Construction of a Metabolic Model for the Human Pathogen *Candida parapsilosis* and Early Identification of Putative Novel Anti-Fungal Drug Targets

Diogo Alexandre Carracedo Couceiro

Thesis to Obtain the Master of Science Degree in

Microbiology

Supervisor: Prof. Dr. Miguel Nobre Parreira Cacho Teixeira

Examination Committee

Chairperson: Prof. Dr. Jorge Humberto Gomes Leitão

Supervisor: Prof. Dr. Miguel Nobre Parreira Cacho Teixeira

Member of the committee: Prof Dr. Nuno Gonçalo Pereira Mira

October 2021

Preface

The work presented in this thesis was performed at the Institute for Bioengineering and Biosciences of Instituto superior Técnico (Lisbon, Portugal), during the period September 2020 – October 2021, under the supervision of Prof. Dr. Miguel Nobre Parreira Cacho Teixeira.

This work was supported by “Fundação para a Ciência e a Tecnologia” (FCT) (Contract PTDC/BII-BIO/28216/2017) and Programa Operacional Regional de Lisboa 2020 (contract LISBOA-01-0145-FEDER-022231, the BioData.pt Research Infrastructure).

Funding received by iBB-Institute for Bioengineering and Biosciences from FCT (UIDB/04565/2020) is also acknowledged.

Declaration

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Abstract

Candida parapsilosis has seen one of the most significant rises in incidence among pathogenic *Candida* spp., often taking second place only to *C. albicans*. Adding to this increased incidence is the rise in resistance to first line antifungals and lack of adequate alternative therapeutics, not only for *C. parapsilosis* but throughout the genus. Genome Scale Metabolic Models (GSMMs) have risen as a powerful *in silico* tool for the understanding of pathogenesis due to their systems view of metabolism and, above all, drug target predictive capacity. In this project a metabolic model for the human pathogen *C. parapsilosis* was constructed – *Cpar*MM – comprising 1112 genes, 598 proteins, 2892 reactions and 2885 metabolites across four compartments. Upon extensive manual curation this model was experimentally validated and proven to quantitatively predict biomass production, as well as predict *C. parapsilosis'* ability to use several metabolites as sole carbon or nitrogen sources. From the resulting validated model, a list of predicted essential genes common to other major pathogenic *Candida* spp. in mimicked host conditions was obtained. Among these predicted drug targets, 18 were found to be entirely new, for which no previous inhibitor was ever assigned. Additionally, Fol1, Abz1, Cab1 and Cab5 seem to be rather promising putative novel drug targets, as they represent the possibility of inducing systemic metabolic impairment by targeting central metabolism.

Key Words: *C. parapsilosis*; Genome Scale Metabolic Model; Resistance; Antifungals; Fungaemia.

Resumo

Candida parapsilosis tem demonstrado o aumento de incidência mais significativo dentro do género, suplantando *C. albicans* nalgumas regiões. Para além de fatores de risco variados, uma panóplia de fatores de virulência ainda por compreender no seu todo aliados ao aumento da resistência antifúngica acrescem à patogenicidade de *C. parapsilosis*. Urge então desenvolver novas alternativas terapêuticas e ferramentas de investigação adequadas. Os Modelos Metabólicos Baseados no Genoma (GSMMs) têm-se demonstrado uma poderosa ferramenta *in silico* não só para a compreensão sistémica do metabolismo, mas acima de tudo pelas suas capacidades preditivas de alvos de drogas.

Deste projeto resultou um modelo metabólico para o patógeno humano *C. parapsilosis* – *Cpar*MM – que se compõe de 1112 genes, 598 proteínas, 2892 reações e 2885 metabolitos distribuídos por quatro compartimentos. Após curadoria manual extensa este modelo foi experimentalmente validado. Daqui, mostrou-se capaz não só de prever de forma quantitativa a produção de biomassa, mas também a capacidade de *C. parapsilosis* usar determinados compostos como fontes únicas de carbono ou azoto. Do modelo validado obteve-se uma lista de essencialidade em condições miméticas do hospedeiro, tendo resultado num total de 18 possíveis alvos inteiramente novos – sem qualquer inibidor previamente descrito. Em particular, as proteínas Fol1, Abz1, Cab1 e Cab5, apresentam-se como de relevado interesse enquanto possíveis novos alvos de drogas por representarem a possibilidade de inibição metabólica sistémica alvejando o metabolismo central.

Palavras-chave: *C. parapsilosis*; Modelos Metabólicos Baseados no Genoma; Resistência; Antifúngicos; Fungémia.

Acknowledgements

To Professor Miguel, for giving me the opportunity to work on this project and for trusting in my abilities. For never allowing the short comings that come with learning make me feel inadequate. For his always constructive and candid criticism, time and encouragement.

To Romeu for his never-ending patience, from whom I learnt an invaluable amount of practical knowledge that I will carry for the rest of my carrier.

To my parents who through enormous difficulties never let anything be amiss and without whom I would not be who I am.

To my cousin Sofia, in whom I saw from early on an example of focus and humanity in academia. Whom through her practiced example and always kind advice guided me at a crucial moment in my path.

To my friends, whose love assured I kept going and reminded me of how ephemeral adversity can be.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Thesis Outline and Objective

The changing climate and social reality of our times calls for the development of new strategies and tools for the study of disease. The past few decades have seen a rise in antimicrobial resistance and change in the incidence dynamics of microbial disease particularly. *Candida* spp. have been a clear example of this phenomenon, with a rise in incidence since the 1980's worldwide and a worrying subversion of historic trends. While *C. albicans* is still the most frequently isolated, other Non-*Candida albicans Candida* Species (NCAC species) - particularly *Candida parapsilosis* - have seen a significant rise in their incidence due to a varied range of factors. The most likely causes being the widespread and persistent use of antifungals, consequent rise in resistance and differing resistance within the genus. There is thus a need to develop new time and cost-effective tools for the identification of new drug targets capable of combining the vast amounts of biological information generated over decades of research. Genome Scale Metabolic Models (GSMMs) have surged as a response to this issue and have proven their accuracy and applicability in biomedical research in the little more than 20 years since their appearance.

It is in this context that this project aims at the construction of a GSMM for the human pathogen *C. parapsilosis*. Ultimately, this model's construction will allow for the surpassing of current challenges not only related to resistance but, for instance, in permitting the implementation of paradigm changes in drug finding - notions such as multiple although weaker targeting possibly being more efficient than full single target inhibition.

The following "Introduction" describes the epidemiological context of fungal and particularly *Candida* spp. infections – from incidence dynamics to risk factors -, characterizes *C. parapsilosis* as a rising human pathogen and concludes with an overview of what are GSMMs, their rationale and applicability.

This Introduction is followed by a "Materials and Methods" chapter where the workflow of this model's construction is schematised and divided into its main stages, each duly described. In the subsequent chapter "Results and Discussion" issues and solutions behind the construction of this model are rationalised and illustrated - some of which via practical examples showing the manual curation steps required for the model's optimization –, and experimental results pertaining to model validation are presented and discussed. This section concludes with the identification and discussion of the predicted putative novel anti-fungal drug targets resulting from this model's exploitation.

Finally, in "Conclusions and Perspectives" this work's main achievements are highlighted and the model's limitations and how they can be surpassed are discussed – from the promising set of identified putative targets to the next great stage of model curation.

1.2 Epidemiology

1.2.1 Fungaemia as a Growing Problem

Humankind's relationship with disease has fundamentally changed in the past few decades. Disease is no longer faced as an inevitable reality to be endured but as an explainable, manageable and above all, surpassable aspect of our lived experience. Nonetheless, this paradigm change is not without its downsides. Moreover, in a world of globalization and climate change, of habitat destruction and species' mass migration, it is inevitable that new interspecies contacts will occur and new species barriers transposed. New pandemics will arise and the need for the development not only of adequate therapeutics but also research tools has, possibly, never been stronger.

With each pandemic, new possibilities for disease surface. Be it due to the more immediate fragility of diseased individuals themselves or to lasting effects on their health. The last 40 years saw the dawn of the still on-going HIV-AIDS pandemic and alongside it the rise in immunocompromised individuals, ever more susceptible to infection [1–4]. In addition, advancements in modern medicine allow for aging populations and extended lifespans of individuals with underlying conditions [1,5,6]. Immunosuppressive therapies, invasive medical procedures - from parenteral feeding to prosthetics - and widespread use of antimicrobials all contributing to increased susceptibility [5–7].

In particular, since the 1980's fungal infections have significantly increased in number [1,8,9]. This increase relates to both nosocomial and community acquired infections that can range from superficial to systemic, such as sepsis and related illnesses like endocarditis, for instance [8,10]. The etiological agents behind this increase are, as fungi themselves, vast and varied. From the major *Aspergillus, Cryptococcus* and, above all, *Candida* spp. to hyaline moulds and non-*Candida* yeasts, such as *Trichosporon* spp. [11,12]. Even so, *Candida* spp. are the most frequently isolated from fungal infection cases worldwide [1,6,11–13].

Epidemiological data for *Candida* spp. infections in Europe are vast but mostly country specific[14]. Nonetheless, meta-analysis of the existing epidemiological data by the European Confederation of Medical Mycology (ECMM) had already observed an increase in *Candida* spp. bloodstream infections (candidemia) by the early 2000's, in particular among hospitalized patients [15]. As of 2019 this general increase was still being observed in hospitalized patients, with incidence estimates of 79 cases per day and fatal outcomes for 29 patients after 30 days of admission [14].

In the United States alone, between 1980 and 1990 Candidemia increased five-fold and by the late 1990's *Candida* spp. accounted for 8 to 10% of nosocomial blood stream infections [4,9,16]. As reported by the American Centre for Disease Control (CDC), between 2009 and 2013 candidemia cases seem to have stabilized overall, although with significant geographical variability, being that decreases in one region occur alongside increases in another [17].

It becomes then apparent that candidemia surges encompassed by a complex multi-variable global phenomenon of increased fungaemia. Not only this, it itself emerges as a dynamic problem with non-negligible geographical variability and moulded by differences in medical practices, host factors and

changing dynamics between different species within the genus, each with specific clinical manifestations[1,5,11,12].

1.2.2 *Candida* spp. as Human Pathogens

Candida is an heterogeneous genus comprising around 200 species of which only a minority has been identified as human pathogens, opportunistic in nature [1,13]. These are environmental yeasts as well as common human commensals, colonizing the gut, the oral cavity and the genitourinary tract[2,3,6,9]. The origin of infection can be both environmental or endogenous, though the latter are relatively more frequent [6]. Infections by *Candida* spp. range from superficial, affecting the skin, nails or the mucosae – like vulvovaginal or oropharyngeal candidiasis – to serious systemic infections - like sepsis and invasive candidiasis affecting major organs and leading, for example, to endocarditis, osteomyelitis or infection of the central nervous system [1,2,4,6,13]. In reality, invasive candida infections have seen a worrying increase, particularly among hospitalized patients both in the United States and foremost in Europe [6,11].

The scope of risk factors for developing *Candida* spp. infections is broad and can be parted into medical care and host-related, both intrinsically linked. The most relevant host-related factor is a growing population with immunocompromise, be it due to underlying disease or advanced age [1,4,11,18]. Medical care related risk factors include invasive medical procedures such as surgery, indwelling medical devices such as prosthetic devices and catheters, immunosuppressive therapy, chemotherapy, haemodialysis, and not least of all, widespread use of antifungals [1,4–7,11,13]. The latter has surged as one of the most relevant since itself has likely changed the dynamics of species incidence within the genus[1,12,13].

Of the known pathogens of this genus, *Candida albicans* is the most frequently isolated, accounting for more than half the cases of invasive candidemia [1,4,6,12,13]. In fact, in the early 20th century *C. albicans* was considered the only pathogen of the known *Candida* spp., possibly relating to its relatively higher frequency of isolation and lack of suitable diagnosis tools for the discrimination of *Candida* spp. [3]. Fluconazole has become widely used as an efficient treatment for systemic *Candida* spp. infections as well as for prophylaxis [5,12,19]. Among other factors, this widespread use allied with heterogenous susceptibility to these antifungals within the genus led to a shift of incidence from *C. albicans* to other NCAC species [1,5,9,20].

1.2.3 Emergence of NCAC Species Infections

The 1990's saw a shift in incidence from *C. albicans* to other NCAC species [5,11,15,21]. Several factors might be behind this change. It is possible that advancements in molecular methods that allow for identification at species level may have led to the further differentiation of new species [1,11]. One example is *C. parapsilosis* which as recently as 2005 consisted of three groups, now differentiated into distinct species [21]. On the other hand, a growing immunocompromised population may have allowed for the emergence of previously non-pathogenic species as opportunistic pathogens [11]. The dominant

factor, though, is likely to be the widespread use of antifungals. In the 1990's fluconazole became widely used not only as effective treatment for infection as well as for prophylaxis in patient groups with added propensity for *Candida* spp. infections [5,20]. Heterogeneous susceptibility within the genus has likely led to the seen shift in incidence [1,12,17,19,22,23].

Although *C. albicans* remains the dominant isolate from invasive infections, *C parapsilosis*, *C. glabrata*, *C. tropicalis* and *C. krusei* have increasingly become more frequent, all five accounting for around 90% of all *Candida* spp. infections[1,6,11,13,15] – figure 1.

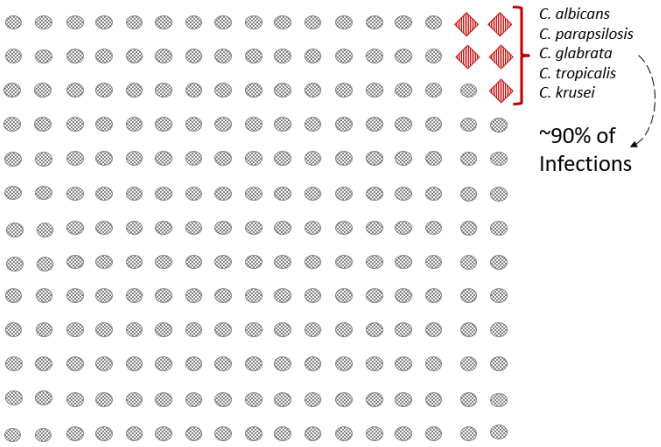

Figure 1: From the around 200 Candida spp. only a small minority have been described as pathogenic. From these, a set of only five account for around 90% of infections – *C. albicans*, *C. parapsilosis*, *C. glabrata*, *C. tropicalis* and *C. krusei.*

The ECMM reported on species distribution in *Candida* spp. infections comprehending the period between September 1997 and December 1999 and involving data from six European national societies – UK, Italy, Spain, Sweden, France and German speaking countries. On this report, *C. albicans* was isolated from 56% of the cases of infection, *C. glabrata* and *C. parapsilosis* from 14% followed by *C. tropicalis* and *C. krusei* isolated from 7 and 2% respectively [6,15]. Nonetheless, these distributions vary significantly with geographic location and above all, with underlying conditions and age. Although *C. albicans* did remain the most frequent isolate, it ranged from 67% in Sweden and German-speaking countries, to 43% in Spain where *C. parapsilosis* accounted for 30% [15] – figure 2. This is consistent with other reports that also indicate higher incidences of *C. parapsilosis* comparatively to other NCAC, even outranking *C. albicans* in some European countries [8,12,24]. In 2007, a Portuguese survey on the frequency of different *Candida* spp. isolates at one of the largest hospitals in the country, Hospital de São João, Porto, had already identified *C. parapsilosis* as the second most frequently isolated (25.6 %), surpassed only by *C. albicans* (35%) [25]. This was still the case between 2010 and 2011 as reported by a multicentre survey involving 10 Portuguese district hospitals and spanning 24 months, in which *C. parapsilosis* comprised 23% of the clinical isolates [26,27]. Besides Europe, this relative increase has also been seen in hospitalized patients in Canada [9]. In Latin American countries and Asia *C. parapsilosis* is the second most frequent isolate from bloodstream infection cases [21].

Other surveillance programs including several countries pertaining to the period between 1997 and 2003 have described similar changes, reporting an increase in *C. parapsilosis* infections, from 4.6% to 7.3% - 1.1 percentual points, the largest increase among the main four NCAC pathogenic species [11]. As of 2008 *C. parapsilosis* was still being reported as the second most frequent isolate from blood cultures [8].

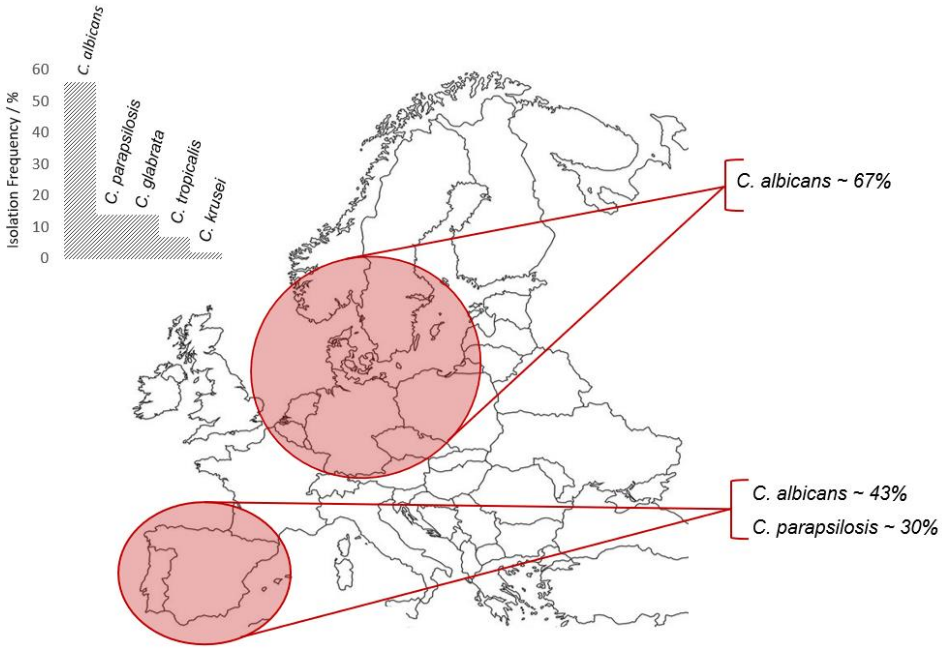

Figure 2: Species distribution within *Candida* is geographically variable. Although *Candida albicans* remains overall the most common isolate – left upper corner graph -, *Candida parapsilosis* has seen a significant rise in its incidence, occupying second place only to *C. albicans* in some European countries. While in Northern Europe *C. albicans* represents around 67% of isolates, in Southern Europe, particularly in Spain, this frequency decreases to 43% with *C. parapsilosis* accounting for 30% of isolates. Percentages according to [15]. Base border map retrieved from [28].

Within the greater context of the variability of incidence in *Candida*, *C. parapsilosis* has increasingly become a significant etiological agent particularly in hospital environments [8,29–31]. Apparently not geographically restricted, with particularities regarding its clinical manifestations and in some contexts subverting historical trends in incidence, there is a strong need to understand and research the biology of *C. parapsilosis*.

1.3 Characterizing *Candida parapsilosis*

In 1928 an American military doctor, Bailey Ashford, first isolated from the stool of a patient with diarrhoea what he then identified as *Monilia parapsilosis*, a distinct species from the more frequent *Monilia psilosis*, now *C. albicans* [32]. In this report *C. parapsilosis* was already proposed as a possible etiological agent for disease although it was only in the 1940's that other case reports definitely established it as a pathogenic organism [1,8,33].

Prior to 2005, *C. parapsilosis* consisted of 3 groups (Group I, II and III), differentiated based on Randomly Amplified Polymorphic DNA (RAPD) profiles. Tavanti *et al.* were able to describe consistent differences in the sequences of three genes allowing for the definition of the three groups as distinct species. Group I maintained the former species nomenclature, *C. parapsilosis*, while Group II and III were named *Candida orthopsilosis* and *Candida metapsilosis* [21]. Nonetheless, among these three species *C. parapsilosis* (previously Group I) is relatively more widely spread geographically and has the higher frequency of isolation [8,21]. This aspect allied to its close evolutionary relation to *C. orthopsilosis* and *C. metapsilosis* highlights the specifically human relationship of *C. parapsilosis*. In fact, although *C. parapsilosis* has been isolated from environmental sources, it is a common human commensal of the skin and is the most common of the fungal isolates from healthy individuals' hands [8,33–35].

1.3.1 Prevalence and Risk Factors

Being a common human commensal, *C. parapsilosis'* pathogenicity has been asserted as being dependent on lesions in the skin and membranes [8,33]. Its entry and further establishment of disease is dependent on invasive phenomena like catheterization, intubation, syringes, among others, on already compromised individuals [8,9].

In fact, *C. parapsilosis* was first established as pathogenic when identified as the cause of fatal endocarditis in several intravenous drug users in the 1940's [33]. It was then found that the cases related to a possibly contaminated batch of heroin and that consequently external introduction of the pathogen was a strong causative factor for infection. Later on, in similar circumstances, *C. parapsilosis* was isolated from the syringes used by a drug user in another case of endocarditis.

On the other hand, several studies particularly among hospitalized patients have taken into account the removal of the catheter as a therapeutical approach in cases of blood stream infection by *C. parapsilosis*. Girmenia *et al.* reported on patients at the Institute of Haematology of the university of La Sapienza in Rome [22]. Of the catheterized patients that developed blood stream infections, the removal of the catheter resulted in 79% presenting negative blood cultures after 24h. Similarly, Almirante *et al.* reported on patients in Barcelona and described intravenous lines as concomitant to 97% of infection cases[9].

Prior colonization seems to be behind some infection cases. Nonetheless, horizontal transmission has been reported as a major factor in infection [8,9]. In the case of hospitalized patients with catheter related infections, contamination is then due either to the patient being colonized or to hand carriage by healthcare workers, the latter being a major transmission vector [8,34,35]. For instance, Lupetti *et al.* reported on a nosocomial *C. parapsilosis* infection at the Ospedale Santa Chiara's neonate intensive care unit in Pisa [36]. At the time of admission, the infant was assessed *C. parapsilosis* negative via conjunctival, oropharyngeal and nasal swabs, urine and stool sampling. This infant later developed severe blood stream infection and the *C. parapsilosis* isolates were assessed to be genetically indistinguishable from those of two nurses caring for the infant. It might have been that the nurses were naturally colonized, or that this infection may have been a result of cross contamination from other

neonates. Nonetheless, vertical transmission is not a probable cause for primary colonization of the infants since *C. parapsilosis* is not a common isolate from the vaginal tract [8,33].

Prevalence of *C. parapsilosis* infections has been shown to vary among age groups but neonates and small infants are the patient groups at greatest risk, with increased rates of infection in neonate intensive care units [8,9,29,31]. This may be due to several aspects, ranging from gestational age, low and extremely low-birth weight to, mainly, extended catheterization, intubation and immune naivety[8,31].

Previous antifungal therapy also seems to influence *C. parapsilosis*' incidence with reports having described the emergence of different *Candida* spp. causing concomitant infection in patients that underwent different antifungal therapies. Cuervo *et al.* described *C. parapsilosis* as the cause of breakthrough candidemia in patients under echinocandin therapy – breakthrough candidemia being defined as candidemia developed whilst under antifungal therapy [18]. Additionally, Almirante *et al.* reported on a Barcelona based population in which patients developing blood stream infections by *C. parapsilosis*, when compared to *C. albicans*, had more commonly been under antifungal therapy – 26% and 7% respectively [9].

Other risk factors are more broad and equally valid for most infections, including open surgery, indwelling medical devices such as prosthetics, and immunosuppressive therapy.

1.3.2 Pathogenicity and Virulence Factors

Virulence factors are defined as the set of characteristics that confer a given organism its pathogenicity [1]. *C. parapsilosis* relies on several such virulence factors to evade the host's immune system and establish infection, from adherence, biofilm formation and extracellular enzyme release to modulating its interaction with the host's immune system via its cell surface morphology [1,8,29,37,38]. Differences in the regulation, structure and chemical nature of these factors are what confer specificity in pathogenicity to all pathogenic *Candida* spp. and to *C. parapsilosis* in particular. All of these are putative novel therapeutic targets and understanding how they work is crucial to shed light on how their biosynthetic and regulatory processes may be disrupted and disease fought.

Adherence and biofilm formation are two intrinsically connected processes. The spread and establishment of infection depend on *C. parapsilosis'* ability to adhere to both abiotic and biotic surfaces. Within the host, adherence is one of the initial steps in the establishment of infection, the latter being dependent on *C. parapsilosis'* cells ability to adhere to the host's tissues [1,8,34]. Catheterization and other similar invasive procedures, as previously mentioned, are major risk factors for *C. parapsilosis* infection. This results from *C. parapsilosis*' ability to adhere to abiotic surfaces such as those of medical devices with subsequent biofilm formation, adding to the relevance of adherence as a virulence factor [1,8,39]. Moreover, relative to *C. albicans*, *C. parapsilosis* has been shown to have higher rates of adhesion to acrylic which may explain the specificity of catheterization as a risk factor [8].

Biofilms are structured and highly regulated communities of cells adhered to a given surface, conferring antifungal resistance by reducing substance penetration and providing protection from the

host's immune response [1,22,37,40]. Thus, biofilms are a quite clinically significant virulence factor with estimates of 65% of microbial disease involving some form of biofilm [39]. The extent of biofilm formation varies among *Candida* spp., which might relate to differences in their pathogenicity [37,40]. In fact, *C. parapsilosis* has been shown to have a comparatively greater propensity to form biofilms compared to other *Candida* spp.[41]. Nejii *et al.* assessed relative biofilm formation among several isolates from *C. parapsilosis* and the closely related and less clinically significant species, *C. orthopsilosis* and *C. metapsilosis*, showing that the former has the highest extent of biofilm formation [37]. One additional and relevant find was the significant *C. parapsilosis* strain dependent variability in biofilm formation. Other studies have also reported this strain dependent variability [40]. This aspect may be relevant in terms of assessing the most clinically significant sources of pathogenic strains, given their respective strains' particular physiological properties. Furthermore, although *C. albicans*' biofilms are comparatively more complex, *C. parapsilosis* blood isolates have been described as consistently forming biofilms, adding to biofilms having a role in virulence and, in particular, their specificity to *C. parapsilosis* infections [42].

Tóth *et al.* reported on several virulence factors from *C. parapsilosis*, describing genes possibly related to biofilm formation [43]. This study was divided into two parts. First, a transcriptomic analysis of given *C. parapsilosis* isolates during pathogen-host interaction resulting in the identification of genes with comparatively high expression levels. Subsequently, several parameters usually associated with virulence – such as biofilm formation and adherence – were assessed in deletion mutants for the previously selected genes. The authors identified three mutants - *CPAR2_200390Δ/Δ, CPAR2_209520Δ/Δ* and *CPAR2_501400Δ/Δ* – that displayed lower capacity for biofilm formation. The high expression levels of these genes during pathogen-host interaction allied with the fact that their deletion resulted in lower biofilm formation ability supports the possible relation between biofilms and pathogenicity.

