Comput Stat (2017) 32:803–843
DOI 10.1007/s00180-017-0730-6

ORIGINAL PAPER

A sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution

Vinny Davies[1] · Richard Reeve[2,3] · William T. Harvey[2,3] · Francois F. Maree[4] · Dirk Husmeier[1]

Received: 12 August 2015 / Accepted: 25 April 2017 / Published online: 13 June 2017

Abstract Understanding how viruses offer protection against closely related emerging strains is vital for creating effective vaccines. For many viruses, multiple serotypes often co-circulate and testing large numbers of vaccines can be infeasible. Therefore the development of an in silico predictor of cross-protection between strains is important to help optimise vaccine choice. Here we present a sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution (SABRE) which can account for the experimental variability in the data and predict antigenic variability. The method uses spike and slab priors to identify sites in the viral protein which are important for the neutralisation of the virus. Using the SABRE method we are able to identify a number of key antigenic sites within several viruses, as well as providing estimates of significant changes in the evolutionary history of the serotypes. We show how our method outperforms alternative established methods; standard mixed effects models, the mixed effects LASSO, and the mixed effects elastic nets. We also propose novel proposal mechanisms for the Markov chain Monte Carlo simulations, which

Electronic supplementary material The online version of this article (doi:10.1007/s00180-017-0730-6) contains supplementary material, which is available to authorized users.

✉ Vinny Davies
V.Davies@leeds.ac.uk

Dirk Husmeier
Dirk.Husmeier@glasgow.ac.uk

1 School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

2 Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow, Glasgow, UK

3 Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK

4 Transboundary Animal Diseases Programme, Onderstepoort Veterinary Institute, Onderstepoort, South Africa

improve mixing and convergence over that of the established component-wise Gibbs sampler.

Keywords Spike and slab prior · Foot-and-mouth disease virus · Influenza virus · Antigenic variability · Bayesian hierarchical models · Mixed-effects models · LASSO · Markov chain Monte Carlo

1 Introduction

Ribonucleic acid (RNA) viruses such as Foot-and-mouth disease virus (FMDV) and Influenza A (H1N1) have been shown to have high genetic variability (Holland et al. 1982). This variability results in changes to the virus proteins that effect recognition by the host immune system, also known as antigenic differences. Differences in these proteins, also known as antigenic proteins, affect how antigenically similar different viruses are. As a consequence of the antigenic variability in the viruses, vaccines are only effective against field strains that are genetically related and antigenically similar to the vaccine strain (Mattion et al. 2004). This feature of FMDV and Influenza makes it important to estimate antigenic similarity among strains and therefore cross-protection, the protection against one strain conferred by previous exposure to another strain by either infection or vaccination (Paton et al. 2005).

RNA virus are classified into serotypes, genetically and antigenically distinct virus lineages between which there is no effective degree of cross-protection. Individual vaccines may protect against large groups of genetically diverse viruses within a serotype, however there are antigenically distinct subtypes against which the vaccines do not work. In FMDV, South African Territories types 1 and 2 (SAT1 and SAT2) are responsible for the majority of FMDV outbreaks in cattle in the region, while the H1N1 virus has been responsible for several major flu outbreaks; Spanish Flu in 1918 and Swine Flu in 2009. Within these serotypes of FMDV and Influenza are significant levels of antigenic variability, which allows us to examine the relationship between genetic and antigenic variation and to determine which protein changes affect recognition by the immune system. Given the importance of these serotypes in the region and the difficulties with vaccination caused by antigenic variation, it is vital to understand how genetic changes affect antigenicity and within-serotype cross-protection.

In the outer capsid or virus shell, proteins influence antigenicity. Many areas of these proteins are exposed on the surface of the capsid and among these are antigenic regions that are recognised by the host immune system. Single amino acid substitutions (mutations) within these antigenic regions can dramatically affect recognition by the immune system. Identifying the specific amino acid residues that comprise these antigenic regions and the substitutions that cause antigenic differences is critical to understanding antigenic similarity among viruses and cross-protection within serotypes. In the FMDV serotypes, both SAT1 and SAT2 are known to share one major antigenic region, the highly flexible VP1 G-H loop (Crowther et al. 1993b), a cord of connected amino acid residues. Additional residues have also been identified on the SAT serotypes (Crowther et al. 1993b; Grazioli et al. 2006), as well as others on related serotypes A, O, C and Asia1 which may also occur within the SAT serotypes (Grazioli

et al. 2013; Kitson et al. 1990; Lea et al. 1994; Saiz et al. 1991). For the H1N1 virus, experimental studies have identified four major antigenic sites (Caton et al. 1982), as well as a number of other sites known to be important (McDonald et al. 2007).

Changes in the antigenic proteins occur as the strains within each serotype evolve. The accumulation of these changes in geographically isolated virus lineages allows for the division of serotypes into topotypes, groups of genetically similar viruses associated with a particular geographic area (Knowles and Samuel 2003). Strains within topotypes share a common evolutionary history that is distinct from strains within other topotypes. Accounting for the genetic differences between topotypes that have arisen due to their significantly different evolutionary paths is necessary for understanding antigenic variability (Reeve et al. 2010). Interpreting the antigenic consequences of genetic differences between topotypes can improve our understanding of the evolutionary history of serotypes, as well as the likely extent of vaccine coverage across topotypes.

In order to infer the antigenic importance of specific genetic changes that have occurred during the evolution of the virus, we require in addition to genetic data, a measure of the antigenic similarity of any two virus strains. Virus Neutralisation (VN) titre and Haemagglutination inhibition (HI) assay give in vitro measures of antigenic similarity between a protective, i.e. a potential vaccine, and a challenge strain, i.e. a potential circulating virus (Hirst 1942; WHO 2011). They approximate the extent to which one strain confers protection against another by recording the maximum dilution at which the virus-specific antibody in a sample of antiserum from a cow (VN titre) or ferret (HI assay) exposed to one strain of the virus (the protective strain) remains able to neutralise a sample of a second virus strain (the challenge strain). Higher titres or assay measures indicate that the antiserum still neutralises the challenge strains at greater dilution and therefore that the protective and challenge strains are more antigenically similar.

In principle, it is possible to identify experimentally how a mutation of the residues affects antigenicity. However, due to the large number of co-circulating virus strains, this is time consuming and expensive. Developing in silico predictors of VN titre and HI assay that are robust and can account for experimental variation can help substantially reduce the number of strains that must be tested in order to select an effective vaccine strain. Previously mixed-effects models have been used on a variety of datasets to model antigenic variability in both FMDV and Influenza by accounting for the experimental variation in the VN titre or HI assay measurements (Reeve et al. 2010, 2016; Harvey et al. 2016).

However in order to identify antigenically important residues, we must infer which explanatory variables are selected in the model. While stepwise regression techniques, such as that of Reeve et al. (2010), can be used to select variables within standard mixed-effects models, they do not explore all variable configurations and can result in a non-optimal solution. An improved method, which allows for simultaneous variable selection, is the Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996), which uses an ℓ_1 penalty to select the variables. Schelldorfer et al. (2011) have recently extended the LASSO to mixed-effects models with a single random effect and we further extend the method here to work with multiple random effects and the elastic net penalty (Zou and Hastie 2005).

A drawback of the LASSO and elastic net, is the ℓ_1 regularisation term itself, equivalent to a Laplace prior in a Bayesian context (Park and Casella 2008). This is computationally efficient and leads to a convex optimisation problem for penalised maximum likelihood or Bayesian maximum a posteriori (MAP) inference. However, ℓ_1 regularisation gives an increased bias from shrinkage while not giving sufficient sparsity, as discussed in Chapter 13 of Murphy (2012).

Spike and slab priors, as proposed in Mitchell and Beauchamp (1988), improve variable selection and avoid excessive shrinkage, but lead to a non-convex optimisation problem. The performance improvement of spike and slab priors over ℓ_1 regularisation methods has previously been reported (Mohamed et al. 2012). Here these priors are integrated into a Bayesian hierarchical mixed-effects model and this has a number of advantages. In particular Bayesian hierarchical models allow consistent inference of all parameters and hyper-parameters, and inference borrows strength by the systematic sharing and combination of information; see Gelman et al. (2013).

The previously proposed models used datasets containing a variety of explanatory variables in order to explain variation in VN titre and HI assay measurements. The datasets, which we also use here, include variables showing the presence or absence of amino acid substitutions (changes in protein composition caused by genetic mutations) at different residues on the surface of the virus. The selection of these variables within any model then indicates the relevance of that residue in determining the antigenic similarity of the virus strains. Additionally, the datasets include variables to correct for the phylogenetic structure of the serotypes, in order to account for the shared evolutionary history of the strains. All the previous methods have used mixed-effects models in order to correct for the experimental variation associated with the data collection. The random effects include information about the strains tested, the animal from which the serum was taken and when the lab work was completed and by whom.

Using these variables and random effects, Reeve et al. (2010) identified a single known antigenic residue, VP3 138, in the VP1 G-H loop from a relatively small amount of SAT1 data. More recently Reeve et al. (2016) explored an extended SAT1 dataset, which includes an increased number of strains and repeated experiments, and identified a number of known antigenic residues. Reeve et al. (2010) also used their methods on a small SAT2 dataset but were unable to select any significant residues in their model. Using a larger H1N1 dataset Harvey et al. (2016) identified a number of known antigenic site from the four major antigenic regions and other known sites.

The main purpose of this paper is to develop a Sparse hierArchical Bayesian model for detecting Relevant antigenic sites in virus Evolution (SABRE) and use it to analyse the SAT1, SAT2 and H1N1 datasets from Reeve et al. (2010, 2016) and Harvey et al. (2016). We propose and evaluate three different versions of the SABRE method and use them to identify a number of previously unidentified residues, as well as predicting a number of sites that could be plausibly antigenic. Moreover we use the data to investigate the antigenic changes in the evolutionary history of the different lineages within the FMDV serotypes and we are able to make reasonable predictions based on biologically estimated topotypes. In addition to these predictions of antigenic change, we also propose a new method for understanding when non-antigenic changes occur in the evolution of the virus.

2 Classical methods

A variety of classical statistical methods have previously been applied in predicting antigenically significant sites. Before we introduce the SABRE method, we review some of these methods and propose variations which are applicable in this context. For all of the methods we use the following notation; bold upper case letter, $\mathbf{X}$, for a matrix, bold lower case letter, $\mathbf{x}$, for a column vector, and non-bold letter, x, for a scalar. We do not distinguish between random variables and their realisations.

For further notational details see Tables 1 and 2 in the online supplementary materials.

2.1 Classical mixed-effects model

We define the response $\mathbf{y} = (y_1, \ldots, y_N)^\top$ to be the log VN titre or log HI assay and denote the explanatory variables, $\mathbf{X}$, to be indicators of mutational changes at different residues and information on the phylogenetic structure (see Sect. 4). The explanatory variables, $\mathbf{X}$, are given as a matrix of $J+1$ columns and N rows, where the first column is a column full of ones for the intercept. Each column j of explanatory variables, $\mathbf{x}_j$, is then given an associated regression coefficient, w_j, to control its influence on the response.

We further set the random-effects design matrix, $\mathbf{Z}$, as the matrix of indicators with N rows and $||\mathbf{b}||$ columns, where $||.||$ indicates the length of the vector. The random-effects design matrix describes experimental conditions which must be accounted for based on the information that is available (see Sect. 4). For our datasets this can include information about the strains tested, the animal from which the serum was taken and when the lab work was completed. The random-effects coefficients are given as $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_G^\top)^\top$, with each $\mathbf{b}_g$ relating to a group $g \in \{1, \ldots, G\}$, e.g. challenge strain. Each $\mathbf{b}_g$ has length $||\mathbf{b}_g||$ and follows a zero mean Gaussian distribution with a group dependent variance, $\mathbf{b}_g \sim \mathcal{N}(\mathbf{b}_g|\mathbf{0}, \sigma_{b,g}^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix. This leads to the random-effects coefficients having the following joint distribution $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$. Here we define $\boldsymbol{\Sigma}_\mathbf{b}$ to be a matrix with $\boldsymbol{\sigma}_\mathbf{b}^2 = (\sigma_{b,1}^2, \ldots, \sigma_{b,1}^2, \sigma_{b,2}^2, \ldots, \boldsymbol{\sigma}_{b,G}^2)^\top$ on the diagonal, i.e. $\boldsymbol{\Sigma}_\mathbf{b} = diag(\boldsymbol{\sigma}_\mathbf{b}^2)$. In this case each $\sigma_{b,g}^2$ is repeated with length $||\mathbf{b}_g||$. See Pinheiro and Bates (2000) for more details on mixed-effects models.

We define the model as:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{\varepsilon}|\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}\right) \tag{1}$$

assigning the model additive iid Gaussian errors. Using a simple application of Gaussian integrals (Bishop 2006), we integrate over $\mathbf{b}$ to give the likelihood:

$$L\left(\mathbf{w}, \sigma_\varepsilon^2, \boldsymbol{\Sigma}_\mathbf{b}|\mathbf{y}, \mathbf{X}, \mathbf{Z}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{X}\mathbf{w}, \mathbf{Z}\boldsymbol{\Sigma}_\mathbf{b}\mathbf{Z}^\top + \sigma_\varepsilon^2\mathbf{I}\right). \tag{2}$$

In classical mixed-effects models, model comparison techniques are often used to choose which variables are included within the model. Reeve et al. (2010) used a mix-

ture of forward inclusion and univariate analysis, making an adjustment for multiple testing using the Holm–Bonferroni correction to ensure a sparse model (Holm 1979). They firstly included terms to account for the evolutionary history of the viruses (see Sect. 4). They then did a univariate test for significance on the residue variables, where a p value of <0.05 corresponded to an antigenically important residue.

2.2 LASSO

A problem with the classical mixed-effects models of Reeve et al. (2010) is their model selection technique, which does not explore all variable configurations and can result in a non-optimal solution. A classical alternative which does allow for simultaneous variable selection is the LASSO of Tibshirani (1996, 2011). The LASSO achieves its variable selection through an ℓ_1 penalty (equivalent to a Bayesian Laplace prior). In the simplest case of linear regression, this gives the following parameter estimates:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \lambda \sum_{j=1}^{J} |w_j| \right\} \tag{3}$$

where we do not penalise the intercept w_0 so that the model remains scale invariant (Hastie et al. 2009). This is a convex optimisation problem where a variety of fast and effective algorithms exist (e.g. Efron et al. 2004; Hastie et al. 2009). The effect of (3) is to simultaneously shrink and prune parameters $\mathbf{w}$, thereby promoting a sparse model. The degree of sparsity depends on the regularisation parameter λ, which can be optimised with cross-validation or information criteria.

A recent extension of the standard LASSO is the mixed-effects LASSO proposed by Schelldorfer et al. (2011), who marginalised over $\mathbf{b}$ to estimate the regression coefficients $\mathbf{w}$, random-effects variances $\boldsymbol{\sigma}_{\mathbf{b}}^2$ and the variance of the noise σ_ε^2 as:

$$\left(\hat{\mathbf{w}}, \hat{\boldsymbol{\sigma}}_{\mathbf{b}}^2, \hat{\sigma}_\varepsilon^2\right) = \underset{\mathbf{w}, \boldsymbol{\sigma}_{\mathbf{b}}^2 > 0, \sigma_\varepsilon^2 > 0}{\operatorname{argmin}} \left\{ \log |\mathbf{V}| + (\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \sum_{j=1}^{J} |w_j| \right\} \tag{4}$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I}$. For our study we choose the penalty parameter λ based on the Bayesian Information Criterion (BIC), as in Schelldorfer et al. (2011).

A problem with the mixed-effects LASSO of Schelldorfer et al. (2011) is that the method has only been developed for one random effect. While it is possible to map the Cartesian product of several random effects onto a single random effect, doing so can lead to over-estimating the complexity of the model. We have therefore developed our own mixed-effects LASSO which is able to handle multiple random effects. Our method uses a conjugate gradient optimisation strategy available in *R* (R Core Team 2013), but requires a tolerance that must be determined by the user. In practise defining this tolerance is easy to do, as for a large λ and standardised data there will be a group of regressors clearly grouped around

zero. The tolerance can then be set such as to force these values to zero, i.e. exclusion from the model, and then λ reduced to create the so called LASSO path. Following the described optimisation strategy may not be as computationally efficient as the purpose-built block coordinate descent scheme proposed in Schelldorfer et al. (2011), but we have found in practise that they achieve the same results.

