Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution

Matteo Bisardi,[1,2,†] Juan Rodriguez-Rivas,[2,†] Francesco Zamponi,[1,*] and Martin Weigt ⓘ[*,2]

[1]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France
[2]Biologie Computationnelle et Quantitative LCQB, Institut de Biologie Paris Seine, CNRS, Sorbonne Université, Paris, France

[†]These authors contributed equally to this work.
*Corresponding authors: E-mails: francesco.zamponi@ens.fr; martin.weigt@sorbonne-universite.fr.
Associate editor: Banu Ozkan

Abstract

During their evolution, proteins explore sequence space via an interplay between random mutations and phenotypic selection. Here, we build upon recent progress in reconstructing data-driven fitness landscapes for families of homologous proteins, to propose stochastic models of experimental protein evolution. These models predict quantitatively important features of experimentally evolved sequence libraries, like fitness distributions and position-specific mutational spectra. They also allow us to efficiently simulate sequence libraries for a vast array of combinations of experimental parameters like sequence divergence, selection strength, and library size. We showcase the potential of the approach in reanalyzing two recent experiments to determine protein structure from signals of epistasis emerging in experimental sequence libraries. To be detectable, these signals require sufficiently large and sufficiently diverged libraries. Our modeling framework offers a quantitative explanation for different outcomes of recently published experiments. Furthermore, we can forecast the outcome of time- and resource-intensive evolution experiments, opening thereby a way to computationally optimize experimental protocols.

Key words: protein evolution, fitness landscapes, sequence space, epistasis, data-driven models.

Introduction

In the course of evolution, biological sequences encoding proteins explore functional sequence space. The observable sequence variability between homologous sequences, that is, sequences connected by common ancestry, results from a delicate balance between mutation and selection. Mutations tend to randomize sequences, whereas natural selection prunes most of those mutations having a deleterious effect on fitness. When analyzing large databases of homologous protein families (Mistry et al. 2021), we therefore find sequences with 70–80% different amino acids, but highly conserved functional and structural properties.

In turn, it is possible to search for statistical patterns in ensembles of homologous proteins (Durbin et al. 1998), using tools borrowed from statistical inference and unsupervised machine learning, and to relate them to selective constraints acting in these proteins. The most prominent signal is conservation; a position in a protein not (or rarely) changing amino acid over extended evolutionary time scales, is likely to play an important role in the protein's function (e.g., active sites in enzymes) or for the protein's structural stability (e.g., residues buried in the protein core).

A second type of statistical signal has received a lot of attention during the last decade (De Juan et al. 2013; Levy et al. 2017; Cocco et al. 2018). The correlations between the amino acids present in pairs of residue positions can be extracted via global statistical models like those used in direct coupling analysis (DCA) (Weigt et al. 2009; Morcos et al. 2011), Gremlin (Balakrishnan et al. 2011), or PSICOV (Jones et al. 2012). This signal of residue–residue coevolution results from epistatic couplings between residues in structural contact in the folded proteins, that is, of residue pairs in direct physical interaction in the 3D structure of the protein, even if possibly located at long distance along the primary amino acid sequence. Coevolutionary methods, in particular when used as input for structurally supervised deep-learning methods like RaptorX (Xu 2019), DeepMetaPSICOV (Greener et al. 2019), AlphaFold (Senior et al. 2020), or trRosetta (Yang et al. 2020), have recently induced a revolution in protein-structure prediction, reaching unprecedented accuracy in computationally predicted structures close to the accuracy of experimentally determined structures (Jumper et al. 2021). Hundreds of previously unknown protein structures have been predicted this way (Ovchinnikov et al. 2017; Tunyasuvunakool et al. 2021).

However, coevolutionary methods rely on the availability of large alignments of homologous but diverged proteins, since they rely on statistical signatures extracted from sequence variability (Haldane and Levy 2019). Recently, two

groups have independently asked the question, if experimentally generated sequences can be used instead of natural homologs for contact prediction (Fantini et al. 2020; Stiffler et al. 2020). To this aim, they have proposed and performed similar experiments. First, starting from a given wildtype sequence, they have iterated several rounds of alternating sequence diversification via error-prone polymerase chain reaction (epPCR) (Cadwell and Joyce 1992), and selection for functionality (antibiotic resistance for both experiments). In contrast to traditional directed evolution (Arnold 1998, 2018), selection was very weak (low antibiotic concentrations), so proteins are not simply optimized for function, but may diversify their sequences while maintaining a certain level of functionality. After a certain number of rounds, the resulting sequence library was sequenced, to provide the data for statistical learning.

The resulting functional sequence libraries were quite diversified, with typical distances of 10–15% of the sequence length from the wildtype protein used as a starting point. This is much less than in natural protein families, characterized typically by average distances of 70–80% between homologs. However, the simultaneous emergence of about 10–40 mutations, and the depth of more than $10^4 - 10^5$ sequences in the experimentally evolved libraries, could make the detection of epistasis, and thus contact prediction, possible (Fantini et al. 2020; Stiffler et al. 2020).

Interestingly, both teams have run plmDCA (Ekeberg et al. 2013), or evCouplings based on plmDCA (Hopf et al. 2019), on the data—with very different results. Although the contact signal in (Fantini et al. 2020) was quite weak, and mostly concentrated to nearby positions along the sequence, (Stiffler et al. 2020) found a sufficiently accurate contact prediction to enable the subsequent construction of a precise structural model.

To understand the differences in results given the similarity in approaches, we have developed a modeling scheme, which allows us to simulate protein evolution in a data-driven sequence landscape. Comparison of simulated and experimental data of both experiments shows that our simulations reproduce quantitatively the experimental observations. Furthermore, the simulation scheme allows us to control important parameters of the experiments, like the evolutionary distance from the wildtype in the final evolved library, the sequencing depth of the library, or the strength of selection. We find that our model is able to explain the difference in contact prediction between the two experiments in terms of sequence divergence and sequencing depth.

The agreement between simulations and experiments suggests that our modeling framework allows for a quantitative analysis of important questions about protein evolution, like the mechanism underlying sequence space exploration and the emergence of signatures of epistasis with sequence divergence, compare also the related Sequence Evolution with Epistatic Contributions (SEEC) model (de la Paz et al. 2020). Beyond such basic questions in evolutionary biology, our framework has also the potential to help in optimizing experimental design. To give an example, our simulations predict that both experiments would have benefited from slightly weaker selection, represented by slightly lower antibiotic concentrations. This would have enabled a faster exploration of the neighborhood of the wildtype sequence and the occurrence of slightly more deleterious mutations, which have a better chance to be coupled by epistasis than the predominantly neutral mutations accepted at strong selection. Such predictions are very interesting, since our computational approach is efficient and can be applied to thousands of protein families, whereas the experiments are expensive in time and resources. Guiding them to increase the success probability may therefore be an impactful strategy. For instance, our approach can be used to explore different protocols, such as alternating cycles of strong and weak selection.

Results

The general procedure of our modeling approach is graphically illustrated in figure 1. In this section, we first describe the data-driven sequence landscape, which is inferred from multiple sequence alignments (MSA) of natural homologs of the experimentally studied wildtype, that is, from data unrelated to the experiment. As a first check of robustness, we show that this landscape represents well the mutational effects of single-residue substitutions when compared with a deep-mutational scanning experiment, and that the inclusion of epistatic couplings in the landscape model is essential for its accuracy. The landscape can thus be used as a proxy for the protein's fitness landscape.

Next, we present a minimal model of evolutionary dynamics, very similar to but more quantitative than SEEC. In this model, mutations appear at the level of the DNA sequence via single-nucleotide mutations, but selection acts exclusively at the protein level, that is, on the amino acid sequence translated from the DNA sequence, via the inferred sequence landscape. We will show that sequences generated in silico by this model reproduce quantitative features of the experimentally generated sequences, like mutational profiles or the fitness distribution.

