Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes

Lucile Vigué[*,1], Giancarlo Croce[*,2,3], Marie Petitjean[1], Etienne Ruppé[1, 4], Olivier Tenaillon[†,1,5], and Martin Weigt[†,6]

[1]Université de Paris, INSERM, Infection Antimicrobials Modelling Evolution — IAME, Paris, France
[2]Department of Oncology, Ludwig Institute for Cancer Research Lausanne, University of Lausanne, Lausanne, Switzerland
[3]Swiss Institute of Bioinformatics — SIB, Lausanne, Switzerland
[4]Laboratoire de Bactériologie, Hôpital Bichat, APHP, Paris, France
[5]Université Sorbonne Paris Nord, INSERM, Infection Antimicrobials Modelling Evolution — IAME, Paris, France
[6]Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Computational and Quantitative Biology — LCQB, Paris, France
[*]These authors contributed equally.
[†]These authors contributed equally.

November 2021

Abstract

Characterizing the effect of mutations is key to understand the evolution of protein sequences and to separate neutral amino-acid changes from deleterious ones. Epistatic interactions between residues can lead to a context dependence of mutation effect. Context dependence constrains the amino-acid changes that can contribute to polymorphism in the short term, and the ones that can accumulate between species in the long term. We use computational approaches to accurately predict the polymorphisms segregating in a panel of 61,157 *Escherichia coli* genomes from the analysis of distant homologues. By comparing a context-aware Direct-Coupling Analysis modelling to a non-epistatic approach, we show that the genetic context strongly constrains the tolerable amino acids in 30% to 50% of amino-acid sites. The study of more distant species suggests the gradual build-up of genetic context over long evolutionary timescales by the accumulation of small epistatic contributions.

Introduction

Understanding how biological diversity emerges and evolves is at the heart of molecular evolutionary biology. The long-standing confrontation between adaptationists [1] and neutralists [2] has oriented the scientific debate towards comparing the relative contributions of natural selection and drift in the process. While the first ones consider most of the differences between organisms to result from adaptation to different environments, the second ones support that polymorphisms reflect random occurrences of equally fit variants.

In recent years, the increasing interest in the role played by historical contingency has revived this old neutral-versus-selective debate [3]. Evolutionary contingency arises when mutations that fix depend on permissive mutations that occurred before. Once fixed, they influence the fate

of upcoming mutations and become increasingly deleterious to remove — a phenomena called entrenchment [4]. The concept of contingency puts epistasis at the forefront of molecular evolution: an amino acid that is neutral or beneficial in a genetic context can be deleterious in another due to epistatic interactions between residues [5]. Characterizing these epistatic interactions is thus key to uncover the context dependence of mutation effects and understand the extent to which contingency shapes molecular evolution. Moreover, predicting which non-synonymous mutations are likely or not to affect a protein is essential in molecular genetics. Though genetic analyses from quantitative trait locus (QTL) analyses to genome-wide association studies (GWAS) successfully identify genomic regions associated to a disease or to a trait of interest, these regions usually encompass multiple neutral mutations in addition to the causative one. An accurate characterization of non-synonymous mutation effects would definitely help identifying the causative mutations.

Deep mutational scans and small adaptive landscape reconstructions allow to experimentally study the effect of mutations or combinations of mutations in a genetic background [3] [6]. They highlight the short-term evolutionary constraints the protein faces and a more general pattern of negative epistasis in which deleterious mutations become more deleterious in combination. However, purifying selection removes these mutations from the population. Consequently, their epistatic interactions may not contribute to long-term protein evolution. Some experiments have unveiled a strong role of positive epistasis over long evolutionary times, by measuring the effect of the same mutation in distant homologues from diverged or ancestral species [7] [8]. For instance, the same amino-acid change can be deleterious in distant backgrounds while being neutral or beneficial in its native background.

Computational approaches can help to bridge the gap between short-term and long-term evolution. On the one hand, simulations can mimick the fixation of amino-acid changes across many generations [4] [9] [10]. Yet, their results rely on the validity of the assumptions made to model protein evolution and the effects of epistasis. On the other hand, data-driven approaches to study protein evolution become possible thanks to the revolution of high-throughput sequencing. The accumulation of closely related and more diverged genome sequences enable us to track the apparition and the fixation of amino-acid changes over different timescales. Instead of simulating evolution, we can analyse the patterns of diversity observed in nature on both short term (polymorphisms within a species) and longer term (fixed differences between diverged species). The computational study of epistasis requires models of amino-acid sequences that account for epistatic interactions between residues. A current tool to model epistasis is Direct-Coupling Analysis (DCA) [11]. DCA aims at modelling epistatic patterns of co-evolution between residues from the analysis of diverged but homologous protein sequence alignments. It successfully identified residue contacts in the three-dimensional protein fold [11], generated new and functional artificial enzymes [12], predicted deep mutational scanning outcomes [13] [14] and was used to investigate amino-acid changes between two closely related genomes [15]. Importantly, DCA epistatic models constantly outperform simpler non-epistatic modelling approaches (independent models, IND).

Here, we intend to use IND and DCA models in a large-scale study of the *Escherichia coli* core genome in order to understand to what extent epistasis constrains the emergence of non-synonymous polymorphisms within a species and how these epistatic constraints are building-up through time. To this end, we gathered a collection of >60,000 *E. coli* genomes and a sample of diverged species ranging from *Escherichia coli* to *Yersinia pestis* to study both polymorphisms arising within a species and fixed differences accumulating with divergence. With that genome scale approach we intend to: (i) test how mutation effect prediction can identify the sites where polymorphisms segregate; (ii) quantify how the genetic background contributes to these predictions; (iii) study the type of epistatic interactions responsible for the background specific effect of mutations and how they build up over evolutionary timescales.

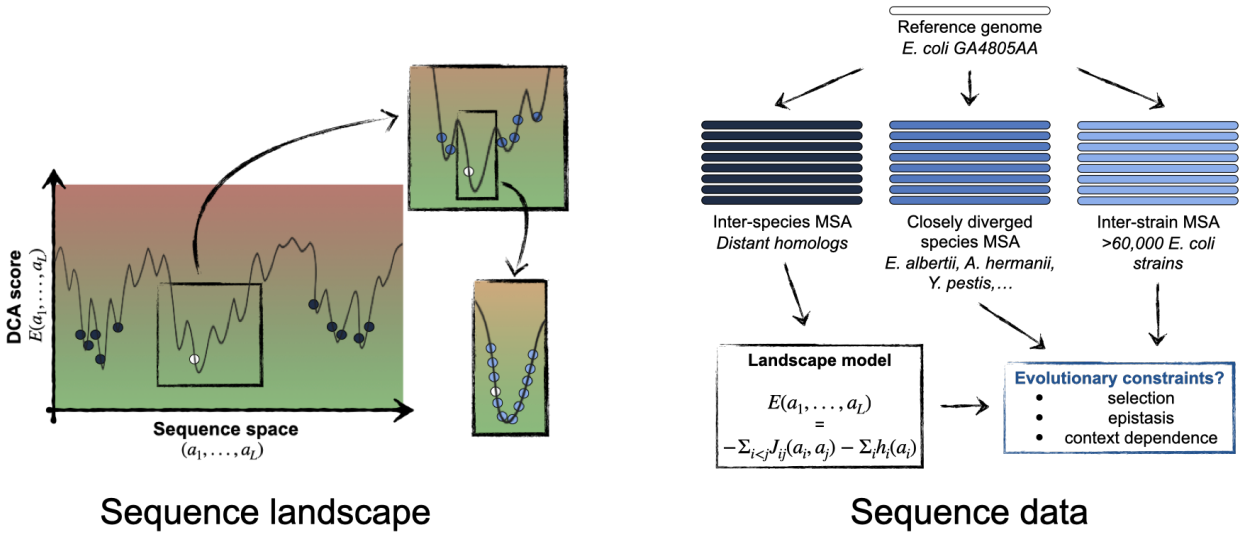


Figure 1: **Schematic representation of the sequence landscape and its relation to sequence data.** The landscape is defined *via* a real-valued function of any aligned sequence, with low values indicating "good" functional sequences (green area), and high values "bad" non-functional sequences (red area). Natural sequences can be seen as samples of low values: close orthologs (light blue) of a reference sequence (in white) form a sample which is localized in sequence space and surrounded by closely diverged species (mid-blue). Distantly diverged homologs (dark blue) form a global sample. All sequence data are aligned relative to the reference sequence. Within our work, the global sample will be used to infer data-driven landscape models for all proteins families present in the *E. coli* core genome, and the variability of the local sample and the closely diverged species will be analyzed for signatures of selection, epistasis and context dependence of natural amino-acid polymorphisms.

Results

Data-driven protein sequence landscapes for the case proteome of *E. coli*

The central concept of our work are amino-acid sequence landscapes, constructed for each protein or protein domain in some reference genome, here *E. coli*. These landscapes associate a DCA score E to any sequence $(a_1, ..., a_L)$. A DCA score is composed of constant terms reflecting amino-acid conservation at each site and pairwise couplings modelling epistatic interactions between pairs of residues. Low DCA scores correspond to good (*i.e.* fully functional) sequences whereas high values to bad (*i.e.* non-functional) ones (Figure 1). We build these amino-acid sequence landscapes by training DCA models on multiple-sequence alignments (MSAs) of distant homologs sampled in diverged species (Methods). These are widely variable sequences (typical sequence identities are around 20-30%), so they may be understood as a global sample of the sequence landscape, cf. the dark blue dots in Figure 1. To avoid biasing the results, we have removed from the MSAs any sequence which is too close to *E. coli*. Therefore, it is not evident that the resulting models are informative about the very local structure of the landscape around the *E. coli* reference sequence (white, light blue and mid-blue dots in Figure 1). The latter might be dominated by idiosyncratic constraints characterizing *E. coli* as a species, while the MSAs of homologs contains the conserved evolutionary constraints of the entire protein family. Thus, we want to investigate whether amino-acid sequence landscapes can unify the study of epistasis on short and long timescales.

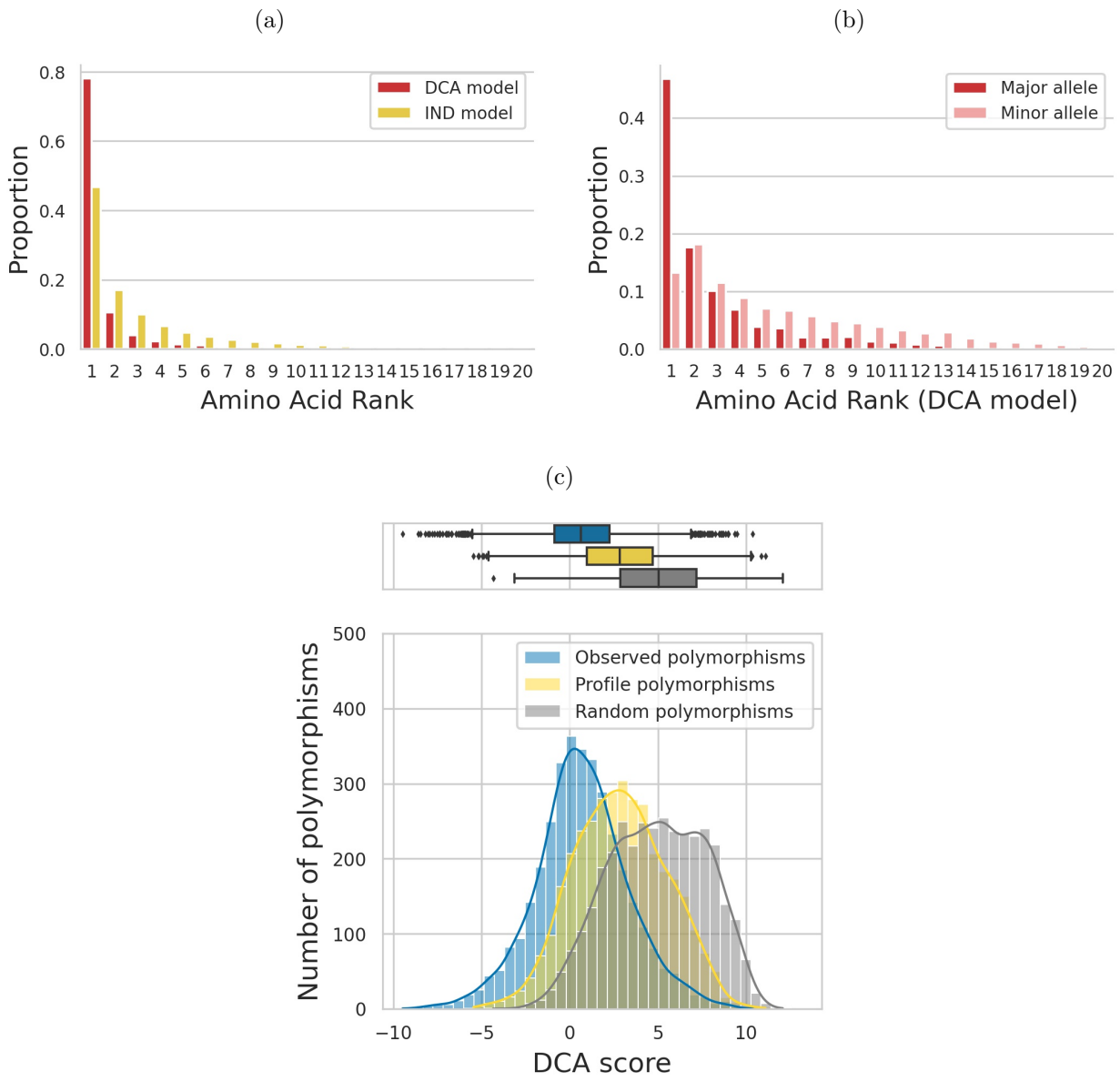


Figure 2: **Predicted effects of observed amino acids using an IND model that neglects epistasis or a DCA model that incorporates pairwise epistasis. (a)** Rank of native amino acid in the reference strain as compared to all 20 possible amino acids. DCA model (red) outperforms IND (yellow) by predicting twice as many native amino acids to be the best possible. **(b)** DCA rank of major and minor allele for all sites that are polymorphic at a >5%-threshold, among all 20 possible amino acids. Major alleles (alleles at frequencies >50%) have better ranks than minor alleles (alleles at frequencies between 5% and 50%). The distribution of consensus alleles peaks at the first rank (46.8% of polymorphic sites have major allele ranking first and 17.6% have second best rank) while the distribution of minor alleles peaks at the second rank (13.2% have best rank against 18.1% that are second-best). **(c)** Distribution of DCA scores of non-synonymous polymorphisms observed at frequencies >5% across the >60,000 strains (blue) compared to mutations sampled from an IND model (yellow) or to random mutations (grey). A large number of possible mutations are predicted to be highly deleterious (positive scores) compared to naturally-occurring polymorphisms that tend to be neutral (blue distribution centered on zero). Polymorphisms predicted from IND are slightly deleterious once epistasis is taken into account (yellow distribution shifted towards positive values).

