A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effect Model

Auriane Gabaut[1,2,3], Rodolphe Thiébaut[1,2,3,4], Cécile Proust-Lima[1], Mélanie Prague[1,2,3*]

[1] : Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France
[2] : Univ. Bordeaux, Inria, Bordeaux, France
[3] : Vaccine Reasearch Institute, Paris, France
[4] : CHU de Bordeaux, Service d'Information Médicale, INSERM, U1219, F-33000 Bordeaux, France
[*] : melanie.prague@inria.fr

1. Abstract

Non-linear mixed effect models (NLMEMs) defined by ordinary differential equations (ODEs) are central to modeling complex biological systems over time, particularly in pharmacometrics, virus dynamics and immunology. However, selecting relevant covariates associated with a dynamics in high-dimensional settings remains a major challenge. This study introduces a novel model-building approach called lasso-SAMBA that integrates Lasso regression with a stability selection algorithm for robust covariate selection within ODE-based NLMEMs.

The method iteratively constructs models by coupling penalized regression with mechanistic model estimation using SAEM algorithm. It extends a prior strategy (SAMBA: Stochastic Approximation for Model Building Algorithm) originally based on stepwise inclusion, by replacing this step with a penalized, stability-driven approach that reduces false discoveries and improves selection robustness. By maintaining the monotonic decrease of the information criterion through a calibrated exploration of penalization parameters, the proposed method outperforms conventional stepwise and Bayesian variable selection alternatives.

Extensive simulation studies, spanning pharmacokinetic and immunological models, demonstrate the superiority of lasso-SAMBA in variable selection fidelity, FDR control, and computational efficiency. The lasso-SAMBA method is implemented in an R package. Applied to a Varicella-Zoster virus vaccination study, the method reveals robust, biologically plausible associations between parameters of the mechanistic model of the humoral immune response with early transcriptomic expressions. These results underscore the practical utility of our method for high-dimensional model building in systems vaccinology and beyond.

KEYWORDS: non-linear mixed effect model - model selection - lasso regression - high-dimension - mechanistic models

2. Introduction

ODE-based models, referred to as mechanistic models, are powerful tools for understanding complex biological processes. They are used in many fields such as virus dynamics [1], epidemic modeling [2] and pharmacometrics [3]. They were introduced to analyze observations repeatedly collected from multiple individuals in a population of interest [4]. These models are defined by systems of differential equations whose parameters often carry clear biological meaning—such as the half-life of a drug, the infection rate of a virus, or the turnover of immune cells. As such, they

enable a causal interpretation of the relationships between compartments and offer a predictive framework that can simulate system dynamics under varying conditions [5, 6]. However, to ensure accurate predictions, it is essential to carefully identify all exogenous relationships with covariates that may influence the parameters and behavior of the model. The availability of high-dimensional data has increased significantly due to advances in data acquisition and analysis methods [7], complicating the problem of mechanistic model building. High-dimensional variable selection, where the number of covariates exceeds the number of observations, is well-documented in standard regression models. However, methodological developments remain limited for its application in mechanistic models.

In a broad sense, the mechanistic model of interest belongs to the family of non-linear mixed effect models (NLME), where the nonlinearity does not necessarily allow for a closed-form solution and is instead defined by an ordinary differential equation (ODE). The variability inherent in the data, captured by the model parameters, can be attributed to different sources (intraindividual, interindividual, and residual). Mixed-effect models on parameters of the ODE allow the study of N individuals exhibiting the same overall behavior. Inter-individual variability in the biological parameters can be partly explained by observed covariates through fixed effects, while residual unexplained variability is captured by individual-specific random effects. This decomposition enables the model to incorporate both structured variation, due to measurable factors (e.g., demographic covariates, biomarkers), and residual variation, reflecting subject-specific deviations. Model inference and selection are essential steps in identifying the most appropriate model to best represent the data and make accurate predictions. In this article, inference methods are not seen as a major problem since many inference methods exist [8, 9, 10]. We mainly focus on the Stochastic Approximation version of the EM algorithm [11, 12, 13] with the selection of the "best" model among the available models done using an extension of the Bayesian information criterion (BICc) that is well suited for NLME [14]. Our main goal in this work is to propose a model building strategy to determine which observed individual covariates have a fixed effect on the mechanistic model parameters when the set of available explanatory covariates of size p is high-dimensional ($N \ll p$).

Several approaches have been proposed in recent years to meet this challenge. The most common strategy involves a two-step process. First, the parameters of the NLME model are estimated. Then, a regression model is fitted to relate these parameters to the covariates using either a stepwise method [15] or lasso [16], including its variations to deal with problems such as multicollinearity [17]. Bertrand et al. [18] introduced an iterative approach in which the parameter estimates are updated after being related to covariates. However, their method relied on Empirical Bayes Estimates (EBEs), which are prone to shrinkage and sensitive to study design [19]. Yuan et al. [20] proposed a correction method for shrinkage but may fail in high-dimensional settings. An alternative to address these limitations is to integrate penalized regression into NLME models, which allow simultaneous selection of covariate and parameter estimation [21]. However, the penalized likelihoods in this approach do not incorporate structured penalties, which means that they do not induce structural sparsity or control information criteria. As a more recent extension, Naveau et al. [22] proposed a Bayesian lasso approach that controls information criteria and combines spike-and-slab priors with the SAEM algorithm. However, this work focus on non-linear mixed effect models with closed-form outcomes. In this article, we will compare our original method with the performance of their method on the published example. However, extending their approach to ODE-based NLME models remains computationally prohibitive, necessitating the development of alternative methods.

In ODE-based NLMEMs, methods for model selection such as SCM (Stepwise Covariate Modeling) [23] and COSSAC (Conditional Sampling for Stepwise Approach based on Correlation tests) [24] were proposed. SCM is a stepwise approach that systematically adds or removes

covariates to optimize model fit, while COSSAC improves this process by considering correlation tests between covariates and parameter values. The SAMBA method (Stochastic Approximation for Model Building Algorithm) has been shown to be faster with similar performance [25]. In the first step, SAMBA performs an efficient screening of covariates by establishing the linear link between covariates and parameters through stochastic approximation. In the second step, it refines the selection by focusing on covariates that have been shown to be significant, thus ensuring a robust final model after several iterations of the two-step process. However, none of these methods behaves well in high-dimensional setting : the SCM or COSSAC methods are too time-consuming to be used and the SAMBA method is based on a stepwise algorithm [26] that performs poorly in high-dimensional environments. Indeed, there is a growing risk of overfitting, multicollinearity and the introduction of bias with the usual variable selection methods such as stepAIC. These issues emphasize the need for variable selection methods that balance model complexity and predictive power, such as penalized methods like lasso [27, 28] within existing model building methods. The method presented in this article develops a selection method for covariates in high-dimensional setting for non-linear mixed effect models by replacing the stepwise algorithm for covariate selection in complex ODE-based SAMBA with lasso-based selection. To improve the robustness of the method, we introduce a stability selection algorithm [29, 30]. The stepwise algorithm originally used in SAMBA enforces a monotonic decrease in the information criterion across iterations, a property that lasso does not guarantee on its own. Therefore, to control the information criterion in the new approach, we explore multiple covariate models stochastically to identify the one that minimizes the information criterion.

The paper is organized as follows. Section 3 introduces the statistical model based on a general ODE-based NLMEM and gives an overview of the SAMBA method, hereafter referred to as the step-SAMBA method, focusing on the aspects relevant to our approach. The new lasso-SAMBA method is then presented. Simulation results are discussed in Section 4 and an application to Varicella-Zoster Virus (VZV) vaccinology study [31] revealing biological association between parameters of the humoral immune response mechanistic model and high-dimensional transcriptomic expressions is proposed in Section 5. Finally, in Section 6, the potential of the method is highlighted and an outlook for future work is given.

3. Methods

3.1. Non-Linear Mixed Effect Model

We consider an outcome variable, eventually multivariate, $Y_i = (Y_{ij})_{j \leq n_i}$ observed at time points t_{ij} with j the occasion ($j = 1, ..., n_i$) and i the individual ($i = 1, \ldots, N$). We assume that these observations come from an individual dynamic process $y_i = f(\cdot, \boldsymbol{\psi}_i)$ that depends on time and on individual-specific model parameters $\boldsymbol{\psi}_i = (\psi_{il})_{l \leq m}$. These dynamics may arise from a mechanistic model defined by an ordinary differential equation (ODE) with parameters $\boldsymbol{\psi}$. Thus, our observations can be expressed as follows:

$$\forall i \leq N, j \leq n_i, \quad u(Y_{ij}) = u(f(t_{ij}, \boldsymbol{\psi}_i)) + g(t_{ij}, \boldsymbol{\psi}_i, \boldsymbol{\xi})\varepsilon_{ij},$$

where ε_{ij} are standardized Gaussian random variables with zero mean and g is a function that possibly depends on the parameters $\boldsymbol{\xi}$ and defines the residual error model. This definition is valid up to a certain pre-determined linked function u, which can be applied to account for nonlinearity or other systematic deviations in the residual structure.

In the context of ODE-based NLME models, it is assumed that the individual parameters $\boldsymbol{\psi}_i$ are normally distributed, or up to a transformation, by a typical population parameter

$\boldsymbol{\psi}_{pop} \in \mathbb{R}^m$, with the covariance matrix $\boldsymbol{\Omega}$. If the individual i is additionally characterized by covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$, we can include covariate effects in the definition of the individual parameters, which leads to the following equation:

$$\forall i \leq N,\ h(\boldsymbol{\psi}_i) = h(\boldsymbol{\psi}_{pop}) + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\eta}_i,$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$, and $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lp})^T$ represents the effect of the p covariates on the parameter $l \leq m$, $\boldsymbol{\eta}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\Omega})$, $i \leq N$, and h is a potential transformation function applied to the covariates or parameters.

Such models, denoted hereafter by $\mathcal{M}$, are typically estimated using the Stochastic Approximation Expectation-Maximization (SAEM) algorithm [12, 13], a stochastic variant of the Expectation-Maximization (EM) algorithm [11]. The EM algorithm aims to maximize the observed-data log-likelihood by alternating between two steps:

- an Expectation step (E-step), which computes the conditional expectation of the complete-data log-likelihood given the current estimate of parameters,
- a Maximization step (M-step), which updates the parameters by maximizing the expected complete-data log-likelihood obtained from the E-step.

However, for complex models such as ODE-based NLME models, this conditional expectation cannot be computed analytically due to latent random effects $\boldsymbol{\eta}_i$. The SAEM algorithm addresses this limitation by introducing a stochastic approximation during the E-step, where individual parameters $\boldsymbol{\eta}_i$ are sampled from their conditional posterior distribution given the current estimate of population parameters, typically using Markov Chain Monte Carlo (MCMC) methods, thus enabling parameter estimation despite the absence of a closed-form likelihood.

3.2. The original step-SAMBA Algorithm

To select covariates into an ODE-based NLME model, the step-SAMBA [25] method creates models iteratively, starting with an empty initial model $\mathcal{M}_0$ that contains no covariate effects or correlations, and typically assumes a combined error model $g(t, \psi, \boldsymbol{\xi} = (a, b)) = a + bf(t, \psi)$. At each iteration k, the algorithm updates the model $\mathcal{M}_{k+1}$ by sequentially:

- estimating the parameters $\boldsymbol{\theta}^{(k)} = \left(\boldsymbol{\psi}_{pop}, \boldsymbol{\Omega}, \boldsymbol{\beta}, \boldsymbol{\xi}\right)$ of $\mathcal{M}_k$ by maximum likelihood using a SAEM algorithm [13] ;
- sampling of individual parameters $(\boldsymbol{\psi}_i^{(k)})_{i \leq N}$ using Markov Chain Monte Carlo (MCMC) methods;
- building three sub-models with $\boldsymbol{\theta}^{(k)}$ and $(\boldsymbol{\psi}_i^{(k)})_{i \leq N}$: a correlation model, an error model and a covariate model (noted $\mathcal{M}^{\mathrm{COV}}$).

Those three sub-models then form $\mathcal{M}_{k+1}$. Complete description of the method is available in Prague et al. [25].

Our key focus here is on the construction of the covariate part, $\mathcal{M}_{k+1}^{\mathrm{COV}}$. Using the generated parameters $(\boldsymbol{\psi}_i^{(k)})_{i \leq N} = (\psi_{i1}^{(k)}, \ldots, \psi_{im}^{(k)})_{i \leq N}$, a regression model is built for each parameter $\boldsymbol{\psi}_l$, where $l \leq m$:

$$h_l(\psi_{li}^{(k)}) = \mu_l + \boldsymbol{X}_i\boldsymbol{\beta}_l + \eta_{il}, \quad i \leq N \tag{1}$$

where for $l \leq m$, η_{il} are Gaussian errors corresponding to the random effects of individual i on parameters l, μ_l the intercept of the regression model, and $\boldsymbol{X}_i$ represents the individual covariates.

These are selected either by a stepwise selection procedure if there are more than 10 covariates, or by an exhaustive search among all possible models if there are less than 10 covariates.

The SAMBA method is an efficient method to improve the model building process in ODE-based NLMEMs as demonstrated in simulation and real-world examples [25]. We propose to replace the stepwise algorithm for selecting covariates in step-SAMBA with a refined lasso-based selection method in order to improve the variable selection process in high-dimensional data ($N \ll p$).

3.3. The stability-enhanced lasso-SAMBA Algorithm

Similar to the step-SAMBA method, our approach proposes to perform variable selection independently for each parameter ψ_l in the ODE-based model. At each iteration $k \in \mathbb{N}$, the first two step remain unchanged : the lasso-SAMBA method first estimates the population parameters and samples the individual parameters. Then, in the lasso-SAMBA method, we estimate coefficients in equation (1) with a lasso regression [27] given a penalty parameter λ_l:

$$(\hat{\mu}_l^{\lambda_l}, \hat{\boldsymbol{\beta}}_l^{\lambda_l}) = \underset{\mu \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i \leq N} \left(h\left(\psi_{li}^{(k)}\right) - \mu - \boldsymbol{X}_i \boldsymbol{\beta} \right)^2 + \lambda_l \|\beta\|_1 \right\}. \tag{2}$$

For simplicity, we now omit the explicit subscript of each parameter l in the remainder of the paper, e.g. in Equation (2) λ_l, $\hat{\beta}_l^{\lambda_l}$ is denoted later by λ, $\hat{\beta}^{\lambda}$.

