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

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

[1] Faculty of Health and Life Sciences, 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 empowering potential applications. However, a significant number of bacterial and fungal species, including *Pneumocystis jirovecii*, remain unculturable, hampering research efforts. *P. jirovecii* is a deadly pathogen of humans that causes life-threatening pneumonia in immunocompromised individuals and transplant patients. Despite the major impact of *Pneumocystis* on human health, we remain ignorant about the pathobiology of this fungus. This is largely due to the fact that its experimental dissection has been constrained by the inability to culture the organism *in vitro*. We present a comprehensive *in silico* genome-scale metabolic model of *Pneumocystis* growth and metabolism, with a view to identifying metabolic requirements and imbalances that hinder growth *in vitro*. We utilise recently published genome data and available information in the literature as well as bioinformatics and software tools to develop and validate the model. In addition, we employ Flux Balance Analysis and Reinforcement Learning approaches to make predictions regarding metabolic fluxes and to identify 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

The experimental dissection of microorganisms, such as bacteria and fungi, is greatly facilitated by access to 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 the potential applications of these organisms. By cultivating bacteria and fungi, we are able to investigate mechanisms that underlie the infectious diseases they elicit, refine diagnostic assays for these diseases, and develop novel vaccines and therapeutic interventions.

A significant number of bacterial and fungal species 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* species 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]. Viable but non-culturable (VBNC) cells represent another group of bacteria that is challenging to culture. 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 skin disorders that include 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]. *Pneumocystis* 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). The fungus is also an emerging issue in non-HIV patients in developed countries, such as UK [14] and USA [15]. Although *Pneumocystis* has a major impact on human health, we remain remarkably ignorant about the biology and epidemiology of this opportunistic pathogen. 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 an organism'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. Potential objectives include maximizing growth or reducing 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, metabolic modelling has been used to design and optimise the composition of culture media. For instance, GEMs have been successfully applied to formulate *in vitro* growth media for bacteria such as *Lactobacillus plantarum* [23] and

Bacteroides caccae [24]. In addition, GEMs have been used to design minimal growth media for the lactic acid bacterium *Lactococcus lactis* [25], a potentially beneficial gut microbe *Faecalibacterium prausnitzii* [26], and 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-based 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 Flux Balance Analysis (FBA), 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 reinforcement learning approach.

METABOLIC MODEL DEVELOPMENT

To develop a metabolic model for *Pneumocystis*, we 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 published information about the organism's phenotypes under various environmental conditions. A 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 nutrient transport 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*). Finally, to get the most accurate results possible, we performed re-evaluation and manual curation of the entire draft reconstruction and the associated network content. We then used the brick GENA (Genome-Based Network Analysis) of Constellab™ platform (see *Materials and Methods*) 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 487 chemical reactions corresponding to 495 enzymes thought to operate in a *Pneumocystis* cell. We compared *Pneumocystis* metabolism with that of the model yeasts *S. pombe* and *S. cerevisiae* using the KEGG Mapping Tool [36] (Fig 2A). This metabolic map clearly shows that *P. murina* has a reduced

genome that lacks metabolic pathways essential in 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 to synthesize a crucial arginine precursor, ornithine, from glutamate and to convert it to citrulline. Furthermore, the production of arginine from aspartate is also blocked due to the lack of the requisite enzymes. Other yeasts possess these enzymes and, therefore, are capable of arginine biosynthesis. Interestingly, according to this metabolic map, *Pneumocystis* can 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 broken down into their 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 organism of interest. 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 the relative fractions 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 the 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.

FLUX BALANCE 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 (Fig 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 quasi-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 are used to reduce the FBA solution space. These constraints rely on biological knowledge or experiment data on the organism. For instance, experimental data can be used to derive intake/excretion fluxes and thus constraints on extracellular fluxes. However, the inability to cultivate *Pneumocystis in vitro* has led to a deficiency of essential empirical information. Thus, we are mostly dependent on the reconstruction, open data collection and curation processes used to contextualize the model before optimization.

Since we are interested in predicting growth in *Pneumocystis*, our biological objective is its biomass production. 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*). Fig 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 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. 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. Fig 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 the differential contributions of particular amino acids to the overall biomass (see *Biomass Equation* 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, for example. These 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 (Fig 6). 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 enzyme serine hydroxymethyltransferase (SHMT, EC 2.1.2.1) catalyses the interconversion of serine and glycine. Therefore, the gene encoding SHMT was conceptually "switched off", to simulate a typical gene knockout experiment in the laboratory. The model predicts that the biomass flux remains relatively unchanged for wild type cells under the medium conditions designed by Merali and co-workers, even when serine, glycine, or both are omitted from the medium (Fig 7A). However, when the SHMT enzyme is also "switched off", the biomass flux drops due to the inability of *Pneumocystis* to interconvert serine into glycine. Essentially, this *in silico* experiment corresponds to the inhibition of serine-glycine interconversion as well as the exclusion of both amino acids from the medium (Fig 7B). Since serine and glycine contribute significantly to the amino acid content of *Pneumocystis* biomass, their elimination is predicted to lead to a decrease in the biomass production.

Critical role of lipids in the Pneumocystis growth

To identify the key lipids in Merali's medium that are crucial for *Pneumocystis* growth, we evaluated their impact on biomass flux using an analogous approach. Not all lipids appear to contribute equally to *Pneumocystis* growth. Fig 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 was the most abundant sterol in the *Pneumocystis* cell membrane [38], [52], whereas palmitic, stearic, and oleic fatty acids were the dominant fatty acids [38]. Moreover, it has been demonstrated experimentally that the pathogen is able to assimilate cholesterol and 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 and co-workers in their growth medium [30]. This discovery led us to formulate the hypothesis that supplementing the medium with extra 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. The model 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 represents a valuable tool to make predictions regarding growth medium composition and can be used to design new culture media or to refine existing growth conditions. Here, we demonstrate how the model can facilitate the optimisation of Merali's medium. In particular, we illustrate how the number of components in Merali's medium might be reduced while maintaining equivalent biomass flux values, without compromising growth performance.

To predict 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) [30], 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 exploiting RL to optimise the composition of Merali's growth medium (see *Materials and Methods*), we were able to decrease the initial set of 155 nutrients by about 76% to only 37 nutrients while maintaining the biomass flux value. During the learning process (300,000 iterations) the RL algorithm was successfully trained to identify and remove non-essential nutrients (Fig S2). The algorithm predicted 815 minimal sets of nutrients of size 37, where 58 of these sets were unique. Interestingly, 29 core nutrients were the same in all minimal sets: the other 8 nutrients came from a list of 19 non-core nutrients that varied from set to set (SI 4, Table S6). Table 2 shows the list of 37 nutrients that appeared most frequently, with a total of 649 occurrences.

We also verified the flux values computed by the metabolic model under Merali's medium conditions and discovered 91 zero fluxes associated with extracellular transport reactions (SI 4, Table S5). These fluxes do not appear to 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 37 nutrients.

By using this task as an example, we showed that our novel RL 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 reinforce our previous findings (see *MODEL PREDICTIONS*, above) 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 (Table 2). 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 core components of Merali's medium required for *Pneumocystis* growth. Indeed, it was shown experimentally that these are dominant lipids in the composition of the pathogen [38], [52] and that *Pneumocystis* cells are able to assimilate cholesterol and these fatty acids from the environment [52], [53].

Moreover, Merali and co-workers reported that *Pneumocystis* growth was observed when p-aminobenzoic acid, putrescine, and protoheme were omitted individually, but that 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 via 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 that it is absolutely required for continuous growth of the fungus [30]. Furthermore, a *Pneumocystis* gene with homology to mitochondrial SAM transporters was discovered, which supported 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 both SAM and L-methionine are 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 their 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 confirm 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 reinforcement 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 has the potential to scavenge nutrients from the host [31]. This implies that the metabolic network of *Pneumocystis* may exhibit genuine gaps that reflect the pathogen's reliance on its host. These gaps should not be filled 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 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 account of regulatory effects such as enzyme activation by protein kinases or gene regulation, nor does it consider stress response or signalling pathways, and so its predictions may not always be accurate. To improve the accuracy of

FBA predictions, 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 experimental data. 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 activities. 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 databases that are not organism-specific, 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 here by the model need to be tested experimentally. Moreover, the growth of *Pneumocystis in vitro* should 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 includes several species, each of which is 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], leading, in the longer-term, to the development of new antifungal therapies that circumvent the emerging resistance of *Pneumocystis* to azole drugs [31], [77], and 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*. 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 relatively closely related to it in phylogenetic terms [78]. In particular, we selected *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* as comparison organisms due to the availability of functional annotations for these species.

To identify homologous genes between these organisms and *P. murina*, we used BLAST [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 [81] and *PomBase* Genome Database [82], respectively.

We then used PFAM [83] (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 [84], [85].

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

This allowed us to generate a comprehensive list of 495 EC numbers corresponding to the enzymes present in *P. murina*, which provided a starting point to build a metabolic reconstruction network.

Constellab™ software tool

To build a metabolic reconstruction network based on the 495 EC numbers that correspond to the enzymes found in *P. murina*, we used a Constellab™ platform (configuration gws_core: 0.7.2, gws_gena: 0.6.3, gws_biota: 0.6.1, exe_gena: 0.2.0, gws_omix: 0.6.1), 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 open 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 of cellular metabolism [87], which are digital representations of the metabolic processes that can be used to simulate, analyse, and optimise their performance and efficiency under various situations. In this way, the Constellab™ platform allowed us 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 open European EMBL-EBI database [88] and the NCBI taxonomy database [89]. The BIOTA 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]. By mapping our data to the BIOTA, we were able to build an initial version of the *P. murina* metabolic network, which consisted of 528 chemical reactions that corresponded to the 495 identified enzymes (see *Bioinformatics tools*, above). The mapping was performed using algorithms available in the open brick GENA.

Manual curation

The initial automated draft reconstruction, comprising 528 reactions, was subjected to additional manual curation (SI 2). As a result, four additional reactions were manually incorporated by drawing information from the KEGG [36] and RHEA [35] databases, as well as from the *S. cerevisiae* model [90] (SI 2, Table S1).

We then assessed whether the network was balanced in terms of mass and charge and identified 52 reactions displaying imbalances in mass and/or charge. Most of these imbalanced reactions contained chemical compounds in a generic form ((1→4)-α-D-glucan being represented as $(1,4\text{-}\alpha\text{-D-glucosyl})_n$, for example), making it impossible to calculate their mass or charge correctly. Such generic reactions are considered non-specific due to insufficient knowledge and biochemical evidence, and hence were typically excluded from the network [18]. As a result, 44 imbalanced reactions were deleted. The other 8 reactions were kept due to their retention in the *S. cerevisiae* [90] and *S. pombe* models [41] as well as their importance according to the literature and likely contribution to the biomass production. The stoichiometry of all reactions was sourced from the RHEA database [35] (SI 2, Table S2).

