Revealing evolutionary constraints on proteins through sequence analysis

Shou-Wen Wang,[1,2,3,*] Anne-Florence Bitbol,[4,†] and Ned S. Wingreen[3,5,‡]

[1] *Beijing Computational Science Research Center, Beijing, 100094, China*
[2] *Department of Engineering Physics, Tsinghua University, Beijing, 100086, China*
[3] *Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*
[4] *Sorbonne Université, CNRS, Laboratoire Jean Perrin (UMR 8237), F-75005 Paris, France*
[5] *Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA*

Alignments of large numbers of protein sequences have revealed "sectors" of collectively coevolving amino acids in several protein families. Here, we show that selection acting on any relevant physical property of a protein, e.g. the elastic energy of an important conformational change, can give rise to such a sector. We demonstrate that the main signature of these physical sectors lies in the small-eigenvalue modes of the covariance matrix of the selected sequences. This simple, generic model leads us to propose a principled method to identify sectors, along with the magnitudes of mutational effects, from sequence data.

Introduction.—Proteins play crucial roles in all cellular processes, acting as enzymes, motors, receptors, regulators, and more. The function of a protein is encoded in its amino-acid sequence, and the recent explosion of available sequences has inspired data-driven approaches to discover the principles of protein operation. At the root of these new approaches is the observation that amino-acid residues which possess related functional roles often evolve in a correlated way. In particular, analyses of large alignments of protein sequences have identified "sectors" of collectively correlated amino acids [1–3, 18, 20], which has enabled successful design of new functional sequences [3]. Sectors are spatially contiguous in the protein structure, and in the case of multiple sectors, each one may be associated with a distinct role [13, 18]. What is the physical origin of these sectors, and can we identify them from sequence data in a principled way?

To address these questions, we developed a general physical model that gives rise to sectors. Specifically, we start from an elastic-network model that quantifies the energetic cost of protein deformations [1], and show that selection acting on such elastic properties yields collective correlation modes in sequence data. Next, we generalize our model to any selected protein property to which mutations contribute additively. We show that the main signature of this selection process lies in the small-eigenvalue modes of the covariance matrix of the selected sequences. Finally, we demonstrate a principled method to identify sectors, and to quantify mutational effects, from sequence data alone.

Elastic-network model.— Functional deformation modes are under selection in proteins [8–10], and dynamical domains possess a signature in sequence data [11]. Elastic-network models have elucidated various protein properties [1, 2, 13, 14], including the emergence of allostery [15–17]. Thus motivated, we begin by building an elastic-network model [1, 2] for a PDZ protein domain [Fig. 1(a,b)] [18, 19], and computing the relationship between its "sequence" and the energetic cost of a functionally-relevant conformational change. Within our elastic model [20], the energetic cost of a small deformation from the equilibrium structure is

$$E = \frac{1}{2}\sum_{i,j}\left(\boldsymbol{r}_i - \boldsymbol{r}_i^0\right) M_{ij} \left(\boldsymbol{r}_j - \boldsymbol{r}_j^0\right) = \frac{1}{2}\,\delta\boldsymbol{r}^T M \delta\boldsymbol{r}, \quad (1)$$

where $\boldsymbol{r}_i$ is the position of the ith carbon atom, $\boldsymbol{r}_i^0$ is its equilibrium position, and the Hessian matrix M contains the second derivatives of the elastic energy with respect to atomic coordinates. Here, we take $\delta\boldsymbol{r}$ to be the conformational change from a ligand-free state (1BFE) to a ligand-bound state (1BE9) of the same PDZ domain [Fig. 1(a)]. This conformational change is central to PDZ function, so its energetic cost has presumably been under selection during evolution. Any other coherent conformational change would also be suitable for our analysis. To mimic simply the effect of mutation and selection, we introduce "mutations" of residues that weaken the spring constants involving their beta carbons by a small fraction ϵ. We represent mutations using a sequence $\vec{S}$ with $S_l \in \{0, 1\}$, where l is the residue index: $S_l = 0$ denotes the reference state, while $S_l = 1$ implies a mutation [Fig. 1(c)]. The sequence $\vec{S}$ and the spring network fully determine the Hessian matrix M, and thus the energy cost E of a conformational change [Eq. (1)].

Additivity of mutational effects.— Focusing on mutations that weakly perturb the elastic properties of a protein, we perform first-order perturbation analysis: $M = M^{(0)} + \epsilon M^{(1)} + o(\epsilon)$. Using Eq. (1) yields $E = E^{(0)} + \epsilon E^{(1)} + o(\epsilon)$, with $E^{(1)} = \delta\boldsymbol{r}^T M^{(1)} \delta\boldsymbol{r}/2$. Both $M^{(1)}$ and $E^{(1)}$ can be expressed as sums of contributions from individual mutations. We define Δ_l as the first-order energy cost $\epsilon E^{(1)}$ of a single mutation at site l of the sequence. To leading order, the effect of mutations

* Present address: Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA; these two authors contributed equally.
† Corresponding author: anne-florence.bitbol@sorbonne-universite.fr; these two authors contributed equally.
‡ Corresponding author: wingreen@princeton.edu

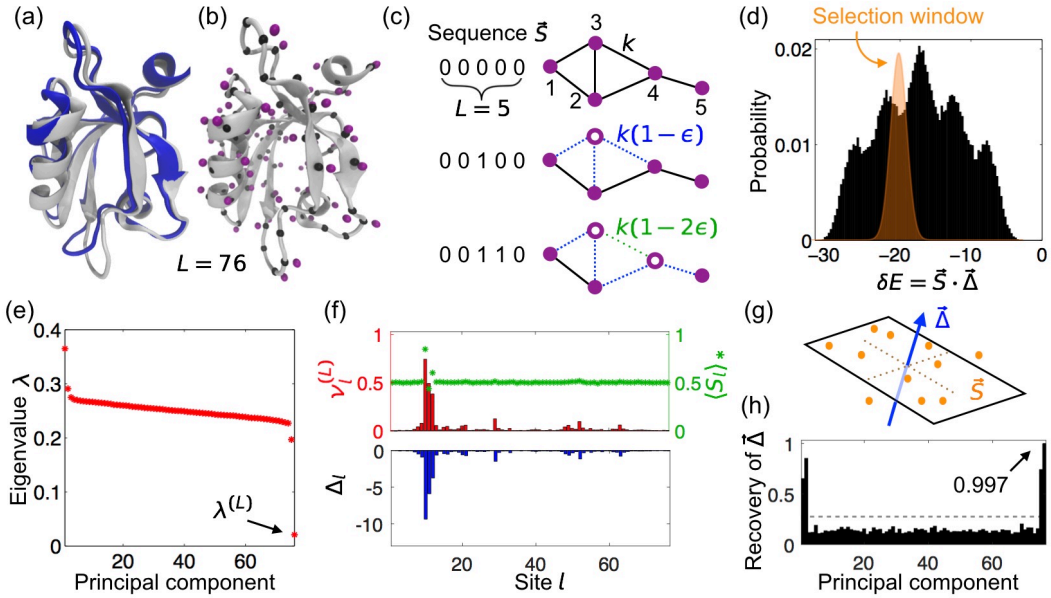

FIG. 1. Selection applied to an elastic protein model leads to a statistical signature among sequences. (a) Cartoon representation of the PDZ domain from the RSCB PDB [21], (gray: ligand free, 1BFE; blue: ligand bound, 1BE9 (ligand not shown)). (b) Elastic network model for 1BFE, where each amino-acid residue is represented by its alpha carbon (Cα, black node) and beta carbon (Cβ, purple node). Nearby nodes interact through a harmonic spring [2, 20]. (c) Relation between protein sequence $\vec{S}$ and elastic network: 0 denotes the reference state, while 1 denotes a mutated residue, which weakens interactions of the corresponding Cβ with all its neighbors by ϵ. (d) Histogram of the energy δE required to deform the domain from its ligand-free to its ligand-bound conformation, for randomly sampled sequences where 0 and 1 are equally likely at each site. Sequences are selectively weighted using a narrow Gaussian window (orange) around δE^*. (e) Eigenvalues of the covariance matrix C for the selectively weighted protein sequences. (f) Upper panel: last principal component $\nu_l^{(L)}$ of C (red) and average mutant fraction $\langle S_l \rangle_*$ (green) at site l after selection; lower panel: effect Δ_l of a single mutation at site l on δE. (g) Schematic representation of the selected ensemble in sequence space, where each dot is a highly-weighted sequence; thus dots are restricted to a narrow region around a plane perpendicular to $\vec{\Delta}$. (h) Recovery of $\vec{\Delta}$ for all principal components $\vec{\nu}^{(j)}$, with maximum Recovery=1 [Eq. (5)]. Gray dashed line: random expectation of Recovery [20].

on the energy cost of a deformation reads

$$\delta E = E - E^{(0)} = \sum_l S_l \Delta_l. \tag{2}$$

This equation, which relates sequence $\vec{S}$ to function δE, is quite general: 1) it constitutes the leading-order contribution of mutational effects, not limited to our elastic-network model; 2) system-specific details are encoded by the single-site mutational effects Δ_l, which in principle can be measured experimentally; 3) "energy" can refer to any physical property of the protein, say the binding affinity, catalytic activity or thermal stability [22], not just the energy cost of a conformational change.

The general sequence-function relation in Eq. (2) suggests a physical definition of a protein sector as the set of sites with dominant mutational effects on some selected property of the protein. For example, the vector $\vec{\Delta}$ of mutational effects on the energy cost of the change from the ligand-unbound to the ligand-bound conformation in our elastic-network model of PDZ is shown in Fig. 1(f). The magnitudes of mutational effects are strongly heterogeneous (see also Fig. S1 [20]). Here, the amino acids with largest effects are those that move significantly upon ligand binding.

How is such a functionally-defined sector reflected in the statistical properties of the sequences that survive evolution? To answer this question, we introduce a selection procedure and analyze the selected sequences.

Selection model and PCA of selected sequences.— For our elastic model of the PDZ domain, the distribution of δEs for random sequences is shown in Fig. 1(d). As an example of a physical selection process, we assume that sequences with a narrower distribution of δEs are selected by evolution, corresponding e.g. to a preferred ligand-binding affinity. Specifically, we weight each sequence $\vec{S}$ according to the distribution $P(\vec{S}) = \exp(w(\vec{S}))/\sum_{\vec{S}} \exp(w(\vec{S}))$, where $w(\vec{S})$ is the fitness of sequence $\vec{S}$:

$$w(\vec{S}) = -\frac{\kappa}{2}\Big(\sum_l \Delta_l S_l - \delta E^*\Big)^2. \tag{3}$$

Here, the selection strength κ sets the width of the selection window, and δE^* is its center. For all selections, we take $\kappa = 10/(\sum_l \Delta_l^2)$, so that the width of the selection window scales with that of the unselected distribution. Importantly, although mutations have additive effects on

δE, selection gives rise to correlations among sites. For instance, if $\delta E^* = 0$ and if $\Delta_l < 0$ for all l, as in Fig. 1, a mutation at a site with large $|\Delta_l|$ will decrease the likelihood of additional mutations at all other sites with large $|\Delta_l|$.

Previous approaches to identifying sectors from real protein sequences have relied on modified forms of Principal Component Analysis (PCA). So we begin by asking: can PCA identify sectors in our physical model? PCA corresponds to diagonalizing the covariance matrix C of sequences: it identifies the principal components (eigenvectors) $\vec{\nu}^{(j)}$ associated with progressively smaller variances (eigenvalues) $\lambda^{(j)}$. We introduce $\langle\cdot\rangle_*$ to denote ensemble averages over the selectively weighted sequences, reserving $\langle\cdot\rangle$ for averages over the unselected ensemble. The mutant fraction at site l in the selected ensemble is $\langle S_l\rangle_* = \sum_{\vec{S}} S_l P(\vec{S})$, and the covariance matrix C reads

$$C_{ll'} = \Big\langle (S_l - \langle S_l\rangle_*)\cdot(S_{l'} - \langle S_{l'}\rangle_*)\Big\rangle_* . \qquad (4)$$

To test the ability of PCA to identify a physical sector, we employed the selection window shown in orange in Fig. 1(d). The resulting eigenvalues are shown in Fig. 1(e). One sees outliers. In particular, why is the last eigenvalue so low? Due to the narrow selection window, the highly-weighted sequences satisfy $\sum_l S_l\Delta_l = \vec{S}\cdot\vec{\Delta} \approx \delta E^*$. This means that in the L-dimensional sequence space, the data points for the highly-weighted sequences lie in a narrow region around a plane perpendicular to $\vec{\Delta}$ [Fig. 1(g)]. Hence, the data has exceptionally small variance in this direction, leading to a particularly small eigenvalue of C. Moreover, the corresponding last principal component $\vec{\nu}^{(L)}$ points in the direction with the smallest variance and is consequently parallel to $\vec{\Delta}$ [Fig. 1(f)]. Formally, in Eq. (3), $\vec{\Delta}$ plays the part of a repulsive pattern in a generalized Hopfield model [6, 7], penalizing sequences aligned with $\vec{\Delta}$, or in our case those that fail to have the selected projection onto $\vec{\Delta}$.

In this example, the last principal component accurately recovers the physical sector corresponding to the largest elements of the mutational-effect vector $\vec{\Delta}$. More generally, to quantify the recovery of $\vec{\Delta}$ by a given vector $\vec{\nu}$, we introduce

$$\text{Recovery} = \frac{\sum_l |\nu_l\Delta_l|}{\sqrt{\sum_l \nu_l^2}\sqrt{\sum_l \Delta_l^2}}, \qquad (5)$$

which is nonnegative, has a random expectation of $(\sqrt{2/\pi L})\sum_l|\Delta_l|/\sqrt{\sum_l\Delta_l^2}$ for $L \gg 1$ [20], and saturates at 1 (including the case of parallel vectors). For our test case, Fig. 1(h) shows Recovery for all principal components. The last one features the highest Recovery, almost 1, confirming that it carries substantial information about $\vec{\Delta}$. The second-to-last principal component and the first two ones also provide a Recovery substantially above the random expectation.

In our model, $\vec{\Delta}$ is fundamentally a direction of *small variance*. So why do the first principal components also carry information about $\vec{\Delta}$? Qualitatively, when variance is decreased in one direction due to a repulsive pattern $\vec{\Delta}$, variance tends to increase in orthogonal directions involving the same sites. To illustrate this effect, let $L = 3$ and $\vec{\Delta} = (-1, 1, 0)$, and consider the sequences $\vec{S}$ satisfying $\vec{\Delta}\cdot\vec{S} = 0$ [namely $(0,0,0)$; $(1,1,0)$; $(0,0,1)$; $(1,1,1)$]. The last principal component is $\vec{\Delta}$, with zero variance, and the first principal component is $(1,1,0)$: Recovery is 1 for both of them. This selection conserves the trace of the covariance matrix (i.e. the total variance), decreasing variance along $\vec{\Delta}$, and therefore increasing it along $(1,1,0)$. This toy model provides an intuitive understanding of why the large-eigenvalue modes of the covariance matrix also carry information about $\vec{\Delta}$.