The main goal is to identify for each parameter of the mechanistic model the true set $\mathcal{S} = \{c \leq p; \beta_c \neq 0\}$ that corresponds to the selection of truly relevant covariates. For a fixed λ, this set can be derived, for example, by:

$$\hat{S}^{\lambda} = \{c \leq p \mid \hat{\beta}_c^{\lambda} \neq 0\}$$

Usually in lasso regression, the penalty parameter, and thus the selected set of covariates, is determined through cross-validation. However, this approach can result in unstable selections, particularly in high-dimensional settings [32, 29]. To address this problem, we rely on a Stability Selection Algorithm [29, 30], a method developed to improve the performance of variable selection algorithms. It has been proven that stability selection achieves consistency in variable selection [29]. The central idea of the stability selection algorithm is to repeat the variable selection process a large number of times for a fixed value of λ on data samples that are half the size of the original dataset. The covariate $c \leq p$ is selected when its selection frequency across batches, denoted as $\hat{\pi}^{\lambda}(c)$, exceeds a given threshold $\pi \in [0, 1]$.

In our case, we draw uniformly n_{SS} random subsamples of individuals $(I_b)_{b \leq n_{SS}}$ from $\{1, \ldots, N\}$, each of size $\lfloor N/2 \rfloor$. For each batch $b \leq n_{SS}$ of sample data $\mathcal{D}_b = (\boldsymbol{X}_i, h(\psi_i^{(k)}))_{i \in I_b}$, we denote $\hat{S}_b^{\lambda} = \{c \leq p, \hat{\beta}_c^{\lambda,b} \neq 0\}$ as the set of selected covariates resulting from the lasso regression presented in Equation (2) in batch b with the penalty parameter $\lambda > 0$. The final selection set for the hyperparameters (λ, π) is finally :

$$\hat{\mathcal{S}}^{\lambda,\pi} = \left\{ c \leq p \mid \hat{\pi}^{\lambda}(c) \geq \pi \right\}. \tag{3}$$

where $\hat{\pi}^{\lambda}(c) = \dfrac{\#\left\{ b \leq n_{SS} \mid \hat{\beta}_c^{\lambda,b} \neq 0 \right\}}{n_{SS}}$.

Ultimately, this method is based on two critical hyperparameters for each parameter of the mechanistic model : $\lambda > 0$ and $\pi \in [0, 1]$.

3.4. Calibration of parameters (λ, π) and BICc

In the step-SAMBA method, the selection of covariates by stepwise algorithm seems to ensure a reduction of the criterion used for model construction. Consequently, the models $\mathcal{M}_k$ constructed during the iterations of the algorithm have a BICc criterion which usually decreases as the iterations progress so that the algorithm can stop at a certain point. However, when using a lasso method, the decrease of the BICc may be lost. This behaviour appears to result from the decrease in the BIC of the individual parameter model (1) during the stepwise covariate selection at each iteration of the step-SAMBA method. To preserve this, we conduct an exploration step for selecting the covariate model $\mathcal{M}^{COV}$ during the covariate model building step of each lasso-SAMBA iteration k. We explore a set $\Lambda \times \Pi$ of hyperparameters (λ, π) that results in models $\mathcal{M}^{COV}(\lambda, \pi)$, among which the one minimizing the BIC is selected as the candidate covariate model $\mathcal{M}^{COV}$. The covariate model is then chosen as $\mathcal{M}^{COV}_{k+1}$ if it has a lower BIC than the previous model $\mathcal{M}^{COV}_k$, otherwise $\mathcal{M}^{COV}_k$ is retained. The schematic pipeline of the procedure of lasso-SAMBA method is presented in Figure 1.

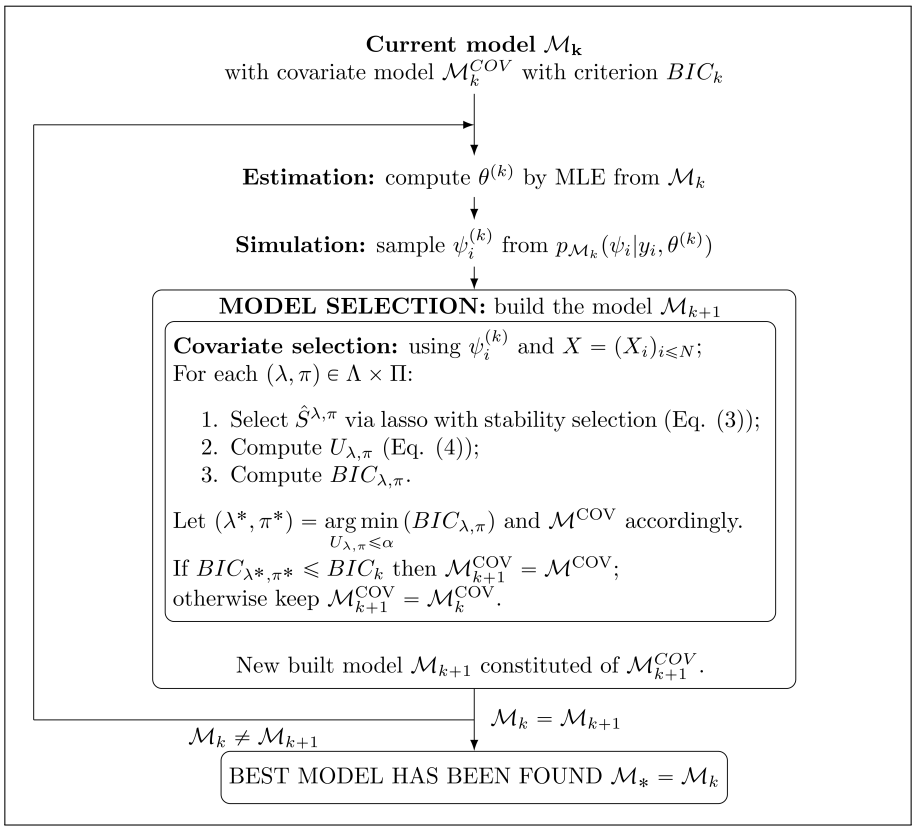


Figure 1: Lasso-SAMBA method pipeline.

In other words, rather than identifying a single set of parameters (λ, π), we aim to explore a space that leads to several covariate models to explore, searching for one that minimizes the BIC.

To characterize a good set $\Lambda \times \Pi$, we use the upper bound for the expected number of false discoveries derived by Meinshausen and Bülhmann [29]. If we denote by $V^{\lambda,\pi}$ the set of falsely

selected variables in $\hat{\mathcal{S}}^{\lambda,\pi}$, we have, for $\pi \in]0.5;1]$:

$$\mathbb{E}[V^{\lambda,\pi}] \leq \frac{1}{2\pi - 1} \times \frac{q_\lambda^2}{p}$$

where $q_\lambda = \overline{|\hat{S}_b^\lambda|}^{b \leq n_{SS}}$ is the average number of features that are selected at least once by the lasso selection with the penalty λ. This upper bound for the expected number of false discoveries can then be derived into an upper bound for the expected proportion of false discoveries, the False Discovery Rate (FDR) :

$$FDR(\lambda,\pi) = \mathbb{E}[FDP(\lambda,\pi)] \leq U_{\lambda,\pi} = \frac{1}{|\hat{\mathcal{S}}^{\lambda,\pi}|} \times \frac{1}{2\pi - 1} \times \frac{q_\lambda^2}{p} \tag{4}$$

Starting from a space of $\lambda \in \Lambda$, which ranges from a dense to a sparse model, as is usually done [33], and $\Pi =]0.5;1]$, we compute $\hat{\mathcal{S}}^{\lambda,\pi}$ for each $\lambda \in \Lambda, \pi \in \Pi$ and the induced upper bounds $U_{\lambda,\pi}$. We first remove from $\Lambda \times \Pi$ any set of parameters that do not fulfill the condition $U_{\lambda,\pi} \leq \alpha$ for a fixed hyperparameter $\alpha \in [0;1]$. For the remaining set of parameters $\Lambda \times \Pi \backslash \{(\lambda,\pi), U_{\lambda,\pi} > \alpha\}$, we choose the final set of selected covariates such that:

$$\hat{\mathcal{S}} = \underset{\substack{(\lambda,\pi)\in\Lambda\times\Pi,\\ U_{\lambda,\pi}\leq\alpha}}{\arg\min} \; BICc_{\lambda,\pi}$$

where $BIC_{\lambda,\pi}$ is the information criterion of the linear model in Equation (1) with $\hat{\mathcal{S}}^{\lambda,\pi}$ for selected covariates.

3.5. Implementation

The new method is implemented in the LSAMBA R [34] package, available at `https://github.com/aurianegbt/LSAMBA`, using the Rsmlx [25, 35] and lixoftConnectors [36] package as a base. For the lasso estimation in lasso-SAMBA, we use the sharp package [37], which implements stability-based variable selection. All code has been run with Lixoft Monolix 2023R1.

4. Simulation Studies

In this section, we evaluate the performance of the lasso-SAMBA method in terms of covariate selection through simulation studies. We conducted a simulation study under two distinct frameworks to assess the performance of the proposed method in sparse settings. The aim was to compare the newly developed procedure to the step-SAMBA [25] and the SAEMVS [22] in terms of variable selection accuracy. The first has been described above in Section 3.2, the latter consists in iteratively applying a SAEM procedure, where the E-step estimates posterior inclusion probabilities of variables using Monte Carlo sampling, and the M-step updates model parameters by maximizing a penalized likelihood, effectively shrinking irrelevant coefficients toward zero for sparse model recovery. We calibrated both literature methods by using the hyperparameters and initialization strategies detailed in the original studies. For our lasso-SAMBA, the only hyperparameter requiring specification was the constraint on the expected number of false discoveries, which we fixed at 10%. A sensitivity analysis over a range of values is reported in Supplementary Material Section C. Data were generated using biologically motivated models presented in each subsection on a toy pharmacokinetics model 4.2 and a humoral immune response model 4.3. For each framework, 100 independent replicates were generated.

4.1. Performance evaluation

In the context of covariate selection, a true positive (TP) is a covariate truly associated with the parameter and correctly identified as such; a false positive (FP) is not associated but incorrectly selected; a false negative (FN) is truly associated but missed; and a true negative (TN) is not associated and correctly excluded. The overarching goal is to have barely any false discoveries. In particular for our method, it is to reduce the False Discovery Rate (FDR) without significantly increasing the False Negative Rate (FNR), defined as:

$$FDR = \frac{FP}{FP + TP},$$

$$FNR = \frac{FN}{FN + TP}.$$

To assess the performance of the model, we compare both the FDR and FNR, while also considering the F_1-score, which provides a single metric that balances the trade-off between FP and FN:

$$F_1 = \frac{2TP}{2TP + FN + FP}.$$

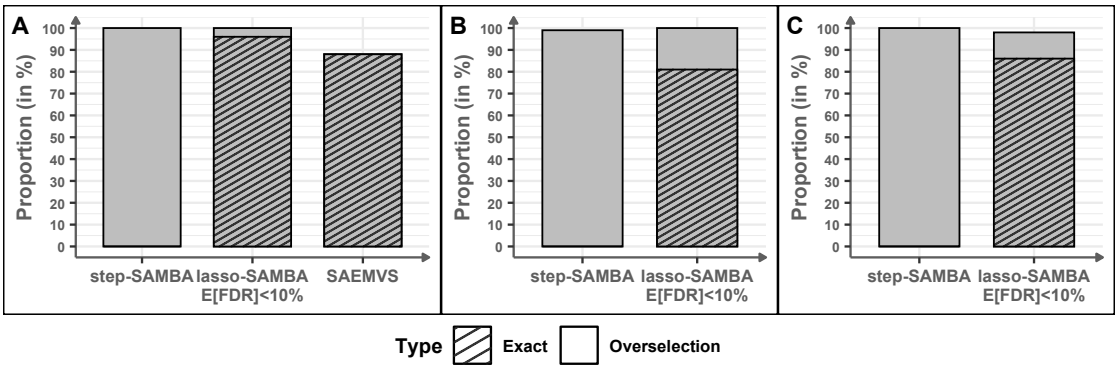


Figure 2: Proportion of exact models (no false positives or false negatives) and models that strictly include the exact model (no false negatives but with false positives). (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

4.2. A Toy Pharmacokinetics Model

We replicate the simulation study provided in Naveau et al. [22]. We simulate a two-compartment pharmacokinetic model with first order elimination. The model describes the concentration y_{ij} of a drug measured at 12 timepoints t_j, $j = 1, ..., 12$, for individual i with $i = 1, \ldots, 200$:

$$y_{ij} = \frac{D\varphi_{1i}}{V\varphi_{1i} - \varphi_{2i}} \times \left(e^{-\varphi_{2i}\frac{t}{V}} - e^{-\varphi_{1i}t}\right) + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

and

$$\begin{cases} \varphi_{1i} &= \varphi_{1pop} + \beta_1^T C_i + \eta_i^1 \\ \varphi_{2i} &= \varphi_{2pop} + \beta_2^T C_i + \eta_i^2 \end{cases},$$

Metric	**PK model** binomial covariates	**Vaccinology model** Gaussian covariates	**Vaccinology model** non-Gaussian covariates
FDR (%) :			
- step-SAMBA	87.5 [78.8;91.7]	72.7 [50.0;83.8]	70.0 [25.0;81.2]
- lasso-SAMBA	0.0 [0.0;14.3]	0.0 [0.0;25.0]	0.0 [0.0;25.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
FNR (%) :			
- step-SAMBA	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- lasso-SAMBA	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
F_1-score (%) :			
- step-SAMBA	22.2 [15.4;34.9]	42.9 [27.9;66.7]	46.2 [31.6;85.7]
- lasso-SAMBA	100.0 [92.3;100.0]	100.0 [85.7;100.0]	100.0 [85.7;100.0]
- SAEMVS	100.0 [0.0;100.0]	ND	ND

Table 1: False Discovery Rate (FDR), False Negative Rate (FNR), and F_1-score computed for each simulation framework, comparing the original step-SAMBA method and the lasso-SAMBA method, with an error control threshold of 10%. (ND: not defined).

with $\boldsymbol{\eta}_i = (\eta_i^1, \eta_i^2) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{\Omega})$ are random effects explaining individual residual heterogeneity and σ^2 is the variance of error measurements. Time points for all individuals are $t_j \in \{0.05, 0.15, 0.25, 0.4, 0.5, 0.8, 1, 2, 7, 12, 24, 40\}$. The administered dose $D = 100$ and the apparent volume of distribution $V = 30$ are known constants. The individual parameters are the individual-specific absorption rate constant φ_1 and individual-specific elimination rate constant φ_2. They are to be estimated and related to individual-specific covariates.

