Unsupervised Inference of Protein Fitness Landscape from Deep Mutational Scan

Jorge Fernandez-de-Cossio-Diaz,[1,2] Guido Uguzzoni,*[3] and Andrea Pagnani[3,4,5]

[1]Systems Biology Department, Center of Molecular Immunology, Havana, Cuba

[2]Laboratory of Physics of the Ecole Normale Superieure, CNRS UMR 8023 & PSL Research, Paris, France

[3]Politecnico di Torino, Torino, Italy

[4]Italian Institute for Genomic Medicine, IRCCS Candiolo, Candiolo, TO, Italy

[5]INFN, Sezione di Torino, Torino, Italy

*Corresponding author: E-mail: guido.uguzzoni@gmail.com.

Associate editor: Rasmus Nielsen

Abstract

The recent technological advances underlying the screening of large combinatorial libraries in high-throughput mutational scans deepen our understanding of adaptive protein evolution and boost its applications in protein design. Nevertheless, the large number of possible genotypes requires suitable computational methods for data analysis, the prediction of mutational effects, and the generation of optimized sequences. We describe a computational method that, trained on sequencing samples from multiple rounds of a screening experiment, provides a model of the genotype–fitness relationship. We tested the method on five large-scale mutational scans, yielding accurate predictions of the mutational effects on fitness. The inferred fitness landscape is robust to experimental and sampling noise and exhibits high generalization power in terms of broader sequence space exploration and higher fitness variant predictions. We investigate the role of epistasis and show that the inferred model provides structural information about the 3D contacts in the molecular fold.

Key words: **computational biology, statistical modeling, fitness landscape, deep mutational scanning, direct-coupling analysis.**

Article

Significance

The continuous interplay between selection and variation is at the basis of Darwinian evolution. Recent advances in experimental techniques allow for the construction of combinatorial libraries of proteins or other biomolecules, whereas high-throughput sequencing technologies are used to characterize their phenotypes. This research line provides a stringent testing grounds for studying genotype–phenotype evolutionary relation under externally controlled selective pressure. Such methods are routinely used to select molecules (e.g., monoclonal and enzymes) with specific properties. Here, we develop a data driven maximum likelihood to model the genotype–phenotype association derived from experiments. The inferred fitness landscape is robust to both experimental and sampling noise and exhibits high generalization power in terms of broader sequence space exploration and higher fitness variant predictions.

Introduction

The continuous interplay between selection and variation is at the basis of Darwinian evolution. Recent advances in experimental techniques allow for a quantitative assessment of evolutionary trajectories at the molecular level (Magurran 2013). From this point of view, the improvement in the construction of combinatorial libraries of proteins or other biomolecules and the high-throughput technologies to characterize their phenotypes (Fowler and Fields 2014; Kemble et al. 2019) provides one of the more stringent testing grounds for studying the genotype–phenotype evolutionary relation under an externally controlled selective pressure. Besides its evident theoretical appeal, this line of research also has a more practical interest: In Directed Evolution experiments, combinatorial libraries of sequences are routinely screened to select molecules with specific biochemical properties such as binding affinity toward a target (e.g., antibodies) (Winter et al. 1994) and catalytic features (e.g., enzymes) (Romero and Arnold 2009; Reetz 2013; Sadler et al. 2018).

The last decades have seen a tremendous boost in the availability of reliable high-throughput selection systems, such as genetic (Tizei et al. 2016), display systems (e.g., phage, SNAP-tag, and mRNA) (Molina-Espeja et al. 2016), cytofluorimetry (e.g., FACS) (Yang and Withers 2009), and microdroplet techniques (Aharoni et al. 2005). Still, a fundamental limitation is the number of variants that can be screened compared with the size of the sequence space of possible mutants. For example, a hundred residue protein has up to

$20^{100} \simeq 10^{130}$ possible variants, whereas the actual massive parallel assay libraries are typically able to handle ranges of variability within 10^{8}–10^{12}, which can be fed to a single high-throughput screening pass.

Thanks to advances in sequencing technologies (especially in terms of reduced cost per read), machine learning methods for the inference of sequence–phenotype associations are starting to show their full potential. In particular, several massively parallel assays, known as deep mutational scanning (DMS) (Fowler and Fields 2014; Kemble et al. 2019; Kinney and McCandlish 2019), are becoming available where typically a large-scale library of protein variants undergo repetitive cycles of selection for a function or an activity. The library is retrieved each round and the counts of each variant are determined by high-throughput sequencing. Such an increasing amount of sequence data demands new algorithms to produce accurate statistical models of genotype–fitness associations.

All computational methods developed so far that make use of DMS sequencing data to learn a genotype–fitness map utilize a supervised approach: A proxy of the fitness of the mutants tested in the experiment is computed from the sequencing reads and a machine learning method solves the regression problem (Rubin et al. 2017; Cadet et al. 2018; Otwinowski et al. 2018; Rollins et al. 2019; Schmiedel and Lehner 2019; Wu et al. 2019; Fantini et al. 2020) (with the only remarkable exception of Otwinowski [2018] that we discuss later). Here, we propose a novel method that shifts the learning approach from a supervised to an unsupervised framework, in the sense that we do not require any biophysical measurements of the molecule variants to train our model (although such data can be incorporated [Barrat-Charlaix et al. 2016]), and use only the abundances of sequences in different selection rounds. We developed a model that describes accurately (after training by likelihood maximization) the fitness landscape of a protein (or other biological sequences) in the context of a selection experiment. This strategy has the advantage of exploiting all the information in the screening experimental data (the time series of sequencing reads), in contrast to alternative methods of analysis which discard sequences affected by sampling noise (Rubin et al. 2017) and cannot predict the fitness of novel unobserved sequences.

The method consists of a probabilistic modeling of the three phases of each experiment cycle: 1) selection, 2) amplification, and 3) sequencing (see Materials and Methods). In brief, what we observe are the reads coming from the sequencing, that is, a sample of the library at a specific time step. The other phases are described in terms of latent variables referring to the number of amplified and selected mutants. The probability that a mutant is selected (e.g., by physical binding to the target) depends on the specific mutant sequence composition. On the other hand, we assume that the probability of a mutant to be amplified depends only on the fraction of mutants present after selection (ignoring possible sources of amplification selection such as codon bias that could however be taken into account in our framework using appropriate priors). We take into account both additive contributions from the individual residues and epistatic contributions in the form of pairwise interactions, although more complex multiresidue interaction schemes could be introduced.

This probabilistic description allows us to define an overall likelihood to observe a time series of reads in an experiment given the parameters involved in the *energetic* contribution to the selection, that is, the genotype–phenotype map. Optimizing the parameters to maximize the likelihood allows us to obtain an effective model of the fitness landscape.

The method has the 2-fold aim of 1) providing an accurate statistical description of the time series (in terms of panning rounds) evolution of the differential composition of the combinatorial library and 2) predicting individual sequences, or rationally designed libraries of increased biophysical activity toward the sought target, that in particular, can be used in the recently proposed machine-learning-guided directed evolution for protein engineering (Wu et al. 2019).

