

# Supporting Information

Qin and Colwell 10.1073/pnas.1711913115

## Simulation Details

**Structural Constraints.** To simulate a proxy for phenotype in protein sequence data that reflects the structural constraints that can be identified using covariance analysis, we adopted a generalized Potts model with pairwise interactions between sequence positions drawn, for example, from a protein contact map. To parameterize a Potts model for sequences of length  $p$  amino acids, we need to specify an interaction matrix  $J$  of size  $p \times p$ , in addition to a mutation matrix  $\Theta$  of size  $q \times q$  that describes transitions between the residues. For protein sequences  $\Theta$  can in principle capture the interactions between amino acids, for example that two amino acids with similar charges will repel each other. This is similar to mutation matrices such as PAM or BLOSUM that are used by other evolutionary models (1), except that here we have the freedom to choose different mutation matrices for different pairs of sequence positions. In the work described in this paper, we do not specialize to a particular choice of mutation matrix to avoid introducing additional potential sources of covariance, but instead make the simplifying assumption that all amino acid mutations are equally likely. The associated amino-sequence probability space is given by

$$E(x) = - \sum_{a,b=1}^q \sum_{i < j} J_{ij} \Theta_{ab} \delta(x_i = a) \delta(x_j = b)$$

$$P(X = x) = \frac{1}{\mathcal{Z}} e^{-E(x)}, \quad [\text{S1}]$$

where  $E(x)$  is the energy of  $x$  and  $\mathcal{Z}$  is the partition function. For the matrix  $\Theta_{a,b}$ , we generalize the Ising model as a system of interacting spins (parallel and antiparallel) by considering a set of spins equally spaced in a circle. This gives the planar Potts model which extends the binary spin states to  $q$  spin states as

$$\Theta_{a,b} = \cos(2\pi(a-b)/q),$$

where  $a, b \in \{1, \dots, q\}$ . We first evolve a randomly generated starting sequence through a number (500) of proposed mutations to ensure that each starting sequence will “see” and be constrained by all of the interactions. Mutations are proposed at random, and the energy change that would result from the proposed mutation is calculated according to Eq. S1. A mutation is accepted with probability

$$P(\text{mutation accepted}) = \min\left(1, e^{-\Delta E}\right),$$

where  $\Delta E$  is the change in energy caused by the mutation. If there are no structural constraints, i.e.,  $J_{ij} = 0 \forall i, j$ , then all mutations will be accepted.

**Expected Covariance.** To estimate the expected covariance for our model, we directly calculate the value by considering the sequence probability distribution shown in Eq. S1. To analyze a single pairwise interaction, we first map the  $q$  states (representing the different amino acids) onto the unit circle of the complex plane, so that each of these state has the same magnitude. The mapping is explicitly

$$F : \{1, \dots, q\} \mapsto \{1, e^{ik\theta}, \dots, e^{i(q-1)k\theta}\}, \quad [\text{S2}]$$

where  $k = 2\pi/q$ . Using this mapping we find that

$$E(x_i x_j) - E(x_i)E(x_j) = \sum_{a,b=1}^q e^{i2\pi(a-b)/q} \frac{e^{J_{ij}\Theta_{ab}}}{\sum_{a,b} e^{J_{ij}\Theta_{ab}}}$$

$$= \frac{1}{\mathcal{Z}_{ab}} \frac{\partial \mathcal{Z}_{ab}}{\partial J_{ij}}, \quad [\text{S3}]$$

where  $\mathcal{Z}_{ab} = \sum_{a,b=1}^q e^{J_{ij}\Theta_{ab}}$ . An important property of Eq. S3 is that as  $J_{ij} \rightarrow \pm\infty$ ,  $E(x_i x_j) \rightarrow \pm 1$ . In other words, the magnitude of the covariance saturates as the strength of the interaction increases and cannot exceed one. If we consider the binary-state case, then Eq. S3 becomes

$$E(x_i x_j) = \frac{1}{e^{J_{ij}} + e^{-J_{ij}}} \frac{\partial}{\partial J_{ij}} \left( e^{J_{ij}} + e^{-J_{ij}} \right)$$

$$= \tanh(J_{ij}) \quad [\text{S4}]$$

which saturates at large values of  $J_{ij}$ . Using this model, the question of what covariance we can expect from interactions drawn from a protein contact map is briefly addressed in the main text.

The saturation of the magnitude of the covariance implies that the largest eigenvalue caused by structural interactions also saturates even as the number and strength of these interactions are allowed to increase. Fig. 4A of the main text shows this saturation as a function of the number of interactions that are included, where the maximum interaction strength is held constant. In Fig. S1 we further explore this phenomenon by varying both the number of interactions included and the magnitude of the maximum interaction strength included. In these simulations the interaction strengths are uniformly distributed on the interval  $[-s, s]$ , where  $s$  is the interaction strength indicated on the  $y$  axis. Fig. S1 shows that the maximum eigenvalue also saturates as the interaction strength increases.

For protein structure prediction, we use a one-hot representation of the sequences, which is the mapping

$$\mathcal{X} : \{1, \dots, q\} \mapsto \{\mathbf{e}_1, \dots, \mathbf{e}_q\}, \quad [\text{S5}]$$

where  $\mathbf{e}_i$  is the  $i$ th basis vector. Then the expected covariance between the  $i$ th and  $j$ th position is  $E(x_i x_j^T) - E(x_i)E(x_j)^T$ ,

$$E(x_i x_j^T) = \sum_{a,b=1}^q \mathbf{e}_a \mathbf{e}_b^T \frac{e^{J_{ij}\Theta_{ab}}}{\sum_{a,b} e^{J_{ij}\Theta_{ab}}}$$

$$\Rightarrow E(x_i x_j^T)_{ab} = P(x_i = a, x_j = b) \quad [\text{S6}]$$

$$E(x_i) = \sum_{a=1}^q \mathbf{e}_a \frac{e^{J_{ij}\Theta_{ab}}}{\sum_{a,b} e^{J_{ij}\Theta_{ab}}} \Rightarrow E(x_i)_a = P(x_i = a). \quad [\text{S7}]$$

Putting this all together gives

$$P(x_i = a, x_j = b) - P(x_i = a)P(x_j = b) = \text{cov}(x_i, x_j)_{ab}. \quad [\text{S8}]$$

This covariance measure is widely used for residue–residue interaction detection.

