METABOLIC MODELLING AS A POWERFUL TOOL TO IDENTIFY CRITICAL COMPONENTS OF THE *PNEUMOCYSTIS* GROWTH MEDIUM

Olga A. Nev[1,2], Elena Zamaraeva[3], Romain De Oliveira[4], Wassim Abou-Jaoudé[4], Djomangan Adama Ouattara[4], Jennifer Claire Hoving[2,5,6], Ivana Gudelj[1], Alistair J. P. Brown[1,2]

[1] Biosciences and Living Systems Institute, University of Exeter, Exeter, United Kingdom

[2] Medical Research Council Centre for Medical Mycology, University of Exeter, Exeter, United Kingdom

[3] Leverhulme Research Centre for Functional Materials Design, Materials Innovation Factory, University of Liverpool, Liverpool, United Kingdom

[4] Gencovery, Lyon, France

[5] CMM AFRICA Medical Mycology Research Unit, Institute of Infectious Diseases and Molecular Medicine (IDM)

[6] Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town 7925, South Africa

ABSTRACT

Establishing suitable *in vitro* culture conditions for microorganisms is crucial for dissecting their biology and potential applications. However, a significant number of bacterial and fungal species, including *Pneumocystis jirovecii*, remain unculturable, hampering research efforts. *Pneumocystis* is a deadly pathogen of humans that causes life-threatening pneumonia in immunocompromised individuals and transplant patients. Despite its major impact on human health, we remain ignorant about the pathobiology of this fungus, largely because its experimental dissection is constrained by the inability to culture *Pneumocystis in vitro*. In this study, we present a comprehensive *in silico* genome-scale metabolic model of *Pneumocystis* growth and metabolism, aiming to identify metabolic requirements and imbalances that hinder its growth in vitro. We utilise recently published genome data, available information in the literature as well as bioinformatics and software tools to develop and validate the model. We also employ Flux Balance Analysis and Machine Learning approaches to make important predictions regarding metabolic flux to predict critical components of the *Pneumocystis* growth medium. Our findings offer insights into the biology of *Pneumocystis* and provide a novel strategy to overcome the longstanding challenge of culturing this pathogen *in vitro*.

INTRODUCTION

To experimentally dissect microorganisms, such as bacteria and fungi, it is necessary to establish suitable *in vitro* culture conditions that enable their independent growth within a laboratory setting. This dissection enhances our comprehension of these microbes, their biology, and their potential applications. By cultivating bacteria and fungi, we are able to investigate mechanisms that underlie the infectious diseases they elicit, refine diagnostics for these diseases, and develop novel vaccines and therapeutic interventions.

A wide range of bacterial and fungal species can be cultured, but a significant number remain unculturable, thereby restricting research into these organisms. Examples of unculturable bacteria include the segmented filamentous bacteria, which are frequently found in the gut of mammals and are thought to modulate the immune system and promote gut-associated lymphoid tissue development [1]. Additionally, unculturable *Oscillospira* spp. in the human gut are thought to be important for fibre degradation,

immune modulation, and metabolic health [2, 3]. There is even a list of the "most wanted" taxa from the human microbiome, which includes numerous bacterial species that are considered highly significant for human health but remain unculturable [4]. Another group of bacteria that is challenging to culture consists of viable but non-culturable (VBNC) cells. These cells are metabolically active but cannot form colonies on standard culture media [5]. Over 80 bacterial species are currently known to belong to this group including members of the Proteobacteria and, in particular, the human pathogen *Vibrio parahaemolyticus* [6]. In the fungal kingdom, the cultivation of obligate intracellular Microsporidial pathogens is not fully standardized in most research laboratories, which hampers the improvement of diagnostic techniques as well as investigations of their pathogenesis [7]. Furthermore, the metabolism and biology of fungi in the *Malassezia* genus are poorly understood because of difficulties in culturing them, despite their importance for human health. These fungi have been associated with several skin disorders, including dandruff and seborrheic dermatitis [8].

Another striking example of a non-culturable fungus is the human pathogen *Pneumocystis jirovecii*. This fungus colonizes the lungs of infants [9, 10] and causes life-threatening pneumonia in immunocompromised and transplant patients [11, 12]. It remains one of the most common and serious infections in HIV/AIDS patients [13] and is a particularly serious problem for developing countries (such as in Sub-Saharan Africa). *Pneumocystis* is also an emerging issue in non-HIV patients in developed countries, such as UK [14] and USA [15]. However, despite its major impact upon human health, we remain remarkably ignorant about the biology and epidemiology of *Pneumocystis*. The inability to culture *Pneumocystis in vitro*, despite three decades of research, makes this pathogen uniquely difficult to study. The lack of well-established *in vitro* culture methods has been recognized as *the* major obstacle in *Pneumocystis* research. So far, attempts to overcome predicted auxotrophic requirements by simple supplementation of growth media has not been sufficient to solve the problem and achieve growth *in vitro* [16]. Therefore, we are employing an alternative approach to address this major bottleneck in *Pneumocystis* research.

To identify metabolic requirements and imbalances that prevent the growth of this major fungal pathogen *in vitro*, we have developed, for the first time, a comprehensive *in silico* genome-scale metabolic model (GEM) of *Pneumocystis* growth and metabolism. GEMs are mathematical representations of a cell's metabolic reactions based on genome annotation data and experimentally obtained information [17, 18]. GEMs are often used in combination with Flux Balance Analysis (FBA) [19], which is a computational method to predict metabolic fluxes across the network to achieve a certain optimisation objective, for instance, to maximize growth or to reduce resource consumption. This powerful approach has been successfully applied to metabolic engineering, strain development, drug discovery, enzyme functional predictions, phenotypic characterizations, interspecies interactions modelling and the understanding of human diseases [20, 21, 22]. Furthermore, this approach has been used to design and optimise the composition of culture media. For instance, GEMs have been successfully employed to formulate *in vitro* growth media for bacteria such as *Lactobacillus plantarum* [23] and *Bacteroides caccae* [24] as well as to design minimal growth media for the lactic acid bacterium *Lactococcus lactis* [25], a potentially beneficial gut microbe *Faecalibacterium prausnitzii* [26], as well as for the human pathogens *Staphylococcus aureus* [27], *Neisseria meningitidis* [28], and *Helicobacter pylori* [29].

In this paper we describe the development of a GEM reconstruction for *Pneumocystis* metabolism. We review and refine the current *Pneumocystis* genome annotation using available information in the literature together with well-established bioinformatics tools, and then organise these data into a well-defined GEM network. We then use literature assumptions and available software tools to describe a process of biomass production by defining a biomass equation. To validate the model, we utilise *in vitro* growth culture medium conditions designed by Merali *et al* [30], which have been reported to support *Pneumocystis* growth for a limited duration of time. By extending this with a Flux Balance Analysis (FBA) approach, we show how the model can be used to make important predictions regarding metabolic flux to identify critical components of the *Pneumocystis* growth medium. Finally, we propose how to optimise Merali's culture medium composition by using a machine learning approach.

METABOLIC MODEL DEVELOPMENT

To develop a metabolic model for *Pneumocystis*, we have followed a well-established protocol for generating a high-quality genome-scale metabolic reconstruction proposed by Thiele and Palsson [18]. The procedure generally consists of the following key steps. Firstly, a draft reconstruction is created. Next, a biomass equation is defined. Finally, the reconstruction is translated into a mathematical format, and the resulting model is verified, evaluated and validated using the information from the experimental literature on the organism's phenotypes under various environmental conditions. The schematic representation of this process is shown in Fig 1A.

Draft reconstruction

To generate a high-quality draft reconstruction for *Pneumocystis*, we used the most recent version of the genomic sequence [31]. In order to gain a more accurate picture of the metabolites and nutrients transports functions in *Pneumocystis*, we reviewed these data and, where necessary, refined them using available information in the literature including recently published RNA-Seq data [32] (see Materials and methods section). Finally, to get the most accurate results possible, we performed re-evaluation and manual curation of the entire draft reconstruction and its network content.
We then used a *Constellab* platform (see Materials and methods section) to organize these data into a well-defined genome-scale metabolic network for *Pneumocystis* using CHEBI [33], BRENDA [34], and RHEA [35] databases to define gene-protein-reaction associations. The schematic representation of the draft reconstruction process is shown in Fig 1B.

As a result, we obtained a network consisting of 839 chemical reactions corresponding to 736 enzymes thought to operate in a *Pneumocystis* cell. A comparison of *Pneumocystis* metabolism with that of the model yeasts *S. pombe* and *S. cerevisiae* is illustrated in Figure 2A using the KEGG Mapping Tool [36]. This metabolic map clearly shows that *P. murina* has a reduced genome that lacks metabolic pathways essential for other yeasts [31]. This metabolic map clearly shows that *P. murina* has a reduced genome that lacks metabolic pathways essential for other yeasts [31]. In particular, *Pneumocystis* has partially lost the gluconeogenesis pathway and glyoxylate cycle and is incapable of synthesising all 20

essential amino acids, as well as some polyamines, lipids, vitamins, and cofactors. For example, as shown in Fig 2B, the fungus lacks the enzymes required for synthesising a crucial arginine precursor, ornithine, from glutamate and to convert it to citrulline. Other yeasts possess these enzymes and, therefore, are capable of arginine biosynthesis. Interestingly, according to this metabolic map, *Pneumocystis* can generate arginine from citrulline and aspartate, as well as convert glutamine into glutamate and vice versa. The metabolic pathways that have been lost or retained in *Pneumocystis* are summarized in Table 1.

Biomass equation

A biomass equation accounts for all known biomass constituents and their fractional contributions to the overall biomass of a cell. The biomass reaction is of utmost significance for a model's analysis as it provides an indication of whether growth is increasing or decreasing in response to specific environmental conditions being studied.

To begin with, the main course-grained components of the biomass are considered and then broken down into smaller elements. Typically, the main biomass precursors include proteins, carbohydrates, lipids, RNA and DNA as well as small molecules, such as cofactors. These macro components can be further decomposed into smaller micro constituents. For example, proteins are essentially a sum of amino acids, whereas DNA and RNA consist of nucleotides, and lipids can be present as a pool of fatty acids and/or sterols.

Ideally, a detailed biomass composition is determined experimentally for the studied organism. However, due to the inability to grow *Pneumocystis in vitro*, it was not feasible to obtain a detailed biomass composition for this fungus. Therefore, we estimated a relative fraction of each biomass precursor using genome data, software tools, pre-existing metabolic models of species closely related to *Pneumocystis*, and the limited information about the pathogen's composition available in the literature.

In particular, we estimated *Pneumocystis* protein and DNA content from genomic data with the aid of a java application [37]. To account for the lipid content, we relied on available experimental data reported on the average composition of phospholipids and sterols in *Pneumocystis* [38, 39, 40]. However, due to the unavailability of experimental data, the majority of the content for the soluble pool of small metabolites was derived from a comprehensive general list [18]. Other assumptions were drawn from the biomass content of *S. pombe* and its energy requirements for polymerisation (ATP) [41] combined with available experimental data on *Pneumocystis* cell wall composition [42, 43] as well as the presence of polyamines and carbohydrates in its cell [44, 45].

The schematic representation of the biomass equation is shown in Fig 3. Detailed information on the process of the biomass reaction reconstruction can be found in Supplementary Information 1.

FBA ANALYSIS

We utilised Flux Balance Analysis (FBA) [19] to construct a model capable of predicting metabolic fluxes throughout a genome-scale metabolic network. This powerful approach has been extensively applied in various physiological and biological studies [20]. For example, FBA has been successfully used to identify and resolve "knowledge gaps" in metabolic networks [46], to study the impact of drugs in pathogens and to identify potential drug targets [47], to simulate multiple gene knockouts [48], to predict the yield of important cofactors [49], and to support metabolic engineering [50, 51].