For each individual i, the covariate matrix $C_i \in \mathbb{R}^{500}$ is simulated from independent binomial distributions with success probability 0.2. The covariates are standardized. The effect parameters (β_1, β_2) is such that covariates 1-3 are associated with φ_{1i} and covariates 3-5 are associated with φ_{2i}. Leaving 495 covariates not associated with any parameters. The parameter values used are detailed in the Supplementary Material section A.

In this initial simulation scenario, designed to evaluate variable selection under controlled conditions, we allows both correlation and covariate model building through the lasso-SAMBA method and step-SAMBA method in order to compare results with SAEMVS algorithm. Here, we observe that the lasso-SAMBA method markedly outperforms step-SAMBA with respect to FDR control. The median FDR achieved by step-SAMBA was 88% (95% CI: 79%, 92%), while lasso-SAMBA reduced this to a striking 0.0% (95% CI: 0.0%, 14.3%), as summarized in Table 1. This considerable improvement in FDR does not come at the expense of reduced power.

Indeed, as illustrated in Figure 3 panel A, which plots covariate selection frequencies across 100 replicated datasets, lasso-SAMBA retains the same number of true positive selections as step-SAMBA, demonstrating that the improved FDR is achieved without sacrificing sensitivity. Consequently, the F_1-score shows a substantial increase: from 22% (95% CI: 15%, 35%) with step-SAMBA to 100% (95% CI: 92%, 100%) with lasso-SAMBA.

Further evaluation of model selection quality is presented in Figure 2 panel A. This figure shows the proportion of exact model recoveries and models that strictly include the correct covariates. The step-SAMBA method fails to recover the exact model in any replicate. In contrast, lasso-SAMBA successfully recovers the exact model structure in 96% of simulations. Notably, neither method yields final models with false negative covariates, indicating consistent inclusion of all relevant associations.

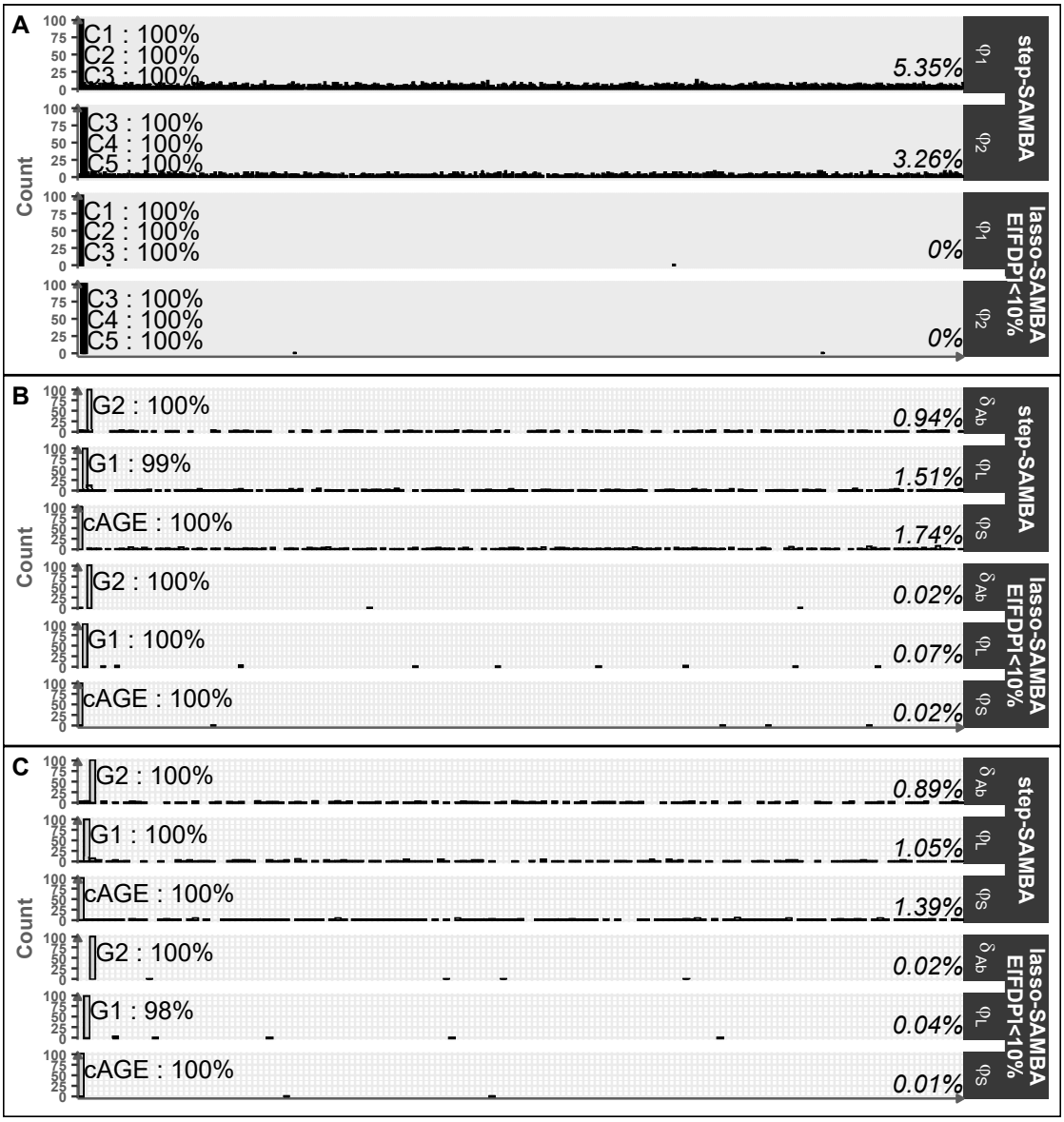


Figure 3: Covariate selection frequency across simulation frameworks for the step-SAMBA method and the lasso-SAMBA method, with an error control threshold of 10%. The mean selection frequency of false discoveries is shown on the right side of each histogram. (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

We also compare the performance of lasso-SAMBA with the SAEMVS method. In this context, SAEMVS achieves exact model recovery in 88% of replicates. However, unlike lasso-SAMBA, which maintains zero false negatives across all replicates, SAEMVS occasionally omits relevant variables. Table 1 also indicates a huge variability in performances as F_1-score can be as low as 0% in some replicates.

4.3. Simulation of Humoral immune response

To emulate a realistic biological process that is of interest for our real-data application [38], we simulated antibody kinetics for 100 individuals based on a mechanistic model incorporating two types of Antibody-Secreting Cells (ASCs): short-lived (S) and long-lived (L), which secrete antibodies (Ab) at rates φ_S and φ_L, respectively, and are characterized by their distinct half-lives δ_S and δ_L. Antibody concentrations decay over time according to a rate parameter δ_{Ab}. The dynamics is driven by an ordinary differential equation as follows:

$$\frac{d}{dt}Ab_i(t_j) = \varphi_{S_i}e^{-\delta_S t_j} + \varphi_{L_i}e^{-\delta_L t_j} - \delta_{Ab_i}Ab_i(t_j)$$

with individual parameters

$$\begin{cases} \log(\varphi_{S_i}) &= \log(\varphi_{S_{pop}}) + \beta_{AGE} \times cAGE_i + \eta_i^S \\ \log(\varphi_{L_i}) &= \log(\varphi_{L_{pop}}) + \beta_{G_1} \times G_{1i} + \eta_i^L \\ \log(\delta_{Ab_i}) &= \log(\delta_{Ab_{pop}}) + \beta_{G_2} \times G_{2i} + \eta_i^{Ab} \end{cases}$$

with $\eta_i^S \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega_S^2)$, $\eta_i^L \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega_L^2)$ and $\eta_i^{Ab} \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega_{Ab}^2)$, and $(\beta_{AGE}, \beta_1, \beta_2)$ the effect parameters. For each individual, the centred age $cAGE$ of participants follows a normal distribution with variance 16, G_1 and G_2 are drawn from standardized normal distribution. The process is observed at times $t_j \in \{0, 7, 21, 123, 180, 300\}$.

To simulate a high-dimensional setting, 200 covariates were generated by adding 197 noise variables alongside the 3 informative ones. Two data generation scenarios were considered to test the robustness of the methods: one involving purely Gaussian covariates and another with a heterogeneous mix of randomly drawn distributions, including both categorical and continuous variables. In both cases, correlation was introduced among the covariates to mimic realistic dependencies observed in biomedical datasets. The full details of the covariate generation process are provided in Supplementary Material Section A.

In the second simulation framework, covariate model building has been done through lasso-SAMBA and step-SAMBA. The lasso-SAMBA method again demonstrates a marked improvement in FDR over the step-SAMBA method. The median FDR decreases from 73% (95% CI: 50%, 84%) to 0.0% (95% CI: 0.0%, 25.0%), as shown in Table 1. Figure 3, panel B, confirms that this improvement in specificity is not achieved at the cost of sensitivity, as the number of true discoveries remains comparable between the two methods. In turn, the F_1-score is drastically increased under lasso-SAMBA, reaching a median of 100% (95% CI: 86%, 100%), again 95% confidence intervals for FDR and F_1-score do not overlap between the methods. Figure 2 panel B further illustrates that, whereas the step-SAMBA method fails to recover the exact model in any replicate, lasso-SAMBA identifies the correct model structure in 81% of cases. It is important to note that the SAEMVS method could not be applied in this simulation framework, as the underlying model involves numerical integration of ordinary differential equations (ODEs), a feature not supported by the current implementation of SAEMVS, which is limited to analytical closed-form problems.

To assess the robustness of the method to deviations from classical assumptions, a third simulation setting was implemented where covariates were randomly drawn from heterogeneous distributions, including categorical and non-Gaussian continuous variables. Even under these more complex and less idealized conditions, lasso-SAMBA continued to substantially outperform the benchmark method. The median FDR drops from 70% (95% CI: 25%, 81%) to 0.0% (95% CI: 0%, 25%), as summarized in Table 1. Figure 3 panel C indicates that the number of true discoveries remains largely unaffected, while the median F_1-score increases from 46% (95% CI: 31%, 85%) to 100% (95% CI: 88%, 100%). Model selection quality is again improved: as shown

in Figure 2 panel C, step-SAMBA fails to identify the true model in any replicate, while lasso-SAMBA achieves exact recovery in 86% of cases.

Finally, a comparative analysis of computational times across all simulation frameworks reveals that lasso-SAMBA is generally faster. Although individual iterations of lasso-SAMBA tend to be slower due to the optimization procedure, the algorithm converges in fewer steps compared to step-SAMBA. In the two more complex vaccinology-inspired frameworks, lasso-SAMBA required a median of two iterations to converge, versus four for step-SAMBA. These runtime differences are visualized in Supplementary Material Section C.

5. Application to VZV Vaccination Data

To illustrate the utility of our approach in a real-world setting, we applied the proposed methodology to a publicly available dataset from a clinical study of immune response to vaccination against the Varicella-Zoster Virus (VZV). This dataset, accessible under the accession number SDY984 on the ImmPort platform [39], stems from the study titled "Zoster vaccine in young and elderly", and includes gene expression and antibody response data following immunization with ZOSTAVAX, a live attenuated vaccine targeting VZV [31].

ZOSTAVAX [40] is employed both for primary prevention of varicella-a childhood illness marked by viremia, fever, and vesicular skin lesions-and for the prevention of herpes zoster, a reactivation of latent VZV that manifests as a localized dermatomal rash. In this study, 35 adult volunteers were administered the vaccine, with longitudinal sampling of whole blood at baseline and days 1, 3, 7, 14, 30, 90, and 180 post-vaccination. Transcriptomic data were obtained at days 0, 1, 3, and 7, while antibody levels were measured at days 0, 7, 14, 30, 90, and 180. Of the 10,086 profiled genes, a subset of 784 protein-coding genes was selected based on functional annotation with roles in Interferon signaling, Type I Interferon response, Neutrophil activation, Inflammation, Cytokine/chemokine activity, and Cell cycle regulation, as defined by Biomart [41] and Li et al. [42].

To characterize antibody kinetics post-vaccination, we utilized the humoral immune response model proposed by Pasin et al. [38], which has been studied in simulations Section 4.3 and is detailed in Supplementary Material Section B. The analysis specifically explored which early gene expression signatures are predictive of early humoral responses, focusing on the first day post-vaccination-a time point previously shown to exhibit maximal transcriptomic variation. We investigate the relationship between these genes and the parameters φ_S, φ_L and δ_S using the step-SAMBA and lasso-SAMBA methods.

The identified gene association exhibit by step-SAMBA and lasso-SAMBA method are presented in Table 2. Step-SAMBA identified 15 genes while the lasso-SAMBA method yielded a sparse model in which LEP and KIFC1 were uniquely associated with the model parameters φ_S and φ_L, respectively. The model resulting from the model building procedure for parameters φ_S and φ_L respectively follows :

$$\log(\varphi_{S_i}) = \log(\varphi_{S_{pop}}) + \beta_{LEP} \times LEP_i + \eta_i^S,$$

and

$$\log(\varphi_{L_i}) = \log(\varphi_{L_{pop}}) + \beta_{KIFC1} \times KIFC1_i + \eta_i^L.$$

The final model estimates, incorporating these genes, is described in detail in Supplementary Material Section B. These findings implicate LEP and KIFC1 as candidate biomarkers of antibody-secreting cell dynamics during early vaccine-induced immunity. The LEP and KIFC1 respectively belong to the Cell cycle pathway. While the first encodes for the leptin hormone, the second

encodes for the production of kinesin protein C1. Their potential role on immunity is discussed in Section 6.

To evaluate model specificity and control for spurious associations, we conducted model selection using a null covariate dataset without any biological signal. This artificial dataset was generated by fitting a multivariate Gaussian distribution to the empirical gene expression values of the 784 preselected genes. Under these conditions, the lasso-SAMBA method did not identify any significant associations with the antibody model parameters, supporting its robustness to spurious associations. In contrast, the step-SAMBA method produced six false-positive associations-one involving φ_S and five involving φ_L.

Further assessment of model stability was performed via nonparametric bootstrap analysis using 500 resampled datasets (Figure 4). The lasso-SAMBA method consistently identified LEP and KIFC1 as the most stable covariates, with selection frequencies of 25.2% and 22.6%, respectively. The step-SAMBA method also frequently selected LEP (45.8%) and KIFC1 (46.2%) but exhibited greater model variability; no other genes surpassed a selection frequency of 12.8%. Additionally, step-SAMBA failed to converge in 7 out of 500 iterations, primarily due to overfitting-related estimation issues.

