Accelerating metabolic models evaluation with statistical metamodels: application to Salmonella infection models

Clémence Frioux, Sylvie Huet, Simon Labarthe, Julien Martinelli, Thibault Malou, David Sherman, Marie-Luce Taupin, Pablo Ugalde-Salas

Accelerating metabolic models evaluation with statistical metamodels: application to *Salmonella* infection models.

Clémence Frioux[1], Sylvie Huet[2], Simon Labarthe[1,3], Julien Martinelli[4,5], Thibault Malou[6], David Sherman[1], Marie-Luce Taupin[7], and Pablo Ugalde-Salas[1]

[1]Inria - INRAE - Université de Bordeaux, 33400 Talence
[2]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[3]INRAE, Univ. Bordeaux, BIOGECO, F-33610 Cestas
[4]INSERM U900, Saint-Cloud, France, Institut Curie, Saint Cloud, France, Paris Saclay University, France, MINES ParisTech, CBIO - Centre for Computational Biology, PSL Research University, Paris, France
[5]Lifeware Group, Inria Saclay Ile-de-France, Palaiseau 91120, France
[6]INSA - Institut de Mathématique de Toulouse, Toulouse, France
[7]Laboratoire LaMME, UEVE and UMR 8071, Université Paris Saclay, Evry, France

Abstract

Mathematical and numerical models are increasingly used in microbial ecology to model the fate of microbial communities in their ecosystem. These models allow to connect in a mechanistic framework species-level informations, such as the microbial genomes, with macro-scale features, such as species spatial distributions or metabolite gradients. Numerous models are built upon species-level metabolic models that predict the metabolic behaviour of a microbe by solving an optimization problem knowing its genome and its nutritional environment. However, screening the community dynamics with these metabolic models implies to solve such an optimization problem by species at each time step, leading to a significant computational load further increased by several orders of magnitude when spatial dimensions are added.

In this paper, we propose a statistical framework based on Reproducing Kernel Hilbert Space (RKHS) metamodels that are used to provide fast approximations of the original metabolic model. The metamodel can replace the optimization step in the system dynamics, providing comparable outputs at a much lower computational cost. We will first build a system dynamics model of a simplified gut microbiota composed of a unique commensal bacterial strain in interaction with the host and challenged by a *Salmonella* infection. Then, the machine learning method will be introduced, and particularly the ANOVA-RKHS that will be exploited to achieve variable selection and model parsimony. A training dataset will be constructed with the original system dynamics model and hyper-parameters will be carefully chosen to provide fast and accurate approximations of the original model. Finally, the accuracy of the trained metamodels will be assessed, in particular by comparing the system dynamics outputs when the original model is replaced by its metamodel. The metamodel allows an overall relative error of 4.71% but reducing the computational load by a speed-up factor higher than 45, while correctly reproducing the complex behaviour occurring during *Salmonella* infection. These results provide a proof-of-concept of the potentiality of machine learning methods to give fast approximations of metabolic model outputs and pave the way towards PDE-based spatio-temporal models of microbial communities including microbial metabolism and host-microbiota-pathogen interactions.

1 Introduction

Modelling in microbial ecology. Microbial ecology focuses on the study of microbial communities, called microbiota, interacting with their environment and regulated by the microbiota host [32, 5]. The gut microbiota is such a symbiotic ecosystem composed of a community of hundreds of microbial species living in the large intestine lumen, referred to as the commensals, and regulated by the epithelial cells of the host colon. The main drivers of the microbiota dynamics are the metabolism of each microbial

species, the interactions between micro-organisms and their spatio-temporal interactions with the host. In the specific case of a pathogenic infection, a new player disturbs the system and tries to shift the microbial environment from an healthy homeostasis favourable to the commensals towards a dysbiotic situation favourable to the pathogen, enabling its colonization [27, 4]. The concept of pathobiome has been introduced [35] as an analysis framework to describe the specific interactions between the commensal microbiota, the host and the pathogen leading to pathogenic infection.

Mathematical and numerical models of the gut microbiota have been recognized as suitable tools for providing mechanistic interpretations of biological observations, predicting the evolution of these ecosystems, for example in pathological situations, or defining controlling actions to lead them towards a targeted state [37, 15, 36, 21]. Mathematical models in microbial ecology are population dynamics models describing the microbial population growth, i.e. their metabolism, microbe-microbe interactions and interactions with their environment, in particular the available nutrients.

FBA framework to model microbial metabolism. A classical modelling framework to represent the microbial metabolism is Flux Balance Analysis (FBA) [24, 29]. FBA relies on metabolic models inferred from microorganism genome: the genes are annotated to identify the biochemical reactions they code for and the whole set of reactions is combined into a genome-scale metabolic network connecting the substrate metabolites the microorganism is able to metabolize to the synthesized biomass and end-products produced by the microbe.

Namely, if we note $(m_i)_{1\leqslant i\leqslant N_m}$ the set of the N_m metabolites that can be found in a micro-organism, and $(r_j)_{1\leqslant j\leqslant N_r}$ the set of the N_r reactions coded in the genome, then mass conservation equations can be written on the internal concentration of the metabolites :

$$\partial_t[m_i] = \sum_{j\in R(m_i)} \theta_{m_i,j}\nu_j \tag{1}$$

In this equation, $R(m_i)$ is the subset of reactions involving the metabolite m_i, $\theta_{m_i,j}$ is the stoichiometric coefficient of the metabolite m_i in the reaction j (negative for consumption reaction, and positive for production reaction) and ν_j is the reaction flux, i.e. the quantity of metabolite involved in the reaction by time and microbial biomass units (the flux unit is $mmol.h^{-1}.g^{-1}$). In FBA models, an additional fictitious biochemical reaction is considered: the biomass reaction r_b, with its corresponding fictional molecule b representing biomass. This comes from an abstraction of the mean content of the cell, and the energetic cost to synthesize it, see for example the works of Battley *et al.* [2]. This reaction connects the biomass precursors to the biomass b with the chemical equation

$$\sum_{i\in M(b)} \theta_{m_i,r_b} m_i \to b$$

where θ_{m_i,r_b} is the stoichiometric coefficient of metabolite m_i in the biomass reaction r_b and $M(b)$ is the subset of metabolites m_i that constitute the biomass, i.e. the metabolites needed by the microorganism for growth (to duplicate the genomic material, the metabolism machinery, the cellular membrane, etc...). The metabolic flux flowing through this biomass equation is noted ν_b and is then the amount of microbial biomass produced by time and biomass unit, with unit ($g.h^{-1}.g^{-1}$ by convention, or h^{-1}).

The FBA models aim to predict this growth rate ν_b while observing biological constraints such as the mass conservation equations (1). To achieve this prediction, the FBA framework makes important simplifying assumptions: 1) *Steady-state assumption.* All internal metabolites are assumed to be at steady-state in the cell, so that the mass conservation equation (1) reduces to a linear system on the flux vector $\nu := (\nu_j)_{1\leqslant j\leqslant N_r}$ gathering the fluxes of the N_r reactions of the metabolic network,

$$A \cdot \nu = 0$$

where A is the reaction matrix, i.e. the matrix of dimension $N_m \times N_r$ with $A_{ij} := \theta_{m_i,j}$ the stoichiometric coefficient of metabolite i in the reaction j, gathering the whole set of conservation equations for the metabolites and reactions involved in the metabolic network; 2) *Biomass maximization.* The microbes are assumed to be instantaneously maximizing the biomass production in a given nutritional context; 3) *Flux constraints.* Every flux are constrained by intrinsic limits, related for example to metabolite transporter capacities, or known enzymatic efficiency. These limits are noted c_{min} and c_{max} so that $c_{min} \leq \nu \leq c_{max}$.

Hence, the biomass production and all the metabolic fluxes in the microbial machinery can be predicted with the constrained optimization FBA problem

$$\text{find } \nu^* \in \mathbb{R}^{N_r}, \text{such that } \nu^* := \begin{array}{c} \arg\max \\ \nu \in \mathbb{R}^{N_r} \\ A \cdot \nu = 0 \\ c_{min} \leq \nu \leq c_{max} \end{array} \quad \nu_b \tag{2}$$

This problem searches for the optimal growth rate represented by the component ν_b, which is the biomass formation flux. It is obtained by the system under mass-balance and flux constraints. Mathematically speaking, this optimization problem is linear and can be solved using linear programming: very efficient solvers exist for such a problem, even for high dimensional problems like this one, where N_r is classically around several thousands. A classical FBA toolbox is the Cobra toolbox (in Matlab environment) [11] or its python equivalent Cobrapy [9].

Nutritional environment described as constraints on uptake fluxes. Important FBA model parameters are constraints on substrate flux from the extracellular compartment into the intracellular compartment, i.e. the first reactions of the metabolic network, enabling nutrients to enter the microbial cell. These constraints represent the possible uptake for the microorganism, hence representing a proxy of the microbe nutritional environment, i.e. the available nutrients for the microbial species to activate its metabolism.

The uptake reactions are exchange reactions, i.e. reactions at the interface between the intra and extracellular media. Indeed, by construction, exchange reactions are reactions

$$m_i \longrightarrow \boldsymbol{m}_i$$

between the extracellular pool $\boldsymbol{m}_i$, i.e. the nutritional environment, and the intracellular pool m_i of the corresponding metabolite.

If we note $c_s^{(up)}$ the upper bound on the uptake fluxes ν_{up} of the N^{up} metabolites in the extra-cellular environment, $c_s^{(up)} \leq \nu_{up} \leq 0$, we get a mapping $\mathcal{F}_s$ between $c_s^{(up)}$ and the FBA solution for the bacterial strain s

$$\mathcal{F}_s : \quad \mathbb{R}^{N^{up}} \longrightarrow \mathbb{R}^{N_r} \tag{3}$$

$$c_s^{(up)} \mapsto \nu^* \tag{4}$$

where ν^* is the FBA solution with the constraints $c_s^{(up)}$ for the strain s. This mapping allows to tune the uptake constraints to adapt the FBA prediction to a specific nutritional environment context. We note that by convention, uptake fluxes are negative due to the exchange reaction orientation.

Dynamic FBA Eq. 4 can be used as the second member of an ordinary differential equation (ODE) to compute the growth or consumption rates of a population dynamics equation in a framework termed dynamic FBA or dFBA [19]. Let us introduce a generic dFBA model describing the dynamics of a microbial population density b growing on a substrate of density s with metabolic fluxes described by a FBA model 2 and the resulting mapping 4. We have

$$\partial_t b = \mathcal{F}_{b,1}(c^{(up)}(s,b))b \tag{5}$$

$$\partial_t s = \mathcal{F}_{b,s}(c^{(up)}(s,b))b \tag{6}$$

In this equation, $c^{(up)}(s,b)$ is a function mapping the state variables b and s to the constraints $c^{(up)}$ on the substrates applied in the FBA model 2. As an example, we can set $c^{(up)}(b,s) = \frac{s}{L_{dt}b}$ to model the fact that the remaining substrate pool s is shared between the current microbial population b at a time rate L_{dt}. We indicate by $\mathcal{F}_{b,1}$ the biomass production flux (index 1) and $\mathcal{F}_{b,s}$ the consumption flux of metabolite s (index s) of the FBA model of b. In the sequel, we will simplify the notations by noting $\mathcal{F}_b(s,b) = \mathcal{F}_b(c^{(up)}(s,b))$.

The dFBA framework is used in an increasing number of system biology models of the gut microbiota[18, 7]. However, dFBA involves the resolution of many FBA optimization problems during the time integration inducing high computational costs that can lead to intractable computations when the dFBA is repeated multiple times, like in several intensive numerical applications such as sensitivity analysis, inference or PDEs, advocating for reduction method.

Outline of the paper. This paper aims to 1) adapt a metamodeling method to the context of metabolic models to accelerate the computation of a population dynamics model coupled to a FBA model such as Eq. (5), 2) benchmark this method in the specific context of an ODE-based model of the gut environment during the infection of an enteric pathogen: *Salmonella enterica* Typhimurium. We want to substitute the FBA optimization problem solved at each time step by an approximate model, built with a *Reproducing Kernel Hilbert Space* (RKHS) metamodeling method. The RKHS metamodel is a machine learning approach: an approximation of the model image is built from the model evaluation in a sample of the state space (i.e. a learning database). This metamodel will be used to predict the model response for new points outside the learning database, with a faster computation than the original optimization problem.

First, we will set up the general framework of the accelerated model using eq. (5) as a toy example to introduce the essential mathematical results for RKHS metamodeling in Sec. 2. Then, we will use the acceleration method on a more evolved population dynamics model of *Salmonella* infection with the host response in Sec. 3. This population dynamic model will be used to produce a learning database to train the metamodel in Sec. 4. Next, the hyperparameters of the learning method will be selected in Sec. 5 in order to provide a good trade-off between prediction accuracy and computation speed. Finally, the RKHS metamodel will be derived with the selected hyperparameters and its accuracy will be assessed in Sec. 6. See Fig. 1 for a sketch image of the overall methodology.

2 Mathematical framework for the RKHS metamodel

2.1 Accelerating a dFBA with a metamodel: general methodology

To accelerate the computation of problem 5, we speed up the evaluation of $\mathcal{F}_b$ by using a metamodel $\hat{\mathcal{F}}_b$, resulting in an overall acceleration for the time integration of (5) (see Fig. 1, left panel). Namely, we solve the following problem.

$$\partial_t b = \hat{\mathcal{F}}_{b,1}(c^{(up)}(s,b))b \tag{7}$$

$$\partial_t s = \hat{\mathcal{F}}_{b,s}(c^{(up)}(s,b))b \tag{8}$$

where $\hat{\mathcal{F}}_b$ is the best approximation of $\mathcal{F}_b$ in a particular functional space, here a specific RKHS called ANOVA-RKHS. We now precise the mathematical framework we use by introducing important results for the global understanding of RKHS metamodeling. We next introduce ANOVA-RKHS that will be used for variable selection. These results are however classical, and we do not provide their proof that can be found in the corresponding references. The main contribution of the paper is the specific adaptations needed for the application of ANOVA-RKHS metamodels to the context of microbial population dynamics models, and in particular the context-specific learning database construction, hyperparameter tuning and selection criteria that will be crucial for tailoring a trade-off between metamodel accuracy and speed-up (see Fig. 1, right panel).

2.2 Metamodeling and Hoeffding decomposition

Let us set up the context of metamodeling for metabolic models. We consider $\mathbf{X}$ a N^{up}-dimensional random vector of possible metabolic constraints for the FBA model inputs with known distribution $P_{\mathbf{X}} = P_1 \times \cdots \times P_{N^{up}}$ on $\mathcal{X}$ and we construct

$$\mathbf{Y}^s = \mathcal{F}_s(\mathbf{X})$$

where $\mathbf{Y}^s$ is a N_r-dimensional vector and s an index designating the bacterial strain related to the FBA model. In this paper, we will consider real-valued meta-models. For a given $1 \leq j \leq N_r$ and a given strain s, building the meta-model m_j of the real-valued function $\mathcal{F}_{s,j}$ amounts to solve in a given functional space $\mathcal{H} \subset L^2(P_{\mathbf{X}})$, the non-parametric Gaussian regression model [13]

$$\mathbf{Y}^s_j = m_j(\mathbf{X}) + \sigma\varepsilon \tag{9}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is independent of (X) and the variance σ^2 is unknown.