Strong signature of selection at the amino-acid level

We first test how accurately DCA can model *E. coli* amino-acid sequences. To work at a genome scale, we focus on 2,087 Pfam domains [16] spanning 284,529 residues among 1,462 core genes (Methods). We also perform the same analysis on 1,029 entire core gene sequences in order to increase site coverage. Results presented in the following sections are those obtained on Pfam domains, results on full core genes are presented in Supplementary Figures 1, 2, 3, 4, 5. The results for full sequences are mostly consistent but of lower quality than those obtained for Pfam domains, since the MSAs used for model training contain less and less diverse sequences.

DCA models provide a score for each amino acid in each position, which depends on the sequence context in *E. coli*. On the contrary, the score of each amino acid in IND models is context-agnostic as it directly derives from its frequency across distant homologs (Methods). To compare model predictions to reality, we gather a database of >60,000 *E. coli* strains where we record all polymorphisms occuring at frequencies >5%. We use a ST131 strain as reference strain, this clonal complex is a public health concern because of its virulence and resistance to antibiotics [17] and has thousands of isolates sequenced in the database.

Amino acids observed in *E. coli* are well predicted by DCA, and to a lesser extent by IND. 78% of amino acids observed in the reference strain rank first with a DCA model while this figure drops to 47% with IND (Figure 2a), in agreement with previous study [15]. Approximately half of the time an amino-acid site is polymorphic the major allele is ranked first by DCA while minor alleles are more likely to rank second (Figure 2b). The DCA score distribution of *E. coli* polymorphisms centers on 0, meaning that DCA predicts them to be close to neutral (blue distribution, Figure 2c). In comparison, DCA predicts that amino acids sampled from distant homologues and inserted in *E. coli* sequences will be deleterious (yellow distribution, Figure 2c), a prediction IND cannot make. These results are consistent with the idea that polymorphisms that fix in a population are close to neutral at the time they occur but can be deleterious in another background. Figure 2c compares these scores with random mutations (grey histogram), predicting them to be more deleterious, since they include never observed mutations that are presumably highly counter-selected.

DCA and IND models predict mutation effects of amino-acid changes. However, the likelihood of observing an amino-acid change also depends on mutational biases. Among the 20 possible amino acids, we cannot obtain more than nine by mutating a given codon only once. On short evolutionary timescales, polymorphisms that require more than one single nucleotide polymorphism (SNP) should rarely occur. If we set the probability of observing them to zero, the power to predict *E. coli* polymorphisms slightly increases for both models (by 5.2% for DCA and 10.5% for IND, Supplementary Figure 6).

These results validate that even though DCA models are trained on distant homologs, they can capture the effect of natural selection at different timescales. Their ability to predict amino acids in the reference strain reflects the action of natural selection in fixing amino acids when *E. coli* diverged from other species. When it comes to predicting polymorphisms, it emphasizes the action of purifying selection on a shorter term. The better performances of DCA over IND highlight the major role played by epistasis in shaping mutation effect and the strong contingency of amino acids observed in *E. coli*. It thus comforts us in using DCA throughout the rest of this work.

The context constrains the predicted site variability in *E. coli*

Focusing on individual amino acids, we have seen that native amino acids fixed in *E. coli* and polymorphisms observed in a wide collection of strains are strongly contingent on the genetic background. Going to an amino-acid site perspective, this raises the question of how much epistasis shapes site variability. When comparing protein sequences from distant species, we observe that some sites are conserved while others vary. However, if mutation effect depends on context, the level of variability observed at an amino-acid site across distant species may not reflect how polymorphic

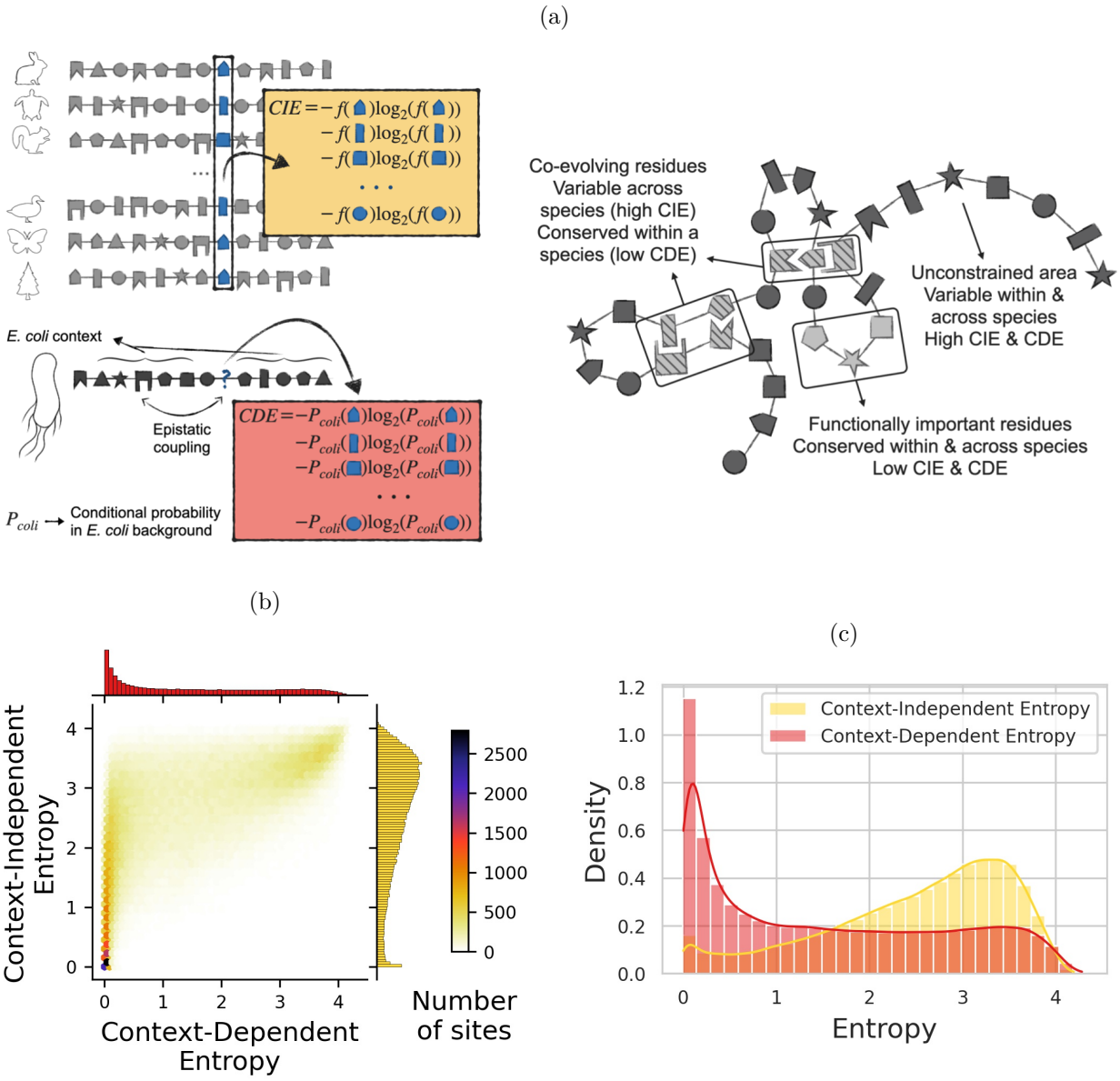


Figure 3: **Predicting the variability of amino-acid sites. (a)** Entropy quantifies the level of variability of an amino-acid site from conserved (entropy ~ 0) to highly variable (entropy ~ 4). It can be computed from a non-epistatic model (Context-Independent Entropy (CIE), yellow) *i.e.* from the frequencies of amino acids observed across distant species, or from an epistatic model (Context-Dependent Entropy (CDE), red) *i.e.* from the conditional probabilities of observing each amino acid in *E. coli* background. Residues that have strong epistatic interactions with others will be lowly polymorphic once the genetic context is fixed (low CDE) but can vary between species (high CIE) by co-evolving with their partners (cf. hatched residues). **(b)** Bivariate histogram of CDE and CIE for all sites in the dataset. Two populations of sites are clearly recognizable, in particular separated by their CDE values. **(c)** Marginal distributions of CDE and CIE for all sites in the dataset. CDE divides amino-acid sites into two populations of similar sizes: conserved (CDE<1) and variable (CDE≥ 1). On the contrary, most of the amino-acid sites have a high CIE, *i.e.* IND predicts them to be highly variable.

this site can be within any specific species.

We use Shannon entropy as an information-theoretic measure quantifying the diversity of amino acids observed at a given site (Figure 3a). It measures the logarithm (in base 2) of the effective number of admissible amino acids at a position, if these were equiprobable. A site with an entropy of zero should only tolerate one amino acid: it is conserved. A value of one can for instance correspond to two amino acids at 50% frequency each. Entropy reaches its maximal value of $\log_2(20) = 4.32$, if all 20 possible amino acids are equally likely. Based on this concept, we can define a Context-Independent Entropy (CIE) from an IND model and an *E. coli* specific Context-Dependent Entropy (CDE) from a DCA model (Methods).

We compute CIE at locus i from the amino-acid frequencies $f_i(a)$ in the column i of the MSA of distant homologs as:

$$CIE_i = -\Sigma_a f_i(a) \log_2 f_i(a)$$

To compute CDE, we first need to determine the probability of observing a certain amino acid a_i in position i, given that the other positions take amino acids $a^0_{\setminus i} = (a_1, ..., a_{i-1}, a_{i+1}, ..., a_L)$ present in the *E. coli* reference sequence. Within our DCA-based modelling framework, this quantity reads:

$$P(a_i|a^0_{\setminus i}) = \exp\left\{h_i(a_i) + \Sigma_{j\neq i} J_{ij}(a_i, a_j)\right\}/z_i,$$

with the normalization z_i chosen such that P becomes a probability distribution over the values of a_i, *i.e.* over the 20 theoretically possible amino acids in position i (gaps are not considered, since we study the effects of amino-acid substitutions and not deletions). CDE is now given by:

$$CDE_i = -\Sigma_{a_i} P_i(a_i|a^0_{\setminus i}) \log_2 P_i(a_i|a^0_{\setminus i}),$$

with $a^0_{\setminus i}$ being the sequence context of the *E. coli* reference strain.

CIE and CDE are both model-predicted quantities, that do not use any *E. coli* polymorphism data to predict variability within this species. CIE corresponds to the level of variability observed across distant species. CDE takes the amino-acid context and the local epistatic couplings of the reference strain into account to predict the level of variability within the *E. coli* sequence background. If epistasis is negligible, CIE and CDE values should be comparable.

Figure 3b shows a bivariate histogram of CIE and CDE over all sites in our dataset. Two distinct communities clearly emerge:

- top-right peak of sites with high CDE and CIE. These sites display very little context-dependence (both entropies have comparable values). They reach entropy values near 4, *i.e.* close to the upper limit of $\log_2(20) = 4.32$. These sites are variable across distant species and predicted to be highly polymorphic in *E. coli*.
- left peak of sites with low CDE and low to high CIE. We predict them to be conserved in *E. coli* (CDE close to 0) but they can vary across distant species (CIE ranging from 0 to more than 3). We expect these sites to display a low level of polymorphism across *E. coli* strains.

CIE and CDE distributions over all sites greatly differ (Figure 3c). While only 10% of sites are conserved across distant species (CIE < 1, corresponding to an effective number of amino acids below 2), we predict 45% of sites to be conserved in *E. coli* (CDE < 1) largely due to local epistatic couplings.

