PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Direct coupling analysis of epistasis in allosteric materials

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

1 Institute of Physics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, **2** Instituto de Fisica, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

* barbarabravi@ymail.com (BB); matthieu.wyart@epfl.ch (MW)

OPEN ACCESS

Citation: Bravi B, Ravasio R, Brito C, Wyart M (2020) Direct coupling analysis of epistasis in allosteric materials. PLoS Comput Biol 16(3): e1007630. https://doi.org/10.1371/journal.pcbi.1007630

Editor: Alexandre V. Morozov, Rutgers University, UNITED STATES

Received: January 15, 2019

Accepted: January 3, 2020

Published: March 2, 2020

Data Availability Statement: All relevant data are within the manuscript, its Supporting Information files, and the Git repository (https://gitlab.com/bbravi/mutations_allosteric_networks).

Funding: MW is supported by Swiss National Science Foundation under grant no. 200021-165509 and the Simons Foundation Grant (454953 Matthieu Wyart). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 in our model the 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 the cost of point mutations but is a rather poor generative model. Strikingly, it can predict short-range epistasis but fails to capture long-range epistasis, in consistence 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 the 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 point mutation costs but underestimates strong long-range epistasis. We provide an argument, based on a simplified model, illustrating the reasons for this discrepancy. Our analysis suggests 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 couplihngs 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 [18–22] and correlate with epistasis [23, 24]. In the context of allostery, there is no statistical evidence for the existence of long-range direct couplings that would reveal allosteric channels [25], in apparent contradiction with the existence of extended sectors reported in [15] and the observation of long-range epistasis [6]. It is therefore an open question why a pairwise model should be successful at predicting protein structure, but not long-range functional dependencies. In this work we propose an explanation for this discrepancy, by benchmarking DCA in models of protein allostery where a material evolves *in silico* to achieve an "allosteric" task [26–32]. We consider recent models incorporating elasticity [27–30, 32], in which long-range co-evolution [29], elongated sectors [29] and long-range epistasis [32] are present and can be interpreted in terms of the propagation of an elastic signal [32]. We focus on materials evolved to optimize cooperative binding over large distances [30], and find that the four types of epistasis (Synergistic, Sign, Antagonistic, Saturation) exist over a wide spatial range. We perform DCA and find that it predicts well the cost of point mutations but is a rather poor generative model. It can predict short-range epistasis but fails to capture long-range effects, in agreement with empirical findings [25]. Moreover, we test this result for one allosteric protein, the PDZ domain, where epistasis was experimentally measured in [12] along with the inference of DCA energetic couplings, showing support for our prediction. We illustrate why it may be so via a simple model, which suggests that neural networks may be better suited than DCA to capture long-range effects.

Model for the evolution of allostery

We follow the scheme of [29, 30] 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 [30], 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

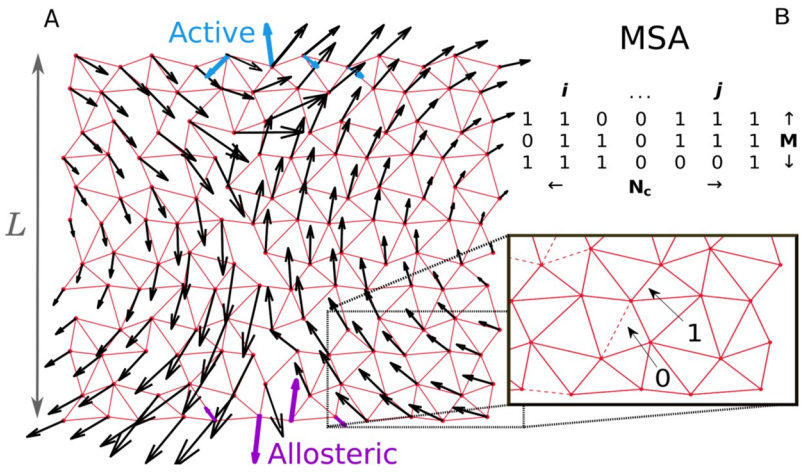

Fig 1. Study of co-evolution in artificial allosteric networks. A: Example of an elastic network made of harmonic springs (red) evolved *in silico* to maximize the cooperativity between the allosteric site (purple) and the active site (blue). The response to binding at the allosteric site is indicated 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.

https://doi.org/10.1371/journal.pcbi.1007630.g001

site only ($\mathcal{A}l$) and at both sites simultaneously ($\mathcal{A}c$, $\mathcal{A}l$). In the limit of weak elastic coupling between allosteric and active site, the fitness can be rewritten approximately as (see Sec. 1 in S1 Text)

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

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. The product $\mathbf{F}^{\mathcal{A}c} \cdot \boldsymbol{R}^{\mathcal{A}l \to \mathcal{A}c}$ is an estimate of the change of mechanical work required for binding the substrate at the active site caused by binding the ligand at the allosteric site. 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 $\beta = 10^4$ as at this temperature networks probed have the highest fitness our protocol can reach [30]. 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 [30]. 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 (see Fig 1A). 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 (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. Single mutation costs $\Delta\mathcal{F}_i$ are expected to be positive since the original network has been selected to have close-to-maximal fitness.

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 $\Delta\Delta\mathcal{F}_{ij} \equiv \Delta\mathcal{F}_{ij} - \Delta\mathcal{F}_i - \Delta\mathcal{F}_j$. We find that generically, the dominant effect of mutations is to affect the propagation of the signal $\boldsymbol{R}^{\mathcal{A}l\to\mathcal{A}c}$, which depends on the arrangement of links in the network. In general, mutations do not affect how binding at the active site locally generates force, as shown in Sec. 1 in S1 Text. Using this observation and following Eq 1, epistasis follows approximately

$$\Delta\Delta\mathcal{F}_{ij} \approx -\mathbf{F}^{\mathcal{A}c} \cdot (\delta\boldsymbol{R}_{ij}^{\mathcal{A}l\to\mathcal{A}c} - \delta\boldsymbol{R}_i^{\mathcal{A}l\to\mathcal{A}c} - \delta\boldsymbol{R}_j^{\mathcal{A}l\to\mathcal{A}c})$$

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 $\delta\mathbf{R}_j^{\mathcal{A}l\to\mathcal{A}c}$.