Although adherence and biofilms are major aspects in *C. parapsilosis*' pathogenesis, they're invariably dependent, first and foremost, on the interaction between the yeasts' cell surface and the host's mucosae, and then on the yeasts' interaction with the host's immune system. In its turn, this interaction is dependent on several factors such as extracellular enzymes - virulence factors by themselves [44]. From proteases to phospholipases, extracellular enzymes are varied in nature and function, some of which not yet fully understood.

Secreted aspartyl proteases (SAPs) are major determinants of virulence common to different *Candida* spp. [1,8,45]. SAPs are a varied group of enzymes with 10 known SAPs in *C. albicans* and inter-species variability in their production possibly accounting for differences in pathogenicity [8]. These are involved in the damage of the host's mucosal membranes – facilitating invasion and adherence – as well as in the yeasts' evasion of the host's immune system – from regulating phagocytosis and phagosome-lysosome maturation to directly degrading functional proteins important in the host's immune response [8,45].

C. parapsilosis has three known SAP encoding genes – *SAPP1*, *SAPP2* and *SAPP3* [8,45]. Although *C. albicans*' SAPs have been extensively studied, several reports recognize the role of *C. parapsilosis*' SAPs as still not fully characterized, with this still being the case as recently as 2019 [1,8,45].

Nonetheless, existing knowledge for *C. albicans* can be used as a reference starting point for these kinds of studies. In one such study, Singh *et al*. reported on the effects of disrupting *C. parapsilosis'* SAP genes on several properties known to be related to SAP activity in *C. albicans* – from adhesion and biofilm formation to given interactions with the host's immune system [45]. One initial and significant observation was that although disruption of the genes did not affect cell viability in YPD or YCB media, it did result in decreased viability in serum – specifically *SAPP1* and *SAPP2*. This supports the role of SAPs as virulence factors, given the suggestion they are required for *C. parapsilosis'* survival within the host. Moreover, *C. albicans*' SAPs are known to degrade proteins with antifungal properties such as the host's complement proteins, mediators of phagocytosis [8,45]. Additionally, *Candida* spp. have been described as being able to replicate and survive within macrophages [45]. Singh *et al.* described similar properties for *C. parapsilosis*' SAPs, like easier mutant phagocytosis and elimination compared to wild-type cells, as well as CR3 and CR4 cleavage by Sapp1p and Sapp2p. These observations then place *C. parapsilosis*' SAPs role in virulence and disease alongside *C. albicans*' more extensively studied SAPs, solidly establishing them as important virulence factors.

Phospholipases are another virulence factor present across *Candida* spp. [44]. Phospholipases hydrolyse ester linkages in glycerophospholipids – major constituents of the glycocalyx, covering epithelia and mucosae. This activity confers them a putative role in aiding invasion and adherence by damaging the host's membranes [1,8,37]. Phospholipases have been shown to relate to *C. albicans'* virulence [44]. Given the similarities previously described for SAPs, it would be intuitive to conclude that since *C. parapsilosis* also produces phospholipases these would likely contribute to its virulence. Even so, this correlation is not yet clear [44]. Neji *et al.* reported on the production of phospholipases among several isolates of *C. parapsilosis*, *C. orthopsilosis* and *C. metapsilosis* [37]. In this report, 63,5% of *C. parapsilosis* isolates produced phospholipases while the other two species did not. Despite this observation, Ping *et al.* reported somewhat contradictory findings, showing *C. metapsilosis* as phospholipase producing, even if in a slightly smaller extent than *C. parapsilosis* [44]. Even so, given the lesser clinical significance of *C. orthopsilosis* and *C. metapsilosis*, their inability or smaller capacity to produce phospholipases when contrasted with *C. parapsilosis*' ability to do so, may account for the latter's higher pathogenicity and support phospholipases as virulence factors in this species.

Lipases catalyse the hydrolysis and synthesis of triacylglycerols [1,8,46]. Lipases are thought to have a broad scope of roles, from adhesion, biofilm formation and nutrient acquisition - by providing fatty acids via lipid degradation - to competing with coexisting microbiomes and being involved in the synthesis of inflammatory mediators - as prostaglandins for instance [42,46,47]. Their role in the virulence of *Candida* spp. does seem undisputed, with several reports describing decreased virulence in different infection models for *C. parapsilosis* and *C. albicans* lipase mutants [46]. Similarly, lipases have been described as important virulence factors for a variety of other fungal as well as bacterial pathogens, adding to lipases' significance as mediators of virulence [8].

C. albicans has 10 lipase encoding genes - from *LIP1* to *LIP10* - whereas *C. parapsilosis* has two described lipase encoding genes - *CpLIP1* and *CpLIP2* –, being that only *CpLIP2* has been shown to encode for an active protein [1,8,42,46]. Lipases are known to have a role in host tissue damage, with

lipase inhibitors leading to a decrease in said damage [42,46]. In fact, reports have described the need for higher doses of *C. parapsilosis* inoculate in order to induce membrane lesions relative to *C. albicans*, which might relate to *C. parapsilosis* encoding fewer and less diverse lipases [33].

Tóth *et al.* reported on how lipases seem to modulate the human immune response to *C. parapsilosis'* infection [46]. For this they assessed macrophages' response to a *CpLIP1* and *CpLIP2* knockout mutant when compared to wild-type *C. parapsilosis*. One observation was that though both strains were phagocytised similarly, the knockout mutants were killed more easily than wild-type cells. The killing process of phagocytosed cells is dependent on several mediators of inflammation. Many of these are lipidic in nature, from prostaglandins to leukotrienes, and the increased mutant susceptibility to macrophage killing may be due to a putative role of lipases in their degradation and consequent impediment in the unfolding of the adequate immunological cascades.

Fungal pathogens, including *C. albicans*, are known to produce inflammatory mediators themselves, such as prostaglandins. These fungal prostaglandins have a role in the modulation of the host's immune response aiding in the establishment of infection [47]. Grózer *et al*. reported on a relevant difference in prostaglandin production between *C. albicans* and *C. parapsilosis* [47]. Both *C. albicans* and *C. parapsilosis* possess one *OLE2* gene, encoding for a protein involved in prostaglandin biosynthesis. While the disruption of this gene in *C. parapsilosis* did not lead to a decrease in prostaglandin production, it significantly impaired *C. albicans'* ability to do so. This observation is quite relevant in the extent that, although both pathogens may have similar virulence factors, it further shows how the selection of putative novel therapeutic targets is dependent on the understanding of the specific aspects of *C. parapsilosis'* biology.

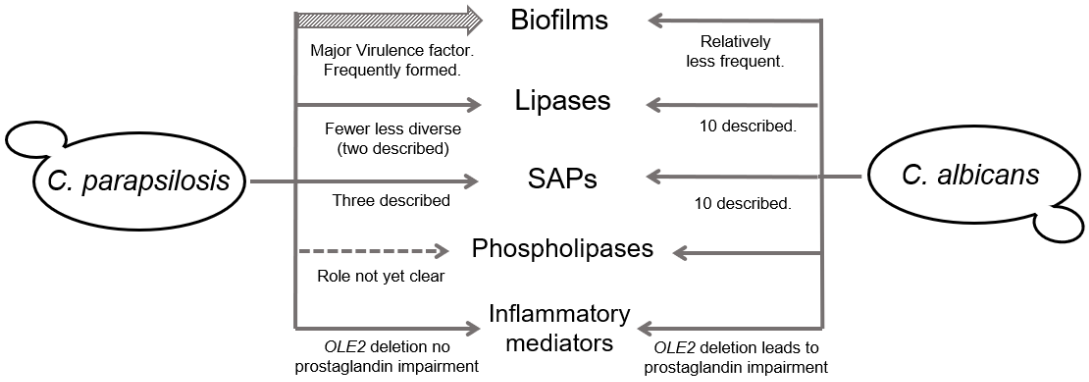

Figure 3: Both *C. parapsilosis* and *C. albicans* have a wide and common range of virulence factors. Among *C. parapsilosis'* virulence factors, biofilms seem to be the most clinically significant, due to its higher frequency of formation relative to other *Candida* spp.. Along with biofilms, secreted proteins such as SAPs, lipases and phospholipases contribute to *C. parapsilosis'* virulence. Nonetheless, many of these, although similar to those of *C. albicans* are not yet fully understood nor described – such as phospholipases, whose role in *C. parapsilosis* specific infections is yet to elucidate. Finally, while both species contain the *OLE2* gene, related to prostaglandin production, only in *C. albicans* does *OLE2* deletion result in prostaglandin production impairment,

In conclusion, although several virulence factors are common to many *Candida* spp., differences in the extent of their production, regulation and diversity within each class are palpable – figure 3. These differences seem to be relevant not only among different *Candida* spp. but as well as among different *C. parapsilosis*' strains, adding to the greater complexity of this emerging pathogen's virulence. However, contradictory reports show how many of these aspects are not yet fully understood. Thus, in an age of rising antifungal resistance and given the therapeutical targetability of virulence factors, understanding *C. parapsilosis'* specific virulence is key to shedding light on how this pathogen can adequately be fought.

1.3.3 Therapeutics and Described Resistance

Antimicrobials were one of the, if not the most impactful advancement in medicine in its millennial history. For the most part of human history not only were we not aware of microorganisms, but not even when they did come to light did we fully understand them as causes of disease or know how to combat them. Then, almost from one moment to the next, we were able to go from a world where too deep a scratch could surely be a death sentence to a world of routine surgery. We went from the centuries long and profoundly dramatic reality of syphilis, for instance, with all social oppression that came with it to the dismantling of antiquated prejudices and relationship moulds. Ultimately, antimicrobials are responsible for the furthering of humankind's emancipation.

Since the first uses of penicillin in the 1930's we've come a long way in the understanding of antimicrobial mechanisms. The scope of antimicrobials at our disposal has been steadily growing and alongside it their use has become more widespread and frequent.

Mindful of this, it becomes clear how worrying it is to already have identified resistance as a growing problem in the less than 100 years since the discovery of antimicrobial molecules.

In particular, resistance to antifungals commonly used to treat *Candida* spp. infections, such as azoles and echinocandins, has been rising [41]. Moreover, as previously mentioned, widespread antifungal therapy allied to heterogeneous susceptibility within the genus has likely lead to observed shifts in incidence among *Candida* spp. causing infection [22,23]. In this context, antifungal resistance encompasses two crucial aspects. It refers both to previously susceptible pathogenic species now being able to survive antifungal therapy, and/or to the enabling of the emergence of new pathogenic species given the subsiding of previously dominant ones. The findings reported by Cuervo *et. al.* on breakthrough Candidemia are one clear example of this [18].

The issue of rising resistance within *Candida* is further worsened by relatively fewer effective antifungals when compared to the greater universe of antibacterials, as a result of the former having to target eukaryotic cells within an also eukaryotic host without risking toxicity [1].

In the 1990s azoles, and fluconazole in particular, emerged as a reliable and efficient antifungal for treating *Candida* spp. infections [5,20]. Ever since, fluconazole has been globally used as the usual first choice antifungal therapy in infections by *Candida* spp. [41,48]. The resulting long term and successive

exposure seems to have potentiated resistance [1,41]. Consequently, echinocandins are now recommended as the front line choice for the treatment of *Candida* spp. infections [41,48,49]. In fact, echinocandins are relatively recent, having come to use in the early 2000's [50,51]. Even so, echinocandin resistance has, already, been described, particularly among strains isolated from immunocompromised individuals under long antifungal exposure [20,41,49].

Fluconazole resistance for *C. parapsilosis* has been identified in several countries – from South Africa, Asia-Pacific and Europe, particularly Italy, to North and Latin America [48,52] – figure 4. This apparently non-geographically restricted emergence of resistance denotes the non-anecdotal nature of this phenomenon and adds to the need for finding alternatives to current common antifungal therapies. One curious observation is that persistent clusters of fluconazole resistant *C. parapsilosis* often occur in countries like Brazil, where poverty has more striking contours [48,53]. This reflects one additional social and economic aspect of this phenomenon since fluconazole is the less expensive option and is still the preferred choice in poorer countries, despite recommendations for echinocandin use.

However, alternative echinocandin treatment for *C. parapsilosis* infections is not as reliable an option as well [20]. Echinocandin minimal inhibitory concentrations (MIC) for *C. parapsilosis* have been shown to be higher relative to other *Candida* spp. as a result of inherently lower susceptibility [1,20,41,49,51]. Consequently, although individuals with systemic infection generally respond well to echinocandin therapy, longer exposure is a significant potentiator of resistant strains due to the need for higher concentrations [41].

Azoles inhibit ergosterol biosynthesis, targeting one of the enzymes involved in its production - the Lanosterol 14-α-Demethylase [13,41,54]. Ergosterol is a major component of the fungal cell membrane, with cholesterol as its analogue in human cells. By interfering with ergosterol biosynthesis, azoles functionally compromise the cell's membrane while leading to the accumulation of toxic14-α-methylated sterols like 14a-methyl-3,6-diol [50,54].

There are three known azole resistance mechanisms described in *Candida* – reduced intracellular accumulation via drug efflux pumps, decreased target affinity and overexpression of the target [13,41]. All three have been identified in *C. parapsilosis* isolates, although with significant geographical variability between them [52]. Reduced target affinity results from one of two identified missense mutations in its coding gene, *ERG11*. Two nucleotide mutations – A395T and A428G – resulting in the Y132F and K143R substitutions lead to a modified azole binding site and consequent impediment of azole binding [41,52,55,56]. However, while the Y132F substitution has been described worldwide - with mutant isolates in Italy, Kuwait, South Korea, Brazil and the United States – the K143R substitution doesn't seem to be as widely distributed, being associated with reports of isolates in the United States and India alone [52,55,56] – figure 4. On the other hand, reduced intracellular azole concentration is due to the overexpression of efflux pumps encoded by *CDR1*/*CDR2* and *MDR1* resulting from Gain of Function (GOF) mutations in the genes encoding their transcription factors, Tac1 and Mrr1 [13,52]. Each of these mechanisms has been described to occur both alone or simultaneously to one another [52].

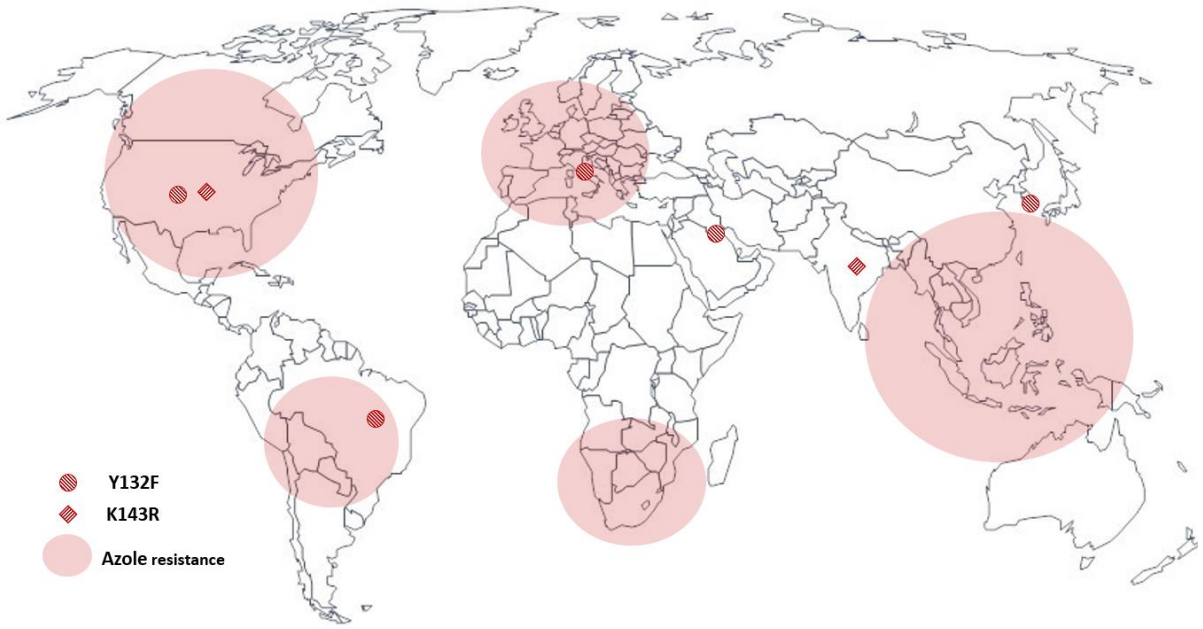

Figure 4: Azole resistance for *Candida* is geographically widespread, denoting its non-anecdotal nature – highlighted regions. Fluconazole resistance in *C. parapsilosis* relates to two missense mutations in the *ERG11* gene, resulting in two amino acid substitutions in its active site – Y132F and K143R. These substitutions do not seem to have the same widespread distribution, with Y132F described in isolates in the United States, Brazil, South Korea, Kuwait and Italy while K143R has been described in the United States and India alone. Base border map retrieved from [57].

Echinocandins target cell wall biosynthesis by inhibiting the β-(1,3)-D-glucan synthase, responsible for the synthesis of a major cell wall polysaccharide - β-(1,3)-D-glucan [13,41,49]. Consequently, the structural integrity of the cell wall is compromised ultimately leading to cell lysis [41]. The β-(1,3)-D-glucan synthase is a complex comprising three proteins and a regulatory subunit, encoded by three different genes – *FKS1*, *FKS2* and *FKS3* [41,49,54]. The P660A amino acid substitution observed in *C. parapsilosis'* Fks1, but not in other *Candida* spp., may be behind *C. parapsilosis* intrinsically lower susceptibility to echinocandins [49].

Antifungals are one of our most important tools in treating *Candida* spp. infections. However, resistance has been increasing and is further aggravated by relatively few effective and reliable antifungal options. Furthermore, the rise in antifungal resistance is a fairly dynamic phenomenon. Geographically widespread and with different mechanisms occurring alone or simultaneously, seemingly depending on the patients' condition. It becomes then apparent that there is an urgent need to further understand the specific pathogenesis of *C. parapsilosis* in order to identify new drug targets, based on which adequate alternative therapeutics may be found.

1.3.4 Nutritional Requirements

The availability of essential nutritional sources is fundamental for a given pathogen's ability to establish infection. Invasive microbial pathogens, like *Candida* spp., are subject to dynamic environments of different compositions within the host [58,59]. Moreover, alternative carbon source usage in *Candida* spp. has been shown to relate to its ability to resist to a variety of stress factors within the host, from cellular wall architecture to internal phagocyte survival [59–61]. Interestingly, different *in vitro* carbon uptake has been shown to be a reliable technique for differentiating some *Candida* spp. [62]. Therefore, understanding the pathogen's nutritional requirements, as well as its ability to adapt to these variations is as relevant for the study of pathogenesis as the understanding of its virulence factors.

One clear example of the role of nutrition in pathogenesis is the relatively higher susceptibility of diabetic individuals to candidiasis. For instance, while susceptibility in immunocompromised individuals is due to inadequate immune response, diabetic individuals are not necessarily immunocompromised and infection is strongly linked to increased glucose levels not only in the blood but in other secretions, like saliva [63].

Table 1 shows the *in vitro* carbon uptake profiles for *C. parapsilosis* and *C. albicans.* Given their relative incidence within the genus, one first relevant observation is that both profiles are indistinguishable. These two species are, as previously described, two of the most clinically significant worldwide. This might relate to their similar carbon usage ability, given that a pathogen's fulfilling and modulation of its nutritional requirements is essential for establishing infection and since, ultimately, all metabolic functions depend on its ability to do so. Moreover, the similarity in carbon uptake profiles could be relevant when trying to find novel therapeutic targets that can transversally affect both species. On the other hand, given the lack of related *C. parapsilosis* specific literature, this similarity provides a good basis for establishing *C. albicans* as a reference organism in this regard. Nonetheless, this gap in knowledge can, as explained further on, be complemented with the construction of a metabolic model for *C. parapsilosis*, further justifying the need for such a model.

Table 1: Carbohydrate uptake profiles for two of the most clinically significant *Candida* spp. - *C. albicans* and *C. parapsilosis*. One relevant observation is the indistinguishability in profile for the two species. Mal - Maltose; Lac – Lactose; Gal – Galactose; Suc – Sucrose; Dul – Dulcitol; Tre – Trehalose; Dex – Dextrose.

	Mal	**Lac**	**Gal**	**Suc**	**Dul**	**Tre**	**Dex**
C. parapsilosis	+	-	+	+	-	+	+
C. albicans	+	-	+	+	-	+	+
References	[62], [64], [65]				[62], [65]		[64], [65]

Host environmental niches are various and dynamic. In healthy individuals, glucose levels are limited and stabilised depending on the site [59]. Thus, in order to survive, *C. albicans*, has been shown to vary its physiology accordingly to environmental niche specificities. Given that *C. parapsilosis*

colonises niches such as the skin, its in-host proliferation is likely to require such a set of adaptive mechanisms as well.

On par with carbon, nitrogen is also an important nutritional requirement as, for instance, a major protein component. Fungi, and *Candida* spp. in particular, can use a wide variety of nitrogen sources, from ammonium and amino acids [66]. Nonetheless, no differently to carbon, uptake of each of these sources is highly regulated depending on availability – be it the kind of, or the lack itself of nitrogen sources in the environment [66]. This regulation encompasses a series of different aspects of *Candida* spp., from the alteration of cell membrane morphology – up or downregulating certain given transporters for instance – to modulating metabolic processes such as protein translation and biosynthesis, SAP secretion and increased flux in so-called nitrogen scavenging pathways [66,67]. *C. parapsilosis* has ammonium sulphate as a preferential nitrogen source relative to isoleucine or other complex nitrogen sources like yeast extract or peptones. For instance, Turner *et al.* reported on the regulation of nitrogen uptake in *C. parapsilosis* when grown in the presence of different nitrogen sources [66]. In this report, over 700 genes were found to be upregulated along with the downregulation of over 200 others upon growth on the non-preferential nitrogen source isoleucine, relative to growth on ammonium sulphate. Among the upregulated genes were varied transporters of nitrogen containing compounds, like urease or several amino acids.

Beyond carbon and nitrogen, the growth and survival of *Candida* spp. also requires other molecules, such as vitamins like biotin. In fact, yeasts such as *Candida* spp. and *C. parapsilosis* in particular, are biotin auxotrophic and as such require biotin to be present in the environment [68].

In conclusion, *Candida* spp. modulate their nutritional requirements according to the colonization environments. The complexity of this modulation is further increased by colonization occurring within a host. These mechanisms seem to be common throughout the genus, although varying according to different preferential nutritional sources. Even so, there seems to be a lack in specific literature for *C. parapsilosis* in this regard. Given the strong link between nutrition and pathogenesis, this gap in knowledge could begin to be filled via the construction of a metabolic model for *C. parapsilosis*, guiding the understanding of its metabolic specificities and shedding light on how to specifically target this pathogen's growth within the host.

1.4 Understanding Genome Scale Metabolic Models

Genome Scale Metabolic models (GSMMs or GEMs) correspond to the *in silico* reconstituted metabolic network of a given organism [69]. In the little more than 20 years since the first published metabolic model, GSMMs have garnered growing attention and have become powerful tools for the study of the cells' metabolism from a Systems Biology perspective [70,71]. These models integrate vast amounts of information, from proteomics to metabolomics and as such, their use and development has been intertwined with the increasing availability of large amounts of data, such as sequenced genomes for instance [71–73]. In fact, the first such model was published in 1999 for *Haemophilus influenzae* RD

which, curiously, was also the first organism to have its genome fully sequenced with around 80% of its ORFs duly annotated at the time of the model's publication [74]. At this stage, modelling was still a fairly insipient affair, still uncertain of genotype-phenotype relations' predictability based on genotypic data. *H. influenzae* RD was followed, in 2003, by the first *Saccharomyces cerevisiae* model, also the first eukaryote with a fully sequenced genome [71]. Then, just four years later in 2007, the first human metabolic model was published [75]. Since then, GSMMs have showed their versatility and applicability to concrete problems, from being used in metabolic engineering, aiding in strain design, to the study of human cancer and drug design [71,76–78].

1.4.1 The Rationale Behind Genome Scale Metabolic Models

The cell's metabolism cannot be seen as a set of individual reactions each catalysed by one single enzyme. Reactions integrate a complex network where they are often compartmentalized, one enzyme can catalyse several different reactions whose products can, in turn, act as substrates for several others, and where there are no clear bounds between metabolic pathways – figure 5. As a result, the disruption of one reaction can affect several others, be it positively or negatively. Hence, being able to predict these interactions before planning any kind of experimental setup is not only a powerful tool for the understanding of complex systems - such as is the metabolism of pathogenic yeasts – but, above all, monetarily and time saving [78–80]. This calls for a paradigm change, from analysing individual elements - Components Biology - to a global panoramic view - Systems Biology -, which aims precisely for the study of how individual components relate to each other within a given system [80].

Genome Scale Metabolic Models surged as one possible solution for this change and are one of Systems Biology's key tools regarding the study of the cells' metabolism [71,73,78].

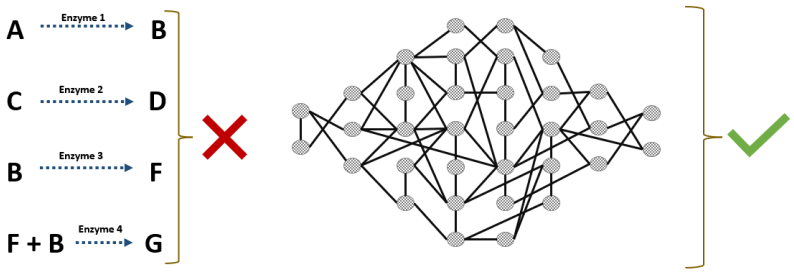

Figure 5: The cell's metabolism is not a set of discrete reactions, each with their own specific enzymes and metabolites. It is rather a fairly complex network with no clear bounds between pathways, where one enzyme can catalyse several different reactions, whose products themselves can act as reagents in several others.

The basic idea behind GSMMs is that all the cell's enzymatic activities are encoded in the genome. In their turn, each of these activities corresponds to specific reactions, ultimately and ideally representing the totality of the cell's metabolism. Thus, GSMMs are primarily based on Gene-Protein-Reaction (GPR) associations and correspond to the reconstituted metabolic network of a given

organism[69,76,78,81] – figure 6. Mathematically, these associations are represented by a $M = m.n$ matrix, where a given metabolite m_n in a given reaction r_n has a mr_n stoichiometric coefficient [69,71]. A positive coefficient corresponds to the production of that given metabolite m_n in the given reaction r_n, while a negative coefficient corresponds to its consumption [69].