ICOD method.— An important concern is whether the last principal component is robust to small and/or noisy datasets. Indeed, other directions of small variance can appear in the data. As a second example, we applied a different selection window, centered in the tail of the distribution of δEs from our elastic model of PDZ [Fig. 2(a), inset]. This biased selection generates strong conservation, $\langle S_l\rangle_* \approx 1$, for some sites with significant mutational effects. Extreme conservation at one site now dictates the last principal component, and disrupts PCA-based recovery of $\vec{\Delta}$ [Fig. 2(a,b)].

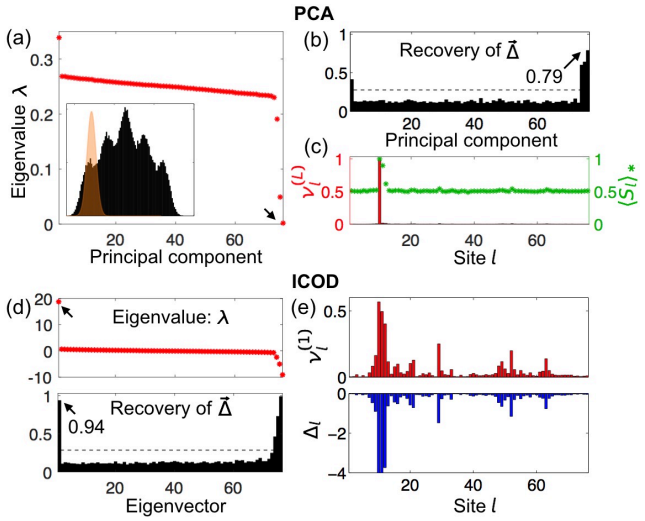

FIG. 2. Recovery of mutational-effect vector $\vec{\Delta}$ from sequence analysis in the case of strongly biased selection. (a-c) Principal Component Analysis (PCA) performs poorly due to strong conservation at some sites of large mutational effect. (a) Eigenvalues of covariance matrix obtained for strongly biased selection around $\delta E^*_{\text{biased}}$ (inset, orange window) for same model proteins as in Fig. 1. (b) Recovery of $\vec{\Delta}$ for all principal components. (c) Last principal component $\nu_l^{(L)}$ (red) and average mutant fraction $\langle S_l\rangle_*$ (green) at site l. (d-e) The ICOD method performs robustly. (d) Eigenvalues of $\tilde{C}^{-1}_{ll'}$ [Eq. (7)] (upper) and Recovery of $\vec{\Delta}$ for all eigenvectors (lower). (e) Leading eigenvector $\nu_l^{(1)}$ (upper) and mutational effect Δ_l at site l [lower, same as in Fig. 1(f)]. Gray dashed lines in (b,d): random expectation of Recovery [20].

To overcome this difficulty, we developed a more robust approach that relies on inverting the covariance matrix. Previously, the inverse covariance matrix was successfully employed in Direct Coupling Analysis (DCA) to identify strongly coupled residues that are in contact within a folded protein [3, 4, 15]. The fitness in our model [Eq. (3)] involves one and two-body interaction terms, and constitutes a particular case of the DCA Hamiltonian [20]. A small-coupling approximation [3–5, 17, 20] gives

$$C^{-1}_{ll'} \approx (1-\delta_{ll'})\ \kappa\Delta_l\Delta_{l'} + \delta_{ll'}\left(\frac{1}{P_l}+\frac{1}{1-P_l}\right), \qquad (6)$$

where P_l denotes the probability that site l is mutated. Since we are interested in extracting $\vec{\Delta}$, we can simply set to zero the diagonal elements of C^{-1}, which are dominated by conservation effects, to obtain a new matrix

$$\tilde{C}^{-1}_{ll'} \approx (1-\delta_{ll'})\kappa\Delta_l\Delta_{l'}. \qquad (7)$$

The first eigenvector of $\tilde{C}^{-1}$ (associated with its largest eigenvalue) should accurately report $\vec{\Delta}$ since, except for its zero diagonal, $\tilde{C}^{-1}$ is proportional to the outer product $\vec{\Delta}\otimes\vec{\Delta}$. We call this approach the *Inverse Covariance Off-Diagonal* (ICOD) method. As shown in Fig. 2(d-e), ICOD overcomes the difficulty experienced by PCA for biased selection, while performing equally well as PCA for unbiased selection (Fig. S2 [20]). Removing the diagonal elements of C^{-1} before diagonalizing is crucial: otherwise, the first eigenvector of C^{-1} is the same as the last eigenvector of C and suffers from the same shortcomings for strong conservation. Here too, both ends of the spectrum contain information about $\vec{\Delta}$ [Figs. 2(b,d)].

Multiple selection.— An important challenge in sector analysis is distinguishing multiple, independently evolving sectors [13, 18, 19]. We can readily generalize our fitness function [Eq. (3)] to allow for multiple selections:

$$w(\vec{S}) = -\sum_i \frac{\kappa_i}{2}\Big(\sum_l \Delta_{i,l}S_l - \delta E^*_i\Big)^2, \qquad (8)$$

where κ_i is the strength of the i-th selection, $\vec{\Delta}_i$ is the corresponding vector of mutational effects, and δE^*_i is the selection bias. For example, $\vec{\Delta}_1$ might measure how mutations change a protein's binding affinity, while $\vec{\Delta}_2$ might be related to its thermal stability, etc.

In Fig. S4 [20], we consider two physical sectors, i.e. a double selection, using artificially-generated mutational-effect vectors $\vec{\Delta}_1$ and $\vec{\Delta}_2$ [20]. ICOD then yields *two* large outlier eigenvalues of the modified inverse covariance matrix $\tilde{C}^{-1}$. The associated eigenvectors accurately recover both $\vec{\Delta}_1$ and $\vec{\Delta}_2$, after a final step of Independent Component Analysis (ICA) [11–13, 20].

Performance.— We further tested the performance of ICOD by systematically varying the selection bias (Fig. 3). ICOD achieves high Recovery of mutational-effect vectors for both single and double selection over a broad range of selection biases δE^*, albeit performance falls off in the limit of extreme bias. This confirms that our physical sectors are encoded in the bottom eigenvectors of the covariance matrix.

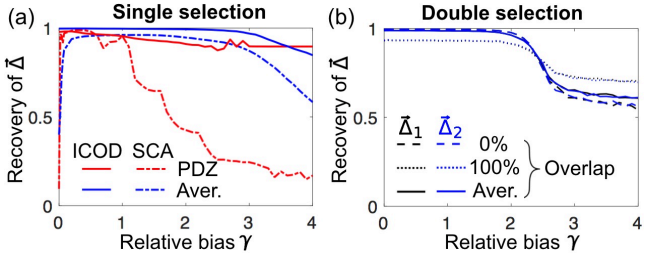

FIG. 3. Average recovery of mutational-effect vectors $\vec{\Delta}$ as a function of relative selection bias $\gamma \equiv (\delta E^* - \langle\delta E\rangle)/\sqrt{\langle(\delta E - \langle\delta E\rangle)^2\rangle}$. (a) Single selection. Different $\vec{\Delta}$s are used to generate sequence ensembles: the elastic-network $\vec{\Delta}$ from Fig. 1 (red); synthetic $\vec{\Delta}$s [20] with number of sites of large mutational effect (sector sites) ranging from 1 to 100, for sequences of length $L = 100$ (blue). Recovery is shown for ICOD (solid curves) and for SCA [13, 18] (dashed curves). (b) Double selection. Different pairs of synthetic $\vec{\Delta}$s [20] are used to generate sequence ensembles (with $L = 100$): "0%" indicates two non-overlapping sectors, each with 20 sites; "100%" indicates two fully overlapping sectors, each with 100 sites; "Aver." indicates average Recovery over 100 cases of double selection, where the single-sector size increases from 1 to 100, and the overlap correspondingly increases from 0 to 100. ICA was applied to improve Recovery [20].

How does ICOD compare with other approaches to identifying sectors? We compared the performance of ICOD with Statistical Coupling Analysis (SCA), the standard PCA-based method [13, 18]. In SCA, the covariance matrix C is reweighted by a site-specific conservation factor ϕ_l, the absolute value is taken, $\tilde{C}^{(\mathrm{SCA})}_{ll'} = |\phi_l C_{ll'} \phi_{l'}|$, and sectors are identified from the leading eigenvectors of $\tilde{C}^{(\mathrm{SCA})}$. We therefore tested the ability of the first eigenvector of $\tilde{C}^{(\mathrm{SCA})}$ to recover $\vec{\Delta}$ for a single selection. We found that the square root of the elements of the first SCA eigenvector can provide high Recovery of $\vec{\Delta}$ (Figs. 3, S10, S11) [20]. However, the performance of SCA relies on conservation through ϕ_l [19]; consequently, for unbiased selection, SCA breaks down [Fig. 3(a), dashed curves] and cannot identify sector sites (Fig. S14 [20]). ICOD does not suffer from such shortcomings, and performs well over a large range of selection. Note that in SCA, only the top eigenvectors of $\tilde{C}^{(\mathrm{SCA})}$ convey information about sectors (Figs. S10, S12).

We also compared ICOD with another PCA-based approach [6], which employs an inference method specific to the generalized Hopfield model, and should thus be well adapted to identifying sectors within our physical model [Eq. (3)]. Overall, this specialized approach performs similarly to ICOD, being slightly better for very localized sectors, but less robust than ICOD for strong selective biases and small datasets [20]. Exactly as for PCA and ICOD, within this method, the top Recovery is obtained for the bottom eigenvector of the (modified)

covariance matrix, consistent with $\vec{\Delta}$ in our model being a repulsive pattern [6], but large Recoveries are also obtained for the top eigenvectors (Fig. S15).

So far we considered binary sequences, with only one type of mutation with respect to the reference state. However, we have also shown that ICOD extends to mutations among the 20 different amino-acid types, plus gap state, relevant to real proteins [20]. On a multiple sequence alignment of PDZ domains, ICOD correctly predicts the majority of experimentally-determined residues [20].

Conclusion.—We have demonstrated how sectors of collectively correlated amino acids can arise from evolutionary constraints on the physical properties of proteins. Our physical sectors are directly associated with the small-eigenvalue modes of the covariance matrix. These modes also contain information about structural contacts [7], which helps explain previously observed correlations between sectors and contacts [33]. Previous works [7, 33] found that sectors could be recovered from the large-eigenvalue modes of the covariance matrix, and some signatures of our physical sectors are indeed obtained in these modes. However, the fundamental link we propose between physical sectors and small-eigenvalue modes is important, since large-eigenvalue modes of the covariance matrix also reflect information about subfamily-specific residues [34] and phylogeny [35].

Acknowledgments.—S.-W. W. and N. S. W. acknowledge the Center for the Physics of Biological Function under NSF Grant PHY-1734030. A.-F. B. and N. S. W. acknowledge the Aspen Center for Physics, which is supported by NSF Grant PHY-1607611. S.-W. W. was supported by the NSFC under Grants No. U1430237 and 11635002. S.-W. W. also acknowledges Tsinghua University for supporting a half-year-visit in Princeton University. N. S. W. was supported by NSF Grant MCB-1344191 and by NIH Grant R01 GM082938.

[1] S. W. Lockless and R. Ranganathan, Science **286**, 295 (1999).
[2] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan, Nat. Struct. Biol. **10**, 59 (2003).
[3] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan, Nature **437**, 512 (2005).
[18] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[20] R. N. McLaughlin Jr, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, Nature **491**, 138 (2012).
[13] O. Rivoire, K. A. Reynolds, and R. Ranganathan, PLoS Comput. Biol. **12**, e1004817 (2016).
[1] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, Annu. Rev. Biophys. **39**, 23 (2010).
[8] W. Zheng, B. R. Brooks, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **103**, 7664 (2006).
[9] S. Lukman and G. H. Grant, Proteins **76**, 588 (2009).
[10] T. E. Saldano, A. M. Monzon, G. Parisi, and S. Fernandez-Alberti, PLoS Comput. Biol. **12**, e1004775 (2016).
[11] D. Granata, L. Ponzoni, C. Micheletti, and V. Carnevale, Proc. Natl. Acad. Sci. U.S.A. **114**, E10612 (2017).
[2] P. De Los Rios, F. Cecconi, A. Pretre, G. Dietler, O. Michielin, F. Piazza, and B. Juanico, Biophys. J. **89**, 14 (2005).
[13] M. Delarue and Y. H. Sanejouand, J. Mol. Biol. **320**, 1011 (2002).
[14] W. Zheng and S. Doniach, Proc. Natl. Acad. Sci. U.S.A. **100**, 13253 (2003).
[15] L. Yan, R. Ravasio, C. Brito, and M. Wyart, Proc. Natl. Acad. Sci. U.S.A. , 201615536 (2017).
[16] S. Dutta, J.-P. Eckmann, A. Libchaber, and T. Tlusty, Proc. Natl. Acad. Sci. U.S.A. **115**, E4559 (2018).
[17] J. W. Rocks, N. Pashine, I. Bischofberger, C. P. Goodrich, A. J. Liu, and S. R. Nagel, Proc. Natl. Acad. Sci. U.S.A. **114**, 2520 (2017).
[18] D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon, Cell **85**, 1067 (1996).
[19] A. Y. Hung and M. Sheng, J. Biol. Chem. **277**, 5699 (2002).
[20] See Supplemental Material.
[21] H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook, Nat. Struct. Biol. **7 Suppl**, 957 (2000), `www.rcsb.org`.
[22] A. D. Cunningham, A. Colavin, K. C. Huang, and D. Mochly-Rosen, Cell Rep **18**, 2592 (2017).
[6] S. Cocco, R. Monasson, and V. Sessak, Phys. Rev. E **83**, 051123 (2011).
[7] S. Cocco, R. Monasson, and M. Weigt, PLOS Comput. Biol. **9**, e1003176 (2013).
[15] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).
[4] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).
[3] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PLoS ONE **6**, e28766 (2011).
[5] T. Plefka, J. Phys. A: Math. Gen. **15**, 1971 (1982).
[17] A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **113**, 12180 (2016).
[19] T. Teşileanu, L. J. Colwell, and S. Leibler, PLoS Comput. Biol. **11**, e1004091 (2015).
[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley and Sons, 2001).
[12] L. K. Hansen, J. Larsen, and T. Kolenda, in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, Vol. 5 (2001) pp. 3197–3200.
[33] O. Rivoire, Phys. Rev. Lett. **110**, 178102 (2013).
[34] G. Casari, C. Sander, and A. Valencia, Nat. Struct. Biol. **2**, 171 (1995).
[35] C. Qin and L. J. Colwell, Proc. Natl. Acad. Sci. U.S.A. **115**, 690 (2018).