Our approach gets inspiration from the direct-coupling analysis (DCA) methods developed to describe statistical coevolutionary patterns of homologous sequences (Morcos et al. 2011). This successful field has provided fundamental tools commonly applied in structure prediction pipelines (Hopf et al. 2019) and more recently to provide mutational effect predictions (Mann et al. 2014; Asti et al. 2016; Figliuzzi et al. 2016, 2018; Hopf et al. 2017; Louie et al. 2018). Other approaches apply different machine learning schemes (Riesselman et al. 2018) on the same framework. The main difference between these unsupervised methods and the present work lies in the input data. The DCA approach learns a statistical description of a multiple sequence alignment of the protein family sequences, treating it as if it were an equilibrium sample drawn from a Potts model, whereas we deal with an out-of-equilibrium time series of screening experiments reads. Moreover, the broadness of the sequence space covered by the input data is different, with a protein family typically reaching an average Hamming distances of around 70% which results from the outcome of millions of years of Darwinian evolution, whereas in the DMS typical combinatorial libraries have at most 4–5 mutations away of the wild types in a few rounds of selection.

This opens several questions on the relevance of the modeling. Notably on the role of epistatic interactions (Miton and Tokuriki 2016; Starr and Thornton 2016; Cadet et al. 2018; Sun et al. 2019) and the extent of applicability of the method as a generative model. The application of DCA methods has provided evidence on the importance of epistatic effects in shaping the homologs distribution over phylogenetic sequence space and recent works have shown their role also for local protein fitness landscapes (Miton and Tokuriki 2016; Starr and Thornton 2016; Cadet et al. 2018; Kemble et al. 2019; Kinney and McCandlish 2019; Sun et al. 2019). Moreover, on the modeling side, it has shown the effectiveness of pairwise models (Socolich et al. 2005; Schneidman et al. 2006) to capture the epistatic contribution and provided useful 3D structural predictions (residue contacts) (Rollins et al. 2019; Schmiedel and Lehner 2019).

Table 1. Different Deep Mutational Scanning Data Sets Used in the Article to Evaluate the Performance of the Model.

Reference	Protein	Target	Length Mutated Part	Selection Rounds	Sequenced Time Points	Unique Mutants	Average Distance from Wild Type	Mutant Coverage
Olson et al. (2014)	GB1	IgG-Fc	55	1	2	536,833	2	588
Wu et al. (2016)	GB1	IgG-Fc	4	2	2	157,161	3.8	422
Fowler et al. (2010)	WW domain	Peptide	25	6	3	572,076	3.4	14
Araya et al. (2012)	WW domain	Peptide	34	3	4	940,730	4.3	11
Boyer et al. (2016)	Ab IgH	PVP, DNA	4	3	3	28,195	3.5	4

In the Results section, we investigate whether the same applies to the output of the experimental assay, first and foremost the reliability of the inferred epistatic interactions and the effectiveness to predict the selectivity of unobserved sequences for the given experimental conditions. Our findings are corroborated by the capacity to predict structural properties (e.g., residue contacts) from the inferred epistatic interaction, as similarly found in Rollins et al. (2019) and Schmiedel and Lehner (2019).

Results

First, we assess the accuracy of the inference method to learn the genotype to fitness map by testing on five DMS data sets, briefly described in the Data Sets section. Also, we investigate the generalization power of the inferred fitness landscapes and the promising potential to generate sequences of high fitness. Finally, we examine the epistatic interactions learned, comparing them to nonspecific epistasis due to a global nonlinear genotype to fitness map (Otwinowski et al. 2018) and analyzing the relevance of the epistatic terms to predict contacts between residues in the 3D molecular structure.

In a typical screening experiment, the selectivity (Rubin et al. 2017) is a measure of the fitness of a protein mutant computed from the sequencing samples of the population. In its simpler form, it is the ratio of the sequence counts at two consecutive rounds. Slightly different definitions of selectivity are present in the literature (Rubin et al. 2017), aiming to reduce the impact of experimental noise on the computed mutants fitness.

There are several sources of noise that affect the reproducibility of a DMS experiment. Sequences that are present in low numbers are more susceptible to statistical fluctuations. This can be due to the uneven initial library composition that can change between different realizations of the same experiment. In addition, the attempts to cover a large sequence space can generate low replicates per mutants, since the availability of particles to carry the mutants is limited (e.g., in practice no more than about 10^{13} phages can be manipulated). Moreover, the sequenced mutants represent a very small subsample of the total diversity of variants in the experiment. Therefore, the reads statistics might not reflect fairly the underlying variant abundances. The magnitude of this sampling error depends on the *mutants coverage* that we define as the ratio of the total number of reads over the number of unique mutants, that is, the number of reads per variant in a hypothetical uniform distribution case. In table 1, we list the mutants coverage for each used data set. The sampling noise affects both the trained model and the selectivity measure but, interestingly, as we see later has a more prominent effect on the latter.

Validation of Mutants Fitness Predictions

To validate the inferred genotype–phenotype map, we do not have access to direct high-throughput measures of the binding energy with the target. Nevertheless, we can assess the reconstructed fitness landscape comparing the predicted binding energy with out-of-sample sequence selectivities. To do so, we perform a leave-one-out 5-fold cross-validation, that is, masking a fifth of the mutants in the library from the learning data and testing on the remaining ones. To mitigate the effect of the sampling noise on the selectivity measure, we filter out sequences with high selectivity error (see Materials and Methods for details) from the test set. In figure 1, we show the correlation of the predicted binding energy and the log-selectivity for each examined data set. In all experiments, we obtain an excellent agreement between the out-of-sample model prediction and the selectivity measure based on read counts (i.e., a proxy of the binding energy). The correlation between the two values steadily increases as we filter out more noisy sequences from the validation set (see Materials and Methods for details on the noise filter).

As previously pointed out, in contrast to other approaches that fit sequence selectivities, we train a model directly on the sequencing reads, maximizing the model likelihood of the full set of read counts from the experiment and obtaining a statistical description of the differential composition of the combinatorial library across rounds. This allows us to obtain an estimate of the binding probability more reliable and robust against the experimental noise. To prove this statement, we create a decimated training set by selecting in each round of the experiment a random subset of the reads, and for each training-set realization we learn the model parameters. We perform the test on the Olson et al. data set, which has a high mutant coverage ($\sim$500).

