PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Inferring protein fitness landscapes from laboratory evolution experiments

Sameer D'Costa[1☯], Emily C. Hinds[1☯], Chase R. Freschlin[1], Hyebin Song[2]*, Philip A. Romero[1,3]*

1 Department of Biochemistry, University of Wisconsin, Madison, Wisconsin, United States of America, **2** Department of Statistics, Pennsylvania State University, State College, Pennsylvania, United States of America, **3** Department of Chemical and Biological Engineering, University of Wisconsin, Madison, Wisconsin, United States of America

☯ These authors contributed equally to this work.
* hps5320@psu.edu (HS); promero2@wisc.ed (PAR)

OPEN ACCESS

Citation: D'Costa S, Hinds EC, Freschlin CR, Song H, Romero PA (2023) Inferring protein fitness landscapes from laboratory evolution experiments. PLoS Comput Biol 19(3): e1010956. https://doi.org/10.1371/journal.pcbi.1010956

Editor: Jinyan Li, University of Technology Sydney, AUSTRALIA

Received: October 5, 2022

Accepted: February 16, 2023

Published: March 1, 2023

Data Availability Statement: Raw sequencing data generated by this experiment has been uploaded to the NCBI Sequence Read Archive (SRA Accession ID PRJNA923701) under the title Laboratory evolution of dihydrofolate reductase (DHFR). Also, software code to reproduce this computational method are available under an open source license at Github: https://github.com/RomeroLab/dhfr_neutral_evolution and archived at https://doi.org/10.5281/zenodo.7622051.

Funding: This work was supported by the US National Institutes of Health (5R35GM119854)

Abstract

Directed laboratory evolution applies iterative rounds of mutation and selection to explore the protein fitness landscape and provides rich information regarding the underlying relationships between protein sequence, structure, and function. Laboratory evolution data consist of protein sequences sampled from evolving populations over multiple generations and this data type does not fit into established supervised and unsupervised machine learning approaches. We develop a statistical learning framework that models the evolutionary process and can infer the protein fitness landscape from multiple snapshots along an evolutionary trajectory. We apply our modeling approach to dihydrofolate reductase (DHFR) laboratory evolution data and the resulting landscape parameters capture important aspects of DHFR structure and function. We use the resulting model to understand the structure of the fitness landscape and find numerous examples of epistasis but an overall global peak that is evolutionarily accessible from most starting sequences. Finally, we use the model to perform an *in silico* extrapolation of the DHFR laboratory evolution trajectory and computationally design proteins from future evolutionary rounds.

Author summary

Laboratory evolution has revolutionized our understanding of protein structure, function, and evolution, and has generated countless useful proteins broad applications in medicine, biocatalysis, and biotechnology. These experiments explore protein sequence space through iterative rounds of mutation and selection and can provide rich data of populations traversing the fitness landscape. In this paper, we present a statistical learning framework that models the evolutionary process and can infer the structure of the underlying protein fitness landscape from multiple snapshots along a laboratory evolution trajectory. We generate a dihydrofolate reductase (DHFR) laboratory evolution data set and apply our modeling approach to infer the landscape parameters. The estimated parameters pinpoint key residues that dictate DHFR structure and function. We use the resulting model

awarded to PAR. The funding agency had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare no competing financial interests.

to understand the local and global structure of the fitness landscape and to perform in silico directed evolution for protein engineering.

Introduction

The mapping from protein sequence to function forms a high-dimensional protein fitness landscape. Knowledge of this landscape is important for understanding and modeling natural evolution, diagnosing genetic diseases, and designing new proteins for applications in biotechnology, human health, and chemistry. This landscape is shaped by highly complex protein conformations, dynamics, and biophysical/biochemical mechanisms, and is defined over an astronomically large number of possible protein sequences. While the sequence-function mapping is challenging to model from a physical perspective, approaches from statistics and machine learning can be leveraged to infer the underlying landscape from sparsely sampled experimental and evolutionary data [1–4].

The statistical approaches to model the protein fitness landscape are built around two common data types that provide either labeled or unlabeled data. Labeled protein data consist of a set of amino acid sequences and how each of those sequences map to a particular protein property of interest, such as thermostability, enzyme activity, or binding affinity. These sequence-function data are commonly generated using protein mutagenesis libraries and medium- or high-throughput assays to assign functional labels [5, 6]. Supervised learning approaches such as linear regression or more sophisticated non-linear models can learn from labeled sequence-function data to infer the mapping from sequence to function [7–10]. Unlabeled protein data consist of natural protein sequences taken from genomic and metagenomic sequencing databases. Unsupervised learning approaches can learn from this unlabeled protein data to infer the fitness landscape [11–13]. Direct coupling analysis (DCA) is an important class of unsupervised learning methods that learn residue coevolution patterns from multiple sequence alignments of related sequences [14, 15]. The DCA approach has been used to predict the three-dimensional structures of proteins [16–18], model the effects of mutations [19], and design new proteins [20].

Directed laboratory evolution applies iterative rounds of mutation and selection to explore the protein fitness landscape [7]. As a population evolves it samples diverse regions of sequence space and generates evolutionary trajectories that can be used to understand the structure of the fitness landscape. Laboratory evolution data consist of protein sequences sampled from evolving populations over multiple sequential generations. These data don't naturally fit into established supervised or unsupervised learning paradigms. Previous work has treated laboratory evolution data similar to natural evolution data and performed unsupervised DCA methods to infer landscape parameters [21, 22]. While these approaches were effective at determining contacting residues in the three-dimensional structure, they ignore the sequential nature of laboratory evolution data and instead treat sequences from multiple generations as independent samples. [23] demonstrate a method to infer epistatic fitness landscapes using data from multiple rounds of deep mutational scanning experiments. Also, [24] use genetic time series data to infer epistatic fitness landscapes. These methods model how the time-dependent data are connected to each other. Observing how the evolutionary process unfolds over time provides valuable information about the structure of the fitness landscape.

In this work, we develop a statistical learning framework to infer the protein fitness landscape from laboratory evolution data. We use population genetics principles to develop a model of the underlying evolutionary process and build a likelihood function to estimate the

landscape parameters from multiple rounds of evolution. We performed 15 rounds of laboratory evolution on the enzyme dihydrofolate reductase (DHFR) to generate a large and diverse data set consisting of sequences sampled from multiple sequential generations. We applied our learning method to infer the DHFR landscape and found the model parameters capture important aspects of DHFR function and reveal landscape epistasis arising from interactions between residues. We used the learned model to understand the global structure of the fitness landscape by running thousands of evolution simulations and found all trajectories converged to the same sequence, suggesting a single global optimum despite many examples of local epistasis. Finally, we applied our model to start from where our experimental DHFR evolution left off and continue the evolutionary process *in silico*. This procedure was used to extrapolate the evolutionary trajectory and design new functional DHFRs that were beyond the training data.

Results

Laboratory evolution to explore the fitness landscape of dihydrofolate reductase

Laboratory evolution applies iterative rounds of mutation and selection to explore the protein fitness landscape. We performed a laboratory evolution experiment on murine dihydrofolate reductase (mDHFR) to search the fitness landscape for diverse sequences encoding DHFR activity (Fig 1). DHFR reduces dihydrofolate to tetrahydrofolate and plays an essential role in purine biosynthesis and cell growth.

We employed a commonly used selection strategy that applies the antibiotic trimethoprim to inhibit *E. coli*'s native DHFR and makes the cells reliant on heterologously expressed mDHFR, which is resistant to trimethoprim. We mutagenized mDHFR using error-prone PCR with a target mutation rate of four nucleotide substitutions per gene. We then transformed this library into *E. coli* and performed a selection to identify DHFR variants capable of supporting *E. coli* growth in the presence of trimethoprim. After this, we extracted plasmid DNA from all surviving variants, remutagenized these variants with error-prone PCR, repeated the selection process, and performed a total of 15 rounds of laboratory evolution. For each round, we kept track of the total number of transformants and the fraction of functional DHFR variants; the product of these two numbers gives an estimate of the population size (S1 Table). Over the 15 rounds of evolution the population had an average size of 300,000 DHFR variants and never went less than 40,000. In contrast to an experiment like [25] which only selects the fittest clone each round of evolution and ramps up the selection pressure as much as possible, our low stringency selection and large population size create a neutral evolutionary process that generates diverse sequences that maintain wild-type-like DHFR activity.

We performed DNA sequencing on directed evolution rounds 1–5 and 15 to obtain a sample of the evolutionary trajectory. From this sequencing data, we see the evolutionary process generates a distribution of sequences with varying Hamming distance from wild-type mDHFR and the evolving population increasingly drifts away from the starting sequence (Fig 1b and S1 Fig).

The round 15 population has an average of 11.9 amino acid substitutions, corresponding to 0.79 amino acid substitutions accumulated per round. We also observed the populations' pairwise Hamming distance increased linearly over the course of evolution. The fact that the average distance from wild-type and the average pairwise distance both increased linearly with each round indicates that most evolutionary trajectories are exploring independent directions on the landscape. We also visualized how the DHFR sequences generated in our directed evolution experiment fit into the larger protein family generated by natural evolution (Fig 1c).