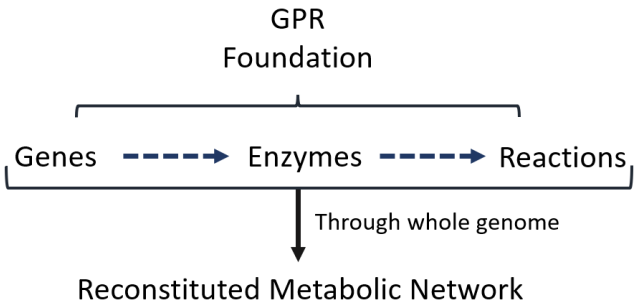

Figure 6: Gene-Protein-Reaction (GPR) associations are the fundamental triad of Genome Scale Metabolic Models. The idea being that the genome encodes for the cell's enzymatic activities (Gene-Protein) which, in their turn, correspond to one or more reactions (Protein-Reaction), thus completing the GPR foundation. Implementing this idea throughout the genome ideally results in the fully reconstructed metabolic network of a given organism.

The first step in the construction of a GSMM is the, increasingly automated, generation of a draft model based on GPR associations [70,73]. Consequently, this step requires first and foremost the organism of interest's genome sequence [71,82]. Ideally, this sequence should be duly annotated and the greater the extent of annotation the greater the quality of reconstruction [73]. Nonetheless, given the relative conservation of protein function between organisms due to possible given common ancestries, existing annotations can be complemented by homology searches against existing databases such as KEGG, BRENDA, among others [72,73,81,83]. This further shows how GSMMs are the result of the concerted effort to concentrate the cumulative knowledge of more than a century of biochemical research [76]. However, given the organism non specificity of many of these databases, this complementary annotation is prone to false-positives, attributing enzymatic activities to sequences that, although present, do not encode for active proteins [73,83]. As a result, the longest and most effort requiring step consists of the manual curation of this draft model.

Curation of the draft model is a time consuming and effortful process, requiring extensive bibliographic research and experimental validation. At this stage, curation looks for several aspects, from reaction mis-connectivity within the network – gap finding – and correct enzyme-reaction assignment, to correct reaction mass balance, reversibility and directionality [72,73]. Nonetheless, before further refinement a biomass equation should be implemented. This equation is usually one of the possible target functions for optimization and plays a role as the final criteria on which model simulations are performed [71]. The biomass equation encompasses the cell's growth requirements – the building blocks necessary for biomass production -, each specified alongside its respective contribution fractions[71,73]. In turn, these requirements are experimentally determined and should be organism

specific. However, specific data might not be available and quantification from closely related and well-studied organisms can be applied which, nonetheless, may be problematic in the extent that it may compromise the models' predictive accuracy [71,73]. At this stage, compartment information can also be implemented.

After this initial curation, which assures the network's connectivity and the production of all components of the biomass equation, comes experimental validation [72]. This consists of comparing the models phenotypic predictions to experimentally described observations – such as carbon and nitrogen sources usage for instance [73]. In essence, this constitutes an added curation step, allowing for the identification of further issues, based on more stringent experimental constraints – given that the model is more explicitly forced to obey experimental data.

Although widely useful and accurate, metabolic models have as a major issue the non-integration of regulatory information. Metabolism is regulated at several different levels. From regulating enzyme concentration via gene expression and inhibition, to structural enzyme modification often exerted by substrates or products of the catalysed reaction, affecting its activity. Given the complexity and in many cases the not yet full understanding of these processes, there is a significant difficulty in the implementation of these processes in metabolic modelling, often affecting their predictive capabilities[71].

1.4.2 GSMMs as Tools for Novel Drug Target Identification

In the little more than 20 years since the publication of the first metabolic model, GSMMs have proven their versatility and applicability, having been used in several fields due to their scaffolding nature[84]. From aiding in genome annotation and metabolic engineering - guiding strain design for biofuel and biopolymer production as well as for CO_2 fixation – to the study of microbiomes and, in particular, by elucidating putative target selection for drug design [70,80,83,84].

GSMMs usefulness consists of their ability to predict the effect of given reaction disruptions in the rate of other given reactions elsewhere in the network [76]. This idea is particularly clear when it comes to understanding how to impair a given microbial pathogen's metabolism and growth. Moreover, when it comes to drug design, metabolic models may aid in integrating the added complexity resulting from disease occurring within the host. These considerations go from possible host toxicity – which becomes particularly relevant when targeting eukaryotic pathogens, like *Candida* spp. – to possible effects that existing microbiomes, as for instance the gut-microbiome, might have on effective host drug exposure[85].

There are two main rationales for drug target analysis using metabolic models – constraints-based flux analysis and topological analysis [80]. The most intuitive one is the constraints-based approach since its aim is to directly impede the pathogen's growth. Assuming mass-balanced reactions and the system's steady state – given fast metabolite turnover –, null intracellular metabolite accumulation can be assumed and reaction fluxes can be restricted to within specific feasible bounds. As a result, these fluxes can be calculated under specific conditions – by specifying essential carbon and nitrogen sources, for instance, and consequently forcing the model to adapt its atom flow through the network – and

essential reactions under those specific conditions can be identified. This analysis is commonly done via Flux Balance Analysis (FBA), consisting of determining the flux of each reaction for optimal growth while forcing a null flux for each reaction one by one [86]. Furthermore, essential reaction identification can also lead to essential gene analysis, due to the GPR basis of the model and, ultimately, allow for essential reaction identification. In other words, this analysis allows for the identification of possible reactions whose impairment compromises the model's ability to fulfil all of the biomass equation's components and consequently, impede its growth. Subsequently, metabolites that modulate these enzymes' activity – so called antimetabolites, consisting of molecules similar to natural ligands that competitively bind the enzyme's active site – can be identified and later validated via further implementation of other techniques like Molecular Docking [76]. Alternatively, topological approaches focus on finding weak points in the metabolic network – from enzymes producing unique metabolites to possible nodes susceptible to metabolic overload. Nonetheless, relative to constraints-based flux analysis, topological analysis has significant fragilities, given it doesn't for instance consider if a given putative target enzyme is active under the conditions of interest [70].

T. Y. Kim *et al.* reported on essential metabolite identification via the analysis of metabolic models of four relevant pathogens - *Helicobacter pylori*, *Staphylococcus aureus*, *Mycobacterium tuberculosis* and *Escherichia coli* [86]. In this analysis the authors focused on an integrative approach, applying a topological analysis - metabolic chokepoint analysis - allied with constraints-based flux analysis. Metabolic chokepoints correspond to given places within a network where a metabolite is surrounded by enzymes that only either produce or consume them. Consequently, that given metabolite is likely an essential metabolite, which can be confirmed via flux-based analysis. The basis for this hybrid approach results from the further complexification of drug target identification arising from the observation that disrupting a given metabolic system in several places simultaneously, even if in relatively lesser extents, is more effective than completely disrupting one specific reaction [86]. This aspect further shows how important it is to have a systems view, given that it truly allows for the identification of molecules capable of systemic metabolic impairment. Thus, the authors aimed to increase the stringency of target identification by scaffolding both the criteria, and subsequently, the results from both topological and flux-based approaches while further reducing the number of putative targets and increasing their accuracy.

From this analysis the authors were successful in identifying not only already described and validated targets – which validates the predictive ability and accuracy of these tools – but as well as in identifying putative novel drug targets. For instance, in the case of *M. tuberculosis* and *H. pylori* two possible targets were identified – mycolate and adphep-DD biosynthesis, respectively. While mycolates are in fact current targets for already in-use anti-tuberculosis drugs, adphep-DD had no described targetability at the time of the paper's publication [86].

GSMMs have been increasingly refined and widely applied to the systemic understanding of not only microbial disease, but also other relevant pathologies like cancer. These models have been proven to accurately predict possible new drug targets. Not only this, they have allowed for the further understanding of the complexity of pathogen-host interactions as well as made possible a paradigm

change when it comes to target identification. From looking for molecules that should be effective while having one unique target, to understanding the increased effectiveness of multiple and simultaneous targetability. It becomes then apparent GSMMs are a powerful tool for the development of molecular medicine.

1.4.3 Docking as Follow-up Exploitation for GSMM Derived Putative Drug Targets

One fundamental idea behind drug design is that molecules structurally similar to the ligands of a given enzymatic target can bind the same sites and consequently competitively inhibit that same enzyme [76]. From the previously described analysis results a set of several putative enzymatic therapeutic targets. Identifying molecules with possibly inhibitory effects on these putative targets is commonly done via Molecular Docking.

Molecular Docking is an *in silico* structure-based analysis of ligand-enzyme interaction and modulation at an atom level [76,87,88]. In essence, Docking consists of considering all possible orientations and conformations of a given molecule within a given binding site and estimating complementarity by implementing a scoring function [88]. In its turn, this function consists of the criteria that distinguish good ligands from inactive molecules based on the non-covalent ligand-enzyme interactions.

Since its first uses in the 1970's, both the technique's scope of application and the understanding of ligand-receptor interactions have changed. From the lock and key theory - considering structurally rigid ligand and binding sites - used in the first biomolecular applications of docking in the 1980's, to the currently used theory of induced fit [87,88]. The latter considers both the ligand and the enzyme's binding site to be flexible. However, given its added complexity, considering both the ligand's and the receptor's flexibility requires significant computational power and, although increasingly applied, this is still one of the techniques greatest challenges [88].

Once again, when considering drug interactions, the environment where these occur has to be taken into account. On one hand, possible adverse parallel interactions with other enzymes within the host – so called off-targets –, risking toxicity have to be considered. On the other hand, parallel interactions might be of interest. As previously mentioned, simultaneous multi-targeting is in fact the more adequate notion guiding drug design. Much like GSMMs, Molecular Docking allows for a systems approach for the identification of ligands with a set of interactions of interest, opening the way for the growing field of polypharmacology [87]. This systemic approach is further made possible by the ability to screen existing databases and widen the range of putative novel active molecules [88].

1.5 Final Remarks

Disease can rarely be seen as an isolated phenomenon. Even beyond its deeply human facet, it is often a result of seemingly unrelated events in apparently distant places. Climate change is now part of our reality. Habitat destruction and consequent species' and particularly human mass migrations are already happening and will only aggravate. New species contacts will occur and species barriers

transposed. Migrant camps without adequate sanitation will furthermore lead to increased susceptibility to disease. Moreover, regarding microbial disease, the decline of historical pathogens is met with the rise of new pathogenic species allied to increasing antimicrobial resistance. *Candida* spp. infections are a clear example of this. Non-geographically restricted and presenting dynamic shifts of species incidence and markedly variable antifungal susceptibility within the genus. In particular, *Candida parapsilosis*, among *Candida* spp., has surged as one of the most significant threats to public health. Thus, the need to find new adequate therapeutics as well as new, adequately paced, research tools, has possibly never been stronger.

The importance of this interconnectivity spans to many aspects of Nature, including at a cellular level. In fact, the metabolic understanding of pathogenesis has greatly benefited from this systemic view, made possible by the development of Genome Scale Metabolic Models. In the little more than 20 years since their first developments, GSMMs have shown their versatility and applicability. They have allowed for paradigm changes, proven their accuracy and, above all, time savingly guided the planning of wet-lab validations. Moreover, GSMMs are a product of the accumulated knowledge of decades of research and the concerted and solidary effort to assimilate it. Although still with many aspects in need of refinement, little by little, research worldwide is establishing GSMMs in the medical field. With this in mind, GSMMs are set to be a tool for the future.

2. Materials and Methods

The herein described metabolic model reports to the yeast *Candida parapsilosis* with the taxonomic ID 5480. Construction was performed using *merlin* 4.0.5 [89] as the basilar software. This reconstruction was a semiautomated process, overall divided into two main parts – metabolic network reconstruction and validation both requiring extensive manual curation. This stage was then followed by essentiality prediction analysis.

Network reconstruction comprised one initial step resulting in a draft model mainly consisting of the implementation of automated tools, followed by extensive manual curation – figure 7. The latter in its turn was divided into two similar stages – pre and post compartmentalization – both consisting of network connectivity solving and refinement – gap-filling, reaction mass balancing, and correcting reaction reversibility and directionality. Compartmentalisation was only implemented on the previously already connected and minimally validated network in order to facilitate post-compartmentalisation issue solving - being that prior to compartmentalization, an initial flux-based analysis step was performed, assessing consumption/production profiles and some carbon and nitrogen source usage.

The final main validation step assessed three main aspects – the model's Consumption *vs.* Production profile, carbon and nitrogen source usage and the model's predicted growth rate relative to the experimentally observed growth rate for a corresponding rate of glucose consumption per grams of Dry Cell Weight per hour.

Finally, the model's reaction and gene essentiality were predicted in simulated RPMI medium and coalesced with predictions from other published models for the two pathogenic yeasts *C. albicans*[90] and *C. glabrata* [91].

Henceforth, this model will be referred to as *Cpar*MM (*C. parapsilosis* Metabolic Model).

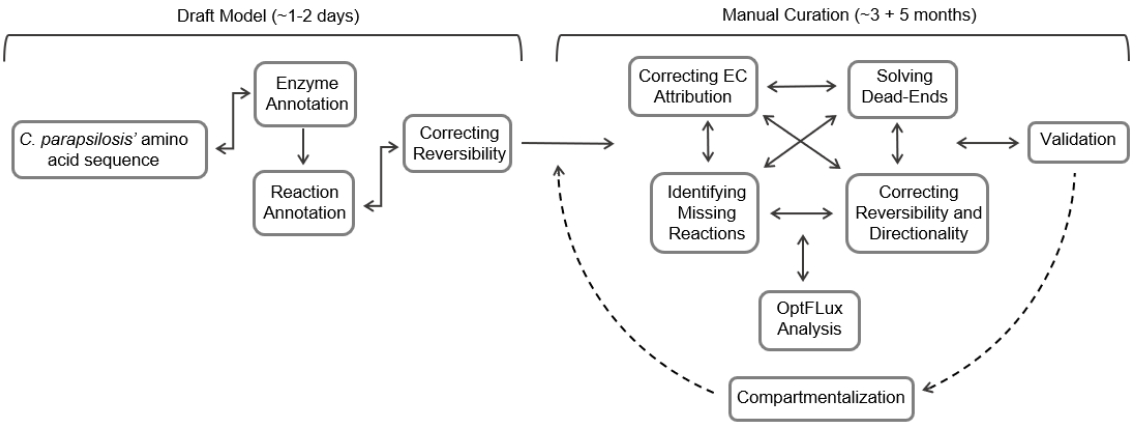

Figure 7 Schematic representation of the general procedure behind the construction of the herein described model, *Cpar*MM. Draft model construction implements mainly automated tools and thus its relatively shorter time span. Manual curation involves a series of non-hierarchical aspects to be assessed and corrected, thus the bidirectional arrows. The workflow of this second stage of curation is similar for both the non- and compartmentalised model.

2.1 Enzyme and Reaction Annotation

The construction of the initial draft model was divided into enzyme and subsequent reaction annotation. These annotations were performed with resort to *merlin*'s integrated tools, allowing for homology alignment searches via the implementation of the Basic Local Alignment Search Tool (BLAST)[92] of *C. parapsilosis'* amino acid sequences against remote databases - UniProt-SwissProt and UniProt-TrEMBL [93]. The amino acid sequences were retrieved from NCBI [94]. Initial automated curation of enzyme annotation was based on phylogenetic proximity as described in Tsui *et. al* 2008 [95]. Reaction annotation was performed via the integration of data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [96] in the model resulting in all reactions and metabolites having KEGG's respective ID.

2.2 Correcting Reaction Reversibility, Directionality and Balance

Correction of reaction reversibility and directionality was performed as a semi-automated process. Initial meta curation of the previously obtained annotated reactions was performed with resort to a *merlin* integrated tool, which implements information from the remote tool eQuilibrator [97]. Further reaction curation was entirely manual and resulted from a sequential procedure of blocked metabolite identification, localization in the network and respective biosynthetic and consuming reaction assessment as explained in the discussion. Corrections were justified with resort to information from MetaCyc [98] and existing literature.

Unbalanced reactions were identified automatically and balancing was performed manually and justified with resort to MetaCyc, CheBi [99], Brenda [100] and existing literature.

2.3 Biomass Equation

In this project the Biomass equation was the optimization target, serving as criterium for the further on described curation of the model. This equation encompasses the cell's biomass components along with their respective and relative contributions – DNA, RNA, carbohydrate, protein, lipid and cofactor content. This model's biomass equation was defined based on both automatically derived data from *C. parapsilosis'* genome and data from existing literature. Due to lack of *C. parapsilosis*' specific literature regarding biomass composition, the herein used biomass equation includes data from *C. parapsilosis, Candida albicans, Candida tropicalis* and *Saccharomyces cerevisiae*. The discerned components of the biomass equation along with their relative content are shown in the supplementary excel file S1.

2.4 Compartmentalisation

Compartmentalisation was implemented with resort to WoLF PSORT [101], a protein localization predictor. Compartmentalization was also only implemented on the already connected non-compartmentalised model in order to simplify respective issue solving.

2.5 Metabolic Network Reconstruction

Metabolic network reconstruction was performed in two main stages – a pre- and post-compartmentalisation stage, the first of which consisting of reconstruction followed by a superficial validation step. The first step aimed at assuring the network connectivity of the automatically obtained reconstruction and consequent production of all biomass components. At this stage the biomass equation served as the first line criterium in the pipeline of blocked or missing reaction identification. Components of the equation whose production was not assured by the initial draft model – gaps - were first identified individually. Each of these was located in the network, blocked or missing reactions were identified and located, and possible issues assigned. The following initial Flux Balance Analysis was performed using OptFlux [102] and was meant to assess and correct the model's nutritional requirements and some intra-network paths for the synthesis of specific metabolites. Compartmentalisation followed this stage and was thus implement on an already connected network in order to facilitate issue solving. Beyond the only addition of a new kind of gap-filling - the new transporter-based as opposed to gene-based gap-filling – reassuring connectivity followed the same rationale as in pre-compartmentalisation *Cpar*MM.

Corrections and reaction addition were justified with resource to literature search, remote databases such as MetaCyc, Brenda, Candida Genome Database (CGD) [103] and UniProt, and to the already validated and published models for *S. cerevisiae* iMM904 [104] and iND750 [105] – all edited reactions are shown in the supplementary excel file S1.

In the face of existing literature supporting ubiquinone-9 as the major ubiquinone in *C. parapsilosis* [106], and in an effort to further specify the model, the previously automatically annotated general ubiquinone biosynthetic pathway – incorporating the general compound "Ubiquinone" with the KEGG ID c00399 – was substituted by a manually constructed ubiquinone-9 biosynthetic pathway termed "Ubiquinone-9 Biosynthesis". The latter was reconstructed from the pathway as shown in MetaCyc [107].

2.6 Validation

The second stage of construction corresponded to validation, similarly performed using OptFlux, which meant to assess *Cpar*MM's predictive reliability by evaluating three main aspects. First its consumption *vs.* production profile, its ability to predict growth using different compounds as sole carbon or nitrogen sources relative to *in vivo* described data, and its predicted specific growth rate ($\mu.h^{-1}$) relative to that observed for a given experimentally determined glucose consumption rate ($mmol.gDCW^{-1}.h^{-1}$). Due to lack of literature data these latter parameters had to be experimentally determined as described in Sauer *et al.* [108].

2.6.1 Growth Media and Strains

C. parapsilosis type strain ATCC 22019 was batch cultured at 30 ºC in orbital agitation (250 rpm) in Yeast Extract-Peptone-Dextrose (YPD) for inoculate cultivation, Synthetic Minimal Media (SMM)

for growth parameter determination, and Yeast Nitrogen Base either without amino acids or ammonium sulphate for carbon and nitrogen source usage assessment.

Media composition are as follows: YPD: 20 g/L glucose (Merck), 20 g/L (Merck) and 10 g/L yeast extract (Merck); SMM: 20g/L (Merck), 2.7 g/L ammonium sulphate, 0.05 g/L magnesium sulphate, 2 g/L potassium dihydrogen phosphate, 0.5 g/L calcium chloride and 100 µg/L biotin; Carbon source assessment with YNB: 5 g/L carbon source and 6.8 g/L YNB without amino acids; Nitrogen source assessment YNB: 5 g/L glucose (Merck), 2.7 g/L nitrogen source and 1.7 g/L YNB without ammonium sulphate. Solid media contained an additional 2 g/L agar (Iberagar).

2.6.2 Curating the Model's Consumption *vs.* Production Profile and Intra-Network flux Distribution

The model's consumption *vs.* production profile and flux distribution was assessed in OptFlux[102] and simulations were performed in simulated SMM – composition as described in 2.6.1. This step consisted of a flux-based analysis that thus follows flux throughout the network and allows for the identification of unduly available or missing intra-network paths. These paths can correspond to biologically non justifiable entry or exit points in the network – reflecting on the model's consumption and production –, or can correspond to paths that bypass certain central pathways. This kind of analysis allows for the following of a given metabolite's consumption throughout the network and thus allows for the identification of these kinds of issues. Curation consists mainly of correcting incorrect reaction reversibility and directionality and in removing unduly annotated reactions or gap-filling.

2.6.3 Assessing Carbon and Nitrogen Source Usage

Regarding carbon and nitrogen source usage, the model's ability to predicted biomass production using several compounds as sole carbon or nitrogen sources was compared to literature based *in vivo* observations. For the few disparate literature based and predicted results, wet-lab growth was assessed using the lab isolate *C. parapsilosis* type strain ATCC 22019. Along with their respective KEGG IDs, the tested sources were ribose (00121), glycerol (00116), succinate (00042), cellobiose (00185) raffinose (00492) and L-Lysine (00047), along with control conditions using glucose and ammonium as respectively carbon and nitrogen sources.

Growth was assessed in solid YNB media. *C. parapsilosis* cell suspensions were grown in YPD until an Optical density at 600 nm (OD_{600nm}) of 0.5 – mid-exponential phase – and subsequently diluted in sterile water to an OD_{600nm} of 0.05. Three consecutive *C. parapsilosis* cell suspensions – 10^{-1}, 10^{-2} and 10^{-3} – were plated as 4 µL spots and the plates were incubated at 30 ºC for 24 hours.

2.6.4 Determining Glucose Consumption Rate in Batch Culture

Cultures for determining the glucose consumption rate per Dry Cell Weight per hour ($Y_{x/s}$.mmol.$gDCW^{-1}.h^{-1}$) were performed in SMM. *C. parapsilosis* cell suspensions were grown in YPD

until an OD_{600nm} of 0.5. Cultures were performed in SMM from an initial OD_{600nm} of 0.3 and incubated at 30 ºC in orbital agitation (250 rpm) for 10 hours. Growth in liquid media was monitored by measuring culture OD_{600nm}.

Samples for determining dry biomass and glucose concentration were taken every 2 hours and centrifuged at 13000 rpm for 3 min. The culture supernatant was collected for glucose concentration determination and the pellet was used for determining dry cell weight. Biomass was measured as the mass difference of each sample tube while empty *vs*. after lyophilization of each sample pellet. Glucose concentration was determined by HPLC on an Aminex HPX-87 H Ion Exchange chromatography column eluted with 0.0005 M H_2SO_4 at a flow rate of 0.6 mL/min at room temperature. Concentrations were determined with resource to the adequate calibration curves.

2.7 Essentiality Prediction and Analysis

Essential gene and reaction annotation was performed using OptFlux [102] and resulted from a Flux Balance Analysis where each reaction's flux was nullified individually and the biomass equation was optimized in each of these conditions. This simulation was performed in simulated RPMI 1640 - composition shown below in table 2 [109]. Gene names and orthologs were retrieved from Candida Genome Database (CGD) [103] and UniProt [93]. Corresponding information regarding existing drugs or inhibiting compounds and their respective investigational state was retrieved from DrugBank [110]. Putative targets of interest were selected and discussed with resource to existing literature.

Table 2: RPMI media composition and constituting metabolite respective KEGG IDs [109] as simulated for gene and reaction essentiality prediction.

Compound	KEGG ID	Compound	KEGG ID	Compound	KEGG ID
Choline	c00114	Arginine	c00062	L-serine	c00065
myo-Inositol	c00137	Histidine	c00135	Aspartate	c00049
Folate	c00504	Phenylalanine	c00079	Glutamate	c00025
Nicotinamide	c00153	D-Tryptophan	c00525	Proline	c00148
Pyridoxine	c00314	Pantothenate	c00864	Threonine	c00188
Thiamine	c00378	L-Asparagine	c00152	Tyrosine	c00082
Biotin	c00120	L-Cysteine	c00491	Valine	c00188
Riboflavin	c00255	Leucine	c00123	Methionine	c00082
Glucose	c00031	Isoleucine	c00407	Glycine	c00183
Glutamine	c00064	L-Lysine	c00047	Glutathione	c00073

3. Results and Discussion

3.1 Model Characteristics

The herein describe metabolic model for *C. parapsilosis* - *Cpar*MM - is a compartmentalised model comprising 1112 genes, 598 proteins, 2892 reactions and 2885 metabolites across four compartments – extracellular, cytoplasm, mitochondria, and peroxisome. Manual curation assessed a total of 408 reactions from which 83 were mass balanced, 162 were corrected in regard to reversibility or directionality, and 163 had their annotation corrected or completed, were added or removed.

Furthermore, CparMM is capable of predicting biomass production with a nutritional cost coherent with that observed for other models for *S. cerevisiae* [105] and *C. albicans* [90] and is reliably predictive of biomass production in different nutritional conditions. However, regarding the model's Consumption *vs.* Production profile, three produced metabolites are of notice. Firstly, the chimeric metabolites Pyridoxal phosphate Residue and L-Histidine Residue, which as explained further on in 3.3.4 result from the mass balancing solution implemented on the reaction with the KEGG ID R10686. Glycolaldehyde production is also of notice and results from the non-solvability of its corresponding dead-end – discussed and exemplified in 3.3.1. This is, due to the absence of duly referenced glycolaldehyde consumption reactions that may connect this metabolite to the network – presenting as unlikely this being an unique metabolic feature of *Candida* spp. when compared to other published and thus validated models.

*Cpar*MM shows a significant and reliable ability to predict biomass production using different compounds as sole carbon or nitrogen sources. From the 34 tested compounds – shown in table 4 – only 5 corresponded to incorrect predictions – translating into 85% correct predictions. Furthermore, from an also quantitative perspective, *Cpar*MM is also reliably predictive of the rate of biomass production relative to a given corresponding glucose consumption rate. For an experimentally determined glucose consumption rate of 2.098 +/- 0.404 $mmol.gDCW^{-1}.h^{-1}$ *Cpar*MM predicts a specific growth rate of 0.180 h^{-1} being the corresponding experimentally determined 0.159 +/- 0.027 h^{-1}. These values have no significant differences – note the experimental uncertainty – and reflect the network's curation extent and fine tuning.

In its turn, the implemented Biomass equation includes data from *C. parapsilosis, Candida albicans, Candida tropicalis* and *Saccharomyces cerevisiae* and is as follows – equation 1 and table 3.