Subsequently, we explore the potential of the experiments by performing simulations under variable conditions for sequence divergence, sequencing depth, or selection strength. This allows us to locate the two experiments in an exhaustively scanned parameter space, to understand the limitations of the experiments, and to propose schemes for overcoming current limitations.

An Epistatic Data-Driven Sequence Landscape Captures Mutational Effects

The basis of our approach is a computationally inferred sequence landscape, used as a proxy to quantify protein fitness and selection acting on proteins. To obtain this landscape, we first use the Pfam protein-family database (Mistry et al. 2021) to extract an MSA of diverged homologs of the wildtype protein used in the experiments. Both studies performed experiments with a member of the beta-lactamase family (Pfam accession PF13354), TEM-1 in (Fantini et al. 2020) and PSE-1 in (Stiffler et al. 2020); the latter work also studied the acetyltransferase AAC6 (PF00583). The details of the MSA
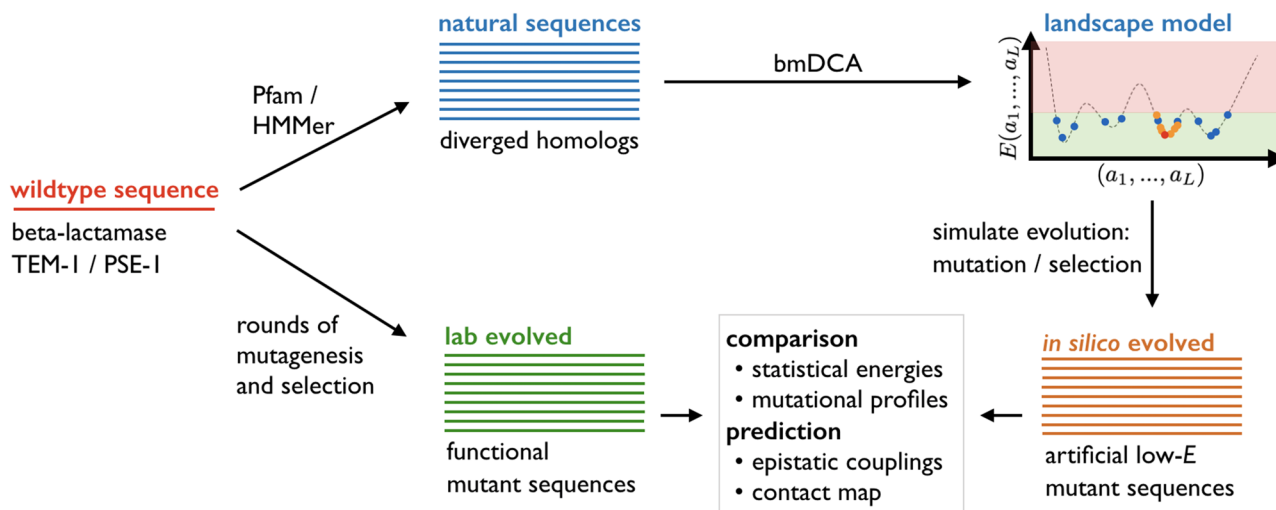
Fig. 1. Scheme of our evolutionary modeling approach: starting from a wildtype sequence (red), we collect a large multiple sequence alignment of naturally diverged homologs (blue), which are used to learn a generative landscape model using bmDCA (Figliuzzi et al. 2018). Evolution is simulated as a Markov process in this landscape, leading to simulated, or in silico evolved mutant sequences. These sequences can be compared with the results of evolution experiments (Fantini et al. 2020; Stiffler et al. 2020) (green), to assess estimated protein fitness (so-called statistical energies, compare below), mutational profiles, and DCA-based epistasis and contact prediction. The simulation scheme also allows for changing experimental control parameters like final sequence divergence, sequencing depth, and selection strength.

construction are given in Materials and Methods below; we find, for example, an MSA of 18,334 beta-lactamase sequences.

The underlying idea of our work is to represent the natural variability of this MSA via a generative statistical model $P(a_1, \ldots, a_L)$, with $(a_1, \ldots, a_L)$ representing an aligned amino acid sequence, that is, the a_i are either one of the 20 natural amino acids, or an alignment gap. Since data are limited, we need to assume some mathematical form for $P(a_1, \ldots, a_L)$. Introducing:

$$P(a_1, \ldots, a_L) = \frac{1}{Z} \exp\left\{ -E(a_1, \ldots, a_L) \right\}, \quad (1)$$

we write the "statistical energy" $E(a_1, \ldots, a_L)$, which is to be seen as a proxy for negative protein fitness (Morcos et al. 2014; Levy et al. 2017), in the form used by DCA (Weigt et al. 2009; Morcos et al. 2011; Cocco et al. 2018),

$$E(a_1, \ldots, a_L) = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j), \quad (2)$$

as a sum over position- and amino acid-specific single-residue biases, or fields, $h_i(a_i)$ and pairwise epistatic residue–residue couplings $J_{ij}(a_i, a_j)$. This model, also known as Potts model, assigns low statistical energy E to "good/fit" sequences of high probability, and high E to "bad/unfit" nonfunctional sequences of low probability. As illustrated in figure 1, we expect to find low statistical energies for both natural and experimentally evolved sequences. The strongest couplings are known to be related to residue–residue contacts in the 3D protein structure, compare with (Morcos et al. 2011).

The model parameters are inferred by the currently most accurate version of DCA, called bmDCA (Figliuzzi et al. 2018), which maximizes the model's likelihood via Boltzmann-machine learning (Ackley et al. 1985). As is known from the

literature (Sutto et al. 2015; Levy et al. 2017; Figliuzzi et al. 2018), this model is generative because sequences sampled from $P(a_1, \ldots, a_L)$ reproduce many statistical properties of the MSA of natural sequences. This does not only concern fitted quantities like one- and two-site amino acid frequencies, but also nonfitted properties like three-residue amino acid frequencies or the clustering of beta-lactamases into subfamilies in sequence space. Note that the epistatic couplings are essential for the model to be generative: a profile model having only fields $h_i(a_i)$ but no couplings $J_{ij}(a_i, a_j)$, that is, a model assuming statistical independence of all positions in the protein, is not generative in the rather strict sense discussed above (Figliuzzi et al. 2018). It misses both nontrivial second- and higher-order correlations and the clustered sequence distribution. Note also that, in a different protein family (chorismate mutase, PF01817), the same modeling approach was recently shown to artificially generate fully in vivo functional protein sequences (Russ et al. 2020).