**Phylogenetic Relatedness.** The central goal of this work is to study the effects of phylogenetic relatedness on the covariance matrix of the resulting sequences. This requires a simulation that allows us to isolate the effects of phylogeny from other potential sources of covariance. To simulate phylogenetic relationships between protein sequences, we consider a simple model of evolution where there are no interactions between sequence positions as used by the structural constraints detailed above, and furthermore there are no amino acid preferences or mutation rate differences between different sites. As above, the only allowed mutation events are amino acid substitutions. In this setup, all proposed mutations will be accepted, as there is no energy difference between different proposed mutations. Initially we work with what we term homogeneous phylogenetic trees, in which all branches have the same length. In this case the number of proposed mutations per branch is the same between duplication

events for all branches so that all members of a generation of sequences will branch or duplicate simultaneously.

We start the simulation with a randomly drawn sequence with the required fixed length  $p$ , which we then simulate through  $m$  mutations representing the initial branch of the tree. The first branching event then occurs—the current sequence is duplicated, and the two copies of the sequence are then independently evolved along two parallel branches of the tree, each for  $m$  mutations. At this point the next branching event occurs—in general a branching event is where all of the sequences present are duplicated, to double the number of sequences creating the next “generation” of sequences. Our simulations also explore heterogeneous trees, where the number of proposed mutations for each branch is drawn from a probability distribution, and the extension of the simulation to this case is straightforward.

**Expected Covariance.** To examine the expected covariance generated by a phylogenetic tree, we first consider two nodes separated by  $2m$  mutation events. We use a substitution model with the assumption that each amino acid can mutate to any other with equal probability; i.e., there are no phenotypic effects. This is a stationary Markov process reminiscent of the Jukes–Cantor model, with the property that if  $\alpha(t) = \mathbf{E}(x(0)x(t))$ , then  $\alpha(m+n) = \alpha(m)\alpha(n)$  (1). This yields an autocorrelation function  $\alpha(t)$  that is proportional to an exponential, with the exponent given by a relaxation rate  $r$ , where  $\mathbf{E}(x(t+1)|x(t)) = (1-r)x(t)$ .

To show this, we note that when there are no preferences for mutation sites, the stationary state of the Markov chain is the uniform distribution. Since the sequences are randomly chosen and so uniformly distributed at the start of the simulation, this Markov chain setup is stationary. The Markov condition is given by

$$\mathbf{P}(x(t+1)|x(t), \dots, x(0)) = \mathbf{P}(x(t+1)|x(t)). \quad \text{[S9]}$$

To see that the phylogenetic process is Markovian, we first note that the state at  $t+1$  can be written as  $x(t+1) = x(t) + v(t+1)$ , where  $v(t)$  is the change induced by the mutation and can be viewed as the discrete velocity. Crucially,  $v(t)$  is dependent only on the state  $x(t)$ . This implies that  $x(t+1)$  is solely dependent on  $x(t)$ , satisfying Eq. S9.

The Markov condition can be rewritten in expectation as

$$\begin{aligned} \mathbf{E}(x_j(t+1)|x_j(t)) &= \mathbf{E}(d_j(t+1)x_j(t)|x_j(t)) \\ &= \underbrace{\mathbf{E}(d_j(t+1)|x_j(t))}_{(A)} x_j(t), \end{aligned} \quad \text{[S10]}$$

where we define  $d_j(t) = x_j(t+1)/x_j(t)$ . Here  $d_j(t)$  and  $x_j(t)$  are the  $j$ th elements of  $d(t)$  and  $x(t)$ , respectively. To derive the relaxation rate for the phylogenetic process we consider the map of  $q$  states onto the unit circle of a complex plane given by Eq. S2. With this mapping (A) becomes

$$\begin{aligned} \mathbf{E}(d_j(t+1)|x_j(t)) &= \frac{p-1}{p} + \frac{1}{p(q-1)} \sum_{j=1}^{q-1} e^{ij k \theta} \\ &= 1 - \frac{1}{p} \frac{q}{q-1}. \end{aligned}$$

Substituting this back into Eq. S10 yields

$$\mathbf{E}(\mathbf{x}(t+1)|\mathbf{x}(t)) = \left(1 - \frac{1}{p} \frac{q}{q-1}\right) \mathbf{x}(t).$$

Using the definition of relaxation rate,  $r$ , given above, we obtain

$$\alpha(t) = \alpha(0) \exp\left(-\frac{q}{q-1} \frac{t}{p}\right). \quad \text{[S11]}$$

The mapping to a complex circle implies that  $\alpha(0) = \text{var}(x) = 1$ . This can be seen if we consider the  $\mathbf{E}(x)$  and  $\mathbf{E}(x^\dagger x)$  sepa-

ately. We note that for a uniform distribution  $\mathbf{E}(x)$  is proportional to the sum of the states on a complex circle which is 0. Similarly, we note that since the magnitude of the nodes on a unit circle is one,  $\mathbf{E}(x^\dagger x) = \text{var}(x) = 1$ . Consequently, for two nodes separated by  $t = 2m$  mutations, the covariance is given by  $\alpha = \exp(-2mq/p(q-1))$ .

**Phylogeny and Interactions.** We also use simulations to investigate how the combination of phylogeny and structural interactions affects the covariance. We use the Potts model described above to model structural constraints and evolve sequences simulated according to this model along the chosen phylogenetic tree. The Potts model is given by

$$P(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp\left(\beta \sum_i h_i x_i + \beta \sum_{i<j} J_{ij} x_i x_j\right), \quad \text{[S12]}$$

where  $h_i$  and  $J_{ij}$  are the field parameters,  $\beta = 1/T$ , and  $\mathcal{Z}$  is the partition function. The temperature,  $T$ , acts as a dial for the simulated annealing process. If the samples are drawn independently, the log-likelihood function can be computed as the following:

$$\begin{aligned} l(X; J, h) &= \beta \sum_i h_i \sum_k x_i^k + \beta \sum_{i<j} J_{ij} \sum_k x_i^k x_j^k - n \log \mathcal{Z} \\ &\propto \beta \sum_i h_i P_i + \beta \sum_{i<j} J_{ij} P_{ij} - \log \mathcal{Z}. \end{aligned} \quad \text{[S13]}$$

