# Fluctuations and the limit of predictability in protein evolution

Saverio Rossi,[1,2] Leonardo Di Bari,[3,4] Martin Weigt,[4] and Francesco Zamponi[1,2]

[1] *Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Rome, Italy*
[2] *Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université Paris Cité, 75005 Paris, France*
[3] *DISAT, Politecnico di Torino, Corso Duca degli Abruzzi, 24, I-10129, Torino, Italy*
[4] *Sorbonne Université, CNRS, Laboratory of Computational and Quantitative Biology, 75005 Paris, France*

(Dated: December 4, 2024)

Protein evolution involves mutations occurring across a wide range of time scales [Di Bari et al., PNAS 121, e2406807121 (2024)]. In analogy with other disordered systems, this dynamical heterogeneity suggests strong correlations between mutations happening at distinct sites and times. To quantify these correlations, we examine the role of various fluctuation sources in protein evolution, simulated using a data-driven epistatic landscape. By applying spatio-temporal correlation functions inspired by statistical physics, we disentangle fluctuations originating from the ancestral protein sequence from those driven by stochastic mutations along independent evolutionary paths. Our analysis shows that, in diverse protein families, fluctuations from the ancestral sequence predominate at shorter time scales. This allows us to identify a time scale over which ancestral sequence information persists, enabling its reconstruction. We link this persistence to the strength of epistatic interactions: ancestral sequences with stronger epistatic signatures impact evolutionary trajectories over extended periods. At longer time scales, however, ancestral influence fades as epistatically constrained sites evolve collectively. To confirm this idea, we apply a standard ancestral sequence reconstruction algorithm and verify that the time-dependent recovery error is influenced by the properties of the ancestor itself.

## I. INTRODUCTION

Proteins play a key role in many biological processes that are essential for life. At the same time, they display a huge evolutionary flexibility, in that a large diversity of protein sequences can fold into the same structure and perform the same biological function. Such proteins are called 'homologous' and grouped into the same 'protein family' [1–4]. During evolution, a single ancestral protein belonging to a given family can diversify its amino acid sequence and, given enough time, explore a portion of the 'neutral space' of equivalently fit protein sequences [5]. Nevertheless, a random mutation in the amino acid sequence has a high probability of negatively impacting the structure and functionality of the protein [6, 7]. Such kind of deleterious mutations are eliminated by natural selection, by which certain protein variants are favored over others due to their functional advantages: neutral (or even beneficial) mutations that maintain (or enhance) protein structure, stability, or function are more likely to persist in a population [8].

The picture gets more complex as one takes into account the concept of epistasis [9–20]: the effect of a mutation on the sequence fitness changes with the 'background' in which the mutation takes place, i.e. the amino acids that are present in the other sites of the protein. As a consequence, a mutation that would be deleterious in a particular background sequence can be instead beneficial in another (also called sign epistasis), allowing evolution to explore different pathways.

Understanding and characterizing the impact of epistasis in evolution requires careful experiments and modeling. In particular, recent developments increased substantially the sequence divergence that can be reached by laboratory evolution experiments [21–25], such that many data are becoming available and many more are expected to come soon. Yet, the sequence diversity of natural evolution still remains out of reach of such experiments, thus leaving an unexplored gap in evolutionary time scales. In order to fill this gap, one can simulate the evolution of protein sequences *in silico* [26–30], relying on the data-driven approach that goes under the name of Direct Coupling Analysis (DCA) [31, 32], in which a fitness landscape (analogous to an energy function in the statistical physics vocabulary) is inferred starting from a Multiple Sequence Alignment (MSA) of natural homologs constituting a given protein family [33–37]. The energy function that results from this inference procedure can then be used to assign a probability to each sequence. The resulting landscape is explored by means of a biologically motivated Monte Carlo Markov-Chain (MCMC) algorithm, providing in silico evolutionary trajectories that quantitatively mimic experimental results [26, 27, 29, 30, 38]. Within this framework, it has been recently shown [30] that different sites evolve with widely different time scales, which also depend on the background sequence, due to epistatic interactions. By analogy with similar phenomena observed in disordered physical systems [39–47], this large heterogeneity in time scales suggests the presence of strong and highly collective correlations between sites that slow down the evolution of the sequence. This is the main observation that motivated this work.

In this paper, we measure such correlations to explore how the heterogeneity of evolutionary time scales relates to the epistatic interactions within the ancestral

sequence. We follow the dynamics of the Hamming distance, i.e., the number of accepted mutations with respect to the ancestral sequence. We characterize how this quantity fluctuates (i) between different evolutionary trajectories originating from the same ancestor, and (ii) between different ancestors. We observe a strong dependence of the dynamics on the ancestral sequence for short enough times, over which the first source of fluctuations is found to be subdominant with respect to the second. On the other hand, at long times the stochasticity of evolutionary trajectories takes over, and any memory of the ancestor is lost. This has important consequences, as it allows us to properly quantify the time scale over which it should be possible to reconstruct the ancestral sequence of a certain set of evolutionary trajectories, and how this time scale depends on the epistatic interactions in the ancestor itself. In fact, we find that the amount of epistatically constrained sites in the ancestral sequence determines this time scale: more epistatic sequences leave their trace on evolutionary dynamics for longer times. We then measure the correlations between the evolution of all pairs of sites at the time scale at which such correlations are stronger, finding a different pattern for each ancestral sequence, that is then reflected in the evolutionary dynamics. Finally, we show that the magnitude of epistatic interactions, which controls the stochasticity of the evolutionary trajectories, is also related to the linear response of the evolutionary dynamics to a change in selective pressure (controlled by the temperature in our Monte Carlo simulations). Hence, we show and quantify how more epistatically constrained ancestors lead to a more complex evolutionary dynamics over longer time scales, which is also more sensitive to perturbations of the environment.

## II. METHODS

### A. Modeling evolution in silico

In order to mimic the evolution of protein sequences, we use the DCA model energy as a proxy for the fitness landscape. We start by building an MSA of naturally occurring sequences for each protein family we want to study. From the MSA we infer the parameters (fields and couplings) of the DCA model via Boltzmann Machine learning (bmDCA) [48, 49]. The resulting model assigns a probability $P(A) = \exp[-\mathcal{H}(A)]/Z$ to each sequence $A = (a_1, \ldots, a_L)$, with $L$ the common length of aligned sequences in the MSA and $a_i$ being a symbol that takes 21 possible values corresponding to the 20 natural amino acids plus the gap symbol needed for alignment. According to the statistical physics language, low energy

$$\mathcal{H}(A) = -\sum_{i<j} J_{ij}(a_i, a_j) - \sum_i h_i(a_i) \qquad (1)$$

corresponds to high probability, hence high fitness. In this expression, epistatic interactions between different amino acids are represented via the pairwise couplings $J_{ij}(a_i, a_j)$ that have been found to be crucial in data-driven statistical models of biological sequences, cf. [32, 50]. Because training of these models has become a standard and well-documented procedure [48, 49], we do not give further details here.

Following [26, 27, 30], we consider a fixed initial sequence $A_0 = (a_1^0, \ldots, a_L^0)$ as the ancestor, and we let many trajectories evolve from it in parallel, hence restricting ourselves to a star phylogeny describing an ensemble of independent evolutionary trajectories of common initialization. Our goal is to characterize the statistical properties of the specific ancestral sequence and the sequence space accessible from it in a given evolutionary time. The sequence evolution is simulated by a Monte Carlo dynamics. In this work, we use the Metropolis algorithm acting on amino acids for simplicity, and we verified that more refined sampling strategies taking into account amino-acid accessibility via the genetic code, insertions and deletions [26, 27, 30] produce qualitatively the same results (see SM). As it was previously shown [50], at large times the generated sequences accurately reproduces many statistical features of the natural sequences used for the training, and have the same probability of being biologically functional in a given experimental platform; hence, the model is generative. Here, we are concerned with what happens at short and intermediate times, where the influence of the ancestral sequence is still important.

The results we report mainly concern the DNA-binding domain (DBD) protein family already studied in a similar setting in Ref. [30] and experimentally in Ref. [24], but for some results we also generalize to other protein families, in particular the WW domain (WW), Chorismate Mutase (CM), Aminoglycoside 6-N-acetyltransferase (AAC6), Dihydrofolate reductase (DHFR), Beta Lactamase (BL), and Serine Protease (SP) families. These families have been chosen because they span several chain lengths, and experimental data obtained either from Deep Mutational Scanning or by *in vitro* evolution are available. The procedures to construct the natural MSAs for these families are given in the SM.

### B. Measures of epistasis

By using as input data the MSA of a given protein family, and inferring the fitness landscape parameters via the DCA model, the authors of [30, 51] have been able to roughly classify the sites of any specific protein sequence belonging to the family in three categories: conserved, mutable, and epistatically constrained. In order to do so, they defined two site-mutability metrics.

- The Context-Independent Entropy (CIE) is obtained by computing the empirical frequency $f_i(a)$ of occurrence of amino acid $a$ on site $i$, for every

$a$ and $i$, in the MSA obtained from the natural sequences. This is then used to compute a standard Shannon entropy as

$$\mathrm{CIE}_i = -\sum_{a=1}^{21} f_i(a) \log_2 f_i(a) \ . \tag{2}$$

This quantity measures to what extent a site is variable or conserved across the input MSA.

- The Context-Dependent Entropy (CDE) is defined for site $i$ in sequence $A$ as

$$\mathrm{CDE}_i^A = -\sum_{a=1}^{21} P_i(a|A_{\setminus i}) \log_2 P_i(a|A_{\setminus i}) \ , \tag{3}$$

using the conditional probability $P_i(a|A_{\setminus i}) = P(a_i = a|A_{\setminus i})$ of having amino acid $a$ on site $i$ given the rest of the sequence $A_{\setminus i} = (a_1, ..., a_{i-1}, a_{i+1}, ..., a_L)$. This quantity cannot be extracted directly from the input MSA and needs to be obtained from the model parameters; due to the epistatic couplings in the energy $\mathcal{H}(A)$ it actually differs from the CIE. As indicated explicitly, the $\mathrm{CDE}_i^A$ depends on the site $i$ and the sequence $A$, hence on the context in which site $i$ finds itself. As a matter of fact, this metric quantifies the local mutability of a site, i.e. its mutability within a certain reference-sequence context $A_{\setminus i}$.