Together, these findings underscore the enhanced robustness and reproducibility of the lasso-SAMBA method in identifying biologically possibly relevant gene–parameter associations in the context of humoral immune modeling.

	Genes selected	
Parameter	step-SAMBA	lasso-SAMBA
φ_S	ASPM, CDC45, **LEP**, NOTCH1, SP110	LEP
φ_L	GDPD5, IKZF2, **KIFC1**, PSORS1C1, SLC16A5, SLC50A1, SRP72	KIFC1
δ_S	ISTNA1, RAB2A, SLC31A1	

Table 2: Results from the step-SAMBA and lasso-SAMBA method for the ZOSTAVAX vaccine. The initial model had an empty covariate structure, a constant error model, and no correlation. Both methods started from the same initial conditions, with covariate model selection being the only difference.

6. Discussion

Model selection remains a central challenge in the development of reliable statistical models, particularly within the context of personalized medicine, where the accurate identification of individual-specific covariates can substantially influence therapeutic decision-making. In this paper, we proposed a method for selecting covariate in high-dimensional settings by integrating a lasso approach for covariate selection within the framework of non-linear mixed effect models described by ordinary differential equations. Although the method was presented for a univariate longitudinal outcome variable, it also applies to any ODE-based model with multivariate longitudinal data.

The motivation for this enhancement stems from the limitations of traditional stepwise selection procedures as implemented in step-SAMBA which, although computationally tractable, are known to be susceptible to overfitting and the inclusion of non-informative covariates, especially in high-dimensional settings [43]. By integrating the lasso, coupled with a stability selection procedure, we aimed to improve the selection of relevant covariates while preserving computational

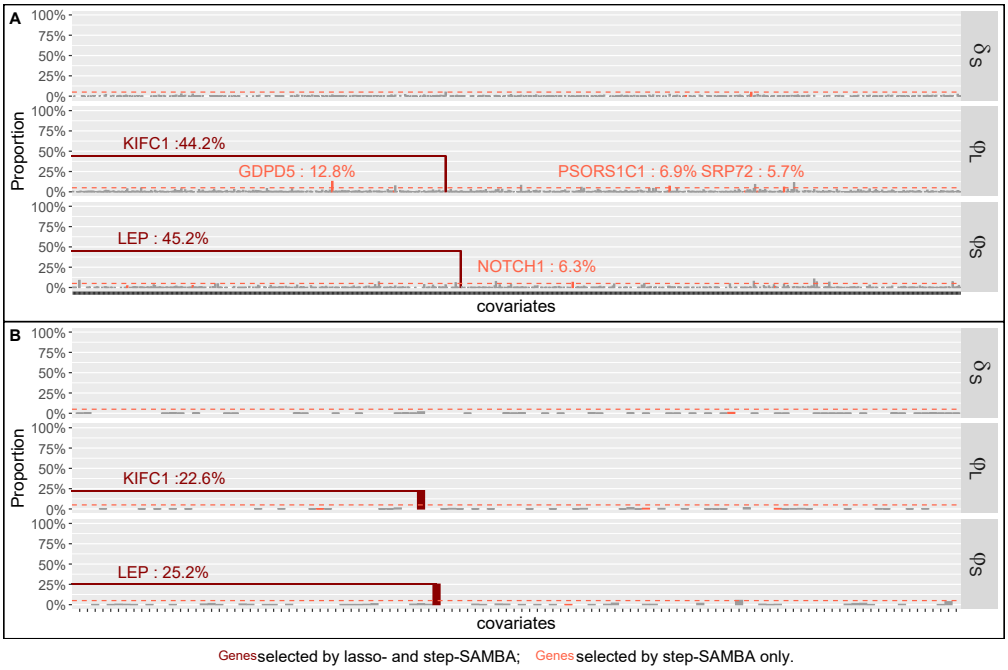


Figure 4: Bootstrap selection frequencies for covariates across all parameter models. The top panel shows results for the step-SAMBA method (493 successful runs), and the bottom panel for the lasso-SAMBA method (500 runs). Selection frequencies above 5% of selected genes by any of the procedures are indicated.

feasibility in contrast with the approach by Naveau et al. [22].

Our simulation studies demonstrated a marked improvement in the control of FDR compared to other procedures, without compromising FNR. This improvement was consistent across diverse pharmacokinetic and vaccinology-inspired model scenarios. These findings reinforce prior conclusions on the robustness of penalized regression approaches in complex hierarchical models [44].

Furthermore, in the real-data application involving antibody response to Varicella-Zoster virus vaccination, the lasso-SAMBA approach exhibited superior performance in identifying biologically plausible biomarkers while producing a sparser covariate selection. This is consistent with the detection of true signals observed in our simulation studies. The LEP gene is a player in several common biological pathways involved in the immune response such as JAK-STAT or NFkB. KIFC1 is invovled in cell proliferation and therefore is not specific of immune response. These biomarkers may however be reflecting the involvment of broader pathways in the establishment of the vaccine response [45]. KIFC1 also has the ability to promote stable mitotic spindle formation during early B cell development where centriole duplication is frequent but must be tightly regulated [46]. These results suggest that our method improves the reliability of gene-covariate relationships in biomedical research involving ODE-based modeling.

The methodological integration we propose holds promise for broader applications, particularly in fields such as pharmacogenomics and systems pharmacology, where model interpretability and variable selection are of paramount importance. Future investigations could extend this approach to non-linear relationships between covariates and parameters.

Data and code availability

Code for running the simulations and applications are available at `https://github.com/aurianegbt/LassoSAMBA`.

Acknowledgment

This study was carried out in the framework of the University of Bordeaux's France 2030 program RRI PHDS. We thank the Human Immunology Project Consortium for providing the ImmPort platform for open-access immunology data. We also thank the investigators who contributed data for study SDY984, supported by the Human Immunology Project Consortium (HIPC) RFA-AI-15-041. We thank Simulations Plus, Lixoft division for the free academic use of the MonolixSuite. Computer time for this study was provided by the computing facilities of the PlaFRIM (Plateforme Fédérative pour la Recherche en Informatique et Mathématiques). We disclose that generative AI tools were used to polish English language.

List of Figures

1 Lasso-SAMBA method pipeline. 6
2 Proportion of exact models (no false positives or false negatives) and models that strictly include the exact model (no false negatives but with false positives). (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates. 8
3 Covariate selection frequency across simulation frameworks for the step-SAMBA method and the lasso-SAMBA method, with an error control threshold of 10%. The mean selection frequency of false discoveries is shown on the right side of each histogram. (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates. 10
4 Bootstrap selection frequencies for covariates across all parameter models. The top panel shows results for the step-SAMBA method (493 successful runs), and the bottom panel for the lasso-SAMBA method (500 runs). Selection frequencies above 5% of selected genes by any of the procedures are indicated. 14

List of Tables

1 False Discovery Rate (FDR), False Negative Rate (FNR), and F_1-score computed for each simulation framework, comparing the original step-SAMBA method and the lasso-SAMBA method, with an error control threshold of 10%. (ND: not defined). 9
2 Results from the step-SAMBA and lasso-SAMBA method for the ZOSTAVAX vaccine. The initial model had an empty covariate structure, a constant error model, and no correlation. Both methods started from the same initial conditions, with covariate model selection being the only difference. 13

References

[1] A. S. Perelson and R. M. Ribeiro, "Introduction to modeling viral infections and immunity," *Immunological reviews*, vol. 285, no. 1, p. 5, 2018.

[2] I. Ganser, D. L. Buckeridge, J. Heffernan, M. Prague, and R. Thiébaut, "Estimating the population effectiveness of interventions against covid-19 in france: A modelling study," *Epidemics*, vol. 46, p. 100744, 2024.

[3] M. Ibarra, M. Lorier, and I. F. Trocóniz, "Pharmacometrics: Population approach," in *The ADME Encyclopedia: A Comprehensive Guide on Biopharmacy and Pharmacokinetics*, pp. 1–13, Springer, 2021.

[4] M. Lavielle, *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC, 2014.

[5] O. O. Aalen, K. Røysland, J. M. Gran, and B. Ledergerber, "Causality, mediation and time: a dynamic viewpoint," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 175, no. 4, pp. 831–861, 2012.

[6] M. Prague, D. Commenges, J. M. Gran, B. Ledergerber, J. Young, H. Furrer, and R. Thiébaut, "Dynamic models for estimating the effect of haart on cd4 in observational studies: Application to the aquitaine cohort and the swiss hiv cohort study," *Biometrics*, vol. 73, no. 1, pp. 294–304, 2017.

[7] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS computational biology*, vol. 13, no. 5, p. e1005457, 2017.

[8] Q. Clairon, C. Pasin, I. Balelli, R. Thiébaut, and M. Prague, "Parameter estimation in nonlinear mixed effect models based on ordinary differential equations: an optimal control approach," *Computational Statistics*, pp. 1–31, 2023.

[9] Y. Wang, "Derivation of various nonmem estimation methods," *Journal of Pharmacokinetics and pharmacodynamics*, vol. 34, pp. 575–593, 2007.

[10] M. Lavielle and F. Mentré, "Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software," *Journal of pharmacokinetics and pharmacodynamics*, vol. 34, pp. 229–249, 2007.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977.

[12] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the em algorithm," *Annals of statistics*, pp. 94–128, 1999.

[13] E. Kuhn and M. Lavielle, "Maximum likelihood estimation in nonlinear mixed effects models," *Computational statistics & data analysis*, vol. 49, no. 4, pp. 1020–1038, 2005.

[14] M. Delattre, M. Lavielle, and M.-A. Poursat, "A note on bic in mixed-effects models," *Electronic Journal Statistics*, vol. 8, no. 1, pp. 456–475, 2014.

[15] T. Lehr, H.-G. Schaefer, and A. Staab, "Integration of high-throughput genotyping data into pharmacometric analyses using nonlinear mixed effects modeling," *Pharmacogenetics and genomics*, vol. 20, no. 7, pp. 442–450, 2010.

[16] J. Ribbing, J. Nyberg, O. Caster, and E. N. Jonsson, "The lasso - a novel method for predictive covariate model building in nonlinear mixed effects models," *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 34, pp. 485–517, 8 2007.

[17] E. Haem, K. Harling, S. M. T. Ayatollahi, N. Zare, and M. O. Karlsson, "Adjusted adaptive lasso for covariate model-building in nonlinear mixed-effect pharmacokinetic models," *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 44, pp. 55–66, 2 2017.

[18] J. Bertrand and D. J. Balding, "Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models," *Pharmacogenetics and genomics*, vol. 23, no. 3, pp. 167–174, 2013.

[19] T. Nguyen, M.-S. Mouksassi, N. Holford, N. Al-Huniti, I. Freedman, A. C. Hooker, J. John, M. O. Karlsson, D. Mould, J. Pérez Ruixo, *et al.*, "Model evaluation of continuous data pharmacometric models: metrics and graphics," *CPT: pharmacometrics & systems pharmacology*, vol. 6, no. 2, pp. 87–109, 2017.

[20] M. Yuan, Z. Zhu, Y. Yang, M. Zhao, K. Sasser, H. Hamadeh, J. Pinheiro, and X. S. Xu, "Efficient algorithms for covariate analysis with dynamic data using nonlinear mixed-effects model," *Statistical Methods in Medical Research*, vol. 30, no. 1, pp. 233–243, 2021.

[21] J. Bertrand, M. De Iorio, and D. Balding, "Integrating dynamic mixed-effect modelling and penalized regression to explore genetic association with pharmacokinetics," *Pharmacogenetics and genomics*, vol. 25, 03 2015.

[22] M. Naveau, G. Kon Kam King, R. Rincent, L. Sansonnet, and M. Delattre, "Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the saem algorithm," *Statistics and Computing*, vol. 34, Dec. 2023.

[23] R. J. Svensson and E. N. Jonsson, "Efficient and relevant stepwise covariate model building for pharmacometrics," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 11, no. 9, pp. 1210–1222, 2022.

[24] G. Ayral, J.-F. Si Abdallah, C. Magnard, and J. Chauvin, "A novel method based on unbiased correlations tests for covariate selection in nonlinear mixed effects models: The cossac approach," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 10, no. 4, pp. 318–329, 2021.

[25] M. Prague and M. Lavielle, "Samba: A novel method for fast automatic model building in nonlinear mixed-effects models," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 11, no. 2, pp. 161–172, 2022.

[26] Z. Zhang, "Variable selection with stepwise and best subset approaches," *Annals of Translational Medicine*, vol. 4, no. 7, 2016.

[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.

[28] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407 – 499, 2004.

[29] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society, Series B*, vol. 72, pp. 417–473, 2010.

[30] R. D. Shah and R. J. Samworth, "Variable selection with error control: another look at stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 1, pp. 55–80, 2013.

[31] S. Li, N. L. Sullivan, N. Rouphael, T. Yu, S. Banton, M. S. Maddur, M. McCausland, C. Chiu, J. Canniff, S. Dubey, K. Liu, V. Tran, T. Hagan, S. Duraisingham, A. Wieland, A. K. Mehta, J. A. Whitaker, S. Subramaniam, D. P. Jones, A. Sette, K. Vora, A. Weinberg, M. J. Mulligan, H. I. Nakaya, M. Levin, R. Ahmed, and B. Pulendran, "Metabolic phenotypes of response to vaccination in humans," *Cell*, vol. 169, pp. 862–877.e17, 5 2017.

[32] P. Bühlmann and S. van de Geer, *Variable selection with the Lasso*, pp. 183–247. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[33] J. Friedman, R. Tibshirani, and T. Hastie, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[34] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[35] F. Mihaljevic, *Rsmlx: R Speaks 'Monolix'*. Lixoft SAS, Antony, France, 2023. R package version 2023.1.5.

[36] LIXOFT, *lixoftConnectors: R connectors for Lixoft Suite (@Lixoft)*. Lixoft SAS, Antony, France, 2023. R package version 2023.1.

[37] B. Bodinier, *sharp: Stability-enHanced Approaches using Resampling Procedures*, 2024. R package version 1.4.6.

[38] C. Pasin, I. Balelli, T. Van Effelterre, V. Bockstal, L. Solforosi, M. Prague, M. Douoguih, and R. Thiébaut, "Dynamics of the Humoral Immune Response to a Prime-Boost Ebola Vaccine: Quantification and Sources of Variation," *Journal of Virology*, vol. 93, no. 18, pp. e00579–19, 2019.