Here,  $P_i$  and  $P_{ij}$  are sufficient statistics for this problem. The maximum-likelihood solution  $J_{ij}(a, b)$  is given by

$$C_{ij}(a, b) = P_{ij}(a, b) - P_i(a)P_j(b), \quad \text{[S14]}$$

where  $P_{ij}(a, b)$  is the empirical probability of character pairs  $(a, b)$  appearing in the  $(i, j)$ th columns.  $P_i(a)$  is the empirical probability of character  $a$  appearing in the  $i$ th column. Eq. S14 can be written in Wishart format if we consider the design matrix  $X$  to be in one-hot format. The one-hot format is defined as

$$\{1, \dots, q\} \mapsto \{\mathbf{e}_1, \dots, \mathbf{e}_q\}, \quad \text{[S15]}$$

$\mathbf{e}_1 = (1, 0, \dots, 0)$ . Eq. S14 is now given by

$$C_{(i-1)q+a, (j-1)q+b} = \frac{1}{n} X^T X - \bar{X}^T \bar{X}, \quad \text{[S16]}$$

where  $\bar{X}_i = X_{:,i}/n$ . Regularization is not applied here. The Frobenius norm is used to summarize the matrix  $C_{ij}$  into a score. The covariance between two phylogenetically related sequences in one-hot format can be easily found using Eq. S11 by finding the corresponding  $\alpha(0)$ , which is the variance. The variance for one-hot-formatted sequences is  $((q-1)/q)^2$ , and thus we can find the analytical form of the spectrum by scaling  $C$  by a factor  $((q-1)/q)^{-2}$ .

### Covariance Structure Caused by Phylogeny

The hierarchical structure of the phylogenetic tree yields a true sequence covariance matrix  $\Sigma_S$  made of nested squares that correspond to the different branching events. To elucidate the nested structure induced by a phylogenetic tree, we consider a homogeneous tree with  $b$  branching events and  $m$  mutations per branch. Each pair of sequences is separated by  $2\bar{b}m$  mutations, where  $\bar{b}$  is the number branching events (generations) since their most recent common ancestor. Hence, using the covariance relation found in Eq. S11 the true covariance matrix  $\Sigma_S$  between a set of sequences generated by a homogeneous tree is given by

$$\Sigma_S = \exp\left(-\frac{q}{p(q-1)} D\right), \quad \text{[S17]}$$

where

$$D = \begin{pmatrix} 0 & 2m & \cdots & 2bm & \cdots & 2bm \\ 2m & 0 & & \vdots & \ddots & \vdots \\ \vdots & & \ddots & 2bm & \cdots & 2bm \\ 2bm & \cdots & 2bm & \ddots & & \vdots \\ \vdots & \ddots & \vdots & & 0 & 2m \\ 2bm & \cdots & 2bm & \cdots & 2m & 0 \end{pmatrix} \quad [\text{S18}]$$

is the distance matrix. The monotonicity of the exponential function means that the nested structure of  $D$  is reflected in  $\Sigma_S$ , the true covariance matrix.

**Eigenvalues of  $\Sigma_S$ .** The eigenvalues of  $\Sigma_S$  in Eq. S17 can be found analytically. This set of eigenvalues has a few distinct features. First, there are  $b+1$  distinct eigenvalues for sequences generated with  $b$  branching events; second, the degeneracy of the eigenvalues increases as their magnitude decreases. The explicit mathematical formula is given by

$$\lambda_i = \begin{cases} 1 + \sum_{j=1}^b 2^{j-1} \alpha^j & i = 0 \\ (1 - \alpha) \left( \sum_{j=0}^{b-i} (2\alpha)^j \right) & i > 0 \end{cases}, \quad [\text{S19}]$$

where  $\lambda_0 \geq \cdots \geq \lambda_b$  ( $\alpha \neq 0$ ). We can view the degeneracy of the eigenvalues as proportional to the probability of drawing a particular eigenvalue; this probability distribution is given by

$$p_i = \begin{cases} 1/n & i = 0 \\ 2^{i-1}/n & i > 0 \end{cases}, \quad [\text{S20}]$$

where  $p_i = \mathbb{P}(\lambda = \lambda_i)$  and  $n = 2^b$ .

**Eigenvectors of  $\Sigma_S$ .** Due to the degeneracy of  $\lambda_i$ , there are  $2^{i-1}$  eigenvectors with the same corresponding eigenvalue. Furthermore, these eigenvectors reflect the events in the phylogenetic tree. The principal eigenvector captures the uniform background noise while all of the other eigenvectors capture duplication events that occur in the phylogenetic tree. The set of eigenvectors associated with  $\lambda_j$ , which we denote  $\mathcal{V}_j$ , captures the duplication events in the  $j - 1$ st generation. This is shown if we consider the outer product of eigenvectors (Fig. S2). In Fig. S2, the  $j$ th and  $k$ th elements of eigenvectors in  $\mathcal{V}_i$  will have opposite sign if the  $j$ th and  $k$ th sequences are leaves produced from different branches immediately after the duplication event; in Fig. S2 red is positive and blue is negative. On the other hand, if the  $j$ th sequence is not a leaf of a certain duplication event, then the  $j$ th element is 0, shown in green.

For example, the expected covariance for a tree with two branching events is given by

$$\Sigma_S = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^2 \\ \alpha & 1 & \alpha^2 & \alpha^2 \\ \alpha^2 & \alpha^2 & 1 & \alpha \\ \alpha^2 & \alpha^2 & \alpha & 1 \end{pmatrix},$$

where  $\alpha = \exp(-2mq/(q-1)p)$ . The eigenvectors for this system are given by

$$\mathcal{V} = \{(1, 1, 1, 1), (1, -1, 0, 0), (0, 0, 1, -1), (1, 1, -1, -1)\}.$$

**Extension of MP.** To analyze the spectrum of the empirical covariance matrix,  $C_S = XX^T/(n-1)$ , where  $X$  is the MSA, we use techniques from RMT. RMT has been applied to a wide range of areas such as quantum mechanics, population genetics, and finance to name a few.