Biomass = 0.5056 e-Protein + 0.3822 e-Carbohydrate + 0.0377 e-Lipid + 0.0096 e-Cofactor + 0.0602 e-RNA + 0.0038 e-DNA **(1)**

Note that the notation e-macromolecule corresponds to a model construction abstraction that serves the purpose of simplifying the representation of each relative contribution of each macro-component of the biomass. The idea then is that one e-molecule is defined by a kind of biomass sub equation that comprises the entirety of the metabolites that correspond to its chemical nature. For instance, one e-lipid comprises all the sterols, fatty acids and phospholipids each with their own sub-contributions to the

totality of the lipid composition of the model. Then, the synthesis of all these sub-components leads to the synthesis of one e-lipid molecule, and the same applies for all other e-molecules.

Table 3: The full *Cpar*MM biomass equation metabolite composition and their corresponding KEGG compound ID. Note that an e-molecule corresponds to a model abstraction, being that these e-molecules are what in fact constitute the biomass equation. The idea then is that the synthesis of the totality of the lipids for instance, corresponds to the synthesis of one e-lipid molecule. The full individual validated contributions of each of these metabolites are shown in the supplementary excel file S1.

e-Metabolite	Metabolite	KEGG ID	e-Metabolite	Metabolite	KEGG ID
Lipids	Lanosterol	C01724	**Proteins**	L-Valine	C00183
	Squalene	C00751		L-Tyrosine	C02515
	Ergosterol	C01694		L-Tryptophan	C00078
	Phosphatidylserine	C02737		L-Threonine	C00188
	1-Phosphatidyl-D-myo-inositol	C01194		L-Serine	C00065
	Phosphatidylcholine	C00157		L-Proline	C00148
	Phosphatidylethanolamine	C00350		L-Phenylalanine	C00079
	Phosphatidic acid	C00416		L-Methionine	C00073
	Phosphatidylglycerol	C00344		L-Lysine	C00047
	Tetradecanoic acid	C06424		L-Leucine	C00123
	Hexadecanoic acid	C00249		L-Isoleucine	C00407
	(9Z)-Hexadecenoic acid	C08362		L-Histidine	C00135
	Octadecanoic acid	C01530		L-Glutamate	C00025
	(9Z)-Octadecenoic acid	C00712		L-Cysteine	C00097
	(9Z,12Z)-Octadecadienoic acid	C01595		L-Aspartate	C00049
	(9Z,12Z,15Z)-Octadecatrienoic acid	C06427		L-Asparagine	C00152
	Triacylglycerol	C00422		L-Arginine	C00062
	Monoacylglycerol	C01885		L-Alanine	C00041
	Diacylglycerol	C00165		Glycine	C00037
	Sterol esters	C01958		L-Glutamine	C00064
Cofactors	Thiamine	C00378	**Carbohydrates**	Chitin	C00461
	Ubiquinone-6	C17568		Mannan	C00464
	NADP+	C00006		β (1,3)-Glucan	C00965
	NAD+	C00003	**RNA**	UTP	C00075
	FMN	C00061		GTP	C00044
	FAD	C00016		CTP	C00063
	CoA	C00010		ATP	C00002
	Biotin	C00120	**DNA**	dTTP	C00459
	Pyridoxal phosphate	C00627		dGTP	C00286
	Tetrahydrofolate	C00440		dCTP	C00458
				dATP	C00131

Moreover, in simulated RPMI growth medium a total of 129 genes and reactions were predicted as essential, and 37 were common between the three foremost pathogenic *Candida* spp. – *C. albicans, C. parapsilosis* and *C. glabrata.*

However, carrying *Cpar*MM through to this stage implied a series of sequential and extensive steps, each requiring the pondering of several different issues and solutions as described below.

3.2 Construction of a Draft Model

The first step in the construction of *Cpar*MM was the obtaining of a draft model based solely on GPR associations. This is an automated process consisting of two fundamental steps - enzyme and subsequent reaction annotation. Thus, the first requirement is the organism's genome sequence [71].

Enzyme annotation resulted from two sequential homology alignment searches. The first ran against UniProt-SwissProt, an organism non-specific curated database which may have resulted in coding sequences without hits. Thus, this annotation was further complemented by integrating another broader and non-curated database – UniProt-TrEMBL - allowing for the identification of additional putative enzymatic activities. This initial annotation resulted in 5810 genes from which 1876 were enzyme encoding. However, given the organism non-specific nature of these databases, the previously performed annotation is prone to false positives [73,83]. This is, the obtained annotation includes both genes pertaining to *C. parapsilosis* as well as orthologs pertaining to other organisms when no information is found for the former, a frequent event due to the unduly annotation of *C. parapsilosis'* sequences. Moreover, at this point each identified gene had several possible hits corresponding to orthologs across different species. This called for a more stringent annotation criterium - in this case based on phylogenetic proximity [95] -, corresponding to the first of many curation steps in the construction of such a model. *merlin* allows for the automated implementation of this criteria based on a manually instructed species order. This step consists of a reiterative process assessing the existence of annotated orthologs for each species in the list, going to the subsequent phylogenetically closest indicated organism if no orthologs were identified for the previous one. As a result, not only is the genome annotation maximised but each annotation has an associated confidence [81].

Reaction annotation followed enzyme annotation, completing the GPR triad of the metabolic model. KEGG contains information on known metabolic reactions, from reactants and their chemical formulae, to their respective enzymes. Integrating this database into *Cpar*MM not only allows for the linking of reactions to the previously annotated enzymes, but furthermore allows for the standardization of nomenclature within *Cpar*MM and to other existing databases used further on in the model's curation.

From this step resulted a putative and minimally curated reconstruction of *C. parapsilosis'* metabolic network that ideally, and assuming the genome's full and curated annotation, would represent the totality of the network. However, the resulting reconstruction includes several non-specific hits, several incorrectly annotated genes, misses non-identified reactions (gaps) and reaction reversibility, directionality and mass balance have to be assessed. Consequently, this automatically derived reconstruction is uncapable of fulfilling its biomass components and results in null-predictive capacity. In fact, from the 61 metabolites encompassed by the biomass equation, only 20 were being produced at this stage. This is, 41 metabolites pended assessment. This called for the next stage in the model's construction - manual curation [80].

3.3 Manual Curation

Although incorporating increasingly automated steps, metabolic model construction does not dispense with extensive, time consuming and effortful manual curation, itself requiring extensive bibliographic research and experimental validation [72,89]. Manual curation comprises two main aspects. On one hand, assuring network connectivity – from gap-filling to reaction mass balancing, and reversibility and directionality correction – and on the other, a stage on its own, validation – from assessing the model's nutritional requirements to its ability to uptake different carbon and nitrogen sources. *Cpar*MM's construction was divided into two main stages. One initially non-compartmentalised stage, followed by a compartmentalised stage in order to facilitate curation, given the cumulative nature of their respective and resulting network issues.

In total, 408 reactions were manually identified and added, removed or corrected in regard to reversibility, directionality, mass balance and chemical equation. All of these reactions are shown in the supplementary excel file S1. Among these, 83 were mass balanced, 162 were corrected in regard to reversibility or directionality, 163 had their annotation corrected or completed, were added or removed – from which 8 correspond to the manually reconstructed Ubiquinone-9 (Q9) biosynthetic pathway.

3.3.1 Assuring Network Connectivity

The first stage in the manual curation of *Cpar*MM aimed at the assurance of production of all the biomass equation's components by assuring network connectivity. This is, securing every metabolite essential for biomass production results from an uninterrupted set of sequential reactions that assure its biosynthesis and or consumption [81] – figure 8A. This step looks for issues in the network such as gaps resulting from previously incomplete annotation, previously automatically and incorrectly attributed reaction reversibility or directionality, missing precursors, dead-ends and incorrect enzyme annotations – or missing transporters and incorrect compartment annotation as described further on.

First, the production of each biomass component was assessed individually and the identified blocked metabolites were then located in the network using KEGG's mapping functionalities. At this point advantages of nomenclature standardization became apparent. For missing and blocked reaction localization, pathways can be truncated by providing precursor metabolites upstream the metabolite whose production is being assessed. The reasoning being that providing an intermediate metabolite downstream the blocked or missing reaction will result in the final metabolite's production while the precursor entering upstream the blocked or missing reaction will maintain the metabolite's blockage – figure 8C.

Considering for the moment a non-compartmentalised model, several issues might explain a given blocked reaction and this analysis calls for the understanding of metabolites as being part of an interconnected network. The first, more intuitive case, is the possibility of gaps resulting from incomplete annotation for instance, this is, no enzyme association – figure 8B. In this situation, it might be the case that either a given metabolite's respective biosynthetic reactions or precursor producing reactions have not been included in the model – solved adding referenced gene-reaction associations, so called gap-

filling. It might also be the case those same reactions, although in the model, have been incorrectly attributed a given directionality resulting in a blockage within a given pathway – figure 8E. Furthermore, one other possibility are dead-ends. These consist of reactions whose products are not connected to the network and/or cannot be exported, and that consequently will not be predicted to take place given the impossibility of metabolite accumulation – due to the assumption of steady-state [83]. If these reactions do not occur, the reaction upstream will also be blocked, unless its reactants or products are connected to somewhere else in the network with a corresponding non-null flux – figure 8D. In such a situation the metabolite whose production would result in its accumulation has to be assessed in regards either to its *de facto* missing connection to somewhere else in the network or to a missing export reaction.

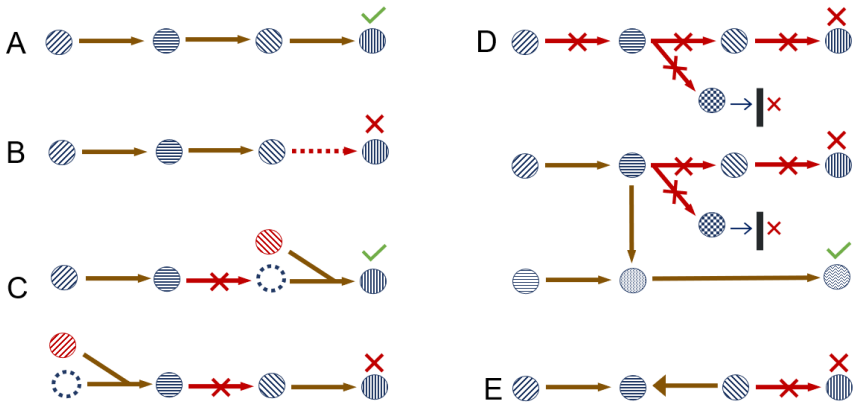

Figure 8: Schematic representation of both the rationale behind missing or blocked reaction localization and possible issues that might result in blocked metabolites. A) A given metabolite's production is made possible by an uninterrupted sequence of reactions; B) A missing reaction results in a blocked metabolite; C) A precursor metabolite provided downstream the non-occurring reaction results in the unblocking of the final metabolite. On the other hand, providing a precursor upstream said reaction maintains blocked metabolite; D) A metabolite headed for a dead-end will impede its synthetic reaction from occurring. If its precursor is not connected to somewhere else in the network its respective synthetic reaction will equally be impeded and so on; E) A reaction with incorrectly attributed directionality will lead to downstream reaction blockage.

However, while the previously described rationale is as valid in a non-compartmentalised *Cpar*MM as in a compartmentalised state, upon compartmentalisation new aspects have to be considered. Compartmentalization corresponds to the process through which the model is instructed on the compartments that harbour each metabolite and enzyme, represented by their respective EC in each specific reaction across the network. While prior to compartmentalization metabolites and ECs coexisted in one single compartment and were consequently available to interact throughout the network, they are now dependent on transport reactions at discrete points in the network. Furthermore, compartmentalization in *Cpar*MM was implemented in an already connected network and consequently **new** gaps result from two main compartment related issues. First, compartmentalization results in given intra-network paths being truncated and having to be reconnected by implementing referenced

transporters – if no automatically annotated transporter already exists - figure 9A to C. This is an additional kind of gap-filling – transporter-based as opposed to the previous gene-based gap-filling. Second, it might also be that given metabolites were incorrectly attributed a given compartment, either down or upstream the true crossing point in the path, resulting in a blocked path – figure 9D. However, it might also happen that a given truncated path may in fact not be solved due to no extant referenceable transporters, redirecting curation to other alternative paths that may contain previously unidentified gene related gaps – figure 9E. Therefore, while all **new** gaps resulting from compartmentalization are compartment related, this step of curation still calls for both kinds of gap-filling.

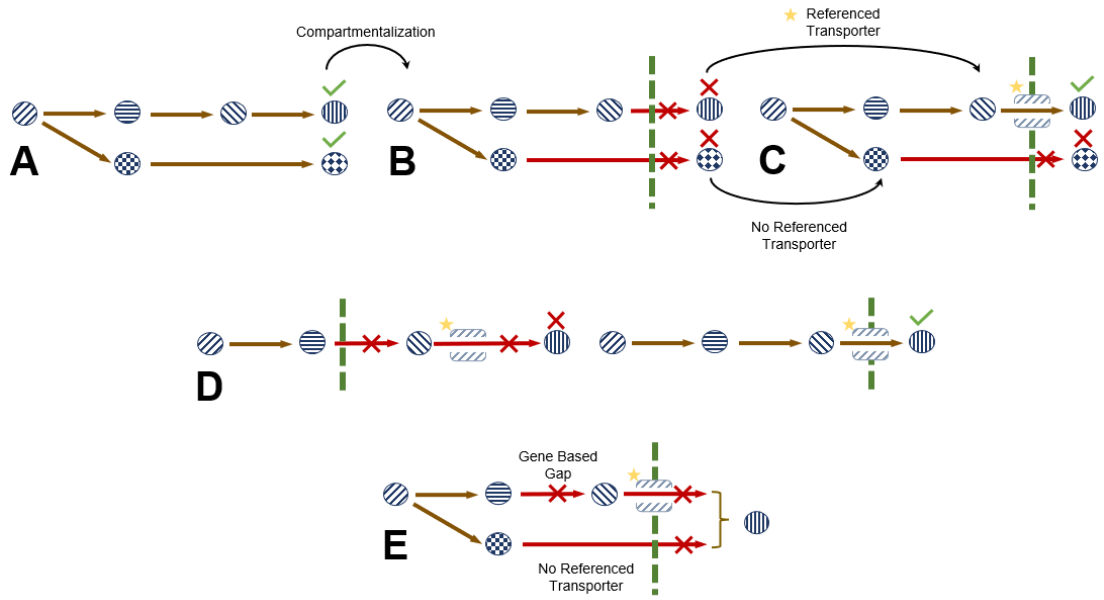

Figure 9: A) Prior to compartmentalization all ECs and metabolites coexist in one single compartment and are consequently available in all reactions throughout the network. B) Compartmentalization results in the separation of given metabolites of given single reactions and results in their blockage – gaps. C) This calls for a transporter-based gap filling, via the implementation of referenced transporter reactions. D) Incorrect metabolite compartment annotation results in the blockage of paths that in fact should not be so. E) Compartmentalization may redirect interest to previously blocked paths containing previously unidentified but solvable gene-based gaps. For means of simplification, metabolites immediately prior to a blocked reaction are assumed to be connected to somewhere else in the network, omitting a possible dead-end situation.

One showing example of what to have in mind when reassuring connectivity in a compartmentalised model is the Tricarboxylic Acid Cycle (TCA). The TCA in *Cpar*MM occurs partially in the cytoplasm and the mitochondria and thus, at some point, certain metabolites will have to cross the mitochondrial membrane. These metabolites are citrate and malate, with the KEGG IDs 00158 and 00149 respectively - mitochondrial entry and export, respectively. The issue though, is that for both these metabolites the most feasibly implementable mitochondrial transporters in all referenced *S. cerevisiae* models are antiporters, and consequently dependent on extant mitochondrial metabolites. Thus, implementing these transporters calls for the consideration of four possibilities – figure 10.

Note though, that biologically compartmentalization necessarily calls for membranes. Even so, and *on par* with other previous abstractions, in this model there is no such thing as a membrane *per se* – cytoplasmic, mitochondrial, peroxisomal and so on - since it only implements compartment information and there is neither a reasonable need for the implementation of a membrane nor does it make sense to do so. A "membrane", then, is defined solely by the set of transporters that allow translocation between the two compartments it supposedly separates. Thus, from this point forward, the mention of any membrane serves only language simplification. The same reasoning goes for transporters themselves, as they are represented simply by sets of metabolite/compartment pairings, much the same as any other reaction.

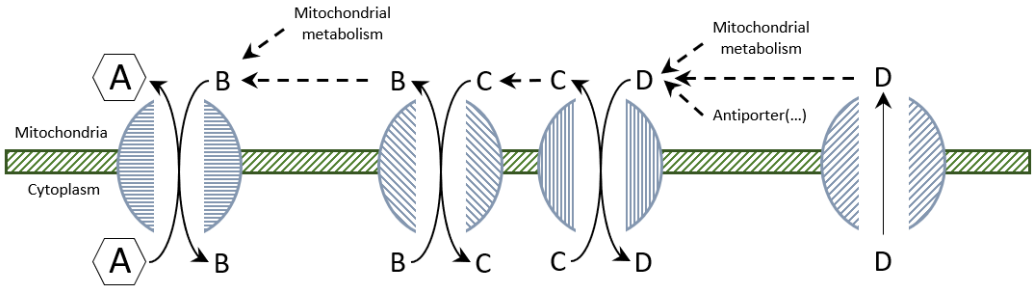

Figure 10: Mitochondrial antiporters rely on extant mitochondrial metabolites. The latter result either from mitochondrial biosynthesis – calling for a reconstructed mitochondrial metabolic network which is absent in CparMM – or import. The metabolite of interest A entries the mitochondria along the antiporter metabolite B. In the absence of B's mitochondrial biosynthesis, it results from mitochondrial import. If the latter is similarly an antiport the issue is incrementally maintained up until the antiporter metabolite is directly imported, represented by metabolite D.

Firstly, the extant mitochondrial metabolites on which the antiport depends may result from mitochondrial metabolism, calling for the time consuming and extensive reconstruction of a mitochondrial metabolic network – a noteworthy effort by itself, at the level of *Cpar*MM's construction. Mitochondrial metabolism in *Cpar*MM consists solely of the mitochondrial segment of the TCA, the Ubiquinone-9 (Q9) Biosynthetic Pathway and the Oxidative Phosphorylation, all of which with little common metabolites besides Q9 or ubiquitous ATP and such. Secondly, it might be that they themselves may be imported into the mitochondria, via whatever mechanisms possible. Third, it might be that the latter is similarly an antiport and the issue is incrementally maintained, with the ideal being that at a certain point transport is direct. If not, it might be necessary to simply implement a chimeric transporter justified by the non-existent mitochondrial metabolic network, and the knowledge that, at least at some level, that given metabolite does have to enter a given compartment – even if dissociating it from that eventual antiporter metabolite. Lastly, the previous issues may be further complicated by the absence of mitochondrial metabolism that upon metabolite import could consume them, and consequently import has to be followed by export. Note once again that in steady-state consumption equals production, resulting in the impossibility of metabolite accumulation – dead-end as shown in figure 8D. This is precisely what happened with citrate, and most evidently with malate.

Malate can exit the mitochondria in two ways – figure 11. On one hand, via an oxaloacetate antiporter (as seen in iMM904 R_MALOAAtp [104]) followed by oxaloacetate mitochondrial export via a proton symporter (as seen in iMM904 R_OAAt2m [104]), resulting in a set of two sequential transport reactions. On the other, malate can aid citrate import by acting as an antiporter metabolite (as seen in iMM904 R_CITtam [104]).

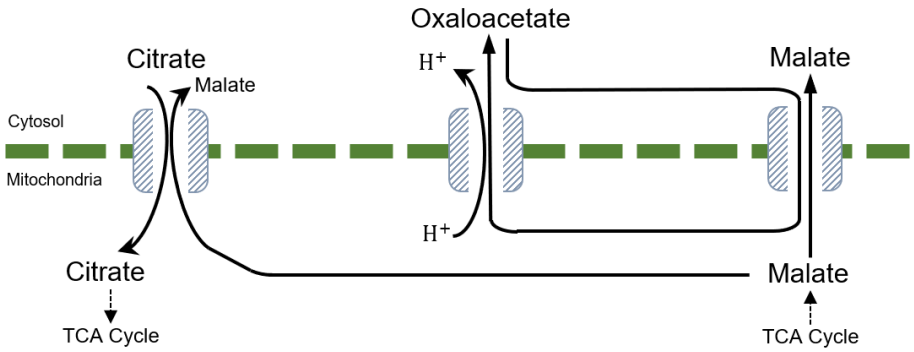

Figure 11: There are two strategies for malate mitochondrial export. On the left, a citrate/malate antiporter. On the right a two-step strategy consisting of an oxaloacetate/malate antiporter followed by oxaloacetate export. On one hand, one strategy turns citrate import dependent on the production of the final metabolite of the mitochondrial fraction of the TCA – malate. On the other, malate export is independent of citrate import, important in the prospect of further reconstruction of mitochondrial metabolism that could hypothetically provide intermediate metabolites

It might seem redundant to have both transport strategies. However, making citrate import and malate export absolutely co-dependent could hinder the TCA cycle further on in the model's curation, in the hypothetical case that mitochondrial metabolism is more fully reconstructed. In this situation intermediate metabolites resulting from other mitochondrial metabolism or transport could hypothetically enter the cycle downstream to citrate, and it is important to allow for the draining of this increased flux – not resulting from citrate import necessarily.

One other relevant example is the Oxidative Phosphorylation Pathway (OP) which is not only highly compartmentalized, but most importantly, dependent on it since it relies on the formation of a proton gradient. Note though that in this model the idea of a gradient does not have the same meaning as a true biological gradient, similarly to what occurs with membranes. Thus, the aforementioned gradient consists simply of the assurance of proton availability in the right compartment (in this case the cytoplasm) and the due connectivity of the pathway itself. The importance then of compartmentalization in the OP pathway is its absolute dependency on assigning correct locations for each proton resulting from each reaction. The Oxidative Phosphorylation pathway in *Cpar*MM was reconstructed as shown in KEGG.

Furthermore, the Oxidative Phosphorylation pathway is also a showing example of how reaction directionality and reversibility possibly are the most important reaction attributes in a metabolic model with no regulatory layer – such as is *Cpar*MM. A practical example of the type issue shown in figure 8E.

3.3.2 Reaction Reversibility and Directionality

All reaction information included in *Cpar*MM upon reaction annotation was retrieved from KEGG, which assumes all reactions to be reversible [81]. The immediate consequence of this is that all metabolites within the network are connected by a given number of reactions, and thus all are ultimately interchangeable. Not only this, incorrect reversibility and/or directionality can lead to blocked pathways or even cases where discrete reactions substitute entire pathways – a showing example of this is the OP pathway, as described further on. Furthermore and in last instance, the set of these issues can compromise the model's own utility in for instance predicting gene and reaction essentiality.

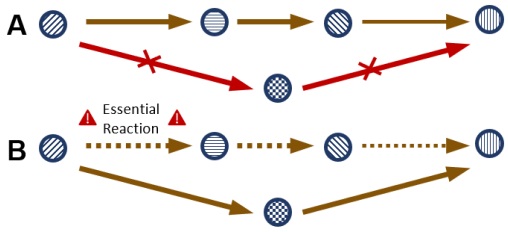

Figure 12: Simplified representation of two possibilities for alternative paths due to a given specific pathway being blocked (A) or to a relatively simpler path being unduly available (B) and compromising reaction essentiality predictions – the latter resulting only due to incorrect reversibility, directionality, or reaction annotation. Solving these issues is made possible by a flux analysis, where the synthesis of individual metabolites can be tracked throughout the network by identifying reactions with non-null fluxes.

Nonetheless, automated curation of reaction reversibility and directionality is not error-proof due to, for instance, information in some instances not being up to date. In total, 162 reactions with incorrectly attributed reversibility or directionality were manually identified and corrected.

As previously described, reaction directionality and reversibility dictate which intra-network paths are available and thus what can and cannot be done – a kind of "static regulation". One initial issue with OP was that although the pathway was in fact available and fully connected, it still had no flux. This was because ATP production was being assured by a series of other discrete reactions annotated with incorrect reversibilities, particularly r03469, which represented a simpler manner of ATP production relative to the 6 sequential reactions that comprise the OP. Upon correction, the OP assured not only a non-null flux, but the highest for ATP production. This is, one single incorrect reversibility made the difference between the fully connected and available OP pathway having a null flux, allowing one single reaction to supply the most of the model's ATP, and the other way around.

However, while r03469 did in fact have an incorrectly attributed reversibility that allowed for ATP production (reversible to irreversible), other reactions such as r01512 can in fact produce ATP. The question then is how come they are not the preferred manner of ATP production, accommodating smaller fluxes than the OP pathway itself. Once again, not only does this relate to correct reaction

directionality and reversibility but also calls for the understanding of each reaction as being part of a network. This is, despite these reactions being connected and available, the flux they can accommodate depends on the reactions to which they are connected – not only upstream reactions that provide reagents, but also downstream reactions due to the model's steady-state. Consequently, in a duly curated network contributions to a given reaction's flux although dynamic, are bound to a given frame – figure 13.

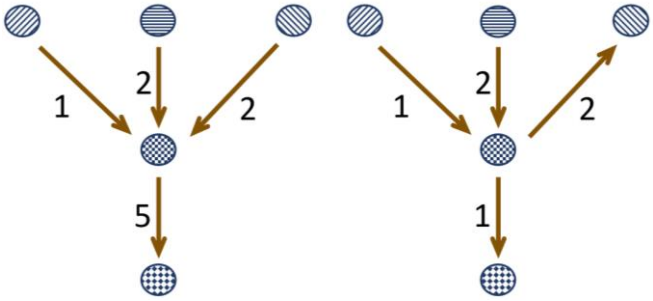

Figure 13: The flux of a given reaction is dependent on the reactions to which it is connected. Correct reaction reversibility and directionality can consequently affect their contribution to that same given reaction's flux. Each hypothetical flux is represented by a given number. The change in one reaction's directionality can significantly reduce the final reaction's flux.

In conclusion, and in liaison with the previous topic 3.3.1, assuring *Cpar*MM's connectivity is a cumulative process in the sense that although at every new added level of information (in this case, compartmentalization) new hypothetical problems arise, the previous ones maintain their validity – transporter-based gap-filling alongside gene-based gap-filling, for instance. The Tricarboxylic Acid Cycle and the Oxidative Phosphorylation pathways sum not only the set of compartmentalization related issues and the rationale behind their solutions, but also highlight the importance of correct reaction directionality and reversibility. From the indissociable idea of steady state, the consequent impossibility of metabolite accumulation and how that relates to the transport strategies at our disposal, to how one incorrect reaction reversibility can substitute an entire central pathway.

3.3.3 A Putative Impact of Non-*C. parapsilosis* Referencing in Gap-Filling

One relevant question which has to be posed at this stage regards the extent in which the systematic use of non-*C. parapsilosis* organisms as references, either for pathway completion or intracellular transporters and metabolite localization, may affect the model's predictive ability. Furthermore, the posing of this putative effect gains an added relevance when noted that one objective of this model's construction is the finding of a coalescence in novel drug targets within *Candida*. Thus, the possibility that the use of other organisms within the genus as references for curation might result in coalescences that are no more than construction artefacts has to be taken into account.

