Characterizing and inferring epistasis via direct evolutionary couplings in models of allostery

Barbara Bravi[a,1], Riccardo Ravasio[a], Carolina Brito[b], and Matthieu Wyart[a,1]

[a]*Institute of Physics, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

[b]*Instituto de Fìsica, Universidade Federal do Rio Grande do Sul, CP 15051, 91501-970 Porto Alegre RS, Brazil*

[1]To whom correspondence should be addressed. E-mail: barbarabravi@ymail.com, matthieu.wyart@epfl.ch

Abstract

In allosteric proteins, the binding of a ligand modifies function at a distant active site. Such allosteric pathways can be used as target for drug design, generating considerable interest in inferring them from sequence alignment data. Currently, different methods lead to conflicting results, in particular on the existence of long-range evolutionary couplings between distant amino-acids mediating allostery. Here we propose a resolution of this conundrum, by studying epistasis and its inference in models where an allosteric material is evolved *in silico* to perform a mechanical task. We find four types of epistasis (Synergistic, Sign, Antagonistic, Saturation), which can be both short or long-range and have a simple mechanical interpretation. We perform a Direct Coupling Analysis (DCA) and find that DCA predicts well mutation costs but is a rather poor generative model. Strikingly, it can predict short-range epistasis but fails to capture long-range epistasis, in agreement with empirical findings. We propose that such failure is generic when function requires subparts to work in concert. We illustrate this idea with a simple model, which suggests that other methods may be better suited to capture long-range effects.

Author summary

Allostery in proteins is the property of highly specific responses to ligand binding at a distant site. To inform protocols of *de novo* drug design, it is fundamental to understand the impact of mutations on allosteric regulation and whether it can be predicted from evolutionary correlations. In this work we consider allosteric architectures artificially evolved to optimize the cooperativity of binding at allosteric and active site. We first characterize the emergent pattern of epistasis as well as the underlying mechanical phenomena, finding four types of epistasis (Synergistic, Sign, Antagonistic, Saturation), which can be both short or long-range. The numerical evolution of these allosteric architectures allows us to benchmark Direct Coupling Analysis, a method which relies on co-evolution in sequence data to infer direct evolutionary couplings, in connection to allostery. We show that Direct Coupling Analysis predicts quantitatively mutation costs but underestimates strong long-range epistasis. We provide an argument, based on a simplified model, illustrating the reasons for this discrepancy and we propose neural networks as more promising tool to measure epistasis.

Introduction

Allosteric regulation in proteins allows for the control of functional activity by ligand binding at a distal allosteric site [1] and its detection could guide drug design [2, 3]. Yet, understanding the principles responsible for allostery remains a challenge. How random mutations dysregulate allosteric communication is a valuable information studied experimentally [4] and computationally [5]. Several analyses have highlighted the non-additivity of mutational effects or *epistasis*. This "interaction" between mutations can span long-range positional combinations [6], results in either beneficial or detrimental effects to fitness [7], and shapes protein evolutionary paths [8]. Given the combinatorial complexity of its characterization, empirical patterns of epistasis are still rather elusive [9–12]. Concomitantly, progress in sequencing has led to an unprecedented increase of availability of data arranged into Multiple Sequence Alignments (MSAs) [13] containing many realizations of the same protein in related species. Different methods have been developed to extract information from sequence variability, e.g. Statistical Coupling Analysis [14,15] was applied to allostery detection in proteins. It was argued that the allosteric pathway was encoded in spatially extended and connected *sectors*, groups of strongly co-evolving amino-acids, supporting that long-range information on the allosteric pathway is contained in the MSA. Another approach, Direct Couplings Analysis (DCA) [16], aims at inferring evolutionary couplings between amino-acids. Direct couplings predict successfully residue contacts [16] so to inform the discovery of new folds [17], allow one to describe evolutionary fitness landscapes [18, 19] and correlate with epistasis [20, 21]. In the context of allostery, there is no statistical evidence for the existence of long-range direct couplings that would reveal allosteric channels [22], in apparent contradiction with the existence of extended sectors reported in [15] and the observation of long-range epistasis [6].

In this work we propose a solution for this discrepancy, by benchmarking DCA in models of protein

allostery where a material evolves *in silico* to achieve an "allosteric" task [23–29]. We consider recent models incorporating elasticity [24–27, 29], in which long-range co-evolution [26], elongated sectors [26] and long-range epistasis [29] are present and can be interpreted in terms of the propagation of an elastic signal [29]. We focus on materials evolved to optimize cooperative binding over large distances [27], and find four types of epistasis (Synergistic, Sign, Antagonistic, Saturation) that exist over a wide spatial range. We perform DCA and find that it predicts well mutation costs but is a rather poor generative model. Strikingly, it can predict short-range epistasis but fails to capture long-range one, in agreement with empirical findings [22]. We illustrate why it may be so via a simple model, which suggests that neural networks are better suited than DCA to capture long-range effects.

Model for the evolution of allostery

We follow the scheme of [26, 27] where a protein is described by an elastic network of size L made of harmonic springs of unit stiffness (here we consider $L = 12$). Binding events are modeled as imposed displacements either at the "allosteric" or at the "active" site (each consisting of several nodes), as shown in color in Fig. 1A. Such imposed displacements elicit an elastic response in the entire protein and cost some elastic energy, which defines our binding energy (see Sec. 1 in S1 Text). Following [27], the fitness $\mathcal{F}$ measures the cooperativity of binding between allosteric and active site and is defined as the energy difference $\mathcal{F} \equiv E^{\mathcal{A}c} - (E^{\mathcal{A}c,\mathcal{A}l} - E^{\mathcal{A}l})$ where $E^{\mathcal{A}c}$, $E^{\mathcal{A}l}$ and $E^{\mathcal{A}c,\mathcal{A}l}$ are respectively the elastic energy of binding at the active site only ($\mathcal{A}c$), at the allosteric site only ($\mathcal{A}l$) and at both sites simultaneously ($\mathcal{A}c, \mathcal{A}l$). The fitness can be rewritten approximately as (see Sec. 1 in S1 Text)

$$\mathcal{F} \approx \mathbf{F}^{\mathcal{A}c} \cdot \boldsymbol{R}^{\mathcal{A}l \to \mathcal{A}c} \tag{1}$$

where $\mathbf{F}^{\mathcal{A}c}$ is the force field imparted by substrate binding on the nodes of the active site, and $\boldsymbol{R}^{\mathcal{A}l \to \mathcal{A}c}$ is the displacement field induced at the active site by ligand binding. Note that each field in Eq. 1 is of dimension $n_0 d$, where $n_0 = 4$ is the number of nodes in the active site and $d = 2$ the spatial dimension.

Such networks are evolved by changing the position of springs according to a Metropolis-Monte Carlo routine to maximize $\mathcal{F}$. At each step, the fitness difference with respect to the previous configuration $\Delta\mathcal{F}$ is computed and the new configuration is accepted with a probability $p = \min(1, \exp \beta\Delta\mathcal{F})$. β is an evolution inverse temperature controlling the selection pressure for high fitness $\mathcal{F}$, we choose $\beta = 10^4$. We sample every 1000 time steps after an initial equilibration time of 10^5 steps. At long times one obtains a cooperative system of typical $\mathcal{F} \sim 0.2$, whose architecture depends on the spatial dimension and boundary conditions [27]. Here we consider a network in $d = 2$ dimensions with periodic boundaries, equivalent to a cylindrical geometry, where the response to binding evolves towards a *shear* mode. With our scheme we can generate thousands of networks with a similar design. A sequence $\boldsymbol{\sigma}$ of 0 and 1, where $\sigma_i = 1$ stands for the presence of a spring at link i and $\sigma_i = 0$ for its absence, can be associated to any network, leading to a Multiple Sequence Alignment (MSA) of networks performing the same function

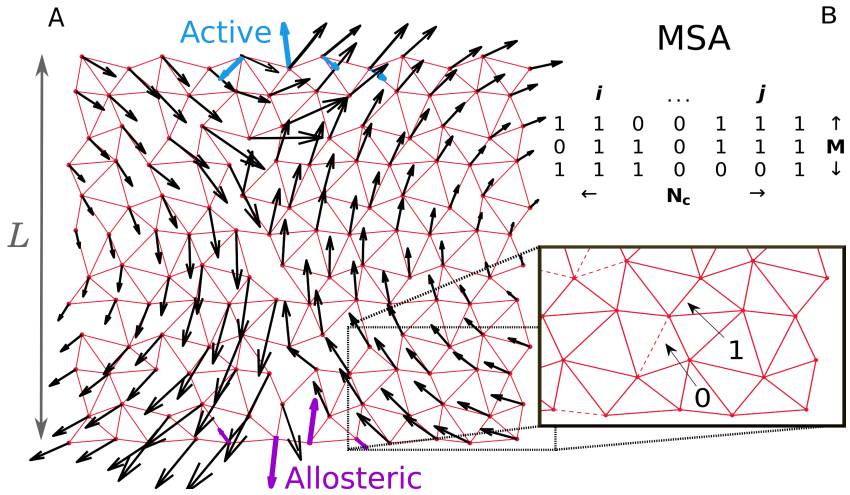

Figure 1: **Study of co-evolution in artificial allosteric networks**. A: Example of an elastic network made of harmonic springs (red) evolved *in silico* to maximize the cooperativity between the allosteric site (purple) and the active site (blue). The response to binding at the allosteric site is indicated by black arrows, and is found to follow a shear motion. B: Each network corresponds to a sequence of 0 and 1 coding for the spring absence or presence. Our scheme allows us to generate a large number M of such sequences, each corresponding to a slightly different shear architecture.

(see Fig. 1B).