The results are shown in figure 2: from panels *a* and *b*, we clearly see that the reliability of the selectivity decays faster as the decimation ratio increases, compared with that of the model which provides accurate predictions also in the highly undersampled regime. In other terms, the selectivity of a sequence derived from an undersampled data set is a worst statistical predictor of the full data set selectivity compared with our model predictor. Even if we use the correction strategy outlined by Rubin et al. (2017), the selectivity measure is severely impacted by sampling noise, whereas the predicted fitness landscape inferred by our model is more robust to undersampling. Finally, the measure of selectivity relies

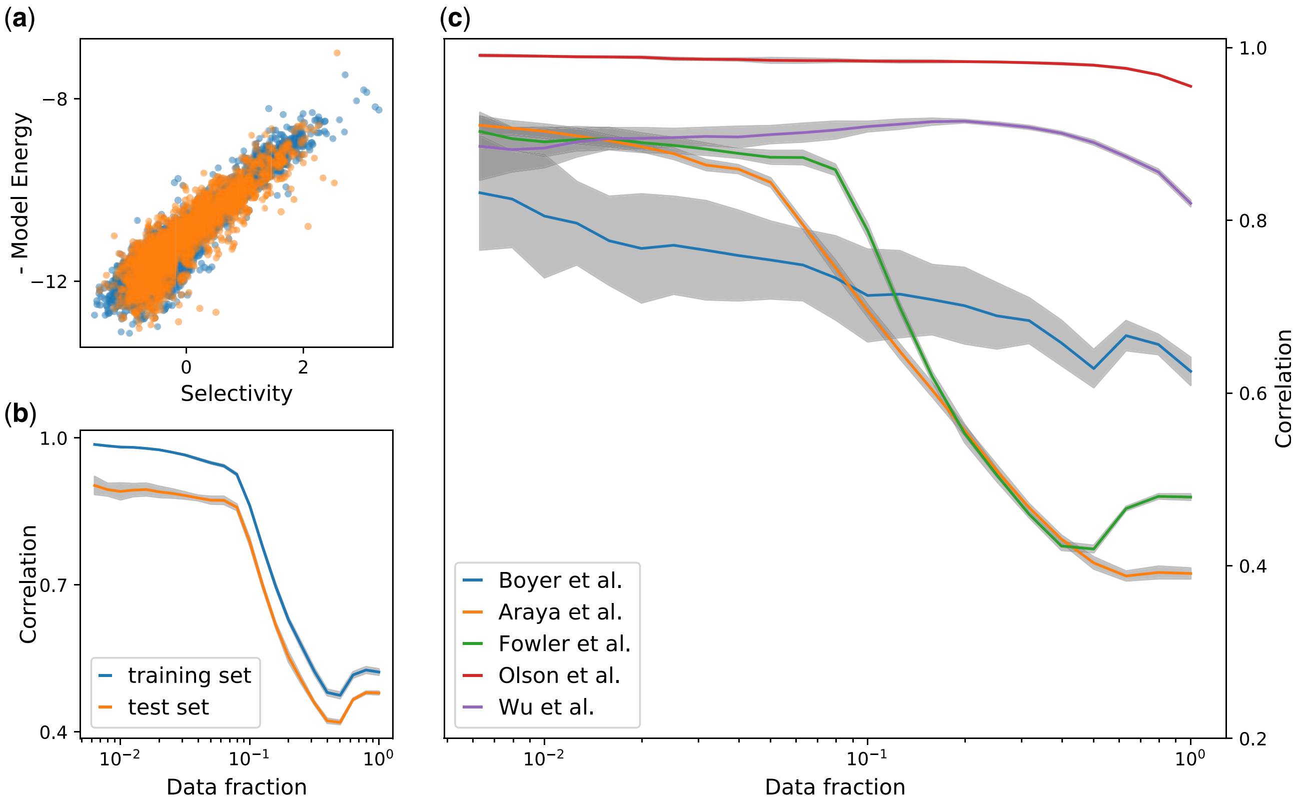

FIG. 1. Overall model performances. Correlation of the predicted binding energies E and the log-selectivity θ computed from the sequencing reads. (*a*) Scatter plot of E versus θ for the in-sample (blue dots) and out-of-sample (orange dots) sequences in the Araya et al. data set. The Pearson correlation is $\rho = 0.81$, after filtering out the noisier data (fraction of used data $f = 0.05$). (*b*) Pearson correlation coefficient between E and θ for different filtering thresholds on sequence errors in the Araya et al. data set. On the x-axis, the fraction of the sequences used to compute the correlation, a lower fraction of sequences account for a more severe filter on noisy sequences and provide higher correlations. The comparison of the in-sample and out-of-sample sets (four-fifths and one-fifth of the mutants, respectively) shows a minor overfitting bias. (*c*) Comparison of the Pearson correlation curves (same of *b*) for the five data sets. Notice that the higher the mutants coverage of the data set (Olson et al. and Wu et al.), the higher the correlation reached (see table 1). In all cases, the noise filter is used after the model learning which uses all sequences in the training test (see Materials and Methods for details).

upon the presence of multiple reads of the same mutant across the rounds, whereas our approach seems to be less impeded by this limitation.

Generalization Extent of the Inferred Fitness Landscape

The fitness landscape refers to the selection process of a specific phenotype of a protein. We investigate to which extent the model is able to extract information about general features of the landscape that can be used to provide reliable predictions on different experiments, different energy spectra, or different sequence space regions with respect to the training ones.

An intriguing question is to what extent different experimental settings can have an impact on the inferred parameters, reducing its generalization power, or whether we can use the learned map to predict new experimental outcomes. Fowler et al. (2010) and Araya et al. (2012) published two experimental data sets, in which the hYAP64 WW domain is selected for binding against its cognate polyproline peptide ligand. In both cases, most variants in the library are on average two a.a. substitutions away from the wild-type sequence. Still, the initial libraries of the two experiments have only about 50% sequences in common, with the rest of the sequences being unique to each data set. The model trained on one data set (discarding the common sequences) provides accurate predictions of the empirical selectivities observed in the other experiment, as shown in figure 2*a*. Interestingly when the common sequences are taken into account, the binding energies inferred from the two data sets show better correlations than the selectivities. This suggests that the inferred energy is more reproducible and robust to noise, and hence is a better estimator of the *true* fitness.

We repeated the same analysis training on the Olson et al. data set and testing on the Wu et al. data set, as they both used the immunoglobulin G (IgG)-binding domain of protein G (GB1) and performed the selection for binding to immunoglobulin G fragment crystallizable (IgG-Fc). In this case, we train on all the sequences at Hamming distance 1 and 2 from the wild type (from the Olson et al. data set) and we ask whether the model can make predictions of sequences three or four mutations away from the wild type (from the Wu et al. data set). The results in supplementary figure S4, Supplementary Material online, show that the model is still able to predict the fitness landscape for more distant mutants (Pearson correlation of $\rho = 0.67$ for Hamming distance 3