To test the quantitative character of our landscape $E(a_1, \ldots, a_L)$, we compare the model predictions $\Delta E = E(\text{mutant}) - E(\text{wildtype})$ for the effect of mutations introduced into a wildtype sequence, with the results of a deep-mutational scan of the beta-lactamase TEM-1 (Firnberg et al. 2014). As is shown in figure 2A and B, the two are highly correlated, with a Spearman rank correlation of -0.77, compare also with (Figliuzzi et al. 2016) and (Hopf et al. 2017) and the scatter plot supplementary figure S1A, Supplementary Material online, directly comparing prediction and experiment. This correlation shows that our landscape $E(a_1, \ldots, a_L)$, even if inferred using distantly diverged TEM-1 homologs, provides quantitative information in the direct vicinity of TEM-1. As expected, low statistical energies correspond to high fitness values. To underline the importance of the epistatic couplings in
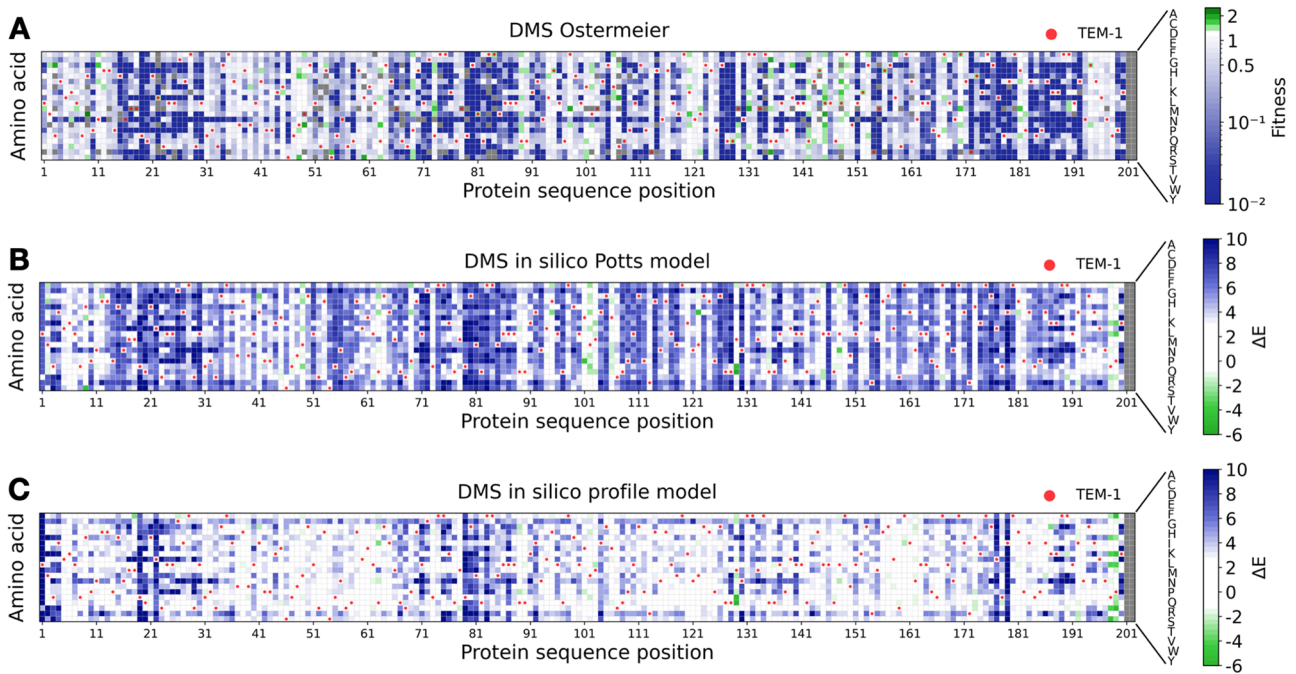
Fig. 2. Experimental and predicted mutational effects in TEM-1: panel (*A*) shows the results of the deep-mutational scanning experiment of (Firnberg et al. 2014), as compared with the computational predictions using the epistatic Potts model (*B*) and the nonepistatic profile model (*C*). Panels (*A*) and (*B*) have a Spearman rank correlation of −0.77, showing that low energies correspond to high fitness. Panels (*A*) and (*C*) have a reduced Spearman correlation of −0.6 due to the absence of epistatic couplings in the profile model.

our model, we also show in figure 2*C* and supplementary figure S1*B*, Supplementary Material online, the predictions of a nonepistatic profile model inferred from the same beta-lactamase MSA: the correlation with the experimental data decreases to −0.6, compare with (Figliuzzi et al. 2016).

This observation is central for our evolutionary model since the selection of sequences with few mutations with respect to the wildtype reference will be modeled by energy differences ΔE as introduced above.

A Model of Evolutionary Dynamics Reproduces Quantitative Features of Experimentally Evolved Sequences

Evolution (natural and experimental) can be seen as a stochastic process in a sequence landscape, with random mutations and phenotypic selection modeled by our statistical energy $E(a_1, \ldots, a_L)$. A minimal model realizing this idea is SEEC (de la Paz et al. 2020): a random site $i \in \{1, \ldots, L\}$ is selected, and an amino acid $b \in \{A, C, \ldots, Y\}$ is selected to substitute a_i with a probability proportional to $\exp\{-\Delta E(a_i \to b)\}$, with ΔE being the statistical-energy difference between the mutated and the unmutated sequences. A nonaccepted or synonymous mutation is characterized by $a_i = b$. Note that deletions and insertions are currently not considered in our model.

Although this model can be used to explore the qualitative influence of epistasis on protein sequence evolution, our analysis requires a more quantitative model taking in particular two differences into account:

- Mutations happen at the "nucleotide" level. As a consequence, not all amino acids are accessible from all amino acids via a single nucleotide mutation; and the set of accessible amino acids depends specifically on the used codon.
- The experiments allow to "vary selection strength." For TEM-1 and PSE-1, this is done by modifying the antibiotic concentration: the same mutation can be more or less strongly favored or suppressed.

To include these factors into our evolutionary model, we introduce two important modifications with respect to SEEC: first, we model evolution at the level of the nucleotide sequence $(n_{11}, n_{12}, n_{13}, \ldots, n_{i1}, n_{i2}, n_{i3}, \ldots, n_{L1}, n_{L2}, n_{L3})$ coding for the amino acid sequence $(a_1, \ldots, a_L)$, that is, the nucleotide triplet (n_{i1}, n_{i2}, n_{i3}) codes for amino acid a_i. For each possible codon $(n_1, n_2, n_3) \in \{A, C, G, T\}^3$ (with the exception of the stop codons), we introduce the set of amino acids $\mathcal{A}_{acc}(n_1, n_2, n_3) \subset \{A, \ldots, Y\}$, which are accessible from (n_1, n_2, n_3) by at most a single nucleotide mutation. Possible substitutions for a_i are now only selected from $\mathcal{A}_{acc}(n_{i1}, n_{i2}, n_{i3})$. Note that also a_i is in $\mathcal{A}_{acc}(n_{i1}, n_{i2}, n_{i3})$, accessible via its original codon and any synonymous mutation.

Second, selection strength will be regulated by a new parameter β, having the form of an inverse temperature $\beta = 1/T$ in statistical physics, which modifies the sequence probability to $P \sim \exp\{-\beta E\}$. The "low-temperature" case $\beta > 1$ ($T < 1$) corresponds to increased selection (e.g., higher antibiotic concentration, or directed evolution), in the limit $\beta \to \infty$ ($T \to 0$) only the best possible amino acid in position i

is accepted. The "high-temperature" case $\beta < 1$ ($T > 1$) corresponds to decreased selection (e.g., lower antibiotic concentration); the limit $\beta \to 0$ ($T \to \infty$) describes the case of mutation-accumulation experiments without selection.

This idea is implemented in the following three steps, which are iterated, compare with Materials and Methods for details:

(1) We randomly select a site $i \in \{1, \dots, L\}$ to be mutated, corresponding to the codon $\mathbf{n}_i = (n_{i1}, n_{i2}, n_{i3})$ and the amino acid a_i.

(2) One of the accessible amino acids $b \in \mathcal{A}_{acc}(\mathbf{n}_i)$ is selected to substitute a_i with a probability $P(b|a_1, \dots, a_{i-1}, a\ i+1, \dots, a_L) \propto \exp\{-\beta \Delta E(a_i \to b)\}$. Due to the epistatic couplings in (equation 2), this probability depends explicitly on the sequence context $(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_L)$.

(3) One out of the possible codons for amino acid b, which differs from $\mathbf{n}_i$ in at most a single nucleotide, is selected uniformly at random.

The resulting nucleotide and amino acid sequences remain thus mutually consistent.

