Impact of phylogeny on the inference of functional sectors from protein sequence data

Nicola Dietler[1,2], Alia Abbara[1,2], Subham Choudhury[1,2], Anne-Florence Bitbol[1,2,*]

1 Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
2 SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
* Corresponding author: anne-florence.bitbol@epfl.ch

Abstract

Statistical analysis of multiple sequence alignments of homologous proteins has revealed groups of coevolving amino acids called sectors. These groups of amino-acid sites feature collective correlations in their amino-acid usage, and they are associated to functional properties. Modeling showed that natural selection on an additive functional trait of a protein is generically expected to give rise to a functional sector. These modeling results motivated a principled method, called ICOD, which is designed to identify functional sectors, as well as mutational effects, from sequence data. However, a challenge for all methods aiming to identify sectors from multiple sequence alignments is that correlations in amino-acid usage can also arise from the mere fact that homologous sequences share common ancestry, i.e. from phylogeny. Here, we generate controlled synthetic data from a minimal model comprising both phylogeny and functional sectors. We use this data to dissect the impact of phylogeny on sector identification and on mutational effect inference by different methods. We find that ICOD is most robust to phylogeny, but that conservation is also quite robust. Next, we consider natural multiple sequence alignments of protein families for which deep mutational scan experimental data is available. We show that in this natural data, conservation and ICOD best identify sites with strong functional roles, in agreement with our results on synthetic data. Importantly, these two methods have different premises, since they respectively focus on conservation and on correlations. Thus, their joint use can reveal complementary information.

Author Summary

Proteins perform crucial functions in the cell. The biological function of a protein is encoded in its amino-acid sequence. Natural selection acts at the level of function, while mutations arise randomly on sequences. In alignments of sequences of homologous proteins, which share common ancestry and common function, the amino acid usages at different sites can be correlated due to functional constraints. In particular, groups of collectively correlated amino acids, termed sectors, tend to emerge due to selection on functional traits. However, correlations can also arise from the shared evolutionary history of homologous proteins, even without functional constraints. This may obscure the inference of functional sectors. By analyzing controlled synthetic data as well as natural protein sequence data, we show that two very different methods allow to identify sectors and mutational effects in a way that is most robust to phylogeny. We suggest that considering both of these methods allows a better identification of functionally important sites from protein sequences. These results have potential impact on the design of new functional sequences.

Introduction

Statistical inference from genome and protein sequence data is currently making great progress. This is due both to the important growth of available genome sequences, and to the development of inference methods ranging from interpretable models inspired by statistical physics and information theory to large deep learning

models. One important type of data sets consists in multiple sequence alignments (MSAs) of homologous proteins. These proteins, which are said to be the members of a protein family, share a common ancestry and common functional properties. Because of this, MSA columns (i.e. residue sites) feature correlations in amino-acid usage. Pairwise correlations due to contacts between amino acids in the three-dimensional structure of proteins have been thoroughly studied [1–23]. Statistical analyses of MSAs have also revealed the existence of groups of collectively correlated amino acids, termed sectors, which have are associated to a functional role and are often spatially close in the protein structure [24–29]. Preserving the pairwise correlations of natural sequences in a sector-based analysis enabled the successful design of new functional sequences in a pioneering work [26]. Sectors can be identified using Statistical Coupling Analysis (SCA), which focuses on the large-eigenvalue modes of a covariance matrix weighted by conservation [27, 30]. It was recently shown in a minimal model that sectors of collectively correlated amino acids emerge from nonlinear selection on an additive trait (i.e. functional property) of the protein [31]. This case should be quite generic, since empirical data can often be modeled as nonlinear functions of a linear trait [32], and since many protein traits involve additive contributions from amino acids [32–34]. These modeling results motivated a principled method, called ICOD (for Inverse Covariance Off-Diagonal), which is designed to identify functional sectors, and more generally mutational effects, from sequence data [31]. Here, the focus is on the large-eigenvalue modes of the ICOD matrix, which essentially correspond to the small-eigenvalue modes of the covariance matrix, making ICOD *a priori* quite distinct from SCA. Another important difference is that ICOD was designed to focus on correlations and eliminate conservation signal as well as possible. Accordingly, a strength of ICOD is its robustness to conservation [31].

Because homologous sequences share a common ancestry, MSAs also comprise correlations coming from phylogeny [27, 35, 36]. These correlations can arise even in the absence of any selection, i.e. in the case where all mutations are neutral. Phylogenetic correlations impair the inference of structural contacts from sequences [4, 36–40], which has motivated empirical corrections aiming at reducing their impact [4, 7–9, 41–43, 43–45]. Disentangling them from collectively correlated groups of amino acids such as functional sectors is bound to be a significant challenge too, perhaps even more. Indeed, phylogeny strongly impacts the covariance matrix of sites of an MSA, and its particular its large-eigenvalue modes [35, 36], while sector identification by SCA focuses on the large-eigenvalue modes of a modified covariance matrix [27, 30]. Accordingly, the top mode identified in [27] was discarded for this reason, and one of the sectors allowed to classify sequences by organism type, which suggests a phylogenetic contribution.

Here, we investigate how phylogeny impacts the inference of functional sectors, and more generally, of mutational effects, from protein sequence data. Fully disentangling correlations from phylogeny and from functional constraints in natural data is challenging. Thus, we propose a minimal model comprising both phylogeny and functional sectors, allowing us to generate synthetic data with controlled amounts of phylogeny and of functional constraints. We use this data to quantify the impact of phylogeny on sector identification and on mutational effect inference by SCA and ICOD, but also by conservation and correlation. We find that ICOD is most robust to phylogeny, but that conservation is also quite robust. Next, we consider MSAs of 30 natural protein families for which deep mutational scan experiments have been performed, yielding measurements of mutational effects on a specific aspect of function [46]. We show that conservation and ICOD best identify sites with strong functional roles, consistently with our results on synthetic data. Importantly, these two methods have different premises, since they respectively focus on conservation and on correlations.

Results

Minimal model of sequences with a functional sector and phylogeny

Functional sector. We consider sequences where each site can take two states, -1 and 1. While our model can be generalized to include 20 states, reflecting the number of different amino acids, restricting to two states allows to retain key features within a minimal model. We focus on a scalar trait, which corresponds to a physical property associated to a given protein function. Examples of traits include the binding activity to a ligand and the catalytic activity of an enzyme. The trait is assumed to be additive: for a given sequence $\vec{\sigma} = (\sigma_1, \ldots, \sigma_L)$ with length L, it reads $\tau(\vec{\sigma}) = \sum_i D_i \sigma_i$ where $\vec{D}$ is the vector of mutational effects. Thus, each site is assumed to contribute independently to the trait, with its state impacting the trait value.

Nonlinear selection on such an additive scalar trait was shown to lead to a sector [31], i.e. a collectively correlated group of amino acids [27]. The sector corresponds to a specific group of functional amino acids in the protein that are particularly important for the trait of interest. In our model, sector sites have a large mutational effect on the trait. Following [31], we assume that a specific value τ^* of the trait is favored by natural selection. Selection is thus performed using the quadratic Hamiltonian

$$H(\vec{\sigma}) = \frac{\kappa}{2}\left(\sum_{i=1}^{L} D_i\sigma_i - \tau^*\right)^2, \tag{1}$$

where κ denotes selection strength. Note that the fitness of a sequence is minus the value of the Hamiltonian. Equilibrium sequences under the Hamiltonian in Eq 1 can be sampled using the Metropolis-Hastings algorithm, see Fig 1**A**. The resulting data sets contain only correlations coming from selection on the sector.

Phylogeny. To incorporate phylogeny, we take an equilibrium sequence and we evolve this ancestral sequence along a binary branching tree, see Fig 1**B**. Along the tree, random mutations changing the state of an amino acids are proposed, and they are accepted or rejected according to the same Metropolis criterion as the one used to generate equilibrium sequences. Thus, selection on the trait is maintained through the phylogeny. A fixed number μ of accepted mutations are performed on each branch. Generated sequences at the leaves of the tree therefore contain correlations from phylogeny and from selection. Their importance are controlled respectively by the parameters μ and κ: a smaller μ means that sequences are more closely related, yielding more phylogenetic correlations, while a larger κ means stronger selection, yielding more correlations arising from the sector. Note that this data generation process with phylogeny is close to the one we used previously [40, 47], but that the Hamiltonian we use here (Eq 1) is specific to the sector model.

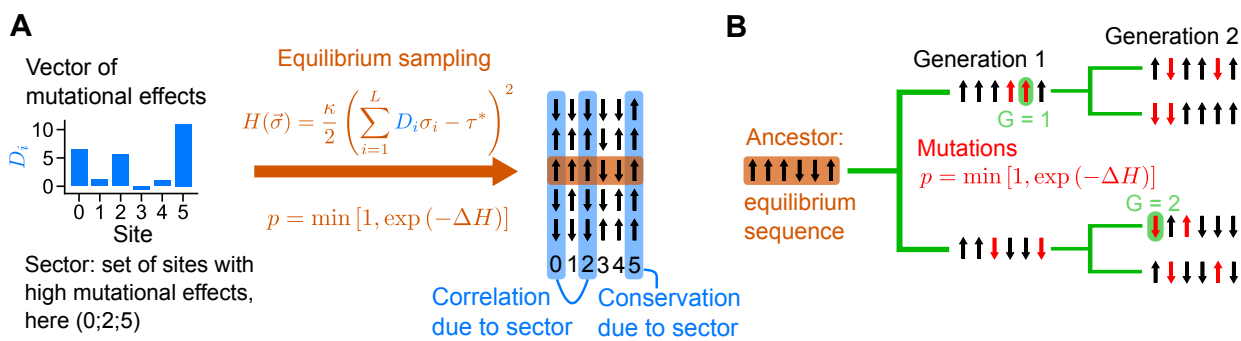


Figure 1: **Data generation process. A:** Generation of sequences with selection only. Given a vector of mutational effects $\vec{D}$ on the trait of interest, we sample independent equilibrium sequences under the Hamiltonian $H(\vec{\sigma})$ in Eq 1. For this, we start from random sequences and we use a Metropolis Monte Carlo algorithm where proposed mutations (changes of state at a randomly chosen site) are accepted with probability p, according to the Metropolis criterion associated to $H(\vec{\sigma})$. The obtained sequences feature pairwise correlations and conservation arising from selection on the trait (via the Hamiltonian $H(\vec{\sigma})$ in Eq 1). **B:** To incorporate phylogenetic correlations, we start from one equilibrium sequence, which becomes the ancestor. We evolve it on a binary branching tree over a fixed number of "generations" (branching events), here 2 generations. A fixed number of mutations μ (red, here $\mu = 2$) are accepted with probability p on each branch of the tree. The earliest generation at which a site mutates with respect to its ancestral state is denoted by G, see examples highlighted in green.

Some previous results on sector identification

SCA. Sectors were first defined in natural data by using Statistical Coupling Analysis (SCA), a method which detects groups of collectively correlated and conserved sites in sequences [27, 30]. In SCA, sectors correspond to the sites that have large components in the eigenvectors associated to the largest eigenvalues of the covariance matrix of sites in the sequences weighted by conservation.

ICOD. Another method to detect sectors was introduced in Ref. [31], and focuses on the eigenvectors with large eigenvalues of the inverse covariance matrix of sites, with diagonal terms set to zero. This method is called ICOD, for Inverse Covariance Off-Diagonal. The eigenvalues of the ICOD matrix are thus close to the inverse of those of the covariance matrix of sites, and large ICOD eigenvalues map to small eigenvalues of the covariance matrix. Hence, ICOD and SCA focus on a different part of the spectrum of the covariance matrix. Besides, ICOD does not use conservation weighting, and further removes conservation signal by setting diagonal terms to zero. It thus focuses on correlations, while SCA combines conservation and correlations. Conceptually, the Hamiltonian in Eq 1 entails that selected sequences satisfy $\vec{D} \cdot \vec{\sigma} \approx \tau^*$. All selected sequences have a similar projection on the vector $\vec{D}$ of mutational effects and lie close to a plane orthogonal to $\vec{D}$. Therefore, $\vec{D}$ is a direction of particularly small variance, which is why the eigenvector with smallest eigenvalue of the covariance matrix, and the one with largest eigenvalue of the ICOD matrix, contains information on $\vec{D}$ and on the sector [31].

Impact of selection and phylogeny on ICOD, covariance and SCA spectra

Signatures of sectors in the spectrum of the ICOD matrix. Before studying the effect of phylogeny, let us analyze the ICOD matrix and its the spectrum without phylogeny, which will serve as baseline. To first order in κ (small selection regime), the off-diagonal elements of the ICOD matrix can be approximated as $\tilde{C}^{-1}_{ij} \approx \kappa D_i D_j$ for all i, j between 1 and L with $i \neq j$ [31]. Here, we show using two different methods that this formula still holds beyond first order in κ: one method extents this formula to second order, and the second one to fourth order (see Supplementary material). Within this robust approximation, if diagonal elements were equal to κD_i^2 instead of 0, the spectrum of the ICOD matrix would comprise a single nonzero eigenvalue $\kappa \sum_{i=1}^{L} D_i^2$ associated to the mutational effect vector $\vec{D}$. Hence, the eigenvector associated to the largest eigenvalue of the ICOD matrix allows to recover $\vec{D}$, as observed in [31].

Consider a mutational effect vector $\vec{D}$ comprising L_S sites with large effects, corresponding to the functional sector, and $L - L_S$ sites with substantially smaller effects. Reordering sites to group together first the sector sites and then the non-sector sites yields an ICOD matrix containing one block of size $L_S \times L_S$ corresponding to the sector, and another of size $(L - L_S) \times (L - L_S)$ which encompasses other sites with low mutational effects. To gain insight on the ICOD spectrum of this matrix, let us ignore the elements that do not belong to either of these blocks by setting them to 0 (see Fig S4). The modified matrix is a block diagonal matrix, and its eigenvalues are the union of the eigenvalues of each block. Apart from its zero diagonal elements, each block is well-approximated by an outer product matrix with elements $\kappa D_i D_j$ whose absolute value is large for the sector block, and small for the non-sector block. If diagonal elements were equal to κD_i^2 instead of 0, the spectrum of the sector (resp. non-sector) block matrix would comprise a single nonzero eigenvalue $\kappa \sum_{i=1}^{L_S} D_i^2$ (resp. $\kappa \sum_{i=L_S+1}^{L} D_i^2$). Setting the diagonal elements to zero entails that eigenvalues in each block need to sum to 0 (as the trace is 0), which leads to some perturbation of the nonzero eigenvalue, but moreover to the appearance of negative eigenvalues that compensate this large positive one. These negative eigenvalues are expected to be larger in absolute value in the sector block than in the non-sector block.

Fig 2 illustrates that the spectrum of the ICOD matrix is reasonably well approximated by that of its block diagonal approximation, if the ICOD matrix is obtained over many sequences. Furthermore, in the block diagonal approximation, we observe that eigenvalues from the sector block yield the largest eigenvalue and the smallest (most negative) eigenvalues, while the non-sector block provides the intermediate eigenvalues. This is in agreement with our expectations coming from the outer product approximation of these blocks. Thus, in addition to possessing a largest eigenvalue coming from the sector and associated to an eigenvector close to the mutational effect vector $\vec{D}$, the ICOD matrix also comprises smallest (most negative) eigenvalues that contain information on the sector.

Note that in the example shown in Fig 2, important mutational effects corresponding to the sector are all positive. We checked that our results are robust to the case where the sector comprises both sites with large positive and and large negative mutational effects: the spectrum is very similar to that in Fig 2, and the top eigenvector of the ICOD matrix successfully recovers $\vec{D}$.

