Biologically-Constrained Parameter Reduction for 5-Species Biofilm Model

Keisuke Nishioka

February 2026

Abstract

This document describes the Proposed Method, a parameter reduction technique for Bayesian estimation of the 5-species biofilm interaction model. By incorporating biological knowledge from experimentally determined interaction networks, the algorithm reduces the parameter space from 20 to 15 free parameters, improving estimation efficiency and biological interpretability.

Keywords: Bayesian Estimation, Biofilm, Parameter Reduction, TMCMC, Biological Constraints, Multi-species Interaction, Inverse Problem, Peri-implantitis, Uncertainty Quantification

Contents

1 Introduction

Understanding the dynamics of multi-species biofilms is crucial for the prevention and treatment of oral diseases. Heine et al. [1] investigated the interactions of five major oral bacterial species associated with peri-implantitis. Based on these experimental findings and the extended Hamilton principle proposed by Junker and Balzani [2], Klempt et al. [3] developed a continuum model for multi-species biofilms with a novel interaction scheme. Furthermore, Fritsch et al. [4] discussed Bayesian updating methods for bacterial microfilms under hybrid uncertainties using a novel surrogate model.

The 5-species biofilm model describes the dynamics of bacterial populations through an interaction matrix $\mathbf{A}$ and decay vector $\mathbf{b}$. However, the standard parameter estimation approach estimates all 20 parameters freely, which can lead to:

- Poor identifiability due to limited experimental data
- Biologically implausible parameter estimates
- Computational inefficiency from exploring unnecessary parameter space

The Proposed Method addresses these issues by constraining certain interaction parameters to zero based on experimental evidence of absent species interactions.

2 Biological Basis

2.1 Species in the Model

The model includes five bacterial species commonly found in oral biofilms:

ID	Species	Abbrev.	Role
0	*Streptococcus oralis*	S.o	Early colonizer
1	*Actinomyces naeslundii*	A.n	Early colonizer
2	*Veillonella* spp.	Vei	Metabolic bridge
3	*Fusobacterium nucleatum*	F.n	Bridge organism
4	*Porphyromonas gingivalis*	P.g	Late colonizer (pathogen)

Table 1: Species included in the 5-species biofilm model.

2.2 Interaction Network (Figure 4C)

Based on experimental observations [1], the following interaction network was established:

2.3 Active Interactions

The following species pairs have direct biological interactions:

2.4 Absent Interactions (Locked)

The following species pairs have no direct interaction according to experimental evidence (Figure 4C). These are locked to zero ($\theta_k = 0$) in the Proposed Method:

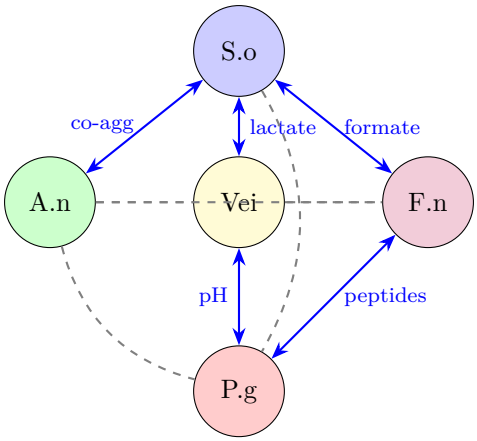

Figure 1: Species interaction network derived from Figure 4C. Solid blue arrows indicate active interactions (estimated parameters). Dashed gray lines indicate absent interactions (locked to zero).

Species Pair	**Mechanism**	**Type**
S. oralis ↔ A. naeslundii	Co-aggregation	Bidirectional
S. oralis ↔ Veillonella	Lactate production/consumption	Bidirectional
S. oralis ↔ F. nucleatum	Formate/Acetate symbiosis	Bidirectional
Veillonella ↔ P. gingivalis	pH rise support	Positive only
F. nucleatum ↔ P. gingivalis	Co-aggregation, peptides	Bidirectional

Table 2: Active species interactions with biological mechanisms.

3 Mathematical Formulation

3.1 Governing Equations

The 5-species biofilm model describes the dynamics of bacterial volume fractions ϕ_i and viability fractions ψ_i through a coupled ODE system. The interaction term for species i is:

$$I_i = \sum_{j=0}^{4} A_{ij}\phi_j\psi_j \tag{1}$$

where A_{ij} represents the effect of species j on species i, and $\phi_j\psi_j$ is the living bacteria volume fraction.

3.2 Symmetric Matrix Assumption

Critical assumption: The interaction matrix $\mathbf{A}$ is symmetric:

$$A_{ij} = A_{ji} \quad \forall i, j \in \{0, 1, 2, 3, 4\} \tag{2}$$

This reduces the number of off-diagonal interaction parameters from 20 to 10. For example, the lactate handover interaction between S. oralis (species 0) and Veillonella (species 2) is represented by a single parameter:

$$A_{02} = A_{20} = \theta_{10} \quad \text{(stored as } a_{13} \text{ in code)} \tag{3}$$

Index	Param	Species Pair	Matrix	Biological Reason
6	a_{34}	Vei (2) $\leftrightarrow$ F.n (3)	$A[2,3] = A[3,2]$	No direct metabolic pathway
12	a_{23}	A.n (1) $\leftrightarrow$ Vei (2)	$A[1,2] = A[2,1]$	No direct metabolic link
13	a_{24}	A.n (1) $\leftrightarrow$ F.n (3)	$A[1,3] = A[3,1]$	No direct interaction
16	a_{15}	S.o (0) $\leftrightarrow$ P.g (4)	$A[0,4] = A[4,0]$	No direct interaction
17	a_{25}	A.n (1) $\leftrightarrow$ P.g (4)	$A[1,4] = A[4,1]$	No direct interaction

Table 3: Absent interactions locked to zero in the Proposed Method. Numbers in parentheses are 0-indexed species IDs.

3.3 Parameter Vector Definition

The full 20-parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{19})^T$ is organized into five blocks corresponding to the model structure:

$$\begin{aligned}\boldsymbol{\theta} = &\underbrace{(a_{11}, a_{12}, a_{22}, b_1, b_2)}_{\text{M1: Species 1–2}} \oplus \underbrace{(a_{33}, a_{34}, a_{44}, b_3, b_4)}_{\text{M2: Species 3–4}} \\ &\oplus \underbrace{(a_{13}, a_{14}, a_{23}, a_{24})}_{\text{M3: Cross 1–2 vs 3–4}} \oplus \underbrace{(a_{55}, b_5)}_{\text{M4: Species 5}} \oplus \underbrace{(a_{15}, a_{25}, a_{35}, a_{45})}_{\text{M5: Cross with Species 5}}\end{aligned} \tag{4}$$

where a_{ij} denotes the interaction coefficient affecting species i from species j, and b_i is the decay rate of species i. Species are 1-indexed in notation (a_{ij}) but 0-indexed in code ($A[i-1, j-1]$).