To apply FBA and calculate metabolic fluxes within the network, a metabolic network reconstruction must first be developed. This reconstruction includes a list of stoichiometrically balanced biochemical reactions which need to be converted into a mathematical model. To this end, a numerical matrix S of a size $(m \times n)$ is constructed, where m rows represent metabolites X_i (with $i \in \{1 \dots m\}$) and n columns represent reactions j (with $j \in \{1 \dots n\}$), and metabolic fluxes v_j (with $j \in \{1 \dots n\}$) are assigned to the corresponding reactions (Figure 4). The matrix S consists of stoichiometric coefficients from the biochemical reactions, with positive numbers representing metabolites produced in the reactions, negative numbers representing metabolites consumed in the reactions, and zero representing metabolites that do not participate in the reactions. With a help of matrix S the system of biochemical reactions can then be rewritten as a system of differential equations describing how metabolite concentrations X change over time in the network: $\frac{dX}{dt} = S \times v$.

FBA assumes that the metabolic system is in a steady state, where metabolite production and consumption are balanced within the cell. Thus, the flux through each reaction can be calculated using the following equation: $S \times v = 0$. This mathematical equation typically has multiple solutions due to the greater number of reactions than metabolites in realistic large-scale metabolic networks. Therefore, certain constraints must be placed on the variables, namely fluxes. Typically, constraints on extracellular fluxes are added in accordance with available experimental data. However, the inability to cultivate *Pneumocystis in vitro* has led to a deficiency of essential empirical information. Thus, we are mostly dependent on optimisation methods to obtain solutions by applying FBA.

Since we are interested in predicting growth in *Pneumocystis*, our biological objective is the production of its biomass. We define this biological objective mathematically as an objective function of the biomass flux v_{BIOMASS} and assume that this flux is equivalent to the exponential growth rate of the organism. Then linear or quadratic programming can be used to identify a particular flux distribution in the network that maximizes the biomass objective function (see Materials and methods section for more details). Figure 4 provides a schematic representation of the FBA approach.

MODEL VALIDATION

A genome-scale metabolic model can be verified, validated, and evaluated using available experimental data [18]. To ensure that the model is free of technical errors, the model is usually tested against known constraints. To test whether the model accurately represents

the metabolism of the organism, the predictions of the model are compared with experimental data obtained under various environmental conditions. Finally, to check whether the model is reliable and applicable for the intended purposes, one can perform simulations using the model and compare the results with the actual biological outcomes. Then the model can be refined and iterated to improve its accuracy and usefulness. These steps help to ensure that the genome-scale metabolic model is a reliable tool for predicting and analysing the metabolism of the organism under various conditions.

Due to the lack of a well-established *in vitro* culture medium for *Pneumocystis*, experimental data are limited and are often obtained under *ex vivo* conditions (i.e. growth with cultured mammalian cells) which are not suitable for the model's validation. To validate our metabolic model, we employed *in vitro* growth culture medium conditions designed by Merali *et al.* [30], as described in the Materials and Methods section. We compared the biomass flux values obtained by the model's simulations under Merali's conditions with those calculated for the culture medium where only glucose is present. Genetic data suggest that *Pneumocystis* relies largely on glucose utilisation via oxidative pathways for energy production [44], but cannot grow on this nutrient alone. Figure 5 shows that the model predicts a higher biomass flux under Merali's conditions, as expected, as compared to glucose alone.

MODEL PREDICTIONS

The model of *Pneumocystis* growth and metabolism can be exploited to analyse the composition of Merali's medium and to identify its most significant components for the growth of *Pneumocystis*.

In silico drop off experiments

To identify the most important amino acids for *Pneumocystis* growth present in Merali's medium, we assessed their contributions to biomass flux by sequentially removing each amino acid individually from the growth medium and calculating the corresponding biomass flux. The results suggest that not all amino acids contribute equally to *Pneumocystis* growth. In particular, the model predicts that a deficiency of isoleucine or histidine in the medium results in a significant reduction in biomass flux, whereas arginine or tyrosine deficiencies do not (Fig 6). One possible explanation may be differences in the contribution of amino acids to the overall biomass (see Biomass equation section for more details). Indeed, according to estimates of protein content from genomic data using a java application [37], both isoleucine and histidine contribute a relatively high weight per gram of biomass, whereas arginine and tyrosine contribute the least to biomass (Supplementary Information 1).

In silico gene knockout experiments

The reasoning we employed in the previous subsection to explain the relationship between various amino acids and the biomass flux was based on their respective weights in the biomass equation and may not be applicable to certain amino acids, such as serine and glycine, as shown in Figure 6. Both amino acids contribute approximately as much to the biomass equation as isoleucine and histidine, and yet their absence from the growth

medium is not predicted to decrease the biomass flux. This observation is consistent with the idea that *Pneumocystis* is capable of synthesising certain amino acids from others. In particular, *Pneumocystis* can interconvert serine and glycine, indicating that the elimination of glycine from the growth medium may not dramatically affect the biomass flux, since the pathogen may still be able to synthesize glycine from serine or *vice versa*. To investigate this hypothesis, we employed our model to simulate an *in silico* gene knockout experiment. The genes encoding the enzyme serine hydroxymethyltransferase (SHMT, EC 2.1.2.1) and the enzyme phosphoserine phosphatase (PSP, 3.1.3.3) were "switched off", simulating a typical gene knockout experiment in the laboratory. The model predicts that the biomass flux remains relatively unchanged under the medium conditions designed by Merali *et al*, even when serine, glycine, or both are omitted from the medium (Fig 7A). However, when the SHNT and PSP enzymes are also "switched off", the biomass flux drops due to the inability of *Pneumocystis* to synthesize serine and convert it into glycine. Essentially, this *in silico* experiment corresponds to the inhibition of serine biosynthesis and its further conversion into glycine as well as exclusion of both serine and glycine from the medium (Fig 7B). Since these amino acids contribute a relatively high weight per gram of biomass, their elimination leads to a decrease in the biomass production.

Critical role of lipids in the Pneumocystis growth

In order to identify the key lipids in Merali's medium that are crucial for *Pneumocystis* growth, we evaluated their impact on biomass flux using the same approach as we did for amino acids. Not all lipids appear to contribute equally to *Pneumocystis* growth. Figure 8 shows that cholesterol and palmitic, stearic, and oleic acids are the most important lipids for *Pneumocystis* growth present in Merali's medium. Indeed, experimental studies showed that cholesterol is the most abundant sterol in *Pneumocystis* cell membrane [38, 52], whereas palmitic, stearic, and oleic fatty acids are the dominant fatty acids [38]. Moreover, it has also been demonstrated experimentally that the pathogen is able to assimilate cholesterol and to assimilate these fatty acids from the medium to generate phospholipids and form new membranes [52, 53].

However, certain sterols and fatty acids, which have been experimentally identified in *Pneumocystis* [38, 39, 40], and therefore included in the biomass reaction, were not included by Merali *et al*. This discovery led us to formulate a hypothesis that supplementing the medium with other lipids might enhance *Pneumocystis* growth. We employed our metabolic model to test this hypothesis by simulating the effects of supplementing Merali's medium with additional lipids upon the resultant biomass flux. This predicted that supplementation with additional lipids would significantly increase the biomass flux compared to previously observed values, suggesting that these components may enhance *Pneumocystis* growth (Fig 5).

GROWTH MEDIUM OPTIMISATION

Our model is also a valuable tool to make predictions regarding growth medium composition and it can be used to design new culture media or to refine existing growth conditions. Here, we demonstrate how the model can help in the optimisation of Merali's medium. In particular, we present a method for reducing the number of components in

Merali's medium while maintaining equivalent biomass flux values, without compromising growth performance.

To identify the minimal set of nutrients that can produce the maximal possible biomass flux value, the irreducible subset(s) of the initial nutrient set must be examined. This is the subset that contains those nutrients that cannot be removed without reducing the biomass flux value. Due to the large number of nutrients in Merali's medium (155), it is impractical to explore every possible nutrient combination. Hence, alternative heuristic approaches are needed. Furthermore, specific nutrients may exhibit synergistic effects, implying that the optimal minimal subset size is not only determined by the initial nutrient set size, but also by the order in which individual nutrients are removed. As a result, a simple exclusion of nutrients from the initial list would not yield the optimal solution, necessitating a more sophisticated approach.

Reinforcement learning (RL) has demonstrated its effectiveness in solving optimisation problems where brute force methods are impractical or computationally infeasible [54]. For instance, this robust technique has been applied to various chemical problems [55] including crystal structure prediction [56], used successfully in drug discovery [57], personalised treatment development and clinical decision making [58, 59], and applied to social and behavioural neurosciences [60].

By employing RL to optimise the composition of Merali's growth medium (see Materials and methods section for details), we were able to decrease the initial 155-nutrient set by about 80% while still predicting the same biomass flux value with only 32 required nutrients (SI 4, Table S6). During the learning process (185,000 iterations) the RL algorithm was successfully trained to identify and remove non-essential nutrients only (Fig S2). As a result, the algorithm predicted 1188 minimal sets of nutrients of size 32, where 85 sets were unique. Interestingly, only 27 core nutrients were the same in all minimal sets, whereas other 5 nutrients from the list of 21 non-core nutrients varied from set to set (SI 4, Table S6). Table 2 shows the list of 32 nutrients that appeared most frequently, with a total of 268 occurrences.

We also verified the flux values computed by the model under Merali's medium conditions and discovered 87 zero fluxes associated with extracellular transport reactions (SI 4, Table S5). These fluxes do not contribute to biomass production and therefore, the corresponding ambient nutrients from the medium can be excluded. We compared the nutrients identified as non-essential for growth by the RL approach with nutrients whose corresponding flux values were zero. The RL algorithm identified a significantly greater number of non-essential nutrients, considerably simplifying the original 155-component list of Merali's medium composition to only 32 nutrients.

By using this task as an example, we showed that our novel approach can be successfully applied to make further predictions regarding growth medium composition as well as to design new culture media or to refine existing growth conditions.

MODEL PREDICTIONS AND EMPIRICAL DATA

Significantly, the results obtained by applying RL approach reinforce our previous findings (see Model predictions section) by showing that crucial amino acids and lipids from Merali's growth medium are present in the final optimal minimal nutrient set predicted by the RL algorithm. Significantly, the model's predictions are consistent with the empirical data available in the literature.

Experimental observations

The RL approach identified cholesterol as well as palmitic, stearic, and oleic fatty acids as essential and core components of Merali's medium required for *Pneumocystis* growth. Indeed, it was shown experimentally that these lipids are dominant in the pathogen's composition [38,52] and that *Pneumocystis* is able to assimilate cholesterol and these fatty acids from the environment [52, 53].

Moreover, it was noted by Merali *et al* that *Pneumocystis* growth was observed when p-aminobenzoic acid, putrescine, and protoheme were omitted individually, but their addition enhanced growth [30]. Our RL approach identified these components of Merali's medium as essential, classifying putrescine and protoheme as core nutrients and p-aminobenzoic acid as non-core, which is consistent with Merali's observations.

The SAM dilemma

S-Adenosyl-L-methionine (SAM) is a co-substrate that is involved in various metabolic processes, playing crucial roles in methylation reactions and polyamine biosynthesis. SAM is synthesised from L-methionine and ATP by the enzyme SAM synthase.

Merali *et al* did not experimentally detect any activity of the SAM synthase and thus suggested that *P. carinii* is a SAM auxotroph [61]. They also reported that the inclusion of SAM in the growth medium enhances *Pneumocystis* growth and is absolutely required for its continuous growth [30]. Furthermore, a *Pneumocystis* gene with homology to mitochondrial SAM transporters was discovered, which supports the hypothesis that the fungus scavenges this critical nutrient from its host [62]. However, others have suggested that a gene encoding SAM synthase is present in *Pneumocystis* [31]. Furthermore, this gene was cloned, and its enzymatic activity was confirmed via expression in *Escherichia coli* [63]. These results suggest that *Pneumocystis* may not depend on its host for the supply of this important cofactor.

We used the model to explore this interesting conundrum relating to SAM biosynthesis and metabolism in *Pneumocystis* [64]. Interestingly, the model predicted that SAM and L-methionine are both essential, but non-core components of Merali's medium. The optimal nutrient sets identified by the RL contained either SAM or L-methionine or both. This supports the view that, based on the metabolic interactions in *Pneumocystis* (Fig 9), the fungus requires either methionine or SAM for growth *in vitro*.