2.3 Elastic net

A potential improvement over the LASSO is the elastic net of Zou and Hastie (2005). It has several advantages including the ability to select more than N variables in a $J > N$ situation, whereas the LASSO saturates to at most N variables. More importantly for our application is that it also deals better with groups of correlated variables. While the LASSO will arbitrarily select one of the correlated variables, the penalty of the elastic net allows it to keep all of the variables in the model. See Section 2.3 of Zou and Hastie (2005) for more information on the grouping effect.

The elastic net combines ℓ_1 and ℓ_2 penalties and in the case of linear regression gives the following parameter estimates:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \alpha\lambda \sum_{j=1}^{J} |w_j| + (1-\alpha)\lambda \sum_{j=1}^{J} |w_j|^2 \right\} \tag{5}$$

where λ is the penalty parameter and α controls the ratio of the ℓ_1 and ℓ_2 penalties. When $\alpha = 1$ the elastic net is equivalent to the LASSO and likewise ridge regression when $\alpha = 0$. We have investigated values of α between 0.2 and 0.8 and found the results varied across different datasets and did not show a clear best choice of α; see Section 8 of the online supplementary materials. We have therefore set $\alpha = 0.3$ for the examples in the main paper following Ruyssinck et al. (2014) and relegated the remaining results to the supplementary materials for conciseness; see Sect. 6.1.

Like the LASSO, we can expand the elastic net into the context of a mixed-effects model, something that we propose here:

$$\left(\hat{\mathbf{w}}, \hat{\boldsymbol{\sigma}}_{\mathbf{b}}^2, \hat{\sigma}_{\varepsilon}^2\right) = \underset{\mathbf{w}, \boldsymbol{\sigma}_{\mathbf{b}}^2 > 0, \sigma_{\varepsilon}^2 > 0}{\operatorname{argmin}} \left\{ \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\mathbf{w})^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \alpha\lambda \sum_{j=1}^{J} |w_j| + (1-\alpha)\lambda \sum_{j=1}^{J} |w_j|^2 \right\} \tag{6}$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{Z}^\top + \sigma_{\varepsilon}^2\mathbf{I}$. We use the same optimisation strategy we proposed for the mixed-effects LASSO.

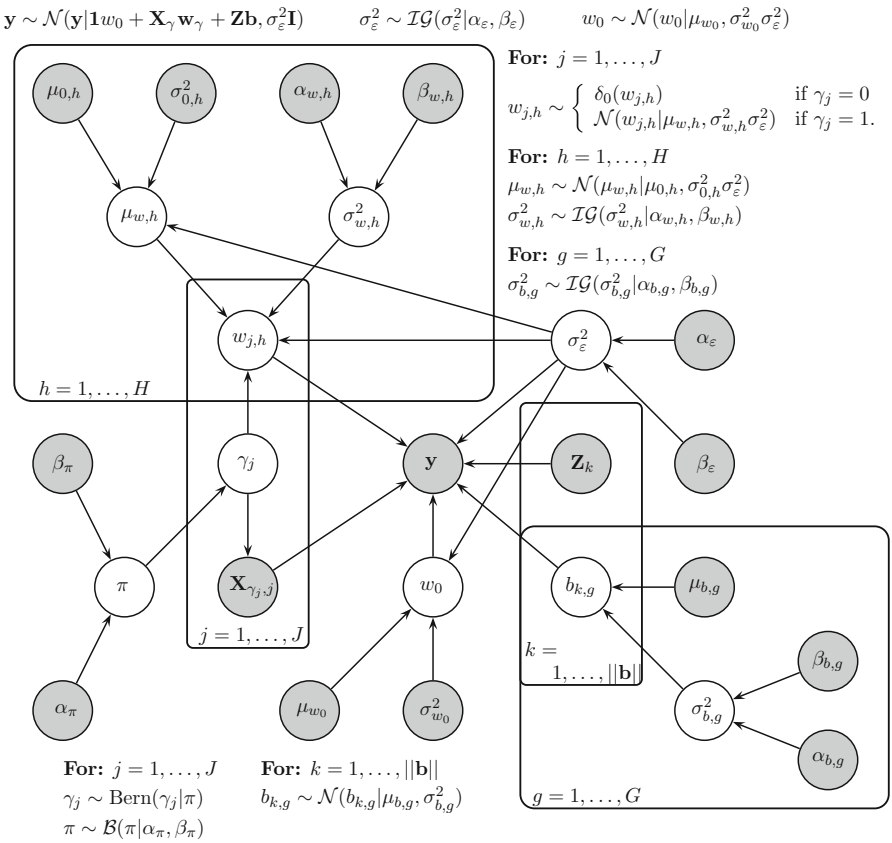

Fig. 1 Compact representation of the conjugate SABRE method as a PGM. The *grey circles* refer to the data and fixed (higher-order) hyperparameters, while the *white circles* refer to parameters and hyperparameters that are inferred. PGMs for the semi-conjugate and conjugate binary mask versions of the SABRE method are given in Figures 1 and 3 of the online supplementary materials

3 SABRE methods

The LASSO and elastic net have multiple weaknesses, as we have discussed in Sect. 1, and they have been shown to be sub-optimal compared to Bayesian approaches (Dalton and Dougherty 2012; Mohamed et al. 2012) such as the spike and slab prior (Mitchell and Beauchamp 1988). In the remainder of this section we incorporate the spike and slab prior into a hierarchical Bayesian model as shown in the Probabilistic Graphical Model (PGM) in Fig. 1. Figure 1 shows a particular version of the SABRE method, the conjugate SABRE method, but this section also discusses two other versions of the SABRE method; the semi-conjugate SABRE method (Figure 1 in the online supplementary materials) and the binary mask SABRE method (Figure 3 of the online supplementary materials). The parameters of the models are sampled from their posterior distributions using Markov chain Monte Carlo (MCMC).

3.1 Likelihood

The likelihood for our Bayesian model is similar to the classical mixed-effects model described in Sect. 2.1, however we include only the relevant residue and phylogenetic tree variables, $\mathbf{X}_{\boldsymbol{\gamma}}$, and regressors, $\mathbf{w}_{\boldsymbol{\gamma}}$. As with classical mixed-effects models we separate the intercept, w_0, from the other regressors such that it is always included in the model; see Fig. 1. We expect the intercept to be high as each strain should offer strong protection against itself and hence there should be high log VN titre or log HI assay, $\mathbf{y}$, when all covariates are equal to zero, i.e. the protective and challenge strains are the same.

$$p\left(\mathbf{y}|w_0, \mathbf{w}_{\boldsymbol{\gamma}}, \mathbf{b}, \sigma_\varepsilon^2, \mathbf{X}_{\boldsymbol{\gamma}}, \mathbf{Z}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{w}_{\boldsymbol{\gamma}} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I}\right). \tag{7}$$

The relevance of the jth column of $\mathbf{X}$ is determined by $\gamma_j \in \{0, 1\}$, where feature j is said to be relevant if $\gamma_j = 1$, giving $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^\top \in \{0, 1\}^J$. We then define $\mathbf{X}_{\boldsymbol{\gamma}}$ to be the matrix of relevant explanatory variables with $||\boldsymbol{\gamma}||$ columns and N rows, where $||\boldsymbol{\gamma}|| = \sum_{j=1}^{J} \gamma_j$ is the number of non-zero elements of $\boldsymbol{\gamma}$. Similarly $\mathbf{w}_{\boldsymbol{\gamma}}$ is given as the column vector of regressors, where the inclusion of each parameter is dependent on $\boldsymbol{\gamma}$. This is demonstrated by the following example:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,2} & x_{3,3} \end{bmatrix}; \ \mathbf{X}_{\boldsymbol{\gamma}} = \begin{bmatrix} x_{1,1} & x_{1,3} \\ x_{2,1} & x_{2,3} \\ x_{3,1} & x_{3,3} \end{bmatrix}; \ \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}; \ \mathbf{w}_{\boldsymbol{\gamma}} = \begin{bmatrix} w_1 \\ w_3 \end{bmatrix};$$
$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 = 1 \\ \gamma_2 = 0 \\ \gamma_3 = 1 \end{bmatrix}. \tag{8}$$

An alternative to this model is the binary mask model; e.g. Chapter 13 of Murphy (2012). In binary mask models the indicator variables, $\boldsymbol{\gamma}$, 'mask' or hide the impact of the non-zero coefficients, $\mathbf{w}$, and explanatory variables, $\mathbf{X}$, when the variable is not selected:

$$p\left(\mathbf{y}|w_0, \mathbf{w}, \boldsymbol{\gamma}, \mathbf{b}, \sigma_\varepsilon^2, \mathbf{X}, \mathbf{Z}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{1}w_0 + \mathbf{X}\boldsymbol{\Gamma}\mathbf{w} + \mathbf{Z}\mathbf{b}, \sigma_\varepsilon^2\mathbf{I}\right) \tag{9}$$

where $\boldsymbol{\Gamma} = diag(\boldsymbol{\gamma})$. We have tested the binary mask version of the model against the other versions of the SABRE method in Sect. 6.1 and found that the results are reasonably similar. For clarity of the paper we have moved a more in depth description of the model into Section 5 of the online supplementary materials, as we believe that the spike and slab model makes more sense theoretically as the variance of the fixed effects, $\sigma_{w,h}^2$, is only calculated based on those variables included in the model.

3.2 Noise and intercept priors

As with the classical methods described in Sect. 2, we assume additive iid Gaussian noise with variance σ_ε^2. In a Bayesian context we wish to infer σ_ε^2, so we specify the conjugate prior:

$$\sigma_\varepsilon^2 \sim \mathcal{IG}\left(\sigma_\varepsilon^2 | \alpha_\varepsilon, \beta_\varepsilon\right) \tag{10}$$

where the hyper-parameters α_ε and β_ε are fixed, as indicated by the grey nodes in Fig. 1.

Additionally we also require a prior on our intercept, w_0;

$$w_0 \sim \mathcal{N}\left(w_0 | \mu_{w_0}, \sigma_{w_0}^2 \sigma_\varepsilon^2\right). \tag{11}$$

We treat the intercept differently from the remaining regressors, wishing to use vague prior settings so as not to penalise this term and effectively make the model scale invariant (Hastie et al. 2009).

The distribution of w_0 also has σ_ε^2 included, which makes the model conjugate rather than semi-conjugate, as discussed in Chapter 3 of Gelman et al. (2013). Additionally, there are relationships between w_0, $\mathbf{w}_{\boldsymbol{\gamma}}$, $\boldsymbol{\mu}_{\mathbf{w}} = (\mu_{w,1}, \ldots, \mu_{w,H})^\top$ (defined in Sect. 3.3) and the error variance, σ_ε^2, increasing information sharing and meaning that the error variance in terms of model fit is reflected in the distribution of the regression coefficients; see Fig. 1. In addition to the increased information sharing, conjugate models also have a computational advantage as the sampling strategy can be improved through using collapsed Gibbs sampling. The difference between the conjugate and semi conjugate SABRE models in terms of accuracy and computational efficiency is discussed in Sect. 6.1. Additionally, the PGM for the semi-conjugate version of our model is given in Figure 1 of the online supplementary materials.

3.3 Spike and slab priors

Spike and slab priors have been used in a number of different contexts and have been shown to outperform ℓ_1 methods both in terms of variable selection and out-of-sample predictive performance (Mohamed et al. 2012). They were originally proposed by Mitchell and Beauchamp (1988) as a mixture of a Gaussian distribution and Dirac spike, but have also been used as a mixture of two Gaussians (George and McCulloch 1993, 1997). Binary mask models (e.g. Jow et al. 2014) have also been used as an alternative to the spike and slab prior in a number of applications, as is discussed in Chapter 13 of Murphy (2012). A binary mask based version of the conjugate SABRE method is compared in Sect. 6.1 and given in Section 5 of the online supplementary materials.

The idea behind the spike and slab prior is that the prior reflects whether the feature is relevant based on the values of $\boldsymbol{\gamma}$. In this way we expect that $w_{j,h} = 0$ if $\gamma_j = 0$, i.e. the feature is irrelevant, and conversely it should be non-zero if the variable is relevant, $w_{j,h} \neq 0$ if $\gamma_i = 1$. For generality, we allow the models the option to have multiple groups of variables $h \in \{1, \ldots, H\}$ which are defined by j, i.e. $w_{j,h}$ is shorthand for w_{j,h_j}. However this is not used in the results reported in Sect. 6. A conjugate prior, with σ_ε^2 added for further conjugacy, is then assigned where the feature is relevant and a Dirac spike at zero where it is not:

$$p\left(w_{j,h}|\gamma_j, \mu_{w,h}, \sigma^2_{w,h}, \sigma^2_\varepsilon\right) = \begin{cases} \delta_0(w_{j,h}) & \text{if } \gamma_j = 0 \\ \mathcal{N}\left(w_{j,h}|\mu_{w,h}, \sigma^2_{w,h}\sigma^2_\varepsilon\right) & \text{if } \gamma_j = 1 \end{cases} \tag{12}$$

for $j \in 1, \ldots, J$ and where δ_0 is the delta function. Here we have a spike at 0 and as $\sigma^2_{w,h}\sigma^2_\varepsilon \to \infty$ the distribution, $p(w_{j,h}|\gamma_j = 1)$, approaches a uniform distribution, a slab of constant height.

The prior for the variance of the parameters selected is then given by:

$$\sigma^2_{w,h} \sim \mathcal{IG}\left(\sigma^2_{w,h}|\alpha_{w,h}, \beta_{w,h}\right). \tag{13}$$

By choosing the same fixed hyper-parameters, $\alpha_{w,h}$ and $\beta_{w,h}$ for each h, we lose information coupling between the different groups, although this could be regained with an additional layer in the hierarchical model.

In addition to $\sigma^2_{w,h}$, we use the hyper-parameters $\mu_{w,h}$ to reflect the likely non-zero means of each group h:

$$\mu_{w,h} \sim \mathcal{N}\left(\mu_{w,h}|\mu_{0,h}, \sigma^2_{0,h}\sigma^2_\varepsilon\right) \tag{14}$$

where the hyper-parameters $\mu_{0,h}$ and $\sigma^2_{0,h}$ are fixed and σ^2_ε is again included in the variance for further conjugacy. This specification comes from our biological understanding of the problem. In the FMDV and H1N1 data we are likely to observe a comparatively large intercept, with negative regression coefficients, $w_{j,h}$, reflecting the fact that any mutational changes are likely to reduce the similarity between virus strains, therefore reducing the measured VN titre or HI assay.

For mathematical convenience we then define the prior distribution of $\mathbf{w}^*_{\boldsymbol{\gamma}} = (w_0, \mathbf{w}^\top_{\boldsymbol{\gamma}})^\top$ as:

$$\mathbf{w}^*_{\boldsymbol{\gamma}} \sim \mathcal{N}\left(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_\varepsilon \boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\right) \tag{15}$$

where $\mathbf{m}_{\boldsymbol{\gamma}} = (\mu_{w_0}, \mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ldots, \mu_{w,H})^\top$ and $\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = diag(\boldsymbol{\sigma}^2_{\mathbf{w}^*})$ with $\boldsymbol{\sigma}^2_{\mathbf{w}^*} = (\sigma^2_{w_0}, \sigma^2_{w,1}, \ldots, \sigma^2_{w,1}, \sigma^2_{w,2}, \ldots, \sigma^2_{w,H})^\top$. Each $\mu_{w,h}$ and $\sigma^2_{w,h}$ is repeated with length $||\mathbf{w}_{\boldsymbol{\gamma},h}||$ dependent on $\boldsymbol{\gamma}$.

The final part of the spike and slab prior is to set a prior for $\boldsymbol{\gamma}$, the parameters which determine the relevance of the variables:

$$p(\boldsymbol{\gamma}|\pi) = \prod_{j=1}^{J} \text{Bern}(\gamma_j|\pi) \tag{16}$$

where π is the probability of the individual variable being relevant. The value of π can either be set as a fixed hyper-parameter as in Sabatti and James (2005), where the authors argue that it should be determined by underlying knowledge of the problem. Alternatively it can be given a conjugate Beta prior:

$$\pi \sim \mathcal{B}(\pi|\alpha_\pi, \beta_\pi) \tag{17}$$

as has been used here. This is a more general model, which subsumes a fixed π as a limiting case for $\alpha_\pi \beta_\pi / ((\alpha_\pi + \beta_\pi)^2(\alpha_\pi + \beta_\pi + 1)) \to 0$ and has also been shown to act as a multiplicity correction in Scott and Berger (2010).

