PNAS RESEARCH ARTICLE | APPLIED MATHEMATICS OPEN ACCESS

On the surprising effectiveness of a simple matrix exponential derivative approximation, with application to global SARS-CoV-2

Gustavo Didier[a], Nathan E. Glatt-Holtz[a], Andrew J. Holbrook[b,1], Andrew F. Magee[b], and Marc A. Suchard[b,c,d]

Edited by James Bull, University of Idaho, Moscow, ID; received November 6, 2023; accepted November 30, 2023

The continuous-time Markov chain (CTMC) is the mathematical workhorse of evolutionary biology. Learning CTMC model parameters using modern, gradient-based methods requires the derivative of the matrix exponential evaluated at the CTMC's infinitesimal generator (rate) matrix. Motivated by the derivative's extreme computational complexity as a function of state space cardinality, recent work demonstrates the surprising effectiveness of a naive, first-order approximation for a host of problems in computational biology. In response to this empirical success, we obtain rigorous deterministic and probabilistic bounds for the error accrued by the naive approximation and establish a "blessing of dimensionality" result that is universal for a large class of rate matrices with random entries. Finally, we apply the first-order approximation within surrogate-trajectory Hamiltonian Monte Carlo for the analysis of the early spread of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) across 44 geographic regions that comprise a state space of unprecedented dimensionality for unstructured (flexible) CTMC models within evolutionary biology.

continuous-time Markov chains | Hamiltonian Monte Carlo | matrix exponential | molecular epidemiology | random matrix theory

Significance

Recent work uses a simplistic approximation to the matrix exponential derivative to apply gold-standard models from evolutionary biology to a collection of challenging data analyses. Whereas one may expect the naive approach to break down with increasing model dimensionality, empirical results show no such failure. Here, we 1) develop rigorous error bounds that improve—in a certain sense—as model dimension grows and 2) demonstrate the scalability of the naive approach to a higher-dimensional analysis of the global spread of the virus responsible for the COVID-19 pandemic.

Phylogeographic methods (1–4) model large-scale viral transmission between human populations as a function of the shared evolutionary history of the viral population of interest. Data take the form of dates, locations, and genome sequences associated with individual viral samples. Spatiotemporal structure interfaces with network structure given by the phylogeny, or family tree, describing the viruses' collective history beginning with the most recent common ancestor. While one cannot directly observe this history, one may statistically reconstruct the phylogenetic tree by positing that changes in the viral genome happen randomly at regular intervals, thereby capturing the intuition that viral samples with more differences between their (aligned) sequences should find themselves further apart on the family tree.

The continuous-time Markov chain (CTMC) (5) represents the gold-standard mathematical model for such evolution of characters (e.g., nucleotides) within a fixed span of evolutionary time. A CTMC defined over a discrete, d-element state space consists of a row vector $\boldsymbol{\pi}_0$ whose individual components describe the probability of inhabiting each of the possible states at time $t = 0$ as well as a $d \times d$ infinitesimal generator (or rate) matrix $\mathbf{Q}$ with nonnegative off-diagonal elements q_{ij}, $i \neq j$, and nonpositive diagonal elements $q_{ii} = -\sum_j q_{ij}$. For any lag $t \geq 0$, the matrix exponential (6, 7) provides the Markov chain's transition probability matrix

$$\mathbf{P}_t := e^{t\mathbf{Q}} := \sum_{n=0}^{\infty} \frac{t^n \mathbf{Q}^n}{n!}, \qquad [1]$$

which has elements $[\mathbf{P}_t]_{ij}$ that dictate the probability of the process jumping from state i to state j after time t. It is straightforward to verify that $\mathbf{P}_t$ is a valid transition matrix, having probability vectors for rows: If $\mathbf{1}$ and $\mathbf{0}$ are the column vectors of ones and zeros, respectively, then $\mathbf{Q1} = \mathbf{0}$ and, therefore, $\mathbf{P}_t\mathbf{1} = \mathbf{1}$. The law of total probability then provides the marginal probability of the process at any time $t \geq 0$ as $\boldsymbol{\pi}_t = \boldsymbol{\pi}_0 e^{t\mathbf{Q}}$.

Whether frequentist (8) or Bayesian (9–12) likelihood-based approaches to phylogenetic reconstruction allow phylogenetic tree branch lengths to parameterize time lags within the CTMC framework. We present the exact statement of the phylogenetic CTMC paradigm below (Section 3). Here, we note that the historical importance of tree-reconstruction from aligned sequences leads to an early emphasis on the sparse specification of $\mathbf{Q}$ based on biologically motivated assumptions (13–15). Classical Markov chain Monte Carlo (MCMC) procedures (16, 17) work well for such low-dimensional models. But the phylogenetic CTMC framework has applications

Author affiliations: [a]Department of Mathematics, Tulane University, New Orleans, LA 70118; [b]Department of Biostatistics, University of California, Los Angeles, CA 90095; [c]Department of Biomathematics, University of California, Los Angeles, CA 90095; and [d]Department of Human Genetics, University of California, Los Angeles, CA 90095

Author contributions: A.J.H. and M.A.S. designed research; G.D., N.E.G.-H., and A.J.H. calculated mathematical results; A.J.H. and A.F.M. performed simulations and data analysis; and G.D., N.E.G.-H., and A.J.H. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: aholbroo@g.ucla.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2318989121/-/DCSupplemental.

Published January 12, 2024.

beyond simple nucleotide substitution models. Within, e.g., Bayesian phylogeography, the work in ref. 1 provides a phylogenetic CTMC model for the spread of avian influenza across $d = 20$ global geographic locations but, for computational reasons, favors a low-dimensional $\mathcal{O}(d)$ parameterization of $\mathbf{Q}$. Similarly, Lemey et al. (2) model the spread of influenza A H1N1 and H3N2 between as many as $d = 26$ geographic regions but—again for computational reasons—fit the model with approximation techniques that provide no inferential guarantees.

Recently, Magee et al. (18) demonstrate the feasibility of approximate gradient-based methods for both maximum a posteriori and full Bayesian inference of flexible and fully-parameterized rate models and apply these methods to a gold-standard $\mathcal{O}(d^2)$ mixed-effect CTMC model for the spread of A H3N2 influenza between $d = 14$ geographic locations. The usual CTMC log-likelihood gradient calculations feature the matrix exponential derivative (Eqs. **13** and **15**)

$$\nabla_{\mathbf{J}} e^{t\mathbf{Q}} := \lim_{\epsilon \to 0} \frac{e^{t(\mathbf{Q}+\epsilon\mathbf{J})} - e^{t\mathbf{Q}}}{\epsilon} \quad [2]$$

$$= e^{t\mathbf{Q}} \sum_{n=0}^{\infty} \frac{t^{n+1}}{(n+1)!} \left(\sum_{\ell=0}^{n} (-1)^{\ell} \binom{n}{\ell} \mathbf{Q}^{\ell} \mathbf{J} \mathbf{Q}^{n-\ell} \right) \quad [3]$$

computed in the direction $\mathbf{J}$ of each of the d^2 natural basis elements $\mathbf{J}_{ij}$ spanning the space of real-valued, $d \times d$ matrices $\mathcal{M}(d) = \mathcal{M}(d, \mathbb{R})$, thereby requiring at least $\mathcal{O}(d^5)$ floating point operations (19).

Within the phylogenetic CTMC models of Section 3, log-likelihood derivative computations that require $\nabla_{\mathbf{J}} e^{t\mathbf{Q}}$ balloon to $\mathcal{O}(KNd^5)$, for N the number of biological specimens observed and K the number of parameters parameterizing $\mathbf{Q}$. To address this overwhelming computational cost, Magee et al. (18) leverage the simplistic approximation obtained by setting $n = 0$ within Eq. **3**:

$$\widetilde{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}} := t e^{t\mathbf{Q}} \mathbf{J}. \quad [4]$$

Ref. 18 show that this approximation helps reduce total cost to $\mathcal{O}(Kd^2 + Nd^3)$ and use this speedup within surrogate-trajectory Hamiltonian Monte Carlo (HMC) (*SI Appendix*) to obtain a 34-fold improvement in effective sample size per second (ESS/s) over random-walk MCMC within their 14-region phylogeographic example. When trying to explain the remarkable empirical performance of the naive approximation, the authors derive an error upper bound (for an arbitrary matrix norm)

$$\| \widetilde{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}} - \nabla_{\mathbf{J}} e^{t\mathbf{Q}} \| \leq \frac{\|\mathbf{J}\| \|\mathbf{Q}\|}{2} (e^{2t} - 2t - 1) \quad [5]$$

that fails to leverage the specific forms of $\mathbf{Q}$ and $\mathbf{J}$. Notably, this bound explodes as either t or $\|\mathbf{Q}\|$ diverges to ∞. Of course, the latter quantity would be expected to grow large with dimension d without more careful structural assumptions, e.g., that $\mathbf{Q}$ is a rate matrix belonging to the class

$$\mathcal{R}(d) := \left\{ \mathbf{Q} \in \mathcal{M}(d) | \mathbf{Q}_{ij} \geq 0 \text{ for } j \neq i, \sum_{j=1}^{d} \mathbf{Q}_{ij} = 0 \right\}. \quad [6]$$

In the following, we use the finer structural properties of $\mathbf{Q}$ to obtain more precise bounds on

$$\mathbf{E}(t) := \nabla_{\mathbf{J}} e^{t\mathbf{Q}} - t e^{t\mathbf{Q}} \mathbf{J}. \quad [7]$$

In Theorem 1, we provide an affine (in t) correction to the approximation Eq. **4** that yields an exponentially tight $t \to \infty$ asymptotic for the error Eq. **7**. Then, in Theorem 3, we establish precise probabilistic bounds in the high-dimensional $d \to \infty$ limit for a large class of randomly drawn rate matrices $\mathbf{Q} \in \mathcal{R}$. Here, we show, for any $\mathbf{Q} \in \mathcal{R}$ whose off-diagonal elements are determined by independently and identically distributed (iid) draws from a positive, sub-exponential distribution F, that all of the nonzero singular values grow as $\sigma_j(\mathbf{Q}) \sim d$ along with asymptotically valid almost sure bounds for these rates as $d \to \infty$.

