Random-Effects Substitution Models for Phylogenetics via Scalable Gradient Approximations

ANDREW F. MAGEE[1], ANDREW J. HOLBROOK[1], JONATHAN E. PEKAR[2,3], ITZUE W. CAVIEDES-SOLIS[4],
FREDRICK A. MATSEN IV[5,6,7,8], GUY BAELE[9], JOEL O. WERTHEIM[10], XIANG JI[11], PHILIPPE LEMEY[9]
AND MARC A. SUCHARD[1,12,13,*]

[1]Department of Biostatistics, Jonathan and Karin Fielding School of Public Health, University of California - Los Angeles, Los Angeles, CA, USA
[2]Bioinformatics and Systems Biology Graduate Program, University of California - San Diego, La Jolla, CA, USA
[3]Department of Biomedical Informatics, University of California - San Diega, La Jolla, CA, USA
[4]Department of Biology, Swarthmore College, Swarthmore, PA, USA
[5]Howard Hughes Medical Institute, Seattle, Washington, USA
[6]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
[7]Department of Genome Sciences, University of Washington, Seattle, Washington, USA
[8]Department of Statistics, University of Washington, Seattle, Washington, USA
[9]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium
[10]Department of Medicine, University of California - San Diego, La Jolla, CA, USA
[11]Department of Mathematics, Tulane University, New Orleans, LA, USA
[12]Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California - Los Angeles, Los Angeles, CA, USA
[13]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California - Los Angeles, Los Angeles, CA, USA
*Correspondence to be sent to: Department of Biomathematics, University of California - Los Angeles, Los Angeles, CA 90095, USA;
msuchard@ucla.edu

Abstract.—Phylogenetic and discrete-trait evolutionary inference depend heavily on an appropriate characterization of the underlying character substitution process. In this paper, we present random-effects substitution models that extend common continuous-time Markov chain models into a richer class of processes capable of capturing a wider variety of substitution dynamics. As these random-effects substitution models often require many more parameters than their usual counterparts, inference can be both statistically and computationally challenging. Thus, we also propose an efficient approach to compute an approximation to the gradient of the data likelihood with respect to all unknown substitution model parameters. We demonstrate that this approximate gradient enables scaling of sampling-based inference, namely Bayesian inference via Hamiltonian Monte Carlo, under random-effects substitution models across large trees and state-spaces. Applied to a dataset of 583 SARS-CoV-2 sequences, an HKY model with random-effects shows strong signals of non-reversibility in the substitution process, and posterior predictive model checks clearly show that it is a more adequate model than a reversible model. When analyzing the pattern of phylogeographic spread of 1441 influenza A virus (H3N2) sequences between 14 regions, a random-effects phylogeographic substitution model infers that air travel volume adequately predicts almost all dispersal rates. A random-effects state-dependent substitution model reveals no evidence for an effect of arboreality on the swimming mode in the tree frog subfamily Hylinae. Simulations reveal that random-effects substitution models can accommodate both negligible and radical departures from the underlying base substitution model. We show that our gradient-based inference approach is over an order of magnitude more time efficient than conventional approaches. [Bayesian inference, Hamiltonian Monte Carlo, phylogeography.]

Along the branches of a phylogenetic tree, discrete characters such as nucleotides, amino acids, or morphologic traits evolve according to some (typically unknown) substitution process. Substitution models are probabilistic representations of the substitution process and are central quantities in phylogenetic and phylodynamic models. Broadly, substitution models describe the relative rates of discrete change from one character state to another.

When inferring phylogenies from character data, the nature of the substitution process is generally not the subject of primary biological interest. Nevertheless, because substitution models stand as the key link between the phylogenetic tree and the observed discrete character data, appropriate modeling remains of paramount importance to avoid bias and their specification has received considerable attention (see, e.g., Tavaré 1986; Suchard et al. 2001, 2003; Woodhams et al. 2015; Abadi et al. 2019; Fabreti and Höhna 2022). There are also cases in which the substitution process is itself of direct interest. In phylogeographic modeling of rapidly evolving pathogens, character states may represent the geographic locations of sampled pathogen sequences and the substitution process describes the spread of the pathogens through geographic space. In this case, inferring an appropriately parameterized substitution model

can deliver insight into the factors driving the spread of disease (Lemey et al. 2014, 2020; Dudas et al. 2017). Questions regarding potentially coevolving traits can also be addressed with substitution models by expanding their state-space, such as to pairs of binary characters (Pagel and Meade 2006).

Popular phylogenetic substitution models are continuous-time Markov chain (CTMC) models parameterized in terms of one or more infinitesimal rate matrices and branch lengths that measure the expected number of substitutions along each branch in the phylogeny. As the number of possible characters S in the data state-space grows, the number of potential parameters in each rate matrix $S \times (S-1)$ quickly becomes large. When inferring a phylogeny from nucleotide sequences, the rate matrix is small, and relatively parameter-rich models have been considered (Tavaré 1986; Yang 1994a). But rate matrices for other data types can easily grow large: there are 20 amino acids, 64 codons, and phylogeographic analyses can easily encompass many dozens of locations (Lemey et al. 2014; Dudas et al. 2017; Gao et al. 2022). Models which account for heterogeneity of the substitution process along branches, such as Markov-modulated models (Baele et al. 2021) may involve hundreds of parameters. In such cases, inferring the unconstrained model, in which all $\mathcal{O}(S^2)$ non-diagonal elements are free parameters, has been historically prohibitive. One reason is because the typical approach to Bayesian inference of substitution models is to use random-walk Metropolis–Hastings-based Markov chain Monte Carlo (MCMC) (Metropolis et al. 1953; Hastings 1970). Such large rate matrices have many parameters which are (potentially) strongly correlated and often only weakly identifiable, rendering random-walk MCMC burdensome.

When confronted with substitution models for large state-spaces, the historical approach has been to find ways to reduce the number of free parameters in the model. Amino acid models are often parameterized empirically (Dayhoff et al. 1978; Whelan and Goldman 2001), requiring no free parameters for inference. Codon models are often represented as combinations of site-level nucleotide models and codon-level processes (Yang et al. 2000), some of which may be measured empirically (Hilton and Bloom 2018). Such approaches reduce the number of parameters that must be inferred to $\mathcal{O}(S)$. Another approach is to parameterize the rate matrix in terms of log-linear functions of observed covariates. This generalized linear model (GLM) approach has been successful in phylogeographic inference, where observed covariates include factors like the distance between locations and air travel volumes (Lemey et al. 2014; Dudas et al. 2017). In addition to making inference tractable, the GLM approach can be used to quantify the strength of evidence for which factors do or do not affect the spread of infectious diseases.

In this paper, we demonstrate the utility of random-effects substitution models. These models extend a wide class of CTMC models to incorporate additional rate variation by representing the original (base) model as fixed-effect model parameters and allowing the additional random-effects to capture deviations from the simpler process. We demonstrate the utility of these models in Bayesian inference on a variety of exemplar evolutionary problems. On a dataset of 583 SARS-CoV-2 genomes, an HKY model with random-effects captures known mutational biases in SARS-CoV-2 and is shown to be superior to the richest, general-time reversible (GTR) model. Applied to a phylogeographic analysis of influenza A subtype H3N2, a GLM substitution model with random-effects provides evidence that air travel volume captures the geographic process of dispersal for all except a small set of pairs for which it underpredicts dispersal. As a test for ecologically dependent trait evolution, a random-effects pairwise-dependent substitution model finds no evidence for an effect of arboreality on the swimming mode in hylid tree frogs. We quantify the performance of random-effects substitution models using simulations.

Before we can perform inference under these models, however, there stand two obstacles that must be overcome. First and foremost, the parameter-space of the random effects models can be very large, and these parameters may be strongly correlated. To overcome the dimensionality, we derive an efficient-to-compute approximation of the gradient of the phylogenetic (log-)likelihood with respect to (wrt) all parameters of an arbitrary CTMC substitution model simultaneously. Notably, the exact gradient is often computationally prohibitive. We implement our approximate gradient in the phylogenetic inference software package BEAST 1.10 (Suchard et al. 2018) and the high-performance computational library BEAGLE 3 (Ayres et al. 2019), enabling the use of Hamiltonian Monte Carlo (HMC), a gradient-based alternative to random-walk MCMC (Neal 2011), for efficient parameter inference. HMC leverages gradients to take bold steps through even highly correlated parameter spaces and can greatly increase MCMC efficiency. Second, to avoid identifiability issues with potentially overparameterized models (such as inferring a 14×14 rate matrix based on a single observed character), we make use of the Bayesian bridge prior (Polson et al. 2014) that is strongly regularizing and allows the data to decide which parameters are important to capture their variability.

The rest of this paper is structured as follows. In the Methods section, we formally introduce the random-effects substitution model and the Bayesian bridge prior distribution. Then we derive our approximate gradient of the phylogenetic log-likelihood with respect to parameters of the substitution model. We also provide an introduction to gradient-based inference. In the Results section, we first investigate the increase in efficiency from using our approximate gradients compared to alternative approaches, both for optimization tasks and full Bayesian inference. Then we apply our

random-effects substitution model to a number of real-world examples and to simulated data. We conclude by contemplating future approaches for improving inference efficiency and additional application areas where random-effects substitution models are likely to be useful.

Methods

In this paper, we assume that there is a (possibly unknown) rooted phylogeny $\mathcal{F}$ with N tips that links the observed character sites and $N-1$ internal nodes. We index the branch lengths and nodes such that the edge connecting node v to its parent u has length t_v. Along each branch of the tree, we assume that characters arise from an alphabet of size S and evolve under a CTMC model with instantaneous rate matrix $\mathbf{Q} = \{\lambda_{ij}\}$, where $\lambda_{ij} \geq 0$ for $i \neq j$ and the diagonal elements are fixed such that row-sums of $\mathbf{Q}$ equal 0. We measure branch length t_v in expected number of substitutions per site according to a probability mass vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)$ over the characters. Often, $\boldsymbol{\pi}$ is taken as the stationary distribution of $\mathbf{Q}$, but this need not be the case. To account for variation in evolutionary rates across sites in the character alignment, finite mixture models (e.g., the discrete-gamma model of Yang 1994b) modulate the expected number of substitutions along all branches at a specific site. Consider that there are R rate categories, the rate scalar in the rth category is γ_r, and the prior probability of being in any particular mixture category is $\mathbb{P}(\gamma_r)$. Then, the finite-time transition probability matrix along branch v in category r is given by $\mathbf{P}_{vr} = \exp(\gamma_r \times \mathbf{Q} \times t_v)$, where we assume that $\mathbf{Q}$ is normalized wrt $\boldsymbol{\pi}$. The matrix $\mathbf{P}_{vr}$ governs the probability of change from state i to state j along branch v in category r. Note that the subscripts on $\mathbf{P}$ do not denote elements of the matrix but rather which of the $R \times (2N - 2)$ distinct transition probability matrices—one for each branch and rate category—is under consideration. In truth, the rate matrix $\mathbf{Q}$ is a function of a vector of estimable parameters $\boldsymbol{\theta}$, specifically $\mathbf{Q}(\boldsymbol{\theta})$, but we suppress this notation for ease of presentation.

Random-effects substitution models

Random-effects substitution models are extensions of simpler CTMC substitution models. We start with a *base model*, which could be as simple as Jukes-Cantor (Jukes and Cantor 1969), as complex as a GLM substitution model (Lemey et al. 2014) or anything in between. This base model carries a rate matrix $\mathbf{B} = \{b_{ij}\}$ and probability mass vector $\boldsymbol{\pi_B}$ over the characters. We define the random-effects substitution model rate matrix $\mathbf{Q}$ using the following log-linear formulation,

$$\log \lambda_{ij} = \log b_{ij} + \epsilon_{ij} \text{ for } i \neq j, \tag{1}$$

and set $\boldsymbol{\pi} = \boldsymbol{\pi_B}$. Intuitively, the random-effects ϵ_{ij} are multiplicative real-valued parameters which enable each non-diagonal element to deviate from the values specified by the base model. For example, $\epsilon_{ij} = \log(2)$ doubles the rate implied by the base model, $\lambda_{ij} = 2b_{ij}$.