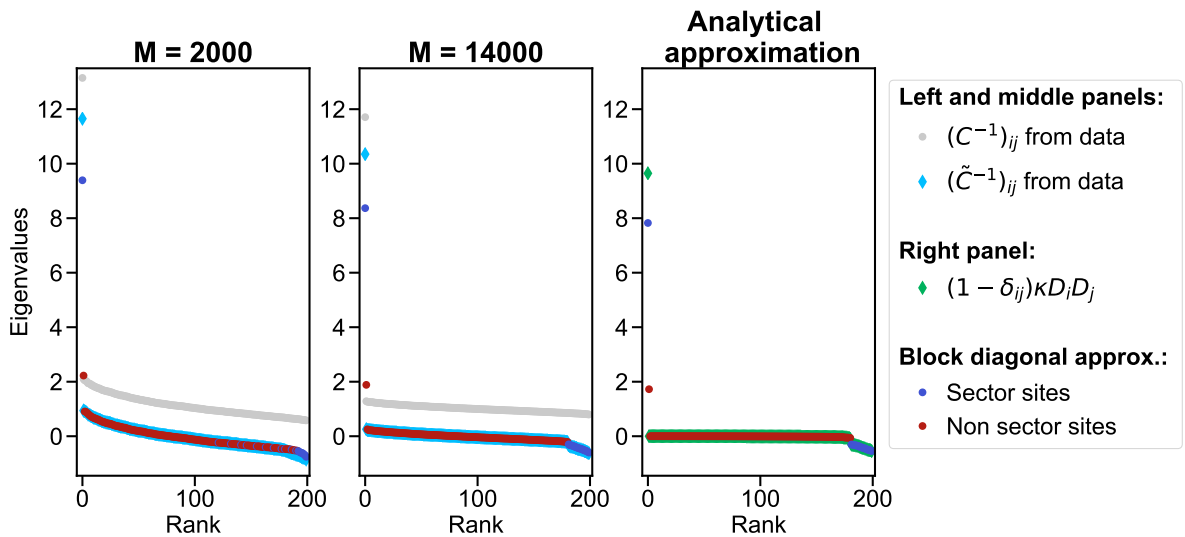


Figure 2: **Spectrum of the ICOD matrix and of its block diagonal approximation.** Left (resp. middle) panel: spectrum of the ICOD matrix $\tilde{C}^{-1}$ computed on 2000 (resp. 14,000) sequences generated independently at equilibrium, and of its block diagonal approximation. The spectrum of the inverse covariance matrix C^{-1} is also shown as a reference. Right panel: spectrum of the analytical approximation of the ICOD matrix, and of its block diagonal approximation. Sequences of length $L = 200$ were sampled independently at equilibrium using the Hamiltonian in Eq 1 with $\kappa = 10/\left(\sum_i D_i^2\right)$ and $\tau^* = 90$. The vector of mutational effect $\vec{D}$ comprises sector sites (the 20 first sites) with components sampled from a Gaussian distribution with mean 5 and variance 0.25, and non-sector sites (the remaining 180 sites) with components sampled from a Gaussian distribution with mean 0.5 and variance 0.25. The analytical approximation $\tilde{C}_{ij}^{-1} \approx (1-\delta_{ij})\kappa D_i D_j$ (see Supplementary material) was computed from the values of κ and $\vec{D}$ used for data generation.

Impact of phylogeny on ICOD, covariance and SCA spectra. How do phylogenetic correlations affect the signature of sectors in the spectra of the ICOD, covariance and SCA matrices? To answer this question, we generate data with only selection, or only phylogeny, or both, within our minimal model.

Fig 3 shows the spectra of the ICOD, covariance and SCA matrices for these data sets. Without phylogeny, a sector gives rise to one large eigenvalue (rank 0) that is an outlier in the spectrum of the ICOD matrix [31]. We observe that this outlier is preserved even with strong phylogeny (small μ, here $\mu = 5$), although its contrast with the rest of the spectrum decreases when μ decreases. Comparing to baseline data without selection confirms that this outlier is due to selection, and also shows that some signal associated to selection is present in the small eigenvalues of the ICOD matrix (rank close to 199), in agreement with our analysis above. This effect of selection also remains visible with phylogeny.

Recall that because of the matrix inversion in ICOD, large ICOD eigenvalues essentially map to small covariance eigenvalues. Accordingly, comparing data with and without selection shows that the main signature of selection is observed for the smallest eigenvalue of the covariance matrix (rank 199). While this outlier is strong without phylogeny, its contrast with the rest of the spectrum decreases as phylogeny is increased, and it is no longer a clear outlier for $\mu = 5$. Meanwhile, the large eigenvalues of the covariance matrix are strongly impacted by phylogeny, in agreement with previous findings [36], but little impacted by selection.

Finally, the large eigenvalues (rank close to 0) of the SCA matrix contain selection signal, as they are outliers with selection, and not without. However, these outliers disappear when phylogeny is strong ($\mu = 5$), and the eigenvalues with and without selection are then very similar. Interestingly, except in the high-phylogeny case, the number of large-eigenvalue outliers in SCA matches the number of sector sites in $\vec{D}$.

Thus, overall, the signatures of selection in the spectrum of ICOD appear to be more robust to phylogeny than those observed in the spectra of the covariance matrix and of the SCA matrix.

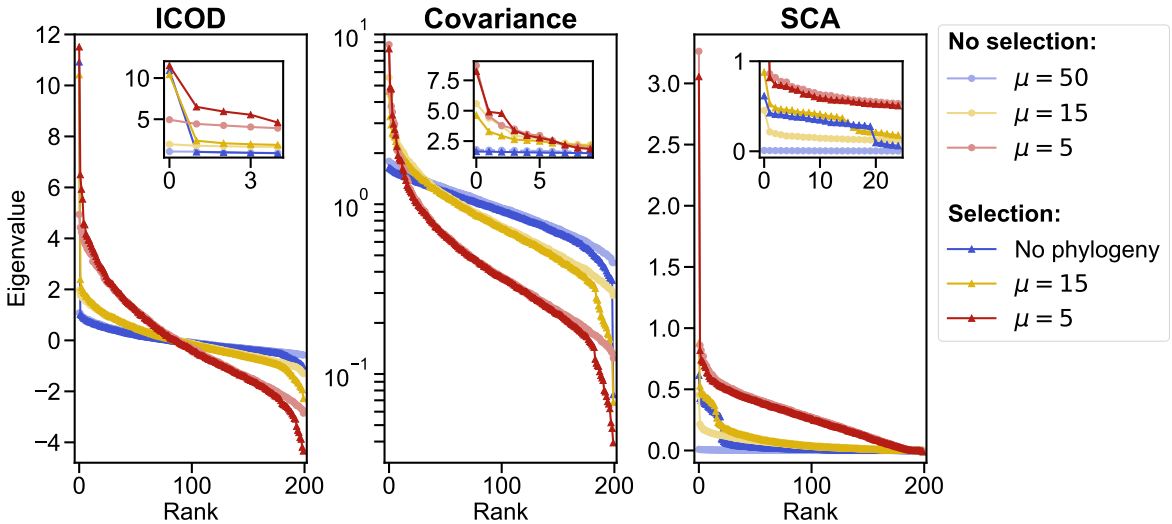


Figure 3: **Impact of phylogeny and selection on ICOD, covariance and SCA spectra.** Eigenvalues of the ICOD, covariance and SCA matrices, sorted from largest to smallest, are shown for sequences generated with only phylogeny (light shades) and both phylogeny and selection (dark shades). We consider different levels of phylogeny by considering different values of μ (shown as different colors). 'No phylogeny' corresponds to sequences generated independently at equilibrium, and thus containing only correlations due to selection. This data set comprises $M = 2048$ sequences generated exactly as in Fig. 2. Data sets without selection are generated by evolving random sequences of length $L = 200$ on a binary branching tree with 11 generations and μ random mutations on each branch, providing $M = 2^{11} = 2048$ sequences. Finally, data sets with phylogeny and selection are generated along a binary branching tree with μ accepted mutations per branch (with acceptance criterion in Eq 2 using the same κ and τ^* as in the no-phylogeny case and as in Fig. 2) and 11 generations again. Insets show a zoom over large eigenvalues. A logarithmic y-scale is used in the center panel for readability.

Impact of phylogeny on mutational effect recovery

How much signal from the sector and the mutational effect vector $\vec{D}$ is contained in the eigenvectors of the ICOD, covariance and SCA matrices? How is this impacted by phylogeny? To quantitatively address this question, we employ the recovery score (see Methods), which is the normalized scalar product of the eigenvector of interest and of the vector $\vec{D}$, both of them with their components replaced by their absolute values. The absolute value is used because our main goal is to identify sites with large mutational effect, whatever its sign, and because eigenvectors are defined up to a global sign.

We consider the eigenvectors associated to the main outliers identified in Fig 3, i.e. the largest eigenvalue for ICOD, and the smallest one for covariance. Furthermore, as we showed that signal about selection also exists at the other end of the spectrum, we investigate it too. In addition to ICOD and covariance, we consider conservation as a baseline, since it is known to be efficient at identifying sites under selection in natural protein sequence data. How well do these different eigenvectors, or conservation, recover $\vec{D}$ when varying the amount of phylogeny? Fig 4 shows the recovery versus the number of mutations per branch μ for ICOD, covariance, and conservation. We observe that the eigenvector associated to the largest eigenvalue Λ_{max} of ICOD, and the one associated to the smallest eigenvalue λ_{min} of covariance (blue), contain more information than those at the opposite end of the respective spectra (purple). This is expected from our theoretical analysis, as the eigenvector with smallest variance should be close to $\vec{D}$ (see above and [31]). Moreover, we observe that ICOD is more robust than covariance to phylogeny, and performs much better than covariance in strong phylogenetic regimes (small μ). Both methods have recoveries close to the maximum possible value of 1 for intermediate to weak phylogeny (larger μ), meaning that they recover almost perfectly $\vec{D}$. Furthermore, ICOD outperforms conservation, except for very small μ (very strong phylogeny). Conservation performs better than covariance with strong phylogeny, while the opposite holds with weaker phylogeny. We note

that the eigenvector associated to the smallest ICOD eigenvalue Λ_{min} also contains information about $\vec{D}$, as it outperforms the null model corresponding to recovery by a random vector (see Methods and [31]). The corresponding eigenvector of the covariance is less informative.

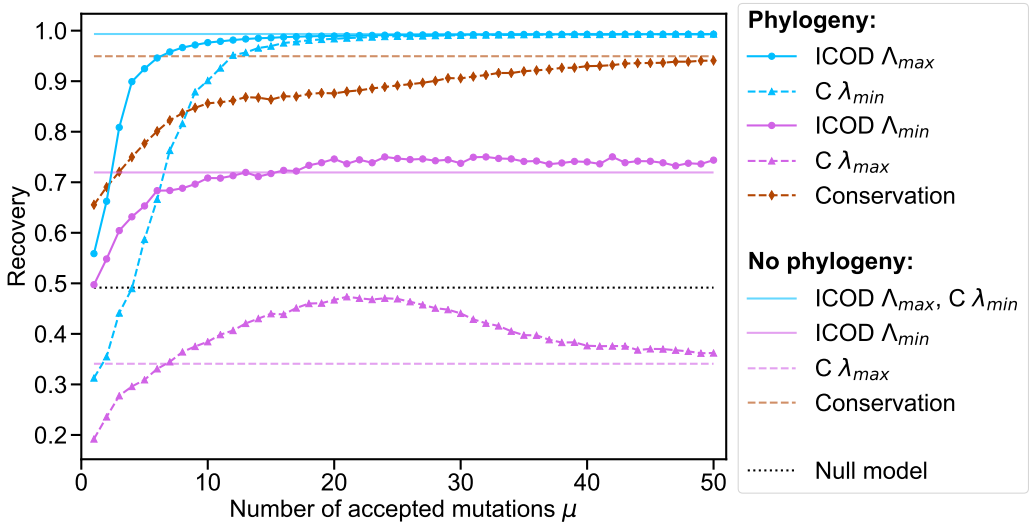


Figure 4: **Impact of phylogeny on mutational effect recovery.** The recovery of the mutational effect vector $\vec{D}$ (see Methods) for specific eigenvectors is shown as a function of the number μ of mutations per branch, using ICOD and covariance. Recovery by conservation is also shown. For ICOD (resp. covariance C), eigenvectors associated to the largest eigenvalue Λ_{max} (resp. λ_{max}) and smallest eigenvalue Λ_{min} (resp. λ_{min}) are considered. Data was generated as in Fig 3, with various values of μ. All results are averaged over 100 realisations of data generation. The null model corresponds to recovery from a random vector (see Methods, Eq 13).

How do these results obtained by ICOD compare to SCA? Recoveries of $\vec{D}$ from SCA are shown in Fig S3, and compared to those of ICOD and conservation. We observe that ICOD is more robust to phylogeny than SCA and generally performs better. Interestingly, we see that conservation yields better recoveries than SCA in strong phylogenetic regimes, while the opposite holds for intermediate to weak phylogeny, where SCA reaches higher recovery scores. Recall that the SCA matrix is constructed using both conservation and covariance.

While we focused on the largest and smallest eigenvalues because they are expected to contain most of the signal about the sector, it is interesting to look at the recovery of all eigenvectors. Fig S2 confirms that the best recoveries (closest to 1) are found in the eigenvector associated to the largest (rank 0) eigenvalue for ICOD and SCA, and to the smallest one (rank 199) for covariance. However, these recoveries decrease with increased phylogeny (i.e. for smaller μ), but ICOD is most robust, consistent with our previous results. For ICOD, we further observe relatively large recovery scores for $L_S - 1$ small eigenvalues (here $L_S = 20$, so the associated ranks are 180-199), in agreement with our formal analysis of the spectrum and our expectation that smallest eigenvalues should regard the sector. Fig S2 shows that this holds both without and with phylogeny.

Impact of selection parameters on mutational effect recovery

Impact of the favored trait value. To what extent is recovery impacted by the favored trait value τ^*, at various levels of phylogeny? To investigate this, we generate data using our minimal model for various values of τ^*. Results are presented in Fig 5 for ICOD, covariance, SCA and conservation for three data sets with various levels of phylogeny. First, we observe that ICOD and covariance perform well at small τ^*, and then recovery decays and gets poor (below null model) around $\tau^* = \sum_i D_i (\approx 196)$ before it gets somewhat better for ICOD at large τ^*. SCA performs rather well at small and intermediate τ^* (except $\tau^* = 0$) but its performance deteriorates as τ^* increases. Finally, conservation performs well overall and reaches a maximum

around $\tau^* = \sum_i D_i (\approx 196)$. Moreover, we observe that ICOD and conservation are more robust to phylogeny than covariance and SCA.

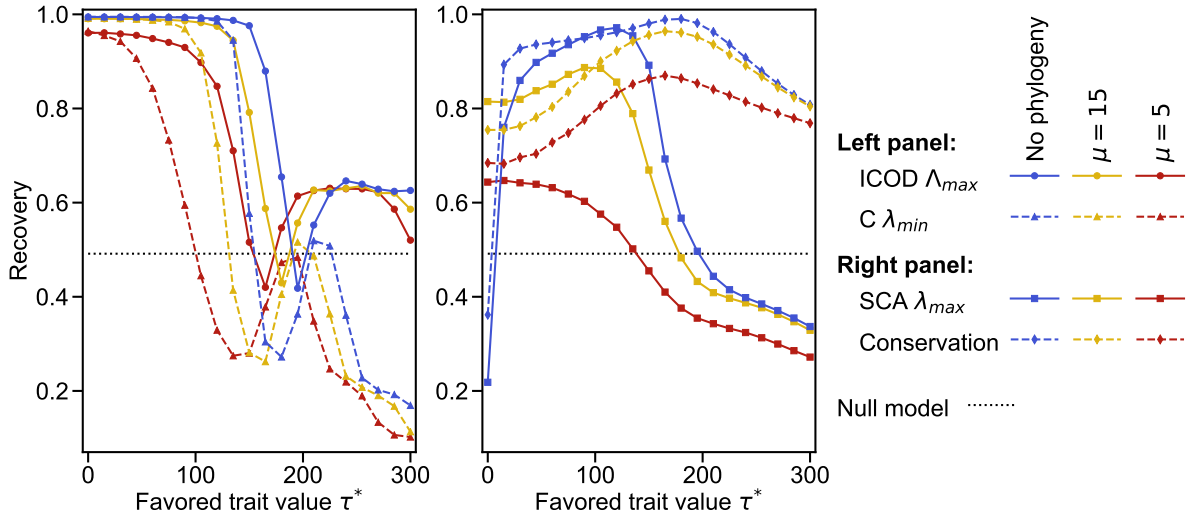


Figure 5: **Impact of the favored trait value on mutational effect recovery.** The mutational effect recovery is shown as a function of the favored trait value τ^* for ICOD and covariance (left panel), and for SCA and conservation (right panel). As in Fig 3, we consider different levels of phylogeny by considering different values of μ (shown as different colors). For ICOD (resp. SCA), eigenvectors associated to the largest eigenvalue Λ_{max} (resp. λ_{max}) are considered, while for covariance C, the eigenvector associated to the smallest eigenvalue λ_{min} is considered. The null model corresponds to recovery from a random vector (see Methods, Eq 13). The data is generated exactly as in Fig 3, except that τ^* is varied. All results are averaged over 100 realisations. Results for $\tau^* \geq 0$ are shown, and those for $\tau^* < 0$ can be obtained by symmetry, see Eq 1.