In regards to this second result, Theorem 3, note that random rate (or "Laplacian") matrices have attracted a great deal of attention from the probability research community, especially in regard to their high-dimensional properties; see e.g., refs. 20–25. In particular, seminal papers such as refs. 26–28 establish broad characterizations of bulk behavior for the Laplacian eigenspectrum. One contribution of this paper, of independent interest, is a short and self-contained construction of useful bounds for the singular values of $\mathbf{Q} \in \mathcal{R}$.

In Theorem 2 and Corollary 1, we show how Theorems 1 and 3 combine to provide a more refined analysis of $\mathbf{E}$ for suitable randomly generated $\mathbf{Q} \in \mathcal{R}(d)$. In Theorem 2, we establish that particular terms appearing in the bound Eq. **24** in Theorem 1 decay with a rate on the order of $1/d$ in the operator norm topology for large d for certain classes of symmetric generators $\mathbf{Q}$. This class includes the random elements considered in Theorem 3. Here, although Corollary 1 applies only for symmetric matrices composed of sub-exponential draws, we provide strong supplemental numerical evidence that our bounds remain valid well beyond this special symmetric, sub-exponential special case (Remark 5 and Fig. 1).

One notable practical implication of Theorem 2, Corollary 1 and Remark 5 is the identification of a further correction to Eq. **4**. Crucially, this correction has the same $\mathcal{O}(d^3)$ computational cost as Eq. **4** while leading to an asymptotically temporally uniformly accurate approximation of $\nabla_{\mathbf{J}} e^{t\mathbf{Q}}$ (Remark 5 and Eq. **46**).

Section 2 contains simulation studies comparing accuracy of matrix exponential derivative approximations for different distributional assumptions on the generator matrix; the posterior distributions obtained using surrogate-trajectory HMC and traditional HMC; and parameter identification under different priors on generator matrix elements.

In Section 3, we follow these theoretical and empirical investigations with an application of the naive, first-order gradient approximation Eq. **4** to a challenge in phylogeography requiring Bayesian inference of a rate-matrix of unprecedented dimensionality. Namely, we apply the approximation to a gold-standard mixed-effects, phylogenetic CTMC model that uses 1,897 parameters to describe the spread of SARS-CoV-2 across a $d = 44$ dimensional state space consisting of different global geographic locations. Such an application complements the empirical studies of ref. 18 in a manner that emphasizes the naive approximation's potential for impact.

1. Rigorous Results

This section lays out our rigorous results Theorems 1–3 and Corollary 1, the proofs of which appear in *SI Appendix*.

In what follows, we adopt the following notational conventions. For any $\mathbf{A} \in \mathcal{M}(d)$, we list the associated (not necessarily

distinct) eigenvalues of $\mathbf{A}$ in ascending order according to their real part, namely,

$$\Re\lambda_1(\mathbf{A}) \le \cdots \le \Re\lambda_d(\mathbf{A}). \quad [8]$$

Similarly, the singular values of $\mathbf{A}$ are written in ascending order

$$\sigma_1(\mathbf{A}) \le \cdots \le \sigma_d(\mathbf{A}). \quad [9]$$

We let $\mathcal{S}(d) = \mathcal{S}(d,\mathbb{R})$ and $\mathcal{S}(d,\mathbb{C})$ represent the spaces of $d \times d$ symmetric and Hermitian matrices, respectively.

We make use of multiple matrix norms leading to materially different bounds as $d \to \infty$ (29). Take

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j=1}^{d} \mathbf{A}_{ij}^2} = \sqrt{\sum_{j=1}^{d} \sigma_j(\mathbf{A})^2} \quad [10]$$

for the Frobenius norm and

$$\|\mathbf{A}\|_{\mathrm{op}} := \sqrt{\lambda_d(\mathbf{A}^*\mathbf{A})} = \sqrt{\lambda_d(\mathbf{A}\mathbf{A}^*)} = \sigma_d(\mathbf{A}) \quad [11]$$

for the operator norm of $\mathbf{A}$. Finally, note that when we simply write $\|\cdot\|$; the statement then holds for any valid matrix norm as in our formulation of Theorem 1.

1.1. Deterministic Bounds on Approximation Error in Time. We begin by deriving a dynamical equation for the error $\mathbf{E}(t)$, defined in Eq. 7. Recall that $\mathbf{X}(t) := e^{t\mathbf{Q}}$ for any $\mathbf{Q} \in \mathcal{M}(d)$ obeys the (matrix-valued) ordinary differential equation

$$\frac{d\mathbf{X}}{dt} = \mathbf{Q}\mathbf{X}, \quad \mathbf{X}(0) = \mathbf{I}. \quad [12]$$

Setting $\mathbf{Y}^\epsilon = \epsilon^{-1}(e^{t(\mathbf{Q}+\epsilon\mathbf{J})} - e^{t\mathbf{Q}})$ and taking a limit as $\epsilon \to 0$, we find that $\mathbf{Y} = \nabla_{\mathbf{J}} e^{t\mathbf{Q}}$ obeys

$$\frac{d\mathbf{Y}}{dt} = \mathbf{Q}\mathbf{Y} + \mathbf{J}\mathbf{X} = \mathbf{Q}\mathbf{Y} + \mathbf{J}e^{t\mathbf{Q}}, \quad \mathbf{Y}(0) = \mathbf{0}. \quad [13]$$

Thus, variation of constants yields that, for any $t \ge 0$,

$$\nabla_{\mathbf{J}} e^{t\mathbf{Q}} = e^{t\mathbf{Q}} \int_0^t e^{-s\mathbf{Q}} \mathbf{J} e^{s\mathbf{Q}} ds \quad [14]$$

$$= e^{t\mathbf{Q}} \sum_{k,m=0}^{\infty} \int_0^t \frac{(-s)^k \mathbf{Q}^k \mathbf{J} s^m \mathbf{Q}^m}{k!m!} ds$$

$$= e^{t\mathbf{Q}} \sum_{n=0}^{\infty} \frac{t^{n+1}}{(n+1)!} \left(\sum_{\ell=0}^{n} (-1)^\ell \binom{n}{\ell} \mathbf{Q}^\ell \mathbf{J} \mathbf{Q}^{n-\ell} \right). \quad [15]$$

Taking the first-order ($n = 0$) approximation in Eq. **15** produces Eq. **4**. Note that, from Eq. **14**, this approximation $\widetilde{\nabla} e^{t\mathbf{Q}}\mathbf{J}$ is evidently exact in the special case when $\mathbf{J}$ and $\mathbf{Q}$ commute.

Next notice that, if we differentiate $\widetilde{\mathbf{Y}} := te^{t\mathbf{Q}}\mathbf{J}$ in t, we find that $\widetilde{\mathbf{Y}}$ obeys

$$\frac{d\widetilde{\mathbf{Y}}}{dt} = \mathbf{Q}\widetilde{\mathbf{Y}} + e^{t\mathbf{Q}}\mathbf{J}, \quad \widetilde{\mathbf{Y}}(0) = \mathbf{0}. \quad [16]$$

Thus, taking the error $\mathbf{E}(t)$ as in Eq. 7 and combining Eq. **13** with Eq. **16** yields

$$\frac{d\mathbf{E}}{dt} = \mathbf{Q}\mathbf{E} + \mathbf{J}e^{t\mathbf{Q}} - e^{t\mathbf{Q}}\mathbf{J}, \quad \mathbf{E}(0) = \mathbf{0}. \quad [17]$$

Hence, again integrating this expression, we find

$$\mathbf{E}(t) = e^{t\mathbf{Q}} \int_0^t (e^{-s\mathbf{Q}} \mathbf{J} e^{s\mathbf{Q}} - \mathbf{J}) ds \quad [18]$$

$$= e^{t\mathbf{Q}} \sum_{n=1}^{\infty} \frac{t^{n+1}}{(n+1)!} \left(\sum_{\ell=0}^{n} (-1)^\ell \binom{n}{\ell} \mathbf{Q}^\ell \mathbf{J} \mathbf{Q}^{n-\ell} \right) \quad [19]$$

as could also be directly deduced from Eqs. **14** to **15**.

