MISSION-T2D

Multiscale Immune System SImulator for the Onset of Type 2 Diabetes integrating genetic, metabolic and nutritional data

**Work Package 3**

**Deliverable 3.6**

# Report on validation and refinement of the stress induced inflammation model in the overall workflow

## Document Information

| Grant Agreement | Nº | 600803 | Acronym | MISSION-T2D |
|---|---|---|---|---|
| Full Title | Multiscale Immune System SImulator for the Onset of Type 2 Diabetes integrating genetic, metabolic and nutritional data | | | |
| Project URL | http://www.mission-t2d.eu | | | |
| EU Project Officer | Name | Dr. Adina Ratoi | | |

| Deliverable | No | 3.6 | Title | Report on validation and refinement of the stress induced inflammation model in the overall workflow |
|---|---|---|---|---|
| Work package | No | 3 | Title | Stress induced inflammation onset modelling |

| Date of delivery | Contractual | 29.02.2016 | Actual | 14.04.2016 |
|---|---|---|---|---|
| Status | Version 1.3 | | Final | |

| Nature | Prototype | | Report | X | Dissemination | | Other | |
|---|---|---|---|---|---|---|---|---|

| Dissemination level | Consortium+EU | |
|---|---|---|
| | Public | X |

| Target Group (If Public) | | Society (in general) | |
|---|---|---|---|
| Specialized research communities | X | Health care enterprises | |
| Health care professionals | | Citizens and Public Authorities | |

| Responsible Author | Name | Pietro Lio, Gialuca Ascolani | Partner | Name |
|---|---|---|---|---|
| | Email | pietro.lio@cl.cam.ac.uk, ga304@cam.ac.uk | | |

**Version Log**

| Issue Date | Version | Author (Name) | Partner |
|---|---|---|---|
| 12.02.2016 | 1.0 | Pietro Liò | UniCAM |
| 27.02.2016 | 1.1 | Gianluca Ascolani | UniCAM |
| 10.03.2016 | 1.2 | Filippo Castiglione | CNR |
| 11.04.2016 | 1.3 | Pietro Liò | UniCAM |

| | |
|---|---|
| **Executive Summary** | In this deliverable we describe the work done in the development of multi omic metabolic models of gut microbiota bacteria to provide clues on the effects of the bacteria on gut environment and the interaction with the immune system inflammation in the gut. Moreover, we have used machine learning techniques to compare clinical and control data as normal autoimmune, diabetes and arthritis-related diseases and identified key biomarkers occurring during infectious diseases such as osteomyelitis. This is a new methodology for the analysis of signaling factors in T2D and autoimmune diseases.<br><br>The study is intended as a refinement of the previous inflammatory model and evidences the relationships between inflammation and gut microbiota. |
| **Keywords** | At least four keywords |

## Contents

# 1 Overview

Among the research activities in the context of Mission-T2D project we have performed research on: 1) development of multi omic metabolic models of gut microbiota bacteria. We have considered beneficial and pathogenic bacteria (pre and post infection metabolism). This provides clues on the effects of the bacteria on gut environment and the interaction with the immune system in the gut. In the following section, we will describe a multi omic metabolic model for pathogenic gut bacteria (clostridium difficile). Our methodology is novel because integrates different information (metabolic map and transcriptome); 2) Machine learning of patients versus normal autoimmune, diabetes and arthritis-related diseases data. We have identified key biomarkers occurring during infectious diseases such as osteomyelitis. We present here a new methodology for the analysis of signaling factors in T2D and autoimmune diseases. This methodology integrates epigenetic and transcriptomic data and is able to identify new disease associated genes. For both aspects publications have been submitted.

# 2 Multi omic metabolic model of gut microbiota bacteria

Clostridium difficile is a bacterium, which can infect various animal species, including humans. Incidence of infection with this bacterium is increasing in both frequency and severity. A better understanding of this organism and the relationship between its genotype and phenotype is essential to the search for an effective treatment. We reconstructed the metabolic model of this bacterium and analyzed it using sensitivity and robustness analysis. The standard metabolic model cannot account for changes in the bacterial metabolism in response to different environmental conditions. To account for this limitation, we integrated transcriptomic data, which details the gene expression of the bacterium, in a wide array of environments. To bridge the gap between gene expression levels and protein abundance, codon usage data was incorporated

in the model and the metabolic fluxes were assumed to be proportional to this protein abundance.

We validated this model by comparing model predictions with biological rationale, formed from analyzing literature on C. difficile. The model was able to predict various facets of the bacterium's metabolism. However, to use the model to gauge the efficacy of treatments in vivo, the metabolic network must be refined through inclusion of additional reactions. We believe this multi-omic model will be an important resource for better understanding this bacterium and devising effective treatments against infection.

## 2.1 Introduction to metabolic models

Clostridium difficile is a gram-positive, spore-forming, anaerobic bacterium, which infects or colonizes various animal species. Clinical manifestations in humans include asymptomatic colonization, mild diarrhea, pseudomembranous colitis, and death. Recently, C. difficile has ben recognized as one of the most dangerous bacteria growing in healthcare facilities worldwide [1]. The primary risk factor for development of C. difficile infection among hospitalized patients is antibiotic use, which disrupts the normal colonic microflora, and provides a niche for C. difficile to multiply and produce toxins [2]. Incidence of infection with C. difficile is increasing in both frequency and severity, and is contributing to considerable morbidity and mortality [3].

A model of the genotype-phenotype relationship of C. difficile's metabolism can be used to understand this bacterium and to identify potential drug targets. Methods to model the genotype-phenotype relationship range from stochastic kinetic models to statistical Bayesian networks. Kinetic models are limited, as extensive experimental data is required to determine the rate laws and kinetic parameters of biochemical reactions. An alternative to kinetic models is metabolic modeling, which has been used to depict a range of cell types from microbes to higher organisms without the need for difficult-to-measure kinetic parameters [4].

Metabolic models have been able to successively predict a range of cellular functions, such as cellular growth capabilities on various substrates and the effect of gene knockouts on the genome scale [5].

Metabolic models require a well-curated genome-scale metabolic network of the cell. Metabolic networks contain all the known metabolic reactions in an organism, along with the genes that encode each enzyme involved in a reaction. These networks are constructed based on genome annotations, biochemical characterizations, and various published literature on the target organism. The different scopes of such networks include metabolism, regulation, signaling, and other cellular processes [5].

Organisms such as Escherichia coli have been well studied and their metabolic networks continuously refined [6]. In contrast, there is a paucity of experimental data and literature on C. difficile strain 630 due, in part, to the organism's pathogenicity, which restricts the number of researchers actively studying it. As such, the first network for this bacterium was published in 2014 [7]. Despite the success of metabolic modeling in capturing large-scale biochemical networks, the approach is limited as it describes cellular phenotype simply in terms of biochemical reaction rates and is thereby disconnected from other biological processes that impact phenotype. Moreover, metabolic models cannot account for changes in the metabolism of the bacterium in response to different environmental conditions. To overcome this limitation, in this work we used the network published in [7] to create a multi-omic model for C. difficile. More specifically, we integrated a wide array of omics data into the metabolic model to improve its predictive power. Recent advances in the omics technologies, such as genomics (genes), transcriptomics (mRNA), and proteomics (proteins), have enabled quantitative monitoring of the abundance of biological molecules at various levels in a high-throughput manner. Transcriptomic data has been shown to be effective in improving model predictions of cellular behaviour in different environmental conditions [8].

To bridge the gap between gene expression data and protein abundance, we accounted for the codon usage bias of the bacterium. During translation of a mRNA to a protein, the information contained in the form of nucleotide triplets (codons) in the RNA is decoded to derive the amino acid sequence of the resulting protein. Most amino acids are coded by (2-6) synonymous codons. These codons, which code for the same amino acid, are surprisingly used differentially in protein-encoding sequences. The codon usage has been found to alter the translation time and the abundance of the resulting protein [9].

Here we present the first integrated model of the metabolism of C. difficile strain 630. We incorporated gene expression data to create context-specific models of metabolism. We also used the Codon Adaptation Index (CAI) to bridge the gap between gene expression levels and protein abundance in genome-scale models. We gauged the model with regards to its global robustness. We assessed the robustness of the model to changes in availability of metabolites and to varying fluxes through essential reactions. To validate the model, we compared model predictions to literature on C. difficile, highlighting strengths and room for improvement of the current version of the model. Finally, we identified metabolic pathways that could yield drug targets using network sensitivity analysis. The model and the associated code are compatible with the COBRA 2.0 toolbox [10].

