Global multivariate model learning from hierarchically correlated data

Edwin Rodriguez Horta,[1,2] Alejandro Lage,[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*

Abstract

The ~~Inverse problem of Statistical Physics~~ inverse statistical physics infers maximum-entropy models ~~compatibles with a corresponding~~ compatible with a set of empirical averages ~~extracted~~ estimated from a high-dimensional dataset of independently distributed equilibrium configurations. ~~Practical interest extended these methods to non-equilibrium data like time series, where detailed balance does not hold.~~ However, in several applications, data ~~samples result from evolutionary processes~~ result from stochastic evolutionary processes, and the relation between configurations is ~~conditioned~~ characterized by a hierarchical ~~structure of the population. This makes sample statistics a superposition of signals coming from both: internal configuration correlations and relatedness correlations produced by common evolutionary history~~correlation structure, typically represented by a tree. ~~How to disentangle both sources of correlation for a better parameter inference is an open question to explore. Here we propose an approach where the evolutive process through a phylogenetic tree is~~ In turn, empirical averages of observables superpose intrinsic signals related to the equilibrium distribution of the studied system, and spurious historical (or phylogenetic) signals characterizing the specific correlated data-generating process. The naive application of maximum-entropy 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 ~~dynamics characterized by gaussian propagator and stationary distributions. Inference of these distributions from phylogenetic non-independent samples is solved in a Bayesian framework. This procedure can be extended to discrete variables by using a binary representation of data and approximating binary sates by continuous variables. Our method proposes a possible way for a better estimation of both, equilibrium and dynamic parameters, making applications more accurate.~~ process defined on a finite tree. Using a Bayesian framework, we can disentangle covariances in the data, which are result of their multivariate Gaussian equilibrium distribution, from those resulting from the historical correlations. Our approach leads to a clear gain in accuracy in the inferred equilibrium distribution, which corresponds to an effective two- to fourfold increase in sample size.

* Correspondence to: Pierre Barrat-Charlaix, **pierre.barrat@unibas.ch**

I. 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 the last years. 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, firing patterns of neurons, individuals in animal or human groups, stock prices etc. Within a static modeling approach, frequently based on the maximum-entropy approach, 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 some, more rare cases, data correspond to observed time series , but the theoretical and algorithmic development is much less advanced than in the case of independent static data, assuming, e.g., that measurements are taken at the microscopic time scale.

One of the biggest application areas of inverse statistical mechanics is the modeling of biological sequences, and we will use them to motivate our otherwise theoretical-methodological study. Typical datasets are so called homologous protein families, collecting ensembles of proteins of common evolutionary ancestry. Despite their strong divergence in amino-acid sequences, the proteins in such families have highly similar biological functions and three-dimensional fold structures. The different sequences can thereby, in a statistical-physics perspective, be seen as microscopically distinct solutions for the same macroscopic behavior.

In such protein families, which are typically represented by large multiple-sequence alignments, at least two complementary types of biological information are contained:

- *Phylogenetic information:* the distances between sequences carry information about the evolutionary time since the common ancestor. Using phylogenetic methods we may reconstruct the evolutionary history of a protein family (and of the species carrying the proteins), and also the sequences of the ancestral proteins.

- *Coevolutionary information* considers the correlated evolution of positions in proteins, frequently related to contacts in the folded protein, even for positions distant along the sequence. This information has been massively used in protein-structure prediction, most recently in combination with structurally supervised deep learning, but also to infer mutational landscapes or networks of interacting proteins.

These two types of information are contained in two complementary features of the data set: phylogenetic inference is based on the comparative analysis of the rows of the best, i.e., of entire proteins; and coevolutionary inference is based on the comparative analysis of MSA columns, which describe the amino-acid usage of the individual positions across proteins.

For methodological reasons, each type of inference assumes the absence of the other information: phylogenetic inference assumes independent evolution of all positions; coevolutionary inference assumes that proteins form an independently and identically distributed sample of the model to be inferred.

It has turned out to be hard to combine the two types of information in a single approach, even if the existence of, e.g, phylogenetic correlations between sequences influences the results of coevolutionary analysis. Here we propose to advance on the question of phylogeny-aware coevolutionary inference using the simplest nontrivial model setting: correlated multivariate Ornstein-Uhlenbeck (OU) processes.

In the last decades, biology has seen impressive progress in experimental techniques which has resulted in a large increase of available data. This is especially visible in the case of biological sequences, with databases now harboring a vast amount of high-quality DNA or protein sequences [1, 2]. This has in turn fuelled a development of quantitative methods to model forces guiding the evolution of sequences or other biological traits [REFS/EXAMPLES?]. A common idea in this context is that it is possible to use characteristics of homologous genes or organisms 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 [3, 4].

Models built in this way usually assume that biological characteristics of related organisms are evolving under the same constraints but are independent from each other in the statistical sense. This allows simple formulation for model learning. However, the phylogenetic relations between evolving entities are in direct contradiction with this simplifying hypothesis, and ignoring them can result in biases when learning models. For instance, it has been shown that phylogenetic relations between protein sequences induce non-trivial correlations that are not related to protein function [5].

In order to correct for these biases, it is necessary to disentangle the effects due to selection forces and those due to phylogeny. However, this becomes a hard problem when dealing with complex models where traits under study are not independent from each other. Accounting for these biases then requires one to understand how interdependent traits under selection evolve along a tree. A historically well-known way to represent such processes is to use Ornstein-Uhlenbeck dynamics (OU), which models traits as a Gaussian vector evolving in a quadratic potential that represents selection forces [6–8]. This method is commonly used in the field of phylogenetic comparative methods (PCM) [9, 10].

This modeling approach is *a priori* limited to continuous traits, but could potentially be used for protein sequences combined with a continuous-variable approximation, that has successfully been used in the past [11–13]. 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 [14–16].

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 II the main characteristics of the multivariate OU process. We then describe the setting of the inference problem that we want to solve in section III A, propose a solution in sections III B and III C. Finally, we present results obtained on simulated data in section IV, with the context of pairwise models of protein sequences in mind.

II. 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,

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

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 Eq. (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, Eq. (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 $\vec{x}_0$ some time Δt in the past. The Fokker-Planck equation corresponding to Eq. (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 Eq. (3) is a multivariate normal distribution [17]:

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

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

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

Eqs. (4) and (5) define a multivariate *Ornstein-Uhlenbeck* (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 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, one immediately recovers Eq. (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 Eqs. (1) and (4),

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

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 Eq. (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, one obtains the following covariance:

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

Eq. (7) allows us to readily distinguish two regimes. Let us call ρ_a the eigenvalues of $\boldsymbol{J}$. The eigenvalues of $\boldsymbol{\Lambda}^{\Delta t}\boldsymbol{C}$ are then equal to $\rho_a^{-1} e^{-\gamma\rho_a \Delta t}$. Since all ρ_a are positive, the eigenvalues of $\boldsymbol{\Lambda}^{\Delta t}\boldsymbol{C}$ vanish exponentially over time. The slowest timescale of exponential decay is set by $\tau_c^{-1} = \gamma\rho_{min}$, with ρ_{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. If this is verified for all pairs of observations $\vec{x}_i$ and $\vec{x}_j$, the regime is that of *uncorrelated* data – the inference of $\boldsymbol{J}$ can simply be performed by inverting the empirical covariance matrix extracted from the data. Inversely, for $\Delta t/\tau_c \ll 1$, $\vec{x}_1$ and $\vec{x}_2$ are highly correlated, defining a *strongly correlated* regime. 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 Eq. (7) becomes irrelevant. Actually, $\vec{x}_1$ and $\vec{x}_2$ coincide at that point, *i.e.* we have $P(\vec{x}_1, \vec{x}_2 |\, \Delta t = 0) = P_{eq}(\vec{x}_1) \times \delta(\vec{x}_1 - \vec{x}_2)$ using the L-dimensional Dirac distribution.