Consider the case where the cost of a double mutation is dominated by the strongest point mutation, i.e. $\Delta\mathcal{F}_{ij} \approx \max(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j)$. It leads to:

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

Interestingly, this situation does capture the main trend of epistasis in our data, especially when it is strong, as shown in Fig 2A (see dashed line). 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. First of all, no epistasis corresponds to purely additive mutations, i.e. $\Delta\Delta\mathcal{F}_{ij} = 0$, see dotted line in Fig 2A. Next, we observe the following regimes

Saturation: We define mutations with $\Delta\mathcal{F} > 0.1$ as "lethal". This somewhat arbitrary definition corresponds to 50% of loss of fitness. Pairs of such lethal mutations (which represent $\sim 0.1\%$ of all pairs, a sparsity in line with experimental findings [24]) 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 [33, 34].

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, 35], whereby, after a first mutation, making a second one results only in a weak additional change. Antagonistic epistasis is also known as positive magnitude epistasis (where positivity indicates that the double mutant is fitter than expected from the additive case).

Sign. In the intermediate range of mutation costs with $\min(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j) < 0.1$, 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

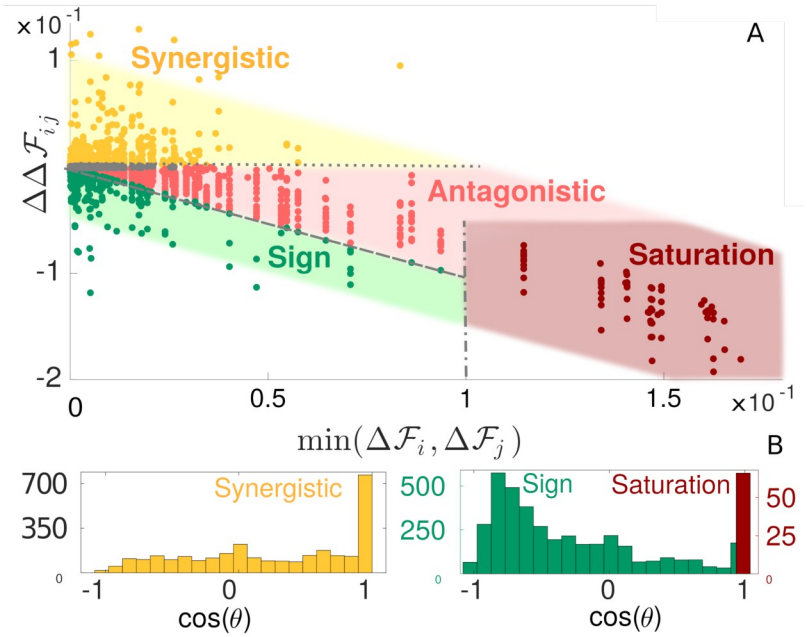

Fig 2. Classification and mechanical characterization of epistasis in our model of allosteric cooperativity. A: Phase diagram of epistasis in our allosteric material. All quantities are averages over 50 configurations obtained in a single run. The shaded area is taken with arbitrary width and a -1 slope as a guide to the eye. We show the lines $\Delta\Delta\mathcal{F}_{ij} = 0$ (dotted style), which corresponds to no epistasis (and divides synergistic from antagonistic/sign epistasis), $\Delta\Delta\mathcal{F}_{ij} = \max(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j)$ (dashed style), separating sign and antagonistic epistasis, and $\min(\Delta\mathcal{F}_i, \Delta\mathcal{F}_j) = 0.1$ (dash-dotted style), the threshold set to distinguish lethal mutations (corresponding to the saturation region). 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.

https://doi.org/10.1371/journal.pcbi.1007630.g002

can become beneficial (i.e. increase the fitness) in presence of another mutation, and this resembles the "sign" epistasis empirically detected [7, 36]. 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 of $\Delta\Delta\mathcal{F}_{ij}$ 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 thus eliminate the possibility of our sequences to display "phylogenetic" effects, i.e. correlations due to a common evolutionary history, known to complicate the inference from sequence data and to require *ad hoc* corrections, see e.g. [37]). 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 $N_c = (3L^2 - 2L) = 408$. Working in such an over-sampling regime (which is generally not the case for real proteins) ensures that the limitations of the inference we find below are not due to sampling, but to the model underlying DCA.

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

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

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

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 only interested in testing the