Results

Nature and classification of epistasis

The cost of a single mutation (i.e. changing the occupancy) at some link i is defined as $\Delta\mathcal{F}_i = \mathcal{F} - \mathcal{F}_i$ where $\mathcal{F}$ is the original fitness and $\mathcal{F}_i$ the one of the network after the mutation. We denote by $\Delta\mathcal{F}_{ij} = \mathcal{F} - \mathcal{F}_{ij}$ the cost of a double mutation at i and j. Epistasis between loci i and j is then defined as $\Delta\Delta\mathcal{F}_{ij} \equiv \Delta\mathcal{F}_{ij} - \Delta\mathcal{F}_i - \Delta\mathcal{F}_j$. Following Eq. 1 and observing that a mutation mostly affects the propagation of the signal $\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}$ and not how binding locally generates force (see Sec. 1 in S1 Text), epistasis follows approximately

$$\Delta\Delta\mathcal{F}_{ij} \approx -\mathbf{F}^{\mathcal{A}c} \cdot \left(\delta\boldsymbol{R}_{ij}^{\mathcal{A}l\to\mathcal{A}c} - \delta\boldsymbol{R}_{i}^{\mathcal{A}l\to\mathcal{A}c} - \delta\boldsymbol{R}_{j}^{\mathcal{A}l\to\mathcal{A}c}\right)$$

where $\delta\mathbf{R}_i^{\mathcal{A}l\to\mathcal{A}c} = \mathbf{R}_i^{\mathcal{A}l\to\mathcal{A}c} - \mathbf{R}^{\mathcal{A}l\to\mathcal{A}c}$, and $\mathbf{R}_i^{\mathcal{A}l\to\mathcal{A}c}$ is the allosteric response at the active site of the protein mutated at link i. $\delta\mathbf{R}_j^{\mathcal{A}l\to\mathcal{A}c}$ and $\delta\mathbf{R}_{ij}^{\mathcal{A}l\to\mathcal{A}c}$ follow analogous definitions. We denote by θ the angle between $\delta\mathbf{R}_i^{\mathcal{A}l\to\mathcal{A}c}$ and $\delta\mathbf{R}_j^{\mathcal{A}l\to\mathcal{A}c}$. Assuming that the cost of a double mutation is dominated by the strongest point mutation, i.e. $\Delta\mathcal{F}_{ij} \approx \max(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j)$ leads to

$$\Delta\Delta\mathcal{F}_{ij} \approx -\min(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j). \tag{2}$$

This assumption does capture a significant part of epistasis, especially when it is strong, as shown in Fig. 2A. This observation suggests to classify pairs of loci in terms of their epistasis and the minimal associated mutation cost $\min(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j)$ as performed in Fig. 2A.

Saturation: We define (somewhat arbitrarily) mutations with $\Delta\mathcal{F} > 0.1$ as lethal. Pairs of such lethal mutations (which represent $\sim 0.1\%$ of all pairs, a sparsity in line with experimental findings [21]) have the strongest epistasis in absolute value, and follow closely Eq. 2, as visible in Fig. 2A. Physically, these mutations essentially shut down signal propagation by themselves with $\mathbf{R}_i^{\mathcal{A}l\to\mathcal{A}c} \approx \mathbf{R}_j^{\mathcal{A}l\to\mathcal{A}c} \approx 0$, in such a way that the double mutation has the effect of a single one with $\mathbf{R}_{ij}^{\mathcal{A}l\to\mathcal{A}c} \approx 0$. This view is confirmed in Fig. 2B by the observation that $\cos(\theta) \approx 1$, as follows from $\delta\mathbf{R}_i^{\mathcal{A}l\to\mathcal{A}c} \approx \delta\mathbf{R}_j^{\mathcal{A}l\to\mathcal{A}c} \approx -\mathbf{R}^{\mathcal{A}l\to\mathcal{A}c}$. Saturation is then a form of very high "diminishing-returns" epistasis, for which evidence from data and support from theoretical models are accumulating [30, 31].

Antagonistic. Further up along the diagonal of Eq. 2 in Fig. 2A, this saturation effect becomes milder. It is more akin to "antagonistic" epistasis [7, 32], whereby, after a first mutation, making a second one results only in a weak additional change.

Sign. In the intermediate range of negative-sign epistasis, more compensatory epistatic interactions can take place, where the fitness cost of a deleterious mutation is diminished by the second mutation (i.e. $\Delta\mathcal{F}_{ij} < \max(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j)$). Thus some mutations can become beneficial (i.e. increase the fitness) in presence of another mutation, and this resembles the "sign" epistasis empirically detected [7,33]. Geometrically, it corresponds to situations where the two mutations deform the signal in opposite directions, so the second one can partially re-establish fitness. In support of this, Fig. 2B shows that for sign epistasis $\cos(\theta)$ tends to be negative.

Synergistic. Positive-sign values indicate "synergistic" epistasis. It occurs if two mutations perturb the elastic signal in the same direction, causing more damage than expected if they were purely additive. As clear from Fig. 2B, $\cos(\theta)$ tends to be positive in this case.

Direct Coupling Analysis

We evolve numerically M configurations maximizing cooperativity $\mathcal{F}$, each yielding a realization of a (variable) shear design. We sample a configuration for every initial condition to avoid introducing a bias in the sampling due to their high similarity. We find that the average Hamming distance among the obtained sequences is $\sim 20\%$ of their length. Our set of sequences is analogous to a protein MSA – importantly, in this analogy the role of an amino-acid is played by a link, which can be stiff ($\sigma_i = 1$) or not ($\sigma_i = 0$, no springs). In practice we take $M = 135000$, much larger than the sequence length $N_c = (3L^2 - 2L) = 408$.

Next, for a statistical analysis of these sequences, we use DCA, which is based on the idea of fitting the observed single-site $\langle\sigma_i\rangle = 1/M \sum_m \sigma_i^m$ and pairwise $\langle\sigma_i\sigma_j\rangle = 1/M \sum_m \sigma_i^m \sigma_j^m$ frequencies of links by the probability distribution $P(\boldsymbol{\sigma})$ with maximal entropy (as this ensures the least biased fit of data under such empirical constraints). In our setup this approach leads to

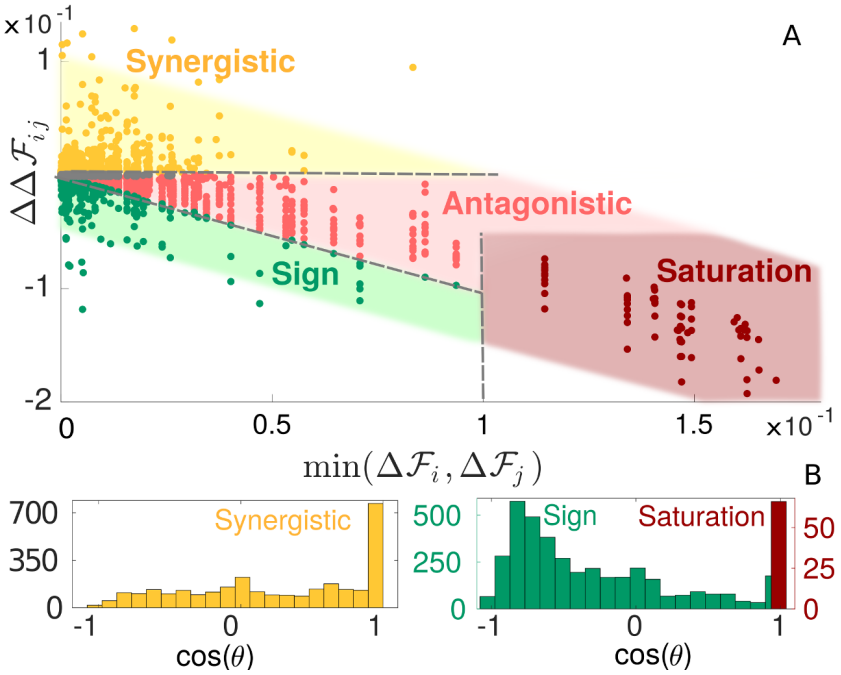

Figure 2: **Classification and mechanical characterization of epistasis in our model of allosteric cooperativity.** A: Phase diagram of epistasis in our allosteric material. All quantities are averages over 50 configurations obtained in a single run. The shaded area is taken with arbitrary width and a -1 slope as a guide to the eye. We show the lines $\Delta\mathcal{F}_{ij} = 0$, which divides synergistic from antagonistic/sign epistasis, $\Delta\mathcal{F}_{ij} = \max(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j)$, separating sign and antagonistic epistasis, and $\min(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j) = 0.1$, the threshold set to distinguish lethal mutations. Points in grey correspond to epistasis $< 5 \times 10^{-4}$ and are excluded from our analysis. B: Histograms of $cos(\theta)$ for synergistic, sign and saturation epistasis.

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(-\mathcal{E}(\boldsymbol{\sigma})\right) \quad (3)$$

$$\mathcal{E}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \quad (4)$$

which is equivalent to an Ising model where $\sigma_i = 0, 1$ would denote the two states (down, up) of spins. In this setting, $\mathcal{E}$ is an estimation of $\beta\mathcal{F}$, β being the inverse evolution temperature. In all the comparisons (e.g. Fig. 3) we omit β as we are interested in the proportionality between $\mathcal{E}$ and $\mathcal{F}$. The "fields" h_i and "couplings" J_{ij} are inferred to match $\langle\sigma_i\rangle$ and $\langle\sigma_i\sigma_j\rangle$. The inference of these parameters can be performed with several algorithms, we focus on ACE (Adaptive Cluster Expansion) [34, 35], an approximate technique developed from statistical physics ideas, combined with maximum likelihood, an exact technique. This approach is extremely accurate and we compare it to a method more approximate, but much faster computationally, as mean field Direct Coupling Analysis (mfDCA) [16], see Methods for details on the implementation.

In this way we can benchmark DCA in the context of allosteric materials and test if it: (i) reproduces accurately the cost of single mutations; (ii) is a good generative model, i.e. if it can generate new sequences with high fitness and (iii) can predict epistasis.

Inferring mutation costs

Fig. 3A shows the map of true mutation costs, indicating a large cost near the allosteric and active sites as well as in the central region where the allosteric response displays high shear (as documented in [27]). DCA enables one to infer this map by computing the estimated mutation cost $\Delta\mathcal{E}_i = \mathcal{E}_i - \mathcal{E}$ for a mutation at a generic link i, Fig. 3B. The comparison is excellent, as evident also from the high correlation revealed by the scatter plot Fig. 3C. Importantly, including pairwise couplings is key for inferring mutation costs, as a model based on conservation alone performs poorly, see inset of Fig. 3C.

