Journal of Statistical Mechanics:
Theory and Experiment

PAPER

Global multivariate model learning from hierarchically correlated data

To cite this article: Edwin Rodríguez Horta *et al J. Stat. Mech.* (2021) 073501

View the article online for updates and enhancements.

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Biological modelling and information

Global multivariate model learning from hierarchically correlated data

Edwin Rodríguez Horta[1,2,*],
Alejandro Lage-Castellanos[2], **Martin Weigt**[1]
and Pierre Barrat-Charlaix[3,*]

[1] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratoire de Biologie Computationnelle et Quantitative-LCQB, Paris, France
[2] Group of Complex Systems and Statistical Physics, Department of Theoretical Physics, University of Havana, Havana, Cuba
[3] Biozentrum, Universität Basel, Basel, Switzerland
E-mail: edwin@fisica.uh.cu and pierre.barrat@unibas.ch

Received 2 March 2021
Accepted for publication 6 May 2021
Published 16 July 2021

Online at stacks.iop.org/JSTAT/2021/073501
https://doi.org/10.1088/1742-5468/ac06c2

J. Stat. Mech. (2021) 073501

Abstract. Inverse statistical physics aims at inferring models compatible with a set of empirical averages estimated from a high-dimensional dataset of independently distributed equilibrium configurations of a given system. However, in several applications, such as biology, data result from stochastic evolutionary processes, and configurations are related through a hierarchical structure, typically represented by a tree, and therefore not independent. In turn, empirical averages of observables superpose intrinsic signals related to the equilibrium distribution of the studied system, and spurious historical (or phylogenetic) signals resulting from the structure underlying the data-generating process. The naive application of inverse statistical physics techniques therefore leads to systematic biases and an effective reduction of the sample size. To advance on the currently open task of extracting intrinsic signals from correlated data, we study a system described by a multivariate Ornstein–Uhlenbeck process defined on a finite tree. Using a Bayesian framework, we can disentangle covariances in the data corresponding to their multivariate Gaussian equilibrium distribution from those resulting from the historical correlations. Our approach leads to a clear gain in accuracy

[*] Author to whom any correspondence should be addressed.

1742-5468/21/073501+30$33.00

in the inferred equilibrium distribution, which corresponds to an effective two to fourfold increase in sample size.

Keywords: co-evolution, inference of graphical models, phylogeny, statistical inference in biological systems

Contents

1. Introduction

With the emergence of large, high-dimensional datasets for complex systems across disciplines, methods of *inverse statistical physics* have seen rapidly growing interest during recent years [1]. In the most standard setting, the data provide observational samples of the 'microscopic' degrees of freedom of the system under study—this can be biological sequences [2, 3], firing patterns of neurons [4, 5], individuals in animal groups

https://doi.org/10.1088/1742-5468/ac06c2

[6, 7], stock markets [8, 9], etc. Within a static modeling approach, frequently based on the maximum-entropy principle [10], data $\vec{x}$ are assumed to be generated independently from some unknown probability distribution $P(\vec{x})$. This distribution describes the underlying interaction patterns between the observed degrees of freedom, and has to be learned from data to unveil the rules governing the system. In more rare cases where data correspond to observed time series, theoretical and algorithmic development is much less advanced than for independent static data [1].

One of the biggest application areas of inverse statistical mechanics is the modeling of biological processes. These applications are fueled by the large amount of available data resulting from the impressive progress in experimental techniques in biology. This is especially visible in the case of biological sequences, with databases now harboring a vast amount of high-quality DNA or protein sequences [11, 12]. A common idea in this context is that it is possible to use characteristics of genes or organisms related by a common ancestry—called *homologous*—to construct models of the selection acting on them. A successful example in this regard is the representation of protein sequences by probabilistic models in the so-called DCA method [2, 3]. The prototypical datasets in this context are multiple-sequence alignments (MSA), with lines being a so-called homologous, i.e. evolutionarily related sequences, and columns specific positions deriving from some common ancestral position [13]. The MSA contains at least two kinds of complementary information:

- *Phylogenetic information*: the distances between sequences carry information about the evolutionary time since their common ancestor. Using phylogeny inference methods [14, 15] we may reconstruct the evolutionary history of our dataset, represented by a phylogenetic tree.
- *Co-evolutionary information*: positions in a sequence typically do not evolve independently, but rather in a correlated way. This co-evolution carries important information about the selection forces acting on evolving entities. This fact has been extensively studied in the case of protein sequences, and used to predict structure, mutational landscapes or networks of interacting proteins [2, 3].

These two types of information are contained in two complementary features of the data: phylogenetic inference is based on the comparative analysis of different sequences, while co-evolutionary information is contained in the correlation of different columns of the MSA. Modeling approaches using one type of information typically neglect the other one: inference of phylogenies generally assumes that all positions in a sequence evolve independently, while co-evolutionary models of proteins assume that sequences in the MSA are independently distributed. This choice is motivated by the fact that taking the two types of correlations into account, i.e. through time with phylogeny and across trait values for co-evolution, results in very hard inference procedures, cf [16, 17]. However, this can lead to biases in the model parameters: it has for instance been shown that phylogenetic relations between protein sequences induce non-trivial correlations that are not related to protein function [18, 19].

In this work, we consider the case of the inverse problem for high-dimensional data showing hierarchical correlations due to a branching generating process. Our motivation for this purely methodological study comes from the modeling of protein sequences discussed above, but the underlying problem is much more general. Instead of sequences of discrete characters, like amino acids or nucleotides, we may consider continuous phenotypic traits. The branching process is not necessarily the phylogeny of species, but it may be the genealogy of populations of the same species, or other branching processes like epidemics spreading or geographic migration.

To address this problem, we use a simple and very general model for the temporal evolution of correlated variables: a historically well-known way to represent such processes is to use Ornstein–Uhlenbeck (OU) dynamics, which models configurations as Gaussian vectors evolving in a quadratic potential [20]. The use of OU processes on trees is a common way of representing biological evolution of continuous traits, such as gene expression levels or macroscopic traits of organisms [14, 21, 22]. In this context, the quadratic potential represents selection forces acting on the evolving entities. This formalism has led to the development of methods to infer the strength of selection from data that is correlated through a phylogenetic tree in the field of phylogenetic comparative methods [23, 24] and for the study of gene expression levels [25–27]. However, these methods are limited to scalar quantities, and can only be applied independently to each variable in the system of interest. More recent works have extended these methods to the case of many variables systems, allowing us to also consider correlations between these variables [28, 29]. These developments are based on the recursive use of Gaussian integration of the OU process along branches of the tree. Compared to [29], the method that we present in this article can be seen as an alternative way to approach the problem of inferring parameters of an OU process that takes place on a tree.

The use of an OU process to model evolution is *a priori* limited to continuous traits. However, it could potentially be used for protein sequences if combined with a continuous-variable approximation, which has successfully been used in the past [30–32]. In this context, the equilibrium distribution reached by the OU process represents the probability distribution given by the DCA method, which can be used to predict non-trivial structural contacts in the protein fold, effects of amino-acid mutations or even designing novel functional sequences [33–35].

In this work, we are interested in constructing an inference method for parameters of an OU process from data correlated through a tree. Our approach is purely methodological, and the data can represent any set of continuous phenotypic traits, e.g. from different organisms, with the tree indicating the phylogenetic relations between data points. Inferred parameters then represent the selection forces without biasing effects from the phylogeny. The manuscript is divided as follows: we first review in section 2 the main characteristics of the multivariate OU process. We then describe the setting of the inference problem that we want to solve in section 3 and propose a solution in section 4. Finally, we present results obtained on simulated data in section 5, with the context of pairwise models of protein sequences in mind.

2. The multivariate Ornstein-Uhlenbeck process

We consider a system characterized by L continuous degrees of freedom and whose state is fully described by an L-dimensional vector $\vec{x} \in \mathbb{R}^L$. These degrees of freedom can be continuous phenotypic traits of some living organism, or the sequence of a gene or a protein if a continuous approximation is made. At equilibrium, $\vec{x}$ is assumed to be normally distributed,

$$P_{\mathrm{eq}}(\vec{x}) = \frac{1}{Z(\boldsymbol{J})} \exp\left\{-\frac{1}{2}\vec{x}^{\mathrm{T}}\boldsymbol{J}\vec{x}\right\}, \tag{1}$$

where $\boldsymbol{J}$ is the symmetric, positive definite *coupling matrix* and $Z(\boldsymbol{J}) = \sqrt{(2\pi)^L/\det\,\boldsymbol{J}}$ is the normalization constant; the means of all components of $\vec{x}$ are set to zero without loss of generality. We are interested in inferring the coupling matrix from a given amount of observed states $\vec{x}$ of the system. If these observations were independent from each other, due to the simple Gaussian form of equation (1), $\boldsymbol{J}$ would simply be equal to the inverse of the empirical *covariance matrix* of the data, written $\boldsymbol{C} = \boldsymbol{J}^{-1}$.

However, we consider the case where observations are not independent. On the contrary, they result from a dynamical process taking place during a finite amount of time, and different data-points are therefore correlated to each other. This dynamical process is described below.

We suppose that the considered system evolves according to the following Langevin equation

$$\gamma^{-1}\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} = -\boldsymbol{J}\vec{x} + \vec{\xi}(t). \tag{2}$$

Here, $\vec{\xi}(t)$ is a vector of uncorrelated white noise, and γ^{-1} is the characteristic timescale governing the dynamics. In short, equation (2) states that the system described by $\vec{x}$ undergoes Brownian motion in a quadratic energy landscape characterized by the coupling matrix $\boldsymbol{J}$.

We are not interested in $\vec{x}$ directly, but rather in its probability distribution $P(\vec{x}|\,\vec{x}_0, \Delta t)$, i.e. in the probability to find the system in state $\vec{x}$ knowing it was in state $\vec{x}_0$ sometime Δt in the past. The Fokker–Planck equation corresponding to equation (2) is straightforward to write,

$$\gamma^{-1}\partial_t P = \left(-\sum_{a,b=1}^{L}\frac{\partial}{\partial x_a}J_{ab}x_b + \sum_{a=1}^{L}\frac{\partial^2}{\partial x_a^2}\right)P, \tag{3}$$

where the parenthesized expression on the right-hand side is understood as an operator acting on P. The solution to equation (3) is a multivariate normal distribution [36]:

$$P(\vec{x}|\vec{x}_0, \Delta t) = \left[(2\pi)^N \det\,\boldsymbol{\Sigma}\right]^{-1/2} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right\}, \tag{4}$$

where we introduce the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ as well as the vector $\vec{\mu}$ as

$$\boldsymbol{\Lambda} = \mathrm{e}^{-\gamma\boldsymbol{J}}, \qquad \vec{\mu} = \boldsymbol{\Lambda}^{\Delta t}\vec{x}_0, \qquad \boldsymbol{\Sigma} = \boldsymbol{J}^{-1}(\mathbb{1} - \boldsymbol{\Lambda}^{2\Delta t}). \tag{5}$$

Equations (4) and (5) define a multivariate OU process.

Note that since matrix $\boldsymbol{\Lambda}$ is an exponential of $\boldsymbol{J}$, it is symmetric, has strictly positive eigenvalues and commutes with $\boldsymbol{J}$. We also underline that $\boldsymbol{\Sigma}$ and $\vec{\mu}$ depend on

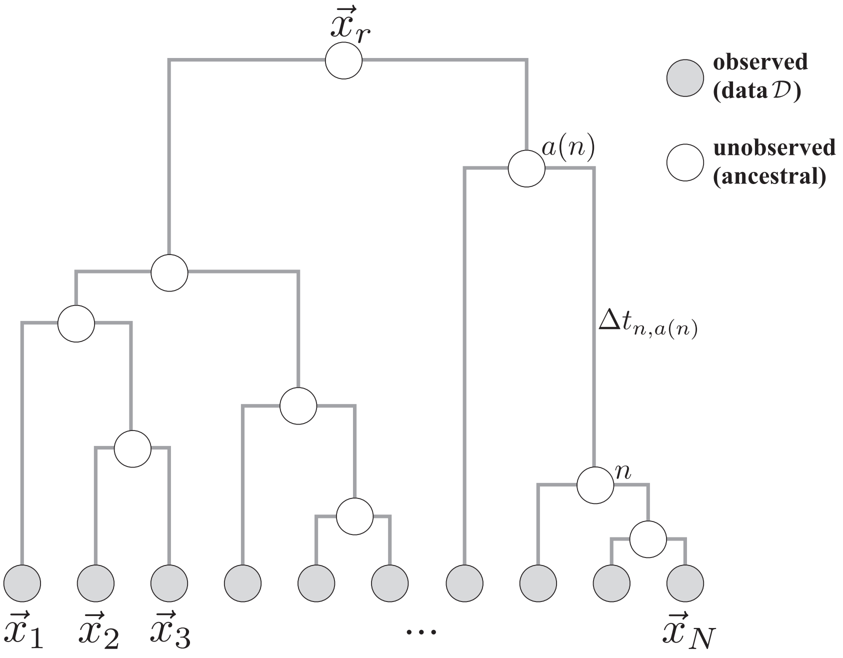

Figure 1. Schematic representation of a tree $\mathcal{T}$ underlying the data-generating process. The process starts at the root node r with a configuration $\vec{x}_r$ sampled from $P_{\text{eq}}(\vec{x}_r)$. The dynamics consist of independent realizations of the OU process on all branches from ancestral nodes $a(n)$ to child nodes n over times corresponding to the branch length $\Delta t_{n,a(n)}$, initialized in the ancestral configuration $\vec{x}_{a(n)}$. The observable data only consist of configurations of the leaf nodes (grey circles in the figure), while configurations of ancestral nodes remain unknown. There are no restrictions on the topology of tree $\mathcal{T}$ and the length of the branches.