Given a rate matrix $\mathbf{Q}$ in $\mathcal{R}(d)$, recall that $\Re\lambda_{d-1}(\mathbf{Q}) \le \lambda_d(\mathbf{Q}) = 0$ by the Gershgorin circle theorem (30, Theorem 6.1.1). Imposing a further non-degeneracy assumption (e.g., that $\Re\lambda_{d-1}(\mathbf{Q}) < 0$) we therefore have an exponential decay in $\mathbf{Q}e^{t\mathbf{Q}}$. This starting point suggests that, under fairly general conditions, we may decompose Eq. **18** into a component where $\mathbf{Q}e^{t\mathbf{Q}}$ induces an exponential decay in time and a complementary component taking the form of a time-affine correction term.

These observations lead to the following theorem, the proof of which appears in *SI Appendix*.

Theorem 1. *Suppose that $\mathbf{Q}, \mathbf{J} \in \mathcal{M}(d)$ for some $d \ge 1$. We assume that we can find an element $\mathbf{Q}^+ \in \mathcal{M}(d)$ such that $\mathbf{Q}$ is a generalized inverse of $\mathbf{Q}^+$, namely,*

$$\mathbf{Q}^+\mathbf{Q}\mathbf{Q}^+ = \mathbf{Q}^+ \quad [20]$$

and such that

$$e^{\tau\mathbf{Q}}(\mathbf{I} - \mathbf{Q}^+\mathbf{Q}) = \mathbf{I} - \mathbf{Q}^+\mathbf{Q}, \ (\mathbf{I} - \mathbf{Q}\mathbf{Q}^+)e^{\tau\mathbf{Q}} = \mathbf{I} - \mathbf{Q}\mathbf{Q}^+, \quad [21]$$

for any $\tau \in \mathbb{R}$. Furthermore, we suppose that $\mathbf{Q}^+$ and $\mathbf{Q}$ commute

$$\mathbf{Q}\mathbf{Q}^+ = \mathbf{Q}^+\mathbf{Q}. \quad [22]$$

Finally, taking $\|\cdot\|$ be any matrix norm, we assume that

$$\|\mathbf{Q}e^{\tau\mathbf{Q}}\| \le C_0 e^{-\kappa\tau}, \quad \text{for all } \tau \ge 0, \quad [23]$$

where the constants $C_0 > 0, \kappa > 0$ are independent of τ. Then, under these circumstances,

$$\|\mathbf{Q}^+\mathbf{J}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^+) - (\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}^+ + t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}} e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\| \le C(1+t)e^{-\kappa t} \quad [24]$$

for any $t \ge 0$. Here, $C > 0$ is a t-independent constant which is given explicitly as

$$C_0\big(\|(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}(\mathbf{Q}^+)^2\| + \|(\mathbf{Q}^+)^2\mathbf{J}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^+)\| + \|\mathbf{Q}^+\mathbf{J}\|\big) + C_0^2\|\mathbf{Q}^+\mathbf{J}\mathbf{Q}^+\|. \quad [25]$$

Remark 1: To illuminate the scope of Theorem 1, we have the following three classes of matrices maintaining the conditions Eqs. **20–23** as follows.

(i) Suppose that $\mathbf{Q} \in \mathcal{M}(d)$ is such that

$$\Re\lambda_{d-1}(\mathbf{Q}) < \Re\lambda_d(\mathbf{Q}) \le 0, \ \lambda_d(\mathbf{Q}) \text{ is simple} \quad [26]$$

and such that, if $\lambda_d(\mathbf{Q})$ has an imaginary component, then its real part is strictly negative. Under these circumstances,

writing $\mathbf{Q}$ in its Jordan canonical form yields, for some $m \geq 1$,

$$\mathbf{Q} = \mathbf{M}\mathrm{diag}(J_1, \ldots, J_{m-1}, \lambda_d(\mathbf{Q}))\mathbf{M}^{-1}. \quad [27]$$

Here, under Eq. **26** each of these blocks J_j must be invertible and so we may take

$$\mathbf{Q}^+ := \mathbf{M}\,\mathrm{diag}(J_1^{-1}, \ldots, J_{m-1}^{-1}, \lambda_d(\mathbf{Q})^+)\,\mathbf{M}^{-1}, \quad [28]$$

where

$$\lambda_d(\mathbf{Q})^+ = \begin{cases} 0 & \text{if } \lambda_d(\mathbf{Q}) = 0, \\ \lambda_d(\mathbf{Q})^{-1} & \text{otherwise.} \end{cases} \quad [29]$$

(ii) We next consider the case where $\mathbf{Q} \in \mathcal{M}(d)$ is diagonalizable and its spectrum lies strictly on the left half plane or at the origin. This time, we can write

$$\mathbf{Q} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^{-1} \text{ where } \mathbf{\Lambda} = \mathrm{diag}(\lambda_1(\mathbf{Q}), \ldots, \lambda_d(\mathbf{Q})) \quad [30]$$

and we set

$$\mathbf{Q}^+ = \mathbf{M}\mathbf{\Lambda}^+\mathbf{M}^{-1} \text{ with } \mathbf{\Lambda}^+ = \mathrm{diag}(\lambda_1(\mathbf{Q})^+, \ldots, \lambda_d(\mathbf{Q})^+). \quad [31]$$

The complex numbers $\lambda_j(\mathbf{Q})^+$, $j = 1, \ldots, d$ are defined as in Eq. **29**.

(iii) Finally, we specialize to the case where $\mathbf{Q} \in \mathcal{S}(d)$ is symmetric. In this case, $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$, where $\mathbf{U}$ is a unitary matrix and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1(\mathbf{Q}), \ldots, \lambda_d(\mathbf{Q}))$, $\lambda_j(\mathbf{Q})$ are its (real) eigenvalues. We suppose that these eigenvalues are all nonpositive,

$$\lambda_d(\mathbf{Q}) \leq 0. \quad [32]$$

Here, we take $\mathbf{Q}^+$ as the Moore-Penrose inverse, namely,

$$\mathbf{Q}^+ = \mathbf{U}\mathbf{\Lambda}^+\mathbf{U}^*, \quad [33]$$

where $\mathbf{\Lambda}^+$ is as in Eq. **31**.

In anticipation of Theorem 3 and our desired application in Section 3, we are preoccupied with the dimensional dependence of the constants in Eqs. **23**–**25** in our formulation of Theorem 1. We next provide some such desirable bounds in case (iii) of Remark 1. Note that analogous results for generators $\mathbf{Q}$ in the classes (i) or (ii) would seemingly require a delicate analysis of the associated eigenspaces, i.e., of the structure of $\mathbf{M}$ in Eq. **27** or Eq. **30** respectively. However, Remark 4 and Fig. 1 provide numerical evidence of a broader scope for dimensionally improving approximations beyond the symmetric case, at least for certain classes of randomly drawn matrices.

Theorem 2. *Let symmetric $\mathbf{Q} \in \mathcal{S}(d)$ be nonpositive, i.e., suppose that* Eq. **32** *holds. Take $\mathbf{Q}^+$ as in* Eq. **33** *and define*

$$d_- = \max\{1 \leq j \leq d | \Re\lambda_j(\mathbf{Q}) < 0\}. \quad [34]$$

Then, for any $\mathbf{J} \in \mathcal{M}(d)$,

$$\begin{aligned} &\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_F \\ &\leq \Big(\sqrt{d_-}|\lambda_1(\mathbf{Q})| \cdot \Big[2\sqrt{d - d_-}\|\mathbf{Q}^+\|_F^2 + \|\mathbf{Q}^+\|_F \\ &\quad + \sqrt{d_-}|\lambda_1(\mathbf{Q})|\|\mathbf{Q}^+\|_F^2\Big](1+t)e^{-t|\lambda_{d-}(\mathbf{Q})|} \\ &\quad + 2\sqrt{d - d_-}\|\mathbf{Q}^+\|_F\Big)\|\mathbf{J}\|_F \end{aligned} \quad [35]$$

with $\|\mathbf{Q}^+\|_F^2 := \sum_{k=1}^{d_-} \frac{1}{\lambda_k^2(\mathbf{Q})}$, whereas

$$\begin{aligned} &\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_{\mathrm{op}} \\ &\leq \Big(\frac{|\lambda_1(\mathbf{Q})|(2 + |\lambda_{d_-}(\mathbf{Q})| + |\lambda_1(\mathbf{Q})|)}{\lambda_{d-}^2(\mathbf{Q})}(1+t)e^{-t|\lambda_{d_-}(\mathbf{Q})|} \\ &\quad + \frac{2}{|\lambda_{d_-}(\mathbf{Q})|}\Big)\|\mathbf{J}\|_{\mathrm{op}}. \end{aligned} \quad [36]$$