Supplemental Material

CONTENTS

I. SUPPLEMENTAL METHODS AND RESULTS FOR ELASTIC NETWORK MODEL OF PDZ DOMAIN

To build the elastic-network model of the PDZ domain, we replace each of the $L = 76$ amino-acid residues by its corresponding alpha carbon Cα and beta carbon Cβ, as shown in Fig. 1(b). Every pair of carbons within a cutoff distance d_c are then connected with a harmonic spring [1]. Following an analysis of the same PDZ domain by De Los Rios et al. [2], we set $d_c = 7.5$ Å and assign spring constants as follows: a) 2 for Cα-Cα pairs if adjacent along the backbone, 1 otherwise; b) 1 for Cα-Cβ pairs; c) 0.5 for Cβ-Cβ pairs. For the following sequence analysis, we consider that mutations decrease the stiffnesses of the springs involving mutated Cβ atoms by a fraction $\epsilon = 0.2$.

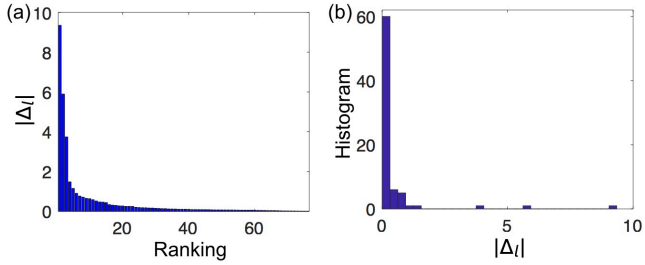

FIG. S1. Magnitude of single-site mutational effects Δ_l for the PDZ domain conformational change from Fig. 1. (a) Magnitudes by rank. (b) Histogram of magnitudes. According to our definition, the sites of large magnitude constitute "sector" sites with respect to selection on the energy cost of this conformational change, while all others are "non-sector" sites.

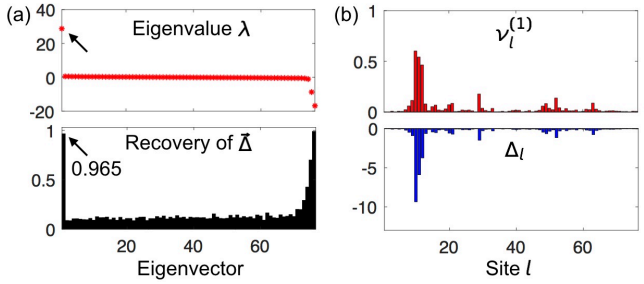

FIG. S2. Performance of ICOD for the selected sequence ensemble from Fig. 1. (a) Eigenvalues for ICOD method (upper) and Recovery of $\vec{\Delta}$ for all eigenvectors (lower). (b) Leading eigenvector $\nu_l^{(1)}$ (upper) and mutational effect Δ_l at site l [lower, same as in Fig. 1(f)]. The excellent performance of ICOD on this unbiased ensemble of sequences supports the general applicability of the ICOD method to both biased and unbiased sequence ensembles.

II. RECOVERY BY A RANDOM VECTOR

Here, we calculate the random expectation of the Recovery of the mutational-effect vector $\vec{\Delta}$ by a generic other vector $\vec{\nu}$, in order to establish a null model to which to compare. For a binary sequence, Recovery, as defined in Eq. (5), can be expressed as

$$\text{Recovery} = \vec{\Delta}' \cdot \vec{\nu'} = \sum_{l=1}^{L} \Delta'_l \, \nu'_l, \tag{S1}$$

with $\Delta'_l = |\Delta_l|/\sqrt{\sum_l \Delta_l^2}$ and $\nu'_l = |\nu_l|/\sqrt{\sum_l \nu_l^2}$. As before, L denotes the length of the sequence and hence the number of components of $\vec{\Delta}$ and $\vec{\nu}$. As $\vec{\nu'}$ is a normalized L-dimensional vector, its components can be expressed in

L-dimensional spherical coordinates using $L-1$ angles θ_i:

$$\nu'_l = \left(\prod_{i=1}^{l-1} \sin\theta_i\right) \cos\theta_l \qquad \forall l \in \{1,\dots,L-1\}\,, \tag{S2}$$

$$\nu'_L = \prod_{i=1}^{L-1} \sin\theta_i\,, \tag{S3}$$

where $\theta_i \in [0,\pi/2]$ for all $i \in \{1,\cdots,L\}$, because all components of $\vec{\nu'}$ are nonnegative. Note that we employ the usual convention that empty products are equal to one: Eq. (S2) yields $\nu'_1 = \cos\theta_1$.

The average Recovery for a random vector $\vec{\nu'}$ with an orientation uniformly distributed in the L-dimensional sphere reads:

$$\langle \text{Recovery} \rangle = \frac{\int_\Omega d\Omega \, \sum_l \Delta'_l \nu'_l}{\int_\Omega d\Omega} = \frac{\sum_l \Delta'_l \, I_l}{\int_\Omega d\Omega}\,, \tag{S4}$$

where the angular element is $d\Omega = \prod_{i=1}^{L-1} d\theta_i \, \sin^{L-i-1}(\theta_i)$, the integration domain is $\Omega = [0,\pi/2]^{L-1}$, and we have introduced $I_l = \int_\Omega d\Omega \, \nu'_l$. Using Eq. (S2), we obtain for $1 \le l \le L-1$

$$I_l = \int_\Omega d\Omega \, \nu'_l = \left(\prod_{i=1}^{l-1} \int_0^{\pi/2} d\theta_i \, \sin^{L-i}(\theta_i)\right) \left(\int_0^{\pi/2} d\theta_l \, \sin^{L-l-1}(\theta_l)\cos(\theta_l)\right) \left(\prod_{i=l+1}^{L-1} \int_0^{\pi/2} d\theta_i \, \sin^{L-i-1}(\theta_i)\right)\,, \tag{S5}$$

and similarly, Eq. (S3) yields

$$I_L = \int_\Omega d\Omega \, \nu'_L = \prod_{i=1}^{L-1} \int_0^{\pi/2} d\theta_i \, \sin^{L-i}(\theta_i)\,. \tag{S6}$$

Using the following results valid for $n > -1$:

$$\int_0^{\pi/2} d\theta \, \sin^n(\theta) = \frac{\sqrt{\pi}}{2} \frac{\Gamma\left(\frac{1+n}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}\,; \qquad \int_0^{\pi/2} d\theta \, \sin^n(\theta)\cos(\theta) = \frac{1}{n+1}\,, \tag{S7}$$

where Γ denotes the Euler Gamma function, which satisfies $\Gamma(x+1) = x\,\Gamma(x)$ for all x, we obtain for $1 \le l \le L$:

$$I_l = \frac{\pi^{(L-1)/2}}{2^{L-1}\,\Gamma\left(\frac{L+1}{2}\right)}, \tag{S8}$$

which is independent of l. Besides,

$$\int_\Omega d\Omega = \frac{\pi^{L/2}}{2^{L-1}\,\Gamma(L/2)}. \tag{S9}$$

Combining Eq. (S4) with Eqs. (S8) and (S9) finally yields

$$\langle \text{Recovery} \rangle = \frac{\sum_l \Delta'_l \, I_l}{\int_\Omega d\Omega} = \frac{\Gamma\left(L/2\right)}{\sqrt{\pi}\,\Gamma\left(\frac{L+1}{2}\right)} \sum_l \Delta'_l = \frac{\Gamma\left(L/2\right)}{\sqrt{\pi}\,\Gamma\left(\frac{L+1}{2}\right)} \frac{\sum_l |\Delta_l|}{\sqrt{\sum_l \Delta_l^2}}. \tag{S10}$$

In particular, in the relevant regime $L \gg 1$, an asymptotic expansion of Γ yields:

$$\langle \text{Recovery} \rangle \approx \sqrt{\frac{2}{\pi L}} \, \frac{\sum_l |\Delta_l|}{\sqrt{\sum_l \Delta_l^2}}. \tag{S11}$$

The maximum expectation of Recovery is obtained when all components of $\vec{\Delta}$, i.e. all mutational effects, are identical:

$$\langle \text{Recovery} \rangle_{\max} = \sqrt{\frac{2}{\pi}} \approx 0.798. \tag{S12}$$

Conversely, the average Recovery becomes minimal when only one component of $\vec{\Delta}$ is nonzero, which constitutes the limit of the case where the mutational effect at one site is dominant:

$$\langle \text{Recovery} \rangle_{\min} = \sqrt{\frac{2}{\pi L}}, \tag{S13}$$

which approaches zero in the limit $L \to \infty$.

III. INVERSE COVARIANCE MATRIX OF OUR SEQUENCE ENSEMBLES

Here, we present a derivation of the small-coupling approximation of the inverse covariance matrix for our artificially-generated sequence ensembles. In this small-coupling limit, the inverse covariance matrix provides an estimate of the energetic couplings used to generate the data. More generally, deducing energetic parameters from observed statistics is a well-known inference problem, also known as an inverse problem. Two-body energetic couplings can be inferred from the one and two-body frequencies observed in the data, using a standard maximum entropy approach. However, the exact calculation of the energetic terms is difficult, and various approximations have been developed. Following Refs [3, 4], we use the mean-field or small-coupling approximation, which was introduced in Ref. [5] for the Ising spin-glass model. For the sake of completeness, we now review the main steps of the calculation, which follow Ref. [4]. Note that we do not use inference methods specific to low-rank coupling matrices [6, 7] because we wish to retain generality, with the application to real sequence data in mind.

We begin with the case of binary sequences, which is discussed in the main text. Following that, we generalize to cases where more than two states are allowed at each site, such as the 21 possible states for real protein sequence (20 amino acids plus gap).

A. Binary sequences

We begin by deriving Eq. (6) from the main text, which provides an approximation for the inverse covariance matrix of the ensembles of our binary artificial sequences. Each sequence $\vec{S}$ is such that $S_l \in \{0, 1\}$ for each site l with $1 \leq l \leq L$, where L is the length of the sequence.

1. *From a sector model for binary sequences to an Ising model*

Recall the fitness w of a binary sequence $\vec{S}$ in our sector model [Eq. (3)]:

$$w(\vec{S}) = -\frac{\kappa}{2}\left(\sum_l \Delta_l S_l - \delta E^*\right)^2 . \tag{S14}$$

We introduce $s_l = 2S_l - 1$: it is an "Ising spin" variable ($S_l = 0 \Leftrightarrow s_l = -1$ and $S_l = 1 \Leftrightarrow s_l = 1$). The fitness in Eq. (S14) can be rewritten as

$$w(\vec{s}) = -\frac{\kappa}{2}\left(\sum_l D_l s_l - \alpha\right)^2 , \tag{S15}$$

with $D_l = \Delta_l/2$ and $\alpha = \delta E^* - \sum_l D_l$. Expanding yields

$$w(\vec{s}) = -\frac{\kappa}{2}\left(\sum_{l\neq p} D_l D_p s_l s_p + \sum_l D_l^2 - 2\alpha \sum_l D_l s_l + \alpha^2\right) , \tag{S16}$$

where we have used the fact that $s_l^2 = 1$. The second term and the last term in Eq. (S16) are both constants, and therefore our fitness is equivalent to

$$w(\vec{s}) = -\frac{\kappa}{2}\left(\sum_{l\neq p} D_l D_p s_l s_p - 2\alpha \sum_l D_l s_l\right) . \tag{S17}$$

This fitness has the form of a standard Ising Hamiltonian with inter-spin couplings and local fields, albeit with the convention difference in overall sign between fitness and energy.

2. *First-order small-coupling expansion*

We next consider the general Ising Hamiltonian with inter-spin couplings and local fields

$$H(\vec{s}) = -\frac{1}{2}\epsilon \sum_{i\neq j} J_{ij} s_i s_j - \sum_i h_i s_i , \tag{S18}$$

where ϵ is a constant to be employed in a small-coupling expansion. With this Hamiltonian, taking thermal energy $k_{\mathrm{B}}T = 1$, the equilibrium probability of finding a particular sequence $\vec{s}$ is

$$P(\vec{s}) = \frac{1}{Z} e^{-H(\vec{s})}, \tag{S19}$$

where $Z = \sum_{\vec{s}} e^{-H(\vec{s})}$.

Introducing $F = -\log Z$, we have

$$\begin{aligned} \frac{\partial F}{\partial h_i} &= -\langle s_i \rangle = -m_i \,, \\ \frac{\partial^2 F}{\partial h_i \partial h_j} &= -\frac{\partial m_i}{\partial h_j} = \langle s_i \rangle \langle s_j \rangle - \langle s_i s_j \rangle = -C'_{ij} \,, \end{aligned} \tag{S20}$$

where, following the Ising terminology, m_i denotes the average magnetization at site i, while C' denotes the covariance matrix in the Ising convention. Note that, using the identity $m_i = 2P_i - 1$, where P_i denotes the probability that $s_i = 1$, we obtain

$$C'_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle = 4\,(P_{ij} - P_i P_j) = 4\,C_{ij} \,, \tag{S21}$$

where P_{ij} is the probability that $s_i = s_j = 1$, and C denotes the covariance matrix in the Potts convention, which is used in the main text because it allows straightforward generalization to the case where more than two states are possible at each site.

Performing a Legendre transform, we introduce $G = F + \sum_i m_i h_i$, yielding

$$\frac{\partial G}{\partial m_i} = h_i \,, \tag{S22}$$

$$\frac{\partial^2 G}{\partial m_i \partial m_j} = \frac{\partial h_i}{\partial m_j} = {C'_{ij}}^{-1} \,. \tag{S23}$$

We now perform a small-coupling expansion and express G to first order in ϵ (see Eq. (S18)): $G(\epsilon) \approx G(0) + \epsilon G'(0)$. Since sites are independent for $\epsilon = 0$, it is straightforward to express $G(0)$ and $G'(0)$ as a function of the one-body expectations, represented by m_i, and of the couplings. We obtain

$$G(0) = \sum_i \frac{m_i + 1}{2} \log\left(\frac{m_i + 1}{2}\right) + \frac{1 - m_i}{2} \log\left(\frac{1 - m_i}{2}\right) \,, \tag{S24}$$

and

$$G'(0) = \frac{\partial G}{\partial \epsilon}(0) = -\frac{1}{2} \sum_{i \neq j} J_{ij} m_i m_j \,. \tag{S25}$$

Using these expressions, and taking $\epsilon = 1$ in the expansion, we obtain the following approximation for G:

$$G \approx \sum_i \frac{m_i + 1}{2} \log\left(\frac{m_i + 1}{2}\right) + \frac{1 - m_i}{2} \log\left(\frac{1 - m_i}{2}\right) - \frac{1}{2} \sum_{i \neq j} J_{ij} m_i m_j \,. \tag{S26}$$

Using Eqs. (S22) and (S23), we obtain the elements of the inverse covariance matrix from Eq. (S26):

$$\begin{aligned} {C'_{kl}}^{-1} &= -J_{kl} \,, \ \forall l \neq k \,, \\ {C'_{ll}}^{-1} &= \frac{1}{2}\left(\frac{1}{1 + m_l} + \frac{1}{1 - m_l}\right) = \frac{1}{4}\left(\frac{1}{P_l} + \frac{1}{1 - P_l}\right) \,, \end{aligned} \tag{S27}$$

where P_l denotes the probability that $s_l = 1$.