Next, we eliminated the "isolates" from the reconstruction, i.e. those reactions that were not topologically linked to biomass production and, thus, were not predicted to contribute

to *Pneumocystis* growth, 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. Two orphan reactions were excluded from the model (SI 2, Table S3). Finally, we derived 5 reactions from the databases and manually added them due to their association with corresponding enzymes that are present in the *Pneumocystis* metabolic network.

The resulting draft reconstruction contained 487 reactions (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 and co-workers (see *Merali's culture medium composition*, below). 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 predicting 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.

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 in 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 were 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*).

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 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 medium exchange twice-daily was crucial for the growth. Furthermore, other researchers 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 (RL) 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 with which the agent interacts 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 this RL approach 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.

An 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 chose the Proximal Policy Optimization (PPO) algorithm [93] as a particular RL method because PPO is widely regarded as a state-of-the-art RL algorithm known for its effectiveness in various domains and it 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*.

Table 2. Most frequently occurring optimal minimal 37-nutrient set predicted by the Reinforcement Learning algorithm.

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 was utilised, reviewed, and refined using bioinformatics tools and the literature. This information was further combined with RNA-Seq data and manually curated to create the list of functional enzymes present in *Pneumocystis*. From this list, the draft reconstruction was generated using the *Gencovery* software Constellab™ and the CHEBI, BRENDA, and RHEA databases.

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

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. Some of these are made up of smaller micro-nutrients that include amino acids, nucleotides, and fatty acids.

The 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* for details). FBA relies on the assumption that the metabolic system is in a steady state. Mathematically, biomass production 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; growth medium designed by Merali *et al* [30] and Merali's medium supplemented with the additional lipids present in the *Pneumocystis* biomass equation. Biomass flux is expressed relative to that observed for Merali's medium.

Figure 6. *In silico* drop off experiment: amino acids.
The biomass flux predicted by the model in the medium designed by Merali *et al* [30]; and in the same medium but where each amino acid is individually omitted. Biomass flux is expressed relative to that observed for Merali's medium.

Figure 7. *In silico* gene knockout experiment: serine and glycine interconversion.
(A) The biomass flux predicted by the model in the medium designed by Merali *et al* [30] and in equivalent medium lacking serine, glycine or both, or where the SHMT enzyme is "switched off". Biomass flux is expressed relative to that observed for Merali's medium.
(B) Schematic representation of the above *in silico* experiment.

Figure 8. *In silico* drop off experiment: lipids.
The biomass flux predicted by the model in the medium designed by Merali *et al* [30] and in Merali's medium lacking each individual lipid. Biomass flux is expressed relative to that observed for Merali's medium.

Figure 9. The SAM dilemma.
Comparison of cysteine and methionine metabolism in *Pneumocystis* and in *S. pombe* and *S. cerevisiae* using the KEGG Mapping Tool. Enzymes present in *Pneumocystis* are highlighted in blue, while those present in model yeasts are 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 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 118 attempts to remove nutrients by episodes. The value 118 represents the minimum number of nutrient deletions necessary to reduce the initial set of 155 nutrients to a minimal set consisting of 37 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.

ACKNOWLEDGEMENTS

We are grateful for support from the MRC Centre for Medical Mycology at the University of Exeter (MR/N006364/2) https://www.ukri.org/councils/mrc/. This study/research is funded by the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC) https://www.exeterbrc.nihr.ac.uk/. O.A.N. was the recipient of a Skills Development Fellowship from the MRC (MR/V006169/1) https://www.ukri.org/councils/mrc/ and an NIHR BRC Exeter Translational Fellowship (NIHR203320) https://www.exeterbrc.nihr.ac.uk/. A.J.P.B was supported by grant from Wellcome (224323/Z/21/Z) https://wellcome.org/. E.Z. acknowledges funding from the Leverhulme Trust via the Leverhulme Research Centre for Functional Materials Design (RC-2015-036) https://www.leverhulme.ac.uk/. L. D. acknowledges funding from the Carnegie Corporation of New York (CCNY) https://www.carnegie.org/. J.C.H. was supported by grant from Wellcome (209293) https://wellcome.org/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
The authors have declared that no competing interests exist.

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

REFERENCES

1. Ivanov II, Atarashi K, Manel N, Brodie EL, Shima T, Karaoz U, et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell. 2009;139(3):485-98.
2. Konikoff T, Gophna U. *Oscillospira*: a central, enigmatic component of the human gut microbiota. Trends in microbiology. 2016;24(7):523-4.
3. Mackie RI, Aminov RI, Hu W, Klieve AV, Ouwerkerk D, Sundset MA, Kamagata Y. Ecology of uncultivated *Oscillospira* species in the rumen of cattle, sheep, and reindeer as assessed by microscopy and molecular approaches. Applied and environmental microbiology. 2003;69(11):6808-15.
4. Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye Y, Hepburn T, et al. The "most wanted" taxa from the human microbiome for whole genome sequencing. 2012.
5. Li L, Mendis N, Trigui H, Oliver JD, Faucher SP. The importance of the viable but non-culturable state in human bacterial pathogens. Frontiers in microbiology. 2014;5:258.
6. Wagley S, Morcrette H, Kovacs-Simon A, Yang ZR, Power A, Tennant RK, et al. Bacterial dormancy: A subpopulation of viable but non-culturable cells demonstrates better fitness for revival. PLoS Pathogens. 2021;17(1).
7. Visvesvara GS. *In vitro* cultivation of microsporidia of clinical importance. Clinical microbiology reviews. 2002;15(3):401-13.
8. Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, et al. Dandruff-associated Malassezia genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. Proceedings of the National Academy of Sciences. 2007;104(47):18730-5.
9. Beard CB, Fox MR, Lawrence GG, Guarner J, Hanzlick RL, Huang L, et al. Genetic differences in *Pneumocystis* isolates recovered from immunocompetent infants and from adults with AIDS: epidemiological implications. The Journal of infectious diseases. 2005;192(10):1815-8.
10. Vargas SL, Hughes WT, Santolaya ME, Ulloa AV, Ponce CA, Cabrera CE, et al. Search for primary infection by *Pneumocystis carinii* in a cohort of normal, healthy infants. Clinical infectious diseases. 2001;32(6):855-61.
11. de Boer MG, de Fijter JW, Kroon FP. Outbreaks and clustering of *Pneumocystis* pneumonia in kidney transplant recipients: a systematic review. Medical mycology. 2011;49(7):673-80.
12. Yiannakis E, Boswell T. Systematic review of outbreaks of *Pneumocystis jirovecii* pneumonia: evidence that *P. jirovecii* is a transmissible organism and the implications for healthcare infection control. Journal of Hospital Infection. 2016;93(1):1-8.
13. Buchacz K, Lau B, Jing Y, Bosch R, Abraham AG, Gill MJ, et al. Incidence of AIDS-defining opportunistic infections in a multicohort analysis of HIV-infected persons in the

United States and Canada, 2000–2010. The Journal of infectious diseases. 2016;214(6):862-72.
14. Maini R, Henderson KL, Sheridan EA, Lamagni T, Nichols G, Delpech V, Phin N. Increasing *Pneumocystis* pneumonia, England, UK, 2000–2010. Emerging infectious diseases. 2013;19(3):386-92.
15. Inagaki K, Blackshear C, Hobbs CV. *Pneumocystis* infection in children: national trends and characteristics in the United States, 1997–2012. The Pediatric Infectious Disease Journal. 2019;38(3):241-7.
16. Cushion MT, Tisdale-Macioce N, Sayson SG, Porollo A. The persistent challenge of *Pneumocystis* growth outside the mammalian lung: past and future approaches. Frontiers in Microbiology. 2021;12:681474.
17. Kim HU, Kim TY, Lee SY. Metabolic flux analysis and metabolic engineering of microorganisms. Molecular BioSystems. 2008;4(2):113-20.
18. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols. 2010;5:93-121.
19. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature biotechnology. 2010;28(3):245-8.
20. Feist AM, Palsson BØ. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nature biotechnology. 2008;26(6):659-67.
21. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. Molecular systems biology. 2009;5.
22. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome biology. 2019;20:1-18.
23. Teusink B, van Enckevort FH, Francke C, Wiersma A, Wegkamp A, Smid EJ, Siezen RJ. *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. Applied and environmental microbiology. 2005;71(11):7253-62.
24. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature biotechnology. 2017;35(1):81-9.
25. Kleerebezem M, Boekhorst J, Van Kranenburg R, Molenaar D, Kuipers OP, Leer R, et al. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proceedings of the National Academy of Sciences. 2003;100(4):1990-5.
26. Heinken A, Khan MT, Paglia G, Rodionov DA, Harmsen HJ, Thiele I. Functional metabolic map of *Faecalibacterium prausnitzii*, a beneficial human gut microbe. Journal of bacteriology. 2014;196(18):3289-302.
27. Becker SA, Palsson BØ. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC microbiology. 2005;5(1):1-12.
28. Baart GJ, Zomer B, de Haan A, van der Pol LA, Beuvery EC, Tramper J, Martens DE. Modeling *Neisseria meningitidis* metabolism: from genome to metabolic fluxes. Genome biology. 2007;8:1-20.
29. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BØ. Genome-scale metabolic model of *Helicobacter pylori* 26695. Journal of Bacteriology. 2002;184(16):4582-93.

