PREDICTING HOST-PATHOGEN INTERACTIONS BETWEEN C. DIFFICILE 630 AND MOUSE

by

Sri Harsha Vishwanath

A Thesis Submitted to the Faculty of the

DEPARTMENT OF MICROBIOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2024

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by: Sri Harsha Vishwanath
titled: Predicting Host-Pathogen Interactions between C. difficile 630 and Mouse.

and recommend that it be accepted as fulfilling the thesis requirement for the Master's Degree.

Gayatri Vedantam — Date: Aug 13, 2024

VK Viswanathan — Date: Aug 13, 2024

Haiquan Li — Date: Aug 13, 2024

Arun K Dhar — Date: Aug 13, 2024

Fiona McCarthy — Date: Aug 14, 2024

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement.

Gayatri Vedantam — Date: Aug 13, 2024
SACBS

ACKNOWLEDGEMENT

I want to thank my advisor, Dr. Fiona McCarthy, for everything she helped me during my graduate career. Thank you for the long discussions and for encouraging me to keep up with my writing. I would also like to thank Dr. Gayatri Vedantam, Dr. V. K. Viswanathan, Dr. Arun Dhar, and Dr. Haiquan Li for their graciously imparted academic mentorship and scientific education.

I want to thank the members of the V & V labs – Dr. Anusha Harishankar, Allison Sullivan, Jason Lindsey, Dr. Jenny Roxas, Bryan Roxas, Katie Cocchi, and Dr. Farhan Anwar. Thank you for all the food, feedback, and help with research.

Finally, I want to thank my family and friends for their support and fortitude throughout my graduate career.

LAND ACKNOWLEDGEMENT

We respectfully acknowledge the University of Arizona is on the land and territories of Indigenous peoples. Today, Arizona is home to 22 federally recognized tribes, with Tucson being home to the O'odham and the Yaqui. Committed to diversity and inclusion, the University strives to build sustainable relationships with sovereign Native Nations and Indigenous communities through education offerings, partnerships, and community service.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	FULL FORM
ABC	ATP binding casette
AD	Activation domain
ADPR	ADP-ribosyltransferase
ANTXR1	ANTXR cell adhesion molecule 1
ANTXR2	ANTXR cell adhesion molecule 2
AP-1	Activator protein 1
APD	Autoprotease domain
CAMP	Cyclic AMP
CD	*Clostridoides difficile*
CDI	*Clostridoides difficile* infection
Cdt	*C. difficile* transferase
CdtA	Binary clostridial toxin A
CdtB	Binary clostridial toxin B
CNN	Convolutional neural network
CROPS	Combined repetitive oligopeptide sequences
CSPG4	Chondroitin sulfate proteoglycan 4
CWP	Cell wall protein
CXCL1	C-X-C motif chemokine ligand 1
DBD	DNA binding domain
DCS	Dendritic cells
DL	Deep learning
EF	Edema factor
ET	Edema toxin
FZD	Frizzled
GI	Gastrointestinal
GTD	Glucosyltransferase domain
GTP	Nucleotide guanosine triphosphate
HMW	High molecular weight
HPI	Host-pathogen interactions
HUPO	Human proteome organization
IL	Interleukin
LCT	Large clostridial toxins
LF	Lethal factor
LMW	Low molecular weight
LR	Logistic regression
LSR	Lipolysis-stimulated lipoprotein receptor
LT	Lethal toxin
MAB	Monoclonal antibody

MAPK	Mitogen-activated protein kinase kinases
MI	Molecular interactions
ML	Machine learning
MLP	Multilayer perceptron
MLST	Multilocus sequence typing
MSCRAMM	Microbial surface components recognizing adhesive matrix molecules
MYD88	Myeloid differentiation primary response 88
NAP	North American pulsed-field
NF-KB	Nuclear factor-κb
NLR	Nod-like receptors
NMR	Nuclear magnetic resonance
NOD	Nucleotide-binding oligomerization domain containing 1
PA	Protective antigen
PACSINS	Protein kinase C and casein kinase substrate in neurons
PCR	Polymerase chain reaction
PFGE	Pulse-field gel electrophoresis
PPI	Protein-protein interactions
PRR	Pattern recognition receptors
PSICQUIC	Proteomics Standards Initiative Common query interface
rCDI	Recurrent *C. difficile* infection
REA	Restriction Endonuclease analysis
RF	Random forest
RL	Reinforcement learning
ROS	Reactive oxygen species
RT	Ribotyping
SBP	Substrate-binding protein component
SLIM	Short linear motif
SLPS	Surface layer proteins
SVM	Support vector machine
TcdA	Toxin A
TcdB	Toxin B
TFPI	Tissue factor pathway inhibitor
TLR	Toll-like receptors
Y2H	Yeast two-hybrid

ABSTRACT

Clostridioides difficile (*C. difficile*) is the causative organism of hospital-acquired infectious diarrhea. *C. difficile* infection (CDI) causes more than 500,000 infections and 12,800 deaths. Antibiotic ablation of the commensal microbiome leads to colonization and infection in CDI patients. Symptomatic diarrheal infection requires the release of *C. difficile* toxins, TcdA and TcdB. Together, these toxins destabilize the colon epithelial cell membrane, causing fluid secretion into the colon, inflammation, and tissue damage. Molecular interaction databases describe *C. difficile* toxin interactions with host cell receptors, establishing these toxins as crucial virulence factors. However, there needs to be more curated and annotated information about non-toxin colonization and infection of these interactions in databases. Comprehensive identification of C. difficile toxin interactions and adhesion and colonization processes with host proteins will advance our understanding of *Clostridioides difficile* infection (CDI) and facilitate the design of targeted rational therapies.

In this study, we predict host-pathogen interactions (HPI) between mouse and *C. difficile* strain 630. We select mice as the host to predict interactions as these animal models provide genetic similarity to humans and can predict human disease responses. Selecting *C. difficile* 630 allows for a comprehensive study due to its well-characterized genome, clinical relevance, and representation of prevalent, virulent strains. The interolog-based approach is leveraged to provide host-pathogen interactions (HPI) rapidly. This method relies on sequence-based homology to transfer experimentally validated interactions to a mouse and *C. difficile* 630 model. The interolog approach predicts protein-protein interactions based on the assumption that if two proteins interact between species, their orthologs will likely

interact in another species' system. PSICQUIC interaction data is used to identify homologs in mouse-*C. difficile*. Mouse orthologs are found in human proteins using Ensembl BioMart. Next, pathogenic bacterial orthologs in *C. difficile* strain 630 are identified using reciprocal BLAST. This approach allows for the inference of potential interactions by mapping known interactions from one organism to another, yielding a set of 1,281 interologs; for extracellular pathogens such as *C. difficile* strain 630, extracellular or secreted proteins are expected to interact with host surface proteins. This pruning results in 100 interologs. CDI occurs in the colon, leading us to our next step: studying the colonic expression of host proteins. We exclude proteins not expressed in the colon as they do not contribute to the disease. Next, we manually assess *C. difficile* proteins in the predicted HPI for biological function and annotation. This results in 37 predicted HPIs. Network analysis of these HPIs further identifies 13 interactions with edge clustering for further investigation.

One predicted HPI shows an interaction between the *C. difficile* strain 630 sortase protein, CD630_27180, which cleaves bacterial surface proteins for adhesion, interacting with mouse protein, Ninjurin-1, an outer surface protein in the host implicated in immune functioning. This interaction is likely as these proteins are present externally in the bacteria. Another prediction identifies an interaction between CD630_03860, a secreted sortase protein, and Col12a1, a collagen protein expressed in the colon and implicated in maintaining cell junction stability. This HPI is likely to occur as CD630_03860 is a secreted protein that may allow for the breakdown of the epithelial tight junction, giving *C. difficile* access to the basolateral layer of colon epithelial cells.

While the interolog approach is rapid and cost-effective, it has some limitations. The interolog method relies on sequence homology to predict interactions. The predicted HPI quality in this study depends on the experimental data available on PSICQUIC. Experimental data on PSICQUIC is sourced from databases that are manually annotated. It only represents some of the available experimental data in the literature that could lead to meaningful predictions.

Additionally, predicted proteins may have different functions than those in the experimental data, leading to biologically incongruent predictions in CDI. For example, *C. difficile* strain 630 protein flagellin (FliC) interacts with TLR5 in humans. However, this interaction cannot be predicted in mice as TLR5 in mice binds different ligands in bacterial proteins. This research advances the application of interologs to HPI prediction by using pruning methods to refine interactions specific to *C. difficile* infection. However, alternative approaches to predict HPIs for *C. difficile* should also be investigated, such as machine learning. Ensemble machine learning in predicting HPI using protein sequence has been used to predict accurate host-pathogen interaction.

Chapter 1 Background and Introduction to *C. difficile* and Available Public Resources Relating to *C. difficile*

1.1 Overview of *Clostridoides difficile* Infection:

Clostridioides difficile is a Gram-positive anaerobe bacteria spore-forming bacillus and is the causative agent of *Clostridioides difficile* Infection (CDI) in the gastrointestinal tract[1]. The clinical presentation of CDI varies, ranging from asymptomatic infection to diarrhea and toxic megacolon. The onset of CDI occurs because of the oral-fecal transmission of sufficient spores of a toxin-producing strain of *C. difficile* within the host colon. A visual summarization of the onset of CDI colonization and infection is depicted in Figure 1[2]. CDI is viewed as a nosocomial infection; however, the hospital transmission rate has declined steadily. Concomitantly, community-associated CDI is increasing, with 53% of CDI infections reported as community-associated in 2019[3]. The US Centers for Disease Control classifies *C. difficile* as an urgent threat as CDI is associated with 500,000 infections and roughly 30,000 deaths annually in the US[4]. Inpatient costs of CDI to the US healthcare system are estimated to be nearly $5 billion annually. Recurrent CDI costs are estimated to be approximately $2.8 billion, placing a severe economic burden on this disease[6].

Positive CDI cases are treated with fidaxomicin, vancomycin, and metronidazole antibiotics. Traditional therapy poses two challenges to addressing CDI, as it can lead to antibiotic resistance and antibiotic-induced intestinal dysbiosis[7]. These challenges lead to recurrent CDI (rCDI), which presents the same challenges as fulminant initial infection. A fecal microbiota transplant is preferred over antibiotic treatment to treat rCDI. Various studies have reported that CDI and rCDI have significant detrimental effects on patients' quality of life that can have long-lasting and emotional impacts[5].

C. difficile is also an important pathogen of veterinary populations, with non-human neonates (piglets, calves, and foals) being particularly affected and susceptible to *C. difficile* infection within

1-14 days of birth[8]. In swine operations, the most significant numbers of *C. difficile* isolates are recovered from suckling piglets inside the farrowing barn[9]. On some farms, >70% of all piglets carry *C. difficile*, and a subset will succumb to the disease. Since neonates have underdeveloped immune systems, classic approaches such as vaccination are not used for CDI prevention. CDI can be treated with antibiotics (ceftiofur, enrofloxacin, apramycin). Due to the resulting gut dysbiosis, the development of a healthy gut microbiota is delayed. Surviving animals are asymptomatically colonized and shed *C. difficile*, contaminating the environment.

Three main events occur during CDI: disruption of the gut microbiota, colonization, and expression of virulence factors[10]. Antibiotic use is the most critical risk factor in CDI. Antibiotics such as ampicillin, amoxicillin, cephalosporins, clindamycin, and fluoroquinolones are routinely prescribed to treat bacterial infections[11]in humans. Prolonged antibiotic usage causes gut dysbiosis, where ablation of host commensal bacteria creates a niche for *C. difficile* to colonize and infect. *C. difficile* colonization is defined as the detection of the organism in the absence of CDI symptoms[12]. With *C. difficile* colonization, bacterial loads can be lower than those for CDI[13]. Infection with *C. difficile* is defined as the presence of *C. difficile* toxin or a toxigenic strain type and clinical manifestations with diarrhea, ileus (disturbed bowel function), and toxic megacolon[14].

Developing new therapeutic strategies is crucial for effectively treating CDI. The two primary challenges in managing CDI are the elevated risk of the bacteria developing antibiotic resistance and the detrimental effects of current treatments on the host's gut microbiome. To address these issues, it is essential to create therapeutics that specifically target the pathogenic mechanisms of *C. difficile*.

1.2 *C. difficile* Molecular Typing:

Molecular typing identifies the genomic relatedness between bacterial isolates, helping distinguish between them. Bacterial isolates can express different colonization and infection factors that can determine the severity of infection in the host. *C. difficile* isolates are classified using molecular typing techniques, aiding in epidemiologic investigations[15]. Polymerase chain reaction (PCR) - ribotyping, pulse-field gel electrophoresis (PFGE), and restriction endonuclease analysis (REA) are commonly used methods[16]. Genetic analysis tools such as Multilocus sequence typing (MLST) are used for evolutionary studies, outbreak detection, and transmission or population structure studies.[17] Using MLST, the known population of *C. difficile* is distributed in eight clades (Clades 1 to 5, plus Clades C-I, C-II, and C-III)[18]. Isolates from Clades 1-5 are associated with humans, and clades C-I to C-III are associated with non-toxin isolates from the environment.[19] *C. difficile* Strain 630 is a clinical isolate responsible for an outbreak of *C. difficile* infection at a Swiss hospital and belongs to PCR ribotype 012 (clade 1)[20]. This ribotype of *C. difficile* is the 8th most common in a recent European hospital-based survey[21] and 1.4% of US strains were found to belong to PCR ribotype 012[22]. After its genome sequence was published in 2006, 630 was rapidly adopted as the reference strain for laboratory-based studies. *C. difficile* strain 630 does not carry the *cdt* locus encoding the binary toxins. Since the early 2000s, the outbreak-associated strain ribotype 027 (RT027) (clade 2), also known as REA type BI or North American Pulsed-Field (NAP) type NAP1, has been linked to increased prevalence and persistence in clinical settings, as well as to causing severe disease[23]. *C. difficile* RT078 (clade 5) is a zoonotic strain associated with transmission between humans and animals such as pigs and cattle[24,25]. Isolates with this ribotype carry all three *C. difficile* toxins.

1.3 Transmission, Infection and Virulence Mechanisms of *C. difficile*:

C. difficile is transmitted via the oral-fecal route and is highly infectious between human hosts. The formation of spores facilitates this high transmissibility[26]. *C. difficile* produces highly infective endospores that infected patients excrete, allowing this oxygen-sensitive bacteria to retain viability outside the host. It is also critical to its life cycle[27]. Upon ingestion, *C. difficile* spores traverse the gastrointestinal tract, enduring the physical barriers of the esophagus and the acidic environment of the stomach[28]. Germination occurs in the host's duodenum, where primary bile acids, such as taurocholate and glycine, activate the CspC germinant receptor on the spores[29]. This process leads to the emergence of vegetative cells via a proteolytic cascade that leads to the breaking down of the spore peptidoglycan, the release of calcium dipicolonic acid, and rehydration of the spore, resulting in the outgrowth of the vegetative cell[30]. The vegetative cell attaches to the caecum and colon of the GI tract through adhesion factors. Vegetative cell attachment allows for colonization, infection, and sporulation of *C. difficile*. The expression and secretion of adhesion, toxin, and proteolytic proteins in this stage present as symptoms in the host.

C. difficile secretes toxins, including TcdA, TcdB, and binary toxins CdtA and CdtB, which cause damage to the colon epithelial lining. These toxins disrupt colonic epithelium cell stability, causing fluid secretion, inflammation, and tissue damage, characteristic of CDI[31]. In addition, *C. difficile* produces a wide range of virulence factors that mediate bacterial attachment to the mucosal surface and are vital in colonization. Table 1 summarizes virulence proteins known to affect host cells.

Table 1. Virulence Proteins and their Roles in Colonization and Infection.

Virulence Protein	Protein Name	Role in Colonization and Infection	Protein Interaction in Public Database
Spore germination related protein	CspC	Spore germination	No data
Toxin A	TcdA	Toxin	No data
Toxin B	TcdB	Toxin	Interactions available on PSICQUIC
Binary toxin	CdtA and CdtB	Toxin	Interactions available on PSICQUIC
Fibronectin binding protein	FbpA	Adherence	No data
Collagen binding protein	CbpA	Adherence	No data
Lipoprotein CD0873	CD0873	Adherence	No data
Surface layer protein	SlpA	Adherence	No data
Proteolytic activity	Cwp84	Adherence	No data
Surface protein	Cwp66	Adherence	No data
Zinc metalloprotease	Zmp1	Adherence	No data
Flagellin	FliC	Adherence and motility	No data
Flagellar cap protein	FliD	Adherence and motility	No data
Type iv pilin	PilA	Adherence and motility	No data
Heat shock protein 60	GroEL	Adherence and chaperone	No data

1.3.1 *C. difficile* Spore Proteins

The pathogenesis of *C. difficile* infection relies on the dormant spore. Because of the anaerobic nature of *C. difficile*, it cannot survive in aerobic environments in the vegetative form[32]. The *C. difficile* spore surface is divergent from other Gram-positive spore-forming bacteria such as *Bacillus subtilis*[33]. The structural layers of the *C. difficile* spore are composed of an exosporium and coat protein. The spore surface layer needs to be better characterized and varies between strains. The *C. difficile* spore surface interacts with the unidentified surface receptor(s) of intestinal epithelium cells[34], which plays a role in host-pathogen interactions. Germinant and co-germinant receptors on the coat layer of the spore facilitate spore germination to vegetative cells. CspC is activated by host bile salts, initiating a signaling cascade that leads to spore outgrowth and the rehydration of vegetative cells[28]. Additionally, CspC activates other co-germinant proteins, CspA and CspB. However, the interactions between CspA, CspB, and the host cell are not well-characterized[35].

1.3.2 *C. difficile* Toxins

Many *C. difficile* sequence types contain genes that encode up to three different toxins, which have been linked to the onset of clinical symptoms[36].

1.3.2.1 Large Clostridial Toxins: TcdA and TcdB.

Toxin A (TcdA) and toxin B (TcdB) are glucosyltransferases, which belong to the large clostridial toxin (LCT) family. When expressed and secreted within the colon, TcdA and TcdB bind host cell receptors, are endocytosed by host cells, and inactivate Rho-family GTPases via glucosylation[31]. Inactivation of Rho GTPases disrupts the host cytoskeleton and accelerates the breakdown of epithelium barrier function[37]. The toxins are encoded by the 'pathogenicity locus' PaLoc. The PaLoc is a 19.6-kb DNA region and encodes the large clostridial toxins A (TcdA) and B (TcdB),

the positive toxin regulator TcdR, the negative toxin regulator TcdC, the holin TcdE and the endolysin fragment TcdL[38–40].

TcdA and TcdB are 308 kDa and 270kDa, respectively. These toxins share a 47% sequence identity[41]. Both toxins contain four functional domains: an amino-terminal (N-terminal) glucosyltransferase domain (GTD), an auto protease domain (APD), a delivery domain with receptor binding, pore formation, and cargo translocation functions, and a domain formed by combined repetitive oligopeptide sequences (CROPS)[42]. The two toxins intoxicate host epithelial cells, beginning with toxins binding host cell receptors, and are endocytosed in either a PACSIN-dynamin manner (TcdA)[43] or clathrin, caveolae-mediated (TcdB)[44]. The endosome is acidified through proton accumulation, leading to pore formation and the rupture of the endosomal membrane, allowing the APD and GTD domains of the *C. difficile* toxins to enter the cytoplasm[45]. This is followed by the cleavage of the autoprotease domain and activation of the enzymatic GT domain that is catalyzed by cytosolic inositol hexakisphosphate[46]. Toxin A and B target the family of signaling G protein called Rho GTPases (e.g., Rho, Ras and Cdc42), where the GTD transfers glucose onto Rho GTPases. These modification switches cause cytopathic effect resulting from rearrangement of the actin cytoskeleton and can lead to apoptosis[44].

The TcdA and TcdB CROPS domains are hypothesized to mediate toxin attachment to the cell surface via glycan binding interactions[47(p3),48(p3)]. However, receptors for Toxin A (TcdA) are not well characterized in infection. Two glycoproteins are reported as receptors for toxin A - sucrase-isomaltase and soluble glycoprotein 96. Sucrase-isomaltase is not expressed in the colon[49] and glycoprotein 96 is found within the endoplasmic reticulum[50(p96)]. Toxin B binds different classes of proteins in the host - chondroitin sulfate proteoglycan 4 (CSPG4)[51], Frizzled proteins (FZD1,

FZD2, FZD7)[52] and Nectin 3 (previously known as poliovirus receptor-like protein 3' or PVRL3) [53] and tissue factor pathway inhibitor (TFPI)[54].

(a) Clostridial Toxins have Shared Functional Domains:

The domain architecture of TcdA and TcdB is structurally similar to that of other large clostridial toxins, which also target Rho GTPases and share similar mechanisms of entry and activation[41]. Although LCTs vary in their clinical manifestation, they all have highly similar structure and function. LCTs are high molecular weight (>200 kDa) single-chain polypeptides, sharing between 36 and 90% sequence identity[55]. To gain entry into cells and access cytosolic GTPases, LCTs utilize their multi-domain architecture, much like other AB toxin families, including diphtheria toxin and botulinum neurotoxin. In brief, using their central translocation and receptor-binding domain, LCTs bind cell-surface receptors and undergo receptor-mediated endocytosis[42]. The translocation pore facilitates the passage of the LCT glycosyltransferase (GTD) and cysteine protease (APD in *C. difficile*) into the cytosol, where the GTD is proteolytically released[55]. The glucosyltransferase domain in TcdA/B is found in other large clostridial toxins, such as Toxin B from *Clostridium perfringens* and Toxin B from *Clostridium sordellii*. The autoprotease domain is structurally and functionally like those in other bacterial toxins that use autoproteolysis for activation, such as the large clostridial toxins mentioned above. These similarities underscore the evolutionary conservation of structure and function among LCTs, reinforcing their role as potent and adaptable virulence factors[41].

1.3.2.2 Binary Toxin: CdtA and CdtB.

The binary toxin consists of two proteins, CdtA and CdtB, and is present in some clades (4 – 45% of strains) of *C. difficile*. CdtA and CdtB bind host cell receptors and catalyze the depolymerization

of actin. This toxin has been found in epidemic strains, including ribotypes 027 and 078, but not in ribotype 012 strains, which includes *C. difficile* 630[56].

The *C. difficile* transferase (CDT) is an ADP-ribosyltransferase (ADPR) two-component toxin encoded by a 6.2kb binary toxin locus (CDTLoc)[57]. The CDTloc contains genes that encode toxins, CdtA and CdtB, as well as the positive toxin regulator, CdtR. The binary toxins, CdtA and CdtB are 43 kDa and 99 kDa[58]. Host cell intoxication begins with CdtB binding to lipolysis-stimulated lipoprotein receptor (LSR)[59] and oligomerization to a heptamer[60]. The CdtB heptamer then engages the pADPRT of the CdtA toxin. This results in a CDT complex, which enters the cell by endocytosis. Endosome acidification triggers the translocation of CdtA into the cytosol[61]. CdtA ADP-ribosylates G-actin, which acts as a cap and inhibits its polymerization. The depolymerization of F-actin at the apical host cell surface promotes aberrant microtubule protrusion, supported by septin proteins[62]. ADP-ribosylation alters host cellular morphology, including the detachment of tight junctions connecting epithelial cells, and culminates in cellular rounding and epithelial tissue shedding[63].

1.3.3 Flagella:

Flagella are responsible for bacterial motility, evasion of host defenses, and colonization of the host cell surface. *C. difficile* flagellum consists primarily of FliC, a 39-kDa flagellar protein, and FliD, a 56-kDa flagellar cap protein, and plays a role in the motility and adherence of the bacteria to surfaces[64]. The components of the flagellum work together to drive bacterial motility. The motor proteins provide the power to the hook–basal-body structure to drive flagellum rotation. The basal body is an integral membrane protein complex[65]. Thus, the basal body with the motor protein is present within the bacterial cell membrane. Bacterial flagella have been implicated in contributing to bacterial pathogenesis by: (i) promoting adherence to host cells; (ii) providing force-driven

motility to nutrients; (iii) promoting biofilm formation; (iv) facilitating translocation of virulence factors across cell membranes; and (v) acting as immunomodulators by triggering proinflammatory cytokines through the Toll-like receptor 5 (TLR5) signaling pathway[66].