Note that Eq. (S26) is a first-order small-coupling (or mean-field) approximation. The expansion can be extended to higher order, and the second-order expansion is known as the Thouless, Anderson, and Palmer (TAP) free energy [5, 8].

3. Application to our sector model

Comparing Eqs. (S17) and (S18) (with $\epsilon = 1$) allows us to identify the couplings in our sector model as

$$J_{kl} = -\kappa\, D_k D_l = -\kappa\, \Delta_k \Delta_l/4\,,\ \forall k \neq l\,. \tag{S28}$$

Note that this expression is in the Ising gauge (also known as the zero-sum gauge). Recall also that the link to the Potts convention is made through $C' = 4\,C$ [Eq. (S21)], which implies ${C'}^{-1} = C^{-1}/4$. Finally, recall that fitness and energy have opposite signs.

Hence, in the Potts convention, Eq. (S27) yields for our sector model:

$$\begin{aligned} C^{-1}_{kl} &= \kappa \Delta_k \Delta_l\,,\ \forall l \neq k\,,\\ C^{-1}_{ll} &= \frac{1}{P_l} + \frac{1}{1-P_l}\,. \end{aligned} \tag{S29}$$

This corresponds to Eq. (6) in the main text.

B. Sequences with q possible states at each site

1. From a sector model to a Potts model for sequences

Motivated by the fact that a real protein sequence has 21 possible states at each site (20 different amino acids plus gap), we now generalize the above result to the case where q states are possible at each site. We denote these states by α with $\alpha \in \{1, .., q\}$. Our sector model can then be mapped to a q-state Potts model. The length-L vector $\vec{\Delta}$ of single-site mutational effects introduced in the two-state case in the main text is replaced by a $(q-1) \times L$ matrix of mutational effects, each being denoted by $\Delta_l(\alpha_l)$. These mutational effects can be measured with respect to a reference sequence $\vec{\alpha}^0$ satisfying $\Delta_l(\alpha_l^0) = 0,\ \forall l \in \{1, \ldots, L\}$: at each site l, the state present in the reference sequence $\vec{\alpha}^0$ serves as the reference with respect to which the mutational effects at that site are measured. For the sake of simplicity, we will take state q as reference state at all sites. This does not lead to any loss of generality, since it is possible to reorder the states for each l.

The generalization of the fitness function Eq. (3) [Eq. (S14)] to our q-state model can be written as

$$w(\vec{\alpha}) = -\frac{\kappa}{2}\left(\sum_{l=1}^{L} \Delta_l(\alpha_l) - \delta E^*\right)^2\,. \tag{S30}$$

Expanding this expression, discarding a constant term, and using the fact that there can only be one state at each site, we find that the fitness of sequences can be expressed as

$$w(\vec{\alpha}) = -\frac{\kappa}{2}\sum_{l\neq k} \Delta_l(\alpha_l)\Delta_k(\alpha_k) - \frac{\kappa}{2}\sum_{l=1}^{L} \Delta_l(\alpha_l)\left(\Delta_l(\alpha_l) - 2\,\delta E^*\right)\,. \tag{S31}$$

This is a particular case of the more general Potts Hamiltonian

$$H(\vec{\alpha}) = -\frac{1}{2}\sum_{l\neq k} e_{lk}(\alpha_l, \alpha_k) - \sum_{l=1}^{L} h_l(\alpha_l)\,, \tag{S32}$$

which is the one usually considered in Direct Coupling Analysis (DCA) [3, 4].

In order to identify Eq. (S31) and Eq. (S32), one must deal with the degeneracies present in Eq. (S32), where the number of independent parameters is $L(q-1) + L(L-1)(q-1)^2/2$ [9]. To lift this degeneracy, we choose the gauge usually taken in mean-field DCA [4]: $e_{lk}(\alpha_l, q) = e_{lk}(q, \alpha_k) = h_l(q) = 0$ for all l, k, α_l, α_k. This choice is consistent with taking state q as the reference state for mutational effects (see above), and we will refer to it as the reference-sequence gauge. This gauge choice enables us to identify the couplings between Eq. (S31) and Eq. (S32):

$$e_{lk}(\alpha_l, \alpha_k) = -\kappa \Delta_l(\alpha_l)\Delta_k(\alpha_k)\,, \tag{S33}$$

for all $l \neq k$, and all α_l, α_k, with $\Delta_l(q) = 0$ for all l (recalling that fitness and energy have opposite signs).

2. First-order small-coupling expansion

The derivation of the first-order mean-field or small-coupling approximation for q-state models is very similar to the Ising case presented above. Hence, we will simply review the main results (see Ref. [4]).

We start with the Hamiltonian

$$H(\vec{\alpha}) = -\frac{\epsilon}{2}\sum_{l\neq k} e_{lk}(\alpha_l,\alpha_k) - \sum_{l=1}^{L} h_l(\alpha_l)\,, \tag{S34}$$

where ϵ has been introduced to perform the small-coupling expansion. Eq. (S34) coincides with Eq. (S32) for $\epsilon = 1$. Considering $F = -\log(Z)$ with $Z = \sum_{\vec{\alpha}} e^{-H(\vec{\alpha})}$, where $H(\vec{\alpha})$ is the Potts Hamiltonian in Eq. (S34), we have for all k and all $\alpha_k < q$:

$$\frac{\partial F}{\partial h_k(\alpha_k)} = -P_k(\alpha_k)\,, \tag{S35}$$

where $P_k(\alpha_k)$ is the one-body probability. Similarly, we have for all k, l and all $\alpha_k < q$ and $\alpha_l < q$:

$$\frac{\partial^2 F}{\partial h_l(\alpha_l)\partial h_k(\alpha_k)} = -\frac{\partial P_k(\alpha_k)}{\partial h_l(\alpha_l)} = -C_{kl}(\alpha_k,\alpha_l)\,, \tag{S36}$$

where we have introduced the covariance $C_{kl}(\alpha_k,\alpha_l) = P_{kl}(\alpha_k,\alpha_l) - P_k(\alpha_k)P_l(\alpha_l)$.

We perform a Legendre transform and introduce $G = F - \sum_i \sum_{\alpha_i} h_i(\alpha_i)P_i(\alpha_i)$, yielding

$$\frac{\partial G}{\partial P_k(\alpha_k)} = h_k(\alpha_k)\,, \tag{S37}$$

$$\frac{\partial^2 G}{\partial P_l(\alpha_l)\partial P_k(\alpha_k)} = \frac{\partial h_l(\alpha_l)}{\partial P_k(\alpha_k)} = C^{-1}_{kl}(\alpha_k,\alpha_l)\,, \tag{S38}$$

for all k, l and all $\alpha_k < q$ and $\alpha_l < q$. Note that, in the latter equation, $C^{-1}_{kl}(\alpha,\beta)$ is shorthand for A^{-1}_{ij}, where A is the $(q-1)L \times (q-1)L$ covariance matrix where terms involving the reference state q have been excluded: $A_{ij} = C_{kl}(\alpha,\beta)$, where $i = (q-1)(k-1)+\alpha$ and $j = (q-1)(l-1)+\beta$, with $\alpha \in \{1,\ldots,q-1\}$ and $\beta \in \{1,\ldots,q-1\}$ [10].

We next perform a first-order expansion of G in ϵ, and take $\epsilon = 1$, yielding:

$$G \approx \sum_l \sum_{\alpha_l} P_l(\alpha_l)\log\left(P_l(\alpha_l)\right) - \frac{1}{2}\sum_{l\neq k}\sum_{\alpha_l,\alpha_k} e_{lk}(\alpha_l,\alpha_k)P_l(\alpha_l)P_k(\alpha_k)\,. \tag{S39}$$

Applying Eqs. (S37), (S38) to Eq. (S39), and using $P_l(q) = 1 - \sum_{\alpha_l<q} P_l(\alpha_l)$ gives

$$\begin{aligned} C^{-1}_{kl}(\alpha_k,\alpha_l) &= -e_{kl}(\alpha_k,\alpha_l),\ \ \forall\, l \neq k\,,\\ C^{-1}_{ll}(\alpha_k,\alpha_l) &= \frac{1}{P_k(q)} + \frac{\delta_{\alpha_k\alpha_l}}{P_k(\alpha_k)}\,. \end{aligned} \tag{S40}$$

This result is the standard one found in DCA [4].

3. Application to our sector model

Combining Eqs. (S33) and (S40), we obtain for our sector model:

$$\begin{aligned} C^{-1}_{kl}(\alpha_k,\alpha_l) &= \kappa\Delta_k(\alpha_k)\Delta_l(\alpha_l),\ \ \forall\, l \neq k\,,\\ C^{-1}_{ll}(\alpha_k,\alpha_l) &= \frac{1}{P_k(q)} + \frac{\delta_{\alpha_k\alpha_l}}{P_k(\alpha_k)}\,. \end{aligned} \tag{S41}$$

For $q = 2$, Eq. (S41) reduces to Eq. (S27) [Eq. (6) in the main text], using $1 - P_l = P_l(q)$.

4. *Multiple selections*

So far, we have mainly discussed the case where there is only one selection (one sector). However, real proteins face various selection pressures. The generalization of the fitness in Eq. (S30) to N simultaneous different selections reads

$$w(\vec{S}) = -\sum_{i=1}^{N} \frac{\kappa_i}{2} \left(\sum_l \Delta_{i,l}(\alpha_l) - \delta E_i^* \right)^2 , \tag{S42}$$

which corresponds to Eq. (8) in the main text. We choose the reference-state gauge, assuming again for simplicity that the reference state is q at each site. The identification to the general Potts Hamiltonian Eq. (S32) (recalling that fitnesses and energies have opposite signs) then yields

$$e_{lk}(\alpha_l, \alpha_k) = -\sum_{i=1}^{N} \kappa_i \Delta_{i,l}(\alpha_l) \Delta_{i,k}(\alpha_k) , \tag{S43}$$

which generalizes Eq. (S33) to the multiple selection case. Using the small-coupling expansion result in Eq. (S40), we obtain the following approximation for the inverse covariance matrix:

$$\begin{aligned} C^{-1}_{kl}(\alpha_k, \alpha_l) &= \sum_{i=1}^{N} \kappa_i \Delta_{i,k}(\alpha_k) \Delta_{i,l}(\alpha_l), \ \ \forall\, l \neq k , \\ C^{-1}_{ll}(\alpha_k, \alpha_l) &= \frac{1}{P_k(q)} + \frac{\delta_{\alpha_k \alpha_l}}{P_k(\alpha_k)} . \end{aligned} \tag{S44}$$

This generalizes Eq. (S41) to the multiple selection case.

IV. ROBUSTNESS OF ICOD

In the main text, we introduced the Inverse Covariance Off-Diagonal (ICOD) method to identify protein sectors from sequence data. The ICOD method exploits the approximate expression derived above for the inverse covariance matrix [Eq. (S41)]; in particular, ICOD makes use of the fact that the off-diagonal elements of C^{-1} are simply related to the elements of the mutational effect vector $\vec{\Delta}$. In this section, we first describe our comparison of ICOD to SCA for single selection, and detail our test of ICOD for double selection, using synthetic binary sequences. Next, we show how the ICOD method can be extended to sequences with more than two states per site, and demonstrate its robustness to gauge choice and pseudocounts.

A. Robustness of ICOD to selection bias and multiple selections

To quantify the performance of ICOD and to compare to SCA over a range of selection biases we focused on binary sequences. To obtain the average curve for single selections in Fig. 3(a), we first generated 100 distinct synthetic $\vec{\Delta}$s, one for each sector size from $n = 1$ to 100, where sector sites are defined as those with large mutational effects. To this end, the mutational effects of the sector sites and the non-sector sites were sampled, respectively, from zero-mean Gaussian distributions with standard deviations 20 and 1. For each sector size and each selection bias we generated a sequence ensemble of 50,000 random sequences and weighted each sequence according to the distribution

$$P(\vec{S}) = \frac{\exp(w(\vec{S}))}{\sum_{\vec{S}} \exp(w(\vec{S}))} , \tag{S45}$$

where $w(\vec{S})$ is the fitness of sequence $\vec{S}$, given by the single selection formula Eq. (3). In general, we wish to employ a selection window whose width in energy (or any other selected variable) scales with the overall width of the unselected distribution. Hence, as mentioned in the main text, we perform all selections with a strength

$$\kappa = \frac{10}{\sum_l \Delta_l^2} . \tag{S46}$$

Then, for each method (ICOD or SCA), performance as measured by Recovery of $\vec{\Delta}$ by the first eigenvector was averaged over the 100 different sector sizes.

Similarly, to obtain the average curve for double selection in Fig. 3(b), we generated 100 distinct pairs of $\vec{\Delta}_1$s and $\vec{\Delta}_2$s, one pair for each sector size from $n = 1$ to 100. Specifically, the sector for $\vec{\Delta}_1$ consisted of the first n sites, while the sector for $\vec{\Delta}_2$ corresponded to the last n sites, so that the two sectors overlap for $n > 50$. As for the single selections, the mutational effects of the sector sites and the non-sector sites were sampled, respectively, from Gaussian distributions with standard deviations 20 and 1. As an example, two synthetic $\vec{\Delta}$s for $n = 20$ are shown in Fig. S3. Again, for each sector size and each selection bias, we generated an ensemble of 50,000 random sequences and weighted them according to Eq. (S45) along with the double selection formula Eq. (S42) (i.e. Eq. (8) in the main text). The performance of ICOD as measured by Recovery of $\vec{\Delta}_1$ and $\vec{\Delta}_2$ by the first two eigenvectors was averaged over the 100 different sector sizes. In Fig. 3(b) we also reported the performance of ICOD for two non-overlapping sectors, each with 20 sites, and for two fully overlapping sectors, each with 100 sites. We followed a protocol similar to that described above, but in each of these cases, we averaged Recovery over 100 realizations using distinct pairs of $\vec{\Delta}_1$ and $\vec{\Delta}_2$.