The model contains 2892 reactions, of which 408 have been manually edited – only around 14% of the whole network. Even so, this percentage could be further increased by the removal of blocked reactions. For instance, an outstanding and hypothetical 1000 blocked reactions could be removed, turning the previous 14% to 22% of manually edited reactions. Nonetheless, from these 22% not all correspond to manually added reactions, with around 162 reactions having been corrected in regards to reversibility and 83 to mass-balance. Both mass-balance and reversibility are not necessarily dependent on species, with the latter foremost dependent on thermodynamics - with the exception of some species possibly having specific enzymes with dual activities that turn some reactions reversible. Thus, these 162 plus 83 reactions can be removed turning the previous 22% to around 9%. It becomes then apparent that although the extent of manual curation by itself is significant, its true relative dimension is shown when considered within the greater context of the network. Therefore, at an initial level, referencing non-*C. parapsilosis* organisms directly impacts only around 9% of the network. Furthermore, and even so, these 9% of manually edited reactions do not correspond to fully reconstructed pathways from other reference organisms, but to discrete reactions used in a gap-filling process itself guided by the initial automatic annotation. The idea being that the analysis of the automatically annotated enzymes within a given pathway indicates the intra-pathway path for the synthesis of a given metabolite. This is, within a given pathway the most complete, or least gapped path allowing for the synthesis of a given metabolite is most likely the intended path for that specific synthesis, and gaps are a result not of the path's *de facto* absence but of incomplete annotation. This indication allied to the identification of likely points of entry of precursors to the pathway allows for a gap-filling process that is fairly guided by the annotation that was itself done based on *C. parapsilosis*' genome sequence.

In conclusion, non-*C. parapsilosis* organisms are referenced not only in a relatively small extent but are so referenced in a way that is conservative and most importantly guided by *C. parapsilosis*' genome annotation, and thus not only counter but turn unlikely the possibility that their systematic referencing may lead to indistinguishable models and result in this model's redundancy.

3.3.4 Reaction Mass Balance

Reaction mass balancing is fundamental in the extent that it can lead to the production or consumption of given metabolites out of nothing. For instance, a reaction with a 1:2 carbon unbalance ultimately represents the synthesis of two carbon atoms from one single carbon atom. The sum of these kinds of unbalances over the scope of the model's nearly 3000 reactions may affect the model's biomass and consequently its predictability [73,81,89]. Aspects to consider go from identifying incomplete reaction equations to assessing correct metabolite chemical formulae. In total 83 reactions were manually balanced – see supplementary excel file S1.

One particularly interesting example of reaction balancing in the context of this model was the reaction with the KEGG ID R10686 – equation 2. In *C. albicans* this reaction is involved in Thiamine biosynthesis and is catalysed by Thi5 [111]. The particularity of this reaction arises from the enzyme itself

acting as a reactant, via the lateral chain of an histidine residue and a pyridoxal phosphate bound to a lysine residue in its active site [112] – figure 14.

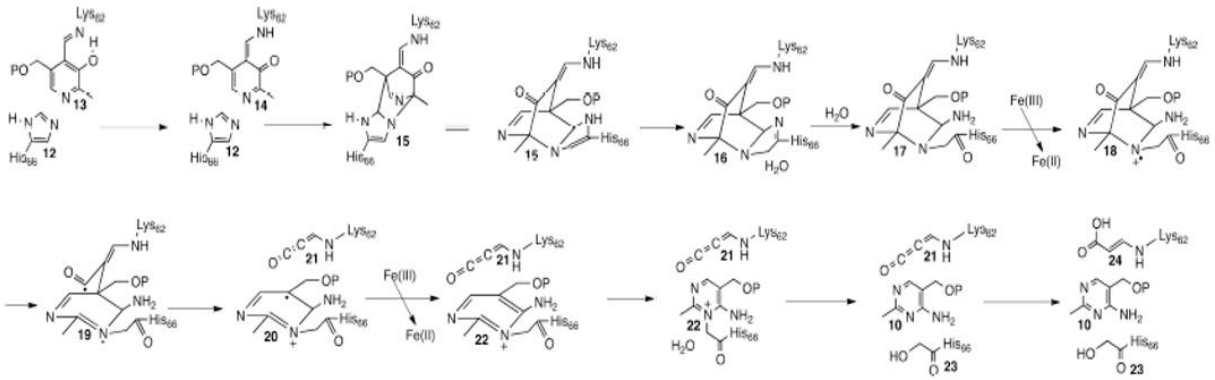

Figure 14: Reaction mechanism for r10686 as shown in Lai et al. 2012 [112].

In the context of this metabolic model, this translates into a reaction whose reactants and products are necessarily not connected to the rest of the network – dead-end, as shown in figure 8D. KEGG simplified this reaction as follows in equation 2:

Pyridoxal phosphate + L-Histidine <=> 4-Amino-2-methyl-5-(phosphooxymethyl)pyrimidine **(2)**

The equation above does not consider the resulting consumed Pyridoxal phosphate and L-Histidine – henceforth named Pyridoxal phosphate residue and L-Histidine residue. On the other hand, while introducing both these residues solves reaction balancing, the issue of lack of a regeneration or consuming reaction and its consequent dead-end nature prevails. Furthermore, it takes the enzymatic histidine for the entirety of the polypeptide chain and considers the pyridoxal phosphate molecule as free from the lysine residue. Therefore, one first possibility would in fact be the restitution of the polypeptidic nature of these residues by considering a chimeric compound with the name R-L-Histidine-Pyridoxal phosphate – where R corresponds to the non-reactant polypeptide chain – as shown in Lai *et al.* 2012 [112]. This would introduce both reactants true chemical formula when bound to the enzyme – figure 14 -, which might balance it. The resulting equation would be as follows in equation 3:

R-L-Histidine-Pyridoxal phosphate <=> 4-Amino-2-methyl-5-(phosphooxymethyl)pyrimidine + R- L-Histidine-Pyridoxal phosphate Residue **(3)**

Although pleasing in the sense that it incorporates the enzymatic nature of both reactants, the issue with this abstraction is the added impossibility of synthesising one molecule of R-L-Histidine-Pyridoxal phosphate since it is an entirely chimeric molecule, nowhere else present in the network. This equation is consequently not a solution. On the other hand, although both are bound to the enzyme, L-

Histidine and Pyridoxal phosphate, both ultimately act as individual reactants, and in the context of this model can be considered as such – since enzymes are implemented only via their EC and not the synthesis of their respective polypeptide chains *per se*. Furthermore, this abstraction would not raise major issues since only the enzymatic residues participate as reactants and not the polypeptide chain itself. Thus, introducing histidine and pyridoxal phosphate as two different reactants solves the issue of reactant connectivity and results in the model's ability to produce them. Therefore, one possibility is the equation as follows in equation 4:

L-Histidine + Pyridoxal phosphate <=> 4-Amino-2-methyl-5-(phosphooxymethyl)pyrimidine + L-Histidine residue + Pyridoxal phosphate residue **(4)**

In this hypothesis, balancing is dependent on the restitution of both molecule's initial structure. This is, figure 14 shows pyridoxal phosphate as being bound by one nitrogen atom where originally there was an oxygen atom – compare with figure 15. Similarly, the amino acid skeleton of L-Histidine has to be incorporated in the equation even though only its side chain acts as a reactant. In this manner, the final chemical formulae of both the L-Histidine and Pyridoxal phosphate residues can be reconstructed. Note that the formulae of both the L-Histidine and the Pyridoxal phosphate residues, as bound to the enzyme are known [112]. Adding three additional water molecules to the reactants, the result is a mass balanced reaction. However, once again, the dead-end is still not solved, due to the resulting Pyridoxal phosphate and L-Histidine residues.

For dead-end solving two main possibilities were pondered. One initial consideration was the construction of a reaction where both the L-Histidine and Pyridoxal phosphate residues would be substituted by chemical formulae homologs that would already be connected to the network – figure 15. This is, each residue would be substituted by other extant metabolites whose atom sum would equal that of the residues. Note that chemical structure is not considered in *Cpar*MM. However, the issue with this possibility was ultimately creating a reaction that would be a chimeric source of given metabolites. This is, creating a reaction that putatively could serve as an alternative synthetic reaction of, for instance, pyruvate, since this was one of the considered homologs. In last instance, this could affect the model's predictability.

The second considered solution was the simple export of the non-connected products – the L-Histidine and Pyridoxal phosphate residues. The current model for *C. albicans*, iRV781 [90] while also implementing the balancing possibility shown in equation 4, solves the dead-end by implementing this second solution. In this model this was also the adopted solution. Note that this possibility is biologically non-justifiable since it is not probable that a given cell would spend energy and resources synthesising a given enzyme only to then discard it. Although requiring certain abstractions, this reaction should emulate as best as possible the original one. The main issue with this solution is then the consequent export of atoms that could in one way or another contribute to biomass formation. The end result is not only a fully chimeric reaction, but a reaction that might implicate a certain predictability cost to the cell's biomass in given conditions, even if only to a slight extent. In the long-term, the separate implementation

of these two solutions might be one interesting target for the assessment of their putative and relative impact on the model's predictability.

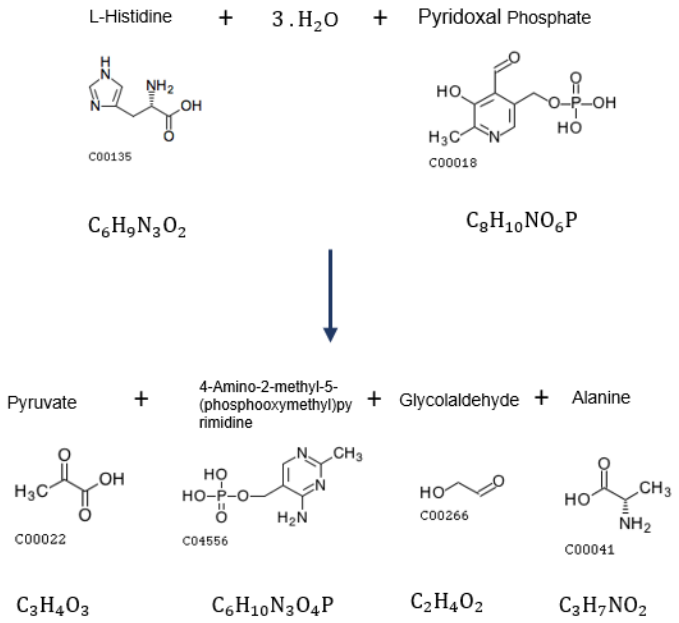

Figure 15: Initially pondered solution for the balancing of r10686. The products are homologs in their chemical formulae to the products of the actual reaction. While being their homologs, in this putative solution they are connected to the network, solving the reaction's dead-end. However, this reaction becomes a chimeric source of relevant metabolites, such as pyruvate for instance. All structures and chemical formulae are retrieved from KEGG. Although in the context of this metabolic model chemical structure is not relevant, in this instance they were included so to facilitate comparison with the structures in figure 14, showing the reaction's actual mechanism and products.

3.4 Validation

In its essence, validation corresponds to the process of assessing and shaping the model's resemblance to an *in vivo* yeast cell. This is, the model's characteristics - from its consumption/production profile to its ability to predict growth or no growth when given certain nutritional sources - are compared to those that have been described *in vivo*, and if need be, corrected. The idea then is that accordance to *in vivo* characteristics is a reflexion of the reliability of the model's predictive capacities.

*Cpar*MM was subjected to two validation steps, both assessing nutritional requirements – a succinct validation step after network connectivity of the non-compartmentalised model was assured, and a more extensive one after compartmentalisation. From these assessments, CparMM was shown to predict biomass production at a nutritional cost that is not only biologically justifiable but coherent with

that of other yeast models – figure 17. Furthermore, *Cpar*MM was also shown to be reliably predictive of the ability to use different compounds as sole carbon or nitrogen sources – with 85% correct predictions from the 34 tested compounds – table 4. Moreover, it was also shown to be accurately predictive of growth from a quantitative perspective, with a predicted specific growth rate of 0.180 h^{-1} (relative to the experimental 0.159 +/- 0.027 h^{-1}) for a glucose consumption rate of 2.098 +/- 0.404 $mmol.gDCW^{-1}.h^{-1}$ – table 8.

3.4.1 The Issue of Consumption vs. Production

Assuring network connectivity and consequent fulfilling of all the biomass equation's components calls for the assessment of the nutritional cost of that connectivity. This is, although all the biomass components are being produced and are consequently connected in one way or another to the network, their production might require *Cpar*MM's consumption and/or production of biologically unjustifiable metabolites for a yeast such as *C. parapsilosis* – consumption of CO_2 along with O_2 production, for instance. It might also happen that given the interconnected nature of the network, biosynthesis of given metabolites is made possible by bypassing specific pathways via alternative parallel reactions whose reversibility and directionality had up to this point not been assessed yet –figure 12. This assessment has to be done via a flux analysis, enabling the tracking of individual paths along the network, since it allows for the identification of reactions with a non-null flux in given conditions. At this stage reactions should be mass-balanced [72,80] since results from this analysis can be affected by cumulative mass-unbalance throughout the network [73].

The simplified consumption/production profile of *Cpar*MM at the first stage of validation, immediately after assuring connectivity of the not yet compartmentalised model is shown in figure 16 - along with the profiles from the iND750 *Saccharomyces cerevisiae* [105] and the iRV781 *C. albicans* [90] models.

*Cpar*MM's profile is markedly different to the profiles of the other two species. One initial and notable incorrection in *Cpar*MM is not only its consumption of CO_2, but also its parallel production of O_2. Additionally, the production of lysophospholipid is also notable. One other relevant observation is the consumption of Fe^{2+} and the production of Fe^{3+}. The latter related to the production of Heme, which was initially incorrectly incorporated as a cofactor essential for biomass. Through its removal from the biomass equation, this issue as well as H^+ consumption was solved. The consumption of biotin, in its turn, while not being present in iND750, results from *Candida* spp. being auxotrophic for this vitamin [68]. The production of O_2 was solved simply by the assessment and correction of reversibility and directionality of reactions where it participated as a reactant.

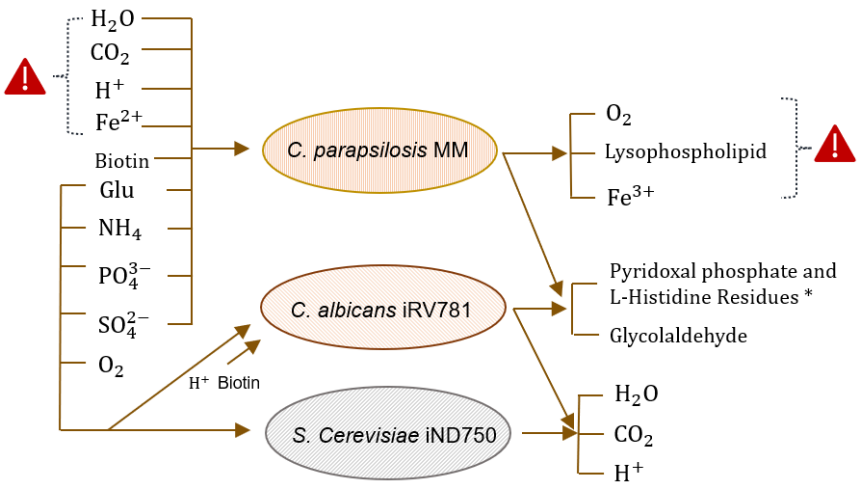

Figure 16: Schematic representation of the consumption/production profiles of the three referenced models. *Cpar*MM's profile is markedly different from both the other, and presents notable issues, such as the consumption of CO2 along with the production of O2. Note also both *Candida* spp. profiles produce Pyridoxal phosphate and L-Histidine residues due to the balancing solution implemented in the reaction with the KEGG ID R10686.

At the first level of validation, prior to any compartmentalization, it might seem paradoxical to try and guide *Cpar*MM's construction by the predictions of models for other organisms, given that one main objective of a GSMM is precisely to identify differences between organisms. Nonetheless, *Cpar*MM at this stage in its construction still needed extensive curation. Many issues, particularly regarding intra-network alternative paths for the synthesis of many metabolites had surely not yet been identified. Moreover, no compartment information had yet been implemented. This implementation, given the consequent separation of metabolites, would likely result in various gaps whose solving would imply the assessment of said alternative paths – which was in fact the case. For instance, lysophospholip production could have initially seemed an interesting differentiating aspect of *Cpar*MM. However, upon further exploration it was understood that its synthesis was only being justified by the need to synthesise one lipid precursor – 1,2-Diacyl-sn-glycerol by the reaction R05333, later corrected from reversible to irreversible. This shows how at this stage, the model possibly had not been curated nor specified in an extent capable of reflecting major species divergences – even if reliably predictive when it came to biomass production under different nutritional conditions, as discussed in 3.4.2.

On the other hand, the compartmentalised *Cpar*MM's final Consumption/Production does not differ from the final simplified profile obtained prior to compartmentalization shown in figure 17, even after the extended and thorough curation it went through afterwards. This is, while initially the similarity of its profile to those of other models could have been argued to result from a still incomplete reconstruction, this was shown to rather reflect the model's early reliability. Furthermore, the question of imposing non-C. *parapsilosis* references systematically on the model putatively resulting in artefacts such as this profile's similarity although valid, is implausible as discussed in 3.3.3.

The final consumption/ production profile of *Cpar*MM resulting from this assessment is shown below – figure 17.

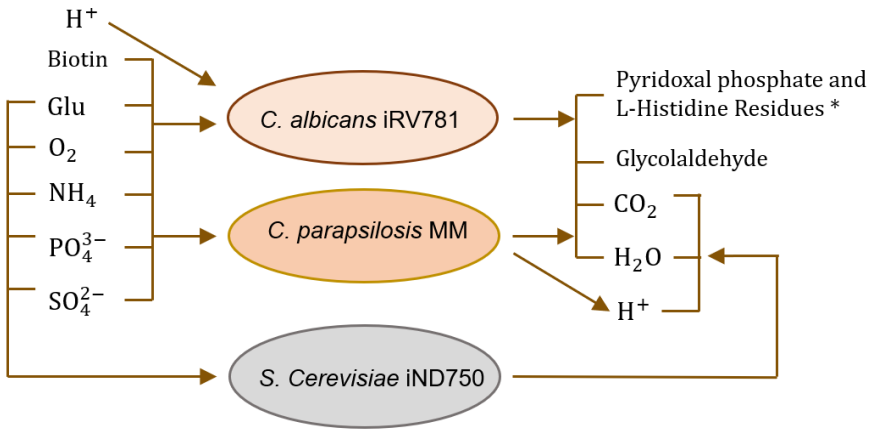

Figure 17: Schematic representation of the consumption/production profiles of the three models, showing their respective similarities and differences. Note both *Candida* spp. models are producing Pyridoxal phosphate and L-Histidine residues, due to the mass balancing solution implemented for the reaction with the KEGG ID R10686. Biotin consumption by both *Candida* models result from *Candida* spp. being biotin auxotrophic.

As previously said, the above shown profiles are simplified, meaning they do not represent their also quantitative nature – this is, their relative compound consumption or production fluxes. These fluxes gain an added relevance when assessing the profile of the latter compartmentalised model since for instance the relative flux of certain pathways had to be adjusted – i.e., the oxidative phosphorylation for instance, as discussed in 3.3.2, which has a direct effect in O_2 consumption. Furthermore, mass unbalanced reactions or particularly incorrections in the relative contributions of each component of the biomass equation will also reflect on these profiles. One example being the cell's amino acid composition directly affecting the model's ammonia consumption, a source of nitrogen indispensable in their synthesis. This meaning that, although the general simplified profile of both the non-compartmentalised and compartmentalised models did not differ, what did significantly change were their relative consumption and production fluxes. All these corrections followed the same rationale as at any other stage of this validation aspect – tracing flux through the network via individual reactions.

3.4.2 Assessing the Predictive Reliability of Carbon and Nitrogen Source Usage

At this stage, not only was *Cpar*MM able to predict the synthesis of all of its biomass components but did so at a biologically justifiable cost, coherent with the general nutritional cost of other yeast models, such as *C. albicans* and *S. cerevisiae*, as discussed previously. Thus, the model's predictive reliability at this stage was further cemented by assessing if when given different sole carbon or nitrogen sources *Cpar*MM behaved as described *in vivo* regarding the ability to produce biomass.

Note that the model's ability to predict or not the use of given nutritional sources results from the availability of given points of entry into the network and of given intra-network paths. These paths may or may not be biologically justifiably available. Thus, a predicted non-null use relative to an observed null use (or vice versa) of a given compound calls for a flux-based analysis as described in 3.4.1.

For some initially contradictory results - literature based *vs.* predicted results – experimental validation was performed by assessing growth in solid media – as described in 2.6.3. The results are shown below in figure 18.

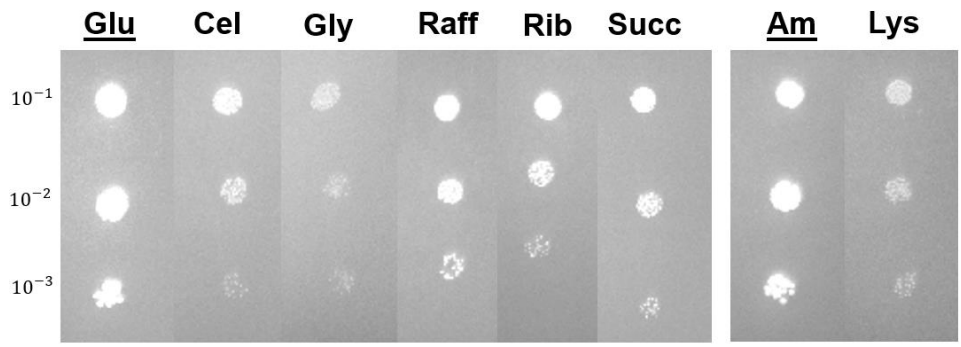

Figure 18: Growth assays on solid YNB medium, 24h incubation at 30ºC, with different sole carbon or nitrogen sources for the experimental validation of contradictory predicted *vs*. literature observed biomass production. **Glu** – glucose (underlined carbon control); **Cel** – cellobiose; **Gly** – glycerol; **Raff** – raffinose; **Rib** – ribose; **Succ** – succinate; **Am** – ammonia (underlined nitrogen control); **Lys** – L- lysine.

One initial observation is that although in different extents *C. parapsilosis* is capable of growth in all the tested conditions. This results in a *de facto* incorrect prediction for raffinose for which *Cpar*MM predicts no biomass. Several reasons might justify this result, from dead-end entry points as shown in figure 8D – due to for instance missing transport or intra-network reactions -, to incorrect reversibilities or directionalities – as shown in figure 8E. For instance, in the case of raffinose although there is in fact an entry point reaction that converts raffinose to sucrose, which in its turn is connected to the network - R01103 with the EC 3.2.1.22 -, this EC was neither annotated for *C. parapsilosis* nor for *C. albicans*. However, contrary to *C. parapsilosis* – figure 18 – raffinose does not seem to support *C. albicans* growth as a sole carbon source [90]. This might reflect a unique metabolic feature of *C. parapsilosis* and a model gap which in the long term might call for further investigation – possibly regarding annotation of underlying reactions' EC numbers. On the other hand, other contradictory results such as that of cellobiose were solvable resulting from simpler issues – such as the addition of a missing extracellular transport reaction.

The case for glycerol as a carbon source not generating biomass might be made, given the significant lesser extent of growth relative to the other platings – figure 18. It could be that growth resulted from residual nutrients from the initial inoculate medium. However, if this was the case the most likely

observation would be that while the first dilution (10^{-1}) would result in faint growth, no growth would be observed for any subsequent dilutions. This is not what is observed for glycerol which, although with an incremental slight decrease, supports growth for all dilutions. Thus, two possible reasons either autonomously or simultaneously might explain this observation – *C. parapsilosis* is relatively less efficient in using glycerol as a carbon source and/or initial inoculate optical density was erroneously lesser.

In total the ability to predict biomass was assessed for 34 different carbon and nitrogen sources simulated within a common minimal synthetic medium – SMM as described in 2.6.1, either substituting the reference glucose or ammonia by the test source. Table 4 shows the different simulation results – biomass *vs.* no biomass.

Table 4: Simulation results for several different carbon and nitrogen sources alongside *in vivo* described data. From the 34 different tested compounds, CparMM correctly predicted growth for 85 %. Biomass production is represented by a plus (+), no biomass production by a minus (-) and prediction disparities are in underlined italic. Experimentally validated sources are noted with an asterisk (*).

Carbon Source	Growth		References	Carbon Source	Growth		References	Nitrogen Source	Growth		References
	in silico	*In vivo*			*in silico*	*In vivo*			*in silico*	*In vivo*	
Glucose*	+	+	[62]	**Ethanol**	+	+	[113]	**Ammonia***	+	+	[113]
Maltose	+	+	[62], [64], [65]	**Methanol**	-	-		**Nitrite**	-	-	
Sucrose	+	+		**Acetic Acid**	+	+		**Nitrate**	-	-	
Lactose	-	-		**Succinate***	+	+	#	**L-Lysine***	+	+	#
Galactose	+	+		**Inulin**	-	-	[113]	***Ethylamine***	-	+	[113]
Raffinose*	-	+	#	**Ribitol**	-	-		**Creatine**	-	-	
Cellobiose*	+	+		**Ribose***	+	+	#	**D-Tryptophan**	-	-	
Galactitol	-	-	[65]	***Mannitol***	-	+	[113]	***Glucosamine***	+	-	
Trehalose	-	-		**L-Arabinose**	-	-		***Cadaverine***	-	+	
Xylose	+	+		**Citrate**	+	+					
Inositol	-	-		**Erythritol**	-	-					
Salicin	-	-	[113]	**Glycerol***	+	+	#				
Arbutin	-	-									

From the 34 tested sources, *Cpar*MM predicts disparate results for five which translates to around 85% correct predictions - and consequently 15% incorrect predictions. Even so, this reflects the already reliable predictive capacity of *Cpar*MM. Note once again the relative extent of manual curation

within the greater universe of the close to 3000 reactions of the model – as described in 3.3.3. Furthermore, it might also be that although unlikely, disparities may result from strain related variability.

Furthermore, and as an interesting additional analysis, beyond qualitative validation – growth *vs.* no growth – these growth assays permit a superficial analysis of *Cpar*MM's quantitative predictive capabilities. As briefly mentioned in 3.4.1, flux-based simulations such as those performed on OptFlux, can not only predict flux distribution throughout the network but also and necessarily relative consumption and production fluxes – one of these corresponding to biomass production. Considering that cell density is proportional to growth, and thus the greater the density the greater the growth. Considering also similar initial inoculate optical densities – as was the case – and similar backgrounds, the relative growth extents shown in figure 18 can be inferred and compared to those predicted by *Cpar*MM.

