and the IUPred scoring method, e.g. IUPred SHORT uses a smaller residue window that could give more precise scores for motifs in linker regions and C- and N-termini. Another option would be to add a machine learning disorder predictor, as they have tended to be better at predicting short disordered regions because their datasets come from short missing regions in crystallography structures (Habchi et al., 2014).

When calculating the disorder of protein regions there is no consensus on which predictors to use. Each kind of predictor has advantages and limitations, so it is also recommended to pick predictors from different kinds as they tend to complement each other. A point in favor of using IUPred is that it calculates disorder based on a reference set of ordered regions instead of a disordered one. This means that it is not biased toward previously characterized disordered regions, which tend to be mainly missing regions in crystal structures (Katuwawala et al., 2020b). However, I also use the pLDDT AlphaFold quality scores to complement the assessment of motif disorder and accessibility.

Quality and variability of structure predictions

In contrast to human proteins*, Toxoplasma gondii* AlphaFold predicted structures might actually have lower pLDDT scores due to bad quality alignments and not due to disordered regions. There might not be enough related sequences to form large alignments which lower the quality of the predictions. This low alignment coverage might not necessarily represent disorder but perhaps new tertiary structures or folds of poorly studied proteins. The quality of AlphaFold structures could also affect accessibility. Models could still have small secondary structures that do not fold into tertiary structures or complete domains. This would mean that single alpha helix and beta sheet will be more exposed and their accessibility predictions will be higher. Thus, low-quality predictions are likely to overestimate the disorder level and accessibility of the motif matches. This might not be a problem as I also consider the IUPred scores. If a motif match is located within a low-quality highly accessible region but has low IUPred scores, they will be still be filtered out. Through this approach it is still possible to filter out motif matches in high-quality ordered regions, which would not be able to bind. A way to check the quality of predictions or complement the models used would be to compare the AlphaFold models

with experimentally predicted structures. There are around 200 structures for *Toxoplasma* in the PDB**.**

As many proteins exist in the cell as part of larger protein complexes, single protein AlphaFold predictions and their associated accessibility values would be less reliable if the motif matches lie in regions that would be buried in the complex. Proteins also change conformation when carrying their function, in different conditions or when binding other molecules, in those cases I would also expect that the accessibility of certain protein regions will change.

Using ColabFold for larger proteins was a first approximation and could be further extended. For now, the length of a protein and the availability of computing resources are limiting steps in predicting larger proteins. One could circumvent this issue by predicting structures for specific regions of the protein, e.g. predicting the structure of individual domains or highly ordered regions within large proteins. The domain mapping filter in my pipeline already takes care of motifs inside known folded domains. The large proteins can also be divided into smaller consecutive nested regions and predict them and forming a complete protein assembly. Nevertheless, certain distant regions within a protein could still be able to interact with one another, so this approach would not be able to completely solve this issue.

Further structural mappings

I mainly relied on protein disorder to obtain good scoring motif matches and the only order information added was the domain mappings. My pipeline could easily integrate data and predictions of other protein features like signal peptides or transmembrane regions. Software like SignalP (Teufel et al., 2022) and TMHMM (Hallgren et al., 2022) are commonly used for this and their prediction are sometimes integrated into the databases. Nevertheless, if the tools have not been trained with proteins from parasites so they could still have lower-quality predictions. PROSITE domains are good for proteins with conserved functions but might not be present in unique parasite proteins, e.g., only 3000 *Toxoplasma* proteins have a Domain mapping.

6.3. Conservation scoring

Sequence data quality

From the beginning of this project it was clear that sequence data quality could be an issue when analyzing the conservation of motifs. When dealing with *Sarcocystidae* genomes we face the challenge of having few high-quality proteomes. Some protein sequences in the databases have been automatically translated from genomes sequences and they still need to be validated experimentally, e.g., through mass spectrometry. This might result in the presence of different errors within protein sequences. They might contain many wrongly called residues, gene models with missing exons, fussed or truncated proteins, with missing N- or C-termini. Gene model software might also predict pseudogenes that are never translated during any stage of the parasite life cycle. By selecting specific reference proteomes, I was able to select the ones with better quality, so I did not need to include steps to remove bad-quality sequences. And by integrating Mass Spec data I was able to focus on proteins translated during the infection process. On the other hand, by having few proteomes from related *Sarcocystidae*, I was not able to use different conservation metrics. In this study I was primarily interested in secreted proteins, which might not be present in other apicomplexans (Barylyuk et al., 2020). For example, there might be proteomes available for *Plasmodium* species but they might not have paralogous dense granule proteins that could be aligned with *Toxoplasma*. So, the current alignments would be already the best possible ones for this specific set of proteins. However, if we want to expand our predictions to parasite proteins with housekeeping functions, then adding more protein sequences from distantly related parasites will be necessary.

Another option to expand my alignments is to add more *Toxoplasma gondii* strains as there are currently more genomes available (Galal et al., 2022). For this pipeline, I avoided using many strain sequences as they are minimally variable and potentially redundant. One way to implement them would be to separate different types of alignments for the conservation analysis. Particularly, I would create separate alignments only for the strains, and alignments with strains and species. Counting with larger strain alignments would help in finding specific residue or motif copy variations that could have an effect on their function. For example, the RON6 protein in Chapter 5, displays a higher copy number of the TRAF6 motifs only among Type I, II and III strains **Figure 5.1.a**.

Improvement of sequence homologous groups

In order to form groups of homologous sequences, I used the best BLAST e-value scored sequences from each strain and species when searching against the reference. Because of the small size of their combined proteomes, in comparison with public databases, it was not an issue to find related sequences. But this approach did not tackle common issues like grouping the correct paralogous sequences. A way to address this, also if the alignments become larger, would be to carry a standard reciprocal best hit procedure. Through this I would form the groups by selecting the sequences that have the best e-value scores for each other. Through this I would then be able to make deeper evolutionary questions, like how often a motif originate, move or disappear among paralogous sequences.