DISCUSSION

We have developed, for the first time, a genome-scale metabolic model for *Pneumocystis*. In order to do so, we reviewed and improved the current *Pneumocystis* genome annotation by incorporating relevant information from the literature and utilizing established bioinformatics tools and then organized this information into a well-defined GEM network. We used literature assumptions and available software tools to generate a biomass equation and validated our metabolic model by utilizing *in vitro* growth culture medium conditions proposed by Merali *et al* [30]. By employing a Flux Balance Analysis (FBA) approach, we calculated biomass flux values to demonstrate that our metabolic model is a powerful tool to investigate biological questions that cannot be fully explored through conventional bioinformatics and experimental approaches. Moreover, using the model, we made novel predictions regarding *Pneumocystis* growth and metabolism. In particular, we analysed Merali's growth medium and identified its critical components as well as suggested how to supplement it to enhance *Pneumocystis* growth. We have also demonstrated that our model can replicate drop-off and gene knockout experiments, highlighting its potential as a valuable tool in the development of new anti-fungal drug targets. Specifically, our model can be used to predict the likely impact of inhibiting specific metabolic targets upon fungal growth. Finally, by employing a machine learning approach, we propose how to simplify Merali's medium to achieve a minimal growth culture medium.

Reconstructing metabolic networks can be a challenging process for organisms with incomplete genome annotations, low sequence homology to other organisms and/or unique pathways [65]. As a result, incomplete reconstructions may occur and the quality of the constraints used to build the model may be compromised, which limits the efficacy and precision of the FBA approach [66, 67]. The remarkable nature of *Pneumocystis*, with its extremely reduced genome and dependence on the host for both nutrients and a stable environment [44], requires special attention in the model development process. Since the pathogen possesses certain nutrients transporters and retains an endocytosis pathway, it is potentially capable to scavenge nutrients from the host [31]. This implies that the metabolic network of *Pneumocystis* may exhibit "actual" gaps that mirror the pathogen's reliance on its host. These gaps should not be filled in automatically during the reconstruction process, as this could lead to inaccurate predictions. The FBA approach, however, can be used to predict how these "actual" metabolic gaps may be complemented to promote the growth of the fungus *in vitro*. This will help to identify the growth phenotypes of *Pneumocystis* under specified media conditions and to suggest ways to supplement the nutrient requirements of the pathogen.

The FBA approach, although well-established, has limitations [68]. For example, it cannot predict nutrient concentrations, since it does not incorporate kinetic parameters. Additionally, in its classical form, FBA is only applicable to determine fluxes at a steady state. Moreover, FBA does not take into account regulatory effects such as enzyme activation by protein kinases or gene expression regulation, nor does it consider stress response or signalling pathways, and so its predictions may not always be accurate. To improve the accuracy of predictions made by FBA, the integration of regulatory and signalling network information into the metabolic model is likely to be useful. This approach represents a

valuable extension to the classical FBA approach and has been previously employed to obtain a more detailed and dynamic picture of biological networks [69, 70].

Experimental data play a crucial role at each stage of developing and refining genome-scale metabolic models [68]. To construct and refine metabolic networks, experimental data on the presence or absence of enzymes, metabolites, and reactions can be integrated with transcriptomic and proteomic data. The construction of the biomass equation, as well as setting constraints for metabolic fluxes, also require data from experiments. Furthermore, to validate the model's predictions, a comparison with experimental data is necessary. This creates a feedback loop between the model and the experiments, allowing for the identification of errors and gaps in the model, and informing and refining the metabolic network. However, in the case of *Pneumocystis*, the inability to grow this pathogen *in vitro* has resulted in a lack of critical empirical data on its composition, physiological and biological characteristics, kinetics of nutrient uptake, and enzymatic activity of its genes. The absence of experimental data has hindered the development and refinement of a comprehensive metabolic model for *Pneumocystis*, highlighting the importance of experimental data in the process.

For example, data on enzyme specificity regarding substrate and cofactor usage is exclusively available in organism-specific biochemical literature or databases. Lack of such information for *Pneumocystis* compelled us to rely on organism-unspecific databases, such as RHEA [35] and BRENDA [34]. Additionally, acquiring data on *Pneumocystis*-specific gene and reaction localization proved to be challenging. Hence, we assumed that the proteins reside in the cytoplasm. Moreover, due to the deficiency of critical experimental data, we considered reactions in our initial draft reconstruction reversible, despite the fact that the directionality of reactions may have significant impact on the model's performance. It is worth noting that the reconstruction proposed in this study is subject to revision and refinement in the subsequent versions of the *Pneumocystis* metabolic model, as more biochemical data become available.

The hypotheses generated by the model need to be tested experimentally. Moreover, the growth of *Pneumocystis in vitro* has to be compared experimentally against its growth *in vivo* in the host (as an important control). Clearly, it is not possible to test this in humans. Therefore, for practical and ethical reasons, we developed a genome-scale metabolic model for the mouse-specific fungal pathogen, *Pneumocystis murina*. However, this platform allows us to generate analogous models for other host-specific pathogens, which will provide the basis for the future development of *in vitro* culture conditions for each of these species. Recent genome sequencing has confirmed that the *Pneumocystis* genus is comprised of several species, each being highly specific to a particular mammalian host [71]: human-specific *Pneumocystis jirovecii*, mouse-specific *Pneumocystis murina*, rat-specific *Pneumocystis carinii* and *Pneumocystis wakefieldiae*, rabbit-specific *Pneumocystis oryctolagi*, macaque-specific *Pneumocystis* macacae, and dog-specific *Pneumocystis canis*. Their genome sequences have been recently reported [31, 72, 73, 74]. Metabolic models developed for these *Pneumocystis* species can also be linked to the corresponding metabolic models for their specific mammalian hosts. This will permit detailed comparison of the metabolic gaps for each of these pathogens, thereby providing invaluable clues about the potential basis for their species-specificity.

The lack of *in vitro* culture methods has been recognized as the major obstacle in *Pneumocystis* research. For example, the diagnosis of *Pneumocystis* infections relies primarily on microscopic detection in respiratory specimens [44]. Using a powerful combination of *in silico* metabolic modelling and experimental phenotyping opens new opportunities to release this major experimental bottleneck. This is extremely important because a reliable *in vitro* culture method will empower the dissection of *Pneumocystis* immunology and pathobiology [75, 76], the analysis of the antifungal drug resistance that is emerging in *Pneumocystis* [77], and the development of new antifungal therapies that circumvent the resistance of *Pneumocystis* to azole drugs [31], leading, in the longer-term, to dramatic advances in *Pneumocystis* biology and therapy.

MATERIALS AND METHODS

Bioinformatics tools

In this study, for practical and ethical reasons, we developed a genome-scale metabolic model for the mouse-specific fungal pathogen, *Pneumocystis murina* (*P. murina*). To review and refine the genome annotation for *P. murina*, we used well-established bioinformatics tools to compare this pathogen with other species that are phylogenetically closely related to it [78]. In particular, we selected *Schizosaccharomyces pombe* (*S. pombe*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) as comparison organisms due to the availability of functional annotations for these species.
To identify homologous genes between these organisms and the mouse pathogen *P. murina*, we used BLAST tool [79] in combination with the UniProtKB sequence database [80] (BLASTp with default parameters, filtering alignment at coverage = 80%, identity/similarity = 80%, p-value >= 0.05) and keeping only fungal identified proteins to avoid potential false positive results.

Protein sequences for *P. murina* (PRJNA70803) were downloaded from NCBI as deposited by Ma *et al.* [31]. The latest version of protein sequences for *S. cerevisiae* S288c (GCA_0001460452) and *S. pombe* (GCA_000002945) were downloaded from the *Saccharomyces* Genome Database [84] and *PomBase* Genome Database [85], respectively.

We then used PFAM [81] (with default parameters, filtering results for E-value >= 0.05) to further annotate functional domains in previously selected genes to better link them to gene ontology (GO) terms and enzyme commission (EC) numbers [82, 83].

We finally checked the obtained list of enzymes against biochemical reaction database BRENDA [34] and RNA Seq data [32], and also used available, albeit limited, experimental data on *Pneumocystis* to curate enzymes manually.

These allowed us to generate a comprehensive list of 736 EC numbers corresponding to the enzymes present in *P. murina*, which was a starting point to build a metabolic reconstruction network.

Constellab software tool

In order to build a metabolic reconstruction network based on the 736 EC numbers that correspond to the enzymes found in *P. murina*, we used a *Constellab* platform, which was developed by *Gencovery*. This company specialises in metabolic modelling and offers state-of-the-art expertise in this field [86]. *Constellab* is a cloud-based bioinformatic platform that enables the efficient analysis of complex biological datasets, such as large-scale omics data. In the context of this study, *Constellab* allowed us to create digital twins [87], which are digital representations of the real-world systems or processes that can be used to simulate, analyse, and optimise their performance and efficiency under various situations. In our study, *Constellab* was used to create a metabolic model for *P. murina*, which served as a digital twin of its metabolism and growth.

To achieve this goal, we established gene-protein-reaction associations using the *Gencovery*'s in-house BIOTA open database, which is a comprehensive and structured repository of omics data sourced from the official open European database (EMBL-EBI) [88] and the NCBI taxonomy database [89]. This database also contains manually curated ontology data as well as molecular data on taxonomy, enzymes, metabolites, and metabolic pathways from comprehensive databases, such as CHEBI [33], BRENDA [34], and RHEA [35]. Finally, by mapping our data to the BIOTA, we were able to build an initial version of the *P. murina* metabolic network, which consisted of 839 chemical reactions that corresponded to previously identified 736 enzymes (see Bioinformatics tools subsection).

Manual curation

The initial automated draft reconstruction, comprising 839 reactions, was subject to additional manual curation (SI 2). As a result, four additional reactions were manually incorporated by drawing information from KEGG [36] and RHEA [35] databases, as well as from the *S. cerevisiae* model [90] (SI 2, Table S1).

Subsequently, we assessed whether the network was balanced in terms of mass and charge and identified 64 reactions displaying imbalances in mass and/or charge. We observed that most of the imbalanced reactions contained chemical compounds in a generic form, making it impossible to calculate their mass or charge correctly. Such generic reactions are considered as non-specific due to insufficient knowledge and biochemical evidence, and hence are typically excluded from the network [18]. As a result, 56 imbalanced reactions have been deleted; 8 reactions were kept due to their presence in the *S. cerevisiae* [90] and *S. pombe* models [41] as well as their importance according to the literature and potential contribution to the biomass production. The stoichiometry of all reactions was sourced from the RHEA database [35], which is a comprehensive and manually curated resource (SI 2, Table S2).

Next, we eliminated the isolates from the reconstruction, which refers to those reactions that were not topologically linked to the biomass production and, thus, did not contribute to *Pneumocystis* growth, which is the primary focus of this study. Furthermore, metabolites that lacked any known association with a metabolic pathway or biological function in the CHEBI database [33] were considered as orphan compounds and subsequently removed

from the metabolic network, as their inclusion could lead to inaccurate or incomplete models. At this step, 13 reactions were excluded from the model (SI 2, Table S3).

The resulting draft reconstruction contained 774 reactions and 1027 metabolites (SI 2, Table S4).

Transport reactions

In order to simulate the model under specific environmental conditions, it is necessary to incorporate transport reactions. In this study, we investigated *Pneumocystis* growth in the culture medium proposed by Merali *et al* (see subsection on Merali's culture medium composition). To achieve this, we compiled a list of all the nutrients present in this culture medium and assumed that *Pneumocystis* could acquire these nutrients from the surrounding environment either through dedicated transporters or through endocytosis, as suggested by genome data [31] (SI 3).