When the input variables $\mathbf{X}$ are independent, and since $m_j \in L^2(P_{\mathbf{X}})$, the classical Hoeffding-Sobol decomposition holds (see [33, 34] section 11.4). The functions m_j can be decomposed with its ANOVA functional expansion

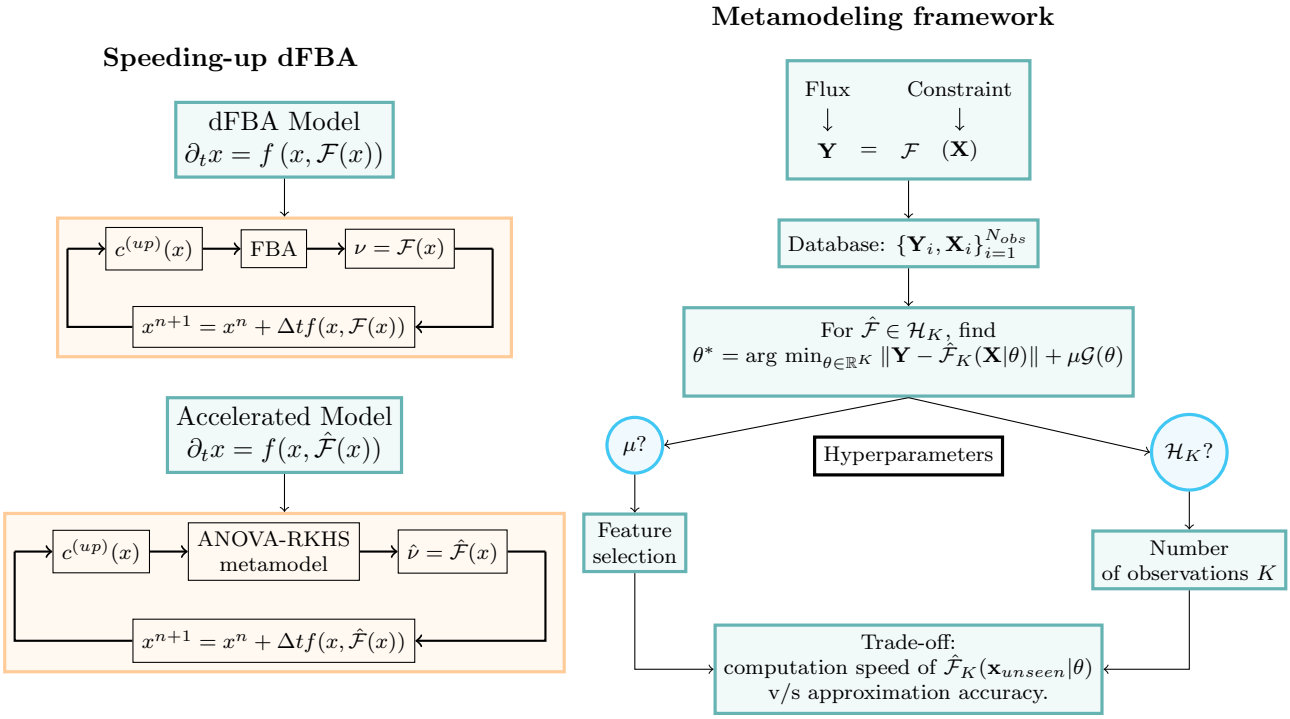


Figure 1: **Sketch of the general methodology.** *Left panel: speeding up dFBA.* The dFBA framework (upper panel) is defined by the coupling of a FBA metabolic model with a dynamic system. Numerically, this remains to loop over a time integration scheme in which a FBA is solved at each time step. We propose a new framework (lower panel) where the FBA model is replaced by a low-computational-cost metamodel speeding up the time integration process. *Right panel: metamodeling framework.* We set up the general statistical framework where the flux $\mathbf{Y}$ is the output of the FBA model $\mathcal{F}$ given the input $\mathbf{X}$. We then assemble a learning dataset by sampling the input space ($\mathbf{X}_i$) and computing the corresponding FBA output $\mathbf{Y}_i$ with N_{obs} observations. The metamodel is then defined as the solution of a non-linear non-parametric regression problem in a finite dimensional functional space $\mathcal{H}_K$ of dimension K with regularization function $\mathcal{G}$. In practice, we will choose a group-lasso regularization to perform feature selection together with the metamodel computation. This regression problem has two hyperparameters to be chosen: the regularization parameter μ, that will tune the number of selected input variables, and the dimension K of the functional space, which is related to the number of observations in the RKHS framework (see Sec. 2 and 5). Selecting lower number of features or lower K decreases the computation load of the metamodel evaluation in a new unseen point $\mathbf{x}_{unseen}$ and thus accelerates the ODE model integration but decreases the metamodel accuracy: a trade-off must be sought (see Sec. 5).

$$m_j(\mathbf{x}) = m_{j,0} + \sum_{p \in \mathcal{P}} m_{j,p}(\mathbf{x}_p)$$

where p is a multi-index, $\mathcal{P}$ the power set of $\{1, \cdots, N^{up}\}$, $\mathbf{x}_p$ denotes the vector with components $\mathbf{x}_j$ for $j \in p$. The functions m_p are $L^2(P_{\mathbf{X}})$ functions centered and orthogonal in $L^2(P_{\mathbf{X}})$, so that the variance of m_j can be decomposed with

$$Var(m_j(\mathbf{x})) = \sum_{p \in \mathcal{P}} Var(m_{j,p}(\mathbf{x}_p)).$$

The Hoeffding decomposition is used to separate principal effects (the function $m_{j,p}$ that involve one unique input variable $\mathbf{x}_i$) from variable interactions (the functions $m_{j,p}$ with $|p| > 1$, i.e. involving more than one input component). The Hoeffding decomposition is widely used for sensitivity analysis, since Sobol index directly derives from it, or for variable selection: the relative contribution of the functions $m_{j,p}$ in the Hoeffding decomposition allows to neglect the less contributive terms which can lead to discard some input variables if all the functions they are involved in are neglected.

2.3 Generalities on RKHS metamodel

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^{N^{up}}$. A definite symmetric kernel is a function

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (x, x') &\mapsto k(x, x') \end{aligned}$$

such that, for all $N \in \mathbb{N}$ and $x_1, \cdots, x_N \in \mathcal{X}^N$, the Gramm matrix $(k)_{i,j} = k(x_i, x_j)$ is symmetric positive definite.

The Moore–Aronszajn's theorem ensures a bijective mapping between the space of positive definite kernels and specific Hilbert spaces termed Reproducing Kernel Hilbert spaces (or RKHS).

Theorem 1 (Moore–Aronszajn [1])**.** *Setting $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a symmetric positive definite kernel, there exists a unique Hilbert space $\mathcal{H}_k$ of real-valued functions on $\mathcal{X}$ defined as the completion of*

$$\tilde{\mathcal{H}_k} := \left\{ f : \mathcal{X} \to \mathbb{R} | f(\cdot) = \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i), \beta_i \in \mathbb{R}, z_i \in \mathcal{X}, \|f\|_{\mathcal{H}_k} < \infty \right\}$$

with respect to the norm $\|\cdot\|_{\mathcal{H}_k}$ induced by the scalar product

$$\left\langle \sum_{i=1}^{\infty} \beta_i k(\cdot, z_i), \sum_{j=1}^{\infty} \alpha_j k(\cdot, y_j) \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i \alpha_j k(y_j, z_i)$$

that endows $\mathcal{H}_k$. The kernel k is termed the Reproducing kernel of the RKHS $\mathcal{H}_k$.

Reciprocally, if $\mathcal{H}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ endowed with its inner product noted $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and if $\forall x \in \mathcal{X}$ the functional $f \mapsto f(x)$ is continuous on $\mathcal{H}$, then $\mathcal{H}$ is a RKHS [6]. The reproducing kernel of $\mathcal{H}$ can be exhibited according to the Riesz theorem: for all $x \in \mathcal{X}$, there exists a unique $k_x \in \mathcal{H}$ such that for all $f \in \mathcal{H}, f(x) = \langle f, k_x \rangle_{\mathcal{H}}$. The reproducing kernel k is then defined as

$$\begin{aligned} k : \quad \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (x, x') &\mapsto k_{x'}(x) = \langle k_x, k_{x'} \rangle_{\mathcal{H}} \end{aligned}$$

and we have by construction the *reproducing property*

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

The RKHS framework is very powerful to approximate solutions of the non-linear regression problem 9 on the basis of N_{obs}-samples $(\mathbf{Y}^s_{j,i}, \mathbf{X}_i), i = 1, \cdots, N_{obs}$ in the RKHS $\mathcal{H}_k$. Namely, we will address the problem of finding

$$m_j^* := \arg\min_{m_j \in \mathcal{H}_k} \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} (\mathbf{Y}_{j,i}^s - m_j(\mathbf{X}_i))^2 + g(\|m_j\|_{\mathcal{H}_k}) \tag{10}$$

where g is a strictly increasing function allowing to regularize the regression problem. As $\mathcal{H}_k$ is a functional space of *a priori* infinite dimension, this problem must be discretized to be solved. In the RKHS framework, the Representer theorem reduces this problem to a N_{obs}-dimensional minimization

Theorem 2 (Representer Theorem [30])**.** *Any function $m_j \in \mathcal{H}_k$ minimizing equation* (10) *admits a representation of the form*

$$m_j(\cdot) = \sum_{i=1}^{N_{obs}} \alpha_i k(\cdot, \mathbf{X}_i)$$

so that problem (10) can be replaced by finding

$$\alpha^* := \arg\min_{\alpha \in \mathbb{R}^{N_{obs}}} \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} \left(\mathbf{Y}_{j,i}^s - \sum_{j=1}^{N_{obs}} \alpha_j k(\mathbf{X}_j, \mathbf{X}_i) \right)^2 + g\left(\left(\sum_{i=1}^{N_{obs}} \sum_{j=1}^{N_{obs}} \alpha_i \alpha_j k(\mathbf{X}_j, \mathbf{X}_i) \right)^{1/2} \right) \tag{11}$$

or, in vectorial form

$$\alpha^* := \arg\min_{\alpha \in \mathbb{R}^{N_{obs}}} \frac{1}{N_{obs}} \|\mathbf{Y}_j^s - K \cdot \alpha\|_F^2 + g\left(\left(\alpha^t K \alpha \right)^{1/2} \right) \tag{12}$$

where K is the Gram matrix obtained with the kernel k and $(\mathbf{X}_i)_{i=1,\cdots,N_{obs}}$.

The inference of α^* uniquely defines the metamodel m^* which can be evaluated in a new point $X \in \mathbb{R}^{N^{up}}$ with

$$m^*(X) := \sum_{i=1}^{N_{obs}} \alpha_i^* k(X, \mathbf{X}_i). \tag{13}$$

We note that the computational load of eq. (13) linearly depends on N_{obs}.

2.4 ANOVA-RKHS

In (12), multidimensional kernels can be chosen to assemble the matrix K, resulting in a simple regression problem if $g = Id$. However, in the context of metabolic modelling, vectors $\mathbf{X}$ can be of high dimension (a.e. in our application $N^{up} = 9$) implying a large number N_{obs} of samples in the learning set to cover this high dimension space. Thinking in term of computational budget for the evaluation of eq. (13) which is linearly tuned by N_{obs}, it is appealing to reduce N^{up} and thus the dimension of the space of state variable involved in the metamodel. For a fixed number N_{obs} allowed by the computational budget, the metamodel approximation accuracy is expected to be better in a reduced state variable space (see Sec. 7.3 for a deeper discussion on this aspect): we then adopt a more evolved method based on variable selection framework introduced in [13] and based on a very specific RKHS introduced in [8], the ANOVA-RKHS. The ANOVA-RKHS $\mathcal{H}$ is built as a direct sum of sub-RKHS $\mathcal{H}_p$ so that a given function $f \in \mathcal{H}$ will have for Hoeffding decomposition its decomposition on the subspaces $\mathcal{H}_p$. Using the ANOVA-RKHS, we will build a metamodel only involving the most significant state variables (i.e. a reduced number N^{up}), reducing the input space dimension and thus increasing the metamodel accuracy and the computational speed-up for a given computational budget fixed by N_{obs}. The goal of the ANOVA-RKHS is not to accelerate the metamodel computation in (12), but rather to speed up the metamodel evaluation in an unseen point in (13).

Let us note $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_{N^{up}}$. For each coordinate $a \in \{1, \cdots, N^{up}\}$, a kernel k_a and its corresponding RKHS $\mathcal{H}_a$ are chosen on $\mathcal{X}_a$, with the additional properties: 1) k_a is $P_a \times P_a$ mesurable on $\mathcal{X}_a \times \mathcal{X}_a$ and 2) $\mathbb{E}_{P_a} \sqrt{k_a(X_a, X_a)} < \infty$.

The RKHS $\mathcal{H}_a$ can be decomposed as $\mathcal{H}_a = \mathcal{H}_{0a} \overset{\perp}{\oplus} \mathcal{H}_{1a}$ where

$$\mathcal{H}_{0a} := \{f_a \in \mathcal{H}_a, \mathbb{E}_{P_a}(f_a(X_a)) = 0\}, \qquad \mathcal{H}_{1a} := \{f_a \in \mathcal{H}_a, f_a(X_a) = C\}$$

the kernel associated to the RKHS $\mathcal{H}_{0a}$ being defined as follows [13] p.8:

$$k_{0a}(X_a, X'_a) = k_a(X_a, X'_a) - \frac{\mathbb{E}_{U\sim P_a}[k_a(X_a, U)]\mathbb{E}_{U\sim P_a}[k_a(X'_a, U)]}{\mathbb{E}_{(U,V)\sim P_a\otimes P_a}[k_a(U, V)]}. \tag{14}$$

The ANOVA kernel is finally defined by

$$k(X, X') = \left(\prod_{a=1}^{N^{up}} (1 + k_{0a}(X_a, X'_a))\right) = 1 + \sum_{p\in\mathcal{P}} k_p(X_p, X'_p) \tag{15}$$

with $k_p(X_p, X'_p) = \prod_{a\in p} k_{0a}(X_a, X'_a)$. The corresponding RKHS is finally

$$\mathcal{H} = \left(\prod_{a=1}^{N^{up}} \mathbb{1} \overset{\perp}{\oplus} \mathcal{H}_{0a}\right) = \mathbb{1} + \sum_{p\in\mathcal{P}} \mathcal{H}_p \tag{16}$$

where $\mathcal{H}_p$ is the RKHS associated to k_p. Let us now take any function f in the ANOVA-RKHS $\mathcal{H}$. We get by the reproducing property and linearity

$$f(x) = \langle f, k(x, .)\rangle_{\mathcal{H}} = f_0 + \sum_{p\in\mathcal{P}} f_p(x), \quad \text{with } f_p(x) = \langle f, k_p(x_p, .)\rangle_{\mathcal{H}} \tag{17}$$

As the functions f_p are centered and uncorrelated by construction, this decomposition is also the Hoeffding decomposition of f. This setting will be used for variable selection: in the following, the numerical problem will be set up, with a group-lasso regularization that will select the important variables and variables interactions.