3.4 Random-effects priors

In mixed-effects models the random effects, $b_{k,g}$, are usually given group dependant Gaussian priors where the group g is defined by k, i.e. $b_{k,g}$ is shorthand for b_{k,g_k}:

$$b_{k,g} \sim \mathcal{N}\left(b_{k,g} | \mu_{b,g}, \sigma^2_{b,g}\right). \tag{18}$$

We define this to have a fixed mean, $\mu_{b,g} = 0$, and a common variance parameter, $\sigma^2_{b,g}$, with a conjugate Inverse-Gamma prior for each random-effects group g, as shown in Fig. 2a:

$$\sigma^2_{b,g} \sim \mathcal{IG}\left(\sigma^2_{b,g} | \alpha_{b,g}, \beta_{b,g}\right) \tag{19}$$

where $\alpha_{b,g}$ and $\beta_{b,g}$ are fixed hyper-parameters for each g and we define $\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$ where $\boldsymbol{\Sigma}_\mathbf{b} = diag(\boldsymbol{\sigma}^2_\mathbf{b})$ with $\boldsymbol{\sigma}^2_\mathbf{b} = (\sigma^2_{b,1}, \ldots, \sigma^2_{b,1}, \sigma^2_{b,2}, \ldots, \sigma^2_{b,G})^\top$ such that each $\sigma^2_{b,g}$ is repeated with length $||\mathbf{b}_g||$.

An alternative to this hierarchical prior setting is the folded-non-central-t prior distribution described in Gelman (2006), which gives a redundant multiplicative reparameterisation to the model given in (18), (19) and Fig. 2a. This prior has several potential advantages over the Inverse-Gamma prior. Firstly it is considered to be a prior that better represents non-informativeness. While the posterior distribution can be sensitive to the fixed hyper-parameter settings of an Inverse-Gamma prior, the impact is reduced when the folded-non-central-t prior is used. In that case the posterior distribution does not have a sharp peak at zero unlike with an vague

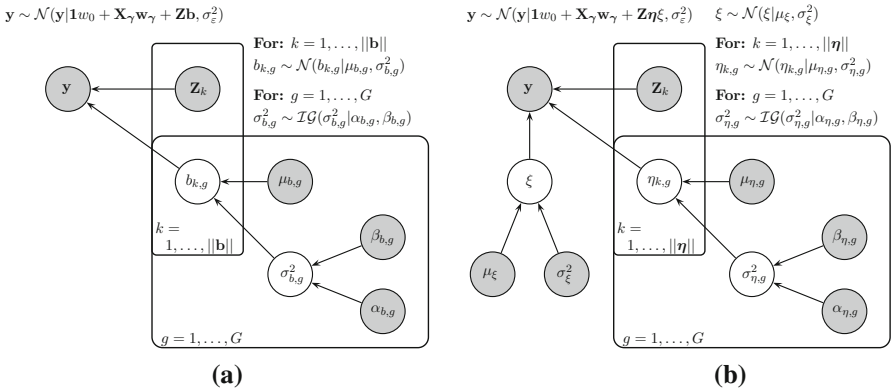

Fig. 2 PGMs for the two different specifications of the hierarchical random-effects model. **a** Classical random-effects model using Gaussian and Inverse-Gamma priors. **b** Half-t prior specified in a hierarchical manner, as suggested by Gelman (2006)

Inverse-Gamma prior, reducing problems with underestimating the variance. Secondly, Gelman (2006) found that the folded-non-central-t prior results in a more realistic posterior distribution of $\sigma^2_{b,g}$ when there are only a few random effects (usually less than 8) in each group g. The author showed that the posterior distribution reflected the marginal distribution well at its low end, but removed its unrealistically heavy tail; see Figure 2 in Gelman (2006). Doing this ensures that $\sigma^2_{b,g}$ is not overestimated and does not lead to non-optimal shrinkage of $\mathbf{b}_g$. Finally the overparameterisation can improve sampling by reducing the dependence between parameters in the hierarchical model leading to improved MCMC convergence (Gelman 2004).

The redundant multiplicative reparameterisation used for this prior specification sets $\mathbf{b} = \boldsymbol{\eta}\xi$ and is given by the following conjugate priors and shown in Fig. 2b:

$$\eta_{k,g} \sim \mathcal{N}\left(\eta_{k,g}|\mu_{\eta,g}, \sigma^2_{\eta,g}\right) \tag{20}$$

$$\xi \sim \mathcal{N}\left(\xi|\mu_{\xi}, \sigma^2_{\xi}\right) \tag{21}$$

where μ_ξ and σ^2_ξ are fixed for identifiability, $\mu_{\eta,g} = 0$, $\eta_{k,g}$ is shorthand for η_{k,g_k} and each $b_{k,g} = \xi\eta_{k,g}$. Following Gelman (2006), we fix $\mu_\xi = 0$ which leads to the half-t distribution. We then set a prior on $\sigma^2_{\eta,g}$:

$$\sigma^2_{\eta,g} \sim \mathcal{IG}\left(\sigma^2_{\eta,g}|\alpha_{\eta,g}, \beta_{\eta,g}\right) \tag{22}$$

where $\alpha_{\eta,g}$ and $\beta_{\eta,g}$ are fixed hyper-parameters. In terms of classical mixed-effects models, the variance is given by $\sigma^2_{b,g} = \xi^2\sigma^2_{\eta,g}$. For convenience we define $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$ when $\mu_{\eta,g} = 0$ for all g and where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = diag(\boldsymbol{\sigma}^2_{\boldsymbol{\eta}})$ with $\boldsymbol{\sigma}^2_{\boldsymbol{\eta}} = (\sigma^2_{\eta,1}, \ldots, \sigma^2_{\eta,1}, \sigma^2_{\eta,2}, \ldots, \sigma^2_{\eta,G})^\top$ where each $\sigma^2_{\eta,g}$ is repeated with length $||\boldsymbol{\eta}_g||$.

3.5 Posterior inference

In order to explore the posterior distribution of the parameters we use an MCMC algorithm. Having chosen conjugate priors where possible means we can run a Gibbs sampler for the majority of parameters (Ripley 1979; Geman and Geman 1984). The only exception is $\boldsymbol{\gamma}$, although it is possible to use component-wise Gibbs sampling with a small adaptation; see Sect. 3.6.1. Additionally we sample the intercept and regression parameters together and define $\mathbf{w}^*_{\boldsymbol{\gamma}} = (w_0, \mathbf{w}^\top_{\boldsymbol{\gamma}})^\top$, $\mathbf{X}^*_{\boldsymbol{\gamma}} = (\mathbf{1}, \mathbf{X}_{\boldsymbol{\gamma}})$, $\mathbf{m}_{\boldsymbol{\gamma}} = (\mu_{w_0}, \mu_{w,1}, \ldots, \mu_{w,1}, \mu_{w,2}, \ldots, \mu_{w,H})^\top$ and $\boldsymbol{\Sigma}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = diag(\boldsymbol{\sigma}^2_{\mathbf{w}^*})$ with $\boldsymbol{\sigma}^2_{\mathbf{w}^*} = (\sigma^2_{w_0}, \sigma^2_{w,1}, \sigma^2_{w,1}, \sigma^2_{w,2}, \ldots, \sigma^2_{w,H})^\top$. Each $\mu_{w,h}$ and $\sigma^2_{w,h}$ is repeated with length $||\mathbf{w}_{\boldsymbol{\gamma},h}||$ dependent on $\boldsymbol{\gamma}$, as indicated below (15).

The conditional distributions for those parameters amenable to standard Gibbs sampling are derived in Section 2 of the online supplementary materials and given here, where by a slight abuse of notation $\boldsymbol{\theta}'$ denotes all the other parameters, excluding the ones on the left of the conditioning bar:

$$\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}\left(\mathbf{w}^*_{\boldsymbol{\gamma}}|\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}(\mathbf{y}-\mathbf{Z}\mathbf{b}) + \mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\mathbf{m}_{\boldsymbol{\gamma}}, \sigma^2_{\varepsilon}\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\right) \tag{23}$$

$$\mathbf{b}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}\left(\mathbf{b}|\tfrac{1}{\sigma^2_{\varepsilon}}\mathbf{V}_{\mathbf{b}}\mathbf{Z}^{\top}\left(\mathbf{y}-\mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}\right), \mathbf{V}_{\mathbf{b}}\right) \tag{24}$$

$$\sigma^2_{b,g}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}\left(\sigma^2_{b,g}|\ ||\mathbf{b}_g||/2 + \alpha_{b,g}, \beta_{b,g} + \tfrac{1}{2}\mathbf{b}^{\top}_g\mathbf{b}_g\right) \tag{25}$$

$$\mu_{w,h}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}\left(\mu_{w,h}|V^{-1}_{\mu_{\gamma},h}\left(\sum\left(\mathbf{w}_{\boldsymbol{\gamma},h}\right)/\sigma^2_{w,h} + \mu_{0,h}/\sigma^2_{0,h}\right), \sigma^2_{\varepsilon}V_{\mu_{\gamma},h}\right) \tag{26}$$

$$\begin{aligned}\sigma^2_{w,h}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}\Big(&\sigma^2_{w,h}|\ ||\mathbf{w}_{\boldsymbol{\gamma},h}||/2 + \alpha_{w,h}, \beta_{w,h} \\ &+ \tfrac{1}{2\sigma^2_{\varepsilon}}\left(\mathbf{w}_{\boldsymbol{\gamma},h} - \mathbf{1}\mu_{w,h}\right)^{\top}\left(\mathbf{w}_{\boldsymbol{\gamma},h} - \mathbf{1}\mu_{w,h}\right)\Big)\end{aligned} \tag{27}$$

$$\sigma^2_{\varepsilon}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}\left(\sigma^2_{\varepsilon}|\left(N + ||\mathbf{w}^*_{\boldsymbol{\gamma}}|| + H\right)/2 + \alpha_{\varepsilon}, \beta_{\varepsilon} + \tfrac{1}{2}R_{\sigma^2_{\varepsilon}}\right) \tag{28}$$

$$\pi|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{B}\left(\pi|\alpha_{\pi} + ||\boldsymbol{\gamma}||, \beta_{\pi} + J - ||\boldsymbol{\gamma}||\right) \tag{29}$$

where we sample $\sigma^2_{b,g}$, $\mu_{w,h}$ and $\sigma^2_{w,h}$ for each g and h respectively. We also define $\mathbf{V}_{\mathbf{w}^*_{\boldsymbol{\gamma}}} = \left(\mathbf{X}^{*\top}_{\boldsymbol{\gamma}}\mathbf{X}^*_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\right)^{-1}$, $\mathbf{V}_{\mathbf{b}} = \left(\frac{1}{\sigma^2_{\varepsilon}}\mathbf{Z}^{\top}\mathbf{Z} + \boldsymbol{\Sigma}^{-1}_{\mathbf{b}}\right)^{-1}$, $V_{\mu_{\gamma},h} = \left(\left(||\mathbf{w}_{\boldsymbol{\gamma},h}||/\sigma^2_{w,h}\right)^{-1} + \left(\sigma^2_{0,h}\right)^{-1}\right)^{-1}$ and $R_{\sigma^2_{\varepsilon}} = \left(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b}\right)^{\top}\left(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{Z}\mathbf{b}\right) + \left(\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}}\right)^{\top}\boldsymbol{\Sigma}^{-1}_{\mathbf{w}^*_{\boldsymbol{\gamma}}}\left(\mathbf{w}^*_{\boldsymbol{\gamma}} - \mathbf{m}_{\boldsymbol{\gamma}}\right) + \sum_{h=1}^{H}(\mu_{w,h} - \mu_{0,h})^2/\sigma^2_{0,h}$ for notational simplicity.

In order to use the half-t prior instead of the standard Inverse-Gamma prior we set $\mathbf{b} = \boldsymbol{\eta}\xi$ and $\sigma^2_{b,g} = \xi^2\sigma^2_{\eta,g}$. This would also need to be done for sampling $\boldsymbol{\gamma}$ in Sect. 3.6. We can then sample $\boldsymbol{\eta}$, ξ and $\sigma^2_{\eta,g}$ from their conditional distributions, replacing (24) and (25):

$$\boldsymbol{\eta}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\eta}|\tfrac{\xi}{\sigma^2_{\varepsilon}}\mathbf{V}_{\boldsymbol{\eta}}\mathbf{Z}^{\top}\left(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}\right), \mathbf{V}_{\boldsymbol{\eta}}\right) \tag{30}$$

$$\xi|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{N}\left(\xi|V_{\xi}\left[\tfrac{\mu_{\xi}}{\sigma^2_{\xi}} + \tfrac{1}{\sigma^2_{\varepsilon}}\boldsymbol{\eta}^{\top}\mathbf{Z}^{\top}\left(\mathbf{y} - \mathbf{X}^*_{\boldsymbol{\gamma}}\mathbf{w}^*_{\boldsymbol{\gamma}}\right)\right], V_{\xi}\right) \tag{31}$$

$$\sigma^2_{\eta,g}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y} \sim \mathcal{IG}\left(\sigma^2_{\eta,g}|||\boldsymbol{\eta}_g||/2 + \alpha_{\eta,g}, \beta_{\eta,g} + \tfrac{1}{2}\boldsymbol{\eta}^{\top}_g\boldsymbol{\eta}_g\right) \tag{32}$$

where $\mathbf{V}_{\boldsymbol{\eta}} = (\frac{\xi^2}{\sigma^2_{\varepsilon}}\mathbf{Z}^{\top}\mathbf{Z} + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\eta}})^{-1}$ and $V_{\xi} = (\frac{1}{\sigma^2_{\xi}} + \frac{1}{\sigma^2_{\varepsilon}}\boldsymbol{\eta}^{\top}\mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\eta})^{-1}$.

Collapsing will lead to improved mixing and convergence, e.g. Andrieu and Doucet (1999). We take advantage of the induced conjugacy to sample the parameters $\boldsymbol{\gamma}$, $\mathbf{w}^*_{\boldsymbol{\gamma}}$, $\boldsymbol{\mu}_{\mathbf{w}} = (\mu_{w,1}, \ldots, \mu_{w,H})^{\top}$, σ^2_{ε} and π as a series of collapsed distributions rather than through Gibbs sampling:

$$\begin{aligned}p\left(\boldsymbol{\gamma}, \mathbf{w}^*_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_{\varepsilon}, \pi\right) &= p\left(\boldsymbol{\gamma}\right)p\left(\pi|\boldsymbol{\gamma}\right)p\left(\sigma^2_{\varepsilon}|\pi, \boldsymbol{\gamma}\right)p\left(\boldsymbol{\mu}_{\mathbf{w}}|\sigma^2_{\varepsilon}, \pi, \boldsymbol{\gamma}\right) \\ &\quad\times\ p\left(\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_{\varepsilon}, \pi, \boldsymbol{\gamma}\right) \qquad (33)\\ &= p\left(\boldsymbol{\gamma}\right)p\left(\pi|\boldsymbol{\gamma}\right)p\left(\sigma^2_{\varepsilon}|\boldsymbol{\gamma}\right)p\left(\boldsymbol{\mu}_{\mathbf{w}}|\sigma^2_{\varepsilon}, \boldsymbol{\gamma}\right)p\left(\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_{\varepsilon}, \boldsymbol{\gamma}\right) \qquad (34)\end{aligned}$$

where the conditionality on $\boldsymbol{\theta}'$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{y}$ has been dropped and the simplification from (33) to (34) follows from the conditional independence relations shown in Fig. 1, exploiting the fact that π is d-separated from the remaining parameters in the argument via $\boldsymbol{\gamma}$. These distributions are achieved by collapsing over parameters as derived in Sections 2, 3 and 4 of the online supplementary materials, and are used for all of our conjugate models.