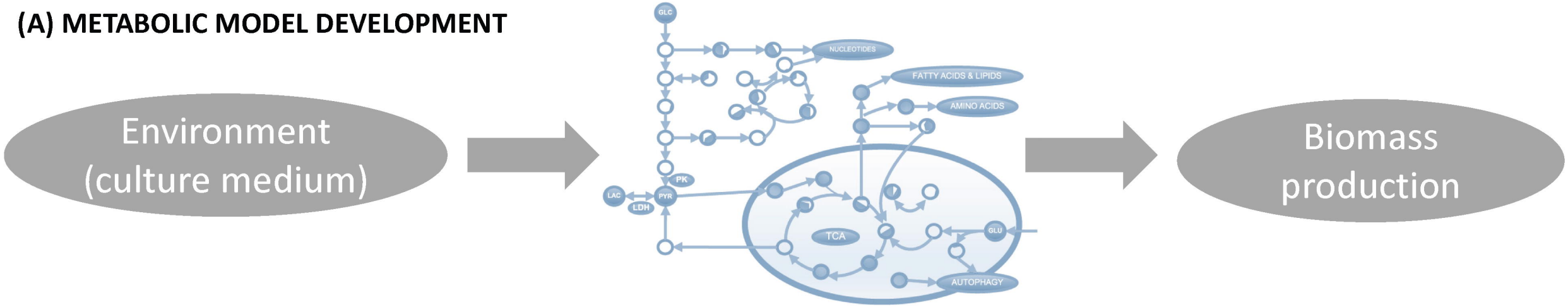
30. Merali S, Frevert U, Williams JH, Chin K, Bryan R, Clarkson ABJ. Continuous axenic cultivation of *Pneumocystis carinii*. Proceedings of the National Academy of Sciences. 1999;96(5):2402-7.
31. Ma L, Chen Z, Huang DW, Kutty G, Ishihara M, Wang H, et al. Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. Nature communications. 2016;7.
32. Eddens T, Elsegeiny W, Ricks D, Goodwin M, Horne WT, Zheng M, Kolls JK. Transcriptomic and proteomic approaches to finding novel diagnostic and immunogenic candidates in Pneumocystis. Msphere. 2019;4(5):10.1128/msphere. 00488-19.
33. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research. 2016;44(D1):D1214-D9.
34. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic acids research. 2021;49(D1):D498-D508.
35. Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L, et al. Rhea, the reaction knowledgebase in 2022. Nucleic acids research. 2022;50(D1):D693-D700.
36. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27-30.
37. Santos S, Rocha I. Estimation of biomass composition from genomic and transcriptomic information. Journal of Integrative Bioinformatics. 2016;13(2).
38. Kaneshiro ES. The lipids of *Pneumocystis carinii*. Clinical microbiology reviews. 1998;11(1):27-41.
39. Guo Z, Kaneshiro ES. Phospholipid composition of *Pneumocystis carinii* carinii and effects of methylprednisolone immunosuppression on rat lung lipids. Infection and immunity. 1995;63(4):1286-90.
40. Kaneshiro ES, Amit Z, Chandra J, Baughman RP, Contini C, Lundgren B. Sterols of *Pneumocystis carinii hominis* organisms isolated from human lungs. Clinical Diagnostic Laboratory Immunology. 1999;6(6):970-6.
41. Sohn SB, Kim TY, Lee JH, Lee SY. Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of *in silico/in vivo* mutant growth. BMC systems biology. 2012;6.
42. Lipschik GY, Masur H, Kovacs JA. Polyamine metabolism in *Pneumocystis carinii*. Journal of Infectious Diseases. 1991;163(5):1121-7.
43. Vassallo R, Standing JE, Limper AH. Isolated *Pneumocystis carinii* cell wall glucan provokes lower respiratory tract inflammatory responses. The Journal of Immunology. 2000;164(7):3755-63.
44. Ma L, Cissé OH, Kovacs JA. A molecular window into the biology and epidemiology of *Pneumocystis spp*. Clinical microbiology reviews. 2018;31(3):e00009-18.
45. Kottom TJ, Hebrink DM, Jenson PE, Gudmundsson G, Limper AH. Evidence for proinflammatory β-1, 6 glucans in the *Pneumocystis carinii* cell wall. Infection and Immunity. 2015;83(7):2816-26.
46. Oberhardt MA, Puchałka J, Fryer KE, Martins dos Santos VA, Papin JA. Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. Journal of Bacteriology. 2008;190(8):2790-803.
47. Raman K, Rajagopalan P, Chandra N. Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. PLoS computational biology. 2005;1(5).

48. Edwards JS, Ibarra RU, Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. Nature biotechnology. 2001;19(2):125-30.
49. Varma A, Palsson BØ. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. Journal of theoretical biology. 1993;165(4):477-502.
50. Koffas M, Stephanopoulos G. Strain improvement by metabolic engineering: lysine production as a case study for systems biology. Current Opinion in Biotechnology. 2005;16(3):361-6.
51. García Sánchez CE, Vargas García CA, Torres Sáez RG. Predictive potential of flux balance analysis of Saccharomyces cerevisiae using as optimization function combinations of cell compartmental objectives. 2012.
52. Kaneshiro ES. Sterol metabolism in the opportunistic pathogen *Pneumocystis*: advances and new insights. Lipids. 2004;39(8):753-61.
53. Paulsrud JR, Queener SF. Incorporation of fatty acids and amino acids by cultured *Pneumocystis carinii*. Journal of Eukaryotic Microbiology. 1994;41(6):633-8.
54. Sutton RS, Barto AG. Reinforcement learning: An Introduction. A Bradford Book, Cambridge, MA, USA. 2018.
55. Gow S, Niranjan M, Kanza S, Frey JG. A review of reinforcement learning in chemistry. Digital Discovery. 2022.
56. Zamaraeva E, Collins CM, Antypov D, Gusev VV, Savani R, Dyer MS, et al. Reinforcement Learning in Crystal Structure Prediction2023:[ChemRxiv p.]. Available from: https://doi.org/10.26434/chemrxiv-2023-4fn8j
57. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Science advances. 2018;4(7).
58. Liu M, Shen X, Pan W. Deep reinforcement learning for personalized treatment recommendation. Statistics in medicine. 2022;41(20):4034-56.
59. Wu X, Li R, He Z, Yu T, Cheng C. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. NPJ Digital Medicine. 2023;6(1).
60. Zhang L, Lengersdorff L, Mikus N, Gläscher J, Lamm C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. Social Cognitive and Affective Neuroscience. 2020;15(6):695-707.
61. Merali S, Vargas D, Franklin M, Clarkson ABJ. S-adenosylmethionine and *Pneumocystis carinii*. Journal of Biological Chemistry. 2000;275(20):14958-63.
62. Perez-Leal O, Moncada C, Clarkson AB, Merali S. *Pneumocystis* S-adenosylmethionine transport: a potential drug target. American Journal of Respiratory Cell and Molecular Biology. 2011;45(6):1142-6.
63. Kutty G, Hernandez-Novoa B, Czapiga M, Kovacs JA. *Pneumocystis* encodes a functional S-adenosylmethionine synthetase gene. Eukaryotic cell. 2008;7(2):258-67.
64. Kaneshiro ES. Biochemical research elucidating metabolic pathways in *Pneumocystis*. Parasite. 2010;17(4):285-91.
65. Pinney JW, Papp B, Hyland C, Wambua L, Westhead DR, McConkey GA. Metabolic reconstruction and analysis for parasite genomes. Trends in Parasitology. 2007;23(11):548-54.
66. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Current opinion in biotechnology. 2003;14(5):491-6.
67. Babaei P, Shoaie S, Ji B, Nielsen J. Challenges in modeling the human gut microbiome. Nature biotechnology. 2018;36(8):682-6.

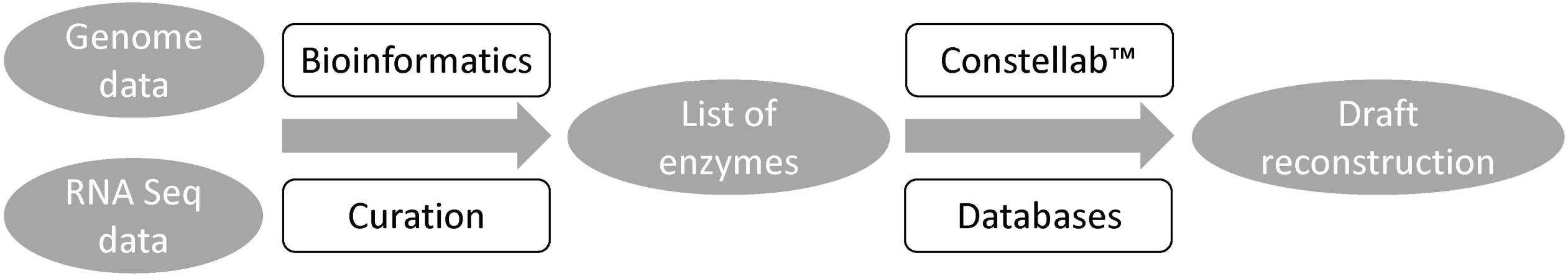
68. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Briefings in bioinformatics. 2009;10(4):435-49.
69. Covert MW, Xiao N, Chen TJ, Karr JR. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. Bioinformatics. 2008;24(18):2044-50.
70. Min Lee J, Gianchandani EP, Eddy JA, Papin JA. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS computational biology. 2008;4(5).
71. Durand-Joly I, Aliouat EM, Recourt Cl, Guyot K, François N, Wauquier Ml, et al. *Pneumocystis carinii f. sp. hominis* is not infectious for SCID mice. Journal of clinical microbiology. 2002;40(5):1862-5.
72. Cissé OH, Pagni M, Hauser PM. De novo assembly of the Pneumocystis jirovecii genome from a single bronchoalveolar lavage fluid specimen from a patient. MBio. 2013;4(1):10.1128/mbio. 00428-12.
73. Slaven BE, Meller J, Porollo A, Sesterhenn T, Smulian AG, Cushion MT. Draft assembly and annotation of the *Pneumocystis carinii* genome. Journal of Eukaryotic Microbiology. 2006;53:S89-S91.
74. Cissé OH, Ma L, Dekker JP, Khil PP, Youn J-H, Brenchley JM, et al. Genomic insights into the host specific adaptation of the Pneumocystis genus. Communications biology. 2021;4(1):305.
75. Susi H, Barrès B, Vale PF, Laine A-L. Co-infection alters population dynamics of infectious disease. Nature communications. 2015;6.
76. Prestes-Carneiro LE, Junior NGS, Oliveira MR, Auer LS, Duarte AJS, Almeida A, Vasconcelos DM. Recurrent pneumocystosis pneumonia/chronic obstructive pulmonary disease and mild immunodeficiency in a human immunodeficiency virus-negative subject. JMM Case Reports. 2014;1(2):1-5.
77. Queener S, Cody V, Pace J, Torkelson P, Gangjee A. Trimethoprim resistance of dihydrofolate reductase variants from clinical isolates of *Pneumocystis jirovecii*. Antimicrobial agents and chemotherapy. 2013;57(10):4990-8.
78. Ames R, Brown AJ, Gudelj I, Nev OA. Analysis of *Pneumocystis* Transcription Factor Evolution and Implications for Biology and Lifestyle. mBio. 2023;14(1):e02711-22.
79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
80. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research. 2018;47(D1):D506-D15. doi: 10.1093/nar/gky1049.
81. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucleic Acids Research. 2011;40(D1):D700-D5. doi: 10.1093/nar/gkr1029.
82. Harris MA, Rutherford KM, Hayles J, Lock A, Bähler J, Oliver SG, et al. Fission stories: using PomBase to understand Schizosaccharomyces pombe biology. Genetics. 2021;220(4). doi: 10.1093/genetics/iyab222.
83. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research. 2015;44(D1):D279-D85. doi: 10.1093/nar/gkv1344.
84. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research. 2019;47(D1):D330-D8. doi: 10.1093/nar/gky1055.
85. (NC-IUBMB) NCotIUoBaMB. Enzyme Nomenclature. Academic Press1992.
86. https://gencovery.com.