## 2.2 Methods and Materials

### 2.2.1 Constraint-based Reconstruction and Modeling Approach

One constraint-based method for simulating the metabolic steady state of a cell is flux-balance analysis (FBA), which can be used to analyse the metabolic network solely on the basis of systemic mass-balance and reaction capacity constraints. FBA simulations have been able to capture microorganism growth, nutritional resource consumption, and waste-product secretion rates of various cell types [11].

The first step of FBA involves representing the metabolic network in the form of a numerical matrix S of size (m x n). This matrix contains the stoichiometric coefficients of each of the m metabolites in the n different reactions. In the matrix, each row represents one unique metabolite and each column represents one reaction. The stoichiometric matrix helps enforce a mass balance constraint on the system. The mass balance on the cell for i=1,...,m metabolites and j =1,...,n reactions constrains the metabolite concentrations xi, as shown in Equation 1, where vj is the flux through reaction j.

$$dx_i/dt=\Sigma_{j=1..n}\, S_{ij}v_j;\ i = 1,..\ m \qquad (1)$$

Under the steady state assumption, dxi/dt = 0; ∀i and the total amount of any compound being produced equals the total amount being consumed:

$$\Sigma_{j=1..n}\, S_{ij}v_j=0;\ i = 1,..\ m \qquad (2)$$

In most metabolic models, there are more reactions than there are compounds. Because there are more unknown variables than equations (n>m), any v that satisfies Equation 2 is considered to be in the null space of S.

FBA can be used to find and determine points within the solution space that are most representative of the biological system using linear programming methods. Studies have revealed that metabolic fluxes in microorganisms are best predicted by maximizing the cellular objectives of growth [11]. To determine the point corresponding to the maximum growth rate within the constrained space, the objective function shown in Equation 3 is maximized, where z is a linear combination of the fluxes. The f is a vector of weights and indicates how much each reaction flux contributes to the biomass objective function. The maximum growth rate can be achieved by determining the flux distribution v that results in maximal biomass flux.

$$z = f^T v \qquad (3)$$

Additional constraints can be added through the upper bound $v_j^U$ and the lower bound $v_j^L$ j for the flux $v_j$ . These bounds mandate the minimum and maximum fluxes allowed for a certain reaction and further decrease the space of allowable flux distributions for the relevant system. The mathematical representation of

the metabolic reactions, the objective function, and the capacity constraints define a linear system as shown in Equation 4.

$$\max f^T v$$
$$\text{subject to } Sv = 0$$
$$v_j^L \leq v_j \leq v_j^U,\ j = 1, \ldots n \quad (4)$$

The model fluxes are usually given units of mmol/gDWh, where gDW is the dry weight of cell mass in grams and h is the reaction time in hours. The bounds enforce thermodynamic constraints by dictating whether reactions are reversible or irreversible. The lower and upper flux bounds were arbitrarily chosen to be -1000 mmol/gDWh and 1000 mmol/gDWh for reversible reactions. For irreversible reactions, $v_j^L$ was chosen to be 0 mmol/gDWh and $v_j^U$ was set to 1000 mmol/gDWh. The biomass production equation was extrapolated from Bacillus subtilis, the closest organism for which biomass constituent data was available [12]. The biomass reaction in the network is scaled such that the flux is equal to the exponential growth rate of the organism, which has units of $h^{-1}$.

### 2.2.2 Multi-objective Optimization in Metabolic Models

One limitation of using only biomass as the objective is that such an approach ignores specific protein costs and benefits, and their effects on the remaining capacity of ribosomes and cytosolic space. The existence of such trade-offs between energy production and protein costs is not accounted for when using only biomass as the objective [37]. Furthermore, the biomass objective vector is usually perpendicular to one of the surfaces of the solution space of the FBA problem. For this reason, biomass maximizing flux states are usually degenerate; there exist multiple flux distributions that yield the same maximal biomass value [38]. To choose between the various flux distributions, additional criteria must be considered. For these reasons, we modelled metabolism as a multiobjective phenomenon. We used a multi-objective optimization approach to address the k conflicting objectives, as shown in Equation 5.

$$\max y = f(x) = (f1(x), f2(x), \ldots, fk(x))$$

subject to $Sv = 0$

$$v_j^L \leq v_j \leq v_j^U,\ j = 1, \ldots n \quad (5)$$

Because the optimality goals in metabolism are often different and are simultaneously competing, the scalar notion of "optimality" does not hold; cells are thought to face a trade-off that is described by the set of Pareto-optimal solutions. In a maximization multi-objective problem, a vector that is part of the feasible space is considered to be Pareto-optimal if all other vectors have the same or a lower value for at least one of the objective functions. Therefore, a Pareto-optimal solution is found when there exists no other feasible solution, which would decrease one objective without increasing another objective. The set of Pareto-optimal solutions constitutes the Pareto-optimal front [13]. In the absence of additional information, no one Pareto-optimal solution can be said to be better than the other; higher-level information is required to choose one of the solutions [14].

As proposed by Costanza et al. [15], to solve this multi-objective optimization problem one can use evolutionary algorithms Istochastic optimization methods that simulate the process of natural evolution.

Evolutionary algorithms are well suited to multi-objective problems because they can generate multiple Pareto-optimal solutions after one run and can use recombination to make use of the similarities of solutions [14].

The input to the evolutionary algorithm is a set of arrays, also called individuals, representing potential solutions to the problem. These arrays are then ranked based on the values of their objective functions.

Potential optimal solutions are generated by retaining the best individuals and by generating new individuals through the use of variation. This process is continued until no further improvements are detected on the Pareto front.

By modelling the metabolism of bacterium as a multi-objective problem, we address a conflict problem whereby maximizing one objective (e.g., biomass) might involve a trade-off in the other objective (e.g., intracellular flux). We used a modified version of the evolutionary algorithm developed by Angione et al. in

[16]. The population size and the number of populations used with this algorithm were 70 and 210, respectively.

We used iMLTC806cdf, an extensively curated reconstructed metabolic network of C. difficile strain 630. The metabolic network consists of 806 genes, 705 metabolites, and 1092 reactions [7]. We investigated the effect of the choice of different objectives on model predictions. The first objective was always maximization of biomass production, because it was found to best predict metabolic fluxes. In this work, we used the IBM ILOG CPLEX CP Optimizer solvers [21].

## 2.3 Sensitivity Analysis

Sensitivity Analysis is used to identify model inputs that have a large influence on the model outputs. To find the metabolic pathways (sets of reactions producing a particular product) that had the largest effect on the model outputs, we used PoSA [15]. The pathways, such as the Citric Acid Cycle and the Pentose Phosphate Pathway, perform certain metabolic tasks for the bacterium. This approach involves genetically manipulating the metabolic model to find the sensitive pathways, which make a large impact on model outputs. In other word, we perturbed pathways by mutating the genes that govern their biochemical reactions and analysed the result on the outputs.

In the knockout vector $y = \{b_1, b_2, ,b_s, b_p\}$, $b_s$ represents the perturbations on the genes governing the metabolic pathway s, where $|b_s| = W_s$ (number of genes partaking in the $s_{th}$ pathway). Because the gene knockouts are represented through the use of binary variables, we perform combinatorial perturbations, namely the bits in bs are switched from 0 to 1 or from 1 to 0; if a gene in $b_s$ is set to 1, this gene is knocked-out in the model. According to [15], the Pathway Elementary Effect (PEE) for the genetic perturbation $b_s$ can be defined as follows:

$$PEEs = F(b_1, b_2, \ldots, b_s^+, \ldots, b_p) - F(y^+)/\Delta s \quad (6)$$

where $b_s^+$ represents the genetic manipulation of the input $b_s$; $y^+$ is the mutation carried out on the knockout vector y; F(y) is the vector v of fluxes as produced by the model; finally, Δs is a scale and is defined as:

$$\Delta s = W^{-1}_{s} \cdot \Sigma_{j=1..Ws}\, b^{+}s(i);\ s = 1,..\ p \qquad (1)$$

Next, the sensitivity indices μ and σ are determined by calculating the mean and the standard deviation of the distribution of the PEE for each input. Pathways with a large σ have a large influence on the output. A large σ indicates an input whose influence highly depends on the value of other inputs. By perturbing the genes through the use of knockouts and comparing the outputs of the model with and without the genetic manipulations, we detect the most sensitive pathways of the metabolic network.

## 2.4 Robustness Analysis

A facet of living organisms is their homeostasis, otherwise known as their ability to remain robust to external and internal perturbations within a certain range. External perturbations include changes in temperature or food supply while internal perturbations include spontaneous mutations. The robustness of biological systems is partly due to the presence of parallel metabolic pathways. Robustness represents the insensitivity of a system to changes in system parameters.