3.4 Complete Parameter Mapping

Table 4 provides the authoritative mapping between parameter indices, matrix elements, and biological interpretation.

3.5 Locked Parameter Indices

The Proposed Method defines the set of locked indices based on absent biological interactions:

$$\mathcal{L} = \{6, 12, 13, 16, 17\} \tag{5}$$

For all $k \in \mathcal{L}$:

$$\theta_k = 0 \quad \text{(fixed, not estimated)} \tag{6}$$

3.6 Prior Bounds

The **base** prior distribution (for Commensal/Dysbiotic Static conditions) is:

$$\theta_k \sim \begin{cases} \text{Uniform}(0, 0) & \text{if } k \in \mathcal{L} \text{ (locked)} \\ \text{Uniform}(0, 1) & \text{if } k = 18 \text{ (Vei} \rightarrow \text{P.g, positive cooperation)} \\ \text{Uniform}(-1, 1) & \text{otherwise (free)} \end{cases} \tag{7}$$

Important: For the Dysbiotic HOBIC condition ("Surge" reproduction), the bounds for index 18 are modified to allow strong negative values:

$$\theta_{18} \sim \text{Uniform}(-3, -1) \quad \text{(Dysbiotic HOBIC only)} \tag{8}$$

This reflects the strong cooperative effect from Veillonella to P. gingivalis required to drive the pathogen surge.

Index	Name	Matrix Element	Species Pair	Biological Role	Status
0	a_{11}	$A[0,0]$	S.o self	Self-regulation	Free
1	a_{12}	$A[0,1] = A[1,0]$	S.o $\leftrightarrow$ A.n	Co-aggregation	Free
2	a_{22}	$A[1,1]$	A.n self	Self-regulation	Free
3	b_1	$b[0]$	S.o	Decay rate	Free
4	b_2	$b[1]$	A.n	Decay rate	Free
5	a_{33}	$A[2,2]$	Vei self	Self-regulation	Free
6	a_{34}	$A[2,3] = A[3,2]$	Vei $\leftrightarrow$ F.n	*No interaction*	**Locked**
7	a_{44}	$A[3,3]$	F.n self	Self-regulation	Free
8	b_3	$b[2]$	Vei	Decay rate	Free
9	b_4	$b[3]$	F.n	Decay rate	Free
10	a_{13}	$A[0,2] = A[2,0]$	S.o $\leftrightarrow$ Vei	**Lactate handover**	Free
11	a_{14}	$A[0,3] = A[3,0]$	S.o $\leftrightarrow$ F.n	Formate symbiosis	Free
12	a_{23}	$A[1,2] = A[2,1]$	A.n $\leftrightarrow$ Vei	*No interaction*	**Locked**
13	a_{24}	$A[1,3] = A[3,1]$	A.n $\leftrightarrow$ F.n	*No interaction*	**Locked**
14	a_{55}	$A[4,4]$	P.g self	Self-regulation	Free
15	b_5	$b[4]$	P.g	Decay rate	Free
16	a_{15}	$A[0,4] = A[4,0]$	S.o $\leftrightarrow$ P.g	*No interaction*	**Locked**
17	a_{25}	$A[1,4] = A[4,1]$	A.n $\leftrightarrow$ P.g	*No interaction*	**Locked**
18	a_{35}	$A[2,4] = A[4,2]$	Vei $\leftrightarrow$ P.g	**pH trigger**	Free*
19	a_{45}	$A[3,4] = A[4,3]$	F.n $\leftrightarrow$ P.g	Co-aggregation	Free

Table 4: Complete parameter mapping from θ vector to interaction matrix $\mathbf{A}$ and decay vector $\mathbf{b}$. Red rows indicate locked parameters ($\theta_k = 0$). *Index 18 bounds vary by condition (see Section 6).

3.7 Effective Parameter Space

The effective number of free parameters is:

$$n_{\text{free}} = 20 - |\mathcal{L}| = 20 - 5 = 15 \tag{9}$$

4 Bayesian Inference Framework

4.1 Forward Model

The forward model $\mathcal{M}(\boldsymbol{\theta})$ maps the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{20}$ to predicted species trajectories. Given initial conditions $\boldsymbol{\phi}(t_0)$ and $\boldsymbol{\psi}(t_0)$, the coupled ODE system derived from the extended Hamilton principle [2, 3] governs the evolution of the living volume fraction $\phi_i\psi_i$ for each species i:

$$\frac{d(\phi_i\psi_i)}{dt} = \left(A_{ii} + \sum_{j \neq i} A_{ij}\, \phi_j\psi_j \right) \phi_i\psi_i - b_i\, \phi_i\psi_i, \quad i = 0, \ldots, 4 \tag{10}$$

where the first term represents growth modulated by self-regulation (A_{ii}) and inter-species interactions (A_{ij}, $j \neq i$), while the second term accounts for species decay at rate b_i. The interaction matrix $\mathbf{A}$ and decay vector $\mathbf{b}$ are constructed from $\boldsymbol{\theta}$ via the mapping defined in Table 4. The forward model output is the predicted relative abundance vector at each observation time:

$$\hat{\mathbf{y}}(t_k; \boldsymbol{\theta}) = \mathcal{M}(\boldsymbol{\theta})\big|_{t=t_k}, \quad k = 1, \ldots, N_t \tag{11}$$

The system (10) represents a generalized Lotka–Volterra competition model with symmetric interactions, a structure that arises naturally from the variational formulation of Klempt et al. [3].

4.2 Bayesian Inverse Problem

The goal of Bayesian parameter estimation is to infer the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})$ of the model parameters $\boldsymbol{\theta}$ given the observed experimental data $\mathbf{y}_{\text{obs}} = \{y_{\text{obs},i}(t_k)\}_{i,k}$. By Bayes' theorem:

$$p(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}) = \frac{p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{y}_{\text{obs}})} \tag{12}$$

where $p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})$ is the likelihood function quantifying the data-model agreement, $p(\boldsymbol{\theta})$ is the prior distribution encoding biological constraints, and $p(\mathbf{y}_{\text{obs}}) = \int p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ is the model evidence (normalizing constant). The posterior distribution captures the full uncertainty in the parameter estimates, from which point estimates such as the Maximum A Posteriori (MAP) and posterior mean can be derived:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}), \qquad \hat{\boldsymbol{\theta}}_{\text{mean}} = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}}] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})\, d\boldsymbol{\theta} \tag{13}$$

4.3 Likelihood Function

Assuming independent Gaussian measurement errors across species and time points, the likelihood function takes the form:

$$p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}) = \prod_{k=1}^{N_t} \prod_{i=0}^{4} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{\left(y_{\text{obs},i}(t_k) - \hat{y}_i(t_k;\boldsymbol{\theta})\right)^2}{2\sigma_i^2} \right) \tag{14}$$

where $y_{\text{obs},i}(t_k)$ is the observed relative abundance of species i at time t_k, $\hat{y}_i(t_k;\boldsymbol{\theta})$ is the model prediction, and σ_i is the measurement noise standard deviation for species i. The corresponding log-likelihood is:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{k=1}^{N_t} \sum_{i=0}^{4} \left[\frac{\left(y_{\text{obs},i}(t_k) - \hat{y}_i(t_k;\boldsymbol{\theta})\right)^2}{\sigma_i^2} + \log(2\pi\sigma_i^2) \right] \tag{15}$$

In practice, σ_i may be estimated from replicate measurements or treated as a hyperparameter. If no replicate data are available, a common choice is to set σ_i equal to a fraction of the observed data range for species i, reflecting the expected measurement uncertainty.