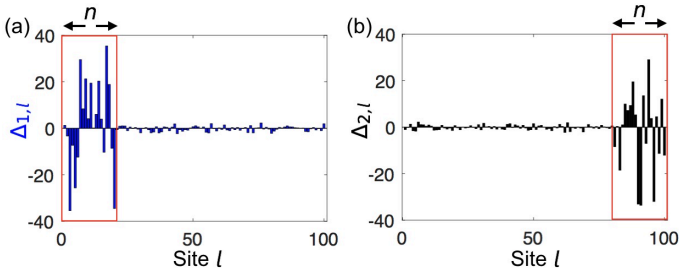

FIG. S3. Example of two synthetic $\vec{\Delta}$s generated for the double selection in Fig. 3(b). (a) Generation of $\vec{\Delta}_1$, where the mutational effects for the first 20 sites and for the last 80 sites are sampled, respectively, from zero-mean Gaussian distributions with a standard deviation of 20 and 1. (b) Generation of $\vec{\Delta}_2$, where the mutational effects for the last 20 sites and for the first 80 sites are sampled, respectively, from zero-mean Gaussian distributions with a standard deviation of 20 and 1.

Unless otherwise stated, data for other plots were generated in the same way, i.e. using 50,000 random sequences, sequence length $L = 100$, selection strength κ in Eq. (S46), and standard deviation 20/1 of Δ_l in the sector/non-sector sites.

Note that to improve Recovery in the case of double selection, we applied Independent Component Analysis (ICA) [11–13] to the first two eigenvectors in order to disentangle the contributions coming from the two constraints. In general, we expect that the first N eigenvectors of the ICOD matrix $\tilde{C}^{-1}$ will report N constraints. However, each of these N eigenvectors is likely to include a mixture of contributions from different constraints. Applying ICA to the first N eigenvectors to recover the individual constraints amounts to assuming that all the constraints are statistically independent. As an example, in Fig. S4, we consider the case of two selections targeting a different set of sites and with different selection windows (one biased, one non-biased). In this case, ICOD plus ICA yields excellent Recovery (Fig. S4). Without ICA, the results are noticeably worse (Fig. S5). Moreover, Fig. 3(b) demonstrates that ICOD plus ICA can achieve a high Recovery for a broad range of overlaps between two sectors.

In Fig. 3(b), one observes a slight decrease of performance of ICOD plus ICA for double selection with overlapping sectors. Does this arise from increasing sector size or from increasing overlap? As expected from Eqs. (5) and (7), Fig. S6(a) shows that Recovery does not fall off with increased sector size. Thus, we tested whether larger sector overlaps could reduce Recovery. Fig. S6(b) shows that this is indeed the case for sequence ensembles subject to two selections each with a fixed sector size of 20, but with different numbers of overlapping sites. However, the reduction of Recovery is quite modest, as even for 100% overlap, Recovery remains above 0.9. It is interesting to note that, independent of sector size and overlap, Recovery decreases faster for double selection than for single selection at large relative biases (see Figs. 3 and S6).

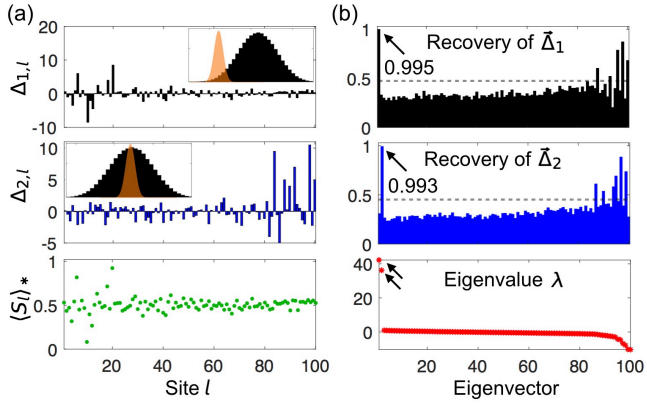

FIG. S4. ICOD method for double selection. (a) Upper panels: Components at each site l of two synthetically generated mutational-effect vectors, with insets showing biased selection around δE_1^* for $\vec{\Delta}_1$ and neutral selection around δE_2^* for $\vec{\Delta}_2$. Lower panel: average mutant fraction $\langle S_l \rangle_*$ at site l after double selection. (b) Performance of ICOD method. Recovery of $\vec{\Delta}_1$ and $\vec{\Delta}_2$ for all eigenvectors (upper) and corresponding eigenvalues (lower). The gray dashed line indicates the random expectation of Recovery [Eq. (S11)].

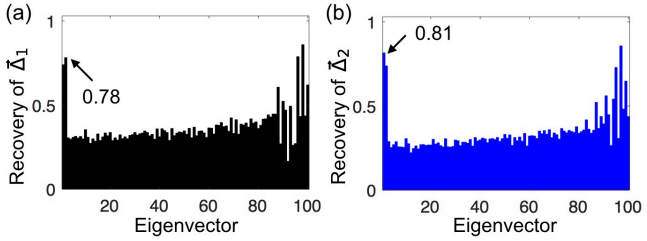

FIG. S5. Performance of ICOD for the two-sector case in Fig. S4, without applying ICA.

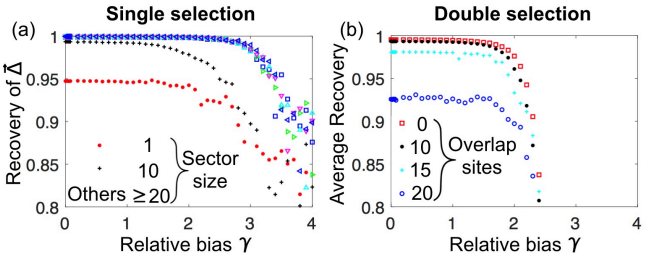

FIG. S6. Performance of ICOD for different sector sizes and sector overlaps. (a) Single selection with varying sector size. Recovery is shown as a function of relative selection bias $\gamma \equiv (\delta E^* - \langle \delta E \rangle)/\sqrt{\langle (\delta E - \langle \delta E \rangle)^2 \rangle}$ for sectors of size 1, 10, 20, 40, 60, 80, and 100 out of 100 sequence sites (*cf.* Fig. 3(a)). Recovery is almost perfect for sectors of size larger than 10, but is substantially lower for sector size 1, which violates the criteria $\Delta_l \ll \sqrt{\sum_{l'} \Delta_{l'}^2}$. (b) Double selection with different degrees of sector overlap. For each selection, the sector size is 20 out of 100 sequence sites, and the overlap varies from 0 to 20 sites. The average Recovery for $\vec{\Delta}_1$ and $\vec{\Delta}_2$ is shown as a function of relative selection bias. The data in (b) is averaged over 20 realizations of $\vec{\Delta}$s.

B. Multiple states per site and alternative gauge choice

In Section III B above, we described how to generalize from binary sequences to sequences with q possible states at each site. Correspondingly, we now generalize the ICOD method to higher values of q. Since we are interested in extracting the single-site mutational effects $\Delta_l(\alpha_l)$ with respect to a reference state at each site, we can simply set to zero the diagonal blocks of C^{-1} in Eq. (S44), yielding the modified inverse covariance matrix

$$\tilde{C}^{-1}_{kl}(\alpha_k, \alpha_l) = (1 - \delta_{kl}) \sum_{i=1}^{N} \kappa_i \Delta_{i,k}(\alpha_k) \Delta_{i,l}(\alpha_l)\,, \tag{S47}$$

for the case of multiple selections, or more simply for a single selection

$$\tilde{C}^{-1}_{kl}(\alpha_k, \alpha_l) = (1 - \delta_{lk})\, \kappa \Delta_k(\alpha_k) \Delta_l(\alpha_l). \tag{S48}$$

This equation generalizes Eq. (7) obtained for $q = 2$ in the main text. As in that case, the first eigenvector of $\tilde{C}^{-1}$ (associated with the largest eigenvalue) should accurately report the single-site mutational effects $\Delta_k(\alpha_k)$. Indeed, Fig. S7 shows that this generalized version of ICOD performs very well on synthetic data generated for the case $q = 21$ relevant to real protein sequences. Note that in the reference-sequence gauge, Recovery generalizes naturally to the q-state model as

$$\text{Recovery} = \frac{\sum_{l,\alpha_l} |\nu_l(\alpha_l) \Delta_l(\alpha_l)|}{\sqrt{\sum_{l,\alpha_l} \nu_l(\alpha_l)^2} \sqrt{\sum_{l,\alpha_l} \Delta_l(\alpha_l)^2}}, \tag{S49}$$

where the sums over states α_l do not include the reference state at each site.

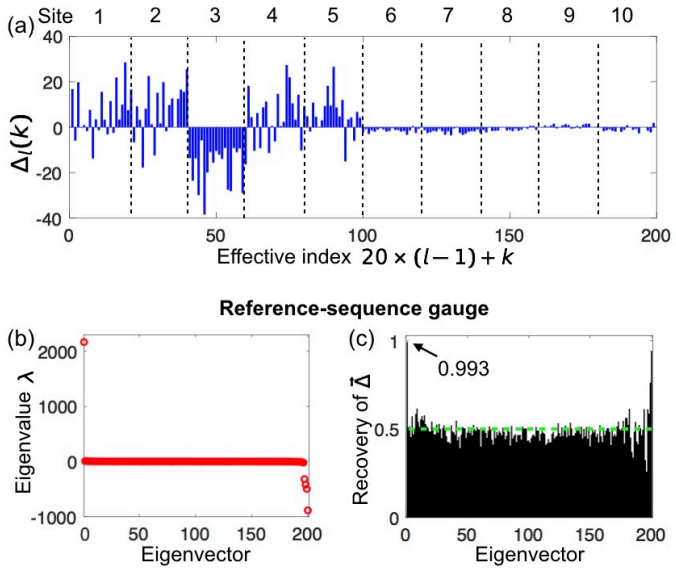

FIG. S7. Performance of ICOD on synthetic sequence data with $q = 21$ possible states at each site. (a) Mutational effects $\Delta_l(k)$ with respect to a reference sequence, chosen for simplicity to be state 21 at every site. The mutational effect at $q = 21$ is not shown. Note that while mutational effects are initially generated from a Gaussian distribution, *relative* mutational effects (calculated with respect to the reference sequence) can have a systematic bias at each site l. (b) Eigenvalues of the ICOD-modified inverse covariance matrix $\tilde{C}^{-1}$ defined in Eq. (S48). (c) Recovery of $\vec{\Delta}$ (see Eq. (S49)). The green dashed line indicates the random expectation of Recovery [Eq. (S11)].

While the reference-sequence gauge is convenient and allows a clear interpretation of the mutational effects, other gauge choices are possible. For instance, in the DCA literature, the zero-sum (or Ising) gauge is often employed [10, 14]. In this gauge, the couplings satisfy

$$\sum_{\alpha} e_{ij}(\alpha, \beta) = \sum_{\beta} e_{ij}(\alpha, \beta) = 0\,, \tag{S50}$$

Qualitatively, the gauge degree of freedom means that contributions to the Hamiltonian in Eq. (S32) can be shifted between the fields and the couplings [15]. In DCA, the focus is on identifying the dominant two-body interactions, so one does not want the couplings to include contributions that can be accounted for by the one-body fields [9]. The zero-sum gauge satisfies this condition because it minimizes the Frobenius norms of the couplings

$$\|e_{ij}\| = \sqrt{\sum_{\alpha,\beta=1}^{q} \left[e_{ij}(\alpha,\beta)\right]^2}. \tag{S51}$$

Hence, the zero-sum gauge attributes the smallest possible fraction of the energy in Eq. (S32) to the couplings, and the largest possible fraction to the fields [10, 15]. In order to transform to the zero-sum gauge defined in Eq. (S50), each coupling $e_{ij}(\alpha,\beta)$ is replaced by

$$\tilde{e}_{ij}(\alpha,\beta) = e_{ij}(\alpha,\beta) - \langle e_{ij}(\zeta,\beta)\rangle_\zeta - \langle e_{ij}(\alpha,\eta)\rangle_\eta + \langle e_{ij}(\zeta,\eta)\rangle_{\zeta,\eta}\,, \tag{S52}$$

where $\langle . \rangle_\zeta$ denotes an average over $\zeta \in \{1, ..., q\}$ [10].

Shifting from the reference-sequence gauge where one state (in our derivations, state q) is taken as a reference at each site to the zero-sum gauge requires the replacement

$$\tilde{\Delta}_l(\alpha) = \Delta_l(\alpha) - \frac{1}{q}\sum_{\beta=1}^{q} \Delta_l(\beta), \tag{S53}$$

The new reference-state-free mutational effects satisfy $\sum_{\beta=1}^{q} \tilde{\Delta}_l(\beta) = 0$, and the associated couplings $\tilde{e}_{lk}(\alpha_l,\alpha_k) = -\kappa\tilde{\Delta}_l(\alpha_l)\tilde{\Delta}_k(\alpha_k)$ (see Eq. (S33)) are related to the initial ones $e_{lk}(\alpha_l,\alpha_k)$ through Eq. (S52).

Importantly, these reference-state-free mutational effects can be used to assess the overall importance of mutations at each particular site in the sequence. To this end, let us introduce the Frobenius norm of the reference-state-free mutational effects:

$$||\Delta_l|| = \sqrt{\sum_{\beta=1}^{q} \left(\tilde{\Delta}_l(\beta)\right)^2}. \tag{S54}$$

This quantity, which we refer to as the "site significance", measures the overall importance of mutational effects at site l. In order to assess site significances from an ensemble of sequences, without investigating the impact of each particular mutation at each site, one can apply the zero-sum gauge to the ICOD-modified inverse covariance matrix (see Eq. (S48)), and compute the Frobenius norm of each 20×20 block associated to each pair of sites (i, j) according to Eq. (S51). The first eigenvector of this compressed $L \times L$ matrix accurately reports the mutational significance of each site, as illustrated in Fig. S8. Specifically, it yields a high Recovery of site significances as defined in Eq. (S54) [see Fig. S8(c)], and it successfully predicts the most important sites, i.e. the sector sites, in our synthetic data [see Fig. S8(d)].