Random-effects substitution models retain the basic structure of the base model that may be biologically or epidemiologically motivated, while allowing for potentially large deviations from this base process. In our phylogeographic example, we start with an epidemiologically motivated model for the spread of rapidly evolving pathogens parameterized using air travel volume between countries with a GLM substitution model (although we could potentially use many more covariates) and allow the random-effects to capture shortcomings of this description. For convenience, we shorthand the random-effects version of a substitution model $\mathcal{M}$ + RE, where $\mathcal{M}$ is the base model (e.g., Hasegawa, Kishino, and Yano model [HKY]; Hasegawa et al. 1985).

Bayesian regularization

Random-effects substitution models are in general overparameterized and as such not identifiable by the likelihood alone. In a Bayesian setting, a prior will provide relief from this and allow for a posterior to be inferred. Nevertheless, such circumstances demand careful thought about the choice of prior, as it will play a key role in determining the posterior. An attractive class of priors for these situations are shrinkage priors such as the Bayesian bridge (Polson et al. 2014) or the horseshoe (Carvalho et al. 2010). Originally developed for handling regression models when there are more parameters than observations, shrinkage priors regularize coefficients (often strongly) pulling them to be near or equal to 0 when the data provide little or no information to the contrary and, otherwise, attempting to impart little bias into the posterior. Shrinkage priors often (though not always) induce sparsity in the model, pulling *many* coefficients to be effectively 0. Shrinkage priors have also found success in phylogenetic contexts, including models for the rate of evolution of the rate of evolution (Fisher, Ji, Nishimura et al. 2021; Fisher, Ji, Zhang et al. 2021), population sizes over time (Faulkner et al. 2020), and rates of speciation and extinction (Magee et al. 2020). In these cases, the prior pull is toward no change, either from a branch to its descendants or from one time interval to the next. In the context of random-effects substitution models, shrinkage is imposed by pulling the random-effects toward the null value of 0 (at 0, there is no deviation from the base model in the $i \rightarrow j$ direction and $\lambda_{ij} = b_{ij}$).

Shrinkage priors also permit us to perform model selection, with the data and prior reconciling which parameters belong in or out of the model. For the random-effects substitution model, a particular random-effect ϵ_{ij} is excluded from the model if it is approximately 0. Where discrete mixture models, such as those used in Bayesian stochastic search variable selection (BSSVS,

Lemey et al. 2009), carry a finite probability that a parameter achieves exactly 0, shrinkage priors instead have a large spike of prior density near 0. While this occasionally makes it more difficult to declare if a parameter belongs in the model or not, the use of purely continuous priors usually yields Markov chains that mix more efficiently and, importantly, permits the use of gradient-based inference.

The Bayesian bridge prior on random-effect ϵ has density,

$$\Pr(\epsilon \mid \tau, \alpha) \propto \exp\left(-\left|\frac{\epsilon}{\tau}\right|^{\alpha}\right), \tag{2}$$

where the global scale τ controls the overall spread of the distribution and the exponent α controls the shape. The Bayesian bridge is perhaps best thought of as family of distributions, modulated by α, some of which fall into the class of sparsity-inducing priors ($\alpha < 1$) and some of which do not. At $\alpha = 2$, the density coincides with a standard normal distribution, while at $\alpha = 1$ it is the density of the Laplace distribution. At lower exponent values, the distribution becomes increasingly peaked around 0 and induces sparsity. We use $\alpha = 1/4$, which in practice imposes a useful level of sparsity without compromising MCMC convergence. The global scale τ also plays an important role in determining the degree of regularization. To both permit the data to inform the strength of regularization and efficient Gibbs sampling procedures, we place a Gamma(shape=1, scale=2) prior on $\tau^{-\alpha}$ (Nishimura and Suchard 2022).

The Bayesian bridge distribution has particularly fat tails for lower α. This can hamper sampling, and it can allow parameter values which are, a priori, unrealistically large (or small). Particularly, large random-effects can also cause numerical instability when exponentiating the substitution-rate matrix. These effects can be ameliorated by the use of the shrunken-shoulder Bayesian bridge (Nishimura and Suchard 2022). This formulation includes a "slab" parameter ζ that controls the tails of the distribution. Specifically, outside of $[-\zeta, \zeta]$, the tails of the shrunken-shoulder Bayesian bridge become Normal(0,ζ^2). We set $\zeta = 2$, which a priori specifies that it is unlikely for a particular element of the rate matrix to be more than $e^2 \approx 7$ times larger or smaller than specified by the base model.

The gradient of the phylogenetic log-likelihood

In this paper, we are interested in the gradient of the phylogenetic log-likelihood with respect to the parameters of the substitution model. The data $\mathbf{Y}$ are a collection of homologous sites (columns in a multiple sequence alignment), $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_C)$. We will write the likelihood $\mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta})$, and its gradient $\nabla \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta})$. The gradient is the collection of derivatives wrt all substitution model parameters,

$$\nabla \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_1} \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \theta_k} \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta})\right)', \tag{3}$$

where $'$ denotes the transpose operator.

Under the common assumption that sites evolve independently and identically, we can express the log-likelihood as a sum across all C sites, and hence derivatives of it as well. We have

$$\frac{\partial}{\partial \theta_k} \log \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta}) = \sum_{c=1}^{C} \frac{\partial}{\partial \theta_k} \log \left(\sum_{r=1}^{R} \mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r) \, \mathbb{P}(\gamma_r) \right)$$

$$= \sum_{c=1}^{C} \frac{\sum_{r=1}^{R} \frac{\partial}{\partial \theta_k} \mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r) \, \mathbb{P}(\gamma_r)}{\sum_{r=1}^{R} \mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r) \, \mathbb{P}(\gamma_r)} \tag{4}$$

The denominator is simply the likelihood of a site $\mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta})$. For simplicity, we will focus on the computation of $\frac{\partial}{\partial \theta_k} \mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r)$ for site c under rate category r.

Partial likelihood vectors and the phylogenetic likelihood We can, at any node v in the tree, compute the likelihood as

$$\mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}) = \sum_{r=1}^{R} [\mathbf{p}'_{vrc} \mathbf{q}_{vrc}] \, \mathbb{P}(\gamma_r). \tag{5}$$

The post-order partial likelihood vector $\mathbf{p}_{vrc}$ describes the probability, at node v, in rate category r, at the cth site, of observing the tip-states in all tips which descend from the node, conditioned on the state at the node. The pre-order partial likelihood vector, $\mathbf{q}_{vrc}$, describes the *joint* probability of observing the tip-states in all tips *not* descended from the node and the state at the node.

The post-order partial likelihood vectors are computed via pruning from the tip to the roots (a post-order traversal), for the tree in Fig. 1, via

$$\mathbf{p}_{urc} = \mathbf{P}_{vr}\mathbf{p}_{vrc} \circ \mathbf{P}_{wr}\mathbf{p}_{wrc}, \tag{6}$$
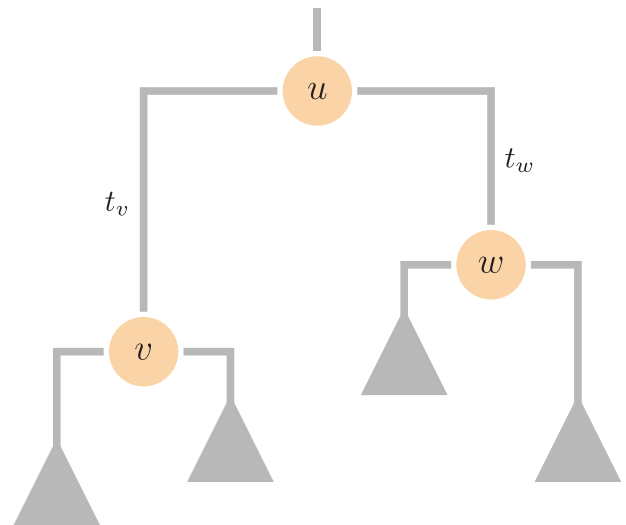

FIGURE 1. A phylogenetic tree highlighting three key nodes. We will take node v as our focal node, which here has parent u and sister w. We index branch lengths by the node which subtends them, such that the branch with length t_v is the branch leading to node v.

where ∘ denotes the Hadamard (element-wise) product. The pre-order partial likelihood vectors are then computed in a root-to-tip pass through the tree (a pre-order traversal) using the relation

$$\mathbf{q}_{vrc} = \mathbf{P}'_{vr}[\mathbf{q}_{urc} \circ \mathbf{P}_{wr}\mathbf{p}_{wrc}]. \tag{7}$$

We note that $\mathbf{p}_{vrc}$ is independent of $\mathbf{P}_{vr}$, while $\mathbf{q}_{vrc}$ is dependent on $\mathbf{P}_{vr}$. At the root $v = 2N - 1$, the pre-order partial likelihood vector is simply the root-frequency vector $\boldsymbol{\pi}_{\text{root}}$, which may or may not be the same as the frequency vector $\boldsymbol{\pi}$ used to normalize the rate matrix.

A naïve derivative We can use the multivariable chain rule to obtain the total derivative of the likelihood wrt θ_k. To do this, we first envision a simple parameter expansion with branch- and root-specific variables η_{vk}. By setting $\eta_{vk} = \theta_k$, we can frame the derivative wrt θ_k in terms of a sum across the all the per-branch contributions from the derivatives wrt η_{vk}. First we rewrite the differential as

$$\frac{\partial}{\partial\theta_k} = \sum_{v=1}^{2N-1} \left(\frac{\partial}{\partial\eta_{vk}}\right)\left(\frac{\partial\eta_{vk}}{\partial\theta_k}\right) = \sum_{v=1}^{2N-1} \frac{\partial}{\partial\eta_{vk}}. \tag{8}$$

Then, the partial likelihood vector representation of the phylogenetic likelihood allows us to isolate the contribution of each branch and $\boldsymbol{\pi}_{\text{root}}$ to this total derivative. In doing so, we also recall that $\mathbf{p}_{vrc}$ is independent of $\mathbf{P}_{vr}$ and $\boldsymbol{\pi}_{\text{root}}$, such that $(\partial/\partial\eta_{vk})\,\mathbf{p}_{vrc} = 0$ for all v. By summing over all branches and the root, we obtain the total derivative as

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r) &= \sum_{v=1}^{2N-1} \frac{\partial}{\partial\eta_{vk}}\mathbf{p}'_{vrc}\mathbf{q}_{vrc} \\
&= \sum_{v=1}^{2N-1} \mathbf{p}'_{vrc}\frac{\partial}{\partial\eta_{vk}}\mathbf{q}_{vrc} \\
&= \sum_{v=1}^{2N-2} \mathbf{p}'_{vrc}\left(\frac{\partial}{\partial\eta_{vk}}\mathbf{P}_{vr}\right)' \\
&\qquad [\mathbf{q}_u \circ \mathbf{P}_{wr}\mathbf{p}_w] + R(\theta_k) \\
&= \sum_{v=1}^{2N-2} \mathbf{p}'_{vrc}\left(\frac{\partial}{\partial\eta_{vk}}\mathbf{P}_{vr}\right)' \\
&\qquad \tilde{\mathbf{q}}_{vrc} + R(\theta_k) \\
&= \sum_{v=1}^{2N-2} \mathbf{p}'_{vrc}\left(\frac{\partial}{\partial\eta_{vk}}e^{\mathbf{Q}\times t_v\times\gamma_r}\right)' \\
&\qquad \tilde{\mathbf{q}}_{vrc} + R(\theta_k) \\
&= \sum_{v=1}^{2N-2} \mathbf{p}'_{vrc}\left(\sum_{i=1}^{S}\sum_{j=1}^{S}\frac{\partial e^{\mathbf{Q}\times t_v\times\gamma_r}}{\partial\lambda_{ij}}\frac{\partial\lambda_{ij}}{\partial\eta_{vk}}\right)' \\
&\qquad \tilde{\mathbf{q}}_{vrc} + R(\theta_k),
\end{aligned} \tag{9}$$

where the contribution from root-frequency vector is

$$R(\theta_k) = \mathbf{p}'_{2N-1,rc}\frac{\partial}{\partial\theta_k}\boldsymbol{\pi}_{\text{root}}. \tag{10}$$