Under the further assumption that $\lambda_d(\mathbf{Q}) = 0$ and

$$\mu_1 d \leq |\lambda_{d-1}(\mathbf{Q})| \leq |\lambda_1(\mathbf{Q})| \leq \mu_2 d \quad [37]$$

for some $0 < \mu_1 \leq \mu_2$, we have

$$\begin{aligned} &\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_F \\ &\leq \Big(\frac{\mu_2}{\mu_1^2} \cdot \Big[2\sqrt{d} + \mu_1 d + \mu_2 d^2\Big](1+t)^{-t\mu_1 d} + \frac{2}{\mu_1\sqrt{d}}\Big)\|\mathbf{J}\|_F \end{aligned} \quad [38]$$

and that

$$\begin{aligned} &\|t(\mathbf{I} - \mathbf{Q}^+\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^+ + \nabla_{\mathbf{J}}e^{t\mathbf{Q}} - te^{t\mathbf{Q}}\mathbf{J}\|_{\mathrm{op}} \\ &\leq \Big(\frac{\mu_2}{\mu_1^2} \cdot \Big[\frac{2}{d} + \mu_1 + \mu_2\Big](1+t)e^{-t\mu_1 d} + \frac{2}{\mu_1 d}\Big)\|\mathbf{J}\|_{\mathrm{op}}. \end{aligned} \quad [39]$$

1.2. High-Dimensional Asymptotics via Random Matrix Theory. We turn to our probabilistic bounds on the singular values of randomly generated rate matrices, Theorem 3. Although interesting in its own right, this result leads to consequences for the bounds in Eqs. **23** and **24** when applied in Theorem 2. Before proceeding, we briefly introduce further mathematical preliminaries associated with the so-called sub-exponential random variables. To avoid confusion, note that the following definition uses the term in the same way as, e.g., ref. 31, but that other definitions that mean quite the opposite (i.e., heavier than exponential tails) appear in the literature (32).

Definition 1: A random variable X is called sub-exponential if there exists some constant $K > 0$ for which its tails satisfy

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t/K}, \quad \forall t \geq 0. \quad [40]$$

In this case, the sub-exponential norm of X is defined by

$$\|X\|_{\psi_1} = \inf\{s > 0 : \mathbb{E}e^{|X|/s} \leq 2\}. \quad [41]$$

The class of sub-exponential distributions is denoted by

$$L_{\psi_1} = \{F_X(dx) : \|X\|_{\psi_1} < \infty\}.$$

Remark 2: In fact, by Vershynin (31, Proposition 2.7.1, p. 33), condition Eq. **40** is equivalent to the existence of some $s_0 > 0$ such that $\mathbb{E}e^{|X|/s_0} \leq 2$, namely, $\|X\|_{\psi_1} < \infty$ in Eq. **41**. As a notable example, if $X \sim \exp(\lambda)$, $\lambda > 0$, then it is easy to see that $X \in L_{\psi_1}$, indeed.

We formulate our second major result as follows.

Theorem 3. *Let $F_X \in L_{\psi_1}$ be a distribution such that $X \geq 0$ a.s., $\mathbb{E}X = \mu > 0$ and $\mathrm{Var}X = \sigma^2 > 0$. We consider a sequence of*

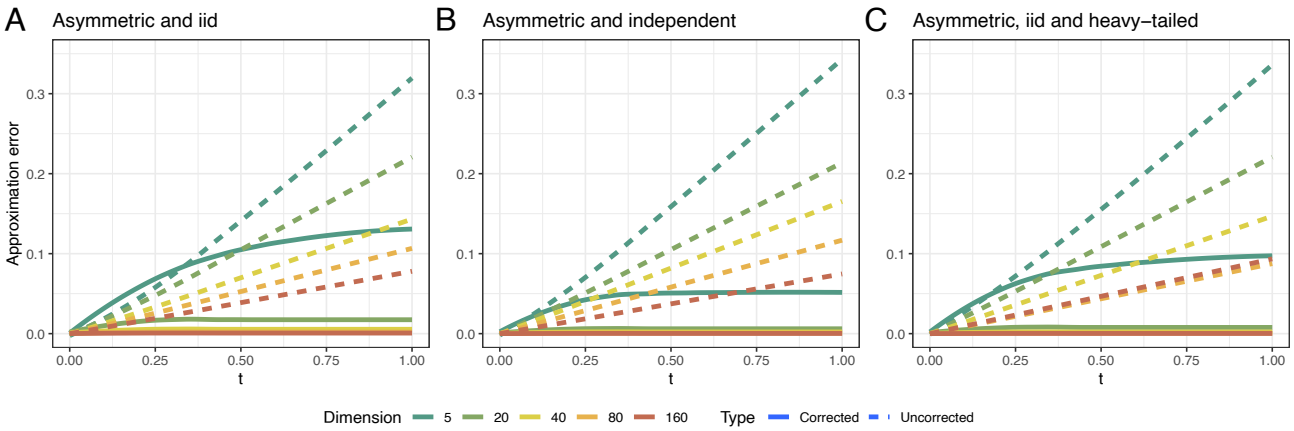

Fig. 1. Frobenius norm errors obtained by first-order approximation $te^{t\mathbf{Q}}\mathbf{J}$ and by affine-corrected first-order approximation $te^{t\mathbf{Q}}\mathbf{J} - t(\mathbf{I}-\mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}$ (Theorem 2) under increasingly relaxed assumptions. Within each assumption set, we average over 20 independent Monte Carlo simulations of random generator matrices for each dimension. Plot *A* corresponds to asymmetric generators with off-diagonal elements having independent and iid standard exponential random variables that correspond to the sub-exponential distribution hypothesis. Plot *B* drops the identical distribution assumption by allowing each row and column of the generator matrix to additively contribute its own mean—itself given by a standard exponential—to its corresponding exponentially distributed entries. Plot *C* features rate matrices with iid Cauchy entries truncated to be positive. Empirically, the results of Corollary 1 extend beyond the symmetric, iid, and sub-exponential hypotheses, suggesting scope of future work.

random matrices $\mathbf{Q} \equiv \mathbf{Q}(d) = \{q_{ij}\}_{i,j=1,\ldots,d} \in \mathcal{R}(d)$, $d \in \mathbb{N}$, *where either*

$$\begin{aligned} \{q_{ij}\}_{i,j=1,\ldots,d,\, i\neq j} &\overset{\text{iid}}{\sim} F_X \text{ and} \\ -q_{ii} &= \sum_{j\in\{1,\ldots,d\}\setminus\{i\}} q_{ij},\ i = 1,\ldots,d \end{aligned} \qquad [42]$$

or we impose that $\mathbf{Q} \in \mathcal{R}(d) \cap \mathcal{S}(d)$ *as*

$$\begin{aligned} \{q_{ij}\}_{i,j=1,\ldots,d,\, i>j} &\overset{\text{iid}}{\sim} F_X, q_{ij} := q_{ji} \text{ for } i < j \text{ and} \\ -q_{ii} &= \sum_{j\in\{1,\ldots,d\}\setminus\{i\}} q_{ij},\ i = 1,\ldots,d. \end{aligned} \qquad [43]$$

Then, in either of these cases, for any $d \in \mathbb{N}\setminus\{1\}$, *we have*

$$\begin{aligned} \mu + O_{\text{a.s.}}\left(\sqrt{\frac{\log d}{d}}\right) &\leq \frac{\sigma_2(\mathbf{Q})}{d} \\ &\leq \frac{\sigma_d(\mathbf{Q})}{d} \leq \mu + O_{\text{a.s.}}\left(\sqrt{\frac{\log d}{d}}\right) \end{aligned} \qquad [44]$$

almost surely.

Remark 3: The bounds constructed in Theorem 3 are strongly reminiscent of the bounds for eigenvalues provided in Theorem 1.5 of the seminal paper (28) (see also, for instance, Corollary 1.6 in ref. 26 and Corollary 1.1 in ref. 27).

Finally, let us observe that Theorems 2 and 3 as well as the fact that

$$\sigma_j(\mathbf{Q}) = |\lambda_{d-j+1}(\mathbf{Q})|, \text{ for e.g., any } \mathbf{Q} \in \mathcal{S}(d) \cap \mathcal{R}(d) \qquad [45]$$

combine to produce the following immediate corollary.

Corollary 1. *Consider any sequence of random matrices* $\mathbf{Q} \equiv \mathbf{Q}(d) = \{q_{ij}\}_{i,j=1,\ldots,d} \in \mathcal{R}(d)$, $d \in \mathbb{N}$, *as in Theorem 3 under the second (symmetric) case* Eq. **43**. *Then, taking* $\mu_1 = \mu_1(d)$ *and* $\mu_2 = \mu_2(d)$ *as the resulting lower and upper bounds defined by* Eq. **44**, *we have that* $\mathbf{Q}$ *satisfies both* Eqs. **38** *and* **39**, *cf.* Eq. **45** *relative to this sequence of* μ_1, μ_2 *for any* $\mathbf{J} \in \mathcal{M}(d)$.