III. METHODS

A. 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. Fig. 1. Formally, we assume that

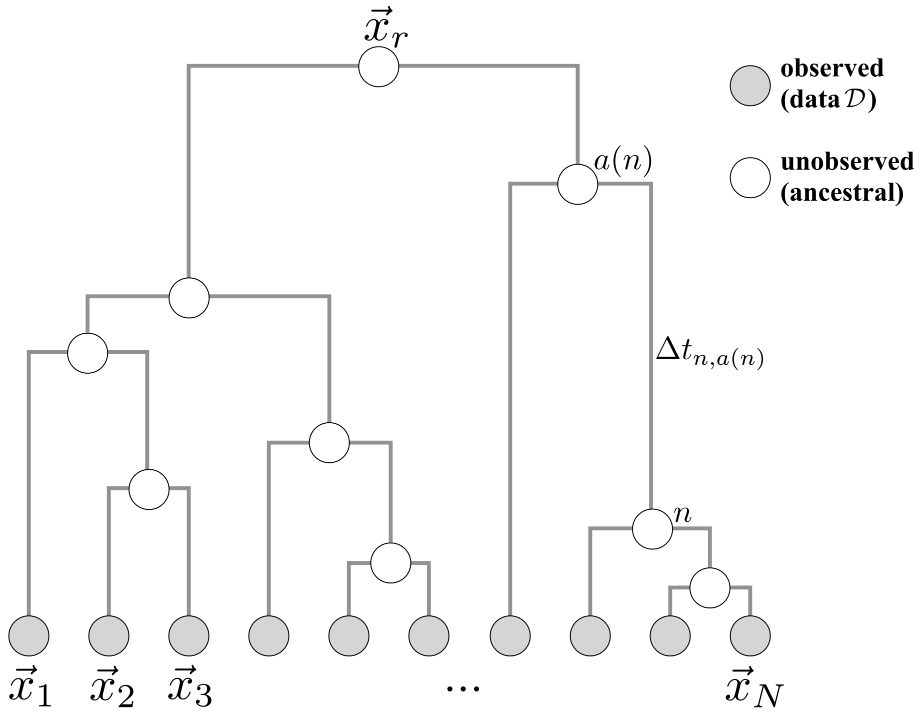

FIG. 1. Schematic representation of a tree $\mathcal{T}$ underlying the data generating process. The process starts at the root node r with a configuration $\vec{x}_r$ sampled from $P_{eq}(\vec{x}_r)$. The dynamics consist in independent realizations of the OU process on all branches from ancestral nodes $a(n)$ 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.

the data consists of N real-valued vectors of length L, denoted $\{\vec{x}_i\} \in \mathbb{R}^L$ with $i = 1, ..., N$. Taken individually, we assume that the $\vec{x}_i$ are distributed according to Eq. (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 $\vec{x} = (x^1, ..., x^L)$ 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 $a, b = 1, ..., L$. 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 $\{\vec{x}_i\}_{i=1,...,N}$ is the result of an Ornstein-Uhlenbeck (OU) process taking place on a tree $\mathcal{T}$, as is illustrated in Fig. 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 Eq. (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 Eq. (4)
- As a consequence, OU processes on branches stemming from 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 biologic traits along a phylogenetic tree, with the leave 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 $i, j \in \{1, ..., N\}$, is given by $P(\vec{x}_i, \vec{x}_j | \, \Delta t_{ij})$ (Eq. (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) is 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 of the OU process. 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. 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 one to hope of reconstructing 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 Eq. (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 $\mathcal{D}=\{\vec{x}_i\}_{i=1,\ldots,N}$ 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.

B. Calculation of the likelihood

The joint distribution of two configurations $\vec{x}_1$ and $\vec{x}_2$ separated by time Δt is given by Eq. (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 Eqs. (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 $\vec{X}=[\vec{x}_1,\ldots,\vec{x}_N]$ to be the concatenation of all configurations in our dataset $\mathcal{D}$, we can write the distribution of $\vec{X}$ as

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

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

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

where the (zero) marginals $\langle x_i^a \rangle$ and $\left\langle x_j^b \right\rangle$ 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 Eq. (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 $\boldsymbol{\Lambda} = e^{-\gamma \boldsymbol{C}^{-1}}$ 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 $\boldsymbol{C}$.

Eq. (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

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

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 Eq. (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. Using now Eq. (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

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

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 $\boldsymbol{G}^a$, $a = 1, ..., L$, with elements

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

In other words, for a given index $1 \leq a \leq L$, $\boldsymbol{G}^a$ 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 $\{\lambda_{ka}, \vec{u}_{ka}\}_{k=1,\ldots,N}$.

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 $\vec{s}_a$ and $\vec{u}_{ka}$, defining vectors $\vec{S}_{ka}$ of dimension

$L \times N$:

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

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 $1 \leq k \leq N$ and $1 \leq a \leq L$. Since $\mathbb{G}$ is the covariance matrix of a Gaussian distribution, we conclude the λ_{ka} to be strictly positive.

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 Eq. (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 diagonlize $\boldsymbol{C}^{-1}$ (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 section SA 4, this calculation simplifies even more when considering a fully balanced 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, Eq. (12) can now be rewritten using the eigen-decomposition of $\mathbb{G}$:

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

Eq. (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 Eq. (14) makes clear that its eigenmodes $\{\lambda_{ka}, \vec{u}_{ka}\}$ depend only of the eigenvalues ρ_a of $\boldsymbol{C}^{-1}$, on γ, as well as of 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 time scale parameter γ and of the pairwise distances on the tree Δt_{ij}.

C. 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 time scale γ. In Eq. ((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 know attempt to maximize the likelihood with respect to the eigenmodes $\{\rho_a, \vec{s}_a\}$ and to the time scale γ.

In order to perform this optimization, we need to compute the gradient of the likelihood with repsect 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 in using $L(L-1)/2$ scalar *Eulerian angles* $\{\theta_{\alpha\beta}\}$ with $1 \leq \alpha < \beta \leq L$ [18, 19]. 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, that we have found faster in practice, is to express the matrix of the $\{\vec{s}_a\}$ as the exponential of a skew-symetric matrix, see section A 3 of the SM. However, this parametrization does not allow a simple analytical expression of the gradient of the likelihood, and we use it along with automatic differentiation ~~.~~[20]. 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 $\{\rho_a, \theta_{\alpha\beta}\}$ 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 Eq. (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 Eq. (17) with respect to the parameters of interest. This gives us three equations:

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

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

and

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

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 [21]:

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

and

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

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 $\{\rho_a, \theta_{\alpha\beta}\}$ 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 [22]. Details are presented in section A 5 of the SM.