1.3.4 Type IV Pili:

The *C. difficile* genome encodes the type IV pili (T4P) system, consisting of nine different pilin genes, assembly, and scaffold proteins. The best-characterized genes are *pilA1* and *pilB1*, which encode the major pilin and pilus assembly ATPase, respectively[67]. *In vitro* studies with mutants of the pilus assembly genes exhibited a reduction in adherence to host cells, illustrating their role in pathogenesis[68].

1.3.5 Surface Layer Proteins:

The outermost surface of most bacteria and nearly all archaea is a 2D sheet of repeating surface-layer proteins or glycoproteins, known as an S-layer[69]. This layer is crucial for cell integrity, enzyme display, and interactions with the host immune system. The *C. difficile* vegetative cell has the typical Gram-positive cell envelope with a surface-exposed S-layer. The S-layer is decorated and functionalized by members of the Cell wall Protein (CWP) family[70]. In *C. difficile*, the S-layer is essential for intestinal colonization, sporulation, toxin production, and resistance to the innate immune system[71]. *C. difficile* S-layer is a para-crystalline layer composed of heterodimers and assembled from a reservoir of surface layer proteins[72]. The *slpA* gene encodes the precursor SlpA, which has three subdomains: an N-terminal secretion signal, low molecular weight SLP (LMW), and high molecular weight SLP (HMW). The SlpA N terminal is cleaved by the cysteine protease Cwp84[73] into HMW and LMW components. The HMW SLP is shown to bind to The LMW component, which is exposed to the environment and recognized by the immune system, showing significant antigenic variation between strains[74]. SlpA is required for interactions with intestinal

mucosal cells and activates Toll-like receptor 4-dependent immune responses. Host recognition of *C. difficile* involves MYD88 and NOD1 pathways[75].

1.3.6 Other Cell Wall Proteins:

The outer surface membrane of *C. difficile* is made up of cell wall proteins (CWP). These proteins are paralogs of the S-layer present in *C. difficile*[71]. Only a small number of these CWPs have been characterized in any detail but several have been shown to play crucial roles in the interaction between *C. difficile* and the host. Cwp66 is a 66 kDa protein with N-terminal CWB2 motifs. The C-terminal domain contains an apparently surface-exposed adhesin that can mediate adherence to Vero cells[76]. Cwp84 has a papain class cysteine protease domain and is involved in the cleavage of SlpA proteins into low molecular weight Slp and high molecular weight Slp[73]. Proteolytic enzymes such as Cwp84 are frequently involved in bacterial colonization process, serving to degrade host proteins including immunoglobulin, nutrient acquisition and processing bacterial proteins necessary in pathogenesis[77]. Purified Cwp84 exhibits proteolytic activity against fibronectin, laminin and type IV collagen, suggestive of a possible role in infection[78]. Another protease, Cwp13 is a paralog of Cwp84. Cwp13 partially substitutes Cwp84 role in SlpA cleavage[79]. CwpV is another identified cell wall protein with a potential function in conferring protection against host immune response or bacteriophage attack. However, further studies are required to confirm this role[80].

a. Sortase Anchored Proteins:

Sortases covalently attach surface proteins to the cell wall in several Gram-positive bacteria. These proteins are essential for nutritional acquisition and are often required for virulence[81]. Eight putative sortase substrates have been identified in *C. difficile* 630, although attachment to the cell wall has only been demonstrated for a few of these[82]. Sortase B, SrtB is a surface anchored sortase

that shows similarity to iron acquisition proteins in *S. aureus*, where the sortase protein recognizes a C-terminal tripartite signal sequence containing a highly conserved pentapeptide cell wall sorting motif, LPxTGz[82]. Proteins are then anchored to the cell wall via the catalytic action of a conserved cysteine residue of the sortase, cleaving the LPxTG motif between the threonine and glycine residues and, subsequently, covalently attaching the substrate protein to PG precursors[83]. Collagen-binding protein A (CbpA) has been identified as a putative sortase substrate due to the presence of a sorting motif (NVQTG)[84]. CbpA is surface exposed. CbpA belongs to the MSCRAMM family, which includes proteins that interact with the host extracellular matrix and display high affinity for collagens I and V, the most common components of fibrils[85]. Additionally, the collagen-binding protein CD2831 and the putative adhesin CD3246 both require sortase activity to attach to the bacterial cell wall[86].

b. FbpA:

Fibronectin-binding protein (Fbp68/FbpA) is another member of the MSCRAMM family and is surface-associated in *C. difficile*[87]. Fbp68 is a manganese-dependent fibronectin-binding protein capable of binding immobilized fibronectin and cultured Vero cells. Anti-Fbp68 antibodies have been found in CDI patient sera, suggesting that Fbp68 may perhaps be a useful component of a *C. difficile* vaccine[88].

c. Lipoprotein CD0873:

The *C. difficile* antigen CD0873 is annotated as a substrate-binding protein component (SBP) of an ATP-binding cassette (ABC) transporter and is an immunoreactive protein in human infection[89]. It is present at the bacterial cell surface and plays a role in the adherence of *C. difficile* to host cells[90].

d. Proline-Proline Endopeptidase-1 (PPEP-1):

PPEP-1 is a secreted zinc metalloprotease reported to cleave CD2831 and CD3246 to mediate better adherence to host cells. Furthermore, PPEP-1 can cleave fibrinogen *in vitro* and fibronectin produced by human fibroblasts, indicating a potential role of extracellular metalloproteases in attachment and infection[91].

e. *C. difficile* Heat Shock Proteins:

Heat shock proteins have been shown to be important for survival in the host for many pathogenic bacteria. GroEL is a member of the Hsp60 chaperonin family, and its expression is upregulated in response to all of these stresses[70]. GroEL acts as an adhesin and is associated with the cell surface despite lacking a signal sequence or clear mechanism for surface association. GroEL is immunogenic, and immunization with recombinant GroEL reduced intestinal colonization by *C. difficile* in mice[92].

1.4 Host Responses to *C. difficile:*

The host response to *C. difficile* is multifaceted, and in this section, we describe these different barriers to colonization and infection.

Physical barriers serve as the initial defense mechanisms against pathogenic attacks. Starting in the oral cavity, the host relies on the mucosal layer as a physical barrier, preventing pathogens from infiltrating the circulatory system[93]. Additionally, the gastrointestinal tract employs non-specific defense mechanisms, including the acidic environment, beginning in the stomach. Saliva in the host typically has an acidic pH of around 6.2 to 7.6, and it contains enzymes like lysozyme, which can break down bacterial cell walls, further contributing to the defense against pathogens[94]. When pathogens are ingested, they can enter the body through contaminated food or environmental

sources. As the ingested material travels to the stomach, it encounters the highly acidic pH of the gastrointestinal tract, primarily composed of hydrochloric acid. This acidity acts as a barrier against pathogens, such as *C. difficile* spores, which could be present due to contamination[95]. Although *C. difficile* spores are tolerant to oxygen, they still have the potential to germinate in the small intestine under specific conditions, potentially resulting in infection. Primary bile acids (specifically cholate derivatives, such as taurocholate, deoxycholate, and glycolate) induce spore germination and taurocholate is a potent germinator[13]. Bile acids are secreted into the duodenum of the small intestine. In the ileum of the small intestine, some of the bile acids are reabsorbed by bile salt transporters returning these acids to the enterohepatic recirculation system. The remaining bile acids are modified by microbiota-mediated transformations forming secondary bile acids. The gut bacteria perform α-dehydroxylation of bile acids forming 7-dehydroxylated secondary bile acids, deoxycholate, and lithocholate which are toxic to *C. difficile* vegetative growth[96].

Vegetative *C. difficile* cells colonize and infect the gut lining of the host. Colon epithelial cells provide a physical barrier to prevent bacterial infiltration of the gut. These cells also sense the presence of pathogenic and commensal microbes through a variety of pattern recognition receptors (PRRs), including Toll-like receptors (TLRs) and Nod-like receptors (NLRs)[97]. Epithelial cells also release retinoic acid (RA), a breakdown product of β-carotene, interleukin (IL)-33, IL-25, and other cytokines to modulate the function of different immune cells, including adaptive immune cells such as the dendritic cells (DCs), and T and B cells. In return, these immune cells provide signals to epithelial stem cells located at the base of the intestinal crypts needed for their proliferation and differentiation[98]. The toxins, TcdA and TcdB activate activation of nuclear factor-κB (NF-κB) and activator protein 1 (AP-1) which in turn secrete pro-inflammatory cytokines and chemokines, such as interleukin 1 (IL-1), IL-8, and CXCL1(C-X-C motif chemokine

ligand 1)[99]. Similarly, the binary toxin CdtB is recognized by TLR2 on eosinophils and was able to initiate the NF-κB pathway[100]. The TLR recognizes non-toxin proteins such as SlpA and FliC. SlpA is recognized by the TLR4, which further stimulates macrophage clearing of bacteria[101]. Purified *C. difficile* FliC specifically acts on TLR5 – through which it induces the NF-κB and P38 activation, and to a lesser degree Erk/2 and JNK MAPKs activation, to stimulate the production and secretion of cytokines[102],[103]. The classical symptom of pseudomembranous colitis in severe CDI is associated with epithelial tissue damage and heavy inflammation from neutrophil infiltration[104].

Adaptive immunity against *C. difficile* colonization has been studied for antibodies against the toxins. Antibodies to TcdA and TcdB have been poor as they do not protect from colonization, but they influence disease susceptibility and, subsequently, the progression from colonization to CDI[13]. For example, anti-toxin A and B antibodies have been associated with protection against recurrent CDI[105]. Vaccination strategies based on antibodies against non-toxin such as SlpA and FliC proteins show protective roles against colonization but do provide complete immunity against the disease[13].

1.5 Animal Models to Study *C. difficile* Infection:

Animal models of CDI have been extensively used for research. Mammalian models such as hamsters, mice, and piglets are commonly used to study the pathogenesis of CDI[106].

1.5.1 Hamster Models

The hamster model of *C. difficile* infection is used in many different areas of research, including the induction of *Clostridioides difficile* infections in experimental models, the evaluation and testing of new therapeutic treatments, the study of population dynamics of microbial communities,

and the investigation of host-pathogen interactions and immune response[107]. In the hamster model, the disease is induced by the administration of antibiotics, which disrupt the normal gut flora; following infection with *C. difficile*, hamsters display many of the pathophysiological features seen in humans[108]. In addition to the overall deterioration in the health of the hamster, there are changes in the appearance of the gastrointestinal tract, which often appear inflamed. The inflammation may manifest as redness, swelling, and the presence of lesions or ulcers on the mucosal surface[109]. This rapidly and uniformly fatal disease pattern is not characteristic of human *C. difficile*, and a key drawback of the model is that hamsters do not typically develop diarrhea. They may occasionally develop a "wet tail," in which the hamster displays symptoms of watery diarrhea, lethargy, irritability, and refuses food, but invariably, this leads to death[110],[111]. Thus, in the context of *C. difficile* treatment experiments, the hamster model is a prevention of death model.

1.5.2 Mice Models

The use of mouse models to study CDI is increasing, mainly due to improved methods of inducing disease susceptibility in mice and the greater availability of mouse-specific reagents to perform detailed host tissue analysis. Untreated mice are relatively resistant to infection with *C. difficile* and do not develop fatal infections[112]. This is most likely due to the colonization resistance provided by the resident microbiota, although these mice can become asymptomatic carriers that persistently shed low numbers of spores[112]. An intoxication model was developed in which toxins were administered to mice intrarectally. This direct delivery of toxins into the colon led to inflammation, upregulation of cytokines and chemokines, and increased tissue damage[104]. In addition to this intoxication model, three different mouse *C. difficile* infection models have been described. The first employs gnotobiotic/germ-free mice[113], the second uses a cocktail of

antibiotics to disrupt the normal gut microbial communities and predispose the mice to infection, and the third uses a single antibiotic to induce susceptibility to CDI[106].

Gnotobiotic mouse models do not need antibiotics to disrupt the gut microbiota, making these models advantageous in understanding the innate immune response. However, since germ-free models begin with no gut microbiota, this model does not reflect the normal situation in humans and animals[106]. The antibiotic cocktail model mirrors key features of CDI in humans, including diarrhea, weight loss, and histological damage. Antibiotic cocktail models showed histopathological outcomes to human disease, such as proliferative ulcerative enteritis with superficial epithelial necrosis[106]. It also showed extensive submucosal edema without submucosal inflammation, mucosal proliferation, and inflammatory cell influx[106].

Single antibiotic mouse models closely resemble human disease as CDI in humans is induced by a single antibiotic to alter the gut microbiome. Mouse models are injected with a single antibiotic, such as clindamycin, cephalosporin, and cefoperazone. Pretreatment of mice with either antibiotic induces susceptibility to CDI and results in disease characterized by diarrhea, weight loss, mortality, and colonic or cecal pathology[106].

A disadvantage of mouse models is the variability in individual species. Mice infected with the same strain have reported different disease manifestations with varying diarrhea symptoms[114]. Disease dosage has been attributed to this variability, as seen in human CDI[106].

While the hamster model provided the foundation for *C. difficile* research and continues to be a useful and important model for studying *C. difficile*, the development of new mouse models, combined with wide access to mouse-specific reagents and tools, offers new opportunities to study subtle features of the disease[106].

1.5.3 Piglet Models

Piglets infected with *C. difficile* mimic key characteristics of the disease observed in humans, and thus, this model is used in investigating why some strains are associated with more severe *C. difficile*[107]. In a study by Steele *et al*.,[115] the piglet model was validated to be a reproducible model of acute or chronic CDI with characteristic pseudomembranous colitis. The disease's clinical manifestations, including gastrointestinal and systemic symptoms and characteristic mucosal lesions of the large bowel (including pseudomembranous colitis), are described. Additionally, the model demonstrated the presence of toxins in feces, body fluids, and serum and a significant elevation in interleukin 8 levels in animals with severe disease. Thus, the piglet model for CDI is suitable for investigating the role of virulence attributes in CDI, drug efficacy, and identifying vaccine candidates[115].

1.6 *C. difficile* Therapeutics:

Antibiotics such as vancomycin, metronidazole, and fidaxomicin are standard treatment options for patients with confirmed CDI. Vancomycin and fidaxomicin are standard-of-care treatments prescribed during severe cases of CDI. Vancomycin inhibits cell wall synthesis in Gram-positive bacteria, resulting in cell lysis. However, prolonged use can contribute to antibiotic resistance. Fidaxomicin blocks RNA transcription, preventing protein synthesis and cell function[116]. Metronidazole disrupts the DNA synthesis in anaerobic bacteria but can cause neuropathy in the patient. Metronidazole is no longer recommended as first-line therapy for *C difficile* infections and is prescribed in cases where vancomycin or fidaxomicin are unavailable[116]. Although often effective, antibiotic treatment prolongs the state of dysbiosis of the intestinal microbiota, and there is a high rate of recurrent disease[117]. Additionally, antibiotics do not directly affect the metabolically inactive *C. difficile* spores that serve as a reservoir for recurrent infection and

symptomatic disease[118]. Another therapeutic used to treat CDI is fecal microbiota transplantation is effective against refractory and recurrent CDI, but has inherent risks associated with the lack of standardization[119]. However, in May 2023, the US Food and Drug Administration approved Seres Therapeutics' SER-109, an oral microbiota therapy, to prevent the recurrence of *Clostridioides difficile* infections[120]. SER-109 is prepared from the stool of healthy donors screened to exclude the presence of a panel of known pathogens. Spores from Bacillota are purified from the donor samples and are used to manufacture the pills[120] Another emerging antibiotic against *C. difficile* is ridinilazole, a narrow spectrum antibiotic. It has demonstrated a lower propensity for collateral damage to the gut microbiome and appears to diminish the production of *C. difficile* toxins and subsequent bowel inflammation, which may prove advantageous in managing severe CDI. However, like vancomycin, it can potentially disrupt the gut flora[121].

Vaccination assays using serum antibodies against *C. difficile* surface components have been used in animal trials therapies. Parenteral or mucosal vaccination with the S-layer proteins led to specific antibody production but only partial protection in the hamster model[122],[123]. Immunization studies with Cwp84 and the flagellar proteins FliC and FliD administered to animals by the mucosal route resulted in a significant decrease in intestinal *C. difficile* colonization in the mouse model and partial protection in the hamster model[124]. These results suggest that antibodies against *C. difficile* surface proteins have a protective role against colonization. At the moment, studies with surface protein-based vaccines to prevent colonization in humans are lacking. Vaccination trials with the two toxins or toxin fragments have not been successful. Sanofi Pasteur recently announced the cessation of its vaccine development program, which was based on toxin antigens alone[125].

Monoclonal antibody (Mab)-based passive immunotherapy directed to toxins was able to protect hamsters from CDI. In humans, two MAbs, one targeting TcdA (actoxumab) and another targeting TcdB (bezlotoxumab), were tested in human clinical trials aimed at the prevention of recurrent disease[126]. Bezlotoxumab is an IgG Mab that binds to toxin B and neutralizes its effects on Mammalian cells. It prevented approximately 40% of recurrences presumably due to limiting epithelial damage and facilitating rapid microbiome recovery[127]. The Clinical Practice Guideline, issued by the Infectious Diseases Society of America, recommends using bezlotoxumab in addition to standard-of-care antibiotics for adult patients with a recurrence of CDI within the past 6 months[128]. However, while effective at reducing recurrences, it does not treat the initial infection and must be used alongside standard antibiotic therapy and specifically targets toxin B, but there are other factors involved in CDI. Additionally, bezlotoxumab is prescribed for patients at high risk of recurrence and is not indicated for all CDI patients.

1.7 Available Resources for *C. difficile*:

To facilitate the study of *C. difficile*, multiple genomic and proteomic resources are available. These resources enable researchers to analyze the many isolates of *C. difficile*, aiding in the understanding of disease spread and the characteristics of different strains.

1.7.1 DiffBase

DiffBase is an online database that classifies Toxin A (TcdA) and Toxin B (TcdB) protein sequences from *Clostridioides difficile* strains. It groups these toxins into distinct subtypes based on sequence similarities through a bioinformatic analysis pipeline. DiffBase contains a phylogenetic analysis of 8,839 *C. difficile* strains for researchers to view and analyze. The goal of this database is to classify TcdA and TcdB into subtypes that allow clinicians and researchers to

categorize and predict functional-immunological variations of future sequenced *C. difficile* isolates[36].

1.7.2 Webribio

Webribio is a web-based database developed in 2008 to support PCR ribotyping[129]. The method involves PCR amplification using the same primers as those used for agarose gel-based ribotyping, with the addition of a fluorescent label on one of the primers. Amplicon sizes are determined using an ABI genetic analyzer. The Webribio database allows users to adjust sequencer settings and primer pairs. By uploading sequencer data files, users can determine known ribotypes[129].

1.7.3 PubMLST

This database contains typing data collected from MLST analysis, isolates with phenotype data, and genome assemblies for some of the isolates in the database. The PubMLST.org website hosts a collection of open-access, curated databases that integrate population sequence data with location and phenotype information for over 100 different microbial species and genera. The databases employ population genomics along with genomic and typing information, which is useful for epidemiological studies[130]. *C. difficile* is one of the microbial species hosted in PubMLST.

1.7.4 BioCyc

BioCyc.org is a microbial genome web portal that combines thousands of genomes with additional information inferred by computer programs, imported from other databases (DBs), and curated from biomedical literature by biologist curators. BioCyc also provides extensive query tools, visualization services, and analysis software. *C. difficile* BioCyc is a part of the larger BioCyc collection. It currently hosts 16 *C. difficile* genomes and associated metabolic pathways. BioCyc compiles data on *C. difficile* strain 630. The pathway information on this resource is based on *C. difficile* 630[131].

1.7.5 EGRIN

EGRIN is the Environment and Gene Regulatory Influence Network, which provides which provides a blueprint for the functioning of *C. difficile* 630 functioning[132]. *C. difficile* adapts to the complex gut environment by modifying its metabolism through a gene regulatory network that responds to various stimuli. *C. difficile* Web Portal is a compiled transcriptional compendium. This resource compiles a wide array of context-specific transcriptomic data for virulent, multidrug-resistant *C. difficile* 630 clinical strain from public repositories. The transcriptional network created in this resource informs EGRIN.

1.7.6 PSICQUIC

PSICQUIC (Proteomics Standards Initiative Common QUery InterfaCe) is a standard web service developed by the HUPO Proteomics Standards Initiative to facilitate integrating and retrieving molecular interaction data from multiple databases[133]. It provides a unified interface for querying diverse data sources, enabling researchers to access comprehensive interaction information quickly. PSICQUIC supports various data formats and promotes data sharing and interoperability among different bioinformatics platforms. This tool is crucial for researchers studying protein-protein interactions, host-pathogen interactions, and other molecular interactions[133],[134]. PSICQUIC contains only 11 protein-protein interactions (PPI) that describe host-pathogen interactions between Clostridioides difficile and humans. These interactions are limited to those involving the toxins TcdB (Toxin B) and CdtB (Binary Toxin B), highlighting a significant gap in our understanding of the broader range of PPIs that may occur during *C. difficile* infections. This limited dataset underscores the need for further research to identify and characterize additional host-pathogen interactions, which could provide deeper insights into the mechanisms of *C. difficile* pathogenicity.

1.8 Host-Pathogen Interactions:

Molecular interactions (MI) form the basis of all biological interactions. Pathogenic infectious mechanisms and host response are mediated via molecular interactions. For example, pathogen-host adherence can occur through molecular interactions. When encountering pathogens, the host responds by releasing reactive oxygen species (ROS), highly reactive molecules derived from oxygen[135]. Protein-protein interactions refer to the specific interactions between two or more proteins. These interactions are essential for various biological functions, including signal transduction, metabolic pathways, and structural support[136]. Host-pathogen interactions (HPI) are a dynamic process that includes the pathogenesis of infectious mechanisms and the host response. Pathogenic infection involves different stages, beginning from entry to host, attachment, invasion of host cell, and proliferation of the causative organism[137]. Host-pathogen interactions underpin infectious disease research as pathogens target host cellular and defense responses to ensure their survival and propagation. Molecular mechanisms mediate these interactions. Identifying these molecular mechanisms provides targets for pharmacological intervention in developing therapeutics and prophylactics[138]. Molecular interactions may be between proteins, nucleotide sequences, and small ligands of hosts and pathogens by physical associations.

1.8.1 Protein-Protein Interactions

A protein-protein interaction (PPI) refers to a physical interaction between two proteins. PPIs are fundamental for biological processes[136]. These interactions are the underlying mechanism of almost all cellular processes, they mediate signaling pathways, structural configurations, and metabolic networks. In terms of physical interactions, proteins directly associate with other proteins, such as those of protein complexes, or via phosphorylation as part of signal transduction. In indirect interactions, proteins regulate functions in cellular pathways[139]. Uncovering the PPIs

involved in the pathogenesis of a disease can help understand its development and progression, leading to the identification of potential diagnostic and therapeutic targets.

1.8.2 Protein-Protein Interaction Detection Methods

Protein-protein interaction detection methods are classified into three types: *in vitro*, *in vivo*, and *in silico*[140]. In *in vitro* techniques, a given procedure is performed in a controlled environment outside a living organism. The *in vitro* methods in PPI detection are tandem affinity purification[141], affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation[140], X-ray crystallography[142], and NMR spectroscopy[142]. In *in vivo* techniques, a given procedure is performed on the whole living organism itself. The *in vivo* methods in PPI detection are yeast two-hybrid (Y2H)[143]. *In silico* techniques are performed on a computer (or) via computer simulation. The *in silico* methods in PPI detection are sequence-based approaches, structure-based approaches, phylogenetic trees, and gene expression-based approaches.