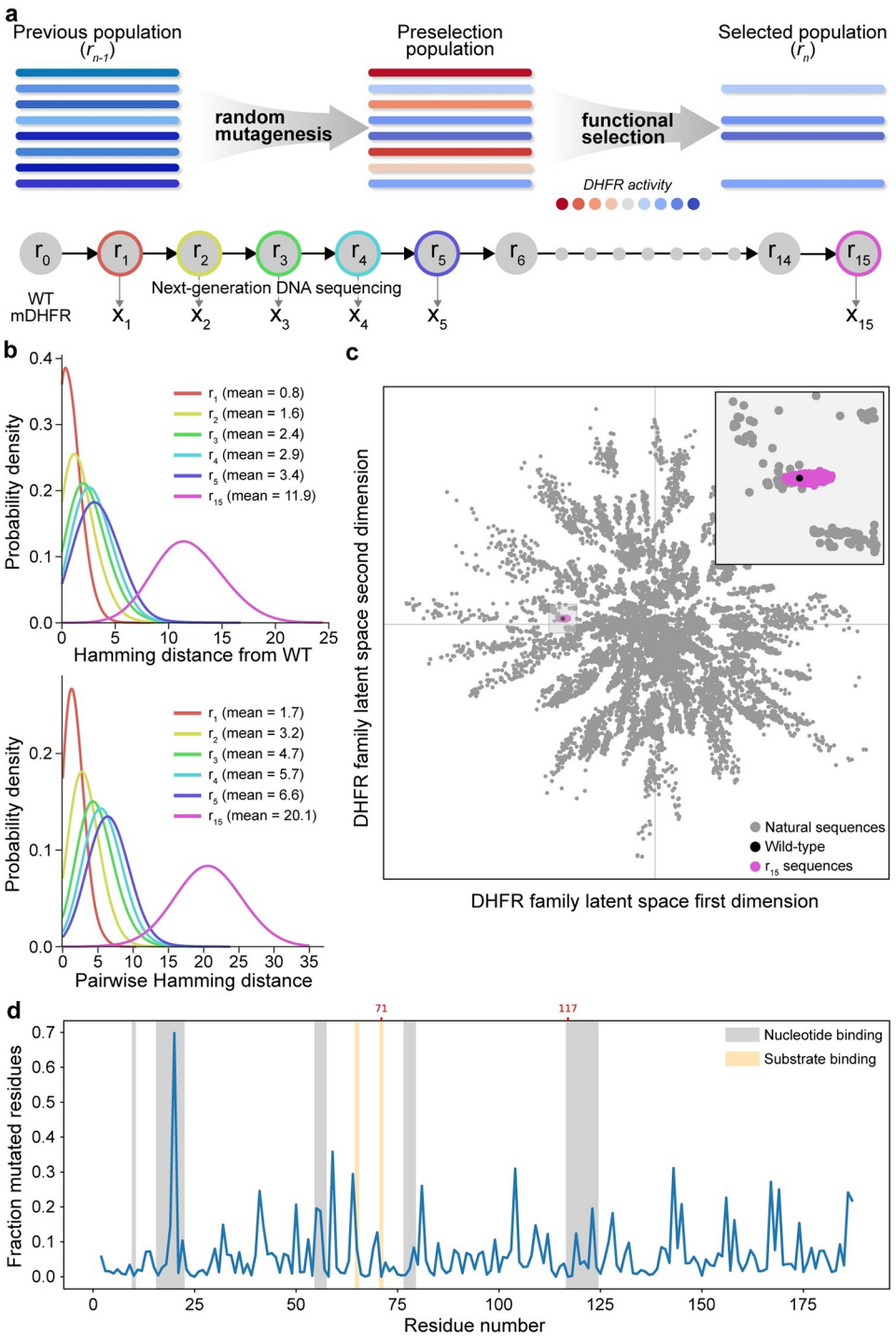

Fig 1. Experimental DHFR laboratory evolution (a) Laboratory evolution combines iterative rounds of random mutagenesis and functional selection to evolve populations of molecules. We performed 15 rounds of evolution on DHFR and sequenced the population at generations 1–5 and 15 to obtain snapshots along the evolution trajectory. (b) Sequence statistics of the experimental evolving populations displayed increasing distance from wild-type DHFR and increasing distance between sequences within a population, suggesting a diffusion-like spread into protein sequence space. (c) Visualization of the round 15 population in the context of natural DHFR sequences. The laboratory evolution experiment explored a small fraction of DHFR sequence space. Sequences were visualized using a two dimensional latent space of a variational auto encoder (VAE) trained on natural DHFR sequences (Details of the VAE are available in S2 Text). (d) The mutational statistics of sequences from the round 15 population displayed lower

mutation rates at active site residues. The mutation N20D overtook the wild-type asparagine residue in the round 15 population.

https://doi.org/10.1371/journal.pcbi.1010956.g001

Our directed evolution experiment started to capture similar variation to some natural sequences but only explored a small fraction of the sequence space spanned by nature.

We further analyzed the round 15 sequences to understand how mutations were distributed across the primary sequence. The mutations generally are distributed across the protein sequence but there is a lower observed mutation rate in the protein core and active site residues. We also observed the N20D mutation overtook the wild-type asparagine residue in the round 15 population, possibly indicating positive selection for this mutation. Residue 20 is in the active site region of mDHFR that binds nucleotide phosphates.

A statistical framework to learn from sequential rounds of experimental evolution

Directed evolution provides a sampling of populations evolving on a protein fitness landscape. We develop a statistical framework to infer the underlying landscape structure from these experimental evolutionary trajectories. We posit that data from sequential rounds of evolution provides inherently more information than data from a single round or data from multiple rounds considered independently. Observing how the evolutionary process unfolds over multiple rounds allows us to make stronger inferences and extrapolate behavior. For a simplified illustrative example, if we observe how the frequencies of amino acids change over each round, we can extrapolate these trajectories to estimate where the evolutionary process will converge. We build a generative model of the laboratory evolution process, parameterize the fitness landscape using a generalized Potts model, and infer the landscape parameters from sequencing data from multiple rounds of evolution.

We model the dynamics of laboratory evolution as a Markov chain process where sequences transition to other sequences according to their mutational accessibility and relative fitness. We make a number of assumptions on the sequence transition mechanism between rounds of evolution. First, we assume the mutation process happens independently at each DNA position and the mutation probability at each position is known from the experiment. Second, we assume that the transition mechanism is time-homogeneous, that is, both the fitness values and the mutation probabilities do not change between rounds. Third, we assume the number of experimental transformants is sufficiently large so that the distribution of sequences at a given round depends only on their relative fitness level. Finally, we assume simplified Markov chain dynamics, which assumes localized competitions between direct descendants, well approximates the true dynamics (see the Markov chain approximation to infinite population dynamics in S1 Text).

We parameterize the fitness landscape with a (generalized) Potts model that describes how all amino acid residues and their pairwise interactions contribute to fitness. Potts models have been used extensively to recover the interaction graph between residues of a protein and strong interactions have been shown to correspond to long-range contacts in the 3D structure of a protein [15, 17]. Directed evolution data consists of sequences sampled from multiple rounds of evolution and the observed sequences at each round are random realizations from the Markov model probability distribution. We can estimate the Potts model parameters by maximizing a statistical likelihood function to obtain the most likely model given the observed evolution data. The details of our evolutionary model and parameter estimation methods are given in the Methods section.

Learned landscape parameters capture protein structure and function

We applied the statistical learning framework developed above to infer DHFR's fitness landscape from our experimental laboratory evolution data. We estimated the Potts model's canonical parameters from evolution rounds 1–5 and 15, and used these parameters to get information about the protein's structure and function.

The learned Potts model reveals how individual amino acid substitutions affect wild-type DHFR's activity (Fig 2a). This learned mutation map clearly highlights the importance of the enzyme's key catalytic residues and displays expected mutational patterns residues with similar