In this superficial analysis only carbon sources are considered given that curation mainly focused on central carbon metabolism pathways. Firstly from figure 18 the control sources – glucose and ammonia – yield the greater growth. This would be expected since these are, as previously described in 1.3.4, the preferential sources of both carbon and nitrogen respectively. On the other hand, ribose and succinate seem to correspond to the second highest growth extents followed by glycerol. Glycerol in its turn supports the least growth extent. Lastly, cellobiose seems to support a growth extent certainly higher than glycerol but possibly lower than succinate and ribose. Note though that raffinose is excluded from this analysis due to its contradictory result. The expected growth extent order would be something as: Control > Ribose ⇔ Succinate > Cellobiose > Glycerol. Quantitative predicted Biomass production is shown below in table 5.

Table 5: Predicted relative Biomass production for the different experimentally tested carbon sources considering a lower bound of -10 and a synthetic minimal media as described in 2.6.1. Control conditions for carbon or nitrogen sources correspond to the same medium composition and thus the predicted biomass for glucose and ammonia are shown together. With the exception of cellobiose that presents a higher predicted biomass than expected relative to observation in figure 18, quantitative biomass predictions seem to be quite coherent with the experimental observations.

Source Compound	**Control**	**Ribose**	**Succinate**	**Glycerol**	**Cellobiose**
Predicted Biomass	0.8583	0.7152	0.4649	0.4649	1.7166

At first glance results shown in table 5 do seem quite coherent to what was discussed above. Glycerol does present the lowest growth extent while ribose presents the second highest compared to the control. On the other hand, the predicted succinate growth extent seems to be a bit lower than expected – it should certainly be higher than that of glycerol but also the most similar to that of ribose. Even so, it is still within acceptable bounds given that it neither surpasses control conditions nor is it lower than glycerol. However, cellobiose's predicted growth extent – significantly higher than that of the

control - cannot be explained if not for the undue availability of certain intra-network paths that putatively allow for cellobiose's flux contribution for the synthesis of certain biomass constituting metabolites. Note though that however important this quantitative facet might be, in the long term and in a possible range of this model's applications, it does not compromise the credibility of the further on developed essentiality discussion, since it deals in growth *vs.* no growth terms.

Additionally, one relevant observation refers to the two different possible results and their relative differences regarding validation power. In an *in silico/in vivo* pairing order and following table 4's notation, a coherent result can either be +/+ or -/-, being that the latter has the greater value for validation. Note how throughout this discussion one of the major issues with network curation has been the more than frequent issue of reactions incorrectly annotated as reversible and consequently constituting artificially available intra-network paths. This is, more often than not reactions were automatically annotated as reversible and thus the more likely observation is a given path being available. Consequently, a coherent +/+ result has some validation power since it is, once again, a correctly predicted result, but ultimately represents a given paths availability. On the other hand, a -/- result represents a given intra-network path is not available and that that same unavailability most likely results either from correct automatic reaction reversibility annotation or manual curation. With this hierarchy in mind, the results shown in table 4 gain another dimension and most importantly, assurance, since coherent pairings (i.e., correct predictions) not only correspond to 85% of the totality of predictions, as from these more than half are coherent -/- pairings.

In conclusion, not only is *Cpar*MM reliably predictive of different nutritional sources usage, as this reliability allied to its biologically justifiable consumption/production profile and to the previously discussed extent of curation reflects its verisimilitude and credibility.

3.4.3 Assessing *Cpar*MM's Growth Parameters: Glucose Consumption Rate *vs.* Specific Growth Rate

Due to lack of literature data regarding *C. parapsilosis'* specific glucose consumption rate and its respective and resulting growth rate in a minimal synthetic media as described in 2.6.1, these parameters were experimentally determined. As previously described, flux-based simulations can predict not only flux through the network but also the model's relative consumption and production fluxes – one of these corresponding to biomass, i.e. the *in silico* predicted specific growth rate. The main objective of this step being the assessment of how similar the predicted and experimentally determined growth rates are for a corresponding and likewise experimentally determined glucose consumption rate.

Glucose consumption, as the consumption of a carbon source, relates to biomass production which translates into a particular specific growth rate. Thus, these two parameters are strongly intertwined. The more similar the experimental and predicted growth rates are for the same glucose consumption rate, the stronger and more refined the predictive reliability of the model. To a certain extent, this similarity is a measure of the whole network's state of curation. Much more than assessing production *vs.* no production or consumption *vs.* no consumption - which although still immensely

relevant are purely qualitative and rely on certain relatively more isolated intra-network paths - the aforementioned similarity assesses how flux distributes throughout the network and how incorrections can affect its quantitative predictions.

These parameters were obtained as the slopes of two linear regressions [108]. On one hand the specific growth rate resulted from a linear regression of the natural logarithm of the concentration of Dry Cell Weight *vs.* Time – ln[DCW] *vs.* t. On the other, the glucose consumption rate resulted from a linear regression of Glucose concentration *vs.* the ratio of Dry Cell Weight concentration by the previously determined specific growth rate – [Glucose] *vs.* [DCW]/μ. The base values used in determining both parameters are shown in table 6 and the resulting linear regressions in figure 19. Linear regression equations and respective determination coefficients are shown in table 7.

Table 6: Experimentally determined base values for the linear regressions used in determining the growth parameters Glucose Consumption Rate and its respective Specific Growth Rate. Growth was carried out in Synthetic Minimal Media (SMM) as described in Materials and Methods – 2.6.4.

t/h	[DCW]/g/L	ln[DCW]	[Glucose]/mM
0	0.3417	-1.0739	114.8584
2	0.7667	-0.2657	102.4434
4	1.0833	0.0800	98.8520
6	1.3000	0.2624	96.3412
8	1.5083	0.4110	93.7752
10	2.0417	0.7138	91.6444

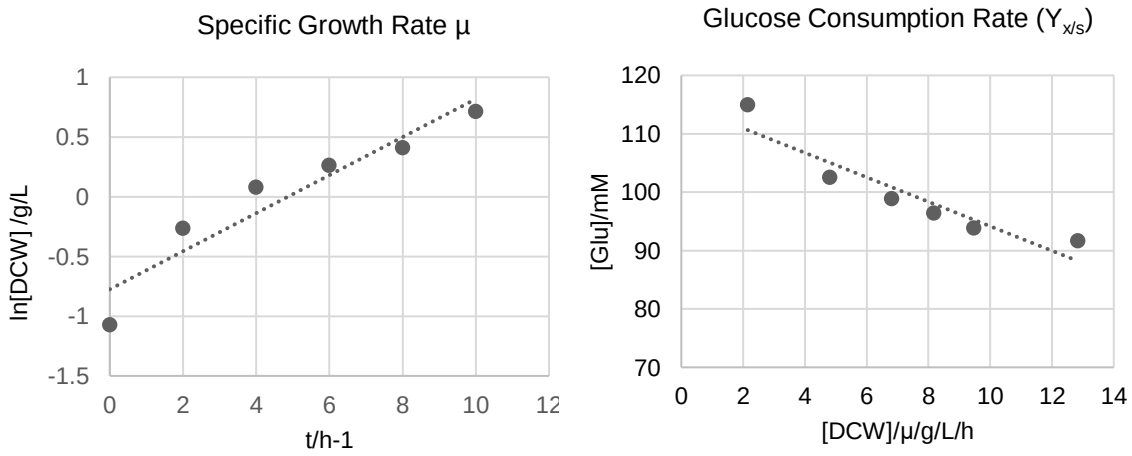

Figure 19: Linear regressions resulting from values shown in Table 6 whose slopes correspond to the growth parameters of interest, specific growth rate (left) – determined to be 0.159 +/- 0.027 h^{-1} - and glucose consumption rate (right) – determined to be 2.098 +/- 0.404 mmol.$gDCW^{-1}.h^{-1}$.

Table 7: Equations and respective determination coefficients of the regressions shown above in figure 19. The growth parameters of interest specific growth rate and glucose consumption rate correspond to the line's slope, here shown to respectively be 0.159 +/- 0.027 h^{-1} and 2.098 +/- 0.404 mmol.gDCW^{-1}.h^{-1}

Specific Growth Rate/h^{-1}	**Glucose Consumption/ mmol.gDCW^{-1}.h^{-1}**
ln[DCW] = 0.1593.t - 0.7752	[Glucose] = -2.0975.[DCW].μ^{-1} + 115.13
R^2 = 0.8995	R^2 = 0.8706

The experimentally determined parameters along with the corresponding *Cpar*MM predicted specific growth rate value are shown in table 8 below.

Table 8: For the experimentally determined glucose consumption rate of 2.098 +/- 0.404 mmol.gDCW^{-1}.h^{-1} CparMM predicts a specific growth rate of 0.180, relative to the experimentally determined rate of 0.159 +/- 0.027 h^{-1}. The predicted growth rate is within the uncertainty interval of the experimentally determined parameter and thus there is no significant difference between both, reflecting CparMM's predicate reliability.

	Experimental	**Predicted**
$Y_{X/S}$/ mmol.gDCW^{-1}.h^{-1}	2.098 +/- 0.404	
μ/h^{-1}	0.159 +/- 0.027	0.180

As shown above, for a glucose consumption rate of 2.098 +/- 0.404 mmol.gDCW^{-1}.h^{-1} *Cpar*MM predicts a specific growth rate of 0.180 h^{-1} relative to the experimentally determined 0.159 +/- 0.027 h^{-1}. This predicted value not only is fairly similar to its experimental counterpart, as it is well within its corresponding uncertainty interval which translates in no significant difference between both. This is, *Cpar*MM is curated to an extent that beyond the qualitative verisimilitude discussed previously, it seems to be strikingly reliable also from a quantitative predictive perspective.

Concluding, at this stage *Cpar*MM is capable of predicting biomass production at a biologically justifiable cost, coherent with that of other validated yeast models, and is reliably predictive of different carbon and nitrogen source usage. Furthermore, it seems to be reliably predictive also from a quantitative point of view. This set of characteristics reflect a duly curated, credible and reliably predictive model ready for the next stage of this project – identifying putative novel antifungal drug targets by network essentiality predictions.

3.5 Predicting and Discussing Gene and Reaction Essentiality

*Cpar*MM's reliability has been assessed and established. The next step is then the obtaining and analysis of *Cpar*MM's essentiality predictions, fulfilling its construction's ultimate purpose. At this stage a list of both essential genes and reactions critical for biomass production was obtained, each of

these representing putative novel anti-fungal drug targets. One initial relevant observation is the distinction between essential genes and essential reactions. Keeping in mind GPR associations, while an essential gene necessarily relates to an essential reaction, the other way around is not necessarily true. This is, an essential reaction may be assured by two or more genes that encode proteins with overlapping activities – or in *Cpar*MM's terms, the same ECs -, and thus present as alternatives in the absence of either one of them. Furthermore, this distinction highlights how the focus of this analysis are the ECs and not the genes themselves, since naturally the genome will not be the target in the context of drug formulation.

A *Cpar*MM simulation in RPMI 1640 medium as described in 2.7 resulted in a total of 129 putative novel targets, from which 87 correspond to essential genes and 42 to essential reactions annotated with multiple genes. The full results are shown in tables A10 and A11 in the annexes.

As discussed previously, incidence has become increasingly dynamic throughout the genus in the past decade. Although *C. albicans* is still generally the most frequent cause of infection, other species such as *C. parapsilosis* and *C. glabrata* have seen relevant rises in incidence, even subverting historical incidence trends in some regions. Furthermore, in some instances infection is of mixed nature, leading to cases of breakthrough candidemia likely due to heterogenous antifungal susceptibility between species [18]. It is then of interest to try and identify putative targets that span these species in order to maximise their applicability in the long term. Thus, focus lays on the intersection of predicted essentiality with that of other published models for *C. albicans* [90] and *C. glabrata* in the same environmental conditions – similarly simulated in RPMI 1640, see table 9. This intersection resulted in a total of 37 common essential ECs, from which 18 are not associated to any active drug or inhibiting compound according to the information retrieved from the DrugBank [110] – representing entirely novel putative drug targets. Interestingly, these interception's results also include a set of essential ECs that have already been validated as essential and have associated in-use drugs – such as *ERG11* and the *FKS* genes, as discussed further on -, conferring an added level of validation of *Cpar*MM's predictive reliability

For the choosing of a set of particularly interesting drug target predictions three general criteria were considered. Firstly and ideally, a drug target should be present in the pathogen and absent in the host. This gains a particularly added relevance when considering a eukaryotic pathogen – *C. parapsilosis* – in an eukaryotic host – human –, reducing the principal source of possible drug toxicity. Secondly, these targets should pertain to already effectively targeted pathways – ergosterol biosynthesis for instance -, pathways directly related to cell structural integrity – fatty acid biosynthesis for example - or related to the synthesis of vital metabolites such as acyl-CoA. Lastly, the predicted targets should ideally already have been associated to pharmacological inhibitors – this is, for organisms other than any *Candida* spp., meaning that those predicted targets have already been validated as effective drug targets in those other organisms.

One relevant observation refers to the lack of *C. parapsilosis'* specific literature regarding *in vivo* described essentiality. In this perspective then, the vast majority of *Cpar*MM's predictions constitute an exciting novel set of results. Furthermore, the aforementioned essentiality predictions refer to specific

conditions of interest that mimic host conditions. Likewise, described *in vivo* essentiality for *C. albicans* often lacks these conditions. Thus, this set of predictions gains an added level of novelty as well.

Table 9: Overlapping essential ECs from the three existing metabolic models for *C. albicans*, *C. parapsilosis* and *C. glabrata*. Essentiality prediction in simulated RPMI medium as described in 2.7. The column "Drug" refers to the existence or absence of compounds effective in inhibitinhg their respective ECs in organisms other than *Candida*.

Gene Name	Human Ortholog	EC	Pathway	Drug
FKS1	#	2.4.1.34	Starch and Sucrose Metabolism	Yes
FOL1	#	2.5.1.15	Folate Biosynthesis	Yes
ABZ1	#	2.6.1.85	Folate Biosynthesis	No
ERG11	*CYP51A1*	1.14.14.154	Ergosterol Biosynthesis	Yes
ERG27	*DHRS11*	1.1.1.270	Ergosterol Biosynthesis	No
ERG26	*NSDHL*	1.1.1.170	Ergosterol Biosynthesis	No
*ERG24 **	*TM7FS2*	1.3.1.70	Ergosterol Biosynthesis	Yes
ERG20	*FDPS*	2.5.1.10	Ergosterol Biosynthesis	Yes
ERG13	*HMGCS*	2.3.3.10	Ergosterol Biosynthesis	No
ERG12	*MVK*	2.7.1.36	Ergosterol Biosynthesis	No
ERG8	*PMVK*	2.7.4.2	Ergosterol Biosynthesis	No
ERG7	*LSS*	5.4.99.7	Ergosterol Biosynthesis	Yes
HMG1	*HMGCR*	1.1.1.34	Ergosterol Biosynthesis	No
IDI1	*IDI1*	5.3.3.2	Ergosterol Biosynthesis	Yes
MVD	*MVD*	4.1.1.33	Ergosterol Biosynthesis	No
CDC21	*TYMS*	2.1.1.45	Pyrimidine Metabolism	Yes
URA1	#	1.3.98.1	Pyrimidine Metabolism	Yes
URA3	*UMPS*	4.1.1.23	Pyrimidine Metabolism	Yes
URA5	*UMPS*	2.4.2.10	Pyrimidine Metabolism	Yes
URA7	*CTPS1*	6.3.4.2	Pyrimidine Metabolism	No
*IMD3 **	*IMPDH2*	1.1.1.205	Purine Metabolism	Yes
ADE1	*PAICS*	6.3.2.6	Purine metabolism	No
ADE12	*ADSS*	6.3.4.4	Purine Metabolism	Yes
ADE13	*ADSL*	4.3.2.2	Purine Metabolism	No
ADE17	*ATIC*	2.1.2.3	Purine Metabolism	No
ADE2	*PAICS*	4.1.1.21	Purine Metabolism	No
ADE4	*PPAT*	2.4.2.14	Purine Metabolism	Yes
ADE5	*GART*	6.3.3.1	Purine Metabolism	No
ADE6	*PFAS*	6.3.5.3	Purine Metabolism	No
ADE8	*GART*	2.1.2.2	Purine Metabolism	Yes
GUA1	*GMPS*	6.3.5.2	Purine Metabolism	Yes
GUK1	*GUK1*	2.7.4.8	Purine Metabolism	Yes
CAB1	*PANK2*	2.7.1.33	Pantothenate and CoA Biosynthesis	Yes
CAB5	*COASY*	2.7.1.24	Pantothenate and CoA Biosynthesis	Yes
PEL1	*PGS1*	2.7.8.5	Glycerophospholipid Metabolism	No
*FAS1 **	#	2.3.1.86	Fatty Acid Biosynthesis	No
ACC1	*ACACB*	6.4.1.2	Fatty Acid Biosynthesis	No

First, the predictive reliability of essentiality in *Cpar*MM can be assessed via three genes– *ERG11*, *FKS* and *ERG24* -, considering also that the list shown above in table 9 results from a coalescence with two published and thus validated models, adding to its credibility.

As discussed in 1.3.3, azoles and echinocandins are the current first-line anti-fungals used in treating *Candida* spp. infections. On one hand, azoles target the Lanosterol 14-α-Demethylase involved in ergosterol biosynthesis and are encoded by the *ERG11* gene. On the other, echinocandins target cell wall biosynthesis by inhibiting the β-(1,3)-D-glucan synthase and are encoded by the *FKS*1, *FKS2* and *FKS3* genes. Furthermore, these genes' corresponding reactions are predicted as essential in both *C. albicans* and *C. glabrata* models. Consequently, the fact that they are predicted as essential by *Cpar*MM represents a strong validation of the predictive power of the model herein constructed and described.

Moreover, beyond *ERG11* and *FKS*, *ERG24* can also be conferred some validation power over the model's predictive reliability. *ERG24* codes the sterol C-14 reductase and participates in ergosterol biosynthesis – highlighting one of the previously mentioned criteria, that priority should be given to pathways containing already validated targets. Its encoded protein has been shown to be inhibited by morpholines in *C. albicans* being that one of these, amorolfine, is used clinically - although restricted to topical use due to quick metabolization by the host [114,115].

A first set of foremost interesting genes are the *FOL1*, *ABZ1*, *FAS1* and *URA1*genes, given the absence of respective human orthologs. Note that the absence of a human ortholog although not an excluding factor, is a preferable attribute since this translates into lower chances of host drug toxicity and may allow for greater freedom of drug design. Adding to this set are the *FAS2, URA3, URA4, URA5,* and *URA7* genes, given the close functional and/or network proximity to the initially presented set as discussed further on.

Fungi and consequently *Candida* spp. rely on folate *de novo* biosynthesis given their inability to uptake folate from the environment [116]. Targeting these kinds of metabolic chokeholds and thinking from an auxotrophy inducing perspective represents a straightforward way of compromising the cell's growth. Both *FOL1* and *ABZ1* encode proteins responsible for two folate precursor synthesising reactions. Furthermore, these genes do not present human orthologs since human metabolism does not synthesise folate, relying on diet derived folate [116]. *FOL1* encodes a dihydropteroate synthase (DHP) with the EC 2.5.1.15, annotated to the reaction R03067. From this reaction results dihydropteroate – KEGG ID c00921 -, a folate precursor with R03067 as its only biosynthetic reaction. In its turn, *ABZ1* encodes a para-aminobenzoate synthetase with the EC 2.6.1.85 annotated to the reaction R01716. Similarly to R03067, R01716 corresponds to an unique synthesising reaction of another folate precursor – 4-amino-4-deoxychorismate with the KEGG ID c11355. Furthermore, folate results from one single branch of its biosynthetic pathway – figure 20 - and thus it is expected that unique biosynthetic reactions be predicted as essential - given the resulting lack of alternative biosynthetic paths.

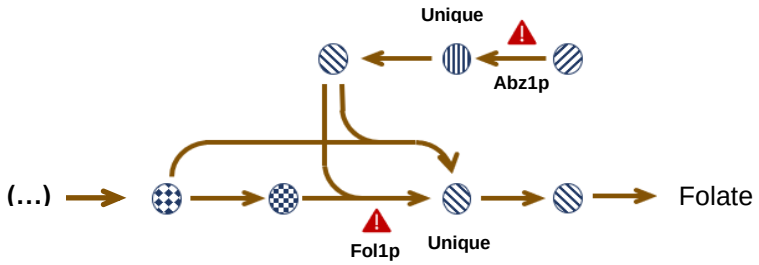

Figure 20: Folate synthesising branch of the Folate Biosynthesis Pathway. Given its isolation from the rest of the pathway, unique precursor synthesising reactions will be predicted as essential, considering furthermore the inability to import extracellular folate that could otherwise complement the resulting auxotrophy.

The dihydropteroate synthase encoded by *FOL1* has been shown to be successfully inhibited by antifolates such as sulfones in a series of microorganisms – from *Escherichia coli* to *Plasmodium falciparum* [110,116]. However, antifolate therapy for *Candida* infections is not particularly effective considering current antifolate compounds [117]. In fact, for *C. albicans* only sulfanilamide is used clinically, although restricted to topical use [118]. Given the efficacy of antifolates in treating infections by other etiologic agents, this might present the opportunity to design new effective antifungal compounds.

In its turn the para-aminobenzoate synthetase encoded by *ABZ*1 does not seem to have any assigned drug neither for *Candida* nor for other organisms, and in that sense represents a fully novel putative drug target. On one hand, this putative target might present an alternative to *FOL1* in the possibility of no effective Fol1p targeting compounds. On the other, this might also present the possibility of a combined targeting strategy, referring back to the idea that multiple and simultaneous targeting, although in smaller extents, might be relatively more effective – as discussed in 1.4.2. This is, it could be that this combined therapy could overcome the current inefficacy of antifolate compounds for treating *Candida* spp. infections.

Fatty acids as the base constituents of membranes are major structural elements of the cell. Much like the reasoning behind impairing ergosterol biosynthesis for instance, impairing fatty acid biosynthesis can compromise the cells structural integrity and survival [119]. However, on one hand while ergosterol is a major structural component in fungal membranes, in human hosts cholesterol functions as its analogue, and it is in these substrate differences that drug design relies. On the other hand, drug design for targeting fungal fatty acid biosynthesis has to fundamentally rely on differences in enzymatic structure – note how in contrast *ERG11* has a human ortholog. Fatty acid synthases (FAS), responsible for fatty acid biosynthesis as the name pre-empts, show significant structural variability between fungi and mammals – respectively, the fungal FAS corresponds to an heterododecamer and the human FAS corresponds to homodimers of comparatively smaller polypeptidic chains [120,121]. This structural difference may then form the basis of putative novel antifungal drug design. The *FAS1* gene – see table 9 - is tightly connected to the *FAS2* gene, together encoding the fungal fatty acid synthase complex [121] - the latter predicted as annotated to a series of essential reactions, gene ID CPAR2_807400, see table A10 in the annexes. Moreover, *C. parapsilosis FAS2* deletion mutants have been shown to result not

only in auxotrophy for certain fatty acids but most importantly in reduced ability to form biofilms [121] – the major characterizing virulence factor in *C. parapsilosis* infections as discussed in 1.3.2. These observations along with the close functional roles of both Fas1p and Fas2p, present the FAS complex as a promising putative novel drug target.

In their turn, additionally to the foremost *URA1* which has no human ortholog, *URA3*, *URA5* and *URA7* have also been predicted to be essential for the three considered *Candida* spp. – table 9. Adding to these is *URA4* predicted as essential in *C. albicans* and *C. parapsilosis* for which effective drugs have been identified for *Escherichia coli*, for instance [110]. Considering the information retrieved from DrugBank[110], from these, Ura1p, Ura5p and Ura3p have already been identified as targetable in other organisms - reflecting their viability as possible new targets in *Candida* - while Ura7p has no associated drug, highlighting its added novelty as a putative drug target. In *Cpar*MM Ura1p, Ura5p and Ura3p – with the respective ECs 1.3.98.1, 2.4.2.10 and 4.1.1.23 – are annotated to three consecutive reactions from one isolated branch of the pyrimidine biosynthesis pathway that results in the synthesis of UMP – KEGG ID c00105 –, which in its turn corresponds to the common precursor of most of the pathway – figure 21. In the absence of this UMP synthesising branch, UMP can be synthesised from extracellular incorporated uridine – via several possible ECs annotated in *Cpar*MM. Given their relative positions in the network, these genes are then predicted to become essential in the absence of uridine – supported by the *in vivo* observation of uridine auxotrophy in *C. albicans URA3* mutants [122]. In its turn *URA4* with the EC 3.5.2.3 is annotated to R01993 which is likewise located consecutively to the previously discussed reactions, justifying thus its inclusion in this set. Furthermore, the gene *URA7* is annotated to EC 6.3.4.2, which corresponds to a CTP synthesising reaction – KEGG ID c00063 – which connects the UMP synthesising branch of the pathway to the pathway itself. Thus its essentiality is justified similarly to that of the previously discussed *URA* genes, the difference being that Ura7p is not annotated to a reaction consecutive to the previously described reactions.

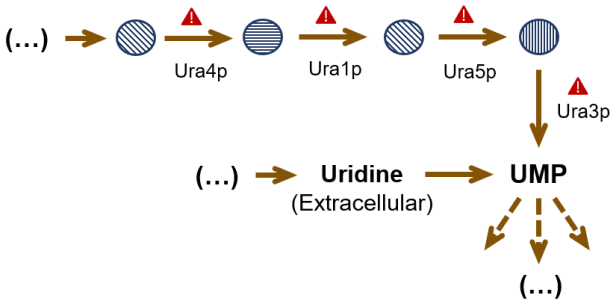

Figure 21: The four predicted URA genes are annotated to consecutive reactions. Furthermore, these reactions comprise an isolated branch that culminates in the sinthesis of the important precursor UMP. Similarly to what was discussed for folate biosynthesis, these unique reactions will be predicted as essential. Furthermore, their added interest stems for the consecutive position of the reactions which could hypothetically guide effective combined therapies.

Concluding, the interest of the *URA* genes comprises not only the novelty of these predicted targets as isolated discrete putatively effective targets, but also the possibility of simultaneous targeting. This is, similarly to *FOL1* and *ABZ1*, and once again given the relative positions of these genes' corresponding reactions, this set of predicted essentiality may represent the possibility of combined and simultaneous targeting – be it via the designing of one, although unlikely, transversally active drug or combined therapies that would integrate a series of drugs that would otherwise not be as effective. In favour of this combined therapy is once again the relative position of these genes' encoded proteins. Note how the four predicted essential ECs are positioned sequentially and how metabolic flux inhibition would be cumulative the further downstream a protein is.