2.5 Discretization of the regression problem and metamodel construction

From the representer theorem 2 and the ANOVA-RKHS reproducing property in eq. (17), we can state the following finite dimension parametric regression problem: for a given $1 \leqslant j \leqslant N_r$ and a given bacterial strain s, find

$$\hat{\theta}^s_{0,j}, (\hat{\theta}^s_{p,j})_{p\in\mathcal{P}} := \underset{\substack{\theta^s_{0,j}\in\mathbb{R} \\ \theta^s_{p,j}\in\mathbb{R}^{N_{obs}}, \forall p\in\mathcal{P}}}{\arg\min} \|\mathbf{Y}^s_j - (\theta^s_{0,j}\mathbb{1} + \sum_{p\in\mathcal{P}} K_p\theta^s_{p,j})\|_2^2 + \mathcal{G}(W, \theta^s_{p,j}) \tag{18}$$

with $K_p \in \mathbb{R}^{N_{obs}\times N_{obs}}$ the Gram matrix such that $(K_{p\ j_1,j_2})_{1\leq j_1,j_2\leq N_{obs}} = k_p(c^{j_1}, c^{j_2})$, the value of the kernel k_p evaluated at constraint points c^{j_1} and c^{j_2}. In this equation, the norm $\|\cdot\|_2$ is the classical l_2 norm: $\|\mathbf{x}\|_2 = \left(\sum_{i=1,\cdots,N_{obs}} x_i^2\right)^{1/2}$. The term $\mathcal{G}$ is a regularization term that writes:

$$\mathcal{G}(W, \theta^s_{p,j}) = N_{obs}\mu \sum_{p\in\mathcal{P}} \|W\theta^s_{p,j}\|_2$$

with μ an hyperparameter and W some weight matrix.

If the weight matrix is $W = K_p^{1/2}$, then $\|W\theta^s_{p,j}\|_2 = \|f_p\|_{\mathcal{H}_p}$ where $f_p := \sum_{i=1}^{N_{obs}} \theta^s_{p,j,i} k_p(\mathbf{X}_i, \cdot)$ with $\mathbf{X}_i$ the i-th row of the learning database $\mathbf{X}$. If the weight matrix is $W = \frac{1}{\sqrt{n}}K_p$, then $\|W\theta^s_{p,j}\|_2 = \|f_p\|$ where $\|\cdot\|$ is the empiric l_2 norm. A composite criteria can be chosen such as the ridge group sparse criteria $\sqrt{N_{obs}}\gamma \sum_{p\in\mathcal{P}} \|K_p\theta^s_p\|_2 + N_{obs}\mu \sum_{p\in\mathcal{P}} \|K_p^{1/2}\theta^s_p\|_2$ as introduced in [13] (formula 17). In this exploratory study, we set $W = Id$, leading to a group-lasso criteria.

To compute K_p, a numerical version of the ANOVA-RKHS is needed, and in particular the computation of k_{a0} and the integrals in Eq. (14). These integrals are computed empirically for all $1 \leq i \leq N_{obs}$ once for all and stored for further use with the formulas:

$$\mathbb{E}_{\mathcal{U}\sim P_a}[k_a(\mathbf{X}_{i,a}, U)] \simeq \frac{1}{N_{obs}} \sum_{j=1}^{N_{obs}} k_a(\mathbf{X}_{i,a}, \mathbf{X}_{j,a}) \quad \text{and} \quad \mathbb{E}_{\mathcal{U}\sim P_a\otimes\mathcal{V}\sim P_a}[k_a(U, V)] \simeq \frac{1}{N_{obs}^2} \sum_{i=1}^{N_{obs}} \sum_{j=1}^{N_{obs}} k_a(\mathbf{X}_{i,a}, \mathbf{X}_{j,a}) \tag{19}$$

where $\mathbf{X}_i$ is the i-th row of the learning database $\mathbf{X}$ and a is the mono-dimensional index. Note that these integrals are respectively mono and bi-dimensional, which limits the computational load.

This estimation problem is a $N_{obs} \times |\mathcal{P}| + 1$-dimensional optimization problem, which can be numerically expensive if N^{up} and N_{obs} are large. The problem can be reduced by considering interactions up to a certain order. However, the minimization problem is done off-line once for all. Then, the function

$\mathcal{F}_{s,j}$ can be approximated in a new point $\tilde{c}^{(up)}$ in the input parameter space by $\hat{\mathcal{F}}_{s,j}(\tilde{c}^{(up)})$ defined with the explicit formula

$$\hat{\mathcal{F}}_{s,j}(\tilde{c}^{(up)}) := \hat{\theta}^s_{0,j} + \sum_{p\in\mathcal{P}} F_p(\tilde{c}^{(up)}) \cdot \hat{\theta}^s_{p,j} \tag{20}$$

where $F_p(\tilde{c}^{(up)})$ is the N_{obs} dimensional vector

$$F_p(\tilde{c}^{(up)}) := \Big(k_p(\mathbf{X}_i, \tilde{c}^{(up)})\Big)_{1\leqslant i\leqslant N_{obs}}$$

i.e., the evaluation of the k_p kernel at $\tilde{c}^{(up)}$ and the N_{obs} learning set points $\mathbf{X}_i$. This analytical formula is fast to compute: it has the complexity of a dot product once k_p are evaluated. In practice, we will use Matern kernels for kernels k_a, $a \in \{1, \cdots, N^{up}\}$, the parameters of the Matern kernel being fixed to *a priori* values so that the kernel are computed with the formula $(c_1, c_2) \mapsto (1 + 2|c_1 - c_2|)e^{(-2|c_1-c_2|)}$.

Note that k_{0a} in eq. (14) is needed to compute k_p: the computation of the first integral $\mathbb{E}_{\mathcal{U}\sim P_a}[k_a(\mathbf{X}_{i,a}, U)]$ in eq. (14) is done empirically through eq. (19) while the others have been computed once for all and stored, reducing the computation time.

3 Population dynamics model of *Salmonella* infection, including host inflammatory response

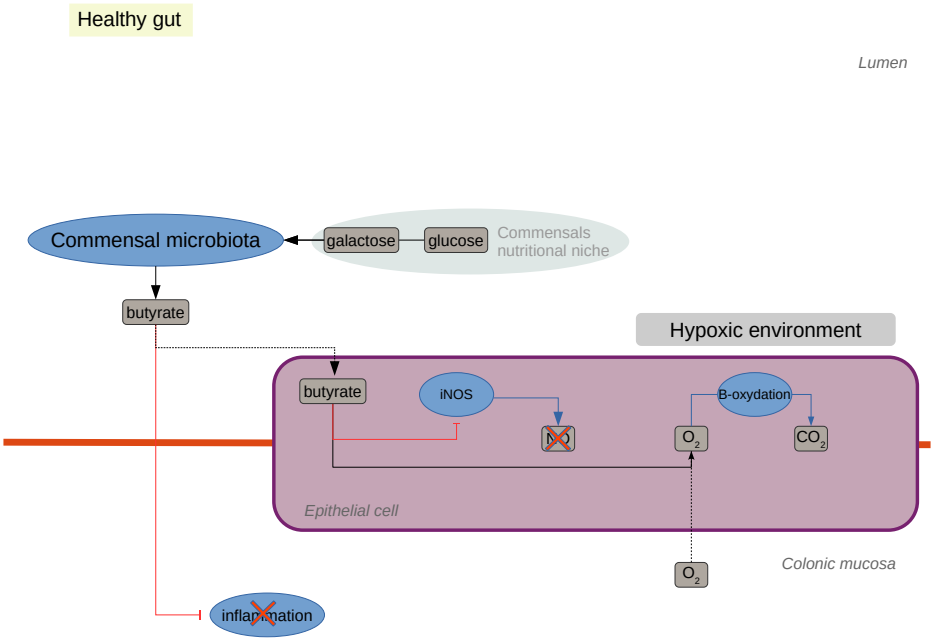
We now contextualize the previous methodology to a dynamic system describing *Salmonella* infection in the gut lumen. This application example is a sound benchmark to show the potentiality of our method because 1) it is a representative example of the intrinsic complexity of a system biology model of the gut by involving two different metabolic models and 10 metabolites screened in time in two compartments, 2) the model involves stiff dynamics after infection, making it sensitive to flux approximation errors and thus more difficult to approximate by a metamodel.

3.1 Biological context of Salmonella infection.

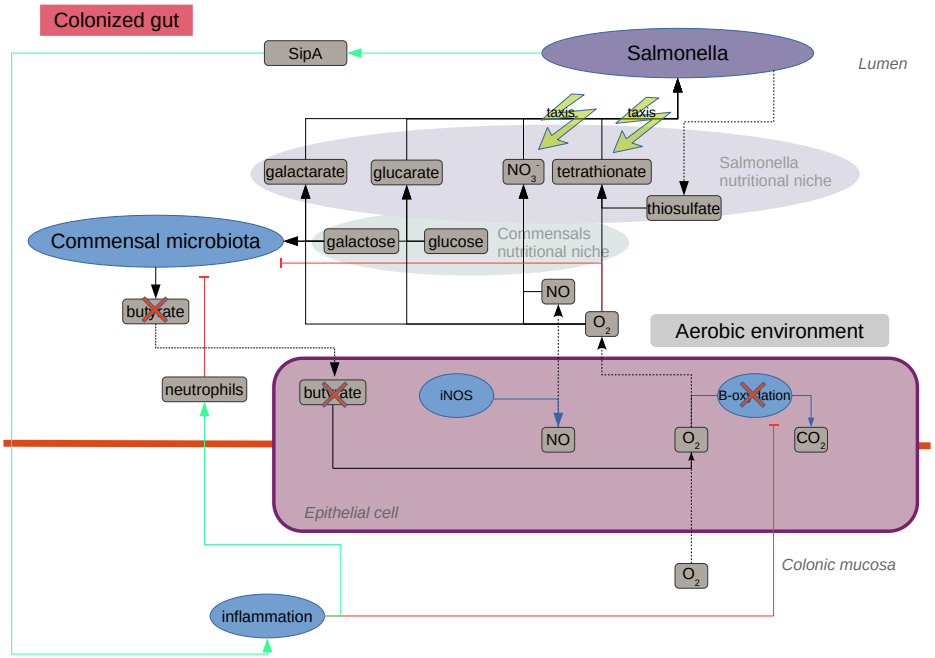
Salmonella Thyphimurium uses a very complex mechanism to invade the gut. Let us characterized the healthy gut homeostasis: it will emphasize by contrast how the pathogen colonizes the intestine lumen.

Healthy gut. The environment of a healthy gut is anaerobic: the commensal micro-organisms are then specialized microbes relying on anaerobic metabolism to grow without oxygen. Actually, a main part of the gut microbiota are strictly anaerobic, meaning that oxygen is harmful to them. With this anaerobic metabolism, the commensal microbiota consumes fibre-derivated sugars (e.g.. glucose and galactose) and produces short-chain fatty acids (SCFA) – mainly butyrate, acetate and propionate – that are absorbed by the host for its own metabolism. The main energetic source for the intestinal cells is butyrate, which is metabolized together with the oxygen carried to the intestine by the blood system. A virtuous cycle is then set up (see Figure 2a): the commensal microbiota produces butyrate that is metabolized by the host with oxygen; consequently, this oxygen does not diffuse to the lumen ensuring hypoxia and a favorable habitat for the butyrate-producing anaerobes. *Salmonella* is not very efficient in an anaerobic environment: the pathogen will have to hack this regulation mechanism, in order to create a favorable niche and permit the invasion of the gut. [4, 27]

Colonized gut. When arrived at the gut lumen, the pathogen releases a virulence factor (sipA) that triggers an inflammation in the epithelial cells (see Figure 2b). The host cells produce neutrophils: these immune cells are sent into the gut lumen where they trap any bacteria they encounter (pathogenic bacteria but also SCFA-producing symbionts). Then, the production of butyrate decreases, and this metabolite is no longer available for the epithelial cells: the oxygen reaching the cells is no longer metabolized and starts flowing in the gut lumen. This oxygen will be harmful for the butyrate-producing anaerobes, which initiates a vicious circle. The oxygen will also oxydize nutrients present in the gut, providing very efficient energetic sources for the pathogen alone, allowing it to take over from the commensal bacteria. Namely, galactose, glucose and thiosulfate will be oxydized into galactarate, glucarate end tetrathionate. In the meantime, inflammation induces the production of nitric oxyde, which is oxydated in nitrate, also very favorable for the pathogen [4, 27]. Figures sketching these mechanisms can be found in Fig. (2a-2b).



(a) Healthy gut at homeostasis: the colon lumen is hypoxic, so that commensal microbiota produces butyrate from sugars, which is consumed by the host with the blood-stream oxygen, regulating anaerobia.



(b) Salmonella colonization process: the pathogen triggers inflammation, decreasing commensals levels. Butyrate production drops down, reducing availability for the host. Epithelial cell metabolism switches from aerobic to anaerobic: blood-stream oxygen is no longer consumed and starts flowing in the gut lumen creating an aerobic niche for the pathogen.

Figure 2: Simplified illustrations recapitulating the biological regulation in an healthy gut, and *S.* Typhimurium colonization mechanisms.

We will first build a population dynamics model of *Salmonella* infection. The commensal microbiota will be represented by a unique strain of butyrate-producing bacteria: *Faecalibacterium Prauznitzii*. This bacteria belongs to one of the dominant genera in the gut microbiota, and is widely studied in the context of probiotic development [20]. The model proposed in this section is an adaptation of the works of Muñoz *et al.* [22], where they model the human colon by dividing it in compartments that are treated as continuous-flow stirred tank reactors (CSTR).

Our adaptation comes from a simplification of the colon into a single CSTR (called luminal compartment), and the novelty comes from the inclusion of FBA for computing the growth rate and the addition of the epithelial compartment representing the epithelium. We need to add the former to our set of equations in order to model the Salmonella infection. Our model aims to reproduce the main steps of the Salmonella infection: 1) the neutrophils (immune system cells that sequester bacteria) release into the colon from the epithelial compartment after the virulence factor triggers the inflammatory response; 2) the resulting drop of butyrate producing bacteria which entails decreased butyrate levels; 3) the metabolic switch, induced by the butyrate drop, of the epithelial cell from aerobic to anaerobic metabolism resulting in oxygen flow into the luminal compartment of the oxygen that is not consumed; 4) the bloom of Salmonella growing in this newly aerobic luminal compartment. A mathematical model describing the infection and the shift from an anaerobic to aerobic environment in the colon has been introduced in [16] at a larger scale. The model we introduce here focuses on the host-microbiota-pathogen metabolic interactions. Many parameters contained in ODE models are normally estimated by fitting the model to experimental data. In view of the lack of it, we will content ourselves to qualitatively representing the 4 steps introduced above that are hallmark of Salmonella infection [27], however, some parameters can be known before hand, such as the hydraulic retention time.