Δt, although this dependence is not explicitly written in our notation to make it less heavy. By taking $\gamma\Delta t \gg 1$ and using the fact that $\boldsymbol{J}$ has strictly positive eigenvalues, we immediately recover equation (1), meaning that the OU process converges to the desired equilibrium distribution.

We can compute the joint distribution of two configurations $\vec{x}_1$ and $\vec{x}_2$ separated by a time Δt by multiplying equations (1) and (4),

$$\begin{aligned} P(\vec{x}_1, \vec{x}_2|\, \Delta t) &= P(\vec{x}_1|\vec{x}_2, \Delta t) \times P_{\text{eq}}(\vec{x}_2) \\ &\propto \exp\left\{ -\frac{1}{2}\left(\vec{x}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\vec{x}_1 + \vec{x}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\vec{x}_2 - 2\vec{x}_1^{\mathrm{T}}\boldsymbol{\Lambda}^{\Delta t}\boldsymbol{\Sigma}^{-1}\vec{x}_2 \right) \right\}. \end{aligned} \quad (6)$$

This equation illustrates the *time reversibility* of the OU process. Indeed, the distribution is symmetric in $\vec{x}_1$ or $\vec{x}_2$ and does not depend on which configuration came first.

Equation (6) allows for computing the joint covariance of the correlated equilibrium configurations $\vec{x}_1$ and $\vec{x}_2$. The probability distribution in equation (6) is normal with an inverse covariance matrix defined by blocks: $\boldsymbol{\Sigma}$ on the diagonal and $-\boldsymbol{\Lambda}^{\Delta t}\boldsymbol{\Sigma}$ off-diagonal. By inverting this block matrix, given that $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}$ commute and are invertible, we obtain the following covariance:

$$\langle \vec{x}_1 \vec{x}_2^{\mathrm{T}} \rangle_{\Delta t} = \boldsymbol{\Lambda}^{\Delta t} \boldsymbol{J}^{-1} = \boldsymbol{\Lambda}^{\Delta t} \boldsymbol{C}. \tag{7}$$

Equation (7) allows us to identify typical timescales associated to the OU process. Let us call ρ_a the eigenvalues of $\boldsymbol{J}$. The eigenvalues of $\boldsymbol{\Lambda}^{\Delta t}\boldsymbol{C}$ are then equal to $\rho_a^{-1}\mathrm{e}^{-\gamma\rho_a\Delta t}$. Since all ρ_a are positive, the eigenvalues of $\boldsymbol{\Lambda}^{\Delta t}\boldsymbol{C}$ vanish exponentially over time over timescales of the form $\rho_a\Delta t$. The slowest timescale of exponential decay is set by $\tau_c^{-1} = \gamma\rho_{\min}$, with $\rho_{\min}$ being the smallest eigenvalue of $\boldsymbol{J}$. Thus, for $\Delta t/\tau_c \gg 1$, $\vec{x}_1$ and $\vec{x}_2$ are uncorrelated, whereas they are highly correlated for $\Delta t/\tau_c \ll 1$. It should be noted that for $\Delta t = 0$, the joint correlation matrix of $\vec{x}_1$ and $\vec{x}_2$ becomes non invertible, and equation (7) becomes irrelevant. Actually, $\vec{x}_1$ and $\vec{x}_2$ coincide at that point, i.e. we have $P(\vec{x}_1, \vec{x}_2|\Delta t = 0) = P_{\mathrm{eq}}(\vec{x}_1) \times \delta(\vec{x}_1 - \vec{x}_2)$ using the L-dimensional Dirac distribution.

3. Statement of the problem

The problem discussed here is the inference of the probability distribution describing samples that are hierarchically correlated by a tree, cf figure 1. Formally, we assume that the data consist of N real-valued vectors of length L, denoted $\{\vec{x}_i\} \in \mathbb{R}^L$ with $i = 1, \ldots, N$. Taken individually, we assume that the $\vec{x}_i$ are distributed according to equation (1), i.e. according to a multivariate Gaussian of zero mean and covariance $\boldsymbol{C}$. By construction, the equilibrium covariance between any pair of elements of a given vector $\vec{x} = (x^1, \ldots, x^L)$ is given by the inverse of the coupling matrix: $\langle x^a x^b\rangle - \langle x^a\rangle\langle x^b\rangle = \boldsymbol{C}_{ab} = (\boldsymbol{J}^{-1})_{ab}$ for all $a, b \in \{1, \ldots, L\}$. This implies that inferring the coupling matrix defining the probability distribution amounts to finding the *equilibrium* covariance matrix $\mathbf{C}$.

However, this covariance cannot be directly measured as we consider observations that are not independently distributed. Instead, the set of measured configurations $\{\vec{x}_i\}_{i=1,\ldots,N}$ is the result of an OU process taking place on a tree $\mathcal{T}$, as is illustrated in figure 1:

- The process starts at the root node r with a state vector $\vec{x}_r$ drawn from the equilibrium distribution P_{eq}.
- On each branch $(n, a(n))$ of length $\Delta t_{n,a(n)}$ connecting node n with its ancestral node $a(n)$, the dynamics follow equation (2), starting from initial condition $\vec{x}_{a(n)}$, and running for time $\Delta t_{n,a(n)}$. In other words, given the state $\vec{x}_{a(n)}$ of the ancestral node, $\vec{x}_n$ is sampled from $P(\vec{x}_n|\vec{x}_{a(n)}, \Delta t_{n,a(n)})$; see equation (4).
- As a consequence, OU processes on branches stemming from a common ancestral node evolve independently, but from an identical initial condition.
- Observed data vectors correspond to the states of the leaves of the tree at the end of this process. The states of the internal nodes are not part of the observed data and remain unknown.

This process is thought to represent the evolution of biological traits along a phylogenetic tree, with the leaf nodes corresponding to traits observed in today's species.

Note that due to the reversible nature of our OU process, the joint probability of any pair of leaf configurations $\vec{x}_i$ and $\vec{x}_j$, with $i, j \in \{1, \dots, N\}$, is given by $P(\vec{x}_i, \vec{x}_j | \Delta t_{ij})$ (equation (6)), with Δt_{ij} denoting the total branch length of the path connecting i and j in the tree.

The OU process is characterized by the quadratic potential $\boldsymbol{J} = \boldsymbol{C}^{-1}$ and the rate γ. Hence, the joint statistics of the leaf configurations $\{\vec{x}_i\}_{i=1,\dots,N}$ (i.e. the data) are fully determined by $\boldsymbol{C}$, γ, and the tree $\mathcal{T}$. The aim of this work is to derive a method for inferring the most likely values of $\boldsymbol{C}$ and γ given the knowledge of the data $\mathcal{D} = \{\vec{x}_i\}_{i=1,\dots,N}$ and the underlying tree $\mathcal{T}$. We consider here that both the topology and the branch lengths of $\mathcal{T}$ are known.

This problem shows two notable extreme cases: the first one is the case where the typical branch length of the tree is short compared to the timescales introduced in section 2. As a consequence, leaf configurations are close to identical to the root, i.e. $\vec{x}_i \simeq \vec{x}_r$, and the inference of $\boldsymbol{C}$ becomes impossible due to lack of independent samples from P_{eq}. The second one is the opposite case where the typical branch length of the tree is long compared to the longest timescale of the OU process τ_c. In this case, the configuration of a child node is close to independent from that of its ancestor, and leaf configurations can be considered as independent samples from the equilibrium distribution P_{eq}. $\boldsymbol{C}$ can then be readily estimated by computing the empirical covariance matrix. We are interested here in the intermediate regime where substantial tree-mediated correlations between data make it impossible to simply estimate $\boldsymbol{C}$ with the empirical covariance, but the depth of the tree introduces enough variability in the data for us to hope to reconstruct the energy potential $\boldsymbol{J}$.

We adopt a Bayesian inference approach by writing the probability of a given set of parameters $\{\boldsymbol{C}, \gamma\}$ given the data $\{\mathcal{D}, \mathcal{T}\}$ using Bayes' equation

$$P(\boldsymbol{C}, \gamma | \mathcal{D}, \mathcal{T}) \propto P(\mathcal{D} | \boldsymbol{C}, \gamma, \mathcal{T}) \cdot P(\boldsymbol{C}, \gamma), \tag{8}$$

with the proportionality constant not depending on the parameters $\{\boldsymbol{C}, \gamma\}$. Here, $P(\boldsymbol{C}, \gamma)$ can be any arbitrarily chosen prior distribution. The difficulty in equation (8) lies in the estimation of the likelihood $P(\mathcal{D} | \boldsymbol{C}, \gamma, \mathcal{T})$, i.e. of the joint probability of the datapoints $\mathcal{D} = \{\vec{x}_i\}_{i=1,\dots,N}$ for an OU process given by its parameters $\{\boldsymbol{C}, \gamma\}$ and the tree $\mathcal{T}$. We detail the computation of this probability in the following section.

4. Methods

4.1. Calculation of the likelihood

The joint distribution of two configurations $\vec{x}_1$ and $\vec{x}_2$ separated by time Δt is given by equation (6) and corresponds to a joint normal distribution. This means that the vector $\vec{X} = [\vec{x}_1, \vec{x}_2]$, i.e. the concatenation of vectors $\vec{x}_1$ and $\vec{x}_2$, follows a normal distribution with zero mean and variance described above in equation (5). Of importance here is that this property of the OU process can be extended to the joint distribution of any subset of nodes in a tree. In other words, if we now define $\vec{X} = [\vec{x}_1, \dots, \vec{x}_N]$ to be the concatenation of all configurations in our dataset $\mathcal{D}$, we can write the distribution of $\vec{X}$

as

$$P(\vec{X}|\, \boldsymbol{C}, \gamma, \mathcal{T}) = \left((2\pi)^{LN} \det \mathbb{G}\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\vec{X}^{\mathrm{T}}\mathbb{G}^{-1}\vec{X}\right\}, \tag{9}$$

where $\mathbb{G}$ is the *joint covariance matrix* and depends on the tree as well as on $\boldsymbol{C}$ and γ.

The joint covariance matrix is a matrix of dimension $(L\cdot N)\times(L\cdot N)$, built by $N\times N$ blocks of size $L\times L$ with entries

$$\mathbb{G}_{ij}(a,b) = \left\langle x_i^a x_j^b\right\rangle - \left\langle x_i^a\right\rangle\left\langle x_j^b\right\rangle, \quad i,j\in\{1,\dots,N\}\,; \quad a,b\in\{1,\dots,L\}, \tag{10}$$

where the (zero) marginals $\langle x_i^a\rangle$ and $\langle x_j^b\rangle$ are explicitly written for clarity. Each block $\mathbb{G}_{ij}$ is describing the connected correlations between two data vectors $\vec{x}_i$ and $\vec{x}_j$, which are separated by time Δt_{ij}, resulting as the sum of all branch lengths of the path connecting i and j on tree $\mathcal{T}$. Because the OU process is time-reversible, we can directly apply equation (7) and give all blocks of $\mathbb{G}$ in closed form,

$$\mathbb{G}_{ij} = \begin{cases} \boldsymbol{C} & \text{if } i=j \\ \boldsymbol{\Lambda}^{\Delta t_{ij}}\boldsymbol{C} & \text{otherwise,}\end{cases} \tag{11}$$

using the (currently unknown) covariance matrix $\boldsymbol{C}$ of a single equilibrium vector $\vec{x}$. We remind here that $\boldsymbol{\Lambda} = \mathrm{e}^{-\gamma\boldsymbol{C}^{-1}}$ depends only on γ and $\boldsymbol{C}$, and commutes with $\boldsymbol{C}$. As a direct consequence, all blocks $\mathbb{G}_{ij}$ commute with each other and with $\boldsymbol{C}$.