Variations of the motif position conservation

In the pipeline I was not able to use standard conservation metrics, due to the lack of sufficient sequences. Instead, I used a reside window to evaluate the presence of motif matches in the different sequences of the alignment. This was very helpful as motifs lie in regions that are poorly conserved regions and the exact positioning of the motifs often changes. Because motifs are short in length, different sequence aligners might not be able to cluster them together. They would have high-scoring penalties if they introduce gaps to optimize motif clustering. In this way, the reside window helped identify small motif position variations (e.g., **Figure 4.5.e**), but its overall reliability was not tested. The one of 15 residues was implemented based on the length distribution of motifs classes and would consider around 90% of the motif classes length (Davey, Van Roey, et al., 2012). But it could be possible to test different residue window changes and evaluate the final organism presence scores as well as the number of filtered hits.

6.4. Further experimental data

Additional supportive data

I added mass spectrometry data in the pipeline in a bimodal way, whether or not a protein had expression data, and if they had their alignments would be produced. Nevertheless, I can still couple the total number of peptides mapped to proteins with the predictions. Protein expression data per stage can also serve to identify proteins that are relevant

during infection, proteins that are overexpressed after host invasion and during the bradyzoite stage. On the other hand, I added phosphorylation data from different experiments. The results from these experiments are dependent on defined set-ups as well as their biological context, e.g., *Toxoplasma gondii* proteins might not be phosphorylated in all conditions or all stages of their life cycle. So, the functionality of a candidate phosphomotif could not be discarded unless taking proper care of analyzing the experimental setup and the nature of the motif interaction.

ToxoDB searches

As mentioned in Chapter 2, ToxoDB has integrative searches that allow users to search for specific genes and proteins using all the available information contained in this resource. This information can be used to filter and expand gene and protein lists to arrive at a certain final list of interesting gene or protein hits, in some cases also chemical compounds. ToxoDB even provides with a tool for searching user-defined motif patterns based on defined REGEX. Knowing this and being familiar with ToxoDB integrative searches one could imagine that this pipeline could be recreated using those exact same searches. While it is true that parts of this pipeline are derived from ToxoDB (location, Mass Spec evidence, PTMs), at the moment the database has not integrated any ELM model or disorder predictions. And while the motif search is useful to find proteins containing a given pattern, it does not treat motif hits as individual entities so one cannot filter out motifs located inside domains or keep the ones inside disorder regions. It is hard to imagine that these additional filters will be integrated in the near future, but having a defined set of motif candidates with a scoring system could be a good addition to the database.

6.5. Prediction pipeline scalability

As mentioned above, I can expand my pipeline to include more data at every step and integrate more software. In summary, I could add multiple disorder predictors, a different set of structure predictions, larger and different types of alignments, as well as more and quantifiable experimental data. All of this will potentially require each of the modules to be rewritten and optimized. One can carry out speed and performance tests using software pipeline managers, also for individual module and parameter testing.

6.6. Filter combination and exploration

Benchmarking

There is not a big reference set for *Toxoplasma* motifs, there are only 14 instances from 5 *Toxoplasma* proteins from 5 distinct ELM motif classes **Table 4.15**. I added some of these instances to the database based on the experiments cited along this work. This does not make for a good training set for benchmarking or machine learning algorithms, especially if the motifs are still not well defined. Besides that, ELM instances have not been integrated with current AlphaFold scores. A reliable motif benchmarking with these scores could be developed and tested further before it could be applied to this work.

Filter exploration

To obtain a better list of motif matches I relied on value filters with different scores. In this sense it was helpful to complement them with one another as they are not mutually exclusive. I also applied all filters to every motif type but as I exemplified in Chapter 4.2 with the RGD motifs, not all of them would apply to each motif case. I could apply specific filtering processes to each motif type, and check if our results are improved. Ultimately, further motif selection will be dependent on the subsequent experiments and the particular context and characteristics of the protein, so there is as much as filtering can do to improve find better motif candidates. We also do not know the extent to which I am excluding true instances, but that can only be assessed once more motifs are validated and my predictions are reevaluated.

6.7. Candidate selection

As we have seen exploring and selecting candidates still require multiple criteria that cannot be assumed for all motif types. I have focused on the ones in secreted proteins but we could start exploring motifs involved in *Toxoplasma* cell processes, like the regulation of the cell cycle or the export and processing of proteins, a topic that has already been reviewed. In fact, I did not explore the motifs of the MOD and CLV types, but there is also the potential to find interesting biology or infection strategies, e.g., compare the MOD sites contained in secreted vs non-secreted proteins. Another way to select groups of candidates is to group them according to Gene Ontology (GO) terms. In a similar fashion

to the motif candidate selection in Chapter 4, we can use the GO terms of each ELM motif class to group relevant proteins with related motifs, and quantify trends for each cellular location or biological process.

6.8. Further candidate testing

Overall, the current approach offers an unbiased way to find and select motifs, without prior expectations we determine their presence in the proteins of *Toxoplasma*. By counting with all motifs in a protein, people can then systematically focus on the process relevant to each protein without missing potentially related ones. "Cherry picking" protein and motif candidates might still bias the type of motifs characterized, so a more thorough approach, also including hypothetical proteins, would be preferred to understand the whole process. In case the number of motifs is too high to make site-directed mutagenesis in the infection context, one could focus on mutating their binding partners in the host cell and see if a phenotype for the parasite is observed. Currently, the dataset is publicly available for researchers to explore but it could still be hard to select candidates without proper knowledge of the motifs. For this reason, it could be helpful to develop a website with exploratory tools that help the researcher in detecting and selecting the list of proteins to test.

6.9. Future perspectives

Motif high throughput research

Motif binding can be assessed in a high throughput fashion by phage display experiments (Ali et al., 2020). Disorder regions of a proteome are fragmented into smaller peptides that are expressed as surface proteins in phages that would then be captured in a domain bait assay. The phages displaying peptides that are enriched are the ones with higher binding affinity and sequences of the inserts can be determined using DNA sequencing. When aligning the peptides that were enriched in the bait experiment one is able to identify motifs instances (Benz et al., 2022). Constructing such a library for *Toxoplasma* and other apicomplexan parasites is possible, in this case it could be from the disorder

regions of the secreted proteins. The result could be used to develop better motif models, and to provide evidence for my candidates or improve selection criteria.

Motif validation through structural modelling

SLiMs and their binding partners can also be modelled by structural prediction. Using AlphaFold, or predictors like RossettaFold (Baek et al., 2021), pairs of motifs containing peptides bound together with their known interacting domains can be determined. One can check if the motif is modelled to bind to the binding pocket of domains. Furthermore, we could use the structures with bound motifs and carry molecular dynamic simulations to know if the interaction is stable (Halpin et al., 2022).