A large value of CIE or CDE means that many mutations are allowed, while a small value means that only one or a few amino acids are tolerated.

In terms of these quantities, a site $i$ in a given background sequence $A$ can thus be classified as follows.

- *Mutable sites* have a large CDE and a large CIE. These sites can tolerate many mutations, both in the natural MSA and in the considered background.

- *Conserved sites* have a small CDE and a small CIE. These sites do not tolerate mutations, neither in the natural MSA nor in the considered background.

- *Epistatically constrained sites* have a small CDE and a large CIE. These sites display a large variability in the natural alignment, but only because the rest of the sequence is mutating at the same time. In fact, in the considered background, only one or a few amino acids are tolerated.

Note that it is very rare that a site has a CDE larger than the CIE, because generally speaking, fixing the background reduces the number of mutations that can be tolerated [51]. We also stress that this classification depends on the background $A$, and Ref. [30] has shown that epistatically constrained sites in a background can be mutable in another background and viceversa (while conserved sites tend to remain so in all backgrounds).

Furthermore, and most importantly for the present work, Ref. [30] has considered a given background $A_0$ as the ancestral sequence, and starting from it has performed many parallel evolutions in silico, looking at how each site diversifies in the library of mutants obtained after a certain evolutionary time. It was found that mutable sites evolve very rapidly and quickly reach their asymptotic mutability, i.e. the CIE. Conserved sites remain so at any time during evolution, hence do not display interesting dynamics. Epistatically constrained sites, instead, are very conserved at short evolutionary times, due to the epistatic constraints in the background of the ancestor, which cause a small CDE. But, as soon as the background sequence mutates enough, they can tolerate more mutations, asymptotically reaching the large CIE that they display in the natural alignment. Their mutation, however, is contingent on several mutations happening in the rest of the sequence, which often require a rather long evolutionary time to take place [30].

Along with the classification of sites discussed above, we also introduce a global measure of the level of epistasis in a given sequence $A$. Because each sequence has a different set of variable and epistatically constrained sites, we measure its overall level of epistatic constraints by averaging the CDE over sites,

$$\mathrm{CDE}^A = \frac{1}{L} \sum_{i=1}^{L} \mathrm{CDE}_i^A \ . \tag{4}$$

Hence, a more epistatically constrained sequence will display a lower $\mathrm{CDE}^A$.

### C. Dynamical fluctuations

Building on the previous work described above, and inspired by works on disordered systems [39–47], we devised a protocol (sketched in Fig. 1) to quantify how the amount of epistasis in the ancestral sequence impacts the subsequent evolution. We consider a series of ancestral sequences, taken from the natural MSA, let us call them $A_0$, $B_0$, $C_0$, etc. Each ancestor has a different level of epistatic constraints, as measured by the average CDE introduced in Eq. (4), as illustrated in Fig. 1a.

Next, we consider a star phylogeny of independent MCMC evolutionary trajectories, all starting from the same ancestor and evolving in parallel (Fig. 1b). Hence, for each ancestor, we construct, as a function of evolutionary time, a MSA of descendants that mimic those obtained through in vitro evolution experiments. As a measure of diversity, we focus on the evolution of the Hamming distance, between the ancestor $A_0$ and the MSA of evolving sequences at time $t$, $A_t = (a_1^t, \cdots, a_L^t)$, defined as

$$H^{A_0}(t) = H(A_t, A_0) = \frac{1}{L} \sum_{i=1}^{L} (1 - \delta_{a_i^t, a_i^0}) \ , \tag{5}$$

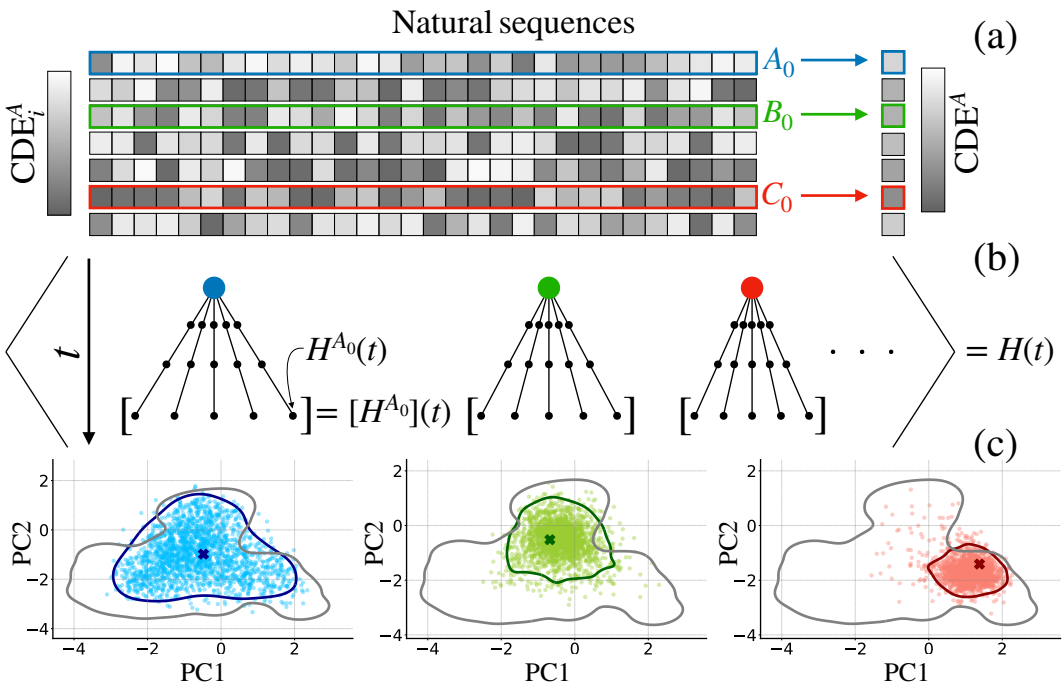


Figure 1. Schematic representation of (a) a MSA in which each site is colored based on its mutability as measured by the $\mathrm{CDE}_i^A$ and (b) the protocol of our evolutionary dynamics with the relevant averages. (c) Depending on the epistatic constraints acting on the ancestral sequence and for equal evolutionary times, we observe a different level of diversity of the evolved sequences (colored), corresponding to a different exploration of the functional sequence space (grey).

introducing the Kronecker delta symbol $\delta_{a_i^t, a_i^0}$, which is zero if site $i$ is mutated between the two sequences $A_t$ and $A_0$, and one otherwise. This quantity corresponds to the 'overlap' in the disordered systems literature and to the number of accepted mutations in evolution. Other measures could be considered as well, but we focus on this one for simplicity. For a given ancestor, simulating many parallel evolutionary trajectories, we thus obtain a set of realizations of the random variable $H^{A_0}(t)$.

As illustrated through a projection in PCA space in Fig. 1c, we observe that the level of epistasis in the ancestor determines the size of the portion of the fitness landscape explored by evolution. For comparable evolutionary time, ancestors with more epistatically constrained sites (e.g., sequence $C_0$ in Fig. 1) lead to a less diverse set of evolved sequences. Conversely, ancestors with less epistatically constrained sites (e.g., $A_0$ in Fig. 1) lead to a widely diverse set of evolved sequences.

To make this observation quantitative, we characterize the statistical properties of the resulting MSAs, and how they depend on the ancestor, by introducing two distinct averages and corresponding fluctuations, inspired from the disordered systems literature [44, 46, 47] and illustrated in Fig. 1b. Recall that the number of accepted mutations $H^{A_0}(t)$ at evolutionary time $t$ for fixed ancestor $A_0$ is a random variabile, whose realizations depend on the stochasticity of the evolutionary trajectory.

- First, we consider averaging over many evolutionary trajectories for fixed ancestor. We denote this average as $[\cdots]$. The variance of $H^{A_0}(t)$ can then be defined as

$$\chi_{\mathrm{dyn}}^{A_0}(t) = [H^2] - [H]^2 \ , \tag{6}$$

where the dependence of $H^{A_0}(t)$ on $A_0$ and $t$ is omitted to simplify the notation. The quantity $\chi_{\mathrm{dyn}}^{A_0}(t)$ depends on the ancestor $A_0$ and on time $t$.

- Second, we consider the average over different ancestors, which we define as $\langle \cdots \rangle$. In particular, we consider the average number of mutations $[H^{A_0}(t)]$ for a given ancestor, that measures the average diversity of the evolved MSA, and we measure how this quantity fluctuates from ancestor to ancestor via the variance

$$\chi_{\mathrm{bg}}(t) = \langle [H]^2 \rangle - \langle [H] \rangle^2 \ , \tag{7}$$

where again the dependence of $[H] = [H^{A_0}(t)]$ on $A_0$ and $t$ is omitted to simplify the notation. This quantity is the variance of $[H]$ associated to the fluctuations in the background of the ancestral sequence $A_0$, hence the suffix 'bg'.

Note that the total variance of $H^{A_0}(t)$ over both sources of randomness, i.e. the random choice of ancestor and the stochasticity of mutations along the evolutionary trajec-

tory, can be decomposed as

$$\begin{aligned}\chi_{\rm tot}(t) &= \langle [H^2]\rangle - \langle [H]\rangle^2 \\ &= \langle [H^2]\rangle - \langle [H]^2\rangle + \langle [H]^2\rangle - \langle [H]\rangle^2 \\ &= \chi_{\rm bg}(t) + \chi_{\rm dyn}(t)\ ,\end{aligned} \tag{8}$$

where

$$\chi_{\rm dyn}(t) = \langle \chi^{A_0}_{\rm dyn}(t)\rangle \tag{9}$$

is the average of the ancestor-dependent $\chi^{A_0}_{\rm dyn}(t)$ over $A_0$. Using these quantities, which are called 'dynamical susceptibilities' in the physics of disordered systems, we can thus quantify the relative importance of the ancestor and the evolutionary noise in determining the diversity of the resulting MSAs at a fixed evolutionary time. In the SM we also study some alternative definitions of susceptibilities, based for example on the Hamming distance between two chains starting from the same ancestor. However, because our MCMC dynamics is time-reversible, these two quantities are related.