Remark 4: Our rigorous formulation of Corollary 1 is limited to symmetric random rate matrices whose above diagonal elements are independent and iid draws from a sub-exponential distribution. However, strong numerical evidence suggests that the scope of the approximations Eqs. **38** and **39** reach far beyond the limitations of Corollary 1 in several different ways. Fig. 1 explores the consequences of relaxing various assumptions of Corollary 1. The third plot involves folded Cauchy random variables, the heavy tails of which violate the sub-exponential assumption (Definition 1). There appears to be no significant departure from the idea that the form $t(\mathbf{I} - \mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}$ corresponds increasingly well to the true approximation error [Eq. 7] as the dimension increases.

Remark 5: Calculation of $\widetilde{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}} := te^{t\mathbf{Q}}\mathbf{J}$ requires $\mathcal{O}(d^3)$ operations by, e.g., computing the spectral decomposition $\mathbf{Q}$ as in Eq. **30**. One may then recycle this decomposition to determine the additional term $t(\mathbf{I} - \mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+}$ for little extra cost. In view of Corollary 1 and Fig. 1,

$$\widehat{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}} := te^{t\mathbf{Q}}\mathbf{J} - t(\mathbf{I} - \mathbf{Q}^{+}\mathbf{Q})\mathbf{J}\mathbf{Q}\mathbf{Q}^{+} \qquad [46]$$

provides an accurate approximation of $\nabla_{\mathbf{J}} e^{t\mathbf{Q}}$ for asymptotically for large d. Thus, we anticipate further computational improvements when fitting large, gold-standard models using this refined approximation. That said, we leave the efficient and scalable application of $\widehat{\nabla}_{\mathbf{J}} e^{t\mathbf{Q}}$ to future work.

2. Empirical Studies

Before applying the naive matrix exponential derivative approximation to the phylogeographic analysis of SARS-CoV-2, we carry out a few targeted studies that illustrate the empirical performance of the approximate derivative and its affine correction Eq. **46** (Fig. 1); agreement between CTMC generator matrix empirical posterior distributions generated by surrogate-trajectory HMC algorithms using approximate derivatives and the truth (Fig. 2); and point estimation of generator matrix element values under different sparsity regimes and different CTMC state space dimensionalities d (Fig. 3).

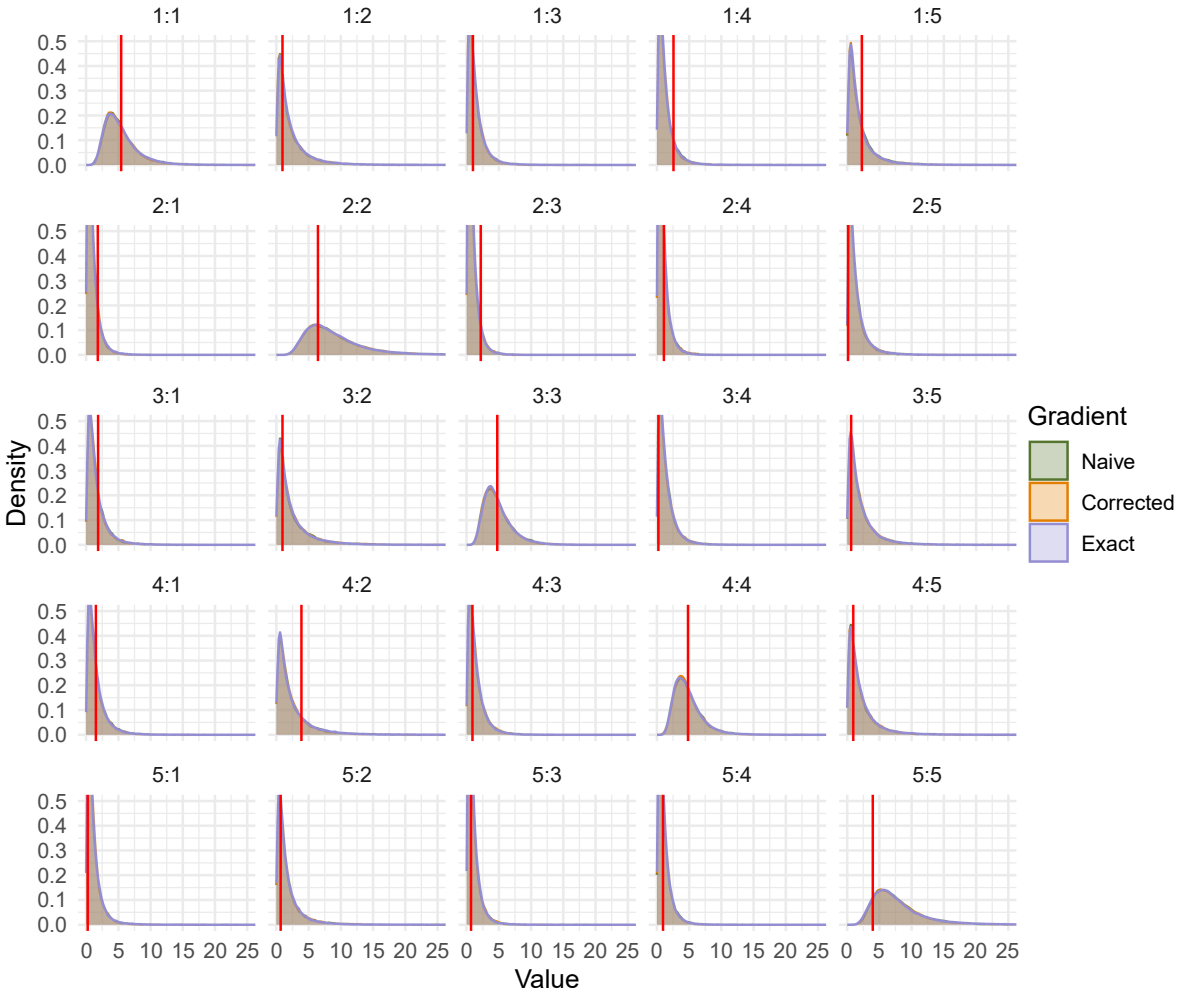

Fig. 2. Posterior density plots for surrogate-trajectory HMC using the naive approximate derivative, the corrected approximation Eq. **46** and the exact matrix exponential derivative for elements of a 5×5 generator matrix. The near-perfect overlap reflects the fact that each algorithm's transition kernel leaves the posterior distribution invariant (33). To generate data, we randomly draw standard normal generator entries once and simulate 20 independent initial/final position pairs from a CTMC with time interval $t = 1$. We show the true value in red and negate diagonal elements to simplify presentation.

Here, we fill in remaining simulation details not included in figure captions. In the simulations contributing to Fig. 1, we randomly generate new, independent direction matrices $\mathbf{J}$ at each time step within each of the 20 independent runs. The results are not sensitive to $\mathbf{J}$ in general. We also note that under Definition 1 the Cauchy distribution is not sub-exponential and therefore represents a deviation from core assumptions of Section 1. Using varying distributions on the elements of generator matrices $\mathbf{Q}$, the

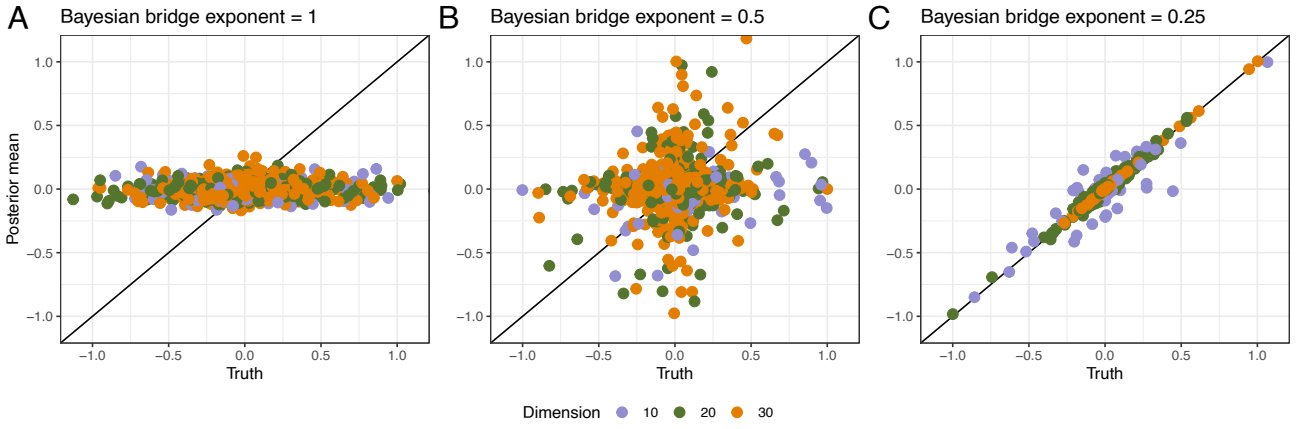

Fig. 3. Posterior means versus truth for CTMC generator matrix elements within differing sparsity regimes and with different dimensionalities d, holding observation count fixed at 300. We generate posteriors using surrogate-trajectory HMC with the naive matrix exponential derivative. To affect sparsity levels, we generate generator matrix entries according to the Bayesian bridge distribution (34) with different exponents ($\alpha \in \{1, 1/2, 1/4\}$ for plots *A*, *B* and *C*, respectively), normalizing by the largest absolute values to ease comparison. Smaller α values encode more peaked distributions with heavier tails and thus enforce greater sparsity. Plot *C* reflects the fact that the Bayesian bridge prior with exponent $\alpha = 1/4$ helps identify non-null parameters in small sample contexts (18). With this intuition in mind, we specify such a prior on generator random effects in Section 3.