Motif hijacking in Apicomplexa

Comparing the types of motifs that Apicomplexans use to invade host cells can tell us which functions are more frequently hijacked. As secreted proteins evolved faster, the motif hijacking strategy might be conserved across species, telling us that some strategies are preferred over others and providing further molecular details of such interactions. Conservation analyses often tend to heavily focus on the whole protein sequences, which might lose small conserved functional modules such as motifs. Thus, applying my pipeline to other apicomplexans could offer venues to systematically compare motif usage among apicomplexans. With motif-domain interaction information, we can even extend infection signaling networks. By creating a systems biology analysis of the signaling networks affected we could capture how the network can be perturbed by a few effectors.

Drugging opportunities

Finally, motif research opens the opportunity to focus on the host biology to come up with drugging strategies. Instead of trying to drug the parasite to disrupt its cellular machinery, we can then drug the cellular machineries of the host that the parasite requires in order to invade the host, blocking the infection process. Strategies like this have been proposed in bacteria to tackle antibiotic resistance (Sámano-Sánchez & Gibson, 2020). This approach also has the advantage that we do not put selective pressures on the pathogen so it is harder to evolve a novel full protein strategy of infection if they rely on host components.

7. Conclusions

In conclusion, I was able to develop a useful computational pipeline for the discovery of motif instances in *Toxoplasma gondii*. The pipeline was able to capture interesting set of motifs in secreted proteins, as well as recapture and complement previously characterized ones. The amount of motif matches found exemplify that there is a large amount of potential true positive matches in *Toxoplasma*, and other parasite, proteins still to be characterized and potentially discovered. This was possible through the integration of different structural software and the latest experimental information for *Toxoplasma*, as well as the most representative list of SLiMs. The application of this approach, as well as the information generated through have the potential to be expanded and to be used to direct research into protein-protein interaction in the host-parasite interface in a more efficient way. Further motif testing is still necessary to improve approaches like the one presented here, but also if we want to understand the extent to which *Toxoplasma* and other Apicomplexans exploit motif for infection in a systemic way.

8. References

Alaganan, A., Fentress, S. J., Tang, K., Wang, Q., & Sibley, L. D. (2014). *Toxoplasma* GRA7 effector increases turnover of immunity-related GTPases and contributes to acute virulence in the mouse. *Proceedings of the National Academy of Sciences*, *111*(3), 1126–1131. https://doi.org/10.1073/pnas.1313501111

Ali, M., Simonetti, L., & Ivarsson, Y. (2020). Screening Intrinsically Disordered Regions for Short Linear Binding Motifs. In B. B. Kragelund & K. Skriver (Eds.), *Intrinsically Disordered Proteins* (Vol. 2141, pp. 529–552). Springer US. https://doi.org/10.1007/978-1-0716-0524-0_27

Almazán, C., Scimeca, R. C., Reichard, M. V., & Mosqueda, J. (2022). Babesiosis and Theileriosis in North America. *Pathogens*, *11*(2), 168. https://doi.org/10.3390/pathogens11020168

Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E. Y., Bażant, W., Belnap, R., Blevins, A. S., Böhme, U., Brestelli, J., Brunk, B. P., Caddick, M., Callan, D., Campbell, L., Christensen, M. B., Christophides, G. K., Crouch, K., Davis, K., DeBarry, J., … Zheng, J. (2022). VEuPathDB: The eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Research*, *50*(D1), D898–D911. https://doi.org/10.1093/nar/gkab929

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Barylyuk, K., Koreny, L., Ke, H., Butterworth, S., Crook, O. M., Lassadi, I., Gupta, V., Tromer, E., Mourier, T., Stevens, T. J., Breckels, L. M., Pain, A., Lilley, K. S., & Waller, R. F. (2020). A Comprehensive Subcellular Atlas of the Toxoplasma Proteome via hyperLOPIT Provides Spatial Context for Protein Functions. *Cell Host & Microbe*, *28*(5), 752-766.e9. https://doi.org/10.1016/j.chom.2020.09.011

Benz, C., Ali, M., Krystkowiak, I., Simonetti, L., Sayadi, A., Mihalic, F., Kliche, J., Andersson, E., Jemth, P., Davey, N. E., & Ivarsson, Y. (2022). Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Molecular Systems Biology*, *18*(1). https://doi.org/10.15252/msb.202110584

Beraki, T., Hu, X., Broncel, M., Young, J. C., O'Shaughnessy, W. J., Borek, D., Treeck, M., & Reese, M. L. (2019). Divergent kinase regulates membrane ultrastructure of the Toxoplasma parasitophorous vacuole. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(13), 6361–6370. https://doi.org/10.1073/pnas.1816161116

Berdoy, M., Webster, J. P., & Macdonald, D. W. (2000). Fatal attraction in rats infected with *Toxoplasma gondii*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *267*(1452), 1591–1594. https://doi.org/10.1098/rspb.2000.1182

Bieniasz, P. D. (2006). Late budding domains and host proteins in enveloped virus release. *Virology*, *344*(1), 55–63. https://doi.org/10.1016/j.virol.2005.09.044

Boothroyd, J. C., & Dubremetz, J.-F. (2008). Kiss and spit: The dual roles of Toxoplasma rhoptries. *Nature Reviews Microbiology*, *6*(1), 79–88. https://doi.org/10.1038/nrmicro1800

Bratek-Skicki, A., Pancsa, R., Meszaros, B., Van Lindt, J., & Tompa, P. (2020). A guide to regulation of the formation of biomolecular condensates. *The FEBS Journal*, *287*(10), 1924–1935. https://doi.org/10.1111/febs.15254

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, *20*(3), 426–427. https://doi.org/10.1093/bioinformatics/btg430

Coffey, M. J., Sleebs, B. E., Uboldi, A. D., Garnham, A., Franco, M., Marino, N. D., Panas, M. W., Ferguson, D. J., Enciso, M., O'Neill, M. T., Lopaticki, S., Stewart, R. J., Dewson, G., Smyth, G. K., Smith, B. J., Masters, S. L., Boothroyd, J. C., Boddey, J. A., & Tonkin, C. J. (2015). An aspartyl protease defines a novel pathway for export of Toxoplasma proteins into the host cell. *ELife*, *4*, e10809. https://doi.org/10.7554/eLife.10809