4.4 Prior Distribution with Biological Constraints

The prior distribution $p(\boldsymbol{\theta})$ encodes both the biological constraints from the interaction network and the parameter locking mechanism. Assuming independence among the prior marginals:

$$p(\boldsymbol{\theta}) = \prod_{k=0}^{19} p(\theta_k) \tag{16}$$

where each marginal prior is:

$$p(\theta_k) = \begin{cases} \delta(\theta_k) & \text{if } k \in \mathcal{L} \quad \text{(Dirac delta: locked to zero)} \\ \dfrac{1}{u_k - l_k}\, \mathbf{1}_{[l_k,u_k]}(\theta_k) & \text{if } k \notin \mathcal{L} \quad \text{(uniform prior on free parameter)} \end{cases} \tag{17}$$

Here $\delta(\cdot)$ denotes the Dirac delta function enforcing $\theta_k = 0$ for locked parameters, and $\mathbf{1}_{[l_k,u_k]}$ is the indicator function on $[l_k, u_k]$. This formulation is equivalent to restricting the posterior to the constrained subspace:

$$\Theta_{\mathcal{L}} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{20} : \theta_k = 0 \ \ \forall\, k \in \mathcal{L} \right\} \tag{18}$$

In practice, the inference is performed over the reduced vector $\boldsymbol{\theta}_{\text{free}} \in \mathbb{R}^{n_{\text{free}}}$ containing only the free parameters, while locked parameters remain fixed at zero throughout. The dimension reduction from $\mathbb{R}^{20}$ to $\mathbb{R}^{n_{\text{free}}}$ directly improves the sampling efficiency of MCMC methods, as the mixing time of Markov chains generally increases with dimensionality [5].

5 Transitional Markov Chain Monte Carlo (TMCMC)

5.1 Algorithm Overview

The Transitional Markov Chain Monte Carlo (TMCMC) algorithm, introduced by Ching and Chen [6], is a sequential Monte Carlo method designed for sampling from complex, potentially multimodal posterior distributions. Unlike standard MCMC methods (e.g., Metropolis–Hastings, Gibbs sampling) that may suffer from poor mixing in high-dimensional or multimodal spaces, TMCMC progressively transforms samples from the prior to the posterior through a sequence of intermediate "tempered" distributions [7].

The key idea is to define a tempering schedule $0 = \beta_0 < \beta_1 < \cdots < \beta_M = 1$ and construct a sequence of intermediate distributions:

$$p_m(\boldsymbol{\theta}) \propto p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})^{\beta_m} \, p(\boldsymbol{\theta}), \quad m = 0, 1, \ldots, M \tag{19}$$

At $\beta_0 = 0$, $p_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ reduces to the prior, and at $\beta_M = 1$, $p_M(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y}_{\text{obs}})$ recovers the full posterior. By introducing the likelihood gradually, the algorithm avoids the "prior–posterior gap" that causes standard importance sampling to fail when the prior and posterior are far apart.

5.2 Adaptive Tempering Schedule

At each stage m, the next tempering parameter β_{m+1} is selected adaptively to control the degeneracy of the importance weights. Following Betz et al. [8], β_{m+1} is chosen such that the coefficient of variation (CoV) of the importance weights satisfies:

$$\text{CoV}\Big[\{w_j^{(m)}\}_{j=1}^N\Big] = \frac{\text{Std}[w_j^{(m)}]}{\text{Mean}[w_j^{(m)}]} = \delta_{\text{target}} \tag{20}$$

where the importance weights are computed as:

$$w_j^{(m)} = p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}_j^{(m)})^{\beta_{m+1} - \beta_m}, \quad j = 1, \ldots, N \tag{21}$$

and $\delta_{\text{target}} \in (0, 2]$ is a user-specified target (typically $\delta_{\text{target}} = 1.0$). Equation (20) is solved for β_{m+1} via bisection on $(\beta_m, 1]$. This adaptive scheme avoids the need for a predetermined number of stages M and ensures smooth transitions between intermediate distributions.

5.3 Resampling and MCMC Mutation

At each stage m, the algorithm proceeds through three steps:

1. **Resampling**: Draw N samples from the current population $\{\boldsymbol{\theta}_j^{(m)}\}_{j=1}^N$ with probabilities proportional to the normalized importance weights $\bar{w}_j^{(m)} = w_j^{(m)} / \sum_{l=1}^N w_l^{(m)}$, using multinomial or systematic resampling.

2. **Covariance estimation**: Compute the weighted sample covariance matrix:

$$\boldsymbol{\Sigma}^{(m)} = \sum_{j=1}^{N} \bar{w}_j^{(m)} \left(\boldsymbol{\theta}_j^{(m)} - \bar{\boldsymbol{\theta}}^{(m)}\right)\left(\boldsymbol{\theta}_j^{(m)} - \bar{\boldsymbol{\theta}}^{(m)}\right)^T \tag{22}$$

   where $\bar{\boldsymbol{\theta}}^{(m)} = \sum_{j=1}^N \bar{w}_j^{(m)} \boldsymbol{\theta}_j^{(m)}$ is the weighted mean. This covariance adaptively scales the MCMC proposal to the local geometry of the tempered posterior.

3. **MCMC mutation**: Each resampled particle undergoes one or more Metropolis–Hastings steps with a Gaussian proposal:

$$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_j) = \mathcal{N}\Big(\boldsymbol{\theta}_j,\ \gamma^2 \boldsymbol{\Sigma}^{(m)}\Big) \tag{23}$$

where $\gamma > 0$ is a scaling factor (typically $\gamma^2 = 0.04$). The acceptance probability is:

$$\alpha = \min\left(1,\ \frac{p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}^*)^{\beta_{m+1}}\, p(\boldsymbol{\theta}^*)}{p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}_j)^{\beta_{m+1}}\, p(\boldsymbol{\theta}_j)}\right) \tag{24}$$

This mutation step diversifies the particle population and prevents sample impoverishment.