In its turn, a second set of interesting predicted essential genes common to the three considered species' models are *CAB5* and *CAB1*, and the *ERG7*, *ERG26* and *ERG27* genes. Respectively, these putative targets' interest is justified by the two previously presented criteria that putative targets should pertain to pathways that synthesise important and widespread biomass precursors – such as Coenzyme A – or to already effectively targeted pathways – ergosterol biosynthesis for instance.

Coenzyme A (CoA) is a major if not the major cosubstrate of many biosynthetic reactions across the cell's metabolism - acting as an acyl group carrier for, among others, fatty acid biosynthesis whose role in cellular integrity was previously discussed. Inhibiting CoA and subsequent Acyl-CoA biosynthesis represents a good manner of inducing general metabolic impairment. The gene *CAB1* annotated to the EC 2.7.1.33 encodes the enzyme Pantothenate Kinase (PanK). The PanK catalyses the first step of CoA biosynthesis from pantothenate, which can be acquired from the environment [123] – in *Cpar*MM this conversion corresponds to R03018. Note that in the simulated conditions of interest environmental pantothenate is provided and thus CoA synthesis related essentially is expected only downstream of pantothenate. This is in fact what is observed, given that similarly to what was described for the *URA* genes in *Cpar*MM CoA results from one isolated branch of the Pantothenate Backbone Biosynthesis Pathway in which pantothenate is the main precursor – from which results R03018's essentiality. In its turn, *CAB5* annotated to the EC 2.7.1.24 corresponds to the last step of this CoA synthesising branch – R00130 – being that its essentiality results from the same reasoning. The targeting of PanK by a range of compounds – pantothenamides - has already been described as effective in infections by *Plasmodium falciparum* for instance [124]. However, the targeting of the last step of CoA biosynthesis seems to be relatively more novel, in that according to DrugBank it has only one described and most importantly putatively effective inhibiting compound for *Escherichia coli*. Nonetheless, both these putative targets represent the fundamental possibility of inducing systemic metabolic impairment by targeting one specific metabolite's synthesis. Furthermore, they have already been described as targetable even if to different extents. In this sense they represent a pair of quite relevant and interesting predicted targets.

Lastly, the *ERG7, ERG26* and *ERG27* genes reflect how trying to identify alternative putative targets within an already effectively targeted pathway may be a good starting point for novel drug target identification. Note once again how *ERG11* pertains to the same pathway as these genes - ergosterol biosynthesis - and how it corresponds to the target of one of the two currently used front line antifungals as discussed previously. In opposition to this criteria, the argument for designing modified drugs aimed

at the already identified effective targets in order to counteract observed resistance mechanisms may be made. This has been a frequent line of action in the designing of, specifically, antibiotics – cephalosporine as an alternative to penicillin for instance. In the case of pathogenic *Candida* spp. *ERG11* could be the target of this kind of reasoning. However, these kinds of approaches seem to not only pose as short-term solutions but may present increasing complexity given their intent on counteracting ever developing resistance mechanisms. In this context then, ideally focus should lay on entirely new putative targets.

The *ERG26* and *ERG27* genes encode respectively a C-3 sterol dehydrogenase (C-4 decarboxylase) and a 3-keto reductase [125,126]. These two enzymes catalyse two consecutive reactions in the ergosterol synthesising branch of the pathway – respectively R07494 and R07495 in *Cpar*MM. The ergosterol biosynthesis pathway consists of one single branch where most reactions are unique. Similarly to what was discussed for the *URA* genes, this results in most of these reactions being predicted as essential. This inference is coherent with *in vivo* observations, being that both the Erg26p and the Erg27p have been shown to be essential in *C. albicans* [125,126]. Furthermore, inhibition of Erg27p has been shown to result in an inhibition of Erg7p at the beginning of the pathway and resulting thus in the impossibility of synthesis of ergosterol precursors early on in the pathway [126]. This observation establishes Erg27p as a rather interesting putative target, since its inhibition is known to not only block the pathway in its discreet position but further upstream. Additionally, according to information retrieved from DrugBank, neither Erg26p nor Erg27p have any effective described inhibitors, presenting these putative targets as fairly novel. On the other hand, according to DrugBank, Erg7p is a target of oxiconazole, a clinically approved drug although restricted to topical use.

*Cpar*MM has reliably predicted a significant set of fairly interesting and novel putative drug targets. In total, the coalescence of predicted essentiality in simulated RPMI for the models of the three main pathogenic *Candida* spp. has resulted in 37 common predicted essential putative targets from which 18 do not seem to have any associate described drug - representing entirely novel predictions. From these common predicted targets, 17 seem to be quite promising not only because they represent putative novel drug targets that satisfy the three main previously discussed criteria, but in the sense that they may allow for paradigm changes when it comes to drug design – from single targeting to the reasoning of new combined therapies.

4. Conclusions and Perspectives

The herein described metabolic model for *C. parapsilosis* corresponds to a compartmentalised model comprising 4 compartments – Extracellular, Cytoplasm, Mitochondria and Peroxisome -, 2892 reactions, 2885 metabolites and 1112 genes and 598 proteins.

This is the result of one year of extensive literature research, manual curation and of reasoning and thinking of new solutions for evermore varied issues. From this work resulted a reliably predictive model. *Cpar*MM is capable of predicting biomass production at a nutritional cost that is not only biologically justifiable but coherent with the cost observed for other published yeast models – namely *C. albicans* and *S. cerevisiae*. It is also reliably predictive of the ability to produce biomass using different compounds as sole carbon or nitrogen sources - from the 34 tested sources, 85% corresponded to correct predictions. Moreover, from a quantitative perspective, for an experimentally determine glucose consumption rate of 2.098 +/- 0.404 $mmol.gDCW^{-1}.h^{-1}$ *Cpar*MM predicts a specific growth rate of 0.180 h^{-1} relative to the experimentally determined and corresponding 0.159 +/- 0.027 h^{-1}. The predicted value being within the uncertainty interval reflects how there is no significant difference. This quantitative predictive accuracy in its turn reflects the fine tuning of the network's curation since it quantitatively indicates how flux can be distributed throughout the model.

*Cpar*MM predicted a set of 37 essential ECs common to those predicted for two other major pathogenic *Candida* spp. – *C. albicans* and *C. glabrata* -, each of these predictions representing putative novel drug targets. From these 37 predicted putative targets 18 were found to be entirely new, with no corresponding assigned drug or inhibiting compound for any species. Note once again the epidemiological context in which this model's construction and this essentiality prediction has been done. Since the 1990's incidence within *Candida* has become increasingly dynamic. *Candia* spp. other than the historically more frequent *C. albicans* have seen significant rises in incidence – namely *C. glabrata* and most relevantly *C. parapsilosis*. Many factors may have been behind these shifts, although heterogeneous resistance to frontline antifungals may be the most likely. It is then of interest to try and identify putative novel drug targets that are effective across the genus so to maintain their applicability in the long term. Thus the coalescence of the predicted essentiality for these three yeast models comprising a predicted total set of 37 putative targets appears quite promising.

Two sets of predicted putative target encoding genes surface as of significant interest, the *FOL1* and *ABZ1* along with the *CAB1* and *CAB5* genes. *FOL1* and *ABZ1* pertain to folate biosynthesis, and inhibition of Fol1p or Abz1p results in a folate auxotrophy that cannot be complemented due to *Candida* spp. inability to uptake folate from the environment. Note though that in contrast human hosts do not rely on *de novo* synthesised folate, but on diet derived folate. Moreover, antifolate compounds have been identified as effective therapies in infections by other organisms, although currently not reliable options for treating *Candida* spp. infections. This observation presents then the exciting possibility of designing new more effective drugs – with the added positive aspect of possibly developing new combined therapies targeting both these proteins even if to lesser extents. In their turn, *CAB1* and *CAB5*

pertain to Coenzyme A biosynthesis, one of the most important biosynthetic precursors of the cell. In this sense they reflect the possibility of inducing systemic metabolic impairment while targeting one specific metabolite's synthesis. In fact, these two gene's encoded proteins have been described as putatively effective drug targets in other species, namely *E.coli.* The next great stage in assessing the targetability of these predicted putative targets will be the implementation of docking methodologies as described in 1.4.3 – which in fact are currently and promisingly being carried out for *FOL1*.

Beyond the pharmacological applicability of *Cpar*MM, such a model could also be used in predicting putative industrially relevant features or characteristics that may shed light on new diagnostic techniques in the form of unique ECs. Note that curation up to this point focused mainly on central and sometimes ubiquitous pathways directly involved in the synthesis of biomass precursors, thus difficulting the identification of putatively unique ECs – this is, considering ECs annotated to connected reactions. However, *Cpar*MM still contains a somewhat considerable number of unconnected reactions – blocked or gap reactions - that resulted from the initial automatic annotation. Their non-connectivity may result from several reasons as discussed previously. Nonetheless, two reasons support their removal not yet having been performed. On one hand, blocked reactions do not affect in any way whatsoever the model's predictability since they cannot accommodate flux. On the other and most significantly, although the model is already at a quite significant stage of curation and its predictive reliability has been firmly established, curation is a continuous process and it might be the case that some of these previously blocked reactions may in the long term be connected – given the constant generation of new biological information and the possibility for the model's constant updating. This would then result in the incorporation of pathways other than those directly essential for biomass production that could in their turn implement truly unique features of interest. Furthermore, one current limitation with the used basilar software – *merlin* [89] – is its inability to differentiated the connectivity of duplicated reactions from different compartments. This is, while one same duplicated reaction may be connected in one compartment, if its duplicate is blocked *merlin* will identify both as being blocked. Adding to the putative manual addition of these blocked ECs' respective reactions, this software limitation further calls for the manual filtering and identification of truly blocked reactions that can then similarly be assessed in regard to possibly being connected to the network.

For the moment this is the greatest issue yet to be resolved in this model's construction. *Cpar*MM at this point of its construction cannot yet make sustained predictions of unique ECs. Not because it does not, nor will not contain any, but because curation although extensive in itself has not yet focused on peripheral pathways - that if connected will not only complete the model but represent added metabolic diversity. Manually assessing and possibly connecting these pathways will undoubtedly correspond to the next great stage of *Cpar*MM's construction.

With this in mind *Cpar*MM presents then as one of many steps forward in establishing GSMMs as important tools for the future.

Bibliography

1. Silva, S. *et al.* Candida glabrata, Candida parapsilosis and Candida tropicalis: Biology, epidemiology, pathogenicity and antifungal resistance. *FEMS Microbiol. Rev.* **36**, 288–305 (2012).

2. Miramón, P., Kasper, L. & Hube, B. Thriving within the host: Candida spp. interactions with phagocytic cells. *Med. Microbiol. Immunol.* **202**, 183–195 (2013).

3. Ghannoum, M. A. & Abu-Elteen, K. H. Pathogenicity determinants of Candida. *Mycoses* **33**, 265–282 (1990).

4. Yapar, N. Epidemiology and risk factors for invasive candidiasis. *Ther. Clin. Risk Manag.* **10**, 95–105 (2014).

5. Fridkin, S. K. The changing face of fungal infections in health care settings. *Clin. Infect. Dis.* **41**, 1455–1460 (2005).

6. Lass-Flörl, C. The changing face of epidemiology of invasive fungal disease in Europe. *Mycoses* **52**, 197–205 (2009).

7. Kojic, E. M. & Darouiche, R. O. Candida Infections of Medical Devices. *Clin. Microbiol. Rev.* **17**, 255–267 (2004).

8. Trofa, D., Gácser, A. & Nosanchuk, J. D. Candida parapsilosis, an emerging fungal pathogen. *Clin. Microbiol. Rev.* **21**, 606–625 (2008).

9. Almirante, B. *et al.* Epidemiology, risk factors, and prognosis of Candida parapsilosis bloodstream infections: Case-control population-based surveillance study of patients in Barcelona, Spain, from 2002 to 2003. *J. Clin. Microbiol.* **44**, 1681–1685 (2006).

10. Pfaller, M. A. Editorial Response: The Epidemiology of Invasive Mycoses— Narrowing the Gap. *Clin. Infect. Dis.* **27**, 1148–1150 (1998).

11. Pfaller, M. A. & Diekema, D. J. Epidemiology of invasive candidiasis: A persistent public health problem. *Clin. Microbiol. Rev.* **20**, 133–163 (2007).

12. Nucci, M. & Marr, K. A. Emerging fungal diseases. *Clin. Infect. Dis.* **41**, 521–526 (2005).

13. Spampinato, C. & Leonardi, D. Candida infections, causes, targets, and resistance mechanisms: Traditional and alternative antifungal agents. *Biomed Res. Int.* (2013) doi:10.1155/2013/204237.

14. Koehler, P. *et al.* Morbidity and mortality of candidemia in Europe: an epidemiologic meta-analysis. *Clin. Microbiol. Infect.* **25**, 1200–1212 (2019).

15. Tortorano, A. M. *et al.* Candidemia in Europe: epidemiology and resistance. *Int. J. Antimicrob.*

Agents **27**, 359–366 (2006).

16. Toda, M. *et al.* Population-Based Active Surveillance for Culture-Confirmed Candidemia - Four Sites, United States, 2012-2016. *MMWR. Surveill. Summ.* **68**, 1–15 (2019).

17. Invasive Candidiasis Statistics. https://www.cdc.gov/fungal/diseases/candidiasis/invasive/statistics.html. Consulted in November, 2020

18. Cuervo, G. *et al.* Breakthrough candidemia in the era of broad-spectrum antifungal therapies. *Clin. Microbiol. Infect.* **22**, 181–188 (2016).

19. Trick, W. E., Fridkin, S. K., Edwards, J. R., Hajjeh, R. A. & Gaynes, R. P. Secular trend of hospital-acquired Candidemia among intensive care unit patients in the United States during 1989-1999. *Clin. Infect. Dis.* **35**, 627–630 (2002).

20. Pappas, P. G. *et al.* Guidelines for treatment of candidiasis. *Infect. Dis. Clin. Pract.* **12**, 245–246 (2004).

21. Tavanti, A., A. D. Davidson, N. A. Gow, M. C. Maiden, and F. C. O. Candida orthopsilosis and Candida metapsilosis spp. nov. To Replace Candida parapsilosis Groups II and III. *Definitions* **43**, 284–292 (2005).

22. Girmenia, C. *et al.* Rising incidence of Candida parapsilosis fungemia in patients with hematologic malignancies: Clinical aspects, predisposing factors, and differential pathogenicity of the causative strains. *Clin. Infect. Dis.* **23**, 506–514 (1996).

23. Fleck, R., Dietz, A. & Hof, H. In vitro susceptibility of Candida species to five antifungal agents in a German university hospital assessed by the reference broth microdilution method and Etest. *J. Antimicrob. Chemother.* **59**, 767–771 (2007).

24. Pfaller, M. A. *et al.* International surveillance of bloodstream infections due to Candida species: Frequency of occurrence and in vitro susceptibilities to fluconazole, ravuconazole, and voriconazole of isolates collected from 1997 through 1999 in the SENTRY antimicrobial surv. *J. Clin. Microbiol.* **39**, 3254–3259 (2001).

25. Costa-De-Oliveira, S., Pina-Vaz, C., Mendonça, D. & Gonçalves Rodrigues, A. A first Portuguese epidemiological survey of fungaemia in a university hospital. *Eur. J. Clin. Microbiol. Infect. Dis.* **27**, 365–374 (2008).

26. Faria-Ramos, I. *et al.* Species distribution and in vitro antifungal susceptibility profiles of yeast isolates from invasive infections during a Portuguese multicenter survey. *Eur. J. Clin. Microbiol. Infect. Dis.* **33**, 2241–2247 (2014).

27. Faria-Ramos, I. *et al.* Erratum to: Species distribution and in vitro antifungal susceptibility profiles of yeast isolates from invasive infections during a Portuguese multicenter survey (Eur J Clin Microbiol Infect Dis, (2014), 33, 12, 2241–47, 1007/s10096-014-2194-8). *Eur. J. Clin. Microbiol.*

Infect. Dis. **34**, 2137 (2015).

28. https://www.vectorstock.com/royalty-free-vector/blank-outline-map-europe-simplified-wireframe-vector-13703065. Consulted in January, 2021

29. Bliss, J. M. Candida parapsilosis: An emerging pathogen developing its own identity. *Virulence* **6**, 109–111 (2015).

30. Enger, L. E. E. *et al.* Cloning and Characterization of a Complex DNA Fingerprinting Probe for Candida parapsilosis. **39**, 658–669 (2001).

31. Garzillo, C. *et al.* Risk factors for Candida parapsilosis bloodstream infection in a neonatal intensive care unit : a case-control study. (2017) doi:10.1186/s13052-017-0332-5.

32. Ashford, B. K. Certain Conditions of the Gastro-Intestinal Tract in Porto Rico and their Relation to Tropical Sprue. *Am. J. Trop. Med.* **8:507**-**538**, (1928).

33. Weems, J. J. Candida parapsilosis : Epidemiology , Pathogenicity , Clinical Manifestations , and Antimicrobial Susceptibility. **14**, 756–766 (1992).

34. Bendel, C. M. Colonization and Epithelial Adhesion in the Pathogenesis of Neonatal Candidiasis. *Semin. Perinatol.* **27**, 357–364 (2003).

35. Bonassoli, L. A., Bertoli, M. & Svidzinski, T. I. E. High frequency of Candida parapsilosis on the hands of healthy hosts. **59**, 159–162 (2005).

36. Lupetti, A. *et al.* Horizontal Transmission of Candida parapsilosis Candidemia in a Neonatal Intensive Care Unit. **40**, 2363–2369 (2002).

37. Neji, S. *et al.* Virulence factors , antifungal susceptibility and molecular mechanisms of azole resistance among Candida parapsilosis complex isolates recovered from clinical specimens. **24**, 67 (2017).

38. Karkowska-Kuleta, J., Satala, D., Bochenska, O., Rapala-Kozik, M. & Kozik, A. Moonlighting proteins are variably exposed at the cell surfaces of Candida glabrata, Candida parapsilosis and Candida tropicalis under certain growth conditions. *BMC Microbiol.* **19**, 1–13 (2019).

39. Ramage, G., Martínez, J. P. & López-Ribot, J. L. Candida biofilms on implanted biomaterials: A clinically significant problem. *FEMS Yeast Res.* **6**, 979–986 (2006).

40. Silva, S. *et al.* Biofilms of non-Candida albicans Candida species: Quantification, structure and matrix composition. *Med. Mycol.* **47**, 681–689 (2009).

41. Pristov, K. E. & Ghannoum, M. A. Resistance of Candida to azoles and echinocandins worldwide. *Clin. Microbiol. Infect.* **25**, 792–798 (2019).

42. Gácser, A., Trofa, D., Schäfer, W. & Nosanchuk, J. D. Targeted gene deletion in Candida parapsilosis demonstrates the role of secreted lipase in virulence. **117**, 3049–3058 (2007).

43. Tóth, R. *et al.* Investigation of Candida parapsilosis virulence regulatory factors during host-pathogen interaction. 1–14 (2018) doi:10.1038/s41598-018-19453-4.

44. Ping, Y. *et al.* In Vitro Evaluation of Phospholipase , Proteinase , and Esterase Activities of Candida parapsilosis and Candida metapsilosis. 429–438 (2011) doi:10.1007/s11046-011-9440-8.

45. Singh, D. K., Papp, A., Heidingsfeld, O., Dostal, J. & Bajtay, Z. Functional Characterization of Secreted Aspartyl Proteases in. **4**, 1–16 (2019).

46. Tóth, A. *et al.* Secreted Candida parapsilosis lipase modulates the immune response of primary human macrophages. 555–562 (2014).

47. Grózer Z, Tóth A, Toth R, Kecskementi A, Vagolgyi C, Nosanchuk JD, Szekeres A, G. A. Candida parapsilosis produces prostaglandins from exogenous arachidonic acid and OLE2 is not required for their synthesis. **6**, 85–92 (2015).

48. Thomaz, D. Y., Nobrega, J. & Jr, D. A. An Azole-Resistant Candida parapsilosis Outbreak : Clonal Persistence in the Intensive Care Unit of a Brazilian Teaching Hospital. **9**, (2018).

49. Davari, A., Haghani, I., Hassanmoghadam, F. & Nabili, M. Echinocandin resistance in Candida parapsilosis sensu stricto : Role of alterations in CHS3 , FKS1 and Rho gene expression. *J. Glob. Antimicrob. Resist.* **22**, 685–688 (2020).

50. François, I. E. J. A. *et al.* Azoles : Mode of Antifungal Action and Resistance Development . Effect of Miconazole on Endogenous Reactive Oxygen Species Production in Candida albicans. **5**, 3–13 (2006).

51. Allana J Sucher, Elias B Chahine, and H. E. B. Echinocandins : The Newest Class of Antifungals. **43**, 1647–1657 (2009).

52. Castanheira, M., Deshpande, L. M., Messer, S. A., Rhomberg, P. R. & Pfaller, M. A. Analysis of global antifungal surveillance results reveals predominance of Erg11 Y132F alteration among azole-resistant Candida parapsilosis and Candida tropicalis and country-specific isolate dissemination. *Int. J. Antimicrob. Agents* **55**, 105799 (2020).

53. Nucci, M. *et al.* Epidemiology of Candidemia in Latin America : A Laboratory-Based Survey. **8**, (2013).

54. Bhattacharya, S., Sae-tia, S. & Fries, B. C. Candidiasis and Mechanisms of Antifungal Resistance. 1–19 (2020).

55. Choi, Y. J. *et al.* Fluconazole-Resistant Candida parapsilosis Bloodstream Isolates with Y132F Mutation in ERG11 Gene, South Korea. **24**, 1768–1770 (2018).

56. Singh, A. *et al.* Emergence of clonal fluconazole-resistant Candida parapsilosis clinical isolates in a multicentre laboratory-based surveillance study in India. 1260–1268 (2019)

doi:10.1093/jac/dkz029.

57. https://aline-goulart.blogspot.com/2020/03/map-of-world-black-and-white-simple.html. Consulted in January, 2021.

58. Chew, S. Y. *et al.* Physiologically relevant alternative carbon sources modulate biofilm formation, cell wall architecture, and the stress and antifungal resistance of candida glabrata. *Int. J. Mol. Sci.* **20**, 1–13 (2019).

59. Brown, A. J. P., Brown, G. D., Netea, M. G. & Gow, N. A. R. Metabolism impacts upon Candida immunogenicity and pathogenicity at multiple levels. *Trends Microbiol.* **22**, 614–622 (2014).

60. Biology, H., Williams, R. B. & Lorenz, M. C. Multiple Alternative Carbon Pathways Combine To Promote Candida albicans Stress Resistance, Immune Interactions, and Virulence. **11**, 1–18 (2020).

61. Papadakis GZ, Millo C, S. C. Overview of carbon and nitrogen catabolite metabolism in the virulence of human pathogenic fungi. **107**, 277–297 (2018).

62. Devadas, S. M. *et al.* Auxanographic carbohydrate assimilation method for large scale yeast identification. *J. Clin. Diagnostic Res.* **11**, DC01–DC03 (2017).

63. Man, A. *et al.* New perspectives on the nutritional factors influencing growth rate of Candida albicans in diabetics . An in vitro study. **112**, 587–592 (2017).

64. Qadri, S. M. H. & Nichols, C. W. Tube carbohydrate assimilation method for the rapid identification of clinically significant yeasts. *Med. Microbiol. Immunol.* **165**, 19–27 (1978).

65. Deorukhkar & Roushani, S. Identification of Candida species: conventional methods in the era of molecular diagnosis. *Ann. Microbiol. Immunol.* **1**, 1002 (2018).

66. Turner, S. A., Ma, Q., Ola, M., Martinez de San Vicente, K. & Butler, G. Dal81 Regulates Expression of Arginine Metabolism Genes in Candida parapsilosis . *mSphere* **3**, 1–13 (2018).

67. Limjindaporn, T., Khalaf, R. A. & Fonzi, W. A. Nitrogen metabolism and virulence of Candida albicans require the GATA-type transcriptional activator encoded by GAT1. *Mol. Microbiol.* **50**, 993–1004 (2003).

68. Sprenger, M. *et al.* Fungal biotin homeostasis is essential for immune evasion after macrophage phagocytosis and virulence. *Cell. Microbiol.* **22**, 1–19 (2020).

69. Zhang, C. & Hua, Q. Applications of genome-scale metabolic models in biotechnology and systems medicine. *Front. Physiol.* **6**, 1–8 (2016).

70. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biol.* **20**, 1–18 (2019).

71. Lopes, H. & Rocha, I. Genome-scale modeling of yeast: chronology, applications and critical

perspectives. *FEMS Yeast Res.* **17**, 1–14 (2017).

72. Santos, F., Boele, J. & Teusink, B. *A practical guide to genome-scale metabolic models and their analysis*. *Methods in Enzymology* vol. 500 (Elsevier Inc., 2011).

73. Ines Thiele and Bernhard Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nucleic Acids Res.* **4**, 1–10 (2010).

74. Edwards, J. S. & Palsson, B. O. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416 (1999).

75. Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1777–1782 (2007).

76. Raškevičius, V. *et al.* Genome scale metabolic models as tools for drug design and personalized medicine. *PLoS One* **13**, 1–14 (2018).

77. Rout, S., Patra, N. P. & Mahapatra, R. K. An in silico strategy for identification of novel drug targets against Plasmodium falciparum. *Parasitol. Res.* **116**, 2539–2559 (2017).

78. Richelle, A. *et al.* Towards a widespread adoption of metabolic modeling tools in biopharmaceutical industry : a process systems biology engineering perspective. *Syst. Biol. Appl.* **6**, (2020).

79. Sigmarsdóttir, Þ., McGarrity, S., Rolfsson, Ó., Yurkovich, J. T. & Sigurjónsson, Ó. E. Current Status and Future Prospects of Genome-Scale Metabolic Modeling to Optimize the Use of Mesenchymal Stem Cells in Regenerative Medicine. *Front. Bioeng. Biotechnol.* **8**, 1–22 (2020).

80. Kim, H. U., Sohn, S. B. & Lee, S. Y. Metabolic network modeling and simulation for drug targeting and discovery. *Biotechnol. J.* **7**, 330–342 (2012).

81. Dias, O., Rocha, M., Ferreira, E. C. & Rocha, I. Reconstructing High-Quality Large-Scale Metabolic Models with merlin. *Methods Mol. Biol.* (2018) doi:10.1007/978-1-4939-7528-0_1.

82. Kell, D. B. Systems biology, metabolic modelling and metabolomics in drug discovery and development. **11**, 1085–1092 (2006).

83. Blais, E. M., Chavali, A. K. & Papin, J. A. Linking genome-scale metabolic modeling and genome annotation. *Methods Mol. Biol.* **985**, 61–83 (2013).

84. Jun, W., Uk, H. & Yup, S. Current state and applications of microbial genome-scale metabolic models. *Curr. Opin. Syst. Biol.* **2**, 10–18 (2017).