The expected covariance matrix,  $\Sigma_S$ , of sequences simulated along a homogeneous tree is given by Eq. S17. Correspondingly, its expected eigenvalue distribution is given by Eqs. S19 and S20. Marčenko and Pastur (45) formulated a connection between the expected eigenvalue distribution and the empirical eigenvalues of  $C_S$ . Here, we describe a way to extend upon Marčenko and Pastur's derivation for independent samples to samples which are dependent via a tree structure. Surprisingly, the parameters of the phylogenetic tree, i.e., number of mutations per branch and number of branching events, control the empirical eigenvalue distribution which we can find analytically.

**Algebraic Random Matrices.** Rao and Edelman (46) coined the term "algebraic random matrices" which refers to random matrices whose spectra are encoded in a polynomial. Here, we show that  $C_S$  is an algebraic random matrix. Marčenko and Pastur derived a connection between the eigenvalue distribution of  $\Sigma_S$ , which we denote as  $T(\lambda)$ , and the spectrum of  $C_S$ ,  $f(\lambda)$ , via its Stieltjes transform,  $G(z)$ . The Stieltjes transform of  $f(\lambda)$  is

$$G(z) = \int_{-\infty}^{\infty} \frac{dF(\lambda)}{\lambda - z}, \quad [\text{S21}]$$

where  $dF(\lambda) = f(\lambda)d\lambda$ . The inversion formula is given by

$$f(\lambda) = \lim_{y \rightarrow 0} \Im \{ G(\lambda + iy) \}. \quad [\text{S22}]$$

Marčenko and Pastur (45) found that  $G(z)$  satisfies the differential equation

$$\frac{-1}{G(z)} = z - c \int_{-\infty}^{\infty} \frac{\lambda dT(\lambda)}{1 + \lambda G(z)}, \quad [\text{S23}]$$

where  $c = n/p$  and  $X$  is a matrix of size  $n \times p$ . This establishes a connection between  $T(\lambda)$  and  $f(\lambda)$  via  $G(z)$ . To apply Eq. S23 we simply use the expressions for the eigenvalues  $\lambda_i$  and their corresponding probabilities  $p_i$  from Eqs. S19 and S20, yielding  $dT(\lambda) = \sum_{i=1}^{b+1} p_i \delta(\lambda - \lambda_i) d\lambda$ . Thus, Eq. S23 becomes

$$\frac{-1}{G(z)} = z - c \sum_{i=1}^{b+1} \frac{p_i \lambda_i}{1 + \lambda_i G(z)}. \quad [\text{S24}]$$

Multiply by  $G(z) \prod_{i=1}^{b+1} (1 + \lambda_i G(z))$  to obtain

$$(zG + 1) \prod_{i=1}^{b+1} (1 + \lambda_i G) - c \sum_{i=1}^{b+1} p_i \lambda_i G \prod_{j \neq i} (1 + \lambda_j G) = 0. \quad [\text{S25}]$$

Using the inversion formula Eq. S22,  $f(\lambda)$  is found as the positive imaginary part of the roots of Eq. S25. One limit to this method is the accuracy of root-finding algorithms for polynomials of high degree. However, we note that the spectrum becomes stationary as  $b$  increases, and as a consequence the spectrum is well approximated by finding the spectrum of a tree with a sufficiently large number of branching events. For example, Fig. S3 shows the change in spectrum as we increase the number of branching events; the change between 6 and 7 branching events is noticeable, while the spectrum is almost exactly the same between 10 and 11 branching events, as suggested above.

**Simple Phylogeny.** The simplest phylogeny has equal branch lengths and just a single branching event—we call this the simple phylogenetic tree. In this case, the expected eigenvalue distribution is

- i)  $p_1 = P(\lambda = 1 + \alpha) = 1/2$
- ii)  $p_2 = P(\lambda = 1 - \alpha) = 1/2$ ,

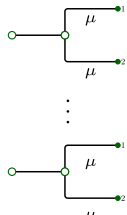
where  $\alpha = \exp(-2mq/p(q-1))$ . As a result, Eq. S25 becomes

$$z(1 - \alpha^2)G^3 + (2z + (1 - c)(1 - \alpha^2))G^2 + (z + 2 - c)G + 1 = 0. \quad [\text{S26}]$$

This is a polynomial of degree three, and hence there are three roots,  $G_1(z)$ ,  $G_2(z)$ , and  $G_3(z)$ . These roots can all be real, or we can have one real root and two complex conjugate roots. The limiting eigenvalue distribution is given by  $f(z) = \Im(G(z))$ .

### Spectra of Inhomogeneous Simple Phylogeny

Here we extend to the case where the branch lengths are no longer equal; instead the number of mutations per branch is drawn from a probability distribution with mean  $\mathbf{E}(m) \equiv \mu$ . To model this, we use the Poisson distribution, which realistically models frequency of events in a time interval. We first recall that if we consider  $n_0$  copies of an initial sequence which all go through a homogeneous simple phylogenetic tree with  $\mu$  mutations per branch, then the expected covariance matrix is given by

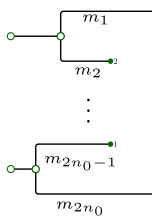


$$\Rightarrow \Sigma_S = \begin{pmatrix} 1 & \alpha_{2\mu} & & & \\ \alpha_{2\mu} & 1 & & & \\ & & \ddots & & \\ & & & 1 & \alpha_{2\mu} \\ & & & \alpha_{2\mu} & 1 \end{pmatrix}, \quad \text{[S27]}$$

where  $\alpha_{2\mu} = e^{-2\mu q/p(q-1)}$ . Using Eq. S25, the Stieltjes transform for this homogeneous simple phylogenetic system satisfies

$$G \left( z - \frac{c}{2} \frac{1 + \alpha_{2\mu}}{1 + (1 + \alpha_{2\mu})G} - \frac{c}{2} \frac{1 - \alpha_{2\mu}}{1 + (1 - \alpha_{2\mu})G} \right) = -1, \quad \text{[S28]}$$

where  $c = n/p$  and  $n = 2n_0$ . For the equivalent inhomogeneous case, we denote the branch lengths drawn from a Poisson distribution with mean  $\mu$  as  $m_1, m_2, \dots, m_{2n_0-1}, m_{2n_0}$  (diagram below), which implies that  $\Sigma_S$  is given by



$$\Rightarrow \Sigma_S = \begin{pmatrix} 1 & \alpha_{i_1} & & & \\ \alpha_{i_1} & 1 & & & \\ & & \ddots & & \\ & & & 1 & \alpha_{i_{n_0}} \\ & & & \alpha_{i_{n_0}} & 1 \end{pmatrix}, \quad \text{[S29]}$$

where the notation  $i_1 = m_1 + m_2, \dots, i_{n_0} = m_{2n_0-1} + m_{2n_0}$  and in all cases  $\alpha_i = e^{-qi/p(q-1)}$ . This notation satisfies three properties:

- i)  $\alpha_{i+j} = \alpha_i \alpha_j$
- ii)  $\alpha_{ij} = \alpha_i^j$
- iii)  $\alpha_i^j = \alpha_j^i$ .