In the third-to-last step, we defined $\tilde{\mathbf{q}}_{vrc} = [\mathbf{q}_u \circ \mathbf{P}^{(w)}\mathbf{p}_w]$ to simplify the notation and focus on the part of the equation which depends on η_{vk} (θ_k). In the last step, we employed the matrix chain rule (Petersen et al. 2008). The term $\partial \exp(\mathbf{Q} \times t_v \times \gamma_r)/\partial\lambda_{ij}$ is the derivative of the matrix exponential with respect to one of the elements of the rate matrix, which we discuss in more detail in the Section "Efficiently approximating the derivative of the matrix exponential." We note that the rate matrix $\mathbf{Q}$ is sometimes defined to be unnormalized, such that the transition probability matrix along a branch is instead given by $\mathbf{P}_{vr} = \exp(\gamma_r \times (-1/(\text{diag}(\mathbf{Q})'\boldsymbol{\pi})) \times \mathbf{Q} \times t_v)$. For simplicity of notation, when normalization is desired we take the rate matrix $\mathbf{Q}$ to be normalized, and allow the $\partial\lambda_{ij}/\partial\eta_{vk}$ term to capture the effect of the normalizing constant on the elements of the rate matrix. We discuss the computation of the gradient of the rate matrix with respect to its parameters and the root contribution in Supplementary Material, "More details on the form of the gradient" section.

As we discuss in Supplementary Material, "Computational complexity of alternative approaches to computing the gradient of a matrix exponential" section, the computational cost of obtaining $\partial \exp(\mathbf{Q}\times t_v\times\gamma_r)/\partial\lambda_{ij}$ is $\mathcal{O}(S^3)$. The sum in Equation 9 requires this quantity for all S^2 elements in $\mathbf{Q}$ and for each of the $2N-2$ branches, making the cost to compute the derivative $\mathcal{O}(NS^5)$. Obtaining the gradient requires using Equation 9 for all K substitution model parameters, making the cost of the gradient $\mathcal{O}(KNS^5)$. For random effects models, this is $\mathcal{O}(NS^7)$. Such costs are prohibitive for even moderate S, so we turn our attention now to improving the computational efficiency of gradient computations.

Reducing the computational complexity We can reformulate the naïve approach of Equation 9 to produce a more efficient gradient computation. By rearranging the order of summation, we can disentangle the derivative of the rate matrix wrt its elements from the derivative of its elements wrt model parameters. Specifically,

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r) &= \sum_{v=1}^{2N-2} \mathbf{p}'_{vrc}\left(\sum_{i=1}^{S}\sum_{j=1}^{S}\frac{\partial e^{\mathbf{Q}\times t_v\times\gamma_r}}{\partial\lambda_{ij}}\frac{\partial\lambda_{ij}}{\partial\eta_{vk}}\right)' \\
&\qquad \tilde{\mathbf{q}}_{vrc} + R(\theta_k) \\
&= \sum_{v=1}^{2N-2}\sum_{i=1}^{S}\sum_{j=1}^{S} \mathbf{p}'_{vrc}\left(\frac{\partial e^{\mathbf{Q}\times t_v\times\gamma_r}}{\partial\lambda_{ij}}\right)' \\
&\qquad \tilde{\mathbf{q}}_{vrc}\frac{\partial\lambda_{ij}}{\partial\eta_{vk}} + R(\theta_k) \\
&= \sum_{v=1}^{2N-2} \mathbf{M}_k \,\text{vec}(\mathbf{D}_v) + R(\theta_k),
\end{aligned} \tag{11}$$

where the vec operator makes the matrix $\mathbf{D}_v$ into a column vector by stacking the columns on top of each other, and we obtain the last line by defining two new

quantities which we will now discuss. The matrix $\mathbf{M}$ is a $K \times S^2$ mapping matrix, which stores in each row k a vector of the partial derivatives of all elements of $\mathbf{Q}$ wrt θ_k,

$$\mathbf{M}_{k\cdot} = \left(\frac{\partial \lambda_{11}}{\partial \theta_k}, \ldots, \frac{\partial \lambda_{SS}}{\partial \theta_k} \right). \qquad (12)$$

The matrix $\mathbf{D}_{vr} = \{d_{vrij}\}$ contains the contribution of branch v to the derivative of the phylogenetic likelihood wrt the ij^{th} entry of $\mathbf{Q}$ in rate category r. Specifically,

$$d_{vrij} = \mathbf{p}'_{vrc} \left(\frac{\partial}{\partial \lambda_{ij}} e^{\mathbf{Q} \times t_v \times \gamma_r} \right)' \tilde{\mathbf{q}}_{vrc}. \qquad (13)$$

We arrive at the entire gradient (as opposed to a single entry) and increase computational efficiency by replacing $\mathbf{M}_{k\cdot}$ with $\mathbf{M}$ in Equation 11 and rearranging,

$$\begin{aligned} \nabla \mathbb{P}(\mathbf{Y}_c \mid \boldsymbol{\theta}, r) &= \sum_v \mathbf{M} \operatorname{vec}(\mathbf{D}_v) + R(\theta_k) \\ &= \mathbf{M} \sum_v \operatorname{vec}(\mathbf{D}_v) + R(\theta_k). \qquad (14) \end{aligned}$$

This approach separates the gradient of the phylogenetic likelihood wrt model parameters into two pieces, a gradient of the phylogenetic likelihood wrt elements of the rate matrix, and a gradient of the elements of the rate matrix wrt the model parameters. The result is the intermediate quantity $\mathbf{D}_v$ that can be obtained with only a single computation of the derivative of a matrix exponential per branch. As this quantity can be summed across the tree prior to mapping it to the substitution model parameters, $\mathcal{O}(N)$ matrix multiplications are avoided. The result is that this approach is $\mathcal{O}(KS^2 + NS^5)$ rather than $\mathcal{O}(KNS^5)$. For random-effects substitution models, this is the difference between a $\mathcal{O}(NS^7)$ computation and a $\mathcal{O}(S^4 + NS^5) = \mathcal{O}(NS^5)$ computation. Note that this approach works for branch-specific models as well, by specifying the mapping matrix $\mathbf{M}$ appropriately.

Efficiently approximating the derivative of the matrix exponential We now turn our attention to an efficient approximation to the derivative of a matrix exponential. The derivative of a matrix exponential can be represented as a power-series (Najfeld and Havel 1995, equation 103),

$$\frac{\partial}{\partial \lambda_{ij}} e^{\mathbf{Q}t} = e^{\mathbf{Q}t} \sum_{x=0}^{\infty} \frac{t^{x+1}}{(x+1)!} \{\mathbf{E}_{ij}, \mathbf{Q}^x\}, \qquad (15)$$

where $\mathbf{E}_{ij}$ is a matrix which is 0 for all but the (ij)th entry, which is 1. The matrix commutator power $\{\mathbf{E}_{ij}, \mathbf{Q}^x\}$ for non-negative integer x is defined recursively (Najfeld and Havel 1995), such that $\{\mathbf{E}_{ij}, \mathbf{Q}^0\} = \mathbf{E}_{ij}$ and $\{\mathbf{E}_{ij}, \mathbf{Q}^x\} = [\{\mathbf{E}_{ij}, \mathbf{Q}^{x-1}\}, \mathbf{Q}]$, where $[\mathbf{A}, \mathbf{B}]$ is the matrix commutator $\mathbf{AB} - \mathbf{BA}$.

The first-order approximation to Equation 15 is taken by keeping only the $x = 0$ term, yielding

$$\frac{\partial}{\partial \lambda_{ij}} e^{\mathbf{Q}t} \approx t e^{\mathbf{Q}t} \mathbf{E}_{ij}. \qquad (16)$$

We can use this first-order approximation to approximate $\mathbf{D}_v$ on each branch. Specifically,

$$\begin{aligned} d_{vij} &\approx \mathbf{p}'_{vrc}[(t_v \times \gamma_r \times e^{\mathbf{Q} \times t_v \times \gamma_r} \mathbf{E}_{ij})'[\mathbf{q}_u \circ (e^{\mathbf{Q}t_w \times \gamma_r} \mathbf{p}_w)]] \\ &= t_v \times \gamma_r \times \mathbf{p}'_{vrc}[\mathbf{E}_{ji}(e^{\mathbf{Q} \times t_v \times \gamma_r})'[\mathbf{q}_u \circ (e^{\mathbf{Q} \times t_v \times \gamma_r} \mathbf{p}_w)]] \\ &= t_v \times \gamma_r \times \mathbf{p}'_{vrc} \mathbf{E}_{ji} \mathbf{q}_{vrc} \\ &= t_v \times \gamma_r \times q_{vi} p_{vj}, \qquad (17) \end{aligned}$$

where we get from line 3 to line 4 by noting that $(e^{\mathbf{Q} \times t_v \times \gamma_r})' = \mathbf{P}'_{vr}$ and applying Equation 7. Intuitively, we have the (approximate) derivative with respect to an $i \to j$ transition depending on the pre-order partial likelihood in state i and the post-order partial likelihood in state j.

Equation 17 means that we can write our approximate $\mathbf{D}_v$ as an outer product,

$$\mathbf{D}_{vrc} \approx t_v \times \gamma_r \times \mathbf{q}_{vrc} \otimes \mathbf{p}_{vrc}. \qquad (18)$$

This means that we can obtain all S^2 entries of $\mathbf{D}_{vrc}$ in $\mathcal{O}(S^2)$, which is much more efficient than the $\mathcal{O}(S^5)$ cost of the non-approximate computation. Thus, with this approximation and the mapping approach outlined in the previous section, the (approximate) substitution gradient can be obtained in $\mathcal{O}(KS^2 + NS^3)$, rather than the $\mathcal{O}(KS^2 + NS^5)$ cost suggested by Equation 14 or the $\mathcal{O}(KNS^5)$ cost suggested by Equation 9. We will denote the approximate gradient that comes from using this approximation to $\mathbf{D}_{vrc}$ in Equation 14 as $\widetilde{\nabla} \log \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta})$.

Hamiltonian Monte Carlo with surrogate trajectories

HMC (Duane et al. 1987; Neal 2011) is an advanced MCMC algorithm that falls broadly within the well-known class of Metropolis–Hastings MCMC (MH-MCMC) algorithms (Metropolis et al. 1953; Hastings 1970). By allowing samples to be drawn (sequentially) from arbitrary target distributions, MH-MCMC algorithms like HMC allow users to approximate distributions that do not have known closed-form densities. Unlike many commonly employed random-walk Metropolis–Hastings proposals, however, HMC uses information captured by the log-posterior gradient to traverse a model's parameter space much more efficiently by proposing (and accepting) states which are farther apart.