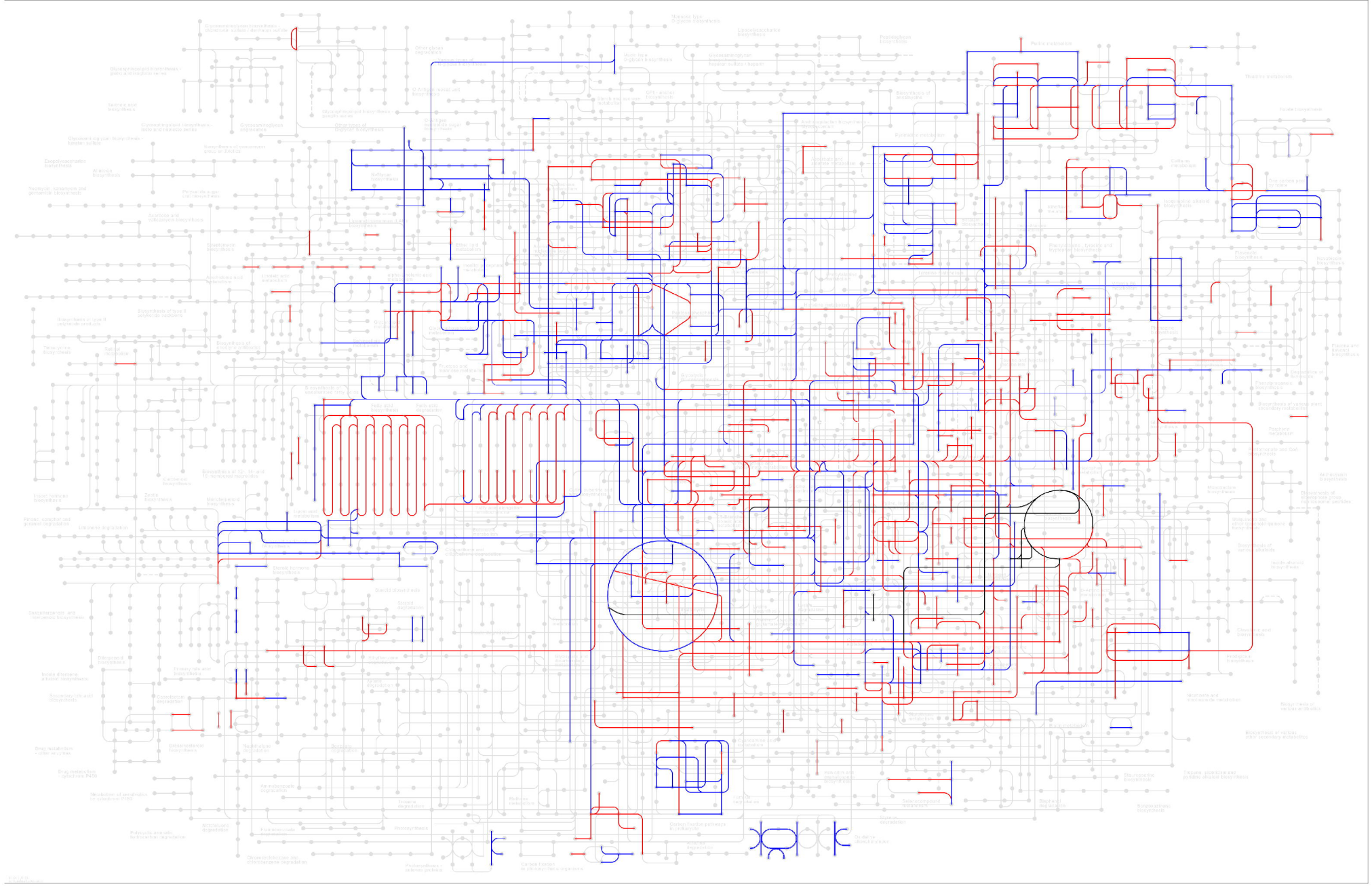
87. Wright L, Davidson S. How to tell the difference between a model and a digital twin. Advanced Modeling and Simulation in Engineering Sciences. 2020;7(1). doi: 10.1186/s40323-020-00147-4.
88. https://www.ebi.ac.uk/.
89. https://www.ncbi.nlm.nih.gov/taxonomy.
90. Mo ML, Palsson BØ, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology. 2009;3.
91. Sobolewska A, Dzbeński T. *In vitro* cultivation of *Pneumocystis* isolated from infected rat lungs. Wiad Parazytol. 2009;55(4):451-7. PubMed PMID: 20209825.
92. Larsen HH, Kovacs JA, Stock F, Vestereng VH, Lundgren B, Fischer SH, Gill VJ. Development of a rapid real-time PCR assay for quantitation of *Pneumocystis carinii f. sp. carinii*. Journal of Clinical Microbiology. 2002;40(8):2989-93. doi: doi:10.1128/jcm.40.8.2989-2993.2002.
93. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal Policy Optimization Algorithms2017 July 01, 2017:[arXiv:1707.06347 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv170706347S.
94. Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-baselines3: reliable reinforcement learning implementations. Journal of Machine Learning Research. 2021;22(268):1-8.

(A) METABOLIC MODEL DEVELOPMENT



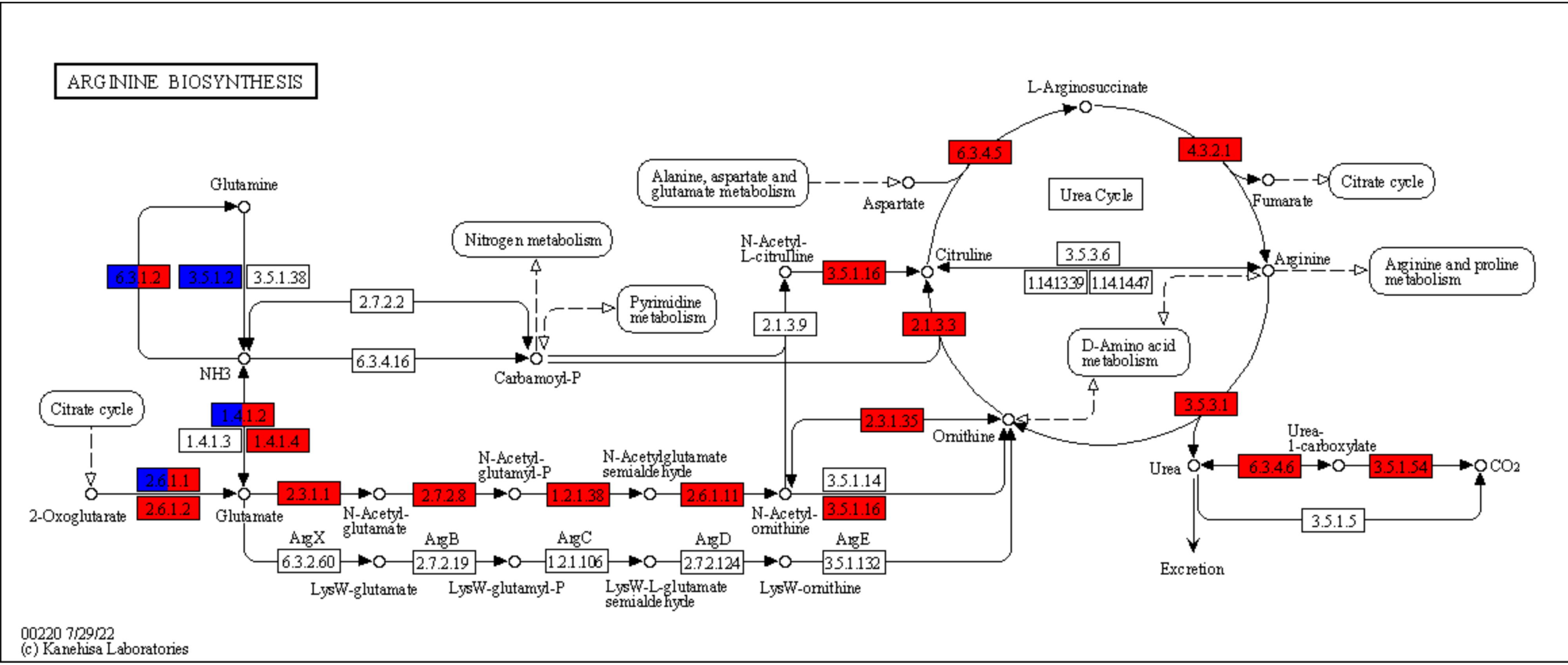
(B) DRAFT RECONSTRUCTION





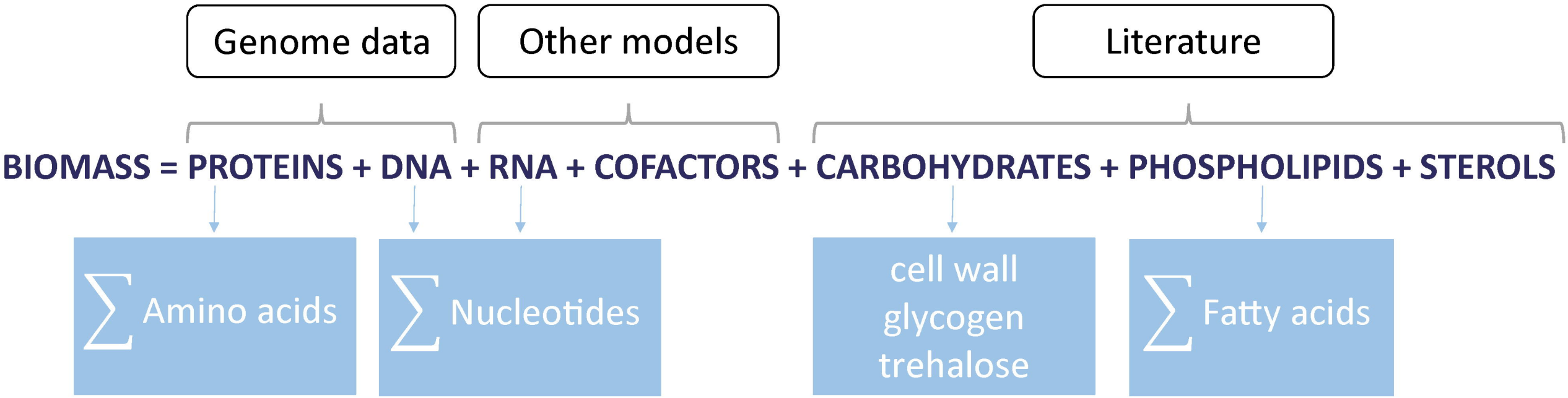
present in *Pneumocystis*

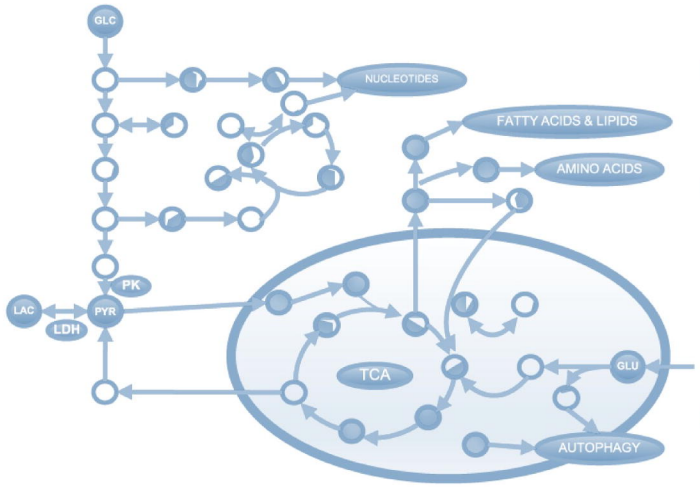
present in model yeasts (*S. pombe* & *S. cerevisiae*), but not in *Pneumocystis*



present in *Pneumocystis*

present in model yeasts (*S. pombe* & *S. cerevisiae*)