The additive property of the Poisson distribution implies that  $i_1, \dots, i_{n_0}$  are independent and identically distributed variables drawn from a Poisson distribution with mean  $2\mu$ , so that if we let  $\rho_i = \mathbf{P}(i_1 = i)$ , then

$$\rho_i = \frac{(2\mu)^i e^{-2\mu}}{i!}. \quad \text{[S30]}$$

Using Eq. S24, we find that  $G$  satisfies

$$G \left( z - \underbrace{\frac{c}{2} \sum_{i=0}^{\infty} \frac{\rho_i(1 + \alpha_i)}{1 + (1 + \alpha_i)G}}_{(A)} - \underbrace{\frac{c}{2} \sum_{i=0}^{\infty} \frac{\rho_i(1 - \alpha_i)}{1 + (1 - \alpha_i)G}}_{(B)} \right) = -1,$$

where  $\rho_i$  is given by Eq. S30. We note that the term (A) can be rearranged into the following:

$$\begin{aligned} \sum_{i=0}^{\infty} \frac{\rho_i(1 + \alpha_i)}{1 + (1 + \alpha_i)G} &= \frac{1}{G} - \frac{1}{G(1 + G)} \sum_{i=0}^{\infty} \frac{\rho_i}{1 + \alpha_i \frac{G}{G+1}} \\ &= \frac{1}{G} - \frac{1}{G(1 + G)} \sum_{i=0}^{\infty} \rho_i \sum_{j=0}^{\infty} \left( \frac{-G}{1 + G} \alpha_i \right)^j \\ &= \frac{1}{G} - \frac{1}{G(1 + G)} \sum_{j=0}^{\infty} \left( \frac{-G}{1 + G} \right)^j \sum_{i=0}^{\infty} \rho_i \alpha_i^j. \end{aligned} \quad \text{[S31]}$$

Furthermore, term (B) can be rearranged in a similar fashion. We can make the approximation

$$\begin{aligned} \sum_{i=0}^{\infty} \rho_i \alpha_i^j &= \mathbf{E}_i(\alpha_i^j) \\ &= \exp \left( 2\mu(e^{-qj/p(q-1)} - 1) \right) \\ &= \exp \left( 2\mu(qj/p(q-1) + o(p^{-2})) \right) \sim \alpha_{2\mu}^j, \end{aligned} \quad \text{[S32]}$$

for large  $p$ . Substituting this back into Eq. S31 gives

$$\begin{aligned} \sum_{i=0}^{\infty} \frac{\rho_i(1 + \alpha_i)}{1 + (1 + \alpha_i)G} &\sim \frac{1}{G} - \frac{1}{G(1 + G)} \sum_{i=0}^{\infty} \left( \frac{-G}{1 + G} \right)^i \alpha_{2\mu}^i \\ &= \frac{1 + \alpha_{2\mu}}{1 + (1 + \alpha_{2\mu})G}. \end{aligned} \quad \text{[S34]}$$

Eq. S34 shows that when  $p$  is sufficiently large, the summation simplifies as a function of the mean of the Poisson distribution, which yields

$$G \left( z - \frac{c}{2} \frac{1 + \alpha_{2\mu}}{1 + (1 + \alpha_{2\mu})G} - \frac{c}{2} \frac{1 - \alpha_{2\mu}}{1 + (1 - \alpha_{2\mu})G} \right) \sim -1.$$

This is the same as Eq. S28. Intuitively, this result tells us that  $G$  can be approximated by the mean of the probability distribution,  $\mu$ , when  $p$  is sufficiently large.

For the spectral plots we noted that the solutions given by the polynomial in Eq. S25 give a delta peak around zero. This set of algebraic polynomials has one convenient property: The behavior in the neighborhood around  $z = 0$  can be analytically identified. In this vicinity,  $G(z)$  is  $\mathcal{O}(1/z)$ . Equating the coefficients for  $\mathcal{O}(1/z^d)$  in the polynomial given in Eq. S25, we can obtain by self-consistency that  $G(z)$  satisfies

$$\begin{aligned} zG(z) + (1 - c) &= 0 \\ \Rightarrow G(z) &= -\frac{1 - c}{z}, \end{aligned} \quad \text{[S35]}$$

where  $c = n/p$ . We know that when  $f(\lambda) = \delta(0)$ , then  $G(z) = -1/z$ , and therefore we can invert Eq. S35 to find that the area for the peak at zero is  $1 - c$ . Consequently, the limiting eigenvalue distribution of  $C_S$  can be divided into two parts,

$$f(\lambda) = f_1(\lambda) + f_2(\lambda), \quad \text{[S36]}$$

where

$$f_1(\lambda) = \begin{cases} (1 - c)\delta(x) & \text{if } 0 \leq c \leq 1 \\ 0 & \text{if } c > 1 \end{cases}. \quad \text{[S37]}$$

The "area" underneath this peak is therefore  $1 - c$  for  $c \leq 1$ ; as a result, the area of the rest of the data, e.g., the histogram bars shown in Fig S3, will scale to be  $c$  for  $c \leq 1$  or  $1$  if  $c > 1$ . Note that  $f_2(\lambda) = f(\lambda)$  when  $\lambda \neq 0$ , and hence we find this nontrivial part of the solution in the exact same way as we find  $f(\lambda)$  given in Eq. S22. This peak tells us about the rank of the empirical covariance matrix but does not offer any information about phylogeny.

### Spectra of Inhomogeneous Trees

For the simple phylogenetic tree we have proved that the spectrum for branches drawn from a Poisson distribution with mean  $\mu$  can be approximated by the spectrum of a homogeneous tree with branch length  $\mu$ . To extend this to other inhomogeneous trees, we proceed via two steps. We first show that Eq. S33,  $\mathbf{E}_i(\alpha_j^i) \sim \alpha_{\mathbf{E}(i)}^j$  where  $\mathbf{E}(i) = 2\mu$ , holds for any probability distribution with a convergent moment-generating function (MGF). The second step is to show that the approximation works for a tree with an arbitrary number of branching events.