In Fig S6, we use symmetrized AUC (area under the receiver operating curve, see methods) to assess performance of sector site identification on the same data with the same methods. While ICOD and covariance display similar behaviours with symmetrized AUC as with recovery, SCA and conservation reach better performance with symmetrized AUC, especially for large τ^* or with little phylogeny. Indeed, conservation identifies sites associated to the sector as highly conserved, but does not recover mutational effect values, and SCA may inherit this property from conservation. However, SCA is quite hindered by phylogeny at small τ^*, which limits the interest of the contribution of conservation.

Since we showed that the other end of the ICOD spectrum, corresponding to small eigenvalues, also contains information on the sector, we also compute recovery and symmetrized AUC from that end of the spectrum for ICOD, covariance and SCA, see Fig S5 and S7. Those results confirm that the smallest ICOD eigenvalue Λ_{min} contains some information on the sector, especially at large values of τ^*. Meanwhile, results for covariance and SCA are poor for that end of the spectrum.

Fig 5, S6, S5 and S7 feature particular behaviours around $\tau^* = 190$. We remark that this value of τ^* is close to $\sum_i D_i \approx 196$, which is the value taken by the trait τ when all sites have state $+1$. Given that sector sites have much stronger mutational effects than others, reaching this value of the trait essentially requires all sector sites have state $+1$, which makes them strongly conserved. Accordingly, the performance of conservation in Fig 5 peaks around $\tau^* = 190$. Indeed, for larger τ^*, even non-sector sites have to be conserved, including those with negative mutational effect, impairing recovery but not symmetrized AUC (as sector sites remain the most conserved ones), see Fig S6.

The ICOD spectrum and the corresponding eigenvectors are analyzed in more detail in Fig S9 for different values of τ^*. In the spectrum, the number of positive outliers increases when τ^* is increased, and these eigenvalues become very large, while the number of negative outliers reduces to one. For large τ^*, sector sites are extremely conserved, and each of them gives rise to a direction of very low variance [31], which results in very large eigenvalues of the ICOD matrix due to the matrix inversion step. In this regime, we do not observe such a split of sector signal at the opposite end of the spectrum (most negative eigenvalue),

which contains information on the sector as well (see above). Accordingly, the eigenvector associated to the largest eigenvalue Λ_{max} of the ICOD matrix recovers the mutational effect vector $\vec{D}$ very well for moderate τ^*, but possesses large components only on a few of the sector sites for larger τ^*. Meanwhile, the eigenvector associated to Λ_{min} has large components on all sector sites for larger τ^*, which is not the case for smaller τ^*.

Impact of selection strength on mutational effect recovery. How do selection strength and phylogeny combine and impact the performance of mutational effect recovery? To address this, we generate data using our minimal model for various values of selection strength κ. Results are shown in Fig 6 for ICOD and covariance, with two values of τ^* and various levels of phylogeny. For $\tau^* = 90$, mutational effect recovery by ICOD and covariance is quite robust to selection strength κ, apart from the very weak selection case where recovery is low because sequences are noisy. Furthermore, covariance is more negatively impacted by phylogeny than ICOD across values of κ. For $\tau^* = 140$, we observe that recovery tends to decrease when κ increases (ignoring the very weak selection regime). However, this only happens with substantial phylogeny ($\mu = 5$) for ICOD, while this effect is always present for covariance, but becomes stronger when phylogeny is stronger. In this case, phylogeny makes recovery by covariance poor, presumably because the combination of strong phylogeny and large τ^* yields strong conservation. Consistently, Fig S8 shows that conservation yields good recovery for $\tau^* = 140$, which does not deteriorate too much due to phylogeny. The latter figure also shows that recovery by SCA is impaired by strong selection and phylogeny. Overall, ICOD and conservation are the most robust methods to stronger selection and phylogeny.

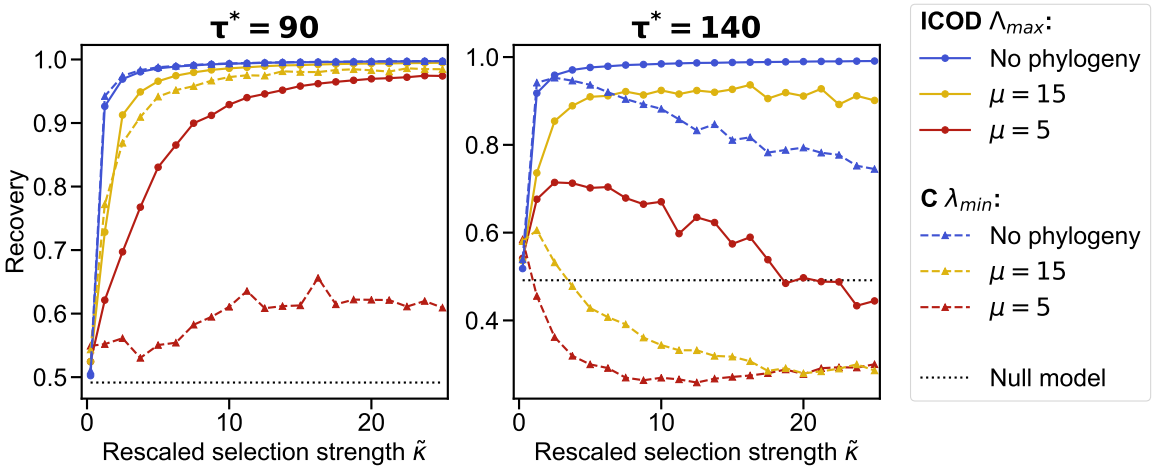


Figure 6: **Impact of selection strength on mutational effect recovery.** Recovery by covariance and ICOD is shown as a function of the rescaled selection strength $\tilde{\kappa} = \kappa \sum_i D_i^2$ for three data sets generated with different levels of phylogeny, and with favored trait value $\tau^* = 90$ (left) or $\tau^* = 140$ (right). For ICOD, the eigenvector associated to the largest eigenvalue Λ_{max} is considered, while for covariance C, the eigenvector corresponding to the smallest eigenvalue λ_{min} is considered. The data is generated as in Fig 3, except that κ is varied. All results are averaged over 100 realisations.

Impact of phylogenetic correlations

How does phylogeny impact the components of the key eigenvectors of the ICOD, covariance and SCA matrices? To investigate this question, we generate data using our minimal model and keeping track of all intermediate ancestral sequences in the phylogenetic tree. Next, we assign to each site of the sequence a score G, which corresponds to the earliest generation when the site mutates (and thus becomes different from the initial state in the ancestral sequence). In Fig 7, we examine how the eigenvector components relate to G, for two data sets, generated respectively with small and large amounts of phylogeny. In the weak phylogeny regime, ICOD, C and SCA all yield much larger eigenvector components for sector sites than for other sites. Conversely, with strong phylogeny, the eigenvector components associated with sector and non-sector sites

strongly overlap when using covariance and SCA. Remarkably, ICOD is still able to disentangle them. We also note that covariance scores are strongly affected by G: sites that mutate late have stronger scores, presumably because they are quite conserved and feature little variance, resulting in a high contribution to the eigenvector associated with the smallest eigenvalue of the covariance matrix. Furthermore, Fig S10 considers the eigenvector associated with the largest eigenvalue of the covariance matrix (other end of the spectrum) under strong phylogeny. There, sites that mutate early in the phylogeny obtain high scores. Thus, the spectrum of the covariance matrix is strongly impacted by phylogeny. Similarly, SCA attributes higher scores to sites that mutate relatively early in the phylogeny and might be less conserved, leading to a higher variance and high contributions to the eigenvector associated with the largest eigenvalue of the SCA matrix.

Fig 7 further shows that sector sites tend to mutate later (higher G score) than other sites. Indeed, selection impedes mutation of sites with high mutational effects, as these mutations tend to substantially deteriorate fitness.

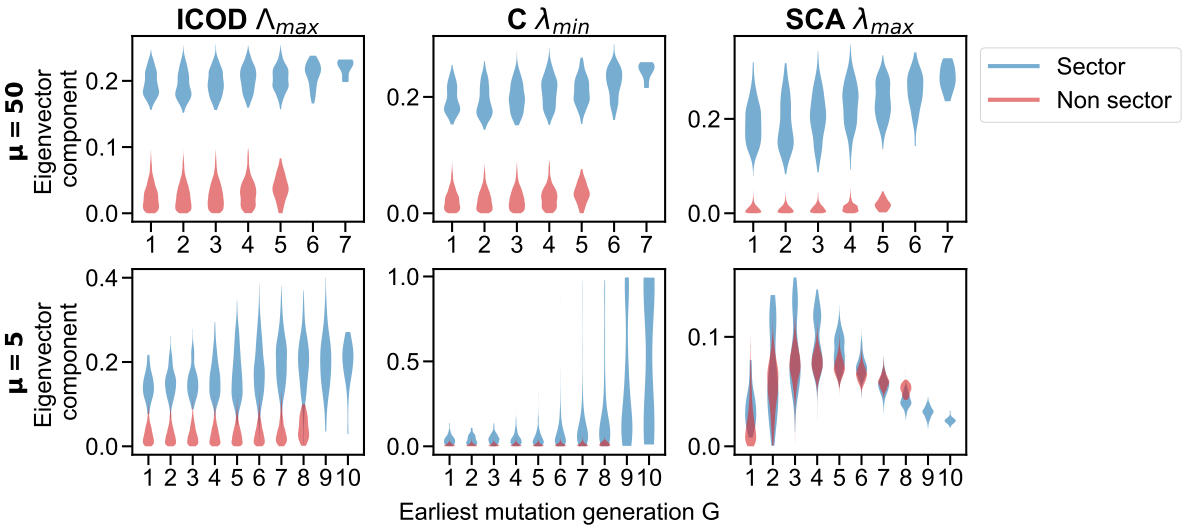


Figure 7: **Impact of earliest mutation generation G on eigenvector components.** Violin plots of the absolute value of components of the key eigenvectors of the ICOD, covariance C and SCA matrices are represented versus the earliest mutation generation G at which the associated site first mutates in the phylogeny. Results are shown for data sets generated with $\mu = 50$ (top panels) and $\mu = 5$ (bottom panels). Other parameter values are the same as in Fig 3. Violin plots are obtained over 100 realisations of data generation.

Identifying functionally important sites in natural protein families

How does ICOD perform at predicting sites with large mutational effects on natural data? Does its robustness to phylogeny give it an advantage? To address this, we consider natural protein families with published Deep Mutational Scan (DMS) experimental data, which we consider as ground truth. We investigate 30 protein families [46] listed in table S1. For each of them, we rank sites by the magnitude of predicted mutational effects using ICOD, SCA, Mutual Information (MI, see [45]), and conservation. We compare these predictions to DMS data. Specifically, for each protein family, we start from the reference sequence mutated in the DMS experiment, and construct a multiple sequence alignment of homologs of this protein whose depth can be varied by selecting the neighbors of the reference sequence in Jukes-Cantor distance up to a given phylogenetic cutoff [45]. Next, for each method, we compute the top eigenvector on MSAs constructed using different phylogenetic cutoffs, and then we summed component by component the eigenvectors corresponding to these different cutoffs. We use the components of the resulting vector as predictors of mutational effects, and the DMS data as ground truth. Fig 8 shows that all methods are able to predict important sites, the most successful ones being conservation and ICOD. Indeed, the average of the symmetrized AUC over all families is 0.45 for conservation, 0.43 for ICOD, 0.35 for MI and 0.26 for SCA. Furthermore, we observe that better

results tend to be obtained for DMS with bimodal distributions of fitness effects (see Fig S12 for examples). This suggests that when the data comprises sites with substantially stronger mutational effects than others for the function probed by the DMS (i.e. a functional sector), then this sector can be well identified by the methods considered here, in particular conservation and ICOD.

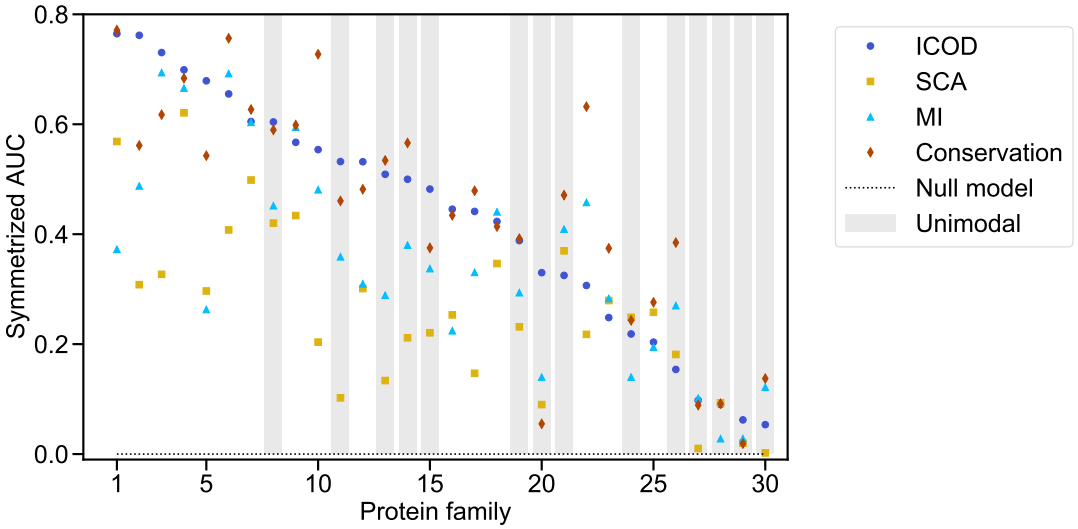


Figure 8: **Identifying functionally important sites in natural protein families.** The symmetrized AUC for the prediction of sites with large mutational effects is computed on 30 protein families, using four different methods: ICOD, SCA, MI and Conservation, using Deep Mutational Scan (DMS) data as ground truth. For ICOD and MI, the average product correction (APC) [4] is applied to the matrix of interest (it was found to improve the average performance over families for these methods, but not for SCA). For ICOD, MI and SCA, the components of the eigenvector associated to the largest eigenvalue are employed to make predictions of mutational effects. Protein families are ordered by decreasing symmetrized AUC for ICOD. The mapping between protein family number and name is given in Table S1. The protein families shaded in grey have DMS data featuring a unimodal shape, the other ones have a bimodal shape.

We note that in many cases, focusing on specific cutoffs yielded better performance than the sum of eigenvectors over cutoffs considered here. However, optimal cutoffs differed between families and methods, making it difficult to exploit this for prediction. For completeness, Fig S11 shows the best performance obtained across cutoffs. The average of the symmetrized AUC over all families is 0.47 for conservation and for ICOD, 0.42 for MI and 0.37 for SCA. Thus, our conclusion that ICOD and conservation perform best is robust to considering the best cutoff instead of summing over cutoffs.

While conservation and ICOD yield the best identification of functionally important sites, they rely on entirely different signals. Indeed, ICOD was designed to focus primarily on correlations and not on conservation – recall that the diagonal of the inverse covariance matrix is set to zero in order to eliminate conservation. Thus, we expect these two methods to find different important sites. Table S1 shows the number of sites identified as functionally important by both ICOD and conservation ("Both") or only by either method ("ICOD" or "Cons."), out of those that are deemed important by the DMS, i.e. the L_S sector sites found by binarising the DMS data and setting a cutoff (see Methods). We find that predictions can indeed differ. For instance, β-lactamase has 12 sites that are predicted by ICOD only and 8 sites by conservation only, although the two methods yield a similar symmetrized AUC. We examined the 3D structures of HRas and PDZ and the location of the sites identified by both methods or by only one method. In these two cases, DMS data is obtained by focusing on ability to bind other molecules. Accordingly, functionally important sites identified in these DMS are located around interaction regions. We observe that the sites detected by ICOD only tend to be closer to these interfaces than those detected by conservation only. These sites have large DMS scores and are localized close to binding interfaces, strongly hinting at their functional importance, while they are not among the most conserved ones. Interestingly, ICOD allows to reveal such sites in these examples.