Generative power of DCA

Once the model of Eqs. 3, 4 is inferred, can it be used to generate new sequences with a high fitness, as previously shown for models of protein folding [36]? To answer this question, we generate new sequences by Monte Carlo sampling from the probability distribution Eq. 3. Fig. 4 shows the fitness of the obtained sequences vs their distance to "consensus" - the consensus being the most representative sequence of the MSA, i.e. where springs occupy the positions with largest mean occupancy. We find that (i) the variability of the MSA, quantified by the distance to consensus, is well reproduced (ii) the fitness is much more variable than for random sequences, with a few sequences that do perform as well as evolved ones (which never occurs for random sequences) but (iii) the mean obtained fitness is rather low, although larger, in a statistically significant way, than the one of random configurations (which is zero). As shown in Fig. 4, these results deteriorate further if a more approximate algorithm as mfDCA is used to infer parameters. We have checked that the generative performance is not improved by lowering the temperature of the

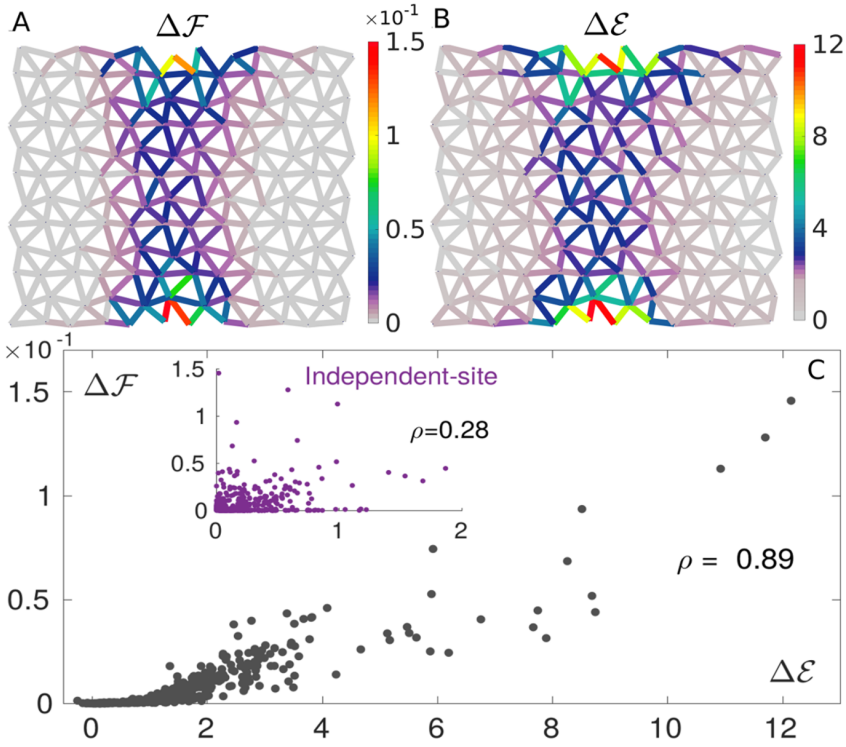

Figure 3: **Prediction of mutation costs by DCA**. Maps of true $\Delta\mathcal{F}$ (A) and DCA-inferred $\Delta\mathcal{E}$ (B) single mutation costs, averaged over 1.5×10^3 configurations randomly chosen from the MSA. Their patterns are very similar, revealing high costs near the allosteric and active sites and in the shear path connecting them. C: Scatter plot showing the strong correlation between $\Delta\mathcal{F}$ and $\Delta\mathcal{E}$ for all links. The estimation of mutation costs based on an independent-site model (i.e. on conservation) correlates poorly with the true cost (inset), proving the need for incorporating correlations for proper prediction of mutation costs.

Monte Carlo sampling. Overall, these results suggest that the generative power of DCA is limited in the context of allostery, in contrast with results for models of protein folding [36]. Thus an Ising model, a quadratic model accounting for conservation and correlations in the MSA (first and second order statistics), although it can capture some features of the shear design (e.g. the inhomogeneous distribution of coordination, as shown in Fig. S2), is a rather drastic approximation for the initial allosteric fitness. Indeed we have tested that higher orders as the third moment are not well reproduced (see Fig. S1). In what follows we shall emphasize in particular the failure of DCA to infer long-range epistasis.

Inferring epistasis with DCA

From Eq. 4 one readily has that the DCA prediction for epistasis follows $\Delta\Delta\mathcal{E}_{ij} = -J_{ij}(2\sigma_i - 1)(2\sigma_j - 1)$, implying $|\Delta\Delta\mathcal{E}_{ij}| = |J_{ij}|$. Hence, within DCA, the epistasis magnitude is simply the one of evolutionary couplings. In the inset of Fig. 5A we show the spatial location of the top 400 pairs of links with highest coupling magnitude, illustrating that long-range couplings are rare. Yet, as implied jointly by Fig. 2A

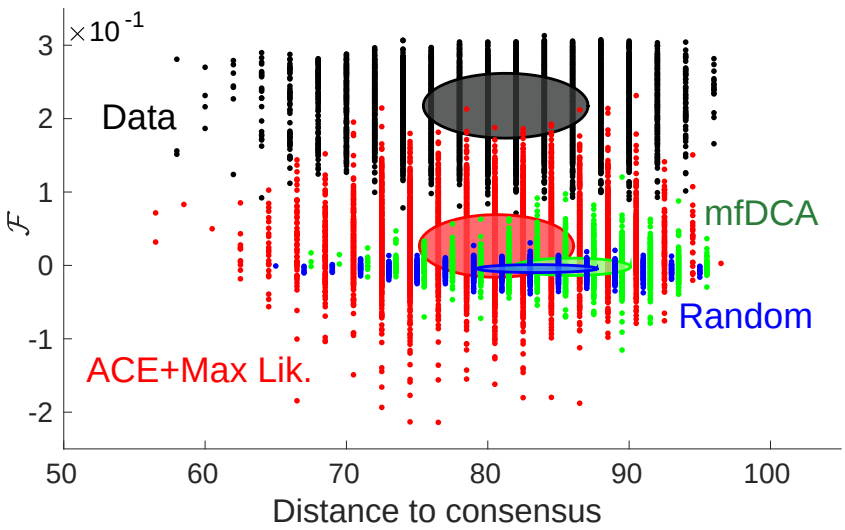

Figure 4: **Generative performance of DCA**. Fitness vs distance to consensus of configurations generated by the inferred model, following the representation of [36]. The sampling is done from $P(\boldsymbol{\sigma})$ of Eq. 4 (a Boltzmann-Gibbs probability distribution), whose parameters have been inferred via ACE + maximum likelihood (red cloud) or mfDCA (green cloud). Original high fitness configurations (black cloud) and random ones (blue) are added as a reference. Each cloud consists of 10^4 sequences and the drawn ellipse gives one standard deviation around the mean in both horizontal and vertical directions. Distances to consensus of ACE + maximum likelihood, mfDCA and random sequences are shifted by respectively $+0.7$, -0.7 and -1.3 for better visibility.

and Fig. 3A, long range epistasis is present in our model, meaning that DCA fails to capture it. This fact is demonstrated quantitatively in Fig. 5A showing the mean epistasis $|\Delta\Delta\mathcal{F}_{ij}|$ and mean DCA prediction $|\Delta\Delta\mathcal{E}_{ij}|$ as a function of distances. The DCA-predicted trend reproduces the original one at small distances but strongly underestimates long-range epistasis. This is further evidenced in Fig. 5B showing that the average fraction of long-range pairs (range > 7) with the largest epistasis which falls in the list of the 400 pairs with largest couplings is much smaller than for short-distance pairs (< 7). However, even at short distance the prediction by $|J_{ij}|$ is not excellent, but it is remarkably improved if, as done in [12, 21], one considers epistasis averaged over several configurations (see Sec. 2 in S1 Text). Our finding is consistent with the lack of empirical evidence for long-range inferred couplings in allosteric proteins [22].

A proposed explanation for the failure of DCA at long-distances

We propose that the failure of DCA at long-range stems from its inability to describe a function that requires many subparts of the system to work in concert, when each subpart can be of different type. For example, in allosteric proteins on short length scales soft regions must exist where shear propagates [27, 37], giving rise to local constraints. Yet, there is flexibility in the exact location of these soft regions. On a larger length scale, these regions must assemble to create an extended soft elastic mode [27, 38, 39],

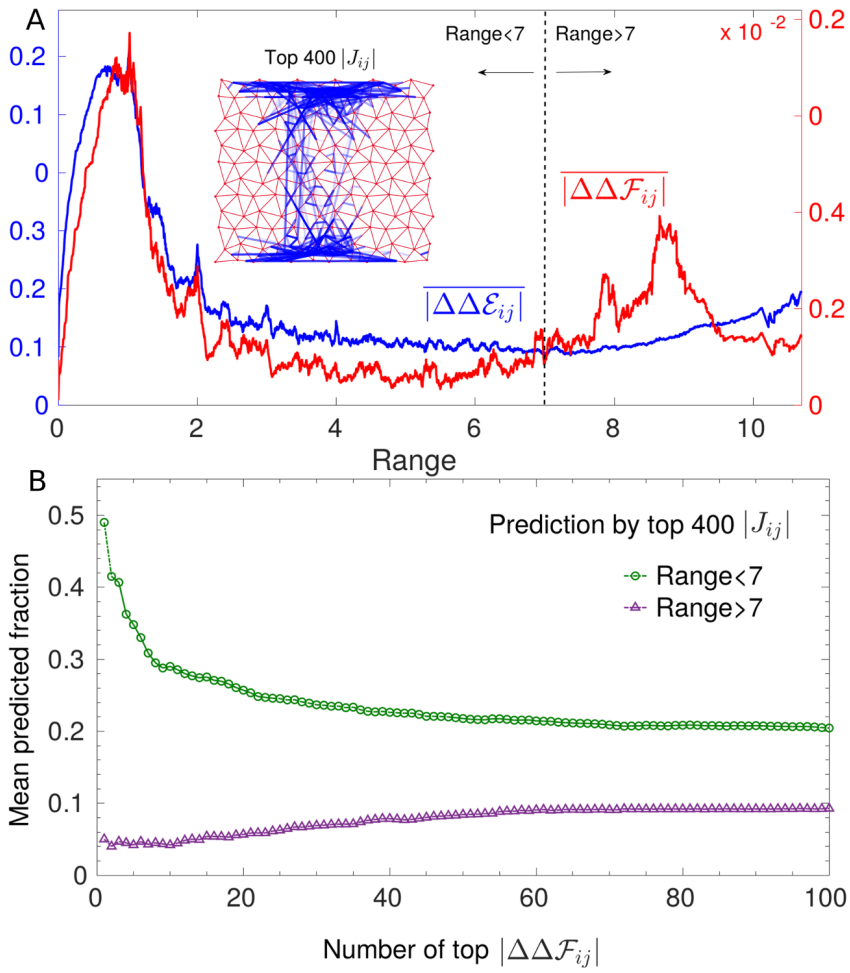

Figure 5: **Prediction of epistasis by DCA**. A: Running average of the absolute value of epistasis $\Delta\Delta\mathcal{F}_{ij}$ and of DCA prediction $\Delta\Delta\mathcal{E}_{ij}$ for 1.5×10^3 configurations as a function of the distance between link i and j. The trends are nearly identical at short distances but at long distance DCA underestimates epistasis. Inset: Top 400 inferred couplings. They are mostly short range with only a few long-range couplings connecting the allosteric and the active site. Next we assess the prediction of epistasis in single configurations by these top 400 couplings. We consider separately long-range (> 7) and short-range (< 7) pairs of links, and rank them respectively in terms of the epistasis magnitude $|\Delta\Delta\mathcal{F}_{ij}|$. B shows which fraction of these pairs - averaged over 100 configurations randomly chosen - belongs to the 400 largest couplings, as a function of the number of pairs with maximal epistasis considered. Clearly coupling magnitude has less predictive power at large distances than at short ones. This feature stays robust also if we increase, e.g. up to 1000, the number of top couplings for prediction (see Fig. S4A).