Dahlgren, S. S., Gouveia-Oliveira, R., & Gjerde, B. (2008). Phylogenetic relationships between Sarcocystis species from reindeer and other Sarcocystidae deduced from ssu rRNA gene sequences. *Veterinary Parasitology*, *151*(1), 27–35. https://doi.org/10.1016/j.vetpar.2007.09.029

Darnay, B. G., Ni, J., Moore, P. A., & Aggarwal, B. B. (1999). Activation of NF-κB by RANK Requires Tumor Necrosis Factor Receptor-associated Factor (TRAF) 6 and NF-κB-inducing Kinase. *Journal of Biological Chemistry*, *274*(12), 7724–7731. https://doi.org/10.1074/jbc.274.12.7724

Davey, N. E., Cowan, J. L., Shields, D. C., Gibson, T. J., Coldwell, M. J., & Edwards, R. J. (2012). SLiMPrints: Conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Research*, *40*(21), 10628–10641. https://doi.org/10.1093/nar/gks854

Davey, N. E., Travé, G., & Gibson, T. J. (2011). How viruses hijack cell regulation. *Trends in Biochemical Sciences*, *36*(3), 159–169. https://doi.org/10.1016/j.tibs.2010.10.002

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Mol. BioSyst.*, *8*(1), 268–281. https://doi.org/10.1039/C1MB05231D

Dhal, A. K., Panda, C., Yun, S.-I., & Mahapatra, R. K. (2022). An update on Cryptosporidium biology and therapeutic avenues. *Journal of Parasitic Diseases*, *46*(3), 923–939. https://doi.org/10.1007/s12639-022-01510-5

Dos Santos Pacheco, N., Brusini, L., Haase, R., Tosetti, N., Maco, B., Brochet, M., Vadas, O., & Soldati-Favre, D. (2022). Conoid extrusion regulates glideosome assembly to control motility and invasion in Apicomplexa. *Nature Microbiology*, *7*(11), 1777–1790. https://doi.org/10.1038/s41564-022-01212-x

Dosztányi, Z., Csizmók, V., Tompa, P., & Simon, I. (2005). The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *Journal of Molecular Biology*, *347*(4), 827–839. https://doi.org/10.1016/j.jmb.2005.01.071

Dubey, J. P. (2021). *Toxoplasmosis of Animals and Humans* (3rd ed.). CRC Press. https://doi.org/10.1201/9781003199373

Duro, N., Miskei, M., & Fuxreiter, M. (2015). Fuzziness endows viral motif-mimicry. *Molecular BioSystems*, *11*(10), 2821–2829. https://doi.org/10.1039/C5MB00301F

English, E. D., & Striepen, B. (2019). The cat is out of the bag: How parasites know their hosts. *PLOS Biology*, *17*(9), e3000446. https://doi.org/10.1371/journal.pbio.3000446

Erdős, G., Pajkos, M., & Dosztányi, Z. (2021). IUPred3: Prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of

evolutionary conservation. *Nucleic Acids Research*, *49*(W1), W297–W303. https://doi.org/10.1093/nar/gkab408

Galal, L., Ariey, F., Gouilh, M. A., Dardé, M.-L., Hamidović, A., Letourneur, F., Prugnolle, F., & Mercier, A. (2022). A unique Toxoplasma gondii haplotype accompanied the global expansion of cats. *Nature Communications*, *13*(1), 5778. https://doi.org/10.1038/s41467-022-33556-7

Gibson, T. J., Dinkel, H., Van Roey, K., & Diella, F. (2015). Experimental detection of short regulatory motifs in eukaryotic proteins: Tips for good practice as well as for bad. *Cell Communication and Signaling: CCS*, *13*, 42. https://doi.org/10.1186/s12964-015-0121-y

Gouw, M., Alvarado-Valverde, J., Čalyševa, J., Diella, F., Kumar, M., Michael, S., Van Roey, K., Dinkel, H., & Gibson, T. J. (2020). How to Annotate and Submit a Short Linear Motif to the Eukaryotic Linear Motif Resource. In B. B. Kragelund & K. Skriver (Eds.), *Intrinsically Disordered Proteins* (Vol. 2141, pp. 73–102). Springer US. https://doi.org/10.1007/978-1-0716-0524-0_4

Guérin, A., Corrales, R. M., Parker, M. L., Lamarque, M. H., Jacot, D., El Hajj, H., Soldati-Favre, D., Boulanger, M. J., & Lebrun, M. (2017). Efficient invasion by Toxoplasma depends on the subversion of host protein networks. *Nature Microbiology*, *2*(10), 1358–1366. https://doi.org/10.1038/s41564-017-0018-1

Habchi, J., Tompa, P., Longhi, S., & Uversky, V. N. (2014). Introducing Protein Intrinsic Disorder. *Chemical Reviews*, *114*(13), 6561–6588. https://doi.org/10.1021/cr400514h

Hakimi, M.-A., Olias, P., & Sibley, L. D. (2017). *Toxoplasma* Effectors Targeting Host Signaling and Transcription. *Clinical Microbiology Reviews*, *30*(3), 615–645. https://doi.org/10.1128/CMR.00005-17

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* [Preprint]. Bioinformatics. https://doi.org/10.1101/2022.04.08.487609

Halpin, J. C., Whitney, D., Rigoldi, F., Sivaraman, V., Singer, A., & Keating, A. E. (2022). Molecular determinants of TRAF6 binding specificity suggest that native interaction

partners are not optimized for affinity. *Protein Science*, *31*(11). https://doi.org/10.1002/pro.4429

Heinemann, S., Biesinger, B., Fleckenstein, B., & Albrecht, J.-C. (2006). NFκB Signaling Is Induced by the Oncoprotein Tio through Direct Interaction with TRAF6. *Journal of Biological Chemistry*, *281*(13), 8565–8572. https://doi.org/10.1074/jbc.M510891200

Henne, W. M., Buchkovich, N. J., & Emr, S. D. (2011). The ESCRT Pathway. *Developmental Cell*, *21*(1), 77–91. https://doi.org/10.1016/j.devcel.2011.05.015