Discussion

It was shown in [31] that nonlinear selection acting on any additive functional trait of a protein gives rise to a functional sector in the sequence data associated to the protein family of interest. While the main signature of the selection process lies in the small-eigenvalue modes of the covariance matrix of sites, motivating ICOD, the large-eigenvalue modes were nevertheless affected [31]. Here, we showed that this arises from the mathematical properties of the covariance matrix provided that some sites have much larger mutational effects than others. We further generalized the analytical approximation of the elements of the ICOD matrix, thereby reinforcing the theoretical bases of ICOD.

Next, using synthetic data from a minimal model where we can tune the amounts of phylogeny and of nonlinear selection on an additive trait, we showed that phylogenetic correlations can substantially impair the inference of functional sectors from sequence data. However, ICOD and conservation are more robust to phylogenetic noise than SCA and covariance. The robustness of ICOD suggests that focusing on the small-eigenvalue modes of the covariance matrix is a successful strategy, as the large-eigenvalue ones tend to be most affected by phylogeny [35, 36]. Conservation has been used a lot to identify functionally important sites, and thus, its good performance is not surprising. In addition, it was argued in [48] that at least in some cases, the success of SCA can be attributed to its use of conservation. Here, we find that SCA's performance is indeed often tied to that of conservation. Importantly, in ICOD, the diagonal terms of the inverse covariance matrix are set to zero in order to eliminate the impact of conservation as much as possible, and focus on correlations. Thus, the robustness of ICOD cannot be explained by that of conservation, and the two methods should provide different and complementary information. While SCA aimed to combine the two important and complementary ingredients of conservation and covariance, our analysis of controlled synthetic data suggests that considering the two separate ingredients and using ICOD instead of covariance to further separate them may yield even more information.

Our analysis of several natural protein families for which DMS data is available showed that conservation and ICOD were the best predictors of mutational effects among the methods we considered. Our results regarding synthetic data suggest that this could be due to the robustness of these methods to phylogenetic correlations, which are bound to exist in natural sequence data. Here, we showed that phylogenetic correlations make the inference of functional sectors challenging, very much like they obscure the inference of structural contacts [4, 36–40]. It is important to note that phylogenetic correlations are nevertheless interesting and provide useful signal e.g. for the inference of protein partners among paralogs [47, 49, 50].

Our model of selection on an additive trait with a quadratic Hamiltonian is formally close to Potts models [7–23] and to generalized Hopfield models [51, 52]. Potts models have allowed mutational effect analysis [46, 53–55]. However, our selection model induces a specific dependence of the fields and of the couplings in the mutational effect vector which does not generically exist in these models. This is what makes it possible to recover mutational effects directly via an eigenvector of the ICOD matrix, instead of needing to compare inferred energies [54]. Generalizing beyond quadratic models would be very interesting [56, 57]. Besides, while we started from a generic model for sectors which involves an additive trait [32–34] under nonlinear natural selection, it would be interesting to analyze specific traits in more detail. One of them [31] is the energetic cost of protein deformations within an elastic-network model [58]. The low-energy deformation modes of protein structures are important in many functionally important protein deformations [59–63], and are robust to sequence variation within protein families [64–66]. It would be interesting to reexamine these cases with a more in-depth analysis of phylogenetic effects.

Our work provides a step towards understanding the impact of the rich structure of biological data on the performance of inference methods [67]. We focused on traditional and principled inference methods, because these interpretable methods allow us to get direct insight into the origin of method performance. However, we believe that the insight gained can also be useful for deep learning approaches. In particular, the question of disentangling purely phylogenetic signal from functional signal is important for all methods. Furthermore, many current deep learning models start from the same data structure as the methods considered here, namely MSAs. For instance, AlphaFold, which has brought major advances to the computational prediction of protein structures from sequences, starts by constructing an MSA of homologs when given a protein sequence as input, and its performance is strongly impacted by MSA properties [68]. Closer to our analysis, DeepSequence [46] employs a variational autoencoder trained on MSAs to predict mutational effects, and reaches strong performance. While protein language models relying on non-aligned protein sequences also

allow successful mutational effect prediction [69], their performance is strongly impacted by the number of homologs that a sequence possesses, and increases with it [70], hinting that homologs remain important in these methods.

Methods

Generating synthetic sequences

To assess the impact of phylogeny on the prediction of mutational effects sectors, we generate sequences using our minimal model, either without phylogeny or with various levels of phylogeny.

Generating independent equilibrium sequences. A Metropolis Monte Carlo algorithm is used to sample equilibrium and independent sequences according to the Hamiltonian in Eq 1, see Fig 1**A**. To generate each sequence, we start from a random sequence and let it evolve by accepting a fixed number of mutations (spin flips), chosen large enough for sequences to equilibrate – in practice, 3000, which yields convergence of the energy of sequences. The probability p that a proposed mutation is accepted is given by the Metropolis criterion

$$p = \min\left[1, \exp\left(-\Delta H\right)\right], \tag{2}$$

where ΔH is the difference of the value of the Hamiltonian in Eq 1 with and without the proposed mutation.

In this way, we generate sequences sampled from the distribution $P(\vec{\sigma}) = \exp\left(-H(\vec{\sigma})\right) / \sum_{\vec{\sigma}} \exp\left(-H(\vec{\sigma})\right)$. The resulting sequences have trait values distributed around a favored value τ^*, and the selection strength κ regulates how much the trait values deviate from the favored value.

Generating sequences with phylogeny. In order to incorporate phylogeny, we take a functional sequence generated as above as the ancestor, and let it evolve on a binary branching tree, see Fig 1**B**. A fixed number μ of accepted mutations are performed independently along each branch, and the sequences sequences at the tree leaves constitute our MSA. Mutations are accepted with probability p from Eq 2, which maintains the same selection on the trait. The resulting sequences contain correlations from phylogeny (controlled by μ) in addition to those coming from selection. If μ is large enough, even sister sequences become independent and equilibrium statistics are recovered. Conversely, if μ is small, sister sequences are very similar and phylogeny is strong. Note that we can also generate sequences wit phylogeny and no selection using the same method, but accepting all proposed mutations.

Natural data

Comparing to experiments. Deep Mutational Scan (DMS) experiments provide a direct measurement of the fitness effect associated to each possible point mutation at each site of a reference sequence. In DMS experiments, selection for a given function (e.g. ability to bind to another molecule) is applied on the mutated and the wild type sequences. The ratio of the number of sequences after and before selection is called r_{WT} for the wild-type and r_{mut} for the mutant. Most often, the fitness effect of the particular mutation is assessed via the logarithm of the enrichment ratio $\log\left(r_{mut}/r_{WT}\right)$. Functionally important sites are those for which mutations are most costly, leading to a small r_{mut} and a negative log enrichment score. Thus, we construct the analogue of our vector $\vec{D}$ of mutational effects by taking for each site the most negative (smallest) enrichment ratio, associated with the most deleterious mutation at this site. We start from the 30 DMS datasets collected in [46]. We binarise the minimum enrichment scores from the DMS data. We observe that most score distributions are in practice either bimodal or unimodal, see Fig S12 for some examples. Thus, we fit for each family either a linear combination of two Gaussian distributions or a single Gaussian distribution to find a cutoff value for the binarisation. For the unimodal families, we use the mean of the fitted Gaussian distribution as the cutoff, while for the bimodal ones we use the position of the minimum between the two peaks. In some cases, there are more than two peaks in the distribution: then, we fit a bimodal distribution to the two most negative peaks, see upper middle panel in Fig S12. In all cases, we define the sector as comprising the sites with values more negative than the cutoff value. The number of thus-defined sector sites is called L_S.

MSA construction. In order to infer sectors in natural data, we need to construct Multiple Sequence Alignments (MSAs) of homologs of the reference sequence of the DMS experiment. For this, we use the method from reference [46]. We search the reference sequence against the UniRef100 database using `jackhmmer` from the HMMER3 suite [71], specifically the command `"jackhmmer --incdomT 0.5*L --cpu 6 -N 5 -A pathsave pathrefseq pathuniref"`, where L is the length of the wild type sequence. We restrict to match states where the reference sequence does not have a gap, remove columns with more than 30% gaps, and remove sequences which have more than 20% gaps. Inspired by reference [45], we further subsample the MSA to generate a collection of MSAs comprising neighbors of the reference sequence up to different phylogenetic cutoffs. Specifically, we only retain sequences within a given Jukes-Cantor distance, termed cutoff distance, of the reference sequence. Each gap of the MSA is then replaced by the amino acid possessed at the same site by the nearest sequence in terms of Jukes-Cantor distance. We employ the following values of cutoff distance: 0.4, 0.6, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9 and 2.0, yielding 15 MSAs for each protein family. Note that for protein families 1, 3, and 27, MSAs were generated only up to cutoffs 1.2, 1.4 and 1.4 respectively, due to computational constraints.

Inferring mutational effects

In an MSA, each column is a protein site and each row is a sequence. All methods considered here rely on single-site and two-site frequencies of amino acids. Single-site frequencies are denoted by $f_i(\sigma_i)$ for a given state (or amino acid) σ_i on the ith site of the protein. Similarly, two-site frequencies are denoted by $f_{ij}(\sigma_i, \sigma_j)$. The state σ_i can take values $(-1, 1)$ for synthetic data and $(0, ..., 19)$ for natural data, representing the 20 natural amino acids. To avoid divergences due to states that do not appear in the data for ICOD or Mutual information, a pseudocount a is introduced [7–9], leading to pseudocount-corrected frequencies

$$\tilde{f}_i(\sigma_i) = a/q + (1-a) f_i(\sigma_i)\,, \tag{3}$$

$$\tilde{f}_{ij}(\sigma_i, \sigma_j) = a/q^2 + (1-a) f_{ij}(\sigma_i, \sigma_j) \quad \text{for} \quad i \neq j\,, \tag{4}$$

$$\tilde{f}_{ii}(\sigma_i, \sigma_j) = \delta_{\sigma_i \sigma_j} \tilde{f}_i(\sigma_i)\,, \tag{5}$$

where q is the number of states, i.e. $q = 2$ for synthetic data and $q = 20$ for natural data.

In the case of synthetic data, we apply each method (ICOD, covariance, SCA, conservation) to the generated MSA. In the case of natural data, we have 15 different MSAs for each protein family. We apply each method (ICOD, MI, SCA, conservation) on each MSA and extract the eigenvector associated to the eigenvalue of focus. This yields 15 different eigenvectors that are added together component by component. Because normalized eigenvectors are defined up to an overall sign, we set the sign of one eigenvector arbitrarily and choose the one of the other eigenvectors such that they have a positive Pearson correlation with the first eigenvector. The overall sign of the sum remains arbitrary, but we employ scoring methods that are invariant to it.

Conservation. We use an entropy-based definition of conservation [48]. For a given site i (ith column in the MSA), it reads

$$\text{Conservation}_i = \log_q(q) - \sum_{\sigma_i} f_i(\sigma_i) \log_q\left(f_i(\sigma_i)\right) = 1 - \sum_{\sigma_i} f_i(\sigma_i) \log_q\left(f_i(\sigma_i)\right)\ , \tag{6}$$

where q is the number of states and $\log_q$ denotes the logarithm of base q. Specifically, this conservation measure is the Kullback-Leibler divergence of the amino acid frequencies with respect to the uniform distribution, which is taken as reference for simplicity (another possibility would be to use the background amino-acid frequencies [48]). The conservation score for each site i can directly be compared to the mutational effect vector $\vec{D}$ for synthetic data, or to the DMS score for natural data.

Covariance. The elements of the covariance matrix of sites can be calculated as

$$C_{ij}(\sigma_i, \sigma_j) = f_{ij}(\sigma_i, \sigma_j) - f_i(\sigma_i) f_j(\sigma_j)\ . \tag{7}$$

Note that for ICOD, we use the pseudocount-corrected frequencies here (see above).

For synthetic data, in the two state case $(-1, 1)$ with $q = 2$, we use for covariance and ICOD the standard covariance definition

$$C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \,, \tag{8}$$

where $\langle \cdot \rangle$ denotes a mean across all sequences of the MSA [31]. Employing it is equivalent to using Eq 7 and restricting to one state, considering the other as reference, which can be done because normalization entails that the second state gives redundant information: $f_i(1) + f_i(-1) = 1$ [31]. For ICOD, the covariance matrix is computed with the pseudocount-corrected frequencies (see Eq 3-5), and thus its elements read $C_{ij}^{(a)} = (1-a)\langle \sigma_i \sigma_j \rangle - (1-a)^2 \langle \sigma_i \rangle \langle \sigma_j \rangle$ for $i \neq j$ and $C_{ii}^{(a)} = (1-a)^2(1 - \langle \sigma_i \rangle^2) + a(2-a)$. Meanwhile, for SCA, the full covariance matrix with all states (i.e. Eq 7) is used (see below).

For natural data, C is a $20 \times L$ by $20 \times L$ matrix where for each pair of sites i, j there is a sub-matrix of size 20×20 comprising each possible pair of amino acids. The Frobenius norm of each of these sub-matrices weighted by conservation is used in SCA (see [30] and below), while in ICOD we use the reference-sequence gauge and eliminate one redundant state (see below).

ICOD. ICOD is a method introduced in [31], where the covariance matrix of sites is inverted (which requires using a pseudocount), and its diagonal terms are all set to zero. The latter allows to mitigate the impact of conservation [31]. Using the mean-field approximation of Potts model inference [7–9] it was shown in [31] that in the two-state case, starting from the covariance matrix in Eq 8, the ICOD matrix can be approximated as

$$(\tilde{C}^{-1})_{ij} \approx (1 - \delta_{ij}) \kappa D_i D_j \,. \tag{9}$$

In practice, we use a pseudocount value $a = 1 \times 10^{-5}$ for synthetic data.

For natural data, a similar construction can be made by employing the covariance matrix in Eq 7 with pseudocount-corrected frequencies (Eq 3 and 4), with a pseudocount value $a = 0.05$. The reference-sequence gauge [31] is used, and we do not compute the frequency of the amino acid in the reference sequence at a given position, yielding a covariance matrix of size $(19 \times L, 19 \times L)$. This matrix is then inverted, the Frobenius norm is taken on each submatrix of size 19×19 corresponding to each pair of sites (i, j). Next, the diagonal is set to zero.

SCA. Statistical Coupling Analysis (SCA), introduced in [27, 30, 72], is a method that combines covariance and conservation. The elements of the SCA matrix are defined as

$$\tilde{C}_{ij}^{SCA}(\sigma_i, \sigma_j) = \phi_i(\sigma_i) \phi_j(\sigma_j) C_{ij}(\sigma_i, \sigma_j) \,, \tag{10}$$

where $C_{ij}(\sigma_i, \sigma_j)$ is an element of the full covariance matrix (Eq 7), while $\phi_i(\sigma_i)$ characterizes amino conservation on site i. The Frobenius norm of each 20×20 block of this matrix, which corresponds to a pair of sites i, j, is then computed, and it is on this matrix that we perform an eigendecomposition. More details on this method can be found in [30], and we use the corresponding implementation, namely the `pySCA` github package (https://github.com/reynoldsk/pySCA). For the natural data, we use default parameters [30]. For the synthetic case, we transformed our data from $(-1, 1)$ states to $(0, 1)$ states, and in `pySCA` we changed the number of states from 20 to 2 and the background frequency to 0.5 for each state.

Mutual Information (MI). The MI of each pair of sites (i, j) is computed as

$$\mathrm{MI}_{ij} = \sum_{\sigma_i, \sigma_j} \tilde{f}_{ij}(\sigma_i, \sigma_j) \log \left(\frac{\tilde{f}_{ij}(\sigma_i, \sigma_j)}{\tilde{f}_i(\sigma_i) \tilde{f}_j(\sigma_j)} \right) , \tag{11}$$

where $\tilde{f}_i(\sigma_i)$ and $\tilde{f}_{ij}(\sigma_i, \sigma_j)$ are the pseudocount-corrected one- and two-body frequencies (see Eq 3-5). We set the pseudocount value to $a = 0.001$. Note that we use frequencies instead of probabilities to estimate mutual information, and do not correct for finite-size effects [73]. This is acceptable because we only compare scores computed on data sets with a given size, affected by the same finite size effects.