Equation (9) allows us to compute the log-likelihood of the data $\vec{X}$ as a function of $\vec{X}$ itself and of the joint covariance matrix. Indeed, taking its logarithm immediately gives

$$\mathcal{L}_{\mathcal{D}}(\mathbb{G}) = -\frac{1}{2}\log\det\mathbb{G} - \frac{1}{2}\vec{X}^{\mathrm{T}}\mathbb{G}^{-1}\vec{X} + \text{const.}\ , \tag{12}$$

but this expression is impractical for any numerical evaluation due to the large dimension of $\mathbb{G}$. However, the particular block structure of $\mathbb{G}$ described in equation (11) allows us to simplify the expression. To do so, we first introduce the eigenvalues and eigenvectors $\{\rho_a, \vec{s}_a\}$ of $\boldsymbol{C}^{-1}$, where the index a runs from 1 to L and vectors $\vec{s}_a$ are of dimension L. By definition, we have $\rho_a > 0$ for all a. Now using equation (11), we immediately see that the vectors $\vec{s}_a$ are also eigenvectors of the individual blocks $\mathbb{G}_{ij}$ with eigenvalues $z(\rho_a, \Delta t_{ij})$ where we introduced

$$z(\rho_a, \Delta t_{ij}) = \rho_a^{-1}\mathrm{e}^{-\gamma\rho_a\Delta t_{ij}}\ . \tag{13}$$

By convention, $\Delta t_{ii} = 0$ and the diagonal blocks are thus included via $z(\rho_a, \Delta t_{ii}) = \rho_a^{-1}$.

As the next step, we introduce $N\times N$-dimensional matrices $\boldsymbol{G}^a, a = 1,\dots,L$, with elements

$$\boldsymbol{G}_{ij}^a = z(\rho_a, \Delta t_{ij})\ , \quad 1\leqslant i,j\leqslant N\ . \tag{14}$$

In other words, for a given index $1\leqslant a\leqslant L$, $\boldsymbol{G}^a$ is the matrix built by replacing all blocks of $\mathbb{G}$ by their respective ath eigenvalue. Matrices $\boldsymbol{G}^a$ are symmetric and have their own eigenmodes, that we denote by $\{\lambda_{ka}, \vec{u}_{ka}\}_{k=1,\dots,N}$.

To obtain the eigenmodes of the joint covariance matrix $\mathbb{G}$ as a function of the $\vec{s}_a$ and $\vec{u}_{ka}$, we construct the direct product of vectors $\vec{u}_{ka}$ and $\vec{s}_a$, defining vectors $\vec{S}_{ka}$ of dimension $L \cdot N$:

$$\begin{aligned}\vec{S}_{ka} &= \vec{u}_{ka} \otimes \vec{s}_a \\ &= [u^1_{ka} \cdot \vec{s}_a, \ldots, u^N_{ka} \cdot \vec{s}_a],\end{aligned} \tag{15}$$

where the superindex $i \in \{1, \ldots, N\}$ refers to components of $\vec{u}_{ka}$. The ith block vector of $\vec{S}_{ka}$ will thus be written as $\vec{S}^i_{ka} = u^i_{ka} \cdot \vec{s}_a$. We can now show that $\vec{S}_{ka}$ are eigenvectors of matrix $\mathbb{G}$ by considering the ith block vector of the product $\mathbb{G} \cdot \vec{S}_{ka}$:

$$\begin{aligned}\left(\mathbb{G} \cdot \vec{S}_{ka}\right)^i &= \sum_{j=1}^{N} \mathbb{G}_{ij} u^j_{ka} \cdot \vec{s}_a \\ &= \sum_{j=1}^{N} z(\rho_a, \Delta t_{ij}) u^j_{ka} \cdot \vec{s}_a \\ &= (\mathbf{G}^a \cdot \vec{u}_{ka})^i \cdot \vec{s}_a \\ &= \lambda_{ka}(u^i_{ka} \cdot \vec{s}_a) \\ &= \lambda_{ka} \vec{S}^i_{ka}.\end{aligned} \tag{16}$$

We have first used the fact that $\vec{s}_a$ is an eigenvector of $\mathbb{G}_{ij}$, then the definition of $\boldsymbol{G}^a$, and finally the fact that $\vec{u}_{ka}$ is an eigenvector of $\boldsymbol{G}^a$. This demonstrates that the eigenmodes of $\mathbb{G}$ are $\left\{\lambda_{ka}, \vec{S}_{ka}\right\}$ with $1 \leqslant k \leqslant N$ and $1 \leqslant a \leqslant L$. Since $\mathbb{G}$ is the covariance matrix of a Gaussian distribution, we conclude the λ_{ka} to be strictly positive. Interestingly, the definition of $\vec{S}_{ka}$ as a direct product between eigenvectors $\vec{s}_a$ of the energy potential and eigenvectors $\vec{u}_{ka}$ reflecting the correlation structure mediated by the tree illustrates how these two types of information are entangled in the covariance matrix of the data.

Note that this decomposition of the eigenvectors leads to a drastic decrease in computational complexity for diagonalizing $\mathbb{G}$ (at given $\boldsymbol{C}$, γ and $\mathcal{T}$), and in consequence also for calculating the likelihood according to equation (12), which depends on the inverse covariance matrix $\mathbb{G}^{-1}$. Matrix $\mathbb{G}$ has linear dimension LN, so the numerical diagonalization or inversion takes time $\mathcal{O}((LN)^3)$. This is hardly achievable for systems of realistic length L of the state vector, and sufficient number N of data points for model learning. Following the above description, we need to first diagonalize $\boldsymbol{C}^{-1}$ (or equivalently $\boldsymbol{C}$), which requires time of $\mathcal{O}(L^3)$, followed by inversion of the L matrices $\boldsymbol{G}^a$, each one having linear dimension N. The total time complexity therefore results in $\mathcal{O}(L^3) + \mathcal{O}(L \cdot N^3)$, and the calculation can be easily achieved even on a standard PC. This observation is essential for inference, since we need to redo this calculation for many realizations of $\boldsymbol{C}$ and γ, in order to find the ones maximizing the likelihood given the data $\mathcal{D}$ and the tree $\mathcal{T}$. As is shown in appendix A.4, this calculation simplifies even more when considering a symmetric and homogeneous tree. In this case, the matrices $\boldsymbol{G}^a$ commute and can be diagonalized simultaneously and analytically for any value of ρ^a.

For the case of arbitrary trees, equation (12) can now be rewritten using the eigendecomposition of $\mathbb{G}$:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\mathbb{G}) &= -\frac{1}{2}\sum_{k=1}^{N}\sum_{a=1}^{L} \log\, \lambda_{ka} - \frac{1}{2}\sum_{k=1}^{N}\sum_{a=1}^{L} \lambda_{ka}^{-1}(\vec{X}\cdot\vec{S}_{ka})^2 \\ &= -\frac{1}{2}\sum_{k,a}\left(\log\, \lambda_{ka} + \lambda_{ka}^{-1}\left(\sum_{i=1}^{N} u_{ka}^{i}\vec{x}_i\cdot\vec{s}_a\right)^2\right).\end{aligned} \tag{17}$$

Equation (17) expresses the likelihood as a function of $\vec{u}_{ka}$, λ_{ka} (resulting from the tree $\mathcal{T}$ and given ρ^a) and $\vec{s}_a$ (resulting from $\boldsymbol{C}$). However, the definition of $\boldsymbol{G}^a$ in equation (14) makes it clear that its eigenmodes $\{\lambda_{ka}, \vec{u}_{ka}\}$ depend only on the eigenvalues ρ_a of $\boldsymbol{C}^{-1}$, on γ, as well as on the structure of the tree through the quantities Δt_{ij}, although this dependence cannot be analytically expressed in a simple manner. This means that the likelihood in equation (17) is in fact a function of $\{\rho_a, \vec{s}_a\}$, i.e. the eigenmodes of $\boldsymbol{C}^{-1}$, of the timescale parameter γ and of the pairwise distances on the tree Δt_{ij}.

4.2. Maximizing the likelihood

As stated at the beginning of this section, our main task is to find the equilibrium covariance matrix $\boldsymbol{C}$ that maximizes the likelihood of the data. We also need to find the optimal timescale γ. In equation (17), the likelihood is expressed as a function of γ and $\{\rho_a, \vec{s}_a\}$, i.e. the eigenvalues and eigenvectors of $\boldsymbol{C}^{-1}$, either directly or through the quantities $\{\lambda_{ka}, \vec{u}_{ka}\}$. We now attempt to maximize the likelihood with respect to the eigenmodes $\{\rho_a, \vec{s}_a\}$ and to the timescale γ.

In order to perform this optimization, we need to compute the gradient of the likelihood with respect to the eigenvectors $\{\vec{s}_a\}$. Since $\boldsymbol{C}^{-1}$ is a symmetric matrix, its eigenvectors form an orthogonal basis of the vector-space of dimension L and their components cannot be changed independently. One possible parametrization for the $\{\vec{s}_a\}$ consists of using $L(L-1)/2$ scalar *Eulerian angles* $\{\theta_{pq}\}$ with $1 \leqslant p < q \leqslant L$ [37, 38]. With the L eigenvalues ρ_a, this results in $L(L+1)/2$ independent values that fully parametrize the $L(L+1)/2$ values of $\boldsymbol{C}^{-1}$. A second possibility, which we have found faster in practice, is to express the matrix of the $\{\vec{s}_a\}$ as the exponential of a skew-symmetric matrix with $L(L-1)/2$ independent values; see appendix A.3. However, this parametrization does not allow a simple analytical expression of the gradient of the likelihood, and we use it along with automatic differentiation [39]. For this reason, we use the Eulerian angles below to express the gradient of the likelihood.

As a first step, we need to compute the gradient of the likelihood $\mathcal{L}_{\mathcal{D}}(\mathbb{G})$ with respect to all parameters $\{\rho_a, \theta_{pq}\}$ and γ. To make explicit the dependences of eigenvalues and eigenvectors of the matrices $\boldsymbol{G}^a$ on these parameters, we introduce the notation $\vec{u}_k(\rho_a, \gamma) = \vec{u}_{ka}$ and $\lambda_k(\rho_a, \gamma) = \lambda_{ka}$. Note that from the definition of $\boldsymbol{G}^a$ in equation (14), its eigenvalues and vectors depend only on the eigenvalues of $\boldsymbol{C}^{-1}$ and not on its eigenvectors. In the same way, we will now write $\boldsymbol{G}(\rho_a, \gamma)$ instead of $\boldsymbol{G}^a$.

The gradient of the likelihood is obtained by differentiating equation (17) with respect to the parameters of interest. This gives us three equations:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \rho_a} = -\frac{1}{2}\sum_{k=1}^{N}\Bigg\{&\frac{\partial \lambda_k}{\partial \rho_a}\lambda_k^{-1} - \frac{\partial \lambda_k}{\partial \rho_a}\lambda_k^{-2}\left(\sum_{i=1}^{N} u_k^i \vec{x}_i\cdot\vec{s}_a\right)^2\\ &+2\lambda_k^{-1}\left(\sum_{i=1}^{N} u_k^i \vec{x}_i\cdot\vec{s}_a\right)\left(\sum_{i=1}^{N}\frac{\partial u_k^i}{\partial \rho_a}\vec{x}_i\cdot\vec{s}_a\right)\Bigg\},\end{aligned} \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{pq}} = -\sum_{k=1}^{N}\lambda_k^{-1}\left(\sum_{i=1}^{N} u_k^i \vec{x}_i\cdot\vec{s}_a\right)\left(\sum_{i=1}^{N} u_k^i \vec{x}_i\cdot\frac{\partial \vec{s}_a}{\partial \theta_{pq}}\right), \tag{19}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \gamma} = -\frac{1}{2}\sum_{k=1}^{N}\Bigg\{&\frac{\partial \lambda_k}{\partial \gamma}\lambda_k^{-1} - \frac{\partial \lambda_k}{\partial \gamma}\lambda_k^{-2}\left(\sum_{i=1}^{N} u_k^i \vec{x}_i\cdot\vec{s}_a\right)^2\\ &+2\lambda_k^{-1}\left(\sum_{i=1}^{N} u_k^i \vec{x}_i\cdot\vec{s}_a\right)\left(\sum_{i=1}^{N}\frac{\partial u_k^i}{\partial \gamma}\vec{x}_i\cdot\vec{s}_a\right)\Bigg\}.\end{aligned} \tag{20}$$

The derivatives of $\vec{u}_k(\rho,\gamma)$ and $\lambda_k(\rho,\gamma)$ with respect to ρ can then be computed using the following equations [40]:

$$\frac{\partial \lambda_i(\rho,\gamma)}{\partial \rho} = \vec{u}_k(\rho,\gamma)^{\mathrm{T}}\frac{\partial \boldsymbol{G}(\rho,\gamma)}{\partial \rho}\vec{u}_k(\rho,\gamma) \tag{21}$$

and

$$\frac{\partial \vec{u}_k(\rho,\gamma)}{\partial \rho} = \sum_{l\neq k}\left(\vec{u}_k(\rho,\gamma)^{\mathrm{T}}\frac{\partial \boldsymbol{G}(\rho,\gamma)}{\partial \rho}\vec{u}_l(\rho,\gamma)\right)(\lambda_k(\rho,\gamma)-\lambda_l(\rho,\gamma))^{-1}\vec{u}_l(\rho,\gamma). \tag{22}$$

Equivalent equations can be written for their derivatives with respect to γ.