IV. RESULTS

In order to evaluate our inference procedure, we generate artificial data corresponding to the process described in section III A. We first build a balanced binary tree $\mathcal{T}$ with $2^9 = 512$ leaves. 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 ~~coupling matrix with approximative~~ sparsified coupling matrix of approximate density 0.3. ~~Null~~ Zero elements of J correspond to variables that do not interact, ~~or~~ 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:

$$\gamma_d = \frac{1}{\Delta t_{av} \rho_{min}} \tag{23}$$

where Δt_{av} is the average branch length separating two leaves of $\mathcal{T}$. For $\gamma \gg \gamma_d$, leaf configurations are on average well decorrelated, whereas for $\gamma \ll \gamma_d$ all leaves will be strongly correlated. By simulating data using different γ in the range $[10^{-2}, 2] \cdot \gamma_d$, we investigate all relevant temporal regimes. For each value of γ, we then sample configurations of leaves of $\mathcal{T}$ using the process described in section A 1 of the supplementary material. 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 $\boldsymbol{C}_{max}$. As a means of comparison, we also compute the empirical covariance matrix $\boldsymbol{C}_{emp}$ as if leaf configurations were independent. ~~Figure~~ Fig. 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 ~~section A 6of the smallest~~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 signal, and both methods perform ~~as~~ equally poorly. However, there exists an intermediate regime where $\boldsymbol{C}_{max}$ is much closer to the actual correlation than $\boldsymbol{C}_{emp}$. In ~~figure~~ Fig. 3, we plot the relative $l2$-error between either covariance matrices in the left panel or coupling matrices in the right panel. 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 $\boldsymbol{C}_{max}$ and $\boldsymbol{C}_{emp}$ being close to singular in this case.

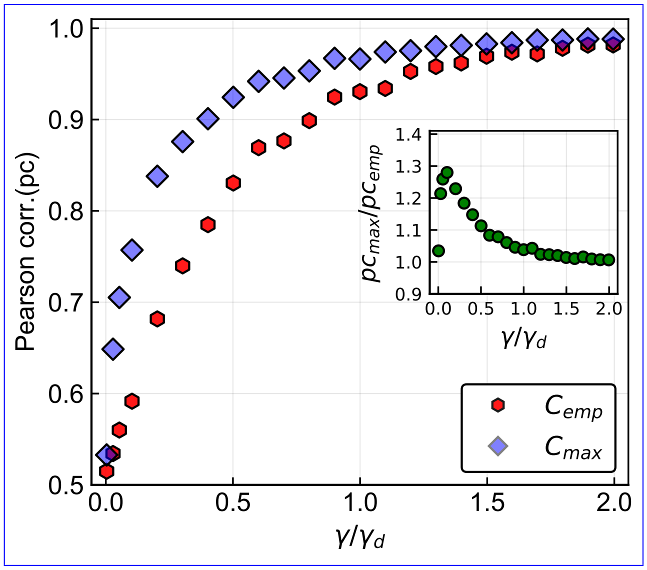

FIG. 2. Pearson correlation between empirical /maximum-likelihood covariance matrices and the true covariance matrix. The inset plot represents the ratio between the Pearson correlation for the maximum-likelihood covariance matrix and the one for the empirical covariance matrix. Simulations are performed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

An interesting way to illustrate the benefits of reconstructing the covariance matrix using knowledge of the tree is to evaluate the ~~effective number of samples that is gained by doing so. Figure S1 shows~~ gain in *effective sample size.* Intuitively, the use of correlated samples

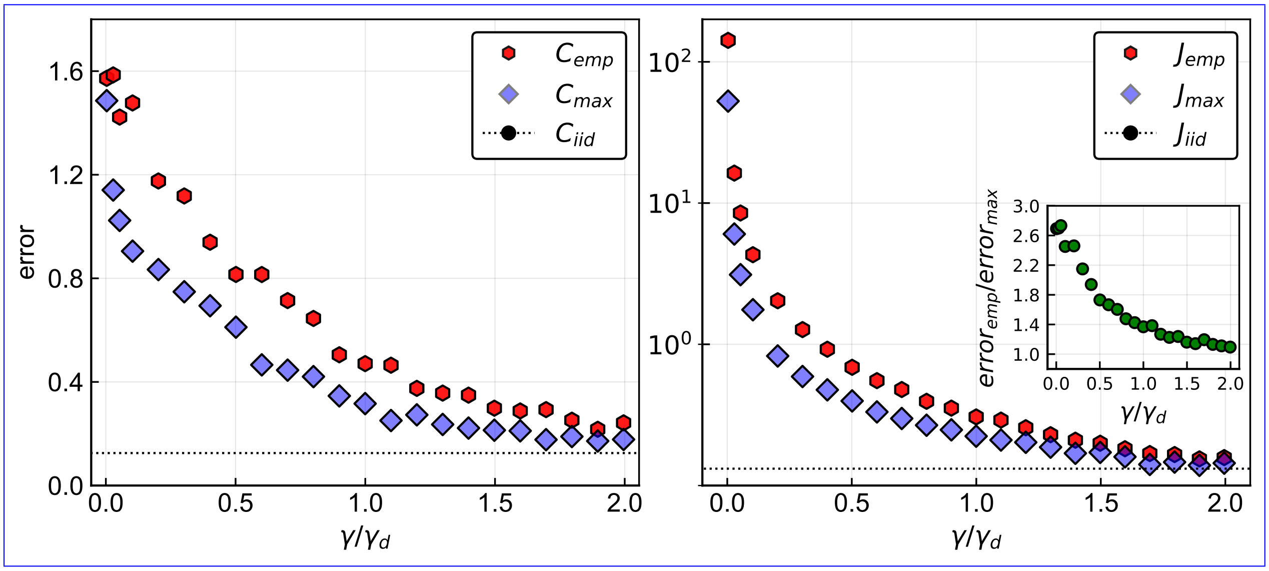

FIG. 3. **Left:**Relative $l2$-error between empirical or maximum-likelihood covariance matrices and the true covariance matrix. **Right:**Relative $l2$-error between empirical /maximum-likelihood coupling matrices and the true coupling matrix. Logarithmic scale is chosen for the y-axis because of large values of the error at low γ. The inset in both panels show the ratio between the two errors.

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 on smaller but $i.i.d.$ samples. To do so, we report in Fig. S1 the $l2$-error between ~~the true covariances and the~~ true and empirical covariances computed from ~~an~~ a $i.i.d.$ ~~sample of size~~ samples of variable sizes N. As expected~~the error is increasing for~~ , the error increases with decreasing values of N. We can use this ~~as a calibration~~ in turn to express values of the $l2$-error in ~~terms of an effective sample size~~correlated samples in terms of effective $i.i.d.$ sample sizes. For example, the error reached by $\boldsymbol{C}_{emp}$ for $\gamma/\gamma_d \in [0.5, 1]$ and $L = 4$ corresponds to the one obtained for an $i.i.d.$ sample of size ~ 16, whereas it corresponds to a sample of size $\sim 32 - 64$ for $\boldsymbol{C}_{max}$. Thus, our correction is equivalent to ~~doubling~~ increasing by a factor 2-4 the number of effective samples.