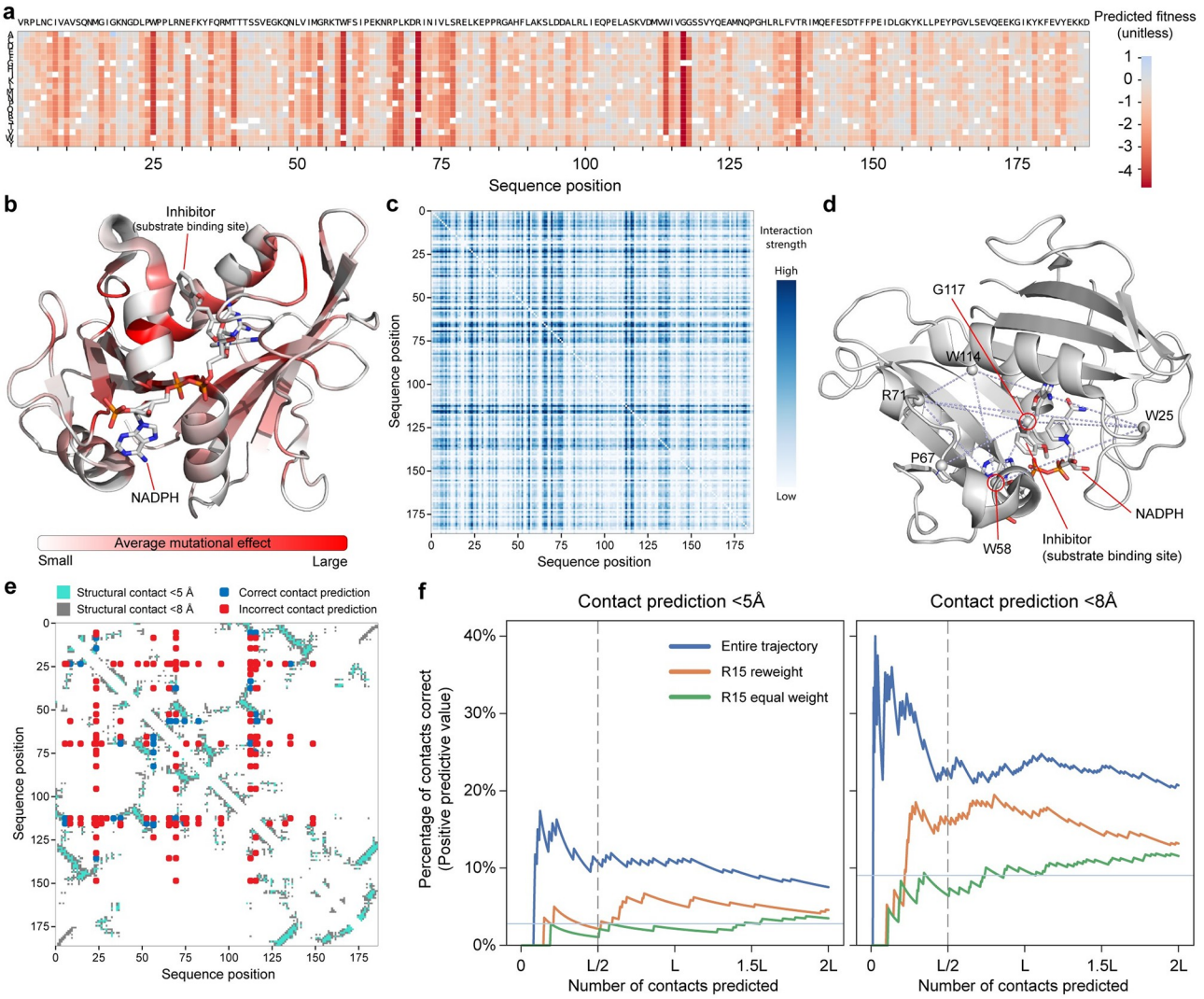

Fig 2. Model parameters relate to DHFR structure and function. (a) A heatmap of the model's predicted mutational effects across the 186 DHFR sequence positions. The wild-type amino acid is colored in white. (b) The average mutational effect at each site mapped onto the DHFR structure (PDB ID: 3K47). The largest magnitude mutations tend to occur in the protein core and the substrate binding site. (c) The interaction strength between all pairs of sites in DHFR. The interaction strength is calculated as the Frobenius norm over the Potts model interaction coefficient all amino acid combinations at each pair of sites. (d) The top ten long range interactions plotted on the DHFR structure (PDB ID: 3K47). Many of these interactions occur through interactions through the substrate. (e) The top L/2 (93) interactions between residues plotted on a contact map showing residues with heavy atoms closer than 5 Å and 8 Å. (f) A comparison of contact prediction for pseudo-likelihood DCA models trained on R15 data, R15 data reweighted to account for evolutionary bias, or using our methods on the evolutionary trajectory. The horizontal line represents random chance and the vertical line is drawn at the commonly used $L/2$ threshold.

https://doi.org/10.1371/journal.pcbi.1010956.g002

physiochemical properties. We compare this mutation map to a mutation map learned from the natural DHFR sequences using DCA [19] and Bayesian VAE using EVE [26] (S2 Fig). The EVE method aims to capture higher order interactions compared to DCA. The Spearman rank based correlation between the our inferred mutation map and that from DCA was 0.54 and to that from EVE was 0.62, showing that the fitness landscape learned from the evolution experiment is similar to but not the same as that learned from the natural sequences. The G18A amino acid substitution has the largest beneficial effect in the inferred model and is located in a loop that interacts with NADPH. We mapped the average mutational effect magnitude onto the three-dimensional DHFR structure to understand how the learned parameters relate to structure (Fig 2b). We observed that mutations in the protein core tend to have larger effects on activity, presumably because mutations at these sites disrupt the enzyme's three-dimensional structure.

The Potts model can also be used to understand landscape epistasis and interactions between residues. We compute the fitness score for all single and double mutants of wild-type and we identified thirteen examples of reciprocal sign epistasis in the landscape out of approximately 6 million possible amino acid and position pairs. One such example in the inferred landscape occurs between K47A and R99E, where the individual mutations are beneficial individually, but when combined result in a decrease in fitness. We observed 24695 examples of sign epistasis in the landscape where one mutation has an opposite effect in the presence of another mutation. We compute a residue-residue interaction score by calculating the Frobenius norm F_{ij} between all interaction parameters between a pair of residues (Fig 2c). The inferred residue interactions capture many contacts from the enzyme's three-dimensional structure and also functional interactions in the enzyme active site. The top interaction is between residues R71 and G117, which are not directly interacting in the 3D structure, but form opposite ends of the nucleotide binding pocket (Fig 2d). Seven of the top ten residue-residue interactions involve one of these key sites. The top 20 interactions are provided in S2 Table.

The Potts model interaction scores can be used to identify residue pairs that are contacting in the three-dimensional protein structure. The residue pairs with the top $L/2$ (93) Frobenius scores correspond to 10 structural contacts at a distance of less than 5 Å and 21 contacts at a distance of less than 8 Å (Fig 2e). We compared our model's contact prediction performance to established methods for inferring contacts from evolutionary data, including the standard DCA modeling procedure [19] and also a modified DCA model that weights sequences based on their distance from wild-type to account for the evolutionary process [22] (Fig 2f and S3 Table). When trained on the round 15 data, neither of these methods were able to outperform the random chance expectation (2.79%) when predicting the top $L/2$ (93) long range contacts at a distance of less than 5 Å. In contrast, our model, which considered the evolutionary trajectory, was able to correctly identify 10 out of the top $L/2$ contacts (10.7%), which is well above random chance. Our model also outperforms the other two methods by recovering more structural contacts with heavy atoms 5–8 Å and less than 8 Å (Fig 2f and S3 Table).

In silico evolutionary simulations to map the global structure of the fitness landscape and extrapolate evolutionary trajectories

Our statistical method infers the underlying fitness landscape from experimental laboratory evolution trajectories. This model can be used to run *in silico* simulations to understand the landscape, evolutionary processes, and engineer new proteins.

We used our model to understand the global structure of the fitness landscape and the convergence of adaptive evolutionary walks. We uniformly sampled 1600 random amino acid

sequences (length 186) across all 20 amino acids to obtain broad sampling of the landscape. For each of these sequences, we performed an adaptive walk by evaluating all single mutants, selecting the most fit variant, and repeating this process until a local fitness peak was reached. We found every single adaptive evolutionary trajectory converged to the same fitness peak and the sequence at this fitness peak is 79 amino acid substitutions from wild-type DHFR. The fact that adaptive walks starting from diverse regions of the landscape converge to the same peak implies a Mt. Fuji-type fitness landscape with few local optima.

We also used the learned model to continue the experimental DHFR evolution process and extrapolate the evolutionary trajectory to future generations (Fig 3b). We started the

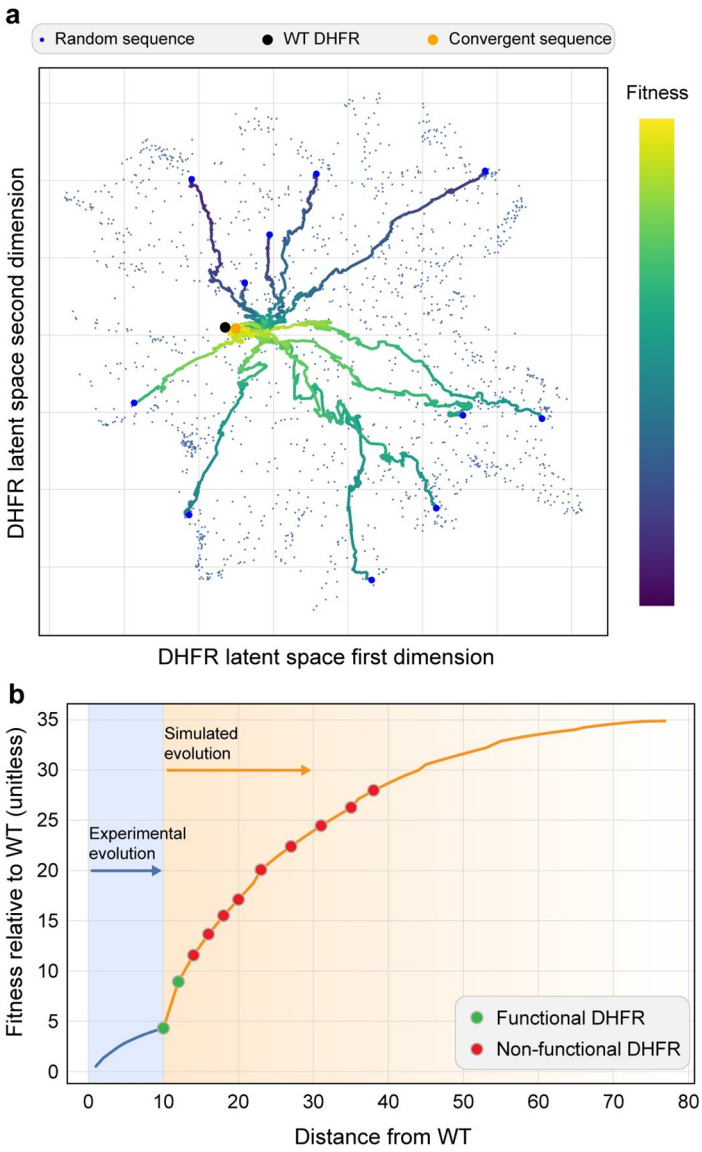

Fig 3. Evolutionary simulations to understand the landscape structure. (a) We uniformly sampled random amino acid sequences of length 186 and performed an adaptive walk until a fitness peak was reached. All adaptive walks converged to the same fitness peak. 11 representative evolutionary trajectories are shown. (b) We continued our laboratory DHFR evolution by continuing the evolutionary process *in silico*. We experimentally tested several DHFR sequences along the evolutionary trajectory and found many were inactive enzymes.

https://doi.org/10.1371/journal.pcbi.1010956.g003

simulation with the most common amino acid sequence observed in the final round of the laboratory evolution experiment. This starting sequence is 10 amino acid substitutions from wild-type DHFR. We performed an adaptive walk by evaluating all single and double mutants, selecting the most fit variant, and repeating this process until there were no further uphill steps. The simulated evolutionary trajectory continued to move away from wild-type DHFR and the round 15 starting sequence and converged to the same fitness peak as found in the global landscape search described above.