State variables. The model is a compartment model: a first compartment describes the gut lumen while the second stands for the epithelial cells. The luminal compartment describes the dynamics of the bacteria S_{th} and F_{prau}, for *Salmonella enterica* Typhimurium and *Faecalibacterium prauznitsii* populations, n_l, the luminal neutrophils, and m_l a vector containing all the metabolites concentrations of interest in the luminal compartment that describe the nutritional environment. Vector m_l is indexed by $i \in \{Gal, Gluc, NO, GalO, GlucO, NO_3, thio, tet, O_2, but\}$ standing for, respectively, luminal galactose, glucose, nitric oxyde, galactarate (i.e. oxydized galactose), glucarate (i.e. oxydized glucose), nitrate, thiosulfate, tetrationate (i.e. oxidized thiosulfate), oxygen and butyrate. F_{prau} consumes glucose and galactose and produces butyrate, whereas S_{th} consumes all of the metabolites except butyrate. The instant rate at which these are consumed or produced is given by the resolution of the FBA problem (see (2)) for each time step for each species. Nitric oxide has a special role in the host response to the pathogen, since it will react with oxygen to form nitrate which boosts the growth of S_{th} and gives an edge to *Salmonella* in the competition for resources [27].The epithelial compartment has 4 state variables: n_e, NO_e, O_{2_e} and but_e representing neutrophils, nitric oxide, oxygen and butyrate, respectively. Each of these state variables is transported to or from the luminal compartment, in order to model the host response effect in the colon to the pathogen invasion. The vector m_e indexed by $\{NO, O_2, but\}$ will gather the epithelial metabolites.

Luminal compartment. The gut lumen is modelled as an open system, meaning that matter flows through it. A working hypothesis is that the volume of the gut lumen is preserved at all times, meaning that a volume entering the gut must be balanced by a volume going out, thus the gut lumen can be modelled as a reactor [10]. The rate of change of the concentration of a component inside the gut lumen depends then on the difference between the input and output flow [22]. More precisely , let s be the concentration of a component of interest, then Q_{in} and Q_{out} be the volumetric input and output flow, s_{in} the concentration of the incoming flow, and V the reactor volume.

$$\partial_t s = \frac{Q_{in}s_{in} - Q_{out}s}{V} + \text{ biological and chemical reactions} + \text{ transport to epithelial compartment.}$$

Particularly, under the constant volume hypothesis $Q_{in} = Q_{out} = Q$. Define $D := \frac{Q}{V}$ as the dilution rate, which is the inverse of the hydraulic retention time. Then we can write $\frac{Q_{in}s_{in} - Q_{out}s}{V} = (s_{in} - s)D$.

Recall from equation (4) that $\mathcal{F}_s(c_s^{(up)})$ maps the upper bound of consumption to the uptake rates of metabolites for $s \in \{S_{th}, F_{prau}\}$. To couple Eq. (4) to the state equation, a relation between the state variable and the consumption upper bound c^{up} is needed. We then define

$$c_m^{(up)}(m_l) = \max\left\{\frac{m_{l,m}}{L_{dt}(S_{th}\mathbb{1}_{S_{th}}(m_l) + F_{prau}\mathbb{1}_{F_{prau}}(m_l)) + \varepsilon}, S_m\right\} \tag{21}$$

where $m_{l,m}$ is the substrate metabolite m of the luminal metabolites m_l, L_{dt} is a characteristic consumption time, $\mathbb{1}_s(m_l)$ is an indicator function indicating whether the bacteria s metabolizes the substrate m, ε is a small regularization parameter and S_m is the maximal substrate uptake when the metabolite m is at saturation in the media. As the upper bound $c_s^{(up)}$ now depends on vector m_l and bacterial densities, we will simply denote $\mathcal{F}_s(m_l, S_{th}, F_{prau})$ the uptake rates of metabolites for species s. Note that this vector also includes the biomass production rate, denoted by $\mathcal{F}_{s,1}(m_l, S_{th}, F_{prau})$. Analogously, vector $\mathcal{F}_{s,m_l}(m_l, S_{th}, F_{prau})$ is assembled from the uptake rates of metabolites in m_l. Finally, we introduce the diag$(\cdot)$ operator, which maps a vector of size n to the corresponding diagonal matrix of size n.

$$\partial_t S_{th} = (\mathcal{F}_{S_{th},1}(m_l, S_{th}, F_{prau}) - \rho n_l - D_{S_{th}})S_{th} \tag{22}$$

$$\partial_t F_{prau} = \left(\mathcal{F}_{F_{prau},1}(m_l, S_{th}, F_{prau}) - \rho n_l - \alpha\frac{O_{2_l}}{K_{O_2} + O_{2_l}} - D_{F_{prau}}\right) F_{prau} \tag{23}$$

$$\partial_t n_l = \gamma_n(n_e - n_l) - d_n n_l - D_n n_l \tag{24}$$

$$\begin{aligned}\partial_t m_l = {} & D(m_{in} - m_l) + \mathcal{F}_{S_{th},m_l}(m_l, S_{th}, F_{prau})S_{th} + \mathcal{F}_{F_{prau},m_l}(m_l, S_{th}, F_{prau})F_{prau} \\ & + \beta m_l O_{2_l} + \text{diag}(\gamma)T_r(m_e, m_l)\end{aligned} \tag{25}$$

where $\mathcal{F}_{S_{th}}$ (resp. $\mathcal{F}_{F_{prau}}$) is the FBA metabolic model of the pathogen (resp. the commensal). The parameter ρ represents the trapping by the neutrophils n_l. The term $\alpha\frac{O_{2_l}}{K_{O_2}+O_{2_l}}$ models the deleterious effect of the oxygen level O_2 on the obligate anaerobe F_{prau}, with a Michaelis-Menten dynamics using tuning parameters α and K_{O_2}. The terms $D_{S_{th}}$ and $D_{F_{prau}}$ indicate the passive dilution plus a bacteria specific death rate. The term $\gamma_n(n_e - n_l)$ represents the transfer process from the epithelial compartment. The term $d_n n_l$ is the death rate of neutrophils. Remark that mathematically we could have added the dilution rate (D_m) of neutrophils to its death rate d_n and have a single term, however since neutrophils also die in the epithelial compartment which has no dilution rate we decided to keep this explicit form. No entry of bacteria takes place, the bacteria getting into the system through initial conditions.

In equation (25), the first term describes the metabolite inflow, with m_{in} a vector containing the concentration in the small intestine of component m_l and D the passive dilution rate common to all the inert metabolites. The terms $\mathcal{F}_{b,m_l}(m_l, S_{th}, F_{prau})b$ for $b \in \{S_{th}, F_{prau}\}$ correspond to the consumption or production of metabolites due to the bacterial metabolism. The term $\beta m_l O_{2_l}$ corresponds to the oxidation reactions, where β is a diagonal matrix with entries only in the index corresponding to the reduced-oxidized pairs, each metabolite of a reduced-oxidized pair have the same coefficient, but with opposite sign, thus ensuring mass conservation. The term diag$(\gamma)T_r(m_e, m_l)$ shows the transport process to the epithelial compartment. We have for the transfer coefficient γ:

$$\text{diag}(\gamma)T_r(m_e, m_l)_i = \begin{cases}\gamma(m_{e,i} - m_{l,i}) & \text{if } i \in NO, O_2, but \\ 0 & \text{otherwise}\end{cases}$$

Epithelial compartment The 4 state variables of the epithelial compartment have the following dynamics

$$\partial_t n_e = C_{but,n} n_e \left(n_e - L_n\frac{but_e}{K_{but} + but_e}\right)(L_n - n_e) - d_n n_e + \gamma_n(n_m - n_e) + VF(S_{th}) \tag{26}$$

$$\begin{aligned}\partial_t NO_e = {} & C_{but,NO} NO_e \left(NO_e - L_{NO}\frac{but_e}{K_{but} + but_e}\right)(L_{NO} - NO_e) - d_{NO} NO_e \\ & + \gamma_{NO}(NO_l - NO_e) + VF(S_{th})\end{aligned} \tag{27}$$

$$\partial_t O_{2_e} = -\lambda_{but} but_e O_{2_e} - d_{O_2} O_{2_e} + L_{O_2} + \gamma(O_{2_l} - O_{2_e}) \tag{28}$$

$$\partial_t but_e = -\lambda_{but} but_e O_{2_e} + \gamma_{but}(but_m - but_e) \tag{29}$$

The term $C_{but,n} n_e \left(n_e - L_n\frac{but_e}{K_{but}+but_e}\right)(L_n - n_e)$ in equation (26) (and the analogue term in eq. (27)) is a bistable term with stable steady-state 0 and L_n, the threshold separating the attraction areas being $L_n\frac{but_e}{K_{but}+but_e}$. The threshold $\frac{but_e}{K_{but}+but_e}$ tends to 1 when butyrate is abundant and drops to zero

when butyrate level drops, pulling the state variable towards 0 or L_n when n_e exceeds this threshold. The term $VF(S_{th})$ is a Heaviside function in order to simulate the virulence factor that *Salmonella* secrets triggering neutrophils and the nitric oxide production. The terms $d_n n_e$, $d_{NO} NO_e$, and $d_{O_2} O_{2_e}$ in equations (26), (27), and (28), respectively, represent death terms. Terms $\gamma_n(n_m - n_e)$ in equation (26) (and all its analogues in other equations) model the transport process towards the luminal compartment, which couple these equations to Eq. (22)-(25). Finally terms $\lambda_{but} but_e O_{2_e}$ in both equations (28) and (29) model the epithelial cell metabolism mainly based on butyrate oxydation.

The system is supplemented with initial conditions Y_0 that can be found in Table A.2. The system was simulated in absence of *Salmonella* for 40 hours, time at which a pulse of *Salmonella* is added and models the initial invasion. The model is solved with custom python scripts (see Sec. A in the Annexe). The FBA models are taken from the literature: the S_{th} model is taken from [26] as provided by Cobrapy [9] while the F_{prau} model is taken from [31]. The parameter values can be found in Table A.1.

In Figure 3 a simulation of the system can be found. The abundance of S_{th}, F_{prau}, and neutrophils is first plotted (Fig. 3.a). Notice how the infection takes place at hour 40 and produces a spike of neutrophils in both the luminal (Fig. 3.a, dark green curve) and epithelial compartment (Fig. 3.e, dark green). After the immune response led by neutrophils we can observe the decline of F_{prau} and the rise of S_{th} achieving colonization. Plots of Fig. 3.b, Fig. 3.c and Fig. 3.d show the metabolite concentrations in time in the luminal compartment. Butyrate starts decreasing after S_{th} infection (Fig. 3.b, orange) because of the drop of F_{prau}, and eventually the media becomes completely aerobic after hour 60 (Fig. 3.b, blue). This can be explained by observing Fig. 3.e which illustrates how in the epithelial compartment the decreasing levels of butyrate allow oxygen to accumulate and flow into the luminal compartment (blue), as shown in Fig. 3.f (blue) plotting the flow between compartments, i.e. $\gamma(m_e - m_l)$. The same can be observed for nitric oxide (Fig. 3.f, green) which starts flowing into the luminal compartment from the beginning of the infection. The growth of S_{th} exhibits two phases (Fig. 3.a, red): a first phase is mainly fueled by the depletion of thiosulfate (Fig. 3.c, purple), while the second is more based on the consumption of oxidized molecules, allowed by the flow of oxygen, and nitrate coming from the oxidation of NO. We note that oxygen actually recycles the end product of the metabolism of the oxydized molecules, maintaining the favourable niche for *Salmonella*. We can see that the dynamical system renders all the four steps of *Salmonella* infection as described in the literature (see Fig. 2b): 1) the inflammation-induced raise of neutrophils 2) the consecutive drop of butyrate-producing bacteria and butyrate, 3) the switch to anaerobic metabolism in the epithelium and the resulting oxygenation of the lumen, favourable niche for 4) the bloom of Salmonella.

In the remainder, we will use the notation

$$Y^{ode} = (S_{th}, F_{prau}, n_l, m_l, n_e, NO_e, O_{2_e}, but_e)$$

to designate the vectorial state variable of the whole dynamical system.

4 Learning database definition

The assembling of the learning database is linked to the question of sampling the feature space of the RKHS method, which has dimension $N_{up} = 9$ in our application. Building a uniform sampling of a nine-dimensional hypercube necessitates a high number of points to cover all the volume of the hypercube. To mitigate the number of samples in the learning database, we adopt a supervised strategy: we sample the feature space in the neighbourhood of feature time-series observed during different solutions of the ODE system (22)-(29). In this way, the feature co-variance of our learning database is closer to the co-variance imposed by the dynamical system structure. We then compute $N_{sim} = 60$ repetitions of the ODE system (22)-(29) with random initial conditions sampled in uniform distributions (cf Table B.3 for parameter values), multiplied for the metabolites of the luminal compartment by a Bernoulli distribution simulating their presence/absence to also simulate cases where a metabolite is not initially present in the system.

From these $N_{sim} = 60$ replicates, we performed a time sampling of the state variables $m_l(i\Delta t)$, $S_{th}(i\Delta t)$ and $F_{prau}(i\Delta t)$, $i = 1, \cdots, N_t$ from which we computed the corresponding FBA constraints using formula (21) to get $\mathbf{X}_1$ after duplicate removal. The matrix $\mathbf{X}_1$ only contains constraints that have been observed during the time course of the system dynamics. To enrich the database around these orbits, we then perturbed $\mathbf{X}_1$ with a multiplicative Gaussian noise ($\sigma = 0.1$), and selected samples with resulting all negative constraints (i.e. substrate uptake) to get $\mathbf{X}_2$. The concatenation $\mathbf{X}_{large}$ of $\mathbf{X}_1$ and $\mathbf{X}_2$ leads to a database of $N_{obs} = 47942$ samples. We subsampled $\mathbf{X}_{large}$ by uniformly picking up 1000