The proposed dynamics can be efficiently implemented, and very large sequence libraries can be simulated over long times. To make these data comparable with the libraries generated by experimental evolution, we need to adapt the simulation parameters: first, the number of mutational steps in our simulation is not directly related to the number of experimental generations (because error-prone PCR may introduce multiple mutations each round); we choose it to reach the same average number of substituted amino acids in the simulated and experimental libraries. In this sense, different experimental mutation rates can be parametrized by the number of steps needed by our dynamics to reach the same number of mutations. Second, the selection strength $\beta = 1/T$ has no evident relation to the antibiotic concentration used in the experiment. We therefore tune the value of $\beta = 1/T$ such that the statistical energy $E(a_1, \dots, a_L)$ of the simulated and the experimental sequences have the same linear slope as a function of the number of substitutions. For the case of PSE-1, shown in figure 3, we find that $T = 1.4$ is a good value, compare figure 3A for the experimental data from (Stiffler et al. 2020), and figure 3B for simulated data. This corresponds to low selection strength $\beta = 1/T < 1$. Even if we adjust only average distance and slope, we find that also the overall distribution is well reproduced. Similar observations for TEM-1 and AAC6 are shown in supplementary figures S2 and S3, Supplementary Material online.

Figure 3C shows that for strong selection $T = 0.05$ ($\beta = 20$) the sequence energy decreases with the number of substitutions, corresponding to an increasing fitness as expected in a directed-evolution scenario. Weak selection, shown in figure 3D for $T = 20$ ($\beta = 0.05$), corresponds to a sharp increase in statistical energy, and thus a loss in fitness, as expected from the accumulation of predominantly deleterious random mutations.

Figures show global measures comparing experimental and simulated sequences: the Hamming distance is the number of substitutions along the entire amino acid sequence, the energy also depends on the entire sequence. To increase our confidence in the quantitative character of our evolutionary model, we compare in figure 4 the site- and amino acid-specific mutational frequencies between experimental and simulated sequence data. To this end, we extract the quantities $f_i(a)$ describing the fraction of sequences in an MSA having amino acid a in position i. Interestingly, also this refined measure of sequence diversity is very similar for simulated and experimental sequences; we observe a high correlation of 86%, compare with supplementary figure S4, Supplementary Material online. These plots highlight the importance of working only with amino acid substitutions accessible via single-nucleotide mutations: many amino acids show zero frequency in both plots due to inaccessibility. The mutational spectrum predicted without considering the accessibility of amino acids is shown in supplementary figure S4, Supplementary Material online: we see that the mutational frequencies are more homogeneously distributed, close-to-zero frequency mutations become very rare as compared with the experimental sequences. The correlation goes down to 65% between simulated and experimental data in this case.

Based on these observations, we conclude that our evolutionary model, which combines mutations at the nucleotide level with selection at the amino acid level, is able to reproduce well the statistical features of the experimental sequences. This conclusion is also confirmed, when using TEM-1 and AAC6 as initial wildtype sequences, compare with supplementary figures S5 and S6, Supplementary Material online.

In Silico Sequence-Space Exploration, and the Emergence of Epistatic Signals

Having developed a quantitative model to simulate experimental evolution, we are now able to explore evolutionary scenarios going well beyond those realized in the experiments. We can systematically analyze the influence of the sequence divergence from wildtype, of the sequenced library depth, and of the selection strength on the accuracy of coevolution-based contact prediction. Each setting of these parameters would require long experiments and would sometimes be inaccessible due to the high number of experimental rounds or the depth of the sequenced library.

Computationally this becomes straightforward although intensive: we have performed many runs of evolutionary simulations, each producing an MSA with specific parameters, simulating the possible outcome of an evolutionary experiment, as represented in figure 5. Each square in these plots corresponds to the average over five simulation runs. Depicted is the positive predictive value (PPV), which measures the fraction of true positive contact predictions within the first 100 contact predictions, compare with Materials and Methods for details. Due to the large number of contact predictions to be performed, we used GaussDCA (Baldassi et al. 2014), a very fast, even if not the most accurate contact
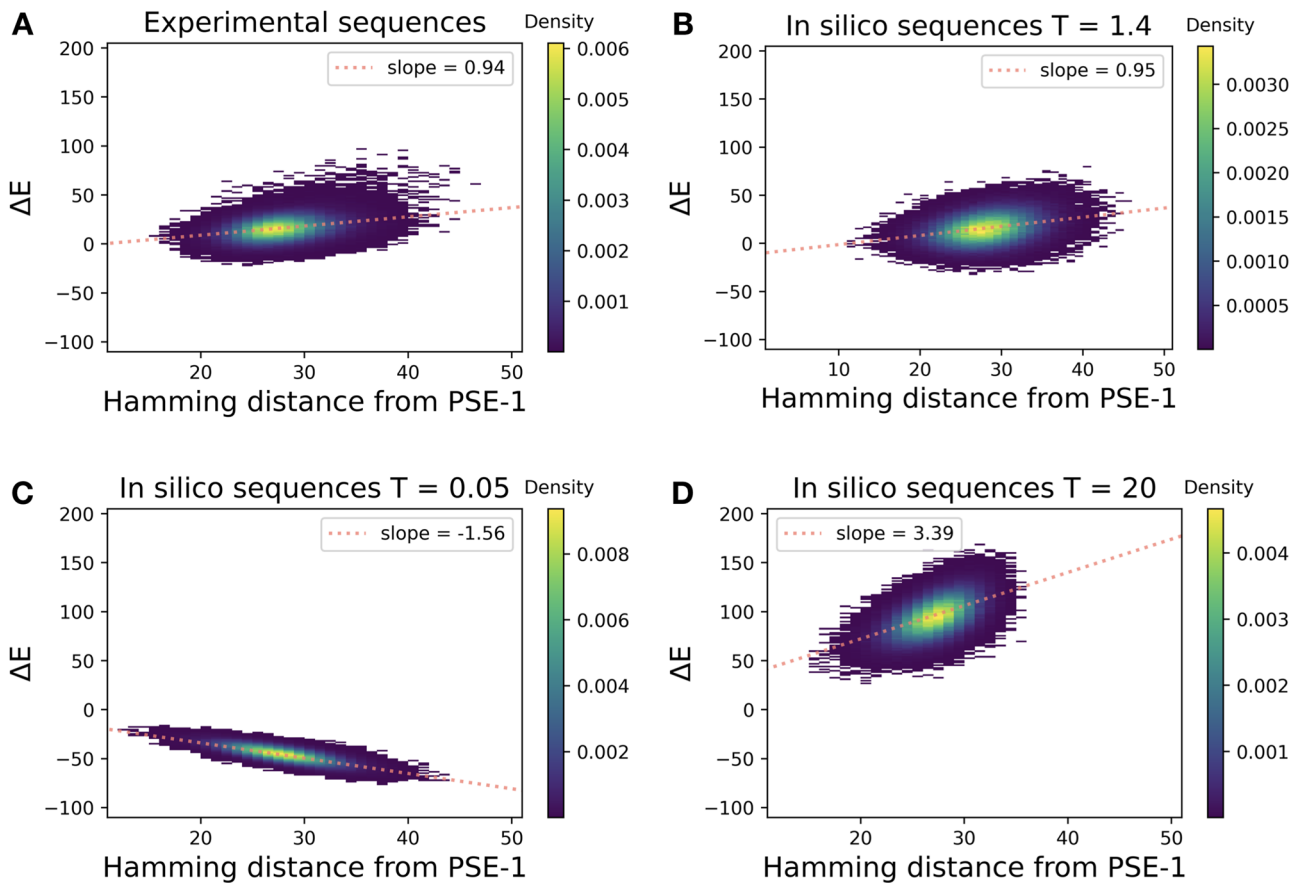
FIG. 3. Statistical energy in dependence of sequence distance from wildtype: panel (*A*) shows the statistical energies of the sequences from generation 20 in Stiffler et al., as a function of the Hamming distance (number of substituted amino acids) from the wildtype PSE-1. Panel (*B*) shows the same quantities for the in silico simulated sequences, where selection strength *T* and the number of simulated evolutionary steps are adjusted to reproduce the average distance and the slope from panel (*A*). Panel (*C*) shows an example of strong selection ($T \ll 1$) leading to optimized sequences having lower statistical energies/higher fitness. Panel (*D*) shows the case of very weak selection ($T \gg 1$) resulting in random, mostly deleterious substitutions strongly increasing statistical energy.