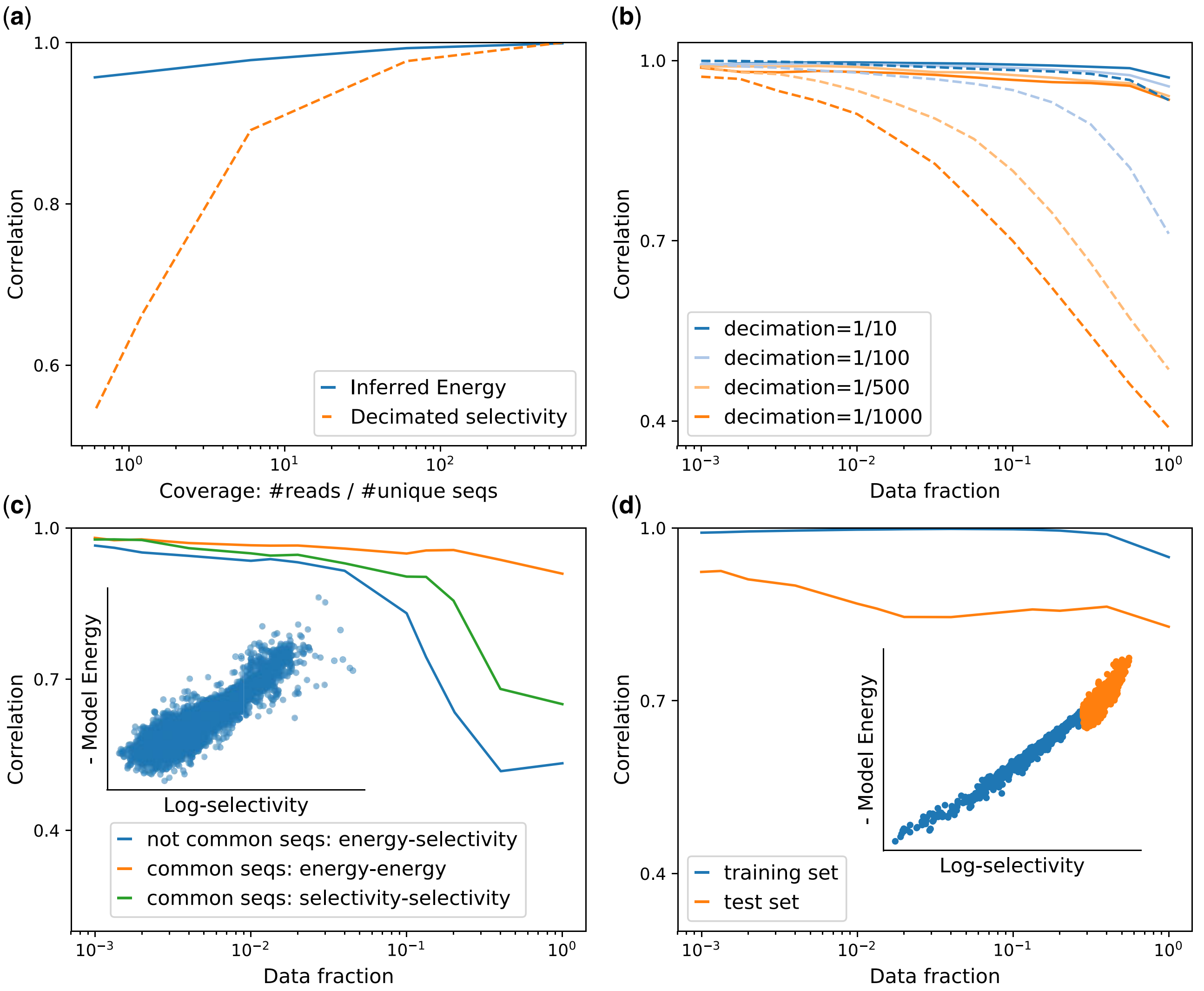

FIG. 2. Model robustness and generalization power. In (*a*) and (*b*) is shown the robustness of the model inference with respect to sampling noise. We reduce the mutants coverage decimating randomly the reads in the Olson et al. data set, obtaining data sets with different numbers of total reads. We use them to infer the binding energy and to compute the selectivity after decimation and we compare them with the full data set selectivity. In (*a*), we show the Pearson correlation coefficient between the full data set selectivity and the predicted energy of the model and the decimated selectivity on a test set, as a function of the coverage of the decimated data sets. In (*b*), we show the correlations of the two measures with the full data set selectivity for different thresholds of filtering of mutants errors on the test set. Although the Pearson correlation between the decimated data set selectivity and the original one reduces drastically upon increasing the decimation rate, the predicted binding energy maintains always a high correlation with the test-set selectivity. (*c*) Correlation of predicted energies E and empirical selectivity θ, when the model is trained on one data set (Araya et al.) and tested on the outcome of a different experiment(Fowler et al.). On the x-axis, the fraction of the sequences used to compute the correlation after filtering on error. The blue curve refers to the model trained on sequences that are not common to the two data sets. In the inset, a scatter plot of E and empirical log-selectivity for a particular choice of filter threshold (data fraction $f = 0.05$, correlation $\rho = 0.91$). The yellow curve corresponds to the correlation between inferred energies of the sequences common to the two data sets, whereas the green one refers to the correlation between selectivity compute on the same sequences in the two data sets, interestingly is lower than the previous correlations suggesting that energy is a more reproducible quantity than the empirical selectivity itself. (*d*) Capacity of the model to predict best binders. Correlation of predicted energies E and empirical selectivity θ trained on low selectivity mutants and test on the top selectivity ones. The high correlation in the test set and the capacity to rank properly the unseen best binders suggest the promising application of the method as a generative model.

and $\rho = 0.55$ for Hamming distance 4), although these predictions deteriorate as the distance to the sequences covered in the training set increases.

Can we learn from sequences in the low binding energy-band a predictive model of the high binding energy-band of the fitness landscape? This is a relevant question if we want to exploit the model to generate, for instance, better binders not originally present in the experiment. To gauge the performance of the model for this task, we use the sequences with low selectivity as training set and the sequences with higher selectivity as the test set.

In contrast to the previous results where the out-of-sample sequences were extracted from the same distribution as the in-sample, in the present computation we selectively learn from low-medium binding energy sequences whereas we test on the top binders. As shown in figure 2c, the predictions are in excellent agreement with the data, showing the capacity to learn the fitness landscape of high-fitness region very accurately.

Epistasis

The role of intragenic epistatic interactions in shaping the fitness landscape is a subject of intense research, with different contrasting results being largely debated in the scientific community (Winter et al. 1994; Rodrigues et al. 2016; Echave and Wilke 2017; Otwinowski et al. 2018; Domingo et al. 2019). Although on the evolutionary scale it is clear that epistasis has an important role in shaping the sequence ensemble of protein family domains across homologs (Miton and Tokuriki 2016; Starr and Thornton 2016; Cadet et al. 2018; Sun et al. 2019), on a more local scale, in the selection of local mutations around a wild-type sequence for binding a target, there is some debate whether such effects are involved and to what extent. It has been shown that the evolution of novel functions in proteins benefits from epistatic interactions, which enable mutational paths that would otherwise not be accessible (Miton and Tokuriki 2016). Due to modeling difficulties, epistasis is often viewed as an challenging obstacle to predicting mutational effects in protein engineering (Cadet et al. 2018; Sun et al. 2019) and is sometimes ignored (Otwinowski 2018). We investigated whether a model without epistasis, where mutation effects are independent in each residue and provide additive contributions to fitness, can reach the same description accuracy of the experiments. (See Materials and Methods for details on the independent site model and the pairwise epistatic one.)