Global Robustness (GR) analysis can be used to survey the parameter space to determine the region where the cell exhibits specific features. More specifically, we perturbed the flux bounds of the metabolic model and observed the resulting effects on the objectives: biomass and intracellular flux. We followed the analysis described in [19]. The Global Robustness was defined as the percentage of trials determined to be robust.

We were also interested in the robustness of the network to a particular reaction. Therefore, robustness was defined as a measure of the change in the maximal flux of the objective function when the optimal flux through any of the reactions was altered [29]. The robustness of the biomass objective function

to availability of glucose and fumarate was determined. We also determined the robustness of the model to changes in the flux through essential reactions, such as the reaction catalyzed by Glutamate Dehydrogenase (GDH) and Ribose-5-phosphate isomerase (Rpi).

## 2.5 Modeling transcriptomics in genome-scale models

To increase the reliability of the model, gene expression was added to the FBA framework, as shown in 1. The flux bounds of the model were continuous functions of the gene expression data available for a specific environmental condition. Through this methodology, a model is tailored for each environmental condition. Each of the reactions in the metabolic model depend on a gene set, which is represented through the use of AND/OR operators. In this formulation, if a gene set is composed of two genes and an AND operator, both genes are required to carry out the corresponding reaction. On the other hand, if two genes connected by OR, one gene is sufficient in carrying out the reaction.

This formulation can be transformed to derive the gene set expression from the expression of individual genes. When two genes are connected through an AND operator, the gene set expression for reaction i, gi, is the minimum of the expression of the individual genes making up the gene set. The gene set expression for two genes connected by an OR operator is the sum of the expression of the individual genes. In an alternative methodology, proposed by [22], the gene set expression for two genes connected by an OR is the maximum of the expression of the individual genes. Following METRADE [22], gene set expression was then used to alter the bounds on $v_j$.

The logarithmic map between gene set expression and the flux bounds was chosen because at high mRNA abundance, an increase in mRNA abundance was found to produce a relatively small increase in the protein synthesis rate. On the other hand, at low mRNA abundance, an increase in mRNA abundance was found to produce a large increase in the protein synthesis rate [23].

## 2.6 Integrating Codon Usage in the FBA Framework

The translation rate of a codon is determined in part by the speed of diffusion of a translationally competent tRNA to the ribosome. Because tRNAs are differentially abundant in the cell, codons pairing to high-abundance tRNAs are translated faster than those pairing to low-abundance tRNAs. Although synonymous codons produce the same amino acid sequence, they can alter the translation speed and the protein expression levels depending on the abundance of their associated tRNA [24]. A study by Gygi et al. [25] revealed that a large codon bias generally resulted in higher protein expression levels.

Therefore, the inclusion of codon bias can help improve the metabolic model predictions by helping link gene expression levels to protein levels.

The codon usage table for C. difficile was obtained from the Kazusa Codon Usage Database, which lists the frequency of different codons in the C. difficile genome [26]. The weights for synonymous codons was determined as the ratio between the observed frequency of the codon k and the frequency of the most preferred synonymous codon for that amino acid. We obtained the mRNA sequence associated with the 806 genes of C. difficile from UniProt [27]. The counts of different codons were determined for each mRNA sequence. To obtain a measure of the codon bias, we calculated the relative Codon Adaptation Index (rCAI) for each gene. The rCAI represents the relative adaptiveness of the codon usage of the relevant gene to the codon usage of highly expressed genes [28]. The rCAI was calculated according to the following:

$$rCAI = e^{\left[\frac{1}{L}\sum_{l=1}^{L} ln(w_{k(l)})\right]}$$

where L is the number of codons in the genes and wk(l) is the weight associated with codon type k for lth codon along the length L of the gene. The value hi obtained after integrating gene expression (Eq. 9) was then scaled by rCAI to account for both gene expression and codon usage bias (equaz 12):

$$g_t^{'} = g_t \cdot (rCAI_t)$$

$$v_j^L h(g_j) \leq v_j \leq v_j^U h(g_j),$$

where

$$h(g_j) = \begin{cases} (1 + |log(g_j)|)^{\frac{g_j - 1}{|g_j - 1|}} & \text{if } g_j \in \mathbb{R}^+ \setminus \{1\} \\ 1 \text{ if } g_i = 1 \end{cases}$$

In each reaction of the model, the associated hj was finally adopted as a multiplicative factor for the flux bounds. The logarithmic map between gene set expression and the flux bounds was chosen because at high mRNA abundance, an increase in mRNA abundance was found to produce a relatively small increase in the protein synthesis rate. On the other hand, at low mRNA abundance, an increase in mRNA abundance was found to produce a large increase in the protein synthesis rate [21].

## 2.7 Analysis of cDNA Microarrays

We used microarray analysis to determine the combination of genes which were up-regulated or down-regulated in different environmental conditions.

We used Limma, a package in Bioconductor 3.1 software, because it allows for rigorous statistical analysis of gene expression [30]. We preprocessed the data through background correction, within-array normalization, and between-array normalization. After normalization, we used filtering to remove probes that did not appear to be expressed in any of the experimental conditions. Next, we used linear models to analyze the microarray data. To conduct statistical analysis and assess differential expression, we used an empirical Bayes method to modulate the standard errors of the log-fold changes. To test for the comparisons of interest, we used an analysis of variance (ANOVA) model. A

false discovery rate is the percentage of expressed genes that are actually non differentially expressed among those that are differentially expressed. We considered probes with false discovery rate less than 0.05 to
be differentially expressed.

### 2.8 Results and Discussion

To construct our metabolic model, we investigated different multi-objective optimization scenarios to find the objectives that best constrained the feasible space. The metabolic model from the multi-objective optimization was then assessed using robustness and network sensitivity analyses. Comparison of model predictions to literature during these analyses was used to evaluate the model. To determine the behaviour of the bacterium in different environments, gene expression and codon usage data were integrated to create a multi-omic model for the bacterium. Due to the lack of experimental data on biomass composition and flux rates for C. difficile, the model can only be used to make predictions about qualitative changes in the organism's behavior in response to different environmental conditions. Furthermore, the lack of experiments that focus on the metabolism and growth rate of C. difficile makes it difficult for the model to yield reliable absolute growth rate predictions. For this reason, applications of this model are limited to areas such as studying the bacterium and designing novel drug targets rather than using the model to optimize production of a certain amount of metabolite.

## 3 Secondary Objective Analysis

In the first experiment, we built a model maximizing biomass while minimizing total intracellular flux. The Pareto front, shown in red in Figure 2a, signifies a trade-off between the two objectives. The Pareto front in both cases is obtained through METRADE [22] by manipulating the gene expression of the bacterium to achieve a range of trade-offs between the two objectives. Biological systems

that perform multiple tasks face an optimality trade-off. A given phenotype cannot be optimal for all environmental conditions and the Pareto front indicates the range of possible phenotypes resulting from these trade-offs [39].

In the second experiment, we built a model which maximizes biomass and succinate production. C. difficile uses the succinate spike in the microbiota, which follows antibiotic-treatments, to expand in the intestine. Organic acids, such as succinate, are common end products of fermentation by the microbiota and are metabolized by C. difficile during its colonization. C. difficile couples reduction of succinate to butyrate with fermentation of dietary carbohydrates [20]. For this reason, an increase in succinate availability would be expected to result in an increase in growth (biomass production). The results of the second experiment are shown in Figure 2. In this case, no clear Pareto front was observed; the red points don't constitute a Pareto front as they are not strictly dominated by any other point. The dominated solution was obtained due to the flux bound for succinate production, 10 mmol/gDWh. The first model, maximizing biomass and minimizing total intracellular flux, was used in subsequent analyses due to the presence of a trade-off, which was thought to better constrain the feasible space. Due to the absence of experimentally-derived metabolic fluxes, validation of the model involved comparing model predictions to biological rationale, findings on various facets of the C. difficile's metabolism.

## 3.1 Robustness Analysis

In our analysis, we determined robustness of the network and of the model with regards to metabolite activity and essential reactions.

### 3.1.1 Global Robustness

To gauge the robustness of the network, we determined the change in the maximal biomass flux in response to external perturbations. Global Robustness (GR) analysis revealed that the biomass production was fully robust to

perturbations in the flux bounds. This facet of the bacterium's metabolism was biologically relevant as bacteria, such as C. difficile are able to grow despite small fluctuations in their physical environment. The intracellular flux was 65.8% robust for a tolerance (ε) and flux bound perturbation (σ) of 1%.