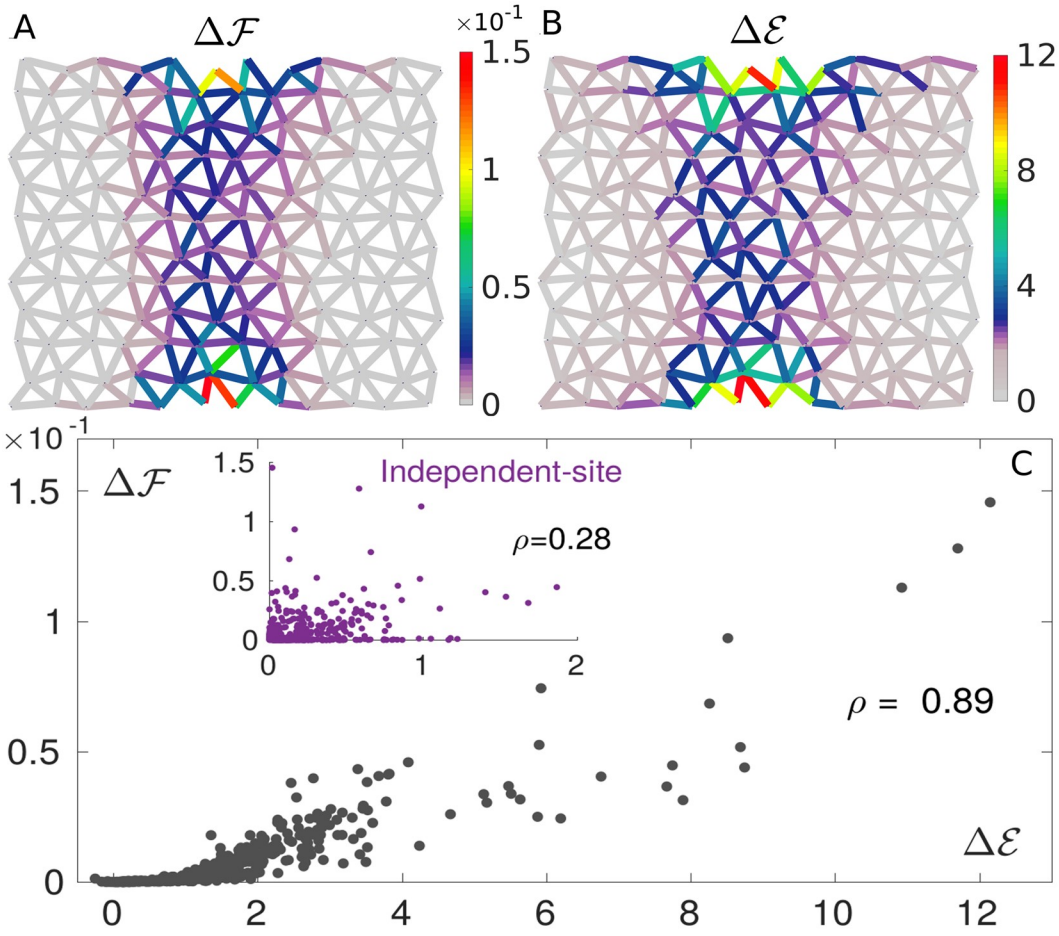

Fig 3. Prediction of mutation costs by DCA. Maps of true $\Delta\mathcal{F}$ (A) and DCA-inferred $\Delta\mathcal{E}$ (B) single mutation costs, averaged over 1.5×10^3 configurations randomly chosen from the MSA. Their patterns are very similar, revealing high costs near the allosteric and active sites 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 (averaged over 1.5×10^3 configurations). 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. The correlation is quantified via the Pearson correlation coefficient, ρ.

https://doi.org/10.1371/journal.pcbi.1007630.g003

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) [38, 39], 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 [30]). 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 (a standard measure of mutation costs, see Methods) performs poorly in this case, see inset of Fig 3C.

Generative power of DCA. Once the model of Eqs 3 and 4 is inferred, can it be used to generate new sequences with a high fitness, as previously shown for models of protein folding [40]? 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

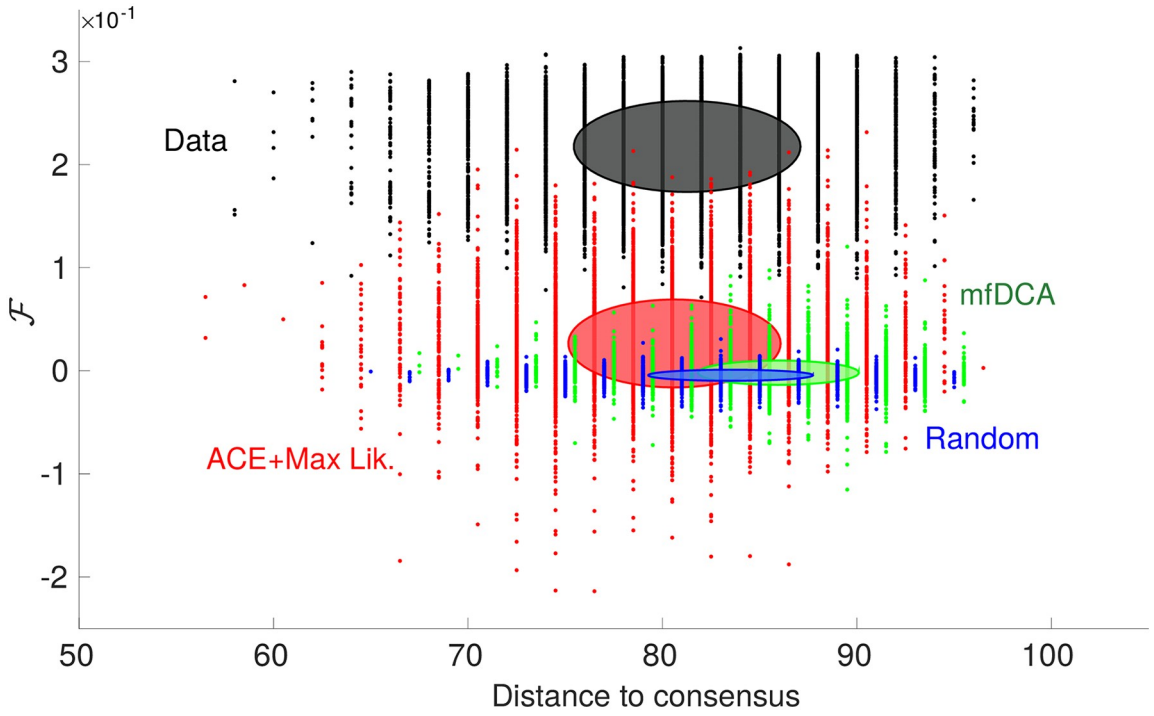

Fig 4. Generative performance of DCA. Fitness vs distance to consensus of configurations generated by the inferred model, following the representation of [40]. The sampling is done from $P(\boldsymbol{\sigma})$ of Eq 4 (a Boltzmann-Gibbs probability distribution), whose parameters have been inferred 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.

https://doi.org/10.1371/journal.pcbi.1007630.g004

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 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 [40]. 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. B in S1 Text), is a rather drastic approximation for the actual allosteric fitness. Indeed we have tested that higher orders as the third moment are not well reproduced (see Fig. A in S1 Text), suggesting that the longer-range correlations induced by allostery are not well captured by a pairwise model. On the other hand, for protein structure predictions, several works as [41] suggest that local correlations between residues in spatial contact are well-captured by a pairwise model, even beyond pairwise correlations. To test our findings, it would be interesting to condition the analysis of e.g. [41] on the distance between residues considered and see if the 3-body correlations are still captured when the residues are further apart. It would also be relevant to restrict the study to allosteric proteins only, to check whether statistical properties are changed, in such a way as to gauge the effect of allosteric vs folding constraints in proteins.