Performance evaluation

In order to quantify how well methods perform, we use the recovery score introduced in [31]. The recovery of the vector $\vec{D}$ of mutational effects by any vector $\vec{v}$ is defined as

$$\text{Recovery} = \frac{\sum_i |v_i D_i|}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i D_i^2}}, \tag{12}$$

The chance expectation of the recovery, corresponding to recovery by a random vector, is [31]

$$\langle \text{Recovery} \rangle \approx \sqrt{\frac{2}{\pi L}} \frac{\sum_i |D_i|}{\sqrt{\sum_i D_i^2}} \ . \tag{13}$$

We also evaluated performance using the AUC, i.e. area under the receiver operating characteristic. More precisely, we used the symmetrized AUC, defined as $2(|\text{AUC} - 0.5|)$. This allows us to account for the fact that eigenvectors are defined up to an overall sign.

For synthetic data, we mainly use the recovery score because the eigenvectors are likely to match the full vector $\vec{D}$. Indeed, there is only selection on one function with a quadratic nonlinearity and it is well captured by our methods. The symmetrized AUC is also studied in some cases and there $\vec{D}$ is binarised accordingly to sites belonging to the sector or not. In general, the symmetrized AUC focuses on the ranking of important sites (sector sites), while recovery focuses on the similarity between the eigenvector considered and $\vec{D}$.

For natural data, we mainly use the symmetrized AUC to evaluate performance, because DMS only test one aspect of function and the measured fitness can involve additional nonlinearities not captured by our simple model.

Acknowledgments

A.-F. B. thanks Ned S. Wingreen and Shou-Wen Wang for inspiring discussions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 851173, to A.-F. B.).

Code accessibility

Our code is freely available at https://github.com/Bitbol-Lab/Phylogeny-sector-inference.

References

[1] U. Gobel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–317, Apr 1994.

[2] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, 271(4):511–523, Aug 1997.

[3] A. S. Lapedes, B. G. Giraud, L. Liu, and G. D. Stormo. Correlated mutations in models of protein sequences: phylogenetic and structural effects. In *Statistics in molecular biology and genetics - IMS Lecture Notes - Monograph Series*, volume 33, pages 236–256. 1999.

[4] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb 2008.

[5] J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, and M. T. Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6):1043–1054, Jun 2008.

[6] L. Burger and E. van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.*, 4:165, 2008.

[7] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):67–72, Jan 2009.

[8] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12):e28766, 2011.

[9] 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. *Proc. Natl. Acad. Sci. U.S.A.*, 108(49):E1293–1301, Dec 2011.

[10] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic. Genomics-aided structure prediction. *Proc. Natl. Acad. Sci. U.S.A.*, 109(26):10340–10345, Jun 2012.

[11] F. Morcos, B. Jana, T. Hwa, and J. N. Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.*, 110(51):20533–20538, Dec 2013.

[12] R. S. Dwyer, D. P. Ricci, L. J. Colwell, T. J. Silhavy, and N. S. Wingreen. Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. *Genetics*, 195(2):443–455, Oct 2013.

[13] R. R. Cheng, F. Morcos, H. Levine, and J. N. Onuchic. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.*, 111(5):E563–571, Feb 2014.

[14] D. Malinverni, S. Marsili, A. Barducci, and P. De Los Rios. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS Comput. Biol.*, 11(6):e1004262, Jun 2015.

[15] Anne-Florence Bitbol, Robert S Dwyer, Lucy J Colwell, and Ned S Wingreen. Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 113(43):12180–12185, 2016.

[16] T. Gueudre, C. Baldassi, M. Zamparo, M. Weigt, and A. Pagnani. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 113(43):12186–12191, 10 2016.

[17] R. R. Cheng, O. Nordesjo, R. L. Hayes, H. Levine, S. C. Flores, J. N. Onuchic, and F. Morcos. Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. *Mol. Biol. Evol.*, 33(12):3054–3064, 12 2016.

[18] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol. Biol. Evol.*, 33(1):268–280, Jan 2016.

[19] G. Croce, T. Gueudré, M. V. Ruiz Cuevas, V. Keidel, M. Figliuzzi, H. Szurmant, and M. Weigt. A multi-scale coevolutionary approach to predict interactions between protein domains. *PLoS Comput Biol*, 15(10):e1006891, 10 2019.

[20] Q. Cong, I. Anishchenko, S. Ovchinnikov, and D. Baker. Protein interaction networks revealed by proteome coevolution. *Science*, 365(6449):185–189, 07 2019.

[21] J. A. de la Paz, C. M. Nartey, M. Yuvaraj, and F. Morcos. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc Natl Acad Sci U S A*, 117(11):5873–5882, 03 2020.

[22] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, and R. Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 07 2020.

[23] A. G. Green, H. Elhabashy, K. P. Brock, R. Maddamsetti, O. Kohlbacher, and D. S. Marks. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat Commun*, 12(1):1396, 03 2021.

[24] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, Oct 1999.

[25] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10(1):59–69, Jan 2003.

[26] Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512, 2005.

[27] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[28] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chakraborty. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl. Acad. Sci. U.S.A.*, 108(28):11530–11535, Jul 2011.

[29] Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138, 2012.

[30] Olivier Rivoire, Kimberly A Reynolds, and Rama Ranganathan. Evolution-based functional decomposition of proteins. *PLoS Comput. Biol.*, 12(6):e1004817, 2016.

[31] S-W Wang, A-F Bitbol, and NS Wingreen. Revealing evolutionary constraints on proteins through sequence analysis. *PLoS Comput Biol*, 15(4):e1007010, 2019.

[32] J. Otwinowski, D. M. McCandlish, and J. B. Plotkin. Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U.S.A.*, 115(32):E7550–E7558, 08 2018.

[33] M. A. DePristo, D. M. Weinreich, and D. L. Hartl. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.*, 6(9):678–687, Sep 2005.

[34] T. N. Starr and J. W. Thornton. Epistasis in protein evolution. *Protein Sci.*, 25(7):1204–1218, 07 2016.

[35] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nat. Struct. Biol.*, 2(2):171–178, Feb 1995.

[36] C. Qin and L. J. Colwell. Power law tails in phylogenetic systems. *Proc. Natl. Acad. Sci. U.S.A.*, 115(4):690–695, Jan 2018.

[37] Susann Vorberg, Stefan Seemayer, and Johannes Söding. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. *PLOS Computational Biology*, 14(11):1–25, 11 2018.

[38] Edwin Rodriguez Horta, Pierre Barrat-Charlaix, and Martin Weigt. Toward inferring potts models for phylogenetically correlated sequence data. *Entropy*, 21(11), 2019.

[39] Edwin Rodriguez Horta and Martin Weigt. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. *PLoS Comput Biol*, 17(5), 2021.

[40] Nicola Dietler, Umberto Lupo, and Anne-Florence Bitbol. Impact of phylogeny on structural contact inference from protein sequence data. *Journal of The Royal Society Interface*, 20(199):20220707, Feb 2023.

[41] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–358, Mar 1996.

[42] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, and E. Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E*, 87(1):012707, Jan 2013.

[43] A. J. Hockenberry and C. O. Wilke. Phylogenetic weighting does little to improve the accuracy of evolutionary coupling analyses. *Entropy*, 21(10), Oct 2019.

[44] D. Malinverni and A. Barducci. Coevolutionary Analysis of Protein Subfamilies by Sequence Reweighting. *Entropy*, 21(11):1127, Jan 2020.

[45] A. Colavin, E. Atolia, A.-F. Bitbol, and K. C. Huang. Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. *Scientific Reports*, 12:820, 2022.

[46] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, Oct 2018.

[47] A. Gerardos, N. Dietler, and A. F. Bitbol. Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences. *PLoS Comput Biol*, 18(5):e1010147, 05 2022.

[48] Tiberiu Teşileanu, Lucy J. Colwell, and Stanislas Leibler. Protein sectors: Statistical coupling analysis versus conservation. *PLOS Computational Biology*, 11(2):e1004091, Feb 2015.

[49] G. Marmier, M. Weigt, and A.-F. Bitbol. Phylogenetic correlations can suffice to infer protein partners from sequences. *PLoS Comput. Biol.*, 15(10):e1007179, Oct 2019.

[50] Carlos A. Gandarilla-Perez, Sergio Pinilla, Anne-Florence Bitbol, and Martin Weigt. Combining phylogeny and coevolution improves the inference of interaction partners among paralogous proteins. *PLoS Comput. Biol.*, 19(3):e1011010, 2023.

[51] S. Cocco, R. Monasson, and V. Sessak. High-dimensional inference with the generalized Hopfield model: principal component analysis and corrections. *Phys Rev E*, 83(5 Pt 1):051123, May 2011.

[52] Simona Cocco, Remi Monasson, and Martin Weigt. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLOS Comput. Biol.*, 9(8):e1003176, 2013.

[53] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta PI Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128, 2017.

[54] J. Rodriguez-Rivas, G. Croce, M. Muscat, and M. Weigt. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc Natl Acad Sci U S A*, 119(4), Jan 2022.

[55] F. Pucci, M. B. Zerihun, M. Rooman, and A. Schug. pycofitness-Evaluating the fitness landscape of RNA and protein sequences. *Bioinformatics*, 40(2), Feb 2024.

[56] W. C. Chen, J. Zhou, J. M. Sheltzer, J. B. Kinney, and D. M. McCandlish. Field-theoretic density estimation for biological sequence space with applications to 5' splice site diversity and aneuploidy in cancer. *Proc Natl Acad Sci U S A*, 118(40), Oct 2021.

[57] J. Zhou, M. S. Wong, W. C. Chen, A. R. Krainer, J. B. Kinney, and D. M. McCandlish. Higher-order epistasis and phenotypic prediction. *Proc Natl Acad Sci U S A*, 119(39):e2204233119, Sep 2022.

[58] Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.*, 39:23–42, 2010.

[59] P. De Los Rios, F. Cecconi, A. Pretre, G. Dietler, O. Michielin, F. Piazza, and B. Juanico. Functional dynamics of PDZ binding domains: a normal-mode analysis. *Biophys. J.*, 89(1):14–21, Jul 2005.

[60] M. Delarue and Y. H. Sanejouand. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J. Mol. Biol.*, 320(5):1011–1024, Jul 2002.

[61] W. Zheng and S. Doniach. A comparative study of motor-protein motions by using a simple elastic-network model. *Proc. Natl. Acad. Sci. U.S.A.*, 100(23):13253–13258, Nov 2003.

[62] Le Yan, Riccardo Ravasio, Carolina Brito, and Matthieu Wyart. Architecture and coevolution of allosteric materials. *Proc. Natl. Acad. Sci. U.S.A.*, 114(10):2526–2531, 2017.

[63] B. Bravi, R. Ravasio, C. Brito, and M. Wyart. Direct coupling analysis of epistasis in allosteric materials. *PLoS Comput Biol*, 16(3):e1007630, 03 2020.

[64] W. Zheng, B. R. Brooks, and D. Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U.S.A.*, 103(20):7664–7669, May 2006.

[65] S. Lukman and G. H. Grant. A network of dynamically conserved residues deciphers the motions of maltose transporter. *Proteins*, 76(3):588–597, Aug 2009.

[66] T. E. Saldano, A. M. Monzon, G. Parisi, and S. Fernandez-Alberti. Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLoS Comput. Biol.*, 12(3):e1004775, Mar 2016.

[67] V. Ngampruetikorn, V. Sachdeva, J. Torrence, J. Humplik, D. J. Schwab, and S. E. Palmer. Inferring couplings in networks across order-disorder phase transitions. *Phys. Rev. Research*, 4:023240, Jun 2022.

[68] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

[69] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

[70] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[71] Sean R. Eddy. Accelerated profile hmm searches. *PLOS Computational Biology*, 7(10):e1002195, Oct 2011.

[72] Steve W. Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, Oct 1999.

[73] William Bialek. *Biophysics: Searching for Principles*. Princeton University Press, 2012.

[74] J. Sjöstrand. Singularités analytiques microlocales. *Astérisque*, 95:III–166, 1982.

[75] Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019.

Supplementary material

Contents

S1 Analytical approximation of the inverse covariance matrix at higher orders **21**
S1.1 Complex saddle point approximation . . . 21
S1.2 High-temperature Plefka expansion of the TAP free energy . . . 22
S1.2.1 TAP free energy expansion . . . 23
S1.2.2 Extremized magnetizations . . . 23
S1.2.3 Inverse covariance matrix . . . 24
S1.3 Comparison to synthetic data . . . 24

S2 Supplementary figures **26**

S3 Supplementary table **33**

S1 Analytical approximation of the inverse covariance matrix at higher orders

S1.1 Complex saddle point approximation

Consider a sequence of L Ising spins $\{\sigma_l\}$, with $\sigma_l \in \{1, -1\}$ for all $l \in [1, L]$, with Hamiltonian

$$H = \frac{\kappa}{2}\Big(\sum_l \sigma_l D_l - \tau^*\Big)^2 + \sum_l \sigma_l f_l, \tag{14}$$

where f_l is a site-specific auxiliary field, which we use for calculations and then set to zero. The partition function is given by

$$Z = \sum_{\{\sigma_l\}} \exp(-H) = \sum_{\{\sigma_l\}} \exp\Big(-\frac{\kappa}{2}\Big[\sum_l \sigma_l D_l - \tau^*\Big]^2 - \sum_l \sigma_l f_l\Big). \tag{15}$$

Then, the magnetization $m_l \equiv \langle \sigma_l \rangle$ at site l reads

$$m_l = -\left.\frac{\partial \ln Z}{\partial f_l}\right|_{\vec{f}=\vec{0}}, \tag{16}$$

and the covariance $C_{ll'} \equiv \langle \sigma_l \sigma_{l'} \rangle - \langle \sigma_l \rangle \langle \sigma_{l'} \rangle$ reads

$$C_{ll'} = \left.\frac{\partial^2 \ln Z}{\partial f_{l'} \partial f_l}\right|_{\vec{f}=\vec{0}} = -\left.\frac{\partial m_l}{\partial f_{l'}}\right|_{\vec{f}=\vec{0}}. \tag{17}$$

To compute the partition function Z, we employ the Hubbard-Stratonovich transform, which involves an auxiliary field h:

$$\exp\Big(-\frac{\kappa}{2}x^2\Big) = \frac{1}{\sqrt{2\pi\kappa}} \int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2\kappa}h^2 - ihx\Big)\, dh. \tag{18}$$

Employing this result, we have

$$\begin{aligned}
Z &= \frac{1}{\sqrt{2\pi\kappa}}\int_{-\infty}^{\infty} dh \sum_{\{\sigma_l\}} \exp\Big(-\frac{1}{2\kappa}h^2 - ih\Big[\sum_l \sigma_l D_l - \tau^*\Big] - \sum_l \sigma_l f_l\Big) \\
&= \frac{1}{\sqrt{2\pi\kappa}}\int_{-\infty}^{\infty} dh \exp\Big(-\frac{1}{2\kappa}h^2 + ih\tau^*\Big)\prod_l \sum_{\sigma_l} \exp\Big(-ih\sigma_l D_l - \sigma_l f_l\Big) \\
&= \frac{2^L}{\sqrt{2\pi\kappa}}\int_{-\infty}^{\infty} dh \exp\Big(-\frac{1}{2\kappa}h^2 + ih\tau^*\Big)\prod_l \cosh\Big(ihD_l + f_l\Big) \\
&= \frac{2^L}{\sqrt{2\pi\kappa}}\int_{-\infty}^{\infty} dh \exp\left(-\frac{h^2}{2\kappa}\right) a(h), \qquad (19)
\end{aligned}$$

where we introduced

$$a(h) \equiv \exp\left[ih\tau^* + \sum_l \ln\cosh\Big(ihD_l + f_l\Big)\right]. \qquad (20)$$

The complex saddle point approximation [74] yields the following expansion when $\kappa \to 0$:

$$Z = 2^L a(0) + \kappa\, 2^{L-1} a^{(2)}(0) + \kappa^2\, 2^{L-3} a^{(4)}(0) + O(\kappa^3)\,, \qquad (21)$$

where $a^{(i)}$ denotes the i-th derivative of a with respect to h. Therefore,

$$\ln Z = \ln\left[2^L a(0)\right] + \frac{\kappa}{2}\frac{a^{(2)}(0)}{a(0)} + \frac{\kappa^2}{8}\left\{\frac{a^{(4)}(0)}{a(0)} - \left[\frac{a^{(2)}(0)}{a(0)}\right]^2\right\} + O(\kappa^3)\,. \qquad (22)$$

Computing the successive derivatives of a (see Eq 20) and employing the definitions in Eq 16 and 17, we obtain

$$m_k = \tau^*\kappa D_k\left[1 + \kappa\left(D_k^2 - \sum_l D_l^2\right)\right] + O(\kappa^3), \qquad (23)$$

$$C_{jk} = -\kappa D_j D_k\left[1 + \kappa\left(D_k^2 + D_j^2 - \sum_l D_l^2\right)\right](1-\delta_{jk}) + \left(1 - \tau^{*2}\kappa^2 D_k^2\right)\delta_{jk} + O(\kappa^3). \qquad (24)$$

$$(25)$$

We also obtain the elements of the inverse correlation matrix:

$$C^{-1}_{jk} = \kappa D_j D_k\,(1-\delta_{jk}) + \left[1 + \kappa^2 D_j^2\left(\sum_l D_l^2 - D_j^2 + \tau^{*2}\right)\right]\delta_{jk} + O(\kappa^3). \qquad (26)$$

To first order in κ, we recover the results of our previous work [31], where we employed the mean-field approximation to first order in the coupling strengths $\kappa D_k D_l$. Moreover, the ICOD matrix reads

$$\tilde{C}^{-1}_{jk} = \kappa D_j D_k\,(1-\delta_{jk}) + O(\kappa^3)\,, \qquad (27)$$

i.e. our previous expression [31] holds to second order in κ, generalizing our previous results.