The GR for intracellular flux is shown in Figure 3 for different values of  and . The GR falls when σ is increased or ε is decreased. The model appears to more sensitive to the than to the level of perturbations incurred. For an ε of 1%, the GR falls sharply as σ increases from 5% of flux bounds.

Metabolite Availability and Essential Reactions For C. difficile, biomass production was expected to increase before leveling off for increased availability of carbon substrates, such as glucose and fructose. Robustness analysis showed an increase in biomass production through increased availability of glucose (Figure 4). Further robustness analysis revealed that the model did not agree with experimental findings for some reactions in the bacterium's metabolisms. For example, one essential reaction for C. difficile is the conversion of -ketoglutarate to glutamate by GDH, which is the sole means of incorporating inorganic nitrogen into carbon backbones for the bacteria [32].

Glutamate Dehydrogenase (GLDH) was experimentally found to be important for growth of C. difficile [34]. However, increasing the metabolic flux through the corresponding reaction actually decreased biomass production, as shown in Figure 4. Furthermore, using the model in a gene deletion case study revealed that knockout of the gene gluD responsible for producing GDH, did not affect model predictions of maximal biomass production. These results contradict a study by Girinathan et al. which found that the gluD mutants grew slower than the parent strain and that complementation of the gluD mutant with the fuctional gluD gene reveresed the growth defect [34].

A different trend was observed for the robustness analysis of the reaction catalyzed by D-proline racemase (PrdF). The Stickland pathway is thought to serve as a primary source of energy in C. difficile. One reaction in this pathway is the racemization of L-proline to D-proline, which is the substrate of D-proline

reductase in the Stickland pathway. This racemization is conducted by PrdF. Experimental findings show that PrdF is partially involved in mediating optimal growth of C.difficile Wu2014. These experimental findings vary starkly with model findings that show a decrease in growth with increasing flux in proline racemase as shown in Figure 4.

## 3.2 Sensitivity Analysis

Sensitivity analysis in genome-scale models has been proposed to assess the key component in the metabolic network. Using PoSA [15], we compute sensitivity indices in a pathway-based fashion. Each pathway is assessed through random perturbations of its reactions, and the average perturbation and the standard deviation are computed as a result.

We performed the pathway-based sensitivity analysis and identified eleven sensitive pathways in the metabolic model as shown in Figure 5. The pathway with the largest σ is the valine, leucine, and isoleucine metabolism pathway. As shown in Table 1, these three amino acids are essential to the bacterium and their metabolism pathway was also expected to be essential. On the other hand, pathways, such as the alanine, aspartate, and glutamate metabolism pathway, were not determined to be sensitive.

Model findings match our expectations that the growth of the bacterium will not be as sensitive to metabolism of the nonessential amino acids, such as alanine, aspartate, and glutamate. The model found the ubiquinone and other terpenoid-quinone biosynthesis to be a sensitive pathway for C. difficile. Terpenoid biosynthesis has been found to be essential for survival of bacteria. Currently, drugs against various bacterial pathogens target enzymes in this pathway [35]. Another pathway determined by literature to be important to C. difficile is the Stickland pathway. Stickland reactions couple oxidation and reduction of amino acid pairs and have been found to be important for the growth of C. difficile as they contribute to generation of ATP and reducing power [36].

The sensitivity and robustness analysis of the model revealed that overall, despite some limitations, the network could be used to make meaningful predictions about the behavior of the bacterium. Therefore, we decided to investigate whether the model could reliably predict C. difficile behavior in different environments.

### 3.3 Essential Amino Acids

We compared model predictions to existing experimental findings on the response of C. difficile to the absence of various amino acids. The results are shown in Table 1. The bacterial essential amino acids were found to be essential by the model as their removal prevented the production of biomass.

The removal of amino acids that were not found to be essential or to affect growth, did not visibly affect model-predicted biomass values. However, the removal of growth-enhancing amino acids did not affect growth predictions. Therefore, while the model is consistent with experimental data regarding the amino acids that are essential or non-essential to C. difficile for survival, it does not capture the effect of growth-enhancing amino acids.

### 3.4 C. difficile Growth Under Different Conditions

We collected transcriptomic profiles from [3], E-GEOD-22423, Janoir2013 E-GEOD-37442, Fimlaid2013Ferreyra2014a Context-specific models for C. difficile were generated by incorporating gene expression data obtained for the bacterium in different environmental conditions. To improve the reliability of the model, we also integrated codon usage data (See Materials and Methods for details). The biomass production after accounting for the bacterium's codon usage varied slightly for some environmental conditions.

Model predictions of these context-specific models were compared to expectations of the organism's behavior in these environmental conditions to further validate the model. C. difficile is an obligate anaerobe so it is sensitive to presence of oxygen in the atmosphere. This experimental finding is also

achieved by the model, which predicted a drastic decline in growth rate in aerobic conditions. However, the bacteria is able to survive in atmospherically exposed (aerobic) cultures, suggesting that the bacteria has some mechanisms for tolerating limited oxidizing conditions [41]. The effect of acid shock on C. difficile is particularly relevant in a clinical setting as proton pump inhibitors (PPI) have been identified as a risk factor for C. difficile associated disease because they elevate pH to greater than 5. Gastric acid (pH 1.5-3.5) has been shown to kill vegetative C. difficile but the bacterium could contribute to disease pathogenesis if it is able to survive in gastric contents after use of PPIs Jump2007 The model predicts that the bacterium can tolerate these elevated pHs, with growth rate at a low around pH 8.5. Model findings agree with experimental findings that show C. difficile can withstand pH shocks between 4.8 and 8.3 [3]. A study further determined that bacterial growth was inhibited when the pH was decreased to 4.8 [42]. The model predicts a low biomass production when the pH is decreased to 4.5. This findings are also replicated in clinical settings as (http://sma.org/southernmedical-journal/article/does-alkaline-colonic-ph-predispose-to-clostridium-difficile-infection/) found a strong association between C. difficile infection and alkaline stool pH, suggesting that alkaline colonic pH predisposes patients to CDI.

We incorporated gene expression data of C. difficile, which was inoculated into Brain-Heart Infusion (BHI) medium containing sublethal concentrations of antibiotics, into our model. Our model predicted a decrease in bacterial growth (1.26-1.27 $h^{-1}$), relative to the BHI medium (1.29 $h^{-}1$). This prediction is supported by [44], which determined that sub-lethal concentrations of these antibiotics will delay C. difficile's growth.

To further illustrate the applicability of these multi-omic models in clinical settings, we decided to conduct a case study. Using the transcriptional response elicited by C. difficile in presence and absence of B. thetaiotaomicron in vivo for polysaccharide-deficient dietary conditions. Bt produces high levels of succinate during fermentation of dietary carbohydrates and C. difficile couples

reduction of this succinate to butyrate with fermentation of dietary carbohydrates. The results of Ng et al. support this hypothesis as shown in Figure 6. After constructing a multi-omic model for the bacterium C. difficile in the presence and absence of Bt, we found that the growth rate was higher in the presence of Bt, agreeing with the experimental findings of Ng2013.

## 3.5 Conclusions and Future Work

In this study, we created the first multi-omic model of metabolism for the organism C. difficile. The global robustness of the model was found to be 100% and 65.8% in predicting biomass production and the intracellular flux, respectively. A sensitivity analysis identified eleven sensitive pathways for the metabolic model, which largely matched expectations from literature. During validation of the model, the model was found to successfully predict essential amino acids.

To predict the bacterium's behavior in different environmental conditions, the model was integrated with transcriptomic and codon usage data to generate reliable and context-specific metabolic flux distributions. We assessed the predictive potential of the model by comparing model prediction with published experimental data, therefore highlighting strong and weak points of the current knowledge of C. difficile metabolism. The comparison of model predictions to literature on C. difficile revealed that the model is very useful but neglects some intricate details of the bacterial metabolism. In these specific cases, to make the model more representative of the bacterium, the metabolic network must be refined by adding more reactions.

Gene expression data in combination with our multi-omic model could provide translational medicine information and predictions of bacterial lifestyle. The multi-omic model provides effective comparison between different medical-related conditions (for example pre-infection, post infection) and could be further refined by adding other intermediate conditions. Our multi-omic model can be used by biomedical researchers to study the bacterium and devise

targeted treatments. It can also be used to simulate the interactions between the human gut microbiota and the host. By accounting for the gut microbiota-host interactions, we can construct a whole gut model response to infections and other inflammatory events, paving the path towards more informed and effective treatments.Interestingly, the comparison of the gene expression data through a multi omic model allows to identify the trajectories of key gene activations. A selection of these trajectories could be used as prognosis biomarkers and as biomarkers of drug actions. For instance, we report that common drug targets/treatments for C. difficile target metabolites that within what we in silico detect as ``sensitive'' pathways.