We wanted to test whether the directed evolution trajectories could be extrapolated *in silico* as an approach to engineer new proteins. We experimentally characterized ten DHFR sequences along the trajectory from the round 15 sequence to the adaptive fitness peak (Fig 3b). The first five sequences were picked consecutively along this trajectory and the next five sequences were picked by skipping every second sequence. We did not sample the entire trajectory. We found the round 15 sequence was an active DHFR enzyme, a double mutant of the round 15 sequence was also active, but all sequences beyond that were inactive and unable to complement *E. coli* growth in the presence of trimethoprim (Fig 3b and S3 Fig).

Discussion

Directed laboratory evolution is a powerful approach to explore the protein fitness landscape. Directed evolution generates data consisting of sequences sampled over multiple generations and provides valuable clues about the underlying fitness landscape structure. Directed evolution data does not naturally fit into established supervised or unsupervised learning methods because they do not consider the evolutionary data generation process. In this work, we developed a statistical learning framework to infer the protein fitness landscape from laboratory evolution data. We built a generative model of the directed evolution process, parameterized the fitness landscape using a generalized Potts model, and inferred the landscape parameters from sequencing data from multiple rounds of evolution. We applied this learning approach to a large and diverse dihydrofolate reductase (DHFR) directed evolution data set. The inferred landscape model revealed numerous examples of epistasis arising from interactions between residues, but an overall global fitness peak that is evolutionarily accessible from most starting sequences. Finally, we explored the potential of the landscape model to extrapolate evolutionary trajectories for protein engineering.

Our laboratory evolution model made two important approximations for the sake of computational tractability. Understanding when these approximations are valid or invalid can help identify the scope and limitations of our method. The first approximation makes the assumption the evolving population has an infinite population size, while in practice all laboratory evolution has a finite population size dictated by experimental constraints. Our actual DHFR experimental population sizes are on the order of 10^5–10^6 and this number is largely determined by the library transformation efficiency into E. coli. With larger population sizes, this approximation gets more accurate.

The second approximation makes the assumption that the population evolves according to simplified Markov chain dynamics (S1 Text) where competition only occurs between sequences who share a direct ancestor. In our actual experiment all sequences are directly competing in the growth selection regardless of their lineage (S4 Fig). The Markov chain dynamics approximation is valid when competing sequences have similar fitness at each stage of the experiment, which is likely if the fitness landscape near the initial wild-type sequence is flat (i.e., most selected mutations are neutral) and the evolution experiment is carried out at lower mutation rates. At low mutation rates the laboratory evolution experiment only explores the wild-type sequence's local neighborhood and all selected mutant sequences will have similar

fitness due to the neutral landscape. In addition, the Markov chain dynamics are a better approximation during the earlier rounds of evolution, while all evolving sequences are still near the initial wild-type sequences.

Once we specified the evolution dynamics, we inferred fitness landscape parameters by approximate moment matching. The approximate recursive formulas for the first and second order moments between round r and $r + 1$ were derived based on a weak dependence assumption between variables of interest with the remaining variables, which allow us to approximate marginals at $r + 1$ round only based on local information in the protein residue graph. This approximation is only exact if the residue graph factorizes, but we expect this provides a reasonable approximation for most of the nodes and/or pair of nodes where the strength of interaction between them and the other variables is limited. We have three main regularization hyperparameters in our inference procedure. The first two influence the magnitude of the main effects and the pairwise effects parameters. These can be set in the same way as they are for the DCA model [17]. The third hyperparameter is the regularization parameter that infers the canonical parameters from the mean parameters. We observed the best performance when it was set to be several times larger than the negative of the smallest eigenvalue in the estimated covariance matrix. Of these three parameters, the results were most sensitive to the last parameter.

Our inferred DHFR landscape parameters were able to identify contacting residues in the three-dimensional protein structure and also numerous functionally coupled residues that indirectly interact through the enzyme active site. Our trajectory learning approach showed significant improvements over established contact prediction methods at the commonly used $L/2$ threshold for residues with heavy atoms closer than 5 Å. This improvement is due to modeling the evolutionary trajectory as well as including sequencing data from earlier rounds of evolution. Several of top interactions (S2 Table) inferred are in residues that are not close in the 3D structure but are in functional regions that interact with either the substrate or the cofactor. These interactions are marked as false positives for contact prediction, however they are likely epistatic interactions arising through ligand interactions or cofactor repositioning [27]. 17 out of the top 20 interacting residue pairs in S2 Table involve a tryptophan residue. Although this represents repeat interactions with just three residues (25W, 58W and 114W), it is possible that this over-representation is from a bias in the model assumptions about the error-prone PCR transition probabilities. However, it is known that tryptophan residues play an important role in structure and stability of E. coli DHFR [28] and specifically Residue 25 is highly conserved in the natural DHFR sequences and has a role in the function of the protein [29]. So an alternative explanation is that these tryptophan residues show up in S1 Table because of their importance.

Previous work has demonstrated the ability to identify residue contacts by applying DCA to the final generation of laboratory evolution experiments [21, 22]. Stiffler et al. applied a reweighing procedure to DCA to account for evolutionary bias and were able to get contact recovery rates exceeding 50% for the top $L/2$ contacts. Previous laboratory experiments using TEM1 β-lactamase [30] were not able to detect local epistatic interactions, however a larger and more recent experiment [21] was able to identify epistatic interactions and detect some contacts using DCA. It's notable that these DCA methods were unable to reliably detect DHFR contacts from our round 15 evolution data (S5 Fig). We applied standard DCA and the reweighting DCA methods of [22] to our round 15 DHFR data and when predicting the top $L/2$ contacts closer than 5 Å in the 3D structure, the percentage of correctly labeled contacts was lower than random chance expectation. There are a number of factors that could contribute to the differences in contact recovery across these data sets, including population sequence diversity, the number of variants sequenced, or the evolution mutation rate. Stiffler et al. provide an

analysis of down-sampled data sets for aminoglycoside acetyl-transferase (AAC6) [22], so we can directly compare to our mDHFR data. At 10^5 sequences, they have an average pairwise distance of 10.9% and identify nearly 40% of the contacts. In contrast, with our mDHFR data with 10^5 sequences, we have an average pairwise distance of 11.1%, but only recover 2.15% of the contacts. Based on these findings, it seems the sequence diversity and data set size are not contributing to differences in contact recovery. The two experiments do differ significantly in their mutation rate, where AAC6 0.8% per round, while DHFR was evolved at 0.4% per round. We hypothesize this varied mutation rate is resulting in different population structures, despite having similar sizes and levels of diversity. Even for a fixed neutral landscape, finite populations following the quasispecies dynamics with different mutation rates will evolve different levels of robustness [31] and this robustness is linked to epistatic interactions [32]. Simulations of these types of laboratory evolution experiments using the simplified markov chain dynamics is provided by [33] and these simulations do not consider the mutation rate as an important parameter. However, since the laboratory evolution experiment likely follows the more complicated quasispecies dynamics, it is possible that the mutation rate plays a more important role.

We used the inferred model parameters to explore epistasis and the global structure of the fitness landscape. We found many examples of additive mutations, approximately 0.4% of mutations that interact through sign epistasis, and only thirteen examples of reciprocal sign epistasis. This frequency of epistatic interactions is consistent with other studies [34, 35]. Despite the landscape epistasis, we found the estimated landscape structure had a overall global fitness peak that was accessible by adaptive walks from most starting sequences. Similar fitness landscape features have been observed in other proteins [35, 36], however, due to the various approximations and limitations in the model's ability to generalize outside its training data, we caution against interpreting the model's estimated landscape features as accurately existing in the actual protein fitness landscape.

We used the inferred Potts model to run an *in silico* directed evolution experiment to design new, previously unobserved DHFR variants. We found the model could design functional DHFRs that were close to the training data regime, but further evolutionary extrapolation resulted in nonfunctional enzymes. This result is consistent with other machine learning based protein engineering studies that show decreased model accuracy while extrapolating away from the training data [10, 37]. This result also shows that the prediction of a fitness peak at 79 mutations is not reliable. The model inaccuracy could be the result of insufficient or low-quality data, approximations we made in the evolutionary model, or computational challenges with parameter estimation. A more accurate evolutionary model would consider finite population sizes and competition between all sequences in a generation, which is referred to as the finite quasispecies model [31, 38]. Another possibility to improve model accuracy would be to use a simpler first order model without interactions so that we need to estimate fewer parameters. Another approach to improve the reliability for protein engineering would be to run multiple independent evolution simulations and test a panel of diverse designs.