HMC constructs an artificial Hamiltonian system by augmenting the parameter space to include an auxiliary Gaussian "momentum" variable $\boldsymbol{\xi} \sim \text{MVN}_K(0, \mathbf{M})$ that is independent from the target variable $\boldsymbol{\theta}$ by construction. Letting $f(\cdot)$ denote the posterior density, the

resulting Hamiltonian energy function is the negative logarithm of the joint distribution over $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. Ignoring normalizing constants we obtain

$$H(\boldsymbol{\theta}, \boldsymbol{\zeta}) = -\log f(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{M}^{-1}\boldsymbol{\zeta},$$

and Hamilton's equations are

$$\dot{\boldsymbol{\theta}} = +\frac{\partial H}{\partial \boldsymbol{\zeta}} = \mathbf{M}^{-1}\boldsymbol{\zeta}$$

$$\dot{\boldsymbol{\zeta}} = -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla \log f(\boldsymbol{\theta}).$$

On the one hand, one may show that the action of the dynamical system that satisfies these equations leaves the target $f(\cdot)$ invariant thanks to the reversibility, volume preservation and energy conservation of Hamiltonian dynamics. On the other hand, closed-form descriptions of these dynamics are rarely available for arbitrary target distributions, leading to the need for computer intensive approximations. In particular, the Störmer–Verlet (velocity Verlet) or leapfrog method (Leimkuhler and Reich 2004) has become the numerical integrator of choice for obtaining discretized trajectories within HMC. Beginning at time τ and letting $\epsilon > 0$ be small, a single leapfrog iteration proceeds thus:

$$\boldsymbol{\zeta}\left(\tau + \frac{\epsilon}{2}\right) = \boldsymbol{\zeta}(\tau) + \frac{\epsilon}{2}\nabla \log f(\boldsymbol{\theta}(\tau)) \qquad (19)$$

$$\boldsymbol{\theta}(\tau + \epsilon) = \boldsymbol{\theta}(\tau) + \epsilon\,\mathbf{M}^{-1}\boldsymbol{\zeta}\left(\tau + \frac{\epsilon}{2}\right)$$

$$\boldsymbol{\zeta}(\tau + \epsilon) = \boldsymbol{\zeta}\left(\tau + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\nabla \log f(\boldsymbol{\theta}(\tau + \epsilon)).$$

Trajectories arising from concatenated leapfrog iterations maintain some of the desirable qualities of the exact Hamiltonian dynamics (reversibility, volume preservation) but no longer conserve energy.

A single iteration of the HMC algorithm features three distinct steps. It starts at the current parameter values $\boldsymbol{\theta}_t$, setting $\boldsymbol{\theta}(\tau) = \boldsymbol{\theta}_t$, and draws an initial momentum $\boldsymbol{\zeta}_t \sim \mathrm{MVN}_K(0, \mathbf{M})$ from the given Gaussian distribution, setting $\boldsymbol{\zeta}(\tau) = \boldsymbol{\zeta}_t$. Then, it uses L steps of the leapfrog integrator for some given step-size ϵ, repeatedly applying Equation 19 to generate a proposed parameter state $\boldsymbol{\theta}^\star = \boldsymbol{\theta}(L\epsilon)$, as well as a final momentum $\boldsymbol{\zeta}^\star = \boldsymbol{\zeta}(L\epsilon)$. Lastly, it either accepts or rejects this new value according to the usual Metropolis–Hastings acceptance rule (Metropolis et al. 1953; Hastings 1970) with acceptance probability $\min\left[1, f(\boldsymbol{\theta}^\star)/f(\boldsymbol{\theta}_t) \times \mathrm{MVN}_K(-\boldsymbol{\zeta}^\star; 0, \mathbf{M})/\mathrm{MVN}_K(\boldsymbol{\zeta}_t; 0, \mathbf{M}))\right]$, where $\mathrm{MVN}_K(x; 0, \mathbf{M})$ indicates the density of the Gaussian momentum distribution evaluated at x. (The final momentum is negated to ensure the proposal is a symmetric (Neal 2011; Vats 2023).) The accept/reject step accounts for integration error and leaves the target distribution invariant.

Indeed, HMC's Metropolis correction allows for additional deviations from Hamiltonian dynamics over

and beyond numerical discretization schemes such as Equation 19. *Surrogate HMC* methods seek to improve computational performance of HMC by approximating the log-posterior gradient with less expensive surrogate models including, e.g., piecewise-approximations (Zhang et al. 2017b), Gaussian processes (Rasmussen 2003; Lan et al. 2016) or neural networks (Zhang et al. 2017a; Li et al. 2019). Directly relevant to the present work, Li et al. (2019) show the validity of replacing the log-posterior gradient $\nabla \log f(\cdot)$ within the leapfrog method (Equation 19) with *any* vector function $\mathbf{g} : \mathbb{R}^K \to \mathbb{R}^K$. In particular, such an approach maintains the reversibility and volume preservation of Hamiltonian dynamics and, when paired with Metropolis corrections, leaves the target posterior distribution invariant. In the present work, we select $\mathbf{g} = \widetilde{\nabla} \log \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\theta}) + \nabla \log \mathbb{P}(\boldsymbol{\theta})$, the approximate posterior gradient obtained by using our approximation to the gradient of the phylogenetic log-likelihood and the true gradient for the prior. In the supplementary Materials, we discuss an alternative justification.

RESULTS

C to T bias in SARS-CoV-2 evolution

The mutational profile of SARS-CoV-2 has been intensely scrutinized, one feature in particular which has been noted is a strongly increased rate of C→T substitutions over the reverse T→C substitutions. We note that while RNA viruses like SARS-CoV-2 use uracil (U) in place of thymine (T), it is generally coded as thymine–the coding of adenosine (A), cytosine (C), and guanine (G) are unchanged. The elevation of one direction of substitution over its reverse is a violation of the common phylogenetic assumption of reversibility made by the GTR (Tavaré 1986) family of substitution models. Random-effects substitution models are suitable for addressing this model violation, in particular we consider an HKY+RE substitution model. In principle, we could choose any GTR-family model. HKY represents a balance between the simplicity of JC+RE (where the random-effects would also have to account for uneven nucleotide frequencies) and the complexity of GTR+RE (where the random-effects only capture nonreversibilities). The rate matrix is

$$\log \lambda_{ij} = \log \kappa \times \mathbb{I}(ij \in \mathcal{T}) + \log \pi_j + \epsilon_{ij}, \qquad (20)$$

where κ is the HKY parameter governing relative rate of transitions to transversions, $\mathbb{I}(ij \in \mathcal{T})$ indicates that the i to j change is a transition, and $\boldsymbol{\pi}$ are the HKY stationary frequencies.

We apply this HKY+RE model to infer both the dynamics of molecular substitution and the phylogeny for 583 SARS-CoV-2 sequences from Pekar et al. (2021). (More information about the model and dataset is in Supplementary Table S1.) Consistent with previous
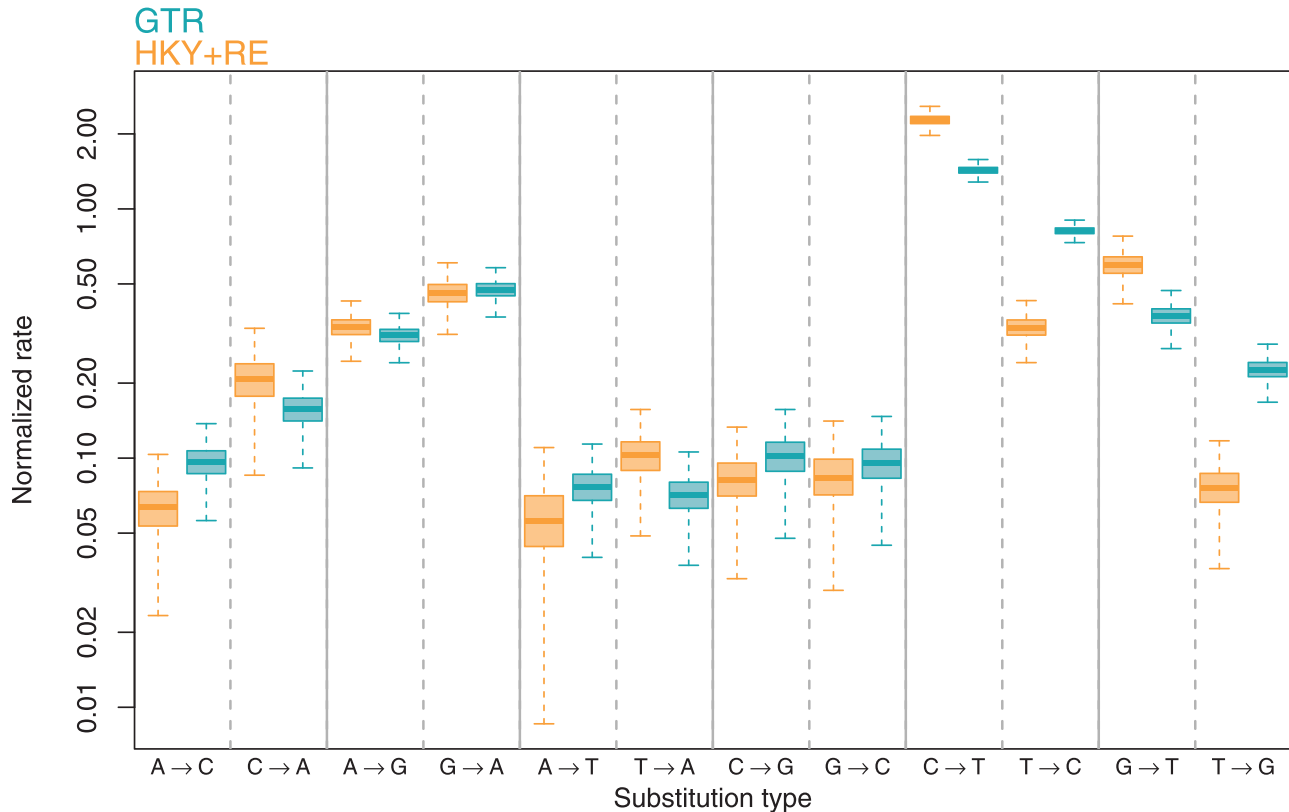
FIGURE 2. Posterior distributions of the 12 non-diagonal elements of the inferred rate matrices for the dataset of Pekar et al. (2021). The solid line is the posterior median, the shaded region the 50% CI. The whiskers extend to the posterior samples farthest from the median but within 1.5× the interquartile range. By comparing HKY+RE (which is not constrained by the assumption of reversibility) to GTR, we can see that the assumption of reversibility leads to the overestimation of the T→C (and T→G) rates and the underestimation of the C→T (and G→T) rates.

studies (*e.g.* Matyášek and Kovařík 2020; Tonkin-Hill et al. 2021), we find evidence for a greatly elevated rate of C→T substitutions, as well as an elevated G→T rate (Figure 2). We can test the support for nonreversibilities, for example the difference between the C→T and T→C rates, with Bayes factors. The fact that a model with the C→T and T→C rates equal (reversible wrt C↔T) is nested within the random-effects model allows us to use the Savage-Dickey ratio (*e.g.* Wagenmakers et al. 2010) to compute the Bayes factor from the posterior distribution of the random-effects model, as we discuss in the Supplemental Section "Assessing the strength of evidence for nonreversibilities." (There is no need to fit any additional models or estimate marginal likelihoods directly, though for models like these where analytical posterior distributions are unavailable, the results are still not entirely free from the numerical difficulties common in Bayes factor estimation.) The Bayes factor provides "very strong" (Kass and Raftery 1995) support for the nonreversibility of C→T and G→T rates (over the reversible model). We can also assess the strength of evidence for nonreversibilities via the posterior sign probability. This is the posterior probability that the sign of a variable is the same as the sign of the posterior

median (this is one minus the tail probability used by Zhang et al. 2021). The estimated sign probability ranges from 0.5 to 1.0, with larger values indicating increasingly strong support that the parameter is non-zero. Here, as with Bayes factors, we are interested in the sign probabilities of the differences in random-effects rather than the random-effects directly. The sign probabilities agree with the Bayes factors that there is strong evidence for the nonreversibility of C→T and G→T rates, with both estimated sign probabilities above 0.99.