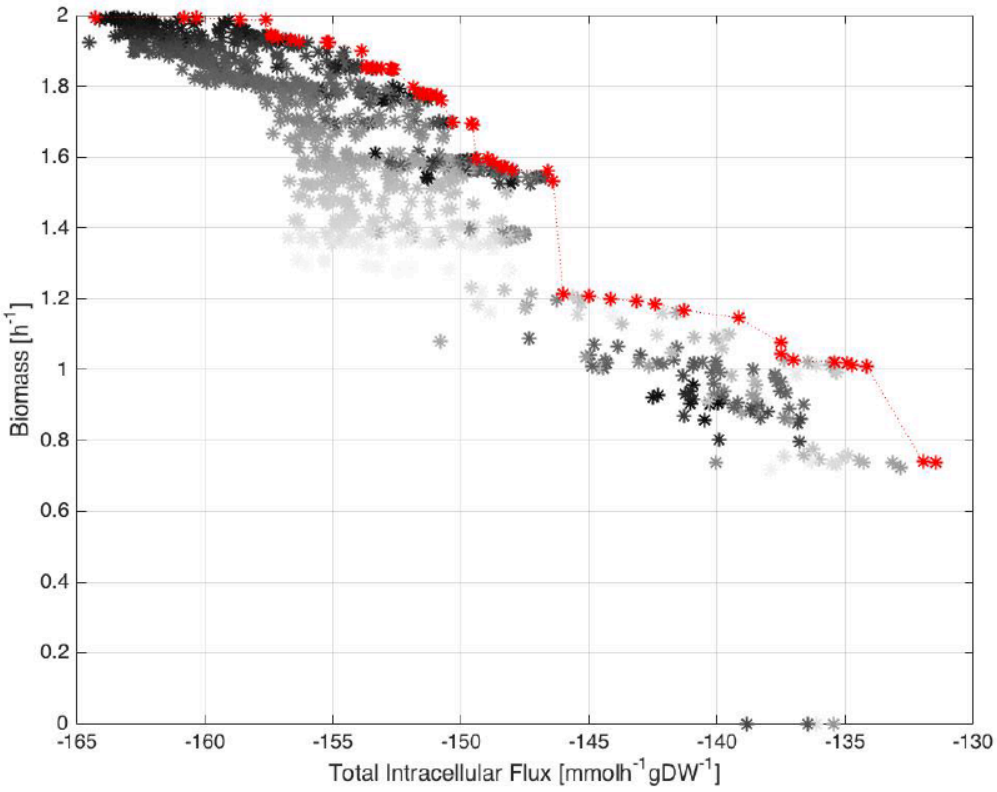
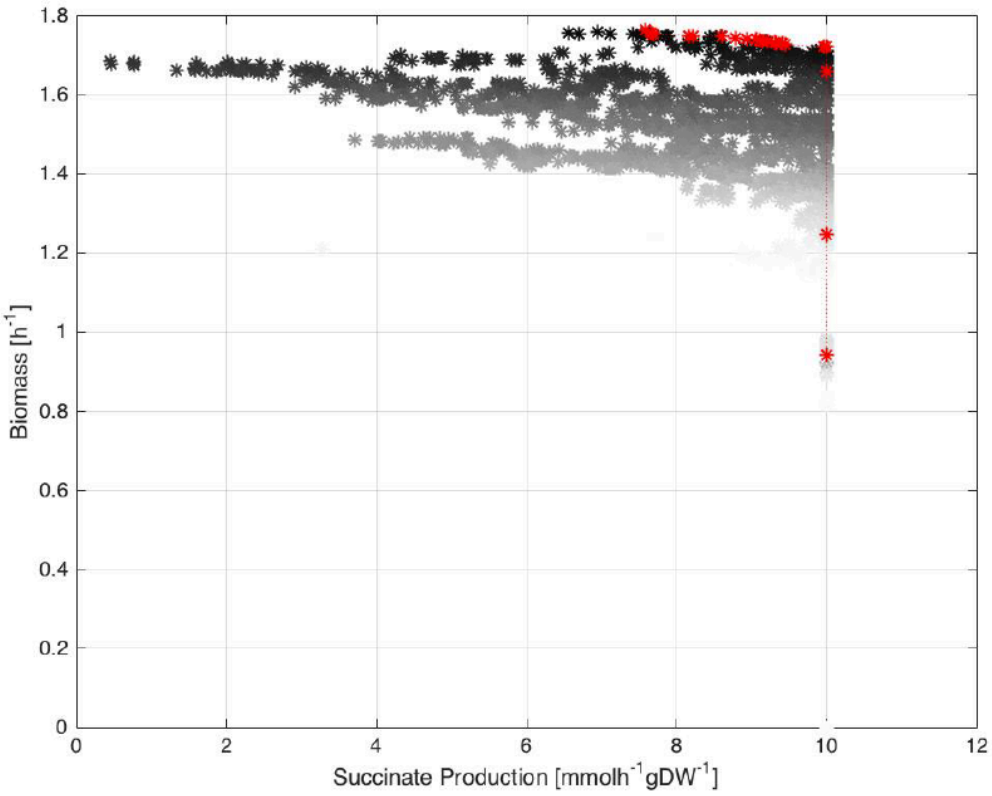



Fig. 1: Three experiments showing the effect of the second objective on model predictions. Gray points represent the solutions obtained by the parallel genetic algorithm at different time steps. The Pareto front and the Pareto-optimal points are shown in red. (Upper) Maximizing biomass while minimizing total intracellular flux. (Below) Maximizing biomass while minimizing redox potential.




Legend: see figure 1 (top).

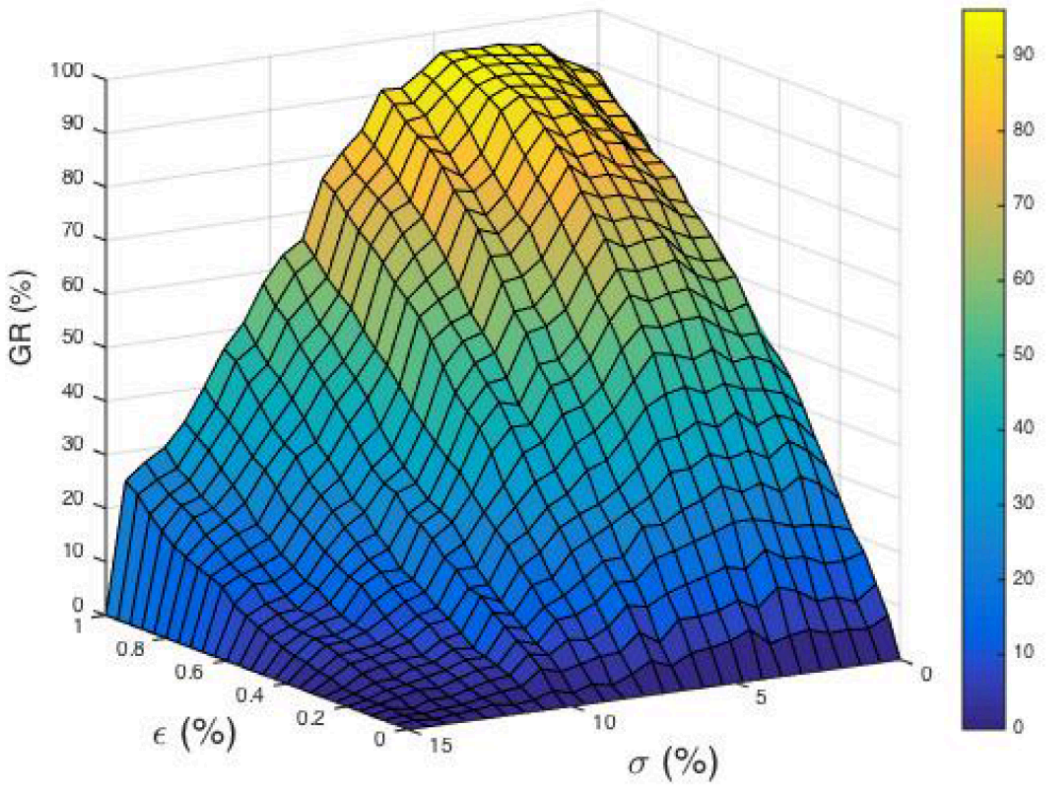



Fig. 2: GR of the second objective, the intracellular flux, for different fluctuations in the bacterium's environment and thereby of the flux bounds. Each fluctuations is represented by a different value of ε and σ