The five data sets considered in this study vary with respect to the broadness of sequence space sampled (how far from the wild type are the mutants in the library) and the length of the mutated part of the sequence, as summarized in table 1. The two opposite limits are the Olson et al. data set where the full length of the GB1 (55aa) is mutated only by a maximum of two amino acids (long sequence, limited broadness) and the Boyer et al. and Wu et al. data sets where only four amino acids are considered but the libraries cover a significant fraction of sequence space. The Araya et al. and Fowler et al. data sets lay in an intermediate regime.

Figure 3*a–c* shows the comparison of the performance of the independent site model and the pairwise epistatic model: The broader the sequence space covered in the experiment, the more crucial becomes the inclusion of the epistatic interactions in the model for a proper description of the experimental outcome.

Recent articles have pointed out that epistatic interactions can arise spuriously from nonlinearities in the genotype–phenotype map (Otwinowski et al. 2018; Domingo et al. 2019). Otwinowski proposed a global epistatic model that infers the parameters of an independent site model together with the nonlinear shape of the map. This method provides a good prediction of the fitness in DMS experiments (see supplementary figure S5, Supplementary Material online, for the performance on all the data sets), considering also the lower number of parameters used. Nonetheless, performing the same analysis to test robustness and generalization highlighted in the previous paragraph, the global epistasis model appears to be sensitive to sampling error (fig. 3*e* and *f*) and fails to predict higher fitness mutants when trained on the low fitness ones (fig. 3*f*).

In recent studies, it has been demonstrated that epistatic interactions, quantified from DMS experiments, can be used to determine 3D contacts in the molecular structures (Rollins et al. 2019; Schmiedel and Lehner 2019; Fantini et al. 2020). This finding provided a strong support to the idea that the observed epistatic interaction does not come only from nonspecific artifacts due to the nonlinearity of the fitness map, but rather reflects the interplay of structural stability and functional binding of the selection process in the experiment.

We investigate whether and to which extent the proposed model provides contact predictions that can be used for 3D structure modeling, on the GB1 domain of protein G using the Olson et al. DMS experiment. To test the predictions, we use a crystal structure of the protein in complex with the Fc domain of human IgG (PDB id 1fcc). As a measure of the epistasis between two positions, we use the average difference in binding energy between the sum of single mutations and the double mutants, similar to the score used by Tubiana et al. (2019) (see the Materials and Methods section for details). We identified the most epistatic pairs by sorting all of the pairs of positions by order of decreasing epistatic score.

In figure 3g, we show the receiver operating characteristic (ROC) curve of the predictions compared with the weighted Mutual Information (MI; a nonparametric measure defined in supplementary eq. S.10, Supplementary Material online). In figure 3*h* are shown the first 20 predictions and the true contact maps. We note that the MI predictions are strongly clustered around the binding surface, whereas the model contact predictions are distributed all over the structure, making them more useful for 3D structure modeling. In the Supplementary Material online, we report the same structural analysis for the WW domain, using the data set from Fowler et al. and Araya et al.

Discussion

Despite advances in high-throughput screening and sequencing techniques, investigating genotype–phenotype relationships remains a substantial challenge due to the enormous size and large dimensionality of the space of possible genotypes. We propose a computational method to obtain a model of the genotype to fitness map, learned from the sequencing data of a DMS experiment. The novelty of the method consists of an unsupervised approach that uses a probabilistic description of the full amplification–selection–sequencing phases of the experiment. One key element lies in the inclusion of pairwise epistatic interactions in the modeling of the specific mutant selection, nevertheless we remark that in the same framework other modeling schemes are possible.

To investigate the properties of the inferred fitness landscape, we used five DMS experiments related to two well-