## III. RESULTS

### A. Mutational dynamics and its fluctuations

The dynamical evolution of the Hamming distance $H^{A_0}(t)$ (number of accepted mutations with respect to the ancestor) is displayed as a function of the number of Monte Carlo sweeps in Fig. 2, with a sweep corresponding, on average, to one attempted mutation per site. More specifically, Fig. 2a shows the average $[H^{A_0}(t)]$

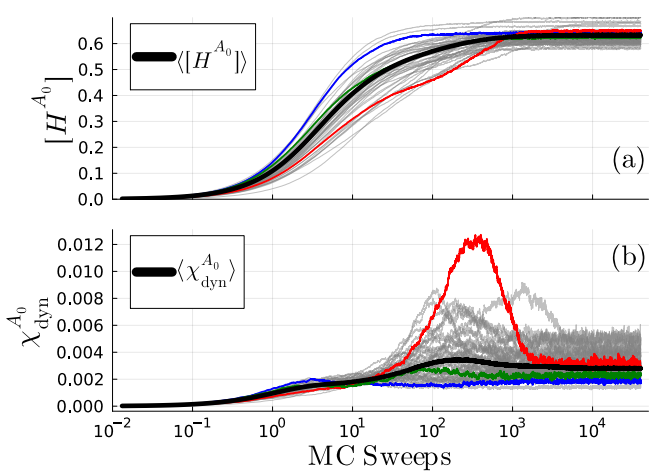


Figure 2. Average (a) and variance (b) of the Hamming distance between the evolving sequence and the ancestor estimated using $10^3$ independent trajectories, for many different ancestors (grey lines). The thick black line represents the average over ancestors, i.e. $\langle [H^{A_0}(t)]\rangle$ and $\chi_{\rm dyn}(t) = \langle \chi^{A_0}_{\rm dyn}(t)\rangle$, over 200 ancestors. The green, blue and red lines highlights the same specific choices of ancestor used in Fig. 1.

over a star phylogeny of $10^3$ parallel evolutions, for 200 choices of the ancestor $A_0$ taken at random (with weights, see SM) from the DBD protein family. Fig. 2b reports the variance $\chi^{A_0}_{\rm dyn}(t) = [H^2] - [H]^2$ defined in Eq. (6) over the same phylogeny, and for the same initial sequences as in Fig. 2a.

Before describing three interesting cases that we highlighted with colors, we discuss the general traits of such dynamics. The average Hamming distance and its variance are both initially null because all chains start from the same ancestral sequence. Because the DCA model is generative and the evolutionary dynamics respects detailed balance, at large times we expect the simulated sequences to be independent samples from the DCA model, which then reproduce statistical features of the natural ones used for training. This means that the average of $H^{A_0}(t)$ and its variance converge, respectively, to the average and the variance of the Hamming distance between the chosen ancestral sequence and the rest of the natural ones. These values can vary significantly (the final average Hamming distance in Fig. 2a varies from roughly 0.55 to 0.65) depending on how close $A_0$ is to the other sequences in the natural MSA. The different curves, corresponding to distinct ancestors $A_0$, show a wide range of behaviors and time scales: some trajectories reach the steady state rapidly while others take much longer, displaying intermediate plateaus as a hallmark of epistasis. In some cases $\chi^{A_0}_{\rm dyn}(t)$ displays a peak and then decreases, while in others the equilibrium value is reached in a monotonic way. The large value reached by $\chi^{A_0}_{\rm dyn}(t)$ for some initial sequences implies strong fluctuations between distinct evolutionary trajectories, making it hard to predict the dynamics.

In both panels of Fig. 2, we highlighted with colors some representative curves. In particular, the green curve corresponds to an ancestor that behaves in a rather 'typical' way, close to the average. The blue curve corresponds to an ancestor that has less epistatically constrained sites, hence the dynamics is faster and less heterogeneous. Finally, the red curve corresponds to a highly epistatically constrained ancestor, which leads to a slower dynamics with an intermediate plateau in $[H^{A_0}(t)]$, and a large peak in $\chi^{A_0}_{\rm dyn}(t)$. The same three ancestors have been used to construct the PCA plots in Fig. 1 that correspond to time $t = 50$ MC Sweeps.

### B. Dynamical heterogeneity across families

In Fig. 3 we generalize our results to many protein families, chosen to have different ranges of sequence length, MSA depth (number of natural sequences) and equilibration time scales. These families are also interesting because for some of them, experimental data from Deep Mutational Scans and in vitro evolution are available.

For compactness, we only display the average dynamical fluctuations, with $\chi_{\rm tot}$ corresponding to the total variance of $H$, $\chi_{\rm dyn}$ to the variance due to the stochatisticy

of the evolution (averaged over the ancestor), and $\chi_{\rm bg}$ to the variance due to the choice of ancestor. All the $\chi_{\rm tot}$ curves display a similar behavior, increasing from zero to a maximum value that is maintained over one or more decades, before finally decreasing to a smaller value corresponding to equilibrium. What matters most to us, however, is the relative importance of the two terms $\chi_{\rm dyn}$ and $\chi_{\rm bg}$ in which $\chi_{\rm tot}$ can be decomposed. The key observation is that background-related fluctuations dominate over dynamical ones at short evolutionary times, while the inverse is true at larger times. This means that, at least over time scales for which $\chi_{\rm bg} \gg \chi_{\rm dyn}$, we can hope to reconstruct with good accuracy the ancestral sequence from which a set of evolutionary trajectories started. Conversely, at larger time scales, the dynamical noise contribution dominates and the trajectory-to-trajectory fluctuations are large enough to hide the signal coming from the ancestral sequence, precluding the possibility to reconstruct it. This behavior is observed in all the protein families that we tested, with the exception of the WW domain in which $\chi_{\rm bg}$ never grows to larger values than $\chi_{\rm dyn}$. We attribute this difference to the small size of the WW domain, which prevents the sequences to accumulate enough epistatic interactions. However, both the time scales and the relative importance of the two terms contributing to the total susceptibility vary greatly from family to family (Fig. 3). In DBD, AAC6, and DHFR the peak of $\chi_{\rm tot}$ is reached quite early in the dynamics and the two contributions $\chi_{\rm dyn}$ and $\chi_{\rm bg}$ are almost non-overlapping. On the other hand, for CM, BL, and SP (the longer families we analyzed) the peak occurs much later due to the two contributions having a significant overlap. Overall, these results point to an increasing complexity of the fitness landscape with the length of the sequence, as intuitively expected.

For a given family and a typical ancestor, the time scale at which $\chi_{\rm dyn}$ and $\chi_{\rm bg}$ cross, the former becoming larger than the latter, defines a characteristic evolutionary time scale, around which the memory of the ancestor is lost.

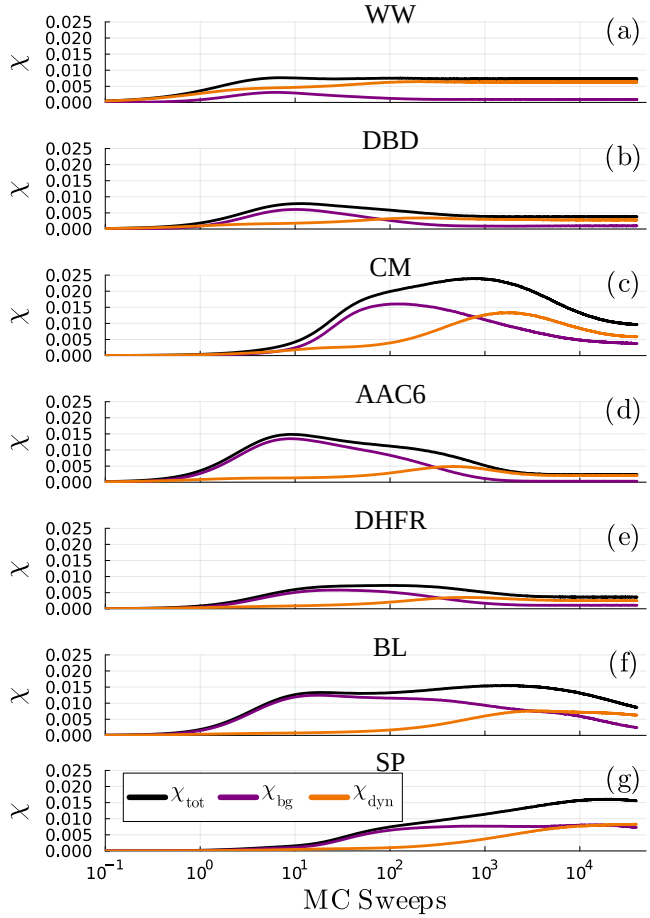


Figure 3. Evolution of the different susceptibilities $\chi_{\rm tot}$, $\chi_{\rm bg}$, and $\chi_{\rm dyn}$, for different protein families. (a) WW domain, (b) DNA-binding domain (DBD), (c) Chorismate Mutase (CM), (d) AAC6, (e) DHFR enzyme, (f) Beta-lactamase (BL), and (g) Serine Protease (SP).

### C. Epistatic constraints and dynamical fluctuations

Up to this point, we discussed the time dependence of the different contributions to the total fluctuations. We argued that, as long as the dominant contribution to $\chi_{\rm tot}$ comes from $\chi_{\rm bg}$, the ancestral sequence is strongly related to the evolving ones. We now want to better understand the transition to the final regime in which dynamical fluctuations, i.e., $\chi_{\rm dyn}$, dominate.