1.8.2.1 Yeast Two-Hybrid Analysis

The Y2H method is an *in vivo* method applied to detecting PPIs[144]. Two protein domains are required in the Y2H assay, which will have two specific functions: (i) a DNA binding domain (DBD) that helps bind to DNA and (ii) an activation domain (AD) responsible for activating transcription of DNA. Both domains are required for the transcription of a reporter gene[145]. Y2H analysis allows the direct recognition of PPI between protein pairs. However, the method may incur a large number of false positive interactions[143].

1.8.3 Interaction Databases

The massive quantity of experimental PPI data generated on a steady basis has led to the construction of computer-readable biological databases to organize and process this data.

Multiple databases house information about these interactions in multiple organisms from various sources, including experimentally validated sources, computational predictions, and automated curations[139]. MINT[146], STRING[147], and BioGRID[148] contain comprehensive information about proteins and protein-protein interactions for a wide range of organisms. IntAct[149] expands upon molecular interactions, such as those involving DNA, RNA, and small molecules. DIP[150] stores information on binary protein interactions with the structure of molecular complexes. HPRD[151] is a collection of protein interactions between human proteins with extensive information, and I2D[152] enhances predicted protein interactions by integrating homologous interactions across species. In addition, BIND[153] contains information on molecular interactions within molecular complexes and pathways, and MPact[154] contains manually curated yeast protein interactions. HPIDB[138] is a curated database that contains host-pathogen interaction data.

The proliferation of molecular interaction databases has led to fragmented and difficult-to-access interaction data. In order to share interaction data and make it accessible to researchers, a global effort was made by these databases to form the International Molecular Exchange (IMEx) Consortium[155]. The IMEx consortium was developed with the goal of providing users with a dataset enhanced with controlled vocabulary (CV) terms to enable scoring, filtering, and sophisticated searching of the information. The full dataset of participating IMEx databases is available through the Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC)[133]. PSICQUIC currently integrates 11 million interactions from 34 interaction databases, encompassing molecular data, including protein-protein interactions and the more specific host-pathogen interactions[134].

1.8.4 Host-pathogen Interaction Data on PSICQUIC

Host-pathogen interaction data described by PPI is accessible on PSICQUIC. Of the 11 million interactions, every 2 out of 5 interactions describe HPI. Several PPI databases are solely created to study host-virus PPI[156,157]. Host-bacterial interactions are underrepresented in molecular interaction databases. For example, out of 11,957,574 interactions available in PSICQUIC, only 3,214 represent Bacillota-host interactions. Bacillota is a phylum within the domain of Bacteria. *Clostridioides difficile* is present within this phylum. A phylogenetically related species to *C. difficile* is *Bacillus anthracis*. *Bacillus anthracis* – human interactions comprise most of the Gram-positive bacterial host-pathogen interactions available on PSICQUIC[158]. This section briefly elucidates the *B. anthracis* infection and its relevance in CDI research.

Bacillus anthracis:

Bacillus anthracis is a spore-forming, Gram-positive bacteria and the causative organism of anthrax, an infectious disease affecting humans and mammals[159]. Anthrax infection occurs when the *B. anthracis* endospores enter the body. Depending on the mode of transmission of these endospores, anthrax can lead to various clinical illnesses, including cutaneous, gastrointestinal, inhalation, and injection anthrax[160]. Among these, inhalation anthrax is the most extensively studied, as it occurs when aerosolized spores are inhaled. In October 2001, letters containing anthrax spores were sent through the United States Postal Service to politicians and media offices in Washington DC, New York, and Florida. The anthrax spores aerosolized from these letters, infecting at least 22 people, resulting in five deaths[161]. The use of anthrax as a bioweapon and its high lethality has made *Bacillus anthracis* a very well-studied pathogen. Gastrointestinal (GI) Anthrax is an acute infectious disease resulting from ingesting *B. anthracis* spores[159]. Ingested spores germinate within the mammal host to produce the vegetative forms and clinically manifest

as ulcerative lesions accompanied by gastrointestinal bleeding, fluid loss, abdominal pain and diarrhea[162].

B. anthracis Pathogenesis Mechanisms:

B. anthracis harbors two large virulence plasmids, pXO1 and pXO2, transcribing proteins that cause toxemia and subsequent septicemia[160]. Anthrax tripartite toxin is composed of the Protective Antigen (PA), Edema Factor (EF), and Lethal Factor (LF). This section uses toxin abbreviations repeatedly; Table 2 summarizes the toxins with their abbreviations used in this thesis. Intoxication of the host cells begins when PA binds either ANTXR cell adhesion molecule 1 (ANTXR1)[163] or ANTXR cell adhesion molecule 2 (ANTXR2)[164]. Table 3 provides a list of human receptors binding protective antigens. PA also binds β1 integrins, which enhance the uptake of PA by macrophages[165]. PA binds host receptors and is cleaved by furin like proteases to form PA63, which oligomerizes and facilitates internalizing of the EF and LF to host cells. Protective antigen oligomerization is necessary for its association with host lipid rafts and clathrin-mediated endocytosis of toxins[166]. When bound to the PA, the EF is referred to as Edema Toxin (ET), and the LF is referred to as Lethal toxin (LT). ET causes a steady elevation in cyclic AMP (cAMP), resulting in activation of signaling pathways through protein kinase A, causing vascular dysfunction and hemorrhaging[167]. LT is a zinc metalloprotease that cleaves most isoforms of mitogen-activated protein kinase kinases (MAPKKs) and prevents the activation of Erk1/2, p38, and JNK pathways[168],[169]. The exotoxins act on several pathways, affecting several physiological functions. Ultimately, toxemia is caused by the secretion of lethal toxin (LT) and edema toxin (ET) encoded by pXO1, and septicemia is caused by an anti-phagocytic capsule produced by gene products encoded by pXO2[170].

Table 2. Anthrax Toxins and their Abbreviations.

Anthrax Toxin Name	Abbreviation
Protective Antigen	PA
Edema Factor	EF
Lethal Factor	LF
Edema Toxin	ET
Lethal Toxin	LT

Table 3. Anthrax Toxins and their Known Receptors in Humans.

Binding with Anthrax Toxin	Human Receptor
Protective Antigen (PA)	ANTXR cell adhesion molecule 1 (ANTXR1)
Protective Antigen (PA)	ANTXR cell adhesion molecule 2 (ANTXR2)
Protective Antigen (PA)	collagen type VI alpha 3 chain (COL6A3)

Anthrax toxin receptors, TEM8 and CMG2 are expressed on many tissues, enabling the exotoxin to act on several systems. These receptors are highly expressed on intestinal epithelia and immune cells[171]. Additionally, the PA binds collagen type VI alpha 3 (COL6A3) and collagen type IV[172],[173] on the epithelial surface. PA binding facilitates secretion of the LT into host cells, which arrests proliferation and induces cytoskeletal rearrangement, impacting intestinal integrity and presenting as intestinal colitis[174].

1.8.5 Computational Protein-Protein Interaction Prediction

Predicting protein-protein interactions can guide early drug development, which is pivotal to determining the clinical application and to initiate drug development campaigns. Since drug discovery experiments are time and cost-intensive, computational efforts to rapidly predict PPI have been developed and continually improved. By predicting putative PPI targets, effective drug development campaigns have been initiated[175].

Prediction methods can be broadly categorized based on this approach:

1. Homology-based approach – interaction between a pair of proteins in one species is expected to be conserved in a related species[176].
2. Structure-based approach – In this approach, structural information is used to identify the similarity between query proteins and infer interactions from template protein-protein interactions[177].
3. Domain/motif interaction-based methods - These methods identify potential protein-protein interactions (PPIs) by matching known interacting protein domains[178]. In motif, some protein interactions are mediated not by interactions between domains but by interactions between a domain in one protein and a short linear motif (SLiM) in the other protein[179].
4. Machine learning-based predictions – This method uses input data such as sequence features, structural information, and functional annotations to build predictive models to infer potential PPI. By training on known interaction datasets, machine learning algorithms can identify patterns and relationships indicative of protein interactions. This approach enables the prediction of novel PPIs with high accuracy, aiding in the understanding of cellular processes, disease mechanisms, and the discovery of new therapeutic targets[180].

PPI prediction was developed to identify intraspecies interactions. However, these approaches can be applied to identify host-pathogen interactions. Experimental methods are often time-consuming and laborious, making it unfeasible to detect all possible host-pathogen PPIs. Therefore, to meet the urgent need to predict HPIs, computational prediction approaches to identify interactions have been implemented. Traditional approaches to predict PPI, such as sequence homology, domain-domain interaction, and domain-motif interaction-based methods, have been directly adapted to predict HPI[181].

1.8.6 Host-Pathogen Protein-Protein Interaction Prediction Using Interologs

An interolog is a conserved interaction between two proteins across different species. If two proteins in one species are known to interact, their homologs (proteins with common ancestry) in another species are predicted to interact similarly[176]. In this approach, an interaction between a pair of proteins in one species is anticipated to be conserved in its related species. This approach can be adapted to host-pathogen interaction prediction, where an interaction is inferred in the query species from a known PPI. Homologs to the known PPI can be said to be interacting. The investigated host-pathogen systems in past studies include *H. sapiens – Helicobacter pylori*[182], *H. sapiens – E. coli*, and *H. sapiens – Yersinia pestis*[183].

The interolog approach has broad applicability, wherein it can be applied to a wide range of organisms, including those with limited experimental data, and allows for rapidly predicting numerous HPIs. The quality of interologs inferred can be impacted by sequence homology, where predictions may be less accurate for distant homologs with significant divergence. Additionally, homology does not always confirm an interaction. This approach can result in interactions that do not physically occur. Finally, the accuracy of the prediction depends on the quality of interaction data from which interologs are predicted.

1.8.7 Network Analysis

Network biology enables a detailed understanding and analysis of biological components by mapping out molecular interactions. Protein-protein interactions are linked extensively in gene regulation, signaling, and transportation processes. By creating intricate maps of these interactions, network biology provides insights into how different molecules within a cell interact with each other, forming complex networks that govern cellular functions. Protein-protein interactions are linked extensively in gene regulation, signaling, and transportation processes. The overall proteome level of interactions constitutes an "interactome"[184]. Networks consist of systems' components, called nodes, and interactions, termed 'edges'. In a protein-protein interaction, nodes are the interacting proteins in the interactome, and edges are the interactions between the proteins. Edges show a direct relationship between proteins.

Elucidating an interactome's physical characteristics and functional interaction properties could reveal novel relationships between host proteins and show proteins targeted by pathogens[184]. Such structural and functional topological features offer valuable insights into the specific roles of proteins and the overall network. Analyzing the network architecture and understanding these topological properties can lead to discovering novel components within complex systems, thereby providing significant biological insights[184]. For instance, network architectural properties can determine the connectivity and the critical distribution of a particular node within a network. These include degree, the number of connections of a node, and betweenness. Network properties identify nodes that can be highly connected. These nodes are called hubs and have been proposed to play important roles in biological processes[185].

1.8.8 Subcellular Localization of Host and Pathogen Proteins

The localization information of pathogen and host proteins may relate to the possibility of their interactions. For extracellular pathogens, secreted proteins present in the extracellular region and proteins with translocation signals are more likely to interact with host extracellular or membrane proteins[186]. This subcellular localization information is often used in pruning of predicted host-pathogen PPIs. In the case of intracellular pathogens, the host proteins interacting with the pathogen are more likely to be found in the same location as the pathogenic proteins[187].

1.9 Rationale and Goal for Interolog Prediction:

Clostridioides difficile (*C. difficile*) is a public health threat, and there is a critical need for easily accessible and comprehensive information for researchers. Although several resources are available to support *C. difficile* research, many are limited in scope and coverage. For instance, resources such as DiffBase and MLST[130] offer valuable genomic information, with DiffBase specifically focusing on toxin information. However, these resources do not encompass the full range of *C. difficile* genomic and proteomic data.

Limitations of Existing Data

BioCyc[131] and EGRIN[188] are metabolic databases that provide information based on only one strain of *C. difficile* 630. While these databases offer some insights into the metabolism of this bacterium, they do not represent the complete genomic and proteomic diversity of *C. difficile*. Additionally, PSICQUIC is a general protein-protein interaction (PPI) database that is not specific to *C. difficile*. It requires manual annotation for data availability, and its coverage of *C. difficile* interactions is limited, failing to capture the full spectrum of protein interactions related to *C. difficile*, currently hosting only 11 PPI that describe toxin interactions with host proteins.

Existing research on CDI has predominantly focused on the role of toxins, such as toxin A and toxin B, as targets for therapeutic interventions[125]. While bezlotoxumab has been identified as a target for Toxin B in CDI caused by strains with toxin A, toxin B, and the binary toxins, therapeutics such as bezlotoxumab is not recommended and cannot be used because it is not intended for initial treatment, is limited to reducing recurrence rather than addressing the infection directly[116]. This toxin-centric approach has significantly informed vaccine development efforts, but unfortunately, these attempts have largely failed to produce effective vaccines or treatments[189]. This situation underscores the need for a broader exploration of HPIs that extend beyond the well-characterized toxin mechanisms.

Project Goals and Objectives

This project aims to address this research gap by predicting a range of novel HPIs that could be crucial for CDI pathogenesis but have not been thoroughly explored in the existing literature. By identifying these potential interactions, we seek to provide a foundational resource for researchers to explore new avenues for therapeutic development. This work will offer a "first-pass" overview of promising HPIs for further testing and validation, with the ultimate goal of guiding the rational design of innovative treatments and vaccine candidates for CDI.

The overarching objective of this project is to predict HPI to develop a more comprehensive approach that considers other critical factors in CDI. By broadening the scope of research, this project aims to open new opportunities for effective therapeutic interventions and advance the development of novel treatments and vaccines for this pressing public health issue.

Chapter 2 Predicting Interologs between Mouse and *C. difficile* 630

2.1 Introduction

Clostridioides difficile is a gram-positive, anaerobic spore-forming bacteria and the causative agent of *Clostridioides difficile* Infection (CDI) in humans and animals[75]. Symptomatic infection causes mild to severe diarrhea and can result in life-threatening conditions such as pseudomembranous colitis and toxic megacolon[190]. CDI is the leading cause of hospital-acquired infectious diarrhea. It results in over 200,000 hospitalizations annually and costs the United States $1 billion in healthcare expenses annually[4].

Three main events occur during CDI: disruption of the gut microbiota, colonization of the causal bacterium, and expression of virulence factors[10]. Positive CDI cases are treated with antibiotics such as fidaxomicin, vancomycin, and metronidazole. As antibiotic treatment is burdened with an important risk of relapse, alternative therapies must be explored[191]. A strategy to rapidly investigate new interactions to identify potential therapeutic targets will aid in the process of drug development. Detailed knowledge of *C. difficile* interactions with host cells can provide more strategies for intervention. *C. difficile* toxins are well-established virulence factors leading to host cell destabilization[31], and toxin interactions are well documented on molecular interaction resources such as PSICQUIC[133]. However, the processes of adherence and colonization are equally critical for the pathogenesis of *C. difficile*[192]. Although there is literature evidence suggesting that *C. difficile* recognizes host cells through accessory enzymes and toxin host cell binding, this information is not available in interaction databases.

Targeting critical host-pathogen interactions (HPI) informs successful hypothesis generation in rational therapeutic and prophylactic design. Comprehensive interaction knowledge on *C. difficile* colonization, infection, and pathogenesis is not easily available and accessible to researchers, impeding successful hypothesis generation. In this study, we computationally predict HPI between

mouse and *C. difficile* to rapidly generate interactions to advance knowledge in colonization and infection mechanisms. We predict interactions in mice due to their genetic and physiological similarity to humans[193]. Mice models have been instrumental in elucidating key aspects of CDI, such as *C. difficile* toxins production, host immune response, and the role of microbiota in the host's susceptibility to infection. Additionally, there are well-established protocols for inducing CDI in mice, making them a standardized and widely adopted model in the field[114].

Drug discovery experiments are time and cost-intensive. Computational alternatives to rapidly predict the primary targets are useful. Since mice are commonly used clinical *in vivo* models, predicting interaction in mice can inform pre-clinical studies in *C. difficile* research. This study uses the interologs to infer protein-protein interactions between mouse and *C. difficile*[176]. Interolog prediction uses sequence homology to make predictions from experimental interaction data. Therefore, the interaction of a pair of proteins in one species can be predicted in its related species. Interolog-predicted interactions are rapid and use relatively less computing power, relying on preexisting experimental data to transfer information to sparsely studied systems. However, the accuracy of these predictions is dependent on the quality and comprehensiveness of the initial experimental data. This accuracy also depends on the quality of orthologs identified in this approach[194]. Despite these challenges, the interolog approach allows for generating hypotheses that can be experimentally validated to guide targeted and efficient research[175].

We identified 1,281 interactions between *C. difficile* strain 630 and mouse using the interolog approach. A visual summarization of the steps used in our method is seen in Figure 2. We examined the subcellular localization of mouse and *C. difficile* proteins to identify and select interactions involving only those proteins that are either surface-exposed or secreted. This resulted in 66 interologs. Next, we manually assessed these interologs for their role in *C. difficile* infection

and identified 37 interactions as the most likely candidates for further investigation. While this approach is rapid and combines biological knowledge of infection with the interolog method, it also demonstrates the necessity of available and accessible curated information to identify HPI relevant to *C. difficile* infection.

2.2 Methods

2.2.1 Identifying Host-Pathogen Interactions from PSICQUIC

We used Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) (Version 1.6.0; August 2022) [133] to identify the initial set of manually curated Mammalian-Bacillota interactions. PSICQUIC uses Molecular Interaction Query Language (MIQL) to query for searches. We build the query using the MIQL field 'taxidA,' assigned to species interactor A, and 'taxidB,' assigned to species interactor B. Thus, we identify host interactions using the Mammalian taxon ID: 40674 and identify pathogen interactions using the Bacillota taxon ID: 1239. We combine the results of the two related searches as 1) "taxidA:40674 AND taxidB:1239" and 2) "taxidA: 1239 AND taxidB:40674". We use the option 'Cluster this query' for each search to remove redundant results. Since curators use taxidA and taxidB interchangeably, we manually scan host and pathogen protein lists to remove duplicate interactions.

2.2.2 Finding Orthologs to Host Proteins

Manual inspection of the host protein list identifies interactions that are predominantly human proteins. Using Ensembl (version 108) Biomart[195], we identify 1:1 human: mouse orthologs. We select the option for 'human genes' and select 'compared to Orthologous Mouse Genes', selecting the Orthology type. We download the results as a tab-separated file and use strict 1:1 mouse orthologs in subsequent steps.

2.2.3 Finding Homologs to Pathogen Proteins

We identified a mixture of bacterial species from a manual inspection of the pathogen protein list. Since ortholog predictions are unavailable for all these species, we used reciprocal BLAST matches to identify homologous pathogen proteins. Using UniProtKB, we created a FASTA sequence file of proteins in the pathogen list identified from PSICQUIC. We also downloaded the FASTA sequence of the *C. difficile* 630 proteome from UniProt Proteome (2022-03 release). These form the two datasets for reciprocal BLAST matching. We then used NCBI- BLAST (v 2.12.0) to index the two FASTA files (makeblastdb) and create the databases[196]. We searched the pathogen protein FASTA against the *C. difficile* 630 protein database and the *C. difficile* 630 protein FASTA against the pathogen protein database. Only matches with Evalue <0.001 are used in subsequent steps. To find the best reciprocal matches, we use a script to look up and compare results from both BLAST searches (Qin 2017).

2.2.4 Matching Host and Pathogen Homologs to Create Interologs

Using the set of manually curated interactions, we matched host proteins to identify mouse orthologs and matched pathogen proteins to identify *C. difficile* 630 reciprocal best matches using Excel's 'VLOOKUP' function. This creates a list of mouse– *C. difficile* 630 interologs.

2.2.5 Identifying *C. difficile* 630 and Mouse Proteins with Potential for Host-Pathogen Interactions

While interologs provide a broad set of potential interactions, this method does not consider the biological context. For example, host and pathogens need to co-localize to interact. We use a combination of Gene Ontology (GO IDs) and SwissProt Keywords (SPKW) to identify *C. difficile* proteins found on the bacterial membrane surface, secreted or surface spore proteins (Table 4).

Likewise, we identify surface or secreted proteins in mice (Table 5). Next, we manually add individual proteins known to meet these criteria, but which have not been annotated (GroEL, Q18CT5; CD630_32460, Q17ZZ0; SlpA, Q183M8) to complete the *C. difficile* 630 list. We match these protein lists to the interolog set using Excel's XLOOKUP function to identify interactions.

Table 4. Annotation Terms Used to Identify *C. difficile* Strain 630 Proteins Likely to be Involved in HPI.

Source of Annotation	Cellular Compartment	Term ID
Gene Ontology terms	Plasma membrane	GO:0005886
	Flagella	GO:0009288
SwissProt Key words	Lipoprotein	KW-0449
	Membrane	KW-0472
	Secreted	KW-0964

Table 5. Annotation Terms Used to Identify Mouse Proteins Likely to be Involved in HPI.

Source of Annotation	Cellular Compartment	Term ID
Gene Ontology terms	extracellular organelle	GO:0005615
	external encapsulating structure	GO:0030312
	plasma membrane	GO:0005886

2.2.6 Manual Evaluation of Predicted Host-Pathogen Interactions

We qualitatively assessed the predicted HPI in the context of *C. difficile* infection. First, we selected only mouse proteins expressed in colon epithelial cells using NCBI Gene Expression Omnibus data to observe protein expression in the colon[198]. Subsequently, we evaluated the ability of mouse proteins involved in the predicted HPI to evoke established responses to *C. difficile*

infection (refer to Table 6). We manually review *C. difficile* membrane proteins to select proteins expressed on the cell membrane and remove transmembrane transporter proteins and multidrug-resistant proteins. Next, we reviewed literature to evaluate if *C. difficile* proteins have known host cell binding partners that have not yet been annotated.

Table 6. GO Biological Processes Used to Assess Mouse Proteins Involved in HPI. Mouse proteins in predicted HPI are assessed for their role in *C. difficile* colonization and infection.

GO Biological Process	**GO ID**
Innate immune response	GO:0045087
Adaptive immune response	GO:0002250
Humoral immune response	GO:0006959 GO:0061844
Inflammatory response	GO:0006954
Cytokine mediated response	GO:0009617 GO:0070098 GO:0019955
Cell death	GO:0006915

2.2.7 Network Analysis of Reviewed Host-Pathogen Interactions

To gain biological insight into the predicted HPI, we used network construction to analyze interactions. The predicted HPI are pairwise, binary interactions showing host-pathogen interactions. Mouse protein-protein interactions (PPI) and *C. difficile* PPI with an Interaction Score > 0.7 were retrieved from the STRING Database[147] and visualized in Cytoscape[199]. The downloaded interactions were merged with the existing HPI set and visualized as a single network. Using the 'Network Analyzer' option, we analyzed the network properties. Cluster analysis using

the EAGLE algorithm[200] was implemented with standard settings to identify sub-networks with biological significance.

2.3 Results

2.3.1 Identifying Interologs from Annotated Host-Pathogen Interactions

We identified 3,214 Mammalian-Bacillota interactions from the PSICQUIC interaction resource. These interactions represent host-pathogen protein-protein interactions only. Manual inspection of the host proteins in interaction results identified that 99% of host proteins are human proteins (Figure 1A). Likewise, 96% of pathogen proteins are from *Bacillus anthracis*, with the remaining 4% from Bacillota families *Clostridiaceae*, *Streptococcaceae*, and *Staphylococcus* (Figure 1B). This set of HPI forms the experimental dataset to infer interologs between mouse – *C. difficile* 630 and contains 1,460 host proteins and 1,013 bacterial proteins. Comparing human: mouse orthologs based on the experimental interactions identifies 1,460 mouse proteins (45%) (Supplementary Table 1). Identifying homologs from the experimental bacterial interactions resulted in 483 *C. difficile* proteins (47%) (Supplementary Table 2). The mouse and *C. difficile* 630 proteins are matched to the PSICQUIC interaction set to generate 1,281 interologs (Supplementary Table 3).