[39] S. Bhattacharya, P. Dunn, C. Thomas, H. Schaefer, J. Chen, Z. Hu, K. Zalocusky, R. Shankar, S. Shen-Orr, E. Thomson, J. Wiser, and A. Butte, "Immport, toward repurposing of open access immunological assay data for translational and clinical research," *Scientific Data*, vol. 5, p. 180015, 02 2018.

[40] A. M. Arvin, "Varicella-zoster virus," *Clinical Microbiology Reviews*, vol. 9, no. 3, pp. 361–381, 1996.

[41] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber, "Biomart and bioconductor: a powerful link between biological databases and microarray data analysis," *Bioinformatics*, vol. 21, pp. 3439–3440, 2005.

[42] S. Li, N. Rouphael, S. Duraisingham, S. Romero-Steiner, S. Presnell, C. Davis, D. S. Schmidt, S. E. Johnson, A. Milton, G. Rajam, S. Kasturi, G. M. Carlone, C. Quinn, D. Chaussabel, A. K. Palucka, M. J. Mulligan, R. Ahmed, D. S. Stephens, H. I. Nakaya, and B. Pulendran, "Molecular signatures of antibody responses derived from a systems biology study of five human vaccines," *Nature immunology*, vol. 15, no. 2, pp. 195–204, 2014.

[43] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection–a review and recommendations for the practicing statistician," *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.

[44] A. Caillebotte, E. Kuhn, and S. Lemler, "Estimation and variable selection in a joint model of survival times and longitudinal outcomes with random effects.." working paper or preprint, June 2023.

[45] T. Hagan, B. Gerritsen, L. E. Tomalin, S. Fourati, M. P. Mulè, D. G. Chawla, D. Rychkov, E. Henrich, H. E. R. Miller, J. Diray-Arce, P. Dunn, A. Lee, A. Deckhut-Augustine, R. Gottardo, E. K. Haddad, D. A. Hafler, E. Harris, D. Farber, S. H. Kleinstein, O. Levy, J. McElrath, R. R. Montgomery, B. Peters, B. Pulendran, A. Rahman, E. F. Reed, N. Rouphael, M. M. Sarwal, R. P. Sékaly, A. Fernandez-Sesma, A. Sette, K. Stuart, A. Togias, J. S. Tsang, O. Levy, R. Gottardo, M. M. Sarwal, J. S. Tsang, M. Suárez-Fariñas, R.-P. Sékaly, S. H. Kleinstein, and B. Pulendran, "Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses," *Nature Immunology*, vol. 23, pp. 1788–1798, 12 2022.

[46] M. A. Schapfl, G. M. LoMastro, V. Z. Braun, M. Hirai, M. S. Levine, E. Kiermaier, V. Labi, A. J. Holland, and A. Villunger, "Centrioles are frequently amplified in early b cell development but dispensable for humoral immunity," *Nature Communications*, vol. 15, no. 1, p. 8890, 2024.

Supplementary Material

A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effect Model

Auriane Gabaut[1,2,3], Rodolphe Thiébaut[1,2,3,4], Cécile Proust-Lima[1], Mélanie Prague[1,2,3*]

[1] : Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, France
[2] : Univ. Bordeaux, Inria, Bordeaux, France
[3] : Vaccine Reasearch Institute, Paris, France
[4] : CHU de Bordeaux, Service d'Information Médicale, INSERM, U1219, F-33000 Bordeaux, France
[*] : melanie.prague@inria.fr

This supplementary material provides additional details and results supporting the main manuscript. It includes (A) the full specification of simulation models and parameters, (B) the mathematical formulation and estimation of the ODE-based model used in the real data application, and (C) complementary analyses figures such as sensitivity assessments, computational benchmarking, and convergence diagnostics.

The method is implemented in the LSAMBA R [R Core Team, 2024] package, available at `https://github.com/aurianegbt/LSAMBA`. Code for running the simulations and applications are available at `https://github.com/aurianegbt/LassoSAMBA`. All code has been run with Lixoft Monolix 2023R1.

Contents

A Simulation generation process **1**
A.1 Pharmacokinetic model . . . 1
A.2 Humoral Immune Response . . . 3
A.3 Covariates Generation . . . 4

B Additional analysis for the real-data application **7**
B.1 Initial model . . . 7
B.2 Final Model . . . 13

C Additional graphs for results **18**

A. Simulation generation process

For all simulated datasets, age of participant, when considered, was centred, following standard preprocessing practices before model fitting.

A.1. Pharmacokinetic model

For the growth model, we used the datasets provided by Naveau et al. [Naveau et al., 2023], available at `https://github.com/Marion-Naveau/Supp_Information_SAEMVS/tree/main`. The phar-

macokinetic model is defined as:

$$y = \frac{D\varphi_1}{V\varphi_1 - \varphi_2}\left(e^{-\varphi_2 t/V} - e^{-\varphi_1 t}\right),$$

where the population parameters include the dose D, the volume V, and the two rate parameters φ_1 and φ_2. Inter-individual variability is introduced on φ_1 and φ_2, such that for each individual $i \leq N$:

$$\begin{cases} \varphi_{1_i} &= \varphi_{1_{pop}} + \beta_{11}C_1 + \beta_{12}C_2 + \beta_{13}C_3 + \eta_i^1 \\ \varphi_{2_i} &= \varphi_{2_{pop}} + \beta_{23}C_3 + \beta_{24}C_4 + \beta_{25}C_5 + \eta_i^2 \end{cases},$$

with $\boldsymbol{\eta}_i = (\eta_i^1, \eta_i^2) \overset{i.i.d.}{\sim} \mathcal{N}(0, \Omega)$.

The error model is constant, yielding the following observation equation:

$$Y_{ij} = \frac{D\varphi_{1_i}}{V\varphi_{1_i} - \varphi_{2_i}}\left(e^{-\varphi_{2_i} t/V} - e^{-\varphi_{1_i} t}\right) + \varepsilon_{ij},$$

where $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Each of the $N = 200$ individuals is observed at 12 timepoints:

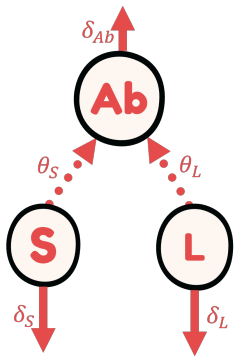
$$\{0.05, 0.15, 0.25, 0.4, 0.5, 0.8, 1, 2, 7, 12, 24, 40\}.$$

The simulation values used are summarized in Table 1.

Parameter	**Definition**	**Mean value**
Fixed Effects		
D	Dose	100
V	Volume	30
$\varphi_{1_{pop}}$	Absorption rate (pop)	6
β_{11}	MF per unit increase in C_1	3
β_{12}	MF per unit increase in C_2	2
β_{13}	MF per unit increase in C_3	1
$\varphi_{2_{pop}}$	Elimination rate (pop)	8
β_{23}	MF per unit increase in C_3	3
β_{24}	MF per unit increase in C_4	2
β_{25}	MF per unit increase in C_5	1
Random Effects		
Ω	Covariance matrix for (φ_1, φ_2)	$\begin{pmatrix} 0.2 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$
Error Model		
σ^2	Error variance	10^{-3}

Supplementary Table 1: Simulation parameters for the pharmacokinetic model based on Naveau et al. [Naveau et al., 2023] (*MF: Multiplicative factor*).

A.2. Humoral Immune Response



Supplementary Figure 1: Structure of the mechanistic model describing humoral immune response.

To simulate antibody dynamics, we adopted a non-linear mixed effects model based on the work of Pasin et al. [Pasin et al., 2019], previously used to characterize vaccine-induced humoral responses (Figure 1). The model integrates short-lived (S) and long-lived (L) Antibody-Secreting Cells (ASCs), each contributing to antibody production with distinct kinetics.

- **Structural Model** :

$$\forall i \leq N, j \leq n_i, \quad \begin{cases} \dfrac{d}{dt} Ab_i(t_{ij}) & = & \varphi_{S_i} e^{-\delta_S t_{ij}} + \varphi_{L_i} e^{-\delta_L t_{ij}} - \delta_{Ab_i} Ab_i(t_{ij}) \\ Ab_i(t_{i0} = 0) & = & Ab_{0i} \end{cases}$$

$$\begin{cases} \log(\varphi_{S_i}) & = & \log(\varphi_{S_{pop}}) + \log(FC\Delta AGE = +1an) \times \left(AGE_i - \overline{AGE}\right) + \eta_i^S \\ \log(\varphi_{L_i}) & = & \log(\varphi_{L_{pop}}) + \beta_{G_1} \times G_{1i} + \eta_i^L \\ \log(\delta_{Ab_i}) & = & \log(\delta_{Ab_{pop}}) + \beta_{G_2} \times G_{2i} + \eta_i^{Ab} \end{cases} \qquad \text{(Mod Ab Str)}$$

- **Statistical Model** :

$$\forall i \leq N, \quad \begin{cases} \eta_i^S & \overset{iid}{\sim} & \mathcal{N}(0, \omega_S^2) \\ \eta_i^L & \overset{iid}{\sim} & \mathcal{N}(0, \omega_L^2) \\ \eta_i^{Ab} & \overset{iid}{\sim} & \mathcal{N}(0, \omega_{Ab}^2) \end{cases} \qquad \text{(Mod Ab Stat)}$$

- **Observation Model** :

$$\forall i \leq N, j \leq n_i, \quad \begin{cases} Y_{ij} & = & \log_{10}(Ab_i(t_{ij})) + \varepsilon_{ij} \\ \varepsilon_i & \overset{iid}{\sim} & \mathcal{N}(0, \Sigma = \sigma_{Ab}^2 I_{n_i}) \end{cases} \qquad \text{(Mod Ab Obs)}$$

The parameter values were based on estimates from the EBOVAC trial [Eurosurveillance editorial team, 2014, Alexandre et al., 2023, Pasin et al., 2019] and are provided in Table 2. Antibody concentrations were simulated at timepoints $t_j \in \{0, 7, 21, 123, 180, 300\}$.

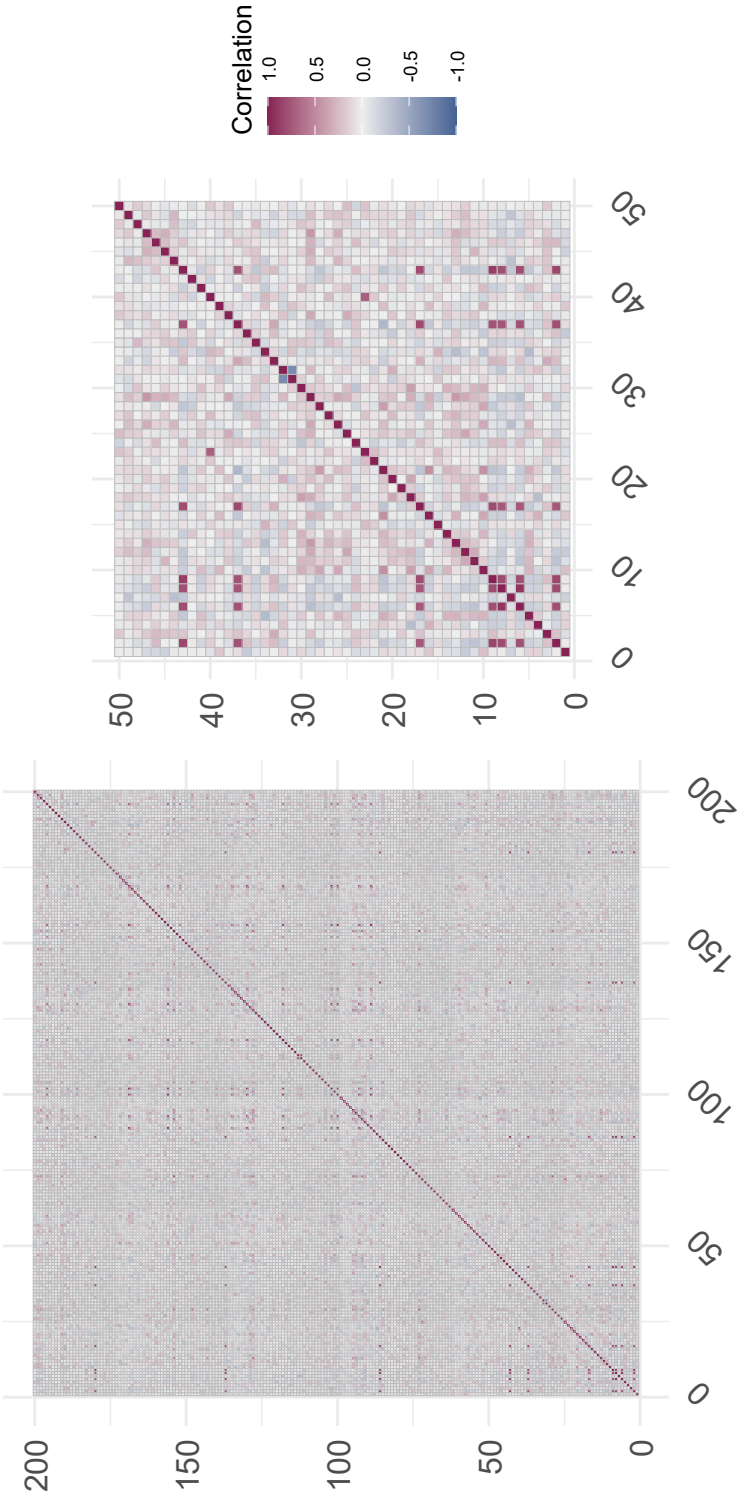
To create high-dimensional covariate settings, we introduced 200 covariates per individual, including 197 noise variables and 3 informative ones (AGE, G_1, G_2). Two covariate generation scenarios were considered: one with standard Gaussian covariates, and one with covariates drawn from randomly selected heterogeneous distributions. Details of the covariate simulation process are provided in Section A.3.

Parameter	Definition	Value
Fixed Effects		
Influx rates		
φ_S	ASC-S antibody production (Eu/mL/day)	3057 at age 35
FC_{AGE}	Fold change per year of age	0.934
φ_L	ASC-L antibody production (Eu/mL/day)	16.6 when $G_1 = 0$
β_{G_1}	Effect of G_1 on φ_L	1.2
Decay rates		
δ_{Ab}	Antibody decay rate (1/day)	0.0251 when $G_2 = 0$
β_{G_2}	Effect of G_2 on δ_{Ab}	0.8
δ_S	ASC-S decay rate	0.23
δ_L	ASC-L decay rate	$3.16 \cdot 10^{-4}$
Random Effects		
ω_S	SD for φ_S	0.92
ω_L	SD for φ_L	0.85
ω_{Ab}	SD for δ_{Ab}	0.30
Observation Error		
σ_{Ab}	SD of error on log antibody levels	0.11

Supplementary Table 2: Parameter values for the humoral immune response model based on the EBOVAC trial [Alexandre et al., 2023, Pasin et al., 2019]. *(FC: fold change; SD: standard deviation)*

A.3. Covariates Generation

To simulate realistic dependencies among covariates, we introduced correlation realistic dependencies observed in biomedical datasets. The theoretical correlation matrix used to generate the datasets in the vaccinology framework is shown in Figure 2.