As it is not possible to culture *Pneumocystis in vitro*, the absence of critical experimental data prevented us from imposing constraints on the model's variables to achieve more accurate results using the FBA approach. Nevertheless, we did constrain the biomass flux and the fluxes associated with its coarse-grained constituents (proteins, DNA, RNA, carbohydrates, phospholipids, sterols, cofactors) to positive values, thereby ensuring that these components were solely produced and not consumed. In the absence of empirical data, we opted not to constrain external fluxes to prevent the cell from being impeded in excreting by-products.

In order to automatically fill "metabolic gaps" in the network, we exclusively relied on the *Pneumocystis* taxonomy (4753) to avoid filling "actual metabolic gaps" that reflect the pathogen's dependence on its host, as this may potentially lead to inaccurate predictions.

FBA algorithm

We applied the FBA approach to study the metabolic pathways of *P. murina* by calculating the fluxes of all reactions in a metabolic network under studied conditions [19]. To this end, we defined an objective function of the biomass flux v_{BIOMASS}, and identified a particular flux distribution in the network that maximizes the biomass objective function (v^*_{BIOMASS}). We formulated our challenge as an optimization problem with a quadratic objective and affine inequality constraints (a quadratic programming method), and we relaxed a quasi-steady state assumption to obtain an optimal solution the following way:

$$v^*_{\text{BIOMASS}} = \arg\min \{-v_{\text{BIOMASS}} + \lambda R^2(v)\}$$
$$lb < v < ub,$$

where v denotes a vector of the metabolic fluxes; lb and ub are lower and upper bounds of the fluxes (we used $lb = -1000$ and $ub = +1000$ where no empirical data was available); $\lambda > 0$ is the relaxation strength (we used $\lambda = 1$); and $R(v) = \|S \times v\|_2$ is the relaxation term with S denoting a stoichiometric matrix (see FBA analysis section).

Merali's culture medium composition

To simulate the model, we used *in vitro* growth culture medium conditions designed by Merali *et al* [30]. These authors reported that continuous axenic culture of *P. carinii* was achieved in the culture medium based on Minimal Essential Medium with Earle's salt supplemented with S-adenosyl-L-methionine, putrescine, ferric pyrophosphate, N-acetyl glucosamine, putrescine, p-aminobenzoic acid, L-cysteine, L-glutamine, and horse serum. The organism grew attached to a collagen-coated porous membrane at normal atmosphere and 31°C with the pH of the medium ranging between 8.8 and 9. Growth was measured by employing three distinct methods, namely, the estimation of total DNA and total ATP, in addition to a conventional microscopic evaluation of Giesma-stained smears. Doubling times varied from 35 to 44 hours in four independent experiments. It was also reported that cultured organisms were able to infect immunosuppressed rats and could be stored frozen and used to reinitiate culture. However, this cultivation method had its limitations. In particular, no growth was observed after dilutions to the point that only 5-10 clusters of *P. carinii* could be observed. There was no growth either in uncoated inserts. Also, the authors reported that a twice-daily medium exchange was crucial for the growth. Furthermore, other scientists worldwide have reported that their attempts to establish the continuous axenic culture system described by Merali for *P. carinii* met with varying degrees of success. In certain cases, researchers documented their successful attempts in achieving axenic cultivation of *Pneumocystis* [91], while others reported the absence of growth [92].

Reinforcement learning algorithm

Reinforcement learning is a machine learning paradigm where the algorithm is focused on goal-directed learning via interactions between the agent and the environment [54]. The agent takes actions in the environment and receives feedback in the form of rewards (Fig S1), which helps it to learn which actions are desirable and which to avoid through trial and error.

The environment which the agent interacts with provides both state and action spaces. The decision of which action to take next is made based on the current state. The policy maps states to actions and determines which action to take in each state. The policy's quality is evaluated and updated based on the rewards received following each step. The objective is to maximize the long-term cumulative reward by balancing the need to attempt new actions to acquire further knowledge about the environment with the need to take actions that will result in the highest reward.

To optimise the growth medium composition, we used the RL approach described above to find the minimal irreducible subset of nutrients from the initial Merali's 155-nutrient set without dropping the biomass flux value. For this purpose, we treated a subset of nutrients as a state in the state space (the set of all possible subsets), the removal of a specific nutrient as an action, and considered a positive reward if the biomass flux value corresponding to the selected nutrient set was equivalent to that of the previous nutrient set.

RL routine typically comprises a series of consequent episodes. In our case, an episode was defined as a sequence of nutrient deletions commencing from the initial nutrient set and lasting until no nutrient could be excluded without dropping the biomass flux value.

We have chosen the Proximal Policy Optimization (PPO) algorithm [93] as a particular RL method. PPO is widely regarded as a state-of-the-art RL algorithm known for its effectiveness in various domains and has been implemented in different RL libraries. We employed the implementation provided by StableBaseline3 [94].

TABLES

Table 1. Summary of metabolic pathways lost or retained in *Pneumocystis*.

PATHWAY	LOST	RETAINED	PARTLY LOST
Nucleotide metabolism	Nucleotide salvage pathway	Nucleotide *de novo* synthesis	
Carbohydrate metabolism	Nucleotide sugar biosynthesis from galactose Pyruvate fermentation Chitin and mannan biosynthesis Beta-glucan biosynthesis and degradation (troph form only) Glucuronate pathway	Glycolysis TCA cycle Pentose phosphate pathway (oxidative part) PRPP biosynthesis Glycogen degradation Nucleotide sugar biosynthesis from glucose UDP-N-acetyl-D-glucosamine biosynthesis Conversion of fructose and mannose to glucose Trehalose utilization Beta-glucan biosynthesis and degradation (cyst form only)	Pentose phosphate pathway (non-oxidative part) Gluconeogenesis Inositol phosphate metabolism Glycogen biosynthesis Glyoxylation
Lipid metabolism	Fatty acids degradation Conversion of fecosterol and episterol to ergosterol Myo-inositol, choline, complex sphingolipids, ether lipids, phosphatidylinositol, phosphatidylcholine *de novo* synthesis Cytosolic pathway for fatty acids Synthesis of glycerol from glycerone-phosphate or monoacylglycerol	Fatty acid biosynthesis and elongation in mitochondria and ER Fecosterol and episterol synthesis Alternative mechanisms to supply cells with phosphatidylinositol, phosphatidylcholine, inositol, choline	Fatty acid biosynthesis, initiation and elongation Fatty acids beta-oxidation Synthesis of ergosterol Synthesis of cholesterol
Cofactor metabolism	Pantothenate *de novo* synthesis Thiamine, biotin, ubiquinone, siderophores *de novo* synthesis Reductive iron assimilation Synthesis of CoA	Conversion of pantothenate to CoA NAD *de novo* synthesis and its salvage Heme biosynthesis Riboflavin biosynthesis	
Amino acid metabolism	20 amino acids *de novo synthesis* Polyamine *de novo* synthesis	Cysteine biosynthesis from homocysteine + serine	Shikimate pathway Glutathione biosynthesis from glutamate
Energy metabolism	Sulfur metabolism V-type ATPase	Cytochrome bc1 complex Succinate dehydrogenase Carbon fixation	Formaldehyde assimilation NADH dehydrogenase Cytochrome c oxidase F-type ATPase
Glycan metabolism		N-glycan precursor biosynthesis O-glycan biosynthesis	N-glycan precursor trimming N-glycosylation by oligosaccharyltransferase N-glycan biosynthesis GPI-anchor biosynthesis
Endocytosis		Clathrin-dependent endocytosis	
CO2 hydration	whole pathway		
Stress response pathway	pH sensing Osmotic stress response Oxidative stress response Cell wall stress response		

Table 2. Most frequently occurring optimal minimal 32-nutrient set predicted by the Reinforcement Learning algorithm.

Nutrient	Core/Non-core nutrient
Cysteine	core
Histidine	core
Isoleucine	core
Leucine	core
Lysine	core
Phenylalanine	core
Threonine	core
Tryptophan	core
Tyrosine	core
Valine	core
Alanine	core
Pantothenate	core
Riboflavin	core
Thiamine	core
Putrescine	core
Protoheme	core
Margaric acid	core
Myristic acid	core
Oleic acid	core
Palmitic acid	core
Palmitoleic acid	core
Stearic acid	core
Arachidonic acid	core
Linoleic acid	core
Sitosterol	core
Cholesterol	core
Proline	core
Methionine	non-core
Glutamate	non-core
Hydrogenphosphate	non-core
S-adenosyl-L-methionine	non-core
Citrulline	non-core

FIGURES

Figure 1. Schematic representation of the metabolic model development and draft reconstruction processes

(A) The metabolic model development process includes the reconstruction of a draft network, formulation of a biomass equation, and validation of the resulting model using experimental data on the organism's phenotypes under specified environmental conditions. **(B)** To create a high-quality draft reconstruction for *Pneumocystis*, the latest genome annotation is utilised, reviewed, and refined using bioinformatics tools and the literature. This information is further combined with the RNA-Seq data and manually curated. As a result, a list of functional enzymes present in *Pneumocystis* is created. From this list, the draft reconstruction is generated using the Gencovery software tool and the CHEBI, BRENDA, and RHEA databases.

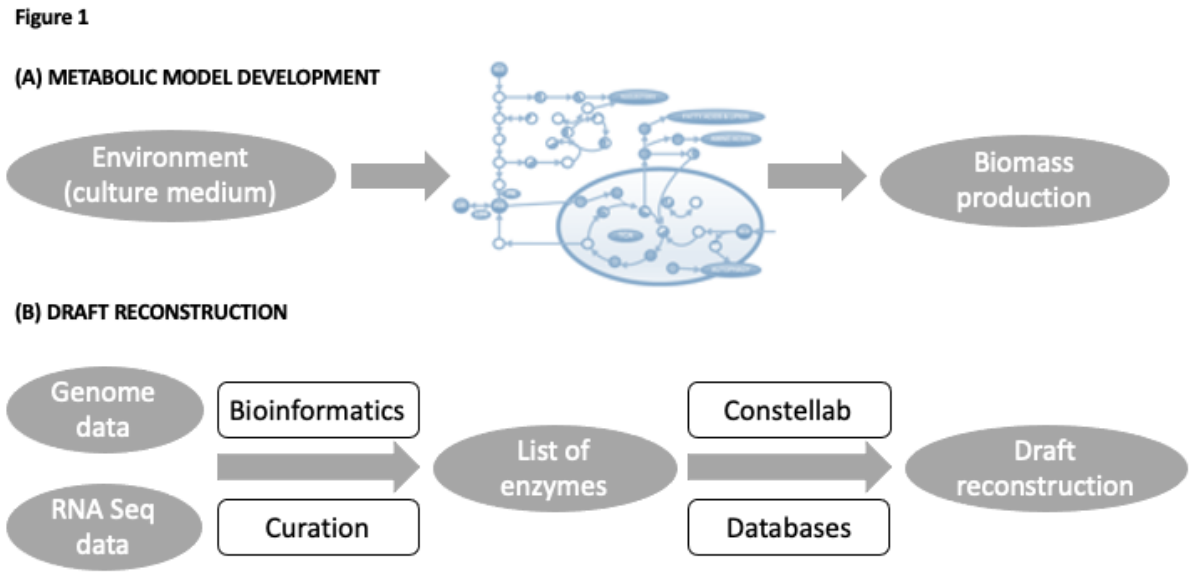
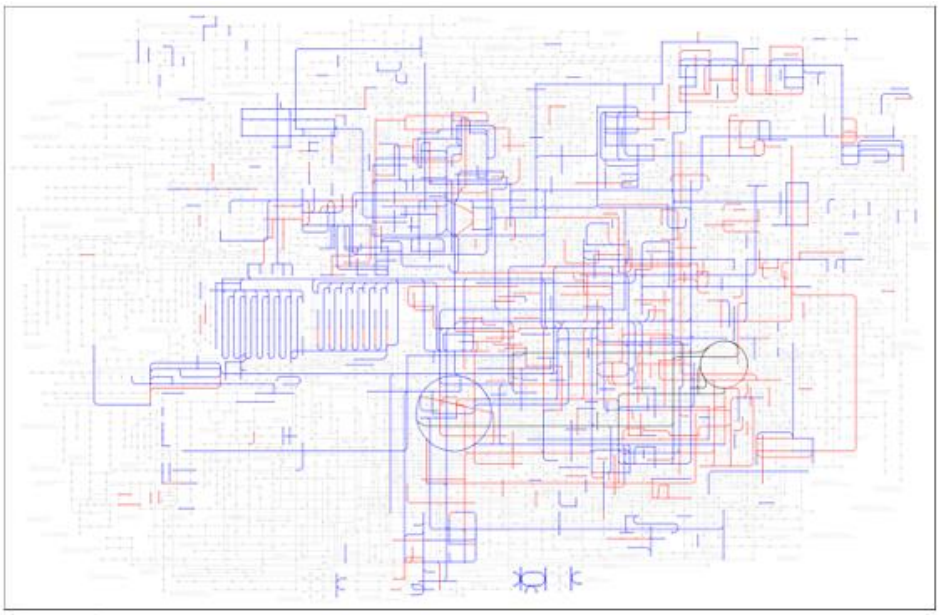