Finally, we assess the performance of our method in improving the prediction of ~~interactions between gaussian~~ 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 $J_{ab} \neq 0$. ~~We~~ 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

~~Figure~~ Fig. 4 shows the PPV as a function of the number of predictions for different values of γ and $L = 4$ (see Fig. S7 for the $L = 10$ case). In this case, the coupling matrix only has 6 independent non-diagonal elements, and only 6 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 $\gamma/\gamma_d \in [0.5, 1]$.

V. DISCUSSION

In this work, we proposed a method for inferring parameters of an Ornstein-Uhlenbeck process using data that is 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 ~~only~~ 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 ~~branch~~ 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

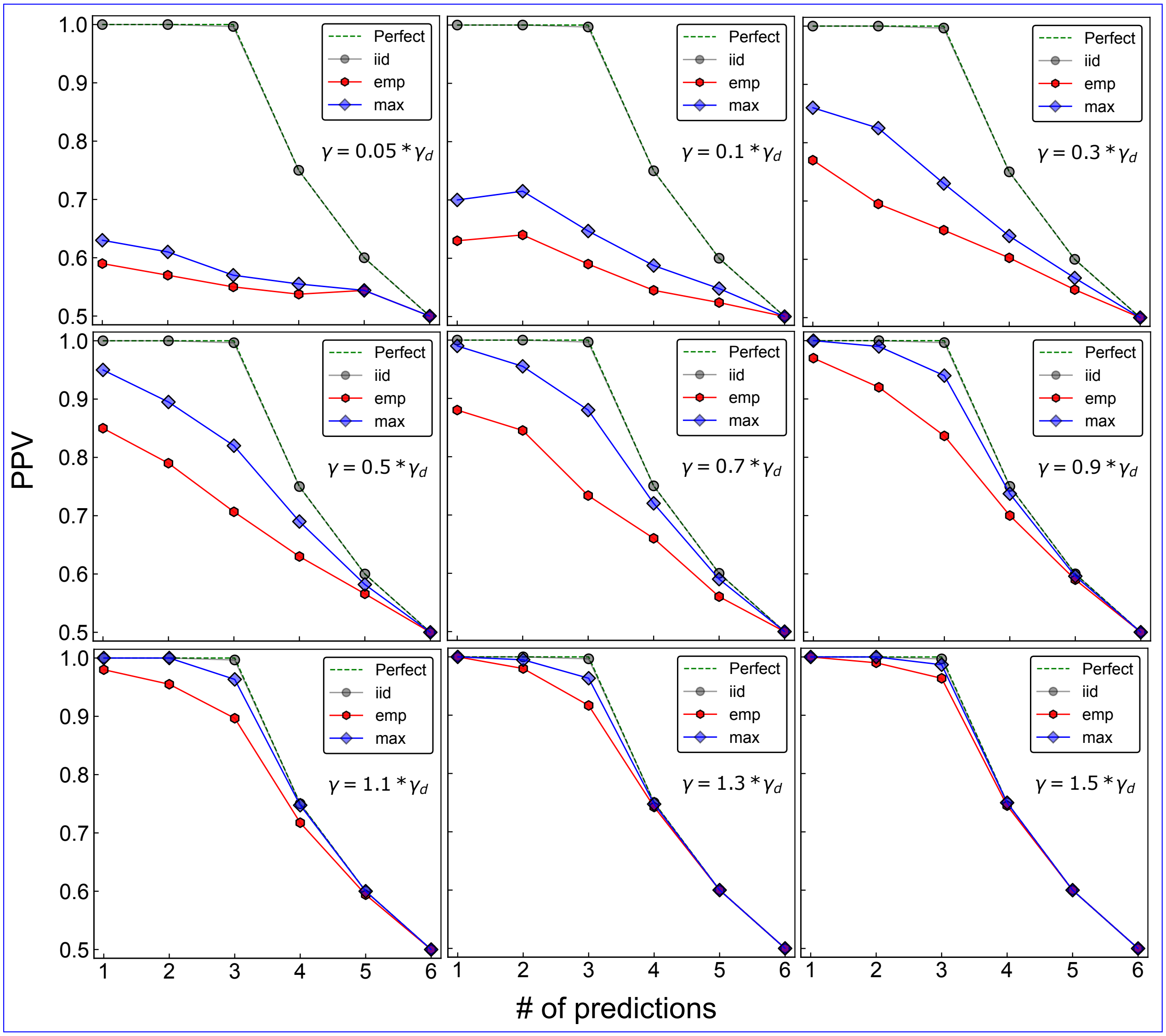

FIG. 4. Quality of prediction of interactions for different values of γ and system size $L = 4$. Interactions are defined as non-zero elements of the coupling matrix. In the $L = 4$ case, there are 6 possible interactions. Predictions are made by taking the largest elements (in absolute terms) of the inferred coupling matrix. The PPV is the fraction of correctly predicted contacts for a given number of predictions.

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 ~~expected~~ 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 ~~is too high~~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 ~~the intermediate regime , our~~ an intermediate regime where intrinsic and historical correlations in the dataset coexist, our tree-aware re-construction of $\mathbf{J}$ ~~that takes the tree into account~~ results in clear benefits over ~~an~~ 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 [10] 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 proteins sequences. Several results in the last years have shown that selection forces shaping the evolution of protein sequences are well described by a pairwise potential. 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 Ornstein-Uhlenbeck 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 [11–13]. Its use in the context of modelling the evolutionary dynamics of protein sequences remains an open question.

Acknowledgments: We acknowledge interesting discussions with Roberto Mulet. PBC and MW acknowlege 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.

[1] Eric W Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi. GenBank. *Nucleic Acids Research*, 47(D1):D94–D99, January 2019.

[2] The UniProtConsortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699, March 2018.

[3] Ronald M Levy, Allan Haldane, and William F Flynn. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current Opinion in Structural Biology*, 43:55–62, April 2017.

[4] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Remi Monasson, and Martin Weigt. Inverse Statistical Physics of Protein Sequences: A Key Issues Review. *Reports on Progress in Physics*, 81(3):032601, March 2018. arXiv: 1703.01222.

[5] Chongli Qin and Lucy J. Colwell. Power law tails in phylogenetic systems. *Proceedings of the National Academy of Sciences*, 115(4):690–695, January 2018.

[6] G. E. Uhlenbeck and L. S. Ornstein. On the Theory of the Brownian Motion. *Physical Review*, 36(5):823–841, September 1930. Publisher: American Physical Society.

[7] Joseph Felsenstein. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1):445–471, November 1988. Publisher: Annual Reviews.

[8] Thomas F. Hansen. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution*, 51(5):1341–1351, 1997. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1558-5646.1997.tb01457.x.

[9] Krzysztof Bartoszek, Jason Pienaar, Petter Mostad, Staffan Andersson, and Thomas F. Hansen. A phylogenetic comparative method for studying multivariate adaptation. *Journal of Theoretical Biology*, 314:204–215, December 2012.

[10] Venelin Mitov, Krzysztof Bartoszek, Georgios Asimomitis, and Tanja Stadler. Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theoretical Population Biology*, 131:66–78, February 2020.

[11] David T. Jones, Daniel W. A. Buchan, Domenico Cozzetto, and Massimiliano Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190, January 2012.