The computation of the gradient of $\mathcal{L}$ for a given set of parameters $\{\rho_a,\theta_{pq}\}$ then goes as follows. For each eigenvalue ρ_a, we compute and diagonalize matrix $\boldsymbol{G}(\rho_a)$ to obtain its eigenmodes $\vec{u}_k(\rho_a)$ and $\lambda_k(\rho_a)$. Using equations (21) and (22) and their equivalent form for γ, we also numerically compute their derivatives with respect to ρ_a and γ. This gives us all the quantities to estimate the gradient of $\mathcal{L}$ with respect to ρ_a using equation (18).

The optimization is performed by a quasi-Newton method (QNM) [41]. Details are presented in appendix A.5.

5. Results

In order to evaluate the accuracy of our inference procedure, we estimate how close the covariance matrix $\boldsymbol{C}_{\max}$ reconstructed by likelihood maximization is to the real

https://doi.org/10.1088/1742-5468/ac06c2

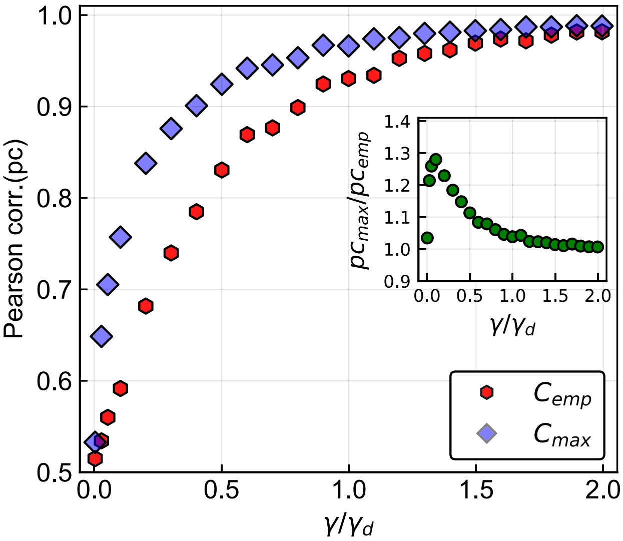

Figure 2. Pearson correlation between empirical/maximum-likelihood covariance matrices and the true covariance matrix as a function of γ/γ_d. The inset plot represents the ratio between the Pearson correlation for the maximum-likelihood covariance matrix and the one for the empirical covariance matrix. Simulations are performed for a tree of 512 leaves and system size $L = 4$.

covariance matrix $\boldsymbol{C}$. To establish a reference point for inference quality, we compare $\boldsymbol{C}$ to the empirical covariance matrix $\boldsymbol{C}_\text{emp}$ computed as if observations were independent, i.e. $\boldsymbol{C}_\text{emp}(a,b) = N^{-1}\sum_{i=1}^{N} x_i^a x_i^b$. Note that it is also possible to estimate the accuracy of the method by comparing the coupling matrices obtained by inverting each covariance matrix, namely $\boldsymbol{J}$, $\boldsymbol{J}_\text{max}$ and $\boldsymbol{J}_\text{emp}$. Here, we describe results obtained on artificial data.

Data is generated according to the process described in section 3. We first build a symmetric binary tree $\mathcal{T}$ with $N = 2^9 = 512$ leaves. Here, binary means that every non-leaf node of the tree has exactly two children, while symmetric means that the two subtrees stemming from any non-leaf node have the exact same topology. The length of each branch of $\mathcal{T}$ is chosen from a uniform distribution in the interval $[0, 1]$. We also sample positive semi-definite coupling matrix $\boldsymbol{J}$ of size $L \times L$ with $L = 4$ or $L = 10$, with entries normally distributed with mean $\mu_J = 0.8$ and $\sigma_J = 0.2$.

In the case of statistical models of protein sequence, a major achievement is the ability of pairwise models to predict contacts in the three-dimensional structure of the protein from an inferred coupling matrix. In order to replicate this setting and to perform interaction prediction, we randomly set to 0 off-diagonal elements of J with probability 0.7, resulting in a sparsified coupling matrix of approximate density 0.3. Zero elements of J correspond to variables that do not interact, in analogy to non-contacts in the case of an application to protein sequences.

In order to investigate the different regimes of tree-induced correlation, we vary the parameter γ around a reference timescale γ_d defined as follows:

$$\gamma_\text{d} = \frac{1}{\Delta t_\text{av}\rho_\text{min}} \tag{23}$$

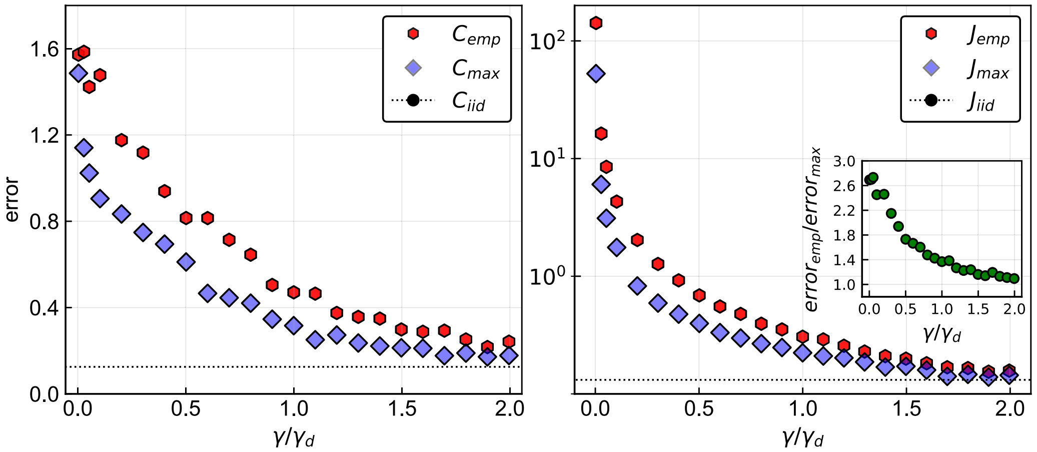

Figure 3. (Left) Relative $l2$-error between empirical or maximum-likelihood covariance matrices and the true covariance matrix as a function of $\gamma/\gamma_{\rm d}$. (Right) Relative $l2$-error between empirical/maximum-likelihood coupling matrices and the true coupling matrix as a function of $\gamma/\gamma_{\rm d}$. Logarithmic scale is chosen for the y-axis because of large values of the error at low γ. The inset shows the ratio between the two errors.

where $\Delta t_{\rm av}$ is the average branch length separating two leaves of $\mathcal{T}$. For $\gamma \gg \gamma_{\rm d}$, leaf configurations are on average well decorrelated, whereas for $\gamma \ll \gamma_{\rm d}$ all leaves will be strongly correlated. By simulating data using a range of values of γ, we can investigate these different regimes. Empirically, we find that the range $\gamma \in [10^{-2}, 2] \cdot \gamma_{\rm d}$ covers all relevant temporal regimes. For each value of γ, we then sample configurations of leaves of $\mathcal{T}$ using the process described in section A1 of the appendix. To avoid statistical noise when assessing the quality of our inference, we repeat the sampling of leaf configurations 100 times for each value of γ.

For each repetition of the sampling process, we perform our maximum likelihood procedure and obtain an inferred covariance matrix $\boldsymbol{C}_{\rm max}$. We also compute the empirical covariance matrix $\boldsymbol{C}_{\rm emp}$ as if leaf configurations were independent. Figure 2 shows the Pearson correlation between the real covariance matrix $\boldsymbol{C} = \boldsymbol{J}^{-1}$ and the empirical or inferred ones in the $L = 4$ case (similar figures for $L = 10$ are in appendix A.6). As expected, both methods perform well in the large γ limit with a correlation close to 1, and worse in the low γ limit. In this latter case, correlations due to phylogeny are too strong for our maximum likelihood method to pick up a signal, and both methods perform equally poorly. However, there exists an intermediate regime where $\boldsymbol{C}_{\rm max}$ is much closer to the actual correlation than $\boldsymbol{C}_{\rm emp}$. In figure 3, we plot the relative $l2$-error between either covariance matrices in the left panel or coupling matrices in the right panel, where the $l2$-error is defined as the $l2$-norm of the difference between two matrices, e.g. $||\boldsymbol{C}_{\rm max} - \boldsymbol{C}||_2$. In both cases, our maximum-likelihood method results in a consistent improvement over the empirical estimator. However, the relative error still reaches high values in the low γ regime, which is likely due to $\boldsymbol{C}_{\rm max}$ and $\boldsymbol{C}_{\rm emp}$ being close to singular in this case.

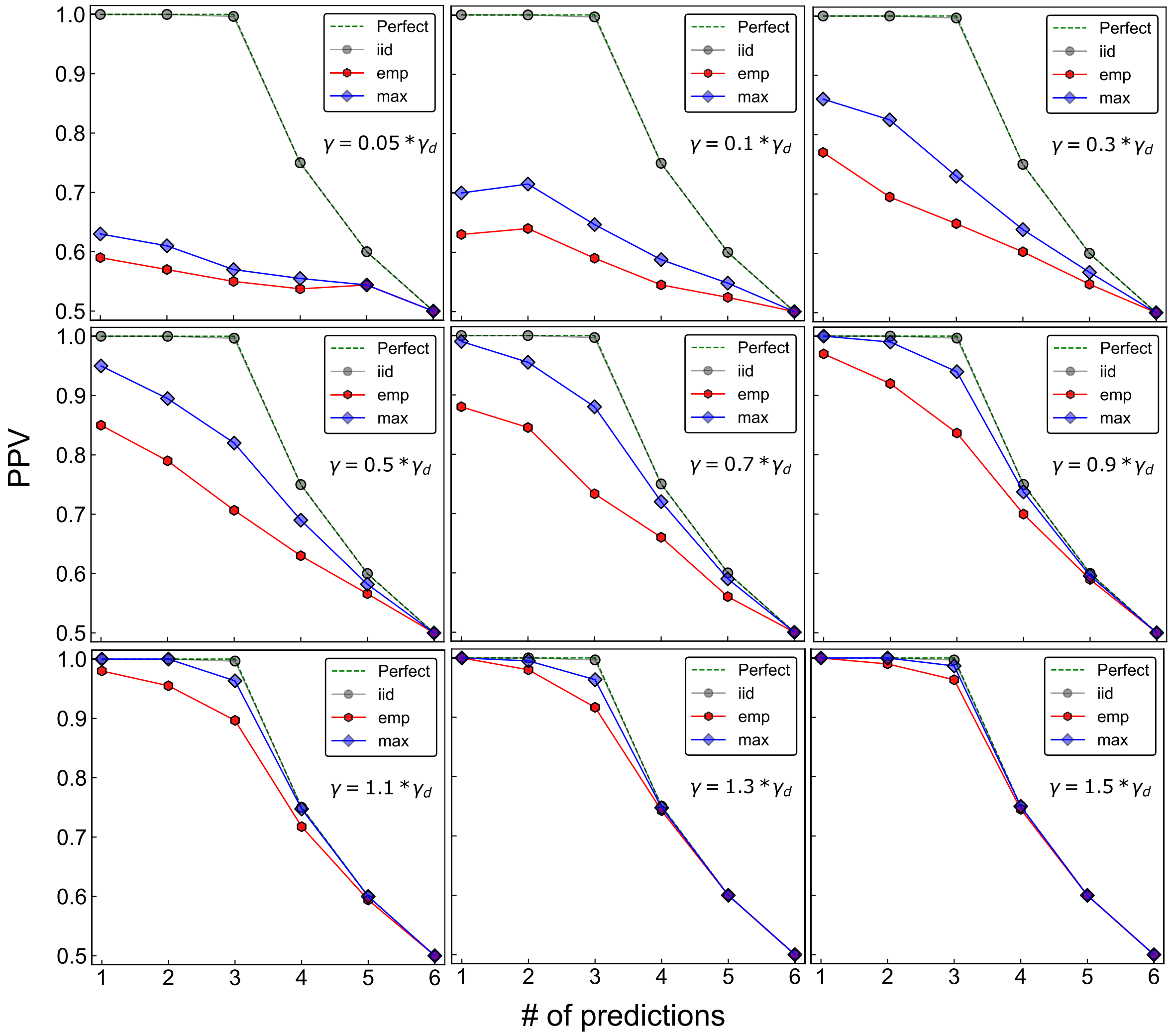

Figure 4. Quality of prediction of interactions for different values of γ and system size $L = 4$. Interactions are defined as non-zero elements of the coupling matrix. In the $L = 4$ case, there are six possible interactions. Predictions are made by taking the largest elements (in absolute terms) of the inferred coupling matrix. The PPV is the fraction of correctly predicted contacts for a given number of predictions. The perfect prediction is the one that would be made using the real coupling matrix $\boldsymbol{J}$.