Our statistical landscape inference approach naturally complements recent advances in continuous directed evolution [39–41]. These experimental methods combine population level mutagenesis and selection in continuously-fed bioreactors to evolve populations without discrete mutation/selection steps. The populations can be sampled and analyzed by next-generation DNA sequencing to observe how the population changes over time and traverses the fitness landscape. Our learning method could infer the landscape from this sequential evolution data to understand protein structure, function, and evolution.

The relationships between protein sequence, structure, and function involve thousands of exquisite molecular interactions that are dynamically coupled over space and time. Machine

learning is revolutionizing our understanding of these relationships by dissecting the complex inner workings of proteins with a scale and resolution beyond human comprehension. Future advances in data-driven protein science will improve our ability to understand natural evolutionary processes, predict genetic disease, and design new proteins for broad applications in biotechnology.

Materials and methods

Overview of statistical method

We model the dynamics of laboratory evolution as a Markov chain process where sequences transition to other sequences according to their mutational accessibility and relative fitness. We consider the set of all possible codon sequences Ω of length L, where each sequence is denoted by $x \coloneqq (x_1, \ldots, x_L)$. Each x_i corresponds to the codon that encodes the ith residue position and each codon is from the set $\mathcal{C} \coloneqq \{\mathrm{ATT}, \mathrm{ATC}, \ldots\}$ of 61 codons that exclude the stop codons. We let $\Pi(x)$ represent the fitness of each sequence $x \in \Omega$, i.e., the number of copies that a sequence makes of itself per *unit* time, and $\pi(x)$ the corresponding prevalence based on relative fitness, defined as $\pi(x) = \Pi(x)/\sum_{u\in\Omega}\Pi(u)$.

Under the assumptions of (S1 Text), we have a transition probability given by

$$p(x \to y) \coloneqq \frac{g(x \to y)\pi(y)}{\sum_{z\in\Omega} g(x \to z)\pi(z)}. \tag{1}$$

where $g(x \to y)$ represents the probability sequence x mutating to sequence y in the absence of selection. Each sequence $x^{(n,\,r)}$ from the round r is then a random sample from the marginal probability $p^{(r)}$ after the rth step transition.

We parameterize the fitness landscape π with a (generalized) Potts model that describes how all amino acid residues and their pairwise interactions contribute to fitness. We parameterize the fitness level $\pi(x) = \pi_\theta(x)$ of sequence x with a Potts model with canonical parameter set on the amino acids $\theta \coloneqq [\{h_i(a)\}_{i\in[L],a\in\mathcal{A}}, \{e_{ij}(a,b)\}_{i,j\in[L],i<j,a,b\in\mathcal{A}}]$ where

$$\pi_\theta(x) = \frac{1}{Z(\theta)}\exp\left\{\sum_{i=1}^{L} h_i(AC(x_i)) + \sum_{i=1}^{L}\sum_{i<} e_{ij}(AC(x_i), AC(x_j))\right\} \tag{2}$$

AC$(\cdot)$ is the mapping from the set of codons $\mathcal{C}$ to the set to amino acids $\mathcal{A} = [A, V, \ldots]$, and $Z(\theta)$ is the normalization constant so that the probabilities sum to 1. This model consists of $q_a L$ main effect parameters (h_i), where $q_a \coloneqq |\mathcal{A}| = 20$, and $q_a^2\binom{L}{2}$ (couplings) interaction effect parameters (e_{ij}).

Directed evolution data consists of sequences sampled from multiple rounds of evolution. These observed sequences $x^{(n,\,r)}$, $n = 1, 2, \ldots, n_r$ at each round r are random realizations from the probability distribution $p^{(r)}(\cdot;\theta)$. We can estimate the Potts model parameters θ by maximizing the following log-likelihood

$$\log\ \mathrm{Lik}(\theta) = \sum_{r\in\mathcal{R}}\sum_{n=1}^{n_r} \log\ p^{(r)}(x^{(n,r)};\theta). \tag{3}$$

where $\mathcal{R}$ denotes the set of experimental rounds at which sequencing data exists and $p(x \to y;\theta)$ in (1) is given by the Markov chain dynamics under the Potts model.

It is challenging to identify the model parameters θ that maximize this log-likelihood function (3). First, even though π_θ forms a Markov Random Field (MRF), the distribution of $p^{(r)}$ no longer factorizes with respect to the graph associated with the fitness distribution π_θ. In

particular, the conditional independence relationships which hold for π_θ do not hold for $p^{(r)}$, $r = 1, 2, \ldots$, which prohibits applying techniques for parametric inference in Markov Random Field settings. In addition, the large dimension of the state space precludes any exact tracing of this Markov chain process. For instance, the dimension of the transition matrix is $|\Omega| \times |\Omega|$, and even computing a single element in the transition matrix for a given θ is computationally infeasible due to the intractable sum in the denominator in (1) (as $|\Omega| = 61^{186}$).

We used an approximate moment matching method to overcome these computational challenges. In particular, we first derived approximate relationships between the first and second order marginals $\mu_i^{(r)}, \mu_{ij}^{(r)}$ at each round r and the marginals μ under π_θ (Sec. Approximate relationships between marginals). Then we sought for $\hat{\mu}$ which matches empirical and expected 1st and 2nd order marginals and is also locally compatible (Sec. Inferring mean parameters). Lastly, we used mean-field DCA [15] using estimated marginals as inputs to get estimates of parameters θ of the Potts model (2) (see estimating canonical parameters in S1 Text).

We use the estimates of the canonical parameters θ to get information about the protein's structure and function. Each parameter $e_{ij}(a, b)$ represents the interaction between amino acid a at residue i and amino acid b at residue j. We compute an interaction score between residues i and j using the canonical parameters e in the same manner as the DCA method [17]. The parameter set e is over parameterized and so we first we convert to the zero-sum gauge and then we compute the Frobenius norm

$$e'_{ij}(a, b) = e_{ij}(a, b) - e_{ij}(\cdot, b) - e_{ij}(a, \cdot) + e_{ij}(\cdot, \cdot) \quad \text{and} \quad F_{ij} = \|e'_{ij}\|.$$

The Frobenius norm F_{ij} is the interaction score between pairs of residues i and j. We only look at long range interactions (i.e. between residues that differ by more than 5 positions in the amino acid sequence).

A schematic diagram with an overview of our method is given in Fig 4. Although the framework we have put together in this way to infer fitness landscapes does not appear in the literature, several components are from previous work. With the Potts model parameterization (2), we can look at the simplified dynamics given by (1) as similar to the evolution dynamics given in [33]. A generalization in our framework is that we allow for the possibility multiple simultaneous mutations where [33] only use a single mutation scheme for efficient implementation of simulations. Also, the component on estimating canonical Potts model parameters follows the fairly standard mean-field DCA [15] method. The components that deal with moment matching (Sec. Approximate relationships between marginals) and (Sec. Inferring mean parameters) should be considered new contributions.

Software code to reproduce this method are available at github.com/RomeroLab/dhfr_neutral_evolution and archived at DOI 10.5281/zenodo.7622051.

Approximate relationships between marginals

We first approximate first and second order marginals $\mu_i^{(r)}, \mu_{ij}^{(r)}$ with a function of first and second order marginals μ of π_θ. Once the approximated moment vectors for all available rounds are obtained, the mean vector μ for π_θ is then estimated by matching the approximated moment vectors with the empirical counts.

We let $X^{(r)}$ be a random variable following the probability distribution $p^{(r)}$. $P(X^{(r)} = x) \coloneqq p^{(r)}(x)$. For any codons $c, d \in \mathcal{C}$, first we define the first and second order marginals at round r as $\mu_i^{(r)}(c) \coloneqq \mathbb{E}[\delta(X_i^{(r)}, c); \theta]$, $\mu_{ij}^{(r)}(c, d) \coloneqq \mathbb{E}[\delta(X_i^{(r)}, c)\delta(X_j^{(r)}, d); \theta]$ where $\delta(\cdot, \cdot)$ is the Kronecker delta function such that $\delta(a, b) = 1$ if $a = b$ and 0 otherwise. We let the moment vector as a

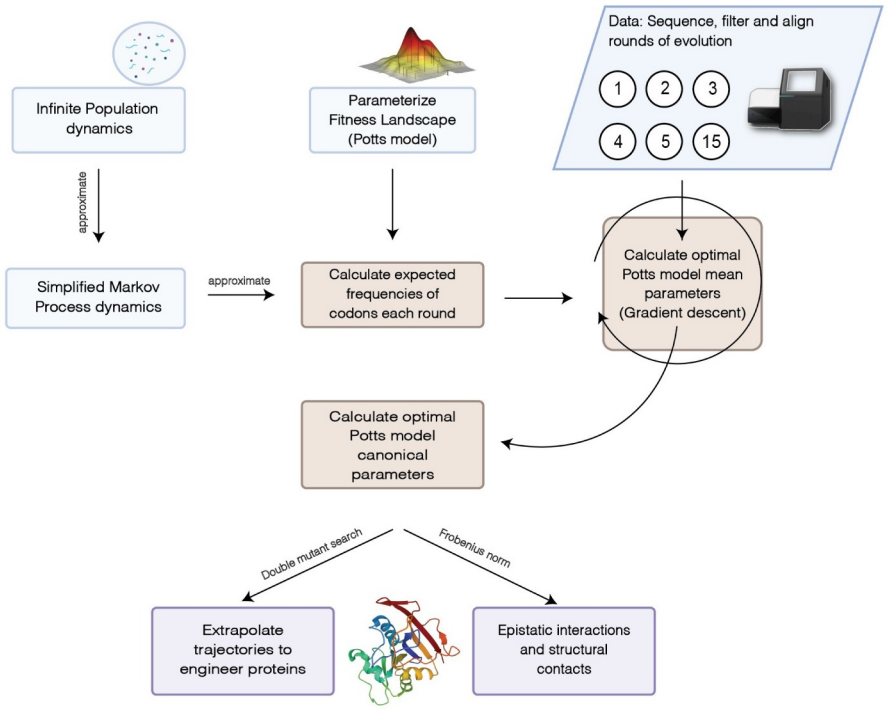

Fig 4. An overview of the algorithm to infer fitness landscape parameters.