3.6 Sampling the latent indicators

Sampling $\boldsymbol{\gamma}$ is more difficult, as it does not naturally take a distribution of standard form. However we can still get a valid conditional distribution and use a variety of techniques to sample from it. Here we have used collapsing methods following Sabatti and James (2005) to achieve faster mixing and convergence:

$$p\left(\boldsymbol{\gamma}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}\right) \propto \int p\left(\boldsymbol{\gamma}, \pi, \sigma^2_\varepsilon, \mathbf{w}^*_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\mathbf{w}}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}\right) d\boldsymbol{\mu}_{\mathbf{w}} d\mathbf{w}^*_{\boldsymbol{\gamma}} d\pi d\sigma^2_\varepsilon \quad (35)$$

$$\propto \int p\left(\boldsymbol{\gamma}|\pi\right) p\left(\pi\right) p\left(\mathbf{y}|\mathbf{w}^*_{\boldsymbol{\gamma}}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \sigma^2_\varepsilon\right) p\left(\mathbf{w}^*_{\boldsymbol{\gamma}}|\boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_\varepsilon\right)$$

$$\times\ p\left(\boldsymbol{\mu}_{\mathbf{w}}|\sigma^2_\varepsilon\right) p\left(\sigma^2_\varepsilon\right) d\boldsymbol{\mu}_{\mathbf{w}} d\mathbf{w}^*_{\boldsymbol{\gamma}} d\pi d\sigma^2_\varepsilon \quad (36)$$

where the factorisation follows from the conditional independence relations depicted in Fig. 1 and the fixed hyper-parameters (given as grey circles in Fig. 1) have been dropped to improve notational clarity. The full distribution is available in Section 3 of the online supplementary materials.

Multiple methods have been proposed for sampling the latent variables, $\boldsymbol{\gamma}$. In this paper we look at two of these in particular; the component-wise Gibbs sampling approach and a Metropolis–Hastings step (Metropolis et al. 1953; Hastings 1970). In the latter we can propose changes to multiple parameters simultaneously for a computational improvement.

3.6.1 Component-wise Gibbs sampling

A component-wise Gibbs sampler can be used to consecutively sample each γ_j from $\boldsymbol{\gamma}$ in a random order dependent on the current state, c, of all the other γs, $\boldsymbol{\gamma}^c_{-j} = \left(\gamma^c_1, \ldots, \gamma^c_{j-1}, \gamma^c_{j+1}, \ldots, \gamma^c_J\right)$. We can define the conditional distribution of the ith iteration of γ_j to be a Bernoulli distribution with probability $p(\gamma_j = 1|\boldsymbol{\theta}', \boldsymbol{\gamma}^c_{-j}, \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y}) = \frac{a}{a+b}$, where we define $a \propto p(\gamma_j = 1, \boldsymbol{\gamma}^c_{-j}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})$ and $b \propto p(\gamma_j = 0, \boldsymbol{\gamma}^c_{-j}|\boldsymbol{\theta}', \mathbf{X}^*_{\boldsymbol{\gamma}}, \mathbf{Z}, \mathbf{y})$ using (36).

3.6.2 Block Metropolis–Hastings sampling

Block Metropolis–Hastings sampling can improve mixing and convergence through proposing sets, S, of latent indicator variables, $\boldsymbol{\gamma}_S$, simultaneously, where $\boldsymbol{\gamma}_S$ denotes

a column vector of all the γ_js where $j \in S$ and $\boldsymbol{\gamma}_{-S}$ its compliment. The proposals are then accepted with the following acceptance rate:

$$\alpha\left(\boldsymbol{\gamma}_S^*, \boldsymbol{\gamma}_S^c | \boldsymbol{\theta}', \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}, \boldsymbol{\gamma}_{-S}^c\right) := \min\left\{\frac{q\left(\boldsymbol{\gamma}_S^c | \boldsymbol{\gamma}_S^*, \pi\right) p\left(\boldsymbol{\gamma}_S = \boldsymbol{\gamma}_S^*, \boldsymbol{\gamma}_{-S}^c | \boldsymbol{\theta}', \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}\right)}{q\left(\boldsymbol{\gamma}_S^* | \boldsymbol{\gamma}_S^c, \pi\right) p\left(\boldsymbol{\gamma}_S = \boldsymbol{\gamma}_S^c, \boldsymbol{\gamma}_{-S}^c | \boldsymbol{\theta}', \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}\right)}, 1\right\} \tag{37}$$

where $q(.)$ is a proposal density, which we set to be: $q(\boldsymbol{\gamma}_S^* | \boldsymbol{\gamma}_S^c, \pi) = \prod_{j \in S} \text{Bern}(\gamma_j^* | \pi)$. Proposed moves for independent sets of randomly ordered inclusion parameters, $\boldsymbol{\gamma}_S^*$, are then accepted if $\alpha(\boldsymbol{\gamma}_S^*, \boldsymbol{\gamma}_S^c | \boldsymbol{\theta}', \mathbf{X}_{\boldsymbol{\gamma}}^*, \mathbf{Z}, \mathbf{y}, \boldsymbol{\gamma}_{-S}^c)$ is greater than a random variable $u \sim \mathcal{U}[0, 1]$, until updates have been proposed for all the latent indicator variables.

4 Data

The antigenic similarity between two viruses can be measured by VN titre or HI assay. To get these measurements an antiserum is created from a single strain, the protective, and measured in its ability to neutralise a sample of a different strain, the challenge. However the resulting measures are affected by a number of experimental effects, which can potentially include the challenge strain, antiserum and date; see Sects. 4.2, 4.3, 4.4 and 4.5 . The experimental effects are accounted for via the random effects, with the random-effects coefficients, $b_{k,g}$, representing an unknown effect of a particular random effect level on the measured log VN titre or log HI assay. Once this has been accounted for, it should then be possible to explain the underlying true log VN titre or log HI assay values by looking at the difference in the protein structure of the two viruses. This difference can be attributed to the presence (1) or absence (0) of an amino acid substitution at each specific residue which is exposed on the surface of the capsid or virus shell. This information can be added into the model as fixed-effects and the selection of a particular residue, or variable, indicates its importance in explaining antigenicity. Residues in the FMDV datasets are given by their protein sequence alignment (Reeve et al. 2010), where for example VP3 138 represents position 138 on the VP3 protein. Residues in the H1N1 dataset are given by their position on the H1 common alignment (Harvey et al. 2016).

However, sometimes we observe antigenic differences between virus lineages that we unable to attribute to amino acid changes at any specific residue. In these cases we wish to relate the changes to the evolutionary history of the virus. We do this through the inclusion of variables related to different properties of the branches of the phylogenetic tree. The reconstruction of phylogenetic trees is not the subject of this article, and therefore we use trees generated from the structural proteins of SAT1, SAT2 and H1N1 viruses presented by Reeve et al. (2010, 2016) and Harvey et al. (2016). Where possible, for each of the branches we include variables related to the effect of the challenge and protective strains, as well as to account for unknown antigenic effects; see Section 6 of the supplementary details for more information.

4.1 Simulated data

To test the effectiveness of the methods described in Sects. 2 and 3 , we have used a simulation study. We simulated 12 sets of simulated data based on the structure of the FMDV datasets. For the first 9 sets of simulated data, the datasets contain 100 measurements for training and 900 measurements for testing. For each of these sets of simulated data we varied the number of variables, $||\mathbf{w}|| \in \{40, 60, 80\}$, and the size of the error, $\sigma_\varepsilon^2 \in \{0.01, 0.1, 0.3\}$, to test the methods under different circumstances. The final 3 sets of simulated data were used to test the model under the $p > N$ setting with each of the datasets containing 50 measurements for training and 900 for testing. We then set $||\mathbf{w}|| = 200$ and again varied the size of the error, $\sigma_\varepsilon^2 \in \{0.01, 0.1, 0.3\}$. Additionally in each of the 12 sets of simulated data we added two groups of random effects to each dataset to represent experimental variation, both with 8 levels.

To reflect the fact that we expect many of the variables to have no influence on the response we drew a probability π from $\mathcal{U}(0.2, 0.4)$ for each of the first 9 sets of data and fixed it to $\pi = 0.05$ for the other 3 sets of data. The range of values for π in the first 9 sets of data reflects the values we expect to see for the FMDV and Influenza datasets, while the value for the other sets of data, $\pi = 0.05$, represents a value found regularly in the literature for other biological problems. With this probability, each of the variables in the dataset was then given a regressor simulated from $\mathcal{U}(-0.4, -0.2)$ and zero otherwise, remembering that we expect the variables to have a negative effect as any mutational changes will reduce the response, VN titre. Each response y_i was then generated with an intercept of 10 and perturbed with $\mathcal{N}(0, 0.02)$ iid additive Gaussian noise.

4.2 SAT1 data

The original SAT1 dataset analysed in Reeve et al. (2010) is made up of 246 VN titre measurements of comparisons between 3 protective and 20 challenge strains. For each of these measurements, there are 754 residues in the amino acid sequence of the structural proteins. Of these 306 are exposed on the surface of the capsid, and 137 are variable between the 20 test viruses, producing usable indicator variables to assess the antigenic effect of amino acid substitutions. The phylogenetic tree contains 38 branches, and we have included variables on the phylogenetic trees to account for the type of effect where possible (see Section 6 of the supplementary materials), resulting in 64 different indicator variables to help determine the effect of each branch. To complete the analysis we removed groups of variables with correlation coefficients of one, leaving only one variable included but using information from the whole group in order to classify the included variable (see Table 3 and Section 7 of the online supplementary materials). This gave the final dataset 138 variables in total. Random effects were included to account for the antiserum and challenge strain.

4.3 Extended SAT1 data

After the original analysis of the SAT1 data in Reeve et al. (2010), more data was collected, including additional strains and repeated experiments, for the SAT1

serotype (Reeve et al. 2016). This data includes the original SAT1 data and consists of a total of 2125 VN titre measurements with 5 protective and 42 challenge strains. Of the 306 surface exposed sites, the amino acid sequence is variable between the viruses at 146. 132 variables were also provided from the phylogenetic structure. Once groups of variables with correlation coefficients of one were removed, 221 variables were left in the model. Random effects were included to account for the antiserum, challenge strain and date of the experiment.

4.4 SAT2 data

The SAT2 data was originally analysed in Reeve et al. (2010) and contains 320 VN titre measurements of 4 protective and 22 challenge strains. It contains data on 128 variable surface exposed residues and 80 variables associated with different types of phylogenetic changes. After removing variables with correlation coefficients of one as before, this left 148 different variables to be included in the model. Random effects were included to account for the antiserum and challenge strain.

4.5 Influenza A (H1N1) data

Harvey et al. (2016) used a H1N1 dataset that contained 506 challenge strains and 43 protective strains. Here we have uses a slightly smaller dataset in order to fully account for the effect of the phylogenetic structure; see Section 6 of the supplementary materials. The dataset used here contains 15,693 HI assay measurements with 43 challenge and 43 protective strains. As this full dataset is too large to analyse using the conjugate SABRE method we have summarised the data to 570 mean HI assay measurement for each combination of challenge and protective strains. For each pair of challenge and protective strains the 279 explanatory variables, 53 surface exposed residues and 226 variables related to the phylogenetic data, remain the same. Doing this however means we cannot use the date of the experiment as a random effect group and additionally the dataset does not contain antiserum data, meaning we have only used the challenge strain as a random effect group.

5 Computational inference

Our code for the classical mixed-effects models has been implemented in *R* (R Core Team 2013) using the package *lme4* (Bates et al. 2015). To choose these models we used forward inclusion while adjusting for multiple testing using the Holm–Bonferroni correction. The code for the mixed-effects LASSO, mixed-effects elastic net and the SABRE methods was written directly into *R* and the software is freely available from the authors upon request.

For our MCMC chains, we sampled 10,000 and 15,000 iterations respectively for the simulated and real data before removing an appropriate portion for burn-in. This was determined by running 4 chains for each model and computing the potential scale reduction factor (PSRF) (Gelman and Rubin 1992) from the within-chain and

between-chain variances (Plummer et al. 2006). We take a PSRF ≤ 1.05 as a threshold for convergence and terminate the burn-in when this is consistently satisfied for 95% of the variables. In general, the fixed hyper-parameters, shown as grey nodes in Fig. 1, were set to give a vague distribution for the flexible (hyper-)parameters, shown as white nodes. The only exception was the prior on π, defined in (17), which was set to be weakly informative such that $\alpha_\pi = 1$ and $\beta_\pi = 4$. This corresponds to prior knowledge that only a small number of residues or branches have a significant antigenic effect.

The following hyper-parameters, shown as grey nodes in Fig. 1, are fixed to give vague distributions: $\alpha_{b,g} = \beta_{b,g} = \alpha_{\eta,g} = \beta_{\eta,g} = 0.001$ and $\mu_{b,g} = \mu_{\eta,g} = 0$ for all g, $\alpha_{w,h} = \beta_{w,h} = 0.001$, $\mu_{0,h} = 0$ and $\sigma^2_{0,h} = 100$ for all h, $\mu_\xi = 0$, $\sigma^2_\xi = 100$, $\mu_{w_0} = max(\mathbf{y})$, $\sigma^2_{w_0} = 100$ and $\alpha_\varepsilon = \beta_\varepsilon = 0.001$. The only unusual choice is $\mu_{w_0} = max(\mathbf{y})$ which follows from us expecting a high intercept with the regression coefficients then having a negative effect on the response. This is a result of strains having high reactivity with themselves, and any changes making the strains less similar, reducing their reactivity.

To analyse the best proposal method we tested the component-wise Gibbs sampler and several specifications of the Metropolis–Hastings sampler on the original SAT1, extended SAT1 and H1N1 datasets. For the Metropolis–Hastings sampler, we proposed the inclusion or exclusion of the variables in groups of 5, 10, 15, 20 and 30. We analysed convergence by monitoring the percentage of variables with a PSRF ≤ 1.05, similar to Grzegorczyk and Husmeier (2013).

For selecting variables in the mixed-effects LASSO and elastic net we used BIC as in Schelldorfer et al. (2011). For the SABRE method there are a variety of techniques that have been used in the literature to choose a cut-off. Often a cut-off of 0.5 is used and this has been shown to be the best predictive model under strict conditions (Barbieri and Berger 2004). Alternatively the top $J\hat{\pi}$ ranked variables have been taken, where J is the number of variables and $\hat{\pi}$ is the posterior mean of π, defined in (16) and (17), i.e. the global probability of variables being included in the model.

6 Results and discussion

To recapitulate, we have introduced a hierarchical Bayesian modelling framework (called SABRE) for selecting relevant antigenic sites in viral evolution. There are two fundamentally different approaches to variable selection: the slab and spike prior, whereby the influence of an input variable is controlled via the prior distribution of its associated regression parameters, and the binary mask model, where variables are put through a binary multiplicative filter. There are also different prior distributions one can choose: a conjugate prior, and a semi-conjugate prior. This gives us three variants of the proposed modelling framework:

- The conjugate SABRE model, with slab and spike prior
- The semi-conjugate SABRE model, with slab and spike prior
- The binary mask SABRE model.

These three variants are depicted as probabilistic graphical models in Fig. 1 of the main paper, and in Figures 1 and 3 of the supplementary material. We

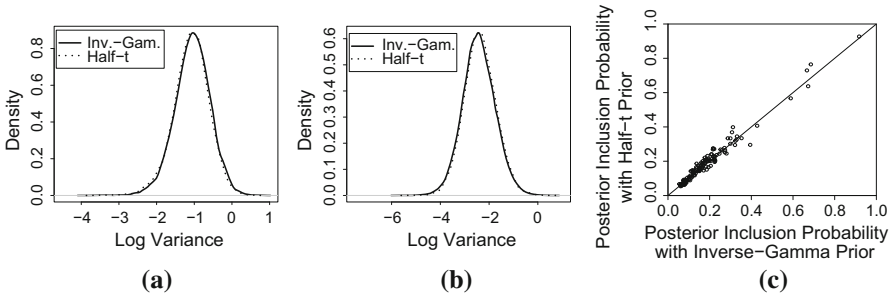

Fig. 3 Gaussian Kernel density estimation plots of random effects variances and a comparison of posterior inclusion probabilities. Gaussian kernel density estimation plots are shown for the sampled posterior densities of the log random effect variance. This is given for the two groups of random effects, **a** challenge strain and **b** antiserum, under a vague Inverse-Gamma prior (*solid*) and the half-t prior (*dotted*) proposed in Gelman (2006). **c** Plot showing the comparative posterior inclusion probability for each variable for the two models

have compared their performance with that of two established methods from the literature: the mixed-effects model with stepwise variable selection, and the mixed-effects LASSO. Since there are indications from the literature that the elastic net offers an improvement over the LASSO, we have also modified the mixed-effects LASSO model from the literature (Schelldorfer et al. 2011) by a novel mixed-effects elastic net model. This gives us three classical methods for comparison:

- Mixed-effects model with stepwise variable selection
- Mixed-effects LASSO model
- Mixed-effects elastic net model.