An interesting way to illustrate the benefits of reconstructing the covariance matrix using knowledge of the tree is to evaluate the gain in *effective sample size*. Intuitively, the use of correlated samples reduces the information contained in the data, as compared to an equally large dataset of i.i.d. configurations. It is therefore interesting to compare the accuracy of our inferences with the accuracy obtained by computing the empirical covariance matrix using i.i.d. samples. This can serve as a baseline for the minimal error that can be achieved for a given sample size N, but also to estimate an effective i.i.d. sample size N_{eff} that would empirically result in the same difference found between $\boldsymbol{C}_{\mathrm{max}}$ and $\boldsymbol{C}$. In other words, to each value of $||\boldsymbol{C}_{\mathrm{max}} - \boldsymbol{C}||_2$ corresponds an i.i.d. sample size N_{eff} for which the same $l2$-error is obtained.

We report in figure S5 the $l2$-error between true and empirical covariances computed from i.i.d. samples of variable sizes N. As expected, the error increases with decreasing values of N. We can use this in turn to express values of the $l2$-error in correlated samples in terms of effective i.i.d. sample sizes. For example, the error reached by $\boldsymbol{C}_{\text{emp}}$ for $\gamma/\gamma_{\text{d}} \in [0.5, 1]$ and $L = 4$ lies in the range $[0.8, 0.5]$ with average value 0.7, corresponding to the one obtained for an i.i.d. sample of size $N_{\text{eff}} \sim 16$ ($l2$-error 0.7). The error obtained in the same range of γ values by using $\boldsymbol{C}_{\text{max}}$ lies in the range $[0.6, 0.3]$ with average value 0.4, corresponding to i.i.d. sample sizes $N_{\text{eff}} \sim 32$–64 (with respective $l2$-errors 0.5 and 0.3). Thus, our correction is equivalent to increasing the number of effective samples by a factor of 2–4.

Finally, we assess the performance of our method in improving the prediction of the network of interactions between the Gaussian variables $\{x_a\}$. We consider that two variables x_a and x_b interact if the corresponding entry in the coupling matrix is non-zero, that is $J_{ab} \neq 0$. Using the data, we predict these interactions by taking the largest n elements (in absolute value) of the inferred coupling matrix, resulting in n predictions. The fraction TP/n of these n predictions that correspond to non-zero entries in the true matrix (TP = true positives) defines the positive predictive value (PPV). This problem is equivalent to the one of predicting contacts in a protein structure [30, 33].

Figure 4 shows the PPV as a function of the number of predictions for different values of γ and $L = 4$ (see figure S10 for the $L = 10$ case). In this case, the coupling matrix only has six independent non-diagonal elements, and only six predictions can be made. Our correction systematically outperforms the predictions from the empirical coupling matrix, with an always larger PPV. This gain is negligible in the extreme regimes of very high γ, where the prediction is close to identical to the one obtained with an i.i.d. sample, or very low γ, where it is essentially random. It is however much larger in the intermediate regime, with a significantly improved prediction in the region $\gamma/\gamma_{\text{d}} \in [0.5, 1]$.

6. Discussion

In this work, we proposed a method for inferring parameters of an OU process using data that are correlated through an evolutionary tree. We kept a very general setting in which data can in principle represent any set of continuous phenotypic traits or potentially discrete sequences if a continuous approximation is made. As such, our approach is purely methodological, and does not directly investigate any particular application.

We showed that due to the Gaussian and time-reversible nature of the OU process, it is possible to write the joint covariance matrix of all data vectors in a simple way. The resulting matrix $\mathbb{G}$ consists of block entries that represent covariances between pairs of leaves. The dependence of these blocks on the coupling matrix $\boldsymbol{J}$ characterizing the OU process and on the tree structure can be written explicitly. Interestingly, $\mathbb{G}$ only depends on the tree structure through the pairwise path length Δt_{ij} separating leaves along the tree.

We then proposed a way to compute the likelihood of the data given the tree and the parameters of the OU process, namely the coupling matrix $\mathbf{J}$ and timescale γ. This method relies on computing the eigenvalues and vectors of the joint covariance matrix in an efficient manner. Indeed, it is possible to separate this calculation in two steps: the first in which we perform the eigen-decomposition of the matrix $\mathbf{J}$, and the second in which we compute eigenvalues and vectors of matrices $\mathbf{G}^a$ that embed the tree structure. This reduces the computational complexity from $\mathcal{O}(L^3N^3)$ for a naive inversion of $\mathbb{G}$ to $\mathcal{O}(L^3)+\mathcal{O}(LN^3)$. We also show that this method can be used to compute the gradient of the likelihood with respect to parameters with the same complexity. This makes the problem of inferring $\mathbf{J}$ amenable to maximum likelihood methods using a gradient ascent approach.

Finally, we showed that this process gives encouraging results on simulated data, with a more accurate reconstruction of parameters than if empirical estimation was performed. These simulations highlight the fact that this method is only useful in the intermediate regime of phylogenetic correlations. If the timescale γ characterizing the branch lengths of the tree is too large, correlation of data points through the tree is weak and an empirical estimation performs well. On the other hand, a very low γ results in strong phylogenetic biases that make recovering $\mathbf{J}$ impossible, basically due to a strong reduction of the information in a too redundant dataset. However, in an intermediate regime where intrinsic and historical correlations in the dataset coexist, our tree-aware reconstruction of $\mathbf{J}$ results in clear benefits over a tree-unaware empirical estimation.

A limitation of our approach remains the long computational time. Even with the efficient computation of the gradient, it was necessary to use small system sizes, $L=10$ at most, to repeat our inference process many times with simulated data in a reasonable time. For this reason, the framework proposed here is limited to a small number of variables. In this respect, it is interesting to note that a different manner of computing the likelihood developed in [29] and based on Gaussian integrations on every branch of the tree results in an asymptotic complexity of $\mathcal{O}(NL^3)$.

Although our method can in principle be used for any set of traits, a major motivation in developing it is its potential application to model of protein sequences. Several results in recent years have shown that selection forces shaping the evolution of protein sequences are well described by a pairwise potential [35, 42]. The estimation of this potential is performed using homologous sequences, and is therefore biased by the phylogenetic relations between these sequences. Results presented here are a first step in disentangling effects due to phylogeny from effects due to selection in a principled way.

However, there remain several challenges in using this framework for protein sequences. First, the computational power required to process actual sequences is much larger than what was needed for the small simulated systems presented here. As an example, a protein of length $L=100$ will be represented by $q\times 100=2000$ Gaussian variables, where $q=20$ is the number of amino acids. This is of course much larger than the $L=10$ system used as an example to test our approach.

A second question is the capacity of a continuous variable approximation, necessary when using OU dynamics, to represent dynamical properties of the landscape protein sequences evolve in. This type of approximation has been successfully used before, but in quite different contexts [30–32]. Its use in the context of modeling the evolutionary dynamics of protein sequences remains an open question.

Code availability

The code used to compute and optimize the likelihood is available at https://github.com/ed-rodh/GaussianPhylogeny.

Acknowledgments

We acknowledge interesting discussions with Roberto Mulet. P B C and M W acknowledge the hospitality of the Department of Theoretical Physics of University of Havana, where part of this work was done. Our work was partially funded by the EU H2020 Research and Innovation Programme MSCA-RISE-2016 under Grant Agreement No. 734439 InferNet.

Appendix A. Description of technical details

A.1. Generating artificial data

We are interested in the case where the dynamics of the L-dimensional OU process takes place on a tree. For example, if configurations $\{\vec{x}\}$ represent quantitative traits of some organisms, the tree can represent the genealogy or phylogeny of these organisms. Therefore, to generate our datasets, we have to be able to simulate the OU process on a tree. In practice, given a rooted tree such as the one shown in figure 1 of the main text, we want to sample a configuration $\vec{x}$ for every node in such a way that equation (6) holds for every pair of nodes, with time Δt being the path length connecting the nodes along the tree.

We use a simple methodology to achieve this. First, note that given an arbitrary configuration $\vec{x}_0$ and a time Δt, we can generate a new configuration $\vec{x}$ distributed according to the propagator equation (4) by exploiting the transformation

$$\vec{x} = \mathbf{\Lambda}^{\Delta t}\vec{x}_0 + \mathbf{\Sigma}^{1/2}\vec{\eta}\ , \tag{A1}$$

where $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ are defined in equation (5), and $\vec{\eta}$ is a vector of uncorrelated variables drawn individually from the normal distribution $\mathcal{N}(0, 1)$. Moreover, if $\vec{x}_0$ is distributed according to the equilibrium distribution equation (1), then $\vec{x}$ and $\vec{x}_0$ are distributed according to the joint distribution equation (6) describing two equilibrium configurations at finite time difference. Note that equation (A1) is quite different from the Langevin equation (2), which describes the instantaneous dynamics of $\vec{x}$ in the potential given by $\boldsymbol{J}$, and which could also be simulated in a more complicated situation where no analytical expression for the propagator can be derived.

Given any already sampled internal node in the tree, equation (A1) allows us to emit a configuration for each of its child nodes. To sample the whole tree, we first draw the root configuration $\vec{x}_r$ from the equilibrium distribution equation (1). By recursive applications of equation (A1), we then simply work our way down the tree until all leaves are sampled. Only the configurations at the leaves form the data set, and the internal configuration remain hidden to our model-learning task.

In our inference scheme, the tree topology and the lengths of the branches where the dynamics takes place are assumed to be arbitrary but known. For the specific simulations presented in the main text, we chose a symmetric binary tree $\mathcal{T}$ with $N = 2^9 = 512$ leaves and the length of each branch Δt drawn independently from a uniform distribution over the interval $[0, 1]$.

A.2. Empirical parameters

A.2.1. Eigenvalues and eigenvectors of $\boldsymbol{C}^{-1}$. We initialize the covariance matrix using the empirical one:

$$\mathbf{C}^{\mathrm{emp}} = \frac{1}{N}\sum_{i=1}^{N} \vec{x}_i \cdot \vec{x}_i^{\mathrm{T}}.$$

Its eigenmodes $\{\rho_a^0, \vec{s}_a^{\,0}\}$ determine the starting point of the optimization. A suitable parametrization of $\vec{s}_a^{\,0}$ in terms of generalized Eulerian angles or a skew symmetric matrix is described below in appendix A.3.

A.2.2. Timescale parameter γ. The optimization also requires that we initialize the timescale γ. For coherence with the last section, we define the empirical value of γ as the optimal one given the data $\vec{X}$, the tree, and the OU process defined by the empirical covariance matrix.