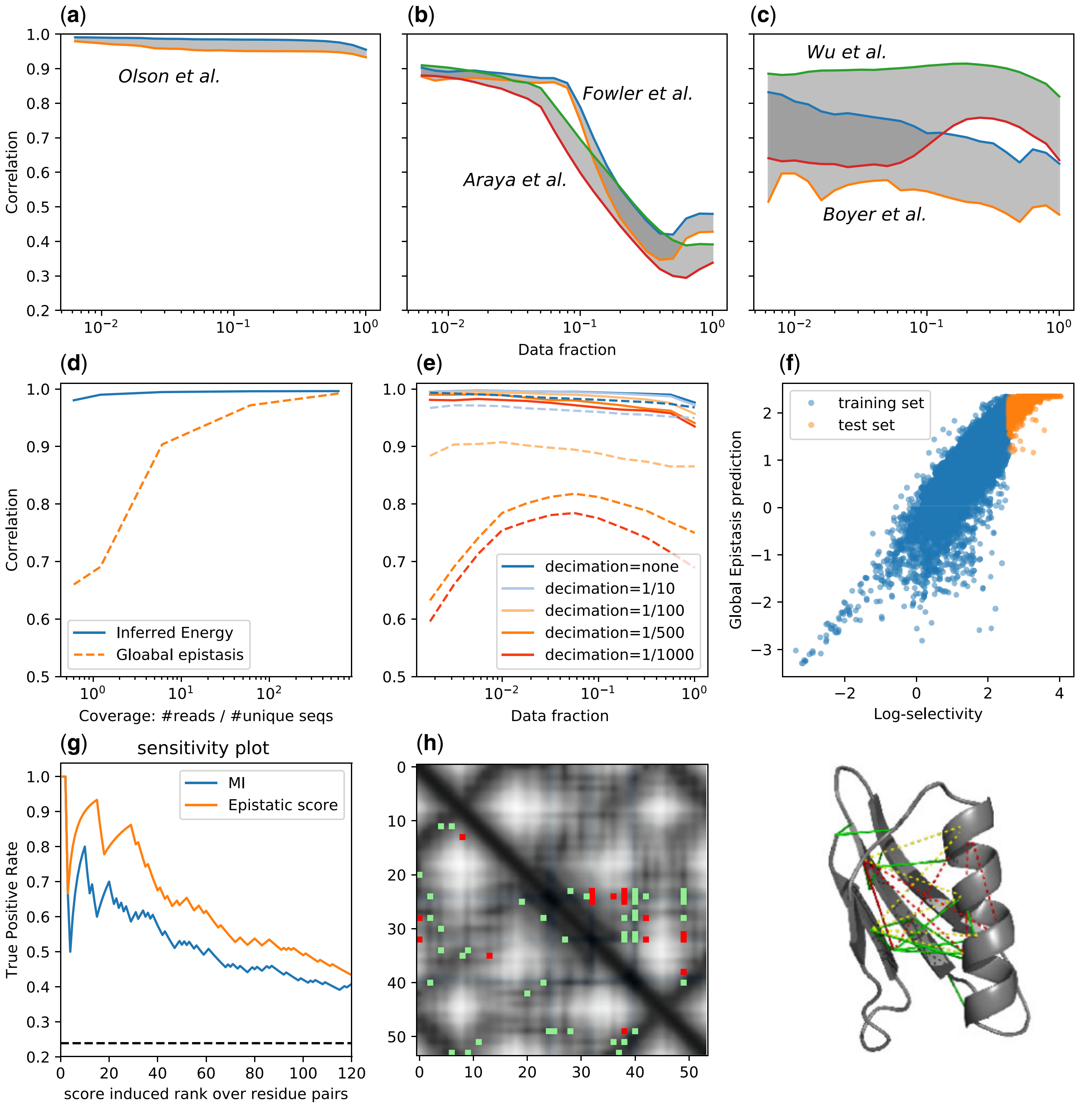

FIG. 3. Epistatic effects. The comparison of the independent site model and the epistatic model is displayed in (*a*)–(*c*). The three panels refer to different characteristics in the broadness of the screened library and the length of the mutated part of the protein sequence. From left to right, broadness increase and the number of the mutated residues reduce (see table 1). For each data set are shown the correlation of the epistatic model (upper line) and the independent site model (lower line) and is highlighted in gray the gap between the two. Broader the library, more distant mutants from the wild type are screened and more the epistatic effects become relevant. In (*d*)–(*f*) are shown the robustness and generalization analysis (same as depicted for the epistatic model in fig. 2) of the global epistasis model (Otwinowski et al. 2018). (*d* and *e*) Display the reduction of the accuracy of GE model when lowering the mutants coverage. (*f*) The low correlation between GE prediction and selectivity for high-fitness mutants depicts the deficiency to generalize the prediction to lower binding energy spectra. (*g*–*i*) The structure contacts predictions for GB1 domain form the Olson et al. DMS experiment (tested on the crystal structure PDB id: 1fcc). (*g*) The ROC curves of the predicted contacts with the epistatic score computed from the inferred model and the weighted MI. (*h*) Contact map of the first predicted contacts. In gray-scale is displayed the distances between residue heavy atoms, in blue are highlighted the residues on the binding surface (<3 Å to the Fc domain of human IgG). The green dots are the true positives (heavy atoms distance <8 Å) and the reds are false positives. In the upper triangular part are shown the MI predictions, whereas in the lower triangular part the epistatic score. The MI predictions are strongly clustered around the binding surface, whereas the model predictions cover the whole structure. (*i*) The same predictions on the molecular structure of G protein in complex with the Fc domain of human IgG.

studied proteins, the WW domain part of the YAP65 protein, the GB1 domain of the IgG-binding protein, and the variable part of a human antibody, all selected for binding to cognate ligands. These experiments differ in several technical characteristics, such as library generation and expression, length of the mutated part of the protein, broadness of the initial library, and sequencing coverage per mutants.

First, we performed a cross-validation test obtaining accurate selectivity predictions for all the five data sets. Second, we investigated the generalization power of the model. We learned on one experiment and predicted correctly the outcome of a second one with same wild-type protein and binding target, obtaining a better experimental reproducibility than from the mutant selectivities themselves. Remarkably the model shows the capacity to predict lower binding energy spectra, we masked the higher fitness mutants from the training and we recover the correct ranking and fitness of the best binders. Moreover, we noticed that the predicted fitness landscape is more robust to experimental noise than the selectivity measures (the fitted quantity in the supervised approach). To demonstrate this, we performed a decimation of sequencing reads and assess the detrimental in the predictions.

Finally, we investigated the reliability of the epistatic interactions in the model. Our results show that when increasing the library's sequence diversity, epistatic interactions become more important to obtain a good fit to the experiments. In addition, we can extract from the inferred epistatic interactions, structural information of the 3D contact proximity.

Recently, it has been pointed out that nonlinearities in the genotype to fitness map can produce spurious epistatic effects, namely nonspecific epistasis or global epistasis (Otwinowski et al. 2018; Domingo et al. 2019). This suggests a limited magnitude of specific epistatic effects in shaping the fitness landscape in local screening assays. We compared the two hypothesis and our analysis suggest that although the spurious global epistasis could have a prominent role where the experiment is selecting *complex* phenotypes (among others: cell growth rate [Roscoe and Bolon 2014; Mishra et al. 2016] or a proxy of expression levels [Sarkisyan et al. 2016]), in the set of experiments we have analyzed, where the selection is upon the binding affinity to a target molecule, the specific epistatic effects account for real genetic interactions.

From all these findings, we speculate that the presented unsupervised approach could be utilized as a generative model to identify novel high-fitness variants and can be included in a machine-learning-assisted Directed Evolution framework where the computational part are included in the cycle to design the combinatorial libraries to be screened (Saito et al. 2018; Yoshida et al. 2018; Yang et al. 2019). Experimental tests of predicted novel sequences are undergoing and will be published in the future.

Materials and Methods

The machine learning method we developed makes use of an unsupervised approach based on likelihood maximization. This approach uses as input data exclusively the sequencing reads and does not require any biophysical measurement of the molecule variants to train it or to ground the results on the robustness of some statistical proxy of the fitness. The cost is to define the likelihood function to observe a time series of reads in an experiment given the parameters involved in the selection. In the following section, we outline the probabilistic description of the experiment in terms of selection, amplification, and sequencing phases. The probability that a protein variant is selected depends on the specific variant amino acid composition and represents the genotype to fitness map or fitness landscape. The model parametrizes the fitness both with additive contributions from the single residues and with the epistatic contributions in the form of pairwise interactions, although different parametrization schemes could be introduced.

Inference

We consider a set of experimental rounds of selection $t \in 0, 1, \ldots, T$, with $t = 0$ referring to the initial combinatorial library, and $t = T$ denoting the last round. At round t, we denote by N_s^t the number of phages displaying sequence s. Each of the phages carrying sequence s has a certain probability p_s of being selected for the next round, called the *selectivity*. This probability is determined by the properties of the molecule affecting its fitness in the experiment, for example, the affinity toward a binding target. The number of phages that will be selected for the next round can then be taken as a binomial distributed random variable, with mean $p_s N_s^t$. Since a large quantity of phages is present initially, we can approximate this as the deterministic selection of a fraction $p_s N_s^t$ of phages for each sequence. Usually, selection is stringent and most phages are washed away. The selected pool must be amplified to recover the initial population size. We model this step as a stochastic multinomial distribution,

$$P(\mathbf{N}^{t+1}|\mathbf{N}^t, \mathbf{p}) = \frac{N^t!}{\prod_s N_s^{t+1}!} \prod_s \left(\frac{p_s N_s^t}{\sum_\sigma p_\sigma N_\sigma^t} \right)^{N_s^{t+1}} \quad (1)$$

with $t = 0, \ldots, T-1$ and $N^t = \sum_s N_s^t$ is the total number of phages at round t, and a bold symbol such as $\mathbf{N}^t$ denotes the vector of all N_s^t for all sequences at round t. The full experiment consists of iterating these two steps (selection and amplification). Finally, at selected rounds, a sample of the amplified population is sequenced. In the limit of large enough sample size, we assume that the read counts are approximately proportional to the frequencies of sequences in the population (see supplementary appendix, Supplementary Material online, for details).