S1.2 High-temperature Plefka expansion of the TAP free energy

We can derive the expression of the inverse covariance matrix in the large temperature limit (or equivalently here, in the limit of small coupling κ), using a *Plefka expansion*. We follow the calculation from [75], starting from the Hamiltonian in Eq 14 with no auxiliary field, and ignoring additive constant terms:

$$\begin{aligned}
H &= \frac{\kappa}{2}\sum_{i,j} D_i D_j \sigma_i\sigma_j - \kappa\sum_j \tau^*\sigma_j D_j \qquad (28) \\
&= \frac{\tilde{\kappa}}{2}\sum_{i,j} \tilde{D}_i \tilde{D}_j \sigma_i\sigma_j - \tilde{\kappa}\sum_j \tilde{\tau}\sigma_j \tilde{D}_j, \qquad (29)
\end{aligned}$$

where we defined rescaled parameters as

$$\tilde{\kappa} = \kappa \sum_j D_j^2, \tag{30}$$

$$\tilde{\tau} = \frac{\tau^*}{\sqrt{\sum_j D_j^2}}, \tag{31}$$

$$\tilde{D}_i = \frac{D_i}{\sqrt{\sum_j D_j^2}}. \tag{32}$$

For simplicity and to match notations from [75], we introduce $J_{ij} = -\tilde{D}_i \tilde{D}_j$, and $h_i = -\tilde{\tau}\tilde{D}_j$.

S1.2.1 TAP free energy expansion

We are interested in the perturbative expansion in $\tilde{\kappa}$ of the *TAP free energy*, which is the free energy $\Phi = \log Z$ constrained to fixed values of the magnetizations $m_i = \langle \sigma_i \rangle$ and variances $v_i = \langle (\sigma_i - m_i)^2 \rangle$, by using Lagrange multipliers. Focusing on binary spins, we have $v_i = 1 - m_i^2$. Ref. [75] suggested a full expression of the perturbative expansion, and proved that it is exact up to order 4. This expansion reads:

$$\Phi(\tilde{\kappa}) = \Phi(0) + \frac{\tilde{\kappa}}{2} \sum_{i,j} J_{ij} m_i m_j - \tilde{\kappa} \sum_i h_i m_i + \sum_{p=1}^{4} \frac{\tilde{\kappa}^p}{2p} \sum_{\substack{i_1,\dots,i_p \\ \text{pairwise distinct}}} J_{i_1 i_2} \dots J_{i_p i_1} \prod_{\alpha=1}^{p} v_{i_\alpha}, \tag{33}$$

where

$$\Phi(0) = -\sum_i \left[\frac{1-m_i}{2} \log\left(\frac{1-m_i}{2}\right) + \frac{1+m_i}{2} \log\left(\frac{1+m_i}{2}\right) \right]. \tag{34}$$

S1.2.2 Extremized magnetizations

The first step is then to extremize this free energy with respect to magnetizations:

$$\begin{aligned} \frac{\partial \Phi}{\partial m_i} &= \frac{1}{2} \log \frac{1+m_i}{1-m_i} + \tilde{\kappa} \sum_j J_{ij} m_j - \tilde{\kappa} h_i - \tilde{\kappa}^2 \sum_{\substack{j \\ j \neq i}} J_{ij}^2 (1-m_j^2) m_i \\ &- \tilde{\kappa}^3 \sum_{\substack{j \neq k \\ j,k \neq i}} J_{ij} J_{jk} J_{ki} (1-m_j^2)(1-m_k^2) m_i - \tilde{\kappa}^4 \sum_{\substack{j \neq k \neq l \\ j,k,l \neq i}} J_{ij} J_{jk} J_{kl} J_{li} (1-m_j^2)(1-m_k^2)(1-m_l^2) m_i = 0. \end{aligned} \tag{35}$$

We write the perturbative expansion $m_i = m_i^{(0)} + \tilde{\kappa} m_i^{(1)} + \tilde{\kappa}^2 m_i^{(2)} + \tilde{\kappa}^3 m_i^{(3)} + \tilde{\kappa}^4 m_i^{(4)}$, where we only keep leading terms in N (in the large N limit) in all $m_i^{(k)}$. Inserting this expression in 35, we obtain

$$m_i^{(0)} = 0 \tag{36a}$$

$$m_i^{(1)} = -h_i = \tilde{\tau} \tilde{D}_i \tag{36b}$$

$$m_i^{(2)} = \sum_j J_{ij} m_j^{(1)} = -\tilde{\tau} \tilde{D}_i \sum_j \tilde{D}_j^2 \tag{36c}$$

$$m_i^{(3)} = \sum_j J_{ij} m_j^{(2)} = \tilde{\tau} \tilde{D}_i \left(\sum_j \tilde{D}_j^2 \right)^2 \tag{36d}$$

$$m_i^{(4)} = \sum_j J_{ij} m_j^{(3)} = -\tilde{\tau} \tilde{D}_i \left(\sum_j \tilde{D}_j^2 \right)^3. \tag{36e}$$

Note that the dominant terms we retained are of order $1/\sqrt{N}$.

S1.2.3 Inverse covariance matrix

From the free energy, we can derive the elements of the inverse covariance matrix through $C_{ij}^{-1} = -\dfrac{\partial^2 \Phi}{\partial m_i \partial m_j}$. Differentiating Eq 35 yields

$$\begin{aligned} C_{ii}^{-1} =& \frac{1}{2}\left(\frac{1}{1+m_i} + \frac{1}{1-m_i}\right) + \tilde{\kappa}^2 \sum_{\substack{j \\ j\neq i}} J_{ij}^2 (1-m_j^2) + \tilde{\kappa}^3 \sum_{\substack{j\neq k \\ j,k\neq i}} J_{ij} J_{jk} J_{ki} (1-m_j^2)(1-m_k^2) \\ &+ \tilde{\kappa}^4 \sum_{\substack{j\neq k\neq l \\ j,k,l\neq i}} J_{ij} J_{jk} J_{kl} J_{li} (1-m_j^2)(1-m_k^2)(1-m_l^2) + O(\tilde{\kappa}^5), \end{aligned} \tag{37}$$

$$\begin{aligned} C_{ij}^{-1} =& -\tilde{\kappa} J_{ij} - 2\tilde{\kappa}^2 J_{ij}^2 m_i m_j - 4\tilde{\kappa}^3 \sum_{\substack{k \\ k\neq j,i}} J_{ij} J_{jk} J_{ki} (1-m_k^2) m_i m_j \\ &- 6\tilde{\kappa}^4 \sum_{\substack{k\neq l \\ k,l\neq i,j}} J_{ij} J_{jk} J_{kl} J_{li} (1-m_k^2)(1-m_l^2) m_i m_j + O(\tilde{\kappa}^5) \quad \text{if } i \neq j. \end{aligned} \tag{38}$$

Injecting Eq 36 finally yields, up to order 4 in $\tilde{\kappa}$:

$$\begin{aligned} C_{ii}^{-1} =& 1 + \tilde{\kappa}^2 \tilde{D}_i^2 \left(\tilde{\tau}^2 + \sum_{\substack{j \\ j\neq i}} \tilde{D}_j^2\right) - \tilde{\kappa}^3 \tilde{D}_i^2 \left(2\tilde{\tau}^2 \sum_j \tilde{D}_j^2 + \sum_{\substack{j\neq k \\ j,k\neq i}} \tilde{D}_j^2 \tilde{D}_k^2\right) \\ &+ \tilde{\kappa}^4 \tilde{D}_i^2 \left(\tilde{\tau}^4 \tilde{D}_i^2 + 3\tilde{\tau}^2 (\sum_k \tilde{D}_k^2)^2 + \sum_{\substack{j\neq k\neq l \\ j,k,l\neq i}} \tilde{D}_j^2 \tilde{D}_k^2 \tilde{D}_l^2\right) + O(\tilde{\kappa}^5), \end{aligned} \tag{39}$$

$$C_{ij}^{-1} = \tilde{\kappa} \tilde{D}_i \tilde{D}_j + O(\tilde{\kappa}^5) \quad \text{if } i \neq j. \tag{40}$$

Except the 1 in the first line, all other terms are of order $1/N$. These expressions are consistent with the second order expansion in κ of the inverse correlation matrix obtained in Eq 26. They further extend our previous results on the ICOD matrix [31], up to order 4 in $\tilde{\kappa}$.

S1.3 Comparison to synthetic data

In Fig S1, we compare our analytical approximations from Eq 39, 40 to the elements of the inverse covariance matrix obtained directly from synthetic data. We find a good agreement, which improves when the number M of synthetic sequences increases, and, in the case of the diagonal elements, with the order of the expansion in $\tilde{\kappa}$. Despite the fact that these expansions are made in the limit of small selection strength $\tilde{\kappa}$, when $\tilde{\kappa}$ is very small, the data is noisy, and agreement is thus less good that when $\tilde{\kappa}$ is somewhat larger. Interestingly, the analytical approximation Eq 40 for off-diagonal elements is very good even for intermediate and large values of $\tilde{\kappa}$, hinting that its validity extends beyond the regime of small $\tilde{\kappa}$ and beyond order 4 in $\tilde{\kappa}$. Meanwhile, for the diagonal terms, the expansion gives good results for relatively small $\tilde{\kappa}$, but agreement becomes less good as $\tilde{\kappa}$ becomes larger, in agreement with expectations for a small-$\tilde{\kappa}$ expansion.

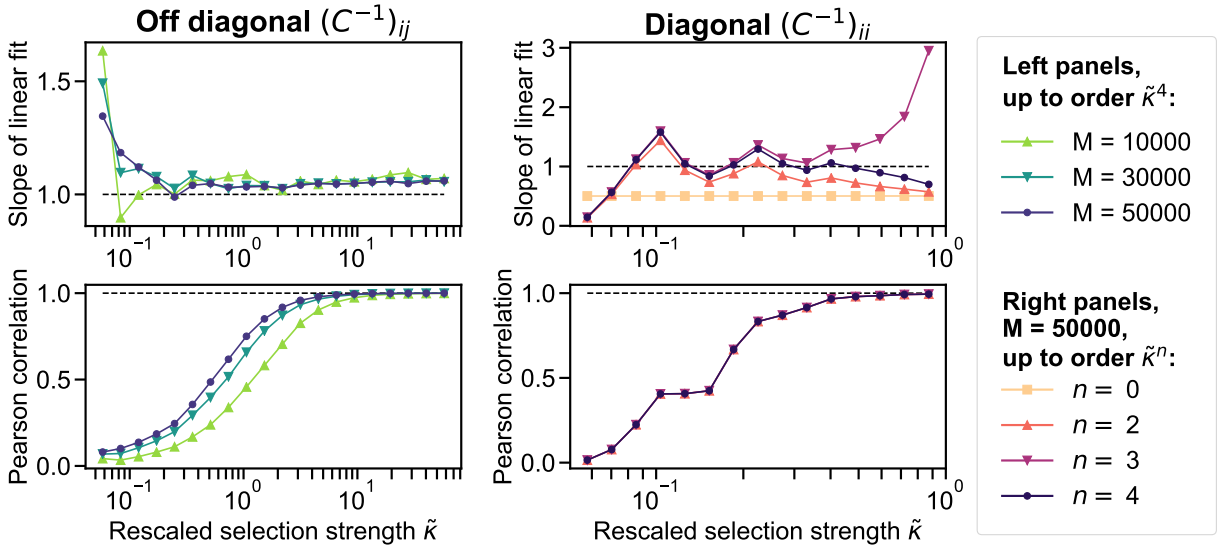


Figure S1: **Inverse covariance matrix: Comparison between analytical approximation and synthetic data.** Left (rep. right) panels show the slope of the linear fit and the Pearson correlation between off-diagonal (resp. diagonal) elements of the inverse covariance matrix given by analytical formulas (Eq 39, 40) and the inverse covariance matrix computed from synthetic data. For the latter, we generate independent sequences at equilibrium as in Fig 2 but the parameter κ is varied and τ^* is taken of order 1 (specifically, $\tau^* = 0.05\sqrt{\sum_i D_i^2} \approx 1$). In the left panels, we construct MSAs with depths $M = 10000, M = 30000, M = 50000$. In the right panels, we take $M = 50000$ sequences and compare the diagonal of the inverse covariance matrix computed from the data to the analytical formula 39 by gradually adding orders in $\tilde{\kappa}$. Note that we do not use a pseudocount ($a = 0$) here, which is acceptable given the large numbers of sequences, and allows a more direct comparison with the analytical formulas. The dashed lines represent the ideal values (slope 1 and Pearson correlation 1), which would indicate a perfect match between the data and the analytical approximation.

S2 Supplementary figures

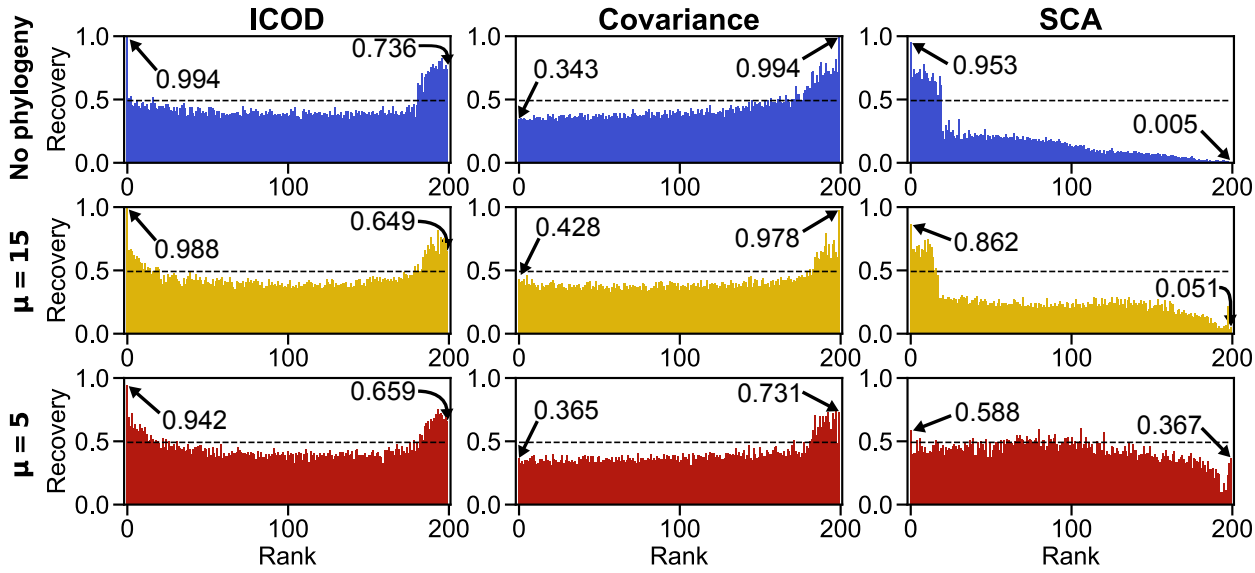


Figure S2: **Mutational effect recovery by each eigenvector.** The mutational effect recovery is shown for three methods: ICOD, Covariance and SCA, using each eigenvectors in the spectrum of the matrix involved in the method. The dashed line shows the chance expectation for the recovery (see Eq 13). Recovery by the eigenvectors associated to the largest and smallest eigenvalues are indicated for each method. Throughout, three data sets are considered, which differ by the amount of phylogeny (No phylogeny, $\mu = 15$, $\mu = 5$). The data is generated as in Fig 3.