The probability distribution P for the configurations of two leaves $\vec{x}_i$ and $\vec{x}_j$ separated by time Δt_{ij} is given by equation (6) of the main text. With this distribution we can analytically calculate the average of the scalar product $\vec{x}_i^{\mathrm{T}} \cdot \vec{x}_j$ of two equilibrium configurations at given time separation:

$$\left\langle \vec{x}_i^{\mathrm{T}} \cdot \vec{x}_j \right\rangle_P = \sum_{a=1}^{L} \left\langle x_i^a x_j^a \right\rangle_P. \tag{A2}$$

The covariance $\langle x_i^a x_j^a \rangle_P$ of two observations separated by time Δt_{ij} is given by equation (7). Using this, we find

$$\begin{aligned} \langle \vec{x}_i^{\mathrm{T}} \cdot \vec{x}_j \rangle_P &= \sum_{a=1}^{L} \left(\boldsymbol{\Lambda}^{\Delta t_{ij}} \mathbf{C}\right)_{aa} \\ &= \mathrm{Tr}\left(\boldsymbol{\Lambda}^{\Delta t_{ij}} \mathbf{C}\right) \\ &= \sum_{a=1}^{L} \rho_a^{-1} \mathrm{e}^{-\gamma \rho_a \Delta t_{ij}}. \end{aligned} \tag{A3}$$

Having initialized the covariance matrix $\mathbf{C}$ with its empirical value, we know the values of all members of the rhs of equation (A3) except the one of γ. To find an initial value of γ which is consistent with the data and the empirical covariance matrix for all pairs of data configurations $i < j$, we search for one that best explains the observed scalar products between configurations. We thus define γ^0 to be the argument minimizing the functional $F(\gamma)$:

$$F(\gamma) = \sum_{1\leqslant i<j\leqslant N} \left[\vec{x}_i^{\mathrm{T}} \cdot \vec{x}_j - \sum_{a=1}^{L} \rho_a^{-1} \mathrm{e}^{-\gamma\rho_a \Delta t_{ij}} \right]. \tag{A4}$$

Since F depends on a single scalar parameter, it is straightforward to minimize it and thereby to initialize γ to an empirically reasonable value.

A.3. Parametrizations of eigenvectors

A.3.1. Parametrization using generalized Eulerian angles. The idea is to write the base vectors $\vec{s}_a$ as columns of an orthogonal matrix $\boldsymbol{T}$, and to parameterize this matrix in terms of $L(L-1)/2$ independent variables θ_{pq} with $1 \leqslant p < q \leqslant L$. These parameters are called generalized Eulerian angles, since they generalize Eulerian angles to $L > 3$.

To construct this matrix, we start from a rotational transformation in a two-dimensional subspace of an L-dimensional space. It is given by an L-dimensional matrix of the form

$$\boldsymbol{a}_{pq} = \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cos\,\theta_{pq} & \cdot & \sin\,\theta_{pq} & \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & -\sin\,\theta_{pq} & \cdot & \cos\,\theta_{pq} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}, \tag{A5}$$

where all diagonal elements are unity except for the diagonal elements in the pth and the qth column, which equal $\cos\theta_{pq}$. All off-diagonal elements are zero, represented as dots above, except for the one corresponding to the intersection of the pth row and the qth column, which is $\sin\theta_{pq}$, and that on the intersection of the qth row and the pth column, which equals $-\sin\theta_{pq}$. There are $L(L-1)/2$ matrices of this form, corresponding to all choices of p and q with $1 \leqslant p < q \leqslant L$.

An arbitrary L-dimensional orthogonal matrix $\boldsymbol{T}$ can be represented as a product of these $L(L-1)/2$ orthogonal matrices with appropriate values of the $L(L-1)/2$ independent parameters θ_{pq}. Raffenetti *et al* [37] exposed a recursive algorithm to efficiently perform the matrix multiplication, as well as the construction of the derivatives of $\boldsymbol{T}$ with respect to parameters θ_{pq}. The main equations are presented below.

The matrix multiplication can be done by a sequence of L steps implied by the following recurrence relations where n goes from 1 to L:

$$A^{(n)} = \boldsymbol{a}_{n,n}\, \boldsymbol{a}_{n-1,n} \ldots \boldsymbol{a}_{2,n}\, \boldsymbol{a}_{1,n}, \tag{A6}$$

$$\boldsymbol{T}^{(1)} = 1, \tag{A7}$$

$$\boldsymbol{t}^{(n)} = \begin{pmatrix} \boldsymbol{T}^{(n-1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}, \tag{A8}$$

$$\boldsymbol{T}^{(n)} = \boldsymbol{A}^{(n)}\boldsymbol{t}^{(n)}, \tag{A9}$$

$$\boldsymbol{T} = \boldsymbol{T}^{(L)}, \tag{A10}$$

where the $\boldsymbol{a}_{n,m}$ matrices are defined by (A5) for $n \neq m$, and $\boldsymbol{a}_{n,n}$ is the identity matrix of dimension n.

The recurrence equations given by (A9) can be explicitly written as

$$T_{kl}^{(n)} = \cos\,\theta_{kn} \cdot t_{kl}^{(n)} - \sin\,\theta_{kn} \cdot z_{kl}^{(n)} \quad \text{with} \quad k, l = 1, \ldots, n \tag{A11}$$

where

$$z_{kl}^{(n)} = \begin{cases} \delta_{\text{ln}} & \text{for } k = 1 \\ \sin\,\theta_{k-1n} \cdot t_{k-1l}^{(n)} + \cos\,\theta_{k-1n} \cdot z_{k-1l}^{(n)} & \text{for } k = 2, \ldots, n \end{cases} \tag{A12}$$

with $\theta_{nn} = \pi/2$.

Thus, if $\boldsymbol{T}^{(n-1)}$ is given, we find $\boldsymbol{t}^{(n)}$ from equation (A8). Then, from elements $t_{kl}^{(n)}$ we get $z_{kl}^{(n)}$ using (A12) and finally from $z_{kl}^{(n)}$ and $t_{kl}^{(n)}$ we obtain $T_{kl}^{(n)}$.

Therefore, eigenvectors $\vec{s}_a$ can be chosen as ath column of matrix $\boldsymbol{T}$:

$$s_a^k = T_{k,a}^{(L)} = \cos\,\theta_{kL} \cdot t_{ka}^{(L)} - \sin\,\theta_{kL} \cdot z_{ka}^{(L)} \;\; \text{for } k = 1, .., L. \tag{A13}$$

A.3.1.1. Determination of parameters for a given matrix. To use this expression, we still need to determine parameters $\boldsymbol{\theta}$ given an orthogonal matrix $\boldsymbol{T}$, such that all equations

$$T_{ij}(\boldsymbol{\theta}) = T_{ij} \tag{A14}$$

are satisfied. This system of nonlinear transcendental equations cannot be solved algebraically. However, it is possible to overcome this issue finding the set of $\boldsymbol{\theta}$ which minimizes the square distance between the target and the parametrized matrices:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i<j} [T_{ij} - T_{ij}(\boldsymbol{\theta})]^2. \tag{A15}$$

This is useful when we initialize parameters $\boldsymbol{\theta}$ for the matrix formed by the eigenvectors of the empirical covariance matrix.

A.3.1.2. Derivatives with respect to the angular parameters. To compute the derivatives of $\boldsymbol{T}$ with respect to the angular parameters θ_{pq}, we first note that it is possible to rewrite the recurrence step of equation (A9) as the following matrix product

$$\boldsymbol{T}^{(L)} = \boldsymbol{B}^{(L)}\boldsymbol{B}^{(L-1)} \ldots \boldsymbol{B}^{(3)}\boldsymbol{B}^{(2)} \tag{A16}$$

where

$$\boldsymbol{B}^{(n)} = \begin{pmatrix} \boldsymbol{A}^{(n)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}^{(L-n)} \end{pmatrix} \tag{A17}$$

is block diagonal, $\boldsymbol{A}^{(n)}$ was defined by equation (A6) and $\boldsymbol{I}^{(L-n)}$ is a unit matrix of $(L - n)$ dimensions.

From the definition of $\boldsymbol{A}^{(n)}$ we note that the terms θ_{pq} for $p = 1, 2, \ldots, q-1$ only occur in the factor $\boldsymbol{B}^{(q)}$ of equation (A16). This allows us to write the derivative of $\boldsymbol{T}$ with respect to θ_{pq} as the matrix product:

$$\frac{\partial \boldsymbol{T}}{\partial \theta_{pq}} = \boldsymbol{B}^{(L)} \boldsymbol{B}^{(L-1)} \ldots \frac{\partial \boldsymbol{B}^{(q)}}{\partial \theta_{pq}} \ldots \boldsymbol{B}^{(3)} \boldsymbol{B}^{(2)} \tag{A18}$$

where

$$\frac{\partial \boldsymbol{B}^{(q)}}{\partial \theta_{pq}} = \begin{pmatrix} \frac{\partial \boldsymbol{A}^{(q)}}{\partial \theta_{pq}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \tag{A19}$$

Therefore, the calculation of the derivative of $\boldsymbol{T}$ could be done with the following three steps:

(a) Calculate the product $\boldsymbol{B}^{(q-1)} \boldsymbol{B}^{(q-2)} \ldots \boldsymbol{B}^{(3)} \boldsymbol{B}^{(2)}$.

(b) Calculate

$$\frac{\partial \boldsymbol{B}^{(q)}}{\partial \theta_{pq}} \boldsymbol{B}^{(q-1)} \ldots \boldsymbol{B}^{(3)} \boldsymbol{B}^{(2)} = \begin{pmatrix} \frac{\partial \boldsymbol{T}^{(q)}}{\partial \theta_{pq}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \tag{A20}$$

(c) Calculate

$$\frac{\partial \boldsymbol{T}}{\partial \theta_{pq}} = \boldsymbol{B}^{(L)} \boldsymbol{B}^{(L-1)} \ldots \boldsymbol{B}^{(q+1)} \begin{pmatrix} \frac{\partial \boldsymbol{T}^{(q)}}{\partial \theta_{pq}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \tag{A21}$$

The $q-3$ recurrence steps for step (a) can be carried out using the same recurrence scheme described before for matrix $\boldsymbol{T}$ construction. For step (b), we need to evaluate

$$\frac{\partial \boldsymbol{T}^{(q)}}{\partial \theta_{pq}} = \begin{cases} -\sin\, \theta_{pq} \sigma_{kl}^{(q)} & k > p \\ -\sin\, \theta_{pq} t_{kl}^{(q)} - \cos\, \theta_{pq} z_{kl}^{(q)} & k = p \\ 0 & k < p \end{cases} \tag{A22}$$

where the quantities

$$\sigma_{kl}^{(q)} = \frac{\partial \boldsymbol{z}_{kl}^{(q)}}{\partial \theta_{pq}} \tag{A23}$$

can be obtained from (A12). Finally, for step (c), we follow $L-q$ recurrence steps described by equations (A11) and (A12) with the two exceptions:

$$z_{1l}^{(n)} = 0 \quad ; \quad 1 \leqslant l < n \tag{A24}$$

$$z_{k+1,n}^{(n)} = 0 \qquad 1 \leqslant k \leqslant n-1. \tag{A25}$$

A.3.2. Parametrization in terms of the exponential of a skew-symmetric matrix. The expression of the eigenvectors of $\boldsymbol{C}^{-1}$ in terms of Eulerian angles allows us to compute the gradient of the likelihood with respect to them, as well as expressing it analytically. In practice, we adopted another parametrization of $\boldsymbol{C}^{-1}$ to accelerate the evaluation of the gradient. This new parametrization consists of expressing $\boldsymbol{C}^{-1}$ as the exponential of a *skew-symmetric* matrix, that is a matrix X such that $\boldsymbol{X} = -\boldsymbol{X}^{\mathrm{T}}$.

A skew-symmetric matrix X of size L has $L(L-1)/2$ independent values, and its exponential is a special orthogonal matrix:

$$\boldsymbol{S} = \exp(\boldsymbol{X}). \tag{A26}$$

This is simply shown by the fact that $\exp(\boldsymbol{X})^{\mathrm{T}} = \exp(\boldsymbol{X}^{\mathrm{T}}) = \exp(-\boldsymbol{X}) = \exp(\boldsymbol{X})^{-1}$ and $\det(\boldsymbol{S}) = \exp(\mathrm{Tr}\,\boldsymbol{X}) = 1$ since $\mathrm{Tr}\,\boldsymbol{X} = 0$ for a skew-symmetric matrix. Furthermore, it is always possible to obtain a skew-symmetric matrix $\boldsymbol{X}$ from a special orthogonal matrix $\boldsymbol{S}$ by inverting the exponential relation, $\boldsymbol{X} = \log \boldsymbol{S}$ [38].