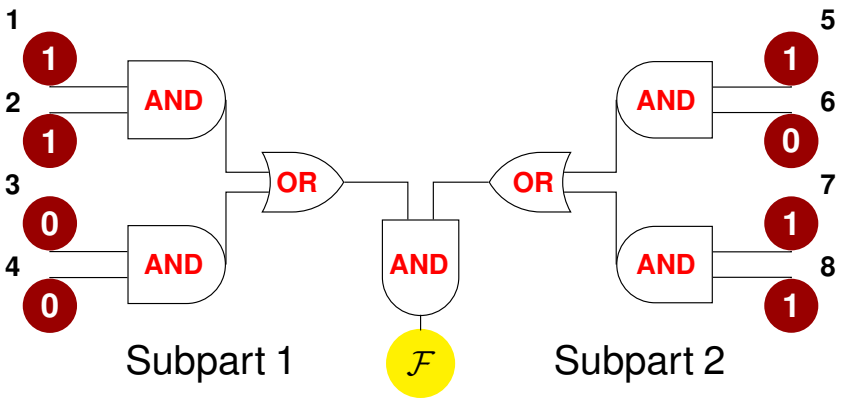

Figure 6: **Sketch of a simple model for protein function**. A system is arranged into 2 subparts which must work jointly to accomplish a given function (AND gate). Each subpart can be of 2 types (OR gate), to work each type must satisfy some constraints (AND gate between single units).

which generates global constraints: for the shear architectures it implies the presence of a soft path between the allosteric and active site, whose position however can fluctuate. We argue that when applied to systems whose function is organized in such a hierarchical way, DCA underestimates long-range constraints. To illustrate this point, we introduce a Boolean model, shown in Fig. 6. A generic "function" is achieved by two subparts that must work in concert (AND gate) and that can be of two different types (OR gate) but each must be functional (AND gate). This model comprises 8 units, taking the value 0 or 1, decomposed into 4 groups: 2 groups are the possible types of subpart 1 (left in Fig. 6) and the other 2 the possible types of subpart 2 (right). A configuration is "functional" if 2 units of the same group are simultaneously in state 1 for each subpart. There are 49 functional configurations, whose fitness is fixed to $\mathcal{F}$, all other configurations have fitness 0. We assume that $\mathcal{F}$ is large in such a way that the sequences in the MSA are only the 49 functional ones, with a uniform distribution. It is straightforward to calculate epistasis in this model, as well as single-site and pairwise frequencies from which couplings J_{ij} and fields h_i can be inferred. In particular we can compare $\Delta\Delta\mathcal{F}_{ij}$ and $\Delta\Delta\mathcal{E}_{ij}$ for units i and j either in the same group, so locally constrained by function (at "short distance", e.g. $i = 1$ and $j = 2$), or in the two different subparts, thus globally constrained (at "long distance" e.g. $i = 1$ and $j = 5$). We obtain (see Sec. 2.1 in S1 Text) that $|\Delta\Delta\mathcal{F}_{12}|/|\Delta\Delta\mathcal{F}_{15}| \approx 2.3$: global and local constraints lead to relatively similar short range and long-range epistasis. Yet we find that epistasis between subparts is noticeably underestimated in contrast to epistasis within subparts. To show this, we look at the DCA prediction for the ratio of epistasis between two pairs of sites divided by the true ratio of epistasis. For pairs of sites belonging to the same subpart, DCA predicts equally well epistasis. For example, considering the pair of sites (1,2) and the pair (1,3), one finds $|\Delta\Delta\mathcal{E}_{13}|/|\Delta\Delta\mathcal{E}_{12}| \times |\Delta\Delta\mathcal{F}_{12}|/|\Delta\Delta\mathcal{F}_{13}| \approx 0.86$ which is close to unity. However if sites belong to different subparts, DCA strongly underestimates epistasis with $|\Delta\Delta\mathcal{E}_{15}|/|\Delta\Delta\mathcal{E}_{12}| \times |\Delta\Delta\mathcal{F}_{12}|/|\Delta\Delta\mathcal{F}_{15}| \approx 0.33$.

Discussion

We have benchmarked DCA in a model of protein allostery where a mechanical task must be achieved over long distances. Such models display a rich pattern of epistasis, which can be both short and long-range and vary in sign. DCA predicts well mutation costs but is not a good generative model. This failure echoes with the drastic underestimation of long-range epistasis by the pairwise couplings inferred by DCA from evolutionary correlations. This finding rationalizes why there is no statistical evidence for long-range couplings in allosteric proteins analyzed by DCA [22], where long-range epistasis and functional effects are however found [6, 12, 15].

Yet, as we show in S1 Text (see Sec. 2), we expect that DCA can capture some aspects of the long-range epistasis pattern in allosteric proteins. Indeed, high-cost mutations exhibit stronger epistasis than low-cost ones (as also seen in RNA sequences [33, 40], in the enzyme TEM-1 β-lactamase [11] and in previous *in silico* evolution work [29]), and are well-predicted by DCA. Testing this DCA prediction for epistasis patterns empirically could be made possible by the increasing availability of deep mutational scans [12, 41].

Finally, we have provided the more general argument, illustrated by a simple model, that a co-evolution based maximum-entropy approach as DCA is not the appropriate inference framework when function requires several, variable parts to work in concert. Can one find better generative models than DCA for such complex functions? Several ways have been proposed to go beyond pairwise models by including nonlinearities, which implicitly take into account correlations at all orders, as nonlinear potentials in Restricted Boltzmann Machines [42], maximum-entropy probability measures with a nonlinear function of the energy [43] or maximum-likelihood inference procedures based on nonlinear functions [44]. As a first test, we have trained a 3-layers feedforward neural network with nonlinear (sigmoid) activation functions to learn the values of fitness in the simple model of Fig. 6. On the validation set, we could reach average mean squared errors on the estimated fitness $\sim 10^{-6} - 10^{-7}$, hence mutation costs and epistasis are correctly captured by this method (see Sec. 2.1.1 in S1 Text). This observation raises the possibility that neural networks may lead to better generative models in proteins, a hypothesis that could also be benchmarked *in silico*.

Methods

Direct Coupling Analysis: inference procedure

In a maximum-entropy approach, extracting information from MSAs can be cast as an inverse problem, i.e. inferring the set of parameters which enable the model (an Ising model in our setup) to reproduce certain observed statistical properties [45, 46]. The exact solution of this problem is found by Maximum Likelihood algorithms, which search for the set of couplings J_{ij} and fields h_i maximizing the likelihood

that the model specified by such parameters produced data with the given statistics (single-site and pairwise frequencies in our case). This exact maximization might often be infeasible, therefore to tackle the inverse problem approximate techniques have been developed: for instance, we resort to the Adaptive Cluster Expansion (ACE), an expansion of the entropy (which indeed corresponds to the likelihood) into contributions from clusters of spins [34, 35, 47]. We use the package made available by Barton https://github.com/johnbarton/ACE. The implementation consists of first a run of ACE followed by a proper maximum likelihood refinement (QLS routine), which takes as starting set of fields and couplings the ACE-inferred ones. Different parameters for the ACE and QLS routines can be set by the user, e.g. γ_2, the L_2-norm regularization strength for couplings which penalizes spurious large absolute values induced by undersampling and for which a natural value is $\gamma_2 = 1/M$ (M being the size of the sample). To help convergence, we have chosen for ACE a higher value $\gamma_2 = 10^{-2}$ and $\theta = 10^{-5}$ (this is the threshold at which the algorithm will run then exit, see [35]). In the further refinement by QLS, we have set *mcb*, the number of Monte Carlo steps used to estimate the inference error, to 200000 and $\gamma_2 = 1/M$. Having full control of the numerical evolution, we have tried to avoid undersampling issues by generating a large number of configurations $M = 135000$, which leads to $\gamma_2 \approx 0.7 \times 10^{-5}$. For the inference we remove from sequences the 6 links at the active and allosteric sites as they are always associated to the symbol 1 (always occupied by a spring), so the number of parameters to infer is $N_c' + N_c'(N_c' - 1)/2 \sim 81000$ with $N_c' = N_c - 6 = 402$. We have verified that low values of the L_2-regularization allow us to obtain the maximal generative performance compatible with the model (in comparison to higher regularization). By default the L_2 regularization of fields is $0.01 \times \gamma_2$. In Fig. S1A, it is shown that the result of the inference is a model perfectly able to reproduce the first and second order statistics (as it should by construction) but that fails at reproducing higher order statistics.

For a comparison, we have considered also mean field Direct Coupling Analysis (mfDCA) [16], derived from a mean-field factorized ansatz for the Boltzmann-Gibbs distribution Eq. 3. Couplings in mfDCA are given by $J_{ij} = -(\boldsymbol{C}^{-1})_{ij}$, where $\boldsymbol{C}_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$ is the covariance of the MSA (we recall that in each sequence $\sigma_i = 1$ stands for the presence of a spring at link i and $\sigma_i = 0$ for its absence). Typically $\boldsymbol{C}$ is not invertible due to undersampling, making it necessary to add a pseudocount λ (see [48]). As shown in [49], a pseudocount also helps correct for the systematic biases introduced by the mean field approximation: for this reason, we have used a pseudocount λ and chosen its value as $\lambda = 0.5$, which allows the best comparison to the ACE and maximum likelihood results, see Fig. S1B. It is noteworthy that in this way a computationally cheap technique as mfDCA yields a pattern of top J_{ij} strikingly similar to the one of a very accurate inference achieved by the combination of ACE and maximum likelihood. Therefore mfDCA, while extremely poor as a generative model, exhibits a good performance at reconstructing the distribution of relevant couplings, as shown in Fig. S1C.

Mutation costs and generative performance in the inferred Ising model

Costs of double mutations, i.e. joint mutations affecting links i and j, can be computed in the original model via fitness changes $\Delta\mathcal{F}_{ij} = \mathcal{F} - \mathcal{F}_{ij}$, where $\mathcal{F}_{ij}$ is the fitness after springs in i and j have been mutated. A double mutation can correspond either to (i) adding two springs at links i and j (i.e. $\sigma_i = \sigma_j = 1$) or removing them (i.e. $\sigma_i = \sigma_j = 0$) or to (ii) moving a spring from link i to link j or viceversa (i.e. $\sigma_i = 0$, $\sigma_j = 1$ or $\sigma_i = 1$, $\sigma_j = 0$). Let us call the former "non-swap" mutations and the latter "swap" mutations. Swap mutations conserve the total amount of springs (360), thus the overall average coordination $\langle z \rangle = 5$, and are the ones performed in the *in silico* evolution. As optimal allosteric configurations maximize fitness with respect to this type of mutations, we stick to them also when we compare mutation costs in terms of fitness and inferred energy (see Fig. 3C): we define "effective" single mutation costs $\Delta\mathcal{F}_i$ and $\Delta\mathcal{E}_i$ by taking, for each link, the swap with a link in the external region (more rigid, as visible in e.g. Fig. S2), where mutations are completely neutral, thus whose cost would be roughly zero.

For the generative step, we implement a Monte Carlo sampling which relocates springs from an occupied to an unoccupied link, i.e. which follows swap-type dynamics as for the original numerical evolution. This allows us to select, from the inferred model, sequences that are structurally as close as possible to the initial data, i.e. with the same average coordination $\langle z \rangle = 5$, to make a consistent comparison with them. We have verified that even relaxing this constraint in the sampling leads to sequences endowed with higher internal variability yet lying in the same range on fitness (hence the inferred model incorporates rather well the information on the fixed amount of springs). The parameters of the Ising model are inferred in such a way as to match single-site occupancy, which reflects the spatial pattern of coordination in the allosteric networks. In Fig. S2 we show that generated sequences, despite having lower fitness, reproduce successfully this property as they should.