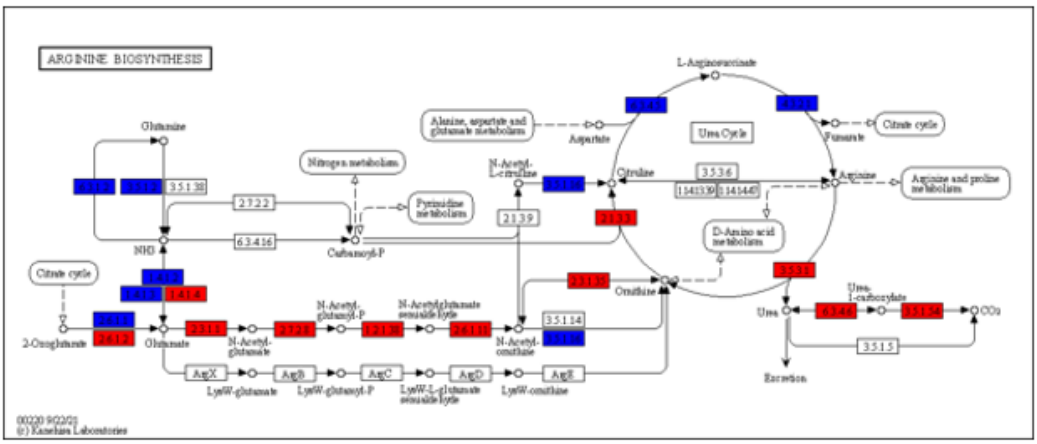
Figure 2. *Pneumocystis* metabolism in comparison with the model yeasts.
(A) Visualisation of *Pneumocystis* metabolism in comparison with that of the model yeasts *S. pombe* and *S. cerevisiae* using the KEGG Mapping Tool. The metabolic pathways present in *Pneumocystis* are highlighted in blue, while the metabolic pathways present in model yeasts, but not in *Pneumocystis*, are featured in red. The arginine biosynthetic pathway is shown in black.
(B) Visualisation of *Pneumocystis* arginine metabolism in comparison with that of the model yeasts *S. pombe* and *S. cerevisiae* using the KEGG Mapping Tool. The enzymes present in *Pneumocystis* are highlighted in blue, while those present in model yeasts, but not in *Pneumocystis*, are featured in red.

Figure 2
(A)



present in *Pneumocystis*
present in model yeasts (*S. pombe* & *S. cerevisiae*), but not in *Pneumocystis*

Figure 2
(B)



present in *Pneumocystis*
present in model yeasts (*S. pombe* & *S. cerevisiae*), but not in *Pneumocystis*

Figure 3. Schematic representation of the biomass equation.
It is assumed that the biomass of *Pneumocystis* consists of the following macro-nutrients: proteins, DNA, RNA, carbohydrates, phospholipids, sterols, and cofactors, which are then broken down into the smaller micro-nutrients, including amino acids, nucleotides, and fatty acids. Contribution of each biomass precursor in *Pneumocystis* is estimated by using genome data, software tools, metabolic models of species closely related to *Pneumocystis*, and available experimental data on *Pneumocystis* composition in the literature.

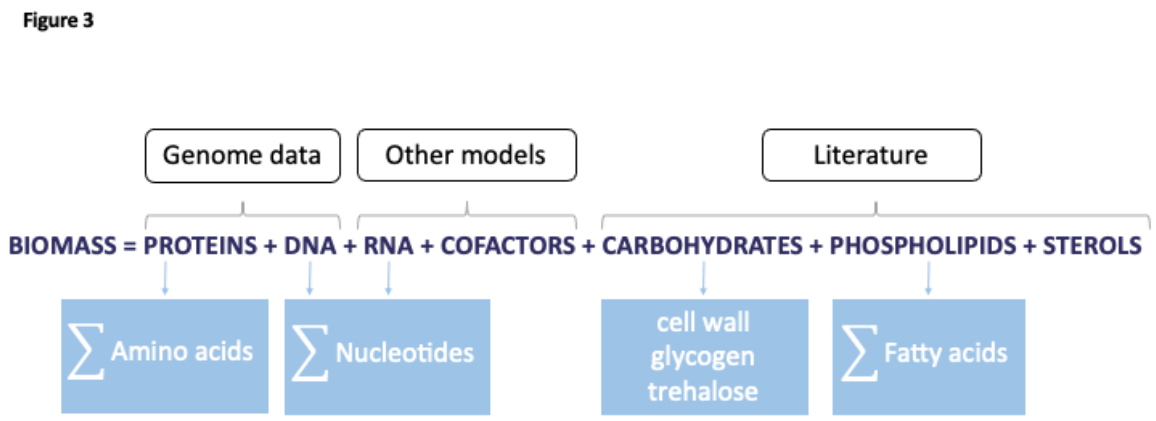


Figure 4. Schematic representation of the FBA approach.
To implement the FBA approach, biochemical reactions are converted into a mathematical model (see Biomass equation section for details). FBA relies on the assumption that the metabolic system is in a steady state. Mathematically, the production of biomass as a biological objective is formulated as a function that maximises the biomass flux, which is equivalent to the exponential growth rate of the organism.

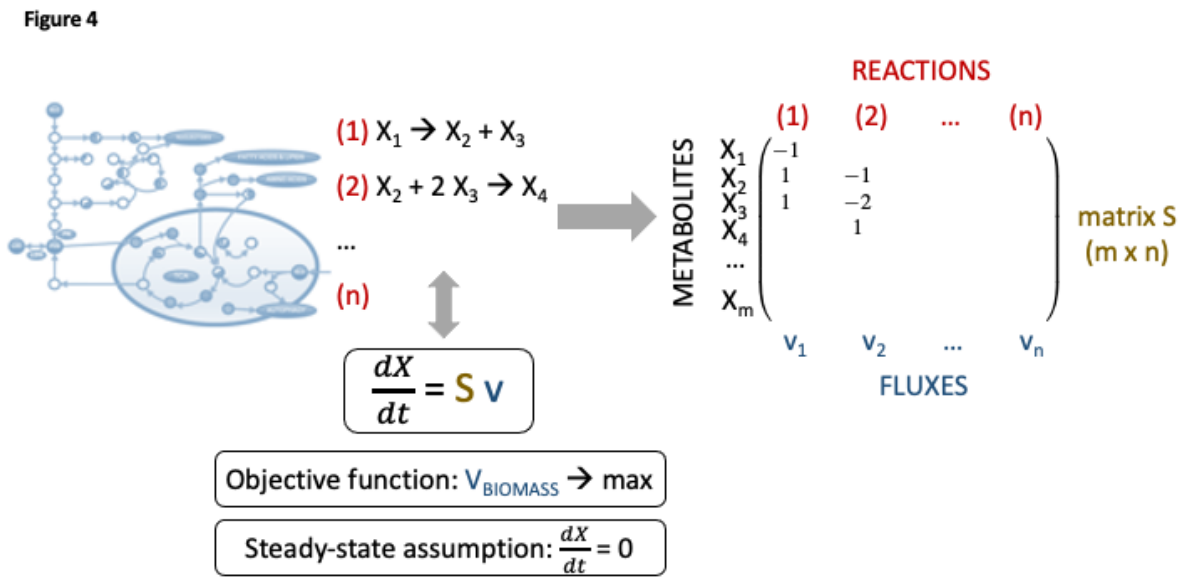


Figure 5. Metabolic model validation and predictions.
The biomass flux predicted by the model under the following conditions: only glucose is present, medium designed by Merali *et al*, and Merali's medium supplemented with lipids which are present in the *Pneumocystis* biomass equation, but not in Merali's medium.

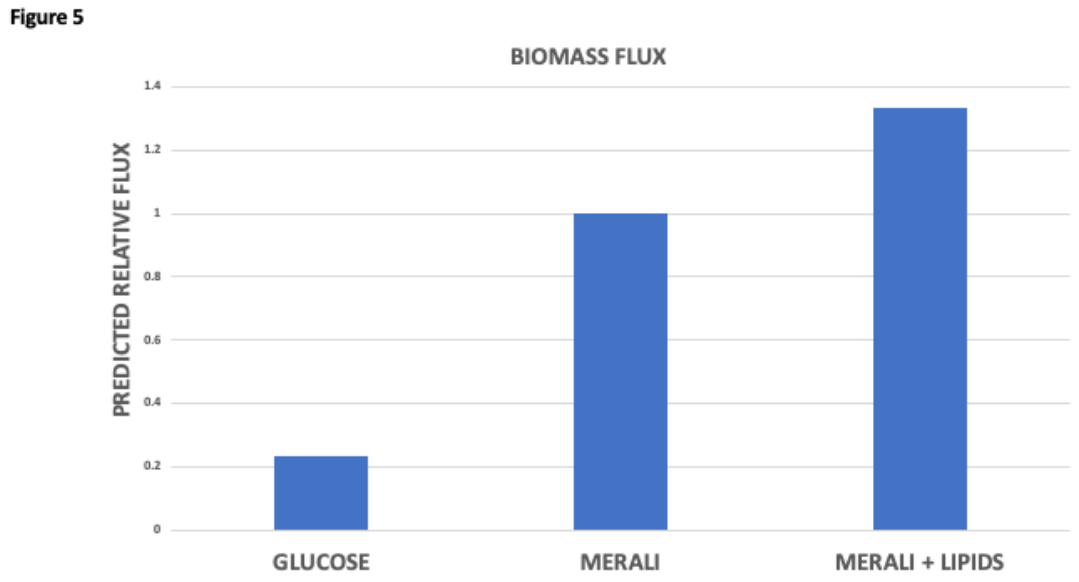


Figure 6. *In silico* drop off experiment: amino acids.
The biomass flux predicted by the model under the medium designed by Merali *et al* and the medium conditions, where each amino acid is individually omitted from Merali's medium.