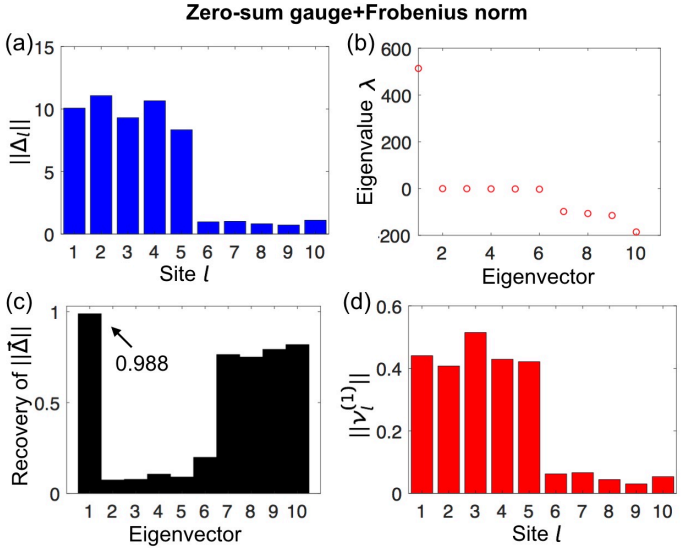

FIG. S8. Assessing site significance for synthetic sequence data. The same synthetic data as in Fig. S7 (with $q = 21$ possible states at each site) is used. (a) Significance $||\Delta_l||$ of each site l, computed directly by applying Eqs. (S53) and (S54) to the mutational effects $\Delta_l(k)$ shown in Fig. S7(a). (b) Eigenvalues of the compressed $(L \times L)$ ICOD-modified inverse covariance matrix, calculated by applying the zero-sum gauge to the ICOD-modified inverse covariance matrix [see Eq. (S48)], and by computing the Frobenius norm of each 20×20 block associated to each pair of sites (i, j) according to Eq. (S51). (c) Recovery of site significances $||\vec{\Delta}||$ from each eigenvector of the compressed ICOD-modified inverse covariance matrix [see panel (a) and Eq. (S49)] (d) Estimated site significances computed from the first eigenvector $\vec{\nu}^{(1)}$ of the compressed ICOD-modified inverse covariance matrix.

C. Pseudocounts

As pseudocounts are often necessary to regularize real sequence data, and as a high fraction of pseudocounts is generally used in DCA, we consider here whether the ICOD method is robust to the addition of pseudocounts.

Until now, we used only raw empirical frequencies obtained from sequence data. For instance, one-body frequencies were obtained by counting the number of sequences where a given state occured at a given site and dividing by the total number M of sequences in the ensemble. Covariances were computed from the empirical single-site frequencies of occurrence of each state α at each site i, denoted by $f_i^e(\alpha)$, and the empirical two-site frequencies of occurrence of each ordered pair of states (α, β) at each ordered pair of sites (i, j), denoted by $f_{ij}^e(\alpha, \beta)$. Specifically, we obtained the covariance matrix as $C_{ij}(\alpha, \beta) = f_{ij}^e(\alpha, \beta) - f_i^e(\alpha) f_j^e(\beta)$ [15].

To avoid issues arising from limited sample size, such as states that never appear at some sites (which present mathematical difficulties, e.g. a non-invertible covariance matrix [4]), one can introduce pseudocounts via a parameter Λ [3, 4, 15, 16]. The one-site frequencies f_i then become

$$f_i(\alpha) = \frac{\Lambda}{q} + (1 - \Lambda) f_i^e(\alpha)\,, \tag{S55}$$

where q is the number of states per site. Similarly, the two-site frequencies f_{ij} become

$$f_{ij}(\alpha, \beta) = \frac{\Lambda}{q^2} + (1 - \Lambda) f_{ij}^e(\alpha, \beta) \text{ if } i \neq j\,, \tag{S56}$$

$$f_{ii}(\alpha, \beta) = \frac{\Lambda}{q} \delta_{\alpha\beta} + (1 - \Lambda) f_{ii}^e(\alpha, \beta) = f_i(\alpha) \delta_{\alpha\beta}\,. \tag{S57}$$

These pseudocount corrections are uniform (i.e. they have the same weight $1/q$ for all states), and their influence relative to the raw empirical frequencies can be tuned through the parameter Λ. In DCA, a high value of f Λ has been found to improve contact prediction: typically $\Lambda \approx 0.5$ [3, 4, 17]. Note that the correspondence of Λ with the parameter λ in Refs. [3, 4, 16] is obtained by setting $\Lambda = \lambda/(\lambda + M)$.

From these quantities, we define the pseudocount-corrected covariances

$$C'_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - f_i(\alpha) f_j(\beta)\,. \tag{S58}$$

We show in Fig. S9 that adding pseudocounts as high as $\Lambda = 0.3$ still allows for accurate extraction of mutational effects (Recovery 0.96) and provides a reliable prediction of sector sites.

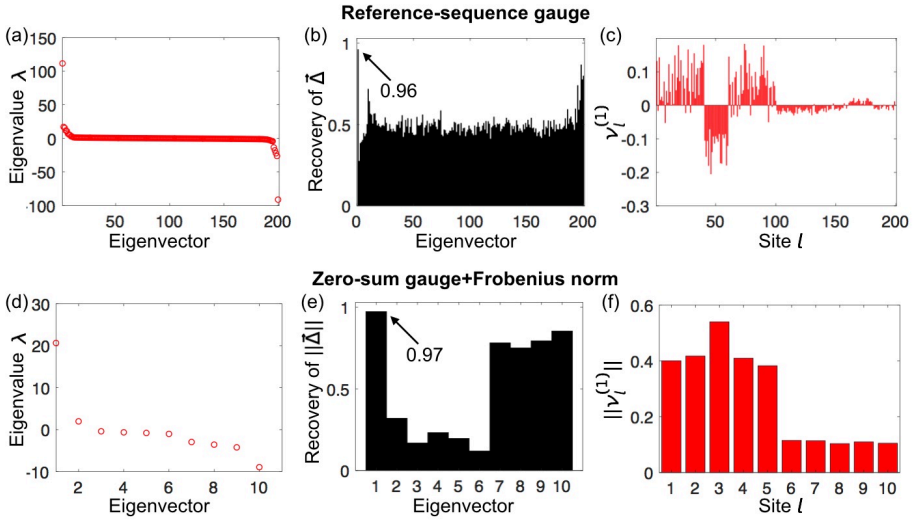

FIG. S9. Effect of pseudocounts on ICOD performance on synthetic sequence data with $q = 21$ possible states at each site. The same synthetic data as in Fig. S7 and S8 is used, but here pseudocounts are employed, with weight $\Lambda = 0.3$. (a-c) Similar analysis as in Fig. S7: (a) Eigenvalues of the ICOD-modified inverse covariance matrix. (b) Recovery of $\vec{\Delta}$ from each eigenvector of the ICOD-modified inverse covariance matrix. (c) First eigenvector of the ICOD-modified inverse covariance matrix. (d-f) Similar analysis as in Fig. S8: (d) Eigenvalues of the compressed ICOD-modified inverse covariance matrix. (e) Recovery of site significances $||\vec{\Delta}||$ from each eigenvector of the compressed ICOD-modified inverse covariance matrix. (f) Estimated site significances computed from the first eigenvector of the compressed ICOD-modified inverse covariance matrix.

V. PERFORMANCE OF SCA

A. Analytical estimates for $\langle S_l\rangle_*$ and $C_{ll'}$ for a single selection with binary sequences

Protein sectors were first discovered from sequence data using a PCA-based method called Statistical Coupling Analysis (SCA) [13, 18]. Interestingly, in SCA, sectors are found from the eigenvectors associated to the largest eigenvalues, while in ICOD they are found from the (modified) eigenvectors associated to the smallest eigenvalues. This difference stems from the fact that SCA and ICOD do not start from the same matrix. For binary sequences, SCA uses the absolute value of a conservation-weighted covariance matrix, $\tilde{C}_{ll'}^{(\mathrm{SCA})} = |\phi_l C_{ll'} \phi_{l'}|$ (see main text and Ref. [18]). When all amino-acid states are accounted for, SCA compresses each block of the conservation-weighted matrix corresponding to two sites to obtain one positive value, e.g. the Frobenius norm of the block [13]. Conversely, ICOD employs the regular covariance matrix, suppressing the diagonal blocks of its inverse at the last step before diagonalization. To better understand the performance of SCA in recovering the site-dependent mutational effects associated with a selective constraint, it is helpful to have analytical estimates for the average mutant fraction $\langle S_l\rangle_*$ at each site l and the covariance matrix $C_{ll'}$ for an ensemble of binary sequences obtained from a single selection using vector of mutational effects $\vec{\Delta}$. To this end, we provide the following two ansatzes:

$$\langle S_l\rangle_* - \langle S_l\rangle \approx (\delta E^* - \langle \delta E\rangle)\frac{\Delta_l}{\sum_l \Delta_l^2}, \tag{S59}$$

$$C_{ll'} \approx \begin{cases} -\frac{\Delta_l \Delta_{l'} \sigma_l^2 \sigma_{l'}^2}{\sum_l \Delta_l^2 \sigma_l^2}, & l \neq l' \\ \sigma_l^2, & l = l', \end{cases} \tag{S60}$$

where $\sigma_l^2 = \langle S_l^2\rangle_* - \langle S_l\rangle_*^2 = \langle S_l\rangle_*(1 - \langle S_l\rangle_*)$ represents the variance of S_l. Recall that $S_l \in \{0, 1\}$, where 0 is the reference state and 1 the mutant state, and that $\langle\cdot\rangle_*$ denotes ensemble averages over the selectively weighted subset of sequences, while $\langle\cdot\rangle$ denotes averages over the unselected (unweighted) ensemble.

Although we have not proven these two ansatzes, numerical tests (Fig. S10) have verified these two relations for ensembles generated from a $\vec{\Delta}$ with multiple sites of comparably large mutational effects so as not to be dominated by a single site, i.e., $\Delta_l/\sqrt{\sum_{l'}\Delta_{l'}^2} \ll 1$ for any l. As a counterexample, the $\vec{\Delta}$ from our elastic network model does not satisfy this condition.

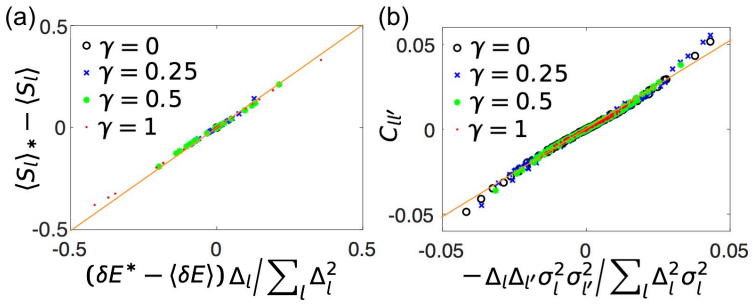

FIG. S10. Numerical verification of the ansatzes in Eq. (S59) and Eq. (S60). We generate a sequence ensemble by considering four values of relative selection bias $\gamma \equiv (\delta E^* - \langle \delta E\rangle)/\sqrt{\langle(\delta E - \langle \delta E\rangle)^2\rangle} = 0, 0.25, 0.5, 1$ and for each case we use a synthetic $\vec{\Delta}$ with a sector size of 20. (a) Numerically computed average bias of the mutant fractions $\langle S_l\rangle_* - \langle S_l\rangle$. Here, $\langle S_l\rangle = 0.5$ for the unselected ensemble. (b) Numerically computed covariances $C_{ll'}$. The results in (a,b) compare well with the analytical predictions (orange lines), provided that $\Delta_l/\sqrt{\sum_l' \Delta_l'^2} \ll 1$ for any l. For each case, 10^6 random sequences were generated to minimize noise from sampling.

B. Analysis of the SCA method

Here, we provide a detailed analysis of the SCA method from Refs. [13, 18]. Following these references, the reweighting factor is chosen to be

$$\phi_l = \frac{\partial D\left(\langle S_l\rangle_*, \langle S_l\rangle\right)}{\partial \langle S_l\rangle_*}, \tag{S61}$$

where, for each site l, $D\left(\langle S_l\rangle_*, \langle S_l\rangle\right)$ is the Kullback-Leibler divergence between the distribution of mutant fractions for the selected sequences and the background distribution:

$$D\left(\langle S_l\rangle_*, \langle S_l\rangle\right) = \langle S_l\rangle_* \log \frac{\langle S_l\rangle_*}{\langle S_l\rangle} + (1 - \langle S_l\rangle_*) \log \frac{1 - \langle S_l\rangle_*}{1 - \langle S_l\rangle}. \tag{S62}$$

In our case, the background distribution is obtained from the unselected sequence ensemble, for which $\langle S_l\rangle = 0.5$. Hence, we have

$$\phi_l = \log\left[\frac{\langle S_l\rangle_* \left(1 - \langle S_l\rangle\right)}{\langle S_l\rangle \left(1 - \langle S_l\rangle_*\right)}\right], \tag{S63}$$

as illustrated in Fig. S11(a). In the regime of relatively weak conservation, i.e. when $\langle S_l\rangle$ is not close to 0 or 1, and $|\langle S_l\rangle_* - \langle S_l\rangle| \ll \langle S_l\rangle$, a first-order expansion yields

$$\phi_l \approx \frac{\langle S_l\rangle_* - \langle S_l\rangle}{\langle S_l\rangle(1 - \langle S_l\rangle)}, \tag{S64}$$

as shown in Fig. S11(b). Employing the ansatz (S59) in this regime, we obtain

$$\phi_l \propto (\delta E^* - \langle \delta E\rangle)\Delta_l. \tag{S65}$$

This relation is verified in Fig. S11(c) for a sequence ensemble generated with a synthetic Δ_l. Hence, the SCA reweighting factor carries information about Δ_l as long as $\delta E^* \neq \langle \delta E\rangle$. In this regime, information about conservation (namely ϕ_l) should thus be sufficient to recover mutational effects and sectors. This was indeed found to be the case for some real proteins with a single sector [19]. However, when the selection bias, $\delta E^* - \langle \delta E\rangle$, is small, random noise due to finite sampling will typically swamp this relationship.