Comparison with conservation

Single-site frequency in protein alignments, informative about local conservation, is a standard measure of mutation costs at a certain position [50] and can be fit by an independent-site Ising model. Energy (Eq. 4) in this case contains only field terms and, once these are inferred from link occupancies $\langle \sigma_i \rangle$, one can compute energy changes $\Delta\mathcal{E}_i$ upon point mutations. The energy cost of a mutation in an independent-site model is then $\Delta\mathcal{E}_i = (2\sigma_i - 1)h_i$, where $h_i = \log(\langle \sigma_i \rangle (1-\bar{\sigma})/\bar{\sigma}(1-\langle \sigma_i \rangle))$ describes how the observed occupancy of a link i, $\langle \sigma_i \rangle$, is biased away from the average occupancy $\bar{\sigma} = 360/408 = 0.88$. In average $\Delta\mathcal{E}_i$ gives also a measure of *conservation* of link i as it is 0 when $\langle \sigma_i \rangle = \bar{\sigma}$ and it increases the more link i tends to be either occupied or vacant. The improvement achieved by the pairwise model over this conservation-based measure of mutation costs is extremely significant (see inset of Fig. 3C). On the one hand, conservation is a purely local measure - it takes into account how a particular position is crucial to the propagation of the allosteric response. Including pairwise couplings proves to be crucial to capture the context-dependence of mutation costs thus for their quantitative prediction. On the other hand, the

degree itself of structural conservation is rather low due to the heterogeneity of the shear-design MSA: the conformation, precise location and size of the shear path, hence the role of each link, can vary from architecture to architecture, leading to low structural conservation (with peaks only around the active and allosteric site). Conservation is found much higher *within* one set of dynamically related solutions (as for Fig. 2A), corresponding to one realization of the shear design among the many included in the MSA.

Acknowledgment:

We acknowledge interesting and stimulating discussions with Eric Aurell, John Barton, Johannes Berg, Simona Cocco, Paolo de Los Rios, Solange Flatt, Joachim Krug, Michael Lassig, Duccio Malinverni, Simone Pompei, Remi Monasson, Martin Weigt, Le Yan, Stefano Zamuner. We are particularly grateful to John Barton, Le Yan, Duccio Malinverni and Stefano Zamuner for help with the codes. M. W. thanks the Swiss National Science Foundation for support under Grant No. 200021-165509 and the Simons Foundation Grant (454953 Matthieu Wyart).

References

1. Jingjing G, Huan-Xiang Z. Protein Allostery and Conformational Dynamics. Chem Rev. 2016;116(11):6503–6515.

2. Dokholyan NV. Controlling Allosteric Networks in Proteins. Chem Rev. 2016;116(11):6463–6487.

3. Guarnera E, Berezovsky IN. Allosteric sites: remote control in regulation of protein activity. Curr Opin Struct Biol. 2016;37:1–8.

4. Tang Q, Fenton AW. Whole protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism. Human Mutation. 2017;38:1132–1143.

5. Ahuja LG, Kornev AP, McClendon CL, Veglia G, Taylor SS. Mutation of a kinase allosteric node uncouples dynamics linked to phosphotransfer. Proc Natl Acad Sci USA. 2017;114(6):E931–E940.

6. Olson CA, Wu NC, Sun R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. Curr Biol. 2014;24(22):2643–2651.

7. De Visser JAGM, Cooper TF, Elena SF. The causes of epistasis. Proc R Soc B. 2011;278(1725):3617–3624.

8. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016;25:1204—1218.

9. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. Science. 2007;317(5844):1544–1548.

10. Natarajan C, Inoguchi N, Weber RE, Fago A, Moriyama H, Storz JF. Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin. Science. 2013;340(6138):1324–1327.

11. Schenk MF, Szendro IG, Salverda ML, Krug J, de Visser JAGM. Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene. Mol Biol Evol. 2013;30(8):1779–1787.

12. Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. eLife. 2018;7:e34300.

13. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press, 1998; 1998.

14. Süel GM, Lockless SW, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Mol Biol. 2003;10:59–69.

15. Reynolds KA, McLaughlin RN, Ranganathan R. Hot Spots for Allosteric Regulation on Protein Surfaces. Cell. 2011;147(7):1564–1575.

16. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011;108(49):E1293–E1301.

17. Ovchinnikov S, Park H, Varghese N, Huang P, Pavlopoulos GA, Kamisetty DEK, et al. Protein Structure Determination using Metagenome sequence data. Science. 2017;355(6322):294–298.

18. Barrat-Charlaix P, Figliuzzi M, Weigt M. Improving landscape inference by integrating heterogeneous data in the inverse Ising problem. Scientific Reports. 2016;6:37812.

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

20. Nelson ED, Grishin NV. Inference of epistatic effects in a key mitochondrial protein. Phys Rev E. 2018;97(062404).

21. Poelwijk FJ, Socolich M, Ranganathan R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Bioarxiv http://dxdoiorg/101101/213835. 2017;.

22. Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. Origins of coevolution between residues distant in protein 3D structures. Proc Natl Acad Sci USA. 2017;114(34):9122–9127.

23. Hemery M, Rivoire O. Evolution of sparsity and modularity in a model of protein allostery. Phys Rev E. 2015;91(4):042704.

24. Rocks JW, Pashine N, Bischofberger I, Goodrich CP, Liu AJ, Nagel SR. Designing allostery-inspired response in mechanical networks. Proc Natl Acad Sci USA. 2017;114(10):2520–2525.

25. Flechsig H. Design of elastic networks with evolutionary optimised long-range communication as mechanical models of allosteric proteins. Biophys J. 2017;113(3):558–571.

26. Yan L, Ravasio R, Brito C, Wyart M. Architecture and coevolution of allosteric materials. Proc Natl Acad Sci USA. 2017;114(10):2526–2531.

27. Yan L, Ravasio R, Brito C, Wyart M. Principles for optimal cooperativity in allosteric materials. Biophys J. 2018;114(12):2787–2798.

28. Tlusty T, Libchaber A, Eckmann JP. Physical model of the sequence-to-function map of proteins. Phys Rev X. 2017;7(021037).

29. Dutta S, Eckmann JP, Libchaber A, Tlusty T. Green function of correlated genes in a minimal mechanical model of protein evolution. Proc Natl Acad Sci USA. 2018;.

30. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. Science. 2011;332(6034):1190–1192.

31. Schoustra S, Hwang S, Krug J, de Visser JAGM. Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. Proc R Soc B. 2016;283:20161376.

32. Desai MM, Weissman D, Feldman MW. Evolution Can Favor Antagonistic Epistasis. Genetics. 2007;177:1001–1010.

33. Lalić J, Elena SF. Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. Heredity. 2012;109(2):71–77.

34. Cocco S, Monasson R. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. Phys Rev Lett. 2011 Mar;106:090601.

35. Cocco S, Monasson R. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. J Stat Phys. 2012;147:252–314.

36. Jacquin H, Gilson A, Shakhnovich E, Cocco S, Monasson M. Benchmarking Inverse Statistical Approaches for Protein Structure and Design with Exactly Solvable Models. PLoS Comput Biol. 2016;12(5):e1004889.

37. Mitchell MR, Tlusty T, Leibler S. Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings. Proc Natl Acad Sci USA. 2016;113(40):5847–5855.

38. De Los Rios P, Cecconi F, Pretre A, Dietler G, Michielin O, Piazza F, et al. Functional dynamics of PDZ binding domains: a normal-mode analysis. Biophys J. 2005;89(1):14–21.

39. Zheng W, Brooks BR, Thirumalai D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci USA. 2006;103(20):7664–7669.

40. Wilke CO, Christoph A. Interaction between directional epistasis and average mutational effects. Proc R Soc B. 2001;268(1475):1469–1474.

41. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11(801):801–807.

42. Tubiana J, Cocco S, Monasson R. Learning protein constitutive motifs from sequence data. arXiv:180308718. 2018;.

43. Humplik J, Tkačik G. Probabilistic models for neural populations that naturally capture global coupling and criticality. PLoS Comput Biol. 2017;13(9):e1005763.

44. Otwinowski J, McCandlish DM, Plotkin JB. Inferring the shape of global epistasis. Proc Natl Acad Sci USA. 2018;115:E7550–E7558.

45. Nguyen R H C R Zecchina, Berg J. Inverse statistical problems: from the inverse Ising problem to data science. Adv Phys. 2017;66(3):1–65.

46. Bachschmid-Romano L, Opper M. A statistical physics approach to learning curves for the inverse Ising problem. J Stat Mech Theory Exp. 2017;2017(6):063406.

47. Barton JP, De Leonardis E, Coucke A, Cocco S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics. 2016;32(20):3089–3097.

48. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys. 2018;81(3):032601.

49. Barton JP, Cocco S, De Leonardis E, Monasson R. Large pseudocounts and L_2-norm penalties are necessary for the mean-field inference of Ising and Potts models. Phys Rev E. 2014;90:012132.

50. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073.

Supplementary Information S1 Text:

Characterizing and inferring epistasis via direct evolutionary couplings in models of allostery

Barbara Bravi[1], Riccardo Ravasio[1], Carolina Brito[2], Matthieu Wyart[1]

[1] *Institute of Physics, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

[2] *Instituto de Fìsica, Universidade Federal do Rio Grande do Sul, CP 15051, 91501-970 Porto Alegre RS, Brazil*