Honnappa, S., Gouveia, S. M., Weisbrich, A., Damberger, F. F., Bhavesh, N. S., Jawhari, H., Grigoriev, I., van Rijssel, F. J. A., Buey, R. M., Lawera, A., Jelesarov, I., Winkler, F. K., Wüthrich, K., Akhmanova, A., & Steinmetz, M. O. (2009). An EB1-Binding Motif Acts as a Microtubule Tip Localization Signal. *Cell*, *138*(2), 366–376. https://doi.org/10.1016/j.cell.2009.04.065

Hu, K., Johnson, J., Florens, L., Fraunholz, M., Suravajjala, S., DiLullo, C., Yates, J., Roos, D. S., & Murray, J. M. (2006). Cytoskeletal Components of an Invasion Machine—The Apical Complex of Toxoplasma gondii. *PLoS Pathogens*, *2*(2), e13. https://doi.org/10.1371/journal.ppat.0020013

Huang, W., Liao, J., Hsiao, T., Wei, T. W., Maestre-Reyna, M., Bessho, Y., & Tsai, M. (2018). Binding and Enhanced Binding between Key Immunity Proteins TRAF6 and TIFA. *ChemBioChem*, cbic.201800436. https://doi.org/10.1002/cbic.201800436

Huber, S., Karagenc, T., Ritler, D., Rottenberg, S., & Woods, K. (2018). Identification and characterisation of a *Theileria annulata* proline-rich microtubule and SH3 domain-interacting protein (TaMISHIP) that forms a complex with CLASP1, EB1, and CD2AP at the schizont surface. *Cellular Microbiology*, *20*(7), e12838. https://doi.org/10.1111/cmi.12838

Hunter, C. A., & Sibley, L. D. (2012). Modulation of innate immunity by Toxoplasma gondii virulence effectors. *Nature Reviews Microbiology*, *10*(11), 766–778. https://doi.org/10.1038/nrmicro2858

Ihara, F., Fereig, R. M., Himori, Y., Kameyama, K., Umeda, K., Tanaka, S., Ikeda, R., Yamamoto, M., & Nishikawa, Y. (2020). Toxoplasma gondii Dense Granule Proteins 7, 14, and 15 Are Involved in Modification and Control of the Immune Response Mediated via NF-κB Pathway. *Frontiers in Immunology*, *11*, 1709. https://doi.org/10.3389/fimmu.2020.01709

Janouškovec, J., Horák, A., Oborník, M., Lukeš, J., & Keeling, P. J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences*, *107*(24), 10949–10954. https://doi.org/10.1073/pnas.1003335107

Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., & Vriend, G. (2011). A series of PDB related databases for everyday needs. *Nucleic Acids Research*, *39*(Database), D411–D419. https://doi.org/10.1093/nar/gkq1105

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Katuwawala, A., Oldfield, C. J., & Kurgan, L. (2020a). Accuracy of protein-level disorder predictions. *Briefings in Bioinformatics*, *21*(5), 1509–1522. https://doi.org/10.1093/bib/bbz100

Katuwawala, A., Oldfield, C. J., & Kurgan, L. (2020b). Accuracy of protein-level disorder predictions. *Briefings in Bioinformatics*, *21*(5), 1509–1522. https://doi.org/10.1093/bib/bbz100

Kessler, H., Herm-Götz, A., Hegge, S., Rauch, M., Soldati-Favre, D., Frischknecht, F., & Meissner, M. (2008). Microneme protein 8 – a new essential invasion factor in *Toxoplasma gondii*. *Journal of Cell Science*, *121*(7), 947–956. https://doi.org/10.1242/jcs.022350

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, *5*(11), 826–837. https://doi.org/10.1038/nrg1471

Koreny, L., Mercado-Saavedra, B. N., Klinger, C. M., Barylyuk, K., Butterworth, S., Hirst, J., Rivera-Cuevas, Y., Zaccai, N. R., Holzer, V. J. C., Klingl, A., Dacks, J. B., Carruthers, V. B., Robinson, M. S., Gras, S., & Waller, R. F. (2022). *Stable and ancient endocytic*

structures navigate the complex pellicle of apicomplexan parasites [Preprint]. Cell Biology. https://doi.org/10.1101/2022.06.02.494549

Krystkowiak, I., Manguy, J., & Davey, N. E. (2018). PSSMSearch: A server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Research*, *46*(W1), W235–W241. https://doi.org/10.1093/nar/gky426

Kumar, M., Michael, S., Alvarado-Valverde, J., Mészáros, B., Sámano-Sánchez, H., Zeke, A., Dobson, L., Lazar, T., Örd, M., Nagpal, A., Farahi, N., Käser, M., Kraleti, R., Davey, N. E., Pancsa, R., Chemes, L. B., & Gibson, T. J. (2022). The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Research*, *50*(D1), D497–D508. https://doi.org/10.1093/nar/gkab975

Laliberté, J., & Carruthers, V. B. (2011). Toxoplasma gondii toxolysin 4 is an extensively processed putative metalloproteinase secreted from micronemes. *Molecular and Biochemical Parasitology*, *177*(1), 49–56. https://doi.org/10.1016/j.molbiopara.2011.01.009

Lamarque, M. H., Papoin, J., Finizio, A.-L., Lentini, G., Pfaff, A. W., Candolfi, E., Dubremetz, J.-F., & Lebrun, M. (2012). Identification of a New Rhoptry Neck Complex RON9/RON10 in the Apicomplexa Parasite Toxoplasma gondii. *PLoS ONE*, *7*(3), e32457. https://doi.org/10.1371/journal.pone.0032457

Lesk, A. (2010). *Introduction to Protein Science: Architecture, Function, and Genomics*. OUP Oxford. https://books.google.de/books?id=QVScAQAAQBAJ

Liu, Q., Li, F.-C., Zhou, C.-X., & Zhu, X.-Q. (2017). Research advances in interactions related to Toxoplasma gondii microneme proteins. *Experimental Parasitology*, *176*, 89–98. https://doi.org/10.1016/j.exppara.2017.03.001

Ma, J. S., Sasai, M., Ohshima, J., Lee, Y., Bando, H., Takeda, K., & Yamamoto, M. (2014). Selective and strain-specific NFAT4 activation by the Toxoplasma gondii polymorphic dense granule protein GRA6. *Journal of Experimental Medicine*, *211*(10), 2013–2032. https://doi.org/10.1084/jem.20131272

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Martorelli Di Genova, B., Wilson, S. K., Dubey, J. P., & Knoll, L. J. (2019). Intestinal delta-6-desaturase activity determines host range for Toxoplasma sexual reproduction. *PLOS Biology*, *17*(8), e3000364. https://doi.org/10.1371/journal.pbio.3000364