2.3.2 Identifying Interologs with Potential Host-Pathogen Interactions Using Known Biological Characteristics

The initial interolog set represents HPI predicted based on sequence similarity to experimental HPI. We further evaluate these interactions in the context of bacterial pathogenesis and infection. For example, *C. difficile* proteins found on the membrane surface, secreted or surface spore proteins are likely to interact with host proteins. We use Gene Ontology (GO) and SwissProt Keywords (SPKW) to identify these *C. difficile* proteins. In addition, since the annotation of *C. difficile* proteins is incomplete, we supplement this list with known virulence factors [201]. Using

this approach, we identified 227 biologically relevant HPI (23% of initial interologs), with a majority of these *C. difficile* proteins being annotated as surface and secreted proteins (Figure 2A). Further, host proteins should also be surface proteins or extracellular proteins. Using a similar approach, we identified the localization of host proteins in these 227 biologically relevant HPI (Figure 2B), resulting in 100 HPI that met these criteria. Our manual review to identify mouse proteins expressed in the colon and *C. difficile* proteins showing potential to interact with the host resulted in 35 HPI. We use these HPI for our subsequent steps in network analysis. e cannot perform statistical tests in this approach due to the lack of ground truth for comparing predicted HPI. While curated HPI data between toxin B and human receptors is available, it does not capture all possible interactions, making it insufficient for developing robust statistical tests[133].

2.3.3 Host-Pathogen Interaction Network Analysis and Submodule Analysis

Network construction and analysis are useful in identifying interactions with biological relevance. We query the STRING database to retrieve mouse protein-protein interactions (PPI) and *C. difficile* PPI with interaction scores > 0.7. Scores closer to 1 have the highest possible confidence, whereas a score of 0.5 indicates that every second interaction may be erroneous. After merging with the predicted HPI, the network contains 578 nodes and 1946 edges (Figure 3). Supplemental Table 4 provides each interaction in the network. Analysis of the network properties identifies proteins with betweenness and degree centrality. Degree centrality highlights proteins with many direct interactions, while betweenness centrality identifies proteins critical for maintaining connectivity. The EAGLE algorithm identifies densely connected regions within the network[200]; the resulting sub-module analysis identifies 27 sub-modules with edge clustering. Of these 27 sub-modules, 13 sub-modules show HPI with edge clustering (Table 7).

Table 7. Identifying Sub-Modules with High Clustering in the HPI Network. Sub-module analysis of the generated network reveals clusters with high connectivity. Thirteen sub-modules in this clustering contain predicted HPI between *C. difficile* 630 and mouse.

***C. difficile* Protein**	**Mouse Protein**	**Nodes**	**Edges**
CD630_18010	Anxa7	58	227
CD630_29630	Rap1b	28	71
CD630_22500	Hfe	30	119
CD630_14710	Cd74	27	58
CD630_14710	Mycbp2	27	58
CD630_16720	Cfd	38	78
CD630_03380	Plscr4	26	44
CD630_05540	Limd1	22	61
CD630_03860	Col12a1	21	81
CD630_27180	Ninj1	24	31
CD630_28370	Il4r	24	98
CD630_27690	Svep1	22	54
CD630_30420	Cask	22	70

2.4 Discussion

2.4.1 HPI Identified Using the Interolog Approach

Leveraging computational techniques to rapidly predict HPIs helps identify potential drug targets and advance our understanding of molecular mechanisms underpinning infection. This study predicts interactions between mice and *C. difficile* 630. We select *C. difficile* strain 630 to predict interactions as it is a well-characterized reference strain extensively used in research to study CDI's genetic and pathogenic mechanisms, providing critical insights into its virulence factors and resistance mechanisms[202]. Its comprehensive genome sequence serves as a benchmark for comparative studies between strains. We predict interactions in mouse as it is a commonly used comparative model to study CDI in humans[193]. By applying the interolog[176] approach to Mammalian-Bacillota interactions from PSICQUIC[133], we use sequence homology to identify 1,281 potential HPI between mouse and *C. difficile* 630. Since the only *C. difficile* proteins

available for host interactions will be on the bacterial cell surface, secreted or the spore surface, we used post-processing steps to identify predicted HPI containing these *C. difficile* proteins. Likewise, the mouse proteins most likely to encounter *C. difficile* are cell surface, secreted or extracellular membrane proteins and must be expressed in the colon, the site of *C. difficile* infection[31]. Based on this information, we found 66 HPI, and from a manual assessment, we identified 37 HPI. Using sub-module analysis of the host-pathogen interaction network, we selected 13 of the predicted HPI to examine in detail.

1) Interaction between CD630_18010 and Anxa7
   CD630_18010 is a putative membrane protein belonging to the YiT family with currently unknown functions. UniProtKB annotates this membrane-anchored protein with uncharacterized domain function[203]. Anxa7 is a host calcium/phospholipid-binding protein that promotes membrane fusion and is involved in exocytosis. Anxa7 is a calcium-dependent transporter protein crucial for ion transport across the host membrane. Proper regulation of Anxa7 is essential for maintaining cell shape and fluid secretion. Dysfunction in this protein can contribute to a loss of cell stability and disrupt fluid secretion[204]. I need further information about the *C. difficile* protein to say further about this predicted HPI.

2) Interaction between CD630_29630 and Rap1b
   CD630_29630 is an L, D-TPase catalytic domain-containing protein. This protein is predicted to be involved in cell wall biology (InterPro prediction) in forming peptidoglycan cross-links, crucial for bacterial cell wall integrity. Mouse protein Rap1b regulates lymphocyte adhesion and migration, with its active form (Rap1-GTP) being crucial for integrin activation and T cell recirculation[205]. Rapb1 deficiency leads to lymphopenia and the generation of pathogenic effector/memory T cells, which are home to the colon and

exacerbate colitis. Colitis disrupts normal gut absorption and secretion, leading to symptoms such as diarrhea. The infiltration of pathogenic T cells compromises the mucosal barrier, further promoting fluid secretion and inflammation[205]. Without further information about the *C. difficile* protein, I cannot say further about this predicted HPI.

3) Interaction between CD630_22500 and Hfe

CD630_22500 is part of the Autoinducer 2E (AI-2E) family, a group of putative transporters within the AI-2 exporter superfamily, with no other functionally characterized members. These proteins, derived exclusively from bacteria, are involved in AI-2-mediated quorum sensing, which promotes biofilm formation in many pathogenic bacteria. AI-2 proteins also regulate virulence factor production and bacterial motility.[206],[207]

The predicted mouse-interacting protein, Hfe, is a member of the MHC class I family and binds to the transferrin receptor, reducing its affinity for iron-loaded transferrin[208]. Located on the apical part of the colon epithelial cell, Hfe plays a critical role in controlling iron homeostasis[208]. Without further information about the *C. difficile* protein, I cannot say further about this predicted HPI.

4) Interaction between CD630_14710 and Mycbp2

CD630_14710 belongs to the Pip/YhgE protein family (InterPro) and is involved in xenobiotic compound transfer. However, as CD630_14710 is poorly characterized, its role in pathogenesis cannot be ascertained. Mouse Mybcp2 is a Myc-binding protein with multifunctional binding with involvement in critical cellular processes in the cytoskeleton which is typical processes that is affected by *C. difficile* toxins. Without further information about the *C. difficile* protein, I cannot say further about this predicted HPI.

5) Interaction between CD630_14710 and Cd74

As stated above, CD630_14710 belongs to the Pip/YhgE protein family involved in xenobiotic compound transfer. Mouse Cd74 is a key player in the host immune response to bacterial infections, primarily through its roles in antigen presentation and immune cell signaling. Its interactions help coordinate the activation and proliferation of immune cells, regulate inflammatory responses, and ensure effective pathogen clearance[209]. Without further information about the *C. difficile* protein, I cannot say further about this predicted HPI.

6) Interaction between CD630_16720 and Cfd

CD630_16720 is a sensor histidine kinase. In bacteria, sensor histidine kinases are membrane-associated proteins that are part of a two-component signal transduction system. They detect environmental signals and transduce these signals into a cellular response through a series of phosphorylation events. Sensor histidine kinases can respond to host immune factors[210]. This is a HAMP domain-containing protein. The predicted mouse interacting partner Cfd is part of the alternative complement pathway, which is one of the three main complement pathways. Cfd is continuously active at a low level and can be rapidly amplified in response to pathogens[211],[212]. Cfd is part of the cascade reaction and is an intermediate protein that activate the alternative complement pathway and may not participate directly in HPI. Without further information about the *C. difficile* protein, I cannot say further about this predicted HPI.

7) Interaction between CD630_03380 and Plscr4

CD630_03380 is also a sensor histidine kinase of the GHK family predicted through InterPro. Mouse protein predicted to interact with CD630_16720 is Plscr4, a phospholipid scramblase protein that regulates the release of cytokines and pro-inflammatory

molecules[213]. In a murine model infected with bacteria, the expression of Plscr4 was found to be reduced[213]. Without further information about the *C. difficile* protein, I cannot say further about this predicted HPI.

8) Interaction between CD630_05540 and Limd1

C. difficile protein CD630_05540 belongs to the signal peptidase family of proteins that cleaves host or bacterial proteins[77],[214]. Mouse Limd1 is a scaffolding protein predicted to be present in adherens junctions and focal adhesion. Toxin-induced disruption of tight junctions can affect LIMD1's role in maintaining the epithelial barrier. However, there is no evidence of Limd1 and CD630_05540 physically interacting.

9) Interaction between CD630_03860 and Col12a1

Col12a1 is a collagen XIIA protein that allows cell adhesion between cells or to a membrane. Col12a1 is expressed in the colon and is predicted to maintain cell junction adhesion in epithelial cells[215]. The role of Col12a1 in host colon epithelial cells is not well characterized. However, as a putative component of cell junction adhesion, *C. difficile* may target the cell junction to access the basal layer of epithelial cells. A study by Fairweather et al., 2017 characterizes CD630_03860 as a sortase that may be secreted to promote *C. difficile* pathogenesis.[83]

10) Interaction between CD630_27180 and Ninj1

Nerve injury-induced protein (Ninjurin [Ninj]) 1 is an adhesion molecule and is induced under inflammatory conditions[216]. Ninj1 is elevated under inflammatory conditions and contributes to inflammation not only by mediating leukocyte migration but also by modulating Toll-like receptor 4-dependent expression of inflammatory mediators[217]. CD630_27180 is a sortase that can cleave bacterial surface proteins and could potentially affect host proteins[82]. Sortases are membrane-bound cysteine transpeptidases that anchor surface proteins to the peptidoglycan cell wall in Gram-positive bacteria. The sorting

process is mediated by a conserved C-terminal cell wall sorting signal on the anchored protein, comprised of a C-terminal recognition sequence (often LPXTG, where X is any amino acid), followed closely by a hydrophobic transmembrane domain and a positively charged tail[82]. Sortases catalyze the covalent attachment of specific surface proteins to the peptidoglycan layer of the bacterial cell wall. These surface proteins often include adhesins, and virulence factors crucial for interacting with host cells. Sortases display various virulence factors on the bacterial surface, such as toxins, enzymes, and factors that modulate host cell signaling. These virulence factors can directly damage host tissues or alter host cell functions to benefit the pathogen[81]. While sortases are surface proteins, their role in HPI is indirect and contributes to colonization and infection while not directly participating in host binding.

11) Interaction between CD630_28370 and Il4r1

C. difficile protein, CD630_28370, belongs to the YkvI family of proteins. It is a membrane protein. Mouse Il4r1 binds to interleukin-4 (Il4) and interleukin-13 (Il13), which are key cytokines in immune regulation[218]. During bacterial infections, Il4r1 signaling typically skews the immune response towards a Th2-type profile, characterized by the production of cytokines such as IL-4, IL-5, and IL-13. This Th2 response is often associated with increased antibody production, eosinophil activation, and macrophage polarization, which can be beneficial in controlling extracellular pathogens but may also lead to chronic inflammation or tissue damage in certain bacterial infections[218].

12) Interaction between CD630_27690and Svep1

CD630_27690 is a biosynthesis protein involved in building the peptidoglycan layer (Uniprot annotation). Mouse Svep1 (Sushi, von Willebrand factor type A, EGF, and

Pentraxin domain-containing protein 1) is poorly characterized in colon epithelial cells. Svep1 is an essential protein in maintaining cell-matrix adhesion in mice[219].

13) Interaction between CD630_30420 and Cask

CD630_30420 protein is a putative aromatic acid exporter with a C-terminal domain-containing protein (from InterPro Annotation). Its role in pathogenesis cannot be determined. Mouse Calcium/calmodulin-dependent serine protein kinase (Cask) is a type of serine/threonine kinase that is regulated by calcium and calmodulin. Cask is a scaffolding protein involved in cascade processes[220].

2.4.2 Limitations of the Interolog Approach

Predicting interologs depends on finding homologous proteins between experimentally confirmed, curated interactions and proteins in the system of interest. Host homology is inferred from orthologous human: mouse proteins. There is a high degree of orthology between human and mice: 80% of human and 72% of mouse protein-coding genes have a one-to-one orthologous relationship[221],[222]. However, only 45% of the human proteins curated on PSICQUIC have 1:1 ortholog in mouse[195]. We lose one in two interactions while transferring HPI to mouse and *C. difficile* 630 in this process. For example, absences in cell surface protein TLRs between human and mouse highlights distinct immune systems between mouse and humans[223]. Transferring interactions is limited to the protein set from PSICQUIC as demonstrated by the transfer of homologous proteins between the initial Mammalian – Bacillota dataset to mouse- *C. difficile* 630. To perform statistical testing, we use ground truth observations as a comparative measure. However, this method does not detect any known interactions between *C. difficile* and mouse to perform statistical testing. Statistical testing compares observed data to expected outcomes (the null hypothesis). Without a ground truth, there is no reference point to define what is expected. This makes it difficult to determine if the observed data deviates from the expected distribution

meaningfully. Additionally, performing statistical tests without a ground truth can lead to misleading interpretations. The results of such tests might show significance or patterns that do not actually reflect experimental evidence.

2.4.3 Phylogenetic Divergence Between *C. difficile* and *B. anthracis* Impacts Quality of Interologs

B. anthracis makes up 96% of the host-pathogen protein-protein interactions found in PSICQUIC. *B. anthracis* is a spore-forming bacteria that causes anthrax in Mammalian hosts[224]. Anthrax is a well-studied respiratory disease, but relevant to this study is GI anthrax, transmitted via ingestion of infected meat[159]. GI anthrax causes severe infectious colitis and diarrhea with symptoms similar to *Clostridioides difficile* infection[170]. Both GI anthrax and *C. difficile* infection involve targeting epithelial host cells mechanistically. The intoxication of host epithelial cells with anthrax lethal toxin (LT) or *Clostridioides difficile* TcdA/B contributes to the breakdown of the gut lining by destabilizing intestinal epithelial cells[162]. Further, LT in GI anthrax and Toxins A and B in CDI elicit proinflammatory response via activation of MAPK proteins, thereby contributing to an infection-induced production of interleukins[225]. Because of these mechanisms, these are areas where the proteins may be functionally similar between *C. difficile* and *B. anthracis*. However, regarding sequence homology, LT and TcdA/B are not homologous proteins. HPI involving these toxins is not transferrable between these systems. The interolog approach cannot be used to transfer interactions between unique proteins, that is, proteins with functional and structural differences. Since there are no toxin homologs between the two bacterial species, we will not identify the HPI of toxins using the interolog approach

Key to predicting HPIs are the cell surface and secreted proteins, which typically constitute 20-30% of the cell's proteome[226]. When we identified homologous matches between *B. anthracis* and *C. difficile*, we found that 24% are cell surface and secreted proteins. Despite the phylogenetic divergence between these bacteria, we do not lose identification of potential HPIs based on the homology of surface and secreted proteins[227]. This homology is demonstrated in flagellin, a filament protein of bacterial flagella that plays a role in bacterial adhesion to the host cell[228]. This protein is homologous between *C. difficile* 630 and *B. anthracis*. It is characterized as interacting with host TLR5. However, the interaction data in PSICQUIC cannot be transferred to our study as interaction with TLR5 is not found in the PSICQUIC experiment HPI data. Highlighting differences between strain-wise homology, the pore-forming anthrax toxin, protective antigen (PA) in *B. anthracis* is a well-known homolog to *C. difficile* binary toxin, CdtB[31], however, *C. difficile* 630 does not encode the binary toxins CdtA and CdtB and therefore, interactions with PA, curated on PSICQUIC, is not transferred to our system of study between mouse and *C. difficile* 630. Some proteins like alpha enolase, DnaK, and GroEL are homologous between *C. difficile* and *B. anthracis*, and we would expect them to be involved in HPI in both bacteria. Still, there is no literature source confirming these interactions in *C. difficile*. The interolog approach can be used to predict interactions rapidly. Still, the quality of predicted interologs depends on the experimental dataset from which we infer interactions, as highlighted in the above examples.

To demonstrate the ability of methods, we predict interologs between humans and *C. difficile* 630. We find 1,489 interologs, a 14% increase in interologs from mouse and *C. difficile* 630. The proteins in the new interologs found between humans and *C. difficile* 630 are nuclear proteins involved in cellular functioning, which are not likely to interact as these proteins are not

extracellular or surface proteins. This brief analysis highlights that the refinement steps in our method result in only a subset of interologs from the experimental interactions.

By extending the ability of interologs with the addition of post-processing steps, this methodology can be adapted to any host-pathogen pair. Focusing on surface and secreted proteins, we identify interactions that are more likely to play a direct role in *C. difficile* adherence and invasion in host cells. The identified interactions provide a valuable starting point for experimental validation and could inform the development of novel therapeutic and prophylactic strategies against CDI. However, the prediction of host-pathogen interactions using interologs is constrained by the limited curated HPI data currently available, making this approach less suitable for predicting interactions between *C. difficile* and mice and highlighting the necessity of curation data that is closely related to Clostridia species. Therefore, alternative approaches for predicting HPI for *C. difficile* should be investigated. For example, advances in machine learning and its application to predicting host-pathogen interactions or molecular interactions using ensemble machine learning can be beneficial. Ensemble machine learning[229] uses several protein features to make predictions that can provide comprehensive HPI predictions.

Chapter 3 Conclusion

3.1 Conclusions from Research Results

Utilizing computational techniques to predict host-pathogen interactions (HPIs) rapidly generates potential drug targets and enhances our understanding of the molecular mechanisms underlying infections. This study predicts interactions between mouse and *Clostridioides difficile* strain 630, a well-characterized reference strain extensively used in research[20] for studying CDI genetic and pathogenic mechanisms. Using the interolog approach[176], the study initially identified 1,281 potential HPIs between mouse and *C. difficile* 630. Pruning based on identifying proteins that would physically interact identified 66 predicted HPIs, and subsequent sub-module analysis identified 13 of these putative interactions. This study highlights the relevance of cell surface and secreted proteins, crucial for bacterial interaction with host cells. Sub-module analysis finds interactions where proteins essential to the network are highlighted. One predicted HPI is between CD630_03860 and Col12a1. The CD630_03860 protein is a sortase-anchored protein involved in bacterial adherence to host cell protein[83]. Sortase proteins (transpeptidases) are bacterial enzymes that anchor proteins to the cell walls of Gram-positive bacteria[81]. They do this by cleaving the sorting signals of secreted proteins to form isopeptide bonds between the proteins and peptidoglycan or polypeptides. The CD630_03860 protein is predicted to bind mouse Col12a1, an extracellular matrix protein that contributes to cell adhesion by interacting with other extracellular membrane proteins. *C. difficile* infection mechanisms target cell adhesion function; thus, this interaction may provide an approach for targeting *C. difficile* colonization. Another potential HPI of interest is between CD630_27180 and mouse Ninj1. CD630_27180 is a predicted sortase that can cleave bacterial surface proteins and potentially affect host proteins[82]. Ninj1 is a plasma membrane protein[230] and plays a role in bacterial-induced expression of inflammatory mediators[217]. Notably, HPIs identified using the interolog approach identify understudied *C.*

difficile proteins. The 13 predicted HPI are not reported in the literature. This is because most *C. difficile* protein information is uncharacterized. UniProtKB reports 3,762 proteins in the *C. difficile* 630 proteome[20]. However, of these 3,762 proteins, only 303 (8%) of *C. difficile* proteins are manually curated and considered reliable, which is the gold standard for protein data. Therefore, it is challenging to say more about HPI without further functional or expression data about these poorly characterized *C. difficile* proteins.

The interolog approach has limitations, primarily the dependency on finding homologous proteins and the limited, curated HPI data. Only 45% of human proteins curated on PSICQUIC[133] have direct orthologs in mice, resulting in a loss of potential interactions as fewer proteins are transferred to make interactions. The phylogenetic divergence between *B. anthracis* and *C. difficile* further complicates the prediction of HPIs[227]. For example, the interolog approach cannot predict interactions with *C. difficile* 630 toxins as they do not have homologous proteins in *B. anthracis*. The *B. anthracis* accessory toxin Protective Antigen (PA) has sequence and functional similarity to the *C. difficile* binary toxin CdtB[58]. Still, this binary toxin is not present in the *C. difficile* 630 proteome and cannot be used to predict CdtB interactions in this strain of *C. difficile*.

Our interolog approach uses a pruning step to identify host and pathogen proteins likely to interact physically. Since *C. difficile* is an extracellular pathogen, the only mouse proteins most likely to encounter *C. difficile* are cell surface, secreted, or extracellular membrane proteins. In addition, these mouse proteins must be expressed in the colon, the site of *C. difficile* infection.

Likewise, the only *C. difficile* proteins that are available for the host to bind will be on the cell surface, secreted, or the spore surface. However, spore protein interactions are challenging to predict as *C. difficile* 630 spore proteins are not annotated in annotation resources such as QuickGO[231] and SwissProt[232] , which we use for the pruning step. We note that the overwhelming

lack of homologs between *C. difficile* and *B. anthracis* and lack of protein functional annotation in *C. difficle* 630 hinders this prediction study. Despite these challenges, this study shows that focusing on surface and secreted proteins identifies interactions with potential for HPI as demonstrated between CD630_27180 and Ninj1 and CD630_3860 and Col12a1. The current limitations underscore the need for more comprehensive curated HPI data and suggests that alternative methods, such as ensemble machine learning, could improve the accuracy and scope of HPI predictions.

3.2 Future Directions for Predicting *C. difficile* HPI

'Machine learning' (ML) refers broadly to the process of fitting predictive models to data or of identifying informative groupings within data. Machine learning is particularly useful when the datasets have many features to process[233]. A feature represents a specific characteristic or measurement in an experiment. For instance, in a gene expression over time experiment aimed at studying the rate of conversion of metabolites, input features could include 'gene expression level' and 'time of conversion.' In contrast, an output feature would be 'conversion rate.' Data from protein-protein interaction experiments frequently contain properties regarding expression level, amino acid sequence, protein domains, and motifs, which can be used as input features. Output features can include interaction pairs, interaction strength, and interaction networks. The classical ML model is trained on a data set of engineered features in a user-supervised or unsupervised manner. The type of data used as input is an important aspect in ML models.

Training data is of two types: labeled and unlabeled data. Labeled data is verified true data. As an example, labeled data in HPI prediction would be HPI confirmed using co-immunoprecipitation or affinity chromatography. Unlabeled data refers to unsorted or unorganized data. Unlabeled data is synonymous with heterogenous data. As an example, unlabeled data would be HPI information

available from a study but not curated for analysis. By training models on extensive datasets of known HPIs, including genomic, proteomic, and other biological features, these models can learn patterns and associations that facilitate the prediction of new, previously uncharacterized interactions. This approach utilizes machine learning techniques, such as supervised learning, where the model is trained on labeled data in which , or unsupervised learning, which can uncover hidden patterns in unlabeled data[234]. ML algorithms which have been applied to predicting HPIs, include are Random Forest (RF)[235], multilayer perceptron (MLP)[236] and kernel-based supervised machine learning[237]. Models are tested for accuracy and maximal area under the curve as measures of success. ML models have been implemented in finding HPI for virus-host interactions[238_240].