Supplementary Figure 2: Correlation matrix used to generate the 100 datasets in the vaccinology simulation framework (left: full matrix; right: zoom on first 50 covariates).

In the first scenario, we generated all covariates from a multivariate normal distribution with the specified correlation structure using the `mvtnorm` package [Genz and Bretz, 2009] in R. In the second scenario, each covariate was drawn from a randomly selected distribution among a predefined set (Gaussian, Gamma, Poisson, Uniform). To automate this process, we implemented a function called `randomCovariate`, available on github `https://github.com/aurianegbt/LassoSAMBA`, which returns the name of a distribution and its associated parameters. The parameters are then sampled to define the distribution, and covariates are generated accordingly. Table 3 summarizes the rules used to define the parameters of each distribution.

To generate covariates with controlled correlations under the mixed-distribution scenario, we used the `simstudy` package [Goldfeld and Wujciak-Jens, 2020]. It enables conditional simulation of variables using a flexible table-based specification.

The data definition table includes the following columns:

- `varname`: name of the variable to simulate.
- `formula`: value or R formula (defining the mean or transformation).
- `variance`: variance or another scale parameter.
- `dist`: name of the distribution (e.g., "gamma", "norm").
- `link`: link function between the formula and distribution parameters (identity, log, or logit).

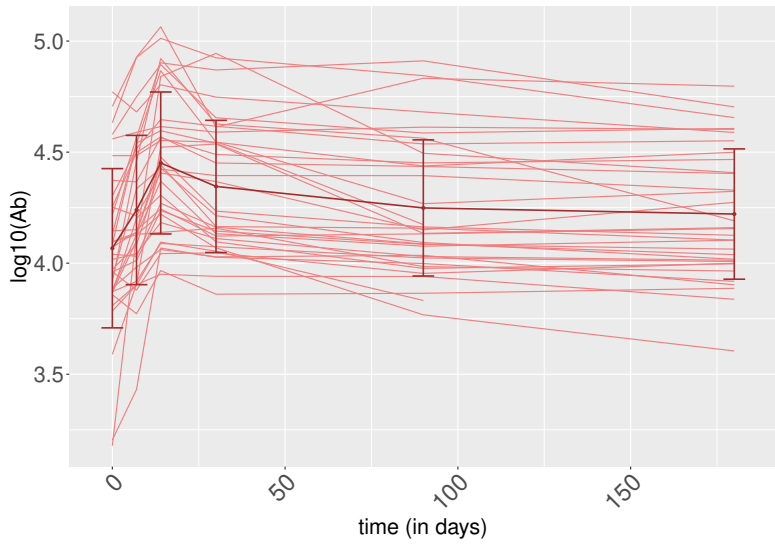
Each variable is conditionally generated using the values or distributions defined above, and the resulting dataset is aligned with the specified correlation matrix using `genCorGen()` from `simstudy`. This approach ensures diverse, realistic high-dimensional covariate structures while maintaining reproducibility.

Suffix for the distribution	**Element(s)**	**Generation**
Gamma distribution		
"gamma"	"shape"	$\mathscr{P}(10)$
	"scale"	$\mathscr{E}(1/2)$
Normal Distribution		
"norm"	"mean"	$\mathscr{N}(0,1)$
	"sd"	$\mathscr{E}(0.1)$
Poisson distribution		
"pois"	"lambda"	$\mathscr{E}(0.1)$
Uniform distribution		
"unif"		$x = \mathscr{N}(0_2, I_2)$
	"min"	$\min(x)$
	"max"	$\max(x)$

Supplementary Table 3: Process of generating various distributions using the implemented `randomCovariate` function.

B. Additional analysis for the real-data application

B.1. Initial model



Supplementary Figure 3: Antibody titre trajectories for the 35 individuals receiving the ZOSTAVAX vaccine.

To describe the antibody kinetics observed in Figure 3, we fit the humoral immune response model introduced by Pasin et al. [Pasin et al., 2019]. The model structure is as follows:

▸▸ **Structural Model** :

$$\forall i \leq N, j \leq n_i, \quad \begin{cases} \dfrac{d}{dt} Ab_i(t_{ij}) & = & \varphi_{S_i} e^{-\delta_{S_i} t_{ij}} + \varphi_{L_i} e^{-\delta_L t_{ij}} - \delta_{Ab} Ab_i(t_{ij}) \\ Ab_i(t_{i0} = 0) & = & Ab_{0i} \end{cases}$$

$$\begin{cases} \log(\varphi_{S_i}) & = & \log(\varphi_{S_{pop}}) + \eta_i^S \\ \log(\varphi_{L_i}) & = & \log(\varphi_{L_{pop}}) + \eta_i^L \\ \log(\delta_{S_i}) & = & \log(\delta_{S_{pop}}) + \eta_i^\delta \end{cases} \tag{1}$$

▸▸ **Statistical Model** :

$$\forall i \leq N, \quad \begin{cases} \eta_i^S & \overset{iid}{\sim} & \mathcal{N}(0, \omega_{\varphi_S}^2) \\ \eta_i^L & \overset{iid}{\sim} & \mathcal{N}(0, \omega_{\varphi_L}^2) \\ \eta_i^\delta & \overset{iid}{\sim} & \mathcal{N}(0, \omega_{\delta_S}^2) \end{cases} \tag{2}$$

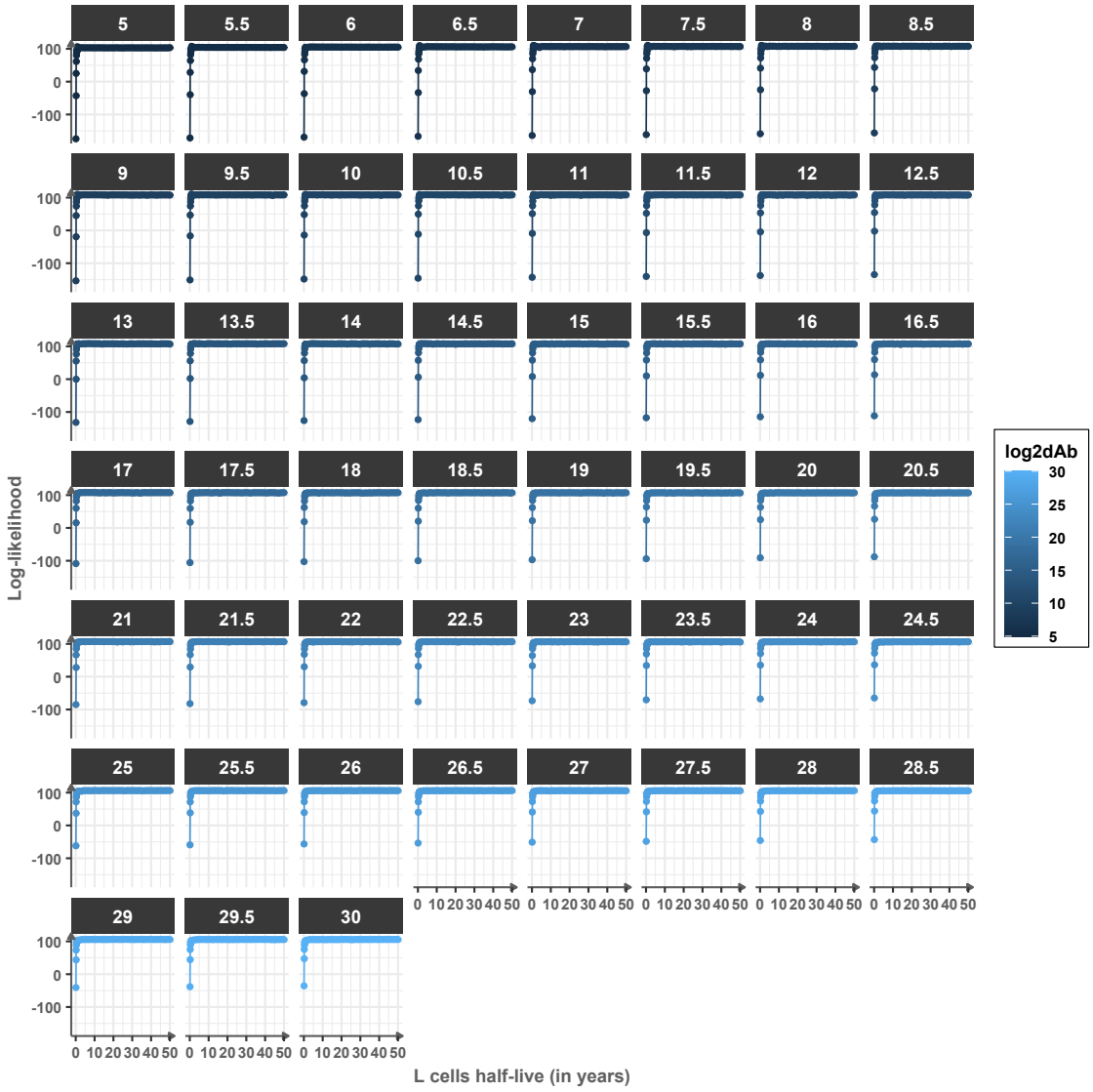
▸▸ **Observation Model** :

$$\forall i \leq N, j \leq n_i, \quad \begin{cases} Y_{ij} & = & \log_{10}(Ab_i(t_{ij})) + \varepsilon_{ij} \\ (\varepsilon_i j)_{j \leq n_i} & \overset{iid}{\sim} & \mathcal{N}(0, \Sigma = \sigma^2_{Ab}) \end{cases} \tag{3}$$

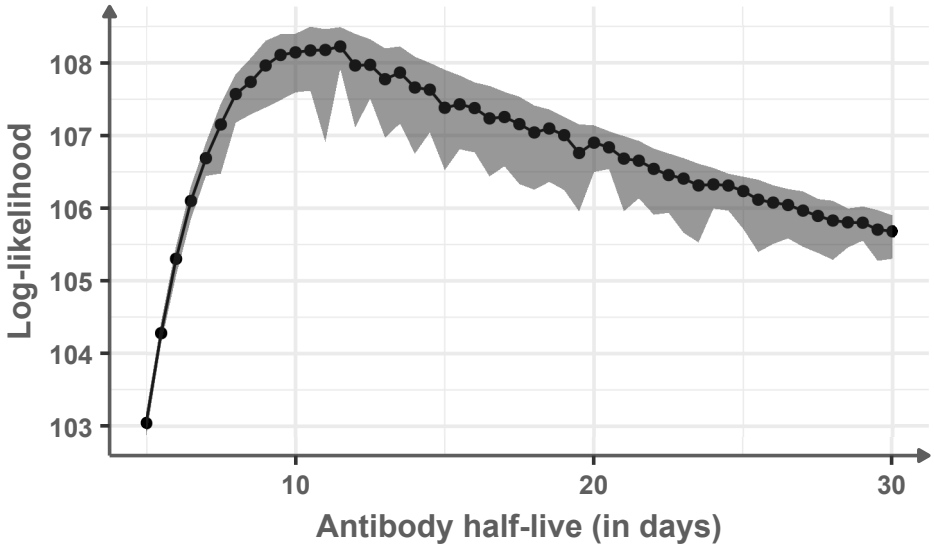
To resolve identifiability issues, we fixed the decay rates δ_L (half-life for long-lived ASCs) and δ_{Ab} (half-life for antibody), based on likelihood profile analyses shown in Figures 4 and 5.

Figure 4 shows that for long-lived ASCs, the log-likelihood stabilizes beyond a half-life of 5 years, indicating a lack of identifiability in the data. We therefore fixed $\delta_L = 0.00019$ day^{-1}, corresponding to a 10-year half-life, to avoid identifiability inversion between ASC-S and ASC-L components, given their symmetrical roles in the model.

Regarding the antibody half-life, Figure 5 displays the mean values and 95% confidence intervals for fixed antibody half-life values and L-cell half-life values exceeding 5 years. The log-likelihood reaches a maximum around 11 days of half-life. We thus fixed $\delta_{Ab} = 0.063$ day^{-1}. The remaining model parameters were estimated via the SAEM algorithm using Monolix, and are presented in Table 4.



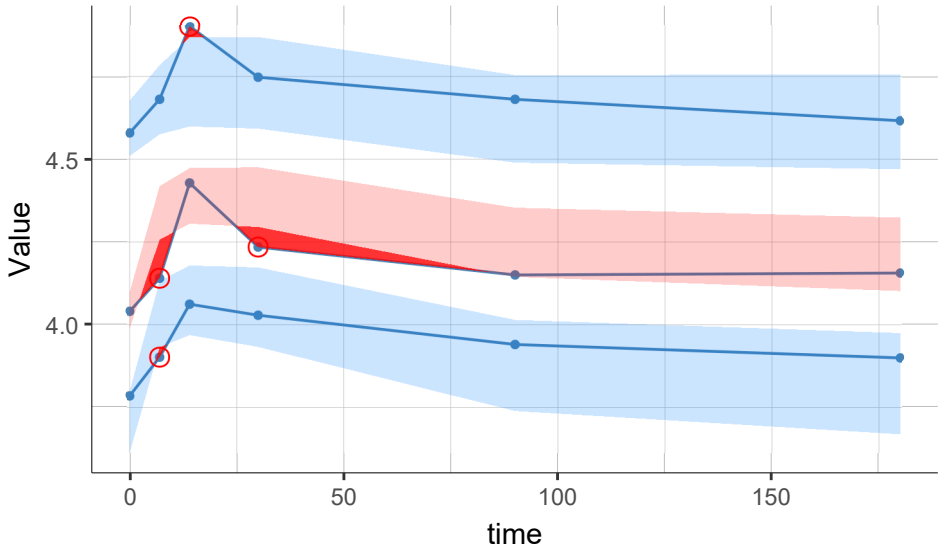
Supplementary Figure 4: Log-likelihood profile as a function of the L-cell half-life, with antibody half-life fixed (facets).