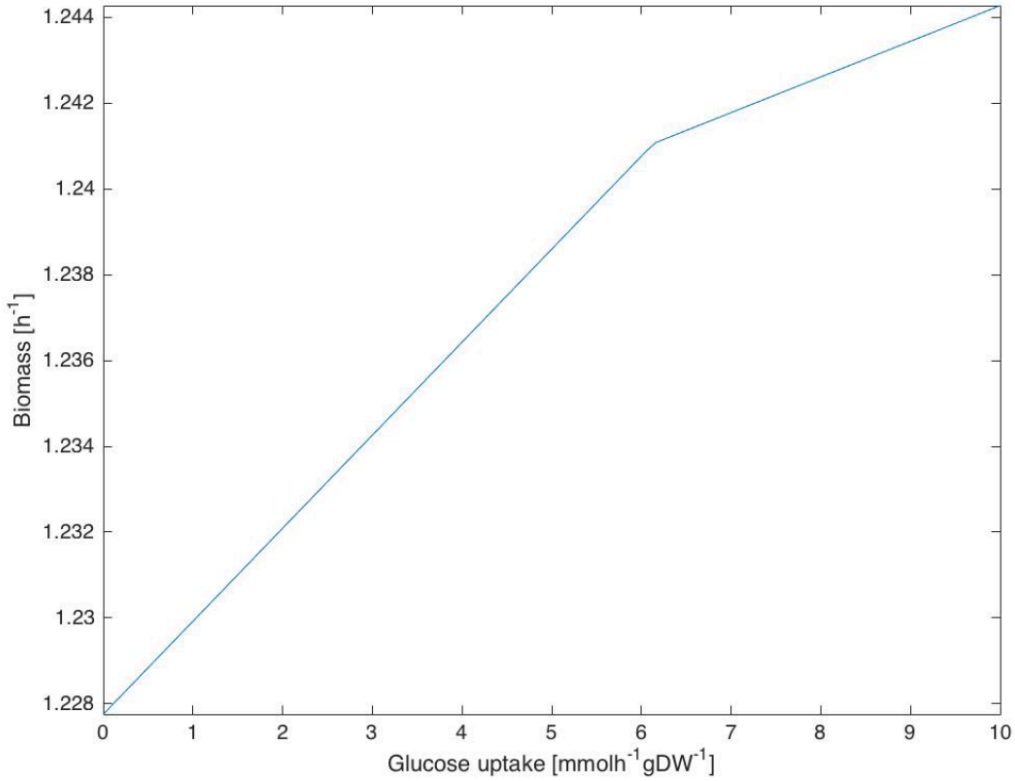

1.244
1.242
1.24
1.238
1.236
1.234
1.232
1.23
1.228
Biomass [h-1]
0
1
2
3
4
5
6
7
8
9
10
Glucose uptake [mmolh-1gDW-1]

Fig. 3: Biomass production as influx of glucose was altered from 0 mmol/gDWh to 10 mmol/gDWh.

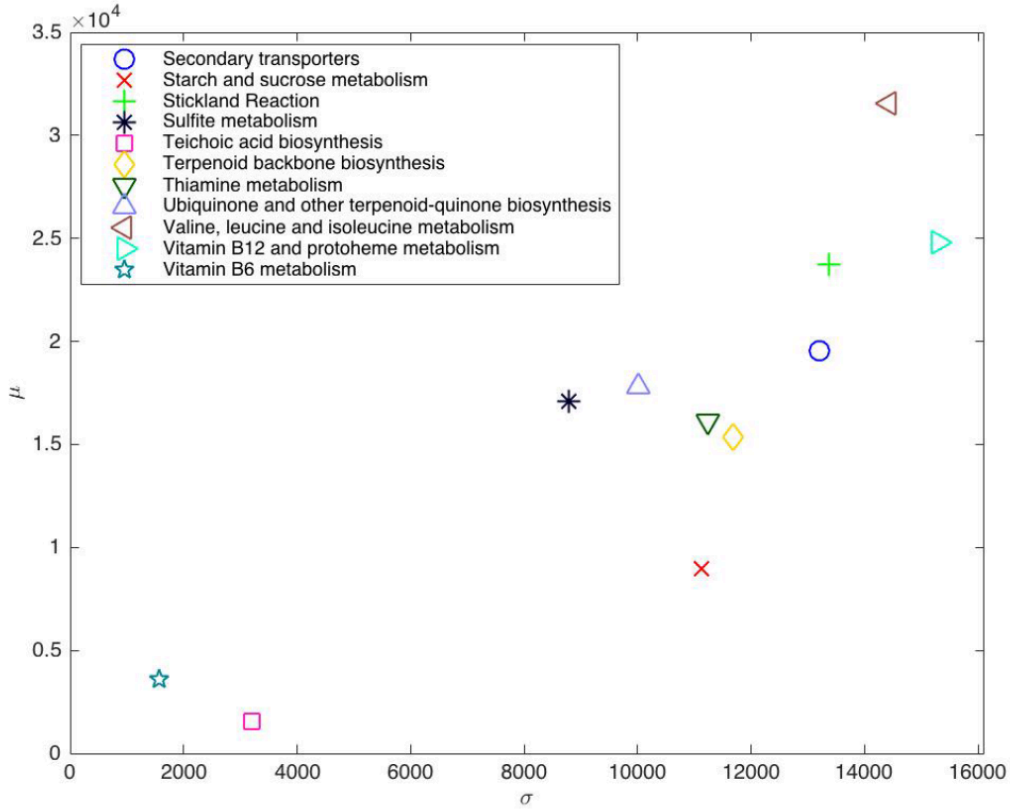



Fig. 4: The mean (μ) and standard deviation (σ) of the PEEs from the pathway-based sensitivity analysis of the C. difficile metabolic model. The output of this analysis indicates the metabolic pathways that have the greatest impact on the model outputs.

| Condition | Biomass ($h^{-1}$) | Accession Number |
|---|---|---|
| pH 7 | 1.36 | E-BUGS-55 |
| Acidic (pH 4.5) | 1.25 | E-BUGS-55 |
| Alkali (pH 8.5) | 1.23 | E-BUGS-55 |
| Aerobic | 1.19 | E-BUGS-55 |
| Anaerobic | 1.30 | E-BUGS-55 |
| High temperature (42°C) | 1.13 | E-BUGS-55 |
| Low temperature (30°C) | 1.12 | E-BUGS-55 |
| BHI (control) | 1.30 | E-BUGS-56 |
| Amoxicillin | 1.15 | E-BUGS-56 |
| Clindamycin | 1.17 | E-BUGS-56 |
| Metronidazole | 1.21 | E-BUGS-56 |

Table 1: Biomass production (growth) under different conditions. Biomass fluxes normalized by the biomass flux in control condition. Unless stated otherwise, the bacteria were grown at 37 C, anaerobic conditions, and pH 7. GE stands for gene expression data and CU stands for codon usage data.

While the antibiotics used in table 1 will control the biomass of the bacteria, it is known that beneficial bacteria will produce gamma amino-butirrate that have a positive influence on T lymphocytes. Pathogens will put generate inflammatory and immune system response. Our recent results and ongoing research show that multi omic metabolic approaches are effective in estimating the amount of disruption caused by the gut infection. Parameters estimated in the models we have developed could be implemented into the models of inflammation. Next we use machine learning methods to investigate the pathways involved in inflammatory and autoimmune diseases.

## 4 Machine learning of inflammatory diseases (sterile and infective)

The interaction of T2D diabetes with autoimmune diseases has been investigated using the adaptive landscape model from Waddington. Here we modeled the adaptive landscape using a Ornstein Uhlenbeck approach. Here

we have focused on Allergy (CD4+ T cell), Asthma (Airway epithelial cell), Ulcerative Colitis (PBMC), Crohns' Disease (PBMC), Rheumatoid Arthritis (PBMC), Chronic Fatigue Syndrome (PBMC), Systemic Lupus Erythematosus (CD14+), Type 2 Diabetes (Pancreatic islets).