Kshirsagar *et al*.[241] proposed a multitask learning-based method for predicting human−bacteria HPIs where the host is fixed, and the pathogens are a mixture of bacterial species. This study is based on the biological hypothesis that proteins from different pathogens essentially target the same critical biological processes in human cells[241]. This study uses Gene Ontology (GO) to predict interactions. However, when put into practice, this method has limitations as the Gene Ontology structure is very complex for feature engineering. Gene Ontology comprises a large vocabulary and a hierarchical structure that is difficult to engineer using ML. Another major challenge in using machine learning to predict HPIs is addressing its biggest pitfall: the lack of non-interaction data for both training and testing[242]. ML approaches in predicting *C. difficile* HPI may not be successful due to the lack of non-interacting data. Additionally, as these approaches are dependent on available experimental HPI data between host and bacteria, the accuracy of predicted HPI will be determined by the quality and quantity of HPI data which can be used to train and run the ML algorithms.

Beyond classical ML, a subset of approaches named Deep Learning (DL) uses architectures like convolutional neural networks (CNN) to provide a methodology[243]. Deep learning (DL) architectures designed for predicting protein-protein interactions (PPIs) can be thought of as advanced tools for extracting meaningful features from complex data. These models function primarily as end-to-end binary classifiers, which means they are built to perform a final task of classifying whether a given protein-protein pair interacts or not. Initially, these DL models extract crucial features from the input data, which might include various details about the proteins such as their sequences, structures, and other biological attributes[243]. After this feature extraction process, the models undergo supervised learning where they are trained on a dataset with known interaction outcomes, learning to differentiate between interacting and non-interacting protein pairs. Once trained, these models can be applied to predict interaction[244].DL methods have outperformed traditional ML methods in predicting human-virus HPIs[244]. In contrast to ML models, deep learning architectures are flexible in the known labeled data, meaning the DL architecture automatically learns complex patterns and representations from the input data without manually adding features. This approach is ideal for predicting HPI in species for which input HPI data is sourced from different experiments. Since experimentally verified HPI data is generated from different experiments, DL approaches can adapt to this heterogenous data. Another advantage to DL is flexible architectures, which are a main component of DL approaches as the architecture can learn from the different data types with capability to learn complex representations automatically. Some commonly used DL architectures such as convolutional neural network (CNN), recurrent neural network (RNN), long-short term memory (LSTM) have been used to predict human-virus protein interactions[244].

Kaundal *et al.*[242] use a DL architecture convolutional neural network to predict HPI between humans and bacteria based on human and *Y. pestis* HPI data. This model uses the sequences of proteins to create feature vectors. For each pair of proteins—one from the host and one from the pathogen—it extracts features and combines them into a single vector. This combined vector represents the interaction between the host and pathogen proteins. However, amongst the training dataset, human-bacteria models faced problems in predicting true HPIs due to low sensitivity. This was attributed to the training data belonging to human-*Yersinia pestis*. The lack of true negative data impacted this model. A DL approach to predicting HPI between the host and *C. difficile* can be successful based on the number of layers and features selected to predict interactions. However, DL models require a large training dataset to predict efficiently. Greener *et al.*[233] find that there is a disproportionate amount of sequence data compared to protein-protein interaction data. For instance, public databases like GenBank and UniProt have many biological sequence data available. In contrast, reliable data on protein interactions are much scarcer and harder to find.

A potential approach to predict HPI in *C. difficile* is using ensemble machine learning. Host-pathogen interaction Prediction (HPiP)[229] is built on an ensemble approach using three different classifiers – support vector machine (SVM), reinforcement learning (RL), and logistic regression (LR). A support vector machine uses a graphical method to transform the original input data such that in their transformed versions (called the 'latent representation'), data belonging to separate categories are divided by a clear gap that is made as wide as possible. This method separates data into categories, highlighting its flexibility. However, SVM can be affected by noise or outlier points. Reinforcement learning is a model in which the classifier (called agent) learns optimal strategies over time. This model improves performance through trial and error. The logistic regression model is a binary classifier that uses probability estimation to categorize interactions

based on a threshold. The HPiP model accurately captured 83% of the test HPI against *Mycobacterium tuberculosis* and human protein sequences[229], showing high performance. This prediction model is unique as it uses several machine learning models to allow the user to select a classifier that is suitable.

3.3 Key Challenges in Machine Learning Prediction

One of the main challenges in addressing the current task as a machine learning problem is the limited training data. The number of known HPIs is usually small and thus not representative enough to ensure the generalizability of trained models. In effect, the trained models might overfit the training data and would give inaccurate predictions[245].

Additionally, the quality of the available data is often a concern. Data may contain noise, errors, or inconsistencies, which can further degrade the model's performance. Ensuring high-quality, clean, and well-labeled data is crucial for building reliable machine-learning models.

The complexity of the model also plays a significant role. Highly complex models, while powerful, are more prone to overfitting, especially with limited data. Simpler models may generalize better but might not capture the underlying patterns in the data effectively.

Quality of PPI data: Molecular interaction data quality can pose a challenge in prediction. There is a lack of standardization in the molecular interaction experimental detection and prediction techniques[133]. Yeast two-hybrid analysis is a rapid and high-throughput proteomic approach for detecting molecular interactions, with the ability to screen large libraries. However, this technique is prone to false positives and may not detect interactions requiring post-translational modifications or specific cellular conditions[143]. Another popular molecular interaction detection technique is co-immunoprecipitation, which detects molecular interactions using antibody

binding. This technique provides cellular context for interactions but requires high-quality antibodies to initiate binding.

Additionally, this study may not find transient interactions, which are important in host-pathogen interactions[246]. The challenges of data quality also impede computational techniques. For example, homology-based techniques can be carried out using existing interaction data. However, predictions rely on the quality of experimental interaction data. Machine learning prediction also faces the same challenge of the quality of experimental data. Experimental PPI data in databases lack standardization and normalization for computational methods. A major difficulty of various techniques can hinder data integration, where combining data from different techniques can be problematic due to varying levels of reliability and relevance[133].

Another point to consider is that the biological significance of an interaction varies depending on the detection method and context. Thus, to address the variety in data, the PSICQUIC developed the PSI Confidence Scoring System (PSISCORE). PSISCORE is based on the concept of decentralization, where individual scoring servers can apply different scoring methods for assessing diverse biological and methodological aspects of interaction data[133]. However, our approach cannot implement the PSICSCORE as a factor in predicting HPI as the reciprocal best hits found between *C. difficile* 630 and pathogen proteins come from a proteomic experiment that uses yeast two-hybrid to predict interactions[158].

Difference to other pathogens: In the DeepHPI approach, predictions are based on *Yersinia pestis*–human interactions. For the negative training dataset, this approach utilizes the Neglog dataset as a source of negative examples [247]. The Neglog interaction database contains data where no interaction occurs between proteins[247]. This helps the model differentiate between true positive

interactions and random noise or non-interactions. By providing a robust set of negative examples, the Neglog dataset aids in training more accurate and reliable predictive models.

With bacterial HPI prediction, interactions with hosts can be more complex, involving a variety of mechanisms, including secretion systems, toxins, and surface proteins. These diverse and multifaceted interactions take more work to capture accurately in models.

Limited information on structure and function of virus proteins: While researchers can retrieve information from many publicly available databases for human proteins to extract features related to their function, semantic annotation, domains, structure, pathway association, and intercellular localization, such information with experimental evidence is not readily available for most bacterial proteins[245].

Computational prediction of host-pathogen interactions (HPIs) rapidly identifies potential interactions. The effectiveness of these prediction methods is highly contingent on the quality of the input data and the quantity of data. The quantity of data available for a given problem profoundly impacts the choice of techniques that can effectively be used. As a very rough guideline, when only small amounts of data (hundreds of or a few thousand examples) are available, one is forced to use more traditional machine learning methods, as these are more likely to produce robust predictions. When larger quantities are available, one can start to consider more highly parameterized models, such as deep neural networks[233]. However, the quality of data determines the sensitivity of the learning model. Interaction experiments such as yeast two-hybrid are challenging data to use in training models as yeast two-hybrid generates noisy data (non-interacting data). Noisy data impacts machine learning models' sensitivity and accuracy as it distorts the regression curves to predict interactions. Challenges inherent in methods such as interolog-based approaches, traditional machine learning techniques, and advanced deep learning

models underscore the need for high-quality data encompassing both interacting and non-interacting protein pairs. High-quality data for these methods should align with standardized 'ground truths' or true, verified information.

In the context of viruses, prediction tasks are relatively straightforward due to their intracellular nature, where the entire viral proteome interacts with host cell proteins. In contrast, bacterial HPIs involve specific attachment and infection mechanisms, making accurate protein characterization crucial for reliable interaction predictions. Thus, successful HPI prediction for bacteria demands comprehensive protein functional annotation and subcellular location annotation to ensure high-quality, biologically relevant predictions that can be validated experimentally.

FIGURES

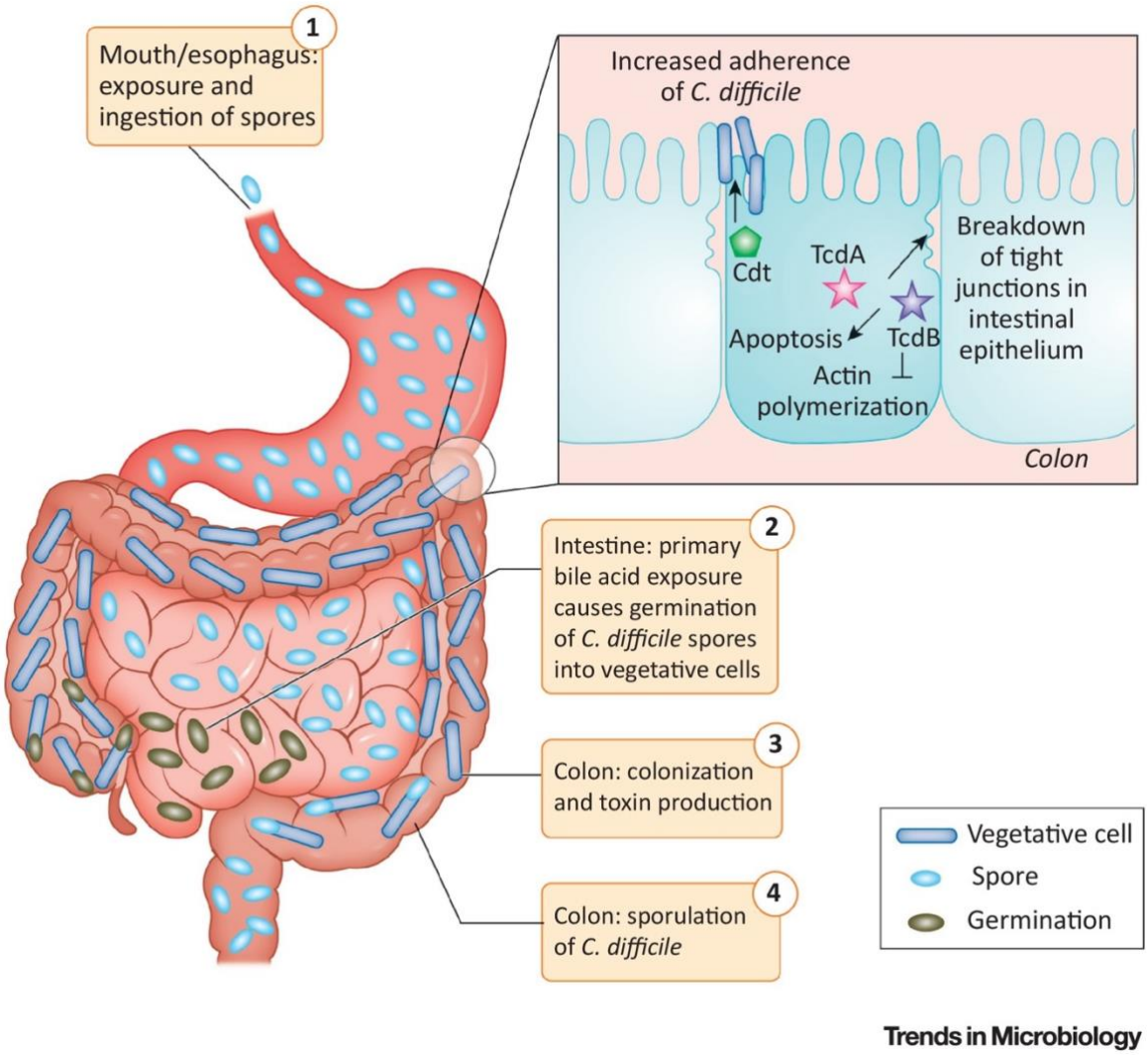


Figure 1. The Pathogenesis of *Clostridioides difficile* Infection. This figure summarizes the infection process by *C. difficile* in the human gastrointestinal tract, highlighting key stages from exposure to colonization and toxin production. The figure is taken from Sandhu, Brindar, and McBride (2018)[2].

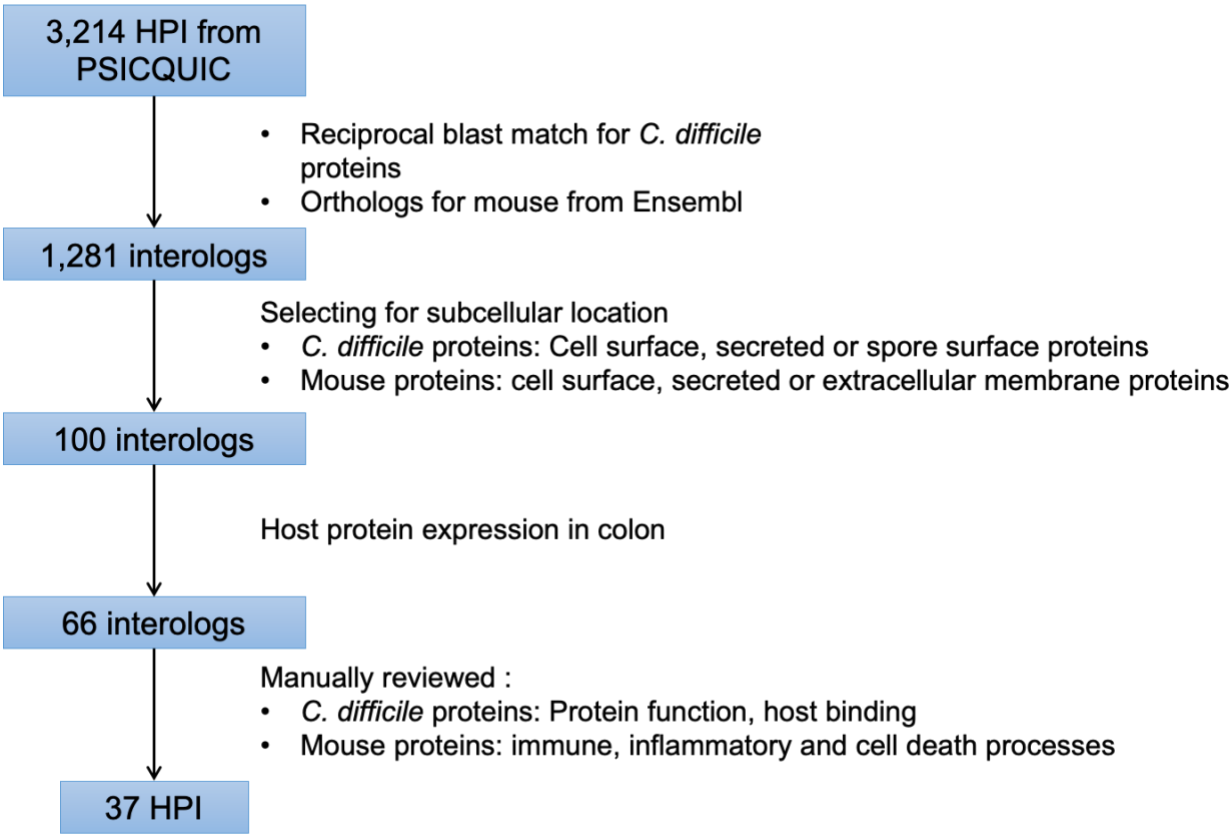


Figure 2. Summary of Sequential Results Predicting Host-Pathogen Interactions between *C. difficile* 630 and Mouse. This figure summarizes the results of each step in predicting HPIs between *C. difficile* 630 and mouse, highlighting key findings at each stage to identify the most relevant interactions for further study.

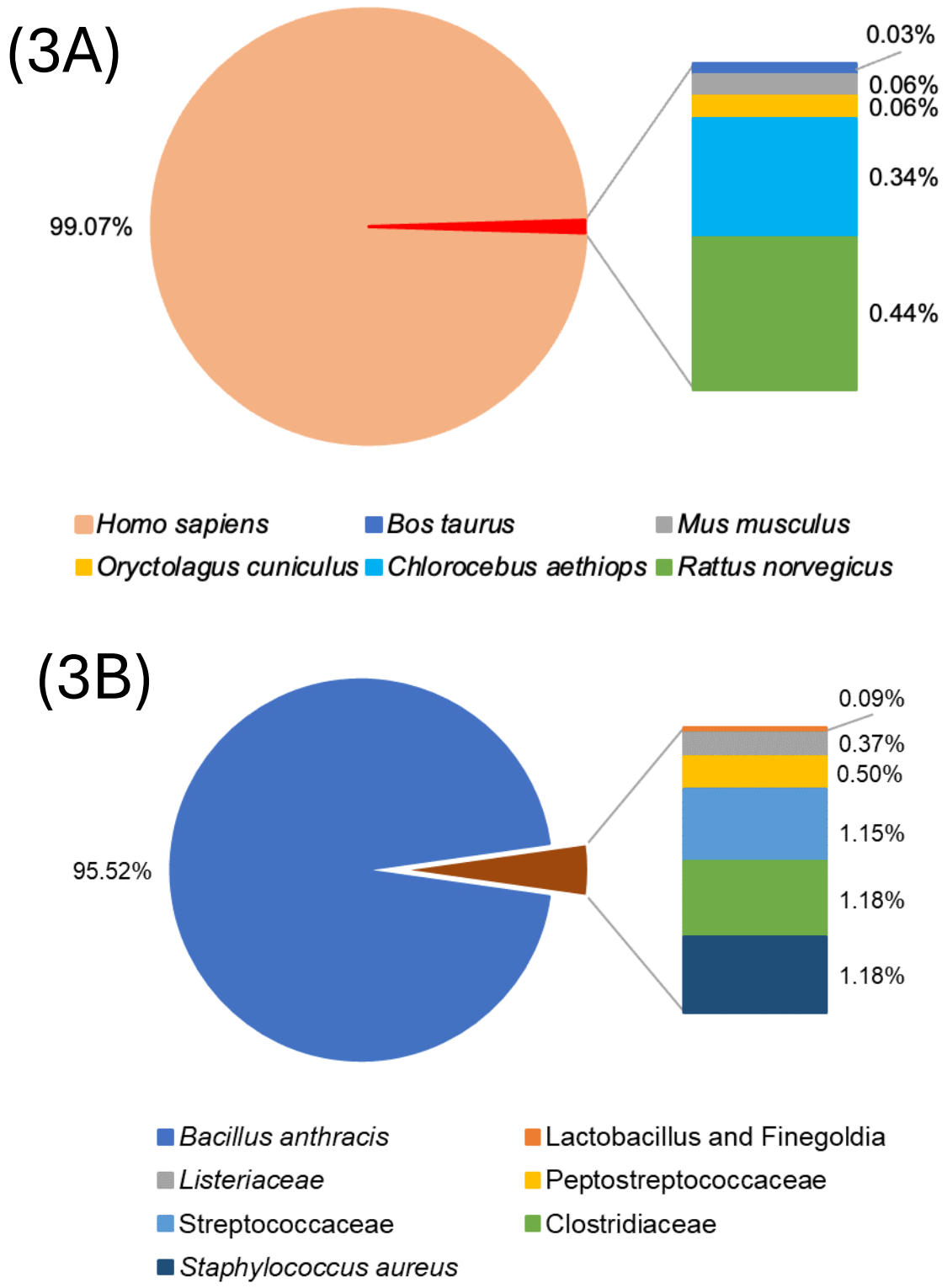
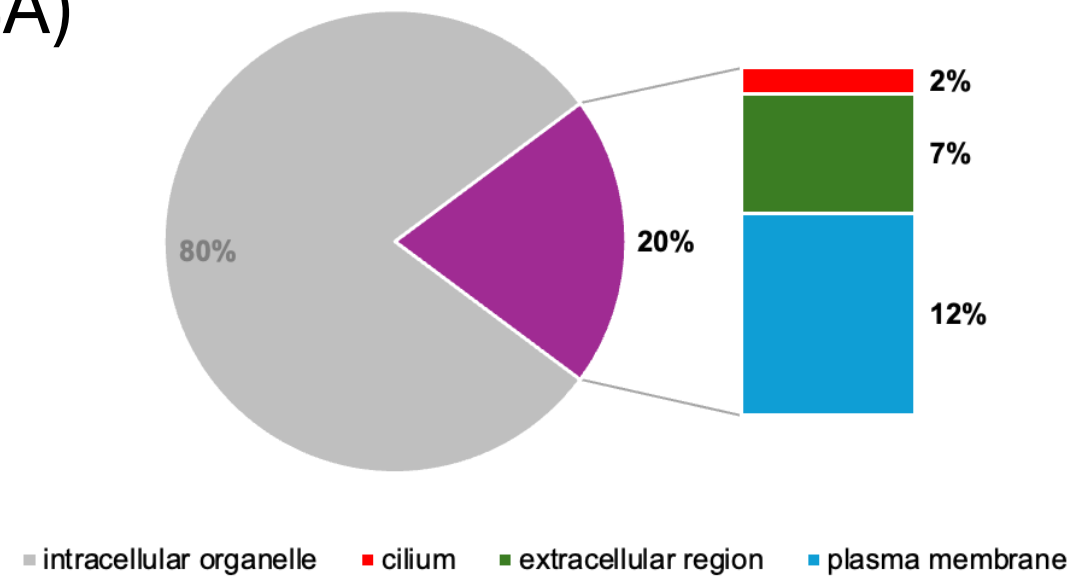


Figure 3. Curated Experimental HPI Obtained from Public Databases. The curated dataset of host-pathogen interactions (HPI) between mammals and bacteria, sourced from PSICQUIC (August 2022), a resource with interaction data from different interaction databases. (A) Mammalian species represented in curated HPI data (B) Firmicute species represented in curated HPI data.