Mathur, V., Kwong, W. K., Husnik, F., Irwin, N. A. T., Kristmundsson, Á., Gestal, C., Freeman, M., & Keeling, P. J. (2021). Phylogenomics Identifies a New Major Subgroup of Apicomplexans, Marosporida *class nov.* , with Extreme Apicoplast Genome Reduction. *Genome Biology and Evolution*, *13*(2), evaa244. https://doi.org/10.1093/gbe/evaa244

Mészáros, B., Erdős, G., & Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, *46*(W1), W329–W337. https://doi.org/10.1093/nar/gky384

Mészáros, B., Kumar, M., Gibson, T. J., Uyar, B., & Dosztányi, Z. (2017). Degrons in cancer. *Science Signaling*, *10*(470), eaak9982. https://doi.org/10.1126/scisignal.aak9982

Mészáros, B., Sámano-Sánchez, H., Alvarado-Valverde, J., Čalyševa, J., Martínez-Pérez, E., Alves, R., Shields, D. C., Kumar, M., Rippmann, F., Chemes, L. B., & Gibson, T. J. (2021). Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. *Science Signaling*, *14*(665), eabd0334. https://doi.org/10.1126/scisignal.abd0334

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Monroe, A., Williams, N. A., Ogoma, S., Karema, C., & Okumu, F. (2022). Reflections on the 2021 World Malaria Report and the future of malaria control. *Malaria Journal*, *21*(1), 154. https://doi.org/10.1186/s12936-022-04178-7

Montoya, J., & Liesenfeld, O. (2004). Toxoplasmosis. *The Lancet*, *363*(9425), 1965–1976. https://doi.org/10.1016/S0140-6736(04)16412-X

Nadipuram, S. M., Thind, A. C., Rayatpisheh, S., Wohlschlegel, J. A., & Bradley, P. J. (2020). Proximity biotinylation reveals novel secreted dense granule proteins of Toxoplasma gondii bradyzoites. *PLOS ONE*, *15*(5), e0232552. https://doi.org/10.1371/journal.pone.0232552

Nebl, T., Prieto, J. H., Kapp, E., Smith, B. J., Williams, M. J., Yates, J. R., Cowman, A. F., & Tonkin, C. J. (2011). Quantitative in vivo analyses reveal calcium-dependent

phosphorylation sites and identifies a novel component of the Toxoplasma invasion motor complex. *PLoS Pathogens*, *7*(9), e1002222. https://doi.org/10.1371/journal.ppat.1002222

Olias, P., Schade, B., & Mehlhorn, H. (2011). Molecular pathology, taxonomy and epidemiology of Besnoitia species (Protozoa: Sarcocystidae). *Infection, Genetics and Evolution*, *11*(7), 1564–1576. https://doi.org/10.1016/j.meegid.2011.08.006

O'Shea, J. P., Chou, M. F., Quader, S. A., Ryan, J. K., Church, G. M., & Schwartz, D. (2013). pLogo: A probabilistic approach to visualizing sequence motifs. *Nature Methods*, *10*(12), 1211–1212. https://doi.org/10.1038/nmeth.2646

Pancsa, R., & Tompa, P. (2012). Structural Disorder in Eukaryotes. *PLoS ONE*, *7*(4), e34687. https://doi.org/10.1371/journal.pone.0034687

Procter, J. B., Carstairs, G. M., Soares, B., Mourão, K., Ofoegbu, T. C., Barton, D., Lui, L., Menard, A., Sherstnev, N., Roldan-Martinez, D., Duce, S., Martin, D. M. A., & Barton, G. J. (2021). Alignment of Biological Sequences with Jalview. In K. Katoh (Ed.), *Multiple Sequence Alignment* (Vol. 2231, pp. 203–224). Springer US. https://doi.org/10.1007/978-1-0716-1036-7_13

Puntervoll, P. (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, *31*(13), 3625–3630. https://doi.org/10.1093/nar/gkg545

Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., Pajkos, M., Lazar, T., Peña-Díaz, S., Santos, J., Ács, V., Farahi, N., Fichó, E., Aspromonte, M. C., Bassot, C., Chasapi, A., Davey, N. E., Davidović, R., Dobson, L., … Piovesan, D. (2022). DisProt in 2022: Improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research*, *50*(D1), D480–D487. https://doi.org/10.1093/nar/gkab1082

Rastogi, S., Cygan, A. M., & Boothroyd, J. C. (2019). Translocation of effector proteins into host cells by Toxoplasma gondii. *Current Opinion in Microbiology*, *52*, 130–138. https://doi.org/10.1016/j.mib.2019.07.002

Rivera-Cuevas, Y., Mayoral, J., Di Cristina, M., Lawrence, A.-L. E., Olafsson, E. B., Patel, R. K., Thornhill, D., Waldman, B. S., Ono, A., Sexton, J. Z., Lourido, S., Weiss, L. M., & Carruthers, V. B. (2021). Toxoplasma gondii exploits the host ESCRT machinery for parasite uptake of host cytosolic proteins. *PLOS Pathogens*, *17*(12), e1010138. https://doi.org/10.1371/journal.ppat.1010138

Rome, M. E., Beck, J. R., Turetzky, J. M., Webster, P., & Bradley, P. J. (2008). Intervacuolar Transport and Unique Topology of GRA14, a Novel Dense Granule Protein in *Toxoplasma gondii*. *Infection and Immunity*, *76*(11), 4865–4875. https://doi.org/10.1128/IAI.00782-08

Rost, B., & Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins: Structure, Function, and Genetics*, *20*(3), 216–226. https://doi.org/10.1002/prot.340200303

Saeij, J. P. J., Coller, S., Boyle, J. P., Jerome, M. E., White, M. W., & Boothroyd, J. C. (2007). Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature*, *445*(7125), 324–327. https://doi.org/10.1038/nature05395

Sámano-Sánchez, H., & Gibson, T. J. (2020). Mimicry of Short Linear Motifs by Bacterial Pathogens: A Drugging Opportunity. *Trends in Biochemical Sciences*, *45*(6), 526–544. https://doi.org/10.1016/j.tibs.2020.03.003