Genome data
Other models
Literature
BIOMASS = PROTEINS + DNA + RNA + COFACTORS + CARBOHYDRATES + PHOSPHOLIPIDS + STEROLS
∑ Amino acids
∑ Nucleotides
cell wall
glycogen
trehalose
∑ Fatty acids



(1) $X_1 \rightarrow X_2 + X_3$

(2) $X_2 + 2\,X_3 \rightarrow X_4$

…

(n)

$$\frac{dX}{dt} = S\,v$$

REACTIONS

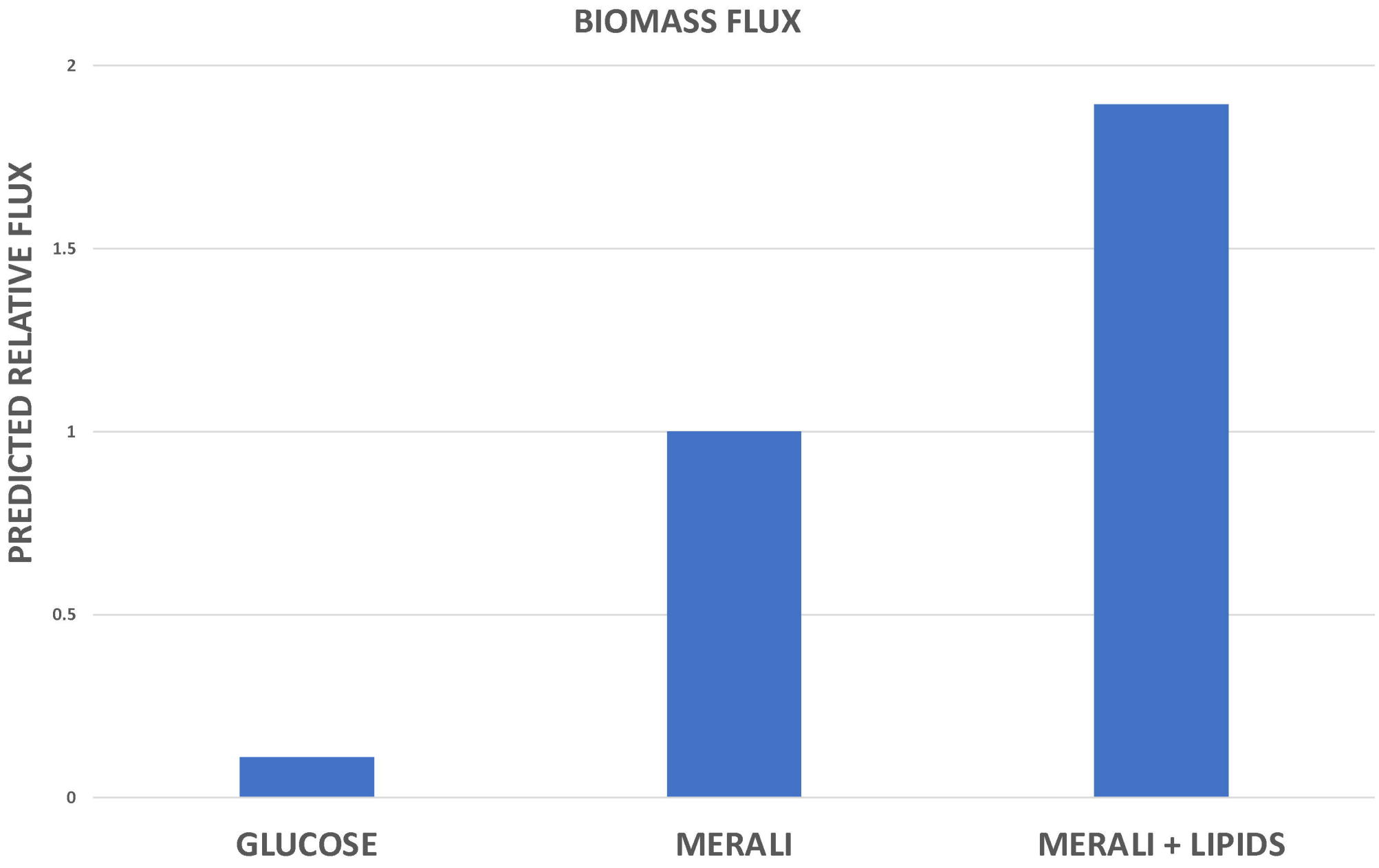
$$\begin{array}{c|c} & \begin{array}{cccc} (1) & (2) & \dots & (n) \end{array} \\ \text{METABOLITES} \begin{array}{c} X_1 \\ X_2 \\ X_3 \\ X_4 \\ \vdots \\ X_m \end{array} & \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ 1 & -2 & & \\ & 1 & & \\ & & & \\ & & & \end{pmatrix} \\ & \begin{array}{cccc} v_1 & v_2 & \dots & v_n \end{array} \end{array}$$