The advantage of expressing $\boldsymbol{S}$ in this form is that $\boldsymbol{X}$ has $L(L-1)/2$ entries that can be varied independently. This allows us to perform the optimization over $L(L-1)/2$ independent parameters, with derivatives with respect to the independent entries of $\boldsymbol{X}$ being defined by

$$\frac{\partial \boldsymbol{S}}{\partial X_{jk}} = \lim_{h\to 0} \frac{1}{h}\left(\exp(\boldsymbol{X} + h\boldsymbol{E}^{jk}) - \exp(\boldsymbol{X})\right) \tag{A27}$$

where $\boldsymbol{E}^{jk}$ for $j > k$ is defined as a skew-symmetric matrix that has only two non-zero entries in positions (j, k) and (k, j):

$$E^{pq}_{jk} = \delta_{pj}\delta_{qk} - \delta_{pk}\delta_{qj}. \tag{A28}$$

It is not possible to give a simple analytical form to equation (A27). However, since $\boldsymbol{S}$ is obtained through a simple algebraic expression (equation (A26)), it is possible to compute its derivative with respect to entries of $\boldsymbol{X}$ through automatic differentiation techniques [43]. We implemented this process using the Julia package Zygote.jl [39].

A.4. Homogeneous and symmetric binary tree

We define a tree as homogeneous if all of its branches have the same length. The symmetric and the binary character of a tree are understood as in the main text: binary means that every non-leaf node of the tree has exactly two children, while symmetric means that the two subtrees stemming from any non-leaf node have the exact same topology.

Let us assume that the tree is symmetric, binary and completely homogeneous with all branches having the length Δt. The covariance matrix for a homogeneous symmetric binary tree with branches of length Δt, $K = 2$ branching levels and four leaves is

$$\mathbb{G} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{C}\Lambda^{2\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} \\ \boldsymbol{C}\Lambda^{2\Delta t} & \boldsymbol{C} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} \\ \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C} & \boldsymbol{C}\Lambda^{2\Delta t} \\ \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{2\Delta t} & \boldsymbol{C} \end{pmatrix}. \tag{A29}$$

The associated matrix $\boldsymbol{G}^a = z(\rho_a, \gamma, \Delta t)$ defined in equation (A29) becomes

$$\boldsymbol{G}^a = \rho_a^{-1} \begin{pmatrix} 1 & \mathrm{e}^{-2\gamma\rho_a\Delta t} & \mathrm{e}^{-4\gamma\rho_a\Delta t} & \mathrm{e}^{-4\gamma\rho_a\Delta t} \\ \mathrm{e}^{-2\gamma\rho_a\Delta t} & 1 & \mathrm{e}^{-4\gamma\rho_a\Delta t} & \mathrm{e}^{-4\gamma\rho_a\Delta t} \\ \mathrm{e}^{-4\gamma\rho_a\Delta t} & \mathrm{e}^{-4\gamma\rho_a\Delta t} & 1 & \mathrm{e}^{-2\gamma\rho_a\Delta t} \\ \mathrm{e}^{-4\gamma\rho_a\Delta t} & \mathrm{e}^{-4\gamma\rho_a\Delta t} & \mathrm{e}^{-2\gamma\rho_a\Delta t} & 1 \end{pmatrix}. \tag{A30}$$

Matrices such as the one in (A30) are called *hyper-geometric*. For dimensions 2^K, they have $K+1$ different eigenvalues given by:

$$\lambda_k(\rho_a, \gamma) = \rho_a^{-1} * \begin{cases} 1 + \displaystyle\sum_{l=1}^{k-1} 2^{l-1}\mathrm{e}^{-2l\gamma\rho_a\Delta t} - 2^{k-1}\mathrm{e}^{-2k\gamma\rho_a\Delta t}, & \text{for} \quad k \in [1, K] \\ 1 + \displaystyle\sum_{l=1}^{K} 2^{l-1}\mathrm{e}^{-2l\gamma\rho_a\Delta t}, & \text{for} \quad k = K+1 \end{cases} \tag{A31}$$

where $\lambda_{K+1} \geqslant \lambda_K \ldots \geqslant \lambda_1$. For $k < K+1$, the degeneracy of eigenvalue λ_k is $d_k = 2^{K-k}$. The associated eigenvectors are independent of the parameter ρ_a and reflect the events in the phylogenetic tree. Each eigenvector $\vec{u}_k$ of length 2^K captures the duplication events in the $(K+1-k)st$ generation:

$$\vec{u}_k = \begin{cases} (\underbrace{\overbrace{1,\ldots,1}^{2^{k-1}}, \overbrace{-1,\ldots,-1}^{2^{k-1}}}_{Q}, 0, \ldots, 0) \bigcup \Gamma(\vec{u}_k), & \text{for} \quad k \in [1, K] \\ (1, 1, 1, \ldots, 1, 1, 1), & \text{for} \quad k = K+1 \end{cases}$$

where $\Gamma(\vec{u}_k)$ represents the d_k combinations obtained by shifting the block of length $Q = 2^k$, generating all eigenvectors corresponding to the eigenvalue λ_k. These eigenvectors are orthogonal to each other, and can be used to construct the eigenvectors of the matrix $\mathbb{G}$ from equation (15)

To compute the gradient of the likelihood, derivatives of $\lambda_k(\rho_a, \gamma)$ with respect to ρ_a and γ can be directly obtained from expression (A31).

A.5. Optimization scheme

The proposed inference scheme was transformed into a multidimensional nonlinear optimization problem for which we can compute the gradient of the optimized quantity $\mathcal{L}$. To solve it, we used a variant of the QNMs. The Newton method differs from the classical gradient ascent technique in that it makes use of the matrix of the second-order derivatives of the quantity to maximize, that is the Hessian matrix $\boldsymbol{H}$. The main feature in QNM when compared to the standard Newton method is that the Hessian matrix is only approximated. When maximizing the likelihood $\mathcal{L}$ with respect to parameters $\vec{\theta}$, the direction of the change of parameters $\Delta\vec{\theta}$ is determined by

$$\Delta\vec{\theta} = \hat{\mathbf{H}}_k \nabla\mathcal{L}(\vec{\theta}_k),$$

where $\vec{\theta}_k$ and $\hat{\mathbf{H}}_k$ respectively represent the parameter values and the approximation of the Hessian at the kth iteration. Various QNMs differ in their approximation of the Hessian matrix. We used the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm [44, 45], which chooses $\hat{\mathbf{H}}_k$ as a positive definite matrix where

$$\hat{\mathbf{H}}_{k+1} = \hat{\mathbf{H}}_k + \frac{y_k y_k^{\mathrm{T}}}{y_k \Delta\theta_k} - \frac{\hat{\mathbf{H}}_k \Delta\theta_k (\hat{\mathbf{H}}_k \Delta\theta_k)^{\mathrm{T}}}{\Delta\theta_k^T \hat{\mathbf{H}}_k \Delta\theta_k} \quad \text{with} \quad y_k = \nabla\mathcal{L}(\vec{x}_k + 1) - \nabla\mathcal{L}(\vec{x}_k).$$

Implementation of the method was done using the NLopt package [41].

A.6. Supporting figures

See figures 5–8 and 10.

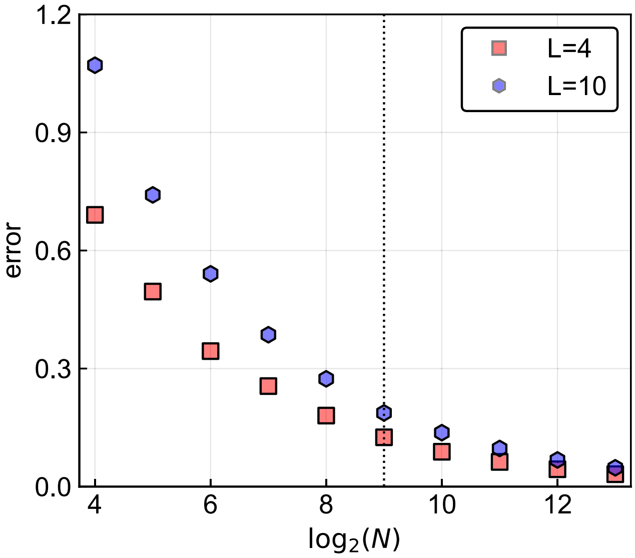

Figure 5. Relative $l2$-error between the empirical covariance matrix calculated from an i.i.d. sample and the true covariance matrix, for system sizes $L = 4$ and $L = 10$, as a function of the sample size N. An effective i.i.d. sample size N_{eff} can be attributed to each value of the $l2$-error. By comparing values of the $l2$-error with results in figures 3 and 9, we can formulate the gain in accuracy in terms of a gain in effective sample size. The dashed vertical line corresponds to the number of leaves of the tree used in the simulations.

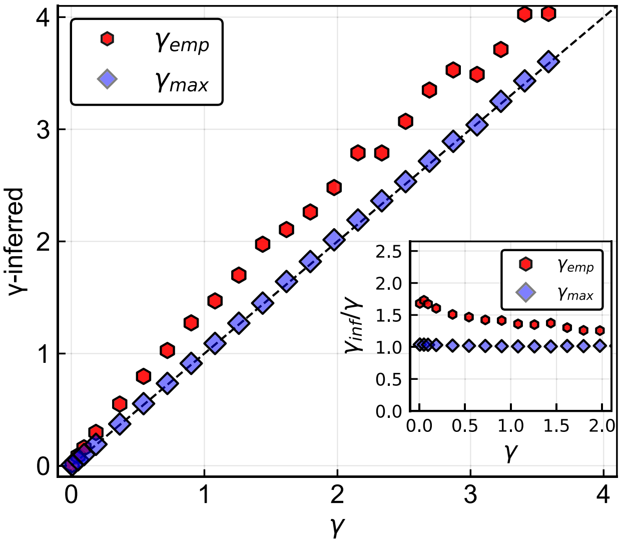

Figure 6. Inferred γ values as a function of real γ, for system size $L = 4$. γ_{emp} is the value obtained by the process described in appendix A.2. $\gamma_{\max}$ is the value inferred by the maximum-likelihood calculation. The inset represents the ratio of both inferred parameters γ_{emp} or $\gamma_{\max}$ to the real γ.

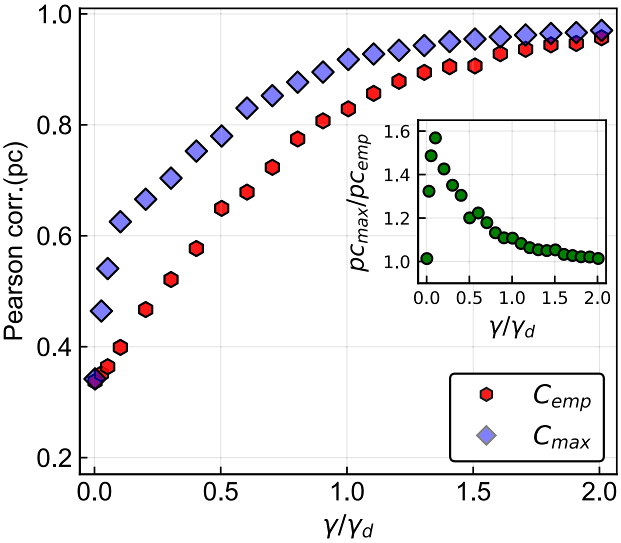

Figure 7. Pearson correlation between empirical/maximum-likelihood covariance matrices and the true covariance matrix as a function of γ/γ_{d}. The inset plot represents the ratio between the person correlation for the maximum-likelihood covariance matrix and the one for the empirical covariance matrix.

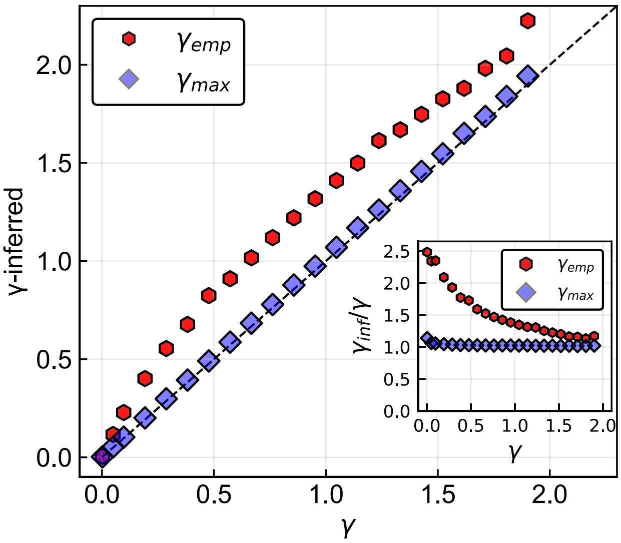

Figure 8. Inferred γ values as a function of real γ, for system size $L = 10$. γ_{emp} is the value obtained by the process described in appendix A.2. γ_{max} is the value inferred by the maximum-likelihood calculation. The inset represents the ratio of both inferred parameters γ_{emp} or γ_{max} to the real γ.