We have applied and assessed the proposed methods with a three-pronged approach. Firstly, we have tested them on a large set of synthetic benchmark data, where the true structure of the model is known, and it is therefore straightforward to quantify the accuracy of inference. This is discussed in Sect. 6.1. Secondly, we have applied the methods to real data for which partial biological prior knowledge is known, which can be used to partially assess the model predictions. These findings are presented in Sect. 6.2. Finally, in Sect. 6.3, we present novel applications to new data, from a less well known serotype of FMDV and the H1N1 serotype of the Influenza virus. Here, little reliable biological prior knowledge is available, and the purpose of our study is new hypothesis generation.

As discussed in Sect. 3.4, we have also tested the choice of the random-effects prior on the SAT2 dataset. Figure 3a, b show posterior samples of the log variance of the two random-effects groups from the conjugate SABRE method comparing the half-t and Inverse-Gamma priors, and shows no notable differences. Similarly Fig. 3c shows that the inclusion probabilities for the two competing models are approximately the same. Based on these findings, we only report the results obtained with the conjugate Inverse-Gamma prior.

6.1 Simulated data with known ground truth

Table 1 compares the different methods in terms of variable selection, Widely Applicable Information Criterion (WAIC) score (Watanabe 2010), predictive performance and fixed effects coefficients inference using the simulated datasets described in Sect. 4.1.[1] To measure variable selection we have ranked the covariates in terms of their significance or influence. For the Bayesian methods, the ranking is defined by the marginal posterior probabilities of inclusion. For the alternative methods, we explain the way the ranking is obtained below. Since for the simulated data the true covariates are known, this ranking can be used to produce a receiver operating characteristic (ROC) curve (e.g. Hanley and McNeil 1982; Section 5.7. of Murphy 2012), where for all possible values of the inclusion threshold, the sensitivity or recall (the relative proportion of true positive covariates: TP/(TP+FN)) is plotted against the complementary specificity (the relative proportion of false positive covariates: FP/(FP+TN)).[2] By numerical integration we obtain the area under the ROC curve (AUROC) as a global measure of accuracy, where larger values indicate a better performance, starting from AUROC = 0.5 to indicate random expectation, to AUROC = 1 for perfect variable identification.

In addition to ranking the covariates to get ROC curves for the SABRE methods, we also need to rank the alternative established methods for a comparison. For the classical mixed-effects models this is done by removing the significance threshold and ranking the edges by order of inclusion. For the mixed-effects LASSO and elastic net we predicted models for a variety of different penalty parameters, λ, to create the so called LASSO path and create a ranking based on when variables become 0. For the mixed-effects elastic net we only show the results for $\alpha = 0.3$ following Ruyssinck et al. (2014), however the remaining results are available in Section 8 of the online supplementary materials. Alternative AUROC values based on using model selection and then ranking the variables based on the absolute values of the regression coefficients (Aderhold et al. 2014), as well as other results, are also available in Section 8 of the online supplementary materials.

Table 1 also measures the accuracy of predicting out of sample observations, $\mathbf{y_{out}}$, and the fixed effects coefficients, $\mathbf{w}$ in terms of Mean Squared Errors (MSEs). For the Bayesian methods, the predictions are made by sampling from the model and then choosing which variables are included based on taking the top $J \times \hat{\pi}$ variables with the highest inclusion probabilities. The model is then sampled with just those variables set to be included and the estimates calculated. For the mixed-effects LASSO, mixed-effects elastic net and classical mixed effects models the regression coefficients can be taken from the chosen model. The random effects coefficients can then be calculated using the best linear unbiased estimator and predictions of the out of sample observations, $\mathbf{y_{out}}$, made.

[1] We do not do a comparison with the classical mixed effects models in the cases where $p > N$. This is a result of it not being possible to complete the model selection procedure in this case as the regression coefficients, $\mathbf{w}$, are unidentifiable.

[2] TP: true positive count, FP: false positive count, TN: true negative count, FN: false negative count.

Table 1 Table of simulation study results

Method	$N = 100$ & $\|\|\mathbf{w}\|\| = 40$			$N = 100$ & $\|\|\mathbf{w}\|\| = 60$			$N = 100$ & $\|\|\mathbf{w}\|\| = 80$			$N = 50$ & $\|\|\mathbf{w}\|\| = 200$		
	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$
AUROC												
Conjugate SABRE	1	0.98	0.90	1	0.98	0.90	1	0.97	0.88	0.66	0.64	0.63
Semi-conjugate SABRE	1	0.98	0.89	1	0.98	0.89	1	0.97	0.87	0.66	0.64	0.61
BM conjugate SABRE	1	0.98	0.90	1	0.98	0.90	1	0.97	0.88	0.64	0.62	0.60
Mixed-effects LASSO	0.95	0.93	0.80	0.91	0.84	0.74	0.90	0.75	0.69	0.50	0.51	0.51
M-E elastic net ($\alpha = 0.3$)	0.93	0.84	0.79	0.88	0.85	0.76	0.84	0.75	0.69	0.54	0.54	0.53
Mixed-effects models	0.99	0.95	0.80	0.99	0.91	0.75	0.95	0.85	0.72	–	–	–
MSE($\mathbf{y_{out}}$)												
Conjugate SABRE	0.15	0.22	0.49	0.18	0.30	0.57	0.26	0.36	0.63	1.15	1.27	1.62
Semi-conjugate SABRE	0.16	0.23	0.48	0.18	0.29	0.57	0.24	0.35	0.63	1.09	1.15	1.47
BM conjugate SABRE	0.16	0.22	0.49	0.18	0.29	0.56	0.24	0.36	0.62	1.18	1.31	1.67
Mixed-effects LASSO	0.06	0.22	0.59	0.13	0.40	0.75	0.31	0.56	1.37	1.44	1.60	1.92
M-E elastic Net ($\alpha = 0.3$)	0.06	0.18	0.60	0.11	0.34	0.75	0.31	0.65	1.81	2.16	2.21	2.43
Mixed-effects models	0.08	0.23	0.53	0.16	0.37	0.68	0.32	0.50	0.77	–	–	–

Table 1 continued

Method	$N = 100$ & $\|\|\mathbf{w}\|\| = 40$			$N = 100$ & $\|\|\mathbf{w}\|\| = 60$			$N = 100$ & $\|\|\mathbf{w}\|\| = 80$			$N = 50$ & $\|\|\mathbf{w}\|\| = 200$		
	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$	$\sigma_\varepsilon^2 = 0.03$	$\sigma_\varepsilon^2 = 0.1$	$\sigma_\varepsilon^2 = 0.3$
MSE($\mathbf{w}^*$)												
Conjugate SABRE	0.019	0.019	0.025	0.017	0.021	0.024	0.021	0.022	0.024	0.012	0.012	0.016
Semi-conjugate SABRE	0.021	0.022	0.022	0.017	0.020	0.025	0.019	0.020	0.025	0.010	0.010	0.012
BM conjugate SABRE	0.020	0.018	0.022	0.016	0.019	0.023	0.019	0.022	0.025	0.013	0.013	0.017
Mixed-effects LASSO	0.003	0.017	0.046	0.009	0.034	0.060	0.020	0.024	0.071	0.010	0.010	0.013
M-E elastic Net ($\alpha = 0.3$)	0.004	0.010	0.045	0.007	0.022	0.052	0.020	0.038	0.112	0.035	0.039	0.037
Mixed-effects Models	0.008	0.020	0.032	0.015	0.031	0.041	0.033	0.040	0.044	–	–	–
WAIC												
Conjugate SABRE	−309.7	−173.2	−100.4	−314.0	−172.2	−100.8	−309.8	−172.8	−103.1	−102.70	−129.39	−93.92
Semi-conjugate SABRE	−308.7	−170.5	−96.8	−312.1	−171.2	−98.5	−310.5	−171.4	−101.3	−159.54	−163.80	−132.99
BM conjugate SABRE	−309.7	−173.5	−98.7	−313.9	−171.9	−101.3	−310.4	−172.0	−103.3	−106.77	−95.09	−148.57

The table gives results for the Conjugate, Semi-Conjugate and Binary Mask (BM) Conjugate SABRE methods, the mixed-effects LASSO, the mixed-effects (M-E) elastic net with $\alpha = 0.3$ and the classical mixed-effects models applied to the simulated data described in Sect. 4.1. The table gives the mean AUROC value based on ordering the variables, the Mean Squared Errors (MSEs) of the out-of-sample observations, $\mathbf{y_{out}}$, the MSEs of the fixed effects coefficients, $\mathbf{w}$, and the mean WAIC scores for each method. An extended version of these results is given in Tables 4–7 in the online supplementary materials

In terms of variable selection, the AUROC values shown in Fig. 4 and Table 1 show that all the SABRE methods outperform the alternative methods; the mixed-effects LASSO, the mixed effects elastic net and the classical mixed effects models. This is achieved across all datasets and is highlighted in Fig. 5, which compares the difference in AUROC values obtained by the different methods and that of the conjugate SABRE method. A negative score signifies a reduction in performance compared to the conjugate SABRE method. Figure 5 shows that the conjugate SABRE method performs significantly better than the mixed-effects LASSO, the mixed-effects elastic net and the classical mixed-effects models in all sets of data.

The performance in terms of predicting out of sample observations and inferring fixed effects coefficients shown in Table 1 again shows the SABRE methods outperforming the alternative methods in most cases. Table 1 shows a huge improvement for the SABRE methods in all cases except where both the error variance and number of variables is small. This is especially the case with the mixed-effects LASSO and the mixed-effects elastic net where the reliance on ℓ_1 regularisation causes a bias which affects both the inference of the fixed effects coefficients and the variable selection, as well as subsequently the out of sample predictions. The alternative methods do outperform the SABRE methods in some sets of data, however this is limited to a small number of cases which can mainly be explained by the model selection technique used with the SABRE methods not accurately selecting the correct cut-off when the AUROC value is close to 1.

We have also explored multiple different versions of the SABRE method, namely the semi-conjugate (Figure 1 of the online supplementary materials), conjugate (Fig. 1) and binary mask conjugate (Figure 3 of the online supplementary materials) SABRE methods. As far as we are aware the quantitative comparison between a spike and slab based method and a binary mask based one is the first of its kind. Our results given in Table 1, as well as Figs. 4 and 5 , show a strong similarity in performance between the methods. The comparison of AUROC values given in Fig. 5 clearly shows a large overlap in both methods' variable selection performance and this is backed up by the paired t-tests given in Table 5 of the online supplementary materials. The only exception to this is the AUROC values for the sets of data with $n = 50$ and $||\mathbf{w} = 200||$, which show a small significant improvement for the spike and slab model. This result is not repeated across different datasets so it is logical to conclude that these models perform similarly in general. Identifying that these methods give similar results is important, as in practise both methods are discussed and used throughout the literature, e.g. Murphy (2012), and Jow et al. (2014).

We have also compared the conjugate and semi-conjugate SABRE models, as depicted in Fig. 1 here and Figure 1 in the online supplementary materials. Overall, our results, shown in Table 1 and Tables 4–7 of the supplementary material, suggest that the two methods perform similarly across the wide range of simulated data sets. A paired t-test, summarised in Table 6 of the supplementary material, identifies two data sets ($||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.3$; $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.3$) where the conjugate SABRE model outperforms the semi-conjugate SABRE model. Formal model selection based on WAIC also shows a slight, but significant preference for the conjugate model in 9 out of 12 sets of data (see Tables 4–7 of the online supplementary material).

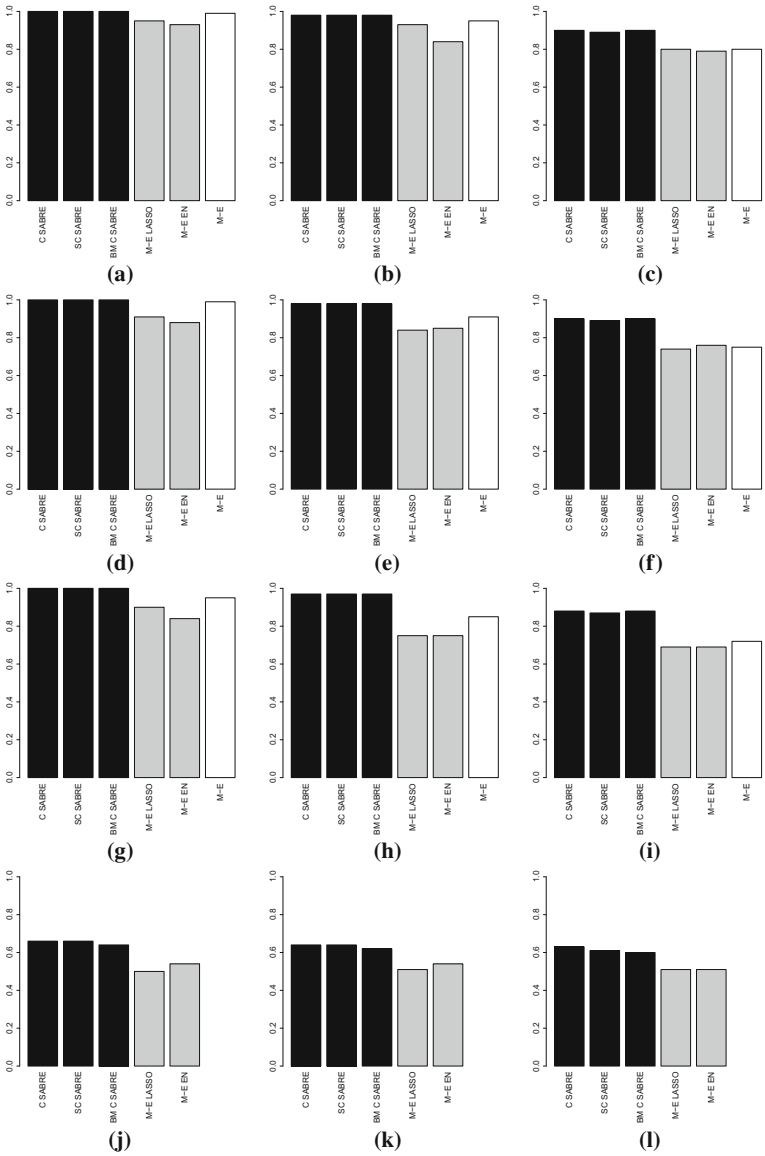

Fig. 4 *Bar plots* of AUROC values from the Simulation Study Results in Table 1. The *bar plots* give AUROC values for the Conjugate (C), Semi-Conjugate (SC) and Binary Mask Conjugate (BM C) SABRE methods (*black bars*), the mixed-effects (M-E) LASSO, the mixed-effects elastic net (M-E EN) with $\alpha = 0.3$ (both *grey bars*) and the classical mixed-effects models (*white bars*) applied to the simulated data described in Sect. 4.1. **a** $N = 100$, $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.03$. **b** $N = 100$, $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.1$. **c** $N = 100$, $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.3$. **d** $N = 100$, $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.03$. **e** $N = 100$, $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.1$. **f** $N = 100$, $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.3$. **g** $N = 100$, $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.03$. **h** $N = 100$, $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.1$. **i** $N = 100$, $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.3$. **j** $N = 50$, $||\mathbf{w}|| = 200$, $\sigma_\varepsilon^2 = 0.03$. **k** $N = 50$, $||\mathbf{w}|| = 200$, $\sigma_\varepsilon^2 = 0.1$. **l** $N = 50$, $||\mathbf{w}|| = 200$, $\sigma_\varepsilon^2 = 0.3$