Shi, J.-H., & Sun, S.-C. (2018). Tumor Necrosis Factor Receptor-Associated Factor Regulation of Nuclear Factor κB and Mitogen-Activated Protein Kinase Pathways. *Frontiers in Immunology*, *9*, 1849. https://doi.org/10.3389/fimmu.2018.01849

Shi, Z., Zhang, Z., Zhang, Z., Wang, Y., Li, C., Wang, X., He, F., Sun, L., Jiao, S., Shi, W., & Zhou, Z. (2015). Structural Insights into Mitochondrial Antiviral Signaling Protein (MAVS)-Tumor Necrosis Factor Receptor-associated Factor 6 (TRAF6) Signaling. *Journal of Biological Chemistry*, *290*(44), 26811–26820. https://doi.org/10.1074/jbc.M115.666578

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2012). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(D1), D344–D347. https://doi.org/10.1093/nar/gks1067

Sparvoli, D., Delabre, J., Penarete-Vargas, D. M., Kumar Mageswaran, S., Tsypin, L. M., Heckendorn, J., Theveny, L., Maynadier, M., Mendonça Cova, M., Berry-Sterkers, L., Guérin, A., Dubremetz, J., Urbach, S., Striepen, B., Turkewitz, A. P., Chang, Y., & Lebrun, M. (2022). An apical membrane complex for triggering rhoptry exocytosis

and invasion in *Toxoplasma*. *The EMBO Journal*, *41*(22). https://doi.org/10.15252/embj.2022111158

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, *40*(7), 1023–1025. https://doi.org/10.1038/s41587-021-01156-3

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., … Zhang, J. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE*, *8*(11), e80635. https://doi.org/10.1371/journal.pone.0080635

Tompa, P., Davey, N. E., Gibson, T. J., & Babu, M. M. (2014). A Million Peptide Motifs for the Molecular Biologist. *Molecular Cell*, *55*(2), 161–169. https://doi.org/10.1016/j.molcel.2014.05.032

Treeck, M., Sanders, J. L., Elias, J. E., & Boothroyd, J. C. (2011). The phosphoproteomes of Plasmodium falciparum and Toxoplasma gondii reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host & Microbe*, *10*(4), 410–419. https://doi.org/10.1016/j.chom.2011.09.004

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Turetzky, J. M., Chu, D. K., Hajagos, B. E., & Bradley, P. J. (2010). Processing and secretion of ROP13: A unique Toxoplasma effector protein. *International Journal for Parasitology*, *40*(9), 1037–1044. https://doi.org/10.1016/j.ijpara.2010.02.014

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., … Velankar, S. (2022). AlphaFold

Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Votýpka, J., Modrý, D., Oborník, M., Šlapeta, J., & Lukeš, J. (2017). Apicomplexa. In J. M. Archibald, A. G. B. Simpson, & C. H. Slamovits (Eds.), *Handbook of the Protists* (pp. 567–624). Springer International Publishing. https://doi.org/10.1007/978-3-319-28149-0_20

Wang, J.-L., Li, T.-T., Elsheikha, H. M., Chen, K., Zhu, W.-N., Yue, D.-M., Zhu, X.-Q., & Huang, S.-Y. (2017). Functional Characterization of Rhoptry Kinome in the Virulent Toxoplasma gondii RH Strain. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00084

White, M. W., & Suvorova, E. S. (2018). Apicomplexa Cell Cycles: Something Old, Borrowed, Lost, and New. *Trends in Parasitology*, *34*(9), 759–771. https://doi.org/10.1016/j.pt.2018.07.006

Woo, Y. H., Ansari, H., Otto, T. D., Klinger, C. M., Kolisko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., Ali, S., Bernal, A., del Campo, J., Cihlář, J., Flegontov, P., Gornik, S. G., Hajdušková, E., Horák, A., Janouškovec, J., … Pain, A. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *ELife*, *4*, e06974. https://doi.org/10.7554/eLife.06974

Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, *16*(1), 18–29. https://doi.org/10.1038/nrm3920

Yamamoto, M., Gohda, J., Akiyama, T., & Inoue, J. (2021). TNF receptor-associated factor 6 (TRAF6) plays crucial roles in multiple biological systems through polyubiquitination-mediated NF-κB activation. *Proceedings of the Japan Academy, Series B*, *97*(4), 145–160. https://doi.org/10.2183/pjab.97.009

Yang, C.-S., Yuk, J.-M., Lee, Y.-H., & Jo, E.-K. (2016). Toxoplasma gondii GRA7-Induced TRAF6 Activation Contributes to Host Protective Immunity. *Infection and Immunity*, *84*(1), 339–350. https://doi.org/10.1128/IAI.00734-15

Ye, H., Arron, J. R., Lamothe, B., Cirilli, M., Kobayashi, T., Shevde, N. K., Segal, D., Dzivenu, O. K., Vologodskaia, M., Yim, M., Du, K., Singh, S., Pike, J. W., Darnay, B. G., Choi, Y., &

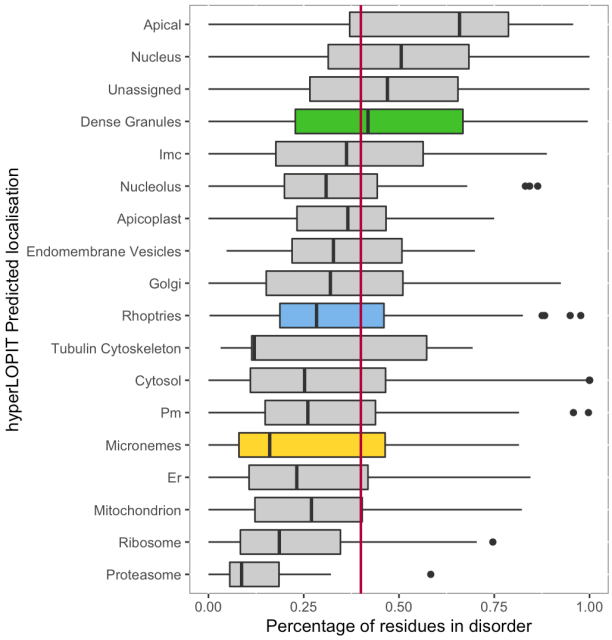
Wu, H. (2002). Distinct molecular mechanism for initiating TRAF6 signalling. *Nature*, *418*(6896), 443–447. https://doi.org/10.1038/nature00888