In Fig. 4 we show, for the three ancestral sequences highlighted in Fig. 2, how the dynamical part of the susceptibility is strongly related to the epistatically constrained sites (defined as in Sec. II B and in Ref. [30]). To make this relation clear, we first checked that the time scale at which the dynamical susceptibility reaches its peak is compatible with the time scale at which epistatically constrained sites evolve. For a given ancestor $A_0$, we consider the five most epistatically constrained sites, i.e. those with the largest $\mathrm{CIE}_i - \mathrm{CDE}_i^{A_0}$. For these sites, we consider at time $t$ the frequency of appearance of amino acid $a$, $f_i^t(a)$, in the set of sequences that evolved from $A_0$, and from it we compute a time-dependent entropy

$$S_i(t) = -\sum_{a=1}^{21} f_i^t(a) \log_2 f_i^t(a) \ . \qquad (10)$$

In Fig. 4, $\chi_{\rm dyn}^{A_0}(t)$ (colored curve) is superposed to $S_i(t)$ of the five sites (gray curves), as a function of evolutionary time $t$. We observe that the peak of $\chi_{\rm dyn}^{A_0}$ is reached just before the equilibration of the epistatically constrained sites, i.e. the time at which $S_i(t)$ approaches the $\mathrm{CIE}_i$. The time of equilibration of epistatically constrained sites also increases from Fig. 4a to Fig. 4c, similarly to what happens for the equilibration of the Hamming distance in the same three sequences in Fig. 2. Epistasis does not only affect the time scale at which the peak of $\chi_{\rm dyn}^{A_0}$ is reached, but also the intensity of the peak, which means

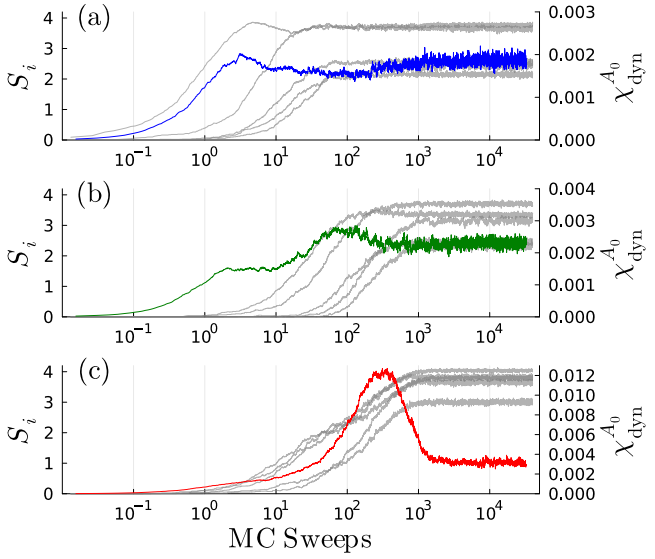


Figure 4. Evolution of the entropy of the most epistatically constrained sites for a specific ancestral sequence of DBD plotted together with $\chi_{\rm dyn}^{A_0}$ of the same sequence. (a), (b), and (c) are for the blue, green, and red sequences in Fig. 2, respectively.

that more epistatic ancestors lead to a more heterogeneous dynamics at intermediate times, when the epistatic sites mutate.

To quantify these observations, in Fig. 5 we consider the same set of families as in Fig. 3, namely WW, DBD, CM, AAC6, DHFR, BL, and SP, spanning a wide range of sequence length (from $L = 31$ in WW to $L = 254$ in SP). For each family, we consider a set of ancestors $A_0$ and we report a scatter plot of the following quantities, each $A_0$ being a point:

- the value of $\mathrm{CDE}^{A_0}$ defined as in Eq. (4),
- the maximum value $\max_t \chi_{\rm dyn}^{A_0}(t)$, indicated as $\max(\chi_{\rm dyn}^{A_0})$ for simplicity,
- and the time $t_{90}$ at which the average $[H^{A_0}]$ over the trajectories reaches 90% of the equilibrium ($t \to \infty$) value.

We observe that both $\max(\chi_{\rm dyn}^{A_0})$ and $t_{90}$ markedly increase as $\mathrm{CDE}^{A_0}$ decreases. Hence, a 'highly epistatic' sequence with a small value of $\mathrm{CDE}^{A_0}$ (with respect to the typical value of the family) has many sites with low context-dependent entropy, which cannot mutate at the beginning of evolution, leading to a slower and more heterogeneous overall dynamics. This is for example the case of the red sequence in Fig. 2, for which we have $\mathrm{CDE}^{A_0} = 1.45$, while for the blue one $\mathrm{CDE}^{A_0} = 2.11$. The three sequences considered in Fig. 2 are indicated with stars of the corresponding colors in Fig. 5b.

Epistatically constrained sites also carry the specific signature of the ancestral sequence, because the conserved sites are roughly the same for every sequence of

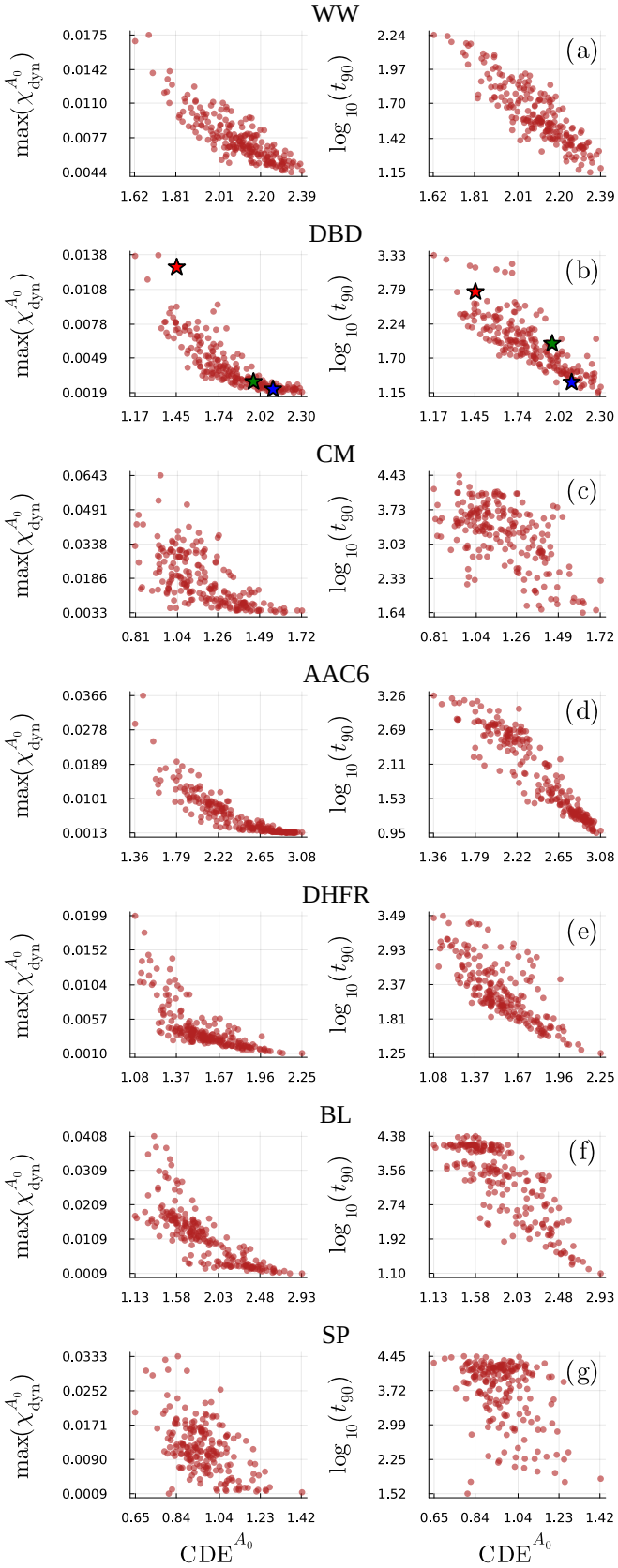


Figure 5. Scatter plot of $\mathrm{CDE}^{A_0}$ against the maximum of $\chi_{\rm dyn}^{A_0}$ (left) and against $\log_{10}(t_{90})$ (right) for each protein family we considered. For each family, the points correspond to 200 sequences $A_0$ extracted from the natural MSA with weights (see SM). In (b) colored stars are used to highlight the sequences that correspond to the blue, green, and red curves in Fig. 2.

the family and the variable ones carry little to no information. These results thus suggest that tracing back an evolutionary trajectory to its ancestral sequence becomes more difficult as one approaches the peak of $\chi_{\rm dyn}$, because this is when the epistatic sites, i.e. the sites that carry information about that initial sequence, start to mutate.

### D. Cooperative mutational dynamics

We have established that, when epistatic sites start to mutate, $\chi_{\rm dyn}^{A_0}$ reaches its peak, and that the intensity of this peak is correlated with the amount of epistatically correlated sites in the ancestral sequence. We now want to show that the peak of $\chi_{\rm dyn}^{A_0}$ is due to dynamical correlations between different sites, caused by the strong interaction between those sites and the context. These correlations result in a cooperative mutational process, in which epistatically constrained sites can only mutate because other such sites mutate, leading to an avalanche of mutations.

To precisely quantify this effect, one can interpret the dynamical susceptibility as the sum of a site-site dynamical correlation function [45]. The Hamming distance is defined in Eq. (5). Inserting its expression into Eq. (6), we obtain

$$\chi_{\rm dyn}^{A_0}(t) = \frac{1}{L^2}\sum_{ij} G_{ij}^{A_0}(t) \ , \tag{11}$$

with

$$G_{ij}^{A_0}(t) = [\delta_{a_i^t,a_i^0}\delta_{a_j^t,a_j^0}] - [\delta_{a_i^t,a_i^0}][\delta_{a_j^t,a_j^0}] \ , \tag{12}$$

recalling that the Kronecker delta symbol $\delta_{a_i^t,a_i^0}$ is zero if site $i$ is mutated between the two sequences $A_t$ and $A_0$, and one otherwise. The matrix $G_{ij}^{A_0}$ is a time-dependent correlation matrix between sites $i$ and $j$. It is high when the sites are dynamically correlated, i.e. if $a_i^t$ is different from $a_i^0$, then $a_j^t$ is likely different from $a_j^0$ as well, and vice versa. It is low when the sites mutate independently. Because $\chi_{\rm dyn}^{A_0}$ is the sum of $G_{ij}^{A_0}$ over all pairs $(i,j)$, a large value of $\chi_{\rm dyn}^{A_0}$ implies that many $(i,j)$ are strongly correlated. Furthermore, by looking at this correlation matrix one can infer which pairs of sites mutate in a correlated way during evolution.

The value of the off-diagonal part of $G_{ij}^{A_0}(t^*)$, computed at the time $t^*$ at which a peak in the dynamical susceptibility is observed, is shown in Fig. 6 for six different starting sequences in the DBD family.