Given the strong evidence for nonreversibilities, we sought to investigate the issue of the adequacy of reversible models (namely GTR) using posterior predictive model checks. In a posterior predictive framework, a summary of the observed dataset is compared with the distribution of summaries of datasets produced by drawing from the posterior distribution on model parameters. Broadly, if the model fits the data well, we expect that the predicted summaries will match the observed values, while if the fit is poor there will be a mismatch. As our test statistics, we consider all pairwise covariances of the proportion of each nucleotide (A, C, G, and T) across the alignment (we discuss this in more detail in Supplementary Material, "Posterior predictive
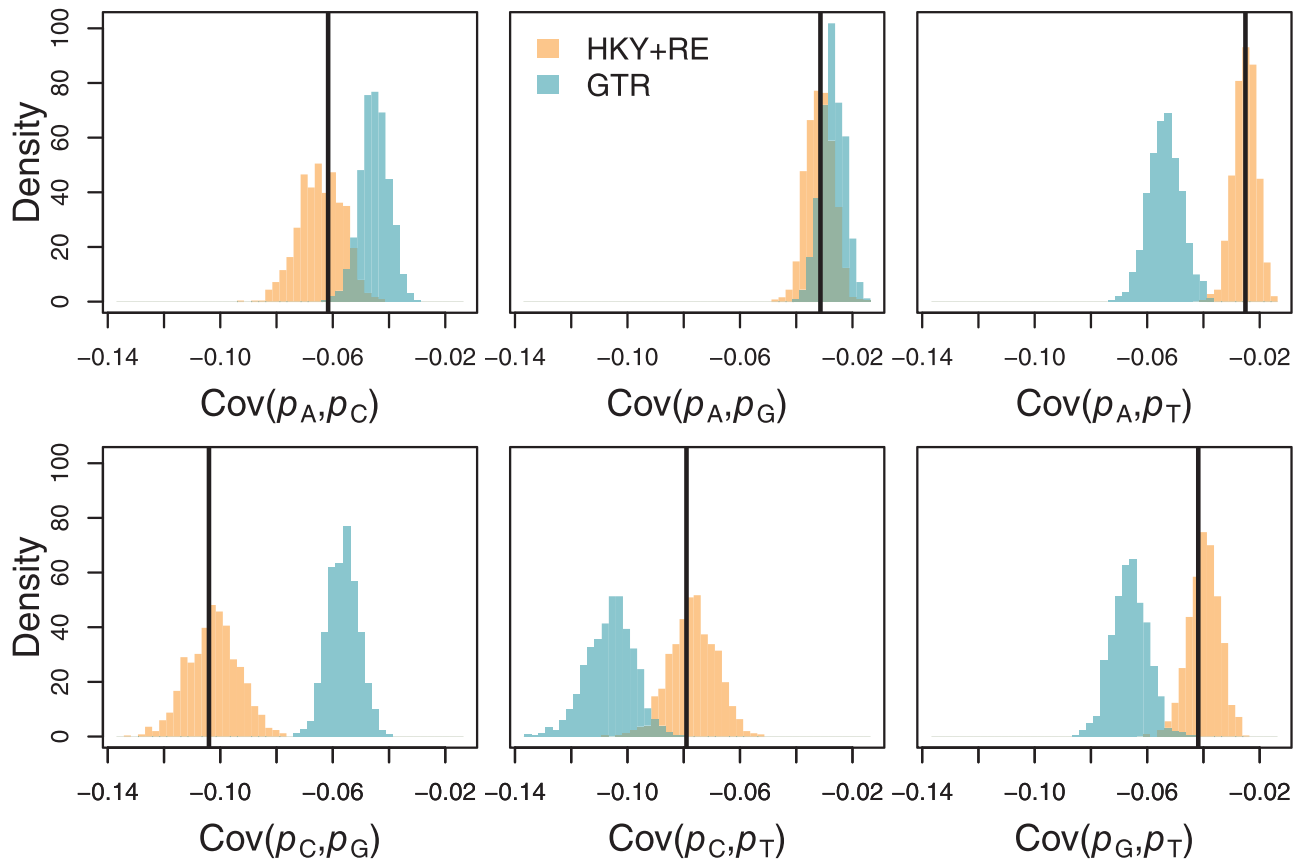
FIGURE 3. Posterior predictive distributions of the covariances of the proportions of each nucleotide (denoted p_A, p_C, p_G, and p_T) across sites in the alignment (histograms) compared with the true values (vertical black lines). The HKY+RE predictive distributions all closely align with the observed values while all but one of the GTR predictive distributions are discordant.

p-values for proportions" section). These test statistics clearly demonstrate that the HKY+RE model better captures the evolutionary processes at hand (Fig. 3). Compared with inference using GTR, the analysis with HKY+RE produces notably higher support for the rootmost divergence (the 95% credible set includes 67 possible resolutions for GTR and 1 for HKY+RE, Supplementary Fig. S8) and infers a root time approximately 5 days earlier.

Phylogeography of influenza from 2002 to 2007

For a larger state-space example of random-effects substitution models, we consider the global spread of human influenza A virus (subtype H3N2) from 2002 to 2007. Lemey et al. (2014) examined the movement patterns between 14 distinct air travel communities using 1529 viral genomes. The authors used a GLM to parameterize the spread of the virus between these communities as a function of a number of covariates, and discovered that the most consistently supported predictor of spread between communities was the volume of air traffic.

We re-analyze this dataset using a GLM substitution model with random-effects. We now briefly review the setup of a GLM substitution model, and our random-effects extension. For each pair of locations i and j, let $\mathbf{X}_{ij}$ be a vector of P predictors of the rate of movement from i to j (these may depend on the source i, the destination j, or both) with associated coefficients $\boldsymbol{\beta}$. A GLM substitution model with random-effects defines the rate matrix through

$$\log \lambda_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + \epsilon_{ij}. \tag{21}$$

This is a log-linear model, in which the GLM defines a substitution rate based on predictors and the estimated coefficients, and the random-effects allow for deviations from the model's predictions.

In particular, we employ a simple GLM with only air traffic included as a predictor. This approach allows us to determine how well air traffic volume predicts the spread of influenza A virus in the mid-2000s. If most random-effects are negligible, then air traffic volume alone is perhaps adequate for modeling the spread of influenza in this time frame. On the other hand, if many or most random-effects are not negligible, although air traffic volume may be an important model component,

it is not sufficient to explain spread, absent random-effects. While Lemey et al. (2014) used spike-and-slab priors on β in a Bayesian model averaging approach, since we are using only predictors identified previously to be important, we use a Normal prior instead (corresponding to the slab in the original study). We apply Bayesian bridge priors for the random-effects. To account for phylogenetic uncertainty, we marginalize our inference over the same empirical distribution of phylogenetic trees used by Lemey et al. (2014).

We find that air traffic volume sufficiently explains the viral spread between most communities. That is, for most community pairs, the posterior distribution of the random-effect indicates that the parameter has been declared "insignificant," and is a spike centered at 0 (Fig. 4). However, for five pairs of communities (from the United States to Japan and South America; from China to the United States and Japan; from Oceania to

the United States), the inferred random-effect is clearly significant (all sign probabilities > 0.99) and strongly positive, indicating 6- to 12-fold higher dispersal than predicted by travel. There is support for an additional six random-effects (from the United States to Oceania, Russia, and Southeast Asia; from China to Oceania; from Japan to Oceania; from Southeast Asia to Oceania) which have sign probabilities between 0.87 and 0.97 and correspond to 2- to 5-fold higher dispersal than predicted by travel. All other area pairs of sign probabilities are less than 0.78. Given the offset seasons between hemispheres, some of these connections likely do not represent biologically meaningful connections, and may potentially be attributed to sampling biases. A comparison of the number of samples in the dataset to the population sizes of the regions (a rough proxy for the number of infections in the regions) reveals that the United States, Oceania, and Japan are
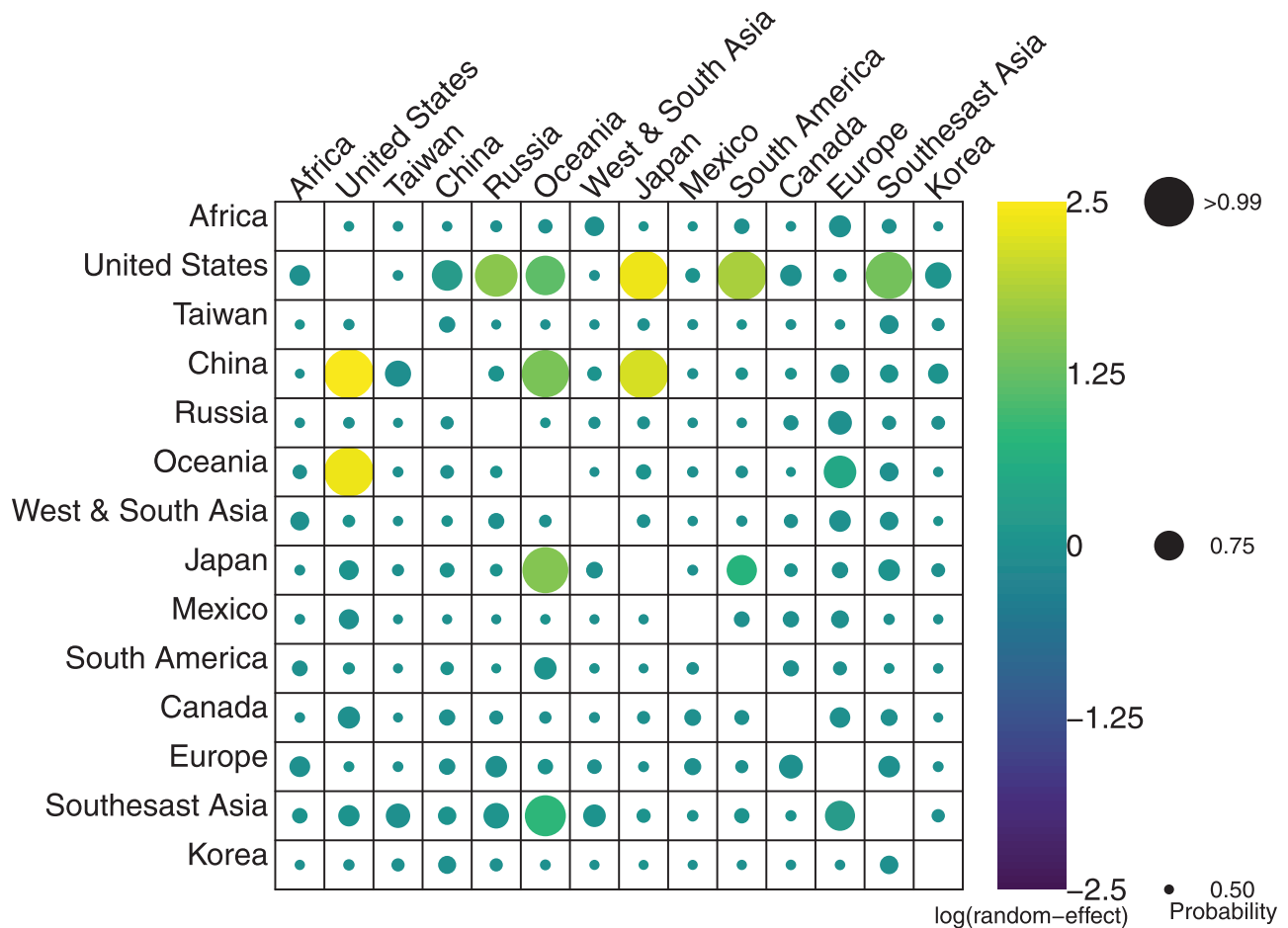

FIGURE 4. Summary of all 182 random-effects for the influenza A virus (subtype H3N2), shown in the format of the rate matrix, with the source in rows and destination in columns. The circle in each square is colored by the posterior median random-effect. The size of the circle denotes how strong the posterior support is that a random-effect is in the model. Specifically, the radius corresponds to the posterior sign probability. When the prior dominates the posterior distribution, a random-effect gains a larger posterior mass at 0 and becomes increasingly symmetric, the median approaches 0, and the posterior sign probability approaches 0.5. When the data are strongly informative, the posterior distribution moves away from 0 and the posterior sign probability gets larger. The random-effects which are most strongly supported are all positive, indicating that air travel underpredicts dispersal for those pairs of locations.

strongly oversampled. Thus, sampling biases likely explains many of the significant random-effects, including the between-hemisphere connections. As China is not particularly oversampled, the elevated rates of transmission from China may represent source-sink dynamics which are not captured by air travel alone, rather than sampling bias.