Zaheer, T., Abbas, R. Z., Imran, M., Abbas, A., Butt, A., Aslam, S., & Ahmad, J. (2022). Vaccines against chicken coccidiosis with particular reference to previous decade: Progress, challenges, and opportunities. *Parasitology Research*, *121*(10), 2749–2763. https://doi.org/10.1007/s00436-022-07612-6

Zhu, W., Li, J., Pappoe, F., Shen, J., & Yu, L. (2019). Strategies Developed by Toxoplasma gondii to Survive in the Host. *Frontiers in Microbiology*, *10*, 899. https://doi.org/10.3389/fmicb.2019.00899

Annex

Supplementary figures



Supplementary Figure 4.1 Disorder levels of the different Toxoplasma gondii cellular locations. Boxplots of the values corresponding to the percentage of residues in disorder context (above the 0.4 IUPred score) for the proteins assigned to different cellular locations. Secretory organelles are colored, and the red line correspond to a proportion of 0.4 residues in disorder.

Code annex

Annex2.1

PROSITE domain information table format:

```
From          Entry Name          Domain [FT]
A0A125YP69    A0A125YP69_TOXGM    DOMAIN      163..464;      /note="Peptidase_M24";
/evidence="ECO:0000259|Pfam:PF00557"
```

Annex2.2

Command

```
python        pLDDT_Values.py     PDB_files_directory
```

PDB values format

```
TITLE     2 (A0A0N5EAW5)
ATOM      2  CA  MET A   1      -23.615 -62.293  -6.667  1.00 32.94           C
ATOM     10  CA  ALA A   2      -20.981 -63.857  -8.079  1.00 34.59           C
…
ATOM   6250  CA  PHE A 821      -77.969 -34.034   9.454  1.00 34.69           C
ATOM   6261  CA  GLY A 822      -78.078 -35.236  13.129  1.00 40.94           C
```

pLDDT values format

```
A0A0N5EAW5 [32.94, 34.59, …, 34.69, 40.94]
```

Annex2.3

Command

```
python       accessibility_Values.py   DSSP_files_directory
```

DSSP values format

```
TITLE     2 (A0A0N5EAW5)
  #  RESIDUE AA STRUCTURE BP1 BP2    ACC     N-H-->O    O-->H-N    N-H-->O …
  1       1 A  M              0   0  240      0, 0.0     0, 0.0     0, 0.0 …
  2       2 A  A        +     0   0   80      0, 0.0     0, 0.0     0, 0.0 …
…
821     821 A  F              0   0  218     -2,-0.2     0, 0.0     1,-0.1 …
822     822 A  G              0   0  151     -2,-0.2    -1,-0.1     0, 0.0 …
```

DSSP values format before normalization

```
A0A0N5EAW5 [240, 80, …, 218, 151]
```

DSSP values format after normalization

```
A0A0N5EAW5 [1, 0.620, …, 0.908, 1]
```

Annex2.4

ToxoDB PTMs table format:

```
Gene ID      Modified Residues Total Modified Residues Total Modifications By Type
TGME49_200280                  S:680, T:881           2    phosphorylation site:2
```

PTMs table format for Motif match integration:

```
key                 seq_id              aminoacid    residue
TGME49_200280_1     TGME49_200280       S            680
TGME49_200280_2     TGME49_200280       T            881
```

Annex2.5

ELM motif classes table format:

```
CLV_C14_Caspase3-7        ([DSTE][^P][^DEWHFYC]D[GSAN])    0.00309374
```

ToxoDB proteome in fasta format:

```
>TGME49_287280-t26_1-p1
MMHLIQKKCPGFPPGFQLPCRLKARRGRLFRHESCTMLFSVALCLTALASFVPFECSTRR…
```

Command for motif match search:

```
Python   MotifMatches_Dis.py  'ALIAS'   TgondiiME49.fasta   Elm_classes.tsv     0.4
```

Motif matches result table format:

```
Protein_ID                      Motif_Name          Match_N
         Motif_Instance
TGME49_287280-t26_1-p1     CLV_C14_Caspase3-7  1            TERDG

         Motif_sSite       Motif_Disorder            Dis_context
         85                0.634                disorder
```

Annex2.6

Command for reformatting motif match sites:

```
Rscript      MotifMatches_Sites.R      ALIAS_MotifMatches_list.txt
```

Motif match site table format:

```
Protein_ID                Motif_Name          Match_N            Motif_sites
TGME49_287280-t26_1-p1    CLV_C14_Caspase3-7  1            85
```

Command to find motif matches in alignments:

```
Python     MotifMatches_InAlignments.py Elm_classes.tsv
           ALIAS_MotifMatches_sites.txt
```

Annex2.7

Motif match presence table format:

```
key                              seq_id              motif               motif_site
TGME49_292920|CLV_C14_Caspase3-7|1     TGME49_292920       CLV_C14_Caspase3-7  1

       presence_org  presence_str presence_spc
       0.7777        1.0          0.5
```

Annex2.8

Command to map pLDDT and accessibility values to motif matches:

```
Python        MotifMatches_pLDDT.py            ALIAS_MotifMatches_list.txt
       TgondiiME49_AF_plddt_values.txt  TgondiiME49_AF_dssp_values.txt
```

Motif match pLDDT and accessibility values table format:

```
Key                              pLDDT         Accessibility
TGME49_287280|CLV_C14_Caspase3-7|1     31.3800             0.7203
```

Annex2.9

Command to phosphosite to motif matches:

```
Python        MotifMatches_PTMs.py       ALIAS_MotifMatches_list.txt
       TgondiiME49_Phosphosites.tab
```

Motif match phosphosite table format:

```
Key                              modsNum             modsSites
TGME49_293300|CLV_C14_Caspase3-7|1     2             S96, S99
```

Annex2.10

Command to add domain mapping to motif matches:

```
Python        MotifMatches_Domains.py    ALIAS_MotifMatches_list.txt
       TgondiiME49_DomainMappings.tsv   TgondiiME49_DBMappings.tsv
```

Motif match domain mapping table format:

```
key                              doms_num      doms_name
TGME49_305460|DOC_PP2A_B56_1|1   0             NA
TGME49_305460|DOC_PP2A_B56_1|2   1             NAPeptidase_M24
```

Annex2.11

Command to combine all the motif matches information:

```
Rscript             MotifMatches_Enrichment.R ALIAS
```