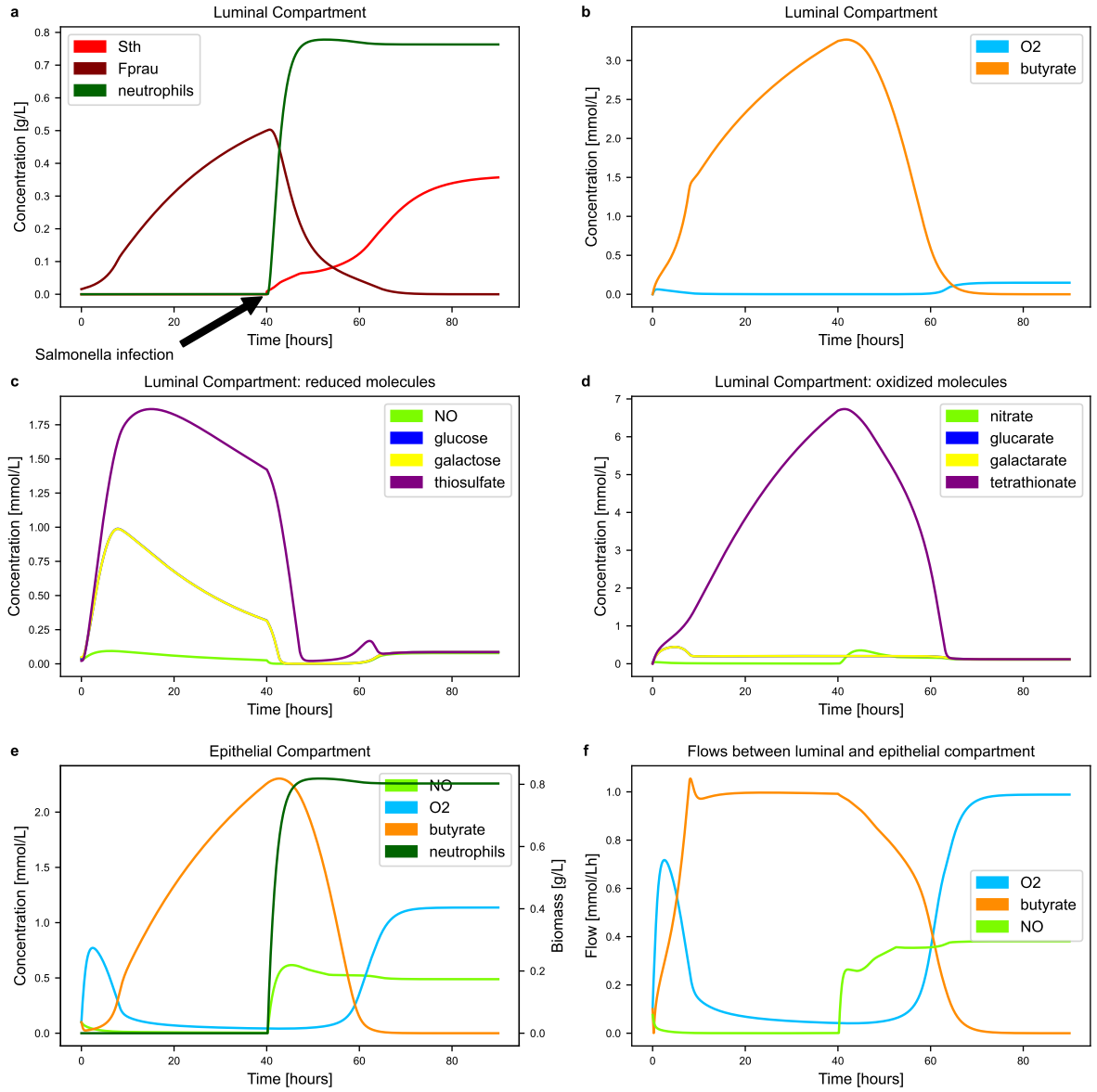


Figure 3: **dFBA model of *Salmonella* infection.** The output of the dFBA model of *Salmonella* infection is plotted. The fate of the different model components is displayed in the luminal and epithelial compartments. Butyrate and oxygen flows between epithelial and luminal compartments is also plotted. The results can be read as follows: Plot **a** shows the ecological dynamics, i.e. the abundance in time of the commensal microbiota and the pathogen. Neutrophils appear after the infection at hour 40, which affects negatively F_{prau} and allows the posterior S_{th} settlement. Plot **b** show how butyrate level drops after the infection, because of the decrease in the F_{prau} population and how the colon becomes aerobic after hour 60. Plots **c** and **d** show the dynamics of the reduced and oxidized metabolites. Notably thiosulfate accumulates until the infection moment, and then is consumed by S_{th}, the rising levels of nitrate are also linked to the presence of S_{th} and the available oxygen to transform nitric oxide in nitrate. Plot **e** shows the behaviour in the epithelial compartment and one can see how the butyrate level drops since the appearance of neutrophils, the oxygen accumulation because of the reduced butyrate levels and the nitric oxide increased explained by the presence of Salmonella that triggers its production. Plot f shows the flows between compartments to show that indeed the accumulations or depletion of metabolites described before is linked to exchanges between compartments. The different stages of the infection are then qualitatively recovered by the ODE system.

samples. Since it is particularly important from a biological point of view to capture the dynamics when a given metabolite is not limiting (i.e. when its concentration is close to S_m in eq. (21)) and when is nearly depleted (i.e. when its concentrations gets close to zero), we selected 1000 additional points by randomly taking $1000/(N_{sub} * 2)$ additional samples in the first and last decile of each columns of $\mathbf{X}_{large}$ to enrich the database in the distribution limits. We then finally obtained a learning database X with $N_{obs} = 2000$ samples. Model outputs $Y^{F_{prau}}$ and $Y^{S_{th}}$ were assembled for each species with the FBA model. The resulting distributions in $\mathbf{X}$ and $\mathbf{X}_{large}$ can be seen in B.9.

5 Hyperparameters selection

We now are ready to learn the metamodel, i.e. to solve (18) in order to find the parameters θ providing the best trade-off between $\mathbf{Y}$ reconstruction and RKHS subspace selection.

5.1 Selection of the group-lasso weight μ

For each species $s = S_{th}, F_{Prau}$ and model output j, we solve the problem (18) for

$$\mu \in \{0.0, .001, .01, .05, 0.075, .1, 0.15, .2, .3, .4, .5, .75, 1.0, 1.5\}$$

and a subsample of $N_{obs} = 400$ observations of $\mathbf{X}$ and $\mathbf{Y}^s$ and compute the loss $\mathcal{L}_{\mu,s,j}$, i.e. the relative reconstruction error on a testing set $(\mathbf{X}_{test}, \mathbf{Y}^s_{test})$ of $N_{obs} = 300$ unseen points of X

$$\mathcal{L}_{\mu,s,j} = \frac{\|\mathbf{Y}^s_{test,j} - \hat{\mathbf{Y}}^s_{test,j|\mu}\|_2}{\|\mathbf{Y}^s_{test}\|_2} \quad \text{where } \hat{\mathbf{Y}}^s_{test,j|\mu} = \hat{\mathcal{F}}_{s,j|\mu}(\mathbf{X}_{test}).$$

We display in Figures 4 and 5 the respective resulting lasso-paths for F_{prau} and S_{th}. Namely, we compute for each μ, species s and output j the norm $n^p_{\mu,s,j} = \|\hat{\theta}^s_{p,j|\mu}\|_2$ for $p \in \mathcal{P}$, where second order interactions only are considered in $\mathcal{P}$, and derive their relative contribution $n^p_{\mu,s,j} / \sum_{p\in\mathcal{P}} n^p_{\mu,s,j}$ that is displayed in Figures 4 and 5. This relative contribution allows to display the groups p of $\hat{\theta}^s_{p,j|\mu}$ that vanish for increasing μ, and the groups that remain non-null indicating input variables that are necessary to reconstruct the output j. In other words, for increasing μ, the group-lasso penalty becomes preponderant, turning off the parameters corresponding to the RKHS subspace p carrying the lower part of signal variance, which remains to perform variable selection. In the meantime, the loss tends to increase when a group of θ is discarded, since the signal is approximated in lower-dimensional subspaces. We are then seeking, for each output j, for the parameter μ providing the best trade-off between signal reconstruction and reduced number of selected groups p, synonym of reduced computational load and speed-up.

For F_{prau} (Fig. 4), we first observe that the lasso paths are very similar for the substrates (all the curves are similar in the glucose and galactose plots), indicating that these sugars have a comparable fate in the FBA model and similar influence on butyrate production (butyrate plots, the blue and orange plots are parallel). To predict the growth (F_{prau} plot), both sugars and their interaction are needed to achieve correct predictions (blue, orange and gray lines): the loss curve (dark blue line with stars) shows sharp increases when a group is dropped off. Due to the reduced number of substrates for F_{prau} ($N_{up} = 2$), all groups are kept for the four model outputs (see Table C.4 for selected μ).

For S_{th} (Fig. 5), input interactions are more complex. We first observe that O_2 intake (blue curve) is always preponderant for all model outputs plots, which is expected for this bacteria able to respire in aerobic environment. Again, glucose and galactose plots are very similar, such as glucarate and galactarate (their oxidated version). For these oxidated sugars, the loss increase (dark blue line with stars) is very limited when groups are dropped-off, indicating that the two groups that are kept (O_2,blue line, and galactarate, brown line) are enough for a correct signal reconstruction. The same kind of observation is made for the nitric oxyde, thiosulfate and tetrathionate plots. We next can see that O_2 and nitrate are badly reconstructed (O_2 and nitrate plots, dark blue line with stars), even with the whole set of subspaces (more than 30% loss). Finally, for S_{th} growth rate (S_{th} plot), we keep several groups of inputs, including O_2, thiosulfate, tetrathionate, glucarate and their interactions (see Table C.4 for selected μ).

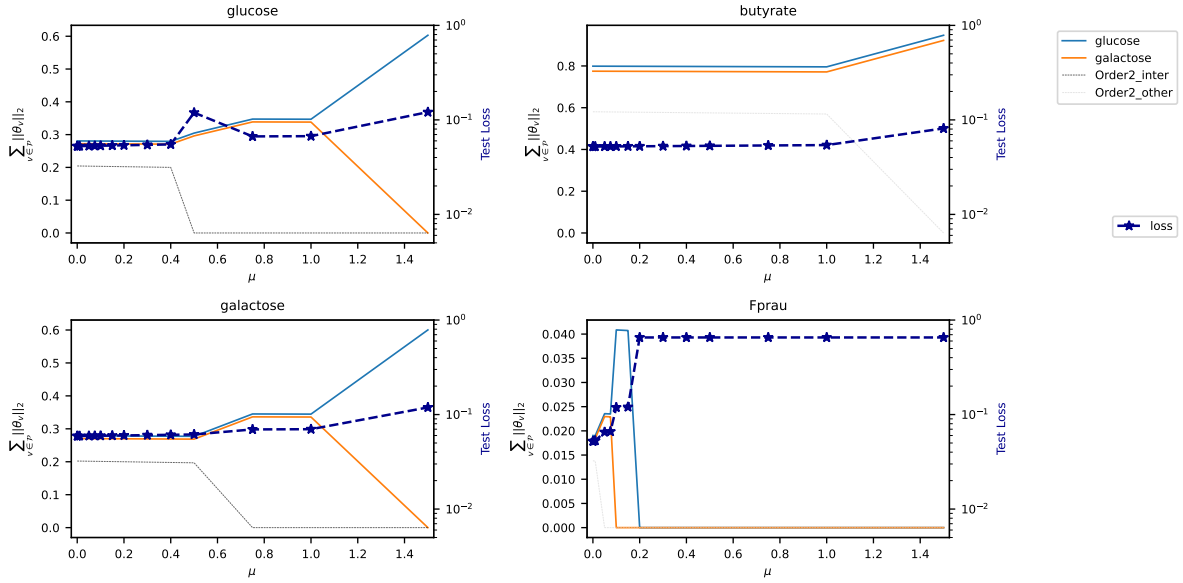


Figure 4: **Lasso path for F_{prau}.** For each metamodel, the lasso path is displayed: the relative contribution of the different blocks to the penalty is plotted for several values of the group lasso penalty μ, together with the loss function value. Namely, we plot $\|\hat{\theta}^s_{p,j|\mu}\|_2 / \sum_{p\in\mathcal{P}} \|\hat{\theta}^s_{p,j|\mu}\|_2$ wrt μ. For increasing μ, the group carrying less information vanish (i.e. its relative contribution goes to zero), indicating that the remaining groups support the main part of the signal. Dashed dark gray lines indicate order 2 interactions involving the displayed compound. Dashed light gray lines indicate order 2 interactions that do not involve the displayed compound (i.e. involving other compounds).

5.2 Selection of the number of functional basis

For given regularization parameters μ, different numbers of functional basis can be involved in the approximation, i.e. according to the Representer theorem 2 different numbers of samples included in the learning set. Again, a trade-off between reconstruction accuracy and computation speed is expected, since more functional basis enlarges the discretized functional space where the optimum is searched in eq. (11), allowing for better approximation, but at the cost of additional computations during each metamodel evaluation in (20).

For the μ previously selected, we then performed additional metamodel learnings for varying $N_{obs} \in \{50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700\}$. We then computed $n_{rep} = 5$ repetitions of the ODE system (22)-(29), for random initial conditions sampled with the same procedure than for the learning set construction (see Sec. 4), and for the FBA model or its metamodel approximation in eq. (22) to (25). The L_2 relative reconstruction error between the dFBA solutions Y^{ode}_{FBA} and their metamodel approximations $Y^{ode}_{mm|N_{obs}}$ is plotted in Fig. 6, together with the computation speed-up, i.e. the computation time ratio using the metamodel in place of the FBA model.

We can observe that the best trade-off between speed-up and reconstruction error is obtained for 500 functional basis. A higher number of basis increases the number of numerical operations and degrades the computation time while a lower number worsens the reconstruction error. More counter-intuitively, the speed-up is decreased for low numbers of functional basis ($N_{obs} \leq 100$). This is due to a higher number of blocks $p \in \mathcal{P}$ that are conserved when the number of observation in the learning basis (i.e. the number of functional basis in the RKHS) is reduced: the block-lasso penalty tends to conserve a higher number of blocks to preserve the data reconstruction, which is mechanically decreased for lower numbers of samples in the learning set.