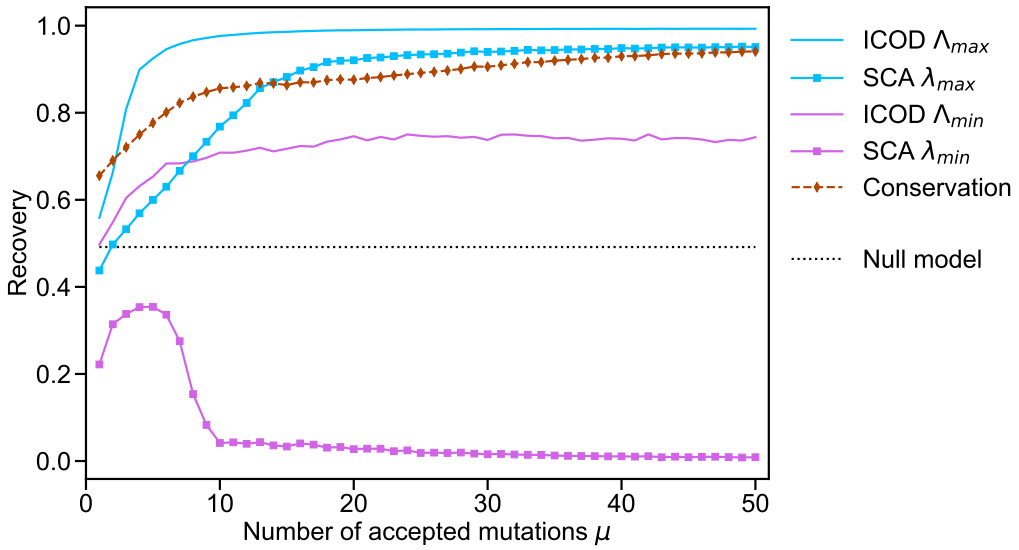


Figure S3: **Impact of phylogeny on mutational effect recovery for SCA.** Same as in Fig 4, but for SCA. For SCA, eigenvectors associated to the largest (λ_{max}) and smallest (λ_{min}) eigenvalues are considered. The results for ICOD and conservation (see Fig 4) are reproduced here for comparison purposes.

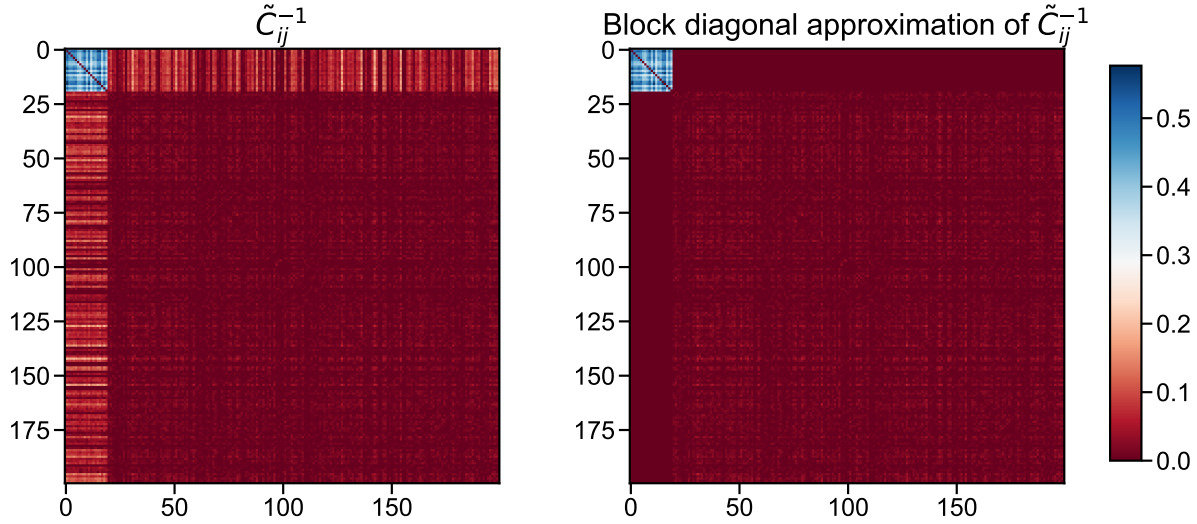


Figure S4: **ICOD matrix and its block diagonal approximation.** The left panel shows the ICOD matrix computed on data generated independently at equilibrium (14000 sequences). Parameters are the same as for the equilibrium ('No phylogeny') data set in Fig 3. Recall that $L_S = 20$ sector sites out of $L = 200$ total sites. The first 20×20 diagonal block (mainly blue) is associated to the sector, while the second one, of size 180×180, is associated to non-sector sites (i.e. sites with small mutational effects). The right panel shows the block diagonal approximation of the ICOD matrix shown in the left panel. Here, the matrix elements that do not belong to either of the two diagonal blocks are set to 0. Meanwhile, elements of these diagonal blocks are the same as in the ICOD matrix.

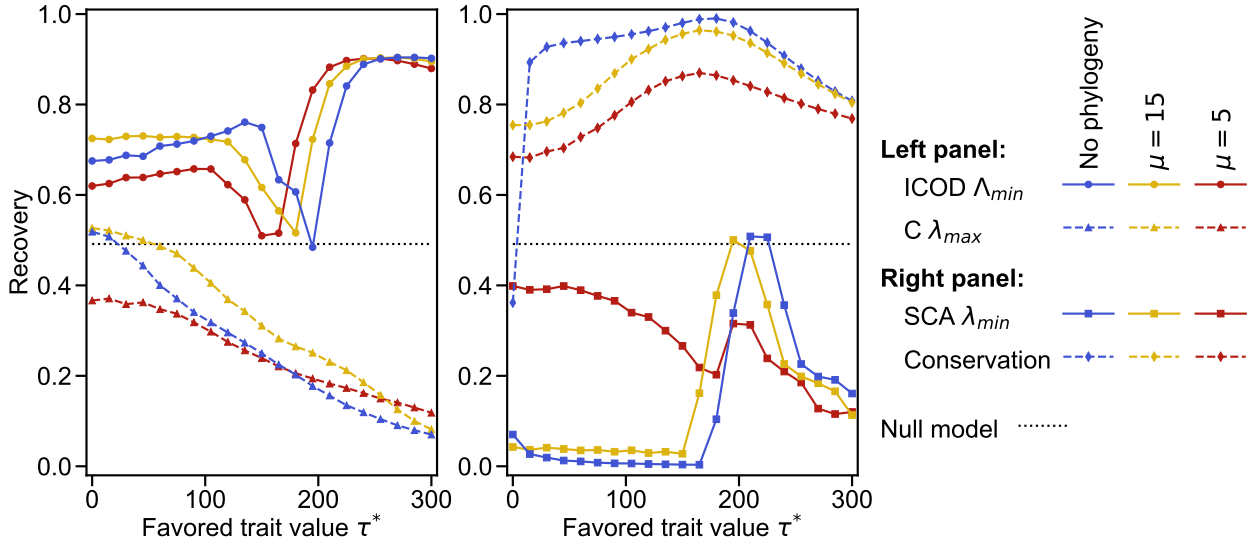


Figure S5: **Impact of the favored trait value on mutational effect recovery at the opposite end of the spectrum.** Same as in Fig 5, but focusing on eigenvectors at the opposite end of the spectrum. For ICOD (resp. SCA), eigenvectors associated to the smallest eigenvalue Λ_{min} and (resp. λ_{min}) are used, while for covariance, the eigenvector associated to the largest eigenvalue λ_{max} is used. Conservation results (see Fig 5) are reproduced here for comparison purposes.

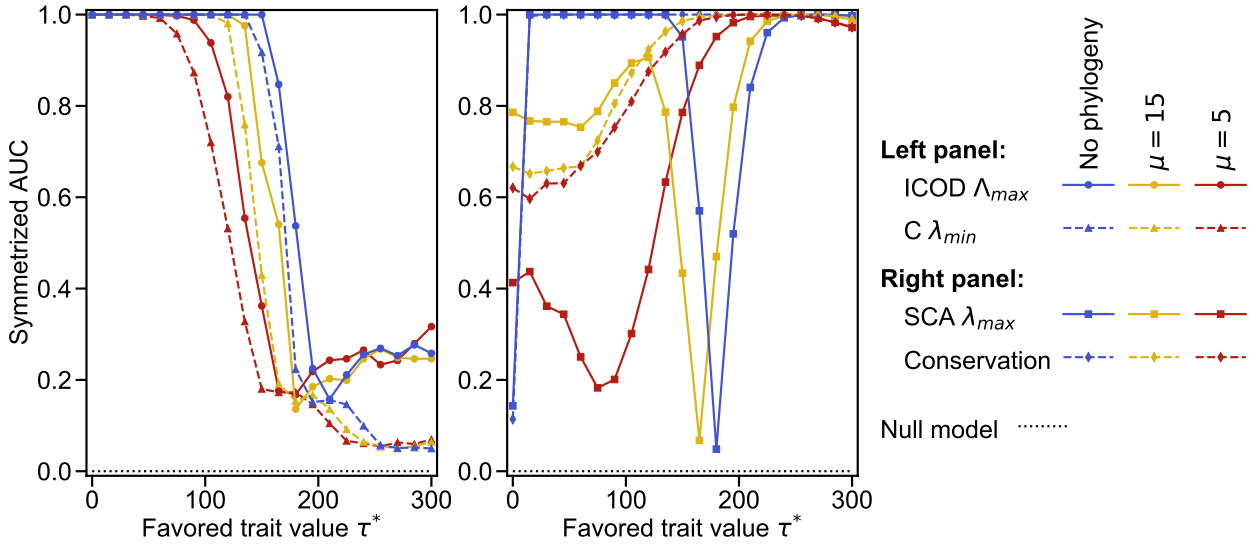


Figure S6: **Impact of the favored trait value on symmetrized AUC.** Same as in Fig 5, but instead of using recovery, we consider the symmetrized AUC between the eigenvectors and the mutational effects.

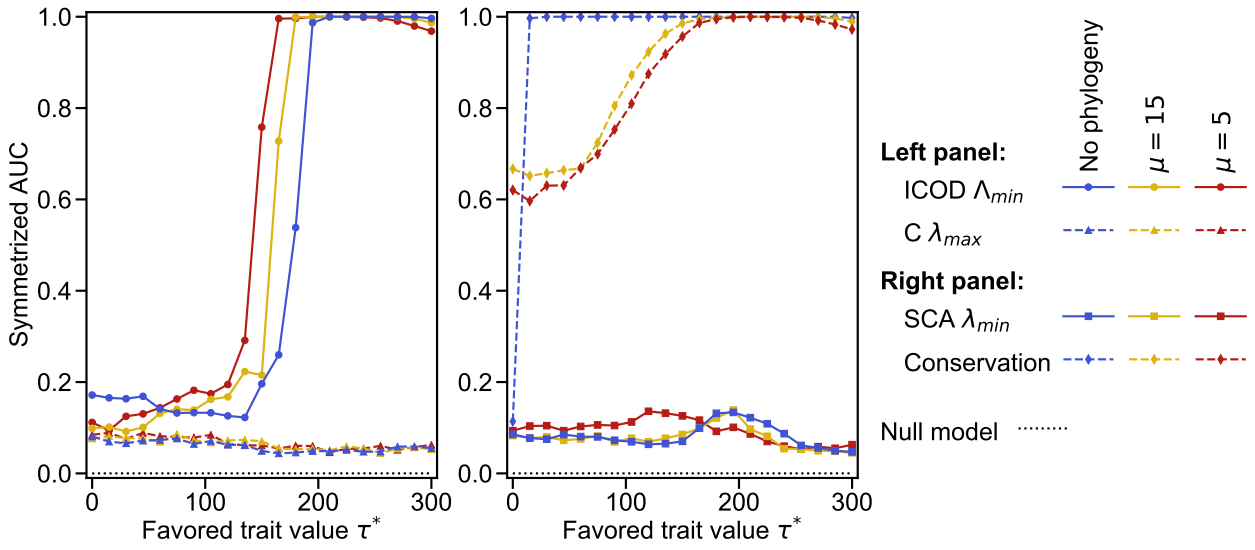


Figure S7: **Impact of the favored trait value on symmetrized AUC at the opposite end of the spectrum.** Same as in Fig S5, but instead of using recovery, we consider the symmetrized AUC between the eigenvectors and the mutational effects.

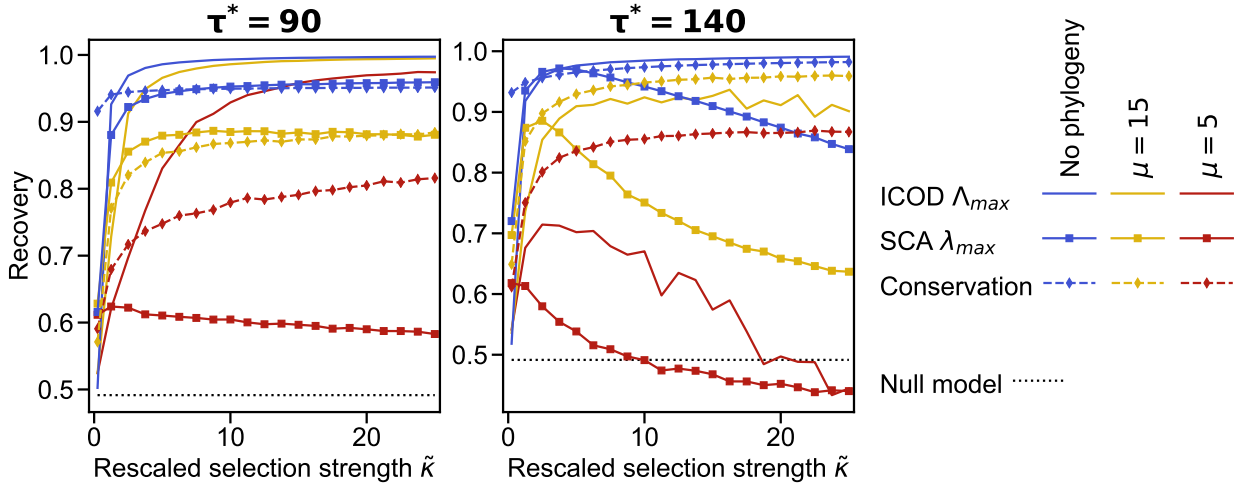


Figure S8: **Impact of selection strength on mutational effect recovery for SCA and conservation.** Same as in Fig 6 but using SCA and conservation. For SCA, the eigenvector associated to the largest eigenvalue (λ_{max}) is used. ICOD results (see Fig 6) are reproduced here for comparison purposes.