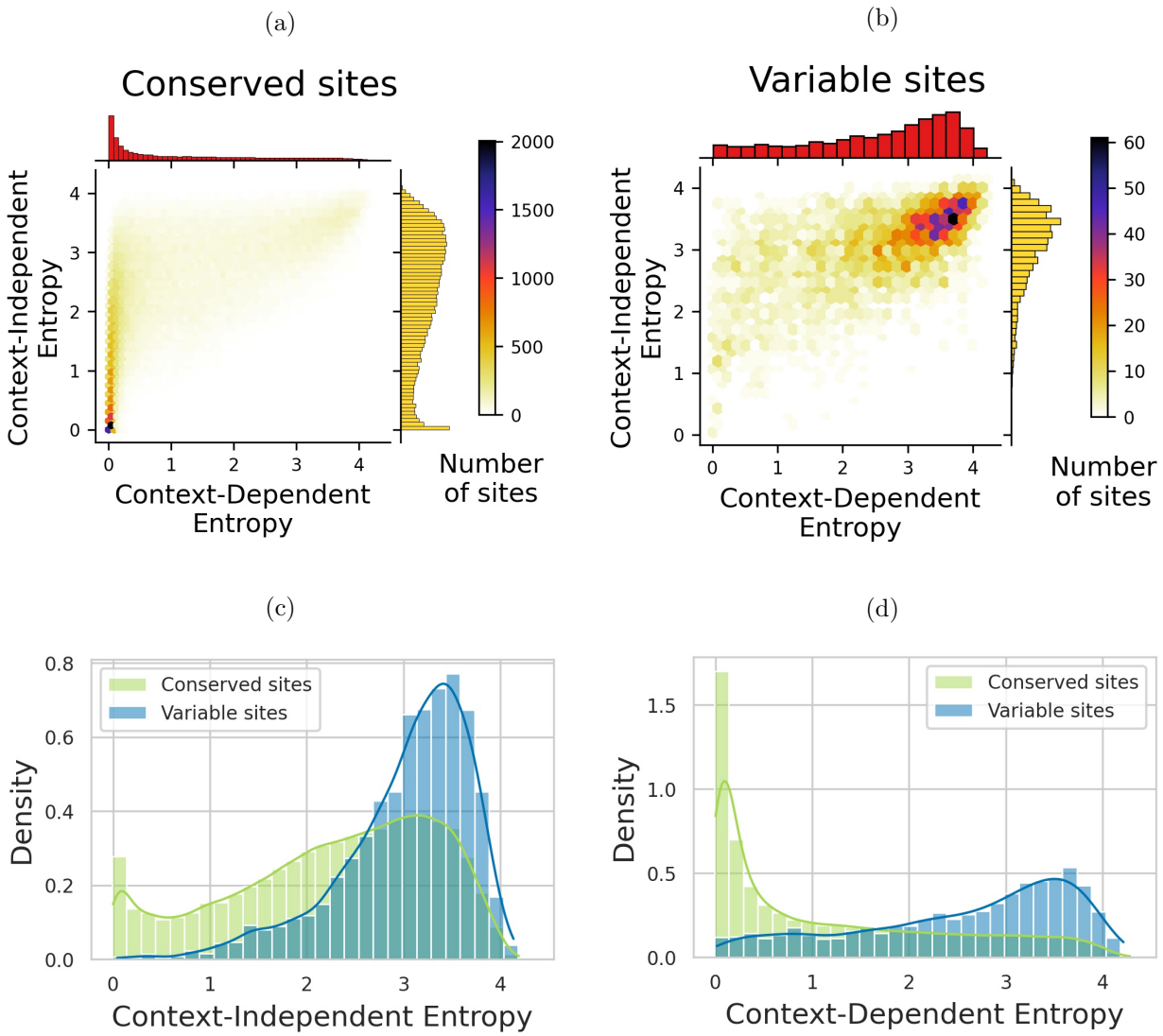


Figure 4: **Predicting amino-acid sites that are conserved or polymorphic in *E. coli*. Comparison of the performances of IND and DCA models.** **(a)** Bivariate histogram of CDE and CIE for sites that are conserved across >60,000 strains of *E. coli*. Most of them cluster on the left peak of low CDE. **(b)** Bivariate histogram of CDE and CIE for sites that are polymorphic at a 5% threshold across >60,000 strains of *E. coli*. Most of them cluster on the right peak of high CDE. **(c)** Distribution of CIE for conserved (green) and polymorphic (blue) sites in *E. coli*. A non-epistatic model fails at distinguishing between both populations. Most of the sites are predicted to have a high entropy so to be highly variable, including those that display no mutation in >60,000 strains of *E. coli* (green distribution). **(d)** Distribution of CDE for conserved (green) and polymorphic (blue) sites in *E. coli*. A model that incorporates pairwise epistasis predicts a low entropy for conserved sites (the green distribution peaks near 0) and a high entropy for variable sites (the blue distribution peaks near 4).

Context-Dependent Entropy accurately predicts polymorphic and constrained sites in *E. coli*

We can now confront these model-based predictions to the observed variability in our dataset of >60,000 *E. coli* strains. To do so, we categorize *E. coli* sites into:

- conserved: no polymorphism observed in any of the strains.
- variable: at least 5% of the strains harbor a mutation with respect to the consensus sequence.

Lowly polymorphic sites (<5%-frequency polymorphisms) can correspond to variable sites but also to conserved sites with sequencing errors or deleterious mutations segregating at low frequencies, so we choose to exclude them from the analysis.

Most of the conserved sites cluster on the left peak of low CDE (Figure 4a) whereas variable sites tend to cluster on the top-right peak of high entropies (Figure 4b). CDE appears more relevant than CIE to discriminate conserved from variable sites. Indeed, only 14.9% of conserved sites have CIE<1 (Figure 4c) while 56.6% have CDE<1 (Figure 4d). If we integrate mutational biases into our analysis, by restricting the computation of entropy to 1-SNP amino-acid mutations (Methods), we find that 70.3% of conserved sites have CDE<1 whereas only 28.2% have CIE<1 (Supplementary Figure 7). Using simulations (Supplementary Methods), we show that the remaining 29.7% of conserved sites that are predicted to be polymorphic (CDE>1) may correspond to random drift limiting the amount of neutral diversity that segregates within a population (Supplementary Figure 8). In other words, polymorphisms may arise on these sites but have not been observed in nature yet.

These results show that CDE accurately predicts the level of variability of an amino-acid site by integrating constraints linked to its function, common to all genetic backgrounds, and local epistatic couplings that are specific to a given genetic context. CIE misses most of the conserved sites, demonstrating how strongly the context reduces the variability, which is possible at an amino-acid site.

Quantifying the level of contingency

We now want to investigate how much the genetic context reduces the diversity of amino acids tolerated at a site. In other words, how contingent on the genetic background the effect of an amino-acid change is. Comparing CIE to CDE allows to quantify contingency, as they both measure site variability with CIE being context-agnostic and CDE being context-aware. We can split amino-acid sites into three categories (Figure 5a):

- 10.0% are conserved across all species as well as in *E. coli* (CIE < 1). They are likely to be functionally essential. Mutating away from the observed amino acid will always be deleterious so the context has no real influence on their level of conservation.
- 54.6% are variable across all species as well as in *E. coli* (CIE ≥ 1, CDE ≥ 1). They are often constrained (CDE $< \log_2(20)$), but allow for a considerable amino-acid variability both in the family and in the specific *E. coli* context: at these positions, we observe both fixed differences between species and polymorphisms within the *E. coli* population.
- 35.4% are conserved in *E. coli* context but variable across species (CIE ≥ 1, CDE < 1). Amino acids observed in distant species will not be tolerated in this specific context: evolution is contingent on the genetic background.

We define the information gain provided by the sequence context as the difference between CIE and CDE (Methods). If both are equal, no information is contained in the context. The lower CDE is compared to CIE, the greater the information gain and the level of contingency. We observe that the majority of sites have a positive gain in information when the sequence context is known

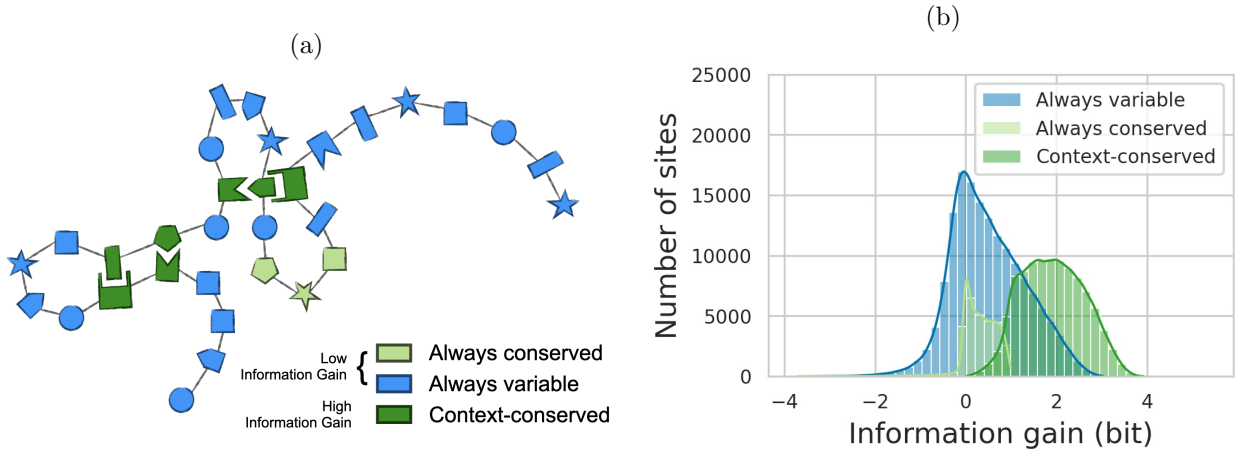


Figure 5: **Quantifying the effect of the context in reducing amino-acid site variability. (a)** The genetic background is expected to differentially impact amino-acid sites. It has a low influence on sites that have the same level of variability in *E. coli* and across distant species (blue and light green). On the contrary, it strongly impacts sites that are variable across distant species but are conserved in *E. coli* due to local epistatic couplings (dark green). **(b)** Information gain quantifies the difference between an amino-acid site variability across distant species and its potential variability in *E. coli*. Sites that are variable across distant species (CIE ≥ 1) but conserved in *E. coli* (CDE < 1) are the ones with the highest information gains (dark green distribution). Note that the information gain is given in bits, 1 bit corresponds to an effective reduction of the available amino acids by a factor 2, 2 bits by a factor 4, and 3 bits by a factor 8.

(Figure 5b). In 47.1% of sites, the effective number of acceptable amino acids in the *E. coli* context is at least a factor two smaller than what a context-independent analysis of distant homologs would predict (information gain > 1 bit). We conclude that roughly 30% to 50% of amino-acid sites show some consistent signal of context dependence.

Epistasis is a diffuse pattern involving a sum of many small couplings

The higher accuracy of DCA over IND in predicting site variability and amino acids observed in *E. coli* proves that epistasis strongly shapes the effect of mutations. Following this observation, we want to use DCA as a tool to study epistasis in natural isolates. First, we look at epistasis between polymorphisms arising jointly in *E. coli*. To do so, we gather all gene sequences with exactly two amino-acid substitutions (other than gaps, *i.e.* deletions or insertions) compared to the reference strain. For each pair of mutations, we compare the DCA-predicted effect of the double mutation to the sum of the effects of each single mutation introduced individually in the reference sequence (Methods). We observe no clear difference between these two quantities (Figure 6a), indicating an absence of strongly-coupled polymorphisms. Two main factors may explain the absence of strong epistatic couplings between polymorphisms in *E. coli*. First, polymorphisms arise on highly variable sites: these sites are poorly constrained by epistasis (high CDE). Second, previous works claim that epistasis is often weak compared to the typical effect size of mutations [18]. This second point does not contradict the strong context dependence of mutations. It suggests that context might be a collective effect arising from the accumulation of many small epistatic couplings. Importantly, these couplings may involve sites that are conserved in *E. coli* but vary across distant species. We use inverse participation ratio (IPR) to estimate the proportion of sites effectively coupled to a locus in amino-acid sequences modelled with DCA (Figure 6b, Methods). We find that each amino-acid site is coupled to about one fourth of the rest of the protein. Taken altogether, these results lead us

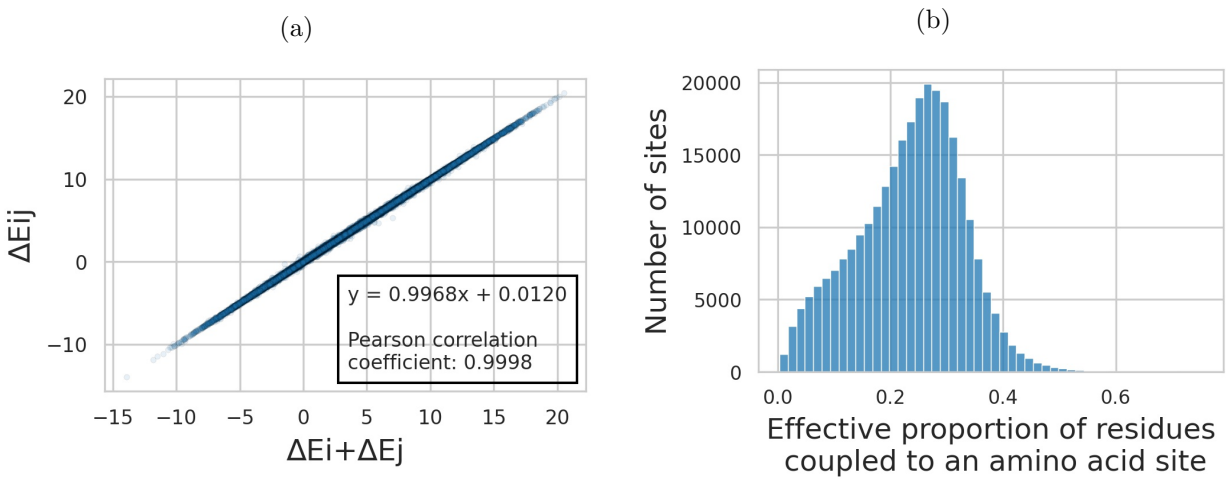


Figure 6: **Epistasis observed in *E. coli*.** **(a)** Mutational effect ΔE_{ij} of observed double mutations with respect to the reference, plotted against the sum $\Delta E_i + \Delta E_j$ of the individual mutation scores. The absence of clear deviations from the diagonal reveals the lack of strong epistatic couplings between pairs of mutations in our strain dataset. **(b)** Histogram of the effective proportion of sites coupled with a given amino acid. It is computed from the inverse participation ratio: $1/(IPR \times$protein length$)$. The median of the distribution is 24%, meaning that amino-acid sites are generally coupled to about one fourth of the other residues in the protein according to DCA modelling of epistasis.

to consider that context dependence of mutations does not rely on a few strong epistatic couplings but on an aggregation of many small couplings accumulated with divergence.