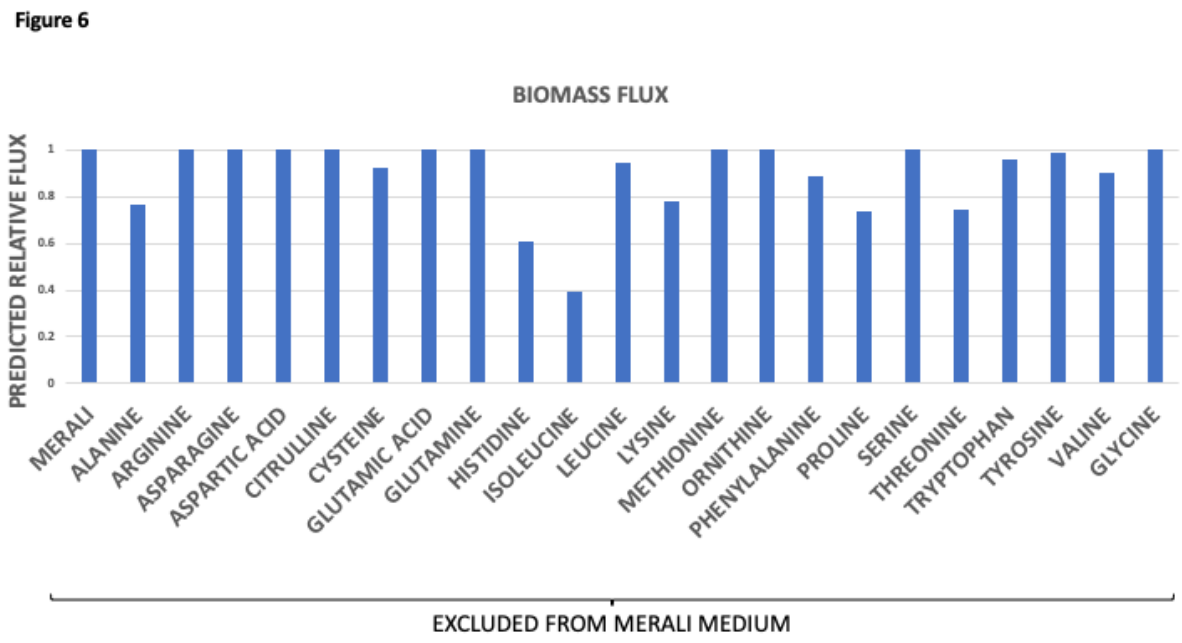
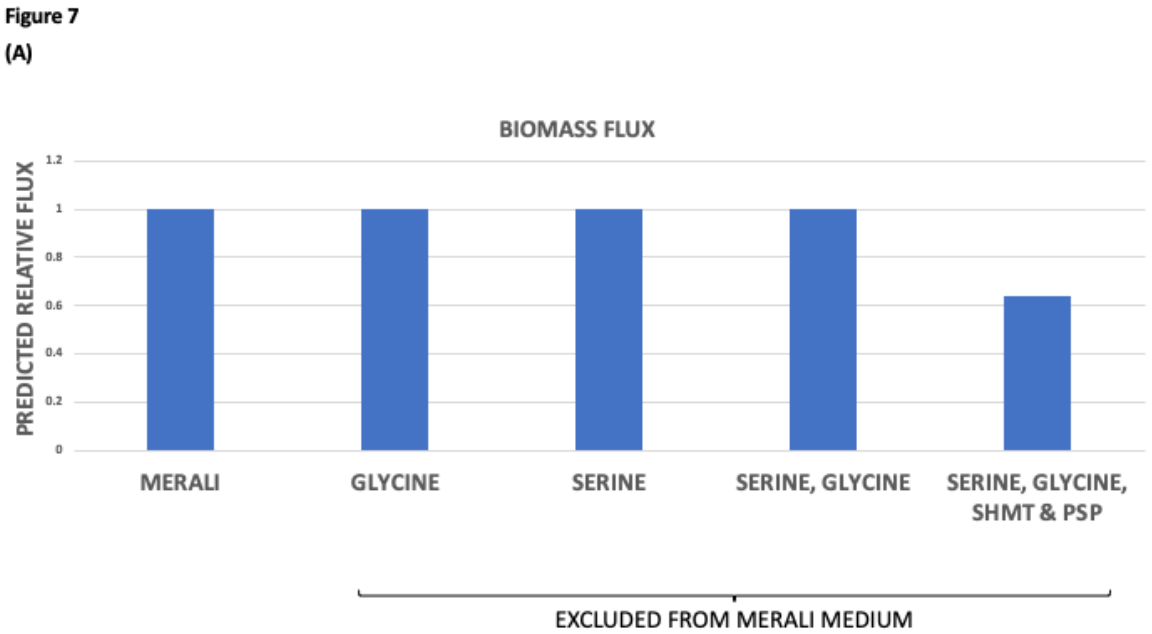


Figure 7. *In silico* gene knockout experiment: serine and glycine interconversion.
(A) The biomass flux predicted by the model under the medium designed by Merali *et al* and the medium conditions, where serine, glycine or both are individually omitted from Merali's medium as well as the conditions, where genes corresponding to SHNT and PSP enzymes are "switched off".
(B) Schematic representation of the *in silico* experiment whereby serine and glycine are excluded from the culture medium and the genes corresponding to SHNT and PSP enzymes are "switched off".



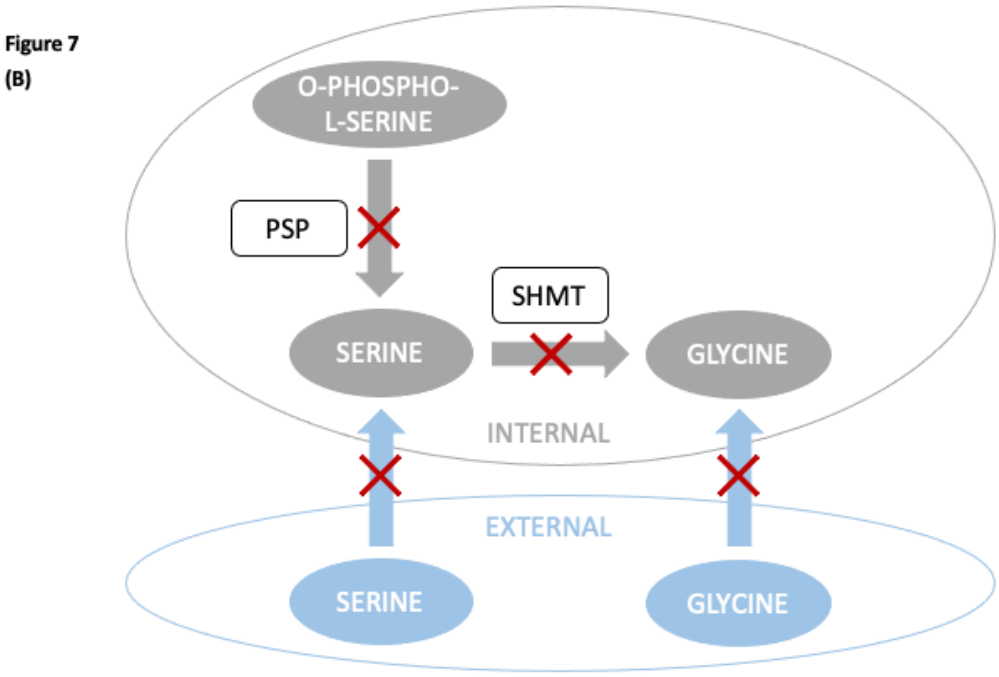


Figure 8. *In silico* drop off experiment: lipids.
The biomass flux predicted by the model under the medium designed by Merali *et al* and the medium conditions, where each lipid is individually omitted from Merali's medium.

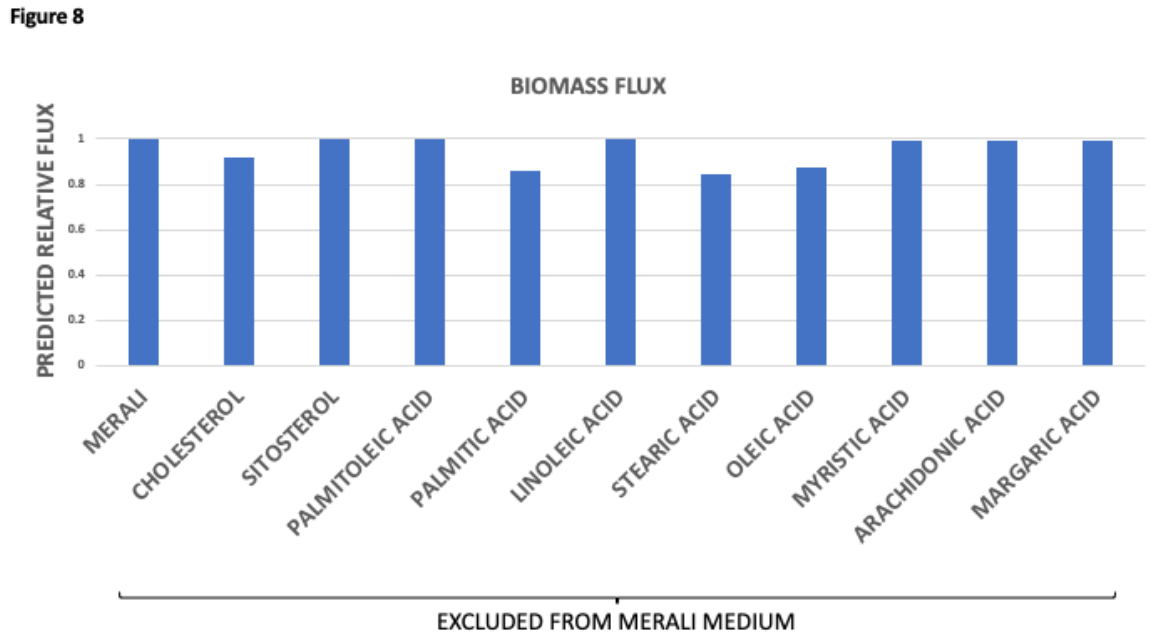
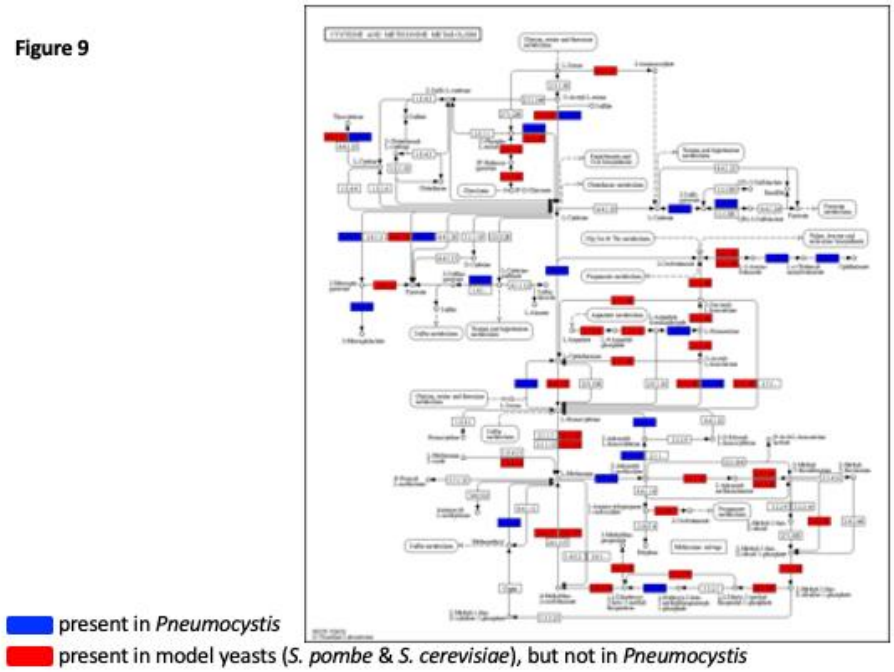


Figure 9. The SAM dilemma.
Visualisation of the *Pneumocystis* cysteine and methionine metabolism in comparison with that of the model yeasts *S. pombe* and *S. cerevisiae* using the KEGG Mapping Tool. The enzymes present in *Pneumocystis* are highlighted in blue, while those present in model yeasts, but not in *Pneumocystis*, are featured in red.



SUPPLEMENTARY INFORMATION

SI 1. Biomass equation reconstruction.

SI 2. Draft reconstruction.

Table S1. Added reactions in the *Pneumocystis* network.
Table S2. Imbalanced reactions in the *Pneumocystis* network.
Table S3. Isolates in the *Pneumocystis* network.
Table S4. Final list of metabolic reactions in the *Pneumocystis* network.

SI 3. List of transport reactions under Merali's medium conditions.

SI 4. Merali's medium optimisation.

Table S5. List of transport reactions under Merali's medium conditions together with the corresponding flux values.
Table S6. List of nutrients in the minimal growth medium predicted by the RL approach.

Figure S1. Schematic representation of the RL approach used in this study.
RL algorithms operate on states, actions, and rewards within an environment. The RL agent decides of which action to take based on the current state. The policy maps states to actions and is updated according to the obtained reward. The objective is to maximize cumulative reward by balancing exploration and high-reward actions.