predictor. Figure 5A shows the plot for the selection strength used in the experiments for PSE-1. The red zone corresponds to inaccurate contact predictions, being sometimes hardly better than random (PPV $\sim$ 0.13). It is found consistently for small sequence libraries, and for sequence libraries of low divergence from wildtype. It becomes evident that we need to go to a sufficient number of simultaneous mutations to be able to detect at least a weak epistatic signal between mutations, which can be used for contact prediction. However, this signal remains weak: we need much larger sequence libraries of at least about 50,000 sequences to reach a reasonable contact prediction. However, even for the largest and most diverged library we have studied, a PPV of only 0.7–0.8 is reached, which remains below the contact prediction reached by using the MSA of natural homologs, which was used before for the inference of our sequence landscape. The latter reaches a PPV of 0.98 using GaussDCA. Figure 5B shows the same observables for experiments starting with the TEM-1 sequence, the overall results are very similar to PSE-1, even if some quantitative details depend on the initial wildtype sequence.

It might be speculated that better contact-prediction algorithms may shift the region of nontrivial predictions down to lower Hamming distances from wildtype, or to lower sequence numbers. Although the computational cost of plmDCA is too high to reproduce the full analysis of figure 5, we have reanalyzed two columns at average Hamming distance 41 and 65. As is shown in supplementary figure S7, Supplementary Material online, for low sequence numbers GaussDCA and plmDCA give very similar low prediction accuracies, whereas the improved accuracy of plmDCA over GaussDCA becomes visible only at sufficiently high sequence numbers. At the resolution of our analysis, no shift in the boundary is observable.

The conditions of the experiments for PSE-1 and TEM-1 are highlighted, in the two panels of figure 5. For PSE-1, 20 rounds of evolution led to an average sequence distance of 27 amino acid substitutions from wildtype, and a sequenced library of 165,000 distinct sequences (Stiffler et al. 2020). Interestingly, this point is located slightly beyond the boundary of emergence of coevolutionary signal. The predicted average PPV of 0.58 is comparable with the 0.65 obtained using
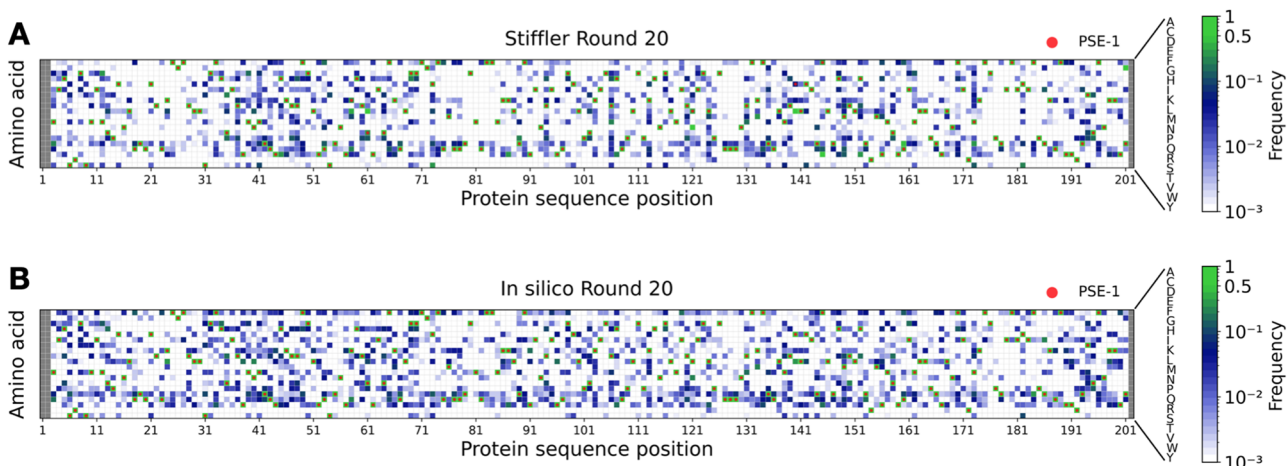
FIG. 4. Position-specific amino acid frequencies for experimental and simulated sequence libraries: panel (*A*) shows the frequencies $f_i(a)$ of usage of amino acid *a* in site *i* in round 20 of experimental PSE-1 evolution, panel (*B*) shows the same quantity for simulated evolution. The Spearman rank correlation between the two frequency spectra is 86%.
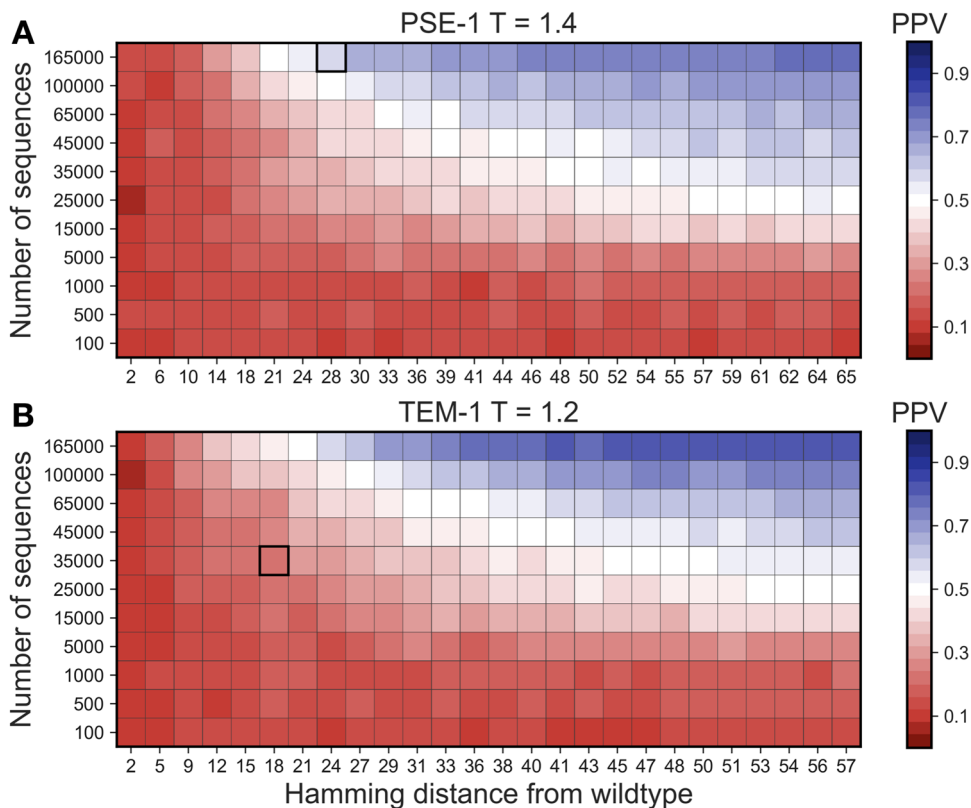

FIG. 5. Accuracy of contact prediction as a function of sequence number and sequence divergence: panel (*A*) shows the accuracy of contact prediction as a function of the average sequence divergence from wildtype PSE-1 and the depth of the sequenced library. The accuracy is measured via the PPV, that is, the fraction of true positive contact predictions in the first 100 DCA-predicted contacts, compare with Materials and Methods for details. The selection strength $T = 1.4$ corresponds to the experimental condition in (Stiffler et al. 2020). The highlighted square indicates an average Hamming distance of about 27 and a sequence library of 165,000, as realized in (Stiffler et al. 2020). Panel (*B*) shows the same quantities for wildtype TEM-1, and for the experimental conditions used in (Fantini et al. 2020).

the experimental MSA, compare with Materials and Methods section.