[12] J. P. Barton, S. Cocco, E. De Leonardis, and R. Monasson. Large pseudocounts and L 2 -norm

penalties are necessary for the mean-field inference of Ising and Potts models. *Physical Review E*, 90(1), July 2014.

[13] Carlo Baldassi, Marco Zamparo, Christoph Feinauer, Andrea Procaccini, Riccardo Zecchina, Martin Weigt, and Andrea Pagnani. Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLoS ONE*, 9(3), March 2014.

[14] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, December 2011.

[15] Matteo Figliuzzi, Herv Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, 33(1):268–280, January 2016.

[16] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, July 2020. Publisher: American Association for the Advancement of Science Section: Report.

[17] Rajesh Singh, Dipanjan Ghosh, and R. Adhikari. Fast bayesian inference of the multivariate ornstein-uhlenbeck process. *arxiv:1706.04961*, 2017.

[18] Richard C. Raffenetti and Klaus. Ruedenberg. Parametrization of an orthogonal matrix in terms of generalized eulerian angles. *International Journal of Quantum Chemistry, Vol.III s,625-634*, 1970.

[19] Ron Shepard, Scott R. Brozell, and Gergely Gidofalvi. The representation and parametrization of orthogonal matrices. *Journal of Physical Chemistry A, 119,7924-7939*, 2015.

[20] Michael Innes. Don't unroll adjoint: Differentiating ssa-form programs. *CoRR*, abs/1810.07951, 2018.

[21] Kaare Brandt Petersen and Michael Syskind Pedersen. *The Matrix Cookbook*. 2015.

[22] Steven G. Johnson. The nlopt nonlinear-optimization package. *http://github.com/stevengj/nlopt.*

Appendix A: Supplementary material

1. Generating artificial data

We are interested in the case where the described Ornstein-Uhlenbeck 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, 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, the time Δt then being the branch length connecting them.

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 in equation (4) by 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 with distributions $\mathcal{N}(0,1)$. Moreover, if $\vec{x_0}$ is distributed according to equation (1), then $\vec{x}$ and $\vec{x}_0$ are distributed according to equation (6). Note that equation (A1) is quite different from the Langevin equation (2), even though they have similar forms. While the Langevin equation describes the motion of $\vec{x}$ in the potential $\boldsymbol{J}$, eq. (A1) directly samples from the OU process. Given an already sampled internal node in the tree, equation (A1) allows to sample a configuration for each of its children. To sample the whole tree, we first sample the root note $\vec{x}_0$ from the equilibrium distribution (1). By recursive applications of (A1), we then simply work our way down the tree until all leaves are sampled.

2. Initializing parameters

a. Eigenvalues and eigenvectors of $\mathbf{C}^{-1}$

The initial value that we take for the covariance matrix is the empirical one

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

Its eigenmodes $\{\rho_a^0, \vec{s}_a^{\,0}\}$ determine the starting point of the optimization.

[Insert here how we go from $\vec{s}_a^{\,0}$ to the corresponding set of Eulerian angles or to the corresponding skew symmetric matrix.]

b. Time scale parameter γ

The optimization also requires that we initialize the time scale γ. For this, we try to find the optimal γ given the data $\mathbf{X}$, 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 calculate analytically the average of the scalar product $\vec{x}_i^T\vec{x}_j$:

$$\begin{aligned}\langle \vec{x}_i^T\vec{x}_j\rangle_P &= \int \mathrm{d}\vec{x}_i \mathrm{d}\vec{x}_j P(\vec{x}_i,\vec{x}_j|\Delta t_{ij}) \sum_{a=1}^{L} x_i^a x_j^a \\ &= \sum_{i=a}^{L} \langle x_i^a x_j^a\rangle_P.\end{aligned}$$

The covariance $\langle x_i^a x_j^a\rangle_P$ between observations separated by a time Δt_{ij} is given by equation (7) of the main text. Using this, we now have

$$\begin{aligned}\langle \vec{x}_i^T\vec{x}_j\rangle_P &= \sum_{a=1}^{L} \big(\mathbf{\Lambda}^{\Delta t_{ij}}\mathbf{C}\big)_{aa} \\ &= \mathrm{Tr}\big(\mathbf{\Lambda}^{\Delta t_{ij}}\mathbf{C}\big) \\ &= \sum_{a=1}^{L} \rho_a^{-1} e^{-\gamma\rho_a\Delta t_{ij}}. \qquad \text{(A2)}\end{aligned}$$

Having initialized the covariance matrix $\mathbf{C}$ at its empirical value, we know the values of all members of the r.h.s. of equation (A2) except that of γ. On the other hand, equivalent versions of equation (A2) can be written for all pairs of configurations i and j. To find an initial value of γ which is consistent with the data and the empirical covariance matrix, 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 $F(\gamma)$:

$$F(\gamma) = \sum_{1\le i<j\le N} \left[\vec{x}_i^T\vec{x}_j - \sum_{a=1}^{L} \rho_a^{-1} e^{-\gamma\rho_a\Delta t_{ij}}\right]. \qquad \text{(A3)}$$

As F depends on one scalar parameter, it is straightforward to minimize it, allowing us to initialize γ to a reasonable value.

3. Parametrizations of eigenvectors

a. Parametrization in term of generalized Eulerian angles

The aim is to parameterize each base vector $\vec{s}_a$ according to the $a-th$ column of an arbitrary orthogonal matrix $\boldsymbol{S}$ parameterized in terms of $L(L-1)/2$ independent variables $\theta_{pq}, p = 1, 2, ..., L; q = 1, 2, ..., L; p < q$, named Generalized Eulerian angles (Eulerian angles is for L=3).

To construct this matrix we start from a transformation in a two-dimensional subspace of an $L - dimensional$ space which is given by an L-dimensional matrix of the form:

$$\boldsymbol{T}_{pq} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \cos\theta_{pq} & & \sin\theta_{pq} & \\ & & & 1 & & \\ & & -\sin\theta_{pq} & & \cos\theta_{pq} & \\ & & & & & 1 \end{pmatrix} \tag{A4}$$

where all diagonal elements are unity except the diagonal elements in the pth column and the qth column, which are $\cos\theta_{pq}$ and all off-diagonal elements are zero except 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 is $-\sin\theta_{pq}$. There are $L(L-1)/2$ matrices of the form indicated in Equation (1), corresponding to all choices of p and q with $p < q$.

An arbitrary $L - dimensional$ orthogonal matrix $\boldsymbol{S}$ 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}.

$$\boldsymbol{S}(\boldsymbol{\theta}) = \prod_{\{pq\}} T_{pq}(\theta_{pq}) \tag{A5}$$

This matrix is in charge of rotations of vectors in $L-\mathit{dimensional}$ space. For $L=3$ we get the classical Eulerian matrix.

In reference [7] is explained an recursive algorithm to efficiently perform this multiplication, as well as the construction of the matrix derivatives respect to parameters θ_{pq}.