https://doi.org/10.1371/journal.pcbi.1010956.g004

collection of these terms

$$\mu^{(r)} \coloneqq [\{\mu_i^{(r)}(c)\}_{i\in[L],c\in\mathcal{C}}, \{\mu_{ij}^{(r)}(c,d)\}_{i,j\in[L],i<j,c,d\in\mathcal{C}}].$$

Similarly, we also define the first and second order marginals corresponding to the fitness landscape π_θ as $\mu_i(c) \coloneqq \mathbb{E}_\pi[\delta(X_i,c);\theta]$, $\mu_{ij}(c,d) \coloneqq \mathbb{E}_\pi[\delta(X_i,c)\delta(X_j,d);\theta]$, as well as the moment vector μ for π_θ as

$$\mu \coloneqq [\{\mu_i(c)\}_{i\in[L],c\in\mathcal{C}}, \{\mu_{ij}(c,d)\}_{i,j\in[L],i<j,c,d\in\mathcal{C}}].$$

If we have sequencing data at round r, we can calculate the individual and pairwise codon frequencies observed for that round by summing over the data.

$$f_i^{(r)}(c) = \frac{1}{n_r}\sum_{n=1}^{n_r}\delta(x_i^{(n,r)},c); \quad f_{ij}^{(r)}(c,d) = \frac{1}{n_r}\sum_{n=1}^{n_r}\delta(x_i^{(n,r)},c)\delta(x_j^{(n,r)},d)$$

We can put these terms together in a frequency vector $f^{(r)}$

$$f^{(r)} \coloneqq [\{f_i^{(r)}(c)\}_{i\in[L],c\in\mathcal{C}}\{f_{ij}^{(r)}(c,d)\}_{i,j\in[L],i<j,c,d\in\mathcal{C}}\}],$$

and when it can be calculated from the data it gives us an estimate of the moment vector $\mu^{(r)}$ for round r.

The fitness landscape itself is never observed but only the dynamics of the experiment over the fitness landscape is observed. So, unlike the moment vectors for the rounds where we have sequencing data and we calculate $f^{(r)}$ as an estimate of $\mu^{(r)}$, we do not have an estimate for the moment vector of the fitness landscape μ. To get around this, for each node i or a pair of nodes (i,j), we approximate the prevalence of codon c at i and the prevalence of the codon pair (c,d)

at i, j at $r+1$ in terms of the moment vector $\mu^{(r)}$ of round r and the moment vector of the fitness landscape μ. In particular, assuming that the dependence between a set of nodes of interest X_S and the remaining variables X_{S^c} are limited for the approximation of $\mu_S^{(r+1)}$, we derive the following recursive relationships in S1 Text:

$$\begin{aligned}\mu_i^{(r+1)}(c) &\approx \sum_{c' \in \mathcal{C}} \frac{g_i(c' \to c)\mu_i(c)\mu_i^{(r)}(c')}{\sum_{c'' \in \mathcal{C}} g_i(c' \to c'')\mu_i(c'')} \\ \mu_{ij}^{(r+1)}(c,d) &\approx \sum_{c',d' \in \mathcal{C}} \frac{g_i(c' \to c)g_j(d' \to d)\mu_{ij}(c,d)\mu_{ij}^{(r)}(c',d')}{\sum_{c'',d'' \in \mathcal{C}} g_i(c' \to c'')g_j(d' \to d'')\mu_{ij}(c'',d'')}.\end{aligned} \tag{4}$$

We note that once $\mu_i^{(0)}$ is specified, all mean vectors $\mu_i^{(r)}$ are subsequently specified as functions of μ via the recursive relations in (4). Since at the start of the experiment, we have only copies of the wild-type sequence present (denoted by $w \in \Omega$) and so its frequency is 1 and the frequency of all other sequences is 0. Therefore, $\mu_i^{(0)}(c) = 1$ and $\mu_{ij}^{(0)}(c,d) = 1$ if $w_i = c$ and $w_j = d$ and it is 0 in all other cases.

Now, we estimate μ by minimizing the aggregated log-loss between the expected first and second-order frequencies $\mu_i^{(r)}$, $\mu_{ij}^{(r)}$ (as functions of μ) and the observed first and second-order frequencies $f_i^{(r)}, f_{ij}^{(r)}$ for all positions $i, j \in [L]$, $i \neq j$. In doing so, we reparametrize μ so that we can optimize the objective function over amino-acid level parameters. We also enforce a local consistency condition, and regularize parameters to prevent overfitting during optimization (Sec. Inferring mean parameters).

Inferring mean parameters

Here, we describe the approximate moment-matching by minimizing the aggregated log-losses between the expected and observed first and second-order frequencies. In other words, we would like to solve the following objective function:

$$\arg\min_{\mu \in \mathcal{M} \cap \mathcal{U}} \left\{ -\sum_{r \in \mathcal{R}} \sum_{i \in [L]} \sum_{c \in \mathcal{C}} f_i^{(r)}(c) \log\, \mu_i^{(r)}(c;\mu) - \sum_{r \in \mathcal{R}} \sum_{\substack{i,j \in [L] \\ i \neq j}} \sum_{c,d \in \mathcal{C}} f_{ij}^{(r)}(c,d)\, \log\, \mu_{ij}^{(r)}(c,d;\mu) \right\}, \tag{5}$$

where $\mathcal{M} \coloneqq \{\mu \in \mathbb{R}^d; \exists p \text{ such that } \mathbb{E}_p[\phi(X)] = \mu\}$ and $\phi(x) \coloneqq [\{\delta(x_i, c)\}_{i \in [L], c \in \mathcal{C}}, \{\delta(x_i, c)\delta(x_j, d)\}_{i,j \in [L], i<j, c,d \in \mathcal{C}}]$ is a sufficient statistic for π_θ, and $\mathcal{U} \coloneqq \{\mu \in \mathbb{R}^d; \mu_i(c) = \mu_i(c')\, \forall i \in [L], \forall c, c' \in \mathcal{C}$ such that $AC(c) = AC(c')$ and $\mu_{ij}(c,d) = \mu_{ij}(c',d')\, \forall i,j \in [L]\, \forall c, c', d, d' \in \mathcal{C}$ such that $AC(c) = AC(c')$ and $AC(d) = AC(d')\}$ In words, the set $\mathcal{M}$ corresponds to the set of the globally consistent mean vectors, i.e., all first-order and pairwise marginal probabilities that can be realized by some distribution over $\{0, 1\}^d$ where d is the dimension of the sufficient statistic ϕ, and the set $\mathcal{U}$ corresponds to the set of mean vectors such that the mean values only depend on the amino acid values of an input sequence.

First, to optimize the objective over amino-acid level parameters, we reparameterize the mean parameters μ of the Potts model on Ω in terms of $\gamma \coloneqq [\{\gamma_i(a)\}_{i \in [L], a \in \mathcal{A}}, \{\gamma_{ij}(a,a')\}_{i,j \in [L], i<j, a,a' \in \mathcal{A}}]$ as

follows:

$$\mu_i(c;\gamma) = \frac{v_i(\mathrm{AC}(c);\gamma)}{\sum_{c'\in\mathcal{C}} v_i(\mathrm{AC}(c');\gamma)} \quad \text{and} \quad v_i(a;\gamma) = \frac{\exp\{\gamma_i(a))\}}{\sum_{a'\in\mathcal{A}}\exp\{\gamma_i(a')\}}$$
$$\mu_{ij}(c,d;\gamma) = \frac{v_{ij}(\mathrm{AC}(c),\mathrm{AC}(d);\gamma)}{\sum_{c',d'\in\mathcal{C}} v_{ij}(\mathrm{AC}(c'),\mathrm{AC}(d');\gamma)} \quad \text{and} \quad v_{ij}(a,b;\gamma) = \frac{\exp\{\gamma_{ij}(a,b))\}}{\sum_{a',b'\in\mathcal{A}}\exp\{\gamma_{ij}(a',b')\}}. \tag{6}$$

We also define $v_{ji}(b, a; \gamma) = v_{ij}(a, b; \gamma)$ for $j > i$.