Figure S1

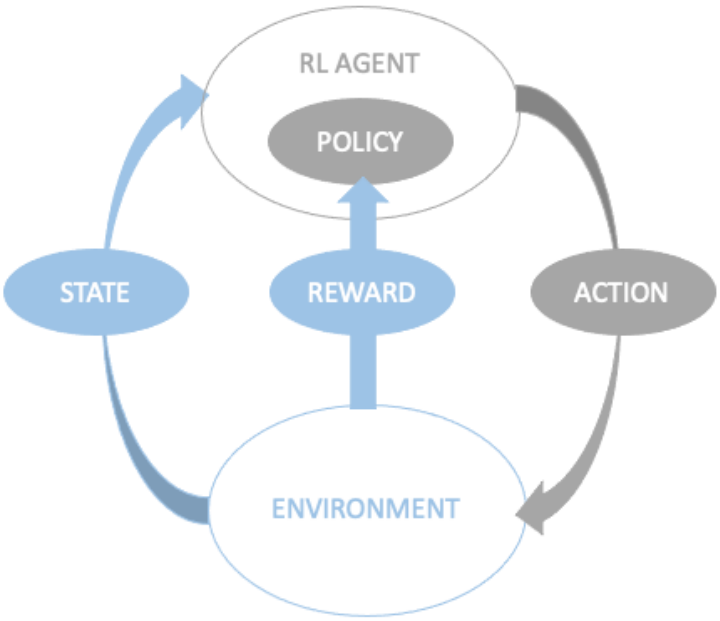


Figure S2. Learning process of the RL approach used in this study.
(A) The learning process of identifying and eliminating non-essential nutrients from the initial list of 155 nutrients is depicted by the blue and orange curves, which represent the precise and average sizes of the nutrient set after 123 attempts to remove nutrients by episodes. The value 123 represents the minimum number of nutrient deletions necessary to reduce the initial set of 155 nutrients to a minimal set consisting of 32 nutrients.
(B) The distribution of the steps on which attempts made to eliminate choline, a non-essential nutrient (depicted on the left), and cholesterol, an essential nutrient (depicted on the right), from the initial list of 155 nutrients in Merali's medium during each episode. The figures indicate that during the learning process the essentiality of cholesterol has been established, suggesting that it should not be removed at the onset of the episode. On the other hand, since choline is non-essential, it is advisable to eliminate it during the middle phase of the episode.

Figure S2

(A)

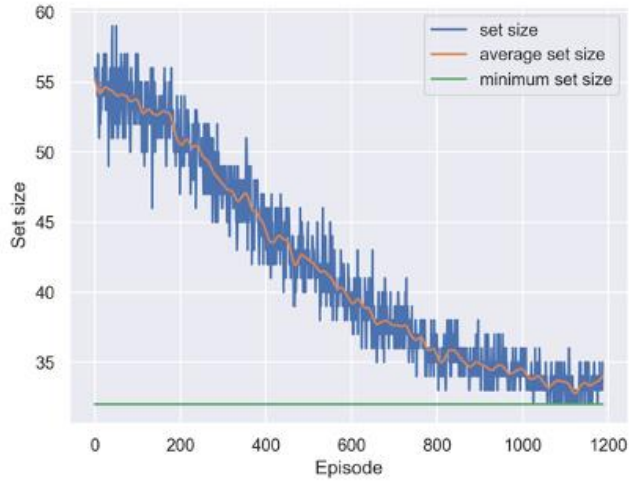
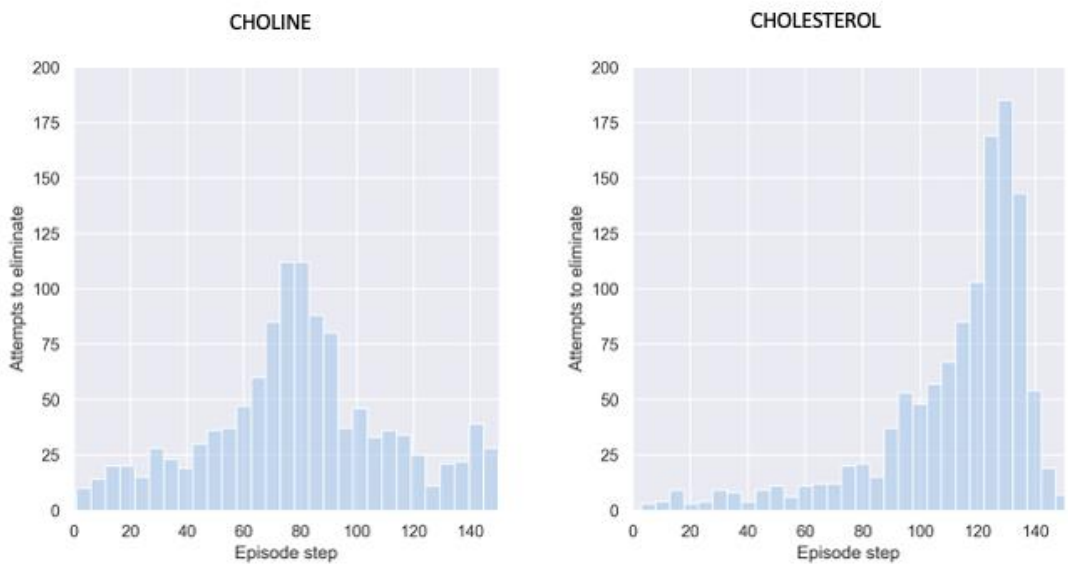


Figure S2

(B)



ACKNOWLEDGEMENTS

We acknowledge funding from the MRC Centre for Medical Mycology at the University of Exeter (MR/N006364/2 and MR/V033417/1) and from the MRC Skills Development Fellowship to O.A.N. (MR/V006169/1). O.A.N. is also grateful for a Ken Haynes Travel Bursary. A.J.P.B was also supported by grants from the MRC (MR/M026663/2) and Wellcome (224323/Z/21/Z). E.Z. is grateful for funding from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors would like to thank the research collaborators who contributed their expertise and knowledge to this project, including Dr Jay K. Kolls, Dr Lucian Duvenage, Dr Thibault Etienne, Benjamin Maisonneuve, and Ilia Ryzhkov for helpful discussions.

REFERENCES

[1] Ivanov II, Atarashi K, Manel N, Brodie EL, Shima T, Karaoz U, Wei D, Goldfarb KC, Santee CA, Lynch SV, Tanoue T, Imaoka A, Itoh K, Takeda K, Umesaki Y, Honda K, Littman DR. 2009. Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell 139(3):485-498
[2] Konikoff T, Gophna U. 2016. *Oscillospira*: a central, enigmatic component of the human gut microbiota. Trends Microbiol 24(7):523–524
[3] Mackie RI, Aminov RI, Hu W, Klieve AV, Ouwerkerk D, Sundset MA, Kamagata Y. 2003. Ecology of uncultivated *Oscillospira* species in the rumen of cattle, sheep, and reindeer as assessed by microscopy and molecular approaches. Appl Environ Microbiol 69(11):6808–6815
[4] Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye Y, Hepburn T, Hu P, Sodergren E, Liolios K, Huot-Creasy H, Birren BW, Earl AM. 2012. The "most wanted" taxa from the human microbiome for whole genome sequencing. PLoS One 7(7):e41294
[5] Li L, Mendis N, Trigui H, Oliver JD, Faucher SP. 2014. The importance of the viable but non-culturable state in human bacterial pathogens. Front Microbiol 5:258
[6] Wagley S, Morcrette H, Kovacs-Simon A, Yang ZR, Power A, Tennant RK, Love J, Murray N, Titball RW, Butler CS. 2021. Bacterial dormancy: a subpopulation of viable but non-culturable cells demonstrates better fitness for revival. PLoS Pathogens 17(1):e1009194
[7] Visvesvara GS. 2002. *In vitro* cultivation of microsporidia of clinical importance. Clin Microbiol Rev 15(3):401-413
[8] Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, Kronstad JW, Deangelis YM, Reeder NL, Johnstone KR, Leland M, Fieno AM, Begley WM, Sun Y, Lacey MP, Chaudhary T, Keough T, Chu L, Sears R, Yuan B, Dawson TL Jr. 2007. Dandruff-associated *Malassezia* genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. Proc Natl Acad Sci U S A 104(47):18730-18735
[9] Beard CB, Fox MR, Lawrence GG, Guarner J, Hanzlick RL, Huang L, del Rio C, Rimland D, Duchin JS, Colley DG. 2005. Genetic differences in *Pneumocystis* isolates recovered from immunocompetent infants and from adults with AIDS: epidemiological implications. J Infect Dis 192:1815–1818
[10] Vargas SL, Hughes WT, Santolaya ME, Ulloa AV, Ponce CA, Cabrera CE, Cumsille F, Gigliotti F. 2001. Search for primary infection by *Pneumocystis carinii* in a cohort of normal, healthy infants. Clin Infect Dis 32:855– 861
[11] de Boer MG, de Fijter JW, Kroon FP. 2011. Outbreaks and clustering of *Pneumocystis* pneumonia in kidney transplant recipients: a systematic review. Med Mycol 49:673–680

[12] Yiannakis EP, Boswell TC. 2016. Systematic review of outbreaks of *Pneumocystis jirovecii* pneumonia: evidence that *P. jirovecii* is a transmissible organism and the implications for healthcare infection control. J Hosp Infect 93:1–8
[13] Buchacz K, Lau B, Jing Y, Bosch R, Abraham AG, Gill MJ, Silverberg MJ, Goedert JJ, Sterling TR, Althoff KN, Martin JN, Burkholder G, Gandhi N, Samji H, Patel P, Rachlis A, Thorne JE, Napravnik S, Henry K, Mayor A, Gebo K, Gange SJ, Moore RD, Brooks JT, North American AIDS Cohort Collaboration on Research and Design (NA-ACCORD) of IeDEA. 2016. Incidence of AIDS-defining opportunistic infections in a multicohort analysis of HIV-infected persons in the United States and Canada, 2000 – 2010. J Infect Dis 214:862– 872
[14] Maini R, Henderson KL, Sheridan EA, Lamagni T, Nichols G, Delpech V, Phin N. 2013. Increasing *Pneumocystis* pneumonia, England, UK, 2000-2010. Emerg Infect Dis 19(3):386–392
[15] Inagaki K, Blackshear C, Hobbs CV. 2019. *Pneumocystis* Infection in Children: National Trends and Characteristics in the United States, 1997-2012. Pediatr Infect Dis J 38(3):241–247
[16] Cushion MT, Tisdale-Macioce N, Sayson SG and Porollo A. 2021. The Persistent Challenge of *Pneumocystis* Growth Outside the Mammalian Lung: Past and Future Approaches. Front Microbiol 12:681474
[17] Kim HU, Kim TY, Lee SY. 2008. Metabolic flux analysis and metabolic engineering of microorganisms. Mol BioSyst 4:113–120
[18] Thiele I, Palsson BØ. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121
[19] Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat Biotechnol 28(3):245–248
[20] Feist AM, Palsson BØ. 2008. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26(6):659–667
[21] Oberhardt MA, Palsson BØ, Papin JA. 2009. Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320
[22] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. 2019. Current status and applications of genome-scale metabolic models. Genome Biol 20(1):121
[23] Teusink B, van Enckevort FHJ, Francke C, Wiersma A, Wegkamp A, Smid EJ, Siezen RJ. 2005. *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. Appl Environ Microbiol 71(11):7253–7262
[24] Magnusdottir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, Greenhalgh K, Jager C, Baginska J, Wilmes P, Fleming RM, Thiele I. 2017. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol 35(1):81–89
[25] Kleerebezem M, Boekhorst J, van Kranenburg R, Molenaar D, Kuipers OP, Leer R, Tarchini R, Peters SA, Sandbrink HM, Fiers MW. 2003. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proc Natl Acad Sci U S A 100(4):1990–1995
[26] Heinken A, Khan MT, Paglia G, Rodionov DA, Harmsen HJ, Thiele I. 2014. Functional metabolic map of *Faecalibacterium prausnitzii*, a beneficial human gut microbe. J Bacteriol 196(18):3289–3302
[27] Becker SA, Palsson BØ. 2005. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. BMC Microbiol 5(1):1
[28] Baart GJ, Zomer B, de Haan A, van der Pol LA, Beuvery EC, Tramper J, Martens DE. 2007. Modeling *Neisseria meningitidis* metabolism: from genome to metabolic fluxes. Genome Biol 8(7):R136
[29] Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BØ. 2002. Genome-scale metabolic model of *Helicobacter pylori* 26695. J Bacteriol 184(16):4582-4593
[30] Merali S, Frevert U, Williams JH, Chin K, Bryan R, Clarkson AB Jr. 1999. Continuous axenic cultivation of *Pneumocystis carinii*. Proc Natl Acad Sci U S A 96(5):2402-2407
[31] Ma L, Chen Z, Huang DW, Kutty G, Ishihara M, Wang H, Abouelleil A, Bishop L, Davey E, Deng R, Deng X, Fan L, Fantoni G, Fitzgerald M, Gogineni E, Goldberg JM, Handley G, Hu X, Huber C, Jiao X,