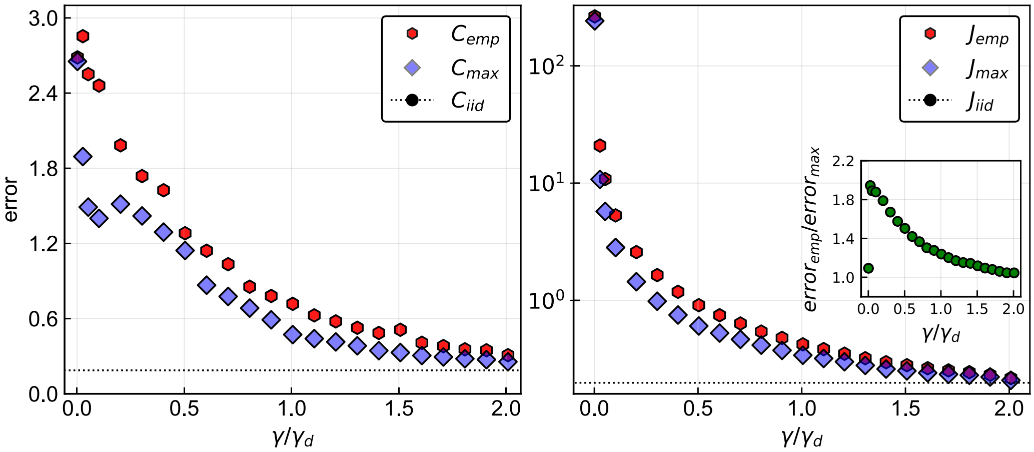

Figure 9. (Left) Relative $l2$-error between empirical or maximum-likelihood covariance matrices and the true covariance matrix as a function of γ/γ_d. (Right) Relative $l2$-error between empirical/maximum-likelihood coupling matrices and the true coupling matrix as a function of $\gamma/\gamma_{\mathrm{d}}$. Logarithmic scale is chosen for the y-axis because of large values of the error at low γ. The inset shows the ratio between the two errors. For system size $L = 10$.

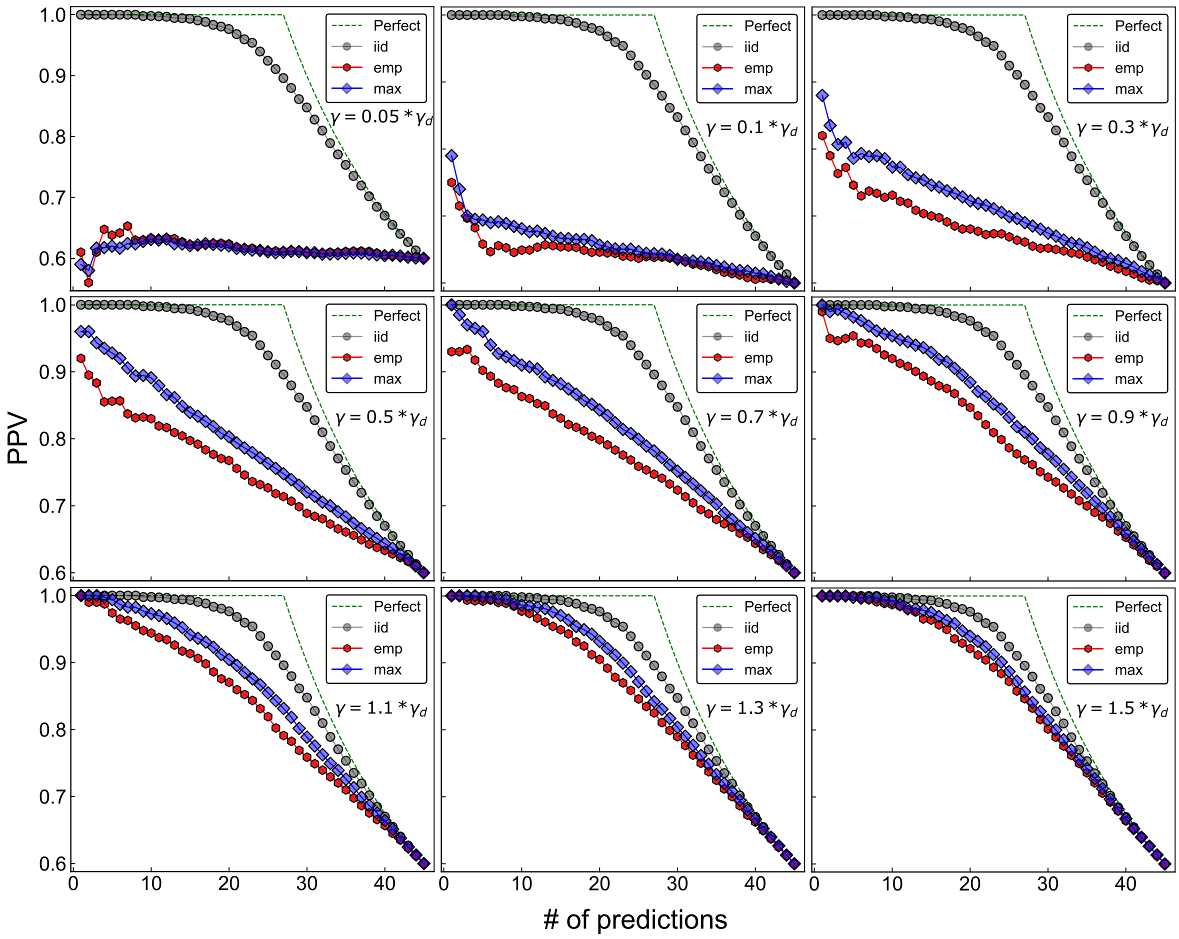

Figure 10. Quality of prediction of interactions for different values of γ and system size $L = 10$. Interactions are defined as non-zero elements of the coupling matrix. In the $L = 10$ case, there are 45 possible interactions. Predictions are made by taking the largest elements (in absolute terms) of the inferred coupling matrix. The PPV is the fraction of correctly predicted contacts for a given number of predictions. The perfect prediction is the one that would be made using the real coupling matrix $\boldsymbol{J}$.

References

[1] Chau N H, Zecchina R N and Berg J 2017 Inverse statistical problems: from the inverse Ising problem to data science *Adv. Phys.* **66** 197–261

[2] Levy R M, Haldane A and Flynn W F 2017 Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness *Curr. Opin. Struct. Biol.* **43** 55–62

[3] Cocco S, Feinauer C, Figliuzzi M, Monasson R and Weigt M 2018 Inverse statistical physics of protein sequences: a key issues review *Rep. Prog. Phys.* **81** 032601

[4] Schneidman E, Berry M J, Segev R and Bialek W 2006 Weak pairwise correlations imply strongly correlated network states in a neural population *Nature* **440** 1007–12

[5] Roudi Y, Tyrcha J and Hertz J 2009 Ising model for neural data: model quality and approximate methods for extracting functional connectivity *Phys. Rev.* E **79** 051915

[6] Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M and Walczak A M 2012 Statistical mechanics for natural flocks of birds *Proc. Natl Acad. Sci.* **109** 4786–91

[7] Cavagna A, Giardina I and Grigera T S 2018 The physics of flocking: correlation as a compass from experiments to theory *Phys. Rep.* **728** 1–62

[8] Bury T 2013 Market structure explained by pairwise interactions *Physica* A **392** 1375–85

[9] Borysov S S, Roudi Y and Balatsky A V 2015 Us stock market interaction network as learned by the Boltzmann machine *Eur. Phys. J.* B **88** 1–14

[10] Jaynes E T 1957 Information theory and statistical mechanics *Phys. Rev.* **106** 620

[11] Sayers E W, Cavanaugh M, Clark K, Ostell J, Pruitt K D and Karsch-Mizrachi I 2019 GenBank *Nucleic Acids Res.* **47** D94–9

[12] The UniProt Consortium 2018 UniProt: the universal protein knowledgebase *Nucleic Acids Res.* **46** 2699

[13] Durbin R, Eddy S R, Krogh A and Mitchison G 1998 *Biological Sequence Analysis* (*Probabilistic Models of Proteins and Nucleic Acids*) (Cambridge: Cambridge University Press)

[14] Felsenstein J 1988 Phylogenies and quantitative characters *Annu. Rev. Ecol. Syst.* **19** 445–71

[15] Felsenstein J 2003 *Inferring Phylogenies* (Oxford: Oxford University Press)

[16] Obermayer B and Levine E 2014 Inverse Ising inference with correlated samples *New J. Phys.* **16** 123017

[17] Horta E R, Barrat-Charlaix P and Weigt M 2019 Toward inferring Potts models for phylogenetically correlated sequence data *Entropy* **21** 1090

[18] Qin C and Colwell L J 2018 Power law tails in phylogenetic systems *Proc. Natl Acad. Sci.* **115** 690–5

[19] Horta E R and Weigt M 2021 Phylogenetic correlations have limited effect on coevolution-based contact prediction in proteins *PLoS Comput. Biol.* **17** e1008957

[20] Uhlenbeck G E and Ornstein L S 1930 On the theory of the Brownian motion *Phys. Rev.* **36** 823–41

[21] Hansen T F 1997 Stabilizing selection and the comparative analysis of adaptation *Evolution* **51** 1341–51

[22] Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W and Pääbo S 2004 A neutral model of transcriptome evolution *PLoS Biol.* **2** e132

[23] Hansen T F, Pienaar J and Hecht Orzack S 2008 A comparative method for studying adaptation to a randomly evolving environment *Evolution* **62** 1965–77

[24] Bartoszek K, Pienaar J, Mostad P, Andersson S and Hansen T F 2012 A phylogenetic comparative method for studying multivariate adaptation *J. Theor. Biol.* **314** 204–15

[25] Bedford T and Hartl D L 2009 Optimization of gene expression by natural selection *Proc. Natl Acad. Sci.* **106** 1133–8

[26] Rohlfs R V, Harrigan P and Nielsen R 2014 Modeling gene expression evolution with an extended Ornstein–Uhlenbeck process accounting for within-species variation *Mol. Biol. Evol.* **31** 201–11

[27] Nourmohammad A, Rambeau J, Held T, Kovacova V, Berg J and Lässig M 2017 Adaptive evolution of gene expression in Drosophila *Cell Rep.* **20** 1385–95

[28] Bastide P, Ané C, Robin S and Mariadassou M 2018 Inference of adaptive shifts for multivariate correlated traits *Syst. Biol.* **67** 662–80

[29] Mitov V, Bartoszek K, Asimomitis G and Stadler T 2020 Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts *Theor. Popul. Biol.* **131** 66–78

[30] Jones D T, Buchan D W A, Cozzetto D and Pontil M 2012 PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments *Bioinformatics* **28** 184–90

[31] Barton J P, Cocco S, De Leonardis E and Monasson R 2014 Large pseudocounts and L2-norm penalties are necessary for the mean-field inference of Ising and Potts models *Phys. Rev.* E **90**

[32] Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M and Pagnani A 2014 Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners *PLoS One* **9** e92721

[33] Morcos F *et al* 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families *Proc. Natl Acad. Sci.* **108** E1293–301

[34] Figliuzzi M, Jacquier H, Schug A, Tenaillon O and Weigt M 2016 Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1 *Mol. Biol. Evol.* **33** 268–80

[35] Russ W P *et al* 2020 An evolution-based model for designing chorismate mutase enzymes *Science* **369** 440–5

[36] Singh R, Ghosh D and Adhikari R 2017 Fast Bayesian inference of the multivariate Ornstein–Uhlenbeck process (arXiv:1706.04961)

[37] Raffenetti R C and Ruedenberg K 1970 Parametrization of an orthogonal matrix in terms of generalized Eulerian angles *Int. J. Quantum Chem.* **4** 625–34

[38] Shepard R, Brozell S R and Gidofalvi G 2015 The representation and parametrization of orthogonal matrices *J. Phys. Chem.* A **119** 7924–39

[39] Innes M 2018 Don't unroll adjoint: differentiating SSA-form programs (arXiv:1810.07951)

[40] Petersen K B and Pedersen M S 2015 *The Matrix Cookbook*

[41] Johnson S G 2014 The NLopt nonlinear-optimization package http://github.com/stevengj/nlopt

[42] Figliuzzi M, Barrat-Charlaix P and Weigt M 2018 How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35** 1018–27

[43] Griewank A 1989 On automatic differentiation *Mathematical Programming* (*Recent Developments and Applications*) (Dordrecht: Kluwer) pp 83–108

[44] Nocedal J 1980 Updating quasi-Newton matrices with limited storage *Math. Comput.* **35** 773

[45] Liu D C and Nocedal J 1989 On the limited memory BFGS method for large scale optimization *Math. Program.* **45** 503–28