Analysis of paired macroevolutionary traits

Random-effects substitution models can also be used to test for dependent substitution processes between multiple characters as follows. Let us assume that we have two characters of interest, $\mathbf{Y}_1$ and $\mathbf{Y}_2$. These characters could be morphological, behavioral, or even ecological traits. If these characters evolved independently along the phylogeny $\mathcal{F}$, we could model this with two rate matrices, $\mathbf{Q}^1$ and $\mathbf{Q}^2$, a (strict) clock rate which defines the rate of change (in substitutions per year or million years) for $\mathbf{Y}_2$, and a relative rate parameter μ which defines how much faster (or slower) $\mathbf{Y}_1$ evolves compared with $\mathbf{Y}_2$. We can define a composite character from $\mathbf{Y}_1$ and $\mathbf{Y}_2$ by considering both states simultaneously. This yields a new character $\mathbf{Y}$ which is the Cartesian product of the two state-spaces, with the combined state-space size $S = S_1 \times S_2$. The rate matrix $\mathbf{Q}$ for the combined character is 0 for any double substitution and for any single substitution is defined by $\mathbf{Q}^1$ or $\mathbf{Q}^2$ depending on which character changes. Written on the log-scale, the (unnormalized) rate matrix is given by

$$\log \lambda_{ij} = \begin{cases} \log \lambda_{ij}^1 & \text{if } i \to j \text{ substitution in } Y_1 \\ \log \mu + \log \lambda_{ij}^2 & \text{if } i \to j \text{ substitution in } Y_2 \\ -\infty & \text{for all double substitutions.} \end{cases}$$

We can test for departures from independent evolution by allowing the state of one character to modulate the rates of change between states in the other through the addition of random-effects.

We employ this random-effects dependent morphological evolution model on a dataset of 29 species of frogs in the family Hylidae (subfamily Hylinae). Taking the phylogeny inferred by Caviedes-Solis et al. (2020) to be fixed, we focus on two traits described in Caviedes-Solis (2019), one ecological and one behavioral. The ecological trait is the habitat, which is characterized as either arboreal or understory. The behavioral trait is the swimming mode, which is characterized by whether the back legs move in an alternating or simultaneous fashion or whether both types are observed. To determine the structure of the underlying independent-trait models, we first fit the independent model using asymmetric rates for both traits. Bayes factors show no evidence for any model more complex than the Mk (Jukes-Cantor-like) model (Lewis 2001).

In particular, we are interested in whether the degree of arboreality, defined as habitat preference, impacts the swimming mode, as canopy-dwelling species move the back legs in an alternating fashion while climbing. Thus, we place random-effects only in the direction of arboreality affecting swimming mode. Letting $\mathbf{Y}_1$ be arboreality and $\mathbf{Y}_2$ be swimming mode, the unnormalized rate matrix for our random-effects substitution model is,

$$\log \lambda_{ij} = \begin{cases} 0 & \text{if } i \to j \text{ substitution in } Y_1 \\ \log \mu + \epsilon_{ij} & \text{if } i \to j \text{ substitution in } Y_2 \text{ and} \\ & Y_1 \text{ indicates canopy-dwelling} \\ \log \mu & \text{if } i \to j \text{ substitution in } Y_2 \text{ and} \\ & Y_1 \text{ indicates understory-dwelling} \\ -\infty & \text{for all double substitutions.} \end{cases}$$

We infer no effect of arboreal habitat on the swimming mode, all posterior sign probabilities are between 0.5 and 0.57, indicating that all random-effects have clearly been deemed insignificant (Fig. 5). We also infer that the rate of habitat evolution is roughly twice that of swimming-mode evolution ($\mu \approx 0.5$, 95% CI 0.23–1.16). There are two important caveats to these results. First, with only 29 species, the power to detect dependent evolution is likely low unless the effect is quite large. Secondly, by only modeling two traits, we are missing out on possible interactions between other aspects of ecology (such as the aquatic environments the species make use of) and morphology (such as the lengths of limbs and digits) which might modulate this relationship.

Performance gains from gradients

For inferring random-effects in nucleotide substitution models, we find a notable improvement in efficiency using HMC with our approximate gradients over using standard random-walk MH-MCMC. For our measure of efficiency, we consider the number of effectively independent samples taken per second (ESS/s). This measure incorporates both the increased ability of HMC to move through parameter space, as well as the increased cost per MCMC move required for repeated evaluation of the gradient. We track the efficiency separately for each random-effect (that is, we use the univariate ESS), and we consider two summaries of efficiency gains from HMC. As an overall measure of efficiency increase, we consider the parameterwise average increase in the efficiency. However, as analyses are constrained by waiting for the slowest-mixing parameter to achieve a sufficiently large ESS, we also consider the improvement in the minimum ESS (regardless of which parameter is slowest-mixing). When applied to nucleotide models (HKY+RE) to infer the tree from sequence data, we observe an average parameterwise increase in efficiency of 6.6-fold, and an increase in the minimum efficiency of 14.8-fold (Fig. 6). For the larger state-space of the flu phylogeographic example (14 discrete areas), where we average across a set of posterior samples of the tree from the original study, we find an average parameterwise increase in efficiency of 20.2-fold and an increase in minimum efficiency of 33.6-fold (Fig. 6). Timing was done
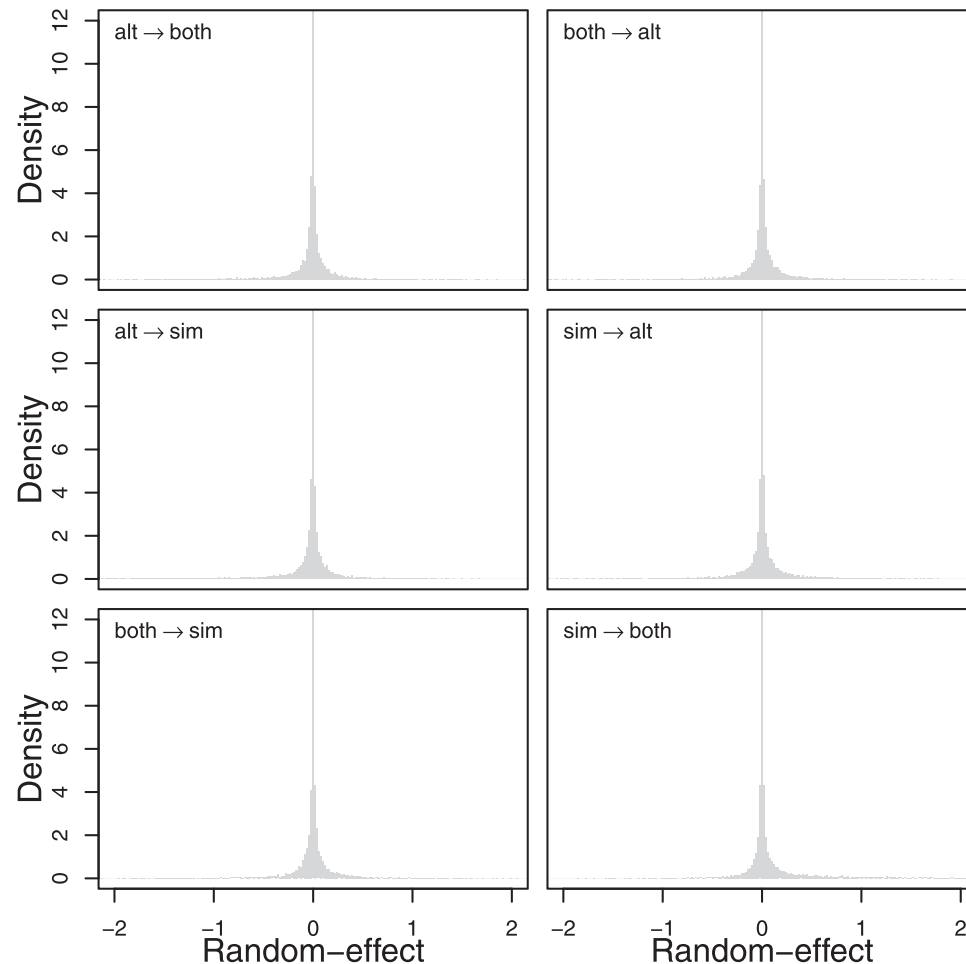
FIGURE 5. Posterior distributions of the six random-effects which capture the effect of arboreality on swimming mode. Abbreviations are "alt" for back legs moving in an alternating fashion, "sim" for back legs moving in a simultaneous fashion, and "both" for both types of movements. Each row groups forward and reverse transitions. The shape of the posterior distributions is strongly indicative of parameters which have been shrunk out of the model via the Bayesian bridge prior, and all posterior sign probabilities are less than 0.57.

on a MacBook Pro with an 8-core CPU M1 Pro chip and 32 GB of memory.

Analyses of simulated data

To assess the performance of random-effects substitution models in estimation of model parameters, we performed a simulation study. We based the simulation setup on our analysis of the SARS-CoV-2 data. In particular, we used the posterior distribution of trees, the HKY κ parameter, and the shape parameter governing the Gamma-distributed among-site rate variation. For the random-effects, we simulated from three groups: null effects, moderate effects, and strong effects. For each of these we drew values from Normal distributions (Supplementary Fig. S3) chosen to reflect the values observed in the real-data posterior distributions. The strong effects were C→T and G→T, which simulated from a Normal(1.50,0.18) distribution. The

moderate effects were A→T and G→A, which were simulated from Normal(−0.68,0.37) and Normal(0.68,0.37) distributions, respectively. The remaining 8 random-effects were classified as null and simulated from Normal(0.0,0.11) distributions. We simulated 100 datasets under this model.

Analyses of simulated datasets were conducted following the analysis of the SARS-CoV-2 dataset, with two exceptions. First, we treated the tree as known. Second, we considered several values for the exponent parameter, each simulated dataset was analyzed four times with $\alpha = 1/8, 1/4, 1/2, 1$. These values range from strongly regularizing priors (small α) to the weakly regularizing Laplace prior ($\alpha = 1$).