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 (showing that pairs of sites with large mutation cost systematically display strong epistasis) and Fig 3A (showing that sites with a large mutation cost can be distant), 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, 24], one considers epistasis averaged over several configurations (see Sec. 2 in S1 Text). (This result is in contrast to the remarkable performance of DCA in residue contact prediction, which guided the discovery of novel protein structures [17]. We recall that couplings inferred by the most accurate DCA algorithms exhibit maximal precision (i.e. number of true predicted contacts divided by the total number of predictions equal to 1) up to a number of contacts

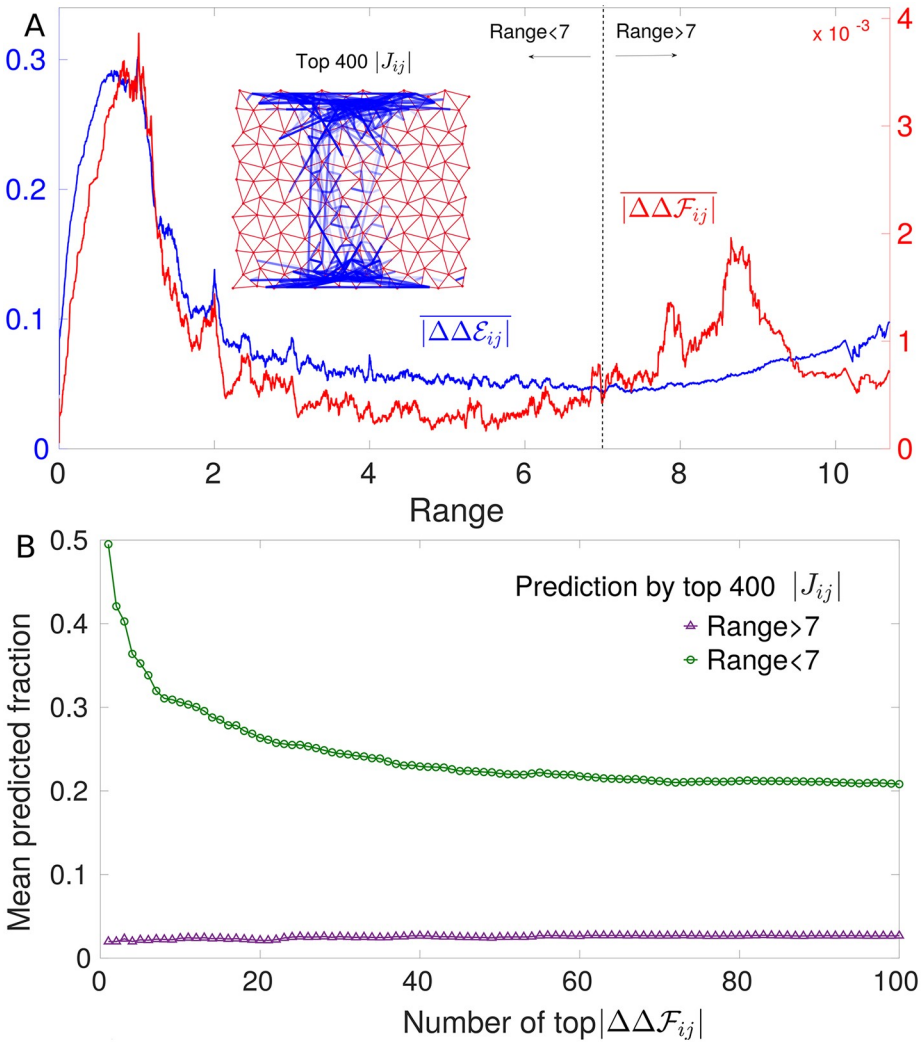

Fig 5. Prediction of epistasis by DCA. A: Running average of the absolute value of epistasis $\Delta\Delta\mathcal{F}_{ij}$ and of DCA prediction $\Delta\Delta\mathcal{E}_{ij}$ for 1.5×10^3 configurations as a function of the distance between link *i* and *j*. The trends are nearly identical at short 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. The random expectations for these mean predicted fractions are 0.0041 for short-range pairs and 0.0009 for long-range ones (they are both significantly lower than the values reported here). This feature stays robust also if we increase, e.g. up to 1000, the number of top couplings for prediction (see Panel A in Fig. D, S1 Text).

https://doi.org/10.1371/journal.pcbi.1007630.g005

comparable with the protein size [42, 43]). Our finding is consistent with the lack of empirical evidence for long-range inferred couplings in allosteric proteins [25].