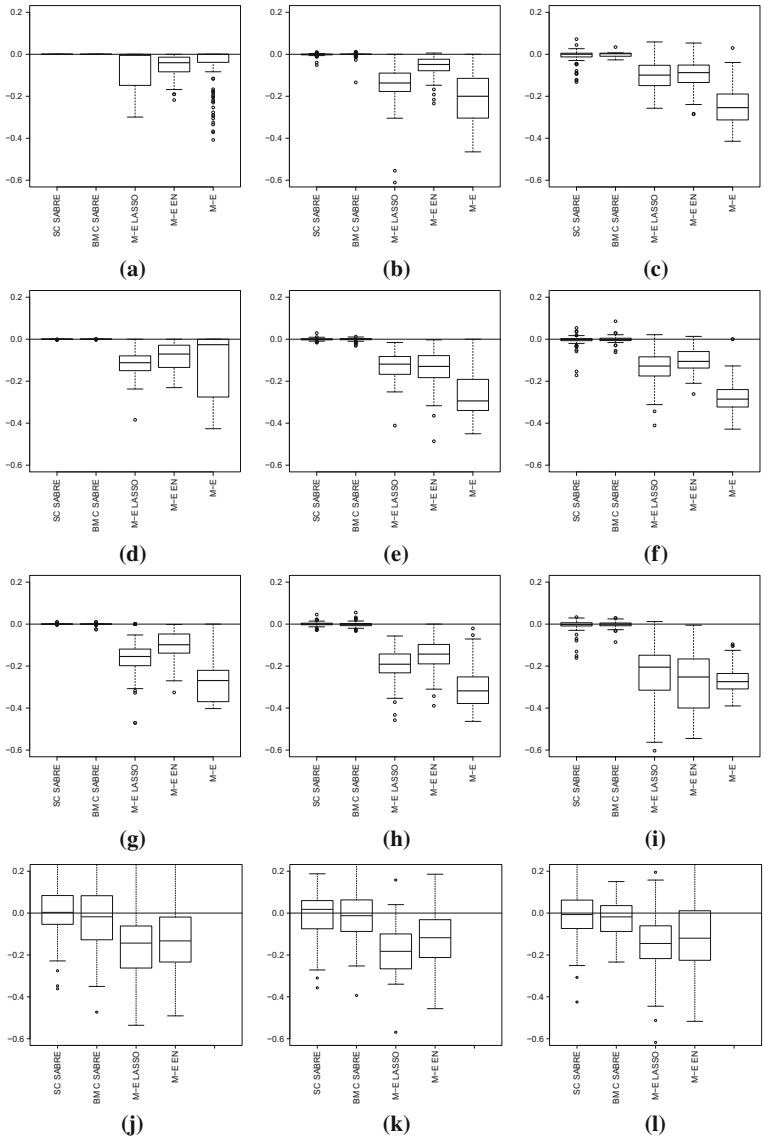

Fig. 5 *Box plots* of the difference in AUROC values for each method in comparison to the conjugate SABRE method. The *box plots* give the difference in AUROC values for each of the methods after the AUROC value of the conjugate SABRE method has been subtracted for the appropriate dataset. *Negative values* indicate that the conjugate method has outperformed the alternative method. Each *box plot* contains 100 datasets as described in Sect. 4.1. The alternative methods are the Semi-Conjugate (SC) and Binary Mask Conjugate (BM C) SABRE methods, the mixed-effects (M-E) LASSO, the mixed-effects elastic net (M-E EN) with $\alpha = 0.3$ and the classical mixed-effects models. **a** $N = 100$, $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.03$. **b** $N = 100$, $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.1$. **c** $N = 100$, $||\mathbf{w}|| = 40$, $\sigma_\varepsilon^2 = 0.3$. **d** $N = 100$, $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.03$. **e** $N = 100$, $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.1$. **f** $N = 100$, $||\mathbf{w}|| = 60$, $\sigma_\varepsilon^2 = 0.3$. **g** $N = 100$, $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.03$. **h** $N = 100$, $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.1$. **i** $N = 100$, $||\mathbf{w}|| = 80$, $\sigma_\varepsilon^2 = 0.3$. **j** $N = 50$, $||\mathbf{w}|| = 200$, $\sigma_\varepsilon^2 = 0.03$. **k** $N = 50$, $||\mathbf{w}|| = 200$, $\sigma_\varepsilon^2 = 0.1$. **l** $N = 50$, $||\mathbf{w}|| = 200$, $\sigma_\varepsilon^2 = 0.3$

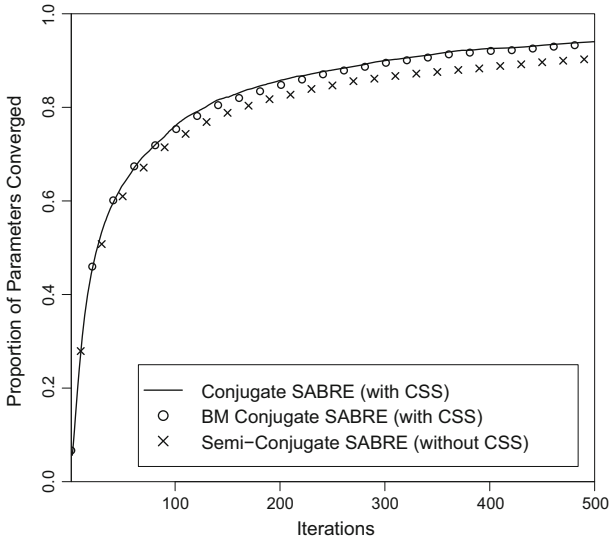

Fig. 6 Convergence diagnostics for the combined simulated datasets. Convergence diagnostics for the conjugate SABRE method with the collapsed sampling scheme (CSS) (*solid line*), the semi-conjugate SABRE method without CSS (*crosses*) and the BM conjugate SABRE method with CSS (*circles*). The *lines* show the proportion of parameters converged (PSRF < 1.05) versus the number of iteration of the 4 MCMC chains. The proportion is based on all of the simulated datasets with $n = 100$ from Sect. 4.1

The final contribution of our simulation study is to test whether the use of the collapsed sampling scheme in conjunction with increased conjugacy achieves an improvement in terms of MCMC mixing and convergence. Figure 6 indicates that a slight improvement is achieved with the conjugate SABRE model over the semi-conjugate one when the methods were tested on the sets of data with $n = 100$. However, this difference is not statistically significant, as becomes clear when considering the confidence intervals (not shown in Fig. 6 to avoid clutter). This finding suggests that the major bottleneck in the MCMC sampling scheme is caused by the latent variables $\boldsymbol{\gamma}$ rather than the regression parameters.

6.2 Real data with partial ground truth

Both SAT1 datasets have been analysed using classical mixed-effects models. Originally Reeve et al. (2010) analysed the original SAT1 dataset (Sect. 4.2) and Reeve et al. (2016) investigated an extended version of this dataset (Sect. 4.3). We have used our method on each of these datasets in order to identify a number of candidate residues which could be considered important for understanding antigenic variability. Knowledge of which residues are antigenically important is partially incomplete. Therefore, for validation purposes, residues were assigned to three different groups, proven, plausible and implausible, based on how likely they are to be antigenic based on experimental results.

For the SAT1 FMDV serotype, residues are included in the experimentally *proven* group for three different reasons. Firstly we include any residues which have been

experimentally validated as important within the SAT1 serotype by monoclonal antibody escape mutant studies (MAbs) (Grazioli et al. 2006). Secondly, we include those residues which are part of cords of connected experimentally validated antigenic residues for four or more different serotypes; VP1 140–169 (part of the VP1 G-H loop), VP1 200–224 (VP1 C terminus), VP2 70–82 (VP2 B-C loop) and VP3 56–61 (VP3 B-B knob) (Aktas and Samuel 2000; Barnett et al. 1989; Crowther et al. 1993a; Baxt et al. 1989; Bolwell et al. 1989; Grazioli et al. 2006, 2013; Lea et al. 1994; Kitson et al. 1990; Mateu 1995; Saiz et al. 1991; Thomas et al. 1988a, b). As antigenic sites have been found in a large number of different individual locations, we include additional information from other serotypes when classifying whole loops due the similar structure of the different serotypes. Finally, we also include a number of topotype-defining branches that are known to represent significant changes in the evolutionary history (Reeve et al. 2010).

We define the *plausible* group to consist of residues from any protein loop where residues have been identified in three or less FMDV serotypes, excluding those residues that are already classified as proven. Additionally, any non-topotype-defining branches of the phylogenetic trees are included in the plausible group, as it is unknown which of the remaining branches may also be significant in evolutionary history of the serotype. Finally we classify any residues not included in these groups as *implausible*.

It is common that variables have correlation coefficients exactly equal to one. In this case we only include one of the variables in the model and use Table 3 of the online supplementary materials to guide the classification into the proven, plausible and implausible groups, as explained in Section 7 of the online supplementary materials.

6.2.1 SAT1 data

The analysis of the original SAT1 dataset has resulted in the identification of 29 residues or branches of importance based on taking the top $J\hat{\pi}$ variables with the highest marginal posterior inclusion probabilities. 9 of the selected residues and 2 of the branches are classified as proven, at the expense of only 1 implausible variable. A full list of selected variables can be found in Section 8.2 of the online supplementary materials. The proportion of the differently classified variables at different cut-off points is shown in Fig. 7a. The proven residues include several that have been validated using MAbs in the SAT1 serotype (Grazioli et al. 2006), as well as others from the VP2 B-C loop, VP1 G-H loop and VP1 C-terminus (the end of the VP1 protein) and we have focused on these proven residues in our analysis.

The residues that have been experimentally validated in the SAT1 serotype are VP3 71 and VP3 77 in the VP3 B-C loop and VP1 144 and VP1 149 in the VP1 G-H loop (Grazioli et al. 2006). Additionally in the VP1 G-H loop, an antigenic loop in every FMDV serotype (Bolwell et al. 1989; Crowther et al. 1993b; Grazioli et al. 2006, 2013; Kitson et al. 1990; Lea et al. 1994) known to distract the host immune systems, the conjugate SABRE method has also identified VP1 143 and VP1 150. These residues are next to the experimentally validated residues in the protein alignment and confirm that the VP1 G-H loop is a highly antigenic part of the SAT1 serotype.

In addition to the residues in the VP3 B-C and VP1 G-H loops, the conjugate SABRE method has additionally selected VP2 74 in the VP2 B-C loop, as well as VP1 216 and

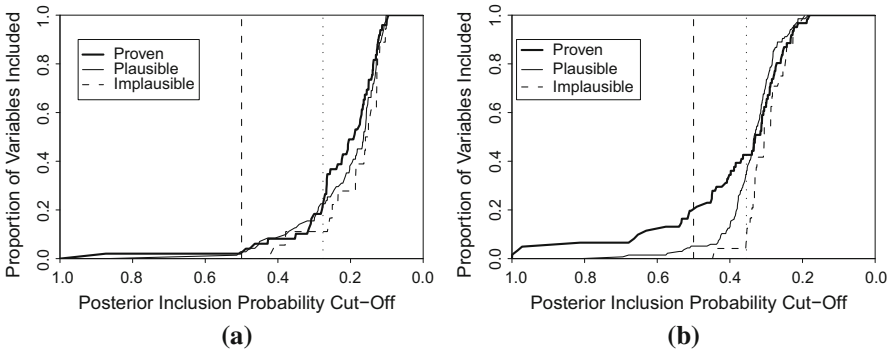

Fig. 7 Proportion of categorised SAT1 variables included based on different cut-off values for posterior inclusion probability. The graph shows the proportion of the experimentally proven (*thick solid line*), plausible (*solid line*) and implausible (*dashed line*) variables based on a cut-off value for the posterior inclusion probability. The variables were classified into groups based on the method outlined in the first 3 paragraphs of Sect. 6.2. Cut-offs are marked at 0.5 posterior inclusion probability (*vertical dashed line*) and the posterior inclusion probability equivalent to the top $J\hat{\pi}$ variables with the highest posterior inclusion probabilities (*vertical dotted line*). **a** Original SAT1, **b** extended SAT1

VP1 219 in the VP1 C-terminus. The VP2 B-C loop is antigenic in all serotypes and contains the highly antigenic VP2 72 residue, which has been experimentally validated in all of the FMDV serotypes except SAT2 (Aktas and Samuel 2000; Crowther et al. 1993a; Grazioli et al. 2006, 2013; Kitson et al. 1990; Lea et al. 1994; Saiz et al. 1991). The VP1 C-terminus has been proven to be antigenic in all but the Asia1 serotype, although it is almost certainly antigenic there also (Aktas and Samuel 2000; Baxt et al. 1989; Grazioli et al. 2006; Mateu 1995).

Figure 8a shows the model predictions for the antigenically significant branches based on using just the branch variables from the original SAT1 dataset. Here we have identified all of the branches known to divide topotypes (Reeve et al. 2010), as well as a number of other branches. Several of the branches, including two topotype defining branches, have been specifically identified as reactivity, immunogenic or antigenic changes, an improvement over previously used models.

6.2.2 Extended SAT1 data

The analysis of the extended SAT1 dataset resulted in selecting 76 variables, which included 24 proven residues, 4 important branches in the evolutionary history and only 2 implausible residues. A full list of the selected variables can again be found in the online supplementary materials, Section 8.2, and the proportion of proven, plausible and implausible residues selected at different cut-offs is shown in Fig. 7b here. The improved results over Sect. 6.2.1 show the advantage of getting a larger dataset through testing an increased number of strains under a variety of different experimental conditions.

The conjugate SABRE method has identified 11 residues in the highly variable VP1 G-H loop (VP1 142, VP1 143, VP1 144, VP1 147, VP1 148, VP1 149, VP1 150, VP1 155, VP1 156, VP1 163 and VP1 164). Finding this many significant residues in this

highly antigenic region while keeping the number of implausible residues low shows that the model is working effectively.

Additionally, like with the original SAT1 dataset in Sect. 6.2.1, the conjugate SABRE method has selected VP2 74 from the VP2 B-C loop. However in addition it has also selected VP2 72 which is antigenic in all FMDV serotypes and VP2 79 which has been experimentally validated in the A, O, Asia1 and SAT2 serotypes (Grazioli et al. 2006, 2013; Mateu 1995). The conjugate SABRE model also again selects several residues from the VP1 C-terminus; VP1 209, VP1 211 and VP1 218.

The final proven residues are from the VP3 B-B knob or have been experimentally validated specifically in the SAT1 serotype (Grazioli et al. 2006). In the VP3 B-B knob the conjugate SABRE method has identified VP3 58 (serotypes A, O, C and Asia1) and VP3 61 (serotype A) (Grazioli et al. 2006; Lea et al. 1994; Mateu 1995). From those residues which have specifically been validated in the SAT1 serotype, again VP3 71 and VP3 77 from the VP3 B-C loop have been selected. However for the extended SAT1 dataset, the conjugate SABRE method has also selected VP3 138, which was also found in Reeve et al. (2010), from VP3 E-F loop.

As well as finding some branches in our overall model (including 4 topotype defining branches identified as representing significant evolutionary changes a priori), we have also compiled a model based only on branches to help us understand the evolutionary history of the serotype. The results of this model are given in Fig. 8b, where the seven branches known to define topotypes are indicated by the vertical line. In order to produce more interpretable results, where larger groups of strains are not separated by a significant evolutionary change (selected branch), we have used a cut-off of 0.5. The full results using a $J\hat{\pi}$ cut-off are given in Figure 5 of the online supplementary materials. The results given in Fig. 8b show that we have been able to identify all but one of the topotype defining branches, while the other is found when the $J\hat{\pi}$ cut-off is used. We have also been able to specify whether the evolutionary changes have affected virus antigenicity, immunogenicity or reactivity, helping us to further understand the underlying biological processes.

6.2.3 Comparison with previous work

To compare the results of the SABRE method against the mixed-effects models used in Reeve et al. (2010, 2016) and Harvey et al. (2016), we examine which categories (proven, plausible or implausible) the various residues selected fall into. Note that to do this we ignore any branch terms that do not directly correspond to a residue term. The full results for variables selected can be found in Tables 6, 8,10 and 13 of Section 8 in the online supplementary materials. For comparison, the results of Reeve et al. (2016) are given in Table 12 of the online supplementary materials, as the results of the equivalent study are not given in the original paper. We also note that the results of Harvey et al. (2016) are not completely comparable as they were obtained from a larger dataset.