simulations contributing to Figs. 2 and 3 both randomly generate N independent initial states $\mathbf{y}_1^0, \ldots, \mathbf{y}_N^0$ according to uniform distributions over their respective CTMC state spaces. For each of these initial states $\mathbf{y}_0$, we then simulate from the CTMC for time interval $t = 1$ to obtain samples $\mathbf{y}_1^1, \ldots, \mathbf{y}_N^1$. The likelihood then takes the form $\prod_n (\mathbf{y}_n^0)^T e^{\mathbf{Q}} \mathbf{y}_n^1$. For the simulation contributing to Fig. 2, we specify standard normal priors and sample from the posterior by generating 100,000 iterations from each algorithm. For the simulation contributing to Fig. 3, we generate 500,000 MCMC samples using surrogate-trajectory HMC with the naive matrix exponential derivative and calculate the posterior mean of each generator matrix element using the final 100,000 samples.

3. Application: Global Spread of SARS-CoV-2

Ref. 18 consider phylogenetic models that involve CTMC priors and show that the first-order approximation Eq. **4** of the matrix exponential derivative Eq. **2** performs remarkably well within surrogate HMC (*SI Appendix*), achieving an over 30-fold efficiency gain compared to random-walk Metropolis for a 14-state model with over 180 model parameters. Here, we demonstrate similar strong performance of the first-order approximation within surrogate HMC for a 44-state model with almost 1,900 model parameters.

3.1. Phylogenetic CTMC. Start with a (possibly unknown) rooted and bifurcating phylogenetic tree $\mathcal{T}$ consisting of N leaf nodes that correspond to observed biological specimens and $N-1$ internal nodes that correspond to unobserved ancestors. The tree also contains $2N-2$ branches of length t_v connecting each child node v to its parent u.

Given $\mathcal{T}$, we model the evolution of d characters along each branch of the tree according to a CTMC model with $d \times d$ generator matrix $\mathbf{Q}$ and stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d) = \lim_{t\to\infty} \widetilde{\boldsymbol{\pi}} e^{t\mathbf{Q}}$, for $\widetilde{\boldsymbol{\pi}}$ any arbitrary probability vector. Examples of characters are the $d = 4$ nucleotides within a set of aligned genome sequences (13) and the set of $d = 15$ geographic regions visited by an influenza subtype (2). We may scale t_v to be raw time (e.g., years) or the expected number of substitutions with respect to $\boldsymbol{\pi}$ depending on the given problem. In the former case, one may augment the model with a rate scalar γ that modulates the expected number of substitutions across all branches, and the finite-time transition probability matrix along branch v becomes $\mathbf{P}_v := e^{\gamma t_v \mathbf{Q}}$. In the following, we further posit $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ a vector of parameters.

Let data $\mathbf{Y}$ be the $d \times N$ matrix with columns $\mathbf{y}_n$, $n \in \{1, \ldots, N\}$, each having a single nonzero entry (set to 1) corresponding to the observed state of the biological specimen. One may use any node v to express the likelihood

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \mathbf{p}_v^T \mathbf{q}_v, \qquad [47]$$

where $\mathbf{p}_v$ and $\mathbf{q}_v$ are the post-order and pre-order partial likelihood vectors, respectively (35). The former describes the probability of the observed states for all observed specimens (i.e., leaf nodes) that descend from node v, conditioned on the state at node v. The latter describes analogous probabilities for all observed specimens not descending from node v. For leaf nodes, $\mathbf{p}_n := \mathbf{y}_n$, $n \in \{1, \ldots, N\}$, and one may specify the root node's pre-order partial likelihood to be any arbitrary probability vector a priori. Let "$\circ$" denote the Hadamard or elementwise product between matrices or vectors of equal dimensions. If we suppose that node u gives rise to two child nodes, v and w, then

$$\mathbf{p}_u = \mathbf{P}_v \mathbf{p}_v \circ \mathbf{P}_w \mathbf{p}_w, \quad \mathbf{q}_v = \mathbf{P}_v^T (\mathbf{q}_u \circ \mathbf{P}_w \mathbf{p}_w). \qquad [48]$$

Using the chain rule and the fact that $\mathbf{p}_v$ does not depend on $\mathbf{P}_v$, one may write the likelihood's derivative with respect to a single parameter θ_k thus:

$$\begin{aligned}\frac{\partial}{\partial \theta_k} p(\mathbf{Y}|\boldsymbol{\theta}) &\propto \sum_{v=1}^{2N-2} \operatorname{tr}\left(\frac{\partial (\mathbf{p}_v^T \mathbf{q}_v)}{\partial \mathbf{P}_v} \frac{\partial \mathbf{P}_v^T}{\partial \theta_k}\right) \\ &= \sum_{v=1}^{2N-2} \operatorname{tr}\left((\mathbf{q}_u \circ \mathbf{P}_w \mathbf{p}_w) \mathbf{p}_v^T \left(\frac{\partial e^{\gamma t_v \mathbf{Q}}}{\partial \theta_k}\right)^T\right) \\ &= \sum_{v=1}^{2N-2} \mathbf{p}_v^T \left(\sum_{i,j=1}^{d} \frac{\partial e^{\gamma t_v \mathbf{Q}}}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial \theta_k}\right)^T (\mathbf{q}_u \circ \mathbf{P}_w \mathbf{p}_w),\end{aligned} \qquad [49]$$

where we suppress the dependence of u and w on v.

Whereas the recursions of Eq. **48** facilitate fast likelihood computation, inferring $\boldsymbol{\theta}$ using gradient-based approaches such as HMC requires a large number of repeated evaluations of the matrix exponential derivative. These computations become particularly onerous when one opts for a gold-standard mixed-effects model (18) and specifies

$$\log q_{ij} = b_{ij} + \epsilon_{ij}, \quad i \neq j. \qquad [50]$$

Here, the fixed effects b_{ij} are elements of some non-random matrix $\mathbf{B}$, and the random effects ϵ_{ij} are mutually independent a priori and inferred as model parameters. The dimension K of $\boldsymbol{\theta}$ in this model is $\mathcal{O}(d^2)$, so the $\mathcal{O}(KNd^5)$ cost of the log-likelihood derivative Eq. **49** becomes a massive $\mathcal{O}(Nd^7)$. In this context, Magee et al. (18) shows that an approximate log-likelihood derivative based on the first-order approximation to the matrix exponential helps achieve a considerable speedup over the exact derivative, requiring only $\mathcal{O}(d^4 + Nd^3)$ floating-point operations yet facilitating high-quality proposals in the context of surrogate HMC.

In the following section, we use this method to analyze the global spread of SARS-CoV-2 and show that the first-order approximation maintains its performance for an even higher-dimensional problem than previously considered.

3.2. Bayesian Analysis of SARS-CoV-2 Contagion. We use the phylogenetic CTMC framework to model the early spread of SARS-CoV-2—the virus responsible for the ongoing COVID-19 pandemic—based on $N = 285$ observed viral samples collected from 31 regions worldwide between 24 December 2019 and 19 March 2020. These regions comprise 13 provinces within China and 18 countries without. Understanding the manner in which viruses travel between human populations is an object of ongoing study, and phylogeographic analyses point to the central role of travel networks including those measured by airline passenger counts (3) or Google mobility data (36). Here, we include three such predictors of travel in our CTMC model by expanding the fixed effects in regression model Eq. **50** to take the form

$$\mathbf{B} = \theta_1 \mathbf{X}_1 + \theta_2 \mathbf{X}_2 + \theta_3 \mathbf{X}_3 \qquad [51]$$

for $\mathbf{X}_1, \mathbf{X}_2$, and $\mathbf{X}_3$ fixed 44×44 matrix predictors and θ_1, θ_2, and θ_3 real-valued regression coefficients. Note we have expanded the