This is in contrast to the TEM-1 experiment of (Fantini et al. 2020), compare with figure 5*B*: the experiment was performed for fewer rounds, leading to less divergence

from TEM-1, and the sequence library was less deeply sequenced. The resulting library, with an average Hamming distance of 18 from TEM-1 and with 34,431 unique sequences, is located slightly below the line of emergence of coevolution signal. This observation provides
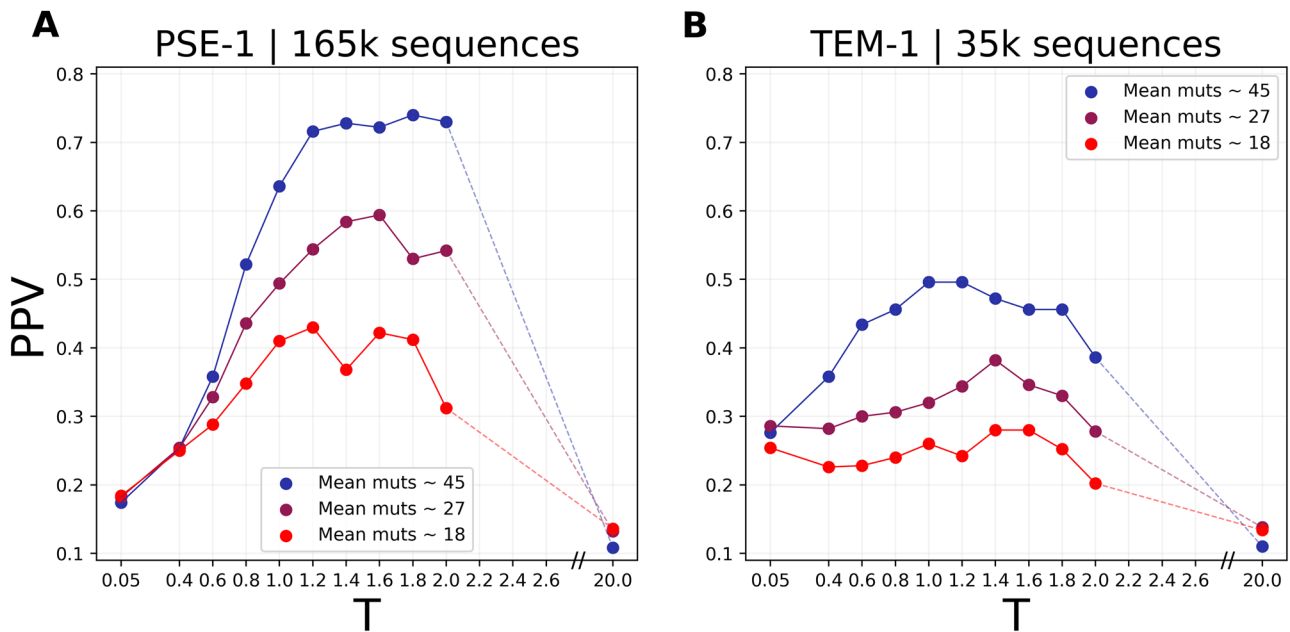
Fig. 6. Dependence of the contact-prediction accuracy on selection strength: we show the PPV (100 predicted contacts) of simulated MSAs at variable selection strength T (panel A for PSE-1, panel B for TEM-1), and for different sequence distances from the wildtype protein. We predict that, for the distances observed in the evolution experiments (27 for PSE-1, 18 for TEM-1), both experiments would have benefited from slightly lower antibiotic concentrations.

a potential explanation for the observed reduced performance in contact prediction.

The AAC6 results show that reduced sequence divergence can, at least partly, be compensated by a strong increase in the number of sequences in the evolved MSA, compare with supplementary figure S8, Supplementary Material online, which confirms original findings of (Stiffler et al. 2020). Even if having only an average Hamming distance of about 8 substitutions, the large library of more than 10^6 sequences allows for the detection of a weak contact-related signal.

The results depend substantially on the strength of selection. Supplementary figure S9, Supplementary Material online, shows the extreme cases of very strong and very weak selection discussed before. Both show inaccurate prediction. An important difference becomes visible when looking at the horizontal axes: all use the same number of simulated evolutionary steps. In the case of strong selection, sequences stay closer to the wildtype, since most mutations are deleterious and selected against, and they stay close to each other. So while being all functional, they do not accumulate sufficient sequence variability to provide a reliable epistatic signal. In the case of extremely weak selection, almost all mutations are acceptable. Sequences are found to diverge strongly from the initial PSE-1 sequence, but the absence of selection causes also an absence of coevolution.

Discussion

The aim of this work was to showcase the potential of evolutionary models in data-driven sequence landscapes. Recent progress in landscape modeling has led to advances in using sequence alignment to predict protein structure, mutational effects, and even to design non-natural but biologically functional sequences. Here we show that, equipped with a simple stochastic dynamics capturing the interplay between mutation and selection, these landscapes lead to models which are able to describe in a quantitatively accurate way the results of evolution experiments. This is not only restricted to proteins, as studied in this work, but similar evolution experiments have been performed for RNA (Zhou et al. 2018) and could therefore be analyzed in an analogous way starting from sequence landscapes for RNA families (Kalvari et al. 2021).

The applications for experimental evolution are evident: we can use our modeling to optimize experimental evolution protocols, for example, when we search for fully functional sequences but at some minimum number of mutations from a starting sequence, or when we want to explore sequence space optimally for contact prediction. In this case, we could, for example, optimize the selection strength. In the case of the beta-lactamases studied in this article, figure 6 shows that a slightly lower selection pressure (i.e., higher selection temperature) would have led to even better contact predictions. However, this potential increase is weak as compared with the one reachable by more diverged sequences.

A possible obstacle in such applications is the fact that the selection temperature T, which we use to model selective pressure, has to be fitted from experimental data via the slope of the statistical energies of the evolved sequences vs. their distance from wildtype. To understand the minimal sequence requirements for reaching robust and accurate slope estimates, we have subsampled the experimental sequence libraries of PSE-1 for rounds 10 and 20. As is shown in supplementary figure S10, Supplementary Material online, we observe: 1) that the slope

can be estimated accurately already from about 200–300 sequences, whereas the estimation error becomes large when using less than 100 sequences, and 2) that the estimates are almost equal for round 10 and round 20. We conclude that the selection temperature T can be reliably determined with moderate experimental effort (low number of sequences, few experimental rounds). Once estimated, the parameters can be used in simulations, which may guide more massive experiments evolving large sequence libraries over many rounds.

We see our current model as a starting point for more detailed evolutionary models. There is space for a substantial gain in accuracy: we can introduce biases in the mutations introduced by error-prone PCR directly into the model (Moore and Maranas 2000; Pritchard et al. 2005), the latter can be derived from data by analyzing synonymous mutations. Furthermore, we can introduce codon bias, the difference between transitions and transversions, the fact that error-prone PCR may introduce simultaneously several mutations before selection, or the emergence of phylogeny in cycles of mutation and selection.

The modeling can also benefit from experimental feedback. If sequence libraries would also be sequenced before and after the selection step, we could establish a better correspondence between statistical energies and selection, up to a gauge of statistical energies vs. antibiotic concentrations.

However, the potential of such evolutionary models in data-driven landscapes goes far beyond the application to experimental evolution. As is shown by SEEC (de la Paz et al. 2020), already the simplest nontrivial evolutionary model allows for illuminating important consequences of epistasis in evolution, like the site- and time-dependence of substitution rates. We anticipate that the proposed modeling framework may capture many of these effects in a highly quantitative way. The relatively simple modeling framework proposed in our paper might also be a starting point for more theoretical–mathematical analyses about, for example, the emergence of epistatic signals in sequence libraries. In this context, it might also be interesting to see in how far more distributed signatures of epistatic signal, possibly related to protein function rather than contacts, become visible in experimentally evolved sequence libraries, compare with (Rivoire et al. 2016), (Shimagaki and Weigt 2019), and (Tubiana et al. 2019).