Gradual construction of the context with divergence

So far, we have gathered evidence that many small couplings accumulate to build a genetic context. This translates into an absence of strong epistatic signature of polymorphisms co-occuring in *E. coli*. However, we expect epistasis patterns to emerge gradually when the number of substitutions increases. To study how the genetic background is building up with divergence, we gather 853 Pfam domains spanning 516 core genes shared by diverged species from *E. coli* to *Yersinia pestis* (Figure 7a, Methods).

We start by comparing pairs of homologous sequences. For each pair, we compute the DCA epistatic cost as being the difference between the DCA score of the fixed differences altogether and the sum of their DCA effects when inserted individually in one of the two genetic backgrounds (Methods). It is worth noting that a negative DCA epistatic cost corresponds to positive epistasis: fixed differences are more beneficial, *i.e.* have a lower DCA score, taken altogether than expected by the sum of their individual effects. As gaps can artificially create a pattern of positive epistasis, we only keep pairs of sequences that have no more than one gap difference. We observe a strong pattern of positive epistasis that increases with divergence (Figure 7b). This is consistent with a model where fixed differences are contingent on previous mutations and entrenched by subsequent ones. Individual couplings are biased towards positive epistasis (pronounced left tail of negative DCA couplings between pairs of fixed differences in Figure 7c). However, their values rarely fall below -1 (note the log scale of the vertical axis), a rather low effect size compared to the most extreme epistatic costs that can be measured between entire sequences in Figure 7b. This is consistent with epistatic patterns emerging gradually by an addition of small couplings accumulated with divergence. The more diverged the sequences, the stronger the epistatic signal because each

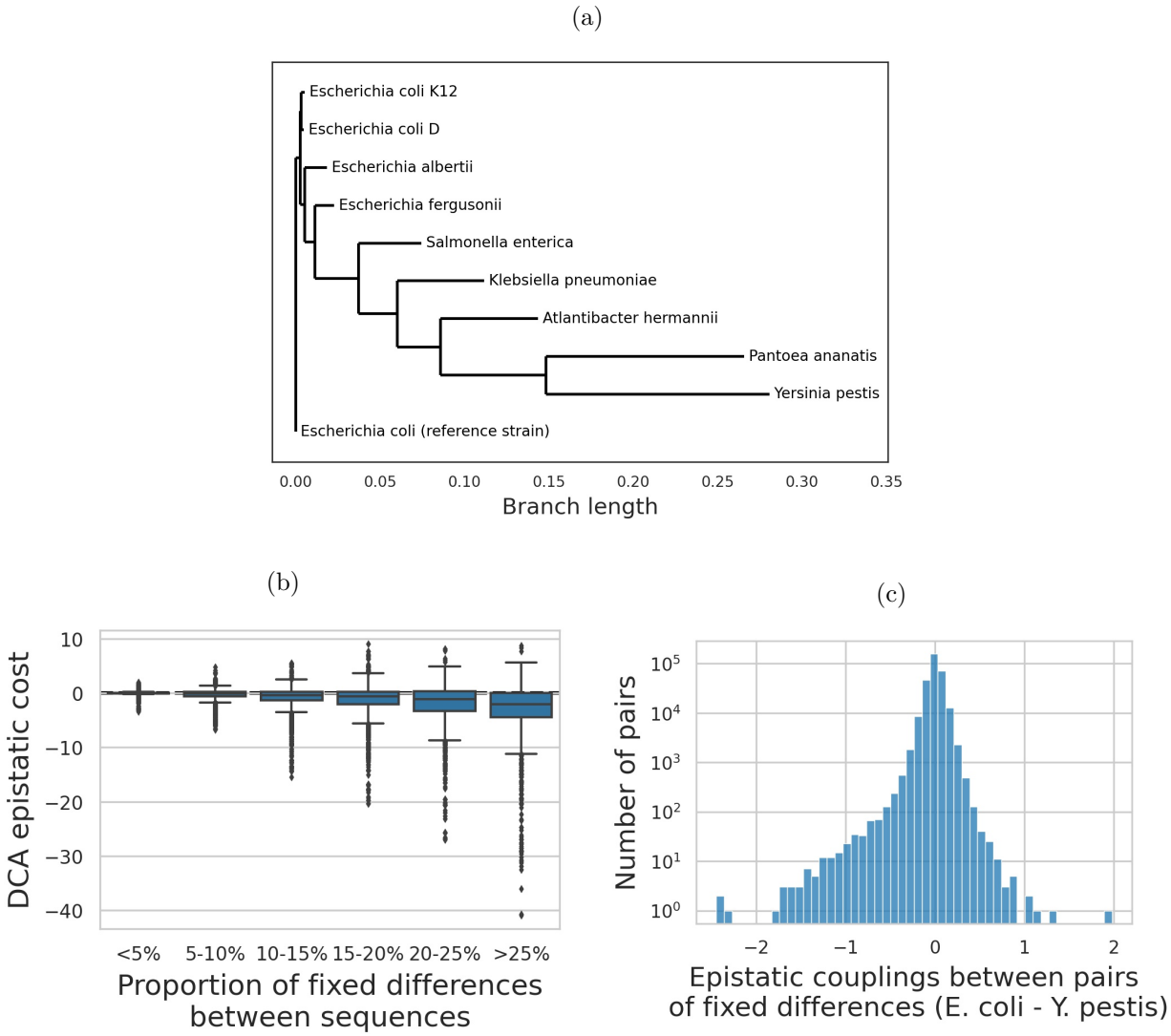


Figure 7: **Epistasis between fixed differences in a panel of diverged species. (a)** Phylogenetic tree of studied strains. Tree built from an amino-acid sequence alignment of 878 core genes. **(b)** DCA epistatic cost decreases with divergence. It is defined as the difference between the total change in statistical energy between pairs of sequences and the sum of single-mutant effects. Negative values correspond to positive epistasis: mutations are more beneficial (lower DCA score) taken altogether than the sum of their individual effects. **(c)** Distribution of epistatic couplings between pairs of fixed differences between *E. coli* and *Y. pestis*. The distribution is shifted towards negative values corresponding to positive epistatic couplings between fixed differences: they are better together than the sum of their individual effects. The relative small values of these couplings as compared to overall epistatic scores measured between entire sequences (Figure 7b) indicates that epistatic patterns build up gradually by an accumulation of many small couplings.

additional fixed difference modifies many couplings. These sequences have evolved naturally since their corresponding species diverged: the over-representation of positive epistatic couplings that we detect is consistent with evolution under long-term purifying selection [4].

rplK: a gene displaying a strong epistatic signal

rplK codes for the L11 protein of 50S subunit of the ribosome. It exhibits a strong signal of positive epistasis among the 14 differences fixed between *E. coli* and *Y. pestis*. This relatively small number of fixed differences offers a good opportunity to investigate how epistasis emerges at an individual protein level.

The range of epistatic couplings between fixed differences (Figure 8a) is consistent with Figure 7c: no very strong couplings but a clear tendency towards negative DCA values (*i.e.* positive epistasis). The strongest epistatic couplings correspond to pairs of residues that are in close vicinity in the 3D folding of the protein (distances <10Å in Figure 8b). We also observe a clear over-representation of couplings near -0.2 — as compared to the number of couplings near 0.2 — the majority of which correspond to more distant pairs of sites. Even if these residues are not necessarily in contact with one another, almost all of them cluster in the same protein domain (red dots in Figure 8c). This suggests that epistasis does not solely arise from direct contacts between few neighboring residues but also from more distant interactions between amino acids that contribute to the stability of the same protein domain. We previously found that DCA predicts about one fourth of amino-acid sites to be effectively coupled to a given residue. This figure clearly exceeds the number of residues that are in physical contact with an amino-acid site but could be explained by the hypothesis that sites belonging to the same protein domain are epistatically coupled with one another even if not in direct contact. These domains of correlated residues that co-evolve over long evolutionary times are reminiscent of protein sectors [19].

Discussion

The adaptationist and neutralist interpretations of biological diversity have long neglected epistasis. The complexity of modelling epistasis certainly contributes to explain why independent site models remain common in molecular evolution. Breen et al. first raised the possibility of epistasis being the primary factor in protein evolution [5]. If their methodology based on dN/dS computations underwent criticism [20], it clearly called for a deeper and more systematic study of epistasis across the genome. Experimental studies of mutations in different genetic backgrounds have confirmed an important role of epistasis in long-term evolution [7] [8]. However, they remain constrained to the analysis of single proteins. As abundant genetic data for both *E. coli* strains and diverged species have become available, data-driven approaches offer new opportunities. Through the concept of DCA-informed amino-acid landscapes, this allows for a large-scale data-driven study of epistasis on both short- and long-term evolution. The systematic analysis of wide genome portions has the potential to unveil much more widespread mechanisms than the potentially idiosyncratic studies led on specific proteins.

We find that DCA overperforms IND in predicting native amino acids as well as observed mutations and amino-acid site variability within *E. coli* species. Native amino acids arise from long-term evolution whereas observed polymorphisms and site variability within *E. coli* strains reflect short-term evolution. Thus, amino-acid landscapes appear relevant to study both short- and long-term evolution even though they are inferred from highly diverged species and can only capture evolutionary forces that are conserved for the entire family. Interestingly, it suggests that local adaptation of some specific strain to some specific ecological niche might add on top of these general constraints but does not dominate evolution. Our data analysis also emphasizes the importance of mutational biases on short evolutionary timescales. Neutral polymorphisms that

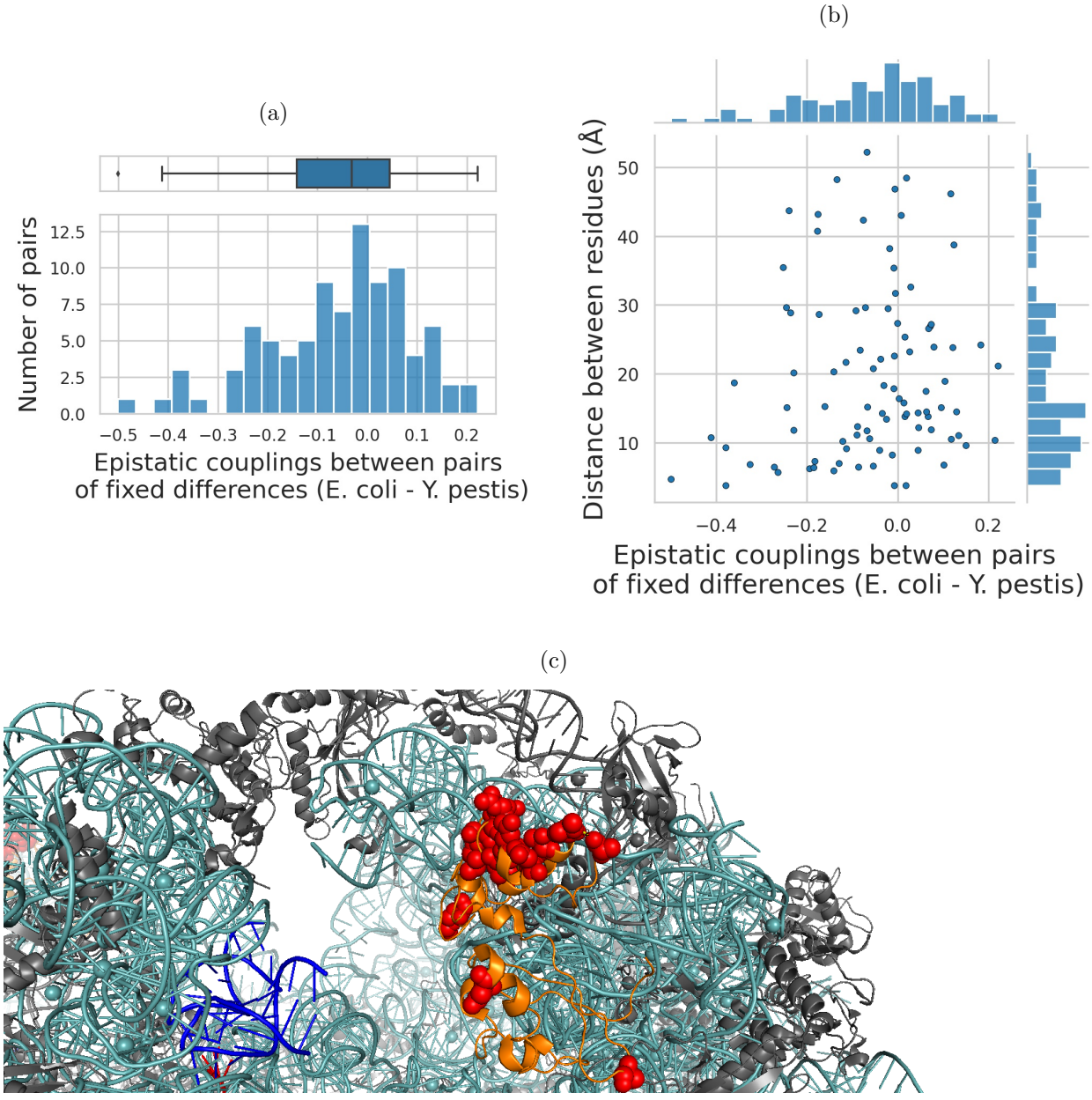


Figure 8: **Epistatic couplings between amino-acid differences that have fixed between *E. coli* and *Y. pestis* in rplK gene. (a)** Distribution of epistatic couplings between pairs of fixed differences. The left tail of negative DCA scores signals an over-representation of positive epistatic couplings. **(b)** Joint distribution of epistatic couplings values between pairs of residues harbouring a fixed difference and their physical distance in the 3D structure of the protein. The strongest couplings corresponds to residues that are in contact (<10Å). However, most of the couplings involve residues that are more distant than 10Å. **(c)** Representation of the 3D structure of rplK protein: the entire protein is coloured in orange, residues that differ between *E. coli* and *Y. pestis* are highlighted with red dots. Most of the fixed differences cluster together in the same domain, explaining why we observe a strong epistatic signal even though most of the pairs of fixed differences are not in physical contact.

require more than one SNP are virtually absent.