**MGF Approximation.** The MGF is given by  $\mathbf{E}_i(\alpha_j^i) = \mathbf{E}_i(e^{-qij/p(q-1)})$ . We want to show that as  $p$  becomes sufficiently large,

$$\mathbf{E}_i(\alpha_j^i) \sim \alpha_j^{\mathbf{E}(i)} = \alpha_{\mathbf{E}(i)}^j. \quad [\text{S38}]$$

Equivalently, we want to show that  $\mathbf{E}(e^{-\delta x}) \sim e^{-\delta \mathbf{E}(x)}$ , where  $\delta = O(1/p)$ . First, we note that the MGF satisfies

$$\mathbf{E}(e^{-\delta x}) = \sum_{i=0}^{\infty} (-1)^i \frac{1}{i!} \delta^i \mathbf{E}(x^i), \quad [\text{S39}]$$

in particular,  $\mathbf{E}(x^2) = \text{var}(x) + \mathbf{E}(x)^2$ . The functional form  $f(x) = e^{-\delta x}$  is convex for positive  $\delta$  and  $x$ . Therefore, we can apply Jensen's inequality, which yields

$$e^{-\delta \mathbf{E}(x)} \leq \mathbf{E}(e^{-\delta x}), \quad [\text{S40}]$$

and this gives a lower bound to Eq. S39. An upper bound can also be found by considering the following inequality:

$$e^{-\delta x} \leq 1 - \delta x + \frac{1}{2!}(\delta x)^2. \quad [\text{S41}]$$

We can apply the expectation operator on both sides to give

$$\begin{aligned} \mathbf{E}(e^{-\delta x}) &\leq 1 - \delta \mathbf{E}(x) + \frac{1}{2!} \delta^2 \mathbf{E}(x)^2 + \frac{1}{2!} \delta^2 \text{var}(x) \\ &= e^{-\delta \mathbf{E}(x)} + \frac{1}{2!} \delta^2 \text{var}(x) + O(\delta^3). \end{aligned} \quad [\text{S42}]$$

Consequently, Eq. S39 is bounded by the following:

$$e^{-\delta \mathbf{E}(x)} \leq \mathbf{E}(e^{-\delta x}) \leq e^{-\delta \mathbf{E}(x)} + O(\delta^2). \quad [\text{S43}]$$

As  $\delta$  becomes sufficiently small, the upper and lower bounds both converge to  $e^{-\delta \mu}$ . Subsequently, we can approximate  $\mathbf{E}(e^{-\delta x})$  by  $e^{-\delta \mu}$  with an error term which is second order with respect to  $\delta$ . Hence, as  $p$  becomes sufficiently large we can use the approximation

$$\mathbf{E}(e^{-\delta x}) \sim e^{-\delta \mathbf{E}(x)}, \quad [\text{S44}]$$

where the error of the approximation is  $O(p^{-2})$ .

**Extension to Arbitrary Phylogeny.** Recall that the Stieltjes transform,  $G$ , for a homogeneous tree with  $b$  branching events and  $\mu$  mutation events per branch is given by

$$G\left(z - c \sum_{i=1}^{b+1} p_i \frac{\lambda_i}{1 + \lambda_i G(z; c)}\right) = -1, \quad [\text{S45}]$$

where  $\lambda_i$  and  $p_i$  are given in Eqs. S19 and S20. The equivalent inhomogeneous tree with  $E$  branches, where the length of each branch is drawn from a distribution with mean  $\mu$ , is given by

$$G\left(z - c \sum_{i=1}^{b+1} p_i \sum_{\mathbf{M}} \frac{\rho_{m_1} \cdots \rho_{m_E} \lambda_i(\alpha_{m_1}, \dots, \alpha_{m_E})}{1 + \lambda_i(\alpha_{m_1}, \dots, \alpha_{m_E}) G}\right) = -1, \quad [\text{S46}]$$

where  $\mathbf{M} = (m_1, \dots, m_E)$  is the set of branch lengths. As for the simple inhomogeneous case, we let  $\rho_i = P(m=i)$ . Here we want to show that  $\lambda_i(\alpha_{m_1}, \dots, \alpha_{m_E})$  satisfies  $\lambda_i(\alpha_\mu, \dots, \alpha_\mu) = \lambda_i$ . We consider an inductive process, where we show that the following is satisfied:

$$\begin{aligned} &\sum_{\mathbf{M} \in \mathbb{N}^E} \frac{\rho_{m_1} \cdots \rho_{m_E} \lambda(\alpha_{m_1}, \dots, \alpha_{m_E})}{1 + \lambda(\alpha_{m_1}, \dots, \alpha_{m_E}) G} \\ &\sim \sum_{\mathbf{M} \in \mathbb{N}^{E-1}} \frac{\rho_{m_1} \cdots \rho_{m_{E-1}} \lambda(a_{m_1}, \dots, \alpha_{m_{E-1}}, \alpha_\mu)}{1 + \rho_{m_{E-1}} \lambda(a_{m_1}, \dots, \alpha_{m_{E-1}}, \alpha_\mu) G}. \end{aligned} \quad [\text{S47}]$$

Toward this end, we consider the Taylor expansion of the function

$$h(x) = \frac{\lambda(\alpha_{m_1}, \dots, \alpha_{m_{E-1}}, x)}{1 + \lambda(\alpha_{m_1}, \dots, \alpha_{m_{E-1}}, x) G} = \sum_{j=0}^{\infty} h_j x^j, \quad [\text{S48}]$$

where the coefficients  $h_j$  depend on  $\alpha_{m_i}$ . This Taylor expansion can be used in the following way:

$$\begin{aligned} &\sum_{i=0}^{\infty} \frac{\rho_i \lambda(\alpha_{m_1}, \dots, \alpha_{m_{E-1}}, \alpha_i)}{1 + \lambda(\alpha_{m_1}, \dots, \alpha_{m_{E-1}}, \alpha_i) G} \\ &= \sum_{i=0}^{\infty} \rho_i \sum_{j=0}^{\infty} h_j \alpha_i^j = \sum_{j=0}^{\infty} h_j \sum_{i=0}^{\infty} \rho_i \alpha_i^j = \sum_{j=0}^{\infty} h_j \sum_{i=0}^{\infty} \rho_i \alpha_i^j \\ &= \sum_{j=0}^{\infty} h_j \mathbf{E}_i(\alpha_j^i). \end{aligned}$$