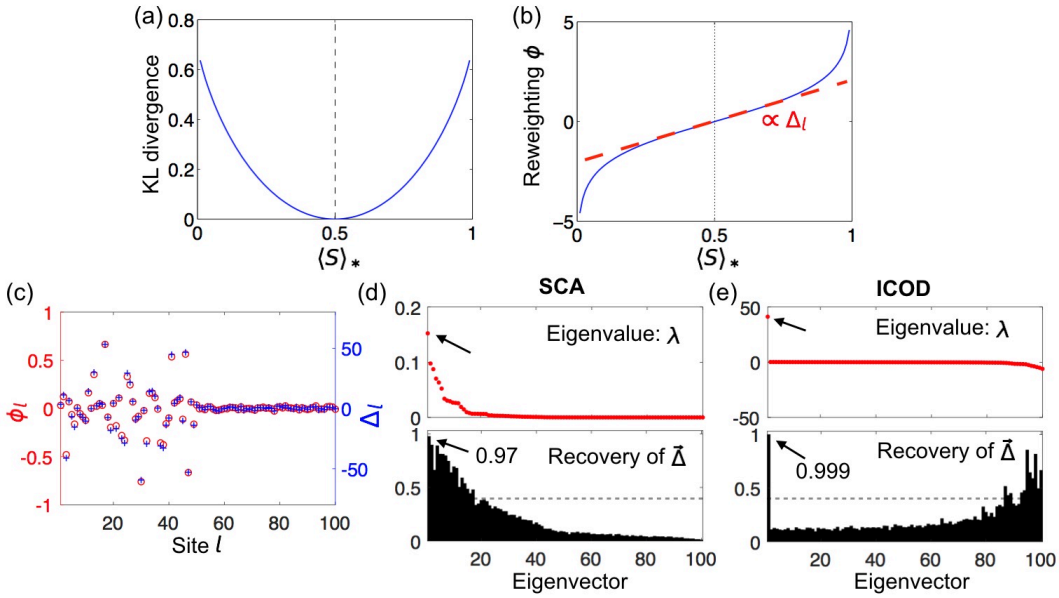

FIG. S11. Underpinnings of Recovery of mutational effect vector $\vec{\Delta}$ by SCA. (a) Kullback-Leibler divergence versus mutant fraction $\langle S \rangle_*$ for background mutant fraction $\langle S \rangle = 0.5$. (b) Reweighting factor ϕ as a function of mutant fraction $\langle S \rangle_*$ for background mutant fraction $\langle S \rangle = 0.5$. (c) Reweighting factor ϕ_l and synthetic Δ_l for an ensemble of sequences generated with a single selection at relative selection bias $\gamma = 1$. $\vec{\Delta}$ was generated with the first 50 sites as sector sites, and 50,000 sequences were employed, as in most of our examples using ICOD (see above). (d-e) Performance of SCA and ICOD for this ensemble, respectively. In computing Recovery using SCA, we use the normalized vector $\sqrt{\nu_l^{(j)}}$ to predict $\vec{\Delta}$. The gray dashed lines in (d) and (e) indicate the random expectation of Recovery [Eq. (S11)].

In Refs. [13, 18], the first eigenvectors of the conservation-reweighted SCA covariance matrix, $\tilde{C}_{ll'}^{(\mathrm{SCA})} = |\phi_l C_{ll'} \phi_{l'}|$, were used to find sectors from sequence data. How does the first eigenvector of $\tilde{C}^{(\mathrm{SCA})}$ relate to the mutational effect vector $\vec{\Delta}$? Utilizing both Eq. (S60) and Eq. (S65), and assuming $\delta E^* \neq \langle \delta E \rangle$, we obtain

$$\tilde{C}_{ll'}^{(\mathrm{SCA})} \propto \begin{cases} \Delta_l^2 \Delta_{l'}^2 \sigma_l^2 \sigma_{l'}^2, & l \neq l' \\ \Delta_l^2 \sigma_l^2, & l = l'. \end{cases} \tag{S66}$$

Apart from the diagonal, the matrix is approximately proportional to the tensor product of $\Delta_l^2 \sigma_l^2$ with itself. If we neglect the contribution from the diagonal elements of $\tilde{C}^{(\mathrm{SCA})}$, the first eigenvector $\vec{\nu}^{(1)}$ satisfies

$$\nu_l^{(1)} \propto \Delta_l^2 \sigma_l^2. \tag{S67}$$

Eq. (S67) explains why $\sqrt{\nu_l^{(1)}}$ carries information about Δ_l. In Fig. S11(d), Recovery using SCA (and Eq. (5) with $\sqrt{\nu_l^{(1)}}$ instead of $\nu_l^{(1)}$) is 0.97, which remains lower than Recovery using ICOD, which is 0.999 here. Besides, Fig. S12 illustrates that Recovery of $\vec{\Delta}$ by SCA is much better using $\sqrt{\nu_l^{(1)}}$ than $\nu_l^{(1)}$.

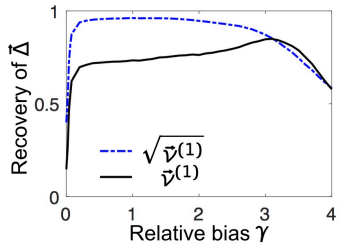

FIG. S12. Recovery of $\vec{\Delta}$ from the first SCA eigenvector using $\vec{\nu}^{(1)}$ or $\sqrt{\vec{\nu}^{(1)}}$. The sequence data are the same as used for the blue curves in Fig. 3(a). As suggested by Eq. (S67), use of the square root of $\vec{\nu}^{(1)}$ significantly improves Recovery.

C. Comparison between ICOD and SCA

In the main text, we compared the performance of ICOD and SCA with respect to Recovery of mutational-effect vectors $\vec{\Delta}$ in synthetic data (see Fig. 3). We found that ICOD performs well over a broader range of relative biases γ than SCA. The failure of SCA at biases close to zero can be explained by the fact that the conservation weights ϕ_l then vanish (see Eq. (S65)). A further example of the failure of SCA for non-biased selections is given by the case studied in Fig. S4, where we considered two selections, a biased one associated to $\vec{\Delta}_1$ and a non-biased one associated to $\vec{\Delta}_2$. Fig. S13 shows that SCA recovers $\vec{\Delta}_1$ well, but performs badly for $\vec{\Delta}_2$, while ICOD recovers both of them very well (see Fig. S4).

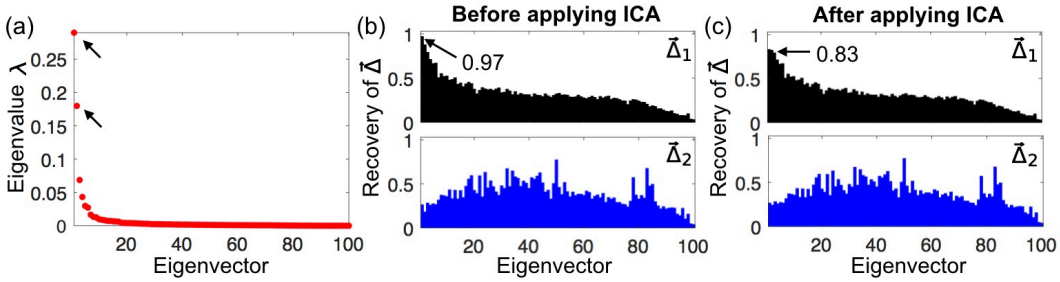

FIG. S13. Performance of SCA for the double selection from Fig. S4. (a) Eigenvalues. (b) Before applying ICA, the first eigenvector has high Recovery of $\vec{\Delta}_1$, but no eigenvector has substantial Recovery of $\vec{\Delta}_2$. This difference matches our observation that SCA performs well for selections of intermediate bias, but not for unbiased selections. (c) Applying ICA on the first two eigenvectors does not improve Recovery.

While the comparison of Recovery favors ICOD, SCA was originally used to identify sectors (in our model, sites with important mutational effects under a given selection) rather than to recover complete mutational effect vectors $\vec{\Delta}$. Hence, in Fig. S14, we compare the ability of ICOD and SCA to predict the n sites with the largest mutational effects. Note that this comparison is independent of whether we use $\vec{\nu}^{(1)}$ or $\sqrt{\vec{\nu}^{(1)}}$ as the predictor in SCA, since the square-root function is increasing and preserves order. Using this criterion, we again find that ICOD performs well over a broad range of relative biases γ, while SCA only works well for sequences selected under moderate biases.

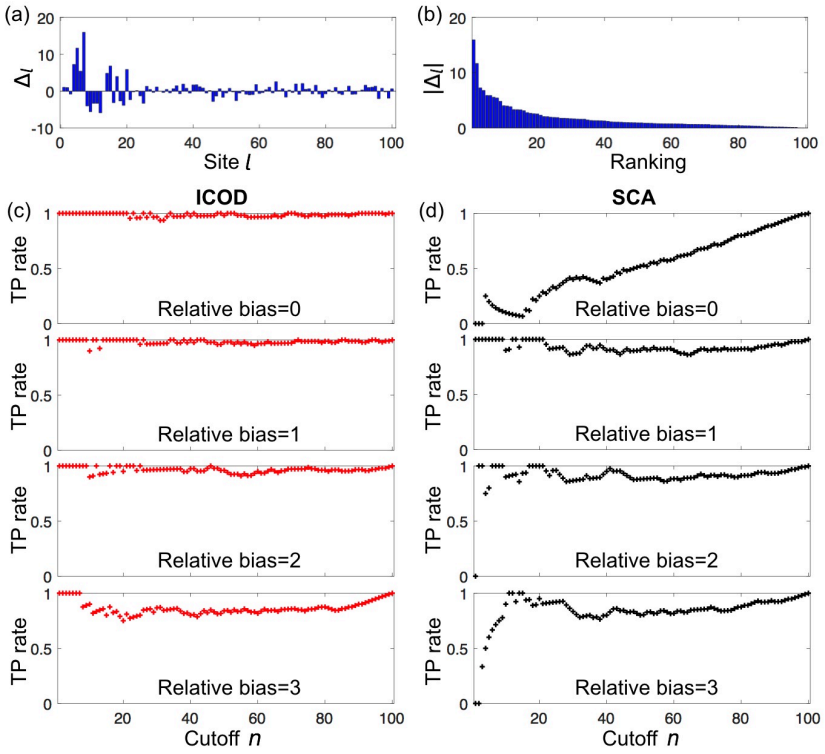

FIG. S14. Comparison of sector-site identification by ICOD and SCA (see also Fig. 3). We use the synthetic $\vec{\Delta}$ in (a) to selectively weight 5,000 random sequences at four relative bias values $\gamma \equiv (\delta E^* - \langle \delta E \rangle)/\sqrt{\langle (\delta E - \langle \delta E \rangle)^2 \rangle} = 0, 1, 2, 3$ and test the ability of ICOD or SCA to correctly predict the sites with the n largest mutational effects. (b) Magnitudes of mutational effects of $\vec{\Delta}$ by rank. (c-d). True Positive (TP) rates obtained by taking the first eigenvector $\vec{v}^{(1)}$ from either ICOD or SCA, generating a ranked list of sites of descending $|\nu_l^{(1)}|$ at each site l, and computing the fraction of the top n sites in this predicted ordering that are also among the top n sites of the actual ordering of mutational effect magnitudes $|\Delta_l|$. The effect of relative bias γ on Recovery is shown in Fig. 3. (c) As expected, the prediction of ICOD is very good under all relative biases. (d) On the other hand, SCA does not perform well at the smallest or largest relative biases.

VI. PERFORMANCE OF A METHOD BASED ON THE GENERALIZED HOPFIELD MODEL

As mentioned in the main text, we also compared ICOD with another PCA-based approach developed in Ref. [6], which employs an inference method specific to the generalized Hopfield model. For L Ising spins ($s_l \in \{-1, 1\}$ for $1 \leq l \leq L$), the Hamiltonian of the generalized Hopfield model reads (see Eq. (3) in Ref. [6])

$$H(\vec{s}) = -\sum_{l=1}^{L} h_l \, s_l - \frac{1}{2L} \sum_{i=1}^{N} \left(\sum_{l=1}^{L} \xi_{i,l} \, s_l \right)^2 + \frac{1}{2L} \sum_{i=1}^{N'} \left(\sum_{l=1}^{L} \xi'_{i,l} \, s_l \right)^2 , \tag{S68}$$

where h_l is the local field at site l, while $\vec{\xi}_i = (\xi_{i,1}, \ldots, \xi_{i,L})$ is an attractive pattern and $\vec{\xi'}_i = (\xi'_{i,1}, \ldots, \xi'_{i,L})$ is a repulsive pattern. Here there are N attractive patterns and N' repulsive ones. In our model, in the single-selection case, the fitness of a sequence $\vec{s}$ in the Ising representation reads [see above, Sec. III A 1, Eq. (S15)]

$$w(\vec{s}) = -\frac{\kappa}{2} \left(\sum_{l=1}^{L} D_l s_l - \alpha \right)^2 = -\frac{\kappa}{2} \left[\left(\sum_{l=1}^{L} D_l s_l \right)^2 - 2\alpha \sum_{l=1}^{L} D_l s_l + \alpha^2 \right] , \tag{S69}$$

with $D_l = \Delta_l/2$ and $\alpha = \delta E^* - \sum_l D_l$. Recalling that fitnesses and Hamiltonians have opposite signs, a comparison of Eqs. (S68) and (S69) shows that $\vec{\Delta}$ plays the part of a repulsive pattern, with the exact correspondence given by $\vec{\xi'} = \vec{\Delta} \sqrt{\kappa L}/2$. Note that in our model the local fields are proportional to the components of $\vec{\Delta}$.

Ref. [6] proposed a method to infer attractive and repulsive patterns from data generated using a generalized Hopfield model Eq. (S68). Introducing the correlation matrix G, which is related to the covariance matrix C through

$$G_{ll'} = \frac{C_{ll'}}{\tilde{\sigma}_l \tilde{\sigma}_{l'}}, \tag{S70}$$

where $\tilde{\sigma}_l^2 = \langle s_l^2 \rangle_* - \langle s_l \rangle_*^2 = 1 - \langle s_l \rangle_*^2$. Ref. [6] found, to lowest order, the following approximation for a single repulsive pattern $\vec{\xi}'$ (see Eq. (9) in Ref. [6]):

$$\xi'_l \approx \sqrt{L\left(\frac{1}{\lambda^{(L)}} - 1\right)}\, \frac{\nu_l^{(L)}}{\tilde{\sigma}_l}, \tag{S71}$$

where $\lambda^{(L)}$ is the smallest (last) eigenvalue of the correlation matrix G and $\nu_l^{(L)}$ is the associated eigenvector. This yields

$$\Delta_l \propto \frac{\nu_l^{(L)}}{\tilde{\sigma}_l}. \tag{S72}$$

Inference of $\vec{\Delta}$ based on Eq. (S72) is referred to as GHI (for Generalized Hopfield Inference) below.

GHI performs very well for the sequence ensembles from the elastic network model used in Fig. 1 and Fig. 2 (Fig. S15). Importantly, just as for simple PCA and for ICOD (see main text), the top Recovery is obtained for the (modified) bottom eigenvector of the covariance matrix, consistently with $\vec{\Delta}$ being a repulsive pattern, but the large-eigenvalue modes also contain some information about $\vec{\Delta}$ (Fig. S15).