Materials and Methods

Sequence Data
Sequences from Experimental Evolution
We include in our analysis the sequence data coming from the experiments of in vitro evolution by (Fantini et al. 2020) on TEM-1 and by (Stiffler et al. 2020) on PSE-1 and AAC6.

The aligned amino acid sequences from (Fantini et al. 2020) were kindly provided by the authors prior to publication, and can also be found at http://laboratoriobiologia.sns.it/supplementary-mbe-2019/ (last accessed November 17, 2021). The raw sequencing reads are available at the National Centre for Biotechnology Information Sequence Read Archive (SRA) with accession code PRJNA528665 (http://www.ncbi.nlm.nih.gov/sra/PRJNA528665, last accessed November 17, 2021). Amino acid sequences with more than six gaps were discarded as a quality control to remove sequences with lower quality.

Stiffler et al. (2020) ran two experiments using the PSE-1 beta-lactamase and the AAC6 acetyltransferase as starting wildtypes. Aligned sequencing reads from the last round of the two experiments (translated into amino acid sequences) can be found at https://github.com/sanderlab/3Dseq (last accessed November 17, 2021). The raw sequencing reads are available at the National Centre for Biotechnology Information Sequence Read Archive (SRA) with accession code PRJNA578762 (http://www.ncbi.nlm.nih.gov/sra/PRJNA578762, last accessed November 17, 2021).

Our models are built for the Pfam-annotated positions using the corresponding Pfam domains PF13354 (Beta-lactamase2) and PF00583 (Acetyltransf1). We realigned the wildtype sequence using the hmmalign command from the HMMer software suite (Eddy 2011) and profile Hidden Markov Models downloaded from Pfam (Mistry et al. 2021). We then removed from the experimental MSAs all columns corresponding to nonmatched states of the wildtype sequence.

The resulting MSAs of experimentally evolved sequences have 202 sites and 165,855 sequences for PSE-1 (round 20), and 34,431 sequences for TEM-1 (generation 12). For AAC6, we find 117 sites and 1,260,048 sequences (round 8).

Natural Homologous Sequences and Preprocessing of the Training Set
The MSAs of natural homologous sequences of the two considered protein families PF13354 (Beta-lactamase2) and PF00583 (Acetyltransf1) were generated running the hmmsearch command from the HMMer software suite (Eddy 2011) on the UniProt database (The UniProt Consortium 2021). Insertions were removed, and sequences with more than 10% gaps and duplicated sequences were excluded to improve the quality of the alignment. Any sequence closer than 80% to the wildtypes TEM-1, PSE-1, or AAC6 was excluded from the alignments to avoid the introduction of biases toward these sequences in the bmDCA learning. The resulting MSAs included 18,333 (43,576) homologous and nonidentical aligned sequences of length 202 (117) for PF13354 (PF00583).

Note that some residues, which are present in the N- and C-terminal regions of the experimental sequences, are not covered by the Pfam domains, and therefore excluded from our analyses. Extending the MSA beyond the borders of the Pfam domains would lead to the inclusion of evolutionarily less conserved positions, and thus to the inclusion of highly gapped columns into the MSA of natural data. Such columns have been previously found to compromise the accuracy of DCA landscapes (Figliuzzi et al. 2016) and are therefore left out in this study.

The natural MSA were used to train two Potts models using bmDCA (Figliuzzi et al. 2018) in the implementation of Barrat-Charlaix et al. (2021), which provides the currently most accurate DCA models.

Evolutionary Model

As already discussed in Results section, our evolutionary model combines mutations at the nucleotide level with selection at the level of aligned amino acid sequences. We therefore need to specify both the nucleotide sequence $\mathbf{n} = (n_{11}, n_{12}, n_{13}, \ldots, n_{i1}, n_{i2}, n_{i3}, \ldots, n_{L1}, n_{L2}, n_{L3})$ and the resulting amino acid sequence $\mathbf{a} = (a_1, \ldots, a_L)$, which is translated from $\mathbf{n}$ using the standard genetic code. Since we consider full-length aligned sequences of Pfam domains, stop codons are not allowed in $\mathbf{n}$. Furthermore, we have to accommodate alignment gaps possibly existing in $\mathbf{a}$: a gap in $\mathbf{a}$ is represented by a triplet of gaps in $\mathbf{n}$. Gaps are not changed during our simulations, our model does consider only single-nucleotide substitutions, but no insertions and no deletions. Note that the gray columns in figure 4 and supplementary figures S5 and S6, Supplementary Material online, correspond to gaps in the wildtype sequence, which are conserved both in the experiment and in the model.

As mentioned before, for each codon $(n_1, n_2, n_3) \in \{A, C, G, T\}^3$, we consider the set of amino acids $\mathcal{A}_{acc}(n_1, n_2, n_3) \subset \{A, \ldots, Y\}$, which are accessible from (n_1, n_2, n_3) by at most a single nucleotide mutation.

Our simulation of sequence evolution proceeds by iterating the following three steps defining a Markov chain (MC) in the space of nucleotide sequences (note that, due to the degeneracy of the genetic code, the process is "not" an MC in amino acid sequence space):

(1) A position $i \in \{1, \ldots, L\}$ is chosen uniformly at random along the amino acid sequence, corresponding to the codon $\mathbf{n}_i = (n_{i1}, n_{i2}, n_{i3})$ and the amino acid a_i. Although $a_i = "-"$, that is, a gap is chosen, we repeat the selection of the position i.

(2) Out of all accessible amino acids $b \in \mathcal{A}_{acc}(\mathbf{n}_i)$, we selected one using the conditional probability $P_\beta(b|\mathbf{a}_{-i})$, which couples the amino acid b explicitly to the sequence context $\mathbf{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_L)$:

$$P_\beta(b|\mathbf{a}_{-i}) = \frac{\exp\left\{\beta h_i(b) + \beta \sum_{j(\neq i)} J_{ij}(b, a_j)\right\}}{z_i(\mathbf{a}_{-i})}, \quad (3)$$

with

$$z_i(\mathbf{a}_{-i}) = \sum_{b \in \mathcal{A}_{acc}(\mathbf{n}_i)} \exp\left\{\beta h_i(b) + \beta \sum_{j(\neq i)} J_{ij}(b, a_j)\right\} \quad (4)$$

being a normalization constant. In difference to Z in equation (1), it can be calculated efficiently by summing over the less than 20 accessible amino acids.

(3) One out of the possible codons for amino acid b, which differs from $\mathbf{n}_i$ in at most a single nucleotide, is selected uniformly at random.

The new amino acid b substitutes a_i in $\mathbf{a}$, and the new codon $\mathbf{n}_i$ in $\mathbf{n}$. We thereby conserve the coherence between nucleotide and amino acid sequence.

To simulate an entire MSA of M sequences, the process is initiated M times in the wildtype reference sequence, and M independent runs of the MC are performed. The number of steps in these MCs is chosen such that the average Hamming distance of the generated amino acid sequences reaches a target number. Note that the Hamming distances may vary from MC to MC, since $\mathcal{A}_{acc}(\mathbf{n}_i)$ contains the case $b=a_i$ accessible via any synonymous mutations. The Hamming distance can therefore assume any value between zero and the number of performed mutational steps.