6 Validation of the selected RKHS metamodel

The accuracy of the selected RKHS metamodel is first assessed by testing the metamodel with the corresponding FBA model on $n_{test} = 1500$ unseen points (Fig. 7a and 7b). We can see that the large majority of points lie in the vicinity of the line $y = x$, providing excellent $R2$ scores, with minimal value

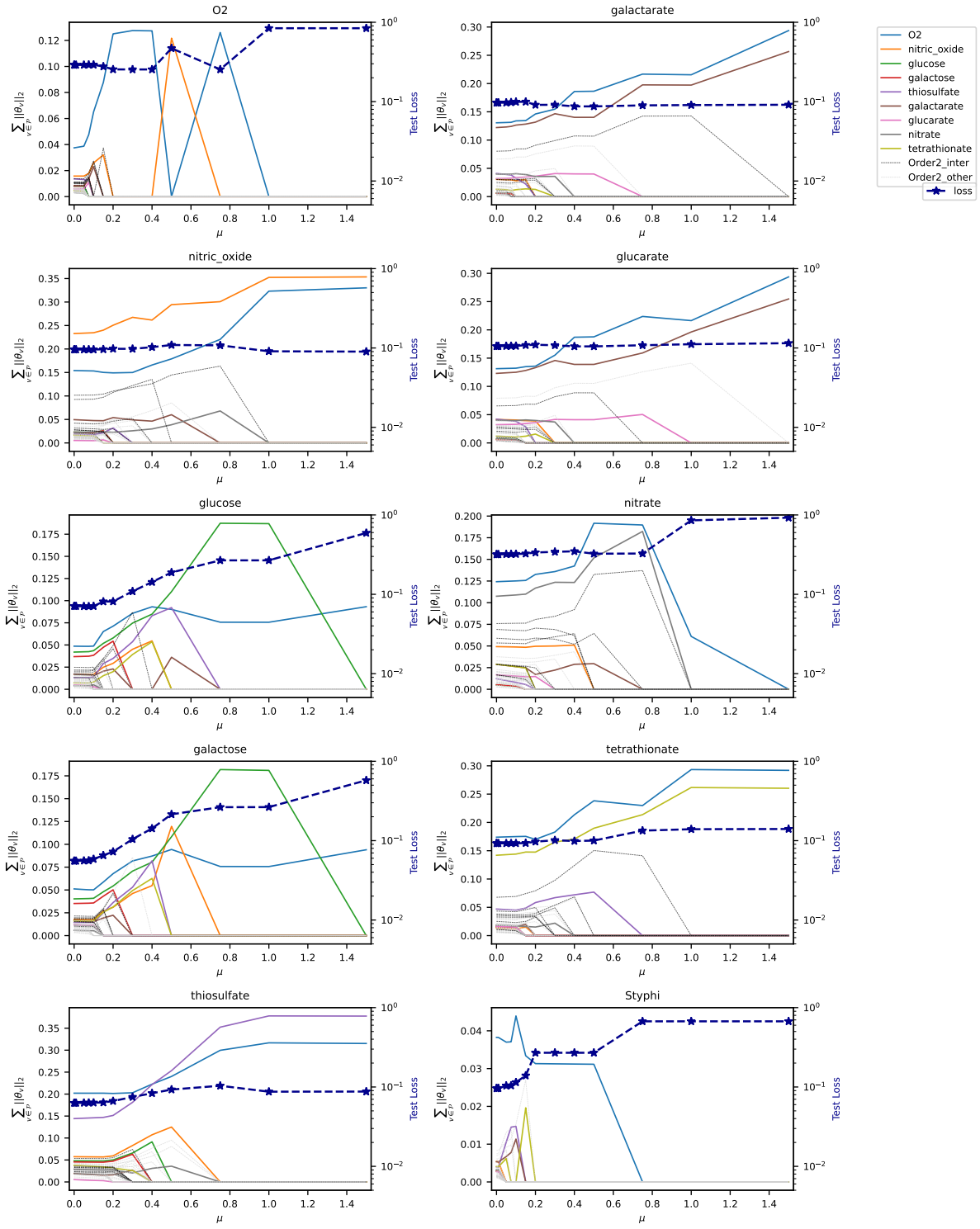


Figure 5: **Lasso path for S_{th}.** For each metamodel, the lasso path is displayed: the relative contribution of the different blocks to the penalty is plotted for several values of the group lasso penalty μ, together with the loss function value (dark blue line with stars). Namely, we plot $\|\hat{\theta}^s_{p,j|\mu}\|_2 / \sum_{p\in\mathcal{P}} \|\hat{\theta}^s_{p,j|\mu}\|_2$ wrt μ. For increasing μ, the groups carrying less information vanish (i.e. its relative contribution goes to zero), indicating that the remaining groups support the main part of the signal. Dashed dark gray lines indicate order 2 interactions involving the displayed compound. Dashed light gray lines indicate order 2 interactions that do not involve the displayed compound (i.e. involving other compounds).

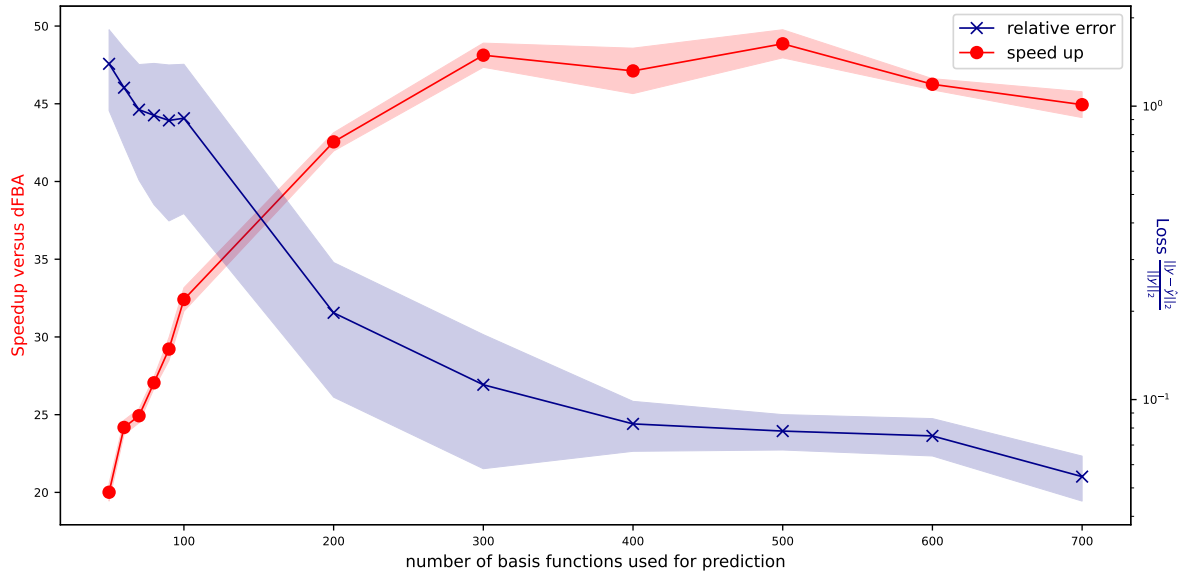


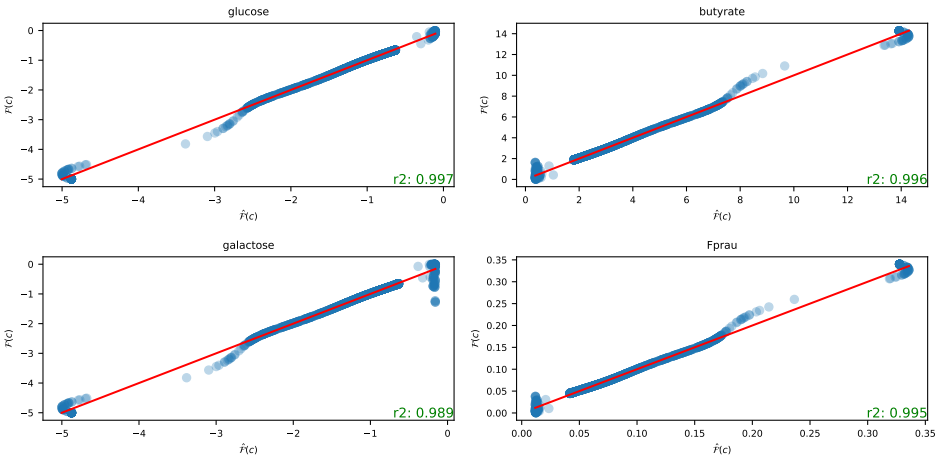
Figure 6: **Trade-off between speed-up and accuracy.** The average speed-up obtained by replacing the model by the metamodel in 5 repetitions is indicated for varying numbers of functional basis included in the ANOVA-RKHS (red line) with the standard deviation. The average relative reconstruction error $\frac{1}{N_{rep}}\sum_{r=1}^{N_{rep}} \frac{\|Y^{ode}_{FBA,r} - Y^{ode}_{mm,r|N_{obs}}\|}{\|Y^{ode}_{FBA,r}\|}$ is also displayed (blue line), for the Froebenius norm, together with the standard deviation.

of 0.922 for the worst reconstructed compound (nitrate for S_{th}).The worst approximation are mainly located near the boundaries of the domain, specially for F_{prau}. When looking at the FBA models responses for varying substrate constraints (Fig. C.10a and C.10b), we can see that the model is quasi-linear for sugar consumption for F_{prau}, but the behaviour is more complex for S_{th}, in particular for sugar consumption: sugar FBA uptake (y-axis) can vanish whereas glucose or galactose remain in the media (non-null constraints, x-axis) indicating metabolic switches. This behaviour is correctly predicted by the metamodel.

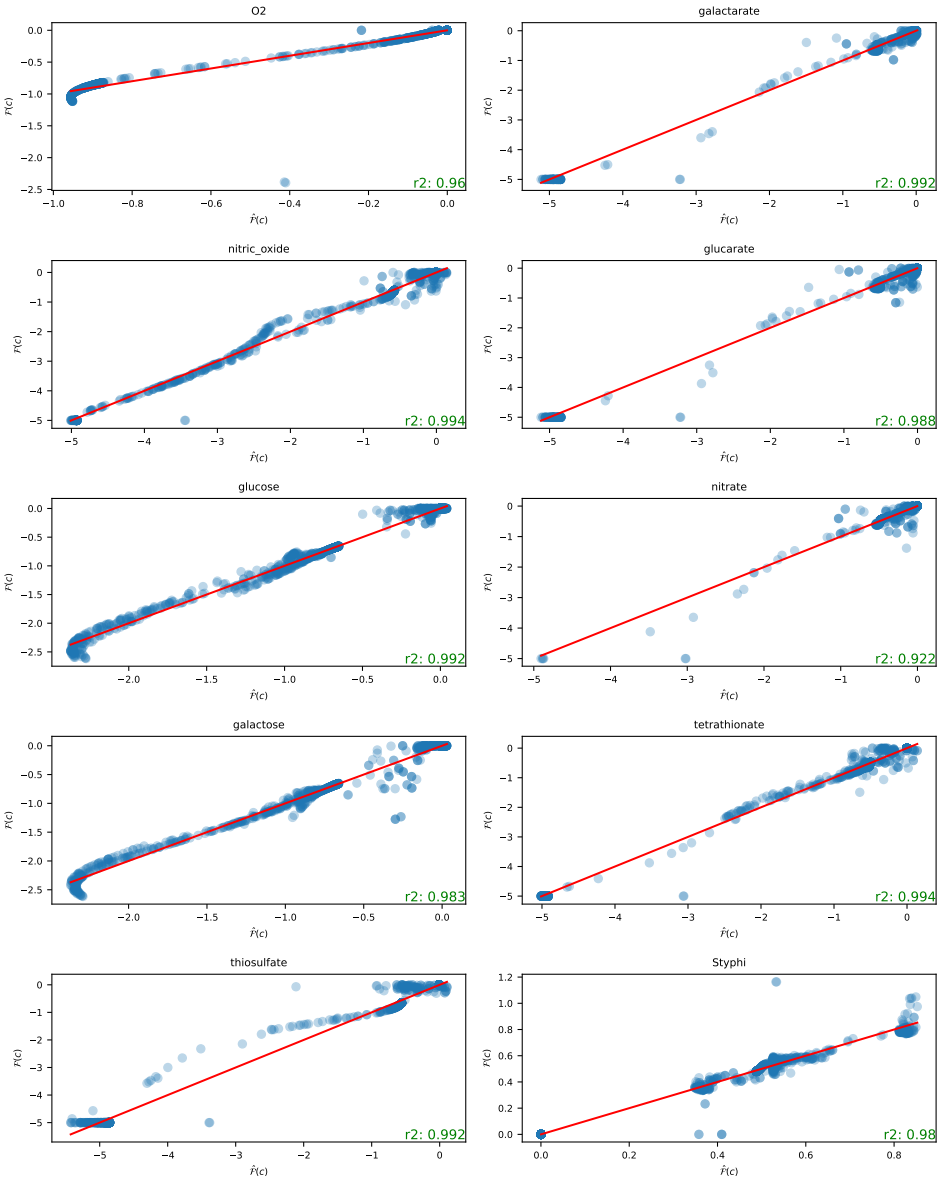
We then assess the metamodel approximation by comparing the ODE simulations with the FBA (plain lines) and the metamodel (dashed lines, Figure 8). Some limited discrepancies can be observed. In Fig. 8.a, *Salmonella* approximation accuracy is reduced in the second phase of growth, when S_{th} takes benefit of the micro-aerobic environment. In the same plot around hour 60, the metamodel is slightly off for F_{prau}, inducing a slight lag for butyrate production around $T = 60$ (Fig. 8.b, orange curves) which is reflected in the epithelial densities (Fig. 8.e, orange) and trans-epithelial flow (plot 6, orange).

For metabolites, the time courses are particularly well reconstructed, except for glucose after $T = 70h$ which goes awry, reflecting that there was little glucose consumption predicted by the metamodel, whereas in the original system it was completely consumed. Thiosulfate and tetrathionate are slightly off as well which might be linked with the oxygen lag observed in Fig. 8.e and f (blue lines). Less oxygen goes into the luminal compartment during the lag and the formation of tetrathionate by the oxidation of thiosulfate becomes impaired. This mechanism should be observed for other reduced-oxidized pairs, however since they are less abundant the effect might be attenuated.

Altogether, the behaviour of the metamodel is satisfactory in reproducing the dFBA system: it produces an overall reconstruction error $\|Y^{ode} - \hat{Y}^{ode}\|^2/\|Y^{ode}\|^2$ of $4,71\%$ and it accurately renders all different phases of S_{th} infection as observed in Fig. 3, such as F_{prau} and consecutive butyrate drop-off, O_2 and NO flows between epithelial and luminal compartments and the resulting two-phase growth of S_{th}. The metamodel furthermore allows computation speed-up by 45, which is a considerable gain.



(a) F_{prau}



(b) S_{ty}

Figure 7: **QQplot.** The FBA model value $\mathcal{F}(c)$ (y-axis) is plotted against its metamodel approximation $\hat{\mathcal{F}}(c)$ (x-axis) for 1600 unseen constraints c. The $r2$ score is indicated for each output

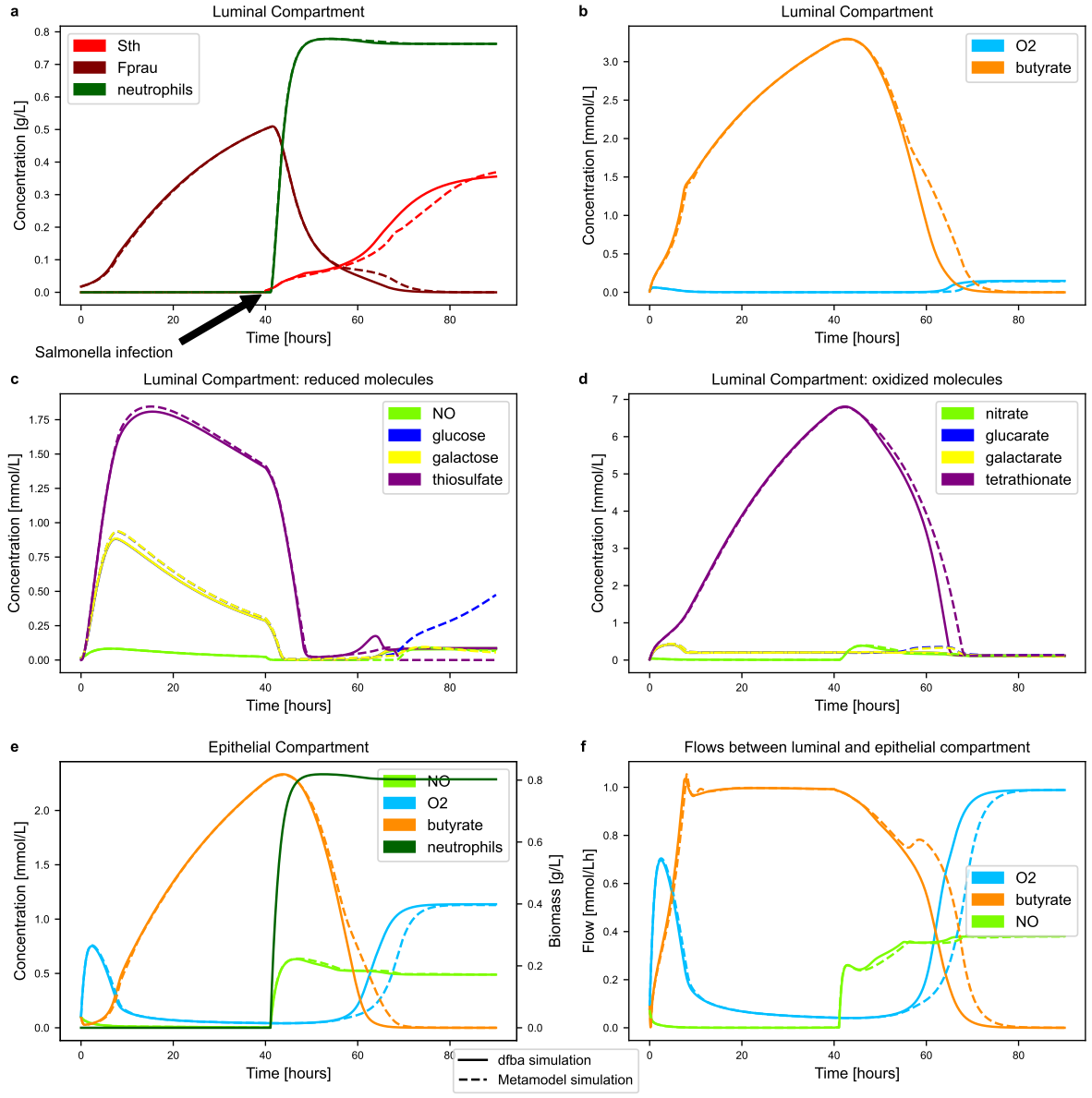


Figure 8: **DFBA and metamodel approximation.** The dFBA model is plotted (plain lines) together with its metamodel approximation, i.e. the ODE model output where the FBA model is replaced by its metamodel (dashed lines).