85. Zimmermann-kogadeeva, M., Zimmermann, M. & Andrew, L. Insights from pharmacokinetic models of host- microbiome drug metabolism. *Gut Microbes* **11**, 587–596 (2020).

86. Kim, T. Y., Kim, H. U. & Lee, S. Y. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metab. Eng.* **12**, 105–111 (2010).

87. Pinzi, L. & Rastelli, G. Molecular docking: Shifting paradigms in drug discovery. *Int. J. Mol. Sci.* **20**, (2019).

88. Meng, X. Y., Zhang, H. X., Mezei, M., & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. Current computer-aided drug design. *Curr. Comput. Aided Drug Des.* **7**, 146–157 (2011).

89. Dias, O., Rocha, M., Ferreira, C. & Rocha, I. Reconstructing genome-scale metabolic models with merlin. **43**, 3899–3910 (2015).

90. Viana, R. *et al.* Genome-scale metabolic model of the human pathogen candida albicans: A promising platform for drug target prediction. *J. Fungi* **6**, 1–19 (2020).

91. Xu, N. *et al.* Reconstruction and analysis of the genome-scale metabolic network of Candida glabrata. *Mol. Biosyst.* **9**, 205–216 (2013).

92. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

93. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

94. https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=5480.

95. Tsui, C. K. M., Daniel, H. M., Robert, V. & Meyer, W. Re-examining the phylogeny of clinically relevant Candida species and allied genera based on multigene analyses. *FEMS Yeast Res.* **8**, 651–659 (2008).

96. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acid Res.* **28**, 27–30 (2000).

97. Flamholz, A., Noor, E., Bar-Even, A. & Milo, R. EQuilibrator - The biochemical thermodynamics calculator. *Nucleic Acids Res.* **40**, 770–775 (2012).

98. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, 459–471 (2014).

99. Degtyarenko, K. *et al.* ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, 344–350 (2008).

100. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **30**, 47–49 (2002).

101. Horton, P. *et al.* WoLF PSORT : protein localization predictor. **35**, 585–587 (2007).

102. Rocha, I. *et al.* OptFlux: An open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.* **4**, (2010).

103. Skrzypek, M. S., Binkley, J. & Sherlock, G. Using the candida genome database. *Methods Mol.*

Biol. **1757**, 31–47 (2018).

104. Zomorrodi, A. R. & Maranas, C. D. Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. *BMC Syst. Biol.* **4**, 178 (2010).

105. Duarte, N. C., Herrgård, M. J. & Palsson, B. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309 (2004).

106. Suzuki, M. & Nakase, T. A phylogenetic study of ubiquinone Q-8 species of the genera Candida, Pichia, and Citeromyces based on 18S ribosomal DNP sequence divergence. *J. Gen. Appl. Microbiol.* **45**, 239–246 (1999).

107. https://biocyc.org/META/NEW-IMAGE?object=PWY-5871. Consulted in November, 2020.

108. Sauer, U. *et al.* Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *J. Bacteriol.* **181**, 6679–6688 (1999).

109. Moore, G. E., Gerner, R. E. & Franklin, H. A. Culture of Normal Human Leukocytes. *JAMA J. Am. Med. Assoc.* **199**, 519–524 (1967).

110. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).

111. Coquille, S., Roux, C., Fitzpatrick, T. B. & Thore, S. The last piece in the vitamin B1 biosynthesis puzzle: Structural and functional insight into yeast 4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate (HMP-P) synthase. *J. Biol. Chem.* **287**, 42333–42343 (2012).

112. Rung-Yi Lai *et al.* Thiamin pyrimidine biosynthesis in Candida albicans: a remarkable reaction between histidine and pyridoxal phosphate. **22**, 9157–9159 (2012).

113. https://wi.knaw.nl/page/fungal_table. Consulted in January, 2021.

114. Jia, N. *et al.* Candida albicans sterol C-14 reductase, encoded by the ERG24 gene, as a potential antifungal target site. *Antimicrob. Agents Chemother.* **46**, 947–957 (2002).

115. Jachak, G. R. *et al.* Silicon Incorporated Morpholine Antifungals: Design, Synthesis, and Biological Evaluation. *ACS Med. Chem. Lett.* **6**, 1111–1116 (2015).

116. Lawrence, M. C. *et al.* The three-dimensional structure of the bifunctional 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/dihydropteroate synthase of Saccharomyces cerevisiae. *J. Mol. Biol.* **348**, 655–670 (2005).

117. Navarro-Martínez, M. D., Cabezas-Herrera, J. & Rodríguez-López, J. N. Antifolates as antimycotics?. Connection between the folic acid cycle and the ergosterol biosynthesis pathway in Candida albicans. *Int. J. Antimicrob. Agents* **28**, 560–567 (2006).

118. Eldesouky, H. E., Mayhoub, A., Hazbun, T. R. & Seleema, M. N. Reversal of azole resistance in

Candida albicans by sulfa antibacterial drugs. *Antimicrob. Agents Chemother.* **62**, 1–12 (2018).

119. Nguyen, L. N. & Nosanchuk, J. D. The inhibitory effect of cerulenin to yeasts is fungicidal. *Commun. Integr. Biol.* **4**, 631–632 (2011).

120. Paiva, P. *et al.* Animal Fatty Acid Synthase: A Chemical Nanofactory. *Chem. Rev.* **121**, 9502–9553 (2021).

121. Nguyen, L. N., Trofa, D. & Nosanchuk, J. D. Fatty acid synthase impacts the pathobiology of Candida parapsilosis in vitro and during mammalian infection. *PLoS One* **4**, (2009).

122. Bain, J. M., Stubberfield, C. & Gow, N. A. R. Ura-status-dependent adhesion of Candida albicans mutants. *FEMS Microbiol. Lett.* **204**, 323–328 (2001).

123. Spry, C., Kirk, K. & Saliba, K. J. Coenzyme A biosynthesis: An antimicrobial drug target. *FEMS Microbiol. Rev.* **32**, 56–106 (2008).

124. Spry, C. *et al.* Pantothenamides Are Potent, On-Target Inhibitors of Plasmodium falciparum Growth When Serum Pantetheinase Is Inactivated. *PLoS One* **8**, (2013).

125. Aaron, K. E., Pierson, C. A., Lees, N. D. & Bard, M. The Candida albicans ERG26 gene encoding the C-3 sterol dehydrogenase (C-4 decarboxylase) is essential for growth. *FEMS Yeast Res.* **1**, 93–101 (2001).

126. Pierson, C. A. *et al.* Isolation, characterization, and regulation of the Candida albicans ERG27 gene encoding the sterol 3-keto reductase. *Med. Mycol.* **42**, 461–473 (2004).

Annexes

Table A10: Predicted essential reactions from CparMM. Note how R03118 is encoded by the FKS genes, encoding for the target of the front-line antifungals echinocandins. Similarly to ERG11 and the respective azoles, these genes are solidly validated as essential and thus have strong prediction validation power. Gene names were retrieved from Candida Genome Database and UniProt. Drug information was retrieved from DrugBank and refers to compounds that have been identified for any species.

Reaction ID	Gene ID	EC	Pathway	Human Ortholog	Drug
R03118	*CPAR2_106400*	2.4.1.34	Starch and Sucrose Metabolism	#	**Echinocandin Target (FKS genes)**
	CPAR2_109680				
	CPAR2_804030				
R07509	*CPAR2_400970*	1.14.18.9	Sterol Biosynthesis	*MSMO1*	Yes
	CPAR2_801410				
R12403	*CPAR2_400970*	1.14.18.9	Sterol Biosynthesis	*MSMO1*	Yes
	CPAR2_801410				
R05639	*CPAR2_502980*	1.3.1.70	Sterol Biosynthesis	*LBR*	Yes
	CPAR2_405900				
R02003	*CPAR2_103950*	2.5.1.10	Sterol Biosynthesis	*FDPS*	Yes
	CPAR2_302840				
R01130	*CPAR2_104580*	1.1.1.205	Purine Metabolism	*IMPDH1/2*	Yes
	CPAR2_701040				
R02017	*CPAR2_212160*	1.17.4.1	Purine Metabolism	#	Yes
	CPAR2_211690				
	CPAR2_204520				
R02019	*CPAR2_212160*	1.17.4.1	Purine Metabolism	*RRM*	Yes
	CPAR2_204520				
	CPAR2_211690				
R02101	*CPAR2_806850*	2.1.1.45	Pyrimidine metabolism	*TYMS*	Yes
	CPAR2_206550				
R00965	*CPAR2_502890*	4.1.1.23	Pyrimidine metabolism	*UMPS*	Yes
	CPAR2_502880				
R02018	*CPAR2_212160*	1.17.4.1	Pyrimidine Metabolism	*RRM*	Yes
	CPAR2_211690				
	CPAR2_204520				
R02024	*CPAR2_204520*	1.17.4.1	Pyrimidine Metabolism	*RRM*	Yes
	CPAR2_211690				
	CPAR2_212160				
R02115	*CPAR2_213090*	3.1.1.13	#	*CEL*	No
	CPAR2_105000				
R01658	*CPAR2_302840*	2.5.1.1	Terpenoid Backbone Biosynthesis	*FDPS*	Yes
	CPAR2_103950				

R03038	*CPAR2_400930*	6.1.1.7	Aminoacyl-tRNA Biosynthesis	*AARS1/2*	No
	CPAR2_601240				
R03664	*CPAR2_501890*	6.1.1.2	Aminoacyl-tRNA Biosynthesis	*WARS*	Yes
	CPAR2_700730				
R03659	*CPAR2_502690*	6.1.1.10	Aminoacyl-tRNA Biosynthesis	*MARS1/2*	Yes
	CPAR2_101150				
R03662	*CPAR2_603540*	6.1.1.11	Aminoacyl-tRNA Biosynthesis	*SARS*	No
	CPAR2_102160				
R03656	*CPAR2_802020*	6.1.1.5	Aminoacyl-tRNA Biosynthesis	*IARS1/2*	Yes
	CPAR2_601570				
R03658	*CPAR2_805320*	6.1.1.6	Aminoacyl-tRNA Biosynthesis	*KARS1*	No
	CPAR2_807790				
R03648	*CPAR2_703810*	6.1.1.22	Aminoacyl-tRNA Biosynthesis	*NARS1/2*	Yes
	CPAR2_802400				
	CPAR2_703840				
R02918	*CPAR2_400810*	6.1.1.1	Aminoacyl-tRNA Biosynthesis	*YARS1/2*	Yes
	CPAR2_802620				
	CPAR2_603160				
R03660	*CPAR2_213040*	6.1.1.20	Aminoacyl-tRNA Biosynthesis	*FARSA/B/2*	Yes
	CPAR2_800830				
	CPAR2_106550				
R03663	*CPAR2_106030*	6.1.1.3	Aminoacyl-tRNA Biosynthesis	*TARS1/3*	Yes
	CPAR2_502300				
R03652	*CPAR2_301110*	6.1.1.18	Aminoacyl-tRNA Biosynthesis	*QARS1*	No
	CPAR2_301040				
R03661	*CPAR2_803080*	6.1.1.15	Aminoacyl-tRNA Biosynthesis	*PARS2*	Yes
	CPAR2_703820				
	CPAR2_802990				
R00858	*CPAR2_806380*	1.8.1.2	Sulfur Metabolism	#	Yes
	CPAR2_104830				
R02222	*CPAR2_406570*	1.14.19.1	Biosynthesis of Unsaturated Fatty Acids	*SCD*	Yes
	CPAR2_206900				
R04355	*CPAR2_807400*	2.3.1.86	Fatty Acid biosynthesis	#	No
	CPAR2_302650				
R04726	*CPAR2_807400*	2.3.1.86	Fatty Acid biosynthesis	#	No
	CPAR2_302650				
R04957	*CPAR2_807400*	2.3.1.86	Fatty Acid biosynthesis	#	No
	CPAR2_302650				
R04963	*CPAR2_807400*	2.3.1.86	Fatty Acid biosynthesis	#	No
	CPAR2_302650				
R04960	*CPAR2_807400*	2.3.1.86	Fatty Acid biosynthesis	#	No
	CPAR2_302650				
R04952	*CPAR2_302650*	2.3.1.86	Fatty Acid biosynthesis	#	No
	CPAR2_807400				
R04968	*CPAR2_302650*	2.3.1.86	Fatty Acid biosynthesis	#	No

	CPAR2_807400				
R00742	*CPAR2_105520*	6.4.1.2	Fatty Acid biosynthesis	*ACACB*	
	CPAR2_804060				
R07758	*CPAR2_211750*	2.3.1.199	Biosynthesis of Unsaturated Fatty Acids	*ELOVL*	No
	CPAR2_201340				
	CPAR2_807310				
R07760	*CPAR2_704290*	4.2.1.134	Biosynthesis of Unsaturated Fatty Acids	*HACD*	Yes
	CPAR2_203400				
	CPAR2_704270				
R01049	*CPAR2_206340*	2.7.6.1	Pentose Phosphate Pathway	*PRPS1/2*	Yes
	CPAR2_401280				
	CPAR2_104910				
	CPAR2_304260				
R07274	*CPAR2_402160*	2.5.1.47	Cysteine and Methionine Metabolism	#	Yes
	CPAR2_106760				
	CPAR2_602080				
	CPAR2_300800				
R00355	*CPAR2_701830*	2.6.1.1	Alanine, Aspartate and Glutamate Biosynthesis	*GOT1*	Yes
	CPAR2_206490				
	CPAR2_400400				
R01826	*CPAR2_107970*	2.5.1.54	Phenylalanine, Tyrosine and Tryptophan Biosynthesis	#	Yes
	CPAR2_211970				

Table A11: Predicted essential genes in CparMM. Simulation performed in RPMI medium as described in methods. Note how ERG11 is one of the predicted essential genes. ERG11 codes for the target of the front-line antifungals azoles and is thus a solidly validated essential gene with strong validation power. Note also how a significant number of putative targets, similarly to ERG11, pertain to ergosterol biosynthesis. By means of ERG11 this is a targetable pathway and thus of notice. In its turn, FOL1 has also been identified as a strong putative novel drug target due to Candida spp.'s inability to uptake folate from the environment relying on its de novo biosynthesis. Gene names were retrieved from Candida Genome Database and UniProt. Drug information was retrieved from DrugBank and refers to compounds that have been identified for any species.

Gene ID	Gene Name			Essential *C. albicans*	EC	Pathway	Human Ortholog	Drug
	C. parapsilosis	***C. albicans***	***S. cerevisiae***					
CPAR2_303740	*ERG11*	*ERG11*	*ERG11*	Yes	1.14.14.154	Ergosterol Biosynthesis	*CYP51A1*	**Azole Target**
CPAR2_801560	#	*ERG27*	*ERG27*	Yes	1.1.1.270	Ergosterol Biosynthesis	*HSD17B7*	No
CPAR2_302110	#	*ERG26*	*ERG26*	No	1.1.1.170	Ergosterol Biosynthesis	*NSDHL*	No
CPAR2_701400	#	*ERG13*	*ERG13*	Yes	2.3.3.10	Ergosterol Biosynthesis	*HMGCS1*	Yes
CPAR2_803530	#	*ERG12*	*ERG12*	Yes	2.7.1.36	Ergosterol Biosynthesis	*MVK*	No
CPAR2_602050	#	#	*CAB1*	Yes	2.7.1.33	Ergosterol metabolic process	*PANK1/2/3*	Yes
CPAR2_210480	#	*ERG1*	*ERG1*	Yes	1.14.14.17	Ergosterol Biosynthesis	*SQLE*	Yes
CPAR2_406760	#	*ERG9*	*ERG9*	Yes	2.5.1.21	Ergosterol Biosynthesis	*FDFT1*	No
CPAR2_400710	#	*ERG8*	*ERG8*	Yes	2.7.4.2	Ergosterol Biosynthesis	*PMVK*	Yes
CPAR2_301800	#	*ERG7*	*ERG7*	No	5.4.99.7	Ergosterol Biosynthesis	*LSS*	Yes
CPAR2_405010	*ERG6*	*ERG6*	*ERG6*	No	2.1.1.41	Ergosterol Biosynthesis	#	Yes
CPAR2_703970	#	ERG5	ERG5	No	1.14.19.41	Ergosterol Biosynthesis	#	Yes
CPAR2_105550	*ERG3*	*ERG3*	*ERG3*	No	1.14.19.20	Ergosterol Biosynthesis	*SC5D*	Yes
CPAR2_301960	#	*NCP1*	*NCP1*	No	1.6.2.4	Ergosterol Biosynthesis	*NPC1*	Yes
CPAR2_401630	#	*IDI1*	*IDI1*	Yes	5.3.3.2	Ergosterol Biosynthesis	*IDI1/2*	Yes
CPAR2_109530	#	*MVD*	*MVD1*	Yes	4.1.1.33	Ergosterol Biosynthesis	*MVD*	Yes
CPAR2_110330	#	*HMG1*	*HMG1*	Yes	1.1.1.34	Ergosterol Biosynthesis	*HMGCR*	No
CPAR2_212310	#	#	*ABZ2*	No	4.1.3.38	Folate Biosynthesis	#	Yes
CPAR2_704170	#	*DFR1*	*DFR1*	Yes	1.5.1.3	Tetrahydrofolate Biosynthesis	*DHFR*	Yes
CPAR2_303390	#	*FOL1*	*FOL1*	No	2.5.1.15	Folate Biosynthesis	#	Yes

CPAR2_101130	#	#	*FOL2*	Yes	3.5.4.16	Folate Biosynthesis	*GCH1*	Yes
CPAR2_403110	#	*ABZ1*	*ABZ1*	No	2.6.1.85	Tetrahydrofolate Biosynthesis	#	Yes
CPAR2_802790	#	*URA5*	*URA5*	Yes	2.4.2.10	Pyrimidine Metabolism	*UMPS*	Yes
CPAR2_100620	#	*URA7*	*URA7*	Yes	6.3.4.2	Pyrimidine Metabolism	*CTP*	No
CPAR2_208400	#	*ADE5,7*	*ADE5,7*	No	6.3.4.13/6.3.3.1	Histidine, Purine and Pyrimidine Biosynthesis	*GART*	No
CPAR2_101390	#	*YNK1*	*YNK1*	No	2.7.4.6	Histidine, Purine and Pyrimidine Biosynthesis	*NME7*	No
CPAR2_805940	#	*ADE2*	*ADE2*	No	4.1.1.21	Histidine, Purine and Pyrimidine Biosynthesis	*PAICS*	Yes
CPAR2_204070	#	*ADE6*	*ADE6*	Yes	6.3.5.3	Histidine, Purine and Pyrimidine Biosynthesis	*PFAS*	No
CPAR2_211620	#	*ADE8*	*ADE8*	No	2.1.2.2	Histidine, Purine and Pyrimidine Biosynthesis	*GART*	Yes
CPAR2_208260	#	*ADE4*	*ADE4*	No	2.4.2.14	Histidine, Purine and Pyrimidine Biosynthesis	*PPAT*	Yes
CPAR2_500190	#	*ADE1*	*ADE1*	Yes	6.3.2.6	Histidine, Purine and Pyrimidine Biosynthesis	*PAICS*	Yes
CPAR2_100500	#	*URA4*	*URA4*	No	3.5.2.3	Histidine, Purine and Pyrimidine Biosynthesis	*CAD*	Yes
CPAR2_804900	#	*URA1*	*URA1*	No	1.3.5.2	Histidine, Purine and Pyrimidine Biosynthesis	*DHODH*	Yes
CPAR2_202250	#	*ADE17*	*ADE17*	Yes	3.5.4.10/2.1.2.3	Histidine, Purine and Pyrimidine Biosynthesis	*ATIC*	No
CPAR2_803560	#	GUA1	GUA1	Yes	6.3.5.2	Histidine, Purine and Pyrimidine Biosynthesis	GMPS	Yes
CPAR2_109130	#	*DUT1*	*DUT1*	No	3.6.1.23	Pyrimidine Metabolism	*DUT*	Yes
CPAR2_803640	#	*ADE12*	*ADE12*	Yes	6.3.4.4	Purine Metabolism	*ADSS1/ADSS2*	Yes
CPAR2_303080	#	*GUK1*	*GUK1*	Yes	2.7.4.8	Purine Biosynthesis	*GUK1*	Yes
CPAR2_204960	#	*ADE13*	*ADE13*	No	4.3.2.2	Purine Metabolism	*ADSL*	Yes
CPAR2_400020	*ARO1*	*ARO1*	*ARO1*	Yes	4.2.1.10	Aromatic Aminoacid Biosynthesis	#	Yes
CPAR2_701850	*ARO2*	*ARO2*	*ARO2*	Yes	4.2.3.5	Aromatic Aminoacid Biosynthesis	#	Yes
CPAR2_804740	#	*TKL1*	*TKL1*	Yes	2.2.1.1	Related to PPP	*TKT*	Yes
CPAR2_205030	#	#	*IFA38*	No	1.1.1.330	FA Elongation and Sphingolipid Biosynthesis	*HSD17B12*	Yes

CPAR2_110100	#	#	*TSC13*	No	1.3.1.93	FA Metabolism	*TECR*	Yes
CPAR2_302650	*FAS1*	*FAS1*	*FAS1*	Yes	Multiple	FA Metabolism	*OLAH/FASN/(…)*	No
CPAR2_602700	#	#	*GEP4*	No	3.1.3.27	Cardiolipin Biosynthesis	*PTPMT1*	Yes
CPAR2_503080	#	#	*DGA1*	No	2.3.1.22	Triacylglycerol Biosynthesis	*MOGAT3*	Yes
CPAR2_805350	#	*PEL1*	*PGS1*	Yes	2.7.8.5	Phospholipid Biosynthesis	*PGS1*	Yes
CPAR2_108790	#	*COQ6*	*COQ6*	No	3.5.-.-	Ubiquinone Biosynthesis	*(?)*	Yes
CPAR2_209250	#	*COQ5*	*COQ5*	Yes	2.1.1.201	Ubiquinone Biosynthesis	*COQ5*	Yes
CPAR2_602300	#	*COQ3*	*COQ3*	No	2.1.1.114/2.1.1.64	Ubiquinone Biosynthesis	*COQ3*	No
CPAR2_200190	#	#	*COQ2*	Yes	2.5.1.39	Ubiquinone Biosynthesis	*COQ2*	Yes
CPAR2_303330	#	#	*CAT5*	No	1.14.13.-	Ubiquinone Biosynthesis	*MICAL1*	Yes
CPAR2_807090	#	#	*CAB2*	Yes	6.3.2.5	Pantothenate and CoA Biosynthesis	#	Yes
CPAR2_800750	#	*CAB3*	*CAB3*	No	4.1.1.36	Pantothenate and CoA Biosynthesis	*PPCDC*	No
CPAR2_502760	#	#	*CAB5*	No	2.7.1.24	Pantothenate and CoA Biosynthesis	*COASY*	Yes
CPAR2_805880	*ASN1*	*ASN1*	*ASN2*	No	6.3.5.4	Asparagine Biosynthesis	*ASNS*	No
CPAR2_107820	#	*GRS1*	*GRS1*	Yes	6.1.1.14	tRNA Regulation	*GARS1*	No
CPAR2_107220	#	#	*MSD1*	Yes	6.1.1.12	tRNA Regulation	*DARS2*	Yes
CPAR2_109290	#	#	*YDR341C*	Yes	6.1.1.19	tRNA Regulation	*RARS1/RARS2*	Yes
CPAR2_100050	#	*HTS1*	*HTS1*	Yes	6.1.1.21	tRNA Regulation	*HARS2*	Yes
CPAR2_801890	#	#	*YNL247W*	Yes	6.1.1.16	tRNA Regulation	*CARS1/2*	No
CPAR2_104470	*VAS1*	*VAS1*	*VAS1*	No	6.1.1.9	tRNA Regulation	*VARS1/VARS2*	No
CPAR2_500260	#	#	*PIS1*	No	2.7.8.11	Drug Transport	*CDIPT*	No
CPAR2_209830	#	*TES1*	*TES1*	No	3.1.2.2	Drug Transport	*THEM5*	Yes
CPAR2_807880	#	*SRB1*	*PSA1*	Yes	2.7.7.13	Plyamine Degradation	*GMPPB*	Yes
CPAR2_202590	#	#	*FMN1*	No	2.7.1.26	Riboflavin, FMN and FAD Biosynthesis	*RFK*	Yes
CPAR2_304080	#	*MET14*	*MET14*	No	2.7.1.25	Sulfate Assimilation	*PAPSS2*	Yes
CPAR2_303180	#	*TRR1*	*TRR1*	No	1.8.1.9	Thioredoxin	*TXNRD1/2/3*	Yes
CPAR2_500130	#	*MET16*	*MET16*	No	1.8.4.8	Sulfate Assimilation	#	Yes

CPAR2_702410	#	#	*PET8*	Yes	#	#	#	Yes
CPAR2_201630	#	*AGM1*	*PCM1*	No	5.4.2.3	UDP-N-acetylglucosamine Biosynthesis	*PGM3*	Yes
CPAR2_211290	#	*GNA1*	*GNA1*	Yes	2.3.1.4	UDP-N-acetylglucosamine Biosynthesis	*GNPNAT1*	Yes
CPAR2_101180	#	*UAP1*	*QRI1*	No	2.7.7.23	UDP-N-acetylglucosamine Biosynthesis	*UAP1*	Yes
CPAR2_805640	#	*CHS1*	*CHS2*	Yes	2.4.1.16	Chitin Biosynthesis	#	Yes
CPAR2_406050	#	*PMI1*	*PMI40*	No	5.3.1.8	Mannitol and Mannose Metabolism	*MPI*	Yes
CPAR2_203830	#	*UGP1*	*UGP1*	Yes	2.7.7.9	UDP-glucose Conversion/Galactose and Sucrose Degradation	*UGP2*	Yes
CPAR2_106460	#	*PMM1*	*SEC53*	No	5.4.2.8	#	*PMM1/2*	Yes
CPAR2_603730	*FAD2*	*FAD2*	#	No	1.14.19.-	#	Multiple	No
CPAR2_801860	*FAD3*	*FAD3*	#	No	1.14.19.-	#	Multiple	No
CPAR2_101880	*NCE103*	*NCE103*	*NCE103*	Yes	4.2.1.1	#	*CA*	Yes
CPAR2_402730	#	#	#	No	1.14.15.7	#	#	Yes
CPAR2_805110	*SER33*	*SER33*	*SER3*	No	1.1.1.95/1.1.1.399	Serine Biosynthesis	*PHGDH*	Yes
CPAR2_500060	*TRP4*	*TRP4*	*TRP4*	No	2.4.2.18	Tryptophan Biosynthesis	#	Yes
CPAR2_109570	#	*TRP1*	*TRP1*	No	5.3.1.24	Phenylalanine, tyrosine and tryptophan Biosynthesis	#	Yes
CPAR2_104790	*SER1*	*SER1*	*SER1*	No	2.6.1.52	Serine Biosynthesis	*PSAT1*	Yes