Although the set $\mathcal{M}$ can be characterized by a finite number of linear inequalities, the number of linear inequalities grows fast depending on the dimension d, and in general, it is known to be extremely difficult to optimize even a linear objective over $\mathcal{M}$ unless the dimension d is small [42]. We proceed by considering the relaxation of the optimization problem (5) by enforcing normalization conditions

$$\sum_{a\in\mathcal{A}} v_i(a) = 1, \sum_{a,b\in\mathcal{A}} v_{ij}(a,b) = 1, \ \forall i,j\in[L], i\neq j \tag{7}$$

and local consistency conditions

$$v_i(a) = \sum_{b\in\mathcal{A}} v_{ij}(a,b), \forall a\in\mathcal{C}, \ \forall i,j\in[L], i\neq j. \tag{8}$$

Note (7) is satisfied by the reparameterization (6). We add a following penalty term $\mathcal{P}(\gamma)$ with a Lagrange multiplier ρ to promote the local consistency conditions (8)

$$\mathcal{P}(v) = \sum_{i=1}^{L}\sum_{a\in\mathcal{A}} \|v_i(a) - \frac{1}{L-1}\sum_{j\neq i}\sum_{b\in\mathcal{A}} v_{ij}(a,b)\|$$

Finally, to handle the high-dimensionality of γ, we add ℓ_2-regularization terms $\mathcal{R}_{\mathrm{main}}$ and $\mathcal{R}_{\mathrm{int}}$ with hyper parameters λ_{main} and λ_{int}

$$\mathcal{R}_{\mathrm{main}}(\gamma) = \sum_{i\in[L]} \|\gamma_i\|^2 \quad \text{and} \quad \mathcal{R}_{\mathrm{int}}(\gamma) = 2\sum_{i,j\in[L];i<j} \|\gamma_{ij}\|^2.$$

In summary, we solve the following optimization problem:

$$\hat{\gamma} = \underset{\gamma_i(a);\gamma_{ij}(a,b)\in\mathbb{R}}{\arg\min} \{ -\sum_{r\in\mathcal{R}}\sum_{i\in[L]}\sum_{c\in\mathcal{C}} f_i^{(r)}(c)\log\ \mu_i^{(r)}(c;\gamma) - \sum_{r\in\mathcal{R}}\sum_{\substack{i,j\in[L]\\ i\neq j}}\sum_{c,d\in\mathcal{C}} f_{ij}^{(r)}(c,d)\log\ \mu_{ij}^{(r)}(c,d;\gamma) + \lambda_{\mathrm{main}}\mathcal{R}_{\mathrm{main}}(\gamma) + \lambda_{\mathrm{int}}\mathcal{R}_{\mathrm{int}}(\gamma) + \rho\mathcal{P}(\gamma)\}. \tag{9}$$

We initialize the parameters γ for optimization to be the log of the pairwise frequencies of the last round that we have sequencing data for. A small pseudocount is added to the frequencies so that this initialization procedure is defined for missing frequencies. We compute first order derivatives of the objective using automatic differentiation. Also, we use a gradient descent optimizer together with an early stopping rule.

After optimization, we get the individual and pairwise estimates $\hat{\gamma}$ and can compute an estimate for the mean parameters $\hat{v}, \hat{\mu}$ using equation (6).

Statistical framework parameters

The statistical inference method was implemented in Python using PyTorch [43]. The optimization method used the Adam optimizer with a learning rate of 0.03 with other learning

parameters set to default and trained for 300 steps. The regularization hyperparameters in (9) were set to $\lambda_{main} = 10^{-3}$, $\lambda_{int} = 10^{-4}$ and $\lambda_{reg} = 50$. The penalty term to ensure the main effect parameters marginalized to the pairwise parameters was set to $\rho = 10^5$.

The exact mutation distribution in this experiment is not known as all rounds were sequenced after growth selection, however, the statistical method seems fairly robust to choice of the mutation distribution as the results look similar with mutation bias distributions from other experiments done with a similar protocol (results not shown). We picked a mutation bias distribution to model mutagenesis from the *Taq* DNA polymerase column of Table 2 in [44], [45] and then scaled to match an average of 4 DNA mutations per round. The final mutation distribution used in the statistical method is given in S4 Table.

mDHFR laboratory evolution

Our selection strain consists of the murine dihydrofolate reductase (mDHFR) gene cloned into the pET22b plasmid and transformed into *E. coli* BL21(DE3). We performed error-prone PCR using the reaction's $MnCl_2$ concentration to tune the mutation rate of Taq DNA polymerase [46]. We determined that a final concentration of 200 μM $MnCl_2$ yielded 3.25±0.74 amino acid substitutions per gene. We performed 15 error-prone PCR cycles, treated the reaction with DpnI overnight to remove template, purified the PCR product with a DNA spin column (Zymo Research), cloned the insert back into pET-22b using circular polymerase extension cloning (CPEC) [47], purified the CPEC reaction using a DNA spin column (Zymo Research), and transformed the CPEC reaction into electrocompetent BL21(DE3) cells (Lucigen). Several dilutions of the transformation were plated to determine the total library size, which was in the range of 10^5–10^6 colony forming units (CFUs). The remainder of the transformation was used as input to a competitive growth selection in 100 mL of LB containing 100 μg/mL carbenicillin, 500 μM IPTG, and 5 μg/mL trimethoprim. We allowed these selection cultures to grow shaking for 16 hours at 37˚C. Approximately 20 ODU of overnight culture was used to harvest plasmid DNA via miniprep. This selected plasmid DNA population was then used as a template for the next round of error-prone PCR. A portion of the post-selection culture was also archived as 15% glycerol stocks and stored at -80˚C.

We determined the fraction of functional variants from each round of evolution by picking colonies from each transformation plate into individual wells of a 96-well plate containing LB broth, 100 μg/mL carbenicillin, 500 μM IPTG, and 5 μg/mL trimethoprim. We incubated these plate cultures shaking for 16 hours at 37˚C, measured the OD600 of each well, and categorized DHFR variants as functional if their OD600 was greater than 0.5, otherwise they were considered nonfunctional.

DNA sequencing of evolved populations. We performed next-generation DNA sequencing on several rounds of laboratory evolution. We analyzed rounds 1–5 using Illumina sequencing and round 15 using Pacific Biosciences sequencing. For the Illumina libraries, we used NdeI and SacI restriction enzymes to remove the DHFR insert from the plasmid, ligated Illumina adaptor sequences to this insert, and submitted the samples to the UW-Madison Biotechnology Center DNA Sequencing Core to run on an Illumina MiSeq instrument using the 2x300 v3 kit. Each sample had 2–5 million reads. For the PacBio sequencing, we removed the DHFR insert with XbaI/PsiI and submitted the samples to the UW-Madison Biotechnology Center DNA Sequencing Core for analysis on their Pacific Biosciences Sequel instrument. The PacBio run returned over 10^5 reads. Raw sequencing data is available at the NCBI Sequence Read Archive (SRA Accession ID PRJNA923701).

Sequence pre processing. For rounds 1–5, the Illumina sequencing data are processed with steps similar to [22]. The forward and reverse reads are first stitched together using the

FLASH program [48]. At the first filtering step, only sequences with a minimum length of 500 and a minimum quality score of 15 for every base were retained. At the second step, sequences with a compound quality score of at least 10 were retained, implying a 90% probability of having no read errors. For round 15, quality filtering was done to keep sequences with a minimum length of 564 and a minimum quality score of 0.99. The remaining sequences were then aligned to the reference sequence using bowtie2 [49].

Experimentally testing evolutionary designed DHFRs. We designed ten DHFR variants using the inferred model to simulate an evolutionary process. The genes encoding these variants were synthesized by Twist Bioscience and cloned into the pET21(+) plasmid. We transformed the plasmids into *E. coli* BL21(DE3) and performed growth measurements to assess the DHFR variant's activity. We performed the growth assays by first inoculating a 5 ml LB starter culture containing 100 μg/mL carbenicillin and growing overnight shaking at 37˚C. We then diluted this starter culture 100x into an LB culture containing 100 μg/mL carbenicillin, 500 μM IPTG, and 5 μg/mL trimethoprim and monitored growth by measuring the OD600 in 30 min. intervals over a 16.5 hour incubation period at 37˚C. These measurements were carried out in triplicate. Inactive DHFR variants displayed no growth under these conditions, while active variants displayed standard growth curves.

Supporting information

S1 Fig. Average Hamming distance from wild-type and average pairwise Hamming distance between sequences per round.
(PDF)

S2 Fig. Mutation effect prediction on natural DHFR sequences.
(PDF)

S3 Fig. Growth curves of sequences designed using the inferred model.
(PDF)

S4 Fig. Diagram of an Idealized Experiment vs Real Experiment.
(PDF)

S5 Fig. DCA methods trained on round 15 DHFR evolution data.
(PDF)

S1 Table. Estimate of Colony Formining Units and Fractional Functional.
(PDF)

S2 Table. The top 20 long range interaction scores between pairs of residues.
(PDF)

S3 Table. Positive predictive value (PPV) of long range contacts recovered by different methods.
(PDF)

S4 Table. Error-prone PCR mutation bias.
(PDF)

S1 Text. Supplementary Methods and Mathematical Details.
(PDF)

S2 Text. VAE architecture and hyperparameters.
(PDF)

Author Contributions

Conceptualization: Sameer D'Costa, Emily C. Hinds, Hyebin Song, Philip A. Romero.

Data curation: Sameer D'Costa, Emily C. Hinds, Chase R. Freschlin, Philip A. Romero.

Formal analysis: Sameer D'Costa, Emily C. Hinds, Hyebin Song.

Funding acquisition: Philip A. Romero.

Investigation: Sameer D'Costa, Emily C. Hinds, Hyebin Song, Philip A. Romero.

Methodology: Sameer D'Costa, Emily C. Hinds, Chase R. Freschlin, Hyebin Song.