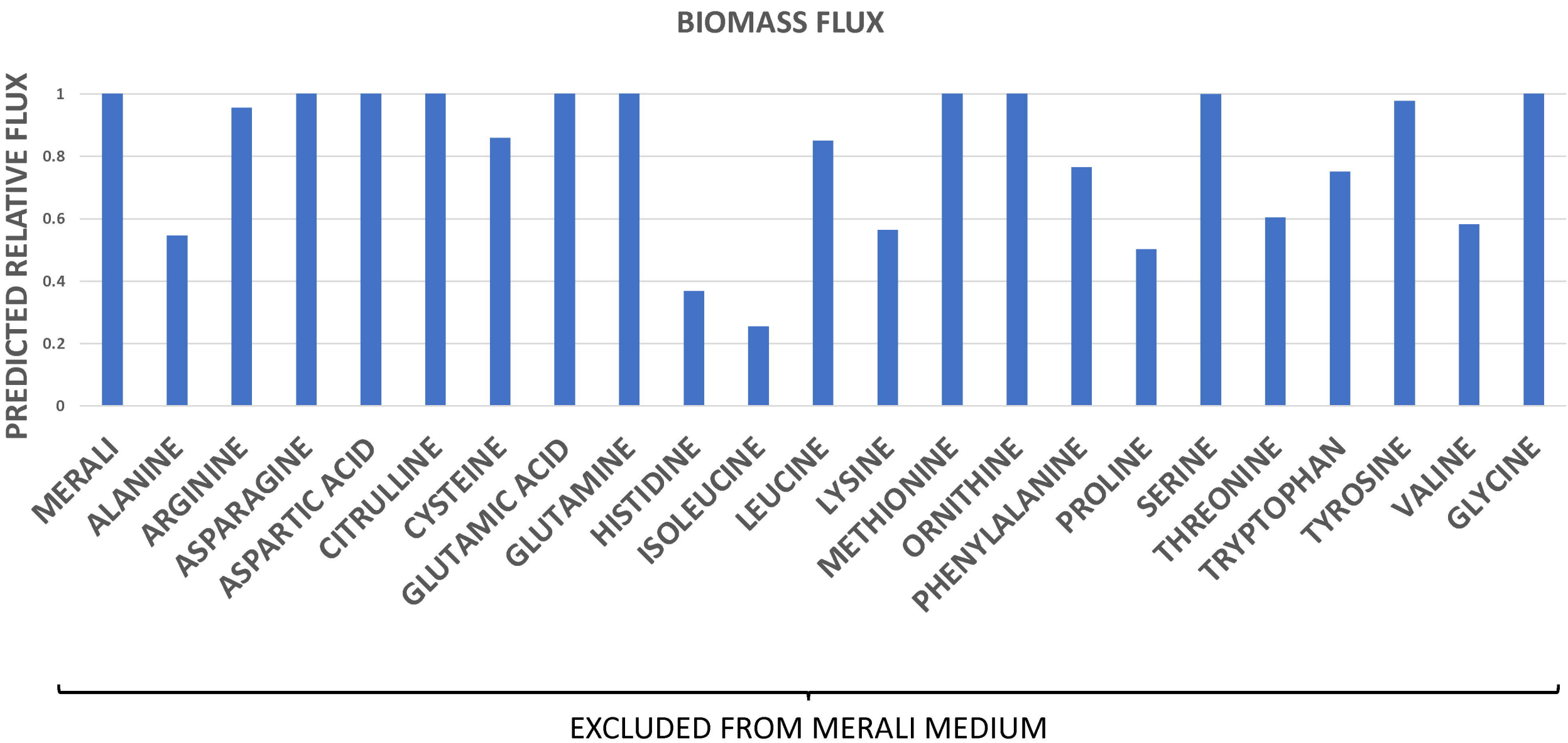
FLUXES

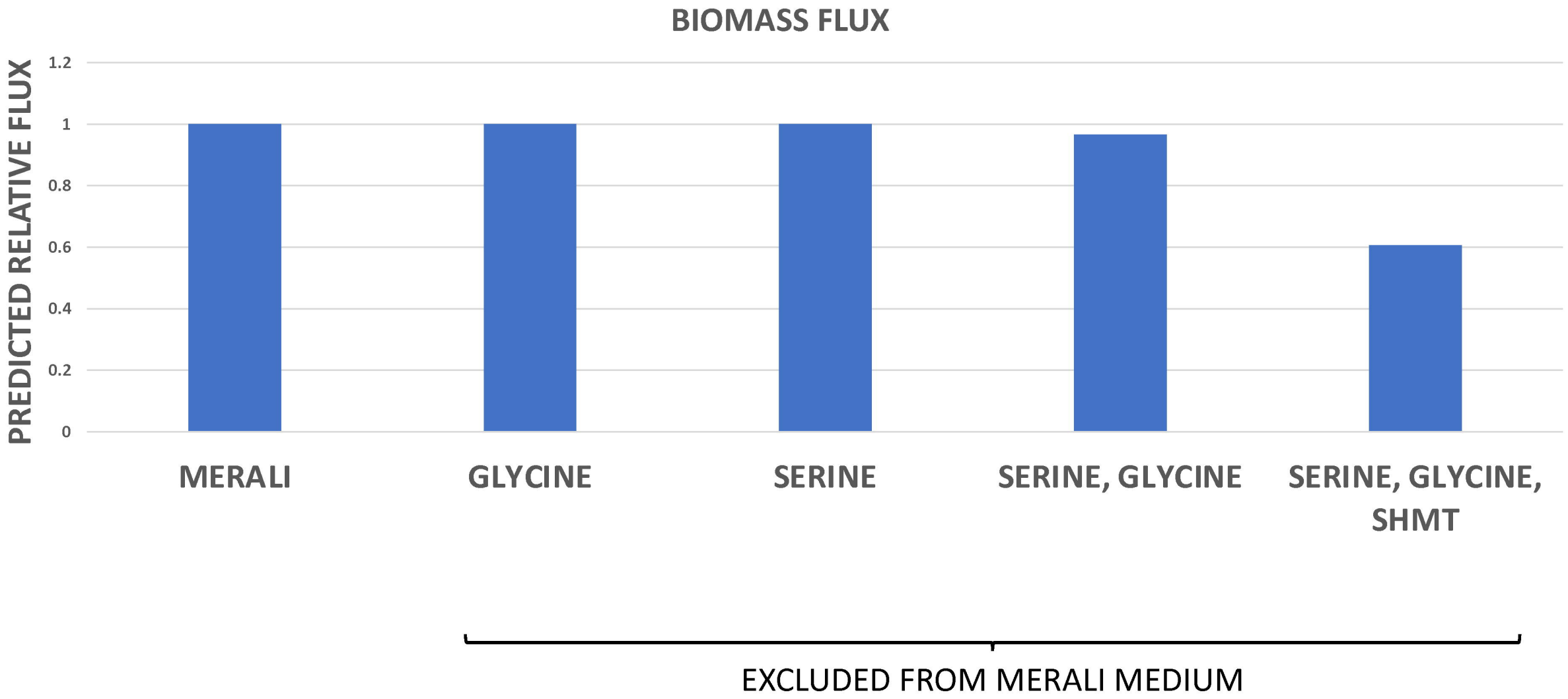
matrix S
(m x n)

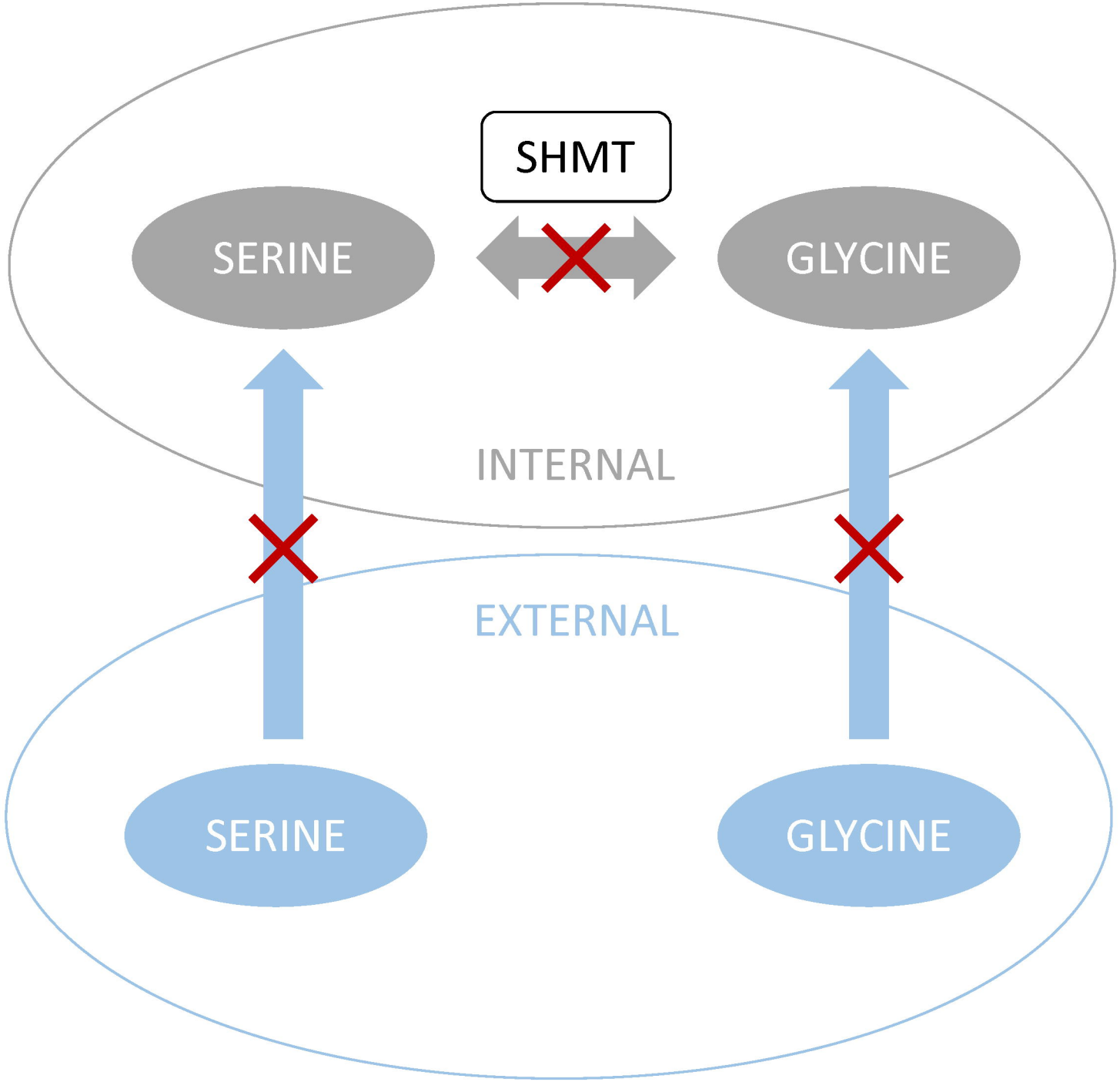
Objective function: $V_{BIOMASS} \rightarrow$ max

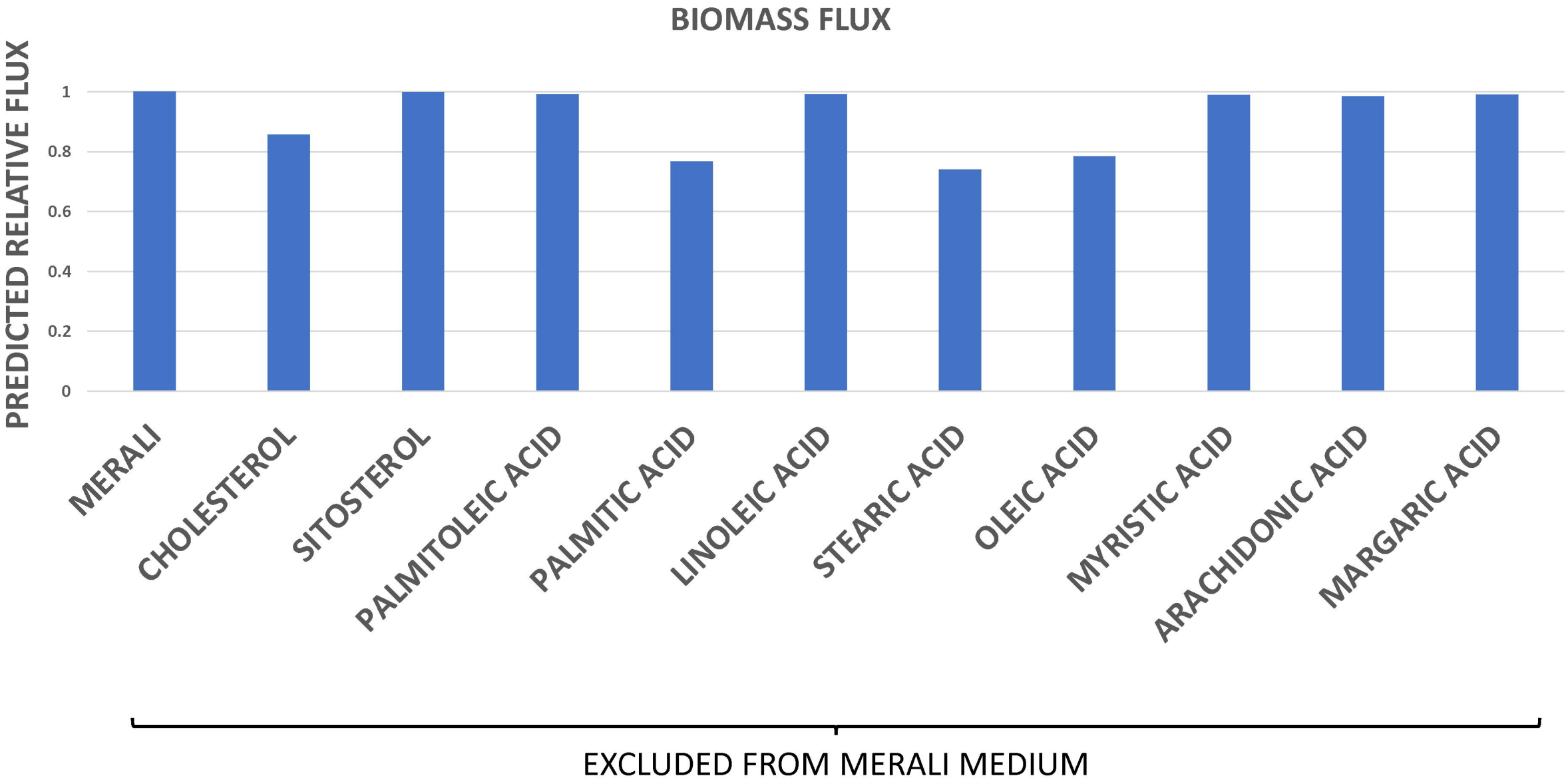
Steady-state assumption: $\frac{dX}{dt} = 0$