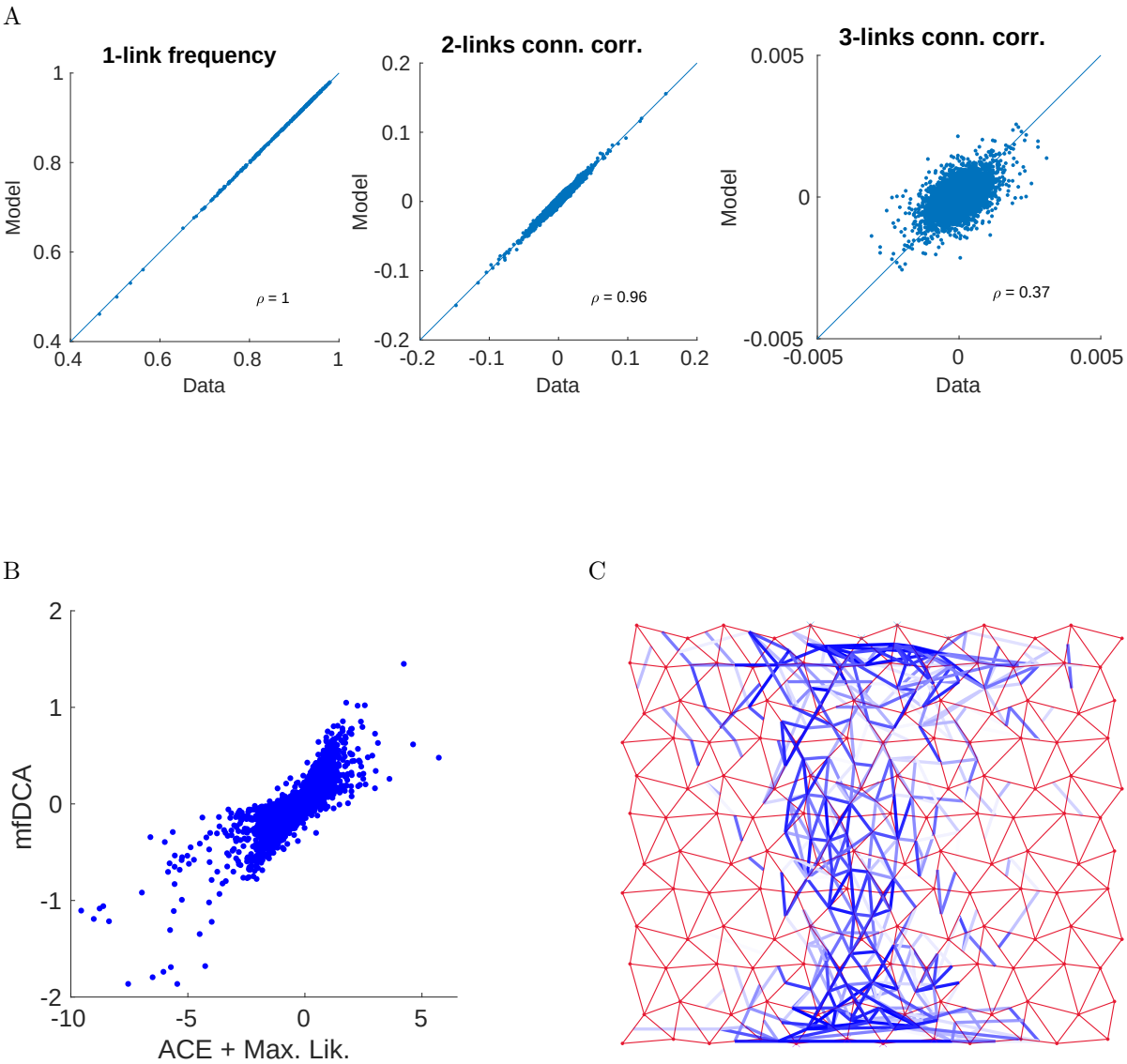

Figure S1: **Performance of the inference procedure**. A: Statistics of the model inferred by combining ACE and Maximum Likelihood. 1-link frequency and 2-links connected correlations are very accurately reproduced, as they should by construction (the relative errors, defined as in [1], are respectively $\epsilon_m = 2.45 \times 10^{-1}$ and $\epsilon_C = 1.30 \times 10^{-1}$). In contrast the third order connected correlations, which are not constrained in the inference, are not well captured (Pearson coefficient $\rho = 0.37$). This is a hint that the Ising model - a pairwise probabilistic model over σ_i - is an approximation which becomes poor for estimating higher order moments. B: Scatter plot comparing J_{ij} inferred via mfDCA to the direct couplings of ACE + Max. Lik.: the pseudocount in mfDCA has been set to $\lambda = 0.5$ in such a way as to obtain the highest correlation between the two. C: Spatial distribution of top 400 mfDCA-inferred couplings on the network. The reconstruction of the topology of relevant couplings is rather robust with respect to the choice of more approximate inference methods as mfDCA. As in Fig. 5A (inset) of the main text, they are concentrated at short range, i.e. they connect links lying close either to the active site or the allosteric site and in the central high-shear path. Long range mfDCA couplings, connecting links around respectively allosteric and active site, are weaker and appear among the top 600-1000 ones, implying an even worse performance at predicting long range epistasis than ACE + Max. Lik.

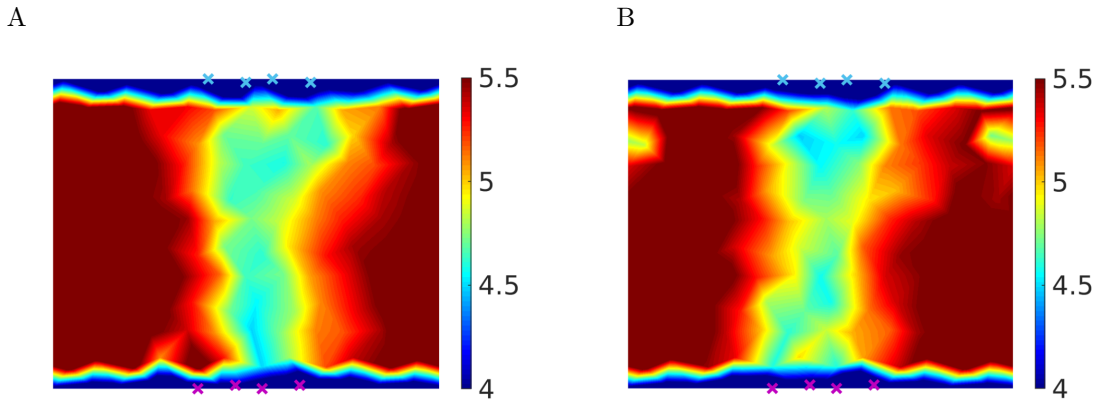

Figure S2: **Properties of generated allosteric sequences**. Coordination map of original sequences (A) and generated ones (B). They both exhibit a softer (i.e. with coordination $z < 5$) central path joining active and allosteric sites (indicated respectively by blue and purple crosses) along which the shear-like sliding takes place. This path is embedded in a more connected, "rigid" region where the coordination $z > 5$. Solutions sampled from the inferred energy landscape have a design but are not maximally fit, showing that more "structural" components, as the distribution of links, are captured but additional information would be needed to reproduce a complex mechanical function as the cooperative fitness.

1 Mechanical interpretation of mutation costs and epistasis

Let us denote by ϵ the set of nodes where ligand binding takes place, e.g. for ligand binding at the allosteric site $\epsilon = (\mathcal{A}l)$ with size $\dim(\epsilon) = n_0$. Such event imposes a displacement $\mathbf{R}^\epsilon$ on the nodes ϵ which imparts locally a force $\boldsymbol{F}^\epsilon$ and induces a response $\mathbf{R}^{\epsilon\to r}$ on all the other nodes r. Clearly $\dim(\epsilon) + \dim(r) = L^d$ where L^d is the total number of nodes for a network of size L in d dimensions; for the example of binding to the allosteric site $r = (\mathcal{A}c, b)$, where b stands for the "bulk" of nodes not belonging neither to the allosteric nor to the active site. (In this paper we consider networks as in Fig. 1A of the main text, with $d = 2$, $L = 12$ and $n_0 = 4$ for both active and allosteric site). The relation between force and overall response field is written in terms of the dynamical matrix $\boldsymbol{\mathcal{M}}$

$$\begin{pmatrix} \boldsymbol{F}^\epsilon \\ \mathbb{0} \end{pmatrix} = \boldsymbol{\mathcal{M}} \begin{pmatrix} \mathbf{R}^\epsilon \\ \mathbf{R}^{\epsilon\to r} \end{pmatrix} \tag{5}$$

hence $\boldsymbol{\mathcal{M}}$ is endowed with a block structure as follows

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \boldsymbol{\mathcal{M}}^{\epsilon,\epsilon} & \boldsymbol{\mathcal{M}}^{\epsilon,r} \\ (\boldsymbol{\mathcal{M}}^{\epsilon,r})^T & \boldsymbol{\mathcal{M}}^{r,r} \end{pmatrix}$$

Forces as well as responses can be calculated solely from the imposed displacement by introducing a matrix $\boldsymbol{\mathcal{Q}}$

$$\boldsymbol{\mathcal{Q}} = \begin{pmatrix} \mathbb{1}^\epsilon & -\boldsymbol{\mathcal{M}}^{\epsilon,r} \\ \mathbb{0} & -\boldsymbol{\mathcal{M}}^{r,r} \end{pmatrix}$$

such that

$$\begin{pmatrix} \boldsymbol{F}^\epsilon \\ \mathbf{R}^r \end{pmatrix} = \boldsymbol{\mathcal{Q}}^{-1}\boldsymbol{\mathcal{M}} \begin{pmatrix} \mathbf{R}^{\epsilon\to r} \\ \mathbb{0} \end{pmatrix} \tag{6}$$

Binding at ϵ costs an elastic energy E^ϵ

$$E^\epsilon = \frac{1}{2}\boldsymbol{F}^\epsilon \cdot \boldsymbol{R}^\epsilon \tag{7}$$

and the cooperative fitness is specified by a combination of such elastic energies

$$\mathcal{F} = E^{\mathcal{A}c} - (E^{\mathcal{A}c,\mathcal{A}l} - E^{\mathcal{A}l}) \tag{8}$$

where $E^{\mathcal{A}c}$, $E^{\mathcal{A}c,\mathcal{A}l}$ and $E^{\mathcal{A}l}$ are given by Eq. 7 with $\epsilon = (\mathcal{A}c)$, $\epsilon = (\mathcal{A}c, \mathcal{A}l)$ and $\epsilon = (\mathcal{A}l)$ respectively. Maximal cooperativity corresponds to making binding of a substrate at the active site energetically favored when already a ligand is bound to the allosteric site, as this reduces its binding energy from $E^{\mathcal{A}c}$ to $(E^{\mathcal{A}c,\mathcal{A}l} - E^{\mathcal{A}l})$. One can express the energy of joint binding at the allosteric and active site $E^{\mathcal{A}c,\mathcal{A}l} = \frac{1}{2}\boldsymbol{F}^{\mathcal{A}c,\mathcal{A}l} \cdot \boldsymbol{R}^{\mathcal{A}c,\mathcal{A}l}$ as

$$\frac{1}{2}\boldsymbol{F}^{\mathcal{A}c,\mathcal{A}l} \cdot \boldsymbol{R}^{\mathcal{A}c,\mathcal{A}l} = \frac{1}{2}\boldsymbol{F}^{\mathcal{A}l} \cdot \boldsymbol{R}^{\mathcal{A}l} + \frac{1}{2}\boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l} \cdot (\boldsymbol{R}^{\mathcal{A}c} - \boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}) \tag{9}$$

i.e. after binding at the allosteric site with an energy cost $\frac{1}{2}\boldsymbol{F}^{\mathcal{A}l} \cdot \boldsymbol{R}^{\mathcal{A}l}$, the elastic energy of binding at the active site is determined by (i) the force there when a ligand is already bound at the allosteric site ($\boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l}$ with subindex $|\mathcal{A}l$); (ii) the displacement imposed at the active site $\boldsymbol{R}^{\mathcal{A}c}$ to which we subtract the

response already caused by ligand binding at the allosteric site $\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}$. Eq. 9 allows us to rewrite Eq. 8 as

$$\mathcal{F} = \frac{1}{2}\,\boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c} + \frac{1}{2}\,\delta\boldsymbol{F}^{\mathcal{A}l\to\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}c} \tag{10}$$

where one has $\boldsymbol{F}^{\mathcal{A}c} - \boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l} = \delta\boldsymbol{F}^{\mathcal{A}l\to\mathcal{A}c}$. If we express $\delta\boldsymbol{F}^{\mathcal{A}l\to\mathcal{A}c}$ and $\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}$ in terms of the imposed displacements by using Eq. 6 and if we assume weak elastic coupling between allosteric and active site, we find that each term in Eq. 10 scales in the same way as

$$\frac{1}{2}\,\boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c} \approx \frac{1}{2}\,\delta\boldsymbol{F}^{\mathcal{A}l\to\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}c} \approx \frac{1}{2}\,(\boldsymbol{R}^{\mathcal{A}c})^T\cdot(\boldsymbol{\mathcal{M}}^{\mathcal{A}c,b})(\boldsymbol{\mathcal{M}}^{b,b})^{-1}(\boldsymbol{\mathcal{M}}^{b,\mathcal{A}l})\cdot\boldsymbol{R}^{\mathcal{A}l} \tag{11}$$