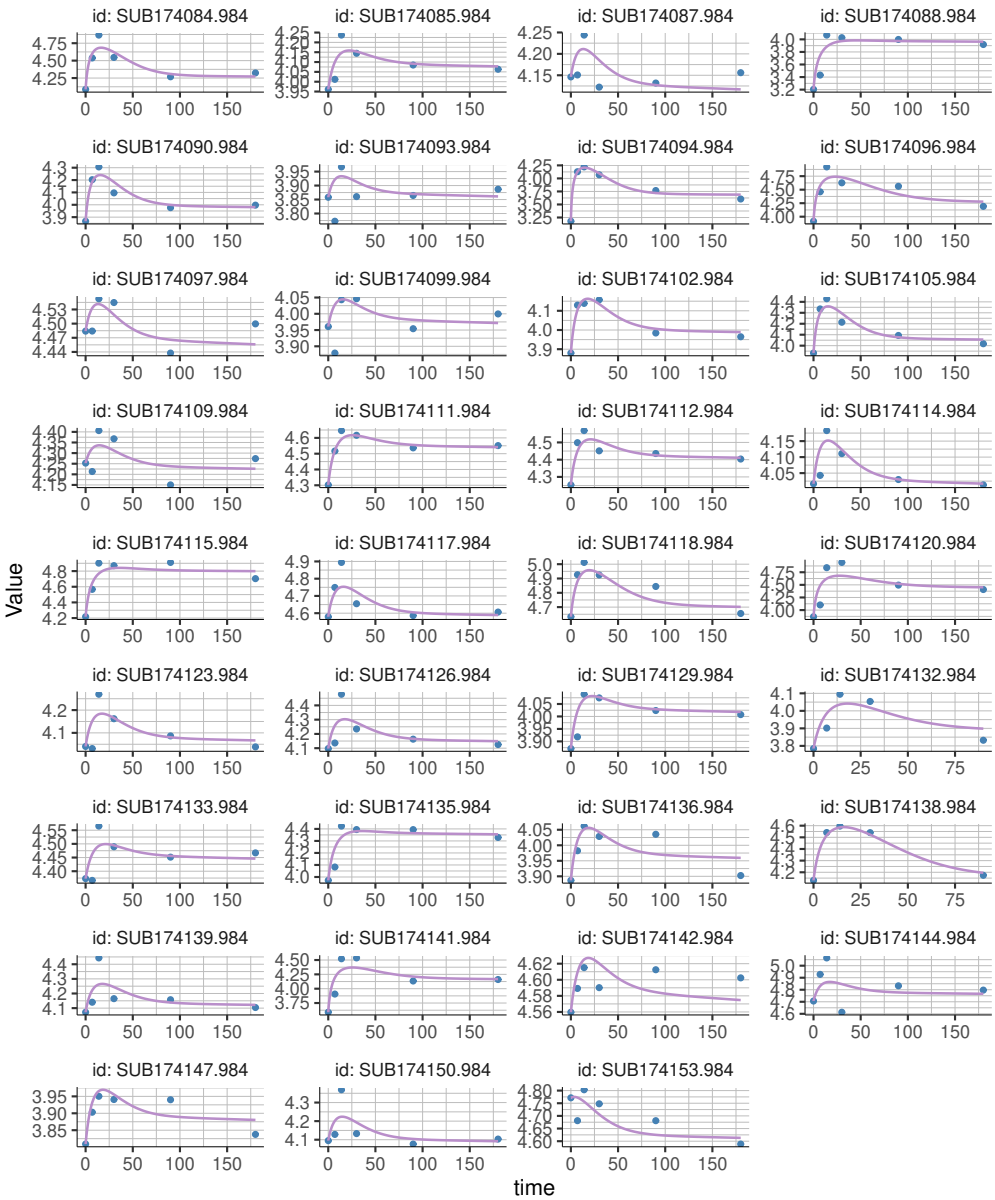
Supplementary Figure 5: Log-likelihood profile as a function of antibody half-life. The line and ribbon indicate mean and 95% confidence bounds over L-cell half-lives > 5 years.

Parameter	**Definition**	**Mean value (standard error)**
Fixed Effects		
Influx		
$\varphi_{\mathbf{Spop}}$	Antibodies influx of ASC S (Eu/mL/day)	907.35 (243.66)
$\varphi_{\mathbf{Lpop}}$	Antibodies influx of ASC L (Eu/mL/day)	1069.24 (127.66)
Decay rates		
$\delta_{\mathbf{Ab}}$	Decay rates of antibodies (per day)	0.063
$\delta_{\mathbf{Spop}}$	Decay rate of ASC S (per day)	0.058 (0.017)
$\delta_{\mathbf{L}}$	Decay rate of ASC L (per day)	1.9×10^{-4}
Random Effects		
$\omega_{\varphi_{\mathbf{S}}}$	Standard deviation of random effects for φ_S	1.16 (0.22)
$\omega_{\varphi_{\mathbf{L}}}$	Standard deviation of random effects for φ_L	0.67 (0.084)
$\omega_{\delta_{\mathbf{S}}}$	Standard deviation of random effects for δ_S	0.50 (0.20)
Error Model		
$\sigma_{\mathbf{Ab}}$	Standard deviation of error model	0.095 (0.0056)

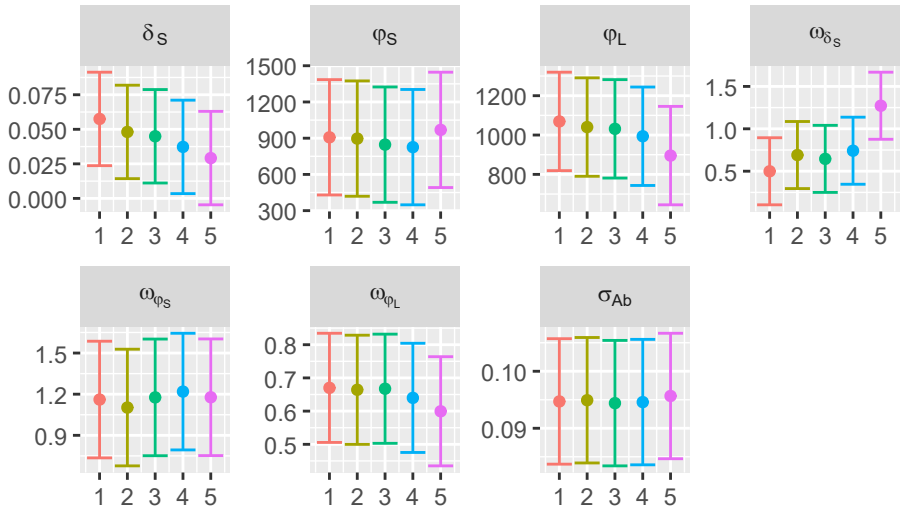
Supplementary Table 4: Estimation of the initial model parameters.



Supplementary Figure 6: Visual predictive check (VPC) of the initial model.



Supplementary Figure 7: Individual fits of the initial model across all 35 subjects.



Supplementary Figure 8: Convergence diagnostics for the initial model: population estimates across 5 independent Monolix runs with 95% CI.

B.2. Final Model

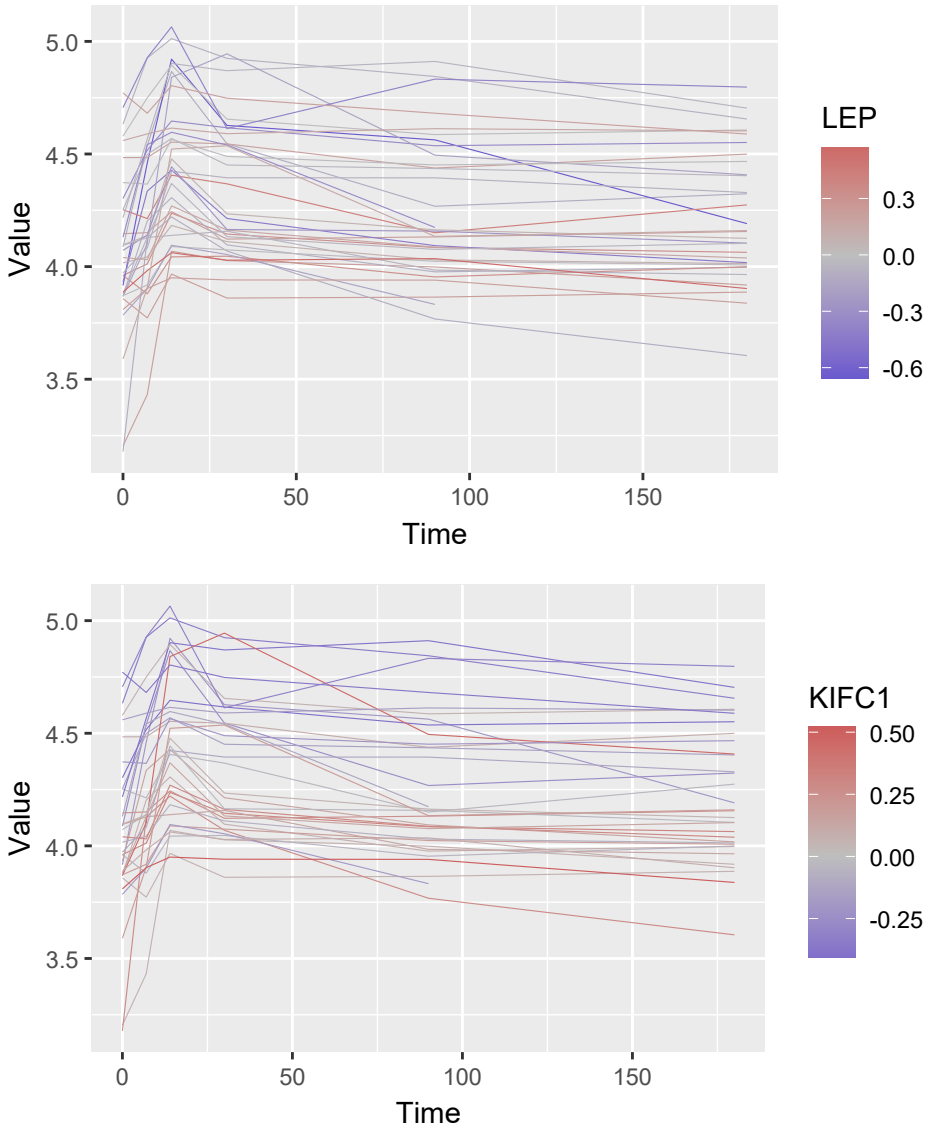
The final model was obtained using the lasso-SAMBA procedure, which selected two genes: **LEP**, associated with φ_S, and **KIFC1**, associated with φ_L:

$$\log(\varphi_{Si}) = \log(\varphi_{S_{pop}}) + \beta_{LEP} \times X_{LEPi} + \eta_i^S,$$

and

$$\log(\varphi_{Li}) = \log(\varphi_{L_{pop}}) + \beta_{KIFC1} \times X_{KIFC1i} + \eta_i^L.$$

Parameters were re-estimated via SAEM in Monolix, summarised in Table 5. Gene expression data were centered before estimation. Figure 9 displays the antibody data stratified by gene expression levels.

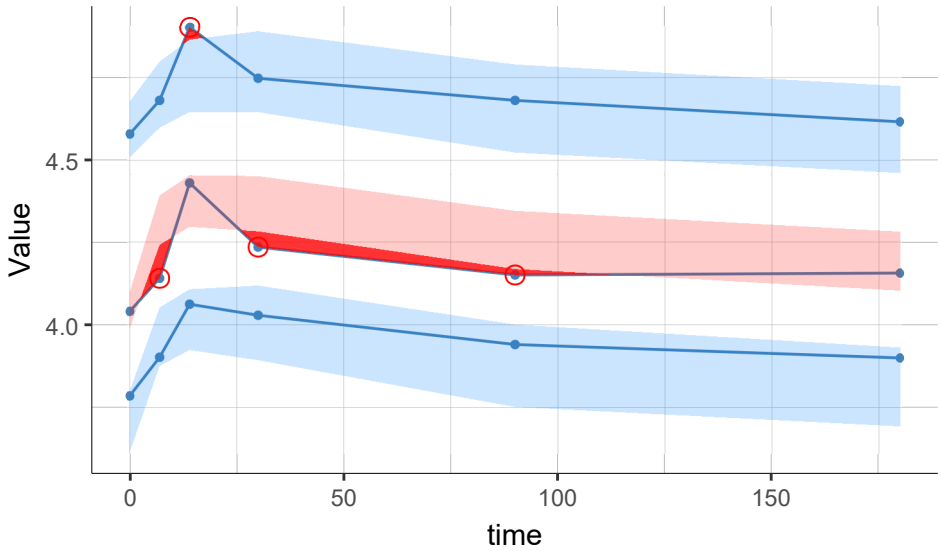


Supplementary Figure 9: Antibody trajectories stratified by gene expression levels of LEP (top) and KIFC1 (bottom).