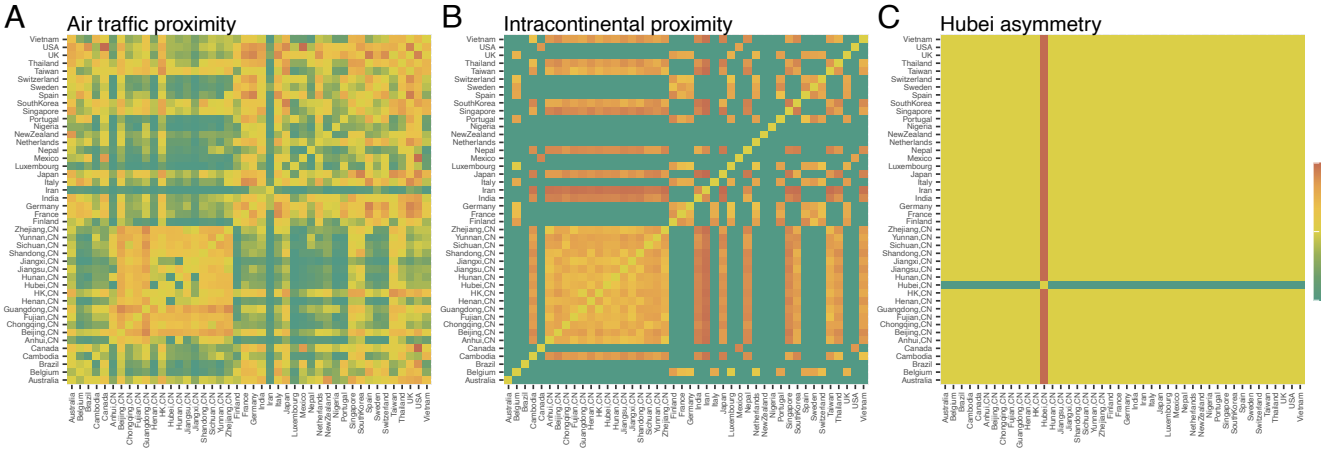

Fig. 4. Matrix predictors described in Eq. **51** combine in a linear manner to form the fixed-effect matrix **B** featured within the mixed-effects regression model Eq. **50**. Air traffic proximities (*A*) are proportional to the number of air passengers exchanged between airports within respective regions (3). Intracontinental proximities (*B*) take values between −1 for regions on different continents to 1 for adjacent regions. The Hubei asymmetry (*C*) roughly characterizes the Hubei quarantine of early 2020.

number of regions between which viruses may travel to $d = 44$ by including two additional Chinese provinces and 11 additional countries. Fig. 4 presents the three predictors of interest: $\mathbf{X}_1$ contains air travel proximities between locations as measured by annualized air passenger counts between airports contained within a region (3); $\mathbf{X}_2$ contains intracontinental proximities arising from physical distances when two regions inhabit the same continent and fixed at the minimum otherwise; finally, $\mathbf{X}_3$ describes the Hubei asymmetry, i.e., the nonexistence of human travel out of the Hubei province in early 2020.

In the context of a Bayesian analysis, we specify independent normal priors on θ_1, θ_2, and θ_3 with means of 0 and variances of 2. We also assume that the 1,892 random effects ϵ_{ij} follow sparsity-inducing Bayesian bridge priors with global scale parameter τ and exponent $\alpha = 0.25$. Here, we follow ref. 37 and specify a Gamma prior on $\tau^{-\alpha}$ with a shape parameter of 1 and a rate parameter of 2. Finally, we place a flat prior on the rate scalar γ. Inferring the posterior distribution of all 1,897 model parameters requires an advanced MCMC strategy. Namely, we adopt an HMC-within-Gibbs approach, updating the scalars γ and τ independently but updating all 1,895 regression parameters (both fixed and random effects) using surrogate HMC accomplished with the first-order approximation

$$\gamma t_v e^{\gamma t_v \mathbf{Q}} \mathbf{J}_{ij} \approx \frac{\partial e^{\gamma t_v \mathbf{Q}}}{\partial q_{ij}}$$

within Eq. **49**.

We generate 8 million MCMC samples in this manner, saving 1 in 10,000 Markov chain states, in order to guarantee a minimum ESS greater than 100. By far, the worst-mixing parameter is the global scale τ, which obtains an ESS of 185. The three fixed-effects regression coefficients θ_1, θ_2, and θ_3 obtain ESS's of 721, 721, and 644, respectively. The median and minimum ESS for the 1,892 random effects $\epsilon_{ss'}$ are 721 and 190, respectively. We note that, after thinning and removing burn-in, the sample only consists of 721 states, so ESS of 721 implies negligible autocorrelation between samples.

The *Left* plot of Fig. 5 presents posterior densities for the fixed-effect coefficients from regression Eq. **50**. The air traffic proximity, intracontinental proximity and Hubei asymmetry coefficients have posterior means and 95% credible intervals of

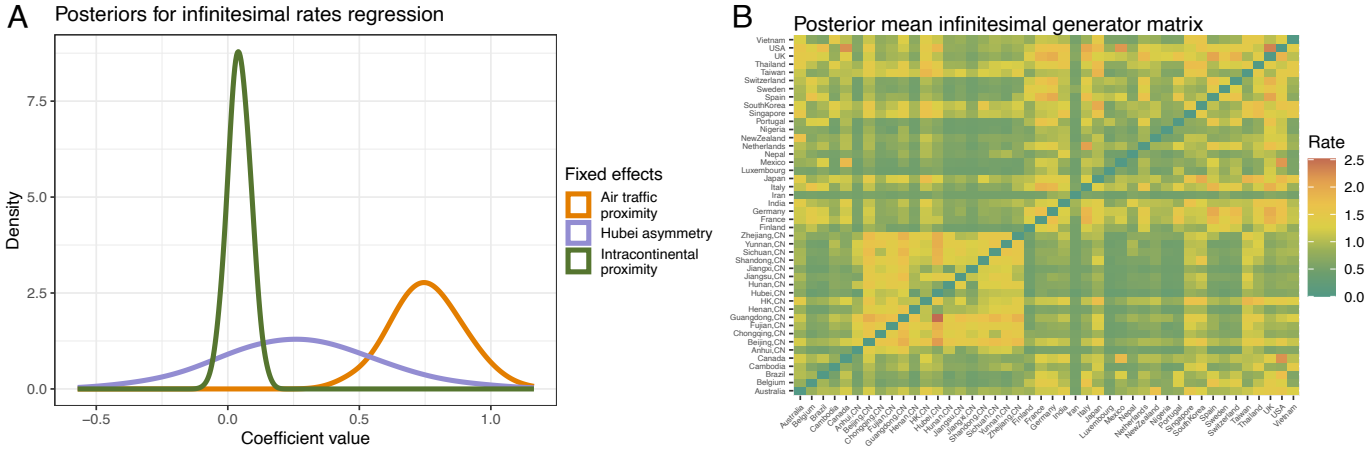

Fig. 5. Posterior inference. Posterior densities (*A*) for the three fixed-effect regression coefficients corresponding to the predictor matrices of Fig. 4 reflect largely positive associations between predictors and infinitesimal rate matrix, although air traffic proximity has the only statistically significant coefficient with posterior mean of 0.76 and 95% credible interval of (0.50, 1.03). The posterior mean (*B*) infinitesimal rate matrix closely resembles the air traffic predictor, while reflecting the Hubei asymmetry to a lesser extent.

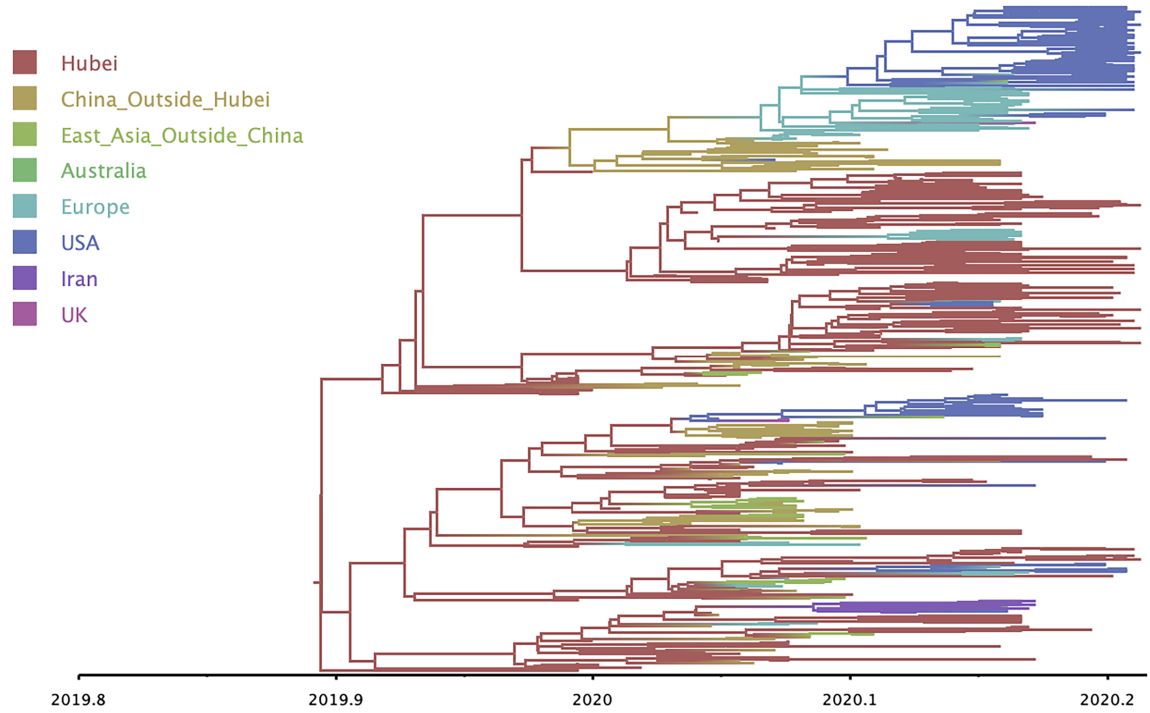

Fig. 6. Posterior predictive modes for (unobserved) ancestral locations color a phylogenetic tree that describes the shared evolutionary history of 285 SARS-CoV-2 samples.