Fig. 6a corresponds to an ancestor (red in Fig. 2) for which a large peak and a longer time to equilibrate are observed in the evolution of $\chi_{\rm dyn}^{A_0}$. The large value of $\chi_{\rm dyn}^{A_0}$, which is just the sum of all matrix elements, implies a large number of positively correlated sites. Fig. 6e corresponds to an ancestor (green in Fig. 2) for which $\chi_{\rm dyn}^{A_0}$ has a smaller peak before saturating. We notice

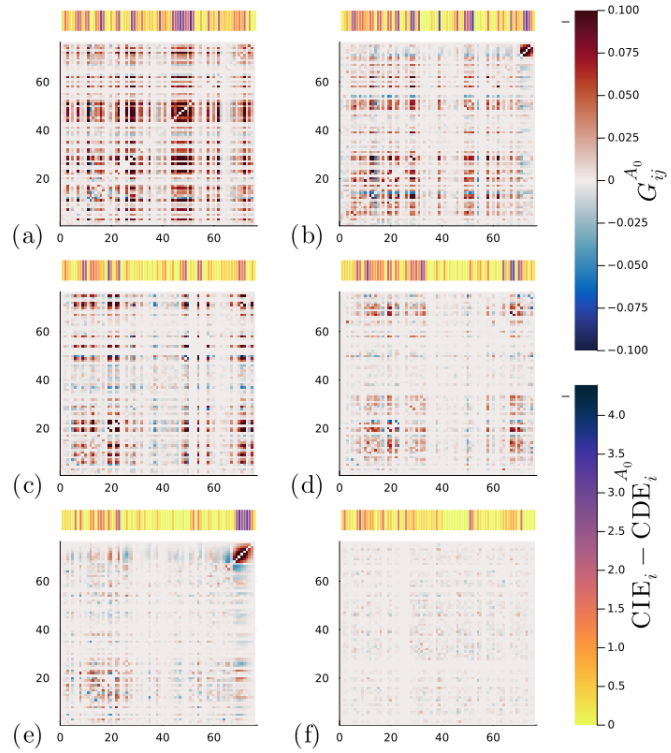


Figure 6. Correlations between sites at time $t^*$ at which $\chi_{\rm dyn}^{A_0}(t)$ reaches its maximum (with a cutoff at $t_{\rm th} = 1000$ sweeps), for different ancestral sequences. The bar above each snapshot encodes the strength of epistatic constraints for the sites in the ancestral sequence.

in this case that $G_{ij}^{A_0}(t^*)$ still displays some strong correlations, but the majority of the sites are uncorrelated. Finally, Fig. 6f corresponds to an ancestor (blue in Fig. 2) for which $\chi_{\rm dyn}^{A_0}$ saturates to a small value. In this case correlations are almost absent at $t^*$, which means that essentially all sites mutate independently. It is thus evident how different initial sequences give rise to different patterns in the $G_{ij}^{A_0}$ matrix. At large times the chains reach equilibrium and are almost statistically indistinguishable from natural sequences. This means that $G_{ij}^{A_0}(t\to\infty) \approx G_{ij,\rm nat}^{A_0}$, with

$$G_{ij,\rm nat}^{A_0} = \langle \delta_{a_i^{\rm nat},a_i^0}\delta_{a_j^{\rm nat},a_j^0}\rangle_{\rm nat} - \langle \delta_{a_i^{\rm nat},a_i^0}\rangle_{\rm nat}\langle \delta_{a_j^{\rm nat},a_j^0}\rangle_{\rm nat},$$

where $\langle \dots \rangle_{\rm nat}$ is the average computed over the natural sequences. We checked that this is indeed the case.

In section III C we showed that the presence of a peak in $\chi_{\rm dyn}^{A_0}$ is correlated to $\mathrm{CDE}^{A_0}$, i.e. to the amount of epistatically constrained sites in the ancestral sequence. In this section, we show that these sites also display large dynamical correlations between themselves. The bar above each plot in Fig. 6 is shaded to represent $\mathrm{CIE}_i - \mathrm{CDE}_i^{\rm A}$, capturing the information about site $i$ encoded in the background sequence $A_{\backslash i}$. This coloring reflects the degree to which each site is constrained by epistatic interactions. The conserved and variable sites

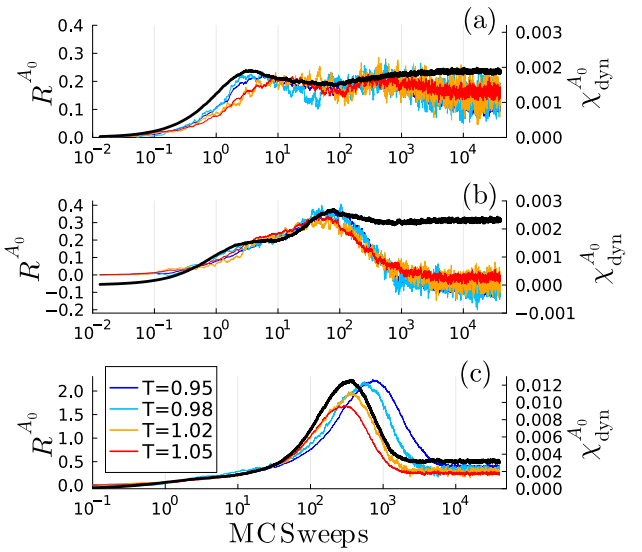


Figure 7. Evolution of the dynamical susceptibility $\chi^{A_0}_{\rm dyn}(t)$ compared with the linear response of the average Hamming distance to a change of temperature $R^{A_0}(t)$. The three panels correspond to the three ancestral sequences highlighted in Fig. 2, (a) blue, (b) green, (c) red.

are highlighted by a bright yellow, while the epistatically constrained ones are indicated by a dark blue. As expected, we see good agreement between these residues and the ones that give a large contribution to $\chi^{A_0}_{\rm dyn}$. Hence, we conclude that epistatically constrained sites mutate cooperatively around the time scale corresponding to the peak of $\chi^{A_0}_{\rm dyn}$.

### E. Response to environmental variation

In the literature on disordered physical systems, it has been established that the dynamical correlations discussed in sections III C and III D are related to the linear response of the dynamics to a change in temperature through a kind of fluctuation-dissipation relation [43].

In the present context, the 'temperature' is a parameter that controls the probability with which mutations are accepted. High temperature corresponds to all mutations being accepted (very low selection), while low temperature corresponds to only the most beneficial mutations being accepted (very strong selection, i.e. directed evolution). The value of $T = 1$ corresponds to the conditions at which the DCA model is trained on natural sequences, hence $T \sim 1$ corresponds to a neutral drift dynamics during which the neutral space of sequences that have comparable fitness to natural ones is explored [27]. Indeed, Ref. [27] has shown that two distinct in vitro evolution experiments, realized with selection pressures comparable to the natural one, can be described by fitting the temperature in a range slightly above $T \sim 1$.

Hence, in our modeling framework, a change of temperature corresponds to a change of selection strength. Following Ref. [43], we check whether the dynamical susceptibilities are related to the response of the average dynamics with respect to the temperature variation. To estimate such linear response, we consider two evolutionary dynamics starting from the same ancestor, one at temperature $T \neq 1$ and the other at temperature $T = 1$, and we compare the average number of accepted mutations $[H^{A_0}(t;T)]$ between the two evolutions. The dynamical linear response is given by

$$R^{A_0}(t) = \lim_{T\to 1} \frac{[H^{A_0}(t;T)] - [H^{A_0}(t;T=1)]}{T-1} \ , \qquad (13)$$

and we emphasize that this quantity depends on the ancestor $A_0$ and on time $t$. In Fig. 7, we compare the time dependence of $\chi^{A_0}_{\rm dyn}(t)$ with $R^{A_0}(t)$ for the same three ancestors as in Fig. 2. Note that we cannot take the limit $T \to 1$ due to statistical noise, and we estimate $R^{A_0}(t)$ by observing that the curves for several values of $T$ close to 1 are almost superimposed to each other. Fig. 7 shows that the fluctuation-dissipation relation seems to hold quite well, such that $\chi^{A_0}_{\rm dyn}(t) \propto R^{A_0}(t)$, at least for times $t$ smaller than the peak of both quantities.

The results in Fig. 7 suggest that more epistatically constrained ancestral sequences, which display a stronger peak in the dynamical susceptibility, will also display a stronger response of the evolutionary dynamics to a small change in selection strength. Such response is stronger around the time of the peak in $\chi^{A_0}_{\rm dyn}(t)$, which is also the time at which epistatically constrained sites mutate cooperatively, see section III D. In other words, *sequences with more epistatically constrained sites are more sensitive to environmental changes*. In the SM, we give an additional analytical argument based on the short-time dynamics, that supports the same conclusion.

### F. Limits of ancestral sequence reconstruction

In an evolutionary process, predictions can be made in two directions: either starting from the ancestral sequence and predicting its future evolution over a given time; or starting from a set of evolved sequences and attempting to infer the ancestral sequence from which the evolution began. An attempt at predicting future evolution has been made in Ref. [27], where however the prediction was limited to small evolutionary time scales. Because we have repeatedly hinted that our analysis can provide insight into the time scale over which reconstructing the ancestral sequence of an evolutionary process is possible, before concluding, we consider here the ancestral sequence reconstruction problem more explicitly.