Jones K, Levin JZ, Liu Y, Macdonald P, Melnikov A, Raley C, Sassi M, Sherman BT, Song X, Sykes S, Tran B, Walsh L, Xia Y, Yang J, Young S, Zeng Q, Zheng X, Stephens R, Nusbaum C, Birren BW, Azadi P, Lempicki RA, Cuomo CA, Kovacs JA. 2016. Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. Nat Commun 7:10740
[32] Eddens T, Elsegeiny W, Ricks D, Goodwin M, Horne WT, Zheng M, Kolls JK. 2019. Transcriptomic and Proteomic Approaches to Finding Novel Diagnostic and Immunogenic Candidates in *Pneumocystis*. mSphere 4(5): e00488-19
[33] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res 44(D1):D1214-D1219
[34] Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D. 2021. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49:D498-D508
[35] Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, Hyka-Nouspikel N, Gasteiger E, Kerhornou A, Neto TB, Pozzato M, Blatter MC, Ignatchenko A, Redaschi N, Bridge A. 2022. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res 50(D1):D693-D700
[36] Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27-30
[37] Santos S, Rocha I. 2016. Estimation of biomass composition from genomic and transcriptomic information. J Integr Bioinform 13(2):285
[38] Kaneshiro ES. 1998. The lipids of *Pneumocystis carinii*. Clin Microbiol Rev 11(1):27-41
[39] Guo Z, Kaneshiro ES. 1995. Phospholipid composition of *Pneumocystis carinii* and effects of methylprednisolone immunosuppression on rat lung lipids. Infect Immun 63(4):1286-1290
[40] Kaneshiro ES, Amit Z, Chandra J, Baughman RP, Contini C, Lundgren B. 1999. Sterols of *Pneumocystis carinii hominis* organisms isolated from human lungs. Clin Diagn Lab Immunol 6(6):970-976
[41] Sohn SB, Kim TY, Lee JH, Lee SY. 2012. Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of *in silico*/*in vivo* mutant growth. BMC Syst Biol 6:49
[42] Lipschik GY, Masur H, Kovacs JA. 1991. Polyamine metabolism in *Pneumocystis carinii*. J Infect Dis 163(5):1121-1127
[43] Vassallo R, Standing JE, Limper AH. 2000. Isolated *Pneumocystis carinii* cell wall glucan provokes lower respiratory tract inflammatory responses. J Immunol 164(7):3755-3763
[44] Ma L, Cissé OH, Kovacs JA. 2018. A Molecular Window into the Biology and Epidemiology of *Pneumocystis* spp. Clin Microbiol Rev 31(3):e00009-18
[45] Kottom TJ, Hebrink DM, Jenson PE, Gudmundsson G, Limper AH. 2015. Evidence for Proinflammatory β-1,6 Glucans in the *Pneumocystis carinii* Cell Wall. Infect Immun 83(7):2816-2826
[46] Oberhardt MA, Puchałka J, Fryer KE, Martins dos Santos VA, Papin JA. 2008. Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. J Bacteriol 190(8):2790-2803
[47] Raman K, Rajagopalan P, Chandra N. 2005. Flux balance analysis of mycolic acid pathway: targets for antitubercular drugs. PLoS Comput Biol 1:e46
[48] Edwards JS, Ibarra RU, Palsson BØ. 2001. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125–130
[49] Varma A, Palsson BØ. 1993. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. Journal of Theoretical Biology 165:477–502
[50] Koffas M, Stephanopoulos G. 2005. Strain improvement by metabolic engineering: lysine production as a case study for systems biology. Curr Opin Biotechnol 16:361–366
[51] García Sánchez CE, Vargas García CA, Torres Sáez RG. 2012. Predictive Potential of Flux Balance Analysis of *Saccharomyces cerevisiae* Using as Optimization Function Combinations of Cell Compartmental Objectives. PLoS One 7(8):e43006

[52] Kaneshiro ES. 2004. Sterol metabolism in the opportunistic pathogen *Pneumocystis*: advances and new insights. Lipids 39(8):753-761
[53] Paulsrud JR, Queener SF. Incorporation of fatty acids and amino acids by cultured *Pneumocystis carinii*. J Eukaryot Microbiol 41(6):633-638
[54] Sutton RS, Barto AG. 2018. Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA.
[55] Gow S, Niranjan M, Kanza S, Frey JG. 2022. A review of reinforcement learning in chemistry, Digital Discovery 1(5):551-567
[56] Zamaraeva E, Collins CM, Antypov D, Gusev VV, Savani R, Dyer MS, Darling GR, Potapov I, Rosseinsky MJ, Spirakis PG. Reinforcement Learning in Crystal Structure Prediction. 2023. ChemRxiv. DOI https://doi.org/10.26434/chemrxiv-2023-4fn8j
[57] Popova M, Isayev O, Tropsha A. 2018. Deep reinforcement learning for de novo drug design. Sci Adv 4(7):eaap7885
[58] Liu M, Shen X, Pan W. 2022. Deep reinforcement learning for personalized treatment recommendation. Stat Med 41(20):4034-4056
[59] Wu X, Li R, He Z, Yu T, Cheng C. 2023. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. NPJ Digit Med 6(1):15
[60] Zhang L, Lengersdorff L, Mikus N, Gläscher J, Lamm C. 2020. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. Soc Cogn Affect Neurosci 15(6):695-707
[61] Merali S, Vargas D, Franklin M, Clarkson AB Jr. 2000. S-adenosylmethionine and *Pneumocystis carinii.* J Biol Chem 275(20):14958-14963
[62] Perez-Leal O, Moncada C, Clarkson AB, Merali S. 2011. *Pneumocystis* S-adenosylmethionine transport: a potential drug target. Am J Respir Cell Mol Biol 45(6):1142-1146
[63] Kutty G, Hernandez-Novoa B, Czapiga M, Kovacs JA. 2008. *Pneumocystis* encodes a functional S-adenosylmethionine synthetase gene. Eukaryot Cell 7(2):258-267
[64] Kaneshiro ES. 2010. Biochemical research elucidating metabolic pathways in *Pneumocystis*. Parasite 17(4):285-291
[65] Pinney JW, Papp B, Hyland C, Wambua L, Westhead DR, McConkey GA. 2007. Metabolic reconstruction and analysis for parasite genomes. Trends Parasitol 23(11):548-554
[66] Kauffman KJ, Prakash P, Edwards JS. 2003. Advances in flux balance analysis. Curr Opin Biotechnol 14:491–496.
[67] Babaei P, Shoaie S, Ji B, Nielsen J. 2018. Challenges in modeling the human gut microbiome. Nat Biotechnol. 36(8):682-686
[68] Raman K, Chandra N. 2009. Flux balance analysis of biological systems: applications and challenges. Briefings in Bioinformatics 10(4):435–449
[69] Covert MW, Xiao N, Chen TJ, Karr JR. 2008. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics 24(18):2044-2050
[70] Min Lee J, Gianchandani EP, Eddy JA, Papin JA. 2008. Dynamic Analysis of Integrated Signaling, Metabolic, and Regulatory Networks. PLOS Comput Biol 4(5):e1000086
[71] Durand-Joly I, Aliouat EM, Recourt C, Guyot K, Francois N, Wauquier M, Camus D, Dei-Cas E. 2002. *Pneumocystis carinii f. sp. hominis* is not infectious for SCID mice. J Clin Microbiol 40:1862–1865
[72] Cissé OH, Pagni M, Hauser PM. 2012. De novo assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. mBio 4:e00428–00412
[73] Slaven BE, Meller J, Porollo A, Sesterhenn T, Smulian AG, Cushion MT. 2006. Draft assembly and annotation of the *Pneumocystis carinii* genome. J Eukaryot Microbiol 53:S89–S91
[74] Cissé OH, Ma L, Dekker JP, Khil PP, Youn JH, Brenchley JM, Blair R, Pahar B, Chabé M, Van Rompay KKA, Keesler R, Sukura A, Hirsch V, Kutty G, Liu Y, Peng L, Chen J, Song J, Weissenbacher-Lang C, Xu J, Upham NS, Stajich JE, Cuomo CA, Cushion MT, Kovacs JA. 2021. Genomic insights into the host specific adaptation of the *Pneumocystis* genus. Commun Biol 4:305

[75] Susi H, Barres B, Vale PF, Laine AL. 2015. Co-infection alters population dynamics of infectious disease. Nat Commun 6:5975
[76] Prestes-Carneiro LE, Junior NGS, Oliveira MR, Auer LS, Duarte AJS, Almeida A, Vasconcelos DM. 2014. Recurrent pneumocystosis pneumonia/chronic obstructive pulmonary disease and mild immuno-deficiency in a human immunodeficiency virus-negative subject. JMM Case Rep 1:1–5
[77] Queener SF, Cody V, Pace J, Torkelson P, Gangjee A. 2013. Trimethoprim resistance of dihydrofolate reductase variants from clinical isolates of *Pneumocystis jirovecii*. Antimicrob Agents Chemother 57:4990–4998
[78] Ames R, Brown AJP, Gudelj I, Nev OA. 2023. Analysis of *Pneumocystis* Transcription Factor Evolution and Implications for Biology and Lifestyle. mBio 14(1):e0271122
[79] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of molecular biology 215(3):403–410
[80] UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic acids research 47(D1):D506-D515
[81] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):D279-85
[82] GO Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic acids research 47(D1):D330-D338
[83] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). 1992. Enzyme Nomenclature. Academic Press
[84] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40(Database issue):D700-D705.
[85] Harris MA, Rutherford KM, Hayles J, Lock A, Bähler J, Oliver SG, Mata J, Wood V. 2022. Fission stories: using PomBase to understand *Schizosaccharomyces pombe* biology. Genetics 220(4):iyab222
[86] https://gencovery.com
[87] Wright L, Davidson S. How to tell the difference between a model and a digital twin. 2020. Adv Model and Simul in Eng Sci 7(1):1-13
[88] https://www.ebi.ac.uk/
[89] https://www.ncbi.nlm.nih.gov/taxonomy
[90] Mo ML, Palsson BO, Herrgård MJ. 2009. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol 3:37
[91] Sobolewska A, Dzbeński TH. 2009. *In vitro* cultivation of *Pneumocystis* isolated from infected rat lungs. Wiad Parazytol 55(4):451-457
[92] Larsen HH, Kovacs JA, Stock F, Vestereng VH, Lundgren B, Fischer SH, Gill VJ. 2002. Development of a rapid real-time PCR assay for quantitation of *Pneumocystis carinii f. sp. carinii*. J Clin Microbiol 40(8):2989-2993
[93] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017. Proximal Policy Optimization Algorithms. arXiv 1707.06347 [cs.LG]. DOI https://doi.org/10.48550/arXiv.1707.06347
[94] Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. 2021. Stable-baselines3: Reliable reinforcement learning implementations. J Mach Learn Res 22(1):12348-55