Project administration: Philip A. Romero.

Resources: Sameer D'Costa, Emily C. Hinds.

Software: Sameer D'Costa.

Supervision: Hyebin Song, Philip A. Romero.

Validation: Sameer D'Costa, Emily C. Hinds, Chase R. Freschlin.

Visualization: Sameer D'Costa.

Writing – original draft: Sameer D'Costa, Philip A. Romero.

Writing – review & editing: Sameer D'Costa, Emily C. Hinds, Hyebin Song, Philip A. Romero.

References

1. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nat Methods. 2019; 16(8):687–694. https://doi.org/10.1038/s41592-019-0496-6 PMID: 31308553
2. Ferguson AL, Ranganathan R. 100th anniversary of macromolecular science Viewpoint: Data-driven protein design. ACS Macro Lett. 2021; 10(3):327–340. https://doi.org/10.1021/acsmacrolett.0c00885 PMID: 35549066
3. Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cell Syst. 2021; 12(6):654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017 PMID: 34139171
4. Freschlin CR, Fahlberg SA, Romero PA. Machine learning to navigate fitness landscapes for protein engineering. Curr Opin Biotechnol. 2022; 75(102713):102713. https://doi.org/10.1016/j.copbio.2022.102713 PMID: 35413604
5. Wittmann BJ, Johnston KE, Almhjell PJ, Arnold FH. EvSeq: Cost-effective amplicon sequencing of every variant in a protein library. ACS Synth Biol. 2022; 11(3):1313–1324. https://doi.org/10.1021/acssynbio.1c00592 PMID: 35172576
6. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014; 11 (8):801–807. https://doi.org/10.1038/nmeth.3027 PMID: 25075907
7. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol. 2009; 10(12):866–876. https://doi.org/10.1038/nrm2805 PMID: 19935669
8. Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-N protein engineering with data-efficient deep learning. Nat Methods. 2021; 18(4):389–396. https://doi.org/10.1038/s41592-021-01100-y PMID: 33828272
9. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol. 2022; 40(7):1114–1122. https://doi.org/10.1038/s41587-021-01146-5 PMID: 35039677
10. Gelman S, Fahlberg SA, Heinzelman P, Romero PA, Gitter A. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. Proceedings of the National Academy of Sciences. 2021; 118(48):e2104878118. https://doi.org/10.1073/pnas.2104878118 PMID: 34815338
11. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021; 118 (15):e2016239118. https://doi.org/10.1073/pnas.2016239118 PMID: 33876751

12. Trinquier J, Uguzzoni G, Pagnani A, Zamponi F, Weigt M. Efficient generative modeling of protein sequences using simple autoregressive models. Nat Commun. 2021; 12(1):5800. https://doi.org/10.1038/s41467-021-25756-4 PMID: 34608136

13. Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, et al. Expanding functional protein sequence spaces using generative adversarial networks. Nature Machine Intelligence. 2021; 3(4):324–333. https://doi.org/10.1038/s42256-021-00310-5

14. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. 2009; 106(1):67–72. https://doi.org/10.1073/pnas.0805923106 PMID: 19116270

15. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

16. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PloS one. 2011; 6(12):e28766. https://doi.org/10.1371/journal.pone.0028766 PMID: 22163331

17. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707 PMID: 23410359

18. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences. 2013; 110(39):15674–15679. https://doi.org/10.1073/pnas.1314045110

19. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nature biotechnology. 2017; 35(2):128–135. https://doi.org/10.1038/nbt.3769 PMID: 28092658

20. Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, et al. An evolution-based model for designing chorismate mutase enzymes. Science. 2020; 369(6502):440–445. https://doi.org/10.1126/science.aba3304 PMID: 32703877

21. Fantini M, Lisi S, De Los Rios P, Cattaneo A, Pastore A. Protein structural information and evolutionary landscape by in vitro evolution. Molecular biology and evolution. 2020; 37(4):1179–1192. https://doi.org/10.1093/molbev/msz256 PMID: 31670785

22. Stiffler MA, Poelwijk FJ, Brock KP, Stein RR, Riesselman A, Teyra J, et al. Protein structure from experimental evolution. Cell Systems. 2020; 10(1):15–24. https://doi.org/10.1016/j.cels.2019.11.008 PMID: 31838147

23. Fernandez-de Cossio-Diaz J, Uguzzoni G, Pagnani A. Unsupervised inference of protein fitness landscape from deep mutational scan. Molecular biology and evolution. 2021; 38(1):318–328. https://doi.org/10.1093/molbev/msaa204 PMID: 32770229

24. Sohail MS, Louie RH, Hong Z, Barton JP, McKay MR. Inferring epistasis from genetic time-series data. Molecular biology and evolution. 2022; 39(10):msac199. https://doi.org/10.1093/molbev/msac199 PMID: 36130322

25. Salverda ML, Dellus E, Gorter FA, Debets AJ, Van Der Oost J, Hoekstra RF, et al. Initial mutations direct alternative pathways of protein evolution. PLoS genetics. 2011; 7(3):e1001321. https://doi.org/10.1371/journal.pgen.1001321 PMID: 21408208

26. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021; 599(7883):91–95. https://doi.org/10.1038/s41586-021-04043-8 PMID: 34707284

27. Miton CM, Buda K, Tokuriki N. Epistasis and intramolecular networks in protein evolution. Current opinion in structural biology. 2021; 69:160–168. https://doi.org/10.1016/j.sbi.2021.04.007 PMID: 34077895

28. Ohmae E, Sasaki Y, Gekko K. Effects of five-tryptophan mutations on structure, stability and function of Escherichia coli dihydrofolate reductase. The Journal of Biochemistry. 2001; 130(3):439–447. https://doi.org/10.1093/oxfordjournals.jbchem.a003004 PMID: 11530021

29. Beard WA, Appleman JR, Huang S, Delcamp TJ, Freisheim JH, Blakley RL. Role of the conserved active site residue tryptophan-24 of human dihydrofolate reductase as revealed by mutagenesis. Biochemistry. 1991; 30(5):1432–1440. https://doi.org/10.1021/bi00219a038 PMID: 1991124

30. Bershtein S, Goldin K, Tawfik DS. Intense neutral drifts yield robust and evolvable consensus proteins. Journal of molecular biology. 2008; 379(5):1029–1044. https://doi.org/10.1016/j.jmb.2008.04.024 PMID: 18495157

31. Van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. Proceedings of the National Academy of Sciences. 1999; 96(17):9716–9720. https://doi.org/10.1073/pnas.96.17.9716 PMID: 10449760

32. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006; 444(7121):929–932. https://doi.org/10.1038/nature05385 PMID: 17122770

33. Bisardi M, Rodriguez-Rivas J, Zamponi F, Weigt M. Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. Molecular biology and evolution. 2022; 39(1):msab321. https://doi.org/10.1093/molbev/msab321 PMID: 34751386

34. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Current biology. 2014; 24(22):2643–2651. https://doi.org/10.1016/j.cub.2014.09.072 PMID: 25455030

35. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. 2016; 5:e16965. https://doi.org/10.7554/eLife.16965 PMID: 27391790

36. Weinreich DM, Delaney NF, DePristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science. 2006; 312(5770):111–114. https://doi.org/10.1126/science.1123539 PMID: 16601193

37. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, et al. Deep diversification of an AAV capsid protein by machine learning. Nature Biotechnology. 2021; 39(6):691–696. https://doi.org/10.1038/s41587-020-00793-4 PMID: 33574611

38. Eigen M. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften. 1971; 58(10):465–523. https://doi.org/10.1007/BF00623322 PMID: 4942363

39. Esvelt KM, Carlson JC, Liu DR. A system for the continuous directed evolution of biomolecules. Nature. 2011; 472(7344):499–503. https://doi.org/10.1038/nature09929 PMID: 21478873

40. Ravikumar A, Arzumanyan GA, Obadi MK, Javanpour AA, Liu CC. Scalable, continuous evolution of genes at mutation rates above genomic error thresholds. Cell. 2018; 175(7):1946–1957. https://doi.org/10.1016/j.cell.2018.10.021 PMID: 30415839

41. Halperin SO, Tou CJ, Wong EB, Modavi C, Schaffer DV, Dueber JE. CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. Nature. 2018; 560(7717):248–252. https://doi.org/10.1038/s41586-018-0384-8 PMID: 30069054

42. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Now Publishers Inc; 2008.

43. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alch -Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024–8035. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

44. Stratagene, Agilent Technologies. GeneMorph II Random Mutagenesis Kit; 2009. Available from: https://www.chem-agilent.com/pdf/strata/200552.pdf.

45. Shafikhani S, Siegel RA, Ferrari E, Schellenberger V. Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. Biotechniques. 1997; 23(2):304–310. https://doi.org/10.2144/97232rr01 PMID: 9266088

46. Romero PA, Tran TM, Abate AR. Dissecting enzyme function with microfluidic-based deep mutational scanning. Proceedings of the National Academy of Sciences. 2015; 112(23):7159–7164. https://doi.org/10.1073/pnas.1422285112 PMID: 26040002

47. Quan J, Tian J. Circular polymerase extension cloning of complex gene libraries and pathways. PloS one. 2009; 4(7):e6441. https://doi.org/10.1371/journal.pone.0006441 PMID: 19649325

48. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011; 27(21):2957–2963. https://doi.org/10.1093/bioinformatics/btr507 PMID: 21903629

49. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4):357. https://doi.org/10.1038/nmeth.1923 PMID: 22388286