5.4 TMCMC Procedure

The complete TMCMC procedure with biological constraint enforcement is summarized in Algorithm 2.

Algorithm 1: Transitional Markov Chain Monte Carlo (TMCMC)

Input: Prior $p(\boldsymbol{\theta})$, likelihood $p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta})$, number of particles N, locked set $\mathcal{L}$, target CoV δ_{target}

Output: Weighted posterior samples $\{\boldsymbol{\theta}_j^{(M)}\}_{j=1}^N$

1. **Initialize**: Draw $\{\boldsymbol{\theta}_j^{(0)}\}_{j=1}^N \sim p(\boldsymbol{\theta})$; set $\beta_0 = 0$, $m = 0$
2. **While** $\beta_m < 1$:
   (a) Find $\beta_{m+1} \in (\beta_m, 1]$ such that $\text{CoV}[\{w_j^{(m)}\}] = \delta_{\text{target}}$ via bisection
   (b) Compute weights $w_j^{(m)} = p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}_j^{(m)})^{\beta_{m+1}-\beta_m}$ for $j = 1, \ldots, N$
   (c) Compute weighted covariance $\boldsymbol{\Sigma}^{(m)}$ from current samples
   (d) Resample N particles from $\{\boldsymbol{\theta}_j^{(m)}\}$ with probabilities $\propto w_j^{(m)}$
   (e) **For each** resampled particle $j = 1, \ldots, N$:
      i. Propose $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\theta}_j, \gamma^2 \boldsymbol{\Sigma}^{(m)})$
      ii. Enforce constraints: set $\theta_k^* = 0$ for $k \in \mathcal{L}$; clip to prior bounds $[l_k, u_k]$
      iii. Accept $\boldsymbol{\theta}_j^{(m+1)} = \boldsymbol{\theta}^*$ with probability α (Eq. 24); otherwise retain $\boldsymbol{\theta}_j^{(m+1)} = \boldsymbol{\theta}_j$
   (f) $m \leftarrow m + 1$
3. **Return** posterior samples $\{\boldsymbol{\theta}_j^{(M)}\}_{j=1}^N$

Figure 2: TMCMC algorithm with biological constraint enforcement. Step 2(e)(ii) ensures that locked parameters remain at zero and all free parameters respect their condition-specific prior bounds throughout the sampling process.

5.5 Model Evidence Estimation

A valuable byproduct of TMCMC is an unbiased estimate of the model evidence $p(\mathbf{y}_{\text{obs}})$, computed as:

$$\hat{p}(\mathbf{y}_{\text{obs}}) = \prod_{m=0}^{M-1} \left(\frac{1}{N} \sum_{j=1}^{N} w_j^{(m)} \right) \tag{25}$$

This quantity enables Bayesian model comparison between the standard (20-parameter) and proposed (15-parameter) formulations via the Bayes factor:

$$B_{10} = \frac{p(\mathbf{y}_{\text{obs}} \mid \mathcal{M}_{\text{proposed}})}{p(\mathbf{y}_{\text{obs}} \mid \mathcal{M}_{\text{standard}})} \tag{26}$$

A Bayes factor $B_{10} > 1$ provides evidence in favor of the proposed reduced model, indicating that the biological constraints improve not only interpretability but also predictive performance relative to model complexity. According to Kass and Raftery's scale, $B_{10} > 3$ constitutes "substantial" evidence and $B_{10} > 20$ "strong" evidence for the reduced model.

6 Experiment Conditions & Parameter Estimation

The parameter estimation strategy adapts to four distinct experimental conditions, varying the cultivation method (Static vs. HOBIC) and the community state (Commensal vs. Dysbiotic). Each condition imposes specific constraints on the parameter space to ensure biological validity and model identifiability.

6.1 Parameter Locking Rules

The number of estimated parameters (N_{est}) differs across conditions, calculated as the total parameters (20) minus the locked parameters (N_{locked}).

Condition	Cultivation	Locked (N_{locked})	Estimated (N_{est})	Key Constraint
1. Commensal	Static	9	11	Match data (Zero interactions)
2. Dysbiotic	Static	5	15	Estimate Pathogen interactions
3. Commensal	HOBIC	8	12	Match data (Zero interactions)
4. Dysbiotic	HOBIC	0	20	**Unlock All (Surge Reproduction)**

Table 5: Parameter estimation counts for each experimental condition.

6.2 Detailed Locking Logic

1. **Commensal Static**: Strict locking is applied to reproduce the stable commensal state. In addition to the standard structural locks (5), growth rates for late colonizers (Red/Purple) and their interactions with early colonizers are locked to zero ($N_{locked} = 9$).

   > "Based on qPCR data showing *P. gingivalis* and *F. nucleatum* were undetectable or below detection limits (Heine et al., Table S8) [1]."

2. **Dysbiotic Static**: Represents the transition to a pathogen-rich state. Locking is relaxed to allow estimation of key pathogen growth and interaction parameters, maintaining only the structural locks ($N_{locked} = 5$).

   > "Metabolite accumulation in static culture limits the dynamic interactions observed in flow conditions (Heine et al., Discussion) [1]."

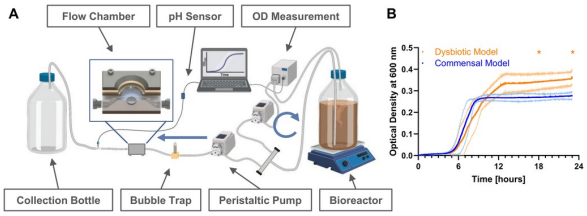

Figure 3: **1. Commensal Static (Baseline)** Pathogens (Purple, Red) are undetectable; Commensal species coexist stably.

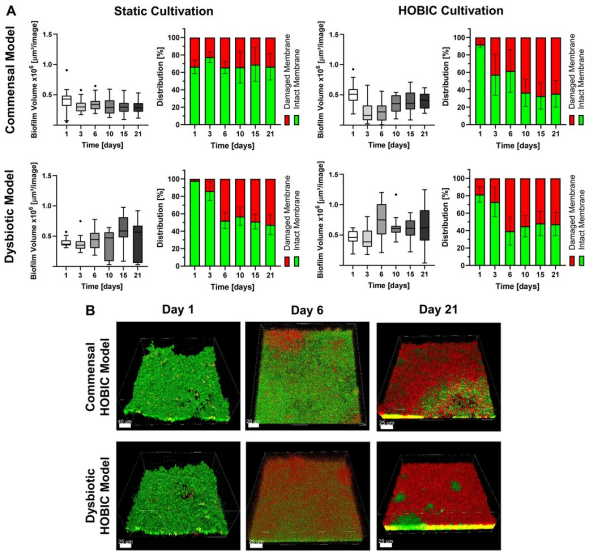

Figure 4: **2. Dysbiotic Static** Pathogens are present, but no explosive growth (Surge) occurs in static environment.