Under these assumptions, it follows that the likelihood of the time series $\{\mathbf{N}^0, \mathbf{N}^1, \ldots, \mathbf{N}^T\}$ of phage abundances for each sequence is given by the product of (1) from $t = 0$ to $t = T - 1$. The inference of the model is carried out by maximizing this likelihood in the parameters p_s.

Genotype to Fitness Map

The selection probabilities can be modeled by a two-state thermodynamic model (bound or unbound), $p_s = 1/(1 + e^{E_s - \mu})$, where E_s is the binding energy of sequence s and μ is the chemical potential, which depends on the concentration of binding target presented in the experiment (Haldane et al. 2014).

As a function of sequence, E_s is a genotype-to-phenotype mapping that assigns a biophysical parameter (binding energy) to each sequence. We assume that E_s decomposes into additive contributions from individual a.a. species in the sequence, plus epistatic contributions from interacting pairs of letters:

$$E_s = -\sum_i h_i(s_i) - \sum_{i<j} J_{ij}(s_i, s_j). \tag{2}$$

The problem becomes that of inferring the parameters $h_i(a), J_{ij}(a,b)$ by maximizing the likelihood (1) over all rounds. In addition, a regularization (e.g., an $\mathcal{Q}_2$-norm) term can be included to prevent overfitting.

Rare Binding Approximation

In typical experiments, the fraction of selected phages is very small, implying that $p_s \ll 1$ for most sequences. This suggests a rare binding approximation, $p_s \approx e^{\mu - E_s}$. Under this approximation, the log-likelihood simplifies to

$$\mathcal{Q} = \sum_{t=0}^{T-1}\sum_s \ell_s^t, \quad \mathcal{Q}_s^t = N_s^{t+1}\ln\left(\frac{N_s^t e^{-E_s}}{\sum_\sigma N_\sigma^t e^{-E_\sigma}}\right). \tag{3}$$

In this limit, the log-likelihood does not depend on μ anymore. It also makes ℓ a concave function of the energies E_s, and hence of the fields $h_i(a)$, $J_{ij}(a,b)$ that we intend to learn. Since our algorithm consists in finding the maximum of (3) with respect to these parameters, concavity guarantees that the solution is unique and that it can be found efficiently with numerical optimization routines. In our implementation, we found that the L-BFGS algorithm performed well.

Independent Site Model

Due to the rare binding approximation, when in the energy terms are considered only the h parameters contribution, each residue contributes independently and there are no epistatic effects as the amino acid changes impact additively to the fitness, $p_s \approx \prod_i e^{h_i}$.

Empirical Selectivity

To compute empirical selectivities, we performed a linear regression of the parameters θ_s, α^t, in a model of the form:

$$\ln N_s^t - \ln N_s^{t-1} = \theta_s + \alpha^t + \epsilon_s^t, \tag{4}$$

where θ_s is the (empirical) log-selectivity, α^t an amplification factor, and ϵ_s^t is a normal distributed measurement noise. We performed a weighted least squares regression, assuming approximate independent variances $1/N_s^t$ for the terms $\ln N_s^t$, given that counts follow Poisson distributions. To mitigate the effect of low counts, we add a pseudo-count of 1/2 to all counts before computing the empirical selectivities and before carrying out the inference (Rubin et al. 2017).

Noise Filter

We estimated error bars for the selectivities θ_s by standard linear regression formulae and used it to filter out sequences from the validation set. Sequences for which the empirical selectivity has an error bar higher than a given threshold are filtered out from the validation set. This approach is based on Rubin et al. (2017), which provides evidence for the robustness of selectivities computed in this way, given an appropriate choice of the threshold. In our analysis, we do not choose an a priori threshold but we consider several threshold values providing increasingly severe filters (the data fraction left after the filtering procedure is shown on x-axis of figs. 1b, 1c, 2b–2d, 3a–3c, and 3e). We stress that the whole filtering procedure does not impact the learning of the model since the filtering is performed on the validation set. All sequences are considered for the model training since the inference procedure is robust against low count noise (as we showed in fig. 2). Nonetheless, the filtering procedure is useful to compare the results with more reliable empirical selectivity measures on the test set.

Structural Contact Predictions

The presence of a large epistatic effect between site positions is related to the 3D proximity of the residues in the protein fold (Morcos et al. 2011). To quantify the strength of the epistatic effect, we computed the difference between the fitness effect of double mutations and the sum of the effects of the two related single mutations, hence the expected additive fitness in absence epistasis. The genotype to fitness map in equation (2), the fitness of a sequence, is minus the energy $f(s) = -E(s)$.

Considering a sequence s, the double mutant ν_{ij} in position i and j and the single mutant ν_i (and ν_j resp.) in position i (and j resp.), we define the epistatic score as

$$S_{ij}(s, \nu_{ij}) = \Delta f(\nu_{ij}) - \Delta f(\nu_i) - \Delta f(\nu_j), \tag{5}$$

where $\Delta f(\nu_{ij}) = f(\nu_{ij}) - f(s) = -E(\nu_{ij}) + E(s) = \log[P(\nu_{ij})/P(s)]$ and similarly for $\Delta f(\nu_i)$ and $\Delta f(\nu_j)$. We substitute in S_{ij} and we average over each sequence s in the data set and for each possible double mutants, obtaining:

$$S_{ij} = \langle \log\left[\frac{P(s)P(\nu_{ij})}{P(\nu_i)P(\nu_j)}\right]\rangle_{s,\nu} = \tag{6}$$

$$= \langle E(\nu_i) + E(\nu_j) - E(s) - E(\nu_{ij})\rangle_{s,\nu}. \tag{7}$$

Data Sets

In order to assess the inferred genotype–phenotype map, we used five data sets of mutational scan studies (Fowler and Fields 2014) that assess experimentally the mutational landscape of three different proteins. In the Olson et al. data set (Olson et al. 2014), the effects of all single and double mutations between all positions in the IgG-binding domain of protein G (GB1) are quantified.

In this study, a library of all possible single and double amino acid substitutions of the 55 sites of the GB1 protein domain is screened in a single round for the binding to an immunoglobulin fragment (IgG-Fc). The same protein and target pair were investigated by Wu et al. (2016), who selected four positions in GB1 and exhaustively randomized them. Two more data sets come from Fowler et al. (2010) and Araya et al. (2012), where a WW domain was randomized and selected for binding against its cognate peptide.

In the two studies, the wild-type protein and the target are the same and, interestingly, the initial randomized libraries in the two data sets have about half of the sequences in common.

Finally, in Boyer et al. (2016), four positions of a variable antibody region are fully randomized and selected for binding against one of two targets: polyvinylpyrrolidone (PVP) or a short DNA loop with three cycles of selection.

The four positions are embedded into one of 23 possible frameworks. The main characteristics of the biological system and of the experimental settings are summarized in table 1. Although in Olson et al. (2014), the high mutant coverage (500) allows us to obtain a low sampling noise, the main limitations are due to the covered sequence space, limited to a maximum Hamming distance of two from the wild-type sequence. In Boyer et al. (2016), the fraction of covered sequence space is significant (16% of all possible sequences) and there are multiple rounds of selection but the obvious limitation comes from the small length of the mutated part of the sequence (4 a.a.).