To better investigate the reasons for this phenomenon in our *in silico* model, we report evolutionary correlations as a function of distance in Fig 6. We find that, although strong long range epistasis occurs, large long-range correlations are absent (a fact in some sense more surprising that not finding long-range couplings, since in principle short-range couplings alone could result in long-range correlations). The absence of long-range correlations suggests that it

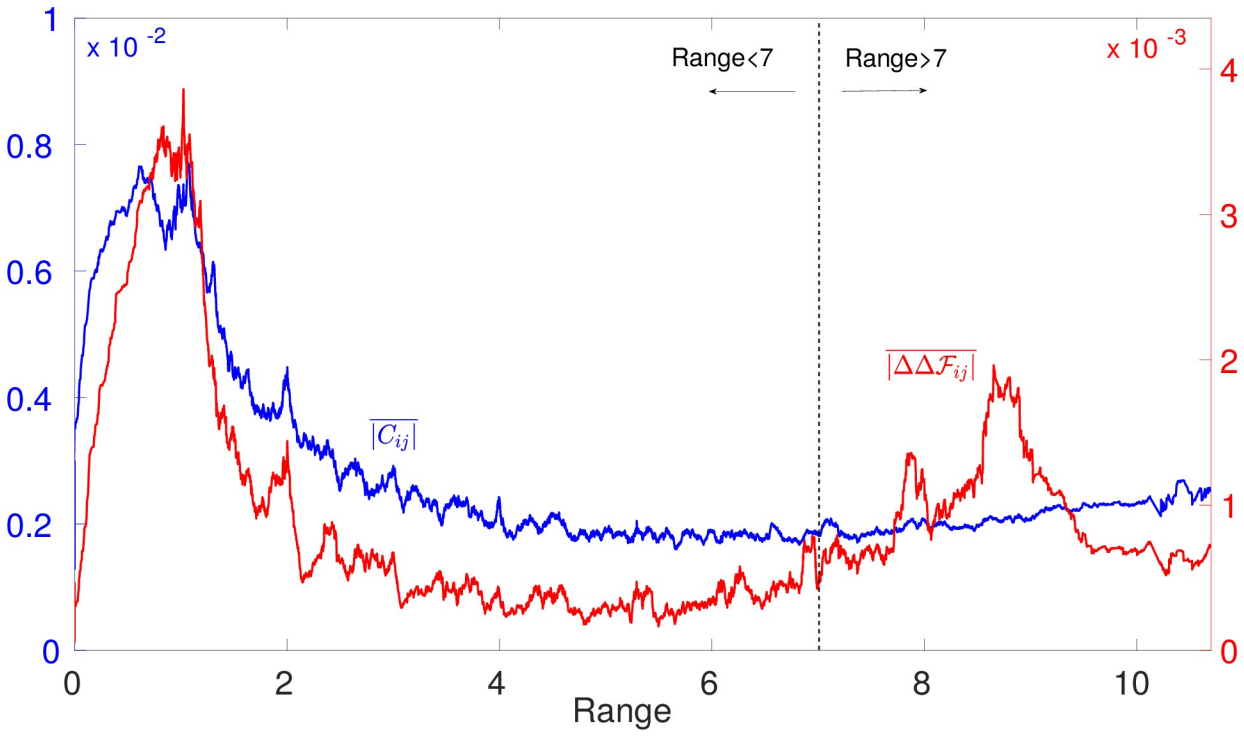

Fig 6. Running average of the absolute value of connected correlations $C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$ and of epistasis $\Delta\Delta\mathcal{F}_{ij}$ for the same 1.5×10^3 configurations of Fig 5A as a function of the distance between link *i* and *j*.

https://doi.org/10.1371/journal.pcbi.1007630.g006

will be particularly challenging to capture long-range functional dependencies from low order statistics of the MSA alone. Consistently with this observation, statistical approaches based on principal components of the MSA covariance such as Sectors [44, 45] or Inverse Covariance Off-Diagonal (ICOD) [46] do not lead overall to better predictions of epistasis in our context, as we show in S1 Text, Sec. 2.2. Among these approaches, we find that the best predictor of long-range epistasis is ICOD, a result that would be interesting to benchmark also in other systems.

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 [30, 47], giving rise to local constraints. Yet, the exact location of these soft regions can vary in space. On a larger length scale, these regions must assemble to create an extended soft elastic mode [30, 48, 49], 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 7. 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 7) and the other 2 the possible types of subpart 2 (right). A configuration is "functional" if 2 units of the same group are

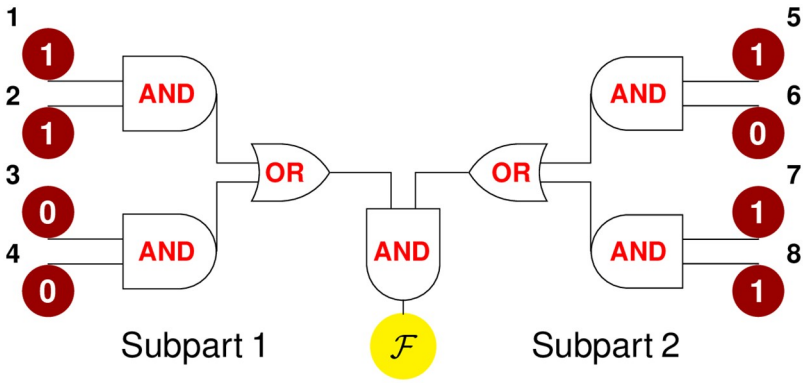

Fig 7. Sketch of a simple model for protein function. A system is arranged into 2 subparts which must work jointly to accomplish a given function (AND gate). Each subpart is composed of 2 groups, i.e. can be of 2 types (OR gate), to work each type must satisfy some constraints (AND gate between single units).

https://doi.org/10.1371/journal.pcbi.1007630.g007

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 (or in the same subpart), 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 by DCA 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 $|\Delta\Delta\mathcal{E}_{15}|/|\Delta\Delta\mathcal{E}_{12}| \times |\Delta\Delta\mathcal{F}_{12}|/|\Delta\Delta\mathcal{F}_{15}| \approx 0.33$, i.e. by 3 fold. In this model as well we find that long-range correlations are essentially absent (they are smaller than 1%), despite long-range epistasis being present. Hence, a functional constraint on the cooperation between subparts potentially far away in the structure, as allosteric and active site, implies strong long-range epistasis, but does not imply strong long-range correlations, which is then reflected in small couplings. To summarize these facts, numerical values for correlation, epistasis and inferred couplings are listed in Table 1. Overall, this situation is

Table 1. Table summarizing true and predicted epistasis magnitude, $|\Delta\Delta\mathcal{F}_{ij}|$ and $|\Delta\Delta\mathcal{E}_{ij}|$, connected correlations C_{ij} and inferred couplings J_{ij} in the simple model for sites i and j in the same group, in the same subpart and in different subparts. For i and j in different subparts (third row) the sizeable magnitude of epistasis is not reflected in the values of correlations, thus of the inferred couplings, in such a way that it is then underestimated by the DCA model. In section Sec. 2.1 in S1 Text, we derive $|\Delta\Delta\mathcal{F}_{ij}| = 21/49\,\mathcal{F}$ for i and j in the same group: since we do not predict the prefactor $\mathcal{F}$, we can fix $21/49\,\mathcal{F} = 1$ and other numbers in the first column follow from this choice.