0.76 (0.50, 1.03), 0.04 (−0.04, 0.12), and 0.27 (−0.39, 0.78), respectively. While these results suggest a positive association between each predictor matrix and the infinitesimal generator matrix $\mathbf{Q}$, the only predictor with a statistically significant association is air traffic proximity. This result agrees with a previous phylogeographic analysis of the global spread of influenza (3). The *Right* plot of Fig. 5 presents the posterior mean for generator $\mathbf{Q}$. As one may expect, the matrix looks similar to that of the air traffic predictor matrix in Fig. 4, but one may also see the influence of the Hubei asymmetry in, e.g., the squares corresponding to travel between Guangdong and Hubei provinces. Finally, we randomly generate regions of unobserved ancestors from their posterior predictive distributions every 100-thousandth MCMC iteration. After collapsing regions into 8 major blocks, Fig. 6 projects the empirical posterior predictive mode of these blocks onto the phylogenetic tree $\mathcal{T}$. The general pattern looks similar to that of figure 1 from ref. 38, although the geographic blocking scheme differs slightly.

In addition to these scientific questions of interest, we are interested in the performance of the first-order approximation as a surrogate gradient for HMC in such a high-dimensional setting. Whereas we know that the surrogate-trajectory HMC transition kernel leaves its target distribution invariant regardless of the approximation quality (33), transitions that rely on poor gradient approximations result in small acceptance rates, more random walk behavior and high autocorrelation between samples. Since ESS is inversely proportional to a Markov chain's asymptotic autocorrelation, larger ESS suggests a useful gradient approximation. To isolate the approximation's performance, we fix the Bayesian bridge global-scale τ at 2.5×10^{-5}. We generate a Markov chain with 80,000 states, saving every tenth state and removing the first 1,000 states as burn-in. Despite the relatively small number of iterations, we observe large ESS that suggest satisfactory accuracy of the first-order approximation within the context of high-dimensional surrogate HMC. The ESS for the three fixed-effect regression coefficients θ_1, θ_2, and θ_3 are 1,053, 1,343, and 498, respectively. The median and minimum ESS for the 1,892 random effects ϵ_{ij} are 1,514 and 1,161, respectively.

4. Discussion

We develop tight probabilistic bounds on the error associated with a simplistic approximation to the matrix exponential derivative for a large class of CTMC infinitesimal generator matrices with random entries. Our "blessing of dimensionality" result shows that this error improves for higher-dimensional matrices. We apply the numerically naive approach to the analysis of the global spread of SARS-CoV-2 using a mixed-effects model of unprecedented dimensions. The results obtained herein suggest the further study of CTMCs through the lens of random matrix theory. Furthermore, this analysis suggests a refinement of the first-order approximation to the matrix exponential derivative that may be particularly useful within modern, high-dimensional settings.

Data, Materials, and Software Availability. Public data have been deposited in Github (39).

ACKNOWLEDGMENTS. G.D. was partially supported by the Simons Foundation collaboration grant #714014. N.E.G.-H. received support for this work under NSF DMS 2108790. A.J.H. is supported by a gift from the Karen Toffler Charitable Trust and by grants NIH K25 AI153816, NSF DMS 2152774, and NSF DMS 2236854. A.F.M. and M.A.S. are partially supported through grants NIH R01AI153044 and R01AI162611. The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the Centers for Disease Control.

1. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
2. P. Lemey *et al*., Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* **10**, e1003932 (2014).
3. A. J. Holbrook *et al*., Massive parallelization boosts big Bayesian multidimensional scaling. *J. Comput. Graph. Stat.* **30**, 11–24 (2021).
4. A. J. Holbrook, X. Ji, M. A. Suchard, From viral evolution to spatial contagion: A biologically modulated Hawkes model. *Bioinformatics* **38**, 1846–1856 (2022).
5. J. R. Norris, *Continuous-Time Markov Chains I. Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press, 1997), pp. 60–107.
6. C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.* **20**, 801–836 (1978).
7. C. Moler, C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* **45**, 3–49 (2003).
8. J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mole. Evol.* **17**, 368–376 (1981).
9. J. S. Sinsheimer, J. A. Lake, R. J. Little, Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* **52**, 193–210 (1996).
10. Z. Yang, B. Rannala, Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**, 717–724 (1997).
11. B. Mau, M. A. Newton, B. Larget, Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55**, 1–12 (1999).
12. M. A. Suchard, R. E. Weiss, J. S. Sinsheimer, Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**, 1001–1013 (2001).
13. T. H. Jukes *et al*., Evolution of protein molecules. *Mamm. Protein Metabol.* **3**, 21–132 (1969).
14. M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
15. M. Hasegawa, H. Kishino, T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
16. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
17. W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
18. A. F. Magee *et al*., Random-effects substitution models for phylogenetics via scalable gradient approximations. arXiv [Preprint] (2023). http://arxiv.org/abs/2303.13642. Accessed 25 September 2023.
19. I. Najfeld, T. F. Havel, Derivatives of the matrix exponential and their computation. *Adv. Appl. Math.* **16**, 321–375 (1995).
20. Y. Takahashi, "Markov chains with random transition matrices" in *Kodai Mathematical Seminar Reports* (Department of Mathematics, Tokyo Institute of Technology, Tokyo, Japan, 1969), vol. 21, pp. 426–447.
21. Z. D. Bai, Methodologies in spectral analysis of large dimensional random matrices, a review. *Stat. Sinica* **9**, 611–677 (1999).
22. D. Chafaï, The Dirichlet Markov ensemble. *J. Multiv. Anal.* **101**, 555–567 (2010).
23. C. Bordenave, P. Caputo, D. Chafaï, Circular law theorem for random Markov matrices. *Probab. Theory Related Fields* **152**, 751–779 (2012).
24. A. Chatterjee, R. S. Hazra, Spectral properties for the Laplacian of a generalized Wigner matrix. *Random Matrices: Theory Appl.* **11**, 2250026 (2022).
25. G. Nakerst, S. Denisov, M. Haque, Random sparse generators of Markovian evolution and their spectral properties. *Phys. Rev. E* **108**, 014102 (2023).
26. W. Bryc, A. Dembo, T. Jiang, Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Ann. Probab.* **34**, 1–38 (2006).
27. X. Ding, T. Jiang, Spectral distributions of adjacency and Laplacian matrices of random graphs. *Ann. Appl. Probab.* **20**, 2086–2117 (2010).
28. C. Bordenave, P. Caputo, D. Chafaï, Spectrum of Markov generators on sparse random graphs. *Commun. Pure Appl. Math.* **67**, 621–669 (2014).
29. L. N. Trefethen, D. Bau, *Numerical Linear Algebra* (Siam, 2022), vol. 181.
30. R. A. Horn, C. R. Johnson, *Matrix Analysis* (Cambridge University Press, 2012).
31. R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge University Press, 2018), vol. 47.
32. C. M. Goldie, C. Klüppelberg, "Subexponential distributions" in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, R. Adler, R. Feldman, M. Taqqu, Eds. (Birkhäuser, Boston, MA, 1998), pp. 435–459.
33. N. E. Glatt-Holtz, J. A. Krometis, C. F. Mondaini, On the accept-reject mechanism for Metropolis-Hastings algorithms. *Ann. Appl. Probab.* **33**, 5279–5333 (2023).
34. N. G. Polson, J. G. Scott, J. Windle, The Bayesian bridge. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **76**, 713–733 (2014).
35. X. Ji *et al*., Gradients do grow on trees: A linear-time O(N)-dimensional gradient for statistical phylogenetics. *Mol. Biol. Evol.* **37**, 3047–3060 (2020).
36. M. Worobey *et al*., The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
37. A. Nishimura, M. A. Suchard, Shrinkage with shrunken shoulders: Gibbs sampling shrinkage model posteriors with guaranteed convergence rates. *Bayes. Anal.* **1**, 1–24 (2022).
38. P. Lemey *et al*., Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
39. A. J. Holbrook, ExpmDerivative: Source materials for "On the surprising effectiveness of a simple matrix exponential derivative approximation, with application to global SARS-CoV-2". GitHub. https://github.com/andrewjholbrook/expmDerivative. Deposited 1 March 2023.