Overall, we find that random-effects are well-estimated and that random-effects which imply strong deviations from the base model (HKY) can be confidently identified using the posterior sign probability. Overall the posterior means are strongly correlated
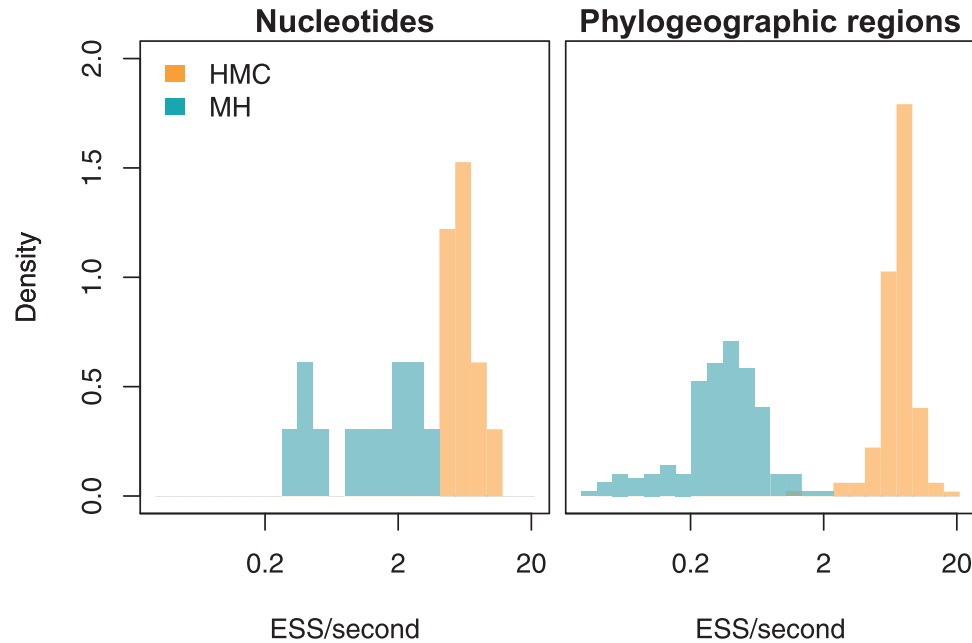
FIGURE 6. Efficiency, in effective samples per second (ESS/s), of HMC versus MH-MCMC for inferring the random-effects substitution models for both the 12 random-effects in an HKY+RE model and the 182 random-effects in the flu discrete phylogeographic analysis. Both datasets show markedly improved estimation efficiency as a result of employing HMC with approximate gradients.

with the true simulating value, ($\rho = 0.95$), although it appears that null and strong effects are generally better-estimated than moderate effects (Fig. 7 and Supplementary Fig. S5). To determine whether a random-effect is significant, one can use a threshold on the sign probability, declaring larger sign probabilities to be evidence for significant effects. Particularly notable deviations from the base model are easy to detect at any chosen threshold. Lower thresholds declare many negligible deviations to be significant, while higher thresholds are somewhat underpowered to detect smaller, but potentially important, deviations. A threshold of around 0.8 (0.75–0.85) provides a good balance between these forces (Supplementary Fig. S4).

Examining the choice of exponent α, we find that values on the order of 1/4 provide a reasonable trade-off between estimation performance and MCMC behavior. The smaller exponents, 1/8 and 1/4, in general produce posteriors which are notably closer to the true values than the larger values (Supplementary Fig. S5). The difference in performance is less notable for the evidence for significance provided by sign probabilities. The smaller exponents perform better for identifying null effects as null, while the larger exponents produce more confident support that moderate effects belong in the model, and all coefficients do well with strong effects (Supplementary Fig. S6). We investigated MCMC efficiency by examining the minimum effective sample size per sample. This efficiency is higher at $\alpha = 1/4$ than any other exponent (Supplementary Fig. S7).
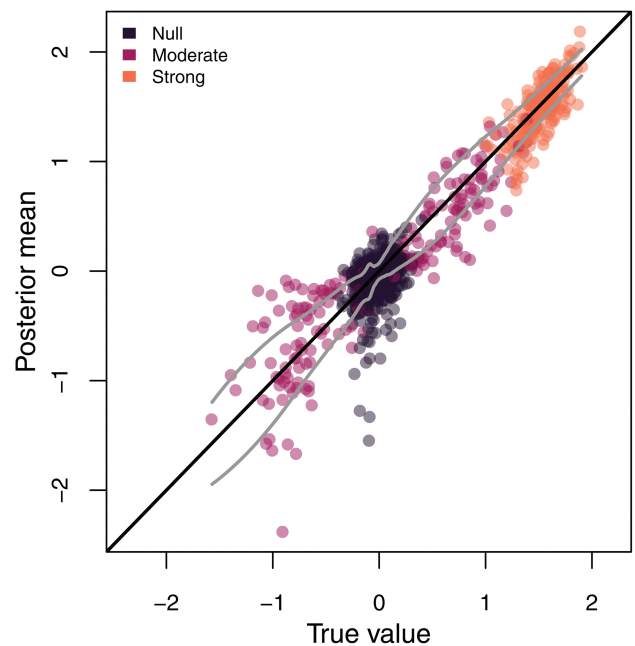

FIGURE 7. Estimation performance of HKY+RE on simulated data. Posterior mean parameter value versus true value simulated, colored by whether the true value was drawn from the distribution on null, moderate, or strong effects. As most posterior means are near the true value (close to the solid black line), parameter estimation is generally good. The solid grey lines display locally smoothed estimates of the standard deviation of the error $\hat{\theta} - \theta_{\text{true}}$, showing that null and strong effects are better-estimated than moderate ones.

Discussion

In this paper, we demonstrated the versatility and usefulness of random-effects substitution models. By wrapping around a simpler base substitution model, random-effects substitution models enable increased flexibility while retaining the useful structure of the base model. Applied to a dataset of 583 SARS-CoV-2 sequences, an HKY+RE model picks up strong C→T and G→T mutational biases and is shown by posterior predictive model checks to be an adequate substitution model where reversible models like GTR fail. Used with a GLM substitution model to analyze the phylogeographic pattern of spread of influenza in humans, the random-effects suggest the air traffic volume alone is a powerful explanation for the spread of influenza from 2002 to 2007. In examination of the evolution of ecological and behavioral characters in hylid tree frogs, a random-effects model shows no evidence for an effect of arboreality on the mode of swimming. Simulations show that random-effects can be accurately estimated and provide guidelines for interpreting whether a random-effect is significant or not.

To enable efficient inference of random-effects substitution models, we derived an approximate substitution gradient. The time-complexity of our approximate approach is cubic in the size of the state-space, while "exact" analytical techniques are quintic. For parameter-rich random-effects substitution models, numerical gradients are also quintic, and our approximate gradients enable *maximum a posteriori* inference of the parameters of an amino-acid substitution model over 50 times faster than numerical gradients ("Inferring the dynamics of amino acid substitution in Metazoa" section in the Supplementary Material available on Dryad). Used in Bayesian inference, we find that HMC using our approximate gradients is 6.6 to 20.2 times more efficient than standard Metropolis–Hastings moves, with yet more substantial gains when comparing the dimension with the most difficult sampling (where the efficiency gains are 14.8- and 33.6-fold). In particular, it appears that the efficiency of HMC with the approximate gradients is roughly invariant to the dimension (Fig. 6). For our SARS-CoV-2 example, with a 4 × 4 rate matrix, the average efficiency of HMC is 6.3 effective samples per second, while for the influenza A virus phylogeographic example, with a 14 × 14 rate matrix, it is 7.6 effective samples per second. However, the efficiency of Metropolis–Hastings moves decreases from 1.5 effective samplers per second to 0.41. We expect this trend to continue as the size of the state-space increases, and that for sufficiently large models (such as codon models or Markov-modulated amino acid models), HMC will be the only approach capable of inferring random-effects substitution models in any reasonable timeframe.

Although the approximate substitution gradient we derived performed very well in our applications, it cannot be expected to perform ideally in every circumstance. Mathematical analysis and in silico experiments suggest that the error in our approximation grows with the branch length measured in genetic distance ("Error in the approximate gradient" section in the Supplementary Material available on Dryad). Thus, we should expect performance to be best where the tree has few substitutions per site. Wertheim et al. (2022) refer to this as the near-perfect regime, and it is common in viral phylodynamic applications. However, we note reasons for optimism in applying our approximate gradients in regimes with larger numbers of substitutions. The influenza phylogeographic example falls outside the near-perfect regime, and the efficiency of HMC using our approximate gradients is still quite good. Similarly, good inference efficiency is observed in *maximum a posteriori* inference of an amino acid model on a Metazoan tree which has over 5.5 substitutions per site on average ("Inferring the dynamics of amino acid substitution in Metazoa" section in the Supplementary Material available on Dryad). It is also important to note that when used for HMC, the accept–reject step ensures correctness even in regimes where the approximation gets poor. It is likely the error bound we have obtained is quite conservative. Furthermore, Didier et al. (2023) establish a more rigorous error bound and show that the error decreases with increasingly large state-spaces, suggesting that phylogeographic analyses are well suited to this approximation.

An open question is to define the regimes where the approximation becomes poor enough that inference becomes inefficient such that other techniques would be preferable. We note two such alternative approaches which could be considered and compared with the efficiency of our approximation in future work. An exact gradient can be obtained from a data-augmentation procedure which jointly samples the complete mutational history along the tree, such as the approach adopted by Lartillot (2006). Within the framework of approximate gradients, an affine correction, as Didier et al. (2023) suggest, may yield smaller approximation error when the expected number of substitutions per branch is relatively large.

There are a number of important extensions of this work. Currently, we have implemented the gradient computations (in BEAST 1.10 (Suchard et al. 2018) and BEAGLE 3 (Ayres et al. 2019)) for use on CPUs; however, GPU-based likelihood computations have proven incredibly efficient in many phylogenetic contexts (Suchard and Rambaut 2009; Dudas et al. 2017; Ayres et al. 2019; Baele et al. 2021; Lemey et al. 2021). In particular, Gangavarapu et al. (2023) recently showed minimum increases of 8-fold and 128-fold for nucleotide and codon models, respectively, when computing gradients with respect to branch rate parameters. A GPU implementation of our approximate gradients would likely produce notable speedups in inference, especially for large state-space models. Mathematically, our approximation holds for any case in which there is a single substitution rate matrix on any edge of the phylogeny (although we have currently only implemented the case

for a single rate matrix across the whole tree). However, the process of geographic spread may be temporally inhomogeneous while applying consistently across all lineages alive at any given time. In such cases, epoch models (Bielejec et al. 2014; Gao et al. 2022) are needed. The epoch times break branches into multiple regimes, which requires matrix convolutions for likelihood computation and thus an extension of our approach.

Random-effects substitution models are a flexible approach for creating more realistic substitution models, but they are not a panacea. They cannot, for example, address gross violations of the underlying assumptions of the CTMC model, such as memorylessness. Nor can they address dependence between characters without carefully predefining the set of (potentially) coevolving characters and expanding the state space of the model. The Bayesian bridge provides a robust framework for regularization, and HMC an efficient framework for inference. However, the additional complexity of random-effects models may occasionally cause challenges for MCMC which require more active user intervention. Consider, for example, the (likely) APOBEC-induced C→T bias observed in our SARS-CoV-2 example, which, in a double-stranded virus, will also lead to a G→A bias (Gigante et al. 2022; O'Toole et al. 2023). Application of HKY+RE to such a dataset will lead to multimodality (caused by ridges in the likelihood) jointly involving five substitution model parameters, κ and the pairs of random-effects $\epsilon_{G \to A}, \epsilon_{C \to T}$ and $\epsilon_{A \to G}, \epsilon_{T \to C}$. Such multimodality does not invalidate the model, and it could be mitigated by the use of TN93+RE or avoided entirely by using a simpler model like F81+RE.

Supplementary material

Data available from the Dryad Digital Repository: https://doi.org/10.5068/D1709N.

Data availability

BEAST XML files for the analyses in this paper are available at, as well as the supplementary text, are available on Dryad, https://doi.org/10.5068/D1709N. The approximate gradients have been implemented in the hmc-clock branch of BEAST (https://github.com/beast-dev/beast-mcmc/tree/hmc-clock/) and the v4.0.0 release of BEAGLE (https://github.com/beagle-dev/beagle-lib/releases/tag/v4.0.0). BEAST XML files for the analyses in this paper are additionally available at https://github.com/suchard-group/approximate_substitution_gradient_supplement.

References

Abadi S., Azouri D., Pupko T., Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nat. Commun. 10(1):1–11.

Ayres D.L., Cummings M.P., Baele G., Darling A.E., Lewis P.O., Swofford D.L., Huelsenbeck J.P., Lemey P., Rambaut A., Suchard M.A. 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. Syst. Biol. 68(6):1052–1061.

Baele G., Gill M.S., Bastide P., Lemey P., Suchard M.A. 2021. Markov-modulated continuous-time Markov chains to identify site-and branch-specific evolutionary variation in BEAST. Syst. Biol. 70(1):181–189.

Bielejec F., Lemey P., Baele G., Rambaut A., Suchard M.A. 2014. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. Syst. Biol. 63(4):493–504.

Carvalho C.M., Polson N.G., Scott J.G. 2010. The horseshoe estimator for sparse signals. Biometrika 97(2):465–480.