	$\lvert\Delta\Delta\mathcal{F}_{ij}\rvert$	C_{ij}	$\lvert\Delta\Delta\mathcal{E}_{ij}\rvert$	J_{ij}
Same group	1	0.061	0.51	1.18
Same subpart	0.33	−0.08	0.14	-1.01
Different subpart	0.43	0.00	0.07	0.40

https://doi.org/10.1371/journal.pcbi.1007630.t001

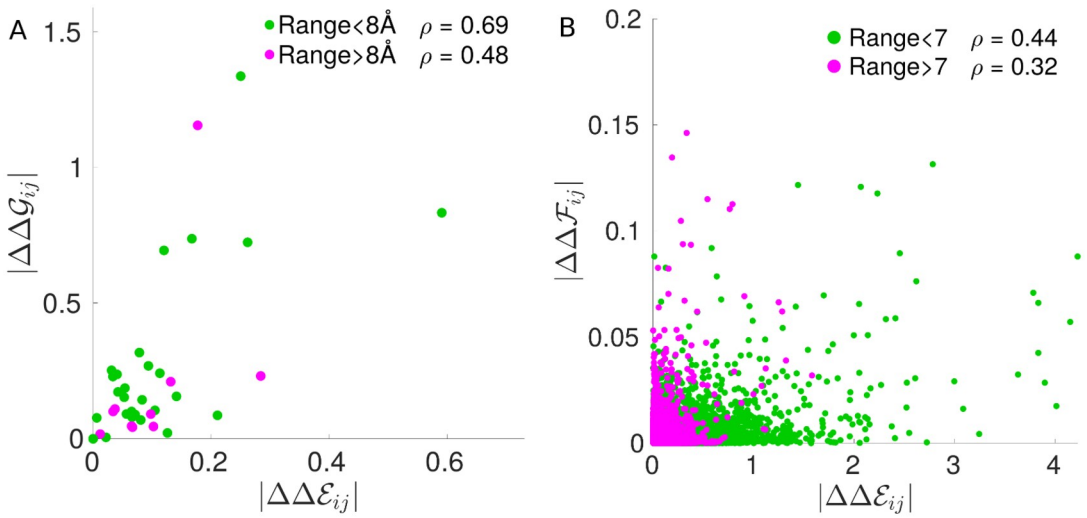

Fig 8. Prediction of experimentally measured epistasis by DCA from [12]. A: Scatter plot of average epistasis magnitude $|\Delta\Delta\mathcal{G}|$ vs DCA-inferred energetic couplings $|\Delta\Delta\mathcal{E}|$, where the color code distinguishes short and long distance pairs of residues on the PDZ α2-helix three-dimensional structure. ρ, the Pearson correlation coefficient, indicates a better performance at short range. As a comparison, in B we show the scatter plot of average epistasis magnitude $|\Delta\Delta\mathcal{F}|$ vs DCA-inferred energetic couplings $|\Delta\Delta\mathcal{E}|$ in our *in silico* evolved networks: similarly to A, the prediction at long distance is poorer than at short distance.

https://doi.org/10.1371/journal.pcbi.1007630.g008

precisely that of the *in silico* allosteric material (Figs 5 and 6), supporting that the present toy model captures the essence of the DCA limitations in more realistic settings.

Empirical evidence. Recently epistasis was measured in an empirical setting by Salinas and Ranganathan [12] with the aid of deep mutational scan techniques applied to the PDZ domain α2-helix (9 residues), which is part of an allosteric regulatory mechanism controlling ligand binding. Five homologs of PDZ domain were considered in the study. There, epistasis is

$$\Delta\Delta\mathcal{G}_{ij}^{xy} = (\Delta\mathcal{G}_{i}^{x} + \Delta\mathcal{G}_{j}^{y}) - \Delta\mathcal{G}_{ij}^{xy} \tag{5}$$

where $\mathcal{G}$ is the binding free energy and x, y correspond to mutations happening at positions i, j, respectively. DCA inference in [12] was performed on an alignment of 1656 eukaryotic PDZ domains (Poole alignment, see [12]), from where the DCA epistasis prediction $|\Delta\Delta\mathcal{E}_{ij}^{xy}|$ could be directly estimated. The authors then considered averages over mutations x, y and the 5 homologs (we denote them simply as $\Delta\Delta\mathcal{E}_{ij}$ and $\Delta\Delta\mathcal{G}_{ij}$); in Fig 8A we show how well $|\Delta\Delta\mathcal{E}_{ij}|$ predict the experimental energetic couplings $|\Delta\Delta\mathcal{G}_{ij}|$ for pairs of residues (i, j) at distance > 8Å and < 8Å, where distances are measured on the known three-dimensional crystal structure of the PDZ α2-helix and averaged over the 5 homologs. We find a stronger correlation between $|\Delta\Delta\mathcal{G}|$ and $|\Delta\Delta\mathcal{E}|$ for short range pairs (Pearson correlation $\rho = 0.69$), than for long range pairs ($\rho = 0.48$), as the long-range strong epistatic interaction between residues 1 and 8 is not captured by the DCA-inferred energetic couplings, see discussions in [12]. $|\Delta\Delta\mathcal{G}_{18}|$ in Fig 8A is the point at largest $|\Delta\Delta\mathcal{G}|$ in the long-range set. This observation is consistent with our model prediction, shown in Figs 5 and 8B, on the limits of DCA in capturing strong long-range epistasis.