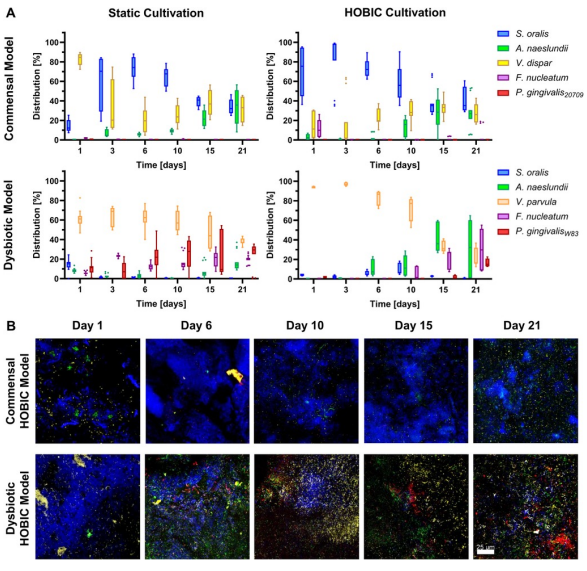

Figure 5: **3. Commensal HOBIC** S. oralis (Blue) increases specifically under high flow (Blue Bloom). Pathogens are suppressed.

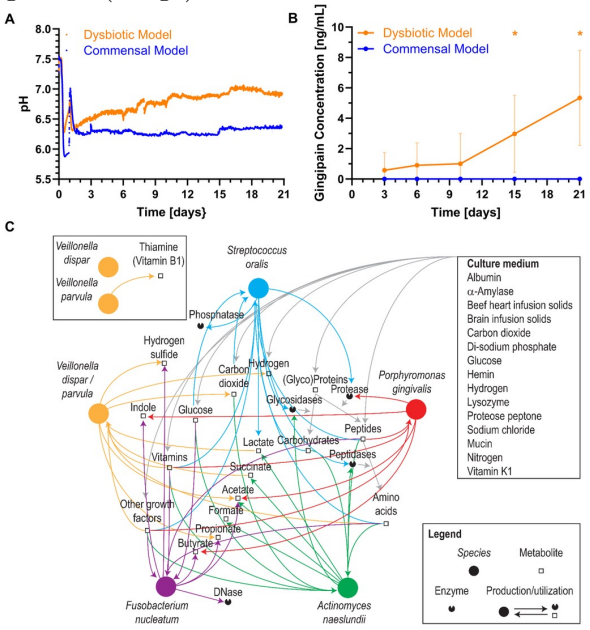

Figure 6: **4. Dysbiotic HOBIC (Surge)** V. parvula (Orange) and P. gingivalis (Red) grow explosively in symbiosis (Discovery Mode).

Figure 7: Experimental observation data [1]. Time course of species composition in each condition (Day 1–21).

3. **Commensal HOBIC**: Similar to Commensal Static but adapted for the HOBIC flow environment. Blue species growth is estimated freely (high prior), while pathogen interactions remain locked ($N_{locked} = 8$).

4. **Dysbiotic HOBIC (The "Surge" Model)**: This is the critical validation case. **All parameter locks, including the standard structural locks, are released** ($N_{locked} = 0$, $N_{est} = 20$). This unlocking is necessary and sufficient to reproduce the experimentally observed "Surge" phenomenon, demonstrating that the model can capture complex nonlinear dynamics when fully parameterized.

   > "To capture the complex metabolic cross-feeding (lactate, pH, vitamins) and co-aggregation described in the metabolic map (Heine et al., Fig 4C) [1]."

   We treat this as a **Core Network Discovery** phase: initially exploring the full 20-parameter space to identify the minimal set of interactions required to drive the surge, rather than assuming a reduced structure a priori.

6.3 4-Stage Sequential Estimation

Parameter estimation is performed sequentially in 4 stages. This configuration has been optimized considering parameter correlations (coupling) and search space dimensionality.

Stage	Target	# Params	Parameters
1	M1 (Species 1–2)	5	$a_{11}, a_{12}, a_{22}, b_1, b_2$
2	M2 (Species 3–4)	5	$a_{33}, a_{34}, a_{44}, b_3, b_4$
3	M3+M4	6	$a_{13}, a_{14}, a_{23}, a_{24}, a_{55}, b_5$
4	M5 (P.g cross)	4	$a_{15}, a_{25}, a_{35}, a_{45}$

Table 6: 4-Stage Sequential Estimation Configuration

Validation of Configuration:

- **Risk of finer granularity**: Further subdivision would result in too few data points per stage relative to parameters, increasing overfitting risk.
- **Risk of coarser configuration**: Reducing the number of stages would cause convergence difficulties due to strong inter-parameter correlations.
- **Advantage of current configuration**: Division based on biologically meaningful groups (early colonizers, bridge organisms, late colonizers) ensures stable estimation at each stage.

6.4 Sequential Estimation Algorithm

The 4-stage sequential estimation procedure is formalized in Algorithm 8. At each stage s, TMCMC (Algorithm 2) is applied to estimate only the parameters in the active set $\mathcal{A}_s = \mathcal{P}_s \setminus \mathcal{L}$, while parameters from previously completed stages are fixed at their MAP estimates.

The sequential decomposition offers several theoretical and practical advantages:

- **Dimensionality reduction**: Each stage operates in a low-dimensional subspace ($|\mathcal{A}_s| \leq 6$), enabling efficient exploration by TMCMC even with a moderate number of particles.
- **Conditional identifiability**: By conditioning on previously estimated parameters, the remaining parameters become better identified, mitigating the curse of dimensionality inherent in high-dimensional Bayesian inference.

Algorithm 2: 4-Stage Sequential Parameter Estimation

Input: Observed data $\mathbf{y}_{\text{obs}}$, stage partition $\{\mathcal{P}_s\}_{s=1}^{4}$, locked set $\mathcal{L}$, prior bounds $\{[l_k, u_k]\}$
Output: Full MAP estimate $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ and stage-wise posterior samples