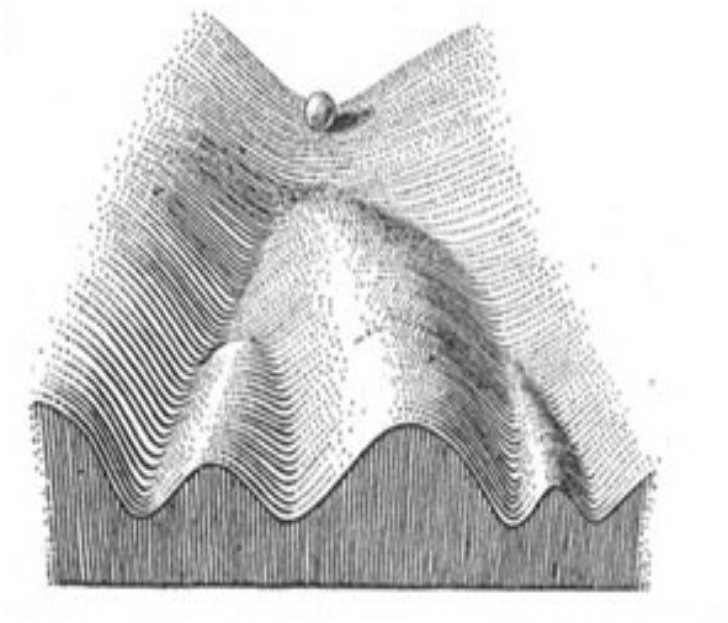


Figure 5. In the health condition the gene regulatory level can be represented as a ball at the bottom of a valley — stochastic noise allows it to vary slightly, but regulatory forces represented by the walls of the valley keep the levels clustered around a single point.

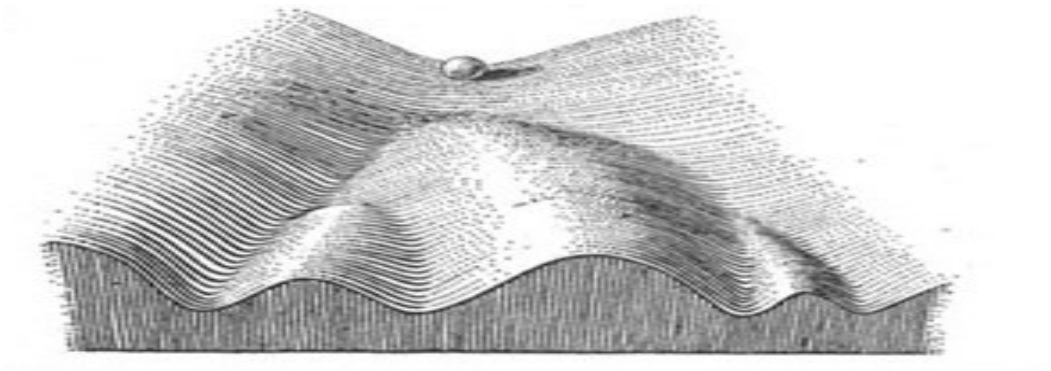


Figure 6. An early disease condition flattens the landscape, leading to more variable regulatory levels and breaking the delicate balance of regulation that maintained the stable gene expression and epigenetic signature in the face of noise.

The Ornstein Uhlenbeck could be described as:

- OU Process: $dX_t = \alpha(\theta - X_t) + \sigma dW_t$
- Deterministic term: $\alpha(\theta - X_t)$
  - Used to model selection
  - $\alpha$ = strength of selection
  - $\theta$ = optimum trait value
  - Force of selection is proportional to the distance of current trait value from the optimum $(\theta - X_t)$
- Diffusion tem: $\sigma dW_t$
  - Used to model random genetic drift
  - $\sigma$ = intensity of drift

The mean and variance relationships could be described as follows:

$$\mathbf{E}[\vec{X}](t) = e^{-\mathbf{A}t}\vec{X}(0) + \left(\mathbf{I} - e^{-\mathbf{A}t}\right)\vec{\theta}$$

$$\mathbf{Var}[\vec{X}](t) = \int_0^t e^{-\mathbf{A}s}\mathbf{\Sigma}\mathbf{\Sigma}^T e^{-\mathbf{A}^T s} ds.$$

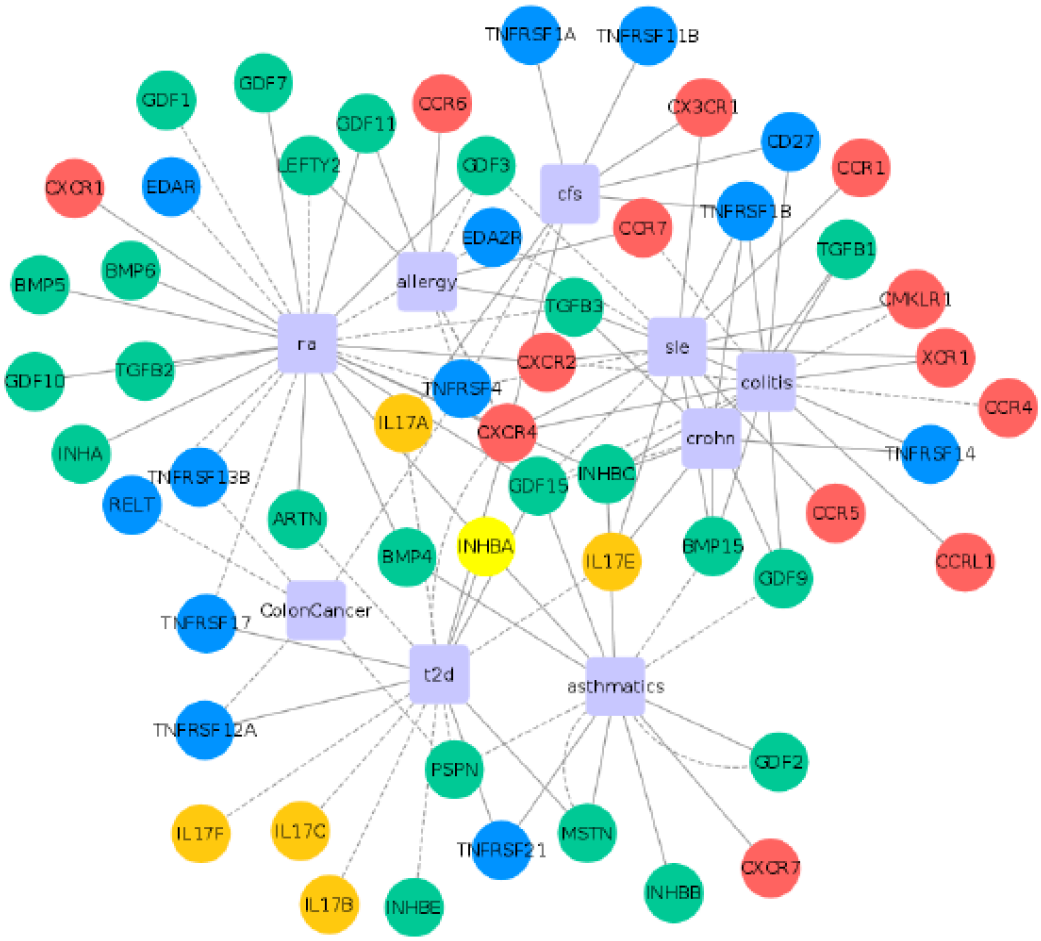



Figure 7 shows the relationship between candidate biomarker genes and autoimmune and T2D diabetes : Allergy (CD4+ T cell), Asthma (Airway epithelial cell), Ulcerative Colitis (PBMC), Crohns' Disease (PBMC), Rheumatoid Arthritis (PBMC), Chronic Fatigue Syndrome (PBMC), Systemic Lupus Erythematosus (CD14+), Type 2 Diabetes (Pancreatic islets). Publication submitted to Frontiers in Physiology. We have used the p value statistics to derive a model of the antagonist approach of proinflammatory and anti inflammatory cytokines. This model , through the AKT/mTOR pathway generate describe the comorbidity of diabetes and arthritis.