For the first, equation (2) is transformed in the recurrence equations:

$$\boldsymbol{S}(\vec{\theta}) = \boldsymbol{T}^{(L)} \tag{A6}$$

with $\theta_{pq}, p=1,2,...,L; q=1,2,...,L; p<q$ and

$$T^{(n)}_{kl} = \cos\theta_{kn} * t^{(n)}_{kl} - \sin\theta_{kn} * s^{(n)}_{kl} \tag{A7}$$

with

$$\begin{aligned} s^{(n)}_{k+1,l} &= \sin\theta_{kn} * t^{(n)}_{kl} + \cos\theta_{kn} * s^{(n)}_{kl} \\ s^{(n)}_{k+1,n} &= \cos\theta_{kn} * s^{(n)}_{kn} \\ s^{(n)}_{1,l} &= -\delta_{ln} \\ \theta_{nn} &= \pi/2 \end{aligned} \tag{A8}$$

and

$$\boldsymbol{t}^{(n)} = \begin{pmatrix} \boldsymbol{T}^{(n-1)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \tag{A9}$$

This mean that the eigenvectors $\vec{s}_a$ of the matrix $\boldsymbol{C}$ with dimensions $L \times L$ can be written as:

$$s^k_a = T^{(L)}_{.,a} = \cos\theta_{kL} * t^{(L)}_{ka} - \sin\theta_{kL} * s^{(L)}_{ka} \quad \text{for } k=1,..,L \tag{A10}$$

Then there is another important problem : how to determine parameters $\boldsymbol{\theta}$ given an orthogonal matrix $\boldsymbol{Q}$ such that all equations

$$S_{ij}(\boldsymbol{\theta}) = Q_{ij}$$

are satisfied. This system of nonlinear transcendental equations can't be algebraically solved, however it's possible overcome this issue finding the set of $\boldsymbol{\theta}$ which minimize the function:

$$f(\boldsymbol{\theta}) = \sum_{i,j} [Q_{ij} - S_{ij}(\boldsymbol{\theta})]^2 \tag{A11}$$

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

b. Parametrization in term of the exponential of a skew-symmetric matrix

The exponential of a skew-symmetric matrix $\boldsymbol{X} = -\boldsymbol{X}^T$:

$$\boldsymbol{S} = \exp(\boldsymbol{X}) \tag{A12}$$

is an orthogonal matrix : $\exp(X)^T = \exp(X^T) = \exp(-X) = \exp(X)^{-1}$.

The derivatives on $L(L-1)/2$ independent variables of $\boldsymbol{X}$ is formally defined by

$$\frac{\partial \boldsymbol{S}}{\partial X_{jk}} = \lim_{h\to 0} \frac{1}{h}\big(\exp(\boldsymbol{X} + h\boldsymbol{E}^{jk}) - \exp(\boldsymbol{X})\big) \tag{A13}$$

where $\boldsymbol{E}^{jk}$ for $j > k$ is defined as a skew-symmetric matrix that has only two nonzero elements:

$$E^{pq}_{jk} = \delta_{pj}\delta_{qk} - \delta_{pk}\delta_{qj} \tag{A14}$$

To obtain a skew-symmetric matrix $\boldsymbol{X}$ from an orthogonal matrix $\boldsymbol{Q}$ is enough invert the exponential relation :

$$\boldsymbol{X} = \log \boldsymbol{Q}$$

If we write eigenvectors of the likelihood function as the columns of the exponential of the skew-symmetric matrix $\boldsymbol{X}$, then we are able to perform the optimization over it's $L(L-1)$ independent variables of $\boldsymbol{X}$.

Furthermore, as the eigenvectors were parameterized by an algebraic function, the complete likelihood is described in terms of arithmetics operation and elementary functions. This allow to compute the gradient via automatic differentiation (AD), making computations in our optimization process faster. The idea is to represent a function as a computational graph, also called a Wengert list , where each node in this list will represent an intermediate result of the computation. The intermediate results can then be assembled using the chain rule to get the final derivative were looking for. There is two main algorithms to traverses the chain rule in AD: forward mode and reverse mode, here we used reverse mode algorithm which is the of choice for back-propagation in deep learning . In particular we implemented it using the Julia package Zygote.jl .

4. Homogeneous and fully balanced tree

Let's assume that the tree is binary, symmetric and completely homogeneous with all branches having the same length Δt. As an example, the covariance matrix for such a tree with $K = 2$ branching events and four leaves is

$$\mathbb{G} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{C}\Lambda^{2\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} \\ \boldsymbol{C}\Lambda^{2\Delta t} & \boldsymbol{C} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} \\ \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C} & \boldsymbol{C}\Lambda^{2\Delta t} \\ \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{4\Delta t} & \boldsymbol{C}\Lambda^{2\Delta t} & \boldsymbol{C} \end{pmatrix}. \tag{A15}$$

The associated matrix $\boldsymbol{G}^a = z(\rho_a, \gamma, \Delta t)$ defined in Eq. (A15) becomes

$$\boldsymbol{G}^a = \begin{pmatrix} \rho_a^{-1} & \rho_a^{-1}e^{-2\gamma\rho_a\Delta t} & \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} & e^{-4\gamma\rho_a\Delta t} \\ \rho_a^{-1}e^{-2\gamma\rho_a\Delta t} & \rho_a^{-1} & \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} & \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} \\ \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} & \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} & \rho_a^{-1} & \rho_a^{-1}e^{-2\gamma\rho_a\Delta t} \\ \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} & \rho_a^{-1}e^{-4\gamma\rho_a\Delta t} & \rho_a^{-1}e^{-2\gamma\rho_a\Delta t} & \rho_a^{-1} \end{pmatrix}. \tag{A16}$$

For hyper-geometric matrices as (A16) of dimension 2^K, there are $K+1$ different eigenvalues given by:

$$\lambda_k(\rho_a, \gamma) = \rho_a^{-1} * \begin{cases} (1 + \sum_{l=1}^{k-1} 2^{l-1} e^{-2l\gamma\rho_a\Delta t} - 2^{k-1} e^{-2k\gamma\rho_a\Delta t}) & k \in [1, K] \\ (1 + \sum_{l=1}^{K} 2^{l-1} e^{-2l\gamma\rho_a\Delta t}) & k = K+1 \end{cases} \tag{A17}$$

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

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

where $\Gamma(\vec{u}_k)$ represents the d_k combinations obtained by shifting the block of length Q, generating all eigenvectors corresponding to the eigenvalue λ_k. The eigenvectors are orthogonal to each other, and can be normalized and arranged horizontally into a matrix U.

To compute the gradient of the likelihood, we need derivatives of $\lambda_k(\rho_a, \gamma)$ with respect to ρ_a and γ, which can be directly obtained from expression (A17).

5. Optimization scheme

The proposed inference scheme was transformed into a multidimensional nonlinear optimization problem for which a variant of quasi-Newton Methods (QNMs) was used. The main feature in QNMs is that Hessian matrix $\hat{H}$ does not need to be computed instead its approximated. The Hessian approximation $\hat{H}$ is chosen to satisfy the secant equation:

$$\nabla L(\vec{x}_k + \Delta\vec{x}) = \nabla L(\vec{x}_k) + \hat{H}\Delta\vec{x}$$

for $n - dimensions$ $\hat{H}$ is undetermined, the various QNMs differ in their choice of the solution to the secant equation. We used LBFGS variant (limited memory version of BFGS), this particular method is based in choosing $\hat{H}$ as a positive definite matrix where

$$\hat{H}_{k+1} = \hat{H}_k + \frac{y_k y_k^T}{y_k \Delta x_k} - \frac{\hat{H}_k \Delta x_k (\hat{H}_k \Delta x_k)^T}{\Delta x_k^T \hat{H}_k \Delta x_k} \quad \text{and} \quad y_k = \nabla L(\vec{x}_k + 1) - \nabla L(\vec{x}_k)$$

The method implementation was done using NLopt package ~~(see reference)~~ [22] .