The better performances of DCA as compared to IND demonstrate the importance of taking epistasis into account to understand the effect of amino-acid changes. Recent achievements in synthetic biology prove that DCA captures enough of protein constraints to predict functional variants having less than 65% identity with amino-acid sequences used to train the DCA model [12]. They also experimentally demonstrate that an IND model fails at generating functional variants. This leads us to question the widespread use of softwares based on independent-site models such as SIFT [21] or Polyphen [22] to predict mutation effects. Here, we use DCA to characterize *E. coli* evolutive history. However, it paves the way to a far broader range of applications such as predicting adaptation or understanding molecular mechanisms underlying genetic diseases. In the latter case, DCA may prove useful at investigating cases of Dobzhansky–Muller incompatibilities [23] where amino-acid changes that have been fixed in distant species would be pathogenic to humans. For more applied purposes, DCA could be used to single out causative mutations associated to diseases in human genetics.

In agreement with Breen et al. [5], we find that context dependence dramatically reduces the variability observed at a given amino-acid site. Epistasis therefore plays an important role in evolution. However, we show that epistatic couplings between pairs of sites remain small compared to the typical effect of a mutation. Our data suggests that the strong context dependence of mutation effect comes from an accumulation of many small couplings. Consequently, most of the polymorphisms that arise within a species should have the same effect in all strains: the amino-acid landscape near a reference strain is locally smooth. On the contrary, the global landscape is rougher, with about one third of amino-acid sites where the effect of mutations drastically varies between distant species. Analysing a panel of closely diverged species through DCA modelling, we are able to show how these epistatic patterns gradually emerge with divergence.

Deep mutational scans have shown that positive epistasis between pairs of amino acids is less common than negative epistasis [3]. However, we show that positive epistatic couplings between residues dominate long-term evolution. Simulating the evolution of argT protein, Shah *et al.* have already noticed that, under purifying selection, mutations that fix are enriched in positive epistatic couplings with the rest of the background [4]. This is because purifying selection favors both mutations that are beneficial in all backgrounds and mutations that are beneficial in a given background due to epistatic couplings with the rest of the sequence. Here, we observe the same phenomena with real data and across hundreds of genes. Quantifying these effects experimentally would require performing deep mutational scans on several homologs at different distances with extremely accurate fitness estimates to detect small effects.

According to our findings, polymorphisms currently occurring in *E. coli* are close to neutral. On the contrary, fixed differences with *Y. pestis* tend to be deleterious in *E. coli* background. These observations perfectly fit a scenario of contingency and entrenchment: mutations are neutral at the time when they appear while being contingent on previous mutations and entrenched by subsequent mutations [4]. However, our approach to analysing context dependence is necessarily limited by the accuracy of DCA at modelling epistatic interactions. We have gathered evidence that DCA correctly captures the local neighborhood near *E. coli* sequences. These results combined with other assessments of DCA predictive power [11] [12] lead us to believe that it should be informative on how context dependence evolves with divergence. We cannot, though, reject the hypothesis that some of our observations are not a true biological signal but more artefacts of DCA modelling.

DCA model performances rely on the quality of the inter-species MSAs on which they are trained. Pfam domains MSAs are larger and more diverse than full gene MSAs because many different proteins across a wide range of organisms can share the same Pfam domain. As a consequence, DCA models trained on Pfam domain MSAs overperform those trained on full gene MSAs in predicting native amino acid and mutation effects (Supplementary Figures 1, 2, 3, 4, 5). However, full gene MSAs cover a larger fraction of the genome and DCA models trained on them

perform well at predicting site variability. The choice of the MSA reveals a trade-off between the DCA model performances and the fraction of the genome that can be covered. Depending on the intended applications, one might be favored over the other.

Since landscape models are inferred one by one for each protein, we can only capture intra-protein epistasis, but not any epistatic interaction between proteins. This is not an intrinsic limitation of the DCA approach, epistatic landscapes connecting two or more proteins may be inferred from joint MSAs [24]. However the size of the model grows quadratically with the number of amino-acid sites, making the inference of a full joint core genome landscape impractical in terms of computational time. Even by restricting to intra-protein epistasis, we obtain amino-acid landscapes that are relevant to study evolution on short and long timescales. The substantial context dependence of mutation effects that we uncover may be enhanced by accounting for inter-protein epistasis.

Materials and Methods

Datasets — inter-strain MSAs

61,157 *E. coli* genomes are downloaded from Enterobase [25]. 298,781,787 coding sequences are detected by Prokka [26]. In all analyses, the reference strain is the GA4805AA genome. For each gene in the reference strain, homologous sequences in the other genomes are retrieved using phmmer [27] (parameters: --popen 0.0001 --pextend 0.01) followed by a curation step where only sequences with less than 10 gaps after being aligned on the reference and more than 90% identity with the reference are kept. All genes with at least 60,000 homologous sequences are kept, these are referred to as core genes. Amino-acid sequences are aligned using mafft [28] and DNA sequences are reverse-aligned from amino-acid sequence alignments to preserve codon alignments. Two types of multiple sequence alignments (MSAs) are generated: one with the full-length core gene sequences (full gene MSAs, produced for genes that are present in at least 61,000 genomes) and one per Pfam domain [16] present in a core gene (Pfam domain MSAs).

Datasets — closely diverged species MSAs

The coding sequences of nine genomes of species closely related to *E. coli* are downloaded from Mage [29]: *Escherichia coli* K12 - chromosome ECK.1, *Escherichia coli* UMN026 - chromosome ESCUM.2, *Escherichia albertii* TW07627 - chromosome ESCAL.1, *Escherichia fergusonii* ATCC 35469T - chromosome EFER.2, *Salmonella enterica* subsp. arizonae serovar 62:z4,z23:– RSK2980 - chromosome NC_010067.1, *Klebsiella pneumoniae* 1162281 - WGS AFQL.1, *Atlantibacter hermannii* 4928STDY7071316 - WGS CABGLB01.1, *Pantoea ananatis* AJ13355 - chromosome NC_017531.1, *Yersinia pestis* Angola - chromosome NC_010159.1. Homologous sequences are retrieved using vsearch [30] usearch_global command against the reference genome (parameters: --strand plus --id 0.5 --query_cov 0.8 --target_cov 0.8 --maxaccepts 1). Only core genes (genes with a homologue in all 9 genomes) are kept. Amino-acid sequences are aligned by mafft [28]. Both full gene MSAs and Pfam domain MSAs are generated. Full genes MSAs are also concatenated to produce a unique MSA used to generate a phylogeny with FastTree [31].

Datasets — inter-species MSAs

For each full gene inter-strain MSA and full gene closely diverged species MSA, the corresponding full gene inter-species MSA is produced by querying the corresponding reference amino-acid sequence against UniRef30 2020-03 [32] using HHblits [33] followed by a curation step where sequences with more than 10% gap are removed from the MSA.

For each Pfam domain inter-strain MSA and Pfam domain closely diverged species MSA, the corresponding Pfam domain inter-species MSA is generated by downloading the full Pfam alignment from the Pfam 34.0 (March 2021) database [16] and aligning the reference sequence to the Pfam HMM using hmmalign [27]. All sites corresponding to inserts in the reference sequence are removed from the reference sequence, sites that are gapped in the reference sequence after aligning it to the Pfam HMM are removed from the Pfam MSA.

DCA and IND models

Direct-Coupling Analysis in the pseudo-likelihood maximization framework (plmDCA) [34] is used to train DCA models.

For each inter-strain MSA, the corresponding inter-species MSA is filtered to remove all sequences with >90% identity with the reference sequence. A DCA model is then trained if the filtered inter-species MSA contains more than 200 sequences.

For each closely diverged species MSA, a tree is built with FastTree [31] from the corresponding inter-species MSA concatenated to the closely diverged species MSA. The most recent common ancestor to the closely diverged species is inferred from this phylogeny. Any sequence of the inter-species MSA that descends from this most recent common ancestor is removed from the inter-species MSA. This is done in order to limit the risk of phylogenetic couplings to interfere with true epistatic interactions when training DCA models. A DCA model is then trained if the filtered inter-species MSA contains more than 200 sequences.

Each time a DCA model is trained, a corresponding IND model is produced from the frequencies of all possible amino acids or gaps at each position in the filtered inter-species MSA used to train the DCA model.

Data analysis

When no particular software is mentioned, analyses are performed using Python3 [35] and Biopython [36]. Amino-acid sites that are gapped in more than 20% of the sequences of the inter-species or intra-species MSAs are never considered.

Individual mutation effect prediction by DCA and IND models

A DCA model trained on an inter-species MSA of length L is composed of two matrices: h and J. They can be used to assign a statistical energy $E(a_1, ..., a_L)$ to any amino-acid sequence $(a_1, ..., a_L)$:

$$E(a_1, ..., a_L) = -\Sigma_{i<j} J_{ij}(a_i, a_j) - \Sigma_i h_i(a_i)$$

The $h_i(a_i)$ are site-dependent biases taking into account the importance of single amino acids in individual sequence positions; the $J_{ij}(a_i, a_j)$ are epistatic couplings connecting the amino acids in pairs of positions. The function E is inferred to maximize the pseudolikelihood of the sequences in the inter-species MSA.

Two amino-acid sequences can be compared to one another by simply making the difference between their statistical energy values. In particular, the DCA score of mutating amino acid α into amino acid β at position i in the amino-acid background $(a_1, ..., a_{i-1}, a_{i+1}, ..., a_L)$ is given by:

$$\Delta E_i = E(a_1, ..., a_{i-1}, \beta, a_{i+1}, ..., a_L) - E(a_1, ..., a_{i-1}, \alpha, a_{i+1}, ..., a_L) = \\ h_i(\alpha) - h_i(\beta) + \Sigma_{j \neq i} J_{ij}(\alpha, a_j) - \Sigma_{j \neq i} J_{ij}(\beta, a_j)$$

The DCA score of the mutation $\alpha \to \beta$ at locus i in the amino-acid background $(a_1, ..., a_{i-1}, a_{i+1}, ..., a_L)$ can be turned into a conditional probability of observing the amino acid β at locus i, given that the other positions take amino acids $a^0_{\setminus i} = (a_1, ..., a_{i-1}, a_{i+1}, ..., a_L)$. Within our DCA-based modelling framework, this quantity reads:

$$P_i(\beta|a^0_{\setminus i}) = \exp\left\{h_i(\beta) + \Sigma_{i\neq j} J_{ij}(\beta, a_i)\right\}/z_i,$$

with the normalization z_i chosen such that P becomes a probability distribution over the values of β, *i.e.* over the 20 theoretically possible amino acids in position i (gaps are not considered, since we study the effects of amino-acid substitutions and not deletions).

The probability of observing amino acid β at locus i in IND is given by the frequency of amino acid β at locus i in the inter-species MSA: $f_i(\beta)$.

Context-Independent and Context-Dependent Entropies

The Context-Independent Entropy (CIE) is the standard column entropy of the inter-species MSAs. It is calculated from the position-specific amino-acid frequencies $f_i(a)$, measuring the fraction of sequences in the inter-species MSA having amino acid a at locus i:

$$CIE_i = -\Sigma_a f_i(a) \log_2 f_i(a).$$

The Context-Dependent Entropy (CDE) is computed from the conditional probabilities of observing the amino acid a at locus i in the amino-acid context of the reference strain $P_i(a|a^0_{\setminus i})$:

$$CDE_i(a^0_{\setminus i}) = -\Sigma_a P_i(a|a^0_{\setminus i}) \log_2 P_i(a|a^0_{\setminus i})$$

The difference between CIE and CDE gives the information gain (IG) provided by the context:

$$IG_i(a^0_{\setminus i}) = CIE_i - CDE_i(a^0_{\setminus i})$$

1-SNP mutations

All codons in the reference genome are analysed in order to record all possible synonymous mutations and non-synonymous mutations that can be obtained by mutating them exactly once. These mutations are referred to as 1-SNP mutations. For non-synonymous mutations, the corresponding amino acids encoded by the mutated codons are also recorded.