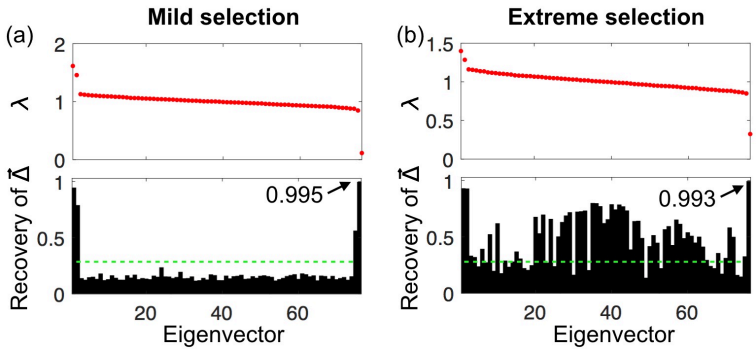

FIG. S15. Performance of GHI on sequence ensembles generated with our elastic-network $\vec{\Delta}$. (a) Eigenvalues of G and Recovery under mild selection bias, as in Fig. 1 in the main text. (b) Eigenvalues of G and Recovery under extreme selection bias, as in Fig. 2 in the main text. The green dashed lines in (a,b) indicate the random expectation of Recovery [Eq. (S11)].

In Fig. S16, we systematically compare all methods discussed in our work to recover $\vec{\Delta}$ from sequence data under various selection biases, using different sector sizes, for selectively weighted ensembles of 50,000 random sequences. We focus on the case of a single selection and compare Recovery of $\vec{\Delta}$ according to:

- ICOD, using the first eigenvector of the modified inverse covariance matrix $\tilde{C}^{-1}$ [see main text, Eq. (7)]
- PCA, using the last principal component of the data (last eigenvector of the covariance matrix, see main text)
- SCA, using the first eigenvector of the absolute value of a conservation-weighted covariance matrix, $\tilde{C}_{ll'}^{(\mathrm{SCA})} = |\phi_l C_{ll'} \phi_{l'}|$ (see main text and Ref. [18])
- GHI, using the reweighted last eigenvector of the correlation matrix [see Eqs. (S70) and (S72)].

Overall, ICOD and GHI perform best. For small selection biases, all methods perform accurately, except SCA, which fails when selection bias vanishes, as explained above. When the sector size is small compared to the sequence length L [Fig. S16 (a-d)], GHI performs a little bit better than ICOD for relatively small selection biases (however Recovery remains $\gtrsim 95\%$ with ICOD). Conversely, GHI is significantly outperformed by ICOD for relatively large selection bias, and the performance of PCA and SCA falls off quite rapidly in this regime. The performances of ICOD, PCA, and GHI become similar when the sector size becomes comparable to the sequence length [Fig. S16 (e, f)].

We further find that GHI is more sensitive to the size of the sequence ensemble than ICOD, although it becomes the most accurate for very large dataset sizes (see Fig. S17). The performance of ICOD is quite robust to dataset size. Note that PCA outperforms other methods when the data size becomes very small (Fig. S17, number of sequences = 500).

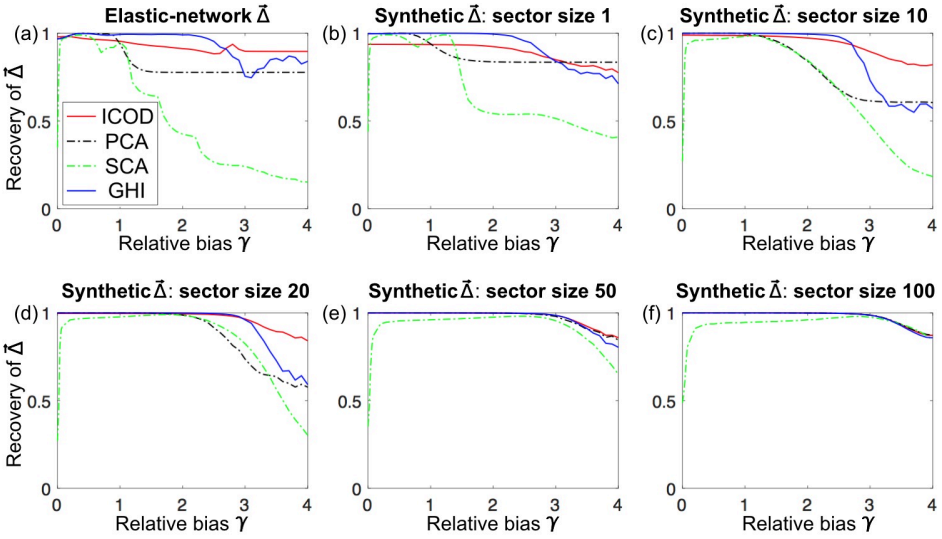

FIG. S16. Comparing Recovery of different methods for various $\vec{\Delta}$s. Here, GHI refers to inference based on Eq. (S72). Curves are obtained by averaging over 100 realizations, each for an ensemble of 50,000 random sequences. For synthetic $\vec{\Delta}$s, each realization corresponds to a new $\vec{\Delta}$.

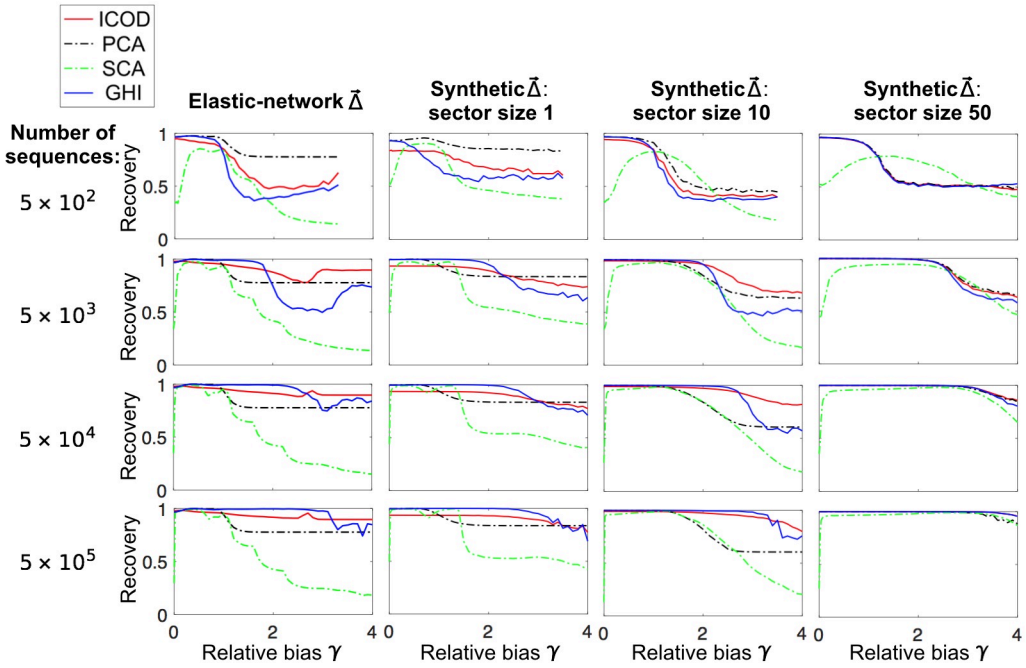

FIG. S17. Effect of dataset size on Recovery of $\vec{\Delta}$. Selectively reweighted ensembles of 5×10^2, 5×10^3, 5×10^4, and 5×10^5 random sequences are generated for the elastic-network $\vec{\Delta}$ and synthetic $\vec{\Delta}$s with sector sizes 1, 10, and 50. All results are averaged over 100 realizations, except those using 5×10^5 sequences, where only 5 realizations were used. For synthetic $\vec{\Delta}$s, each realization employs a different $\vec{\Delta}$ with the same sector size. For the case of 500 sequences, some Recoveries were not computed at high biases due to numerical instabilities.

Overall, we find that GHI is very well suited to infer $\vec{\Delta}$ from very large synthetic datasets. However, ICOD is more robust to variation of dataset size and to selection bias, which should be an advantage in the application to real protein data.

VII. APPLICATION OF ICOD TO A MULTIPLE SEQUENCE ALIGNMENT OF PDZ DOMAINS

Our general physical model for sectors provides insights into the statistical signatures of sectors in sequence data. In particular, we have found that the primary signature of physical sectors lies in the modes associated with the smallest eigenvalues of the covariance matrix, even though there is often additional signal from these sectors in the large eigenvalue modes, as studied more conventionally, e.g. in SCA. The success of ICOD on synthetic data demonstrates that information about sectors can indeed be extracted from the small eigenvalue modes of the covariance matrix.

How well does ICOD perform on real sequence data? Here, we apply ICOD to an actual alignment of sequences of PDZ domains from the Pfam database (`https://pfam.xfam.org/`) containing 24,934 sequences of length $L = 79$. In Ref. [20], sites important for the specific binding of PDZ to peptide ligands were identified experimentally via complete single-site mutagenesis. In particular, 20 sites showing particularly high mutational effects were deemed functionally significant [20]. It was further shown that 15 among the 20 sector amino acids found by SCA (i.e. 75%) were also functionally significant sites.

In order to compute the empirical covariance matrix of the data, we first removed sites with more than 30% gaps (6 sites out of 79). To eliminate the confounding effects of very rare residues at particular sites, we used a pseudocount weight $\Lambda = 0.005$.

Next, we performed both SCA and ICOD using this empirical covariance matrix:

- For SCA, we computed the conservation reweighting factors as in Refs. [13, 18], using the background frequency values from Ref. [18]. We compressed the conservation-reweighted covariance matrix using the Frobenius norm, and we focused on the first eigenvector of this reweighted and compressed covariance matrix in order to predict sector sites (see above, Section V B, and Ref. [13]).
- For ICOD, we used as reference the most abundant state (residue or gap) at each site. We inverted the covariance matrix, set its diagonal to zero, changed the gauge to the zero-sum gauge, and used the Frobenius norm to compress the matrix. Then we focused on the first eigenvector of this reduced ICOD-modified inverse covariance matrix in order to predict site significances and sector sites (see above, Section IV B, especially Fig. S9).

Fig. S18 shows the results obtained by applying SCA and ICOD to our multiple sequence alignment of PDZ domains. In both cases, we find one strong outlying eigenvalue [see Fig. S18(a) and (c)], thus confirming that PDZ has only one sector [20]. Comparing the top n sites identified from the top eigenvector to the 20 experimentally-identified sites with strong mutational effect [20] shows that SCA identifies experimentally important sites somewhat better than ICOD [see Fig. S18(b) and (d)]. For instance, over the $n = 20$ top sites identified by SCA (resp. ICOD), we find that 75% (resp. 55%) of them are also among the 20 experimentally important sites. Note that for SCA, we recover the result from Ref. [20]. Importantly, both methods perform much better than the random expectation, which is $20/73 = 27\%$. Hence, both ICOD and SCA should be useful to identify functionally important sites. The slightly better performance of SCA on this particular dataset might come from the fact that many of the experimentally-identified functionally important sites in PDZ are strongly conserved [19], which makes the SCA conservation reweighting advantageous. Moreover, the real sequences are related by phylogeny, which interacts with functional constraints in ways which require more study. We emphasize that the main goal of this paper is to provide insight into the possible physical origins of sectors, and into the statistical signatures of these physical sectors in sequence data. A more extensive application of ICOD and related methods to real sequence data will be an interesting subject for future work.

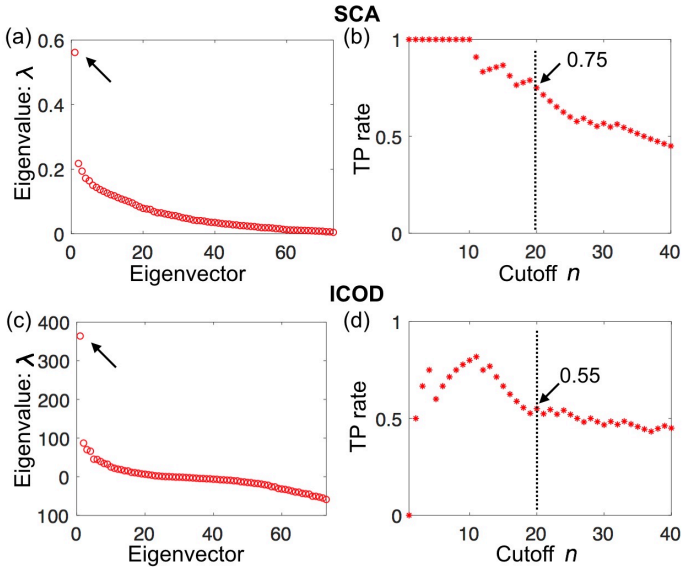

FIG. S18. Performance of SCA and ICOD in predicting the 20 sites with largest experimentally-determined mutational effects [20] in PDZ sequence data. (a) Eigenvalues of the SCA matrix. (b) True Positive (TP) rates obtained by taking the first eigenvector $\vec{\nu}^{(1)}$ from the SCA matrix, generating a ranked list of sites of descending $|\nu_l^{(1)}|$ at each site l, and computing the fraction of the top n sites in this predicted ordering that are also among the top 20 experimentally-important sites [20]. (c) Eigenvalues of the reduced ICOD-modified inverse covariance matrix. (d) TP rates from ICOD, computed as in panel (b) for SCA. In panels (b) and (d), the TP rate values obtained for the top 20 predicted sites are indicated by arrows. A pseudocount ratio $\Lambda = 0.005$ was used throughout this analysis.

[1] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, Annu. Rev. Biophys. **39**, 23 (2010).
[2] P. De Los Rios, F. Cecconi, A. Pretre, G. Dietler, O. Michielin, F. Piazza, and B. Juanico, Biophys. J. **89**, 14 (2005).
[3] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PLoS ONE **6**, e28766 (2011).
[4] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).
[5] T. Plefka, J. Phys. A: Math. Gen. **15**, 1971 (1982).
[6] S. Cocco, R. Monasson, and V. Sessak, Phys. Rev. E **83**, 051123 (2011).
[7] S. Cocco, R. Monasson, and M. Weigt, PLOS Comput. Biol. **9**, e1003176 (2013).
[8] D. J. Thouless, P. W. Anderson, and R. G. Palmer, Philos. Mag. **35**, 593 (1977).
[9] M. Ekeberg, T. Hartonen, and E. Aurell, J. Comput. Phys. **276**, 341 (2014).
[10] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Phys. Rev. E **87**, 012707 (2013).
[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley and Sons, 2001).
[12] L. K. Hansen, J. Larsen, and T. Kolenda, in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, Vol. 5 (2001) pp. 3197–3200.
[13] O. Rivoire, K. A. Reynolds, and R. Ranganathan, PLoS Comput. Biol. **12**, e1004817 (2016).
[14] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, PLoS ONE **9**, e92721 (2014).
[15] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).
[16] A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt, PLoS ONE **6**, e19729 (2011).
[17] A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **113**, 12180 (2016).
[18] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[19] T. Teşileanu, L. J. Colwell, and S. Leibler, PLoS Comput. Biol. **11**, e1004091 (2015).
[20] R. N. McLaughlin Jr, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, Nature **491**, 138 (2012).