Caviedes-Solis I.W. 2019. Intertwined evolution of swimming, morphology and microhabitat in tree frogs from the subfamily Hylinae. [PhD thesis]. University of Washington. https://digital.lib.washington.edu/researchworks/handle/1773/44728

Caviedes-Solis I.W., Kim N., Leaché A.D. 2020. Species IUCN threat status level increases with elevation: a phylogenetic approach for neotropical tree frog conservation. Biodivers. Conserv. 29(8):2515–2537.

Dayhoff M., Schwartz R., Orcutt B. 1978. A model of evolutionary change in proteins. In: Dayhoff M.O., editor. Atlas of protein sequence and structure. Washington, (DC): National Biomedical Research Foundation. p. 345–352.

Didier G., Glatt-Holtz N.E., Holbrook A.J., Magee A.F., Suchard M.A. 2024. On the surprising effectiveness of a simple matrix exponential derivative approximation, with application to global SARS-CoV-2. Proc. Natl. Acad. Sci. USA 16;121(3). http://dx.doi.org/10.1073/pnas.2318989121.

Duane S., Kennedy A.D., Pendleton B.J., Roweth D. 1987. Hybrid Monte Carlo. Phys. Lett. B, 195(2):216–222.

Dudas G., Carvalho L.M., Bedford T., Tatem A.J., Baele G., Faria N.R., Park D.J., Ladner J.T., Arias A., Asogun D., Bielejec F., Caddy S.L., Cotten M., D'Ambrozio J., Dellicour S., Di Caro A., Diclaro J.W., Duraffour S., Elmore M.J., Fakoli L.S., Faye O., Gilbert M.L., Gevao S.M., Gire S., Gladden-Young A., Gnirke A., Goba A., Grant D.S., Haagmans B.L., Hiscox J.A., Jah U., Kugelman J.R., Liu D., Lu J., Malboeuf C.M., Mate S., Matthews D.A., Matranga C.B., Meredith L.W., Qu J., Quick J., Pas S.D., Phan M.V.T., Pollakis G., Reusken C.B., Sanchez-Lockhart M., Schaffner S.F., Schieffelin J.S., Sealfon R.S., Simon-Loriere E., Smits S.L., Stoecker K., Thorne L., Tobin E.A., Vandi M.A., Watson S.J., West K., Shannon W., Wiley M.R., Winnicki S.M., Wohl S., Wölfel R., Yozwiak N.L., Andersen K.G., Blyden S.O., Bolay F., Carroll M.W., Dahn B., Diallo B., Formenty P., Fraser C., Gao G.F., Garry R.F., Goodfellow I., Günther S., Happi C.T., Holmes E.C., Kargbo B., Keïta S., Kellam P., Koopmans M.P.G., Kuhn J.H., Loman N.J., Magassouba N.F., Naidoo D., Nichol S.T., Nyenswah T., Palacios G., Pybus O.G., Sabeti P.C., Sall A., Ströher U., Wurie I., Suchard M.A., Lemey P., Rambaut Andrew. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544(7650):309.

Fabreti L.G., Höhna S. 2022. Bayesian inference of phylogeny is robust to substitution model over-parameterization. bioRxiv. https://doi.org/10.1101/2022.02.17.480861, preprint: not peer reviewed.

Faulkner J.R., Magee A.F., Shapiro B., Minin V.N. 2020. Horseshoe-based Bayesian nonparametric estimation of effective population size trajectories. Biometrics 76(3):677–690.

Fisher A.A., Ji X., Nishimura A., Lemey P., Suchard M.A. 2021. Shrinkage-based random local clocks with scalable inference. Mol Biol Evol 40(11):msad242. https://doi.org/10.1093/molbev/msad242.

Fisher A.A., Ji X., Zhang Z., Lemey P., Suchard M.A. 2021. Relaxed random walks at scale. Syst. Biol. 70(2):258–267.

Gangavarapu K., Ji X., Baele G., Fourment M., Lemey P., Matsen IV F.A., Suchard M.A. 2024. Many-core algorithms for high-dimensional gradients on phylogenetic trees. Bioinformatics 40(2):btae030. http://dx.doi.org/10.1093/bioinformatics/btae030

Gao J., May M.R., Rannala B., Moore B.R. 2022. New phylogenetic models incorporating interval-specific dispersal dynamics improve inference of disease spread. Mol. Biol. Evol., 39(8):msac159.

Gigante C.M., Korber B., Seabolt M.H., Wilkins K., Davidson W., Rao A.K., Zhao H., Smith T.G., Hughes C.M., Minhaj F., Waltenburg M.A., Theiler J., Smole S., Gallagher G.R., Blythe D., Myers R., Schulte J., Stringer J., Lee P., Mendoza R.M., Griffin-Thomas L.A., Crain J., Murray J., Atkinson A., Gonzalez A.H., Nash J., Batra D., Damon I., McQuiston J., Hutson C.L., McCollum A.M., Li Y. 2022. Multiple lineages of monkeypox virus detected in the United States, 2021–2022. Science 378(6619):560–565.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22(2):160–174.

Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109.

Hilton S.K., Bloom J.D. 2018. Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence. Virus Evol. 4(2):vey033.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press, p. 21–132.

Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Statist. Assoc. 90 (430):773–795.

Lan S., Bui-Thanh T., Christie M., Girolami M. 2016. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. J. Comput. Phys. 308:81–101.

Lartillot N. 2006. Conjugate Gibbs sampling for Bayesian phylogenetic models. J. Comput. Biol. 13(10):1701–1722.

Leimkuhler B., Reich S. 2004 *Simulating Hamiltonian dynamics*. Number 14 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge, England: Cambridge University Press.

Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5(9):e1000520.

Lemey P., Rambaut A., Bedford T., Faria N., Bielejec F., Baele G., Russell C.A., Smith D.J., Pybus O.G., Brockmann D., Suchard M.A. 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathogens 10(2):e1003932.

Lemey P., Hong S.L., Hill V., Baele G., Poletto C., Colizza V., O'Toole Á., McCrone J.T., Andersen K.G., Worobey M., Nelson M.I., Rambaut A., Suchard M.A. 2020. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat. Commun. 11(1):1–14.

Lemey P., Ruktanonchai N., Hong S.L., Colizza V., Poletto C., Van den Broeck F., Gill M.S., Ji X., Levasseur A., Oude Munnink B.B., Koopmans M., Sadilek A., Lai S., Tatem A.J., Baele G., Suchard M.A., Dellicour S. 2021. Untangling introductions and persistence in COVID-19 resurgence in Europe. Nature 595(7869): 713–717.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50(6):913–925.

Li L., Holbrook A., Shahbaba B., Baldi P. 2019. Neural network gradient Hamiltonian Monte Carlo. Comput. Statist. 34(1):281–299.

Magee A.F., Höhna S., Vasylyeva T.I., Leaché A.D., Minin V.N. 2020. Locally adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts. PLoS Comput. Biol. 16(10):e1007999.

Matyášek R., Kovařík A. 2020. Mutation patterns of human SARS-CoV-2 and bat RaTG13 coronavirus genomes are strongly biased towards c>u transitions, indicating rapid evolution in their hosts. Genes 11(7):761.

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

Najfeld I., Havel T.F. 1995. Derivatives of the matrix exponential and their computation. Adv. Appl. Math. 16(3):321–375.

Neal R.M. MCMC using Hamiltonian dynamics. In: Brooks S., Gelman A., Jones G., Meng X.-L. editors. Handbook of Markov chain Monte Carlo. Chapman & Hall/CRC Press.

Nishimura A., Suchard M.A. 2022. Shrinkage with shrunken shoulders: Gibbs sampling shrinkage model posteriors with guaranteed convergence rates. Bayesian Anal. 18(2):367–390.

O'Toole Á., Neher R.A., Ndodo N., Borges V., Gannon B., Gomes J.P., Groves N., King D.J., Maloney D., Lemey P., Lewandowski K., Loman N., Myers R., Suchard M.A., Worobey M., Chand M., Ihekweazu C., Ulaeto D., Adetifa I., Rambaut Á. 2023. Putative APOBEC3 deaminase editing in MPXV as evidence for sustained human transmission since at least 2016. BioRXiv, p. 2023–01. https://doi.org/10.1101/2023.01.23.525187, preprint: not peer reviewed.

Pagel M., Meade, A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. 167(6):808–825.

Pekar J., Worobey M., Moshiri N., Scheffler K., Wertheim J.O. 2021. Timing the SARS-CoV-2 index case in Hubei province. Science 372 (6540):412–417.

Petersen K.B., Pedersen M.S. 2008. The matrix cookbook. Technical University of Denmark 7(15):510.

Polson N.G., Scott J.G., Windle J. 2014. The Bayesian bridge. J. R. Statist. Soc. Ser. B Statist. Methodol. 76(4):713–733.

Rasmussen C.E. 2003. Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. Seventh Valencia International Meeting, dedicated to Dennis V. Lindley. Oxford, England: Oxford University Press. p. 651–659.

Suchard M.A., Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25(11):1370–1376.

Suchard M.A., Weiss R.E., Sinsheimer J.S. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18(6):1001–1013.

Suchard M.A., Weiss R.E., Dorman K.S., Sinsheimer J.S. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model. J. Am. Statist. Assoc. 98(462):427–437.

Suchard M.A., Lemey P., Baele G., Ayres D.L., Drummond A.J., Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4(1):vey016.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17(2): 57–86.

Tonkin-Hill G., Martincorena I., Amato R., Lawson A.R.J., Gerstung M., Johnston I., Jackson D.K., Park N., Lensing S.V., Quail M.A., Gonçalves S., Ariani C., Chapman M.S., Hamilton W.L., Meredith L.W., Hall G., Jahun A.S., Chaudhry Y., Hosmillo M., Pinckert M.L., Georgana I., Yakovleva A., Caller L.G., Caddy S.L., Feltwell T., Khokhar F.A., Houldcroft C.J., Curran M.D., Parmar S.; COVID-19 Genomics UK (COG-UK) Consortium; Alderton A., Nelson R., Harrison E.M., Sillitoe J., Bentley S.D., Barrett J.C., Torok M.E., Goodfellow I.G., Langford C., Kwiatkowski D.; Wellcome Sanger Institute COVID-19 Surveillance Team. 2021. Patterns of within-host genetic diversity in SARS-CoV-2. Elife, 10: e66857.

Vats D. 2023. Hamiltonian Monte Carlo for (physics) dummies. https://dvats.github.io/assets/pdf/HMCforDummies.pdf.

Wagenmakers E.-J., Lodewyckx T., Kuriyal H., Grasman R. 2010. Bayesian hypothesis testing for psychologists: a tutorial on the savage–dickey method. Cognit. Psychol. 60(3):158–189.

Wertheim J.O., Steel M., Sanderson M.J. 2022. Accuracy in near-perfect virus phylogenies. Syst. Biol. 71(2):426–438.

Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18(5):691–699.

Woodhams M.D., Fernández-Sánchez J., Sumner J.G. 2015. A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. Syst. Biol. 64(4):638–650.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39(1):105–111.

Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39(3):306–314.

Yang Z., Nielsen R., Goldman N., Pedersen A.M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155(1):431–449.

Zhang C., Shahbaba B., Zhao H. 2017a. Hamiltonian Monte Carlo acceleration using surrogate functions with random bases. Statist. Comput. 27(6):1473–1490.

Zhang C., Shahbaba B., Zhao H. 2017b. Precomputing strategy for Hamiltonian Monte Carlo method based on regularity in parameter space. Comput. Statist. 32(1):253–279.

Zhang Z., Nishimura A., Bastide P., Ji X., Payne R.P., Goulder P., Lemey P., Suchard M.A. 2021. Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models. Ann. Appl. Statist. 15(1):230–251. doi: 10.1214/20-AOAS1394. https://doi.org/10.1214/20-AOAS1394.