For the original SAT1 dataset, Reeve et al. (2010) selected 0 proven, 0 plausible and 0 implausible residues using the method described in Sect. 2 (i.e. when the Holm–Bonferroni correction was used). These results compare to 1 proven, 1 plausible and 0 implausible residues when the conjugate SABRE method was used and selecting

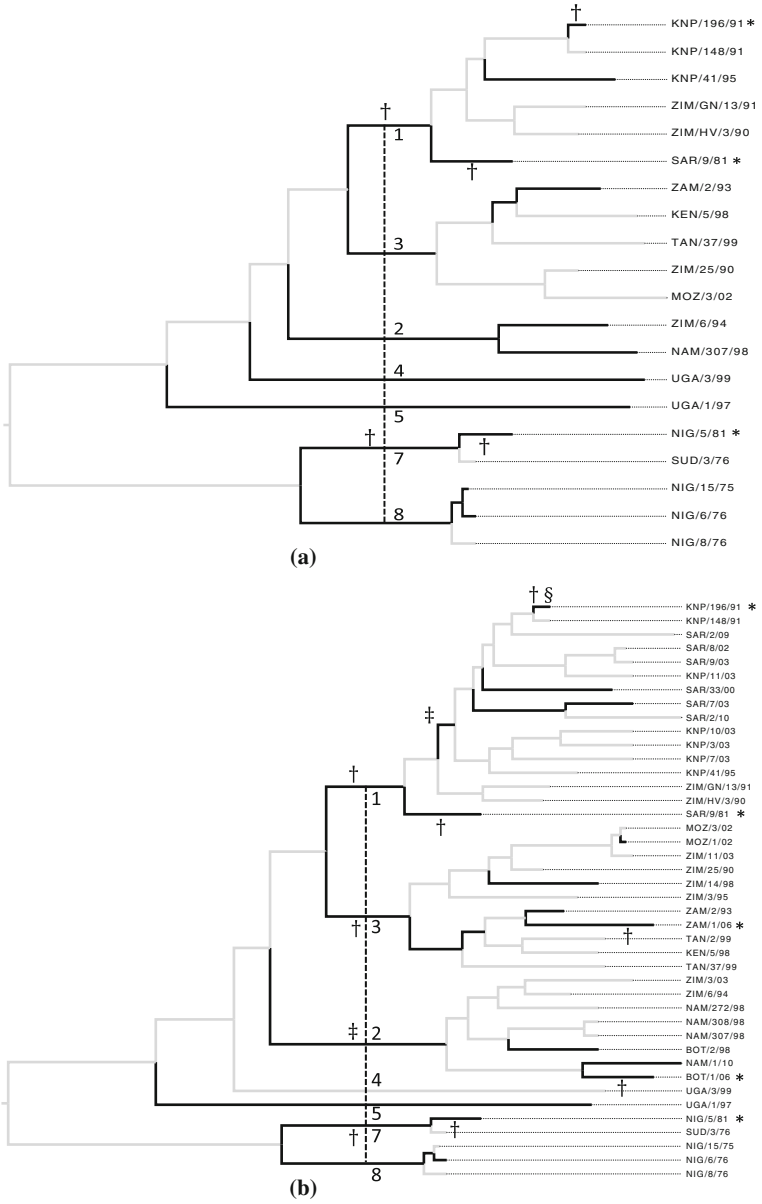

Fig. 8 Phylogenetic trees indicating significant branches in the evolutionary history of the SAT1 serotype. Phylogenetic trees were created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. Marked on the tree are protective strains (*asterisk*) and topotype defining branches (*dashed vertical line*). Branches inferred by the SABRE method are highlighted (*black*). *Symbols* indicate whether this was inferred to be a change in virus antigenicity (*dagger*), virus reactivity (*double dagger*) or virus immunogenicity (*section mark*). Where a highlighted branch has no symbol, an associated change in antigenicity or reactivity could not be discriminated between. The cut-off for significance is discussed in Sects. 6.2.1 and 6.2.2. **a** Original SAT1, **b** extended SAT1

any residue variables with a marginal posterior inclusion probability of greater than or equal to 0.5.[3] We have also looked at how well the methods do before selecting an implausible variable or before a p value of greater than 0.05 (before the Holm–Bonferroni correction was used) was reached (in Reeve et al. (2010) the variable selection process was stopped as soon as a 0.05 p value was reached). In this situation again the conjugate SABRE method offers an improvement, selecting 5 proven, 5 plausible and 0 implausible residues compared to 1, 1 and 0 respectively for the classical mixed-effects models. The difference in these results shows an advantage for the conjugate SABRE method over the classical mixed-effects models.

In the extended SAT1 dataset, Reeve et al. (2016) used the method in Sect. 2 to select 5 proven, 0 plausible and 0 implausible residues, or 8, 1 and 0, respectively, if the method continued until selecting the first implausible residue. The conjugate SABRE method selected 11 proven, 3 plausible and 0 implausible residues when taking any variables with marginal posterior inclusion probabilities of greater than or equal to 0.5, or 15, 4 and 0, respectively, before selecting the first implausible residue.[4] It can again be seen that the power of the proposed SABRE method has improved over the method of Reeve et al. (2010).

6.2.4 Sampling of latent indicators

Figure 9 compares component-wise Gibbs sampling (Sect. 3.6.1) against block Metropolis–Hastings sampling (Sect. 3.6.2) in terms of speed of convergence. To do this we ran 4 chains for the component-wise Gibbs sampler and each of the 5 variations of the Metropolis–Hastings sampler, monitoring the PSRFs for each parameter in the different methods. Figure 9 shows the proportion of parameters with PSRFs < 1.05 in each case compared with the CPU time taken to get that number of samples. The higher the proportion of parameters with PSRFs < 1.05, the better the method is said to have performed.

The results from Fig. 9 support the advantage of a block Metropolis–Hastings sampler over a component-wise Gibbs sampler. In all of the datasets the block Metropolis–Hastings samplers have outperformed the component-wise Gibbs sampler, with the exception of when more than 40 or 50 variables were sampled at a time (not shown in the diagrams for clarity). This shows that even sampling a reasonably large number of variables simultaneously, where the acceptance rate is likely to be low, can still yield a notable improvement. The results[5] in Fig. 9 suggest that as a rule of thumb, sampling about 10 or 15 (or around 7%) of the variables at a time will lead to effective sampling with the quickest convergence.

[3] The power can be further improved (12 proven and 9 plausible residues) by inferring the selection threshold and selecting the top $J\hat{\pi}$ variables, at the expense of the selection of 1 implausible residue.

[4] The power can be further improved (24 proven and 15 plausible residues) by inferring the selection threshold and selecting the top $J\hat{\pi}$ variables, at the expense of the selection of 2 implausible residues.

[5] The best performing samplers in Fig. 9 are as follows: Metropolis–Hastings samplers with 10 (7.2%) or 15 (10.9%) variables at a time for the original SAT1 dataset and with 10 (4.5%) or 15 (6.8%) variables at a time for the extended SAT1 dataset.

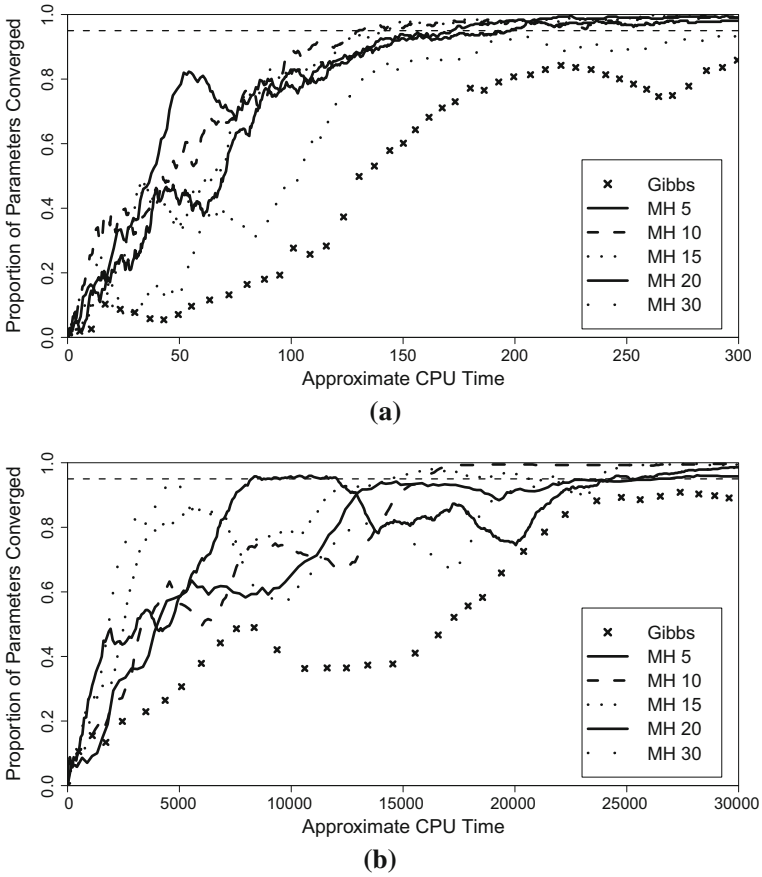

Fig. 9 Convergence diagnostics for the original SAT1 and extended SAT1 datasets. The *lines* show the proportion of parameters that have converged (PSRF < 1.05) versus the average CPU time (second) when using component-wise Gibbs sampling (*crosses*) and Metropolis–Hastings sampling proposing 5 (*solid*), 10 (*dashed*), 15 (*dotted*), 20 (*thick solid*) and 30 (*thick dotted*) inclusion parameters simultaneously. **a** Original SAT1, **b** extended SAT1

6.3 Real data for novel predictions

Little knowledge is available on how mutational changes affect antigenic variability for the SAT2 FMDV serotype and the H1N1 Influenza virus. We have therefore applied our conjugate SABRE method as a tool for new hypothesis generation.

6.3.1 SAT2 data

Although knowledge of the SAT2 FMDV serotype is minimal, for validation purposes we can exploit knowledge gained from other serotypes of FMDV and previous work on the SAT2 serotype. Grazioli et al. (2006) and Crowther et al. (1993b) has found evidence for antigenicity of the following three areas of the SAT2 capsid: VP1 140–

169 (part of the VP1 G-H loop), VP1 200–224 (VP1 C terminus) and VP2 70–82 (VP2 B-C loop). A full list of the variables selected by the conjugate SABRE method can be found in the online supplementary materials, Section 8.2.

Firstly in the VP2 B-C loop, the conjugate SABRE method has identified 5 residues that are antigenic; VP2 71, VP2 72, VP2 78, VP2 79 and VP2 80 (Grazioli et al. 2006, 2013; Kitson et al. 1990; Lea et al. 1994; Saiz et al. 1991). Of these VP2 78 has been experimentally identified using MAbs (Grazioli et al. 2006). Additionally VP2 72 is known to be antigenic in all other serotypes and these results suggest it is also antigenic in the SAT2 serotype (Grazioli et al. 2006, 2013; Mateu 1995).

The second region in which antigenically significant residues have been found is in the VP1 G-H loop. The VP1 G-H loop is known to be a highly variable distracter site designed to confuse the host immune system (Crowther et al. 1993b) and is antigenic in all of the FMDV serotypes. In this loop, the conjugate SABRE method has specifically identified VP1 144 and VP1 166, where it is notable that VP1 166 lies directly between several residues that have been experimentally validated in the SAT2 serotype using MAbs (Crowther et al. 1993b).

The final known antigenic region that has been identified by the conjugate SABRE method is part of the VP1 C-terminus, the end of the VP1 protein. In the VP1 C-terminus we have identified VP1 207, VP1 208, VP1 209, VP1 210 and VP1 211 which are part of a region known to be antigenic in all FMDV serotypes except Asia1 (Aktas and Samuel 2000; Grazioli et al. 2006; Lea et al. 1994; Saiz et al. 1991). With the conjugate SABRE method identifying all these neighbouring residues, it suggests that this section of the protein is a highly antigenic part of the SAT2 serotype.

Figure 10 gives the phylogenetic tree for the SAT2 serotype with the predicted significant evolutionary changes. Unlike the SAT1 serotype, there is no prior knowledge of which residues and branches are antigenically relevant and we therefore apply our method to generate genuinely new hypotheses. The results presented give our best prediction for the significant branches and show a couple of potentially interesting groupings which could represent functional groups for the SAT2 serotype.

6.3.2 Influenza A (H1N1) data

Knowledge of the H1N1 Infuenza serotype is restricted to a few experimental results and some knowledge about where antigenic sites are likely to occur. In general the virus can be divided into the head and stalk domains, with the head domain most likely to contain antigenic residues within four major antigenic sites (Sa, Sb, Ca and Cb) (Caton et al. 1982). But knowledge is incomplete as to which areas of the head domain contain antigenic residues beyond that. We have applied the conjugate SABRE method to the H1N1 dataset and a full list of the variables selected can be found in the online supplementary materials, Section 8.3.

Of those variables selected by the conjugate SABRE method, one residue was identified on the Receptor Binding Site (position 187 on the H1 common alignment), the main binding site for the H1N1 virus (Skehel and Wiley 2000). Additionally we have identified 4 other residues that are nearby (positions 130, 153, 189 and 190), including one on the Sa antigenic site, two on the Sb and one that is known to be a location of a major antigenic change (130) (Harvey et al. 2016). 6 other residues are

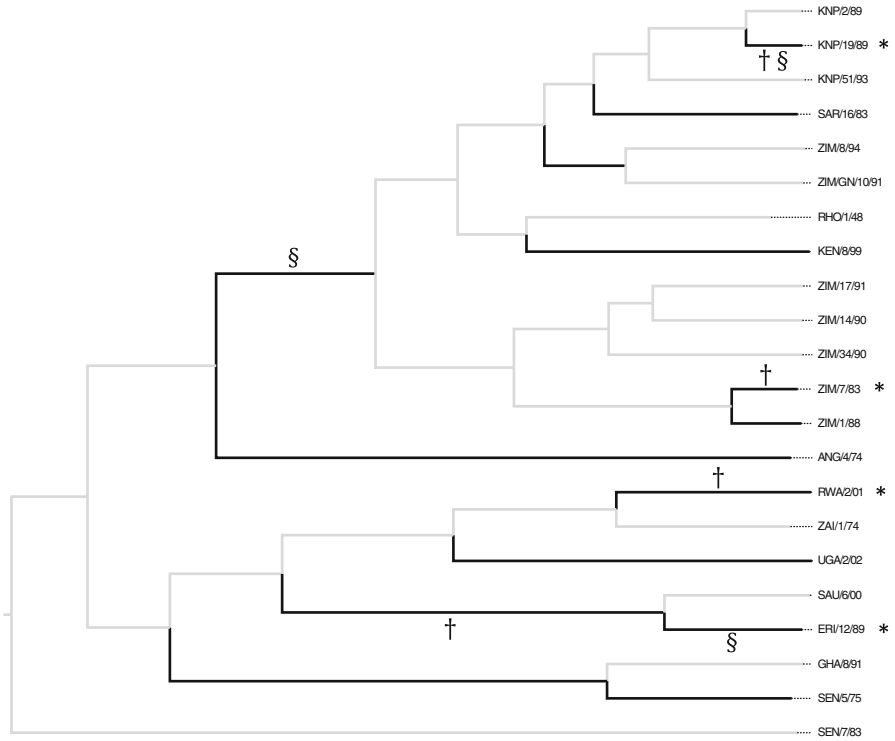

Fig. 10 Phylogenetic trees indicating significant branches in the evolutionary history of the SAT2 serotype. The phylogenetic tree was created using BEAST v1.7.2 and FigTree v1.4.2 from aligned nucleotide sequence data with date of isolation. Marked on the tree are protective strains (*asterisk*). Branches associated with a change in virus phenotype are highlighted (*black*). *Symbols* indicate whether this was inferred to be a change in virus antigenicity (*dagger*), virus reactivity (none-identified) or virus immunogenicity (*section mark*). Where a highlighted branch has no symbol, an associated change in antigenicity or reactivity could not be discriminated between. The cut-off for significance was taken to be the $J\hat{\pi}$ variables with the highest marginal inclusion probability

identified on the Ca and Cb antigenic site (positions 69, 72 and 74 on the Cb; positions 139, 141 and 142 on the Ca). Other residues were also found near to the Ca and Cb antigenic sites, but with a lack of experimental knowledge about the H1N1 virus it is impossible to tell whether they are antigenic sites without experimental validation.

7 Conclusion

We have addressed the problem of identifying the residues within the SAT1 and SAT2 serotypes of FMDV and Influenza A (H1N1) that are responsible for changes in antigenic variability. This allows us to identify which residues must remain the same in order for two strains to cross react and for one strain to potentially be used as an effective vaccine against another. Identifying such residues can reduce the number of strains that must be tested as a vaccine, potentially reducing the time and cost associated with the selection procedure.