For each of the three sequences we analyzed in Fig. 2, we simulate 200 independent evolutionary trajectories, thus obtaining several MSAs of 200 evolved sequences at different evolutionary times $t$. Then, we use the FastML online tool [52] to infer the ancestral sequence starting from those MSAs and using a star-shaped phylogenetic

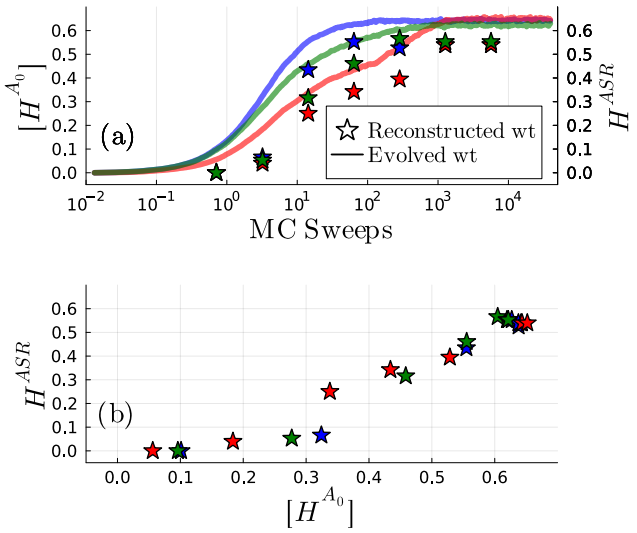


Figure 8. (a) The full lines represent the evolution of the average Hamming distance for the same three ancestors as in Fig. 2. The star symbols represent the Hamming distance $H^{\mathrm{ASR}}$ between the ground truth ancestor and the one reconstructed using the simulated MSAs at different times. (b) Hamming distance $H^{\mathrm{ASR}}$ between the reconstructed ancestor and the ground truth as a function of the average distance $[H^{A_0}(t)]$ between the ancestor and the MSA used for the reconstruction. The colors refer to the same sequences of panel (a).

tree. We then check how close the reconstructed sequence is to the actual ancestor.

The results are shown in Fig. 8a, where we plot the evolution of the average Hamming distance for the same sequences as in Fig. 2, and we compare it with the Hamming distance (i.e. the percentage of errors) between the ground truth ancestor and the reconstructed one. At the beginning of the dynamics the FastML tool is able to reconstruct perfectly the original sequence in all of the three cases. This is not surprising, as at short times few mutations appear. The situation is similar at large times, where the MSAs generated from the three starting sequences are statistically indistinguishable and the FastML tool performs equally bad in all of them. Thus, it is more insightful to look at intermediate times, where a significant difference between the evolving sequences is present. After $\sim 10$ sweeps the variable sites begin to mutate and the ancestral sequence reconstruction tool cannot reproduce the initial sequence perfectly. The leap in the error percentage with respect to the previous point is however larger for the blue sequence, which is the one with fewer epistastic sites, while the red one is still recovered reasonably well. This difference is present for the subsequent points as well, until the error saturates. In Fig. 8b we show that the Hamming distance (percentage of errors) of the sequence reconstruction results grows roughly proportionally to the average sequence divergence of the evolutionary trajectories for all of the three studied sequences, but more so for the more epistatic red sequence. We conclude that, at least with the FastML tool, the possibility of reconstructing ancestral sequences is mostly a matter of divergence, with more errors accumulating when using datasets of more diverged sequences.

Our analysis shows that the amount of diversity at a given evolutionary time depends strongly on the ancestor, due to highly non-trivial epistatic dynamical correlations. More epistatically constrained ancestors give rise to less diversity, thus allowing for reconstruction over longer evolutionary times (Fig. 8a). Yet, at comparable amount of diversity, more epistatic ancestors are more difficult to reconstruct (Fig. 8b), at least using the FastML algorithm that neglects correlations between sites [53]. These results confirm once more that the evolutionary time limit for predictability is linked to the presence of epistatic sites that carry information about the starting sequence only up to a certain sequence divergence. We further observe that more advanced tools (yet to be fully developed) for ancestral sequence reconstruction might exploit the epistatic correlations to achieve a better reconstruction [54], which is not done in FastML and similar algorithms. We expect that such tools would perform better on more epistatic ancestors such as the red one in Fig. 8.

## IV. CONCLUSIONS

In this study, we explore the role of disorder and stochasticity in protein evolution, focusing on the interplay between epistasis, site correlations, and sequence-dependent fluctuations. We simulated evolution in silico using DCA to model the fitness landscape, and a MCMC dynamics to mimic the process of mutation and selection, a methodology that has been validated in previous work [26, 27, 30]. By employing tools from statistical physics [39–47], we quantified how the ancestral amino acid sequence significantly influences early evolutionary dynamics, as measured by the dynamical susceptibility. Our main results are the following.

- Our analysis shows that, during the initial phase of evolution, there is very significant heterogeneity in the dynamics, depending on the choice of ancestral sequence. Some ancestral sequences lead to a smooth evolution where each site mutates almost independently of the others, while others lead to a more complex dynamics characterized by intermediate plateaus and significant fluctuations (Sec. III A).

- We have shown that, for a variety of protein families, the noise arising from the starting sequence dominates over stochastic evolutionary fluctuations, implying that one can, in principle, trace the evolutionary trajectory back to its origin (Sec. III B). This effect is more or less pronounced,

depending on the family, and our tools allow one to quantify it. However, as time progresses and epistatically constrained sites evolve, this traceability is lost due to the growing influence of the mutational stochasticity.

- The heterogeneity of the initial evolutionary dynamics can be traced back to the amount of epistatic constraints in the ancestral sequence. We introduced a quantity, the average over sites of the context-dependent entropy, and we have shown that this quantity is strongly correlated with the time at which dynamical heterogeneity reaches its peak and with the strength of the fluctuations at the peak (Sec. III C).

- The amplitude of the global fluctuations can be expressed in terms of a sum of pairwise dynamical correlations between pairs of residues. More epistatically constrained ancestral sequences show groups of residues that mutate collectively, and those patterns can be identified from the correlation matrix $G_{ij}^{A_0}(t)$ we introduced in Sec. III D. The observed correlations between sites thus reflect a complex epistatic landscape, where certain residues evolve in a highly context-dependent manner.

- We demonstrated that these epistatic sites, which evolve over long time scales, significantly affect the response of the dynamics to a change in environmental conditions. More epistatically constrained sequences lead to a larger response to a change in environment, suggesting potential implications on sequence evolvability imposed by environmental pressures, such as antibiotic concentration (Sec. III E).

- Finally, we presented a preliminary study of the performance of an Ancestral Sequence Reconstruction (ASR) algorithm (here, FastML), in light of our previous findings. We found that more epistatically constrained ancestors lead to less diversity at comparable time scales, which facilitates their reconstruction. Yet, at comparable diversity, they are more difficult to reconstruct. We believe that this analysis will be instrumental in improving the efficiency of ASR algorithms, which could in principle exploit the correlations identified in this work.

More generally, our findings extend the analogy between protein evolution and disordered physical systems, reinforcing the idea that protein dynamics exhibit characteristics of complex and strongly correlated systems. However, when comparing these results with statistical physics models exhibiting glassy dynamics, we also found qualitative differences, emphasizing the unique constraints imposed by natural evolution on proteins. Future work should aim to further elucidate the connection between epistasis, evolvability, and environmental selection, which may offer insights into evolutionary strategies across diverse biological systems. Furthermore, we believe that most of the ideas presented in this work will soon be amenable to experimental testing, thanks to the increased power of in vitro evolution platforms.

## ACKNOWLEDGMENTS

We thank Maria Chiara Angelini, Ludovic Berthier, Pierre Barrat-Charlaix, Simona Cocco, Dongkyu Lee, Arvind Murugan, Clément Nizak, Misaki Ozawa, Andrea Pagnani, Olivier Rivoire, Joe Thornton, Nobuhiko Tokuriki, and Alya Zeinaty for fruitful discussions. This research has been supported by first FIS (Italian Science Fund) 2021 funding scheme (FIS783 - SMaC - Statistical Mechanics and Complexity) from MUR, Italian Ministry of University and Research and from the PRIN funding scheme (2022LMHTET - Complexity, disorder and fluctuations: spin glass physics and beyond) from MUR, Italian Ministry of University and Research.

[1] M. Blum, A. Andreeva, L. C. Florentino, S. R. Chuguransky, T. Grego, E. Hobbs, B. L. Pinto, A. Orr, T. Paysan-Lafosse, I. Ponamareva, *et al.*, Nucleic Acids Research , gkae1082 (2024).

[2] T. Paysan-Lafosse, A. Andreeva, M. Blum, S. R. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, F. Llinares-López, L. Meng-Papaxanthos, *et al.*, Nucleic Acids Research , gkae997 (2024).

[3] T. U. Consortium, Nucleic Acids Research **51**, D523 (2022).

[4] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, P. A. Craig, G. V. Crichlow, K. Dalenberg, J. M. Duarte, S. Dutta, M. Fayazi, Z. Feng, J. W. Flatt, S. Ganesan, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. P. Hudson, I. Khokhriakov, C. L. Lawson, Y. Liang, R. Lowe, E. Peisach, I. Persikova, D. W. Piehl, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, B. Vallat, M. Voigt, B. Webb, J. D. Westbrook, S. Whetstone, J. Y. Young, A. Zalevsky, and C. Zardecki, Nucleic Acids Research **51**, D488 (2022).

[5] M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1985).

[6] D. M. Fowler and S. Fields, Nature methods **11**, 801 (2014).

[7] P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, *et al.*, Advances in Neural Information Processing Systems **36** (2024).

[8] D. L. Hartl, A. G. Clark, and A. G. Clark, *Principles of population genetics*, Vol. 116 (Sinauer associates Sunderland, MA, 1997).