Hence, by using that $\frac{1}{2}\,\delta\boldsymbol{F}^{\mathcal{A}l\to\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}c} \approx \frac{1}{2}\,\boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}$, we obtain from Eq. 10

$$\mathcal{F} \approx \mathbf{F}^{\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c} \tag{12}$$

since $\boldsymbol{F}^{\mathcal{A}c}_{|\mathcal{A}l}$ can be approximated by $\boldsymbol{F}^{\mathcal{A}c}$ in the weak coupling limit.

If we denote by $\mathbf{F}^{\mathcal{A}c}_i$ and $\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i$ forces and displacements after a mutation at link i, the cost of one mutation can be expressed approximatively (see Fig. S3B) as $\Delta\mathcal{F}_i \approx \Delta(\mathbf{F}^{\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c})_i$, where $\Delta(\mathbf{F}^{\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c})_i = \mathbf{F}^{\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c} - \mathbf{F}^{\mathcal{A}c}_i\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i$. This can be further rewritten as

$$\Delta(\mathbf{F}^{\mathcal{A}c}\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c})_i \approx -\big(\mathbf{F}^{\mathcal{A}c}\cdot\delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i + \delta\mathbf{F}^{\mathcal{A}c}_i\cdot\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c} + \delta\mathbf{F}^{\mathcal{A}c}_i\cdot\delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i\big) \tag{13}$$

having defined changes in force as $\delta\mathbf{F}^{\mathcal{A}c}_i = \mathbf{F}^{\mathcal{A}c}_i - \mathbf{F}^{\mathcal{A}c}$ in analogy to changes in displacement $\delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i$ introduced in the main text. We find numerically that the cost of single mutations, when it is not too small, is dominated by the changes in displacement at the active site

$$\Delta\mathcal{F}_i \approx -\mathbf{F}^{\mathcal{A}c}\cdot\delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i \tag{14}$$

as shown in Fig. S3C. As a consequence, epistasis between mutations at i and j with significant magnitude can be written $\Delta\Delta\mathcal{F}_{ij} \approx -\mathbf{F}^{\mathcal{A}c}\cdot\big(\delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_{ij} - \delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_i - \delta\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}_j\big)$, as presented in the main text. Displacement vectors and their changes upon high-cost mutations at the active site are schematically depicted in Fig. S3A.

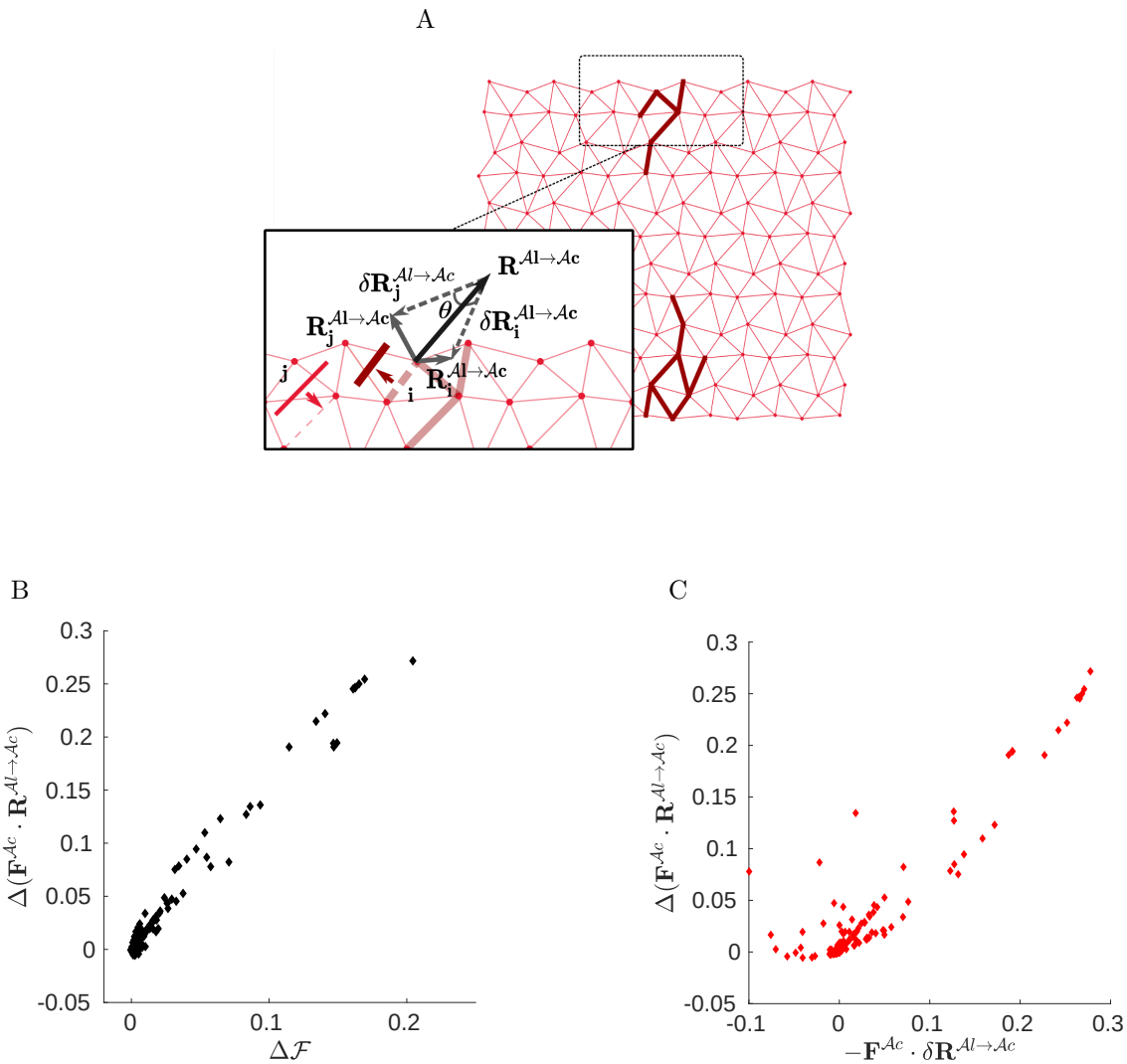

Figure S3: **Mechanics of mutations**. A: The geometry of mutation costs is illustrated in the zoom on the active site region (note that for simplicity of visualization we consider only one of the $n_0 = 4$ nodes). Thick, dark red lines highlight links whose disruption would be lethal for the allosteric fitness. These few links, crucial to the long-distance propagation of the allosteric response, are located around active and allosteric site and exhibit maximal epistasis along with maximal single mutation costs (i.e. they populate the saturation region of Fig. 2A in the main text). After a lethal mutation consisting in removing a spring at link i, the displacement at the active site $\mathbf{R}_i^{Al\to Ac}$ is significantly reduced with respect to the original optimal displacement $\mathbf{R}^{Al\to Ac}$ and their difference is given by $\delta\mathbf{R}_i^{Al\to Ac}$ (dashed arrow). When a second lethal mutation at j occurs, we denote by θ the angle between $\delta\mathbf{R}_i^{Al\to Ac}$ and $\delta\mathbf{R}_j^{Al\to Ac}$; for lethal mutations $\cos(\theta) \approx 1$ (see Fig. 2B in the main text), i.e. they all tend to have a homogeneous direction of action which is precisely the one opposite to the displacement at the active site. B: Numerical test of the approximation $\Delta\mathcal{F}_i \approx \Delta(\mathbf{F}^{Ac}\cdot\boldsymbol{R}^{Al\to Ac})_i$ and of $\Delta(\mathbf{F}^{Ac}\cdot\boldsymbol{R}^{Al\to Ac})_i \approx -\mathbf{F}^{Ac}\cdot\delta\boldsymbol{R}_i^{Al\to Ac}$ (C). The latter is valid only for medium-high mutation costs.

2 Prediction of epistasis

The scaling of epistasis (Eq. 2 in the main text) suggests a measure simply based on the inferred single mutation costs, i.e. $|\Delta\Delta\mathcal{F}_{ij}| \propto \min(\Delta\mathcal{E}_i, \Delta\mathcal{E}_j)$. We have verified that this improves extremely the prediction of long-range epistasis in our model for allostery, both for single configurations and for the average epistatic pattern, as shown in respectively in Fig. S4B and C. The measure of epistasis via top $|J_{ij}|$ requires the inferred model to be performant at capturing local information via local parameters; on the other hand, the estimation of single mutation costs incorporates all the local parameters inferred from the statistics. These results support the view that more functional information (related to non-local modes) is embedded in weaker couplings which would be excluded by applying the contact-prediction criterion of looking at the largest ones (usually as many as the system size): for example recently [2] has found that the prediction of functional cooperativity between distant sites could be improved by considering several "non-contacting" DCA couplings.