The probability of observing an amino acid can be computed from an IND model restricted to 1-SNP mutations, by setting to 0 all entries of the $f_i(a)$ vector that do not correspond to 1-SNP mutations and re-normalizing $f_i(a)$. These new probabilities can be used to compute a CIE that is restricted to 1-SNP mutations.

The probability of observing an amino acid can be computed from a DCA model restricted to 1-SNP mutations, by setting to 0 all entries of the $P_i(a|a^0_{\setminus i})$ vector that do not correspond to 1-SNP mutations and re-normalizing $P_i(a|a^0_{\setminus i})$. These new probabilities can be used to compute a CDE that is restricted to 1-SNP mutations.

Epistatic cost

Epistasis is defined as the deviation from additivity of mutational effects. Having two mutations in sites i and j of a protein, the total mutational effect ΔE_{ij}, defined as the difference in statistical energy between the double mutant and the reference sequences, can be compared to the sum $\Delta E_i + \Delta E_j$ of the effects of the two single-site mutations, individually inserted into the reference sequence. The epistatic cost for substituting the reference residues α_i, α_j with β_i, β_j is the difference:

$$\Delta\Delta E_{ij} = \Delta E_{ij} - \Delta E_i - \Delta E_j = J_{ij}(\alpha_i, \beta_j) + J_{ij}(\beta_i, \alpha_j) - J_{ij}(\beta_i, \beta_j) - J_{ij}(\alpha_i, \alpha_j)$$

Similarly, the epistatic cost of an arbitrary number of mutations is the difference between the total mutational effect $\Delta E_{ij...n}$ of the mutations altogether (*i.e.* the difference in statistical energy between the mutant and the reference sequences) and the sum $\Delta E_i + \Delta E_j + ... + \Delta E_n$ of the effects of the all single-site mutations, individually inserted into the reference sequence:

$$\Delta\Delta E_{ij...n} = \Delta E_{ij...n} - \Delta E_i - \Delta E_j - ... - \Delta E_n$$

For each inter-strain MSA, sequences with exactly two mutations compared to the reference sequence and no gap are gathered. The total mutational effect ΔE_{ij} of each pair of mutations in the reference sequence is computed and compared to the sum $\Delta E_i + \Delta E_j$ of the effects of the two single-site mutations, individually inserted into the reference sequence. For all pairs of fixed differences between *Y. pestis* and the reference sequences, the epistatic coupling ΔE_{ij} is also recorded.

For closely diverged species MSAs, the epistatic cost between each pair of homologous sequences with no more than one gap difference (but any arbitrary number of other missense mutations) is computed as well as the proportion of fixed differences between them.

When comparing epistatic cost between pairs of fixed differences in rplK to the distance between these residues in the 3D structure of the protein, the 4V6E PDB structure is used [37]. It is displayed using PyMOL [38] .

Effective proportion of residues coupled to an amino-acid site

DCA models are based on a matrix J of pairwise epistatic couplings between residues in a sequence. The Inverse Participation Ratio (IPR) quantifies how diffuse epistatic couplings involving a residue at position i are. It is computed as follow:

$$IPR_i = \Sigma_{j\neq i}(J_{ij}(a_i, a_j)^2 / \Sigma_{k\neq i} J_{ik}(a_i, a_k)^2)^2, \text{ with } (a_1, ..., a_L) \text{ being the reference sequence}$$

IPR_i corresponds to the inverse of the effective number of sites that are epistatically coupled with a position i. The effective proportion of residues coupled to an amino-acid site at position i in a sequence of size L is derived from IPR_i as being $1/(IPR_i.L)$.

Code availability

Codes in Python are available at `https://github.com/GiancarloCroce/DCA_polymorphism_Ecoli`.

Data availability

Data is available on Zenodo (DOI 10.5281/zenodo.5774192).

Contributions

L.V., G.C., O.T. and M.W. designed the analyses and wrote the paper. L.V. and G.C. performed the analyses. M.P. and E.R. gathered and prepared genetic sequence data.

Acknowledgements

We are thankful to Alaksh Choudhury for help with protein 3D structure visualization. We also wish to thank Juan Rodriguez-Rivas.

Our work was partially funded by the French Agence Nationale pour la Recherche ANR GeWiEp (ANR-18-CE35-0005-01, to L.V. and O.T.), the French Fondation pour la Recherche Médicale (EQU201903007848, to L.V. and O.T.), the PhD program AMX of École polytechnique and Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (to L.V.) and EU H2020 Research and Innovation Programme MSCA-RISE-2016 (Grant Agreement No. 734439 InferNet, to M.W.).

References

[1] Ernst Mayr. How to carry out the adaptationist program? *The American Naturalist*, 121(3):324–334, 1983. Publisher: University of Chicago Press.

[2] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[3] Tyler N. Starr and Joseph W. Thornton. Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218, July 2016.

[4] Premal Shah, David M. McCandlish, and Joshua B. Plotkin. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*, 112(25):E3226–E3235, June 2015.

[5] Michael S. Breen, Carsten Kemena, Peter K. Vlasov, Cedric Notredame, and Fyodor A. Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, October 2012.

[6] J. Arjan G.M. de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, July 2014.

[7] Mark Lunzer, G. Brian Golding, and Antony M. Dean. Pervasive Cryptic Epistasis in Molecular Evolution. *PLOS Genetics*, 6(10):e1001162, October 2010. Publisher: Public Library of Science.

[8] Jamie T. Bridgham, Eric A. Ortlund, and Joseph W. Thornton. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature*, 461(7263):515–519, September 2009. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7263 Primary_atype: Research Publisher: Nature Publishing Group.

[9] Jose Alberto de la Paz, Charisse M. Nartey, Monisha Yuvaraj, and Faruck Morcos. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proceedings of the National Academy of Sciences*, 117(11):5873–5882, March 2020.

[10] Matteo Bisardi, Juan Rodriguez-Rivas, Francesco Zamponi, and Martin Weigt. Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. *arXiv:2106.02441 [cond-mat, q-bio]*, June 2021. arXiv: 2106.02441.

[11] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, December 2011.

[12] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, July 2020.

[13] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, 33(1):268–280, January 2016.

[14] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017. Bandiera_abtest:

a Cg_type: Nature Research Journals Number: 2 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational models;Molecular evolution;Mutation;Protein function predictions Subject_term_id: computational-models;molecular-evolution;mutation;protein-function-predictions.

[15] Alejandro Couce, Larissa Viraphong Caudwell, Christoph Feinauer, Thomas Hindré, Jean-Paul Feugeas, Martin Weigt, Richard E. Lenski, Dominique Schneider, and Olivier Tenaillon. Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proceedings of the National Academy of Sciences*, 114(43):E9026–E9035, October 2017.

[16] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, and Erik LL Sonnhammer. The Pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004. Publisher: Oxford University Press.

[17] Benjamin A. Rogers, Hanna E. Sidjabat, and David L. Paterson. Escherichia coli O25b-ST131: a pandemic, multiresistant, community-associated strain. *Journal of Antimicrobial Chemotherapy*, 66(1):1–14, January 2011.

[18] C. Anders Olson, Nicholas C. Wu, and Ren Sun. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, 24(22):2643–2651, November 2014.

[19] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138(4):774–786, August 2009.

[20] David M. McCandlish, Etienne Rajon, Premal Shah, Yang Ding, and Joshua B. Plotkin. The role of epistasis in protein evolution. *Nature*, 497(7451):E1–E2, May 2013.

[21] Pauline C. Ng and Steven Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, July 2003. Publisher: Oxford Academic.

[22] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, 76(1):7–20, 2013. Publisher: Wiley Online Library.

[23] Alexey S. Kondrashov, Shamil Sunyaev, and Fyodor A. Kondrashov. Dobzhansky–Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences*, 99(23):14878–14883, November 2002. Publisher: National Academy of Sciences Section: Biological Sciences.

[24] Hendrik Szurmant and Martin Weigt. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Current Opinion in Structural Biology*, 50:26–32, June 2018.

[25] Zhemin Zhou, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, Agama Study Group, and Mark Achtman. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny and Escherichia core genomic diversity. *Genome Research*, page gr.251678.119, December 2019. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[26] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014. Publisher: Oxford University Press.

[27] Robert D. Finn, Jody Clements, and Sean R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl_2):W29–W37, July 2011.

[28] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002. Publisher: Oxford Academic.

[29] David Vallenet, Alexandra Calteau, Stéphane Cruveiller, Mathieu Gachet, Guillaume Gautreau, Adrien Josso, Aurélie Lajus, Jordan Langlois, Jonathan Mercier, and Hugo Pereira. MICROSCOPE: an integrated platform for the Exploration and Curation of Microbrial Genomes. *Biologie, Informatiqueet Mathématiques*, page 119, 2017.

[30] Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, October 2016. Publisher: PeerJ Inc.

[31] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, March 2010.

[32] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. Publisher: Oxford University Press.

[33] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2):173–175, 2012. Publisher: Nature Publishing Group.

[34] Magnus Ekeberg, Tuomo Hartonen, and Erik Aurell. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, 276:341–356, November 2014.

[35] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

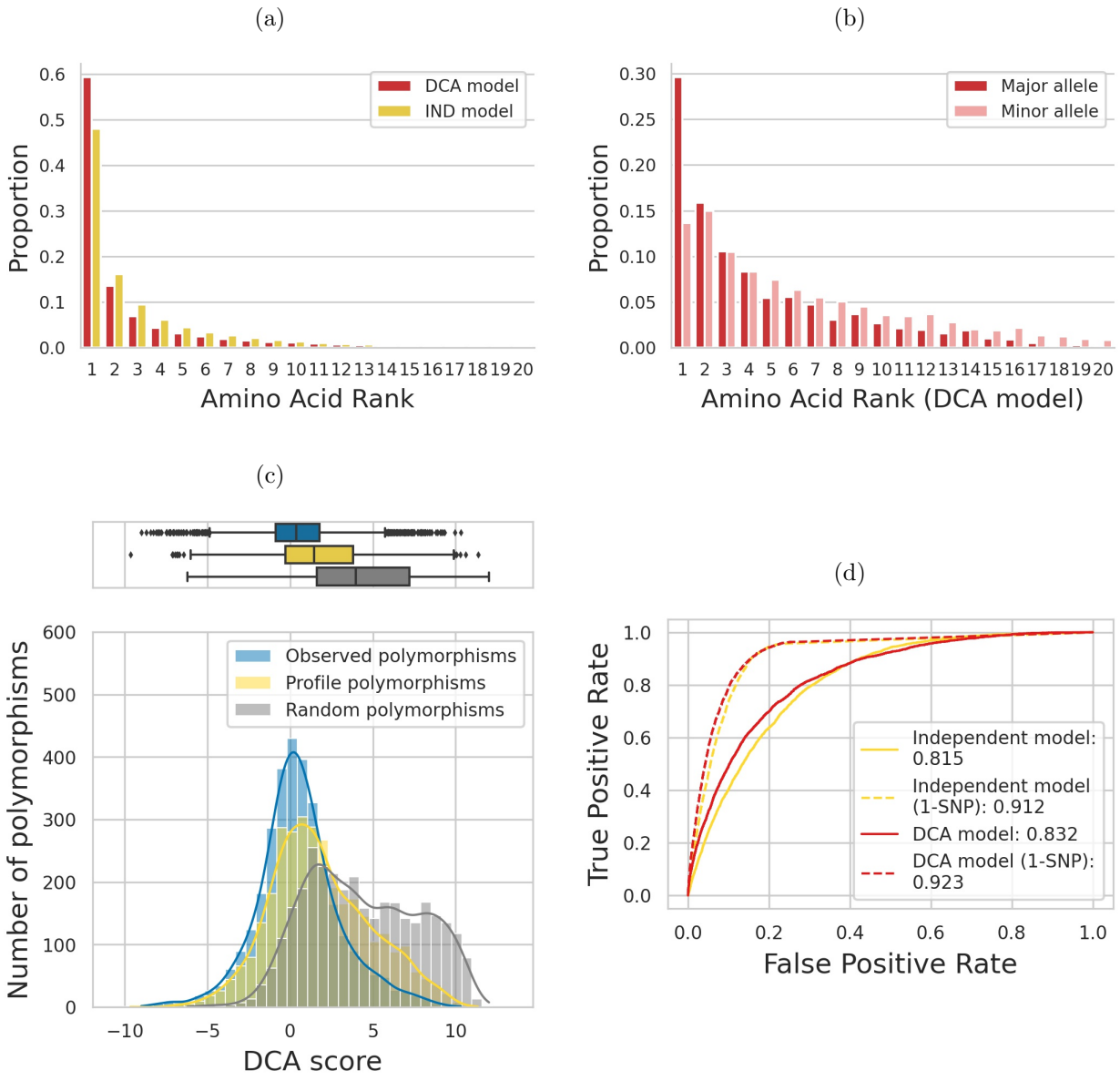
[36] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, and Bartek Wilczynski. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. Publisher: Oxford University Press.