It would be important to test more broadly this predicted effect, which may be possible thanks to the advances of deep mutational scans.

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 [25], where long-range epistasis and functional effects are however found [6, 12, 15], as tested here with the data from [12].

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 [36, 50], in the enzyme TEM-1 β-lactamase [11] and in previous *in silico* evolution work [32]), and are well-predicted by DCA. Specifically, the scaling of epistasis of Eq 2 suggests as approximation $|\Delta\Delta\mathcal{F}_{ij}| \propto \min(\Delta\mathcal{E}_i, \Delta\mathcal{E}_j)$ where $\Delta\mathcal{E}$ are inferred by DCA. Testing this prediction for epistasis patterns empirically could be made possible by the increasing availability of deep mutational scans [12, 51].

Moreover, 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 [52], maximum-entropy probability measures with a nonlinear function of the energy [53], maximum-likelihood inference procedures based on nonlinear functions [54] and, finally, deeper architectures [55, 56]. 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 7 and we have obtained that mutation costs and epistasis can be 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*.

Finally, as a future direction it would be interesting to extend our model by considering the constraint that the protein must fold to operate, in addition to the allosteric constraint considered here. It could be done for example in the spirit of [40] by considering that nodes are amino-acids, and that the stiffness of the spring between two adjacent amino-acids as well as their contribution to the total folding energy depend on the identity of that pair. Although we believe that such a model will lead to similar results as presented here for long-range coupling, it will presumably differ significantly in the statistics of short range ones. In particular, it may capture why 3-body correlations are well described by 2-body correlations in real proteins, and lead to stronger conservation overall [55].

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 [57, 58]. 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 [38, 39, 42]. 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 [39]). 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 Panel A in Fig. A of S1 Text, 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 [37]). As shown in [59], 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 Panel B in Fig. A of S1 Text. 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 Panel C, Fig. A in S1 Text.

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. B of S1 Text), 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. B of S1 Text 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 [60] 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, and thus must be included 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 (see in particular Fig. 4G in [30]).

Supporting information

S1 Text. Supporting information for "Direct coupling analysis of epistasis in allosteric materials".
(PDF)

Acknowledgments

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 and to Rama Ranganathan for making available data from [12].

Author Contributions

Conceptualization: Barbara Bravi, Riccardo Ravasio, Carolina Brito, Matthieu Wyart.

Data curation: Barbara Bravi, Riccardo Ravasio, Carolina Brito.

Formal analysis: Barbara Bravi, Riccardo Ravasio, Carolina Brito, Matthieu Wyart.

Funding acquisition: Matthieu Wyart.

Investigation: Barbara Bravi, Riccardo Ravasio, Carolina Brito, Matthieu Wyart.

Methodology: Barbara Bravi, Riccardo Ravasio, Carolina Brito, Matthieu Wyart.

Software: Barbara Bravi, Riccardo Ravasio, Carolina Brito.

Supervision: Carolina Brito, Matthieu Wyart.

Visualization: Barbara Bravi, Riccardo Ravasio.

Writing – original draft: Barbara Bravi, Matthieu Wyart.

Writing – review & editing: Barbara Bravi, Riccardo Ravasio, Carolina Brito, Matthieu Wyart.

References

1. Jingjing G, Huan-Xiang Z. Protein Allostery and Conformational Dynamics. Chem Rev. 2016; 116(11):6503–6515. https://doi.org/10.1021/acs.chemrev.5b00590
2. Dokholyan NV. Controlling Allosteric Networks in Proteins. Chem Rev. 2016; 116(11):6463–6487. https://doi.org/10.1021/acs.chemrev.5b00544 PMID: 26894745
3. Guarnera E, Berezovsky IN. Allosteric sites: remote control in regulation of protein activity. Curr Opin Struct Biol. 2016; 37:1–8. http://dx.doi.org/10.1016/j.sbi.2015.10.004. PMID: 26562539
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. https://doi.org/10.1002/humu.23231 PMID: 28407397
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. https://doi.org/10.1073/pnas.1620667114 PMID: 28115705
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. http://dx.doi.org/10.1016/j.cub.2014.09.072. PMID: 25455030
7. De Visser JAGM, Cooper TF, Elena SF. The causes of epistasis. Proc R Soc B. 2011; 278(1725):3617–3624. https://doi.org/10.1098/rspb.2011.1537 PMID: 21976687
8. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016; 25:1204–1218. https://doi.org/10.1002/pro.2897. PMID: 26833806
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. https://doi.org/10.1126/science.1142819 PMID: 17702911
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. https://doi.org/10.1126/science.1236862 PMID: 23766324
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. https://doi.org/10.1093/molbev/mst096 PMID: 23676768
12. Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. eLife. 2018; 7:e34300. https://doi.org/10.7554/eLife.34300 PMID: 30024376

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. http://dx.doi.org/10.1038/nsb881.

15. Reynolds KA, McLaughlin RN, Ranganathan R. Hot Spots for Allosteric Regulation on Protein Surfaces. Cell. 2011; 147(7):1564–1575. https://doi.org/10.1016/j.cell.2011.10.049 PMID: 22196731

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. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

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. https://doi.org/10.1126/science.aah4043 PMID: 28104891

18. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. Immunity. 2013; 38(3):606–617. https://doi.org/10.1016/j.immuni.2012.11.022 PMID: 23521886

19. Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, et al. The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing. PLOS Computational Biology. 2014; 10(8):e1003776. https://doi.org/10.1371/journal.pcbi.1003776 PMID: 25102049

20. Barrat-Charlaix P, Figliuzzi M, Weigt M. Improving landscape inference by integrating heterogeneous data in the inverse Ising problem. Scientific Reports. 2016; 6:37812. http://dx.doi.org/10.1038/srep37812. PMID: 27886273

21. 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. https://doi.org/10.1093/molbev/msv211 PMID: 26446903

22. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation Effects Predicted from Sequence Co-Variation. Nature Biotechnology. 2017; 35(2):128–135. https://doi.org/10.1038/nbt.3769 PMID: 28092658

23. Nelson ED, Grishin NV. Inference of epistatic effects in a key mitochondrial protein. Phys Rev E. 2018; 97(062404). https://doi.org/10.1103/PhysRevE.97.062404

24. Poelwijk FJ, Socolich M, Ranganathan R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Bioarxiv. 2017; http://dx.doi.org/10.1101/213835.