Compared with the previous data sets, the Fowler et al. and Araya et al. data sets have intermediate features where the covered sequence space is wider than in Olson et al. (average distance 3.4 and 4.3) and the number of selection rounds is respectively 3 and 4, but the sequencing depth is lower showing greater sampling noise.

Supplementary Material

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

Acknowledgments

The authors acknowledge financial support from Marie Skłodowska-Curie (grant agreement no. 734439) (INFERNET) and the Centro de Inmunologia Molecular of Cuba and the Department of Physics of University of Havana, for their kind hospitality. We also thank C. Nizak, O. Rivoire, and K. Leon for many interesting discussions.

Author Contributions

All authors contributed equally.

References

Aharoni A, Griffiths AD, Tawfik DS. 2005. High-throughput screens and selections of enzyme-encoding genes. *Curr Opin Chem Biol.* 9(2):210–216.

Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S. 2012. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci U S A.* 109(42):16858–16863.

Asti L, Uguzzoni G, Marcatili P, Pagnani A. 2016. Maximum-entropy models of sequenced immune repertoires predict antigen–antibody affinity. *PLoS Comput Biol.* 12(4):e1004870.

Barrat-Charlaix P, Figliuzzi M, Weigt M. 2016. Improving landscape inference by integrating heterogeneous data in the inverse Ising problem. *Sci Rep.* 6(1):37812.

Boyer S, Biswas D, Soshee AK, Scaramozzino N, Nizak C, Rivoire O. 2016. Hierarchy and extremes in selections from pools of randomized proteins. *Proc Natl Acad Sci U S A.* 113(13):3482–3487.

Cadet F, Fontaine N, Li G, Sanchis J, Chong MNF, Pandjaitan R, Vetrivel I, Offmann B, Reetz MT. 2018. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci Rep.* 8(1):1–15.

Domingo J, Baeza-Centurion P, Lehner B. 2019. The causes and consequences of genetic interactions (epistasis). *Annu Rev Genomics Hum Genet.* 20(1):433–460.

Echave J, Wilke CO. 2017. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Annu Rev Biophys.* 46(1):85–103.

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.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S. 2010. High-resolution mapping of protein sequence–function relationships. *Nat Methods.* 7(9):741–746.

Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat Methods.* 11(8):801–807.

Haldane A, Manhart M, Morozov AV. 2014. Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol.* 10(7):e1003683.

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.

Kemble H, Nghe P, Tenaillon O. 2019. Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. *Evol Appl.* 12(9):1721–1742.

Kinney JB, McCandlish DM. 2019. Massively parallel assays and quantitative sequence–function relationships. *Annu Rev Genomics Hum Genet.* 20(1):99–127.

Louie RH, Kaczorowski KJ, Barton JP, Chakraborty AK, McKay MR. 2018. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc Natl Acad Sci U S A.* 115(4):E564–E573.

Magurran AE. 2013. Measuring biological diversity. Oxford: Blackwell Ltd.

Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, Ndung'u T. 2014. The fitness landscape of hiv-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput Biol.* 10(8):e1003776.

Mishra P, Flynn JM, Starr TN, Bolon DN. 2016. Systematic mutant analyses elucidate general and client-specific aspects of hsp90 function. *Cell Rep.* 15(3):588–598.

Miton CM, Tokuriki N. 2016. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* 25(7):1260–1272.

Molina-Espeja P, Vina-Gonzalez J, Gomez-Fernandez BJ, Martin-Diaz J, Garcia-Ruiz E, Alcalde M. 2016. Beyond the outer limits of nature by directed evolution. *Biotechnol Adv.* 34(5):754–767.

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.

Olson CA, Wu NC, Sun R. 2014. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol*. 24(22):2643–2651.

Otwinowski J. 2018. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol Biol Evol*. 35(10):2345–2354.

Otwinowski J, McCandlish DM, Plotkin JB. 2018. Inferring the shape of global epistasis. *Proc Natl Acad Sci U S A*. 115(32):E7550–E7558.

Reetz MT. 2013. Biocatalysis in organic chemistry and biotechnology: past, present, and future. *J Am Chem Soc*. 135(34):12480–12496.

Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods*. 15(10):816–822.

Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI. 2016. Biophysical principles predict fitness landscapes of drug resistance. *Proc Natl Acad Sci U S A*. 113(11):E1470–E1478.

Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, Marks DS. 2019. Inferring protein 3D structure from deep mutation scans. *Nat Genet*. 51:1170–1176.

Romero PA, Arnold FH. 2009. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol*. 10(12):866–876.

Roscoe BP, Bolon DN. 2014. Systematic exploration of ubiquitin sequence, e1 activation efficiency, and experimental fitness in yeast. *J Mol Biol*. 426(15):2854–2870.

Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM. 2017. A statistical framework for analyzing deep mutational scanning data. *Genome Biol*. 18(1):150.

Sadler JC, Green L, Swainston N, Kell DB, Currin A. 2018. Fast and flexible synthesis of combinatorial libraries for directed evolution. *Methods Enzymol*. 608:59–79.

Saito Y, Oikawa M, Nakazawa H, Niide T, Kameda T, Tsuda K, Umetsu M. 2018. Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth Biol*. 7(9):2014–2022.

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al. 2016. Local fitness landscape of the green fluorescent protein. *Nature* 533(7603):397–401.

Schmiedel JM, Lehner B. 2019. Determining protein structures using deep mutagenesis. *Nat Genet*. 51(7):1177–1186.

Schneidman E, Berry II, MJ, Segev R, Bialek W. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440(7087):1007–1012.

Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. 2005. Evolutionary information for specifying a protein fold. *Nature* 437(7058):512–518.

Starr TN, Thornton JW. 2016. Epistasis in protein evolution. *Protein Sci*. 25(7):1204–1218.

Sun Z, Liu Q, Qu G, Feng Y, Reetz MT. 2019. Utility of b-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem Rev*. 119(3):1626–1665.

Tizei PA, Csibra E, Torres L, Pinheiro VB. 2016. Selection platforms for directed evolution in synthetic biology. *Biochem Soc Trans*. 44(4):1165–1175.

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

Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR. 1994. Making antibodies by phage display technology. *Annu Rev Immunol*. 12(1):433–455.

Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* 5:e16965.

Wu Z, Kan SJ, Lewis RD, Wittmann BJ, Arnold FH. 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A*. 116(18):8852–8858.

Yang G, Withers SG. 2009. Ultrahigh-throughput FACS-based screening directed enzyme evolution. *ChemBioChem* 10(17):2704–2715.

Yang KK, Wu Z, Arnold FH. 2019. Machine-learning-guided directed evolution for protein engineering. *Nat Methods*. 16(8):687–694.

Yoshida M, Hinkley T, Tsuda S, Abul-Haija YM, McBurney RT, Kulikov V, Mathieson JS, Reyes SG, Castro MD, Cronin L. 2018. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* 4(3):533–543.