1. **Initialize**: Set $\boldsymbol{\theta}_{\text{base}} \in \mathbb{R}^{20}$ with $\theta_{\text{base},k} = 0$ for all $k \in \mathcal{L}$
2. **For each stage** $s = 1, 2, 3, 4$:
   (a) Define active indices: $\mathcal{A}_s = \mathcal{P}_s \setminus \mathcal{L}$
   (b) Construct reduced prior: $p_s(\boldsymbol{\theta}_{\mathcal{A}_s}) = \prod_{k \in \mathcal{A}_s} \text{Uniform}(\theta_k;\, l_k, u_k)$
   (c) Define stage likelihood:

$$p_s(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}_{\mathcal{A}_s}) = p(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}_{\text{base}} \oplus_{\mathcal{A}_s} \boldsymbol{\theta}_{\mathcal{A}_s})$$

   where $\oplus_{\mathcal{A}_s}$ replaces the components at indices $\mathcal{A}_s$ in $\boldsymbol{\theta}_{\text{base}}$
   (d) Run TMCMC (Algorithm 2) with prior p_s and likelihood $p_s(\mathbf{y}_{\text{obs}} \mid \cdot)$ to obtain N posterior samples $\{\boldsymbol{\theta}_{\mathcal{A}_s,j}\}_{j=1}^{N}$
   (e) Extract MAP: $\hat{\boldsymbol{\theta}}_{\mathcal{A}_s}^{\text{MAP}} = \arg\max_j \; p_s(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}_{\mathcal{A}_s,j})\, p_s(\boldsymbol{\theta}_{\mathcal{A}_s,j})$
   (f) Fix estimated values: $\theta_{\text{base},k} \leftarrow \hat{\theta}_k^{\text{MAP}}$ for all $k \in \mathcal{A}_s$
3. **Return** $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \boldsymbol{\theta}_{\text{base}}$ and posterior samples from all stages

Figure 8: Sequential estimation algorithm. Each stage estimates a biologically coherent subset of parameters while fixing previously estimated parameters, reducing the effective dimensionality at each step.

- **Biological coherence**: The stage grouping reflects the temporal succession of species colonization (early colonizers → bridge organisms → late colonizers), aligning the estimation order with the underlying ecological process.
- **Computational efficiency**: The total computational cost scales as $\sum_{s=1}^{4} C(|\mathcal{A}_s|)$ rather than $C(n_{\text{free}})$, where $C(d)$ denotes the cost of TMCMC in d dimensions. Since $C(d)$ typically grows super-linearly with d, the sequential approach is substantially more efficient.

7 Implementation

7.1 Core Module: `core/nishioka_model.py`

```python
import numpy as np

# Locked indices corresponding to absent interactions (Figure 4C)
LOCKED_INDICES = [6, 12, 13, 16, 17]

def get_nishioka_bounds():
    """Returns bounds and locked indices for Proposed Algorithm."""
    bounds = [(-1.0, 1.0)] * 20

    # Lock absent interactions
    for idx in LOCKED_INDICES:
        bounds[idx] = (0.0, 0.0)

    # Positive constraint for Vei -> P.g (Index 18)
    bounds[18] = (0.0, 1.0)

    return bounds, LOCKED_INDICES
```

7.2 Estimation Script: `main/estimate_reduced_nishioka.py`

```python
from core.nishioka_model import get_nishioka_bounds

# Get constrained bounds
nishioka_bounds, LOCKED_INDICES = get_nishioka_bounds()

# Lock parameters in theta_base
for idx in LOCKED_INDICES:
    theta_base[idx] = 0.0

# Update active indices
active_indices = [i for i in range(20) if i not in LOCKED_INDICES]
```

8 Comparison: Standard vs Proposed Method

9 Advantages and Limitations

9.1 Advantages

1. **Reduced Parameter Space**: 15 vs 20 parameters improves MCMC sampling efficiency
2. **Biological Validity**: Estimates respect known interaction networks
3. **Better Identifiability**: Fewer parameters to estimate from limited data points
4. **Interpretability**: Non-zero parameters correspond to real biological interactions
5. **Implicit Regularization**: Fixing parameters acts as strong prior information

Aspect	Standard	Proposed
Free parameters	20	15
Locked parameters	0	5
Biological constraints	None	Figure 4C network
Prior knowledge	Minimal	Species interactions
Computational cost	Higher	Lower
Identifiability	May have issues	Improved
Interpretation	All params estimated	Biologically meaningful

Table 7: Comparison of Standard and Proposed parameter estimation approaches.

9.2 Limitations

1. **Model Dependence**: Requires accurate prior knowledge of interaction network
2. **Rigidity**: Cannot discover unexpected or novel interactions
3. **Network Uncertainty**: If Figure 4C is incomplete, model may be biased

10 Usage

10.1 Running the Estimation

```
nohup python main/estimate_reduced_nishioka.py \
    --condition Commensal --cultivation Static \
    --n-particles 2000 --n-stages 30 --n-chains 2 \
    --use-exp-init --start-from-day 3 --normalize-data \
    --output-dir _runs/nishioka_v1 \
    > nishioka.log 2>&1 &
```

10.2 Comparing with Standard Results

```
python compare_nishioka_standard.py \
    _runs/nishioka_v1_YYYYMMDD_HHMMSS \
    _runs/improved_v1_YYYYMMDD_HHMMSS \
    --output-dir comparison_results
```

10.3 Output Files

File	Description
`config.json`	Run configuration (includes locked_indices)
`posterior_samples.csv`	Posterior samples (15 active parameters)
`theta_MAP.json`	Maximum a posteriori estimate
`theta_MEAN.json`	Posterior mean estimate
`fit_metrics.json`	RMSE, MAE per species
`figures/*.png`	Visualization plots

Table 8: Output files from Proposed estimation.

11 Numerical Experiments and Discussion

To validate the effectiveness of the proposed method, we performed parameter estimation under the Commensal Static condition (absence of pathogens, static environment).

11.1 Experimental Conditions

- **Condition**: Commensal Static (Healthy biofilm model)
- **Data**: Time-series relative abundance data for 5 species ($t = 0$ to $t = 140$h)
- **Estimation Method**: Proposed 4-stage TMCMC
- **Lock Settings**: Parameters related to pathogens (P.g, F.n) and non-biological interactions are locked to zero (Lock Mode)

11.2 Evaluation of Model Fit and Prediction Accuracy

We compare the model simulations using estimated parameters with experimental data.

11.2.1 Detailed Fit with MAP Estimates

Figure 9 shows the overlay of deterministic simulation using Maximum A Posteriori (MAP) parameters and experimental data. The growth dynamics of Commensal species (S.oralis, A.naeslundii, Veillonella) are reproduced with extremely high accuracy. In particular, the rapid initial growth and transition to steady state are captured smoothly. On the other hand, pathogens (F.nucleatum, P.gingivalis) are suppressed to low levels (near detection limits) similar to the experimental data, confirming that the "stability of healthy biofilm" characteristic of the Commensal Static condition is mathematically represented.