6. Supplementary figures

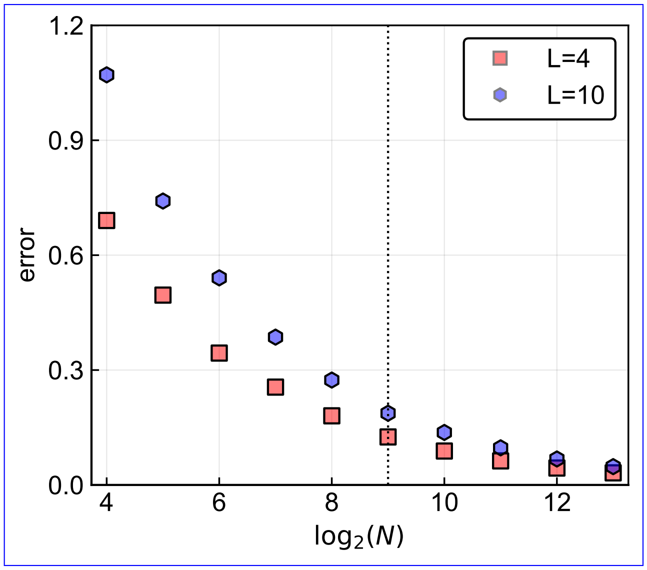

Figure S 1. Relative $l2$-error between the empirical covariance matrix calculated from an *i.i.d.* sample and the true covariance matrix, for system sizes $L = 4$ and $L = 10$. The dashed vertical line corresponds to the number of leaves of the tree used in the simulations.

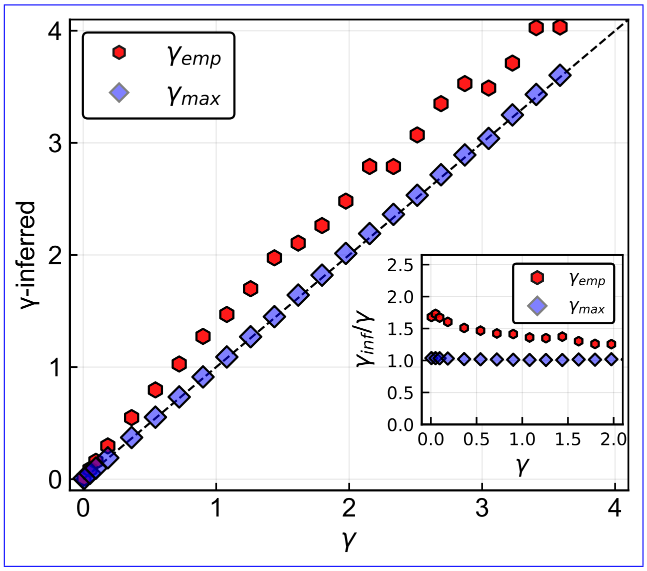

Figure S 2. Inferred γ values as a function of real γ, for system size $L = 4$. γ_{emp} is the value obtained by the process described in section A 2 of the SM. γ_{max} is the value inferred by the maximum-likelihood calculation. The inset represents the ratio of both inferred parameters γ_{emp} or γ_{max}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

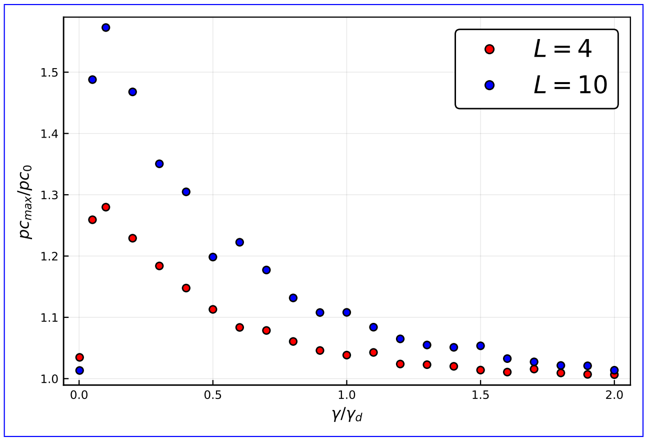

Figure S 3. Ratio between Pearson correlation of the maximum-likelihood and true covariance matrix to the Pearson correlation of the empirical and true covariances matrices for the two system sizes $L = 4$ and $L = 10$.

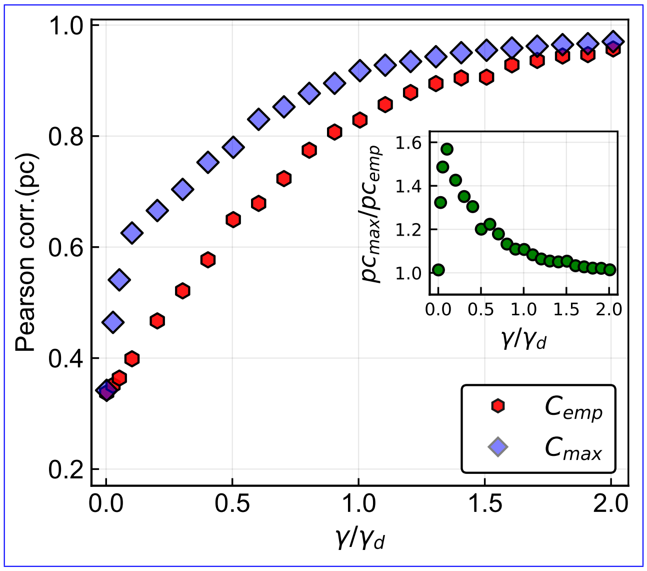

Figure S 4. Pearson correlation between empirical /maximum-likelihood covariance matrices and the true covariance matrix, the inset plot represent the ratio between the person correlation for the maximum-likelihood covariance matrix and the one for the empirical covariance matrix.

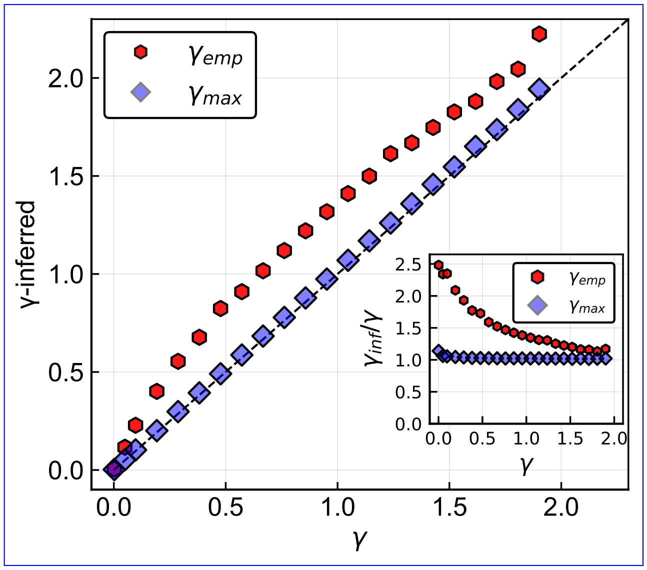

Figure S 5. Inferred γ values as a function of real γ, for system size $L = 10$. γ_{emp} is the value obtained by the process described in section A 2 of the SM. γ_{max} is the value inferred by the maximum-likelihood calculation. The inset represents the ratio of both inferred parameters γ_{emp} or γ_{max} to the real γ.