[9] J. A. G. De Visser and J. Krug, Nature Reviews Genetics **15**, 480 (2014).
[10] M. J. Harms and J. W. Thornton, Nature Reviews Genetics **14**, 559 (2013).
[11] T. N. Starr and J. W. Thornton, Protein science **25**, 1204 (2016).
[12] J. Domingo, P. Baeza-Centurion, and B. Lehner, Annual review of genomics and human genetics **20**, 433 (2019).
[13] M. S. Johnson, G. Reddy, and M. M. Desai, BMC biology **21**, 120 (2023).
[14] M. Lunzer, G. B. Golding, and A. M. Dean, PLOS Genetics **6**, 1 (2010).
[15] A. Biswas, A. Haldane, E. Arnold, and R. M. Levy, eLife **8**, e50524 (2019).
[16] F. J. Poelwijk, M. Socolich, and R. Ranganathan, Nature communications **10**, 4213 (2019).
[17] J. Z. Chen, M. Bisardi, D. Lee, S. Cotogno, F. Zamponi, M. Weigt, and N. Tokuriki, bioRxiv , 2023.11.27.568921 (2023).
[18] F. J. Poelwijk, V. Krishna, and R. Ranganathan, PLoS computational biology **12**, e1004771 (2016).
[19] G. Reddy and M. M. Desai, Elife **10**, e64740 (2021).
[20] J. Otwinowski, Molecular biology and evolution **35**, 2345 (2018).
[21] M. Fantini, S. Lisi, P. De Los Rios, A. Cattaneo, and A. Pastore, Molecular biology and evolution **37**, 1179 (2020).
[22] M. A. Stiffler, F. J. Poelwijk, K. P. Brock, R. R. Stein, A. Riesselman, J. Teyra, S. S. Sidhu, D. S. Marks, N. P. Gauthier, and C. Sander, Cell Systems **10**, 15 (2020).
[23] A. N. Erdoğan, P. Dasmeh, R. D. Socha, J. Z. Chen, B. E. Life, R. Jun, L. Kiritchkov, D. Kehila, A. W. Serohijos, and N. Tokuriki, bioRxiv , 2023 (2023).
[24] Y. Park, B. P. Metzger, and J. W. Thornton, Science **376**, 823 (2022).
[25] G. Rix, R. L. Williams, H. Spinner, V. J. Hu, D. S. Marks, and C. C. Liu, bioRxiv:2023.11.13.566922 (2023).
[26] J. A. de la Paz, C. M. Nartey, M. Yuvaraj, and F. Morcos, Proceedings of the National Academy of Sciences **117**, 5873 (2020).
[27] M. Bisardi, J. Rodriguez-Rivas, F. Zamponi, and M. Weigt, Molecular biology and evolution **39**, msab321 (2022).
[28] S. Alvarez, C. M. Nartey, N. Mercado, J. A. de la Paz, T. Huseinbegovic, and F. Morcos, Proceedings of the National Academy of Sciences **121**, e2308895121 (2024).
[29] A. Biswas, I. Choudhuri, E. Arnold, D. Lyumkis, A. Haldane, and R. M. Levy, Proceedings of the National Academy of Sciences **121**, e2316662121 (2024).
[30] L. D. Bari, M. Bisardi, S. Cotogno, M. Weigt, and F. Zamponi, Proceedings of the National Academy of Sciences **121**, e2406807121 (2024).
[31] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proceedings of the National Academy of Sciences **108**, E1293 (2011).
[32] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Reports on Progress in Physics **81**, 032601 (2018).
[33] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, and A. K. Chakraborty, Immunity **38**, 606 (2013).
[34] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, Molecular biology and evolution **33**, 268 (2016).
[35] R. M. Levy, A. Haldane, and W. F. Flynn, Current opinion in structural biology **43**, 55 (2017).
[36] A. Couce, L. V. Caudwell, C. Feinauer, T. Hindré, J.-P. Feugeas, M. Weigt, R. E. Lenski, D. Schneider, and O. Tenaillon, Proceedings of the National Academy of Sciences **114**, E9026 (2017).
[37] L. Vigué and O. Tenaillon, Proceedings of the National Academy of Sciences **120**, e2304177120 (2023).
[38] S. Alvarez, C. Nartey, N. Mercado, and F. Morcos, Biophysical Journal **121**, 45a (2022).
[39] T. R. Kirkpatrick and D. Thirumalai, Physical Review A **37**, 4439 (1988).
[40] S. Franz, C. Donati, G. Parisi, and S. C. Glotzer, Philosophical Magazine B **79**, 1827 (1999).
[41] S. Franz and G. Parisi, Journal of Physics: Condensed Matter **12**, 6335 (2000).
[42] J.-P. Bouchaud and G. Biroli, Physical Review B—Condensed Matter and Materials Physics **72**, 064204 (2005).
[43] L. Berthier, G. Biroli, J.-P. Bouchaud, L. Cipelletti, D. E. Masri, D. L'Hôte, F. Ladieu, and M. Pierno, Science **310**, 1797 (2005).
[44] S. Franz, G. Parisi, F. Ricci-Tersenghi, and T. Rizzo, The European Physical Journal E **34**, 1 (2011).
[45] L. Berthier, G. Biroli, J.-P. Bouchaud, L. Cipelletti, and W. van Saarloos, *Dynamical heterogeneities in glasses, colloids, and granular media*, Vol. 150 (OUP Oxford, 2011).
[46] S. Franz, H. Jacquin, G. Parisi, P. Urbani, and F. Zamponi, The Journal of chemical physics **138** (2013).
[47] B. Seoane and F. Zamponi, Soft Matter **14**, 5222 (2018).
[48] M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt, Molecular Biology and Evolution **35**, 1018 (2018).
[49] A. P. Muntoni, A. Pagnani, M. Weigt, and F. Zamponi, BMC bioinformatics **22**, 1 (2021).
[50] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt, *et al.*, Science **369**, 440 (2020).
[51] L. Vigué, G. Croce, M. Petitjean, E. Ruppé, O. Tenaillon, and M. Weigt, Nature Communications **13**, 4030 (2022).
[52] H. Ashkenazy, O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, and T. Pupko, Nucleic Acids Research **40**, W580 (2012).
[53] J. Felsenstein, *Inferring phylogenies* (Oxford University Press, 2003).
[54] M. De Leonardis, A. Pagnani, and P. Barrat-Charlaix, bioRxiv , 2024 (2024).
[55] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman, Nucleic Acids Research **49**, D412 (2020).
[56] F. Calvanese, C. N. Lambert, P. Nghe, F. Zamponi, and M. Weigt, Nucleic Acids Research **52**, 5465–5477 (2024).

# SUPPLEMENTAL MATERIAL

## S1. CONSTRUCTION OF THE NATURAL MSAS

The alignments of natural sequences used in this work have been constructed as follows. We believe that the details of the alignment procedure do not affect the results presented in the paper.

- WW: we downloaded the alignment corresponding to the PF00397 family from Pfam [55] and excluded sequences with more than 20% gaps.
- DBD: we used as seed the alignment from Ref. [24] (221 sequences) and ran `hmmsearch` on `uniref90`, excluding sequences with more than 20% gaps.
- Chorismate Mutase (CM): we used the alignment from Ref. [50].
- AAC6: we used the same alignment as in Ref. [27].
- DHFR: we downloaded the alignment corresponding to the PF00186 family from Pfam [55] and excluded sequences with more than 20% gaps.
- BL: we used the same alignment as in Ref. [27].
- SP: the alignment comes from a private communication with the group of Clément Nizak in Paris.

To compute the empirical frequencies used to train the bmDCA model, each sequence is assigned a weight, given by the inverse of the number of other sequences at distance smaller than 20%. The effective number of sequences is the sum of such weights. See e.g. Ref. [49] for details.

In table S1 we show for each family the length of the aligned protein sequences $L$, the depth (number of aligned sequences) $M$ of the alignment, and its effective depth $M_{\text{eff}}$, which serves as a proxy for estimating the diversity of sequences within the alignment.

| **Family** | $L$ | $M$ | $M_{\text{eff}}$ |
|---|---|---|---|
| WW | 31 | 157809 | 10791 |
| DBD | 76 | 24944 | 3129 |
| CM | 96 | 1259 | 937 |
| AAC6 | 117 | 43576 | 21873 |
| DHFR | 160 | 36612 | 8785 |
| BL | 202 | 18334 | 6875 |
| SP | 254 | 6740 | 1207 |

Table S1. Schematic summary of the characteristics of the multiple sequence alignments used for our analysis.

The three DBD sequences that correspond to the blue, green, and red colors are identified, respectively, with the code `UniRef90_E4X5B7/28-107`, `UniRef90_O61854/33-105`, and `UniRef90_A0A818T1L9/14-89`. They come, respectively, from the organisms *Oikopleura dioica (Tunicate)*, *Caenorhabditis elegans*, and *Rotaria sp. Silwood1*.

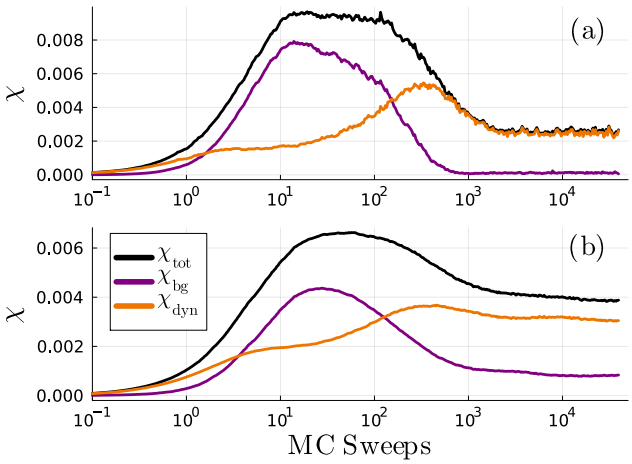


Figure S1. Evolution of $\chi_{\text{bg}}$, $\chi_{\text{dyn}}$ and $\chi_{\text{tot}}$ as a function of Monte Carlo sweeps using a sparse edge-activated version [56] of the Potts model (a) and using the dynamics of Ref. [30] implementing insertions (b), deletions and single nucleotide substitutions for the DBD family.

## S2. DIFFERENT MODELS AND EVOLUTIONARY SCHEMES

Different variants of the bmDCA model that we employed in the main text have recently been suggested. We performed simulations using a sparse edge-activated [56] version for the DBD family, without any significant change in the results (Fig. S1a).

As we already pointed out, the evolutionary protocol that we implemented is a simple Metropolis Monte Carlo defined over amino acid sequences, which is far from being realistic. A more detailed approach was recently described in [30], where it was shown that an algorithm working with single-nucleotide substitutions, insertions and deletions was able to correctly reproduce the dynamics both at long and short time scales. We checked that this more refined algorithm, which was shown to overperform a simple Metropolis at short time scales [27], produces qualitatively similar results for the quantities of interest for our work (Fig. S1b).