We have proposed a sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution (SABRE) and shown how it offers improvement over the classical mixed-effects model, the mixed-effects LASSO and the mixed-effects elastic net. There are four reasons for this improvement. The proposed hierarchical modelling framework with slab-and-spike prior (1) avoids the bias inherent in Lasso-type methods, (2) genuinely and consistently achieve sparsity, (3) properly accounts for uncertainty at all levels of inference, and (4) borrows strength from information coupling, whereby all parameters are systematically and iteratively inferred in the context of all other parameters. In some more detail: (1) The shrinkage effect inherent in the ℓ_1 penalty term introduces a bias by which the regression parameters are systematically underestimated. This bias is avoided with the slab and spike prior that we use. (2) The LASSO is known to only give sparse solutions at the MAP (maximum a posteriori) configuration, but not when sampling parameters from the posterior distribution. From a Bayesian perspective, the MAP is methodologically inconsistent, as it is not guaranteed to represent the region in parameter space with the highest probability mass. The spike-and-slab prior, which we use, avoids this methodological inconsistency and achieves sparsity in a sound Bayesian inference context. (3) In our hierarchical Bayesian models, all sources of uncertainty are properly accounted for. The higher-level hyperparameters have their own distributions, which are systematically inferred from the data. In contrast, the regularisation parameters of the established methods are typically fixed, set e.g. by cross-validation, but without taking their uncertainty into account (see also Chapter 5 in Gelman et al. (2013) for a more detailed discussion). (4) In our approach, we explicitly model all dependencies among the variables, and inference is carried out within the context of the whole system. This systematically borrows strength from information coupling and avoids the piecemeal approach of established methods.

There are two fundamentally different approaches to variable selection in Bayesian hierarchical models: the slab-and-spike prior, whereby the influence of an input variable is controlled via the prior distribution of its associated regression parameters, and the binary mask model, where variables are put through a binary multiplicative filter. The difference is depicted in Fig. 1 here and Figure 3 in the online supplementary materials. Which method is better? Standard textbooks, like Murphy (2012), describe both methods (see Chapter 13), but do not offer a comparative evaluation, and in the literature, authors rather arbitrarily tend to opt for one method or another (see e.g. Heydari et al. (2016)). We have carried out a systematic comparison to properly quantify the difference in terms of accuracy and computational efficiency, and found it to be negligible. We have also systematically evaluated the influence of the prior, comparing a conjugate with a non-conjugate prior, as depicted by Fig. 1 here and Figure 1 in the online supplementary materials, and we have assessed its influence systematically in terms of accuracy, computational efficiency, and formal model selection preference in Table 1. The differences in accuracy are negligible (see e.g. Fig. 5). The conjugate model has slightly better computational efficiency (Fig. 6), but this difference is not significant; this finding indicates that the bottleneck in the computational procedure is the sampling of the latent variables rather than the regression parameters. The conjugate model shows a slight but significant improvement over the non-conjugate model in a number of the model selection scores based on WAIC, as

seen from Table 1 here and Table 7 of the online supplementary materials, but this has little immediate impact on the variable selection. Overall, our findings demonstrate a remarkable robustness of the proposed hierarchical modelling framework with respect to minor model modifications, which boosts our confidence in the predictions and in the variable ranking.

Further to this we have investigated the sampling of latent inclusion variables. We have shown that by proposing multiple variables simultaneously through Metropolis–Hastings sampling it is possible to give a significant computational improvement over the conventional component-wise Gibbs sampler (Fig. 9). We have shown this improvement in a number of different datasets and have offered a general rule of thumb that proposing 10 or 15 (or around 7%) of the variables at a time will lead to good mixing within MCMC chains for a variety of different datasets.

Through the use of this new model with the improved sampling techniques we have been able to identify an increased number of known antigenic sites in the SAT1 serotype of FMDV (Grazioli et al. 2006) compared to Reeve et al. (2010) and Reeve et al. (2016), while incurring no (for the default selection threshold 0.5) or only a very small number (for the inferred selection threshold $J\hat{\pi}$) implausible residues. Very little biological knowledge exists about the SAT2 serotype, and a previous in silico application has failed to make any predictions at all (Reeve et al. 2010). To our knowledge, our study is the first time that specific new hypotheses about genetic-antigenic associations have been made with an in silico model based on the currently available data. Additionally we have provided an insight into the evolutionary history of the SAT serotypes (Figs. 8 and 10) and have provided a novel way of interpreting the biological effects of these virus mutations. Finally we have identified a number of significant antigenic sites in the H1N1 Influenza virus and provided new hypotheses for this virus.

7.1 Future work

Further work to follow on from this paper comes in several forms. Firstly we would like to find a way of effectively using the proposed model on the full H1N1 dataset used in Harvey et al. (2016), without having to summarise and reduce the data as we have done here. With 15,693 measurements, currently the model is too computationally expensive to get a reasonable number of posterior samples. More computationally efficient methods, such as variational inference or expectation propagation (Minka 2001), can offer a solution to this problem at the expense of an approximation. Methods for using spike and slab priors have been proposed for both of these techniques, e.g. Titsias and Lázaro-Gredilla (2011) and Hernández-Lobato et al. (2013), and these can be extended to create alternative, faster versions of the SABRE methods.

Secondly we would like to investigate how different types of mutations affect antigenic variability. At present the data simply consists of indicators of mutational changes that occur without any regard to the type of mutation, something which is addressed in Reeve et al. (2016). Adding this information will dramatically increase the number of variables on which selection must be made and is likely to make inferring $\boldsymbol{\gamma}$ more difficult. To address this we will likely require improved proposal distributions

for $\boldsymbol{\gamma}$, as has been used for continuous variables in Haario et al. (2006), which account for the estimated posterior correlations between these latent inclusion parameters. A method to generate correlated binary variables has been proposed in Leisch et al. (1988). However we would require a density function to put into the Metropolis–Hastings ratio in order to use this within MCMC sampling.

With more information, and therefore more variables, relating to the mutations being included in the models, it may be necessary to add additional information sharing between the latent indicator variables, $\boldsymbol{\gamma}$. Latent Gaussian processes can be used to model this, where inference can be achieved in a variety of ways, e.g. Filippone et al. (2013) and Andersen et al. (2014). The use of latent Gaussian processes would allow us to introduce correlations between mutations of the same type or mutations occurring in similar location on the surface of the virus shell. This can potentially allow us to identify which types of mutations are important, as well as identifying complete antigenic regions rather than just individual residues.

References

Aderhold A, Husmeier D, Grzegorczyk M (2014) Statistical inference of regulatory networks for circadian regulation. Stat Appl Genet Mol Biol 13(3):227–273

Aktas S, Samuel AR (2000) Identification of antigenic epitopes on the foot and mouth disease virus isolate O-1/Manisa/Turkey/69 using monoclonal antibodies. Sci Tech Rev Office Int Epizoot 19(3):744–753

Andersen MR, Winther O, Hansen LK (2014) Bayesian inference for structured spike and slab priors. Adv Neural Inf Process Syst 27:1745–1753

Andrieu C, Doucet A (1999) Joint bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. IEEE Trans Signal Process 47(10):2667–2676

Barbieri L, Berger J (2004) Optimal predictive model selection. Ann Stat 32(3):870–897

Barnett P, Ouldridge E, Rowlands D, Brown F, Parry N (1989) Neutralizing epitopes of type O foot-and-mouth disease virus. I. Identification and characterization of three functionally independent, conformational sites. J Gen Virol 70(Pt 6):1483–1491

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. doi:10.18637/jss.v067.i01

Baxt B, Vakharia V, Moore D, Franke A, Morgan D (1989) Analysis of neutralizing antigenic sites on the surface of type A12 foot-and-mouth disease virus. J Virol 63(5):2143–2151

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Bolwell C, Brown A, Barnett P, Campbell R, Clarke B, Parry N, Ouldridge E, Brown F, Rowlands D (1989) Host cell selection of antigenic variants of foot-and-mouth disease virus. J Gen Virol 70(Pt 1):45–57

Caton AJ, Brownlee GG, Yewdell JW, Gerhard W (1982) The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 31(2 Pt 1):417–427

Crowther J, Farias S, Carpenter W, Samuel A (1993a) Identification of a fifth neutralizable site on type O foot-and-mouth disease virus following characterization of single and quintuple monoclonal antibody escape mutants. J Gen Virol 74(Pt 8):1547–1553

Crowther J, Rowe C, Butcher R (1993b) Characterization of monoclonal antibodies against a type SAT 2 foot-and-mouth disease virus. Epidemiol Infect 111(2):391–406

Dalton L, Dougherty E (2012) Exact sample conditioned MSE performance of the Bayesian MMSE estimator for classification error—part II: consistency and performance analysis. IEEE Trans Signal Process 60(5):2588–2603

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499
Filippone M, Zhong M, Girolami M (2013) A comparative evaluation of stochastic-based inference methods for Gaussian process models. Mach Learn 93:93–114
Gelman A (2004) Parameterization and bayesian modeling. J Am Stat Assoc 99(466):537–545
Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Anal 1(3):515–534
Gelman A, Rubin D (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–511
Gelman A, Carlin JB, Stern HS, Dunson DB, Ventari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman & Hall, London
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741
George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88(423):881–889
George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. Stat Sin 7:339–373
Grazioli S, Moretti M, Barbieri I, Crosatti M, Brocchi E (2006) Use of monoclonal antibodies to identify and map new antigenic determinants involved in neutralisation on FMD viruses type SAT 1 and SAT 2. In: Report of the session of the research group of the standing technical committee of the European commission for the control of foot-and-mouth disease, pp 287–297, appendix 43
Grazioli S, Fallacara F, Brocchi E (2013) Mapping of antigenic sites of foot-and-mouth disease virus serotype Asia 1 and relationships with sites described in other serotypes. J Gen Virol 94(3):559–569
Grzegorczyk M, Husmeier D (2013) Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. Mach Learn 91:105–151
Haario H, Laine M, Mira A, Saksman E (2006) DRAM: efficient adaptive MCMC. Stat Comput 16(4):339–354
Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
Harvey WT, Gregory V, Benton DJ, Hall JP, Daniels RS, Bedford T, Haydon DT, Hay AJ, McCauley JW, Reeve R (2016) Identifying the genetic basis of antigenic change in influenza A (H1N1). arXiv preprint arXiv:1404.4197
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, Berlin
Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109
Hernández-Lobato D, Hernández-Lobato JM, Dupont P (2013) Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. J Mach Learn Res 14(1):1891–1945
Heydari J, Lawless C, Lydall DA, Wilkinson DJ (2016) Bayesian hierarchical modelling for inferring genetic interactions in yeast. J R Stat Soc Ser C (Appl Stat) 65(3):367–393
Hirst GK (1942) The quantitative determination of influenza virus and antibodies by means of red cell agglutination. J Exp Med 75(1):49–64
Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S (1982) Rapid evolution of RNA genomes. Science 215:1577–1585
Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
Jow H, Boys RJ, Wilkinson DJ (2014) Bayesian identification of protein differential expression in multi-group isobaric labelled mass spectrometry data. Stat Appl Genet Mol Biol 13(5):531–551
Kitson J, McCahon D, Belsham G (1990) Sequence analysis of monoclonal antibody resistant mutants of type O foot and mouth disease virus: evidence for the involvement of the three surface exposed capsid proteins in four antigenic sites. Virology 179(1):26–34
Knowles N, Samuel A (2003) Molecular epidemiology of foot-and-mouth disease virus. Virus Res 91:65–80
Lea S, Hernandez J, Blakemore W, Brocchi E, Curry S, Domingo E, Fry E, Abu Ghazaleh R, King A, Newman J, Stuart D, Mateu M (1994) The structure and antigenicity of a type C foot-and-mouth disease virus. Structure 2(2):123–139
Leisch F, Weingessel A, Hornik K (1988) On the generation of correlated artificial binary data. Working paper series, Working paper no. 13. SFB "Adaptive information systems and modelling in economics and management science". Vienna University of Economics and Business Administration, Wien, Austria. http://www.wu-wien.ac.at/am
Mateu M (1995) Antibody recognition of picornaviruses and escape from neutralization: a structural view. Virus Res 38(1):1–24
Mattion N, König G, Seki C, Smitsaart E, Maradei E, Robiolo B, Duffy S, León E, Piccone M, Sadir A, Bottini R, Cosentino B, Falczuk A, Maresca R, Periolo O, Bellinzoni R, Espinoza A, Torre J, Palma

E (2004) Reintroduction of foot-and-mouth disease in Argentina: characterisation of the isolates and development of tools for the control and eradication of the disease. Vaccine 22:4149–4162

McDonald NJ, Smith CB, Cox NJ (2007) Antigenic drift in the evolution of H1N1 influenza A viruses resulting from deletion of a single amino acid in the haemagglutinin gene. J Gen Virol 88(Pt 12):3209–3213

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092

Minka TP (2001) Expectation propagation for approximate Bayesian inference. In: Proceedings of the seventeenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 362–369

Mitchell T, Beauchamp J (1988) Bayesian variable selection in linear regression. J Am Stat Assoc 83(404):1023–1032

Mohamed S, Heller K, Ghahramani Z (2012) Bayesian and l_1 approaches for sparse unsupervised learning. In: Proceedings of the 29th international conference on machine learning (ICML-12), pp 751–758

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge

Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103(482):681–686

Paton D, Valarcher J, Bergmann I, Matlho O, Zakharov V, Palma E, Thomson G (2005) Selection of foot and mouth disease vaccine strains—a review. Rev Sci Tech 24:981–993

Pinheiro JC, Bates D (2000) Mixed-effects models in S and S-PLUS. Springer, Berlin

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R News 6(1):7–11

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Reeve R, Blignaut B, Esterhuysen JJ, Opperman P, Matthews L, Fry EE, de Beer TAP, Theron J, Rieder E, Vosloo W, O'Neill HG, Haydon DT, Maree FF (2010) Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus. PLoS Comput Biol 6(12):e1001027

Reeve R, Borley DW, Maree FF, Upadhyaya S, Lukhwareni A, Esterhuysen JJ, Harvey WT, Blignaut B, Fry EE, Parida S, Paton DJ, Mahapatra M (2016) Tracking the antigenic evolution of foot-and-mouth disease virus. PloS ONE 11(7):1–17

Ripley B (1979) Algorithm AS 137: simulating spatial patterns: dependent samples from a multivariate density. J R Stat Soc Ser C 28(1):109–112

Ruyssinck J, Huynh-Thu V, Geurts P, Dhaene T, Demeester P, Saeys Y (2014) NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms. PLoS ONE 9(3):e92709

Sabatti C, James GM (2005) Bayesian sparse hidden components analysis for transcription networks. Bioinformatics 22(6):739–746

Saiz JC, Gonzalez MJ, Borca MV, Sobrino F, Moore DM (1991) Identification of neutralizing antigenic sites on VP1 and VP2 of type A5 foot-and-mouth disease virus, defined by neutralization-resistant variants. J Virol 65(5):2518–2524

Schelldorfer J, Bühlmann P, van de Geer S (2011) Estimation for high-dimensional linear mixed-effects models using $\ell 1$-penalization. Scand J Stat 38(2):197–214

Scott JG, Berger JO (2010) Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. Ann Stat 38(5):2587–2619

Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Ann Rev Biochem 69(1):531–569

Thomas A, Woortmeijer R, Barteling S, Meloen R (1988a) Evidence for more than one important, neutralizing site on foot-and-mouth disease virus. Brief report. Arch Virol 99(3–4):237–242

Thomas A, Woortmeijer R, Puijk W, Barteling S (1988b) Antigenic sites on foot-and-mouth disease virus type A10. J Virol 62(8):2782–2789

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288

Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective (with comments). J R Stat Soc Ser B 73(3):273–282

Titsias MK, Lázaro-Gredilla M (2011) Spike and slab variational inference for multi-task and multiple kernel learning. In: Advances in neural information processing systems, pp 2339–2347

Watanabe S (2010) Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 11:3571–3594

WHO (2011) Manual for the laboratory diagnosis and virological surveillance of influenza. http://whqlibdoc.who.int/publications/2011/9789241548090_eng.pdf

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67(2):301–320