7 Discussion

7.1 Machine learning for accelerated computations of metabolic models.

An increasing number of studies [3, 18, 7] address the problem of modelling a community of micro-organisms by concatenating strain-level genome-scale metabolic models. If this strategy is well-established for well-mixed communities when one unique metabolic model can render the metabolic behaviour of the whole population of a specific strain discarding any spatial heterogeneities, it faces computational difficulties in contexts with important spatial structures: the metabolic model must be repeated at each spatial step, increasing linearly the computational load with the number of cells in the spatial mesh. This observation grounds the need for numerical accelerations of the metabolic model evaluations.

In this study, we adapted a machine learning method to the context of metabolic models, approximating the metabolic model output at reduced computational costs. We provided a proof-of-concept showing that RKHS-based metamodels are able to capture some non-linear effects exhibited by metabolic models (see Fig. C.10b), so that replacing the FBA metabolic model by its metamodels only marginally impacts the time-course of a system dynamics involving a metabolic model (Fig. 8). The metamodel drastically speeds up the overall time integration of the ODE system since integrating eq. (22)-(29) took in average 22 min and 27 s with the FBA model but only 28 s with the metamodel. We expect that this approximation remains valid in a PDE system.

The deployment of the RKHS method necessitates a careful selection of hyperparameters that strongly impacts the trade-off between accuracy and computation load. The block-lasso regularization penalty mitigates the number of blocks needed to provide accurate model reconstruction, which reduces the number of numerical operations during metamodel evaluation, thus speeding-up the overall computations. Likewise, the number of samples in the learning database is directly linked to the number of functional basis approximating each ANOVA-RKHS subspaces: if a higher number of observations increases the accuracy, it mechanically degrades the computation time. This tuning directly depends on the learning database and must be reproduced when the learning set is changed.

7.2 Learning dataset construction.

Metamodeling is specific in the framework of machine learning in that the learning dataset is not imposed to the user: the user keeps the hand on the assembly of the learning dataset. Ones can then search for sound experimental planning by placing the points of the learning set in strategic areas of the state space. One 'agnostic' approach consists in sampling uniformly hypercubes of the input space: after defining upper and lower bounds on the inputs, uniform sampling methods such as Latin Hypercube Sampling (LHS) or fast99 methods [25, 12, 28] can be deployed which provides suitable property for sensitivity analysis and computation of descriptive index such as Sobol Index. We opted for a more 'supervised' approach by sampling the feature space around time trajectories of the ODE system we want to approximate: several time integrations are performed based on random initial conditions which allows to compute FBA model inputs through eq. (21) that samples the feature space. The learning database was further enriched by randomly sampling around these trajectories, and by oversampling the borders of the hypercube (see fig. B.9).

Other strategies could be explored, by defining a generative statistical model of the points around the ODE trajectories. For example, ones could simulate these point clouds with copulas, by coupling uniform sampling of hypercubes with simulations of the empirical marginals of the observed points during the ODE time course.

7.3 Why using ANOVA-RKHS in our approach.

In this study, we opted for a specific RKHS method, based on ANOVA-RKHS. Unlike classical RKHS metamodel that approximates the model in a unique functional space through the Representer theroem 2, the ANOVA-RKHS method provides a theoretical metamodel the decomposition of which corresponds to its Hoeffding decomposition. The metamodel approximation with a penalized least square method enables the selection of the main effects and their interactions, leading to a more parsimonious metamodel. If this strategy is more complex from mathematical and computational points of view, it allows reducing the dimension of the input space by selecting the input variables that most influence the output variability. Besides the biological interpretations that can be done based on this input-output interactions or the Sobol index that are directly given by the ANOVA-RKHS method, variable selection also provides a better trade-off between reconstruction accuracy and computation load. Indeed, the fixed number of

samples in the learning dataset is more likely to cover the feature space with reduced dimensions. In our context, the feature space has 9 dimensions for S_{th}, and we could provide accurate predictions with 500 points. Working directly with classical 9-dimensional RKHS might have necessitated a higher number of training samples to provide the same accuracy. On the contrary, 500 points provides a good sampling of 1 or 2-dimensional feature spaces as observed in the f_p of eq. (17). Benchmarking ANOVA-RKHS with other RKHS and other machine learning methods is kept as a perspective for this work.

Additionally, ANOVA-RKHS could be compared or enriched with other functional spaces. In particular, as the response curves of the metabolic models are quite regular except near the origin (see Figs. C.10a and C.10a), other approximation methods could be investigated, such as polynomial regression models. This kind of models could provide faster evaluations by compensating a lower number of functional basis by higher priors on the response shape. Again, variable selection approaches could speed up metamodel evaluation on unseen points.

7.4 Exploring other regularization penalties.

In eq. (11), we selected a classical group lasso penalty to regularize the optimization problem. This penalty could be problematic in practice since it does not involve the ANOVA-RKHS norm, which is the norm that theoretically ensures the existence of a solution through the Representer theorem 2. However, these difficulties did not occur in the context of the computations presented here. Other regularizations were explored in [14, 13] and could be introduced in the future in our package. However, computing the ANOVA-RKHS norm involves the computation of the square root of large (N_{obs}^2) dense matrices (as many matrices as $card(\mathcal{P})$), which can be expensive in computational time and memory, specially if high-order interactions are considered in the Hoeffding decomposition. Hence, dimension reduction techniques or active learning could be coupled with the ANOVA-RKHS method to select at the same time input variables (with the ridge-group-sparse penalty introduced in [13]) and the most informative samples in the testing test.

8 Conclusion

In this study, we provided a proof-of-concept of the potentiality of machine learning methods to provide fast approximations of metabolic model outputs: these metamodels could replace FBA models in large systems biology models necessitating a massive number of FBA computations such as spatio-temporal models of microbial communities. We leveraged existing metamodeling methods (ANOVA-RKHS), provided strategies for the assembling of the testing dataset, set a framework for hyperparameter selection and assessed the accuracy of the metamodel. Replacing the original FBA models by their metamodel in an ODE system dynamics model of *Salmonella* infection in an healthy gut accelerated the computations by 45 with a relative error of about 5%. This result makes reachable PDE models of microbial communities involving genome-scale metabolic models such as FBA models, by approximating them with their metamodel.

9 Acknowledgments

Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr).

10 Funding

This study received fundings from Inria through the Exploratory Action SLIMMEST (for further information see https://www.inria.fr/en/slimmest). Simon Labarthe got support for this study from the France-Berkeley Fund through the project Articulate.

References

[1] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[2] Edwin H Battley. "The development of direct and indirect methods for the study of the thermodynamics of microbial growth". In: *Thermochimica Acta* 309.1-2 (1998), pp. 17–37.

[3] Eugen Bauer et al. "BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities". In: *PLoS computational biology* 13.5 (2017), e1005544.

[4] Andreas J Bäumler and Vanessa Sperandio. "Interactions between the microbiota and pathogenic bacteria in the gut". In: *Nature* 535.7610 (2016), pp. 85–93.

[5] Seth R Bordenstein and Kevin R Theis. "Host biology in light of the microbiome: ten principles of holobionts and hologenomes". In: *PLoS Biol* 13.8 (2015), e1002226.

[6] Felipe Cucker and Steve Smale. "On the mathematical foundations of learning". In: *Bulletin of the American mathematical society* 39.1 (2002), pp. 1–49.

[7] Ilija Dukovski et al. "A metabolic modeling platform for the computation of microbial ecosystems in time and space (COMETS)". In: *Nature protocols* 16.11 (2021), pp. 5030–5082.

[8] Nicolas Durrande et al. "ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis". In: *Journal of Multivariate Analysis* 115 (2013), pp. 57–67.

[9] Ali Ebrahim et al. "COBRApy: constraints-based reconstruction and analysis for python". In: *BMC systems biology* 7.1 (2013), pp. 1–6.

[10] Jean-Jacques Godon et al. "Overview of the oldest existing set of substrate-optimized anaerobic processes: digestive tracts". In: *BioEnergy Research* 6.3 (2013), pp. 1063–1081.

[11] Laurent Heirendt et al. "Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0". In: *Nature Protocols* 14.3 (2019), pp. 639–702. ISSN: 1754-2189. DOI: 10.1038/s41596-018-0098-2.

[12] Jon Herman and Will Usher. "SALib: an open-source Python library for sensitivity analysis". In: *Journal of Open Source Software* 2.9 (2017), p. 97.

[13] Sylvie Huet and Marie-Luce Taupin. "Metamodel construction for sensitivity analysis". In: *ESAIM: Proceedings and Surveys* 60 (2017), pp. 27–69.

[14] Halaleh Kamari, Sylvie Huet, and Marie-Luce Taupin. "RKHSMetaMod: An R package to estimate the Hoeffding decomposition of an unknown function by solving RKHS ridge group sparse optimization problem". In: *arXiv preprint arXiv:1905.13695* (2019).

[15] Simon Labarthe et al. "A mathematical model to investigate the key drivers of the biogeography of the colon microbiota". In: *Journal of theoretical biology* 462 (2019), pp. 552–581.

[16] Simon Labarthe et al. "A multi-scale epidemic model of salmonella infection with heterogeneous shedding". In: *ESAIM: Proceedings and Surveys* 67 (2020), pp. 261–284.

[17] Mireia Lopez-Siles et al. "Faecalibacterium prausnitzii: from microbiology to diagnostics and prognostics". In: *The ISME journal* 11.4 (2017), pp. 841–852.

[18] Stefanía Magnúsdóttir et al. "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota". In: *Nature biotechnology* 35.1 (2017), pp. 81–89.

[19] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle III. "Dynamic flux balance analysis of diauxic growth in Escherichia coli". In: *Biophysical journal* 83.3 (2002), pp. 1331–1340.

[20] Rebeca Martín et al. "Functional Characterization of Novel Faecalibacterium prausnitzii Strains Isolated from Healthy Volunteers: A Step Forward in the Use of F. prausnitzii as a Next-Generation Probiotic". In: *Frontiers in Microbiology* 8 (2017), p. 1226. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.01226.

[21] Arun S Moorthy et al. "A spatially continuous model of carbohydrate digestion and transport processes in the colon". In: *PloS one* 10.12 (2015), e0145309.

[22] Rafael Muñoz-Tamayo et al. "Mathematical modelling of carbohydrate degradation by human colonic microbiota". In: *Journal of theoretical biology* 266.1 (2010), pp. 189–201.

[23] Sean-Paul Nuccio and Andreas J Bäumler. "Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut". In: *MBio* 5.2 (2014), e00929–14.

[24] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. "What is flux balance analysis?" In: *Nature Biotechnology* 28.3 (2010), pp. 245–248. ISSN: 1087-0156. DOI: 10.1038/nbt.1614.

[25] Gilles Pujol et al. "The sensitivity Package". In: *R package version* 1 (2007), p. 878.

[26] Anu Raghunathan et al. "Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction". In: *BMC systems biology* 3.1 (2009), pp. 1–16.

[27] Fabian Rivera-Chávez and Andreas J Bäumler. "The pyromaniac inside you: Salmonella metabolism in the host gut". In: *Annual review of microbiology* 69 (2015), pp. 31–48.

[28] Andrea Saltelli, Stefano Tarantola, and KP-S Chan. "A quantitative model-independent method for global sensitivity analysis of model output". In: *Technometrics* 41.1 (1999), pp. 39–56.

[29] Jan Schellenberger et al. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0". In: *Nature protocols* 6.9 (2011), p. 1290.

[30] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. "A generalized representer theorem". In: *International conference on computational learning theory*. Springer. 2001, pp. 416–426.

[31] Ibrahim E El-Semman et al. "Genome-scale metabolic reconstructions of Bifidobacterium adolescentis L2-32 and Faecalibacterium prausnitzii A2-165 and their interaction". In: *BMC systems biology* 8.1 (2014), pp. 1–11.

[32] Jean-Christophe Simon et al. "Host-microbiota interactions: from holobiont theory to analysis". In: *Microbiome* 7.1 (2019), pp. 1–5.

[33] I.M. Sobol. "Sensitivity estimates for nonlinear mathematical models". In: *Math. Model. Comput. Exp* 1.4 (1993), pp. 407–414.

[34] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[35] Muriel Vayssier-Taussat et al. "Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics". In: *Frontiers in cellular and infection microbiology* 4 (2014), p. 29.

[36] Barbora Waclawiková et al. "Gut microbiota-motility interregulation: insights from in vivo, ex vivo and in silico studies". In: *Gut microbes* 14.1 (2022), p. 1997296.

[37] Stefanie Widder et al. "Challenges in microbial ecology: building predictive understanding of community function and dynamics". In: *The ISME journal* 10.11 (2016), pp. 2557–2568.

A model parameters and code availability

The system dynamics (22)-(29) is parametrized with the coefficients included in Table A.1 and initial conditions as indicated in Table A.2. The python code used for ODE system computation, and RKHS learning is available at https://gitlab.inria.fr/slimmest/cemracs_results.git together with a tutorial on a toy model.

The FBA models are taken from the literature: the S_{th} model is taken from [26] as provided by Cobrapy [9]. Metabolite names were modified to match with [23]. The F_{prau} model is taken from [31]. The metabolite IDs were also changed to keep consistent with the S_{th} model. Import reactions were further modified for consistency: all sugar exchange reactions of the original model were knock-out, and import reactions were allowed for sugars known to be metabolized by F_{prau} in the gut as described in [17].