[37] Wen Zhang, Jack A Dunkle, and Jamie HD Cate. Structures of the ribosome in intermediate states of ratcheting. *Science*, 325(5943):1014–1017, 2009.

[38] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

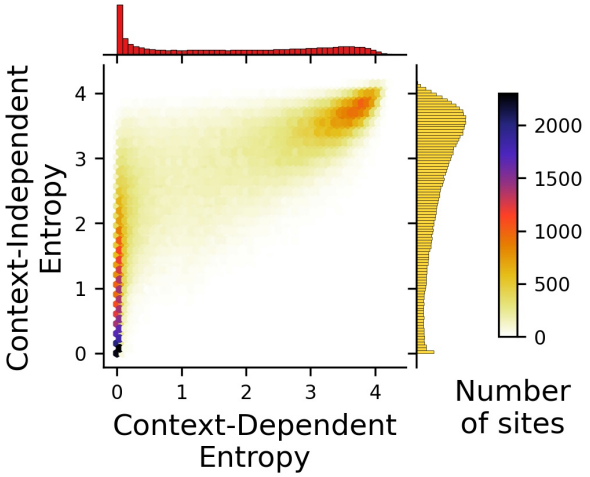
Supplementary Information

Supplementary Figures

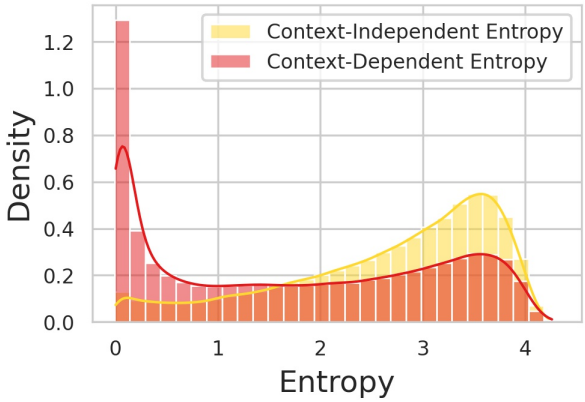


Supplementary Figure 1: **Predicted effects of observed amino acids using an IND model that neglects epistasis or a DCA model that incorporates pairwise epistasis. Models trained on full genes.** **(a)** Rank of native amino acid in the reference strain as compared to all 20 possible amino acids. **(b)** DCA rank of major and minor allele for all sites that are polymorphic at a >5%-threshold, among all 20 possible amino acids. **(c)** Distribution of DCA scores of non-synonymous polymorphisms observed at frequencies >5% across the >60,000 strains (blue) compared to mutations sampled from an IND model (yellow) or to random mutations (grey). **(d)** ROC curves of different models for predicting polymorphisms observed at >5% frequency in *E. coli*.

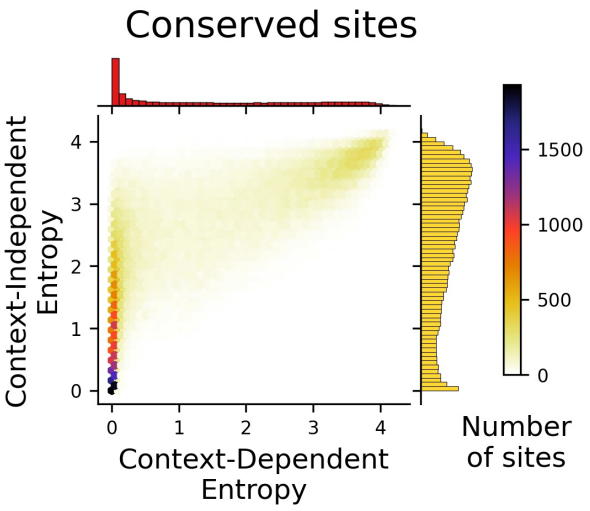
(a)



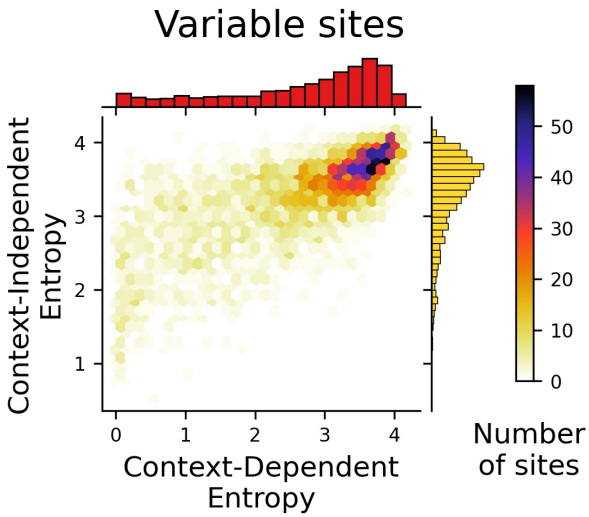
(b)



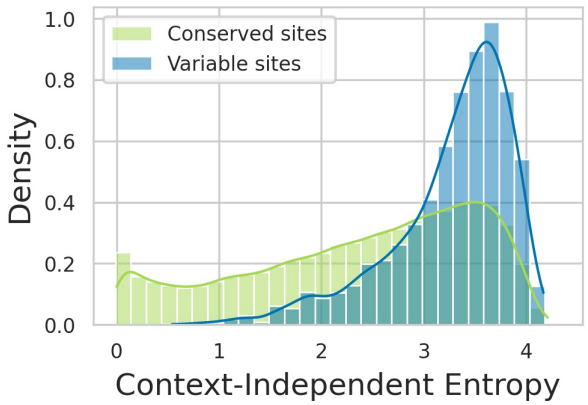
(c)



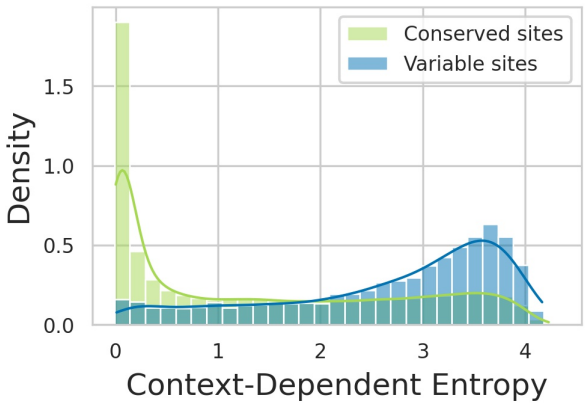
(d)

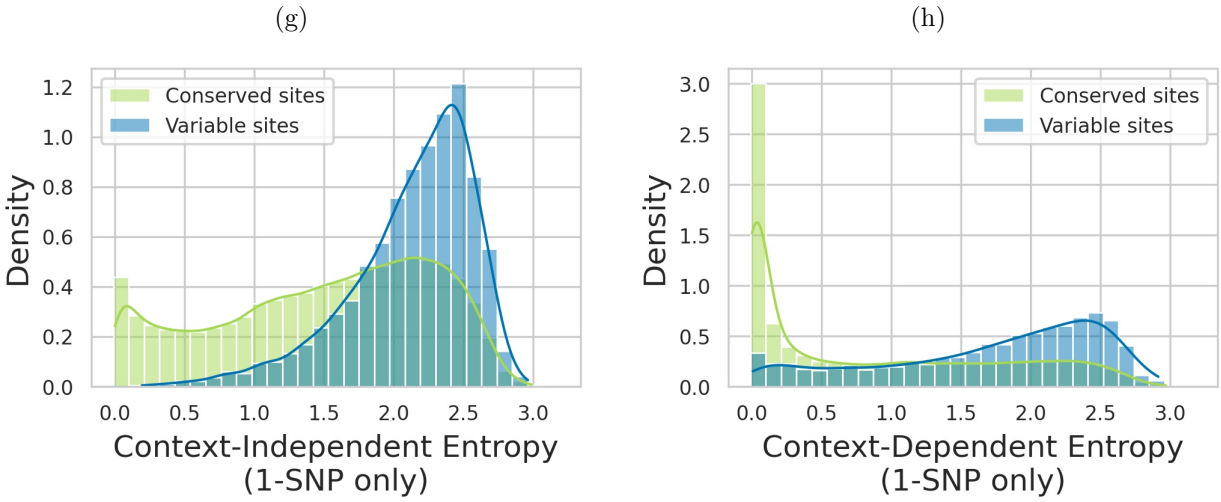


(e)

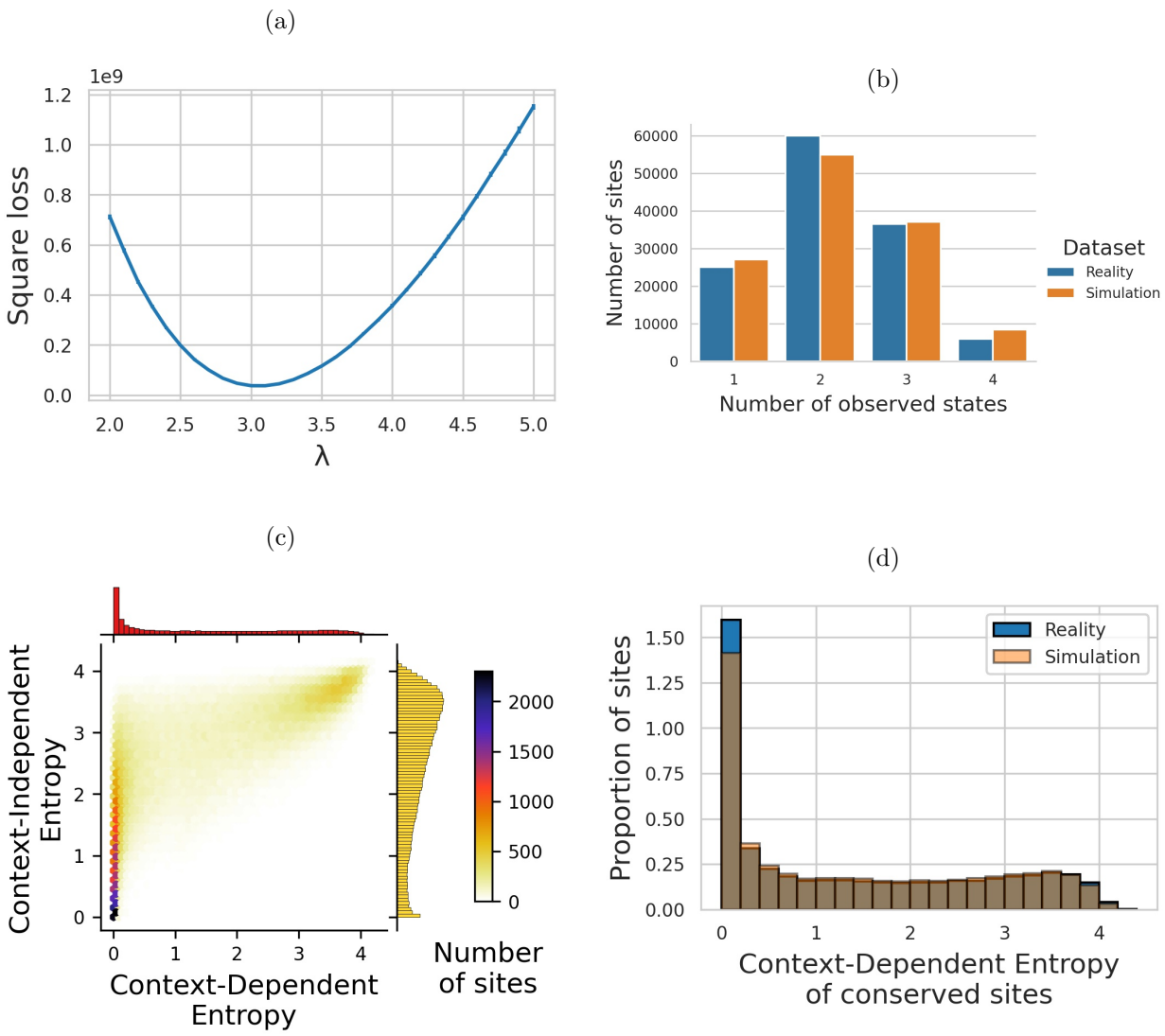


(f)

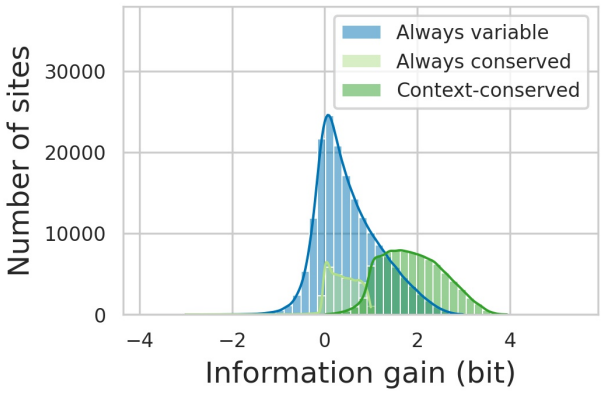




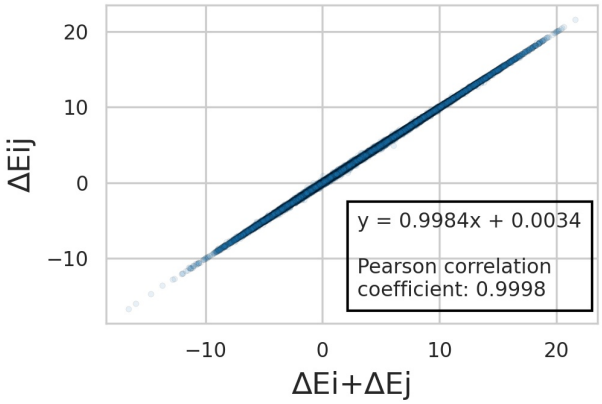
Supplementary Figure 2: **Predicting the variability of amino-acid sites and amino-acid sites that are conserved or polymorphic in *E. coli*. Comparison of the performances of an IND and a DCA models trained on full genes.** **(a)** Bivariate histogram of CDE and CIE for all sites in the dataset. **(b)** Marginal distributions of CDE and CIE for all sites in the dataset. **(c)** Bivariate histogram of CDE and CIE for sites that are conserved across >60,000 strains of *E. coli*. **(d)** Bivariate histogram of CDE and CIE for sites that are polymorphic at a 5% threshold across >60,000 strains of **E. coli**. **(e)** Distribution of CIE for conserved (green) and polymorphic (blue) sites in *E. coli*. **(f)** Distribution of CDE for conserved (green) and polymorphic (blue) sites in *E. coli*. **(g)** CIE for 1-SNP mutations. **(h)** CDE for 1-SNP mutations.