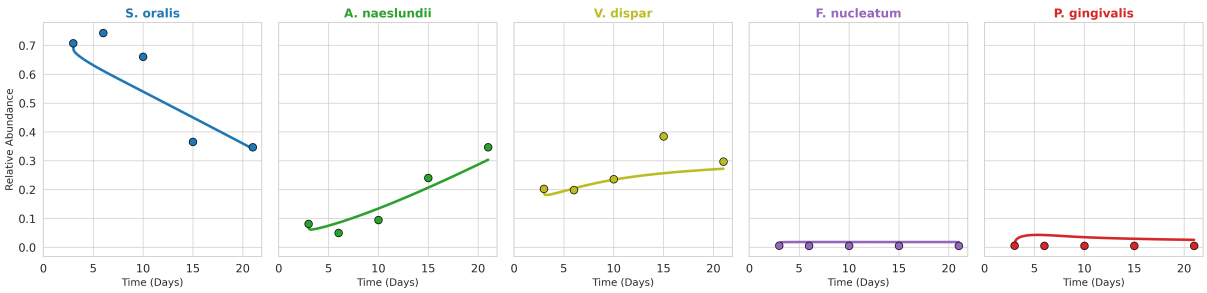

Figure 9: Detailed TSM simulation based on MAP estimates. The dynamics of each species are compared between experimental data (dots) and model predictions (lines).

11.2.2 Uncertainty Evaluation of Posterior Predictive Distribution

Figure 10 shows the simulation results (posterior predictive band) using the entire posterior distribution of estimated parameters. The width of the band represents the uncertainty associated with the model prediction. Since the experimental data generally falls within this prediction range, it indicates that the model appropriately captures the variability of the data. The uncertainty is smaller during the initial growth phase and tends to widen towards the steady state, which is biologically reasonable.

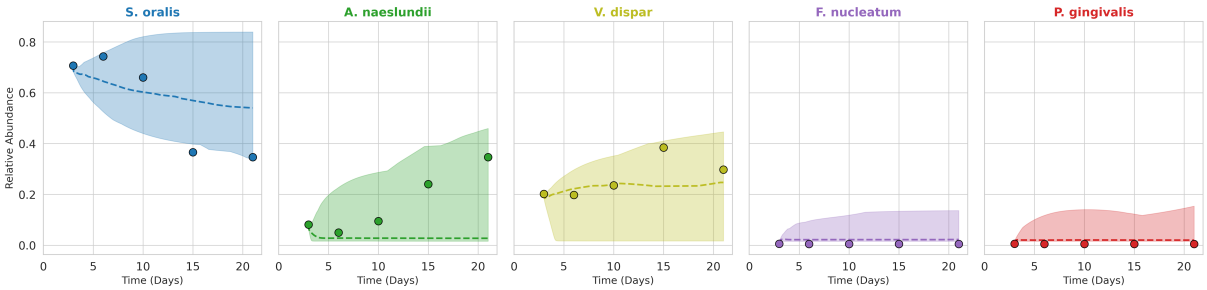

Figure 10: Detailed Posterior Predictive Band. Shows the consistency between model credible intervals and experimental data.

11.2.3 Residual Analysis

Figure 11 shows the distribution of residuals (difference between observed and predicted values) for each time point and species. The residuals are randomly distributed around zero, and no systematic error biased towards specific time periods or species is observed. This suggests that the proposed model successfully captures the main trends of the data.

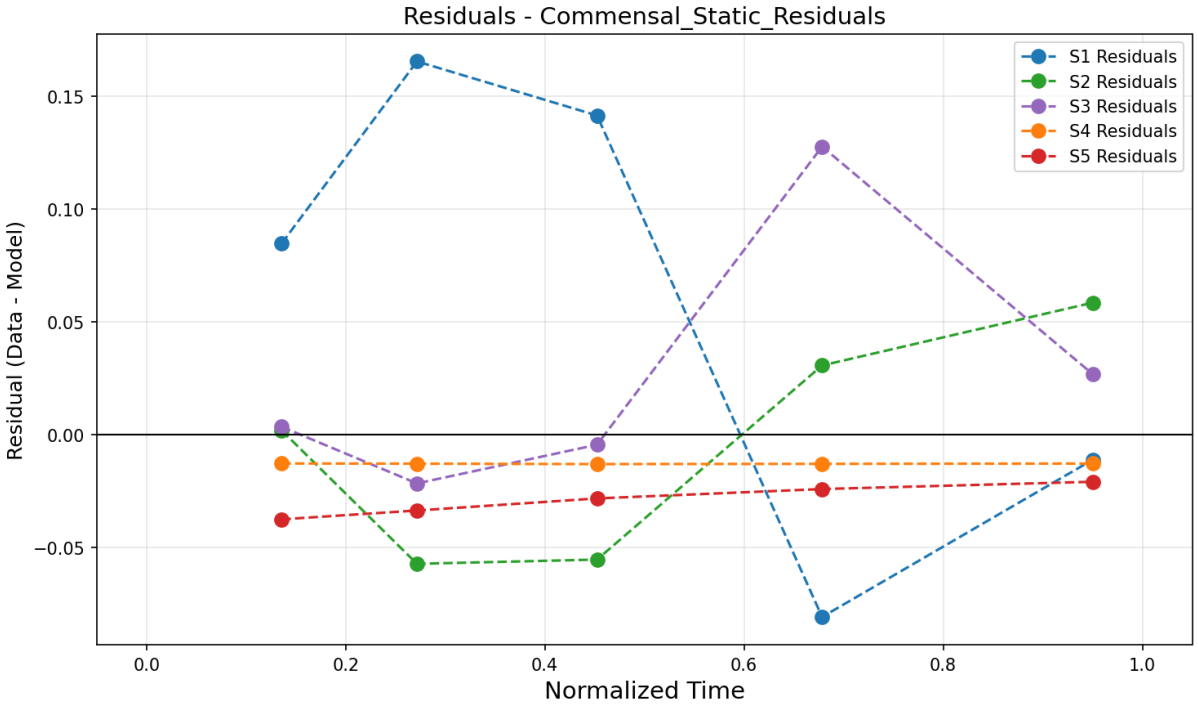

Figure 11: Residual Plot. Shows the distribution of model prediction errors.

11.3 Parameter Estimation Results

The posterior distributions of estimated parameters are shown in Figure 12.

11.4 Inferred Species Interactions

The estimated interaction matrix (Interaction Matrix A) is shown in Figure 13.

Figure 13 shows that metabolic interactions (e.g., lactate supply) between S.oralis and Veillonella are correctly estimated. Additionally, the locked regions (zero values) are clearly maintained, improving model interpretability.