(4A)



(4B)

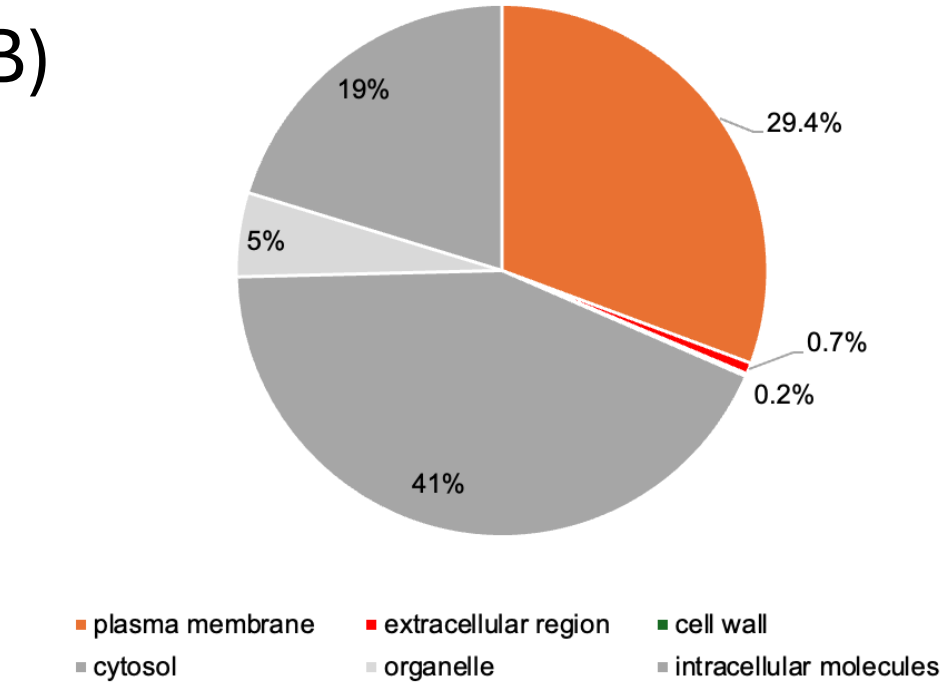


Figure 4. Subcellular Locations are Used to Assess Proteins Likely Involved in HPI. HPI proteins located on cell surfaces or secreted are most likely to interact. Localizations of proteins are identified using GO Cellular Compartment terms (A) Cellular Compartment distribution of mouse proteins in predicted HPI. (B) Cellular compartment distribution of *C. difficile* proteins in predicted HPI.

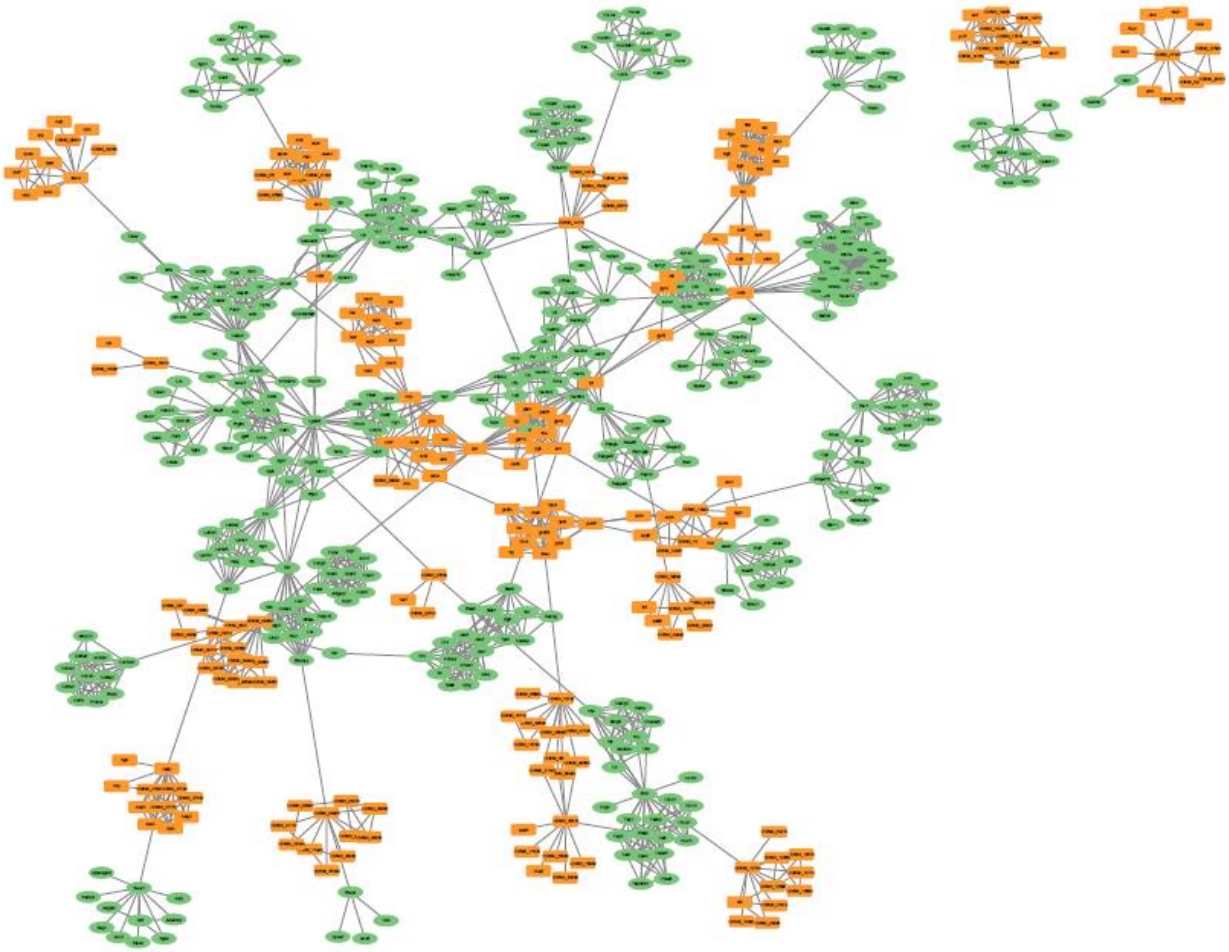


Figure 5. Combination of Predicted and Curated Interactions to Model *C. difficile* Infection in Mouse. This network diagram is generated from the predicted HPI and supplemented with intraspecies interactions identified for these proteins (accessed from STRING, April 2024). Mouse proteins are depicted with green boxes, while *C. difficile* 630 proteins are depicted with orange boxes. The lines between proteins represent interactions. Individual interactions used to create this network are available in Supplementary Table 4.

APPENDIX A: SUPPLEMENTAL DATA

Supplementary Table 1. **Human: Mouse Orthologs.** Orthologs were identified from Ensembl (Version 108) BioMart, and only strict (1:1) ortholog types were selected.

https://osf.io/27b4h/?view_only=4c043508e87440199fa8176a3d57c20d

Supplementary Table 2. **Reciprocal Best Matches for *C. Difficile* Strain 630 Vs Firmicute Proteins.** Reciprocal best hits were identified using BLASTP.

https://osf.io/27b4h/?view_only=4c043508e87440199fa8176a3d57c20d

Supplementary Table 3. **Mouse - *C. difficile* 630 Interologs.** These are based on experimental HPI derived from PSICQUIC.

https://osf.io/27b4h/?view_only=4c043508e87440199fa8176a3d57c20d

Supplementary Table 4. **Host-Pathogen Interactions between Mouse-*C. difficile* strain 630 Identified Using Network Analysis**

https://osf.io/27b4h/?view_only=4c043508e87440199fa8176a3d57c20d

REFERENCES

1. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. Clostridium difficile infection. *Nat Rev Dis Primer*. 2016;2(1):16020. doi:10.1038/nrdp.2016.20

2. Sandhu BK, McBride SM. Clostridioides difficile. *Trends Microbiol*. 2018;26(12):1049-1050. doi:10.1016/j.tim.2018.09.004

3. Spatz ES, Gottlieb M, Wisk LE, Anderson J, Chang AM, Gentile NL, Hill MJ, Huebinger RM, Idris AH, Kinsman J, Koo K, Li SX, McDonald S, Plumb ID, Rodriguez RM, Saydah S, Slovis B, Stephens KA, Unger ER, Wang RC, Yu H, Hota B, Elmore JG, Weinstein RA, Venkatesh A. Three-Month Symptom Profiles Among Symptomatic Adults With Positive and Negative Severe Acute Respiratory Syndrome Coronavirus 2 Tests: A Prospective Cohort Study From the INSPIRE Group. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2023;76(9):1559-1566. doi:10.1093/cid/ciac966

4. Centers for Disease Control and Prevention (U.S.). *Antibiotic Resistance Threats in the United States, 2019*. Centers for Disease Control and Prevention (U.S.); 2019. doi:10.15620/cdc:82532

5. Feuerstadt P, Theriault N, Tillotson G. The burden of CDI in the United States: a multifactorial challenge. *BMC Infect Dis*. 2023;23(1):132. doi:10.1186/s12879-023-08096-0

6. Guh Alice Y., Mu Yi, Winston Lisa G., Johnston Helen, Olson Danyel, Farley Monica M., Wilson Lucy E., Holzbauer Stacy M., Phipps Erin C., Dumyati Ghinwa K., Beldavs Zintars G., Kainer Marion A., Karlsson Maria, Gerding Dale N., McDonald L. Clifford. Trends in U.S. Burden of Clostridioides difficile Infection and Outcomes. *N Engl J Med*. 2020;382(14):1320-1330. doi:10.1056/NEJMoa1910215

7. Stewart D, Anwar F, Vedantam G. Anti-virulence strategies for Clostridioides difficile infection: advances and roadblocks. *Gut Microbes*. 2020;12(1):1802865. doi:10.1080/19490976.2020.1802865

8. Weese JS. Clostridium (Clostridioides) difficile in animals. *J Vet Diagn Investig Off Publ Am Assoc Vet Lab Diagn Inc*. 2020;32(2):213-221. doi:10.1177/1040638719899081

9. O'Shaughnessy RA, Habing GG, Gebreyes WA, Bowman AS, Weese JS, Rousseau J, Stull JW. Clostridioides difficile on Ohio swine farms (2015): A comparison of swine and human environments and assessment of on-farm risk factors. *Zoonoses Public Health*. 2019;66(7):861-870. doi:10.1111/zph.12637

10. Kirk JA, Banerji O, Fagan RP. Characteristics of the Clostridium difficile cell envelope and its importance in therapeutics. *Microb Biotechnol*. 2017;10(1):76-90. doi:10.1111/1751-7915.12372

11. Leffler DA, Lamont JT. Clostridium difficile infection. *N Engl J Med*. 2015;372(16):1539-1548. doi:10.1056/NEJMra1403772

12. Guerrero DM, Becker JC, Eckstein EC, Kundrapu S, Deshpande A, Sethi AK, Donskey CJ. Asymptomatic carriage of toxigenic *Clostridium difficile* by hospitalized patients. *J Hosp Infect*. 2013;85(2):155-158. doi:10.1016/j.jhin.2013.07.002

13. Crobach MJT, Vernon JJ, Loo VG, Kong LY, Péchiné S, Wilcox MH, Kuijper EJ. Understanding Clostridium difficile Colonization. *Clin Microbiol Rev*. 2018;31(2):e00021-17. doi:10.1128/CMR.00021-17

14. Debast SB, Bauer MP, Kuijper EJ. European Society of Clinical Microbiology and Infectious Diseases: Update of the Treatment Guidance Document for Clostridium difficile Infection. *Clin Microbiol Infect*. 2014;20:1-26. doi:10.1111/1469-0691.12418

15. Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M, Riedel T, Bunk B, Spröer C, Overmann J, Blaschitz M, Indra A, von Müller L, Kohl TA, Niemann S, Seyboldt C, Klawonn F, Kumar N, Lawley TD, García-Fernández S, Cantón R, del Campo R, Zimmermann O, Groß U, Achtman M, Nübel U. A publicly accessible database for Clostridioides difficile genome sequences supports tracing of transmission chains and epidemics. *Microb Genomics*. 2020;6(8):mgen000410. doi:10.1099/mgen.0.000410

16. Stabler RA, Gerding DN, Songer JG, Drudy D, Brazier JS, Trinh HT, Witney AA, Hinds J, Wren BW. Comparative phylogenomics of Clostridium difficile reveals clade specificity and microevolution of hypervirulent strains. *J Bacteriol*. 2006;188(20):7297-7305. doi:10.1128/JB.00664-06

17. Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, Fung R, Golubchik T, Harding RM, Jeffery KJM, Jolley KA, Kirton R, Peto TE, Rees G, Stoesser N, Vaughan A, Walker AS, Young BC, Wilcox M, Dingle KE. Multilocus Sequence Typing of Clostridium difficile. *J Clin Microbiol*. 2010;48(3):770-778. doi:10.1128/jcm.01796-09

18. Janezic S, Potocnik M, Zidaric V, Rupnik M. Highly Divergent Clostridium difficile Strains Isolated from the Environment. *PLOS ONE*. 2016;11(11):e0167101. doi:10.1371/journal.pone.0167101

19. Martínez-Meléndez A, Morfin-Otero R, Villarreal-Treviño L, Baines SD, Camacho-Ortíz A, Garza-González E. Molecular epidemiology of predominant and emerging *Clostridioides difficile* ribotypes. *J Microbiol Methods*. 2020;175:105974. doi:10.1016/j.mimet.2020.105974

20. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MT, Wright A, Churcher C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat Genet*. 2006;38(7):779-786. doi:10.1038/ng1830

21. Bauer MP, Notermans DW, Van Benthem BH, Brazier JS, Wilcox MH, Rupnik M, Monnet DL, Van Dissel JT, Kuijper EJ. Clostridium difficile infection in Europe: A hospital-based survey. *The Lancet*. 2011;377(9759):63-73. doi:10.1016/S0140-6736(10)61266-4

22. Tickler IA, Goering RV, Whitmore JD, Lynn ANW, Persing DH, Tenover FC, for the Healthcare Associated Infection Consortium. Strain Types and Antimicrobial Resistance Patterns of Clostridium difficile Isolates from the United States, 2011 to 2013. *Antimicrob Agents Chemother*. 2014;58(7):4214-4218. doi:10.1128/aac.02775-13

23. Anwar F, Roxas BAP, Shehab KW, Ampel NM, Viswanathan VK, Vedantam G. Low-toxin Clostridioides difficile RT027 strains exhibit robust virulence. *Emerg Microbes Infect*. 11(1):1982-1993. doi:10.1080/22221751.2022.2105260

24. Patterson L, Wilcox MH, Fawley WN, Verlander NQ, Geoghegan L, Patel BC, Wyatt T, Smyth B. Morbidity and mortality associated with *Clostridium difficile* ribotype 078: a case–case study. *J Hosp Infect*. 2012;82(2):125-128. doi:10.1016/j.jhin.2012.07.011

25. Goorhuis A, Debast SB, van Leengoed LAMG, Harmanus C, Notermans DW, Bergwerff AA, Kuijper EJ. Clostridium difficile PCR Ribotype 078: an Emerging Strain in Humans and in Pigs? *J Clin Microbiol*. 2008;46(3):1157-1158. doi:10.1128/jcm.01536-07

26. Deakin LJ, Clare S, Fagan RP, Dawson LF, Pickard DJ, West MR, Wren BW, Fairweather NF, Dougan G, Lawley TD. The Clostridium difficile spo0A Gene Is a Persistence and Transmission Factor. *Infect Immun*. 2012;80(8):2704-2711. doi:10.1128/IAI.00147-12

27. Lawley TD, Croucher NJ, Yu L, Clare S, Sebaihia M, Goulding D, Pickard DJ, Parkhill J, Choudhary J, Dougan G. Proteomic and genomic characterization of highly infectious Clostridium difficile 630 spores. *J Bacteriol*. 2009;191(17):5377-5386. doi:10.1128/JB.00597-09

28. Paredes-Sabja D, Shen A, Sorg JA. Clostridium difficile spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol*. 2014;22(7):406-416. doi:10.1016/j.tim.2014.04.003

29. Bhattacharjee D, Francis MB, Ding X, McAllister KN, Shrestha R, Sorg JA. Reexamining the Germination Phenotypes of Several Clostridium difficile Strains Suggests Another Role for the CspC Germinant Receptor. *J Bacteriol*. 2015;198(5):777-786. doi:10.1128/JB.00908-15

30. Shrestha R, Sorg JA. Terbium chloride influences Clostridium difficile spore germination. *Anaerobe*. 2019;58:80-88. doi:10.1016/j.anaerobe.2019.03.016

31. Kordus SL, Thomas AK, Lacy DB. Clostridioides difficile toxins: mechanisms of action and antitoxin therapeutics. *Nat Rev Microbiol*. Published online November 26, 2021. doi:10.1038/s41579-021-00660-2

32. Jump RLP, Pultz MJ, Donskey CJ. Vegetative Clostridium difficile survives in room air on moist surfaces and in gastric contents with reduced acidity: a potential mechanism to explain

the association between proton pump inhibitors and C. difficile-associated diarrhea? *Antimicrob Agents Chemother*. 2007;51(8):2883-2887. doi:10.1128/AAC.01443-06

33. Henriques AO, Moran CP. Structure, assembly, and function of the spore surface layers. *Annu Rev Microbiol*. 2007;61:555-588. doi:10.1146/annurev.micro.61.080706.093224

34. Paredes-Sabja D, Sarker MR. Adherence of Clostridium difficile spores to Caco-2 cells in culture. *J Med Microbiol*. 2012;61(Pt 9):1208-1218. doi:10.1099/jmm.0.043687-0

35. Shrestha R, Lockless SW, Sorg JA. A Clostridium difficile alanine racemase affects spore germination and accommodates serine as a substrate. *J Biol Chem*. 2017;292(25):10735-10742. doi:10.1074/jbc.M117.791749

36. Mansfield MJ, Tremblay BJM, Zeng J, Wei X, Hodgins H, Worley J, Bry L, Dong M, Doxey AC. Phylogenomics of 8,839 Clostridioides difficile genomes reveals recombination-driven evolution and diversification of toxin A and B. *PLOS Pathog*. 2020;16(12):e1009181. doi:10.1371/journal.ppat.1009181

37. Jafari NV, Kuehne SA, Minton NP, Allan E, Bajaj-Elliott M. Clostridium difficile-mediated effects on human intestinal epithelia: Modelling host-pathogen interactions in a vertical diffusion chamber. *Anaerobe*. 2016;37:96-102. doi:10.1016/j.anaerobe.2015.12.007

38. Ransom EM, Kaus GM, Tran PM, Ellermeier CD, Weiss DS. Multiple factors contribute to bimodal toxin gene expression in Clostridioides (Clostridium) difficile. *Mol Microbiol*. 2018;110(4):533-549. doi:10.1111/mmi.14107

39. Dupuy B, Raffestin S, Matamouros S, Mani N, Popoff MR, Sonenshein AL. Regulation of toxin and bacteriocin gene expression in Clostridium by interchangeable RNA polymerase sigma factors. *Mol Microbiol*. 2006;60(4):1044-1057. doi:10.1111/j.1365-2958.2006.05159.x

40. Matamouros S, England P, Dupuy B. Clostridium difficile toxin expression is inhibited by the novel regulator TcdC. *Mol Microbiol*. 2007;64(5):1274-1288. doi:10.1111/j.1365-2958.2007.05739.x

41. Orrell KE, Melnyk RA. Large Clostridial Toxins: Mechanisms and Roles in Disease. *Microbiol Mol Biol Rev*. 2021;85(3):10.1128/mmbr.00064-21. doi:10.1128/mmbr.00064-21

42. Pruitt RN, Chambers MG, Ng KKS, Ohi MD, Lacy DB. Structural organization of the functional domains of Clostridium difficile toxins A and B. *Proc Natl Acad Sci U S A*. 2010;107(30):13467-13472. doi:10.1073/pnas.1002199107

43. Chandrasekaran R, Kenworthy AK, Lacy DB. Clostridium difficile Toxin A Undergoes Clathrin-Independent, PACSIN2-Dependent Endocytosis. *PLoS Pathog*. 2016;12(12):e1006070. doi:10.1371/journal.ppat.1006070

44. Papatheodorou P, Barth H, Minton N, Aktories K. Cellular Uptake and Mode-of-Action of Clostridium difficile Toxins. *Adv Exp Med Biol*. 2018;1050:77-96. doi:10.1007/978-3-319-72799-8_6

45. Geny B, Popoff MR. Bacterial protein toxins and lipids: pore formation or toxin entry into cells. *Biol Cell*. 2006;98(11):667-678. doi:10.1042/BC20050082

46. Reineke J, Tenzer S, Rupnik M, Koschinski A, Hasselmayer O, Schrattenholz A, Schild H, von Eichel-Streiber C. Autocatalytic cleavage of Clostridium difficile toxin B. *Nature*. 2007;446(7134):415-419. doi:10.1038/nature05622

47. Krivan HC, Clark GF, Smith DF, Wilkins TD. Cell surface binding site for Clostridium difficile enterotoxin: evidence for a glycoconjugate containing the sequence Gal alpha 1-3Gal beta 1-4GlcNAc. *Infect Immun*. 1986;53(3):573-581. doi:10.1128/iai.53.3.573-581.1986

48. Teneberg S, Lönnroth I, Torres López JF, Galili U, Halvarsson MO, Angström J, Karlsson KA. Molecular mimicry in the recognition of glycosphingolipids by Gal alpha 3 Gal beta 4 GlcNAc beta-binding Clostridium difficile toxin A, human natural anti alpha-galactosyl IgG and the monoclonal antibody Gal-13: characterization of a binding-active human glycosphingolipid, non-identical with the animal receptor. *Glycobiology*. 1996;6(6):599-609. doi:10.1093/glycob/6.6.599

49. Pothoulakis C, Gilbert RJ, Cladaras C, Castagliuolo I, Semenza G, Hitti Y, Montcrief JS, Linevsky J, Kelly CP, Nikulasson S, Desai HP, Wilkins TD, LaMont JT. Rabbit sucrase-isomaltase contains a functional intestinal receptor for Clostridium difficile toxin A. *J Clin Invest*. 1996;98(3):641-649. doi:10.1172/JCI118835

50. Na X, Kim H, Moyer MP, Pothoulakis C, LaMont JT. gp96 is a human colonocyte plasma membrane binding protein for Clostridium difficile toxin A. *Infect Immun*. 2008;76(7):2862-2871. doi:10.1128/IAI.00326-08

51. Yuan P, Zhang H, Cai C, Zhu S, Zhou Y, Yang X, He R, Li C, Guo S, Li S, Huang T, Perez-Cordon G, Feng H, Wei W. Chondroitin sulfate proteoglycan 4 functions as the cellular receptor for Clostridium difficile toxin B. *Cell Res*. 2015;25(2):157-168. doi:10.1038/cr.2014.169

52. Tao L, Zhang J, Meraner P, Tovaglieri A, Wu X, Gerhard R, Zhang X, Stallcup WB, Miao J, He X, Hurdle JG, Breault DT, Brass AL, Dong M. Frizzled proteins are colonic epithelial receptors for C. difficile toxin B. *Nature*. 2016;538(7625):350-355. doi:10.1038/nature19799

53. LaFrance ME, Farrow MA, Chandrasekaran R, Sheng J, Rubin DH, Lacy DB. Identification of an epithelial cell receptor responsible for Clostridium difficile TcdB-induced cytotoxicity. *Proc Natl Acad Sci U S A*. 2015;112(22):7073-7078. doi:10.1073/pnas.1500791112

54. Tian S, Xiong X, Zeng J, Wang S, Tremblay BJM, Chen P, Chen B, Liu M, Chen P, Sheng K, Zeve D, Qi W, Breault DT, Rodríguez C, Gerhard R, Jin R, Doxey AC, Dong M. Identification of TFPI as a receptor reveals recombination-driven receptor switching in

Clostridioides difficile toxin B variants. *Nat Commun*. 2022;13(1):6786. doi:10.1038/s41467-022-33964-9

55. Orrell KE, Mansfield MJ, Doxey AC, Melnyk RA. The C. difficile toxin B membrane translocation machinery is an evolutionarily conserved protein delivery apparatus. *Nat Commun*. 2020;11(1):432. doi:10.1038/s41467-020-14306-z

56. Stare BG, Delmée M, Rupnik M. Variant forms of the binary toxin CDT locus and tcdC gene in Clostridium difficile strains. *J Med Microbiol*. 2007;56(Pt 3):329-335. doi:10.1099/jmm.0.46931-0

57. Perelle S, Gibert M, Bourlioux P, Corthier G, Popoff MR. Production of a complete binary toxin (actin-specific ADP-ribosyltransferase) by Clostridium difficile CD196. *Infect Immun*. 1997;65(4):1402-1407.