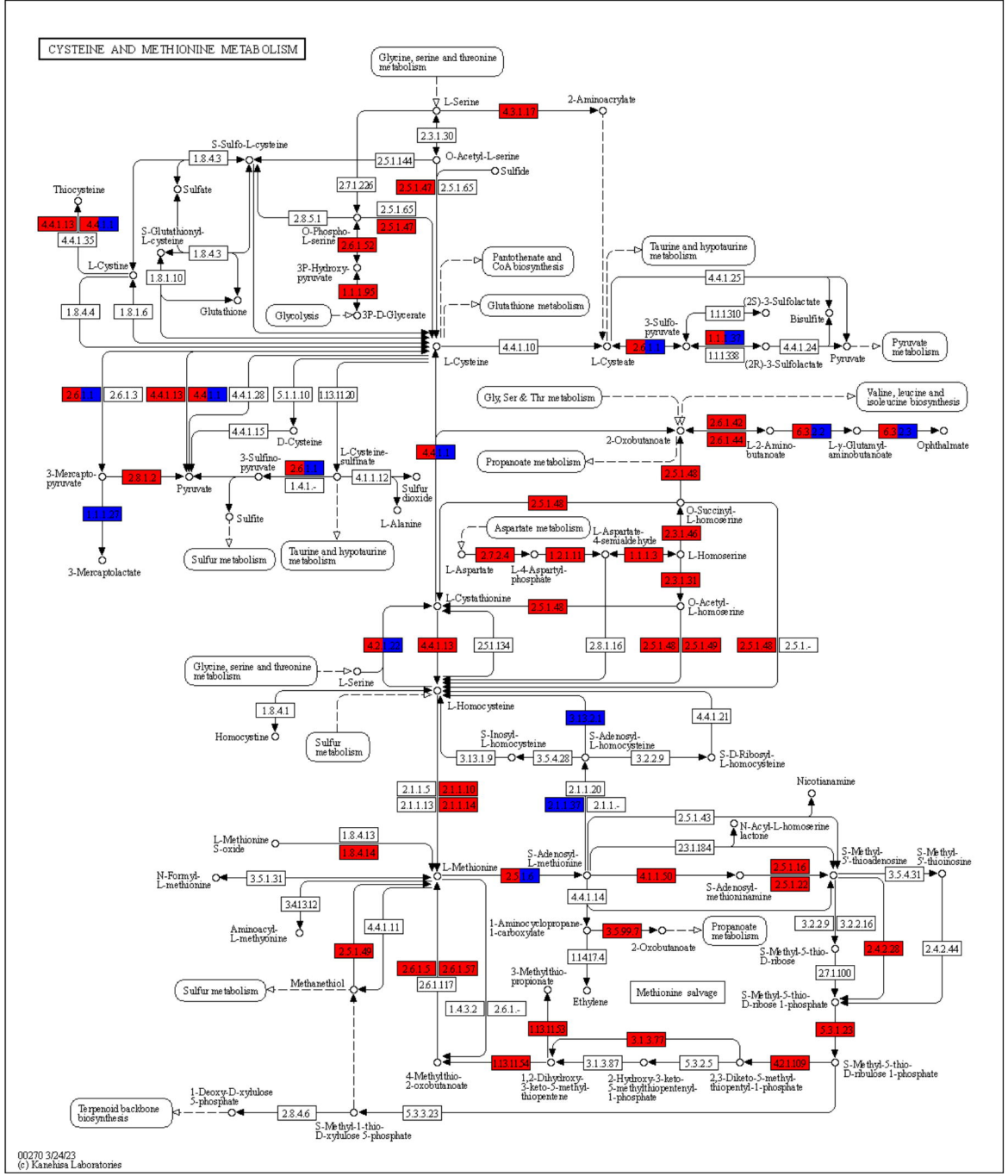

BIOMASS FLUX
PREDICTED RELATIVE FLUX
2
1.5
1
0.5
0
GLUCOSE
MERALI
MERALI + LIPIDS


BIOMASS FLUX
PREDICTED RELATIVE FLUX
1
0.8
0.6
0.4
0.2
0
MERALI
ALANINE
ARGININE
ASPARAGINE
ASPARTIC ACID
CITRULLINE
CYSTEINE
GLUTAMIC ACID
GLUTAMINE
HISTIDINE
ISOLEUCINE
LEUCINE
LYSINE
METHIONINE
ORNITHINE
PHENYLALANINE
PROLINE
SERINE
THREONINE
TRYPTOPHAN
TYROSINE
VALINE
GLYCINE
EXCLUDED FROM MERALI MEDIUM


BIOMASS FLUX
PREDICTED RELATIVE FLUX
1.2
1
0.8
0.6
0.4
0.2
0
MERALI
GLYCINE
SERINE
SERINE, GLYCINE
SERINE, GLYCINE, SHMT
EXCLUDED FROM MERALI MEDIUM