25. 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. https://doi.org/10.1073/pnas.1702664114. PMID: 28784799

26. Hemery M, Rivoire O. Evolution of sparsity and modularity in a model of protein allostery. Phys Rev E. 2015; 91(4):042704. https://doi.org/10.1103/PhysRevE.91.042704

27. 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. https://doi.org/10.1073/pnas.1612139114 PMID: 28223534

28. 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. https://doi.org/10.1016/j.bpj.2017.06.043 PMID: 28793211

29. Yan L, Ravasio R, Brito C, Wyart M. Architecture and coevolution of allosteric materials. Proc Natl Acad Sci USA. 2017; 114(10):2526–2531. https://doi.org/10.1073/pnas.1615536114 PMID: 28223497

30. Yan L, Ravasio R, Brito C, Wyart M. Principles for optimal cooperativity in allosteric materials. Biophys J. 2018; 114(12):2787–2798. https://doi.org/10.1016/j.bpj.2018.05.015 PMID: 29925016

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

32. 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. https://doi.org/10.1073/pnas.1716215115

33. Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. Science. 2011; 332(6034):1190–1192. https://doi.org/10.1126/science.1203799 PMID: 21636771

34. 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. http://dx.doi.org/10.1098/rspb.2016.1376. PMID: 27559062

35. Desai MM, Weissman D, Feldman MW. Evolution Can Favor Antagonistic Epistasis. Genetics. 2007; 177:1001–1010. https://doi.org/10.1534/genetics.107.075812 PMID: 17720923

36. Lalić J, Elena SF. Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. Heredity. 2012; 109(2):71–77. http://dx.doi.org/10.1038/hdy.2012.15. PMID: 22491062

37. 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. https://doi.org/10.1088/1361-6633/aa9965 PMID: 29120346

38. Cocco S, Monasson R. Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data. Phys Rev Lett. 2011; 106:090601. https://doi.org/10.1103/PhysRevLett.106.090601 PMID: 21405611

39. Cocco S, Monasson R. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. J Stat Phys. 2012; 147:252–314. https://doi.org/10.1007/s10955-012-0463-4

40. 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. https://doi.org/10.1371/journal.pcbi.1004889. PMID: 27177270

41. Figliuzzi M, Barrat-Charlaix P, Weigt M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? Mol Biol Evol. 2017; 354:1018–1027.

42. 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. https://doi.org/10.1093/bioinformatics/btw328 PMID: 27329863

43. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys Rev E. 2013; 87:012707. https://doi.org/10.1103/PhysRevE.87.012707

44. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999; 286(5438):295–299. https://doi.org/10.1126/science.286.5438.295 PMID: 10514373

45. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009; 138(4):774–786. https://doi.org/10.1016/j.cell.2009.07.038 PMID: 19703402

46. Wang SW, Bitbol AF, Wingreen NS. Revealing evolutionary constraints on proteins through sequence analysis. PLoS Comput Biol. 2019; 15(4):e1007010. https://doi.org/10.1371/journal.pcbi.1007010. PMID: 31017888

47. 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. https://doi.org/10.1073/pnas.1609462113

48. 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. https://doi.org/10.1529/biophysj.104.055004 PMID: 15821164

49. 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. https://doi.org/10.1073/pnas.0510426103 PMID: 16682636

50. Wilke CO, Christoph A. Interaction between directional epistasis and average mutational effects. Proc R Soc B. 2001; 268(1475):1469–1474. https://doi.org/10.1098/rspb.2001.1690 PMID: 11454290

51. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014; 11(801):801–807. http://dx.doi.org/10.1038/nmeth.3027. PMID: 25075907

52. Tubiana J, Cocco S, Monasson R. Learning Protein Constitutive Motifs from Sequence Data. eLife. 2019; 8:e39397. https://doi.org/10.7554/eLife.39397 PMID: 30857591

53. Humplik J, Tkačik G. Probabilistic models for neural populations that naturally capture global coupling and criticality. PLoS Comput Biol. 2017; 13(9):e1005763. https://doi.org/10.1371/journal.pcbi.1005763. PMID: 28926564

54. Otwinowski J, McCandlish DM, Plotkin JB. Inferring the shape of global epistasis. Proc Natl Acad Sci USA. 2018; 115:E7550–E7558. https://doi.org/10.1073/pnas.1804015115 PMID: 30037990

55. Riesselman AJ, Ingraham JB, Marks DS. Deep Generative Models of Genetic Variation Capture the Effects of Mutations. Nature Methods. 2018; 15(10):816–822. https://doi.org/10.1038/s41592-018-0138-4 PMID: 30250057

56. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. Nature Methods. 2019; 16(12):1315–1322. https://doi.org/10.1038/s41592-019-0598-1 PMID: 31636460

57. Nguyen RHCZecchina R, Berg J. Inverse statistical problems: from the inverse Ising problem to data science. Adv Phys. 2017; 66(3):1–65. https://doi.org/10.1080/00018732.2017.1341604

58. 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. https://doi.org/10.1088/1742-5468/aa727d

59. 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. https://doi.org/10.1103/PhysRevE.90.012132

60. 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. http://dx.doi.org/10.1038/nprot.2009.86. PMID: 19561590