58. Gonçalves C, Decré D, Barbut F, Burghoffer B, Petit JC. Prevalence and Characterization of a Binary Toxin (Actin-Specific ADP-Ribosyltransferase) from Clostridium difficile. *J Clin Microbiol*. 2004;42(5):1933-1939. doi:10.1128/JCM.42.5.1933-1939.2004

59. Papatheodorou P, Carette JE, Bell GW, Schwan C, Guttenberg G, Brummelkamp TR, Aktories K. Lipolysis-stimulated lipoprotein receptor (LSR) is the host receptor for the binary toxin Clostridium difficile transferase (CDT). *Proc Natl Acad Sci U S A*. 2011;108(39):16422-16427. doi:10.1073/pnas.1109772108

60. Anderson DM, Sheedlo MJ, Jensen JL, Lacy DB. Structural insights into the transition of Clostridioides difficile binary toxin from prepore to pore. *Nat Microbiol*. 2020;5(1):102-107. doi:10.1038/s41564-019-0601-8

61. Kaiser E, Kroll C, Ernst K, Schwan C, Popoff M, Fischer G, Buchner J, Aktories K, Barth H. Membrane Translocation of Binary Actin-ADP-Ribosylating Toxins from Clostridium difficile and Clostridium perfringens Is Facilitated by Cyclophilin A and Hsp90 ▿. *Infect Immun*. 2011;79(10):3913-3921. doi:10.1128/IAI.05372-11

62. Schwan C, Stecher B, Tzivelekidis T, van Ham M, Rohde M, Hardt WD, Wehland J, Aktories K. Clostridium difficile Toxin CDT Induces Formation of Microtubule-Based Protrusions and Increases Adherence of Bacteria. *PLoS Pathog*. 2009;5(10):e1000626. doi:10.1371/journal.ppat.1000626

63. Wegner A, Aktories K. ADP-ribosylated actin caps the barbed ends of actin filaments. *J Biol Chem*. 1988;263(27):13739-13742.

64. A Clostridium difficile gene encoding flagellinThe GenBank accession numbers for the sequences reported in this paper are AF065259 (strain 79-685) and AF077341 (strain VPI 10463). | Microbiology Society. Accessed April 15, 2024. https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-146-4-957

65. Macnab RM. How Bacteria Assemble Flagella. *Annu Rev Microbiol*. 2003;57(Volume 57, 2003):77-100. doi:10.1146/annurev.micro.57.030502.090832

66. Duan Q, Zhou M, Zhu L, Zhu G. Flagella and bacterial pathogenicity. *J Basic Microbiol*. 2013;53(1):1-8. doi:10.1002/jobm.201100335

67. Maldarelli GA, Piepenbrink KH, Scott AJ, Freiberg JA, Song Y, Achermann Y, Ernst RK, Shirtliff ME, Sundberg EJ, Donnenberg MS, von Rosenvinge EC. Type IV pili promote early biofilm formation by Clostridium difficile. *Pathog Dis*. 2016;74(6):ftw061. doi:10.1093/femspd/ftw061

68. Purcell EB, McKee RW, Bordeleau E, Burrus V, Tamayo R. Regulation of Type IV Pili Contributes to Surface Behaviors of Historical and Epidemic Strains of Clostridium difficile. *J Bacteriol*. 2016;198(3):565-577. doi:10.1128/JB.00816-15

69. Bharat TAM, von Kügelgen A, Alva V. Molecular Logic of Prokaryotic Surface Layer Structures. *Trends Microbiol*. 2021;29(5):405-415. doi:10.1016/j.tim.2020.09.009

70. Waligora AJ, Hennequin C, Mullany P, Bourlioux P, Collignon A, Karjalainen T. Characterization of a Cell Surface Protein ofClostridium difficile with Adhesive Properties. *Infect Immun*. Published online April 1, 2001. doi:10.1128/IAI.69.4.2144-2153.2001

71. Fagan RP, Fairweather NF. Biogenesis and functions of bacterial S-layers. *Nat Rev Microbiol*. 2014;12(3):211-222. doi:10.1038/nrmicro3213

72. Ravi J, Fioravanti A. S-layers: The Proteinaceous Multifunctional Armors of Gram-Positive Pathogens. *Front Microbiol*. 2021;12:663468. doi:10.3389/fmicb.2021.663468

73. Kirby JM, Ahern H, Roberts AK, Kumar V, Freeman Z, Acharya KR, Shone CC. Cwp84, a Surface-associated Cysteine Protease, Plays a Role in the Maturation of the Surface Layer of Clostridium difficile *. *J Biol Chem*. 2009;284(50):34666-34673. doi:10.1074/jbc.M109.051177

74. Calabi E, Calabi F, Phillips AD, Fairweather NF. Binding of Clostridium difficile Surface Layer Proteins to Gastrointestinal Tissues. *Infect Immun*. 2002;70(10):5770-5778. doi:10.1128/iai.70.10.5770-5778.2002

75. Nibbering B, Gerding DN, Kuijper EJ, Zwittink RD, Smits WK. Host Immune Responses to Clostridioides difficile: Toxins and Beyond. *Front Microbiol*. 2021;12:804949. doi:10.3389/fmicb.2021.804949

76. Calabi E, Ward S, Wren B, Paxton T, Panico M, Morris H, Dell A, Dougan G, Fairweather N. Molecular characterization of the surface layer proteins from Clostridium difficile. *Mol Microbiol*. 2001;40(5):1187-1199. doi:10.1046/j.1365-2958.2001.02461.x

77. Caminero A, Guzman M, Libertucci J, Lomax AE. The emerging roles of bacterial proteases in intestinal diseases. *Gut Microbes*. 2023;15(1):2181922. doi:10.1080/19490976.2023.2181922

78. Janoir C, Grénery J, Savariau-Lacomme MP, Collignon A. [Characterization of an extracellular protease from Clostridium difficile]. *Pathol Biol (Paris)*. 2004;52(8):444-449. doi:10.1016/j.patbio.2004.07.025

79. de la Riva L, Willing SE, Tate EW, Fairweather NF. Roles of cysteine proteases Cwp84 and Cwp13 in biogenesis of the cell wall of Clostridium difficile. *J Bacteriol*. 2011;193(13):3276-3285. doi:10.1128/JB.00248-11

80. Sekulovic O, Ospina Bedoya M, Fivian-Hughes AS, Fairweather NF, Fortier LC. The Clostridium difficile cell wall protein CwpV confers phase-variable phage resistance. *Mol Microbiol*. 2015;98(2):329-342. doi:10.1111/mmi.13121

81. Cascioferro S, Totsika M, Schillaci D. Sortase A: an ideal target for anti-virulence drug development. *Microb Pathog*. 2014;77:105-112. doi:10.1016/j.micpath.2014.10.007

82. Donahue EH, Dawson LF, Valiente E, Firth-Clark S, Major MR, Littler E, Perrior TR, Wren BW. Clostridium difficile has a single sortase, SrtB, that can be inhibited by small-molecule inhibitors. *BMC Microbiol*. 2014;14:219. doi:10.1186/s12866-014-0219-1

83. Peltier J, Shaw HA, Wren BW, Fairweather NF. Disparate subcellular location of putative sortase substrates in Clostridium difficile. *Sci Rep*. 2017;7(1):9204. doi:10.1038/s41598-017-08322-1

84. Tulli L, Marchi S, Petracca R, Shaw HA, Fairweather NF, Scarselli M, Soriani M, Leuzzi R. CbpA: a novel surface exposed adhesin of Clostridium difficile targeting human collagen. *Cell Microbiol*. 2013;15(10):1674-1687. doi:10.1111/cmi.12139

85. Foster TJ. The MSCRAMM Family of Cell-Wall-Anchored Surface Proteins of Gram-Positive Cocci. *Trends Microbiol*. 2019;27(11):927-941. doi:10.1016/j.tim.2019.06.007

86. Hensbergen PJ, Klychnikov OI, Bakker D, van Winden VJC, Ras N, Kemp AC, Cordfunke RA, Dragan I, Deelder AM, Kuijper EJ, Corver J, Drijfhout JW, van Leeuwen HC. A Novel Secreted Metalloprotease (CD2830) from *Clostridium difficile* Cleaves Specific Proline Sequences in LPXTG Cell Surface Proteins. *Mol Cell Proteomics*. 2014;13(5):1231-1244. doi:10.1074/mcp.M113.034728

87. Hennequin C, Janoir C, Barc MC, Collignon A, Karjalainen T. Identification and characterization of a fibronectin-binding protein from Clostridium difficile. *Microbiol Read Engl*. 2003;149(Pt 10):2779-2787. doi:10.1099/mic.0.26145-0

88. Péchiné S, Gleizes A, Janoir C, Gorges-Kergot R, Barc MC, Delmée M, Collignon A. Immunological properties of surface proteins of Clostridium difficile. *J Med Microbiol*. 2005;54(Pt 2):193-196. doi:10.1099/jmm.0.45800-0

89. Wright A, Drudy D, Kyne L, Brown K, Fairweather NF. Immunoreactive cell wall proteins of Clostridium difficile identified by human sera. *J Med Microbiol*. 2008;57(Pt 6):750-756. doi:10.1099/jmm.0.47532-0

90. Kovacs-Simon A, Leuzzi R, Kasendra M, Minton N, Titball RW, Michell SL. Lipoprotein CD0873 Is a Novel Adhesin of Clostridium difficile. *J Infect Dis*. 2014;210(2):274-284. doi:10.1093/infdis/jiu070

91. Hensbergen PJ, Klychnikov OI, Bakker D, Dragan I, Kelly ML, Minton NP, Corver J, Kuijper EJ, Drijfhout JW, van Leeuwen HC. Clostridium difficile secreted Pro-Pro endopeptidase PPEP-1 (ZMP1/CD2830) modulates adhesion through cleavage of the collagen binding protein CD2831. *FEBS Lett*. 2015;589(24PartB):3952-3958. doi:10.1016/j.febslet.2015.10.027

92. Péchiné S, Hennequin C, Boursier C, Hoys S, Collignon A. Immunization using GroEL decreases Clostridium difficile intestinal colonization. *PloS One*. 2013;8(11):e81112. doi:10.1371/journal.pone.0081112

93. Turner JR. Intestinal mucosal barrier function in health and disease. *Nat Rev Immunol*. 2009;9(11):799-809. doi:10.1038/nri2653

94. Vancamelbeke M, Vermeire S. The intestinal barrier: a fundamental role in health and disease. *Expert Rev Gastroenterol Hepatol*. 2017;11(9):821-834. doi:10.1080/17474124.2017.1343143

95. Glomski IJ, Piris-Gimenez A, Huerre M, Mock M, Goossens PL. Primary Involvement of Pharynx and Peyer's Patch in Inhalational and Intestinal Anthrax. *PLOS Pathog*. 2007;3(6):e76. doi:10.1371/journal.ppat.0030076

96. Shen A. Clostridioides difficile Spore Formation and Germination: New Insights and Opportunities for Intervention. *Annu Rev Microbiol*. 2020;74(Volume 74, 2020):545-566. doi:10.1146/annurev-micro-011320-011321

97. Emami CN, Petrosyan M, Giuliani S, Williams M, Hunter C, Prasadarao NV, Ford HR. Role of the Host Defense System and Intestinal Microbial Flora in the Pathogenesis of Necrotizing Enterocolitis. *Surg Infect*. 2009;10(5):407-417. doi:10.1089/sur.2009.054

98. Forchielli ML, Walker WA. The role of gut-associated lymphoid tissues and mucosal defence. *Br J Nutr*. 2005;93 Suppl 1:S41-48. doi:10.1079/bjn20041356

99. Warny M, Keates AC, Keates S, Castagliuolo I, Zacks JK, Aboudola S, Qamar A, Pothoulakis C, LaMont JT, Kelly CP. p38 MAP kinase activation by Clostridium difficile toxin A mediates monocyte necrosis, IL-8 production, and enteritis. *J Clin Invest*. 2000;105(8):1147-1156. doi:10.1172/JCI7545

100. Cowardin CA, Buonomo EL, Saleh MM, Wilson MG, Burgess SL, Kuehne SA, Schwan C, Eichhoff AM, Koch-Nolte F, Lyras D, Aktories K, Minton NP, Petri WA. The binary toxin CDT enhances Clostridium difficile virulence by suppressing protective colonic eosinophilia. *Nat Microbiol*. 2016;1(8):16108. doi:10.1038/nmicrobiol.2016.108

101. Ryan A, Lynch M, Smith SM, Amu S, Nel HJ, McCoy CE, Dowling JK, Draper E, O'Reilly V, McCarthy C, O'Brien J, Eidhin DN, O'Connell MJ, Keogh B, Morton CO, Rogers TR,

Fallon PG, O'Neill LA, Kelleher D, Loscher CE. A Role for TLR4 in Clostridium difficile Infection and the Recognition of Surface Layer Proteins. *PLOS Pathog*. 2011;7(6):e1002076. doi:10.1371/journal.ppat.1002076

102. Batah J, Denève-Larrazet C, Jolivot PA, Kuehne S, Collignon A, Marvaud JC, Kansau I. Clostridium difficile flagella predominantly activate TLR5-linked NF-κB pathway in epithelial cells. *Anaerobe*. 2016;38:116-124. doi:10.1016/j.anaerobe.2016.01.002

103. Lynch M, Walsh TA, Marszalowska I, Webb AE, Mac Aogain M, Rogers TR, Windle H, Kelleher D, O'Connell MJ, Loscher CE. Surface layer proteins from virulent Clostridium difficile ribotypes exhibit signatures of positive selection with consequences for innate immune response. *BMC Evol Biol*. 2017;17(1):90. doi:10.1186/s12862-017-0937-8

104. Sun X, Hirota SA. The roles of host and pathogen factors and the innate immune response in the pathogenesis of Clostridium difficile infection. *Mol Immunol*. 2015;63(2):193-202. doi:10.1016/j.molimm.2014.09.005

105. Warny M, Vaerman JP, Avesani V, Delmée M. Human antibody response to Clostridium difficile toxin A in relation to clinical course of infection. *Infect Immun*. 1994;62(2):384-389. doi:10.1128/iai.62.2.384-389.1994

106. Hutton ML, Mackin KE, Chakravorty A, Lyras D. Small animal models for the study of Clostridium difficile disease pathogenesis. *FEMS Microbiol Lett*. 2014;352(2):140-149. doi:10.1111/1574-6968.12367

107. Best EL, Freeman J, Wilcox MH. Models for the study of Clostridium difficile infection. *Gut Microbes*. 2012;3(2):145-167. doi:10.4161/gmic.19526

108. Delmée M, Avesani V. Virulence of ten serogroups of Clostridium difficile in hamsters. *J Med Microbiol*. 1990;33(2):85-90. doi:10.1099/00222615-33-2-85

109. George WL, Sutter VL, Goldstein EJ, Ludwig SL, Finegold SM. Aetiology of antimicrobial-agent-associated colitis. *Lancet Lond Engl*. 1978;1(8068):802-803. doi:10.1016/s0140-6736(78)93001-5

110. Razaq N, Sambol S, Nagaro K, Zukowski W, Cheknis A, Johnson S, Gerding DN. Infection of hamsters with historical and epidemic BI types of Clostridium difficile. *J Infect Dis*. 2007;196(12):1813-1819. doi:10.1086/523106

111. Sambol SP, Tang JK, Merrigan MM, Johnson S, Gerding DN. Infection of hamsters with epidemiologically important strains of Clostridium difficile. *J Infect Dis*. 2001;183(12):1760-1766. doi:10.1086/320736

112. Lawley TD, Clare S, Walker AW, Goulding D, Stabler RA, Croucher N, Mastroeni P, Scott P, Raisen C, Mottram L, Fairweather NF, Wren BW, Parkhill J, Dougan G. Antibiotic treatment of clostridium difficile carrier mice triggers a supershedder state, spore-mediated transmission, and severe disease in immunocompromised hosts. *Infect Immun*. 2009;77(9):3661-3669. doi:10.1128/IAI.00558-09

113. Pawlowski SW, Calabrese G, Kolling GL, Platts-Mills J, Freire R, AlcantaraWarren C, Liu B, Sartor RB, Guerrant RL. Murine model of Clostridium difficile infection with aged gnotobiotic C57BL/6 mice and a BI/NAP1 strain. *J Infect Dis*. 2010;202(11):1708-1712. doi:10.1086/657086

114. Fachi JL, Vinolo MAR, Colonna M. Reviewing the Clostridioides difficile Mouse Model: Insights into Infection Mechanisms. *Microorganisms*. 2024;12(2):273. doi:10.3390/microorganisms12020273

115. Steele J, Feng H, Parry N, Tzipori S. Piglet models of acute or chronic Clostridium difficile illness. *J Infect Dis*. 2010;201(3):428-434. doi:10.1086/649799

116. Mada PK, Alam MU. Clostridioides difficile infection. In: *StatPearls*. StatPearls Publishing; 2024. Accessed August 8, 2024. http://www.ncbi.nlm.nih.gov/books/NBK431054/

117. Stevens VW, Nelson RE, Schwab-Daugherty EM, Khader K, Jones MM, Brown KA, Greene T, Croft LD, Neuhauser M, Glassman P, Goetz MB, Samore MH, Rubin MA. Comparative Effectiveness of Vancomycin and Metronidazole for the Prevention of Recurrence and Death in Patients With Clostridium difficile Infection. *JAMA Intern Med*. 2017;177(4):546-553. doi:10.1001/jamainternmed.2016.9045

118. Cornely OA, Miller MA, Louie TJ, Crook DW, Gorbach SL. Treatment of first recurrence of Clostridium difficile infection: fidaxomicin versus vancomycin. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2012;55 Suppl 2(Suppl 2):S154-161. doi:10.1093/cid/cis462

119. Wilcox MH, McGovern BH, Hecht GA. The Efficacy and Safety of Fecal Microbiota Transplant for Recurrent Clostridium difficile Infection: Current Understanding and Gap Analysis. *Open Forum Infect Dis*. 2020;7(5):ofaa114. doi:10.1093/ofid/ofaa114

120. Carvalho T. First oral fecal microbiota transplant therapy approved. *Nat Med*. 2023;29(7):1581-1582. doi:10.1038/d41591-023-00046-2

121. Cho JC, Crotty MP, Pardo J. Ridinilazole: a novel antimicrobial for Clostridium difficile infection. *Ann Gastroenterol*. 2019;32(2):134-140. doi:10.20524/aog.2018.0336

122. Bruxelle JF, Mizrahi A, Hoys S, Collignon A, Janoir C, Péchiné S. Immunogenic properties of the surface layer precursor of *Clostridium difficile* and vaccination assays in animal models. *Anaerobe*. 2016;37:78-84. doi:10.1016/j.anaerobe.2015.10.010

123. Ní Eidhin DB, O'Brien JB, McCabe MS, Athié-Morales V, Kelleher DP. Active immunization of hamsters against Clostridium difficile infection using surface-layer protein. *FEMS Immunol Med Microbiol*. 2008;52(2):207-218. doi:10.1111/j.1574-695X.2007.00363.x

124. Ghose C, Eugenis I, Sun X, Edwards AN, McBride SM, Pride DT, Kelly CP, Ho DD. Immunogenicity and protective efficacy of recombinant Clostridium difficile flagellar protein FliC. *Emerg Microbes Infect*. 2016;5(1):1-10. doi:10.1038/emi.2016.8

125. de Bruyn G, Saleh J, Workman D, Pollak R, Elinoff V, Fraser NJ, Lefebvre G, Martens M, Mills RE, Nathan R, Trevino M, van Cleeff M, Foglia G, Ozol-Godfrey A, Patel DM, Pietrobon PJ, Gesser R, H-030-012 Clinical Investigator Study Team. Defining the optimal formulation and schedule of a candidate toxoid vaccine against Clostridium difficile infection: A randomized Phase 2 clinical trial. *Vaccine*. 2016;34(19):2170-2178. doi:10.1016/j.vaccine.2016.03.028

126. Sferra TJ, Merta T, Neely M, Murta de Oliveira C, Lassaletta A, Fortuny Guasch C, Dorr MB, Winchell G, Su FH, Perko S, Fernsler D, Waskin H, Holden SR. Double-Blind, Placebo-Controlled Study of Bezlotoxumab in Children Receiving Antibacterial Treatment for Clostridioides difficile Infection (MODIFY III). *J Pediatr Infect Dis Soc*. 2023;12(6):334-341. doi:10.1093/jpids/piad031

127. Dieterle MG, Young VB. Reducing Recurrence of *C. difficile* Infection. *Cell*. 2017;169(3):375. doi:10.1016/j.cell.2017.03.039

128. Johnson S, Lavergne V, Skinner AM, Gonzales-Luna AJ, Garey KW, Kelly CP, Wilcox MH. Clinical Practice Guideline by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA): 2021 Focused Update Guidelines on Management of Clostridioides difficile Infection in Adults. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2021;73(5):e1029-e1044. doi:10.1093/cid/ciab549

129. Indra A, Huhulescu S, Schneeweis M, Hasenberger P, Kernbichler S, Fiedler A, Wewalka G, Allerberger F, Kuijper EJ. Characterization of Clostridium difficile isolates using capillary gel electrophoresis-based PCR ribotyping. *J Med Microbiol*. 2008;57(Pt 11):1377-1382. doi:10.1099/jmm.0.47714-0

130. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3:124. doi:10.12688/wellcomeopenres.14826.1

131. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, Ong WK, Paley SM, Subhraveti P. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform*. 2019;20(4):1085-1093. doi:10.1093/bib/bbx085

132. Arrieta-Ortiz ML, Immanuel SRC, Turkarslan S, Wu WJ, Girinathan BP, Worley JN, DiBenedetto N, Soutourina O, Peltier J, Dupuy B, Bry L, Baliga NS. Predictive regulatory and metabolic network models for systems analysis of *Clostridioides difficile*. *Cell Host Microbe*. 2021;29(11):1709-1723.e5. doi:10.1016/j.chom.2021.09.008

133. Aranda B, Blankenburg H, Kerrien S, Brinkman FSL, Ceol A, Chautard E, Dana JM, De Las Rivas J, Dumousseau M, Galeota E, Gaulton A, Goll J, Hancock REW, Isserlin R, Jimenez RC, Kerssemakers J, Khadake J, Lynn DJ, Michaut M, O'Kelly G, Ono K, Orchard S, Prieto C, Razick S, Rigina O, Salwinski L, Simonovic M, Velankar S, Winter A, Wu G, Bader GD, Cesareni G, Donaldson IM, Eisenberg D, Kleywegt GJ, Overington J, Ricard-Blum S, Tyers M, Albrecht M, Hermjakob H. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods*. 2011;8(7):528-529. doi:10.1038/nmeth.1637

134. del-Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J, Breuer K, Ono K, Salwinski L, Hermjakob H. A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res*. 2013;41(Web Server issue):W601-606. doi:10.1093/nar/gkt392

135. Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. *Cell*. 2006;124(4):783-801. doi:10.1016/j.cell.2006.02.015

136. Kuzmanov U, Emili A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Med*. 2013;5(4):37. doi:10.1186/gm441

137. Kamaladevi A, Marudhupandiyan S, Balamurugan K. Model system based proteomics to understand the host response during bacterial infections. *Mol Biosyst*. 2017;13(12):2489-2497. doi:10.1039/C7MB00372B

138. Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host-pathogen interactions. *Database J Biol Databases Curation*. 2016;2016. doi:10.1093/database/baw103

139. Nakajima N, Akutsu T, Nakato R. Databases for Protein–Protein Interactions. In: Cecconi D, ed. *Proteomics Data Analysis*. Springer US; 2021:229-248. doi:10.1007/978-1-0716-1641-3_14

140. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. *Int J Proteomics*. 2014;2014(1):147648. doi:10.1155/2014/147648

141. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*. 1999;17(10):1030-1032. doi:10.1038/13732

142. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, Robinson M, Raghibizadeh S, Hogue CWV, Bussey H, Andrews B, Tyers M, Boone C. Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science*. 2001;294(5550):2364-2368. doi:10.1126/science.1065810

143. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*. 2000;403(6770):623-627. doi:10.1038/35001009

144. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int J Mol Sci*. 2009;10(6):2763-2788. doi:10.3390/ijms10062763

145. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci*. 2001;98(8):4569-4574. doi:10.1073/pnas.061034498

146. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*. 2012;40(Database issue):D857-861. doi:10.1093/nar/gkr930

147. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*. 2023;51(D1):D638-D646. doi:10.1093/nar/gkac1000

148. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, Zhang F, Dolma S, Willems A, Coulombe-Huntington J, Chatr-aryamontri A, Dolinski K, Tyers M. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019;47(Database issue):D529-D541. doi:10.1093/nar/gky1079

149. Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E, Perfetto L, How K, Ratan P, Shirodkar G, Lu O, Mészáros B, Watkins X, Pundir S, Licata L, Iannuccelli M, Pellegrini M, Martin MJ, Panni S, Duesbury M, Vallet SD, Rappsilber J, Ricard-Blum S, Cesareni G, Salwinski L, Orchard S, Porras P, Panneerselvam K, Hermjakob H. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res*. 2022;50(D1):D648-D653. doi:10.1093/nar/gkab1006

150. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*. 2004;32(Database issue):D449-451. doi:10.1093/nar/gkh086

151. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human Protein Reference Database--2009 update. *Nucleic Acids Res*. 2009;37(Database issue):D767-772. doi:10.1093/nar/gkn892

152. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*. 2007;8(5):R95. doi:10.1186/gb-2007-8-5-r95

153. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, Hurrell L, Isserlin R, Jack F, Juma F, Khan A, Kon T, Konopinsky S, Le V, Lee E, Ling S, Magidin M, Moniakis J, Montojo J, Moore S, Muskat B, Ng I, Paraiso JP, Parker B, Pintilie G, Pirone R, Salama JJ, Sgro S, Shan T, Shu Y, Siew J, Skinner D, Snyder K, Stasiuk R, Strumpf D, Tuekam B, Tao S, Wang Z, White M, Willis R, Wolting C, Wong S, Wrong A, Xin C, Yao R, Yates B, Zhang S, Zheng K, Pawson T, Ouellette BFF, Hogue CWV. The Biomolecular

Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*. 2005;33(Database issue):D418-424. doi:10.1093/nar/gki051

154. Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stümpflen V. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*. 2006;34(Database issue):D436-441. doi:10.1093/nar/gkj003

155. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FSL, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock REW, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods*. 2012;9(4):345-350. doi:10.1038/nmeth.1931

156. Dey L, Mukhopadhyay A. DenvInt: A database of protein–protein interactions between dengue virus and its hosts. *PLoS Negl Trop Dis*. 2017;11(10):e0005879. doi:10.1371/journal.pntd.0005879

157. Cook HV, Doncheva NT, Szklarczyk D, von Mering C, Jensen LJ. Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. *Viruses*. 2018;10(10):519. doi:10.3390/v10100519

158. Dyer MD, Neff C, Dufford M, Rivera CG, Shattuck D, Bassaganya-Riera J, Murali TM, Sobral BW. The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. *PloS One*. 2010;5(8):e12089. doi:10.1371/journal.pone.0012089

159. Beatty ME, Ashford DA, Griffin PM, Tauxe RV, Sobel J. Gastrointestinal anthrax: review of the literature. *Arch Intern Med*. 2003;163(20):2527-2531. doi:10.1001/archinte.163.20.2527

160. Liu S, Moayeri M, Leppla SH. Anthrax lethal and edema toxins in anthrax pathogenesis. *Trends Microbiol*. 2014;22(6):317-325. doi:10.1016/j.tim.2014.02.012

161. Jernigan JA, Stephens DS, Ashford DA, Omenaca C, Topiel MS, Galbraith M, Tapper M, Fisk TL, Zaki S, Popovic T, Meyer RF, Quinn CP, Harper SA, Fridkin SK, Sejvar JJ, Shepard CW, McConnell M, Guarner J, Shieh WJ, Malecki JM, Gerberding JL, Hughes JM, Perkins BA, Anthrax Bioterrorism Investigation Team. Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States. *Emerg Infect Dis*. 2001;7(6):933-944. doi:10.3201/eid0706.010604

162. Fang H, Sun C, Xu L, Owen RJ, Auth RD, Snoy PJ, Frucht DM. Neutrophil Elastase Mediates Pathogenic Effects of Anthrax Lethal Toxin in the Murine Intestinal Tract. *J Immunol*. 2010;185(9):5463-5467. doi:10.4049/jimmunol.1002471

163. Bradley KA, Mogridge J, Mourez M, Collier RJ, Young JAT. Identification of the cellular receptor for anthrax toxin. *Nature*. 2001;414(6860):225-229. doi:10.1038/n35101999

164. Scobie HM, Rainey GJA, Bradley KA, Young JAT. Human capillary morphogenesis protein 2 functions as an anthrax toxin receptor. *Proc Natl Acad Sci*. 2003;100(9):5170-5174. doi:10.1073/pnas.0431098100

165. Martchenko M, Jeong SY, Cohen SN. Heterodimeric integrin complexes containing β1-integrin promote internalization and lethality of anthrax toxin. *Proc Natl Acad Sci*. 2010;107(35):15583-15588. doi:10.1073/pnas.1010145107

166. Abrami L, Liu S, Cosson P, Leppla SH, van der Goot FG. Anthrax toxin triggers endocytosis of its receptor via a lipid raft–mediated clathrin-dependent process. *J Cell Biol*. 2003;160(3):321-328. doi:10.1083/jcb.200211018

167. Hong J, Doebele RC, Lingen MW, Quilliam LA, Tang WJ, Rosner MR. Anthrax Edema Toxin Inhibits Endothelial Cell Chemotaxis via Epac and Rap1*. *J Biol Chem*. 2007;282(27):19781-19787. doi:10.1074/jbc.M700128200

168. Duesbery NS, Webb CP, Leppla SH, Gordon VM, Klimpel KR, Copeland TD, Ahn NG, Oskarsson MK, Fukasawa K, Paull KD, Vande Woude GF. Proteolytic Inactivation of MAP-Kinase-Kinase by Anthrax Lethal Factor. *Science*. 1998;280(5364):734-737. doi:10.1126/science.280.5364.734

169. During RL, Li W, Hao B, Koenig JM, Stephens DS, Quinn CP, Southwick FS. Anthrax Lethal Toxin Paralyzes Neutrophil Actin-Based Motility. *J Infect Dis*. 2005;192(5):837-845. doi:10.1086/432516

170. Xie T, Auth RD, Frucht DM. The Effects of Anthrax Lethal Toxin on Host Barrier Function. *Toxins*. 2011;3(6):591-607. doi:10.3390/toxins3060591

171. Lowe DE, Glomski IJ. Cellular and Physiological Effects of Anthrax Exotoxin and Its Relevance to Disease. *Front Cell Infect Microbiol*. 2012;2. doi:10.3389/fcimb.2012.00076

172. Nanda A, Carson-Walter EB, Seaman S, Barber TD, Stampfl J, Singh S, Vogelstein B, Kinzler KW, St. Croix B. TEM8 Interacts with the Cleaved C5 Domain of Collagen α3(VI). *Cancer Res*. 2004;64(3):817-820. doi:10.1158/0008-5472.CAN-03-2408

173. Young JAT, Collier RJ. Anthrax Toxin: Receptor Binding, Internalization, Pore Formation, and Translocation. *Annu Rev Biochem*. 2007;76(Volume 76, 2007):243-265. doi:10.1146/annurev.biochem.75.103004.142728

174. Moayeri M, Leppla SH. Cellular and systemic effects of anthrax lethal toxin and edema toxin. *Mol Aspects Med*. 2009;30(6):439-455. doi:10.1016/j.mam.2009.07.003

175. Sydow D, Burggraaff L, Szengel A, van Vlijmen HWT, IJzerman AP, van Westen GJP, Volkamer A. Advances and Challenges in Computational Target Prediction. *J Chem Inf Model*. 2019;59(5):1728-1742. doi:10.1021/acs.jcim.8b00832

176. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M. Identification of potential interaction networks using sequence-based searches for conserved

protein-protein interactions or “interologs.” *Genome Res*. 2001;11(12):2120-2126. doi:10.1101/gr.205301

177. Davis FP, Barkan DT, Eswar N, McKerrow JH, Sali A. Host pathogen protein interactions predicted by comparative modeling. *Protein Sci Publ Protein Soc*. 2007;16(12):2585-2596. doi:10.1110/ps.073228407

178. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein-protein interactions. *Bioinforma Oxf Engl*. 2007;23(13):i159-166. doi:10.1093/bioinformatics/btm208

179. Edwards RJ, Davey NE, Shields DC. SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PloS One*. 2007;2(10):e967. doi:10.1371/journal.pone.0000967

180. Sen R, Nayak L, De RK. A review on host–pathogen interactions: classification and prediction. *Eur J Clin Microbiol Infect Dis*. 2016;35(10):1581-1599. doi:10.1007/s10096-016-2716-7

181. Lian X, Yang S, Li H, Fu C, Zhang Z. Machine-Learning-Based Predictor of Human–Bacteria Protein–Protein Interactions by Incorporating Comprehensive Host-Network Properties. *J Proteome Res*. 2019;18(5):2195-2205. doi:10.1021/acs.jproteome.9b00074

182. Tyagi N, Krishnadev O, Srinivasan N. Prediction of protein-protein interactions between Helicobacter pylori and a human host. *Mol Biosyst*. 2009;5(12):1630-1635. doi:10.1039/b906543c

183. Krishnadev O, Srinivasan N. Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int J Biol Macromol*. 2011;48(4):613-619. doi:10.1016/j.ijbiomac.2011.01.030

184. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. *Nat Commun*. 2018;9(1):2312. doi:10.1038/s41467-018-04632-8

185. Chapter 5: Network Biology Approach to Complex Diseases | PLOS Computational Biology. Accessed July 25, 2024. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002820

186. Krishnadev O, Srinivasan N. A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite. *In Silico Biol*. 2008;8(3-4):235-250.

187. Lee SA, Chan C hsiung, Tsai CH, Lai JM, Wang FS, Kao CY, Huang CYF. Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics*. 2008;9 Suppl 12(Suppl 12):S11. doi:10.1186/1471-2105-9-S12-S11

188. Mishra V, Xavier JB. Unclouding Clostridiodes difficile virulence with systems biology. *Cell Host Microbe*. 2021;29(11):1608-1610. doi:10.1016/j.chom.2021.10.005

189. Basak S, Deb D, Narsaria U, Kar T, Castiglione F, Sanyal I, Bade PD, Srivastava AP. In silico designing of vaccine candidate against Clostridium difficile. *Sci Rep*. 2021;11(1):14215. doi:10.1038/s41598-021-93305-6

190. Donskey CJ. Update on Clostridioides difficile Infection in Older Adults. *Infect Dis Clin North Am*. 2023;37(1):87-102. doi:10.1016/j.idc.2022.10.001

191. Piccioni A, Rosa F, Manca F, Pignataro G, Zanza C, Savioli G, Covino M, Ojetti V, Gasbarrini A, Franceschi F, Candelli M. Gut Microbiota and Clostridium difficile: What We Know and the New Frontiers. *Int J Mol Sci*. 2022;23(21):13323. doi:10.3390/ijms232113323

192. Awad MM, Johanesen PA, Carter GP, Rose E, Lyras D. Clostridium difficile virulence factors: Insights into an anaerobic spore-forming pathogen. *Gut Microbes*. 2014;5(5):579-593. doi:10.4161/19490976.2014.969632

193. Sarkar S, Heise MT. Mouse Models as Resources for Studying Infectious Diseases. *Clin Ther*. 2019;41(10):1912-1922. doi:10.1016/j.clinthera.2019.08.010

194. Yu H. Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. *Genome Res*. 2004;14(6):1107-1118. doi:10.1101/gr.1774904

195. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database J Biol Databases Curation*. 2011;2011:bar030. doi:10.1093/database/bar030

196. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

197. Qin H. hongqin/Simple-reciprocal-best-blast-hit-pairs. Published online March 8, 2017. Accessed September 18, 2023. https://github.com/hongqin/Simple-reciprocal-best-blast-hit-pairs

198. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert

R, Hansen RS, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014;515(7527):355-364. doi:10.1038/nature13992

199. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

200. Wang J, Zhong J, Chen G, Li M, Wu F xiang, Pan Y. ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12(4):815-822. doi:10.1109/TCBB.2014.2361348

201. Abt MC, McKenney PT, Pamer EG. Clostridium difficile colitis: pathogenesis and host defence. *Nat Rev Microbiol*. 2016;14(10):609-620. doi:10.1038/nrmicro.2016.108

202. Monot M, Boursaux-Eude C, Thibonnier M, Vallenet D, Moszer I, Medigue C, Martin-Verstraete I, Dupuy B. Reannotation of the genome sequence of Clostridium difficile strain 630. *J Med Microbiol*. 2011;60(Pt 8):1193-1199. doi:10.1099/jmm.0.030452-0

203. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2023;51(D1):D523-D531. doi:10.1093/nar/gkac1052

204. Caohuy H, Srivastava M, Pollard HB. Membrane fusion protein synexin (annexin VII) as a Ca2+/GTP sensor in exocytotic secretion. *Proc Natl Acad Sci*. 1996;93(20):10797-10802. doi:10.1073/pnas.93.20.10797

205. Ishihara S, Nishikimi A, Umemoto E, Miyasaka M, Saegusa M, Katagiri K. Dual functions of Rap1 are crucial for T-cell homeostasis and prevention of spontaneous colitis. *Nat Commun*. 2015;6(1):8982. doi:10.1038/ncomms9982

206. Rettner RE, Saier MH. The autoinducer-2 exporter superfamily. *J Mol Microbiol Biotechnol*. 2010;18(4):195-205. doi:10.1159/000316420

207. Li X, Fan X, Shi Z, Xu J, Cao Y, Zhang T, Pan D. AI-2E Family Transporter Protein in Lactobacillus acidophilus Exhibits AI-2 Exporter Activity and Relate With Intestinal Juice Resistance of the Strain. *Front Microbiol*. 2022;13:908145. doi:10.3389/fmicb.2022.908145

208. Nairz M, Metzendorf C, Vujic-Spasic M, Mitterstiller AM, Schroll A, Haschka D, Hoffmann A, von Raffay L, Sparla R, Huck CW, Talasz H, Moser PL, Muckenthaler MU, Weiss G. Cell-specific expression of Hfe determines the outcome of Salmonella enterica serovar Typhimurium infection in mice. *Haematologica*. 2020;106(12):3149-3161. doi:10.3324/haematol.2019.241745

209. Schröder B. The multifaceted roles of the invariant chain CD74--More than just a chaperone. *Biochim Biophys Acta*. 2016;1863(6 Pt A):1269-1281. doi:10.1016/j.bbamcr.2016.03.026

210. Buschiazzo A, Trajtenberg F. Two-Component Sensing and Regulation: How Do Histidine Kinases Talk with Response Regulators at the Molecular Level? *Annu Rev Microbiol*. 2019;73:507-528. doi:10.1146/annurev-micro-091018-054627

211. Qi H, Wei J, Gao Y, Yang Y, Li Y, Zhu H, Su L, Su X, Zhang Y, Yang R. Reg4 and complement factor D prevent the overgrowth of E. coli in the mouse gut. *Commun Biol*. 2020;3(1):483. doi:10.1038/s42003-020-01219-2

212. Sekine H, Machida T, Fujita T. Factor D. *Immunol Rev*. 2023;313(1):15-24. doi:10.1111/imr.13155

213. McLean KC, Oppenheimer KH, Sweet LM, Phillippe M. Phospholipid scramblase expression in the pregnant mouse uterus in LPS-induced preterm delivery. *Reprod Sci Thousand Oaks Calif*. 2012;19(11):1211-1218. doi:10.1177/1933719112446078

214. Silbergleit M, Vasquez AA, Miller CJ, Sun J, Kato I. Oral and intestinal bacterial exotoxins: Potential linked to carcinogenesis. *Prog Mol Biol Transl Sci*. 2020;171:131-193. doi:10.1016/bs.pmbts.2020.02.004

215. Xiang Z, Li J, Song S, Wang J, Cai W, Hu W, Ji J, Zhu Z, Zang L, Yan R, Yu Y. A positive feedback between IDO1 metabolite and COL12A1 via MAPK pathway to promote gastric cancer metastasis. *J Exp Clin Cancer Res CR*. 2019;38(1):314. doi:10.1186/s13046-019-1318-5

216. Sheng Y, Wu L, Chang Y, Liu W, Tao M, Chen X, Zhang X, Li B, Zhang N, Ye D, Zhang C, Zhu D, Zhao H, Chen A, Chen H, Song J. Tomo-seq identifies NINJ1 as a potential target for anti-inflammatory strategy in thoracic aortic dissection. *BMC Med*. 2023;21(1):396. doi:10.1186/s12916-023-03077-1

217. Jennewein C, Sowa R, Faber AC, Dildey M, von Knethen A, Meybohm P, Scheller B, Dröse S, Zacharowski K. Contribution of Ninjurin1 to Toll-like receptor 4 signaling and systemic inflammation. *Am J Respir Cell Mol Biol*. 2015;53(5):656-663. doi:10.1165/rcmb.2014-0354OC

218. Keegan AD, Leonard WJ, Zhu J. Recent advances in understanding the role of IL-4 signaling. *Fac Rev*. 2021;10:71. doi:10.12703/r/10-71

219. Sato-Nishiuchi R, Nakano I, Ozawa A, Sato Y, Takeichi M, Kiyozumi D, Yamazaki K, Yasunaga T, Futaki S, Sekiguchi K. Polydom/SVEP1 is a ligand for integrin α9β1. *J Biol Chem*. 2012;287(30):25615-25630. doi:10.1074/jbc.M112.355016

220. Sanford JL, Mays TA, Rafael-Fortney JA. CASK and Dlg form a PDZ protein complex at the mammalian neuromuscular junction. *Muscle Nerve*. 2004;30(2):164-171. doi:10.1002/mus.20073

221. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. *Nat Rev Genet*. 2017;18(7):425-440. doi:10.1038/nrg.2017.19

222. Foster LJ. The adult mouse proteome. *Nat Methods*. 2022;19(7):792-793. doi:10.1038/s41592-022-01546-8

223. Rehli M. Of mice and men: species variations of Toll-like receptor expression. *Trends Immunol*. 2002;23(8):375-378. doi:10.1016/s1471-4906(02)02259-7

224. Bower WA, Hendricks KA, Vieira AR, Traxler RM, Weiner Z, Lynfield R, Hoffmaster A. What Is Anthrax? *Pathogens*. 2022;11(6):690. doi:10.3390/pathogens11060690

225. Mamareli P, Kruse F, Friedrich C, Smit N, Strowig T, Sparwasser T, Lochner M. Epithelium-specific MyD88 signaling, but not DCs or macrophages, control acute intestinal infection with Clostridium difficile. *Eur J Immunol*. 2019;49(5):747-757. doi:10.1002/eji.201848022

226. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305(3):567-580. doi:10.1006/jmbi.2000.4315

227. Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, Tettelin H, Fouts DE, Eisen JA, Gill SR, Holtzapple EK, Okstad OA, Helgason E, Rilstone J, Wu M, Kolonay JF, Beanan MJ, Dodson RJ, Brinkac LM, Gwinn M, DeBoy RT, Madpu R, Daugherty SC, Durkin AS, Haft DH, Nelson WC, Peterson JD, Pop M, Khouri HM, Radune D, Benton JL, Mahamoud Y, Jiang L, Hance IR, Weidman JF, Berry KJ, Plaut RD, Wolf AM, Watkins KL, Nierman WC, Hazen A, Cline R, Redmond C, Thwaite JE, White O, Salzberg SL, Thomason B, Friedlander AM, Koehler TM, Hanna PC, Kolstø AB, Fraser CM. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature*. 2003;423(6935):81-86. doi:10.1038/nature01586

228. Batah J, Kansau I. Intestinal Epithelial Cell Response to Clostridium difficile Flagella. In: Roberts AP, Mullany P, eds. *Clostridium Difficile: Methods and Protocols*. Springer; 2016:103-116. doi:10.1007/978-1-4939-6361-4_8

229. Rahmatbakhsh M, Moutaoufik MT, Gagarinova A, Babu M. HPiP: an R/Bioconductor package for predicting host–pathogen protein–protein interactions from protein sequences using ensemble machine learning approach. *Bioinforma Adv*. 2022;2(1):vbac038. doi:10.1093/bioadv/vbac038

230. Kayagaki N, Kornfeld OS, Lee BL, Stowe IB, O'Rourke K, Li Q, Sandoval W, Yan D, Kang J, Xu M, Zhang J, Lee WP, McKenzie BS, Ulas G, Payandeh J, Roose-Girma M, Modrusan Z, Reja R, Sagolla M, Webster JD, Cho V, Andrews TD, Morris LX, Miosge LA, Goodnow CC, Bertram EM, Dixit VM. NINJ1 mediates plasma membrane rupture during lytic cell death. *Nature*. 2021;591(7848):131-136. doi:10.1038/s41586-021-03218-7

231. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinforma Oxf Engl*. 2009;25(22):3045-3046. doi:10.1093/bioinformatics/btp536

232. Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, UniProt Consortium T. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*. 2017;33(21):3454-3460. doi:10.1093/bioinformatics/btx439

233. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*. 2022;23(1):40-55. doi:10.1038/s41580-021-00407-0

234. Yakimovich A. Machine Learning and Artificial Intelligence for the Prediction of Host–Pathogen Interactions: A Viral Case. *Infect Drug Resist*. 2021;14:3319-3326. doi:10.2147/IDR.S292743

235. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. ; 1995:278-282 vol.1. doi:10.1109/ICDAR.1995.598994

236. Crammer K, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines.

237. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983-987. doi:10.1038/nbt.4235

238. Zheng N, Wang K, Zhan W, Deng L. Targeting Virus-host Protein Interactions: Feature Extraction and Machine Learning Approaches. *Curr Drug Metab*. 2019;20(3):177-184. doi:10.2174/1389200219666180829121038

239. Liu-Wei W, Kafkas Ş, Chen J, Dimonaco NJ, Tegnér J, Hoehndorf R. DeepViral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes. *Bioinforma Oxf Engl*. 2021;37(17):2722-2729. doi:10.1093/bioinformatics/btab147

240. Eid FE, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein–protein interaction prediction. *Bioinformatics*. 2016;32(8):1144-1150. doi:10.1093/bioinformatics/btv737

241. Kshirsagar M, Carbonell J, Klein-Seetharaman J. Multitask learning for host-pathogen protein interactions. *Bioinforma Oxf Engl*. 2013;29(13):i217-226. doi:10.1093/bioinformatics/btt245

242. Kaundal R, Loaiza CD, Duhan N, Flann N. deepHPI: a comprehensive deep learning platform for accurate prediction and visualization of host–pathogen protein–protein interactions. *Brief Bioinform*. 2022;23(3):bbac125. doi:10.1093/bib/bbac125

243. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

244. Yang X, Yang S, Ren P, Wuchty S, Zhang Z. Deep Learning-Powered Prediction of Human-Virus Protein-Protein Interactions. *Front Microbiol*. 2022;13. doi:10.3389/fmicb.2022.842976

245. Dong TN, Brogden G, Gerold G, Khosla M. A multitask transfer learning framework for the prediction of virus-human protein–protein interactions. *BMC Bioinformatics*. 2021;22(1):572. doi:10.1186/s12859-021-04484-y

246. Lin JS, Lai EM. Protein–Protein Interactions: Co-Immunoprecipitation. In: Journet L, Cascales E, eds. *Bacterial Protein Secretion Systems*. Vol 1615. Methods in Molecular Biology. Springer New York; 2017:211-219. doi:10.1007/978-1-4939-7033-9_17

247. Mei S, Zhang K. Neglog: Homology-Based Negative Data Sampling Method for Genome-Scale Reconstruction of Human Protein–Protein Interaction Networks. *Int J Mol Sci*. 2019;20(20):5075. doi:10.3390/ijms20205075