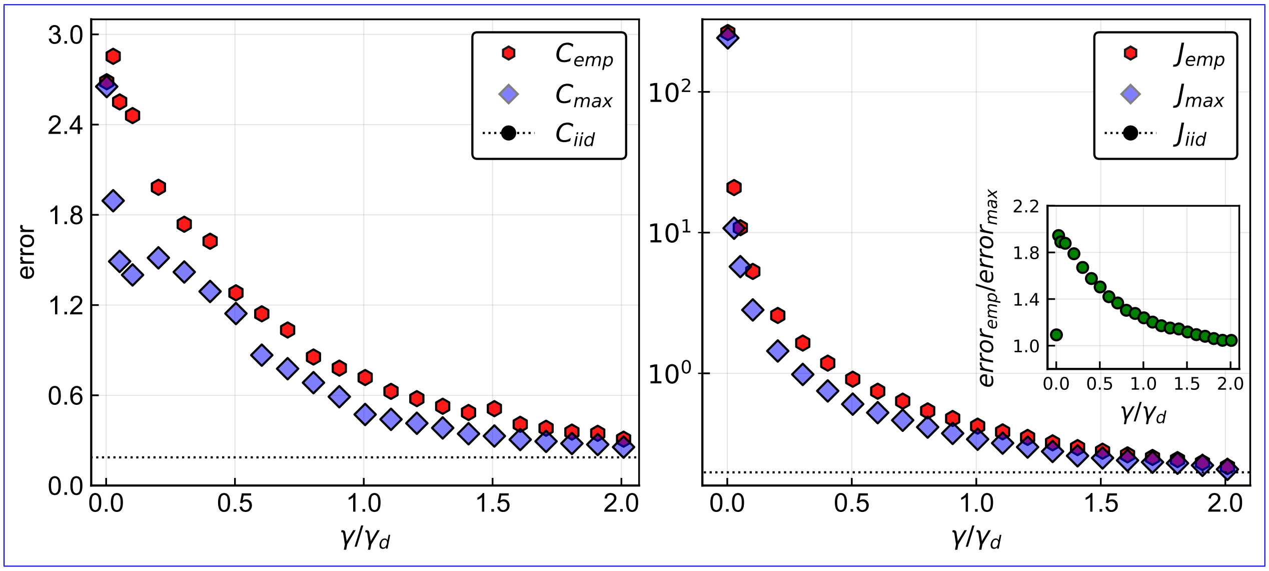

Figure S 6. **Left:** Relative $l2$-error between empirical or maximum-likelihood covariance matrices and the true covariance matrix. **Right**: Relative $l2$-error between empirical /maximum-likelihood coupling matrices and the true coupling matrix. Logarithmic scale is chosen for the y-axis because of large values of the error at low γ. The inset in both panels show the ratio between the two errors. For system size $L = 10$.

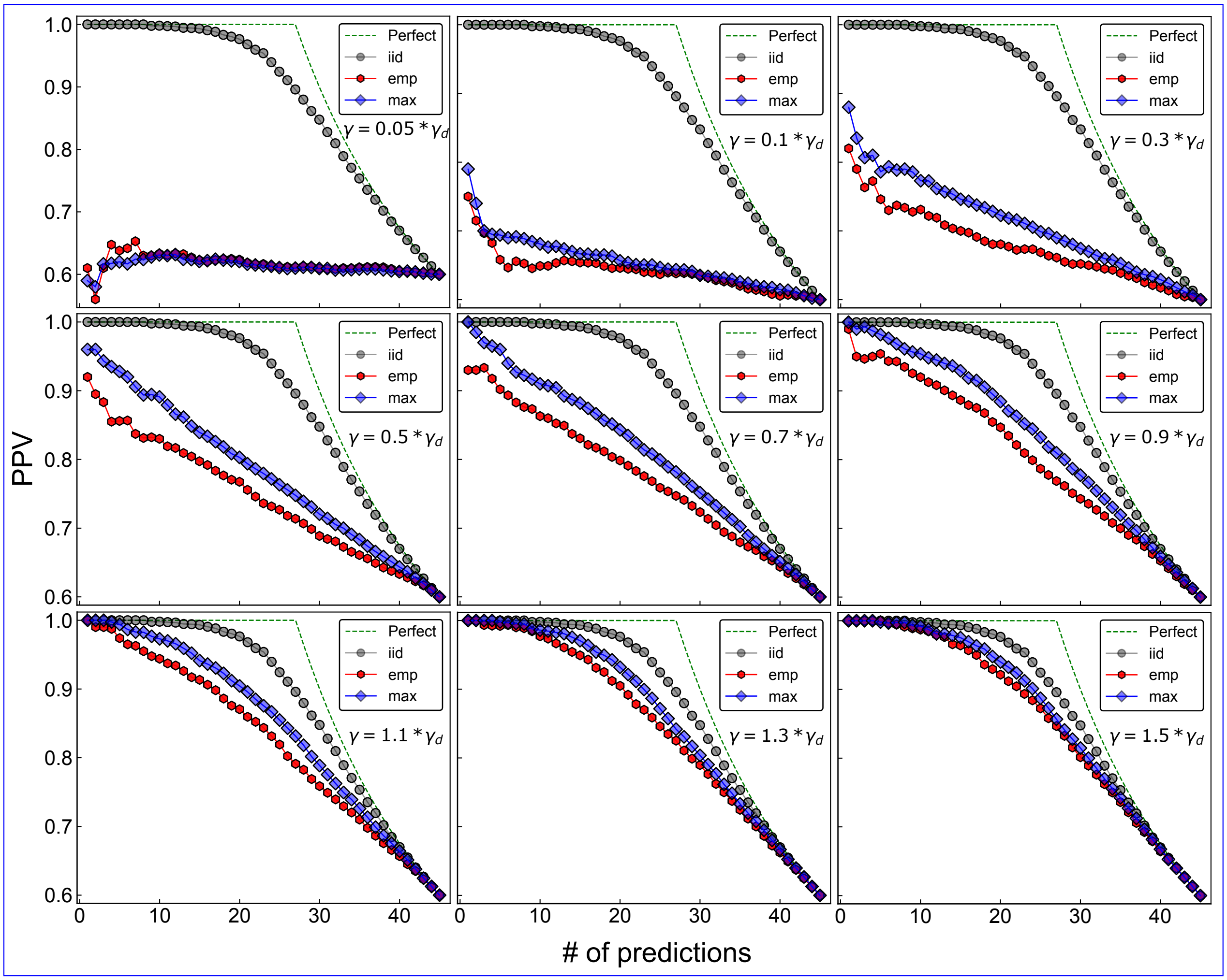

Figure S 7. Quality of prediction of interactions for different values of γ and system size $L = 4$. Interactions are defined as non-zero elements of the coupling matrix. In the $L = 4$ case, there are 6 possible interactions. Predictions are made by taking the largest elements (in absolute terms) of the inferred coupling matrix. The PPV is the fraction of correctly predicted contacts for a given number of predictions.