B Learning database distribution

In this section, we indicate the parameters used for uniform sampling of the initial conditions of the 60 repetitions of the ODE system in the learning database definition in Table B.3. We then present the distribution of the whole database (60 repetitions that are sampled in time, and enriched with perturbed inputs observed during ODEs, see Sec. 4), and after sub-selection and enrichment near the boundaries in Fig. B.9.

C Model and metamodel responses

We present in this section the value of the regularization parameter μ and the metamodel response for selected μ compared with the FBA model response for a testing database of unseen points in Fig. C.10a and C.10b.

Parameter	Description	Units	Value [reference]
ρ	Death rate by unit of neutrophils	[1/day]	0.3
α	Maximum rate of oxygen's noxious effect on Fprau	[1/day]	0.2
K_s	Half saturation constant of oxygen's noxious effect on Fprau	[mmol/l]	0.1
γ_{O_2}	Transfer coefficient of oxygen between compartments	[1/day]	1
γ_{NO}	Transfer coefficient of nitric oxide between compartments	[1/day]	1
γ_{but}	Transfer coefficient of butyrate between compartments	[1/day]	1 [22]
γ_N	Transfer coefficient of neutrophils between compartments	[1/day]	1
β_s $s \in \{Gal, Gluc, thio, NO\}$	Coefficient for the rate of oxidation	[day $\cdot$ mmol/l] $^{-1}$	10
D_s $s\{Gal, Gluc, thio\}$	Influx of molecules to the luminal compartment	[mmol/l]/[day]	1/24
d_n	death rate of neutrophils	[1/day]	0.01
d_{NO}	degradation rate of NO_e in cells	[1/day]	0.01
d_{O_2}	degradation rate of O_{2_e} in cells	[1/day]	0.01
d_{but}	degradation rate of butyrate in cells	[1/day]	0.01
K_{but}	Half-saturation for the inhibition by butyrate	[mmol/l]	1.5
L_N	Source term of neutrophils in epithelium	[g/l]	0.1
L_{NO}	Source term of nitric oxide in epithelium	[mmol/l]	0.01
L_{O_2}	Source term of oxygen in epithelium	[mmol/l]	1

Table A.1: Values from literature are scarce. Most parameters were fitted manually and measuring their actual value is beyond the scope of this work. The work of Muñoz *et al.* [22] fitted some parameters such as the exchange rate for butyrate in the colon, so it was assumed as the value of the transfer coefficient of other products. Note particularly that parameter D represents the inverse of the hydraulic retention rate, which for a gut should be approximately 24 hours.

Parameter	Description	Units	Value [reference]
F_{prau}	*Faecalibacterium prauznitsii*	[g/l]	$1.56 \cdot 10^{-2}$
S_{th}	*Salmonella enterica Typhimurium*	[g/l]	0 at $t = 0$ and $8.64 \cdot 10^{-3}$ at $t = 40h$
m_{l,O_2}	Luminal oxygen	[mmol/l]	0
$m_{l,Gal}$	Luminal galactose	[mmol/l]	$7.6 \cdot 10^{-3}$
$m_{l,GalO}$	Luminal galactarate	[mmol/l]	$4.91 \cdot 10^{-2}$
$m_{l,Gluc}$	Luminal glucose	[mmol/l]	$2.00 \cdot 10^{-2}$
$m_{l,GlucO}$	Luminal glucarate	[mmol/l]	$4.02 \cdot 10^{-2}$
$m_{l,NO}$	Luminal nitric oxide	[mmol/l]	$2.45 \cdot 10^{-2}$
m_{l,NO_3}	Luminal nitrate	[mmol/l]	$3.10 \cdot 10^{-2}$
$m_{l,thio}$	Luminal thiosulfate	[mmol/l]	0
$m_{l,tet}$	Luminal tetrathionate	[mmol/l]	$2.19 \cdot 10^{-2}$
$m_{l,but}$	Luminal butyrate	[mmol/l]	0
n_l	Luminal neutrophils	[mmol/l]	0
n_e	Epithelial neutrophils	[mmol/l]	0
$m_{e,NO}$	Epithelial nitric oxide	[mmol/l]	0
m_{e,O_2}	Epithelial O_2	[mmol/l]	0
$m_{e,but}$	Epithelial butyrate	[mmol/l]	0

Table A.2: **Initial conditions**. Initial conditions have been sampled randomly as described in Sec. 4. The resulting sampling is given here that were used in Fig. 3 and 8.

State variable	lower bound	upper bound	Bernouilli parameter
F_{prau}	0	0.02	-
S_{th}	0	0.02	-
m_{l,O_2}	0.001	0.05	0.85
$m_{l,Gal}$	0.001	0.05	0.85
$m_{l,GalO}$	0.001	0.05	0.85
$m_{l,Gluc}$	0.001	0.05	0.85
$m_{l,GlucO}$	0.001	0.05	0.85
$m_{l,NO}$	0.001	0.05	0.85
m_{l,NO_3}	0.001	0.05	0.85
$m_{l,thio}$	0.001	0.05	0.85
$m_{l,tet}$	0.001	0.05	0.85
$m_{l,but}$	0.001	0.05	0.85
n_l	0	0	-
n_e	0	0	-
$m_{e,NO}$	0	0	-
m_{e,O_2}	0	0	-
$m_{e,but}$	0	0	-

Table B.3: **Parameter of the random functions describing the intial conditions of the 60 repetitions of the ODEs computed for the learning database**. The lower and upper bounds of the uniform distributions are indicated, together with the Bernouilli parameter that models the presence/absence of the metabolite at $t = 0$ when relevant.

Species	Glucose	Galactose	Butyrate	F_{prau}	Glucarate	Galactarate	NO	Nitrate	Thiosulf.	Tetrathin.	O2	S_{th}
F_{prau}	0.05	0.05	0.05	0.05	-	-	-	-	-	-	-	-
S_{th}	0.2	0.2	-	-	1.0	1.0	1.0	0.75	1.0	0.5	0.3	0.075

Table C.4: **Selected regularization parameter μ**. Selected hyperparameter μ that tunes the group-lasso penalty is indicated for each species (rows) and each model output (columns). This parameter provides the best trade-off between signal reconstruction and reduced number of RKHS subspace that are kept for reconstruction.

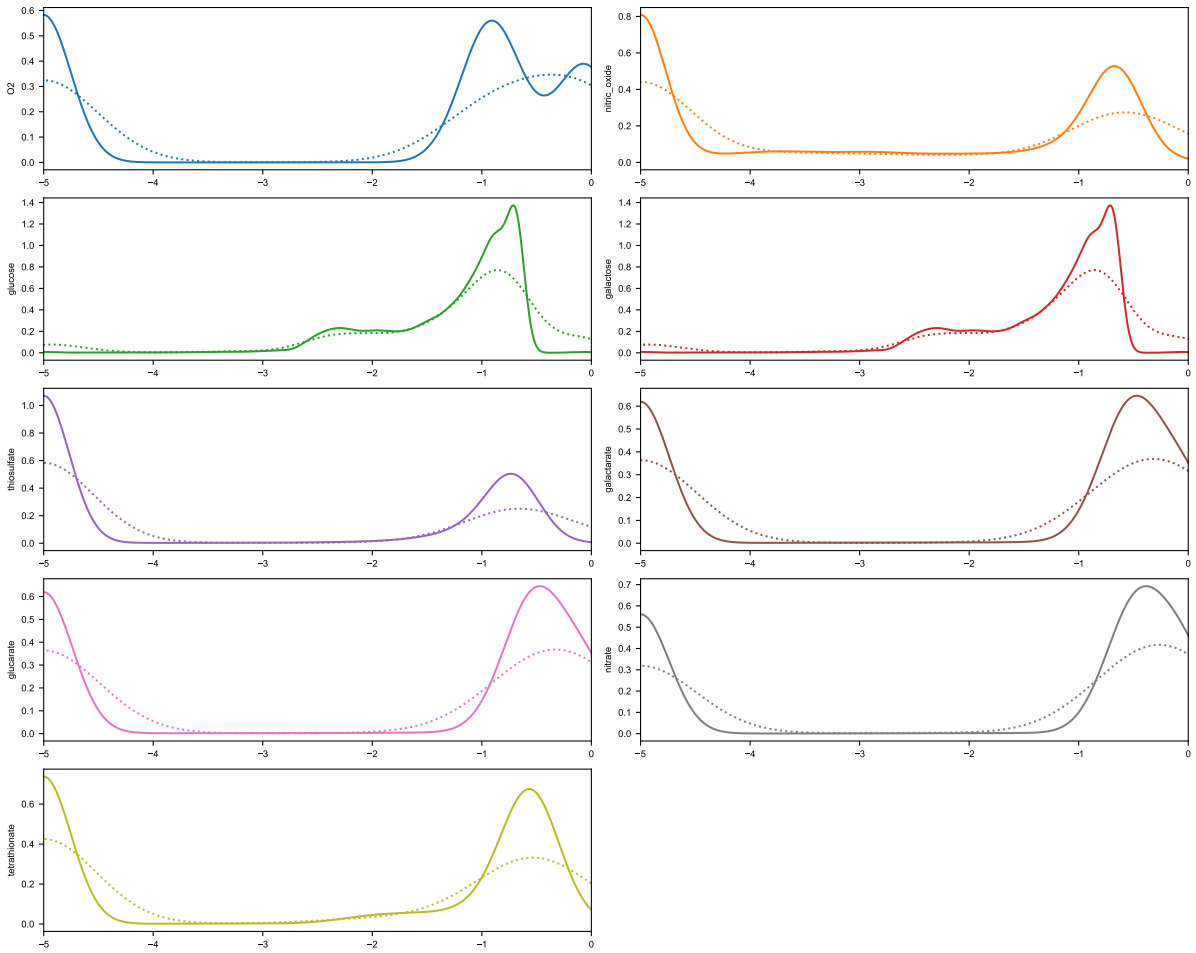
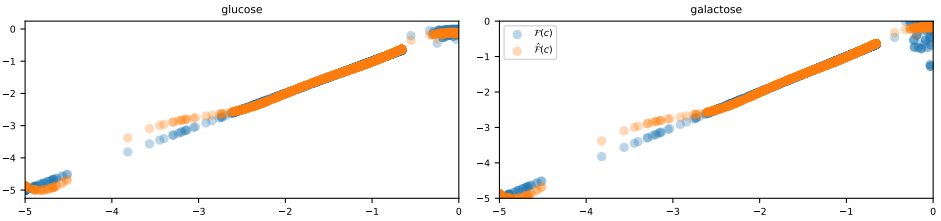
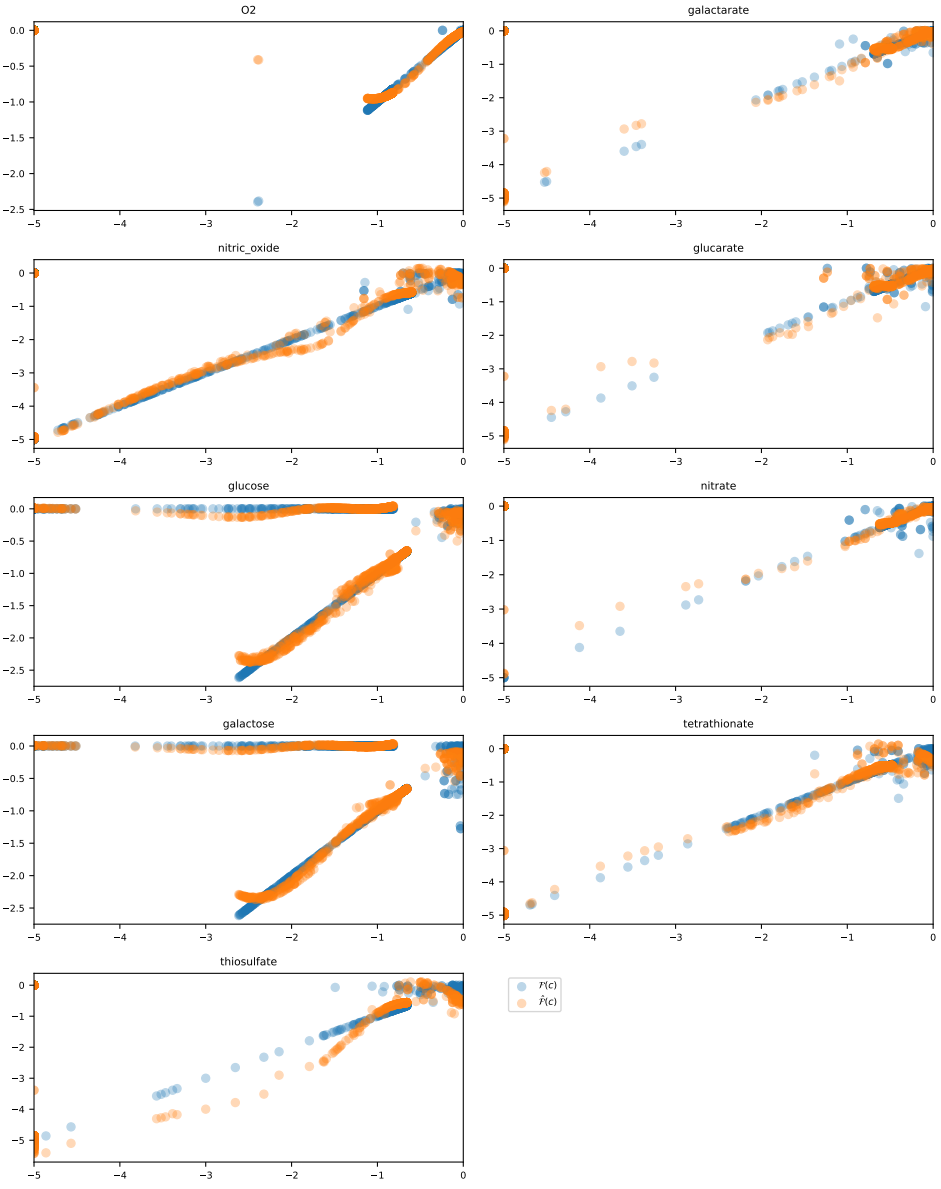


Figure B.9: **Marginal distributions in the learning database.** We display for each column $1 \leqslant c \leqslant N_{up}$ of the database $\mathbf{X}_{large}$ its marginal distribution (plain lines) together with the marginal distribution of $\mathbf{X}$ (dashed lines) obtained after subsampling and enrichment near the boundaries of $\mathbf{X}_{large}$. As expected, the main modes of $\mathbf{X}_{large}$ are conserved in $\mathbf{X}$, while points in the first and last deciles (near the boundaries) are over-represented by construction in $\mathbf{X}$.



(a) F_{prau}



(b) S_{ty}

Figure C.10: **Model response.** The FBA model value $\mathcal{F}(c)$ (blue dots) is plotted with its metamodel approximation $\hat{\mathcal{F}}(c)$ (orange dots, y-axis) for 1600 unseen constraints c (x-axis).