Using the approximation in Eq. S38, we find

$$\sum_{j=0}^{\infty} h_j \mathbf{E}_i(\alpha_j^i) \approx \sum_{j=0}^{\infty} h_j \alpha_\mu^j = h(\alpha_\mu), \quad [\text{S49}]$$

and thus Eq. S47 is satisfied. We can repeat this process  $E$  times (once for each branch of the tree), yielding

$$\begin{aligned} \sum_{\mathbf{M} \in \mathbb{N}^E} \frac{\rho_{m_1, \dots, m_E} \lambda(\alpha_{m_1}, \dots, \alpha_{m_E})}{1 + \lambda(\alpha_{m_1}, \dots, \alpha_{m_E}) G} &\sim \frac{\lambda(\alpha_\mu, \dots, \alpha_\mu)}{1 + \lambda(\alpha_\mu, \dots, \alpha_\mu) G} \\ &= \frac{\lambda}{1 + \lambda G}. \end{aligned}$$

Substituting this back into Eq. S46, we find that this equation is approximated by Eq. S45 as required.

### Power Law Induced by Phylogeny

For protein covariance matrices, we find that a power law tail occurs in the eigenvalue distribution because of the phylogenetic structure. This structure causes all but the largest two eigenvalues to be degenerate. As detailed in Eq. S20, the degeneracy increases as the size of the eigenvalue decreases. To deduce the power law in the tail of the empirical eigenvalue distribution, we note that the expected eigenvalue distribution in Eq. S19 can be rewritten as

$$\begin{aligned} \lambda(r) &= (1 - \alpha) \left(1 + 2\alpha + \cdots + (2\alpha)^k\right) \\ \lambda(2r) &= (1 - \alpha) \left(1 + 2\alpha + \cdots + (2\alpha)^{k-1}\right), \end{aligned} \quad [\text{S50}]$$

where  $r > 1$  is an index that runs over all copies of each eigenvalue given by Eq. S20, and  $\lambda(r) > \lambda(2r)$ . If we consider the  $b$ th generation of sequences where  $b$  is sufficiently large, we can evaluate the gradient of  $\log(\lambda)$  as a function of  $\log(r)$  by taking the approximation that  $\lambda(r) \sim O((2\alpha)^{k+1})$  if  $2\alpha > 1$ . This gives the following approximation:

$$\nabla \log(\lambda) \sim \frac{\log((2\alpha)^{k+1}) - \log((2\alpha)^k)}{\log(r) - \log(2r)} \quad [\text{S51}]$$

$$= -\frac{\log(2\alpha)}{\log(2)}. \quad [\text{S52}]$$

For the case where  $2\alpha \leq 1$  we note that  $\nabla \log(\lambda) \leq 1$ , since the gradient increases as the mutation rate decreases. We can further simplify this by considering  $\alpha$  as a function of the mutation rate  $m/p$ ,  $\alpha = e^{-2qm/(p(q-1))}$ , which gives

$$\nabla \log(\lambda) = \begin{cases} \frac{2q}{\log(2)(q-1)} \frac{m}{p} - 1 & 2\alpha < 1 \\ 0 & \text{otherwise} \end{cases}. \quad [\text{S53}]$$

This shows that the power law induced by phylogeny in the tail of the eigenvalue distribution is controlled by the average branch length,  $\lambda(r) \propto r^{F(m/p)}$ . The slope is between 0 and 1 with lower mutation rates generating a steeper gradient, reflecting the fact that phylogenetic effects are stronger for lower mutation rates.

We note that many previous works use a phylogeny correction that involves down-weighting aligned sequences that are more similar to each other than some user-defined threshold—such as a Hamming distance of 0.7. If this type of threshold is used, then this correction modifies the contribution to the covariance made by those highly similar sequences found at the leaves of the tree. This contribution does not significantly affect the spectrum of the sample covariance matrix. To demonstrate this, we compare raw and filtered sequences alignments, where the filtered alignments are pruned so that no two sequences are more similar than the relevant threshold. Fig. S4 shows that the eigenvalue spectra of the raw and filtered alignments cannot be distinguished.

In contrast, the phylogeny power law presented here accounts for the contribution to the covariance made by the deep branching events that occur early in the tree. This is demonstrated most clearly by Fig. 4 C and D of the main text, which shows the eigenvalues of the sample covariance matrix plotted on a log scale for the cases (Fig. 4C) with phenotypic interactions only, where the slope is zero, and (Fig. 4D) with both phenotypic interactions

and phylogeny, where the slope is nonzero. This is further illustrated by Fig. S5, which shows that eigenvalues of the covariance matrix for sequences simulated with phenotypic interactions in the absence of phylogeny follow the MP law, unlike those in Fig. S4.

### Truncating Principal Eigenvectors

To examine the effects of removing modes from the covariance, we consider the eigendecomposition of the covariance measure  $C$  in Eq. S16, given by

$$C = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \cdots + \lambda_r \mathbf{v}_r \mathbf{v}_r^T, \quad [\text{S54}]$$