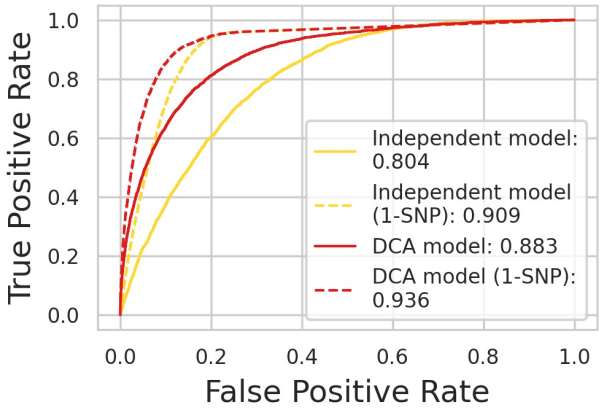
Supplementary Figure 3: **Simulations of synonymous and non-synonymous diversity occuring on full genes.** **(a)** Simulation of synonymous diversity. For each λ ranging from 2 to 5 with a 0.1 step-size, 20 simulations are run. The square loss between the amount of simulated synonymous diversity and the real amount observed in the dataset is computed. The best λ parameter is 3.1. **(b)** Simulation of synonymous diversity. Average results of the 20 simulations of synonymous diversity with $\lambda = 3.1$. We have focussed on sites where there are exactly four possible 1-SNP synonymous mutations. As we can see synonymous diversity is not saturated (sites with all four possible synonymous codons observed in the dataset are rare). Simulations achieve very good fit of the observed reality even with a basic model like JC69 that ignores differences in mutation rates between nucleotide pairs. **(c)** Simulation of non-synonymous diversity. Bivariate histogram of CDE and CIE for sites that are conserved in the simulated dataset produced with parameter $\lambda = 3.1$. Most of the sites cluster on the left peak of low CDE. However, as observed in the real dataset, some of the sites where no mutation occurred have a high CDE. **(d)** Simulation of non-synonymous diversity. Comparison of CDE distributions of real conserved sites (sites conserved across >60,000 strains in the dataset) and simulated conserved sites (sites where no mutation was simulated).



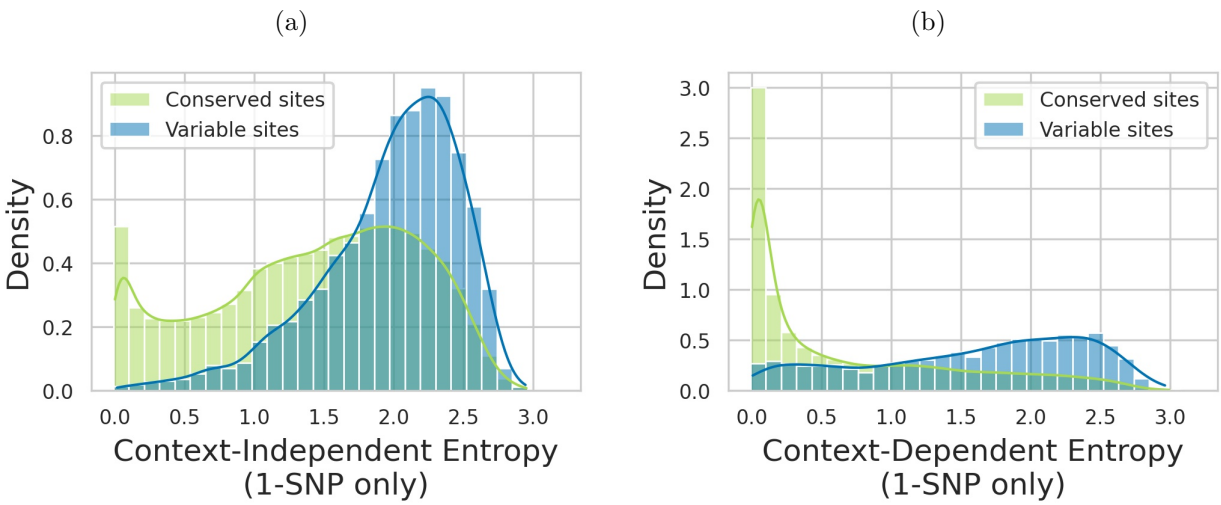
Supplementary Figure 4: **Quantifying the effect of the context in reducing amino-acid site variability with models trained on full genes.** Information gain quantifies the difference between an amino-acid site variability across distant species and its potential variability in *E. coli*. Sites that are variable across distant species (CIE$\geq$1) but conserved in *E. coli* (CDE$<$1) are the ones with the highest information gains (dark green distribution).



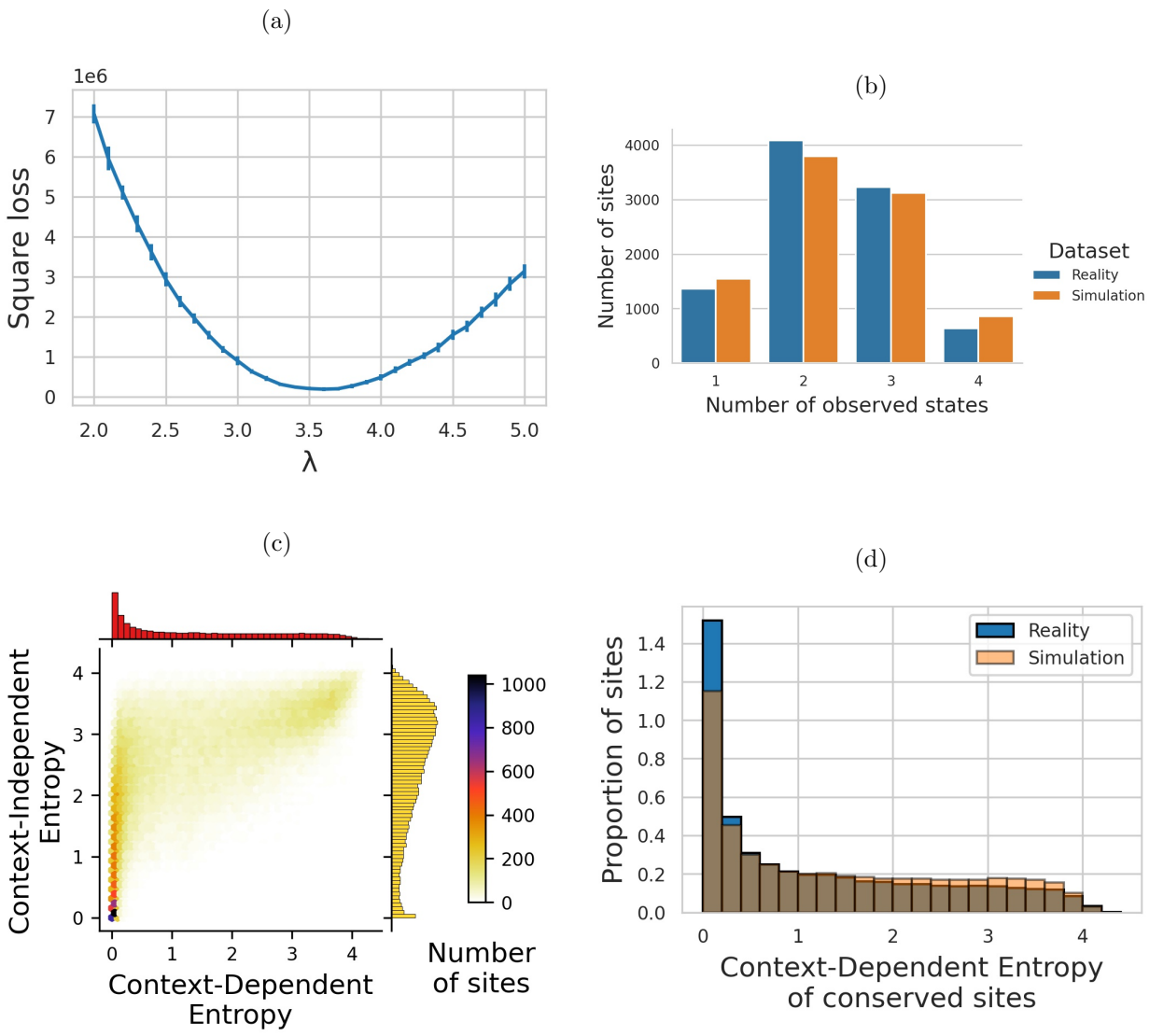
Supplementary Figure 5: **Epistasis observed in *E. coli* with models trained on full genes.** Mutational effect ΔE_{ij} of observed double mutations with respect to the reference, plotted against the sum $\Delta E_i + \Delta E_j$ of the individual mutations.



Supplementary Figure 6: **ROC curves of different models for predicting polymorphisms observed at >5% frequency in *E. coli*. Models trained on Pfam domains.**



Supplementary Figure 7: **CIE and CDE computed for 1-SNP mutations from the reference codon with models computed on Pfam domains.** The total number of amino acids that can be observed with no more than 1 SNP never exceeds 9, corresponding to a maximal entropy value of about 3.2. **(a)** CIE for 1-SNP mutations. **(b)** CDE for 1-SNP mutations.



Supplementary Figure 8: **Simulations of synonymous and non-synonymous diversity occuring on Pfam domains.** **(a)** Simulation of synonymous diversity. For each λ ranging from 2 to 5 with a 0.1 step-size, 20 simulations are run. The square loss between the amount of simulated synonymous diversity and the real amount observed in the dataset is computed. The best λ parameter is 3.6. **(b)** Simulation of synonymous diversity. Average results of the 20 simulations of synonymous diversity with $\lambda = 3.6$. We have focussed on sites where there are exactly four possible 1-SNP synonymous mutations. As we can see synonymous diversity is not saturated (sites with all four possible synonymous codons observed in the dataset are rare). Simulations achieve good fit of the observed reality even with a basic model like JC69 that ignores differences in mutation rates between nucleotide pairs. **(c)** Simulation of non-synonymous diversity. Bivariate histogram of CDE and CIE for sites that are conserved in the simulated dataset produced with parameter $\lambda = 3.6$. Most of the sites cluster on the left peak of low CDE. However, as observed in the real dataset, some of the sites where no mutation occurred have a high CDE. **(d)** Comparison of CDE distributions of real conserved sites (sites conserved across >60,000 strains in the dataset) and simulated conserved sites (sites where no mutation was simulated).

Supplementary Methods

Simulations: The simulations are based on Jukes-Cantor 1969 model (JC69). Two sets of simulations are performed. The first one is led on synonymous mutations in order to calibrate the mutation rate parameter. Simulations of synonymous and non-synonymous mutations under selection are then performed using the previously-inferred mutation rate parameter and DCA score as a proxy for fitness cost of non-synonymous mutations. Synonymous mutations are supposed to be neutral (DCA score of zero). Only sites where the reference codon is the major allele are considered.

For each codon with exactly three synonymous 1-SNP mutations, a random number N is sampled from a Poisson distribution of parameter λ: it corresponds to the total number of synonymous mutations occuring at this site. N codons are then sampled with replacement from the three synonymous mutations possible at this site (with equiprobability). Each of these codons is kept with an acceptance probability of 50%. The number of different codons that are accepted at each site is recorded. Its minimal value is one (the reference codon alone) and the maximal value it can take is four (the reference codon and all three others synonymous mutations). 20 simulations for each λ ranging from two to five with a 0.1 step-size are run to select the value of λ for which the average number of synonymous mutations per site is the closest to what is observed in the >60,0000-strain dataset.

Synonymous and non-synonymous mutations are then simulated for all the sites of the dataset. For each site, a total number of mutations, N, is sampled from a Poisson distribution of parameter λ (using the λ estimated with synonymous mutations). N codons are sampled with replacement from the nine possible codons (with equiprobability). Each of these codons is kept with an acceptance probability p = P(observing derived amino acid at locus $i|a^0_{\setminus i}$)/(P(observing derived amino acid at locus $i|a^0_{\setminus i}$) + P(observing reference amino acid at locus $i|a^0_{\setminus i}$)), where P(observing a given amino acid at locus $i|a^0_{\setminus i}$) is the conditional probability of observing this amino acid at locus i given the amino-acid context of the reference strain, computed with DCA.