## 5 Bibliography

[1] P. Durre. World Scientific, 2014.
[2] T. Janvilisri, J. Scaria, A. D. Thompson, et al., Journal of Bacteriology, vol. 191, no. 12, pp. 3881-3891, 2009, issn: 00219193. doi: 10.1128/JB.00222-09.
[3] J. E. Emerson, R. A. Stabler, B. W. Wren, et al., Journal of Medical Microbiology, vol. 57, no. 6, pp. 757-764, 2008, issn: 00222615. doi: 10.1099/jmm.0.47657-0.
[4] A. Bordbar, J. M. Monk, Z. A. King, et al., Nature Reviews. Genetics, vol. 15, no. 2, pp. 107{20, 2014, issn: 1471-0064. doi: 10.1038/nrg3643. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24430943.
[5] E. J. O'Brien, J. M. Monk, and B. O. Palsson, Cell, vol. 161, no. 5, pp. 971-987, 2015, issn: 00928674. doi: 10 . 1016 / j . cell . 2015 . 05 . 019. [Online]. http : / / linkinghub . elsevier.com/retrieve/pii/S0092867415005681.
[6] J. S. Edwards and B. O. Palsson, Proceedings of the National Academy of Sciences of the United States of America, vol. 97, no. 10, pp. 5528{5533, 2000, issn: 0027-8424. doi: 10.1073/pnas.97. 10.5528.
[7] M. Larocque, T. Chenard, and R. Najmanovich, BMC Systems Biology, vol. 8, no. 1, p. 117, 2014, issn: 1752-0509. doi: 10.1186/s12918- 014- 0117- z. [Online]. Available: http://www. biomedcentral.com/1752-0509/8/117.
[8] A. Zelezniak, S. Sheridan, and K. R. Patil, PLoS Computational Biology, vol. 10, no. 4, 2014, issn: 15537358. doi: 10.1371/journal.pcbi.1003572.
[9] S. Klumpp, J. Dong, and T. Hwa, PLoS ONE, vol. 7, no. 11, 2012, issn: 19326203. doi: 10.1371/ journal.pone.0048542.
[10] J. Schellenberger, R. Que, R. M. Fleming, et al., Nature protocols, vol. 6, no. 9, pp. 1290{1307, 2011.
[11] R. U. Ibarra, J. S. Edwards, and B. O. Palsson, Nature, vol. 420, no. 6912, pp. 186{189, 2002, issn: 0028-0836. doi: 10.1038/nature01149.
[12] Y. K. Oh, B. O. Palsson, S. M. Park, et al., Journal of Biological Chemistry, vol. 282, no. 39, pp. 28 791{28 799, 2007, issn: 00219258. doi: 10.1074/jbc.M703759200.
[13] L. T. Bui and S. Alam, in Multi-objective Optimization in Computational Intelligence: Theory and Practice, 2008, pp. 1{19, isbn: 9781599044989. doi: 10.4018/978-1-59904-498-9.
[14] I. F. Sbalzarini, S. Muller, and P. Koumoutsakos, in Proceedings of the Summer Program 2000, pp. 63{74. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.24.3567\&rep=rep1\&type=pdf.
[15] J. Costanza, G. Carapezza, C. Angione, et al., Bioinformatics, vol. 28, no. 23, pp. 3097{3104, 2012, issn: 13674803. doi: 10.1093/bioinformatics/bts590.
[16] Angione Claudio , Costanza Jole, Carapezza Giovanni, Lio Pietro and N. Giuseppe, PLoS ONE, 2015, issn: 15455963. doi: 10.1109/TCBB.2013.95.
[17] W. M. Van Gulik and J. J. Heijnen, Biotechnology and Bioengineering, vol. 48, no. 6, pp. 681-698,1995, issn: 00063592. doi: 10.1002/bit.260480617.
[18] H. P. Bonarius, V. Hatzimanikatis, K. P. Meesters, et al., Biotechnology and Bioengineering, vol. 50, no. 3, pp. 299{318, 1996, issn: 0006-3592. doi: 10.1002/(SICI)1097-0290(19960505)50: 3<299::AID-BIT9>3.0.CO;2-B.

[19] A. L. Knorr, R. Jain, and R. Srivastava, Bioinformatics, vol. 23, no. 3, pp. 351{357, 2007, issn: 13674803. doi: 10.1093/bioinformatics/btl619.
[20] J. Ferreyra, K. Wu, A. Hryckowian, et al., Cell Host & Microbe, vol. 16, no. 6, pp. 770-777, 2014, issn: 19313128. doi: 10.1016/j.chom.2014.11.003. [Online]. Available: http: //linkinghub.elsevier.com/retrieve/pii/S1931312814004193.
[21] P. Laborie, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5547 LNCS, 2009, pp. 148-162, isbn: 3642019285. doi: 10.1007/978-3-642-01929-6_12.
[22] C Angione and P Lio, Scientific Reports, vol. 5, p. 15 147, 2015.
[23] H. Firczuk, S. Kannambath, J. Pahle, et al., Molecular Systems Biology, vol. 9, no. 635, p. 635, 2013,issn: 1744-4292. doi: 10.1038/msb.2012.73. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564266\&tool=pmcentrez\&rendertype=abstract.
[24] T. E. Gorochowski, Z. Ignatova, R. A. L. Bovenberg, et al., Nucleic Acids Research, pp. 1{11, 2015, issn: 0305-1048. doi: 10.1093/nar/gkv199. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv199.
[25] S. P. Gygi, Y. Rochon, R. Franza, et al., Molecular and Cellular Biology, vol. 19, no. 3, pp. 1720-1730, 1999, issn: 0270-7306.
[26] Y. Nakamura, T. Gojobori, and T. Ikemura, Nucleic Acids Research, vol. 27, no. 1, p. 292, 1999, issn: 03051048. doi: 10.1093/nar/27.1.292.
[27] A. Bairoch, R. Apweiler, C. H. Wu, et al., Nucleic Acids Research, vol. 33, 2005, issn: 03051048. doi: 10.1093/nar/gki070.
[28] S. Lee, S. Weon, S. Lee, et al., Evolutionary Bioinformatics, vol. 2010, no. 6, pp. 47{55, 2010, issn: 11769343. doi: 10.4137/EBO.S4608.
[29] J. S. Edwards and B. O. Palsson, Biotechnology Progress, vol. 16, no. 6, pp. 927{939, 2000, issn: 8756-7938. doi: 10.1021/bp0000712.
[30] M. E. Ritchie, B. Phipson, D. Wu, et al., pp. 1-13, 2015. doi: 10.1093/nar/gkv007.
[31] D. Knorr, A. Froehling, H. Jaeger, et al., 2011. Do 10.1146/annurev.food.102308.124129.
[32] Z. Wen and M. Morrison, Applied and Environmental Microbiology, vol. 62, no. 10, pp. 3826-3833, 1996, issn: 00992240.
[33] A. K. Roos. 2007, isbn: 9789155469528.
[34] B. P. Girinathan, S. E. Braun, and R. Govind, Microbiology (United Kingdom), vol. 160, no. PART 1, pp. 47{55, 2014, issn: 13500872. doi: 10.1099/mic.0.071365-0.
[35] S. Heuston, M. Begley, C. G. M. Gahan, et al., Microbiology, vol. 158, no. Pt 6, pp. 1389-1401, 2012, issn: 1350-0872. doi: 10.1099/mic.0.051599-0.
[36] L. Bouillaut, W. T. Self, and A. L. Sonenshein, Journal of Bacteriology, vol. 195, no. 4, pp. 844-854, 2013, issn: 00219193. doi: 10.1128/JB.01492-12.

[37] D. Molenaar, R. van Berlo, D. de Ridder, et al., Molecular Systems Biology, vol. 5, no. 323, p. 323, 2009, issn: 1744-4292. doi: 10.1038/msb.2009.82. [Online]. Available: http://dx.doi.org/10. 1038/msb.2009.82.
[38] J. S. Lee, T. Nishikawa, and A. E. Motter, Discrete and Continuous Dynamical Systems, vol. 32, no. 8, pp. 2937{2950, 2012, issn: 10780947. doi: 10.3934/dcds.2012.32.2937. arXiv: 1206.0766.
[39] O. Shoval, H. Sheftel, G. Shinar, et al., 2012. doi: 10.1126/science.1217405.
[40] C. Colijn, A. Brandes, J. Zucker, et al., PLoS computational biology, vol. 5, no. 8, e1000489, 2009.
[41] A. N. Edwards, J. M. Suarez, and S. M. McBride, Journal of Visualized Experiments : JoVE, no. 79, e50787, 2013, issn: 1940-087X. doi: 10.3791/50787. arXiv: NIHMS150003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24084491.
[42] T. Yamamoto-Osaka, S. Kamiya, S. Sawamura, et al., Journal of Medical Microbiology, vol. 40, no. 3, pp. 179{187, 1994, issn: 00222615. doi: 10.1099/00222615-40-3-179.
[43] T. Wiegert, G. Homuth, S. Versteeg, et al., Molecular Microbiology, vol. 41, no. 1, pp. 59-71, 2001, issn: 0950382X. doi: 10.1046/j.1365-2958.2001.02489.x.
[44] L. J. Drummond, D. G. E. Smith, and I. R. Poxton, Journal of Medical Microbiology, vol. 52, no.12, pp. 1033{1038, 2003, issn: 00222615. doi: 10.1099/jmm.0.05387-0.
[45] D. Karolchik, A. S. Hinrichs, and W. J. Kent, Current Protocols in Human Genetics, no. SUPPL. 71, 2011, issn: 19348266. doi: 10.1002/0471142905.hg1806s71.