## S3. PROFILE MODELS

As epistatic sites play an important role in our description, it is interesting to study a dynamics in which they are totally absent. This can be done by using a profile model in which only a local field acts on each site, eliminating the coupling between different protein

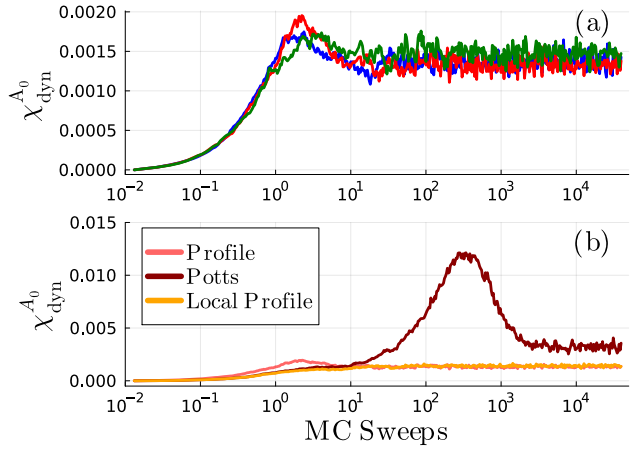


Figure S2. Evolution of the dynamical susceptibility for the profile models. In (a) we show the evolution of $\chi_{\rm dyn}^{A_0}$ for the same starting sequences used in the main text studied with the global profile model, while in (b) we compare the results for the red sequence obtained using the two profile models and the full one.

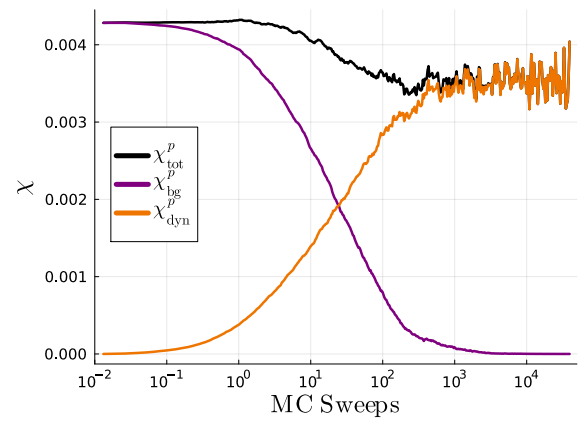


Figure S3. Evolution of $\chi_{\rm bg}^p$, $\chi_{\rm dyn}^p$ and $\chi_{\rm tot}^p$ as a function of Monte Carlo sweeps for the DBD family.

residues. Such profile models can be built in two ways, analogously to what has been done in Ref. [30]. We call *global profile* the model in which the local field is set equal to $h_i^{\rm GP}(a) = -\log f_i(a)$, in such a way as to reproduce the one-point frequencies. In this case, the sites evolve following their local field independently from one another. In Fig. S2a we show how the dynamical susceptibility evolves for the same three sequences we studied in the main text. We saw above that the blue sequence had a smaller $\chi_{\rm dyn}^{A_0}$ and that it converged to the long-time limit quite fast. Differently, the red sequence showed a large peak and only converged after many sweeps. In this figure we see that when using a global profile model the starting sequence does not affect the $\chi_{\rm dyn}^{A_0}$ anymore. As the interactions are set to zero, the initial condition plays a much less important role and all the curves seem to behave in a rather similar way, showing a small peak at a homogeneous time scale. Couplings between residues and hence epistasis are needed to observe the qualitative behavior presented in the main text.

One can also define a *local profile* model, in which the fields acting on the residues are obtained as

$$h_i^{A_0}(a_i) = h_i(a_i) + \sum_j J_{ij}(a_i, a_j^0) \ , \tag{S1}$$

and therefore depend on the initial sequence $A_0$. We compare both these profile models and the Potts one Fig. S2b for one sequence $A_0$. At the very beginning of the dynamics the three models are very similar, as only few mutations occurred. Already after few steps, less than a sweep, the global profile model displays the small peak discussed above, due to the uncorrelated mutations occurring at different residues. The local profile takes into account the role of the initial condition and hence it follows a dynamics similar to the full Potts model for a longer time ($\sim$ 10 sweeps). However, for even longer times, in the full Potts model the evolution of the background changes the local field acting on each residue, allowing sites to evolve in a coordinated manner. This results in the presence of the peak, as discussed in the main text. In the local profile this correlated evolution cannot take place, as the local fields are kept constant. Hence, the dynamical susceptibility quickly saturates to its long time limit value.

## S4. DIFFERENT DEFINITION OF SUSCEPTIBILITY

It is also interesting to consider a different definition of susceptibility. In particular we already defined $H_r^{A_0}(t,0)$ as the Hamming distance between the sequence that evolved for $t$ steps under the predefined evolutionary noise $r$ starting from the initial sequence $A_0$ and the initial sequence itself. We can define two supplementary Hamming distances, this time computed at fixed Monte Carlo sweeps between different independent simulations of the system. In particular we define

$$H_{r,r'}^{A,B}(t) = 1 - \frac{1}{L}\sum_i \delta_{a_{r,i}^A(t), a_{r',i}^B(t)} \tag{S2}$$

as the Hamming distance between sequences that have independently evolved from different initial sequences $A$ and $B$ for the same time $t$. The $[\dots]$ average is then computed by summing over all different realizations of the couples of evolutionary noises $(r, r')$, while the $\langle\dots\rangle$ average is computed by summing over all couples of different initial sequences $(A, B)$. It is interesting to compare the particular case $H_{r,r'}^{A,A}(t)$ with what we showed before. Because our dynamics is time-reversible, we expect that $\langle[H_{r,r'}^{A,A}(t/2)]\rangle \sim \langle[H^A(t)]\rangle$. We checked that this is indeed the case.

In the main text we showed that $\chi_{\rm dyn}^{A_0}$ can be expressed as the sum of a correlation function between sites (see

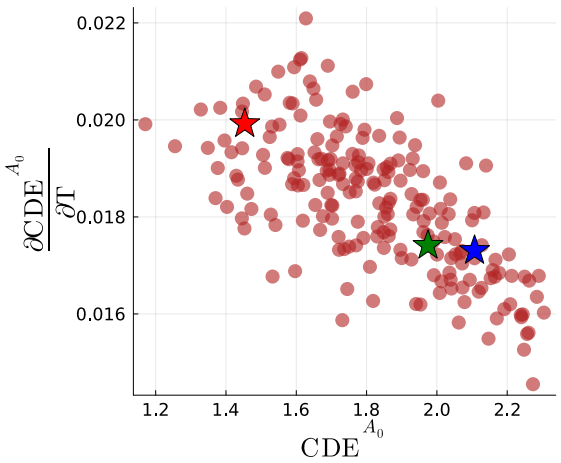


Figure S4. Derivative of the context-dependent entropy with respect to temperature (i.e. inverse selection pressure), as a function of the sequence mutability $\mathrm{CDE}^{A_0}$ for 200 sequences of the DBD family. The colored stars refer to the three sequences of Fig. 2.

Eqs. 11 and 12). If two sites both mutated after a time $t$ with respect to the initial sequence $A_0$, then they presented a high correlation. In this case, $\chi^{A,B}_{\mathrm{dyn}}$ can be still seen as the sum of site correlations, but, differently from the previous case, two sites $i$ and $j$ are highly correlated if after time $t$ the realizations starting from the two sequences $A$ and $B$ display the same amino acids in those sites, namely

$$G^{A,B}_{ij,\mathrm{dyn}}(t) = [\delta_{a_i^{t,A},a_i^{t,B}}\delta_{a_j^{t,A},a_j^{t,B}}] - [\delta_{a_i^{t,A},a_i^{t,B}}][\delta_{a_j^{t,A},a_j^{t,B}}].$$

Analogously to what is presented in the main text, we can define $\chi^p_{\mathrm{bg}}$, $\chi^p_{\mathrm{dyn}}$ and $\chi^p_{\mathrm{tot}}$ where $p$ apex indicates that we are considering a pairwise metric. Because $H^{A,B}_{r,r'}(t)$ strongly depends on which couples $A, B$ are chosen to initialize the two dynamical evolutions, we can easily expect that the impact of the background initialization $\chi^p_{\mathrm{bg}}$ will be dominant at the beginning of evolution. Instead, when equilibrium is reached, $H^{A,B}_{r,r'}(t)$ tends to the mean pairwise hamming between the sequences in the training set as the model is generative. Hence, at later times the memory of the initial condition is lost and $\chi^p_{\mathrm{bg}}$ vanishes. This is confirmed by Fig. S3.

## S5. RESPONSE TO ENVIRONMENTAL VARIATION

To complement the analysis presented in the main text concerning the response to a change in selection pressure, we checked how much the mutability of the ancestral sequence $\mathrm{CDE}^{A_0}$ changes with respect to a shift in selection pressure. In Fig. S4 we plot such derivative as a function of $\mathrm{CDE}^{A_0}$ for a set of sequences of the DBD family. It is clear that epistatically constrained sequences have a higher change in mutability when subject to a change in selection pressure (i.e. a change in the simulation temperature). As a matter of fact, the average context-dependent entropy of a given sequence $A_0$ can be expressed as a function of the local fields $h_i^{A_0}(a_i)$ defined in Eq. (S1) as follows

$$\mathrm{CDE}^{A_0} = -\frac{1}{L}\sum_i \sum_{a=1}^{21} \frac{e^{-\beta h_i^{A_0}(a)}}{z_i^{A_0}} \log_2 \frac{e^{-\beta h_i^{A_0}(a)}}{z_i^{A_0}} \ , \quad \text{(S3)}$$

with $z_i^{A_0} = \sum_{a=1}^{21} e^{-\beta h_i^{A_0}(a)}$. Hence, the derivative with respect to inverse selective pressure can be analytically computed as follows:

$$\frac{\partial \mathrm{CDE}^{A_0}}{\partial T} \propto \frac{1}{L}\sum_i \sum_{a=1}^{21} \langle h_i^{A_0}(a)^2\rangle - \langle h_i^{A_0}(a)\rangle^2 \ , \quad \text{(S4)}$$

where $\langle \bullet \rangle = \sum_{a=1}^{21} \bullet e^{-\beta h_i^{A_0}(a)}/z_i^{A_0}$. This expression indicates that sites having a greater variance of local fields are more prone to change their mutability as a sudden change in selective pressure takes place. This analysis offers interesting insights into the evolvability of protein sequences which should be corroborated by evolution experiments.