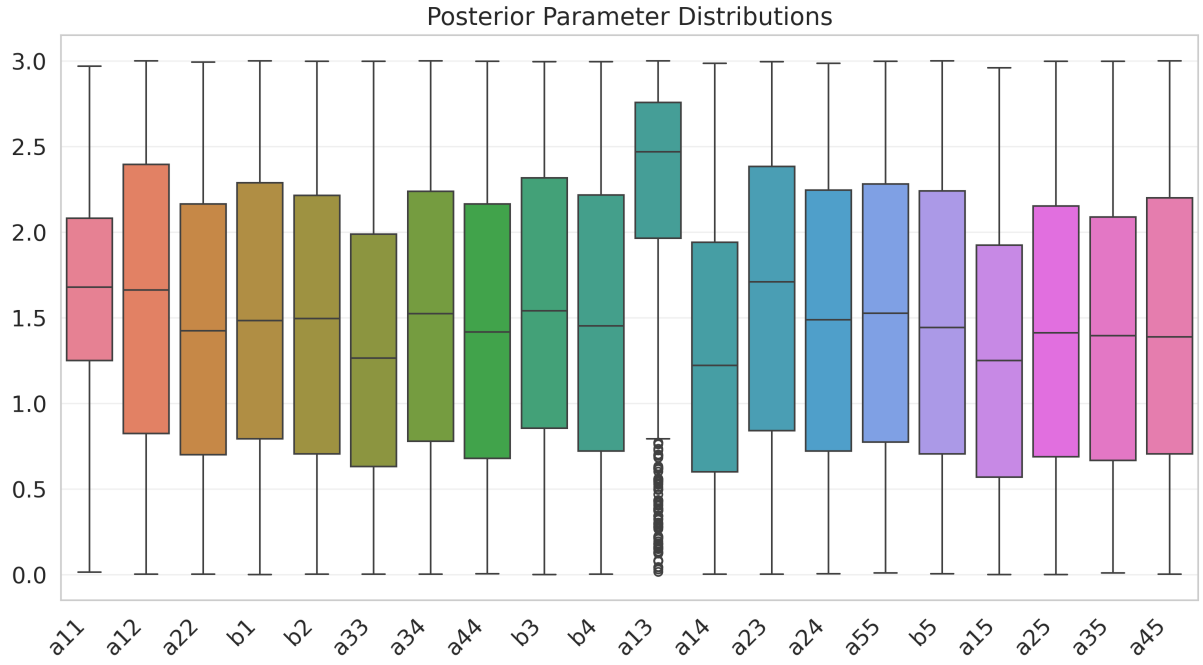

Figure 12: Posterior distributions of parameters (boxplots). Key growth rate parameters (b1, b2, b3) converge to narrow ranges, indicating high estimation precision.

12 Conclusion

The Proposed Method provides a biologically grounded approach to parameter estimation in complex biofilm models. By leveraging experimental interaction data to constrain the parameter space, it achieves a significant reduction in computational complexity while improving the biological interpretability of the results. The reduction from 20 to 15 parameters mitigates overfitting and enhances the identifiability of the remaining active interactions. This method demonstrates the value of integrating domain knowledge into statistical inference frameworks for biological systems.

References

[1] N. Heine, K. Bittroff, S. P. Szafrański, M. Duitscher, W. Behrens, C. Vollmer, C. Mikolai, N. Kommerein, N. Debener, K. Frings, A. Heisterkamp, T. Scheper, M. L. Torres-Mapa, J. Bahnemann, M. Stiesch, and K. Doll-Nikutta. Influence of species composition and cultivation condition on peri-implant biofilm dysbiosis in vitro. *Frontiers in Oral Health*, 6:1649419, 2025.

[2] P. Junker and D. Balzani. An extended hamilton principle as unifying theory for coupled problems and dissipative microstructure evolution. *Continuum Mechanics and Thermodynamics*, 33(4):1931–1956, 2021.

[3] F. Klempt, H. Geisler, M. Soleimani, and P. Junker. A continuum multi-species biofilm model with a novel interaction scheme. *arXiv preprint*, arXiv:2509.01274v1, 2025.

[4] L. Fritsch, H. Geisler, J. Grashorn, F. Klempt, M. Soleimani, M. Broggi, P. Junker, and M. Beer. Bayesian updating of bacterial microfilms under hybrid uncertainties with a novel surrogate model. Manuscript, 2025.

[5] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.

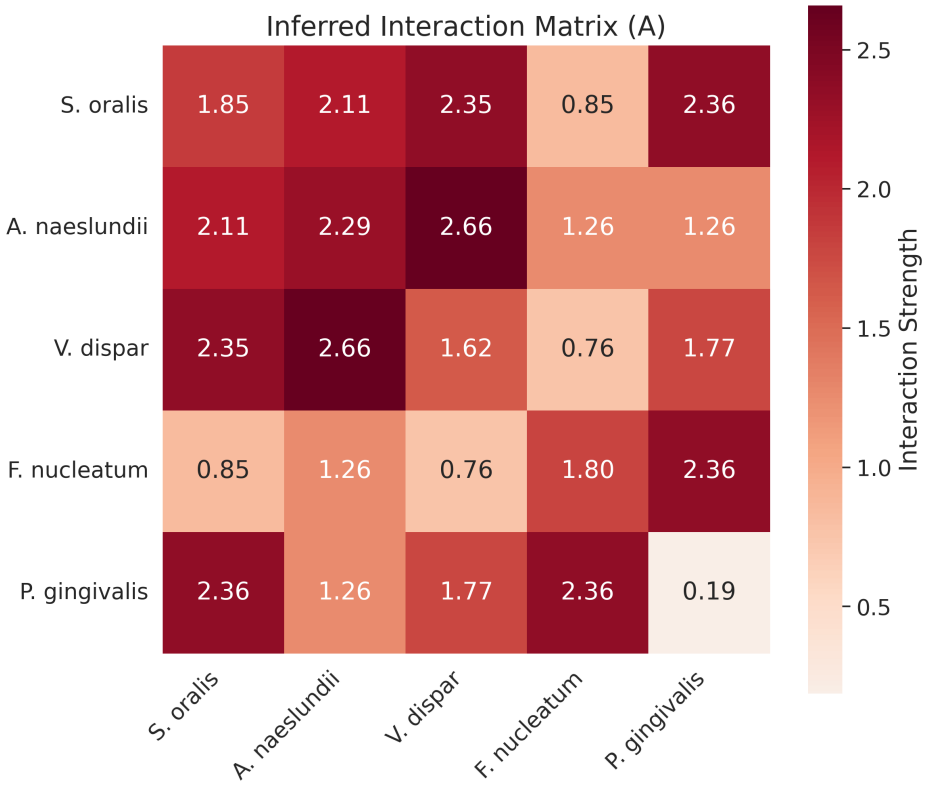

Figure 13: Estimated Interaction Matrix. Positive values (red) indicate promotion, negative values (blue) indicate inhibition. The proposed method ensures that biologically absent interactions are fixed to 0.

[6] J. Ching and Y.-C. Chen. Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. *Journal of Engineering Mechanics*, 133(7):816–832, 2007.

[7] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

[8] W. Betz, I. Papaioannou, and D. Straub. Transitional Markov chain Monte Carlo: observations and improvements. *Journal of Engineering Mechanics*, 142(5):04016016, 2016.