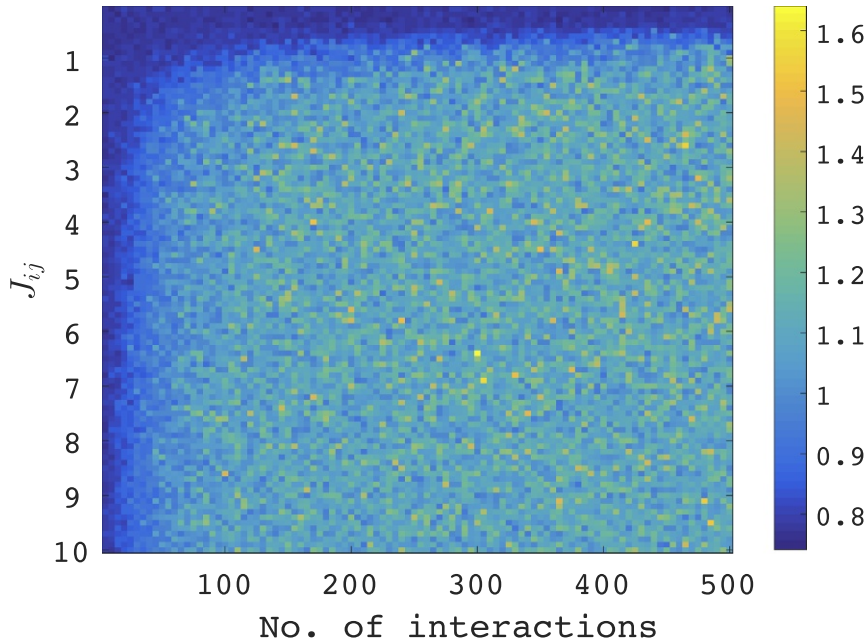
where  $r = p(q-1)$  is the rank of the matrix and  $\lambda_1 \geq \cdots \geq \lambda_r$  are the empirical eigenvalues. To distinguish the effects of phylogeny on the eigenvectors, we note that the bulk of the eigenvalue distribution for independent sequences with only phenotypic signals is roughly MP, shown in Fig. S5. This implies that the majority of the true eigenvalues are around one. Thus, we use the following truncation scheme:

$$C(t) = \mathbf{v}_t \mathbf{v}_t^T + \cdots + \mathbf{v}_r \mathbf{v}_r^T. \quad [\text{S55}]$$

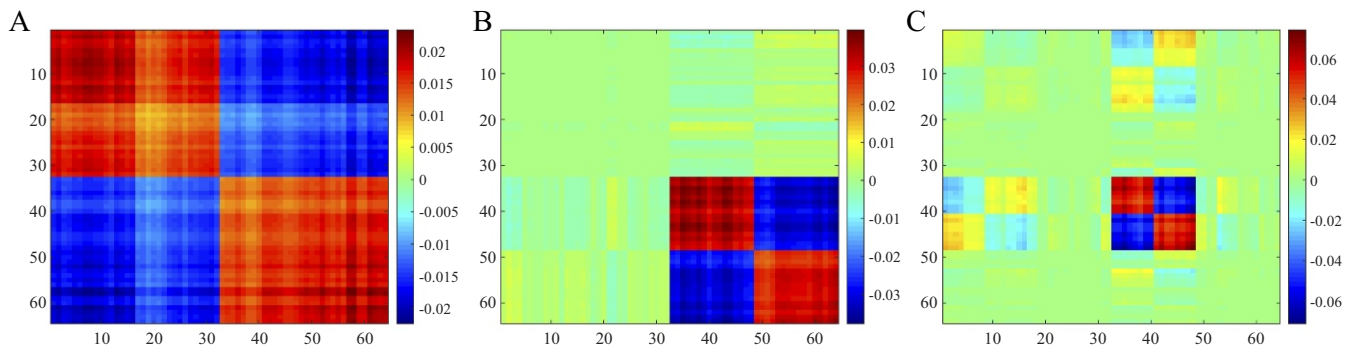
The maximum-likelihood estimator in Eq. S14 makes the critical assumption that the sequences are independent. Fig. 1 shows that this assumption significantly curbs the accuracy of contact prediction. Over recent years, methods such as DCA (13, 15) have improved the accuracy of contact prediction by using the inverse of  $C_{ij}(a, b)$ , otherwise known as the mean-field approximation (MFA). To see how MFA relates to removing the principal eigenvectors, we note that the MFA approximation is given by

$$C^{-1} = \frac{1}{\lambda_1} \mathbf{v}_1 \mathbf{v}_1^T + \cdots + \frac{1}{\lambda_r} \mathbf{v}_r \mathbf{v}_r^T. \quad [\text{S56}]$$

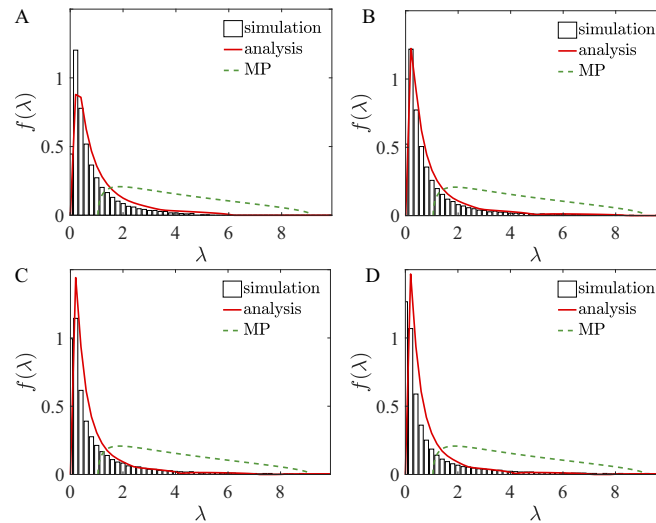
Crucially, this approximation effectively removes the largest eigenvalues since  $1/\lambda$  is negligible in these cases.



**Fig. S1.** Here we use simulations to examine the largest eigenvalue produced with different phenotypic parameter values. The plot shows a heat map of the largest eigenvalue of the covariance matrix produced for simulations where we vary the number of structural interactions between 0 and 500, where the maximum strength of these interactions varies on the y axis between 0 and 10. These simulations extend those shown in Fig. 4 A and B of the main text. We note that the saturation behavior observed in Fig. 4A is replicated here, both as the number of interactions increases and as the maximum interaction strength increases.



**Fig. S2.** Eigenvectors caused by phylogeny. The configuration used to produce these plots is  $m = 30$ ,  $p = 1,000$ , and  $b = 6$ . A–C show the outer products of eigenvectors corresponding to (A)  $\lambda_6$ , (B)  $\lambda_5$ , and (C)  $\lambda_4$  in Eq. S19.



**Fig. S3.** The eigenvalue distributions produced by phylogeny differ substantially from the MP distribution (shown in green). As the number of branching events increases, the resulting eigenvalue distributions become more similar to one another. Here 8,096 sequences of length 100 are generated using (A) 6, (B) 7, (C) 10, and (D) 11 branching events. Analytical solutions using the analysis presented here are shown in red. There are 300 interactions of strength uniformly distributed between  $-5$  and  $5$ .