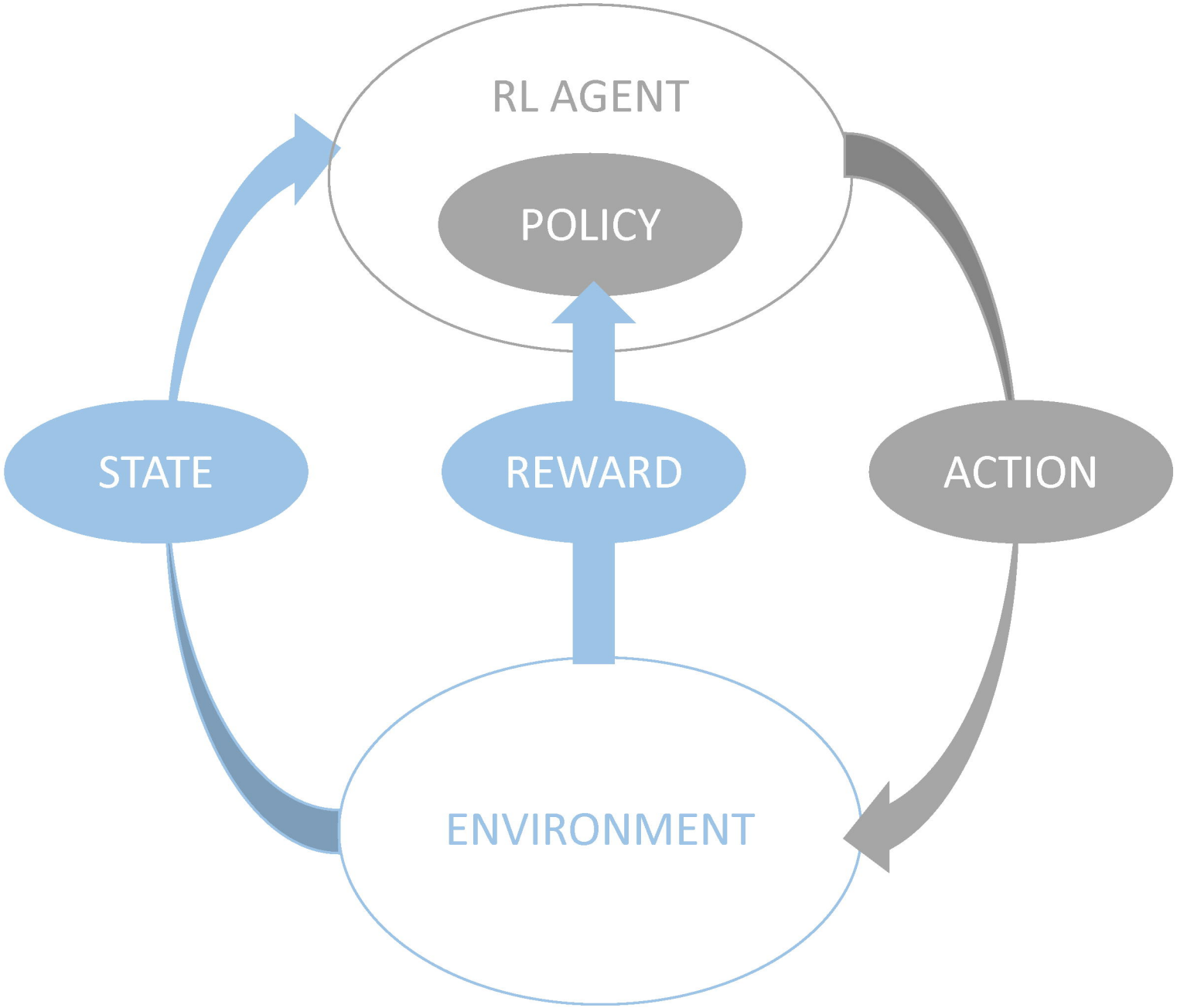

SHMT
SERINE
GLYCINE
INTERNAL
EXTERNAL
SERINE
GLYCINE

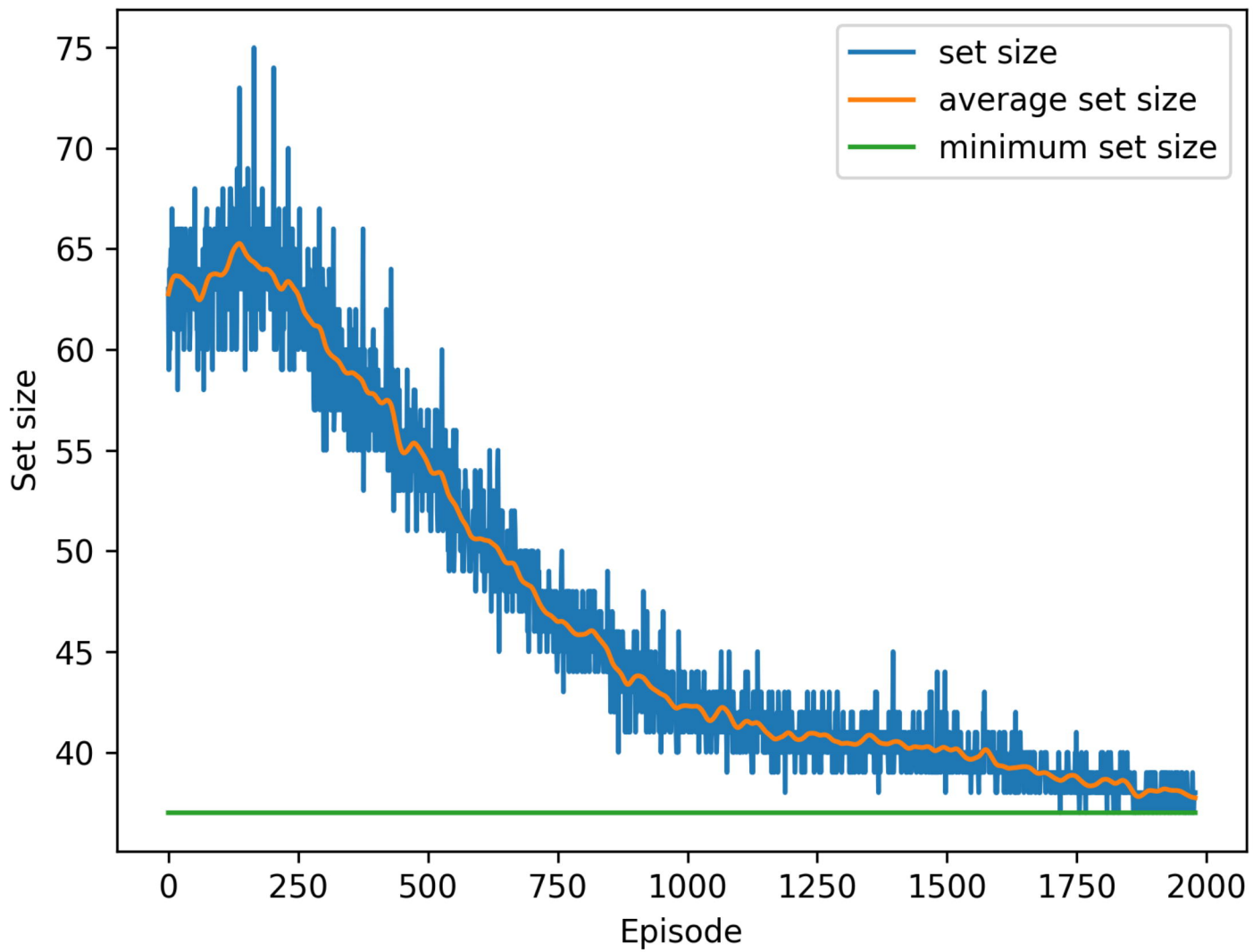

BIOMASS FLUX
PREDICTED RELATIVE FLUX
1
0.8
0.6
0.4
0.2
0
MERALI
CHOLESTEROL
SITOSTEROL
PALMITOLEIC ACID
PALMITIC ACID
LINOLEIC ACID
STEARIC ACID
OLEIC ACID
MYRISTIC ACID
ARACHIDONIC ACID
MARGARIC ACID
EXCLUDED FROM MERALI MEDIUM

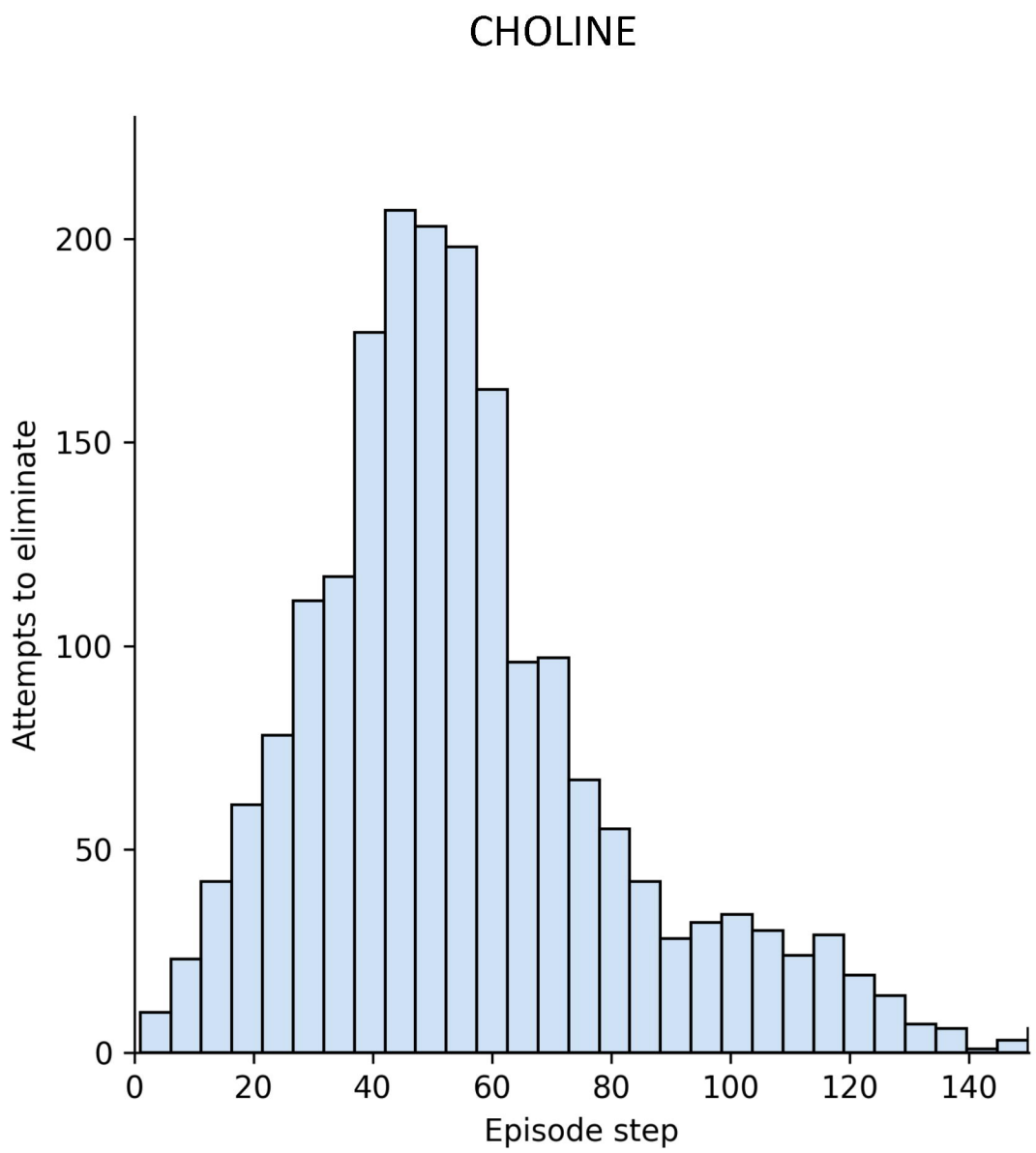


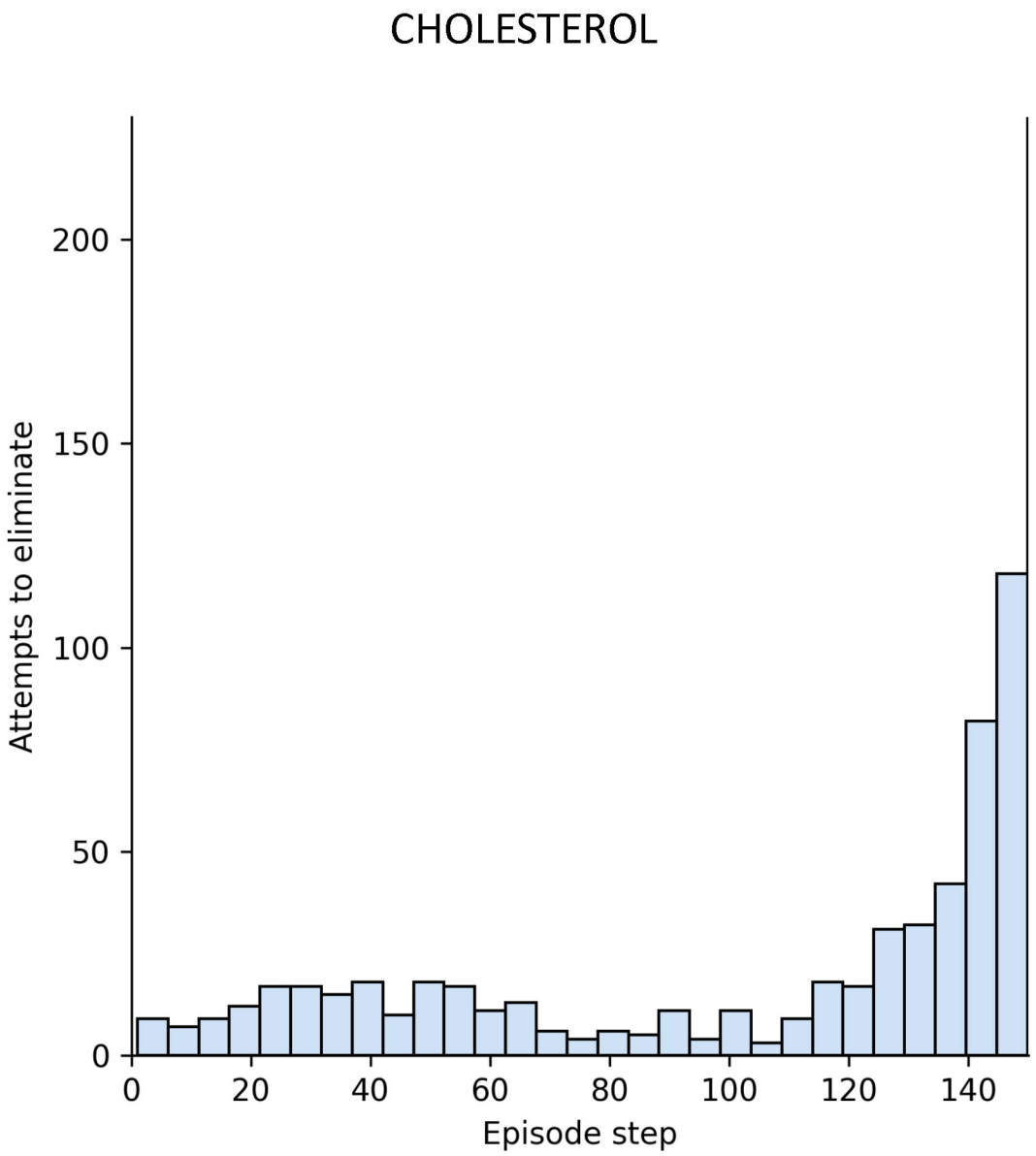
present in *Pneumocystis*

present in model yeasts (*S. pombe* & *S. cerevisiae*)


RL AGENT
POLICY
STATE
REWARD
ACTION
ENVIRONMENT


set size
average set size
minimum set size
Set size
75
70
65
60
55
50
45
40
0
250
500
750
1000
1250
1500
1750
2000
Episode


CHOLINE
Attempts to eliminate
Episode step
0
50
100
150
200
0
20
40
60
80
100
120
140


CHOLESTEROL
Attempts to eliminate
Episode step
0
50
100
150
200
0
20
40
60
80
100
120
140