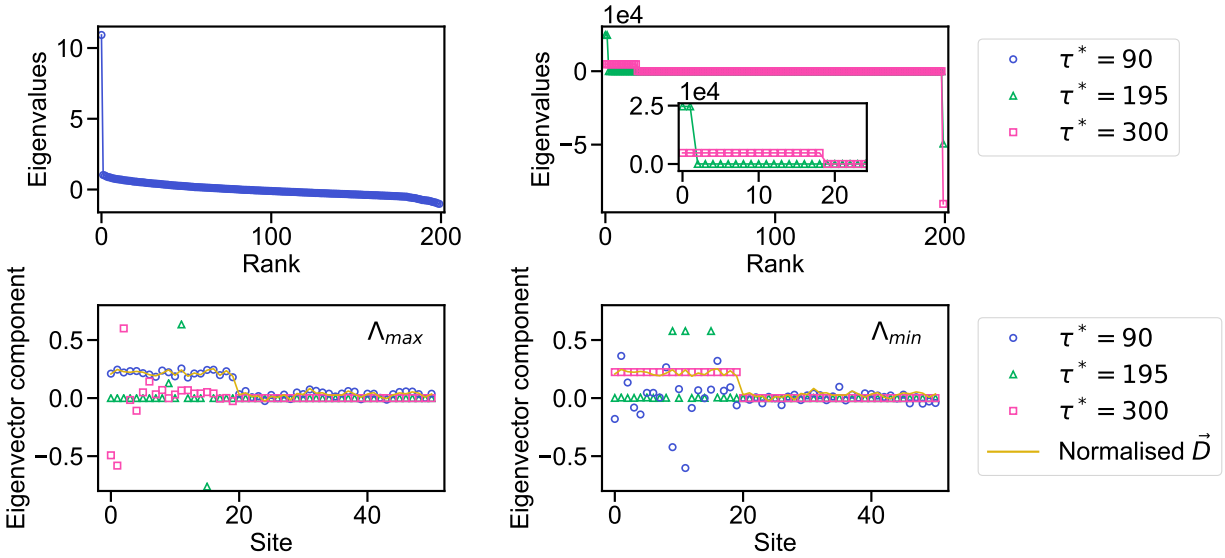


Figure S9: **Impact of large favored trait values on ICOD spectrum and eigenvectors.** Top panels show the eigenvalues of the ICOD matrix at different values of favored trait value τ^* (left: $\tau^* = 90$; right: $\tau^* = 195$ and $\tau^* = 300$). Inset in the upper right panel: zoom on the 20 first eigenvalues. Bottom panels show the normalised eigenvectors associated to the largest eigenvalue (Λ_{max}) on the left and smallest eigenvalue (Λ_{min}) on the right panel, for the same three values of τ^* as in the top panels. In the bottom panels, we only show the first 50 sites of the eigenvectors, which comprise the 20 sites associated to the sector and 30 non-sector sites. The vector of mutational effects $\vec{D}$ is normalised and shown for comparison with the eigenvectors. Data is generated as in the equilibrium (no phylogeny) case in Fig 3, but using different values of τ^*.

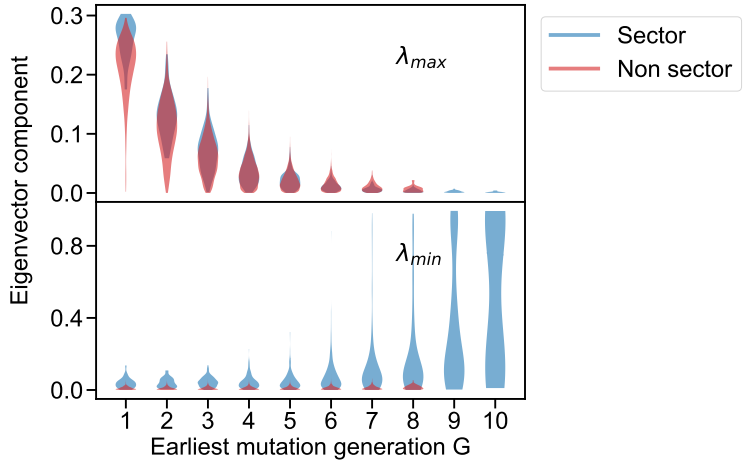


Figure S10: **Impact of earliest mutation generation G on covariance eigenvector components.** Same as in Fig 7, restricting to the data set with $\mu = 5$ and to the covariance method. Top panel: eigenvector associated to the largest eigenvalue λ_{max} of the covariance matrix. Bottom panel: results for the eigenvector associated to the smallest eigenvalue λ_{min} (see Fig 7) are reproduced here for comparison purposes.

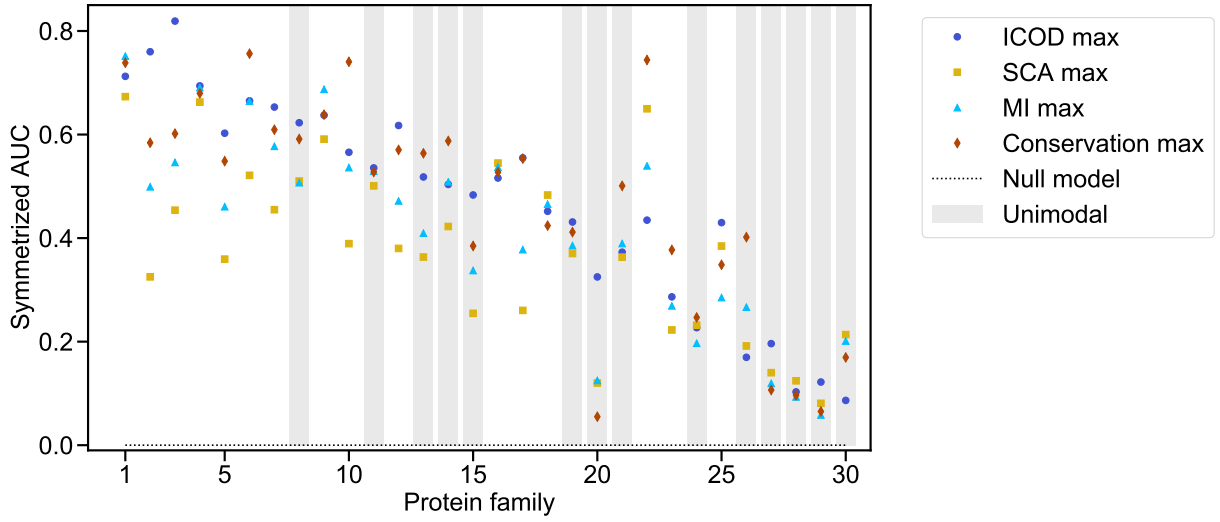


Figure S11: **Identifying functionally important sites in natural protein families: Maximum performance over phylogenetic cutoffs.** Same as in Fig 8, but focusing on the MSA phylogenetic cutoffs that maximize the symmetrized AUC. The mapping between protein family number and name is given in Table S1.

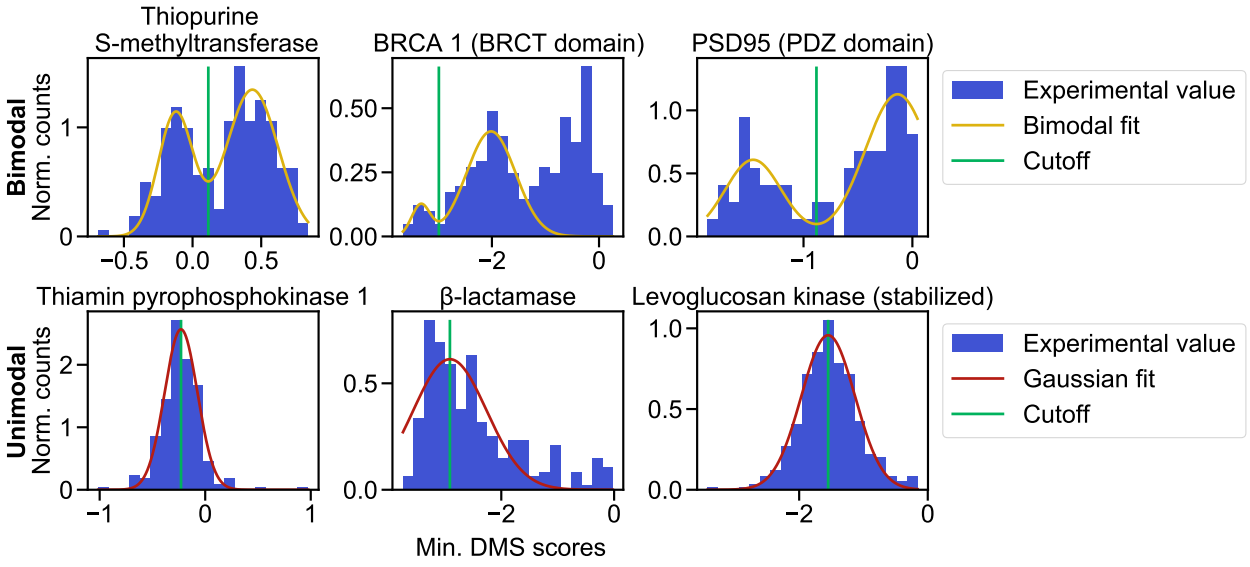


Figure S12: **Distribution of DMS scores for 6 example protein families.** Histograms of the minimum DMS scores (see Methods) are shown in six example cases. Top panels: three examples of scores with bimodal shape (or more complex shape - middle panel, see Methods). Bottom panels: three examples of scores with unimodal shape. In each case, bimodal Gaussian (in top panels) or Gaussian fits (in bottom panels) are shown together with their respective cutoffs, which correspond either to the location of the minimum value between peaks for bimodal fits, or to the mean for Gaussian fits (see Methods).

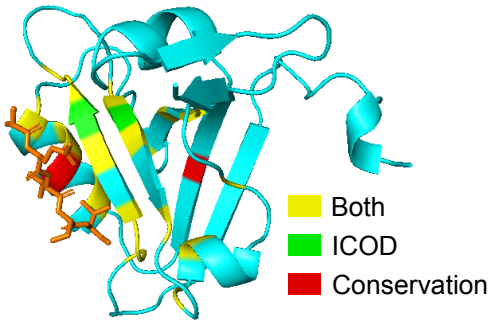
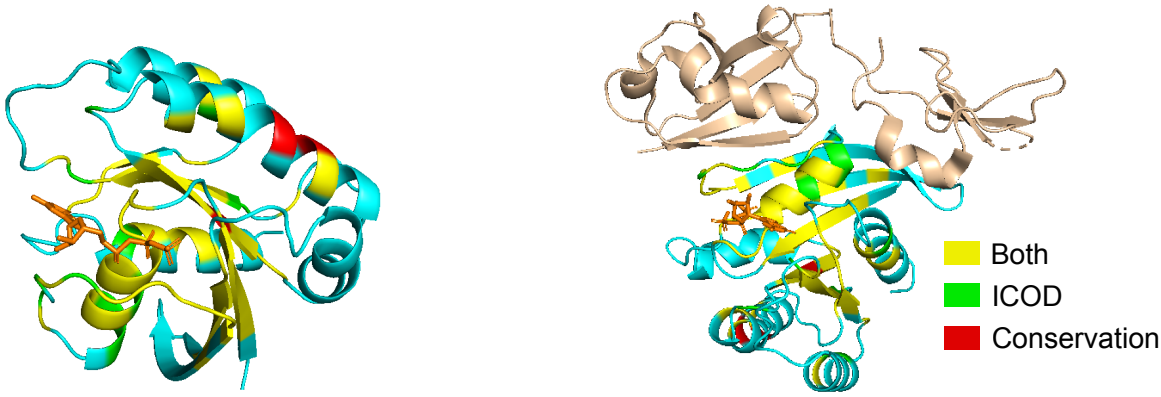


Figure S13: **Sector of the PDZ domain inferred by ICOD and by conservation.** The structure of the PDZ domain is represented in cyan and the CRIPT molecule is shown in orange. True positive sites of the sector found both by ICOD and by conservation are colored in yellow, while TP sites found by ICOD only are shown in green and TP sites found by conservation only are shown in red. The PDB identifier of this structure is 1BE9.



(a) Ras and GNP structure

(b) GNP-Ras-Raf complex

Figure S14: **Sector of the Ras protein inferred by ICOD and by conservation.** In both panels, the structure of Ras is shown in cyan. True positive sites of the sector found both by ICOD and by conservation are colored in yellow, while TP sites found by ICOD only are shown in green and TP sites found by conservation only are shown in red. (a) Structure of Ras in interaction with the GNP molecule (colored in orange). PDB identifier: 5P21. (b) Structure of Ras in complex with the RBD and CRD domains of Raf (colored in wheat). The GNP molecule is again colored in orange. PDB identifier: 6XI7.

S3 Supplementary table

			Symmetrized AUC				True positives				
Name	Label	DMS	ICOD	Cons.	SCA	MI	Both	ICOD	Cons.	L_S	L
GAL4 (DNA-binding domain)	1	2	0.76	**0.77**	0.57	0.37	29	1	2	34	43
Translation initiation factor IF1	2	2	**0.76**	0.56	0.31	0.49	12	5	2	23	69
PABP singles (RRM domain)	3	2	**0.73**	0.62	0.33	0.69	15	3	1	25	71
Kanamycin kinase APH(3')-II	4	2	**0.70**	0.68	0.62	0.67	119	12	9	158	244
HRas	5	2	**0.68**	0.54	0.30	0.26	55	11	3	83	158
HSP90 (ATPase domain)	6	2	0.66	**0.76**	0.41	0.69	67	4	7	94	216
PSD95 (PDZ domain)	7	2	0.60	**0.63**	0.50	0.60	18	2	2	28	77
β-lactamase	8	1	**0.60**	0.59	0.42	0.45	63	12	8	107	256
BRCA 1 (RING domain)	9	2	0.57	**0.60**	0.43	0.59	13	1	1	23	70
HIV env protein (BF520)	10	2	0.55	**0.73**	0.20	0.48	242	7	27	295	356
Influenza polymerase PA subunit	11	1	**0.53**	0.46	0.10	0.36	209	35	22	346	683
DNA methylase HaeIII	12	2	**0.53**	0.48	0.30	0.31	129	14	4	179	306
Aliphatic amide hydrolase	13	1	0.51	**0.53**	0.13	0.29	100	17	15	163	316
HIV env protein (BG505)	14	1	0.50	**0.57**	0.21	0.38	95	18	16	167	363
Small ubiquitin-related modifier 1	15	1	**0.48**	0.38	0.22	0.34	18	5	3	35	82
UBE4B (U-box domain)	16	2	**0.45**	0.43	0.25	0.22	34	6	5	52	76
β-glucosidase	17	2	0.44	**0.48**	0.15	0.33	219	24	29	314	441
BRCA 1 (BRCT domain)	18	2	0.42	0.41	0.35	**0.44**	1	2	1	11	208
Influenza hemagglutinin	19	1	**0.39**	**0.39**	0.23	0.29	123	18	21	241	544
Hepatitis C NS5A	20	1	**0.33**	0.05	0.09	0.14	71	4	4	80	85
SUMO-conjugating enzyme UBC9	21	1	0.32	**0.47**	0.37	0.41	34	3	5	60	136
Ubiquitin	22	2	0.31	**0.63**	0.22	0.46	28	4	8	45	70
Thiopurine S-methyltransferase	23	2	0.25	**0.37**	0.28	0.28	44	3	6	84	210
Levoglucosan kinase (stabilized)	24	1	0.22	0.24	**0.25**	0.14	91	19	15	179	361
YAP1 (WW domain)	25	2	0.20	**0.28**	0.26	0.19	10	1	1	17	30
PTEN	26	1	0.15	**0.38**	0.18	0.27	62	7	18	123	255
Mitogen-activated protein kinase 1	27	1	**0.10**	0.09	0.01	**0.10**	55	3	4	109	214
Levoglucosan kinase	28	1	**0.09**	**0.09**	**0.09**	0.03	83	18	19	187	361
Thiamin pyrophosphokinase 1	29	1	**0.06**	0.02	0.02	0.03	53	8	6	115	222
Calmodulin-1	30	1	0.05	**0.14**	0.00	0.12	25	7	9	64	137

Table S1: **Identifying functionally important sites in natural protein families: detailed results.** The first two columns present the name of natural protein families that are considered and the corresponding label used in Fig 8 and S11. Next, the 'DMS' column indicates the shape of the DMS data, 1 for unimodal and 2 for bimodal. The symmetrized AUC columns correspond to the results shown in Fig 8 – conservation is abbreviated by 'Cons.'. The best score among the methods is highlighted in bold for each family. Finally, the number of true positive sites found both by ICOD and by conservation ('Both'), by ICOD only ('ICOD') and by conservation only ('Cons.') are given. The last columns provide the size (or length) L_S of the sector and the protein length L.