Simulated Sequence Data for Contact Prediction

Our evolutionary algorithm has three input parameters adding to the wildtype sequence and the statistical-energy model: the number of sequences M, the number N_{MC} of steps of our evolutionary MC model, and the selection temperature T. Given this triplet of numbers it outputs an MSA obtained simulating evolution for N_{MC} iterations starting from the wildtype sequence, repeating the sampling independently M times at temperature $T = 1/\beta$.

For each wildtype sequence, we simulated the outcome of different protein evolution experiments by scanning these three input parameters within a range of interest. For MSA generated starting from TEM-1 or PSE-1 (AAC6), we varied M in the range $100 - 165,000$ ($500 - 1,250,000$), N_{MC} in the range 5–255 (4–120), and T in the range 0.05–20.

To save resources and time, given the computational cost of sampling, we opted for a scheme that would allow us to reduce the number of independent MC chains needed to simulate evolution. For each temperature T, we run 165,000 (1,250,000) independent MCs for TEM-1 and PSE-1 (AAC6) and printed MSAs at the desired number of MC steps until 255 (120) MC steps. The MSAs with less sequences were obtained by randomly subsampling without replacement from the MSA with 165,000 (1,250,000) sequences. To produce more statistics, we ran the same simulations five times.

Contact Prediction

Contact prediction was performed using GaussDCA (Baldassi et al. 2014) for all MSAs, included, for coherence, the experimental ones. GaussDCA is the computationally most efficient implementation of DCA. Its accuracy of contact prediction is slightly inferior to plmDCA or bmDCA. However, we use it: in our analysis, we had to predict contacts for a large number of partially deep simulated MSAs (cf., fig. 5) to explore multiple combinations of sampling time, sample size, and selection strengths.

The reweighting parameter was set to 0 for contact prediction of in silico MSAs, as this reduces computational time and is coherent with the independence of the simulated MCs. On the other hand, contact prediction of experimental MSAs

was performed using the default option ":auto" of GaussDCA for reweighting. These different treatments of simulated and experimental sequences are based on the fact, that simulations generate statistically independent sequences (conditioned to wildtype initialization), whereas the experiments may generate sequence ensembles having nontrivial phylogenetic effects. The pseudocount was set to 0.6 (0.5) for PSE-1 and TEM-1 (AAC6) empirically, as we found it to be a good intermediate value for MSAs with very different statistics.

Intrachain atomic distances for both families were obtained by running the single-protein mode of the code provided by Pfam Interactions (https://doi.org/10.5281/zenodo.4080947, last accessed November 17, 2021), we used the shortest distance between heavy atoms of the two amino acids among all structures of the Protein Data Bank (PDB) (Burley et al. 2021) listed in Pfam. Following standards in coevolutionary contact prediction, all pairs with distance below 8 Å and a minimum separation of 5 positions along the sequence are kept as contacts for the calculation of the PPV. For AAC6, we used a more stringent cutoff of 5.5 Å, since the structural variability across the protein family is already well represented in the PDB.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Data Availability

The sequence data underlying this article are available on Github at https://github.com/matteobisardi/DataSeqEvol. The datasets were derived from sources in the public domain (see Materials and Methods). The software implementation for the evolutionary simulations, SeqEvol, is available on Github at https://github.com/matteobisardi/SeqEvol, together with the instructions to reproduce the results of this work.

References

Ackley DH, Hinton GE, Sejnowski TJ. 1985. A learning algorithm for Boltzmann machines. *Cogn Sci*. 9(1):147–169.
Arnold FH. 1998. Design by directed evolution. *Acc Chem Res*. 31(1):125–131.

Arnold FH. 2018. Directed evolution: bringing new chemistry to life. *Angew Chem Int Ed Engl*. 57(16):4143–4148.
Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. 2011. Learning generative models for protein fold families. *Proteins* 79(4):1061–1078.
Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. 2014. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* 9(3):e92721.
Barrat-Charlaix P, Muntoni AP, Shimagaki K, Weigt M, Zamponi F. 2021. Sparse generative modeling via parameter reduction of Boltzmann machines: application to protein-sequence families. *Phys Rev E*. 104(2):024407.
Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, et al. 2021. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*. 49(D1):D437–D451.
Cadwell RC, Joyce GF. 1992. Randomization of genes by PCR mutagenesis. *PCR Methods Appl*. 2(1):28–33.
Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. 2018. Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys*. 81(3):032601.
De Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein coevolution. *Nat Rev Genet*. 14(4):249–261.
de la Paz JA, Nartey CM, Yuvaraj M, Morcos F. 2020. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc Natl Acad Sci U S A*. 117(11):5873–5882.
Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, United Kingdom: Cambridge University Press.
Eddy SR. 2011. Accelerated profile hmm searches. *PLoS Comput Biol*. 7(10):e1002195.
Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys*. 87(1):012707.
Fantini M, Lisi S, De Los Rios P, Cattaneo A, Pastore A. 2020. Protein structural information and evolutionary landscape by in vitro evolution. *Mol Biol Evol*. 37(4):1179–1192.
Figliuzzi M, Barrat-Charlaix P, Weigt M. 2018. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol Biol Evol*. 35(4):1018–1027.
Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol*. 33(1):268–280.
Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol*. 31(6):1581–1592.
Greener JG, Kandathil SM, Jones DT. 2019. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun*. 10(1):1–13.
Haldane A, Levy RM. 2019. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Phys Rev E*. 99(3):032405.
Hopf TA, Green AG, Schubert B, Mersmann S, Schärfe CP, Ingraham JB, Toth-Petroczy A, Brock K, Riesselman AJ, Palmedo P, et al. 2019. The evcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 35(9):1582–1584.
Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. 2017. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. 35(2):128–135.
Jones DT, Buchan DW, Cozzetto D, Pontil M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021.

Highly accurate protein structure prediction with alphafold. *Nature* 596(7873):583–589.

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res*. 49(D1):D192–D200.

Levy RM, Haldane A, Flynn WF. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr Opin Struct Biol*. 43:55–62.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 49(D1):D412–D419.

Moore GL, Maranas CD. 2000. Modeling DNA mutation and recombination for directed evolution experiments. *J Theor Biol*. 205(3):483–503.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*. 108(49):E1293–E1301.

Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc Natl Acad Sci U S A*. 111(34):12408–12413.

Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* 355(6322):294–298.

Pritchard L, Corne D, Kell D, Rowland J, Winson M. 2005. A general model of error-prone PCR. *J Theor Biol*. 234(4):497–509.

Rivoire O, Reynolds KA, Ranganathan R. 2016. Evolution-based functional decomposition of proteins. *PLoS Comput Biol*. 12(6):e1004817.

Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, Hilvert D, Monasson R, Cocco S, Weigt M, et al. 2020. An evolution-based model for designing chorismate mutase enzymes. *Science* 369(6502):440–445.

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706–710.

Shimagaki K, Weigt M. 2019. Selection of sequence motifs and generative Hopfield-Potts models for protein families. *Phys Rev E*. 100(3):032128.

Stiffler MA, Poelwijk FJ, Brock KP, Stein RR, Riesselman A, Teyra J, Sidhu SS, Marks DS, Gauthier NP, Sander C. 2020. Protein structure from experimental evolution. *Cell Syst*. 10(1):15–24.

Sutto L, Marsili S, Valencia A, Gervasio FL. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci U S A*. 112(44):13567–13572.

The UniProt Consortium. 2021. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 49(D1): D480–D489.

Tubiana J, Cocco S, Monasson R. 2019. Learning protein constitutive motifs from sequence data. *Elife* 8:e39397.

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596(7873):590–596.

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A*. 106(1):67–72.

Xu J. 2019. Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci U S A*. 116(34):16856–16865.

Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A*. 117(3):1496–1503.

Zhou Q, Kunder N, De la Paz JA, Lasley AE, Bhat VD, Morcos F, Campbell ZT. 2018. Global pairwise RNA interaction landscapes reveal core features of protein recognition. *Nat Commun*. 9(1):1–10.