2.1 Simple model illustrating the failure of DCA

To explain the discrepancy between short-range and long-range DCA-predictions of epistasis, we resort to the simple model of Fig. 6 (main text). We assign to all the 49 functional configurations the same fitness $\mathcal{F}$, all the other $2^8 - 49$ configurations would not belong to the sample of optimal configurations and are taken with zero fitness, thus $\Delta\mathcal{F} = 0$ if a mutation (single or double) results in a configuration still belonging to the optimal sample and $\Delta\mathcal{F} = \mathcal{F}$ otherwise. We can estimate average mutation costs by counting how frequently mutations would lead to a configuration outside of the optimal sample, yielding

$$\Delta\Delta\mathcal{F}_{12} = \Delta\mathcal{F}_{12} - \Delta\mathcal{F}_1 - \Delta\mathcal{F}_2 = 21/49\mathcal{F} - 21/49\mathcal{F} - 21/49\mathcal{F} = -21/49\mathcal{F} \tag{15}$$

$$\Delta\Delta\mathcal{F}_{15} = 33/49\mathcal{F} - 21/49\mathcal{F} - 21/49\mathcal{F} = -9/49\mathcal{F} \tag{16}$$

$$\frac{|\Delta\Delta\mathcal{F}_{12}|}{|\Delta\Delta\mathcal{F}_{15}|} = 21/9 \approx 2.3 \tag{17}$$

Next, by a simple likelihood maximization we infer the set of J_{ij} and h_i compatible with $\langle\sigma_i\rangle$ and $\langle\sigma_i\sigma_j\rangle$, single-site and pairwise frequencies of the optimal sample. We estimate $J_{12} = 1.18$ and $J_{15} = 0.40$, thus the prediction by DCA

$$\frac{|\Delta\Delta\mathcal{E}_{12}|}{|\Delta\Delta\mathcal{E}_{15}|} = \frac{|J_{12}(2\langle\sigma_1\rangle + 2\langle\sigma_2\rangle - 4\langle\sigma_1\sigma_2\rangle - 1)|}{|J_{15}(2\langle\sigma_1\rangle + 2\langle\sigma_5\rangle - 4\langle\sigma_1\sigma_5\rangle - 1)|} = \frac{|J_{12}(-21/49)|}{|J_{15}(-9/49)|} \approx 6.9 \tag{18}$$

i.e. the DCA prediction is significantly biased towards short-range epistasis. Due to symmetry of our model, epistasis and the DCA-prediction for any combination of units in the two subparts is the same as for units 1 and 5; similarly, the result for 2 units within the same group is given by the values for units 1 and 2. For the remaining combinations of units, i.e. the ones belonging the same subpart but to different groups (e.g. $i = 1$ and $j = 3$) we obtain that epistasis is weaker compared to units within the same group

$$\frac{|\Delta\Delta\mathcal{F}_{12}|}{|\Delta\Delta\mathcal{F}_{13}|} = \frac{|-21/49\mathcal{F}|}{|-7/49\mathcal{F}|} = 3 \tag{19}$$

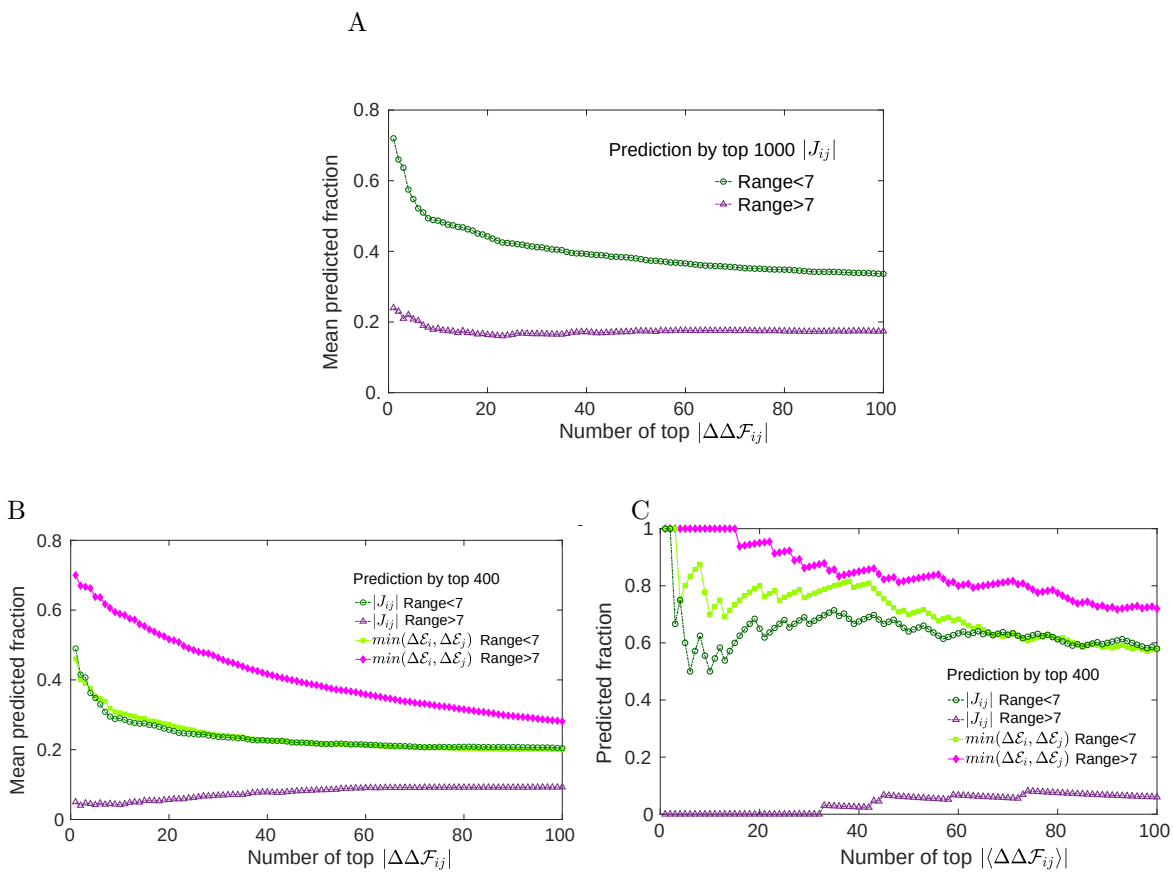

Figure S4: **Prediction of epistasis by the DCA-inferred model**. A: Same plot as in Fig. 5B (main text) where we show the fraction of top rank epistasis $|\Delta\Delta\mathcal{F}_{ij}|$ predicted by top 1000 $|J_{ij}|$, averaged over 100 configurations. In comparison to Fig. 5B, here we consider a higher number of the largest in magnitude couplings to predict epistasis: the mean predicted fraction increases both for short range and long range epistasis, yet a clear difference between their values remains. B: Same plot as Fig. 5B (main text) where we added curves for the prediction by $\min(\Delta\mathcal{E}_i, \Delta\mathcal{E}_j)$ - the minimum between average single mutation costs at i and j - as implied by scaling 2 in the main text. As in Fig. 5B, we rank separately long-range (> 7) and short-range (< 7) pairs of links i and j in terms of $|\Delta\Delta\mathcal{F}_{ij}|$ and we plot the fraction of these pairs - averaged over 100 configurations randomly chosen - falling either into the top 400 $|J_{ij}|$ (empty symbols) or into the top 400 values of $\min(\Delta\mathcal{E}_i, \Delta\mathcal{E}_j)$ (filled symbols). This second measure improves only slightly the estimation of strong short-range epistasis but it does so dramatically for long-range one. C: Same plot as B where we show the fraction of the average epistasis $\langle\Delta\Delta\mathcal{F}_{ij}\rangle$ (estimated from 1.5×10^3 randomly chosen configurations of the MSA) that one would predict either via $|J_{ij}|$ or $\min(\Delta\mathcal{E}_i, \Delta\mathcal{E}_j)$. The prediction at short distance is rather accurate, with the predicted fraction reaching 1 for the maximally epistatic pairs; at long distance, signal on long-range epistasis captured by $|J_{ij}|$ is almost absent while the prediction by $\min(\Delta\mathcal{E}_i, \Delta\mathcal{E}_j)$ stands out for its precision.

Since each subpart can be of different type (OR gate), units from different groups (i.e. types) are less tightly constrained by function. The DCA-prediction does not underestimate epistasis as for units of different subparts (i.e. at long distance) with

$$\frac{|\Delta\Delta\mathcal{E}_{12}|}{|\Delta\Delta\mathcal{E}_{13}|} = \frac{|J_{12}(-21/49)|}{|J_{13}(-7/49)|} \approx 3.5 \tag{20}$$

where $J_{13} = -1.01$. From Eq. 17, Eq. 18, Eq. 19 and Eq. 20 it is straightforward to calculate $|\Delta\Delta\mathcal{E}_{13}|/|\Delta\Delta\mathcal{E}_{12}| \times |\Delta\Delta\mathcal{F}_{12}|/|\Delta\Delta\mathcal{F}_{13}| \approx 0.86$ and $|\Delta\Delta\mathcal{E}_{15}|/|\Delta\Delta\mathcal{E}_{12}| \times |\Delta\Delta\mathcal{F}_{12}|/|\Delta\Delta\mathcal{F}_{15}| \approx 0.33$.

2.1.1 Feedforward neural network

To understand which machine learning tools could improve the prediction of epistasis in the simple model, we have built a feedforward neural network performing least squares regression of sequence data based on their fitness (see Fig. S5). For data in the training set, we provide the network with both the input sequence and the target answer, i.e. a label 1 (standing for fitness $\mathcal{F}$) or 0. We vary the size of the training set from 50% to 80% of the $2^8 = 256$ total sequences and we keep 20% of the sample for validation of the accuracy of prediction. We learn the weights, i.e. the connections between layers, which minimize the mean squared error between the output of the network and the target answers by stochastic gradient descent from a random orthogonal initialization; only relatively few trainings (about 1 in 10) find a high quality solution. We obtain that the mean squared error between true and estimated fitness, averaged over 100 of such high-quality trainings, ranges between $\sim 2 \times 10^{-6}$ for a training set with 50% of the sample to $\sim 2 \times 10^{-7}$ with 80%. Therefore, when the network is presented with an optimal sequence mutated at some position, the network can predict the value of its fitness with extreme accuracy in such a way as to predict $\Delta\mathcal{F} \sim 0$ when it still belongs to the optimal sample or $\Delta\mathcal{F} \sim 1$ if it does not. This ensures that also epistasis would be accurately predicted at any range.

References

1. Cocco S, Monasson R. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. J Stat Phys. 2012;147:252–314.

2. Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. eLife. 2018;7:e34300.

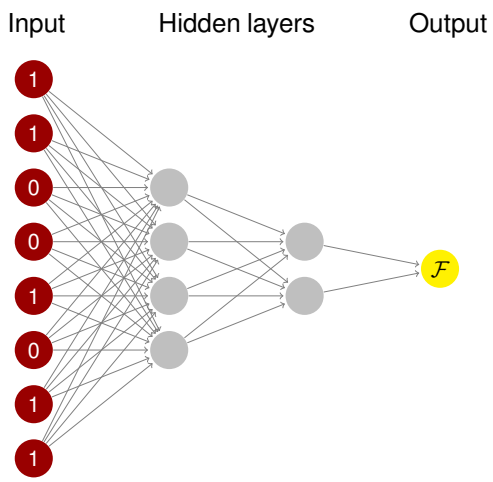

Figure S5: **Graphical representation of the feedforward neural network for regression in the simple model**. The size of the input layer is 8, as the size of the system. We add two hidden layers of 4 and 2 units and the final one-unit output is 1 if the input sequence has fitness $\mathcal{F}$ and 0 otherwise. The activation function from one layer to the successive one is a sigmoid and the weights are dense (all units in one layer are connected to all units of the successive one).