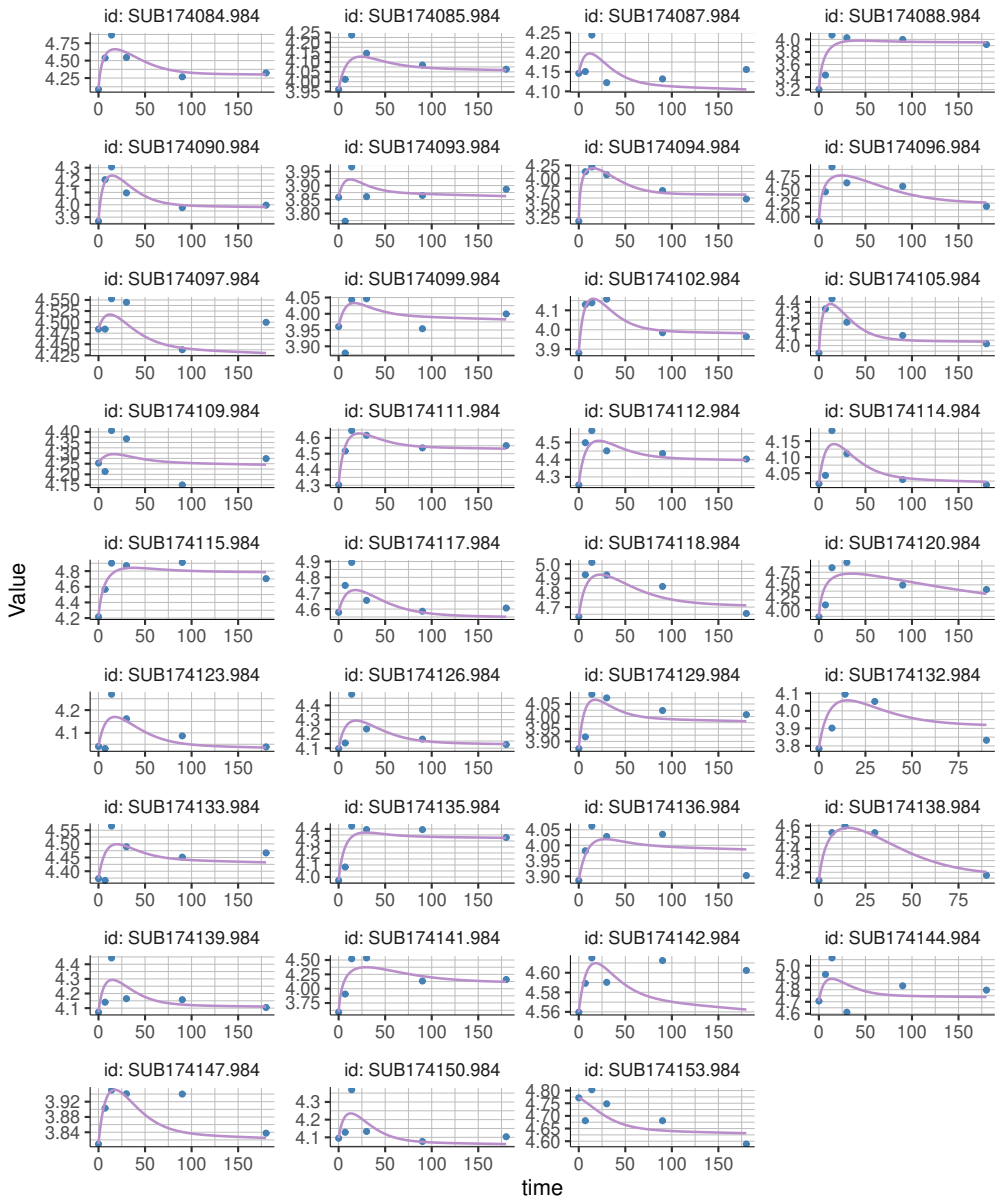
Parameter	Definition	Mean value (standard error)
Fixed Effects		
Influx		
$\varphi_{\mathbf{Spop}}$	Antibodies influx of ASC S (Eu/mL/day)	945.43 (193.78)
$\beta_{\mathbf{LEP}}$		-3.40 (0.71)
$\varphi_{\mathbf{Lpop}}$	Antibodies influx of ASC L (Eu/mL/day)	1007.21 (105.63)
$\beta_{\mathbf{KIFC1}}$		-2.05 (0.43)

Decay rates		
$\delta_{\mathbf{Ab}}$	Decay rates of antibodies (per day)	0.063
$\delta_{\mathbf{Spop}}$	Decay rate of ASC S (per day)	0.047 (0.018)
$\delta_{\mathbf{L}}$	Decay rate of ASC L (per day)	1.9×10^{-4}
Random Effects		
$\omega_{\varphi_{\mathbf{S}}}$	Standard deviation of random effects for φ_S	0.83 (0.15)
$\omega_{\varphi_{\mathbf{L}}}$	Standard deviation of random effects for φ_L	0.66 (0.089)
$\omega_{\delta_{\mathbf{S}}}$	Standard deviation of random effects for δ_S	0.72 (0.27)
Error Model		
$\sigma_{\mathbf{Ab}}$	Standard deviation of error model	0.094 (0.0056)

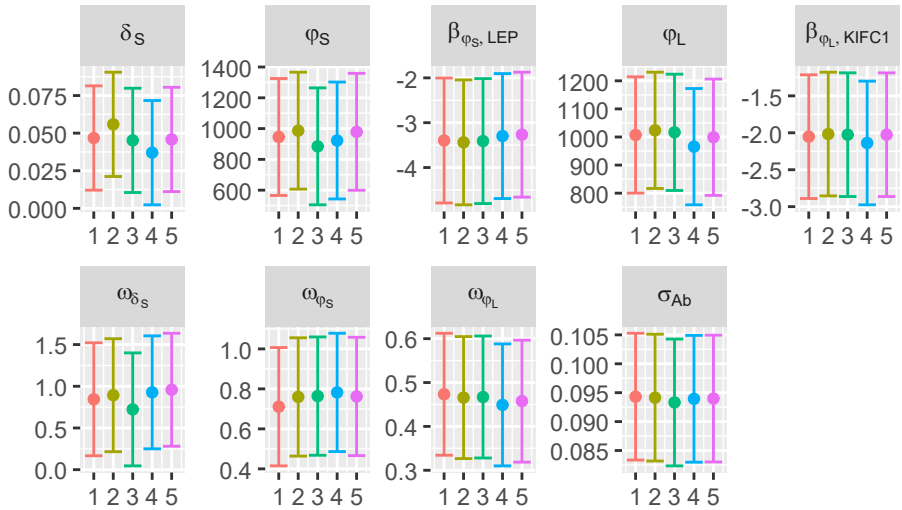
Supplementary Table 5: Estimation of the final model parameters.



Supplementary Figure 10: Visual predictive check of the final estimated model for real-data illustration.

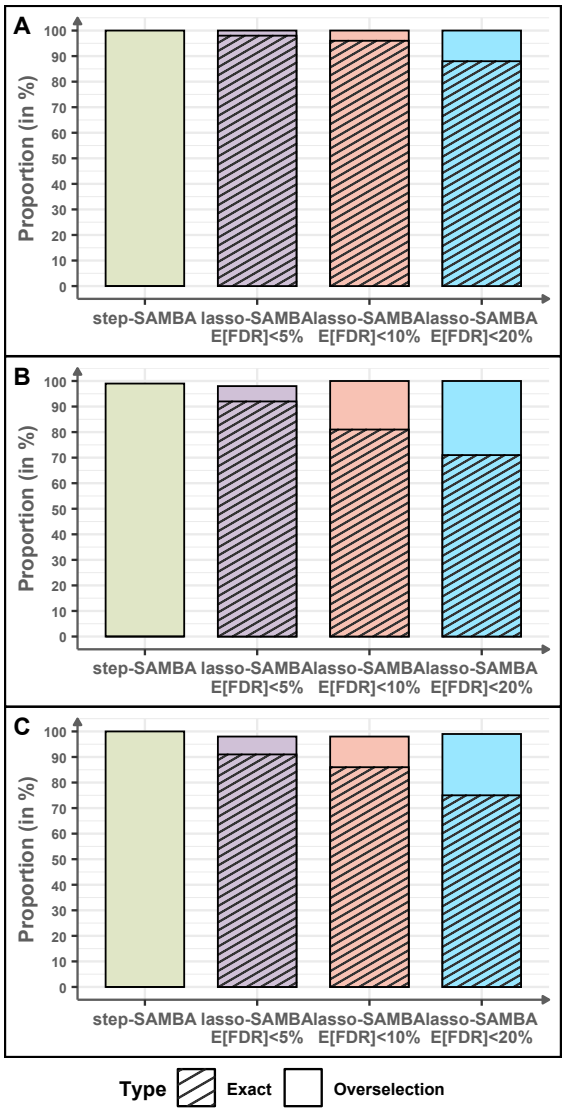


Supplementary Figure 11: Individual fits of the final estimated model for the real-data illustration.



Supplementary Figure 12: Convergence assessment plot with estimated population parameters and 95% confidence interval for 5 runs through monolix software, for the final model.

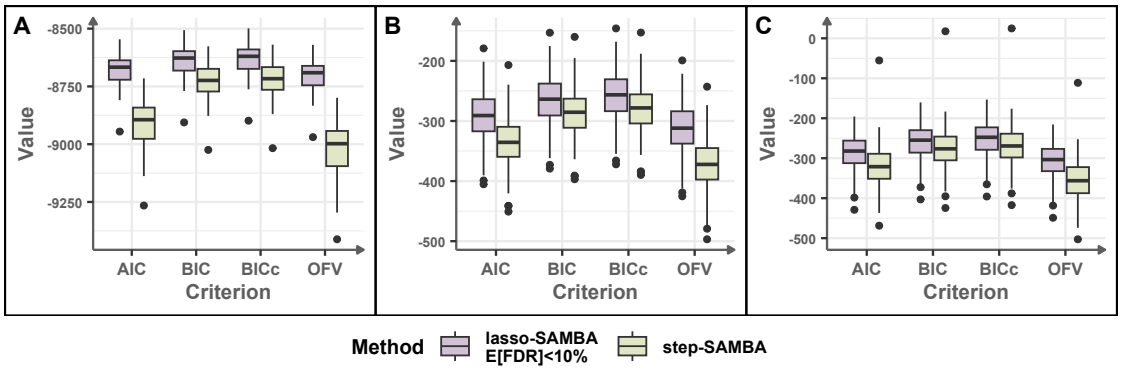
C. Additional graphs for results



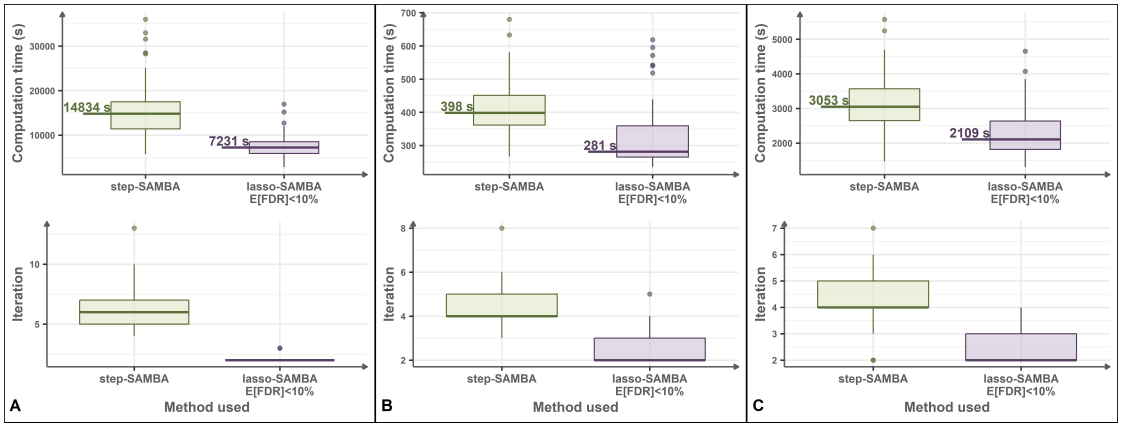
Supplementary Figure 13: Proportion of exact models (no false positives or false negatives) and models that strictly include the exact model (no false negatives but with false positives) for step-SAMBA and lasso-SAMBA, using different control thresholds. (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

Rate	**PK model** binomial covariates	**Vaccinology model** Gaussian covariates	**Vaccinology model** non-Gaussian covariates
FDR (%) :			
- step-SAMBA	87.5 [78.8;91.7]	72.7 [50.0;83.8]	70.0 [25.0;81.2]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.05$	0.0 [0.0;0.0]	0.0 [0.0;25.0]	0.0 [0.0;25.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.1$	0.0 [0.0;14.3]	0.0 [0.0;25.0]	0.0 [0.0;25.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.2$	0.0 [0.0;14.3]	0.0 [0.0;40.0]	0.0 [0.0;25.0]
- SAEMVS	0.0 [0.0;100.0]	ND	ND
FNR (%) :			
- step-SAMBA	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.05$	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.1$	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.2$	0.0 [0.0;0.0]	0.0 [0.0;0.0]	0.0 [0.0;0.0]
- SAEMVS	0.0 [0.0;0.6]	ND	ND
F1-score (%) :			
- step-SAMBA	22.2 [15.4;34.9]	42.9 [27.9;66.7]	46.2 [31.6;85.7]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.05$	100.0 [100.0;100.0]	100.0 [85.7;100.0]	100.0 [85.7;100.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.1$	100.0 [92.3;100.0]	100.0 [85.7;100.0]	100.0 [85.7;100.0]
- lasso-SAMBA $U_{\lambda,\pi} \leq 0.2$	100.0 [92.3;100.0]	100.0 [75.0;100.0]	100.0 [85.7;100.0]
- SAEMVS	100.0 [0.0;100.0]	ND	ND

Supplementary Table 6: False Discovery Rate (FDR), False Negative Rate (FNR), and $F1$-score computed for each simulation framework, comparing the original step-SAMBA method and the lasso-SAMBA method, with various error control thresholds. (ND: not defined).



Supplementary Figure 14: Distribution of final model selection scores (AIC, BIC, BICc, OFV) across replicates in each framework, with the step-SAMBA method and the lasso-SAMBA method, with error control thresholds set to 10%. (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.



Supplementary Figure 15: Comparison of computation time distributions across simulation frameworks, with the step-SAMBA method and the lasso-SAMBA method, setting the error control thresholds to 10%. (A) Pharmacokinetics model with categorical covariates; (B) Vaccinology framework with Gaussian-correlated covariates; (C) Vaccinology framework with randomly drawn covariates.

References

[Alexandre et al., 2023] Alexandre, M., Prague, M., McLean, C., Bockstal, V., Douoguih, M., Thiébaut, R., Effelterre, T., Solforosi, L., and Dari, A. (2023). Prediction of long-term humoral response induced by the two-dose heterologous ad26.zebov, mva-bn-filo vaccine against ebola. *npj Vaccines*, 8.

[Eurosurveillance editorial team, 2014] Eurosurveillance editorial team, C. (2014). The innovative medicines initiative launches call on ebola and other filoviral haemorrhagic fevers. *Eurosurveillance*, 19(45).

[Genz and Bretz, 2009] Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

[Goldfeld and Wujciak-Jens, 2020] Goldfeld, K. and Wujciak-Jens, J. (2020). simstudy: Illuminating research methods through data generation. *Journal of Open Source Software*, 5(54):2763.

[Naveau et al., 2023] Naveau, M., Kon Kam King, G., Rincent, R., Sansonnet, L., and Delattre, M. (2023). Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the saem algorithm. *Statistics and Computing*, 34(1).

[Pasin et al., 2019] Pasin, C., Balelli, I., Van Effelterre, T., Bockstal, V., Solforosi, L., Prague, M., Douoguih, M., and Thiébaut, R. (2019). Dynamics of the Humoral Immune Response to a Prime-Boost Ebola Vaccine: Quantification and Sources of Variation. *Journal of Virology*, 93(18):e00579–19.

[R Core Team, 2024] R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.